Optimal $h$-$p$ finite element methods

J. Tinsley Oden

Texas Institute for Computational and Applied Mathematics, the University of Texas at Austin, USA

Received 14 November 1992
Revised manuscript received 8 January 1993
EDITORIAL ADDRESSES

John H. ARGYRIS
Institut für Computer Anwendungen
Pfaffenwaldring 27
D-70569 STUTTGART
Germany
( Editorial Office)

Department of Aeronautics
Imperial College of Science
and Technology
Prince Consort Road
LONDON S.W. 7
UK

Thomas J.R. HUGHES
Division of
Applied Mechanics
Durand Building
Room No. 281
Stanford University
STANFORD
CA 94305-4040, USA

J. Tinsley ODEN
The University of Texas
The Texas Institute for
Computational and
Applied Mathematics
3500 West Balcones
Center Drive
AUSTIN
TX 78759, USA

ASSOCIATE EDITORS

K. APPA, Hawthorne, CA
I. BABUŠKA, College Park, MD
A.J. BAKER, Knoxville, TN
T. BELYTSCHKO, Evanston, IL
L. DEMKOWICZ, Krakow
R.E. EWING, Texas, TX
M. FEINGOLD, Marly-le-Roy

R.H. GALLAGHER, Potsdam, NY
R. GLOWINSKI, Houston, TX
H.-O. KREISS, Los Angeles, CA
J.L. LIONS, Paris
H. LOMAX, Moffet Field, CA
C.E. MASSONNET, Liège

L.S.D. MORLEY, Farnborough
K.S. PISTER, Berkeley, CA
G. STRANG, Cambridge, MA
G. VOSKODENSKY, Moscow
W.H. YANG, Ann Arbor, MI
O.C. ZIENKIEWICZ, Swansea

ADVISORY EDITORS

J.F. ABEL, Ithaca, NY
H. ARMEN, Bethpage, NY
K.J. BATE, Cambridge, MA
P.G. BERGAN, Hovik
J.F. BESSELING, Delft
H. BUFLER, Stuttgart
H. CABANES, Paris
C. CANUTO, Torino
G.F. CARRIER, Cambridge, MA
T. CEBECI, Long Beach, CA
A.S.L. CHAN, London
J.L. CHENOT, Valbonne
H. CHRISTIANSEN, Provo, UT
T.J. CHUNG, Huntsville, AL
P.G. CIARLET, Paris
H. COHEN, New York, NY
M.Z. COHN, Waterloo, Ont.
J. DONEA, Ispra
P.R. EISEMAN, New York, NY
B. ENGQUIST, Los Angeles, CA
C.A. FELIPPA, Boulder, CO
K. FENG, Beijing
I. FRIED, Boston, MA
R.A. GELATLY, San Leandro, CA

M. GERADIN, Liège
R. GRUBER, Manno
K.K. GUPTA, Edwards, CA
R.W. HAMMING, Monterey, CA
F.H. HARLOW, Low Alamos, NM
E.J. HAUG, Iowa City, IA
J.C. HEINRICH, Tucson, AZ
M. HOOG, Liège
I. HOLAND, Trondheim
C. JOHNSON, Göteborg
B.Z. KAPLAN, Beer-Sheva
T. KAWAI, Tokyo
J. KESTENS, Brussels
S.W. KEY, La Cañada-Flintridge, CA
W.C. KHUDSON, Sunnyvale, CA
F.A. LECKIE, Santa Barbara, CA
R.W. LEWIS, Swannee
K. LINKWITZ, Stuttgart
Luo Shi-jun, Xi'an
G. MAIER, Milano
J.L. MEEK, St. Lucia, Queensland
A.J. MORRIS, Cranfield
A. NEEDLEMAN, Providence, RI
M.P. NIELSEN, Lyngby
A.K. NOOR, Hampton, VA

This volume is printed on acid-free paper.
Optimal $h$-$p$ finite element methods

J. Tinsley Oden

Texas Institute for Computational and Applied Mathematics, the University of Texas at Austin, USA

Received 14 November 1992
Revised manuscript received 8 January 1993

A theory of optimal Petrov-Galerkin, $h$-$p$ version, finite element approximations is presented. The optimal scheme is defined relative to a fine mesh solution space and relative to an arbitrary symmetric bilinear form. The optimal method leads to a symmetric, positive-definite stiffness matrix which is independent of the coefficients of the given problem, exhibits 'extra superconvergence' properties, and has a relative error that can be calculated exactly, at each point in the problem domain. Various generalizations are also discussed, including the connection of these methods with certain preconditioning schemes.

1. Introduction

Early literature on Galerkin formulations of boundary- and initial-value problems made a distinction between the space of trial functions, in which the solution is sought, and the space of test functions, against which the residual acts. The notion of trial function probably had its origin in optimization theory, in which one sampled functions from a class of admissible functions in an attempt to find one that minimized a cost functional. Test functions may be inspired by the theory of distributions, but the notion of test functions is more primitive. In problems of physics, the fluxes are characterized by their actions on arbitrary test devices; stress, heat flux, etc., are only measurable by actions on measuring devices or transducers (test functions), and the space of test functions represents a mathematical characterization of this notion.

In the mid-1970s, distinctions were made in the finite element literature between Bubnov-Galerkin methods, in which the spaces of trial functions and test functions coincide, and Petrov-Galerkin methods in which the spaces of trial and test functions are different. Indeed, beginning with some of the more fundamental literature of the 1970s, the idea emerged of designing special trial or test functions to achieve a particular property of an algorithm.

Among the early Petrov-Galerkin techniques for finite elements was the upwinding scheme of Wahlbin [1] for first order linear hyperbolic problems. The subsequent work of Dendy [2] dealt with studies of upwinding schemes for linear problems. The paper of Christie et al. [3] dealt with optimal upwinding of linear convection problems and made clear the role of special test functions for stabilizing numerical schemes for such problems. These schemes were predecessors of the SUPG (streamline upwinded Petrov-Galerkin) methods of Brooks and Hughes [4], studied in some detail by Johnson et al. (e.g., [5]). Barrett and Morton [6] discussed the exact construction of optimal test functions for one-dimensional elliptic boundary-value problems and Demkowicz and Oden [7, 8] developed optimal trial and test functions for both one- and two-dimensional linear parabolic problems. Recently, the development of special test functions to deliver specific properties has been exploited by several authors. We mention in particular the ELLAM methods (Euler–Lagrange Local Adjoint Methods) of Celia et al. [9] (see also the references in this work). More recently, Brezzi et al. [10] showed the role of internal 'bubble functions' in producing algorithms which essentially mimic properties of SUPG schemes. With few exceptions (e.g., [4, 8, 10]), the papers mentioned deal with one space dimension and, thus, are not immediately extendable to two- and three-dimensional cases.
In the present exposition, we present a general theory of optimal Petrov-Galerkin methods for \( h-p \) finite element approximations of linear boundary-value problems. We introduce the notion of \( a \)-optimal trial functions and \( a \)-optimal test functions relative to a fine mesh approximation \( \mathcal{V}^h \) of the space \( \mathcal{V} \) to which the exact solution belongs. Here 'a-optimal' refers to optimizations with respect to an arbitrary symmetric positive-definite bilinear form defined on \( \mathcal{V} \), independently of the bilinear form characterizing the boundary-value problem under consideration. The \( a \)-optimal Petrov-Galerkin approximation belongs to a space \( \mathcal{V}^H \subset \mathcal{V}^h \) and corresponds to a coarse-mesh approximation of the problem.

We prove that such \( a \)-optimal \( h-p \) approximations possess a number of remarkable properties. First, one can calculate the \( a \)-optimal solution using a symmetric positive-definite matrix, the entries of which are independent of the coefficients of the given problem. The exact relative error between the \( a \)-optimal and the fine mesh approximation can be precisely calculated anywhere in the domain of the problem, it exhibits 'extra superconvergence' properties in that, independent of the regularity of the exact solution, it vanishes at degrees of freedom shared by the fine- and coarse-mesh approximations. This represents a generalization of the various one-dimensional constructions designed to produce approximations with zero error at nodal points. In addition, in the present construction, exact expressions are derived for the pointwise relative error at any point in the domain, and, of course, from such expressions, the exact relative error can be evaluated in any norm. In addition, it is shown that local (elementwise) versions of \( a \)-optimal trial and test functions can be constructed, so that even though \( a \)-optimality is a global property valid for the connected finite element mesh, essential properties of such approximations can be produced by local element shape functions.

Finally, the relationship between matrix condensation and \( a \)-optimality is established. It is shown in particular, that appropriate constructions of Schur complements of submatrices of the global and local stiffness matrices effectively reproduce the \( a \)-optimal properties of the optimal Petrov-Galerkin scheme. This explains, in part, why the condensation of bubble functions reported by Brezzi et al. [10] led to schemes similar to upwinded Petrov-Galerkin method.

Following the presentation of the theory of \( a \)-optimal approximations in the context of a model convection-dominated elliptic problem in two dimensions, several generalizations of the theory are discussed.

2. Model problem: 2D equilibrium problems with small coefficients

We begin this discussion with the consideration of a model class of steady two-dimensional boundary-value problems of the form

\[
-\nabla \cdot \epsilon \nabla u + c \cdot \nabla u + \sigma u = f \quad \text{in } \Omega, \\
u = 0 \quad \text{on } \partial \Omega, \tag{2.1}
\]

where the coefficient \( \epsilon \) of the diffusion term is assumed to be very small in comparison with the other coefficients,

\[
0 < \epsilon \ll \sigma, \quad \epsilon \ll |c|.
\]

Thus, if \( \epsilon \neq 0 \), problem (2.1) is 'convection dominated,' but even if \( \epsilon = 0 \), the fact that \( \epsilon \) can be very small in comparison with \( \sigma \) leads to notorious computational difficulties. To simplify the discussion, the coefficients \( \epsilon, c, \sigma \) are taken to be constants, although variable coefficients present no conceptual difficulties in the analysis. We are also interested in the non-convective case \( (\epsilon = 0) \) since other techniques for treating convection-dominated problems are also discussed. In (2.1), the domain is simply a rectangular region in \( \mathbb{R}^2 \) with boundary \( \partial \Omega \) and the source term \( f \in L^2(\Omega) \).

In the sequel, we denote by \( H^r(\Omega) \) the usual Sobolev spaces of functions defined on \( \Omega \) with distributional derivatives of order \( r \geq 0 \) in \( L^2(\Omega) \) and with norm
\[ \|v\|_{r,A} = \left\{ \sum_{i=1}^{n} \int_{\Omega} |D^{r}u|^{2} \, dx \right\}^{1/2} \]

with \( dx = dx_{1}, dx_{2}, \) and \( D^{r}u = a^{i_{1}i_{2}}u / \partial x_{i_{1}}^{2} \partial x_{i_{2}}^{2} \). \( \alpha_{i} = \text{integer} \geq 0, \ i = 1, 2 \). When the domain under consideration is clear from the context, we simply write \( \|v\|_{r,A} \).

To recast (2.1) in a functional setting, we introduce the space

\[ \mathcal{V} = H^{1}_{0}(\Omega) = \{ v = v(x) \in H^{1}(\Omega): v = 0 \text{ on } \partial \Omega \} \]

with norm

\[ |v|_{1} = \left\{ \int_{\Omega} \nabla v \cdot \nabla v \, dx \right\}^{1/2} \]

with the boundary values of \( v \) interpreted in the sense of traces of \( H^{1} \)-functions, and we introduce the bilinear and linear forms

\[ \mathcal{B} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}: \mathcal{B}(u, v) = \int_{\Omega} (\varepsilon \nabla u \cdot \nabla v + c \cdot \nabla u \cdot \nabla v + \sigma uv) \, dx \]

\[ \mathcal{L} : v \to \mathbb{R}: \mathcal{L}(v) = \int_{\Omega} fu \, dx \]

The weak formulation of (2.1) is thus: Find \( u \in \mathcal{V} \) such that

\[ \mathcal{B}(u, v) = \mathcal{L}(v) \quad \forall v \in \mathcal{V} \]

REMARK 2.1. It is not difficult to verify that \( \mathcal{L} \) and \( \mathcal{B} \) are continuous on \( \mathcal{V} \). In particular, a constant \( m_{0} > 0 \) exists such that for any \( u, v \in \mathcal{V} \),

\[ \mathcal{B}(u, v) \leq m_{0} |u|_{1} |v|_{1} \]

Also, \( \mathcal{B}(\cdot, \cdot) \) satisfies a Gårding inequality,

\[ \mathcal{B}(v, v) \geq c_{0} \varepsilon |v|_{1}^{2} - \left( \frac{|c|^{2}}{(1 - c_{0}) \varepsilon} - \sigma \right) \|v\|_{0}^{2} \]

for any \( v \) in \( \mathcal{V} \), where \( c_{0} \) is any real number such that \( 0 < c_{0} < 1 \) and \( \|v\|_{0}^{2} = \int_{\Omega} v^{2} \, dx \). Thus, if \( c = 0 \) or convection is small enough that \( |c|^{2} \leq (1 - c_{0}) \varepsilon \sigma \), then

\[ \mathcal{B}(v, v) \geq c_{0} \varepsilon |v|_{1}^{2} \]

whereas if \( |c| \neq 0 \) and \( |c| > (1 - c_{0}) \varepsilon \sigma \), then

\[ \mathcal{B}(v, v) + \mu |v|_{0}^{2} \geq c_{0} \varepsilon |v|_{1}^{2} \]

where \( \mu = -\sigma + |c|^{2} / (1 - c_{0}) \varepsilon \geq 0 \).

REMARK 2.2. It is easily shown that a unique solution \( u \) exists to problem (2.4) and, therefore, \( u \) also the solution to (2.1) in the sense of distributions.
REMARK 2.3. The presence of small $\varepsilon$ in the definition of the model problem strongly influences the structure and stability of the solution $u$. If $|c|^2 \leq (1 - c_0)\varepsilon a$,

$$|u|_1 \leq \frac{1}{c_0\varepsilon} \|f\|_0.$$ 

Thus, any change in $f$ is magnified by $(c_0\varepsilon)^{-1}$ in bounds on the corresponding change in $u$. If $\varepsilon = 0$, the $L^2$-term, $\sigma(u,v)$, dominates $\Omega$ and the addition of small diffusion, $\varepsilon(\nabla u, \nabla v)$, means roughly speaking that any gradients in the solution contribute to the energy only in terms of order $\varepsilon$. In the case of pure convection ($\varepsilon = 0$), the problem becomes hyperbolic. For the time-dependent case, boundary data from a portion of $\partial \Omega$ propagate along the characteristics, the lines $dx/dt = c$. The addition of small diffusion then generally produces a continuous solution in which the (possible) jumps across characteristics are smoothed and spread out over a region of width $O(\sqrt{\varepsilon})$. In this layer, $u$ and $\nabla u$ may undergo rapid changes and, depending on the boundary conditions, very thin boundary layers can be developed along $\partial \Omega$ of width $O(\varepsilon)$.

3. Discretization of the model problem

We consider two $h$-$p$ finite element approximations of problem (2.4): an approximation $u_h$ of $u$ on a 'fine mesh' on $\Omega$, and a coarse mesh approximation $u_l$ on a 'coarser' submesh. The terms 'fine mesh' and 'coarse mesh' may have nothing to do with the fineness or coarseness of the FE mesh, but merely indicate an $h$-$p$ approximation spaces each involving different numbers of degrees of freedom, one being a proper subspace of the other. We will ultimately design $u_l$ so that it is an optimal approximation of $u_h$ on the coarse mesh, in a way that will be made precise later.

3.1. Shape functions and mesh

Let $\{\mathcal{D}_h\}$ denote a family of partitions of $\tilde{\Omega}$ into meshes of regular quadrilateral finite elements. To define a mesh (a partition) $\mathcal{D}_h$, we consider first a master element $\Omega$ which, for present purposes, can be taken to the square, $\tilde{\Omega} = [-1,1]^2$. The fine mesh is produced by a family of $N_{\text{elem}}$ affine invertible maps $F_K$, where

$$F_K: \tilde{\Omega} \rightarrow \tilde{\Omega}_K, \quad F_K((\xi, \eta)) = T_K \begin{bmatrix} \xi \\ \eta \end{bmatrix} + a_K = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

$$\tilde{\Omega} = \bigcup_{1 \leq K \leq N_{\text{elem}}} \tilde{\Omega}_K, \quad \Omega_K \cap \Omega_J = \emptyset \text{ if } K \neq J.$$ 

Here $T_K$ is an invertible matrix and $a_K$ is a translation vector. Standard properties of these partitions are assumed to exist, e.g. (1) $\tilde{\Omega}_K \cap \tilde{\Omega}_J$ is either a common vertex or side or is empty; (2) if $\tilde{\Omega}_K \cap \tilde{\Omega}_J = \Gamma_{KJ}$ is a common side, then $F_K^{-1}$ and $F_J^{-1}$ coincide on this side; (3) vertices of $\Omega$ coincide with those of some $\Omega_K$, etc.

On $\tilde{\Omega}$ we construct families of hierarchic shape functions organized into three categories:

(1) Vertex (or node) functions

$$\hat{\psi}_1(\xi, \eta) = \frac{1}{4}(1 + \xi)(1 + \eta), \quad \hat{\psi}_2(\xi, \eta) = \frac{1}{4}(1 - \xi)(1 + \eta),$$

$$\hat{\psi}_3(\xi, \eta) = \frac{1}{4}(1 - \xi)(1 - \eta), \quad \hat{\psi}_4(\xi, \eta) = \frac{1}{4}(1 + \xi)(1 - \eta),$$

(3.1a)
(2) Edge functions

\[
\begin{align*}
\hat{\xi}_{k}^{(1)}(\xi, \eta) &= \frac{(-1)^{k}}{2} (1 + \eta) \rho_{k}(\xi), \quad k = 2, \ldots, p, \\
\hat{\xi}_{k}^{(2)}(\xi, \eta) &= \frac{(-1)^{k}}{2} (1 - \xi) \rho_{k}(\eta), \quad k = 2, \ldots, p, \\
\hat{\xi}_{k}^{(3)}(\xi, \eta) &= \frac{1}{2} (1 - \eta) \rho_{k}(\xi), \quad k = 2, \ldots, p, \\
\hat{\xi}_{k}^{(4)}(\xi, \eta) &= \frac{1}{2} (1 + \eta) \rho_{k}(\eta), \quad k = 2, \ldots, p.
\end{align*}
\]

Here \(\rho_{k}\) are the integrated Legendre polynomials.

\[
\rho_{k}(\xi) = \sqrt{\frac{2k-1}{2}} \int_{-1}^{1} P_{k-1}(s) \, ds,
\]

where \(P_{k-1}\) is the Legendre polynomial of order \(k-1\).

(3) Bubble (or internal) functions

\[
\hat{\delta}_{ij}(\xi, \eta) = \rho_{i}(\xi) \rho_{j}(\eta), \quad 2 \leq i, j \leq p.
\]

When mapped to a typical element \(\Omega_{k}\) in the mesh, these functions become the local element shape functions for the element and are denoted

\[
\begin{align*}
\psi_{i}^{K} &= \hat{\psi}_{i} \circ F_{K}^{-1}, \quad i = 1, 2, 3, 4, \\
\hat{\xi}_{k}^{(i)} &= \hat{\xi}_{k}^{(i)} \circ F_{K}^{-1}, \quad i = 1, 2, 3, 4, \quad k = 2, \ldots, p, \\
\hat{b}_{ij}^{K} &= \hat{b}_{ij} \circ F_{K}^{-1}, \quad i, j = 2, \ldots, p.
\end{align*}
\]

**Remark 3.1.** Numerous generalizations of (3.1), (3.3) are possible. For example, polynomials of differing degrees \(\rho_{i}^{K}\) could be used in the definition of each edge function \(\hat{\xi}_{k}^{(i)}\) and polynomials of different degree \(q_{1}\) and \(q_{2}\) could be used in the product functions \(\rho_{i}(\xi) \rho_{j}(\eta), \quad i = 2, \ldots, q_{1}, \quad j = 2, \ldots, q_{2}\). This would result in \(4 + \sum_{i=1}^{q_{1}} (p_{i}^{K} - 1) + (q_{1} - 1)(q_{2} - 1)\) degrees of freedom for an element. Also, irregular \(h\)-meshes could be used in which interelement constraints are imposed to enforce continuity of basis functions across interelement boundaries, such as described in [11]. While data structures that employ this level of sophistication are used in our applications, they are unnecessary for our present purposes, which is to demonstrate basic properties of \(h-p\) methods on a simple model problem.

An interpolation function defined on \(\hat{\Omega}_{K}\) as a linear combination of the local shape functions is of the form

\[
\begin{align*}
u_{n}^{K}(x) &= \sum_{i=1}^{4} \left( v_{i}^{K} \psi_{i}^{K}(x) + \sum_{k=2}^{p} u_{k}^{K} \hat{\xi}_{k}^{(i)}(x) \right) + \sum_{i,j=2}^{p} u_{ij}^{K} \hat{b}_{ij}^{K}(x), \\
x \in \hat{\Omega}_{K}, \quad \hat{\Omega}_{K} \subset \mathbb{R}_{h}, \quad 1 \leq K \leq N_{ELEM}.
\end{align*}
\]

Globally, over the connected mesh of finite elements, we demand that the finite element representations be continuous. Let \(\psi_{i}^{K}, \hat{\xi}_{k}^{(i)}\) and \(\hat{b}_{ij}^{K}\) denote the *global basis functions* associated with the shape functions \(\psi_{i}^{K}, \hat{\xi}_{k}^{(i)}\) and \(\hat{b}_{ij}^{K}\), respectively, i.e.,

\[
\psi_{i}(x)|_{\Delta} = \psi_{i}^{K}(x) \quad \text{if node } \Delta \text{ coincides with node } i \text{ of element } \Omega_{K}.
\]
etc. Then the space of \( h-p \) finite element functions generated on \( \mathcal{D} \) is
\[
\mathcal{V}^h = H^1_0(\Omega) \cap \text{span}\{ \{ \psi_i \}, \{ Z_k^{(i)} \}, \{ B_p \} : \Delta = 1, \ldots, N_{\text{NODES}} \}, \quad \alpha = 1, \ldots, N_{\text{ELEM}}; \quad k = 2, \ldots, p_N; \quad i, j = 2, \ldots, p_K \}.
\] (3.7)

and a function \( v_h \) in \( \mathcal{V}^h \) is a continuous function of the form
\[
v_h(x) = \sum_{\Delta=1}^{N_{\text{NODES}}} v^\Delta \psi_\Delta(x) + \sum_{N=1}^{N_{\text{EDGES}}} \sum_{k=2}^{p_N} v^{N,k} Z_N^{(k)}(x) + \sum_{\alpha=1}^{N_{\text{ELEM}}} \sum_{i,j=2}^{p_K} v_{ij}^\alpha B^\alpha_{ij}(x)
\] (3.8)

and
\[
|v_h|_{\Delta, \Omega} = 0.
\] (3.9)

The fine mesh approximation of (2.4) is now characterized by the following problem: Find \( u_h \in \mathcal{V}^h \) such that
\[
\mathcal{B}(u_h, v_h) = \mathcal{L}(u_h) \quad \forall v_h \in \mathcal{V}^h.
\] (3.10)

REMARK 3.2. In (3.8), \( p_N \) is the highest degree of the polynomial edge shape functions sharing a common edge on an interelement boundary and \( p_K \) is the highest order polynomial bubble function for element \( \Omega_K \in \mathcal{D}_h \). Thus, (3.8) provides for the use of edge functions \( Z_k^{(i)} \) of different degrees for different edges (recall Remark 3.1) and, therefore, for a nonuniform distribution of spectral orders \( p \) over the mesh.

3.2. A priori error estimates

The space \( \mathcal{V}^h \) possesses standard interpolation properties of \( h-p \) finite element methods (e.g., [12]). Thus, for element \( \Omega_K \), if \( h_K = \text{diam}(\Omega_K) \) and \( p_K \) is the lowest degree of the complete polynomials contained in the span of shape functions associated with this element, then, for any function \( w \in H^r(\Omega_K), \ r > 0 \), an interpolant \( \tilde{w}_h \) of the form (3.8) can be found such that
\[
\| w - \tilde{w}_h \|_{r,\Omega_K} \leq c \frac{h_K^{r-s}}{p_K^r} \| w \|_{r,\Omega_K} = \min(p_K + s - r, r - s).
\] (3.11)

where \( s = 0, 1 \) and \( c \) is independent of \( h_K, p_K \) and \( w \). Then, globally,
\[
\| w - \tilde{w}_h \|_{r,\Omega} \leq c \left( \sum_{K=1}^{N_{\text{ELEM}}} \frac{h_K^{2r+s}}{p_K^{2(r-2)}} \right) \| w \|_{r,\Omega}^2
\]
\[
\leq c \mu^2(h, p) \| w \|_{r,\Omega}^2.
\] (3.12)

where \( \mu = \sup_K(h_K^{r+s}/p_K^{r-2}) \). If \( r > p_K + s \) for all \( K \) and if \( h = \max h_K \) and \( p = \min p_K \) for \( \mathcal{D}_h \), then, asymptotically as \( h \to 0, p \to \infty \), we may assume that
\[
\| w - \tilde{w}_h \|_{r,\Omega} \leq c \frac{h^{p+s-r}}{p^{r-s}} \| w \|_{r,\Omega}.
\] (3.13)

Since \( \mathcal{V}^h \subset \mathcal{V} \), we can set \( v = v_h \) in (2.4), subtract (3.10), and obtain the orthogonality condition,
\[
\mathcal{B}(e_h, v_h) = 0 \quad \forall v_h \in \mathcal{V}^h.
\] (3.14)

where \( e_h \) is the fine mesh approximation error,
\[
e_h = u - u_h.
\] (3.15)
J.T. Oden, Optimal $h$-$p$ finite element methods

By standard arguments (e.g., [13, 14]), we know that when (2.5) holds

$$|e_h|_{1,0} \leq \left(1 + \frac{m_0}{c_0}\right) \inf_{w_h \in \mathcal{V}_h} |u - w_h|_{1,0}$$

and, therefore, if $u \in H'(\Omega) \cap H_0^0(\Omega)$, the following a priori estimate of the error holds:

$$|e_h|_{1,0} \leq c \left(1 + \frac{m_0}{c_0}\right) \mu(h, p) \|u\|_{1,0} .$$

(3.1f)

We note that this error bound, unfortunately, depends upon $\varepsilon^{-1}$ which may be a very large quantity.

3.3. Simplified case

In order to simplify the discussion and to better manage the notation, we focus on a simplified specific case depicted in Fig. 1. Here each element has nine degrees of freedom and is the affine image of master element $\tilde{\Omega}$ equipped with the following shape functions:

1. **Vertex**
   $$\hat{\psi}_{1,2,3,4}(\xi, \eta) = \frac{1}{2}(1 \pm \xi)(1 \pm \eta) ;$$
   (3.17a)

2. **Edge**
   $$\hat{\xi}_{1,3}(\xi, \eta) = \pm \frac{1}{2}(1 \pm \eta) \rho(\xi) , \quad \hat{\xi}_{2,4}(\xi, \eta) = \mp \frac{1}{2}(1 \pm \eta) \rho(\xi) ,$$
   $$\rho(s) = \sqrt{\frac{3}{8}} (s^2 - 1) ;$$
   (3.17b)

3. **Bubble**
   $$\hat{b}(\xi, \eta) = \rho(\xi) \rho(\eta) .$$
   (3.17c)

As a further simplification, we denote

$$\hat{\xi}_1 = \hat{\psi}_5 , \quad \hat{\xi}_2 = \hat{\psi}_6 , \quad \hat{\xi}_3 = \hat{\psi}_7 , \quad \hat{\xi}_4 = \hat{\psi}_8 ,$$

so that functions $\hat{u}$ defined over these shape functions are represented as the linear combinations

$$\hat{u}(\xi, \eta) = \sum_{i=1}^{8} \hat{u}_i \hat{\psi}_i(\xi, \eta) + \hat{u}_6 \hat{b}(\xi, \eta) .$$

(3.18)

For element $\Omega_K$,

$$u^K_h(x) = \sum_{i=1}^{8} v^i \hat{\psi}_i(x) + v^b \hat{b}(x)$$

(3.19)

with $\psi_i = \hat{\psi}_i \circ F_K^{-1}$, $b^K = \hat{b} \circ F_K^{-1}$, and globally,

$$u_h \in \mathcal{V}_h : \quad u_h(x) = \sum_{N=1}^{M_f} V_N^N \Psi_N(x) + \sum_{a=1}^{N_{\text{ELEM}}} V_a^N B_a(x)$$

(3.20)

($b^K(x) = \delta^K B_a(x), x \in \Omega_K$), and $u_h|_{\partial\Omega} = 0$. For the mesh shown in Fig. 1, $M_f = N_{\text{NODES}} + N_{\text{EDGES}} = 4 \cdot 12 = 16$ and $N_{\text{ELEM}} = 9$. 
The fine mesh solution of (3.10) is of the form

\[ u_h(x) = \sum_{N=1}^{M} U_0^N \psi_N(x) + \sum_{K=1}^{N_{ELEM}} U_0^K B_a(x) . \] \hspace{1cm} (3.21)

so that upon substitution of (3.21) and (3.22), we arrive at the system of equations

\[ K_{NM} U_0^M + K_{Na} U_0^a = F_N , \quad K_{aM} U_0^N + K_{bb} U_b^b = F_b , \] \hspace{1cm} (3.22)

where we sum on repeated indices, \( 1 \leq N,M \leq M_F \), \( 1 \leq \alpha, \beta \leq N_{ELEM} \), and

\[ K_{NM} = \mathcal{B}(\psi_N, \psi_M) , \quad K_{Na} = \mathcal{B}(\psi_N, B_a) , \quad F_N = \mathcal{L}(\psi_N) , \]
\[ K_{aN} = \mathcal{B}(B_a, \psi_N) , \quad K_{bb} = \mathcal{B}(B_a, B_b) , \quad F_b = \mathcal{L}(B_b) . \] \hspace{1cm} (3.23)

Arranging the vertex and edge degrees of freedom into a vector \( \mathbf{U}_0 \) and the internal degrees of freedom into a vector \( \mathbf{U}_b \), we write (3.22) in the matrix form

\[ K_{UU} \mathbf{U}_0 + K_{UB} \mathbf{U}_b = \mathbf{F}_U , \quad K_{BU} \mathbf{U}_0 + K_{BB} \mathbf{U}_b = \mathbf{F}_b . \] \hspace{1cm} (3.24)

Then, eliminating \( \mathbf{U}_b \) by 'matrix condensation', we have

\[ (K_{UU} - K_{UB} K_{bb}^{-1} K_{BU}) \mathbf{U}_0 = \mathbf{F}_U - K_{UB} K_{bb}^{-1} \mathbf{F}_b , \]
\[ \mathbf{U}_b = K_{bb}^{-1} \mathbf{F}_b - K_{bb}^{-1} K_{BU} \mathbf{U}_0 . \] \hspace{1cm} (3.25)
The coefficient matrix of $U_0$ is the Schur complement of $K_{UU}$ and we refer to the right-hand side vector as the Schur complement of $F_U$. Denoting these by

$$\mathcal{S}(K_{UU}) = K_{UU} - K_{UB}K_{BB}^{-1}K_{BU}, \quad \mathcal{S}(F_U) = F_U - K_{UB}K_{BB}^{-1}F_B.$$  \hspace{1cm} (3.2)

(3.25) becomes

$$\mathcal{S}(K_{UU})U_0 = \mathcal{S}(F_U), \quad K_{BB}U_B = F_B - K_{BB}U_0.$$  \hspace{1cm} (3.2')

A similar construction can be made at the element level. Since

$$\mathcal{B}(u, v) = \sum_K \mathcal{B}_K(u, v), \quad \mathcal{L}(v) = \sum_K \mathcal{L}_K(v).$$

$$\mathcal{B}_K(u, v) = \int_{\Omega_K} (\varepsilon \nabla u \cdot \nabla v + c \cdot \nabla u v + \sigma u v) \, dx,$$  \hspace{1cm} (3.2)

$$\mathcal{L}_K(v) = \int_{\partial \Omega_K} f v \, ds.$$  \hspace{1cm} (3.2')

the fine mesh solution satisfies the local equation

$$\mathcal{B}_K(u_h, v_h) = \mathcal{L}_K(v_h) + \int_{\partial \Omega_K} \varepsilon \frac{\partial u}{\partial n} v_h \, ds \quad \forall v_h \in \mathcal{V}^h.$$  \hspace{1cm} (3.2')

Thus,

$$k^{K}_{ij} u_i^K + k^{K}_{ii} u_i^K = f_i^K + \sigma_i^K, \quad (\text{no sum on } K \text{ or } b).$$  \hspace{1cm} (3.3)

$$k^{K}_{ib} u_i^K + k^{K}_{bb} u_b^K = f_b^K + \sigma_b^K,$$  \hspace{1cm} (3.3)

where repeated indices $i, j$ are summed from 1 to 8 and

$$k^{K}_{ij} = \mathcal{B}_K(\psi_i^K, \psi_j^K), \quad k^{K}_{ii} = \mathcal{B}_K(\psi_i^K, \psi_i^K), \quad \ldots \quad \sigma_b^K = \int_{\partial \Omega_K} \varepsilon \frac{\partial u}{\partial n} b^K \, ds.$$  \hspace{1cm} (3.3)

Then, in obvious symbolic notations, for each element $\Omega_K \in \mathcal{O}_h$,

$$\begin{aligned}
(k^{K}_{UU} - k^{K}_{UB}(k^{K}_{bb})^{-1}k^{K}_{BU})u_{0K} &= f^K_0 + \sigma^K_0 - k^{K}_{UB}(k^{K}_{bb})^{-1}(f_b^K + \sigma^K_b), \\
u_b^K &= (k^{K}_{bb})^{-1}[f_b^K + \sigma^K_b - k^{K}_{bb}u_{0K}], \\
\mathcal{S}(k^{K}_{UU})u_{0K} &= \mathcal{S}(f^K_0 + \sigma^K_0), \quad k^{K}_{bb}u_{0K} = f_b^K + \sigma^K_b - k^{K}_{bb}u_{0K}.
\end{aligned}$$  \hspace{1cm} (3.3)

4. Optimal $h$-$p$ finite elements: the coarse-mesh approximation

We now consider a 'coarse-mesh' approximation $u_\mu$ of (2.4) which has the following features:
- $u_\mu$ belongs to a subspace $\mathcal{U}^\mu \subset \mathcal{V}^h$ of trial functions. We will design $\mathcal{U}^\mu$ to have special properties relative to $\mathcal{V}^h$.
- $u_\mu$ will be constructed using an optimal Petrov–Galerkin approximation of (2.4), where the meaning of 'optimal' has to do with the specific design of spaces of trial and test functions and is to be made clear below.
Towards constructing such approximations, we introduce formally two coarse-mesh spaces:

\[ U_h^\prime = \text{span}\{X_N\}_{n=1}^{M} = \text{space of trial functions}, \]
\[ W_h^\prime = \text{span}\{\Phi_N\}_{n=1}^{M} = \text{space of test functions}, \]
\[ U_h \subset U_h^\prime, \quad W_h \subset W_h^\prime, \quad \dim U_h = \dim W_h = M_f. \]

If \( \dim W_h = M_f + N_b \) (\( N_b \) being the number of bubble functions), we will denote by \( W_h \) the \( M_f \)-dimensional subspace of \( W_h \) spanned by the functions \( \{\Phi_N\}_{n=1}^{M} \). The precise definitions of the basis functions \( X_N \) and \( \Phi_N \) are given shortly.

The coarse-mesh Petrov–Galerkin approximation of (2.4) is characterized by the following problem: Find \( u_{II} \in U_h \) such that

\[ B(u_{II}, v_{II}) = L(v_{II}) \quad \forall v_{II} \in W_h. \]  

(4.2)

The approximation error \( e_{II} = u - v_{II} \) satisfies the orthogonality condition

\[ B(e_{II}, v_{II}) = 0 \quad \forall v_{II} \in W_h. \]

(4.3)

and, since

\[ e_{II} = u - u_{II} = u - u_h + u_h - u_{II} = e_h + e_{II}, \]

(4.4)

the relative error between the fine mesh and the coarse mesh approximation,

\[ e_{II} = u_h - u_{II}, \]

(4.5)

satisfies the orthogonality condition

\[ B(e_{II}, v_{II}) = 0 \quad \forall v_{II} \in W_h. \]

(4.6)

(because \( W_h \subset W_h^\prime \)).

4.1. Optimal Petrov–Galerkin methods

We now design the trial and test functions to be optimal relative to a specified bilinear form on \( \mathcal{V} \). In particular, let \( a: \mathcal{V} \times \mathcal{V} \to \mathbb{R} \) be any bilinear form on \( \mathcal{V} \) possessing the following properties:

(a.i) \( a(\cdot, \cdot) \) is symmetric,
(a.ii) \( a(\cdot, \cdot) \) is independent of \( B(\cdot, \cdot) \); i.e., \( a(\cdot, \cdot) \) is independent of \( \varepsilon, \sigma \) and \( \sigma \),
(a.iii) \( a(\cdot, \cdot) \) is positive definite.

For example, one choice of \( a(\cdot, \cdot) \) is

\[ a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx. \quad u, v \in \mathcal{V}. \]

(4.7)

The norm associated with \( a(\cdot, \cdot) \) is denoted

\[ \|v\|_a = \sqrt{a(v, v)}. \]

(4.8)

For example, if \( a(\cdot, \cdot) \) is given by (4.7), then \( \|v\|_a = \|v\| \). Also, restrictions of \( a(\cdot, \cdot) \) to elements \( \Omega_K \) are denoted \( a_K(\cdot, \cdot) \); i.e., \( a(u, v) = \Sigma_K a_K(u, v) \).

In the discussion to follow, \( a(\cdot, \cdot) \) will be a bilinear form satisfying conditions (a.i), (a.ii) and (a.iii).
**DEFINITION 4.1:** a-Optimal trial functions. A set of linearly independent trial functions \( \{X_N\}_{N=1}^{M_F} \) said to be a-optimal relative to \( \mathcal{V}^h \) if and only if

\[
a(X_N, w_h) = 0 \quad \forall w_h \in \mathcal{V}^h - \mathcal{V}^H.
\]

(4.4)

Let \( \{X_N\}_{N=1}^{M_F} \) be such a set of a-optimal trial functions and let \( \mathcal{A} \) denote the triple, \( \{a(\cdot, \cdot), \mathcal{B}(\cdot, \cdot), \{X_N\}\} \).

**DEFINITION 4.2:** a-Optimal test functions. For any triple \( \mathcal{A} \), a set of linearly independent test functions \( \{\Phi_N\}_{N=1}^{M_L} \) is said to be a-optimal relative to \( \mathcal{A} \) and \( \mathcal{V}^h \) if and only if

\[
\mathcal{B}(w_h, \Phi_N) = a(w_h, X_N) \quad \forall w_h \in \mathcal{V}^h \text{ and } \forall X_N, 1 \leq N \leq M_F.
\]

(4.10)

Henceforth, we refer to sets of functions \( \{X_N\}, \{\Phi_N\} \) satisfying (4.4) and (4.5) as simply a-optimal trial and test functions.

For given \( \mathcal{B}(\cdot, \cdot), a(\cdot, \cdot) \) and \( \mathcal{V}^h \), such a-optimal trial and test functions always exist. It is seen that a-optimal trial functions are a-orthogonal to the functions in the fine mesh space \( \mathcal{V}^h \) that do not belong to \( \mathcal{V}^H \). In the case of the simplified examples given in Section 3, the functions \( X_N \) would be orthogonal to the bubble functions \( B_a \). Then (4.9) becomes

\[
a(B_a, X_N) = 0, \quad 1 \leq \alpha \leq M_{\text{ELEM}}, \quad 1 \leq N \leq M_F.
\]

(4.11)

It is noted that the \( X_N \) are independent of \( \mathcal{B}(\cdot, \cdot) \) and, therefore, do not depend upon \( \epsilon \) or \( \alpha \) or \( \mathcal{C} \).

The a-optimal test functions \( \Phi_N \) do indeed depend upon \( \mathcal{B}(\cdot, \cdot) \) and, therefore, on \( \epsilon \) and \( \alpha \) and \( \mathcal{C} \). The construction (4.10) can be regarded as symmetrizing \( \mathcal{B}(\cdot, \cdot) \) (if \( \epsilon \neq 0 \)) or, in any case, rendering the discrete problem to one independent of \( \epsilon \) (and, of course, \( \mathcal{C} \)).

A number of obvious properties of a-optimal Petrov–Galerkin approximations can be listed, but there are also several not so obvious properties that are very important. We record these in the next section. But we first show how such optimal functions can be constructed.

### 4.2. Calculation and localization of a-optimal trial and test functions

We now record two theorems which provide concrete formulas for calculating a-optimal trial and test functions and for the construction of local a-optimal trial and test functions for each element.

**THEOREM 4.1.** (a-Optimal trial functions). Let \( a: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R} \) be a bilinear form satisfying condition (a.i), (a.ii) and (a.iii), and let \( \mathcal{V}^h = \text{span}\{\{\Psi_N\}_{N=1}^{M_F}, \{B_a\}_{a=1}^{M_{\text{ELEM}}}\} \). Then

(i) the functions

\[
X_N(x) = \Psi_N(x) - a_{\Psi_N}a^{\Psi_N}B(x), \quad 1 \leq N \leq M_F, \quad 1 \leq \alpha, \beta \leq N_{\text{ELEM}}
\]

(4.12)

are a-optimal trial functions with respect to \( \mathcal{V}^h \), where

\[
a_{\Psi_N} = a(\Psi_N, B_a), \quad a^{\Psi_N} = (a(B_{\alpha}, B_{\beta}))^{-1}
\]

(4.13)

and repeated indices are summed over their ranges.

(ii) The restrictions of the function \( X_N \) to element \( \Omega_K \subset \mathcal{D} \) are the local a-optimal functions,

\[
\psi_i^K(x) = \psi_i^K(x) - a_{\beta}^{\Psi_N}(a_{\beta})^{-1}b^K(x), \quad x \in \mathcal{D}_K.
\]

(4.14)

and satisfy

\[
a_{\epsilon}(b^K, \psi_i^K) = 0, \quad 1 \leq i \leq 8.
\]

(4.15)
The global matrix
\[ a(\Psi_N, X_M) = a_{NM} - a_{NA} a^{\alpha \phi} a_{BM} = \mathcal{F}(a_{NM}). \]  
\hspace{1cm} (4.16)

where \( a_{NM} = a(\Psi_N, \Psi_M) \) and \( \mathcal{F}(a_{NM}) \) is the Schur complement of \( a_{NM} \).

PROOF. (i) A direct substitution of (4.11) reveals that
\[ a(B, X_N) = a_{AN} - a_{NA} a^{\alpha \phi} a_{BN} = a_{AN} - a_{NA} = 0. \]

Thus, \( X_N \) satisfies (4.11).

(ii) Because \( a_{\alpha \beta} = a(B_{\alpha}, B_{\beta}) \) and \( a^{\alpha \beta} \) are block diagonal, the restriction of \( X_N \) to \( \tilde{\Omega}_h \) is precisely the function defined in (4.14). Local orthogonality condition (4.15) follows by direct substitution.

(iii) This also follows by direct substitution of \( X, \tilde{\Psi} \) into \( a(\Psi_N, X_M) \).

A similar collection of results holds for \( a \)-optimal test functions.

THEOREM 4.2. (\( a \)-optimal test functions). Let \( a: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R} \) be a bilinear form satisfying conditions (a.i), (a.ii) and (a.iii), with \( \mathcal{B}: \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R} \) given by (2.3a), and let \( \mathcal{V}^h = \text{span}\{(\Psi_{\alpha})_{\alpha=1}^{M_{fR}}, (B_{\alpha})_{\alpha=1}^{N_{ELEM}}\} \). Then

(i) the functions
\[ \Phi_N(x) = A^R \Gamma_N(x), \quad 1 \leq N, R \leq M_{fR}. \]  
\hspace{1cm} (4.17)

are \( a \)-optimal test functions relative to \( (a(\cdot, \cdot), \mathcal{B}(\cdot, \cdot), \mathcal{V}^h) \), where
\[ \Gamma_N(x) = \Psi_N(x) - K_{BR} K^{\alpha \beta} B_{\alpha}(x) \]  
\hspace{1cm} (4.18)

and
\[ A^R = \mathcal{F}(K_{MN})^{-1} \mathcal{F}(a_{MN}) \]  
\hspace{1cm} (4.19)

and repeated indices are summed throughout their ranges. Here \( K_{NN}, K_{AN}, K_{NR} \) and \( K_{AB} \) are the stiffness arrays defined in (3.23).

\[ K^{\alpha \beta} = (K_{\alpha \beta})^{-1}, \]  
\hspace{1cm} (4.20)

\[ \mathcal{F}(k_{NM}) = \text{Schur complement of } K_{NM} = K_{NM} - K_{Na} K^{\beta \alpha} K_{BM} \]  
\hspace{1cm} (4.21)

and \( \mathcal{F}(a_{MN}) \) is the Schur complement of \( a_{MN} \) defined in (4.16).

(ii) Moreover, the functions \( \Phi_N \) satisfy the orthogonality condition
\[ \mathcal{B}(B_{\alpha}, \Phi_N) = 0, \quad 1 \leq \alpha \leq N_{ELEM}. \]  
\hspace{1cm} (4.22)

(iii) For each element \( \Omega_k \subset \mathcal{A}_k \), the local shape functions
\[ \phi^K_j(x) = \alpha^K_j \gamma^K_j(x), \quad x \in \tilde{\Omega}_k, \quad 1 \leq i, j \leq 8, \]  
\hspace{1cm} (4.23)

\( j \) summed over its range, where
\[ \gamma^K_j(x) = \psi^K_j(x) - k^{K}_{\beta \phi}(k^{K}_{\phi \beta})^{-1} b^K(x) \]  
\hspace{1cm} (4.24)

and the constants \( \alpha^K_j \) satisfy
\[ \mathcal{F}(k^K_{\alpha \beta}) \alpha^K_j = \mathcal{F}(a_{ik}) \]  
\hspace{1cm} (4.25)
are local a-optimal test functions in the sense that
\[
\mathcal{B}_K(b^K, \varphi_i^K) = 0, \quad 1 \leq i \leq 8, \quad \forall \Omega_K \in \mathcal{Q}_h
\]
and
\[
\phi_N(x)|_{\tilde{\Omega}_K} = \varphi_i^K(x) \quad \text{if } x \in \tilde{\Omega}_K
\]
and index N corresponds to index i of \( \Omega_K \).

**PROOF.** (i) An arbitrary function \( v_h \in \mathcal{V}^h \) is of the form
\[
v_h = V_0^M \psi + V_B^A B_h
\]
Thus, a direct substitution of (4.17) and (4.19) yields
\[
\mathcal{B}(v_h, \phi_N) = \mathcal{B}(B_h, \phi_N) = A_N^R \mathcal{B}(B_h, \psi_M) - K_{MR} \kappa_{MR} \mathcal{B}(\psi_M, B_h) + A_N^R \mathcal{B}(B_h, \psi_R) - K_{RR} \kappa_{RR} \mathcal{B}(\psi_R, B_h)
\]
\[
= A_N^R \mathcal{B}(B_h, \psi_M) + A_N^R \mathcal{B}(B_h, \psi_R) = A_N^R \mathcal{F}(K_{MR}) + 0
\]
\[
= A_N^R \mathcal{F}(a_{MN}) \quad \text{(by (4.19))}
\]
\[
= a(v_h, X_N) \quad \text{(by (4.16))}
\]
In view of (4.9), \( \phi_N \) is an a-optimal test function.

(ii) It is obvious that
\[
\mathcal{B}(B_a, \phi_N) = A_N^R \mathcal{B}(B_a, \psi_R - K_{RR} \kappa_{RR} B_h)
\]
\[
= A_N^R (K_{RR} - K_{RR} \kappa_{RR}) = 0.
\]

(iii) Owing to the block-diagonal structure of \( \mathcal{B}(B_a, B_b) \), it is easily seen that the function \( \gamma_i^K \) of (4.2) is the restriction of \( \Gamma_N \) of (4.18) to \( \tilde{\Omega}_K \), with an appropriate relabeling on indices. The coefficients \( \alpha^K_i \) correspond to element matrices which, when assembled, produce the global matrix \( A^K_N \).

**REMARK 4.1.** Let \( A : \mathcal{V} \rightarrow L^2(\Omega) \) be the operator defined by
\[
\mathcal{B}(u, v) = (Au, v) = (u, A^*v),
\]
where \( A^* \) is the adjoint of \( A \). Since (by virtue of (4.26))
\[
\mathcal{B}_K(b^K, \varphi_i^K) = (b^K, A_Z^k \varphi_i^K) = 0,
\]
the local optimal test functions are designed so that the local adjoint \( A_K \) maps them into a spa.
orthogonal to the bubble function $b^K$ in $L^2(\Omega_K)$. The related Local Adjoint Method and the ELLAM approach (see [9]) choose test functions in the kernel of $A^*_K$.

5. Properties of $a$-optimal methods

Reviewing the notations of the preceding section, we denote
- $a : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ a bilinear form satisfying conditions (a.i), (a.ii) and (a.iii).
- $\{X_N\}_{N=1}^{M}$, a set of $a$-optimal trial functions relative to a fine-mesh finite element space $\mathcal{V}_h$.
- $\mathcal{V}_h = \text{span}\{X_N\}_{N=1}^{M}$, a set of $a$-optimal test functions relative to the bilinear form $\mathcal{B}(\cdot , \cdot )$, the space $\mathcal{V}_h$, and the $a$-optimal trial functions $\{X_N\}_{N=1}^{M}$.
- $\mathcal{W} = \text{span}\{\Phi_N\}_{N=1}^{M}$, the coarse mesh Petrov-Galerkin approximation satisfying (4.2).
- $e_{h} = u_h - u_h$ the relative approximation error.

The most obvious property of the $a$-optimal Petrov-Galerkin approximation is that it is characterized by a symmetric positive-definite system of equations. We record this in the following proposition.

**PROPOSITION 5.1 (Symmetry).** If $u_h(x) = u_h X_N(x)$ is the solution of (4.2), then the coefficients are determined by the system

$$u_h A_{MN} = \mathcal{L}(\Phi_N), \quad 1 \leq M, N \leq M,$$

where $A_{MN}$ is the symmetric positive-definite array

$$A_{MN} = a(X_M, X_N)$$

and $X_N$ and $\Phi_N$ are $a$-optimal trial and test functions, respectively.

**PROOF.** This follows immediately from the fact that $\mathcal{B}(u_h, \Phi_N) = a(u_h, X_N) = u_h A_{MN}$. □

**REMARK 5.1.** Note that $A_{MN} = a(\Psi_M - a_M a^\alpha B_\alpha, X_N) = a(\Psi_M, X_N) = \mathcal{S}(a_{MN})$.

The advantages in computing the coefficients $U_h^M$ using a symmetric positive-definite matrix are, unfortunately, nullified by the work needed to compute the right-hand side. Indeed, using the notations of (3.23) and (3.26),

$$\mathcal{L}(\Phi_N) = A_N^\alpha \mathcal{L}(\Psi_R - K_{R\alpha} K^{\alpha\alpha} B_{\alpha})$$

$$= \mathcal{S}(K_{RM})^{-1} \mathcal{S}(a_{MN})(F_R - K_{R\alpha} K^{\alpha\alpha} F_{\alpha})$$

$$= [\mathcal{S}(K_{RM})^{-1}(F_R - K_{R\alpha} K^{\alpha\alpha} F_{\alpha})][\mathcal{S}(a_{MN})]$$

$$= U_0^M \mathcal{S}(a_{MN}).$$

(5.3)

where $U_0^M$ are the coefficients of the fine mesh vertex and edge functions. Thus, it appears that one must essentially solve the original fine-mesh problem to obtain the vector $\mathcal{L}(\Phi_N)$. Fortunately, in many cases this needs to be done only once locally, for one typical element, and then used successively for each element.

**PROPOSITION 5.2 (Coincidence of coefficients).** Let $u_h(x) = U_0^M \Psi_M(x) + U_R^\alpha B_{\alpha}(x)$ and $u_h(x) = u_h X_M(x)$ be the fine mesh and $a$-optimal coarse mesh approximations of (2.4), respectively. Then

$$u_h^M = U_0^M, \quad 1 \leq M \leq M.$$
J.T. Oden. Optimal h-p finite element methods

**PROOF.** From (5.1) and (5.3), \( u^M_{h} A_{MN} = u^M_{h} \mathcal{S}(a_{hN}) = \mathcal{L}(\Phi_N) = U^M_{h} \mathcal{S}(a_{MN}) \). Thus, (5.4) holds because \( \mathcal{S}(a_{MN}) \) is positive-definite. \( \Box \)

As an intermediate result, we state the following.

**PROPOSITION 5.3** (Independence of coefficients). The relative error \( e_{hH} \) satisfies the inequality

\[
\|e_{hH}\|_a \leq \inf_{w_H \in U_H} \|u_h - w_H\|_a,
\]

where \( \| \cdot \|_a \) and \( w_H \) are independent of the coefficients \( \varepsilon \), \( |\varepsilon| \), \( \alpha \).

**PROOF.** This follows from the definition (4.5) of the \( \alpha \)-optimal test functions, the orthogonal condition (4.3), and the fact that

\[
\|e_{hH}\|_a^2 = a(e_{hH}, e_{hH}) \leq \|e_{hH}\|_a \|u_h - w_H\|_a,
\]

where \( w_H = w_H^N X_N \) (sum on \( N \)) is an arbitrary function in \( \mathcal{W}^H \). \( \Box \)

A much stronger characterization of \( e_{hH} \) can be derived with some simple calculations. Returning (3.8) or to the simplified case (3.20), we recall that any function \( u_h \) in \( \mathcal{W}^h \) can be represented as bilinear combination of vertex, edge, and bubble functions. Thus, the relative error can be written

\[
e_{hH}(x) = E_0^H \psi_N(x) + E_0^a B_a(x),
\]

with \( N \) summed from 1 to \( M_f \) and \( a \) from 1 to \( N_{ELEM} \). The next result establishes that the use of \( \alpha \)-optimal trial and test functions for computing \( u_h \) leads to a type of superconvergence of the method

**PROPOSITION 5.4** (Extrasuperconvergence). If \( u_H \) is the \( \alpha \)-optimal, coarse mesh, Petrov-Galerkin approximation satisfying (4.2), then the relative error \( e_{hH} \) of (5.3) is such that

\[
E_0^N = 0, \quad 1 \leq N \leq M_f.
\]

that is, the relative error at the vertex and edge degrees of freedom is zero.

**PROOF.** Since the \( \alpha \)-optimal test functions \( \Phi_M \in \mathcal{W}^H \),

\[
\begin{align*}
0 = & \mathcal{B}(e_{hH}, \Phi_M) \\
= & \mathcal{B}(E_0^N \psi_N + E_0^a B_a, \Phi_M) \\
= & E_0^N a(\psi_N, X_M) + E_0^a a(B_a, X_M) \quad \text{(by (4.5))} \\
= & E_0^N \mathcal{S}(a_{NM}) \quad \text{(by (4.4))},
\end{align*}
\]

where the Schur complement \( \mathcal{S}(a_{NM}) \) of \( a_{NM} \) is necessarily positive-definite. \( \Box \)

**REMARK 5.2.** In the one-dimensional case, it is well known that optimal trial and test functions \( \psi \) can be designed to make the total error, \( e_H = u - u_h \) precisely zero at the nodal points (see, e.g., [7]). This generalization to two-dimensional cases, with \( e_H \) replaced by \( e_{hH} \), was introduced in [8].

**REMARK 5.3.** Condition (5.7) is easily extended to the more general case in which (3.8) holds. The if
with obvious sums on repeated indices, we have

\[ E^e = 0, \quad \Delta = 1, \ldots, N_{\text{NODES}} \]

\[ E^w = 0, \quad N = 1, \ldots, N_{\text{EDGES}}, \quad k = 2, \ldots, P_N. \]  

**Remark 5.4.** There is another way to arrive at (5.7). Starting with (5.6), we note that

\[
E^w\Psi_M(x) + E^w_bB_k(x) = u_h(x) - u_M(x) \\
= U^w_M\Psi_M(x) + U^w_bB_k(x) - u_M^w(\Psi_M(x) - a_Ma^\alpha b_k(x)) \\
= (U^w_0 - u_M^w)\Psi_M(x) + (U^w_b - u_M^w a_M a^\alpha)B_k(x).
\]

Thus, in view of (5.4), \( E^w_0 = 0 \) and

\[ e_{hT}(x) = (U^w_b - u_M^w a_M a^\alpha)B_k(x). \]  

Next, we construct a precise local a posteriori formula for the relative error \( e_{hT} \).

**Proposition 5.5 (A posteriori error estimates).** Let \( r^K_h \) denote the element residual for element \( \Omega_K \):

\[
r^K_h \coloneqq (f + \nabla \cdot \varepsilon u_h - \varepsilon \cdot \nabla u_h - \epsilon \sigma u_h)|_{\partial K}.
\]  

Then the relative error \( e_{hT} \) at any point \( x \in \Omega_K \) is given by

\[ e_{hT}(x) = (k^{bb}_K)^{-1}(r^K_h, b^K)_K b^K(x) \quad \forall \Omega_K \in \mathcal{K}_h. \]  

where

\[ k^{bb}_K = \mathcal{B}_K(b^K, b^K) \]

and

\[ (r^K_h, b^K)_K = \int_{\Omega_K} r^K_h b^K dx. \]  

**Proof.** On each element \( \Omega_K \), the fine mesh approximation satisfies the equality

\[ \mathcal{B}_K(u_h, v_h) = \mathcal{L}_K(v_h) + \int_{\partial K} \varepsilon \frac{\partial u}{\partial n_K} v_h ds, \]

where \( u \) is the exact solution. Since \( u_h = u_M + e_{hT} \), we have

\[
\mathcal{B}_K(e_{hT}, v_h) = L_K(v_h) - \mathcal{B}_K(u_M, v_h) + \int_{\partial K} \varepsilon \frac{\partial u}{\partial n_K} v_h ds \\
= (r^K_h, v_h)_K + \int_{\partial K} \varepsilon \frac{\partial u}{\partial n_K} v_h ds.
\]

According to (5.7), \( e_{hT}(x) = E^K_h b^K(x) \). Setting \( v_h = b^K \), we obtain

\[ E^K_h \mathcal{B}_K(b^K, b^K) = (r^K_h, b^K)_K \]

from which (5.11) follows. \( \square \)
REMARK 5.5. In the more general case in which \( u_h \) is given by (3.6), the elementwise error is easily shown to be of the form

\[
e_{hh} = (B^K(b^K_{ij}, b^K_{il}))^{-1}r^K_{ij}(x),
\]

\[
r^K_{ij} = \mathcal{L}_K(b^K_{ij}) - B_K(u_h, b^K_{il}), \quad 1 \leq i, j, k, l \leq p, \quad x \in \Omega_K. \tag{5.15}
\]

Several useful results follow from (5.12). For example:

**Directional error indicator.** The gradient of \( e_{hh} \) in \( \Omega_K \) is, of course, calculated immediately from (5.12). Let

\[
\gamma^K_b = (k^{k-1}_b)(r^K_{ii}, b^K)K.
\]

Then the change in \( e_{hh} \) in the direction of a unit vector \( \nu \) is

\[
\nabla e_{hh} \cdot \nu = \gamma^K_b \nu b^K \cdot \nu. \tag{5.17}
\]

A more useful indication of the direction of maximum change of the local relative error may be to identify \( \mathcal{Q}^K \) of elements surrounding \( \Omega_K \).

\[
\mathcal{Q}^K = \bigcup_{j \in \mathcal{Q}^K} \{ \bar{\Omega}_j : \bar{\Omega}_j \cap \Omega_K \neq \emptyset \}.
\]

Since \( \| b' \|_{L^\gamma(\Omega_j)} = \| \hat{b} \|_{L^\gamma(\Omega_j)} \), the number

\[
\eta = \| b \|_{L^\gamma(\Omega_j)} \max_{j \in \mathcal{Q}^K} | \gamma^K_b - \gamma'_{ij} | \tag{5.18}
\]

gives an indication of the change in \( e_{hh} \) as one progresses from \( \Omega_K \) to other elements in \( \mathcal{Q}^K \).

**Local \( L^p \)-estimates.** For \( 1 \leq p < \infty \) and \( \Omega_K \in \mathcal{Q}_h \), we have

\[
\| e_{hh} \|_{L^p(\Omega_K)} = \gamma^K_b \| b^K \|_{L^p(\Omega_K)}. \tag{5.19}
\]

Numerous other examples (e.g., negative- or fractional-norm estimates) could be easily obtained from (5.12).

The next result shows the special role that bubble functions have in stabilizing finite element methods.

**PROPOSITION 5.6.** Let the \( a \)-optimal coarse mesh Petrov--Galerkin approximation \( u_h \) of problem (2.4) be characterized by

\[
u_h(x) = u^{N}_h X^{N}_N(x), \quad x \in \Omega.
\]

\[
u_h|_{\Omega_K}(x) = u^{N}_h X^{N}_N(x), \quad x \in \Omega_K.
\]

where \( X^{N}_N \) are \( a \)-optimal global and local trial functions satisfying (4.9) and (4.15), respectively. Any repeated indices \( N \) and \( i \) are summed throughout their ranges, \( 1 \leq N \leq M_p, 1 \leq i \leq 8 \). Then the coefficient \( u^{i}_{hk} \) are precisely those of the fine mesh approximation obtained through matrix condensation, as in (3.30).

**PROOF.** Note that for every element \( \Omega_K \).
\[ \beta_k(u_{iK}^i, \varphi^K_j) = \mathcal{L}_K(\varphi^K_j) + \int_{\partial \Omega_k^n} \varepsilon \frac{\partial u}{\partial n} \varphi^K_j \, ds . \]  
(5.20)

But

\[ \beta_k(u_{iK}^i, \varphi^K_j) = u_{iK}^i \beta_k \left( \psi^K_i - \frac{\alpha_k(\psi^K_i, b^K)}{\alpha_k(b^K, b^K)} b^K, \varphi^K_j \right) = u_{iK}^i \beta_k \left( \psi^K_i, \varphi^K_j \right) = u_{iK}^i \beta_k \left( \psi^K_i - k_{bb}(k_{bb})^{-1} b^K \right) \alpha^k_j = u_{iK}^i \beta_k \left( k_{ik} - k_{bb}(k_{bb})^{-1} k_{ib} \right) \alpha^k_j \]  
(5.21)

and

\[ \mathcal{L}_K(\varphi^K_j) + \int_{\partial \Omega_k^n} \varepsilon \frac{\partial u}{\partial n} \varphi^K_j \, ds = (f^k + \alpha^k_j - k_{bb}(k_{bb})^{-1} f^k) \alpha^k_j . \]  
(5.22)

Introducing (5.21) and (5.22) into (5.20) gives, equivalently, the first equation in (3.33), multiplied by the coefficients \( \alpha^k_j \) defined in (4.25) which was obtained from (3.30) by matrix condensation.

### 6. Generalizations and discussion

#### 6.1. Interpretation of bubble functions

A review of the orthogonality properties that define \( a \)-optimal trial functions (recall (4.9)) reveals that \( X_N \) need only be \( a \)-orthogonal to the functions in \( \mathcal{V}^h - \mathcal{V}^H \). One could, for example, define \( \mathcal{V}^H \) as merely the space spanned by the nodal basis functions \( \mathcal{V}_n \mathcal{V}_n, \Delta = 1, \ldots, N_{\text{NODES}} \) (recall (3.8)). Then the \( a \)-optimal trial functions \( X_\alpha \) are \( a \)-orthogonal to both the internal bubble functions \( B_\alpha \) and the edge functions \( Z^N_\alpha \). All of our global results still hold, including the supersuperconvergence properties (5.7) and (5.9). Only the error characterization (5.12) is based on the fact that the functions \( b^K \) vanish on the element boundary \( \partial \Omega_k \), but (5.10) is not. Even so, the definition of this boundary on which internal functions vanish is somewhat arbitrary. Indeed, if two adjacent elements share edge functions along a common side (see Fig. 2(a)), then these edge functions qualify as interior bubble functions for the composite element.

Many generalizations of this idea are possible. If one chooses any patch of elements \( \Omega^h_\alpha \subset \Omega \) in an arbitrary \( h-p \) mesh, all of the global basis functions with supports contained in the interior of \( \Omega^h_\alpha \) qualify as bubble functions. This is illustrated in Fig. 2(b). In an adaptive algorithm setting, a cascading sequence of patches could be identified with the structure of these internal functions and, therefore, of the \( a \)-optimal solutions, changing with each adaptation.

#### 6.2. Three-dimensional problems

Generalizations of the theory to three-dimensional cases is straightforward, but, of course, three-dimensional implementations carry a significant increase in complexity and computing effort. In the three-dimensional case, if \( \Omega = [-1, 1]^3 \), we now have vertex, edge and face shape functions as well as the internal bubble functions:

1. **Vertex**

\[ \Psi_{1,2, \ldots, 8}(\xi, \eta, \zeta) = \frac{1}{8}(1 \pm \xi)(1 \pm \eta)(1 \pm \zeta) . \]

2. **Edge**

\[ \xi_{1,2,3,4}(\xi, \eta, \zeta) = \frac{1}{4}(1 \pm \xi)(1 \pm \eta) \rho_k(\zeta) , \]

\[ \xi_{5,6,7,8}(\xi, \eta, \zeta) = \frac{1}{4}(1 \pm \xi)(1 \pm \eta) \rho_k(\eta) , \]

\[ \xi_{9,10,11,12}(\xi, \eta, \zeta) = \frac{1}{4}(1 \pm \eta)(1 \pm \zeta) \rho_k(\xi) . \]
Fig. 2. (a) Bubble functions for a composite element defined by adjacent elements sharing a common side: (b) a patch $\Omega_h^p$ elements in a non-uniform $h$-$p$ mesh for which all interior basis functions vanishing on $\partial \Omega_h^p$ are effectively bubble functions.

(3) Face

$$\delta_{ij}^{1,1} (\xi, \eta, \zeta) = \frac{1}{2} (1 \pm \xi) \rho_i(\eta) \rho_j(\zeta),$$

$$\delta_{ij}^{3,1} (\xi, \eta, \zeta) = \frac{1}{2} (1 \pm \eta) \rho_i(\xi) \rho_j(\zeta),$$

$$\delta_{ij}^{3,5} (\xi, \eta, \zeta) = \frac{1}{2} (1 \pm \zeta) \rho_i(\xi) \rho_j(\eta).$$

(4) Bubble (internal)

$$\delta_{ijk} (\xi, \eta, \zeta) = \rho_i(\xi) \rho_j(\eta) \rho_k(\zeta).$$
with \( 2 \leq i, j, k \leq p \) and \( p \) the integrated Legendre polynomials given by (3.2).

The local function \( \hat{u} \) are thus of the form

\[
\hat{u}(\xi, \eta, \zeta) = \sum_{i=1}^{p} v_{0}^{i} \hat{y}_{i}(\xi, \eta, \zeta) + \sum_{i=1}^{p} \sum_{k=2}^{p} v_{Eh}^{i} \hat{z}_{i}^{(k)}(\xi, \eta, \zeta)
+ \sum_{i=2}^{p} \sum_{j=2}^{p} \sum_{k=2}^{p} v_{j}^{i} \hat{b}_{jk}(\xi, \eta, \zeta).
\]

The remaining steps in defining \( u_{h}, u_{N}, X_{N}, \Phi_{N}, \) etc., for a given form \( a(\cdot, \cdot) \) parallel those discussed for the two-dimensional case.

### 6.3. \( a \)-Optimality and preconditioning

The use of \( a \)-optimality for preconditioning deserves some discussion. Considering again the model problem (2.1) and a bilinear form \( a(\cdot, \cdot) \) with properties (a.i), (a.ii) and (a.iii), we address the issue of poor conditioning of the stiffness matrices that arises with increasing the spectral order \( p \) of the element shape functions. What is sought is a preconditioner that stabilizes the problem in the sense that the condition number grows very slowly with increases in \( p \). We are interested in 'preconditioning by linear elements' (e.g., [13]) or 'nodal preconditioning', in which the stiffness matrices corresponding to nodal degrees of freedom (which generally are well-conditioned with respect to \( p \)) are left intact when designing a preconditioner. Other preconditioning methods, such as those in [14], involve related coarse- and fine-grid computations and could conceivably be brought within the framework of this theory.

In most cases, it is sufficient to focus on a typical element \( \Omega_{K} \subseteq \Omega_{a} \) or the master element \( \hat{\Omega} = F_{K}^{-1}(\hat{\Omega}_{K}) \). A main source of ill-conditioning is the interaction between the edge shape functions \( \zeta_{j}^{K} \) and the bubble functions \( b^{K} \). This problem can be resolved on the element level by using \( a \)-orthogonal edge functions \( \zeta_{j}^{K} \) defined by

\[
\zeta_{j}^{K}(x) = \zeta_{j}^{K}(x) - a_{b}^{K}(a_{b}^{K})^{-1} b^{K}(x),
\]

\[
x \in \hat{\Omega}_{K}, \quad a_{b}^{K} = a_{K}(\zeta_{j}^{K}, b^{K}), \quad 1 \leq j \leq \sum_{r=1}^{4} p_{r}^{E},
\]

where \( p_{r}^{E} \) is the largest polynomial degree of a shape function associated with edge \( r \). Then,

\[
a_{K}(\zeta_{j}^{K}, b^{K}) = 0 \quad \forall j, \forall K.
\]

This is a legitimate definition of local edge shape functions because the values of \( \zeta_{j}^{K} \) and \( \zeta^{K} \) remain the same on the element boundary so that the corresponding global basis functions \( \tilde{Z}_{N} \) (or \( \tilde{Z}_{N}^{(K)} \)) are continuous throughout \( \Omega \).

We now regard the coarse mesh approximation space \( V^{H} \) as that corresponding to only the nodal degrees of freedom; the edge functions and bubble functions lie in \( V^{h} - V^{H} \). For element \( \Omega_{K} \), the local \( a \)-optimal trial functions are of the form

\[
\chi_{\alpha}^{K}(x) = \psi_{\alpha}^{K}(x) - a_{\alpha}^{K}(a_{b}^{K})^{-1} b^{K}(x),
\]

\[
1 \leq \alpha \leq 4, \quad 1 \leq j, \quad k \leq \sum_{r=1}^{4} p_{r}^{E}, \quad x \in \hat{\Omega}_{K}.
\]

where now
\[ a_{nj}^K = a_K(\phi_a^K, \zeta_j^K) , \quad a^{jk} = (a_K(\zeta_j^K, \zeta_k^K))^{-1} \]

eetc. Thus, for all \( a, j \),
\[ a_K(b^K, \chi_a^K) = 0 \quad \text{and} \quad a_K(\zeta_j^K, \chi_a^K) = 0. \] (6.4)

To compute the corresponding \( a \)-optimal test functions, we define
\[ m_{jk}^K = k_{jk}^K + k_{jk}(k_{bb}^K)^{-1} k_{bk}^K , \]
\[ m_j^K = (m_{jk}^K)^{-1} , \]
\[ n_{ja}^K = k_{ja}^K - k_{jk}(k_{bb}^K)^{-1} k_{ba}^K . \] (6.5)

and
\[ \varphi_a^K(x) = \alpha_a^{KB} [\psi_a^K(x) - m_{jk}^K n_{jk}^K \zeta(x) - (k_{bb}^K)^{-1} (k_{bk}^K + k_{kb}^K) m_{jk}^K n_{jk}^K b^K(x)] . \] (6.6)

Then, for arbitrary \( w_h \in \mathcal{V}^h \) \( (w_h(x) = w_0^a \psi_a^K(x) + w_i^j \zeta_j^K(x) + w_k^b b^K(x) ) \), we have
\[ \mathcal{B}(w_h, \phi_a^K) = \sum_{K=1}^{N_{ELEM}} \mathcal{B}_K(w_h, \varphi_a^K) = \sum_{K=1}^{N_{ELEM}} w_0^a \alpha_a^{KB} \mathcal{F}(k_{aa}^K) . \] (6.7)

where
\[ \mathcal{F}(k_{aa}^K) = k_{aa}^K - (k_{bb}^K)^{-1} (k_{bk}^K + k_{kb}^K) m_{jk}^K n_{jk}^K k_{aa}^K - k_{ba}^K n_{jk}^K k_{bb}^K . \] (6.8)

and, as usual, repeated indices are summed (except, of course, the bubble label \( b \)). Then
\[ s_{K=1}^{N_{ELEM}} w_0^a \alpha_a^{KB} \mathcal{F}(k_{aa}^K) = s_{K=1}^{N_{ELEM}} w_0^a \mathcal{F}(a_{ab}^K) . \] (6.9)

where \( s \) is the assembly operation and now
\[ \mathcal{F}(a_{ab}^K) = a_{ab}^K - a_{ab}^K a^{jk} a_{ab}^K - a_{ab}^K (a_{bb}^K)^{-1} a_{ab}^K . \] (6.10)

Remark 6.1. In (6.5), we assume that \( k_{jk}^K \) is invertible: if this is not the case, the indicated operation must be done globally, on the assembled matrices.

Remark 6.2. In (6.5)–(6.9), it is understood that the edge matrices \( k_{jk}^K, k_{jk}^K, k_{jk}^K \) are computed using the \( a \)-orthogonal functions \( \zeta_j^K \) of (6.1) and thus differ from the matrices described in Sections 3–.

Throughout, repeated indices are summed over their ranges. \( 1 \leq a, \beta, \nu \leq 4; 1 \leq i, j, k \leq 5; \gamma \leq 4; p, p \) being the highest degree polynomial edge function on edge \( r \). \( r = 1, 2, 3, 4 \).

The operations leading to (6.9) render the matrix problem of the coefficients \( w_0^a \) (or \( U_i \) independent of the coefficients \( e \), \( k \), \( \sigma \). However, the conditioning of the problem may still grow rapidly with increasing \( p \). To resolve this problem, we introduce the form
\[ c(u_h, v_h) = \sum_{K=1}^{N_{ELEM}} c_K(u_v^a \psi_a^K + u_v^b \psi_a^K) + v_h^b \zeta(x) + v_h^b b^K(x) . \] (6.11)

where
\[ c_K(u_v^a, v_h) = a_K(u_v^a \psi_a^K + u_v^b \psi_a^K) + \sum_i a_K(u_v^i \psi_a^K, v_h^i \zeta(x)) + a_K(u_v^i b^K, v_h b^K) = u_v^a v_h \alpha_{ab}^K + \sum_i u_v^i v_h^i a^K + u_v v_h a^K . \] (6.12)
Here, the parentheses surrounding \((j)\) indicate that repeated index \(j\) is not to be summed within the summation sign.

The major property of the form \(c(\cdot, \cdot)\) is embodied in the following proposition which is proved in [13] and which generalizes a similar analysis for \(h\)-version methods developed in [14].

**PROPOSITION 6.1.** For the element shape functions \(\psi^K_x, \tilde{\zeta}^K_j, b^K, \) with \(\tilde{\zeta}^K_j\) given by (6.1), positive constants \(m_1, m_2\) exists such that

\[
m_1 a(\psi_h, \psi_h) \leq c(\psi_h, \psi_h) \leq m_2 a(\psi_h, \psi_h) \quad \forall \psi_h \in \mathcal{V}^h,
\]

where \(c(\cdot, \cdot)\) is defined by (6.11). Moreover, if a uniform \(p\)-version approximation is used with \(p\) the degree of the largest complete polynomial in the space of element shape functions, or, asymptotically as \(p \to \infty\), \(p\) denotes the largest degree of element shape functions in a family of meshes with \(h = \text{const.}\), then

\[
\frac{m_2}{m_1} \leq c(1 + \log^2 p),
\]

where \(c\) is independent of \(p\) and the partition \(\mathcal{D}_h\).

**REMARK 6.3.** The number \(m_2/m_1\) represents the ‘relative condition number’ (cf. [13]) of the forms \(a(\cdot, \cdot)\) and \(c(\cdot, \cdot)\). Suppose \(c = 0\). Then \(\mathcal{B}(\cdot, \cdot)\) is symmetric-positive definite and the analysis in [13, 14] apply to our model problem. In [13], the preconditioning form is taken to be

\[
c(u_h, v_h) = \sum_{K=1}^{N_{ELM}} \left[ \mathcal{B}_K(u_h^a, v_h^a) + \sum_{i} \mathcal{B}_K(u_h^{(i)} \tilde{\zeta}^K, v_h^{(i)} \tilde{\zeta}^K) + \mathcal{B}_K(u_h b^K, v_h b^K) \right]
\]

(6.15)

Then (6.13) holds with \(a(\cdot, \cdot)\) replaced by \(\mathcal{B}(\cdot, \cdot)\). In this case the constant \(c\) in (6.14) depends upon \(\varepsilon\) and may be very large. The present results thus generalize this result in the following ways: \(\mathcal{B}(\cdot, \cdot)\) can be unsymmetrical and \(c\) in (6.14) is independent of \(\varepsilon\) as well as \(p\) and \(\mathcal{D}_h\).

**REMARK 6.4.** Returning to the construction of \(a\)-optimal trial functions, let us select for the form \(a(\cdot, \cdot)\) the bilinear form \(c(\cdot, \cdot)\) defined in (6.15). Then

\[
a^K_x(\tilde{\zeta}^K, \psi^K_a) = 0 \quad \text{and} \quad a^K_x(b^K, \psi^K_a) = 0,
\]

so that the usual bilinear shape functions associated with nodal points qualify as \(a\)-optimal trial functions:

\[
\chi^K_a(x) = \psi^K_a(x), \quad x \in \Omega_K.
\]

Suppose, again, that \(c = 0\) so that \(\mathcal{B}(\cdot, \cdot)\) is symmetric and positive definite. Then, the functions \(\tilde{\zeta}^K_j\) are \(\mathcal{B}\)-orthogonal to the bubble functions and \(k_{ib} = k^K_{ib} = 0\). In this case, (6.9) degenerates to

\[
\mathcal{S}_{k=1}^{N_{ELM}} w^a_{\alpha \beta} \mathcal{S}(k_{aa}^{K}) = \mathcal{S}_{k=1}^{N_{ELM}} w^a_{\alpha \beta} k^K_{ab},
\]

(6.18)

where

\[
\mathcal{S}(k_{aa}^{K}) = k_{aa}^{K} - k_{ak}^{K} (k_{jk}^{K})^{-1} k_{ja}^{K} - k_{ab}^{K} (k_{bb}^{K})^{-1} k_{b}^{K},
\]

(6.19)
and the global optimal test functions are

$$\phi^{(\lambda)}(x) = \mathcal{S}(K_{\text{r} \lambda})^{-1}K_{\text{r} \lambda}\Psi^{(\lambda)}(x),$$

$$x \in \tilde{\Omega}, \quad 1 \leq \lambda \leq N_{\text{NODES}}.$$  \hspace{1cm} (6.2)

where $K_{\text{r} \lambda}$ is the global stiffness matrix corresponding to nodal degrees of freedom. In other words, the matrices associated with the problem

$$\mathcal{B}(u_h, \phi^{(\lambda)}) = \mathcal{S}(\phi^{(\lambda)}), \quad 1 \leq \lambda \leq N_{\text{NODES}}$$

will have a condition number that grows like $c(1 + \log^3 p)$ with $c$ independent of $p$.

Acknowledgment

The work presented here profited from numerous conversations with Leszek Demkowicz whose help is gratefully acknowledged. The support of several agencies for work on this subject is also gratefully acknowledged; in particular, my work was supported by ONR under Contract N00014-J-92-1161, ARL under Contract DAAH04-92-1-0253, NSF under Grant ASC-9111540, and the Aerothermomechanic Laboratory at NASA Langley under Contract NASA-18746. This work was completed under the support of DARPA Contract DABT63-92-C-0042. Partial support from IBM through an agreement dated January 14, 1991 is also gratefully acknowledged.

References