IN-64
61952
p-18

NASA Technical Memorandum 105382
ICOMP-91-29

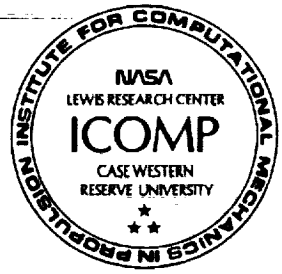
Optimal Least-Squares Finite Element Method for Elliptic Problems

Bo-Nan Jiang
Institute for Computational Mechanics in Propulsion
Lewis Research Center
Cleveland, Ohio

and

Louis A. Povinelli
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio

December 1991



NASA

OPTIMAL LEAST-SQUARES FINITE ELEMENT METHOD FOR ELLIPTIC PROBLEMS

Bo-Nan Jiang
Institute for Computational Mechanics in Propulsion
Lewis Research Center
Cleveland, Ohio 44135

and

Louis A. Povinelli
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135

Summary

In this paper, we propose an optimal least-squares finite element method for 2D and 3D elliptic problems and discuss its advantages over the mixed Galerkin method and the usual least-squares finite element method. In the usual least-squares finite element method, the second-order equation $-\nabla \cdot (\nabla u) + u = f$ is recast as a first-order system $-\nabla \cdot \mathbf{p} + u = f$, $\nabla u - \mathbf{p} = 0$. Our error analysis and numerical experiments show that, in this usual least-squares finite element method, the rate of convergence for flux $\mathbf{p}$ is one-order lower than optimal. In order to get an optimal least-squares method, the irrotationality $\nabla \times \mathbf{p} = 0$ should be included in the first-order system.

1. Introduction

The least-squares finite element method (LSFEM) discussed here is based on minimizing the L_2 norm of the residuals of partial differential equations. In order to use C^0 elements, the second-order 2D elliptic partial differential equation is reduced to a system of three first-order differential equations by introducing two more unknown variables (flux). This idea was first proposed by Lynn and Arya[1] and Zienkiewicz[2], and was an important contribution to the development of least-squares finite element methods. This procedure has long been considered as a standard way to develop least-squares methods.

In this paper, we present both theoretical analysis and numerical results to show that, this simple procedure of reduction destroys ellipticity and the usual LSFEM is not optimal, that is, the rate of convergence for flux is one-order lower than optimal. In order to get an optimal LSFEM, the compatibility condition (the irrotationality) should be included in the first-order system.

The plan of the presentation is as follows. In Section 2, we introduce the model problem and notations. In Section 3, we give a short summary on the related mixed Galerkin method for the purpose of comparison. In Section 4, we analyse the usual LSFEM and explain where the trouble comes from. In Section 5, an optimal LSFEM and the error estimation are presented. In Section 6, we discuss how to deal with some more general elliptic problems. In Section 7, numerical results are given.

2. Preliminaries and Notations

In this paper, we present the essential idea of the optimal least-squares finite element method by solving the following second-order elliptic boundary-value problem:

$$\begin{aligned} -\nabla \cdot \nabla u + u &= f(\mathbf{x}) \qquad in\ \Omega, \\ \nabla u \cdot \mathbf{n} &= g(\mathbf{x}) \qquad on\ \Gamma, \end{aligned} \tag{1}$$

where $\Omega \subset \Re^n$ ($n = 2$ *or* 3) is an open bounded convex domain with a piecewise C^1 boundary Γ, $\mathbf{x} = (x_1, x_2, x_3)$ is a point in Ω, $\mathbf{n} = (n_1, n_2, n_3)$ is a unit outward normal vector on the boundary, and $f(\mathbf{x})$ and $g(\mathbf{x})$ are given functions. Without loss of generality, we shall hereafter consider only the homogeneous boundary condition for simplicity, that is, we shall take $g(\mathbf{x}) \equiv 0$. The primal variable u can be, for instance, temperature for heat conduction; potential for incompressible and irrotational flow; or electric potential for electromagnetics, etc.

Throughout this paper, we use the following notations. $L_2(\Omega)$ denotes the space of

square-integrable functions defined on Ω equipped with the inner product

$$(u,v) = \int_\Omega uvd\mathbf{x} \qquad u,v \in L_2(\Omega)$$

and the norm

$$\|u\|_0^2 = (u,u) \qquad u \in L_2(\Omega).$$

$H^r(\Omega)$ denotes the Sobolev space of functions with square-integrable derivatives of order up to r. $|\cdot|_r$ and $\|\cdot\|_r$ denote the usual seminorm and norm for $H^r(\Omega)$, respectively. For vector-valued function $\mathbf{p}$ with n components, we have the product spaces

$$(L_2(\Omega))^n,\ (H^r(\Omega))^n,$$

and the corresponding norm

$$\|\mathbf{p}\|_0^2 = \sum_{j=1}^n \|p_j\|_0^2,\ \|\mathbf{p}\|_r^2 = \sum_{j=1}^n \|p_j\|_r^2.$$

Further we define the function spaces

$$H = \{v \in H^1(\Omega)\},$$

$$S = \{\mathbf{q} \in (H^1(\Omega))^n | \mathbf{q}\cdot\mathbf{n} = 0 \ on\ \Gamma\},$$

$$W = \{\mathbf{q} \in (L_2(\Omega))^n | \mathbf{q}\cdot\mathbf{n} = 0 \ on\ \Gamma\},$$

and the corresponding finite element subspaces H_h, S_h and W_h, i.e., H_h and S_h are the spaces of continuous piecewise polynomial functions of order k, and W_h is the space of continuous piecewise polynomial functions of order $k-1$. Here the parameter h represents the maximal diameter of the elements. By the finite element interpolation theory[3,4], we have: Given a function $u \in H^{k+1}(\Omega)$ and a function $\mathbf{p} \in (H^{k+1}(\Omega))^n$, there exist an interpolant $\hat{u}^h \in H_h$ and $\hat{\mathbf{p}} \in S_h$ such that

$$\begin{aligned}
\|u - \hat{u}^h\|_0 &\le Ch^{k+1}\|u\|_{k+1},\\
\|u - \hat{u}^h\|_1 &\le Ch^{k}\|u\|_{k+1},\\
\|\mathbf{p} - \hat{\mathbf{p}}^h\|_0 &\le Ch^{k+1}\|\mathbf{p}\|_{k+1},\\
\|\mathbf{p} - \hat{\mathbf{p}}^h\|_1 &\le Ch^{k}\|\mathbf{p}\|_{k+1},
\end{aligned} \tag{2}$$

here and below C denotes a constant independent of the mesh parameter h, with possibly different values in each appearance.

We would also like to write down Green's formula

$$(\nabla \cdot \mathbf{q}, v) + (\mathbf{q}, \nabla v) = \int_\Gamma v\mathbf{q} \cdot \mathbf{n} ds. \tag{3}$$

3. The Mixed Galerkin Method

The most commonly used method for problem (1) is the classical Galerkin method. However, a posteriori numerical differentiation is required to obtain dual variables (flux for heat transfer; velocity for fluid flow; or electric field intensity for electromagnetics) which are often of most interest. In general, the accuracy of so computed dual variables is one-order lower than that of the primal variable.

Mixed Galerkin methods were devised in the hope of getting better accuracy for dual variables[3,5]. Here the term "mixed" means that both the primal variable and the dual variables are approximated as fundamental unknowns.

In mixed methods, problem (1) is decomposed into an equivalent first-order system:

$$\begin{aligned} -\nabla \cdot \mathbf{p} + u &= f \qquad in\ \Omega, \\ \nabla u - \mathbf{p} &= 0 \qquad in\ \Omega, \\ \mathbf{p} \cdot \mathbf{n} &= 0 \qquad on\ \Gamma. \end{aligned} \tag{4}$$

Then the Galerkin principle is applied to problem (4). This leads to the mixed Galerkin weak statement: Find $(u, \mathbf{p}) \in H \times W$ such that

$$\begin{aligned} \int_\Omega (uv + \nabla v \cdot \mathbf{p}) d\mathbf{x} &= \int_\Omega f v d\mathbf{x} \qquad in\ \Omega \qquad \forall v \in H, \\ \int_\Omega (\nabla u \cdot \mathbf{q} - \mathbf{p} \cdot \mathbf{q}) d\mathbf{x} &= 0 \qquad in\ \Omega \qquad \forall \mathbf{q} \in W. \end{aligned} \tag{5}$$

It is well known that problem (5) corresponds to a saddle-point variational problem, and thus in order to guarantee the existence of the solution, the pair $(u, \mathbf{p})$ must satisfy the following Babuška-Brezzi condition:

$$\sup_{\substack{u \in H \\ u \neq 0}} \left(\int_\Omega \nabla u \cdot \mathbf{p} d\mathbf{x} \right) (\|u\|_1)^{-1} \geq C \|\mathbf{p}\|_0 \qquad \forall \mathbf{p} \in W. \tag{6}$$

The Babuška-Brezzi condition precludes the application of simple equal-order finite elements. It can be proved that the finite element spaces H_h and W_h satisfy the discrete Babuska-Brezzi condition (6), and if the solution $(u, \mathbf{p})$ of (4) belongs to $H^{k+1}(\Omega) \times (H^k(\Omega))^n$, we have the following error estimate[5]:

$$|u - u_h|_1 + \|\mathbf{p} - \mathbf{p}_h\|_0 \leq Ch^k(|u|_{k+1} + \|\mathbf{p}\|_k). \tag{7}$$

The estimate (7) tells us that in this mixed method the accuracy for flux $\mathbf{p}$ is always one-order lower than that for the primal variable u.

By inspecting equation (5), we know that the matrix associated with the mixed method is non-positive. This makes the use of iterative methods to solve large-scale problems very difficult.

4. The Usual LSFEM

The usual LSFEM [1,2] is also based on first-order system (4). For 2D problems, the first-order system in (4) consists of three equations and three unknown functions. The first-order system with an odd number of unknowns and an odd number of equations cannot form an elliptic system in the ordinary sense. For 3D problems, although the first-order system in (4) has four unknowns and four equations, it is easy to verify that the system is not elliptic in the ordinary sense. This fact makes us suspect that the LSFEM based on system (4) will not be optimal.

Now let us analyse the usual LSFEM which minimizes the following functional:

$$I : H \times S \rightarrow \Re,$$

$$I(u, \mathbf{p}) = \| - \nabla \cdot \mathbf{p} + u - f\|_0^2 + \|\nabla u - \mathbf{p}\|_0^2. \tag{8}$$

Taking variation of I with respect to u and $\mathbf{p}$, and letting $\delta I = 0$, $\delta u = v$ and $\delta \mathbf{p} = \mathbf{q}$ lead to a least-squares weak statement: Find $U = (u, \mathbf{p}) \in H \times S$, such that

$$B(U, V) = L(V) \qquad \forall V = (v, \mathbf{q}) \in H \times S, \tag{9}$$

where

$$B(U, V) = (-\nabla \cdot \mathbf{p} + u, -\nabla \cdot \mathbf{q} + v) + (\nabla u - \mathbf{p}, \nabla v - \mathbf{q}),$$
$$L(V) = (f, -\nabla \cdot \mathbf{q} + v).$$

The corresponding finite element problem is then to find $U_h = (u_h, \mathbf{p}_h) \in H_h \times S_h$, such that

$$B(U_h, V_h) = L(V_h) \qquad \forall V_h = (v_h, \mathbf{q}_h) \in H_h \times S_h, \tag{10}$$

where

$$B(U_h, V_h) = (-\nabla \cdot \mathbf{p}_h + u_h, -\nabla \cdot \mathbf{q}_h + v_h) + (\nabla u_h - \mathbf{p}_h, \nabla v_h - \mathbf{q}_h),$$
$$L(V_h) = (f, -\nabla \cdot \mathbf{q}_h + v_h).$$

It is easy to verify that

$$B(U, V) \leq C\|U\| \cdot \|V\|, \tag{11}$$

where $\|U\|^2 = \|u\|_1^2 + \|\mathbf{p}\|_0^2 + \|\nabla \cdot \mathbf{p}\|_0^2$. Thus, $B(U, V)$ is continuous on $H \times S$. Since $B(U, V)$ is symmetric, the inequality in the Lax-Milgram theorem reduces to the single coercivity requirement: There exists a constant $\alpha > 0$ such that for $V \in H \times S$

$$B(V, V) \geq \alpha\|V\|^2. \tag{12}$$

Let us prove the coercivity (12). We know that

$$B(V, V) = \| - \nabla \cdot \mathbf{q} + v\|_0^2 + \|\nabla v - \mathbf{q}\|_0^2. \tag{13}$$

Consequently,

$$B(V, V) \geq \|\nabla \cdot \mathbf{q} - v\|_0^2 = \|\nabla \cdot \mathbf{q}\|_0^2 + \|v\|_0^2 - 2(\nabla \cdot \mathbf{q}, v), \tag{14}$$

$$B(V, V) \geq \|\nabla v - \mathbf{q}\|_0^2 = \|\nabla v\|_0^2 + \|\mathbf{q}\|_0^2 - 2(\mathbf{q}, \nabla v). \tag{15}$$

By combining (14) and (15) together and using Green's formula (3) and the boundary condition, we obtain the coercivity:

$$B(V, V) \geq 0.5(\|\nabla v\|_0^2 + \|v\|_0^2 + \|\nabla \cdot \mathbf{q}\|_0^2 + \|\mathbf{q}\|_0^2),$$

or

$$B(V, V) \geq 0.5(\|v\|_1^2 + \|\nabla \cdot \mathbf{q}\|_0^2 + \|\mathbf{q}\|_0^2). \tag{16}$$

REMARK. We may say that the coercivity (16) is incomplete or deficient, because the derivatives of $\mathbf{q}$ are not completely controlled in general. This explains why the usual LSFEM often does not behave well.

Therefore, the following theorem about the rate of convergence of the usual least-squares finite element solutions can be derived.

THEOREM 4.1. Assume that $f(\mathbf{x}) \in L_2(\Omega)$, the solution $(u, \mathbf{p})$ of (4) belongs to $H^{k+1}(\Omega) \times (H^{k+1}(\Omega))^n$, and the finite element interpolation estimates (2) hold. Then for the approximate solution associated with (10), we have the error estimate:

$$\|u - u_h\|_1 + \|\nabla \cdot (\mathbf{p} - \mathbf{p}_h)\|_0 + \|\mathbf{p} - \mathbf{p}_h\|_0 \leq Ch^k(\|u\|_{k+1} + \|\mathbf{p}\|_{k+1}). \tag{17}$$

This theorem shows that the accuracy of the flux $\mathbf{p}$ for the usual LSFEM is one-order lower than optimal. Even so, the usual LSFEM has significant advantages over the mixed Galerkin method. Namely the usual LSFEM is not subject to the Babuška-Brezzi condition and thus can accommodate simple equal-order elements, and the resulting matrix is symmetric and positive definite and thus simple iterative methods, such as conjugate gradient methods, can be employed and vectorization and parallelization are trivial.

5. The Optimal LSFEM

Conservative laws and constitutive laws in physics are in general governed by first-order systems. For historic reasons (convenience for hand calculation and analysis), the equations in a first-order system are combined into a high-order partial differential equation (or equations) with one or less unknowns. For example, for incompressible and irrotational flows, by introducing the potential (or the stream function in 2D cases), the incompressibility and the irrotationality are combined into a second-order Laplace or Poisson equation.

We believe that in the computer age this transformation is unnecessary. We may solve directly the original first-order governing equations by LSFEM. For flows considered in this paper, the governing equations are the following:

$$-\nabla \cdot \mathbf{p} + u = f \qquad in\ \Omega, \tag{18.1}$$

$$\nabla \times \mathbf{p} = 0 \qquad in\ \Omega, \tag{18.2}$$

$$\nabla u - \mathbf{p} = 0 \qquad in\ \Omega, \tag{18.3}$$

$$\mathbf{p} \cdot \mathbf{n} = 0 \qquad on\ \Gamma. \tag{18.4}$$

Here (18.1) is the mass conservation, (18.2) represents the irrotationality, and (18.3) is the constitutive relation.

For 2D problems, the first-order system in (18) consists of three unknowns and four equations. At first glance, one may think that this is an overdetermined system. In fact, this is a determined system in the sense that the components p_1 and p_2 of $\mathbf{p}$ in (18.3) are not completely independent. They must satisfy (18.2).

Let us show that system (18) is determined and elliptic. In 2D cases, we introduce a dummy variable ϕ (this technique was first pointed out to us by C.L.Chang for 2D problems), and rewrite system (18) as

$$\frac{\partial p_1}{\partial x_1} + \frac{\partial p_2}{\partial x_2} - u = -f \qquad in\ \Omega, \tag{19.1}$$

$$-\frac{\partial p_1}{\partial x_2} + \frac{\partial p_2}{\partial x_1} = 0 \qquad in\ \Omega, \tag{19.2}$$

$$\frac{\partial u}{\partial x_1} - \frac{\partial \phi}{\partial x_2} - p_1 = 0 \qquad in\ \Omega, \tag{19.3}$$

$$\frac{\partial u}{\partial x_2} + \frac{\partial \phi}{\partial x_1} - p_2 = 0 \qquad in\ \Omega, \tag{19.4}$$

$$p_1 n_1 + p_2 n_2 = 0 \qquad on\ \Gamma, \tag{19.5}$$

$$\phi = 0 \qquad on\ \Gamma. \tag{19.6}$$

$\partial(19.3)/\partial x_2 - \partial(19.4)/\partial x_1$ leads to $\partial^2\phi/\partial x_1^2 + \partial^2\phi/\partial x_2^2 = 0$. We have already specified that $\phi = 0$ *on* Γ, thus $\phi \equiv 0$ *in* Ω. This means that system (18) with three unknowns and four equations is indeed equivalent to system (19) with four unknowns and four equations.

Now we write (19) in a standard matrix form:

$$\mathbf{A}_1 \frac{\partial \mathbf{u}}{\partial x_1} + \mathbf{A}_2 \frac{\partial \mathbf{u}}{\partial x_2} + \mathbf{A}\mathbf{u} = \mathbf{f} \qquad in\ \Omega, \tag{20}$$

in which

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{A}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} f \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} p_1 \\ p_2 \\ u \\ \phi \end{pmatrix}.$$

Since

$$det(\mathbf{A}_1\xi + \mathbf{A}_2\eta) = det \begin{pmatrix} \xi & \eta & 0 & 0 \\ -\eta & \xi & 0 & 0 \\ 0 & 0 & \xi & -\eta \\ 0 & 0 & \eta & \xi \end{pmatrix} = (\xi^2 + \eta^2)^2 \neq 0$$

for all nonzero real pair (ξ, η), system (19) and thus system (18) is determined and elliptic, as contended.

In 3D cases, by introducing dummy unknowns $\phi, \omega_1, \omega_2, \omega_3$, we write the following first-order system with eight unknowns and eight equations:

$$
\begin{aligned}
\frac{\partial p_1}{\partial x_1} + \frac{\partial p_2}{\partial x_2} + \frac{\partial p_3}{\partial x_3} - u &= -f \qquad in\ \Omega, \\
-\frac{\partial p_2}{\partial x_3} + \frac{\partial p_3}{\partial x_2} + \frac{\partial \phi}{\partial x_1} + \omega_1 &= 0 \qquad in\ \Omega, \\
-\frac{\partial p_3}{\partial x_1} + \frac{\partial p_1}{\partial x_3} + \frac{\partial \phi}{\partial x_2} + \omega_2 &= 0 \qquad in\ \Omega, \\
-\frac{\partial p_1}{\partial x_2} + \frac{\partial p_2}{\partial x_1} + \frac{\partial \phi}{\partial x_3} + \omega_3 &= 0 \qquad in\ \Omega, \\
-\frac{\partial \omega_2}{\partial x_3} + \frac{\partial \omega_3}{\partial x_2} + \frac{\partial u}{\partial x_1} - p_1 &= 0 \qquad in\ \Omega, \\
-\frac{\partial \omega_3}{\partial x_1} + \frac{\partial \omega_1}{\partial x_3} + \frac{\partial u}{\partial x_2} - p_2 &= 0 \qquad in\ \Omega, \\
-\frac{\partial \omega_1}{\partial x_2} + \frac{\partial \omega_2}{\partial x_1} + \frac{\partial u}{\partial x_3} - p_3 &= 0 \qquad in\ \Omega, \\
\frac{\partial \omega_1}{\partial x_1} + \frac{\partial \omega_2}{\partial x_2} + \frac{\partial \omega_3}{\partial x_3} &= 0 \qquad in\ \Omega, \\
p_1 n_1 + p_2 n_2 + p_3 n_3 &= 0 \qquad on\ \Gamma, \\
\phi &= 0 \qquad on\ \Gamma, \\
\omega_1 = \omega_2 = \omega_3 &= 0 \qquad on\ \Gamma,
\end{aligned}
\tag{21}
$$

It is not difficult to verify that $\phi = \omega_1 = \omega_2 = \omega_3 \equiv 0$, and thus system (18) with four unknowns and seven equations is indeed equivalent to system (21) with eight unknowns and eight equations.

We may write system (21) in a standard matrix form:

$$
\mathbf{A}_1 \frac{\partial \mathbf{u}}{\partial x_1} + \mathbf{A}_2 \frac{\partial \mathbf{u}}{\partial x_2} + \mathbf{A}_3 \frac{\partial \mathbf{u}}{\partial x_3} + \mathbf{A}\mathbf{u} = \mathbf{f} \qquad in\ \Omega, \tag{22}
$$

in which

$$
\mathbf{A}_1 = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{pmatrix}, \quad
\mathbf{A}_2 = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix},
$$

$$\mathbf{A}_3 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{f} = \begin{pmatrix} f \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ u \\ \phi \\ \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix}.$$

Since

$$det(\mathbf{A}_1\xi+\mathbf{A}_2\eta+\mathbf{A}_3\zeta) = det \begin{pmatrix} \xi & \eta & \zeta & 0 & 0 & 0 & 0 & 0 \\ 0 & -\zeta & \eta & 0 & \xi & 0 & 0 & 0 \\ \zeta & 0 & -\xi & 0 & \eta & 0 & 0 & 0 \\ -\eta & \xi & 0 & 0 & \zeta & 0 & 0 & 0 \\ 0 & 0 & 0 & \xi & 0 & 0 & -\zeta & \eta \\ 0 & 0 & 0 & \eta & 0 & \zeta & 0 & -\xi \\ 0 & 0 & 0 & \zeta & 0 & -\eta & \xi & 0 \\ 0 & 0 & 0 & 0 & 0 & \xi & \eta & \zeta \end{pmatrix} = -(\xi^2+\eta^2+\zeta^2)^4 \neq 0$$

for all nonzero real triplet (ξ, η, ζ), system (21) and thus system (18) is determined and elliptic.

Now we may use the error analysis developed by Aziz, Kellogg and Stephens[6] for general elliptic systems to show that the LSFEM based on system (18) is optimal. However, their method involves high-level mathematics. Here we use elementary analysis to give a poof of the optimality. The key point of this proof is the following technical lemma:

LEMMA 5.1. Every function $\mathbf{q} \in S = \{\mathbf{q} \in (H^1(\Omega))^n | \mathbf{q} \cdot \mathbf{n} = 0 \text{ on } \Gamma\}$ satisfies:

$$|\mathbf{q}|_1^2 \leq \|\nabla \cdot \mathbf{q}\|_0^2 + \|\nabla \times \mathbf{q}\|_0^2. \tag{23}$$

The lemma is discussed in Girault and Raviart[7], and the complete proof can be found in Grisvard[8]. Here we give an elementary proof for 2D rectangular domains. For 3D rectangular domains, the proof is similar.

Proof.

Since

$$\|\nabla\cdot\mathbf{q}\|_0^2+\|\nabla\times\mathbf{q}\|_0^2=|\mathbf{q}|_1^2+\int_\Omega(\frac{\partial q_1}{\partial x_1}\frac{\partial q_2}{\partial x_2}+\frac{\partial q_1}{\partial x_1}\frac{\partial q_2}{\partial x_2}-\frac{\partial q_2}{\partial x_1}\frac{\partial q_1}{\partial x_2}-\frac{\partial q_2}{\partial x_1}\frac{\partial q_1}{\partial x_2})d\mathbf{x},$$

using integration by parts, we have

$$\|\nabla\cdot\mathbf{q}\|_0^2+\|\nabla\times\mathbf{q}\|_0^2=|\mathbf{q}|_1^2+\int_\Gamma(n_1q_1\frac{\partial q_2}{\partial x_2}+n_2q_2\frac{\partial q_1}{\partial x_1}-n_1q_2\frac{\partial q_1}{\partial x_2}-n_2q_1\frac{\partial q_2}{\partial x_1})ds.$$

The boundary term in the foregoing equation is equal to zero by virtue of the boundary condition $\mathbf{q}\cdot\mathbf{n}=0$. For example, for the part of boundary with $n_1=1$, $n_2=0$, the boundary condition is $q_1=0$, and thus $\partial q_1/\partial x_2=0$. Therefore, the boundary integration associated with this part of boundary is equal to zero.

The optimal LSFEM minimizes the following functional:

$$I:H\times S\rightarrow\Re,$$

$$I(u,\mathbf{p})=\|-\nabla\cdot\mathbf{p}+u-f\|_0^2+\|\nabla\times\mathbf{p}\|_0^2+\|\nabla u-\mathbf{p}\|_0^2. \tag{24}$$

Taking variation of I with respect to u and $\mathbf{p}$, and letting $\delta I=0$, $\delta u=v$ and $\delta\mathbf{p}=\mathbf{q}$ lead to a least-squares weak statement: Find $U=(u,\mathbf{p})\in H\times S$, such that

$$B(U,V)=L(V)\qquad\forall V=(v,\mathbf{q})\in H\times S, \tag{25}$$

where

$$B(U,V)=(-\nabla\cdot\mathbf{p}+u,-\nabla\cdot\mathbf{q}+v)+(\nabla\times\mathbf{p},\nabla\times\mathbf{q})+(\nabla u-\mathbf{p},\nabla v-\mathbf{q}),$$
$$L(V)=(f,-\nabla\cdot\mathbf{q}+v).$$

The corresponding finite element problem is then to find $U_h=(u_h,\mathbf{p}_h)\in H_h\times S_h$, such that

$$B(U_h,V_h)=L(V_h)\qquad\forall V_h=(v_h,\mathbf{q}_h)\in H_h\times S_h, \tag{26}$$

where

$$B(U_h,V_h)=(-\nabla\cdot\mathbf{p}_h+u_h,-\nabla\cdot\mathbf{q}_h+v_h)+(\nabla\times\mathbf{p}_h,\nabla\times\mathbf{q}_h)+(\nabla u_h-\mathbf{p}_h,\nabla v_h-\mathbf{q}_h),$$
$$L(V_h)=(f,-\nabla\cdot\mathbf{q}_h+v_h).$$

It is easy to verify that

$$B(U,V)\leq C\|U\|\cdot\|V\|, \tag{27}$$

where

$$\|U\|^2 = \|u\|_1^2 + \|\mathbf{p}\|_1^2$$

Thus, $B(U,V)$ is continuous on $H \times S$. Since $B(U,V)$ is symmetric, the inequality in the Lax-Milgram theorem reduces to the single coercivity requirement: There exists a constant $\alpha > 0$ such that for $V \in H \times S$

$$B(V,V) \geq \alpha \|V\|^2. \tag{28}$$

Following the similar argument as in Section 3, we may get

$$B(V,V) \geq 0.5(\|\nabla v\|_0^2 + \|v\|_0^2 + \|\nabla \cdot \mathbf{q}\|_0^2 + \|\nabla \times \mathbf{q}\|_0^2 + \|\mathbf{q}\|_0^2). \tag{29}$$

The combination of (29) and Lemma 5.1 yields the coercivity (28).

Once the coercivity is proved, the derivation of the following theorem is trivial.

THEOREM 5.2. Assume that $f(\mathbf{x}) \in L_2(\Omega)$, the solution $(u, \mathbf{p}) \in H^{k+1}(\Omega) \times (H^{k+1}(\Omega))^n$ and the finite element interpolation estimates (2) hold. Then for the approximate solution associated with (26), we have the error estimate:

$$\|u - u_h\|_1 + \|\mathbf{p} - \mathbf{p}_h\|_1 \leq Ch^k(\|u\|_{k+1} + \|\mathbf{p}\|_{k+1}). \tag{30}$$

This theorem shows that the rate of convergence (in H^1 norm) of the LSFEM based on the full first-order system (18) is optimal for all variables. The optimal L_2 convergence can be obtained by Aubin-Nitsche method[3]. The optimality attributes to the fact that the optimal LSFEM controls not only the divergence, but also the curl of the error of flux.

6. Discussion

If there is no u term in the first equation of (1), the corresponding (4) becomes the so called div-grad system and the corresponding (18) is the div-grad-curl system. If the boundary condition is still only related to flux $\mathbf{p}$, then the calculation of u and $\mathbf{p}$ can be separated. We may use the LSFEM based on the div-curl system to obtain $\mathbf{p}$ first, then use $\mathbf{p}$ to calculate u. The LSFEM based on the div-curl system is optimal[9,10].

We also would like to discuss the optimal LSFEM for some more general cases. For example, 2D seepage can be modelled by a system of first-order partial differential equations of the form:

$$\frac{\partial p_1}{\partial x_1} + \frac{\partial p_2}{\partial x_2} = f \qquad in\ \Omega, \tag{31.1}$$

$$p_1 = a_{11}\frac{\partial u}{\partial x_1} + a_{12}\frac{\partial u}{\partial x_2} \qquad in\ \Omega, \tag{31.2}$$

$$p_2 = a_{21}\frac{\partial u}{\partial x_1} + a_{22}\frac{\partial u}{\partial x_2} \qquad in\ \Omega, \tag{31.3}$$

$$p_1 n_1 + p_2 n_2 = g \qquad on\ \Gamma, \tag{31.4}$$

where u denotes the hydraulic head, p_1 and p_2 are the components of seepage velocity, and f and g are given functions.

For simplicity, we consider the case in which the coefficient a_{ij} are constant, and the matrix (a_{ij}) is symmetric and positive definite. Thus, the constitutive relations (31.2) and (31.3) are invertible:

$$\frac{\partial u}{\partial x_1} = \frac{1}{(a_{11}a_{22} - a_{12}a_{21})}(a_{22}p_1 - a_{12}p_2), \tag{32.1}$$

$$\frac{\partial u}{\partial x_2} = \frac{1}{(a_{11}a_{22} - a_{12}a_{21})}(a_{11}p_2 - a_{21}p_1). \tag{32.2}$$

From (32.1) and (32.2) we can obtain the compatibility condition:

$$a_{22}\frac{\partial p_1}{\partial x_2} - a_{12}\frac{\partial p_2}{\partial x_2} - a_{11}\frac{\partial p_2}{\partial x_1} + a_{21}\frac{\partial p_1}{\partial x_1} = 0. \tag{33}$$

The LSFEM based on equation (31.1), (31.2), (31.3) and (33) will be optimal. 3D seepage can be similarly treated.

7. Numerical Results

As the first example, we chose

$$f = (2x-1)(\frac{y^2}{2} - \frac{y^3}{3}) + (2y-1)(\frac{x^2}{2} - \frac{x^3}{3}) + (\frac{x^2}{2} - \frac{x^3}{3})(\frac{y^2}{2} - \frac{y^3}{3}) \qquad in\ \Omega,$$

where $\Omega = \{(x,y) \in \Re^2 : 0 < x < 1, 0 < y < 1\}$ is the unit square with the boundary Γ. The boundary conditions are

$$p_1 = 0 \qquad on\ \Gamma_1 = \{(x,y) \in\ \Gamma : x = 0\},$$

$$p_1 = 0 \qquad on\ \Gamma_3 = \{(x,y) \in\ \Gamma : x = 1\},$$

$$p_2 = 0 \qquad on\ \Gamma_2 = \{(x,y) \in\ \Gamma : y = 0\},$$

$$p_2 = 0 \qquad on\ \Gamma_4 = \{(x,y) \in\ \Gamma : y = 1\}.$$

The exact solution should be

$$u = -(\frac{x^2}{2} - \frac{x^3}{3})(\frac{y^2}{2} - \frac{y^3}{3}),$$

$$p_1 = (x^2 - x)(\frac{y^2}{2} - \frac{y^3}{3}),$$

$$P_2 = (y^2 - y)(\frac{x^2}{2} - \frac{x^3}{3}).$$

Numerical experiments were carried out using bilinear elements on uniform meshes with $1/h = 4, 9, 20, 29$. We calculated the L_2 errors for u and $\mathbf{p}$:

$$e_u = \|u - u_h\|_0, \qquad e_p = (\|p_1 - p_{1h}\|_0^2 + \|p_2 - p_{2h}\|_0^2)^{\frac{1}{2}}.$$

The numerical results on the rates of convergence are given in Fig.1 (a) and (b). As expected, the rate of convergence of flux $\mathbf{p}$ for the usual LSFEM is $O(h)$, which is one-order lower than optimal. The rates of convergence are $O(h^2)$ for both the primal variable u and the dual variables $\mathbf{p}$ for the optimal LSFEM.

As the second example, we chose

$$f = (5\pi^2 + 1)cos(2\pi x)cos(\pi y)$$

with the same boundary conditions as in example 1. The exact solution is

$$u = -cos(2\pi x)cos(\pi y),$$

$$p_1 = 2\pi sin(2\pi x)cos(\pi y),$$

$$p_2 = \pi cos(2\pi x)sin(\pi y).$$

The numerical rates of convergence are shown in Fig.1. (c) and (d). It seems that the rate of convergence of $\mathbf{p}$ for the usual LSFEM is getting better when the mesh is refined. However, the error of $\mathbf{p}$ for the usual LSFEM is quite large.

Here we should mention that in all of our calculations, 2×2 Gaussian quadrature was used for finite element solutions, and 3×3 Gaussian quadrature was used for error evaluation. The LSFEM with numerical quadrature is equivalent to a weighted collocation least-squares method. We may use this idea to choose an appropriate number of Gaussian points. The usual LSFEM with one-point quadrature will not work, because it corresponds to solving a underdetermined algebraic system. The optimal LSFEM with

one-point quadrature works. However, the computed nodal values of **p** have oscillations; but the values of **p** at Gaussian points are correct.

References

1. P.P. Lynn and S.K. Arya, Finite elements formulated by the weighted discrete least squares method, Internat. J. Numer. Methods Engrg. 8 (1974) 71-79.
2. O.C. Zienkiewicz, The finite element method (McGraw-Hill, New York, 1983).
3. J.T. Oden and G.F. Carey, Finite elements: mathematical aspects, Vol.IV (Prentice-Hall, Englewood Cliffs, NJ, 1983).
4. P.G. Ciarlet, Basic error estimates for elliptic problems, in: P.G. Ciarlet and J.L.Lions, eds., Handbook of numerical analysis, Vol II, Finite element methods (Part 1) (North-Holland, Amsterdam, 1991).
5. J.E. Roberts and J.-M. Thomas, Mixed and hybrid methods, in: P.G. Ciarlet and J.L.Lions, eds., Handbook of numerical analysis, Vol II, Finite element methods (Part 1) (North-Holland, Amsterdam, 1991).
6. A.K. Aziz, R.B. Kellogg and A.B. Stephens, Least squares methods for elliptic systems, Math. Comp. 44 (1985) 53-70.
7. V. Girault and P.-A. Raviart, Finite element method for Navier-Stokes equations (Springer, Berlin, 1986).
8. P. Grisvard, Boundary value problems in non-smooth domains, Univ. of Maryland, Dept. of Math. Lecture Notes No. 19. (1985).
9. G.J. Fix and M.E. Rose, A comparative study of finite element and finite difference methods for Cauchy-Riemann type equations, SIAM J.Numer.Anal. 22 (1985) 250-260.
10. B.N. Jiang, Least-squares finite element methods with element-by-element solution including adaptive refinement, PhD dissertation, The University of Texas at Austin, 1986.

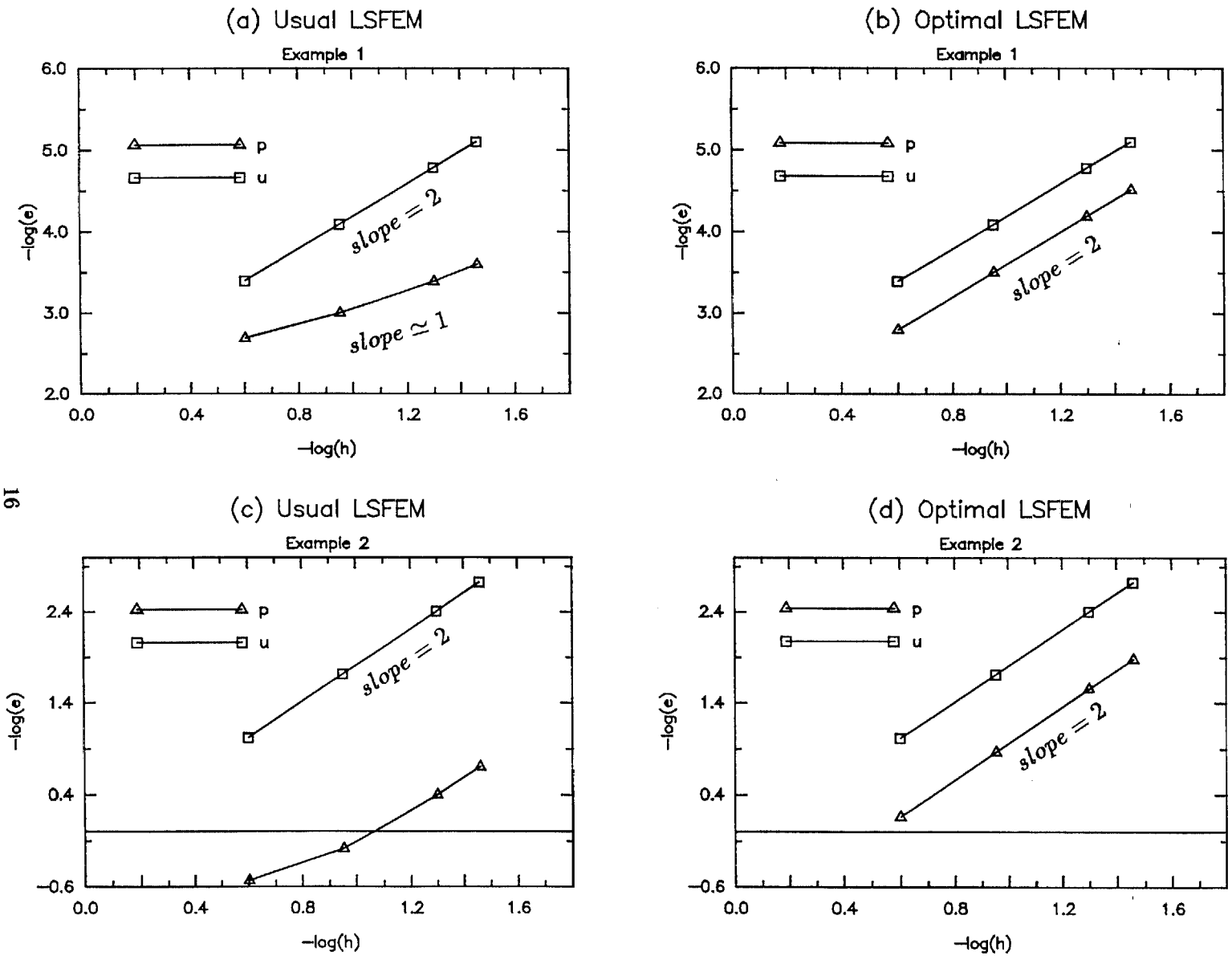


Fig. 1. Computed rates of convergence.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1991	Technical Memorandum

4. TITLE AND SUBTITLE

Optimal Least-Squares Finite Element Method for Elliptic Problems

5. FUNDING NUMBERS

WU-505-62-21

6. AUTHOR(S)

Bo-Nan Jiang and Louis A. Povinelli

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135-3191

8. PERFORMING ORGANIZATION REPORT NUMBER

E-6769

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, D.C. 20546-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA TM-105382
ICOMP-91-29

11. SUPPLEMENTARY NOTES

Bo-Nan Jiang, Institute for Computational Mechanics in Propulsion, Lewis Research Center (work funded under Space Act Agreement C-99066G). Louis A. Povinelli, NASA Lewis Research Center and Space Act Monitor, (216) 433-5818.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category 64

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this paper, we propose an optimal least-squares finite element method for 2D and 3D elliptic problems and discuss its advantages over the mixed Galerkin method and the usual least-squares finite element method. In the usual least-squares finite element method, the second-order equation $-\nabla \cdot (\nabla u) + u = f$ is recast as a first-order system $-\nabla \cdot \mathbf{p} + u = f$, $\nabla u - \mathbf{p} = 0$. Our error analysis and numerical experiments show that, in this usual least-squares finite element method, the rate of convergence for flux p is one-order lower than optimal. In order to get an optimal least-squares method, the irrotationality $\nabla \times \mathbf{p} = 0$ should be included in the first-order system.

14. SUBJECT TERMS

Finite element; Least-squares; First-order partial differential equation; Mixed method; Error analysis; Elliptic

15. NUMBER OF PAGES

18

16. PRICE CODE

A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	