ELSEVIER Journal of Computational and Applied Mathematics 74 (1996) 155–173

JOURNAL OF
COMPUTATIONAL AND
APPLIED MATHEMATICS

Least-squares methods for Stokes equations based on a discrete minus one inner product[1]

James H. Bramble[a,*], Joseph E. Pasciak[b]

[a] *Department of Mathematics, Texas A&M University, College Station, TX 77843-3404, USA*
[b] *Department of Applied Science, Brookhaven National Laboratory, Upton, NY 11973, USA*

Received 15 July 1995; revised 16 February 1996

Abstract

The purpose of this paper is to develop and analyze least-squares approximations for Stokes and elasticity problems. The major advantage of the least-squares formulation is that it does not require that the classical Ladyzhenskaya–Babŭska–Brezzi (LBB) condition be satisfied. We provide two methods. The first is posed in terms of the velocity–pressure pair without the introduction of additional variables. The second adds a vorticity variable. In both cases, we employ least-squares functionals which involve a discrete inner product which is related to the inner product in $H^{-1}(\Omega)$ (the Sobolev space of order minus one on Ω). The use of such inner products (applied to second order problems) was proposed in an earlier paper by Bramble, Lazarov and Pasciak (1994).

Keywords: Least-squares; Stokes equations; Elasticity equations; Preconditioners

AMS classification: 65N30; 65F10

1. Introduction

There is a great deal of literature dealing with approximation schemes for Stokes equations and the equations of linear elasticity (see, [14,16,22,33] and the included references). Mixed finite element methods involving a pair of approximation spaces are commonly used to handle the Stokes equation and avoid locking in linear elasticity problems. These spaces cannot be chosen independently of one another and, for stability, need to satisfy the so-called Ladyzhenskaya–Babuška–Brezzi condition [2,15,29]. To compute the resulting discrete approximation one must

[*] Corresponding author. E-mail: bramble@math.tamu.edu.

[1] This manuscript has been authored under contract number DE-AC02-76CH00016 with the US Department of Energy. Accordingly, the US Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for US Government purposes. This work was also supported in part under the National Science Foundation Grant No. DMS-9007185 and by the US Army Research Office through the Mathematical Sciences Institute, Cornell University.

PII S0377-0427(96)00022-2

solve saddle point systems. Although much progress has been made in the development of efficient iterative procedures for solving such saddle point problems [9, 32], they still pose some difficulties.

In this paper, we shall define and analyze two least-squares approximation techniques for solving the equations of Stokes and linear elasticity. The first scheme involves the same variables as the original mixed formulation, i.e., the velocity and pressure. The second scheme involves an additional variable, the vorticity. Since the second method involves more degrees of freedom in the algebraic system, it is only useful when the vorticity is desired as output from the computation.

The least-squares approach provides a number of advantages over the usual mixed finite element discretizations. First of all, the pairs (triples) of approximation spaces are not required to satisfy the LBB condition. Thus, for example, one can use conforming spaces of piecewise linear functions on the same mesh for both velocity and pressure approximation. This simplifies the implementation. In addition, the algebraic system which must be solved to compute the discrete solution is symmetric and positive definite. Iteration schemes for positive definite systems are most effective and best understood.

There is a substantial literature dealing with the application of least-squares methods to Stokes and Navier–Stokes equations [4, 5, 19, 20, 26–28, 35] In this paper, we will only consider the equations of Stokes and elasticity with zero velocity boundary conditions. Extensions of the ideas developed here will be addressed in future work.

The approach of this paper is new in that it is based on using a discrete negative norm for one of the terms defining the least-squares functional and is an extension of some of the work done in [13]. The use of such norms gives rise to two important advantages. They result in approximation methods which are optimal both in the order of approximation as well as required regularity. In addition, the corresponding algebraic systems can be easily preconditioned. This means that efficient iterative schemes can be developed to compute the discrete least-squares approximation.

The outline of the remainder of the paper is as follows. The least-squares methods which are given in this paper are based on a number of stability estimates. These stability estimates are proved in Section 2. Section 3 describes and analyzes the first least-squares method (without vorticity) while Section 4 studies the least-squares method involving the vorticity. Because of the discrete negative norm, these methods give rise to algebraic systems which are full and hence it is not feasible to assemble the matrices. Nevertheless, it is still possible to efficiently solve these systems since it is inexpensive to compute the action of the matrix corresponding to the system. The development and implementation of effective iterative solvers for the new least-squares systems is given in Section 5.

2. Stokes problem and some stability estimates

In this section, we provide some a priori inequalities which will be of critical importance for the stability and convergence of the least-squares methods studied in the remainder of this paper. We first present some function spaces and define the Stokes and elasticity problems. Next we introduce some finite element spaces and derive the stability estimates.

Let Ω be a Lipschitz domain with polygonal or polyhedral boundary in d dimensional Euclidean space for $d = 2$ or 3 . The velocity–pressure formulation of the Stokes problem is: Find $\boldsymbol{u}$ and p satisfying

$$-2\sum_{j=1}^{d} \frac{\partial}{\partial x_j}\varepsilon_{ij}(\boldsymbol{u}) + \frac{\partial p}{\partial x_i} = \boldsymbol{F}_i \qquad \text{in } \Omega, \tag{1}$$

$$\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0 \qquad \text{in } \Omega, \tag{2}$$

$$\boldsymbol{u} = 0 \qquad \text{on } \partial\Omega, \tag{3}$$

$$\int_\Omega p = 0, \tag{4}$$

for $i = 1,\ldots,d$. Here $\varepsilon_{ij}(\boldsymbol{u})$ is the symmetric strain tensor defined by

$$\varepsilon_{ij}(\boldsymbol{u}) \equiv \frac{1}{2}\left\{\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right\}.$$

To get the equations of linear elasticity, we replace Eq. (2) by

$$\boldsymbol{\nabla} \cdot \boldsymbol{u} = -(1-2v)p. \tag{5}$$

In Eq. (5), $0 < v < \frac{1}{2}$ is Poisson's ratio. If $v = \frac{1}{2}$ in (5), we recover the Stokes case (2).

Remark 1. Condition (3) corresponds to a fixed boundary where the material is clamped. Although we only treat these boundary conditions, other types of boundary conditions are interesting. In particular, for elasticity problems, one often considers mixed conditions of the form

$$\boldsymbol{u} = 0 \qquad \text{on } \partial\Omega/\tilde{\Gamma}, \tag{6}$$

$$\sum_{j=1}^{d} \varepsilon_{ij}(\boldsymbol{u})n_j - p\ n_i = 0 \qquad \text{on } \tilde{\Gamma}, \tag{7}$$

where $\tilde{\Gamma}$ is a subset of $\partial\Omega$. In this case, $\tilde{\Gamma}$ is a segment of the boundary where the material is allowed to deform. Conditions of the form of (7) will be addressed in a subsequent paper.

The stability and regularity properties for solutions of the above problem are most naturally described in terms of Sobolev spaces. Let $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product and $\|\cdot\|$ the corresponding norm. We will use the same inner product and norm notation for vector valued functions in the product space $(L^2(\Omega))^d$. For positive values of s, let $H^s(\Omega)$ denote the Sobolev space of order s and $\|\cdot\|_s$ denote the corresponding norm (cf. [25,30]). Let $H_0^1(\Omega)$ be the set of functions in $H^1(\Omega)$ with vanishing trace on $\partial\Omega$. The Dirichlet form on Ω is defined by

$$D(v,w) \equiv \int_\Omega \boldsymbol{\nabla} v \cdot \boldsymbol{\nabla} w \,\mathrm{d}x \qquad \text{for all } v,w \in H^1(\Omega).$$

Since functions in $H_0^1(\Omega)$ vanish on $\partial\Omega$, the Poincaré inequality implies that the norm

$$\| v \|_1 = D(v,v)^{1/2} \qquad \text{for all } v \in H_0^1(\Omega),$$

is equivalent to the usual Sobolev norm. It shall be convenient to use the above norm in our subsequent analysis. Let $\mathbf{H}_0^1$ denote the product space $(H_0^1(\Omega))^d$. Its norm is induced from the form

$$\mathbf{D}(\boldsymbol{w},\boldsymbol{v}) \equiv \sum_{j=1}^{d} D(w_j, v_j). \tag{8}$$

Without ambiguity, we will use $\|\cdot\|_1$ to denote the norms in both $H_0^1(\Omega)$ and $\mathbf{H}_0^1(\Omega)$. We shall also use Sobolev spaces with negative indices. In particular, the space $\mathbf{H}^{-1}(\Omega)$ is defined to be those distributions for which the norm

$$\| \boldsymbol{V} \|_{-1} = \sup_{\boldsymbol{W}\in \boldsymbol{H}_0^1(\Omega)} \frac{(\boldsymbol{V},\boldsymbol{W})}{\| \boldsymbol{W} \|_1}$$

is finite. Let $C_0^\infty(\Omega)$ denote the infinitely differentiable functions with support in Ω and let $\mathbf{C}_0^\infty(\Omega) = (C_0^\infty(\Omega))^d$. The spaces $C_0^\infty(\Omega)$ and $\mathbf{C}_0^\infty(\Omega)$ are dense in $H_0^1(\Omega)$ and $\mathbf{H}_0^1(\Omega)$, respectively. Finally, we denote by Π the set of functions in $L^2(\Omega)$ with zero mean value on Ω.

Eq. (1) can be rewritten

$$-\boldsymbol{\nabla}^2\boldsymbol{u} + (2-2\nu)\boldsymbol{\nabla} p = \boldsymbol{F} \quad \text{in } \Omega. \tag{9}$$

Note that (9) implies that $\boldsymbol{u}$ and a rescaled p satisfy

$$-\boldsymbol{\nabla}^2\boldsymbol{u} + \boldsymbol{\nabla} p = \boldsymbol{F}, \tag{10}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} + \gamma p = 0. \tag{11}$$

Here $\gamma = 0$ in the Stokes case and $\gamma = (1-2\nu)/(2-2\nu)$ for the elasticity problem.

To approximately solve (1)–(5), we introduce a pair of subspaces $\mathbf{V}_h, \Pi_h$ indexed by h in the interval $0 < h < 1$. We do this by partitioning the region $\Omega = \bigcup_i \bar{\tau}_i$ into triangles or tetrahedrons of quasi-uniform size. With some abuse of semantics, we shall refer to τ_i as a triangle in both the two and three dimensional case. As usual, the boundaries of two triangles or tetrahedrons will intersect at either a vertex, an entire edge or an entire face. Spaces defined with respect to rectangular or parallelpiped partitioning of Ω pose no added difficulty. For some integer $r \geqslant 2$, let V_h denote the functions which are piecewise polynomials of degree less than r with respect to the triangles, continuous on Ω and vanish on $\partial\Omega$ and define $\mathbf{V}_h = (V_h)^d$. Let Π_h denote a space of functions which are piecewise smooth with respect to the triangles defining the mesh.

Remark 2. As a typical example, we could define Π_h as (continuous or discontinuous) functions which are piecewise polynomials of degree less than $r-1$.

The following approximation properties are well known for the finite element spaces just defined.

(1) For $\boldsymbol{v} \in (H^r(\Omega) \cap H_0^1(\Omega))^d$,

$$\inf_{\boldsymbol{W}\in \boldsymbol{V}_h} \| \boldsymbol{v} - \boldsymbol{W} \|_1 \leqslant Ch^{r-1} \| \boldsymbol{v} \|_r . \tag{12}$$

(2) For $\boldsymbol{v} \in H^{r-1}(\Omega) \cap \Pi$,

$$\inf_{W\in \Pi_h} \| \boldsymbol{v} - W \| \leqslant Ch^{r-1} \| \boldsymbol{v} \|_{r-1} . \tag{13}$$

The constant C appearing above is independent of the approximation parameter h. It follows from [7, 12] and (12) above that given $\boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega)$, there exists $\boldsymbol{W} \in \mathbf{V}_h$ and constants C_1, C_2 not depending on h or $\boldsymbol{v}$ such that

$$\| \boldsymbol{W} \|_1 \leqslant C_1 \| \boldsymbol{v} \|_1, \tag{14}$$

and

$$\| \boldsymbol{W} - \boldsymbol{v} \| \leqslant C_2 h \| \boldsymbol{v} \|_1 . \tag{15}$$

To describe the first a priori estimate, we shall need some operators and additional norms on the discrete spaces just defined. The operators $\boldsymbol{\nabla}_h^2 : \mathbf{H}_0^1(\Omega) \mapsto \mathbf{V}_h$ and $\nabla_h : L^2(\Omega) \mapsto \mathbf{V}_h$ are defined by $\boldsymbol{\nabla}_h^2 \boldsymbol{v} = \boldsymbol{W}$ where $\boldsymbol{W}$ is the unique function in $\mathbf{V}_h$ satisfying

$$(\boldsymbol{W}, \boldsymbol{X}) = -\mathbf{D}(\boldsymbol{v}, \boldsymbol{X}) \quad \text{for all } \boldsymbol{X} \in \mathbf{V}_h$$

and $\nabla_h q = \boldsymbol{V}$, where $\boldsymbol{V}$ is the unique function in $\mathbf{V}_h$ satisfying

$$(\boldsymbol{V}, \boldsymbol{X}) = -(q, \boldsymbol{\nabla} \cdot \boldsymbol{X}) \quad \text{for all } \boldsymbol{X} \in \mathbf{V}_h.$$

Let $\{\Gamma_i\}$ be the collection of interior edges (respectively, faces) in the partitioning of Ω into triangles and define the semi-norm $\{\Gamma_i\}$ by

$$|\theta|_h = \left(\sum_i \int_{\Gamma_i} \theta(s)^2 \, \mathrm{d}s \right)^{1/2}.$$

We shall also use the notation

$$\| \boldsymbol{V} \|_h = \left(\sum_i \int_{\tau_i} | \boldsymbol{V}(x)|^2 \, \mathrm{d}x \right)^{1/2}.$$

This provides a norm for functions which are piecewise smooth with respect to the triangulation. For example, even though a second order distributional derivative of a function in $\mathbf{V}_h$ may not be in $(L^2(\Omega))^d$, its corresponding $\| \cdot \|_h$–norm is finite. Finally, for $\boldsymbol{v} \in \mathbf{H}^{-1}(\Omega)$, we define the semi-norm

$$\| \boldsymbol{v} \|_{-1,h} = \sup_{\boldsymbol{W} \in \mathbf{V}_h} \frac{(\boldsymbol{v}, \boldsymbol{W})}{\| \boldsymbol{W} \|_1}.$$

We can now state and prove the first a priori estimate.

Theorem 1. *There is a positive constant C which does not depend on h or γ and satisfies*

$$\begin{aligned} \| P \| + \| \boldsymbol{V} \|_1 \leqslant C \{ \| -\boldsymbol{\nabla}_h^2 \boldsymbol{V} + \nabla_h P \|_{-1,h} + h \| -\boldsymbol{\nabla}^2 \boldsymbol{V} + \boldsymbol{\nabla} P \|_h \\ + h^{1/2} [[P]]_h + \| \boldsymbol{\nabla} \cdot \boldsymbol{V} + \gamma P \| \}, \end{aligned} \tag{16}$$

for all $P \in \Pi_h$ and $\boldsymbol{V} \in \mathbf{V}_h$. Here $[[P]]$ denotes the jump in P across the interior edges $\{\Gamma_i\}$ of the triangulation defining the subspace Π_h.

Proof. Here and in the remainder of this paper, C with or without subscript will denote a generic positive constant independent of h and γ. These constants may represent different values in different occurrences.

We start by deriving an estimate for $\| P \|$ for functions $P \in \Pi_h$. It is well known (see, e.g., [22]) that for all $p \in \Pi$,

$$\| p \| \leqslant C \sup_{v \in \mathbf{H}_0^1(\Omega)} \frac{(p, \nabla \cdot v)}{\| v \|_1}. \tag{17}$$

Fix $P \in \Pi_h$, $v \in \mathbf{H}_0^1(\Omega)$ and let W satisfy (14) and (15). Then,

$$\begin{aligned} |(P, \nabla \cdot v)| &\leqslant |(P, \nabla \cdot (v - W))| + |(P, \nabla \cdot W)| \\ &= |(P, \nabla \cdot (v - W))| + |(\nabla_h P, W)|. \end{aligned} \tag{18}$$

For the first term of (18), integrating by parts gives

$$|(P, \nabla \cdot (v - W))| \leqslant \sum_i \left| \int_{\tau_i} \nabla P \cdot (v - W) \, \mathrm{d}x \right| + \sum_j \left| \int_{\Gamma_j} [[P]](v - W) \cdot n \, \mathrm{d}s \right|.$$

Here n denotes a unit normal vector on the edge Γ_j. Using the well known inequality

$$\int_{\partial \tau_i} |\theta|^2 \, \mathrm{d}s \leqslant C(h^{-1} \| \theta \|_{L^2(\tau_i)}^2 + h \| \theta \|_{H^1(\tau_i)}^2), \tag{19}$$

it follows from (14) and (15) that

$$\sum_j \left| \int_{\Gamma_j} [[P]](v - W) \cdot n \, \mathrm{d}s \right| \leqslant Ch^{1/2} |[[P]]|_h \| v \|_1 .$$

In addition, (15) implies that

$$\sum_i \left| \int_{\tau_i} \nabla P \cdot (v - W) \, \mathrm{d}x \right| \leqslant Ch \| \nabla P \|_h \| v \|_1 .$$

Combining the above estimates and obvious manipulations gives

$$\| P \| \leqslant C(h^{1/2} [[P]]_h + h \| \nabla P \|_h + \| \nabla_h P \|_{-1,h}). \tag{20}$$

We next estimate $\mathbf{D}(U, U)$ for $U \in \mathbf{V}_h$. Fix $U \in \mathbf{V}_h$ and $P \in \Pi_h$. We clearly have that

$$\mathbf{D}(U, U) = (-\nabla_h^2 U + \nabla_h P, U) + (P, \nabla \cdot U + \gamma P) - \gamma(P, P).$$

Thus,

$$\mathbf{D}(U, U) \leqslant \| -\nabla_h^2 U + \nabla_h P \|_{-1,h} \| U \|_1 + \| P \| \| \nabla \cdot U + \gamma P \|. \tag{21}$$

We note that by (20),

$$\begin{aligned} \| P \| \leqslant C(&h^{1/2} [[P]]_h + h \| -\nabla^2 U + \nabla P \|_h + h \| \nabla^2 U \|_h + \| -\nabla_h^2 U + \nabla_h P \|_{-1,h} \\ &+ \| \nabla_h^2 U \|_{-1,h}). \end{aligned} \tag{22}$$

It follows from the quasi-uniformity of the mesh that

$$h \| \nabla^2 U \|_h \leqslant C \| U \|_1 . \tag{23}$$

Moreover, by definition

$$\| \boldsymbol{\nabla}_h^2 \boldsymbol{U} \|_{-1,h} \leqslant \| \boldsymbol{U} \|_1 . \tag{24}$$

Combining (21)–(24) and obvious manipulations gives

$$\begin{aligned}\mathbf{D}(\boldsymbol{U},\boldsymbol{U})^{1/2} \leqslant C\{\| -\boldsymbol{\nabla}_h^2 \boldsymbol{U} + \nabla_h P \|_{-1,h} + h \| -\boldsymbol{\nabla}^2 \boldsymbol{U} + \boldsymbol{\nabla} P \|_h \\ + h^{1/2}[[P]]_h + \| \boldsymbol{\nabla} \cdot \boldsymbol{U} + \gamma P \| \}.\end{aligned} \tag{25}$$

The theorem follows by combining (22)–(25).

We will also study least-squares approximations of the Stokes and elasticity equations utilizing the vorticity

$$\theta = \nabla \times \boldsymbol{u}.$$

In the case of $d = 2$, the vorticity vector is a scalar. In that case, we set $\mathscr{V}$ to be the functions in $L^2(\Omega)$. If $d = 3$, the vorticity is a vector with three components and we take $\mathscr{V} = (L^2(\Omega))^3$.

It immediately follows from the identity

$$-\boldsymbol{\nabla}^2 \boldsymbol{v} = \nabla \times \nabla \times \boldsymbol{v} - \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{v}) \tag{26}$$

that (9) can be rewritten as

$$\nabla \times \nabla \times \boldsymbol{u} + (3 - 4\nu)\boldsymbol{\nabla} p = \boldsymbol{F} \quad \text{in } \Omega. \tag{27}$$

Note that (27) implies that $\boldsymbol{u}$ and a rescaled p satisfy

$$\nabla \times \nabla \times \boldsymbol{u} + \boldsymbol{\nabla} p = \boldsymbol{F}, \qquad \boldsymbol{\nabla} \cdot \boldsymbol{u} + \bar{\gamma} p = 0.$$

Here $\bar{\gamma} = 0$ in the Stokes case and $\bar{\gamma} = \gamma/(1 + \gamma) = (1 - 2\nu)/(3 - 4\nu)$ for the elasticity problem. Substituting the vorticity variable above gives the system

$$\nabla \times \theta + \boldsymbol{\nabla} p = \boldsymbol{F}, \qquad \boldsymbol{\nabla} \cdot \boldsymbol{u} + \bar{\gamma} p = 0, \qquad \theta - \nabla \times \boldsymbol{u} = 0. \tag{28}$$

The a priori estimate which we shall use in our vorticity least-squares approximations is given in the following theorem.

Theorem 2. *For all $q \in \Pi$, $\phi \in \mathscr{V}$ and $\boldsymbol{v} \in \mathbf{H}_0^1(\Omega)$,*

$$\| \phi \| + \| q \| + \| \boldsymbol{v} \|_1 \leqslant C[\| \nabla \times \boldsymbol{v} - \phi \| + \| \nabla \times \phi + \boldsymbol{\nabla} q \|_{-1} + \| \boldsymbol{\nabla} \cdot \boldsymbol{v} + \bar{\gamma} q \|].$$

Proof. Let $\boldsymbol{v}$ be in $\mathbf{C}_0^\infty(\Omega)$, ϕ be in $(C_0^\infty(\Omega))^{2d-3}$ and q be in $\Pi \cap C^\infty(\Omega)$. Then by (26),

$$\begin{aligned}\mathbf{D}(\boldsymbol{v},\boldsymbol{v}) &= (\nabla \times \nabla \times \boldsymbol{v}, \boldsymbol{v}) + \| \boldsymbol{\nabla} \cdot \boldsymbol{v} \|^2 \\ &= (\nabla \times \phi + \boldsymbol{\nabla} q, \boldsymbol{v}) + (\nabla \times \boldsymbol{v} - \phi, \nabla \times \boldsymbol{v}) + \| \boldsymbol{\nabla} \cdot \boldsymbol{v} \|^2 + (q, \boldsymbol{\nabla} \cdot \boldsymbol{v}).\end{aligned} \tag{29}$$

Now

$$\begin{aligned} \| \nabla \cdot \boldsymbol{v} \|^2 + (q, \nabla \cdot \boldsymbol{v}) &= (\nabla \cdot \boldsymbol{v} + \bar{\gamma} q, \nabla \cdot \boldsymbol{v}) + (1-\bar{\gamma})(q, \nabla \cdot \boldsymbol{v} + \bar{\gamma} q) - (1-\bar{\gamma})\bar{\gamma} \| q \|^2 \\ &\leqslant \| \nabla \cdot \boldsymbol{v} + \bar{\gamma} q \| \, (\| \boldsymbol{v} \|_1 + \| q \|). \end{aligned} \tag{30}$$

Combining (29) and (30) and obvious manipulations gives

$$\begin{aligned} \mathbf{D}(\boldsymbol{v},\boldsymbol{v}) \leqslant 3(& \| \nabla \times \phi + \nabla q \|_{-1}^2 + \| \nabla \times \boldsymbol{v} - \phi \|^2 \\ & + \| \nabla \cdot \boldsymbol{v} + \bar{\gamma} q \|^2) + 2 \| q \| \| \nabla \cdot \boldsymbol{v} + \bar{\gamma} q \| . \end{aligned} \tag{31}$$

Now, for $\boldsymbol{W} \in \mathbf{H}_0^1(\Omega)$,

$$-(q, \nabla \cdot \boldsymbol{W}) = (\nabla q + \nabla \times \phi, \boldsymbol{W}) - (\nabla \times \phi - \nabla \times \nabla \times \boldsymbol{v}, \boldsymbol{W}) - (\nabla \times \nabla \times \boldsymbol{v}, \boldsymbol{W})$$

and hence

$$|(q, \nabla \cdot \boldsymbol{W})| \leqslant (\| \nabla q + \nabla \times \phi \|_{-1} + \| \phi - \nabla \times \boldsymbol{v} \| + \| \boldsymbol{v} \|_1) \, \| \boldsymbol{W} \|_1 .$$

Thus, by (17),

$$\| q \| \leqslant C(\| \nabla q + \nabla \times \phi \|_{-1} + \| \phi - \nabla \times \boldsymbol{v} \| + \| \boldsymbol{v} \|_1). \tag{32}$$

Combining (31) and (32) gives

$$\mathbf{D}(\boldsymbol{v},\boldsymbol{v}) \leqslant C(\| \nabla \times \phi + \nabla q \|_{-1}^2 + \| \nabla \times \boldsymbol{v} - \phi \|^2 + \| \nabla \cdot \boldsymbol{v} + \bar{\gamma} q \|^2). \tag{33}$$

We finally note that

$$\| \phi \| \leqslant \| \nabla \times \boldsymbol{v} - \phi \| + \| \boldsymbol{v} \|_1 . \tag{34}$$

The theorem then follows from (32)–(34) and density.

3. Least-squares methods without vorticity

In this section, we consider a least-squares method for the Stokes equation which only involves approximations to the original variables p and $\boldsymbol{v}$. The stability properties of this method result from Theorem 1.

Let $< \cdot, \cdot >_h$, $(\cdot,\cdot)_h$ and $(\cdot,\cdot)_{-1,h}$ denote the forms which correspond to the norms $| \cdot |_h$, $\| \cdot \|_h$, and $\| \cdot \|_{-1,h}$, respectively. We clearly have that

$$< P, Q >_h = \sum_i \int_{\Gamma_i} P(s) Q(s) \, \mathrm{d}s$$

and

$$(\boldsymbol{V}, \boldsymbol{W})_h = \sum_i \int_{\tau_i} \boldsymbol{V}(x) \cdot \boldsymbol{W}(x) \, \mathrm{d}x.$$

It is not difficult to show that

$$(\boldsymbol{v},\boldsymbol{w})_{-1,h}=(\mathbf{T}_h\boldsymbol{v},\boldsymbol{w}), \tag{35}$$

where $\mathbf{T}_h:\mathbf{H}^{-1}(\Omega)\mapsto\mathbf{V}_h$ is the solution operator defined by

$$\mathbf{D}(\mathbf{T}_h\boldsymbol{v},\boldsymbol{X})=(\boldsymbol{v},\boldsymbol{X})\quad\text{for all } \boldsymbol{X}\in\mathbf{V}_h. \tag{36}$$

The identity (35) holds for all $\boldsymbol{v},\boldsymbol{w}\in\mathbf{H}^{-1}(\Omega)$.

Let $\mathbf{B}_h:\mathbf{H}^{-1}(\Omega)\mapsto\mathbf{V}_h$ be a preconditioner for the problem (36). This means that $\mathbf{B}_h$ is symmetric and positive semi-definite on $L^2(\Omega)$ and is spectrally equivalent to $\mathbf{T}_h$; i.e.,

$$C_0(\mathbf{T}_h\boldsymbol{V},\boldsymbol{V})\leqslant(\mathbf{B}_h\boldsymbol{V},\boldsymbol{V})\leqslant C_1(\mathbf{T}_h\boldsymbol{V},\boldsymbol{V}) \tag{37}$$

for all $\boldsymbol{V}$ in $\mathbf{V}_h$, with constants C_0,C_1 independent of h. Obviously, $\mathbf{B}_h$ can be taken equal to $\mathbf{T}_h$. However, it is often more efficient to make some other choice.

There is a vast literature of techniques for developing preconditioners for symmetric positive definite problems, especially in the case of a discretization of an elliptic partial differential equation (see, e.g., [1,3,6,17,18,23,24]). The best preconditioners satisfy (37) with constants C_0 and C_1 independent of the mesh parameter. In addition, a good preconditioner is economical to evaluate. This means that the cost of computing the action of $\mathbf{B}_h$ applied to an arbitrary vector should be much less than that of applying $\mathbf{T}_h$. For our application, low cost preconditioners are known for which (37) holds with C_0 and C_1 independent of the mesh size and hence the number of unknowns (see, e.g., [6,8,10,11,21,31,34]).

The least-squares method which we shall consider is based on the form

$$\begin{aligned}\ll (q,\boldsymbol{v}),(r,\boldsymbol{w})\gg_1\equiv&(\mathbf{B}_h(-\boldsymbol{\nabla}_h^2\boldsymbol{v}+\nabla_h q),-\boldsymbol{\nabla}_h^2\boldsymbol{w}+\nabla_h r)\\&+h^2(-\boldsymbol{\nabla}^2\boldsymbol{v}+\boldsymbol{\nabla}q,-\boldsymbol{\nabla}^2\boldsymbol{w}+\boldsymbol{\nabla}r)_h\\&+h<[[q]],[[r]]>+(\boldsymbol{\nabla}\cdot\boldsymbol{v}+\gamma q,\boldsymbol{\nabla}\cdot\boldsymbol{w}+\gamma r).\end{aligned} \tag{38}$$

The least-squares solution is the pair $(P,\boldsymbol{U})\in\Pi_h\times\mathbf{V}_h$ satisfying

$$\ll (P,\boldsymbol{U}),(R,\boldsymbol{W})\gg_1=(\mathbf{B}_h\boldsymbol{F},-\boldsymbol{\nabla}_h^2\boldsymbol{W}+\nabla_h R)+h^2(\boldsymbol{F},-\boldsymbol{\nabla}^2\boldsymbol{W}+\boldsymbol{\nabla}R)_h \tag{39}$$

for all $(R,\boldsymbol{W})$ in $\Pi_h\times\mathbf{V}_h$. It is a direct consequence of Theorem 1 and (37) that for $\boldsymbol{F}\in(L^2(\Omega))^d$, the solution $(P,\boldsymbol{U})$ of (39) exists and is unique. The following theorem shows that the approximate solution $(P,\boldsymbol{U})$ is close to the solution pair $(p,\boldsymbol{u})$ satisfying (10) and (11).

Theorem 3. *Let $(P,\boldsymbol{U})$ solve (39) and $(p,\boldsymbol{u})$ be the solution of (10), (11). Let $\mathbf{V}_h$ and Π_h be as described in the previous section with $r\geqslant 2$. Assume that $\boldsymbol{F}\in(L^2(\Omega))^d$ and that the solution $(p,\boldsymbol{u})$ is in $H^{r-1}(\Omega)\times(H^r(\Omega)\cap H_0^1(\Omega))^d$. Then*

$$\|\boldsymbol{U}-\boldsymbol{u}\|_1+\|P-p\|\leqslant Ch^{r-1}(\|\boldsymbol{u}\|_r+\|p\|_{r-1}).$$

Proof. The solution $(p,\boldsymbol{u})$ satisfies

$$\mathbf{D}(\boldsymbol{u},\boldsymbol{w})-(p,\boldsymbol{\nabla}\cdot\boldsymbol{w})=(\boldsymbol{F},\boldsymbol{w})\quad\text{for all } \boldsymbol{w}\in\boldsymbol{H}_0^1(\Omega). \tag{40}$$

Thus, for $\boldsymbol{W} \in \mathbf{V}_h$,

$$(-\boldsymbol{\nabla}_h^2\boldsymbol{u} + \nabla_h p, \boldsymbol{W}) = (\boldsymbol{F}, \boldsymbol{W}). \tag{41}$$

In addition, by (40) and density,

$$(-\boldsymbol{\nabla}^2\boldsymbol{u} + \boldsymbol{\nabla} p, \boldsymbol{w}) = (\boldsymbol{F}, \boldsymbol{w}) \quad \text{for all } \boldsymbol{w} \in (L^2(\Omega))^d. \tag{42}$$

It follows from (41), (42) and (39) that

$$\ll (e,\boldsymbol{E}),(R,\boldsymbol{W}) \gg_1 = 0 \quad \text{for all } (R,\boldsymbol{W}) \in \Pi_h \times \mathbf{V}_h, \tag{43}$$

where $e = p - P$ and $\boldsymbol{E} = \boldsymbol{u} - \boldsymbol{U}$. Let $\tilde{P}$ and $\tilde{\boldsymbol{U}}$ satisfy

$$\| p - \tilde{P} \| \leqslant Ch^{r-1} \| p \|_{r-1} \tag{44}$$

and

$$\| \boldsymbol{u} - \tilde{\boldsymbol{U}} \| + h \| \boldsymbol{u} - \tilde{\boldsymbol{U}} \|_1 \leqslant Ch^r \| \boldsymbol{u} \|_r . \tag{45}$$

That such a $\tilde{\boldsymbol{U}}$ exists follows from [7] and [12]. Setting $(\tilde{e},\tilde{\boldsymbol{E}}) = (\tilde{P} - P, \tilde{\boldsymbol{U}} - \boldsymbol{U})$, Theorem 1, (37) and (43) give that

$$\begin{aligned} \| \tilde{e} \|^2 + \| \tilde{\boldsymbol{E}} \|_1^2 &\leqslant C \ll (\tilde{e},\tilde{\boldsymbol{E}}),(\tilde{e},\tilde{\boldsymbol{E}}) \gg_1 \\ &= C \ll (\tilde{P} - p, \tilde{\boldsymbol{U}} - \boldsymbol{u}),(\tilde{e},\tilde{\boldsymbol{E}}) \gg_1 . \end{aligned}$$

It immediately follows that

$$\begin{aligned} \| \tilde{e} \|^2 + \| \tilde{\boldsymbol{E}} \|_1^2 &\leqslant C \ll (\tilde{P} - p, \tilde{\boldsymbol{U}} - \boldsymbol{u}),(\tilde{P} - p, \tilde{\boldsymbol{U}} - \boldsymbol{u}) \gg_1 \\ &\leqslant C(\| -\boldsymbol{\nabla}_h^2(\tilde{\boldsymbol{U}} - \boldsymbol{u}) + \nabla_h(\tilde{P} - p) \|_{-1,h}^2 \\ &\quad + h^2 \| -\boldsymbol{\nabla}^2(\tilde{\boldsymbol{U}} - \boldsymbol{u}) + \boldsymbol{\nabla}(\tilde{P} - p) \|_h^2 \\ &\quad + h |[\tilde{P} - p]|_h^2 + \| \boldsymbol{\nabla} \cdot (\tilde{\boldsymbol{U}} - \boldsymbol{u}) + \gamma(\tilde{P} - p) \|^2). \end{aligned} \tag{46}$$

The last inequality above follows from (37).

We now bound the terms on the right-hand side of (46). For the first term, we clearly have that

$$\| \boldsymbol{\nabla}_h^2(\tilde{\boldsymbol{U}} - \boldsymbol{u}) \|_{-1,h} \leqslant \| \tilde{\boldsymbol{U}} - \boldsymbol{u} \|_1 \quad \text{and} \quad \| \nabla_h(\tilde{P} - p) \|_{-1,h} \leqslant \| \tilde{P} - p \| .$$

Thus,

$$\| \boldsymbol{\nabla}_h^2(\tilde{\boldsymbol{U}} - \boldsymbol{u}) + \nabla_h(\tilde{P} - p) \|_{-1,h}^2 \leqslant Ch^{r-1}(\| \boldsymbol{u} \|_r + \| p \|_{r-1}). \tag{47}$$

Let $\bar{\Pi}_h$ denote the set of discontinuous piecewise polynomial functions of degree less than r with respect to the triangulation defining Π_h. For the second term in (46), we note that

$$h \| \nabla(\tilde{P} - p) \|_h \leqslant h \| \nabla(\tilde{P} - \bar{P}) \|_h + h \| \nabla(\bar{P} - p) \|_h,$$

where $\bar{P}$ is the $L^2(\Omega)$ projection of p into $\tilde{\Pi}_h$. Since the mesh is quasi-uniform, it follows that

$$\begin{aligned} h \parallel \nabla(\tilde{P} - p) \parallel_h &\leqslant \parallel (\tilde{P} - \bar{P}) \parallel + Ch^{r-1} \parallel p \parallel_{r-1} \\ &\leqslant Ch^{r-1} \parallel p \parallel_{r-1} . \end{aligned} \tag{48}$$

A similar argument gives that

$$h \parallel \boldsymbol{\nabla}^2(\tilde{\boldsymbol{U}} - \boldsymbol{u}) \parallel_h \leqslant Ch^{r-1} \parallel \boldsymbol{u} \parallel_r . \tag{49}$$

For the third term of the right-hand side of (47), we again use $\bar{P}$ as defined above and get

$$h^{1/2}|\tilde{P} - p|_h \leqslant h^{1/2}|\tilde{P} - \bar{P}|_h + h^{1/2}|\bar{P} - p|_h.$$

Since $\bar{P}$ is defined as a local $L^2(\Omega)$ projection on each triangle or tetrahedron,

$$\begin{aligned} \int_{\partial\tau_i} (\bar{P}(s) - p(s))^2 \,\mathrm{d}s &\le C\left(h^{-1} \int_{\tau_i} (\bar{P}(x) - p(x))^2 \,\mathrm{d}x + h \int_{\tau_i} |\nabla(\bar{P}(x) - p(x))|^2 \,\mathrm{d}x\right. \\ &\leqslant Ch^{2r-3} \parallel p \parallel^2_{H^{r-1}(\tau_i)} . \end{aligned}$$

Summing the above inequality gives

$$h^{1/2}|\bar{P} - p|_h \leqslant Ch^{r-1} \parallel p \parallel_{r-1} .$$

In addition, since $\tilde{P} - \bar{P}$ is a polynomial on τ_i, standard reference element mapping arguments imply that

$$h \int_{\partial\tau_i} (\tilde{P}(s) - \bar{P}(s))^2 \,\mathrm{d}s \leqslant C \int_{\tau_i} (\tilde{P}(x) - \bar{P}(x))^2 \,\mathrm{d}x$$

and hence

$$h^{1/2}|\tilde{P} - \bar{P}|_h \leqslant C \parallel \tilde{P} - \bar{P} \parallel_h \leqslant Ch^{r-1} \parallel p \parallel_{r-1} .$$

Combining the above inequalities shows that

$$h^{1/2}|\tilde{P} - p|_h \leqslant Ch^{r-1} \parallel p \parallel_{r-1} . \tag{50}$$

For the last term on the right-hand side of (46), (44) and (45) immediately imply that

$$\parallel \boldsymbol{\nabla} \cdot (\tilde{\boldsymbol{U}} - \boldsymbol{u}) + \gamma(\tilde{P} - p) \parallel \leqslant Ch^{r-1}(\parallel \boldsymbol{u} \parallel_r + \parallel p \parallel_{r-1}). \tag{51}$$

Combining (46)–(51) gives

$$\parallel \tilde{e} \parallel^2 + \parallel \tilde{\boldsymbol{E}} \parallel_1^2 \leqslant Ch^{r-1}(\parallel \boldsymbol{u} \parallel_r + \parallel p \parallel_{r-1}).$$

The theorem follows from (44), (45) and the triangle inequality.

Remark 3. The assumption that $\boldsymbol{F} \in (L^2(\Omega))^d$ can be relaxed to $\boldsymbol{F} \in \mathbf{H}^{-1}(\Omega)$ provided that $\boldsymbol{F}$ is replaced by $\mathbf{Q}_h\boldsymbol{F}$ in the h^2 term on the right-hand side of (39). Here $\mathbf{Q}_h$ denotes the $(L^2(\Omega))^d$ projector onto $\mathbf{V}_h$. In this case, the method is stable and convergent (with the expected rates) for $(p,\boldsymbol{u})$ in $H^{s-1}(\Omega) \times (H^s(\Omega) \cap H^1_0(\Omega))^d$ for $1 \leqslant s \leqslant r$. The proof of this result is essentially contained in the proof of Corollary 3.1 of [13].

4. A least-squares method involving the vorticity

In this section, we consider a least-squares method for the Stokes which involves approximation to the vorticity θ as well as the original variables p and $\boldsymbol{v}$. The stability properties of this method result from Theorem 2.

We need an additional approximation space for the vorticity. Let S_h denote the space of continuous piecewise polynomial functions of degree less than $r-1$ with respect to the mesh of Section 2. If $r=2$, we take S_h to be the set of continuous piecewise linear functions with respect to the triangulation. For implementation convenience, one also can use the larger space of continuous piecewise polynomials of degree less than r when $r>2$. The vorticity approximation subspace $\mathscr{V}_h$ is defined to be one or three copies of S_h depending on whether $d=2$ or $d=3$, respectively.

The least-squares method involving the vorticity variable is defined from (28). We start by considering the quadratic form

$$\begin{aligned}\ll (q,\boldsymbol{v},\phi),(r,\boldsymbol{w},\psi) \gg_2 \equiv & (\nabla\times\phi+\boldsymbol{\nabla} q, \nabla\times\psi+\boldsymbol{\nabla} r)_{-1}+(\nabla\times\boldsymbol{v}-\phi, \nabla\times\boldsymbol{w}-\psi)\\ & +(\boldsymbol{\nabla}\cdot\boldsymbol{v}+\bar{\gamma}q, \boldsymbol{\nabla}\cdot\boldsymbol{w}+\bar{\gamma}r). \qquad (52)\end{aligned}$$

Here $(\cdot,\cdot)_{-1}$ is the quadratic form corresponding to the norm on $(H^{-1}(\Omega))^d$. Although the above form corresponds exactly to the the right-hand side of the a priori estimate provided in Theorem 2, it is not computationally useful. This is because evaluation of the inner product $(\boldsymbol{v},\boldsymbol{w})_{-1}$ requires the solution of the continuous problem, i.e., $(\boldsymbol{v},\boldsymbol{w})_{-1}=(\boldsymbol{x},\boldsymbol{w})$ where $\boldsymbol{x}\in\mathbf{H}_0^1(\Omega)$ is the solution of

$$\mathbf{D}(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{v},\boldsymbol{y}) \quad \text{for all } \boldsymbol{y}\in\mathbf{H}_0^1(\Omega).$$

To circumvent this difficulty, we replace the $(\cdot,\cdot)_{-1}$ inner product above by a computable alternative (cf. [13]).

Following [13], we replace the $(\cdot,\cdot)_{-1}$ inner product by

$$(\tilde{\mathbf{T}}_h\cdot,\cdot)\equiv((h^2\mathbf{I}+\mathbf{B}_h)\cdot,\cdot), \qquad (53)$$

where $\mathbf{B}_h$ is a preconditioner as discussed in the previous section and $\mathbf{I}$ denotes the identity operator. For convenience, we require that Π_h and $\mathscr{V}_h$ be subspaces of $H^1(\Omega)$ and $(H^1(\Omega))^{2d-3}$, respectively, although this restriction can be relaxed (see Remark 4). We then define the least-squares form

$$\begin{aligned}\ll (q,\boldsymbol{v},\phi),(r,\boldsymbol{w},\psi) \gg_3 \equiv & (\tilde{\mathbf{T}}_h(\nabla\times\phi+\boldsymbol{\nabla} q), \nabla\times\psi+\boldsymbol{\nabla} r)+(\nabla\times\boldsymbol{v}-\phi, \nabla\times\boldsymbol{w}-\psi)\\ & +(\boldsymbol{\nabla}\cdot\boldsymbol{v}+\bar{\gamma}q, \boldsymbol{\nabla}\cdot\boldsymbol{w}+\bar{\gamma}r). \qquad (54)\end{aligned}$$

The least-squares approximation to (28) is given by the triple $(P,\boldsymbol{U},\Theta)\in\Pi_h\times\mathbf{V}_h\times\mathscr{V}_h$ satisfying

$$\ll (P,\boldsymbol{U},\Theta),(R,\boldsymbol{W},\Psi) \gg_3 = (\tilde{\mathbf{T}}_h\boldsymbol{F}, \nabla\times\Psi+\boldsymbol{\nabla} R) \qquad (55)$$

for all $(R,\boldsymbol{W},\Psi)$ in $\Pi_h\times\mathbf{V}_h\times\mathscr{V}_h$. Let $\boldsymbol{x}\in(L^2(\Omega))^d$, $\boldsymbol{v}\in\mathbf{H}_0^1(\Omega)$ and $\boldsymbol{W}\in\mathbf{V}_h$ satisfy (15). Then

$$\begin{aligned}(\boldsymbol{x},\boldsymbol{v})^2 & \leqslant 2((\boldsymbol{x},\boldsymbol{v}-\boldsymbol{W})^2+(\boldsymbol{x},\boldsymbol{W})^2)\\ & \leqslant Ch^2 \parallel \boldsymbol{x} \parallel^2 \parallel \boldsymbol{v} \parallel_1^2 +2(\boldsymbol{x},\boldsymbol{W})^2.\end{aligned}$$

Thus by (35) and (37),

$$\| \boldsymbol{x} \|_{-1}^2 \leqslant C(h^2 \| \boldsymbol{x} \|^2 + \| \boldsymbol{x} \|_{-1,h}^2) \leq C(\tilde{\mathbf{T}}_h \boldsymbol{x}, \boldsymbol{x}). \tag{56}$$

Combining (56) and Theorem 2 implies that if $\boldsymbol{F} \in (L^2(\Omega))^d$, then the solution $(P, \boldsymbol{U}, \Theta)$ of (55) exists and is unique. The following theorem shows that the approximate solution $(P, \boldsymbol{U}, \Theta)$ is close to the solution triple $(p, \boldsymbol{u}, \theta)$ satisfying (28).

Theorem 4. *Let $(P, \boldsymbol{U}, \Theta)$ solve (55) with $\mathbf{B}_h$ satisfying (37). Let $(p, \boldsymbol{u}, \theta)$ be the solution of (28) where Π_h and $\mathscr{V}_h$ are as discussed above and $\mathbf{V}_h$ is as described in the Section 3(with $r \geqslant 2$). Assume that $\boldsymbol{F} \in (L^2(\Omega))^d$ and that the solution $(p, \boldsymbol{u}, \theta)$ is in $H^{r-1}(\Omega) \times (H^r(\Omega) \cap H_0^1(\Omega))^d \times (H^{r-1}(\Omega))^{2d-3}$. Then*

$$\| \boldsymbol{U} - \boldsymbol{u} \|_1 + \| P - p \| + \| \Theta - \theta \| \leqslant Ch^{r-1}(\| \boldsymbol{u} \|_r + \| p \|_{r-1} + \| \theta \|_{r-1}).$$

Proof. Let $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{E}}$ be as in the proof of Theorem 3. Let $\tilde{P}$ in Π_h and $\tilde{\Theta}$ in $\mathscr{V}_h$ satisfy

$$\| p - \tilde{P} \| + h \| p - \tilde{P} \|_1 \leqslant Ch^{r-1} \| p \|_{r-1}$$

and

$$\| \theta - \tilde{\Theta} \| + h \| \theta - \tilde{\Theta} \|_1 \leqslant Ch^{r-1} \| \theta \|_{r-1} .$$

Define $\tilde{e} = \tilde{P} - p$ and $\tilde{\mathscr{E}} = \tilde{\Theta} - \Theta$. As in the proof of Theorem 3, it follows from (55), (56) and Theorem 2 that

$$\begin{aligned} &\| \tilde{e} \|^2 + \| \tilde{\boldsymbol{E}} \|_1^2 + \| \tilde{\mathscr{E}} \|^2 \\ &\quad \leqslant C \ll (\tilde{P} - p, \tilde{\boldsymbol{U}} - \boldsymbol{u}, \tilde{\Theta} - \theta), (\tilde{P} - p, \tilde{\boldsymbol{U}} - \boldsymbol{u}, \tilde{\Theta} - \theta) \gg_3 . \end{aligned}$$

It is immediate that

$$\begin{aligned} &\| \nabla \times (\tilde{\boldsymbol{U}} - \boldsymbol{u}) - (\tilde{\Theta} - \theta) \| + \| \boldsymbol{\nabla} \cdot (\tilde{\boldsymbol{U}} - \boldsymbol{u}) + \tilde{\gamma}(\tilde{P} - p) \| \\ &\quad \leqslant Ch^{r-1}(\| \boldsymbol{u} \|_r + \| p \|_{r-1} + \| \theta \|_{r-1}). \end{aligned}$$

For the remaining term we have by (37),

$$\begin{aligned} &(\tilde{\mathbf{T}}_h(\nabla \times (\tilde{\Theta} - \theta) + \nabla(\tilde{P} - p)), \nabla \times (\tilde{\Theta} - \theta) + \nabla(\tilde{P} - p)) \\ &\quad \leqslant C(h^2 \| \nabla \times (\tilde{\Theta} - \theta) + \nabla(\tilde{P} - p) \|^2 + \| \nabla \times (\tilde{\Theta} - \theta) + \nabla(\tilde{P} - p) \|_{-1,h}). \end{aligned} \tag{57}$$

We clearly have that

$$h^2 \| \nabla \times (\tilde{\Theta} - \theta) + \nabla(\tilde{P} - p) \|^2 \leqslant Ch^{2r-2}(\| \theta \|_{r-1}^2 + \| p \|_{r-1}^2).$$

Moreover

$$\| \nabla \times (\tilde{\Theta} - \theta) \|_{-1,h} \leqslant \| \tilde{\Theta} - \theta \| \leqslant Ch^{r-1} \| \theta \|_{r-1}$$

and

$$\| \nabla(\tilde{P} - p) \|_{-1,h} \leqslant \| \tilde{P} - p \| \leqslant Ch^{r-1} \| p \|_{r-1} .$$

The theorem follows combining the above inequalities.

Remark 4. Following [13], if one replaces $\boldsymbol{F}$ by $\mathbf{Q}_h\boldsymbol{F}$ in (55) where $\mathbf{Q}_h$ denotes the $(L^2(\Omega))^d$ projector onto $\mathbf{V}_h$ then one gets a method which is stable for solutions with less regularity. Assume that the solution $(p,\boldsymbol{u},\theta)$ is in $H^{s-1}(\Omega)\times(H^s(\Omega)\cap H_0^1(\Omega))^d\times(H^{s-1}(\Omega))^{2d-3}$ for $1\leqslant s\leqslant r$. Then if $(P,\boldsymbol{U},\Theta)$ is defined replacing $\boldsymbol{F}$ by $\mathbf{Q}_h\boldsymbol{F}$ in (55),

$$\| \boldsymbol{U}-\boldsymbol{u}\|_1+\| P-p\|+\|\Theta-\theta\| \leqslant Ch^{s-1}(\|\boldsymbol{u}\|_s+\| p\|_{s-1}+\|\theta\|_{s-1}).$$

The assumption $\boldsymbol{F}\in(L^2(\Omega))^2$ can be relaxed to $\boldsymbol{F}\in\mathbf{H}^{-1}(\Omega)$.

Remark 5. The first term in (54) can be expressed in the notation of the previous section. More precisely, let $\nabla_h\times:(L^2(\Omega))^d\mapsto\mathbf{V}_h$ be the operator defined by

$$(\nabla_h\times\boldsymbol{v},\boldsymbol{U})=(\boldsymbol{v},\nabla\times\boldsymbol{U})\qquad\text{for all } \boldsymbol{U}\in\mathbf{V}_h.$$

Then,

$$\begin{aligned}(\tilde{\mathbf{T}}_h(\nabla\times\phi+\boldsymbol{\nabla} q),\nabla\times\psi+\boldsymbol{\nabla} r) =&h^2(\nabla\times\phi+\boldsymbol{\nabla} q,\nabla\times\psi+\boldsymbol{\nabla} r)\\ &+(\mathbf{B}_h\nabla_h\times\phi+\nabla_h q,\nabla_h\times\psi+\nabla_h r).\end{aligned}$$

Remark 6. It is also possible to relax the continuity conditions on the spaces Π_h and $\mathscr{V}_h$ by adding terms to the form. If either of these spaces consist of discontinuous piecewise polynomials, then one can use a least-squares method based on the form

$$\begin{aligned}\ll (q,\boldsymbol{v},\phi),(r,\boldsymbol{w},\psi)\gg_4\equiv\ &(\mathbf{B}_h(\nabla_h\times\phi+\nabla_h q),\nabla_h\times\psi+\nabla_h r)\\ &+h^2(\nabla\times\phi+\boldsymbol{\nabla} q,\nabla\times\psi+\boldsymbol{\nabla} r)_h+h<[[\phi]],[[\psi]]>_h\\ &+h<[[q]],[[r]]>_h+(\nabla\times\boldsymbol{v}-\phi,\nabla\times\boldsymbol{w}-\psi)\\ &+(\boldsymbol{\nabla}\cdot\boldsymbol{v}+\bar{\gamma}q,\boldsymbol{\nabla}\cdot\boldsymbol{w}+\bar{\gamma}r).\end{aligned}$$

A result analogous to Theorem 4.1 holds for the corresponding least-squares method. We get a method which is stable and convergent in lower norms if $\boldsymbol{F}$ is replaced by $\mathbf{Q}_h\boldsymbol{F}$ in the $h^2(\boldsymbol{F},\nabla\times\psi+\boldsymbol{\nabla} r)_h$ term on the right-hand side corresponding to this form.

5. Implementation and iterative solution of the least-squares system

In this section we consider the implementation aspects of the least-squares methods described in the preceding two sections. As already noted, the algebraic systems associated with these least-squares forms are full. Nevertheless, we shall see that effective preconditioned iterative schemes can be generated which converge rapidly to the desired discrete solution and avoid assembly of the full matrix. To be specific, we consider problem (39). The case of (55) is similar.

Let Π_h and $\mathbf{V}_h$ consist respectively of discontinuous piecewise constant functions and continuous piecewise linear functions. The implementation of higher order spaces is completely analogous.

There are three major aspects involved in setting up the algebraic system and its subsequent solution by preconditioned iteration. All of these operations are performed with respect to a computational basis. Let $\{\theta_i\}$ and $\{\Theta_i\}$ denote the local nodal bases for Π_h and $\mathbf{V}_h$ respectively. These can be combined into a global basis $\{\Psi_j\} = \{(\theta_i, 0)\} \cup \{(0, \Theta_i)\} = \{(\phi_j, \Phi_j)\}$ for $\Pi_h \times \mathbf{V}_h$. Let n_1 and n_2 respectively denote the dimension of Π_h and $\mathbf{V}_h$ and set $n = n_1 + n_2$.

We seek the solution of the discrete problem

$$\tilde{M}\tilde{c} = \tilde{d}, \tag{58}$$

where

$$\tilde{M}_{ij} = \ll \Psi_i, \Psi_j \gg_1 .$$

The right-hand side of (58) is given by

$$\tilde{d}_i = (\mathbf{B}_h \boldsymbol{F}, -\boldsymbol{\nabla}_h^2 \Phi_i + \nabla_h \phi_i) + h^2(\boldsymbol{F}, -\boldsymbol{\nabla}^2 \Phi_i + \boldsymbol{\nabla} \phi_i)_h,$$

for $i = 1, \ldots, n$.

In previous sections of this paper, we defined $\mathbf{B}_h$ as a symmetric positive definite operator on $\mathbf{V}_h$. In terms of the implementation, the preconditioner can be more naturally thought of in terms of a $n_2 \times n_2$ matrix N. The operator $\mathbf{B}_h$ is defined in terms of this matrix as follows. Fix $\boldsymbol{V} \in \mathbf{V}_h$ and expand

$$\mathbf{B}_h \boldsymbol{V} = \sum_i G_i \Theta_i.$$

Then,

$$NG = \tilde{G} \tag{59}$$

where

$$\tilde{G}_i = (\boldsymbol{V}, \Theta_i). \tag{60}$$

The operator $\mathbf{B}_h$ is a good preconditioner for $\mathbf{T}_h$ provided that the matrix $N^{-1}\tilde{N}$ has small condition number. Here $\tilde{N}$ is the stiffness matrix for the form $\mathbf{D}(\cdot,\cdot)$, i.e.,

$$\tilde{N}_{ij} = \mathbf{D}(\Theta_i, \Theta_j).$$

The matrix N need not explicitly appear in the computation of the action of the preconditioner. Instead, one often has a process or algorithm which acts on the vector $\tilde{G}$ and produces the vector G, i.e., computes $N^{-1}\tilde{G}$. Thus, the practical application of the preconditioner on a function in $\boldsymbol{V}$ reduces to a predefined algorithm for computing the action of N^{-1} and the evaluation of the vector $\tilde{G}$ defined by (60).

The first step in computing the coefficient vector $\tilde{c}$ solving (58) is to compute the right-hand side vector $\tilde{d}$. We shall assume that some method for computing integrals of the form

$$\int_{\tau_i} \boldsymbol{F} \cdot \eta \, \mathrm{d}x \tag{61}$$

is available when η is a vector valued polynomial. Here τ_i is a triangle in the mesh. Thus, we can compute the data $(\boldsymbol{F}, \Theta_j)$ for $j = 1, \ldots, n_2$ from which $\mathbf{B}_h \boldsymbol{F}$ can be computed as discussed above.

With $\mathbf{B}_h F$ known, the right-hand side vector $\tilde{d}$ reduces to more integrals of the form of (61) and the integration of polynomials over the triangles $\{\tau_i\}$. These actions are local in the sense that the result for each Θ_j only involves the triangles containing the support of Θ_j.

The next action required for the implementation of the preconditioning iteration is the application of $\tilde{M}$ to arbitrary vectors $c \in R^n$. The vector c represents the coefficients of a function pair

$$(\delta, V) = \sum_{i=1}^{n} c_i(\phi_i, \Phi_i).$$

We are required to evaluate

$$\begin{aligned}(Mc)_j &= \ll (\delta, V), (\phi_j, \Phi_j) \gg_1 \\ &= (\mathbf{B}_h(-\boldsymbol{\nabla}_h^2 V + \nabla_h \delta), -\boldsymbol{\nabla}_h^2 \Phi_j + \nabla_h \phi_j) + h^2(-\boldsymbol{\nabla}^2 \boldsymbol{v} + \boldsymbol{\nabla}\delta, -\boldsymbol{\nabla}^2 \Phi_j + \boldsymbol{\nabla}\phi_j)_h \\ &\qquad + h < [[\delta]], [[\phi_j]] >_h + (\boldsymbol{\nabla} \cdot \boldsymbol{v} + \gamma\delta, \boldsymbol{\nabla} \cdot \Phi_j + \gamma\phi_j),\end{aligned} \tag{62}$$

for $j = 1, \ldots, n$. The data for the preconditioner evaluation is

$$(-\boldsymbol{\nabla}_h^2 V + \nabla_h \delta, \Theta_i) = \mathbf{D}(V, \Theta_i) - (\delta, \boldsymbol{\nabla} \cdot \Theta_i)$$

and reduces to integrals of polynomials over the triangles $\{\tau_l\}$. After application of the preconditioning process, the coefficients for the function $\mathbf{B}_h(-\boldsymbol{\nabla}_h^2 V + \nabla_h \delta)$ are known. All quantities appearing on the right-hand side of (62) can then be computed by integrals of polynomials over the triangles and faces. The work required for computing $(MG)_j$, $j = 1, \ldots, n$ is on the order of n plus the work involved in applying the preconditioning process.

The final step required for a preconditioned iteration is the action of an appropriate preconditioning matrix M. Note that by Theorem 2.1 and the quasi-uniformity of the mesh, there exist positive constants C_0 and C_1 not depending on h satisfying

$$\begin{aligned}C_0(\| Q \|^2 + \| V \|_1^2) &\leqslant \ll (Q, V), (Q, V) \gg_1 \\ &\leqslant C_1(\| Q \|^2 + \| V \|_1^2).\end{aligned} \tag{63}$$

The above inequalities hold for all (Q, V) in the product space $\Pi_h \times \mathbf{V}_h$. Consequently, the task of defining a preconditioner for $\tilde{M}$ is the same as finding a preconditioner for the block diagonal system

$$\begin{pmatrix} \tilde{N}_0 & 0 \\ 0 & \tilde{N} \end{pmatrix},$$

where $\tilde{N}_0$ is the Gram matrix

$$(\tilde{N}_0)_{ij} = (\phi_i, \phi_j) \text{ for } i, j = 1, \ldots, n_1.$$

Define

$$M = \begin{pmatrix} h^2 I & 0 \\ 0 & N \end{pmatrix},$$

where I denotes the identity matrix on R^{n_1}. It follows from (63) that the condition number of $M^{-1}\tilde{M}$ is uniformly bounded independent of h. Thus, the reduction rate per step in, for example, the preconditioned conjugate gradient iteration can be bounded independently of h. The application of M^{-1} involves weighting the Π_h data by h^{-2} and the application of the preconditioning process to the $\mathbf{V}_h$ data. The work involved in one step of the conjugate gradient iteration is on the order of n plus twice the cost of the application of the preconditioning process N^{-1}.

The above discussion is summarized in the following algorithm for the solution of (39).

Algorithm 1. The solution of (39) involves the following two steps.

(1) The computation of the right-hand side vector $\tilde{d}$ of (58).

   (a) Compute $\{(\boldsymbol{F},\Theta_j)\}$ by assembling the quantities given in (61) (Work $\sim$ O(n_2)).

   (b) Solve the preconditioning problem (60) with data $\tilde{G}=\{(\boldsymbol{F},\Theta_j)\}$. This gives the coefficients for $\mathbf{B}_h\boldsymbol{F}$.

   (c) Compute $\tilde{d}$. This involves additional integrals of the form (61) (Work $\sim$ O(n_2)).

(2) Compute $\tilde{c}$ solving (58) by preconditioned conjugate gradient iteration. The entries of $\tilde{c}$ are the coefficients for the solution of (39). Each iteration step requires the evaluation of the matrix operator and preconditioner.

   (a) Evaluation of the matrix operator on a given vector $\{c_i\}$ corresponding to the function pair (δ, V).

      (i) Compute the data $\tilde{G}=\{\mathbf{D}(V,\Theta_i)-(\delta,\boldsymbol{\nabla}\cdot\Theta_i)\}$ for the $\mathbf{B}_h$ evaluation (Work $\sim$ O(n)).

      (ii) Apply the preconditioning process to obtain the coefficients for $\mathbf{B}_h(-\boldsymbol{\nabla}_h^2 V + \nabla_h\delta)$.

      (iii) Compute the quantities $(Mc)_j$ $j=1,\ldots,n$ given in (62) (Work $\sim$ O(n)).

   (b) Evaluation of the block preconditioner on a given vector $\{\tilde{c}_j,\ j=1,\ldots,n\}$. This involves multiplying the first n_1 coefficients by h^2 (Work $\sim$ O(n_1)) and evaluating the action of the preconditioning process.

Remark 7. As already noted, the appearance of $\mathbf{B}_h$ gives rise to a full lower right hand block in the stiffness matrix for the least-squares operator. Consequently, it is not feasible to assemble the matrix. However, some efficiency may be gained by assembling parts of the matrix. For example, it would be feasible to assemble a sparse matrix for all terms in (62) excluding the one involving $\mathbf{B}_h$. Additionally, to more efficiently compute the first term of (62) one could assemble the matrices $\{\mathbf{D}(\Theta_j,\Theta_i)\}$ and $\{(\nabla_h\phi_j,\Theta_i)\}$.

Remark 8. By Theorem 2.2 and the quasi-uniformity of the mesh, there exist positive constants C_0 and C_1 not depending on h satisfying

$$\begin{aligned} C_0(\| Q \|^2 + \| V \|_1^2 + \| \Phi \|^2) &\leqslant \ll (Q,V,\Phi),(Q,V,\Phi) \gg_3 \\ &\leqslant C_1(\| Q \|^2 + \| V \|_1^2 + \| \Phi \|^2). \end{aligned}$$

The above inequalities hold for all (Q,V,Φ) in the product space $\Pi_h \times \mathbf{V}_h \times \mathcal{V}_h$. Consequently, to solve (55), we can use a preconditioner involving three diagonal blocks. The blocks corresponding to Π_h and $\mathcal{V}_h$ can be taken to be h^2 times the identity. For the $\mathbf{V}_h$ block, we once again use N. The rest of the implementation in the case of (55) is analogous to the implementation for (39) described above.

References

[1] O. Axelsson, A generalized conjugate gradient, least squares method, *Numer. Math.* **51** (1987) 209–228.

[2] I. Babuška, On the Schwarz algorithm in the theory of differential equations of mathematical physics, *Tchecosl. Math. J.* **8** (1958) 328–342 (in Russian).

[3] G. Birkhoff and A. Schoenstadt, Eds., *Elliptic Problem Solvers II* (Academic Press, New York, 1984).

[4] P.B. Bochev and M.D. Gunzburger, Accuracy of least-squares methods for the Navier–Stokes equations, *Comput. Fluids*, **22** (1993) 549–563.

[5] P.B. Bochev and M.D. Gunzburger, Analysis of least-squares finite elements methods for the Stokes equations. *Math. Comp.*, **63** (1994) 479–505.

[6] J.H. Bramble, *Multigrid Methods,* Pitman Research Notes in Mathematics Series (Longman Scientific & Technical, London. Copublished with Wiley, New York, 1993).

[7] J.H. Bramble, Interpolation between Sobolev spaces in Lipschitz domains with an application to multigrid theory, *Math. Comp.*, **64** (1995) 1359–1366.

[8] J.H. Bramble and J.E. Pasciak, New convergence estimates for multigrid algorithms, *Math. Comp.* **49** (1987) 311–329.

[9] J.H. Bramble and J.E. Pasciak, A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems, *Math. Comp.* **50** (1988) 1–17.

[10] J.H. Bramble and J.E. Pasciak, New estimates for multigrid algorithms including the V-cycle, *Math. Comp.* **60** (1993) 447–471.

[11] J.H. Bramble, J.E. Pasciak, J. Wang and J. Xu, Convergence estimates for product iterative methods with applications to domain decompositions, *Math. Comp.* **57** (1991) 1–21.

[12] J.H. Bramble and R. Scott, Simultaneous approximations in scales of Banach spaces, *Math. Comp.* **32** (1978) 947–954.

[13] J.H. Bramble, R.D. Lazarov and J.E. Pasciak, A least-squares approach based on a discrete minus one inner product for first order systems, Brookhaven National Laboratory Report BNL-60624, 1994.

[14] S.C. Brenner and L.R. Scott, *The mathematical Theory of Finite Element Methods* (Springer, New York, 1994).

[15] F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising form Lagrange multipliers, *R.A.I.R.O.* (1974) 129–151.

[16] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer, New York, 1991).

[17] T.F. Chan, R. Glowinski, J. Periaux and O.B. Widlund, Eds., *Third Internat. Symp. on Domain Decomposition Methods for Partial Differential Equations* (SIAM, Philadelphia, PA, 1990).

[18] T. Chan, R. Glowinski, J. Periaux and O.B. Widlund, Eds., *Domain Decomposition Methods* (SIAM, Philadelphia, PA. 1989).

[19] T.F. Chen, On the least-squares approximations to compressible flow problems, *Numer. Methods PDE's* **2** (1986) 207–228.

[20] T.F. Chen, and G.J. Fix, Least-squares finite element simulation of transonic flows, *Appl. Numer. Math.* **2** (1986) 399–408.

[21] M. Dryja and O. Widlund, An additive variant of the Schwarz alternating method for the case of many subregions, Technical Report, Courant Institute of Mathematical Sciences, 339, 1987.

[22] V. Girault and P.A. Raviart, *Finite Element Approximation of the Navier–Stokes Equations*, Lecture Notes in Math., Vol. 749 (Springer, New York, 1981).

[23] R. Glowinski, G.H. Golub, G.A. Meurant and J. Periaux, Eds., *First Internat. Symp. on Domain Decomposition Methods for Partial Differential Equations* (SIAM, Philadelphia, PA, 1988).

[24] R. Glowinski, Y.A. Kuznetzov, G. Meurant and J. Periaux, Eds., *Fourth Internat. Symp. on Domain Decomposition Methods for Partial Differential Equations* (SIAM, Philadelphia, PA, 1991).

[25] P. Grisvard, *Elliptic Problems in Nonsmooth Domains* (Pitman, Boston, 1985).

[26] T.J.R. Hughes and L.P. Franca, A new finite element formulation for computational fluid dynamics. VII. The Stokes problems with various well-posed boundary conditions: symmetric formulation that converges for all velocity pressure spaces, *Comput. Methods. Appl. Mech. Eng.* **65** (1987) 85–96.

[27] T.J.R. Hunhes and L.P. Franca, 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. Methods Appl. Mech. Eng.* **59** (1986) 85–99.

[28] B.N. Jiang and C. Chang, Least-squares finite elements for the Stokes problem, *Comput. Mech. Eng.* **81** (1990) 13–37.

[29] O.A. Ladyzhenskaya, *The Methematical Theory of Viscous Incompressible Flows* (Gordon and Breach, London, 1969).

[30] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques* (Academia, Prague, 1967).

[31] P.E. Bjørstad and O.B. Widlund, Solving elliptic problems on regions partitioned into substructures, in: G. Birkhoff and A. Schoenstadt, Eds., *Elliptic problem Solvers II* (Academic Press, New York, 1984) 245–256.

[32] T. Rusten and R. Winther, A preconditoned iterative methods for saddle point problems, *SIAM J. Matrix Anal. Appl.* **13** (1992) 489–512.

[33] R. Temam, *Navier–Stokes Equations* (North-Holland, New York, 1977).

[34] P. Vassilevski, Hybrid V-cycle algebraic multilevel method for visco-elastic fluid flow problems, 1987.

[35] K.C. Wang and G.F. Carrey, A least-squares method for visco-elastic fluid flow problems. *Internat. J. Numer. Methods Fluids*, **17** (1993) 943–953.