NOTES ON A SADDLE POINT REFORMULATION OF MIXED VARIATIONAL PROBLEMS

CONSTANTIN BACUTA, DANIEL HAYES, AND JACOB JACAVAGE

Abstract. We summarize some general ideas regarding approximation of mixed variational problems using saddle point reformulation. We consider the concepts of optimal and almost optimal (or α) test norm and provide estimates for the continuity and stability constants. A preconditioning strategy for solving the discrete mixed formulations is used in combination with the special test norms. We further provide a choice for a discrete trial space, that depends on the choice of a standard test space and leads to discrete stability, when using the appropriate test norm. Examples to illustrate how the stability of the saddle point discretization can be improved using special test norms are included.

1. Introduction

Saddle point reformulations for the Petrov-Galerkin method has become a common methodology in dealing with discretization of PDEs, especially for the Discontinuous Petrov-Galerkin (DPG) community and the Saddle Point Least Squares (SPLS) group. The main idea is to use an auxiliary variable that represents the residual of the original equation on the test space and another simple equation that leads to a (square) saddle point system that is more suitable for analysis and discretization. It turns out that in the saddle point reformulation, the main variable of interest is also the least squares solution of the representation of the original equation on the test space. It is very possible that this idea was used in many other particular discretizations of various variational problems a long time ago, see e.g., [10, 21, 23, 26]. We try to summarize and generalize the main ideas of the SP approach in an abstract general setting. Many of the results regarding this reformulation are common to both the DPG approach [15, 19, 22, 24, 25, 28] and the SPLS approach developed in [6, 7, 8, 11].

We also expand on the concept of optimal test norm [18, 20, 21, 22, 25, 26, 28] that seems to be often involved in the DPG methodology and combine it with a general preconditioning technique, introduced in [6], in order to improve the stability of the saddle point discrete formulations.

2000 *Mathematics Subject Classification.* 74S05, 74B05, 65N22, 65N55.

Key words and phrases. least squares, saddle point systems, mixed methods, optimal test norm, Uzawa conjugate gradient, preconditioning.

The work was supported by NSF-DMS 2011615.

The goal of the paper is to present some of the common concepts and approaches of SPLS discretization that can be used not only by the DPG-SPLS community, but also by any practitioner interested in finite element approximation of variational formulations. We investigate the SP reformulation and discretization of the following general Petrov-Galerkin problem: Given $F \in V^*$, find $p \in Q$ such that

$$b(v, p) = \langle F, v \rangle \quad \text{for all } v \in V, \tag{1.1}$$

where V and Q are Hilbert spaces and $b(\cdot, \cdot)$ is a continuous bilinear form on $V \times Q$ satisfying an $\inf - \sup$ condition.

The paper is organized as follows. In Section 2, we introduce the notation and present the SPLS formulation and special test norms at the continuous level. Section 3 provides the approximation theory and proposes a solver for the SPLS discretization. We present the SPLS preconditioning theory and an iterative solver in Section 4. In Section 5 we consider a general choice of a discrete trial space that depends on the choice of a discrete test space, which is problem dependent, but is always compatible with the trial space. In this section, we also analyze the stability of the proposed discrete spaces using the special test norms. We consider one example of SPLS formulation with an optimal test norm and one example with an almost optimal norm in Section 6. The proofs of the two lemmas regarding special test norms are included in the Appendix.

2. The notation and the general SPLS approach

We now review the main ideas and concepts for the SPLS discretion of a general mixed variational formulation.

2.1. The abstract variational formulation at the continuous level.

We consider the Petrov-Galerkin formulation (1.1). We assume that the inner products $a_0(\cdot, \cdot)$ and $(\cdot, \cdot)_Q$ induce the norms $|\cdot|_V = |\cdot| = a_0(\cdot, \cdot)^{1/2}$ and $\|\cdot\|_Q = \|\cdot\| = (\cdot, \cdot)_Q^{1/2}$. We denote the dual of V by V^* and the dual pairing on $V^* \times V$ by $\langle \cdot, \cdot \rangle$. We assume that $b(\cdot, \cdot)$ is a continuous bilinear form on $V \times Q$ satisfying the $\sup - \sup$ condition

$$\sup_{p \in Q} \sup_{v \in V} \frac{b(v, p)}{|v| \, \|p\|} = M < \infty, \tag{2.1}$$

and the $\inf - \sup$ condition

$$\inf_{p \in Q} \sup_{v \in V} \frac{b(v, p)}{|v| \, \|p\|} = m > 0. \tag{2.2}$$

With the form b, we associate the operators $B : V \to Q$ and $B^* : Q \to V^*$ defined by

$$(Bv, q)_Q = b(v, q) = \langle B^* q, v \rangle \quad \text{for all } v \in V, q \in Q.$$

We define V_0 to be the kernel of B, i.e.,

$$V_0 := Ker(B) = \{v \in V |\ Bv = 0\}.$$

Under assumptions (2.1) and (2.2), the operator B is a bounded surjective operator from V to Q, and V_0 is a closed subspace of V. With the inner product on V, we associate the operator $A_0 : V \to V^*$ defined by

$$\langle A_0 u, v\rangle = a_0(u, v) \quad \text{for all } u, v \in V.$$

We will also assume that the data $F \in V^*$ satisfies the *compatibility condition*

$$\langle F, v\rangle = 0 \quad \text{for all } v \in V_0 = Ker(B). \tag{2.3}$$

The following result describes the well posedness of (1.1) and can be used at the continuous and discrete levels, see e.g., [1, 3, 13, 14].

Proposition 2.1. *If the form $b(\cdot, \cdot)$ satisfies* (2.1) *and* (2.2)*, and the data $F \in V^*$ satisfies the compatibility condition* (2.3)*, then the problem* (1.1) *has unique solution that depends continuously on the data F.*

It is also stated in a few papers, see [10, 11, 12, 21], that, under the *compatibility condition* (2.3), solving the mixed problem (1.1) reduces to solving a standard saddle point formulation: Find $(w, p) \in V \times Q$ such that

$$\begin{array}{lclll} a_0(w, v) & + & b(v, p) & = \langle F, v\rangle & \text{for all } v \in V, \\ b(w, q) & & & = 0 & \text{for all } q \in Q. \end{array} \tag{2.4}$$

In fact, p is the unique solution of (1.1) *if and only if* $(w = 0, p)$ solves (2.4), and the result remains valid if the form $a_0(\cdot, \cdot)$ in (2.4) is replaced by any other symmetric bilinear form $a(\cdot, \cdot)$ on V that leads to an equivalent norm on V.

The Schur complement associated with the SP system (2.4) is $S : Q \to Q$, defined by $S := BA_0^{-1}B^*$. Furthermore, see e.g., [3], S is a bounded symmetric operator on Q, and the spectrum of S satisfies

$$m^2, M^2 \in \sigma(S) \subset [m^2, M^2].$$

In general, for a symmetric positive definite operator S on a Hilbert space $(Q, (\cdot, \cdot)_Q)$ we define $\|q\|_S := (Sq, q)_Q^{1/2}$, see Sections 3.1 and 4.2.

2.2. **The concept of optimal test norm.** If we assume that the operator $B : V \to Q$ is injective ($V_0 = Ker(B) = \{0\}$) then, as in [18, 20, 21, 22, 26], we can define the following operator dependent norm on V,

$$|v|_{opt} := \sup_{p \in Q} \frac{b(v, p)}{\|p\|} = \sup_{p \in Q} \frac{(Bv, p)_Q}{\|p\|} = \|Bv\|\ .$$

Since B is a bounded bijective operator between Hilbert spaces, we have that $|\cdot|_{opt}$ is indeed an equivalent norm on V. We will refer to this norm on V as the *optimal test norm.*

Lemma 2.2. *Assume that the form $b(\cdot,\cdot)$ satisfies* (2.1) *and* (2.2) *and B is injective. By considering the optimal norm $|v|_{opt} = \|Bv\|$ on V, we have that both the continuity constant M_{opt} and the* $\inf-\sup$ *constant m_{opt} are equal to* 1. *Consequently, by replacing the form $a_0(\cdot,\cdot)$ in* (2.4) *with the inner product induced by the optimal test norm $a_{opt}(u,v) := (Bu,Bv)_Q$, we obtain that the Schur complement of the new saddle point system is the identity operator. Hence, the stability of the new saddle point formulation is optimal.*

For the proof, please see the appendix.

2.3. **The concept of almost optimal test norm.** In the case $Ker(B) \neq \{0\}$, we introduce the notion of an *almost optimal test norm*, or α *norm* $|\cdot|_\alpha$. It is defined as the norm induced by the inner product

$$a_\alpha(u,v) := \alpha^2 a_0(u,v) + (Bu,Bv)_Q, \tag{2.5}$$

where α is a positive parameter. The following estimates for the continuity constant M_α and the $\inf-\sup$ constant m_α can be deduced in a similar way that was done in [2] in the context of the Augmented Lagrangian method for Stokes type systems.

Lemma 2.3. *Assume that the form $b(\cdot,\cdot)$ satisfies* (2.1) *and* (2.2) *and consider the almost optimal norm $|v|_\alpha = \left(\alpha^2 a_0(v,v) + (Bv,Bv)_Q\right)^{1/2}$ on V. Then the corresponding continuity and* $\inf-\sup$ *constants satisfy*

$$m_\alpha^2 = \frac{m^2}{m^2+\alpha^2} \quad \text{and} \quad M_\alpha^2 = \frac{M^2}{M^2+\alpha^2}. \tag{2.6}$$

Consequently, by replacing the form $a_0(\cdot,\cdot)$ in (2.4) *with the inner product induced by the almost optimal test norm $a_\alpha(u,v) := \alpha^2 a_0(u,v) + (Bu,Bv)_Q$, we obtain that the condition number of the Schur complement S_α of the new saddle point system is given by*

$$\kappa(S_\alpha) = \frac{M^2}{m^2}\frac{m^2+\alpha^2}{M^2+\alpha^2}.$$

Furthermore, for $\alpha \in (0,m]$ we have $\kappa(S_\alpha) \in (1,2)$.

For the proof, please see the appendix.

Remark 2.4. *The connection with the Augmented Lagrangian method: When solving a Stokes type system, with the variational formulation given by* (2.4), *by the Augmented Lagrangian method, see e.g.,* [2], *we replace the form $a_0(\cdot,\cdot)$ in* (2.4) *with the inner product induced by*

$$a_\rho(u,v) := a_0(u,v) + \rho^2 (Bu,Bv)_Q, \ \text{where } \rho > 0 \text{ is a parameter.} \tag{2.7}$$

Since the inner product (2.7) *is a rescaling of the inner product* (2.5), *inverting or preconditioning the operators associated with these inner products has the same difficulty. Also, when solving a Stokes type system the compatibility condition* (2.3) *does not necessarily hold, and the variable w is an essential (non zero in general) variable. On the other hand, when solving* (2.4) *as an*

SPLS reformulation of (1.1)*, we have that w is an auxiliary variable that is zero at the continuous level due to* (2.3)*. This allows for standard saddle point discretization and enables the use of known solving techniques, such as Uzawa type algorithms.*

3. Saddle point least squares discretization

We assume that the inner product on V is given by the continuous bilinear form $a_0(\cdot,\cdot)$ that leads to the norm $a_0(\cdot,\cdot)^{1/2}$ on V. Let $V_h \subset V$ and $\mathcal{M}_h \subset Q$ be finite dimensional approximation spaces and A_h be the discrete version of the operator A_0, i.e., A_h satisfies

$$\langle A_h w_h, v_h\rangle = a_0(w_h, v_h) \quad \text{for all } w_h, v_h \in V_h.$$

We define the discrete operators $B_h : V_h \to \mathcal{M}_h$ and $B_h^* : \mathcal{M}_h \to V_h^*$ by

$$(B_h v_h, q_h)_Q = b(v_h, q_h) = \langle B_h^* q_h, v_h\rangle \quad \text{for all } v_h \in V_h, q_h \in \mathcal{M}_h.$$

Note that the operator B_h is defined using the inner product on $\mathcal{M}_h$ and not with the duality on $\mathcal{M}_h^* \times \mathcal{M}_h$. Thus, we can define the discrete Schur complement $S_h : \mathcal{M}_h \to \mathcal{M}_h$ as $S_h = B_h\, A_h^{-1} B_h^*$. We further assume the following discrete $\inf - \sup$ condition holds for the pair of spaces $(V_h, \mathcal{M}_h)$:

$$\inf_{p_h \in \mathcal{M}_h} \sup_{v_h \in V_h} \frac{b(v_h, p_h)}{|v_h|\,\|p_h\|} = m_h > 0. \tag{3.1}$$

As in the continuous case, it is known that the spectrum of S_h satisfies

$$m_h^2, M_h^2 \in \sigma(S_h) \subset [m_h^2, M_h^2],$$

where

$$M_h := \sup_{p_h \in \mathcal{M}_h} \sup_{v_h \in V_h} \frac{b(v_h, p_h)}{|v_h|\,\|p_h\|} \le M < \infty. \tag{3.2}$$

We define

$$V_{h,0} := \{v_h \in V_h \,|\, b(v_h, q_h) = 0, \quad \text{for all } q_h \in \mathcal{M}_h\} = Ker(B_h),$$

to be the kernel of the discrete operator B_h and $F_h \in V_h^*$ to be the restriction of F to V_h, i.e., $\langle F_h, v_h\rangle := \langle F, v_h\rangle$ for all $v_h \in V_h$.

In the case $V_{h,0} \subset V_0$, the compatibility condition (2.3) implies the discrete compatibility condition

$$\langle F, v_h\rangle = 0 \quad \text{for all } v_h \in V_{h,0}.$$

Hence, under assumption (3.1), the problem of finding $p_h \in \mathcal{M}_h$ such that

$$b(v_h, p_h) = \langle F, v_h\rangle,\ v_h \in V_h,\ \text{or } B_h^* p_h = F_h,\ \text{or } A_h^{-1} B_h^* p_h = A_h^{-1} F_h, \tag{3.3}$$

has a unique solution. In general, we might not have $V_{h,0} \subset V_0$. Consequently, even though the continuous problem (1.1) is well posed, the discrete problem (3.3) might not be well-posed. However, if the form $b(\cdot,\cdot)$ satisfies (3.1), then the problem of finding $(w_h, p_h) \in V_h \times \mathcal{M}_h$ satisfying

$$\begin{array}{llll} a_0(w_h, v_h) & + \quad b(v_h, p_h) & = \langle f, v_h\rangle & \text{for all } v_h \in V_h, \\ b(w_h, q_h) & & = 0 & \text{for all } q_h \in \mathcal{M}_h, \end{array} \tag{3.4}$$

does have a unique solution. Solving for p_h from (3.4), we obtain

$$S_h\, p_h = B_h(A_h^{-1}B_h^*)\, p_h = B_h A_h^{-1} F_h. \tag{3.5}$$

Since the Hilbert transpose of B_h is $B_h^T = A_h^{-1}B_h^*$, we note that (3.5) is the least squares formulation of (the last version) of (3.3). Thus, we call the component p_h of the solution (w_h, p_h) of (3.4) the saddle point *least squares* approximation of the solution p of the original mixed problem (1.1). The following error estimate for $\|p - p_h\|$ was proved in [11].

Theorem 3.1. *Let $b : V \times Q \to \mathbb{R}$ satisfy* (2.1) *and* (2.2) *and assume that $F \in V^*$ is given and satisfies* (2.3). *Assume that p is the solution of* (1.1) *and $V_h \subset V$, $\mathcal{M}_h \subset Q$ are chosen such that the discrete* $\inf - \sup$ *condition* (3.1) *holds. If (w_h, p_h) is the solution of* (3.4)*, then the following error estimate holds:*

$$\frac{1}{M}|w_h| \le \|p - p_h\| \le \frac{M}{m_h} \inf_{q_h \in \mathcal{M}_h} \|p - q_h\|. \tag{3.6}$$

The considerations made so far in this section remain valid if the form $a_0(\cdot,\cdot)$, as an inner product on V_h, is replaced by another inner product $a(\cdot,\cdot)$ which gives rise to an equivalent norm on V_h. Certainly, the definitions of A_h, S_h, M_h, and m_h will change accordingly with the new norm induced by the inner product $a(\cdot,\cdot)$. In particular, the error estimate (3.6) remains valid with the corresponding new definition for the constant m_h.

3.1. **An Uzawa CG iterative solver.**

In the previous sections, we discussed the possibility of having more than one norm or inner product on V. We will assume next that the inner product on V is given by a generic continuous bilinear form $a(\cdot,\cdot)$ that leads to an equivalent norm on V, $a(\cdot,\cdot)^{1/2}$. Note that a global linear system may be difficult to assemble or solve when $a_0(\cdot,\cdot)$ is replaced by $a(\cdot,\cdot)$ in (3.4). Nevertheless, we can solve (3.4) and avoid building a basis for $\mathcal{M}_h$ by using an Uzawa type algorithm, e.g., the Uzawa Conjugate Gradient (UCG) algorithm.

Algorithm 3.2. *(UCG) Algorithm*

Step 1: Choose any $p_0 \in \mathcal{M}_h$. **Compute** $u_1 \in V_h$, $q_1, d_1 \in \mathcal{M}_h$ *by*

$$\begin{array}{lll} a(u_1, v_h) & = \langle F, v_h\rangle - b(v_h, p_0) & \textit{for all } v_h \in V_h, \\ (q_1, q_h)_Q & = b(u_1, q_h) & \textit{for all } q_h \in \mathcal{M}_h, \qquad d_1 := q_1. \end{array}$$

Step 2: For $j = 1, 2, \ldots,$ **compute** $h_j, \alpha_j, p_j, u_{j+1}, q_{j+1}, \beta_j, d_{j+1}$ *by*

$$\begin{array}{lrll}
(\mathbf{UCG1}) & a(h_j, v_h) = & -\, b(v_h, d_j) & \textit{for all } v_h \in V_h \\
(\mathbf{UCG}\alpha) & \alpha_j = & -\,\dfrac{(q_j, q_j)_Q}{b(h_j, q_j)} & \\
(\mathbf{UCG2}) & p_j = & p_{j-1} + \alpha_j\ d_j & \\
(\mathbf{UCG3}) & u_{j+1} = & u_j + \alpha_j\ h_j & \\
(\mathbf{UCG4}) & (q_{j+1}, q_h)_Q = & b(u_{j+1}, q_h) & \textit{for all } q_h \in \mathcal{M}_h \\
(\mathbf{UCG}\beta) & \beta_j = & \dfrac{(q_{j+1}, q_{j+1})_Q}{(q_j, q_j)_Q} & \\
(\mathbf{UCG6}) & d_{j+1} = & q_{j+1} + \beta_j d_j. &
\end{array}$$

Note that the only inversions needed in the algorithm involve the form $a(\cdot, \cdot)$ in **Step 1** and (**UCG1**). In operator form, these steps become

$$u_1 = A_h^{-1}(F_h - B_h^* p_0), \qquad \text{and} \qquad h_j = -A_h^{-1}(B_h^* d_j), \tag{3.7}$$

respectively. As mentioned in [6], Algorithm 3.2 recovers the steps of the standard conjugate gradient algorithm for solving problem (3.5). Due to assumption (3.1), the Schur complement S_h is a symmetric positive definite operator. Consequently, the UCG-iterations p_j converge to the solution p_h of (3.5), and the rate of convergence for the iteration error $\|p_j - p_h\|_{S_h}$ or $\|p_j - p_h\|$ depends on the condition number of S_h, which is $\kappa(S_h) = \frac{M_h^2}{m_h^2}$. The following *sharp error estimation*, proved in [4], entitles the computed quantity $\|q_j\|$ as an efficient iteration error estimator.

Theorem 3.3. *If* (w_h, p_h) *is the discrete solution of* (3.4) *and* (u_j, p_{j-1}) *is the* j^{th} *iteration for Algorithm 3.2, then* $(u_j, p_{j-1}) \to (w_h, p_h)$ *and*

$$\begin{aligned}
\frac{1}{M^2}\,\|q_j\| \le \|p_{j-1} - p_h\| \le \frac{1}{m_h^2}\,\|q_j\|, \\
\frac{m_h}{M^2}\,\|q_j\| \le |u_j - w_h| \le \frac{M}{m_h^2}\,\|q_j\|.
\end{aligned} \tag{3.8}$$

We note that when $w_h = 0$ (e.g., when $V_{h,0} \subset V_0$), we also obtain, from (3.8), that $|u_j|$ an efficient error estimator for $\|p_{j-1} - p_h\|$. In addition, since u_j satisfies

$$a(u_j, v_h) = \langle F, v_h\rangle - b(v_h, p_{j-1}) = b(v_h, p - p_{j-1}) \quad \text{for all } v_h \in V_h,$$

we have that $|u_j|$ is an estimator for the discrete error $\|p - p_h\|$.

In order to build an efficient solver for (1.1), we would like to modify Algorithm 3.2 by replacing the action of A_h^{-1} with the action of a suitable preconditioner. The analysis for the resulting algorithm can be done using standard SP theory and is presented in the next section.

4. Preconditioning the SPLS discretization

In this section, we summarize a general preconditioning framework to approximate the solution of (1.1) that is presented in [6]. We plan to combine this framework with the new concepts of *optimal* and *almost optimal* test norm. The approach is based on the SPLS formulation (3.4) and on elliptic preconditioning of the operator associated with the inner product on V_h. We will assume that the inner product on V is given by a generic continuous bilinear form $a(\cdot,\cdot)$ that leads to an equivalent norm on V_h (equipped with the original norm $a_0(\cdot,\cdot)^{1/2}$). More precisely, we replace the original form $a(\cdot,\cdot)$ in (3.4) with a uniformly equivalent form $\tilde{a}(\cdot,\cdot)$ on V_h that leads to an implementably fast operator $\tilde{A}_h^{-1}$. We assume that $V_h \subset V$ and $\mathcal{M}_h \subset Q$ are finite dimensional approximation spaces satisfying (3.1) and (3.2).

4.1. The preconditioned saddle point problem.

First, we introduce a general preconditioner operator $P_h : V_h^* \to V_h$ that is equivalent to A_h^{-1} in the sense that

$$\langle g, P_h f\rangle = \langle f, P_h g\rangle \quad \text{for all } f, g \in V_h^*, \tag{4.1}$$

and

$$m_1^2 |v_h|^2 \le a(P_h A_h v_h, v_h) \le m_2^2 |v_h|^2, \tag{4.2}$$

where the positive constants m_1^2, m_2^2 are the smallest and largest eigenvalues of $P_h A_h$, respectively. Assumption (4.1) is equivalent with the fact that $P_h A_h$ is a symmetric operator with respect to the $a(\cdot,\cdot)$ inner product, and condition (4.2) is equivalent with the fact that the condition number of $P_h A_h$ satisfies

$$\kappa(P_h A_h) = \frac{m_2^2}{m_1^2}. \tag{4.3}$$

With the preconditioner $P_h : V_h^* \to V_h$, we define the form $\tilde{a} : V_h \times V_h \to \mathbb{R}$ by

$$\tilde{a}(w_h, v_h) := a((P_h A_h)^{-1} w_h, v_h) \quad \text{for all } w_h, v_h \in V_h. \tag{4.4}$$

It is easy to check under assumptions (4.1) and (4.2) that $\tilde{a}(\cdot,\cdot)$ is a symmetric bilinear form that induces an equivalent norm on V_h (originally equipped with the norm $a(\cdot,\cdot)^{1/2}$). The equivalence constants are independent of h provided that the constants m_1 and m_2 are independent of h. We let $|v_h|_P := \tilde{a}(v_h, v_h)^{1/2}$ be the norm induced by the inner product $\tilde{a}(\cdot,\cdot)$ and define the operator $\tilde{A}_h : V_h \to V_h^*$ by

$$\langle \tilde{A}_h u_h, v_h\rangle := \tilde{a}(u_h, v_h) \quad \text{for all } u_h, v_h \in V_h.$$

Note that $\tilde{A}_h = A_h (P_h A_h)^{-1} = P_h^{-1}$. We will call $\tilde{a}(\cdot,\cdot)$ a *preconditioned version* of the form $a(\cdot,\cdot)$.

The *preconditioned discrete saddle point problem* is: Find $(u_h, p_h) \in V_h \times \mathcal{M}_h$ such that

$$\begin{array}{lllll} \tilde{a}(u_h, v_h) & + & b(v_h, p_h) & = \langle f, v_h \rangle & \text{for all } v_h \in V_h, \\ b(u_h, q_h) & & & = 0 & \text{for all } q_h \in \mathcal{M}_h. \end{array} \tag{4.5}$$

Using that $V_h \subset V$ and $\mathcal{M}_h \subset Q$ satisfy (3.1) and (3.2), with M_h and m_h defined using the form $a(\cdot,\cdot)^{1/2}$, we obtain

$$\tilde{m}_h := \inf_{p_h \in \mathcal{M}_h} \sup_{v_h \in V_h} \frac{b(v_h, p_h)}{|v_h|_P \, \|p_h\|} \geq m_1 \, m_h > 0, \tag{4.6}$$

and

$$\tilde{M}_h := \sup_{p_h \in \mathcal{M}_h} \sup_{v_h \in V_h} \frac{b(v_h, p_h)}{|v_h|_P \, \|p_h\|} \leq m_2 \, M_h \leq m_2 \, M. \tag{4.7}$$

Hence, the *preconditioned saddle point least squares* formulation (4.5) has a unique solution.

The Schur complement associated with problem (4.5) is

$$\tilde{S}_h = B_h \tilde{A}_h^{-1} B_h^* = B_h P_h B_h^*.$$

Solving for p_h from (4.5), we obtain

$$\tilde{S}_h \, p_h = B_h (P_h B_h^*) \, p_h = B_h P_h F_h. \tag{4.8}$$

We call the component p_h of the solution (w_h, p_h) of (4.5) the *(preconditioned) saddle point least squares* approximation of the solution p of the original mixed prolem (1.1). To estimate $\|p - p_h\|$ in this case, we have a similar result and estimate as presented in Theorem 3.1, namely

$$\frac{1}{M} \frac{1}{m_2^2} |w_h| \leq \|p - p_h\| \leq \frac{M}{m_h} \frac{m_2}{m_1} \inf_{q_h \in \mathcal{M}_h} \|p - q_h\|. \tag{4.9}$$

The details can be found in [6].

4.2. **An iterative solver for the preconditioned variational formulation.** We use a modified version of Algorithm 3.2 to solve (4.5) by replacing the form $a(\cdot,\cdot)$ by $\tilde{a}(\cdot,\cdot)$ in **Step 1** and (**UCG1**). With this modification, we obtain the following Uzawa Preconditioned Conjugate Gradient (UPCG) algorithm for mixed methods.

Algorithm 4.1. *(UPCG) Algorithm for Mixed Methods*

Step 1: Choose any $p_0 \in \mathcal{M}_h$. **Compute** $u_1 \in V_h$, $q_1, d_1 \in \mathcal{M}_h$ *by*

$$\begin{aligned} u_1 &= P_h(F_h - B_h^* p_0) \\ q_1 &= B_h u_1, \quad d_1 := q_1. \end{aligned}$$

Step 2: For $j = 1, 2, \ldots,$ **compute** $h_j, \alpha_j, p_j, u_{j+1}, q_{j+1}, \beta_j, d_{j+1}$ *by*

$$
\begin{aligned}
&(\mathbf{PCG1}) && h_j = -\, P_h(B_h^* d_j)\\
&(\mathbf{PCG}\alpha) && \alpha_j = -\,\frac{(q_j, q_j)_Q}{b(h_j, q_j)}\\
&(\mathbf{PCG2}) && p_j = p_{j-1} + \alpha_j\ d_j\\
&(\mathbf{PCG3}) && u_{j+1} = u_j + \alpha_j\ h_j\\
&(\mathbf{PCG4}) && q_{j+1} = B_h u_{j+1},\\
&(\mathbf{PCG}\beta) && \beta_j = \frac{(q_{j+1}, q_{j+1})_Q}{(q_j, q_j)_Q}\\
&(\mathbf{PCG6}) && d_{j+1} = q_{j+1} + \beta_j d_j.
\end{aligned}
$$

We note that at each step of UPCG only the actions of P_h, B_h, and B_h^* are needed. Similar to the convergence of UCG, we have that the Schur complement $\tilde{S}_h$ of (4.5) is a symmetric positive definite operator. Consequently, the UPCG iterations p_j converge to the solution p_h of (4.8). The rate of convergence for $\|p_j - p_h\|_{\tilde{S}_h}$ or $\|p_j - p_h\|$ depends on the condition number of $\tilde{S}_h$, which is $\kappa(\tilde{S}_h) = \frac{\tilde{M}_h^2}{\tilde{m}_h^2}$. Using estimates (4.6) and (4.7), we obtain the following result.

Proposition 4.2. *The condition number of the Schur complement* $\tilde{S}_h = B_h P_h B_h^*$ *satisfies*

$$\kappa(\tilde{S}_h) \le \frac{M_h^2}{m_h^2}\frac{m_2^2}{m_1^2} = \kappa(S_h) \cdot \kappa(P_h A_h). \tag{4.10}$$

The following result, proved in [6], is the analogous form of Theorem 3.3, and entitles the computed quantity $\|q_j\|$ as an efficient iteration error estimator.

Theorem 4.3. *If* (w_h, p_h) *is the discrete solution of* (4.5) *and* (u_j, p_{j-1}) *is the* j^{th} *iteration for Algorithm 4.1, then* $(u_j, p_{j-1}) \to (w_h, p_h)$ *and*

$$
\begin{aligned}
\frac{1}{M^2}\frac{1}{m_2^2}\,\|q_j\| \le \|p_{j-1} - p_h\| &\le \frac{1}{m_h^2}\frac{1}{m_1^2}\,\|q_j\|,\\
\frac{m_h}{M^2}\frac{m_1^2}{m_2^2}\,\|q_j\| \le |u_j - w_h| &\le \frac{M}{m_h^2}\frac{m_2^2}{m_1^2}\,\|q_j\|.
\end{aligned}
\tag{4.11}
$$

5. An operator dependent discrete trial space

Up to this point, we have presented a general theory for mixed variational formulations using abstract spaces at the continuous and discrete levels. In this section, we will still consider abstract spaces, but provide a possible choice of a trial space $\mathcal{M}_h$ that depends on the choice of a trial space V_h and the operator B that defines the problem. We assume that the inner product and the norm on V are given by a bilinear form $a(\cdot,\cdot) = a_0(\cdot,\cdot)$. We plan to show that the family of pairs $(V_h, \mathcal{M}_h)$ with our operator dependent trial space is stable if the right (optimal or almost optimal) norm is chosen for the discrete test space.

Let V_h be a *finite element subspace* of V. As presented in [7, 8], using the current notation, we provide a general trial space $\mathcal{M}_h$ that can be considered for the SPLS discretization (2.4). We define $\mathcal{M}_h$ by

$$\mathcal{M}_h := BV_h \subset Q.$$

From the definition of $B_h : V_h \to \mathcal{M}_h$, we have that

$$B_h v_h = B v_h, \qquad \text{for all } v_h \in V_h,$$

i.e., B_h is the restriction of B to V_h. Consequently, from the choice of $\mathcal{M}_h$, we have that B_h is onto $\mathcal{M}_h$. It is also easy to verify that $V_{h,0} \subset V_0$, where $V_{h,0} := Ker(B_h)$.

5.1. **The discrete** $\inf - \sup$ **condition.** As presented in [8], a discrete $\inf - \sup$ condition holds. For completeness we include the (short) proof. Using a generic representation for $p_h = Bw_h \in \mathcal{M}_h$, with $w_h \in V_{h,0}^{\perp}$, and the fact that $V_{h,0}^{\perp}$ is a finite dimensional space, we have

$$\begin{aligned} m_h = \inf_{p_h \in \mathcal{M}_h} \sup_{v_h \in Vh} \frac{b(v_h, p_h)}{\|p_h\| \, |v_h|} &= \inf_{w_h \in V_{h,0}^{\perp}} \sup_{v_h \in Vh} \frac{(Bv_h, Bw_h)_Q}{\|p_h\| \, |v_h|} \\ \geq \inf_{w_h \in V_{h,0}^{\perp}} \frac{\|Bw_h\|^2}{\|Bw_h\| \, |w_h|} &= \inf_{w_h \in V_{h,0}^{\perp}} \frac{\|Bw_h\|}{|w_h|} > 0. \end{aligned}$$

5.2. **Approximability.** Using that $V_{h,0} \subset V_0$ and Proposition 2.1 on the discrete pair $(V_h, \mathcal{M}_h)$, the variational formulation (3.3) is well posed and has a unique solution $p_h \in \mathcal{M}_h$. Furthermore, by using the brief remarks after Proposition 2.1 for the discrete pair $(V_h, \mathcal{M}_h)$, we obtain that $(w_h = 0, p_h)$ is the solution of (2.4). Let p be the solution of (1.1), and let p_h be the solution of (3.3), which is the same with the second component of the SPLS solution of (2.4). Then, using (1.1) and (3.3), we obtain

$$0 = b(v_h, p - p_h) = (Bv_h, p - p_h)_Q, \qquad \text{for all } \ v_h \in V_h.$$

Thus, we have that p_h is the orthogonal projection of p onto $\mathcal{M}_h$, and consequently,

$$\|p - p_h\| = \inf_{q_h \in \mathcal{M}_h} \|p - q_h\|. \tag{5.1}$$

We can start with a finite element test space V_h with good approximation properties for functions in V. Since the space $\mathcal{M}_h = BV_h$ might not be a standard finite element space, we might not know how well functions in Q can be approximated by elements of $\mathcal{M}_h$. However, using that $B : V \to Q$ is a surjective operator, we can represent $p = Bw$ for some $w \in V$ and write a generic $q_h \in \mathcal{M}_h$ as $q_h = Bv_h$ for some $v_h \in V_h$. From (5.1), we obtain

$$\|p - p_h\| = \inf_{v_h \in V_h} \|Bw - Bv_h\| \leq M \inf_{v_h \in V_h} \|w - v_h\|, \tag{5.2}$$

which says that the best approximation of the solution p with functions in the trial space $\mathcal{M}_h$ reduces, up to the factor constant M, to the best approximation of (a smooth) representation w (such that $p = Bw$) with more familiar test functions in V_h.

5.3. **Stability by using optimal or almost optimal test norms.** Since the discretization (3.3) might be difficult to solve, we can consider the SPLS discretization of (1.1) by solving (2.4) using the UCG algorithm. As presented in Section 3, the approximation of the discrete SPLS solution p_h depends on the condition number $\kappa(S_h) = \frac{M_h^2}{m_h^2}$. Even though we proved that in general $m_h > 0$, $\kappa(S_h)$ can be large as $h \to 0$. To overcome this, we propose to replace the original norm $a_0(\cdot,\cdot)^{1/2}$ on V_h by the *optimal test norm* in the case $Ker(B) = \{0\}$ and by the *almost optimal, or α norm,* in the general case $Ker(B) \neq \{0\}$. Using that $B_h : V_h \to \mathcal{M}_h$ is a surjective operator and the restriction of B to V_h, we can simply apply Lemma 2.2 and Lemma 2.3 on $V_h \times \mathcal{M}_h$ to estimate the spectral properties of the new discrete Schur complement.

If $Ker(B) = \{0\}$ and the inner product $a_0(\cdot,\cdot)$ in (2.4) is replaced by the *optimal test norm* induced by $a_{opt}(u,v) = (Bu, Bv)_Q$, then by applying Lemma 2.2 we conclude that $M_{h,opt} = m_{h,opt} = 1$, and the Schur complement for the SPLS disccstretization with optimal test norm is the identity operator. In this case, we have optimal discrete stability and optimal approximability. On the other hand, if $Ker(B) \neq \{0\}$ and the inner product $a_0(\cdot,\cdot)$ in (2.4) is replaced by the *almost optimal test norm* induced by $a_\alpha(u,v) = \alpha^2 a_0(u,v) + (Bu, Bv)_Q$, by applying Lemma 2.3 we conclude that

$$m_{h,\alpha}^2 = \frac{m_h^2}{m_h^2 + \alpha^2} \quad \text{and} \quad M_{h,\alpha}^2 = \frac{M_h^2}{M_h^2 + \alpha^2}. \tag{5.3}$$

Consequently, the condition number of the Schur complement $S_{h,\alpha}$ of the new saddle point system becomes

$$\kappa(S_{h,\alpha}) = \frac{M_h^2}{m_h^2} \frac{m_h^2 + \alpha^2}{M_h^2 + \alpha^2},$$

and for $\alpha \in (0, m_h]$, we have $\kappa(S_{h,\alpha}) \in (1, 2)$.

Remark 5.1. *The estimates in this subsection for the discrete continuity constants and the discrete* $\inf - \sup$ *constants when using optimal or almost optimal test norms hold true for any pair of spaces $(V_h, \mathcal{M}_h)$ that satisfy* (3.1) *and* (3.2)*, provided that we use the h dependent norm induced by*

$$a_{opt,h}(w_h, v_h) = (B_h u, B_h v)_Q, \qquad \textit{for all } w_h, v_h \in V_h,$$

in the $Ker(B_h) = \{0\}$ case and

$$a_{\alpha,h}(w_h, v_h) = \alpha^2 a_0(w_h, v_h) + (B_h u, B_h v)_Q, \qquad \textit{for all } w_h, v_h \in V_h,$$

for the $Ker(B_h) \neq \{0\}$ case. For independence of h of these norms, we would need $\|B_h v_h\|_Q = \|B v_h\|_Q$ for all $v_h \in V_h$ or $\|B_h v_h\|_Q \approx \|B v_h\|_Q$ with equivalence on V_h independent of h.

Applying the UCG algorithm with the form $a(\cdot,\cdot)$ replaced by $a_{opt}(\cdot,\cdot)$ or $a_\alpha(\cdot,\cdot)$ leads to an efficient iterative process (if α is properly chosen in the second case) with the number of iterations independent of h. The difficulty here is shifted to the inversion of the operators associated with $a_{opt}(\cdot,\cdot)$ or $a_\alpha(\cdot,\cdot)$ (see the examples in Section 6). However, in light of Section 4 we only need to use preconditioners for these symmetric and positive definite bilinear forms. Since the theory of preconditioning symmetric positive definite operators (with or without parameters) is well developed in the finite element community, we consider that *paying with efficient preconditioning in order to get stability* is worth trying, especially when finding stable pairs in the standard norms is more difficult.

6. Examples of optimal and almost optimal test norms

In this section, we consider one example of SPLS formulation with optimal test norm and one example with almost optimal test norm. We do not try to find a new or best discretization approach for the two examples. Rather, the goal is to emphasize how stability for mixed formulation can be gained and provide a way to choose the appropriate preconditioner towards finding an efficient solver for the mixed formulation and discretization.

6.1. **Optimal SPLS test norm for the reaction diffusion problem.**

We consider the following reaction diffusion problem

$$\left\{ \begin{aligned} -\varepsilon\,\Delta u + cu = &\ f \quad \text{in} \quad \Omega,\\ u = &\ 0 \quad \text{on}\ \partial\Omega, \end{aligned} \right. \tag{6.1}$$

for $\varepsilon > 0$ and $c(x) \geq c_0 > 0$ on Ω, a bounded domain in $\mathbb{R}^d$. In what follows, $(\cdot,\cdot)$ and $\|\cdot\|$ will denote the standard L^2 inner product and norm, respectively.

A standard variational formulation for (6.1) is: Find $u \in H_0^1(\Omega)$ such that

$$\varepsilon(\nabla u, \nabla v) + (cu, v) = (f, v) \quad \text{for all } v \in H_0^1(\Omega). \tag{6.2}$$

To obtain a mixed formulation that is suitable to the SPLS framework, we let $V := H_0^1(\Omega)$, and Q be the graph of the operator $\varepsilon\nabla : H_0^1(\Omega) \to L^2(\Omega)^d$, i.e.,

$$Q := G(\varepsilon\nabla) = \left\{ \left(\begin{smallmatrix} v \\ \varepsilon\nabla v \end{smallmatrix}\right) \mid v \in H_0^1(\Omega) \right\}.$$

We define the bilinear form $b : V \times Q \to \mathbb{R}$ as

$$b(v, \left(\begin{smallmatrix} w \\ \varepsilon\nabla w \end{smallmatrix}\right)) := (cw, v) + \varepsilon(\nabla w, \nabla v) \quad \text{for all } v \in V, \left(\begin{smallmatrix} w \\ \varepsilon\nabla w \end{smallmatrix}\right) \in Q,$$

and the linear functional $F \in V^*$ as

$$\langle F, v\rangle := (f, v) \quad \text{for all } v \in H_0^1(\Omega).$$

With this setting, the standard variational formulation (6.2) can be reformulated in the mixed form: Find $\mathbf{p} = \left(\begin{smallmatrix} u \\ \varepsilon\nabla u \end{smallmatrix}\right) \in Q$ such that

$$b(v, \mathbf{p}) = (cu, v) + \varepsilon(\nabla u, \nabla v) = (f, v) \quad \text{for all } v \in V. \tag{6.3}$$

On V, we consider first the standard inner product defined by

$$a_0(u, v) = (\nabla u, \nabla v) \quad \text{for all } u, v \in V,$$

and on Q, we consider the weighted inner product

$$(\left(\begin{smallmatrix} u \\ \varepsilon\nabla u \end{smallmatrix}\right), \left(\begin{smallmatrix} v \\ \varepsilon\nabla v \end{smallmatrix}\right))_Q = (cu, v) + \varepsilon(\nabla u, \nabla v) \quad \text{for all } \left(\begin{smallmatrix} u \\ \varepsilon\nabla u \end{smallmatrix}\right), \left(\begin{smallmatrix} v \\ \varepsilon\nabla v \end{smallmatrix}\right) \in Q.$$

The corresponding norm is

$$\| \left(\begin{smallmatrix} v \\ \varepsilon\nabla v \end{smallmatrix}\right) \|_Q = \left(\|c^{1/2}v\|^2 + \|\varepsilon^{1/2}\nabla v\|^2\right)^{1/2}.$$

For the standard norm on $V = H_0^1(\Omega)$, the $\inf - \sup$ condition on $V \times Q$ holds with a constant m that depends on ε. The operator $B : V \to Q$ is given by

$$Bv = \left(\begin{smallmatrix} v \\ \varepsilon\nabla v \end{smallmatrix}\right) \quad \text{for all } v \in V.$$

Thus, the *optimal test norm* on V is induced by the inner product

$$a_{opt}(u, v) = (Bu, Bv)_Q = \varepsilon(\nabla u, \nabla v) + (cu, v) \quad \text{for all } u, v \in V,$$

which gives rise to the norm

$$|v|_{opt} = \left(\|c^{1/2}v\|^2 + \|\varepsilon^{1/2}\nabla v\|^2\right)^{1/2}.$$

The compatibility condition (2.3) is automatically satisfied as

$$V_0 = \mathrm{Ker}(B) = \{v \in H_0^1(\Omega) \mid Bv = 0\} = \{0\}.$$

In addition, according to Section 2.2, we obtain $M = m = 1$. This leads to optimal continuity and $\inf - \sup$ constants. However, inverting the operator associated with $| \cdot |_{opt}$ coincides with solving the original problem. Fortunately, at the discrete level we can replace $a_{opt}(\cdot, \cdot)$ by a preconditioned form.

For discretization we can choose $V_h \subset V = H_0^1(\Omega)$ to be the space of continuous piecewise polynomials of degree k with respect to a mesh $\mathcal{T}_h$ on Ω and let $\mathcal{M}_h$ the operator dependent choice

$$\mathcal{M}_h := BV_h = \begin{pmatrix} I \\ \varepsilon\nabla \end{pmatrix} V_h,$$

where $I : V_h \to V_h$ is the identity operator and the inner product is chosen to coincide with the inner product on Q. According to Section 5.3, for the above choice of trial space, we also have $M_h = m_h = 1$.

Using Remark 5.1 for a more general choice of $\mathcal{M}_h$ (compatible with V_h), we still have $M_h = m_h = 1$ provided that we are using the *optimal test norm* induced by $a_{opt,h}(\cdot, \cdot)$ on V_h. In order to come up with an efficient UPCG solver, we will need to find robust (with respect to h and ε) preconditioners for the discrete *optimal norm* on V_h. For quasi-uniform meshes, such preconditioners are available, see e.g., [16, 31]. For non-uniform mesh discretization, such as the use of Shishkin meshes [34], such theory seems to not be developed.

6.2. **Numerical results.** We solved (6.1) on the unit square with variable coefficient $c = 2(1 + x^2 + y^2)$ and f computed such that the exact solution is given by

$$\begin{aligned} u(x, y) = x(1 - x) &\left(1 - e^{-y/\sqrt{\varepsilon}}\right)\left(1 - e^{(y-1)/\sqrt{\varepsilon}}\right) \\ &+ y(1 - y)\left(1 - e^{-x/\sqrt{\varepsilon}}\right)\left(1 - e^{(x-1)/\sqrt{\varepsilon}}\right), \end{aligned}$$

as considered in [29]. For this problem, the solution has boundary layers on all sides of the unit square. To this end, we use a Shishkin mesh for discretization purposes. Following [33], we review the construction of the Shishkin mesh.

Assume N is an integer multiple of 8. This parameter will refer to the number of mesh intervals in the x and y directions. The mesh itself is the tensor product of two one-dimensional Shishkin meshes $\mathcal{T}_x \times \mathcal{T}_y$. The process for obtaining $\mathcal{T}_x$ (and $\mathcal{T}_y$) is as follows. The interval $[0, 1]$ is first decomposed into three subintervals $[0, \lambda]$, $[\lambda, 1 - \lambda]$, and $[1 - \lambda, 1]$, where

$$\lambda = \min\left\{\frac{1}{4}, 2\sqrt{\frac{\varepsilon}{c^*}}\ln N\right\} \quad \text{with} \quad 0 < c^* < c. \tag{6.4}$$

The intervals $[0, \lambda]$ and $[1 - \lambda, 1]$ are then partitioned into $N/4$ subintervals of length $\frac{4\lambda}{N}$, while the interval $[\lambda, 1-\lambda]$ is partitioned into $N/2$ subintervals of length $\frac{2(1-2\lambda)}{N}$. The triangular mesh is obtained by drawing diagonals from the top left to bottom right of each quadrilateral. See [27] or Section 6 of [5] for figures showing examples of the Shishkin mesh generated using $\varepsilon = 10^{-4}$ and $c^* = \sqrt{1/2}$ for $N = 16$ and $N = 32$. With this setup, we choose the test space $V_h = V_N \subset H_0^1(\Omega)$ to be the space of continuous piecewise linear polynomials with respect to the Shishkin mesh $\mathcal{T}_h$ and take $\mathcal{M}_h := BV_h$.

We performed computations using the UCG algorithm with different choices for the inner product $a(\cdot, \cdot)$ on V:

(a) $a(u, v) = \varepsilon(\nabla u, \nabla v) + (u, v)$, see Table 1
(b) $a(u, v) = (\nabla u, \nabla v)$, see Table 2
(c) $a(u, v) = \sqrt{\varepsilon}(\nabla u, \nabla v) + (u, v)$, see Table 3
(d) $a(u, v) = \sqrt{\varepsilon}(\nabla u, \nabla v) + (cu, v)$, see Table 3

The stopping criterion used for the UCG algorithm was

$$\|q_j\|_Q \leq 10^{-12},$$

in all cases. Also, we measured the SPLS solution in a balanced norm instead of the norm on Q. This is due to the fact that for small ε the L^2 part of the norm on Q dominates, leading to an unbalanced norm not adequate to accurately measure the error, see [30, 33]. More precisely, we approximate

$$\mathrm{error_h} := \mathrm{error_N} := \left(\|u-u_h\|^2 + \varepsilon^{1/2}\|\nabla u - \nabla u_h\|^2\right)^{1/2},$$

where u_h is in fact u_N. According to [30, 33], standard Galerkin methods for (6.2) lead to a covergence rate of $\mathcal{O}(N^{-1}\ln N)$ using piecewise linear approximation [30, 33]. We expect the same order of convergence for SPLS discretization.

Numerical tests using the UCG algorithm and the optimal norm induced by $a_{opt}(u,v) = \varepsilon(\nabla u, \nabla v)+(cu,v)$ were performed in [27] and [5]. The results are very close with the results of Table 1 with the main observation that the number of iterations is always one when using the optimal norm. This is always the case when using the UCG algorithm and the inner product $a(\cdot,\cdot) = a_{opt}(\cdot,\cdot)$. However, in this case we have to invert (in **Step 1** of UCG) an operator that corresponds to the original problem.

The purpose of our numerical tests is to show that by using closely related norms on V (that might be easier to precondition) we preserve the order of approximation, and the price to pay is an increase in the number of iterations as $\varepsilon \to 0$. Choosing inner product (a), which is close to $a_{opt}(\cdot,\cdot)$, preserves the order of convergence and slightly increases the number of iterations as ε decreases as shown in Table 1. By choosing inner product (b), which is independent of ε, we note a huge increase of the number of iterations. This is because the $\inf-\sup$ condition constant, and hence the condition number of the Schur complement, increases as $\varepsilon \to 0$, see Table 2. The results for choices (c) and (d), shown in Table 3, demonstrate that a larger penalty on the term $(\nabla\cdot,\nabla\cdot)$ could lead to an intermediate number of iterations if compared to the previous two cases. Table 3 also demonstrates that the number of iterations decreases the closer to $a_{opt}(\cdot,\cdot)$ we are (by using the exact component $(c\,\cdot,\cdot)$).

N	$\varepsilon = 1$			$\varepsilon = 10^{-2}$			$\varepsilon = 10^{-4}$		
	error	rate	it	error	rate	it	error	rate	it
16	0.0189		6	0.0682		16	0.132		24
32	0.0095	1.472	6	0.0341	1.471	16	0.088	0.854	25
64	0.0047	1.356	6	0.0171	1.356	16	0.054	0.946	26
128	0.0023	1.286	6	0.0085	1.285	16	0.032	0.984	26
256	0.0012	1.239	6	0.0042	1.239	16	0.018	0.996	27

N	$\varepsilon = 10^{-8}$			$\varepsilon = 10^{-12}$			$\varepsilon = 10^{-16}$		
	error	rate	it	error	rate	it	error	rate	it
16	0.133		24	0.133		23	0.133		22
32	0.089	0.859	26	0.089	0.859	25	0.089	0.860	24
64	0.055	0.951	28	0.055	0.951	27	0.055	0.951	25
128	0.032	0.988	28	0.032	0.988	27	0.032	0.988	26
256	0.018	0.999	29	0.018	0.999	27	0.018	0.999	26

Table 1: Results for the case $a(u, v) = \epsilon(\nabla u, \nabla v) + (u, v)$.

N	$\varepsilon = 1$			$\varepsilon = 10^{-2}$			$\varepsilon = 10^{-4}$		
	error	rate	it	error	rate	it	error	rate	it
16	0.0189		6	0.0682		18	0.1317		93
32	0.0094	1.472	7	0.0341	1.470	20	0.0882	0.854	110
64	0.0047	1.356	7	0.0170	1.356	19	0.0544	0.945	118
128	0.0023	1.285	7	0.0085	1.285	20	0.0320	0.983	124
256	0.0011	1.238	7	0.0042	1.238	20	0.0183	0.995	126

N	$\varepsilon = 10^{-6}$			$\varepsilon = 10^{-8}$		
	error	rate	it	error	rate	it
16	0.1332		302	0.1335		426
32	0.0889	0.858	607	0.0891	0.859	1301
64	0.0547	0.950	825	0.0548	0.950	2796
128	0.0321	0.987	869	0.0322	0.987	4018
256	0.0183	0.998	868	0.0184	0.998	4819

Table 2: Results for the case $a(u, v) = (\nabla u, \nabla v)$.

6.3. **SPLS discretization of a** $\operatorname{div}-\operatorname{curl}$ **system.** We describe how one can apply the general SPLS theory with almost optimal test norm for a model $\operatorname{div}-\operatorname{curl}$ problem on a polyhedral domain $\Omega \subset \mathbb{R}^3$. For given data, we are looking to find the vector function $\mathbf{h} \in \mathbf{L}^2(\Omega)$ such that

$$\begin{aligned} \nabla\times\ (\mu^{-1}\mathbf{h}) &= \mathbf{j} \quad \text{in } \Omega \\ \nabla\cdot\ \mathbf{h} &= g \quad \text{in } \Omega \\ \mathbf{h}\cdot\mathbf{n} &= \sigma \quad \text{on } \Gamma := \partial\Omega, \end{aligned} \tag{6.5}$$

N	$\varepsilon = 1$				$\varepsilon = 10^{-2}$			
	error	rate	it (c)	it (d)	error	rate	it (c)	it (d)
16	0.0189		6	1	0.0682		19	18
32	0.0094	1.472	6	1	0.0341	1.470	20	19
64	0.0047	1.356	6	1	0.0170	1.356	21	19
128	0.0023	1.285	6	1	0.0085	1.285	21	19
256	0.0011	1.238	6	1	0.0042	1.238	21	19

N	$\varepsilon = 10^{-4}$				$\varepsilon = 10^{-6}$			
	error	rate	it (c)	it (d)	error	rate	it (c)	it (d)
16	0.1317		90	68	0.1332		197	125
32	0.0882	0.854	104	75	0.0889	0.858	358	202
64	0.0544	0.945	112	82	0.0547	0.950	476	266
128	0.0320	0.983	115	82	0.0321	0.987	502	279
256	0.0183	0.995	116	83	0.0183	0.998	507	282

N	$\varepsilon = 10^{-8}$			
	error	rate	it (c)	it (d)
16	0.1335		246	130
32	0.0891	0.859	516	261
64	0.0548	0.950	892	418
128	0.0322	0.987	1265	585
256	0.0184	0.998	1568	734

Table 3: Results for cases $a(u,v) = \sqrt{\varepsilon}(\nabla u, \nabla v) + (u,v)$ and $a(u,v) = \sqrt{\varepsilon}(\nabla u, \nabla v) + (cu,v)$.

where $\mu = \mu(x)$ is a given scalar L^2 function satisfying $0 < \mu_0 \le \mu(x) \le \mu_1$, for a.e. $x \in \Omega$. We consider the variational formulation for (6.5) that is presented in [17]. We multiply the first equation in (6.5) by $\mathbf{w} \in \mathbf{H}_0^1(\Omega)$ and the second equation by $\varphi \in H^1(\Omega)/\mathbb{R}$. We assume enough regularity for the data (hence, the solution) in order to be able to integrate by parts. After we integrate by parts, we obtain

$$\begin{aligned} (\mu^{-\mathbf{1}}\mathbf{h}, \nabla\times\ \mathbf{w}) &= (\mathbf{j}, \mathbf{w}) && \text{for all} \quad \mathbf{w} \in \mathbf{H}_0^1(\Omega) \\ (\mathbf{h}, \nabla\varphi) &= (-g, \varphi) + (\sigma, \varphi)_\Gamma && \text{for all} \quad \varphi \in H^1(\Omega)/\mathbb{R}, \end{aligned} \tag{6.6}$$

where $(\cdot,\cdot)$ denotes the standard L^2 type inner product. If we define the spaces $V := \mathbf{H}_0^1(\Omega) \times H^1(\Omega)/\mathbb{R}$ and $Q := \mathbf{L}^2(\Omega)$, and the form "$b(\mathbf{v}, p)$" on (V, Q) by

$$b((\mathbf{w}, \varphi), \mathbf{h}) := (\mu^{-1}\nabla\times\ \mathbf{w} + \nabla\varphi, \mathbf{h}), \text{ for all } (\mathbf{w}, \varphi) \in V,\ \mathbf{h} \in Q,$$

the variational formulation for (6.5) becomes: Find $\mathbf{h} \in Q$ such that

$$b((\mathbf{w}, \varphi), \mathbf{h}) = \langle F, (\mathbf{w}, \varphi)\rangle := (\mathbf{j}, \mathbf{w}) + (-g, \varphi) + (\sigma, \varphi)_\Gamma, (\mathbf{w}, \varphi) \in V. \tag{6.7}$$

The inner product on Q is the weighted inner product $(\cdot,\cdot)_{\mu^{-1}}$ in order to take advantage of the orthogonality between the (weighted) gradient fields and the curl fields. On $V := \mathbf{H}_0^1(\Omega) \times H^1(\Omega)/\mathbb{R}$, the inner product is chosen such that the corresponding induced norm is

$$\|(\mathbf{w},\varphi)\|_V^2 := a_{\mu^{-1}}(\mathbf{w},\mathbf{w}) + a_{\mu}(\varphi,\varphi) := \int_\Omega \mu^{-1}|\nabla \mathbf{w}|^2 + \int_\Omega \mu |\nabla\varphi|^2.$$

With these choices of inner products and norms, the continuity constant M for $b(\cdot,\cdot)$ is $M = 1$. A continuous $\inf-\sup$ condition holds with a constant $m > \frac{c_0}{\mu_1}$, where c_0 depending only on the domain Ω, see Section 4 of [17]. The corresponding operator $B : V \to Q$ is

$$B\,(\mathbf{w},\varphi) = \operatorname{curl}\mathbf{w} + \mu\,\nabla\varphi. \tag{6.8}$$

By using the orthogonality between $\operatorname{curl}\mathbf{w}$ and $\mu\,\nabla\varphi$ in the weighted inner product $(\cdot,\cdot)_{\mu^{-1}}$, we can verify that

$$V_0 := Ker(B) = \{(\mathbf{w},0) \in V \mid \operatorname{curl}\mathbf{w} = 0\}.$$

If the data $(\mathbf{j}, g, \sigma)$ is such that the compatibility condition (2.3) is satisfied, then (6.6) is a well-posed problem and our SPLS discretization can be considered. An *almost optimal or α norm* on V is defined by

$$\begin{aligned}\|(\mathbf{w},\varphi)\|_\alpha^2 := \alpha^2\left((\mu^{-1}\nabla\mathbf{w},\nabla\mathbf{w}) + (\mu\nabla\varphi,\nabla\varphi)\right) + \\ (\mu^{-1}\nabla\times\mathbf{w},\nabla\times\mathbf{w}) + (\mu\nabla\varphi,\nabla\varphi).\end{aligned}$$

We can also try to precondition the almost optimal norm at the continuous level first. To demonstrate this, we take $\mu = 1$ for simplicity. Then

$$\|(\mathbf{w},\varphi)\|_\alpha^2 = (1+\alpha^2)\|\operatorname{curl}\mathbf{w}\|^2 + \alpha^2\|\operatorname{div}\mathbf{w}\|^2 + (1+\alpha^2)\|\nabla\varphi\|^2,$$

which, for α small, is equivalent (or can be preconditioned by)

$$\|(\mathbf{w},\varphi)\|_{\alpha,prec}^2 = \|\operatorname{curl}\mathbf{w}\|^2 + \alpha^2\|\operatorname{div}\mathbf{w}\|^2 + \|\nabla\varphi\|^2.$$

Conforming SPLS discretization using the above norm on V reduces to inverting or preconditioning the operators associated with the forms $\|\nabla\varphi\|^2$ and $\|\operatorname{curl}\mathbf{w}\|^2 + \alpha^2\|\operatorname{div}\mathbf{w}\|^2$. The first form is parameter free and can benefit from classical preconditioning theory for the Laplace operator. In the two dimensional case, the second form is uniformly equivalent with $\|\nabla\mathbf{w}\|^2 + \alpha^2\|\operatorname{div}\mathbf{w}\|^2$ and one can benefit from the preconditioning theory developed for the Augmented Lagrangian method for Stokes systems.

For *SPLS discretization*, we can also choose $V_h \subset V$ to be any good approximation finite element space, e.g., the space of continuous piecewise functions of fixed degree with the appropriate boundary conditions for each component of V, and for $\mathcal{M}_h$ we can consider the general choice $\mathcal{M}_h := BV_h$ as in [9, 11]. If an estimate c_h for the discrete constant m_h is available, then a practical choice for α for an *almost optimal norm* is $\alpha = \min\{\frac{1}{\mu_1}, c_h\}$. If a family $\{(V_h,\mathcal{M}_h)\}$ of discrete spaces is available such that the discrete $\inf-\sup$ constant m_h (using the original norm on V) satisfies $m_h > c_1 > 0$

for all h, then a practical choice for α when using the UCG algorithm with *almost optimal test norm* is $\alpha = \frac{1}{\mu_1}$.

Numerical results done in [11, 32] without preconditioning show that the number of iterations increases due to mesh size h or the discontinuity jump size of μ. Preconditioning the α *test norm* would lead to a more efficient UPCG algorithm for the div − curl system. The construction of such preconditioners for a specific test spaces V_h and norms is a challenging problem and will not be discussed in this paper.

7. Conclusion

We presented some general results and ideas regarding a saddle point (least squares) reformulation of mixed variational problems. The results can be useful for further development of both DPG and SPLS methodologies. We considered the concept of optimal test norm (when $B : V \to Q$ is injective), as presented in [25, 28], and extended it to the case when $B : V \to Q$ might not be injective by introducing the concept of *almost optimal test norm.* A general preconditioning strategy and an iterative process for solving the discrete mixed formulations are reviewed in light of the special test norms. We also presented examples of improving stability and solver efficiency by preconditioning as well as the use of special trial norms.

Acknowledgment. The first author is grateful for an Oden Institute fellowship which allowed him to visit UT Austin in the Fall of 2019.

8. Appendix

Proof of Lemma 2.2

Proof. Using the definitions of $|v|_{opt}$ and $b(v,p)$, we obtain

$$\sup_{v\in V} \frac{b(v,p)}{|v|_{opt}} = \sup_{v\in V} \frac{(Bv,p)_Q}{||Bv||_Q} = \|p\|.$$

Thus, the continuity and inf-sup constants satisfy

$$M_{opt} = \sup_{p\in Q} \sup_{v\in V} \frac{b(v,p)}{|v|_{opt}||p||} = \sup_{p\in Q} \frac{||p||}{||p||} = 1,$$

and

$$m_{opt} = \inf_{p\in Q} \sup_{v\in V} \frac{b(v,p)}{|v|_{opt}||p||} = \inf_{p\in Q} \frac{||p||}{||p||} = 1,$$

as desired. From the remarks at the end of Section 2.1, we can conclude that the corresponding Schur complement S_{opt} is the identity on V. Another way to see that is by considering the new SP system where $a_0(\cdot,\cdot)$ in (2.4) is replaced by $a_{opt}(\cdot,\cdot)$,

$$\begin{array}{lllll} (Bw,Bv)_Q & + & b(v,p) & = \langle F, v\rangle & \text{for all } v \in V, \\ b(w,q) & & & = 0 & \text{for all } q \in Q. \end{array} \tag{8.1}$$

The corresponding operator system is

$$\begin{array}{lcrcl} B^*Bw & + & B^*p & = & F \\ Bw & & & = & 0. \end{array} \tag{8.2}$$

Therefore, using that B (hence B^*) is invertible, we quickly recover that the Schur complement is the identity.

□

Proof of Lemma 2.3

Proof. Denote $A_\alpha : V \to V^*$ to be the operator associated with $a_\alpha(u,v)$, i.e., $A_\alpha = \alpha^2 A_0 + B^*B$. We will need the following two identities:

$$\alpha^2(I + \alpha^2 C^{-1})^{-1} = C - C(\alpha^2 I + C)^{-1}C, \tag{8.3}$$

for any linear bounded operator C on Q, and

$$\alpha^2(\alpha^2 A_0 + B^*B)^{-1} \;\; = \;\; A_0^{-1} - A_0^{-1}B^*(\alpha^2 I + BA_0^{-1}B^*)^{-1}BA_0^{-1}, \tag{8.4}$$

where A_0 and B are the operators defined in Section 2. Both identities can be easily justified by checking that the proposed algebraic inverse satisfies the definition of the corresponding inverse operator. The identity (8.4) is a version of the Sherman-Morrison-Woodbury formula.

If we pre and post compose (8.4) with B^* and B respectively, and combine the result with (8.3) with $C = BA_0^{-1}B^*$, we obtain

$$BA_\alpha^{-1}B^* = (I + \alpha^2(BA_0^{-1}B^*)^{-1})^{-1}.$$

It is known that, see e.g. [2, Lemma 2.1],

$$\sup_{v \in V} \frac{b(v,p)^2}{a_\alpha(v,v)} = (BA_\alpha^{-1}B^*p, p).$$

Replacing $BA_\alpha^{-1}B^*$ from the above identity, we have

$$\begin{aligned} M_\alpha^2 &= \sup_{p\in Q} \frac{(BA_\alpha^{-1}B^*p,p)}{(p,p)} = \sup_{p\in Q} \frac{((I+\alpha^2(BA_0^{-1}B^*)^{-1})^{-1}p,p)}{(p,p)} \\ &= \left(\inf_{q\in Q} \frac{((I+\alpha^2(BA_0^{-1}B^*)^{-1})q,q)}{(q,q)}\right)^{-1} = \left(1+\alpha^2 \inf_{q\in Q}\frac{(BA_0^{-1}B^*q,q)}{(q,q)}\right)^{-1} \\ &= \left(1+\alpha^2\left(\sup_{r\in Q}\frac{(BA_0^{-1}B^*r,r)}{(r,r)}\right)^{-1}\right)^{-1} = \frac{1}{1+\frac{\alpha^2}{M^2}} = \frac{M^2}{M^2+\alpha^2}. \end{aligned}$$

The proof for m_α^2 proceeds similarly. It then follows immediately that the condition number of S_α is given by

$$\kappa(S_\alpha) = \frac{M_\alpha^2}{m_\alpha^2} = \frac{M^2}{m^2}\frac{m^2+\alpha^2}{M^2+\alpha^2}.$$

□

References

[1] A. Aziz and I. Babuška. Survey lectures on mathematical foundations of the finite element method. *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. Aziz, editor*, 1972.

[2] C. Bacuta. A unified approach for Uzawa algorithms. *SIAM J. Numer. Anal.*, 44(6):2633–2649, 2006.

[3] C. Bacuta. Schur complements on Hilbert spaces and saddle point systems. *J. Comput. Appl. Math.*, 225(2):581–593, 2009.

[4] C. Bacuta. Cascadic multilevel algorithms for symmetric saddle point systems. *Comput. Math. Appl.*, 67(10):1905–1913, 2014.

[5] C. Bacuta and J. Jacavage. Saddle point least squares for the reaction diffusion problem. *Results in Applied Mathematics*, Available online 7 May 2020:1–20, 2019.

[6] C. Bacuta and J. Jacavage. Saddle point least squares preconditioning of mixed methods. *Computers & Mathematics with Applications*, 77(5):1396–1407, 2019.

[7] C. Bacuta and J. Jacavage. Least squares preconditioning for mixed methods with nonconforming trial spaces. *Applicable Analysis*, Available online Feb 27, 2019:1–20, 2020.

[8] C. Bacuta and J. Jacavage. A non-conforming saddle point least squares approach for an elliptic interface problem. *Computational Methods in Applied Mathematics*, doi:10.1515/cmam-2018-0202, 2019.

[9] C. Bacuta, J. Jacavage, K. Qirko, and F.J. Sayas. Saddle point least squares iterative solvers for the time harmonic maxwell equations. *Comput. Math. Appl.*, 70(11):2915–2928, 2017.

[10] C. Bacuta and P. Monk. Multilevel discretization of symmetric saddle point systems without the discrete LBB condition. *Appl. Numer. Math.*, 62(6):667–681, 2012.

[11] C. Bacuta and K. Qirko. A saddle point least squares approach to mixed methods. *Comput. Math. Appl.*, 70(12):2920–2932, 2015.

[12] C. Bacuta and K. Qirko. A saddle point least squares approach for primal mixed formulations of second order PDEs. *Comput. Math. Appl.*, 73(2):173–186, 2017.

[13] D. Boffi, F. Brezzi, L. Demkowicz, R. G. Durán, R. Falk, and M. Fortin. *Mixed finite elements, compatibility conditions, and applications*, volume 1939 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin; Fondazione C.I.M.E., Florence, 2008. Lectures given at the C.I.M.E. Summer School held in Cetraro, June 26–July 1, 2006, Edited by Boffi and Lucia Gastaldi.

[14] D Boffi, F Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2013.

[15] T. Bouma, J. Gopalakrishnan, and A. Harb. Convergence rates of the DPG method with reduced test space degree. *Comput. Math. Appl.*, 68(11):1550–1561, 2014.

[16] J.H. Bramble, E.J. Pasciak, and P. Vassilevski. Computational scales of Sobolev norms with application to preconditioning. *Math. Comp.*, 69(230):463–480, 2000.

[17] J.H. Bramble and J.E. Pasciak. A new approximation technique for div-curl systems. *Math. Comp.*, 73:1739–1762, 2004.

[18] Dirk Broersen and Rob Stevenson. A robust Petrov-Galerkin discretisation of convection-diffusion equations. *Comput. Math. Appl.*, 68(11):1605–1618, 2014.

[19] L. Demkowicz C. Carstensen and J. Gopalakrishnan. Breaking spaces and form for the DPG method and applications including maxwell equations. *Computers and Mathematics with Applications*, 72:494–522, 2016.

[20] J. Chan, N. Heuer, T. Bui-Thanh, and L. Demkowicz. A robust DPG method for convection-dominated diffusion problems II: adjoint boundary conditions and mesh-dependent test norms. *Comput. Math. Appl.*, 67(4):771–795, 2014.

[21] A. Cohen, W. Dahmen, and G. Welper. Adaptivity and variational stabilization for convection-diffusion equations. *ESAIM Math. Model. Numer. Anal.*, 46(5):1247–1273, 2012.

[22] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. Part I: the transport equation. *Comput. Methods Appl. Mech. Engrg.*, 199(23-24):1558–1572, 2010.

[23] L. Demkowicz and L. Vardapetyan. Modelling electromagnetic/scattering problems using hp-adaptive finite element methods. *Comput, Methods Appl. Mech. Engrg. Numerical Mathematics*, 152:103 – 124, 1998.

[24] Leszek Demkowicz and Jay Gopalakrishnan. A class of discontinuous Petrov–Galerkin methods. ii. optimal test functions. *Numerical Methods for Partial Differential Equations*, 27(1):70–105, 2011.

[25] Jay Gopalakrishnan. Five lectures on DPG methods. *arXiv 1306.0557*, 2013.

[26] K. W. Morton J. W. Barrett, and. Optimal Petrov-Galerkin methods through approximate symmetrization. *IMA J. Numer. Anal.*, 1(4):439–468, 1981.

[27] J. Jacavage. *A Least Squares Method for Mixed Variational Formulations of Partial Differential Equations.* ProQuest Dissertations Publishing, University of Delaware, 2019. Thesis (Ph.D.)–University of Delaware.

[28] N. Heuer L. Demkowicz, T. Führer and X. Tian. The Double Adaptivity Paradigm (how to circumvent the discrete inf-sup conditions of Babuška and Brezzi). Technical report, Oden Institue for Computational Engineering and Sciences, The University fo Texas at Austin, May 2019.

[29] J. Li. Convergence and superconvergence analysis of finite element methods on highly nonuniform anisotropic meshes for singularly perturbed reaction–diffusion problems. *Applied Numerical Mathematics*, 36(2):129–154, 2001.

[30] R. Lin and M. Stynes. A balanced finite element method for singularly perturbed reaction-diffusion problems. *SIAM Journal on Numerical Analysis*, 50(5):2729–2743, 2012.

[31] K-A. Mardal and R. Winther. Uniform preconditioners for the time dependent Stokes problem. *Numer. Math.*, 98(2):305–327, 2004.

[32] K. Qirko. *A Saddle Point Least Squares Method for Systems of Linear PDEs.* ProQuest LLC, Ann Arbor, MI, 2017. Thesis (Ph.D.)–University of Delaware.

[33] H.G. Roos and M. Schopf. Convergence and stability in balanced norms of finite element methods on shishkin meshes for reaction-diffusion problems: Convergence and stability in balanced norms. *ZAMM Journal of applied mathematics and mechanics: Zeitschrift für angewandte Mathematik und Mechanik*, 95(6):551–565, 2014.

[34] G.I. Shishkin. Grid approximation of singularly perturbed boundary value problems with a regular boundary layer. *Sov. J. Numer. Anal. Math. Model.*, 4(5):397–417, 1989.

University of Delaware, Mathematical Sciences, 501 Ewing Hall, Newark, DE 19716
Email address: bacuta@udel.edu

University of Delaware, Department of Mathematics, 501 Ewing Hall 19716
Email address: dphayes@udel.edu

Lafayette College, Department of Mathematics, Pardee Hall, Easton, PA 18042
Email address: jacavagj@lafayette.edu