Communications in Applied Mathematics and Computational Science

Volume 1 No. 1 2006

mathematical sciences publishers

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

PROBLEM REDUCTION, RENORMALIZATION, AND MEMORY

ALEXANDRE J. CHORIN AND PANAGIOTIS STINIS

We present methods for the reduction of the complexity of computational problems, both time-dependent and stationary, together with connections to renormalization, scaling, and irreversible statistical mechanics. Most of the methods have been presented before; what is new here is the common framework which relates the several constructions to each other and to methods of theoretical physics, as well as the analysis of the approximate reductions for time-dependent problems. The key conclusions are: (i) in time dependent problems, it is not in general legitimate to average equations without taking into account memory effects and noise; (ii) mathematical tools developed in physics for carrying out renormalization group transformations yield effective block Monte Carlo methods; (iii) the Mori–Zwanzig formalism, which in principle yields exact reduction methods but is often hard to use, can be tamed by approximation; and (iv) more generally, problem reduction is a search for hidden similarities.

1. Introduction

There are many problems in science that are too complex for numerical solution as they stand. Examples include turbulence, molecular dynamics, and other problems where multiple scales must be taken into account. Such problems must be reduced to more amenable forms before one computes. In the present paper we would like to summarize some of the reduction methods that have been developed in recent years, together with an account of what was learned in the process. It is obvious that the problem has not been fully solved, but we think that the examples and the conclusions reached so far are useful.

In general terms, a reduction to a more amenable form is a renormalization group transformation, as in physics — a transformation of a problem into a more tractable form while keeping quantities of interest invariant. A renormalization group transformation involves an incomplete similarity transformation, and thus a

MSC2000: 65C20, 65Z05, 82B80, 76F30.

Keywords: problem reduction, renormalization, irreversible statistical mechanics, memory, Monte Carlo.

This work was supported in part by the National Science Foundation under Grant DMS 04-32710, and by the Director, Office of Science, Computational and Technology Research, U.S. Department of Energy under Contract No. DE-AC03-76SF000098.

reduction method is a search for hidden similarities. This is a general feature of reduction methods, and it will be illustrated in the examples. A successful problem reduction produces a new problem which must in some asymptotic sense be similar to the original problem. For general background on renormalization, see, e.g., [5; 20; 39].

In problems with strong time dependence, reduction methods resemble methods for the analysis of thermodynamic systems not in equilibrium; indeed, those aspects of the problem that are ignored in a reduced description conspire to destroy order and increase entropy. Problem reduction for time-dependent problems is basically renormalization group theory for non-equilibrium statistical mechanics. For background on such theory, see, e.g., [3; 22; 8; 44].

The content of the paper is as follows: In section 2 we consider Hamiltonian systems and their conditional expectations. In section 3 we narrow the discussion to statistically stationary Hamiltonian systems and recover Kadanoff real-space renormalization groups and an interesting block Monte Carlo method. In section 4 we display an example that exhibits and also extends the main features of this analysis in simple form.

In section 5 we explain the Mori–Zwanzig formalism for the reduction of statistically time-dependent problems. The analysis shows that averaging the equations is in general not enough; one must take into account noise and a temporal memory. The Mori–Zwanzig formalism is rather dense, and in the sections that follow we present various special cases in which it can be simplified, in particular when the memory is very short or very long. We wish to draw the reader's attention in particular to the "t-model", for which we present a new analysis; it seems to us that it represents a step forward in modeling for a relatively small price in added computational complexity.

One of our goals in exploring the connections between problem reduction and irreversible statistical mechanics is to point out some of the places where the knowledge acquired in statistical mechanics still awaits its proper integration into computational practice.

The paper [21] is a survey of reduction methods organized along different lines and can be profitably read in tandem with the present paper.

For the sake of readability, we remind the reader of the rudiments of similarity theory [3]. Suppose a variable a is a function of variables $a_1, a_2, \dots, a_m, b_1, b_2, \dots, b_k$, where $a_1, \dots, a_m$ have independent units, for example units of length and mass, while the units of $b_1, \dots, b_k$, can be formed from the units of $a_1, a_2, \dots, a_m$. Then there exist dimensionless variables $\Pi = \frac{a}{a_1^{\alpha_1} \cdots a_m^{\alpha_m}}$, $\Pi_i = \frac{b_i}{a_1^{\alpha_{i1}} \cdots a_m^{\alpha_{im}}}$, $i = 1, \dots, k$, where the α_i, α_{ij} are simple fractions, such that Π is a function of the Π_i:

$$\Pi = \Phi(\Pi_1, \dots, \Pi_k). \tag{1}$$

This is just a consequence of the requirement that a physical relationship be independent of the size of the units of measurement. At this stage nothing can be said about the function Φ. Now suppose the variables Π_i are small or large, and assume that the function Φ has a non-zero finite limit as its arguments tend to zero or to infinity; then $\Pi \sim$ constant, and one finds a power monomial relation between a and the a_i. This is a complete similarity relation. If the function Φ does not have the assumed limit, it may happen that for Π_1 small or large, $\Phi(\Pi_1) = \Pi_1^\alpha \Phi_1(\Pi_1) + \cdots$, where the dots denote lower order terms, α is a constant, the other arguments of Φ have been omitted and Φ_1 has a finite non-zero limit. One can then obtain a scaling expression for a in terms of the a_i and b_i, with undetermined powers which must be found by means other than dimensional analysis. The resulting power relation is an incomplete similarity relation. Of course one may well have functions Φ with neither kind of similarity.

Incomplete similarity expresses what is invariant under a renormalization group; all renormalization group transformations involve incomplete similarity. The exponent α is called an anomalous exponent.

2. Averaging a Hamiltonian system

We begin by examining what happens when one tries to reduce the complexity of a Hamiltonian system by averaging (see also [15; 16; 38; 2]). This first section is partially historical – this is how our group in Berkeley started working on problem reduction; part of this development has been superseded by the theory in the section on the Mori–Zwanzig formalism below. It seems to us that this is still the right place to start, because the conclusions here explain the (less than intuitively obvious) need to go beyond averaging to a more complicated theory, and also because the theory in this section is the basis for the analysis of the stationary case in the two sections that follow.

Consider a system of nonlinear ordinary differential equations,

$$\begin{aligned} \frac{d}{dt}\varphi(t) &= R(\varphi(t)), \\ \varphi(0) &= x, \end{aligned} \tag{2}$$

where φ and x are n-dimensional vectors with components φ_i and x_i, and R is a vector-valued function with components R_i; t is time. To each initial value x in (2) corresponds a trajectory $\varphi(t) = \varphi(x, t)$.

Suppose that we only want to find m of the n components of the solution vector $\varphi(t)$ without finding the $n - m$ others. One has to assume something about the variables that are not evaluated, and we assume that at time t=0 we have a joint probability density $f(x)$ for all the variables. The variables we keep will have definite initial values $x_1, x_2, \ldots, x_m$, and the rest of variables will then have a conditional probability density $f_m = f(x_1, \ldots, x_m, x_{m+1}, \ldots)/Z_m$, where $Z_m = \int_{-\infty}^{+\infty} f(x_1, \ldots, x_m, x_{m+1}, \ldots) dx_{m+1} dx_{m+2} \cdots$ is a normalization constant. Without some assumption about the missing variables the problem is meaningless; this particular assumption is reasonable because in practice f can often be estimated from previous experience or from general considerations of statistical mechanics. The question is how to use this prior knowledge in the evaluation of $\varphi(t)$.

Partition the vector x so that $\hat{x} = (x_1, x_2, \ldots, x_m)$, $\tilde{x} = (x_{m+1}, \ldots, x_n)$ and $x = (\hat{x}, \tilde{x})$, and similarly $\varphi = (\hat{\varphi}, \tilde{\varphi})$, $R = (\hat{R}, \tilde{R})$. In general, the first m components of R depend on all the components of φ, $\hat{R} = \hat{R}(\varphi) = \hat{R}(\hat{\varphi}, \tilde{\varphi})$; if they do not we have a system of m equations in m variables and nothing further needs to be done. We want to calculate only the variables $\hat{\varphi}$; then $(d/dt)\hat{\varphi}(t) = \hat{R}(\varphi(t))$ where the right-hand side depends on the variables $\tilde{\varphi}$ which are unknown at time t. We shall call the variables $\hat{\varphi}$ the "resolved variables" and the remaining variables $\tilde{\varphi}$ the "unresolved variables".

Consider in particular a Hamiltonian system as in [15],[16]. There exists then by definition a Hamiltonian function $H = H(\varphi)$ such that for i odd R_i, the i-th component of the vector R in (2) satisfies $R_i = \partial H \big/ \partial \varphi_{i+1}$, while for i even, one has $R_i = -\partial H \big/ \partial \varphi_{i-1}$, with n, the size of the system, even. Assume furthermore that f, the initial probability density, is $f(\varphi) = Z^{-1} \exp(-H/T)$ where T is a parameter, known in physics as the "temperature", which will be set equal to one in much, but not all, of the discussion below. In physics this density appears naturally and is known as the "canonical" density; the normalizing constant $Z = Z(T)$ is the "partition function". This density f is invariant, i.e., sampling it and evolving the system in time commute.

A numerical analyst who wants to approximate the solution of an equation usually starts by approximating the equation. If one solves for the resolved variables one has values for the variables $\hat{\varphi}$ available at each instant t and the best approximation should be a function of these variables; it is natural to seek a best approximation in the mean square sense with respect to the invariant density f at each time; the best approximation in this sense is the conditional expectation $E[R(\varphi)|\hat{\varphi}] = \int Re^{-H} d\tilde{\varphi} \big/ \int e^{-H} d\tilde{\varphi}$ (note that we set $T = 1$ for simplicity). This conditional expectation is the orthogonal projection of R onto the space of functions of $\hat{\varphi}$ with respect to the inner product $(u, v) = E[uv] = \int u(\varphi) v(\varphi) f(\varphi) d\varphi$, where $d\varphi$ denotes integration over all the components of φ. We then try to approximate the system (2) by:

$$\begin{aligned} \frac{d}{dt}\hat{\varphi}(t) &= E[R(\varphi(t))|\hat{\varphi}(t)], \\ \hat{\varphi}(0) &= \hat{x}. \end{aligned} \tag{3}$$

It has been shown in [13; 15; 11] that: (i) the new system (3) is also Hamiltonian:

$$E\left[\frac{\partial H}{\partial \varphi_i}|\hat{\varphi}(t)\right]=\int\frac{\partial H}{\partial \varphi_i}\exp(-H)d\tilde{\varphi}\Big/\int\exp(-H)d\tilde{\varphi}=\frac{\partial \hat{H}}{\partial \varphi_i}, \tag{4}$$

where $i \leq m =$ the dimension of $\hat{\varphi}$, and

$$\hat{H}=-log\int\exp(-H)d\tilde{\varphi} \tag{5}$$

is the new Hamiltonian.

(ii) the new canonical density $\hat{f}=Z^{-1}\exp(-\hat{H})$ is invariant in the evolution of the new, reduced, system.

(iii) when the data are sampled from the canonical distribution, the distribution of $\hat{\varphi}$ in the new system is its marginal distribution in the old system; equivalently, the partition function Z is the same for the old system and for the new system.

Now the question is, what does the solution $\hat{\varphi}(t)$ of (3) represent? Having averaged the equations, one could hope that the result is an average of the solution, of course constrained by the initial data $\hat{x}$, i.e., that the solution of equations (3) is $E[\hat{\varphi}(t)|\hat{x}]$. This is the case for linear systems (where averaging and time integration commute), and is approximately the case for limited time in some other special situations – nearly linear systems and some systems where the "unresolved variables" are fast. However, in general this is not the case. On the other hand, the solution of equations (3) does not approximate the true values of $\hat{\varphi}(t)$ in the full system either – the latter depend strongly on the missing data $\tilde{x}$ while the former does not. We shall see below that a reduced description of the solution of nonlinear systems in time requires in general a "noise" (which describes the fluctuations in $\tilde{\varphi}(t)$) and a "memory" (which depends on the temporal fluctuations of the noise and on the history of the solutions).

The fact that the solution of the averaged equations is not the average of the solutions can be understood by the following physics argument. In physics, a system in which the values of all the variables are drawn from a canonical distribution is a system in thermal equilibrium. The assignment of definite values $\hat{x}$ to the variables $\hat{\varphi}$ at time $t=0$ amounts to taking the system out of equilibrium at $t=0$; if the system is ergodic it will then decay to equilibrium in time, so that all the variables become randomized and acquire the joint density f. Thus the predictive value of the partial initial data $\hat{x}$ decreases in time; all averages of the $\hat{\varphi}$ approach equilibrium averages. However, the reduced system (3) is Hamiltonian, and the solutions it produces oscillate forever.

In Figure 1 we consider the Hald Hamiltonian system [13] with

$$H=\frac{1}{2}\left(\varphi_1^2+\varphi_2^2+\varphi_3^2+\varphi_4^2+\varphi_1^2\varphi_3^2\right) \tag{6}$$

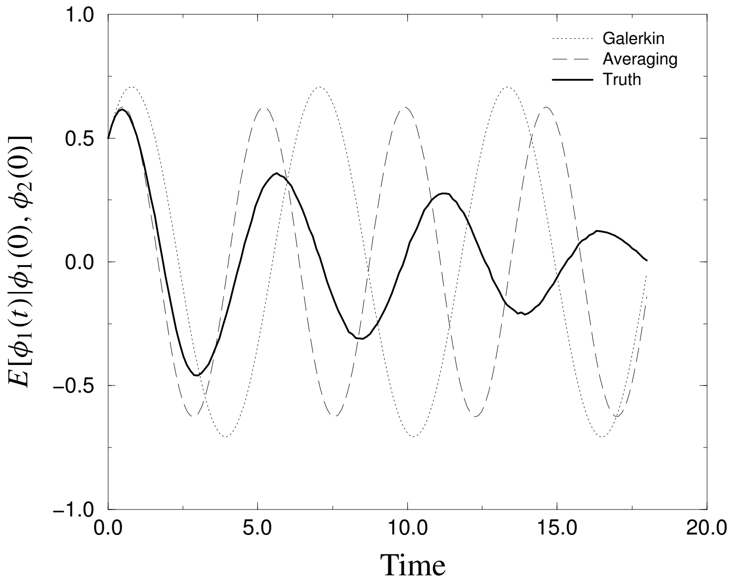


Figure 1. Comparison of the evolution of $E[\phi_1(t)|\phi_1(0), \phi_2(0)]$ (truth), to the prediction by the "Galerkin" approximation and the prediction by the averaging procedure described in the text.

(physically, two linear oscillators with a nonlinear coupling). We assume that $\varphi_1(0), \varphi_2(0)$ are given and sample the two other initial data from the canonical distribution with $T = 1$.

Figure 1 displays (1) the result for φ_1 of a "Galerkin" calculation in which the unresolved variables are set to zero (this is what is implicitly done in many unresolved computations); (2) the result of the averaging procedure just described, and (3) the true $E[\varphi_1(t)|\hat{x}]$, calculated by repeatedly sampling the initial data, solving the full system, and averaging. As one can see, averaging is initially better than the null "Galerkin" method, but in the long run the truth decays but the solution of the averaged system oscillates forever. For more detail, see [13].

Consider now the current practice of "large-eddy simulation" in hydrodynamics (see, e.g., [31]). One defines there as $\hat{\varphi}(t)$ "filtered" (i.e., locally averaged) variables and one finds for the time evolution of these variables equations obtained by relating various averaged terms in the Navier–Stokes equations to the filtered variables at one time. The result can be exactly equivalent to equations (3), as in [30], or indeed it could be an even worse approximation, because the conditional expectation of R is the best approximation of R by a function of the $\hat{\varphi}$. One should consider the possibility that some of the well-known difficulties of large-eddy simulation are due to basic flaws in this procedure, and we will offer a possible alternative below. For a description of special cases, with small fluctuations and particular structures, where the use of equations (3) is legitimate, see [21].

3. Prediction with no data and block Monte Carlo

There is, however, a case where the construction of the preceding section can be very useful – when one tries to predict the future with no initial information. All the data are then sampled from the canonical density, which is invariant. If the system is ergodic, the solutions of equations (2) sample the space of solutions and their time average equals their average with respect to the canonical density. The system then simply samples the canonical density, and the reduction by conditional expectation of the previous section creates a smaller system whose variables have the same probability density after reduction as they had before reduction, and can be sampled at lower cost. This is the starting point for some interesting analysis as well as for block sampling methods (see [38; 2] for applications to molecular dynamics).

To see in detail what the reduction by conditional expectations of the previous section accomplishes under these circumstances, suppose the variables φ_i are associated with nodes on a regular lattice, for example, they may represent spins in a solid, or originate in the spatial discretization of a partial differential equation.

Divide the lattice into blocks of some fixed shape (for example, divide a regular one-dimensional lattice into groups of two contiguous nodes). We have not yet specified how the variables are to be divided into resolved and unresolved. Now decide to "resolve" one variable per block, and leave the others in the same block unresolved. The transformation between the old variables and the smaller set of resolved variables is a Kadanoff renormalization group transformation exactly as the latter are defined in [28] even if the steps which lead to it are presented differently; the Hamiltonian $\hat{H}$ defined above in equation (5) is the renormalized Hamiltonian in the sense of Kadanoff. This is an easy instance of our general claim that problem reduction is renormalization.

Suppose the system described by the Hamiltonian is translation invariant. The equations of motion at any one point, say at the location labeled by 1, have the same form as the equations of motion at any other point. The relation between the right-hand side of the reduced system and the right-hand side of the old system can be rewritten as:

$$\frac{\partial \hat{H}}{\partial \varphi_1} = E[\frac{\partial H}{\partial \varphi_1}|\hat{\varphi}], \tag{7}$$

where the expected value is with respect to the invariant density as before. This relation is the starting point for the evaluation of $\hat{H}$.

The key to success is to expand H and $\hat{H}$ in series, so that the calculation of the conditional expectations becomes easier for each term than it is for the Hamiltonians themselves. We use here a version of what is known in physics as an expansion in successive couplings (see [28]). The Hamiltonians are functions of the variables φ and can be expanded in the form:

$$H = \sum_j a_j \psi_j, \tag{8}$$

where the ψ_j are "elementary Hamiltonians". In a translation invariant system, where each equation has the same form as any other, the Hamiltonian is made up of sums over i of terms of the form $h(\varphi_i \varphi_{i+j})$ for various values of j, where h is some function; these terms represent "couplings" between variables j apart; one can then choose the elementary Hamiltonians to be polynomials in $x_i x_{i+j}$ with a fixed j in each ψ_j, i.e., one segregates the couplings between variables j apart into separate terms.

In a homogeneous system where there is only one variable per site, it is enough to satisfy (7) for one variable, say for φ_1. Define $\psi'_j = \frac{\partial}{\partial \varphi_1} \psi_j$, noting that though each ψ_j for a homogeneous system is necessarily a function with at least as many arguments as there are components on φ, ψ'_j can be sparse in the sense that it depends only on a few of the variables (for example, if $\psi_0 = \sum_i \varphi_i^2$, then $\psi'_0 = 2\varphi_1$). Equation (7) reduces to

$$\frac{\partial \hat{H}}{\partial \varphi_1} = \sum_j a_j P \psi'_j(\varphi) \tag{9}$$

with the projection P defined as before by $Pg(\varphi) = E[g|\hat{\varphi}]$ for any function g of φ. Now we're almost done. Pick a basis in $\hat{L}_2$, the subspace of square integrable functions that depend only on the variables $\hat{\varphi}$, made up of a subset of the set of functions ψ'_j. The right-hand side of equation (9) is then again a linear combination of ψ'_j; integration with respect to φ_1 requires only the erasure of the primes and yields a series for $\hat{H}$. The elements of $\tilde{\varphi}$ are now gone, and one can relabel the remaining variables $\hat{\varphi}$ so that the terms in the series have exactly the same form as before; the calculation can then be repeated, yielding a sequence of Hamiltonians with ever fewer variables: H, $H^{(1)} = \hat{H}$, $H^{(2)} = \hat{H}^{(1)}, \ldots$. The corresponding densities $f^{(n)} = Z^{-1} \exp(-H^{(n)}/T)$ can be sampled by any sampling scheme, for example, by Metropolis sampling (see, e.g., [10]).

At this point we have reduced the number of variables by a factor L equal to the number of variables in each block, but this may well seem to be a Pyrrhic victory. The Hamiltonians one usually encounters are simple in the sense that they involve few couplings – finite differences typically link a few neighboring variables, and so do the usual spin Hamiltonians in physics. As one reduces the number of variables, the new Hamiltonians become more complex, with more terms in the series (8); the cost per time step of solving the equations in time or the cost per move in a Metropolis sampling typically increases quickly as well. To see what has been gained one must again turn to the physics literature (see, e.g., [28],[24]).

Consider the spatial correlation length ℓ which measures the range of values of $|j|$ over which the spatial covariances $E[\varphi_i\varphi_{i+j}]$ are non negligible, and the correlation time τ for which the temporal covariances $E[\varphi_i(t)\varphi_j(t+s)]$ are non-negligible. For very large and very small values of the temperature T (the variance parameter in the density f) both the correlation time and the correlation length are usually small (see [28],[17]); the properties of the system can then be found from calculations with a small number of variables and it is not urgent to reduce the number of variables. There is a range of intermediate values of T for which the correlation length and time are large and then the reduction is worthwhile. There often is a value T_c of T, the "critical value", for which $\ell = \infty$. Values of T around T_c are often of great interest.

Now we can see what the reduction can accomplish. If one tries to compute averages with T near T_c one finds that the cost of computation is proportional to τ and to some positive power of ℓ – one has to compute long enough to obtain independent samples of φ, and a new independent sample will not appear until a time $\sim \tau$ has passed. The reductions above produce a system with smaller ℓ and τ and therefore computation takes less time. Though we started with the declared goal of reducing the number of variables, what has been produced is more interesting: a new system with shorter correlations which is more amenable to computation. It is not the raw number of variables that matters. It is important to notice that what started as a scheme for winnowing out variables has ended up by producing a new system related to the original system by a scaling transformation.

The renormalization can be used with a multigrid scheme, in which one runs up and down on different levels of renormalization, on the finer ones to achieve accuracy and the cruder ones to move fast from one macroscopic configuration to another. It is well known that multigrid schemes require that one store conditional expectations (see, e.g., [7]), and the physicists' expansion in successive linkages provides an effective way to do so; for details see [10],[35].

An alternative method for obtaining the expansion coefficients for the renormalized Hamiltonians was proposed in [42]. The method is based on the maximization of the likelihood of the renormalized density. The maximization of the likelihood leads to a moment-matching problem. The moments in this case are the expectation values of the "elementary Hamiltonians" (see above) with respect to the renormalized density. The solution of the moment matching problem yields the expansion of the renormalized Hamiltonian.

The systematic development of the links of probability with renormalization began with Jona–Lasinio (see, e.g., [26]). The connection of renormalization with incomplete similarity is too well known (see [3; 28; 22]) to require further comment here. The analysis of this section provides a striking example of the benefits to be found in applying to computation ideas drawn from experience in statistical physics.

4. An example: The Korteveg–deVries–Burgers equation

As a further illustration of the ideas in the previous section, consider the equation

$$u_t + uu_x = \epsilon u_{xx} - \beta u_{xxx}, \tag{10}$$

with boundary conditions

$$u(-\infty) = u_0, \quad u(+\infty) = 0, \quad u_x(-\infty) = 0, \tag{11}$$

where the subscripts denote differentiation, x is the spatial variable, t is time, $\epsilon > 0$ is a diffusion coefficient, $\beta > 0$ is a dispersion coefficient, and $u_0 > 0$ is a given constant. The boundary conditions create a traveling wave solution moving to the right (towards $+\infty$) with velocity $u_0/2$ which becomes steady in a moving framework as $t \to \infty$. In nondimensional form the equation can be written as:

$$u_t + uu_x = \frac{1}{R}u_{xx} + u_{xxx}, \tag{12}$$

with $u_x(-\infty) = 0$, $u(+\infty) = 0$, $u(-\infty) = 1$; $R = \sqrt{\beta u_0}/\epsilon$ is a "Reynolds number". For $R \leq 1$ the traveling wave has a monotonic profile, while for $R > 1$ the profile is oscillatory, with oscillations whose wave length is of order 1 [6]. At zero diffusion ($R = \infty$) the stationary asymptotic wave train extends to infinity on the left. For finite R the wave train is damped and the solution tends to 1 as x decreases.

The steady wave profile can be found by noting that it satisfies an ordinary differential equation, whose solution connects a spiral singularity at $x = -\infty$ to a saddle point at $x = +\infty$. At the steady state we average the solution at each point x over the region $(x - \ell/2, x + \ell/2)$ and call the result $\bar{u}$. The task we set ourselves is to find an effective equation $g(v, v_x, v_{xx}, \ldots) = 0$ whose solution v approximates $\bar{u}$; v can be expected to be smoother than the solution of (12) and thus require fewer mesh points for an accurate numerical solution; this is analogous to finding a renormalized Hamiltonian further from the critical point so that the solution of the corresponding problem has lower fluctuations, as we did in the previous section; note that the problem of this present section is not Hamiltonian.

We now make an analogy between the conditional expectations which define the renormalized variables in the previous sections and an averaging in space which defines "renormalized" variables for solutions of the KdVB equations that are stationary in a moving frame. Averaging over an increasing length scale corresponds either to more renormalization steps or, equivalently, to renormalization with a greater number of variables grouped together. We pick a class of equations in which to seek the "effective" equation, the one whose solutions best approximate the averages of the true solution in the mean square sense; the choice of mean-square approximation in the KdVB case corresponds to the use of L_2 norms implied by the use of conditional expectations in the previous sections, and the choice of a class of equations in which to look for the effective equation is analogous to the choice of a basis for the representation of the Hamiltonian; the calculation of the best coefficients in the chosen class of "effective" equations corresponds to the evaluation of the coefficients in the series for the renormalized Hamiltonians. In the Hamiltonian case we average the right-hand sides of the equations and in the analogous KdVB case we attempt to average the solutions; this must be so because in the KdVB case we do not have theorems which guarantee that averaging the right-hand sides produces the correct statistics for the solutions.

We can look for an effective equation in the class of equations of the form

$$-cv_x + vv_x = \epsilon_{eff} v_{xx} + v_{xxx} + \beta |v_x|^{\alpha} v_{xx} + \cdots , \tag{13}$$

where $\epsilon \geq 0$, $\alpha \geq 0$, $\beta \geq 0$ are constants and $c = 1/2$ is the velocity of propagation of the steady wave (see also [4]). This expansion is analogous to the expansion in successive linkages (8) of the previous section; in a continuum limit, a series of partial Hamiltonians, whose derivatives have larger and larger "stencils" across which variables are connected, can be reorganized into an expansion in higher and higher derivatives of the unknown. One knows a priori that u and v propagate at the same velocity, which helps fix some of the parameters (i.e., expansion coefficients) at the outset. The problem is to find the values of the parameters in the effective equation which minimize

$$I = \int_{-\infty}^{+\infty} |\bar{u}(x) - v(x)|^2 dx. \tag{14}$$

One finds numerically that the last terms have little effect on the minimum of I when $\ell \geq 5$ (in physics terminology, they are "irrelevant"). The effective equation is thus a Burgers equation with a value of the dimensionless diffusion coefficient ϵ_{eff} different from $1/R$.

The minimization in (14) was carried out in [9], and it showed that the minimum was achieved when $\epsilon_{eff} = R^{\nu}\Phi(\ell)$, with the exponent $\nu \sim 0.75$. Note that when the diffusion coefficient $\epsilon \to 0$, then $\epsilon_{eff} \to \infty$! This is an incomplete similarity relation, as advertised, relating a "bare" Reynolds number R to a "dressed" Reynolds number ϵ_{eff}^{-1}. The form of the effective equation could conceivably have been found by averaging the original equation, but the relation between the original ϵ and ϵ_{eff} requires some form of renormalization-like reasoning.

5. The Mori–Zwanzig formalism

We now return to the problem we started investigating in section 2: how to determine the evolution of a subset $\hat{\varphi}$ of components of a vector φ described by a nonlinear set of equations of the form (2). This is a nonlinear closure problem of a type much studied in physics, and a variety of formalisms is available for the job. We choose the Mori–Zwanzig formalism of irreversible statistical mechanics [19; 23; 33; 46; 34], because it homes in on the basic difficulty, which is the description of the memory in the system; the relation of this formalism to other nonlinear formalisms is described in [14]. That a reduced description of a nonlinear system involves a memory should be intuitively obvious: suppose you have $n > 3$ billiard balls moving about on top of a table and are trying to describe the motion of just three; the second ball may strike the seventh ball at a time t_1 and the seventh ball may then strike the third ball at a later time. The third ball then "remembers" the state of the system at time t_1, and if this memory is not encoded in the explicit knowledge of where the seventh ball is at all times, then it has to be encoded in some other way. We are no longer assuming that the system is Hamiltonian nor that we know an invariant density.

It is much easier to do theory for linear equations, and we start by finding a linear equation equivalent to (not approximating!) the system (2). Introduce the linear Liouville operator $L = \sum_{i=1}^{n} R_i(x)\frac{\partial}{\partial x_i}$, and the Liouville equation:

$$\begin{aligned}\frac{\partial}{\partial t}u(x,t) &= Lu(x,t)\\ u(x,0) &= g(x),\end{aligned} \tag{15}$$

with initial data $g(x)$. This is the partial differential equation for which (2) is the set of characteristic equations. One can verify that the solution of the Liouville equation is $u(x,t) = g(\varphi(x,t))$ (see, e.g., [11]). In particular, if $g(x) = x_i$, the solution is $u(x,t) = \varphi_i(x,t)$, the i-th component of the solution of (2). This linear partial differential equation is thus equivalent to the nonlinear system (2). The linearity of equation (15) greatly facilitates the analysis.

Introduce the semigroup notation $u(x,t) = (e^{tL}g)(x) = g(\varphi(x,t))$, where e^{tL} is the evolution operator associated with the operator L; therefore $e^{tL}g(x) = g(e^{tL}x)$, and one can also verify that $e^{tL}L = Le^{tL}$ (this can be seen to be a change of variables formula). Equation (15) becomes

$$\frac{\partial}{\partial t}e^{tL}g = Le^{tL}g = e^{tL}Lg.$$

We suppose that as before we are given the initial values of the m coordinates $\hat{x}$, and that the distribution of the remaining $n-m$ coordinates $\tilde{x}$ is the conditional density, f conditioned by $\hat{x}$, where f is initially given.

We define a projection operator P by $Pg = E[g|\hat{x}]$. The conditioning variables are the initial values of $\hat{\varphi}$; in section 2 the conditioning variables were the values of $\hat{\varphi}(t)$, which are unusable here when we do not know the probability density at time t. Quantities such as $P\hat{\varphi}(t) = E[\hat{\varphi}(t)|\hat{x}]$ are by definition the best estimates of the future values of the variables $\hat{\varphi}$ given the partial data $\hat{x}$ and are often the quantities of greatest interest.

Consider a resolved coordinate $\varphi_j(x,t) = e^{tL}x_j$ $(j \leq m)$, and split its time derivative, $R_j(\varphi(x,t)) = e^{tL}Lx_j$ as follows:

$$\frac{\partial}{\partial t}e^{tL}x_j = e^{tL}Lx_j = e^{tL}PLx_j + e^{tL}QLx_j, \tag{16}$$

where $Q = I - P$. Define $\hat{R}_j(\hat{x}) = (PR_j)(\hat{x})$; the first term is $e^{tL}PLx_j = \hat{R}_j(\hat{\varphi}(x,t))$ and is a function of the resolved components only (but it is a function of the whole vector of initial data). Note that if Q were zero we would recover something that looks like the crude approximation of an earlier section; however the conditioning variables are not the same. We shall see that the term in Q is essential.

We further split the remaining term $e^{tL}QLx_j$. This splitting will bring it into a very useful form: a noise term, and a memory term whose kernel depends on the correlations of the noise term. The fact that such a splitting is possible is the essence of "fluctuation-dissipation" theorems (see, e.g., [29]).

The evolution operators e^{tL} and e^{tQL} satisfy the Duhamel relation

$$e^{tL} = e^{tQL} + \int_0^t e^{(t-s)L}PLe^{sQL}\,ds.$$

Hence,

$$e^{tL}QLx_j = e^{tQL}QLx_j + \int_0^t e^{(t-s)L}PLe^{sQL}QLx_j\,ds. \tag{17}$$

Collecting terms, we find

$$\frac{\partial}{\partial t}e^{tL}x_j = e^{tL}PLx_j + \int_0^t e^{(t-s)L}PLe^{sQL}QLx_j\,ds + e^{tQL}QLx_j \tag{18}$$

The first term on the right-hand side is the Markovian contribution to $\partial_t\varphi_j(x,t)$—it depends only on the instantaneous value of the resolved $\hat{\varphi}(x,t)$. The second term depends on x through the values of $\hat{\varphi}(x,s)$ at times s between 0 and t, and embodies a memory—a dependence on the past values of the resolved variables. Finally, the third term, which depends on full knowledge of the initial conditions x, lies in the null space of P and can be viewed as noise.

It is important to see that equation (18) is an identity. The memory and noise terms have not been added artificially, their presence is a direct consequence of the original equations of motion. However tempting it may be to average equations by taking one-time averages, the results will, in general, be wrong; one must add a memory and a noise as well. Note that the first term in equation (18) is, apart from the change of conditioning variables, the same as the right-hand side in equations (3).

If what is desired is $P\hat{\varphi}(t)$, the conditional expectation of $\hat{\varphi}(t)$ given $\hat{x}$ (the best approximation in the sense of L_2 to $\hat{\varphi}$ given the partial data $\hat{x}$), then one can premultiply equation (18) by P; the noise term then drops out and we find

$$\frac{\partial}{\partial t} Pe^{tL}x_j = Pe^{tL}PLx_j + P\int_0^t e^{(t-s)L}PLe^{sQL}QLx_j\,ds. \tag{19}$$

Even if the system we start with is Hamiltonian, the Langevin equation (18) is not; the memory and the noise allow the system to forget its initial values and decay to “thermal equilibrium” as it should (see section 2).

Let $w(x,t) = e^{tQL}QLx_j$; by definition w, the noise, is a solution of the initial value problem:

$$\begin{aligned} \frac{\partial}{\partial t} w(x,t) = QLw(x,t) \; &= \; Lw(x,t) - PLw(x,t) \\ w(x,0) = QLx_j. & \end{aligned} \tag{20}$$

If for some function $h(x)$, $Ph = 0$, then $Pe^{tQL}h = 0$ for all time t, i.e., e^{tQL} maps the null space of P into itself. The solution of the equations (20) defines the “orthogonal dynamics” for the system (2) with with data $\hat{x}$ and the given joint density for all the data at the initial time. The initial data for the orthogonal dynamics, $QLx_j = (I-P)R_j = R_j - E[R_j|\hat{x}]$ can be thought of as the fluctuations in the initial values of the R_j. The range of the projection P is everything that can be expressed as a function of $\hat{x}$, i.e., everything that can be predicted from the knowledge of $\hat{x}$; one can think of the range of P as the “resolved space”. One can think of the range of Q as the “noise space”. The orthogonal dynamics modify the temporal evolution that starts from QLx_j by continuously removing from the evolutes any component that can be resolved or predicted; the result always remains in the noise space.

We now show that the memory term is a functional of the temporal covariances of the noise (i.e., of covariances of stochastic processes confined to the noise space). To save on writing we restrict ourselves to cases where the operator L is skew-symmetric, i.e, $(Lu, v) = -(u, Lv)$, (remember $(u, v) = E[uv]$). The skew-symmetry holds in particular for Hamiltonian systems with canonical data, see [13],[18]; however, here the assumption of skew-symmetry is only an excuse to reduce the number of symbols, not a return to the Hamiltonian case. Pick an orthonormal basis $\{h_k = h_k(\hat{x}), k = 1, \dots\}$ in the range of P, which is the space of functions of $\hat{x}$ (for example, the h_k could be Hermite polynomials in the variables $\hat{x}$). The projection of any function $\psi(x, t)$ can be written as $\psi = \sum_k (\psi(x, t), h_k) h_k(\hat{x})$, and in particular,

$$P(LQe^{sQL}QLx_j) = \sum_k (LQe^{sQL}QLx_j, h_k) h_k(\hat{x}), \tag{21}$$

where a factor Q has been inserted before the exponential, harmlessly because the operators that follow it all live in the null space of P. The memory term now becomes

$$\begin{aligned}\int_0^t e^{(t-s)L} PLe^{sQL} QLx_j ds &= \int_0^t \sum_k e^{(t-s)L} (LQe^{sQL} QLx_j, h_k) h_k(\hat{x}) ds \\ &= \sum_k \int_0^t (LQe^{sQL} QLx_j, h_k) h_k(\hat{\varphi}(t-s)) ds. \end{aligned} \tag{22}$$

In the last identity we used the fact that the inner product in parentheses is independent of time and therefore commutes with the time evolution operator e^{tQL}, and also the fact that $e^{(t-s)L} h_k(\hat{x}) = h_k(\hat{\varphi}(t-s))$. Now $(LQe^{sQL}QLx_j, h_k(\hat{x})) = -(e^{sQL}QLx_j, QLh_k(\hat{x}))$ by the symmetry of Q and the assumed skew-symmetry of L; each term on the right-hand side of equation (22) is the ensemble average of the product of the value of the stochastic process $e^{tQL}QLx_j$ at time $s = t$, with the value of the stochastic process $e^{tQL}QLh_k(\hat{x})$ evaluated at time $s = 0$, i.e., it is a temporal correlation. All these stochastic processes are in the range of Q for all t, and are therefore components of the noise. Remember that by definition $Lx_j = R_j$ (a right-hand side in equations (2)). PLx_j is then an average of the right-hand side of (2) and $QLx_j = R_j - E[R_j|\hat{x}]$ is the initial fluctuation in that right-hand side.

The first, "Markovian", term in equations (18) looks straightforward, but perils lurk there as well. In general R_j in equations (2) is nonlinear, and so is $PLx_j = E[R_j|\hat{x}]$. $e^{tL}PLx_j$ is a nonlinear function of the functions $\hat{\varphi}(t)$ that depends on all the components of x, not only on $\hat{x}$. Some way of approximating this function must be found. If one looks for conditional expectations, one must find a way to commute P with a nonlinear function; for a discussion, see [13]. This bullet was dodged in section 2 when the conditioning variables were chosen to be $\hat{\varphi}(t)$ which change in time, but it may be hard to dodge in general.

The task now at hand is to extract something usable from these rather cumbersome formulas. A very detailed presentation of the analysis in this section can be found in [17].

6. Fluctuation-dissipation theorems

We have established a relation between kernels in the memory term and the noise (the former is made up of covariances of the latter). This is the mathematical content of what are known as "fluctuation-dissipation theorems" in physics. A key difficulty is that the kernels in the memory term consist of covariances of the orthogonal dynamics, whose determination requires in principle the solution of the orthogonal dynamics equations (20), which can be very onerous. However, in the physics literature fluctuation-dissipation theorems are presented in a way that does not stress this difficulty, and we take a moment to explain how the usual physics versions of the theorems come about; they are worth understanding because even though they camouflage the orthogonal dynamics issue they contain significant additional insights.

In the physics literature one often takes a restricted basis in the range of P consisting of the coordinate functions $x_1, ..., x_m$ (the components of $\hat{x}$). The resulting projection is called the " linear projection" as if P as defined above were not linear. The use of this projection is appropriate when the amplitude of the functions $\hat{\phi}(t)$ is small. One then has $h_k(\hat{x}) = x_k$ for $k \leq m$. The covariances in equation (22) are then simply the temporal covariances of the fluctuations in the resolved variables only – all the other covariances have been set to zero. This is known as the fluctuation-dissipation theorem of the second kind. The fluctuations of course obey the orthogonal dynamics equation.

Specialize further to a situation where there is a single resolved variable, say ϕ_1, so that $m = 1$ and $\hat{\phi}$ has a single component. The Mori–Zwanzig equation becomes:

$$\frac{\partial}{\partial t} e^{tL} x_1 = e^{tL} PLx_1 + e^{tQL} QLx_1 + \int_0^t e^{(t-s)L} PLe^{sQL} QLx_1 ds,$$

or,

$$\begin{aligned}\frac{\partial}{\partial t}\phi_1(x,t) &= (Lx_1, x_1)\phi_1(x,t) + e^{tQL} QLx_1 \\ &\qquad + \int_0^t (LQe^{sQL} QLx_1, x_1)\phi_1(x, t-s) ds \\ &= (Lx_1, x_1)\phi_1(x,t) + e^{tQL} QLx_1 - \int_0^t (e^{sQL} QLx_1, QLx_1)\phi_1(x, t-s) ds,\end{aligned} \tag{23}$$

where we have again inserted a harmless factor Q in front of e^{QL}, assumed that L was skew-symmetric as above, and for the sake of simplicity also assumed $(x_1, x_1) = 1$ (if the last statement is not true the formulas can be adjusted appropriately). Take the inner product of equation (23) with x_1, you find:

$$\begin{aligned}\frac{\partial}{\partial t}(\phi_1(x,t),x_1) &= (Lx_1,x_1)(\phi_1(x,t),x_1)\\ &\quad + (e^{tQL}QLx_1,x_1) - \int_0^t (e^{sQL}QLx_1,\,QLx_1)\phi_1(x,t-s)ds\\ &= (Lx_1,x_1)(\phi_1(x,t),x_1) - \int_0^t (e^{sQL}QLx_1,\,QLx_1)(\phi_1(x,t-s),x_1)ds, \end{aligned} \tag{24}$$

because $Pe^{tQL}QLx_1 = (e^{tQL}QLx_1,x_1)x_1 = 0$ and hence $(e^{tQL}QLx_1,x_1)=0$. Multiply equation (24) by x_1, and remember that $P\phi_1(x,t) = (\phi_1(x,t),x_1)x_1$. You find:

$$\frac{\partial}{\partial t}P\phi_1(x,t) = (Lx_1,x_1)P\phi_1(x,t) - \int_0^t (e^{sQL}QLx_1,\,QLx_1)P\phi_1(x,t-s)ds. \tag{25}$$

Observe that the covariance $(\phi_1(x,t),x_1)$ and the projection of ϕ_1 onto x_1 obey the same homogeneous linear integral equation. This is the fluctuation-dissipation theorem of the first kind, which embodies the Onsager principle, according to which spontaneous fluctuations in a system decay at the same rate as perturbations imposed by external means, when both are small (so that the linear projection is adequate). This reasoning can be extended to cases where there are multiple resolved variables, and this is usually done with the added simplifying assumption that $(x_i,x_j)=0$ when $i \neq j$. We omit the details. Finally, if one makes short-memory approximations as in the next section, the issue of orthogonal dynamics disappears completely, as we shall now see.

7. Short-range memory

We have already pointed out that a salient difficulty in using the Mori–Zwanzig equations (18) is the need to solve the orthogonal dynamics equation. We wish now to examine what happens if one bypasses these equations by replacing the orthogonal dynamics by the real dynamics, i.e., if one sets:

$$e^{tQL} \cong e^{tL}. \tag{26}$$

We will show that this is a reasonable approximation under some important circumstances, and that the approximation leads to greatly simplified equations.

First, some heuristic comments. If the resolved dynamics (what happens in the range of P) have no effect on the noise, then the assumption (26) should be valid, for then the unresolved variables interact just with each other; the resulting noise remains unpredictable from the knowledge of $\hat{x}$ and thus remains in the noise space; e^{tQL} and e^{tL} acting on a vector in the noise space should be the same. The effect of the resolved variables on the noise is small in particular if (i) the memory (i.e., the range of t's for which the covariances in the memory term is significant) is short, or (ii) the memory is long. The noise $e^{tQL}QLx_j$ starts out in the noise space by construction, and if the memory is short the operator e^{tL} can take the quantities QLx_j only a small distance out of the noise space before it becomes irrelevant for the evaluation of the covariances; in this short time $e^{tQL}QLx_j$ and $e^{tL}QLx_j$ are the same. If the memory is long, the noise goes on unaffected by the resolved variables. We therefore examine the approximation (26) in these two opposite cases.

In the present section we examine the case of short memory. The memory term in the Mori–Zwanzig equations (18) can be rewritten as

$$\int_0^t e^{(t-s)L}PLe^{sQL}QLx_j\,ds = \int_0^t e^{(t-s)L}PLQe^{sQL}QLx_j\,ds, \tag{27}$$

where the insertion of the extra Q is harmless. Adding and subtracting equal quantities, we find:

$$PLe^{sQL}QLx_j = PLQe^{sL}QLx_j + PLQ(e^{sQL}-e^{sL})QLx_j; \tag{28}$$

a Taylor series yields:

$$e^{sQL}-e^{sL} = I+sQL+\cdots-I-sL-\cdots = -sPL+O(s^2), \tag{29}$$

and therefore, using $QP=0$, we find:

$$\int_0^t e^{(t-s)L}PLe^{sQL}QLx_j\,ds = \int_0^t e^{(t-s)L}PLQe^{sL}QLx_j\,ds + O(t^3). \tag{30}$$

If P is a finite rank projection then

$$PLe^{sQL}QLx_j = \sum_k (QLe^{sQL}QLx_j, h_k)h_k(\hat{x}), \tag{31}$$

where, as before, one can write $(QLe^{sQL}QLx_j, h_k)$ as $-(e^{sQL}QLx_j, QLh_k)$ when L is skew-symmetric. If the covariances $(e^{sQL}QLx_j, QLh_k)$ and also the covariances $(e^{sL}QLx_j, QLh_k)$ are significant only over short times t_0, the approximation (26) provides an approximation with an error $O(t_0^3)$ without requiring the solution of the orthogonal dynamics equation; this is still a short covariance time approximation but it can be preferable to a white noise approximation (see [41] for an application to the dimensional reduction of the Kuramoto–Sivashinsky equation and [2] for an application to molecular dynamics).

One important short-memory situation where the Mori–Zwanzig formalism simplifies even more is when the noise can be viewed as white noise. This is a valid approximation in a number of important cases, in particular when there is scale separation between the resolved and unresolved variables or when these variables are weakly coupled (for recent reviews see, e.g., [21],[32], [40]). These situations are often encountered in applications, but we do not survey them here because their analysis does not require all of our machinery.

If the noise can indeed be viewed as white, one sets:

$$e^{tQL}QLx_j = A_j w_j'(t), \tag{32}$$

where the prime denotes a derivative, the $w_j(t)$ are independent unit Brownian motions so that that the w' are white noises, and the A_j are constants that must be derived from some prior knowledge. The covariances of the noise are then delta functions (thus the memory is vanishingly short). If one assumes further that the projection P is well represented by the physicists' "linear" projection, then the integral in the memory term can be easily seen to reduce to a constant times the unknowns, and equations (18) become stochastic ordinary differential equations of the usual kind. As usual (see, e.g., [27]), the corresponding probability densities can be found via Fokker–Planck formalisms (or Kolmogorov equations, in mathematicians' language).

It is important to note that the assumption of white noise does not require that the linear projection be used. More noise terms appear when one uses a more general linear projection, and one encounters situations where the additional noise terms can no longer be viewed as white and their uses detracts from the overall accuracy (see, e.g., [41; 42; 32]). These papers also include suggestions as to how to pick the best number of terms to use in the projections. Projections other than linear are important for mode-coupling theory in condensed matter physics, see, e.g., [45].

There is a comment to be added here. White noise and delta memory constitute an important special case. However, this is not the general case and maybe not even the usual case. It is rather surprising that 40 years after Alder and Wainwright [1] demonstrated the long-range memory in a common physical system, years during which physicists have learned how to model systems with arbitrary memory, most numerical treatments of dimensional reduction seem to assume that all memory is ultra-short. It is also surprising that most papers on dynamic renormalization (see, e.g., [24]) assume that the noise is white without comment, making it pointless to compare the schemes below with this dynamical renormalization literature.

Finally, it should be obvious that very short memory is very different from no memory, i.e., from situations where the memory term is absent altogether.

8. Long-range memory and the t-model

We examine now the validity of the ansatz $e^{tQL} \cong e^{tL}$ for cases with slowly decaying memory. Write the memory term in the Mori–Zwanzig equation (18) as

$$\begin{aligned}\int_0^t e^{(t-s)L} PLe^{sQL} QLx_j ds = \int_0^t Le^{(t-s)L} e^{sQL} QLx_j ds \\ - \int_0^t e^{(t-s)L} e^{sQL} QLQLx_j ds,\end{aligned}$$

where we have used the commutation of L and QL with e^{tL} and e^{sQL}, respectively. At this point, make the approximation (26), which eliminates the s dependence of both integrands and we obtain:

$$\int_0^t e^{(t-s)L} PLe^{sQL} QLx_j ds \cong te^{tL} PLQLx_j. \tag{33}$$

All that remains of the integration in time is the coefficient t. To estimate the error, consider the difference between the full memory term and its approximation:

$$\begin{aligned}\int_0^t e^{(t-s)L} PLe^{sQL} QLx_j ds - te^{tL} PLQLx_j = \\ \int_0^t [e^{(t-s)L} PLe^{sQL} - e^{tL} PL] QLx_j ds.\end{aligned}$$

Adding and subtracting equal quantities, we find

$$e^{(t-s)L} PLe^{sQL} = e^{tL} PL + e^{tL}[e^{-sL} PLe^{sQL} - PL],$$

and a Taylor series around $s = 0$ gives

$$e^{-sL} PLe^{sQL} - PL = (I - sL + \ldots)PL(I + sQL + \ldots) - PL = O(s). \tag{34}$$

This implies

$$\int_0^t e^{(t-s)L} PLe^{sQL} QLx_j ds = te^{tL} PLQLx_j + O(t^2).$$

To understand this estimate, examine an alternate derivation of (33). Expand the integrand of the memory term of the Mori–Zwanzig equation around $s = 0$ and retain only the leading term, finding

$$\begin{aligned}\int_0^t e^{(t-s)L} PLe^{sQL} QLx_j ds &= \int_0^t [e^{tL} PLQLx_j + O(s)] ds \\ &= te^{tL} PLQLx_j + O(t^2).\end{aligned}$$

If we retain only the leading term, we do not keep any information about the time evolution of the integrand, which in turn means no information about the evolution of the resolved component and of the coupling to the orthogonal dynamics (through the term $(LQe^{sQL}QLx_j, h_k)$). Such a drastic approximation is expected to be appropriate in cases where the memory term integrand is slowly decaying, so that information about its initial value is sufficient to make predictions.

We have just seen that if the memory is long the ansatz $e^{tQL} \cong e^{tL}$ reduces the memory to a Markovian term with a time-dependent coefficient. Thus the assumption $e^{tQL} \cong e^{tL}$ greatly simplifies the equations, as expected. The resulting equations were introduced in [13] and are known as the "t-model".

As an example, consider again the Hald model whose Hamiltonian is

$$H(\phi) = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2 + \phi_1^2\phi_3^2). \tag{35}$$

The resulting equations of motion are:

$$\begin{aligned}\frac{d\phi_1}{dt} &= \phi_2\\ \frac{d\phi_2}{dt} &= -\phi_1(1+\phi_3^2)\\ \frac{d\phi_3}{dt} &= \phi_4\\ \frac{d\phi_4}{dt} &= -\phi_3(1+\phi_1^2).\end{aligned}$$

Suppose one wants to solve only for $\hat{\phi} = (\phi_1, \phi_2)$, with initial data $\hat{x} = (x_1, x_2)$. Assume the initial data x_3, x_4 are sampled from a canonical density with temperature $T = 1$. A quick calculation yields $E[x_3^2|x_1, x_2] = 1/(1+x_1^2)$. The advance in time described by the multiplication by e^{tL} requires just the substitution $\hat{x} \to \hat{\phi}$. If one commutes the nonlinear function evaluation and the conditional averaging, i.e., writes $Pf(\hat{\phi}) = f(P\hat{\phi})$ (a "mean-field approximation"), and writes furthermore $\Phi(t) = P\hat{\phi} = E[\hat{\phi}|\hat{x}]$ one finds $Pe^{tL}PLx_1 = \Phi_2$, $Pe^{tL}PLx_2 = -\Phi_1(1 + 1/(1 + \Phi_2^2))$; one can calculate $Pe^{tL}LQLx_j$ for $j = 1, 2$ and finally one finds:

$$\begin{aligned}\frac{d}{dt}\Phi_1 &= \Phi_2\\ \frac{d}{dt}\Phi_2 &= -\Phi_1(1+\frac{1}{1+\Phi_1^2}) - 2t\frac{\Phi_1^2\Phi_2}{(1+\Phi_1^2)^2}.\end{aligned} \tag{36}$$

The last term represents the damping due to the loss of predictive power of partial data; the coefficient of the last term increases in time and one may worry that this last term eventually overpowers the equations and leads to some odd behavior. This

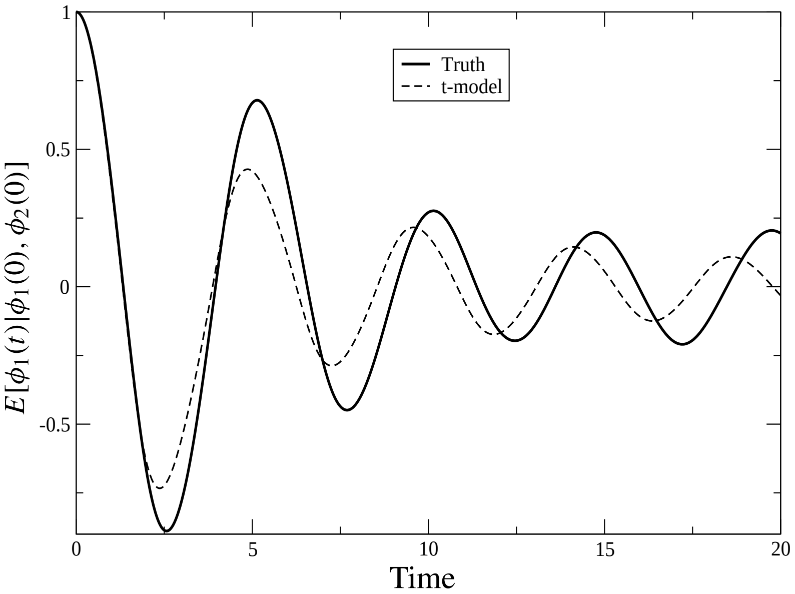


Figure 2. Comparison of the evolution of $E[\phi_1(t)|\phi_1(0), \phi_2(0)]$ (truth) with the prediction by the t-model; for comments, see the text.

is not the case. Indeed, one can prove the following general result: If the system one starts from, equation (2) is Hamiltonian with Hamiltonian H, and if the initial data are sampled from an initial canonical density conditioned by partial data $\hat{x}$, and if $\hat{H}$ is the renormalized Hamiltonian (in the sense of section 2), then $(d/dt)\hat{H} \leq 0$, showing that the components of $\hat{\phi}$ decay as they should. The proof requires a technical assumption (that the Hamiltonian H can be separated into the sum of a function of the momenta and a function of the position, a condition commonly satisfied) and we omit it (see [13]).

The solution of the t-model with the mean-field approximation for the Hald model is presented in Figure 2. The applicability of the approximation suffers from the fact that at the temperature $T = 1$ the Hald system is not ergodic. To see what has been gained, contrast this figure with Figure 1.

If the t-model is not sufficient for the approximation of a given problem, one can try to generalize it. Indeed, we have just seen that the t-model is the zero-th order term in a Taylor expansion (around $s = 0$) of the integrand of the memory term in (18). However, nothing prevents us from keeping more terms in this expansion. Let

$$K(\hat{\varphi}(t-s), s) = e^{(t-s)L} P L e^{sQL} Q L x_j$$

and expand K around $s=0$, i.e.,

$$K(\hat{\varphi}(t-s),s)=K(\hat{\varphi}(t),0)+s\frac{\partial K}{\partial s}|_{s=0}+\frac{1}{2}s^2\frac{\partial^2 K}{\partial s^2}|_{s=0}+O(s^3).$$

In the case when P is the finite-rank projection and the density used to define the projection is invariant, the derivatives of K at $s=0$ are equal-time (static) covariances. In mode-coupling theory, such expressions are known as "sum rules". One can assume a functional form for the memory term integrand around $s=0$, e.g., a Gaussian ae^{-bs^2}, and use the derivatives of K at $s=0$ to estimate a,b (see [37] for more on sum rules and mode-coupling theory). This is potentially another place where current ideas in physics can be helpful in numerical modeling.

The usefulness of the t-model depends on the range of the memory; this raises the question of what this range depends on and whether it can be modified. If the number of resolved variables is small, the range of the memory depends on the range of the memory in the full system (2)- indeed, if there are no resolved variables, as in section 3 above, the dynamics and the orthogonal dynamics are the same. However, in the general case, is it possible to have a reduced model with very short or very long memory, depending on how one coarse-grains a particular system at hand? In [41] evidence was presented that, for the Kuramoto–Sivashinsky equation, the range of the memory of a reduced model can vary dramatically, depending on whether all the unstable modes in the system are resolved or not. The construction of a reduced model corresponds to renormalization, and the two extreme cases can be interpreted as two fixed points of a renormalization scheme. In which one a reduced model will end up depends on how one renormalizes. How to formalize these remarks and put them to use remains a topic for further work.

Both the long memory approximation and the short memory approximation have been derived from the assumption $e^{tQL}\cong e^{tL}$, but this assumption has been used differently. In the short memory case one first makes this substitution in the memory term and then one performs the projection in that term; in the long memory case one performs these two operations in the reverse order. This leads to different results.

Finally, we go back to the remark at the end of section 2. We believe that the t-model is a sound basis for large eddy simulation in hydrodynamics; the equations are relatively simple and the memory is taken into account. We are acting on this basis and expect to publish results soon.

9. Intermediate-range memory

There are intermediate cases where the memory cannot be viewed as either short or long so that neither model above can be used. At present, it is not known how to deal effectively with such cases. In a series of papers [11]-[13] we presented special cases and their solutions. In particular in [13] we presented a detailed analysis of the Hald system without the t-model assumptions. We showed that the memory decays roughly at the same rate as the solution itself (this is the general case in the absence of separation of scales). We expanded the various covariances at equilibrium (i.e., when there are no resolved variables) in Hermite polynomials, evaluated the coefficients in the expansions by Monte Carlo once and for all, and then obtained a system of integro-differential approximations to equations (18) which we then solved in various cases. This is a legitimate procedure which may be useful when the same system of equations has to be solved repeatedly. These calculations do exhibit a salient feature of model reduction in time-dependent problems, which is that its set-up costs are often very high. The future remedy, if there is one, will surely lie in a deeper understanding of dynamical renormalization and, in particular, of the way memory depends on scale.

10. Conclusions

We have made two sets of claims. First, theoretical claims: If one assumes that a probability density is initially available for all the degrees of freedom in a complex problem, then the problem of following the evolution of just a few degrees of freedom becomes a problem in statistical mechanics, of the equilibrium kind for problems with stationary densities, and of the non-equilibrium kind otherwise. Finding an equivalent problem with lesser complexity is equivalent to a renormalization, and a successful reduction in complexity corresponds to uncovering a similarity relation between the full problem and the reduced problem. Physics is often a good guide to what should be done.

On the practical side, reduction by conditional expectation is a powerful tool. In the stationary case we have used it to generate block Monte Carlo algorithms and effective equations for mean solutions. In the time dependent case it leads to the Mori–Zwanzig formalism, generalized Langevin equations, and promising approximation schemes. We have high hopes for the usefulness of one particular approximation scheme, the t-model, which yields good approximations in interesting cases with a relatively low overhead.

11. Acknowledgements

We would like to thank Prof. G.I. Barenblatt, Prof. O. Hald and Prof. R. Kupferman for all that we have learned from them; their ideas are present everywhere in this paper.

References

[1] B. Alder and T. Wainright, *Decay of the velocity correlation function*, Phys. Rev. A **1** (1970), 1–12.

[2] J. Barber, *Application of optimal prediction to molecular dynamics*, Ph.D. thesis, University of California, Berkeley, 2005.

[3] G. I. Barenblatt, *Scaling*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2003. MR 2005e:00011

[4] G. I. Barenblatt, M. Ya. Ivanov, and G. I. Shapiro, *On the structure of wave fronts in nonlinear dissipative media*, Arch. Rational Mech. Anal. **87** (1985), no. 4, 293–303. MR 86d:76024

[5] G. Benfatto and G. Gallavotti, *Renormalization group*, Physics Notes, vol. 1, Princeton University Press, Princeton, NJ, 1995. MR 97e:82001

[6] J. L. Bona and M. E. Schonbek, *Travelling-wave solutions to the Korteweg-de Vries-Burgers equation*, Proc. Roy. Soc. Edinburgh Sect. A **101** (1985), no. 3-4, 207–226. MR 87k:35208 Zbl 0594.76015

[7] A. Brandt and D. Ron, *Renormalization multigrid (rmg): Statistically optimal renormalization group flow and coarse-to-fine monte carlo acceleration*, J. Stat. Phys. **102** (2001), 231–257.

[8] L. Chen, P. Debenedetti, C. Gear, and I. Kevrekidis, *From molecular dynamics to coarse self-similar solutions: a simple example using equation-free computation*, J. Non-Newt. Fluid. Mech. **120** (2004), 215.

[9] A. J. Chorin, *Averaging and renormalization for the Korteveg-deVries-Burgers equation*, Proc. Natl. Acad. Sci. USA **100** (2003), no. 17, 9674–9679 (electr.). MR 2004f:35149

[10] ———, *Conditional expectations and renormalization*, Multiscale Model. Simul. **1** (2003), no. 1, 105–118 (electronic). MR 2004i:82043 Zbl 02139971

[11] A. J. Chorin, O. H. Hald, and R. Kupferman, *Optimal prediction and the Mori–Zwanzig representation of irreversible processes*, Proc. Natl. Acad. Sci. USA **97** (2000), no. 7, 2968–2973 (electronic). MR 2000m:82045 Zbl 0968.60036

[12] ———, *Non-Markovian optimal prediction*, Monte Carlo Methods Appl. **7** (2001), no. 1-2, 99–109. MR 2002a:65016 Zbl 0982.65011

[13] ———, *Optimal prediction with memory*, Physica D **166** (2002), no. 3-4, 239–257. Reviewed in: MR 2003e:62150 Zbl 1017.60046

[14] A. J. Chorin, O. H. Hald, and R. Kupferman, *Prediction from partial data, renormalization and averaging*, To appear in J. Sci. Comp., 2005.

[15] A. J. Chorin, A. P. Kast, and R. Kupferman, *Optimal prediction of underresolved dynamics*, Proc. Natl. Acad. Sci. USA **95** (1998), no. 8, 4094–4098 (electronic). MR 99a:65195 Zbl 0904.65117

[16] A. J. Chorin, R. Kupferman, and D. Levy, *Optimal prediction for Hamiltonian partial differential equations*, J. Comput. Phys. **162** (2000), no. 1, 267–297. MR 2001i:65137 Zbl 0960.65012

[17] A.J. Chorin and O. Hald, *Stochastic tools for mathematics and science*, Springer-Verlag, New York, 2005.

[18] D. Evans and G. Morriss, *Statistical mechanics of nonequilibrium liquids*, Academic Press, London, 1990.

[19] E. Fick and G. Sauermann, *The quantum statistics of dynamic processes*, Springer Series in Solid-State Sciences, vol. 86, Springer-Verlag, Berlin, 1990. MR 91m:82001

[20] M. E. Fisher, *Renormalization group theory: its basis and formulation in statistical physics*, Rev. Modern Phys. **70** (1998), no. 2, 653–681. MR 99e:82041

[21] D. Givon, R. Kupferman, and A. Stuart, *Extracting macroscopic dynamics: model problems and algorithms*, Nonlinearity **17** (2004), no. 6, R55–R127. MR 2097022

[22] N. Goldenfeld, *Lectures on phase transitions and the renormalization group*, Perseus Books, Reading, Mass., 1992.

[23] H. Grabert, *Projection operator techniques in nonequilibrium statistical mechanics*, Springer Tracts Modern Physics, vol. 95, Springer-Verlag, Berlin, 1982. MR 84k:82001

[24] P. Hohenberg and B. Halperin, *Theory of dynamical critical phenomena*, Rev. Mod. Phys. **49** (1977), 435–479.

[25] E. Ingerman, *Modeling the loss of information in optimal prediction*, Ph.D. thesis, University of California, Berkeley, 2003.

[26] G. Jona-Lasinio, *The renormalization group: a probabilistic view*, Nuovo Cimento B (11) **26** (1975), 99–119. MR 51 #4908

[27] W. Just, H. Kantz, C. Rödenbeck, and M. Helm, *Stochastic modelling: replacing fast degrees of freedom by noise*, J. Phys. A **34** (2001), no. 15, 3199–3213. MR 1836464

[28] L. P. Kadanoff, *Statistical physics*, World Scientific Publishing Co. Inc., River Edge, NJ, 2000. MR 2003f:82001

[29] L. D. Landau and E. M. Lifshitz, *Statistical physics, Part I*, Butterworth–Heinemann, 1980. MR 84m:82003a

[30] J. A. Langford and R. D. Moser, *Optimal LES formulations for isotropic turbulence*, J. Fluid Mech. **398** (1999), 321–346. MR 2000h:76103 Zbl 0983.76043

[31] M. Lesieur, P. Compte, and J. Zinn-Justin, *Mécanique des fluides numérique: Les houches, session lvix, 28 juin-30 juillet 1993*, North-Holland Publishing Co., Amsterdam, 1996. Reviewed in MR 97f:76001

[32] A. J. Majda, I. Timofeyev, and E. Vanden Eijnden, *A mathematical framework for stochastic climate models*, Comm. Pure and Applied Math. **54** (2001), no. 8, 891–974. MR 2002k:86012 Zbl 1017.86001

[33] H. Mori, *Transport, collective motion and brownian motion*, Prog. Theor. Phys. **33** (1965), 423–450.

[34] S. Nordholm, , and R. Zwanzig, *A systematic derivation of exact generalized Brownian motion theory*, J. Statist. Phys. **13** (1975), no. 4, 347–371. MR 56 #10694

[35] P. Okunev, *Application of optimal prediction to spin systems and to economics*, Ph.D. thesis, University of California, Berkeley, 2005.

[36] G. C. Papanicolaou, *Asymptotic analysis of stochastic equations*, Studies in probability theory, MAA Stud. Math., vol. 18, Math. Assoc. America, Washington, D.C., 1978, pp. 111–179. MR 80j:60095 Zbl 0443.60049

[37] Y. Pomeau and P. Resibois, *Time dependent correlation functions and mode-mode coupling theories*, Physics Reports C **2** (1975), 63–139.

[38] B. Seibold, *Optimal prediction in molecular dynamics*, Monte Carlo Methods Appl. **10** (2004), no. 1, 25–50. MR 2004m:81286 Zbl 02124459

[39] H. E. Stanley, *Scaling, universality and renormalization, three pillars of modern critical phenomena*, Rev. Mod. Phys. **71** (1999), S358–S366.

[40] P. Stinis, *A comparative study of two stochastic mode reduction methods*, To appear Physica D.

[41] ———, *Stochastic optimal prediction for the Kuramoto-Sivashinsky equation*, Multiscale Model. Simul. **2** (2004), no. 4, 580–612 (electronic). MR 2113171 Zbl 02148122

[42] ———, *A maximum likelihood algorithm for the estimation and renormalization of exponential densities*, J. Comp. Phys. **208** (2005), 691–703.

[43] R. Swendsen, *Monte-carlo renormalization group*, Phys. Rev. Lett. **42** (1979), 859–861.

[44] K. Theodoropoulos, Y.-H. Qian, and I.G. Kevrekidis, *“Coarse" stability and bifurcation analysis using timesteppers: a reaction diffusion example*, Proc. Natl. Acad. Sci. USA **97** (2000), 9840–9843 (electronic).

[45] R. van Zon and J. Schofield, *Mode-coupling theory for multiple-point and multiple-time correlation functions*, Phys. Rev. E (3) **65** (2002), no. 1, part 1, 011106, 17. MR 1892224

[46] R. Zwanzig, *Nonlinear generalized langevin equations*, J. Stat. Phys. **9** (1973), 215–220.

Received April 13, 2005.

ALEXANDRE J. CHORIN: chorin@math.berkeley.edu
Department of Mathematics, University of California, Berkeley CA 94720-3840,
United States
http://math.berkeley.edu/∼chorin

PANAGIOTIS STINIS: pstinis@lbl.gov
Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 50A-1148,
Berkeley, CA 94720-1148, United States

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

ACTIVE CONTROL FOR STATISTICALLY STATIONARY TURBULENT PREMIXED FLAME SIMULATIONS

JOHN B. BELL, MARCUS S. DAY,
JOSEPH F. GRCAR AND MICHAEL J. LIJEWSKI

The speed of propagation of a premixed turbulent flame correlates with the intensity of the turbulence encountered by the flame. One consequence of this property is that premixed flames in both laboratory experiments and practical combustors require some type of stabilization mechanism to prevent blow-off and flashback. The stabilization devices often introduce a level of geometric complexity that is prohibitive for detailed computational studies of turbulent flame dynamics. Furthermore, the stabilization introduces additional fluid mechanical complexity into the overall combustion process that can complicate the analysis of fundamental flame properties. To circumvent these difficulties we introduce a simple, heuristic feedback control algorithm that allows us to computationally stabilize a turbulent premixed flame in a simple geometric configuration. For the simulations, we specify turbulent inflow conditions and dynamically adjust the integrated fueling rate to control the mean location of the flame in the domain. We outline the numerical procedure, and illustrate the behavior of the control algorithm on methane flames at various equivalence ratios in two dimensions. The simulation data are used to study the local variation in the speed of propagation due to flame surface curvature.

1. Introduction

A well-known property of turbulent premixed flames is that their speed of propagation correlates to the turbulent intensity in the unburned mixture. See Bradley [5] and Peters [28] for a discussion of this issue. As a consequence, premixed flames are inherently unstable when propagating against a turbulent flow whose intensity increases upstream but decays downstream. To have a stable flame for either laboratory analysis or for a practical combustor requires some type of flame stabilization mechanism. A variety of approaches are used to stabilize premixed turbulent flames in the laboratory [10]. For example, the Twenty-Ninth Combustion Symposium includes studies by Sattler et al. [33] of a turbulent V-flame, Shepherd et al. [34] of a swirl-stabilized flame, Most et al. [24] of a bluff-body stabilized flame,

MSC2000: 80A25, 93C20, 65M06, 65M50.

Keywords: active control, statistically stationary, turbulent flames.

and Chen et al. [9] of Bunsen and stagnation flames. These stabilization mechanisms are necessary to control the flame location so that data can be collected. Each stabilization mechanism has advantages and disadvantages. Bluff-body stabilized flames, V-flames and Bunsen flames are fluid-mechanically fairly simple but there is substantial flow tangential to the flame and the flame encounters different levels of turbulence further from the burner nozzle. The low-swirl geometry produces a statistically nearly flat flame but the fluid mechanics of the stabilization are quite complex. Stagnation plate flames are geometrically and fluid mechanically simple but the flame experiences a substantial mean strain and heat loss to the plate. In each case, the additional complexity introduced by the stabilization complicates both the analysis of the flame data and the implication of the results to the characterization of premixed turbulent flames.

For the most part, computational studies of premixed flames that include detailed chemistry and transport and resolve the relevant fluid-mechanical scales have not included any of these stabilization mechanisms. For an exception, see the model of a three-dimensional (3D) turbulent V-flame by Bell et al. [4]. The computational demands of these types of simulations combined with the specialized numerical algorithms typically used for direct numerical simulations make including physical stabilization mechanisms prohibitively expensive.

The idealized configuration that we use for the present study is a modified version of one used frequently in the numerical study of premixed turbulent flames. A flat steady laminar premixed flame is initialized in a computational domain and allowed to propagate toward a boundary where turbulent perturbations have been superimposed on a mean inflow. After the turbulence interacts with the flame for a sufficient period of time, statistics are gathered from the solution to quantify the extent to which the turbulent fluctuations modify the flame structure. There is an extensive literature on computational studies of this type in 2D, both with simplified and detailed chemistry. Examples germane this configuration include Baum et al. [2] who studied turbulent flame interactions for detailed hydrogen chemistry, and Haworth et al. [19] who examined the effect of inhomogeneous reactants for propane–air flames using detailed propane chemistry. Analogous studies in 3D have been performed by Rutland and Trouvé [32], Trouvé and Poinsot [38], Zhang and Rutland [41], and Chakraborty and Cant [7]. All of these 3D studies were based on simplified chemistry. More recently Tanahashi et al. [36; 37] have performed simulations of this type for turbulent premixed hydrogen flames with detailed hydrogen chemistry. Bell et al. [3] performed a similar study for a turbulent methane flame.

Computational studies involving the idealized flow configuration suffer from a fundamental instability that prevents stabilization of the computed flames. If the flame begins to propagate faster than the specified inflow velocity, then the flame migrates upstream nearer the stronger turbulence which further increases its speed. Similarly, a propagation speed slower than the inflow velocity causes the flame to migrate downstream into further decayed turbulence where the flame propagation is even slower. Thus the flame may not encounter a given turbulence intensity long enough to gather statistics about its behavior at that intensity. Moreover, since the flame is not statistically stationary in the computational domain, it will often migrate to either the domain inflow or outflow boundary, terminating the simulation.

In this paper, we apply a simple, heuristic feedback control algorithm to automatically adjust the inflow velocity to stabilize flames in the idealized configuration. The control algorithm allows long-time simulation of statistically stationary flames in a configuration free of complications associated with physical boundary conditions. In the next section, we briefly describe the basic simulation methodology for low-speed reacting flows, and describe the feedback control procedure. We then demonstrate the ability of the algorithm to stabilize premixed methane flames in 2D. We next demonstrate the utility of this algorithm by exploring global burning statistics and correlations in localized burning with flame geometry.

2. Computational methodology

2.1. ***Simulation methodology.*** The simulations presented here are based on a low Mach number formulation of the reacting flow equations. The methodology treats the fluid as a mixture of perfect gases. We use a mixture-averaged model for differential species diffusion and ignore Soret, Dufour, gravity and radiative transport processes. Unless explicitly stated otherwise, the chemical kinetics are modeled using the GRI-Mech 3.0 methane mechanism [15; 35] with 53 species and 325 fundamental reactions. The basic discretization combines a symmetric operator-split coupling of chemistry and diffusion processes with a density-weighted approximate projection method. The projection method incorporates the constraint on the velocity divergence that arises in the low Mach number formulation. The resulting integration of the advective terms proceeds on the time scale of the relatively slow advective transport. Faster diffusion and chemistry processes are treated time-implicitly. This integration scheme is embedded in a parallel adaptive mesh refinement algorithm framework based on a hierarchical system of rectangular grid patches. The complete integration algorithm is second-order accurate in space and time, and discretely conserves species mass and enthalpy. The reader is referred to [13] for details of the low Mach number model and its numerical implementation and to [3] for previous applications of this methodology to the simulation of premixed turbulent flames.

2.2. ***Flow Configuration.*** The flow configuration we consider initializes a flat laminar flame in a domain oriented so that the flame propagates downward. Since there is no gravitational force included, up and down are for orientation only. A cold fuel-air premixture enters the domain through bottom boundary, and hot combustion products exit the domain through the top. The remaining computational boundaries are periodic. Along the inflow face we specify both a mean inflow velocity and turbulent fluctuations that are superimposed on the mean inflow. A control algorithm is used to adjust the mean inflow rate to hold the flame in the domain indefinitely. As a result, the calculation essentially is carried out in a Lagrangian frame, moving with the intrinsic mean speed of the flame for a particular fuel, stoichiometry, and turbulence intensity. The following section details the strategy for computing the mean inflow rate needed to hold the flame statistically steady in the simulation domain.

2.3. ***Control Methodology.*** The inflow stream has turbulent fluctuations that interact with the flame to cause fluctuations in the turbulent flame speed. To control the flame location, we need to develop a control algorithm that will dynamically adjust the inflow rate to compensate for these variations in the flame speed. Because the types of flame simulation we want to control are extremely costly, it is infeasible to develop the control algorithm directly in terms of actual simulations. As an alternative, we will develop a simplified model to describe the flame dynamics in 1D, and then develop the control algorithm for the multidimensional flame in the context of that simplified model. The mean vertical flame location, $h(t)$, is computed from the evolving 2D solution by integrating the instantaneous mass of fuel in the domain and dividing this result by the product of the fuel density and inlet area at the inflow boundary. This averaged flame location propagates downward at some effective turbulent flame speed, s, relative to the mean fluid motion. The control problem is to specify a mean inflow velocity $v_{\text{in}}(t)$ that automatically pushes the flame from an initial flame location, $h(0) = \alpha$, to the target flame location, $h(t) = \beta$.

The dynamics of the average flame position can be modeled using a stochastic differential equation of the form

$$dh = (v_{\text{in}}(t) - s(h))dt + d\omega \tag{1}$$

where the effective flame speed, $s(h(t))$ is a function of the time-dependent flame position, and must be estimated as part of the control process. The final term, $d\omega$, represents high-frequency fluctuations in the turbulent flame speed due to stochastic fluctuations in the inflow stream.

Given a quadratic cost functional associated with equation (1) and assumptions about the noise term, there are well-known procedures for deriving optimal control strategies: see Kushner [23], Caines [6] and Chen, Chen and Hsu [8]. However, in the present case, we do not have a closed-form characterization of the fluctuations. Further, since the control velocity, $v_{\text{in}}(t)$, determines the boundary condition for the low Mach number solver, we need to impose additional constraints on the profile. In particular, we need $v_{\text{in}}(t)$ to be smooth in time and we need to impose limits on how rapidly it can change. These heuristic constraints are chosen so that we do not introduce instabilities or inaccuracies into the simulation algorithm or subject the flame to large accelerations that could induce spurious fluid dynamical behavior from Rayleigh-Taylor instabilities.

For each time step in the algorithm, we will take as an ansatz that $v_{\text{in}}(t)$ is linear over the entire AMR coarse time step and limited such that the inflow velocity cannot change dramatically during a time step. These smoothness criteria constrain how rapidly v_{in} can respond to changes in h and to noise. Consequently, we need to introduce a time scale, τ, which is the target lag for reaching the control state. We want to estimate Δv, the change in v from time t_0 to $t_0 + \tau$, so that h reaches β over the period τ. We assume that τ is sufficiently large that the noise $d\omega$ has mean zero over the interval $[t_0, t_0 + \tau]$, yet assume that the turbulent flame speed, s, is slowly varying. We are given a flame location, $h(t_0)$ and an inflow velocity, $v_{\text{in}}(t_0)$, at the beginning of the time step and an estimate s_{est} of the mean flame speed over the interval. Assuming Δv is constant over the interval t_0 to $t_0 + \tau$, we can integrate equation (1) and rearrange to obtain:

$$\beta = h(t_0) + \tau(v_{\text{in}}(t_0) - s_{\text{est}}) + \frac{\tau}{2}\Delta v$$

For the purposes of computing this integral, we estimate s_{est} from the change in fuel mass in the domain during the previous time step. To enforce the smoothness required by the flow solver we then limit Δv so that over a time step the velocity does not change by more than 0.1 $\max\{v_{\text{in}}(t_0), v_{\text{min}}\}$ where v_{min} is a minimum velocity scale of the problem that can be computed from the post-flame velocity of the laminar flame propagating into fluid at rest. Also, we find our simulation methodology to be more robust if we avoid outflows at the inflow boundary by requiring $v_{\text{in}} \geq 0$. (Note that this strategy therefore relies on burning to move the flame in the upstream direction.) We then represent $v_{in}(t) = v_{in}(t_0) + t\Delta v/\tau$ for the current AMR time step. At the beginning of the next time step, we recompute Δv based on the new flame location and the estimated flame speed.

2.4. ***Determination of Control Parameters.*** Robustness of this control algorithm depends strongly on the heuristic parameters. As note earlier, the cost of the computations rules out using actual flame simulations to calibrate the control. Instead, in order to explore the implications of these parameters, we specify a synthetic turbulent flame speed model and noise term into the model equation (1) and perform tests of the algorithm for this synthetic turbulent ”flame” with parameters chosen to reflect conditions of a typical flame simulation. Experimental and computational data suggests that the effective propagation speed of a turbulent premixed flame correlates with the intensity of the turbulence. We expect, therefore, that the closer the flame is to the inlet boundary (turbulence source) the faster it will propagate. In our configuration, this suggests that $s'(h) < 0$. For our model, we set

$$s(h) = \bar{s}\,(1 - \gamma\,(h - \beta))$$

so that $s'(h) = -\gamma\bar{s}$. For our tests, we take the remaining parameters to reflect values corresponding to a lean premixed methane flame: $\bar{s} = 0.3$, $\gamma = 0.1$, $\beta = 0.005$, $\alpha = 0.001$ and $\Delta t = 0.00002$. This Δt is typical of timestep sizes found on the coarsest meshes in our adaptive mesh refinement algorithm for low Mach number flows; we refine in both space and time so the finer, refined meshes have proportionately smaller time steps. To simulate noise due to the turbulent fluctuations, we used uniformly distributed random perturbations of $\pm 33\%$.

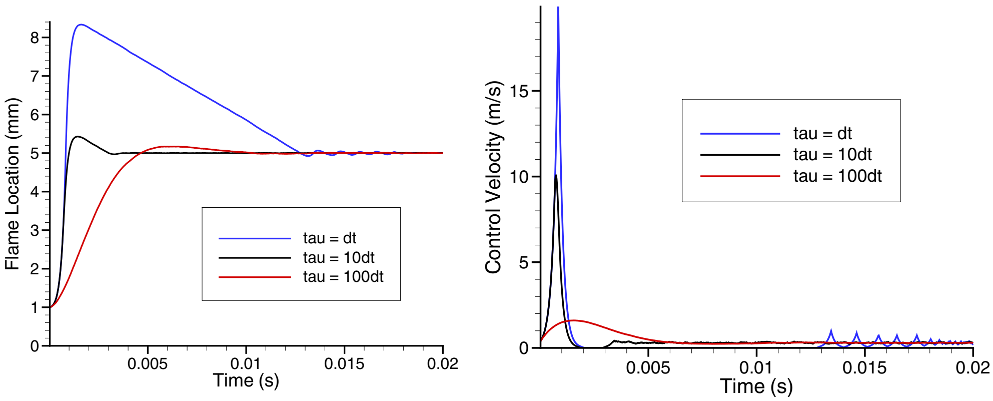


Figure 1. *Synthetic flame control simulations. Left: flame location. Right: control velocity.*

Simulation results showing the "flame location," h, and control velocity computed by the algorithm for various values of τ are shown in Figure 1. From the results, if τ is too small, corresponding to quickly controlling the flame, then the restrictions on changing the velocity lead to fluctuations on the synthetic flame location that persist for considerable time. If τ is too large, the system is well-behaved but a relatively long time is required to reach the desired state. Our test indicate that $\tau = 10\,\Delta t$ appears to provide a robust control that relatively rapidly controls the flame to the desired location. We note that even if the control is started at the correct location and correct velocity, setting $\tau = \Delta t$ does not provide satisfactory results. The interplay of perturbations from the noise and the restrictions on changing v_{in} lead to fairly large oscillations as indicated in Figure 2. We note that the parameters selected here were chosen to introduce more variation in both noise and flame speed than we expect to find in practice. Additional tests, however, have demonstrated that the parameters continue to perform effectively over a range of conditions.

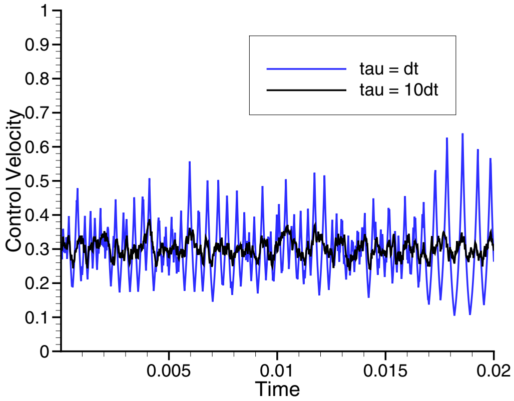


Figure 2. *Synthetic flame control simulations starting from correct flame location and speed.*

3. Controlled methane flames

3.1. ***2-step mechanism.*** We validate the control algorithm using a representative time-dependent simulation of premixed methane combustion. A simplified combustion model reduces the cost of integration so that the control algorithm can be observed over a long time period in order fully characterize the resulting performance and system response. This simplified calculation assumes a unity Lewis number [29] for transport and it has just 6 chemical species: CH_4, O_2, CO_2, H_2O, CO_2, N_2. The 2-step kinetics mechanism (see [26], Model “2”, with Arrhenius rates given by [14; 40; 42]) incorporates a global reaction step for methane oxidation, and a reversible reaction to convert CO to CO_2 in the product stream. The fuel equivalence ratio of the methane-air mixture is $\phi = 1$. For additional computational convenience here, the flame chemistry and transport were modified to artificially thicken the flame so that the thermal laminar flame thickness is $\delta_L = 0.7$ mm, and to adjust its propagation speed to $s_L = 36$ cm/s. These values approximately match those of the corresponding laminar flame computed with a more detailed transport model and the GRI-Mech 3.0 [15; 35] mechanism. The modifications were accomplished by uniformly increasing all transport coefficients by a factor of 2, and reducing chemical production rates uniformly by a factor of 3, following ideas discussed in [12].

The time-dependent calculation is performed using the flame sheet configuration discussed above. Flow enters a 2D domain through the bottom boundary, proceeds upward through a dynamically wrinkling flame surface, and exits the top outflow boundary. The side walls are periodic. The length of the inlet face $L = 28.6\,\delta_L$, and the height of the domain $H = 2L$. The fluctuations are generated in an auxiliary calculation prior to the controlled flame simulation. A random velocity field is generated on a $L \times L$ domain with an energy spectrum of the form

$$E(k) = \frac{(\frac{k}{k_i})^4}{[1+(\frac{k}{k_i})^2]^{\frac{17}{6}}} \exp\left[-\frac{9}{4}\left(\frac{k}{k_d}\right)^{\frac{4}{3}}\right]$$

where k is the wavenumber, $k_d = 1/(2\Delta x)$, and k_i is the peak frequency, which is adjusted empirically to give the desired integral scale.. This form is characteristic of 2D decaying isotropic turbulence [21]. Rather than using the random field directly, we first evolve it for several eddy turnover times using an incompressible Navier-Stokes solver [1] at resolution comparable to the finest meshes in the reacting flow simulations to ensure that the phases are realistic (see below). To accommodate this evolution the initial field is generated at a somewhat higher turbulence intensity and the incompressible evolution is continued until the turbulent intensity reaches the desired level. The resulting fluctuations have an effective integral scale length $\ell_t \sim 2.6\,\delta_L$ and turbulent intensity $u' \sim 1.7 s_L$. They are added to a mean vertical flow given dynamically by the feedback control algorithm to model the advection of turbulence through the inflow boundary. By cycling through the periodic fluctuation data, this technique provides an endless source of fluctuations with a periodicity length L. The amount of corresponding time for cycling through the data is dependent on the (time-dependent) control velocity. The system is on the boundary between the corrugated and distributed flamelet regime [28], but also very nearly laminar.

The simulation is carried out with a three-level adaptive grid hierarchy. The refinement criteria is such that the flame surface remains resolved with a uniform grid spacing at the finest level of $\Delta x = 39\ \mu$m. The base grid covering the entire domain is a factor of four coarser, and an intermediate level a factor of two finer than the base grid is used to resolve the turbulent fluctuations between the inlet boundary and the flame surface.

A steady solution obtained from the PREMIX code [22] and the identical transport and chemistry models is used to initialize a flat flame parallel to the inlet face. The flame position is initially below the target height of $\beta = 5$ mm above the inlet boundary. The flame is evolved using the control algorithm to automatically adjust the inflow rate.

Figure 3 shows the flame location and control velocity as a function of time over approximately 75 integral-scale eddy turnover periods, $\tau_t = \ell_t/u' \approx 1.8$ ms. The initial transient indicates that the control quickly increases the inflow rate to shift the flame upward. The flame overshoots the target so the inflow velocity is adjusted automatically to zero for a short time interval. After the flame burns back upstream to the set point, both the control and the burning speed briefly settle into a value, about 38 cm/s, that is near the speed of a flat laminar flame. During this initial phase the inflowing mixture is carrying decaying turbulence toward the flame, which is only slightly wrinkled. At approximately 10 ms, the fluctuations begin to wrinkle the flame causing a dramatic increase in flame surface area and a corresponding increase in the burning speed. The control algorithm increases the inflow rate in response to flame surface area perturbations so as to maintain a constant volume of unburned mixture. Note that the large periodic transients in fuel consumption correspond to flame topology changes such as localized necking and pinching off of flame fragments, but that the volume of unburned mixture is steady as indicated by the nearly constant mean flame position.

This example demonstrates that for atmospheric stoichiometric premixed methane flames in this corrugated flamelet regime, our control algorithm is sufficiently robust to stabilize the flame in the computational domain, allowing the collection of detailed flame statistics. In Figure 3, we observe that after the initial transients, the flame speed exhibits a cyclic repetition with a period of approximately 17 τ_t, corresponding to the time to traverse the auxiliary file of turbulent fluctuations. With our current approach for introducing turbulent fluctuations, the size of the auxiliary fluctuation file effectively places an upper bound on the scales of temporal dynamics that are representable; however, there are several potential strategies for

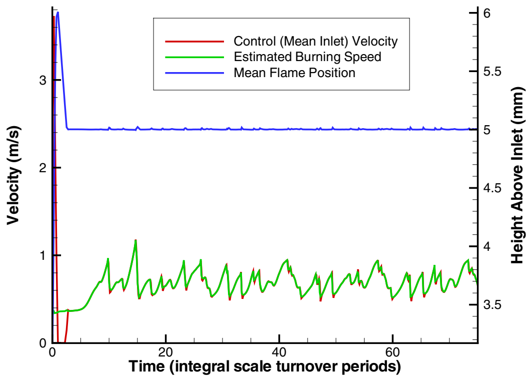


Figure 3. *Performance of control algorithm for* $\phi = 1.0$ *case with simplified chemistry.*

modifying the turbulence description and continuing the simulation if a longer integration or a more diverse set of temporal scales are required.

3.2. ***GRI-Mech 3.0 Mechanism.*** We now apply the control methodology described above to a series of methane flames modeled in significantly greater detail, using the GRI-Mech 3.0 chemistry mechanism (53 species, 325 reactions) and a mixture-averaged diffusive transport model. Three flames are chosen to highlight variations observed in a methane flame's response to flowfield flame surface curvature (see, for example, Tseng et al. [39]). The three cases have stoichiometries, $\phi = 0.55$, 0.75, 1.0. Table 1 lists various properties of the corresponding steady laminar one-dimensional flame solutions computed using the PREMIX [22] code. As before, the computational domain in all three cases is periodic in the horizontal direction with inflow on the bottom face and outflow at the top. In all three cases, the computational domains have dimensions $L \times H = 46\,\delta_L \times 92\,\delta_L$. The fluctuations in the inflow stream were generated for each case separately using a process identical to that discussed in the first example. The resulting fluctuations had an effective integral scale length $\ell_t \sim 2.6\delta_L$ and turbulent intensity $u' \sim 1.7 s_L$, measured with respect to the properties of each flame.

Adaptive mesh refinement was used in all the simulations to maintain approximately 22 uniform grid cells across the thermal width of the flames throughout their evolution. Dynamic refinement for these simulations was based on the magnitude of vorticity and on a flame marker, CH_3. In each case, we waited until the flame height stabilized before collecting the statistical analysis data. The time-dependent data represents snapshots of the three cases taken at uniform intervals over approximately five τ_t.

Table 1. *Characteristics of the three laminar methane-air flames of different stoichiometries at* 1 *atmosphere. Thermal flame thickness is calculated as the change in temperature through the flame divided by the maximum temperature gradient,* $\delta_L = (T_{\max} - T_{\min}) / \max \|\nabla T\|$.

fuel equivalence ratio ϕ	thermal flame thickness δ_L (μm)	flame speed s_L (cm / s)	fuel consumption rate (g / cm s)	isotherm of peak heat release (K)	peak local fuel consumption (mg / mL s)
1.00	433	36.2	0.2380	1684	134
0.75	584	22.34	0.1070	1516	51.3
0.55	1313	7.62	0.0273	1379	7.03

4. Analysis of the GRI-Mech 3.0 flames

4.1. ***Appearance of the flames.*** Representative snapshots of the temperature fields are shown in Figure 4. The three flames of different stoichiometries appear qualitatively similar, as expected given that the flames are at the same point on the regime diagram for premixed turbulent combustion, the so-called Borghi diagram [27]. At any instant in time, the flame surface shows the characteristic wrinkling expected of a turbulent premixed flame, namely, regions where the flame is smoothly bowed toward the reactants separated by sharper cusps protruding into the burned region. Since the bows are the larger geometric feature, they consume more of the unburned mixture whose amount in the domain is kept constant by the control. Thus the bows are relatively stable in the frame of reference of the computational domain. The behavior at the cusps is more dynamic. Cusps are observed to periodically grow into elongated channels after which there is period of apparent rapid movement when the sides of the channel close upon each other and the cusp returns to a more typical position relative to the rest of the flame. Occasionally in this process, a channel will burn through in its center detaching a bubble of unburned fuel surrounded by products. An example of this is shown in the snapshot of the $\phi = 1$ flame in Figure 4 where an elongated channel extends through the periodic boundary. Here, the unburned mixture at the cusp is about to detach. Extinction, marked by

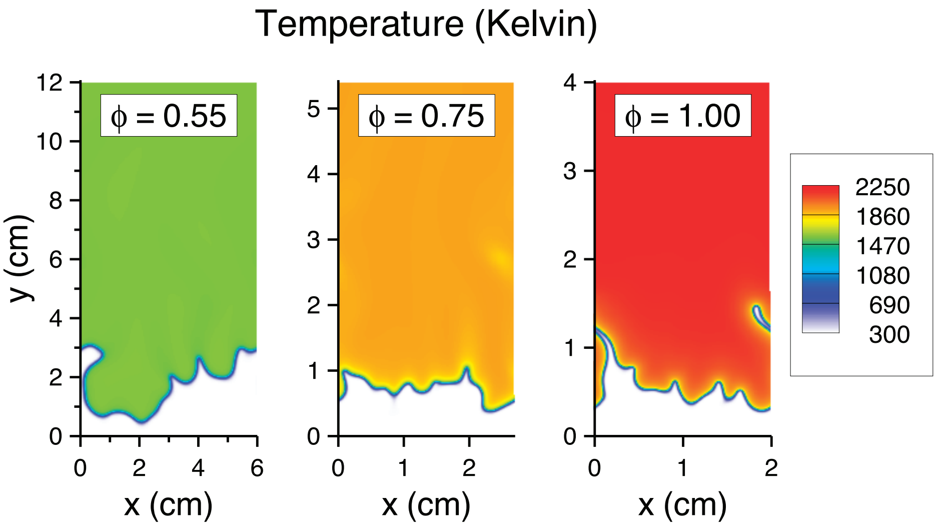


Figure 4. *Temperature in the three flames.*

dramatic and sudden reductions in fuel consumption along the flame surface, was not observed in any of the cases.

We examine the distribution of the curvature of the flame over the course of the simulation to quantitatively demonstrate the similarity of the flames. As indicated in Table 1, we associate the location of the flame with a particular isotherm. The vector field of unit normals to all the isotherms throughout the domain can be calculated as a $\hat{n} = -\nabla T / \|\nabla T\|$ using centered differences on the underlying uniform, rectangular meshes. Note these normals have been chosen to point toward the cold, unburned mixture. The curvature of the isotherms is then $\kappa = \nabla \cdot \hat{n}$ again evaluated throughout the domain using centered differences. We then interpolate this κ to the isotherm corresponding to the peak heat release from the laminar flame solution, which we use as the operational definition of the flame surface. With this definition, the curvature is negative at cusps and positive in the bowed regions.

When the curvature is scaled to the laminar flame thermal thickness, the probability density functions (PDFs) of curvature for all three flames are coincident, indicating that all three flames are experiencing the same degree of wrinkling. See Figure 5. These curves are the probability of finding a portion of flame with the given value of curvature while the flame evolves through several hundred time steps (spanning at least five eddy turnover periods) once reaching a statistically stationary state. We note that the distributions peak slightly to the positive side of zero. In general there is a greater probability of finding positive curvature (the bowed regions), but at high curvature the distributions show a strong bias toward negative values (the cusps). This skewness, emphasized here by the choice a log scale on the ordinate, is typical of turbulent flames, as noted above. Finally, we note that a nontrivial fraction of the flame surface is subject to curvature that is not "small."

These flame dynamics are all consistent with the regime diagram's characterization of these flames as being in the flamelet regime. Flames in the corrugated and wrinkled flamelet regimes tend to maintain a well-defined flame front structure with nearly parallel isocontours of species and temperature. A detailed attempt to base the regime diagram on observations of 2D direct numerical simulations was carried out by Poinsot, Veynante, and Candel [30] using interactions between flames and single vortex pairs. Their work could be successfully extended to long-duration observations of flames in more complicated, stochastic flow fields using the control strategy developed here.

4.2. ***Global Turbulent Burning Speed.*** For the initial analysis of the results, we look first at the effective turbulent flame speed S_c^G, defined in terms of the integrated

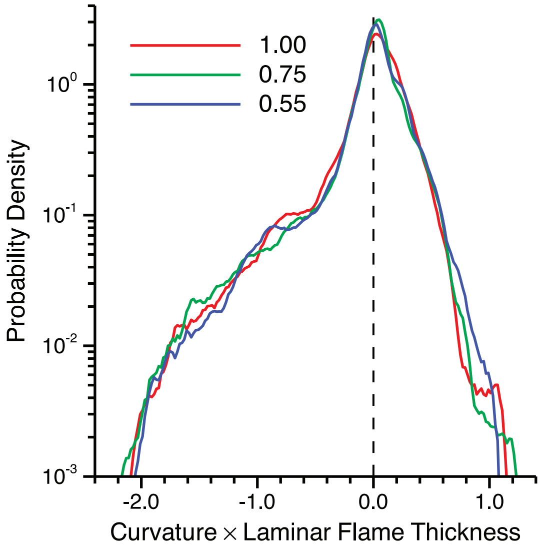


Figure 5. *Probability density of curvature scaled by laminar flame thickness for the three flames of different stoichiometries. Density is calculated by a moving average over 5 intervals of width 0.02 (nondimensional) on the horizontal axis.*

fuel consumption

$$S_c^G = \frac{1}{A_L \left(\rho Y_{\mathrm{CH}_4}\right)_{\mathrm{in}}} \int_\Omega \rho\omega_{\mathrm{CH}_4}\, d\Omega$$

where A_L is the area of the flat laminar flame (ie, the width of the domain, L), $\left(\rho Y_{\mathrm{CH}_4}\right)_{\mathrm{in}}$ is the inflowing methane mass density and $\rho\omega_{\mathrm{CH}_4}$ is the rate of methane mass consumption. In Figure 6 we plot S_c^G, normalized by S_L, versus time, normalized by τ_t, for each case. In these figures, the dramatic drops in turbulent speed correspond to rapid flame area loss at the burning of long thin channels, and to the rapid consumption of detached pockets of unburnt material. The plot demonstrates a large (20–50%) variability in the instantaneous turbulent flame speeds for all cases. When examined at the length and time scales representative of the computation, it makes little sense to talk about turbulent flame speed as a single number. More revealing data may be the PDFs of turbulent flame speed shown in Figure 7. These PDFs are centered at 200–250% of S_L, and are quite broad. The $\phi = 0.75$ case appears bimodal; however, it is not clear if this is a real effect or evidence of a lack of adequate statistics.

We now explore the relationship between aggregate fuel consumption rate and the flame area resulting from wrinkling due to the inflow fluctuations. Figure 8 shows a scatter plot of S_c^G versus the instantaneous flame area A^G (or, length of isotherm

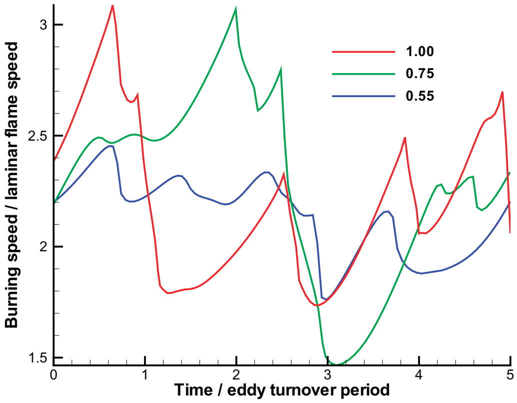


Figure 6. *Turbulent flame speed for the three flames.*

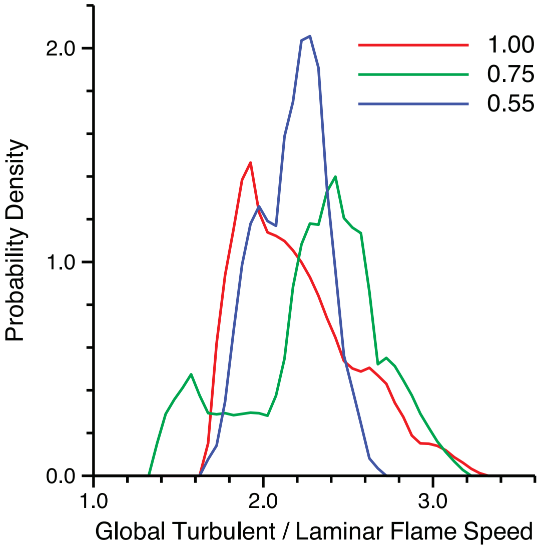


Figure 7. *Probability density of overall turbulent speedup for the three flames of different stoichiometries. Density is calculated by a moving average over 5 intervals of width 0.05 (nondimensional) on the horizontal axis.*

contour we associated with the flame surface at that instant in time). The symbols represent data from solutions taken at uniform intervals throughout the sample period. To a very good approximation, the fuel consumption rate in the domain scales with the overall area of the flame for all three stoichiometries. Thus, at least on average, the turbulent flame speed is directly proportional to the flame area, even

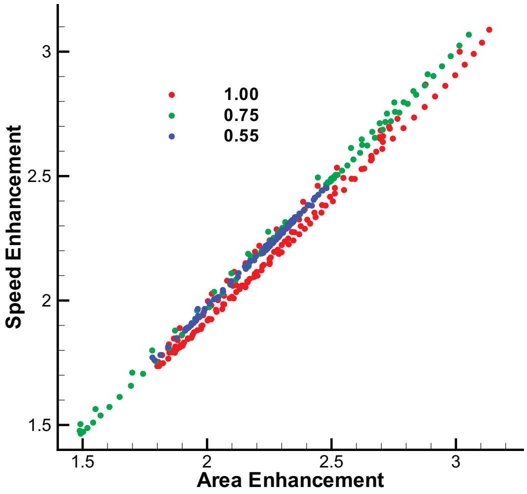


Figure 8. *Turbulent flame speed versus flame area.*

across the large excursions in turbulent flame speed. Note that the stoichiometric flame is slightly slower than predicted by its area and the laminar flame speed. This reflects associated changes in Markstein number with ϕ, which are discussed in more detail in the next section.

4.3. ***Local Burning Speed Behavior.*** In this section, we look at the local flame behavior in more detail. To refine the analysis of flame speed we look at the variation in fuel consumption along the flame surface for each of the three cases. Figure 9 shows representative samples for each flame with a blow up of a localized region of high curvature. For the $\phi = 1.0$ flame, we see a dramatic enhancement in fuel consumption at the cusps, which corresponds to a region of large negative curvature. We observe a comparable reduction in fuel consumption in regions of large positive curvature. Similar but less pronounced behavior is observed for $\phi = 0.75$; however, for $\phi = 0.55$ the observed trends reverse with higher fuel consumption in regions of positive curvature and lower fuel consumption in regions of negative curvature.

We would like to relate this change in the behavior of the fuel consumption to the behavior of the local flame speed. There are several potential definitions of local flame speed; see, e.g., Poinsot and Veynante [29] for a discussion of possible choices. Here we will define a local flame speed based on integrated local fuel consumption in the following way. To define the integrals we will define local coordinates near the flame using arclength along the flame and a normal coordinate defined in terms of a progress variable, c, defined such that $c = 0$ in the unburned reactants, and $c = 1$ in the products. The progress variable may be based on any scalar variable that is

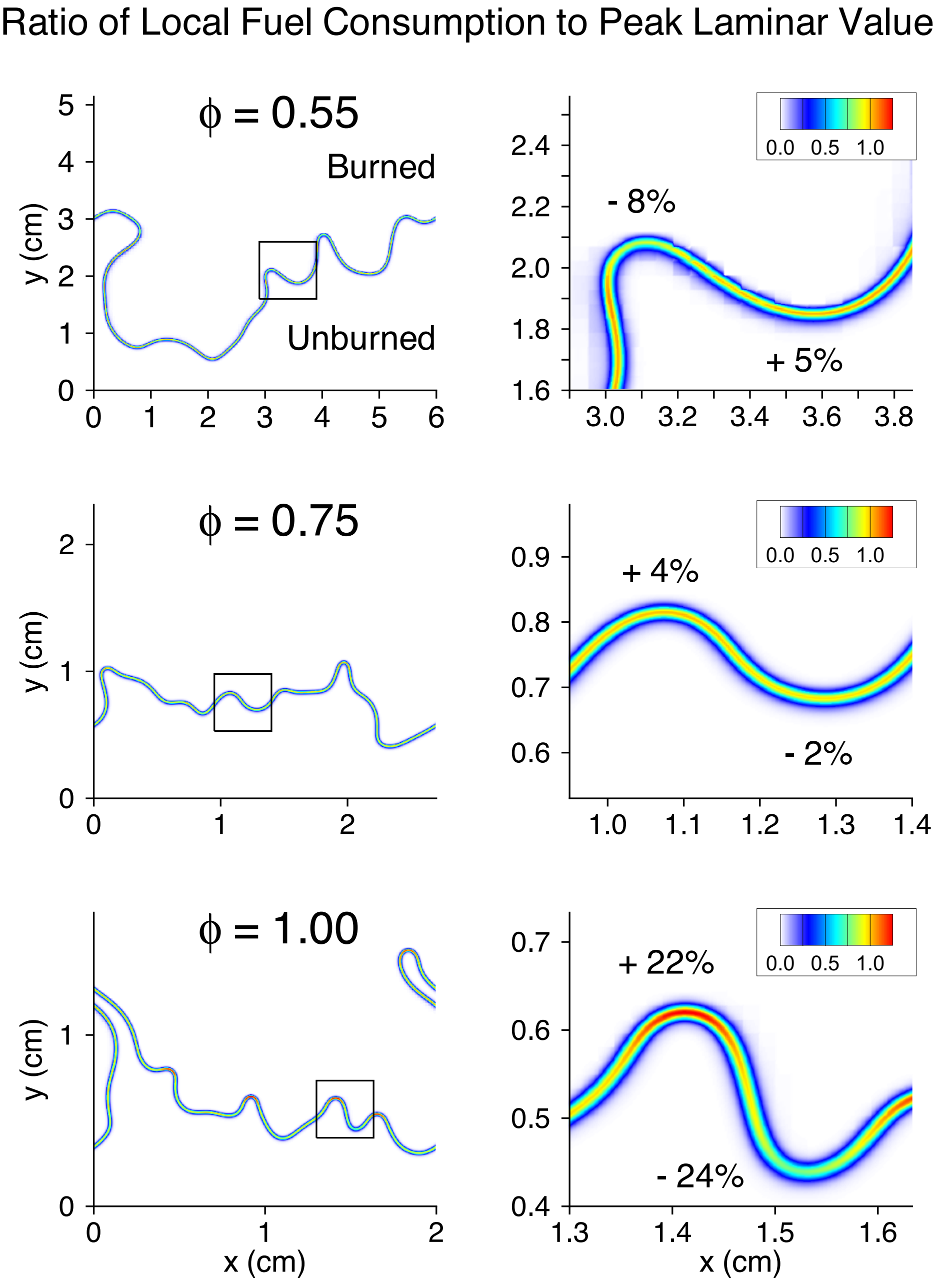


Figure 9. *Fuel consumption is often used as a measure of local flame speed. This figure depicts the ratio of local methane consumption to peak consumption in unstretched laminar flames of identical fuel equivalence ratios. Reference values are given in Table 1.*

monotonic across the flame surface; here, we will use normalized temperature to the define the progress variable.

At uniform intervals along the flame, we extend local normals by following integral curves of the gradient of c toward both the products and fuel. These normals define a series of adjacent disjoint wedge-shaped volumes, Ω, surrounding the flame, and extending well beyond the region of high chemical reactivity. A local burning speed may then be defined over each of these volume:

$$S_c^\ell = \frac{1}{A^\ell \left(\rho Y_{\mathrm{CH}_4}\right)_{\mathrm{in}}} \int_\Omega \rho \omega_{\mathrm{CH}_4} d\Omega \tag{2}$$

where A^ℓ is the area (length) of the intersection of Ω with the flame.

A typical example of a set of such normals, and the resulting wedge-shaped volumes is depicted in Figure 10. The example is taken from the $\phi = 1.0$ case, and includes the instantaneous advection streamlines superimposed for reference. Defining the local speed in this way has the property that the turbulent burning speed is its area-weighted average:

$$S_c^G = \sum_{i=1}^{Nwedges} S_c^{\ell,i} \frac{A^{\ell,i}}{A^G} .$$

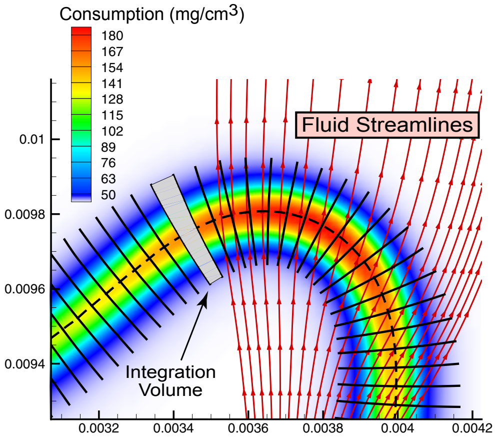


Figure 10. *Construction of local control volumes at a typical flame surface. The volumes are centered on the flame, and extend normal to the local isopleths in progress variable at uniform intervals along the flame. Adjacent flame normals define a volume over which we define the local consumption-based burning speed.*

where *Nwedges* is the number of discrete wedge-shaped volumes tiling the entire flame surface. In addition to preserving the total integral of fuel consumption, these integrals are relatively easy to evaluate accurately and provide a fairly robust characterization of the local burning. Evaluation of the intersected area, $A^{\ell,i}$, is sensitive to the definition of the flame surface (in this case, the choice of progress variable and of its isocontour) which can introduce a small bias into the curvature correlation. In the present cases, however, the fuel consumption profile takes on non-negligible values over a relatively limited range in temperature, so we can minimize this bias by ensuring the flame isotherm is centered near this narrow peak. The values chosen in Table 1 correspond to the peak in heat release for the corresponding steady flat flame solution.

The data in Figure 9 shows a clear dependence of the local fuel consumption on the local flame curvature. To make the notion more precise, we form the consumption-based local flame speed S_c^ℓ at each segment along the flame as discussed above and form scatter plots, shown in Figure 11, of the local flame speed normalized by the laminar flame speed with the curvature normalized by thermal flame thickness. The scatter plots confirm the trend shown in Figure 9, namely, that the $\phi = 1.0$ flame correlates negatively with curvature while the $\phi = 0.55$ flame correlates positively. In addition, the relative insensitivity of the $\phi = 0.75$ flame is apparent. If we associate a curvature Markstein number, $\mathcal{M}_\kappa$, with the slope of the correlation for each case in Figure 11, then the data matches the trend reported in [39], including the change of sign of the Markstein number near $\phi = 0.75$. The magnitude of the Markstein number is sensitive to the definitions of flame thickness, burning speed, flame isopleth, etc. For this reason, it is difficult in general to make detailed quantitative comparisons with the results from other numerical and experimental studies.

Each of the scatter plots shows a number of outlier points, most notably around normalized curvature of -1. To explain this phenomena, we note that in rare situations the regions used to define the integrated local flame speed can become overly distorted or poorly defined. These correspond to regions where an elongated cusp closes, or when the sides of an elongated cusp burn together and change the local topology of the flame. In both cases, an ambiguity in definition of the wedge-shaped regions develops approximately when the flame thickness is equal to the local radius of curvature, that is, where the magnitude of the normalized curvature is unity.

From wrinkled flame theory, as explained for example by Peters [28] and Poinsot and Veynante [29], we expect the local flame speed to correlate with stretch, which combines the effects of curvature, κ, and strain tangential to the flame surface, $\mathcal{S}_t = \hat{t} \cdot \nabla\vec{v} \cdot \hat{t}$, where $\hat{t} \perp \hat{n}$ is the unit vector locally tangent to the flame. The evaluation of stretch in an idealized setting of an “infinitely” thin flame propagating

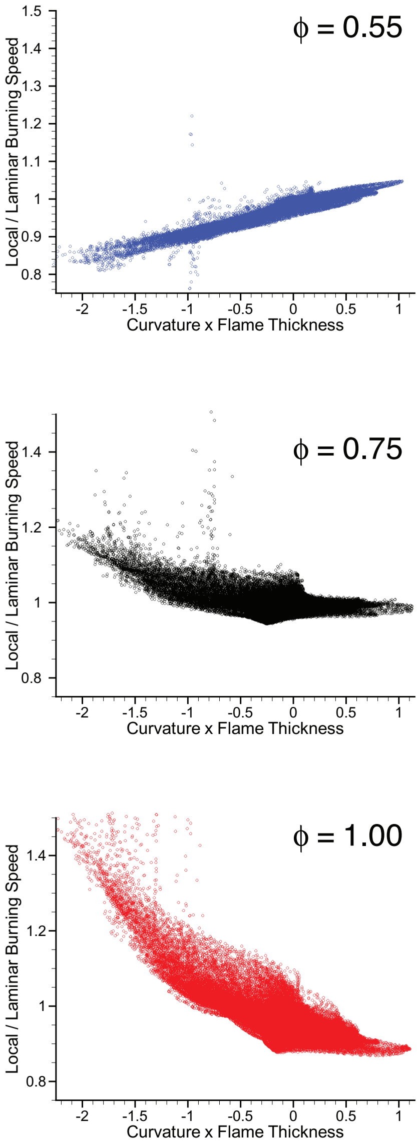


Figure 11. *Scatter plot of local turbulent flame speed scaled by laminar flame speed versus of curvature scaled by laminar flame thickness, for the three flames of different stoichiometries.*

through a fluid is fairly straightforward. In the present setting, where the flame is being resolved and has a finite thickness, it is unclear how to evaluate the strain term in the definition of stretch. There is a large literature on the generalization of classical flame theory to"thick" flames, see [11; 17; 16; 18] and references therein. We pursued several possible approaches to computing stretch; however, local definitions of the stretch appear to be highly sensitive to the method of evaluating the strain rate. Furthermore, for the approaches we considered, the effects of strain on speed-versus-stretch correlations were entirely explained by the correlation of strain with curvature. A similar observation was made by Haworth and Poinsot [20]. Pope [31] also discusses the interrelationship of curvature and strain. Consequently, at least for the flames considered here, the variation in consumption speed along the flame is essentially a function of curvature alone. The difficulties with defining a local strain rate for the definition of stretch suggests that an integral-based approach, as for example [18; 25], is needed to obtain a more robust and physically meaningful method for computing stretch.

5. Conclusions

We have introduced a new computational tool based on applying a feedback mechanism to control and stabilize a turbulent flame in a simple two dimension geometry without introducing a geometric stabilization mechanism such as a flow obstruction or a stagnation plate. We have used this tool to study the behavior of premixed turbulent methane flames in two dimensions. For these simulations we examined both the global flame behavior and the dependence of the local flame speed on flame curvature. By using the control algorithm, we are able to hold the flame at conditions that are statistically stationary, enabling us to obtain detailed diagnostics for an ensemble of snapshots of the flame at the same turbulent conditions. For the methane flame considered here, the simulations show that although the global burning speed correlates well with the global flame area, there is substantial variation in local burning speed over the flame for $\phi = 0.55$ and $\phi = 1.00$. These variations are shown to correlate well with curvature: the negative correlation at $\phi = 1.00$ and a positive correlation at $\phi = 0.55$ reflect a change in Markstein number for methane combustion as a function of equivalence ratio. In future work, we will present a more detailed analysis of local flame dynamics and flame chemistry. In addition, the methodology presented here extends in a straightforward fashion to three dimensions. Applications to three-dimensional turbulent flames will also be presented in future work.

Acknowledgments

The authors would like to thank Jonathan Goodman and Alan Laub for helpful discussions in the course of this work. This research used resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The authors were supported by the Office of Science through the Office of Advanced Scientific Computing Research, Mathematical, Information, and Computational Sciences Division under U.S. Department of Energy contract DE-AC03-76SF00098.

References

[1] Ann S. Almgren, John B. Bell, Phillip Colella, Louis H. Howell, and Michael L. Welcome, *A conservative adaptive projection method for the variable density incompressible Navier-Stokes equations*, J. Comput. Phys. **142** (1998), no. 1, 1–46. MR 99k:76096

[2] M. Baum, T. J. Poinsot, D. C. Haworth, and N. Darabiha, *Direct numerical simulation of $H_2/O_2/N_2$ flames with complex chemistry in two-dimensional turbulent flows*, J. Fluid Mech. **281** (1994), 1–32.

[3] J. B. Bell, M. S. Day, and J. F. Grcar, *Numerical simulation of premixed turbulent methane combustion*, Proc. Combust. Inst. **29** (2002), 1987–1993.

[4] J. B. Bell, M. S. Day, I. G. Shepherd, M. Johnson, R. K. Cheng, J. F. Grcar, V. E. Beckner, and M. J. Lijewski, *Numerical simulation of a laboratory-scale turbulent V-flame*, Proc. Natl. Acad. Sci. USA **102** (2005), no. 29, 10006–10011.

[5] D. Bradley, *How fast can we burn?*, Proc. Combust. Inst. **24** (1992), 247–262.

[6] Peter E. Caines, *Linear stochastic systems*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1988. MR 89f:93037

[7] N. Chakraborty and S. Cant, *Unsteady effects of strain rate and curvature on turbulent premixed flames in an inflow outflow configuration*, Combust. Flame **137** (2004), 129–147.

[8] Guanrong Chen, Goong Chen, and Shih-Hsun Hsu, *Linear stochastic control systems*, CRC Press, Boca Raton, FL, 1995. MR 97g:93001

[9] Y.-C. Chen, P. A. M. Kalt, R. W. Bilger, and N. Swaminathan, *Effects of mean flow divergence on turbulent scalar flux and local flame structure in premixed turbulent combustion*, Proc. Combust. Inst. (2002), 1863–1871.

[10] R. K. Cheng and I. G. Shepherd, *The influence of burner geometry on premixed turbulent flame propagation*, Combust. Flame **85** (1991), 7–26.

[11] S. H. Chung and C. K. Law, *An integral analysis of the structure and propagation of stretched premixed flames*, Combust. Flame **72** (1988), 325–336.

[12] O. Colin, F. Ducros, D. Veynante, and T. Poinsot, *A thickened flame model for large eddy simulations of turbulent premix ed combustion*, Phys. Fluids **12** (2000), no. 7, 1843–1863.

[13] M. S. Day and J. B. Bell, *Numerical simulation of laminar reacting flows with complex chemistry*, Combust. Theory Modelling **4** (2000), 535–556.

[14] F. L. Dryer and I. Glassman, *High-temperature oxidation of CO and CH_4*, Proc. Combust. Inst. **14** (1972), 987–1003.

[15] M. Frenklach, H. Wang, M. Goldenberg, G. P. Smith, D. M. Golden, C. T. Bowman, R. K. Hanson, W. C. Gardiner, and V. Lissianski, *GRI-Mech—An optimized detailed chemical reaction mechanism for methane combustion*, Gas Research Institute Report GRI-95/0058, Gas Research Institute, 1995, See also [35].

[16] L. P. H. de Goey, R. M. M. Mallens, and J. H. M. ten Thije Boonkkamp, *An evaluation of different contributions to flame stretch for stationary premixed flames*, Combust. Flame **110** (1997), 54–66.

[17] L. P. H. de Goey and J. H. M. ten Thije Boonkkamp, *A mass-based definition of flame stretch for flames with fine thickness*, Combustion Science and Technology **122** (1997), 399–405.

[18] ———, *A flamelet description of premixed laminar flames and the relation with flame stretch*, Combust. Flame **119** (1999), 253–271.

[19] D. C. Haworth, R. J. Blint, B. Cuenot, and T. J. Poinsot, *Numerical simulation of turbulent propane-air combustion with nonhomogeneous reactants*, Combust. Flame **121** (2000), 395–417.

[20] D. C. Haworth and T. J. Poinsot, *Numerical simulations of Lewis number effects in turbulent premixed flames*, J. Fluid Mech. **244** (1992), 405–436.

[21] J. O. Hinze, *Turbulence*, 2 ed., McGraw-Hill, 1975.

[22] R. J. Kee, J. F. Grcar, M. D. Smooke, and J. A. Miller, *PREMIX: A fortran program for modeling steady, laminar, one-dimensional premixed flames*, Technical Report SAND85-8240, Sandia National Laboratories, Livermore, 1983.

[23] Harold Kushner, *Introduction to stochastic control*, Holt, Rinehart and Winston, Inc., New York, 1971. MR 43 #5969 Zbl 0293.93018

[24] D. Most, F. Dinkelacker, and A. Leipertz, *Lifted reaction zones in premixed turbulent bluff-body stabilized flames*, Proc. Combust. Inst. **29** (2002), 1801–1806.

[25] J. A. van Oijen, G. R. A. Groot, R. J. M. Bastiaans, and L. P. H. de Goey, *A flamelet analysis of the burning velocity of premixed turbulent expanding flames*, Proc. Combust. Inst. **30** (2005), no. 1, 657–664.

[26] R. B. Pember, L. H. Howell, J. B. Bell, P. Colella, W. Y. Crutchfield, W. A. Fiveland, and J. P. Jessee, *An adaptive projection method for unsteady, low-Mach number combustion*, Combust. Sci. Technol. **140** (1998), 123–168.

[27] N. Peters, *Laminar flamelet concepts in turbulent combustion*, Proc. Combust. Inst. **21** (1986), 1231–1250.

[28] Norbert Peters, *Turbulent combustion*, Cambridge Monographs on Mechanics, Cambridge University Press, Cambridge, 2000. MR 2001j:80007

[29] T. Poinsot and D. Veynante, *Theoretical and numerical combustion*, R. T. Edwards, Inc., Philadelphia, 2001.

[30] T. Poinsot, D. Veynante, and S. Candel, *Diagrams of premixed turbulent combustion based on direct simulation*, Proc. Combust. Inst. **23** (1990), 613–619.

[31] S. B. Pope, *The evolution of surfaces in turbulence*, Internat. J. Engrg. Sci. **26** (1988), no. 5, 445–469. MR 89c:76073 Zbl 0641.76054

[32] C. J. Rutland and A. Trouvé, *Direct simulations of premixed turbulent flames with non-unit Lewis numbers*, Combust. Flame **94** (1993), 41–57.

[33] S. S. Sattler, D. A. Knaus, and F. C. Gouldin, *Determination of three-dimensional flamelet orientation distributions in turbulent V-flames from two-dimensional image data*, Proc. Combust. Inst. **29** (2002), 1785–1795.

[34] I. G. Shepherd, R. K. Cheng, T. Plessing, C. Kortschik, and N. Peters, *Premixed flame front structure in intense turbulence*, Proc. Combust. Inst. **29** (2002), 1833–1840.

[35] G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, W. C. Gardiner Jr., V. V. Lissianski, and Z. Qin, *GRI-Mech 3.0*, Unpublished. See also [15].

[36] M. Tanahashi, M. Fujimura, and T. Miyauchi, *Coherent fine scale eddies in turbulent premixed flames*, Proc. Combust. Inst. **28** (2000), 529–535.

[37] M. Tanahashi, Y. Nada, Y. Ito, and T. Miyauchi, *Local flame structure in the well-stirred reactor regime*, Proc. Combust. Inst. **29** (2002), 2041–2049.

[38] Arnaud Trouvé and Thierry Poinsot, *The evolution equation for the flame surface density in turbulent premixed combustion*, J. Fluid Mech. **278** (1994), 1–31. MR 95g:80012 Zbl 0825.76899

[39] L.-K. Tseng, M. A. Ismail, and G. M. Faeth, *Laminar burning velocities and Markstein numbers of hydrocarbon / air flames*, Combust. Flame **95** (1993), 410–425.

[40] C. K. Westbrook and F. L. Dryer, *Simplified reaction mechanisms for the oxidation of hydrocarbon fuels in flames*, Combust. Sci. Technol. **27** (1981), 31–43.

[41] S. Zhang and C. J. Rutland, *Premixed flame effects on turbulence and pressure-related terms*, Combust. Flame **102** (1995), 447–461.

[42] V. L. Zimont and Y. M. Trushin, *Total combustion kinetics of hydrocarbon fuels*, Comb. Expl. Shock Wave **5** (1969), 391–194.

Received August 30, 2005. Revised November 10, 2005.

JOHN B. BELL: jbbell@lbl.gov
Lawrence Berkeley National Laboratory, Mail Stop 50A-1148, 1 Cyclotron Road, Berkeley, CA 94720-8142, United States

MARCUS S. DAY: msday@lbl.gov
Lawrence Berkeley National Laboratory, Mail Stop 50A-1148, 1 Cyclotron Road, Berkeley, CA 94720-8142, United States
http://seesar.lbl.gov/ccse/people/marc/index.html

JOSEPH F. GRCAR: jfgrcar@lbl.gov
Lawrence Berkeley National Laboratory, Mail Stop 50A-1148, 1 Cyclotron Road, Berkeley, CA 94720-8142, United States
http://seesar.lbl.gov/ccse/people/grcar/index.html

MICHAEL J. LIJEWSKI: mjlijewski@lbl.gov
Lawrence Berkeley National Laboratory, Mail Stop 50A-1148, 1 Cyclotron Road, Berkeley, CA 94720-8142, United States
http://seesar.lbl.gov/ccse/people/lijewski/index.html

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

ESTIMATING HYDRODYNAMIC QUANTITIES IN THE PRESENCE OF MICROSCOPIC FLUCTUATIONS

ALEJANDRO L. GARCIA

This paper discusses the evaluation of hydrodynamic variables in the presence of spontaneous fluctuations, such as in molecular simulations of fluid flows. The principal point is that hydrodynamic variables such as fluid velocity and temperature must be defined in terms of mechanical variables such as momentum and energy density). Because these relations are nonlinear and because fluctuations of mechanical variables are correlated, care must be taken to avoid introducing a bias when evaluating means, variances, and correlations of hydrodynamic variables. The unbiased estimates are formulated; some alternative, incorrect approaches are presented as cautionary warnings. The expressions are verified by numerical simulations, both at thermodynamic equilibrium and at a nonequilibrium steady state.

1. Introduction

Particle simulations are a useful tool in the study of continuum mechanics, especially fluid mechanics [15; 16], and a variety of particle-based algorithms (e.g., molecular dynamics [7], particle-in-cell (PIC) [12], direct simulation Monte Carlo (DSMC) [4], dissipative particle dynamics (DPD) [10], and lattice gas automata (LGA) [24]) are available to simulate hydrodynamic phenomena. In such simulations, the quantities of interest are not the precise trajectories of the particles but rather the hydrodynamic variables such as density, fluid velocity, temperature, pressure, etc. Compared to macroscopic systems, the number of particles in a simulation is small (typically fewer than 10^7) so the number of particles in a volume element is typically on the order of 10 to 100. For this reason, the spontaneous fluctuations in a volume element are significant and statistical samples are taken. The purpose of this paper is to establish the correct construction for measuring hydrodynamic variables and to point out some common errors that lead to biased results.

The bias described in this paper has already been studied in detail by Tysanner and Garcia [26; 25] for the measurement of mean fluid velocity. This paper extends that work in two important directions. First, we consider other hydrodynamic variables,

MSC2000: primary 82C80; secondary 76M28.
Keywords: fluctuations, molecular simulations, hydrodynamics.

most significantly temperature. Second, the study of hydrodynamic fluctuations is an important topic in a variety of fields ranging from nanoscale fluid mechanics [5; 13] to molecular biology [14; 23]. We therefore also consider the measurement of hydrodynamic fluctuations, such as the variance of fluid velocity and the correlation of density and temperature fluctuations.

The paper is organized as follows: Section 2 defines mechanical densities and relates them to hydrodynamic variables, specifically how the mean values of the latter are defined in terms of the former. Variances and correlations of hydrodynamic quantities are similarly described in Section 3. The bias observed when hydrodynamic quantities are measured incorrectly is described in Section 4 where the effects are illustrated by numerical results from simulations. Section 5 summarizes the main points and concludes with general remarks.

2. Mean values

First let us establish some notation: Consider a fluid of particles of mass m. The position and velocity of particle k are $\mathbf{r}_k$ and $\mathbf{v}_k$. The measurement of mechanical variables in a cell, namely the instantaneous densities of mass, momentum, and kinetic energy, may be written as,

$$\rho = \frac{1}{V} \sum_{\mathbf{r}_k \in C} m \tag{1}$$

$$\mathbf{J} = \frac{1}{V} \sum_{\mathbf{r}_k \in C} m\mathbf{v}_k \tag{2}$$

$$K = \frac{1}{V} \sum_{\mathbf{r}_k \in C} \tfrac{1}{2} m |\mathbf{v}_k|^2 \tag{3}$$

where the sums are over particles located within cell C, which has volume V. One may define other mechanical variables but these suffice for the present discussion. For the equations of fluid dynamics these are the fundamental conserved variables.

For any of these mechanical variables ($\mathcal{M} = \rho$, $\mathbf{J}$, or K) we may write the sample mean as the average over S samples, that is,

$$\langle \mathcal{M} \rangle_s = \frac{1}{S} \sum_{j=1}^{S} \mathcal{M}_j \tag{4}$$

where the subscript j indicates individual samples, which may be from an ensemble of runs or, for steady state problems, samples taken at different times (i.e., a time average). In the limit of infinitely many samples, this sample mean goes to the mean value, that is,

$$\overline{\mathcal{M}} = \langle \mathcal{M} \rangle_\infty \equiv \lim_{S\to\infty} \langle \mathcal{M} \rangle_S. \tag{5}$$

It is important to keep in mind that we are *not* considering the "thermodynamic limit" because our interest is in the measurement of fluid variables in relatively small volumes, so the number of particles, $N = \rho V/m = O(10^1\text{–}10^2)$, is by no approximation approaching infinity. Of course it is not necessary to take the thermodynamic limit in order to treat thermodynamic or hydrodynamic variables; one simply has to be careful to retain terms that are $O(1/N)$.

From the sample measurements of the mechanical variables one may obtain estimates of hydrodynamic variables, such as fluid velocity and temperature. However it is important to understand that for a hydrodynamic variable, $\mathcal{H}$, the mean is defined in terms of the means of mechanical variables. Specifically,

$$\overline{\mathcal{H}} = \mathcal{H}(\overline{\rho}, \bar{\mathbf{J}}, \overline{K}) \neq \lim_{S\to\infty} \langle \mathcal{H}(\rho, \mathbf{J}, K) \rangle_S, \tag{6}$$

With this in mind, we introduce the notation

$$\langle \mathcal{H} \rangle_S^* = \mathcal{H}(\langle \rho \rangle_S, \langle \mathbf{J} \rangle_S, \langle K \rangle_S) \tag{7}$$

with $\overline{\mathcal{H}} = \langle \mathcal{H} \rangle_\infty^*$. The asterisk reminds us that the estimated mean of a hydrodynamic variable is constructed from the sample means of mechanical variables.

Landau and Lifshitz (§49, [18]) warn of this subtlety in defining quantities such as temperature and pressure: "Strictly speaking, in a system which is not in thermodynamic equilibrium, such as a fluid with velocity and temperature gradients, the usual definitions of thermodynamic quantities are no longer meaningful, and must be ... defined as being the same functions of [mechanical variables] "as they are in thermal equilibrium. [...] The introduction of any further terms (for example, the inclusion in the mass flux density of terms proportional to the gradient of density or temperature) has no physical meaning.... Worse still, the inclusion of such terms may violate the necessary conservation laws." Such a violation is demonstrated in [26] and is discussed here in Section 4.1.

Intensivity (i.e., invariance with volume) is an important property that is lost when hydrodynamic variables are measured incorrectly. Intensive and extensive variables are familiar from equilibrium statistical mechanics, temperature and entropy being examples of each, respectively. The property of intensivity requires that for two volume elements A and B for which $\overline{\mathcal{M}}_\mathrm{A} = \overline{\mathcal{M}}_\mathrm{B}$, we have $\overline{\mathcal{H}}_\mathrm{A+B} = \overline{\mathcal{H}}_\mathrm{A} = \overline{\mathcal{H}}_\mathrm{B}$ if $A+B$ is the union of the two elements.Intensivity is guaranteed when hydrodynamic variables are defined in terms of mechanical densities as $\overline{\mathcal{H}} = \mathcal{H}(\overline{\mathcal{M}})$. On the other hand,

$$\langle \mathcal{H}(\mathcal{M}) \rangle_\infty = \overline{\mathcal{H}} + \tfrac{1}{2}\overline{\delta\mathcal{M}^2}\left(\frac{\partial^2\mathcal{H}}{\partial\mathcal{M}^2}\right)_{\overline{\mathcal{M}}} + \dots \tag{8}$$

where $\delta\mathcal{M} = \mathcal{M} - \overline{\mathcal{M}}$ is the fluctuation of mechanical variables, and $\overline{\delta\mathcal{M}^2}$ is their covariance. Because the covariance is *not* intensive (e.g., $\overline{\delta\rho^2} = m\overline{\rho}/V$ for a dilute gas at equilibrium) one cannot guarantee that $\langle\mathcal{H}(\mathcal{M})\rangle_\infty$ remains intensive (though in some cases, typically at thermodynamic equilibrium, $\langle\mathcal{H}(\mathcal{M})\rangle_\infty = \overline{\mathcal{H}}$). This generic analysis is illustrated in the next two subsections for the specific examples of fluid velocity and temperature.

2.1. ***Fluid Velocity.*** The simplest example of a hydrodynamic variable is fluid velocity, which from the development of the equation of continuity (§1, [18]) is defined as

$$\overline{\mathbf{u}} = \frac{\overline{\mathbf{J}}}{\overline{\rho}} = \lim_{S\to\infty} \frac{\langle \mathbf{J}\rangle_S}{\langle\rho\rangle_S} \tag{9}$$

The unbiased sample mean for the fluid velocity is

$$\langle\mathbf{u}\rangle_S^* = \frac{\langle\mathbf{J}\rangle_S}{\langle\rho\rangle_S} = \frac{S^{-1}\sum_j^S \mathbf{J}_j}{S^{-1}\sum_j^S \rho_j}, \tag{10}$$

so $\overline{\mathbf{u}} = \lim_{S\to\infty}\langle\mathbf{u}\rangle_S^*$.

It is important to note that

$$\langle\mathbf{u}\rangle_S^* \neq \langle\hat{\mathbf{u}}\rangle_S = \frac{1}{S}\sum_j^S \hat{\mathbf{u}}(\rho_j, \mathbf{J}_j, K_j) \tag{11}$$

where $\hat{\mathbf{u}}$ is any general function that defines an instantaneous fluid velocity in terms of the instantaneous mechanical state.

Specifically, note that the instantaneous center-of-mass velocity, $\hat{\mathbf{u}}_j = \mathbf{J}_j/\rho_j$, when averaged over samples, may be written as

$$\langle\hat{\mathbf{u}}\rangle_S = \frac{1}{S}\sum_{j=1}^S \hat{\mathbf{u}}_j = \frac{1}{S}\sum_{j=1}^S \frac{\mathbf{J}_j}{\rho_j} = \left\langle\frac{\mathbf{J}}{\rho}\right\rangle_S, \tag{12}$$

so one might be tempted to define fluid velocity as the center of mass velocity. This definition, though commonly used (eg. §9-4-1, [12]) for fluid velocity, is problematic for two reasons.

First, there is an ambiguity since $\hat{\mathbf{u}}_j$ is not well defined for samples at which $\rho_j = \mathbf{J}_j = 0$, that is, when the instantaneous number of particles N_j is zero. There are twoways to remove this ambiguity: One could take $\hat{\mathbf{u}}_j = 0$ for those samples, an unacceptable approach because it introduces a bias proportional to $1 - S_0/S$ where S_0 is the number of samples for which $\rho_j = 0$ (see equation (61)). The acceptable approach is to define

$$\langle \hat{\mathbf{u}} \rangle_s = \frac{1}{S - S_0} \sum_{j=1}^{S} \frac{\mathbf{J}_j}{\rho_j} (1 - \delta_{0,N_j}) \tag{13}$$

that is, to skip those samples with zero particles, which we shall implicitly assume is how the averaging of samples is performed.

The second and far more serious issue is that using (12) to define fluid velocity is biased when the fluid is not at equilibrium. To see why, recall that

$$\langle \hat{\mathbf{u}} \rangle_s = \left\langle \frac{\mathbf{J}}{\rho} \right\rangle_s \neq \frac{\langle \mathbf{J} \rangle_s}{\langle \rho \rangle_s} = \langle \mathbf{u} \rangle_s^*, \tag{14}$$

The inequality should not be surprising since the instantaneous values of ρ and $\mathbf{J}$ are correlated (e.g., if the instantaneous mass is greater than average then most likely so is the instantaneous momentum). These correlations happen to cancel out at equilibrium (even when $\overline{\mathbf{u}} \neq 0$) but out of equilibrium (e.g., temperature gradient) the measurement of fluid velocity as $\langle \hat{\mathbf{u}} \rangle_s$ is biased and incorrect. This effect is discussed further in Section 4.1.

2.2. ***Temperature.*** Next we consider the measurement of temperature (or more specifically of translational temperature), which is defined from the principle of equipartition of kinetic energy as

$$\overline{T} = \frac{1}{c_v \overline{\rho}} \left(\overline{K} - \frac{|\overline{\mathbf{J}}|^2}{2\overline{\rho}} \right), \tag{15}$$

where $c_v = d\, k_B/2m$ is the heat capacity per unit mass due to the d translational degrees of freedom. From the discussion above, the unbiased sample mean for temperature is

$$\langle T \rangle_s^* = \frac{1}{c_v \langle \rho \rangle_s} \left(\langle K \rangle_s - \frac{|\langle \mathbf{J} \rangle_s|^2}{2 \langle \rho \rangle_s} \right) \tag{16}$$

$$= \frac{1}{c_v} \left(\frac{\langle K \rangle_s}{\langle \rho \rangle_s} - \frac{1}{2} |\langle \mathbf{u} \rangle_s^*|^2 \right), \tag{17}$$

so$\lim_{S \to \infty} \langle T \rangle_s^* = \overline{T}$.

There are several alternative (and incorrect) hydrodynamic definitions for temperaturein common use. The most naive is to define temperature in terms of the instantaneous internal energy per particle:

$$\hat{T}_j = \frac{1}{c_v \rho_j}\left(K_j - \frac{|\mathbf{J}_j|^2}{2\rho_j}\right) \tag{18}$$

$$= \frac{1}{c_v}\left(\frac{K_j}{\rho_j} - \frac{1}{2}|\hat{\mathbf{u}}_j|^2\right). \tag{19}$$

Note that this definition is problematic if $\rho_j = 0$, in the same fashion as already discussed for $\hat{\mathbf{u}}_j$, so the evaluation of the mean value should exclude those samples. A more serious flaw with this definition of temperature is that it is biased, even at equilibrium with $\overline{\mathbf{u}} = 0$, because it fails to account for the fluctuations of the center-of-mass velocity, as shown in Section 4.2. This definition appears in the standard literature of computational statistical mechanics (e.g., §2.4,[2]) and its use is appropriate in the canonical ensemble (fixed N) but not in general.

A simple modification improves the above definition. Arguing that the unbiased estimate of variance must account for the statistical degree of freedom lost in estimating $\hat{\mathbf{u}}_j$, one writes the improved estimate thus:

$$\hat{T}_j = \frac{K_j - \frac{1}{2}\rho_j|\hat{\mathbf{u}}_j|^2}{c_v(\rho_j - m/V)} = \frac{K_j - \frac{1}{2}\rho_j|\hat{\mathbf{u}}_j|^2}{c_v m(N_j - 1)/V}. \tag{20}$$

Note that in this case averages are computed omitting samples where $N_j = 0$ or 1. This construction may be used in equilibrium simulations (e.g., §4.1, [7]) but in Section 4.2 we show that it is biased out of equilibrium.

2.3. ***Other Hydrodynamic Variables.*** In this paper we focus on the hydrodynamic variables of fluid velocity and translational temperature, but there are many others. If the molecules have internal structure, one may separately define temperatures for other degrees of freedom (e.g., rotational, vibrational) [4]. Here we only consider a single species fluid but the more general case would include concentration as a hydrodynamic variable.

The pressure in a fluid is defined by the equation of state, which may be quite complicated in general. A simple case, however, is the ideal gas law $\overline{P} = \overline{\rho}R\overline{T}$, where $R = k_B/m$ is the gas constant and k_B is Boltzmann's constant. Using mechanical variables, the unbiased sample estimate of the mean pressure is then

$$\langle P\rangle_s^* = \frac{R}{c_v}\left(\langle K\rangle_s - \frac{|\langle\mathbf{J}\rangle_s|^2}{2\langle\rho\rangle_s}\right) = \frac{R}{c_v}\left(\langle K\rangle_s - \frac{1}{2}\langle\rho\rangle_s|\langle\mathbf{u}\rangle_s^*|^2\right). \tag{21}$$

The stress tensor and heat flux are also complicated in general, but for an ideal gas they may be expressed in terms of moments of the molecular velocity distribution.

Evaluating means and variances from sample averages of instantaneous hydrodynamic variables isprone to the biases found for fluid velocity and temperature.

Since the analysis for other variables follows the same lines, for brevity we simply reiterate that unbiased estimates are only guaranteed when defining means and variances in terms of mechanical variables.

3. Variances and correlations

To formulate the measurement of variances and correlations, recall that our hydrodynamic variables are defined in terms of mechanical variables as $\overline{\mathcal{H}} = \mathcal{H}(\overline{\rho}, \bar{\mathbf{J}}, \overline{K})$. We define a fluctuation in $\mathcal{H}$ as

$$\begin{aligned}
\delta\mathcal{H} &= \mathcal{H}(\rho, \mathbf{J}, K) - \mathcal{H}(\overline{\rho}, \bar{\mathbf{J}}, \overline{K}) && (22)\\
&= \mathcal{H}(\overline{\rho}+\delta\rho, \bar{\mathbf{J}}+\delta\mathbf{J}, \overline{K}+\delta\mathbf{J}) - \mathcal{H}(\overline{\rho}, \bar{\mathbf{J}}, \overline{K}) && (23)\\
&= \delta\rho\, \frac{\partial\mathcal{H}}{\partial\rho}\bigg|_{\overline{\rho},\bar{\mathbf{J}},\overline{K}} + \delta\mathbf{J}\cdot\frac{\partial\mathcal{H}}{\partial\mathbf{J}}\bigg|_{\overline{\rho},\bar{\mathbf{J}},\overline{K}} + \delta K\, \frac{\partial\mathcal{H}}{\partial K}\bigg|_{\overline{\rho},\bar{\mathbf{J}},\overline{K}} + O(\delta\mathcal{M}^2), && (24)
\end{aligned}$$

Note that

$$\langle\delta\mathcal{H}\rangle^*_s = \mathcal{H}(\langle\rho\rangle_s, \langle\mathbf{J}\rangle_s, \langle K\rangle_s) - \mathcal{H}(\overline{\rho}, \bar{\mathbf{J}}, \overline{K}), \tag{25}$$

so $\lim_{S\to\infty}\langle\delta\mathcal{H}\rangle^*_s = \overline{\delta\mathcal{H}} = 0$. In general, the exact means are unknown so for estimating $\delta\mathcal{H}$ we implicitly take $\overline{\mathcal{M}} = \langle\mathcal{M}\rangle_s$ and also drop the higher order terms. This construction allows us to formulate the variance of hydrodynamic variables in terms of the variances of mechanical variables, which may be estimated from samples. The remainder of this section presents expressions for variances and correlations involving fluid velocity and temperature.

3.1. ***Fluid Velocity Fluctuations.*** First consider fluid velocity, whose fluctuations are expressed in terms of fluctuations of mechanical variables as

$$\delta\mathbf{u} = \frac{\delta\mathbf{J}}{\overline{\rho}} - \frac{\bar{\mathbf{J}}}{\overline{\rho}^2}\,\delta\rho = \frac{1}{\overline{\rho}}\left(\delta\mathbf{J} - \bar{\mathbf{u}}\,\delta\rho\right), \tag{26}$$

or for the x-component,

$$\delta u_x = \frac{1}{\overline{\rho}}\left(\delta J_x - \bar{u}_x\,\delta\rho\right). \tag{27}$$

The correlation of mass density fluctuations in cell C and fluid velocity fluctuations in cell C' is

$$\langle\delta\rho\,\delta u'_x\rangle^*_s = \frac{1}{\overline{\rho}'}\left(\langle\delta\rho\,\delta J'_x\rangle_s - \bar{u}'_x\,\langle\delta\rho\,\delta\rho'\rangle_s\right) \tag{28}$$

with similar expressions for the other components.

The sample estimated variance of the x-component of fluid velocity is

$$\langle \delta u_x^2 \rangle_s^* = \frac{1}{\overline{\rho}^2} \langle (\delta J_x - \bar{u}_x \, \delta\rho)^2 \rangle_s \tag{29}$$

$$= \frac{1}{\overline{\rho}^2} \left(\langle \delta J_x^2 \rangle_s - 2\bar{u}_x \langle \delta\rho \, \delta J_x \rangle_s + \bar{u}_x^2 \langle \delta\rho^2 \rangle_s \right). \tag{30}$$

If the system is isotropic (i.e., $\bar{\mathbf{u}} = 0$), then $\overline{|\delta\mathbf{u}|^2} = d\,\overline{\delta u_x^2} = d\,\overline{\delta J_x^2}/\overline{\rho}^2$, where d is the dimensionality. The correlations of velocity components are similarly obtained, for example,

$$\langle \delta u_x \, \delta u_y' \rangle_s^* = \frac{1}{\overline{\rho}\,\overline{\rho}'} \left(\langle \delta J_x \, \delta J_y' \rangle_s - \bar{u}_x \, \langle \delta\rho \, \delta J_y' \rangle_s - \bar{u}_y' \, \langle \delta\rho' \, \delta J_x \rangle_s + \bar{u}_x \bar{u}_y' \, \langle \delta\rho \, \delta\rho' \rangle_s \right), \tag{31}$$

with similar results for the other components.

3.2. ***Temperature Fluctuations.*** In terms of mechanical variables, the fluctuation of temperature may be written as

$$\delta T = \frac{1}{c_v \overline{\rho}} \left\{ \delta K - \bar{\mathbf{u}} \cdot \delta\mathbf{J} - \left(c_v \overline{T} - \tfrac{1}{2} |\bar{\mathbf{u}}|^2 \right) \delta\rho \right\}$$

$$= \frac{1}{c_v \overline{\rho}} \left\{ \delta K - \delta G - \overline{Q} \, \delta\rho \right\}, \tag{32}$$

where $\delta G \equiv \bar{\mathbf{u}} \cdot \delta\mathbf{J}$ and $\overline{Q} \equiv c_v \overline{T} - \frac{1}{2} |\bar{\mathbf{u}}|^2$. From this, the estimated sample correlation of temperature fluctuations is

$$\begin{aligned} \langle \delta T \delta T' \rangle_s^* = \frac{1}{c_v^2 \overline{\rho}\,\overline{\rho}'} \big\{ &\langle \delta K \delta K' \rangle_s + \langle \delta G \delta G' \rangle_s + \overline{Q}\,\overline{Q}' \langle \delta\rho \, \delta\rho' \rangle_s \\ &- \langle \delta K \delta G' \rangle_s - \langle \delta G \delta K' \rangle_s - \overline{Q}' \langle \delta K \delta\rho' \rangle_s - \overline{Q} \langle \delta\rho \, \delta K' \rangle_s \\ &+ \overline{Q}' \langle \delta G \, \delta\rho' \rangle_s + \overline{Q} \langle \delta\rho \, \delta G' \rangle_s \big\}. \end{aligned} \tag{33}$$

The covariance of density and temperature fluctuations is

$$\langle \delta\rho \, \delta T' \rangle_s^* = \frac{1}{c_v \overline{\rho}'} \left\{ \langle \delta\rho \, \delta K' \rangle_s - \langle \delta\rho \, \delta G' \rangle_s - \overline{Q}' \langle \delta\rho \, \delta\rho' \rangle_s \right\}. \tag{34}$$

The covariance of fluid velocity and temperature is

$$\begin{aligned} \langle \delta u_x \, \delta T' \rangle_s^* = \frac{1}{c_v \overline{\rho}\,\overline{\rho}'} \Big\{ &\langle \delta J_x \, \delta K' \rangle_s - u_x \langle \delta\rho \, \delta K' \rangle_s \\ &- \langle \delta J_x \, \delta G' \rangle_s + u_x \langle \delta\rho \, \delta G' \rangle_s - \overline{Q}' \langle \delta J_x \, \delta\rho' \rangle_s + u_x \overline{Q}' \langle \delta\rho \, \delta\rho' \rangle_s \Big\}. \end{aligned} \tag{35}$$

4. Biases due to fluctuations

We now consider the possible bias in the statistical measurements of hydrodynamic variables due to fluctuations. To derive and illustrate these results we consider four separate approaches, two for equilibrium and two for nonequilibrium systems. The first is the direct evaluation of statistical means at thermodynamic equilibrium; this methodology is straightforward and details of the calculations are collected in Appendix A. Results for the variances and correlations are compared with fluctuating hydrodynamic theory, which is summarized in Appendix B. The second approach is similar to the first but uses stochastic numerical simulations to generate random samples (see Appendix C). These numerical results illustrate the predicted phenomena and verify the accuracy of various approximate results.

For nonequilibrium systems, various definitions for mean values of fluid velocity and temperature are compared to quadratic order in fluctuations, indicating how a bias may be introduced by nonequilibrium correlations. The predicted bias is confirmed by the fourth approach—molecular simulations of a dilute gas in a closed system with a temperature gradient (see Appendix D). Note that the four approaches are intertwined in the presentation below.

4.1. ***Bias for Fluid Velocity.*** First we consider two ways to estimate the mean value of fluid velocity, $\langle\mathbf{u}\rangle_s^*$ and $\langle\hat{\mathbf{u}}\rangle_s$, as introduced in Section 2.1. By direct evaluation (see (55), (56) and (59)) we find that both definitions are unbiased at equilibrium (even if $\overline{\mathbf{u}} \neq 0$), a result confirmed by numerical simulation. However, $\langle\mathbf{u}\rangle_s^*$ and $\langle\hat{\mathbf{u}}\rangle_s$ are not equivalent out of equilibrium. To see why, note that the sample mean of the center-of-mass velocity from equation (12) may be written as

$$\begin{aligned}
\langle\hat{\mathbf{u}}\rangle_s &= \left\langle\frac{\mathbf{J}}{\rho}\right\rangle_s = \left\langle\frac{\bar{\mathbf{J}}+\delta\mathbf{J}}{\bar{\rho}+\delta\rho}\right\rangle_s && (36)\\
&= \frac{\bar{\mathbf{J}}}{\bar{\rho}}\left\langle\left(1+\frac{\delta\mathbf{J}}{\bar{\mathbf{J}}}\right)\left(1-\frac{\delta\rho}{\bar{\rho}}+\frac{\delta\rho^2}{\bar{\rho}^2}\right)\right\rangle_s + O(\delta\mathcal{M}^3) && (37)\\
&= \overline{\mathbf{u}}\left(1+\frac{\langle\delta\rho^2\rangle_s}{\bar{\rho}^2}\right) - \frac{\langle\delta\rho\,\delta\mathbf{J}\rangle_s}{\bar{\rho}^2} + O(\delta\mathcal{M}^3). && (38)
\end{aligned}$$

From (26), $\delta\mathbf{J} = \bar{\rho}\delta\mathbf{u} + \overline{\mathbf{u}}\delta\rho$, so in the limit where the number of samples $S \to \infty$,

$$\langle\hat{\mathbf{u}}\rangle_\infty = \overline{\mathbf{u}} - \frac{\overline{\delta\rho\,\delta\mathbf{u}}}{\bar{\rho}} + O(\delta\mathcal{M}^3). \qquad (39)$$

The correlation $\overline{\delta\rho\,\delta\mathbf{u}}$ is zero at equilibrium (see Appendix B) but, in general, nonzero for nonequilibrium systems [20]. The correlation $\overline{\delta\rho\delta\mathbf{u}} \propto \nabla T$ and the fact that $\langle\hat{\mathbf{u}}\rangle_\infty{\neq}0$ in a closed system indicates a violation of mass conservation, as cautioned by Landau and Lifshitz (see Section 2 above). Finally, since $\overline{\delta\rho\,\delta\mathbf{u}} \propto V^{-1}$, the quantity $\langle\hat{\mathbf{u}}\rangle_\infty$ is not an intensive variable.

This bias of the center-of-mass fluid velocity is studied at length in [26] where it is shown that the nonequilibrium correlation $\overline{\delta\rho\,\delta\mathbf{u}}$ leads to an anomalous flow, as measured by $\langle\hat{\mathbf{u}}\rangle_s$, in closed systems.[1] For the simulation parameters listed in Appendix D the anomalous flow velocity is about $10^{-4}c$ for the large system and $10^{-3}c$ for the small system, where c is the sound speed.

At equilibrium, the variance of fluid velocity is (see Appendix B),

$$\overline{|\delta\mathbf{u}|^2} = d\,\frac{k_B\overline{T}}{\overline{\rho}V} = d\,\frac{C_T^2}{\overline{N}}. \tag{40}$$

By direct evaluation, the definition based on the variances of mechanical variables is found to be unbiased, that is $\langle|\delta\mathbf{u}|^2\rangle_\infty^* = \overline{|\delta\mathbf{u}|^2}$, (see equation (62)) whereas the center-of-mass definition gives (see equation (66)),

$$\langle|\delta\hat{\mathbf{u}}|^2\rangle_\infty \approx \overline{|\delta\mathbf{u}|^2}\left(1 + \frac{\overline{\delta N^2}}{\overline{N}^2}\right). \tag{41}$$

Figure 1 shows the fractional errors in the sample estimate for the variance of fluid velocity, that is

$$\frac{\langle|\delta\mathbf{u}|^2\rangle_s^* - \overline{|\delta\mathbf{u}|^2}}{\overline{|\delta\mathbf{u}|^2}} \qquad\text{and}\qquad \frac{\langle|\delta\hat{\mathbf{u}}|^2\rangle_s - \overline{|\delta\mathbf{u}|^2}}{\overline{|\delta\mathbf{u}|^2}}.$$

In the simulations N is Poisson-distributed, so $\overline{\delta N^2} = \overline{N}$; thus the error goes roughly as $1/\overline{N}$. Note that this fractional error is significant (e.g., about 5% for $\overline{N} = 20$).

4.2. ***Bias for Temperature.*** Section 2.2 introduced three definitions for the sample mean temperature, specifically the definition in terms of mean values of mechanical variables, $\langle T\rangle_s^*$ (equation (16)), and two definitions based on instantaneous temperature. The latter may be combined and written as

$$\hat{T}_\alpha = \frac{K - \frac{1}{2}\rho|\hat{\mathbf{u}}|^2}{c_V(\rho - \alpha m/V)}, \tag{42}$$

where $\alpha = 0$ for equation (18) and $\alpha = 1$ for equation (20).

By direct evaluation (see (73), (76)), we find that $\langle T\rangle_\infty^* = \langle\hat{T}_1\rangle_\infty = \overline{T}$ at equilibrium, while

$$\langle\hat{T}_0\rangle_\infty \approx \left(1 - \frac{1}{\overline{N}}\right)\overline{T}. \tag{43}$$

[1] In [26] the quantity $\langle\mathbf{u}\rangle_s^*$ is referred to as the Cumulative-Averaged-Measurement (CAM) of fluid velocity and $\langle\hat{\mathbf{u}}\rangle_s$ is called the Sample-Averaged-Measurement (SAM) of velocity.

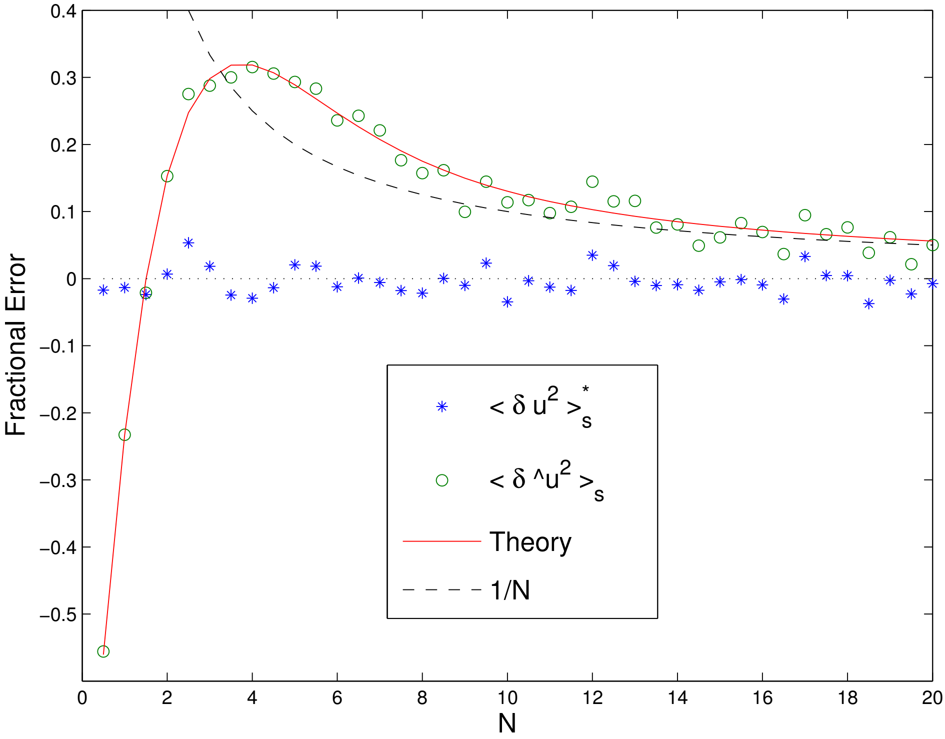


Figure 1. Fractional error in the sample variance of fluid velocity versus $\bar{N}$ for: $\langle|\delta\mathbf{u}|^2\rangle_s^*$ (asterisks); $\langle|\delta\hat{\mathbf{u}}|^2\rangle_s$ (circles). Solid line given by equation (63); dashed line is $1/\bar{N}$ (dashed line).

Figure 2 confirms these results, showing the fractional error in the sample mean of temperature (relative to $\bar{T}$) versus the mean number $\bar{N}$ of particles. Note that the fractional error for $\langle\hat{T}_0\rangle_\infty$ is significant (e.g., about 5% for $\bar{N}=20$).

For a more general result, applicable to nonequilibrium cases, we write the sample mean of instantaneous temperature as

$$\begin{aligned}\langle\hat{T}_\alpha\rangle_s &= \frac{1}{c_V}\left\langle\frac{K-\frac{1}{2}\rho|\hat{\mathbf{u}}|^2}{\rho-\alpha m/V}\right\rangle_s \qquad (44)\\ &= \left(1+\frac{\alpha m}{\bar{\rho}V}\right)\bar{T}-\frac{1}{\bar{\rho}c_V}\left\langle\frac{\delta\rho}{\bar{\rho}}\left(\delta K-\frac{1}{2}\,\delta\rho|\bar{\mathbf{u}}|^2-\rho\bar{\mathbf{u}}\cdot\delta\mathbf{u}\right)\right\rangle_s+O(\delta\mathcal{M}^3).\end{aligned}$$

Using the results from Section 3, after some algebra, we find

$$\langle\hat{T}_\alpha\rangle_s=\left[1+\frac{\alpha}{\bar{N}}-\frac{\langle\delta\rho^2\rangle_s}{\bar{\rho}^2}\right]\bar{T}-\frac{\langle\delta\rho\,\delta T\rangle_s^*}{\bar{\rho}}+O(\delta\mathcal{M}^3). \qquad (45)$$

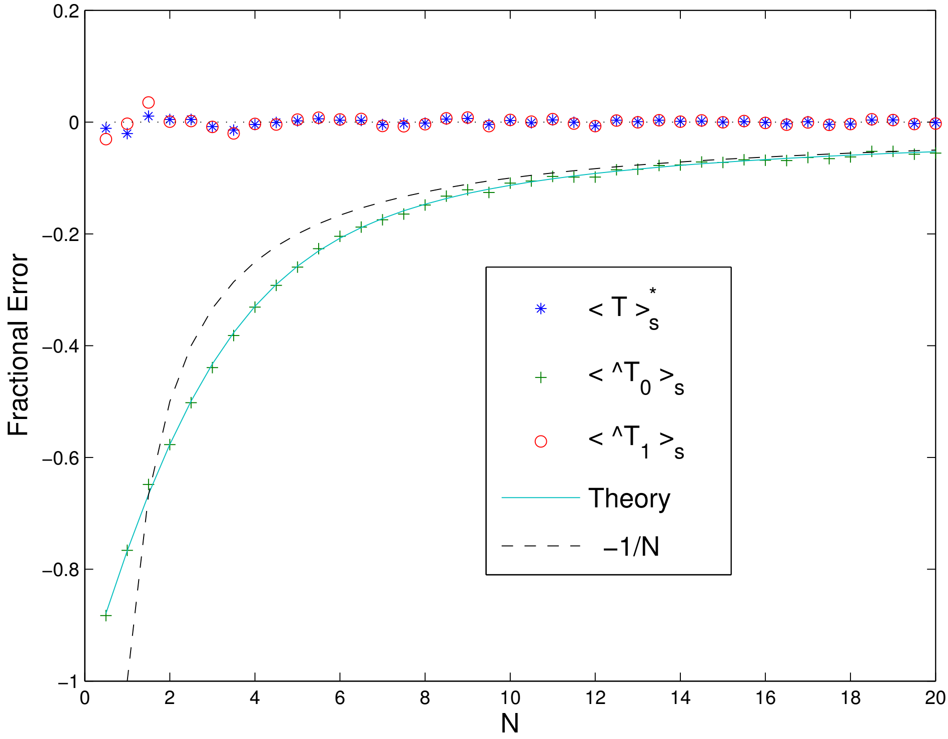


Figure 2. Fractional error in the sample mean of temperature versus $\overline{N}$ for: $\langle T\rangle_s^*$ (asterisks), $\langle\hat{T}_0\rangle_s$ (crosses) and $\langle\hat{T}_1\rangle_s$ (circles). Solid line given by equation (73); dashed line is $-1/\overline{N}$.

At equilibrium $\overline{\delta\rho\delta T} = 0$ so by comparison with the results from direct evaluation we have

$$\langle\hat{T}_1\rangle_\infty = \overline{T} - \frac{\overline{\delta\rho\,\delta T}}{\overline{\rho}} + O(\delta\mathcal{M}^3). \tag{46}$$

This result is verified by molecular simulations of a nonequilibrium system at a steady state, specifically a dilute gas between a pair of thermal walls at different temperatures (see Appendix D). The predicted bias from (46) is in good agreement with the bias measured in both the large (132 particles per sample cell) and small (8.2 particles per sample cell) systems. In the latter case the absolute temperature bias is a few Kelvin (about 1% of the mean), while in the large system the bias is smaller by a factor of $132/8.2 \approx 16$, since $\overline{\delta\rho\,\delta T} \propto V^{-1}$. This result confirms the warning given in Section 2 that the means of instantaneous hydrodynamic variables are not intensive quantities.

Finally, we consider the measurement of temperature fluctuations, choosing among the many possible examples the correlation of density and temperature

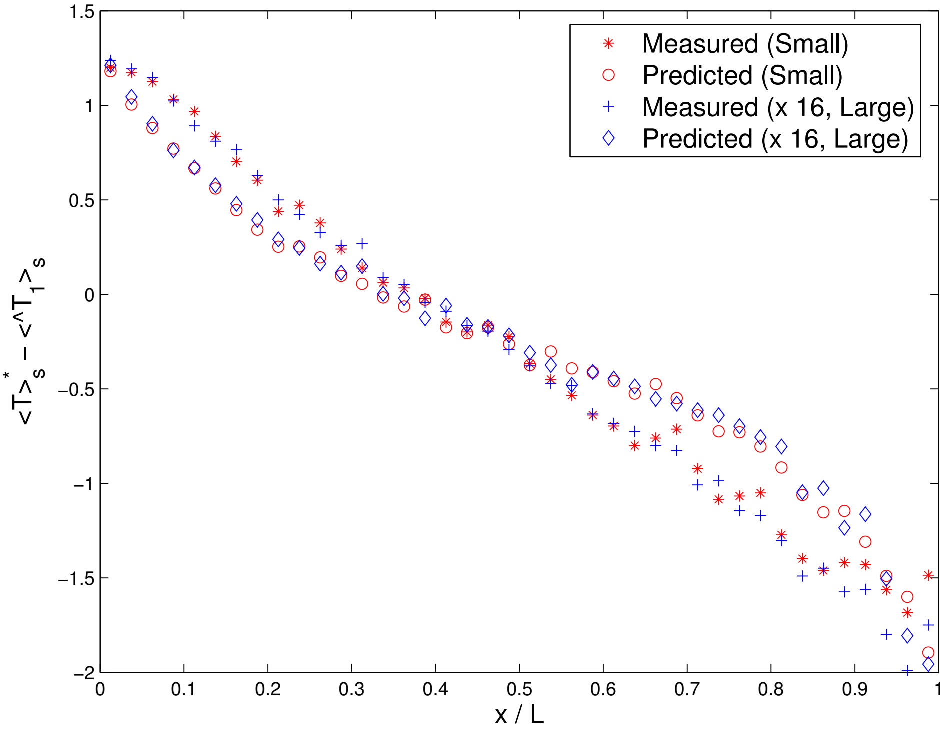


Figure 3. Measured temperature difference $\langle T\rangle_s^* - \langle \hat{T}_1\rangle_s$ [small (asterisks), large (crosses) systems] and theory prediction, $\langle\delta\rho\,\delta T\rangle_s^*/\langle\rho\rangle_s$ [small (circles), large (diamonds) systems] versus position. Results for the large system are scaled by a multiplicative factor of 16. Wall temperatures are 273 and 809 Kelvin; see Appendix D for other parameters.

fluctuations. As mentioned above, at equilibrium $\overline{\delta\rho\,\delta T} = 0$; by direct evaluation we get$\langle\delta\rho\,\delta T\rangle_\infty^* = 0$ (see Appendix A), while for the two definitions of instantaneous temperature we find (see eqns. (78) and (79)),

$$\langle\delta\rho\,\delta\hat{T}_0\rangle_\infty = \overline{\rho}\overline{T}\sum_{N=1}^{\infty}\left(\frac{N-1}{\overline{N}} - \frac{N-1}{N}\right)\frac{P(N)}{1-P(0)} \tag{47}$$

$$\approx \overline{\rho}\overline{T}\,\frac{\overline{\delta N^2}}{\overline{N}^3} \tag{48}$$

and

$$\langle\delta\rho\,\delta\hat{T}_1\rangle_\infty = \frac{\overline{\rho}\overline{T}}{\overline{N}}\left(\frac{\overline{N}P(0) + (\overline{N}-1)P(1)}{1-P(0)-P(1)}\right), \tag{49}$$

where $P(N)$ is the probability distribution for N. When this is the Poisson distribution, then

$$\langle \delta\rho\, \delta\hat{T}_0 \rangle_\infty \approx \frac{\overline{\rho}\overline{T}}{\overline{N}^2} \tag{50}$$

and

$$\langle \delta\rho\, \delta\hat{T}_1 \rangle_\infty = \overline{\rho}\overline{T}\; \overline{N} e^{-\overline{N}}. \tag{51}$$

These results are illustrated and verified in Figure 4 where the scaled error (relative to $(\overline{\delta\rho^2}\,\overline{\delta T^2})^{1/2}$) in the correlation of density and temperature versus $\overline{N}$ is presented for equilibrium simulation measurements (see Appendix C). The bias for $\langle \delta\rho\, \delta\hat{T}_0 \rangle_\infty$ is significant (scaled error of about 7% for $\overline{N} = 20$) while the bias for $\langle \delta\rho\, \delta\hat{T}_1 \rangle_\infty$ decreases quickly with $\overline{N}$ (scaled error is less than 1% for $\overline{N} = 10$). On the other hand, the bias in the variance $\langle \delta\hat{T}_1^2 \rangle_\infty$ turns out to be significant (e.g., over 10% for $\overline{N} = 20$).

5. Summary and concluding remarks

In this paper we demonstrate that in the presence of spontaneous fluctuations the statistical measurement of hydrodynamic quantities, such as fluid velocity and translational temperature, should be done by sampling mechanical variables, such as momentum and kinetic energy densities. The correct constructions for means and variances are given in sections 2 and 3, respectively. In those sections we caution that using definitions based on instantaneous fluid velocity and instantaneous temperature leads to biased statistical results (as shown in Section 4).

Molecular simulations have been used in the study of fluids for nearly half a century, so why are the results presented in this paper not well known? First, one should recall that most molecular dynamics simulations are of equilibrium systems for the purpose of computing thermodynamic properties, such as the equation of state. The computation of means and fluctuations of thermodynamic quantities in the various ensembles of statistical mechanics is certainly well known [2; 7].

Molecular dynamics simulations of hydrodynamic phenomena are more recent (e.g. [17]) and often focus on qualitative features (e.g., appearance of vortex shedding).[2] Other molecular algorithms, such as direct simulation Monte Carlo [4] and lattice gases [24], have always been applied to nonequilibrium flows, yet, as with molecular dynamics, the biases due to fluctuations were not identified. Errors due to these biases were either dismissed as small numerical artifacts (e.g., finite time step effects) or masked by other errors (e.g., large statistical uncertainties). Since the bias in the mean values is usually quite small (about 0.1 Kelvin for the large system in Figure 3) either possibility is plausible.

[2] Evan's nonequilibrium molecular dynamics (NEMD) approach is not designed for hydrodynamic flows but rather is a method for obtaining transport properties, such as viscosity [6].

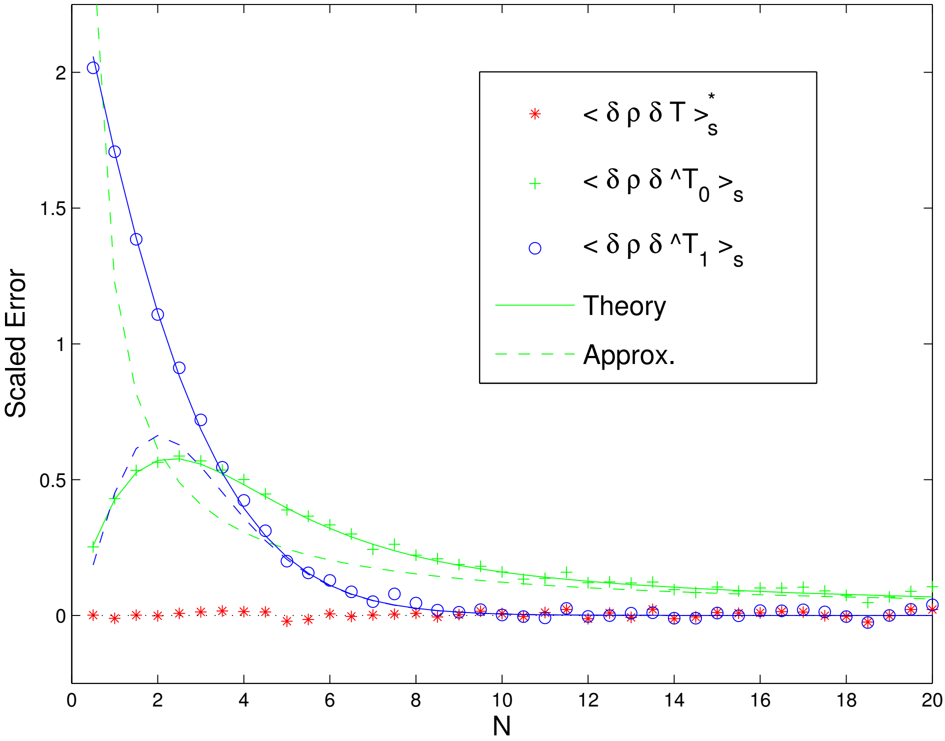


Figure 4. Scaled error in the correlation of density and temperature (relative to $(\overline{\delta\rho^2}\,\overline{\delta T^2})^{1/2}$) versus $\bar{N}$ for: $\langle\delta\rho\,\delta T\rangle_s^*$ (asterisks), $\langle\delta\rho\,\delta\hat{T}_0\rangle_s$, (crosses) and $\langle\delta\rho\,\delta\hat{T}_1\rangle_s$ (circles). Solid lines are theoretical predictions (47) and (49); dashed lines are approximations (50) and (51).

Another possibility is that, in some cases, no errors were made in measuring hydrodynamic quantities because the sampling happened to be equivalent to the unbiased formulation using mechanical variables (e.g., programs in [4]). Unfortunately, one rarely finds a detailed description in the literature of how statistical measurements are performed, especially for fluid velocity.

In molecular simulations of hydrodynamic flows, variances are usually measured only for the purpose of estimating error bars [11]. As such, the effects described in this paper are unlikely to have been noticed by many computational scientists. On the other hand, my own research is in the field of nonequilibrium fluctuations, which is how these effects came to my attention. The recent computational studies of nano-scale and multi-scale flows, as well as of Brownian motors, may also profit from this paper's analysis regarding the measurement of microscopic fluctuations in molecular simulations. The importance of these fluctuations is appreciated by noting that a typical molecular motor protein consumes ATP at a power of roughly 10^{-16} watts while operating in a background of 10^{-8} watts of thermal noise power, which has been said tobe “as difficult as walking in a hurricane is for us.” [3]

Finally, we have focused on the effect of fluctuations in particle-based simulations, yet these effects have a physical rather than numerical origin so the discussion also applies to continuum methods for stochastic partial differential equations. The deterministic hydrodynamic equations can be augmented by the inclusion of stochastic fluxes due to thermal fluctuations. These fluctuating hydrodynamic equations [18] accurately capture equilibrium and nonequilibrium effects and can be computed numerically (see [9] for a simple, finite-difference scheme). Any numerical computation of hydrodynamic phenomena that includes spontaneous fluctuations may be susceptible to the effects presented in this paper. *Caveat ratiocinator*.

Appendix A: Direct evaluation at equilibrium

In this appendix we obtain, by direct evaluation, mean values and variances of mechanical and hydrodynamic variables at thermodynamic equilibrium. To perform this analysis, we first need to say something about the probability distributions for the fluid particles, specifically, $P(\mathbf{v})$, the probability that a particle has velocity $\mathbf{v}$ and $P(N)$, the probability that a cell has N particles.

From the principle of equipartition, at thermodynamic equilibrium the velocities of classical particles are Gaussian-distributed with mean $\overline{\mathbf{v}} = \overline{\mathbf{u}}$ and variance $\overline{|\mathbf{v} - \overline{\mathbf{v}}|^2} = \overline{|\delta \mathbf{v}|^2} = d\,C_T^2 = d\,k_B \overline{T}/m$ where C_T is the thermal speed. Note that thermodynamic equilibrium does *not* imply $\overline{\mathbf{u}} = 0$ since a system is in equilibrium in all inertial frames of reference.

The distribution for N depends on the equation of state for the fluid. For the present analysis we only require the mean $\overline{N} = \overline{N}$ and variance $\overline{\delta N^2} = \sigma_N^2$. In dense fluids σ_N^2 is small since it is proportional to the fluids’ compressibility; in the case of a dilute gas, N is Poisson-distributed with $\sigma_N^2 = \overline{N}$.

For some definitions of instantaneous variables (e.g., eqns. (12) and (18)) we need to exclude the state $N = 0$, in which case we use the distribution

$$P_0(N) = \frac{1}{1 - P(0)}\, P(N) \tag{52}$$

for $N = 1, \ldots, \infty$. For the alternative temperature definition, equation (20), we need to exclude the states $N = 0$ or 1, in which case we use the distribution

$$P_{01}(N) = \frac{1}{1 - P(0) - P(1)}\, P(N) \tag{53}$$

for $N = 2, \ldots, \infty$.

Mean values may be obtained by direct evaluation,

$$\langle X\rangle_\infty = \sum_{N=0}^{\infty}\int d\mathbf{v}_1\ldots\int d\mathbf{v}_N X(N,\mathbf{v}_1,\ldots,\mathbf{v}_N)P(N)P(\mathbf{v}_1)\ldots P(\mathbf{v}_N), \tag{54}$$

with the minor modification that the sum starts at $N = 1$ or $N = 2$ if P_0 or P_{01} is used in place of $P(N)$. For the mechanical variables, we easily find

$$\begin{aligned}\langle\rho\rangle_\infty &= \frac{1}{V}\sum_{N=0}^{\infty}\int d\mathbf{v}_1\ldots\int d\mathbf{v}_N\left(\sum_{k=1}^{N} m\right)P(N)P(\mathbf{v}_1)\ldots P(\mathbf{v}_N)\\ &= \frac{1}{V}\sum_{N=0}^{\infty}\left(Nm\right)P(N) = \frac{m\overline{N}}{V} = \overline{\rho}.\end{aligned} \tag{55}$$

Similarly,

$$\begin{aligned}\langle\mathbf{J}\rangle_\infty &= \frac{1}{V}\sum_{N=0}^{\infty}\int d\mathbf{v}_1\ldots\int d\mathbf{v}_N\left(\sum_{k=1}^{N} m\mathbf{v}_k\right)P(N)P(\mathbf{v}_1)\ldots P(\mathbf{v}_N)\\ &= \frac{1}{V}\sum_{N=0}^{\infty}\left(Nm\overline{\mathbf{v}}\right)P(N) = \frac{m\overline{N}}{V}\overline{\mathbf{v}} = \overline{\rho}\,\overline{\mathbf{u}}\end{aligned} \tag{56}$$

and

$$\begin{aligned}\langle K\rangle_\infty &= \frac{1}{V}\sum_{N=0}^{\infty}\int d\mathbf{v}_1\ldots\int d\mathbf{v}_N\left(\sum_{k=1}^{N} \tfrac{1}{2}m|\mathbf{v}_k|^2\right)P(N)P(\mathbf{v}_1)\ldots P(\mathbf{v}_N)\\ &= \frac{m\overline{N}}{V}\;\tfrac{1}{2}\,\overline{|\mathbf{v}|^2} = \overline{\rho}(c_v\overline{T} + \frac{1}{2}|\overline{\mathbf{u}}|^2),\end{aligned} \tag{57}$$

confirming the expected result that $\langle\mathcal{M}\rangle_\infty = \overline{\mathcal{M}}$.

The variances and covariances of the mechanical variables may be evaluated directly. For example,

$$\begin{aligned}\langle\delta\rho^2\rangle_\infty &= \frac{1}{V^2}\sum_{N=0}^{\infty}\int d\mathbf{v}_1\ldots\int d\mathbf{v}_N\left[\left(\sum_{k=1}^{N} m\right)-\overline{\rho}\right]^2 P(N)P(\mathbf{v}_1)\ldots P(\mathbf{v}_N)\\ &= \frac{1}{V^2}\sum_{N=0}^{\infty}\left[Nm-\overline{N}m\right]^2 P(N) = \frac{m^2}{V^2}\overline{\delta N^2} = \overline{\rho}^2\frac{\sigma_N^2}{\overline{N}^2}.\end{aligned} \tag{58}$$

The procedure is straightforward (though tedious) for the other variables; the results are the same as in eqns. (83)–(88) in Appendix B.

Fluid Velocity. From the results above, the mean fluid velocity

$$\langle \mathbf{u} \rangle_\infty^* = \langle \mathbf{J} \rangle_\infty / \langle \rho \rangle_\infty = \bar{\mathbf{J}} / \overline{\rho} = \overline{\mathbf{u}}.$$

At equilibrium we find for the center-of-mass velocity,

$$\langle \hat{\mathbf{u}} \rangle_\infty = \sum_{N=1}^{\infty} \int d\mathbf{v}_1 \ldots \int d\mathbf{v}_N \left(\frac{\mathbf{v}_1 + \ldots \mathbf{v}_N}{N} \right) P_0(N) P(\mathbf{v}_1) \ldots P(\mathbf{v}_N) \quad (59)$$

$$= \sum_{N=1}^{\infty} \frac{N\overline{\mathbf{v}}}{N} P_0(N) = \overline{\mathbf{u}}, \quad (60)$$

where the $N = 0$ case is excluded. An alternative approach would be to take $\hat{\mathbf{u}}_j = 0$ when $N_j = 0$ which gives

$$\begin{aligned}\langle \hat{\mathbf{u}}' \rangle_\infty &= \sum_{N=0}^{\infty} \frac{N\overline{\mathbf{u}}}{N} (1 - \delta_{N,0}) P(N) = \overline{\mathbf{u}} \left(\sum_{N=0}^{\infty} P(N) \right) - \overline{\mathbf{u}} P(0) \\ &= (1 - P(0)) \overline{\mathbf{u}}, \end{aligned} \quad (61)$$

so at equilibrium this definition for the mean of the center-of-mass velocity does not equal the fluid velocity except when $\overline{\mathbf{u}} = 0$.

From (29), the variance of fluid velocity as obtained from mechanical variables is

$$\langle |\delta \mathbf{u}|^2 \rangle_\infty^* = \frac{1}{\overline{\rho}^2} (\overline{|\delta \mathbf{J}|^2} - 2\overline{\mathbf{u}} \cdot \overline{\delta\rho \mathbf{J}} + |\overline{\mathbf{u}}|^2 \overline{\delta\rho^2}). \quad (62)$$

Using (80), (83), and (86), we find $\langle |\delta \mathbf{u}|^2 \rangle_\infty^* = d C_T^2 / \overline{N} = \overline{|\delta \mathbf{u}|^2}$. By direct evaluation, the variance of the center-of-mass velocity is

$$\begin{aligned}\langle |\delta \hat{\mathbf{u}}|^2 \rangle_\infty &= \sum_{N=1}^{\infty} \int d\mathbf{v}_1 \ldots \int d\mathbf{v}_N \left| \frac{\mathbf{v}_1 + \ldots + \mathbf{v}_N}{N} - \overline{\mathbf{u}} \right|^2 P_0(N) P(\mathbf{v}_1) \ldots P(\mathbf{v}_N) \\ &= \sum_{N=1}^{\infty} \frac{N \overline{|\delta \mathbf{v}|^2}}{N^2} P_0(N) = d\, C_T^2 \sum_{N=1}^{\infty} \frac{1}{N} P_0(N). \end{aligned} \quad (63)$$

By Jensen's inequality

$$\sum_{N=1}^{\infty} \frac{1}{N} P_0(N) > \sum_{N=1}^{\infty} \frac{1}{N} P(N) \geq \left(\sum_{N=1}^{\infty} N P(N) \right)^{-1} = \frac{1}{\overline{N}}, \quad (64)$$

with equality only if $P_0(N) = \delta_{N,\overline{N}}$. Excluding this trivial case, $\langle |\delta \hat{\mathbf{u}}|^2 \rangle_\infty > \overline{|\delta \mathbf{u}|^2}$. Since

$$\overline{N^{-1}} = \frac{1}{\overline{N}} \left(1 + \frac{\sigma_N^2}{\overline{N}^2} + O(\delta N^3) \right), \quad (65)$$

we have

$$\langle|\delta\hat{\mathbf{u}}|^2\rangle_\infty = \overline{|\delta\mathbf{u}|^2}\left(1+\frac{\sigma_N^2}{\overline{N}^2}+O(\delta N^3)\right). \tag{66}$$

If N is Poisson-distributed, then

$$\langle|\delta\hat{\mathbf{u}}|^2\rangle_\infty = \overline{|\delta\mathbf{u}|^2}\left(1+\frac{1}{\overline{N}}+O(\delta N^3)\right) \tag{67}$$

Finally, note that we may write

$$\langle|\delta\hat{\mathbf{u}}|^2\rangle_\infty = \sum_{N=1}^{\infty}\langle|\delta\hat{\mathbf{u}}|_N^2\rangle_\infty P_0(N), \tag{68}$$

where

$$\langle|\delta\hat{\mathbf{u}}|_N^2\rangle_\infty = \frac{d\,C_T^2}{N} \tag{69}$$

is the variance of the center-of-mass velocity for a given value of N, a result used below.

Temperature. From (16), (55), (56), and (57) we find $\langle T\rangle_\infty^* = \overline{T}$. Turning to the two definitions of instantaneous temperature, equation (18) and (20), note that they may be combined as

$$\hat{T}_{\alpha;j} = \frac{1}{2c_V(N_j-\alpha)}\sum_k^{N_j}|\mathbf{v}_{k,j}-\hat{\mathbf{u}}_j|^2 \tag{70}$$

where $\alpha = 0$ or 1 and

$$\hat{\mathbf{u}}_j = \frac{1}{N_j}\sum_k^{N_j}\mathbf{v}_{k,j} \tag{71}$$

is the instantaneous center-of-mass velocity. First, consider the case $\alpha = 0$, by direct evaluation the mean value is,

$$\begin{aligned}\langle\hat{T}_0\rangle_\infty &= \frac{1}{2c_V}\sum_{N=1}^{\infty}\int d\mathbf{v}_1\ldots\int d\mathbf{v}_N\left(\frac{1}{N}\sum_{k=1}^{N}|\mathbf{v}_k-\hat{\mathbf{u}}|^2\right)P_0(N)P(\mathbf{v}_1)\ldots P(\mathbf{v}_N)\\ &= \frac{1}{2c_V}\sum_{N=1}^{\infty}\left(\overline{|\mathbf{v}|^2}-\overline{|\hat{\mathbf{u}}|^2}\right)P_0(N)\end{aligned} \tag{72}$$

In general $\langle \hat{T}_0 \rangle_\infty < \overline{T}$ since $|\delta \hat{\mathbf{u}}|^2 \to 0$ only in the limit $\overline{N} \to \infty$. From the above result for the variance of the center-of-mass velocity,

$$\langle \hat{T}_0 \rangle_\infty = \overline{T}\left(1 - \sum_{N=1}^{\infty} \frac{1}{N} P_0(N)\right) \tag{73}$$

$$\approx \left(1 - \frac{1}{\overline{N}} - \frac{\sigma_N^2}{\overline{N}^3}\right)\overline{T}, \tag{74}$$

so to leading order the bias for this definition of temperature is $O(1/\overline{N})$.

For the alternative definition of instantaneous temperature, equation (20), we have

$$\begin{aligned}\langle \hat{T}_1 \rangle_\infty &= \frac{1}{2c_V} \sum_{N=2}^{\infty} \int d\mathbf{v}_1 \ldots \int d\mathbf{v}_N \left(\frac{1}{N-1} \sum_{k=1}^{N} |\mathbf{v}_k - \hat{\mathbf{u}}|^2\right) P_{01}(N) P(\mathbf{v}_1) \ldots P(\mathbf{v}_N) \\ &= \overline{T} \sum_{N=2}^{\infty} \frac{N}{N-1}\left(1 - \frac{\overline{|\delta \hat{\mathbf{u}}|_N^2}}{d\, C_T^2}\right) P_{01}(N),\end{aligned} \tag{75}$$

where $\overline{|\delta \hat{\mathbf{u}}|_N^2}$ is the variance of the center-of-mass velocity for a given value of N. From (69),

$$\langle \hat{T}_1 \rangle_\infty = \frac{d\, C_T^2}{2c_V} \sum_{N=2}^{\infty} \frac{N}{N-1}\left(1 - \frac{1}{N}\right) P_{01}(N) \; = \; \overline{T} \tag{76}$$

so using this definition gives the correct mean value.

Finally, consider thecorrelation of density and temperature fluctuations; from (34) and the results for mechanical variables, $\langle \delta\rho\, \delta T \rangle_\infty^* = \overline{\delta\rho\, \delta T}$. To obtain the correlation for instantaneous temperature, we use $\langle \delta\rho\, \delta \hat{T}_\alpha \rangle_\infty = \langle \rho \hat{T}_\alpha \rangle_\infty - \overline{\rho} \langle \hat{T}_\alpha \rangle_\infty$; direct evaluation for $\hat{T}_0$ equation (18) gives

$$\begin{aligned}\langle \rho \hat{T}_0 \rangle_\infty &= \frac{m}{2c_V V} \sum_{N=1}^{\infty} \int d\mathbf{v}_1 \ldots \int d\mathbf{v}_N N \left(\frac{1}{N} \sum_{k=1}^{N} |\mathbf{v}_k - \hat{\mathbf{u}}|^2\right) P_0(N) P(\mathbf{v}_1) \ldots P(\mathbf{v}_N) \\ &= \frac{m\overline{T}}{V} \sum_{N=1}^{\infty} N\left(1 - \frac{\overline{|\delta \hat{\mathbf{u}}|_N^2}}{d\, C_T^2}\right) P_0(N) = \overline{\rho}\overline{T} \sum_{N=1}^{\infty} \frac{N-1}{\overline{N}} P_0(N),\end{aligned} \tag{77}$$

so

$$\langle \delta\rho\, \delta \hat{T}_0 \rangle_\infty = \overline{\rho}\overline{T} \sum_{N=1}^{\infty} \left(\frac{N-1}{\overline{N}} - \frac{N-1}{N}\right) P_0(N). \tag{78}$$

For the alternative definition of instantaneous temperature (equation (20)) we get

$$
\begin{aligned}
&\langle \rho \hat{T}_1 \rangle_\infty \\
&= \frac{m}{2c_V V} \sum_{N=1}^{\infty} \int d\mathbf{v}_1 \ldots \int d\mathbf{v}_N N \left(\frac{1}{N-1} \sum_{k=2}^{N} |\mathbf{v}_k - \hat{\mathbf{u}}|^2 \right) P_{01}(N) P(\mathbf{v}_1) \ldots P(\mathbf{v}_N) \\
&= \frac{m\overline{T}}{V} \sum_{N=2}^{\infty} \frac{N^2}{N-1} \left(1 - \frac{\overline{|\delta \hat{\mathbf{u}}|_N^2}}{d\, C_T^2} \right) P_{01}(N) \\
&= \overline{\rho}\overline{T} \sum_{N=2}^{\infty} \frac{N}{\overline{N}} P_{01}(N) = \frac{\overline{\rho}\overline{T}}{\overline{N}} \left(\frac{\overline{N} - P(1)}{1 - P(0) - P(1)} \right),
\end{aligned}
$$

so

$$
\langle \delta\rho\, \delta\hat{T}_1 \rangle_\infty = \frac{\overline{\rho}\overline{T}}{\overline{N}} \left(\frac{\overline{N} P(0) + (\overline{N} - 1) P(1)}{1 - P(0) - P(1)} \right). \qquad (79)
$$

Appendix B: Variances from fluctuating hydrodynamics

This appendix lists the variances and covariances of mechanical and hydrodynamic variables in the case of thermodynamic equilibrium at the mean state, $\overline{\rho}$, $\overline{\mathbf{u}}$, and $\overline{T}$. These results are from the theory of fluctuating hydrodynamics (§132, [18]) as developed from equilibrium statistical mechanics (§112, [19]).

The variance of mass density depends on the compressibility (i.e., the equation of state) of the fluid. In general,

$$
\overline{\delta\rho^2} = \overline{\rho}^2 \, \frac{\sigma_N^2}{\overline{N}^2}, \qquad (80)
$$

where $\overline{N} = \overline{\rho} V / m$ and σ_N^2 is the variance of N at equilibrium. For example, for an ideal gas N is Poisson-distributed so $\sigma_N^2 = \overline{N}$ and $\overline{\delta\rho^2} = \overline{\rho}^2 / \overline{N}$. The more general result is $\sigma_N^2 = -(k_B \overline{T}\, \overline{N}^2 / V^2)(\partial V / \partial \overline{P})_T$.

The variances of fluid velocity and temperature are

$$
\overline{|\delta\mathbf{u}|^2} = d\, \frac{k_B \overline{T}}{\overline{\rho} V} = d\, \frac{C_T^2}{\overline{N}} \qquad (81)
$$

$$
\overline{\delta T^2} = \frac{k_B \overline{T}^2}{c_v \overline{\rho} V} = \frac{C_T^2 \overline{T}}{c_v \overline{N}} \qquad (82)
$$

where $C_T = \sqrt{k_B \overline{T} / m}$ is the thermal speed (and the standard deviation of the Maxwell-Boltzmann distribution). The covariances are $\overline{\delta\rho\, \delta\mathbf{u}} = \overline{\delta\rho\, \delta T} = \overline{\delta\mathbf{u}\, \delta T} = 0$.

From the results above and those formulated in Section 3, the variances and covariances of the mechanical densities at equilibrium are

$$\overline{\delta\rho\delta\mathbf{J}} = \overline{\rho}\bar{\mathbf{J}}\Delta_\rho, \tag{83}$$

$$\overline{\delta\rho\delta K} = \overline{\rho}\bar{K}\Delta_\rho, \tag{84}$$

$$\overline{\delta J_\alpha \delta J_\beta} = \bar{J}_\alpha \bar{J}_\beta \Delta_\rho + \overline{\rho}^2 C_T^2 \Delta_u \delta_{\alpha,\beta}, \tag{85}$$

$$\overline{|\delta\mathbf{J}|^2} = |\bar{\mathbf{J}}|^2 \Delta_\rho + d\,\overline{\rho}^2 C_T^2 \Delta_u, \tag{86}$$

$$\overline{\delta\mathbf{J}\,\delta K} = \bar{\mathbf{J}}\,\bar{K}\Delta_\rho + \bar{\mathbf{J}}\,\rho C_T^2 \Delta_u, \tag{87}$$

$$\overline{\delta K^2} = \bar{K}^2 \Delta_\rho + |\bar{\mathbf{J}}|^2 C_T^2 \Delta_u + c_v^2 \overline{\rho}^2 \overline{T}^2 \Delta_T, \tag{88}$$

where $\bar{\mathbf{J}} = \overline{\rho}\,\bar{\mathbf{u}}$ and $\bar{K} = c_v \overline{\rho}\overline{T} + \frac{1}{2}\overline{\rho}|\bar{\mathbf{u}}|^2$; the dimensionless variances are defined by (80), (81), and (82) normalized as $\Delta_\rho = \overline{\delta\rho^2}/\overline{\rho}^2$, $\Delta_u = \overline{\delta u_x^2}/C_T^2$, and $\Delta_T = \overline{\delta T^2}/\overline{T}^2$.

Appendix C: Equilibrium simulations

Simple stochastic simulations of a dilute gas at thermodynamic equilibrium were performed to verify and illustrate the results obtained by direct evaluation (see Appendix A). Sample means and variances of fluid velocity and temperature, using the various definitions, were computed and compared with theoretical predictions, as shown in the figures in Section 4.

From the principle of equipartition, at thermodynamic equilibrium the velocities of the particles are Maxwell–Boltzmann-distributed,

$$P(\mathbf{v}) = \left(\frac{m}{2\pi k_B \overline{T}}\right)^{d/2} \exp(-m|\mathbf{v}_{k,j} - \bar{\mathbf{u}}|^2/2k_B\overline{T}), \tag{89}$$

with mean $\bar{\mathbf{v}} = \bar{\mathbf{u}}$ and variance $\overline{|\mathbf{v} - \bar{\mathbf{v}}|^2} = \overline{|\delta\mathbf{v}|^2} = d\,C_T^2$ where $C_T \equiv \sqrt{k_B\overline{T}/m}$ is the thermal speed. Note that this distribution is not restricted to a dilute gas but applies to any classical fluid at equilibrium. Also note that thermodynamic equilibrium does *not* imply $\bar{\mathbf{u}} = 0$ since a system is in equilibrium in all inertial frames of reference.

The number of particles in a given sample, N_j, is a random variable whose distribution depends on the equation of state for the fluid. For the simulations we take the case of a dilute gas, so N_j is Poisson-distributed,

$$P(N_j) = \frac{e^{-\overline{N}}\,\overline{N}^{N_j}}{N_j!} \tag{90}$$

with mean $\overline{N} = \overline{N}$ and variance $\overline{\delta N^2} = \overline{N}$.

Each simulation run consisted of $S = 5000$ samples for fixed $\overline{N}$, varying from 0.5 to 20, and arbitrary $\bar{\mathbf{u}}$ and $\overline{T}$. For each sample, given (90), a random value of N_j was generated and then that many random particle velocities were generated according to (89). Means, variances, and correlations were estimated by the various definitions presented in sections 2 and 3; note that for some definitions (e.g., (12), (18), (20)) samples containing zero or one particle are omitted in evaluating sample means.

Appendix D: Non-equilibrium simulations

In Section 4.2 the mean instantaneous temperature $\langle \hat{T}_1 \rangle_s$ is predicted to have a bias due to nonequilibrium correlations of density-temperature fluctuations. To test this prediction, molecular simulations of a dilute gas were performed to measure $\langle T \rangle_s^*$, $\langle \hat{T}_1 \rangle_s$, and $\langle \delta\rho, \delta T \rangle_s^*$ (see equation (46) and Figure 3). The simulations were of a nonequilibrium state, specifically a temperature gradient produced by parallel thermal walls at different temperatures. Similar simulations in [26] verified the predicted bias in the instantaneous center-of-mass fluid velocity (see equation (39)).

The simulations used the direct simulation Monte Carlo (DSMC) algorithm, a well-known method for computing gas dynamics at the molecular scale; see [1; 8] for pedagogical expositions on DSMC, [4] for a complete reference, and [27] for a proof of the method's equivalence to the Boltzmann equation. As in molecular dynamics, the state of the system in DSMC is given by the positions and velocities of particles. In each time step, the particles are first moved as if they did not interact with each other. After moving the particles and imposing any boundary conditions, collisions are evaluated by a stochastic process, conserving momentum and energy and selecting the postcollision angles from their kinetic theory distributions. DSMC is a stochastic algorithm but the statistical variation of the physical quantities has nothing to do with the "Monte Carlo" portion of the method. The equilibrium and nonequilibrium variations in DSMC are the physical spectra of spontaneous thermal fluctuations, as confirmed by excellent agreement with fluctuating hydrodynamic theory [9; 20] and molecular dynamics simulations [21; 22].

The nonequilibrium system we consider is a dilute monatomic hard-sphere gas between a pair of parallel thermal walls. The left wall is at the reference temperature of 273 Kelvin and the right wall's temperature is three times greater. Two cases, hydrodynamically equivalent, are simulated. The distance between the walls is the same in the two cases, but one system is 16 times larger in volume (and has 16 times more particles) than the other. All other parameters (e.g., mean free path, transport coefficients) were the same in the two systems (see Table 1). Samples are taken in forty rectangular cells sliced parallel to the thermal walls; in the large system these cells are 16 times larger than in the small system. Starting near the steady state (approximately linear temperature profile) the simulations of these two systems are run for 2.5×10^7 time steps to dissipate any initial transients. After

Molecular diameter (Argon)	3.66×10^{8}
Molecular mass (Argon)	6.63×10^{23}
Reference mass density	1.78×10^{-3}
Reference temperature	273
Sound speed	33700
Specific heat c_v	3.12×10^{6}
Wall temperature (left)	273
Wall temperature (right)	819
System length	1.25×10^{4}
Reference mean free path	6.26×10^{-6}
System volume (large)	1.96×10^{-16}
System area (small)	1.23×10^{-17}
Number of particles (large)	5265
Number of particles (small)	329
Number of sampling cells	40
Number of samples, S	2.5×10^{7}
DSMC time step	1.0×10^{-11}
DSMC grid size	2.09×10^{-6}

Table 1. System parameters (in cgs units) for DSMC simulations of a dilute gas between thermal walls.

allowing the systems to relax, samples are taken at each time step for a total of $S = 2.5 \times 10^{7}$ samples.

Acknowledgements

The author wishes that thank M. Malek Mansour, M. Tysanner, and W. Wagner for helpful discussions. This work was completed in part during the author's visit to Weierstrass Institute for Applied Analysis and Stochastics and partially funded by the Department of Energy through Lawrence Berkeley National Laboratory subcontract #6719615.

References

[1] F. J. Alexander and A. L. Garcia, *The direct simulation Monte Carlo method*, Computers in Physics **11(6)** (1997), 588–593.

[2] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Clarendon Press, 1987.

[3] R. D. Astumian and P. Hanggi, *Brownian motors*, Physics Today (November 2002), 33–39.

[4] G. A. Bird, *Molecular gas dynamics and the direct simulation of gas flows*, Oxford Engineering Science Series, vol. 42, The Clarendon Press Oxford University Press, New York, 1995. MR 97e:76078

[5] J. Eggers, *Dynamics of liquid nanojets*, Physical Review Letters **89(8)** (2002), 084502.

[6] D. J. Evans and G. P. Morriss, *Statistical mechanics of nonequilibrium liquids*, Academic Press, 1990.

[7] D. Frenkel and B. Smit, *Understanding molecular simulation*, Academic Press, 2002.

[8] A. L. Garcia, *Numerical methods for physics*, Prentice Hall, 2000.

[9] A. L. Garcia, M. M. Mansour, G. C. Lie, M. Mareschal, and E. Clementi, *Hydrodynamic fluctuations in a dilute gas under shear*, Physical Review A **36** (1987), 4348–4355.

[10] R. D. Groot and P. B. Warren, *Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation*, The Journal of Chemical Physics **107(11)** (1997), 4423–4435.

[11] Nicolas G. Hadjiconstantinou, Alejandro L. Garcia, Martin Z. Bazant, and Gang He, *Statistical error in particle simulations of hydrodynamic phenomena*, J. Comput. Phys. **187** (2003), no. 1, 274–297. MR 2004c:76113 Zbl 1047.76578

[12] R. W. Hockney and J. W. Eastwood, *Computer simulation using particles*, Inst. of Physics Publ., 1988.

[13] George Em Karniadakis and Ali Beskok, *Micro flows*, Springer-Verlag, New York, 2002. MR 2002i:76050 Zbl 1035.76052

[14] M. Karplus and J. Kuriyan, *Molecular dynamics and protein function*, Proc Natl Acad Sci U S A **102(19)** (2005), 6679–85.

[15] J. Koplik and J. R. Banavar, *Continuum deductions from molecular hydrodynamics*, Annual Review of Fluid Mechanics **27** (1995), 257–292.

[16] Petros Koumoutsakos, *Multiscale flow simulations using particles*, Annual review of fluid mechanics. Vol. 37, Annu. Rev. Fluid Mech., vol. 37, Annual Reviews, Palo Alto, CA, 2005, pp. 457–487. MR 2005i:76085 Zbl 02212509

[17] G. C. Lie L. Hannon and E. Clementi, *Molecular dynamics simulation of flow past a plate*, J. Sci. Comput. **1(2)** (1986), 145–150.

[18] L. D. Landau and E. M. and Lifshitz, *Fluid mechanics*, Translated from the Russian by J. B. Sykes and W. H. Reid. Course of Theoretical Physics, Vol. 6, Pergamon Press, London, 1959. MR 21 #6839 Zbl 0655.76001

[19] E. M. Lifshitz and L. P. Pitaevskiĭ, *Course of theoretical physics*, vol. 9, Pergamon Press, Oxford, 1980. MR 84m:82003a Zbl 0655.76001

[20] G. C. Lie M. M. Mansour, A. L. Garcia and E Clementi, *Fluctuating hydrodynamics in a dilute gas*, Physical Review Letters **58** (1987), 874–877.

[21] J. W. Turner M. Malek Mansour, A. L. Garcia and M. Mareschal, *On the scattering function of simple fluids in finite systems*, J. Stat. Phys. **52** (1988), 295.

[22] G. Sonnino M. Mareschal, M. M. Mansour and E.Kestemont, *Dynamic structure factor in a nonequilibrium fluid: A molecular-dynamics approach*, Physical Review A **45** (1992), 7180–7183.

[23] G. Oster, *Darwin's motors*, Nature **417** (2002), 25.

[24] J.-P. Rivet and J. P. Boon, *Lattice gas hydrodynamics*, Cambridge Nonlinear Science Series, vol. 11, Cambridge University Press, Cambridge, 2001. MR 2002c:82077

[25] M. Tysanner and A. L. Garcia, *Non-equilibrium behavior of equilibrium reservoirs in molecular simulations.*, International Journal of Numerical Methods in Fluids, (to appear) 2005.

[26] ———, *Measurement bias of fluid velocity in molecular simulations*, Journal of Computational Physics **196** (2004), 173–183.

[27] W. Wagner, *A convergence proof for bird's direct simulation monte carlo method for the boltzmann equation*, Journal of Statistical Physics **66** (1992), 1011.

Received June 19, 2005. Revised January 12, 2006.

ALEJANDRO L. GARCIA: algarcia@algarcia.org
Department of Physics, San José State University, San José, CA 95192-0106, United States
www.algarcia.org

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

BIFURCATED EQUILIBRIA AND MAGNETIC ISLANDS IN TOKAMAKS AND STELLARATORS

PAUL R. GARABEDIAN

The magnetohydrodynamic variational principle is employed to calculate equilibrium and stability of toroidal plasmas without two-dimensional symmetry. Differential equations are solved in a conservation form that describes force balance correctly across islands that are treated as discontinuities. The method is applied to both stellarators and tokamaks, and comparison with observations is favorable in both cases. Sometimes the solution of the equations turns out not to be unique, and there exist bifurcated equilibria that are nonlinearly stable when other theories predict linear instability. The calculations are consistent with recent measurements of high values of the pressure in stellarators. For tokamaks we compute three-dimensionally asymmetric solutions that are subject to axially symmetric boundary conditions.

1. Introduction

A community of industrialized nations is planning construction of the International Thermonuclear Experimental Reactor (ITER). A facility has been designed to test the concept of fusing deuterium and tritium ions so as to form helium and release energetic neutrons that can produce electric power at commercially viable cost [1]. This is to be achieved by confining a very hot plasma of ions and electrons in a strong magnetic field with toroidal geometry and a major radius of 6m. The magnetic fusion configuration preferred for ITER is a tokamak, which is axially symmetric and requires net toroidal current for confinement of the plasma. An alternate concept that seems to be more stable is the stellarator, which has fully three-dimensional geometry generating a poloidal field that eliminates the need for induced current.

Recent advances in high performance computing have led to significant progress in the theory of equilibrium, stability and transport for fusion plasmas in three dimensions. This has made it possible to design stellarators that are competitive

MSC2000: 82D10, 70K50, 82C40.

Keywords: magnetic fusion, plasma physics, equilibrium and stability.

This work has been supported by Grant DE-FG02-86ER53223 from DOE. I wish to express my appreciation for helpful discussions with Frances Bauer, Leon Shohet, Alan Turnbull and Harold Weitzner.

with tokamaks as candidates for a fusion reactor. The work we shall describe in this direction is based on the NSTAB, VMEC and TRAN computer codes [2; 4; 9; 12; 17]. In particular, we consider simulations of anomalous thermal transport in tokamaks that result from calculations of bifurcated equilibria that do not have two-dimensional symmetry. For both tokamaks and stellarators difficult mathematical problems are encountered because accurate solutions of the relevant differential equations turn out to have discontinuities associated with islands and current sheets in the plasma (see Figure 1).

We begin with a study of weak solutions of the partial differential equations governing magnetohydrodynamic (MHD) equilibrium in three dimensions. Then we examine the role played by the magnetic spectrum in estimating the prompt loss of α particles in a reactor. Finally, we discuss candidates for a demonstration of the magnetic fusion concept after the ITER project is completed.

2. Computation of force balance

The KAM theory of dynamical systems predicts that smooth solutions of the partial differential equations describing MHD equilibrium of a toroidal plasma cannot be found in the absence of two-dimensional symmetry [2]. Let B be the magnetic field, $J = \nabla \times B$ be the current density, $p = p(s)$ be the scalar pressure, s be the toroidal flux, $\theta + \iota\phi$ and ϕ be invariant poloidal and toroidal angles, and ι be the rotational transform measuring how far a magnetic line circulates poloidally during one transit the long way around the torus. We call the Fourier coefficients B_{mn} in a representation

$$1/B^2 = \sum B_{mn}(s) \cos\bigl(m\theta - [n - \iota m]\phi\bigr)$$

of the magnetic field strength the magnetic spectrum. For stellarators an elementary manipulation of the MHD equations leads to a corresponding formula

$$\frac{J \cdot B}{B^2} = p' \sum \frac{m B_{mn}}{n - \iota m} \cos\bigl(m\theta - [n - \iota m]\phi\bigr)$$

for the parallel current in which the term with $m = n = 0$ is omitted. In this context the small denominators $n - \iota m$ explain why continuous solutions of the fully three-dimensional equilibrium problem do not exist under the hypothesis that the plasma is covered by nested toroidal flux surfaces s = const., which is important for good confinement.

To handle discontinuous solutions of the MHD equilibrium equations we write them in the conservation form

$$\nabla \cdot B = \nabla \cdot T = 0\,,$$

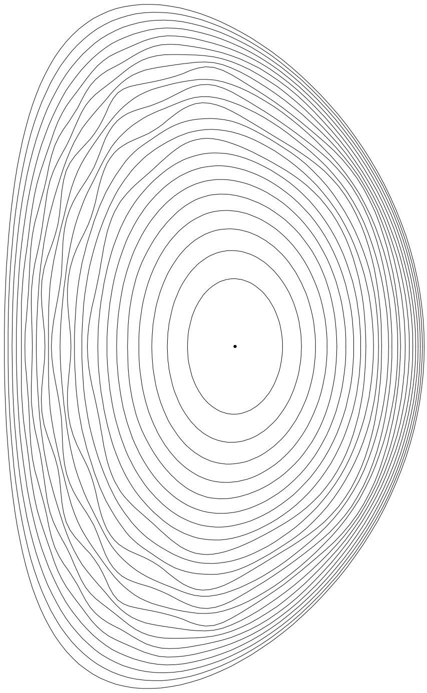


Figure 1. Poincaré section of the flux surfaces of a bifurcated ITER equilibrium at $\beta = 0.03$ with $p = p_0(1 - s^{1.1})^{1.1}$ and with net current bringing the rotational transform into the interval $0.8 > \iota > 0.2$. Ripples in the surfaces represent helical islands in this fully converged three-dimensional solution of an axially symmetric MHD problem.

where

$$T = BB - (B^2/2 + p)I$$

is the Maxwell stress tensor. Then force balance over any test volume of plasma reduces by the divergence theorem to the assertion that the surface integral

$$\iint T \cdot N \, dS = 0$$

vanishes over the boundary. Numerical methods that employ an analogous discrete conservation form of the equations provide an accurate approximation of force balance because when they are similarly summed over any collection of mesh points the result telescopes down to a corresponding statement at the boundary.

We illustrate the way conservation form captures discontinuities in weak solutions of the MHD equations by considering a one-dimensional example of a reversed field pinch (RFP) in slab geometry [8]. In a rectangular coordinate system we conceive of x as a radius and y and z as toroidal and poloidal angles. Let $(0, \Psi_x, C)$, $(0, 0, \Psi_{xx})$ and $\eta(\Psi_{xxx}, 0, 0)$ represent the magnetic field, the current density and an artificial resistivity, respectively, where Ψ is a flux function depending only on x, and C and η are constants. The MHD equilibrium equations reduce to an ordinary differential equation that we write in the conservation form

$$(\Psi_x^2)_x = \eta\Psi_{xxx} \,,$$

and we seek a solution on the interval $-1 \leq x \leq 1$ satisfying the boundary conditions

$$\Psi(-1) = \Psi(+1) = 0 \,, \quad \Psi_x(-1) = 1 \,.$$

The finite difference approximation

$$(\Psi_{n+1} - \Psi_n)^2 - (\Psi_n - \Psi_{n-1})^2 = \eta(\Psi_{n+2} - 3\Psi_{n+1} + 3\Psi_n - \Psi_{n-1})$$

of the RFP equation is in conservation form and defines iterations that converge in the limiting case $\eta = 0$ to the correct answer $\Psi = 1 - |x|$. This has a jump in its derivative at the origin, but satisfies force balance there because Ψ_x^2 remains continuous. It is easy to find less symmetrical difference schemes for the same boundary value problem that are not in conservation form and therefore give results that violate force balance significantly. The numerical example we have presented is also applied in computational fluid dynamics to show that conservation form is required to capture shock waves accurately [3].

The NSTAB computer code calculates toroidal equilibrium of stellarators and tokamaks by a numerical scheme that is in a conservation form associated with the MHD variational principle [4; 17]. Good convergence is achieved by applying the spectral method to describe dependence of the solution on the poloidal and toroidal angles and by using an exceptionally accurate finite difference approximation in the radial coordinate s. The high resolution of the radial scheme has been established by comparing numerical results with exact solutions [2]. The NSTAB code models magnetic fusion configurations effectively using a suitable Fourier series to represent the fixed boundary of the plasma.

Linear and nonlinear stability are tested by looking for bifurcated equilibria that do not have symmetries occurring in conventional models. This method has provided acceptable simulations of experiments for stellarators that exceed stability predictions of linear theory [6]. More specifically, our computations agree with recent observations [18] in the Large Helical Device (LHD) at the National Institute for Fusion Science (NIFS) in Japan of values of the performance parameter $\beta = 2p/B^2$ as high as 4%. The equilibria we examine for the LHD at such values of

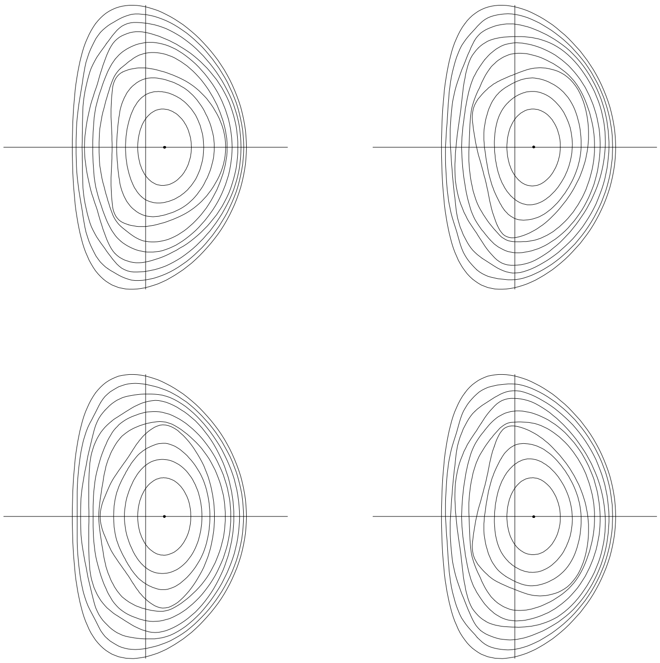


Figure 2. Four cross sections of the flux surfaces over half the torus of a bifurcated DIII-D equilibrium at $\beta = 0.02$ with $p = p_0(1-s^{1.1})^{1.1}$ and with net current bringing the rotational transform into the interval $0.9 > \iota > 0.3$. There is a large $m = 3, \ n = 2$ magnetic island at $\iota = 2/3$ in the solution that models an observed mode.

β tend to be linearly unstable, but nonlinearly stable, so that a better understanding of bifurcated solutions becomes desirable [11].

Our computational method has been applied to study neoclassical tearing modes (NTM) in the Doublet III-D (DIII-D) tokamak at General Atomics (GA) with the net current limited so that $\iota < 1$. Three-dimensional equilibria are calculated by at first imposing, but much later releasing, a suitable constraint in runs of the NSTAB code chosen to find bifurcated solutions that cannot be obtained without introducing discontinuous alterations in the topology of the magnetic surfaces. Figure 2 displays Poincaré sections of the flux surfaces of such a bifurcated equilibrium. Solutions like this are related to observations of NTM modes made in the experiment [5; 13]. On crude radial grids the computations are capable of capturing slender islands whose widths are comparable to the mesh size. The physical significance of finding many three-dimensional MHD equilibria in axially symmetric tokamaks needs

m	n	B_{mn}
0	0	0.997
1	0	0.525
2	0	0.144
3	0	0.058
4	0	0.025
0	1	0.015
3	2	0.010
1	1	0.009
4	2	0.008
1	-1	0.007

Table 1. Nontrivial coefficients in the spectrum of an optimized MHH2 configuration with a prompt loss of α particles below 10%.

further investigation. More specifically, one can ask how much their effect might contribute to the prompt loss of α particles or to disruptions.

3. Prompt loss of α particles

Neoclassical transport in tokamak and stellarator plasmas with three-dimensional geometry can be evaluated by tracking guiding center orbits of charged particles that are subjected to a random walk representing collisions. The TRAN computer code implements such a method that employs equilibria obtained from NSTAB calculations, which are needed to estimate the magnetic spectrum [9]. Substantial agreement has been found between computations of thermal transport from runs of the TRAN code and experimental observations in tokamaks and stellarators [7]. An algorithm determining the electric potential from quasineutrality in three-dimensional equilibria has been applied successfully. This theory has been used to demonstrate the advantage for stellarator transport of a magnetic spectrum with quasihelical symmetry (QHS), where only the diagonal coefficients B_{mm} are large, or with quasiaxial symmetry (QAS), where the first column of coefficients B_{m0} dominate [10; 16]. The computational approach facilitates designing new configurations that may bring the concept of magnetic fusion closer to construction of a commercially viable reactor.

The TRAN code has been modified to estimate the prompt loss of α particles in a fusion plasma at reactor conditions. This is defined to be the percentage of α particles that escape from the plasma during one slowing down time after they are born. Samples of between 128 and 1024 particles are adequate to give a meaningful answer, but that requires significant resources on a commodity work station. For stellarators the spectrum again plays a decisive role in the computations. Experience shows that only a significant improvement in the quasisymmetry required for satisfactory thermal transport can produce a loss of α particles as low as 10% that might be acceptable in the design of a fusion reactor. Table 1 lists averages with respect to s of the largest coefficients B_{mn} in a two field period configuration that has been optimized for such an application [8]. The three-dimensional asymmetry is seen to fall below half a percent if it is measured in units of the field strength B itself rather than $1/B^2$. To achieve this level of quasisymmetry presents a challenge not only to the accuracy of the codes that are used, but also to the precision of the hardware that must be fabricated.

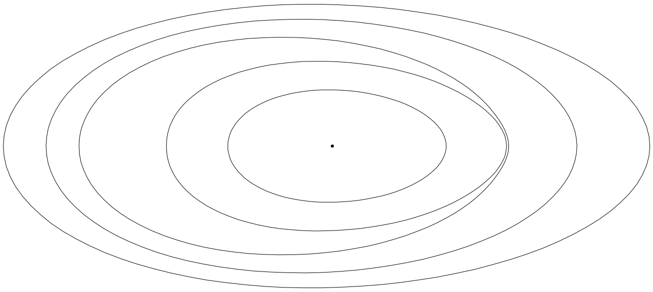


Figure 3. Zero β calculation of a Poincaré section of flux surfaces for a stellarator with reversed poloidal field. Two magnetic surfaces touch each other at an X-point where the rotational transform ι changes sign. They surround a magnetic island that would otherwise be obscured by the nested surface hypothesis implemented in the NSTAB code.

4. Magnetic islands

The NSTAB code captures islands successfully despite a nested surface hypothesis made in the coordinate system that is employed [11]. The resolution of the code can be checked by applying it to the vacuum field of stellarators where islands are known to exist in equilibria found by line tracing [14]. Figures 3 and 4 display calculations of an example of this phenomenon in which the rotational transform changes sign so that a sizeable island appears at $\iota = 0$. The same numerical construction produces helical islands in tokamaks like the DIII-D and ITER. When such three-dimensional solutions of the tokamak problem were used in computations of the energy confinement time, anomalous transport was not observed in the results [9].

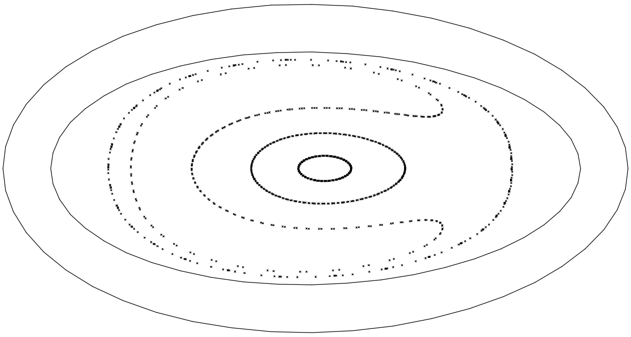


Figure 4. Tracing of magnetic lines through a Poincaré section of a stellarator with reversed poloidal field. A large magnetic island is seen where the rotational transform ι changes sign. The plasma surface is plotted together with a control surface used for Biot–Savart computation of the vacuum magnetic field.

The problem of ideal MHD equilibrium is singular in two dimensions and includes a continuous spectrum in the analysis of stability, and in three dimensions the KAM theory shows that no differentiable solutions exist [2]. So we introduce weak solutions of the kind constructed numerically by the NSTAB code, which have magnetic islands that appear as discontinuities like current sheets. Artificial resistivity implicit in the code captures the islands in a realistic fashion because of the conservation form of the MHD equations that is employed. That results in turn from a mixed Euler–Lagrange coordinate system featuring the toroidal flux as a radius. The method produces three-dimensional tokamak equilibria with small magnetic islands whose cumulative effect simulates experimental observations better than two-dimensional models do [9].

5. Configurations for a fusion reactor

A tokamak like ITER is the candidate of choice by the fusion community for a reactor. Disadvantages are that MHD instability tends to trigger disruptions, and it is hard to control the induced net current in a steady state. The calculations of NTM in the DIII-D at GA that we have described suggest that bifurcated equilibria with three-dimensional asymmetries may turn out to be important in attacking these problems [13]. Of special interest for reactors is that nuclear engineers may ultimately come to prefer a stellarator-tokamak hybrid with good quasisymmetry and small aspect ratio.

It is relatively easy to reduce the prompt loss of α particles in stellarators that have good QHS, such as the Helically Symmetric Experiment (HSX) at the University of Wisconsin with four field periods or the Wendelstein 7-X (W7-X) at Greifswald in Europe with five field periods [16]. We have studied a QHS version of the W7-X with rotational transform in the interval $1 < \iota < 5/4$ that has favorable properties of thermal transport and MHD stability. The prompt loss of α particles can be brought down to several percent by readjustment of the coefficients B_{mn} in the spectrum, but many twisted modular coils are required to maintain an equilibrium with low toroidal ripple of the magnetic field strength because the aspect ratio of the plasma is large.

Most of our theoretical work with the NSTAB and TRAN computer codes has been focused on QAS stellarators like the National Compact Stellarator Experiment (NCSX) at the Princeton Plasma Physics Laboratory (PPPL) with three field periods and the Modular Helias-like Heliac 2 (MHH2) with just two field periods [10]. The NCSX is a principal candidate for the ARIES_CS compact stellarator study of magnetic fusion reactors [15] funded by the United States Department of Energy (DOE). It is difficult to reduce the prompt loss of α particles in both the NCSX and the MHH2 because the necessary calculations are sensitive to small changes in the magnetic spectrum [8]. Net current that raises the rotational transform is helpful in these optimizations. For that one attractive configuration is a hybrid version of the MHH2 shown in Figure 5, which has major radius 8m and plasma radius 3m.

It is hard to find modular coils that generate an external magnetic field compatible with a plasma equilibrium optimized to bring the loss of α particles below 10% at reactor conditions. One possibility is to determine the solution inside the plasma from an equilibrium calculation and then apply the Biot–Savart law to match that with a vacuum field defined by a distribution of current on a suitably chosen control surface where the coils are to be placed [14]. This method could be applied to smooth out unrealistic surface current on the separatrix of an alternate approximation found by solving a free boundary value problem. The analysis taxes the resolution of the best computer codes that are available because there is a high degree of magnetic quasisymmetry required in the answer. Moreover, the harmonics specifying the coils have to be filtered judiciously to eliminate erroneous excursions. The concept is elucidated by Runge's theorem, which asserts that an analytic function can be approximated by polynomials in any simply connected region of the complex plane.

The MHH2 configuration that has been optimized to reduce the prompt loss of α particles is a good candidate for a stellarator experiment to achieve ion temperatures competitive with those in tokamaks. Moreover, three-dimensional equilibria are found numerically in tokamaks, so two-dimensional models may be less realistic. Because truncation error in the computations is insignificant compared to sources hitting the plasma in an experiment, observations may exhibit effects associated with three-dimensional asymmetries in a bifurcated solution of the problem.

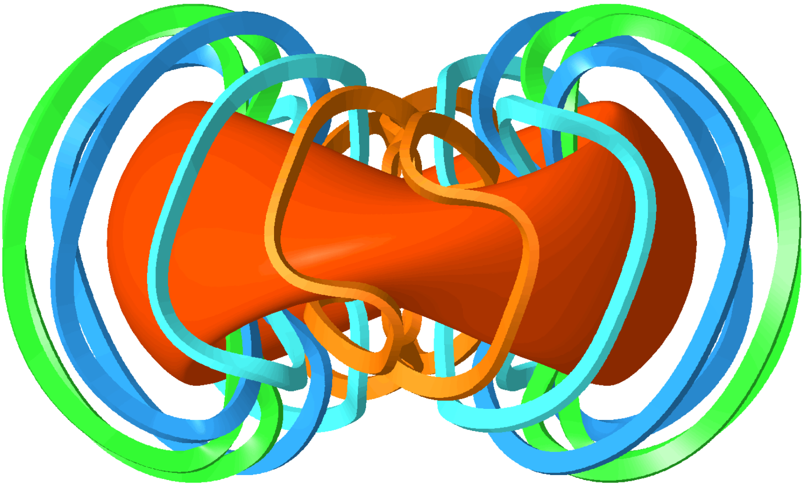


Figure 5. Diagram of a fusion reactor with low prompt loss of α particles in a magnetic field given by the Biot–Savart law. Sixteen moderately twisted modular coils produce robust flux surfaces that do not deteriorate when changes are made in the vertical and toroidal fields. This optimized configuration with two field periods has stellarator stability and tokamak transport. (Courtesy of Tak-Kuen Mau and Tsueren Wang.)

6. Conclusion

The NSTAB code has been applied to calculate a variety of bifurcated equilibria in tokamaks with axially symmetric boundary conditions. The KAM theory of dynamical systems displays small denominators at rational surfaces of 3D solutions, and analysis of the continuous spectrum shows that linear stability of tokamaks is singular. This is consistent with observations of sawtooth oscillations, NTM and ELMS, and disruptions. Desirable 3D solutions of the MHD equations for equilibrium may not exist, may not be unique, and may not depend continuously on the data. Yet success of the DIII-D and LHD experiments fosters a belief that it is possible to design a magnetic fusion reactor. Perhaps a QAS stellarator of very low aspect ratio is the answer, since it is helpful if some of the rotational transform comes from the external magnetic field.

References

[1] R. Atkinson and F. Houtermans, *Zuer frage der aufbaumöglichkeit der Elemente in Sternen*, Zeit. Phys. **54** (1929), 656–665.

[2] F. Bauer, O. Betancourt, and P. Garabedian, *Magnetohydrodynamic equilibrium and stability of stellarators*, Springer, New York, 1984.

[3] F. Bauer, P. Garabedian, D. Korn, and A. Jameson, *Supercritical wing sections ii*, Springer, New York, 1975.

[4] Octavio Betancourt, *BETAS, a spectral code for three-dimensional magnetohydrodynamic equilibrium and nonlinear stability calculations*, Comm. Pure Appl. Math. **41** (1988), no. 5, 551–568. MR 89h:76045 Zbl 0633.76124

[5] D. Brennan, R. La Haye, A. Turnbull, M. Chu, T. Jensenand L. Lao, S. Kruger, and D. Schnack, *A mechanism for tearing onset near ideal stability boundaries*, Phys. Plasmas **10** (2003), 1643–1652.

[6] W. Cooper, *Stability of a compact three-period stellarator with quasiaxial symmetry features*, Phys. Plasmas **7** (2000), 2546–2553.

[7] A. Komori et. al., *Overview of the large helical device*, Plasma Phys. Control. Fusion **42** (2000), 1165–1177.

[8] P. Garabedian and L. Ku, *Reactors with stellarator stability and tokamak transport*, Fusion Sci. Technol. **47** (2005), 400–405.

[9] P. Garabedian and M. Taylor, *Tokamak transport driven by quasineutrality and helical asymmetry*, Nucl. Fusion **32** (1992), 265–270.

[10] Paul Garabedian and Long-Poe Ku, *Quasiaxially symmetric stellarators with three field periods*, Phys. Plasmas **6** (1999), no. 3, 645–648. MR 99k:76161

[11] Paul R. Garabedian, *Computational mathematics and physics of fusion reactors*, Proc. Natl. Acad. Sci. USA **100** (2003), no. 24, 13741–13745. MR 2004j:76177 Zbl 1063.76112

[12] S. Hirshman and O. Betancourt, *Preconditioned descent algorithm for rapid calculations of magnetohydrodynamic equilibria*, J. Comp. Phys. **96** (1991), 99–109.

[13] T. Luce, M. Wadi, J. Ferron, P. Politzer, A. Hyatt, A. Sips, and M. Murakami, *High performance stationary discharges in the diii-d tokamak*, Phys. Plasmas **11** (2004), 2627–2636.

[14] P. Merkel, *Solution of stellarator boundary value problems with external currents*, Nucl. Fusion **27** (1987), 867–871.

[15] F. Najmabadi, *Exploration of compact stellarators as power plants: Initial results from aries_cs study*, Fusion Sci. Technol. **47** (2005), 406–413.

[16] J. Nuehrenberg and R. Zille, *Quasi-helically symmetric toroidal stellarators*, Physics Letters **129A** (1988), 113–116.

[17] Mark Taylor, *A high performance spectral code for nonlinear MHD stability*, J. Comput. Phys. **110** (1994), no. 2, 407–418. MR 94m:76104 Zbl 0795.76063

[18] K. Watanabe, H. Yamada, and A. Komori, *Mhd properties of high β plasma and recent results in lhd experiments*, Proc. U.S.-Japan Workshop on New Approaches in Plasma Confinement Experiments in Helical Systems, Kyoto University, 2004.

Received August 16, 2005. Revised December 21, 2005.

PAUL R. GARABEDIAN: paul.garabedian@nyu.edu
Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, United States
http://www.math.nyu.edu/faculty/garabedi

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

ON THE ACCURACY OF FINITE DIFFERENCE METHODS FOR ELLIPTIC PROBLEMS WITH INTERFACES

J. THOMAS BEALE AND ANITA T. LAYTON

In problems with interfaces, the unknown or its derivatives may have jump discontinuities. Finite difference methods, including the method of A. Mayo and the immersed interface method of R. LeVeque and Z. Li, maintain accuracy by adding corrections, found from the jumps, to the difference operator at grid points near the interface and by modifying the operator if necessary. It has long been observed that the solution can be computed with uniform $O(h^2)$ accuracy even if the truncation error is $O(h)$ at the interface, while $O(h^2)$ in the interior. We prove this fact for a class of static interface problems of elliptic type using discrete analogues of estimates for elliptic equations. Moreover, we show that the gradient is uniformly accurate to $O(h^2 \log(1/h))$. Various implications are discussed, including the accuracy of these methods for steady fluid flow governed by the Stokes equations. Two-fluid problems can be handled by first solving an integral equation for an unknown jump. Numerical examples are presented which confirm the analytical conclusions, although the observed error in the gradient is $O(h^2)$.

1. Introduction

Often in problems of fluid flow or wave propagation an interface between different regions exerts a force on the material, or an interface separates regions of different material properties. The static problem is formulated as an elliptic partial differential equation with possible discontinuities in the coefficients and nonhomogeneous terms, and with possible jump conditions for the unknown and its derivative across the interface. For the numerical solution a finite difference method is straightforward away from the interface, but accuracy will be lost near the interface unless special care is taken. A class of practical methods has been developed, including the method of A. Mayo [32; 34; 31] and the immersed interface method of R. LeVeque and Z. Li [24; 27; 26], in which the specified jumps at the

MSC2000: 35R05, 65N06, 65N15.

Keywords: elliptic equations, interfaces, discontinuous coefficients, finite differences, immersed interface method.

This work was supported in part by the National Science Foundation under grants DMS-0404765 (Beale) and DMS-0340654 (Layton).

interface are used to derive corrections to the difference operator when the stencil crosses the interface, and, if needed, modification of the difference operator as well [24; 27; 26]. Using Taylor expansions and incorporating jumps, the truncation error is corrected to a desired order. It has long been observed that, with grid size h and $O(h^2)$ truncation error in the interior, but only $O(h)$ truncation error near the interface, the solution is still uniformly accurate to $O(h^2)$. In this paper we provide a rigorous explanation for this fact in certain cases. Although we treat steady problems here, this class of methods is naturally suited for time-dependent problems with moving boundaries such as Stokes flow of a viscous fluid; see [25].

We consider a problem in a rectangular region Ω in R^d, $d = 2$ or 3, of the form

$$\beta_-\Delta u_- = f_- \quad \text{in } \Omega_-, \qquad \beta_+\Delta u_+ = f_+ \quad \text{in } \Omega_+, \tag{1–1}$$

$$[u] = g_0 \quad \text{on } S, \qquad [\beta\partial_n u] = g_1 \quad \text{on } S \tag{1–2}$$

in which a closed curve S ($d = 2$), or a closed surface S ($d = 3$), separates an inner region Ω_- from an outer region Ω_+, with $\Omega = \Omega_- \cup S \cup \Omega_+$. Here $[u] = u_+ - u_-$ on S and similarly for $\beta\partial_n u = \beta\partial u/\partial n$, where n is the normal to S, outward from Ω_-. We assume here that $\beta_\pm$ are positive constants, although operators in divergence form with variable coefficients are dealt with by the immersed interface method. We suppose u is given on $\partial\Omega$. If the problem is given in free space, the solution might first be computed on $\partial\Omega$ from an integral representation (see Section 4). Our results hold for other boundary conditions as well; the simplest would be periodicity on $\partial\Omega$.

We first treat the case $\beta_- = \beta_+$; in that case $[\partial_n u]$ is known on S. In Sections 2 and 3 we prove that, with the truncation error as above, the computed solution is uniformly $O(h^2)$ accurate, and moreover the gradient can be found uniformly to $O(h^2 \log(1/h))$. We verify that this result holds for the methods of Mayo and of LeVeque and Li. The gain in accuracy is shown to be a consequence of two facts. First, since the $O(h)$ truncation error is on a set of relative size $O(h)$, it can be written as the discrete divergence of a function which is only $O(h^2)$ in magnitude. Second, the gain in regularity in solving the discrete elliptic problem means that this part of the truncation error contributes an error to the solution which is $O(h^2)$ in a higher norm. To make this plausible, we consider an analogous estimate with continuous variable: If v is a localized function of $x \in R^d$ and $\Delta v = \sum_{k=1}^d \partial_k F_k$, then $v = \sum \partial_k G \star F_k$, where G is the fundamental solution, $\partial_k = \partial/\partial x_k$, and $\star$ denotes convolution. The kernel $\partial_k G$ is locally integrable, and if F_k is bounded, then v is bounded. Moreover, estimates for $\partial_\ell\partial_k G$ show that $\partial_\ell v \in L^p$ for any $p < \infty$. We follow a related line of argument for the discrete problem, using a discrete Green's function.

Various extensions and applications are discussed in Section 4. For the case where β_-, β_+ are unequal positive constants, the problem (1–1), (1–2) in free space can be treated by first solving an integral equation on S for $[\partial_n u]$ and then proceeding as before. The theory of Sections 2 and 3 shows that the immersed interface method for steady fluid flow governed by the Stokes equations as in [25] is second-order accurate. The two-fluid case can again be treated by first solving an integral equation. The analysis can be applied to higher-order methods; use of the nine-point Laplacian in two dimensions, rather than the usual five-point Laplacian, leads to uniform $O(h^4)$ accuracy. Mayo [32] noted that a boundary value problem could be treated as an interface problem by writing the solution as a layer potential on S and first solving a classical integral equation for the strength of the potential. A different but related method introduced by Mayo [33] and expanded on in [2] for solving interface problems or boundary value problems can also be viewed with the present analysis. This approach is to compute the solution near S as a nearly singular integral, form the discrete Laplacian, and then invert. Computational examples of the several types of problems are given in Section 5. We observe $O(h^2)$ accuracy in the gradient, indicating that the $O(h^2 \log(1/h))$ estimate proved here may not be sharp.

The gain in accuracy which is established here has been noted and analyzed since these methods were introduced [32; 33; 24]. The ideas in the Appendix of [33] are related to those used here. In [32] it was shown that, with $O(h)$ truncation error at the irregular points, the error in L^2 norm is at most $O(h^{3/2})$. Proofs of $O(h^2)$ accuracy for general equations in one dimension have been given in [3; 37; 17]. Theorems with a conclusion similar to the present one were proved in [27], Theorems 5.1 and 5.2, for a more general class of equations with Dirichlet boundary condition, using the maximum principle and comparison functions. However, this result required a hypothesis, related to the position of the interface with respect to the grid points, which does not hold in general. In particular, the hypothesis implies that, where the slope of the curve is close to horizontal, the curve cannot cross a vertical grid line closer than $C_0 h$ to a grid point, or the curve must be within $C_1 h^{1+\sigma}$ of the grid point for some $\sigma > 0$ independent of h. This hypothesis is violated for any parabola $x_2 = ax_1^2 + b$ for arbitrarily small h; this is shown in the Appendix.

Related but different methods for solving Dirichlet problems in general regions by embedding in a larger domain and using a regular grid have been used since the 1930's. At internal grid points the standard discrete Laplacian can be used, but a modified stencil must be used at the boundary of the region. A line of analysis beginning with Gerschgorin, and presented in [15], Section 23, shows that the order of accuracy can exceed that of the truncation error at the boundary by 2 under certain circumstances; for example, the accuracy of the solution can be $O(h^2)$

when the truncation error is $O(h^2)$ in the interior, but only $O(1)$ near the boundary. The method of proof is based on the maximum principle, and the gain in accuracy depends on the modification of the difference operator at the boundary. A general approach for such results using discrete Green's functions was developed in [5; 6], and a convergence proof for a class of methods with interpolation at the boundary was given by Böhmer [4]. For a recent review, see Jomaa and Macaskill [21]. Analysis and examples in [21] indicate that an $O(h)$ truncation error at the interface is preferable despite the theoretical results. As noted in [32],[37], the interface methods studied here can be used to solve boundary value problems, extending past the boundary to a computational box. This approach has the important difference from the one just described that the stencil of the differential operator is not modified at the boundary.

Elliptic problems with interfaces can be solved by finite element methods. Convergence results include [9; 11; 29]. In [28] a Cartesian grid method using a finite element formulation is introduced, and the various numerical approaches to interface problems are discussed and compared. Discrete elliptic estimates like Lemma 2.3 below are well known for finite element approximations to elliptic problems; see [10], Chapter 8 and [12], Section 21.

The main result is presented in Section 2 as Theorem 2.1. It gives the error estimate for the solution of Equations (1–1), (1–2) with $\beta_\pm = 1$, assuming estimates for the truncation error. The theorem follows from two facts: Lemma 2.2 shows that a grid function localized near the interface can be written as the divergence of a function smaller in norm, and Lemma 2.3 gives a maximum norm estimate for a discrete elliptic problem with a nonhomogeneous term of divergence form. The applicability to the methods of Mayo, LeVeque, and Li is explained, including a discussion of smoothness properties needed to justify the truncation error. The lemmas are proved in Section 3. Extensions and applications are given in Section 4, and computational examples are presented in Section 5.

2. Main results

We consider the interface problem with $\beta_+ = \beta_-$, a positive constant. For simplicity, we assume $\beta_\pm = 1$. We write the problem (1–1), (1–2) as

$$\Delta u_\pm = f_\pm \quad \text{in } \Omega_\pm, \qquad [u] = g_0 \quad \text{on } S, \quad [\partial_n u] = g_1 \quad \text{on } S \tag{2–1}$$

where $S \subseteq \Omega$, $\overline{\Omega}_- = \Omega_- \cup S$, $\overline{\Omega}_+ = \Omega_+ \cup S$; $f_\pm, u_\pm$ are defined on $\overline{\Omega}_\pm$; and g_0, g_1 are on S. To complete the problem we assume u is specified on $\partial\Omega$,

$$u = u_0 \quad \text{on } \partial\Omega, \tag{2–2}$$

although other boundary conditions are considered below. We assume that $f_\pm$, g_0, g_1, and S are fairly smooth, with a possible jump in $f_\pm$ at S. For appropriate u_0 on $\partial\Omega$, or for other boundary conditions, it follows that $u_\pm$ is smooth on $\overline{\Omega}_\pm$, as discussed below. In order to estimate truncation errors in difference schemes, we suppose for now that $u_\pm$ is C^4 on $\overline{\Omega}_\pm$ and also that S is C^4. It follows that each of $u_\pm$ has a C^4 extension to an open set containing S; this fact will be used to justify the corrections at the interface. Sufficient conditions for the regularity of u are given in Lemma 2.4.

To discuss discretization, we write the region Ω as

$$\Omega = \{x \in R^d : 0 < x_k < A_k,\ 1 \le k \le d\}. \tag{2–3}$$

For simplicity, we assume ratios of the lengths A_k are rational, so that the domain can be partitioned by grid cubes of size h for arbitrarily small h. We assume h is chosen so that $A_k = N_k h$ with integer N_k for each k. The computational domain is

$$\Omega_h = \{jh \in hZ^d : 1 \le j_k \le N_k - 1,\ 1 \le k \le d\}, \tag{2–4}$$

with boundary

$$\partial\Omega_h = \{jh : 0 \le j_k \le N_k,\ 1 \le k \le d\,;\, j_k = 0 \text{ or } N_k \text{ for some } k\}. \tag{2–5}$$

The closure is $\overline{\Omega}_h = \Omega_h \cup \partial\Omega_h$. We also need the partial boundary

$$\partial^0\Omega_h = \{jh : 0 \le j_k \le N_k - 1,\ 1 \le k \le d\,;\, j_k = 0 \text{ for some } k\}. \tag{2–6}$$

We use the usual second-order discrete Laplacian, defined for a function u^h on $\overline{\Omega}_h$ as

$$\Delta_h u^h = \sum_{k=1}^{d} D_k^- D_k^+ u, \tag{2–7}$$

where $D_k^\pm$ is the usual forward or backward difference operator in the k-th direction; for example, with $d = 2$,

$$D_1^+ u(j_1 h, j_2 h) = \big(u((j_1+1)h, j_2 h) - u(j_1 h, j_2 h)\big)/h.$$

We write $\nabla_h^\pm u$ for the discrete gradient whose components are $D_k^\pm$. We will use the discrete L^p norm and maximum norm,

$$\|u^h\|_{p,\Omega_h} = \Bigg(\sum_{jh\in\Omega_h} |u^h(jh)|^p h^d\Bigg)^{1/p}, \quad \|u^h\|_{\max,\Omega_h} = \max_{jh\in\Omega_h} |u^h(jh)|. \tag{2–8}$$

Now suppose the grid size h is chosen and each grid point $jh \in \overline{\Omega}_h$ is labeled as a point in $\overline{\Omega}_+$ or $\overline{\Omega}_-$; points lying on S can be assigned arbitrarily. We say a grid point is *regular* with respect to S if all grid points in the stencil of the discrete

Laplacian at that point are in the same closed region. Otherwise it is *irregular*. Let u^e be the exact solution of (2–1), (2–2). At each regular point we have the usual truncation error

$$\Delta_h u^e(jh) = f_\pm(jh) + \tau^h(jh), \qquad |\tau^h(jh)| \le Ch^2, \tag{2–9}$$

with $f_\pm$ chosen according to whether $jh \in \overline{\Omega}_+$ or $\overline{\Omega}_-$. This holds even if there are boundary points within h of jh, since the $u^e_\pm$ have smooth extensions independent of h; the usual Taylor expansion applies to the extended $u^e_\pm$, once h is small enough. Next we consider the error at the irregular points. Suppose we identify the leading terms in $\Delta_h u^e(jh)$, as is done in the methods under discussion, and explained further below see (2–23)–(2–26), with a first order error remaining. That is, we find $T^h(jh)$, determined by the jumps, so that

$$\Delta_h u^e(jh) = f_\pm(jh) + T^h(jh) + \tau^h(jh), \qquad |\tau^h(jh)| \le Ch. \tag{2–10}$$

(If $jh \in S$, $f_\pm$ is chosen to be consistent with the labeling of jh.) Now define f^h on Ω_h by

$$f^h(jh) = \begin{cases} f_\pm(jh) + T^h(jh), & jh \text{ irregular,} \\ f_\pm(jh), & jh \text{ regular.} \end{cases} \tag{2–11}$$

Finally, as in [32; 34; 24], we take u^h to be the solution of

$$\Delta_h u^h = f^h \quad \text{in } \Omega_h, \qquad u^h = u_0 \quad \text{on } \partial\Omega_h. \tag{2–12}$$

Then the error $u^h - u^e$ satisfies

$$\Delta_h(u^h - u^e) = -\tau^h \quad \text{in } \Omega_h, \qquad u^h - u^e = 0 \quad \text{on } \partial\Omega_h. \tag{2–13}$$

We can now state our main result. We assume that (2–9) and (2–10) hold, rather than making assumptions about the smoothness of the problem. After the theorem and related lemmas, we describe the assumptions which guarantee the needed smoothness and then review the derivation of (2–10). The theorem implies that the error in (2–13) is uniformly $O(h^2)$, with a similar estimate for the discrete gradient.

Theorem 2.1. *Let u^e be the exact solution of the problem* (2–1), (2–2) *with S at least C^1. Suppose $\Delta^h u^e$ has the form given by* (2–9), (2–10), *with $|\tau_h(jh)| \le Ch$ at irregular grid points and $|\tau_h(jh)| \le Ch^2$ at regular grid points. Let u^h be the solution of* (2–11), (2–12). *Then*

$$|u^h(jh) - u^e(jh)| \le C_0 h^2, \qquad jh \in \Omega_h \tag{2–14}$$

and for $1 \le \ell \le d$,

$$|D^+_\ell u^h(jh) - D^+_\ell u^e(jh)| \le C_1 h^2 \log(1/h), \qquad jh \in \Omega_h \cup \partial^0\Omega_h \tag{2–15}$$

with C_0, C_1 *dependent on* u^e *but independent of* h.

The discrete gradient estimate (2–15) can be interpreted as an estimate for

$$D_\ell^-(u^h - u^e)$$

at a slightly different set of points, and thus a similar estimate also holds for centered differences on Ω_h. An accurate approximation to ∇u^e can thus be found; see Corollary 2.5 and Equation (2–27) below.

Theorem 2.1 will follow directly from the next two lemmas, which are proved in Section 3.

Lemma 2.2. *Suppose* f^{irr} *is a function on* Ω_h *which is nonzero only on the set of irregular points. Assume* S *is* C^1. *Then there exist functions* F_k *on* $\Omega_h \cup \partial^0\Omega_h$, $1 \le k \le d$, *such that* $F_k = 0$ *on* $\partial^0\Omega_h$,

$$f^{\mathrm{irr}} = \sum_{k=1}^{d} D_k^- F_k \quad \text{in } \Omega_h \tag{2–16}$$

and

$$\|F_k\|_{\max,\Omega_h\cup\partial^0\Omega_h} \le Ch\|f^{\mathrm{irr}}\|_{\max,\Omega_h}, \quad 1 \le k \le d, \tag{2–17}$$

where C *depends on* S *but is independent of* h.

Lemma 2.3. *Suppose*

$$\Delta_h v = f^{\mathrm{reg}} + \sum_{k=1}^{d} D_k^- F_k \quad \text{in } \Omega_h, \qquad v = 0 \quad \text{on } \partial\Omega_h, \tag{2–18}$$

where

$$v : \overline{\Omega}_h \to R, \quad f^{\mathrm{reg}} : \Omega_h \to R, \tag{2–19}$$

$$F_k : \Omega_h \cup \partial^0\Omega_h \to R, \quad 1 \le k \le d \tag{2–20}$$

and $F_k(jh) = 0$ *for each* $jh \in \partial^0\Omega_h$ *with* $j_\ell = 0$ *for some* $\ell \ne k$. *Then*

$$\|v\|_{\max,\Omega_h} \le C_0\left(\|f^{\mathrm{reg}}\|_{2,\Omega_h} + \sum_{k=1}^{d}\|F_k\|_{\max,\Omega_h\cup\partial^0\Omega_h}\right), \tag{2–21}$$

$$\|D_\ell^+ v\|_{\max,\Omega_h\cup\partial^0\Omega_h} \le C_1 \log(1/h)\left(\|f^{\mathrm{reg}}\|_{\max,\Omega_h} + \sum_{k=1}^{d}\|F_k\|_{\max,\Omega_h\cup\partial^0\Omega_h}\right) \tag{2–22}$$

for $1 \le \ell \le d$, *where* C_0, C_1 *depend only on the lengths* A_k.

To derive the theorem, we set f^{irr} equal to the restriction of τ_h to the irregular points and use Lemma 2.2, concluding that $F_k = O(h \cdot h) = O(h^2)$. Then we apply Lemma 2.3 to $v = u^h - u^e$, using (2–13) with f^{reg} equal to the regular part of τ_h. The entire right side of (2–21) is $O(h^2)$, and similarly for (2–22). Theorem 2.1 and the lemmas also hold with periodic or Neumann boundary conditions, rather than Dirichlet, as discussed below. For the discrete Dirichlet problem (2–18), it is well known that the maximum of v can be estimated by the maximum of the right side, using the discrete maximum principle, but (2–21) is sharper in dependence on F_k.

In order to verify Equations (2–9), (2–10) we need general conditions on the problem (2–1), (2–2) to ensure the smoothness of $u_\pm$. An existence and regularity theorem for a general class of interface problems is given in [22], Section 16. The statement of higher regularity given below for the present case is based on potential theory and the classical Schauder estimates for elliptic equations. A brief justification is given in Section 3. This statement can be extended to the case with a discontinuous coefficient in the jump in normal derivative; see Section 4. We say that $f \in C^{m+\alpha}(\overline{\Omega})$, for integer m and $0 < \alpha < 1$, if $f \in C^m(\overline{\Omega})$ and $D^m f$ is uniformly Hölder continuous with exponent α on $\overline{\Omega}$.

Lemma 2.4. *Suppose $u_\pm$ in Equation* (2–1) *is the restriction to Ω of a solution to the extended problem in R^d. Suppose S is $C^{4+\alpha}$,*

$$f_- \in C^{2+\alpha}(\overline{\Omega}_-), \quad f_+ \in C^{2+\alpha}(R^d - \Omega_-), \quad g_0 \in C^{4+\alpha}(S), \quad \textit{and} \quad g_1 \in C^{3+\alpha}(S),$$

for some $0 < \alpha < 1$. *Then* $u_\pm \in C^{4+\alpha}(\overline{\Omega}_\pm)$.

We now describe the derivation of (2–10) as in Mayo's method [32; 34; 31], the related work of Wiegmann and Bube [37], or the immersed interface method of LeVeque and Li [24; 27; 26]. All these methods start with the observation that jumps in higher derivatives of $u_\pm$ in (2–1) can be found by differentiating the jumps in $u_\pm$, $\partial_n u_\pm$ along S and using $\Delta u_\pm = f_\pm$. To be specific, we emphasize Mayo's point of view. For dimension 2, writing $(x, y) \in R^2$, the jumps in first and second derivatives are

$$[u_x] = x' g_0' + y' g_1, \qquad [u_y] = y' g_0' - x' g_1, \tag{2–23}$$

$$[u_{xx}] = g_2 + y'^2 [f], \qquad [u_{yy}] = -g_2 + x'^2 [f], \tag{2–24}$$

where

$$g_2 \equiv 2\kappa x' y' g_0' + (x'^2 - y'^2)(g_0'' - \kappa g_1) + 2x' y' g_1', \tag{2–25}$$

and where primes denote arclength derivative d/ds along S and κ is the curvature $\kappa = x'' y' - x' y''$. These jump formulas, or equivalent ones, are used to find the corrections T_h at the irregular grid points. Suppose, for example, that

$$(j_1 h, j_2 h) \in \overline{\Omega}_- \quad \text{but} \quad ((j_1 + 1)h, j_2 h) \in \overline{\Omega}_+.$$

To correct $\Delta_h u(j_1 h, j_2 h)$, we find a point $((j_1+\theta)h, j_2 h) \in S,\ 0 \le \theta \le 1$. A Taylor expansion gives

$$\begin{aligned} u_+((j_1+1)h, j_2h) - u_-(j_1h, j_2h) &= hu_{-,x} + \tfrac{1}{2}h^2 u_{-,xx} \\ &\quad + [u] + (1-\theta)h[u_x] + \tfrac{1}{2}(1-\theta)^2 h^2 [u_{xx}] + O(h^3), \end{aligned} \tag{2–26}$$

where $u_{-,x}$,$u_{-,xx}$ are evaluated at (j_1h, j_2h) and the jumps are located at

$$((j_1+\theta)h, j_2h).$$

This expression is valid even if S intersects the segment at more than one point; the Taylor expansion for $u_\pm$ applies to the extended functions under the smoothness assumptions of Lemma 2.4. To approximate $\Delta_h u(j_1h, j_2h)$ we consider four such segments, finding jump terms if needed, add expressions similar to (2–26), and divide by h^2, to obtain an equation in the form (2–10), thus identifying $T^h(j_1h, j_2h)$. The procedure for the immersed interface method [24] is very similar, but for each irregular point (j_1h, j_2h), one nearby boundary point is chosen, and a Taylor expansion in (x, y) about this point is used for each of the points in the stencil. In either case the derivation of (2–9), (2–10) is justified, and Theorem 2.1 applies:

Corollary 2.5. *For the problem* (2–1), (2–2), *with the smoothness assumptions of Lemma 2.4*, *either Mayo's method* [32; 34] *or the immersed interface method of LeVeque and Li* [24], *with corrections of the form* (2–10), *gives a computed solution u_h with $|u_h - u_e| \le Ch^2$ uniformly. Moreover*, *∇u_e can be found on Ω_h from u_h with error uniformly $O(h^2 \log(1/h))$.*

It remains to verify the last statement of the corollary. For regular points the centered difference of u_h gives a value of ∇u_e accurate to $O(h^2 \log(1/h))$, according to (2–15). At irregular points we can correct the centered difference to the same order using formulas such as (2–26). For example, suppose (j_1h, j_2h) and $((j_1-1)h, j_2h)$ are in $\overline{\Omega}_-$ but $((j_1+1)h, j_2h) \in \overline{\Omega}_+$. We find, for the exact solution,

$$\begin{aligned} &u_+\big((j_1+1)h, j_2h\big) - u_-\big((j_1-1)h, j_2h\big) \\ &\quad = 2hu_{-,x}(j_1h, j_2h) + [u] + (1-\theta)h[u_x] + \tfrac{1}{2}(1-\theta)^2h^2[u_{xx}] + O(h^3). \end{aligned} \tag{2–27}$$

From this we obtain a computed value of ∇u which is again accurate to

$$O(h^2 \log(1/h)).$$

Similar results hold if we impose a boundary condition on $\partial\Omega$ other than (2–2). No change is needed if we use the homogeneous Dirichlet condition $u = 0$ on $\partial\Omega$, provided f is the restriction to Ω of an odd, periodic function, with period $2A_k$ in direction k, which is smooth except for the jump at S and its reflections.

(For example, this would be true if $f_+ = 0$ near $\partial\Omega$.) The solution is then smooth, since the problem extends to R^d with u odd and periodic. Alternatively, we could use periodic boundary conditions for u on $\overline{\Omega}$, if f extends smoothly to a periodic function with periods A_k. In this case we have the necessary condition

$$\int_{\Omega_-} f_- + \int_{\Omega_+} f_+ + \int_S [g_1]\, dS = 0 \tag{2–28}$$

and u has an arbitrary constant. Finally, we could impose the Neumann, or no-flux, condition

$$\partial_n u = 0 \quad \text{on } \partial\Omega, \tag{2–29}$$

again with condition (2–28), if f has a smooth, even, periodic extension. In this case we solve for u^h on $\overline{\Omega}_h$, with u^h extended past $\partial\Omega_h$ so that

$$u(-h, j_2 h) = u(h, j_2 h),$$

etc., consistent with (2–29). The exact and discrete Neumann problems both extend to even, periodic problems, and the analysis for the periodic case applies to this case as well.

We discuss the modifications of the analysis for the periodic boundary condition. We cannot solve $\Delta_h u^h = f^h$ exactly with u^h periodic; instead we solve

$$\Delta_h u^h = f^h - f_0^h, \tag{2–30}$$

where f_0^h is the mean value of f^h. Since $\Delta_h u^h$ has mean value zero, and the number of irregular points is $O(h^{-d+1})$, it follows from (2–9), (2–10) that

$$f_0^h = O(h^2),$$

so that this term does not affect the error estimate. Lemma 2.2 must be replaced by the version below. The proof of Lemma 2.3 is similar to the earlier case but simpler. The new term F_0 is treated in the theorem like the term f^{reg}.

Lemma 2.6. *Suppose f^{irr} is a function on Ω_h which is nonzero only on the set of irregular points. Assume S is C^1. Then there exist periodic functions F_k on $\Omega_h \cup \partial^0\Omega_h$, $0 \le k \le d$, so that $F_k = 0$ on $\partial^0\Omega_h$ for $1 \le k \le d$,*

$$f^{\text{irr}} = F_0 + \sum_{k=1}^{3} D_k^- F_k \quad \textit{in } \Omega_h \tag{2–31}$$

and

$$\|F_k\|_{\max,\Omega_h \cup \partial^0\Omega_h} \le Ch \|f^{\text{irr}}\|_{\max,\Omega_h}, \quad 0 \le k \le d \tag{2–32}$$

where C depends on S but is independent of h.

3. Proofs of the lemmas

Proof of Lemma 2.2. For simplicity, we assume dimension $d = 3$. We wish to work with pieces of S for which one spatial coordinate can be written as a function of the others. We can localize using a partition of unity (see, for example, [14], p. 13): Since S is C^1 and compact, there are finitely many open sets $U_i, V_i \subseteq \Omega$ and C^1 functions $\zeta_i \geq 0$ on Ω so that $\overline{V_i} \subseteq U_i$; the V_i cover S; each ζ_i is supported in V_i; $\sum_i \zeta_i(x) = 1$ for each x in an open neighborhood $\mathcal{N}$ of S; and for each i we can choose one coordinate, say x_3, so that the part of S in U_i consists of

$$S \cap U_i = \{(x_1, x_2, Z_i(x_1, x_2)) : (x_1, x_2) \in U_i'\}, \tag{3–1}$$

where U_i' is an open subset of R^2 and $Z : U_i' \to R$ is a C^1 function. Since the irregular points are within distance h of S, they are contained in $\mathcal{N}$ once h is small enough. For f^{irr} as specified, we can then write $f^{\text{irr}} = \sum_i \zeta_i f^{\text{irr}}$. It will suffice to prove the lemma for each $f^{(i)} = \zeta_i f^{\text{irr}}$.

Having localized the problem to considering $f^{(i)}$ on V_i, we first estimate the number of irregular points in V_i with given projection on U_i'. Let V_i' be the projection of V_i on U_i'. Suppose $x' = jh = (j_1 h, j_2 h) \in V_i'$. If $p = (x', z) \in V_i$ is an irregular point, then there is some $q \in S$ with $|q - p| \leq h$, say $q = (x'', z'')$ with $x'' \in U_i'$. Then $|x'' - x'| \leq h$, $|z'' - z| \leq h$, and $z'' = Z_i(x'')$. If M is a bound for $|\nabla Z_i|$, then $|Z_i(x'') - Z_i(x')| \leq Mh$, and

$$|z - Z_i(x')| \leq |z - z''| + |Z_i(x'') - Z_i(x')| \leq (1 + M)h. \tag{3–2}$$

Thus z is restricted to an interval of length $2(M + 1)h$, and the number of irregular points in V_i projecting onto x' is at most $C_1 \equiv 2M + 3$, a number bounded independent of $x' = jh$.

We will write $f^{(i)}$ as $D_3^- F^{(i)}$ for some $F^{(i)}$. We set $F^{(i)} = 0$ on $\partial^0 \Omega_h$, and for $(jh, kh) = (j_1 h, j_2 h, kh) \in \Omega_h$ we define

$$F^{(i)}(jh, kh) = \sum_{\ell=1}^{k} f^{(i)}(jh, \ell h)\, h. \tag{3–3}$$

Then, since kh is the third coordinate,

$$D_3^- F^{(i)} = f^{(i)} \quad \text{in } \Omega_h. \tag{3–4}$$

The function $f^{(i)}$ can only be nonzero at irregular points, and as noted above, the number of such points contributing to the sum (3–3) has a uniform upper bound. The estimate (2–17) for $F^{(i)}$ follows, and the proof is completed by summing over i. □

In proving Lemma 2.3 we will use a discrete Green's function G_h on hZ^d, satisfying

$$\Delta_h G_h(x) = \delta_h(x), \quad x \in hZ^d, \tag{3–5}$$

where $\delta_h(x) = h^{-d}$ for $x = 0$ and $\delta_h(x) = 0$ for $x \neq 0$. For $d = 2$ or 3 such G_h exists, with pointwise estimates analogous to those for the fundamental solution of the exact Laplacian,

$$|G_h(x)| \leq C_{00} + C_0 |\log(|x| + h)|, \quad d = 2, \tag{3–6}$$

$$|G_h(x)| \leq C_0(|x| + h)^{-1}, \qquad d = 3, \tag{3–7}$$

and for the first and second differences in directions k or ℓ, $1 \leq k, \ell \leq d$,

$$|D_k^+ G_h(x)| \leq C_1(|x| + h)^{1-d}, \qquad d = 2, 3, \tag{3–8}$$

$$|D_\ell^+ D_k^+ G_h(x)| \leq C_2(|x| + h)^{-d}, \qquad d = 2, 3. \tag{3–9}$$

For example, for $h = 1$, G_1 is introduced in [23] in terms of the expected number of visits to x by a random walk on Z^d starting at 0. The estimates (3–6)–(3–9) follow from those in [23], (pp. 32, 40), after rescaling G_1 to G_h. (For $d = 2$, G_h must also be adjusted by a constant. For second differences, [23] gives an estimate for a repeated difference in any direction, but $D_k^+ D_\ell^+$ can be reduced to this case by writing, with $h = 1$,

$$(S_k - I)(S_\ell - I) = \tfrac{1}{2}\big((S_k - I)^2 + (S_\ell - I)^2 - S_\ell^2 (S_k S_\ell^{-1} - I)^2\big) \tag{3–10}$$

where S_k is the forward shift in direction k.) If w is a function on hZ^d supported in a bounded set, then

$$w(x) = \sum_{y \in hZ^d} G_h(x - y)(\Delta_h w)(y)\, h^d. \tag{3–11}$$

This follows from (3–5) and the uniqueness of solution of the discrete Poisson problem.

We will need estimates for norms of G_h, $D_k^+ G_h$, and $D_\ell^- D_k^+ G_h$ which follow directly from the pointwise estimates (3–6)–(3–9). With $B_h(R) = \{x \in hZ^d : |x| < R\}$, we have

$$\|G_h\|_{2,B_h(R)} \leq C_0(R), \quad \|D_k^+ G_h\|_{1,B_h(R)} \leq C_1(R), \tag{3–12}$$

$$\|D_\ell^+ D_k^+ G_h\|_{1,B_h(R)} \leq C_2(R) \log(1/h), \tag{3–13}$$

with constants depending on R. Discrete Green's functions for more general elliptic operators and domains were constructed by Bramble et al. [7] and pointwise estimates for G_h were found using the maximum principle [8].

Proof of Lemma 2.3. First we check that

$$\|\nabla_h^+ v\|_{2,\Omega_h \cup \partial^0 \Omega_h} \le C\big(\|f^{\mathrm{reg}}\|_{2,\Omega_h} + \sum_k \|F_k\|_{2,\Omega_h}\big). \tag{3–14}$$

To show this, we multiply by v in (2–18), sum over Ω_h, and then sum by parts on the left and in the F_k terms, using the boundary conditions for v and F_k, to obtain

$$\langle \nabla_h^+ v, \nabla_h^+ v\rangle_{\Omega_h \cup \partial^0 \Omega_h} = -\langle f^{\mathrm{reg}}, v\rangle_{\Omega_h} + \sum_k \langle F_k, D_k^+ v\rangle_{\Omega_h}, \tag{3–15}$$

where brackets denote the usual discrete inner product, for example,

$$\langle v, w\rangle_{\Omega_h} = \sum_{jh \in \Omega_h} v(jh) w(jh) h^d, \qquad \|v\|_{2,\Omega_h} = \langle v, v\rangle_{\Omega_h}^{1/2}. \tag{3–16}$$

We can then derive (3–15) from the Cauchy–Schwarz inequality and the discrete Poincaré inequality, valid since $v = 0$ on $\partial\Omega$,

$$\|v\|_{2,\Omega_h} \le \|\nabla_h^+ v\|_{2,\Omega_h \cup \partial^0 \Omega_h}. \tag{3–17}$$

Next we extend the Poisson equation from Ω_h to hZ^d. Let $\tilde{f}$ be the odd, periodic extension of f^{reg}, with period $2N_k h$ in direction k, with $\tilde{f} = 0$ on the faces

$$j_k h = 0, N_k h$$

and their images. Let $\tilde{\phi}_k$ be the similar odd periodic extension of $D_k^- F_k$, and $\tilde{v}$ the odd periodic extension of v. Then

$$\Delta_h \tilde{v} = \tilde{f} + \sum_k \tilde{\phi}_k \quad \text{in } hZ^d. \tag{3–18}$$

We want to write $\tilde{\phi}_k$ as D_k^- of some extension $\tilde{F}_k$ of F_k. For example, if $k = 1$ and $d = 3$, for $1 \le j_1 \le N_1$ and $0 \le j_k \le N_k - 1$, $k = 2, 3$, we define

$$\tilde{F}_1(-j_1 h, j_2 h, j_3 h) = F_1((j_1 - 1)h, j_2 h, j_3)h.$$

We then extend $\tilde{F}_1$ to all $j_1 h$, with period $2N_1 h$. Finally we extend $\tilde{F}_1$ to be odd and periodic in $j_2 h$, $j_3 h$, with $\tilde{F}_1(j_1 h, j_2 h, j_3 h) = 0$ if $j_k h$ is a multiple of $N_k h$ for $k = 2$ or 3. With this definition, and a similar one for each $\tilde{F}_k$, we have

$$\tilde{\phi}_k = D_k^- \tilde{F}_k \quad \text{in } hZ^d. \tag{3–19}$$

We can now derive the maximum estimate for v. Choose a smooth function $\zeta : R^d \to [0, 1]$ with $\zeta(x) = 1$ for an open set containing $\overline{\Omega}$ and $\zeta = 0$ outside a bounded set B. Then

$$\Delta_h(\zeta \tilde{v}) = \zeta \tilde{f} + \zeta \nabla_h^- \cdot \tilde{F} - \nabla_h^{\pm} \zeta \cdot \nabla_h^{\pm} \tilde{v} - (\Delta_h \zeta)\tilde{v} \quad \text{in } hZ^d, \tag{3–20}$$

where $\tilde{F}$ is the vector with components $\tilde{F}_k$ and ± indicates two terms. We use the discrete Green's function G_h to write, for $x \in \Omega_h$,

$$v(x) = T_1 + T_2 + T_3 + T_4, \tag{3–21}$$

where

$$T_1 = \sum_{y \in hZ^d} G_h(x-y)\zeta(y)\tilde{f}(y)\, h^d, \quad T_2 = \sum_{y \in hZ^d} G_h(x-y)\zeta(y)\nabla_h^- \cdot \tilde{F}(y)\, h^d$$

or, after summation by parts,

$$T_2 = \sum_{y \in hZ^d} (\nabla_h^+ G_h)(x-y)\zeta(y)\tilde{F}(y)\, h^d - \sum_{y \in hZ^d} G_h(x-y)(\nabla_h^+ \zeta)(y)\tilde{F}(y)\, h^d \tag{3–22}$$

and similarly T_3, T_4 are discrete convolutions of G_h with $\nabla_h^\pm \zeta \cdot \nabla_h^\pm \tilde{v}$ and $(\Delta_h \zeta)\tilde{v}$.

To estimate these terms, let $\tilde{B} \subseteq R^d$ be a bounded set which contains all points $x - y$ with $x \in \overline{\Omega}$ and $y \in B$, and let $B_h = B \cap hZ^d$, $\tilde{B}_h = \tilde{B} \cap hZ^d$. Then for each $x \in \Omega_h$,

$$\begin{aligned} |T_1| &\le \|G_h\|_{2,\tilde{B}_h} \|\tilde{f}\|_{2,B_h}, \\ |T_2| &\le \big(\|\nabla_h^+ G_h\|_{1,\tilde{B}_h} + C_2 \|G_h\|_{2,\tilde{B}_h}\big) \end{aligned} \tag{3–23}$$

and

$$\begin{aligned} |T_3| &\le C_3 \|G_h\|_{2,\tilde{B}_h} \big(\|\nabla_h^+ \tilde{v}\|_{2,B_h} + \|\nabla_h^- \tilde{v}\|_{2,B_h}\big), \\ |T_4| &\le C_4 \|G_h\|_{2,\tilde{B}_h} \|\tilde{v}\|_{2,B_h}. \end{aligned} \tag{3–24}$$

The extension of f and F was such that

$$\|\tilde{f}\|_{2,B_h} \le C \|f^{\text{reg}}\|_{2,\Omega_h}, \quad \|\tilde{F}\|_{\max,B_h} \le C \|F\|_{\max,\Omega_h \cup \partial^0 \Omega_h} \tag{3–25}$$

and using (3–12) we get

$$|T_1| + |T_2| \le C \left(\|f^{\text{reg}}\|_{2,\Omega_h} + \|F\|_{\max,\Omega_h \cup \partial^0 \Omega_h}\right). \tag{3–26}$$

Also v was extended so that

$$\|\tilde{v}\|_{2,B_h} \le C \|v\|_{2,\Omega_h}, \qquad \|\nabla_h^\pm \tilde{v}\|_{2,B_h} \le C \|\nabla_h^+ v\|_{2,\Omega_h \cup \partial^0 \Omega_h}. \tag{3–27}$$

Combining this with (3–24), (3–14), (3–17), and (3–12), we see that T_3, T_4 have the same estimate as in (3–26), and (2–21) is now established.

The proof of (2–22) is very similar. We apply D_ℓ^+ to (3–21) with T_2 in the form (3–22); in each term D_ℓ^- acts on the x-variable in G_h. In T_3 and T_4, $D_\ell^+ G_h$ is uniformly bounded for $x \in \Omega_h$ since the support of $\nabla_h^\pm \zeta$ is away from Ω_h. □

Proof of Lemma 2.4. We first reduce to the case $f_\pm = 0$, as in [18]. From the Schauder regularity theory, the presumed solution u_+ is $C^{4+\alpha}$ away from S. Using this fact and the Schauder theory, we see that there exists v_+ in $C^{4+\alpha}(R^d - \Omega_-)$ such that $\Delta v_+ = f_+$ and $v_+ = 0$ on S, and there exists v_- in $C^{4+\alpha}(\overline{\Omega}_-)$ such that $\Delta v_- = f_-$ and $v_- = 0$ on S. Subtracting $v_\pm$, we now consider the reduced problem with $f_\pm = 0$. We can write a solution as the sum of a double layer potential and a single layer, with strengths g_0 and g_1 respectively. The double layer potential has boundary values on each side of S in $C^{4+\alpha}$, the same as for g_0, and it follows from the Schauder theory that it has the desired regularity in $\overline{\Omega}_\pm$. A similar remark applies to the Neumann boundary condition for the single layer potential. This solution may not be the same as $u_\pm$, since we have not imposed a condition at infinity, but the difference is harmonic throughout and therefore is smooth. □

Proof of Lemma 2.6. We proceed as in the proof of Lemma 2.2, but in place of (3–3), we set $F^{(i)}(jh, 0) = 0$ and

$$F^{(i)}(jh, kh) = \sum_{\ell=1}^{k} \bigl(f^{(i)}(jh, \ell h) - F_0^{(i)}(jh)\bigr)\, h, \quad 1 \le k \le N_3, \qquad (3\text{–}28)$$

where $F_0^{(i)}(jh)$ is the average of $f_0^{(i)}(jh, \ell h)$ over ℓ, that is,

$$F_0^{(i)}(jh) = A_3^{-1} \sum_{\ell=0}^{N_3-1} f^{(i)}(jh, \ell h)\, h \qquad (3\text{–}29)$$

for $jh \in V_i'$ and $F_0^{(i)}(jh) = 0$ otherwise. Then $F^{(i)}(jh, N_3 h) = 0$, so that $F^{(i)}$ extends periodically, and

$$f^{(i)} = D_3^- F^{(i)} + F_0^{(i)}. \qquad (3\text{–}30)$$

For each j, there are at most $C_1 = O(1)$ terms in the sum (3–28), and thus

$$\|F_0^{(i)}\|_{\max} \le A_3^{-1} C_1 h \|f^{(i)}\|_{\max} \le A_3^{-1} C_1 h \|f^{\mathrm{irr}}\|_{\max}. \qquad (3\text{–}31)$$

Then (2–17) holds for $F^{(i)}$, as defined in (3–28). Finally, we sum over i. □

4. Applications and extensions

Piecewise constant coefficients. In Section 2 we treated the problem (1–1), (1–2) in the special case $\beta_+ = \beta_-$. We now return to the problem where β_+, β_- are unequal, positive constants, perhaps representing different material properties. The important change is that $[\partial_n u]$ is not known, although $[\beta \partial_n u]$ is known. One possible approach is to enlarge the system of equations for the discretized elliptic system ([37; 27]). Here we use a different strategy, assuming the problem is in free space: We first solve an integral equation on S for the unknown $[\partial_n u]$, based on an integral representation for the solution, thus reducing the problem to the earlier case. A similar strategy is used below for Stokes flow with two fluids, using such a representation, as described, for example, in [36].

Suppose the problem (1–1), (1–2) is the restriction to Ω of a problem in R^2 in which $f_+ = 0$ outside Ω and $u \to 0$ at infinity. We will assume u is continuous across S, that is, $g_0 = 0$ in (1–2), but $[\beta\partial_n u] = g_1$ may be nonzero. The extra step of solving for $[\partial_n u]$ is needed even if $g_1 = 0$. The unknown u can be thought of as a weak solution of

$$\nabla\cdot(\beta\nabla u) \;=\; f \,+\, g_1\delta_S, \tag{4–1}$$

where δ_S is the measure that restricts to S. A recent analytical treatment of such problems can be found in [18]. The solution has the form

$$u(x) \;=\; \int_\Omega G(x-y)\frac{f(y)}{\beta(y)}\,dy \;+\; \int_S G(x-y)q(y)\,ds(y) \tag{4–2}$$

for some q defined on S, where $G(x) = (2\pi)^{-1}\log|x|$, $\beta(y) = \beta_\pm$ for $y \in \Omega_\pm$ and $f = f_\pm$. The last term is a single layer potential with strength q, to be determined. From potential theory we have an expression for the normal derivative of $u_\pm$ at S:

$$\partial_n u_\pm(x) = \\ \int_\Omega \partial_{n(x)}G(x-y)\frac{f(y)}{\beta(y)}\,dy + \int_S \partial_{n(x)}G(x-y)q(y)\,ds(y) \pm \tfrac{1}{2}\,q(x). \tag{4–3}$$

Here

$$\partial_n(x)G(x-y) \;=\; n(x)\cdot\nabla G(x-y), \quad \nabla G(x-y) \;=\; \frac{x-y}{2\pi|x-y|^2}. \tag{4–4}$$

Subtracting, we see that

$$[\partial_n u(x)] \;=\; q(x) \tag{4–5}$$

so that, once q is known, we have reduced the problem to one of the earlier type for the unknown u. To find q we multiply (4–3) by $\beta_\pm$, subtract, and use the second condition in (1–2), obtaining the integral equation

$$\tfrac{1}{2}(\beta_+ + \beta_-)q \;+\; (\beta_+ - \beta_-)\int_S(\partial_n G)q\,ds \;= \\ g_1 \;-\; (\beta_+ - \beta_-)\int_\Omega(\partial_n G)\,(f/\beta)\,dy \tag{4–6}$$

([36], Section 5.3). The equation has a unique solution since

$$|(\beta_+ + \beta_-)/(\beta_+ - \beta_-)| > 1$$

(see, for example, [36], Section 5.4). In this two-dimensional case, $\partial_n(x)G(x-y)$ is smooth for $x, y \in S$, whereas the second integrand in (4–2) has an integrable singularity. If $f_+ = 0$ near $\partial\Omega$, u can be found on $\partial\Omega$ from (4–2) in a routine way, and the solution u can be found as in Section 2 using (4–5). We see from (2–23)–(2–26) that we need to solve for q with accuracy $O(h^2)$ in order to obtain $O(h)$ truncation error near S, and thereby $O(h^2)$ accuracy for the solution u, according to the theory of Section 2. The solution of (4–6) is discussed further in Section 5; see (5–10)–(5–11).

Higher order accuracy. In principle, the theory of Sections 2 and 3 can be applied to higher order methods. In dimension $d = 2$, the nine-point Laplacian (see, for example, [19], Section 7.3) has truncation error $O(h^2)$ proportional to the Laplacian, with remaining error $O(h^4)$. The right-hand side of the discrete Poisson equation can be modified so that the truncation error is $O(h^4)$. In this way the methods outlined in Section 2 can then be improved so that the error in the solution is uniformly $O(h^4)$. The jump conditions of (2–23)-(2–26) can be carried to the fourth derivatives so that the truncation error τ^h in (2–10) remaining at the irregular points after correction is $O(h^3)$, while at the regular points the truncation error is $O(h^4)$. The analogue of Lemma 2.3 holds for the nine-point Laplacian; the estimates (3–6), (3–8), (3–9) apply to the discrete Green's function for this operator, as can be seen from Theorem 2 in [16], and thus (3–12), (3–13) hold as well. It then follows from Lemma 2.2 and the modified version of Lemma 2.3 that the conclusion of Theorem 2.1 holds, with h^4 in place of h^2 in the estimates (2–14), (2–15). Fourth order methods of this type have been given in [30] and [20].

Nearly singular integrals. Mayo [33] suggested a procedure to solve a problem for a harmonic function with prescribed jumps, such as (2.1) with $f_\pm = 0$, distinct from the approach of [32]. The first step is to write the solution as a layer potential and calculate it at grid points near the interface, directly as a nearly singular integral. The discrete Laplacian is then formed at the irregular points from these values and extended to be zero at regular points. Finally, a fast Poisson solver is used to find the solution at all grid points. Mayo was able to solve a boundary value problem in this manner by regarding the boundary as an interface and solving an integral equation for the strength of the layer potential. Beale and Lai [2] developed a method for computing nearly singular integrals and used the approach of [33] to solve Dirichlet problems ([2], Section 4). An error estimate for the solution resulting from this procedure is justified by the theory of Section 2: Suppose we find the solution at the irregular points, accurate to $O(h^3)$, as was essentially done in [2], Section 4. The discrete Laplacian formed from these values at the irregular grid points near the interface is accurate at least to $O(h)$. The discrete Laplacian is set to zero elsewhere, with truncation error $O(h^2)$. Thus it follows from Theorem

2.1 that the computed solution is uniformly accurate to $O(h^2)$. The estimate for the gradient applies as well. Mayo gave an argument for a similar conclusion in the Appendix of [33].

Stokes flow. The methods studied here have been used to solve the Stokes equations, describing creeping flow of a very viscous fluid, with an interface separating regions. The immersed interface method of LeVeque and Li was applied to problems with moving interfaces in one fluid in a periodic region [25]. Here we discuss a very similar method for the steady problem in free space. We emphasize the implications of the present results for the error estimates. We see that choices for corrections as in [25] lead to uniform $O(h^2)$ accuracy for both pressure and velocity. We first discuss the case of 2D flow with one fluid, and then explain how the procedure can be extended to the two-fluid case by first solving an integral equation on the interface.

We write the problem as

$$-\mu\Delta v + \nabla p = f\delta_S, \quad \nabla\cdot v = 0, \tag{4–7}$$

where $v = (v_1, v_2)$ is the fluid velocity, μ is the viscosity, p is the pressure, and $f = (f_1, f_2)$ is a specified force on the interface S. The associated stress tensor is

$$\sigma_{ij} = -p\delta_{ij} + \mu\Big(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\Big), \tag{4–8}$$

with $i, j = 1, 2$. We assume for now that μ has the same value on both sides, but later we consider the case of different viscosities. We always assume the velocity is continuous across S. The delta function terms in the Stokes equations amount to a jump condition on σ (see [35]),

$$[\sigma_{ij}]n_j = -f_i, \qquad i = 1, 2 \tag{4–9}$$

with sum over j understood. From the jump in stress we obtain jump conditions for p and $\partial v/\partial n$ [35; 25]

$$[p] = f\cdot n, \qquad \mu\left[\frac{\partial v}{\partial n}\right] = -(f\cdot\tau)\tau. \tag{4–10}$$

We also need the jump condition for $\partial p/\partial n$, derived in [25],

$$\left[\frac{\partial p}{\partial n}\right] = \frac{\partial(f\cdot\tau)}{\partial s}, \tag{4–11}$$

where s is the arclength parameter on S.

To solve (4–7) we proceed in steps, as in [25], solving first for p and then for v. We choose a computational rectangle Ω containing S and use a square grid as before. We solve the free space problem, assuming decay at infinity. On the computational boundary we prescribe the exact solution, which is known in integral form (for example, see [36; 13]). The pressure is

$$p(x) = \int_S \nabla G(x-y)\cdot f(y)\,ds(y), \tag{4–12}$$

with ∇G as in (4–4). The velocity is

$$v_i(x) = \frac{1}{\mu}\int_S V_{ij}(x-y) f_j(y)\,ds(y), \tag{4–13}$$

$$V_{ij}(x) = -\frac{\delta_{ij}}{4\pi}\log|x| + \frac{x_i x_j}{4\pi|x|^2}. \tag{4–14}$$

The pressure p is determined by a problem of the form (2–1), with $\Delta p = 0$ in $\Omega_\pm$ and jump conditions for $p, \partial p/\partial n$ given in (4–10), (4–11). We solve for p using the procedure of Section 2, adding corrections to the discrete Laplacian. In this way we obtain a solution p^h with error uniformly $O(h^2)$. Next we solve for the velocity components v_1^h, v_2^h. For the exact v we have $\mu\Delta v = \nabla p$ in $\Omega_\pm$. We find a computed velocity v^h as the solution of

$$\mu\Delta_h v^h = \nabla^h p^h + T^h, \tag{4–15}$$

where ∇^h is the centered difference operator for ∇. In T^h we include correction terms to account for the jumps in $\partial v/\partial n$ and in ∇p, given in (4–10), (4–11), as well as corrections for the difference approximation to ∇p, as in (2–27). According to Theorem 2.1, the resulting v^h would be uniformly second-order accurate if p^h were exact. However, since $p^h - p^e = O(h^2)$ uniformly, the error on the right-hand side of the form $\nabla^h(p^h - p^e)$ contributes an error to the solution which is uniformly $O(h^2)$, according to Lemma 2.3. (As noted in [25], we only need correct the difference $\nabla^h p$ to $O(h)$ near S to obtain $O(h^2)$ accuracy for the velocity.)

Stokes flow with two fluids. Next we consider the case of two different viscosities, $\mu_\pm$. With the Stokes equations (4–7) otherwise the same, we have the same jump condition (4–9) for the normal stress. The solution can be written in integral form, derived in [36], Section 5.3:

$$p(x) = \mu_\pm \int_S \nabla G(x-y)\cdot q(y)\,ds(y), \tag{4–16}$$

$$v_i(x) = \int_S V_{ij}(x-y) q_j(y)\,ds(y). \tag{4–17}$$

Here $q = (q_1, q_2)$ is a function on S which solves the integral equation

$$\tfrac{1}{2}q_i(x) = \alpha n_k(x)\int_S T_{ijk}(x-y) q_j(y)\,ds(y) + \beta f_i(x), \tag{4–18}$$

with

$$T_{ijk} = -\frac{x_i x_j x_k}{\pi |x|^4}, \qquad \alpha = \frac{\mu_+ - \mu_-}{\mu_+ + \mu_-}, \qquad \beta = \frac{1}{\mu_+ + \mu_-}. \tag{4–19}$$

(See equation (5.3.9) in [36], noting the factors of 4π should be replaced by 2π in the two-dimensional case.) To solve the flow problem, we begin by solving this integral equation for q. The solvability is discussed in [36], Section 5.4. The kernel $n_k T_{ijk}$ is smooth on S, and the integrals can be computed in a standard way. After solving for q, we can think of v, $p/\mu_\pm$ on $\Omega_\pm$ as the solution of the Stokes equations with $\mu_\pm$ replaced by 1 and with

$$[\sigma_{ij}^{(1)}] n_j = -q_i, \tag{4–20}$$

where

$$\sigma_{ij}^{(1)} = -\frac{p}{\mu_\pm} \delta_{ij} + \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right). \tag{4–21}$$

It follows that v, $p/\mu_\pm$ have jumps as in (4–10), (4–11) but with f replaced by q. Once these jumps are known, we can solve for $p/\mu_\pm$ and then v as in the earlier one-fluid case. In view of (2–23)–(2–26), we need to find q to accuracy $O(h^3)$ to obtain an $O(h)$ truncation error near S in the problem for p of the form (4–10), in order to solve for p, and then v, with $O(h^2)$ accuracy. It is not difficult to solve the integral equation (4–18) to this accuracy provided S and f are smooth enough.

5. Numerical examples

Interface problem with $\beta_+ = \beta_-$. In the first set of examples, we consider the interface problem with $\beta_\pm = 1$:

$$\Delta u_\pm = f_\pm \text{ in } \Omega_\pm, \tag{5–1}$$

where the interface S is given by the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \tag{5–2}$$

and $\Omega = [-1.1, 1.1] \times [-1.1, 1.1]$.

The first example we consider has a solution given by

$$u_- = \sin x \cos y, \; u_+ = 0. \tag{5–3}$$

With $u_\pm$ specified, $f_\pm$ in (5–1) and the jump conditions g_0 and g_1 can be determined. Two choices for the semi-axes of the ellipse were used, first, $(a, b) = (0.7, 0.9)$, and then $(a, b) = (0.9, 0.1)$. In the latter, the curvature $\kappa = -90$ at $(\pm a, 0)$, leading to a more severe test. The solution and its gradient were computed using the technique of Mayo [32] and using the immersed interface method of LeVeque and Li [24]. Results for the two ellipses are reported in Table 1. Solutions were obtained for $N = 40$, 80, 160, 320, and 640, where N denotes half of the number of subintervals in each dimension. Normalized errors in the L^r-norm, defined as $\|u^h - u^e\|_r / \|u^h\|_r$, are shown for $r = 2$ and ∞. These results show $O(h^2)$

N	u		u_x		u_y	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
Mayo’s technique						
40	5.106E-5	3.451E-5	1.448E-4	1.370E-4	2.120E-4	1.436E-4
80	1.415E-5	1.045E-5	3.565E-5	3.275E-5	5.475E-5	3.594E-5
160	3.416E-6	2.458E-6	8.980E-6	8.269E-6	1.336E-5	8.965E-6
320	8.217E-7	5.728E-7	2.262E-6	2.089E-6	3.269E-6	2.352E-6
640	1.180E-7	8.756E-8	4.189E-7	3.945E-7	5.532E-7	4.186E-7
Immersed interface method						
40	2.345E-5	1.773E-5	1.107E-4	1.035E-4	1.362E-4	1.302E-4
80	2.748E-5	2.632E-5	2.748E-5	2.632E-5	3.613E-5	3.510E-5
160	1.510E-6	1.139E-6	6.912E-6	6.656E-6	8.856E-6	8.989E-6
320	3.722E-7	2.805E-7	1.731E-6	1.664E-6	2.199E-6	2.323E-6
640	9.732E-8	7.645E-8	4.305E-7	4.166E-7	5.184E-7	5.967E-7

N	u		u_x		u_y	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
Mayo’s technique						
20	2.377E-5	1.209E-5	4.926E-4	4.900E-4	1.506E-3	9.577E-4
40	6.020E-6	2.730E-6	1.231E-4	1.219E-4	3.637E-4	2.269E-4
80	2.261E-6	1.003E-6	3.057E-5	3.001E-5	1.424E-4	1.376E-4
160	5.730E-7	2.340E-7	7.637E-6	7.532E-6	3.649E-5	3.331E-5
320	1.297E-7	5.360E-8	1.915E-6	1.951E-6	8.189E-6	8.309E-6
Immersed interface method						
20	7.134E-4	6.933E-4	1.601E-3	3.051E-3	6.267E-2	7.441E-2
40	1.311E-5	7.441E-6	1.221E-4	1.611E-4	9.386E-4	1.559E-3
80	4.750E-6	2.043E-6	2.993E-5	3.538E-5	3.019E-4	2.047E-4
160	1.210E-6	5.156E-7	7.468E-6	7.949E-6	7.696E-5	5.145E-5
320	2.972E-7	1.249E-7	1.869E-6	2.036E-6	1.881E-5	1.277E-5

Table 1. Results for interface problem with $\beta = 1$, example (5–3). Normalized errors in computed solution and first derivatives. Top: $a = 0.7$, $b = 0.9$; bottom: $a = 0.9$, $b = 0.1$.

convergence in the solution u, consistent with Theorem 2.1. In Section 2, we proved that ∇u can be approximated from u^h with error uniformly $O(h^2 \log(1/h))$. However, results in both tables show $O(h^2)$ accuracy in ∇u^h.

We then consider a second example where the solution is given by

$$u_- = x^9 y^8, \quad u_+ = 0. \tag{5–4}$$

This example is constructed such that the solution has large high-order derivatives. In particular, $|\partial^3 u/\partial x^3|$ and $|\partial^3 u/\partial y^3|$, which occur in the lowest-order uncorrected terms in both Mayo's method and the immersed interface method, are large. Tables 2 and 3 show normalized errors in the solution and its gradient. Results in Table 2, computed for the ellipse $(a, b) = (0.7, 0.9)$, show $O(h^2)$ convergence, although the magnitude of the errors is larger in this example. In particular, as in the previous example, $O(h^2)$ accuracy was obtained for ∇u^h.

The ellipse $(a, b) = (0.9, 0.1)$ used in the next example has large curvature $|\kappa| \le 90$, compared to $|\kappa| < 1.84$ in the previous example. As shown in Table 3, the computed solution has large errors, compared to all previous examples. In particular, solution errors are $> 100\%$ for $N = 40$. These large errors can be attributed to the $O(h^3)$ error terms neglected in (2–26) by Mayo's technique and in the analogous expression by the immersed interface method. The magnitude of these $O(h^3)$ error terms is proportional to κ and to $\nabla^3 u$ — both of which are large in this example by

N	u		u_x		u_y	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
Mayo's technique						
40	2.370E-2	6.401E-3	2.489E-2	1.994E-2	1.340E-2	1.046E-2
80	7.469E-3	2.764E-3	5.895E-3	5.520E-3	3.249E-3	2.385E-3
160	1.362E-3	4.802E-4	1.508E-3	1.411E-3	7.825E-4	6.834E-4
320	3.232E-4	9.426E-5	3.858E-4	3.532E-4	1.985E-4	1.768E-4
640	8.363E-5	2.952E-5	9.389E-5	9.304E-5	4.839E-5	4.211E-5
Immersed interface method						
40	2.140E-2	5.699E-3	2.557E-2	2.096E-2	1.390E-2	1.144E-2
80	6.963E-3	2.478E-3	6.139E-3	5.840E-3	3.546E-3	2.695E-3
160	1.236E-3	3.719E-4	1.565E-3	1.489E-3	8.571E-4	7.456E-4
320	2.815E-4	7.053E-5	3.992E-4	3.749E-4	2.170E-4	1.931E-4
640	7.858E-5	2.502E-5	9.742E-5	9.795E-5	5.365E-5	4.618E-5

Table 2. Results for interface problem with $\beta = 1$, example (5–4). Normalized errors in computed solution and first derivatives, obtained for $a = 0.7$, $b = 0.9$.

construction. Thus, for a sufficiently coarse grid, these uncorrected $O(h^3)$ error terms result in large solution errors. Nonetheless, for sufficiently large N, the approximations show $O(h^2)$ convergence, although the error magnitude remains large.

Interface problem with piecewise-constant $\boldsymbol{\beta}$. Next we consider the problem of an interface with piecewise-constant coefficients $\beta_\pm$

$$\beta_\pm \Delta u_\pm = f_\pm \quad \text{in } \Omega_\pm, \tag{5–5}$$

$$[u] = 0, \qquad [\beta \partial_n u] = g_1, \tag{5–6}$$

where the interface S is given by an ellipse (5–2) with $a = 0.9$ and $b = 0.7$, and $\Omega = [-1.3, 1.3] \times [-1.3, 1.3]$. The solution is given in elliptic coordinates to be

$$\begin{aligned} u_- &= a_0^3 \left(\cosh^2 \rho \sinh \rho \cos^2 \theta \sin \theta + \sinh^3 \rho \sin^3 \theta\right), \\ u_+ &= ce^{-3\rho} \sin 3\theta + de^{-\rho} \sin \theta, \end{aligned} \tag{5–7}$$

where $\rho \in [0, \infty)$ and $\theta \in [0, 2\pi]$; ρ and θ are defined by the conformal mapping

$$x + \imath y = a \cosh(\rho + \imath\theta) \tag{5–8}$$

such that

$$x = a_0 \cosh \rho \cos \theta,\ y = a_0 \sinh \rho \sin \theta. \tag{5–9}$$

N	u		u_x		u_y	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
Mayo's technique						
40	5.068E0	2.356E0	4.146E0	3.575E0	8.513E-1	6.492E-1
80	6.787E-1	1.946E-1	6.443E-1	4.322E-1	2.004E-1	1.523E-1
160	1.388E-1	3.359E-2	1.236E-1	1.107E-1	4.859E-2	4.095E-2
320	3.777E-2	9.627E-3	2.523E-2	2.869E-2	1.120E-2	1.128E-2
640	8.629E-3	2.126E-2	5.548E-3	7.501E-3	3.003E-3	3.019E-3
Immersed interface method						
40	4.683E0	2.291E0	4.183E0	3.663E0	8.542E-1	6.511E-1
80	6.262E-1	1.791E-1	6.452E-1	3.264E-1	2.009E-1	1.527E-1
160	1.391E-1	3.015E-2	1.251E-1	1.133E-1	4.871E-2	4.101E-2
320	3.493E-2	8.695E-3	2.549E-2	2.944E-2	1.202E-2	1.130E-2
640	8.665E-3	1.921E-3	5.731E-3	7.663E-3	3.009E-3	3.024E-3

Table 3. Results for interface problem with $\beta = 1$, example (5–4). Normalized errors in computed solution and first derivatives, obtained for $a = 0.9$, $b = 0.1$.

Note that u_+ is harmonic, that is, $f_+ = 0$. The coefficients c and d in (5–7) are set to 1.26713535 and 1.12854242, respectively, so that $[u] = 0$.

To compute the solution for (5–5) and (5–6), $[\partial_n u] \equiv q$ is first computed by solving (4–6) iteratively:

$$q^{[k+1]} = \frac{2}{\beta_+ + \beta_-} \bigl(g_1 - (\beta_+ - \beta_-)\bigl(\int_S (\partial_n G) q^{[k]}\, ds + \int_{\Omega_-} (\partial_n G)(f_-/\beta)\, dy\bigr)\bigr) \tag{5–10}$$

Because $\partial_{n(x)} G(x-y)$ is nearly singular for y near (though not on) S, a naïve integration of the second integral containing $\partial_n G$ yields only $O(h)$ accuracy. To attain $O(h^2)$ accuracy, we follow a standard procedure and subtract

$$\partial_{n(x)} G(x-y) f_-(x)/\beta$$

from the integrand, where $x \in S$ is the point at which $q(x)$ in (5–10) is being evaluated; then we add an $O(h^2)$ approximation to $f_-(x)/\beta$ times

$$\int_\Omega \partial_{n(x)} G(x-y)\, dy = -\int_S n(x) \cdot n(y) G(x-y)\, ds(y). \tag{5–11}$$

The resulting interface condition $[\partial_n u] = q$, together with $[u] = 0$, is then used to solve (5–5): $f_\pm$ is divided by $\beta_\pm$, and then u is computed as in the previous example with constant coefficient $\beta = 1$.

Normalized errors in u are shown in Table 4 for two pairs of coefficients. In the first case, $\beta_+ = 2$ and $\beta_- = 0.5$; in the second case the difference between β's is increased substantially: $\beta_+ = 100$ and $\beta_- = 0.2$. Mayo's technique was used to compute correction terms for the finite-difference stencil. The results in Table 4 suggest that, for this problem, not only is $O(h^2)$ accuracy obtained as predicted by Theorem 2.1, but the accuracy of the method is insensitive to the difference between the β's. We did observe a small increase (∼20%) in the number

N	L^2	L^∞
$\beta_+ = 2$, $\beta_- = 0.5$		
40	1.979E-2	7.252E-2
80	5.637E-4	1.058E-3
160	1.381E-4	2.565E-4
320	3.485E-5	6.507E-5

N	L^2	L^∞
$\beta_+ = 100$, $\beta_- = 0.2$		
40	2.002E-2	7.298E-2
80	4.618E-4	9.524E-4
160	1.123E-4	2.302E-4
320	2.847E-5	5.857E-5

Table 4. Normalized errors in computed solution for the interface problem with piecewise-constant $\beta_\pm$. N denotes half of the number of subintervals in each dimension and along the interface S.

of iterations required for (5–10) to converge, when the ratio β_+/β_- was increased from 4 to 50. Similar results were also obtained for the immersed interface method, and for the cases where $\beta_- > \beta_+$.

Stokes equations. In the third example we solved the Stokes equations (4–7) for two fluids. In [13], Cortez derived analytic solutions for the one-fluid case where the enclosed boundary is a unit circle; see examples 4a and 4b in [13]. In each of those examples, the boundary force has either a normal or a tangential component. To obtain a more general example with nontrivial jumps $[p]$, $[\partial p/\partial n]$, $[\partial v/\partial n]$, we combined those two examples by adding the two solutions, and extended the resulting example to the two-fluid case. To that end, we assumed the same v as in [13], scaled p by the appropriate viscosity $\mu_\pm$, and computed boundary forces using (4–8) and (4–9). The resulting pressure and velocities are given by

$$p(r,\theta) = \begin{cases} \mu_+ r^{-3}(\sin 3\theta - \cos 3\theta), & r \ge 1, \\ -\mu_- r^3(\sin 3\theta + \cos 3\theta), & r < 1, \end{cases} \tag{5–12}$$

$$v_1(r,\theta) = \begin{cases} \tfrac{1}{8}r^{-2}(\sin 2\theta - \cos 2\theta) + \tfrac{1}{16}r^{-4}(-3\sin 4\theta + 5\cos 4\theta) \\ \qquad\qquad + \tfrac{1}{4}r^{-2}(\sin 4\theta - \cos 4\theta), & r \ge 1, \\ \tfrac{1}{8}r^{2}(3\sin 2\theta + \cos 2\theta) + \tfrac{1}{16}r^{4}(\sin 4\theta + \cos 4\theta) \\ \qquad\qquad + \tfrac{1}{4}r^{4}(-\sin 2\theta - \cos 2\theta), & r < 1, \end{cases} \tag{5–13}$$

$$v_2(r,\theta) = \begin{cases} \tfrac{1}{8}r^{-2}(\sin 2\theta + \cos 2\theta) + \tfrac{1}{16}r^{-4}(5\sin 4\theta + 3\cos 4\theta) \\ \qquad\qquad + \tfrac{1}{4}r^{-2}(-\sin 4\theta - \cos 4\theta), & r \ge 1, \\ \tfrac{1}{8}r^{2}(3\sin 2\theta - \sin 2\theta) + \tfrac{1}{16}r^{4}(\sin 4\theta - \cos 4\theta) \\ \qquad\qquad + \tfrac{1}{4}r^{4}(\sin 2\theta - \cos 2\theta), & r < 1. \end{cases} \tag{5–14}$$

With p and v chosen, the boundary force f is determined by (4–8), (4–9). The viscosities μ^+ and μ^- were set to 0.5 and 2, respectively. The computational domain Ω was chosen to be $[-3.0, 3.0] \times [-3.0, 3.0]$.

The solution was computed following the procedure described in Section 4. As noted there, the integral equation (4–18) must be solved to $O(h^3)$ accuracy so that the corrections at the interface lead to an $O(h^2)$ solution of the problem. The integral in (4–18) was approximated using the trapezoid rule, providing the necessary accuracy when S is a unit circle. Table 5 shows normalized errors in the solution obtained using Mayo's technique [32]. These results show evidence of the expected $O(h^2)$ convergence. The immersed interface method yielded similar accuracy.

N	p		u		v	
	L^2	L^∞	L^2	L^∞	L^2	L^∞
40	4.769E-2	7.646E-2	5.358E-2	4.558E-2	4.004E-2	3.726E-2
80	1.264E-2	2.192E-2	1.687E-2	1.377E-2	9.338E-3	8.745E-3
160	3.233E-3	5.485E-3	2.597E-3	2.268E-3	2.712E-3	2.564E-3
320	7.811E-4	1.352E-3	6.641E-4	5.825E-4	6.003E-4	5.458E-4
640	1.973E-4	3.385E-4	1.507E-4	1.326E-4	1.605E-4	1.480E-4

Table 5. Normalized errors in computed solution for the Stokes equations. N denotes half of the number of subintervals in each dimension and along the interface S. Results show second-order convergence.

Appendix

The following lemma shows that a parabola $y = ax^2 + b$ can cross a vertical grid line, near the vertex, such that the vertical distance from a grid point is of any specified order in h for small h. Thus a hypothesis such as in Theorem 5.2 of [27] is often violated.

Lemma A.1. *Given $a, b, \sigma \in R$, with $a \neq 0$ and $0 < \sigma < 1$, there are infinitely many integers $N > 0$ such that, with $h = 1/N$, there is a point $(x, y) \in R^2$ on the curve $y = ax^2 + b$ of the form*

$$x = jh, \quad y = kh + ch^{1+\sigma} \tag{A–1}$$

for some j, k, c depending on N, where j and k are integers, $\frac{1}{2} < c < 2$, and $x = jh = O(h^{(1+\sigma)/2})$.

Proof. According to a theorem of Dirichlet (see [1], Section 6.1, for instance), if b is irrational, there are infinitely many fractions m/N such that $b = m/N + \theta/N^2$ with $|\theta| < 1$. If b is irrational, we choose these N; if b is rational, we can choose infinitely many N so that $b = m/N$ for some m, and we take $\theta = 0$ in the argument to follow.

Substituting for x, y and b in $ax^2 + b = y$, we seek j, k, c so that

$$aj^2h^2 + mh + \theta h^2 = kh + ch^{1+\sigma} \tag{A–2}$$

or, multiplying by $N^2 = h^{-2}$,

$$aj^2 + mN + \theta = kN + cN^{1-\sigma}. \tag{A–3}$$

We choose $k = m$ and divide by a, so that the equation is now

$$j^2 + \theta/a = (c/a)N^{1-\sigma}. \tag{A–4}$$

We will first choose j as an approximate solution, ignoring θ and c, and then choose c. Let $r = \sqrt{N^{1-\sigma}/a}$, and let j be the greatest integer $\leq r$. Finally, define c so that (A.4) holds, that is, $c = (j^2 + \theta/a)/r^2$. It is easy to check that $c \to 1$ as $r \to \infty$, that is, as $N \to \infty$. Finally, $j = O(N^{1/2-\sigma/2})$, and $x = jh = O(N^{-1/2-\sigma/2})$. □

References

[1] A. Baker, *A concise introduction to the theory of numbers*, Cambridge University Press, Cambridge, 1984. MR 86f:11001 Zbl 0554.10001

[2] J. T. Beale and M.-C. Lai, *A method for computing nearly singular integrals*, SIAM J. Numer. Anal. **38** (2001), no. 6, 1902–1925. MR 2002f:65033 Zbl 0988.65025

[3] R. P. Beyer and R. J. LeVeque, *Analysis of a one-dimensional model for the immersed boundary method*, SIAM J. Numer. Anal. **29** (1992), no. 2, 332–364. MR 93a:65123 Zbl 0762.65052

[4] K. Böhmer, *Asymptotic expansion for the discretization error in linear elliptic boundary value problems on general regions*, Math. Z. **177** (1981), 235–255. MR 82d:65064 Zbl 0443.65073

[5] J. H. Bramble and B. E. Hubbard, *On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation*, Numer. Math. **4** (1962), 313–327. MR 26 #7157 Zbl 0135.18102

[6] ———, *A theorem on error estimation for finite difference analogues of the Dirichlet problem for elliptic equations*, Contributions to Differential Equations **2** (1963), 319–340. MR 27 #2114 Zbl 0196.50901

[7] J. H. Bramble, B. E. Hubbard, and V. Thomée, *Convergence estimates for essentially positive type discrete Dirichlet problems*, Math. Comp. **23** (1969), no. 107, 695–709. MR 42 #1350 Zbl 0217.21902

[8] J. H. Bramble and V. Thomée, *Pointwise bounds for discrete Green's functions*, SIAM J. Numer. Anal. **6** (1969), 583–590. MR 41 #7870 Zbl 0212.17904

[9] J. H. Bramble and J. T. King, *A finite element method for interface problems in domains with smooth boundaries and interfaces*, Adv. Comput. Math. **6** (1996), no. 2, 109–138 (1997). MR 98e:65094 Zbl 0868.65081

[10] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, 2nd ed., Texts in Applied Mathematics, no. 15, Springer, New York, 2002. MR 2003a:65103 Zbl 1012.65115

[11] Z. Chen and J. Zou, *Finite element methods and their convergence for elliptic and parabolic interface problems*, Numer. Math. **79** (1998), no. 2, 175–202. MR 99d:65313 Zbl 0909.65085

[12] P. G. Ciarlet, *Basic error estimates for elliptic problems*, Handbook of numerical analysis, Vol. II, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991, pp. 17–351. MR 1115237 Zbl 0875.65086

[13] R. Cortez, *The method of regularized Stokeslets*, SIAM J. Sci. Comput. **23** (2001), no. 4, 1204–1225. MR 2002k:76102 Zbl 1064.76080

[14] G. B. Folland, *Introduction to partial differential equations*, 2nd ed., Princeton University Press, Princeton, NJ, 1995. MR 96h:35001 Zbl 0841.35001

[15] G. E. Forsythe and W. R. Wasow, *Finite-difference methods for partial differential equations*, Applied Mathematics Series, Wiley, New York, 1960. MR 23 #B3156 Zbl 0099.11103

[16] Y. Fukai and K. Uchiyama, *Potential kernel for two-dimensional random walk*, Ann. Probab. **24** (1996), no. 4, 1979–1992. MR 97m:60098 Zbl 0879.60068

[17] H. Huang and Z. Li, *Convergence analysis of the immersed interface method*, IMA J. Numer. Anal. **19** (1999), no. 4, 583–608. MR 2000j:65073 Zbl 0940.65114

[18] J. Huang and J. Zou, *Some new a priori estimates for second-order elliptic and parabolic interface problems*, J. Differential Eq. **184** (2002), 570–586. MR 2003h:35026 Zbl 1012.35013

[19] A. Iserles, *A first course in the numerical analysis of differential equations*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 1996. MR 97m:65003 Zbl 0841.65001

[20] K. Ito, Y. Kyei, and Z. Li, *Higher-order, cartesian grid based finite difference schemes for elliptic equations and interface problems*, preprint.

[21] Z. Jomaa and C. Macaskill, *The embedded finite difference method for the Poisson equation in a domain with an irregular boundary and Dirichlet boundary conditions*, J. Comput. Phys. **202** (2005), no. 2, 488–506. MR 2005m:65241 Zbl 1061.65107

[22] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and quasilinear elliptic equations*, Academic Press, New York, 1968. MR 39 #5941 Zbl 0164.13002

[23] G. F. Lawler, *Intersections of random walks*, Probability and its Applications, Birkhäuser, Boston, 1991. MR 92f:60122

[24] R. J. LeVeque and Z. L. Li, *The immersed interface method for elliptic equations with discontinuous coefficients and singular sources*, SIAM J. Numer. Anal. **31** (1994), no. 4, 1019–1044. MR 95g:65139 Zbl 0811.65083

[25] R. J. LeVeque and Z. Li, *Immersed interface methods for Stokes flow with elastic boundaries or surface tension*, SIAM J. Sci. Comput. **18** (1997), no. 3, 709–735. MR 98b:76054 Zbl 0879.76061

[26] Z. Li, *An overview of the immersed interface method and its applications*, Taiwanese J. Math. **7** (2003), no. 1, 1–49. MR 2004c:65120 Zbl 1028.65108

[27] Z. Li and K. Ito, *Maximum principle preserving schemes for interface problems with discontinuous coefficients*, SIAM J. Sci. Comput. **23** (2001), no. 1, 339–361. MR 2002h:65166 Zbl 1001.65115

[28] Z. Li, T. Lin, and X. Wu, *New cartesian grid methods for interface problems using the finite element formulation*, Numer. Math. **96** (2003), no. 1, 61–98. MR 2005c:65104 Zbl 1055.65130

[29] X.-D. Liu and T. C. Sideris, *Convergence of the ghost fluid method for elliptic equations with interfaces*, Math. Comp. **72** (2003), no. 244, 1731–1746. MR 2004h:65107 Zbl 1027.65140

[30] A. Mayo, *A decomposition finite difference method for the fourth order accurate solution of poisson's equation on general regions*, Int. J. High Speed Computing **3** (1991), 89–106.

[31] A. Mayo and A. Greenbaum, *Fast parallel iterative solution of Poisson's and the biharmonic equations on irregular regions*, SIAM J. Sci. Statist. Comput. **13** (1992), no. 1, 101–118. MR 92k:65194 Zbl 0752.65080

[32] A. Mayo, *The fast solution of Poisson's and the biharmonic equations on irregular regions*, SIAM J. Numer. Anal. **21** (1984), no. 2, 285–299. MR 85i:65142

[33] ———, *Fast high order accurate solution of Laplace's equation on irregular regions*, SIAM J. Sci. Statist. Comput. **6** (1985), no. 1, 144–157. MR 86i:65066 Zbl 0559.65082

[34] ———, *The rapid evaluation of volume integrals of potential theory on general regions*, J. Comput. Phys. **100** (1992), no. 2, 236–245. MR 93c:65032 Zbl 0772.65012

[35] C. S. Peskin and B. F. Printz, *Improved volume conservation in the computation of flows with immersed elastic boundaries*, J. Comput. Phys. **105** (1993), no. 1, 33–46. MR 93k:76081 Zbl 0762.92011

[36] C. Pozrikidis, *Boundary integral and singularity methods for linearized viscous flow*, Cambridge Texts in Appl. Math., Cambridge University Press, New York, 1992. MR 93a:76027 Zbl 0772.76005

[37] A. Wiegmann and K. P. Bube, *The explicit-jump immersed interface method: finite difference methods for PDEs with piecewise smooth solutions*, SIAM J. Numer. Anal. **37** (2000), no. 3, 827–862. MR 2001b:65117 Zbl 0948.65107

Received July 19, 2005.

J. THOMAS BEALE: `beale@math.duke.edu`
Department of Mathematics, Duke University, Box 90320, Durham NC 27708, United States

ANITA T. LAYTON: `alayton@math.duke.edu`
Department of Mathematics, Duke University, Box 90320, Durham NC 27708, United States

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

THE FAST SINC TRANSFORM AND IMAGE RECONSTRUCTION FROM NONUNIFORM SAMPLES IN k-SPACE

LESLIE GREENGARD, JUNE-YUB LEE AND SOUHEIL INATI

A number of problems in image reconstruction and image processing can be addressed, in principle, using the sinc kernel. Since the sinc kernel decays slowly, however, it is generally avoided in favor of some more local but less precise choice. In this paper, we describe the fast sinc transform, an algorithm which computes the convolution of arbitrarily spaced data with the sinc kernel in $O(N \log N)$ operations, where N denotes the number of data points. We briefly discuss its application to the construction of optimal density compensation weights for Fourier reconstruction and to the iterative approximation of the pseudoinverse of the signal equation in MRI.

1. Introduction

A number of imaging modalities require the inversion of the equation

$$s(n) = \int_V \rho(\mathbf{r}) e^{2\pi \imath \mathbf{k}(n)\cdot\mathbf{r}} d\mathbf{r}\,, \tag{1}$$

where $\mathbf{k}(n)$ denotes the location of the nth measurement in the frequency domain ("k-space") and $\mathbf{r}$ denotes position in the image domain. It will be convenient below to write this in operator form as

$$\mathbf{s} = \mathcal{H}\rho(\mathbf{r})\,, \tag{2}$$

where $\mathcal{H}$ is the "continuous-to-discrete" Fourier operator which maps the image space to the signal space. (In standard magnetic resonance imaging, $\rho(\mathbf{r})$ is the proton spin density.)

We are particularly concerned with nonuniform sampling schemes, where the points $\{\mathbf{k}(n)\}$ do not lie on a regular grid. The inversion of (2) is, of course, an

Keywords: sinc interpolation, fast transform, nonuniform fast Fourier transform, density compensation weights, iterative methods, Fourier analysis, image reconstruction, magnetic resonance imaging (MRI).

Greengard was supported in part by the U.S. Department of Energy under contract DE-FG02-00-ER25053. Inati was supported in part by the McKnight Foundation, the Seaver Foundation, and the Swartz Foundation.

inherently ill-posed problem; the space of all possible densities $\rho(\mathbf{r})$ is infinite dimensional, while the vector of measurements $\{s(n)\}$ is finite dimensional. In the present paper, we concentrate on two possible approaches to reconstruction, leaving a more general discussion to [12].

Scheme 1. The first reconstruction scheme relies on the inverse Fourier transform

$$\rho(\mathbf{r}) = \iint s(\mathbf{k}) e^{-2\pi \imath \mathbf{k}\cdot\mathbf{r}} d\mathbf{k}\,, \tag{3}$$

or, more precisely, its approximation at M locations $\mathbf{r}_m$ via

$$\rho(\mathbf{r}_m) \approx \sum_{n=1}^{N} s(n) e^{-2\pi \imath \mathbf{k}(n)\cdot\mathbf{r}_m} w_n. \tag{4}$$

The computation of the sum (4) for every location appears to require $O(NM)$ operations. Using the nonuniform fast Fourier transform (NUFFT), however, this can be accomplished using $O((N+M)\log(N+M))$ operations. This is now a relatively mature technology [2; 6; 8; 9; 11; 14; 15; 17].

In (4), the values $\{w_n\}$ can be considered quadrature weights, and it is shown in [3; 12] that an optimal set of weights is given by the formula

$$\frac{1}{w_n} = \sum_{m=1}^{N} \operatorname{sinc}^2(\mathbf{k}(m) - \mathbf{k}(n)). \tag{5}$$

Here, $\operatorname{sinc}(k) \equiv \sin(\pi k)/\pi k$ and, in d dimensions, we define $\operatorname{sinc}(\mathbf{k}) = \operatorname{sinc}(k_1) \cdot \operatorname{sinc}(k_2) \cdots \operatorname{sinc}(k_d)$, where $\mathbf{k} = (k_1, k_2, \ldots, k_d)$. While the evaluation of these weights appears to require $O(N^2)$ operations, the fast sinc^2 transform, described below, makes use of the NUFFT to reduce the cost to $O(N \log N)$.

Scheme 2. A second class of reconstruction schemes is based directly on the signal equation (2). The minimum-norm least-squares solution to this problem, denoted by $\hat{\rho}(\mathbf{r})$, can be found by applying $\mathcal{H}^+$, the pseudo-inverse [10] of the operator $\mathcal{H}$, to the signal. Following the discussion of [19], we write

$$\hat{\rho}(\mathbf{r}) = \mathcal{H}^+ \mathbf{s} = \mathcal{H}^\dagger (\mathcal{H}\mathcal{H}^\dagger)^+ \mathbf{s}, \tag{6}$$

where $\mathcal{H}^\dagger$ is the adjoint of $\mathcal{H}$

$$[\mathcal{H}^\dagger \mathbf{a}](\mathbf{r}) = \sum_n e^{-2\pi \imath \mathbf{k}(n)\cdot\mathbf{r}} a(n),$$

where $\mathbf{a} = (a(1), \ldots, a(n))$ and $(\mathcal{H}\mathcal{H}^\dagger)^+$ is the pseudoinverse of $\mathcal{H}\mathcal{H}^\dagger$.

Given the N sample points $\{\mathbf{k}(n)\}$ in k-space, it is straightforward to verify that

$$(\mathcal{H}\mathcal{H}^\dagger)_{mn} = \operatorname{sinc}(\mathbf{k}_m - \mathbf{k}_n). \tag{7}$$

For notational convenience, we let $M = \mathcal{H}\mathcal{H}^\dagger$. The desired function $\hat{\rho}(\mathbf{r})$ in (6) can then be computed in two steps:

(1) Solve the system

$$M\mathbf{a} = \mathbf{s} \tag{8}$$

(2) Compute

$$\hat{\rho}(\mathbf{r}) = \mathcal{H}^\dagger \mathbf{a}.$$

The matrix M, however, may be ill-conditioned, with the precise condition number depending on the distribution of the sample points. (If two sample points coincide, for example, M is actually singular.) Thus, it is natural, as in [19], to use the pseudoinverse construction

$$\mathbf{a} = M^+\mathbf{s}\,,$$

which can be computed using the singular value decomposition (SVD) at a cost of $O(N^3)$ work. For this, some additional assumptions need to be made as to the choice of regularization [10]. Alternatively, one can attempt an iterative solution of (8). In [5], the authors suggest applying the conjugate gradient method, which is suitable for symmetric positive definite matrices. Since the cost of applying M to a vector is $O(N^2)$ work, the total cost of solving the system is of the order $O(J \cdot N^2)$, where J denotes the number of iterations.

Remark 1. Note that (4) can be written as

$$\rho(\mathbf{r}) \approx \mathcal{H}^\dagger W\mathbf{s}$$

where W is the diagonal matrix of quadrature weights. Thus, the quadrature approach based on the inverse Fourier transform can be viewed as a diagonal approximation ($W\mathbf{s}$) of the pseudoinverse construction ($M^+\mathbf{s}$). As a result, W serves as a good preconditioner for the conjugate gradient method applied to (8).

In summary, both Scheme 1 and Scheme 2 would benefit from appropriate fast algorithms: the optimal weights require a single convolution with the kernel $\mathrm{sinc}^2(\mathbf{k})$ and the iterative solution of the signal equation requires repeated convolution with the kernel $\mathrm{sinc}(\mathbf{k})$.

2. The fast sinc transform

Suppose now that we are given two sets of points $\{\mathbf{k}_n = (k_n^1, k_n^2, \ldots, k_n^d) |\ n = 1, \ldots, N\}$, and $\{\mathbf{v}_m = (v_m^1, v_m^2, \ldots, v_m^d) |\ m = 1, \ldots, M\}$, which we think of as located in the frequency domain in d dimensions. The point sets $\{\mathbf{k}_n\}$ and $\{\mathbf{v}_m\}$ may or may not coincide. We define the d-dimensional sinc and sinc^2 transforms by

$$U_m = \sum_{n=1}^{N} q_n \operatorname{sinc}(\mathbf{k}_n - \mathbf{v}_m) \quad \text{and} \quad W_m = \sum_{n=1}^{N} q_n \operatorname{sinc}^2(\mathbf{k}_n - \mathbf{v}_m), \tag{9}$$

respectively. Clearly, the naive computation of either U_m or W_m from q_n requires $O(M \cdot N)$ work. Since the transforms take the form of convolutions, it is perhaps not surprising that the Fourier transform will play a role in the fast algorithm. Since the data are not assumed to lie on a regular mesh, however, an essential ingredient will be the nonuniform fast Fourier transform (NUFFT), mentioned above. In its most general form, the NUFFT of "type 3" computes sums of the form

$$G_j = \sum_{p=1}^{P} g_p \, e^{-i\mathbf{k}_j \cdot \mathbf{x}_p} \,, \tag{10}$$

for $j = 1, \ldots, J$ or

$$g_p = \sum_{j=1}^{J} G_j \, e^{+i\mathbf{k}_j \cdot \mathbf{x}_p} \,, \tag{11}$$

for $p = 1, \ldots, P$ in $O((J + P)\log(J + P))$ operations to any desired precision. We will refer to equation (10) as the forward NUFFT. We can think of it as a discretization of the continuous Fourier transform,

$$G(\mathbf{k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) \, e^{-2\pi i \mathbf{x}\cdot\mathbf{k}} \, d\mathbf{x} = \mathcal{F} g(\mathbf{x}), \tag{12}$$

using nonuniformly sampled discretization points and evaluated at nonuniformly sampled frequencies. We will refer to (11) as the adjoint NUFFT. We can think of it as a discretization of the continuous inverse Fourier transform,

$$g(\mathbf{x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} G(\mathbf{k}) \, e^{2\pi i \mathbf{x}\cdot\mathbf{k}} \, d\mathbf{k} = \mathcal{F}^{-1} G(\mathbf{k}), \tag{13}$$

using nonuniformly sampled frequencies and evaluated at nonuniformly sampled discretization points.

Remark 2. The nomenclature *inverse* NUFFT would be misleading since, in the discrete case with nonuniform points, it is not the inverse of the forward transform.

Remark 3. The NUFFT has been used previously in order to accelerate iterative methods for the signal equation (1). In [5; 20], for example, the signal equation (or an analog) was discretized and the resulting linear system was solved using the conjugate gradient method applied to the normal equations. Their approach, however, did not make use of the sinc kernel.

The development of the fast sinc or sinc^2 transform now follows. For the sake of concreteness, we restrict our attention to the two-dimensional case, but the approach is clearly independent of dimension. First, we observe that the first equation in (9) can be viewed as the evaluation of the function

$$U(\mathbf{v}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \operatorname{sinc}(\mathbf{v}-\mathbf{k})H(\mathbf{k})\,d\mathbf{k} \tag{14}$$

at the points $\mathbf{v}_m$, due to the singular "source" distribution

$$H(\mathbf{k}) = \sum_{n=1}^{N} q_n\delta(\mathbf{k}-\mathbf{k}_n).$$

This follows from the elementary properties of the δ-function. From the convolution theorem we have that $U(\mathbf{v})$ is given by

$$U(\mathbf{v}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(\mathbf{x})\,e^{-2\pi i\mathbf{x}\cdot\mathbf{v}}\,d\mathbf{x} \tag{15}$$

with

$$u(\mathbf{x}) = \mathscr{F}^{-1}\operatorname{sinc}(\mathbf{k})\cdot\mathscr{F}^{-1}H(\mathbf{k}). \tag{16}$$

The latter two functions can be easily computed. The inverse Fourier transform $\mathscr{F}^{-1}\operatorname{sinc}(\mathbf{k})$ in two dimensions is simply

$$\Pi(\mathbf{x}) = \begin{cases} 0 & \text{if } |x_1| > 1/2 \text{ or } |x_2| > 1/2, \\ 1 & \text{if } |x_1| < 1/2 \text{ and } |x_2| < 1/2, \end{cases} \tag{17}$$

where $\mathbf{x} = (x_1, x_2)$. Further, it is easy to see that

$$h(\mathbf{x}) = \mathscr{F}^{-1}H(\mathbf{k}) = \sum_{n=1}^{N} q_n e^{2\pi i\mathbf{x}\cdot\mathbf{k}_n}. \tag{18}$$

Thus, we can compute $U(\mathbf{v}_m)$ from (15)-(18):

$$U(\mathbf{v}_m) = \int_{-1/2}^{1/2}\int_{-1/2}^{1/2} h(\mathbf{x})\,e^{-2\pi i\mathbf{x}\cdot\mathbf{v}_m}\,d\mathbf{x}. \tag{19}$$

This result is certainly classical.

2.1. ***Quadrature considerations.*** Equation (19) is an exact relation, and it remains only to discretize the integral. Doing so is straightforward, because $h(\mathbf{x})$ consists of a collection of exponential functions with maximum frequency given by $K_{\max} = \max_n \|\mathbf{k}_n\|_{L^1}$. Furthermore, we are only interested in computing $U(\mathbf{v}_m)$ itself up to the frequency $K_{\max}$ so that the term $e^{-2\pi i\mathbf{x}\cdot\mathbf{v}_m}$ also has a maximal frequency content given by $K_{\max}$. Thus, the integrand in (19) is a band-limited function with band limit $2 \cdot K_{\max}$. Nyquist sampling (two points per oscillation) requires that an accurate quadrature in two dimensions use at least $(4K_{\max})^2$ points.

Gauss–Legendre quadrature [4] is particularly useful in this context. While this approach is more involved than the trapezoidal rule or the rectangle rule, it achieves much higher order accuracy. More precisely, the P-point Gauss–Legendre rule can be defined by P weights $\{q_p\}$ and nodes $\{x_p\}$ so that the relation

$$\int_{-1}^{1} x^n dx = \sum_{p=1}^{P} q_p x_p^n$$

is exactly satisfied for $n = 0, \dots, 2P-1$. This yields a $2P \times 2P$ nonlinear system. Fortunately, the weights and nodes are easy to compute using standard software such as the Fortran routine `gaussq.f` from NETLIB (http://www.netlib.org). The weights are positive, but the nodes are not equally spaced, tending to cluster at the endpoints of the interval $[-1, 1]$. Given these weights and nodes, one of the remarkable features of Gauss–Legendre quadrature,

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{p=1}^{P} q_p f(x_p),$$

is that it satisfies the error estimate:

$$E = \left| \int_{-1}^{1} f(x)\,dx - \sum_{p=1}^{P} q_p f(x_p) \right| < \frac{2^{2P+1}(P!)^4}{(2P+1)[(2P)!]^3} \cdot \max |f^{2P}(x)|,$$

where $f^{2P}(x)$ denotes the $2P$-th derivative of the integrand. If the function $f(x)$ is band-limited by $2K_{\max}$, then $|f^{2P}(x)| < (4\pi\ K_{\max})^{2P}$. A modest amount of algebra then shows that the error E in using the P-point rule satisfies:

$$E < \frac{2\sqrt{\pi}\sqrt{P}}{(2P+1)} \left(\frac{1}{2e}\right)^{4P} \left(\frac{4\pi\ K_{\max}}{P}\right)^{2P} < \left(\frac{1}{e}\right)^{4P} \left(\frac{\pi\ K_{\max}}{P}\right)^{2P}.$$

Thus we see that, once P exceeds $\pi\ K_{\max}$, the error decays exponentially.

Using a tensor-product rule for the double integral, we have

$$\begin{aligned} U(\mathbf{v}_m) &= \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} h(\mathbf{x})\, e^{-2\pi i \mathbf{x} \cdot \mathbf{v}_m}\, d\mathbf{x} \\ &\approx \sum_{p_1=1}^{P} \sum_{p_2=1}^{P} h(x_{p_1}, x_{p_2}) e^{-2\pi i (x_{p_1}, x_{p_2}) \cdot \mathbf{v}_m}\, q_{p_1} q_{p_2} \end{aligned} \tag{20}$$

The error, as in the one-dimensional case, decays at an exponential rate once P exceeds $\pi K_{\max}$.

Remark 4. The usual weights and nodes are defined on the interval $[-1, 1]$ and we have scaled them to the interval $[-1/2, 1/2]$.

In summary, the fast sinc transform requires the adjoint NUFFT to compute $h(\mathbf{x})$ via (18) at the tensor product quadrature points. Given the values $h(x_{p_1}, x_{p_2})$, the forward NUFFT is used to compute (20). The amount of work is of the order $O((N+M+P^2)\log(N+M+P^2)) = O((N+M+K_{\max}^2)\log(N+M+K_{\max}^2))$. Since the quadrature used is spectrally accurate, the error in the fast sinc transform is dominated by the tolerance requested of the NUFFT. In most implementations, this is a user-controlled parameter and affects only the constant prefactor implicit in the $O((N+M)\log(N+M))$ notation.

Remark 5. For low accuracy, one could use the trapezoidal rule (a uniform mesh on $[-1/2, 1/2]$). The NUFFTs in this case are slightly more efficient, with a net savings of a factor of two or so in CPU time. The error is of the order $O(1/P^2)$, however, rather than exponentially small.

2.2. ***The fast* sinc^2 *transform.*** The theory underlying the sinc^2 transform is almost identical. The only change is that the inverse Fourier transform of $\operatorname{sinc}^2(\mathbf{k})$ in two dimensions is $t(x_1)\cdot t(x_2)$ where

$$t(x) = \begin{cases} 0 & \text{if } |x| > 1, \\ 1-|x| & \text{if } |x| < 1. \end{cases}$$

We therefore need to compute

$$W(\mathbf{k}) = \int_{-1}^{1}\int_{-1}^{1} h(x_1, x_2)\, t(x_1)\, t(x_2)\, e^{-2\pi i (x_1,x_2)\cdot\mathbf{k}}\, dx_1 dx_2. \tag{21}$$

Since the integrand is piecewise smooth, we maintain high accuracy by using four tensor product Gaussian quadrature rules (each with $P^2 > (\pi\, K_{\max})^2$ points) on the four quadrants $[-1,0]\times[-1,0]$, $[-1,0]\times[0,1]$, $[0,1]\times[-1,0]$, $[0,1]\times[0,1]$.

In summary, the fast sinc^2 transform requires the adjoint NUFFT to compute $h(\mathbf{x})$ via (18) at the tensor product quadrature points, followed by the forward NUFFT to compute (21) using tensor product Gaussian quadrature. The amount of work is again of the order $O((N+M+K_{\max}^2)\log(N+M+K_{\max}^2))$. Related algorithms that rely on the NUFFT for other convolution kernels are described in [16].

3. Results

We illustrate the performance of the algorithm in the context of magnetic resonance image reconstruction (MRI). In MRI, one seeks to produce a spatial map of the effective spin density $\rho(\mathbf{r})$ from raw complex-valued data $s(n) = s(\mathbf{k}(n))$ acquired in the Fourier domain. When the points $\mathbf{k}(n)$ are located on a Cartesian grid, the

N	T_{FST}	$T_{\mathrm{FS^2T}}$	T_{dir}	Error
4096	0.05	0.26	9.1	$< 10^{-3}$
16384	0.11	0.36	145	$< 10^{-3}$
4096	0.09	0.34	9.1	$< 10^{-5}$
16384	0.19	0.61	145	$< 10^{-5}$

Table 1. Timing results for FST and FS^2T on an Archimedean spiral with $K_{\max} = 64$. N denotes the number of sampling points along the spiral, T_{FST} denotes the time required for the fast sinc transform, $T_{\mathrm{FS^2T}}$ denotes the time required for the fast sinc^2 transform, and T_{dir} is the time required for the direct calculation. The direct calculation for sinc and sinc^2 are essentially the same, so only one timing is listed. Error is the requested tolerance for the NUFFT and is an upper bound on the L_2 error in the transform data. Calculations were carried out on a laptop computer with a 1.2GHz Pentium processor.

FFT is typically used to reconstruct the image according to (4) with constant weights $\{w_n\}$. Many modern techniques in MRI, however, including functional MRI, MR angiography, and abdominal imaging, use nonuniform samplings in **k**-space which allow for significantly faster data acquisition rates [1; 18].

One prototypical acquisition scheme is the Archimedean spiral, which we truncate at $K_{\max} = 64$ and sample at N points according to the formula

$$\mathbf{k}_n = K_{\max}\sqrt{\frac{n}{N}}\left(\cos\left(3\pi K_{\max}\sqrt{\frac{n}{N}}\right), \sin\left(3\pi K_{\max}\sqrt{\frac{n}{N}}\right)\right).$$

Before discussing the image reconstruction process itself, we first use this sampling pattern in order to test the efficiency of our fast transforms. For this, the points $\mathbf{k}_n$ serve as both the "source" locations and as the targets (the $\mathbf{v}_m$ in the earlier discussion). Sample timings are given in Table 1.

While the fast transform timings scale as expected with problem size, they rely on the NUFFT algorithm from [14], which has not yet been fine-tuned for performance. We believe that an order of magnitude improvement can be obtained through careful code optimization.

To illustrate the performance of the algorithm in terms of image quality, we generate synthetic data $s(n)$ according to (1) from a standard test image (the Shepp–Logan phantom [7; 13]), depicted in Figure 1. We then consider two data acquisition patterns: an integer Cartesian grid truncated at $K_{\max} = 64$ and the Archimedean spiral above. Both data sets contained $128^2 = 16,384$ (complex) values. The resulting reconstructions are shown in Figure 1. The top figure is based on the FFT using the Cartesian data, the lower left figure is based on the optimal weight reconstruction (Scheme 1) using (4), (5), and the lower right figure is obtained by using 5 iterations of the preconditioned conjugate gradient method (Scheme 2), with a diagonal preconditioner defined by the quadrature weights (5), as discussed in Remark 1. The total time for image reconstruction was approximately 0.2 seconds using the quadrature method (the lower left figure) and 1 second using the approximate pseudoinverse (lower right), the latter requiring 5 sinc transforms,

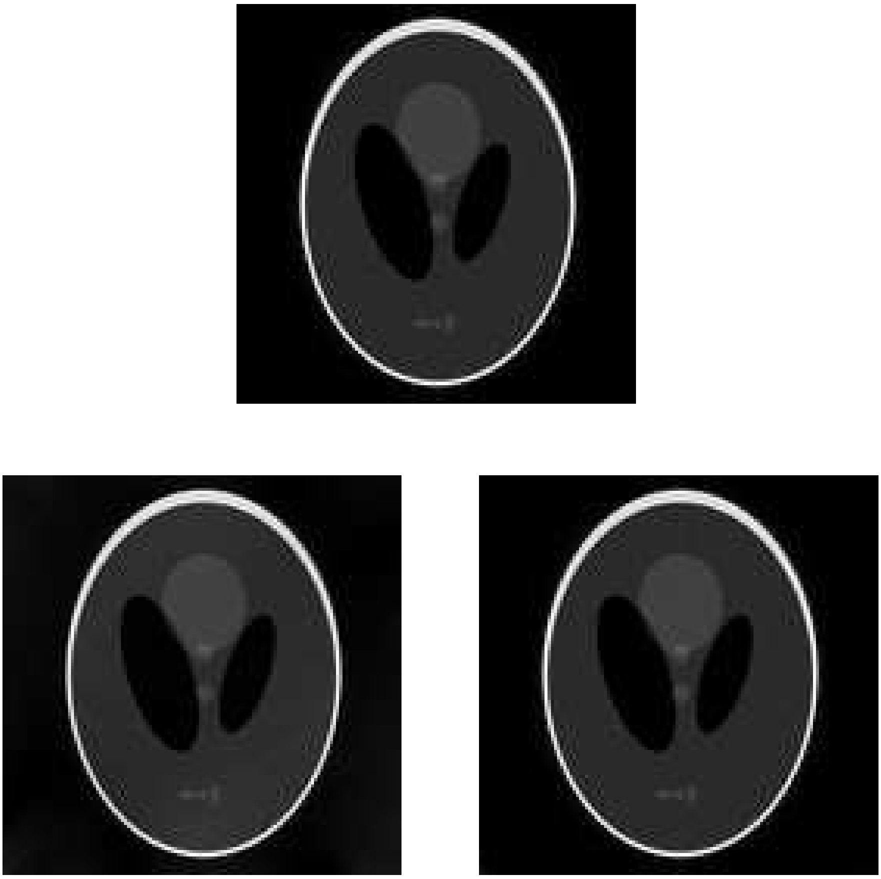


Figure 1. Image reconstruction from Cartesian (top) and spiral (bottom) k-space sampling. Note that the quadrature approximation (left) gives a very reasonable image. The pseudoinverse approximation (right) is nearly identical to that obtained in the Cartesian case.

one sinc^2 transform and one final NUFFT to apply the adjoint $\mathscr{H}^\dagger$. The FFT reconstruction is, of course, much faster — it required less than 0.01 seconds.

4. Discussion

We have constructed a fast algorithm for the (discrete) sinc and sinc^2 transforms which have immediate application in MR image reconstruction. The two algorithms will also accelerate, for example, the band-limited interpolation method of [3]. Since sinc convolution arises naturally in many signal and image processing contexts, we expect that the algorithms described here will be of fairly broad utility.

References

[1] C. B. Ahn, J. H. Kim, and Z. H. Cho, *High speed spiral-scan echo planar imaging*, IEEE Trans. Med. Imag. **M1-5** (1986).

[2] G. Beylkin, *On the fast Fourier transform of functions with singularities*, Appl. Comput. Harmon. Anal. **2** (1995), no. 4, 363–381. MR 96i:65122 Zbl 0838.65142

[3] H. Choi and D. C. Munson, *Analysis and design of minimax-optimal interpolators*, IEEE Trans. Signal Processing **46** (1998), 1571–1579.

[4] P. J. Davis and P. Rabinowitz, *Methods of numerical integration*, Computer Science and Applied Mathematics, Academic Press Inc., Orlando, FL, 1984. MR 86d:65004

[5] B. Desplanques, D. J. Cornelis, E. Achten, R. Van de Walle, and I. Lemahieu, *Iterative reconstruction of magnetic resonance images from arbitrary samples in k-space*, IEEE Trans. Nuc. Sci. **49** (2002), 2268–2273.

[6] A. Dutt and V. Rokhlin, *Fast Fourier transforms for nonequispaced data*, SIAM J. Sci. Comput. **14** (1993), no. 6, 1368–1393. MR 95d:65114 Zbl 0791.65108

[7] C. L. Epstein, *Mathematics of medical imaging*, Prentice-Hall, Englewood Cliffs, NJ, 2003.

[8] J. A. Fessler and B. P. Sutton, *Nonuniform fast Fourier transforms using min-max interpolation*, IEEE Trans. Signal Process. **51** (2003), no. 2, 560–574. MR 2003m:94024

[9] K. Fourmont, *Non-equispaced fast Fourier transforms with applications to tomography*, J. Fourier Anal. Appl. **9** (2003), no. 5, 431–450. MR 2005b:65154 Zbl 1073.65151

[10] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins Series in the Mathematical Sciences, vol. 3, Johns Hopkins University Press, Baltimore, MD, 1989. MR 90d:65055

[11] L. Greengard and J.-Y. Lee, *Accelerating the nonuniform fast Fourier transform*, SIAM Rev. **46** (2004), no. 3, 443–454. MR MR2115056 Zbl 1064.65156

[12] S. Inati, J.-Y. Lee, L. Fleysher, R. Fleysher, and L. Greengard, *Iterative reconstruction of magnetic resonance images from non-uniform samples in k-space*, In preparation.

[13] A. C. Kak and M. Slaney, *Principles of computerized tomographic imaging*, IEEE Press, New York, 1988. MR 90h:92005 Zbl 0721.92011

[14] J.-Y. Lee and L. Greengard, *The type 3 nonuniform FFT and its applications*, J. Comput. Phys. **206** (2005), no. 1, 1–5. MR MR2135833 Zbl 1072.65170

[15] A. Nieslony and G. Steidl, *Approximate factorizations of Fourier matrices with nonequispaced knots*, Linear Algebra Appl. **366** (2003), 337–351. MR 2004e:15016 Zbl 1018.65154

[16] D. Potts, G. Steidl, and A. Nieslony, *Fast convolution with radial kernels at nonequispaced knots*, Numer. Math. **98** (2004), no. 2, 329–351. MR 2005f:65184

[17] D. Potts, G. Steidl, and M. Tasche, *Fast Fourier transforms for nonequispaced data: a tutorial*, Modern sampling theory, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2001, pp. 247–270. MR MR1865690

[18] M. Schmitt, *On the sample complexity for neural trees*, Algorithmic learning theory (Otzenhausen, 1998), Lecture Notes in Comput. Sci., vol. 1501, Springer, Berlin, 1998, pp. 375–384. MR MR1683440

[19] R. Van de Walle, H. H. Barrett, K. J. Meyers, M. I. Altbach, B. Desplanques, A. F. Gmitro, J. Cornelis, and I. Lemahieu, *Reconstruction of mr images from data acquired on a general nonregular grid by pseudoinverse calculation*, IEEE Trans. Med. Imaging **19** (2000), 1160–1167.

[20] R. C. Wittmann, B. K. Alpert, and M. H. Francis, *Near-field antenna measurements using nonideal measurement locations*, IEEE Trans. Antennas Propagat. **46** (1998), 716–722.

Received November 30, 2005.

LESLIE GREENGARD: greengard@cims.nyu.edu
Courant Institute, New York University, New York, NY 10012, United States

JUNE-YUB LEE: jyllee@ewha.ac.kr
Department of Mathematics, Ewha Women's University, Seoul 120-750, Korea

SOUHEIL INATI: souheil.inati@nyu.edu
Center for Neural Science and Department of Psychology, New York University,
New York, NY 10003, United States

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

ON INTERPOLATION AND INTEGRATION IN FINITE-DIMENSIONAL SPACES OF BOUNDED FUNCTIONS

PER-GUNNAR MARTINSSON, VLADIMIR ROKHLIN AND MARK TYGERT

We observe that, under very mild conditions, an n-dimensional space of functions (with a finite n) admits numerically stable n-point interpolation and integration formulae. The proof relies entirely on linear algebra, and is virtually independent of the domain and of the functions to be interpolated.

1. Introduction

Approximation of functions and construction of quadrature formulae constitute an extremely well-developed area of numerical analysis; in most situations one is likely to encounter in practice, standard tools are satisfactory. Much of the research concentrates on obtaining powerful results under strong assumptions — designing interpolation and quadrature formulae for smooth functions on subspaces of $\mathbb{R}^n$, manifolds, etc. In this note, we make a very general observation that, given a finite set of bounded functions $f_1, f_2, \ldots, f_{n-1}, f_n$ (either real- or complex-valued) defined on a set S, there exists an interpolation formula that is exact on all linear combinations of $f_1, f_2, \ldots, f_{n-1}, f_n$, is numerically stable, and is based on n nodes in S (to be denoted $x_1, x_2, \ldots, x_{n-1}, x_n$). If, in addition, S is a measure space and the functions $f_1, f_2, \ldots, f_{n-1}, f_n$ are integrable, then there exists a quadrature formula based on the nodes $x_1, x_2, \ldots, x_{n-1}, x_n$ that is exact on all the functions $f_1, f_2, \ldots, f_{n-1}, f_n$, and is also numerically stable. Both of these statements are purely linear-algebraic in nature, and do not depend on the detailed properties of S, or of the functions $f_1, f_2, \ldots, f_{n-1}, f_n$.

It should be pointed out that all of the statements in this note follow easily from the analysis found in [4]; moreover, Theorem 2 can be found (in a slightly different form) in [7] and in [3]. Due to [3], the points used for interpolation in Theorem 2 are often called (nonelliptic) Fekete points, at least when the functions to be interpolated are polynomials. While we cannot cite earlier works where these

MSC2000: 65-02, ,

Keywords: interpolation, numerical integration, quadrature.

The authors were partially supported by the U.S. DoD under AFOSR Award #5710001605 and AFOSR Grant #F49620–03–C–0041, and by the U.S. NSF under Award #DMS–0139914.

observations are published, it seems unlikely that they had not been made a long time ago (perhaps in contexts other than numerical analysis).

This note has the following structure: Section 2 defines notation used in later sections, Section 3 provides a numerically stable interpolation scheme, Section 4 provides a numerically stable quadrature scheme, Section 5 provides a stronger result on the numerical stability of the interpolation scheme from Section 3, and Section 6 provides a couple of extensions to the techniques described in this note.

2. Notation

Throughout this note, S denotes an arbitrary set, n denotes a positive integer, and $f_1, f_2, \ldots, f_{n-1}, f_n$ denote bounded complex-valued functions on S (all results of this note also apply in the real-valued case, provided that the word "complex" is replaced with "real" everywhere). For any n points $x_1, x_2, \ldots, x_{n-1}, x_n$ in S, we define $A = A(x_1, x_2, \ldots, x_{n-1}, x_n)$ to be the $n \times n$ matrix defined via the formula

$$A_{j,k} = f_j(x_k) \tag{1}$$

with $j, k = 1, 2, \ldots, n-1, n$; we define the function g_k on S to be the ratio of the determinant of $A(x_1, x_2, \ldots, x_{k-2}, x_{k-1}, x, x_{k+1}, x_{k+2}, \ldots, x_{n-1}, x_n)$ to the determinant of $A(x_1, x_2, \ldots, x_{n-1}, x_n)$, via the formula

$$g_k(x) = \frac{\det A(x_1, x_2, \ldots, x_{k-2}, x_{k-1}, x, x_{k+1}, x_{k+2}, \ldots, x_{n-1}, x_n)}{\det A(x_1, x_2, \ldots, x_{n-1}, x_n)} \tag{2}$$

(here, the numerator is the same as the denominator, but with x in place of x_k). We define $D(x_1, x_2, \ldots, x_{n-1}, x_n)$ to be the modulus of the determinant of $A(x_1, x_2, \ldots, x_{n-1}, x_n)$, via the formula

$$D(x_1, x_2, \ldots, x_{n-1}, x_n) = \left|\det A(x_1, x_2, \ldots, x_{n-1}, x_n)\right|. \tag{3}$$

We define B to be the supremum of $D(x_1, x_2, \ldots, x_{n-1}, x_n)$ taken over all sets of n points $x_1, x_2, \ldots, x_{n-1}, x_n$ in S, via the formula

$$B = \sup_{x_1, x_2, \ldots, x_{n-1}, x_n \text{ in } S} D(x_1, x_2, \ldots, x_{n-1}, x_n). \tag{4}$$

For any x in S, we define $u = u(x)$ to be the $n \times 1$ column vector defined via the formula

$$u_k = f_k(x) \tag{5}$$

with $k = 1, 2, \ldots, n-1, n$, and we define $v = v(x)$ to be the $n \times 1$ column vector defined via the formula

$$v_k = g_k(x) \tag{6}$$

with $k = 1, 2, \ldots, n-1, n$.

3. Interpolation

Theorem 2 below asserts the existence of numerically stable n-point interpolation formulae for any set of n bounded functions; first we will need the following lemma.

Lemma 1. *Suppose that n is a positive integer, S is an arbitrary set containing at least n points, and $f_1, f_2, \dots, f_{n-1}, f_n$ are complex-valued functions on S that are linearly independent.*

Then, there exist n points $x_1, x_2, \dots, x_{n-1}, x_n$ in S such that the vectors $u(x_1)$, $u(x_2), \dots, u(x_{n-1}), u(x_n)$ defined in (5) *are linearly independent.*

Proof. We apply the modified Gram–Schmidt process (see, for example, [2]) to the set of all $n \times 1$ column vectors $u(x)$ defined in (5) for all x in S, while ensuring that all pivot vectors are nonzero via appropriate column-pivoting. □

Theorem 2. *Suppose that S is an arbitrary set, n is a positive integer, $f_1, f_2, \dots, f_{n-1}, f_n$ are bounded complex-valued functions on S, and ε is a positive real number such that*

$$\varepsilon \leq 1. \tag{7}$$

Then, there exist n points $x_1, x_2, \dots, x_{n-1}, x_n$ in S and n functions $g_1, g_2, \dots, g_{n-1}, g_n$ on S such that

$$|g_k(x)| \leq 1 + \varepsilon \tag{8}$$

for all x in S and $k = 1, 2, \dots, n-1, n$, and

$$f(x) = \sum_{k=1}^{n} f(x_k)\, g_k(x) \tag{9}$$

for all x in S and any function f defined on S via the formula

$$f(x) = \sum_{k=1}^{n} c_k\, f_k(x), \tag{10}$$

for some complex numbers $c_1, c_2, \dots, c_{n-1}, c_n$.

Proof. Without loss of generality, we assume that the functions $f_1, f_2, \dots, f_{n-1}, f_n$ are linearly independent.

Then, due to Lemma 1, B defined in (4) is strictly positive. Since the functions $f_1, f_2, \dots, f_{n-1}, f_n$ are bounded, $D(x_1, x_2, \dots, x_{n-1}, x_n)$ (defined in (3)) is also bounded, and hence the supremum B is not only strictly positive, but also finite. Therefore, by the definition of a supremum, there exist n points $x_1, x_2, \dots, x_{n-1}, x_n$ in S such that

$$B - D(x_1, x_2, \dots, x_{n-1}, x_n) \leq \frac{B}{2}\varepsilon \tag{11}$$

and $D(x_1, x_2, \dots, x_{n-1}, x_n)$ is strictly positive.

Defining $g_1, g_2, \dots, g_{n-1}, g_n$ via (2), we obtain (9) from the Cramer rule applied to the linear system

$$Av = u, \tag{12}$$

where $A = A(x_1, x_2, \dots, x_{n-1}, x_n)$ is defined in (1), $v = v(x)$ is defined in (6), and $u = u(x)$ is defined in (5). Due to the combination of (11) and (7),

$$\frac{B}{2} \leq D(x_1, x_2, \dots, x_{n-1}, x_n), \tag{13}$$

and due to the combination of (11) and (13),

$$\frac{B}{D(x_1, x_2, \dots, x_{n-1}, x_n)} - 1 \leq \varepsilon; \tag{14}$$

we also observe that, due to (4),

$$D(x_1, x_2, \dots, x_{k-2}, x_{k-1}, \; x, \; x_{k+1}, x_{k+2}, \dots, x_{n-1}, x_n) \leq B \tag{15}$$

for all x in S. Now, (8) is an immediate consequence of (2), (3), (15), (14). □

Remark 3. Due to (8), the interpolation formula (9) is numerically stable.

Remark 4. When calculations are performed using floating-point arithmetic, it is often desirable to "normalize" the set of functions $f_1, f_2, \dots, f_{n-1}, f_n$ before applying Theorem 2, by replacing this set with the set of functions $\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_{n-1}, \tilde{f}_n$, where $\tilde{f}_k$ is the function defined on S via the formula

$$\tilde{f}_k(x) = \frac{f_k(x)}{\sum_{j=1}^{n} |f_j(x)|}, \tag{16}$$

for example.

Remark 5. The proof of Theorem 2 does not specify a computational means for choosing the points $x_1, x_2, \dots, x_{n-1}, x_n$ so that (11) is satisfied (that is, so that the interpolation scheme from Theorem 2 and the quadrature scheme from Theorem 8 are guaranteed to be numerically stable). However, combining the algorithms described in [1], [4] with appropriate discretizations of S yields methods for choosing the points that are proven to work, both in theory and in practice.

Remark 6. It is not hard to see that, if m is a positive integer with $m < n$ such that any set of strictly more than m of the n functions $f_1, f_2, \dots, f_{n-1}, f_n$ is linearly dependent, then only m summands are required in (9); all but m of the functions $g_1, g_2, \dots, g_{n-1}, g_n$ can be arranged to vanish identically at all points in their domain S. Slight variations on the algorithms in [5], [6] yield efficient, effective computational methods for taking full advantage of this fact. For an application of this fact, see [1].

Remark 7. When S is compact and the functions f_1, f_2, …, f_{n-1}, f_n are continuous, Theorem 2 holds with $\varepsilon = 0$ rather than $\varepsilon > 0$, since a continuous function D on a compact space attains its maximal value. Analogously, Theorem 2 holds with $\varepsilon = 0$ when $S = \mathbb{R}^d$ for some positive integer d, the functions f_1, $f_2, \dots, f_{n-1}$, f_n are continuous, and $f_k(x) \to 0$ as $|x| \to \infty$ for all $k = 1, 2, \dots, n-1, n$.

4. Quadratures

The following theorem formalizes the obvious observation that integrating both sides of (9) yields numerically stable quadrature formulae.

Theorem 8. *Suppose that S is a measure space, w is a nonnegative real-valued integrable function on S (that serves as the weight for integration), n is a positive integer, $f_1, f_2, \dots, f_{n-1}, f_n$ are bounded complex-valued integrable functions on S, and $\varepsilon \le 1$ is a positive real number.*

Then, there exist n complex numbers $w_1, w_2, \dots, w_{n-1}, w_n$ such that

$$|w_k| \le (1+\varepsilon) \int w(x)\,dx \tag{17}$$

for all $k = 1, 2, \dots, n-1, n$, and

$$\int f(x)\,w(x)\,dx = \sum_{k=1}^{n} w_k\, f(x_k) \tag{18}$$

for any function f defined on S via (10), *where $x_1, x_2, \dots, x_{n-1}, x_n$ are the n points in S chosen in Theorem 2.*

Proof. For each $k = 1, 2, \dots, n-1, n$, we define w_k via the formula

$$w_k = \int g_k(x)\,w(x)\,dx, \tag{19}$$

where $g_1, g_2, \dots, g_{n-1}, g_n$ are the functions from Theorem 2. Then, (17) is an immediate consequence of (19) and (8). Moreover, (18) is an immediate consequence of (9) and (19). □

Remark 9. Needless to say, the weight function w in the above theorem is superfluous; it could be absorbed into the measure on S. However, we found the formulations of Theorems 8 and 12 involving w to be convenient in applications. While Theorems 8 and 12 require the functions $f_1, f_2, \dots, f_{n-1}, f_n$ to be bounded (perhaps after "normalizing" them as in Remark 4 or otherwise rescaling them), these theorems do not require the weight function w to be bounded.

Remark 10. Theorem 8 asserts the existence under very mild conditions of numerically stable quadratures that integrate linear combinations of n functions using the values of these linear combinations tabulated at n appropriately chosen points. In contrast, construction of optimal "generalized Gaussian" quadratures, which tabulate the linear combinations at fewer nodes than the number of functions, requires more subtle techniques (see, for example, the references cited in [8]).

5. Strengthened numerical stability

Theorem 2 provides the bound (8) under the rather weak assumption that the functions $f_1, f_2, \dots, f_{n-1}, f_n$ are bounded (in fact, this assumption is necessary for (8)). Theorem 12 below provides a stronger bound under the additional assumption that there exists a measure with respect to which the functions $f_1, f_2, \dots, f_{n-1}$, f_n are orthonormal. This stronger bound can be obtained by first using Theorem 2 to reconstruct the function f defined in (10) on its entire domain S from its values $f(x_1)$, $f(x_2)$, $\dots$, $f(x_{n-1})$, $f(x_n)$, as per (9). Then, the coefficients c_1, $c_2, \dots, c_{n-1}, c_n$ in (10) can be calculated by taking the appropriate inner products with the reconstruction of f just obtained. Finally, f can be reconstructed on its entire domain S via (10), using the values of $c_1, c_2, \dots, c_{n-1}, c_n$ just obtained, and the values $f_1(x), f_2(x), \dots, f_{n-1}(x), f_n(x)$, which are assumed to be known for any x in S. (However, please note that the proof of Theorem 12 given below follows this prescription only implicitly.) First, we will need the following lemma, stating that the relation (9) determines the functions $g_1, g_2, \dots, g_{n-1}, g_n$ uniquely, provided that the functions $f_1, f_2, \dots, f_{n-1}, f_n$ are linearly independent.

Lemma 11. *Suppose that n is a positive integer, S is an arbitrary set containing at least n points, $f_1, f_2, \dots, f_{n-1}, f_n$ are bounded complex-valued functions on S, and $\varepsilon \le 1$ is a positive real number. Suppose in addition that $f_1, f_2, \dots, f_{n-1}$, f_n are linearly independent, and that $h_1, h_2, \dots, h_{n-1}, h_n$ are functions on S such that*

$$f(x) = \sum_{k=1}^{n} f(x_k)\, h_k(x) \tag{20}$$

for all x in S and any function f defined on S via (10), *where $x_1, x_2, \dots, x_{n-1}$, x_n are the n points in S chosen in Theorem 2.*

Then,

$$h_k(x) = g_k(x) \tag{21}$$

for all x in S and $k = 1, 2, \dots, n-1, n$, where $g_1, g_2, \dots, g_{n-1}, g_n$ are defined in (2).

Proof. For any x in S, due to (20),

$$At = u, \tag{22}$$

where $A = A(x_1, x_2, \dots, x_{n-1}, x_n)$ is defined in (1), $u = u(x)$ is defined in (5), and $t = t(x)$ is defined to be an $n \times 1$ column vector via the formula

$$t_k = h_k(x) \tag{23}$$

with $k = 1, 2, \dots, n-1, n$; subtracting (12) from (22),

$$A(t - v) = 0, \tag{24}$$

where $v = v(x)$ is defined in (6). Due to Lemma 1, B defined in (4) is strictly positive, so that A defined in (1) is invertible, and therefore, due to (24),

$$t(x) = v(x) \tag{25}$$

for all x in S. Then, (21) is an immediate consequence of (25), (23), (6). □

Theorem 12. *Suppose that n is a positive integer, S is a measure space containing at least n points, w is a nonnegative real-valued integrable function on S (that serves as the weight for integration), $f_1, f_2, \dots, f_{n-1}, f_n$ are bounded complex-valued square-integrable functions on S, and $\varepsilon \le 1$ is a positive real number. Suppose further that $f_1, f_2, \dots, f_{n-1}, f_n$ are orthonormal, that is,*

$$\int |f_k(x)|^2 \, w(x) \, dx = 1 \tag{26}$$

for all $k = 1, 2, \dots, n-1, n$, and

$$\int \overline{f_j(x)} \, f_k(x) \, w(x) \, dx = 0 \tag{27}$$

whenever $j \neq k$.

Then,

$$|g_k(x)| \le (1 + \varepsilon) \sqrt{\int w(y) \, dy} \sum_{j=1}^{n} |f_j(x)| \tag{28}$$

for all x in S and $k = 1, 2, \dots, n-1, n$, where $g_1, g_2, \dots, g_{n-1}, g_n$ are defined in (2), *with the n points $x_1, x_2, \dots, x_{n-1}, x_n$ in S chosen in Theorem 2.*

Proof. In order to prove (28), for each $k = 1, 2, \dots, n-1, n$, we define the function h_k on S via the formula

$$h_k(x) = \sum_{j=1}^{n} f_j(x) \int \overline{f_j(y)} \, g_k(y) \, w(y) \, dy \tag{29}$$

and demonstrate both that (21) holds with the functions $h_1, h_2, \dots, h_{n-1}, h_n$ defined in (29), and that

$$|h_k(x)| \le (1+\varepsilon) \sqrt{\int w(y)\, dy} \sum_{j=1}^{n} |f_j(x)| \tag{30}$$

for all x in S and $k = 1, 2, \dots, n-1, n$.

We first show that (21) holds with the functions $h_1, h_2, \dots, h_{n-1}, h_n$ defined in (29), by demonstrating that $h_1, h_2, \dots, h_{n-1}, h_n$ satisfy the hypotheses of Lemma 11. Suppose that f is defined via (10). To verify that (20) holds with the functions $h_1, h_2, \dots, h_{n-1}, h_n$ defined in (29), we substitute (29) into the right hand side of (20) and exchange the orders of summation and integration, obtaining that

$$\sum_{k=1}^{n} f(x_k)\, h_k(x) = \sum_{j=1}^{n} f_j(x) \int \overline{f_j(y)} \sum_{k=1}^{n} f(x_k)\, g_k(y)\, w(y)\, dy \tag{31}$$

for all x in S. Due to the combination of (31) and (9),

$$\sum_{k=1}^{n} f(x_k)\, h_k(x) = \sum_{j=1}^{n} f_j(x) \int \overline{f_j(y)}\, f(y)\, w(y)\, dy \tag{32}$$

for all x in S. Then, (20) is an immediate consequence of applying (10), (26), and (27) to the right hand side of (32).

Furthermore, the functions $f_1, f_2, \dots, f_{n-1}, f_n$ are linearly independent, since they are assumed to be orthonormal. Thus, all of the hypotheses of Lemma 11 are satisfied, so (21) holds with the functions $h_1, h_2, \dots, h_{n-1}, h_n$ defined in (29).

Finally, due to the Cauchy–Schwarz inequality,

$$\left| \int \overline{f_k(y)}\, g_k(y)\, w(y)\, dy \right| \le \sqrt{\int |f_k(y)|^2\, w(y)\, dy} \sqrt{\int |g_k(y)|^2\, w(y)\, dy}, \tag{33}$$

and, due to (8),

$$\sqrt{\int |g_k(y)|^2\, w(y)\, dy} \le (1+\varepsilon) \sqrt{\int w(y)\, dy} \tag{34}$$

for all $k = 1, 2, \dots, n-1, n$. Then, (30) is an immediate consequence of (29), (33), (26), (34), and then (28) is an immediate consequence of (21) and (30). □

Remark 13. Due to (28), the interpolation formula (9) is numerically stable. While the numerical stability guaranteed by (8) is sufficient under most conditions, sometimes the bound (28) is more useful. The bound (28) is stronger than the bound (8) in the sense that, if all of the values $|f_1(x)|, |f_2(x)|, \dots, |f_{n-1}(x)|, |f_n(x)|$ are small at some point x in S, then all of the values $|g_1(x)|$, $|g_2(x)|$, ..., $|g_{n-1}(x)|$, $|g_n(x)|$ are accordingly small at that point x.

Remark 14. Theorem 12 generalizes easily to the case when the functions f_1, f_2, ..., f_{n-1}, f_n are not precisely orthonormal, but only "close" to orthonormal, in the sense that the condition number of their Gram matrix is reasonably small.

6. Concluding remarks

The following remarks pertain to some fairly obvious but nonetheless useful extensions of the techniques described in this note.

Remark 15. One often encounters infinite-dimensional spaces of functions that are finite-dimensional to a specified precision. A typical situation of this kind involves the range of a compact operator, and the usual way to construct the finite-dimensional approximation is via the Singular Value Decomposition (see, for example, [8]). When combined with this observation, the apparatus of the present note becomes applicable to many infinite-dimensional spaces of functions.

Remark 16. In numerical practice, rather than dealing directly with functions that have finite mass or finite energy but are nevertheless unbounded, we often instead treat the bounded functions obtained from the unbounded ones via either spectral/pseudospectral transforms or localized averaging (involving convolution with kernels that are bounded or have finite energy, for example).

7. Acknowledgements

We would like to thank Reiji Suda and Masaaki Sugihara for pointing out the standard term "Fekete points". In addition, we would like to thank the anonymous reviewers for helpful editorial suggestions.

References

[1] H. Cheng, Z. Gimbutas, P.-G. Martinsson, and V. Rokhlin, *On the compression of low-rank matrices*, SIAM J. Sci. Comput. **26** (2005), no. 4, 1389–1404.

[2] G. Dahlquist and Å. Björck, *Numerical methods*, Dover, Mineola, NY, 2003. MR 2004a:65001

[3] M. Fekete, *Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten*, Mathematische Zeitschrift **17** (1923), 228–249.

[4] M. Gu and S. C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput. **17** (1996), no. 4, 848–869.

[5] G. W. Stewart, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Process. **40** (1992), no. 6, 1535–1541.

[6] ———, *Updating a rank-revealing ULV decomposition*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 2, 494–499.

[7] R. Suda, *Stability analysis of the fast Legendre transform algorithm based on the fast multipole method*, Proc. Estonian Acad. Sci. Phys. Math. **53** (2004), no. 2, 107–115.

[8] N. Yarvin and V. Rokhlin, *Generalized Gaussian quadratures and singular value decompositions of integral operators*, SIAM J. Sci. Comput. **20** (1998), no. 2, 699–718.

Received December 10, 2005.

PER-GUNNAR MARTINSSON: per-gunnar.martinsson@colorado.edu
Department of Applied Mathematics, University of Colorado, 526 UCB, Boulder, CO 80309, United States

VLADIMIR ROKHLIN: *Departments of Computer Science, Mathematics, and Physics,
Yale University, New Haven, CT 06511, United States*

MARK TYGERT: mark.tygert@yale.edu
Department of Mathematics, Yale University, New Haven, CT 06511, United States

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

NONLOCAL DAMAGE ACCUMULATION AND FLUID FLOW IN DIATOMITES

GRIGORY ISAAKOVICH BARENBLATT,
MICHIEL BERTSCH AND CARLO NITSCH

We investigate a new model for fluid flows in diatomite formations recently introduced by Barenblatt, Patzek, Prostokishin and Silin. We provide numerical evidences of the existence of a sharp front between the damaged and the undamaged rock, and we study the structure of this front. Finally, we simulate some infield operations and we set up a qualitative model control problem to maximize the profit during the oil extraction.

1. Introduction

In the last decades oil-bearing diatomite formations have attracted special attention. An example of this kind of deposits is the giant oil fields of Lost Hills and Belridge in California. The development of such deposits has some characteristic properties due to certain peculiarities of diatomaceous rocks: high porosity, very low permeability in the pristine state, and fragility. Because of the high porosity, the diatomite oil reservoirs are often very rich, but, in view of the low permeability, the wells placed in this kind of oil reservoirs have very low productivity unless the hydraulic fracture technique is applied. This technique, which basically consists of injection under high pressure of very viscous fluid from the wells inside the reservoir, increases rock permeability. However, because of fragility, a long-term fracturing process causes subsidence phenomena, with very serious consequences for the safety of the wells themselves. Therefore, though the damaging process is necessary to increase production, it has to be monitored to avoid well collapse.

Recently a new model of fluid flows through diatomaceous rocks was introduced by Barenblatt, Patzek, Prostokishin and Silin [2], taking into account the microstructural changes in diatomites that occur during the filtration of fracturing fluid. In Section 2 we shall review its physical background. The model contains several functional relationships and coefficients which cannot be chosen quantitatively

MSC2000: primary 76S05; secondary 74A45, 35K65.
Keywords: flow in porous media, oil engineering, continuum damage mechanics, degenerate parabolic system.

without further experimental study. In this paper we restrict ourselves to power-type functional relationships. In this case, the model leads to the system

$$\begin{cases} \partial_t \omega \;=\; \Big[\Lambda^2\, \nabla\big(\omega^{\mu}(p-I)_+^{\beta}\, \nabla\omega\big) \;+\; A(1-\omega)(p-I)_+^{\gamma}\Big]_+ \,, \\ \partial_t p \;=\; K\, \nabla\big(\omega^{\alpha}\, \nabla p\big), \end{cases}$$

where the subscript "+" indicates the positive part. The equations are obtained by averaging over the height of the diatomaceous stratum, and thus the problem becomes two-dimensional; the pressure of the fluid at a point $\mathbf{x}$ and time t is denoted by $p(\mathbf{x}, t)$, and, as usual in continuum damage mechanics, $\omega(\mathbf{x}, t)$ is the so-called "damage parameter", with values between 0 and 1. The constant I takes into account the strength of the rock; μ is a nonnegative constant and Λ, K, α, β, γ, and A are positive constants.

As can be easily observed from the second equation, we have made the assumption that in the undamaged rocks, where $\omega = 0$, the permeability vanishes. Already in [2] it was conjectured that this assumption leads to a free boundary problem. In other words, we expect that there exists an *a priori* unknown front that separates the damaged and undamaged rocks. A justification of this assumption is supplied by experimental observations (see the satellite photograph in Figure 1). Throughout this paper we shall show numerical simulations which validate such a conjecture,

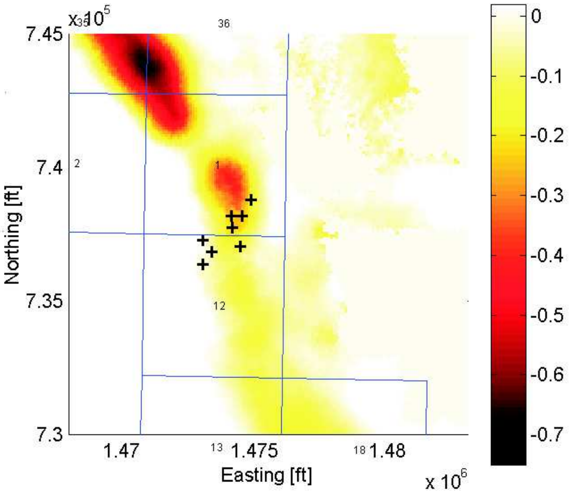


Figure 1. The wells that failed in South Belridge diatomaceous deposit in the year 2000 were located outside the large subsidence bowl. Surface subsidence in mm/day. Courtesy of T. W. Patzek, D. B. Silin (UC Berkeley, LBNL) and E. J. Fielding (NASA JPL).

and for a slightly simplified form of the system it is even possible to perform an analytical investigation (see [5; 4]) which leads to the same conclusion.

This model of fluid flow in diatomites has a wide range of applications, and in this paper we show some of them by simulating infield operations. We simulate several wells placed in five points formation and compute the accumulation of damage and the fluid extraction rate. Moreover we shall set up some model control problems for which it makes sense to look for an optimal strategy, if we impose some feasible restriction on the fluid pressure variations at the wells.

2. Physical background

We remind that the model we shall consider has been specifically introduced in order to deal with the filtration of fluids in a very special kind of rocks called diatomites. A typical picture of the microstructure of the diatomite as it is observed with an electronic microscope is given in Figure 2. The peculiarities of the diatomite in its pristine state are: a very large bulk porosity m that can be up to 70%, and a very low permeability k, of the order of 0.1–1 md (10^{-12}–10^{-11} cm^2) and even less; therefore, for practical purposes, in its pristine state it can be considered as impermeable. For this reason some of the wells in the Lost Hills field in California at the beginning of the primary recovery had a very low, practically zero, production. In fact, the oil production in diatomaceous deposits started only when the technique of hydraulic fracturing was applied.

Hydraulic fracturing, a technology developed in the 1950s, gave producers the possibility to extract more oil out of newly discovered and existing fields. Powerful pumps at the surface inject a fluid (in the beginning a viscous fluid carrying sand was

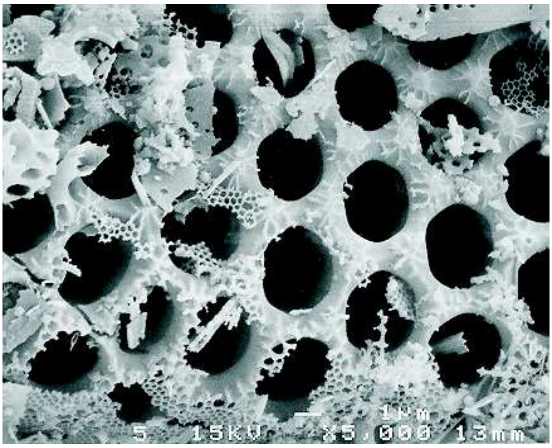


Figure 2. The fragile microstructure of the diatomite rock in SEM microphotograph. Courtesy of Prof. T. W. Patzek.

used, the so-called "fracture fluid") into the reservoir rocks. The pressure exerted by the fluid exceeds the compressive stress of the rock, opening fractures which constitute paths of increased permeability. Thus, when the pressure is released, the sand supports the crack opening and the fluid can flow more easily. Sometimes injection and extraction are performed through the same well.

In diatomaceous oil-bearing formations like Lost Hills, hydraulic fracturing is performed by injecting water. Owing to the peculiarities of the diatomite, the microstructure of this rock has to change to get any appreciable fluid flow. Actually, during field operations the stress in the rock leads to the collapse of wall pores, resulting in a network of microcracks that increases the permeability of the diatomite. Eventually, the microcrack net connects with the macrofracture. Such an interaction of fractures at different scales, which goes down to the microscopic level, is a peculiarity of the diatomite oil-bearing formations, and motivated the development of a new model based on the continuum damage mechanics approach.

We start from the assumption that the filtration of the fluid and the accumulation of cracks are strongly coupled, so that we have to derive a model which solves simultaneously the macroscopic fluid flow and the microstructural changes of diatomite.

We begin from the filtration equation of the fluid in the diatomaceous stratum. We make the following simplifying assumptions:

A1. *Water and oil are not distinguished.* Thereafter we use the word "fluid" to refer to both species.

A2. *The diatomaceous stratum is homogeneous, with constant height and depth, and bounded from above and below by impermeable rocks.*

A3. *Inside the reservoir, during geologic times, the pressure p of the fluid and the mean geostatic stress σ assumed, respectively, the constant values p_i and σ_i.* (The mean geostatic stress is $\frac{1}{3}(\sigma_x + \sigma_y + \sigma_z)$, that is, one-third of the first invariant of the stress tensor.)

A4. *The deposit is "deep".* This assumption together with A3 implies that also if we perturb the fluid pressure p from its equilibrium initial value p_i, the sum $p + \sigma$ remains constant and is equal to $p_i + \sigma_i$ during the whole process.

A5. *The diatomite is a weakly compressible elastic porous medium.*

A6. *The fluid is weakly compressible, so its density ρ is a linear function of pressure.*

Under such hypotheses, following [2] (see also [1]), we obtain from the continuity equation $\partial_t(m\rho) + \nabla(\rho\mathbf{u}) = 0$ (where $\mathbf{u}$ is the filtration velocity) and from Darcy's law $\mathbf{u} = -(k/\mu)\,\nabla p$ (where μ is the fluid viscosity) the equation for the pressure:

$$\partial_t p = \nabla(\mathcal{K}\,\nabla p). \tag{1}$$

Here the “piezo-diffusivity” coefficient $\mathcal{K}$ is defined as

$$\mathcal{K} = \frac{k}{\mu m c},$$

where c is a constant taking into account the compressibility coefficient of the fluid and the compressibility coefficient of the rock porosity. Equation (1) is defined in two spatial dimensions, and all the quantities involved (pressure, porosity, permeability, compressibility, etc.) have to be interpreted as averaged over the height of the diatomaceous stratum.

The key idea in [2] was to consider permeability no longer as a fixed quantity, but as a function of the rock damage, $k = k(\omega)$. In subterranean mechanics sometimes one uses pressure-dependent permeability, but we will neglect this dependence. We will also neglect the contribution to the porosity, given by the microcracks opened during the damage accumulation, because the volume fraction of cracks is small.

The basic equation now becomes

$$\partial_t p = \nabla\big(\mathcal{K}(\omega)\,\nabla p\big). \tag{2}$$

Here $\mathcal{K}(\omega)$ is a fast growing function of ω, and $\mathcal{K}(0) = 0$ by hypothesis. In [2] no further assumption was made about the form of $\mathcal{K}(\,\cdot\,)$ due to the lack of experimental evidence. However, in the following, in order to perform a numerical and analytical investigation, we will assume as a first step that $\mathcal{K}(\omega) = K\omega^{\alpha}$, $\alpha > 0$. Clearly, the permeability is the function of time and space, but here we assume that the space and time variability of permeability is due only to space and time variability of damage.

To complete the problem formulation, it is now necessary to add an equation for the damage accumulation. Initially the rock is considered pristine ($\omega = 0$), with the possible exception of small regions around the wells that appear during the drilling. When the water is pumped into the wells and starts to filtrate in the diatomaceous rock, the pressure in the pores eventually increases above a certain critical value I and microcracks start to appear. The exact value of I is unknown and has to be determined by infield experiments. We claim that it has to be not less then p_i, since during geological time no damage has been accumulated (we neglect seismic and tectonic effects).

It is natural to make the basic assumption that locally the damage accumulation rate $\partial_t\omega$ is proportional to a certain power of $(p-I)_+$, and also proportional to the fraction of undamaged bonds $1-\omega$. Therefore, as in classical continuum damage mechanics, the process of damage accumulation is governed locally by a kinetic equation of the following type:

$$\frac{d\omega}{dt} = A(1-\omega)(p-I)_+^{\gamma},$$

where A is a constant. In addition, together with this bulk mechanism, we also consider a nonlocal damage diffusion process. We expect, in fact, that fluid wedging take place in the microcracks. Again, we are focused on qualitative evaluation of the equations, and we choose a very simple nonlocal damage evolution equation, in the form

$$\partial_t \omega = \left[\nabla\big(D(\omega, p)\,\nabla\omega\big) + A(1-\omega)(p-I)_+^\gamma\right]_+, \tag{3}$$

where the positive part on the right hand side avoids a nonphysical damage healing. In particular, we will use the expression

$$D(\omega, p) = \Lambda^2\,\omega^\mu (p-I)_+^\beta$$

for the damage diffusivity coefficient, where Λ is constant.

3. The mathematical problem formulation

Equations (2) and (3) are the starting point for a mathematical formulation of the problem of fluid flow in diatomaceous rocks, leading to the following nonlocal damage accumulation and pressure evolution model:

$$\begin{cases} \partial_t \omega = \left[\Lambda^2\,\nabla\big(\omega^\mu (p-I)_+^\beta \nabla\omega\big) + A(1-\omega)(p-I)_+^\gamma\right]_+, \\ \partial_t p = K\,\nabla\big(\omega^\alpha\,\nabla p\big), \end{cases} \tag{4a}$$

$$\omega(\mathbf{x}, 0) = \omega_0(\mathbf{x}), \tag{4b}$$

$$p(\mathbf{x}, 0) = p_0(\mathbf{x}), \tag{4c}$$

where $\omega_0(\mathbf{x})$ and $p_0(\mathbf{x})$ are given initial distributions of the damage and pressure. We will refer to system (4) as the *2D formulation of the "diatomite problem"*.

We remind that ω is the vertically averaged damage in the oil-bearing layer of diatomite rock, while p represents the averaged pressure of fluid contained in the stratum. Moreover, I is related to the strength of the material and represents a threshold pressure under which no damage accumulation or diffusion occurs. In the *analytical* investigation, we assume for simplicity that $p = 0$ corresponds to the rest pressure of the fluid, p_i, in the undamaged zone. Therefore we shall also assume $I \geq 0$. The constants α, β, γ, μ and A satisfy

$$\alpha, \beta, \gamma, A > 0 \qquad \text{and} \qquad \mu \geq 0. \tag{5}$$

System (4) exhibits two major mathematical difficulties:

- the nonnegativity of $\partial_t \omega$, which physically represents the condition of no healing but which mathematically renders the equation for ω "fully nonlinear";
- the degeneracy of the diffusion coefficients $\omega^\mu (p-I)_+^\beta$ and ω^α if $\omega = 0$ and $p \leq I$.

We also consider the one-dimensional version of Equation (4a)

$$\begin{cases} \partial_t \omega = \Big[\Lambda^2\, \partial_x\big(\omega^\mu (p-I)_+^\beta\, \partial_x \omega\big) \;+\; A(1-\omega)(p-I)_+^\gamma\Big]_+ \\ \partial_t p = K\, \partial_x\big(\omega^\alpha \partial_x p\big). \end{cases} \tag{6}$$

We will refer to it as the *1D formulation of the diatomite problem.* In spite of its simplification, such a formulation is traditional. Let us consider for example a huge number of wells placed along a straight line and operating simultaneously: this is customary for oil and water filtration problems, see the book [1] as well as the comprehensive book [3]. Therefore it might be useful to replace a discrete representation of the wells by a homogeneous distribution density of wells along the line. This representation is called *drainage gallery,* and it works quite fine as soon as we are not too close to the wells. The 1D formulation can very likely describe the case where a drainage gallery is orthogonal to the x axis.

4. Numerical examples for the two-dimensional formulation

In this section we investigate some numerical examples for the 2D formulation of the diatomite problem (4). Just to understand what happens in a very simple case, we present a first example in which a single well placed in the center of an ideal squared oil field injects fluid at constant pressure. This will help to capture the qualitative behavior of the pressure and damage evolution inside the diatomite stratum.

Subsequently, we simulate an oil field composed by five wells placed in diamond formation (five points scheme). In this formation four of them are located in the corners of a square and one is placed in the center of this square. This formation is frequently used in oil fields, where usually the wells placed in the corners are injectors, and the one in the center is a production well. We will simulate this situation, but we will also do the opposite, using the corner wells as production ones and the central one as injection well. For numerical simulations, we idealize the oil field as a square with vertices $(\pm L, \pm L)$. The wells are represented by circles of radius $L/10$. In the diamond formation, four of them are centered in $(\pm 0.4 \cdot L, \pm 0.4 \cdot L)$, the fifth is centered in the origin. For the computation we first rewrite the problem (4) in a nondimensional form. The spatial variables (x, y) are replaced by nondimensional ones: $(\eta, \zeta) \equiv (L^{-1}x,\ L^{-1}y)$. In the new coordinates, the gradient is $\tilde{\nabla} \equiv L\nabla$. We use the nondimensional pressure $P \equiv p/\tilde{p}$, where $\tilde{p}$ is a certain constant with the dimension of pressure. Finally, we define the nondimensional time $\theta = t\Lambda^2\,\tilde{p}^{\beta} L^{-2}$. If we choose $\tilde{p}$ such that $\tilde{p}^{\beta} = K/\Lambda^2$, we get

$$\begin{cases} \partial_\theta \omega \;=\; \Big[\tilde{\nabla}\big(\omega^{\mu}\,(P-\mathbb{I})_{+}^{\beta}\,\tilde{\nabla}\omega\big) \;+\; \tilde{a}\,(1-\omega)\,(P-\mathbb{I})_{+}^{\gamma}\Big]_{+}\,, \\ \partial_\theta P \;=\; \tilde{\nabla}\big(\omega^{\alpha}\tilde{\nabla}P\big) \end{cases} \tag{7}$$

Here, $\mathbb{I} = \dfrac{I}{\tilde{p}}$ and $\tilde{a} = \dfrac{A\,\tilde{p}^{\gamma-\beta}\,L^2}{\Lambda^2}$.

The pressure is initially everywhere equal to a constant $P_0 < \mathbb{I}$. When an injection well starts working, the pressure inside the corresponding circle is raised to a value $P_{\max} > \mathbb{I}$. In these examples the values of parameters are: $P_0 = 2$, $P_{\max} = 10$, $P_{\min} = 0$ and $\mathbb{I} = 5$.

In the circle corresponding to a production well, the pressure is lowered to a value $P_{\min} < P_0$. The initial damage is prescribed to be equal to 0 in the whole domain, except in the wells where it is equal to 1. We have chosen for these examples $\alpha = 2$, $\gamma = 5$, $\beta = \tilde{a} = 1$, $\mu = 0$, and we prescribed no flux boundary conditions on the sides of the square. Finally, we have discretized the side of the domain in 161 points, and in order to handle the nonlinearity and the degeneracy we have used an implicit finite difference scheme.

Example 4.1. The first example represents a single injection well placed in the center of a squared region. The idea is to demonstrate the sharp front that separates damaged and undamaged regions and that coincides with the pressure front. We remind to the reader that the formation of the sharp front separating perturbed and unperturbed regions is a well known feature for degenerate parabolic equations such as "porous medium equations" (see [1] and references therein). In the following, it will be clear that the degeneracies involved in system (7) make the problem much more complicated than standard ones. The simulation runs from $\theta = 0$ to $\theta = 0.6$. At time $\theta = 0$ the well starts working and the damaged region coincides with the one occupied by the well. Figure 3 shows pictures corresponding to $\theta = 0.09$ and $\theta = 0.6$.

Example 4.2. In this second example a production well is placed in the center of a square, while the injection wells are in the corners. The configuration is schematically represented in Figure 4.

The simulation runs from $\theta = 0$ to $\theta = 0.21$. At time $\theta = 0$, the injection wells start working, while the production well initially is at rest and starts working at time $\theta = 0.15$. The reason for this delay is that we wait until the damaging zone propagating from the injection wells reaches the production well. The results are represented in Figures 5–8. Figure 7 (bottom) shows the vector field of the velocity of the fluid inside the oilfield at $\theta = 0.21$. Figure 8 shows on the same graph the flux of fluid injected and extracted.

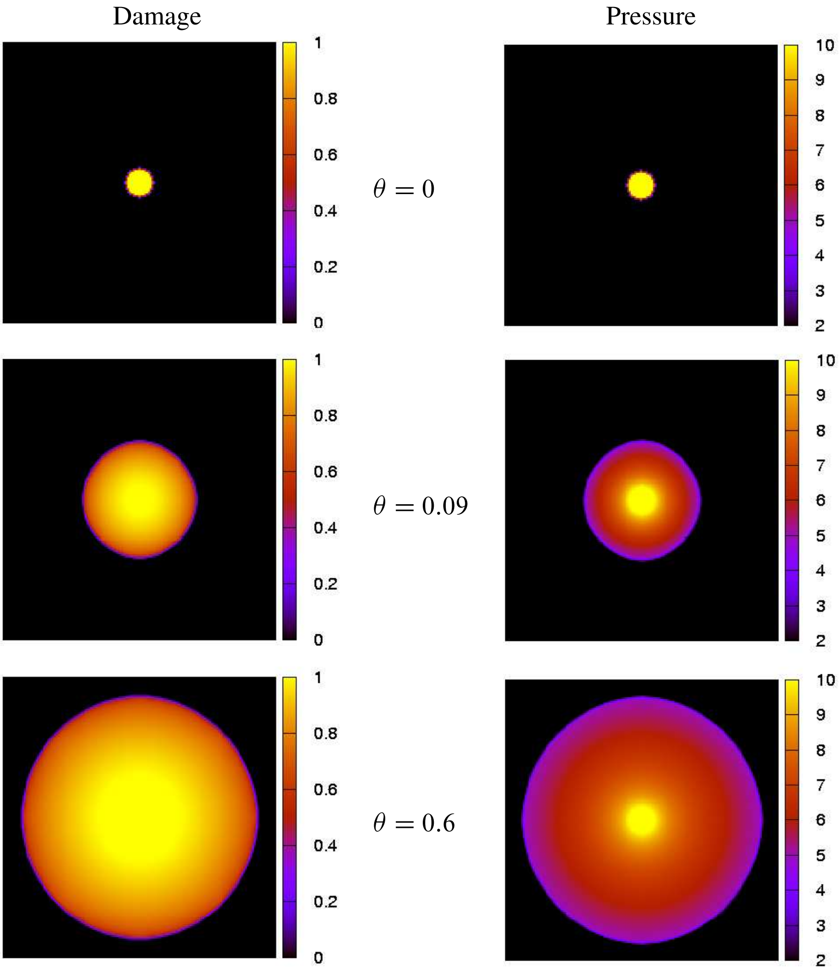


Figure 3. Extension of damage and pressure with increasing time θ.

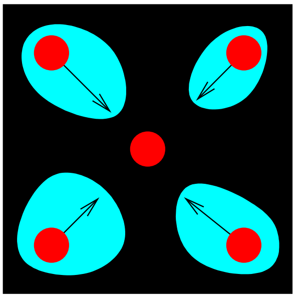


Figure 4. Schematic configuration of wells in the oil field. The center spot indicates the production well, the rest injection wells.

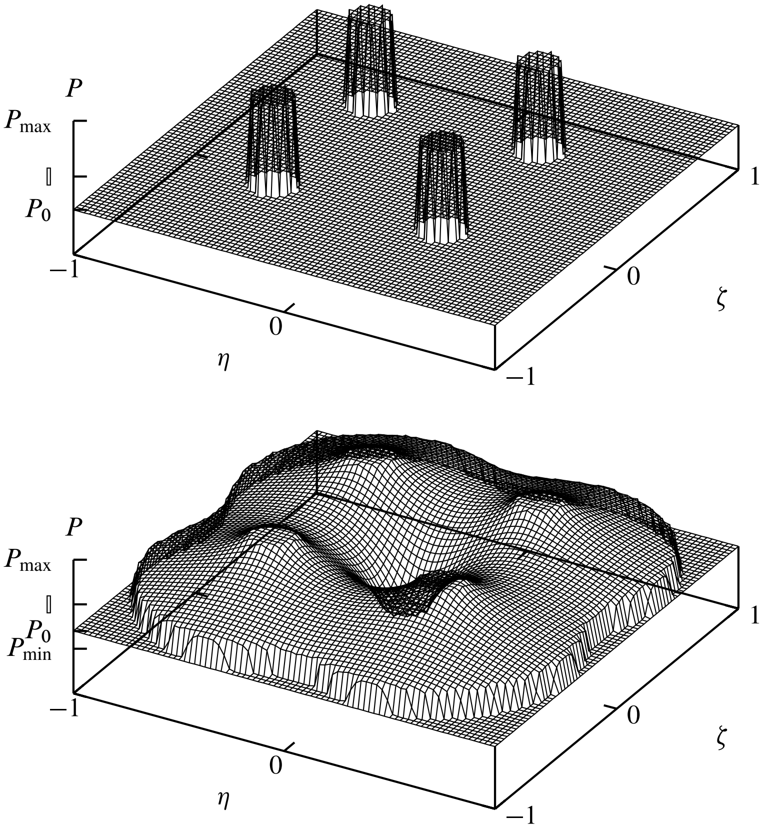


Figure 5. Initial pressure (top) and pressure at $\theta = 0.21$ (bottom). All the wells are active.

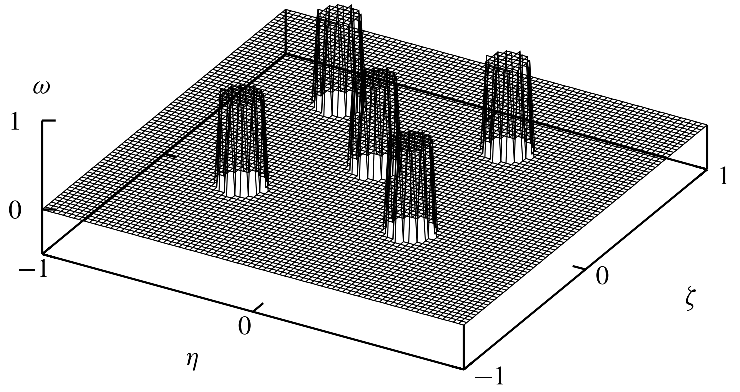
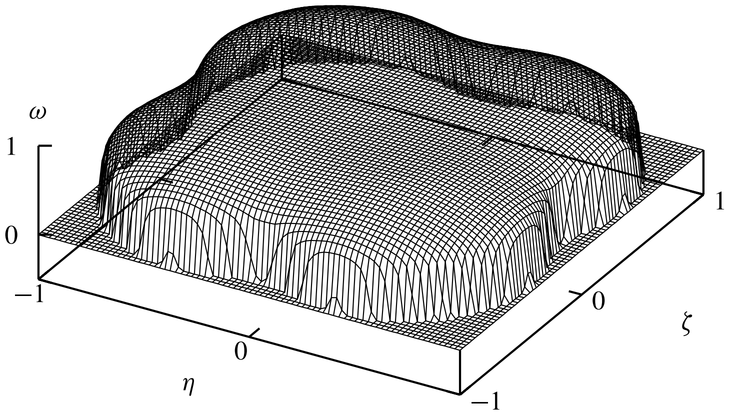


Figure 6. The initial damage ($\theta = 0$) is concentrated around the five wells.



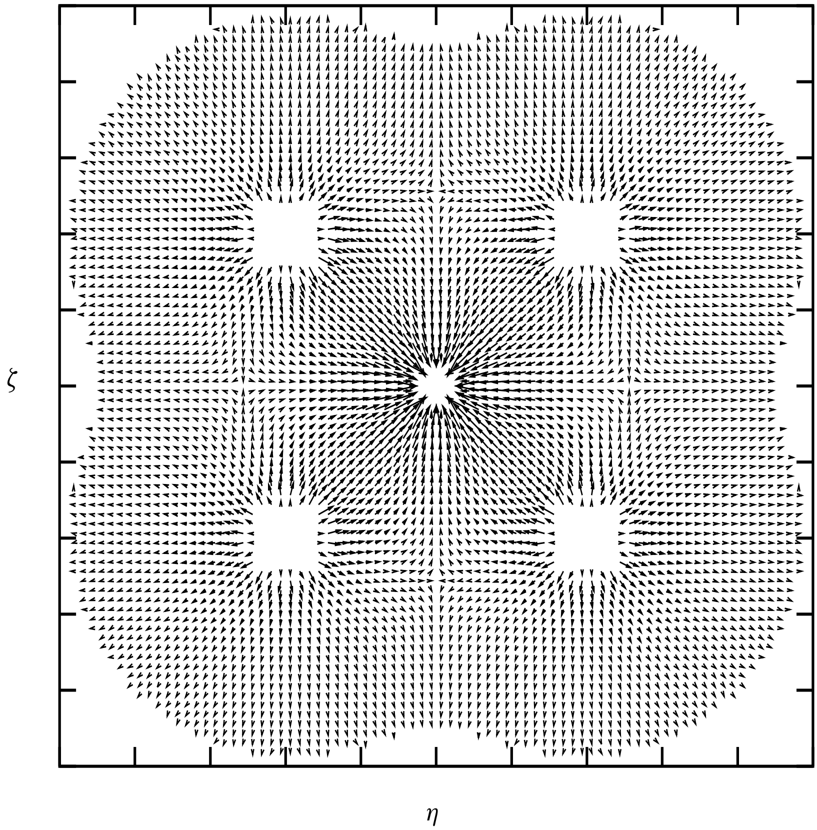


Figure 7. Damage (top) and fluid velocity field (bottom) at $\theta = 0.21$.

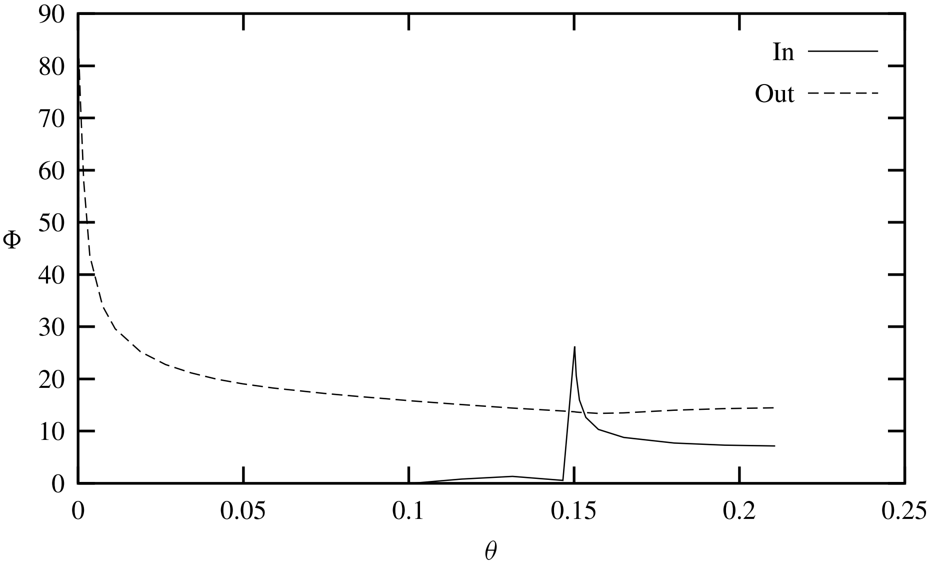


Figure 8. The total amount of fluid recovered per unit of time (Out), and the total amount of fluid injected per unit time (In), are plotted as functions of the nondimensional time θ. At $\theta \approx 0.1$ the damage front reaches the production well. At $\theta = 0.15$ the production well starts working.

Example 4.3. In this example, the well placed in the middle is an injection well, while those placed in the corners are production wells. The configuration is schematically represented in Figure 9.

The simulation runs from $\theta = 0$ to $\theta = 0.56$. At time $\theta = 0$, only the injection well is working, while the production wells start working at $\theta = 0.2$. The results

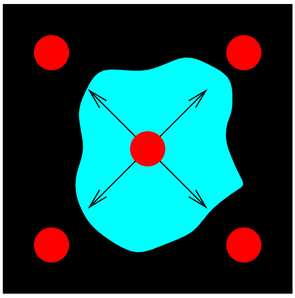


Figure 9. The schematic configuration of the wells in the oil field. The spot in the center corresponds to the injection well. The spots in the corners are the production wells.

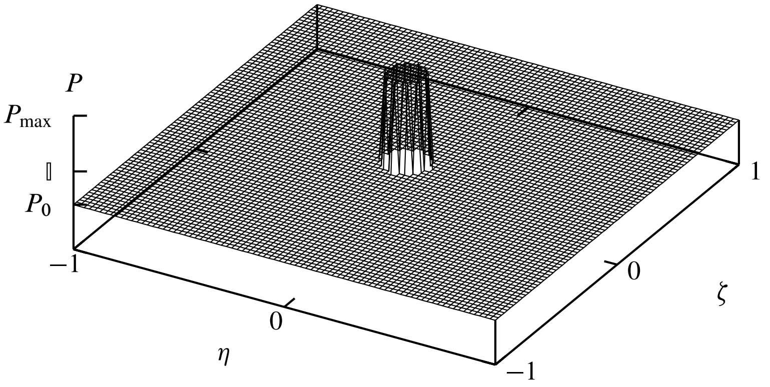


Figure 10. The initial pressure. Only the central well is active, and injecting water.

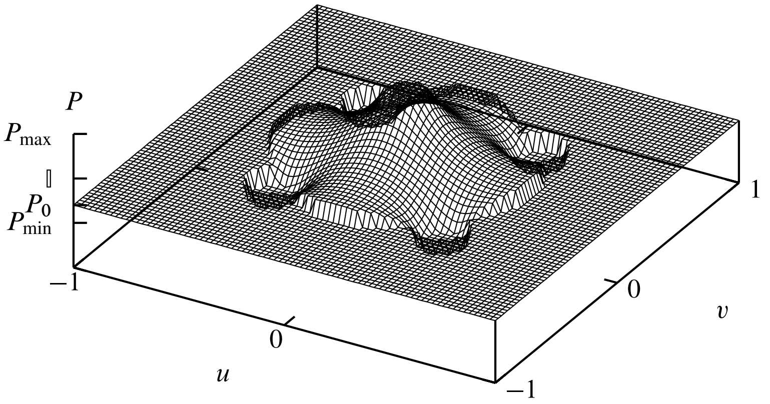


Figure 11. The pressure at $\theta = 0.56$ when all the wells are active.

are presented in Figures 10 through 13. In Figure 12 (bottom) we notice that the flux line starting from the injection well reaches the production wells. Actually, Figure 13 clearly shows that after a certain transition time the injected fluid flux is equal to the extracted fluid flux.

These simple examples are the first step of a numerical investigation of the 2D model. First of all, the numerical simulations confirm the conjecture that there is a sharp front between the damaged and undamaged zones. We emphasize again (see [1]) that such sharp fronts are a common feature of the degenerate parabolic equations of the type considered in the present work. In addition, in Example 4.3 we observe the main consequence of the threshold constant I. As expected, the front propagation stops when the pressure at the front goes below the value ▯; see Figure

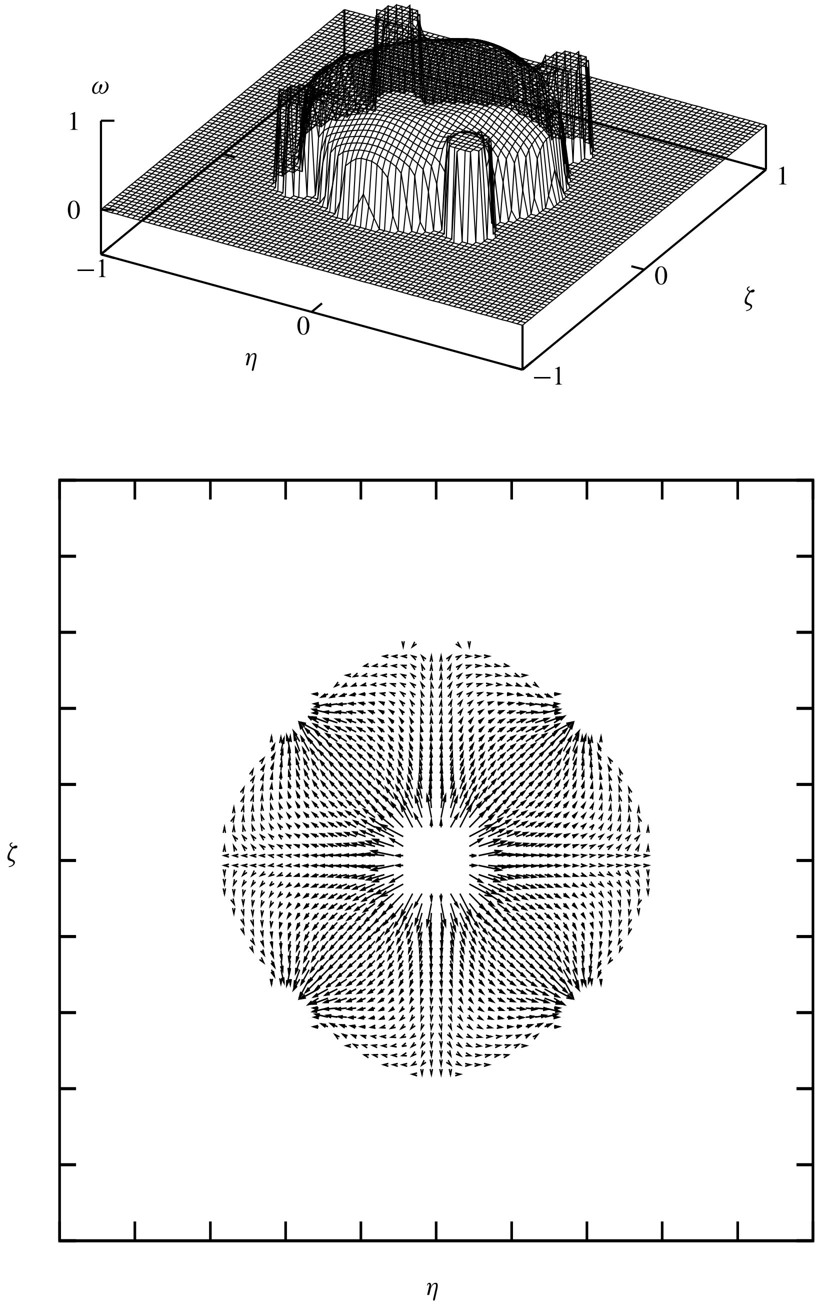


Figure 12. Damage (top) and fluid velocity field (bottom) at $\theta = 0.56$. The initial damage is the same as the one in Figure 6.

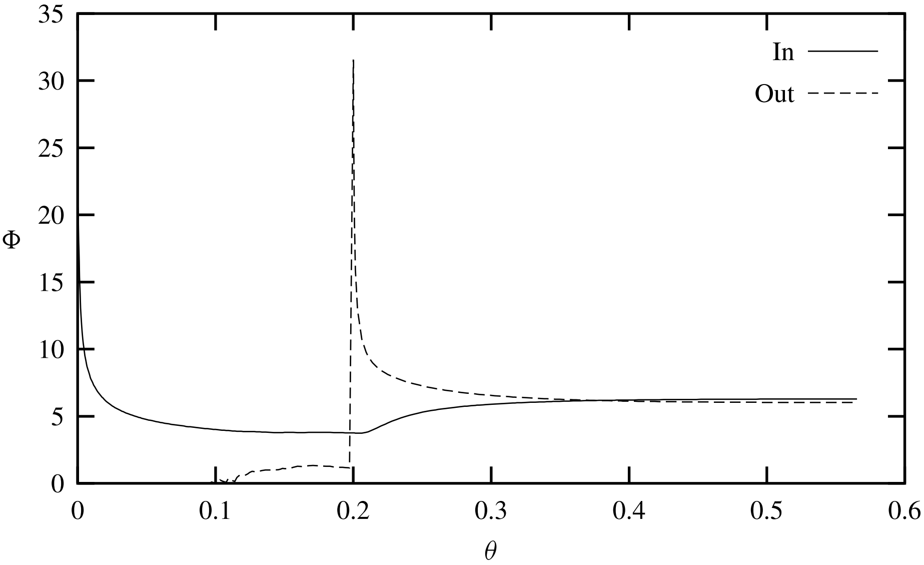


Figure 13. The total amount of fluid recovered per unit of time (Out), and the total amount of fluid injected per unit time (In), as functions of the nondimensional time θ. At $\theta \approx 0.1$ the damage front reaches the production wells. At $\theta = 0.2$ the production wells start working.

11. After a while the system seems to reach a stationary configuration, where both pressure and damage stop to evolve. As a consequence, all the fluid injected in the reservoir is caught by the production wells, as can be seen from Figures 12 and 13. So, from a mathematical point of view, the numerical simulations suggest that the 2D problem is well-posed, and that it has the structure of a free boundary problem, as was suggested in [2]. In this sense, the numerical experiments form a basis for further analytic studies. On the other hand, we are short of systematic quantitative studies of the microstructural changes of diatomites subjected to mechanical stress, and, in particular, the values of all the involved parameters are still unknown. Even merely a rough estimate of these parameters would make it possible to begin the simulation of real oilfields, a most important challenge for the future.

5. The one-dimensional formulation

In the following, we will deal with a flow to or from a drainage gallery described by the system (6). The numerical scheme adopted to solve this system is essentially the same as the one used for the 2D formulation. We begin our investigation by giving an overview of the behavior of the solutions. For this purpose we consider on the interval [0, 1] the system

$$\begin{cases} \partial_t \omega = \partial_x\big(\omega^{\mu}(p-I)_+^{\beta}\,\partial_x\omega\big) + A(1-\omega)(p-I)_+^{\gamma}, \\ \partial_t p = \partial_x\big(\omega^{\alpha}\,\partial_x p\big), \end{cases} \tag{8}$$

where, without loss of generality, we have fixed $\Lambda = K = 1$. We choose initial data

$$\omega_0(x) = 0 \quad \text{for } x \in [0,1]; \qquad p_0(x) = 0.1 \quad \text{for } x \in [0,1],$$

and boundary conditions

$$\partial_x\omega(0,t) = \partial_x\omega(1,t) = 0, \qquad p(0,t) = 1.1 \quad \text{and} \quad p(1,t) = 0.1.$$

We observe that the physical condition of positiveness of $\partial_t\omega$ expressed in (6) has been removed, since it is never violated for these sets of initial and boundary conditions. We expect a damage-pressure front propagating from the left to the right.

We solved the problem numerically for several values of α, β, γ, μ, I. We selected some specific values that exhibit different types of behavior of the corresponding solutions and illustrate strong dependence on the parameters. This strong dependence will become even clearer when we investigate the traveling waves (section Section 5.1).

In Figure 14 (top) we present the numerical approximations of the evolution of damage and pressure with $\alpha = 2$, $\beta = 0.5$, $\gamma = \mu = 1$ and $I = 0.8$. The simulation suggests that the damage is discontinuous across the free boundary. Moreover, the pressure seems to decrease linearly on the left side of the free boundary, towards a value which is reasonably close to the value $I \equiv 0.8$.

In the middle row of the same figure we selected instead $\alpha = 0.5$, $\beta = 2$, $\gamma = \mu = 1$ and $I = 0.6$, and the simulation suggests that ω goes down to zero smoothly near the free boundary.

Finally, in the bottom part of the figure we have chosen $\alpha = 2$, $\beta = 2$, $\gamma = \mu = 3$ and $I = 0.6$, and it seems that p jumps across the free boundary to a value greater than $I \equiv 0.6$, while ω seems to be continuous across the boundary.

Actually the numerical scheme is surprisingly stable near the free boundary, and preserves the sharpness of the boundary without smoothing it.

5.1. ***Traveling waves.*** A numerical investigation, the results of which were presented in the previous section, confirmed the conjecture that there exists a sharp front separating damaged ($\omega > 0$) and undamaged ($\omega = 0$) regions. Therefore, to investigate the *local structure* of such fronts (that is, the behavior of the flow characteristics close to the front), the studying of the traveling waves is appropriate. Naturally, the stretching of the horizontal coordinate is assumed, as it is always done when the structure of the fronts is considered (for example, the structure of

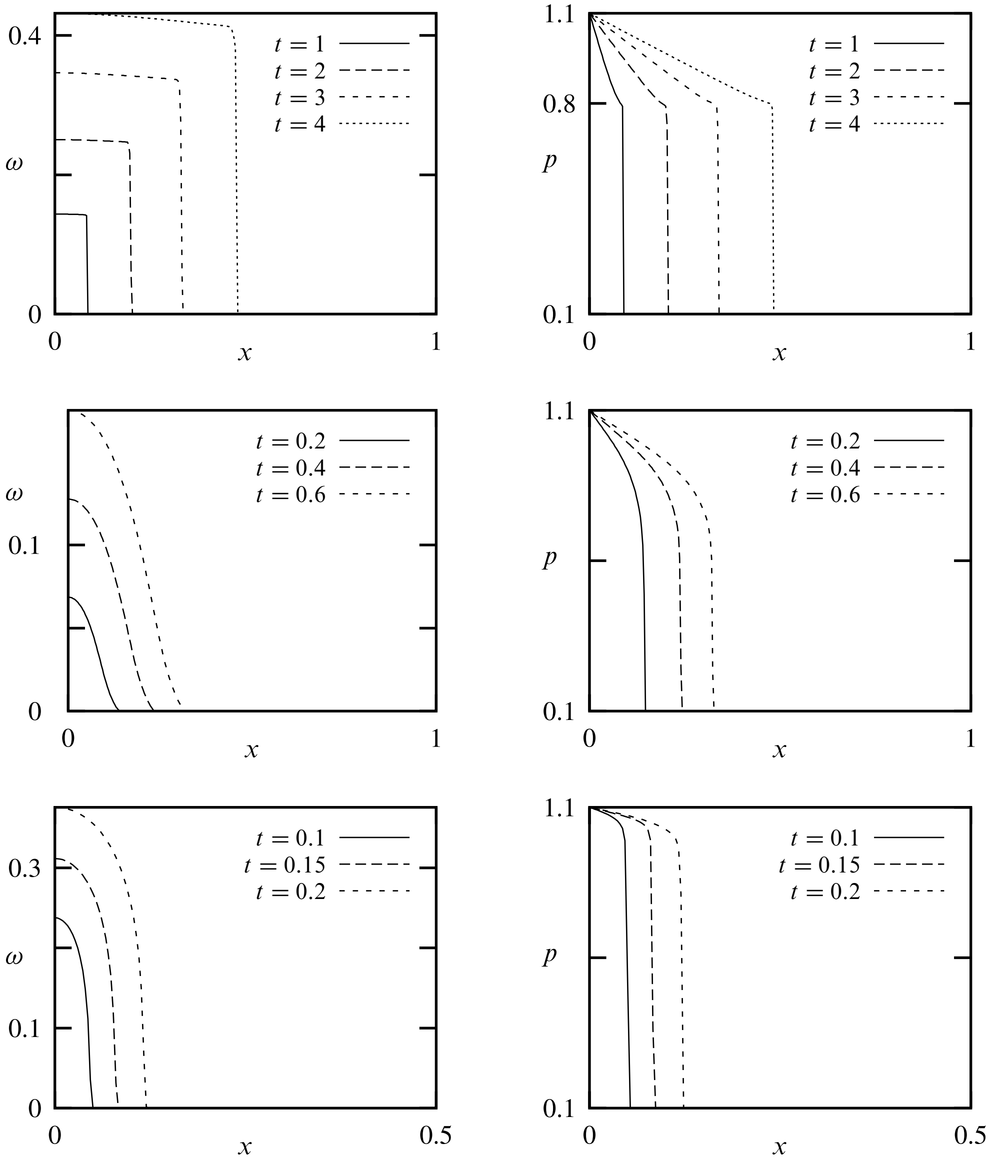


Figure 14. Numerical experiment. Left: damage; right: pressure. Top: $\alpha = 2,\ \beta = 0.5,\ \gamma = \mu = 1,\ I = 0.8$; middle: $\alpha = 0.5$, $\beta = 2,\ \gamma = \mu = 1,\ I = 0.6$; bottom: $\alpha = 2,\ \beta = 2,\ \gamma = 1,\ \mu = 3$, $I = 0.6$.

the shock waves in gas dynamics). We focus our attention to the case of moving fronts. The rigorous mathematical investigation of the moving traveling waves can be found in [4]. At first, we refer to [5] and present some general mathematical results concerning the existence of appropriate solutions for the system (8). In that paper, a solution (ω, p) is constructed in the case

$$\begin{aligned} &\omega_0 > 0,\ p_0 > 0 && \text{in an interval } (a,b),\\ &\omega_0 \ = \ p_0 \ = \ 0 && \text{in } (-\infty,a) \text{ and } (b,\infty). \end{aligned}$$

It is proved in [5] that there exist two fronts, $x = a(t)$ and $x = b(t)$ (where $a(t)$ and $b(t)$ are continuous functions with $a(0) = a,\ b(0) = b$) which separate the damaged and undamaged rocks:

$$\begin{aligned} &\omega(x,t) > 0,\ p(x,t) > 0 && \text{if } a(t) < x < b(t),\\ &\omega(x,t) \ = \ p(x,t) \ = \ 0 && \text{if } x < a(t) \text{ or } x > b(t). \end{aligned}$$

The product $\omega(p-I)_+$ is, generally speaking, continuous across the fronts at $t > 0$, which implies that at least one of the two functions ω and $(p-I)_+$ is continuous across the fronts at such times t. Moreover, if $p(x,0) > I$ in the interval (a,b), then $p(x,t) > I$ if $a(t) < x < b(t),\ t > 0$. We emphasize that still very little is known mathematically about the general behavior of the solutions near the free boundary, and about the dependence of the solutions on the various parameters in the problem—this will be the problem of our further studies. In particular, it is important to know if ω or $(p-I)_+$ can be discontinuous across the free boundaries, as was suggested by the numerical simulations, and, if so, for which parameters.

Therefore we look for traveling waves of constant speed V, that is, for the solutions to the system (8) of the type $p = p(\xi),\ \omega = \omega(\xi),\ \xi = x - Vt$. For definiteness' sake, we assume that V is positive, and that both p and ω vanish when $\xi > 0$, and that $0 < \omega < 1$ and $p > I$ when $\xi_0 < \xi < 0$ for some negative ξ_0. Under these assumptions, the system (8) leads to the system of ordinary differential equations

$$\begin{aligned} &-V\frac{d\omega}{d\xi} = \frac{d}{d\xi}\Big(\omega^\mu (p-I)^\beta \frac{d\omega}{d\xi}\Big) \ + \ A(1-\omega)(p-I)^\gamma && \text{if } \xi_0 \le \xi < 0,\\ &-Vp = \omega^\alpha \frac{dp}{d\xi} && \text{if } \xi_0 \le \xi < 0, \end{aligned}$$

with the properties that $0 < \omega < 1,\ p > I,\ d\omega/d\xi \ < \ 0$ if $\xi_0 \le \xi < 0$, and

$$\lim_{\xi \nearrow 0} \omega(p-I) \ = \ 0 \ = \ \omega^\mu (p-I)^\beta \frac{d\omega}{d\xi} \ + \ V\omega.$$

In [4] it has been shown that, for each choice of the values of the parameters and for each given value of the velocity $V > 0$, there exists a one-parameter family of solutions, and that the behavior of the traveling waves depends strongly on the values of the parameters. In particular, it has been shown that, for every $V > 0$, ω

can be discontinuous across the free boundary ($\xi = 0$) if and only if $0 < \beta < 1$ and $I > 0$. In such a case the traveling waves solutions behave near $\xi = 0$ as

$$\omega \approx \omega^* + \frac{V^{1-\beta}}{(1-\beta)I^\beta} \left(\omega^*\right)^{1+\alpha\beta-\mu} |\xi|^{1-\beta}, \qquad p \approx I + V\left(\omega^*\right)^{-\alpha} I\,|\xi|,$$

where $\omega^* > 0$ is a free parameter. These analytic results are in agreement with the numerical results of Figure 14(a).

The question is more complicated if we consider for which parameter values there exist traveling waves for which $(p - I)_+$ is discontinuous across the interface. If $\mu \le 1$ and $\alpha < \mu$, there again exists, for every velocity $V > 0$, a one-parameter family of traveling waves which, near the interface $\xi = 0$, behave as

$$\omega \approx B\,|\xi|^{1/\mu}, \qquad p \approx p^* + C|\xi|^{1-\alpha/\mu}.$$

Here $p^* > I$ is the free parameter, and B and C are constants determined by V, p^* and by the parameters in the equations. But, if $\mu \ge 1$ and $\alpha < 1$, there exists as well, for every $V > 0$, a two-parameter family of solutions which behave as

$$\omega \approx B_1|\xi|, \qquad p \approx p^* + C_1|\xi|^{1-\alpha}.$$

Here $p^* > I$ is one of the free parameters, and B_1 and C_1 are constants determined by V, p^* and by the parameters in the equations (but not by the second free parameter!).

The latter family has a remarkable property not satisfied by the former ones. Returning to the original variables of problem (6) we obtain new coefficients corresponding to B_1 and C_1, and it turns out that they do not depend on the parameter Λ. As a matter of fact they coincide with the coefficients of the traveling wave solution of (6) if we put $\Lambda = 0$. Indeed, a straightforward calculation shows that if $\Lambda = 0$ and $0 < \alpha < 1$, then for any $p^* > I$ and $V > 0$ problem (6) has a traveling wave solution which, near the interface $x = Vt$ $(x < Vt)$, behaves as

$$\omega \approx \frac{A(p^* - I)^\gamma}{V}\,|x - Vt|,$$
$$p \approx p^* \frac{V^{1+\alpha} p^*}{(1-\alpha)KA^a(p^* - I)^{\alpha\gamma}}|x - Vt|^{1-\alpha}.$$

We refer to [4] for a detailed discussion on families of traveling wave solutions.

5.2. ***Optimization of oil recovery, a qualitative example of control problem.*** In this section we consider an application of the 1D formulation (6) of the problem of fluid flow. Let the domain be the finite interval $0 \le x \le 1$. We are describing a flow to the “drainage galleries” placed in the section $x = 0$. We assume that such a drainage gallery can work both as an injection and a production one. This means that it can work both at higher and lower pressure than the initial pressure of the oil reservoir p_i. It is feasible to suppose that $p_i \leq I$, since the value of I is the threshold pressure above which the microcracking starts to accumulate, and in the pristine state there are no cracks in the rocks. The system can be controlled by prescribing the pressure at $x = 0$, that is, in the gallery: we denote this value by $P(t)$. To start to extract oil, some amount of water is pumped through the wells (P is raised above I) in the diatomaceous stratum. Microcracks will appear inside the stratum and consequently the permeability of the rock will increase. After this initial process of water pumping, the pressure P is lowered below p_i in order to extract oil from the wells.

The problem of finding the best strategy to maximize the amount of extracted oil, as it was formulated, is ill-posed. In principle, a very long (in time) pulse, or a very high pulse for the function $P(t)$, would allow to damage regions very far from the wells. Moreover, there is no limitation on the amount of oil that can be taken out from this damaged zone, if we are able to decrease the pressure enough and wait for a long time.

However, the real conditions are far from this idealization, and there are several aspects that we have to take into account. First of all, a realistic assumption is that the device which controls the pressure in the wells, works only in a certain bounded range of pressure. Moreover, the whole process of oil extraction has several aspects to be taken into account by constructing a suitable "cost function" C. Therefore, we make the following assumptions:

(1) There is a fixed cost k_1 per unit of time to maintain the gallery operating.

(2) The cost per unit of time for injecting or extracting fluid is proportional to the power (work per unit time)

$$k_2\big[-(P(t)-p_i)\,\Phi(t)\big]_+,$$

where p_i is the pressure of the wells at rest (we simplified the problem by considering the rest pressure in the wells equal to the initial pressure in the deposit), $\Phi(t)$ is the flux of liquid, and k_2 is a conversion factor.

(3) The profit $F(t)$ is assumed to be proportional to the volume of extracted liquid $E(t)$:

$$F(t) = k_3 E(t).$$

Therefore, the total cost is $C(t) = k_1 t + k_2 \int_0^t \big[-(P(s)-p_i)\Phi(s)\big]_+ \, ds$. On the other hand,

$$\begin{aligned} F(t) &= -k_3\Big(\int_0^t K\omega^\alpha(0,t)\,\partial_x p(0,t)\,dt\Big)_+ \\ &= k_3\Big(\int_0^1 p(x,t)\,dx - \int_0^1 p(x,0)\,dx\Big)_+ \equiv (V(t))_+. \end{aligned}$$

Here $V(t)$ is the difference between the actual total volume and the initial total volume of fluid in the reservoir. Hence, the net profit is

$$G(t) = F(t) - C(t). \tag{9}$$

Our problem is to find the function $P(t)$ (with $0 \le t \le T$) that realizes the maximum of $G(T)$. Here T is some instant of time that can be fixed a priori, or can be included in the unknowns of the problem. We repeat that our formulation of the control problem is a schematic one, and is used here only for presenting and illustrating the basic idea.

We still need to specify the set of allowed control functions $P(t)$. We introduce two numerical examples where we make strong assumptions on the possible shape of $P(t)$.

Example 5.1. We assume that the device that regulates the pressure can only perform a single cycle consisting of the following steps:

(1) *Injection and damaging.* Starting from the initial value of pressure p_i, the drainage gallery imposes a pulse of maximum amplitude $p_{\max}$;

(2) *Extraction.* After returning to the value p_i, the drainage gallery decreases the pressure linearly until a certain value $p_{\min}$ is achieved, and keeps this value constant as long as the incoming flux ensures a profit;

(3) *Back to the beginning.* The pressure increases linearly to its initial value p_i, so that the cycle is completed.

This cycle can be repeated as long as it is profitable. A cycle of $P(t)$ can be represented as follows (see Figure 15):

$$P(t) = \begin{cases} \big(p_{\max} - p_i\big)\ \sin\Big(\pi \dfrac{t}{t_1}\Big) + p_i & \text{if } 0 \le t \le t_1 \\ \big(-p_{\min} + p_i\big)\Big(\dfrac{t_2 - t}{t_2 - t_1}\Big)_+ + p_{\min} & \text{if } t_1 \le t \le t_3 \text{ and } t_2 < t_3 \\ \big(+p_{\min} - p_i\big)\Big(\dfrac{t_4 - t}{t_4 - t_3}\Big) + p_i & \text{if } t_3 \le t \le t_4. \end{cases} \tag{10}$$

The intervals t_1, $t_2 - t_1$, and $t_4 - t_3$ are fixed a priori, but the interval $t_2 - t_3$ is not known a priori and is chosen by the program to optimize the profit in the cycle. In particular we keep the pressure at the drainage gallery equal to $p_{\min}$ until the profit

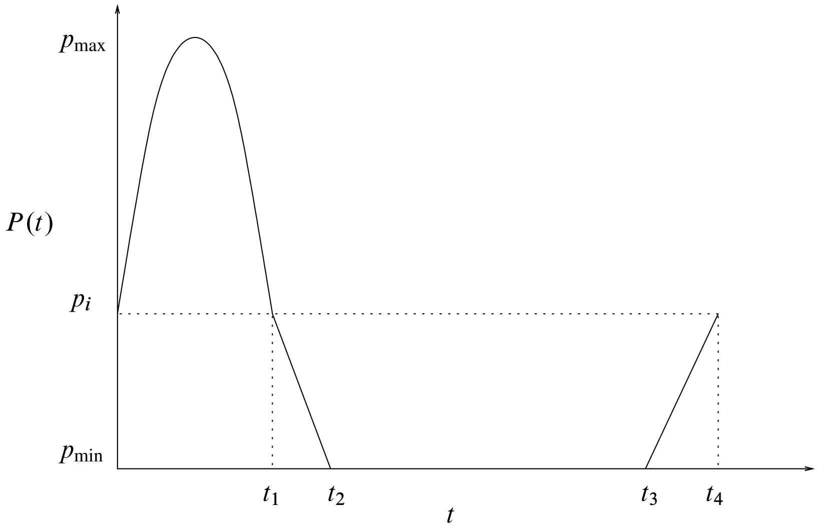


Figure 15. The cycle of the boundary pressure $P(t)$.

reaches a maximum ($G'(t_3) = 0$). Our choice is to optimize every single cycle, although, in case of several cycles, this is not necessarily the best global strategy.

For the numerical simulation we have chosen:

$$\begin{array}{ll} K = \Lambda = A = 1; & I = 0.2; \\ \alpha = \beta = 1; & p_i = 0.1,\ p_{\max} = 1,\ \text{and}\ p_{\min} = 0; \\ \gamma = 5; & t_1 = 0.2,\ t_2 - t_1 = t_4 - t_3 = 0.04. \\ \mu = 0 & \end{array}$$

To compute the cost function we have chosen:

$$k_1 = 0.01, \qquad k_2 = 1, \qquad k_3 = 20.$$

We have performed 5 cycles. As Figure 16 clearly shows, the volume extracted at the end of the cycles increases with the number of cycles performed. However, we have to take into account the cost function $C(t)$, represented in Figure 17. In fact, looking at the actual profit $G(t)$ in Figure 18, we conclude that the best strategy consists in performing only two cycles.

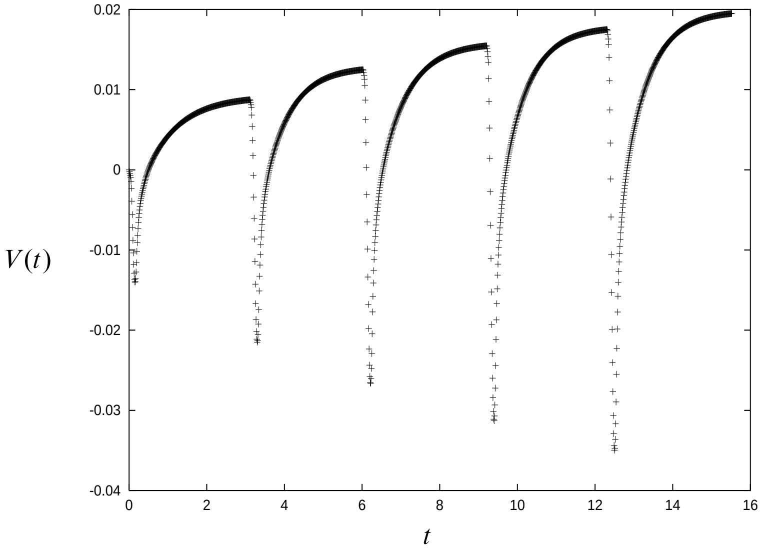


Figure 16. The volume function $V(t)$ is the difference between the volume extracted and injected. It increases with the number of pulses.

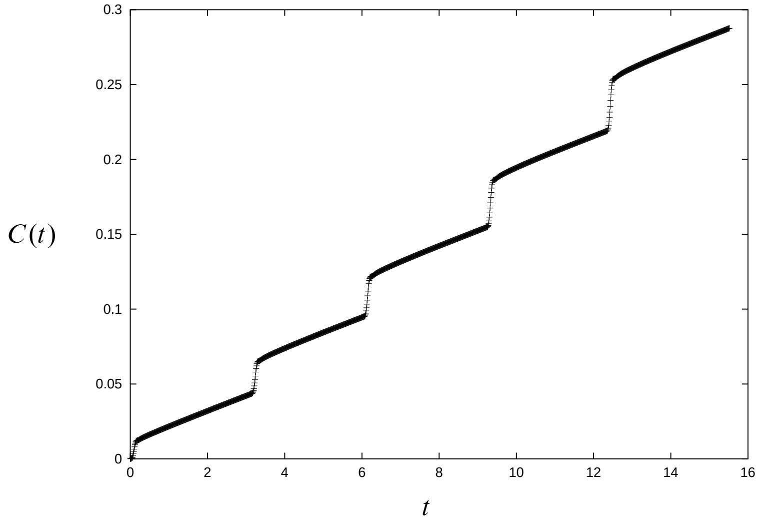


Figure 17. The cost function $C(t)$. The jumps correspond to fluid injections. During the injection, the main contribution is given by the integral term of the cost function. Between two jumps, instead, the cost increases almost linearly. This means that the main contribution is given by the time-linear term.

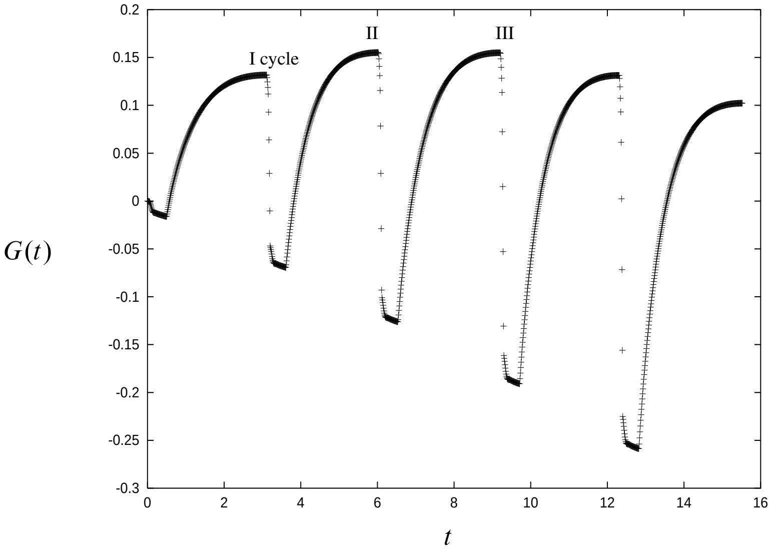


Figure 18. The profit function $G(t)$. The II pulse reaches the maximal profit, although the difference between the II and III pulses is very small (1.5511 versus 1.5496).

Example 5.2. We fix the total time T of the process and prescribe the form of the function $P(t)$ as in Figure 19:

$$P(t) = \begin{cases} \big(p_{\max} - p_i\big)\ \sin\Big(\pi \dfrac{t}{t_1}\Big) \ + \ p_i & \text{if } 0 \le t \le t_1, \\ \big(-p_{\min} + p_i\big)\Big(\dfrac{t_2 - t}{t_2 - t_1}\Big) \ + \ p_{\min} & \text{if } t_1 \le t \le t_2, \\ 0 & \text{if } t_2 \le t \le T. \end{cases}$$

In this example t_1 is the unknown parameter which we use to optimize the profit function $G(T)$. For the sake of simplicity we prescribe the duration of the interval $t_2 - t_1$ to be equal to $T/10$. For the numerical simulation we have chosen

$$\begin{aligned} &K = \Lambda = A = 1, && I = 0.2, \\ &\alpha = \beta = 1, && p_i = 0.1,\ p_{\max} = 1 \text{ and } p_{\min} = 0, \\ &\gamma = 5, && T = 3,\ t_2 - t_1 = 0.3. \\ &\mu = 0, && \end{aligned}$$

We have evaluated numerically the profit function G for several values of the time t_1. The results are shown in Figure 20.

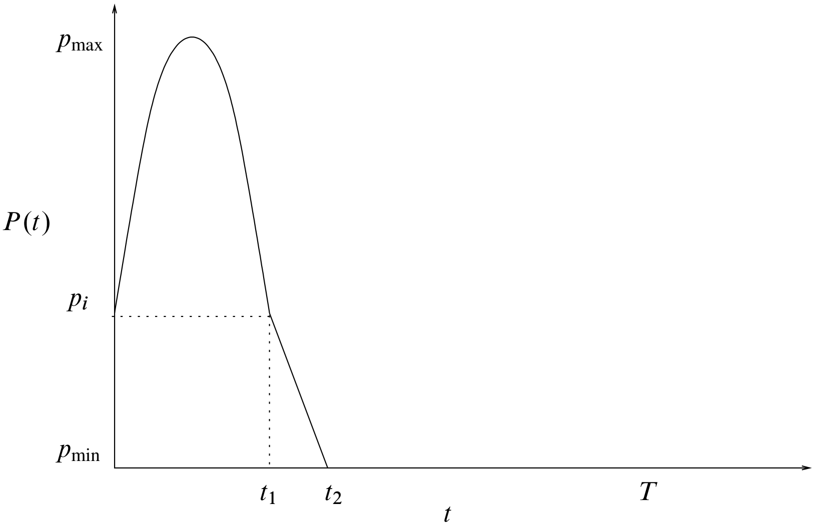


Figure 19. The boundary pressure $P(t)$.

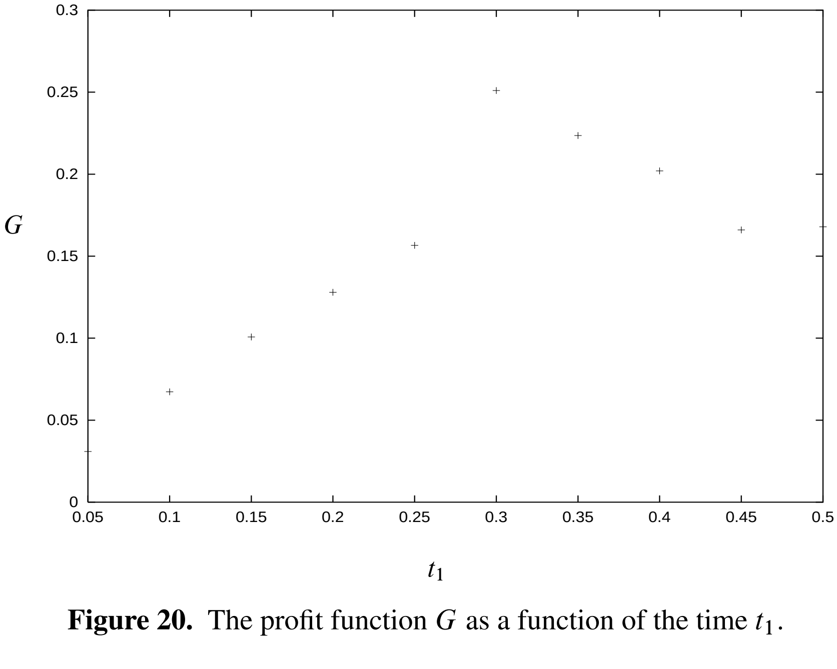


Figure 20. The profit function G as a function of the time t_1.

Acknowledgments. This work was supported in part by the Director, Office of Science, Computational and Technology Research, U.S. Department of Energy, under Contract no. DE–AC03–76SF00098. Nitsch thanks the Lawrence Berkeley National Laboratory, where he carried most of his research, for its great hospitality.

All of us express our deep gratitude to Professor A. J. Chorin for his interest in this work, valuable advice and help. Our discussions with Professor T. W. Patzek, Dr. V. M. Prostokishin and Dr. D. B. Silin were very important for us in performing this work, and we express our deep gratitude to them.

References

[1] G. I. Barenblatt, V. M. Entov, and R. V. M., *Theory of fluid flows through natural rocks*, Theory and applications of transport in porous media, vol. 3, Kluwer Academic Publishers, 1990.

[2] G. I. Barenblatt, T. W. Patzek, V. M. Prostokishin, and D. B. Silin, *SPE75230: Oil deposit in diatomites: A new challenge for subterranean mechanics*, SPE/DOE Improved Oil Recovery Symposium, 2002.

[3] J. Bear and Y. Bachmat, *Introduction to modeling of transport phenomena in porous media*, Kluwer Academic Publishers, 1991.

[4] M. Bertsch and C. Nitsch, *Traveling wave solutions of a nonlinear degenerate parabolic system from petroleum engineering*, preprint, 2006.

[5] M. Bertsch, R. Dal Passo, and C. Nitsch, *A system of degenerate parabolic nonlinear PDE's: a new free boundary problem*, Interfaces Free Bound. **7** (2005), no. 3, 255–276. MR 2006f:35296 Zbl 1079.35102

Received September 28, 2005.

GRIGORY ISAAKOVICH BARENBLATT: gibar@math.berkeley.edu
Department of Mathematics, University of California, Berkeley CA 94720

and

Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 50A/1148,
Berkeley, CA 94720, United States
http://math.lbl.gov/barenblatt/barenblatt.html

MICHIEL BERTSCH: m.bertsch@iac.cnr.it
Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche,
Viale del Policlinico, 137, I-00161 Roma, Italy

and

Dipartimento di Matematica, Universitá degli Studi di Roma "Tor Vergata", Via della Ricerca
Scientifica, I-00133 Roma, Italy

CARLO NITSCH: carlo.nitsch@dma.unina.it
Dipartimento di Matematica e Applicazioni "R. Caccioppoli", Universitá degli Studi di Napoli
"Federico II", Via Cintia, Monte S. Angelo, I-80126 Napoli, Italy
http://wpage.unina.it/c.nitsch/

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

ON THE SPECTRAL DEFERRED CORRECTION OF SPLITTING METHODS FOR INITIAL VALUE PROBLEMS

THOMAS HAGSTROM AND RUHAI ZHOU

Spectral deferred correction is a flexible technique for constructing high-order, stiffly-stable time integrators using a low order method as a base scheme. Here we examine their use in conjunction with splitting methods to solve initial-boundary value problems for partial differential equations. We exploit their close connection with implicit Runge–Kutta methods to prove that up to the full accuracy of the underlying quadrature rule is attainable. We also examine experimentally the stability properties of the methods for various splittings of advection-diffusion and reaction-diffusion equations.

1. Introduction

Dutt et al. [7] have introduced a method of spectral deferred correction which allows one to automatically increase the accuracy of a low order time-stepping method. Defect and/or deferred correction methods for initial value problems have been known for some time [23; 9]. The main innovation in [7] is the use of spectral integration on Gaussian quadrature nodes to construct the corrections. This avoids instabilities and conditioning problems associated with repeated differentiations. They show that if forward or backward Euler methods are used as the base scheme (2–5), stable and stiffly-stable methods of very high order result. More recently, Auzinger et al. [3; 2] have analyzed similar algorithms and suggested various improvements.

Our interest here is in the use of SDC methods in conjunction with operator and/or dimensional splitting to solve initial-boundary value problems for partial differential equations. In a series of papers [21; 6; 19] Minion et al. have explored the use splitting methods as the base scheme in an SDC approach. However, their implementations have mainly been designed to achieve an order of accuracy

MSC2000: 65L06, 65M20.

Keywords: splitting methods, deferred correction, stability regions.

Hagstrom was supported in part by NASA Contract NRA-96-LeRC-2, NSF Grants DMS-9600146, DMS-9971772, DMS-0306285, Lawrence Livermore National Laboratory Subcontract B547968, and by ICOMP, NASA Glenn Research Center. Any conclusions or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of NSF, LLNL, or NASA.

Zhou was supported in part by NASA Grant NRA-96-LeRC-2 and NSF Grant DMS-0308019.

approximately equal to the number of time levels stored. (For an exception see [17].) In our approach we exploit the close connection of SDC methods with implicit Runge-Kutta methods to prove that up to double this accuracy is attainable, though only for the approximate solutions at the boundaries of each correction interval. (See also [3].) Again this result is expected from the perspective of the Runge-Kutta methods as the approximate solutions on the interior nodes correspond to the Runge-Kutta stage variables. We believe this possibility for enhanced accuracy is of importance for large, memory-bound applications; our experiments indicate that the efficiency of the proposed higher-order methods is essentially the same as for the methods proposed in the above-cited works.

We also examine the stability properties of the methods for various special cases. We begin by constructing stability domains for the standard model of operator splitting applied to advection-diffusion problems (e.g., [1]). We have considered all of the typical quadratures (Gauss–Legendre, Gauss–Lobatto, and Gauss–Radau) and a variety of starting methods (2–4). We also compare consistent and inconsistent correction methods (2–5). Second, we consider what we call preconditoned splitting methods for both linear and nonlinear problems. We have used such techniques to develop fourth and higher order solvers for complex models of reacting gases [12; 11; 24]. We note that Layton and Minion [20] have carried out an extensive stability study for SDC applied to splitting methods. We will compare our results to theirs, in essence assessing the effect of our additional corrections on the stability domains.

Finally, we verify the properties of the methods in nonlinear settings through experiments with simpler reaction-diffusion and advection-diffusion problems, focusing on the requirements on the preconditioner to obtain good accuracy and stability. (A similar study for lower order splitting methods is presented in [22].) Given the difficulties in fully analyzing splitting methods for complex problems, such studies seem necessary to validate any proposed methods.

2. Spectral deferred correction with splitting

We consider the initial-value problem:

$$\frac{du}{dt} = F(u,t), \quad u(t_0) = u_0, \quad u, F \in \mathbb{R}^{\kappa} \tag{2–1}$$

and recall the well-known fact that given $t_0 = T_0 < T_1 < \cdots$ (2–1) can be reformulated as a sequence of integral equations:

$$u(t) = u(T_j) + \int_{T_j}^{t} F(u(\tau),\tau)d\tau, \quad t \in [T_j, T_{j+1}]. \tag{2–2}$$

Our formulation of spectral deferred correction of some splitting method for approximating (2–1) has three essentially independent ingredients. First we introduce two splittings of F:

$$F = \tilde{F}_I + \tilde{F}_E, \quad F = F_I + F_E, \tag{2–3}$$

and two associated time-stepping formulas; a pth order multistep method (e.g., an IMEX method [1]) which we call the *starting method*:

$$\begin{aligned}\sum_{j=-1}^{k-1} \alpha_j v(t_{n-j}) &= h_n \sum_{j=-1}^{k-1} \beta_j^{(n)} \tilde{F}_I(v(t_{n-j}), t_{n-j}) \\ &\quad + h_n \sum_{j=0}^{k-1} \tilde{\beta}_j^{(n)} \tilde{F}_E(v(t_{n-j}), t_{n-j}),\end{aligned} \tag{2–4}$$

and the first order method,

$$v(t_{n+1}) = v(t_n) + h_n F_I(v(t_{n+1}), t_{n+1}) + h_n F_E(v(t_n), t_n), \tag{2–5}$$

which we call the *correction method.* Second, we introduce, as in [7], a collocation method for approximating (2–2). Setting $\Delta T = T_{j+1} - T_j$ we introduce m nodes:

$$t_{jk} = T_j + c_k \Delta T, \quad 0 \le c_1 < c_2 < \cdots < c_m \le 1. \tag{2–6}$$

A solution of the polynomial collocation approximation defined by these nodes is a set of vectors v_{jk} satisfying:

$$\begin{aligned} v_{jk} &= v(T_j) + \int_{T_j}^{T_j + c_k \Delta T} \psi_j(t) dt, && (2–7) \\ &= v(T_j) + \Delta T \sum_{\alpha=1}^{m} S_{k\alpha} F(v_{j\alpha}, T_j + c_\alpha \Delta T)), && (2–8) \end{aligned}$$

where $\psi_j(t)$ is the unique degree-($m-1$) interpolant of the data

$$(T_j + c_k \Delta T, F(v_{jk}, T_j + c_k \Delta T)), \quad k = 1, \ldots, m.$$

Here, following [7], we note that the matrix S whose entries are $S_{k\alpha}$ is a well-conditioned $m \times m$ spectral integration matrix.

The evolution of the approximate solution from T_j to T_{j+1} now proceeds as follows:

i: Compute approximations, v_{jk}^0, using m steps of (2–4) with the appropriate reduced time steps,

$$h_k = (c_k - c_{k-1}) \Delta T, \quad c_0 = 0. \tag{2–9}$$

(Note: a multistep starting method may make use of data at points $t_{j-1,k}$.)

ii: Given our lth approximation, v^l_{jk}, we define residuals, r^l_{jk}, using the spectral integration matrix, S:

$$r^l_{jk} = v(T_j) + \Delta T \sum_{\alpha=1}^{m} S_{k\alpha} F(v^l_{j\alpha}, t_{j\alpha}) - v^l_{jk}, \tag{2–10}$$

$$= v(T_j) + \int_{T_j}^{T_j+c_k\Delta T} \psi^l_j(t)dt - v^l_{jk}.$$

Here $\psi^l_j(t)$ is the unique degree-$(m-1)$ interpolant of the data

$$(T_j + c_k\Delta T, F(v^l_{jk}, T_j + c_k\Delta T)), \quad k = 1, \ldots, m.$$

iii: With the residual in hand, (2–5) is used to update the approximation. The idea here is to write $v^{l+1} = v^l + \delta^l$ and note that the correction can be viewed as an approximate solution to the perturbed equation:

$$\frac{d\delta^l}{dt} = F(v^l + \delta^l, t) - F(v^l) + \frac{dr^l}{dt}, \;\; \delta^l(T_j) = 0. \tag{2–11}$$

The most straightforward approach, used by Dutt et al. [7] and Minion [21], is to apply (2–5) directly to (2–11) to obtain the correction formula:

$$\begin{aligned} \delta^l_{jk} &= \delta^l_{j,k-1} + r^l_{jk} - r^l_{j,k-1} \\ &+ h_k \left(F_I(v^l_{jk} + \delta^l_{jk}, t_{jk}) - F_I(v^l_{jk}, t_{jk})\right) \\ &+ h_k \left(F_E(v^l_{j,k-1} + \delta^l_{j,k-1}, t_{j,k-1}) - F_E(v^l_{j,k-1}, t_{j,k-1})\right), \end{aligned} \tag{2–12}$$

$$\delta^l_{j0} = 0, \tag{2–13}$$

$$v^{l+1}_{jk} = v^l_{jk} + \delta^l_{jk}. \tag{2–14}$$

iv: Stop the process after L steps and define the solution update by:

$$v(T_{j+1}) = v(T_j) + \int_{T_j}^{T_{j+1}} \psi^L_j(s)ds. \tag{2–15}$$

Obviously, the description above leaves room for a wide range of implementations, some of which we will discuss below.

Concerning the starting and correction methods, we note that in many cases it is possible to choose F_I and/or $\tilde{F}_I$ to be linear in v, in which case the methods are called linearly implicit. Also we assume that $p \leq m$ where m is the number of nodes in the quadrature formula underlying the correction process. Although we will prove that the overall method typically attains an order $q > m$, our analysis indicates that there is no benefit to choosing $p > m$. Moreover, as multistep starting methods may use stage values, we are limited by the stage order, which is m. We emphasize that we could use the correction method as our starting method, and indeed that is what has been done in the references mentioned herein. Also, as the time steps will not be equally spaced, the coefficients in (2–4) will depend on n if a truly multistep method is used. (See, though, [3; 2] for a method which allows an equispaced temporal grid while maintaining the accuracy of the Gaussian quadrature rules.) Lastly, it is possible to replace (2–5) and/or (2–4) with a multisplitting or fractional step scheme as in [6], but for simplicity we will focus on the simpler splittings for our analysis here.

Concerning the nodes, Dutt et al. [7] take them to be the Gauss–Legendre nodes. Minion [21], on the other hand, suggests Gauss–Lobatto nodes, and we will also consider right-handed Gauss–Radau nodes. In [20] uniform nodes are shown to be feasible from the standpoint of stability, but their use would preclude the attainment of the higher order accuracy which is a focus of the current work.

Lastly we note that alternative correction formulas are also possible. The correction method we have employed in [12; 11; 24; 25] follows [9]:

$$\bar{v}^l_{jk} = \bar{v}^l_{j,k-1} + \delta t_k F_I(\bar{v}^l_{jk}, t_{jk}) + \delta t_k F_E(\bar{v}^l_{j,k-1}, t_{j,k-1}) + r^l_{j,k-1} - r^l_{jk}. \tag{2–16}$$

where

$$\bar{v}^l_{j0} = v(T_j). \tag{2–17}$$

Then set:

$$v^{l+1}_{jk} = v^0_{jk} + v^l_{jk} - \bar{v}^l_{jk}. \tag{2–18}$$

However, the theoretical results in this paper only apply in general to corrections based on (2–12)-(2–14). To apply them to corrections based on (2–16) we must make the starting method and the correction method coincide, which is the case in [12; 11; 24; 25]. Under those conditions, and for linear problems, the two correction methods are mathematically identical.

3. Order of accuracy

In [7; 21] the correction process is carried out until an error on the order of the truncation error in the approximations to (2–2) is attained. This leads to methods of order m. More general analyses of the convergence and accuracy of this process appear in recent manuscripts by Hansen and Strain [15; 16], where both single step and multistep correction formulas are considered. However, their approach does not take account of the full order of accuracy of the underlying quadrature rules which is our aim here. (We note that we could use SDC methods as analyzed in [15; 16] as our starting methods.)

In this work we focus on the classical Gauss-type quadrature methods which are of orders between $2m-2$ and $2m$. We prove here that if further corrections are made the order of accuracy of the underlying quadrature rule is in fact attainable, albeit only for the approximate solutions at the coarse grid points, T_j. (See also [3].) In practice, this allows the construction of higher order methods which are more efficient from the perspective of the number of time levels which must be stored. We also see that while the order of accuracy of the starting method affects the number of corrections needed, the accuracy of the correction method does not. This is in contrast with the results of [15; 16], which show that gains in accuracy commensurate with the order of the correction method are possible until an order m method is produced.

The accuracy result follows from the observation that if the residuals were zero, that is if the related collocation approximations to (2–2) were constructed, then the method would be equivalent to a standard implicit Runge–Kutta method (e.g., [13, Chapter II]), which has the accuracy we claim. Thus we need only show that similar conclusions follow from making these residuals sufficiently high order. We note that one could more directly use the size of the residual as a basis for terminating the correction process, as suggested in [17], but we do not consider that possibility here.

To study the local truncation error we assume $v(T_j)=u(T_j)$ and, for a multistep starting scheme, $v_{j-1,k}=u(t_{j-1,k})$. Set:

$$V^l(t)=u(T_j)+\int_{T_j}^{t}\psi_j^l(s)ds, \tag{3–1}$$

noting that $v(T_{j+1})=V^L(T_{j+1})$. First we prove:

Lemma 3.1. *Suppose F is smooth and that the polynomial quadrature rule based on* (2–6) *has order q. Then there exists a constant, C, independent of ΔT, such that for a sufficiently smooth solution, u, and a sufficiently smooth solution of the SDC method, V^l,*

$$|u(T_{j+1})-V^l(T_{j+1})|\le C\Delta T\max_k|r_{jk}^l|+O(\Delta T^{q+1}).$$

Proof. Define the defect, $d(t)$, by:

$$d(t)=\frac{dV^l}{dt}-F(V^l(t),t)=\psi_j^l(t)-F(V^l(t),t). \tag{3–2}$$

Note that by (2–10):

$$V^l(t_{jk})-v_{jk}^l=r_{jk}^l, \tag{3–3}$$

and by definition

$$\psi_j^l(t_{jk})=F(v_{jk}^l,t_{jk}). \tag{3–4}$$

Thus by the Lipschitz continuity of F:

$$|d(t_{jk})| \le K|r^l_{jk}|. \tag{3–5}$$

Let the matrix $\Phi(t,\tau,V^l(\tau))$ be defined by:

$$\Phi = D_{V^l(\tau)}w, \tag{3–6}$$

where for $t > \tau$:

$$\frac{dw}{dt} = F(w,t), \quad w(\tau) = V^l(\tau). \tag{3–7}$$

We note that standard results on the differentiability of solutions of ordinary differential equations with respect to their initial data imply that the derivatives of Φ with respect to τ can be bounded in terms of the derivatives of V^l which we have assumed (and will subsequently prove) to be bounded independent of l and ΔT. We then have the following nonlinear variation-of-constants formula, known as the Alekseev–Gröbner Lemma [13, Chapter I]:

$$V^l(T_{j+1}) - u(T_{j+1}) = \int_{T_j}^{T_{j+1}} \Phi(T_{j+1},\tau,V^l(\tau))d(\tau)d\tau. \tag{3–8}$$

Now replace the integral by the quadrature rule associated with the nodes. We have:

$$\begin{aligned} V^l(T_{j+1}) - u(T_{j+1}) = \Delta T \sum_k \omega_k \Phi(T_{j+1},t_{jk},V^l(t_{jk}))\cdot d(t_{jk}) \\ + \int_{T_j}^{T_{j+1}} \Phi(T_{j+1},\tau,V^l(\tau))d(\tau)d\tau \\ -\Delta T \sum_k \omega_k \Phi(T_{j+1},t_{jk},V^l(t_{jk}))\cdot d(t_{jk}). \end{aligned} \tag{3–9}$$

Using (3–5) the first term is bounded by $C\Delta T \max|r^l_{jk}|$ while the difference of the second and third is $O(\Delta T^{q+1})$ by our assumption on the accuracy of the quadrature rule and the smoothness of V^l. This completes the proof of Lemma 3.1. □

To complete our analysis, we need only prove that the residual is reduced by the correction process and that the approximate degree-m polynomial solution, V^l, has derivatives bounded independent of the time step.

Lemma 3.2. *For a sufficiently smooth solution*, u, *and smooth functions*, F, F_I, F_E, $\tilde{F}_I$, $\tilde{F}_E$, *a starting method of order* $p \le m$, *and corrections based on* (2–12)-(2–14), *there exist constants*, C_l *and* $M_{r,l}$, *independent of* ΔT *sufficiently small such that the residuals satisfy*:

$$\max_k |r^l_{jk}| \le C_l \Delta T^{p+1+l}, \tag{3–10}$$

and, for $l \geq m-p-2$ *and* $0 \leq r \leq m$:

$$\max_{t\in[T_j,T_{j+1}]} \left| \frac{d^r V^l}{dt^r} \right| \leq M_{r,l}. \tag{3–11}$$

Proof. Denote by $\tilde{V}(t)$ the exact degree m polynomial solution of the collocation equations. We have by [13, Theorem 7.10]:

$$\frac{d^r u}{dt^r} - \frac{d^r \tilde{V}}{dt^r} = O(\Delta T^{m+1-r}), \quad 0 \leq r \leq m. \tag{3–12}$$

Also, the residual satisfies:

$$\begin{aligned} r^l_{jk} &= \int_{T_j}^{t_{jk}} (\psi^l_j(s) - \tilde{V}'(s))ds + \tilde{V}(t_{jk}) - v^l_{jk} \\ &= \Delta T \sum_{\alpha=1}^{m} S_{k\alpha}(F(v^l_{j\alpha}, t_{j\alpha}) - F(\tilde{V}(t_{j\alpha}), t_{j\alpha})) + \tilde{V}(t_{jk}) - v^l_{jk}. \end{aligned} \tag{3–13}$$

We proceed by induction on l. For $l = 0$ (3–10) follows directly from the consistency of (2–4). Precisely, since $p \leq m$, $u(t_{jk}) - v^0_{jk} = O(\Delta T^{p+1})$, (3–12) implies $\tilde{V}(t_{jk}) - v^0_{jk} = O(\Delta T^{p+1})$ with (3–10) following from the Lipschitz assumptions.

Denoting by R^l the residual vector,

$$R^l = \begin{pmatrix} r^l_{j1} \\ \vdots \\ r^l_{jm} \end{pmatrix} \in \mathbb{R}^{m\kappa}, \tag{3–14}$$

we recast the correction process as a fixed point iteration:

$$R^{l+1} = C(R^l), \tag{3–15}$$

and analyze the Jacobian derivative $D_R C(0)$. From (3–13) and (2–14) we have:

$$r^{l+1}_{jk} = r^l_{jk} - \delta^l_{jk} + \Delta T \sum_{\alpha=1}^{m} S_{k\alpha}(F(v^l_{j\alpha} + \delta^l_{j\alpha}, t_{j\alpha}) - F(v^l_{j\alpha}, t_{j\alpha})). \tag{3–16}$$

Taking the difference of (3–16) for consecutive values of k and using (2–12) we arrive at the formula:

$$\begin{aligned} r^{l+1}_{jk} = r^{l+1}_{j,k-1} &+ \Delta T \sum_{\alpha=1}^{m} (S_{k\alpha} - S_{k-1,\alpha})(F(v^l_{j\alpha} + \delta^l_{j\alpha}, t_{j\alpha}) - F(v^l_{j\alpha}, t_{j\alpha})) \\ &- h_k \left(F_I(v^l_{jk} + \delta^l_{jk}, t_{jk}) - F_I(v^l_{jk}, t_{jk}) \right) \\ &- h_k \left(F_E(v^l_{j,k-1} + \delta^l_{j,k-1}, t_{j,k-1}) - F_E(v^l_{j,k-1}, t_{j,k-1}) \right), \end{aligned} \tag{3–17}$$

where $S_{0\alpha} = 0$. Set:

$$G_{kk'} = D_{r^l_{jk'}} r^{l+1}_{jk}, \quad H_{kk'} = D_{r^l_{jk'}} \delta^l_{jk}, \tag{3–18}$$

where the derivatives are evaluated at $R^l = 0$. Note that $G_{kk'}, H_{kk'} \in \mathbb{R}^{\kappa\times\kappa}$ with $G_{kk'}$ being the block entries of $D_R C(0)$. Noting that $\delta^l_{jk} = 0$ if $R^l = 0$ we have from (2–12):

$$\begin{aligned} H_{kk'} &= H_{k-1,k'} + \varepsilon_{kk'} I - \varepsilon_{k-1,k'} I \\ &\quad + h_k \Big(D_u F_I(\tilde V(t_{jk}), t_{jk}) H_{kk'} + D_u F_E(\tilde V(t_{j,k-1}), t_{j,k-1}) H_{k-1,k'} \Big), \end{aligned} \tag{3–19}$$

$$H_{0k'} = 0. \tag{3–20}$$

(Here we are using ε_{ij} to denote the Kronecker δ to avoid confusion with the correction vector.) Combining (3–19) with (3–20) and solving in increasing k we conclude that $H_{kk'} = O(1)$. Moreover,

$$H_{kk'} = 0, \quad k < k'. \tag{3–21}$$

(This fact proves to be a barrier to accelerating convergence; see the remark below.) Differentiating (3–17) on the other hand we find:

$$\begin{aligned} G_{kk'} &= G_{k-1,k'} + \Delta T \sum_{\alpha=1}^{m} (S_{k\alpha} - S_{k-1,\alpha}) D_u F(\tilde V(t_{j\alpha}), t_{j\alpha}) H_{\alpha k'} \\ &\quad - h_k \left(D_u F_I(\tilde V(t_{jk}), t_{jk}) H_{kk'} + D_u F_E(\tilde V(t_{j,k-1}), t_{j,k-1}) H_{k-1,k'} \right), \end{aligned} \tag{3–22}$$

$$G_{0k'} = 0. \tag{3–23}$$

Solving (3–22) it is clear that $G_{kk'} = O(\Delta T)$ which is sufficient to prove (3–10).

Finally we note that a direct consequence of the Lipschitz conditions and the expression of V^l and $\tilde V$ in Lagrange form is that:

$$\left| \frac{d^r V^l}{dt^r} - \frac{d^r \tilde V}{dt^r} \right| \le C \Delta T^{1-r} \max_k |F(V^l(t_{jk}), t_{jk}) - F(\tilde V(t_{jk}), t_{jk})|. \tag{3–24}$$

Using (3–13) and (3–10) we find:

$$\left| \frac{d^r V^l}{dt^r} - \frac{d^r \tilde V}{dt^r} \right| \le C \Delta T^{p+l+2-r}. \tag{3–25}$$

By (3–12), (3–11) holds so long as $p + l + 2 \ge m$. This completes the proof of Lemma 3.2. □

Remarks. The proof of Lemma 3.2 makes no use of the assumption that $F_E + F_I = F$ and thus holds for inconsistent methods (2–5). We also see that the matrix on the righthand side of (3–22) cannot in general be $o(\Delta T)$ since it is the difference between nonzero full and block lower triangular matrices. Thus 3.2 is not directly related to the accuracy of (2–5). However, the choice of (2–5) does effect the stability of the overall method, though it may still be more efficient to use an inconsistent formula in some cases. We note that these results differ from those presented in [15; 16], where gains in accuracy commensurate with the order of the correction method are proved. A difference is that we are proving higher order convergence - in particular higher order than is attained at the interior quadrature nodes. In [17] it is shown, for linear problems, that by using GMRES to accelerate the correction process only half as many corrections are needed to attain the full accuracy. In addition, they show that the use of GMRES improves the accuracy for stiff problems.

We note that if (2–16)-(2–18) are used, then the analogue of δ^l is given by $\bar{\delta}^l = v^0 - \bar{v}^l$. This correction satisfies:

$$\begin{aligned}\bar{\delta}^l_{jk} = \bar{\delta}^l_{j,k-1} + r^l_{jk} - r^l_{j,k-1} + v^0_{jk} - v^0_{j,k-1} \\ -h_k\Big(F_I(v^0_{jk} - \bar{\delta}^l_{jk}, t_{jk}) + F_E(v^0_{j,k-1} - \bar{\delta}^l_{j,k-1}, t_{j,k-1})\Big). \quad (3\text{–}26)\end{aligned}$$

We see that unless v^0 satisfies (2–5), that is unless the correction method and the starting method coincide, the correction does not approach zero with the residual. Hence the method cannot be interpreted as an approximation to an implicit Runge–Kutta method. However, if they do coincide Lemma 3.2 also holds. The proof follows essentially line for line, so we omit it.

Lastly we remark that our proof, relying as it does on the Lipschitz continuity of F, fails in the stiff limit, though we will show that the order of accuracy is attained for some stiff problems. In [2] experimental studies are presented of the convergence of deferred correction of the backward Euler method to the underlying implicit Runge–Kutta method.

Combining Lemma 3.1 and Lemma 3.2 we have proven our main theorem.

Theorem 3.3. *For a sufficiently smooth solution*, u, *Lipschitz continuous functions*, F, F_I, F_E, $\tilde{F}_I$, $\tilde{F}_E$, *a starting method of order* $p \le m$, *and* $l \ge m - p - 2$, *there exists a constant* C *independent of* ΔT *sufficiently small such that*:

$$|u(T_{j+1}) - V^l(T_{j+1})| \le C \Delta T^{\min(p+l+1, q+1)}.$$

Note that we have assumed that after the desired corrections are made the solution is updated by (2–15). Of course, if T_{j+1} is a node, then nothing needs to be done; we simply take v^L_{jm} as the value at T_{j+1}.

If, on the other hand, T_{j+1} is not a node, we may typically replace (2–15) by simple polynomial extrapolation using $(T_j, v(T_j))$ and all data on the quadrature nodes, thus saving m evaluations of F. In particular, even if $c_1 \neq 0$, that is if T_j is not a node, Theorem 3.3 is still valid. To see this, define the polynomial

$$\phi(t) = v(T_j) + \int_{T_j}^{t} \psi_j^L(s)ds - r^L(t), \tag{3–27}$$

where $r^L(t)$ is the polynomial that interpolates $(T_j, 0)$ and (t_{jk}, r_{jk}^L), for $k = 1, 2, \ldots, m$. Obviously, the degree of $\phi(t)$ is m. Suppose T_j is not a node. Then since $\phi(t_{jk}) = v_{jk}^L$, $\phi(t)$ is exactly the polynomial that interpolates $(T_j, v(T_j))$ and $(t_{jk}, v_{jk}^L), k = 1, 2, \ldots, m$. The update given by (2–15) is

$$v(T_{j+1}) = \phi(T_{j+1}) + r(T_{j+1}), \tag{3–28}$$

while the update given by extrapolation is $\phi(T_{j+1})$. So the difference between these two updates is controlled by $r(T_{j+1})$. Therefore, they have the same order, though the stability characteristics may be altered.

4. Efficiency and linear stability of sample methods

We now consider, experimentally, the accuracy, efficiency and linear stability of some simple examples of the methods discussed above. For simplicity, following [1], we consider a Dahlquist-type problem modeling operator splitting applied to a spatially discretized advection-diffusion equation. Precisely we consider:

$$u' = (\alpha + i\beta)u, \quad \alpha, \beta \in R, \quad \alpha \leq 0, \tag{4–1}$$

and take

$$F_E(u) = \tilde{F}_E(u) = i\beta u \quad \text{or} \quad F_E = 0, \tag{4–2}$$

$$F_I(u) = \tilde{F}_I(u) = \alpha u. \tag{4–3}$$

The range of methods tested includes:

(1) Gauss–Lobatto, Gauss–Legendre and right-handed Gauss–Radau quadrature with $m = 3, \ldots, 10$ nodes, encompassing method orders from 5 through 20;

(2) Multistep IMEX methods [1] of orders 1 through $m - 1$ as starting methods (2–4);

(3) Consistent correction methods with $F_E = i\beta u$ and inconsistent correction methods with $F_E = 0$;

We note that the split multistep methods we use are based on backward differentiation with F_E extrapolated to the new time level. Thus they are the natural generalizations of the SBDF methods of [1] to nonequispaced grids. Their order is k and we take $k = 1, \dots, m-1$. When $k > 1$ we are using values, $v^{L}_{j-1,r}$, which are only accurate to order m. Also in that case we need to consider the eigenvalues of the amplification matrix, A, mapping between values used in subsequent starting formulas. The stability properties of the SBDF methods themselves are not directly at issue and have not been studied, though we expect they are unstable at high order.

As the number of methods considered is in the hundreds we will limit our discussion to a few representative cases. Mainly we will display results obtained using the Gauss–Legendre and Gauss–Radau nodes. In most instances the behavior of the Gauss–Lobatto methods was essentially the same. An exception is the stability regions, where all three will be compared.

4.1. ***Accuracy and efficiency.*** We first verify that the methods attain the design order even if inconsistent corrections are used. Precisely we consider (4–1) with $\alpha = -1/20$, $\beta = -2\pi$, and solve up to $T = 20$. See Figure 1 for experiments with Gauss–Legendre nodes and Figure 2 for experiments with Gauss–Radau nodes. In each case we observe convergence at the correct rate, though the use of inconsistent corrections clearly leads to less accurate results for a fixed time step. The accuracy is insensitive to the order of the starting method, indicating higher efficiency with the use of higher order starting methods.

From the point of view of computational effort, the efficiency of an ode solver is typically measured by the number of function evaluations required to attain a given accuracy. An emphasis of the current work is the possibility to achieve higher order for fixed m than in earlier implementations of SDC, presumably with significant savings in memory. However, these savings could conceivably be lost if the methods proposed here turn out to be less efficient. Thus we wish to compare the efficiency of different variations of SDC, including methods where m is chosen larger than necessary to achieve the design order.

To facilitate comparisons with previously published results we consider here, following [21], a nonstiff van der Pol equation for $0 \le t \le 4$:

$$u_1' = u_2, \quad u_2' = -u_1 + (1 - u_1^2)u_2, \tag{4–4}$$

$$u_1(0) = 2, \quad u_2(0) = \frac{2}{3}. \tag{4–5}$$

As the equations are not stiff, we do not split them and simply use explicit starting and correction methods.

We consider three comparisons. First we fix the starting method (first order) and the number of corrections while varying m. Thus we are comparing methods of the same order. The results, shown in Figure 3, show that efficiency measured in this way is essentially independent of m; it is apparently determined by the average step

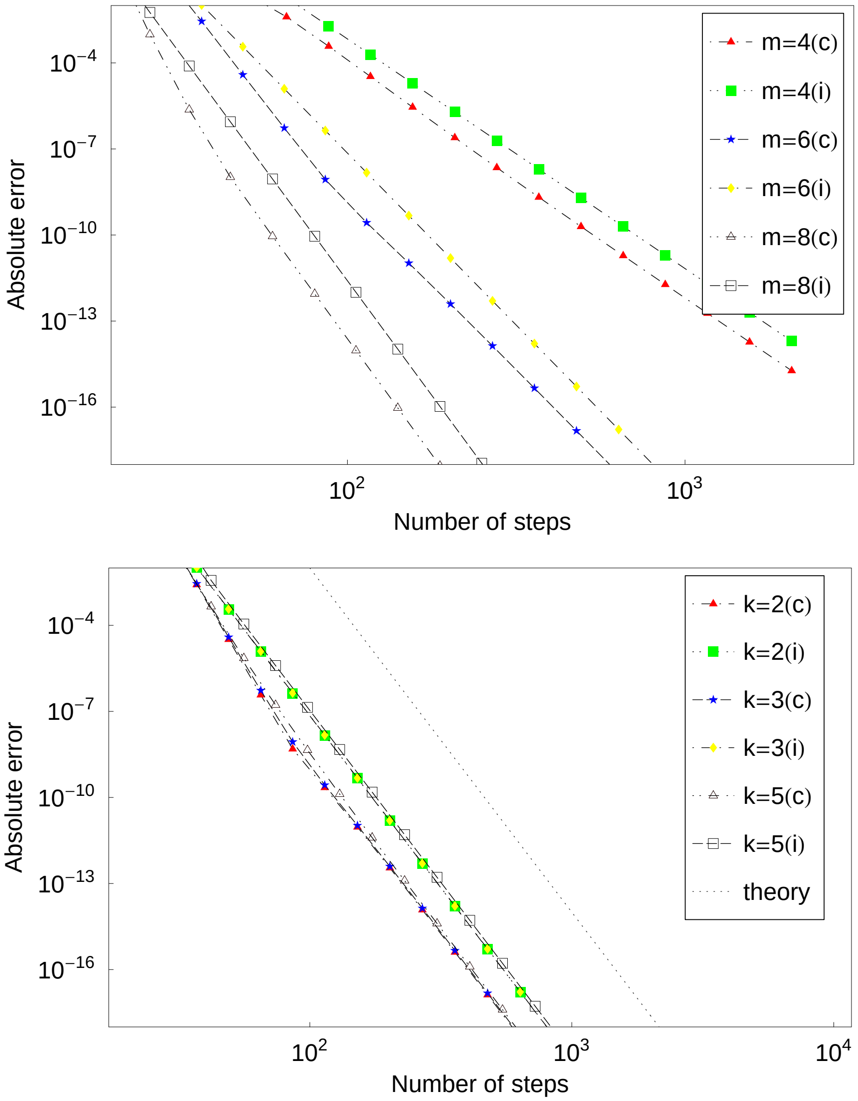


Figure 1. The top graph shows the accuracy of the split SDC methods using the same starting method (3rd order SBDF) and different numbers of Gauss–Legendre nodes m. The bottom graph shows the accuracy of SDC methods with 6 Gauss–Legendre nodes, but different starting methods (k indicates the kth-order SBDF starting method, c,i indicates consistent and inconsistent corrections, respectively). The dotted line shows the theoretical convergence order 12.

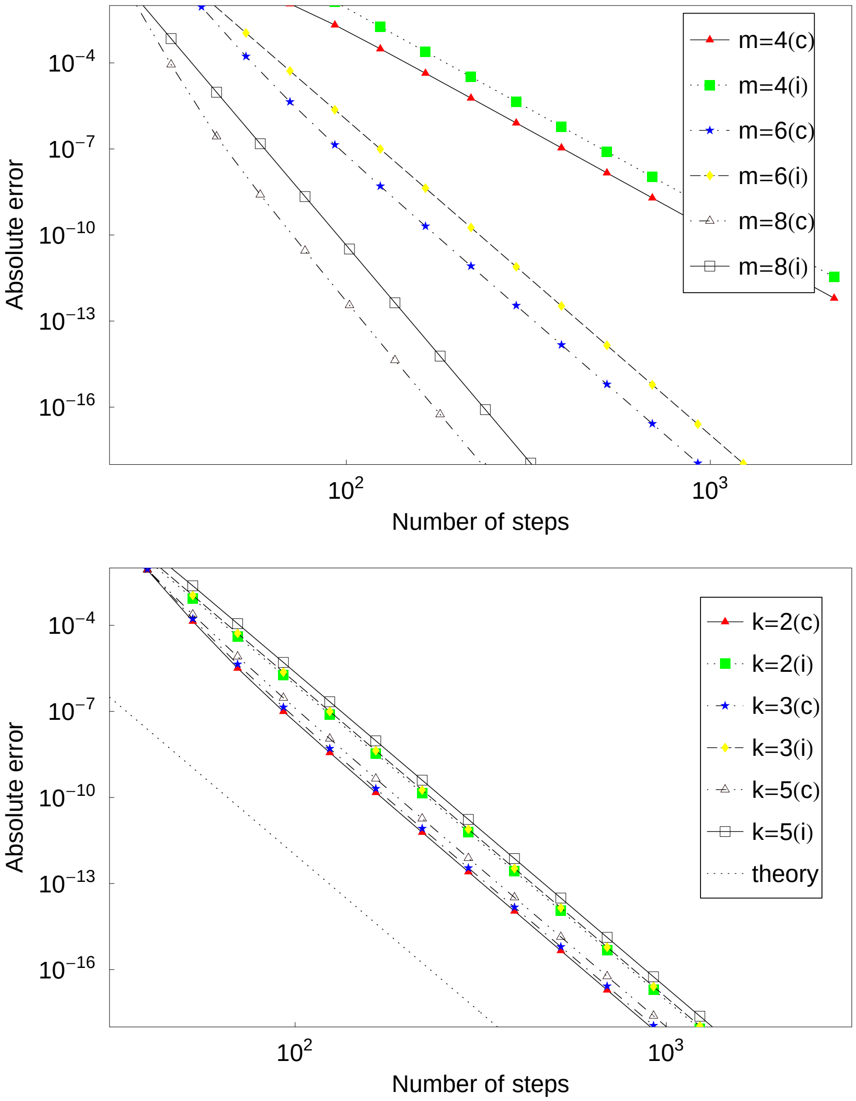


Figure 2. The top graph shows the accuracy of the split SDC methods using the same starting method (3rd order SBDF) and different numbers of Gauss–Radau nodes m. The bottom graph shows the accuracy of split SDC methods with 6 Gauss–Radau nodes, but different starting methods (k indicates the kth-order SBDF starting method, c,i indicates consistent and inconsistent corrections, respectively). The dotted line shows the theoretical convergence order 11.

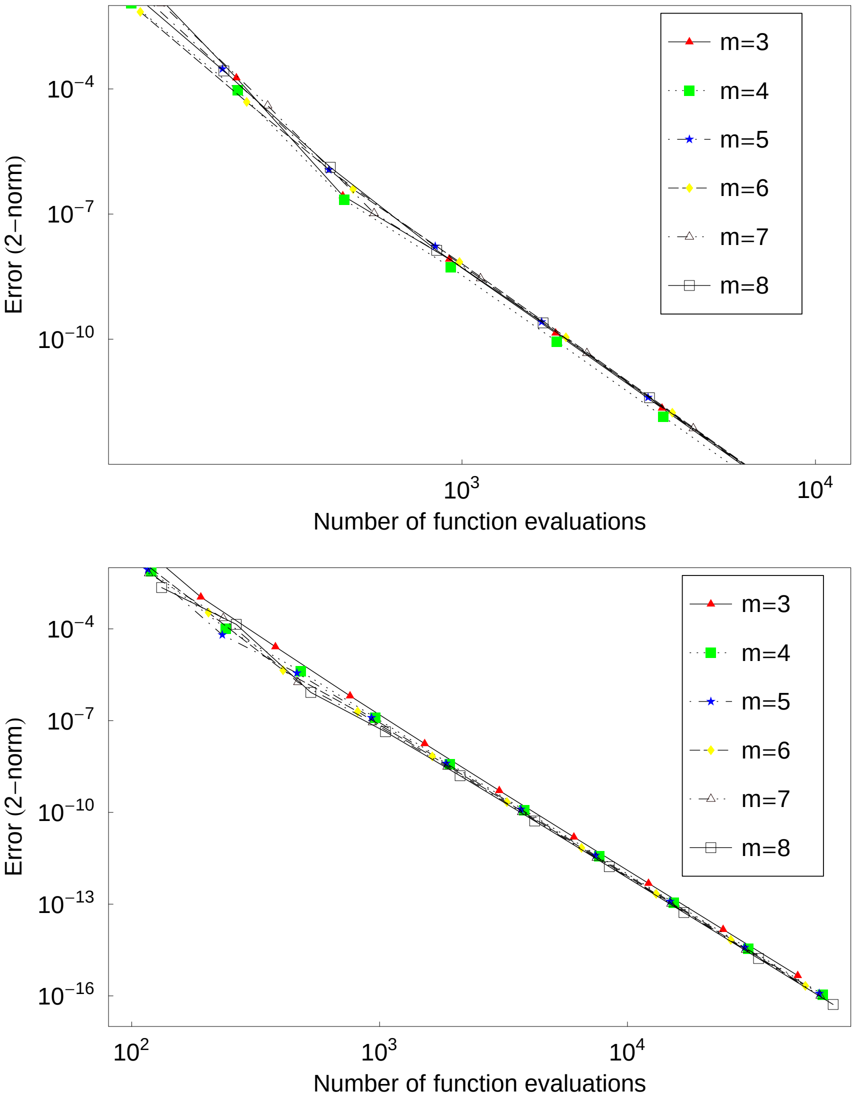


Figure 3. Efficiency in terms of function evaluations for 6th order Gauss–Legendre and 5th order Gauss–Radau methods with first order starting methods and varying numbers of quadrature nodes, m.

size and the number of corrections. Thus the gains in memory utilization resulting from the exploitation of the full order of the quadrature rules are fully realized, at least for this example.

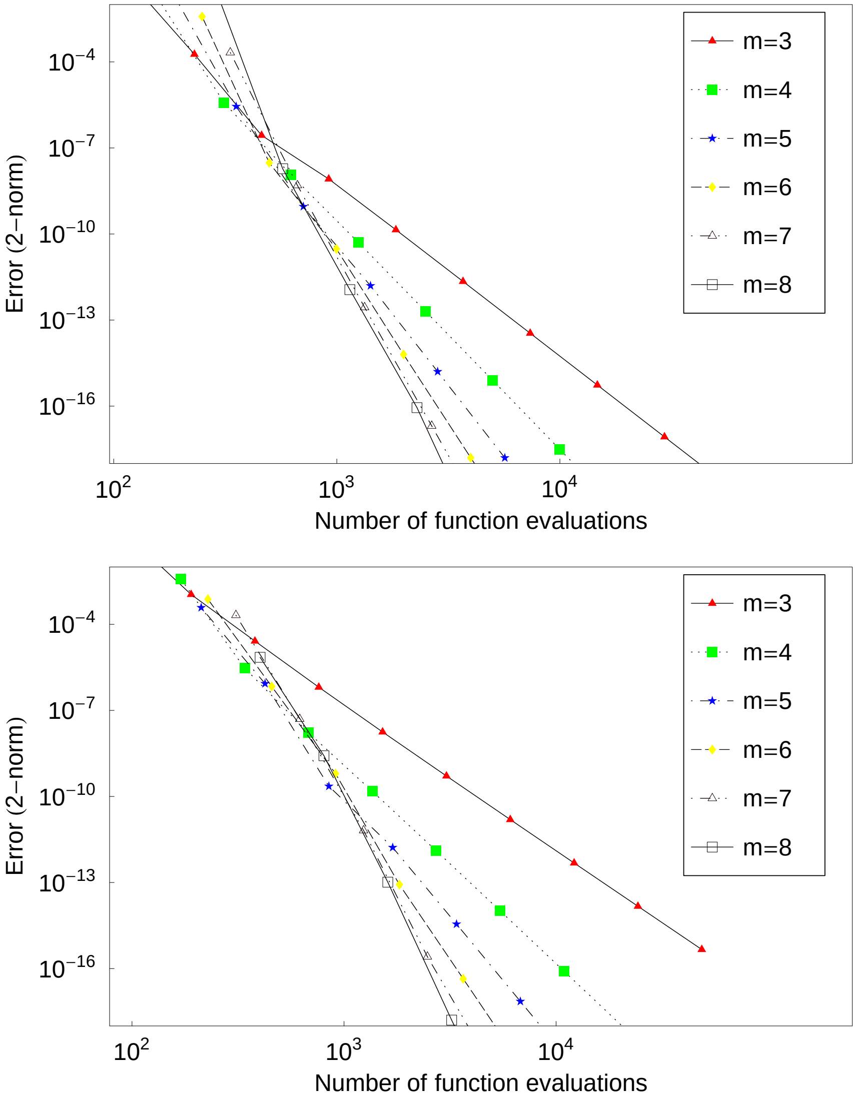


Figure 4. The efficiency of Legendre and Radau methods for various orders. In all cases we use a first order starting method.

Second, in Figure 4 we compare efficiency for differing orders and quadratures, in each case using the full order of the quadrature rule. As expected, the "optimal" method order depends on the desired tolerance, with the higher order methods favored as the tolerances decrease.

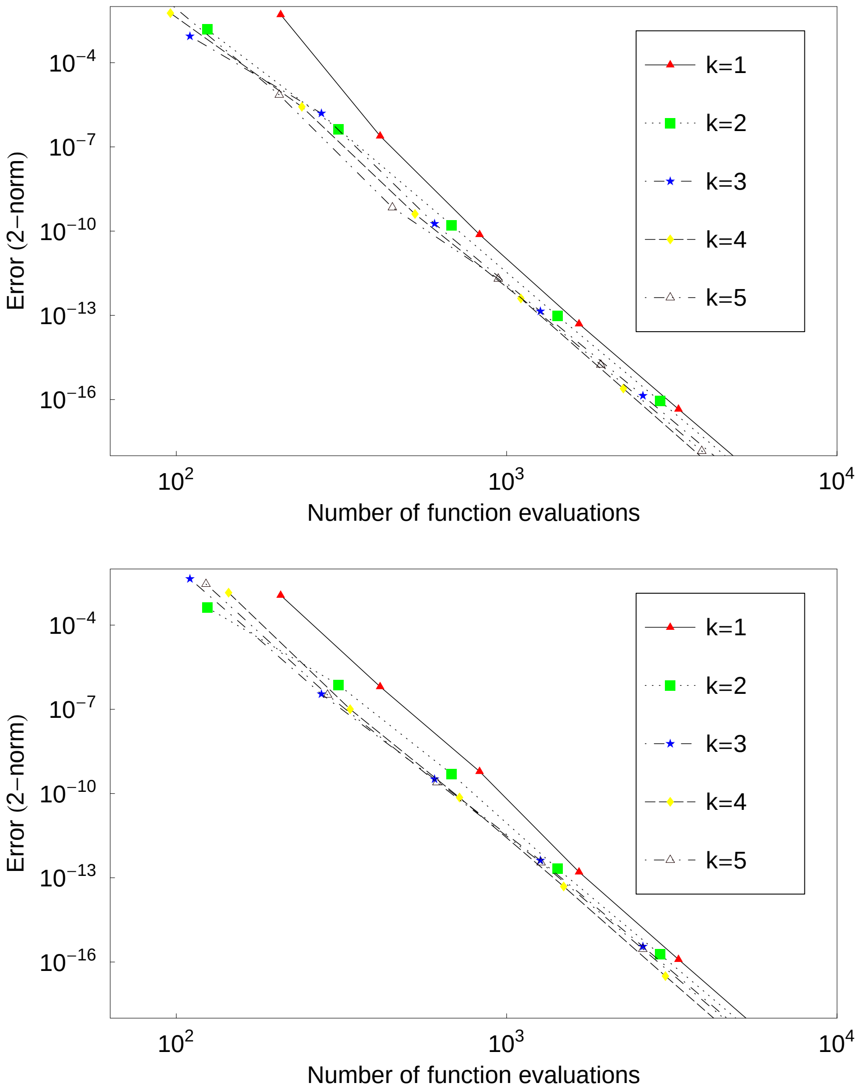


Figure 5. The efficiency of 12th order Legendre and 11th order Radau methods varying the order of the starting method.

Lastly we consider the effect of varying the order of the starting method, fixing $m = 6$. The results, shown in Figure 5, demonstrate a substantial gain in efficiency as the starting method order is increased from 1 to 2 with modest, but measurable, gains when it is increased from 2 to 3. Beyond third order there seems to be no advantage in further increases.

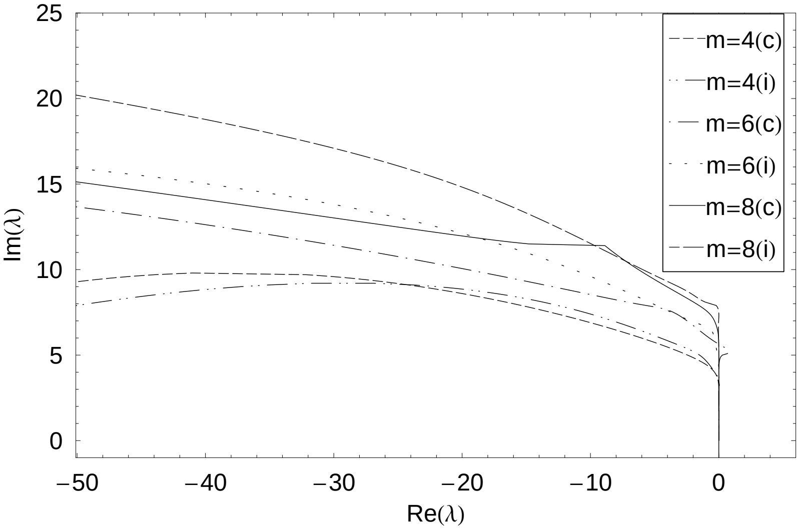
4.2. ***Stability for*** **(4–1).** The analysis of the stability of splitting methods in general is difficult. Even for linear, constant coefficient systems, the fact that the matrices defining the split operators cannot be expected to commute limits the predictive value of analyzing a split version of Dahlquist's model problem. Nonetheless, at least to establish some basis for the comparison of stability for our different splitting procedures, we will follow [1] and plot experimentally determined stability domains associated with (4–1). In the following sections we will consider a more general stability problem motivated by what we call preconditioned splitting methods.

We note that a more interesting definition of stability domains for splitting methods has been proposed by Frank et al. [8]. Their idea is to consider stability for a scalar, split system under the assumption that the time step is chosen so that the explicit method is stable. This allows a clean definition of a stability domain for split methods and a generalization of many of the standard notions of $A(\alpha)$ and $L(\alpha)$ stability. Layton and Minion [20] have applied this definition to the SDC of splitting methods and shown that quadrature rules excluding the left endpoint such as the Gauss–Legendre and righthand Gauss–Radau rules lead to $L(\alpha)$-stable methods with $\alpha \approx \pi/2$. However, this analysis is not general as one might often want to use methods in regimes where the explcit scheme by itself is unstable. The simple case of (4–1) illustrates this; the stability domain of explicit Euler contains no points on the imaginary axis except the origin. Thus with the splitting considered here the consistent explicit correction method is always unstable, so the results of [8; 20] do not apply. Nonetheless, as in [20] we find that the stability properties of the Gauss–Lobatto methods are clearly inferior to those based on Gauss–Legendre or Gauss–Radau quadrature.

Figures 6 and 7 show stability domains for various Legendre and Radau-based methods. The overall results are quite similar. The stability domains are somewhat larger if inconsistent rather than consistent corrections are used. The domains increase in size with increasing m but decrease with increasing k. Obviously, except for k large, they contain a very large region near the negative real axis.

Lastly in Figure 8 we compare the stability of 8th order methods with $m = 5$. We clearly see that the stability domain obtained using Gauss–Radau nodes is slightly larger than that obtained with Gauss–Legendre nodes, but both are much larger than the stability domain obtained using Gauss–Lobatto nodes.

4.3. ***Relative accuracy.*** Lastly we make some relative accuracy comparisons fixing m and the order of the methods in Figure 9. Note that we are thus not carrying out the full number of corrections when Legendre or Radau nodes are used. We find that under these restrictions the Legendre nodes yield the most accurate results, followed by the Radau nodes.



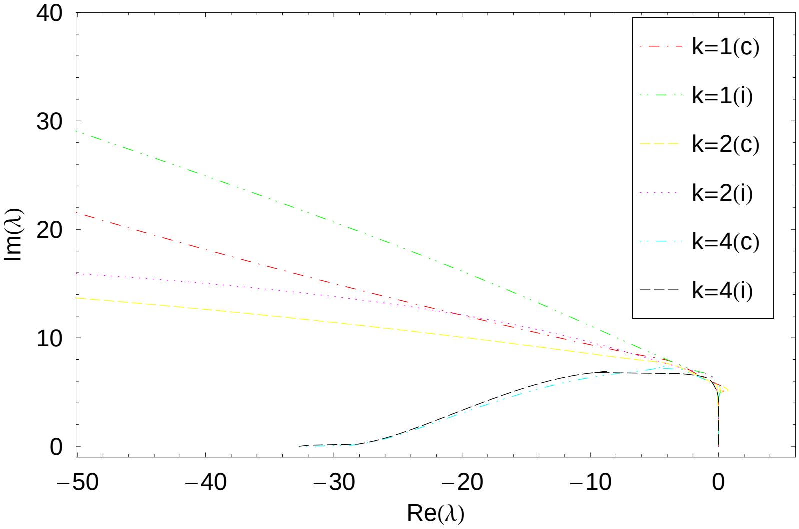


Figure 6. The top graph shows the stability of the SDC methods using same starting method (2nd order SBDF) and different numbers of Gauss–Legendre nodes m. The bottom graph shows the stability of SDC methods with 6 Gauss–Legendre nodes, but different starting methods (k indicates the kth-order SBDF starting method).

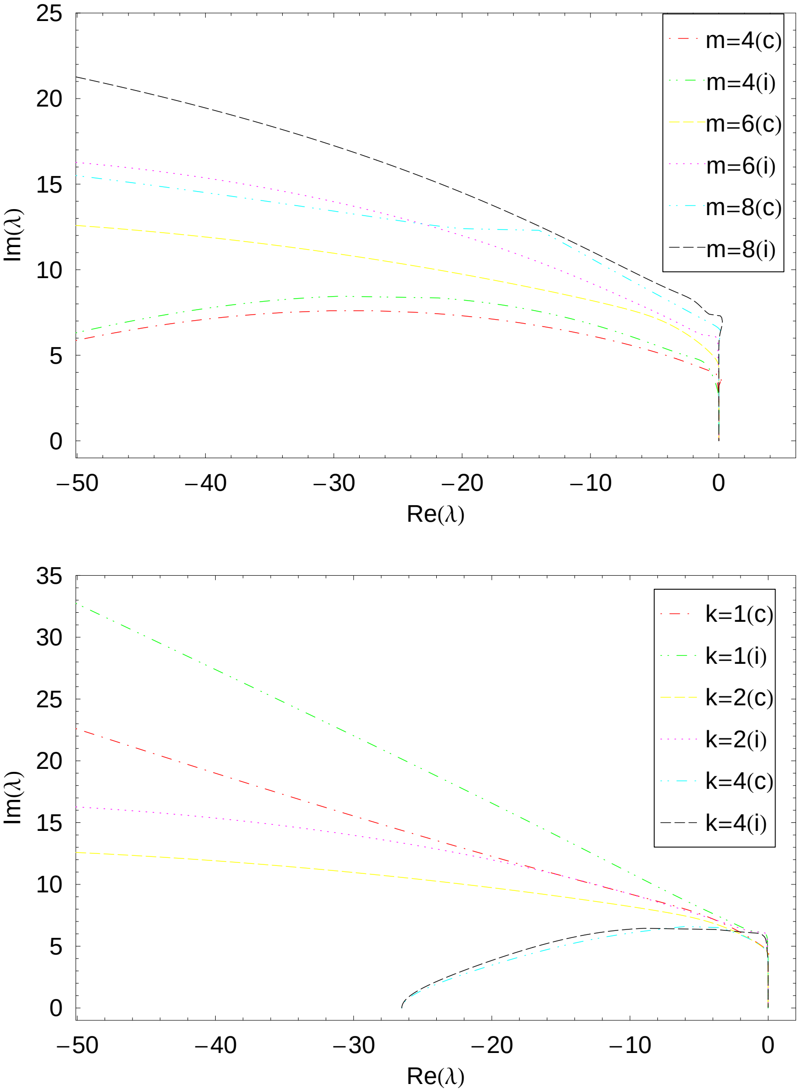


Figure 7. The top graph shows the stability of the SDC methods using same starting method (2nd order SBDF) and different numbers of Gauss–Radau nodes m. The bottom graph shows the stability of SDC methods with 6 Gauss–Radau nodes, but different starting methods (k indicates the kth-order SBDF starting method).

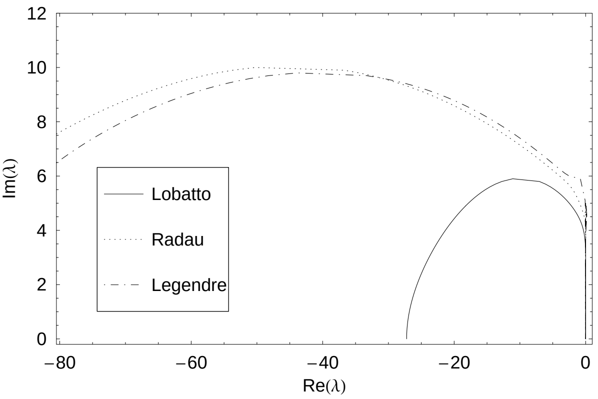


Figure 8. Stability domains for 8th order methods (2nd order SBDF starting method and 6 consistent corrections) with $m = 5$ and various quadrature nodes.

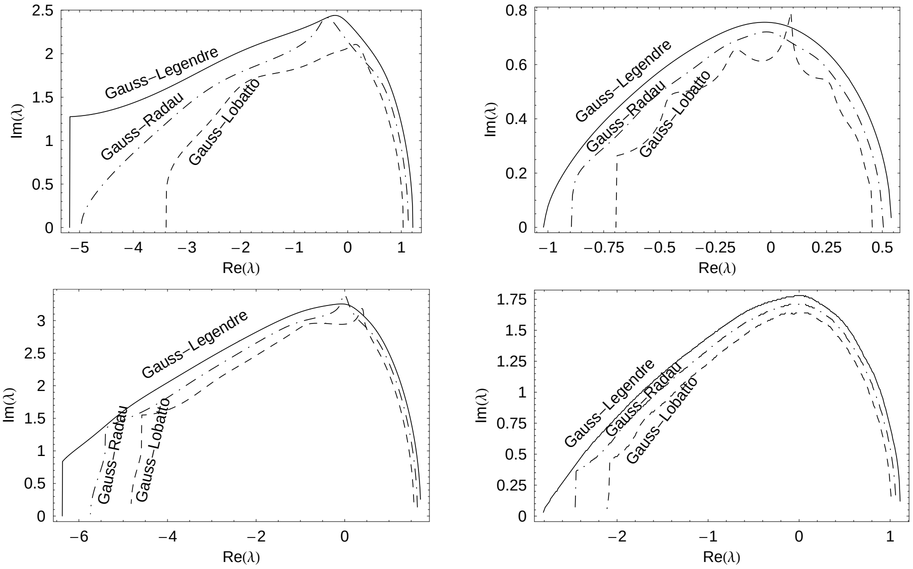


Figure 9. In all figures, we use the single step starting method with consistent corrections. For the top two figures, $m = 5$, the order is 8 for all methods. The top-left figure shows the accuracy region for $\epsilon = 10^{-4}$; while for the top-right figure, $\epsilon = 10^{-9}$. For the bottom two figures, $m = 9$, the order is 16 for all methods. For the bottom-left figure, $\epsilon = 10^{-9}$, while for the bottom-right figure, $\epsilon = 10^{-14}$.

5. Preconditioned splitting methods

Although the stability and accuracy of the SDC methods used in conjunction with advection-diffusion splittings is reasonable, we believe it is worthwhile to pursue a more general and flexible approach. Introduce a preconditioning matrix P and the splitting:

$$F_I = -Pv, \quad F_E = F + Pv. \tag{5–1}$$

Here, P will generally be dependent on t and local values of v. It should satisfy the requirements:

i: $P + P^* \geq 0$;

ii: $I + h_k P$ inexpensively invertible;

iii: the split methods have good stability and accuracy properties.

Of course the difficult property to satisfy is the third. For a simple model problem, we will see that stability of the base method is ensured by choosing P sufficiently large compared with the Jacobian of F, and then good accuracy follows from not making it too large. (We suspect that generalizations of the stability analysis to nonlinear problems satisfying appropriate one-sided Lipschitz conditions would be straightforward.)

5.1. *Linear stability for the scalar problem.*

We repeat the stability analysis for Dahlquist's equation (4–1) but now with the general preconditioner:

$$P = \mu + i\eta, \quad \mu, \eta \in R, \quad \mu \geq 0. \tag{5–2}$$

Note that we are allowing an imaginary part in P, corresponding to the inclusion of a linear advection term in the preconditioner. The amplification factor of the first order splitting (2–5) is then given by:

$$r^2 = \left| \frac{1 + h(\alpha + \mu + i(\beta + \eta))}{1 + h(\mu + i\eta)} \right|^2 = \frac{(1 + h(\alpha + \mu))^2 + h^2(\beta + \eta)^2}{(1 + h\mu)^2 + h^2\eta^2}. \tag{5–3}$$

For $\eta = 0$ we have A-stability if:

$$\mu \geq \frac{\alpha^2 + \beta^2}{2|\alpha|}. \tag{5–4}$$

For a discretized advection-diffusion equation with Peclet number Pe we obtain:

$$P \geq C\left(-\frac{1}{Pe} D_x^2 + Pe\right), \tag{5–5}$$

where D_x^2 is an approximation to the Laplacian and the inequality is in the usual sense of matrices. For large Peclet number such a choice is likely to have a negative impact on accuracy as the preconditioner is large. Of course one can give up on A-stability. For example if $\mu \geq |\alpha|/2$ we have:

$$h \leq \frac{2|\alpha|}{\beta^2} = O(Pe^{-1}), \tag{5–6}$$

independent of the spatial mesh width Δx, which is acceptable if Pe is not too large.

Much better results can be obtained if we choose η to be nonzero and of the opposite sign of β. Then we have A-stability if:

$$\mu \geq \frac{|\alpha|}{2}, \quad |\eta| \geq \frac{|\beta|}{2}, \quad \alpha\beta \leq 0. \tag{5–7}$$

We will show nonlinear examples where a linear advection term is included in P. Of course the sign condition can be difficult to satisfy where the advection term nearly vanishes, but then the local Peclet number is not large. A-stability is not generally preserved when the correction process is included, but we will see below that the stability domains can be quite large.

We again note that we do not in general expect that our time step is chosen so that the explicit method is stable. Thus the stability analyses of [8; 20] are not directly applicable.

5.2. ***Linear stability domains of the preconditioned methods.*** In Figures 10, 11, and 12 we plot linear stability domains of the consistently corrected methods assuming:

$$\mu = d_r\alpha, \quad \eta = d_i\beta, \tag{5–8}$$

with d_r and d_i chosen from $\{\frac{3}{4}, \frac{5}{4}\}$. Clearly, d_i determines stability along the imaginary axis and d_r along the real axis.

As before, the stability characteristics of the Gauss–Legendre and Gauss–Radau methods are quite similar, with the stability domains of the Gauss–Legendre methods being generally a little larger. The Gauss–Lobatto methods, on the other hand, show superior stability along the real axis for $d_r = \frac{3}{4}$.

We also tested multistep starting methods and inconsistent corrections. Except in the case of second order starting methods, the stability domains are significantly reduced when consistent corrections are used. However, with inconsistent corrections and $d_r = d_i = \frac{5}{4}$ they are enlarged. We plot below (Figure 13) results using an 11th order Radau scheme with the 3rd order starting method.

6. Nonlinear numerical experiments

Finally we consider the actual accuracy and stability of one of the methods discussed above for a collection of nonlinear parabolic initial-boundary value problems in

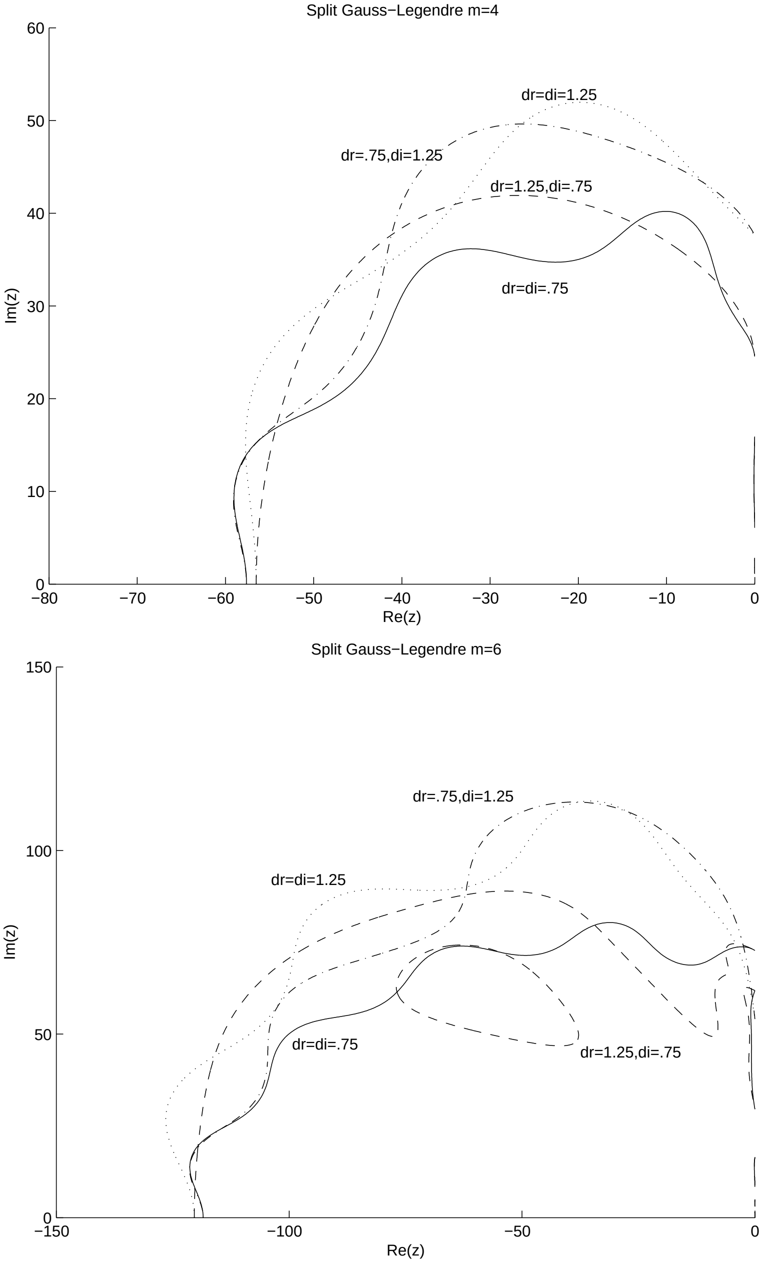
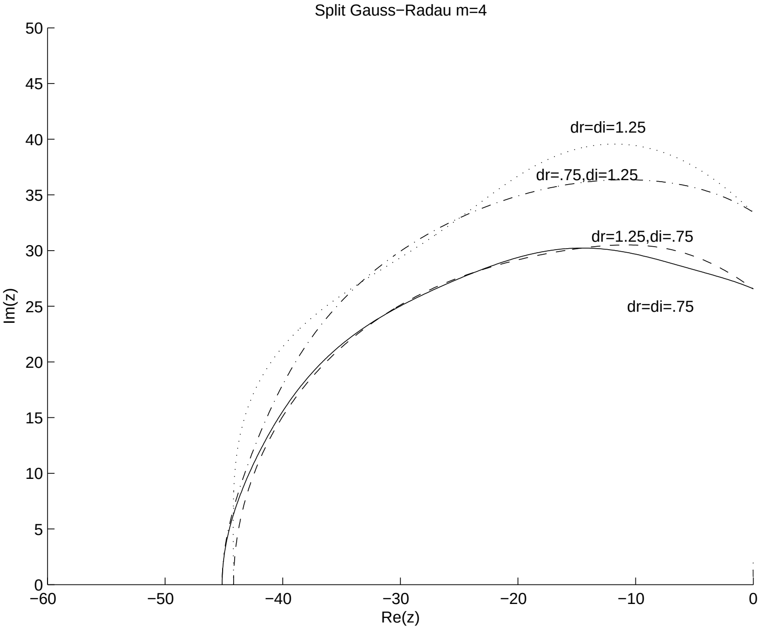


Figure 10. Stability domains for Gauss–Legendre methods with preconditioned splittings.



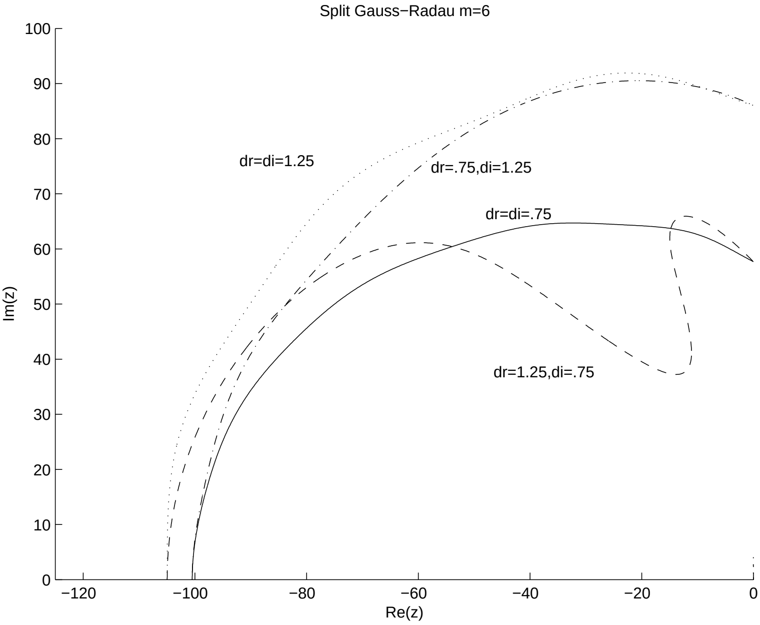
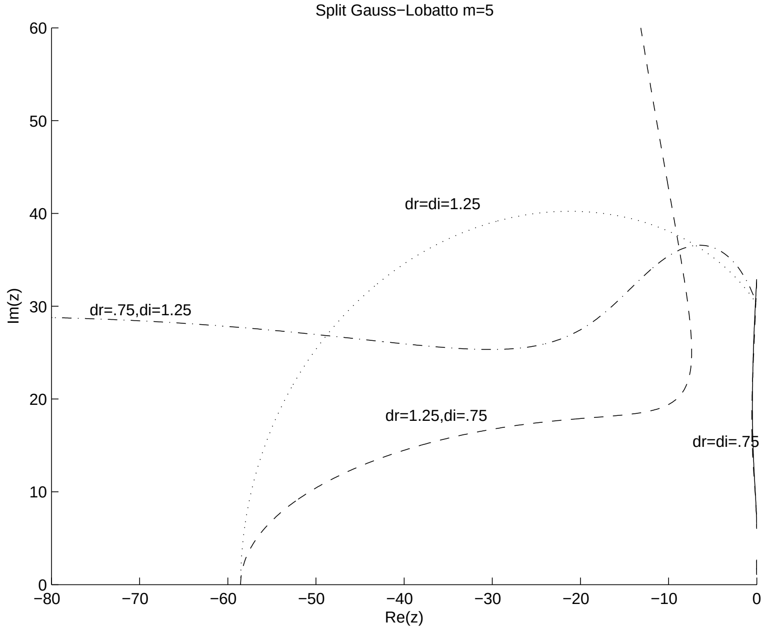


Figure 11. Stability domains for Gauss–Radau methods with preconditioned splittings.



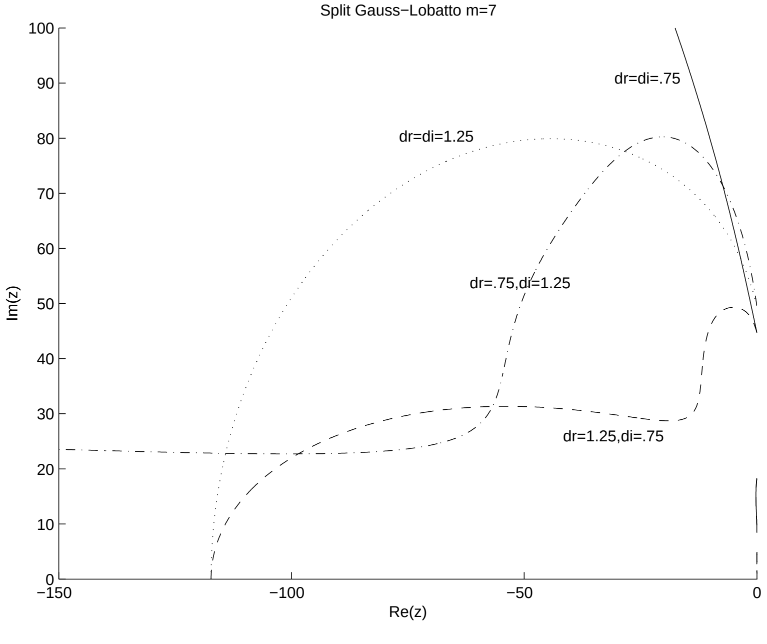


Figure 12. Stability domains for Gauss–Lobatto methods with preconditioned splittings.

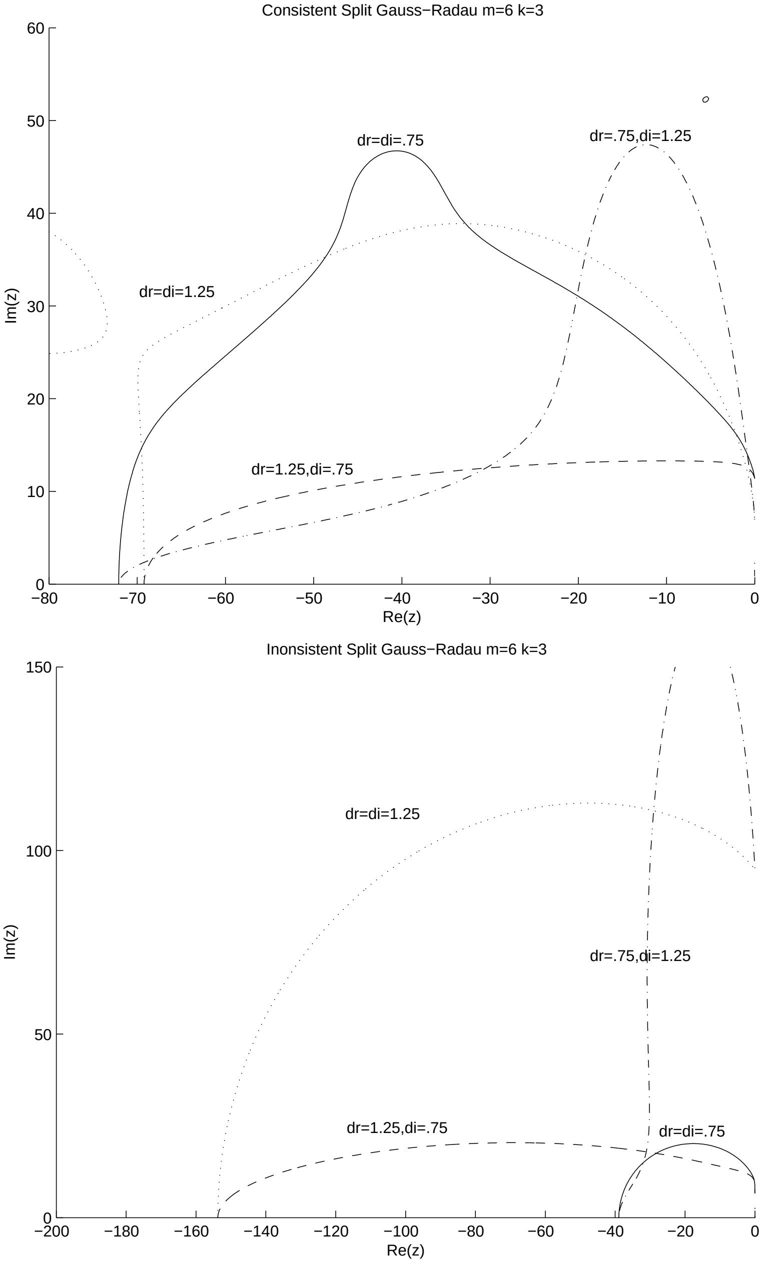


Figure 13. Stability domains for 7th order Gauss–Radau methods with preconditioned splittings and a 3rd order multistep starting method.

$1+1$ dimensions:

$$u_t = F(u, u_x, u_{xx}), \quad x \in (x_L, x_R), \tag{6–1}$$

supplemented by boundary conditions. In all cases we use the 7th order Radau method with a multistep SBDF preconditioned starting method and consistent corrections. Our spatial discretizations are 8th order; one-sided differencing at the boundaries is stabilized by the addition of a single sub-cell point at $x_L + 0.2\Delta x$ and $x_R - 0.2\Delta x$ with central differences used in the interior. See [10] for details.

Various preconditioners are considered, but in all cases the spatial differencing used in the preconditioning is limited to a 3-point stencil to minimize bandwidth. In the interior, then, we are preconditioning 8th order differences by multiples, γ_d, γ_c, of 2nd order differencing. To better correlate the results with the simple linear stability domains shown above we note that the symbols of the qth order central difference approximations to $\frac{d^j}{dx^j}$, $\hat{d}_{j,q}$, satisfy:

$$\max_{|\omega h| \leq \pi} \frac{\hat{d}_{1,8}(\omega)}{\hat{d}_{1,2}(\omega)} \approx 2.66, \tag{6–2}$$

$$\max_{|\omega h| \leq \pi} \frac{\hat{d}_{2,8}(\omega)}{\hat{d}_{2,2}(\omega)} \approx 1.68. \tag{6–3}$$

Thus, for example, a preconditioned approximation to the heat equation is positive independent of time step only if the damping factor, γ, is chosen to be larger than 1.68.

Of course the examples are primarily meant to illustrate a viable preconditioning strategy and to provide some experience in the method's performance under a variety of conditions. With experience for a given system we would expect that better preconditioners could be found leading to further improvements in efficiency. Most of our examples would benefit from the use of an adaptive spatial mesh, but here we simply employ sufficiently fine uniform discretizations. We also compare our results with those obtained using a standard second order Strang splitting in time (e.g [22]) and a time step chosen so that the number of evaluations of the nonlinearities are comparable. For example, if we use a fourth order starting method an entire SDC step entails sixteen substeps, so we choose the time step for the Strang method to be $1/16$ times that of the SDC method. However, recall that our method is linearly implicit while the Strang splitting employs Newton iterations; thus the SDC method is noticeably faster for the time steps compared. We note, of course, that we could have used the Strang splitting as our starting method or even as our correction method, but we have not yet implemented this.

Table 1. Error data for the Brusselator problem.

ΔT	Δx	γ_d	$e_{\max}(u)$	$q(u)$	$e_{\max}(v)$	$q(v)$
$2E(-1)$	$2E(-2)$	2	3.13(−4)		1.51(−4)	
$1E(-1)$	$1E(-2)$	2	2.10(−6)	7.2	1.73(−6)	6.4

6.1. ***Brusselator.*** We consider for $(x, t) \in (0, 1) \times (0, 10)$:

$$u_t = 1 + u^2 v - 4u + 2 \cdot 10^{-3} u_{xx}, \tag{6–4}$$

$$v_t = 3u - u^2 v + 2 \cdot 10^{-3} v_{xx}, \tag{6–5}$$

$$u(x, 0) = 1 + \sin 20\pi x, \quad v(x, 0) = 3, \tag{6–6}$$

with Dirichlet boundary conditions. With this data the solution is known to oscillate; see the graph of the fine grid solution in Figure 14 as well as [14; 22].

Here there is no convective term to be included in the preconditioner, but the Jacobian of the reaction terms is included along with the scaled three point diffusion approximation. That is with $\nu = 2 \cdot 10^{-3}$:

$$P_i = -\begin{pmatrix} 2u_i v_i - 4 + \nu\gamma_d d_{2,2} & u_i^2 \\ 3 - 2u_i v_i & -u_i^2 + \nu\gamma_d d_{2,2} \end{pmatrix}. \tag{6–7}$$

We employ a fourth order starting method with three correction steps.

The results, displayed in Table 1, are consistent with the design accuracy. Error data is obtained by comparison with a solution computed using $\Delta T = 2.5E(-4)$ and $\Delta x = 1E(-3)$.

By way of comparison, with $\Delta x = 2E(-2)$ and $\Delta T = 1.25(-2)$ the maximum errors with Strang splitting were $(1.70(-3), 9.60(-4))$, about six times larger than those reported above. Halving the grid and step sizes the Strang errors are reduced by about a factor of four to $(3.56(-4), 2.19(-4))$, about two orders of magnitude larger than were obtained with the SDC time stepping.

We also determined apparent time step stability limits. For $\gamma_d = 2$ these were weakly dependent on Δx, but we could always take rather large steps; $\Delta T = \frac{1}{2}$ for $\Delta x = \frac{1}{50}$, $\Delta T = \frac{1}{3}$ for $\Delta x = \frac{1}{100}$ and $\Delta T = \frac{1}{8}$ for $\Delta x = \frac{1}{150}$. For $\gamma_d = 1$, on the other hand, they clearly took the form $\Delta T \leq c\Delta x^2$. With $\Delta x = \frac{1}{150}$ it was necessary to take $\Delta T = \frac{1}{101}$.

6.2. ***Smoothed angiogenesis model.*** Here we consider a smoothed verison of a tumor angiogenesis model presented in [18]:

$$\rho_t = 10^{-3}\rho_{xx} - .75(\rho c_x)_x + 10^2 \rho(1 - \rho)K(c) - 4\rho, \tag{6–8}$$

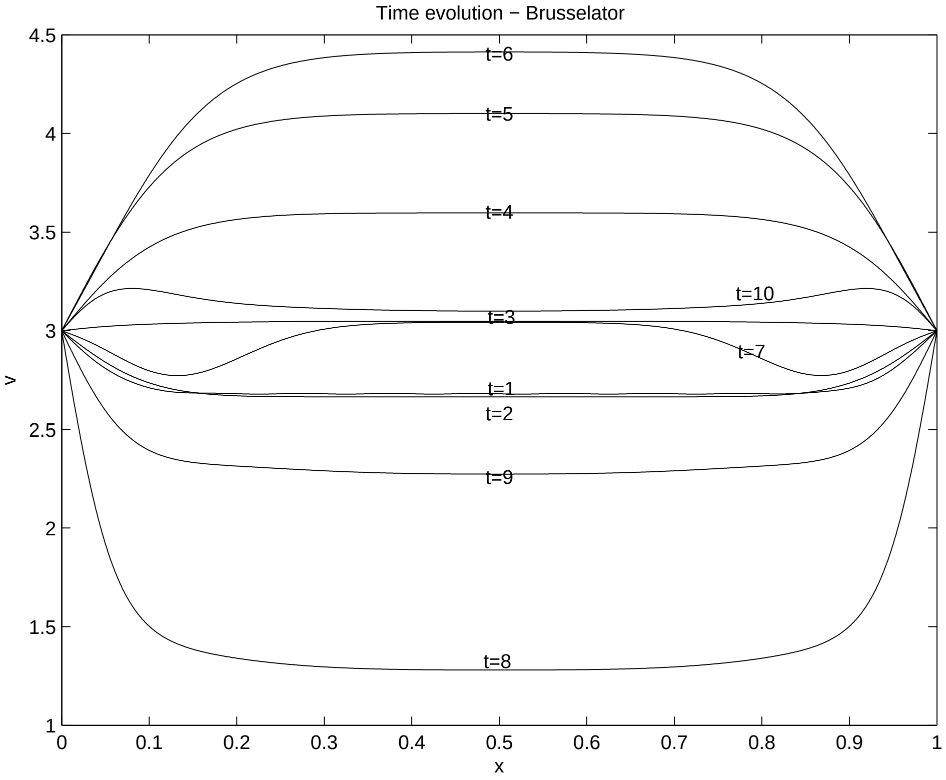


Figure 14. Fine grid solution of the Brusselator equation: v.

$$c_t = c_{xx} - c - \frac{10\rho c}{1+c}, \tag{6–9}$$

where $(x,t) \in (0,1) \times (0,.7)$ and:

$$K(c) = 5 \cdot 10^{-3}(100(c-.2) + \ln(\cosh(100(c-.2)))), \tag{6–10}$$

$$\rho(x,0) = e^{-288(x-1)(x-1.08\bar{3})}, \quad c(x,0) = \cos \pi x/2, \tag{6–11}$$

and $\rho(0,t) = c(1,t) = 0$, $\rho(1,t) = c(0,t) = 1$.

The evolution of ρ for a fine grid solution computed with $\Delta T = 5E(-4)$, $\Delta x = 1E(-3)$ is shown in Figure 15. Comparison with the figures in [18] show that the smoothing has had little effect on the solution.

As these equations involve both first and second order spatial derivatives both γ_d and γ_c must be chosen. Precisely we used a block diagonal preconditioner:

$$P_{i,11} = -\left(10^{-3}\gamma_d d_{2,2} - .75\gamma_c(d_{1,2}c)_i d_{1,2} - 10^2(1-2\rho_i)K(c_i) - 4\right), \tag{6–12}$$

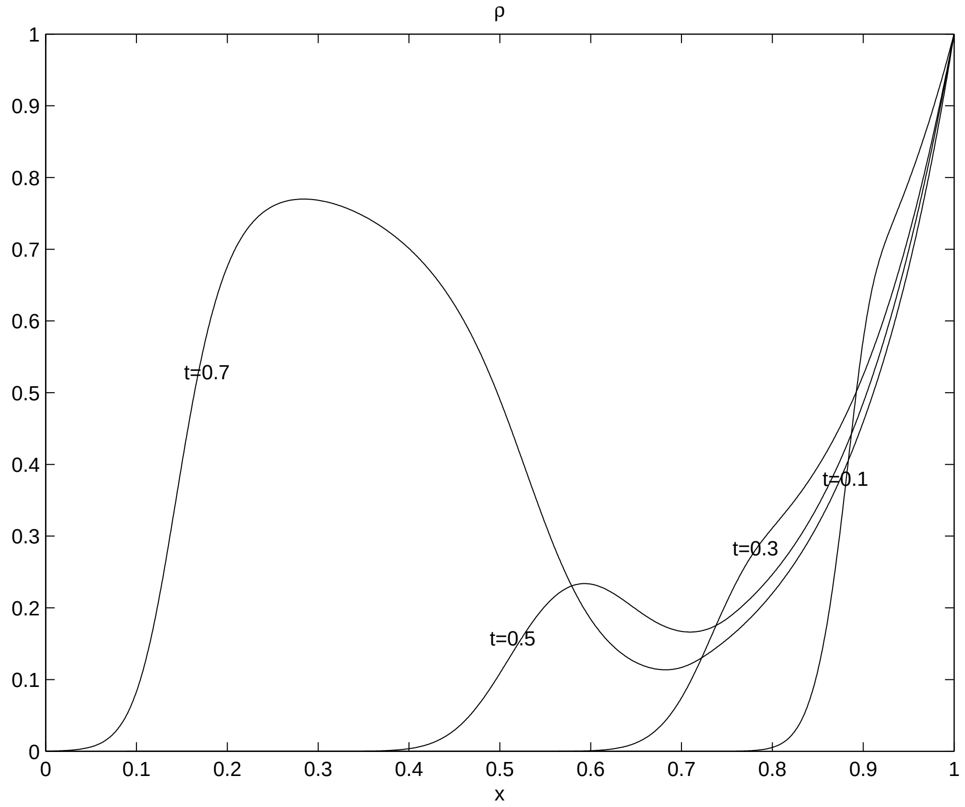


Figure 15. Fine grid solution of the angiogenesis equation: ρ.

$$P_{i,22} = -\left(\gamma_d d_{2,2} - \frac{10\rho_i}{(1+c_i)^2} - 1\right). \tag{6–13}$$

Due, we believe, to the presence of the additional second order term, acceptable stability results led us to use a second order rather than a fourth order starting method. We tested for stability with $\Delta x = 1E(-2)$ and $\Delta x = 1E(-3)$ for $\gamma_d = 1, 2$ and $\gamma_c = 1, 2, 3$. As in the previous example, with $\gamma_d = 1$ it was necessary to take $\Delta T \propto \Delta x^2$. For $\gamma_d = 2$, on the other hand, it was possible to choose ΔT independent of Δx. However, in contrast with the previous case, it was not possible to take ΔT large. With $\gamma_c = 2$ we found $\Delta T \leq 1E(-2)$ while with $\gamma_c = 0, 1$ we could choose $\Delta T \leq 1.7E(-2)$. However, we did observe better accuracy for large steps with $\gamma_c = 0$ than with $\gamma_c = 1$.

The accuracy of the computed solutions with large steps, $\gamma_d = 2$, and $\gamma_c = 0$ is displayed in Table 2. Obviously the results are consistent with the design accuracy.

Table 2. Error data for the angiogenesis problem: ρ.

ΔT	Δx	$e_{\max,\rho}$	q_ρ	$e_{\max,c}$	q_c
$1.\bar{6}E(-2)$	$1E(-2)$	$4.2(-3)$		$1.4(-4)$	
$8.\bar{3}E(-3)$	$5E(-3)$	$1.5(-5)$	8.1	$1.5(-6)$	6.6

As we now require more corrections, Strang splitting was carried out with 24 times as many steps as taken by the SDC solver - precisely 1008 and 2016 steps compared with the 42 and 84 which produced the results in Table 2. For the coarser grid, the results with Strang splitting were slightly more accurate than those obtained with the proposed method. On the finer grid, however, the SDC results were about an order of magnitude better.

6.3. ***Pulsating flame with stiff kinetics.*** Lastly, we consider a simplified thermo-diffusive combustion model with a stiff, intermediate reaction (e.g., [4]):

$$Y_t = \frac{1}{\mathscr{L}_Y}\left(Y_{xx} + \frac{1}{x}Y_x\right) - \frac{V}{x}Y_x - k_1, \tag{6–14}$$

$$W_t = \frac{1}{\mathscr{L}_W}\left(W_{xx} + \frac{1}{x}W_x\right) - \frac{V}{x}W_x + k_1 - k_2, \tag{6–15}$$

$$\Theta_t = \Theta_{xx} + \frac{1}{x}\Theta_x - \frac{V}{x}\Theta_x + k_1 + \gamma(k_2 - k_1), \tag{6–16}$$

$$k_1 = A_1 Y \exp\left(E_1 \frac{(1-\sigma)(\Theta-1)}{\sigma + \Theta(1-\sigma)}\right). \tag{6–17}$$

$$k_2 = A_2 W^2 \exp\left(E_2 \frac{(1-\sigma)(\Theta-1)}{\sigma + \Theta(1-\sigma)}\right). \tag{6–18}$$

The parameters are taken to be:

$$V = 11.7, \quad E_1 = 40, \quad E_2 = 2, \quad \mathscr{L}_Y = 2, \quad \mathscr{L}_W = \frac{3}{2}, \tag{6–19}$$

$$\sigma = \frac{1}{2}, \quad A_1 = 2 \times 10^2, \quad A_2 = 2 \times 10^6. \tag{6–20}$$

Note that if we assume the fast reaction is in balance with the slow reaction, that is if we assume $k_1 = k_2$, we obtain a reduced model with one species at a Lewis number, $\mathscr{L}_Y = 2$, with a pulsating solution (e.g., [5]). Initial and Dirichlet boundary conditions were obtained by interpolating a pulsating solution of the reduced problem on the spatial domain $5 \leq x \leq 35$. The initial W profile is then obtained through the quasiequilibrium assumption and the full system is evolved

Table 3. Observed maximum relative errors for $t \leq 20$, $m = 4$ preconditioned Radau methods applied to the pulsating flame problem. The order is calculated by $q = \log{(e_{N_2}/e_{N_1})}/\log{(N_1/N_2)}$ where N is the number of time steps and e_N is the maximum absolute error.

$\Delta T = 3.2E(-3)$, $\Delta x = 1.2E(-2)$			$\Delta T = 1.6E(-3)$, $\Delta x = 6.0E(-3)$			q		
Y	W	Θ	Y	W	Θ	Y	W	Θ
3.1(−4)	5.0(−3)	1.0(−4)	2.5(−5)	3.2(−4)	7.9(−6)	3.6	4.0	3.7

up to $t = 20$. Plots of the computed profiles on the finest grids, $\Delta T = 4.1\bar{6}E(-5)$, $\Delta x = 1.25E(-3)$, illustrating the flame oscillation are presented in Figure 16.

Note that around $t = 3.5$ the quasisteady initial flame destabilizes and moves towards the fuel source. An oscillation is set up between times 13 and 19.

The preconditioner in this case is as in the previous examples; second order spatial derivatives are approximated by $\gamma_d d_{2,2}$ and first order by $\gamma_c d_{1,2}$. We also include the Jacobian of the reaction terms. As in the preceding case we found it better to use a second order starting method. Choosing $\gamma_d = 2$ the time step limits were independent of γ_c and Δx, with a minimum time step of approximately $\Delta T = 3.8E(-3)$. Choosing $\gamma_d = 1$, on the other hand, required $\Delta T \propto \Delta x^2$ as in the previous examples.

We compare the accuracy of results obtained with $\Delta T = 3.2E(-3)$, $\Delta x = 1.2E(-2)$ and $\Delta T = 1.6E(-3)$, $\Delta x = 6E(-3)$. Here we have taken $\gamma_d = \gamma_c = 2$. The observed maximum errors, listed in Table 3, are consistent with 4th rather than 7th order convergence. This is the order of convergence expected for highly stiff problems, being equal to the stage order of the associated Runge–Kutta method. We note that, as might be expected for an oscillatory solution, the maximum errors are out of phase and occur at very different times for the two resolutions. Recently, Huang et al. [17] have shown how the order reduction phenomenon can be eliminated through the use of GMRES-based convergence acceleration which would no doubt improve our results in this case.

We have also solved this problem using Strang splitting and 24 times as many steps. As in the previous example, the results are slightly more accurate for the coarse resolution but less accurate for the fine resolution. We are confident that an improved implementation of the SDC method as in [17] would prove to be significantly more efficient than the traditional method.

7. Conclusion

In summary, we have shown that spectral deferred correction applied to a first order splitting method can:

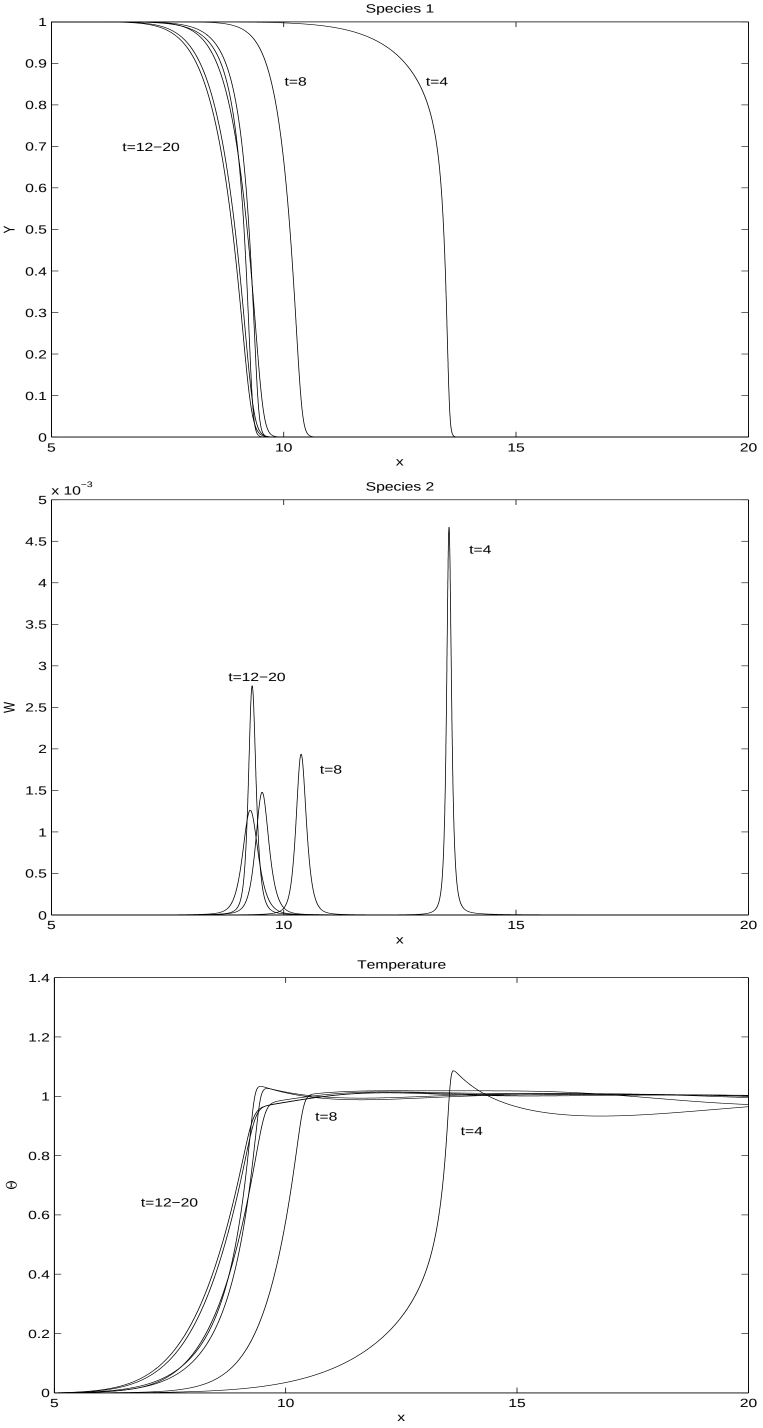


Figure 16. Fine grid solution of the flame equation.

i: Attain the full accuracy of the underlying quadrature rule;

ii: Have large stability domains.

We have also explored a general and flexible technique based on the concept of splitting by preconditioning. We have demonstrated the effectiveness of a particular instance of this strategy for reaction-advection-diffusion equations in one space dimension where high order difference approximations were preconditioned by lower order approximations with far narrower bandwidths. So long as the preconditioner was large enough in comparison with the true Jacobian, time step stability constraints independent of the spatial deiscretization were observed. This is in line with our experience solving complex combustion models [12; 11; 24]. Moreover, despite the very simple choice for the preconditioner and the fact that no convergence acceleration was employed, the methods were always as efficient and in some instances far more efficient than the standard Strng splitting approach.

Of course the greatest potential payoffs in terms of efficiency are for problems in multiple space dimensions. The fundamental issue is how simple (that is inexpensive) a preconditioner can be used without sacrificing too much accuracy or stability. It is also of interest to combine the preconditioned time-stepping strategy with the GMRES-based acceleration techniques described in [17]. We believe these issues deserve further study.

References

[1] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Numer. Anal. **32** (1995), no. 3, 797–823. MR 96j:65076 Zbl 0841.65081

[2] W. Auzinger, H. Hofstätter, W. Kreuzer, and E. Weinmüller, *Modified defect correction algorithms for ODEs, II: Stiff initial value problems*, Tech. Report ANUM 2/03, Vienna University of Technology, 2003.

[3] ———, *Modified defect correction algorithms for ODEs, I: General theory*, Tech. report, Numer. Alg., 2004.

[4] A. Bayliss, M. Garbey, and B. Matkowsky, *Adaptive pseudo-spectral domain decomposition and the approximation of multiple layers*, J. Comput. Phys. **119** (1995), 132–141.

[5] A. Bayliss, D. Gottlieb, B. Matkowsky, and M. Minkoff, *An adaptive pseudo-spectral method for reaction diffusion problems*, J. Comput. Phys. **81** (1989), 421–443.

[6] A. Bourlioux, A. T. Layton, and M. L. Minion, *High-order multi-implicit spectral deferred correction methods for problems of reactive flow*, J. Comput. Phys. **189** (2003), no. 2, 651–675. MR 2004f:76084 Zbl 1061.76053

[7] A. Dutt, L. Greengard, and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT **40** (2000), no. 2, 241–266. MR 2001e:65104 Zbl 0959.65084

[8] J. Frank, W. Hundsdorfer, and J. G. Verwer, *On the stability of implicit-explicit linear multistep methods*, Appl. Numer. Math. **25** (1997), no. 2-3, 193–205. MR 98m:65126 Zbl 0887.65094

[9] R. Frank, J. Hertling, and H. Lehner, *Defect correction algorithms for stiff ordinary differential equations*, Defect correction methods (Oberwolfach, 1983), Comput. Suppl., no. 5, Springer, Vienna, 1984, pp. 33–41. MR 86e:65094

[10] A. Hagstrom and G. Hagstrom, *Grid stabilization of high-order one-sided differencing, I: First order hyperbolic systems*, preprint, 2005.

[11] T. Hagstrom, K. Radhakrishnan, S. Steinberg, and R. Zhou, *Simulation of unsteady combustion phenomena using complex models*, Tech. Report 99-2397, AIAA, 1999.

[12] T. Hagstrom, K. Radhakrishnan, and R. Zhou, *Computation of steady and unsteady laminar flames: theory*, Tech. Report 1998-3246, AIAA, 1998.

[13] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations. I*, Springer Series in Computational Mathematics, no. 8, Springer, Berlin, 1993. MR 94c:65005

[14] E. Hairer and G. Wanner, *Solving ordinary differential equations. II*, Springer Series in Computational Mathematics, no. 14, Springer, Berlin, 1996. MR 97m:65007

[15] A. Hansen and J. Strain, *Convergence theory for spectral deferred correction*, Preprint, 2006.

[16] ———, *On the order of deferred correction*, Preprint, 2006.

[17] J. Huang, J. Jia, and M. Minion, *Accelerating the convergence of spectral deferred correction methods*, J. Comput. Phys. **214** (2006), 633–656.

[18] W. Hundsdorfer and J. Verwer, *Numerical solution of time-dependent advection-diffusion-reaction equations*, Springer Series in Computational Mathematics, no. 33, Springer, Berlin, 2003. MR 2004g:65001

[19] A. T. Layton and M. L. Minion, *Conservative multi-implicit spectral deferred correction methods for reacting gas dynamics*, J. Comput. Phys. **194** (2004), no. 2, 697–715. MR 2004k:76089 Zbl 02056059

[20] ———, *Implications of the choice of quadrature nodes for Picard integral deferred corrections methods for ordinary differential equations*, BIT **45** (2005), no. 2, 341–373. MR 2006h:65087 Zbl 1078.65552

[21] M. L. Minion, *Semi-implicit spectral deferred correction methods for ordinary differential equations*, Commun. Math. Sci. **1** (2003), no. 3, 471–500. MR 2005f:65085 Zbl 1088.65556

[22] D. L. Ropp, J. N. Shadid, and C. C. Ober, *Studies of the accuracy of time integration methods for reaction-diffusion equations*, J. Comput. Phys. **194** (2004), no. 2, 544–574. MR MR2034857 Zbl 1039.65069

[23] P. E. Zadunaisky, *On the estimation of errors propagated in the numerical integration of ordinary differential equations*, Numerische Math. **27** (1976/77), no. 1, 21–39. MR 55 #4691 Zbl 0324.65035

[24] R. Zhou, T. Hagstrom, K. Radhakrishnan, and S. Steinberg, *Numerical methods for reaction-diffusion equations with complex models*, In preparation, 2006.

[25] R. Zhou, M. G. Forest, and Q. Wang, *Kinetic structure simulations of nematic polymers in plane Couette cells. I. The algorithm and benchmarks*, Multiscale Model. Simul. **3** (2005), no. 4, 853–870. MR 2006c:76010 Zbl 02212496

Received August 23, 2005.

THOMAS HAGSTROM: hagstrom@math.unm.edu
Department of Mathematics and Statistics, The University of New Mexico, Albuquerque, NM 87131, United States

RUHAI ZHOU: *Department of Mathematics and Statistics, Old Dominion University,*
Norfolk, VA 23529, United States
rzhou@odu.edu

COMM. APP. MATH. AND COMP. SCI.
Vol. 1, No. 1, 2006

A COMPARISON OF THE EXTENDED FINITE ELEMENT METHOD WITH THE IMMERSED INTERFACE METHOD FOR ELLIPTIC EQUATIONS WITH DISCONTINUOUS COEFFICIENTS AND SINGULAR SOURCES

BENJAMIN LEROY VAUGHAN, JR.,
BRYAN GERARD SMITH AND DAVID L. CHOPP

We compare the Immersed Interface Method (IIM) with the Extended Finite Element Method (X-FEM) for elliptic equations with singular sources and discontinuous coefficients. The IIM has been compared favorably with a number of other competing methods. These methods are of particular interest because they allow for the solution of elliptic equations with internal boundaries on nonconforming meshes. In the context of moving interface problems, the emphasis in this paper is placed on accuracy of solutions and their normal derivatives on the interface. These methods are briefly described and the results for benchmark problems are compared.

1. Introduction

Consider the elliptic equation

$$\nabla \cdot (\beta \nabla u) + \kappa u = f \tag{1}$$

in a domain Ω in two dimensions. Embedded within Ω, there is an interface Γ_I (see Figure 1). The coefficients β, κ, and f may be discontinuous across Γ_I and jump conditions are given on the interface.

This type of problem arises in a broad spectrum of mathematical models and hence, a wide range of numerical methods have been devised to solve it. Often, the location of Γ_I varies in time. As a result, methods which are easily adapted to an arbitrary Γ_I are important. Of particular note in this area is the Immersed Interface Method (IIM) [7], which has been shown to perform very well against competing

MSC2000: 65N06, 65N30, 65N50.

Keywords: cartesian grids, discontinuous coefficient, elliptic equation, extended finite element method, finite difference methods, finite element methods, immersed interface method, immersed boundary method, irregular domain, level set methods, singular source term.

Vaughan's work was supported in part by a grant from the NIH under contract #R01-GM067248. Smith's work was supported in part by the DoD NDSEG Fellowship Program and the Chicago Chapter of the ARCS Foundation.

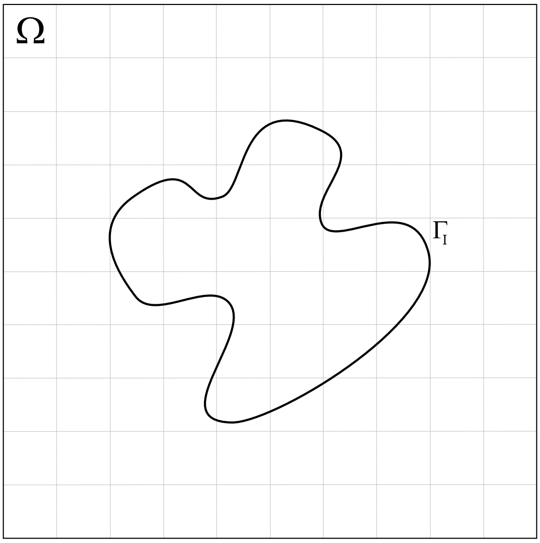


Figure 1. Domain Ω with interface Γ_I.

algorithms [7; 9]. It is representative of a class of methods that are constructed to be globally second order but locally first order on the interface.

In this paper, we compare the Extended Finite Element Method (X-FEM) [11; 3] and the IIM. The X-FEM is a variation on the partition of unity method [10] and has been used for the solution of crack growth problems [11; 2; 17; 15], arbitrary fixed material interfaces and voids [16], solidification problems [5; 4], and modeling rigid particles in Stokes flow [18].

These two methods offer similar advantages in that they both produce accurate solutions without the need for a conforming mesh. This makes them particularly attractive for coupling to methods for moving interfaces, e.g. the level set method [12].

This paper is organized as follows: Sections 2 and 3 discuss the IIM and X-FEM, respectively. A comparison of the numerical results for various types of problems is given in Section 4. Finally, Section 5 gives a summary and concluding remarks.

2. The immersed interface method

The Immersed Interface Method is a finite difference method for approximating the solution to (1). It was introduced in [7] and a detailed overview can be found in [9].

The method solves (1) with singular sources and discontinuous coefficients as well as jump conditions given on the interface by using a regular cartesian grid that does not conform to the interface. For grid points away from the interface, the standard five-point finite difference stencil is used. As a result, the method is second order away from the interface. For grid points near the interface, a six-point stencil and correction terms are added to the right hand side in order to maintain global second order accuracy.

2.1. ***Stencil generation.*** For simplicity, suppose the domain Ω is a square with space step of length h in both the x and y directions, and let the grid points be located at points (x_i, y_j). In general, the goal is to develop a finite difference equation of the form

$$\gamma_{1,0}u_{i+1,j} + \gamma_{-1,0}u_{i-1,j} + \gamma_{0,1}u_{i,j+1} + \gamma_{0,-1}u_{i,j-1} \\ + \gamma_{0,0}u_{i,j} + \gamma_{\pm1,\pm1}u_{i\pm1,j\pm1} + \kappa_{i,j}u_{i,j} = f_{i,j} + C_{i,j}$$

for the grid point at $\left(x_i, y_j\right)$. Here, only one combination of ±1 is used in the subscripts above which corresponds to the extra point in the stencil as described below.

For points away from the interface, i.e. a point where the interface does not come between any points in the standard five-point stencil, the standard five-point stencil

$$\frac{1}{h}\left(\left(\beta_{i+1/2,j}\frac{u_{i+1,j}-u_{i,j}}{h} - \beta_{i-1/2,j}\frac{u_{i,j}-u_{i-1,j}}{h}\right)\right. \\ \left. + \left(\beta_{i,j+1/2}\frac{u_{i,j+1/2}-u_{i,j}}{h} - \beta_{i,j-1/2}\frac{u_{i,j}-u_{i,j-1}}{h}\right)\right) + \kappa_{i,j}u_{i,j} = f_{i,j},$$

with $C_{i,j} = \gamma_{\pm1,\pm1} = 0$, is used.

For a grid point which bounds a square cut by the interface, the finite difference equation is generated by using a first order expansion of the equation about some point (x^*, y^*) on the interface. The point is chosen to be the point on the interface closest to the grid point (x_i, y_i) as shown in Figure 2. To achieve global second order accuracy, a set of equations is solved to generate the coefficients $\gamma_{k,\ell}$ and $C_{i,j}$.

First, a new transformed coordinate system is introduced. Let θ be the angle between the x-axis and the normal direction as shown in Figure 2. The transformation is:

$$\xi = \left(x - x_i^*\right)\cos\theta + \left(y - y_j^*\right)\sin\theta$$
$$\eta = -\left(x - x_i^*\right)\sin\theta + \left(y - y_j^*\right)\cos\theta$$

After the transform, the truncation error is of the form

$$T_{i,j} = a_1u^- + a_2u^+ + a_3u_\xi^- + a_4u_\xi^+ + a_5u_\eta^- + a_6u_\eta^+ + a_7u_{\xi\xi}^- + a_8u_{\xi\xi}^+ \\ + a_9u_{\eta\eta}^- + a_{10}u_{\eta\eta}^+ + a_{11}u_{\xi\eta}^- + a_{12}u_{\xi\eta}^+ + \kappa^-u^- - f^- - C_{i,j} + O(h)$$

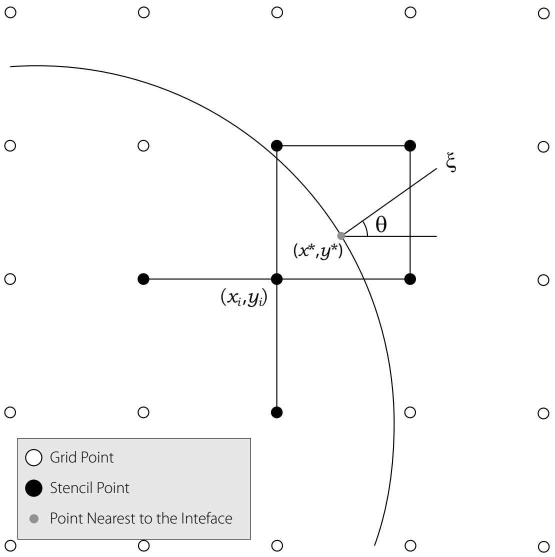


Figure 2. Geometry at a grid point (i, j) near the interface.

where a_j is given by

$$\begin{aligned}
a_1 &= \sum_{k\in K^-} \gamma_k & a_2 &= \sum_{k\in K^+} \gamma_k \\
a_3 &= \sum_{k\in K^-} \xi_k\gamma_k & a_4 &= \sum_{k\in K^+} \xi_k\gamma_k \\
a_5 &= \sum_{k\in K^-} \eta_k\gamma_k & a_6 &= \sum_{k\in K^+} \eta_k\gamma_k \\
a_7 &= \frac{1}{2}\sum_{k\in K^-} \xi_k^2\gamma_k & a_8 &= \frac{1}{2}\sum_{k\in K^+} \xi_k^2\gamma_k \\
a_9 &= \frac{1}{2}\sum_{k\in K^-} \eta_k^2\gamma_k & a_{10} &= \frac{1}{2}\sum_{k\in K^+} \eta_k^2\gamma_k \\
a_{11} &= \sum_{k\in K^-} \xi_k\eta_k\gamma_k & a_{12} &= \sum_{k\in K^+} \xi_k\eta_k\gamma_k
\end{aligned}$$

and the sets K^+ and K^- are defined as

$$K^{\pm} = \{k : (\xi_k, \eta_k) \text{ is on the } \pm \text{ side of } \Gamma_I\}$$

In order to ensure $T_{i,j} = O(h)$, the coefficients of u^-, u^+, u_ξ^-, u_η^-, $u_{\xi\xi}^-$, $u_{\xi\eta}^-$, and $u_{\eta\eta}^-$ must vanish as well as the constant terms. This gives the following six equations for the unknowns $\gamma_k, \cdots, \gamma_k$:

$$a_1 + a_2 - a_8 [\kappa] / \beta^+ = 0 \tag{2}$$

$$\begin{aligned} a_3 + \rho a_4 + a_8 \left(\beta_\xi^- - \rho\beta_\xi^+ - [\beta]\chi''\right) / \beta^+ & \\ + a_{10} [\beta] \chi'' / \beta^+ + a_{12} \left(\beta_\eta^- - \rho\beta_\eta^+\right) / \beta^+ &= \beta_\eta^- \end{aligned} \tag{3}$$

$$a_5 + a_6 - a_8 \left[\beta_\eta\right] / \beta^+ + a_{12} (1-\rho) \chi'' = \beta_\eta^- \tag{4}$$

$$a_7 + a_8 \rho = \beta^- \tag{5}$$

$$a_9 + a_{10} + a_8 (\rho - 1) = \beta^- \tag{6}$$

$$a_{11} + a_{12} \rho = 0 \tag{7}$$

where $\rho = \beta^- / \beta^+$ and χ'' is the curvature of the interface at (x^*, y^*).

Once the γ_j's are computed, $C_{i,j}$ can be obtained from

$$\begin{aligned} C_{i,j} = a_2 w + a_{12} \frac{v'}{\beta^+} + \left(a_6 - a_8 \frac{\beta_\xi^+}{\beta^+} + a_{12}\chi''\right) w' + a_{10} w'' & \\ + \frac{1}{\beta^+}\left(a_4 + a_8\left(\chi'' - \frac{\beta_\xi^+}{\beta^+}\right) - a_{10}\chi'' - a_{12}\frac{\beta_\eta^+}{\beta^+}\right) v & \\ + a_8 \left(\frac{[f]}{\beta^+} - \frac{\kappa^+}{\beta^+} w - w''\right) & \end{aligned} \tag{8}$$

where w and v are defined from the jump conditions on the interface:

$$w(\eta) = u^+ - u^-$$

$$v(\eta) = \beta^+ \frac{\partial u}{\partial \hat{n}}^+ - \beta^- \frac{\partial u}{\partial \hat{n}}^-$$

For a detailed derivation of these equations, see [7].

To summarize, using (2)–(7) to solve for γ_k and (8) for $C_{i,j}$, the stencil and the right hand side corrections are obtained. For continuous coefficients β and κ, the five-point stencil is obtained while a six-point stencil is needed if they are discontinuous. These stencils and the correction term, $C_{i,j}$, is used to assemble the linear system to solve for the values $u_{i,j}$ at the grid points.

3. The extended finite element method

The second method used in this paper is the Extended Finite Element Method (X-FEM). Like the Immersed Interface Method, the X-FEM can use a regular cartesian mesh that does not conform to the interface. Note that the X-FEM can also be used on arbitrary triangulated meshes as well. Since there is no comparable review article discussing this method in detail, we provide here a little more detail on its implementation.

In contrast to finite element meshes, where the mesh conforms to the interface, the X-FEM uses a fixed mesh which does not need to conform to the interface. This is done by extending the standard finite element approximation with extra basis functions on certain "enriched" nodes that capture the behavior of the solution near the interface. This is particularly useful for problems involving moving interfaces where the mesh would otherwise require regeneration every time step. We present here a summary of the method described in [2; 3; 11] with some slight modifications. While the discussion here will focus on 2D problems, it should be noted that this method can be readily applied to 3D as well.

Consider solving (1) on a rectangular domain Ω in two dimensions with Dirichlet boundary conditions applied to the domain boundary $\partial\Omega$. The X-FEM approximation of u is

$$u^h(x,y) = \sum_{n_i \in N} \phi_i(x,y)\, u_i + \sum_{n_j \in N_E} \phi_j(x,y)\, \psi(\varphi)\, a_j \tag{9}$$

where n_i and n_j are the i-th and j-th nodes of their respective sets, N is the set of all nodes in the domain, N_E is the set of enriched nodes, ϕ is a standard finite element basis function (i.e., bilinear or biquadratic), ψ is the enrichment function (described in Section 3.1), and φ is the signed distance function from the interface. The variables u_i and a_j are the unenriched and enriched degrees of freedom, respectively. Also, multiple enrichment functions can be used in the same X-FEM approximation while in this paper, only one is used at a time.

The domain Ω may be meshed by an arbitrary finite element mesh, but in this paper it is meshed with regular rectangular elements independent of the interface. The interface Γ_I is represented by a signed distance function φ and within each element cut by the interface, Γ_I is interpolated as a single line segment.

3.1. ***Enrichments.*** To include the interface's effect, enrichment functions are added to the standard finite element approximation for each element cut by the interface (Figure 3). The choice of enrichment function is based on the behavior of the solution near the interface. In this paper, two enrichment functions are used: a discontinuous, generalized Heaviside function or step function [17] and a continuous ramp function [5]. More application specific enrichment functions can also be used, e.g., a square root singularity function around crack tips [1]. Each of these is a function of the signed distance from the interface given as

$$\varphi(\boldsymbol{x}) = \pm \min_{\boldsymbol{X} \in \Gamma_I} \|\boldsymbol{x} - \boldsymbol{X}\|$$

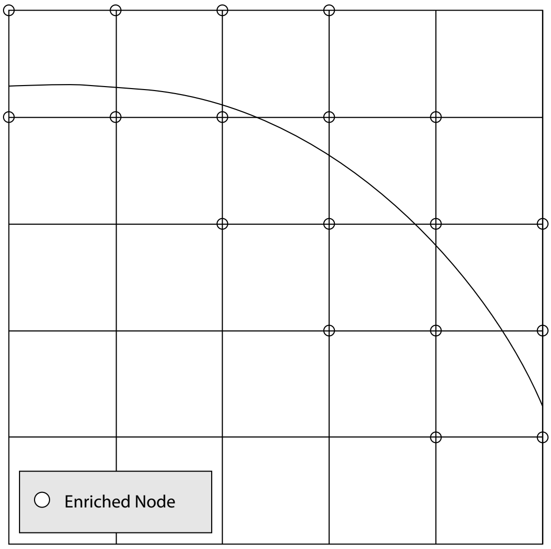


Figure 3. Enriched nodes.

where the sign is positive (negative) if $\boldsymbol{x}$ is outside (inside) the region enclosed by the interface Γ_I. For moving interface problems, the signed distance function is provided directly by the Level Set Method.

The step enrichment function is defined as:

$$\psi_{\text{Step}}(\varphi) = \begin{cases} 1 & \varphi > 0 \\ -1 & \varphi \le 0 \end{cases}$$

This enrichment function can yield a continuous or discontinuous solution across the interface but requires Lagrange multipliers to apply the Dirichlet jump condition.

The ramp function is defined as:

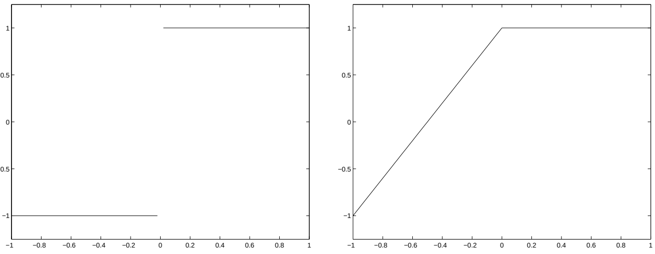
$$\psi_{\text{Ramp}}(\varphi) = \begin{cases} 1 & \varphi > 0 \\ 1 - 2\varphi & \varphi \le 0 \end{cases}$$

This enrichment function yields only continuous solutions. The advantage is that it automatically satisfies the continuity condition $[\![u]\!] = 0$ and does not require the use of Lagrange multipliers.

3.2. ***Element matrices.*** Using the weak form of (1), there are two types of integral terms: domain and interface.

All the matrices computed from the integral terms are block matrices of the form

$$A = \begin{bmatrix} A^{UU} & A^{UA} \\ A^{AU} & A^{AA} \end{bmatrix}$$



(a) Step function (b) Ramp function

Figure 4. Enrichment functions.

where A^{UU} is the standard FEM element matrix. The A^{UA}, A^{AU}, and A^{AA} matrices are the new matrix terms that arise from the addition of the enriched degrees of freedom. Note that the enriched matrix terms only appear when an element has enriched degrees of freedom and are much smaller than the standard FEM matrix term.

The vector terms also have the same form

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}^U \\ \boldsymbol{v}^A \end{bmatrix}$$

where $\boldsymbol{v}^U$ is the standard FEM element vector and $\boldsymbol{v}^A$ is the vector term from the enriched degrees of freedom.

3.2.1. *Domain integrals.* The following domain integral terms come from the Laplacian operator $\nabla \cdot (\beta \nabla u)$:

$$K_{i,j}^{UU} = -\int_{\Omega_E} \beta \left[\nabla\phi_i \cdot \nabla\phi_j\right] \partial\Omega_E$$

$$K_{i,j}^{UA} = -\int_{\Omega_E} \beta \left[\nabla\phi_i \cdot \nabla\left(\phi_j \psi_j\right)\right] \partial\Omega_E = K_{j,i}^{AU}$$

$$K_{i,j}^{AA} = -\int_{\Omega_E} \beta \left[\nabla\left(\phi_i \psi_i\right) \cdot \left(\nabla\phi_j \psi_j\right)\right] \partial\Omega_E$$

From the mass operator κu, the matrices are:

$$M_{i,j}^{UU} = \int_{\Omega_E} \kappa \phi_i \phi_j \partial\Omega_E$$

$$M_{i,j}^{UA} = \int_{\Omega_E} \kappa \phi_i \phi_j \psi_j \partial\Omega_E = M_{j,i}^{AU}$$

$$M_{i,j}^{AA} = \int_{\Omega_E} \kappa \phi_i \psi_i \phi_j \psi_j \partial\Omega_E$$

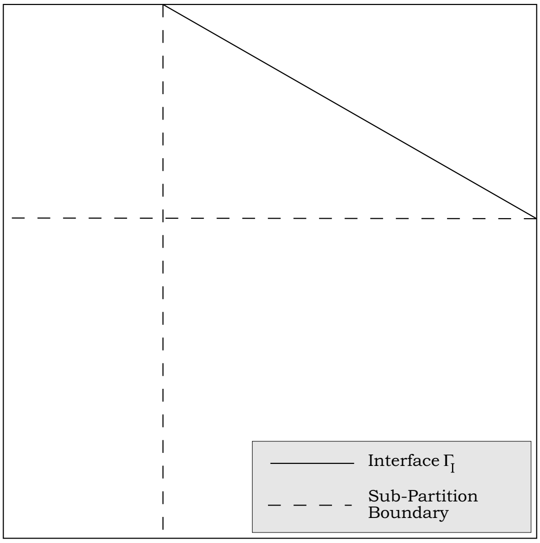


Figure 5. Element subpartitions.

and from the force operator f, the vectors:

$$f_i^U = \int_{\Omega_E} \phi_i f(x, y)\, \partial\Omega_E$$

$$f_i^A = \int_{\Omega_E} \phi_i \psi_i f(x, y)\, \partial\Omega_E$$

3.2.2. *Element integration.* Evaluating the domain integral terms requires a numerical quadrature method. Elements away from the interface are evaluated using standard Gaussian quadrature in two dimensions.

Elements that are cut by the interface must be treated differently due to discontinuities in the coefficients and enrichment functions. The interface is first interpolated as a line segment and the element is then divided into triangles and quadrilaterals that conform to the interface as illustrated in Figure 5. The subdivisions are for integration only and do not introduce any extra degrees of freedom. This method is slightly different than the method used in [3] in that the elements are not partitioned strictly into triangles. In this method, quadrilaterals are used with triangles transformed into quads for integration using the method given in [14].

3.3. ***Interface conditions.*** After creating the element matrices for each element, the only remaining terms arise from the interface conditions. Enforcing the Dirichlet jump conditions are discussed in Section 3.4.

The Neumann jump condition, $[\![\beta\hat{n}\cdot\nabla u]\!] = v\,(x, y)$, is enforced by introducing a line source term with strength v. The term is of the form

$$\int_{\Gamma_I} v\,(x, y)\,\delta\,(\boldsymbol{x} - \boldsymbol{X}\,(\boldsymbol{s}))\,\partial\Gamma_I \tag{10}$$

where $\boldsymbol{X}\,(\boldsymbol{s})$ is the parameterized coordinates of the interface and the direction of integration is such that the normal points from the positive domain into the negative domain. This term is only added if a source term is not already in the equation and the Neumann jump condition is an external constraint.

Integrating (10) over each element yields the vector terms

$$\gamma_i^U = \int_{\Gamma_I} \phi_i v\,(x, y)\,\partial\Gamma_I$$

$$\gamma_i^A = \psi_i\,(0^-)\int_{\Gamma_I} \phi_i v\,(x, y)\,\partial\Gamma_I$$

where $\psi_i\,(0^-)$ indicates that the enrichment function is evaluated on the negative side of the interface.

3.4. ***Lagrange multipliers.*** Since the Dirichlet jump condition on the interface has not been satisfied when using step enrichments, Lagrange multipliers are used to enforce this condition.

Equations (1) and (10) are combined and rewritten as

$$\nabla\cdot(\beta\nabla u) + \kappa u + [\![u]\!]\,\lambda = f + \int_{\Gamma_I} v\,(x, y)\,\delta\,(\boldsymbol{x} - \boldsymbol{X}\,(\boldsymbol{s}))\,\partial\Gamma_I \tag{11}$$

where $v = \left[\!\!\left[\beta\frac{\partial u}{\partial n}\right]\!\!\right]$ and λ is the Lagrange multiplier used to enforce the jump in the solution.

First, a one dimensional mesh is laid down along the interface as shown in Figure 6 by using a piecewise linear interpolation of the interface within each rectangular element. Next, the Lagrange multipliers are approximated using a 1D finite element approximation

$$\lambda^h = \sum_{m_i\in M} \theta_i\lambda_i$$

where M is the set of all Lagrange multiplier nodes [6].

The jump in the solution $[\![u]\!] = w\,(x, y)$ yields

$$C = \begin{bmatrix} 0 \\ C^A \end{bmatrix}$$

where

$$C_{i,j}^A = \int_{\Gamma_I} \theta_j\phi_i\,[\![\psi_i]\!]\,\partial\Gamma_I$$

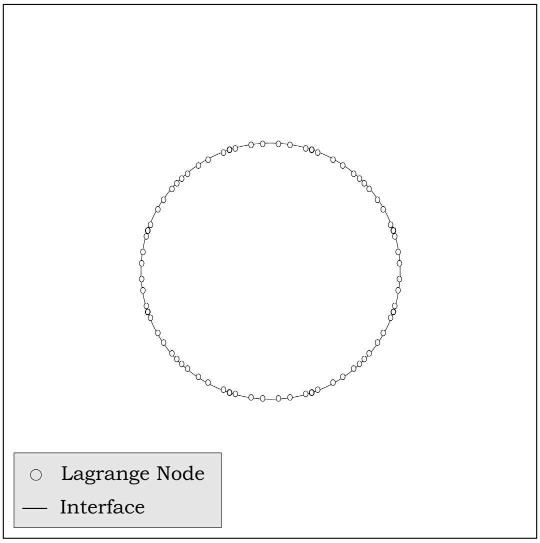


Figure 6. Lagrange multiplier mesh.

and the vector term

$$g_i = \int_{\Gamma_I} \theta_i w \partial\Gamma_I$$

3.5. ***Linear system.*** The resulting linear system contains terms (see Section 3.2) of the form

$$K = \begin{bmatrix} K^{UU} & K^{UA} \\ K^{AU} & K^{AA} \end{bmatrix}$$

$$M = \begin{bmatrix} M^{UU} & M^{UA} \\ M^{AU} & M^{AA} \end{bmatrix}$$

$$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{f}^U \\ \boldsymbol{f}^A \end{bmatrix}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}^U \\ \boldsymbol{\gamma}^A \end{bmatrix}$$

and the Lagrange multiplier terms C and $\boldsymbol{g}$.

The final assembled linear system is $A\boldsymbol{x} = \boldsymbol{b}$ where

$$A = \begin{bmatrix} K + M & C \\ (C)^T & 0 \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{a} \\ \boldsymbol{\lambda} \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{f} + \boldsymbol{\gamma} \\ \boldsymbol{g} \end{bmatrix}$$

and K, M, and C are all block matrices and $\boldsymbol{f}$ and $\boldsymbol{\gamma}$ are block vectors. When using ramp enrichments, the Lagrange multiplier terms, C and $\boldsymbol{g}$, along with the Lagrange degrees of freedom, λ, are not needed.

4. Results

In this section, the Immersed Interface Method and the X-FEM are compared on three example problems that are originally from [7]. For all the examples, a square domain is used with an embedded circular interface (Figure 7). Also, the results are confined to be on the interface since the results there are the most important for moving interface problems, and both methods become their standard counterparts away from the interface. In addition, since the solution of the linear system with the X-FEM requires very little time compared with the construction of the system, a direct linear solver is used for the example problems.

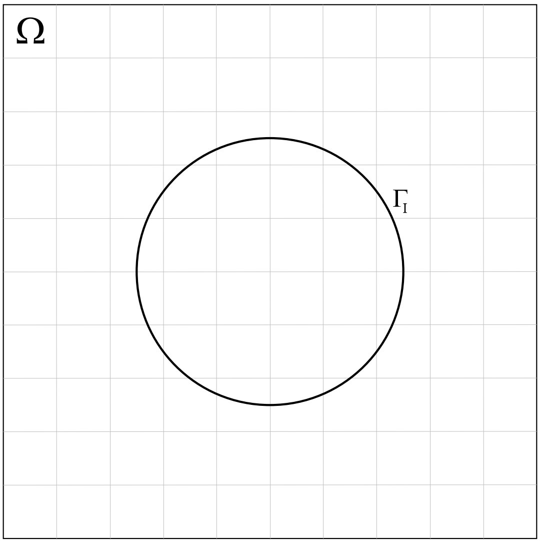


Figure 7. Domain Ω with interface Γ_I.

4.1. ***Example 1.*** The first example has a singular source on Γ_I. The differential equation is:

$$\nabla^2 u = \int_{\Gamma_I} \delta\left(r - R_I\right) \partial\Gamma_I \tag{12}$$

where Γ_I is a circle of radius $R_I = 1/2$ and δ is the Dirac delta function.

The solution to this equation is continuous, $[\![u]\!] = 0$, but the line source gives a jump in the normal

$$\left[\!\!\left[\frac{\partial u}{\partial n}\right]\!\!\right] = -2$$

The exact solution to (12) is:

$$u(x, y) = \begin{cases} 1 & r \leq \frac{1}{2} \\ 1 + \log(2r) & r > \frac{1}{2} \end{cases}$$

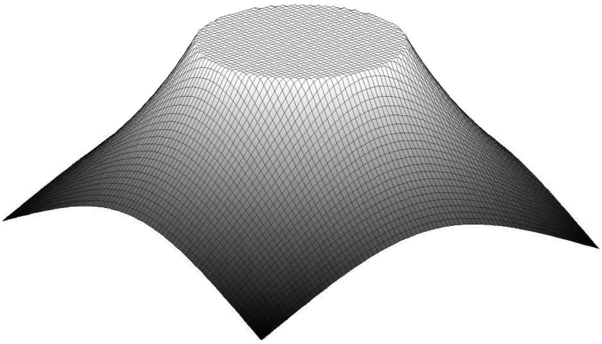


Figure 8. Solution for example 1.

Table 1 shows the convergence results for the X-FEM using four node bilinear elements with step and ramp enrichments. Piecewise constant Lagrange multipliers are used to enforce the Dirichlet jump conditions at the interface when using step enrichments. For comparison, convergence results for the Immersed Interface Method and the Immersed Boundary Method (IBM) [13] are shown. The error values for the Immersed Boundary Method data are taken from [7]. The error given is the maximum error at the nodes defined as

$$\|T_n\|_\infty = \max_{n_i \in N} \left\{ \left| u(x_i, y_i) - u_i^h \right| \right\}$$

where n_i is the i-th node with coordinates (x_i, y_i), N is the set of all nodes, and u_i^h is the computed solution at that node. In addition, the ratio of successive errors is given as $\|T_{2n}\| \,/\, \|T_n\|$.

Note that this is one of the two error measures that are given in [7]. The second, E_n, is the measure of the error at the nodes away from the interface. In this paper,

T_n is used since the error near or on the interface is of more concern for moving interface problems. In addition, the results for our implementation of the IIM differ from the results given in [7] possibly due to the representation of the interface, a signed distance function, and the choice of the point on the interface for computing the irregular stencil. This implementation does not converge as nicely but the error values are much smaller.

n	Step Enrichment		Ramp Enrichment	
	$\Vert T_n \Vert_\infty$	ratio	$\Vert T_n \Vert_\infty$	ratio
19	3.8397×10^{-3}		7.8138×10^{-3}	
39	9.3782×10^{-4}	4.0943	3.9577×10^{-3}	1.9743
79	2.3034×10^{-4}	4.0715	1.9029×10^{-3}	2.0798
159	6.4061×10^{-5}	3.5956	9.3797×10^{-4}	2.0287
319	1.5619×10^{-5}	4.1015	4.7646×10^{-4}	1.9686

n	IIM		IBM	
	$\Vert T_n \Vert_\infty$	ratio	$\Vert E_n \Vert_\infty$	ratio
19	3.1207×10^{-2}		3.6140×10^{-1}	
39	4.3918×10^{-3}	7.1057	2.6467×10^{-2}	12.7939
79	3.2066×10^{-3}	1.3696	1.3204×10^{-2}	2.0045
159	8.9322×10^{-4}	3.5899	6.6847×10^{-3}	1.9753
319	3.4105×10^{-4}	2.6190	3.3393×10^{-3}	2.0018

Table 1. Numerical results for example 1.

From Table 1, the X-FEM is shown to be first order with ramp enrichments and second order with step enrichments coupled with bilinear elements. Ramp enrichments give accuracy comparable with IIM but are only first order. On the other hand, step enrichments show second order accuracy and an order of magnitude improvement over IIM. The first order convergence for the IIM is expected since the error measure includes all the nodes near the interface where the approximation is only first order. Away from the interface both IIM and X-FEM converge second order. As shown before in [7], the IIM outperforms the IBM and consequently, the X-FEM is more accurate than IBM.

Notice that with the X-FEM, the choice of enrichments can change the convergence rate of the method. For this example, ramp enrichments converge first order while step enrichments converge second order. The cause of this is a subject of current research but it seems that extending the region where nodes are enriched, ie enriching nodes a certain distance from the interface but whose support is not necessarily cut by the interface, can regain the second order convergence for certain enrichments.

The X-FEM does show a slight increase in the linear system size. Table 2 gives the linear system size and its sparsity. It is seen that the enrichments and Lagrange multipliers introduce only a small number of new degrees of freedom (less than 2% for a 319×319 mesh).

n	Step Enrichment		Ramp Enrichment		IIM	
	Sys. Size	% Sparse	Sys. Size	% Sparse	Sys. Size	% Sparse
19	520	2.07396%	480	2.15625%	400	1.09250%
39	1,840	0.54277%	1,760	0.55191%	1,600	0.29234%
79	6,880	0.13840%	6,720	0.13940%	6,400	0.07558%
159	26,560	0.03490%	26,240	0.03501%	25,600	0.01921%
319	104,320	0.00876%	103,680	0.00877%	102,400	0.00484%

Table 2. System sizes for example 1.

Table 3 shows the errors interpolated on the interface using (9) for the X-FEM and the method described in [8] for the IIM. The interpolated value on the interface is important if the method is to be coupled with methods for evolving interfaces where the interface velocity is tied to the value at the interface. The interface is parameterized and the errors are computed at 10,000 evenly spaced points on the interface. It is seen that the X-FEM still maintains an order of magnitude improvement over IIM when using step enrichments and both maintain their respective convergence rates.

n	Step Enrichment		Ramp Enrichment		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	5.1857×10^{-3}		2.1871×10^{-2}		6.1970×10^{-2}	
39	1.2444×10^{-3}	4.1672	1.1708×10^{-2}	1.8680	7.5111×10^{-3}	8.2505
79	3.0043×10^{-4}	4.1421	6.0996×10^{-3}	1.9482	3.3766×10^{-3}	2.2245
159	8.8146×10^{-5}	3.4083	3.1101×10^{-3}	1.9612	1.1298×10^{-3}	2.9887
319	1.9315×10^{-5}	4.5636	1.6142×10^{-3}	1.9267	3.6684×10^{-4}	3.0798

Table 3. Interface results for example 1.

Table 4 gives the error in the normal derivative on the interface. This data is quite important when the speed of an evolving interface depends on the gradient of the solution at the interface, eg when the speed is derived from a potential. With the X-FEM using ramp enrichments and IIM, the normal derivative is not accurately captured with $O(1)$ errors, which is expected since the IIM is only an $O(h)$ method on the interface and taking the derivative costs the method an order of accuracy. On the other hand, using X-FEM with step enrichments captures the normal derivative with first order accuracy.

n	Step Enrichment		Ramp Enrichment		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	4.1828×10^{-1}		1.8292×10^{-0}		4.6176	
39	1.6067×10^{-1}	2.6033	1.6479×10^{-0}	1.1100	4.4095	1.0472
79	9.3826×10^{-2}	1.7124	1.3096×10^{-0}	1.2583	4.2222	1.0444
159	4.5301×10^{-2}	2.0712	1.4733×10^{-0}	0.8889	4.1219	1.0243
319	2.2290×10^{-2}	2.0323	1.3818×10^{-0}	1.0662	4.0640	1.0142

Table 4. Interface derivative results for example 1.

Since using step enrichments with the X-FEM yields much better accuracy while only slightly increasing the system size, the remaining examples will only use step enrichments with the X-FEM.

4.2. ***Example 2.*** The second example has discontinuous coefficients along with a singular source term. The equation is

$$\nabla \cdot (\beta \nabla u) = f + C \int_{\Gamma_I} \delta\left(\boldsymbol{x} - \boldsymbol{X}(s)\right) \partial \Gamma_I \tag{13}$$

where

$$f(x, y) = 8\left(x^2 + y^2\right) + 4$$

and

$$\beta(x, y) = \begin{cases} r^2 + 1 & r \leq \frac{1}{2} \\ b & r > \frac{1}{2} \end{cases}$$

with the following jump conditions

$$[u] = 0$$

$$\left[\left[\beta \frac{\partial u}{\partial n}\right]\right] = 0$$

The exact solution to (13) is:

$$u(x, y) = \begin{cases} r^2 & r \leq \frac{1}{2} \\ \frac{1}{4}\left(1 - \frac{1}{8b} - \frac{1}{b}\right) + \frac{1}{b}\left(\frac{r^4}{2} + r^2\right) + C \log(2r) & r > \frac{1}{2} \end{cases}$$

with $b = 10$ and $C = 0.1$.

Table 5 shows the results for the IIM and the X-FEM. Both methods handle the discontinuous variable coefficient with the IIM still being first order while the X-FEM achieves second order accuracy. The results are similar for interpolation on

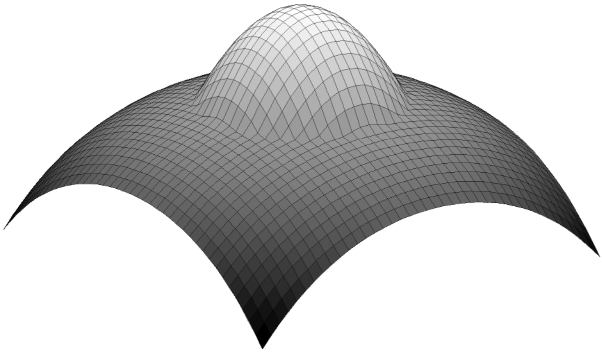


Figure 9. Solution for example 2.

n	X-FEM with steps		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	1.7613×10^{-3}		2.5520×10^{-2}	
39	4.1771×10^{-4}	4.2166	8.4159×10^{-3}	3.0324
79	1.0289×10^{-4}	4.0598	3.5290×10^{-3}	2.3848
159	3.0164×10^{-5}	3.4110	2.1227×10^{-3}	1.6625
319	6.7960×10^{-6}	4.4385	9.8789×10^{-4}	2.1487

Table 5. Numerical results for example 2.

the interface as show in Table 6. In addition, Table 7 shows the same convergence results as the previous example problem for evaluating the normal derivative on the interface with no convergence for the IIM and first order convergence for the X-FEM.

4.3. ***Example 3.*** For the third example, jumps in the function u are imposed on the interface Γ_I. The differential equation is

$$\nabla^2 u = 0 \tag{14}$$

with the jump conditions

$$[u] = e^x \cos y$$

$$\left[\left[\frac{\partial u}{\partial n}\right]\right] = 2e^x\,(x \cos y - y \sin y)$$

The exact solution of (14) is:

n	X-FEM with steps		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	1.6517×10^{-3}		2.5988×10^{-2}	
39	3.3824×10^{-4}	4.8832	8.8692×10^{-3}	2.9301
79	8.2238×10^{-5}	4.1129	3.6100×10^{-3}	2.4568
159	3.1568×10^{-5}	2.6051	2.1768×10^{-3}	1.6584
319	7.4612×10^{-6}	4.2310	1.0004×10^{-4}	2.1759

Table 6. Interface results for example 2.

n	X-FEM with steps		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	2.7307×10^{-1}		4.6176×10^{-0}	
39	1.2776×10^{-1}	2.1374	4.4095×10^{-0}	1.0472
79	6.1203×10^{-2}	2.0875	4.2222×10^{-0}	1.0444
159	4.8216×10^{-2}	1.2694	4.1219×10^{-0}	1.0243
319	2.4790×10^{-2}	1.9450	4.0640×10^{-0}	1.0142

Table 7. Interface derivative results for example 2.

$$u(x,y) = \begin{cases} e^x \cos y & r \le \frac{1}{2} \\ 0 & r > \frac{1}{2} \end{cases} \tag{15}$$

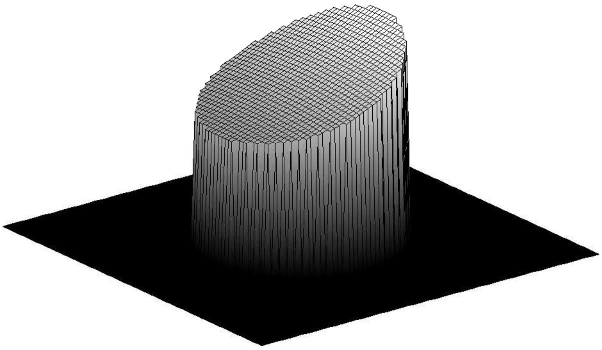


Figure 10. Solution for example 3.

Since (14) does not have a line source term explicitly, the equation is modified for the X-FEM to include one that yields the correct jump in the normal derivative. The new equation is

$$\nabla^2 u = \int_{\Gamma_I} 2e^x \left(x \cos y - y \sin y\right) \delta \left(\boldsymbol{x} - \boldsymbol{X}\left(x\right)\right) \partial\Gamma_I \tag{16}$$

n	X-FEM with steps		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	1.7648×10^{-4}		3.6253×10^{-3}	
39	6.0109×10^{-5}	2.9360	4.6278×10^{-4}	7.8337
79	1.7769×10^{-5}	3.3828	3.0920×10^{-4}	1.4967
159	4.8626×10^{-6}	3.6542	1.1963×10^{-4}	2.5846
319	1.2362×10^{-6}	3.9335	4.5535×10^{-5}	2.6272

Table 8. Numerical results for example 3.

n	X-FEM with steps		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	4.7842×10^{-4}		4.0230×10^{-3}	
39	1.0659×10^{-4}	4.4884	5.7563×10^{-4}	6.9889
79	2.8361×10^{-5}	3.7583	3.1617×10^{-4}	1.8206
159	7.3603×10^{-6}	3.8532	1.2004×10^{-4}	2.6339
319	2.0634×10^{-6}	3.5671	4.5526×10^{-5}	2.6367

Table 9. Interface results for example 3.

n	X-FEM with steps		IIM	
	$\|T_n\|_\infty$	ratio	$\|T_n\|_\infty$	ratio
19	5.6520×10^{-2}		$3.0009 \times 10^{+1}$	
39	2.4190×10^{-2}	2.3365	$5.5185 \times 10^{+1}$	0.5438
79	9.4512×10^{-3}	2.5595	$1.2034 \times 10^{+2}$	0.5392
159	7.1671×10^{-3}	1.3187	$2.6466 \times 10^{+2}$	0.4547
319	2.6865×10^{-3}	2.6678	$5.2870 \times 10^{+2}$	0.5006

Table 10. Interface derivative results for example 3.

Table 8 gives the results for the X-FEM and the IIM with the X-FEM is second order while the IIM is first order with the X-FEM giving about an order of magnitude better improvement at the nodes. The conclusions are similar for errors taken on the interface as given in Table 9. Table 10 contains the errors in the normal derivative on the interface for the X-FEM and the IIM. Once again, the X-FEM is first order when computing the normal derivative and the IIM is unable give an accurate evaluation of the normal derivative at the interface.

5. Conclusion

In this paper, the Extended Finite Element Method and the Immersed Interface Method were compared. Both methods use a regular cartesian mesh, which does not conform to an internal interface.

The Immersed Interface Method is a finite difference method that handles interfaces by using a six point stencil where needed, along with correction terms on the right hand side, to handle the jump conditions. It is second order accurate at the grid points away from the interface and first order accurate at the grid points near the interface.

The Extended Finite Element Method is a finite element method where extra "enriched" basis functions are added to the standard finite element approximation. These enrichment functions add discontinuities that approximate the behavior near the interface. These enrichments coupled with the enforcement of the interface conditions yields accurate results both near and away from the interface. In addition, the X-FEM is not restricted to enforcing only jump conditions on the interface in its formulation. The lack of this restriction allows explicit boundary conditions to be applied, which is a subject of current research.

Overall, the X-FEM performed well compared to the IIM. It provides second order accuracy at the nodes and on the interface while more accurately capturing the gradient on the interface for each of the problems. Against other methods like the IIM, which are constructed as second order methods away from the interface but only have a local $O(h)$ truncation error near the interface, the X-FEM maintains an advantage due it being second order on all the nodes including the ones near the interface. This is an advantage because an accurate approximation of the gradient at the interface is important for moving interface problems where the velocity is often derived from a gradient of the velocity potential. This makes the X-FEM a more attractive choice for coupling with moving interface methods such as the Level Set Method.

References

[1] T. Belytschko and T. Black, *Elastic crack growth in finite element with minimal remeshing*, International Journal of Numerical Methods in Engineering **45** (1999), 601–620.

[2] J. E. Dolbow, N. Moës, and T. Belytschko, *Discontinuous enrichment in finite elements with a partition of unity method*, Finite Elements in Analysis and Design **36** (2000), 235–260.

[3] ———, *An extended finite element method for modeling crack growth with frictional contact*, Computational Methods in Applied Mechanics and Engineering **190** (2001), 6825–6846.

[4] J. C. et. al., *The extended finite element method (xfem) for solidification problems.*, International Journal of Numerical Methods in Engineering **53** (2002), 1959–1977.

[5] H. Ji, D. Chopp, and J. E. Dolbow, *A hybrid extended finite element/level set method for modeling phase transformation*, International Journal of Numerical Methods in Engineering **54** (2002), 1209–1233.

[6] H. Ji and J. E. Dolbow, *On strategies for enforcing interfacial constraints and evaulating jump conditions with the extended finite element method*, International Journal of Numerical Methods in Engineering **61** (2004), 2508–2535.

[7] R. J. LeVeque and Z. Li, *The immersed interface method for elliptic equations with discontinuous coefficients and singular sources*, SIAM Journal Numerical Analysis **31** (1994), 1019–1044.

[8] R. J. LeVeque and Z. Li, *Immersed interface methods for stokes flow with elastic boundaries or surface tension*, SIAM Journal Scientific Computing **18** (1997), no. 3, 709–735.

[9] Z. Li, *An overview of the immersed interface method and its applications*, Taiwanese Journal of Mathematics **7** (2003), 1–49.

[10] J. M. Melenk and I. Babuška, *The partition of unity finite element method: Basic theory and applications*, Computational Methods in Applied Mechanics and Engineering **139** (1996), 289ï¿½314.

[11] N. Moës, J. Dolbow, and T. Belytschko, *A finite element method for crack growth without remeshing*, International Journal of Numerical Methods in Engineering **46** (1999), 131–150.

[12] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: algorithms base on hamilton-jacobi formulations*, Journal of Computational Physics **79** (1988), 12–49.

[13] C. S. Peskin, *Numerical analysis of blood flow in the heart*, Journal of Computational Physics **25** (1977), 220–252.

[14] H. T. Rathod, K. V. N. B. Venkatesudu, and N. L. Ramesh, *Gauss legendre quadrature over a triangle*, Journal Indian Institute of Science **84** (2004), 183–188.

[15] M. Stolarska, D. L. Chopp, N. Moës, and T. Belytschko, *Modelling crack growth by level sets in the extended finite element method*, International Journal of Numerical Methods in Engineering **51** (2001), 943–960.

[16] N. Sukumar, D. Chopp, N. Moës, and T. Belytschko, *Modeling holes and inclusions by level sets in the extended finite element method*, International Journal of Numerical Methods in Engineering **48** (2000), 1549–1570.

[17] N. Sukumar, D. L. Chopp, and B. Moran, *Extended finite element method and fast marching method for three-dimensional fatigue crack propagation*, Engineering Fracture Mechanics **70** (2003), 29–48.

[18] G. J. Wagner, N. Moës, W. K. Liu, and T. Belytschko, *The extended finite element method for rigid particles in stokes flow*, International Journal of Numerical Methods in Engineering **51** (2001), 293–313.

Received January 4, 2006.

BENJAMIN LEROY VAUGHAN, JR.: b-vaughan@northwestern.edu
Engineering Sciences and Applied Mathematics Dept., Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, United States

BRYAN GERARD SMITH: b-smith7@northwestern.edu
Engineering Sciences and Applied Mathematics Dept., Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, United States

DAVID L. CHOPP: chopp@northwestern.edu
Engineering Sciences and Applied Mathematics Dept., Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, United States

Communications in Applied Mathematics and Computational Science

Volume 1 No. 1 2006

1559-3940(2006)1:1;1-9