COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING 63 (1987) 97–112
NORTH-HOLLAND

A NEW FINITE ELEMENT FORMULATION FOR COMPUTATIONAL FLUID DYNAMICS: VI. CONVERGENCE ANALYSIS OF THE GENERALIZED SUPG FORMULATION FOR LINEAR TIME-DEPENDENT MULTIDIMENSIONAL ADVECTIVE-DIFFUSIVE SYSTEMS*

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

Received 12 December 1986

An SUPG-type finite element method for linear symmetric multidimensional advective-diffusive systems is described and analyzed. Optimal and near optimal error estimates are obtained for the complete range of advective-diffusive behavior.

1. Introduction

We consider linear multidimensional advective-diffusive *systems*. Based on the work of Johnson and his colleagues [4, 5, 7], we extend the semidiscrete SUPG formulation of Hughes and Mallet [3] (hereafter simply referred to as Part III) to a fully discrete space-time formulation using a discontinuous Galerkin method in time. A crucial ingredient in our formulation is the *matrix of intrinsic time scales*, defined in Part III, which enables us to obtain uniform error estimates for systems, analogous to the scalar case [5], encompassing the full spectrum of advective and diffusive behavior. For systems, the possibility of advection- and diffusion-dominated 'modes' occuring simultaneously engenders analytical difficulties not observed in the scalar case, for which the limits of pure advection and pure diffusion provide an essentially complete picture. Overcoming these difficulties is a main motivation behind the development of the procedures described herein.

Linear symmetric advective-diffusive systems are of interest in their own right and serve as model equations for many important nonlinear systems encountered in science and engineering. For example, when written in terms of entropy variables (see, e.g., [2] and references therein) the compressible Euler and Navier–Stokes equations fall within this class, the former being a symmetric hyperbolic system and the latter a symmetric incompletely parabolic

* This research was sponsored by the NASA Langley Research Center under Grant NASA-NAG-1-361 and the IBM Almaden Research Center under Grant No. 604912.
** Presently at Avions Marcel Dassault-Breguet Aviation, St. Cloud.

system. Thus it seems self-evident that any method proposed for these complex nonlinear equations (and there have been many) should, at the very least, be able to be mathematically verified in the context of simpler linear advective-diffusive system models. It is indicative of the primitive state of affairs in this area of numerical analysis that, to the best of our knowledge, no methods have attained this modest objective besides the one described herein.

An outline of the paper follows. In Section 2 we begin by studying the hyperbolic limit. This case has been previously analyzed in [5] and nothing fundamentally new is described here. Our purpose is to establish for subsequent use some results and the method of proof in as simple a setting as possible. In Section 3 we consider the parabolic case. We first analyze the situation in which diffusion is uniformly small in the sense that all 'modes' are advection dominated. The error estimates we obtain are essentially the same as for the hyperbolic limit. Finally, we consider the general situation in which advection and diffusion may dominate different 'modes' simultaneously. We obtain error estimates which are optimal under these circumstances. The form of the matrix of intrinsic time scales plays a key role in the analysis. In section 4 we draw conclusions.

We do not specifically consider the incompletely parabolic case herein. We believe that, for appropriately specified boundary conditions, the present methodology will apply to this case as well.

For recent progress in the analysis of the *nonlinear* hyperbolic case see [6].

2. Symmetric hyperbolic systems

2.1. Preliminaries

Let Ω be an open bounded region in $\mathbb{R}^d$ where d is the number of space dimensions. The boundary of Ω is denoted by Γ and is assumed to be smooth. The unit outward normal vector to Γ is denoted by $\boldsymbol{n} = (n_1, n_2, \ldots, n_d)$. We denote by I the open interval $]0, T[$. Consider a partition of $]0, T[$ of the form $0 = t_0 < t_1 < \cdots < t_N = T$. We denote by $I_n \subset I$ the open interval $]t_n, t_{n+1}[$. Clearly $I = \cup_{n=0}^{N-1} I_n \cup \{t_1, t_2, \ldots, t_{N-1}\}$. We will also use the following notation:

$$Q = \Omega \times I\,, \qquad Q_n = \Omega \times I_n\,, \qquad P = \Gamma \times I\,, \qquad P_n = \Gamma \times I_n\,. \tag{1}$$

2.2. Problem statement

The problem consists of finding $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{x}, t)$ $\forall \boldsymbol{x} \in \Omega$, $\forall t \in I$, such that:

$$\mathscr{L}\boldsymbol{V} = \boldsymbol{A}_0 \boldsymbol{V}_{,t} + \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{V} + \boldsymbol{C}\boldsymbol{V} = \boldsymbol{0} \quad \text{in } Q\,, \tag{2}$$

$$(\boldsymbol{M} - \boldsymbol{D})\boldsymbol{V} = \boldsymbol{0} \quad \text{on } P\,, \tag{3}$$

$$\boldsymbol{V}(\boldsymbol{x}, 0) = \boldsymbol{G}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \Omega\,, \tag{4}$$

where $\boldsymbol{A}_0$, the $\tilde{\boldsymbol{A}}_i$s (the matrices composing $\tilde{\boldsymbol{A}}^{\mathrm{t}} = [\tilde{\boldsymbol{A}}_1, \tilde{\boldsymbol{A}}_2, \ldots, \tilde{\boldsymbol{A}}_d]$[1]), and $\boldsymbol{C}$ are given

[1] More details on the compact notation used here may be found in Part III.

symmetric constant $m \times m$ matrices, and $\boldsymbol{D} = \sum_{i=1}^{d} n_i \tilde{\boldsymbol{A}}_i$. We assume that $\boldsymbol{A}_0$ is positive-definite, $\boldsymbol{C}$ and $\boldsymbol{M}$ are positive semi-definite, and

$$\operatorname{Ker}(\boldsymbol{D} - \boldsymbol{M}) + \operatorname{Ker}(\boldsymbol{D} + \boldsymbol{M}) = \mathbb{R}^m \quad \text{on } P . \tag{5}$$

An appropriate choice for $\boldsymbol{M}$ is the absolute value of $\boldsymbol{D}$. This amounts to specifying boundary conditions on incoming characteristics only.

Remark

It follows from (5) that a vector $\boldsymbol{V}$ may be decomposed as

$$\boldsymbol{V} = \boldsymbol{V}_1 + \boldsymbol{V}_2 , \tag{6}$$

where $\boldsymbol{V}_1 \in \operatorname{Ker}(\boldsymbol{D} - \boldsymbol{M})$, $\boldsymbol{V}_2 \in \operatorname{Ker}(\boldsymbol{D} + \boldsymbol{M})$. By definition,

$$\boldsymbol{D}\boldsymbol{V}_1 = \boldsymbol{M}\boldsymbol{V}_1 , \qquad \boldsymbol{D}\boldsymbol{V}_2 = -\boldsymbol{M}\boldsymbol{V}_2 , \tag{7}$$

and thus,

$$\begin{aligned} \boldsymbol{V}_1 \cdot \boldsymbol{M}\boldsymbol{V}_2 &= \tfrac{1}{2}\boldsymbol{V}_1 \cdot \boldsymbol{M}\boldsymbol{V}_2 + \tfrac{1}{2}\boldsymbol{V}_1 \cdot \boldsymbol{M}\boldsymbol{V}_2 \\ &= \tfrac{1}{2}\boldsymbol{V}_1 \cdot \boldsymbol{D}\boldsymbol{V}_2 + \tfrac{1}{2}\boldsymbol{V}_1 \cdot \boldsymbol{M}\boldsymbol{V}_2 \\ &= \tfrac{1}{2}\boldsymbol{V}_1 \cdot (\boldsymbol{D} + \boldsymbol{M})\boldsymbol{V}_2 \\ &= 0 , \end{aligned} \tag{8}$$

from which it follows that

$$\boldsymbol{V}_1 \cdot \boldsymbol{M}\boldsymbol{V}_1 \leq \boldsymbol{V} \cdot \boldsymbol{M}\boldsymbol{V} , \qquad \boldsymbol{V}_2 \cdot \boldsymbol{M}\boldsymbol{V}_2 \leq \boldsymbol{V} \cdot \boldsymbol{M}\boldsymbol{V} . \tag{9}$$

These results are useful in the error analysis.

2.3. Finite element formulation

The generalized SUPG formulation for the time-dependent symmetric hyperbolic system is:

Find $\boldsymbol{V}^h \in \mathscr{V}_n^h$, such that $\forall \boldsymbol{W}^h \in \mathscr{V}_n^h$

$$B_n(\boldsymbol{W}^h, \boldsymbol{V}^h) = L_n(\boldsymbol{W}^h) , \quad n = 0, 1, \ldots, N-1 , \tag{10}$$

where

$$\begin{aligned} B_n(\boldsymbol{W}^h, \boldsymbol{V}^h) = {} & \int_{Q_n} \boldsymbol{W}^h \cdot \mathscr{L}\boldsymbol{V}^h \, \mathrm{d}Q + \int_{\Omega} \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0 \boldsymbol{V}^h(t_n^+) \, \mathrm{d}\Omega \\ & + \tfrac{1}{2} \int_{P_n} \boldsymbol{W}^h \cdot (\boldsymbol{M} - \boldsymbol{D}) \boldsymbol{V}^h \, \mathrm{d}P \\ & + \int_{Q_n} \mathscr{L}\boldsymbol{W}^h \cdot \boldsymbol{\tau} \mathscr{L}\boldsymbol{V}^h \, \mathrm{d}Q , \quad n = 0, 1, \ldots, N-1 , \end{aligned} \tag{11}$$

$$L_n(\boldsymbol{W}^h) = \int_\Omega \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0 \boldsymbol{V}^h(t_n^-)\,\mathrm{d}\Omega\,, \quad n = 1, \ldots, N-1\,, \tag{12}$$

$$L_0(\boldsymbol{W}^h) = \int_\Omega \boldsymbol{W}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{G}\,\mathrm{d}\Omega\,, \tag{13}$$

in which $\boldsymbol{V}^h$ is the trial solution, $\boldsymbol{W}^h$ is the weighting function, $\mathcal{V}_n^h \subset (H^1(Q_n))^m$ is the space of trial and weighting functions consisting of typical C^0 finite element functions, and τ is the *matrix of intrinsic time scales*, a stabilizing symmetric positive-definite matrix. See Part III for elaboration on $\boldsymbol{\tau}$. In order to avoid confusion the explicit dependence on t of some variables is displayed meaning

$$\boldsymbol{U}(t_n^+) = \lim_{s\to 0^+} \boldsymbol{U}(\boldsymbol{x}, t_n + s)\,, \tag{14}$$

$$\boldsymbol{U}(t_n^-) = \lim_{s\to 0^-} \boldsymbol{U}(\boldsymbol{x}, t_n + s)\,. \tag{15}$$

Remarks

(1) The method is applied in one time slab at a time. The solution obtained in the previous time slab is used as input to the next time slab as one can see from (12).

(2) The method is a fully discrete generalization of the semidiscrete formulation presented in Part III. It combines the discontinuous Galerkin method in time with the generalized SUPG method in space. The only differences between this method and the one discussed in [5] are the presence here of the Riemannian metric $\boldsymbol{A}_0$ and the matrix of intrinsic time scales $\boldsymbol{\tau}$. The scalar case is analyzed in detail in [7].

(3) In the present formulation the boundary conditions are treated *weakly* and thus are not required in the definition of the finite element space $\mathcal{V}_n^h$.

Equations (10)–(13) are the representation of the method from the implementational point of view. To facilitate the analysis of the method we sum over the N time slabs:

Find $\boldsymbol{V}^h \in \mathcal{V}^h$ such that $\forall \boldsymbol{W}^h \in \mathcal{V}^h$

$$B(\boldsymbol{W}^h, \boldsymbol{V}^h) = L(\boldsymbol{W}^h)\,, \tag{16}$$

where

$$\begin{aligned} B(\boldsymbol{W}^h, \boldsymbol{V}^h) &= \sum_{n=1}^{N-1} (B_n(\boldsymbol{W}^h, \boldsymbol{V}^h) - L_n(\boldsymbol{W}^h)) + B_0(\boldsymbol{W}^h, \boldsymbol{V}^h) \\ &= \sum_{n=0}^{N-1} \int_{Q_n} \boldsymbol{W}^h \cdot \boldsymbol{A}_0 \boldsymbol{V}^h_{,t}\,\mathrm{d}Q + \sum_{n=1}^{N-1} \int_\Omega \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0 [\![\boldsymbol{V}^h(t_n)]\!]\,\mathrm{d}\Omega \\ &\quad + \int_\Omega \boldsymbol{W}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{V}^h(0^+)\,\mathrm{d}\Omega + \int_Q \boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{V}^h\,\mathrm{d}Q + \int_Q \boldsymbol{W}^h \cdot \boldsymbol{C} \boldsymbol{V}^h\,\mathrm{d}Q \\ &\quad + \tfrac{1}{2} \int_P \boldsymbol{W}^h \cdot (\boldsymbol{M} - \boldsymbol{D}) \boldsymbol{V}^h\,\mathrm{d}P + \sum_{n=0}^{N-1} \int_{Q_n} \mathcal{L}\boldsymbol{W}^h \cdot \boldsymbol{\tau} \mathcal{L} \boldsymbol{V}^h\,\mathrm{d}Q\,, \end{aligned} \tag{17}$$

$$L(\boldsymbol{W}^h) = L_0(\boldsymbol{W}^h) = \int_\Omega \boldsymbol{W}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{G}\,\mathrm{d}\Omega\,, \tag{18}$$

in which $\mathscr{V}^h \subset (H^1(\cup_{n=0}^{N-1} Q_n))^m$ is the space of trial and weighting functions and

$$[\![\boldsymbol{V}^h(t_n)]\!] = \boldsymbol{V}^h(t_n^+) - \boldsymbol{V}^h(t_n^-) . \tag{19}$$

2.4. Error analysis

Let $\tilde{\boldsymbol{V}}^h \in \mathscr{V}^h$ denote an interpolant of $\boldsymbol{V}$. Let $\boldsymbol{\eta} = \tilde{\boldsymbol{V}}^h - \boldsymbol{V}$ denote the interpolation error. We need to make the following assumptions:

(A1) $\boldsymbol{\eta}$ is a t-continuous function in the sense that

$$\sum_{n=1}^{N-1} \int_\Omega [\![\boldsymbol{\eta}(t_n)]\!] \cdot \boldsymbol{A}_0 [\![\boldsymbol{\eta}(t_n)]\!] \, \mathrm{d}\Omega = 0 . \tag{20}$$

(H1) If $\boldsymbol{V} \in (H^{k+1}(Q))^m$, $\boldsymbol{\eta}$ satisfies:

$$\int_Q \boldsymbol{\eta} \cdot \boldsymbol{A}_0 \boldsymbol{\eta} \, \mathrm{d}Q + h^2 \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{\eta} \cdot \boldsymbol{A}_0^{-1} \mathscr{L}\boldsymbol{\eta} \, \mathrm{d}Q \leq C(\boldsymbol{V}) h^{2k+2} , \tag{21}$$

$$\sum_{n=1}^{N} \int_\Omega \boldsymbol{\eta}(t_n^-) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(t_n^-) \, \mathrm{d}\Omega + \int_\Omega \boldsymbol{\eta}(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(0^+) \, \mathrm{d}\Omega + \int_P \boldsymbol{\eta} \cdot \boldsymbol{M} \boldsymbol{\eta} \, \mathrm{d}P \leq C(\boldsymbol{V}) h^{2k+1} , \tag{22}$$

in which h is a space-time mesh parameter, and $C(\boldsymbol{V})$ denotes a function of the exact solution which is independent of h and which may take on different values in the inequalities above and subsequently.

(H2) The stabilizing matrix $\boldsymbol{\tau}$ is of order h (see Part III), meaning in particular that one can find constants c_1 and c_2 such that

$$\sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{W}^h \cdot \boldsymbol{\tau} \mathscr{L}\boldsymbol{W}^h \, \mathrm{d}Q \leq c_1 h \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{W}^h \cdot \boldsymbol{A}_0^{-1} \mathscr{L}\boldsymbol{W}^h \, \mathrm{d}Q , \tag{23}$$

$$\int_Q \boldsymbol{\eta} \cdot \boldsymbol{\tau}^{-1} \boldsymbol{\eta} \, \mathrm{d}Q \leq c_2 \frac{1}{h} \int_Q \boldsymbol{\eta} \cdot \boldsymbol{A}_0 \boldsymbol{\eta} \, \mathrm{d}Q . \tag{24}$$

We will also use the following result which follows from the properties of $\boldsymbol{C}$ and $\boldsymbol{A}_0$:

$$\int_Q \boldsymbol{\eta} \cdot \boldsymbol{C} \boldsymbol{\eta} \, \mathrm{d}Q \leq c_3 \int_Q \boldsymbol{\eta} \cdot \boldsymbol{A}_0 \boldsymbol{\eta} \, \mathrm{d}Q , \tag{25}$$

where c_3 is a constant.

Let us define the norm, $|||\cdot|||$, in which we prove convergence in the theorem below:

$$|||\boldsymbol{W}^h|||^2 = \tfrac{1}{2} \sum_{n=1}^{N-1} \int_\Omega [\![\boldsymbol{W}^h(t_n)]\!] \cdot \boldsymbol{A}_0 [\![\boldsymbol{W}^h(t_n)]\!] \, \mathrm{d}\Omega + \tfrac{1}{2} \int_P \boldsymbol{W}^h \cdot \boldsymbol{M} \boldsymbol{W}^h \, \mathrm{d}P$$
$$+ \tfrac{1}{2}\Big(\int_\Omega \boldsymbol{W}^h(T^-) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(T^-) \, \mathrm{d}\Omega + \int_\Omega \boldsymbol{W}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(0^+) \, \mathrm{d}\Omega \Big)$$
$$+ \int_Q \boldsymbol{W}^h \cdot \boldsymbol{C} \boldsymbol{W}^h \, \mathrm{d}Q + \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L} \boldsymbol{W}^h \cdot \boldsymbol{\tau} \mathscr{L} \boldsymbol{W}^h \, \mathrm{d}Q \,. \qquad (26)$$

We now establish some preliminary results.

LEMMA 2.1.

$$\sum_{n=0}^{N-1} \int_{Q_n} \boldsymbol{W}^h \cdot \boldsymbol{A}_0 \boldsymbol{W}^h_{,t} \, \mathrm{d}Q = -\tfrac{1}{2} \sum_{n=1}^{N-1} \int_\Omega [\![\boldsymbol{W}^h(t_n) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(t_n)]\!] \, \mathrm{d}\Omega$$
$$+ \tfrac{1}{2}\Big(\int_\Omega \boldsymbol{W}^h(T^-) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(T^-) \, \mathrm{d}\Omega$$
$$- \int_\Omega \boldsymbol{W}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(0^+) \, \mathrm{d}\Omega \Big) \,. \qquad (27)$$

PROOF.

$$\sum_{n=0}^{N-1} \int_{Q_n} \boldsymbol{W}^h \cdot \boldsymbol{A}_0 \boldsymbol{W}^h_{,t} \, \mathrm{d}Q = \sum_{n=0}^{N-1} \int_{Q_n} \Big(\frac{\boldsymbol{W}^h \cdot \boldsymbol{A}_0 \boldsymbol{W}^h}{2} \Big)_{,t} \, \mathrm{d}Q$$
$$= \tfrac{1}{2} \sum_{n=0}^{N-1} \int_\Omega (\boldsymbol{W}^h(t_{n+1}^-) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(t_{n+1}^-) - \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(t_n^+)) \, \mathrm{d}\Omega$$
$$= \tfrac{1}{2}\Big(\sum_{n=1}^{N} \int_\Omega \boldsymbol{W}^h(t_n^-) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(t_n^-) \, \mathrm{d}\Omega$$
$$- \sum_{n=0}^{N-1} \int_\Omega \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(t_n^+) \, \mathrm{d}\Omega \Big) \,. \qquad (28)$$

Extracting the last term of the first summation and the first term of the second summation leads to the desired result.

LEMMA 2.2 (Stability).

$$|||\boldsymbol{W}^h|||^2 = B(\boldsymbol{W}^h, \boldsymbol{W}^h) \quad \forall \boldsymbol{W}^h \in \mathscr{V}^h \,. \qquad (29)$$

PROOF.

From the definition of $B(\cdot\,,\cdot)$,

$$B(\boldsymbol{W}^h, \boldsymbol{W}^h) = \sum_{n=0}^{N-1} \int_{Q_n} \boldsymbol{W}^h \cdot \boldsymbol{A}_0 \boldsymbol{W}^h_{,t} \, \mathrm{d}Q + \sum_{n=1}^{N-1} \int_{\Omega} \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0 [\![\boldsymbol{W}^h(t_n)]\!] \, \mathrm{d}\Omega$$
$$+ \int_{\Omega} \boldsymbol{W}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{W}^h(0^+) \, \mathrm{d}\Omega + \int_Q \boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{W}^h \, \mathrm{d}Q$$
$$+ \tfrac{1}{2} \int_P \boldsymbol{W}^h \cdot (\boldsymbol{M} - \boldsymbol{D}) \boldsymbol{W}^h \, \mathrm{d}P + \int_Q \boldsymbol{W}^h \cdot \boldsymbol{C} \boldsymbol{W}^h \, \mathrm{d}Q$$
$$+ \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L} \boldsymbol{W}^h \cdot \boldsymbol{\tau} \mathscr{L} \boldsymbol{W}^h \, \mathrm{d}Q \,, \tag{30}$$

and noting from integration by parts in space that

$$\int_Q \boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{W}^h \, \mathrm{d}Q = \tfrac{1}{2} \int_P \boldsymbol{W}^h \cdot \boldsymbol{D} \boldsymbol{W}^h \, \mathrm{d}P \,, \tag{31}$$

the lemma follows by substituting (31) and Lemma 2.1 into (30), and comparing with the definition of $|||\cdot|||$ in (26).

LEMMA 2.3 (Orthogonality of the error).

$$B(\boldsymbol{W}^h, \boldsymbol{E}) = 0 \quad \forall \boldsymbol{W}^h \in \mathscr{V}^h \,, \tag{32}$$

in which $\boldsymbol{E} = \boldsymbol{V}^h - \boldsymbol{V}$.

PROOF. It is an immediate consequence of (16)–(18).

THEOREM 2.4.(Johnson, Nävert and Pitkäranta [5]) *Assuming* (A1), (H1), *and* (H2) hold, then

$$|||\boldsymbol{E}|||^2 \leq C(\boldsymbol{V}) h^{2k+1} \,. \tag{33}$$

PROOF. Let us define $\boldsymbol{E}^h = \boldsymbol{V}^h - \tilde{\boldsymbol{V}}^h$. Then

$$\boldsymbol{E} = \boldsymbol{E}^h + \boldsymbol{\eta} \,. \tag{34}$$

We will use the following results in the sequel:

$$\sum_{n=0}^{N-1} \int_{Q_n} \boldsymbol{E}^h \cdot \boldsymbol{A}_0 \boldsymbol{\eta}_{,t} \, \mathrm{d}Q = - \sum_{n=0}^{N-1} \int_{Q_n} (\boldsymbol{A}_0 \boldsymbol{E}^h_{,t}) \cdot \boldsymbol{\eta} \, \mathrm{d}Q - \sum_{n=1}^{N-1} \int_{\Omega} [\![\boldsymbol{E}^h(t_n) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(t_n)]\!] \, \mathrm{d}\Omega$$
$$+ \int_{\Omega} (\boldsymbol{E}^h(T^-) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(T^-) - \boldsymbol{E}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(0^+)) \, \mathrm{d}\Omega \,, \tag{35}$$

$$\int_Q \boldsymbol{E}^h \cdot \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{\eta} \, \mathrm{d}Q = - \int_Q (\tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{E}^h) \cdot \boldsymbol{\eta} \, \mathrm{d}Q + \int_P \boldsymbol{E}^h \cdot \boldsymbol{D} \boldsymbol{\eta} \, \mathrm{d}P \,, \tag{36}$$

$$\int_P \boldsymbol{E}^h \cdot \boldsymbol{D} \boldsymbol{\eta} \, \mathrm{d}P = \int_P \boldsymbol{E}^h \cdot \boldsymbol{M} \boldsymbol{\eta}_1 \, \mathrm{d}P - \int_P \boldsymbol{E}^h \cdot \boldsymbol{M} \boldsymbol{\eta}_2 \, \mathrm{d}P \,. \tag{37}$$

Result (35) follows from integration by parts and reordering as in Lemma 2.1, (36) follows from integration by parts, and (37) follows by decomposing $\boldsymbol{\eta}$ as in (6).

Proceeding,

$$
\begin{aligned}
|||\boldsymbol{E}^h|||^2 &= B(\boldsymbol{E}^h, \boldsymbol{E}^h) && \text{(by Lemma 2.2)}\\
&= B(\boldsymbol{E}^h, \boldsymbol{E} - \boldsymbol{\eta}) && \text{(by (34))}\\
&= -B(\boldsymbol{E}^h, \boldsymbol{\eta}) && \text{(by Lemma 2.3)}\\
&\leqslant |B(\boldsymbol{E}^h, \boldsymbol{\eta})|\\
&= \Big| \sum_{n=0}^{N-1} \int_{Q_n} \boldsymbol{E}^h \cdot \boldsymbol{A}_0 \boldsymbol{\eta}_{,t} \, \mathrm{d}Q + \sum_{n=1}^{N-1} \int_\Omega \boldsymbol{E}^h(t_n^+) \cdot \boldsymbol{A}_0 [\![\boldsymbol{\eta}(t_n)]\!] \, \mathrm{d}\Omega\\
&\quad + \int_\Omega \boldsymbol{E}^h(0^+) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(0^+) \, \mathrm{d}\Omega + \int_Q \boldsymbol{E}^h \cdot \tilde{\boldsymbol{A}} \cdot \nabla \boldsymbol{\eta} \, \mathrm{d}Q + \tfrac{1}{2} \int_P \boldsymbol{E}^h \cdot (\boldsymbol{M} - \boldsymbol{D}) \boldsymbol{\eta} \, \mathrm{d}P\\
&\quad + \int_Q \boldsymbol{E}^h \cdot \boldsymbol{C} \boldsymbol{\eta} \, \mathrm{d}Q + \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{E}^h \cdot \boldsymbol{\tau} \mathscr{L} \boldsymbol{\eta} \, \mathrm{d}Q \Big| && \text{(by definition of } B)\\
&= \Big| - \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{E}^h \cdot \boldsymbol{\eta} \, \mathrm{d}Q - \sum_{n=1}^{N-1} \int_\Omega [\![\boldsymbol{E}^h(t_n)]\!] \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(t_n^-) \, \mathrm{d}\Omega\\
&\quad + \int_\Omega \boldsymbol{E}^h(T^-) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(T^-) \, \mathrm{d}\Omega + \tfrac{1}{2} \int_P \boldsymbol{E}^h \cdot \boldsymbol{M}(2\boldsymbol{\eta}_2) \, \mathrm{d}P + 2 \int_Q (\boldsymbol{C}\boldsymbol{E}^h) \cdot \boldsymbol{\eta} \, \mathrm{d}Q\\
&\quad + \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{E}^h \cdot \boldsymbol{\tau} \mathscr{L} \boldsymbol{\eta} \, \mathrm{d}Q \Big| && \text{(by (35)–(37))}\\
&= \Big| - \sum_{n=0}^{N-1} \int_{Q_n} \Big(\frac{1}{\sqrt{2}} \boldsymbol{\tau}^{1/2} \mathscr{L}\boldsymbol{E}^h \Big) \cdot \sqrt{2} \boldsymbol{\tau}^{-1/2} \boldsymbol{\eta} \, \mathrm{d}Q\\
&\quad - \sum_{n=1}^{N-1} \int_\Omega \Big(\frac{1}{\sqrt{2}} \boldsymbol{A}_0^{1/2} [\![\boldsymbol{E}^h(t_n)]\!] \Big) \cdot \sqrt{2} \boldsymbol{A}_0^{1/2} \boldsymbol{\eta}(t_n^-) \, \mathrm{d}\Omega\\
&\quad + \int_\Omega \Big(\frac{1}{\sqrt{2}} \boldsymbol{A}_0^{1/2} \boldsymbol{E}^h(T^-) \Big) \cdot \sqrt{2} \boldsymbol{A}_0^{1/2} \boldsymbol{\eta}(T^-) \, \mathrm{d}\Omega\\
&\quad + \tfrac{1}{2} \int_P (\boldsymbol{M}^{1/2} \boldsymbol{E}^h) \cdot \boldsymbol{M}^{1/2} (2\boldsymbol{\eta}_2) \, \mathrm{d}P + 2 \int_Q \Big(\frac{1}{\sqrt{2}} \boldsymbol{C}^{1/2} \boldsymbol{E}^h \Big) \cdot \sqrt{2} \boldsymbol{C}^{1/2} \boldsymbol{\eta} \, \mathrm{d}Q\\
&\quad + \sum_{n=0}^{N-1} \int_{Q_n} \Big(\frac{1}{\sqrt{2}} \boldsymbol{\tau}^{1/2} \mathscr{L}\boldsymbol{E}^h \Big) \cdot \sqrt{2} \boldsymbol{\tau}^{1/2} \mathscr{L} \boldsymbol{\eta} \, \mathrm{d}Q \Big|\\
&\leqslant \tfrac{1}{2} |||\boldsymbol{E}^h|||^2 + \int_Q \boldsymbol{\eta} \cdot \boldsymbol{\tau}^{-1} \boldsymbol{\eta} \, \mathrm{d}Q + \sum_{n=1}^{N} \int_\Omega \boldsymbol{\eta}(t_n^-) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(t_n^-) \, \mathrm{d}\Omega + \int_P \boldsymbol{\eta}_2 \cdot \boldsymbol{M} \boldsymbol{\eta}_2 \, \mathrm{d}P\\
&\quad + 2 \int_Q \boldsymbol{\eta} \cdot \boldsymbol{C} \boldsymbol{\eta} \, \mathrm{d}Q + \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{\eta} \cdot \boldsymbol{\tau} \mathscr{L} \boldsymbol{\eta} \, \mathrm{d}Q \, .
\end{aligned}
$$

Thus, combining with the left-hand side:

$$
\begin{aligned}
|||\boldsymbol{E}^h|||^2 &\leq 2\int_Q \boldsymbol{\eta}\cdot\boldsymbol{\tau}^{-1}\boldsymbol{\eta}\,\mathrm{d}Q + 2\sum_{n=1}^{N}\int_\Omega \boldsymbol{\eta}(t_n^-)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(t_n^-)\,\mathrm{d}\Omega \\
&\quad + 2\int_P \boldsymbol{\eta}_2\cdot\boldsymbol{M}\boldsymbol{\eta}_2\,\mathrm{d}P + 4\int_Q \boldsymbol{\eta}\cdot\boldsymbol{C}\boldsymbol{\eta}\,\mathrm{d}Q + 2\sum_{n=0}^{N-1}\int_{Q_n} \mathscr{L}\boldsymbol{\eta}\cdot\boldsymbol{\tau}\mathscr{L}\boldsymbol{\eta}\,\mathrm{d}Q \\
&\leq 2\frac{c_2}{h}\int_Q \boldsymbol{\eta}\cdot\boldsymbol{A}_0\boldsymbol{\eta}\,\mathrm{d}Q + 2\sum_{n=1}^{N}\int_\Omega \boldsymbol{\eta}(t_n^-)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(t_n^-)\,\mathrm{d}\Omega + 2\int_P \boldsymbol{\eta}_2\cdot\boldsymbol{M}\boldsymbol{\eta}_2\,\mathrm{d}P \\
&\quad + 4c_3\int_Q \boldsymbol{\eta}\cdot\boldsymbol{A}_0\boldsymbol{\eta}\,\mathrm{d}Q + 2c_1 h\sum_{n=0}^{N-1}\int_{Q_n} \mathscr{L}\boldsymbol{\eta}\cdot\boldsymbol{A}_0^{-1}\mathscr{L}\boldsymbol{\eta}\,\mathrm{d}Q \qquad \text{(by (H2))} \\
&\leq C(\boldsymbol{V})h^{2k+1}\,. \qquad \text{(by (H1) and (9))}
\end{aligned} \tag{38}
$$

In addition, we have

$$
\begin{aligned}
|||\boldsymbol{\eta}|||^2 &= \tfrac{1}{2}\sum_{n=1}^{N-1}\int_\Omega [\![\boldsymbol{\eta}(t_n)]\!]\cdot\boldsymbol{A}_0[\![\boldsymbol{\eta}(t_n)]\!]\,\mathrm{d}\Omega + \tfrac{1}{2}\int_P \boldsymbol{\eta}\cdot\boldsymbol{M}\boldsymbol{\eta}\,\mathrm{d}P \\
&\quad + \tfrac{1}{2}\Big(\int_\Omega \boldsymbol{\eta}(T^-)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(T^-)\,\mathrm{d}\Omega + \int_\Omega \boldsymbol{\eta}(0^+)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(0^+)\,\mathrm{d}\Omega\Big) \\
&\quad + \int_Q \boldsymbol{\eta}\cdot\boldsymbol{C}\boldsymbol{\eta}\,\mathrm{d}Q + \sum_{n=0}^{N-1}\int_{Q_n} \mathscr{L}\boldsymbol{\eta}\cdot\boldsymbol{\tau}\mathscr{L}\boldsymbol{\eta}\,\mathrm{d}Q \qquad \text{(by (26))} \\
&\leq \tfrac{1}{2}\Big(\int_\Omega \boldsymbol{\eta}(T^-)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(T^-)\,\mathrm{d}\Omega + \int_\Omega \boldsymbol{\eta}(0^+)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(0^+)\,\mathrm{d}\Omega\Big) \\
&\quad + \tfrac{1}{2}\int_P \boldsymbol{\eta}\cdot\boldsymbol{M}\boldsymbol{\eta}\,\mathrm{d}P + c_3\int_Q \boldsymbol{\eta}\cdot\boldsymbol{A}_0\boldsymbol{\eta}\,\mathrm{d}Q + hc_1\sum_{n=1}^{N-1}\int_{Q_n} \mathscr{L}\boldsymbol{\eta}\cdot\boldsymbol{A}_0^{-1}\mathscr{L}\boldsymbol{\eta}\,\mathrm{d}Q \\
&\qquad \text{(by (A1) and (H2))} \\
&\leq C(\boldsymbol{V})h^{2k+1}\,. \qquad \text{(by (H1))}
\end{aligned} \tag{39}
$$

The theorem follows by combining (38) and (39), viz.

$$
|||\boldsymbol{E}|||^2 = |||\boldsymbol{E}^h + \boldsymbol{\eta}|||^2 \leq 2(|||\boldsymbol{E}^h|||^2 + |||\boldsymbol{\eta}|||^2) \leq C(\boldsymbol{V})h^{2k+1}\,. \tag{40}
$$

3. Symmetric parabolic systems

3.1. Problem statement

We are now interested in finding $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{x}, t)$ $\forall \boldsymbol{x} \in \Omega$, $\forall t \in I$ such that:

$$
\mathscr{L}\boldsymbol{V} - \nabla\cdot\boldsymbol{K}\nabla\boldsymbol{V} = \boldsymbol{0} \qquad \text{in } Q\,, \tag{41}
$$

$$
\boldsymbol{V} = \boldsymbol{0} \qquad \text{on } P\,, \tag{42}
$$

$$
\boldsymbol{V}(\boldsymbol{x}, 0) = \boldsymbol{G}(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Omega\,, \tag{43}
$$

where $\boldsymbol{K}$, the diffusivity matrix, is assumed to be a positive-definite and symmetric $m \cdot d \times m \cdot d$ matrix. Note that we consider the case of homogeneous Dirichlet boundary conditions.

3.2. Finite element formulation

The formulation is similar to (10)–(13) of Section 2.3. It is given as follows:

Find $\boldsymbol{V}^h \in \mathcal{V}_n^h$ such that $\forall \boldsymbol{W}^h \in \mathcal{V}_n^h$

$$B_n(\boldsymbol{W}^h, \boldsymbol{V}^h) = L_n(\boldsymbol{W}^h)\,, \quad n = 0, 1, \ldots, N-1\,, \tag{44}$$

where

$$B_n(\boldsymbol{W}^h, \boldsymbol{V}^h) = \int_{Q_n} \boldsymbol{W}^h \cdot \mathscr{L}\boldsymbol{V}^h \,\mathrm{d}Q + \int_{Q_n} \nabla\boldsymbol{W}^h \cdot \boldsymbol{K}\nabla\boldsymbol{V}^h \,\mathrm{d}Q + \int_{\Omega} \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0\boldsymbol{V}^h(t_n^+)\,\mathrm{d}\Omega$$
$$+ \sum_{e=1}^{n_{\mathrm{el}}} \int_{Q_n^e} \mathscr{L}\boldsymbol{W}^h \cdot \boldsymbol{\tau}[\mathscr{L}\boldsymbol{V}^h - \nabla\cdot\boldsymbol{K}\nabla\boldsymbol{V}^h]\,\mathrm{d}Q\,, \quad n = 0, 1, \ldots, N-1\,, \tag{45}$$

and $L_n(\boldsymbol{W}^h)$ is defined by (12) and (13), Q_n^e is the interior of element $\Omega^e \times]t_n, t_{n+1}[$, and the notations are otherwise the same as in Section 2.3.

Remark

This time we assume that the boundary conditions are *strongly* enforced by being included in the definition of the finite element space $\mathcal{V}_n^h$.

The method may again be written as a sum over the N time slabs, viz.

$$B(\boldsymbol{W}^h, \boldsymbol{V}^h) = \sum_{n=0}^{N-1} \int_{Q_n} \boldsymbol{W}^h \cdot \boldsymbol{A}_0\boldsymbol{V}^h_{,t}\,\mathrm{d}Q + \sum_{n=1}^{N-1} \int_{\Omega} \boldsymbol{W}^h(t_n^+) \cdot \boldsymbol{A}_0[\![\boldsymbol{V}^h(t_n)]\!]\,\mathrm{d}\Omega$$
$$+ \int_{\Omega} \boldsymbol{W}^h(0^+) \cdot \boldsymbol{A}_0\boldsymbol{V}^h(0^+)\,\mathrm{d}\Omega + \int_{Q} \boldsymbol{W}^h \cdot \tilde{\boldsymbol{A}}\cdot\nabla\boldsymbol{V}^h\,\mathrm{d}Q + \int_{Q} \boldsymbol{W}^h \cdot \boldsymbol{C}\boldsymbol{V}^h\,\mathrm{d}Q$$
$$+ \sum_{n=0}^{N-1} \int_{Q_n} \nabla\boldsymbol{W}^h \cdot \boldsymbol{K}\nabla\boldsymbol{V}^h\,\mathrm{d}Q + \sum_{n=0}^{N-1}\sum_{e=1}^{n_{\mathrm{el}}} \int_{Q_n^e} \mathscr{L}\boldsymbol{W}^h \cdot \boldsymbol{\tau}[\mathscr{L}\boldsymbol{V}^h - \nabla\cdot\boldsymbol{K}\nabla\boldsymbol{V}^h]\,\mathrm{d}Q\,, \tag{46}$$

and $L(\boldsymbol{W}^h)$ is given in (18).

3.3. Error analysis

Assumption (H1) needs to be modified as follows:

(P1)
$$\int_Q \boldsymbol{\eta}\cdot \boldsymbol{A}_0\boldsymbol{\eta}\,\mathrm{d}Q + h^2\int_Q \nabla\boldsymbol{\eta}\cdot \operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta}\,\mathrm{d}Q + h^2\sum_{n=0}^{N-1}\int_{Qn}\mathscr{L}\boldsymbol{\eta}\cdot\boldsymbol{A}_0^{-1}\mathscr{L}\boldsymbol{\eta}\,\mathrm{d}Q$$
$$+ h^4\sum_{e=1}^{n_{\mathrm{el}}}\sum_{n=0}^{N-1}\int_{Q_n^e}(\nabla\cdot\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta})\cdot\boldsymbol{A}_0^{-1}(\nabla\cdot\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta})\,\mathrm{d}Q \tag{47}$$
$$\leq C(V)h^{2k+2},$$

$$\sum_{n=1}^{N}\int_\Omega \boldsymbol{\eta}(t_n^-)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(t_n^-)\,\mathrm{d}\Omega + \int_\Omega \boldsymbol{\eta}(0^+)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(0^+)\,\mathrm{d}\Omega \leq C(V)h^{2k+1}, \tag{48}$$

where

$$\operatorname{diag}\boldsymbol{A}_0 = \begin{pmatrix} \boldsymbol{A}_0 & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_0 & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{A}_0 \end{pmatrix} \tag{49}$$

is a block-diagonal matrix of size $m\cdot d \times m\cdot d$.

We wish to first establish an error estimate for the case in which diffusion is uniformly small with respect to the space-time discretization. Later on we will return to the general case. We assume the following.

(P3) The diffusivity matrix is of order h in the sense that there exists constants c_4 and c_5 such that

$$\int_Q \nabla\boldsymbol{\eta}\cdot\boldsymbol{K}\nabla\boldsymbol{\eta}\,\mathrm{d}Q \leq c_4 h\int_Q \nabla\boldsymbol{\eta}\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta}\,\mathrm{d}Q, \tag{50}$$

$$\sum_{e=1}^{n_{\mathrm{el}}}\sum_{n=0}^{N-1}\int_{Q_n^e}(\nabla\cdot\boldsymbol{K}\nabla\boldsymbol{\eta})\cdot\boldsymbol{\tau}(\nabla\cdot\boldsymbol{K}\nabla\boldsymbol{\eta})\,\mathrm{d}Q$$
$$\leq c_5 h^3\sum_{e=1}^{n_{\mathrm{el}}}\sum_{n=0}^{N-1}\int_{Q_n^e}(\nabla\cdot\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta})\cdot\boldsymbol{A}_0^{-1}(\nabla\cdot\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta})\,\mathrm{d}Q. \tag{51}$$

The following lemma is a crucial ingredient in the error analysis. It depends on the detailed definition of $\boldsymbol{\tau}$ presented in Part III.

LEMMA 3.1.

$$\int_{Q_n^e}\mathscr{L}\boldsymbol{W}^h\cdot\boldsymbol{\tau}(\nabla\cdot\boldsymbol{K}\nabla\boldsymbol{W}^h)\,\mathrm{d}Q \leq \tfrac{1}{2}\int_{Q_n^e}\mathscr{L}\boldsymbol{W}^h\cdot\boldsymbol{\tau}\mathscr{L}\boldsymbol{W}^h\,\mathrm{d}Q + \tfrac{1}{2}\int_{Q_n^e}\nabla\boldsymbol{W}^h\cdot\boldsymbol{K}\nabla\boldsymbol{W}^h\,\mathrm{d}Q. \tag{52}$$

PROOF. The proof of a similar result is presented in Appendix A of Part III. The interested reader is urged to consult the original source for details.

We now define the norm in which we prove convergence:

$$
\begin{aligned}
|||W^h|||^2 = {} & \tfrac{1}{2} \sum_{n=1}^{N-1} \int_\Omega [\![W^h(t_n)]\!] \cdot A_0 [\![W^h(t_n)]\!] \, \mathrm{d}\Omega + \tfrac{1}{2} \int_\Omega W^h(T^-) \cdot A_0 W^h(T^-) \, \mathrm{d}\Omega \\
& + \tfrac{1}{2} \int_\Omega W^h(0^+) \cdot A_0 W^h(0^+) \, \mathrm{d}\Omega + \int_Q W^h \cdot C W^h \, \mathrm{d}Q \\
& + \tfrac{1}{2} \sum_{n=0}^{N-1} \int_{Q_n} \nabla W^h \cdot K \nabla W^h \, \mathrm{d}Q + \tfrac{1}{2} \sum_{n=0}^{N-1} \int_{Q_n} \mathcal{L} W^h \cdot \boldsymbol{\tau} \mathcal{L} W^h \, \mathrm{d}Q \, .
\end{aligned}
\tag{53}
$$

LEMMA 3.2 (Stability).

$$
|||W^h|||^2 \leqslant B(W^h, W^h) \quad \forall W^h \in \mathcal{V}^h \, .
\tag{54}
$$

PROOF. Employing Lemma 3.1, the result follows in analogous fashion to the proof of Lemma 2.2.

LEMMA 3.3. (Orthogonality of the error).

$$
B(W^h, E) = 0 \quad \forall W^h \in \mathcal{V}^h \, .
\tag{55}
$$

PROOF. The result is a direct consequence of the definitions of B and L.

THEOREM 3.4. Assume (A1), (P1), (H2) *and* (P3) *hold. Then*

$$
|||E|||^2 \leqslant C(V) h^{2k+1} \, .
\tag{56}
$$

PROOF. Proceeding as in the proof of Theorem 2.4, we get:

$$
\begin{aligned}
|||E^h|||^2 \leqslant {} & |B(E^h, \boldsymbol{\eta})| \\
= {} & \Big| \sum_{n=0}^{N-1} \int_{Q_n} E^h \cdot A_0 \boldsymbol{\eta}_{,t} \, \mathrm{d}Q + \sum_{n=1}^{N-1} \int_\Omega E^h(t_n^+) \cdot A_0 [\![\boldsymbol{\eta}(t_n)]\!] \, \mathrm{d}\Omega \\
& + \int_\Omega E^h(0^+) \cdot A_0 \boldsymbol{\eta}(0^+) \, \mathrm{d}\Omega \\
& + \int_Q E^h \cdot \tilde{A} \cdot \nabla \boldsymbol{\eta} \, \mathrm{d}Q + \int_Q E^h \cdot C \boldsymbol{\eta} \, \mathrm{d}Q + \sum_{n=0}^{N-1} \int_{Q_n} \nabla E^h \cdot K \nabla \boldsymbol{\eta} \, \mathrm{d}Q \\
& + \sum_{n=0}^{N-1} \int_{Q_n} \mathcal{L} E^h \cdot \boldsymbol{\tau} \mathcal{L} \boldsymbol{\eta} \, \mathrm{d}Q - \sum_{e=1}^{n_{\mathrm{el}}} \sum_{n=0}^{N-1} \int_{Q_n^e} \mathcal{L} E^h \cdot \boldsymbol{\tau} (\nabla \cdot K \nabla \boldsymbol{\eta}) \, \mathrm{d}Q \Big|
\end{aligned}
$$

$$
\begin{aligned}
&= \Bigg| \sum_{n=0}^{N-1} \int_{Q_n} -\mathscr{L}\boldsymbol{E}^h \cdot \boldsymbol{\eta}\, \mathrm{d}Q - \sum_{n=1}^{N-1} \int_{\Omega} [\![\boldsymbol{E}^h(t_n)]\!] \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(t_n^-)\, \mathrm{d}\Omega \\
&\quad + \int_{\Omega} \boldsymbol{E}^h(T^-) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(T^-)\, \mathrm{d}\Omega \\
&\quad + 2\int_{Q} (\boldsymbol{C}\boldsymbol{E}^h) \cdot \boldsymbol{\eta}\, \mathrm{d}Q + \sum_{n=0}^{N-1} \int_{Q_n} \nabla \boldsymbol{E}^h \cdot \boldsymbol{K} \nabla \boldsymbol{\eta}\, \mathrm{d}Q \\
&\quad + \sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{E}^h \cdot \boldsymbol{\tau} \mathscr{L}\boldsymbol{\eta}\, \mathrm{d}Q - \sum_{e=1}^{n_{\mathrm{el}}} \sum_{n=0}^{N-1} \int_{Q_n^e} \mathscr{L}\boldsymbol{E}^h \cdot \boldsymbol{\tau}(\nabla \cdot \boldsymbol{K} \nabla \boldsymbol{\eta})\, \mathrm{d}Q \Bigg| \\
&= \Bigg| -\sum_{n=0}^{N-1} \int_{Q_n} \left(\frac{1}{\sqrt{6}} \boldsymbol{\tau}^{1/2} \mathscr{L}\boldsymbol{E}^h \right) \cdot \sqrt{6}\boldsymbol{\tau}^{-1/2} \boldsymbol{\eta}\, \mathrm{d}Q \\
&\quad - \sum_{n=1}^{N-1} \int_{\Omega} \left(\frac{1}{\sqrt{2}} \boldsymbol{A}_0^{1/2} [\![\boldsymbol{E}^h(t_n)]\!] \right) \cdot \sqrt{2}\boldsymbol{A}_0^{1/2} \boldsymbol{\eta}(t_n^{-1})\, \mathrm{d}\Omega \\
&\quad + \int_{\Omega} \left(\frac{1}{\sqrt{2}} \boldsymbol{A}_0^{1/2} \boldsymbol{E}^h(T^-) \right) \cdot \sqrt{2}\boldsymbol{A}_0^{1/2} \boldsymbol{\eta}(T^-)\, \mathrm{d}\Omega \\
&\quad + 2\int_{Q} \left(\frac{1}{\sqrt{2}} \boldsymbol{C}^{1/2} \boldsymbol{E}^h \right) \cdot \sqrt{2}\boldsymbol{C}^{1/2} \boldsymbol{\eta}\, \mathrm{d}Q \\
&\quad + \sum_{n=0}^{N-1} \int_{Q_n} \left(\frac{1}{\sqrt{2}} \boldsymbol{K}^{1/2} \nabla \boldsymbol{E}^h \right) \cdot \sqrt{2}\boldsymbol{K}^{1/2} \nabla \boldsymbol{\eta}\, \mathrm{d}Q \\
&\quad + \sum_{n=0}^{N-1} \int_{Q_n} \left(\frac{1}{\sqrt{6}} \boldsymbol{\tau}^{1/2} \mathscr{L}\boldsymbol{E}^h \right) \cdot \sqrt{6}\boldsymbol{\tau}^{1/2} \mathscr{L}\boldsymbol{\eta}\, \mathrm{d}Q \\
&\quad - \sum_{e=1}^{n_{\mathrm{el}}} \sum_{n=0}^{N-1} \int_{Q_n^e} \left(\frac{1}{\sqrt{6}} \boldsymbol{\tau}^{1/2} \mathscr{L}\boldsymbol{E}^h \right) \cdot \sqrt{6}\boldsymbol{\tau}^{1/2} (\nabla \cdot \boldsymbol{K} \nabla \boldsymbol{\eta})\, \mathrm{d}Q \Bigg| \\
&\leqslant \tfrac{1}{2} |||\boldsymbol{E}^h|||^2 + 3\int_{Q} \boldsymbol{\eta} \cdot \boldsymbol{\tau}^{-1} \boldsymbol{\eta}\, \mathrm{d}Q + \sum_{n=1}^{N} \int_{\Omega} \boldsymbol{\eta}(t_n^-) \cdot \boldsymbol{A}_0 \boldsymbol{\eta}(t_n^-)\, \mathrm{d}\Omega + 2\int_{Q} \boldsymbol{\eta} \cdot \boldsymbol{C}\boldsymbol{\eta}\, \mathrm{d}Q \\
&\quad + \sum_{n=0}^{N-1} \int_{Q_n} \nabla \boldsymbol{\eta} \cdot \boldsymbol{K} \nabla \boldsymbol{\eta}\, \mathrm{d}Q + 3\sum_{n=0}^{N-1} \int_{Q_n} \mathscr{L}\boldsymbol{\eta} \cdot \boldsymbol{\tau} \mathscr{L}\boldsymbol{\eta}\, \mathrm{d}Q \\
&\quad + 3\sum_{e=1}^{n_{\mathrm{el}}} \sum_{n=0}^{N-1} \int_{Q_n^e} (\nabla \cdot \boldsymbol{K} \nabla \boldsymbol{\eta}) \cdot \boldsymbol{\tau}(\nabla \cdot \boldsymbol{K} \nabla \boldsymbol{\eta})\, \mathrm{d}Q\,.
\end{aligned}
\tag{57}
$$

Thus, combining with the left-hand side,

$$
\begin{aligned}
|||\boldsymbol{E}^h|||^2 \leqslant\ & 6\int_Q \boldsymbol{\eta}\cdot\boldsymbol{\tau}^{-1}\boldsymbol{\eta}\,\mathrm{d}Q + 2\sum_{n=1}^{N}\int_\Omega \boldsymbol{\eta}(t_n^-)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(t_n^-)\,\mathrm{d}\Omega + 4\int_Q \boldsymbol{\eta}\cdot\boldsymbol{C}\boldsymbol{\eta}\,\mathrm{d}Q \\
&+ 2\sum_{n=0}^{N-1}\int_{Q_n} \nabla\boldsymbol{\eta}\cdot\boldsymbol{K}\nabla\boldsymbol{\eta}\,\mathrm{d}Q + 6\sum_{n=0}^{N-1}\int_{Q_n} \mathscr{L}\boldsymbol{\eta}\cdot\boldsymbol{\tau}\mathscr{L}\boldsymbol{\eta}\,\mathrm{d}Q \\
&+ 6\sum_{e=1}^{n_{\mathrm{el}}}\sum_{n=0}^{N-1}\int_{Q_n^e} (\nabla\cdot\boldsymbol{K}\nabla\boldsymbol{\eta})\cdot\boldsymbol{\tau}(\nabla\cdot\boldsymbol{K}\nabla\boldsymbol{\eta})\,\mathrm{d}Q \\
\leqslant\ & 6\frac{c_2}{h}\int_Q \boldsymbol{\eta}\cdot\boldsymbol{A}_0\boldsymbol{\eta}\,\mathrm{d}Q + 2\sum_{n=1}^{N}\int_\Omega \boldsymbol{\eta}(t_n^-)\cdot\boldsymbol{A}_0\boldsymbol{\eta}(t_n^-)\,\mathrm{d}\Omega + 4c_3\int_Q \boldsymbol{\eta}\cdot\boldsymbol{A}_0\boldsymbol{\eta}\,\mathrm{d}Q \\
&+ 2c_4h\sum_{n=0}^{N-1}\int_{Q_n} \nabla\boldsymbol{\eta}\cdot\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta}\,\mathrm{d}Q + 6c_1h\sum_{n=0}^{N-1}\int_{Q_n} \mathscr{L}\boldsymbol{\eta}\cdot\boldsymbol{A}_0^{-1}\mathscr{L}\boldsymbol{\eta}\,\mathrm{d}Q \\
&+ 6c_5h^3\sum_{e=1}^{n_{\mathrm{el}}}\sum_{n=0}^{N-1}\int_{Q_n^e} (\nabla\cdot\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta})\cdot\boldsymbol{A}_0^{-1}(\nabla\cdot\operatorname{diag}\boldsymbol{A}_0\nabla\boldsymbol{\eta})\,\mathrm{d}Q \\
\leqslant\ & C(\boldsymbol{V})h^{2k+1}\,.
\end{aligned} \tag{58}
$$

We also have

$$|||\boldsymbol{\eta}|||^2 \leqslant C(\boldsymbol{V})h^{2k+1}\,, \tag{59}$$

and therefore,

$$|||\boldsymbol{E}|||^2 \leqslant 2(|||\boldsymbol{E}^h|||^2 + |||\boldsymbol{\eta}|||^2) \leqslant C(\boldsymbol{V})h^{2k+1}\,. \tag{60}$$

This completes the proof of Theorem 3.4.

We now wish to consider the case in which the smallness assumption on the diffusivity, i.e., (P3), is dropped. In this case the best result we can hope for is that we will lose only one power of h in (56) because now the norm $|||\cdot|||$ becomes at least as strong as the $H^1(\cup_{n=0}^{N-1} Q_n)$ norm. Examining the previous proof reveals an ostensible problem: the last term in (57) appears no better than $\mathrm{O}(h^{2k-1})$. In fact, a more careful analysis reveals that the actual result is optimal, namely $\mathrm{O}(h^{2k})$. It depends on the following special case of general inverse estimates for the interpolation error $\boldsymbol{\eta}$ (Arnold, Falk and Scott [1]).

LEMMA 3.5. Let $\boldsymbol{\tau}$ have the following spectral decomposition:

$$\boldsymbol{\tau} = \sum_{i=1}^{m} \boldsymbol{\varphi}_i\tau_i\boldsymbol{\varphi}_i^{\mathrm{t}}\,,$$

where the $\boldsymbol{\varphi}_i$s are the $\boldsymbol{A}_0$-orthogonal eigenvectors of $\boldsymbol{\tau}$ (see Part III *for further details). It is convenient to express $\boldsymbol{K}$ in terms of its $m\times m$ submatrices:*

$$K = [K_{jk}] = \begin{pmatrix} K_{11} & \dots & K_{1d} \\ \vdots & \ddots & \vdots \\ K_{d1} & \dots & K_{dd} \end{pmatrix}, \tag{61}$$

and define

$$\zeta_{ij} = \sum_{k=1}^{d} \varphi_i^{t} K_{jk} \frac{\partial \eta}{\partial x_k} . \tag{62}$$

Assume $V \in (H^{k+1}(Q))^m$ *and that the collection of finite element spaces considered is quasi-uniform. Then, there exists an* $h_0 = h_0(V)$, *and a constant* c_6 *independent of* h *and* V *such that for all* $h \leqslant h_0$,

$$\int_{Q_n^e} \Big(\sum_{j=1}^{d} \zeta_{ij,j}\Big)^2 \,\mathrm{d}Q \leqslant \frac{c_6}{h^2} \int_{Q_n^e} \sum_{j=1}^{d} (\zeta_{ij})^2 \,\mathrm{d}Q , \quad i = 1, 2, \dots, m . \tag{63}$$

LEMMA 3.6.

$$\int_{Q_n^e} \mathscr{L}E^h \cdot \tau(\nabla \cdot K\nabla\eta) \,\mathrm{d}Q \leqslant \tfrac{1}{12} \int_{Q_n^e} \mathscr{L}E^h \cdot \tau \mathscr{L}E^h \,\mathrm{d}Q + 3 \int_{Q_n^e} \nabla\eta \cdot K\nabla\eta \,\mathrm{d}Q . \tag{64}$$

PROOF. This result follows from Lemma 3.5 and an argument similar to that presented in Appendix A of Part III. Again, the reader may wish to consult Part III for further details.

We need also note that (H2) must be modified because in the case when K is not necessarily small, diffusion-dominated modal components of τ will be $\mathrm{O}(h^2)$ whereas advection-dominated components, as before, will be $\mathrm{O}(h)$. Therefore, $\tau = \mathrm{O}(h)$, but $\tau^{-1} = \mathrm{O}(h^{-2})$. Consequently, (23) under (H2) is still in force, but (24) needs to be modified as follows:

$$\text{(P2)} \qquad \int_Q \eta \cdot \tau^{-1} \eta \,\mathrm{d}Q \leqslant \frac{c_7}{h^2} \int_Q \eta \cdot A_0 \eta \,\mathrm{dQ} , \tag{65}$$

where c_7 is a positive constant.

THEOREM 3.7. Assume (A1), (P1), *and* (P2) *hold. Then*

$$|||E|||^2 \leqslant C(V) h^{2k} . \tag{66}$$

PROOF. The proof is the same as the proof of Theorem 3.4, except this time we use (65) and Lemma 3.6 to improve upon (57). Specifically, the last term of (57) may be replaced by the last term of (64). The proof of the theorem is then immediate.

Remarks

(1) The theorems proved in this paper reduce to those quoted in Part III for the steady case.

(2) The error estimates hold in the presence of prescribed source terms and inhomogeneous boundary conditions. This amounts to a straightforward generalization of the methods presented.

(3) We are not aware of any literature on the treatment of well-posed boundary conditions for the general incompletely parabolic case. We conjecture that the present methods will generalize straightforwardly.

(4) The analysis of the parabolic case confirms the appropriateness of the definition of $\boldsymbol{\tau}$ presented in Part III, although it would certainly be desirable to simplify the definition. However, it appears that the decomposition of $\boldsymbol{\tau}$ described in Part III is necessary to establish Lemmas 3.1 and 3.6 which are key results in the analysis.

4. Conclusions

We have developed and analyzed a space-time finite element method for linear symmetric multidimensional advective-diffusive systems. For the general case in which the different modes of the system are simultaneously advection and diffusion dominated, we have obtained optimal error estimates. We believe that our results are the first for the type of equations considered.

Acknowledgment

During the course of this work Leopoldo P. Franca was partially supported by Brazilian Government Fellowship CAPES Proc. 4093/82-5, and M. Mallet was partially supported by an IBM Graduate Fellowship and a contract with Avions Marcel Dassault-Breguet Aviation.

We wish to thank Doug Arnold, Rick Falk and Ridgway Scott for helping us by establishing the general result implying Lemma 3.5.

References

[1] D. Arnold, R. Falk and R. Scott, Private communication, 1986.

[2] T.J.R. Hughes, L.P. Franca and M. Mallet, A new finite element method 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.

[3] T.J.R. Hughes and M. Mallet, A new finite element method for computational fluid dynamics: III. The generalized streamline operator for multidimensional advection-diffusion systems, Comput. Meths. Appl. Mech. Engrg. 58 (1986) 305–328.

[4] C. Johnson, Streamline methods for problems in fluid mechanics, in: R.H. Gallagher, G.F. Carey, J.T. Oden and O.C. Zienkiewicz, eds., Finite Elements in Fluids, Vol. 6 (Wiley, Chichester, U.K., 1986) 251–261.

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

[6] C. Johnson and A. Szepessy, On the convergence of streamline diffusion finite element methods for hyperbolic conservation laws, in: T.E. Tezduyar and T.J.R. Hughes, eds., Numerical Methods for Compressible Flows—Finite Difference, Element and Volume Techniques, AMD-78 (ASME, New York, 1986) 75–91.

[7] U. Nävert, A finite element method for convection-diffusion problems, Ph.D. Thesis, Department of Computer Science, Chalmers University of Technology, Göteborg, Sweden, 1982.