INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 7, 371–394 (1987)

THE CONSISTENT GALERKIN FEM FOR COMPUTING DERIVED BOUNDARY QUANTITIES IN THERMAL AND/OR FLUIDS PROBLEMS

PHILIP M. GRESHO AND ROBERT L. LEE

Lawrence Livermore National Laboratory, University of California, Livermore, California 94550, U.S.A.

AND

ROBERT L. SANI, MIRIAM K. MASLANIK AND BRIAN E. EATON

Cooperative Institute for Research in Environmental Sciences and Department of Chemical Engineering, University of Colorado/NOAA, Boulder, Colorado 80309, U.S.A.

SUMMARY

A consistent, accurate and reasonably simple method of obtaining derived quantities when the conventional Galerkin finite element method (GFEM) is used to obtain the primary quantities is defined and demonstrated, both theoretically and numerically.

KEY WORDS Consistent Flux Heat Flux Derived Quantities FEM

INTRODUCTION

Oftentimes the focal point of the analysis of a physical system is a derived quantity such as a flux, or force. In the corresponding numerical simulation the generation of such quantities directly from the solution can be plagued with accuracy and continuity problems; for example in deriving nodal fluxes related to a typical gradient transport phenomenon, a C^0 finite element representation of the solution leads to discontinuous nodal fluxes. The latter plus the superconvergence phenomena achieved in some problems on regular meshes has led to the common practice of using 'Gauss-point' fluxes. However, various authors, for example Wheeler,[1] Larock and Herrmann,[2] Marshall *et al.*,[3] Gresho *et al.*,[4] Thornton,[5] Kjaran and Sigurdsson[6] and Lynch,[7–9] have suggested and employed an alternative technique, herein referred to as the consistent (flux) method, which can lead to more accurate results (see also References 10–12). The application of such a technique to advection–diffusion as well as fluid flow problems is described in detail and is illustrated by selected examples.

THEORY

Steady-state heat conduction

Poisson equation with Dirichlet boundary conditions. To set the stage, we begin with the simplest

0271–2091/87/040371–24$12.00

Received 4 July 1986
Revised 10 November 1986

of problems: find $T(\mathbf{x})$ in Ω such that

$$\nabla^2 T + S(\mathbf{x}) = 0 \quad \text{in} \quad \Omega, \tag{1a}$$

$$T = T_0 \quad \text{on} \quad \Gamma, \tag{1b}$$

where the source term, $S(\mathbf{x})$, and T_0 are given and Γ is the boundary of Ω. Since we are interested in an FEM approximation to the solution of (1), we begin by restating the problem in the weak form. Let H_1 be the space of functions with L_2 derivatives in Ω and let H_1^{N+M} be a finite dimensional $(N+M)$ subspace of H_1; $H_1^1 \subset H_1^2 \subset H_1^3 \cdots \subset H_1$, such that, as $M, N \to \infty$, $H_1^{N+M} \to H_1$. The definitions of N and M are necessarily vague at this point; they will be made precise in due course. Let $\{\Phi_i\}$ be a set of functions which spans H_1^{N+M} (i.e. a basis). Also, let H_1^0 be a subspace of H_1 in which the functions vanish on Γ and let $H_1^{0,N}$ be the analogous finite-dimensional (N) subspace of H_1^0. Finally, let $\{\phi_i\}$ be a set of functions which spans $H_1^{0,N}$. Assuming that $\{\Phi_i\}$ and $\{\phi_i\}$ are constructed so that $\{\phi_i\} \subset \{\Phi_i\}$, we let $\{\Gamma_i\} \equiv \{\Phi_i\} - \{\phi_i\}$. The functions, $\{\Gamma_i\}$, of which only M are non-zero, are members of H_1^{N+M} and will play a key role in the analysis.

The finite-dimensional weak form of (1) may now be expressed as

$$\int_\Omega \nabla\phi \cdot \nabla\tilde{T} = \int_\Omega \phi S, \quad \forall\, \phi \in H_1^{0,N}, \tag{2}$$

where $\tilde{T} = T_\Omega + T_s$ is the approximate solution, $T_\Omega \in H_1^{0,N}$ is the 'interior' solution, and $T_s(\mathbf{x})$ is an H_1 extension of $T_0(\Gamma)$ into Ω such that $T_s(\Gamma)$ is 'close to' T_0 on Γ (in practice, $T_s(\Gamma)$ will interpolate $T_0(\Gamma)$ via the piecewise polynomial basis functions of FEM; hence $T_s(\Gamma)$ will be identical to $T_0(\Gamma)$ at a finite number (M) of points). Since T_s is presumed known, (2) can be rewritten as

$$\int_\Omega \nabla\phi \cdot \nabla T_\Omega = \int_\Omega \phi S - \int_\Omega \nabla\phi \cdot \nabla T_s, \quad \forall\, \phi \in H_1^{0,N}. \tag{3}$$

But since $\{\phi\}$ form a basis in $H_1^{0,N}$, $\phi = \sum_{j=1}^N a_j \phi_j$ for some $\{a_j\}$ and we therefore have, since ϕ is an arbitrary function in $H_1^{0,N}$,

$$\int_\Omega \nabla\phi_i \cdot \nabla T_\Omega = \int_\Omega \phi_i S - \int_\Omega \nabla\phi_i \cdot \nabla T_s; \quad i = 1, 2, \ldots, N. \tag{4}$$

In accordance with the Galerkin method, we expand T_Ω in the $\{\phi_i\}$ basis as

$$T_\Omega = \sum_{j=1}^N T_j \phi_j(\mathbf{x}), \tag{5}$$

and, for 'convenience', we express T_s as

$$T_s = \sum_{j=1}^M T_j^s \Gamma_j(\mathbf{x}), \tag{6}$$

where $\{T_j\}$ are to be determined and $\{T_j^s\}$ are presumed known (e.g. via interpolation). Inserting (5) and (6) into (4) leads to the Galerkin equations,

$$\sum_{j=1}^N T_j \int_\Omega \nabla\phi_i \cdot \nabla\phi_j = \int_\Omega \phi_i S - \sum_{j=1}^M T_j^s \int_\Omega \nabla\phi_i \cdot \nabla\Gamma_j; \quad i = 1, 2, \ldots, N, \tag{7}$$

for the amplitude coefficients, $\{T_j\}$. In the Galerkin FEM, $\{\Phi_i\}$ and $\{\phi_i\}$ are of course taken to be the piecewise polynomials associated with a discretized mesh of finite elements which approximate Ω, and for our purposes they are assumed to be taken such that there are N nodes in Ω and M nodes on Γ. In this case it is common and convenient to let (6) represent a basis function interpolant of T_0 on Γ and an H_1 extension of T_0 into Ω in such a way that T_s goes to zero as 'quickly as possible' away from Γ.

This in fact is the conventional GFEM, and (7) leads to the usual matrix problem, $\mathbf{KT} = \mathbf{f}$, the solution of which gives the global N-vector of the $\{T_i\}$, an approximation to the actual temperature in Ω.

We now address the problem of obtaining a consistent approximation to the 'derived variable',

$$\tilde{q} \equiv -\mathbf{n}\cdot\nabla\tilde{T}|_{\Gamma}, \tag{8}$$

the outward-directed normal heat flux on Γ ($\mathbf{n}$ is the outward unit normal vector on Γ), hereafter referred to simply as the heat flux. To this end, we construct another weak form of (1), in which the $\{\Gamma_i\}$ are employed as the test functions, i.e. we begin by assuming sufficient smoothness and write

$$\int_{\Omega} \Gamma_i \nabla^2 \tilde{T} + \int_{\Omega} \Gamma_i S = 0; \quad i = 1, 2, \ldots, M, \tag{9}$$

which, upon integration by parts and using (8), yields the appropriate weak form,

$$\int_{\Omega} \Gamma_i \tilde{q} = \int_{\Omega} \Gamma_i S - \int_{\Omega} \nabla\Gamma_i \cdot \nabla\tilde{T}; \quad i = 1, 2, \ldots, M, \tag{10}$$

where $\tilde{T} = \sum_{j=1}^{N} T_j \phi_j(\mathbf{x}) + \sum_{j=1}^{M} T_j^{\mathrm{s}} \Gamma_j(\mathbf{x})$ is known (via solving (7)) and we now need only C^0 smoothness for $\tilde{T}$.

Remark. As N and $M \rightarrow \infty$ for a fixed domain size (Ω), i.e. as the element size $\rightarrow 0$, the source term domain integral in (10)$\rightarrow 0$ whereas the other two terms converge to recover (8). For finite N and M, however, equation (10) represents the *consistently derived* heat flux, in that (i) it is the only heat flux which, if imposed as a Neumann boundary condition, will lead to the same $\{T_j\}$ as from the original Dirichlet problem (7), and (ii) it guarantees the appropriate approximation to the global heat balance implied by (1), namely

$$\int_{\Gamma} q = \int_{\Omega} S. \tag{11}$$

This is true simply as a consequence of the fact that

$$\sum_{j=1}^{N} \phi_j(\mathbf{x}) + \sum_{j=1}^{M} \Gamma_j(\mathbf{x}) = 1{\cdot}0 \tag{12}$$

both in Ω *and* on Γ (wherein every member of the first sum is zero). It is important to note that $\sum_{j=1}^{N} \phi_j(\mathbf{x}) \neq 1$ near Γ. Thus, the simple addition of all N equations of (4) and all M equations of (10) yields, using (12), the heat balance given by (11)—with, of course, the exact heat flux, q, replaced by the approximate heat flux, $\tilde{q}$. This is easier to see if (4) is first rewritten as

$$0 = \int_{\Omega} \phi_i S - \int_{\Omega} \nabla\phi_i \cdot \nabla\tilde{T}; \quad i = 1, 2, \ldots, N, \tag{13}$$

and all equations in (10) and (13) summed, i.e. (11) is *implied* by (4) and (10).

By contrast, the more common approximations to q, such as

$$q = -\left(\sum_{j=1}^{N} T_j \mathbf{n}\cdot\nabla\phi_j + \sum_{j=1}^{M} T_j^{\mathrm{s}} \mathbf{n}\cdot\nabla\Gamma_j\right)\Bigg|_{\Gamma}, \tag{14}$$

cannot be demonstrated to satisfy (11), and they generally will not. In fact, the use of (14) generally requires special procedures, since the derivatives of the basis functions usually do not display interelement continuity; typically either boundary Gauss points on Γ are employed (the recommended technique) or, if nodal values are desired, some sort of averaging procedure is usually required. The consistent flux method obviates these special procedures and accounts for the fact that we are dealing with a *weak* solution on a *finite* mesh. Another advantage of the consistent flux method is that the requirement for actually constructing $\mathbf{n}$ (which is generally tedious and often quite ambiguous, such as at sharp corners on Γ) is also completely eliminated; the method generates a best estimate to the normal flux at nodes even if the boundary is not smooth.

Thus far the flux calculation has been somewhat 'theoretical'. In order to actually *compute* the heat flux from (10), we expand q into the set $\{\Gamma_i\}$ *evaluated on* Γ (in which case the $\{\Gamma_i\}$ form a *basis* for the M-dimensional subspace of H_1 on Γ) as

$$\tilde{q} = \sum_{j=1}^{M} q_j \Gamma_j|_{\Gamma}, \tag{15}$$

which, when inserted into (10) yields

$$\sum_{j=1}^{M} q_j \int_{\Gamma} \Gamma_i \Gamma_j = \int_{\Omega} \Gamma_i S - \int_{\Omega} \nabla\Gamma_i \cdot \nabla\tilde{T}; \quad i = 1, 2, \ldots, M. \tag{16}$$

The contributions to (16) can be formed in the usual way (at element level) except that only those elements with nodes on Γ need be considered, owing to the compact support of $\{\Gamma_i\}$. This linear system can also be written as

$$\mathbf{Bq} = \mathbf{b}, \tag{17}$$

where B_{ij} is the boundary 'mass' matrix which couples q_i to its nearest neighbours of Γ. The solution of (17) gives the nodal values of the flux, q_j, $j = 1, 2, \ldots, M$, wherein it is noteworthy that the consistent boundary flux depends on more than just the normal gradient of T at the boundary. Shortly we will discuss how this apparently costly procedure can often be greatly simplified, but first we generalize the consistent flux method to a problem with mixed boundary conditions.

Poisson equation with mixed boundary conditions (*Dirichlet and Neumann*). Consider now the slight generalization of (1) given by: find $T(\mathbf{x})$ in Ω and on $\partial\Omega_2$ such that

$$\nabla^2 T + S(\mathbf{x}) = 0 \quad \text{in} \quad \Omega, \tag{18a}$$

$$T = T_0 \quad \text{on} \quad \partial\Omega_1 \tag{18b}$$

and

$$\mathbf{n}\cdot\nabla T = -q_2 \text{ on } \partial\Omega_2, \tag{18c}$$

where S, T_0 and q_2 (the outward normal flux on $\partial\Omega_2$) are given and $\Gamma = \partial\Omega_1 \oplus \partial\Omega_2$ is the boundary of Ω. In this case, we take H_1 as before, but H_1^0 is now the subspace of H_1 in which the functions vanish only on $\partial\Omega_1$. Defining again the appropriate finite-dimensional (piecewise polynomial) subspace leads to the following weak form:

$$\int_{\Omega} \nabla\phi_i \cdot \nabla\tilde{T} = \int_{\Omega} \phi_i S - \int_{\partial\Omega_2} \phi_i q_2; \quad i = 1, 2, \ldots, N, \tag{19}$$

where

$$\tilde{T} = T_{\Omega,\partial\Omega_2} + T_s = \sum_{j=1}^{N} T_j\phi_i(\mathbf{x}) + \sum_{j=1}^{M} T_j^s\Gamma_j(\mathbf{x}), \tag{20}$$

where the N nodes now include those in Ω *and* those on $\partial\Omega_2$; the M boundary nodes are on $\partial\Omega_1$. Inserting (20) into (19) yields the conventional GFEM equations to be solved for $\{T_j\}$; this is 'part 1' of the calculation. (Note that, as usual, (18c) has been incorporated into the approximate solution as a natural boundary condition.) Part 2 involves the computation of the consistent heat flux on $\partial\Omega_1$, given that $\tilde{T}$ is available. In a similar way as before, the appropriate weak form for $\tilde{q}_1 = -\mathbf{n}\cdot\nabla\tilde{T}|_{\partial\Omega_1}$ is

$$\int_{\partial\Omega_1}\Gamma_i\tilde{q}_1 = \int_{\Omega}\Gamma_i S - \int_{\Omega}\nabla\Gamma_i\cdot\nabla\tilde{T} - \int_{\partial\Omega_2}\Gamma_i q_2; \quad i = 1, 2, \ldots, M, \tag{21}$$

which, in conjunction with

$$\tilde{q}_1 = \sum_{j=1}^{M} q_{1j}\Gamma_j|_{\partial\Omega_1}, \tag{22}$$

leads to a system of M equations for $\{q_{1j}\}$, the nodal values of q_1 on $\partial\Omega_1$ (i.e. insert (22) into (21)). It is noteworthy that in (21), the contribution of $\int_{\partial\Omega_2}\Gamma_i q_2$ is zero except at those nodes which 'join' $\partial\Omega_1$ to $\partial\Omega_2$ (Γ_i is zero over most of $\partial\Omega_2$).

In this case, the appropriate global balance associated with (18) is realized by adding all of the equations in (19) and (21) to give, using (12) which is now valid in Ω, on $\partial\Omega_1$ and on $\partial\Omega_2$,

$$\int_{\partial\Omega_1}\tilde{q}_1 + \int_{\partial\Omega_2} q_2 = \int_{\Omega} S. \tag{23}$$

Note that, whereas $\tilde{q}_1$ must be expanded into the basis set $\{\Gamma_i|_{\partial\Omega_1}\}$, q_2 need not be so expanded (via interpolation) into the analogous functions $\{\Gamma_i|_{\partial\Omega_2}\}$ (e.g. if q_2 is given in functional form, it may be retained in this form, presumably leading to more accuracy, in (19) and (21)).

The additional important point associated with this mixed boundary condition problem is the close relationship between the consistent flux calculation of 'part 2' and the treatment of the natural boundary condition of 'part 1'. In fact, (21) is the appropriate equation for $\tilde{T}$ if the flux were specified on *all* of Γ; here of course Γ_i must be replaced by ϕ_i, $\tilde{q}_1$ by q_1, and the appropriate redefinitions of the subspaces used. These observations lend even more credibility to the claim that this is the *consistent* flux method.

These interpretations also lead to an alternative manner in which to view the *combined* problem (temperature and boundary heat flux) since (on the boundary) *either* the primary variable (temperature) is unknown (on $\partial\Omega_2$) *or* the secondary variable (normal heat flux) is unknown (on $\partial\Omega_1$), i.e. consider the following 'generalization' of the previous problem: find $T(\mathbf{x})$ in Ω and on $\partial\Omega_2$ and find q_1 on $\partial\Omega_1$ such that (18) is satisfied. The appropriate weak form is now generated using the larger space of test functions, $\{\Phi_i\}\in H_1^{N+M}$ and is obtained from (18a) via integration by parts and, using (18c) and $\tilde{q}_1 = -\mathbf{n}\cdot\nabla\tilde{T}|_{\partial\Omega_1}$, as

$$\int_{\Omega}\nabla\Phi_i\cdot\nabla\tilde{T} + \int_{\partial\Omega_1}\Phi_i\tilde{q}_1 = \int_{\Omega}\Phi_i S - \int_{\partial\Omega_2}\Phi_i q_2; \quad i = 1, 2, \ldots, N+M, \tag{24}$$

where $\tilde{T}$ is given by (20), $\tilde{q}_1$ by (22) and the identification of the $N+M$ nodes (and associated basis functions) is the same as given previously. Equation (24) appears to place the combined problem (for $\tilde{T}$ and $\tilde{q}_1$) into a single function space setting. The automatic satisfaction of the global heat balance, (23), now follows immediately since $\sum_{i=1}^{N+M}\Phi_i(\mathbf{x}) = 1{\cdot}0$ in Ω, on $\partial\Omega_1$, and on $\partial\Omega_2$. There is now one unknown (T_j or q_{1j}) for *each* node in the system and, correspondingly, one equation associated with each node. It turns out, however, upon closer inspection of the individual equations of (24), that indeed the nodal equations for the primary variables are (necessarily) independent of and uncoupled from those for the derived variables (the converse of course is not true). Hence, a two-part solution procedure 'falls out' automatically: (i) solve the first N equations ($i = 1, 2, \ldots, N$) for the primary variables, i.e. (19), and (ii) using these results, solve the remaining M equations for the derived variables, i.e. (21).

Finally, we present (noting that there is no restriction to a Poisson problem) an algorithmic way in which to view, and perhaps implement, the consistent 'flux' method:

(i) Initially, form *all* of the boundary nodal equations as if there were to be imposed the most general type of natural boundary condition at each node (for the Laplace operator considered thus far, it could be $\mathbf{n}\cdot\nabla T + h(T - T_0) + q = 0$, for example).

(ii) Modify the boundary node equations for the particular problem at hand, e.g. for Dirichlet data, the nodal equation can be omitted entirely (after transposing the appropriate coupling information to the right-hand side), although it should also be 'saved' (e.g. on a disk file) for later use in step (iv). For simpler natural boundary conditions, the proper deletions are made (e.g. h, T_0 or q in the current problem).

(iii) Assemble and solve the conventional GFEM equations for the primary variables.

(iv) Recall the nodal equation for which Dirichlet data are employed, simplify the general boundary condition to that relating the primary and derived variables ($q = -\mathbf{n}\cdot\nabla T$ for the current problem) in each equation, and solve for the consistently derived variables.

The last step is, of course, generally optional; it is required only if the derived variables are. Also, it may sometimes be more convenient to perform this step by looping through the appropriate boundary elements and reconstructing the boundary nodal equations (rather than saving and later retrieving the original Dirichlet boundary node equations).

Consistent mass vs. lumped mass solution

As mentioned previously, the consistent flux method leads to the requirement of solving a linear system, represented in general by (17) for the derived quantities when the *consistent* (boundary) mass matrix is employed. It is possible, however, to avoid the use of the consistent mass matrix while still retaining the major benefits of the consistent flux method; consistent flux does not necessarily require consistent mass. First, however, we note that even if the consistent mass matrix is employed, the cost of the calculation is generally small compared to the cost of solving for the original (primary) variables because the dimensionality of the problem is lower (a two-dimensional problem for $\tilde{T}$ gives only a one-dimensional problem for $\tilde{q}_1$, etc.); concomitantly, M is significantly smaller than N in a practical calculation. The most expensive calculation of the derived variable is that from a Dirichlet problem; not only are there more 'boundary nodes' to consider, but the usually sparsely banded matrix structure of $\mathbf{M}$ is then (effectively) full, owing to the 'closed loop' of boundary elements. In this case, a skyline (or profile) solution method[13] would be particularly beneficial if Gaussian elimination is used.

The alternative to solving the linear system is (obviously) to invoke 'mass lumping', in which case the mass matrix [see (16)],

$$B_{ij} = \int_{\Gamma} \Gamma_i \Gamma_j, \tag{25}$$

is rendered diagonal via, for instance, the row sum technique (where applicable),

$$B_{ij} = \delta_{ij} \int_{\Gamma} \Gamma_i, \tag{26}$$

where δ_{ij} is the Kronecker delta (mass lumping is normally done at element level[14]). If (26) is employed, the nodal fluxes in (16) (for example) are uncoupled and are simply

$$q_{1i} = \left(\int_{\Omega} \Gamma_i S - \int_{\Omega} \nabla \Gamma_i \cdot \nabla \tilde{T} \right) \Big/ \int_{\Gamma} \Gamma_i, \quad i = 1, 2, \ldots, M. \tag{27}$$

Although such a procedure is somewhat *ad hoc*, and (perhaps) generally less accurate than that using consistent mass, it can significantly simplify the calculation of the consistently derived quantities. By comparing consistent and lumped mass, Thornton[5] has verified that the lumped mass approach is generally a viable alternative in that it is simpler, sometimes more accurate, and generally more cost effective; also, and importantly, the correct global balances are still obtained. Although the consistent mass approach should probably be employed if one desires the most accurate fluxes (and/or the appropriate 'wiggle signals' *à la* Gresho and Lee[15]) available from the approximation, the lumped mass results are nearly as simple to compute (and are often significantly more accurate) as those from the common methods based on 'basis function derivatives evaluated on Γ' and this alone is probably sufficient reason to advocate it.

Finally we point out that if mass lumping is used to compute $\tilde{q}$, it would also be required if the Dirichlet problem were to be resolved as a Neumann problem to verify that the same solution ($\tilde{T}$) is obtained.

Time-dependent advection diffusion

We now consider an important prototypical equation for problems involving incompressible fluid flow. The advection–diffusion equation is a useful stepping stone to the more complicated Navier–Stokes (and Boussinesq) equations which we ultimately consider; it involves fluid motion which, although presumed to be given, introduces sufficient additional complexities to merit separate consideration.

The problem under consideration is the following: find $T(\mathbf{x}, t)$ in $(\Omega \times [0, \Theta])$ such that

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T + \beta T \nabla \cdot \mathbf{u} = \kappa \nabla^2 T + S(\mathbf{x}, t) \quad \text{in} \quad \Omega, \tag{28a}$$

where $\nabla \cdot \mathbf{u} = 0$ in Ω,

$$T = T_0 \quad \text{on} \quad \partial\Omega_1, \tag{28b}$$

$$\kappa \mathbf{n} \cdot \nabla T + h(T - T_2) + q_2 = 0 \quad \text{on} \quad \partial\Omega_2 \tag{28c}$$

and

$$T(\mathbf{x}, 0) = g(\mathbf{x}); \tag{28d}$$

here $\mathbf{u}(\mathbf{x}, t)$, κ, S, T_0, h, T_2, q_2 and g are given and $\Gamma = \partial\Omega_1 \oplus \partial\Omega_2$ is the boundary of Ω. The parameter, β, although irrelevant if $\nabla \cdot \mathbf{u} = 0$, will be seen to be important in the approximate solution, since in the discretized approximation, wherein $\mathbf{u}$ is typically approximated via piecewise polynomials similar to those used for T, $\nabla \cdot \mathbf{u}$ can, at best, vanish only in a weak sense (not pointwise)—generally.

Introducing the same finite element function spaces and basis functions as in the previous problem, the appropriate weak form of (28) is

$$\int_\Omega \phi_i\left[\frac{\partial\tilde{T}}{\partial t} + (\mathbf{u}\cdot\nabla\tilde{T} + \beta\tilde{T}\nabla\cdot\mathbf{u})\right] + \int_\Omega \kappa\nabla\phi_i\cdot\nabla\tilde{T} + \int_{\partial\Omega_2} h\phi_i\tilde{T}$$

$$= \int_\Omega \phi_i S + \int_{\partial\Omega_2} \phi_i(hT_2 - q_2); \quad i = 1,2,\ldots,N, \tag{29}$$

where $\tilde{T}$ is given by (20). Inserting (20) into (29) and transposing the terms in $\{T_j^s\}$ to the right-hand side, leads to the conventional GFEM equations for $\{T_j\}$, this time in the form of a coupled linear set of first-order ordinary differential equations (ODEs) in time. Integration of these ODEs yields the approximate solution for the primary variable—'part 1' (conventional GFEM) of the solution procedure.

Once $\tilde{T}$ is available, the consistent heat flux on $\partial\Omega_1$ can be computed, as before, by considering the following weak form of (28):

$$\int_\Omega \Gamma_i\left[\frac{\partial\tilde{T}}{\partial t} + \mathbf{u}\cdot\nabla\tilde{T} + \beta\tilde{T}\nabla\cdot\mathbf{u}\right] + \int_\Omega \kappa\nabla\Gamma_i\cdot\nabla\tilde{T} + \int_{\partial\Omega_2} h\Gamma_i\tilde{T}$$

$$= \int_\Omega \Gamma_i S + \int_{\partial\Omega_2} \Gamma_i(hT_2 - q_2) - \int_{\partial\Omega_1} \Gamma_i\tilde{q}_1; \quad i = 1,2,\ldots,M, \tag{30}$$

which, when (22) is employed, is a system of Galerkin equations for the (time-dependent) nodal heat fluxes $\{q_{1i}\}$; this is 'Part 2'. It is noteworthy that $\tilde{q}_1$ depends on *much* more data that simply the normal derivative of $\tilde{T}$ on Γ, which is (effectively) a portion of the term $\int_\Omega \kappa\nabla\Gamma_i\cdot\nabla\tilde{T}$; we will elucidate this point later.

To obtain the correct global heat balance implied by (28) we begin, as before, by summing the $N + M$ equations of (29) and (30) to obtain, using (12),

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{T} + \int_\Omega (\mathbf{u}\cdot\nabla\tilde{T} + \beta\tilde{T}\nabla\cdot\mathbf{u}) = \int_\Omega S - \int_{\partial\Omega_1} \tilde{q}_1 - \int_{\partial\Omega_2} [q_2 + h(\tilde{T} - T_2)], \tag{31}$$

which is almost, but not quite, the correct global balance. The 'culprit' is the advection term which, using the divergence theorem, can be rewritten as

$$\int_\Omega \mathbf{u}\cdot\nabla\tilde{T} = \int_\Gamma \tilde{T}\mathbf{n}\cdot\mathbf{u} - \int_\Omega \tilde{T}\nabla\cdot\mathbf{u}$$

to yield

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{T} = \int_\Omega S - \int_{\partial\Omega_1} \tilde{q}_1 - \int_{\partial\Omega_2} [q_2 + h(\tilde{T} - T_2)] - \int_\Gamma \tilde{T}\mathbf{n}\cdot\mathbf{u} + (1 - \beta)\int_\Omega \tilde{T}\nabla\cdot\mathbf{u}. \tag{32}$$

Except for the last term, each member of the right-hand side is an appropriate contribution to the global energy balance: the first term is the total heat generation rate, the second term is the net flow of heat leaving by conduction through $\partial\Omega_1$, the third term is the net flow of heat leaving by conduction and convection (i.e. Newton's law of cooling) through $\partial\Omega_2$ and the fourth term is the net loss of heat by advection through Γ (this term is zero for a contained flow since then $\mathbf{n}\cdot\mathbf{u}|_\Gamma = 0$; in the more general flow-through domain, it is (appropriately) non-zero, even though $\int_\Gamma \mathbf{u}\cdot\mathbf{n}$ is always zero for incompressible flows).

In most discretized approximations, the term $\int_\Omega \tilde{T}\nabla\cdot\mathbf{u}$ will not be zero, as would be the case in the continuum, although presumably it is usually small. If $\beta = 0$ we are dealing with the so called advective form of the equation[16] and the last term in (32) represents a spurious source (or sink, since $\int_\Omega \tilde{T}\nabla\cdot\mathbf{u}$ is indefinite) which *precludes* the proper global energy balance. On the other hand, the energy balance can be recovered merely by setting $\beta = 1$; this is called the flux divergence form since $\mathbf{u}\cdot\nabla\tilde{T}+\tilde{T}\nabla\cdot\mathbf{u}=\nabla\cdot(\mathbf{u}\tilde{T})$ and $\mathbf{u}\tilde{T}$ is the advective flux of internal energy. Thus, only if β is set to 1 in (29) and (30) will the resulting values of $\{\tilde{q}_{1i}\}$ be consistent in the sense of satisfying the appropriate global energy balance. *Another requirement for consistency then is that the flux divergence form be used for the advection terms* (unless $\nabla\cdot\mathbf{u}\equiv 0$, in which case β is irrelevant). Finally we remark that (i) steady-state advection–diffusion is of course just a special case of the above and is obtained simply by omitting all time derivatives, (ii) if T in (28a) represents a species mass fraction rather than temperature, only the consistent flux method can guarantee species mass conservation.

Remark. An alternative formulation (for $\beta=1$), suggested by a referee, is one that includes the advective flux in the boundary condition (28c), i.e. $\kappa\mathbf{n}\cdot\nabla T$ is replaced by $\mathbf{n}\cdot(\kappa\nabla T-\mathbf{u}T)$. Then, via integration by parts of (also) the advection term in (28a), an appropriate weak form that incorporates the new boundary condition is obtained. The resulting global heat balance (32) has $\tilde{q}_1$ replaced by $(\tilde{q}_1+\mathbf{n}\cdot\mathbf{u}\tilde{T})$ and the global advection term (the integral over Γ) is omitted; the third term then represents the heat loss through $\partial\Omega_2$ via advection, diffusion and 'convection'.

Navier–Stokes (Boussinesq) equations

Having detailed the consistent GFEM procedures for advection-diffusion, it is now a relatively simple matter to extend these results to the incompressible Navier–Stokes equations, written in two-dimensional Cartesian co-ordinates for simplicity: find $\mathbf{u}(\mathbf{x},t)=(u,v)$, $P(\mathbf{x},t)$ and $T(\mathbf{x},t)$ such that

$$\rho\left[\frac{\partial u}{\partial t}+\mathbf{u}\cdot\nabla u+\beta u\nabla\cdot\mathbf{u}\right]=\nabla\cdot\boldsymbol{\tau}_x \quad \text{in} \quad \Omega, \tag{33a}$$

$$\rho\left[\frac{\partial v}{\partial t}+\mathbf{u}\cdot\nabla v+\beta v\nabla\cdot\mathbf{u}\right]=\nabla\cdot\boldsymbol{\tau}_y+\rho\gamma g T \quad \text{in} \quad \Omega, \tag{33b}$$

$$\nabla\cdot\mathbf{u}=0 \quad \text{in} \quad \Omega, \tag{33c}$$

where

$$\boldsymbol{\tau}_x=\mathbf{i}\left(2\mu\frac{\partial u}{\partial x}-P\right)+\mathbf{j}\mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right), \tag{33d}$$

$$\boldsymbol{\tau}_y=\mathbf{i}\mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)+\mathbf{j}\left(2\mu\frac{\partial v}{\partial y}-P\right), \tag{33e}$$

P is the pressure, μ is the viscosity, ρ is the density, γ is the coefficient of volumetric expansion, T is the temperature deviation from a reference value and g is gravity. Equations (33), along with (28) for thermally coupled flows and the following initial and boundary conditions on velocity, are sufficient to yield the primary variables, $\mathbf{u}$, P and T:

$$\mathbf{u}(\mathbf{x},0)=\mathbf{u}_0(\mathbf{x}) \quad \text{where} \quad \nabla\cdot\mathbf{u}_0=0, \tag{34a}$$

$$\mathbf{u}=\mathbf{u}_1 \quad \text{on} \quad \partial\Omega_1^u, \tag{34b}$$

$$\mathbf{n}\cdot\boldsymbol{\tau}_x=n_x\left(2\mu\frac{\partial u}{\partial x}-P\right)+n_y\mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)=f_{x_2} \quad \text{on} \quad \partial\Omega_2^u \tag{34c}$$

$$v=v_1 \quad \text{on} \quad \partial\Omega_1^v \tag{34d}$$

and

$$\mathbf{n}\cdot\boldsymbol{\tau}_y = n_x\mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right)+n_y\left(2\mu\frac{\partial v}{\partial y}-P\right)=f_{y_2} \quad \text{on} \quad \partial\Omega_2^v, \tag{34e}$$

where the x- and y-components of the traction force (exerted by the boundary on the fluid), f_{x_2} and f_{y_2}, are given, as are u_1, v_1 and $\mathbf{u}_0$. Also, the boundary of Ω is $\Gamma = \partial\Omega_1^u \oplus \partial\Omega_2^u = \partial\Omega_1^v \oplus \partial\Omega_2^v = \partial\Omega_1^T \oplus \partial\Omega_2^T$ (the last term is related to (28b, c)). If desired, the (natural) traction boundary conditions, given by (34c, e) could be generalized *à la* (28c) to boundary conditions of the third kind, sometimes called Robin conditions (which could, for example, permit slip along a 'solid' wall) in the form (for 34c)

$$\mathbf{n}\cdot\boldsymbol{\tau}_x = f_{x_2}+\alpha(u-u_2) \quad \text{on} \quad \partial\Omega_2^u, \tag{35}$$

where α and u_2 as well as f_{x_2}, are given.[17]

Proceeding as before, we first write the weak form appropriate to the primary variables, which is

$$\int_\Omega \rho\phi_i^u\left[\frac{\partial\tilde{u}}{\partial t}+\tilde{\mathbf{u}}\cdot\nabla\tilde{u}+\beta\tilde{u}\nabla\cdot\tilde{\mathbf{u}}\right]=\int_{\partial\Omega_2^u}\phi_i^u f_{x_2}-\int_\Omega\tilde{\boldsymbol{\tau}}_x\cdot\nabla\phi_i^u; \quad i=1,2,\ldots,N_u, \tag{36a}$$

$$\int_\Omega \rho\phi_i^v\left[\frac{\partial\tilde{v}}{\partial t}+\tilde{\mathbf{u}}\cdot\nabla\tilde{v}+\beta\tilde{v}\nabla\cdot\tilde{\mathbf{u}}\right]=\int_{\partial\Omega_2^v}\phi_i^v f_{y_2}-\int_\Omega\tilde{\boldsymbol{\tau}}_y\cdot\nabla\phi_i^v+\int_\Omega\rho\gamma g\phi_i^v\tilde{T}; \quad i=1,2,\ldots,N_v, \tag{36b}$$

$$\int_\Omega \psi_i\nabla\cdot\tilde{\mathbf{u}}=0; \quad i=1,2,\ldots,N_p \tag{36c}$$

and (29) in which $\mathbf{u}$ is replaced by $\tilde{\mathbf{u}}$, N is replaced by N_T and ϕ_i by ϕ_i^T (which vanishes on $\partial\Omega_1^T$); here

$$\tilde{u}=u_{\Omega,\partial\Omega_2^u}+u_{\mathrm{s}}=\sum_{j=1}^{N_u}u_j\phi_j^u+\sum_{j=1}^{M_u}u_j^{\mathrm{s}}\Gamma_j^u, \tag{37a}$$

where $\{\phi_j^u\}$ vanish on $\partial\Omega_1^u$,

$$\tilde{v}=v_{\Omega,\partial\Omega_2^v}+v_{\mathrm{s}}=\sum_{j=1}^{N_v}v_j\phi_j^v+\sum_{j=1}^{M_v}v_j^{\mathrm{s}}\Gamma_j^v, \tag{37b}$$

where $\{\phi_j^v\}$ vanish on $\partial\Omega_1^v$,

$$\tilde{P}=\sum_{j=1}^{N_p}P_j\psi_j \tag{37c}$$

and (20) in which $\{\phi_j,\Gamma_j\}$ are replaced by $\{\phi_j^T,\Gamma_j^T\}$, N is replaced by N_T, M by M_T and $\partial\Omega_2$ by $\partial\Omega_2^T$. Here u_{s} is the interpolant of u_1 on $\partial\Omega_1^u$ and similarly for v_{s} on $\partial\Omega_1^v$ and T_{s} on $\partial\Omega_1^T$, $\{\psi_i\}$ are the basis functions for pressure (from the finite-dimensional (N_p) subspace of L_2, which permits discontinuous, C^{-1}, approximation) and

$$\tilde{\boldsymbol{\tau}}_x\cdot\nabla\phi_i^u=\left(2\mu\frac{\partial\tilde{u}}{\partial x}-\tilde{P}\right)\frac{\partial\phi_i^u}{\partial x}+\mu\left(\frac{\partial\tilde{u}}{\partial y}+\frac{\partial\tilde{v}}{\partial x}\right)\frac{\partial\phi_i^u}{\partial y}, \tag{37d}$$

$$\tilde{\boldsymbol{\tau}}_y\cdot\nabla\phi_i^v=\mu\left(\frac{\partial\tilde{u}}{\partial y}+\frac{\partial\tilde{v}}{\partial x}\right)\frac{\partial\phi_i^v}{\partial x}+\left(2\mu\frac{\partial\tilde{v}}{\partial y}-\tilde{P}\right)\frac{\partial\phi_i^v}{\partial y}. \tag{37e}$$

There is a total of $N+M$ nodes for velocity and temperature and N_p nodes for pressure. Also, N_u comprises the 'u-nodes' in Ω and on $\partial\Omega_2^u$ and similarly for N_v and N_T, M_u comprises the u-nodes on $\partial\Omega_1^u$ and similarly for M_v and N_T; finally,

$$N_u + M_u = N_v + M_v = N_T + M_T = N + M. \tag{38}$$

Inserting (37) into (36) and (20) into (29) leads to the conventional GFEM equations for $\{u_j\}$, $\{v_j\}$, $\{P_j\}$ and $\{T_j\}$; considerable effort is of course required to obtain these primary variables since the GFEM equations are coupled, non-linear, first-order differential equations in time (for effective solution procedures, see Reference 18).

Nevertheless, in principle, the 'first part' of the problem is now solved (for $\tilde{\mathbf{u}}$, $\tilde{P}$ and $\tilde{T}$) and we move on to 'Part 2', the appropriate weak form of which is

$$\int_\Omega \rho\Gamma_i^u\left[\frac{\partial\tilde{u}}{\partial t} + \tilde{\mathbf{u}}\cdot\nabla\tilde{u} + \beta\tilde{u}\nabla\cdot\tilde{\mathbf{u}}\right] = \int_{\partial\Omega_1^u}\Gamma_i^u\tilde{f}_{x_1} + \int_{\partial\Omega_2^u}\Gamma_i^u f_{x_2} - \int_\Omega \tilde{\boldsymbol{\tau}}_x\cdot\nabla\Gamma_i^u;.\quad i = 1, 2, \ldots, M_u, \tag{39a}$$

$$\int_\Omega \rho\Gamma_i^v\left[\frac{\partial\tilde{v}}{\partial t} + \tilde{\mathbf{u}}\cdot\nabla\tilde{v} + \beta\tilde{v}\nabla\cdot\tilde{\mathbf{u}}\right] = \int_{\partial\Omega_1^v}\Gamma_i^v\tilde{f}_{y_1} + \int_{\partial\Omega_2^v}\Gamma_i^v f_{y_2} - \int_\Omega \tilde{\boldsymbol{\tau}}_y\cdot\nabla\Gamma_i^v + \int_\Omega \rho\gamma g\Gamma_i^v\tilde{T};$$
$$i = 1, 2, \ldots, M_v, \tag{39b}$$

and (30) in which $\mathbf{u}$ is replaced by $\tilde{\mathbf{u}}$, Γ_i is replaced by Γ_i^T and M by M_T. The only unknowns in these equations are $\tilde{f}_{x_1}$, $\tilde{f}_{y_1}$ and $\tilde{q}_1$; thus, using the appropriate expansions

$$\tilde{f}_{x_1} = \sum_{j=1}^{M_u} f_{x_{1j}}\Gamma_j^u|_{\partial\Omega_1^u}, \tag{40a}$$

$$\tilde{f}_{y_1} = \sum_{j=1}^{M_v} f_{y_{1j}}\Gamma_j^v|_{\partial\Omega_1^v} \tag{40b}$$

and (22) with Γ_j replaced by Γ_j^T, $\partial\Omega_1$ by $\partial\Omega_1^T$ and M by M_T, (39) and (30) can be used to obtain consistent (nodal) forces in the x- and y-directions and the consistent (nodal) heat flux on the respective Dirichlet portions of the boundary.

Although considerable effort would be required, it could be shown that (39) converges, as $h \rightarrow 0$, to the appropriate continuum boundary forces (or surface tractions); namely the $\partial\Omega_1$ analogues of (34c) and (34e). For finite h, however, the above equations provide a more accurate 'local force balance'. Finally, as is often the case, if normal and tangential components of the boundary force are desired, then generally one must employ the techniques discussed by Engelman *et al.*,[19] in which the momentum equations are rotated through the appropriate angle to achieve consistent results, i.e. in contrast to the situation for thermal problems, which are scalar, it is necessary and important, for general polygonal boundaries, to carefully define a *consistent* normal direction for incompressible flow problems. This situation is particularly relevant and evident for the 4/1 element (bilinear velocity, piecewise constant pressure) when a domain corner is 'defined' by a single element—a common occurrence (for *any* element). Since the pressure is constant over the element, it is clear that this element is incapable of representing pressure gradients within a single element. Whereas both f_x and f_y from (39) and (40) will contain P in the corner element, it is clear that the resolution of the boundary force vector at the corner node into normal and tangential components must be done 'properly'; the definition of proper requires that P shows up (as usual) in the normal force, but *not* in the tangential force. Not surprisingly, this unique normal *direction* is identical to that obtained in the 'consistent normal' described in Reference 19.

It is also noteworthy that consistent boundary *forces* can only be obtained using the (given) stress-divergence form of the Navier–Stokes equations. If, for example, $\nabla\cdot\boldsymbol{\tau}_x$ in (33a) were replaced [using (33c)] by its continuum equivalent, $\mu\nabla^2 u - \partial P/\partial x$, the associated natural boundary conditions no longer represent physical forces and it is then not possible to compute consistent boundary forces. (The computed results would be only portions of the total forces).

This of course is not to say that the alternative form of the Navier–Stokes equations is actually wrong or 'illegal'; in fact it is often a useful formulation method, especially for outflow boundaries.[20,21] It does indicate, however, that the stress-divergence form is more physically "consistent".

It is to be emphasized that the forces obtained from (39) and (40) are in fact the components of the total force vector acting in the x- and y-co-ordinate directions. If the normal and tangential components of this force are desired, they can be obtained from f_x and f_y using the appropriate transformation (rotation) *after* the consistent normal and tangential directions are obtained in the manner described by Engelman *et al.*[19] Note that the normal vector is not required if only the Cartesian components of the force are desired.

Finally, to demonstrate the global balances implied by (33) and implicitly contained in the consistently obtained results, we first add all of the equations of (36a) to those of (39a) and similarly for (36b) and (39b) to obtain, using the appropriately generalized version of (12),

$$\rho\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{u} = \int_{\partial\Omega_1^u}\tilde{f}_{x_1} + \int_{\partial\Omega_2^u} f_{x_2} - \rho\int_\Omega(\tilde{\mathbf{u}}\cdot\nabla\tilde{u} + \beta\tilde{u}\nabla\cdot\tilde{\mathbf{u}}) \tag{41a}$$

and

$$\rho\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{v} = \int_{\partial\Omega_1^v}\tilde{f}_{y_1} + \int_{\partial\Omega_2^v} f_{y_2} + \rho\gamma g\int_\Omega \tilde{T} - \rho\int_\Omega(\tilde{\mathbf{u}}\cdot\nabla\tilde{v} + \beta\tilde{v}\nabla\cdot\tilde{\mathbf{u}}), \tag{41b}$$

which can also be written as

$$\rho\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{u} = \int_\Gamma(\tilde{f}_x - \tilde{u}\mathbf{n}\cdot\tilde{\mathbf{u}}) + \rho(1-\beta)\int_\Omega \tilde{u}\nabla\cdot\tilde{\mathbf{u}}, \tag{42a}$$

$$\rho\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \tilde{v} = \int_\Gamma(\tilde{f}_y - \tilde{v}\mathbf{n}\cdot\tilde{\mathbf{u}}) + \rho\gamma g\int_\Omega \tilde{T} + \rho(1-\beta)\int \tilde{v}\nabla\cdot\tilde{\mathbf{u}}. \tag{42b}$$

If $\nabla\cdot\tilde{\mathbf{u}} = 0$ or $\beta = 1$ (only the latter of which is generally feasible in the discretized version), these are the appropriate global force balances in the x- and y-directions.[22] The first term $(\tilde{f}_x, \tilde{f}_y)$ on the right-hand side is the total force exerted by the boundary on the fluid, the second (in the first integral) is the next flux of momentum leaving Ω through Γ and the third term in (42b) is the total upward buoyancy force (which is always largely balanced by a hydrostatic portion of the pressure—a part of $\int_\Omega \tilde{f}_y$). The last term is of course spurious and β must (usually) be set to 1 if true consistency is desired.

As with advection–diffusion, the consistent steady-state equations are obtained from the above simply by omitting all time derivatives.

EXAMPLES

Further clarification of the consistent flux method may be best presented using sample problems.

One-dimensional steady heat conduction

As an extremely simple but illustrative introductory example, consider the one-dimensional version of (1) and/or (18):

$$\frac{\mathrm{d}^2 T}{\mathrm{d}x^2} + S(x) = 0; \quad 0 \leqslant x \leqslant 3, \tag{43a}$$

$$T = 0 \quad \text{at} \quad x = 0 \tag{43b}$$

and either

$$T=0 \quad \text{at} \quad x=3 \tag{43c}$$

or

$$-\frac{\mathrm{d}T}{\mathrm{d}x}=5 \quad \text{at} \quad x=3, \tag{43d}$$

where $S(x)=0$ for $0<x<2$ and $S(x)=6$ for $2<x<3$. The exact solution to (43) is

$$T=x, \quad 0\leqslant x\leqslant 2, \tag{44a}$$

$$T=x(13-3x)-12, \quad 2\leqslant x\leqslant 3, \tag{44b}$$

which is shown as the lower solid curve in Figure 1.

We will now obtain the approximate solution using just three linear elements, each of unit length. Consider first the Dirichlet boundary condition (43c), for which the two nodal equations (for T_1 and T_2, the interior nodes; see Figure 1) are, from (7),

$$\frac{1}{l}(2T_1-T_2)=0 \tag{45a}$$

and

$$\frac{1}{l}(-T_1+2T_2)=Sl/2. \tag{45b}$$

Hence $T_1=1$, $T_2=2$; the solution is exact at the nodes (the dashed line in the lower curve of Figure 1).

Suppose now that the heat flux, $q=-\mathrm{d}T/\mathrm{d}x$, at $x=3$ is desired. If the conventional method is imployed, *a la* (14), we obtain

$$q=-(T_3-T_2)/l=2, \tag{46a}$$

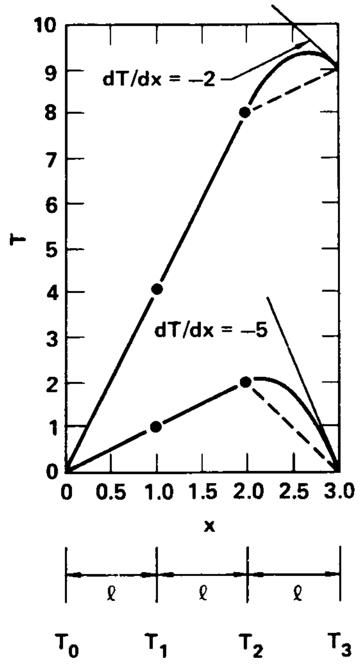

Figure 1. Steady heat conduction in one dimension

whereas the consistent flux, from (10) or (16), is

$$q = Sl/2 - (T_3 - T_2)/l = 5, \tag{46b}$$

which agrees with the exact solution. It is clear that although (46a) is in fact the true slope of the FEM solution, (46b) properly accounts for both the heat generated in the finite (i.e. not infinitesimal) element and heat conduction at $x = 3$.

To further demonstrate consistency, let us re-solve the problem using the Neumann boundary condition, (43d), except that we use the *derived* value for q—from (46). The appropriate nodal equations for this case, from (19) and (20), are

$$\frac{1}{l}(2T_1 - T_2) = 0, \tag{47a}$$

$$\frac{1}{l}(-T_1 + 2T_2 - T_3) = Sl/2, \tag{47b}$$

$$\frac{1}{l}(T_3 - T_2) = Sl/2 - q. \tag{47c}$$

(Note the 'similarity' between (46b) and (47c)). The solution to (47) for $q = 2$(46a) is $T_1 = 4$, $T_2 = 8$, $T_3 = 9$ and is shown as the (dashed) upper curve in Figure 1 (the solid line is the exact solution for $q = 2$). On the other hand, $q = 5$ from (46b) gives the original result, $T_1 = 1$, $T_2 = 2, T_3 = 0$, which is also the solution using (43d)—i.e. the exact solution. Hence, only the consistently derived flux can be used as a natural boundary condition to recover the original solution (obtained with Dirichlet boundary conditions). Finally, it is simple to demonstrate that only (46b) gives a heat flux which yields a global heat balance; see (11).

One-dimensional time-dependent heat conduction

For this example, we consider the one-dimensional version of (28) with u and $S = 0$, in order to demonstrate the effect of time-dependent boundary conditions on the computation of the derived variable; namely

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2}; \quad 0 \leqslant x \leqslant L, \tag{48a}$$

$$T = T_i(x) \quad \text{at} \quad t = 0, \tag{48b}$$

$$T = T_0 \quad \text{constant at} \quad x = 0 \tag{48c}$$

and

$$T = T_L(t) \quad \text{at} \quad x = L. \tag{48d}$$

Here we assume that the approximate solution, $\tilde{T}(x, t)$, has been computed (via (29), etc.), and we desire the heat flux at $x = L$. Denoting by l the length of the last (linear) element, the conventional flux is simply

$$q = -\kappa(T_L(t) - T_N(t))/l \tag{49a}$$

where $T_N(t)$ is the computed nodal temperature at the first node in from $x = L$. The consistent flux result, via (30), is

$$q = -\kappa\{[T_L(t) - T_N(t)]/l\} - (l/6)[2\dot{T}_L(t) + \dot{T}_N(t)], \tag{49b}$$

which, again via a proper energy balance, accounts for more than just the slope of $T(x)$. For example, suppose that $T_L(t)$ is given by

$$T_L(t) = T_s(1 - e^{-t/\tau}), \tag{50}$$

where τ is a given time-constant. In this case, (49b) gives

$$q = -\kappa[(T_L - T_N)/l] - (l/6)[(2T_s/\tau)e^{-t/\tau} + \dot{T}_N(t)], \tag{51}$$

where T_N and $\dot{T}_N(t)$ are available from the solution to 'part 1'. Clearly the consistent incorporation of a specified time-dependent boundary temperature would be significant for $t = O(\tau_\varepsilon)$ where $\tau_\varepsilon \equiv l^2/\kappa \ll \tau$ is the element time constant. (Also, *á la* Gresho and Lee,[15] one should not even *seek* a heat flux result for $t \ll \tau_\varepsilon$.)

Finally we note that if the lumped mass approximation were employed (for $\partial T/\partial t$) to compute the nodal temperatures, then the associated consistent flux equation is

$$q = -\kappa[(T_L - T_N)/l] - l\dot{T}_L/2 \tag{49c}$$

rather than (49b), i.e. the mass must also be lumped when computing q. Equation (49c) is also the 'consistent' flux which would be obtained using the finite difference method and introducing an 'image point' outside of $x = L$ (at a distance l), thus suggesting that these techniques can also be adopted for finite difference approximations.

Steady two-dimensional heat conduction

We now consider the problem given by (1) where $S = -10$ and Ω is the unit square. If $T_0(\Gamma)$ is specified properly, it is easy to demonstrate that an exact solution is

$$T(x, y) = (2x + y)^2, \tag{52}$$

which is also the FEM solution at the nodes, using, for example, the 4-node bilinear element and a mesh of rectangles. The corresponding heat flux is

$$\mathbf{q} = -\nabla T = -2(2x + y)(2\mathbf{i} + \mathbf{j}), \tag{53}$$

where $\mathbf{i}$, $\mathbf{j}$ are the unit vectors in the x- and y-directions, respectively. The boundary flux is $q = \mathbf{n}\cdot\mathbf{q}|_\Gamma$ and is shown in Figure 2 as the heavy solid lines. By comparison, two approximate heat flux results are also shown which were obtained on the graded mesh of 64 elements shown in the Figure. The dots represent the flux obtained using (27), i.e. the consistent heat flux in the lumped mass approximation. The triangles represent the flux as computed from (14) evaluated at the centre of the boundary of each element (boundary Gauss point).

Clearly the consistent flux (even using lumped mass) is significantly more accurate. Also, whereas the consistent flux identically satisfies (11), the results using (14) give $\int_\Gamma q = -8{\cdot}75$, which is a 12·5 per cent error in the global heat balance. It is also noteworthy that the consistent method yields a (continuous) value of q at each of the four corners, where the actual normal is undefined and the exact solution displays a jump in the normal flux. This apparent deficiency is of course related to the C^0 (smoothing) approximation employed for the boundary flux. Finally, only if the consistent flux results were used as prescribed fluxes for the equivalent Neumann problem, would the exact solution at the nodes be again obtained.

Two-dimensional time-dependent advection–diffusion

Applying the consistent flux equation (30) to the boundary two-patch of four-node elements

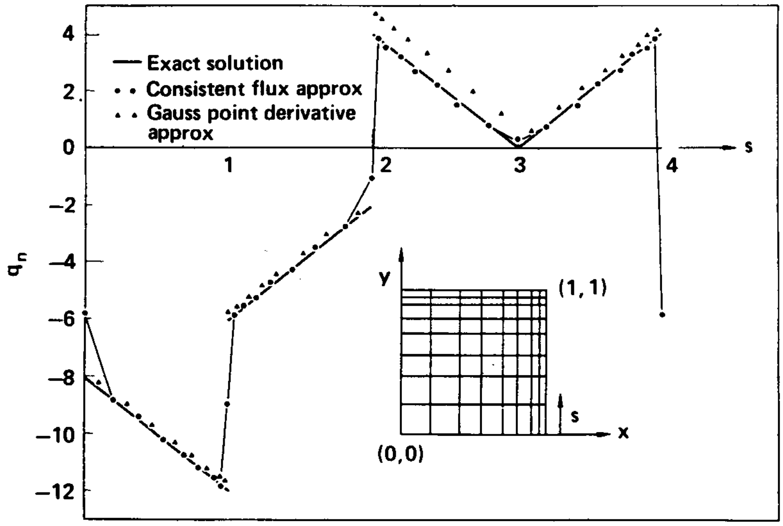

Figure 2. Boundary heat flux for steady two-dimensional heat conduction

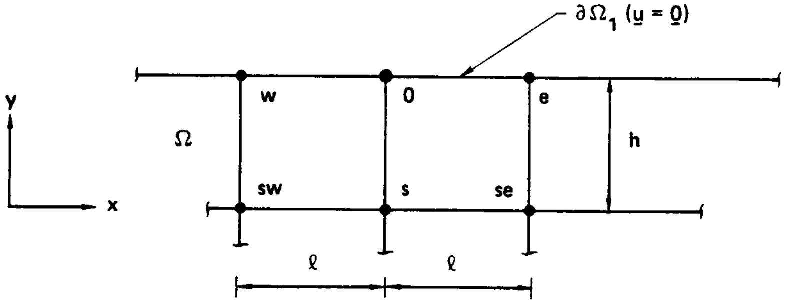

Figure 3. A two-patch of elements on the boundary

shown in Figure 3, yields [for $\beta = 1$, $S = 0$, using (22), and lumping the mass *à la* (26)],

$$
\begin{aligned}
-q_0 = {} & \frac{h}{36}(\dot{T}_{\mathrm{sw}} + 2\dot{T}_{\mathrm{w}} + 4\dot{T}_{\mathrm{s}} + 8\dot{T}_0 + 2\dot{T}_{\mathrm{e}} + \dot{T}_{\mathrm{se}}) \\
& + \frac{h}{72l}[U_{\mathrm{se}}(2T_{\mathrm{se}} + 2T_{\mathrm{e}} + T_{\mathrm{s}} + T_0) + U_{\mathrm{s}}(T_{\mathrm{se}} - T_{\mathrm{sw}} + T_{\mathrm{e}} - T_{\mathrm{w}}) \\
& - U_{\mathrm{sw}}(2T_{\mathrm{sw}} + 2T_{\mathrm{w}} + T_{\mathrm{s}} + T_0)] - \tfrac{1}{72}[V_{\mathrm{sw}}(2T_{\mathrm{sw}} + 2T_{\mathrm{s}} + T_{\mathrm{w}} + T_0) \\
& + V_{\mathrm{s}}(2T_{\mathrm{sw}} + 2T_{\mathrm{se}} + 12T_{\mathrm{s}} + 6T_0 + T_{\mathrm{e}} + T_{\mathrm{w}}) + V_{\mathrm{se}}(2T_{\mathrm{s}} + 2T_{\mathrm{se}} + T_0 + T_{\mathrm{e}})] \\
& + \frac{\kappa h}{6l^2}(4T_0 + 2T_{\mathrm{s}} - 2T_{\mathrm{e}} - 2T_{\mathrm{w}} - T_{\mathrm{se}} - T_{\mathrm{sw}}) \\
& + \frac{\kappa}{6h}[(T_{\mathrm{w}} - T_{\mathrm{sw}}) + 4(T_0 - T_{\mathrm{s}}) + (T_{\mathrm{e}} - T_{\mathrm{se}})], \qquad (54)
\end{aligned}
$$

where we have divided the original equation (30) by $\int_{\partial\Omega_1} \Gamma_0 = l$ in order to display the individual terms as finite differences. The above equation has the following Taylor series interpretation:

$$-q_0=\frac{h}{2}\left[\frac{\partial T}{\partial t}+\frac{\partial}{\partial x}(uT)+\frac{\partial}{\partial y}(vT)-\kappa\frac{\partial^2 T}{\partial x^2}\right]_0+\kappa\left.\frac{\partial T}{\partial y}\right|_0+O(h)[1+O(l^2)], \tag{55}$$

wherein the respective term-by-term identifications are obvious. By contrast, the corresponding result obtained via differentiation of the element-level basis functions and averaging the results is

$$-q_0=\frac{\kappa}{4h}[(T_w-T_{sw})+2(T_0-T_s)+(T_e-T_{se})]=\kappa\left.\frac{\partial T}{\partial y}\right|_0+O(h). \tag{56}$$

Thus, formally, the consistent flux is no more accurate than the conventional flux approximations. In practice, however, for finite h and l, it appears that (54) will yield more accurate results than (56) (which does not even recognize the effect of a change in l) by virtue of a reasonable (i.e. consistent) accounting of *all* the physical processes occurring in the *neighbourhood* of the point in question.

For comparison, the 'semi-consistent' flux equation, using the advection form ($\beta=0$) yields, for the advection terms,

$$\frac{h}{72l}[U_{sw}(T_s-T_{sw}+T_0-T_w)+2U_s(T_{se}-T_{sw}+T_e-T_w)+U_{se}(T_{se}-T_s+T_e-T_0)]$$
$$+\tfrac{1}{72}[(V_{sw}+V_s)(T_w-T_{sw})+(V_{sw}+6V_s+V_{se})(T_0-T_s)+(V_s+V_{se})(T_e-T_{se})],$$

which replaces the second and third terms in (54) and has the Taylor series interpretation,

$$\frac{h}{2}\left[u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial y}\right]+O(h)[1+O(l^2)].$$

We re-emphasize, however, that this form cannot generate a global energy balance.

Heat flux discontinuity

The following related example is presented to (i) show how a heat flux discontinuity is treated consistently and (ii) to further clarify the behaviour of the basis functions $\{\phi_i\}$ and $\{\Gamma_i\}$ on the boundary. We consider the advection–diffusion problem described by (28) with $h=S=0$ for simplicity and examine the appropriate two-patch (four-node element again) in the region near the intersection of $\partial\Omega_1$ and $\partial\Omega_2$, as shown in Figure 4. Here T is specified on node 0, w,...

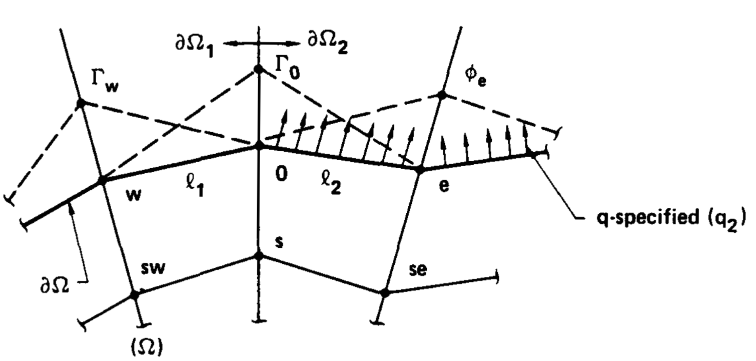

Figure 4. Discontinuity in heat flux caused by a change in boundary conditions

(i.e. on $\partial\Omega_1$) and the normal heat flux q_2 is specified at and to the right of node 0 (on $\partial\Omega_2$). The dashed lines indicate the appropriate basis functions on Γ, i.e. $\{\phi_i\}$ is zero on $\partial\Omega_1$, $\{\Gamma_i\}$ is zero over most of $\partial\Omega_1$ (and Ω) except that Γ_0 is non-zero in the neighbourhood of node 0 (the two corresponding elements), and $\sum_j \phi_j + \sum_j \Gamma_j = 1$ on Γ (and in Ω).

In order to compute the jump in heat flux at node 0 caused by the discontinuity in the boundary condition there, the consistent nodal equation is formed, as before, which yields (upon division by $\int_\Omega \Gamma_0 = \frac{1}{2}(l_1 + l_2)$)

$$\frac{2}{l_1 + l_2}\left[\int_{\partial\Omega_1} \Gamma_0 q_1 + \int_{\partial\Omega_2} \Gamma_0 q_2\right] = \text{RHS}, \tag{57}$$

where 'RHS' is the appropriate set of contributions from (30) and will resemble (54). For convenience and ease of exposition, we will first interpolate the given flux (q_2) into the basis set $\{\Gamma_j\}$ to give, using (22),

$$\frac{1}{6}\left(\frac{2}{l_1 + l_2}\right)[l_1 q_{\mathrm{w}}^{(1)} + 2l_1 q_0^{(1)} + 2l_2 q_0^{(2)} + l_2 q_{\mathrm{e}}^{(2)}] = \text{RHS}, \tag{58}$$

where $q_0^{(2)}$ is the given (specified) flux at node 0 and $q_0^{(1)}$ is the corresponding flux which is to be computed; $q_0^{(1)} + q_0^{(2)}$ is the total flux at node 0 and $q_0^{(1)} - q_0^{(2)}$ is the jump in flux. For further simplicity we now lump the boundary mass matrices *à la* (26) and (27), to give

$$\frac{l_1 q_0^{(1)} + l_2 q_0^{(2)}}{l_1 + l_2} = \text{RHS}, \tag{59}$$

which can be used to compute both $q_0^{(1)}$ and the flux discontinuity

$$q_0^{(1)} - q_0^{(2)} = (1 + l_2/l_1)(\text{RHS} - q_0^{(2)}). \tag{60}$$

Of course the conventional method may also be used to estimate the jump in flux, and the final equation looks much the same as (60), the main difference being the manner in which RHS is evaluated (e.g. from (54) *vis-à-vis* (56)).

Viscous flow

This final example illustrates the computation of boundary forces exerted by a viscous fluid in two-dimensional Stokes flow about a cylinder situated close to a moving wall. The velocity field from an exact solution for a semi-infinite domain[23,24] is used to generate the velocities prescribed on the boundary of the computational domain. (See Reference 25 for correction of an error in Reference 23). The latter allows a comparison between the exact and numerical solutions, both the primitive variables and the boundary stresses—obtained via the consistent method at boundary nodes or at the boundary Gauss points via derivation from primitive variables. A sequence of four mesh refinement experiments (363, 833, 1083, 1365 nodes) was conducted in order to assess the rate of convergence on isoparametric meshes of elements which differ mainly in size; both bilinear velocity, piecewise constant pressure (hereafter referred to as 4/1) and biquadratic velocity and C^{-1} locally linear pressure elements were investigated.[26] The 1083 node mesh of bilinear elements is displayed in Figure 5.

The velocity field and streamlines shown in Figures 6 and 7 were computed using the 4/1 element. Almost identical results were obtained with the 9/3 element, i.e. they were graphically indistinguishable. (Both mixed and penalty formulations agreed to order of the penalty parameter, as theory predicts.) Defining the average element size as the square root of the total area of the

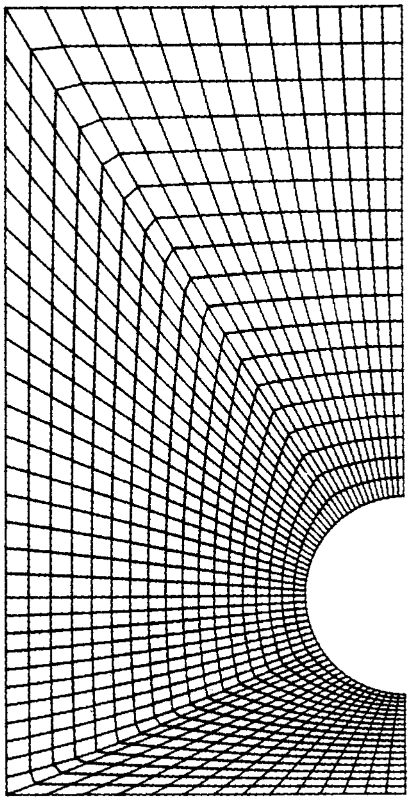

Figure 5. Mesh of 4/1 elements

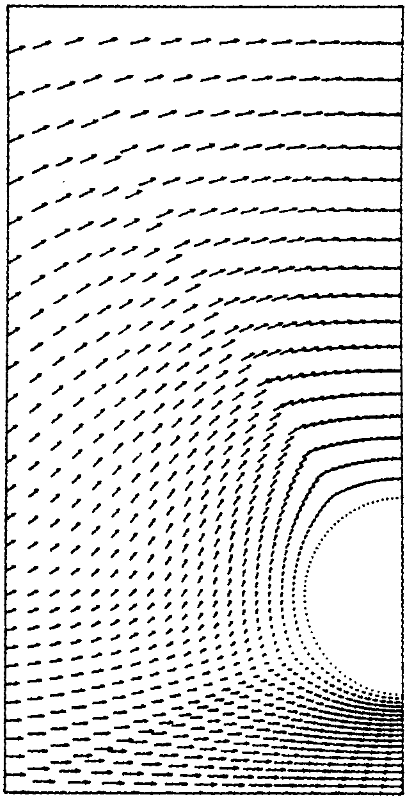

Figure 6. Velocity vectors

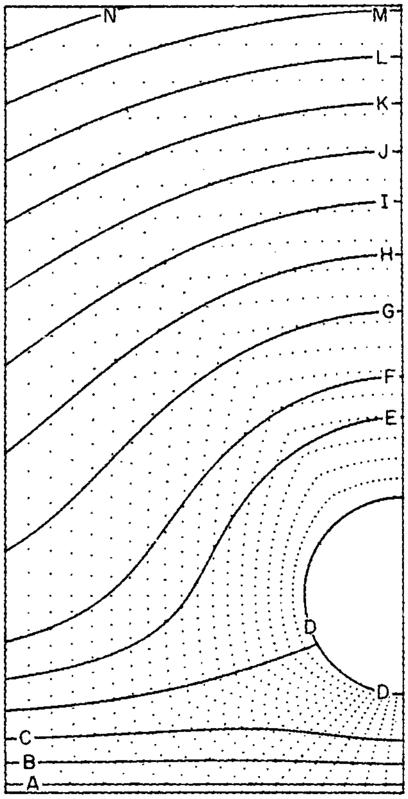

Figure 7. Streamlines. Increment between streamlines A–F is 0·05 and between F–N is 0·1

domain divided by the number of elements, $h \equiv (A/N_e)^{1/2}$, we find that the discrete r.m.s. norm of the error of each velocity component, i.e. $\|u - u_e\| \equiv (1/N)\sqrt{[\sum_i (u_i - u_i^e)^2]}$ where N is the total number of nodes and u_i^e is the exact solution at node i, converges like $O(h^{1\cdot9})$ for bilinear and $O(h^{3\cdot3})$ for biquadratic elements whereas the corresponding norms for the pressure are $O(h^{2\cdot8})$ and $O(h^2)$. (See Reference 25 for details.) The corresponding tractions, f_x and f_y, exerted by the cylinder on the fluid are displayed in Figures 8–11 in which the exact, consistent, and Gauss point values are plotted.

Note added in proof: (A plotting error causes the exact solution to appear discontinuous *near* $\theta = 0$ and π, rather than *at* $\theta = 0$ and π.)

The results of the mesh refinement experiments establish that the consistent method with mass lumping (row sum) is super-convergent, $O(h^4)$, for the bilinear element and optimally convergent, $O(h^2)$, for the biquadratic element for both f_x and f_y when the two end point nodes of the interval along the cylinder are excluded from the calculations of the norm. The inclusion of such end points, where both f_x and f_y suffer discontinuities, leads to a significant convergence rate degradation, $O(h^{1\cdot5})$ and $O(h)$ for f_x and $O(h)$ and $O(h)$ for f_y, respectively. The latter is the ramification of the Gibbs jump phenomenon which occurs in the approximation of a discontinuity and causes the oscillations which are clearly evident in Figures 8 and 9, especially in the consistent

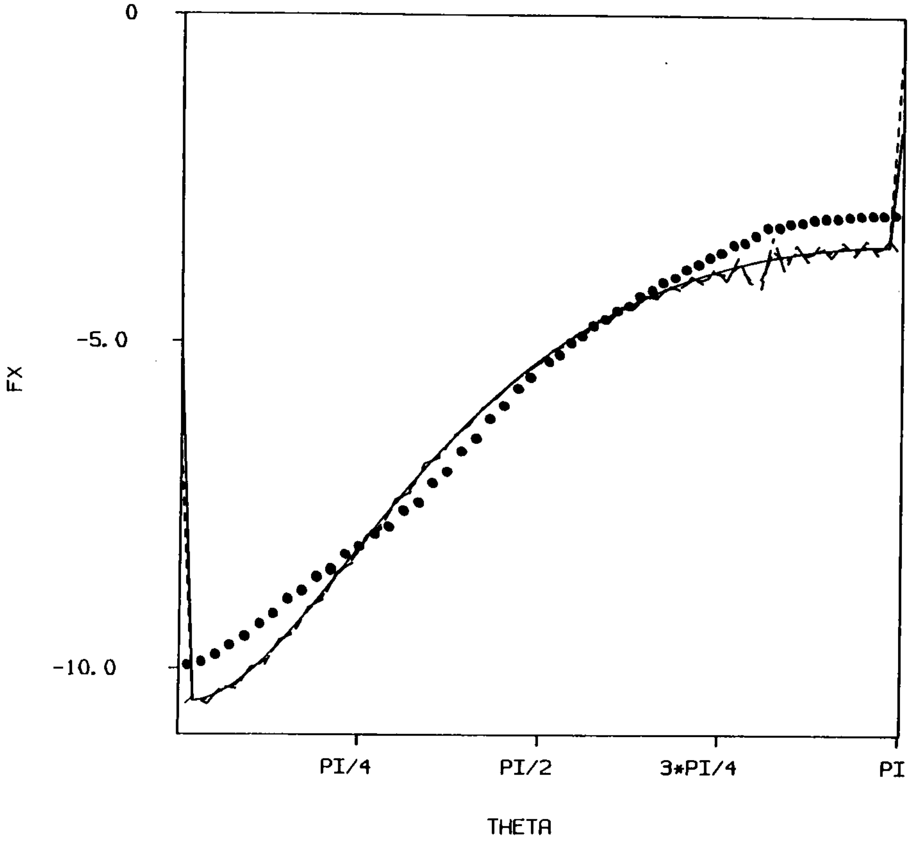

Figure 8. x-component of the force on a cylinder for a 4/1 element as a function of angle θ. $\theta = 0$ at the bottom of the cylinder: exact solution ———; consistent method with lumped mass -----; consistent method with consistent mass————; Gauss point method ● ● ● ● ● ●

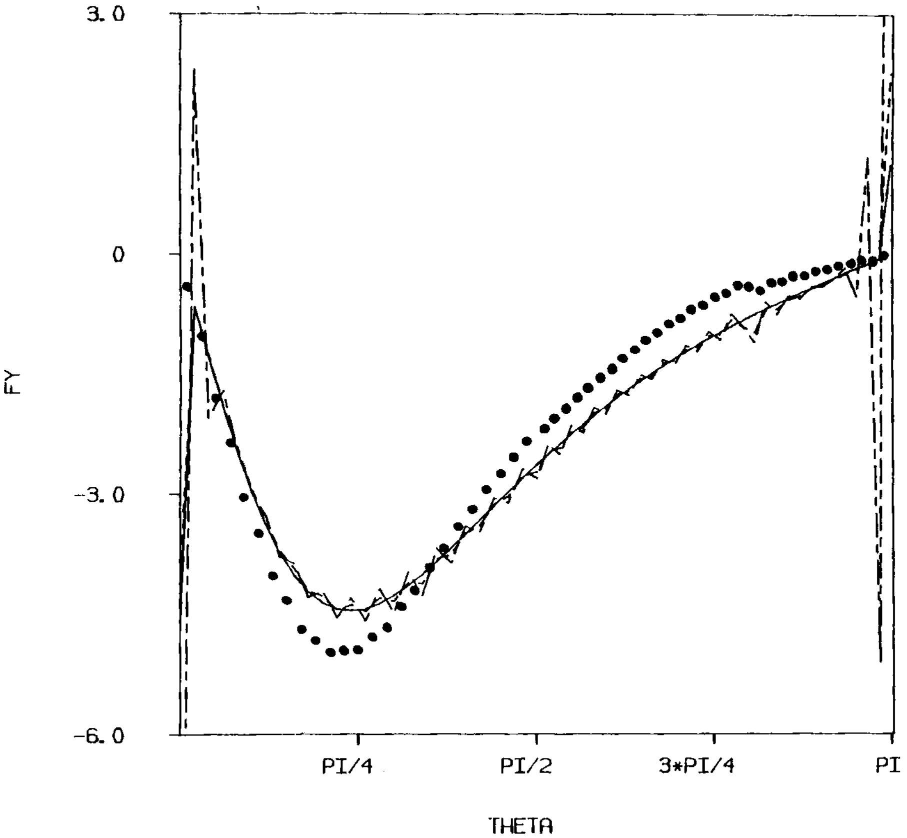

Figure 9. y-component of the force on a cylinder for 4/1 element as a function of angle θ. Curve designations are the same as in Figure 8

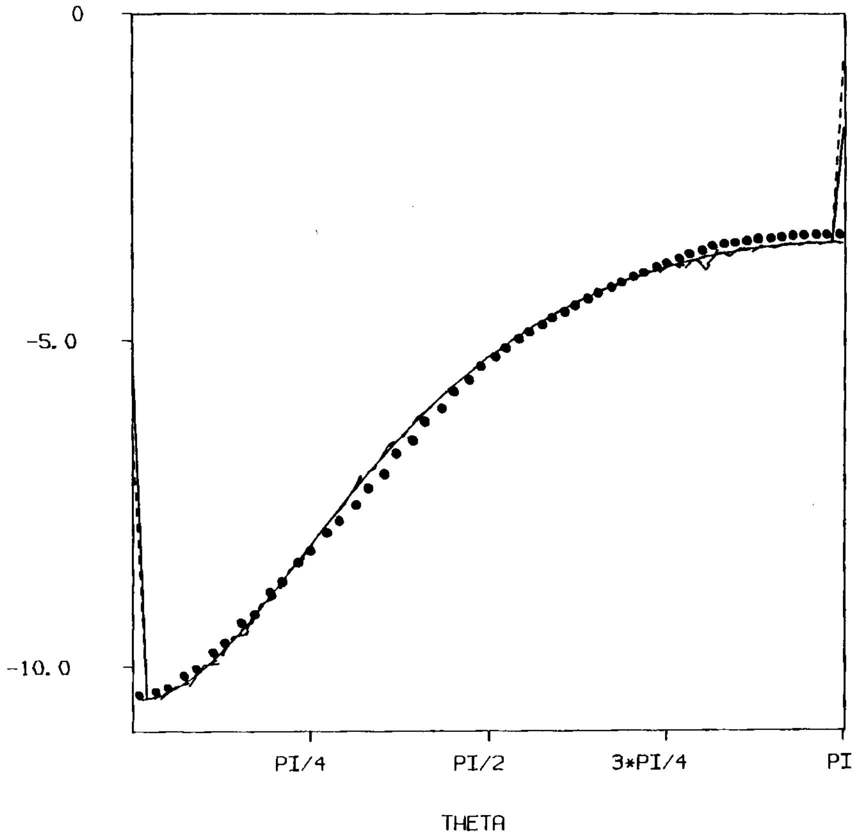

Figure 10. x-component of the force on a cylinder for a 9/3 element as a function of angle θ. Curve designations are the same as in Figure 8

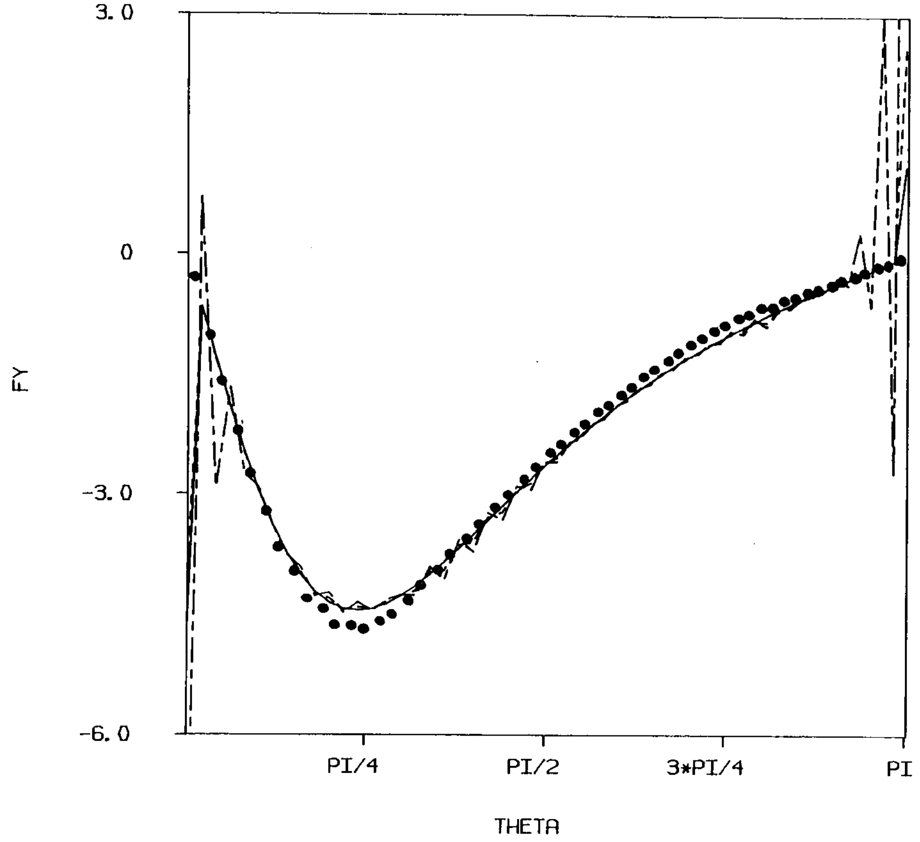

Figure 11. y-component of the force on a cylinder for a 9/3 element as a function of angle θ. Curve designations are the same as in Figure 8

mass formulation. The corresponding convergence rates for the conventional Gauss point evaluation of f_x and f_y are $O(h^{0.85})$ for bilinear elements and $O(h^{1.5})$ for biquadratic elements and it is evident from Figures 8–11 that this method is inferior to the consistent method except possibly near end-points. In addition to the overall accuracy, other noteworthy features in Figures 8–11 which also carry over to the other cases studied are the smoothness of the numerical results except near end-points and the faithfulness in reproducing the general trend of the exact solution. Also noteworthy are the oscillations which appear to be excited near $\theta = 0.75\pi$; their cause appears to be the *poor* elements adjoining the straight line directed from the upper left-hand corner of the domain to the boundary of the cylinder which is apparent in Figure 5. The change in shape in adjoining elements is less than subtle, a feature which degrades the finite element interpolants in this region and generates a noticeable local error in the derived forces on the cylinder. Of course, mesh refinement or *a priori* mesh smoothing via, for example, an equipotential method, diminishes this effect.[25]

CONCLUDING REMARKS

A consistent finite element method for computing certain derived boundary quantities has been considered. The method has been demonstrated to lead to more accurate results on both regular and isoparametric meshes than the conventional Gauss point evaluation of such quantities in both thermal and flow problems. The application of such methods has also been recently shown by Lynch[8] to be very advantageous in modelling the moving interface in phase change problems.

Since the direction and often the magnitude of such derived quantities are discontinuous at surface points with discontinuous normals, such as corners, the method tends to exhibit a Gibbs jump phenomenon and often local oscillations; consequently, the local accuracy of the technique is usually degraded in such situation. However, away from these points optimal or in some cases superconvergent results are obtained.

ACKNOWLEDGEMENTS

During the development of these methods we have benefited from useful and enlightening discussions with D. Griffiths (University of Dundee, Scotland). Also we would like to acknowledge the assistance of Mr. G. Hardin and Dr. A. Peskin in the numerical experiments. RLS, BEE and MKM would also like to acknowledge support from the U.S. Army Research Office (Grant DAAG29-83-K-0112) and the U.S. National Science Foundation (Grant CBT-8513570). The contributions of PMG and RLL were performed under the auspices of the U.S. Department of Energy at the Lawrence Livermore National Laboratory (Contract W-7405-Eng-48). We also thank the Institute for Computational Studies at Colorado State University for computing resources and consultation.

REFERENCES

1. J. Wheeler, 'Simulation of heat transfer from a warm pipeline buried in permafrost', *74th National Meeting of The American Institute of Chemical Engineers*, New Orleans, March 1973.
2. B. E. Larock and L. R. Herrmann, 'Improved flux prediction using low order finite elements', in W. G. Gray and G. F. Pinder (eds), *Proc. First Intl. Conf. On Finite Elements In Water Resources*, July 1976, Pentech Press, London, 1977.
3. R. S. Marshall, J. C. Heinrich and O. C. Zienkiewicz, 'Natural convection in a square enclosure by a finite-element, penalty function method using primitive fluid variables', *Num. Heat Trans.*, **1**, 315–330 (1978).
4. P. M. Gresho, R. L. Lee and R. L. Sani, 'The Consistent Method for Computing Derived Boundary Quantities when the Galerkin FEM is Used to Solve Thermal and/or Fluids Problems', in R. W. Lewis *et al.* (eds), *Proc. 2nd Int. Conf. on Numerical Methods in Thermal Problems*, Pineridge Press, 1981, pp. 663–675.
5. E. A. Thornton, 'Computation of consistent boundary quantities in finite element thermal-fluid solutions', in T. Kawai (ed.), *Finite Element Flow Analysis*, University of Tokyo Press, Tokyo, Japan, 1982, pp. 263–270.
6. S. P. Kjaran and S. T. Sigurdsson, 'Treatment of time derivative and calculation of flow when solving groundwater flow problems by GFEM', *Adv. Water Resources*, **4**, 23–33 (1981).
7. D. R. Lynch, 'Mass conservation in finite element groundwater models', *Adv. Water Resources*, **7**, 67–75 (1984).
8. D. R. Lynch, 'Heat conservation in deforming element phase change simulation', *J. Comp. Phys.*, **57**, 303–317 (1985).
9. D. R. Lynch, 'Mass balance in shallow water simulations', *Communication in Applied Numerical Methods*, **1**, 153–159 (1985).
10. J. Douglas Jr., T. Dupont and M. Wheeler, 'A Galerkin procedure for approximating the flux on the boundary for elliptic and parabolic boundary value problems', *R.A.I.R.O.*, **R-2**, 47–59 (1974).
11. G. F. Carey, D. Humphrey and M. F. Wheeler, 'Galerkin and collocation-Galerkin methods with superconvergence and optimal fluxes', *Int. J. numer. methods eng.*, **17**, 393–950 (1981).
12. T. F. Dupont, 'A short survey of parabolic Galerkin methods', in D. F. Griffiths, (ed.), *The Mathematical Basis of Finite Element Methods*, Clarendon Press, 1984, pp. 27–40.
13. Y. Hasbani and M. Engelman, 'Out of core solution of linear equations with non-symmetric coefficient matrix', *Comp. and Fluids*, **7**, 13 (1979).
14. R. L. Lee, P. M. Gresho and R. L. Sani, 'Smoothing techniques for certain primitive variable solutions of the Navier–Stokes equations', *Int. j. numer. methods eng.*, **14**, 1785–1804 (1979).
15. P. M. Gresho and R. L. Lee, 'Don't suppress the wiggles—they're telling you something!', *Computers and Fluids*, **9**, (2), 223–255 (1981).
16. R. L. Lee, P. M. Gresho, S. T. Chan and R. L. Sani, 'A comparison of several conservative forms for finite element formulations of the incompressible Navier–Stokes or Boussinesq equations', in R. H. Gallagher, D. N. Norrie, J. T. Oden and O. C. Zienkiewicz (eds), *Finite Elements in Fluids, Vol. 4*, Wiley, 1982, pp. 21–45.
17. H. Saito and L. E. Scriven, 'Study of coating flow by finite element method', *J. Comp. Phys.*, **42**, 53- (1981).
18. P. M. Gresho, R. L. Lee, S. Chan and R. L. Sani, 'Solution of the time-dependent incompressible Navier–Stokes and

Boussinesq equations using the Galerkin finite element method', *Lecture Notes in Mathematics, No. 771*. Springer-Verlag 1980, pp. 203–223.

19. M. S. Engelman, R. L. Sani and P. M. Gresho, 'The implementation of normal and tangential velocity boundary conditions in finite element codes for incompressible fluid flow', *Int. j. numer. methods fluids*, **2,** 225–238 (1982).
20. A. G. Hutton and R. M. Smith, 'The prediction of laminar flow over a downstream-facing step by the finite element method', *Central Electricity Generating Board report no. RD/B/N3660*, Research Division, Berkeley Nuclear Laboratory, Berkeley, U.K., April 1979.
21. J. Leone and P. M. Gresho, 'Finite element simulations of steady two dimensional, viscous incompressible flow over a step', *J. Comp. Phys.*, **41,** 167–191 (1981).
22. A. Chorin and F. Marsden, *A Math Intro to Fluid Dynamics*, Springer-Verlag, 1979.
23. G. H. Wannier, 'A contribution to the hydrodynamics of lubrication', *Quart. Appl. Math.*, **8,** 1–32 (1950).
24. D. J. Jeffrey and Y. Onishi, 'The slow motion of a cylinder next to a plane wall', *Q. J. Mech. Appl. Math.*, **34,** 129–137 (1981).
25. M. K. Maslanik, R. L. Sani and P. M. Gresho, 'An isoparametric finite element Stokes flow test problem', to be submitted to *Communications in Applied Numerical Methods.*
26. M. S. Engelman, R. L. Sani, P. M. Gresho and M. Bercovier, 'Consistent vs. reduced integration penalty methods for incompressible media using several old and new elements', *Int. j. numer. methods fluids*, **2,** 25–42 (1982).