On the Finite Element Modeling of the Scalar Transport Equation

Robert L. Sani* — Philip M. Gresho**

* *Department of Chemical Engineering*
University of Colorado
Boulder, Colorado

** *Lawrence Livermore National Lab*
Livermore, California USA

ABSTRACT. *This material was presented (by R.L.S.) as part of a three hour lecture on finite element modeling of incompressible and Boussinesq flows at the summer school organized by IUSTI, Université de provence.*

KEY WORDS : *scalar transport, conversation, incompressible flow.*

1. Introduction

An appropriate starting point for the study and numerical simulation of incompressible fluid flow is that of the simpler but important equation of advection-diffusion, in which the velocity field is presumed to be known. Indeed, many fluid flow simulations are primarily (or ultimately) concerned with the transport and diffusion of scalar quantities such as 'heat' (temperature) or concentration (e.g., air pollution). Unfortunately, even in these cases, the more-difficult-to-obtain velocity field must usually be computed first. Here, however, we shall assume that the velocity is known, either analytically or from a numerical solution of the incompressible Navier-Stokes equations; we will turn to the problem of computing the velocity field itself subsequently. Finally, since the advection-diffusion equation is, in many ways, prototypical of the (much more difficult) Navier–Stokes equations, it is useful to study it first but in less detail since it will also be addressed by other lecturers. The focus here will be on some *special topics*. Most of the material presented herein represents a small portion of our hopefully soon to be completed book.

2. The Continuum Scalar Transport Equation

2.1 *The Advective (Convective) Form*

The conservation principle for energy or chemical species (mass) can often be well approximated by the following partial differential equation (PDE)—the scalar transport equation—written here in terms of temperature (T):

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \nabla \cdot (\mathbf{K} \cdot \nabla T) + S \quad , \qquad [1]$$

where the velocity field, $\mathbf{u}(x, t)$ is given and satisfies $\nabla \cdot \mathbf{u} = 0$, as is the diffusion tensor, $\mathbf{K}$, and the source term, S.

The solution to [1] will generally be sought within a bounded domain, Ω, with boundary, $\partial\Omega$. Given an initial distribution of temperature, Equation [1] can, in principle, be solved subject to an appropriate set of boundary conditions (BC's), which typically are:

$$T = T_D \quad \text{on } \Gamma_D \qquad [2]$$

and

$$\mathbf{n} \cdot (\mathbf{K} \cdot \nabla T) + H\,(T - \tilde{T}) = q \quad \text{on } \Gamma_N \quad , \qquad [3]$$

where $\partial\Omega$ is composed of the two non-overlapping segments, Γ_D and Γ_N; T_D, $\tilde{T}$, H (heat transfer coefficient, and q (specified normal heat flux into Ω are given functions on the appropriate portion of the boundary, and $\mathbf{n}$ is the outward pointing unit normal vector.

While $\mathbf{K}$ can, in general, be a full (but symmetric and positive-definite) second-order tensor (a 2×2 matrix of coefficients in 2D, and 3×3 in 3D) representing anisotropic diffusion, it is usually much simpler; e.g., a diagonal matrix or even a scalar. Since this presentation is largely introductory, we shall often consider the simplest case of a scalar (and constant) diffusion coefficient κ.

Finally, the statement of the scalar transport problem (abbreviated henceforth by AD; Advection-Diffusion) is completed by specifying an initial condition (IC):

$$T(x, 0) = T_0(x) \quad \text{in } \Omega \quad , \qquad [4]$$

where T_0 is a given function of position.

Before continuing, we make several remarks:

1. The IC need not (and generally does not) satisfy the BC's, but if it does, the resulting solution will be smoother; i.e., possess higher-order spatial derivatives—especially if [4] satisfies [2], the Dirichlet BC. (This "flexibility' regarding IC's and BC's will be partially lost when we advance to the NS equations.)
2. A practical application of BC [3] occurs when Γ_N is a wall (at which $\mathbf{u} = \mathbf{0}$, usually) containing a heater (for $q > 0$) and on the other side of which flows a fluid at temperature $\tilde{T}$.

3. Another practical and very common use of [3] occurs when Γ_N represents an 'outflow' ($\mathbf{n} \cdot \mathbf{u} > 0$) boundary that is usually artificial/synthetic in the real world but very real in the mathematical modeling world. Here the use of $H = 0$ and $q = 0$ is often effective as an *approximation* to the true coupling with the rest of the universe.
4. If $\mathbf{K} = \mathbf{0}$, we have the limiting case of pure advection, a hyperbolic equation for which no BC can be specified at outflow; i.e., BC [3] must be *dropped* in this situation, because the theory of characteristics tells us that T must be specified at inlet points on Γ ($\mathbf{n} \cdot \mathbf{u} < 0$) but that there is no BC at outlet points—at these points, the PDE itself prevails.
5. There will generally occur a singularity at the junction of Γ_D and Γ_N, at which certain derivatives of T (e.g., heat flux) will fail to exist (be unbounded).

The unique solution to [1]–[4], when it exists, is called a classical solution to distinguish it from the weak solution to be presented later. In particular, given sufficiently smooth data ($\mathbf{u}$, $\mathbf{K}$, T_D, $\bar{T}$, q, T_0, *and* $\partial\Omega$), it will possess at least two continuous spatial derivatives—at least for $t > 0$. (It will do so at $t = 0$ only if certain compatibility conditions are satisfied.)

2.2 *The Divergence (Conservation) Form*

Since $\nabla \cdot \mathbf{u} = 0$, an equivalent form of [1]–[4] is (for $\mathbf{K} \to \kappa =$ constant) is

$$\begin{aligned} \frac{\partial T}{\partial t} + \nabla \cdot (\mathbf{u}T) &= \kappa \nabla^2 T + S \quad , \quad \text{or} \\ \frac{\partial T}{\partial t} + \nabla \cdot (\mathbf{u}T - \kappa \nabla T) &= S \quad , \end{aligned} \tag{5}$$

which is called the (flux-) divergence form, since

$$\mathbf{q}_A \equiv \mathbf{u}T \tag{6}$$

is the advective flux vector and

$$\mathbf{q}_D \equiv -\kappa \nabla T \tag{7}$$

is the diffusive flux vector. That is, with $\mathbf{q}_T \equiv \mathbf{q}_A + \mathbf{q}_D$, the total flux vector, [5] is clearly

$$\frac{\partial T}{\partial t} + \nabla \cdot \mathbf{q}_T = S \quad , \tag{8}$$

which is called a *conservation* form because integration over Ω gives directly, via the divergence theorem, the following global conservation 'law':

$$\frac{d}{dt} \int T = \int S - \int_{\Gamma} \mathbf{n} \cdot \mathbf{q}_T \quad ; \tag{9}$$

i.e., the total energy (or mass if T represents a concentration or mass fraction) changes (decreases) only by the net flux of T out of the domain through the boundary—except of course for the source term.

Here and hereafter we employ the imprecise but convenient notation that $\int(\cdot)$ means integration of $(\cdot)$ over Ω and $\int_\Gamma(\cdot)$ to denote integration of $(\cdot)$ over the boundary of Ω.

Now it is clear that the same global conservation law could also have been derived from [1] because $\nabla \cdot \mathbf{u} = 0$. So, one may reasonably ask, what is the reason for discussing the divergence form? The detailed answer will come later and is two parts, which we merely hint at now: (1) In the weak formulations of the transport equation, the two forms thus far discussed—advective form and divergence form—*can* differ owing to different natural BC's, and (2) in the spatially-discretized equations, we generally do *not* obtain $\nabla \cdot \mathbf{u} = 0$ pointwise, with the result that only the divergence form can assure global conservation—an assertion we shall later prove. This leads naturally to the subject of the next section.

2.3 *Conservation Laws*

Oftentimes one of the goals of approximate solutions to PDE's, in addition to the principal goal of finding a cost-effective approximate solution that is close to the continuum solution, is the assurance that the approximate solution will satisfy *discrete* approximations to certain *global* conservation laws that are satisfied by the continuum solution *and* that are basically independent of the 'local error'; i.e., they are satisfied on the coarsest of meshes. The principal reason for this goal is the desire to attain stable and bounded numerical solutions, independently of the issue of accuracy.

Toward this end then, we present next a brief discussion of the relevant conservation laws for the AD equation, so that we can set our sights toward the proper *goals* when later generating numerical approximations. The first of these, global conservation of T, has already been derived—in [9], which we restate in expanded form:

$$\frac{d}{dt}\int T = \int S - \int_\Gamma \mathbf{n} \cdot (\mathbf{u}T - \kappa \nabla T) \quad , \qquad [10]$$

showing that *internal* transport (i.e., within Ω) of T via the principle transport processes (advection and diffusion) makes *no* contribution to the *global* change of T—they merely redistribute it within Ω.

Invoking BC [3] in [10] yields another equivalent form of the global energy conservation statement:

$$\frac{d}{dt}\int T = \int S + \int_{\Gamma_N} \left[q + H(\tilde{T} - T)\right] + \int_{\Gamma_D} \kappa \frac{\partial T}{\partial n} - \int_\Gamma \mathbf{u} \cdot \mathbf{n} T \quad , \qquad [11]$$

in which the individual boundary contributions are more clearly displayed.

If a *steady* solution is sought for the somewhat special case of $\Gamma_D = 0$, $H = 0$, and $\mathbf{u} \cdot \mathbf{n} = 0$ on Γ, [11] yields a constraint on the *data*; i.e., it states that $0 = \int S + \int_\Gamma q$. If this solvability condition is not satisfied, the problem is ill-posed and no solution exists—because the given data preclude a global balance and are thus inconsistent.

Another energy-like quantity that is often of interest—is a quadratic one: How does $E \equiv \int T^2$, a positive-definite quantity, behave? (Note that $\int T$ could be well-behaved even if T is locally 'poorly-behaved'; e.g., small regions of large negative T could be cancelled by small regions of large positive T.) To answer this question, we first multiply [5] by T and integrate over Γ:

$$\int T\frac{\partial T}{\partial t} + \int T\nabla \cdot (\mathbf{u}T - \kappa\nabla T) = \int ST \quad .$$

Application of the divergence theorem after an integration by parts of the two transport terms yields, with $\nabla \cdot \mathbf{u} = 0$,

$$\frac{1}{2}\frac{d}{dt}\int T^2 = \int ST - \kappa \int \nabla T \cdot \nabla T - \frac{1}{2}\int_\Gamma \mathbf{n} \cdot (\mathbf{u}T^2 - \kappa\nabla T^2) \quad , \qquad [12]$$

which shows the following:

1. If $S > 0$, the source term will act to increase (decrease) E if $T > 0$ (< 0).
2. Dissipation—the second term on the RHS—will (try to) decrease E monotonically (because $\int \nabla T \cdot \nabla T \geq 0$) and is the reason that diffusional processes are called dissipative. It is noteworthy that this type of 'damping' is present in the T^2 equation, but not in the T equation—internal diffusion acts to equalize T, conserve its integral, and decrease $\int T^2$.
3. The boundary terms show that T^2, like T, is subject to inflow/outflow along Γ by (again, like T) both transport processes.
4. If $\mathbf{n}\cdot\mathbf{u} = 0$ (contained flow) and $\mathbf{n}\cdot\nabla T = 0$ on Γ ('insulated' container), we have $\frac{d}{dt}\int T = \int S$ and $\frac{1}{2}\frac{d}{dt}\int T^2 = \int ST - \kappa\int \nabla T\cdot\nabla T$. In a situation with no source term, $\int T = \int T_0$—where $T_0(x)$ is the initial temperature—and E decays monotonically, showing that $\dot{E} \to 0$ and $T \to$ constant as $t \to \infty$; i.e., a steady state will be attained in which the constant final temperature is the same as the average initial temperature.
5. Finally, if $\kappa = 0$ (pure advection) and $\mathbf{n} \cdot \mathbf{u} = 0$ on Γ, the sourceless situation will conserve *all* powers of T; i.e., it then follows that $\int T^m = \int T_0^m$, $m = 1, 2, \ldots$

These results can be regarded as some goals for the approximate (numerical) solutions. We will later return to these conservation issues after deriving the numerical approximations—both semi-discrete, which leads to a set of ordinary differential equations (ODE's) in time, and fully-discrete in which a time-marching method has been selected.

2.4 *Weak Forms of the PDE's/Natural Boundary Conditions*

The next step in a GFEM solution is to recast the governing PDE—either [1] or [5]—into the weak (or Galerkin) form, sometimes also referred to as a variational form. Here and hereafter, when we speak (loosely, sometimes) of the weak form of an *equation* (PDE), we are usually referring to a weak *formulation* of the spatial part of the *problem* (PDE plus BC's); i.e,. weak forms generally come with BC's. We also state at the outset that while the classical statement of a problem (PDE+BC's+IC's, also often referred to as an IBVP—Initial-Boundary-Value Problem) is generally unique and unambiguous, there is usually no unique *weak* statement of the same problem. But while there may exist alternate weak formulations of a *given* problem, they are actually *equivalent*—at least when a classical solution *exists*, in which case the solution is said to be 'sufficiently smooth.' Some weak formulations, however, are more useful than others because (at least) they more efficiently and more 'naturally' and take account of the BC's. Part of the 'game,' therefore, is to find the most appropriate weak form—a task that is often non-obvious and non-trivial—especially when we consider the NS equations in the next chapter. (Thus, the FDM problems of 'how to discretize each operator and how to treat each term at a boundary?' is replaced by the FEM problem of 'selection of the weak form.')

Find $T(\mathbf{x}, t)$ in H_E^1 such that

$$\int \left[w \left(\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T \right) + \nabla w \cdot (\mathbf{K} \cdot \nabla T) \right] =$$
$$\int wS + \int_{\Gamma_N} w \left[q - H(T - \tilde{T}) \right] \quad , \qquad \forall w \in H_0^1 \quad ,$$

which we rearrange to place the unknown boundary temperature on the LHS; i.e., find $T(x, t) \in {H_E}^1$ such that

$$\int \left[w \left(\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T \right) + \nabla w \cdot (\mathbf{K} \cdot \nabla T) \right] + \int_{\Gamma_N} wHT = \int wS + \int_{\Gamma_N} w(q + H\tilde{T})$$

$$\forall w \in H_0^1 \quad , \qquad [13]$$

where ${H_E}^1$ is that set of piecewise once-differentiable functions in Ω that satisfy the *essential* BC, [2], on Γ_D. This is the final weak form of [1]–[4]; it incorporates *automatically* BC [3] and *can be solved*, in principle at least, for $T(x, t)$ once the initial data, [4], are supplied at $t = 0$. The *weak* solution, can (but need not) now reside in a larger function space than do solutions of [1] since the weak solution need not even possess second spatial derivatives, at least in the classical sense. A final comment on this weak formulation is that the BC [3] has been incorporated into the solution in a (relatively) natural way and is the reason (or *one* reason, at least that such BC is called a natural boundary

condition (NBC); it (the Neumann BC for the Laplacian operator) is 'natural' to this weak formulation—and [13] is the 'natural' weak form of [1]–[4].

If the solution of the weak form of the problem is sufficiently smooth, then that solution is also a classical solution of the same problem. The key word is IF, a word that is *missing* when going the other direction—i.e., a classical solution is *always* also a weak solution. This distinction is actually rather important in practice, wherein most (or at least many) problems posed do not satisfy all of the smoothness requirements in order that a classical solution exist. (It is perhaps also worth pointing out that conventional finite difference approximate solutions to such problems can also converge only to a weak or generalized solution in such cases.)

Before taking the next step toward a finite element solution, we digress to consider *another* weak formulation: If we start with the conservation form of the PDE, [5], to generate the weak form, it may seem natural (although it is not necessary) to also integrate the other divergence term, $w\nabla\cdot(\mathbf{u}T)$, by parts, so that the *total* flux, $\mathbf{u}T - \mathbf{K}\cdot\nabla T$, is thus so treated. The result is

$$\int\left[w\frac{\partial T}{\partial t} + \nabla w\cdot(\mathbf{K}\cdot\nabla T - \mathbf{u}T)\right] = \int wS + \int_{\Gamma_N} w\mathbf{n}\cdot(\mathbf{K}\cdot\nabla T - \mathbf{u}T) \quad ,$$

in which the total flux appears in both domain and boundary integrals. This suggests, properly, that *if* the Neumann/Robin BC were

$$\mathbf{n}\cdot(\mathbf{K}\cdot\nabla T - \mathbf{u}T) + H(T-\tilde{T}) = q \text{ on } \Gamma_N \quad , \qquad [14]$$

instead of [3], then the *appropriate* weak formulation would be:

Find $T(x,t)$ in H^1_E such that

$$\int\left[w\frac{\partial T}{\partial t} + \nabla w\cdot(\mathbf{K}\cdot\nabla T - \mathbf{u}T)\right] + \int_{\Gamma_N} wHT = \int wS + \int_{\Gamma_N} w(q+H\tilde{T}) \;\; \forall w\in H^1_0, \qquad [15]$$

rather than that given by [13]. Remarks:

1. If the advection term (in the flux-divergence form) were not integrated by parts, the resulting weak form would be equivalent to that derived earlier—with $\mathbf{u}\cdot\nabla T$ replaced by $\nabla\cdot(\mathbf{u}T)$ in [13]—and would satisfy the natural BC implied by it; i.e., [3] rather than [14] above.
2. While either form of the PDE (advective or flux divergence) could actually be used to solve the AD equation with either (natural) BC—[3] or [14]—in the weak form, the former is a more natural choice if the BC is [3] and the latter if it is [14].
3. We will have more to say regarding the choice of weak formulation and associated BC's after we 'discretize the weak form' via the finite element method.
4. The weak formulations of the scalar transport equation presented above contain several important and simpler (usually) special cases; examples:

(i) $\mathbf{u} = \mathbf{0}$ gives the transient heat equation (parabolic equation) in a weak form, (ii) $\mathbf{u} = \mathbf{0}$ and $\frac{\partial T}{\partial t} = 0$ gives a weak formulation of a Poisson problem (elliptic equation), (iii) $\mathbf{K} = \mathbf{0}$ gives the pure advection equation (hyperbolic) in a weak form, which also requires $g = 0$ and $H = 0$.

5. We shall concentrate mainly on the first weak form, [13], for reasons that will be explained later.

3. The Finite Element Equations/Discretization of the Weak Form

3.1 *Advective Form*

We now address the issue of 'solving'—albeit approximately—the weak form of the problem, [13], and mention up front that thus far it is far from obvious that solving [13] is any easier than solving [1]–[4]. But it is the weak formulation upon which the finite element method—a weighted residual method (see Finlayson and Scriven (1966)) or a projection method (see Reddy (1984))—is based. *The FEM is a general and systematic technique for obtaining approximate solutions of weak formulations;* i.e., it is a method of 'discretizing' the weak form with the result that the underlying function spaces become finite—and thus amenable to representation via computer. Of course the resulting solution in this 'truncated' function space is only an approximation to the true weak solution; i.e., of the solution to [13] in this case. In addition, the approximate solution is based on the *approximation of functions*—be they given or need to be found—via piecewise-polynomials (PP's) defined on the spatial domain of the problem. Different FEM solutions—by which here and hereafter we always mean *approximate* solutions to a given continuous problem—arise from the use of different PP's and/or different weak forms, all of which are ostensibly trying to solve the same IBVP. But once the choice of weak formulation is made and once the choice of PP (e.g., linear, or quadratic, or cubic) is made, there are virtually *no more choices* available to the analyst (except of course the difficult and important one of how many and what distribution of PP's are to be used—i.e., the number and distribution of nodes/elements in the mesh), such as 'How should I treat this term or that term near and/or at this curved boundary, or at that corner?'

The PP's of the FEM are also called *basis* functions or shape functions and are said to 'span the space': any function in this finite-dimensional subspace is presumed to be representable by an appropriate linear combination of these basis functions. When the test functions, $\{w\}$, are also represented by a linear combination of the *same* basis functions used to approximate the solution, which we assume to be the case herein, the FEM that evolves is called the Galerkin FEM or GFEM (see, e.g., Finlayson and Scriven (1966)). If the test functions differ from the basis functions, we have a so-called Petrov-Galerkin method which leads to different numerical approximations; we'll say more about some of these methods later. And this is our next task—viz., to apply the GFEM to the weak form of the AD equation given by [13]. Since the

integrals in [13] involve no derivatives of higher order than one, we can (and do) employ the simplest class of PP's called C^0 functions; the basis functions are (piecewise) continuous (and linearly independent) and their first derivatives, while discontinuous [they typically suffer (finite) jumps at node points], are square integrable—and that is all that is needed for the terms (viz. the diffusion term) in [13] to 'make sense'; i.e., they can be evaluated. Second and higher order derivatives are not required to 'make sense' nor even to exist.

The next step toward a GFEM solution is to represent the unknown function, $T(x, t)$, in [13] as a linear combination of (known) basis functions (PP's) with unknown amplitude coefficients that are to be determined in such a way that the resulting approximate solution function, which we call $T^h(x, t)$, represents $T(x, t)$ from [13] in a reasonable way. The generic symbol h is used both to represent a typical (or maximum) element size (length) on the discrete mesh and to remind us that we are henceforth dealing with an *approximate* solution—$T^h(x, t) \neq T(x, t)$ from [13], but we hope that $T^h(x, t) - T(x, t)$ is 'small.' Thus, we write

$$T^h(x, t) = \hat{T}(x, t) + \sum_{j=1}^{N} T_j(t)\phi_j(x) \quad , \qquad [16]$$

where ϕ_j is the j-th basis function, $T_j(t)$ is the j-th unknown (to-be-determined) amplitude coefficient, N is the number of nodes (in Ω and on Γ_N) at which T^h is to be determined, and $\hat{T}(x, t)$ is a given function (to be discussed in detail below) whose purpose is to ensure that $T^h(x, t)$ satisfies the Dirichlet (essential) BC, [2], since a property of the $\{\phi_j\}$, inherited from that of $\{w\}$, is that $\phi_j = 0$ for $x \in \Gamma_D$; i.e., for points located on Γ_D, [16] gives $T^h(x, t) = \hat{T}(x, t) \approx T_D$ of [2]. (It may be worthwhile to emphasize that N is not the *total* number of nodes in the mesh; it does not include those on Γ_D.)

Thus, the finite dimensional/GFEM statement of [13] is obtained by inserting [16] into the finite dimensional analog of [13] to obtain the following set of ordinary differential equations (ODE's) for the amplitude coefficients (nodal values of T):

$$\sum_{j=1}^{N} \left\{ \dot{T}_j \int_{\Omega} \phi_i \phi_j + T_j \left[\int_{\Omega} \phi_i \mathbf{u} \cdot \nabla \phi_j + \nabla \phi_i \cdot (\mathbf{K} \cdot \nabla \phi_j)] + T_j \int_{\Gamma_N} H \phi_i \phi_j \right\} = \right.$$

$$\int_{\Omega} \phi_i S + \int_{\Gamma_N} \phi_i (q + H\tilde{T}) - \left\{ \int_{\Omega} \left[\phi_i \left(\frac{\partial \hat{T}}{\partial t} + \mathbf{u} \cdot \nabla \hat{T} \right) + \nabla \phi_i \cdot (\mathbf{K} \cdot \nabla \hat{T}) \right] \right.$$

$$\left. + \int_{\Gamma_N} \phi_i H \hat{T} \right\} \qquad \text{for } i = 1, 2, \ldots, N \quad , \qquad [17]$$

where $\dot{T} \equiv dT_j/dt$, and we note that the entire term in curly brackets on the RHS is actually a formal method of enforcing the essential BC, [2], and is not nearly so cumbersome in practice as it appears at first glance.

Further remarks:

1. The solution of [17] approximates that of [13] which represents a generalized solution of [1]–[4].
2. The entire approximate solution (once IC's are set) of the scalar transport problem [1]–[4] is contained in this set of equations—both at *all* points in Ω (via solving [17] for nodal values and by using [16] elsewhere) and at *all* points on Γ_N (where T is also an unknown function owing to the derivative BC, [3]).
3. Hopefully the dual use of the symbol N—for Neumann (Γ_N) and for the number of nodal unknowns—will not cause a problem.
4. It is noteworthy (and significant, and perhaps even somewhat amazing) that none of the individual basis functions satisfies the NBC of [3], yet the solution of [17] will do—albeit approximately (as indeed is the entire solution an approximate one) and more closely as N is increased—even when Γ_N has a complicated shape. This is in fact one of the major advantages of approximating the weak form rather than the strong form. (See Strang and Fix (1973), for more detailed discussions of the theory behind such 'unstable' BC's.)
5. The ODE's become algebraic equations ($\dot{T}_j = 0$) if the steady AD equation is being solved via the GFEM—a linear system of N equation in N unknowns.
6. An 'implicit' method of obtaining [17] that is sometimes used goes as follows: In the finite dimensional subspace associated with [13], the generic test function, w^h can be represented as $w^h = \sum_{i=1}^{N} a_i \phi_i$ and the statement, 'for every $w^h \epsilon H_0^1$' is replaced by 'Where the a_i are arbitrary', which leads to the following version of [17]: $\sum_{i-1}^{N} a_i \{$ LHS - RHS $\} = 0$, where LHS is the left hand side of [17], etc; and [17] then follows immediately since the a_i's are abitrary coefficients.

Before moving on to the discussion of the *solution* of [17], we must address two more issues: (i) The function $\hat{T}(x, t)$ and (ii) IC's.

The main job of $\hat{T}(x, t)$, as alluded to earlier, is to ensure that the approximate solution satisfies (closely if not exactly) the essential/Dirichlet/stable BC of [15]; there is no free lunch for these BC's. Wait and Mitchell (1985, pp. 88-91) present an interesting sample problem in which a comparison is made of 'blending functions' (which exactly satisfy the essential BC's) and finite element PP basis functions (which *interpolate* the BC's and are therefore exact only at the nodes). The result is that both 'work' quite well and neither is clearly superior. We shall (therefore?) follow common practice and use the *same* class of PP's to interpolate $\hat{T}(x, t)$ for $x \in \Gamma_D$ that are used to approx-

imate the solution (and the test functions, in the Galerkin method). That is, we take

$$\hat{T}(x, t) = \sum_{j=N+1}^{N_T} T_D(x_j, t)\phi_j(x) \quad \text{for } x_j \in \Gamma_D \quad , \qquad [18]$$

where N_T is the *total* number of nodes in the finite element mesh, and we quickly note that the implied ordering/numbering of the nodes (i.e., $j = 1, 2, \ldots, N, N+1, N+2, \ldots, N_T$) is definitely *not* appropriate for solution by the computer–it merely simplifies the presentation of the 'theory.'

The advantages of this choice (the interpolation of T_D) are several:

1. Simplicity; it is 'natural' to the FEM technique, and code writing is much easier.
2. *All* of the amplitude coefficients, $\{T_j(t)\}$—those in Ω and those on $\partial\Omega = \Gamma_D + \Gamma_N$—represent the *value* of the function $T^h(x, t)$ at the nodes. (This is not true if blending or other functional forms are employed.)
3. The function $\hat{T}(x, t)$ is of compact support; it is non-zero only on those elements that are contiguous to Γ_D and zero elsewhere. The bracketed term on the RHS of [17] is thus *zero* over most of the domain.

Turning now to the subject of initial conditions (finally), we mention that again at least one alternative to 'interpolation via the basis functions' exists, but that there is usually not a sufficiently compelling reason to introduce this more complicated technique—which is: Compute the 'consistent' IC's by setting $T^h(x, 0) = T_0(x)$ *weakly*; i.e., from [16] we obtain

$$\int T^h(x, o)\phi_i = \int \hat{T}(x, 0)\phi_i + \int \sum_{j=1}^{N} T_j(0)\phi_j\phi_i = \int T_o(x)\phi_i \quad \text{for } i = 1, 2, \ldots N \qquad [19]$$

which, via [18], is an NxN linear system for $\{T_j(0)\}$. We leave as an exercise the proof that this is the same $T^h(x, 0)$ that minimizes the following functional,

$$Q = \int [T^h(x, 0) - T_o(x)]^2 \quad , \qquad [20]$$

where $T^h(x, 0)$ is again expressed via [16] and [18]. The initial values thus obtained will generally *not* agree with $T_0(x)$ at the nodal points, but the resulting $T^h(x, 0)$ will be as close as possible—in the least squares sense—to $T_0(x)$ in Ω. (See Swartz and Wendroff (1969) for further discussion of these issues.) While this IC computation is indeed more consistent, we shall generally again follow precedent/common practice, and simply *interpolate* the initial data via

$$T_j(0) \equiv T_0(x_j), \quad j = 1, 2, \ldots N \quad , \qquad [21]$$

which again simplifies code writing and is usually sufficiently accurate (indeed, the error is zero at each node, so that the only error is that caused by interpolation).

A final remark on IC's: if $T_0(x)$ and $T_D(x, 0)$ disagree at the nodal points on Γ_D, the BC must prevail; i.e., it is necessary that $T_j(0) = T_D(x_j, 0)$ for all nodes on Γ_D. (If the IC and BC are the same on Γ_D, there is no jump there at $t = 0$ and the resulting soution will be smoother.)

The total GFEM problem has now been posed; viz. using [18], solve [17] for $T_j(t)$ with IC's obtained from [19] and (optionally, and rarely done in practice) use [16] to obtain $T^h(x, t)$, the full finite element solution. But 'solve [17]' is easier said than done—even though we now have only a *finite* number of unknowns.. We will thus later devote a fair amount of attention to methods for solving the ODE's of [17]—but first, we shall spend some time studying the ODE system that has been generated. To begin, we re-write the GFEM problem in the more compact matrix-vector form; viz.,

$$M\dot{T} + [N(\mathbf{u}) + K]T = f, \text{ for } t > 0 \quad , \qquad [22]$$

where $T \equiv (T_1, T_2, \ldots, T_N)^T$ is an N-vector of the nodal values, $T_j(t)$, which satisfy $T_j(0) = T_0(x_j)$ at $t = 0$. Also, M, $N(u)$, and K are sparse $N \times N$ matrices $(i, j = 1, 2, \ldots, N)$:

$$M_{ij} \equiv \int \phi_i \phi_j, \quad \text{the mass matrix,} \qquad [23]$$

$$N_{ij}(\mathbf{u}) \equiv \int \phi_i \mathbf{u}(x, t) \cdot \nabla \phi_j, \quad \text{the advection matrix,} \qquad [24]$$

$$K_{ij} \equiv K_{ij}^D + K_{ij}^B, \quad \text{where} \qquad [25]$$

$$K_{ij}^D \equiv \int \nabla \phi_i \cdot (\mathbf{K} \cdot \nabla \phi_j) \quad \text{is the diffusion matrix, and} \qquad [26]$$

$$K_{ij}^B \equiv \int_{\Gamma_N} H \phi_i \phi_j \quad \text{is the boundary matrix.} \qquad [27]$$

Finally, f is an N-vector that comprises the entire RHS of [17]; i.e., it incorporates the internal source term, the specified boundary heat flux, the remainder (specified portion) of the Robin BC, and it contains information that *couples* the Dirichlet BC to the rest of the problem (the term in curly brackets).

Remarks:

1. K_{ij}^B is zero for most i, j: it is only non-zero for those nodes (i) on Γ_N that 'see' node j (via the support of the basis function).
2. M is symmetric and positive-definite (SPD), which causes $\frac{\partial T^h}{\partial t}(x, t)$ to be a best (least squares) fit to the data: $\nabla \cdot (K \nabla T^h) + S - \mathbf{u} \cdot \nabla T^h$ and the NBC of [3].
3. K is always symmetric; it is SPD *unless* $\Gamma_D = 0$ *and* $H = 0$; i.e., K is symmetric but singular if Neumann data prevail on all of $\partial\Omega$.
4. $N(\mathbf{u})$ is unsymmetric and indefinite; it is also time-dependent when u is.

5. Variable coefficients—especially $\mathbf{u}(x, t)$ or, more commonly $\mathbf{u}(x)$—are usually interpolated via the basis functions before performing the integrations in [21]. We will have more to say on this important topic later.

3.2 *Divergence Form*

It is now a very simple matter to write the GFEM equations in flux-divergence form, a la [5]; just change the definition of the advection matrix—from [24] to

$$N_{ij}(\mathbf{u}) \equiv \int \phi_i \nabla \cdot [\phi_j \mathbf{u}(x, t)] \quad . \tag{28}$$

But since $\nabla \cdot (\phi_j \mathbf{u}) = \mathbf{u} \cdot \nabla \phi_j + \phi_j \nabla \cdot \mathbf{u}$ and the velocity field is allegedly divergence-free, one may properly query, 'Why bother with the divergence form since the results are the same?' While the detailed answer can only be provided after we have discussed the GFEM solution of the NS equations in the next section, it is appropriate to point out here that $\nabla \cdot \mathbf{u} \neq 0$ when $\mathbf{u}$ is obtained from the approximate (GFEM) solution of the NS equations, and that the velocity field that drives the scalar transport equation often, if not usually, is obtained from just these equations. So we must face the case where the velocity divergence is small but not zero. (The velocity is generally only discretely divergence-free.) Hence, we do *not* require that N_{ij} from [28] be the same as N_{ij} from [24]. The consequence of this is that only the use of [22] can assure global conservation of T in the GFEM solution, an assertion that we shall soon prove.

If, of course, the Robin BC were [14] rather than [3], then the GFEM would (or at least, should) be based on the weak form given by [15] rather than on that given by [13]. The resulting semi-discretized equations would differ from those in [17] in the following ways:

1. $\int \phi_i \mathbf{u} \cdot \nabla \phi_j$ is replaced by $-\int \nabla \phi_i \cdot (\phi_j \mathbf{u})$; i.e., in this case, $N_{ij}(\mathbf{u}) \equiv -\int \phi_j \mathbf{u} \cdot \nabla \phi_i$, vis-a-vis [24] and [22].
2. The same replacement must be made in the advection part of the Dirichlet BC term on the RHS; i.e., $\int \phi_i \mathbf{u} \cdot \nabla \hat{T}$ is replaced by $-\int \hat{T} \mathbf{u} \cdot \nabla \phi_i$.

3.3 *Conservation Laws*

We now attempt to mimic the analyses presented in Section 1.3, this time for the semi-discrete system of GFEM equations. But before we can do so conveniently, we must modify/generalize/augment our GFEM in the following way (see, e.g., Mizukami (1986), Gresho et al. (1987)): Rather than stating the problem a la [13], [16], [17], and [18]; i.e., find $T^h(x, t)$ in Ω and on Γ_N from

$$\int \left[\phi_i \left(\frac{\partial T^h}{\partial t} + \mathbf{u} \cdot \nabla T^h \right) + \nabla \phi_i \cdot (\mathbf{K} \cdot \nabla T^h) \right] + \int_{\Gamma_N} \phi_i H T^h$$

$$= \int \phi_i S + \int_{\Gamma_N} \phi_i (q + H\tilde{T}) \quad \text{for } i = 1, 2, \ldots N \quad , \qquad [29]$$

we generalize this weak formulation in three ways;

1. Replace $\mathbf{u} \cdot \nabla T^h$ by $\mathbf{u} \cdot \nabla T^h + \beta T^h \nabla \cdot \mathbf{u}$, where the scalar β will be defined below,
2. Introduce a *new* unknown, the heat flux into Ω through Γ_D (on which T^h is specified), as follows:

$$q_D^h = \sum_{j=N+1}^{N_T} q_{D_j} \phi_j \qquad [30]$$

where the $\{q_{D_j}\}$ are to be determined, and

3. Increase the size of the space of test functions, from those in Ω and on Γ_N to those in Ω and on $\Gamma = \Gamma_D + \Gamma_N = \partial\Omega$. The generalized weak formulation is then: Find $T^h(x,t)$ in Ω and on Γ_N *and* find q_D^h on Γ_D from

$$\int \left[\phi_i \left(\frac{\partial T^h}{\partial t} + u \cdot \nabla T^h + \beta T^h \nabla \cdot \mathbf{u} \right) + \nabla \phi_i \cdot (\mathbf{K} \cdot \nabla T^h) \right] + \int_{\Gamma_N} \phi_i H T^h =$$

$$\int \phi_i S + \int_{\Gamma_N} \phi_i (q + H\tilde{T}) + \int_{\Gamma_D} \phi_i q_D^h \quad \text{for } i = 1, 2, \ldots N_T \quad , \qquad [31]$$

where (still) T^h is given by [16] and $\hat{T}$ by [18], and we immediately point out that [31] naturally 'decomposes' into two sets of equations—the first set given by [29], which (as before) can be used to solve the N ODE's for $T^h(x,t)$ from the first N equations, and the second set by the last $N_T - N$ *algebraic* equations of [31], which can be used to solve for the $N_T - N$ values of q_{D_i} (with T^h *known*). The reason this decomposition occurs is that the first N equations are independent of the rest (the converse, of course, is not true). The reason we *introduced* this additional complexity is that it is a nice way to ensure (or to demonstrate) that the *total* GFEM solution (T^h in Ω and on Γ_N and q_D^h on Γ_D) can be made to satisfy a global energy balance, as we demonstrate below, after making the additional

Remarks:

1. q_D^h from [30] and [31] is called the *consistent* heat flux because, in addition to yielding global conservation (shown below), it is the *only* heat flux that permits reversibility; i.e., if the Dirichlet BC, $T = T_D$ on Γ_D were to be replaced by a Neumann BC, $\mathbf{n} \cdot (\mathbf{K} \cdot \nabla T) = q_D$ with q_D specified, on Γ_D, only q_D^h *as computed from* [30] would produce the *same* T^h as did the original problem.
2. The actual value of q_D^h is seen to depend on much more than just the normal component of $\mathbf{K} \cdot \nabla T$ on Γ_D—at least on a finite mesh; but in the limit of $h \to 0 (N_T \to \infty)$, all of the terms in the last $N_T - N$ equations of

[31] would vanish ($\rightarrow 0$) except $\int \nabla\phi_i \cdot (\mathbf{K}\cdot\nabla T^h)$ on the LHS and $\int_{\Gamma_D} \phi_i q_D^h$ on the RHS.

We will return to these (non-obvious) issues later; for now we just allege their veracity so that we can get on with the problem at hand—the derivation of global conservation laws. To this end, we now note the final reason for introducing the generalized problem of [31]: the sum of all N_T basis functions is unity,

$$\sum_{i=1}^{N_T} \phi_i(x) = 1.0 \quad \forall x \quad , \qquad [32]$$

a result that is *crucial* to the establishment of global conservation statements. (Note that $\sum_{i=1}^{N} \phi_i(x) \neq 1$ nearΓ_D.) The important property [32] leads easily to the following result when all N_T equations of [31] are summed:

$$\int \left(\frac{\partial T^h}{\partial t} + \mathbf{u}\cdot\nabla T^h + \beta T^h \nabla\cdot\mathbf{u}\right) + \int_{\Gamma_N} HT^h = \int S + \int_{\Gamma_N} (q + H\tilde{T}) + \int_{\Gamma_D} q_D^h \quad ,$$

which we rearrange to

$$\frac{d}{dt}\int T^h = \int S + \int_{\Gamma_N} H(\tilde{T} - T^h) + \int_{\Gamma_D} q_D^h + \int_{\Gamma_N} q - \int (\mathbf{u}\cdot\nabla T^h + \beta T^h \nabla\cdot\mathbf{u}), \quad [33]$$

and note that all but the last term on the RHS are in the desired form (cf. [11]); viz., the second term is the heat input from Newton's 'law of cooling,' the third is the heat flux into Γ_D that results from the specified temperature there, and the fourth term is the original applied heat flux on Γ_N. To finish, we use $\int \mathbf{u}\cdot\nabla T^h = \int \nabla\cdot(\mathbf{u}T^h) - \int T^h \nabla\cdot\mathbf{u} = \int_\Gamma \mathbf{n}\cdot\mathbf{u}T^h - \int T^h\nabla\cdot\mathbf{u}$ to obtain, finally,

$$\frac{d}{dt}\int T^h = \int S + \int_{\Gamma_N} [(q + H(\tilde{T} - T^h)] + \int_{\Gamma_D} q_D - \int_\Gamma \mathbf{n}\cdot\mathbf{u}T^h + (1-\beta)\int T^h \nabla\cdot\mathbf{u} \quad , \qquad [34]$$

which now properly mimics [11] *except* for the last term which should but does not vanish unless $\nabla\cdot\mathbf{u} = 0$ or $\beta = 1$. Since, as mentioned earlier, we often must solve the scalar transport equation using velocity fields that have small (hopefully) but indefinite divergence, we conclude that *for these cases it is necessary to set* $\beta = 1$ *if we wish to assure global conservation of our scalar field,* T^h.

But since $\nabla\cdot(\mathbf{u}T^h) = \mathbf{u}\cdot\nabla T^h + T^h\nabla\cdot\mathbf{u}$, we see from [31] that $\beta = 1$ is nothing but the flux-divergence form of the advective term ($\beta = 0$ being the advective form). Thus, while the advective form cannot assure (and will not attain) global conservation of energy when $\nabla\cdot\mathbf{u} \neq 0$, the divergence form can

and will. See Gresho et al. (1980) for some demonstrations of these facts in the case of an ideal fluid (zero diffusion coefficient for the AD equation and zero viscosity in the Boussinesq equations, whose computed velocity field drives the T-field.)

So at this point, it seems clear that the 'proper' GFEM for the scalar transport equation that is driven by a GFEM-computed velocity field (or any other for which $\nabla \cdot \mathbf{u} \neq 0$) should *not* use the simpler advective form; the (higher cost) flux divergence (conservation) form is clearly preferred. Or is it? What about *quadratic* conservation, $E^h \equiv \int (T^h)^2$, and the associated stability/boundedness that it supposedly guarantees? To answer this question, we must attempt to duplicate the steps that led to [12] for the continuum. We begin with [31] again, and with the observation that T^h is a linear combination of all (N_T) basis functions. Thus, we form (in principle) that *same* linear combination of the N_T equations of [31] to obtain (in fact, just replace ϕ_i by T^h)

$$\int \left[T^h \left(\frac{\partial T^h}{\partial t} + \mathbf{u} \cdot \nabla T^h + \beta T^h \nabla \cdot \mathbf{u} \right) + \nabla T^h \cdot (\mathbf{K} \cdot \nabla T^h) \right] + \int_{\Gamma_N} H (T^h)^2 =$$

$$\int T^h S + \int_{\Gamma_N} T^h (q + H \tilde{T}) + \int_{\Gamma_D} T^h q_D \quad , \qquad [35]$$

an equation that is also satisfied if [31] is (it is *implied* by [31]).

Next, recall BC [3] and rearrange the above equation to obtain

$$\frac{1}{2} \frac{d}{dt} \int (T^h)^2 = \int T^h S - \int \nabla T^h \cdot (\mathbf{K} \cdot \nabla T^h)$$

$$+ \int_{\Gamma_N} T^h \mathbf{n} \cdot (\mathbf{K} \cdot \nabla T^h) + \int_{\Gamma_D} T^h q_D$$

$$- \int T^h (\mathbf{u} \cdot \nabla T^h + \beta T^h \nabla \cdot \mathbf{u}) \quad , \qquad [36]$$

which is seen to agree, except for the advection term, with the continuum version, [12], once we generalize ($\kappa \rightarrow \mathbf{K}$) and then replace $\frac{1}{2} \int_\Gamma \underline{n} \cdot (\mathbf{K} \cdot \nabla T^2)$ by $\int_{\Gamma_N} T \mathbf{n} \cdot (\mathbf{K} \cdot \nabla T) + \int_{\Gamma_D} T q_D$ there.

The final step is to invoke the identity

$$\int T^h \mathbf{u} \cdot \nabla T^h = \frac{1}{2} \int \mathbf{u} \cdot \nabla (T^h)^2 = \frac{1}{2} \int_\Gamma \mathbf{n} \cdot \mathbf{u} (T^h)^2 - \frac{1}{2} \int (T^h)^2 \nabla \cdot \mathbf{u} \qquad [37]$$

to obtain

$$\frac{1}{2}\frac{d}{dt}\int (T^h)^2 = \int ST^h - \int \nabla T^h \cdot (\mathbf{K}\cdot\nabla T^h) + \int_{\Gamma_N} T^h \mathbf{n}\cdot(\mathbf{K}\cdot\nabla T^h) + \int_{\Gamma_D} T^h q_D$$
$$-\frac{1}{2}\int_{\Gamma} \mathbf{n}\cdot\mathbf{u}(T^h)^2 + \left(\frac{1}{2}-\beta\right)\int_{\Gamma}(T^h)^2\nabla\cdot\mathbf{u} \quad , \qquad [38]$$

in which agreement with [12], and *the assurance of global conservation of* $\int (T^h)^2$ *can now only be obtained (when* $\nabla\cdot\mathbf{u}\neq 0$*) by choosing* $\beta = \frac{1}{2}$*!* A dilemma, to be sure: for $\nabla\cdot\mathbf{u}\neq 0$, $\beta = 0$ conserves 'nothing,' $\beta = \frac{1}{2}$ conserves T^2 but not T, and $\beta = 1$ conserves T but not T^2. Which β (and associated form) should we choose, and why?

In the experiments performed by Gresho et al. (1980), these discouraging aspects of global conservation when $\nabla\cdot\mathbf{u}\neq 0$ were indeed verified. But they also reported that they would still not switch from the simpler ($\beta = 0$) and (slightly) less expensive advective form. Why is this? Besides computer costs and laziness, the reasons are basically these:

1. Any decent solution of the NS equations, in which $\int \psi\nabla\cdot\mathbf{u} = 0\ \forall\ \psi$ where ψ is a pressure test function, will generate only small (and of variable sign, generally) values of $\nabla\cdot\mathbf{u}$, so that the offending terms are probably always pretty small—although they definitely tend to cause instability in the absence of diffusion.
2. Any real (physical) solution—i.e., one with non-zero diffusion coefficients–should provide sufficient physical dissipation to control most potential instabilities related to the (indefinite) term $\int (T^h)^2\nabla\cdot\mathbf{u}$. Experience, both our own and that of many others, suggests that this is indeed true—usually. So, for now at least, we shall leave the issue of 'β-selection' open—except perhaps for the rare hyperbolic ($\mathbf{K} = \mathbf{0}$) case, wherein $\beta = 1/2$ is to be preferred to ensure stability of the ODE's.

3.4 *A Finite Difference Interpretation*

The GFEM equations are in 'weighted residual' form and it is of some interest to 'undo' the Galerkin weighting so that the equations can be more readily *interpreted* as finite difference equations, a procedure that can unfortunately lead also to *mis*-interpretations—as we shall demonstrate. In this section, we will convert [17], or [22], to an equivalent form that more readily permits such an interpretation.

Since the GFEM equations are formed by the process, 'multiply by each basis (test) function and integrate the result over the domain,' we now consider the effect of *dividing* the final results by the same test functions integrated over the domain; i.e., by $\int_\Omega \phi_i$. For reasons that will become more clear later, we define

$$M_{L_{ij}} \equiv \delta_{ij}\int \phi_i \quad , \qquad [39]$$

where δ_{ij} is the Kronecker delta. Whereas $\int \phi_i$, $i = 1, 2 \ldots N_T$ is, of course, a *vector*, this diagonal *matrix* representation is more convenient for our purposes, because often M_L is also the so-called lumped mass matrix, about which we will say more later. Noting that M_L^{-1} is trivial to compute, we multiply [22] by M_L^{-1} to get

$$A\dot{T} + M_L^{-1}[N(\mathbf{u}) + K]T = M_L^{-1} f \ , \qquad [40]$$

where

$$A \equiv M_L^{-1} M \qquad [41]$$

is, in fact, an *averaging* matrix; i.e.,

$$A_{ij} = \sum_{k=1}^{N_T} (\delta_{ik} \int \phi_k)^{-1} \int \phi_k \phi_j = \int \phi_i \phi_j \Big/ \int \phi_i \qquad [42]$$

is dimensionless and has the property that the sum over each row is unity (since $\sum_{j=1}^{N_T} \phi_j = 1.0$), so that each element of the vector $A\dot{T}$ represents a *particular weighted average* of the elements of the N_T-vector $\dot{T}$. In fact, the i-th row of $A\dot{T}$ is

$$\begin{aligned} A\dot{T}\Big|_i &= \sum_{j=1}^{N_T} \int \phi_i (\dot{T}_j \phi_j) \Big/ \int \phi_i \\ &= \int \phi_i \dot{T}^h(x,t) \Big/ \int \phi_i \quad , \end{aligned}$$

and it is now clear that we did indeed undo the Galerkin weighting—at least for the time derivative.

Next, note that $M_L^{-1} N(u)$ corresponds to (represents) a weak advection operator and $M_L^{-1} K$ corresponds to a weak Laplacian operator [at least when the heat transfer coefficient—see [27]—is zero. Thus,

$$\begin{aligned} M_L^{-1} N_A(u) T\Big|_i &= \sum_{j=1}^{N_T} \left(\int \phi_i \mathbf{u} \cdot \nabla \phi_j \right) T_j \Big/ \int \phi_i = \int \phi_i \mathbf{u} \cdot \nabla T^h \Big/ \int \phi_i \\ M_L^{-1} K^D T\Big|_i &= \sum_{j=1}^{N_T} \left(\int \nabla \phi_i \cdot \mathbf{K} \cdot \nabla \phi_j \right) T_j \Big/ \int \phi_i \\ &= \int \nabla \phi_i \cdot \mathbf{K} \cdot \nabla T^h \Big/ \int \phi_i \end{aligned}$$

Similarly, $M_L^{-1} f$—see [31]—is

$$M_L^{-1} f\Big|_i = \left[\int \phi_i S + \int_{\Gamma_N} \phi_i (q + H\tilde{T}) + \int_{\Gamma_D} \phi_i q_D^h \right] \Big/ \int \phi_i \quad ,$$

and the finite difference interpretation is now available; viz., *except for the time-derivative term, each of the other terms now corresponds (in some sense) to a point-wise approximation of the corresponding term in the original PDE*, because ϕ_i is the 'proper' piecewise polynomial of the FEM—the number of neighboring nodes (j) that couple with the node in question (i) being only a function of the support of the basis function, ϕ_i. Namely, a bilinear approximation in 2D will couple (generally, and away from $\partial\Omega$) eight neighbors to each node. The time derivative term is 'special' in the sense that the GFEM performs a weighted average of all the $\dot{T}_j$ in the neighborhood of node i to approximate $\partial T/\partial t$ at $x = x_i$—again the details depend on the support of the basis functions, but the key point to note is that only $A\dot{T}|_i$ is *not* a pointwise approximation in [40].

A final remark: If ϕ_i belongs to a node on Γ, the interpretation is somewhat trickier. It turns out that if $x_i \in \Gamma_N$ the nodal equation is actually (as for the original GFEM) an approximation to the Neumann (Robin) BC, [3], and will approach this exactly as the mesh is refined; i.e., all other terms will $\to 0$ as $N_T \to \infty$. Finally, if $x_i \in \Gamma_D$, a similar result is obtained, with only $M_L^{-1}K^D T$ and $\int \phi_i q_D^h / \int \phi_i$ remaining significant as $N_T \to \infty$, wherein they give $\mathbf{n} \cdot (\mathbf{K} \cdot \nabla T) = q_D$.

3.5 *A Control Volume FEM*

In this section we develop one form of a non-Galerkin weighted residual method that has been gaining in popularity—probably at the 'expense' of both the GFEM and FDM's. Called the control volume finite element method (CVFEM), it is a subdomain method of weighted residuals (see Finlayson and Scriven (1966)), and seems to have been spearheaded by, among others, Professor Suhas Patankar and colleagues at least for incompressible flow. While most of the papers we have seen involve more than a simple change in test function (such as directional upwinding and mass lumping), herein we develop and present the CVFEM as a fully legitimate (no cheating) *alternate finite element* technique, beginning with the appropriate weak formulation and introducing the CVFEM version of natural boundary conditions (NBC's). It's extension from the advection-diffusion equation to the Navier-Stokes equations will be considered in the next chapter.

Crucial to the CVFEM is the conservation (divergence) form of the PDE, [5] in this case, because it is *based* on 'conservation of T' *at control volume level*. The weak form of this AD equation begins, as usual, by multiplication by a test (weighting) function and integration over the domain. In this case, *the test function is piecewise constant*; it is unity over a particular subdomain (control volume or, in 2D, the only case we consider in detail, a control *area*) and zero over the rest of Ω. Calling the test function for subdomain 'i' ψ_i, we have

$$\int_{\Omega} \psi_i \left(\frac{\partial T}{\partial t} - S \right) + \int_{\Omega} \psi_i \nabla \cdot (\mathbf{u}T - \kappa \nabla T) = 0, \quad i = 1, 2, \ldots N \quad ,$$

where N is now the number of non-overlapping subdomains covering Ω. But owing to the nature of the test functions, the above equation is equivalent, via the divergence theorem, to

$$\int_{\Omega_i} \left(\frac{\partial T}{\partial t} - S \right) + \int_{\Gamma_i} \mathbf{n} \cdot (\mathbf{u}T - \kappa \nabla T) = 0, \quad i = 1, 2, \ldots N \quad , \qquad [43]$$

where Γ_i is the boundary of subdomain Ω_i. This simple set of equations—each representing an energy balance over one subdomain—is the starting point for the finite element discretization; it is the desired weak form.

What about boundary conditions? They are not nearly as apparent here as in the GFEM weak form, a la [15] and [17]. But the answer is simple: If (and only if) Γ_i includes a portion of the full domain boundary, Γ, 'special procedures' need to be introduced so that both Dirichlet and Robin/Neumann data are properly incorporated. These procedures are in fact little different from those already discussed and will later be presented in some detail. Suffice it to say here that the simple *looking* equation of the weak form is, in practice, only slightly simpler than that from GFEM.

The next step then is to approximate the solution in the finite element spirit: Expand the 'solution' in the same set of PP's used in the GFEM, a la [16] and [18], which converts [43] to the final control volume weak form:

$$\int_{\Omega_i} \sum_{j=1}^{N} \left(\dot{T}_j(t) - S_j \right) \phi_j + \int_{\Gamma_i} \sum_{j=1}^{N} \mathbf{n} \cdot (\mathbf{u}\phi_j - \kappa \nabla \phi_j) T_j =$$

$$- \int_{\Omega_i} \frac{\partial \widehat{T}}{\partial t} - \int_{\Gamma_i} \mathbf{n} \cdot (\mathbf{u}\widehat{T} - \kappa \nabla \widehat{T}), \qquad i = 1, 2, \ldots N \quad , \qquad [44]$$

where, for generality, we have also expanded the source term into the FEM basis functions, $S = \sum_{j=1}^{N} S_j \phi_j$, via interpolation. Note that the RHS is only non-zero at points where Ω_i and Γ_i involve Γ_D; i.e., the (now less simple looking) weak formulation now *does* account for the Dirichlet BC.

Note that j also ranges over 1 to N, where N is the number of nodes (in Ω and on Γ_N, as before) at which T_j is to be determined; there must be one subdomain for each unknown nodal temperature. Thus we have reduced the weak form of the continuous problem (obtained via $N \to \infty$ in [44]) to one of finite dimension. All that remains to be addressed prior to programming is the precise definition of Ω_i and Γ_i for $i = 1, 2, \ldots N$, in such a way that the test functions retain linear independence. We do this first in the 2D context in which the basis functions $\{\phi_j\}$ are bilinear. Consider the 4-patch of isoparametric

elements shown below, surrounding a generic node (i) in the domain:

The subdomain Ω_i (control 'volume') is that formed by joining the element centroids ($x_0 = \frac{1}{4}\sum_{j=1}^{4} x_j$ is the x-coordinate of a centroid, etc) with 8 straight line segments; each of which passes through the midside of the appropriate element.

It should now be apparent 'how to build' a CVFEM code: Each internal node's control volume—for the integration of 'volume quantities' ($\partial T/\partial t$ and S above)—is composed of pieces of neighboring elements (4 for a 4-patch, 2 for a 2-patch, etc.); each internal node's control volume *boundary*—for the integration of flux quantities (like $\mathbf{u}T$ and $\kappa\nabla T$)—is made up from two internal segments from *each* element that has something to contribute. It may also be apparent that this method is more 'localized' than GFEM, owing to the nature of the test functions; i.e, CVFEM will give more weight to node i relative to its neighbors than does GFEM. In local coordinates (ξ, η), these line segments are simply pieces of the coordinate lines themselves, e.g., $\xi = 0$ or $\eta = 0$, and this fact actually makes the 'boundary' calculations easier to perform since the general bilinear interpolation becomes simply linear on each of these segments. (The volume quantities are not simplified, however, and conventional element-level matrices (or the equivalent) need to be constructed.)

This is a sufficient exposition of the method at this point. Later we will actually present the resulting semi-discrete equations and compare them with those from the GFEM. Suffice it to say here that there are more similarities than differences.

How do the two schemes compare theoretically? Numerically? While we do not have many answers here, (indeed we have not (yet) programmed a CVFEM), some conjectures, opinions, and assertions are offered here:

1. Because of the particular conservation formulation, the CVFEM has the nice property that 'whatever exits one CV through its boundary surface enters the neighboring CV.' This physically appealling property—which also assures global conservation—accounts in part for the increasing popularity of the method. (The GFEM does not, in general, incorporate control volume or element-level conservation properties.)
2. There is no assurance that quadratic quantities are conserved (e.g., $\int T^2$) and in general they will not be; hence, boundedness of the solutions is not *a priori* guaranteed—as indeed it is not for the analogous ($\beta = 1$) GFEM.
3. For situations in which variational principles apply—which generally require $\mathbf{u} = \mathbf{0}$—the GFEM is guaranteed to produce the *most accurate* solution possible on a given mesh, at least when errors are measured in the appropriate norm. for example, for the elliptic (Poisson) problem PDE, $\nabla^2 T = -S$, the error from GFEM, $e \equiv T - T^h$, is a minimum in the 'energy' (or 'heat flux') norm, $\int \nabla e \cdot \nabla e$.
4. The dispersion (phase speed) error in the advection-dominated situation is smaller for GFEM than CVFEM, as we will demonstrate.

5. The mathematical theory is well-developed for the GFEM; while perhaps simpler, it is nearly absent for the CVFEM.

3.6 *Outflow Boundary Conditions (OBC's)*

We now address, for but one of many times in this text, the important issue of outflow (or, more generally, open) boundary conditions—a special case (usually) of the NBC's associated with the weak form.

In many simulations of interest in fluid mechanics, the fluid—and the 'load' that it carries, here the scalar T—flows *through* (i.e., both into and out of) the computational domain, a situation necessitated by the fact that the true (physical) domain of interest is (much) too large to even be *considered* in the numerical simulation. For an engineering example, consider a physical laboratory in which the experiment of interest is flow past an obstacle—a cylinder in a channel, or an airplane in a wind tunnel—and the flow is forced via a pump or fan/compressor; to attempt to model the entire closed loop would be expensive. For a geophysical application, consider the problem of trying to predict the air pollution from a (dirty) factory that is located (to make the problem more interesting) in mountainous terrain; to attempt to model the entire atmosphere of the earth would be expensive.

So we must consider inflow/outflow situations in which our computational domain is truncated and *some* BC's necessarily applied at these artificial/synthetic 'boundaries;' i.e., the PDE doesn't know that we are truncating the universe—all it knows is that BC's on Γ are *required* in order to 'solve for T.' The general goal, then, is to apply BC's at inflow ($\mathbf{n} \cdot \mathbf{u} < 0$) and—especially—at outflow ($\mathbf{n} \cdot \mathbf{u} > 0$) that are both mathematically legitimate and computationally useful. But what does 'useful' mean? While necessarily vague, it is basically this: Useful BC's are those that lead to good results in the 'smallest' truncated domain. But what does 'good' mean? What does 'smallest' mean? Good results are those that cause the solution in the 'subdomain of principle interest' to change little when the computational domain is made larger and that would agree well with those from the true (physical) domain. The smallest truncated domain is often (but not always) the largest domain that one can afford to model. Naturally, all of these issues are rather qualitative in nature—a necessary consequence of domain truncation. But it is a very real fact of life that many CFD simulations *must* deal with the open-boundary situation.

4. Streamline–Upwind Petrov–Galerkin and Least Squares Formulations

When using a Galerkin–finite element projection technique, spurious oscillations can appear in the solution if the discretization is too coarse to resolve the local physics. This is especially true in convective transport dominated

regimes. If these oscillations remain localized, then this is an obvious indication that some local mesh resolution is necessary; however, many times these oscillations invade the domain and pollute the solution. Mesh refinement allows resolution of the physics which is desirable but sometimes unaffordable or impossible. Another discretization scheme which addresses this issue directly is the Petrov–Galerkin finite element scheme which utilizes different basis and test functions; that is, the weak formulation is the same as [13] except the basis set used in representing the solution T is different from the test functions $w \in H_0^1$. The "symmetric" Galerkin–finite element scheme is replaced by the non–symmetric Petrov–Galerkin scheme. If the test functions are chosen to have more "weight" upstream than downstream as, for example, by Hughes and Brooks (1983), then the effect is to add artificial viscosity primarily in the streamline direction. Ideally the test functions are adapted locally relative to the amount of artificial viscosity needed. This technique is similar in effect and behavior to that of adding a tensor diffusivity, i.e., a tensor proportional to $\boldsymbol{uu}$ (Gresho et al. (1984)), to the transport equation for T. In the latter case, this term arises naturally in the case of transient solutions as a term to balance the time truncation of a forward Euler time integration step; the concept can be extended to a steady–state formulation at the expense of adding an arbitrary parameter, usually a function of the local grid Péclet number.

While there have been numerous other proposed stabilization schemes, here we focus only on two which are relatively new and carry over to the Navier–Stokes and Boussinesq problems. These are the Galerkin/least–squares and least–squares techniques, both implemented via the finite element technique.

The Galerkin/least squares finite element method is similar to the Galerkin finite element formulation except additional least squares terms are added. Thus, the formulation for a steady problem becomes:

Find $T_{\boldsymbol{x}}^h$ in $V^h \subset H_E^1$ such that

$$\int \left[\nabla w^h \cdot \left(K \cdot \nabla T^h - \boldsymbol{u} T^h\right)\right] + \int_{\Gamma_N} w^h H T^h$$

$$-\alpha^2 h \sum_e \int_e \left[\nabla \cdot \left(K \cdot \nabla w^h - \boldsymbol{u} w^h\right)\right] \left[\nabla \cdot \left(\boldsymbol{K} \nabla T^h - \boldsymbol{u} T^h\right)\right]$$

$$= \int w^h S + \int w(q + H\tilde{T}) \qquad \forall w^h \in V^h \quad , \qquad [45]$$

where the sum is over interior elements, α is a stability parameter, and h is a measure of element size. This formulation can be considered a generalization of the streamline–upwind Petrov–Galerkin technique since the least squares addition has a similar streamwise stabilizing effect. Another noteworthy point is that these additional terms involve the transport equation as a factor and hence the system is totally consistent in that an exact solution to the original continuum problem with mesh refinement. This formulation does require the introduction of a stabilizing factor α which must be specified but there are

guidelines. This tehcnique can also be extended to transient problems via either the standard spatial finite element discretization or the newer space–time finite element discretization techniques. (See, for example, Hughes, Franca, and Balestra (1986) or Tezduyar (1992).)

The least squares technique leads to the weak formulation: Find $T^h(\boldsymbol{x})$ in $V^h \subset H_E^1$ and $\boldsymbol{q}^h(\boldsymbol{x})$ in $\boldsymbol{V}^h \subset \boldsymbol{H}_E^1$ such that

$$\int [(\boldsymbol{u} \cdot \nabla T^h - \nabla \cdot \boldsymbol{q}^h)(\boldsymbol{u} \cdot \nabla w^h - \nabla \cdot \boldsymbol{r}^h) + (\boldsymbol{q}^h - K \nabla T^h)(\boldsymbol{r}^h - K \nabla w^h)] = 0 \qquad [46]$$

for $\forall w^h \in V^h$ and $\forall \boldsymbol{r}^h \in \boldsymbol{V}^h$. Here the number of unknowns has increased since the flux $\boldsymbol{q}^h$ must also be included; however, the discrete problem is now symmetric in contrast to the Galerkin/least–squares technique which leads to an unsymmetric system. The former can make the solution via iterative techniques simpler and more cost effective. Here, as in the previous method, the numerical smoothing is of the streamline–upwind form but here no stability parameter must be specified. (See B. Jiang (1991) and B. Jiang (1992) for recent work and a relevant bibliography.)

5. Conclusions

We have presented here a general resumé of some features of the continuum and discretized scalar transport equation associated with an incompressible flow. We have focused on some rather special topics in an effort to supplement the usual discussions of the topic.

6. References

[1] Finlayson, B. and Scriven, L.E., "The Method of Weighted Residuals—A Review," *Appl. Mech. Reviews* **19**, No. 9, 735 (1966).

[2] Gresho, P. M., Chan, S., Upson, C., and Lee, R., "A Modified Finite Element Method for Solving the Time–dependent Incompressible Navier–Stokes Equations," *Int'l. J. Num. Meth. in Fluids*, Part 1: Theory, **4**, 557–598; Part 2: Applications, **4**, 619 (1984).

[3] Gresho, P. M., Lee, R. L., Chan, S., and Sani R. L., "A Comparison of Several Conservation Forms for the Finite Element Formulations of the Incompressible Navier–Stokes or Boussinesq Equations," Proceedings of Third Int'l. Conf. on Finite Element Flow Problems, Banff, Canada (1980).

[4] Gresho, P. M. and Sani, R.L., "On Pressure Boundary Conditions for the Incompressible Navier–Stokes Equations," *Int'l. J. Num. Meth. Fluids* **7**, 1111–1145 (1987).

[5] Hughes, T. J. R. and Brooks, A., "A Theoretical Framework for Petrov–Galerkin Methods with Discontinuous Weighting Functions: Application to the Streamline Upwind Procedure," *Finite Elements in Fluids*, R. Gallagher, ed. **4**, Wiley (1983).

[6] Hughes, T. J. R., Franca, L.P., and Balestra, M., "A New Finite Element Formulation for Computational Fluid Dynamics: V.," *Comp. Meth. in Appl. Mech. and Engng.* **59**, 85–99 (1986).

[7] Jiang, B., "The L_1 Finite Element Method for Pure Convection Problems," NASA Tech. Memor. 103773 (1991).

[8] Jiang, B., "A Least Squares Finite Element Method for Incompressible Navier–Stokes Problems," *Int'l. J. Num. Meth. in Fluids* **14**, 843–859 (1992).

[9] Mitchell, A. R. and Wait, R., *The Finite Element Method in Partial Differential Equations*, John Wiley, London (1977).

[10] Mizukami, A., "A Stream Function–Vorticity Finite Element Formulation for Navier–Stokes Equations in Multi–Connected Domains," *Int'l. J. Num. Meth. Eng.* **19**, 1403–9 (1983).

[11] Swartz, B. and Wendroff, B., "Generalized Finite–Difference Schemes," *Math. of Comp.* **23**, No. 105, 37–49 (1969).

[12] Tezduyar, T. E., "Stabilized Finite Element Formulations for Incompressible Flow Computations," *Adv. Appl. Mech.* **28**, 1–43 (1992).