THE UNIVERSITY OF READING

Numerical Variational Techniques on Adjustable Meshes

M.J. Baines

Numerical Analysis Report 4/96

The University of Reading
Department of Mathematics
P O Box 220
Reading RG6 2AX
UK

DEPARTMENT OF MATHEMATICS

Numerical Variational Techniques on Adjustable Meshes

M.J.Baines

Department of Mathematics,
University of Reading

Abstract

In this report we study the extension of some variational methods of interest to the numerical analyst to include the numerical generation of grids as well as approximation of solutions.

1. Introduction

Variational problems of interest to the numerical analyst include that of minimising errors in the approximation of functions as well as minimising an energy in the approximate solution of elliptic PDEs [1]. Variational principles also exist in fluid mechanics and gasdynamics whereby the equations of motion can be generated by finding stationary points [2]. In this report methods of obtaining optimal approximate solutions and meshes for variational problems are discussed together with properties of associated differential equations. The main interest is in the generation of optimal meshes for steady problems.

In Section 2 some standard variational analysis is given together with examples. The analysis is then repeated with stretching of the abscissa is allowed. In the subsequent section the finite-dimensional case is considered including weak forms and finite element approximations. Matrix-vector forms are displayed which show clearly the structure of the analysis. Section 4 is concerned with grid adaptivity in the finite-dimensional case, including weak forms, matrix-vector forms and adaptive finite elements. The relationship with the Moving Finite Element method [4] is brought out in Section 5.

In Section 6 the argument is repeated in the context of finite differences. The functionals are then augmented in Section 7 to include iteration of the solution (in pseudo-time) of implicit time stepping methods.

In Section 8 algorithms are developed for exploiting the error reduction property studied in the previous sections. Finally, in Section 9 the relationship between the methods described here and equidistribution is discussed.

2. Background

2.1. Variational analysis

Let $u(x,t)$ be a function twice differentiable in the space variable x and once differentiable in the timelike variable t (here used as an iteration variable) and let $F(x,u,u_x)$ be a once differentiable function of its arguments. Define the functional

$$\mathcal{I}(u) = \int_a^b F(x,u,u_x)dx \tag{2.1}$$

whose value at time t is

$$I(t) = \int_a^b F(x,u(x,t),u_x(x,t))dx. \tag{2.2}$$

Two examples are the quadratic functions

$$F(x,u,u_x) = \frac{1}{2}k(x)\left(u - f(x)\right)^2 + \frac{1}{2}D(x)\left(u_x - g_x(x)\right)^2, \tag{2.3}$$

where $k(x) \geq 0, D(x) \geq 0$, and

$$F(x,u,u_x) = \frac{1}{2}\left(p(x)u_x^2 + q(x)u^2\right) - r(x)u, \tag{2.4}$$

where $p(x) > 0, q(x) \geq 0, r(x) \geq 0$. A non-quadratic example from Shallow Water flow in a channel (in which u is the depth) is

$$F(x,u) = B(x)\left(\frac{Q^2}{2u} - \frac{1}{2}gu^2 + E(x)u\right) \tag{2.5}$$

where Q, g are positive constants and $B(x), E(x)$ are given (breadth and energy) functions. The function F is convex if $u^3 - Q^2/g > 0$ (supercritical) and concave if $u^3 - Q^2/g < 0$ (subcritical) but switches when this quantity passes through zero.

Differentiating $I(t)$ we have

$$\begin{aligned}\frac{dI}{dt} &= \int_a^b \left(\frac{\partial F}{\partial u}u_t + \frac{\partial F}{\partial u_x}u_{xt}\right)dx \\ &= \int_a^b \left(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u_x}\right)u_t dx,\end{aligned} \tag{2.6}$$

using integration by parts and assuming that $u_t(a) = u_t(b) = 0$. (Note that $\frac{\partial F}{\partial u_x}$ is a function of u and u_x which are both differentiable with respect to x.) If $\frac{dI}{dt} = 0$ for all u_t then u satisfies the PDE

$$\frac{\partial F}{\partial u} = \frac{d}{dx}\frac{\partial F}{\partial u_x}. \tag{2.7}$$

We consider situations in which the integral (2.6) is non-positive so that $I(t)$ is non-increasing. In particular this is true if u satisfies the time-dependent PDE

$$u_t = -\frac{\partial F}{\partial u} + \frac{d}{dx}\frac{\partial F}{\partial u_x} \tag{2.8}$$

with $u(a) = A, u(b) = B$, constants. Then, from (2.6),

$$\frac{dI}{dt} = -\int_a^b u_t^2 dx \leq 0 \tag{2.9}$$

with equality only if $u_t = 0$, i.e. when u satisfies (2.7). Hence I decreases strictly with time unless (2.7) holds, in which case it is stationary. If F is bounded below then so is I which must therefore tend to a limit. At a limit point $\frac{dI}{dt} = 0$ and $u_t = 0$; we are at steady state and (2.7) holds. These statements are also true when the right hand side of (2.8) is multiplied by a positive constant.

The argument parallels the derivation of the Euler equation for the stationary value of $\mathcal{I}(u)$, for which the first variation

$$\delta\mathcal{I} = \int_a^b \left(\frac{\partial F}{\partial u}\delta u + \frac{\partial F}{\partial u_x}\delta u_x\right) dx \tag{2.10}$$

$$= \int_a^b \left(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u_x}\right) \delta u dx$$

assuming that $\delta u(a) = \delta u(b) = 0$. By Lagrange's lemma, if $\delta\mathcal{I} = 0$ for all δu then (2.7) holds. Moreover, selecting the u variation such that

$$\delta u = \left(-\frac{\partial F}{\partial u} + \frac{d}{dx}\frac{\partial F}{\partial u_x}\right) \delta\tau, \tag{2.11}$$

where $\delta\tau$ is a positive constant, gives

$$\delta\mathcal{I} = -\delta\tau \int_a^b (\delta u)^2 dx \leq 0 \tag{2.12}$$

with zero only if $\delta u = 0$, i.e. only if (2.7) holds. The two arguments are identical if the displacements in the functions are thought of as being brought about in an infinitesimal time δt (or $\delta\tau$).

These properties also hold in higher dimensions. If $u(\mathbf{x}, t)$ is a function twice differentiable in the components of the vector space variable $\mathbf{x}$ and once differentiable in t, and $F(\mathbf{x}, u, \nabla u)$ is a once differentiable function of its arguments in some domain Ω of $\mathbf{x}$ space, then

$$I(t) = \int_\Omega F(\mathbf{x}, u(\mathbf{x}, t), \nabla u(\mathbf{x}, t)) d\Omega$$

and

$$\frac{dI}{dt} = \int_\Omega \left(\frac{\partial F}{\partial u} u_t + \frac{\partial F}{\partial \nabla u} \nabla u_t \right) d\Omega$$

$$= \int_\Omega \left(\frac{\partial F}{\partial u} - \nabla . \frac{\partial F}{\partial \nabla u} \right) u_t d\Omega \tag{2.13}$$

provided that $u_t = 0$ on the boundary $\partial\Omega$ of Ω. If $\frac{dI}{dt} = 0$ for all u_t then the stationary function u satisfies

$$\frac{\partial F}{\partial u} = \nabla . \frac{\partial F}{\partial \nabla u}. \tag{2.14}$$

Also, if u satisfies the time-dependent PDE

$$u_t = -\frac{\partial F}{\partial u} + \nabla . \frac{\partial F}{\partial \nabla u} \tag{2.15}$$

in Ω, then I is a non-increasing function of t, stationary only when (2.14) holds. Assuming that the functional F is bounded below then I approaches a limit as $t \to \infty$ at which $\frac{dI}{dt} = 0$ corresponding to a solution of the steady state equation (2.14). An example is

$$F(x, u, \nabla u) = \frac{1}{2} k(\mathbf{x}) \left(u - f(\mathbf{x}) \right)^2 + \frac{1}{2} D(\mathbf{x}) \left(\nabla (u - g(\mathbf{x})) \right)^2 . \tag{2.16}$$

In a similar way the variation

$$\delta u = \left(-\frac{\partial F}{\partial u} + \nabla . \frac{\partial F}{\partial \nabla u} \right) \delta\tau \tag{2.17}$$

(where $\delta\tau$ is positive) induces a non-positive variation of $\mathcal{I}(u)$, zero only when (2.14) holds.

2.2. Examples

(i) An example of a function F which is bounded below in this way is the convex functional

$$F(\mathbf{x}, u, \nabla u) = \frac{1}{2} \left(u^2 + (\nabla u)^2 \right) \tag{2.18}$$

for which $\mathcal{I}(u)$ is the Sobolev norm

$$\mathcal{I}(u) = \frac{1}{2} \left(a(u, u) + b(u, u) \right) \tag{2.19}$$

where

$$a(u, v) = \int_\Omega \nabla u . \nabla v d\Omega, \; b(u, v) = \int_\Omega uv d\Omega.$$

In terms of t, $\mathcal{I}(u)$ takes the value

$$I(t) = \frac{1}{2}\int_\Omega \left(u(t)^2 + \nabla u(t)^2\right) d\Omega$$

and from (2.14) the stationary value of u when $\frac{dI}{dt} = 0$ satisfies

$$\nabla^2 u = u. \tag{2.20}$$

Moreover, if u satisfies the PDE $u_t = \nabla^2 u - u$, $I(t)$ is a non-increasing function, stationary only if $u_t = 0$ when (3.2) holds. (In this particular case, if u satisfies $u_t = -cu$ where c is a positive constant, so that $\nabla u_t = -c\nabla u$, we have

$$\frac{dI}{dt} = \int_\Omega (uu_t + \nabla u.\nabla u_t)\, d\Omega = -c\int_\Omega \left(u^2 + (\nabla u)^2\right) d\Omega = -2cI$$

in which case $I(t)$ decreases exponentially with t.)

In the more general case

$$F(\mathbf{x}, u, \nabla u) = \frac{1}{2}p(\mathbf{x})\,(\nabla u)^2 + \frac{1}{2}q(\mathbf{x})u^2 - r(\mathbf{x})u \tag{2.21}$$

(cf.(2.4)) the equation for the stationary value of u is

$$\nabla.\,(p(\mathbf{x})\nabla u) = q(\mathbf{x})u - r(\mathbf{x}) \tag{2.22}$$

and the PDE for which $I(t)$ is non-increasing is

$$u_t = \nabla.\,(p(\mathbf{x})\nabla u) - q(\mathbf{x})u + r(\mathbf{x}). \tag{2.23}$$

(ii) A similar example is (2.16) for which the stationary value satisfies

$$k(\mathbf{x})(u - f(\mathbf{x})) = \nabla.\,(D(\mathbf{x})\nabla(u - g(\mathbf{x})) \tag{2.24}$$

and the PDE for which $I(t)$ is a non-increasing function of t is

$$u_t = -k(\mathbf{x})(u - f(\mathbf{x})) + \nabla.\,(D(\mathbf{x})\nabla(u - g(\mathbf{x}))\,. \tag{2.25}$$

(iii) In the case of the shallow water example (2.5) the stationary function u satisfies the algebraic equation

$$-\frac{Q^2}{2u^2} - gu + E(x) = 0 \tag{2.26}$$

and the differential equation for which $I(t)$ is non-increasing is

$$u_t = B(x)\left(\frac{Q^2}{2u^2} + gu - E(x)\right) \tag{2.27}$$

(in the supercritical case where $u^3 > Q^3/g$).

(iv) Another nonlinear example is

$$F(\mathbf{x}, u, \nabla u) = \frac{1}{4}u^4 + \frac{1}{2}\,(\nabla u)^2 \tag{2.28}$$

for which the stationary function satisfies $\nabla^2 u = u^3$ and I is non-increasing if $u_t = \nabla^2 u - u^3$.

2.3. Stretching of the abscissae

Suppose now that the x variable participates in the variation through dependence on t. We make the transformation

$$x = \widehat{x}(\xi,\tau),\ t = \tau,\ u(x,t) = u(\widehat{x}(\xi,\tau),\tau) = \widehat{u}(\xi,\tau) \tag{2.29}$$

where ξ is an unstretched reference variable and τ is also timelike. By the use of the chain rule

$$\widehat{u}_\tau = u_t + u_x \widehat{x}_\tau, \tag{2.30}$$

$$\widehat{u}_\xi = u_x \widehat{x}_\xi. \tag{2.31}$$

(An interpretation of the first of these equations is that u_t is a Lagrangian derivative which is converted into the Eulerian derivative $\widehat{u}_\tau$ by the addition of the $\widehat{x}_\tau$ term.)

Replacing x, u by $\widehat{x}$,$\widehat{u}$ in the function F of (2.2) gives

$$I(\tau) = \int_{\xi_a}^{\xi_b} F\left(\widehat{x}, \widehat{u}, \frac{\widehat{u}_\xi}{\widehat{x}_\xi}\right) \widehat{x}_\xi d\xi. \tag{2.32}$$

Then it can be shown that

$$\frac{dI}{d\tau} = \int_a^b \left(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u_x}\right)(\dot{u} - u_x\, \dot{x})\, dx \tag{2.33}$$

where $\dot{u} = u_\tau$ and $\dot{x} = x_\tau$, i.e.

$$\frac{dI}{d\tau} = \int_a^b \left(\frac{\partial F}{\partial u} - \frac{d}{dx}\frac{\partial F}{\partial u_x}\right) u_t dx. \tag{2.34}$$

If $\frac{dI}{d\tau} = 0\ \forall\ \dot{u}, \dot{x}$ then the stationary values of $\widehat{u}$ and $\widehat{x}$ satisfy

$$\frac{\partial F}{\partial u} = \frac{d}{dx}\frac{\partial F}{\partial u_x} \tag{2.35}$$

twice, suggesting higher-order contact at the stationary point. Moreover, if $\widehat{u}$ and $\widehat{x}$ satisfy the τ-dependent PDEs

$$\dot{u} = -\frac{\partial F}{\partial u} + \frac{d}{dx}\frac{\partial F}{\partial u_x} \tag{2.36}$$

and

$$\dot{x} = -u_x\left(-\frac{\partial F}{\partial u} + \frac{d}{dx}\frac{\partial F}{\partial u_x}\right) \tag{2.37}$$

then from (2.33) $I(t)$ is a non-increasing function, zero only if $\dot{u} = \dot{x} = 0$ when (2.35) holds. If F is bounded below I tends to a limit as $\tau \to \infty$. In a similar way the corresponding increment version gives

$$\delta I = \int \left(-\frac{\partial F}{\partial u} + \frac{d}{dx}\frac{\partial F}{\partial u_x}\right)(\delta\widehat{u} - u_x \delta\widehat{x})\, dx.$$

If F contains an isolated discontinuity at $x = s$, say, but is otherwise differentiable as before, then $\frac{dI}{d\tau}$ will also contain terms from the variation of s of the form

$$\left[F - u_x \frac{\partial F}{\partial u_x}\right]_s s_\tau + \left[\frac{\partial F}{\partial u_x}\right]_s u_\tau \tag{2.38}$$

where $[F]_s$ denotes the jump in F in passing from $x = s-$ to $x = s+$. The jump notation becomes important when we consider piecewise continuous functions in the numerical investigation which follows.

In higher dimensions it can be shown in the same way that

$$\frac{dI}{d\tau} = \int_\Omega \left(\frac{\partial F}{\partial u} - \nabla \frac{\partial F}{\partial \nabla u}\right)(\widehat{u}_\tau - \nabla u.\widehat{\mathbf{x}}_\tau)\, d\Omega = \int_\Omega \left(\frac{\partial F}{\partial u} - \nabla \frac{\partial F}{\partial \nabla u}\right) u_t d\Omega \tag{2.39}$$

(cf. (2.33),(2.34)). If $\frac{dI}{d\tau} = 0\ \forall \widehat{u}_\tau, \widehat{\mathbf{x}}_\tau$ then the the stationary values of $\widehat{u}$ and $\widehat{x}$ satisfy

$$\frac{\partial F}{\partial u} = \nabla . \frac{\partial F}{\partial \nabla u} \tag{2.40}$$

at least twice, once again suggesting higher-order contact at the stationary point. Moreover, if $\widehat{u}$,$\widehat{\mathbf{x}}$ satisfy the τ-dependent PDEs

$$\dot{u} = -\frac{\partial F}{\partial u} + \nabla . \frac{\partial F}{\partial \nabla u} \tag{2.41}$$

and

$$\dot{\mathbf{x}} = -\left(-\frac{\partial F}{\partial u} + \nabla . \frac{\partial F}{\partial \nabla u}\right) \nabla u \tag{2.42}$$

then I is a non-increasing function of τ, zero only if $\dot{u}$=$\dot{\mathbf{x}}$= 0 when (2.40) holds. For F bounded below we again have I tending to a limit as $\tau \to \infty$.

3. The Finite-Dimensional Case and Finite Elements

Suppose now that in one dimension the function $u(x,t)$ is written in terms of a finite number of basis functions $\psi(x)$ $(j = 1, 2, ..., J)$ as

$$u(x,t) \approx U(x,t) = \sum_{j=1}^{J} U_j(t)\psi_j(x) \tag{3.1}$$

where $U_t(a) = U_t(b) = 0$. Then, assuming that $\psi_j(x)$ is piecewise twice differentiable and that $U_j(t)$ is differentiable, and with F, I and $\mathcal{I}$ defined as in section 2.1 with u replaced by U,

$$\frac{dI}{dt} = \sum_{j=1}^{J} \frac{\partial \mathcal{I}}{\partial U_j} \dot{U}_j = \sum_{j=1}^{J} \dot{U}_j \int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) \psi_j(x) dx \tag{3.2}$$

(cf. (2.6)), where the notation $\frac{\partial F}{\partial U}$ stands for $\frac{\partial F}{\partial u}\Big|_{u=U}$. If $\frac{dI}{dt}$ is zero $\forall\ \dot{U}_j$ then for each j the stationary value of the approximation U satisfies

$$\int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx}\frac{\partial F}{\partial U_x}\right)\psi_j(x)dx = 0. \tag{3.3}$$

If every term in the sum in (3.2) is non-positive then $I(t)$ is a non-increasing function of t and this will be true if for example

$$\dot{U}_j\,(t) = \int_a^b \left(-\frac{\partial F}{\partial U} + \frac{d}{dx}\frac{\partial F}{\partial U_x}\right)\psi_j(x)dx \tag{3.4}$$

$\forall j$, in which case

$$\frac{dI}{dt} = -\sum_{j=1}^{J} \dot{U}_j^2 \leq 0,$$

zero only if $\dot{U}_j = 0\ \forall j$ when (3.3) holds. Alternatively, if in a finite element context U satisfies the Galerkin form of the differential equation (2.8) in the form

$$\int_a^b U_t\psi_i(x)dx = \int_a^b \left(-\frac{\partial F}{\partial U} + \frac{d}{dx}\frac{\partial F}{\partial U_x}\right)\psi_i(x)dx \tag{3.5}$$

$\forall i$, then since (from(3.1))

$$U_t = \sum_j \dot{U}_j\,(t)\psi_j(x)$$

equation (3.2) becomes

$$\frac{dI}{dt} = -\sum_{j=1}^{J} \dot{U}_j\,(t)\int_a^b U_t\psi_j(x)dx = -\sum_{j=1}^{J}\sum_{i=1}^{J} \dot{U}_j\,(t)\,\dot{U}_i\,(t)\int_a^b \psi_j(x)\psi_i(x)dx \tag{3.6}$$

so that $I(t)$ is again a non-increasing function provided that the quadratic form in (3.6) is non-negative definite.

3.1. Weak Forms

Similar results hold if U is only once differentiable in x, in which case $\frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right)$ may not exist. In that case we refrain from integrating by parts in deriving $\frac{\partial \mathcal{I}}{\partial U_j}$, giving instead

$$\frac{\partial \mathcal{I}}{\partial U_j} = \int_a^b \left(\frac{\partial F}{\partial U}\psi_j + \frac{\partial F}{\partial U_x}\psi_j'\right)dx$$

so that (3.2) becomes

$$\frac{dI}{dt} = \sum_{j=1}^{J} \frac{\partial \mathcal{I}}{\partial U_j}\,\dot{U}_j = \sum_{j=1}^{J} \dot{U}_j \int_a^b \left(\frac{\partial F}{\partial U}\psi_j(x) + \frac{\partial F}{\partial U_x}\psi_j'(x)\right)dx \tag{3.7}$$

$$= -\sum_{j=1}^{J} U_j b_j(U), \tag{3.8}$$

say, where

$$b_j(U) = -\int_a^b \left(\frac{\partial F}{\partial U}\psi_j(x) + \frac{\partial F}{\partial U_x}\psi_j'(x) \right) dx. \tag{3.9}$$

If $\frac{dI}{dt} = 0 \ \forall \ \dot{U}_j$ then from (3.7) the stationary function U must satisfy

$$\int_a^b \left(\frac{\partial F}{\partial U}\psi_j(x) + \frac{\partial F}{\partial U_x}\psi_j'(x) \right) dx = 0. \tag{3.10}$$

If we suppose that U_j satisfies the time-dependent PDE

$$\dot{U}_j = b_j(U) \tag{3.11}$$

$\forall j$ or if, in the Galerkin form,

$$\int_a^b U_t \psi_j(x) dx = b_j(U) \tag{3.12}$$

$\forall j$, then from (3.8) (provided that the resulting mass matrix in (3.12) is positive definite) $I(t)$ is a non-increasing function of t, stationary only when $\dot{U}_j = 0$ when the weak form (3.10) holds. With F bounded below I tends to a limit as $t \to \infty$ at which U satisfies (3.10).

3.2. Algebraic Forms

To see these results in matrix-vector form, denote by $\mathbf{U}(t)$ the vector of coefficients $U_j(t)$ and by $\mathbf{b}(\mathbf{U})$ the vector of coefficients $b_j(U)$. Then equation (3.10) may be written

$$\mathbf{b}(\mathbf{U}) = \mathbf{0}. \tag{3.13}$$

(see (3.9)). Also equation (3.11) takes the form

$$\dot{\mathbf{U}} = \mathbf{b}(\mathbf{U}) \tag{3.14}$$

in which case, defining $\mathrm{I}(\mathbf{U}) = \mathcal{I}(U) = I(t)$, equation (3.8) becomes

$$\frac{d\mathrm{I}(\mathbf{U})}{dt} = -\dot{\mathbf{U}}^T \mathbf{b}(\mathbf{U}). \tag{3.15}$$

A feature of equations (3.15) is that $\mathbf{b}(\mathbf{U})$ is a search direction for the minimisation of $\mathrm{I}(\mathbf{U})$ by the technique of steepest descent.

If U is given by (3.11),

$$\frac{d\mathrm{I}(\mathbf{U})}{dt} = -\left\| \dot{\mathbf{U}} \right\|^2 \tag{3.16}$$

then $\mathrm{I}(\mathbf{U})$ is a non-increasing function of t, stationary only if $\dot{\mathbf{U}}= \mathbf{0}$. Alternatively, denoting by A the matrix with elements

$$A_{ij} = \int_a^b \psi_i(x)\psi_j(x)dx, \tag{3.17}$$

equation (3.12) may be written

$$A\,\dot{\mathbf{U}}= \mathbf{b}(\mathbf{U}) \tag{3.18}$$

(cf. (3.14)), so that in this case

$$\frac{d\mathrm{I}(\mathbf{U})}{dt} = -\,\dot{\mathbf{U}}^T\,\mathbf{b}(\mathbf{U}) = -\,\dot{\mathbf{U}}^T\,A\,\dot{\mathbf{U}} \tag{3.19}$$

and again, provided that A is positive definite, $\mathrm{I}(\mathbf{U})$ is a decreasing function of t, stationary only if $\dot{\mathbf{U}}= \mathbf{0}$ when (3.13) holds. If F is bounded below, $\mathrm{I}(\mathbf{U})$ tends in both cases to a limit as $t \to \infty$ at which $\dot{\mathbf{U}}= \mathbf{0}$, i.e. $\mathbf{U}$ is a solution of (3.13).

If we are only interested in convergence as $t \to \infty$, (3.16) and (3.19) are equally valid as equations which possess the correct limit. Indeed, all we need is an equation of the form of (3.18) with a positive definite A. In practice A will need to be inverted to obtain a convergent sequence of $\dot{U}$'s, so it makes sense to choose a matrix A which is easily invertible. The unit matrix is the one used in (3.16) but this choice lacks any useful scaling properties which are possessed by (3.19). A good compromise is therefore to choose the diagonal of the matrix A in (3.19). We shall therefore also consider the discretisation for which

$$D\,\dot{\mathbf{U}}\,(t) = \mathbf{b}(\mathbf{U}) \tag{3.20}$$

where $D = diag\{A\}$, which may be thought of as being brought about by tampering with the test function on the left hand side of (3.12).

An important special case is where $\psi_j(x)$ are the once differentiable piecewise linear finite element hat functions, $\alpha_j(x)$ say, and U is the Galerkin linear finite element approximation (with nodal values $U_j(t)$), satisfying (3.12) with ψ replaced by α. Then A is the tridiagonal positive definite finite element mass matrix $A_\alpha = \{A_{ij}\}$ with

$$A_{ij} = \int_a^b \alpha_i(x)\alpha_j(x)dx \tag{3.21}$$

which may readily be inverted using preconditioned conjugate gradients (the matrix D^{-1} being a useful preconditioner).

3.3. Higher Dimensions

The results extend to higher dimensions as in Section 2.1. If $F = F(\mathbf{x}, U, \nabla U)$ and

$$u \sim U = \sum_{j=1}^{J} U_j(t)\psi_j(\mathbf{x})$$

then, as in (3.7),

$$\frac{dI}{dt} = \sum_{j=1}^{J} \frac{\partial \mathcal{I}}{\partial U_j} \dot{U}_j = \sum_{j=1}^{J} \dot{U}_j \int_\Omega \left(\frac{\partial F}{\partial U} \psi_j(\mathbf{x}) + \frac{\partial F}{\partial \nabla U} . \nabla \psi_j(\mathbf{x}) \right) d\Omega \tag{3.22}$$

$$= - \sum_{j=1}^{J} \dot{U}_j \; b_j(U) \tag{3.23}$$

where

$$b_j(U) = - \int_\Omega \left(\frac{\partial F}{\partial U} \psi_j(\mathbf{x}) + \frac{\partial F}{\partial \nabla U} . \nabla \psi_j(\mathbf{x}) \right) d\Omega. \tag{3.24}$$

If (3.22) vanishes for all $\dot{U}_j$ then the stationary function satisfies

$$\int_\Omega \left(\frac{\partial F}{\partial U} \psi_j(\mathbf{x}) + \frac{\partial F}{\partial \nabla U} . \nabla \psi_j(\mathbf{x}) \right) d\Omega = 0. \tag{3.25}$$

Moreover, if $U_j(t)$ is chosen to satisfy

$$\dot{U}_j = b_j(U) \tag{3.26}$$

or the weak form

$$\int_\Omega U_t \psi_j(\mathbf{x}) d\Omega = b_j(U) \tag{3.27}$$

$\forall j$, then, provided that the matrix $A = \{A_{ij}\}$ where

$$A_{ij} = \int_\Omega \psi_i(\mathbf{x}) \psi_j(\mathbf{x}) d\Omega \tag{3.28}$$

is positive definite, the function $I(t)$ is a non-increasing function of t, stationary only if $\dot{U}_j = 0$ so that $U_t = 0$ and (3.25) holds. If F is bounded below, then so is I which therefore tends to a limit as $t \to \infty$ at which U is a solution of (3.25). Linear finite elements again provide the main example, for which the matrix A is readily inverted in a similar manner to the one-dimensional case.

The matrix-vector forms of Section 3.2 hold as before.

4. Adaptivity in the Finite-Dimensional Case

The adaptive mesh methods we consider are finite-dimensional versions of the variational methods with stretched abscissae considered in Section 2. It is convenient to expand each of the functions $u \sim U$ and $x \sim X$ in two frames of reference, in terms of the physical and computational coordinates x and ξ respectively, as

$$U = \sum_{j=1}^{J} U_j(t) \psi_j(x) = \sum_{j=1}^{J} \widehat{U}_j(\tau) \widehat{\psi}_j(\xi) \tag{4.1}$$

and

$$X = \sum_{j=1}^{J} X_j(t)\psi_j(x) = \sum_{j=1}^{J} \widehat{X}_j(\tau)\widehat{\psi}_j(\xi). \tag{4.2}$$

Then, assuming that the ψ_j and $\widehat{\psi_j}$ functions are piecewise twice differentiable and that the U_j and $\widehat{U}_j$ functions are differentiable, and with $I(\tau)$ defined as in Section 2.3 with u, x replaced by U, X, then as in the derivation of (3.2),

$$\frac{dI}{d\tau} = \int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) \sum_j (\dot{U}_j - U_x \dot{X}_j)\psi_j(x)dx \tag{4.3}$$

where $\dot{U}_j = \frac{d\widehat{U}_j}{d\tau}, \dot{X}_j = \frac{d\widehat{X}_j}{d\tau}$.

If $\frac{dI}{d\tau} = 0$ for all $\dot{U}_j, \dot{X}_j$, then the stationary values of U and X satisfy the weak forms

$$\int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) \psi_j(x)dx = 0, \tag{4.4}$$

$$\int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) (-U_x)\psi_j(x)dx = 0. \tag{4.5}$$

Moreover, putting

$$\dot{U}_j = \int_a^b \left(-\frac{\partial F}{\partial U} + \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) \psi_j(x)dx \tag{4.6}$$

$$\dot{X}_j = \int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) (-U_x)\psi_j(x)dx \tag{4.7}$$

ensures that $I(t)$ is a non-increasing function of t, stationary only if $\dot{U}_j = \dot{X}_j = 0$ when (4.4) and (4.5) hold. Hence, if F is bounded below, I tends to limit as $t \to \infty$. Alternatively, the Galerkin forms

$$\int_a^b U_t\psi_j(x)dx = \int_a^b \left(-\frac{\partial F}{\partial U} + \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) \psi_j(x)dx \tag{4.8}$$

$$\int_a^b U_t(-U_x)\psi_j(x)dx = \int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx}\frac{\partial F}{\partial U_x}\right) (-U_x)\psi_j(x)dx \tag{4.9}$$

have the same property, provided that the resulting mass matrix is positive definite.

4.1. Weak Forms

If U is only once differentiable in x then $\frac{d}{dx}\frac{\partial F}{\partial U_x}$ may not exist but it can still be shown [3] that

$$\frac{dI}{d\tau} = \int_{\xi_a}^{\xi_b} \left\{ \left(\frac{\partial F}{\partial X}X_\tau + \frac{\partial F}{\partial U}U_\tau + \frac{\partial F}{\partial U_x}U_{x\tau}\right) X_\xi + FX_{\xi\tau} \right\} d\xi$$

$$= \int_a^b \sum_j \left(\frac{\partial F}{\partial X}\, \dot{X}_j + \frac{\partial F}{\partial U}\, \dot{U}_j \right) \psi_j(x)dx + \int_a^b \sum_j \left(\frac{\partial F}{\partial U_x}\, \dot{U}_j + F\, \dot{X}_j \right) \psi_j'(x)dx$$

$$= \sum_j \int_a^b \left\{ \left(\frac{\partial F}{\partial U}\psi_j(x) + \frac{\partial F}{\partial U_x}\psi_j'(x) \right) \dot{U}_j + \left(\frac{\partial F}{\partial X}\psi_j(x) + F\psi_j'(x) \right) \dot{X}_j \right\} dx \tag{4.10}$$

$$= -\sum_j \left(\dot{U}_j\, b_j(U,X) + \dot{X}_j\, \theta_j(U,X) \right), \tag{4.11}$$

where

$$b_j(U,X) = -\int_a^b \left(\frac{\partial F}{\partial U}\psi_j(x) + \frac{\partial F}{\partial U_x}\psi_j'(x) \right) dx \tag{4.12}$$

$$\theta_j(U,X) = -\int_a^b \left(\frac{\partial F}{\partial X}\psi_j(x) + F\psi_j'(x) \right) dx. \tag{4.13}$$

In this case, if $\frac{dI}{d\tau} = 0$ for all $\dot{X}_j, \dot{U}_j$ the stationary functions U, X satisfy

$$\int_a^b \left(\frac{\partial F}{\partial U}\psi_j(x) + \frac{\partial F}{\partial U_x}\psi_j'(x) \right) dx = \int_a^b \left(\frac{\partial F}{\partial X}\psi_j(x) + F\psi_j'(x) \right) dx = 0. \tag{4.14}$$

Moreover, if we define

$$\dot{U}_j = b_j(U,X), \tag{4.15}$$

$$\dot{X}_j = \theta_j(U,X), \tag{4.16}$$

$\forall j$, then from (4.11) $I(t)$ is a non-increasing function of τ, stationary only if $\dot{U}_j = \dot{X}_j = 0$ when (4.14) holds. Similarly, the Galerkin forms

$$\int_a^b U_t\psi_j(x)dx = \sum_i \int_a^b \left(\dot{U}_j - U_x\, \dot{X}_j \right) \psi_i(x)\psi_j(x)dx = b_j(U,X), \tag{4.17}$$

$$\int_a^b U_t(-U_x)\psi_j(x)dx = \sum_i \int_a^b \left(\dot{U}_j - U_x\, \dot{X}_j \right) (-U_x)\psi_i(x)\psi_j(x)dx = \theta_j(U,X), \tag{4.18}$$

have the same property, provided that the corresponding mass matrix is positive definite. In each of these cases, if F is bounded below I tends to a limit as $\tau \to \infty$.

If F does not depend on U_x equation (4.11) reduces to

$$\frac{dI}{d\tau} = \sum_j \int_a^b \left\{ \frac{\partial F}{\partial U}\psi_j(x)\, \dot{U}_j + \left(\frac{\partial F}{\partial X}\psi_j(x) + F\psi_j'(x) \right) \dot{X}_j \right\} dx.$$

$$= \sum_j \left\{ \int_a^b \frac{\partial F}{\partial U}\, (\dot{U}_j - U_x\, \dot{X}_j)\psi_j(x)dx + [F\psi_j(x)]_j\, \dot{X}_j \right\} \tag{4.19}$$

since

$$\frac{\partial F}{\partial X} = \frac{dF}{dX} - \frac{\partial F}{\partial U} U_x. \tag{4.20}$$

The boundary terms in (4.19) arise from the integration by parts when F and/or ψ are discontinuous, which is allowable in this case. The stationary functions satisfy the weak forms

$$\int_a^b \frac{\partial F}{\partial U} \psi_j(x) dx = 0 \tag{4.21}$$

(c.f.(3.10)) and

$$\int_a^b \frac{\partial F}{\partial U} (-U_x) \psi_j(x) dx + [F \psi_j(x)]_j = 0. \tag{4.22}$$

If, on the other hand, F is independent of U, the corresponding weak forms are

$$\int_a^b \frac{\partial F}{\partial U_x} \psi_j'(x) dx = 0 \tag{4.23}$$

and the second of (4.14) again.

4.2. Algebraic Forms

In matrix-vector form, writing $\mathbf{X}(\tau)$ as the vector of coefficients $X_j(\tau)$ and $\mathrm{I}(\mathbf{U}, \mathbf{X}) = \mathcal{I}(U, X) = I(t)$, equation (4.11) of the previous section takes the form

$$\frac{d\mathrm{I}(\mathbf{U}, \mathbf{X})}{d\tau} = -\mathbf{b}(\mathbf{U}, \mathbf{X})^T \dot{\mathbf{U}} - \mathbf{\Theta}(\mathbf{U}, \mathbf{X})^T \dot{\mathbf{X}}\,, \tag{4.24}$$

where $\dot{\mathbf{X}} = (\dot{X}_1, \dot{X}_2, ..., \dot{X}_J)$ and $\mathbf{b}(\mathbf{U}, \mathbf{X}) = \{b_j(U, X)\}, \mathbf{\Theta}(\mathbf{U}, \mathbf{X}) = \{\theta_j(U, X)\}$. If $\frac{dI}{d\tau} = 0$ then the algebraic forms of the equations for the stationary values are

$$\mathbf{b}(\mathbf{U}, \mathbf{X}) = \mathbf{\Theta}(\mathbf{U}, \mathbf{X}) = \mathbf{0}. \tag{4.25}$$

Introducing the composite notation

$$\begin{aligned} \mathbf{Y} &= \{U_1, X_1, U_2, X_2, ..., U_J, X_j\}^T, \\ \dot{\mathbf{Y}} &= \{\dot{U}_1, \dot{X}_1, \dot{U}_2, \dot{X}_2, ..., \dot{U}_J, \dot{X}_j\}^T, \\ \mathbf{g}(\mathbf{Y}) &= \{b_1, \theta_1, b_2, \theta_2, ..., b_J, \theta_J\}^T, \end{aligned} \tag{4.26}$$

and writing $\mathrm{I}(\mathbf{U}, \mathbf{X}) = I(\mathbf{Y})$, equation (4.24) may be written concisely in the form

$$\frac{dI(\mathbf{Y})}{d\tau} = -\dot{\mathbf{Y}}^T \mathbf{g}(\mathbf{Y}). \tag{4.27}$$

Equation (4.27) shows that $\mathbf{g}(\mathbf{Y})$ may be regarded as a search direction for the minimisation of $I(\mathbf{Y})$ by the method of steepest descent.

Defining

$$\dot{\mathbf{Y}} = \mathbf{g}(\mathbf{Y}) \tag{4.28}$$

ensures that

$$\frac{dI}{d\tau} = -\left\|\dot{\mathbf{Y}}\right\|^2 \leq 0.$$

Alternatively we may use the Galerkin forms (4.17),(4.18) which lead to the matrix system

$$A(\mathbf{Y})\ \dot{\mathbf{Y}} = \mathbf{g}(\mathbf{Y}), \tag{4.29}$$

where $A(\mathbf{Y})$ is a mass matrix $\{A_{ij}\}$ in which each A_{ij} is a block 2×2 submatrix

$$\begin{pmatrix} \int_a^b \psi_i \psi_j dx & \int_a^b (-U_x)\psi_i \psi_j dx \\ \int_a^b (-U_x)\psi_i \psi_j dx & \int_a^b (U_x)^2 \psi_i \psi_j dx \end{pmatrix}. \tag{4.30}$$

Then we have from (4.27)

$$\frac{dI(\mathbf{Y})}{dt} = -\dot{\mathbf{Y}}^T A(\mathbf{Y})\ \dot{\mathbf{Y}}\ . \tag{4.31}$$

Provided that $A(\mathbf{Y})$ is positive definite, $I(\mathbf{Y})$ is a non-increasing function of τ which is zero only if $\dot{\mathbf{Y}} = 0$, i.e. at steady state. If F is bounded below then $I(\mathbf{Y})$ tends to a limit as $\tau \to \infty$, at which $\mathbf{g}(\mathbf{Y}) = 0$, i.e. from (4.26), (4.12) and (4.13), which is equivalent to (4.14).

A third alternative to (4.28) and (4.29) is (cf. (3.20))

$$D(\mathbf{Y})\ \dot{\mathbf{Y}} = \mathbf{g}(\mathbf{Y}), \tag{4.32}$$

where $D(\mathbf{Y})$ is the 2×2 (block) diagonal of A. This matrix is generally trivial to invert.

If the $\psi_j(x)$ are the once differentiable piecewise linear finite element hat functions $\alpha_j(x)$ and U, X are the corresponding Galerkin linear finite element approximations, then A is a block tridiagonal positive semi-definite mass matrix, the so-called MFE matrix (see [4]) with D as its diagonal. Both A and D may be made positive definite by adding regularisation terms to the left hand side of (4.18) (under the appropriate test function).

4.3. Higher Dimensions

In higher dimensions the forms of b_j and θ_j and the stationary values are given by the equations

$$b_j(U, \mathbf{X}) = -\int_\Omega \left(\frac{\partial F}{\partial U}\psi_j(\mathbf{x}) + \frac{\partial F}{\partial \nabla U}.\nabla\psi_j(\mathbf{x}) \right) d\Omega = 0 \tag{4.33}$$

and

$$\theta_j(U, \mathbf{X}) = -\int_\Omega \left\{ (\nabla_{\mathbf{x}} F)\,\psi_j + F\nabla\psi_j - \left(\frac{\partial F}{\partial \nabla U}.\nabla\psi_j \right) \nabla U \right\} d\Omega = 0, \tag{4.34}$$

(cf. (4.12),(4.13)). The PDEs corresponding to (4.15) and (4.16) are

$$\dot{U}_j = b_j(U,X) \text{ and } \dot{X}_j = \theta_j(U,X) \tag{4.35}$$

while the Galerkin weak forms corresponding to (4.17) and (4.18) are

$$\int_\Omega U_t \psi_j d\Omega = b_j(U,\mathbf{X}) \tag{4.36}$$

and

$$\int_\Omega U_t(-\nabla U)\psi d\Omega = \theta_j(U,\mathbf{X}). \tag{4.37}$$

In either case $I(t)$ is non-increasing provided that the corresonding mass matrix is positive definite. In the present case $A = \{A_{ij}\}$ where the blocks A_{ij} are

$$\begin{pmatrix} \int_\Omega \psi_i\psi_j d\Omega & \int_\Omega(-\nabla U)\psi_i\psi_j d\Omega \\ \int_\Omega(-\nabla U)\psi_i\psi_j d\Omega & \int_\Omega(\nabla U)^2\psi_i\psi_j d\Omega \end{pmatrix}. \tag{4.38}$$

If F is bounded below $I(\mathbf{Y})$ again tends to a limit as $t \to \infty$ at which U satisfies (4.33), (4.34).

If the ψ_j are the piecewise linear finite element basis functions α_j (pyramid functions in two dimensions) then A is the positive semi-definite MFE matrix. It may again be made positive definite by adding regularisation terms to the left hand side of (4.37).

The algebraic forms of the previous section hold.

4.4. Examples

(i) If F is given by (2.18), the weak forms (4.33) and (4.34) become

$$\int_\Omega \left(U\psi_j d\Omega + \nabla U.\nabla\psi_j\right) d\Omega = 0 \tag{4.39}$$

and

$$\int_\Omega \left[\frac{1}{2}\left\{U^2 + (\nabla U)^2\right\}\nabla\psi_j - (\nabla U.\nabla\psi_j)\,\nabla U\right] d\Omega = 0. \tag{4.40}$$

(ii) If F is given by (2.16) they are

$$-\int_\Omega \left(k(\mathbf{x})(u - f(\mathbf{x}))\psi_j(\mathbf{x}) + D(\mathbf{x})(\nabla U - \nabla g(\mathbf{x})).\nabla\psi_j(\mathbf{x})\right) d\Omega = 0 \tag{4.41}$$

and

$$-\int_\Omega \Big[(\nabla_{\mathbf{x}} F)\,\psi_j + \left\{k(\mathbf{x})(u - f(\mathbf{x}))^2 + D(\mathbf{x})(\nabla U - \nabla g(\mathbf{x}))^2\right\}\nabla\psi_j -$$

$$\left\{D(\mathbf{x})(\nabla U - \nabla g(\mathbf{x})).\nabla\psi_j\right\}.\nabla U\Big]\, d\Omega = 0. \tag{4.42}$$

If $k(x) = 1$ and $D(x) = 0$ equations (4.33) and (4.34) reduce to

$$\int_\Omega (U - f(\mathbf{x}))\psi_j d\Omega = 0, \tag{4.43}$$

showing that U is the best fit to $f(\mathbf{x})$ with adjustable nodes in the L_2 norm, and

$$\int_\Omega \left[-(U - f(\mathbf{x}))(\nabla f(\mathbf{x}))\psi_j + \frac{1}{2}(U - f(\mathbf{x}))^2 \nabla \psi_j \right] d\Omega = 0$$

which in one dimension becomes from (4.22)

$$-\int_a^b (U - f(x))U_x \psi_j(x)dx + [(U - f(x))^2 \psi_j(x)]_j = 0. \tag{4.44}$$

If $k(x) = 0$ and $D(\mathbf{x}) = 1$ equations (4.33) and (4.34) reduce to

$$\int_\Omega (\nabla U - \nabla g(\mathbf{x})).\nabla \psi_j(\mathbf{x}) d\Omega = 0, \tag{4.45}$$

i.e. U is the best fit to $g(\mathbf{x})$ with adjustable nodes in the H^1 semi-norm, and

$$\int_\Omega [2(\nabla U - \nabla g(\mathbf{x})).\nabla g(\mathbf{x})\psi_j +$$

$$\left\{ (\nabla U - \nabla g(\mathbf{x}))^2 \nabla \psi_j - (\nabla U - \nabla g(\mathbf{x})).\nabla \psi_j \right\} .\nabla U \Big] d\Omega = 0. \tag{4.46}$$

(iii) If F is given by (2.5) then (4.21) and (4.22) give

$$\int_a^b B(x) \left(\frac{Q^2}{2U^2} + gU - E(x) \right) \psi_j(x) dx = 0 \tag{4.47}$$

and

$$\int_a^b B(x) \left(\frac{Q_2}{2U^2} + gU - E(x) \right) U_x \psi_j(x) dx +$$

$$\left[\left(\frac{Q^2}{2U} - \frac{1}{2} g U^2 + E(x)U \right) \psi_j(x) \right]_j = 0. \tag{4.48}$$

Iterations for the solutions of these pairs of equations, based upon $I(t)$ being a non-increasing function, are considered in Section 8. First we discuss the link between the Galerkin forms above and the Moving Finite Element method [4].

5. Moving Finite Elements

In contrast to the method used in Section 4, in the moving finite element approach to adaptation $\alpha_j(x, \mathbf{X}(t))$ denotes the piecewise linear finite element hat functions on the moving grid

$$\mathbf{X}(t) = \{X_1(t), X_2(t), ..., X_J(t)\},$$

so that

$$U = \sum_{j=1}^{J} U_j \alpha_j(x, \mathbf{X}(t)). \tag{5.1}$$

Then it can be shown that

$$U_t = \sum \left(\dot{U}_j \; \alpha_j(x, \mathbf{X}(t)) + \dot{X}_j \; \beta_j(x, \mathbf{X}(t)) \right), \tag{5.2}$$

where the $\beta_j(x, \mathbf{X}(t))$ are so-called second type basis functions satisfying

$$\beta_j = -U_x \alpha_j \tag{5.3}$$

which are also piecewise linear with the same support as $\alpha_j(x, \mathbf{X}(t))$ but are discontinuous at node j. Then

$$\frac{dI}{dt} = \int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx} \frac{\partial F}{\partial U_x} \right) u_t dx$$

$$= \int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx} \frac{\partial F}{\partial U_x} \right) \sum_j (\dot{U}_j \; \alpha_j + \dot{X}_j \; \beta_j) dx \tag{5.4}$$

so that $\frac{dI}{dt} = 0$ implies that

$$\int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx} \frac{\partial F}{\partial U_x} \right) \alpha_j dx = \int_a^b \left(\frac{\partial F}{\partial U} - \frac{d}{dx} \frac{\partial F}{\partial U_x} \right) \beta_j dx = 0. \tag{5.5}$$

If $\frac{d}{dx} \frac{\partial F}{\partial U_x}$ does not exist (which will generally be the case with linear finite elements) we cannot obtain (5.4) but equivalence with the weak forms (4.14) (with ψ replaced by α) has been demonstrated by Jimack [5] using the standard smoothing of U (see [4]). The algebraic forms for (5.5) and the associated PDEs which make $I(t)$ a non-increasing function are therefore the same as before, namely (4.24) - (4.31) where ψ is now replaced by α.

In the two-dimensional case the elements are linears on triangles and the function β_j is a vector which takes the form

$$\beta_j = -(\nabla U) \, \alpha_j \tag{5.6}$$

where α_j is the standard linear finite element basis function (the "pyramid" function in two dimensions) and β_j again has the same support as α_j but is discontinuous at the point j. Once again, equivalence with (4.33) and (4.34) has been shown in [5].

The algebraic form of the MFE equations are again the same as in (4.29) with ψ replaced by α, namely

$$A \, \dot{\mathbf{Y}} = \mathbf{g}(\mathbf{Y}) \tag{5.7}$$

with A modified accordingly. If only the steady limit is of interest, $A(\mathbf{Y})$ may be replaced by $D(\mathbf{Y}) = diag\{A(\mathbf{Y})\}$. The matrices A and D are only positive semi-definite because two rows may be identical if the ∇U's coincide in adjacent elements.

Jimack [5] has used a regularised version of (4.29) to drive $\mathbf{g}(\mathbf{Y})$ to zero, obtaining locally optimal solutions and meshes in a particular example. Although the MFE solution is generally altered by regularisation the steady limit is unaffected. He has also shown that (in the unregularised case) the gradient of $I(\mathbf{Y}) = I(U, X)$ with respect to $\mathbf{Y}$ has the property

$$\nabla_{\mathbf{Y}} I(\mathbf{Y}) = -\mathbf{g}(\mathbf{Y}) \tag{5.8}$$

so that, as long as A is positive definite, from (4.27) and (4.29)

$$\frac{dI(\mathbf{Y})}{dt} = -\dot{\mathbf{Y}}^T \mathbf{g}(\mathbf{Y}) = -\dot{\mathbf{Y}}^T \nabla_{\mathbf{Y}} I(\mathbf{Y}) = -\dot{\mathbf{Y}}^T A^{-1} \dot{\mathbf{Y}} \leq 0. \tag{5.9}$$

It follows that the rate of change of $I(\mathbf{Y})$ is non-positive and also that its magnitude is largest when $\dot{\mathbf{Y}}$ is aligned with $\mathbf{g}(\mathbf{Y})$. Jimack uses (5.8) to prove that a stable steady solution of (4.33),(4.34) is optimal in the sense that the Hessian of $\mathcal{I}$ is positive definite.

It is instructive to rewrite the method as an iterative method in terms of increments. Instead of (5.2) write

$$\delta U = \sum (\delta U_j \alpha_j(x, \mathbf{X}(t)) + \delta X_j \beta_j(x, \mathbf{X}(t)))$$

where δ indicates a small increment (cf. (2.17)). Then (4.27) becomes

$$\delta I(\mathbf{Y}) = -\delta \mathbf{Y}^T \mathbf{g}(\mathbf{Y}) \tag{5.10}$$

where

$$\delta \mathbf{Y} = (\delta U_1, \delta X_1, \delta U_2, \delta X_2, ..., \delta U_J, \delta X_J).$$

The Galerkin weak forms associated with (5.5) in the version of (4.17) and (4.18) become

$$\int_a^b \delta U \alpha_j dx = -\int_a^b \left(\frac{\partial F}{\partial U} \alpha_j + \frac{\partial F}{\partial U_x} \alpha_j' \right) dx \delta\tau \tag{5.11}$$

$$\int_a^b \delta U (-U_x) \alpha_j dx = -\int_a^b \left(\frac{\partial F}{\partial X} \alpha_j + F \alpha_j' \right) dx \delta\tau \tag{5.12}$$

(cf. (4.12),(4.13)) with similar equations in higher dimensions (cf. (4.36),(4.37)). In algebraic terms these can be written as

$$A(\mathbf{Y}) \delta \mathbf{Y} = \mathbf{g}(\mathbf{Y}) \delta\tau, \tag{5.13}$$

(cf. (5.7)) so that, from (5.10),

$$\delta I(\mathbf{Y}) = -\delta \mathbf{Y}^T \mathbf{g}(\mathbf{Y}) = -\delta \mathbf{Y}^T A(\mathbf{Y}) \delta \mathbf{Y} \delta\tau^{-1} \leq 0 \tag{5.14}$$

provided that $A(\mathbf{Y})$ is positive definite (see (2.11) and (4.31)).

5.1. Two-Stage MFE

An alternative formulation of the MFE method is to express the unknown function in the *discontinuous* linear form

$$U = \sum_{k=1}^{K} \left(W_{k,1}(t)\phi_{k,1}(x) + W_{k,2}(t)\phi_{k,2}(x) \right) \tag{5.15}$$

where $\phi_{k-1,2}, \phi_{k,1}$ are the two halves of the linear basis function α_j of the previous section, to be regarded as fixed in time, instantaneously coincident with the half α_j functions. Then

$$U_t = \sum_k \left(\dot{W}_{k,1}\ \phi_{k,1} + \dot{W}_{k,2}\ \phi_{k,2} \right) \tag{5.16}$$

and, comparing (5.16) with (5.2), there exists a relationship between the $\dot{W}$'s and the $\dot{U}$'s, $\dot{X}$'s of the form

$$\begin{aligned} \dot{U}_j - (U_x)_L\ \dot{X}_j &= \dot{W}_L \\ \dot{U}_j - (U_x)_R\ \dot{X}_j &= \dot{W}_R \end{aligned} \tag{5.17}$$

where suffices L, R refer to elements to the left and right of node j.

The mapping from $\dot{W}$ to $\dot{U}, \dot{X}$ (essentially a coordinate change) has an obvious singularity (referred to above) when $(U_x)_L = (U_x)_R$ which occurs naturally and may lead to infinite speeds.

The rate of change of $I(t)$ is now

$$\frac{dI}{dt} = \sum_k \int_a^b \left[\frac{\partial F}{\partial U} \left(\dot{W}_{k,1}\ \phi_{k,1} + \dot{W}_{k,2}\ \phi_{k,2} \right) + \frac{\partial F}{\partial U_x} \left(\dot{W}_{k,1}\ \phi'_{k,1} + \dot{W}_{k,2}\ \phi'_{k,2} \right) \right] dx \tag{5.18}$$

which vanishes if

$$\int_a^b \frac{\partial F}{\partial U} \left(\dot{W}_{k,\upsilon}\ \phi_{k,\upsilon} + \dot{W}_{k,\upsilon}\ \phi_{k,\upsilon} \right) dx = 0 \tag{5.19}$$

($\upsilon = 1, 2$). In vector form this is

$$\frac{dI(\mathbf{W})}{dt} = -\ \dot{\mathbf{W}}\ \mathbf{d}(\mathbf{W}), \tag{5.20}$$

where

$$\mathbf{W} = (W_{11}, W_{12}, W_{21}, W_{22}, ..., W_{K1}, W_{K2})$$

and $\mathbf{d}(\mathbf{W})$ is the vector with components

$$d_i = \int_a^b \frac{\partial F}{\partial U} \left(\dot{W}_{k,\upsilon}\ \phi_{k,\upsilon} + \dot{W}_{k,\upsilon}\ \phi_{k,\upsilon} \right) dx. \tag{5.21}$$

Equation (5.20) shows that $\dot{\mathbf{W}}$ is a search direction for

$$\mathbf{d}(\mathbf{W}) = \mathbf{0} \tag{5.22}$$

by the method of steepest descent.

As argued in previous sections, the rate of change of $I(\mathbf{W})$ is non-positive when $\dot{\mathbf{W}} = \mathbf{d}(\mathbf{W})$ or, in Galerkin form,

$$E\ \dot{\mathbf{W}} = \mathbf{d}(\mathbf{W}), \tag{5.23}$$

so that, from (5.20),

$$\frac{dI(\mathbf{W})}{dt} = -\ \dot{\mathbf{W}}^T\ E\ \dot{\mathbf{W}} \tag{5.24}$$

(cf. (4.31)). Equation (5.24) converts to (5.9) using (5.17) via a square assembly matrix.

As before, assuming that the matrix E is positive definite (true if the element areas are positive and there is no mesh tangling), I is a non-increasing function, stationary only at steady state at which (5.19) holds.

Evaluation of the integrals involving U_x requires a similar smoothing technique as that for MFE. Thus the function $F(x, U, U_x)$ is in effect replaced by $F(x, U, R(U_x))$ where R is a recovery operator (typically Hermite cubic interpolation).

Algebraically, (4.29) is replaced by

$$E\ \dot{\mathbf{W}} = \mathbf{d}(\mathbf{W}) \tag{5.25}$$

where E is a local elementwise mass matrix having diagonal blocks

$$\begin{pmatrix} \int_a^b \phi_{k\upsilon}\phi_{l\mu}dx & \int_a^b(-U_x)\phi_{k\upsilon}\phi_{l\mu}dx \\ \int_a^b(-U_x)\phi_{k\upsilon}\phi_{l\mu}dx & \int_a^b(U_x)^2\phi_{k\upsilon}\phi_{l\mu}dx \end{pmatrix} \tag{5.26}$$

$\dot{\mathbf{W}} = (\dot{W}_{k1}, \dot{W}_{k1}, \dot{W}_{k1}, \dot{W}_{k1}, ..., \dot{W}_{k1}, \dot{W}_{k1})^T$, while we can write (5.17) as

$$X\ \dot{\mathbf{Y}} = \dot{\mathbf{W}} \tag{5.27}$$

where X is a block diagonal matrix whose blocks are the coefficients of (5.17). This leads to

$$X^T E X\ \dot{\mathbf{Y}} = X^T\ \dot{\mathbf{W}}$$

from which we can identify

$$A = X^T E X \text{ and } \mathbf{g}(\mathbf{Y}) = X^T \mathbf{d}(\mathbf{W}) \tag{5.28}$$

(see (5.7)). Moreover D, the diagonal of A, is given by

$$D = X^T diag\{E\} X. \tag{5.29}$$

This formulation of MFE has a correspondence with the approach taken in Section 4, which we now discuss.

5.2. The MBF approach

A variant of the MFE approach used as an iterative method is to again represent U as a *discontinuous* function, expressing it in terms of the half basis functions $\phi_{k,\nu}(x)$ in the projection stage, as in (5.15), but instead of using continuity of U to define the adaptation through (5.17), allowing the minimisation itself to position the nodes [3]. The price of the extra freedom is the possibility of a discontinuous U on the adapted grid but this is less serious than it appears and there are positive advantages.

To describe the approach (in increment form) let δU be

$$\delta U = \sum_{k=1}^{K} \left(\delta W_{k,1} \phi_{k,1}(x, \mathbf{X}) + \delta W_{k,2} \phi_{k,2}(x, \mathbf{X}) \right), \tag{5.30}$$

(cf. (5.16). Then it can be shown [3] that

$$\delta I = -\sum_{k=1}^{K} \sum_{\upsilon=1}^{2} \int_a^b \left(\frac{\partial F}{\partial U} \phi_{k,\upsilon} + \frac{\partial F}{\partial U_x} \phi'_{k,\upsilon} \right) \delta W_{k,\nu} dx - \sum_{j=1}^{J} [F]_j \, \delta X_j \tag{5.31}$$

where $[F]_j$ denotes the jump in F at node j (cf. (2.38) and (4.19)) and where δU variations are constrained along the graph of U as X_j varies. Hence $\delta I = 0$ implies that U, X satisfy

$$\int_a^b \left(\frac{\partial F}{\partial U} \phi_{l,\upsilon} + \frac{\partial F}{\partial U_x} \phi'_{l,\upsilon} \right) dx = 0 \text{ and } [F]_j = 0. \tag{5.32}$$

Note that if F is independent of U_x one solution of $[F]_j = 0$ is $[U] = 0$ (returning continuity of U).

If the $\delta W_{k,\nu}$ are chosen such that

$$\delta W_{l,\upsilon} = \int_a^b \left(\frac{\partial F}{\partial U} \phi_{l,\upsilon} + \frac{\partial F}{\partial U_x} \phi'_{l,\upsilon} \right) dx, \tag{5.33}$$

or, in Galerkin form,

$$\int_a^b \delta W \phi_{l,\upsilon} dx = \int_a^b \left(\frac{\partial F}{\partial U} \phi_{l,\upsilon} + \frac{\partial F}{\partial U_x} \phi'_{l,\upsilon} \right) dx \tag{5.34}$$

and δX_j is chosen such that

$$\delta X_j = [F]_j \, \delta\sigma, \tag{5.35}$$

where $\delta\sigma$ is positive, then δI in (5.31) is non-increasing, zero only when (5.32) holds. The non-increasing property of I also holds if

$$\int_a^b \left(\frac{\partial F}{\partial U} \phi_{l,\upsilon} dx + \frac{\partial F}{\partial U_x} \phi'_{l,\upsilon} \right) \delta W dx \geq 0 \text{ and/or } [F]_j \, \delta X_j \geq 0. \tag{5.36}$$

The algebraic form of (5.31) is

$$\delta I = -\mathbf{d}.\delta\mathbf{W} - \mathbf{\Theta}.\delta\mathbf{X} \tag{5.37}$$

where $\mathbf{d}$ is given by (5.21), $\mathbf{\Theta} = \{\Theta_j\}$, with

$$\Theta_j = [F]_j \tag{5.38}$$

and

$$\delta\mathbf{W} = (\delta W_{k1}, \delta W_{k1}, \delta W_{k1}, \delta W_{k1}, ..., \delta W_{k1}, \delta W_{k1})^T.$$

If (5.36) holds, then

$$\delta I = -\delta\mathbf{W}^T E \delta\mathbf{W} \delta\tau^{-1} - \sum_{j=1}^{J} [F]_j^2 \, \delta\sigma, \tag{5.39}$$

where E is the same matrix as in (5.26). The δI of (5.39) is nonpositive, zero only at the steady state at which $\delta\mathbf{W} = 0, [F]_j = 0$ holds (provided that the matrix E is positive definite).

The same properties hold if E is replaced by any positive definite matrix such as $D = diag\{E\}$.

In higher dimensions

$$\delta U = \sum_{\upsilon=1}^{d+1} \delta W_{k,\nu} \phi_{k,\nu}, \tag{5.40}$$

where d is the number of dimensions, and, as in the derivation of (5.31), the stationary values satisfy the local problems

$$\int_{\Omega_k} \left(\frac{\partial F}{\partial U} \phi_{k,\upsilon} + \frac{\partial F}{\partial \nabla U} \phi'_{k,\upsilon} \right) \delta W_{k,\nu} d\Omega_k = 0 \tag{5.41}$$

$\forall k$, $\upsilon = 1$ to $(d+1)$ and

$$\sum_{k=1}^{K_j} \int_{\partial\Omega_k} F \alpha_j \mathbf{n}_k . \delta \mathrm{X}_j ds = 0, \tag{5.42}$$

$\forall j$ where $\delta \mathrm{X}_j = (\delta X_j, \delta Y_j)$. Here $\mathbf{n}_k$ is a unit vector along the outward normal to the sides $\partial\Omega_k$ of the K_j triangles Ω_k surrounding node j, with the δU variation in (5.42) constrained to move on the graph (i.e. the planes) of U as X is varied [3]. Moreover, by choosing

$$\delta W_{k,\nu} = -\sum_{k=1}^{K_j} \int_{\Omega_k} \left(\frac{\partial F}{\partial U} \phi_{k,\upsilon} + \frac{\partial F}{\partial \nabla U} \phi'_{k,\upsilon} \right) d\Omega_k \tag{5.43}$$

or, in Galerkin form,

$$\sum_{k=1}^{K_j} \int_{\Omega_k} \delta W \phi_{k,\upsilon} d\Omega = -\sum_{k=1}^{K_j} \int_{\Omega_k} \left(\frac{\partial F}{\partial U} \phi_{k,\upsilon} + \frac{\partial F}{\partial \nabla U} \phi'_{k,\upsilon} \right) d\Omega \tag{5.44}$$

as well as

$$\delta \mathrm{X}_j = -\sum_{k=1}^{K_j} \int_{\partial\Omega_k} F\alpha_j \mathbf{n}_k ds, \tag{5.45}$$

I is a non-increasing function, stationary only when $\delta W_{k,v} = 0$, $\delta \mathbf{X}_j = 0$ in which case (5.41) and (5.42) hold.

In Section 8 we show how this approach can be made the basis of an iterative algorithm. In the next two sections, however, we consider two related issues. These can be skipped without affecting the argument.

6. A Finite Difference Version

Consider replacing the integral in (2.1) by a sum, giving

$$I(t) = \sum_{j=1}^{J} F_j \tag{6.1}$$

where

$$F_j = F\left(x_j(t), U_j(t), \left(\frac{\Delta U(t)}{\Delta x}\right)_j\right) \tag{6.2}$$

in which

$$\left(\frac{\Delta U(t)}{\Delta x}\right)_j = \frac{1}{2}\left(\frac{U_{j+1}(t) - U_j(t)}{X_{j+1}(t) - X_j(t)} + \frac{U_j(t) - U_{j-1}(t)}{X_j(t) - X_{j-1}(t)}\right). \tag{6.3}$$

Then

$$\frac{dI}{dt} = \sum_{j=1}^{J}\left[\frac{\partial F_j}{\partial U_j}\dot{U}_j + \frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta x}\right)_j}\frac{1}{2}\left(\frac{U_{j+1}(t) - U_j(t)}{X_{j+1}(t) - X_j(t)} + \frac{U_j(t) - U_{j-1}(t)}{X_j(t) - X_{j-1}(t)}\right)\right]$$

$$= \sum_{j=1}^{J}\left[\frac{\partial F_j}{\partial U_j} + \frac{\Delta}{\Delta x}\left(\frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta x}\right)_j}\right)_j\right]\dot{U}_j \tag{6.4}$$

using rearrangement or summation by parts, where it has been assumed that $\dot{U}_0 = \dot{U}_J = 0$.

Hence, $\frac{dI}{dt} = 0$ implies that

$$\frac{\partial F_j}{\partial U_j} + \frac{\Delta}{\Delta x}\left(\frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta x}\right)_j}\right)_j = 0. \tag{6.5}$$

Moreover, if

$$u_t = -\frac{\partial F}{\partial u} + \frac{d}{dx}\frac{\partial F}{\partial u_x}$$

is semi-discretised in a finite difference manner as

$$\dot{U}_j = -\left[\frac{\partial F_j}{\partial U_j} + \frac{\Delta}{\Delta x}\left(\frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta x}\right)_j}\right)_j\right] \tag{6.6}$$

(see (6.3) then

$$\frac{dI}{dt} = -\sum_{j=1}^{J} \dot{U}_j^2 \leq 0,$$

and I is a non-increasing function of t vanishing only when $\dot{U}_j = 0\ \forall j$, i.e. at a steady state where (6.5) holds. Introducing a vector $\mathbf{b}$ with components equal to the expression in square brackets in (6.4), then we again have

$$\frac{d\mathrm{I}(\mathbf{U})}{dt} = -\dot{\mathbf{U}}^T \mathbf{b}(\mathbf{U}) \tag{6.7}$$

(cf. (3.16)) which is greatest in magnitude when $\dot{\mathbf{U}}^T$ is in the same direction as $\mathbf{b}(\mathbf{U})$.

In two dimensions we write

$$F_j = F\left(x_j(t), y_j(t), U_j(t), \left(\frac{\Delta U(t)}{\Delta x}\right)_j, \left(\frac{\Delta U(t)}{\Delta y}\right)_j\right)$$

where

$$\left(\frac{\Delta U(t)}{\Delta x}\right)_j = \frac{1}{K_j}\sum \frac{\sum U_k\,(Y_l - Y_m)}{\sum X_k\,(Y_l - Y_m)}$$

$$\left(\frac{\Delta U(t)}{\Delta y}\right)_j = \frac{1}{K_j}\sum \frac{\sum U_k\,(X_l - X_m)}{\sum Y_k\,(X_l - X_m)} \tag{6.8}$$

in which the main sums run over the vertices of each of the K_j triangles surrounding the point j. Then

$$\frac{dI}{dt} = \sum_{j=1}^{J}\left[\frac{\partial F_j}{\partial U_j}\dot{U}_j + \frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta x}\right)_j}\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum \dot{U}_k\,(Y_l - Y_m)}{\sum X_k\,(Y_l - Y_m)}\right.$$

$$\left. + \frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta y}\right)_j}\frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum \dot{U}_k\,(X_l - X_m)}{\sum Y_k\,(X_l - X_m)}\right]$$

$$= \sum_{j=1}^{J}\left[\frac{\partial F_j}{\partial U_j} - \frac{1}{K_j}\sum_{k=1}^{K_j}\left.\frac{\sum \frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta x}\right)_j}(Y_l - Y_m)}{\sum X_k\,(Y_l - Y_m)}\right|_{k=j}\right.$$

$$-\frac{1}{K_j}\sum_{k=1}^{K_j}\left.\frac{\sum\frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta y}\right)_j}(X_l-X_m)}{\sum Y_k\,(X_l-X_m)}\right|_{k=j}\dot{U}_j \tag{6.9}$$

where l, m in the inner sums are the second and third vertices of the triangle in which k or j is the first. It has been assumed that $\dot{U}_j = 0$ at the boundary nodes. Clearly $\frac{dI}{dt} = 0$ implies that

$$\frac{\partial F_j}{\partial U_j} - \frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum\frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta x}\right)_j}(Y_l-Y_m)}{\sum X_k\,(Y_l-Y_m)} - \frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum\frac{\partial F_j}{\partial\left(\frac{\Delta U(t)}{\Delta y}\right)_j}(X_l-X_m)}{\sum Y_k\,(X_l-X_m)} = 0. \tag{6.10}$$

If the differential equation

$$u_t = -\frac{\partial F}{\partial u} + \nabla\frac{\partial F}{\partial \nabla u}$$

is discretised in space using the expression in square brackets in (6.9), then

$$\frac{dI}{dt} = -\sum_{j=1}^{J}\dot{U}_j^2 \leq 0$$

with zero only when $\dot{U}_j = 0\ \forall j$, i.e. at steady state where (6.10) holds. If F is a positive function then I tends to a limit as $t \to \infty$ at which U_j is a steady state solution satisfying (6.10).

For example, if

$$F(x, U, \frac{\Delta U}{\Delta x}) = \frac{1}{2}\left(U^2 + \left(\frac{\Delta U}{\Delta x}\right)^2\right) \tag{6.11}$$

(6.10) becomes

$$\sum_{j=1}^{J}\left[U_j - \frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum\frac{\Delta U}{\Delta x}(Y_{k2}-Y_{k3})}{\sum X_{k1}\,(Y_{k2}-Y_{k3})} - \frac{1}{K_j}\sum_{k=1}^{K_j}\frac{\sum\frac{\Delta U}{\Delta x}(X_{k2}-X_{k3})}{\sum Y_{k1}\,(X_{k2}-X_{k3})}\,\dot{U}_j\right] = 0$$

(see (6.8)) and the norm

$$\mathcal{I}(U) = \frac{1}{2}\sum_{j=1}^{J}\left(U^2 + \left(\frac{\Delta U}{\Delta x}\right)^2\right)$$

is minimised locally.

7. Time Discretisation

Although it is not possible to generate PDEs with first derivatives from the minimisation in Section 2, it is possible to generate discretisations of such PDEs. Suppose that we consider the discretised version of (2.15) using implicit finite differences in time, giving

$$\frac{u^{n+1}-u^n}{\Delta t} = \left(-\frac{\partial F}{\partial u} + \nabla.\frac{\partial F}{\partial \nabla U}\right)^{n+1}. \tag{7.1}$$

Write $u(x,\tau) = u^{n+1}$ and extend the function $F(\mathbf{x}, u, \nabla u)$ to

$$G(\mathbf{x}, u, \nabla u) = F(\mathbf{x}, u, \nabla u) + \frac{1}{2}\frac{(u-u^n)^2}{\Delta t}, \tag{7.2}$$

defining also the functional $\mathcal{J}$ as

$$\mathcal{J}(u) = \int_\Omega G(\mathbf{x}, u, \nabla u)d\Omega = \int_\Omega F(\mathbf{x}, u, \nabla u)d\Omega + \frac{1}{2\Delta t}\int_\Omega (u-u^n)^2\, d\Omega \tag{7.3}$$

whose value at time τ as

$$J(\tau) = \int_\Omega G(\mathbf{x}(\tau), u(\mathbf{x},\tau), \nabla u(\mathbf{x},\tau))d\Omega.$$

Differentiating $J(\tau)$ we have

$$\begin{aligned}\frac{dJ}{d\tau} &= \int_\Omega \left(\frac{\partial G}{\partial u}u_\tau + \frac{\partial G}{\partial \nabla u}.\nabla u_\tau\right) d\Omega \\ &= \int_\Omega \left[\frac{\partial G}{\partial u} - \nabla.\left(\frac{\partial G}{\partial \nabla u}\right)\right] u_\tau d\Omega.\end{aligned}$$

$\frac{dJ}{d\tau} = 0$ implies that

$$\frac{\partial G}{\partial u} - \nabla.\left(\frac{\partial G}{\partial \nabla u}\right) = 0 \tag{7.4}$$

which is equivalent ot (7.1).

If u satisfies the differential equation

$$\begin{aligned}u_\tau &= -\frac{\partial G}{\partial u} + \nabla.\left(\frac{\partial G}{\partial \nabla u}\right) \\ &= -\frac{(u-u^n)}{\Delta t} - \frac{\partial F}{\partial u} + \nabla.\left(\frac{\partial F}{\partial \nabla u}\right)\end{aligned} \tag{7.5}$$

then

$$\frac{dJ}{d\tau} = -\int_\Omega u_\tau^2 d\Omega \le 0$$

with zero only if $u_\tau = 0$, i.e. when

$$\frac{(u-u^n)}{\Delta t} = -\frac{\partial F}{\partial u} + \nabla . \left(\frac{\partial F}{\partial \nabla u} \right) \tag{7.6}$$

where u is the required solution of (7.1). From (7.2) if F is a non-negative function then so is G and therefore J. Hence J tends to a limit as $t \to \infty$ at which $u_\tau = 0$ and (7.6) holds.

For example, if

$$F(\mathbf{x}, u, \nabla u) = \frac{1}{2}\left(u^2 + (\nabla u)^2\right)$$

then

$$G(\mathbf{x}, u, \nabla u) = \frac{1}{2}\left(u^2 + (\nabla u)^2\right) + \frac{1}{2\Delta t}(u-u^n)^2$$

and the differential equation is

$$u_\tau = -\frac{(u-u^n)}{\Delta t} - u + \nabla^2 u \tag{7.7}$$

which has the steady limit

$$\frac{(u-u^n)}{\Delta t} = -u + \nabla^2 u$$

This differential equation gives no information about the way in which to discretise (2.15) in time but simply concerns the convergence of the solution of the implicit equation resulting from the a given time-stepping. The approach is easily generalised to Crank-Nicolson.

The independent nature of this extension allows all previous cases to be incorporated, including both fixed and moving finite elements as well as finite differences.

For example, in the fixed finite element case,

$$J(\tau) = \int_\Omega \left[F(\mathbf{x}, U, \nabla U) + \frac{1}{2\Delta t}(U-U^n)^2 \right] d\Omega \tag{7.8}$$

which is a non-increasing function of τ when U satisfies the weak form

$$\int U_\tau \psi_j(\mathbf{x}) d\Omega = \int_\Omega \left[\left(\frac{\partial F}{\partial U} + \frac{U-U^n}{\Delta t} \right) \psi_j(\mathbf{x}) + \frac{\partial F}{\partial \nabla U} . \nabla \psi_j(\mathbf{x}) \right] d\Omega \tag{7.9}$$

(cf.(3.27)). As $\tau \to \infty$ the 'pseudo-steady state' solution satisfies

$$\int_\Omega \left[\left(\frac{\partial F}{\partial U} + \frac{(U-U^n)}{\Delta t} \right) \psi_j(\mathbf{x}) + \frac{\partial F}{\partial \nabla U} . \nabla \psi_j(\mathbf{x}) \right] d\Omega = 0 \tag{7.10}$$

which is the appropriate time-discretisation of the PDE (7.1).

The Galerkin equation (7.10), itself a discretisation of (3.27), is therefore a result of $\mathcal{J}(U)$ being stationary with respect to U. Similarly, in the adaptive case the corresponding discretisations of (4.36) and (4.37) arise from making $\mathcal{J}(u)$ stationary with respect to U, X. In the extension to the MBF method the discretisations of the Galerkin forms (5.41) and (5.42)) simply have F replaced by G.

8. Iteration to Steady State

The property of the previous sections that $I(t)$ is non-increasing and tends to a limit as $t \to \infty$ holds only in a semi-discrete sense. Notwithstanding the results in Section 7, a further (time-like) discretisation is necessary to obtain a practical algorithm for reaching the steady state which may invalidate this property, although it will hold for sufficiently small Δt. However, the property can be used to construct steepest descent algorithms converging (in the sense that I converges) to a stationary point and hence to a corresponding steady state solution of the weak form.

Such an iteration may be expected to converge progressively more slowly as the limit is approached. However, it is desirable to be able to take as large a pseudo-time step $\Delta\tau$ as possible, consistent with reaching convergence. A standard approach is to accelerate the convergence by switching to Newton's method when possible (a strategy which is the basis of many packages for the solution of nonlinear equations).

We illustrate the approach on the fixed finite element method of section 2. The aim is to solve the weak form (3.5) (with ψ replaced by α) in its algebraic form (3.13), i.e.

$$\mathbf{b}(\mathbf{U}) = 0. \tag{8.1}$$

From

$$\frac{dI(\mathbf{U})}{dt} = -\mathbf{b}(\mathbf{U})\,\dot{\mathbf{U}} \tag{8.2}$$

we have a steepest descent property for $\mathbf{b}(\mathbf{U}) = 0$. The steepest descent iteration is

$$\mathbf{U}^{p+1} - \mathbf{U}^p = \Delta t \mathbf{b}(\mathbf{U})^p, \tag{8.3}$$

corresponding to (3.14), where Δt is a suitable step, while an iteration based on the Galerkin form (3.18) is

$$A_\alpha(\mathbf{U}^{p+1} - \mathbf{U}^p) = \Delta t \mathbf{b}(\mathbf{U})^p \tag{8.4}$$

or its diagonal variant

$$D_\alpha(\mathbf{U}^{p+1} - \mathbf{U}^p) = \Delta t \mathbf{b}(\mathbf{U})^p, \tag{8.5}$$

both of the same form as Newton's iteration

$$-J^p(\mathbf{U}^{p+1} - \mathbf{U}^p) = \mathbf{b}(\mathbf{U})^p \tag{8.6}$$

where

$$J = \frac{\partial \mathbf{b}}{\partial \mathbf{U}}. \tag{8.7}$$

The A and D matrices guarantee descent for sufficiently small Δt but the J matrix does not.

(If F is quadratic these arguments are redundant since the matrix equation (8.1) is linear. In all the adaptive cases as well as (2.5) and (2.28), however, the equation (8.1) is nonlinear.)

8.1. MBF iterations

We have seen in the MBF approach in Sections 5.2 -5.4 that an optimal solution U may be sought in the space of discontinuous piecewise linear functions. It is then necessary to solve the pair of equations () and (), namely

$$\mathbf{d}(\mathbf{W}) = \mathbf{0} \tag{8.8}$$

and

$$[F]_j = 0 \tag{8.9}$$

$\forall j$ (with variations δU in the latter case constrained to lie on the graph of U), for $\mathbf{W}$ and X simultaneously. This leads to a natural two-step iteration scheme on these two equations alternately whereby solutions for one of the variables are obtained while the other variable is held fixed. That is, we seek a solution of (8.8) for $\mathbf{W}$ with nodal positions held fixed, and then solve (8.9) for X_j (interpolating or extrapolating W along the piecewise linear graph of the current solution). The procedure is then repeated to convergence. The iteration has the property that $I(t)$ is non-increasing at each stage of the iteration. If (8.8),(8.9) cannot be solved outright a fallback position is to use steepest descent iterations, namely

$$E(\mathbf{W}^{p+1} - \mathbf{W}^p) = \Delta\tau \mathbf{d}(\mathbf{W})^p \tag{8.10}$$

or

$$D_E(\mathbf{W}^{p+1} - \mathbf{W}^p) = \Delta\tau \mathbf{d}(\mathbf{W})^p \tag{8.11}$$

and

$$X_j^{p+1} - X_j^p = -[F]_j^p \Delta\sigma \tag{8.12}$$

(under the above constraint) with $\Delta\tau$ and $\Delta\sigma$ chosen such that $I(t)$ is non-increasing.

8.2. Best Fits using Direct Minimisation

For example, in [6] the case (2.3) is treated in this way using both piecewise linear and piecewise constant approximation on a line and on triangles. In the one-dimensional case it is possible to solve both (8.8) and (8.9) outright, the appropriate equations (from (5.32)) being

$$\int_a^b (U - f(x))\phi_{l,\nu} dx = 0 \tag{8.13}$$

$\forall l, \nu$ and

$$[(U - f(x))^2]_j = 0 \tag{8.14}$$

$\forall j$ where in the latter equation variations are confined to lie on the graph of the current approximation to U. The procedure is to solve (8.13) for U and then to solve (8.14) for X, picking the solution which reduces I, repeating to convergence. The converged solution gives continuity of U almost everywhere with no tangling of the grid. Convergence may be accelerated by Newton's method.

In the two-dimensional case it is possible to solve only (8.8), now in the form

$$\int_\Omega (U - f(\mathbf{x}))\phi_{l,\nu} d\Omega = 0, \tag{8.15}$$

outright while (8.9), now

$$\sum_{k=1}^{K_j} \int_{\partial\Omega_k} F\mathbf{n}.\delta\mathbf{X}_k ds = 0, \tag{8.16}$$

(see (5.42)) must be replaced by (8.12) in the form

$$\mathrm{X}_j^{p+1} - \mathrm{X}_j^p = -\Delta\sigma \sum_{k=1}^{K_j} \int_{\partial\Omega_k} F\alpha_j \mathbf{n} ds \tag{8.17}$$

where $\mathrm{X}_{j=}(X_j, Y_j)$. Additional precautions have to be taken to ensure that the grid does not tangle. There are special problems with convergence in two dimensions in that the error cannot be driven down to machine accuracy because of the inflexibility of the grid topology. However, by using an edge-swapping routine and a technique for small element removal this can be achieved (see [6]). The optimal approximation U is no longer continuous but the jumps are small a.e.

In these best fit problems it is straightforward to calculate $\mathbf{W}$ from (8.8) because the functional is quadratic. However, in the shallow water equation case (2.5), for example, we cannot solve (8.8) outright except by iteration (using Newton's method, say). The alternative available is to go back to (8.10) or (8.11) and do a steepest descent step.

The iteration proposed here, therefore, is to use (8.11) and (8.12) in turn to reduce the functional I at each step, converging towards a limit at which the approximation is optimal. The method is confined to functionals with F bounded below.

9. Relation with Equidistribution

Various authors have discussed properties of optimal grids in the limit of large numbers of nodes, showing that in one dimenson the distribution asymptotically possesses an equidistribution property [7]. That is to say, there exists a function $\mathcal{E}(x)$ whose increment across an interval is constant over the grid. Algebraically this can be stated as $\Delta\mathcal{E} =$ constant or, if Δx is the length of the interval, as

$$\overline{\mathcal{E}}\Delta x = \text{ constant},$$

$\overline{\mathcal{E}}$ being a mean value of $\mathcal{E}(x)$ in the interval. This can also be regarded as a discretisation of

$$\mathcal{E}(x)x_\xi = \text{ constant},$$

where ξ is the computational coordinate, or of the nonlinear elliptic differential equation

$$(\mathcal{E}(x)x_\xi)_\xi = 0 \tag{9.1}$$

with boundary conditions $x(a) = 0$, $x(b) = 1$, say.

In particular, one of the properties of the MFE method when run to steady state includes the generation of grids which equidistribute functions of the solution $U(x)$. For example, in the case of the heat equation, the steady state grid (if reached) equidistributes $|u_{xx}|^{2/3}$ in the limit of large numbers of nodes, and there is a corresponding result for convection-diffusion [4]. Jimack [7] has shown that, for a general number of nodes, the grid for the heat equation is optimal in the sense that it corresponds to the best L_2 fit with adjustable nodes by piecewise linear functions to the steady state approximation U in the H^1 semi-norm. This is equivalent to the best fit by piecewise constant functions in the L_2 norm. A similar result holds (for a different equation) for best fits by piecewise linears in the L_2 norm. In the limit of large numbers of nodes, therefore, there is an equivalence between the best fit and the equidistribution principle. The latter is of only limited use for small numbers of nodes, however, although it may be used to create grids which give a first guess for the iterative algorithms discussed in Section 8.

In higher dimensions there is no unambiguous equidistribution principle but an obvious generalisation of (9.1) is

$$\nabla.(E(\mathbf{x})\nabla\mathbf{x}) = 0, \tag{9.2}$$

which again can be used to construct an initial guess for the grid in the iterations of Section 8.

10. Conclusions

In this report we have shown how some standard variational methods of interest to the numerical analyst may be generalised to include grid adaptivity. A practical

iterative method is proposed which is viable in any number of dimensions.

11. References

[1] **Burden, R.L. and Faires, J.D.**, Numerical Analysis. Wadworth (1981).

[2] **Seliger, R.L. and Whitham, G.B.**, Variational Principles in Continuum Mechanics. Proc. Roy. Soc. A, 305,1-25 (1968).

[3] **Baines, M.J.**, Algorithms for Optimal Discontinuous Piecewise Linear and Constant L_2 Fits to Continuous Functions with Adjustable Nodes in One and Two Dimensions. Math.Comp., 62, 645-669 (1994).

[4] **Baines, M.J.**, Moving Finite Elements. OUP (1994).

[5] **Jimack, P.K.**, On the Stability of the Moving Finite Element Method for a Class of Parabolic Partial Differential Equations. Proceedings of Paris Conference (to appear).

[6] **Tourigny, Y. and Baines, M.J.**, Analysis of an Algorithm for Generating Locally Optimal Meshes for L_2 Approximation by Discontinuous Piecewise Polynomials. Math.Comp. (to appear).

[7] **de Boor, C.**, Good Approximation by Splines with Variable Knots. Springer Lecture Notes Series 363, Springer-Verlag (1973).

[7] **Jimack, P.K.**, A Best Approximation Property of the Moving Finite Element Method. School of Computer Studies Research Report 93.35, University of Leeds. (To appear in SIAM J. Numer. An.).