

A Note on the Least–Squares Finite Elements for Mixed Elliptic Problems

Boško Jovanović¹, R.D. Lazarov², and Ivan Šestak³

¹ University of Belgrade, Faculty of Mathematics,
Studentski trg 16, P.O.B. 550, 11001 Belgrade, Yugoslavia
e-mail: bosko@matf.bg.ac.yu

² Texas A&M University, Department of Mathematics,
College Station, Texas 77843–3368, USA
e-mail: lazarov@math.tamu.edu

³ University of Belgrade, Technical Faculty, Bor,
Vojske Jugoslavije 12, P.O.B. 50, 19210 Bor, Yugoslavia

Abstract. A theoretical analysis of the mixed least–squares finite element approximations to Dirichlet problem for second–order linear elliptic equations with variable coefficients in bounded domains is presented. Three different least–squares functionals are introduced and the coercivity of the corresponding weak forms is proved. It is shown that the finite element approximation are stable and yield symmetric positive definite systems. The conditioning of the linear systems is discussed and error estimates for the approximate solutions are obtained.

1 Introduction

Many classical boundary value problems for second order linear elliptic equations can be transformed into the first order systems of equations, involving the gradient of solution as a new unknown. Such formulation of the problem is called mixed formulation (see [9]).

The unknown functions in the mixed formulations often have appropriate physical interpretation (temperature and flux, displacement and stress etc.) and give direct information about the considered quantities. For example, the vector variable (the flux) belongs to the space $H(\operatorname{div}; \Omega)$ and has a trace (in a weak sense) of the normal component on each surface and therefore the finite element space will consist of vector functions with continuous normal component at the inter-element boundaries. From numerical point of view, the nodal values of the solution gradient are obtained directly from the discrete problem, rather than by post-processing in the standard formulation.

Ritz–Galerkin formulation of the finite element method for mixed problems requires relatively weak smoothness of the input data, which in turn requires weak assumptions for the smoothness of finite element approximations. On the

* AMS Subject Classifications (1991): 65N30

** Key words: least–squares, mixed finite elements, error estimates

other hand, the finite dimensional spaces used in the Ritz–Galerkin method must satisfy so called LBB condition (Ladyzhenskaya–Babuška–Brezzi, see [20], [3], [7]) which substantially restrict the choice of feasible finite element spaces. Moreover, the solution is a stationary point of the saddle point functional, while the corresponding linear system is symmetric but indefinite. This seriously restrict the possibilities of a good choice of a solution algorithm.

Contrary to the Ritz–Galerkin method, the least–squares method for mixed problems is not subject to LBB condition, while the corresponding linear system is symmetric and positive definite. This approach has been applied to various second order problems [2], [10], [11], [14], [17], [18], [23], [24], [25], Stokes equations [5], [8], [13], [15], [16], and Navier–Stokes equations [4]. In this note we discuss three particular formulations of the least-squares finite element method for symmetric second order elliptic problems. Similarly to the mixed finite element method, we have a direct approximation of the vector variable. The main thrust of our paper is the second formulation (case II of the next section). In this case the finite element solution is either continuous or has continuous normal to the inter-element boundaries component. The main advantage of the proposed formulation is that the finite element spaces are not subject to the LBB condition and the obtained system is symmetric and positive definite.

Let Ω be a bounded domain in \mathbb{R}^N , $N = 2$ or 3 , with boundary $\Gamma = \partial\Omega$. For the sake of simplicity, in the sequel we suppose that Ω is a polygon in \mathbb{R}^2 , or polyhedron in \mathbb{R}^3 . Let us consider the Dirichlet boundary value problem

$$\begin{aligned} \mathcal{L}u &\equiv -\operatorname{div}(A \operatorname{grad} u) = f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma, \end{aligned} \quad (1)$$

where, as usual

$$\begin{aligned} \operatorname{grad} u &= \left(\frac{\partial u}{\partial x_1}, \dots, \frac{\partial u}{\partial x_N} \right)^T, \\ \operatorname{div} \mathbf{q} &= \operatorname{div}(q_1, \dots, q_N)^T = \frac{\partial q_1}{\partial x_1} + \dots + \frac{\partial q_N}{\partial x_N}, \end{aligned}$$

and $A = A^T = (a_{ij}(x))_{i,j=1}^N$, $x \in \bar{\Omega}$. We assume that the coefficients of the matrix A are bounded functions, $a_{ij} \in L^\infty(\Omega)$, while A is a positive definite matrix, i.e. there exist positive constants α_0 and α_1 such that

$$\alpha_0 \mathbf{y}^T \mathbf{y} \leq \mathbf{y}^T A \mathbf{y} \leq \alpha_1 \mathbf{y}^T \mathbf{y} \quad (2)$$

for all $\mathbf{y} \in \mathbb{R}^N$ and all $x \in \bar{\Omega}$.

Denoting $-A \operatorname{grad} u = \mathbf{p} = (p_1, \dots, p_N)^T$ the Dirichlet problem (1) reduces to a first–order system for u and \mathbf{p} :

$$\begin{aligned} \mathbf{p} + A \operatorname{grad} u &= 0 && \text{in } \Omega, \\ \operatorname{div} \mathbf{p} - f &= 0 && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma. \end{aligned} \quad (3)$$

Note that in various engineering applications the vector variable \mathbf{p} represents a physical quantity (for example, heat or mass flux, phase velocity, current, etc.) which might be of main interest in the particular application.

Remark. In various applications the right hand side f might be considered as linear bounded functional on $H_0^1(\Omega)$, i.e. an element in the space $H^{-1}(\Omega) \equiv (H_0^1(\Omega))'$. Then, we can assume that it has the form (see [27]):

$$f = f_0 + \operatorname{div} \mathbf{f}, \quad \text{where} \quad f_0, f_1, \dots, f_N \in L^2(\Omega). \quad (4)$$

In this case instead of (3) one can introduce a slightly different splitting of the elliptic equation into a system of first order, namely, $\mathbf{p} + A \operatorname{grad} u = \mathbf{f}$, and $\operatorname{div} \mathbf{p} - f_0 = 0$. This splitting will give better balanced system which in case II will be easier to analyze.

Let V and \mathbf{W} be two Banach spaces such that $u \in V$, $\mathbf{p} \in \mathbf{W}$. Based on the relations (3), for $v \in V$ and $\mathbf{q} \in \mathbf{W}$ we introduce the functional

$$J(v, \mathbf{q}) = \|\operatorname{div} \mathbf{q} - f\|_{(1)}^2 + \|\mathbf{q} + A \operatorname{grad} v\|_{(2)}^2, \quad (5)$$

where the norms $\|\cdot\|_{(1)}$ and $\|\cdot\|_{(2)}$ will be defined later. Thus, the system (3) can be replaced by the minimization problem for the least-squares functional (5):

$$\begin{aligned} \text{find } (u, \mathbf{p}) \in V \times \mathbf{W} \quad \text{such that} \\ J(u, \mathbf{p}) = \inf_{(v, \mathbf{q}) \in V \times \mathbf{W}} J(v, \mathbf{q}). \end{aligned} \quad (6)$$

Obviously, the solution of the original problem (1) will make this functional zero, if the corresponding norms in (5) are finite. A particular choice of the norms in the quadratic functional (5) will define a particular least-squares method. We shall consider three different settings which will lead to three different least-squares methods.

The first choice of norms is the most popular and has been used and studied by many authors (see, e.g. [2], [10], [11], [17], [18], [23], [24], [25]). One simply takes $\|\cdot\|_{(1)}$ and $\|\cdot\|_{(2)}$ to be the L^2 -norms in the corresponding Banach spaces of scalar and vector functions. This formulation, included in our paper as case I, is natural and leads to a standard finite element method. A weak point of this approach is the fact that the error estimates are not optimal with respect to the required regularity of the solution (see Theorem 9). For example, the convergence rate in H^1 -norm for u is $O(h)$ for solutions in $H^3(\Omega)$. Secondly, there is no efficient method for solving the resulting system of linear equations.

Analyzing this fact, Bramble et al. [6] have come to a least-squares formulation in which the norms in (5) are chosen in more natural way. Namely, $\|\cdot\|_{(1)}$ is the H^{-1} -norm in V and $\|\cdot\|_{(2)}$ is the L^2 -norm in \mathbf{W} . This formulation has several advantages compared with the previous one. First, the least-squares functional is well defined for $f \in H^{-1}(\Omega)$. Second, the finite element method for this formulation is optimal with respect to the regularity of the solution. Finally, the finite element system is easily preconditioned. However, a major difficulty had to be overcome in [6]: how to replace the H^{-1} -norm in the finite element formulation with a discrete and efficiently computable norm, which is equivalent to H^{-1} for function in the finite element space. This problem has been successfully

resolved in [6] using the recent advances in the preconditioning techniques for elliptic problems.

In this note we propose a simpler reformulation of the H^{-1} -least squares of Bramble et al. [6] for the case of symmetric second order elliptic problems. This reformulation (presented as case II) leads to a least-squares finite element method of optimal with respect to the regularity and the mesh-size h parameter error estimates.

Finally, for completeness, we also consider a third choice (case III) of the norms in (5): $\|\cdot\|_{(1)}$ is the standard L^2 -norm in V and $\|\cdot\|_{(2)}$ is the $H(\operatorname{div}; \Omega)$ -norm in \mathbf{W} . This leads to a C^1 -finite element method with optimal error estimates, but with condition number $O(h^{-4})$ (here h is the mesh-size parameter).

In the sequel, by $(\cdot, \cdot)_{s, \Omega}$, $|\cdot|_{s, \Omega}$ and $\|\cdot\|_{s, \Omega}$ we shall denote respectively the inner product, semi-norm and norm of the Sobolev space $H^s(\Omega)$ (see [1]), while by $(\cdot, \cdot)_{s, \Omega; N}$, $|\cdot|_{s, \Omega; N}$ and $\|\cdot\|_{s, \Omega; N}$ we will denote the inner product, semi-norm and norm of the space $(H^s(\Omega))^N$.

By C and C_i we shall denote positive generic constants, which can take different values in different formulas.

2 Existence and Uniqueness of the Solution

Let us consider three cases, according to different choice of norms in (5).

Case I. We set

$$\begin{aligned} \|v\|_{(1)} &= \|v\|_{0, \Omega}, & \|\mathbf{q}\|_{(2)} &= \|A^{-1/2}\mathbf{q}\|_{0, \Omega; N}, \\ V = V_1 &= H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma\}, & \text{and} & \\ \mathbf{W} = \mathbf{W}_1 &= H(\operatorname{div}; \Omega) = \{\mathbf{q} \in (L^2(\Omega))^N : \operatorname{div} \mathbf{q} \in L^2(\Omega)\}. \end{aligned} \quad (7)$$

The norm in \mathbf{W}_1 is defined by

$$\|\mathbf{q}\|_{H(\operatorname{div}; \Omega)} = (\|\mathbf{q}\|_{0, \Omega; N}^2 + \|\operatorname{div} \mathbf{q}\|_{0, \Omega}^2)^{1/2}.$$

Taking variations of the functional (5) with respect to v and \mathbf{q} we obtain the following weak statement of the problem (3):

$$\begin{aligned} \text{find } (u, \mathbf{p}) &\in V_1 \times \mathbf{W}_1 \quad \text{such that} \\ a_1(u, \mathbf{p}; v, \mathbf{q}) &= l_1(v, \mathbf{q}) \quad \text{for all } (v, \mathbf{q}) \in V_1 \times \mathbf{W}_1, \end{aligned} \quad (8)$$

where

$$a_1(u, \mathbf{p}; v, \mathbf{q}) = (\operatorname{div} \mathbf{p}, \operatorname{div} \mathbf{q})_{0, \Omega} + (A^{-1} \mathbf{p} + \operatorname{grad} u, \mathbf{q} + A \operatorname{grad} v)_{0, \Omega; N}, \quad (9)$$

$$l_1(v, \mathbf{q}) = (f, \operatorname{div} \mathbf{q})_{0, \Omega}. \quad (10)$$

The following assertions hold:

Theorem 1. —(see [24])—. *Bilinear form (9) is coercive on $V_1 \times \mathbf{W}_1$, i.e. there exists a constant $C > 0$ such that for all $(v, \mathbf{q}) \in V_1 \times \mathbf{W}_1$*

$$a_1(v, \mathbf{q}; v, \mathbf{q}) \geq C (\|v\|_{1, \Omega}^2 + \|\mathbf{q}\|_{H(\text{div}; \Omega)}^2). \quad (11)$$

Theorem 2. —(see [24])—. *Let $f \in L^2(\Omega)$. Then the problem (8) has a unique solution $(u, \mathbf{p}) \in V_1 \times \mathbf{W}_1$ and the following estimate holds*

$$\|u\|_{1, \Omega} + \|\mathbf{p}\|_{H(\text{div}; \Omega)} \leq C \|f\|_{0, \Omega}.$$

Case II. Let

$$\begin{aligned} \|v\|_{(1)} &= \|v\|_{-1, \Omega} = (\mathcal{L}^{-1} v, v)_{0, \Omega}^{1/2}, & \|\mathbf{q}\|_{(2)} &= \|A^{-1/2} \mathbf{q}\|_{0, \Omega; N}, \\ V = V_2 &= H_0^1(\Omega) & \text{and} & \quad \mathbf{W} = \mathbf{W}_2 = (L^2(\Omega))^N. \end{aligned} \quad (12)$$

Under previous assumptions on the matrix A , there exists the inverse operator $\mathcal{L}^{-1} : H^{-1}(\Omega) \equiv (H_0^1(\Omega))' \rightarrow H_0^1(\Omega)$ (see [19]), so the norm $\|v\|_{(1)}$ is well defined. Further

$$\begin{aligned} \|\text{div } \mathbf{p} - f\|_{-1, \Omega}^2 &= \|\mathcal{L} u - f\|_{-1, \Omega}^2 = (\mathcal{L}^{-1} (\mathcal{L} u - f), \mathcal{L} u - f)_{0, \Omega} \\ &= (\mathcal{L} u, u)_{0, \Omega} - 2(u, f)_{0, \Omega} + (\mathcal{L}^{-1} f, f)_{0, \Omega} \\ &= (A \text{ grad } u, \text{ grad } u)_{0, \Omega; N} - 2(u, f)_{0, \Omega} + \|f\|_{-1, \Omega}^2 \\ &= (A^{-1} \mathbf{p}, \mathbf{p})_{0, \Omega; N} - 2(u, f)_{0, \Omega} + \|f\|_{-1, \Omega}^2. \end{aligned}$$

In such a manner, the functional (5) reduces to

$$\begin{aligned} J(v, \mathbf{q}) &= (A^{-1} \mathbf{q}, \mathbf{q})_{0, \Omega; N} + (\mathbf{q} + A \text{ grad } v, A^{-1} \mathbf{q} + \text{ grad } v)_{0, \Omega; N} \\ &\quad - 2(v, f)_{0, \Omega} + \|f\|_{-1, \Omega}^2. \end{aligned} \quad (13)$$

Above we have used the definition of the vector variable through the gradient of the solution u of the original problem (1). However, at this point there is no guarantee that the pair (u, \mathbf{p}) will minimize the quadratic functional (13). Nevertheless, the minimization of the quadratic functional (13), leads to the following weak formulation:

$$\begin{aligned} \text{find } (w, \mathbf{r}) &\in V_2 \times \mathbf{W}_2 \quad \text{such that} \\ a_2(w, \mathbf{r}; v, \mathbf{q}) &= l_2(v, \mathbf{q}) \quad \text{for all } (v, \mathbf{q}) \in V_2 \times \mathbf{W}_2, \end{aligned} \quad (14)$$

where

$$a_2(w, \mathbf{r}; v, \mathbf{q}) = (A^{-1} \mathbf{r}, \mathbf{q})_{0, \Omega; N} + (\mathbf{r} + A \text{ grad } w, A^{-1} \mathbf{q} + \text{ grad } v)_{0, \Omega; N}, \quad (15)$$

$$l_2(v, \mathbf{q}) = (v, f)_{0, \Omega}. \quad (16)$$

Now we can study the relationship between the pair (w, \mathbf{r}) which satisfies the integral identity (14) and the solution (u, \mathbf{p}) of the original problem (3). In order to compare these two pairs we have to find the Euler equation for the problem

(14). Taking first $(0, \mathbf{q})$ and next $(v, \mathbf{0})$ for $v \in V_2$ and $\mathbf{q} \in \mathbf{W}_2$ we get the following two integral identities:

$$(\mathbf{r} + A \operatorname{grad} w, \operatorname{grad} v)_{0, \Omega; N} - (f, v)_{0, \Omega} = 0, \text{ for all } v \in V_2$$

and

$$(2\mathbf{r} + A \operatorname{grad} w, A^{-1} \mathbf{q})_{0, \Omega; N} = 0, \text{ for all } \mathbf{q} \in \mathbf{W}_2.$$

By simple manipulations one verifies that $w = 2u$ and $\mathbf{r} = \mathbf{p}$. From this relation one can easily recover the solution of the problem (3) from the solution of the problem (14). Similar analysis can be done in the case when the right-hand side f includes a term $\operatorname{div} \mathbf{f}$.

Similarly to the previous case one shows that the bilinear form (15) is coercive and the problem (14) has a unique solution.

Theorem 3. *Bilinear form (15) is coercive on $V_2 \times \mathbf{W}_2$, i.e. there exists a constant $C > 0$ such that for all $(v, \mathbf{q}) \in V_2 \times \mathbf{W}_2$*

$$a_2(v, \mathbf{q}; v, \mathbf{q}) \geq C (\|v\|_{1, \Omega}^2 + \|\mathbf{q}\|_{0, \Omega; N}^2). \quad (17)$$

Proof. From (15), it follows

$$\begin{aligned} a_2(v, \mathbf{q}; v, \mathbf{q}) &= 2(A^{-1} \mathbf{q}, \mathbf{q})_{0, \Omega; N} + 2(\mathbf{q}, \operatorname{grad} v)_{0, \Omega; N} + (A \operatorname{grad} v, \operatorname{grad} v)_{0, \Omega; N} \\ &\geq \frac{1}{3}(A^{-1} \mathbf{q}, \mathbf{q})_{0, \Omega; N} + \frac{1}{3}(A \operatorname{grad} v, \operatorname{grad} v)_{0, \Omega; N} \\ &\geq \frac{1}{3\alpha_1} \|\mathbf{q}\|_{0, \Omega; N}^2 + \frac{\alpha_0}{3} \|\operatorname{grad} v\|_{0, \Omega; N}^2. \end{aligned}$$

Using (2) and Poincaré–Friedrichs inequality (see [12])

$$\|v\|_{0, \Omega}^2 \leq C \|\operatorname{grad} v\|_{0, \Omega; N}^2$$

we easily get (17). We note that the constant in (17) depends on the domain Ω and linearly on the ratio α_1/α_0 .

Theorem 4. *If f is an element in $H_0^{-1}(\Omega)$ then the problem (14) has a unique solution $(u, \mathbf{p}) \in V_2 \times \mathbf{W}_2$ and*

$$\|u\|_{1, \Omega} + \|\mathbf{p}\|_{0, \Omega; N} \leq C \|f\|_{-1, \Omega}.$$

Proof. We define the norm in $V_2 \times \mathbf{W}_2$ by

$$\|v\|_{1, \Omega} + \|\mathbf{q}\|_{0, \Omega; N}, \quad v \in V_2, \quad \mathbf{q} \in \mathbf{W}_2.$$

From (17) and (2) it follows that the bilinear form (15) is coercive and bounded on $V_2 \times \mathbf{W}_2$. Obviously, the linear form (16) is bounded on $V_2 \times \mathbf{W}_2$ and the result follows from the Lax–Milgram lemma (see [22], [12]). \square

Case III. Let

$$\begin{aligned} \|v\|_{(1)} &= \|v\|_{0,\Omega}, & \|\mathbf{q}\|_{(2)} &= (\|A^{-1/2}\mathbf{q}\|_{0,\Omega;N}^2 + \|\operatorname{div}\mathbf{q}\|_{0,\Omega}^2)^{1/2}, \\ V = V_3 &= H^2(\Omega) \cap H_0^1(\Omega), & \text{and} & \quad \mathbf{W} = \mathbf{W}_3 = H(\operatorname{div}; \Omega). \end{aligned} \quad (18)$$

Then the functional (5) reduces to

$$\begin{aligned} J(v, \mathbf{q}) &= \|\operatorname{div}\mathbf{q} - f\|_{0,\Omega}^2 + \|A^{-1/2}(\mathbf{q} + A \operatorname{grad} v)\|_{0,\Omega;N}^2 \\ &\quad + \|\operatorname{div}(\mathbf{q} + A \operatorname{grad} v)\|_{0,\Omega}^2. \end{aligned} \quad (19)$$

The weak formulation of the problem (3) is

$$\begin{aligned} \text{find } (u, \mathbf{p}) &\in V_3 \times \mathbf{W}_3 \quad \text{such that} \\ a_3(u, \mathbf{p}; v, \mathbf{q}) &= l_3(v, \mathbf{q}) \quad \text{for all } (v, \mathbf{q}) \in V_3 \times \mathbf{W}_3, \end{aligned} \quad (20)$$

where

$$\begin{aligned} a_3(u, \mathbf{p}; v, \mathbf{q}) &= (\operatorname{div}\mathbf{p}, \operatorname{div}\mathbf{q})_{0,\Omega} + (\mathbf{p} + A \operatorname{grad} u, A^{-1}\mathbf{q} + \operatorname{grad} v)_{0,\Omega;N} \\ &\quad + (\operatorname{div}\mathbf{p} + \operatorname{div}(A \operatorname{grad} u), \operatorname{div}\mathbf{q} + \operatorname{div}(A \operatorname{grad} v))_{0,\Omega}, \end{aligned} \quad (21)$$

$$l_3(v, \mathbf{q}) = (f, \operatorname{div}\mathbf{q})_{0,\Omega}. \quad (22)$$

The following assertions hold:

Theorem 5. *Assume that $\frac{\partial a_{ij}}{\partial x_k} \in L^\infty(\Omega)$, $i, j, k = 1, \dots, N$ and the domain Ω is a convex polygon (polyhedron). Then there exists a constant $C > 0$ such that for all $(v, \mathbf{q}) \in V_3 \times \mathbf{W}_3$*

$$a_3(v, \mathbf{q}; v, \mathbf{q}) \geq C (\|v\|_{2,\Omega}^2 + \|\mathbf{q}\|_{H(\operatorname{div}; \Omega)}^2). \quad (23)$$

Proof. According to Theorem 1, we have for (21)

$$a_3(v, \mathbf{q}; v, \mathbf{q}) \geq C (\|v\|_{1,\Omega}^2 + \|\mathbf{q}\|_{H(\operatorname{div}; \Omega)}^2) + \|\operatorname{div}\mathbf{q} - \mathcal{L}v\|_{0,\Omega}^2. \quad (24)$$

From here it follows

$$\begin{aligned} a_3(v, \mathbf{q}; v, \mathbf{q}) &\geq C (\|v\|_{1,\Omega}^2 + \|\mathbf{q}\|_{0,\Omega;N}^2 + \|\operatorname{div}\mathbf{q}\|_{0,\Omega}^2) \\ &\quad + \|\operatorname{div}\mathbf{q}\|_{0,\Omega}^2 - 2(\operatorname{div}\mathbf{q}, \mathcal{L}v)_{0,\Omega} + \|\mathcal{L}v\|_{0,\Omega}^2 \\ &\geq \frac{C}{C+1} (\|v\|_{1,\Omega}^2 + \|\mathbf{q}\|_{0,\Omega;N}^2 + \|\mathcal{L}v\|_{0,\Omega}^2). \end{aligned} \quad (25)$$

Since the partial derivatives of the coefficients a_{ij} , $i, j = 1, \dots, N$, are bounded functions and the domain is a convex polygon (polyhedron), the so called "second fundamental inequality" (see [21]) is valid:

$$\|v\|_{2,\Omega}^2 \leq C_1 \|\mathcal{L}v\|_{0,\Omega}^2 + C_2 \|v\|_{1,\Omega}^2. \quad (26)$$

Combining (24), (25) and (26) we obtain (23). \square

Theorem 6. *Let the assumptions of the Theorem 5 hold and $f \in L^2(\Omega)$. Then the problem (20) has a unique solution $(u, \mathbf{p}) \in V_3 \times \mathbf{W}_3$ and*

$$\|u\|_{2, \Omega} + \|\mathbf{p}\|_{H(\text{div}; \Omega)} \leq C \|f\|_{0, \Omega}.$$

Proof. Let us define the norm in the space $V_3 \times \mathbf{W}_3$ by

$$\|v\|_{2, \Omega} + \|\mathbf{q}\|_{H(\text{div}; \Omega)}, \quad v \in V_3, \quad \mathbf{q} \in \mathbf{W}_3.$$

From (23) and (2) it follows that the bilinear form (21) is coercive and bounded on $V_3 \times \mathbf{W}_3$. For $f \in L^2(\Omega)$ the linear form (22) is bounded on $V_3 \times \mathbf{W}_3$ and by Lax–Milgram lemma (see [22], [12]) the result of the theorem easily follows. \square

3 Finite Element Approximation

Let \mathcal{T}_h be a partition of the domain Ω into finite elements, $\Omega = \cup_{K \in \mathcal{T}_h} K$, and $h = \max\{\text{diam}(K) : K \in \mathcal{T}_h\}$. Let $V_{m, h}$ and $\mathbf{W}_{m, h}$ be finite-dimensional subspaces of V_m and \mathbf{W}_m , $m = 1, 2, 3$, which have the following approximation properties:

$$\inf_{v_h \in V_{m, h}} \|v - v_h\|_{1, \Omega} \leq C h^k \|v\|_{k+1, \Omega}, \quad m = 1, 2 \quad (27)$$

$$\inf_{v_h \in V_{m, h}} \|v - v_h\|_{2, \Omega} \leq C h^{k-1} \|v\|_{k+1, \Omega}, \quad m = 3 \quad (28)$$

$$\inf_{\mathbf{q}_h \in \mathbf{W}_{m, h}} \|\mathbf{q} - \mathbf{q}_h\|_{0, \Omega; N} \leq C h^{l+1} \|\mathbf{q}\|_{l+1, \Omega; N}, \quad m = 2 \quad (29)$$

$$\inf_{\mathbf{q}_h \in \mathbf{W}_{m, h}} \|\mathbf{q} - \mathbf{q}_h\|_{H(\text{div}; \Omega)} \leq C h^l \|\mathbf{q}\|_{l+1, \Omega; N}, \quad m = 1, 3 \quad (30)$$

where $k > 0$ and $l \geq 0$ are integers. Standard choices for $V_{m, h}$ and $\mathbf{W}_{m, h}$ are spaces of continuous piecewise polynomial functions, i.e.,

$$V_{m, h} = \{v_h \in V_m : v_h|_K \in \mathcal{P}_k(K), \quad \forall K \in \mathcal{T}_h\},$$

$$\mathbf{W}_{m, h} = \{\mathbf{q}_h \in \mathbf{W}_m : q_{h, i}|_K \in \mathcal{P}_l(K), \quad \forall K \in \mathcal{T}_h; \quad i = 1, \dots, N\}.$$

Here $\mathcal{P}_s(K)$ is the space of polynomials of degree not greater than s on K . In case II for the vector variable one can also use the Raviart-Thomas mixed finite elements [26] which satisfy (29) with $l \geq 0$.

In the following we will assume that the domain Ω is covered by finite elements exactly, and that integration is exact. This assumption will eliminate the difficulties related to numerical integration and approximation of the domain.

The finite element approximation of the problems (8), (14) and (20) is:

$$\begin{aligned} \text{find } (u_h, \mathbf{p}_h) \in V_{m, h} \times \mathbf{W}_{m, h} \quad \text{such that} \\ a_m(u_h, \mathbf{p}_h; v_h, \mathbf{q}_h) = l_m(v_h, \mathbf{q}_h) \quad \text{for all } (v_h, \mathbf{q}_h) \in V_{m, h} \times \mathbf{W}_{m, h}, \end{aligned} \quad (31)$$

where m takes respectively the values 1, 2 and 3. From Theorems 1, 3 and 5 follows the uniqueness of the solution of the problem (31) for $m = 1, 2$ and 3. Moreover, in all three cases the error satisfies the orthogonality condition

$$\begin{aligned} a_m(u - u_h, \mathbf{p} - \mathbf{p}_h; v_h, \mathbf{q}_h) &= 0 \\ \text{for all } (v_h, \mathbf{p}_h) &\in V_{m,h} \times \mathbf{W}_{m,h}. \end{aligned} \quad (32)$$

Now we estimate the condition number of the linear systems obtained in (31). Suppose that the finite element partition is quasi-regular (see [12]), i.e. there exists a constant $\delta > 0$ such that

$$\delta h \leq \text{diam}(K) \leq h \quad (33)$$

for all $K \in \mathcal{T}_h$ and all sufficiently small h .

Let $\varphi_{m,1}, \dots, \varphi_{m,L_m}$ and $\Psi_{m,1}, \dots, \Psi_{m,M_m}$ be sets of nodal basis functions in $V_{m,h}$ and $\mathbf{W}_{m,h}$, $m = 1, 2, 3$. We suppose that there exist positive independent of the grid-size parameter h constants $A_{m,1}, A_{m,2}, B_{m,1}$ and $B_{m,2}$ such that for all real vectors $(\alpha_{m,1}, \dots, \alpha_{m,L_m})$ and $(\beta_{m,1}, \dots, \beta_{m,M_m})$

$$A_{m,1} h^N \sum_{i=1}^{L_m} \alpha_{m,i}^2 \leq \left\| \sum_{i=1}^{L_m} \alpha_{m,i} \varphi_{m,i} \right\|_{0,\Omega}^2 \leq A_{m,2} h^N \sum_{i=1}^{L_m} \alpha_{m,i}^2, \quad (34)$$

$$B_{m,1} h^N \sum_{j=1}^{M_m} \beta_{m,j}^2 \leq \left\| \sum_{j=1}^{M_m} \beta_{m,j} \Psi_{m,j} \right\|_{0,\Omega;N}^2 \leq B_{m,2} h^N \sum_{j=1}^{M_m} \beta_{m,j}^2. \quad (35)$$

Notice that such inequalities are fulfilled for all well-known finite element spaces for quasi-uniform partitions.

The following assertions hold:

Theorem 7. (see [24])— *If (34) and (35) are satisfied the condition number of linear system (31) for $m = 1$ is $O(h^{-2})$.*

Theorem 8. *If (34) and (35) are satisfied the condition number of linear system (31) is $O(h^{-2})$ for $m = 2$, and $O(h^{-4})$ for $m = 3$.*

Proof. From coercivity and boundness of the bilinear form (15) follows

$$\begin{aligned} C_1 (\|v_h\|_{1,\Omega}^2 + \|\mathbf{q}_h\|_{0,\Omega;N}^2) &\leq a_2(v_h, \mathbf{q}_h; v_h, \mathbf{q}_h) \\ &\leq C_2 (\|v_h\|_{1,\Omega}^2 + \|\mathbf{q}_h\|_{0,\Omega;N}^2). \end{aligned} \quad (36)$$

From (33) and the inverse estimate (see [12])

$$|v_h|_{1,K} \leq C h^{-1} \|v_h\|_{0,K}, \quad \forall v_h \in V_{2,h} \subset H^1(\Omega), \quad \forall K \in \mathcal{T}_h \quad (37)$$

follows that

$$\|v_h\|_{0,\Omega} \leq \|v_h\|_{1,\Omega} \leq C h^{-1} \|v_h\|_{0,\Omega}.$$

Substituting in (36),

$$\begin{aligned} C_1 (\|v_h\|_{0,\Omega}^2 + \|\mathbf{q}_h\|_{0,\Omega;N}^2) &\leq a_2(v_h, \mathbf{q}_h; v_h, \mathbf{q}_h) \\ &\leq C_3 h^{-2} \|v_h\|_{0,\Omega}^2 + C_2 \|\mathbf{q}_h\|_{0,\Omega;N}^2 \leq C_4 h^{-2} (\|v_h\|_{0,\Omega}^2 + \|\mathbf{q}_h\|_{0,\Omega;N}^2). \end{aligned} \quad (38)$$

Since the system (31) is symmetric, setting in (38) $v_h = \sum_{i=1}^{L_2} \alpha_{2,i} \varphi_{2,i}$ and $\mathbf{q}_h = \sum_{j=1}^{M_2} \beta_{2,j} \Psi_{2,j}$ and using relations (33) and (35) we obtain the first part of the assertion.

For $m = 3$, instead of (36) we have

$$\begin{aligned} C_1 (\|v_h\|_{2,\Omega}^2 + \|\mathbf{q}_h\|_{H(\text{div};\Omega)}^2) &\leq a_3(v_h, \mathbf{q}_h; v_h, \mathbf{q}_h) \\ &\leq C_2 (\|v_h\|_{2,\Omega}^2 + \|\mathbf{q}_h\|_{H(\text{div};\Omega)}^2), \end{aligned}$$

where from, using (33) and (37), one obtains

$$\begin{aligned} C_1 (\|v_h\|_{0,\Omega}^2 + \|\mathbf{q}_h\|_{0,\Omega;N}^2) &\leq a_3(v_h, \mathbf{q}_h; v_h, \mathbf{q}_h) \\ &\leq C_5 h^{-4} (\|v_h\|_{0,\Omega}^2 + \|\mathbf{q}_h\|_{0,\Omega;N}^2). \end{aligned}$$

From here, in the same manner as in the previous case, we obtain the second part of the assertion. \square

4 Error Estimates

The error estimates are obtained in a standard way using the coercivity and the boundness of the bilinear forms (9), (15) and (21) and the orthogonality condition (32):

$$\begin{aligned} C (\|u - u_h\|_{V_m}^2 + \|\mathbf{p} - \mathbf{p}_h\|_{\mathbf{W}_m}^2) &\leq a_m(u - u_h, \mathbf{p} - \mathbf{p}_h; u - u_h, \mathbf{p} - \mathbf{p}_h) \\ &= a_m(u - u_h, \mathbf{p} - \mathbf{p}_h; u - v_h, \mathbf{p} - \mathbf{q}_h) \\ &\leq C_1 (\|u - u_h\|_{V_m}^2 + \|\mathbf{p} - \mathbf{p}_h\|_{\mathbf{W}_m}^2)^{1/2} (\|u - v_h\|_{V_m}^2 + \|\mathbf{p} - \mathbf{q}_h\|_{\mathbf{W}_m}^2)^{1/2}, \end{aligned} \quad (39)$$

where v_h and \mathbf{q}_h are arbitrary elements in V_m and \mathbf{W}_m and $\|\cdot\|_{V_m}$ and $\|\cdot\|_{\mathbf{W}_m}$ denote norms in spaces V_m and \mathbf{W}_m , defined by (7), (12), or (18), respectively.

From (39), for $m = 1$, using (27) and (30), one gets

$$\begin{aligned} \|u - u_h\|_{1,\Omega} + \|\mathbf{p} - \mathbf{p}_h\|_{H(\text{div};\Omega)} &\leq C (h^k \|u\|_{k+1,\Omega} + h^l \|\mathbf{p}\|_{l+1,\Omega;N}) \\ &\leq C h^s (\|u\|_{k+1,\Omega} + \|\mathbf{p}\|_{l+1,\Omega;N}), \end{aligned}$$

where $s = \min\{k, l\}$. In such a manner, in the case I one obtains the optimal convergence rate estimate when $k = l = s$.

We summarize this result in the following theorem:

Theorem 9. —(see [24])—. *The solution of the problem (31) for $m = 1$ and $l = k$ satisfies the error estimate*

$$\|u - u_h\|_{1,\Omega} + \|\mathbf{p} - \mathbf{p}_h\|_{H(\text{div};\Omega)} \leq C h^k (\|u\|_{k+1,\Omega} + \|\mathbf{p}\|_{k+1,\Omega;N}).$$

Now, let us estimate the error for the cases II and III, i.e. $m = 2, 3$. First, for case II from (39), (12), (27) – (30) we obtain

$$\begin{aligned} \|u - u_h\|_{1, \Omega} + \|\mathbf{p} - \mathbf{p}_h\|_{0, \Omega; N} &\leq C (h^k \|u\|_{k+1, \Omega} + h^{l+1} \|\mathbf{p}\|_{l+1, \Omega; N}) \\ &\leq C h^s (\|u\|_{k+1, \Omega} + \|\mathbf{p}\|_{l+1, \Omega; N}), \end{aligned}$$

where $s = \min \{k, l + 1\}$.

Finally, for case III from (39), (18), (28) and (30) we obtain

$$\begin{aligned} \|u - u_h\|_{2, \Omega} + \|\mathbf{p} - \mathbf{p}_h\|_{H(\text{div}; \Omega)} &\leq C (h^{k-1} \|u\|_{k+1, \Omega} + h^l \|\mathbf{p}\|_{l+1, \Omega; N}) \\ &\leq C h^{s-1} (\|u\|_{k+1, \Omega} + \|\mathbf{p}\|_{l+1, \Omega; N}), \end{aligned}$$

where $s = \min \{k, l + 1\}$. Thus, one obtains optimal with respect to the discretization parameter h order error estimates for both cases II and III when $k = l + 1 = s$. We summarize these error estimates in the following assertion:

Theorem 10. *The finite element solutions of the problems (31) for $m = 2$ and 3 and $l = k - 1$ satisfy respectively the error estimates*

$$\begin{aligned} \|u - u_h\|_{1, \Omega} + \|\mathbf{p} - \mathbf{p}_h\|_{0, \Omega; N} &\leq C h^k (\|u\|_{k+1, \Omega} + \|\mathbf{p}\|_{k, \Omega; N}), & m = 2, \\ \|u - u_h\|_{2, \Omega} + \|\mathbf{p} - \mathbf{p}_h\|_{H(\text{div}; \Omega)} &\leq C h^{k-1} (\|u\|_{k+1, \Omega} + \|\mathbf{p}\|_{k, \Omega; N}), & m = 3. \end{aligned}$$

Note that the error estimate for case II (i.e. $m = 2$) is optimal also with respect to the regularity of the solution u . For example, the first estimate from Theorem 10 holds true for $k = 1$ and $l = 0$, i.e. for piece-wise linear elements for the scalar function u and Piece-wise constant functions for the vector functions \mathbf{p} . This estimate is optimal with respect to both the regularity of the solution and the order of the step-size parameter h . However, this is not an interesting application of the proposed methods, since it coincides with the standard Galerkin method with piece-wise linear functions. We can achieve continuity of the pressure gradient by choosing piece-wise quadratic elements for u and piece-wise linear elements for \mathbf{p} . An obvious choice is piece-wise linear elements for both the u and \mathbf{p} . However, in this case the error estimates will be not optimal with respect to the order of approximation, i.e. the finite element solution \mathbf{p}_h converges slower than the best approximation of \mathbf{p} in the finite element space. An optimal error we can get by using Raviart-Thomas elements [26]. The gain in this case is symmetric and positive definite matrix of the finite element system.

Remark. Using interpolation one can easily obtain estimates of order $O(h^\alpha)$ for all $\alpha \in [0, k]$ in the case II (i.e. $m = 2$) and $\alpha \in [0, k - 1]$ in the case III.

5 Numerical Experiments

Below we present some numerical results for the least-squares method described in case II. We have considered Poisson equation in a unit square (i.e. $A = I$

and $\Omega = (0, 1)^2$) with homogeneous Dirichlet boundary data and known smooth solutions. Namely, we consider:

Example 1: $f = 8x(1 - x) + 8y(1 - y)$ and $u = 4x(1 - x)y(1 - y)$.

Example 2: $f = 2\pi^2 \sin(\pi x)\sin(\pi y)$ and $u = \sin(\pi x)\sin(\pi y)$.

In Tables 1 and 2 we report the computational results for Examples 1 and 2, respectively. The l_2 -error for u is computed as $(\sum(u(P) - u_h(P))^2)^{1/2}h$, where the summation is over all nodes P of the mesh and is divided by the maximum value of the solution over the domain Ω . In a similar way the max -error is computed as $\max|u(P) - u_h(P)|$, where the maximum is over all nodes P of the mesh and divided by the maximum value of the solution over the domain Ω . In a similar manner we define the errors for the vector field \mathbf{p} .

Table 1. Error of the least-squares solution (case II) for Example 1

	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$	$h = 1/160$	\approx order
l_2 -error for u	2.10e-2	5.60e-3	1.44e-3	3.64e-4	9.15e-5	2
max -error for u	5.69e-2	1.55e-2	3.97e-3	1.00e-3	2.52e-4	2
l_2 -error for \mathbf{p}	3.22e-2	1.18e-2	4.24e-3	1.50e-3	5.32e-4	1.5
max -error for \mathbf{p}	1.02e-1	7.12e-2	4.02e-2	2.12e-2	1.11e-2	1
# unknowns	239	793	3 113	12 361	50 066	

Table 2. Error of the least-squares solution (case II) for Example 2

	$h = 1/10$	$h = 1/20$	$h = 1/40$	$h = 1/80$	$h = 1/160$	\approx order
l_2 -error for u	2.36e-2	6.28e-3	1.61e-3	4.06e-4	1.02e-4	2
max -error for u	7.29e-2	2.04e-2	5.27e-3	1.33e-3	3.34e-4	2
l_2 -error for \mathbf{p}	6.37e-2	2.26e-2	7.87e-3	2.74e-3	9.59e-4	1.5
max -error for \mathbf{p}	2.30e-1	1.36e-1	7.09e-2	3.58e-2	1.79e-2	1
# unknowns	239	793	3 113	12 361	50 066	

From the computational results one can conclude that for smooth solutions the error behaves as predicted by the theory. A better convergence rate is observed in the L^2 -norm for the vector-field \mathbf{p} . We have not considered any post-processing of the results, neither we have searched for superconvergence points.

6 Acknowledgment

The work of the first author was supported in part by MST of Republic of Serbia under grant # 04M03/C. The second author was supported by part by NSF grant # DMS-9626567 and by EPA grant # R 825207-01-1.

References

1. Adams, R.A.: Sobolev spaces. Academic Press, New York – San Francisco – London 1975.
2. Aziz, A.K., Kellogg, R.B. and Stephens, A.B.: Least-squares methods for elliptic systems, *Math. Comp.*, **44(169)** (1985), 53–70.
3. Babuška, I.: The finite element method with Lagrange multipliers. *Numer. Math.* **20** (1973), 179–192.
4. Bochev, P. B., Gunzburger, M.D.: Accuracy of least-squares methods for the Navier–Stokes equations, *Comput. Fluids*, **22** (1993), 549–563.
5. Bochev, P. B., Gunzburger, M.D.: Analysis of least-squares finite element methods for the Stokes equations, *Math. Comp.*, **63** (1995), 479–505.
6. Bramble, J.H., Lazarov, R.D., and Pasciak, J.E.: A least-squares approach based on a discrete minus one inner Product for first order systems, *Math. Comp.*, **66** (1997) (to appear).
7. Brezzi, F.: On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers. *RAIRO Sér. Anal. Numér.* **8** (1974), 129–151.
8. Brezzi, F., Douglas, J.Jr.: Stabilized mixed method for the Stokes problem. *Numer. Math.* **53** (1988), 225–235.
9. Brezzi, F., Fortin, M.: *Mixed and hybrid finite element methods*. Springer–Verlag, New York 1991.
10. Cai, Z., Lazarov, R., Manteuffel, T., and McCormick, S.: First-order system least squares for partial differential equations: Part I, *SIAM J. Numer. Anal.*, **31** (1994), 1785–1799.
11. Cai, Z., Manteuffel, T. and McCormick, S.: First-order system least squares for partial differential equations: Part II, *SIAM J. Numer. Anal.*, **34** (1997), 425–454.
12. Ciarlet, P.G.: *The finite element method for elliptic problems*, North Holland, Amsterdam – New York – Oxford 1978.
13. Douglas, J.Jr., Wang, J.P.: An absolutely stabilized finite element method for the Stokes problem. *Math. Comput.* **52** (1989), 495–508.
14. Franca, L. P., Stenberg, R.: Error analysis of some Galerkin–least-squares method for the elasticity equations, Report #1054, 1989, INRIA, 1–21.
15. Hughes, T. J. R., Franca, L. P. and Bulestra, M.: A new finite element formulation for computational fluid dynamics. V. Circumventing the Babuška–Brezzi condition: a stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations, *Comput. Meth. Appl. Mech. Engrg.*, **59** (1986), 85–99.
16. Jiang, B. N., Chang, C.: Least-squares finite elements for the Stokes problem, *Comput. Meth. Appl. Mech. Engrg.*, **81** (1990), 13–37.
17. Jespersen, D.C.: A least-square decomposition method for solving elliptic systems, *Math. Comp.*, **31(140)** (1977), 873–880.
18. Jiang, B.N., Povinelli, L.A.: Optimal least-squares finite element method for elliptic problems, *Comput. Meth. Appl. Mech. Engrg.*, **102** (1993), 199–212.

19. Jovanović, B.: Partial differential equations. BS Processor – MF, Belgrade 1993. (Serbian)
20. Ladyzhenskaya, O.A.: The mathematical theory of viscous incompressible flows. Gordon and Breach, London 1969.
21. Ladyzhenskaya, O.A., Ural'tseva, N.N.: Linear and quasilinear equations of elliptic type. Nauka, Moscow 1964. (Russian)
22. Lax, P., Milgram, A.N.: Parabolic equations. Annals of Mathematics Studies **33** (1954), Princeton Univ. Press, Princeton, 167–190
23. Neittaanmäki, P. and Saranen, J.: On finite element approximation of the gradient for the solution to Poisson equation, Numer. Math., **37** (1981), 131–148.
24. Pehlivanov, A.I., Carey, G.F., Lazarov, R.D.: Least-squares mixed finite elements for second-order elliptic problems, SIAM J. Numer. Anal. **31** (1994), 1368–1377.
25. Pehlivanov, A.I., Carey, G.F., Lazarov, R.D. and Shen, Y.: Convergence of least squares finite elements for first order ODE systems, Computing, **51** (1993), 111–123.
26. Raviart, P.A., Thomas, J.M.: A mixed finite element method for second order elliptic equations, Mathematical Aspects of the Finite Element method, Lecture Notes in Mathematics, Springer-Verlag, **606** (1977), 292–315.
27. Wloka, J.: Partial differential equations. Cambridge Univ. Press, Cambridge 1987.