COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING 58 (1986) 305–328
NORTH-HOLLAND

A NEW FINITE ELEMENT FORMULATION FOR COMPUTATIONAL FLUID DYNAMICS: III. THE GENERALIZED STREAMLINE OPERATOR FOR MULTIDIMENSIONAL ADVECTIVE-DIFFUSIVE SYSTEMS*

Thomas J.R. HUGHES and Michel MALLET
Division of Applied Mechanics, Durand Building, Stanford University, Stanford, CA 94305, U.S.A.

Received 5 December 1985

A finite element method based on the SUPG concept is presented for multidimensional advective-diffusive systems. Error estimates for the linear case are established which are valid over the full range of advective-diffusive phenomena.

1. Introduction

In this paper we construct an appropriate generalization of the streamline-upwind/Petrov–Galerkin concept (SUPG) for coupled, multidimensional advective-diffusive systems. Efforts in this direction heretofore have failed to appropriately treat the independent modal components of even one-dimensional systems. After describing the basic equations that we deal with in Section 2, we address the one-dimensional system case in Section 3, beginning with an illustration of the shortcomings of previous SUPG-type formulations. In passing, we show that Lax–Wendroff/Taylor–Galerkin methods also fail for systems for precisely the same reasons. The present method correctly treats the individual components of one-dimensional systems and is suggestive of the required structure of the multidimensional system case. It is pointed out that a generalization of the concept of the absolute value of a matrix needs to be made for the case in which the matrix is rectangular in order that the methodology be correctly extendable to multidimensional systems. A successful generalization is developed and the multidimensional formulation is shown to satisfy three design requirements: (i) it reduces correctly to the optimal one-dimensional system case; (ii) it is equivalent to SUPG for a scalar, multidimensional advection-diffusion equation; and (iii) it reduces to SUPG on each uncoupled component of multidimensional, diagonalizable advective-diffusive systems. Error estimates for the linear case are presented which are valid over the entire range of advective-diffusive behavior.

The next paper in this series deals with the development of a discontinuity-capturing operator for multidimensional advective-diffusive systems. This operator increases the robust-

* This research was sponsored by the NASA Langley Research Center under Grant NASA-NAG-1-361.

ness of the method in the presence of sharp internal and boundary layers in the exact solution (see [7] for related ideas in the scalar case).

2. Basic equations

Consider

$$\boldsymbol{U}_{,t} + \boldsymbol{A}\cdot\boldsymbol{\nabla}\boldsymbol{U} = \boldsymbol{\nabla}\cdot\boldsymbol{K}\boldsymbol{\nabla}\boldsymbol{U} + \mathscr{F}\,, \tag{1}$$

where

$$\boldsymbol{U}^{\mathrm{t}} = (\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_m)\,, \tag{2}$$

$$\boldsymbol{\nabla}^{\mathrm{t}} = (\boldsymbol{I}_m\partial_1, \boldsymbol{I}_m\partial_2, \ldots, \boldsymbol{I}_m\partial_d)\,, \tag{3}$$

$$\boldsymbol{\nabla}\boldsymbol{U}^{\mathrm{t}} = (\boldsymbol{U}^{\mathrm{t}}_{,1}, \boldsymbol{U}^{\mathrm{t}}_{,2}, \ldots, \boldsymbol{U}^{\mathrm{t}}_{,d})\,, \tag{4}$$

$$\boldsymbol{A}^{\mathrm{t}} = (\boldsymbol{A}^{\mathrm{t}}_1, \boldsymbol{A}^{\mathrm{t}}_2, \ldots, \boldsymbol{A}^{\mathrm{t}}_d)\,, \tag{5}$$

$$\boldsymbol{A}\cdot\boldsymbol{\nabla}\boldsymbol{U} = \boldsymbol{A}^{\mathrm{t}}\boldsymbol{\nabla}\boldsymbol{U} = \sum_{i=1}^{d}\boldsymbol{A}_i\boldsymbol{U}_{,i} = \boldsymbol{A}_i\boldsymbol{U}_{,i}\,, \tag{6}$$

$$\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_{11} & \cdots & \boldsymbol{K}_{1d} \\ \vdots & \ddots & \vdots \\ \boldsymbol{K}_{d1} & \cdots & \boldsymbol{K}_{dd} \end{pmatrix}, \tag{7}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{K}\boldsymbol{\nabla}\boldsymbol{U} = \boldsymbol{\nabla}^{\mathrm{t}}\boldsymbol{K}\boldsymbol{\nabla}\boldsymbol{U} = \sum_{i,j=1}^{d}(\boldsymbol{K}_{ij}\boldsymbol{U}_{,j})_{,i} = (\boldsymbol{K}_{ij}\boldsymbol{U}_{,j})_{,i}\,, \tag{8}$$

in which $\boldsymbol{I}_m$ is the $m \times m$ identity matrix, $\partial_i = \partial/\partial x_i$, a comma denotes partial differentiation (e.g., $\boldsymbol{U}_{,t} = \partial\boldsymbol{U}/\partial t$ and $\boldsymbol{U}_{,i} = \partial\boldsymbol{U}/\partial x_i$), $\boldsymbol{A}_i$ and $\boldsymbol{K}_{ij}$ are $m \times m$ matrices, and $\mathscr{F}$ is a source term. As may be deduced from (6) and (8), the summation convention is assumed for repeated spatial indices. To override the summation convention, we underline the repeated index.

REMARK 2.1. (1) $\boldsymbol{A}\cdot\boldsymbol{\nabla}\boldsymbol{U}$ plays the role of a generalized advection term and $\boldsymbol{\nabla}\cdot\boldsymbol{K}\boldsymbol{\nabla}\boldsymbol{U}$ plays the role of a generalized diffusion operator.

(2) Equation (1) is called a *hyperbolic system* if $\boldsymbol{K} = \boldsymbol{0}$ and, for all $c_i \in \mathbb{R}$ $(1 \leqslant i \leqslant d)$, $c_i\boldsymbol{A}_i$ has real eigenvalues and a complete set of eigenvectors.

(3) The compressible Navier–Stokes equations can be written in the form (1) in terms of so-called 'conservation variables',

$$\boldsymbol{U}^{\mathrm{t}} = (\rho, \rho\boldsymbol{u}^{\mathrm{t}}, \rho e)\,, \tag{9}$$

where ρ is the density, $\boldsymbol{u}$ is the velocity, and e is the energy. However, this does not seem to be the most canonical form.

For many physical systems, and in particular for the compressible Navier–Stokes equations, a change of variables exists such that (1) can be written in the following symmetric form, see [3, 4, 6, 11]:

$$\boldsymbol{A}_0\boldsymbol{V}_{,t} + \tilde{\boldsymbol{A}}\cdot\nabla\boldsymbol{V} = \nabla\cdot\tilde{\boldsymbol{K}}\nabla\boldsymbol{V} + \mathscr{F}\,, \tag{10}$$

where

$$\boldsymbol{A}_0 = \partial\boldsymbol{U}/\partial\boldsymbol{V}\,, \tag{11}$$

$$\tilde{\boldsymbol{A}}_i = \boldsymbol{A}_i\boldsymbol{A}_0\,, \tag{12}$$

$$\tilde{\boldsymbol{K}} = \begin{pmatrix} \tilde{\boldsymbol{K}}_{11} & \cdots & \tilde{\boldsymbol{K}}_{1d} \\ \vdots & \ddots & \vdots \\ \tilde{\boldsymbol{K}}_{d1} & \cdots & \tilde{\boldsymbol{K}}_{dd} \end{pmatrix}, \tag{13}$$

$$\tilde{\boldsymbol{K}}_{ij} = \boldsymbol{K}_{ij}\boldsymbol{A}_0\,. \tag{14}$$

In the above, $\boldsymbol{A}_0$ is symmetric and positive-definite, the $\tilde{\boldsymbol{A}}_i$ are symmetric, and $\tilde{\boldsymbol{K}}$ is symmetric and positive-semidefinite.

REMARK 2.2. (1) Weak solutions of (1), as well as strong, are preserved by the symmetric form (10).

(2) If $\tilde{\boldsymbol{K}} = \boldsymbol{0}$, then (10) is called *symmetric hyperbolic*, or a *Friedrichs' system* (e.g., the compressible Euler equations). If $\tilde{\boldsymbol{K}}$ is positive-definite, then (10) is called *symmetric parabolic*. If $\tilde{\boldsymbol{K}}$ is positive-semidefinite, but not positive-definite, then (10) is called *symmetric incompletely parabolic* (e.g., the compressible Navier–Stokes equations).

(3) The matrix $\boldsymbol{A}_0$ plays the role of a *metric tensor* on $\mathbb{R}^m$. If $\boldsymbol{A}_0 = \boldsymbol{I}_m$, we shall say that the system (10) has *Euclidean metric*; if $\boldsymbol{A}_0 \neq \boldsymbol{I}_m$ the system has *Riemannian metric*.

Consider now a given *linear constant-coefficient system* of the form (10). If the system has Riemannian metric it can always be transformed to one with Euclidean metric. This may be seen as follows: Let

$$\boldsymbol{A}_0 = \boldsymbol{L}\boldsymbol{L}^{\mathrm{t}}\,. \tag{15}$$

Note that this representation is not unique. For example, $\boldsymbol{L}$ could equal the Cholesky factor of $\boldsymbol{A}_0$, the square root of $\boldsymbol{A}_0$, etc. Employing (15) in (10), we have

$$\boldsymbol{L}\boldsymbol{L}^{\mathrm{t}}\boldsymbol{V}_{,t} + \tilde{\boldsymbol{A}}_i\boldsymbol{V}_{,i} = (\tilde{\boldsymbol{K}}_{ij}\boldsymbol{V}_{,j})_{,i} + \mathscr{F}\,, \tag{16}$$

which is equivalent to

$$\boldsymbol{L}^{\mathrm{t}}\boldsymbol{V}_{,t} + (\boldsymbol{L}^{-1}\tilde{\boldsymbol{A}}_i\boldsymbol{L}^{-\mathrm{t}})\boldsymbol{L}^{\mathrm{t}}\boldsymbol{V}_{,i} = ((\boldsymbol{L}^{-1}\tilde{\boldsymbol{K}}_{ij}\boldsymbol{L}^{-\mathrm{t}})\boldsymbol{L}^{\mathrm{t}}\boldsymbol{V}_{,j})_{,i} + \boldsymbol{L}^{-1}\mathscr{F}\,, \tag{17}$$

or

$$\boldsymbol{X}_{,t} + \hat{\boldsymbol{A}}_i\boldsymbol{X}_{,i} = (\hat{\boldsymbol{K}}_{ij}\boldsymbol{X}_{,j})_{,i} + \hat{\mathscr{F}}\,, \tag{18}$$

where

$$\boldsymbol{X} = \boldsymbol{L}^{\mathrm{t}}\boldsymbol{V}\,, \tag{19}$$

$$\hat{\boldsymbol{A}}_i = \boldsymbol{L}^{-1}\tilde{\boldsymbol{A}}_i\boldsymbol{L}^{-\mathrm{t}}\,, \tag{20}$$

$$\hat{\boldsymbol{K}}_{ij} = \boldsymbol{L}^{-1}\tilde{\boldsymbol{K}}_{ij}\boldsymbol{L}^{-\mathrm{t}}\,, \tag{21}$$

$$\hat{\mathscr{F}} = \boldsymbol{L}^{-1}\mathscr{F}\,. \tag{22}$$

Thus we may write

$$\boldsymbol{X}_{,t} + \hat{\boldsymbol{A}}\cdot\nabla\boldsymbol{X} = \nabla\cdot\hat{\boldsymbol{K}}\nabla\boldsymbol{X} + \hat{\mathscr{F}}\,, \tag{23}$$

REMARK 2.3. (1) In the sequel, it will prove convenient to develop our theory in terms of (23), and then transform back to the case of (10) which is viewed as canonical in some sense. For example, it might lead to the most natural stability estimate.

(2) It is pertinent to ask whether or not an even more canonical system than those above can be defined. For example, can the matrices $\boldsymbol{I}_m, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_d$ be simultaneously diagonalized by the same similarity transformation? For most multidimensional (i.e., $d \geq 2$) systems of physical interest, the answer seems to be no.

(3) The main intention of the present work is to develop a finite element formulation of (10) which embodies a correct generalization of the streamline concept to coupled multi-dimensional systems. Throughout the remainder of the paper we assume the linear, constant-coefficient case. However, the final formulation obtained has more general applicability.

3. One-dimensional systems

Consider the following one-dimensional hyperbolic system:

$$\boldsymbol{U}_{,t} + \boldsymbol{A}\boldsymbol{U}_{,x} = \boldsymbol{0}\,, \tag{24}$$

where we assume $\boldsymbol{A}$ is constant. Corresponding to the alternate forms introduced in Section 2, we have

$$\boldsymbol{A}_0\boldsymbol{V}_{,t} + \tilde{\boldsymbol{A}}\boldsymbol{V}_{,x} = \boldsymbol{0} \tag{25}$$

and

$$\boldsymbol{X}_{,t} + \hat{\boldsymbol{A}}\boldsymbol{X}_{,x} = \boldsymbol{0}\,. \tag{26}$$

The following eigenvalue problems are useful for diagonalizing (24)–(26):

$$(\boldsymbol{A} - \lambda\boldsymbol{I}_m)\boldsymbol{\psi} = \boldsymbol{0}\,, \tag{27}$$

$$(\tilde{\boldsymbol{A}} - \lambda\boldsymbol{A}_0)\boldsymbol{\varphi} = \boldsymbol{0}\,, \tag{28}$$

$$(\hat{\boldsymbol{A}} - \lambda\boldsymbol{I}_m)\boldsymbol{v} = \boldsymbol{0}\,. \tag{29}$$

Note that in each case the eigenvalues are the same. The eigenvectors are related by

$$\boldsymbol{\psi} = \boldsymbol{A}_0 \boldsymbol{\varphi} = \boldsymbol{L} \boldsymbol{v} \,. \tag{30}$$

The developments for the one-dimensional case could be presented in terms of any one of (24)–(26). Our plan is to work with (26) and eventually convert the results to (25). Let

$$\boldsymbol{Y} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m] \,. \tag{31}$$

It follows from (29) that

$$\hat{\boldsymbol{A}} = \boldsymbol{Y} \boldsymbol{\Lambda} \boldsymbol{Y}^{-1} \,, \tag{32}$$

where

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) \,. \tag{33}$$

By virtue of the symmetry of $\hat{\boldsymbol{A}}$ the eigenvectors are orthogonal and can be scaled such that

$$\boldsymbol{Y}^{\mathrm{t}} \boldsymbol{Y} = \boldsymbol{Y} \boldsymbol{Y}^{\mathrm{t}} = \boldsymbol{I}_m \,. \tag{34}$$

Likewise, (32) can be written as

$$\hat{\boldsymbol{A}} = \boldsymbol{Y} \boldsymbol{\Lambda} \boldsymbol{Y}^{\mathrm{t}} \,. \tag{35}$$

The change of variables

$$\boldsymbol{X} = \boldsymbol{Y} \boldsymbol{\chi} \tag{36}$$

can be employed to diagonalize (26):

$$\boldsymbol{\chi}_{,t} + \boldsymbol{\Lambda} \boldsymbol{\chi}_{,x} = \boldsymbol{0} \,. \tag{37}$$

It is also interesting to consider at this point a counterpart of (26) with diffusion and a source term:

$$\boldsymbol{X}_{,t} + \hat{\boldsymbol{A}} \boldsymbol{X}_{,x} = \kappa \boldsymbol{X}_{,xx} + \hat{\mathscr{F}} \,, \tag{38}$$

where κ is constant. The diffusion term has been taken to be particularly simple in that $\boldsymbol{K} = \kappa \boldsymbol{I}_m$. The fully general case will be considered later on. Equation (38) can also be diagonalized by employing (36):

$$\boldsymbol{\chi}_{,t} + \boldsymbol{\Lambda} \boldsymbol{\chi}_{,x} = \kappa \boldsymbol{\chi}_{,xx} + \boldsymbol{f} \,, \tag{39}$$

where

$$\boldsymbol{f} = \boldsymbol{Y}^{\mathrm{t}} \hat{\mathscr{F}} \,. \tag{40}$$

Each componential equation of (39) has the form

$$\chi_{i,t} + \lambda_i \chi_{i,x} = \kappa \chi_{i,xx} + f_i \,, \tag{41}$$

where $1 \leqslant i \leqslant m$. The underlining of the indices indicates that *no* sum is implied. This is a scalar advection-diffusion equation which can be solved numerically by the streamline-upwind/Petrov–Galerkin method (SUPG). It is our intent to derive a generalization of SUPG for (38) such that each componential equation is treated in an optimal manner.

3.1. *Variational equation for the symmetric form with Euclidean metric*

Consider the following semidiscrete SUPG-type variational equation for (38):

$$\int_0^L (\boldsymbol{Z}^h \cdot \hat{\boldsymbol{A}} \boldsymbol{X}^h_{,x} + \boldsymbol{Z}^h_{,x} \cdot (\kappa \boldsymbol{X}^h_{,x})) \, dx + \sum_{e=1}^{n_{el}} \int_{\Omega^e} \boldsymbol{P}^h \cdot (\hat{\boldsymbol{A}} \boldsymbol{X}^h_{,x} - (\kappa \boldsymbol{X}^h_{,x})_{,x} - \hat{\mathscr{F}} + \boldsymbol{X}^h_{,t}) \, dx$$
$$= \int_0^L \boldsymbol{Z}^h \cdot (\hat{\mathscr{F}} - \boldsymbol{X}^h_{,t}) \, dx + H(\boldsymbol{Z}^h) \,, \tag{42}$$

where $\boldsymbol{X}^h$ is the trial solution, $\boldsymbol{Z}^h$ is the weighting function, $\boldsymbol{P}^h(\boldsymbol{Z}^h)$ is the 'perturbation' to the weighting function, Ω^e is the interior of the eth element, n_{el} is the number of elements, and $H(\boldsymbol{Z}^h)$ is a prescribed natural boundary condition term. We assume that the trial solution and weighting function are expressed in the usual way in terms of typical C^0 finite element basis functions (e.g., piecewise linears). Note that if $\boldsymbol{P}^h \equiv \boldsymbol{0}$ we have the classical Galerkin formulation, otherwise we have a Petrov–Galerkin formulation. For simplicity, assume that Dirichlet boundary conditions are specified. This assumption results in $H(\boldsymbol{Z}^h) = 0$.

In order to complete the specification of the method, a definition of $\boldsymbol{P}^h(\boldsymbol{Z}^h)$ needs to be given. As a *candidate* definition take

$$\boldsymbol{P}^h(\boldsymbol{Z}^h) = \tau \hat{\boldsymbol{A}} \boldsymbol{Z}^h_{,x} \,, \tag{43}$$

where τ is a parameter which, for the moment, is left undefined. Use of (43) has been made previously (see e.g. [8, 9]). However, as we shall see, (43) is *not adequate* in the present situation. To analyze the effect of (43), we employ the following change of variables:

$$\boldsymbol{X}^h = \boldsymbol{Y} \boldsymbol{\chi}^h \,, \qquad \boldsymbol{Z}^h = \boldsymbol{Y} \boldsymbol{\zeta}^h \,. \tag{44}$$

Substituting (44) into (42) leads to

$$\int_0^L (\boldsymbol{\zeta}^h \cdot \boldsymbol{\Lambda} \boldsymbol{\chi}^h_{,x} + \boldsymbol{\zeta}^h_{,x} \cdot (\kappa \boldsymbol{\chi}^h_{,x})) \, dx + \sum_{e=1}^{n_{el}} \int_{\Omega^e} \tau \boldsymbol{\zeta}^h \cdot \boldsymbol{\Lambda} (\boldsymbol{\Lambda} \boldsymbol{\chi}^h_{,x} - (\kappa \boldsymbol{\chi}^h_{,x})_{,x} + \boldsymbol{\chi}^h_{,t} - \boldsymbol{f}) \, dx$$
$$= \int_0^L \boldsymbol{\zeta}^h \cdot (\boldsymbol{f} - \boldsymbol{\chi}^h_{,t}) \, dx + 0 \,. \tag{45}$$

Observe that (45) represents an uncoupled system of scalar variational equations. Consequently, they can be examined one at a time. For this purpose let

$$(\boldsymbol{\zeta}^h)^t = (0, \ldots, 0, \underset{\substack{\uparrow \\ i\text{th component}}}{\zeta^h}, 0, \ldots, 0) \,. \tag{46}$$

With this choice of weighting function, (45) becomes

$$\int_0^L (\zeta^h \lambda_i \chi^h_{i,x} + \zeta^h_{,x}(\kappa \chi^h_{i,x}))\,dx + \sum_{e=1}^{n_{el}} \int_{\Omega^e} \tau \zeta^h_{,x} \lambda_i (\lambda_i \chi^h_{i,x} - (\kappa \chi^h_{i,x})_{,x} + \chi^h_{i,t} - f_i)\,dx$$

$$= \int_0^L \zeta^h (f_i - \chi^h_{i,t})\,dx + 0\,. \tag{47}$$

For definiteness, assume the steady case, a uniform mesh of piecewise linear elements, and f_i is constant. It is well known (see e.g. [1]) that in this case optimal (i.e., nodally exact) behavior is attained if τ is set as follows:

$$\tau = \tfrac{1}{2} h \tilde{\xi}(\alpha_i)/|\lambda_i|\,, \tag{48}$$

where

$$\alpha_i = \tfrac{1}{2}|\lambda_i| h/\kappa \quad \text{(element Peclet number of the } i\text{th component)}\,, \tag{49}$$

$$\tilde{\xi}(\alpha_i) = \coth(\alpha_i) - \alpha_i^{-1}\,, \tag{50}$$

and h is the element length. (The function (50) is alluded to throughout this paper. Note that $\tilde{\xi}(\alpha): \mathbb{R}^+ \to \mathbb{R}^+$, where $\mathbb{R}^+ = [0, \infty[$, and $\tilde{\xi}(\alpha) \to 1$ as $\alpha \to \infty$ and $\tilde{\xi}(\alpha) \to \frac{1}{3}\alpha$ as $\alpha \to 0$. If $\alpha = 0$, $\tilde{\xi}(\alpha) = 0$ and the method reduces to the Galerkin formulation which is optimal for this case. For alternatives, see [7, Appendix A, (A.6)].) Because we have only one τ to work with, there is no possibility of simultaneously attaining optimal behavior with respect to all m components of the system. If τ is too small for a particular component, spurious oscillations will result for that component. If τ is too big, overly diffuse results ensue. Clearly, a *distinct* τ is needed for each component.

This reasoning leads us to consider the following definition of $\boldsymbol{P}^h(\boldsymbol{Z}^h)$:

$$\boldsymbol{P}^h(\boldsymbol{Z}^h) = \hat{\boldsymbol{\tau}} \hat{\boldsymbol{A}} \boldsymbol{Z}^h_{,x}\,, \tag{51}$$

where the *matrix* $\hat{\boldsymbol{\tau}}$ is defined by

$$\hat{\boldsymbol{\tau}} = \boldsymbol{Y}\,\mathrm{diag}(\tau_1, \ldots, \tau_m)\boldsymbol{Y}^{\mathrm{t}}\,, \tag{52}$$

in which

$$\tau_i = \tfrac{1}{2} h \tilde{\xi}(\alpha_i)/|\lambda_i|\,, \quad 1 \leq i \leq m\,. \tag{53}$$

Note that

$$\hat{\boldsymbol{\tau}} \hat{\boldsymbol{A}} = \boldsymbol{Y}\,\mathrm{diag}(\tau_1, \ldots, \tau_m)\boldsymbol{Y}^{\mathrm{t}}\boldsymbol{Y}\boldsymbol{\Lambda}\boldsymbol{Y}^{\mathrm{t}} = \boldsymbol{Y}\,\mathrm{diag}(\lambda_1\tau_1, \ldots, \lambda_m\tau_m)\boldsymbol{Y}^{\mathrm{t}}\,. \tag{54}$$

Substituting (51) into (42), and making use of (44) and (54) leads to (47) with τ replaced by τ_i. This is precisely the expression needed to simultaneously obtain optimal behavior in each of the componential equations. In the case of unequal element lengths, τ_i needs to be set according to

$$\tau_i^e = \tfrac{1}{2} h^e \tilde{\xi}(\alpha_i^e)/|\lambda_i|\,, \tag{55}$$

$$\alpha_i^e = \tfrac{1}{2}|\lambda_i| h^e/\kappa \,. \tag{56}$$

If a general diffusion matrix, $\hat{\boldsymbol{K}}$, is present, κ in (56) should be replaced by

$$\kappa_i = \boldsymbol{v}_i^{\mathrm{t}} \hat{\boldsymbol{K}} \boldsymbol{v}_i \,. \tag{57}$$

The rationale for selecting this expression is discussed more fully in the subsequent section.

For future reference it is helpful to introduce a few additional notations. If a matrix such as $\hat{\boldsymbol{A}}$ can be written in the form (32), then for any function $f: \mathbb{R} \to \mathbb{R}$, we may define a function $\boldsymbol{f}: \mathbb{R}^{m^2} \to \mathbb{R}^{m^2}$ by

$$\boldsymbol{f}(\hat{\boldsymbol{A}}) = \boldsymbol{Y} \operatorname{diag}(f(\lambda_1), \ldots, f(\lambda_m)) \boldsymbol{Y}^{-1} \,. \tag{58}$$

For example, if $f(x) = |x|$, then

$$|\hat{\boldsymbol{A}}| = \boldsymbol{Y} \operatorname{diag}(|\lambda_1|, \ldots, |\lambda_m|) \boldsymbol{Y}^{-1} \,. \tag{59}$$

Likewise, if $f(x) = \operatorname{sgn} x$, then

$$\operatorname{sgn} \hat{\boldsymbol{A}} = \boldsymbol{Y} \operatorname{diag}(\operatorname{sgn} \lambda_1, \ldots, \operatorname{sgn} \lambda_m) \boldsymbol{Y}^{-1} \,. \tag{60}$$

3.2. *Examination of* $\boldsymbol{P}^h(\boldsymbol{Z}^h)$

Combining (51)–(54) yields

$$\begin{aligned} \boldsymbol{P}^h(\boldsymbol{Z}^h) &= \boldsymbol{Y} \operatorname{diag}(\tau_1 \lambda_1, \ldots, \tau_m \lambda_m) \boldsymbol{Y}^{\mathrm{t}} \boldsymbol{Z}^h_{,x} \\ &= \tfrac{1}{2} h \boldsymbol{Y} \operatorname{diag}(\tilde{\xi}_1 \operatorname{sgn} \lambda_1, \ldots, \tilde{\xi}_m \operatorname{sgn} \lambda_m) \boldsymbol{Y}^{\mathrm{t}} \boldsymbol{Z}^h_{,x} \\ &= \tfrac{1}{2} h \boldsymbol{Y} \operatorname{diag}(\tilde{\xi}_1, \ldots, \tilde{\xi}_m) \boldsymbol{Y}^{\mathrm{t}} \boldsymbol{Y} \operatorname{diag}(\operatorname{sgn} \lambda_1, \ldots, \operatorname{sgn} \lambda_m) \boldsymbol{Y}^{\mathrm{t}} \boldsymbol{Z}^h_{,x} \\ &= \tfrac{1}{2} h \tilde{\boldsymbol{\xi}} (\operatorname{sgn} \hat{\boldsymbol{A}}) \boldsymbol{Z}^h_{,x} = \tfrac{1}{2} h (\operatorname{sgn} \hat{\boldsymbol{A}}) \tilde{\boldsymbol{\xi}} \boldsymbol{Z}^h_{,x} \,, \end{aligned} \tag{61}$$

where $\tilde{\xi}_i = \tilde{\xi}(\alpha_i)$ and $\tilde{\boldsymbol{\xi}} = \boldsymbol{Y} \operatorname{diag}(\tilde{\xi}_1, \ldots, \tilde{\xi}_m) \boldsymbol{Y}^{\mathrm{t}}$. In the advection-dominated case in which $\tilde{\xi}_i \to 1$ $(1 \leq i \leq m)$, $\tilde{\boldsymbol{\xi}} \to \boldsymbol{I}_m$, and therefore

$$\boldsymbol{P}^h(\boldsymbol{Z}^h) = \hat{\boldsymbol{\tau}} \hat{\boldsymbol{A}} \boldsymbol{Z}^h_{,x} \to \tfrac{1}{2} h (\operatorname{sgn} \hat{\boldsymbol{A}}) \boldsymbol{Z}^h_{,x} \,. \tag{62}$$

This expression is to be used for the *hyperbolic* case. It is apparent from (62) that, assuming $|\hat{\boldsymbol{A}}|$ has rank m,

$$\hat{\boldsymbol{\tau}} \to \tfrac{1}{2} h |\hat{\boldsymbol{A}}|^{-1} \tag{63}$$

in the advection-dominated case.

3.3. Artificial diffusivity

The term

$$(\hat{\tau}\hat{A}\boldsymbol{Z}^h_{,x})\cdot(\hat{A}\boldsymbol{X}^h_{,x}+\cdots)=\boldsymbol{Z}^h_{,x}\cdot\hat{A}\hat{\tau}\hat{A}\boldsymbol{X}^h_{,x}+\cdots \tag{64}$$

in the variational equation gives rise to what may be interpreted as 'artificial diffusivity' and thus

$$\hat{\boldsymbol{K}}^e=\hat{A}\hat{\tau}\hat{A} \tag{65}$$

represents an *artificial diffusivity* matrix. Employing (53) and (54) in (65) results in

$$\hat{\boldsymbol{K}}^e=\tfrac{1}{2}h\tilde{\xi}(\operatorname{sgn}\hat{A})\hat{A}=\tfrac{1}{2}h\tilde{\xi}|\hat{A}|=\tfrac{1}{2}h|\hat{A}|\,\tilde{\xi}\,. \tag{66}$$

In general, $\hat{\boldsymbol{K}}^e$ is symmetric and positive-semidefinite.

In the advection-dominated limit,

$$\hat{\boldsymbol{K}}^e\to\tfrac{1}{2}h|\hat{A}|\,. \tag{67}$$

3.4. Variational equation for the symmetric form with Riemannian metric

Making use of the transformation rules developed in the previous section, (19)–(22), and making use of the change of variables

$$\boldsymbol{X}^h=\boldsymbol{L}^{\mathrm{t}}\boldsymbol{V}^h\,,\qquad \boldsymbol{Z}^h=\boldsymbol{L}^{\mathrm{t}}\boldsymbol{W}^h\,, \tag{68}$$

we can convert (42) to the desired form:

$$\int_0^L(\boldsymbol{W}^h\cdot\tilde{A}\boldsymbol{V}^h_{,x}+\boldsymbol{W}^h_{,x}\tilde{\boldsymbol{K}}\boldsymbol{V}^h_{,x})\,\mathrm{d}x+\sum_{e=1}^{n_{\mathrm{el}}}\int_{\Omega^e}(\tilde{\tau}\tilde{A}\boldsymbol{W}^h_{,x})\cdot(\tilde{A}\boldsymbol{V}^h_{,x}-(\tilde{\boldsymbol{K}}\boldsymbol{V}^h_{,x})_{,x}-\mathscr{F}+\boldsymbol{A}_0\boldsymbol{V}^h_{,t})\,\mathrm{d}x$$

$$=\int_0^L\boldsymbol{W}^h\cdot(\mathscr{F}-\boldsymbol{A}_0\boldsymbol{V}^h_{,t})\,\mathrm{d}x+\text{natural boundary condition terms}, \tag{69}$$

where

$$\tilde{\tau}=\boldsymbol{L}^{-\mathrm{t}}\hat{\tau}\boldsymbol{L}^{-1}=\boldsymbol{L}^{-\mathrm{t}}\boldsymbol{Y}\operatorname{diag}(\tau_1,\ldots,\tau_m)\boldsymbol{Y}^{\mathrm{t}}\boldsymbol{L}^{-1}=\boldsymbol{\Phi}\operatorname{diag}(\tau_1,\ldots,\tau_m)\boldsymbol{\Phi}^{\mathrm{t}}\,, \tag{70}$$

$$\boldsymbol{\Phi}=[\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_m]\,. \tag{71}$$

The $\boldsymbol{\varphi}_i$ are the eigenvectors of problem (28) and we have made use of (30) in deriving (70). Note that in going from (42) to (69) we have replaced the simple isotropic diffusivity with a general diffusivity matrix. Clearly, in (69) and (70) we have used the definition (51). Note that the κ_i employed to define the τ_i are invariant in the sense that

$$\kappa_i=\boldsymbol{v}_i^{\mathrm{t}}\hat{\boldsymbol{K}}\boldsymbol{v}_i=(\boldsymbol{L}^{\mathrm{t}}\boldsymbol{\varphi}_i)^{\mathrm{t}}\boldsymbol{L}^{-1}\tilde{\boldsymbol{K}}\boldsymbol{L}^{-\mathrm{t}}(\boldsymbol{L}^{\mathrm{t}}\boldsymbol{\varphi}_i)=\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{\varphi}_i\,. \tag{72}$$

An artificial diffusivity matrix for the present case is seen to be

$$\tilde{K}^e = \tilde{A}\tilde{\tau}\tilde{A} \,. \tag{73}$$

Equation (69) is a suitable semidiscrete SUPG-type formulation for one-dimensional systems. In the next section, we shall generalize this formulation to multidimensional systems.

3.5. *Physical interpretation of the* τ_i

The τ_i are intrinsic time scales of the various components of the solution. In the advection-dominated limit, the τ_i represent the transit times for information to be advected over a distance equal to one-half the element length. These times are reduced by the presence of diffusion. The τ_i approach zero as diffusion begins to dominate in keeping with the instantaneous propagation time of diffusive phenomena. Due to these facts we shall refer to $\tilde{\tau}$ as the *matrix of intrinsic time scales*.

3.6. *Why Lax–Wendroff-type methods are inadequate for systems*

Lax–Wendroff-type methods and their finite element counterparts, Taylor–Galerkin methods [2], because of their simplicity and generality, are widely used for analyzing systems of the type considered herein. We wish to point out that when $m > 1$ and the eigenvalues are not all identical, methods of this type simply *cannot* perform well in general. To construct an example which illustrates this point make the following assumptions: (i) the diffusion term is small in the sense that the element Peclet number is very large, (ii) there is no source term, (iii) a steady solution has been reached, and (iv) a uniform mesh of piecewise linear elements is employed. Assumption (i) enables us to neglect diffusion terms in the discrete equations. However, the fact that diffusion is not identically zero allows retaining Dirichlet boundary conditions which lead to sharp boundary layers. Employing the assumptions in (42) and selecting $\boldsymbol{P}^h(\boldsymbol{Z}^h)$ according to the scalar-τ definition, (43), yields

$$0 = \int_0^L (\boldsymbol{Z}^h \cdot \hat{\boldsymbol{A}}\boldsymbol{X}^h_{,x} + \boldsymbol{Z}^h_{,x} \cdot \tau \hat{\boldsymbol{A}}^2 \boldsymbol{X}^h_{,x}) \, \mathrm{d}x \,. \tag{74}$$

With $\tau = \frac{1}{2}\Delta t$, this is the Taylor–Galerkin, or Lax–Wendroff, finite element equation governing the discrete solution. This value of τ may be compared with the optimal value for the ith componential equation, namely $\tau_i = \frac{1}{2}h/|\lambda_i|$. Clearly, if the values of $|\lambda_i|$ vary considerably, the single value $\tau = \frac{1}{2}\Delta t$ cannot simultaneously achieve good behavior with respect to each component. As Lax–Wendroff-type methods are typically explicit, a Courant condition must be respected. Assuming 'lumped' mass is employed, this takes the form

$$\mathrm{Cr} = \frac{\Delta t}{h} \max_i |\lambda_i| \leq 1 \,. \tag{75}$$

In other words, numerical stability dictates that Δt be sufficiently small. In terms of the τ_i, (75) may be expressed as

$$\tau \leq \tau_i \quad (1 \leq i \leq m) \,. \tag{76}$$

Thus oscillatory behavior will be present in all components for which $\tau < \tau_i$. (This seems to explain why computer programs based on the Lax–Wendroff concept invariably employ additional artificial viscosity mechanisms such as, for example, Lapidus or MacCormack viscosity.) Observe that as $\tau/\tau_i \to 0$, the method becomes the Galerkin method which is potentially wildly oscillatory. Thus reducing Δt below the critical value actually has an adverse effect.

Even if the stability problem is circumvented, enabling a larger value of τ, simultaneous control of all components of a system is impossible, because

$$\tau < \tau_i \Rightarrow \text{oscillatory behavior},$$
$$\tau = \tau_i \Rightarrow \text{optimal behavior},$$
$$\tau > \tau_i \Rightarrow \text{overly diffuse behavior}.$$

Thus, assuming distinct τ_i, only one component will behave optimally at most. The only way to get around this fundamental difficulty seems to be to adopt the philosophy presented herein and replace the scalar τ by a matrix designed to exercise optimal control over each component of the system independently.

4. Multidimensional systems

Our plan of attack is the same as in Section 3. We will first derive an SUPG-type formulation for the symmetric form of the equations with Euclidean metric, then transform to the case of Riemannian metric. We begin with the variational equation for the multidimensional Euclidean-metric case (cf. (42)):

$$\int_\Omega (\mathbf{Z}^h \cdot \hat{\boldsymbol{A}} \cdot \nabla \mathbf{X}^h + \nabla \mathbf{Z}^h \cdot (\hat{\boldsymbol{K}} \nabla \mathbf{X}^h))\, d\Omega$$
$$+ \sum_{e=1}^{n_{el}} \int_{\Omega^e} (\hat{\tau} \hat{\boldsymbol{A}} \cdot \nabla \mathbf{Z}^h) \cdot (\hat{\boldsymbol{A}} \cdot \nabla \mathbf{X}^h - \nabla \cdot (\hat{\boldsymbol{K}} \nabla \mathbf{X}^h) - \hat{\mathscr{F}} + \mathbf{X}^h_{,t})\, d\Omega$$
$$= \int_\Omega \mathbf{Z}^h \cdot (\hat{\mathscr{F}} - \mathbf{X}^h_{,t})\, d\Omega + \text{natural boundary conditions terms}. \tag{77}$$

A definition of $\hat{\tau}$ is all that is needed to complete the specification of the method. Note that this time the artificial diffusivity matrix may be written as

$$\hat{\boldsymbol{K}}^e = \hat{\boldsymbol{A}} \hat{\tau} \hat{\boldsymbol{A}}^t, \tag{78}$$

where $\hat{\tau}$ is assumed symmetric and positive-semidefinite. Note also that $\hat{\tau}$ has dimension $m \times m$ irrespective of the number of space dimensions (i.e., 'd'), whereas $\hat{\boldsymbol{K}}^e$ has dimension $(m \cdot d) \times (m \cdot d)$.

Temporarily we will assume that

$$\hat{\boldsymbol{K}} = \kappa \boldsymbol{I}_{m \cdot d} \tag{79}$$

and that κ is 'small'. Thus we will attempt to define $\hat{\tau}$ for the advection-dominated case. Recall that in one dimension, simple and suggestive formulae were developed for this case, namely:

$$\hat{\tau} = \tfrac{1}{2} h |\hat{\boldsymbol{A}}|^{-1}, \qquad \hat{\boldsymbol{K}} = \tfrac{1}{2} h |\hat{\boldsymbol{A}}| \quad \text{(one-dimensional, advection-dominated limit).} \tag{80}$$

Recall that the first part of (80) is only meaningful when $|\hat{\boldsymbol{A}}|$ has rank m. It is apparent from (80) that the notion of the absolute value of a matrix needs to be generalized. By virtue of the fact that when $d > 1$, $\hat{\boldsymbol{A}}$ is not a square matrix, the generalization is not immediately obvious.

We shall require that the definition of $\hat{\tau}$ for multidimensional systems satisfies the following design conditions:

(i) For $m > 1$ and $d = 1$ (i.e., one-dimensional advective-diffusive systems) it should reduce to the formulation presented in Section 3.

(ii) For $m = 1$ and $d > 1$ (i.e., the multidimensional scalar advection-diffusion equation) it should coincide with the SUPG formulation presented in Part II of this series of papers (i.e., [7]).

(iii) For $m > 1$, $d > 1$, and 'simultaneously diagonalizable' coefficient matrices, that is, the $\hat{\boldsymbol{A}}_i$ satisfy

$$\hat{\boldsymbol{A}}_i = \boldsymbol{Y} \operatorname{diag}(a_i^1, \ldots, a_i^m) \boldsymbol{Y}^{-1}, \tag{81}$$

where $\boldsymbol{Y}$ is orthogonal, the definition of $\hat{\tau}$ should reduce to the SUPG formulation of Part II, [7], with respect to each uncoupled scalar equation.

4.1. *Definition of $\hat{\tau}$ for the advection-dominated case*

Let us assume we are working with isoparametric finite elements and that $\boldsymbol{x}^e = \boldsymbol{x}^e(\boldsymbol{\xi})$ denotes the geometric mapping from the parent domain to $\Omega^e \subset \mathbb{R}^d$. The inverse mapping is written $\boldsymbol{\xi}^e = \boldsymbol{\xi}^e(\boldsymbol{x})$. The following definitions are basic to the multidimensional formulation:

$$\hat{\boldsymbol{B}}_i = (\partial \xi_i / \partial x_j) \hat{\boldsymbol{A}}_j, \tag{82}$$

$$\hat{\boldsymbol{B}}^{\mathrm{t}} = (\hat{\boldsymbol{B}}_1^{\mathrm{t}}, \ldots, \hat{\boldsymbol{B}}_d^{\mathrm{t}}). \tag{83}$$

Observe that if the element under consideration is a rectangular parallelopiped aligned with the global Cartesian coordinates and that if the nodal points are uniformly spaced, then

$$\hat{\boldsymbol{B}}_i = (2/h_i) \hat{\boldsymbol{A}}_i, \tag{84}$$

where h_i is the length of the side aligned with the x_i-direction. In particular, if the element is a d-cube, then $\partial \xi_i / \partial x_j = (2/h) \delta_{ij}$, where δ_{ij} is the Kronecker delta, and therefore

$$\hat{\boldsymbol{B}}_i = (2/h) \hat{\boldsymbol{A}}_i. \tag{85}$$

The $m \times m$ *matrix-valued p-norm of* $\hat{\boldsymbol{B}}$ is defined as the following symmetric, positive-semidefinite matrix:

$$|\hat{\boldsymbol{B}}|_p = \left(\sum_{i=1}^{d} |\hat{\boldsymbol{B}}_i|^p \right)^{1/p} . \tag{86}$$

Two representations of $\hat{\boldsymbol{\tau}}$ are useful:
(a) $|\hat{\boldsymbol{B}}|_p$ *has rank* m. In this case the inverse exists and we simply define $\hat{\tau}$ by

$$\hat{\boldsymbol{\tau}} = |\hat{\boldsymbol{B}}|_p^{-1} . \tag{87}$$

(b) $|\hat{\boldsymbol{B}}_p|$ *has rank* $l \leqslant m$. This is the fully general case. Consider the eigenproblem

$$(|\hat{\boldsymbol{B}}|_p - \mu \boldsymbol{I}_m)\boldsymbol{v} = \boldsymbol{0} . \tag{88}$$

We can assume without loss of generality that $\mu_i > 0$ for $1 \leqslant i \leqslant l$, and $\mu_i = 0$ for $l+1 \leqslant i \leqslant m$. By virtue of (88), $|\hat{\boldsymbol{B}}_p|$ has the representation

$$|\hat{\boldsymbol{B}}|_p = \sum_{i=1}^{m} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{t}} = \sum_{i=1}^{l} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{t}} . \tag{89}$$

We define $\hat{\boldsymbol{\tau}}$ to be the inverse of $|\hat{\boldsymbol{B}}|_p$ on its nondegenerate subspace, that is

$$\hat{\boldsymbol{\tau}} = \sum_{i=1}^{l} \mu_i^{-1} \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{t}} . \tag{90}$$

In matrix notation this can be written as:

$$\hat{\boldsymbol{\tau}} = \boldsymbol{Y}\boldsymbol{M}\boldsymbol{Y}^{\mathrm{t}} = \bar{\boldsymbol{Y}}\bar{\boldsymbol{M}}\bar{\boldsymbol{Y}}^{\mathrm{t}} , \tag{91}$$

where

$$\boldsymbol{Y} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m] \quad (m \times m) , \qquad \bar{\boldsymbol{Y}} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_l] \quad (m \times l) , \tag{92}$$

$$\boldsymbol{M} = \operatorname{diag}(\mu_1^{-1}, \ldots, \mu_l^{-1}, 0, \ldots, 0) \quad (m \times m) ,$$
$$\bar{\boldsymbol{M}} = \operatorname{diag}(\mu_1^{-1}, \ldots, \mu_l^{-1}) \quad (l \times l) . \tag{93}$$

When $l = m$, (91) is equivalent to (87).

4.2. *Verification of design conditions* (i)–(iii)

Now let us consider the requirements that the definition is to satisfy:
(i) $m > 1$, $d = 1$. In this case

$$\hat{\boldsymbol{A}} = \hat{\boldsymbol{A}}_1 \quad (m \times m) , \tag{94}$$

$$\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_1 = (\partial \xi / \partial x)\hat{\boldsymbol{A}}_1 = (2/h)\hat{\boldsymbol{A}} , \tag{95}$$

$$|\hat{\boldsymbol{B}}|_p = (2/h)|\hat{\boldsymbol{A}}|_p = (2/h)|\hat{\boldsymbol{A}}|\,. \tag{96}$$

Assuming that the rank of $|\hat{\boldsymbol{A}}|$ is m,

$$\hat{\boldsymbol{\tau}} = \tfrac{1}{2}h|\hat{\boldsymbol{A}}|^{-1} \quad (m \times m)\,, \tag{97}$$

$$\hat{\boldsymbol{K}}^e = \hat{\boldsymbol{A}}\hat{\boldsymbol{\tau}}\hat{\boldsymbol{A}} = \tfrac{1}{2}|\hat{\boldsymbol{A}}| \quad (m \times m)\,. \tag{98}$$

If the rank of $|\hat{\boldsymbol{A}}|$ is less than m, $\hat{\boldsymbol{\tau}}$ may be defined as in (91). In any event, (98) is still valid. Thus condition (i) is satisfied.

(ii) $m = 1$, $d > 1$. In this case

$$\hat{\boldsymbol{A}}_i = a_i \quad (1 \times 1)\,, \tag{99}$$

$$\hat{\boldsymbol{A}}^{\mathrm{t}} = (a_1, \ldots, a_d) = \boldsymbol{a}^{\mathrm{t}} \quad (d \times 1)\,, \tag{100}$$

$$\hat{\boldsymbol{B}}_i = (\partial\xi_i/\partial x_j)\hat{\boldsymbol{A}}_j = (\partial\xi_i/\partial x_j)a_j \overset{\text{def}}{=} b_i \quad (1 \times 1)\,, \tag{101}$$

$$\hat{\boldsymbol{B}}^{\mathrm{t}} = (b_1, \ldots, b_d) = \boldsymbol{b}^{\mathrm{t}} \quad (d \times 1)\,, \tag{102}$$

$$|\hat{\boldsymbol{B}}|_p = \Big(\sum_{i=1}^{d} |\hat{\boldsymbol{B}}_i|^p\Big)^{1/p} = \Big(\sum_{i=1}^{d} |b_i|^p\Big)^{1/p} \overset{\text{def}}{=} |\boldsymbol{b}|_p \quad (1 \times 1)\,, \tag{103}$$

$$\hat{\boldsymbol{\tau}} = |\hat{\boldsymbol{B}}|_p^{-1} = |\boldsymbol{b}|_p^{-1} \quad (1 \times 1)\,, \tag{104}$$

$$\hat{\boldsymbol{K}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{\tau}}\hat{\boldsymbol{A}}^{\mathrm{t}} = |\boldsymbol{b}|_p^{-1}\boldsymbol{a}\boldsymbol{a}^{\mathrm{t}} \quad (d \times d, \text{'streamline diffusion'})\,. \tag{105}$$

These results agree with the formulation presented in [7] and thus condition (ii) is satisfied.

(iii) $m > 1$, $d > 1$, *simultaneously diagonalizable coefficient matrices*. Let us first calculate $\hat{\boldsymbol{\tau}}$:

$$\begin{aligned}\hat{\boldsymbol{B}}_i &= (\partial\xi_i/\partial x_j)\hat{\boldsymbol{A}}_j = (\partial\xi_i/\partial x_j)\boldsymbol{Y}\,\mathrm{diag}(a_j^1, \ldots, a_j^m)\boldsymbol{Y}^{\mathrm{t}} \\ &= \boldsymbol{Y}\,\mathrm{diag}((\partial\xi_i/\partial x_j)a_j^1, \ldots, (\partial\xi_i/\partial x_j)a_j^m)\boldsymbol{Y}^{\mathrm{t}} = \boldsymbol{Y}\,\mathrm{diag}(b_i^1, \ldots, b_i^m)\boldsymbol{Y}^{\mathrm{t}}\,,\end{aligned} \tag{106}$$

$$\begin{aligned}|\hat{\boldsymbol{B}}|_p &= \Big(\sum_{i=1}^{d} |\hat{\boldsymbol{B}}_i|^p\Big)^{1/p} = \Big(\boldsymbol{Y}\sum_{i=1}^{d} |\mathrm{diag}(b_i^1, \ldots, b_i^m)|^p\boldsymbol{Y}^{\mathrm{t}}\Big)^{1/p} \\ &= \boldsymbol{Y}\Big(\sum_{i=1}^{d} |\mathrm{diag}(b_i^1, \ldots, b_i^m)|^p\Big)^{1/p}\boldsymbol{Y}^{\mathrm{t}} = \boldsymbol{Y}\,\mathrm{diag}(|\boldsymbol{b}^1|_p, \ldots, |\boldsymbol{b}^m|_p)\boldsymbol{Y}^{\mathrm{t}}\,,\end{aligned} \tag{107}$$

where

$$(\boldsymbol{b}^i)^{\mathrm{t}} = (b_1^i, \ldots, b_d^i)\,. \tag{108}$$

Therefore

$$\hat{\boldsymbol{\tau}} = |\hat{\boldsymbol{B}}|_p^{-1} = \boldsymbol{Y}\,\mathrm{diag}(|\boldsymbol{b}^1|_p^{-1}, \ldots, |\boldsymbol{b}^m|_p^{-1})\boldsymbol{Y}^{\mathrm{t}}\,. \tag{109}$$

To check that this definition correctly uncouples the variational equation, consider the change of variables

$$\boldsymbol{X}^h = \boldsymbol{Y}\boldsymbol{\chi}^h\,, \qquad \boldsymbol{Z}^h = \boldsymbol{Y}\boldsymbol{\zeta}^h\,, \tag{110}$$

in the various terms of (77):

$$\boldsymbol{Z}^h \cdot \hat{\boldsymbol{A}}\nabla\boldsymbol{X}^h = \boldsymbol{Z}^h \cdot \hat{\boldsymbol{A}}_i \boldsymbol{X}^h_{,i} = \boldsymbol{\zeta}^h \cdot (\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{A}}_i\boldsymbol{Y})\boldsymbol{\chi}^h_{,i} = \boldsymbol{\zeta}^h \cdot \operatorname{diag}(a_i^1, \ldots, a_i^m)\boldsymbol{\chi}^h_{,i}\,, \tag{111}$$

$$\begin{aligned}
\nabla\boldsymbol{Z}^h \cdot \hat{\boldsymbol{A}}\hat{\boldsymbol{\tau}}\hat{\boldsymbol{A}}^{\mathrm{t}}\nabla\boldsymbol{X}^h &= \boldsymbol{Z}^h_{,i} \cdot \hat{\boldsymbol{A}}_i\hat{\boldsymbol{\tau}}\hat{\boldsymbol{A}}_j\boldsymbol{X}^h_{,j} \\
&= \boldsymbol{\zeta}^h_{,i} \cdot (\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{A}}_i\boldsymbol{Y})(\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{\tau}}\boldsymbol{Y})(\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{A}}_j\boldsymbol{Y})\boldsymbol{\chi}^h_{,j} \\
&= \boldsymbol{\zeta}^h_{,i} \cdot \operatorname{diag}(a_i^1, \ldots, a_i^m)\operatorname{diag}(|\boldsymbol{b}^1|_p^{-1}, \ldots, |\boldsymbol{b}^m|_p^{-1})\operatorname{diag}(a_j^1, \ldots, a_j^m)\boldsymbol{\chi}^h_{,j} \\
&= \boldsymbol{\zeta}^h_{,i} \cdot \operatorname{diag}(|\boldsymbol{b}^1|_p^{-1}a_i^1a_j^1, \ldots, |\boldsymbol{b}^m|_p^{-1}a_i^ma_j^m)\boldsymbol{\chi}^h_{,j}\,.
\end{aligned} \tag{112}$$

It is also a worthwhile exercise at this point to check that the diffusion terms also uncouple appropriately. For this purpose we need to assume

$$\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{K}}_{ij}\boldsymbol{Y} = \delta_{ij}\operatorname{diag}(\kappa_1, \ldots, \kappa_m)\,. \tag{113}$$

(Equation (113) is less restrictive than (79).) With this assumption, the diffusion terms of (77) become:

$$\nabla\boldsymbol{Z}^h \cdot \hat{\boldsymbol{K}}\nabla\boldsymbol{X}^h = \boldsymbol{Z}^h_{,i} \cdot \hat{\boldsymbol{K}}_{ij}\boldsymbol{X}^h_{,j} = \boldsymbol{\zeta}^h_{,i} \cdot (\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{K}}_{ij}\boldsymbol{Y})\boldsymbol{\chi}^h_{,j} = \boldsymbol{\zeta}^h_{,i} \cdot \operatorname{diag}(\kappa_1, \ldots, \kappa_m)\boldsymbol{\chi}^h_{,i}\,, \tag{114}$$

$$\begin{aligned}
\nabla\boldsymbol{Z}^h \cdot \hat{\boldsymbol{A}}\hat{\boldsymbol{\tau}}\nabla \cdot \hat{\boldsymbol{K}}\nabla\boldsymbol{X}^h &= \boldsymbol{Z}^h_{,k} \cdot \hat{\boldsymbol{A}}_k\hat{\boldsymbol{\tau}}(\hat{\boldsymbol{K}}_{ij}\boldsymbol{X}^h_{,j})_{,i} \\
&= \boldsymbol{\zeta}^h_{,k} \cdot (\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{A}}_k\boldsymbol{Y})(\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{\tau}}\boldsymbol{Y})((\boldsymbol{Y}^{\mathrm{t}}\hat{\boldsymbol{K}}_{ij}\boldsymbol{Y})\boldsymbol{\chi}_{,j})_{,i} \\
&= \boldsymbol{\zeta}^h_{,k} \cdot \operatorname{diag}(a_k^1, \ldots, a_k^m)\operatorname{diag}(|\boldsymbol{b}^1|_p^{-1}, \ldots, |\boldsymbol{b}^m|_p^{-1}) \\
&\quad \times (\operatorname{diag}(\kappa_1, \ldots, \kappa_m)\boldsymbol{\chi}_{,i})_{,i}\,.
\end{aligned} \tag{115}$$

Examination of (111), (112), (114), and (115) reveals that the terms of (77) fully uncouple and on each component agree with the SUPG procedure. Recall that the definition of $\hat{\boldsymbol{\tau}}$ is only valid for the case in which diffusion is small. However, when this assumption is dropped and $\hat{\boldsymbol{\tau}}$ is suitably generalized, conditions (i)–(iii) will still hold.

4.2.1. Remarks

(1) The following values of p are of particular interest: $p = 1, 2$, and ∞. The 1-norm requires the absolute values of each of the $\hat{\boldsymbol{B}}_i$, viz.

$$|\hat{\boldsymbol{B}}|_1 = \sum_{i=1}^{d} |\hat{\boldsymbol{B}}_i|\,.$$

The 2-norm involves calculating the square root of a matrix:

$$|\hat{\boldsymbol{B}}|_2 = \left(\sum_{i=1}^{d} |\hat{\boldsymbol{B}}_i|^2\right)^{1/2} = \left(\sum_{i=1}^{d} \hat{\boldsymbol{B}}_i^2\right)^{1/2} = (\hat{\boldsymbol{B}}^{\mathrm{t}}\hat{\boldsymbol{B}})^{1/2} . \tag{116}$$

It is interesting to note that, with the aid of the Cayley–Hamilton theorem, (116) can be calculated directly avoiding solution of the eigenproblem (see [5; 10, p. 55]). However, in order to properly account for diffusive effects, the eigenvectors seem essential.

The ∞-norm of a matrix is apparently difficult to compute. Our choice so far in practical computing has been to work with the 2-norm.

(2) The $\hat{\boldsymbol{B}}_i$ play the same role with respect to ξ-coordinates that the $\hat{\boldsymbol{A}}_i$ play with respect to x-coordinates:

$$\hat{\boldsymbol{A}}_j \frac{\partial}{\partial x_j} = \frac{\partial \xi_i}{\partial x_j} \hat{\boldsymbol{A}}_j \frac{\partial}{\partial \xi_i} = \hat{\boldsymbol{B}}_i \frac{\partial}{\partial \xi_i} \quad \Leftrightarrow \quad \hat{\boldsymbol{B}} \cdot \nabla_\xi = \hat{\boldsymbol{A}} \cdot \nabla . \tag{117}$$

(3) Define an artificial diffusivity with respect to ξ-coordinates by requiring that

$$\nabla_\xi \boldsymbol{Z}^h \cdot \hat{\boldsymbol{K}}_\xi^e \nabla_\xi \boldsymbol{X}^h = \nabla \boldsymbol{Z}^h \cdot \hat{\boldsymbol{K}}^e \nabla \boldsymbol{X}^h . \tag{118}$$

It immediately follows that

$$\hat{\boldsymbol{K}}_\xi^e = \hat{\boldsymbol{B}} \hat{\boldsymbol{\tau}} \hat{\boldsymbol{B}}^{\mathrm{t}} , \tag{119}$$

which, in the advection-dominated limit, becomes

$$\hat{\boldsymbol{K}}_\xi^e = \hat{\boldsymbol{B}} |\hat{\boldsymbol{B}}|_p^{-1} \hat{\boldsymbol{B}}^{\mathrm{t}} . \tag{120}$$

(4) Consider (120) for the case $p = 2$. We claim that

$$\hat{\boldsymbol{K}}_\xi^e = (\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^{\mathrm{t}})^{1/2} . \tag{121}$$

This can be established as follows:

$$\hat{\boldsymbol{K}}_\xi^e = \hat{\boldsymbol{B}} |\hat{\boldsymbol{B}}|_2^{-1} \hat{\boldsymbol{B}}^{\mathrm{t}} = \hat{\boldsymbol{B}} (\hat{\boldsymbol{B}}^{\mathrm{t}} \hat{\boldsymbol{B}})^{-1/2} \hat{\boldsymbol{B}}^{\mathrm{t}} . \tag{122}$$

Therefore

$$(\hat{\boldsymbol{K}}_\xi^e)^2 = (\hat{\boldsymbol{B}} (\hat{\boldsymbol{B}}^{\mathrm{t}} \hat{\boldsymbol{B}})^{-1/2} \hat{\boldsymbol{B}}^{\mathrm{t}})(\hat{\boldsymbol{B}} (\hat{\boldsymbol{B}}^{\mathrm{t}} \hat{\boldsymbol{B}})^{-1/2} \hat{\boldsymbol{B}}^{\mathrm{t}}) = \hat{\boldsymbol{B}} \hat{\boldsymbol{B}}^{\mathrm{t}} , \tag{123}$$

which verifies (121). *This result is clearly a generalization of the absolute-value operator for systems*. The following remark provides additional emphasis of this point.

(5) Assume an isoparametric d-cube in which nodal points are uniformly spaced. Recall from (85) that $\hat{\boldsymbol{B}}_i = (2/h)\hat{\boldsymbol{A}}_i$. In this case

$$\hat{\boldsymbol{K}}^e = \tfrac{1}{2} h \hat{\boldsymbol{A}} |\hat{\boldsymbol{A}}|_p^{-1} \hat{\boldsymbol{A}}^{\mathrm{t}} , \tag{124}$$

and, in particular, if $p = 2$, then

$$\hat{\boldsymbol{K}}^e = \tfrac{1}{2}h(\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}^{\mathrm{t}})^{1/2} . \tag{125}$$

Often, the absolute-value operator is defined as the square root of the square of the original matrix. Equation (125) extends this idea to the case in which the original matrix, namely $\hat{\boldsymbol{A}}$, is rectangular.

(6) Equation (125) arises if we *define* $\hat{\boldsymbol{\tau}}$ by

$$\hat{\boldsymbol{\tau}} = \tfrac{1}{2}h(\hat{\boldsymbol{A}}^{\mathrm{t}}\hat{\boldsymbol{A}})^{-1/2} . \tag{126}$$

If the elements under consideration are not too distorted with respect to a d-cube, then a simple expression for h in (126) should suffice (see, e.g., Fig. 1). Under these circumstances (126) has calculational advantages compared with

$$\hat{\boldsymbol{\tau}} = (\hat{\boldsymbol{B}}^{\mathrm{t}}\hat{\boldsymbol{B}})^{-1/2} , \tag{127}$$

in that $\hat{\boldsymbol{A}}$ is typically more sparse than $\hat{\boldsymbol{B}}$. However, it is clear that (126) fails to correctly represent severely elongated, or otherwise highly-distorted elements. The virtue of (127) is that all concerns with complex element geometry are obviated from the outset.

(7) The inverses of $\hat{\boldsymbol{A}}^{\mathrm{t}}\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}^{\mathrm{t}}\hat{\boldsymbol{B}}$ have appeared in several expressions. They exist, respectively, if $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{B}}$ have rank m.

4.3. *Variational equation for the symmetric form with Riemannian metric*

In transforming to the weak form applicable to the symmetric case with Riemannian metric, we will simultaneously enlarge the framework to include a general physical diffusivity matrix $\tilde{\boldsymbol{K}}$. (The only restriction is that it be symmetric and positive-semidefinite.)

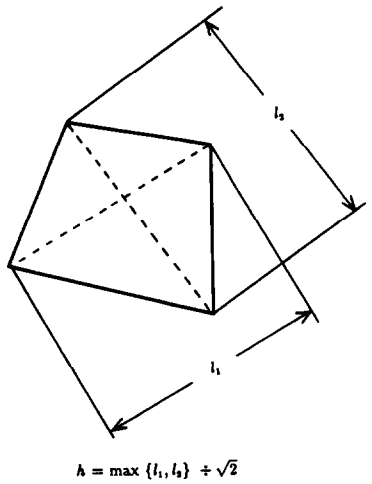

Fig. 1. A simplified definition of the mesh parameter, h, can be used for elements which closely approximate a d-dimensional cube.

Once again employing the change of variables (68), we obtain the desired weak form:

$$\int_\Omega (\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{V}^h + \nabla \boldsymbol{W}^h \cdot (\tilde{\boldsymbol{K}} \nabla \boldsymbol{V}^h))\, \mathrm{d}\Omega$$

$$+ \sum_{e=1}^{n_{\mathrm{el}}} \int_{\Omega^e} (\tilde{\tau} \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{W}^h) \cdot (\tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{V}^h - \nabla \cdot (\tilde{\boldsymbol{K}} \nabla \boldsymbol{V}^h) - \mathscr{F} + \boldsymbol{A}_0 \boldsymbol{V}^h_{,t})\, \mathrm{d}\Omega$$

$$= \int_\Omega \boldsymbol{W}^h \cdot (\mathscr{F} - \boldsymbol{A}_0 \boldsymbol{V}^h_{,t})\, \mathrm{d}\Omega + \text{natural boundary condition terms}, \tag{128}$$

where

$$\tilde{\tau} = \boldsymbol{L}^{-\mathrm{t}} \hat{\tau} \boldsymbol{L}^{-1} = \boldsymbol{L}^{-\mathrm{t}} \bar{\boldsymbol{Y}}\, \mathrm{diag}(\tau_1, \ldots, \tau_l)\, \bar{\boldsymbol{Y}} \boldsymbol{L}^{-1} = \bar{\boldsymbol{\Phi}}\, \mathrm{diag}(\tau_1, \ldots, \tau_l)\, \bar{\boldsymbol{\Phi}}^{\mathrm{t}}, \tag{129}$$

$$\bar{\boldsymbol{\Phi}} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_l], \tag{130}$$

$$\boldsymbol{\varphi}_i = \boldsymbol{L}^{-\mathrm{t}} \boldsymbol{v}_i, \tag{131}$$

$$\tau_i = \tilde{\xi}(\alpha_i)/\mu_i, \tag{132}$$

$$\alpha_i = \mu_i/\sigma_i, \tag{133}$$

$$\sigma_i = \frac{1}{d} \begin{pmatrix} \boldsymbol{\varphi}_i \\ \vdots \\ \boldsymbol{\varphi}_i \end{pmatrix}^{\mathrm{t}} \tilde{\boldsymbol{K}}_\xi \begin{pmatrix} \boldsymbol{\varphi}_i \\ \vdots \\ \boldsymbol{\varphi}_i \end{pmatrix} = \frac{1}{d} \sum_{j,k=1}^{d} \boldsymbol{\varphi}_i^{\mathrm{t}} \tilde{\boldsymbol{K}}_{jk} \boldsymbol{\varphi}_i, \tag{134}$$

$$\tilde{\boldsymbol{K}}_\xi = \begin{pmatrix} (\tilde{\boldsymbol{K}}_\xi)_{11} \cdots (\tilde{\boldsymbol{K}}_\xi)_{1d} \\ \vdots \quad \ddots \quad \vdots \\ (\tilde{\boldsymbol{K}}_\xi)_{d1} \cdots (\tilde{\boldsymbol{K}}_\xi)_{dd} \end{pmatrix}, \tag{135}$$

$$(\tilde{\boldsymbol{K}}_\xi)_{jk} = \sum_{j',k'=1}^{d} \frac{\partial \xi_j}{\partial x_{j'}} \frac{\partial \xi_k}{\partial x_{k'}} \tilde{\boldsymbol{K}}_{j'k'}. \tag{136}$$

The μ_i and $\boldsymbol{\varphi}_i$ are determined by solving an eigenproblem. Several equivalent forms of the eigenproblem are possible (cf. (88)). Let:

$$\tilde{\boldsymbol{B}}_i = \boldsymbol{B}_i \boldsymbol{A}_0 = \frac{\partial \xi_i}{\partial x_j} \boldsymbol{A}_j \boldsymbol{A}_0 = \frac{\partial \xi_i}{\partial x_j} \tilde{\boldsymbol{A}}_j, \tag{137}$$

from which it follows that

$$\hat{\boldsymbol{B}}_i = \boldsymbol{L}^{-1} \tilde{\boldsymbol{B}}_i \boldsymbol{L}^{-\mathrm{t}}. \tag{138}$$

Subsequent results are restricted to the case in which *p is an even integer*. This restriction enables us to considerably simplify the formula for $|\hat{\boldsymbol{B}}|_p$, viz.

$$|\hat{\boldsymbol{B}}|_p = \left(\sum_{i=1}^{d} |\hat{\boldsymbol{B}}_i|^p\right)^{1/p} = \left(\sum_{i=1}^{d} \hat{\boldsymbol{B}}_i^p\right)^{1/p}$$

$$= \left(\sum_{i=1}^{d} (\boldsymbol{L}^{-1}\tilde{\boldsymbol{B}}_i\boldsymbol{L}^{-t})^p\right)^{1/p} = \left(\boldsymbol{L}^{-1}\sum_{i=1}^{d} ((\tilde{\boldsymbol{B}}_i\boldsymbol{A}_0^{-1})^{p-1}\tilde{\boldsymbol{B}}_i)\boldsymbol{L}^{-t}\right)^{1/p}$$

$$= \left(\boldsymbol{L}^{-1}\left(\sum_{i=1}^{d} \boldsymbol{B}_i^p\right)\boldsymbol{A}_0\boldsymbol{L}^{-t}\right)^{1/p} = \left(\boldsymbol{L}^{-1}\left(\sum_{i=1}^{d} \boldsymbol{B}_i^p\right)\boldsymbol{L}\right)^{1/p}. \tag{139}$$

The equivalent forms of the eigenproblem may be expressed with the aid of (139):

$$\left(\sum_{i=1}^{d} \hat{\boldsymbol{B}}_i^p - \mu^p\boldsymbol{I}_m\right)\boldsymbol{v} = \boldsymbol{0}, \tag{140}$$

$$\left(\sum_{i=1}^{d} (\tilde{\boldsymbol{B}}_i\boldsymbol{A}_0^{-1})^{p-1}\tilde{\boldsymbol{B}}_i - \mu^p\boldsymbol{A}_0\right)\boldsymbol{\varphi} = \boldsymbol{0}, \tag{141}$$

$$\left(\sum_{i=1}^{d} \boldsymbol{B}_i^p - \mu^p\boldsymbol{I}_m\right)\boldsymbol{\psi} = \boldsymbol{0}. \tag{142}$$

As in the one-dimensional case, $\boldsymbol{\psi} = \boldsymbol{A}_0\boldsymbol{\varphi} = \boldsymbol{L}\boldsymbol{v}$ (cf. (27)–(29)). In the practically important case of $p = 2$, (140)–(142) specialize to, respectively:

$$(\hat{\boldsymbol{B}}^t\hat{\boldsymbol{B}} - \mu^2\boldsymbol{I}_m)\boldsymbol{v} = \boldsymbol{0}, \tag{143}$$

$$(\tilde{\boldsymbol{B}}^t\,\mathrm{diag}(\boldsymbol{A}_0^{-1}, \ldots, \boldsymbol{A}_0^{-1})\tilde{\boldsymbol{B}} - \mu^2\boldsymbol{A}_0)\boldsymbol{\varphi} = \boldsymbol{0}, \tag{144}$$

$$\left(\sum_{i=1}^{d} \boldsymbol{B}_i^2 - \mu^2\boldsymbol{I}_m\right)\boldsymbol{\psi} = \boldsymbol{0}, \tag{145}$$

where $\mathrm{diag}(\boldsymbol{A}_0^{-1}, \ldots, \boldsymbol{A}_0^{-1})$ is an $(m \cdot d) \times (m \cdot d)$ matrix with d copies of $\boldsymbol{A}_0^{-1}$ along the diagonal and zeros elsewhere. We wish to reiterate a remark made previously. That is, for the *hyperbolic case* the eigenvalue problem can be circumvented. For example, instead of (129) we have

$$\tilde{\boldsymbol{\tau}} = \boldsymbol{A}_0^{-1}\left(\sum_{i=1}^{d} \boldsymbol{B}_i^2\right)^{-1/2}, \tag{146}$$

which can be obtained directly following Marsden and Hughes [10, p. 55] and Hoger and Carlson [5].

Equation (128) *with* $\tilde{\boldsymbol{\tau}}$ *defined as above constitutes the generalized semidiscrete streamline method for advective-diffusive multidimensional systems written in symmetric form with Riemannian metric.*

4.3.1. Remarks

(1) For an isoparametric d-cube (or approximation thereof; see Fig. 1),

$$\partial \xi_i / \partial x_j = (2/h)\delta_{ij} . \tag{147}$$

In this case several simplifications ensue. (Recall that the $\hat{\boldsymbol{A}}_i$ are typically more sparse than the $\hat{\boldsymbol{B}}_i$.) Corresponding to (140)–(146), respectively, we have

$$\left(\sum_{i=1}^{d} \hat{\boldsymbol{A}}_i^p - |\lambda|^p \boldsymbol{I}_m\right)\boldsymbol{v} = \boldsymbol{0} , \tag{148}$$

$$\left(\sum_{i=1}^{d} (\tilde{\boldsymbol{A}}_i \boldsymbol{A}_0^{-1})^{p-1}\tilde{\boldsymbol{A}}_i - |\lambda|^p \boldsymbol{A}_0\right)\boldsymbol{\varphi} = \boldsymbol{0} , \tag{149}$$

$$\left(\sum_{i=1}^{d} \boldsymbol{A}_i^p - |\lambda|^p \boldsymbol{I}_m\right)\boldsymbol{\psi} = \boldsymbol{0} , \tag{150}$$

$$(\hat{\boldsymbol{A}}^{\mathrm{t}}\hat{\boldsymbol{A}} - |\lambda|^2 \boldsymbol{I}_m)\boldsymbol{v} = \boldsymbol{0} , \tag{151}$$

$$(\tilde{\boldsymbol{A}}^{\mathrm{t}} \operatorname{diag}(\boldsymbol{A}_0^{-1}, \ldots, \boldsymbol{A}_0^{-1})\tilde{\boldsymbol{A}} - |\lambda|^2 \boldsymbol{A}_0)\boldsymbol{\varphi} = \boldsymbol{0} , \tag{152}$$

$$\left(\sum_{i=1}^{d} \boldsymbol{A}_i^2 - |\lambda|^2 \boldsymbol{I}_m\right)\boldsymbol{\psi} = \boldsymbol{0} , \tag{153}$$

$$\tilde{\boldsymbol{\tau}} = \tfrac{1}{2} h \boldsymbol{A}_0^{-1}\left(\sum_{i=1}^{d} \boldsymbol{A}_i^2\right)^{-1/2} . \tag{154}$$

By virtue of the fact that:

$$\mu_i = (2/h)|\lambda_i| \tag{155}$$

and

$$\sigma_i = (2/h)^2 \kappa_i , \tag{156}$$

where

$$\kappa_i = \frac{1}{d}\begin{pmatrix} \boldsymbol{\varphi}_i \\ \vdots \\ \boldsymbol{\varphi}_i \end{pmatrix}^{\mathrm{t}} \tilde{\boldsymbol{K}} \begin{pmatrix} \boldsymbol{\varphi}_i \\ \vdots \\ \boldsymbol{\varphi}_i \end{pmatrix} = \frac{1}{d}\sum_{j,k=1}^{d} \boldsymbol{\varphi}_i^{\mathrm{t}} \tilde{\boldsymbol{K}}_{jk} \boldsymbol{\varphi}_i , \tag{157}$$

the element Peclet numbers can be written as:

$$\alpha_i = \frac{\mu_i}{\sigma_i} = \frac{(2/h)|\lambda_i|}{((2/h)^2 \kappa_i)} = \frac{h|\lambda_i|}{2\kappa_i} . \tag{158}$$

(2) The present definitions of $\tilde{\boldsymbol{\tau}}$ are consistent with the advection-dominated limits presented previously. In addition, design conditions (i)–(iii) now hold for the case of a diffusion matrix which is no longer necessarily small.

(3) In applications in which global conservation of the advective flux is important, the first term in (128) should be written in 'conservation form', namely:

$$\int_\Omega \boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{V}^h \, \mathrm{d}\Omega = -\int_\Omega \nabla \boldsymbol{W}^h \cdot \tilde{\boldsymbol{F}}(\boldsymbol{V}^h) \, \mathrm{d}\Omega + \int_\Gamma \boldsymbol{W}^h \cdot \tilde{\boldsymbol{F}}_n(\boldsymbol{V}^h) \, \mathrm{d}\Gamma , \tag{159}$$

where

$$\tilde{\boldsymbol{F}}^{\mathrm{t}} = (\tilde{\boldsymbol{F}}_1(\boldsymbol{V}^h)^{\mathrm{t}}, \tilde{\boldsymbol{F}}_2(\boldsymbol{V}^h)^{\mathrm{t}}, \ldots, \tilde{\boldsymbol{F}}_d(\boldsymbol{V}^h)^{\mathrm{t}}) , \tag{160}$$

$$\tilde{\boldsymbol{F}}_i(\boldsymbol{V}^h) = \tilde{\boldsymbol{A}}_i \boldsymbol{V}^h , \tag{161}$$

$$\tilde{\boldsymbol{F}}_n(\boldsymbol{V}^h) = \tilde{\boldsymbol{F}}_i(\boldsymbol{V}^h) n_i , \tag{162}$$

and Γ is the boundary of Ω and n_i is the unit outward normal vector to Γ. The left- and right-hand sides of (159) are, of course, analytically identical. However, *use of the right-hand side preserves conservation of flux under approximate element numerical integration*. This turns out to be crucial in certain situations.

(4) The definition of the σ_i enables us to establish the following stability estimate (see Appendix A for a proof):

$$\int_{\Omega^e} \nabla \boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}} \tilde{\boldsymbol{\tau}} (\nabla \cdot \tilde{\boldsymbol{K}} \nabla \boldsymbol{W}^h) \, \mathrm{d}\Omega \leqslant \tfrac{1}{2} \int_{\Omega^e} \nabla \boldsymbol{W}^h \cdot (\tilde{\boldsymbol{A}} \tilde{\boldsymbol{\tau}} \tilde{\boldsymbol{A}}^{\mathrm{t}} + \tilde{\boldsymbol{K}}) \nabla \boldsymbol{W}^h \, \mathrm{d}\Omega , \tag{163}$$

for all weighting functions $\boldsymbol{W}^h$. This is a key result which, when combined with the techniques of Johnson et al. [9], enables us to prove convergence of the methods developed herein for *steady, constant-coefficient* cases. In particular, the following error estimate may be shown to hold: let $\boldsymbol{E} = \boldsymbol{V}^h - \boldsymbol{V}$ denote the error in the finite element solution. Then $\boldsymbol{E}$ satisfies

$$\int_{\Omega^e} \nabla \boldsymbol{E} \cdot (\tilde{\boldsymbol{A}} \tilde{\boldsymbol{\tau}} \tilde{\boldsymbol{A}}^{\mathrm{t}} + \tilde{\boldsymbol{K}}) \nabla \boldsymbol{E} \, \mathrm{d}\Omega \leqslant C(\boldsymbol{V}) h^{2k} , \tag{164}$$

where k is the order of the complete polynomial present in the element interpolation function and c, $C(\boldsymbol{V})$ are constants which are independent of h. In the *hyperbolic case*, we obtain

$$\int_{\Omega^e} (c\boldsymbol{E} \cdot \boldsymbol{A}_0 \boldsymbol{E} + \nabla \boldsymbol{E} \cdot \tilde{\boldsymbol{A}} \tilde{\boldsymbol{\tau}} \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{E}) \, \mathrm{d}\Omega \leqslant C(\boldsymbol{V}) h^{2k+1} . \tag{165}$$

(Recall $\tilde{\boldsymbol{\tau}} = \mathrm{O}(h)$ in the hyperbolic limit.) For the *parabolic case*, in the diffusion-dominated limit, the method becomes the Galerkin method and the usual optimal error estimates are attained.

Localization results and error estimates of the form (164), (165) for the fully discrete version of the theory (not presented herein) may also be established (see [9] for further details).

5. Conclusions

In this paper we have developed a generalization of the SUPG concept for multidimensional advective-diffusive systems. Error estimates for the steady, linear case are presented which are valid over the entire range of advective-diffusive phenomena. As far as we are aware, this is the first methodology for this general class of problems for which precise mathematical convergence results are available.

The SUPG method represents an excellent methodology for 'smooth' solutions of advective-

diffusive systems. When sharp internal and/or boundary layers are present, additional robustness may be gained by introducing a discontinuity-capturing operator. See [7] for discussion of the scalar case. The generalization of this idea to multidimensional advective-diffusive systems will be described in the next paper in this series.

Appendix A

In this appendix we establish the key stability estimate needed to prove convergence of the method proposed herein.

A.1. Inverse estimate

We need to use an inverse estimate which takes on the following form:

$$\int_{\Omega^e} \boldsymbol{\varphi}_i^{\mathrm{t}}(\tilde{\boldsymbol{K}}_{jk}\boldsymbol{W}^h_{,k})_{,j}\boldsymbol{\varphi}_i^{\mathrm{t}}(\tilde{\boldsymbol{K}}_{pq}\boldsymbol{W}^h_{,q})_{,p}\,\mathrm{d}\Omega \leq C^2h^{-2}\int_{\Omega^e}(\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}_{jk}\boldsymbol{W}^h_{,k})(\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}_{jq}\boldsymbol{W}^h_{,q})\,\mathrm{d}\Omega\,, \tag{A.1}$$

where C is a constant independent of h.

We will need to assume:

$$\tilde{\xi}(\alpha)/\alpha \leq 4/(C^2ld)\,, \tag{A.2}$$

where C is the constant in (A.1) and l and d are positive integers. This assumption requires ξ to vanish sufficiently fast as $\alpha \to 0$ (cf. (50)).

We shall simplify the setting by assuming an isoparametric d-cube for which:

$$\tau_i = \tfrac{1}{2}h\tilde{\xi}(\alpha_i)/|\lambda_i|\,, \qquad \alpha_i = \tfrac{1}{2}h|\lambda_i|/\kappa_i\,. \tag{A.3}$$

This assumption simplifies the proof, but is actually unnecessary. The stability estimate holds in the general case.

THEOREM A.1. Assuming (A.1), (A.2), *and* (A.3),

$$\int_{\Omega^e} \nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\tilde{\boldsymbol{\tau}}(\nabla\cdot\tilde{\boldsymbol{K}}\nabla\boldsymbol{W}^h)\,\mathrm{d}\Omega \leq \tfrac{1}{2}\int_{\Omega} \nabla\boldsymbol{W}^h\cdot(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{\tau}}\tilde{\boldsymbol{A}}^{\mathrm{t}} + \tilde{\boldsymbol{K}})\nabla\boldsymbol{W}^h\,\mathrm{d}\Omega\,. \tag{A.4}$$

We need to establish a preliminary result. Let

$$\boldsymbol{v}_i^{\mathrm{t}} = (\boldsymbol{\varphi}_i^{\mathrm{t}}, \ldots, \boldsymbol{\varphi}_i^{\mathrm{t}}) \quad (1 \times m\cdot d)\,. \tag{A.5}$$

Recall

$$\kappa_i = \frac{1}{d}\,\boldsymbol{v}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{v}_i\,, \tag{A.6}$$

where $\tilde{\boldsymbol{K}}$ is symmetric and positive-semidefinite.

LEMMA A.2.

$$\tilde{\boldsymbol{K}}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{t}}\tilde{\boldsymbol{K}} \leqslant d\kappa_i\tilde{\boldsymbol{K}} . \tag{A.7}$$

PROOF. Let us consider an $(m \cdot d)$-dimensional vector $\boldsymbol{u}$. The result is established by the following calculation:

$$\boldsymbol{u}^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{v}_i\boldsymbol{v}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{u} = (\boldsymbol{u}^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{v}_i)^2 \leqslant \boldsymbol{u}^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{u}\boldsymbol{v}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{v}_i = d\kappa_i\boldsymbol{u}^{\mathrm{t}}\tilde{\boldsymbol{K}}\boldsymbol{u} . \tag{A.8}$$

PROOF OF THEOREM A.1. Let

$$a_i = Ch^{-1}(\kappa_i ld)^{1/2}\tau_i\nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\boldsymbol{\varphi}_i , \tag{A.9}$$

$$b_i = \frac{1}{Ch^{-1}(\kappa_i ld)^{1/2}}\,\boldsymbol{\varphi}_i^{\mathrm{t}}(\tilde{\boldsymbol{K}}_{jk}\boldsymbol{W}^h_{,k})_{,j} , \tag{A.10}$$

where C is the constant in the inverse estimate. The following calculation completes the proof of the stability estimate:

$$\begin{aligned}
\int_{\Omega^e}\nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\tilde{\boldsymbol{\tau}}(\nabla\cdot\tilde{\boldsymbol{K}}\nabla\boldsymbol{W}^h)\,\mathrm{d}\Omega &= \sum_{i=1}^{l}\int_{\Omega^e}\nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\boldsymbol{\varphi}_i\tau_i\boldsymbol{\varphi}_i^{\mathrm{t}}(\tilde{\boldsymbol{K}}_{jk}\boldsymbol{W}^h_{,k})_{,j}\,\mathrm{d}\Omega \\
&= \sum_{i=1}^{l}\int_{\Omega^e} a_i b_i\,\mathrm{d}\Omega \\
&\leqslant \sum_{i=1}^{l}\int_{\Omega^e}\tfrac{1}{2}(a_i^2 + b_i^2)\,\mathrm{d}\Omega \\
&= \sum_{i=1}^{l}\Big(\tfrac{1}{2}\int_{\Omega^e}(C^2h^{-2}\kappa_i ld\tau_i\nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\boldsymbol{\varphi}_i\tau_i\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{A}}^{\mathrm{t}}\nabla\boldsymbol{W}^h \\
&\qquad + C^{-2}h^2(\kappa_i ld)^{-1}\boldsymbol{\varphi}_i^{\mathrm{t}}(\tilde{\boldsymbol{K}}_{jk}\boldsymbol{W}^h_{,k})_{,j}\boldsymbol{\varphi}_i^{\mathrm{t}}(\tilde{\boldsymbol{K}}_{pq}\boldsymbol{W}^h_{,q})_{,p})\,\mathrm{d}\Omega\Big) \\
&\leqslant \sum_{i=1}^{l}\Big(\tfrac{1}{2}\int_{\Omega^e}(\tfrac{1}{4}C^2ld(\tilde{\xi}(\alpha_i)/\alpha_i)\nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\boldsymbol{\varphi}_i\tau_i\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{A}}^{\mathrm{t}}\nabla\boldsymbol{W}^h \\
&\qquad + (\kappa_i ld)^{-1}\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}_{jk}\boldsymbol{W}^h_{,k}\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{K}}_{jq}\boldsymbol{W}^h_{,q})\,\mathrm{d}\Omega\Big) \\
&\leqslant \sum_{i=1}^{l}\Big(\tfrac{1}{2}\int_{\Omega^e}(\nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\boldsymbol{\varphi}_i\tau_i\boldsymbol{\varphi}_i^{\mathrm{t}}\tilde{\boldsymbol{A}}^{\mathrm{t}}\nabla\boldsymbol{W}^h \\
&\qquad + (\kappa_i ld)^{-1}\nabla\boldsymbol{W}^h \cdot \tilde{\boldsymbol{K}}\boldsymbol{v}_i\boldsymbol{v}_i\tilde{\boldsymbol{K}}\nabla\boldsymbol{W}^h)\,\mathrm{d}\Omega\Big) \\
&\leqslant \tfrac{1}{2}\int_{\Omega^e}\nabla\boldsymbol{W}^h \cdot (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{\tau}}\tilde{\boldsymbol{A}}^{\mathrm{t}} + \tilde{\boldsymbol{K}})\nabla\boldsymbol{W}^h\,\mathrm{d}\Omega .
\end{aligned} \tag{A.11}$$

Acknowledgment

We wish to thank Leo Franca, Claes Johnson, Akira Mizukami, and Farzin Shakib for helpful comments. Michel Mallet was supported by an IBM Graduate Fellowship.

References

[1] A.N. Brooks and T.J.R. Hughes, Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations, Comput. Meths. Appl. Mech. Engrg. 32 (1982) 199–259.

[2] J. Donéa, A Taylor–Galerkin method for convective transport problems, Internat. J. Numer. Meths. Engrg. 20 (1) (1984) 101–119.

[3] P.K. Dutt, Stable boundary conditions and difference schemes for Navier–Stokes type equations, Ph.D. Thesis, University of California, Los Angeles, CA, 1985.

[4] A. Harten, On the symmetric form of systems of conservation laws with entropy, J. Comput. Phys. 49 (1983) 151–164.

[5] A. Hoger and D.E. Carlson, Determination of the stretch and rotation in the polar decomposition of the deformation gradient, Quart. Appl. Math. 42 (1) (1984) 113–117.

[6] T.J.R. Hughes L.P. Franca and M. Mallet, A new finite element formulation for computational fluid dynamics: I. Symmetric forms of the compressible Euler and Navier–Stokes equations and the second law of thermodynamics, Comput. Meths. Appl. Mech. Engrg. 54 (1986) 223–234.

[7] T.J.R. Hughes, M. Mallet and A. Mizukami, A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, Comput. Meths. Appl. Mech. Engrg. 54 (1986) 341–355.

[8] T.J.R. Hughes and T.E. Tezduyar, Finite element methods for first-order hyperbolic systems with particular emphasis on the compressible Euler equations, Comput. Meths. Appl. Mech. Engrg. 45 (1984) 217–284.

[9] C. Johnson, U. Nävert and J. Pitkäranta, Finite element methods for linear hyperbolic problems, Comput. Meths. Appl. Mech. Engrg. 45 (1984) 285–312.

[10] J.E. Marsden and T.J.R. Hughes, Mathematical Foundations of Elasticity (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[11] E. Tadmor, Skew-selfadjoint forms for systems of conservation laws, J. Math. Anal. Appl. 103 (1984) 428–442.