Convection dominated problems – finite element approximations to the convection–diffusion equation

2.1 Introduction

In this chapter we are concerned with the steady-state and transient solutions of equations of the type

\[
\frac{\partial \Phi}{\partial t} + \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_i} + Q = 0
\]  

(2.1)

where in general \( \Phi \) is the basic dependent, vector-valued variable, \( Q \) is a source or reaction term vector and the flux matrices \( F \) and \( G \) are such that

\[
F_i = F_i(\Phi)
\]  

(2.2a)

and in general

\[
G_i = G_i\left(\frac{\partial \Phi}{\partial x_j}\right)
\]  

(2.2b)

\[
Q = Q(x_i, \Phi)
\]

In the above, \( x_i \) and \( i \) refer in the indicial manner to cartesian coordinates and quantities associated with these.

Equations (2.1) and (2.2) are conservation laws arising from a balance of the quantity \( \Phi \) with its fluxes \( F \) and \( G \) entering a control volume. Such equations are typical of fluid mechanics which we have discussed in Chapter 1. As such equations may also arise in other physical situations this chapter is devoted to the general discussion of their approximate solution.

The simplest form of Eqs (2.1) and (2.2) is one in which \( \Phi \) is a scalar and the fluxes are linear functions. Thus

\[
\Phi = \phi \quad Q = Q(x_i)
\]

\[
F_i = F_i = U_i \phi \quad G_i = -k \frac{\partial \phi}{\partial x_i}
\]  

(2.3)
We now have in cartesian coordinates a scalar equation of the form

\[
\frac{\partial \phi}{\partial t} + \frac{\partial (U_i \phi)}{\partial x_i} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q = 0
\]

which will serve as the basic model for most of the present chapter.

In the above equation \( U \), in general is a known velocity field, \( \phi \) is a quantity being transported by this velocity in a convective manner or by diffusion action, where \( k \) is the diffusion coefficient.

In the above the term \( Q \) represents any external sources of the quantity \( \phi \) being admitted to the system and also the reaction loss or gain which itself is dependent on the concentration \( \phi \).

The equation can be rewritten in a slightly modified form in which the convective term has been differentiated as

\[
\frac{\partial \phi}{\partial t} + U_i \frac{\partial \phi}{\partial x_i} + \frac{\partial (U_i \phi)}{\partial x_i} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q = 0
\]

We will note that in the above form the problem is self-adjoint with the exception of a convective term which is underlined. The third term disappears if the flow itself is such that its divergence is zero, i.e. if

\[
\frac{\partial U_i}{\partial x_i} = 0 \quad \text{(summation over } i \text{ implied)}
\]

In what follows we shall discuss the scalar equation in much more detail as many of the finite element remedies are only applicable to such scalar problems and are not transferable to the vector forms. As in the CBS scheme, which we shall introduce in Chapter 3, the equations of fluid dynamics will be split so that only scalar transport occurs, where this treatment is sufficient.

From Eqs (2.5) and (2.6) we have

\[
\frac{\partial \phi}{\partial t} + U_i \frac{\partial \phi}{\partial x_i} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q = 0
\]

We have encountered this equation in Volume I [Eq. (3.11), Sec. 3.13 in connection with heat transport, and indeed the general equation (2.1) can be termed the transport equation with \( F \) standing for the convective and \( G \) for diffusive flux quantities.

With the variable \( \Phi \) (Eq. 2.1) being approximated in the usual way:

\[
\Phi \approx \tilde{\Phi} = \mathbf{N} \tilde{\Phi} = \sum N_k \tilde{\Phi}_k
\]

the problem could be presented following the usual (weighted residual) semi-discretization process as

\[
\mathbf{M} \dot{\tilde{\Phi}} + \mathbf{H} \tilde{\Phi} + \mathbf{f} = 0
\]

but now even with standard Galerkin (Bubnov) weighting the matrix \( \mathbf{H} \) will not be symmetric. However, this is a relatively minor computational problem compared
with inaccuracies and instabilities in the solution which follow the arbitrary use of this
weighting function.

This chapter will discuss the manner in which these difficulties can be overcome and
the approximation improved.

We shall in the main address the problem of solving Eq. (2.4), i.e. the scalar form,
and to simplify matters further we shall often start with the idealized one-dimensional
equation:

\[
\frac{\partial \phi}{\partial t} + U \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q = 0
\]  

(2.10)

The term \( \phi \partial U / \partial x \) has been removed here for simplicity. The above reduces in steady
state to an ordinary differential equation:

\[
U \frac{d\phi}{dx} - \frac{d}{dx} \left( k \frac{d\phi}{dx} \right) + Q = 0
\]

(2.11)
in which we shall often assume \( U, k \) and \( Q \) to be constant. The basic concepts will be
evident from the above which will later be extended to multidimensional problems,
still treating \( \phi \) as a scalar variable.

Indeed the methodology of dealing with the first space derivatives occurring in
differential equations governing a problem, which as shown in Chapter 3 of
Volume 1 lead to non-self-adjointness, opens the way for many new physical
situations.

The present chapter will be divided into three parts. Part I deals with steady-state
situations starting from Eq. (2.11), Part II with transient solutions starting from Eq.
(2.10) and Part III dealing with vector-valued functions. Although the scalar problem
will mainly be dealt with here in detail, the discussion of the procedures can indicate
the choice of optimal ones which will have much bearing on the solution of the general
case of Eq. (2.1). We shall only discuss briefly the extension of some procedures to the
vector case in Part III as such extensions are generally heuristic.

**Part I: Steady state**

### 2.2 The steady-state problem in one dimension

#### 2.2.1 Some preliminaries

We shall consider the discretization of Eq. (2.11) with

\[
\phi \approx \sum N_{i} \hat{\phi}_{i} = N\hat{\phi}
\]

(2.12)

where \( N_{i} \) are shape functions and \( \hat{\phi} \) represents a set of still unknown parameters.
Here we shall take these to be the nodal values of \( \phi \). This gives for a typical internal
node \( i \) the approximating equation

\[
K_{ij} \hat{\phi}_{j} + f_{i} = 0
\]

(2.13)
where

\[ K_{ij} = \int_0^L W_i U \frac{dN_j}{d\xi} d\xi + \int_0^L W_i \frac{dW_j}{d\xi} k \frac{dN_j}{d\xi} d\xi \] (2.14)

\[ f_i = \int_0^L W_i Q d\xi \]

and the domain of the problem is \(0 < x < L\).

For linear shape functions, Galerkin weighting \(W_i = N_i\) and elements of equal size \(h\), we have for constant values of \(U\), \(k\) and \(Q\) (Fig. 2.1) a typical assembled equation

\[-(Pe - 1)\tilde{\phi}_{i-1} + 2\tilde{\phi}_i + (Pe - 1)\tilde{\phi}_{i+1} + \frac{Qh^2}{k} = 0\] (2.15)

where

\[ Pe = \frac{Uh}{2k} \] (2.16)

is the element Peclet number. The above is, incidentally, identical to the usual central finite difference approximation obtained by putting

\[ \frac{d\phi}{dx} \approx \frac{\tilde{\phi}_{i+1} - \tilde{\phi}_{i-1}}{2h} \] (2.17a)

and

\[ \frac{d^2\phi}{dx^2} \approx \frac{\tilde{\phi}_{i+1} - 2\tilde{\phi}_i + \tilde{\phi}_{i-1}}{h^2} \] (2.17b)

The algebraic equations are obviously non-symmetric and in addition their accuracy deteriorates as the parameter \(Pe\) increases. Indeed as \(Pe \to \infty\), i.e. when only convective terms are of importance, the solution is purely oscillatory and bears no relation to the underlying problem, as shown in the simple example where \(Q\) is zero of Fig. 2.2 with curves labelled \(\alpha = 0\). (Indeed the solution for this problem is now only possible for an odd number of elements and not for even.)

Of course the above is partly a problem of boundary conditions. When diffusion is omitted only a single boundary condition can be imposed and when the diffusion is small we note that the downstream boundary condition \((\phi = 1)\) is felt in only a very small region of a boundary layer evident from the exact solution

\[ \phi = \frac{1 - e^{UL/k}}{1 - e^{UL/\kappa}} \] (2.18)
Motivated by the fact that the propagation of information is in the direction of velocity $U$, the finite difference practitioners were the first to overcome the bad approximation problem by using one-sided finite differences for approximating the first derivative. Thus in place of Eq. (2.17a) and with positive $U$, the approximation was put as

$$\frac{d\phi}{dx} \approx \frac{\tilde{\phi}_i - \tilde{\phi}_{i-1}}{h}$$  \hfill (2.19)
changing the central finite difference form of the approximation to the governing
equation as given by Eq. (2.15) to

\[ (-2Pe - 1)\phi_{i-1} + (2 + 2Pe)\phi_{i} - \phi_{i+1} + \frac{Oh^2}{k} = 0 \] (2.20)

With this upwind difference approximation, realistic (though not always accurate)
solutions can be obtained through the whole range of Peclet numbers of the example
of Fig. 2.2 as shown there by curves labelled \( \alpha = 1 \). However, now exact nodal solutions are only obtained for pure convection (\( Pe = \infty \)), as shown in Fig. 2.2, in a similar
way as the Galerkin finite element form gives exact nodal answers for pure diffusion.

How can such upwind differencing be introduced into the finite element scheme and
generalized to more complex situations? This is the problem that we shall now
address, and indeed will show that again, as in self-adjoint equations, the finite
element solution can result in exact nodal values for the one-dimensional approxima-
tion for all Peclet numbers.

2.2.2 Petrov-Galerkin methods for upwinding in one dimension

The first possibility is that of the use of a Petrov-Galerkin type of weighting in which
\( W_i \neq N_i \). Such weightings were first suggested by Zienkiewicz et al.\(^6\) in 1975 and
used by Christie et al.\(^7\). In particular, again for elements with linear shape functions
\( N_i \), shown in Fig. 2.1, we shall take, as shown in Fig. 2.3, weighting functions
constructed so that

\[ W_i = N_i + \alpha W^*_i \] (2.21)

where \( W^*_i \) is such that

\[ \int_{\Omega_i} W^*_i \, dx = \pm \frac{h}{2} \] (2.22)

\[ N_i \]
\[ h \]
\[ i \]
\[ W_i^* \]
\[ \text{or} \]

Fig. 2.3 Petrov-Galerkin weight function \( W_i = N_i + \alpha W^*_i \). Continuous and discontinuous definitions.
the sign depending on whether $U$ is a velocity directed towards or away from the node.

Various forms of $W_i'$ are possible, but the most convenient is the following simple definition which is, of course, a discontinuous function (see the note at the end of this section):

$$\alpha W_i' = \alpha \frac{h}{2} \frac{dN_i}{dx} (\text{sign } U)$$

(2.23)

With the above weighting functions the approximation equivalent to that of Eq. (2.15) becomes

$$[-Pe(\alpha + 1) - 1] \tilde{\phi}_{i-1} + [2 + 2\alpha(Pe)] \tilde{\phi}_i + [-Pe(\alpha - 1) - 1] \tilde{\phi}_{i+1} + \frac{Qh^2}{k} = 0$$

(2.24)

Immediately we see that with $\alpha = 0$ the standard Galerkin approximation is recovered [Eq. (2.19)] and that with $\alpha = 1$ the full upwinded discrete equation (2.20) is available, each giving exact nodal values for purely diffusive or purely convective cases respectively.

Now if the value of $\alpha$ is chosen as

$$|\alpha| = \alpha_{\text{opt}} = \coth |Pe| - \frac{1}{|Pe|}$$

(2.25)

then exact nodal values will be given for all values of $Pe$. The proof of this is given in reference 7 for the present, one-dimensional, case where it is also shown that if

$$|\alpha| > \alpha_{\text{crit}} = 1 - \frac{1}{|Pe|}$$

(2.26)

oscillatory solutions will never arise. The results of Fig. 2.2 show indeed that with $\alpha = 0$, i.e. the Galerkin procedure, oscillations will occur when

$$|Pe| > 1$$

(2.27)

Figure 2.4 shows the variation of $\alpha_{\text{opt}}$ and $\alpha_{\text{crit}}$ with $Pe$.

Although the proof of optimality for the upwinding parameter was given for the case of constant coefficients and constant size elements, nodally exact values will also be given if $\alpha = \alpha_{\text{opt}}$ is chosen for each element individually. We show some typical solutions in Fig. 2.5 for a variable source term $Q = Q(x)$, convection coefficients $U = U(x)$ and element sizes. Each of these is compared with a standard Galerkin solution, showing that even when the latter does not result in oscillations the accuracy is improved. Of course in the above examples the Petrov–Galerkin weighting must be applied to all terms of the equation. When this is not done (as in simple finite difference upwinding) totally wrong results will be obtained, as shown in the finite difference results of Fig. 2.6, which was used in reference 11 to discredit upwinding methods. The effect of $\alpha$ on the source term is not apparent in Eq. (2.24) where $Q$ is constant in the whole domain, but its influence is strong when $Q = Q(x)$.

**Continuity requirements for weighting functions**

The weighting function $W_i$ (or $W_i'$) introduced in Fig. 2.3 can of course be discontinuous as far as the contributions to the convective terms are concerned [see Eq. (2.14)].

* Subsequently $Pe$ is interpreted as an absolute value.
Convection dominated problems

![Critical (stable) and optimal values of the 'upwind' parameter \( \alpha \) for different values of \( Pe = \frac{Uh}{2k} \).](image)

\[ \alpha_{\text{opt}} = \coth Pe - 1/Pe \text{ (optimal)} \]

\[ \alpha_{\text{crit}} = 1 - 1/Pe \text{ (critical)} \]

Fig. 2.4 Critical (stable) and optimal values of the 'upwind' parameter \( \alpha \) for different values of \( Pe = \frac{Uh}{2k} \).

Clearly no difficulty arises at the discontinuity in the evaluation of the above integrals. However, when evaluating the diffusion term, we generally introduce integration by parts and evaluate such terms as

\[ \int_0^L W_i \frac{dF}{dx} \, dx \quad \text{or} \quad \int_0^L W_i U \frac{dN_i}{dx} \, dx \]

in place of the form

\[ \int_0^L W_i \frac{d}{dx} \left( k \frac{dN_i}{dx} \right) \, dx \]

Here a local infinity will occur with discontinuous \( W_i \). To avoid this difficulty we modify the discontinuity of the \( W_i^* \) part of the weighting function to occur within the element and thus avoid the discontinuity at the node in the manner shown in Fig. 2.3. Now direct integration can be used, showing in the present case zero contributions to the diffusion term, as indeed happens with \( C_0 \) continuous functions for \( W_i^* \) used in earlier references.

### 2.2.3 Balancing diffusion in one dimension

The comparison of the nodal equations (2.15) and (2.16) obtained on a uniform mesh and for a constant \( Q \) shows that the effect of the Petrov–Galerkin procedure is equivalent to the use of a standard Galerkin process with the addition of a diffusion

\[ k_h = \frac{1}{2} \alpha Uh \quad (2.28) \]

to the original differential equation (2.11).
The reader can easily verify that with this substituted into the original equation, thus writing now in place of Eq. (2.11)

\[ U \frac{d\phi}{dx} - \frac{d}{dx} \left[ (k + k_b) \frac{d\phi}{dx} \right] + Q = 0 \]  

we obtain an identical expression to that of Eq. (2.24) providing \( Q \) is constant and a standard Galerkin procedure is used.
Convection dominated problems

Fig. 2.6 A one-dimensional pure convective problem \((k = 0)\) with a variable source term \(Q\) and constant \(U\). Petrov–Galerkin procedure results in an exact solution but simple finite difference upwinding gives substantial error.

Such balancing diffusion is easier to implement than Petrov–Galerkin weighting, particularly in two or three dimensions, and has some physical merit in the interpretation of the Petrov–Galerkin methods. However, it does not provide the modification of source terms required, and for instance in the example of Fig. 2.6 will give erroneous results identical with a simple finite difference, upwind, approximation.

The concept of artificial diffusion introduced frequently in finite difference models suffers of course from the same drawbacks and in addition cannot be logically justified.

It is of interest to observe that a central difference approximation, when applied to the original equations (or the use of the standard Galerkin process), fails by introducing a negative diffusion into the equations. This ‘negative’ diffusion is countered by the present, balancing, one.

2.2.4 A variational principle in one dimension

Equation (2.11), which we are here considering, is not self-adjoint and hence is not directly derivable from any variational principle. However, it was shown by Guymon et al.\(^1\) that it is a simple matter to derive a variational principle (or ensure self-adjointness which is equivalent) if the operator is premultiplied by a suitable function \(p\). Thus we write a weak form of Eq. (2.11) as

\[
\int_0^L wp \left[ U \frac{d\phi}{dx} - \frac{d}{dx} \left( k \frac{d\phi}{dx} \right) + Q \right] \, dx = 0 \tag{2.30}
\]
where \( p = p(x) \) is as yet undetermined. This gives, on integration by parts,

\[
\int_0^L \left[ W \frac{d}{dx} \left( pU + k \frac{dp}{dx} \right) + \frac{dW}{dx} (kp) \frac{d}{dx} + WpQ \right] dx + \left. \left( Wpk \frac{d}{dx} \right) \right|_0^L = 0
\]

(2.31)

Immediately we see that the operator can be made self-adjoint and a symmetric approximation achieved if the first term in square brackets is made zero (see also Chapter 3 of Volume 1, Sec. 3.11.2, for this derivation). This requires that \( p \) be chosen so that

\[
pU + k \frac{dp}{dx} = 0
\]

or that

\[
p = \text{constant} \times e^{-U/k} = \text{constant} \times e^{-2(Pe)x/k}
\]

(2.32b)

For such a form corresponding to the existence of a variational principle the ‘best’ approximation is that of the Galerkin method with

\[
W = N_i \quad \phi = \sum N_i \phi_j
\]

(2.33)

Indeed, as shown in Volume 1, such a formulation will, in one dimension, yield answers exact at nodes (see Appendix H of Volume 1). It must therefore be equivalent to that obtained earlier by weighting in the Petrov–Galerkin manner. Inserting the approximation of Eq. (2.33) into Eq. (2.31), with Eqs (2.32) defining \( p \) using an origin at \( x = x_i \), we have for the \( i \)th equation of the uniform mesh

\[
\int_{-h}^{h} \left[ \frac{dN_i}{dx} (k e^{-2(Pe)x/k}) \frac{dN_j}{dx} \phi_j + N_i \frac{d}{dx} (e^{-2(Pe)x/k} Q) \right] dx = 0
\]

(2.34)

with \( j = i - 1, i, i + 1 \). This gives, after some algebra, a typical nodal equation:

\[
(1 - e^{-2(Pe)\Delta x}) \tilde{\phi}_{i+1} + (e^{-2(Pe)} - e^{-2(Pe)\Delta x}) \tilde{\phi}_i - (1 - e^{-2(Pe)}) \tilde{\phi}_{i+1}
\]

\[
- \frac{Qh^2}{2(Pe)k} (e^{Pc} - e^{-Pc})^2 = 0
\]

(2.35)

which can be shown to be identical with the expression (2.24) into which \( \alpha = \alpha_{\text{opt}} \) given by Eq. (2.25) has been inserted.

Here we have a somewhat more convincing proof of the optimality of the proposed Petrov–Galerkin weighting. However, serious drawbacks exist. The numerical evaluation of the integrals is difficult and the equation system, though symmetric overall, is not well conditioned if \( p \) is taken as a continuous function of \( x \) through the whole domain. The second point is easily overcome by taking \( p \) to be discontinuously defined, for instance taking the origin of \( x \) at point \( i \) for all assemblies as we did in deriving Eq. (2.35). This is permissible by arguments given in Sec. 2.2 and is equivalent to scaling the full equation system row by row. Now of course the total equation system ceases to be symmetric.

The numerical integration difficulties disappear, of course, if the simple weighting functions previously derived are used. However, the proof of equivalence is important as the problem of determining the optimal weighting is no longer necessary.
2.2.5 Galerkin least square approximation (GLS) in one dimension

In the preceding sections we have shown that several, apparently different, approaches have resulted in identical (or almost identical) approximations. Here yet another procedure is presented which again will produce similar results. In this a combination of the standard Galerkin and least square approximations is made.\textsuperscript{15,16}

If Eq. (2.11) is rewritten as

\[ L\phi + Q = 0 \quad \phi \approx \tilde{\phi} = N\tilde{\phi} \]  

(2.36a)

with

\[ L = U \frac{d}{dx} - \frac{d}{dx} \left( k \frac{d}{dx} \right) \]  

(2.36b)

the standard Galerkin approximation gives for the kth equation

\[ \int_0^L N_k L(N\tilde{\phi}) \, dx + \int_0^L N_k Q \, dx = 0 \]  

(2.37)

with boundary conditions omitted for clarity.

Similarly, a least square residual minimization (see Chapter 3 of Volume 1, Sec. 3.14.2) results in

\[ R = L\tilde{\phi} + Q \quad \text{and} \quad \frac{1}{2} \frac{d}{d\tilde{\phi}_k} \int_0^L R^2 \, dx = \int_0^L \frac{d(L\tilde{\phi})}{d\tilde{\phi}_k} (L\tilde{\phi} + Q) \, dx = 0 \]  

(2.38)

or

\[ \int_0^L \left( U \frac{dN_k}{dx} - \frac{d}{dx} \left( k \frac{d}{dx} \right) N_k \right) (L\tilde{\phi} + Q) = 0 \]  

(2.39)

If the final approximation is written as a linear combination of Eqs (2.37) and (2.39), we have

\[ \int_0^L \left( N_k + \lambda U \frac{dN_k}{dx} - \lambda \frac{d}{dx} \left( k \frac{d}{dx} \right) N_k \right) (L\tilde{\phi} + Q) \, dx = 0 \]  

(2.40)

This is of course, the same as the Petrov–Galerkin approximation with an undetermined parameter \( \lambda \). If the second-order term is omitted (as could be done assuming linear \( N_k \) and a curtailment as in Fig. 2.3) and further if we take

\[ \lambda = \frac{|\alpha| h}{2|U|} \]  

(2.41)

the approximation is identical to that of the Petrov–Galerkin method with the weighting given by Eqs (2.21) and (2.22).

Once again we see that a Petrov–Galerkin form written as

\[ \int_0^L \left( N_k + \frac{|\alpha| h}{2 |U|} \frac{dN_k}{dx} \right) \left( U \frac{d\tilde{\phi}}{dx} - \frac{d}{dx} \left( k \frac{d\tilde{\phi}}{dx} \right) + Q \right) \, dx = 0 \]  

(2.42)
The steady-state problem in one dimension

is a result that follows from diverse approaches, though only the variational form of Sec. 2.2.4 explicitly determines the value of $\alpha$ that should optimally be used. In all the other derivations this value is determined by an \textit{a posteriori} analysis.

### 2.2.6 The finite increment calculus (FIC) for stabilizing the convective–diffusion equation in one dimension

As mentioned in the previous sections, there are many procedures which give identical results to those of the Petrov–Galerkin approximations. We shall also find a number of such procedures arising directly from the transient formulations discussed in Part II of this chapter; however there is one further simple process which can be applied directly to the steady-state equation. This process was suggested by Oñate in 1998 and we shall describe its basis below.

We shall start at the stage where the conservation equation of the type given by Eq. (2.5) is derived. Now instead of considering an infinitesimal control volume of length \('dx'\) which is going to zero, we shall consider a finite length $\delta$. Expanding to one higher order by Taylor series (backwards), we obtain instead of Eq. (2.11)

\[
-U \frac{d\phi}{dx} + \frac{d}{dx} \left( k \frac{d\phi}{dx} \right) + Q - \frac{\delta}{2} \left[ -U \frac{d\phi}{dx} + \frac{d}{dx} \left( k \frac{d\phi}{dx} \right) + Q \right] = 0 \tag{2.43}
\]

with $\delta$ being the finite distance which is smaller than or equal to that of the element size $h$. Rearranging terms and substituting $\delta = \alpha h$ we have

\[
U \frac{d\phi}{dx} - \frac{d}{dx} \left[ k \left( \frac{\alpha h U}{2} \right) \frac{d\phi}{dx} \right] + Q - \frac{\delta}{2} \frac{dQ}{dx} = 0 \tag{2.44}
\]

In the above equation we have omitted the higher order expansion for the diffusion term as in the previous section.

From the last equation we see immediately that a stabilizing term has been recovered and the additional term $\alpha h U/2$ is identical to that of the Petrov–Galerkin form (Eq. 2.28).

There is no need to proceed further and we see how simply the finite increment procedure has again yielded exactly the same result by simply modifying the conservation differential equations. In reference 17 it is shown further that arguments can be brought to determine $\alpha$ as being precisely the optimal value we have already obtained by studying the Petrov–Galerkin method.

### 2.2.7 Higher-order approximations

The derivation of accurate Petrov–Galerkin procedures for the convective diffusion equation is of course possible for any order of finite element expansion. In reference 9 Heinrich and Zienkiewicz show how the procedure of studying exact discrete solutions can yield optimal upwind parameters for quadratic shape functions. However, here the simplest approach involves the procedures of Sec. 2.2.4, which
Convection dominated problems

are available of course for any element expansion and, as shown before, will always give an optimal approximation.

We thus recommend the reader to pursue the example discussed in that section and, by extending Eq. (2.34), to arrive at an appropriate equation linking the two quadratic elements of Fig. 2.7.

For practical purposes for such elements it is possible to extend the Petrov–Galerkin weighting of the type given in Eqs (2.21) to (2.23) now using

\[ \alpha_{\text{opt}} = \coth Pe \frac{1}{Pe} \quad \text{and} \quad \alpha W_i^* = \frac{h}{4} \frac{dN_i}{dx} (\text{sign } U) \]

This procedure, though not as exact as that for linear elements, is very effective and has been used with success for solution of Navier–Stokes equations.\(^{18}\)

In recent years, the subject of optimal upwinding for higher-order approximations has been studied further and several references show the developments.\(^{19,20}\) It is of interest to remark that the procedure known as the \textit{discontinuous Galerkin} method avoids most of the difficulties of dealing with higher-order approximations. This procedure was recently applied to convection–diffusion problems and indeed to other problems of fluid mechanics by Oden and coworkers.\(^{21-23}\) As the methodology is not available for lowest polynomial order of unity we do not include the details of the method here but for completeness we show its derivation in Appendix B.

\section*{2.3 The steady-state problem in two (or three) dimensions}

\subsection*{2.3.1 General remarks}

It is clear that the application of standard Galerkin discretization to the steady-state scalar convection–diffusion equation in several space dimensions is similar to the problem discussed previously in one dimension and will again yield unsatisfactory answers with high oscillation for local Peclet numbers greater than unity.

The equation now considered is the steady-state version of Eq. (2.7), i.e.

\[ U_x \frac{\partial \phi}{\partial x} + U_y \frac{\partial \phi}{\partial y} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) - \frac{\partial}{\partial y} \left( k \frac{\partial \phi}{\partial y} \right) + Q = 0 \]
The steady-state problem in two (or three) dimensions

in two dimensions or more generally using indicial notation

\[ U_i \frac{\partial \phi}{\partial x_i} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q = 0 \quad (2.46b) \]

in both two and three dimensions.

Obviously the problem is now of greater practical interest than the one-dimensional case so far discussed, and a satisfactory solution is important. Again, all of the possible approaches we have discussed are applicable.

2.3.2 Streamline (Upwind) Petrov–Galerkin weighting (SUPG)

The most obvious procedure is to use again some form of Petrov–Galerkin method of the type introduced in Sec. 2.2.2 and Eqs (2.21) to (2.25), seeking optimality of \( \alpha \) in some heuristic manner. Restricting attention here to two-dimensions, we note immediately that the Peclet parameter

\[ Pe = \frac{U h}{2k} \quad U = \left\{ \begin{array}{c} U_1 \\ U_2 \end{array} \right\} \quad (2.47) \]

is now a ‘vector’ quantity and hence that upwinding needs to be ‘directional’.

The first reasonably satisfactory attempt to do this consisted of determining the optimal Petrov–Galerkin formulation using \( \alpha W^* \) based on components of \( U \) associated to the sides of elements and of obtaining the final weight functions by a blending procedure.\(^8,9\)

A better method was soon realized when the analogy between balancing diffusion and upwinding was established, as shown in Sec. 2.2.3. In two (or three) dimensions the convection is only active in the direction of the resultant element velocity \( U \), and hence the corrective, or balancing, diffusion introduced by upwinding should be anisotropic with a coefficient different from zero only in the direction of the velocity resultant. This innovation introduced simultaneously by Hughes and Brooks\(^{24,25} \) and Kelly et al.\(^{10} \) can be readily accomplished by taking the individual weighting functions as

\[ W_k = N_k + \alpha W^*_k = N_k + \frac{\alpha h}{2} \frac{U_1(\partial N_k/\partial x_1) + U_2(\partial N_k/\partial x_2)}{|U|} \]

\[ = N_k + \frac{\alpha h}{2} \frac{U_i}{|U|} \frac{\partial N_k}{\partial x_i} \quad (2.48) \]

where \( \alpha \) is determined for each element by the previously found expression (2.22) written as follows:

\[ \alpha = \alpha_{\text{opt}} = \coth Pe - \frac{1}{Pe} \quad (2.49) \]

with

\[ Pe = \frac{|U|h}{2k} \quad (2.50a) \]

and

\[ |U| = (U_1^2 + U_2^2)^{1/2} \quad \text{or} \quad \sqrt{U_i U_j} \quad (2.50b) \]
The above expressions presuppose that the velocity components $U_1$ and $U_2$ in a particular element are substantially constant and that the element size $h$ can be reasonably defined.

Figure 2.8 shows an assembly of linear triangles and bilinear quadrilaterals for each of which the mean resultant velocity $\mathbf{U}$ is indicated. Determination of the element size $h$ to use in expression (2.50) is of course somewhat arbitrary. In Fig. 2.8 we show it simply as the maximum size in the direction of the velocity vector.

The form of Eq. (2.48) is such that the ‘non-standard’ weighting $W^*$ has a zero effect in the direction in which the velocity component is zero. Thus the balancing diffusion is only introduced in the direction of the resultant velocity (convective) vector $\mathbf{U}$. This can be verified if Eq. (2.46) is written in tensorial (indicial) notation as

$$U_i \frac{\partial \phi}{\partial x_i} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q = 0$$

In the discretized form the ‘balancing diffusion’ term [obtained from weighting the first term of the above with $W$ of Eq. (2.48)] becomes

$$\int_{\Omega} \frac{\partial N}{\partial x_i} \tilde{k}_{ij} \frac{\partial N}{\partial x_j} \, d\Omega$$

with

$$\tilde{k}_{ij} = \frac{\alpha U_i U_j}{|\mathbf{U}|} \frac{h}{2}$$

This indicates a highly anisotropic diffusion with zero coefficients normal to the convective velocity vector directions. It is therefore named the streamline balancing diffusion$^{10,24,25}$ or streamline upwind Petrov–Galerkin process.

The streamline diffusion should allow discontinuities in the direction normal to the streamline to travel without appreciable distortion. However, with the standard finite element approximations actual discontinuities cannot be modelled and in practice some oscillations may develop when the function exhibits ‘shock like’ behaviour. For this reason it is necessary to add some smoothing diffusion in the direction normal to the streamlines and some investigators make appropriate suggestions.$^{26–29}$
The mathematical validity of the procedures introduced in this section has been established by Johnson et al.\textsuperscript{30} for $\alpha = 1$, showing convergence improvement over the standard Galerkin process. However, the proof does not include any optimality in the selection of $\alpha$ values as shown by Eq. (2.49).

Figure 2.9 shows a typical solution of Eq. (2.46), indicating the very small amount of 'cross-wind diffusion', i.e. allowing discontinuities to propagate in the direction of flow without substantial smearing.\textsuperscript{31}

A more convincing 'optimality' can be achieved by applying the exponential modifying function, making the problem self-adjoint. This of course follows precisely the procedures of Sec. 2.2.4 and is easily accomplished if the velocities are constant in the element assembly domain. If velocities vary from element to element, again the exponential functions

$$ p = e^{-Ux'/k} $$

with $x'$ orientated in the velocity direction in each element can be taken. This appears to have been first implemented by Sampaio\textsuperscript{31} but problems regarding the origin of
coordinates, etc., have once again to be addressed. However, the results are essentially similar here to those achieved by Petrov–Galerkin procedures.

### 2.3.3 Galerkin least squares (GLS) and finite increment calculus (FIC) in multidimensional problems

It is of interest to observe that the somewhat intuitive approach to the generation of the 'streamline' Petrov–Galerkin weight functions of Eq. (2.48) can be avoided if the least square Galerkin procedures of Sec. 2.2.4 are extended to deal with the multi-dimensional equation. Simple extension of the reasoning given in Eqs (2.36) to (2.42) will immediately yield the weighting of Eq. (2.48).

Extension of the GLS to two or three dimensions gives (again using indicial notation)

\[
\int_{\Omega} \left( N_k + \lambda U_i \frac{\partial N_k}{\partial x_i} \right) \left( U_j \frac{\partial \phi}{\partial x_j} - \frac{\partial}{\partial x_j} \left( k \frac{\partial \phi}{\partial x_j} \right) + Q \right) d\Omega = 0 \tag{2.53}
\]

In the above equation, higher-order terms are omitted for the sake of simplicity. As in one dimension (Eq. 2.40) we have an additional weighting term. Now assuming

\[
\lambda = \frac{\alpha h}{2 |U|} \tag{2.54}
\]

we obtain an identical stabilizing term to that of the streamline Petrov–Galerkin procedure (Eq. 2.51).

The finite increment calculus method in multidimensions can be written as

\[
U_j \frac{\partial \phi}{\partial x_j} - \frac{\partial}{\partial x_j} \left( k \frac{\partial \phi}{\partial x_j} \right) + Q - \delta_i \frac{\partial}{\partial x_i} \left[ U_i \frac{\partial \phi}{\partial x_j} - \frac{\partial}{\partial x_j} \left( k \frac{\partial \phi}{\partial x_j} \right) + Q \right] = 0 \tag{2.55}
\]

Note that the value of \( \delta_i \) is now dependent on the coordinate directions. To obtain streamline-oriented stabilization, we simply assume that \( \delta_i \) is the projection oriented along the streamlines. Now

\[
\delta_i = \frac{\delta U_i}{|U|} \tag{2.56}
\]

with \( \delta = \alpha h \). Again, omitting the higher order terms in \( k \), the streamline Petrov–Galerkin form of stabilization is obtained (Eq. 2.51). The reader can verify that both the GLS and FIC produce the correct weighting for the source term \( Q \) as of course is required by the Petrov–Galerkin method.

### 2.4 Steady state – concluding remarks

In Secs 2.2 and 2.3 we presented several currently used procedures for dealing with the steady-state convection–diffusion equation with a scalar variable. All of these translate essentially to the use of streamline Petrov–Galerkin discretization, though
of course the modification of the basic equations to a self-adjoint form given in Sec. 2.2.4 produces the full justification of the special weighting. Which of the procedures is best used in practice is largely a matter of taste, as all can give excellent results. However, we shall see from the second part of this chapter, in which transient problems are dealt with, that other methods can be adopted if time-stepping procedures are used as an iteration to derive steady-state algorithms.

Indeed most of these procedures will again result in the addition of a diffusion term in which the parameter $\alpha$ is now replaced by another one involving the length of the time step $\Delta t$. We shall show at the end of the next section a comparison between various procedures for stabilization and will note essentially the same forms in the steady-state situation.

In the last part of this chapter (Part III) we shall address the case in which the unknown $\phi$ is a vector variable. Here only a limited number of procedures described in the first two parts will be available and even so we do not recommend in general the use of such methods for vector-valued functions.

Before proceeding further it is of interest to consider the original equation with a source term proportional to the variable $\phi$, i.e. writing the one-dimensional equation (2.11) as

$$U \frac{d\phi}{dx} - \frac{d}{dx} \left( k \frac{d\phi}{dx} \right) + m\phi + Q = 0 \quad (2.57)$$

Equations of this type will arise of course from the transient Eq. (2.10) if we assume the solution to be decomposed into Fourier components, writing for each component

$$Q = Q^* e^{i\omega t}, \quad \phi = \phi^* e^{i\omega t} \quad (2.58)$$

which on substitution gives

$$U \frac{d\phi^*}{dx} - \frac{d}{dx} \left( k \frac{d\phi^*}{dx} \right) + i\omega \phi^* + Q^* = 0 \quad (2.59)$$

in which $\phi^*$ can be complex.

The use of Petrov–Galerkin or similar procedures on Eq. (2.57) or (2.59) can again be made. If we pursue the line of approach outlined in Sec. 2.2.4 we note that

(a) the function $p$ required to achieve self-adjointness remains unchanged;

and hence

(b) the weighting applied to achieve optimal results (see Sec. 2.2.3) again remains unaltered – providing of course it is applied to all terms.

Although the above result is encouraging and permits the solution in the frequency domain for transient problems, it does not readily ‘transplant’ to problems in which time-stepping procedures are required.

Some further points require mentioning at this stage. These are simply that:

1. When pure convection is considered (that is $k = 0$) only one boundary condition – generally that giving the value of $\phi$ at the inlet – can be specified, and in such a case the violent oscillations observed in Fig. 2.2 with standard Galerkin methods will not occur generally.
2. Specification of no boundary condition at the outlet edge in the case when $k > 0$, which is equivalent to imposing a zero conduction flux there, generally results in quite acceptable solutions with standard Galerkin weighting even for quite high Peclet numbers.

Part II: Transients

2.5 Transients – introductory remarks

2.5.1 Mathematical background

The objective of this section is to develop procedures of general applicability for the solution by direct time-stepping methods of Eq. (2.1) written for scalar values of $\phi$, $F_i$, and $G_i$:

$$\frac{\partial \phi}{\partial t} + \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_i} + Q = 0$$

(2.60)

though consideration of the procedure for dealing with a vector-valued function will be included in Part III. However, to allow a simple interpretation of the various methods and of behaviour patterns the scalar equation in one dimension [see Eq. (2.10)], i.e.

$$\frac{\partial \phi}{\partial t} + U \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q = 0$$

(2.61a)

will be considered. This of course is a particular case of Eq. (2.60) in which $F = F(\phi)$, $U = \partial F/\partial \phi$ and $Q = Q(\phi, x)$ and therefore

$$\frac{\partial F}{\partial x} = \frac{\partial F}{\partial \phi} \frac{\partial \phi}{\partial x} = U \frac{\partial \phi}{\partial x}$$

(2.61b)

The problem so defined is non-linear unless $U$ is constant. However, the non-conservative equations (2.61) admit a spatial variation of $U$ and are quite general.

The main behaviour patterns of the above equations can be determined by a change of the independent variable $x$ to $x'$ such that

$$dx_i' = dx_i - U_i dt$$

(2.62)

Noting that for $\phi = \phi(x', t)$ we have

$$\frac{\partial \phi}{\partial t} \bigg|_{x \text{ const}} = \frac{\partial \phi}{\partial x'_i} \frac{dx'_i}{dt} + \frac{\partial \phi}{\partial t} \bigg|_{x' \text{ const}} = -U_i \frac{\partial \phi}{\partial x'_i} + \frac{\partial \phi}{\partial t} \bigg|_{x' \text{ const}}$$

(2.63)

The one-dimensional equation (2.61a) now becomes simply

$$\frac{\partial \phi}{\partial t} - \frac{\partial}{\partial x'} \left( k \frac{\partial \phi}{\partial x'} \right) + Q(x') = 0$$

(2.64)
and equations of this type can be readily discretized with self-adjoint spatial operators and solved by procedures developed previously in Volume 1.

The coordinate system of Eq. (2.62) describes characteristic directions and the moving nature of the coordinates must be noted. A further corollary of the coordinate change is that with no conduction or heat generation terms, i.e. when $k = 0$ and $Q = 0$, we have simply

$$\frac{\partial \phi}{\partial t} = 0$$

or

$$\phi(x') = \phi(x - Ut) = \text{constant}$$

along a characteristic [assuming $U$ to be constant, which will be the case if $F = F(\phi)$]. This is a typical equation of a wave propagating with a velocity $U$ in the $x$ direction, as shown in Fig. 2.10. The wave nature is evident in the problem even if the conduction (diffusion) is not zero, and in this case we shall have solutions showing a wave that attenuates with the distance travelled.

### 2.5.2 Possible discretization procedures

In Part I of this chapter we have concentrated on the essential procedures applicable directly to a steady-state set of equations. These procedures started off from somewhat heuristic considerations. The Petrov–Galerkin method was perhaps the most rational but even here the amount and the nature of the weighting functions were a matter of guess-work which was subsequently justified by consideration of the numerical error at nodal points. The Galerkin least square (GLS) method in the same way provided no absolute necessity for improving the answers though of course the least square method would tend to increase the symmetry of the equations and thus could be proved useful. It was only by results which turned out to be remarkably similar to those obtained by the Petrov–Galerkin methods that we have deemed this method to be a success. The same remark could be directed at the finite increment calculus (FIC) method and indeed to other methods suggested dealing with the problems of steady-state equations.

For the transient solutions the obvious first approach would be to try again the same types of methods used in steady-state calculations and indeed much literature
Convection dominated problems

has been devoted to this. Petrov-Galerkin methods have been used here quite extensively. However, it is obvious that the application of Petrov-Galerkin methods will lead to non-symmetric mass matrices and these will be difficult to use for any explicit method as lumping is not by any means obvious.

Serious difficulty will also arise with the Galerkin least squares (GLS) procedure even if the temporal variation is generally included by considering space-time finite elements in the whole formulation. This approach to such problems was made by Nguyen and Reynen, Carey and Jieng, Johnson and coworkers and others. However the use of space-time elements is expensive as explicit procedures are not available.

Which way, therefore, should we proceed? Is there any other obvious approach which has not been mentioned? The answer lies in the wave nature of the equations which indeed not only permits different methods of approach but in many senses is much more direct and fully justifies the numerical procedures which we shall use. We shall therefore concentrate on such methods and we will show that they will lead to artificial diffusions which in form are very similar to those obtained previously by the Petrov-Galerkin method but in a much more direct manner which is consistent with the equations.

The following discussion will therefore be centred on two main directions: (I) the procedures based on the use of the characteristics and the wave nature directly leading to so-called characteristic Galerkin methods which we shall discuss in Sec. 2.6; and then (2) we shall proceed to approach the problem through the use of higher-order time approximations called Taylor-Galerkin methods.

Of the two approaches the first one based on the characteristics is in our view more important. However for historical and other reasons we shall discuss both methods which for a scalar variable can be shown to give identical answers.

The solutions of convective scalar equations can be given by both approaches very simply. This will form the basis of our treatment for the solution of fluid mechanics equations in Chapter 3, where both explicit iterative processes as well as implicit methods can be used.

Many of the methods for solving the transient scalar equations of convective diffusion have been applied to the full fluid mechanics equations, i.e. solving the full vector-valued convective–diffusive equations we have given at the beginning of the chapter (Eq. 2.1). This applies in particular to the Taylor–Galerkin method which has proved to be quite successful in the treatment of high-speed compressible gas flow problems. Indeed this particular approach was the first one adopted to solve such problems. However, the simple wave concepts which are evident in the scalar form of the equations do not translate to such multivariant problems and make the procedures largely heuristic. The same can be said of the direct application of the SUPG and GLS methods to multivariant problems. We have shown in Volume 1, Chapter 12 that procedures such as GLS can provide a useful stabilization of difficulties encountered with incompressibility behaviour. This does not justify their widespread use and we therefore recommend the alternatives to be discussed in Chapter 3.

For completeness, however, Part III of this chapter will be added to discuss to some extent the extension of some methods to vector-type variables.
2.6 Characteristic-based methods

2.6.1 Mesh updating and interpolation methods

We have already observed that, if the spatial coordinate is ‘convected’ in the manner implied by Eq. (2.62), i.e. along the problem characteristics, then the convective, first-order, terms disappear and the remaining problem is that of simple diffusion for which standard discretization procedures with the Galerkin spatial approximation are optimal (in the energy norm sense).

The most obvious use of this in the finite element context is to update the position of the mesh points in a lagrangian manner. In Fig. 2.11(a) we show such an update for the one-dimensional problem of Eq. (2.61) occurring in an interval $\Delta t$.

For a constant $x'$ coordinate

$$dx = U \, dt$$

\begin{align*}
t'^{n+1} &= t^n + \Delta t \\
\Delta t &= \text{characteristic} \\
\text{Updated node position} \\
\text{Initial node position} \\
\end{align*}

(a) Forward

\begin{align*}
t'^{n+1} &= t^n - \Delta t \\
\Delta t &= \text{characteristic} \\
\text{Updated node position} \\
\text{Initial node position} \\
\end{align*}

(b) Backward

Fig. 2.11 Mesh updating and interpolation: (a) Forward; (b) Backward.
and for a typical nodal point $i$, we have

$$x_i^{n+1} = x_i^n + \int_{t_n}^{t_{n+1}} U \, dt$$

(2.67)

where in general the ‘velocity’ $U$ may be dependent on $x$. However, if $F = F(\phi)$ and $U = \frac{\partial F}{\partial \phi} = U(\phi)$ then the wave velocity is constant along a characteristic by virtue of Eq. (2.65) and the characteristics are straight lines.

For such a constant $U$ we have simply

$$x_i^{n+1} = x_i^n + U \Delta t$$

(2.68)

for the updated mesh position. This is not always the case and updating generally has to be done with variable $U$.

On the updated mesh only the time-dependent diffusion problem needs to be solved, using the methods of Volume 1. These we need not discuss in detail here.

The process of continuously updating the mesh and solving the diffusion problem on the new mesh is, of course, impracticable. When applied to two- or three-dimensional configurations very distorted elements would result and difficulties will always arise on the boundaries of the domain. For that reason it seems obvious that after completion of a single step a return to the original mesh should be made by interpolating from the updated values, to the original mesh positions.

This procedure can of course be reversed and characteristic origins traced backwards, as shown in Fig. 2.11(b) using appropriate interpolated starting values.

The method described is somewhat intuitive but has been used with success by Adey and Brebbia and others as early as 1974 for solution of transport equations. The procedure can be formalized and presented more generally and gives the basis of so-called characteristic–Galerkin methods.

The diffusion part of the computation is carried out either on the original or on the final mesh, each representing a certain approximation. Intuitively we imagine in the updating scheme that the operator is split with the diffusion changes occurring separately from those of convection. This idea is explained in the procedures of the next section.

### 2.6.2 Characteristic–Galerkin procedures

We shall consider that the equation of convective diffusion in its one-dimensional form (2.61) is split into two parts such that

$$\phi = \phi^* + \phi^{**}$$

(2.69)

and

$$\frac{\partial \phi^*}{\partial t} + U \frac{\partial \phi}{\partial x} = 0$$

(2.70a)

is a purely convective system while

$$\frac{\partial \phi^{**}}{\partial t} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q = 0$$

(2.70b)
represents the self-adjoint terms [here $Q$ contains the source, reaction and term $(\partial U/\partial x)\phi$].

Both $\phi^*$ and $\phi^{**}$ are to be approximated by standard expansions

$$\hat{\phi}^* = N\phi^* \quad \hat{\phi}^{**} = N\phi^{**}$$  (2.71)

and in a single time step $t^n$ to $t^n + \Delta t = t^{n+1}$ we shall assume that the initial conditions are

$$t = t^n \quad \phi^* = 0 \quad \phi^{**} = \phi^{*n}$$  (2.72)

Standard Galerkin discretization of the diffusion equation allows $\phi^{**n+1}$ to be determined on the given fixed mesh by solving an equation of the form

$$M\Delta \hat{\phi}^{**n+1} = \Delta t H(\hat{\phi}^n + \theta \Delta \hat{\phi}^{**n}) + f$$  (2.73)

with

$$\hat{\phi}^{**n+1} = \hat{\phi}^{**n} + \Delta \hat{\phi}^{**n}$$

In solving the convective problem we assume that $\phi^*$ remains unchanged along the characteristic. However, Fig. 2.12 shows how the initial value of $\phi^{**n}$ interpolated by standard linear shape functions at time $n$ [see Eq. (2.71)] becomes shifted and distorted. The new value is given by

$$\phi^{**n+1} = N(y)\hat{\phi}^{**n} \quad y = x + U\Delta t$$  (2.74)

As we require $\phi^{**n+1}$ to be approximated by standard shape functions, we shall write a projection for smoothing of these values as

$$\int_{\Omega} N^T(N\phi^{**n+1} - N(y)\phi^{**n}) \, dx = 0$$  (2.75)

giving

$$M\hat{\phi}^{**n+1} = \int_{\Omega} [N^TN(y) \, dx] \hat{\phi}^n$$  (2.76a)

where $N = N(x)$ and $M$ is

$$M = \int_{\Omega} N^TN \, dx$$  (2.76b)

The evaluation of the above integrals is of course still complex, especially if the procedure is extended to two or three dimensions. This is generally performed numerically and the stability of the formulation is dependent on the
Convection dominated problems

The accuracy of such integration.\textsuperscript{46} The scheme is stable and indeed exact as far as the convective terms are concerned if the integration is performed exactly (which of course is an unreachable goal). However, stability and indeed accuracy will even then be controlled by the diffusion terms where several approximations have been involved.

\subsection*{2.6.3 A simple explicit characteristic–Galerkin procedure}

Many variants of the schemes described in the previous section are possible and were introduced quite early. References \textsuperscript{45–56} present some successful versions. However, all methods then proposed are somewhat complex in programming and are time consuming. For this reason a simpler alternative was developed in which the difficulties are avoided at the expense of conditional stability. This method was first published in \textsuperscript{1984\textsuperscript{57}} and is fully described in numerous publications.\textsuperscript{58–61} Its derivation involves a local Taylor expansion and we illustrate this in Fig. 2.13.

We can write Eq. (2.61a) along the characteristic as

\[
\frac{\partial \phi}{\partial t} (x'(t), t) = \frac{\partial}{\partial x'} \left( k \frac{\partial \phi}{\partial x'} \right) - Q(x') = 0
\]  

(2.77)

As we can see, in the moving coordinate $x'$, the convective acceleration term disappears and source and diffusion terms are averaged quantities along the characteristic. Now the equation is self-adjoint and the Galerkin spatial approximation is optimal. The time discretization of the above equation along the characteristic (Fig. 2.13) gives

\[
\frac{1}{\Delta t} (\phi^{n+1} - \phi^n)|_{(x-\delta)} \approx \theta \left[ \frac{\partial \phi}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) - Q \right]^{n+1} + (1 - \theta) \left[ \frac{\partial \phi}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) - Q \right]^{n}|_{(x-\delta)}
\]  

(2.78)

where $\theta$ is equal to zero for explicit forms and between zero and unity for semi- and fully implicit forms. As we know, the solution of the above equation in moving coordinates leads to mesh updating and presents difficulties, so we will suggest

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig213.png}
\caption{A simple characteristic–Galerkin procedure.}
\end{figure}
alternatives. From the Taylor expansion we have

\[ \phi^n|_{(x-\delta)} \approx \phi^n - \delta \frac{\partial \phi^n}{\partial x} + \frac{\delta^2}{2} \frac{\partial^2 \phi^n}{\partial x^2} + O(\Delta t^3) \]  

(2.79)

and assuming \( \theta = 0.5 \)

\[ \frac{1}{2} \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)|_{(x-\delta)} \approx \frac{1}{2} \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^n - \delta \frac{\partial}{\partial x} \left( \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^n \right) + O(\Delta t^2) \]  

(2.80a)

\[ \frac{1}{2} Q^n|_{(x-\delta)} = \frac{Q^n}{2} - \frac{\delta}{2} \frac{\partial Q^n}{\partial x} \]  

(2.80b)

where \( \delta \) is the distance travelled by the particle in the \( x \)-direction (Fig. 2.13) which is

\[ \delta = \bar{U} \Delta t \]  

(2.81)

where \( \bar{U} \) is an average value of \( U \) along the characteristic. Different approximations of \( \bar{U} \) lead to different stabilizing terms. The following relation is commonly used\(^{62,63}\)

\[ \bar{U} = U^n - U^n \Delta t \frac{\partial U^n}{\partial x} \]  

(2.82)

Inserting Eqs (2.79)–(2.82) into Eq. (2.78) we have

\[ \phi^{n+1} - \phi^n = -\Delta t \left\{ U \frac{\partial \phi^n}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^{n+1/2} + Q^{n+1/2} \right\} \]

\[ + \Delta t \left\{ \frac{\Delta t}{2} \frac{\partial}{\partial x} \left[ U^n \frac{\partial \phi}{\partial x} \right] - \frac{\Delta t}{2} U \frac{\partial^2}{\partial x^2} \left( k \frac{\partial \phi}{\partial x} \right) + \frac{\Delta t}{2} U \frac{\partial Q}{\partial x} \right\} \]  

(2.83a)

where

\[ \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^{n+1/2} = \frac{1}{2} \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^{n+1} + \frac{1}{2} \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^n \]  

(2.83b)

and

\[ Q^{n+1/2} = \frac{Q^{n+1} + Q^n}{2} \]  

(2.83c)

In the above equation, higher-order terms (from Eq. 2.80) are neglected. This, as already mentioned, is of an identical form to that resulting from Taylor–Galerkin procedures which will be discussed fully in the next section, and the additional terms add the stabilizing diffusion in the streamline direction. For multidimensional problems, Eq. (2.83a) can be written in indicial notation and approximating \( n + 1/2 \) terms with \( n \) terms (for the fully explicit form)

\[ \phi^{n+1} - \phi^n = -\Delta t \left\{ U_{ij} \frac{\partial \phi}{\partial X_j} - \frac{\partial}{\partial X_j} \left( k \frac{\partial \phi}{\partial X_j} \right) + Q \right\} \]

\[ + \Delta t \left\{ \frac{\Delta t}{2} \frac{\partial}{\partial X_j} \left[ U_{ij} \frac{\partial \phi}{\partial X_j} \right] - \frac{\Delta t}{2} U_{ik} \frac{\partial}{\partial X_k} \left[ \frac{\partial}{\partial X_j} \left( k \frac{\partial \phi}{\partial X_j} \right) \right] + \frac{\Delta t}{2} U_{ij} \frac{\partial Q}{\partial X_j} \right\} \]  

(2.84)
Convection dominated problems

An alternative approximation for \( \bar{U} \) recently recommended is\(^{62} \)

\[
\bar{U} = \frac{U^{n+1} + U^n|_{(x-\delta)}}{2}
\]  

(2.85)

Using the Taylor expansion

\[
U^n|_{(x-\delta)} \approx U^n - \Delta t U^n \frac{\partial U^n}{\partial x} + O(\Delta t^2)
\]  

(2.86)

from Eqs (2.78)-(2.81) and Eqs (2.85) and (2.86) with \( \theta \) equal to 0.5 we have

\[
\frac{1}{\Delta t} (\phi^{n+1} - \phi^n) = - U^{n+1/2} \frac{\partial \phi^n}{\partial x} + \Delta t U^n \frac{\partial U^n}{\partial x} \frac{\partial \phi^n}{\partial x} + \Delta t U^{n+1/2} U^{n+1/2} \frac{\partial^2 \phi}{\partial x^2}
\]

\[
+ \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^{n+1/2} - \frac{\Delta t}{2} U^{n+1/2} \frac{\partial}{\partial x} \left[ \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right)^n \right]
\]

\[
- \frac{\Delta t}{2} U^{n+1/2} \frac{\partial Q}{\partial x}
\]  

(2.87)

where

\[
U^{n+1/2} = \frac{U^{n+1} + U^n}{2}
\]  

(2.88)

We can further approximate, as mentioned earlier, \( n + 1/2 \) terms using \( n \), to get the fully explicit version of the scheme. Thus we have

\[
U^{n+1/2} = U^n + O(\Delta t)
\]  

(2.89)

and similarly the diffusion term is approximated. The final form of the explicit characteristic-Galerkin method can be written as

\[
\Delta \phi = \phi^{n+1} - \phi^n = - \Delta t \left[ U^n \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n
\]

\[
+ \frac{\Delta t^2}{2} U^n \frac{\partial}{\partial x} \left[ U^n \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n
\]  

(2.90)

Generalization to multidimensions is direct and can be written in indicial notation for equations of the form Eq. (2.5):

\[
\Delta \phi = - \Delta t \left[ \frac{\partial (U_j \phi)}{\partial x_j} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q \right]^n
\]

\[
+ \frac{\Delta t^2}{2} U^n \frac{\partial}{\partial x_k} \left[ \frac{\partial (U_j \phi)}{\partial x_j} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q \right]^n
\]  

(2.91)

The reader will notice the difference in the stabilizing terms obtained by two different approximations for \( \bar{U} \). However, as we can see the difference between them is small and when \( U \) is constant both approximations give identical stabilizing terms. In the rest of the book we shall follow the latter approximation and always use the conservative form of the equations (Eq. 2.91).
As we proved earlier, the Galerkin spatial approximation is justified when the characteristic–Galerkin procedure is used. We can thus write the approximation

\[ \phi = N \tilde{\phi} \]  

and use the weighting \( N^T \) in the integrated residual expression. Thus we obtain

\[ M(\tilde{\phi}^{n+1} - \tilde{\phi}^n) = -\Delta t([C \tilde{\phi}^n + K \tilde{\phi}^n + f^n] - \Delta t(K_u \tilde{\phi}^n + f_i^n]) \]

in explicit form without higher-order derivatives and source terms. In the above equation,

\[ M = \int_{\Omega} N^T N \, d\Omega \quad C = \int_{\Omega} N^T \frac{\partial}{\partial x_i} (U_i N) \, d\Omega \]

\[ K = \int_{\Omega} \frac{\partial N^T}{\partial x_i} k \frac{\partial N}{\partial x_i} \, d\Omega \quad f = \int_{\Omega} N^T Q \, d\Omega + \text{b.t.} \]

and \( K_u \) and \( f_i^n \) come from the new term introduced by the discretization along the characteristics. After integration by parts, the expression of \( K_u \) and \( f_i \) is

\[ K_u = -\frac{1}{2} \int_{\Omega} \frac{\partial}{\partial x_i} (U_i N^T) \frac{\partial}{\partial x_i} (U_i N) \, d\Omega \]  

\[ f_i = -\frac{1}{2} \int_{\Omega} \frac{\partial}{\partial x_i} (U_i N^T) Q \, d\Omega + \text{b.t.} \]

where b.t. stands for integrals along region boundaries. Note that the higher-order derivatives are not included in the above equation.

The approximation is valid for any scalar convected quantity even if that is the velocity component \( U_i \) itself, as is the case with momentum-conservation equations. For this reason we have elaborated above the full details of the spatial approximation as the matrices will be repeatedly used.

It is of interest that the explicit form of Eq. (2.93) is only conditionally stable. For one-dimensional problems, the stability condition is given as (neglecting the effect of sources)

\[ \Delta t \leq \Delta t_{\text{crit}} = \frac{h}{|U|} \]  

for linear elements.

In two-dimensional problems the criteria time step may be computed as

\[ \Delta t_{\text{crit}} = \frac{\Delta t_o \Delta t_o}{\Delta t_o + \Delta t_o} \]

where \( \Delta t_o \) is given by Eq. (2.97) and \( \Delta t_o = h^2/2k \) is the diffusive limit for the critical one-dimensional time step.

Further, with \( \Delta t = \Delta t_{\text{crit}} \) the steady-state solution results in an (almost) identical diffusion change to that obtained by using the optimal streamline upwinding procedures discussed in Part I of this chapter. Thus if steady-state solutions are the main objective of the computation such a value of \( \Delta t \) should be used in connection with the \( K_u \) term.
A fully implicit form of solution is an expensive one involving unsymmetric matrices. However it is often convenient to apply $\theta \geq 1/2$ to the diffusive term only. We call this a nearly (or semi) implicit form and if it is employed we return to the stability condition

$$\Delta t_{\text{crit}} = \frac{h}{|U|}$$

which can present an appreciable benefit.

Figure 2.14 shows the stability limit variation prescribed by Eq. (2.97) with a lumped mass matrix.

It is of considerable interest to examine the behaviour of the solution when the steady state is reached – for instance, if we use the time-stepping algorithm of Eq. (2.93) as an iterative process. Now the final solution is given by taking

$$\tilde{\phi}^{n+1} = \tilde{\phi}^n = \tilde{\phi}$$

which gives

$$[(C + K - \Delta tK_u)\tilde{\phi} + f - \Delta tf_s = 0$$

(2.100)

Inspection of Secs 2.2 and 2.3 shows that the above is identical in form with the use of the Petrov–Galerkin approximation. In the latter the matrix $C$ is identical and the matrix $K_u$ includes balancing diffusion of the amount given by $\frac{1}{2} \alpha Uh$. However, if we take

$$\frac{1}{2} \alpha Uh = \frac{U^2 \Delta t}{2}$$

(2.101)

the identity of the two schemes results. This can be written as a requirement that

$$\alpha = \frac{U \Delta t}{h} = C$$

(2.102)

where $C$ is the Courant number.
In Fig. 2.14 we therefore plot the optimal value of $\alpha$ as given in Eq. (2.25) against $Pe$. We note immediately that if the time-stepping scheme is operated at or near the critical stability limit of the lumped scheme the steady-state solution reached will be close to that resulting from the optimal Petrov–Galerkin process for the steady state. However, if smaller time steps than the critical ones are used, the final solution, though stable, will tend towards the standard Galerkin steady-state discretization and may show oscillations if boundary conditions are such that boundary layers are created. Nevertheless, such small time steps result in very accurate transients so we can conclude that it is unlikely that optimality for transients and steady state can be reached simultaneously.

Examination of Eqs (2.93) shows that the characteristic Galerkin algorithm could have been obtained by applying a Petrov–Galerkin weighting

$$N^T + \frac{\Delta t}{2} U_i \frac{\partial N^T}{\partial x_i}$$

to the various terms of the governing equation (2.60) excluding the time derivative $\partial \phi / \partial t$ to which the standard Galerkin weighting of $N^T$ is attached. Comparing the above with the steady-state problem and the weighting given in Eq. (2.48) the connection is obvious.

A two-dimensional application of the characteristic–Galerkin process is illustrated in Fig. 2.15 in which we show pure convection of a disturbance in a circulating flow. It is remarkable to note that almost no dispersion occurs after a complete revolution. The present scheme is here contrasted with the solution obtained by the finite difference scheme of Lax and Wendroff which for a regular one-dimensional mesh gives a scheme identical to the characteristic–Galerkin except for mass matrix, which is lumped in the finite difference scheme.

It seems that here the difference is entirely due to the proper form of the mass matrix $M$ now used and we note that for transient response the importance of the consistent mass matrix is crucial. However, the numerical convenience of using the lumped form is overwhelming in an explicit scheme. It is easy to recover the performance of the consistent mass matrix by using a simple iteration. In this we write Eq. (2.93) as

$$M \Delta \tilde{\phi}^n = \Delta t S^n$$  \hspace{1cm} (2.103)

with $S^n$ being the right-hand side of Eq. (2.93) and

$$\tilde{\phi}^{n+1} = \tilde{\phi}^n + \Delta \tilde{\phi}^n$$

Substituting a lumped mass matrix $M_L$ to ease the solution process we can iterate as follows:

$$\begin{align*}
(\Delta \tilde{\phi})^n_l &= M_L^{-1} [\Delta t S^n - M(\Delta \tilde{\phi})^n_{l-1}] + (\Delta \tilde{\phi})^n_{l-1}
\end{align*}$$  \hspace{1cm} (2.104)

where $l$ is the iteration number. The process converges very rapidly and in Fig. 2.16 we show the dramatic improvements of results in the solution of a one-dimensional wave propagation with three such iterations done at each time step. At this stage the results are identical to those obtained with the consistent mass matrix.
2.6.4 Boundary conditions – radiation

As we have already indicated the convection–diffusion problem allows a single boundary condition of the type

$$\phi = \bar{\phi} \text{ on } \Gamma_u$$

(2.105a)
Characteristic-based methods

Fig. 2.16 Characteristic–Galerkin method in the solution of a one-dimensional wave progression. Effect of using a lumped mass matrix and of consistent iteration.

\[ -k \left( \frac{\partial \phi}{\partial n} \right) = \bar{q} \text{ on } \Gamma_q \quad (2.105b) \]

(where $\Gamma = \Gamma_u \cup \Gamma_q$) to be imposed, providing the equation is of second order and diffusion is present.

In the case of pure convection this is no longer the case as the differential equation is of first order. Indeed this was responsible for the difficulty of obtaining a solution in the example of Fig. 2.2 when $Pe \to \infty$ and an exit boundary condition of the type given by Eq. (2.105a) was imposed. In this one-dimensional case for pure convection only the inlet boundary condition can be given; at the exit no boundary condition needs to be prescribed if $U$, the wave velocity, is positive.

For multidimensional problems of pure convection the same wave specification depends on the value of the normal component of $U$. Thus if

\[ U, n, > 0 \quad (2.106) \]

where $n_i$ is the normal direction vector, the wave is leaving the problem and then no boundary condition is specified. If the problem has some diffusion, the same
specification of 'no boundary condition' is equivalent to putting

$$-k \left( \frac{\partial \phi}{\partial n} \right) = 0$$

(2.107)

at the exit boundary.

In Fig. 2.17 we illustrate, following the work of Peraire, how cleanly the same wave as that specified in the problem of Fig. 2.15 leaves the domain in the uniform velocity field when the correct boundary condition is imposed.

Fig. 2.17 A gaussian distribution advected in a constant velocity field. Boundary condition causes no reflection.
2.7 Taylor–Galerkin procedures for scalar variables

In the Taylor–Galerkin process, the Taylor expansion in time precedes the Galerkin space discretization. Firstly, the scalar variable $\phi$ is expanded by the Taylor series in time:

$$
\phi^{n+1} = \phi^n + \Delta t \frac{\partial \phi^n}{\partial t} + \frac{\Delta t^2}{2} \frac{\partial^2 \phi^n}{\partial t^2} + O(\Delta t^3)
$$  \hspace{1cm} (2.108)

From Eq. (2.61a) we have

$$
\frac{\partial \phi^n}{\partial t} = \left[ -U \frac{\partial \phi}{\partial x} + \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n
$$  \hspace{1cm} (2.109)

and

$$
\frac{\partial^2 \phi^n}{\partial t^2} = \frac{\partial}{\partial t} \left[ -U \frac{\partial \phi}{\partial x} + \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n
$$  \hspace{1cm} (2.110)

Substituting Eqs (2.109) and (2.110) into Eq. (2.108) we have

$$
\phi^{n+1} - \phi^n = -\Delta t \left[ U \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n - \frac{\Delta t^2}{2} \frac{\partial}{\partial t} \left[ U \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial t} \right) + Q \right]^n
$$  \hspace{1cm} (2.111)

Assuming $U$ and $k$ to be constant we have

$$
\phi^{n+1} - \phi^n = -\Delta t \left[ U \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n - \frac{\Delta t^2}{2} \frac{\partial}{\partial x} \left[ U \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n
$$  \hspace{1cm} (2.112)

Inserting Eq. (2.109) into Eq. (2.112) and neglecting higher-order terms

$$
\phi^{n+1} - \phi^n = -\Delta t \left[ U \frac{\partial \phi}{\partial x} - \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + Q \right]^n + \frac{\Delta t^2}{2} \frac{\partial}{\partial x} \left[ U^2 \frac{\partial \phi}{\partial x} - U \frac{\partial}{\partial x} \left( k \frac{\partial \phi}{\partial x} \right) + UQ \right]^n + O(\Delta t^3)
$$  \hspace{1cm} (2.113)

As we can see the above equation, having assumed constant $U$ and $k$, is identical to Eq. (2.83a) derived from the characteristic approach. Clearly for scalar variables both characteristic and Taylor–Galerkin procedures give identical stabilizing terms. Thus selection of a method for a scalar variable is a matter of taste. However, the sound mathematical justification of the characteristic–Galerkin method should be mentioned here.

The Taylor–Galerkin procedure for the convection–diffusion equation in multi-dimensions can be written as

$$
\phi^{n+1} - \phi^n = -\Delta t \left\{ U_j \frac{\partial \phi}{\partial x_j} - \frac{\partial}{\partial x_i} \left( k \frac{\partial \phi}{\partial x_i} \right) + Q \right\} - \frac{\Delta t^2}{2} \frac{\partial}{\partial x_i} \left[ U_i U_j \frac{\partial \phi}{\partial x_j} - U_i \frac{\partial}{\partial x_j} \left( k \frac{\partial \phi}{\partial x_j} \right) + U_i Q \right]\} \hspace{1cm} (2.114)
$$
again showing the complete similarity with the appropriate characteristic–Galerkin form and identity when \( U_i \) and \( k \) are constant. The Taylor–Galerkin method is the finite element equivalent of the Lax–Wendroff method developed in the finite difference context.$^{64}$

2.8 Steady-state condition

Both the Taylor–Galerkin and characteristic–Galerkin methods give an answer which compares directly with SUPG and GLS giving additional streamline diffusion (higher-order derivatives are omitted)

$$\frac{\Delta t^2}{2} \frac{\partial}{\partial x_i} \left[ U_i U_j \frac{\partial \phi}{\partial x_j} \right]$$

with \( \Delta t \) replacing the coefficient \( \alpha h \). With the characteristic–Galerkin method being the only method that has a full mathematical justification, we feel that even for steady state problems this should be considered as an appropriate solution technique.

2.9 Non-linear waves and shocks

The procedures developed in the previous sections are in principle of course available for both linear and non-linear problems (with explicit procedures of time stepping being particularly efficient for the latter). Quite generally the convective part of the equation, i.e.

$$\frac{\partial \phi}{\partial t} + \frac{\partial F_i}{\partial x_i} \equiv \frac{\partial \phi}{\partial t} + U_i \frac{\partial \phi}{\partial x_i} = 0$$

will have the vector \( U_i \) dependent on \( \phi \). Thus

$$U_i \equiv \frac{\partial F_i}{\partial \phi} = U_i(\phi)$$

In the one-dimensional case with a scalar variable we shall have equations of the type

$$\frac{\partial \phi}{\partial t} + \frac{\partial F}{\partial x} = \frac{\partial \phi}{\partial t} + U(\phi) \frac{\partial \phi}{\partial x} = 0$$

corresponding to waves moving with a non-uniform velocity \( U \). A typical problem in this category is that due to Burger, which is defined by

$$\frac{\partial \phi}{\partial t} + \frac{\partial }{\partial x} \left( \frac{1}{2} \phi^2 \right) = \frac{\partial \phi}{\partial t} + \phi \frac{\partial \phi}{\partial x} = 0$$

In Fig. 2.18 we illustrate qualitatively how different parts of the wave moving with velocities proportional to their amplitude cause it to steepen and finally develop into a shock form. This behaviour is typical of many non-linear systems and in Chapter 6 we shall see how shocks develop in compressible flow at transonic speeds.
Fig. 2.18 Progression of a wave with velocity $U - \phi$.

Fig. 2.19 Development of a shock (Burger equation).
To illustrate the necessity for the development of the shock, consider the propagation of a wave with an originally smooth profile illustrated in Fig. 2.19(a). Here as we know the characteristics along which \( \phi \) is constant are straight lines shown in Fig. 2.19(b). These show different propagation speeds intersecting at time \( t = 2 \) when a discontinuous shock appears. This shock propagates at a finite speed (which here is the average of the two extreme values).

In such a shock the differential equation is no longer valid but the conservation integral is. We can thus write for a small length \( \Delta s \) around the discontinuity

\[
\frac{\partial}{\partial t} \int_{\Delta s} \phi \, ds + F(s + \Delta s) - F(s) = 0 \tag{2.120}
\]

or

\[
C \Delta \phi - \Delta F = 0 \tag{2.121a}
\]

where \( C = \lim \Delta s/\Delta t \) is the speed of shock propagation and \( \Delta \phi \) and \( \Delta F \) are the discontinuities in \( \phi \) and \( F \) respectively. Equation (2.121a) is known as the Rankine–Hugoniot condition.

We shall find that such shocks develop frequently in the context of compressible flow and shallow-water flow (Chapters 6 and 7) and can often exist even in the presence of diffusive terms in the equation. Indeed, such shocks are not specific to transients but can persist in the steady state. Clearly, approximation of the finite element kind in which we have postulated in general a \( C_0 \) continuity to \( \phi \) can at best smear such a discontinuity over an element length, and generally oscillations near such a discontinuity arise even when the best algorithms of the preceding sections are used.

Figure 2.20 illustrates the difficulties of modelling such steep waves occurring even in linear problems in which the physical dissipation contained in the equations is incapable of smoothing the solution out reasonably, and to overcome this problem artificial diffusivity is frequently used. This artificial diffusivity must have the following characteristics:

1. It must vanish as the element size tends to zero.
2. It must not affect substantially the smooth domain of the solution.

A typical diffusivity often used is a finite element version of that introduced by Lapidus\(^6\) for finite differences, but many other forms of local smoothing have been proposed.\(^6\) The additional diffusivity is of the form

\[
\tilde{k} = C_{\text{Lap}} h^2 \left| \frac{\partial \phi}{\partial x} \right| \tag{2.122}
\]

where the last term gives the maximum gradient.

In Fig. 2.21 we show a problem of discontinuous propagation in the Burger equation and how a progressive increase of the \( C_{\text{Lap}} \) coefficient kills spurious oscillation, but at the expense of rounding of a steep wave.
Non-linear waves and shocks

Fig. 2.20 Propagation of a steep wave by Taylor–Galerkin process: (a) Explicit methods $C = 0.5$, step wave at $Pe = 12500$; (b) Explicit methods $C = 0.1$, step wave at $Pe = 12500$.

For a multidimensional problem with a multidimensional $\phi$ a degree of anisotropy can be introduced and a possible expression generalizing (2.122) is

$$\tilde{K}_{ij} = C_{Lap} h^2 \frac{|V_i V_j|}{|V|}$$

(2.123)

where

$$V_i = \frac{\partial \phi}{\partial x_i}$$

Other possibilities are open here and much current work is focused on the subject of 'shock capture'. We shall return to these problems in Chapter 6 where its importance in the high-speed flow of gases is paramount.
Part III: Vector-valued functions

2.10 Vector-valued variables

2.10.1 The Taylor–Galerkin method used for vector-valued variables

The only method which adapts itself easily to the treatment of vector variables is that of the Taylor–Galerkin procedure. Here we can repeat the steps of Sec. 2.8 but now addressed to the vector-valued equation with which we started this chapter (Eq. 2.1). Noting that now $\varphi$ has multiple components, expanding $\Phi$ by a Taylor series in time we have

$$\Phi^{n+1} = \Phi^n + \Delta t \left. \frac{\partial \Phi}{\partial t} \right|_n + \frac{\Delta t^2}{2} \left. \frac{\partial^2 \Phi}{\partial t^2} \right|_{n+\theta}$$

(2.124)

where $\Theta$ is a number such that $0 \leq \Theta \leq 1$.

From Eq. (2.1),

$$\left[ \frac{\partial \Phi}{\partial t} \right]_n = - \left[ \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_i} + Q \right]_n$$

(2.125a)
and differentiating
\[
\left[ \frac{\partial^2 \Phi}{\partial t^2} \right]_{n+\theta} = -\frac{\partial}{\partial t} \left[ \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right]_{n+\theta} \tag{2.125b}
\]
In the above we can write
\[
\frac{\partial}{\partial t} \left( \frac{\partial F_i}{\partial x_i} \right) = \frac{\partial}{\partial x_i} \left( \frac{\partial F_i}{\partial \Phi} \frac{\partial \Phi}{\partial t} \right) = -\frac{\partial}{\partial x_i} \left[ A_i \left( \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right) \right] \tag{2.125c}
\]
where \( A_i = \partial F_i / \partial \Phi \) and if \( Q = Q(\Phi, x) \) and \( \partial Q / \partial \Phi = S \),
\[
\frac{\partial Q}{\partial t} = \frac{\partial Q}{\partial \Phi} \frac{\partial \Phi}{\partial t} = -S \left( \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right) \tag{2.125d}
\]
We can therefore approximate Eq. (2.124) as
\[
\Delta \Phi^n = \Phi^{n+1} - \Phi^n
= -\Delta t \left[ \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right]_n + \Delta t^2 \left\{ \frac{\partial}{\partial x_i} \left[ A_i \left( \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right) \right] \right\}_n
+ \frac{\partial}{\partial t} \frac{\partial G_i}{\partial x_j} + S \left( \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right) \right\}_{n+\theta} \tag{2.126}
\]
Omitting the second derivatives of \( G_i \) and interpolating the \( n + \theta \) between \( n \) and \( n + 1 \) values we have
\[
\Delta \Phi \equiv \Phi^{n+1} - \Phi^n
= -\Delta t \left[ \frac{\partial F_i}{\partial x_i} + Q \right]_n - \Delta t \left[ \frac{\partial G_i}{\partial x_i} \right]_{n+\theta} + \frac{\partial G_i}{\partial x_i} \right\}_{n+\theta} (1 - \theta)
+ \frac{\Delta t^2}{2} \left\{ \frac{\partial}{\partial x_i} \left[ A_i \left( \frac{\partial F_i}{\partial x_i} + Q \right) \right] + S \left( \frac{\partial F_i}{\partial x_i} + Q \right) \right\}_n \tag{2.127}
\]
At this stage a standard Galerkin approximation is applied which will result in a discrete, implicit, time-stepping scheme that is unconditionally stable if \( \theta \geq \frac{1}{2} \). As the explicit form is of particular interest we shall only give the details of the discretization process for \( \theta = 0 \). Writing as usual
\[
\Phi \approx N \tilde{\Phi}
\]
we have
\[
\left( \int_{\Omega} N^T N \, d\Omega \right) \tilde{\Phi} = -\Delta t \left[ \int_{\Omega} N^T \left( \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right) \, d\Omega \right.
- \frac{\Delta t}{2} \int_{\Omega} N^T \frac{\partial}{\partial x_i} \left\{ A_i \left( \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right) \right\} \, d\Omega
+ \frac{\Delta t}{2} \int_{\Omega} N^T S \left( \frac{\partial F_i}{\partial x_i} + \frac{\partial G_i}{\partial x_j} + Q \right) \, d\Omega \right]_{n} \tag{2.128}
\]
This can be written in a compact matrix form similar to Eq. (2.93) as

\[ M \Delta \Phi = -\Delta t [(C + K_u + K) \Phi + f]^n \]  

(2.129a)

in which, with

\[ G_{ij} = -k_{ij} \frac{\partial \Phi}{\partial x_j} \]

we have (on omitting the third derivative terms and the effect of S) matrices of the form of Eq. (2.94), i.e.

\[
C = \int \Omega N^T A_i \frac{\partial N}{\partial x_i} \, d\Omega \\
K_u = \int \Omega \frac{\partial N^T}{\partial x_i} \left(A_i A_j \frac{\Delta t}{2}\right) \frac{\partial N}{\partial x_j} \, d\Omega \\
K = \int \Omega \frac{\partial N^T}{\partial x_i} k_{ij} \frac{\partial N}{\partial x_j} \, d\Omega \\
f = \int \Omega \left( N^T + \frac{\Delta t}{2} A_i \frac{\partial N^T}{\partial x_i} \right) Q \, d\Omega + \text{boundary terms} \\
M = \int \Omega N^T N \, d\Omega
\]

With \( \theta = \frac{1}{2} \) it can be shown that the order of approximation increases and for this scheme a simple iterative solution is possible.\(^{71}\) We note that with the consistent mass matrix M the stability limit for \( \theta = \frac{1}{2} \) is increased to \( C = 1 \).

Use of \( \theta = \frac{1}{2} \) apparently requires an implicit solution. However, similar iteration to that used in Eq. (2.104) is rapidly convergent and the scheme can be used quite economically.

### 2.10.2 Two-step predictor–corrector methods. Two-step Taylor–Galerkin operation

There are of course various alternative procedures for improving the temporal approximation other than the Taylor expansion used in the previous section. Such procedures will be particularly useful if the evaluation of the derivative matrix A can be avoided. In this section we shall consider two predictor–corrector schemes (of Runge–Kutta type) that avoid the evaluation of this matrix and are explicit.

The first starts with a standard Galerkin space approximation being applied to the basic equation (2.1). This results in the form

\[
M \frac{d\Phi}{dt} = M \ddot{\Phi} = P_C + P_D + f = \psi
\]

(2.130)
where again $M$ is the standard mass matrix, $f$ are the prescribed ‘forces’ and

$$P_c(\Phi) = \int_\Omega N^T \frac{\partial F_i}{\partial x_i} \, d\Omega \quad (2.131a)$$

represents the convective ‘forces’, while

$$P_d(\Phi) = \int_\Omega N^T \frac{\partial G_i}{\partial x_i} \, d\Omega \quad (2.131b)$$

are the diffusive ones.

If an explicit time integration scheme is used, i.e.

$$MA_{n+1} = M(A_n + \Delta t \psi_n(\Phi_n'))$$ \quad (2.132)

the evaluation of the right-hand side does not require the matrix product representation and $A_n$ does not have to be computed.

Of course the scheme presented is not accurate for the various reasons previously discussed, and indeed becomes unconditionally unstable in the absence of diffusion and external force vectors.

The reader can easily verify that in the case of the linear one-dimensional problem the right-hand side is equivalent to a central difference scheme with $\Phi_n'$ and $\Phi_n''$ only being used to find the value of $\Phi_n^+1$, as shown in Fig. 2.22(a).

The scheme can, however, be recast as a two-step, predictor–corrector operation and conditional stability is regained. Now we proceed as follows:

**Step 1.** Compute $\Phi_n^{+1/2}$ using an explicit approximation of Eq. (2.132), i.e.

$$\Phi_n^{+1/2} = \Phi_n + \frac{\Delta t}{2} M^{-1} \psi_n$$ \quad (2.133)

![Progression of information in explicit one- and two-step schemes.](image-url)
Convection dominated problems

Step 2. Compute $\tilde{\Phi}^{n+1}$ inserting the improved value of $\tilde{\Phi}^{n+1/2}$ in the right-hand side of Eq. (2.132), giving

$$\tilde{\Phi}^{n+1} = \Phi^n + \Delta t M^{-1} \psi^{n+1/2}$$  \hspace{1cm} (2.134)

This is precisely equivalent to the second-order Runge–Kutta scheme being applied to the ordinary system of differential equations (2.130). Figure 2.22(b) shows in the one-dimensional example how the information ‘spreads’, i.e. that now $\tilde{\Phi}_i^{n+1}$ will be dependent on values at nodes $i-2, \ldots, i+2$.

It is found that the scheme, though stable, is overdiffusive and numerical results are poor.

An alternative is possible, however, using a two-step Taylor–Galerkin operation. Here we return to the original equation (2.1) and proceed as follows:

Step 1. Find an improved value of $\Phi^{n+1/2}$ using only the convective and source parts. Thus

$$\tilde{\Phi}^{n+1/2} = \Phi^n - \frac{\Delta t}{2} \left( \frac{\partial \Phi^n}{\partial x_i} + Q^n \right)$$  \hspace{1cm} (2.135a)

which of course allows the evaluation of $F_i^{n+1/2}$.

We note, however, that we can also write an approximate expansion as

$$F_i^{n+1/2} = F_i^n + \frac{\Delta t}{2} \frac{\partial F_i^n}{\partial t} = F_i^n - \frac{\Delta t}{2} A^n_i \frac{\partial \Phi^n}{\partial t}$$

$$= F_i^n - \Delta t A^n_i \left( \frac{\partial F_i^n}{\partial x_j} + \frac{\partial G_j^n}{\partial x_j} + Q^n \right)$$  \hspace{1cm} (2.135b)

This gives

$$A^n_i \left( \frac{\partial F_i^n}{\partial x_j} + \frac{\partial G_j^n}{\partial x_j} + Q^n \right) = -\frac{2}{\Delta t} (F_i^{n+1/2} - F_i^n)$$  \hspace{1cm} (2.135c)

Step 2. Substituting the above into the Taylor–Galerkin approximation of Eq. (2.128) we have

$$M \Delta \Phi = -\Delta t \left[ \int_\Omega N^T \left( \frac{\partial F_i^n}{\partial x_i} + \frac{\partial G_j^n}{\partial x_j} + Q^n \right) d\Omega + \int_\Omega N^T \frac{\partial}{\partial x_i} (F_i^{n+1/2} - F_i^n) d\Omega \right]$$

$$+ \int_\Omega N^T S (F_i^{n+1/2} - F_i^n) d\Omega$$  \hspace{1cm} (2.135d)

and after integration by parts of the terms with respect to the $x_i$ derivatives we obtain simply

$$M \Delta \Phi = -\Delta t \left\{ \int_\Omega \frac{\partial N_i^T}{\partial x_i} (F_i^{n+1/2} + G_i^n) d\Omega + \int N^T [Q + S (F_i^{n+1/2} - F_i^n)] d\Omega \right\}$$

$$+ \int_\Gamma N^T (F_i^{n+1/2} + G_i^n) n_i d\Gamma$$  \hspace{1cm} (2.136)
We note immediately that:

1. The above expression is identical to using a standard Galerkin approximation on Eq. (2.1) and an explicit step with \( \mathbf{F} \) values updated by the simple equation (2.135a).

2. The final form of Eq. (2.136) does not require the evaluation of the matrices \( \mathbf{A} \), resulting in substantial computation savings as well as yielding essentially the same results. Indeed, some omissions made in deriving Eqs (2.129) did not occur now and presumably the accuracy is improved.

A further practical point must be noted:

3. In non-linear problems it is convenient to interpolate \( \mathbf{F} \) directly in the finite element manner as

\[
\mathbf{F}_i = \mathbf{N} \tilde{\mathbf{F}}_i
\]

rather than to compute it as \( \mathbf{F}_i(\tilde{\mathbf{F}}) \).

Thus the evaluation of \( \mathbf{F}^{n+1/2}_i \) need only be made at the quadrature (integration) points within the element, and the evaluation of \( \tilde{\mathbf{F}}^{n+1/2} \) by Eq. (2.135a) is only done on such points. For a linear triangle element this reduces to a single evaluation of \( \tilde{\mathbf{F}}^{n+1/2} \) and \( \mathbf{F}^{n+1/2} \) for each element at its centre, taking of course \( \tilde{\mathbf{F}}^{n+1/2} \) and \( \mathbf{F}^{n+1/2} \) as the appropriate interpolation average there.

In the simple one-dimensional linear example the information progresses in the manner shown in Fig. 2.22(c). The scheme, which originated at Swansea, can be appropriately called the Swansea two step,\(^{57,65,72-80}\) and has found much use in the direct solution of compressible high-speed gas flow equations. We shall show some of the results obtained by this procedure in Chapter 6. However in Chapter 3 we shall discuss an alternative which is more general and has better performance. It is of interest to remark that the Taylor-Galerkin procedure can be used in contexts other than direct fluid mechanics. The procedure has been used efficiently by Morgan et al.\(^{81,82}\) in solving electromagnetic wave problems.

### 2.10.3 Multiple wave speeds

When \( \phi \) is a scalar variable, a single wave speed will arise in the manner in which we have already shown at the beginning of Part II. When a vector variable is considered, the situation is very different and in general the number of wave speeds will correspond to the number of variables. If we return to the general equation (2.1), we can write this in the form

\[
\frac{\partial \Phi}{\partial t} + A_i \frac{\partial \Phi}{\partial x_i} + \frac{\partial G_i}{\partial x_i} + \mathbf{Q} = 0
\]

\[(2.137)\]

where \( A_i \) is a matrix of the size corresponding to the variables in the vector \( \Phi \). This is equivalent to the single convective velocity component \( A = U \) in a scalar problem and is given as

\[
A_i \equiv \frac{\partial \mathbf{F}_i}{\partial \Phi}
\]

\[(2.138)\]
Convection dominated problems

This in general may still be a function of \( \Phi \), thus destroying the linearity of the problem.

Before proceeding further, it is of interest to discuss the general behaviour of Eq. (2.1) in the absence of source and diffusion terms. We note that the matrices \( \Lambda_i \), if the matrices \( X_i \) are such that

\[
X_i = X
\]

which is always the case in a single dimension, then Eq. (2.137) can be written (in the absence of diffusion or source terms) as

\[
\frac{\partial \Phi}{\partial t} + \mathbf{X} \Lambda_i X^{-1} \frac{\partial \Phi}{\partial x_i} = 0
\]

Premultiplying by \( X^{-1} \) and introducing new variables (called Riemann invariants) such that

\[
\phi = X^{-1} \Phi
\]

we can write the above as a set of decoupled equations in components \( \phi \) of \( \Phi \) and corresponding \( \Lambda \) of \( \Lambda \):

\[
\frac{\partial \phi}{\partial t} + \Lambda_i \frac{\partial \phi}{\partial x_i} = 0
\]

each of which represents a wave-type equation of the form that we have previously discussed. A typical example of the above results from a one-dimensional elastic dynamics problem describing stress waves in a bar in terms of stresses \( (\sigma) \) and velocities \( (v) \) as

\[
\frac{\partial \sigma}{\partial t} - E \frac{\partial v}{\partial x} = 0
\]

\[
\frac{\partial v}{\partial t} - \frac{1}{\rho} \frac{\partial \sigma}{\partial x} = 0
\]

This can be written in the standard form of Eq. (2.1) with

\[
\Phi = \{ \sigma \ \nu \} \quad F = \{ E \nu \ \sigma/\rho \}
\]

The two variables of Eq. (2.142) become

\[
\phi_1 = \sigma - cv \quad \phi_2 = \sigma + cv
\]

where \( c = \sqrt{E/\rho} \) and the equations corresponding to (2.143) are

\[
\frac{\partial \phi_1}{\partial t} + c \frac{\partial \phi_1}{\partial x} = 0
\]

\[
\frac{\partial \phi_2}{\partial t} - c \frac{\partial \phi_2}{\partial x} = 0
\]

representing respectively two waves moving with velocities \( \pm c \).
Unfortunately the condition of Eq. (2.140) seldom pertains and hence the determination of general characteristics and therefore decoupling is not usually possible for more than one space dimension. This is the main reason why the extension of the simple, direct procedures is not generally possible for vector variables. Because of this we shall in Chapter 3 only use the upwinding characteristic-based procedures on scalar systems for which a single wave speed exists and this retains justification of any method proposed.

### 2.11 Summary and concluding remarks

The reader may well be confused by the variety of apparently unrelated approaches given in this chapter. This may be excused by the fact that optimality guaranteed by the finite element approaches in elliptic, self adjoint problems does not automatically transfer to hyperbolic non-self adjoint ones.

The major part of this chapter is concerned with a scalar variable in the convection-diffusion-reaction equation. The several procedures presented for steady-state and transient equations yield identical results. However the characteristic-Galerkin method is optimal for transient problems and gives identical stabilizing terms to that derived by the use of Petrov-Galerkin, GLS and other procedures when the time step used is near the stability limit. For such a problem the optimality is assured simply by splitting the problem into the self-adjoint part where the direct Galerkin approximation is optimal and an advective motion where the unknown variable remains fixed in the characteristic space.

Extension of the various procedures presented to vector variables has been made in the past and we have presented the Taylor-Galerkin method in this context; however its justification is more problematic. For this reason we recommend that when dealing with equations such as those arising in the motion of a fluid an operator split is made in a manner separating several scalar convection-diffusion problems for which the treatment described is used. We shall do so in the next chapter when we introduce the CBS algorithm using the characteristic-based split.

### References

6. O.C. Zienkiewicz, R.H. Gallagher and P. Hood. Newtonian and non-Newtonian viscous incompressible flow. Temperature induced flows and finite element solutions. in *The


