

# ICES REPORT 17-09

---

May 2017

## Discrete least-squares finite element methods

by

Brendan Keith, Socratis Petrides, Federico Fuentes, and Leszek Demkowicz



**The Institute for Computational Engineering and Sciences**  
The University of Texas at Austin  
Austin, Texas 78712

*Reference: Brendan Keith, Socratis Petrides, Federico Fuentes, and Leszek Demkowicz, "Discrete least-squares finite element methods," ICES REPORT 17-09, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, May 2017.*

# DISCRETE LEAST-SQUARES FINITE ELEMENT METHODS

DEDICATED TO J. T. ODEN ON THE OCCASION OF HIS 80<sup>TH</sup> BIRTHDAY.

BRENDAN KEITH, SOCRATIS PETRIDES, FEDERICO FUENTES, AND LESZEK DEMKOWICZ

**ABSTRACT.** A finite element methodology for large classes of variational boundary value problems is defined which involves discretizing two linear operators: (1) the differential operator defining the spatial boundary value problem; and (2) a Riesz map on the test space. The resulting linear system is overdetermined. Two different approaches for solving the system are suggested (although others are discussed): (1) solving the associated normal equation with linear solvers for symmetric positive-definite systems (e.g. Cholesky factorization); and (2) solving the overdetermined system with orthogonalization algorithms (e.g. QR factorization). The finite element assembly algorithm for each of these approaches is described in detail. The normal equation approach is usually faster for direct solvers and requires less storage. The second approach reduces the condition number of the system by a power of two and is less sensitive to round-off error. The rectangular stiffness matrix of second approach is demonstrated to have condition number  $\mathcal{O}(h^{-1})$  for a variety of formulations of Poisson's equation. The stiffness matrix from the normal equation approach is demonstrated to be related to the monolithic stiffness matrices of least-squares finite element methods and it is proved that the two are identical in some cases. An example with Poisson's equation indicates that the solutions of these two different linear systems can be nearly indistinguishable (if round-off error is not an issue) and rapidly converge to each other. The orthogonalization approach is suggested to be beneficial for problems which induce poorly conditioned linear systems. Experiments with Poisson's equation in single-precision arithmetic as well as the linear acoustics problem near resonance in double-precision arithmetic verify this conclusion. The methodology described here was developed as an outgrowth of the discontinuous Petrov-Galerkin (DPG) methodology of Demkowicz and Gopalakrishnan [29]. The strength of DPG is highlighted throughout, however, the majority of theory presented is general. Extensions to constrained minimization principles are also considered throughout but are not analyzed in experiments.

## 1. INTRODUCTION

Least-squares finite element methods have been demonstrated to be an auspicious class of methods for a wide variety of boundary value problems of engineering interest. These methods are attractive for many challenging problems because of their general simplicity of implementation, their adherence to simple minimization principles and, therefore, numerical stability, and their built-in *a posteriori* error estimator. As we demonstrate in Section 5.1, the monolithic coefficient matrix in a least-squares finite element method can be identified with the coefficient matrix in the normal equation corresponding to a feasible linear least-squares problem. For a thorough study of least-squares methods and the most significant references, we refer to Bochev and Gunzburger [11].

In this article, we present *discrete least-squares* (DLS) finite element methods. As the name suggests, they can be related to least-squares finite element methods, however, the principles governing DLS methods are much more general. Least-squares finite element methods always induce monolithic, Hermitian positive-definite stiffness matrices. On the other hand, DLS methods can generate an overdetermined system of equations, as well as a related normal equation with a Hermitian positive-definite stiffness matrix.

All DLS methods involve a minimum residual principle and a discretization procedure where two operators are discretized: (1) The differential operator defining the spatial boundary value problem; and (2) a test space Riesz map. As we show in Section 3, this induces a discrete minimum residual principle for the solution vector  $\mathbf{u}$  which can be expressed as a least-squares problem of the form

$$\min_{\mathbf{u} \in \mathbb{R}^N} \|\mathbf{B}\mathbf{u} - \mathbf{l}\|_{\mathbf{G}^{-1}} \iff \mathbf{u} = (\mathbf{B}^* \mathbf{G}^{-1} \mathbf{B})^{-1} \mathbf{B}^* \mathbf{G}^{-1} \mathbf{l},$$

where the field  $\mathbb{F}$  is either  $\mathbb{R}$  or  $\mathbb{C}$ ,  $G \in \mathbb{F}^{M \times M}$  is Hermitian positive (semi-)definite and  $B \in \mathbb{F}^{M \times N}$  is rectangular,  $M > N$ . Another useful perspective is that, by duality, all DLS methods can be identified with the special equilibrium system [68]:

$$(1.1) \quad \begin{bmatrix} G & B \\ B^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \end{bmatrix}.$$

Here,  $\mathbf{u}$  is the subvector of primary interest,  $G$  can often be chosen by the numerical analyst, and there is always the orthogonality condition in the dual variable:  $B^* \mathbf{p} = 0$ . This leads to many special features of DLS not common in most finite element saddle-point problems.

Our principal examples will pertain to a specific class of DLS methods: discontinuous Petrov-Galerkin (DPG) methods [29]. In this setting, the matrix  $G$  can be made block-diagonal, no matter the order of the interpolation spaces used in discretization. It is this feature that makes DPG methods the most practical and general DLS setting we are presently aware of. However, because the discrete saddle-point problem (1.1) can appear in several non-DPG finite element settings [73, 24, 54, 20, 14, 55], it is appropriate to consider this entire class of methods from a comprehensive perspective.

In some DLS methods, both subvectors in (1.1) are jointly solved for because  $G$  cannot be efficiently inverted [24, 20]. Because  $G$  *can* be efficiently inverted with DPG, in that setting, much smaller problems, posed solely in the primal variable  $\mathbf{u}$ , can be solved directly. This has been performed for many problems of engineering interest [65, 21, 67, 33, 37, 49, 62, 30, 40, 38, 35].

In the DPG literature, the minimum residual principle is usually introduced in an idealized scenario where the test space Riesz map is known exactly—that is, it has not been discretized—and from this preternatural position, many notable properties of the method are inferred. Only later, when presenting a computer implementation of the method, a tunable discretization of the test space is proposed with the assumption that all properties of the idealized method—such as stability—will be inherited once the test space discretization is rich enough. In some scenarios, these assumptions can be proven [45, 18, 19, 35, 40, 39] or analyzed [56], but for difficult problems, they can only be speculated. We will not devote our attention in this article to such implications of different discretizations, but rather deliberate the structure of these equations mainly from a numerical linear algebra perspective. Finally, when the Riesz map in a DPG method is discretized, this is referred to as the “practical” DPG method by the majority of the community. It is only in this sense that DPG can be considered a DLS method.

One advantage of the perspective we have taken is clear in Section 3.1, where DPG is identified with an overdetermined system of equations. It is probably because the computer implementation of the practical DPG method has only begun to be investigated in its own right for efficiency and accuracy in large-scale computations [64, 66, 4], that this connection with an overdetermined discrete least-squares problem was overlooked, in the past. Furthermore, by exposing the connection with DLS, this article demonstrates that the established manner of assembling a practical DPG problem can affect the round-off error in the solution in a substantial way.

*Critical assumptions.* An additional benefit of the generality of this work is its subtle suggestions at efficient possibilities for DLS outside the DPG setting. Indeed, even in a DPG method, the “ideal” Gram matrix  $G_{\text{ideal}}$ , coming from an idealized norm, is usually substituted with a more easily computable Gram matrix  $G$  coming from a similar norm. If the norms are equivalent, then only the *converged* solution of the dual variable is affected by this modification of the test norm, and not the *converged* solution of the primal variable. Because the exact  $G_{\text{ideal}}^{-1}$  and the exact  $G^{-1}$  could vary greatly—indeed, one could be dense and the other sparse—it is possible that satisfactory estimates of  $G_{\text{ideal}}^{-1}$  may be obtainable without using a block diagonal  $G$ .

Therefore, we state that the three principal scenarios where DLS is most appropriate are:

- (1) When test functions are constructed so that they are orthogonal under the chosen test space norm. In this situation, the Gram matrix  $G$  would be diagonal and its inverse would be trivial to compute.
- (2) When the test space is broken—as is the case with DPG methods—the Gram matrix  $G$  can be directly inverted, because the computation would be local and feasible.
- (3) When the Gram matrix  $G$  has a sparse inverse [69], or it is only approximately inverted but its approximate inverse is sparse.

*Specific remarks for the DPG community.* As far as we, the authors, are aware, in all published papers on the DPG method/methodology, the so-called “practical” DPG method for ultraweak formulations (with broken test spaces) is identical to solving the normal equation of the DLS system presented within. There appear to be no published manuscripts which derive or consider the possible advantages inherent in solving the overdetermined system of equations.

We note that in specific instances where the test space is  $L^2$  in at least one component—and therefore, for most alternative variational formulations [19, 48]—the implementation of DPG permits some ambiguity. For instance, as developed in [48] and contrary to the DLS approach, one can choose not to discretize the  $L^2$ -Riesz map and so construct a least-squares (strong variational formulations) or least-squares hybrid (all other variational formulations) discretization.

The authors believe that this ultimately leads to some confusion. Different discretizations will involve different analysis, different advantages and limitations, and ultimately different results, however slight. We argue to adopt a convention to distinguish between these approaches. That is, the discretization advocated for here should aptly be called the *discrete* least-squares discretization, and the least-squares and hybrid discretizations developed in [48] should be identified as such.

We further advocate that the “ideal” DPG methodology remain as the consummate—however, often impractical—method which derives from the union of a minimum residual principle with broken test spaces and a Riesz map which is manipulated *analytically*. With this distinction, the ideal DPG methodology applied to a strong formulation is identical to traditional least-squares formulations while in the DLS setting, it is a different and unique numerical method.

*Other pertinent remarks, definitions, and notation.* Since we are synthesizing ideas from different fields, we require a certain amnesty on behalf of the reader throughout this article. For example, let  $M \in \mathbb{F}^{M \times N}$  be full rank matrix,  $M > N$ , and  $\mathbf{b} \in \mathbb{F}^M$ . In the least-squares context, the Gram (or Gramian) matrix is usually defined to be the matrix  $M^*M$  in the normal equation

$$M^*M\mathbf{x} = M^*\mathbf{b}.$$

In this article, we will require significantly different notion of Gram matrix. Accordingly, because our perspective is towards finite element methods, we will generally refer to matrices of the form  $M^*M$  as *stiffness matrices*. Analogously, rectangular matrices like  $M$  will be referred to as an *enriched stiffness matrices*.

For most methods we have in mind, the trial functions  $\mathbf{u} = (u_1, \dots, u_n)$  and test functions  $\mathbf{v} = (v_1, \dots, v_m)$ , have many components. The boundary value problems will be posed on simply connected Lipschitz domains  $\Omega \subset \mathbb{R}^d$ ,  $d > 0$ , with sufficiently smooth boundaries  $\partial\Omega$ . The symbol  $\mathcal{T}$  will denote regular partitioning of  $\Omega$  into elements  $K \in \mathcal{T}$ ,  $\bigcup_{K \in \mathcal{T}} \bar{K} = \bar{\Omega}$  and  $K \cap K' = \emptyset$  for all  $K \neq K' \in \mathcal{T}$ . Here, all elements,  $K$  will be assumed to have sufficient smooth boundaries  $\partial K$ , themselves.

For a matrix  $N \in \mathbb{F}^{M \times N}$  of rank  $R < \min(M, N)$ , the condition number of  $N$  is defined as

$$\text{cond}(N) = \frac{\sigma_1}{\sigma_R},$$

where  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R > 0$  are the nonzero singular values of  $N$ . As it will be useful later, note that if  $N = M^*M$ , then  $\text{cond}(N) = \text{cond}(M)^2$  [44].

**Outline.** We attempt to keep the theory of discrete least-squares conspicuously general in Sections 2 and 3. This allows us to draw connections with many other methods in the vast finite element literature as well as highlight the advantages of the DPG setting. In Section 2, we define the Riesz map and introduce notation which allows us to rigorously analyze residual minimization over Hilbert spaces. Here, we give several examples of methods which fall into the DLS framework. In Section 3, we identify DLS with a generalized linear least-squares problem and reduce it to an overdetermined system of equations and associated normal equation. Here, we also demonstrate how to augment these different DLS systems to incorporate boundary conditions as well as compare different direct techniques to solve them. Lastly, we generalize static condensation for the overdetermined system. In Sections 4 and 5, we narrow our focus to DPG methods. In Section 4, we present assembly algorithms in the DPG context. Here, we distinguish between two separate general algorithms; assembly of the normal

equation and assembly of the overdetermined system. In Section 5, we illustrate the advantage of the discrete least-squares perspective in several numerical examples including the Poisson equation and the linear acoustics problem. In Section 6 we discuss future possibilities for DLS methods and in Section 7 we summarize our findings.

Appendix A briefly discusses extensions of the DLS approach to constrained minimization problems in quadratic programming and Appendix B provides a second perspective on the DLS static condensation procedure of Section 3.4.

## 2. MINIMUM RESIDUAL PRINCIPLES

Due to some technicalities which we wish to avoid for clarity, in this section we consider  $\mathbb{F} = \mathbb{R}$ . We leave the generalization to  $\mathbb{F} = \mathbb{C}$  for the reader. We now begin by defining the vital operators.

**2.1. The Riesz map and variational boundary value problems.** In the abstract setting, a variational boundary value problem is defined with a bilinear form,  $b : \mathcal{U} \times \mathcal{V} \rightarrow \mathbb{R}$ , where  $\mathcal{U}$  is the *trial space*, and  $\mathcal{V}$  is the *test space*. Assuming that  $\mathcal{U}$  and  $\mathcal{V}$  are both Hilbert, they are complete in their assigned norms which are induced by inner products and denoted  $\|\cdot\|_{\mathcal{U}} = (\cdot, \cdot)_{\mathcal{U}}^{1/2}$  and  $\|\cdot\|_{\mathcal{V}} = (\cdot, \cdot)_{\mathcal{V}}^{1/2}$ , respectively.

By the Riesz representation theorem, an inner product induces a unique isometric isomorphism between a Hilbert space and its dual [23, 57]. Specifically, letting  $\mathcal{W}$  be either  $\mathcal{U}$  or  $\mathcal{V}$ , there exists a unique bijective linear map  $\mathcal{R}_{\mathcal{W}} : \mathcal{W} \rightarrow \mathcal{W}'$  such that  $\|\mathcal{R}_{\mathcal{W}} \mathbf{w}\|_{\mathcal{W}'} = \|\mathbf{w}\|_{\mathcal{W}}$  for all  $\mathbf{w} \in \mathcal{W}$ . This operator is called the Riesz map and is defined by the variational equation

$$(2.1) \quad \langle \mathcal{R}_{\mathcal{W}} \mathbf{w}, \delta \mathbf{w} \rangle = (\delta \mathbf{w}, \mathbf{w})_{\mathcal{W}} \quad \text{for all } \mathbf{w}, \delta \mathbf{w} \in \mathcal{W}.$$

Following from the theorem,  $\|\mathbf{w}'\|_{\mathcal{W}'} = \|\mathcal{R}_{\mathcal{W}}^{-1} \mathbf{w}'\|_{\mathcal{W}}$  for all  $\mathbf{w}' \in \mathcal{W}'$ ,  $\mathcal{R}_{\mathcal{W}} = \mathcal{R}'_{\mathcal{W}}$ , and, moreover,  $\mathcal{R}_{\mathcal{W}'} = \mathcal{R}_{\mathcal{W}}^{-1}$  (under the identification  $\mathcal{W} \sim \mathcal{W}''$ ).

Consider, for instance, the circumstance that  $\mathcal{W} = L^2(\Omega)$  for some domain  $\Omega \subset \mathbb{R}^N$ . Here,  $\mathcal{R}_{L^2(\Omega)}$  defines the natural identification between  $L^2(\Omega)$  and its dual:

$$(2.2) \quad \langle \mathcal{R}_{L^2(\Omega)} v, w \rangle = (w, v)_{L^2(\Omega)} = \int_{\Omega} wv \quad \text{for all } w, v \in L^2(\Omega).$$

Since this is a point of confusion at times, notice that  $\mathcal{R}_{L^2(\Omega)} v \notin L^2(\Omega)$ . That is, it is technically *not* an identity operator, but is instead an element of  $(L^2(\Omega))'$ .

An abstract variational boundary value problem is a problem of finding a unique trial space element  $\mathbf{u} \in \mathcal{U}$  complementing a given load  $\ell \in \mathcal{V}'$  such that

$$b(\mathbf{u}, \mathbf{v}) = \ell(\mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathcal{V}.$$

Similar to the inner product in (2.1), the bilinear form  $b$  induces an important linear operator from the trial space to the dual of the test space,  $\mathcal{B} : \mathcal{U} \rightarrow \mathcal{V}'$ :

$$(2.3) \quad \langle \mathcal{B}\mathbf{u}, \mathbf{v} \rangle = b(\mathbf{u}, \mathbf{v}) \quad \text{for all } \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}.$$

In Section 3, we will see that the Gram matrix and enriched stiffness matrix formed in DLS methods are discrete analogues of Hilbert space operators  $\mathcal{R}_{\mathcal{V}}$  and  $\mathcal{B}$ . First, however, we must explore the idealized minimum residual principles which DLS methods derive from. These principles lead to problems which must be discretized and it is the choice made in their discretization which distinguishes DLS as a unique finite element methodology.

**2.2. Minimum residual principles on linear subspaces.** Here, we derive the idealized Euler-Lagrange equation which DLS emanates from. As we will remark again, later, this framework is presented in a general setting which can be applied to many different problems including that of least-squares finite element methods.

Let  $\mathcal{U}_h \subset \mathcal{U}$ , be the discrete trial space, where the computed solution will reside. A minimum residual method delivers the optimal discrete solution,  $\mathbf{u}_h^{\text{opt}}$ , of the residual minimization problem

$$(2.4) \quad \mathbf{u}_h^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}'}^2.$$

This is the principle supporting Ritz methods, least-squares methods, DPG methods, and general DLS methods, as we will remark later. Using the Riesz map for the test space, (2.4) can be recast in the following form:

$$(2.5) \quad \mathbf{u}_h^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \langle \mathcal{R}_{\mathcal{V}}^{-1}(\mathcal{B}\mathbf{u}_h - \ell), \mathcal{B}\mathbf{u}_h - \ell \rangle_{\mathcal{V} \times \mathcal{V}'}.$$

For the rest of the document we omit the subscripts on the duality pairing  $\langle \cdot, \cdot \rangle$ , and the spaces are understood based upon the context.

From here, the vanishing Gâteaux derivative at the global minimum implies a variational equation for  $\mathbf{u}_h^{\text{opt}}$ , namely,

$$(2.6) \quad \langle \mathcal{R}_{\mathcal{V}}^{-1} \mathcal{B}\mathbf{u}_h^{\text{opt}}, \mathcal{B}\delta\mathbf{u}_h \rangle = \langle \mathcal{R}_{\mathcal{V}}^{-1} \ell, \mathcal{B}\delta\mathbf{u}_h \rangle \quad \text{for all } \delta\mathbf{u}_h \in \mathcal{U}_h.$$

By introducing the operator  $\mathcal{A} = \mathcal{B}' \mathcal{R}_{\mathcal{V}}^{-1} \mathcal{B}$  and the modified load  $f = \mathcal{B}' \mathcal{R}_{\mathcal{V}}^{-1} \ell \in \mathcal{U}'$ , the variational equation (2.6) can be expressed as

$$\langle \mathcal{A}\mathbf{u}_h^{\text{opt}}, \delta\mathbf{u}_h \rangle = f(\delta\mathbf{u}_h) \quad \text{for all } \delta\mathbf{u}_h \in \mathcal{U}_h.$$

It is helpful to notice that this new operator,  $\mathcal{A} = (\mathcal{R}_{\mathcal{V}}^{-1/2} \mathcal{B})'(\mathcal{R}_{\mathcal{V}}^{-1/2} \mathcal{B})$ , is positive definite and can be identified with an inner product on  $\mathcal{U}$ ,

$$a(\delta\mathbf{u}, \mathbf{u}) = \langle \mathcal{A}\mathbf{u}, \delta\mathbf{u} \rangle \quad \text{for all } \mathbf{u}, \delta\mathbf{u} \in \mathcal{U}.$$

Therefore, it has the interpretation of a Riesz map on the trial space.

**2.3. Examples of minimum residual principles.** Here, we represent some recognizable methods with the minimum residual principle as well as illustrate some convenient properties of the DPG setting.

*Least-squares.* In a least-squares finite element method, a symmetric bilinear form is usually derived in the following way: Beginning with a linear operator  $\mathcal{L} : \mathcal{U} \rightarrow L^2(\Omega)$  and a load function  $f \in L^2(\Omega)$ , seek the solution of the minimization problem

$$(2.7) \quad \mathbf{u}_h^{\text{LS}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \|\mathcal{L}\mathbf{u}_h - f\|_{L^2(\Omega)}^2.$$

By identifying  $\mathcal{V} = L^2(\Omega)$ ,  $\mathcal{B} = \mathcal{R}_{L^2(\Omega)} \mathcal{L}$ , and  $\ell = \mathcal{R}_{L^2(\Omega)} f$ , we see that (2.4) and (2.7) are equivalent. Indeed, observe that the bilinear form  $b$  can be expressed as

$$b(\mathbf{u}, \mathbf{v}) = \langle \mathcal{R}_{L^2(\Omega)} \mathcal{L}\mathbf{u}, \mathbf{v} \rangle = (\mathcal{L}\mathbf{u}, \mathbf{v})_{L^2(\Omega)}, \quad \text{for all } \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V},$$

and likewise,  $\ell(\mathbf{v}) = (f, \mathbf{v})_{L^2(\Omega)}$ . Moreover, the residual in (2.4) can be expressed

$$\begin{aligned} \|\mathcal{B}\mathbf{u}_h - \ell\|_{L^2(\Omega)'}^2 &= \langle \mathcal{R}_{L^2(\Omega)}^{-1} \mathcal{R}_{L^2(\Omega)}(\mathcal{L}\mathbf{u}_h - f), \mathcal{R}_{L^2(\Omega)}(\mathcal{L}\mathbf{u}_h - f) \rangle \\ &= \langle \mathcal{R}_{L^2(\Omega)}(\mathcal{L}\mathbf{u}_h - f), \mathcal{L}\mathbf{u}_h - f \rangle \\ &= (\mathcal{L}\mathbf{u}_h - f, \mathcal{L}\mathbf{u}_h - f)_{L^2(\Omega)} \\ &= \|\mathcal{L}\mathbf{u}_h - f\|_{L^2(\Omega)}^2. \end{aligned}$$

Therefore,  $\mathbf{u}_h^{\text{LS}} = \mathbf{u}_h^{\text{opt}}$ . Similarly, variational problem (2.6) is equivalent to

$$(2.8) \quad (\mathcal{L}\mathbf{u}_h, \mathcal{L}\delta\mathbf{u}_h)_{L^2(\Omega)} = (f, \mathcal{L}\delta\mathbf{u}_h)_{L^2(\Omega)}, \quad \text{for all } \mathbf{u}_h \in \mathcal{U}_h,$$

and the operator  $\mathcal{A} = \mathcal{R}_{\mathcal{U}} \mathcal{L}^* \mathcal{L}$ , where  $\mathcal{L}^* = \mathcal{R}_{\mathcal{U}}^{-1} \mathcal{L}' \mathcal{R}_{L^2(\Omega)} = \mathcal{R}_{\mathcal{U}}^{-1} \mathcal{B}'$  is the Hilbert-adjoint operator of  $\mathcal{L}$ .

*Galerkin methods for symmetric, coercive boundary value problems.* An important variety of linear variational problems, also known as Ritz methods, can be derived from an energy minimization principle

$$(2.9) \quad \mathbf{u}_h^{\text{en}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \mathcal{J}(\mathbf{u}_h),$$

where  $\mathcal{J}(\mathbf{u}) = \frac{1}{2}a(\mathbf{u}, \mathbf{u}) - \ell(\mathbf{u})$ . Here,  $a : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$  is a symmetric, coercive bilinear form, and  $\ell \in \mathcal{U}'$  is a linear form. The corresponding variational equation is

$$(2.10) \quad a(\mathbf{u}_h^{\text{en}}, \delta \mathbf{u}_h) = \ell(\delta \mathbf{u}_h), \quad \text{for all } \delta \mathbf{u}_h \in \mathcal{U}_h.$$

In many physical problems, duality methods can be used to equate the minimization principle of (2.9) with a saddle point problem

$$\begin{pmatrix} \mathcal{C}^{-1} & \mathcal{D} \\ \mathcal{D}^* & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} g \\ f \end{pmatrix},$$

where  $\mathcal{D} : \mathcal{U} \rightarrow L^2(\Omega)$  is differential operator,  $\mathbf{p} \in L^2(\Omega)$ , and  $\mathcal{C} : L^2(\Omega) \rightarrow L^2(\Omega)$  is symmetric and positive definite. (See [12, 68] for more generality.) In this case,

$$a(\mathbf{u}, \mathbf{u}) = (\mathcal{C} \mathcal{D} \mathbf{u}, \mathcal{D} \mathbf{u})_{L^2(\Omega)} \quad \text{and} \quad \ell(\mathbf{u}) = (\mathcal{C} \mathcal{D} \mathbf{u}, g)_{L^2(\Omega)} - (f, \mathbf{u})_{L^2(\Omega)}.$$

If  $f = 0$ , (2.9) can be identified with the least-squares problem

$$(2.11) \quad \mathbf{u}_h^{\text{en}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \|\mathcal{C}^{1/2} \mathcal{D} \mathbf{u} - \mathcal{C}^{1/2} g\|_{L^2(\Omega)}^2,$$

which can, likewise, be equated with a minimum residual principle. Although this formulation has been exploited in some finite element contexts [12, 73], the applicability of an energy principle of the form (2.11) is severely limited.

We now outline a more fruitful procedure. Analogous to (2.3), we may define the operator

$$\langle \mathcal{B} \mathbf{u}, \mathbf{v} \rangle = a(\mathbf{u}, \mathbf{v}), \quad \text{for all } \mathbf{u}, \mathbf{v} \in \mathcal{U}.$$

We may also identify  $a$  with a norm on the “test space”,  $\mathcal{V} = \mathcal{U}$ , viz.,

$$\|\mathbf{v}\|_{\mathcal{V}}^2 = a(\mathbf{v}, \mathbf{v}), \quad \text{for all } \mathbf{v} \in \mathcal{U}.$$

Thus,  $\mathcal{R}_{\mathcal{V}} = \mathcal{B}$ . With these definitions we see that

$$\|\mathcal{B} \mathbf{u}_h - \ell\|_{\mathcal{V}'}^2 = \langle \mathcal{B}^{-1}(\mathcal{B} \mathbf{u}_h - \ell), \mathcal{B} \mathbf{u}_h - \ell \rangle = \langle \mathcal{B} \mathbf{u}_h - \ell, \mathbf{u}_h \rangle - \langle \mathcal{B} \mathbf{u}_h - \ell, \mathcal{B}^{-1} \ell \rangle = 2\mathcal{J}(\mathbf{u}_h) + \langle \ell, \mathcal{B}^{-1} \ell \rangle.$$

Likewise,

$$\mathbf{u}_h^{\text{en}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \mathcal{J}(\mathbf{u}_h) = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \|\mathcal{B} \mathbf{u}_h - \ell\|_{\mathcal{V}'}^2 = \mathbf{u}_h^{\text{opt}}.$$

We have just shown that (2.9) can be expressed as a minimum residual problem and a similar construction is possible with more general (e.g. non-symmetric) variational equations. Unfortunately, the expense of forming a discrete analogue of  $\mathcal{R}_{\mathcal{V}}^{-1} = \mathcal{B}^{-1}$  usually surpasses the expense of discretizing (2.10). For this reason, analysis of such a problem in a minimum residual setting is useless unless we move to a broken variational formulation and a DPG method, which we now present the main highlights of.

*DPG methods.* Consider a mesh,  $\mathcal{T}$ , where  $\bigcup_{K \in \mathcal{T}} \overline{K} = \overline{\Omega}$  and  $K \cap K' = \emptyset$  for all  $K \neq K' \in \mathcal{T}$ . We require the following definitions:

- A test space,  $\mathcal{V}^{\text{br.}}$ , is *broken* if it does not require in its members any form of continuity (conformity) across element boundaries,  $\partial K$ .
- A test space norm,  $\|\cdot\|_{\mathcal{V}^{\text{loc.}}}$ , is *localizable* if it induces a finite orthogonal direct sum,  $\mathcal{V}^{\text{loc.}} = \bigoplus_{K \in \mathcal{T}} \mathcal{V}_K$ , where  $\mathbf{v}_K|_{K'} = 0$  for all  $\mathbf{v}_K \in \mathcal{V}_K$ ,  $K \neq K'$ .

In a DPG method, the test space,  $\mathcal{V}^{\text{DPG}}$ , must be broken and the corresponding test space norm,  $\|\cdot\|_{\mathcal{V}^{\text{DPG}}}$ , must be localizable. Here, the methodology is often derived with precisely the minimum residual principle (2.4) as well as these two assumptions, although other perspectives are also possible [29].

The broken test space assumption allows the use of discontinuous basis functions in DPG methods, but it is the localizable norm assumption which ensures that the resulting Gram matrix will be block-diagonal (see Section 3.1).

In general, with a slight abuse of notation, the trial space in a DPG method involves two components:  $\mathcal{U}^{\text{DPG}} = \mathcal{U}^{\text{fld.}} \times \hat{\mathcal{U}}$ , where  $\hat{\mathcal{U}}$  can be interpreted as a space of Lagrange multipliers due to the broken nature of the test space. Similarly, the operator  $\mathcal{B}^{\text{DPG}}$  in a DPG method can be decomposed into two terms [19]:

$$(2.12) \quad \mathcal{B}^{\text{DPG}}(\mathbf{u}^{\text{fld.}}, \hat{\mathbf{u}}) = \mathcal{B}^{\text{fld.}}\mathbf{u}^{\text{fld.}} + \hat{\mathcal{B}}\hat{\mathbf{u}}.$$

At the infinite-dimensional level, the direct sum decomposition from a localizable norm allows the action of the Riesz map to decouple into its independent element contributions. For instance, given a fixed element  $K \in \mathcal{T}$ , an element-local test function  $\mathbf{v}_K \in \mathcal{V}_K$ , and an arbitrary test function decomposed into element-local contributions,  $\delta\mathbf{v}^{\text{DPG}} = \sum_{K' \in \mathcal{T}} \delta\mathbf{v}_{K'}$ , then

$$(2.13) \quad \langle \mathcal{R}_{\mathcal{V}^{\text{DPG}}} \mathbf{v}_K, \delta\mathbf{v}^{\text{DPG}} \rangle = \sum_{K' \in \mathcal{T}} \langle \mathbf{v}_K, \delta\mathbf{v}_{K'} \rangle_{\mathcal{V}^{\text{DPG}}} = \langle \mathbf{v}_K, \delta\mathbf{v}_K \rangle_{\mathcal{V}^{\text{DPG}}} = \langle \mathcal{R}_{\mathcal{V}_K} \mathbf{v}_K, \delta\mathbf{v}_K \rangle.$$

That is,  $\mathcal{R}_{\mathcal{V}^{\text{DPG}}} \mathbf{v}_K : \mathcal{V}^{\text{DPG}} \rightarrow \mathbb{F}$  only responds to the nonzero contribution of  $\delta\mathbf{v}^{\text{DPG}}$  within  $K$ .

Notably, the locality of the Riesz map on a broken test space also exists in its inverse; that is,  $\mathcal{R}_{\mathcal{V}^{\text{DPG}}}^{-1} = \bigoplus_{K \in \mathcal{T}} \mathcal{R}_{\mathcal{V}_K}^{-1}$ . Therefore, the residual can be decomposed into a *single* sum over the elements of  $\mathcal{T}$ :

$$(2.14) \quad \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}^{\text{DPG}}}^2 = \sum_{K \in \mathcal{T}} \langle \mathcal{B}\mathbf{u}_h - \ell, \mathcal{R}_{\mathcal{V}_K}^{-1}(\mathcal{B}\mathbf{u}_h - \ell) \rangle = \sum_{K \in \mathcal{T}} \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}_K}^2.$$

Here, each individual term making up the right-hand sum in (2.14) induces an *a posteriori* error estimate which, for each element  $K \in \mathcal{T}$ , is denoted  $\eta_K^2 = \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}_K}^2$ . These error estimators are often incorporated into adaptive mesh refinement strategies [18].

To identify a connection with least-squares methods, observe that  $L^2$  is a broken test space and  $\mathcal{R}_{L^2(\Omega)} = \bigoplus_{K \in \mathcal{T}} \mathcal{R}_{L^2(K)}$  and  $\hat{\mathcal{B}} = 0$ . Likewise,

$$\|\mathcal{L}\mathbf{u}_h - f\|_{L^2(\Omega)}^2 = \sum_{K \in \mathcal{T}} \|\mathcal{L}\mathbf{u}_h - f\|_{L^2(K)}^2,$$

and each localized residual above,  $\eta_K = \|\mathcal{L}\mathbf{u}_h - f\|_{L^2(K)}$ , can be used as an *a posteriori* error estimator [6].

**2.4. Non-homogeneous boundary conditions and linear equality constraints.** The minimum residual principle for a variational boundary value problem with non-homogeneous essential boundary conditions generalizes naturally from (2.4). In many physical problems, a selection of components of the trace of the solution are explicitly declared on the boundary of  $\Omega$ . That is,  $u_i = \hat{u}_i$  on  $\partial\Omega$ , for some number of prescribed functions  $\hat{u}_i$ ,  $i \in \mathcal{I} \subset \{1, \dots, n\}$ . The set of all possible prescribed data, in all components, is often a Hilbert space itself—which we will denote  $\hat{\mathcal{U}} = \hat{\mathcal{U}}_1 \times \dots \times \hat{\mathcal{U}}_n$ —with an associated continuous linear surjection  $\text{tr}_{\mathcal{U}} = \text{tr}_{\mathcal{U}_1} \times \dots \times \text{tr}_{\mathcal{U}_n} : \mathcal{U} \rightarrow \hat{\mathcal{U}}$ , called a trace map. With the trace map on  $\mathcal{U}$ , we may take into account the boundary conditions to hone the solution space into only the affine set  $\mathcal{K} = \bigcap_{i \in \mathcal{I}} \text{tr}_{\mathcal{U}_i}^{-1}\{\hat{u}_i\} \subset \mathcal{U}$ .

Observe that if any single element  $\mathbf{u}^{\text{lift}} \in \mathcal{K}$  is isolated, this so-called “lift” can be used to recharacterize the affine set:  $\mathcal{K} = \mathcal{U}^{\text{hom.}} + \mathbf{u}^{\text{lift}} \subset \mathcal{U}$ , where  $\mathcal{U}^{\text{hom.}} = \bigcap_{i \in \mathcal{I}} \text{tr}_{\mathcal{U}_i}^{-1}\{0\} = \bigcap_{i \in \mathcal{I}} \text{Null}(\text{tr}_{\mathcal{U}_i})$  is a linear subspace of  $\mathcal{U}$ . This suggests a procedure which is often performed in precomputation for standard finite element methods. Indeed, it is common practice to isolate a discrete lift,  $\mathbf{u}_h^{\text{lift}}$ , which optimally interpolates each component of the solution onto its associated boundary condition in a discrete solution space  $\mathcal{U}_h$ . (See [27] for an abstract description of traditional finite element methods.) With the lift computed, the problem can then be posed over  $\mathcal{K}_h = \mathcal{U}_h^{\text{hom.}} + \mathbf{u}_h^{\text{lift}} \subset \mathcal{U}_h$ , where  $\mathcal{U}_h^{\text{hom.}} = \mathcal{U}_h \cap \mathcal{U}^{\text{hom.}}$ .



Recall that  $b : \mathcal{U} \times \mathcal{V} \rightarrow \mathbb{R}$ . Abstractly, for such explicit essential boundary conditions, we may therefore begin with an affine subspace  $\mathcal{U}_h^{\text{hom.}} + \mathbf{u}_h^{\text{lift}} \subset \mathcal{U}$  and consider the variational problem

$$(2.15) \quad \begin{cases} \text{Find } \mathbf{u}_h \in \mathcal{U}_h^{\text{hom.}} + \mathbf{u}_h^{\text{lift}} : \\ b(\mathbf{u}_h, \mathbf{v}) = \ell(\mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathcal{V}. \end{cases}$$

The corresponding minimum residual principle is

$$(2.16) \quad \mathbf{u}_h^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h^{\text{hom.}} + \mathbf{u}_h^{\text{lift}}} \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}'}^2 = \arg \min_{\mathbf{u}_h^{\text{hom.}} \in \mathcal{U}_h^{\text{hom.}}} \|\mathcal{B}\mathbf{u}_h^{\text{hom.}} - (\ell - \mathcal{B}\mathbf{u}_h^{\text{lift}})\|_{\mathcal{V}'}^2 + \mathbf{u}_h^{\text{lift}}.$$

As is plain to see, the minimum residual principle of the explicit boundary condition with an associated lift falls under the subspace minimization theory of Section 2.2.

Unfortunately, not all important boundary conditions can be represented by an explicit prescription of the values of the solution components on  $\partial\Omega$ . A more general scenario requires the existence of a surjective linear operator  $\hat{\mathcal{C}} : \hat{\mathcal{U}} \rightarrow \hat{\mathcal{W}}$ , where  $\hat{\mathcal{W}}$  is an unspecified Hilbert space, for the sake of abstraction. With this operator, we may define a broad range of possible boundary conditions,  $\hat{\mathcal{C}}(\text{tr}_{\mathcal{U}}\mathbf{u}) = \hat{\mathbf{g}} \in \hat{\mathcal{W}}$ . Moreover, we can restrict the possible solutions to the affine subspace  $\mathcal{K} = \text{tr}_{\mathcal{U}}^{-1}\{\hat{\mathcal{C}}^{-1}\{\hat{\mathbf{g}}\}\} \subset \mathcal{U}$ .

Now, unless  $\hat{\mathcal{C}}$  is the projection map,  $\hat{\pi}_{\mathcal{I}} : \hat{\mathcal{U}} \rightarrow \prod_{i \in \mathcal{I}} \hat{\mathcal{U}}_i$ —in which case we arrive with the explicit boundary conditions above—it is generally difficult to estimate the corresponding lift and homogenous subspace decomposition,  $\mathcal{K}_h = \mathcal{U}_h^{\text{hom.}} + \mathbf{u}_h^{\text{lift}}$ , as before. Because of this issue, we require a generalization of the minimum residual principle (2.4). As it extends naturally to the case of more general linear equality constraints that one may wish to satisfy, we deal with the circumstance in full generality.

Consider the linear operator  $\mathcal{C} : \mathcal{U} \rightarrow \mathcal{W}'$  and the constraint  $\mathcal{C}\mathbf{u} = d$ . A convenient perspective to characterize the associated discrete solution is to search for a unique solution in the implicitly defined, weakly consistent affine subspace  $\mathcal{K}_h = \mathcal{K}(\mathcal{U}_h, \mathcal{W}_h) = \{\mathbf{u}_h \in \mathcal{U}_h \mid \langle \mathcal{C}\mathbf{u}_h, \mathbf{w}_h \rangle = d(\mathbf{w}_h) \forall \mathbf{w}_h \in \mathcal{W}_h\}$ . The corresponding discrete variational problem is

$$(2.17) \quad \begin{cases} \text{Find } \mathbf{u}_h \in \mathcal{K}_h : \\ b(\mathbf{u}_h, \mathbf{v}) = \ell(\mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathcal{V}, \end{cases}$$

and its complementary constrained minimum residual principle is

$$(2.18) \quad \mathbf{u}_h^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{K}_h} \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}'}^2,$$

which can equivalently be expressed as

$$(2.19) \quad \mathbf{u}_h^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \sup_{\mathbf{w}_h \in \mathcal{W}_h} \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}'}^2 + \langle \mathcal{C}\mathbf{u}_h - d, \mathbf{w}_h \rangle.$$

Assuming  $\mathcal{K}_h \neq \emptyset$ , since  $\mathcal{K}_h$  is convex, (2.19) will always have at least one solution [32]. Indeed, existence of a solution to the minimization problem (2.18) can be demonstrated for any convex set  $\mathcal{K}_h$  (see Appendix A).

Choosing an inappropriate balance of discrete function spaces  $\mathcal{U}_h$  and  $\mathcal{W}_h$  can affect the rate of convergence of the optimal solution in (2.19) [12]. In some scenarios, however,  $\mathcal{W}$  is finite dimensional and the constraint space  $\mathcal{W}_h = \mathcal{W}$  is obvious and will not affect convergence rates when paired with most traditional choices for  $\mathcal{U}_h$  (for an example, see [34]).

To avoid potential issues with balancing  $\mathcal{U}_h$  and  $\mathcal{W}_h$ , it is sometimes convenient to consider the following secondary interpretation of optimal solutions to constrained minimization problems. Begin by considering the constraint equation as a second variational problem that we are equally inclined to solve. That is,

$$(2.20) \quad \begin{cases} \text{Find } \mathbf{u}_h \in \mathcal{U}_h : \\ b(\mathbf{u}_h, \mathbf{v}) = \ell(\mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathcal{V} \end{cases} \quad \text{and} \quad \begin{cases} \text{Find } \mathbf{u}_h \in \mathcal{U}_h : \\ c(\mathbf{u}_h, \mathbf{w}) = d(\mathbf{w}) \quad \text{for all } \mathbf{w} \in \mathcal{W}, \end{cases}$$

where  $c(\mathbf{u}, \mathbf{v}) = \langle \mathcal{C}\mathbf{u}_h, \mathbf{v}_h \rangle$  is the bilinear form corresponding to the constraint operator. The optimal solution is then the penalized sum of the residuals from both problems:

$$(2.21) \quad \mathbf{u}_h^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} (\|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}'}^2 + \|\mathcal{C}(\mathbf{u}_h) - d\|_{\mathcal{W}'}^2).$$

This is a penalized minimum residual principle. In many scenarios, neither of the problems in (2.20) will have a unique solution. This does not preclude uniqueness of  $\mathbf{u}_h^{\text{opt}}$  in (2.21).

### 3. DISCRETIZATION

Unlike in the previous section, from now on we consider both cases,  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ , jointly.

**3.1. The overdetermined system.** Denote an ordered basis for  $\mathcal{U}_h$  as  $\mathcal{U}_h = \{\mathbf{u}_i\}_{i=1}^N$ . Similarly, denote a basis for the test space  $\mathcal{V}$  as  $\mathcal{V} = \{\mathbf{v}_i\}_{i \in \mathcal{I}}$  and denote the first  $M > N$  of a countable subset of these functions as the ordered basis,  $\mathcal{V}_r = \{\mathbf{v}_i\}_{i=1}^M$ , for the discrete test space  $\mathcal{V}_r$ . Likewise, the discrete Riesz map,  $\mathcal{R}_{\mathcal{V}_r} : \mathcal{V}_r \rightarrow \mathcal{V}_r'$ , is defined

$$\langle \mathcal{R}_{\mathcal{V}_r} \mathbf{v}, \delta \mathbf{v} \rangle = (\delta \mathbf{v}, \mathbf{v})_{\mathcal{V}} \quad \text{for all } \mathbf{v}, \delta \mathbf{v} \in \mathcal{V}_r = \text{span}(\{\mathbf{v}_i\}_{i=1}^M).$$

With this definition, we formulate the practical minimization problem

$$(3.1) \quad \mathbf{u}_{h,r}^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \|\mathcal{R}_{\mathcal{V}_r}^{-1}(\mathcal{B}\mathbf{u}_h - \ell)\|_{\mathcal{V}}^2.$$

In (3.1) above,  $\mathcal{R}_{\mathcal{V}_r}^{-1} : \mathcal{V}' \rightarrow \mathcal{V}$  is an abuse of notation naturally representing the extension of  $\mathcal{R}_{\mathcal{V}_r}^{-1} : \mathcal{V}_r' \rightarrow \mathcal{V}_r$  by zero. Or, in other terms,  $\mathcal{R}_{\mathcal{V}_r}^{-1} = \mathcal{R}_{\mathcal{V}}^{-1} \circ \iota_r'$ , where  $\iota_r'$  is the dual of the canonical embedding,  $\iota_r : \mathcal{V}_r \rightarrow \mathcal{V}$ .

Although (3.1) generally represents the quadratic minimization problem actually solved in a computer implementation, it is perhaps more naturally expressed in terms of the actual finite element matrices. In such a representation, it can be identified with a (discrete) generalized least-squares problem on the coefficients,  $\mathbf{u} = [\mathbf{u}_i]_{i=1}^N \in \mathbb{F}^N$ , of the solution  $\mathbf{u}_h \in \mathcal{U}_h$  represented in the  $\{\mathbf{u}_i\}_{i=1}^N$  basis,

$$\mathbf{u}_h = \sum_{i=1}^N \mathbf{u}_i \mathbf{u}_i \in \mathcal{U}_h.$$

Indeed, define  $B_{ij} = b(\mathbf{u}_j, \mathbf{v}_i)$ ,  $G_{ij} = (\mathbf{v}_i, \mathbf{v}_j)_{\mathcal{V}}$ , and  $\mathbf{l} = \ell(\mathbf{v}_i)$  for each ordered basis element  $\mathbf{u}_j \in \mathcal{U}$  and  $\mathbf{v}_i, \mathbf{v}_j \in \mathcal{V}_r$ , and define the optimal coefficients to be

$$(3.2) \quad \mathbf{u}^{\text{opt}} = \arg \min_{\mathbf{u} \in \mathbb{F}^N} (\mathbf{B}\mathbf{u} - \mathbf{l})^* \mathbf{G}^{-1} (\mathbf{B}\mathbf{u} - \mathbf{l}).$$

Of course, for  $\mathbf{u}^{\text{opt}} = [\mathbf{u}_i^{\text{opt}}]_{i=1}^N$ ,

$$\mathbf{u}_{h,r}^{\text{opt}} = \sum_{i=1}^N \mathbf{u}_i^{\text{opt}} \mathbf{u}_i,$$

as the reader may verify.

Now, reflect upon the matrix  $\mathbf{G}$ . In place of a general basis  $\mathcal{V}_r$ , had we used an orthonormal basis  $\tilde{\mathcal{V}}_r = \{\tilde{\mathbf{v}}_i\}_{i=1}^M$ , then minimization problem (3.2) would be equivalent to an ordinary linear least-squares problem. That is,

$$(\tilde{\mathbf{B}}\mathbf{u} - \tilde{\mathbf{l}})^* \tilde{\mathbf{G}}^{-1} (\tilde{\mathbf{B}}\mathbf{u} - \tilde{\mathbf{l}}) = \|\tilde{\mathbf{B}}\mathbf{u} - \tilde{\mathbf{l}}\|_2^2,$$

where  $\tilde{B}_{ij} = b(\mathbf{u}_j, \tilde{\mathbf{v}}_i)$ ,  $\tilde{G}_{ij} = (\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j)_{\mathcal{V}} = \delta_{ij}$ , and  $\tilde{\mathbf{l}} = \ell(\tilde{\mathbf{v}}_i)$ , and therefore

$$(3.3) \quad \mathbf{u}^{\text{opt}} = \arg \min_{\mathbf{u} \in \mathbb{F}^N} \|\tilde{\mathbf{B}}\mathbf{u} - \tilde{\mathbf{l}}\|_2^2.$$

As we will now demonstrate, precomputing an orthonormal basis  $\tilde{\mathcal{V}}_r$  is not necessary and can be performed implicitly in the process of inverting  $\mathbf{G}$ .

Indeed, since  $\mathbf{G}$  is positive definite, it is congruent to the identity matrix, so one can arrive at (3.3) by defining  $\tilde{\mathbf{B}} = \mathbf{W}^{-1}\mathbf{B}$  and  $\tilde{\mathbf{l}} = \mathbf{W}^{-1}\mathbf{l}$ , where  $\mathbf{W}$  is any matrix solving the equation  $\mathbf{W}\mathbf{W}^* = \mathbf{G}$ . Such a factorization is

naturally related to finding a discretely  $\mathcal{V}$ -orthogonal change-of-basis transformation. For instance, consider an arbitrary basis vector  $\mathbf{v}_i \in \mathfrak{V}_r$  represented in an orthonormal basis  $\tilde{\mathfrak{V}}_r$ ,

$$\mathbf{v}_i = \sum_{j=1}^M W_{ij} \tilde{\mathbf{v}}_j,$$

where  $(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j)_{\mathcal{V}} = \delta_{ij}$  for all  $\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j \in \tilde{\mathfrak{V}}_r$ . With this expression, the Gram matrix can be represented as

$$G_{ij} = (\mathbf{v}_i, \mathbf{v}_j)_{\mathcal{V}} = \sum_{k,l=1}^M W_{ik} \overline{W_{jl}} (\tilde{\mathbf{v}}_k, \tilde{\mathbf{v}}_l) = \sum_{k=1}^M W_{ik} \overline{W_{jk}}.$$

i.e.  $G = WW^*$ .

If  $M > 1$ , the equation  $WW^* = G$  has an infinite number of solutions but a convenient choice is the (unique) Cholesky factorization,  $G = LL^*$ . For lower-triangular matrix  $W_{\text{Chol.}} = L$ , we can rewrite

$$\begin{aligned} \mathbf{u}^{\text{opt}} &= \arg \min_{\mathbf{u} \in \mathbb{R}^N} (W_{\text{Chol.}}^{-1} (\mathbf{B}\mathbf{u} - \mathbf{l}))^* W_{\text{Chol.}}^{-1} (\mathbf{B}\mathbf{u} - \mathbf{l}) \\ (3.4) \quad &= \arg \min_{\mathbf{u} \in \mathbb{R}^N} \|L^{-1}(\mathbf{B}\mathbf{u} - \mathbf{l})\|_2^2, \end{aligned}$$

and the matrices  $\tilde{B} = L^{-1}B$  and  $\tilde{l} = L^{-1}l$  can be efficiently computed by back-substitution. This choice is usually assumed throughout this paper. In the statistics and signal processing communities, the factoring and row-weighting procedure described above is known as “whitening” or “sphering” and we refer the reader to [50] to explore the benefits of the other possibilities for the change-of-basis matrices  $W$  in that context.

*Synopsis.* Ultimately, discretization of the Riesz map in a minimum residual principle induces a generalized least-squares problem

$$\min_{\mathbf{u} \in \mathbb{R}^N} \|\mathbf{B}\mathbf{u} - \mathbf{l}\|_{G^{-1}},$$

where  $G$ , the Gram matrix, is the discretization of the Riesz map. This is equivalent to the special equilibrium system, or saddle-point problem,

$$(3.5) \quad \begin{bmatrix} G & B \\ B^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \end{bmatrix}.$$

Since  $\mathbf{u}$  is the subvector of primary importance, the statically-condensed equations

$$(3.6) \quad \mathbf{A}\mathbf{u} = \mathbf{f},$$

where  $\mathbf{A} = B^*G^{-1}B$  and  $\mathbf{f} = B^*G^{-1}l$ , can be used to characterize the solution. If  $G$  can be efficiently factored  $G = WW^*$ —this is the case which we are predominantly interested in—it is also useful to rewrite this problem as

$$\min_{\mathbf{u} \in \mathbb{R}^N} \|W^{-1}(\mathbf{B}\mathbf{u} - \mathbf{l})\|_2^2.$$

The normal equation of are then (3.6).

We will discuss efficient solution algorithms for these equations in Section 3.3, but first we mention some special properties of the DPG setting for this class of problems and also discuss the imposition of boundary conditions and linear constraints in general terms.

*DLS for traditional methods.* Beginning with any standard finite element method, it is tempting to consider the corresponding stiffness matrix  $B \in \mathbb{R}^{M \times N}$  and load vector  $\mathbf{l} \in \mathbb{R}^N$ , with  $M = N$ . Defining the Gram matrix to be the identity matrix,  $G = I$ —or any invertible matrix, really—the full rank linear system of equations

$$\mathbf{B}\mathbf{u} = \mathbf{l},$$

is equivalent to each of the formulations above.

In some methods, it is also quite straightforward to extend the associated test space. Therefore, let us assume again that  $M > N$ . In this situation, the naive choice  $G = I$  naturally induces the linear system (3.6),

but the condition number of  $A = B^*B$  would then be the square of the condition number of  $B$ . Since the condition number of  $B$  is often  $\mathcal{O}(h^{-2})$ , this would usually be a gross impediment to practicality. Fortunately, with an appropriate choice of test norm or Gram matrix, the condition number can be far better behaved. Indeed, in Section 5 we demonstrate condition number growth of  $A$  to be  $\mathcal{O}(h^{-2}) \neq \mathcal{O}(h^{-4})$ —and, thus,  $\text{cond}(\tilde{B}) = \mathcal{O}(h^{-1})$ —in several practical settings.

*DPG methods.* As mentioned previously, if the test space is broken and localizable (see Section 2.3), the Gram matrix  $G$  can be constructed to be block diagonal. Indeed, let  $\mathcal{V}^{\text{DPG}} = \bigoplus_{K \in \mathcal{T}} \mathcal{V}_K$  be the DPG test space and let  $\mathfrak{V}_r^{\text{DPG}} = \bigoplus_{K \in \mathcal{T}} \mathfrak{V}_K$ ,  $\mathfrak{V}_K \subset \mathcal{V}_K$ , be a broken and localized basis for the finite-dimensional subspace  $\mathcal{V}_r^{\text{DPG}} \subset \mathcal{V}^{\text{DPG}}$ . That is, for each  $i \in \{1, \dots, M\}$ ,  $\mathbf{v}_i \in \mathcal{V}_K$  for some  $K \in \mathcal{T}$  where  $\mathbf{v}_i|_{K'} = 0$  for all  $K' \neq K$ .

Fixing the previous  $i$  and  $K$ , observe that, for each  $j \in \{1, \dots, M\}$ ,

$$G_{ij} = (\mathbf{v}_i, \mathbf{v}_j)_{\mathcal{V}^{\text{DPG}}} = (\mathbf{v}_i, \mathbf{v}_j)_{\mathcal{V}_K},$$

by (2.13). Therefore,  $G = \text{diag}(G_K)$ , where each  $G_K$  is the local Gram matrix,

$$(G_K)_{ij} = (\mathbf{v}_i, \mathbf{v}_j)_{\mathcal{V}_K},$$

for each  $\mathbf{v}_i, \mathbf{v}_j \in \mathfrak{V}_K$ .

As also alluded to previously, in some circumstances, many, or all,  $G_K$  can be made be *diagonal* if the basis is properly compatible with the test norm. For regular enough elements, this is often possible when  $\mathcal{V} = L^2$ —indeed, this was verified in the FOSLS experiments in Section 5.1. The interested reader may consult [7, 36] and references therein, to explore such orthogonal or sparsity-optimized shape functions.

Recall from (2.12),  $\mathcal{B}^{\text{DPG}}(\mathbf{u}^{\text{fld}}, \hat{\mathbf{u}}) = \mathcal{B}^{\text{fld}} \mathbf{u}^{\text{fld}} + \hat{\mathcal{B}} \hat{\mathbf{u}}$ . Through an abuse of notation, denote the coefficients of  $\mathbf{u}^{\text{fld}}$  as  $\mathbf{u}$  and the coefficients of  $\hat{\mathbf{u}}$  as  $\hat{\mathbf{u}}$ . Similarly define  $B$  and  $\hat{B}$ . With these definitions, the DPG analog of the saddle-point problem (3.5) is

$$(3.7) \quad \begin{bmatrix} G & B & \hat{B} \\ B^* & 0 & 0 \\ \hat{B}^* & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{u} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \\ 0 \end{bmatrix}.$$

**3.2. Boundary conditions and linear equality constraints.** Recall the matrices  $G, B, \tilde{B}$ , and  $A$ , and vectors  $\mathbf{l}, \tilde{\mathbf{l}}$ , and  $\mathbf{f}$  from Section 3.1. As demonstrated in Section 2.4, in principle, a constrained minimum residual problem will involve two test spaces,  $\mathcal{V}$  and  $\mathcal{W}$ . We have already fixed a truncated basis for  $\mathcal{V}$  spanning  $\mathcal{V}_r \subset \mathcal{V}$ . Here, we additionally require a similar basis for  $\mathcal{W}_s \subset \mathcal{W}$ , denoted  $\mathfrak{W}_s = \{\mathbf{w}_i\}_{i=1}^L$ . Drawing from  $\mathfrak{W}_s$ , we now define  $C_{ij} = c(\mathbf{w}_j, \mathbf{w}_i)$ ,  $H_{ij} = (\mathbf{w}_i, \mathbf{w}_j)_{\mathcal{W}}$ , and  $\mathbf{d}_i = d(\mathbf{w}_i)$ .

Only in discretizing the penalized problem, (2.21), will we require the second Gram matrix,  $H$ . In the saddle-point problem, (2.19), however, it will be required to solve for the constraint coefficients,  $\mathbf{w} = [\mathbf{w}_i]_{i=1}^L \in \mathbb{R}^L$ .

Proceeding as before, the saddle-point problem can be expressed

$$(3.8) \quad \mathbf{u}^{\text{opt}} = \arg \min_{\mathbf{u} \in \mathbb{R}^N} \sup_{\mathbf{w} \in \mathbb{R}^L} (\|\mathbf{B}\mathbf{u} - \mathbf{l}\|_{G^{-1}} + \mathbf{w}^*(\mathbf{C}\mathbf{u} - \mathbf{d})).$$

Or, in block matrix form, as

$$(3.9) \quad \begin{bmatrix} G & B & 0 \\ B^* & 0 & C^* \\ 0 & C & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}^{\text{opt}} \\ \mathbf{u}^{\text{opt}} \\ \mathbf{w}^{\text{opt}} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \\ \mathbf{d} \end{bmatrix} \iff \begin{bmatrix} A & C^* \\ C & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{\text{opt}} \\ \mathbf{w}^{\text{opt}} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{d} \end{bmatrix}.$$

Notably, this problem is well-posed if and only if the linear system  $\mathbf{C}\mathbf{u} = \mathbf{d}$  is consistent,  $C$  has full row rank, and  $\text{Null}(A) \cap \text{Null}(C) = \{0\}$  [8, Chapter 5]. Moreover, the right-hand equation in (3.9) is fundamentally different from the DLS saddle-point problem (3.5), even when  $\mathbf{d} = 0$ . This is not only because we cannot expect  $A$  to have any easily exploitable structure, but because we are predominately interested in the first subvector,  $\mathbf{u}^{\text{opt}}$ , not  $\mathbf{w}^{\text{opt}}$ .

Alternatively, the penalized problem is

$$(3.10) \quad \mathbf{u}^{\text{opt}} = \arg \min_{\mathbf{u} \in \mathbb{R}^N} (\|\mathbf{B}\mathbf{u} - \mathbf{l}\|_{G^{-1}}^2 + \|\mathbf{C}\mathbf{u} - \mathbf{d}\|_{H^{-1}}^2),$$

resulting in the saddle point problem

$$(3.11) \quad \begin{bmatrix} G & B & 0 \\ B^* & 0 & C^* \\ 0 & C & H \end{bmatrix} \begin{bmatrix} \mathbf{p}^{\text{opt}} \\ \mathbf{u}^{\text{opt}} \\ \mathbf{w}^{\text{opt}} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \\ \mathbf{d} \end{bmatrix},$$

or, equivalently,

$$(3.12) \quad (A + E)\mathbf{u}^{\text{opt}} = \mathbf{f} + \mathbf{e},$$

where  $E = C^*H^{-1}C$  and  $\mathbf{e} = C^*H^{-1}\mathbf{d}$ . This is well-posed if and only if  $\text{Null}(A) \cap \text{Null}(C) = \{0\}$ .

Most often, the Gram matrix  $H$  has a tunable penalization parameter:  $H = \alpha^{-2}H_0$ . The magnitude of  $\alpha$  determines the bias on the constraint in (3.10). Intuitively, the solution of (3.11) converges to the solution of (3.9) as  $\alpha \rightarrow \infty$ , and indeed, this convergence can be made precise [5]. Unfortunately, as  $\alpha$  grows, so will the condition number of  $E$ , and therefore, so will the round-off error in (3.12).

Often,  $H_0 = I$ . In this scenario, (3.11) can also be expressed as the least-squares problem for an overdetermined system:

$$(3.13) \quad \min_{\mathbf{u} \in \mathbb{R}^N} \left\| \begin{bmatrix} \alpha C \\ \tilde{B} \end{bmatrix} \mathbf{u} - \begin{bmatrix} \alpha \mathbf{d} \\ \tilde{\mathbf{l}} \end{bmatrix} \right\|_2,$$

which is less sensitive to the penalization parameter than the normal equation when solved with QR. In the least-squares context, this is referred to as the “method of weighting” [71]. And, the extension to scenarios when the Gram matrix can be factored,  $H = MM^*$ , should be clear.

In the special, but common, case of equality constraints in the solution coefficients, we can decompose the vector  $\mathbf{u} = [\mathbf{u}_{\text{hom.}} \mid \mathbf{u}_{\text{lift}}]^T$  into the *free* homogeneous coefficients,  $\mathbf{u}_{\text{hom.}}$ , and the *fixed* lifted coefficients,  $\mathbf{u}_{\text{lift}}$ . We may then decompose the identity matrix  $I = [Q_{\text{hom.}} \mid Q_{\text{lift}}]$  into two submatrices with (clearly) orthonormal columns,  $Q_{\text{hom.}}^* \mathbf{u} = \mathbf{u}_{\text{hom.}}$  and  $Q_{\text{lift}}^* \mathbf{u} = \mathbf{u}_{\text{lift}}$ .

Now, observe that

$$B\mathbf{u} = BQ_{\text{hom.}}Q_{\text{hom.}}^*\mathbf{u} + BQ_{\text{lift}}Q_{\text{lift}}^*\mathbf{u} = B_{\text{hom.}}\mathbf{u}_{\text{hom.}} + B_{\text{lift}}\mathbf{u}_{\text{lift}},$$

where  $B_{\text{hom.}} = BQ_{\text{hom.}}$  and  $B_{\text{lift}} = BQ_{\text{lift}}$ . Then, the optimal solution coefficients can then be characterized as

$$\mathbf{u}^{\text{opt}} = \arg \min_{\mathbf{u}_{\text{hom.}} \in \mathbb{R}^{N_{\text{hom.}}}} \|B_{\text{hom.}}\mathbf{u}_{\text{hom.}} - \mathbf{l}_{\text{lift}}\|_{G^{-1}} + \mathbf{u}_{\text{lift}} = \arg \min_{\mathbf{u}_{\text{hom.}} \in \mathbb{R}^{N_{\text{hom.}}}} \|\tilde{B}_{\text{hom.}}\mathbf{u}_{\text{hom.}} - \tilde{\mathbf{l}}_{\text{lift}}\|_2 + \mathbf{u}_{\text{lift}},$$

where  $\mathbf{l}_{\text{lift}} = \mathbf{l} - B_{\text{lift}}\mathbf{u}_{\text{lift}}$ ,  $\tilde{B}_{\text{hom.}} = \tilde{B}Q_{\text{hom.}}$ , and  $\tilde{\mathbf{l}}_{\text{lift}} = \tilde{\mathbf{l}} - \tilde{B}Q_{\text{lift}}\mathbf{u}_{\text{lift}}$ .

In the construction above, introducing the matrices  $Q_{\text{hom.}}$  and  $Q_{\text{lift}}$  was seemingly unnecessary because  $B = [B_{\text{hom.}} \mid B_{\text{lift}}]$ . However, the solution coefficients will generally not be ordered as above and this detail allows us to connect to the more general constraint case, by identifying  $C$  with  $Q_{\text{lift}}^*$ :

$$\begin{bmatrix} G & B & 0 \\ B^* & 0 & Q_{\text{lift}} \\ 0 & Q_{\text{lift}}^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}^{\text{opt}} \\ \mathbf{u}^{\text{opt}} \\ \mathbf{w}^{\text{opt}} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \\ \mathbf{u}_{\text{lift}} \end{bmatrix} \iff \begin{bmatrix} A & Q_{\text{lift}} \\ Q_{\text{lift}}^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{\text{opt}} \\ \mathbf{w}^{\text{opt}} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{u}_{\text{lift}} \end{bmatrix}.$$

After multiplying the first row of the second equation by  $Q_{\text{hom.}}^*$ , it is easy to see that these equations imply

$$A_{\text{hom.}}\mathbf{u}_{\text{hom.}}^{\text{opt}} = B_{\text{hom.}}^*G^{-1}\mathbf{l}_{\text{lift}},$$

where  $A_{\text{hom.}} = Q_{\text{hom.}}^*AQ_{\text{hom.}} = B_{\text{hom.}}^*G^{-1}B_{\text{hom.}}$ .

*Post-processing precomputed solutions.* In some situations, large subvectors of the coefficients of the solution,  $\mathbf{u}_{\text{sol.}}$ , corresponding to entire components,  $u_i$ ,  $i \in \mathcal{I}_{\text{sol.}} \subsetneq \{1, \dots, n\}$ , of the discrete solution may already be available at the outset of computation. For instance, this is possible if an auxiliary finite element method was used to compute  $u_i$ . In this scenario, the remaining solution coefficients can be computed using the minimum residual framework with equality constraints as described above and defining the associated  $Q_{\text{sol.}} = C^*$ ,  $C\mathbf{u} = \mathbf{u}_{\text{sol.}}$ .

**3.3. Solution algorithms.** There are several prevalent strategies to solve generalized linear least-squares problems

$$(DLS) \quad \min_{\mathbf{u} \in \mathbb{F}^N} \|\tilde{\mathbf{B}}\mathbf{u} - \tilde{\mathbf{l}}\|_{G^{-1}}.$$

Under infinite precision arithmetic, they are each essentially equivalent, however, algorithmically speaking, each approach is significantly different. We briefly survey the most important direct methods in this subsection.

*The normal equation.* In order to solve for  $\mathbf{u}^{\text{opt}}$ , we could begin by forming the normal equation:

$$(NE) \quad \mathbf{A}\mathbf{u}^{\text{opt}} = \mathbf{f}, \quad \text{where} \quad \mathbf{A} = \mathbf{B}^* \mathbf{G}^{-1} \mathbf{B} \quad \text{and} \quad \mathbf{f} = \mathbf{B}^* \mathbf{G}^{-1} \mathbf{l}.$$

Beginning in this way is standard practice in the DPG community and is advantageous in that the stiffness matrix and load vector,  $\mathbf{A}$  and  $\mathbf{f}$ , can be constructed locally and assembled in sparse format using standard finite element assembly algorithms. Moreover,  $\mathbf{A} = \tilde{\mathbf{B}}^* \tilde{\mathbf{B}}$  is Hermitian positive-definite and therefore has a structure amenable to efficient linear solvers not usually available for many challenging problems. Unfortunately, the condition number of  $\mathbf{A}$  will grow quadratically with the condition number of  $\tilde{\mathbf{B}}$  and, likewise, so will the upper bound on the round-off error of the normal equation solution. However, this is not to say that when the normal equation are formed the growth of the condition number will be expected to be faster than most standard finite element methods. For instance, when using a first-order system (FOS) formulation, some methods using the normal equation can be proven to have condition number growth  $\mathcal{O}(h^{-2})$ , where  $h$  is the element size in a quasi-uniformly refined mesh [45]. Similarly, this has also been shown to be true in a more general context for the linear systems arising from FOS least-squares (FOSLS) methods [10, 11] which the normal equation above are related to (see Section 5.1).

Nevertheless, the scaling constant controlling the condition number of the stiffness matrix is often large in a FOS setting due to the additional equations and unknowns. This is an impediment to producing accurate solutions to very difficult problems and can induce large and sometimes overwhelming numerical errors. This is one reason that it is sometimes convenient to consider other approaches which deal explicitly with the matrices  $\mathbf{B}$ ,  $\mathbf{W}$ , and  $\mathbf{l}$ , and avoid the normal equation altogether.

*Orthogonal decomposition methods.* The most practical alternative to the normal equation when solving for  $\mathbf{u}^{\text{opt}}$  is to deal directly with the matrices  $\tilde{\mathbf{B}}$  and  $\tilde{\mathbf{l}}$  coming from the (sparsely weighted) linear least-squares problem

$$(LS) \quad \min_{\mathbf{u} \in \mathbb{F}^N} \|\tilde{\mathbf{B}}\mathbf{u} - \tilde{\mathbf{l}}\|_2, \quad \text{where} \quad \tilde{\mathbf{B}} = \mathbf{W}^{-1} \mathbf{B}, \quad \tilde{\mathbf{l}} = \mathbf{W}^{-1} \mathbf{l}, \quad \text{and} \quad \mathbf{W}\mathbf{W}^* = \mathbf{G}.$$

The most common of these approaches is the orthogonalization algorithm called QR-factorization (Householder, Givens, MGS) first introduced for least-squares problems in [41]. Other direct approaches are SVD, complete orthogonal decomposition, and Peter-Wilkinson as well as various hybrid methods [8]. Each of these approaches are usually less efficient than solving via the normal equation, but are often preferred because they are more numerically stable.

For the purposes we have in mind, the matrix  $\tilde{\mathbf{B}}$  will be large and sparse and so, because not all of the methods above are well suited for sparse matrices or amenable to parallel computing, we will focus only on the QR approach. As shown in various textbooks [8, 44, 70], the relative error in the solution from a least-squares QR solve is controlled by  $\text{cond}(\tilde{\mathbf{B}}) + \rho(\tilde{\mathbf{B}}, \tilde{\mathbf{l}}) \cdot \text{cond}(\tilde{\mathbf{B}})^2$ , where

$$(3.14) \quad \rho(\tilde{\mathbf{B}}, \tilde{\mathbf{l}}) = \frac{\|\tilde{\mathbf{B}}\mathbf{u}^{\text{opt}} - \tilde{\mathbf{l}}\|_2}{\|\tilde{\mathbf{B}}\|_2 \|\mathbf{u}^{\text{opt}}\|_2},$$

and therefore depends upon the load vector.

Due to consistency of the problem,  $\ell \in \text{Range}(\mathbf{B})$ . Therefore, the residual (the numerator in (3.14)) is expected to tend to zero as the mesh is refined. That is,  $\rho(\tilde{\mathbf{B}}, \tilde{\mathbf{l}}) \rightarrow 0$  as  $h \rightarrow 0$ . Indeed, the reader may observe that the  $\rho(\tilde{\mathbf{B}}, \tilde{\mathbf{l}})$  even vanishes if  $\tilde{\mathbf{l}} \in \text{Range}(\tilde{\mathbf{B}})$ . Of course, the validity of this convergence as well as its rate will be determined by the interpolation spaces used in the discretization. However, in many common scenarios the *a priori* bound can be proven to decrease at a rate of at least  $\mathcal{O}(h)$ . Indeed, for many cases we have in mind, this

rate is of the form  $\mathcal{O}(h^p)$  where  $p \geq 1$  is a parameter indicating the polynomial order used in the discretization. Therefore, intuition is thus: the quadratic condition number term controlling the round-off error in a QR solve will be offset by a converging solution.

For instance, recall that  $\text{cond}(\tilde{B}) = \text{cond}(A)^{1/2}$ . Therefore, for many common FOS formulations, the condition number growth in  $\tilde{B}$  is only  $\mathcal{O}(h^{-1})$ . Moreover, if the residual converges to zero as described above, then  $\rho(\tilde{B}, \tilde{l}) \cdot \text{cond}(\tilde{B})^2$  can be no worse than  $\mathcal{O}(h^{-1})$ . In such conventional scenarios, the numerical sensitivity of the least-squares solution is controlled only by the inverse of the mesh size and will be far more accurate than any normal equation approach!

Precisely, in the typical FOS scenario, we expect a QR-based algorithm will deliver an error bound of

$$(3.15) \quad \|\mathbf{e}\|_2 \leq \epsilon_{mach.} \|\mathbf{u}\|_2 C h^{-1},$$

where  $\mathbf{e}$  is the round-off error in the computation of the least-squares solution,  $\epsilon_{mach.}$  is machine precision, and  $C$  is a mesh-independent constant.

*An ill-conditioned Gram matrix.* Unfortunately, although the QR is guaranteed to deliver a more accurate solution than solving with the normal equation, there is potentially still a concealed obstacle. As many authors have pointed out, explicitly forming a product of two matrices before solving a least-squares problem posed with the matrix product  $\tilde{B} = W^{-1}B$  is still not backwards stable [8]. Indeed, even when  $B$  is sparse and the matrix  $W$  is diagonal—but extremely ill-conditioned—this can be a potential issue [9, 47]. Because of this concern, several algorithms exist in the numerical linear algebra literature for this very class of problems [60, 2, 46, 72]. Nevertheless, we believe that such precautions are unwarranted in all but the most exceptional problems that can be expected with a DLS method.

Recall the critical assumptions in Section 1 on the structure of the Gram matrix. We implicitly assume that the conditioning of  $W$  should not be badly behaved as the problem size grows. Indeed, in the cases where the Gram matrix comes from a DPG method (i.e. is block-diagonal) or from some other technique (perhaps a preconditioner estimate),  $G^{-1}$  can usually be generated using local element or patch information. This motivates us to assume that some measure of its local rank structure should stay constant or be uniformly bounded (with respect to the mesh size,  $h$ ) as the mesh is refined. If this is true, preconditioning the Gram matrix—which itself often acts like a preconditioner—with its diagonal entries,

$$(3.16) \quad G \mapsto D^{-1/2} G D^{-1/2}, \quad B \mapsto D^{1/2} B, \quad l \mapsto D^{1/2} l,$$

where  $D = \text{diag}(G)$ , before locally factoring into  $WW^*$  and performing back-substitution, should lead to robust results.

This diagonal preconditioning procedure has been more than adequate in all of our experiments thus far. However, another possibility for improving the condition of the Gram matrix is the modified Lagrangian approach suggested in [42]. Here,

$$(3.17) \quad G \mapsto G + BSB^*,$$

where  $S$  is a user-defined Hermitian matrix. Because the equivalent equilibrium systems

$$\begin{bmatrix} G & B \\ B^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \end{bmatrix} \iff \begin{bmatrix} G + BSB^* & B \\ B^* & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{l} \\ 0 \end{bmatrix}$$

have the same solution, a well-chosen matrix  $S$  may significantly improve the condition number of the  $(1, 1)$  block. It may also be useful to handle the case of a singular  $G$ . We have experimented with this technique in our experiments, however, we have not made any certain conclusions for an effective  $S$ . For the reader willing to apply this technique in the DPG setting, we note that  $BSB^*$  should probably hold the same block diagonal structure as  $G$ .

If these procedures fail, a third alternative approach is obviously to avoid condensing the system. This is similar to some of the saddle-point finite element methods mentioned in Section 1. For completeness, we now summarize the highlights of this strategy.

*Generalized least-squares.* For the most badly behaved problems, [8, 44] suggest beginning with the generalized least-squares problem

$$(GLS) \quad \min_{\mathbf{u} \in \mathbb{R}^N, \mathbf{r} \in \mathbb{R}^M} \|\mathbf{r}\|_2 \quad \text{subject to} \quad \mathbf{B}\mathbf{u} + \mathbf{W}\mathbf{r} = \mathbf{l}, \quad \text{where} \quad \mathbf{W}\mathbf{W}^* = \mathbf{G}.$$

For a direct method, the solution coefficients  $\mathbf{u}$  are then suggested to be computed using a QR factorization approach which was first described in [58]. Although there are a couple of strict advantages of this idea—including that the Gram matrix  $\mathbf{G}$  need no longer be invertible—the size of the resulting saddle-point system is much larger than the systems in the previously methods. Not to mention, the QR-based solution algorithm given in [58] is seemingly impractical for any reasonably large sparse system because it involves storing and applying large and probably dense orthogonal matrices. In [59], an efficient algorithm is proposed for the case that  $\mathbf{G}$  is block diagonal, however, we are unaware of any multi-frontal implementations so have not explored it in our experiments.

Notably, this saddle-point approach is analogous to the finite element methods described in [20] and [25]. Moreover, although this generalized least-squares starting point may be too expensive to be practical for direct linear solvers, it may have benefits for iterative solution algorithms [74, 5, 43]. One such very promising technique in the DPG context, inspired by PDE-constrained optimization, is developed in [14] in a similar setting where  $\mathbf{G}$  is not factored.

*Further discussion.* In choosing the best algorithm to solve the least-squares problem coming from a DLS method, many factors are important to consider. For instance, the normal equation have been demonstrated to be adequate when the methodology has been applied to many DPG problems [64, 65, 21, 67, 33, 37, 49, 62, 30, 40, 38, 35]. Indeed, in many reasonable circumstances, the round-off error in the solution from the associated linear solve cannot be expected to be nearly as large as the truncation error due to the finite element discretization. Direct solvers for Hermitian positive-definite systems are also usually several times faster than their associated QR-based counterparts for DLS problems.<sup>1</sup> Moreover, the normal equation, itself, generally requires less storage than the modified coefficient matrix and load vector.

Nevertheless, there are many anticipated circumstances where other least-squares methods would be especially useful. Archetypal examples include, but are not limited to, singular perturbation problems, problems with large material contrast, high-order PDEs, penalty methods, and nonlinear problems where the linear approximation may become singular or very ill-conditioned.

It is also important to mention that practical direct methods for constrained DLS methods have additional complications. Indeed, a naive direct QR-based algorithm for a problem

$$(CDLS) \quad \min_{\mathbf{u} \in \mathbb{R}^N} \|\tilde{\mathbf{B}}\|_{\mathbf{G}^{-1}}, \quad \text{subject to} \quad \mathbf{C}\mathbf{u} = \mathbf{d}.$$

where both  $\tilde{\mathbf{B}}$  and  $\mathbf{C}$  are sparse will probably not be practical. This is because such methods involve transforming the minimization problem above to an equivalent problem posed over the null space of  $\mathbf{C}$ . This sequence of operations will involve representing  $\tilde{\mathbf{B}}$  in a basis where it may no longer have a similar sparsity. The same issues appear for problems with inequality constraints [8, Section 6.8]. Therefore, for direct methods, (CDLS) requires either solving the condensed system

$$\begin{bmatrix} \mathbf{A} & \mathbf{C}^* \\ \mathbf{C} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{\text{opt}} \\ \mathbf{w}^{\text{opt}} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{d} \end{bmatrix},$$

or compromising on a penalized constraint and performing QR on (3.13).

<sup>1</sup>We make this statement based upon personal experience on a single-core machine with the available software at this time. Because most complexity arguments in the literature are usually pessimistic and proved for dense matrices (they do not incorporate the expected sparse matrix structure) and because the time-to-solution may depend heavily upon implementation or architecture, we believe this is more valid than predictions which could be made with well-known complexity results.



**3.4. Static condensation.** A common procedure which is often used to improve the solving time of linear systems is called static condensation. Here, the degrees of freedom associated with the element interior nodes (bubbles) are eliminated from the linear system. In practice, using a Schur complement procedure, small and independent blocks of the original stiffness matrix and load vectors are removed, and the original system is changed into a smaller-but-modified linear system with fewer unknowns. Often, this procedure of condensing, solving, and then recovering the global solution is much faster than solving the original system outright with standard means.

A similar procedure can be developed for overdetermined discrete least-squares problems without forming the normal equation—although it can be identified with static condensation therein (see Appendix B). To illustrate this application in practical DLS problems like (LS), consider (3.4) where we have separated  $\mathbf{u} = [\mathbf{u}_{\text{bubb.}} \mid \mathbf{u}_{\text{interf.}}]^T$  into bubble and interface components, respectively. Here, the bubble coefficients are assigned to functions which have support entirely within a single element, while the interface coefficients are assigned to basis functions with support across multiple elements or those which may only be defined at element interfaces. With this distinction, we may rewrite the minimization problem of (3.4) as a sequence of two independent minimizations:

$$\min_{\mathbf{u} \in \mathbb{R}^N} \mathcal{R}(\mathbf{u}) = \min_{\mathbf{u}_{\text{interf.}} \in \mathbb{R}^{N_{\text{interf.}}}} \min_{\mathbf{u}_{\text{bubb.}} \in \mathbb{R}^{N_{\text{bubb.}}}} \mathcal{R}(\mathbf{u}) = \min_{\mathbf{u}_{\text{bubb.}} \in \mathbb{R}^{N_{\text{bubb.}}}} \min_{\mathbf{u}_{\text{interf.}} \in \mathbb{R}^{N_{\text{interf.}}}} \mathcal{R}(\mathbf{u}),$$

where we recall  $\mathcal{R}(\mathbf{u}) = \|\tilde{\mathbf{B}}\mathbf{u} - \tilde{\mathbf{l}}\|_2^2$ . Indeed, we may first solve for  $\mathbf{u}_{\text{interf.}}^{\text{opt}}$  using an implicit expression for the unknown  $\mathbf{u}_{\text{bubb.}}^{\text{opt}}$  in terms of the variable  $\mathbf{u}_{\text{interf.}}$ . Once  $\mathbf{u}_{\text{interf.}}^{\text{opt}}$  has been computed, the second minimization problem,

$$\mathbf{u}_{\text{bubb.}}^{\text{opt}} = \arg \min_{\mathbf{u}_{\text{bubb.}} \in \mathbb{R}^{N_{\text{bubb.}}}} \mathcal{R}([\mathbf{u}_{\text{bubb.}} \mid \mathbf{u}_{\text{interf.}}^{\text{opt}}]),$$

can be solved locally.

Representing the distinction in coefficients,  $\mathbf{u} = [\mathbf{u}_{\text{bubb.}} \mid \mathbf{u}_{\text{interf.}}]^T$ , with the operator  $\tilde{\mathbf{B}}$  as well, we define two sub-matrices  $\tilde{\mathbf{B}}_{\text{bubb.}}$  and  $\tilde{\mathbf{B}}_{\text{interf.}}$  where  $\tilde{\mathbf{B}} = [\tilde{\mathbf{B}}_{\text{bubb.}} \mid \tilde{\mathbf{B}}_{\text{interf.}}]$ . Naturally, the dimensions of  $\tilde{\mathbf{B}}_{\text{bubb.}}$  are  $M \times N_{\text{bubb.}}$  and the dimensions of  $\tilde{\mathbf{B}}_{\text{interf.}}$  are  $M \times N_{\text{interf.}}$ . Recall that we are solving

$$(3.18) \quad \tilde{\mathbf{B}}_{\text{bubb.}} \mathbf{u}_{\text{bubb.}} + \tilde{\mathbf{B}}_{\text{interf.}} \mathbf{u}_{\text{interf.}} \stackrel{\text{LS}}{=} \tilde{\mathbf{l}},$$

where  $\stackrel{\text{LS}}{=}$  indicates equality in the least-squares sense. If  $\mathbf{u}_{\text{interf.}}$  were fixed, then the bubble contribution of the least-squares solution would be represented as

$$(3.19) \quad \mathbf{u}_{\text{bubb.}}^{\text{opt}} = \tilde{\mathbf{B}}_{\text{bubb.}}^+ \left( \tilde{\mathbf{l}} - \tilde{\mathbf{B}}_{\text{interf.}} \mathbf{u}_{\text{interf.}} \right),$$

where  $\tilde{\mathbf{B}}_{\text{bubb.}}^+ = \left( \tilde{\mathbf{B}}_{\text{bubb.}}^* \tilde{\mathbf{B}}_{\text{bubb.}} \right)^{-1} \tilde{\mathbf{B}}_{\text{bubb.}}^*$  is the pseudoinverse of  $\tilde{\mathbf{B}}_{\text{bubb.}}$ . Exploiting this expression for  $\mathbf{u}_{\text{bubb.}}^{\text{opt}}$ , we can rewrite (3.18) as a least-squares problem entirely in the variable  $\mathbf{u}_{\text{interf.}}$  with  $\mathbf{u}_{\text{bubb.}} = \mathbf{u}_{\text{bubb.}}^{\text{opt}}$  fixed, viz.,

$$(3.20) \quad (\mathbf{I} - \mathbf{P}_{\text{bubb.}}) \tilde{\mathbf{B}}_{\text{interf.}} \mathbf{u}_{\text{interf.}} \stackrel{\text{LS}}{=} (\mathbf{I} - \mathbf{P}_{\text{bubb.}}) \tilde{\mathbf{l}},$$

where  $\mathbf{I}$  is the identity matrix and  $\mathbf{P}_{\text{bubb.}} = \tilde{\mathbf{B}}_{\text{bubb.}} \left( \tilde{\mathbf{B}}_{\text{bubb.}}^* \tilde{\mathbf{B}}_{\text{bubb.}} \right)^{-1} \tilde{\mathbf{B}}_{\text{bubb.}}^*$ . The reader may observe that  $\mathbf{P}_{\text{bubb.}}^2 = \mathbf{P}_{\text{bubb.}}$  and likewise for  $\mathbf{P}_{\text{interf.}} = \mathbf{I} - \mathbf{P}_{\text{bubb.}}$ . These two matrices can be identified with orthogonal projectors onto the space of bubble coefficients and interface coefficients, respectively.

Applying  $\mathbf{P}_{\text{interf.}}$  to  $\tilde{\mathbf{B}}_{\text{interf.}}$  can be performed in parallel because  $\tilde{\mathbf{B}}_{\text{bubb.}}$  will always be block diagonal. Then, without actually computing  $\mathbf{u}_{\text{bubb.}}^{\text{opt}}$ , expression (3.20) can be used to solve for the optimal interface coefficients,

$$\mathbf{u}_{\text{interf.}}^{\text{opt}} = \arg \min_{\mathbf{u}_{\text{interf.}} \in \mathbb{R}^{N_{\text{interf.}}}} \mathcal{R}([\mathbf{u}_{\text{bubb.}}^{\text{opt}} \mid \mathbf{u}_{\text{interf.}}]).$$

Afterwards, once  $\mathbf{u}_{\text{interf.}}^{\text{opt}}$  are available, (3.19) can be used to extract  $\mathbf{u}_{\text{bubb.}}^{\text{opt}}$ , locally.

Before closing, we remark that with a QR-decomposition of  $\tilde{\mathbf{B}}_{\text{bubb.}}$ ,

$$\tilde{\mathbf{B}}_{\text{bubb.}} = [\mathbf{Q}_{\text{bubb.}} \mid \mathbf{Q}_{\text{interf.}}] \begin{bmatrix} \mathbf{R}_{\text{bubb.}} \\ 0 \end{bmatrix}$$

where  $[Q_{\text{bubb.}} \mid Q_{\text{interf.}}]$  is unitary and  $R_{\text{bubb.}}$  is upper-triangular, (3.19) can be rewritten as

$$\mathbf{u}_{\text{bubb.}}^{\text{opt}} = R_{\text{bubb.}}^{-1} Q_{\text{bubb.}}^* \left( \tilde{\mathbf{l}} - \tilde{B}_{\text{interf.}} \mathbf{u}_{\text{interf.}} \right).$$

Meanwhile,  $P_{\text{bubb.}} = Q_{\text{bubb.}} Q_{\text{bubb.}}^*$  and  $P_{\text{interf.}} = I - Q_{\text{bubb.}} Q_{\text{bubb.}}^* = Q_{\text{interf.}} Q_{\text{interf.}}^*$ .

Exploiting the QR-decomposition will minimize the round-off error introduced in modifying the original system as well as in recovering  $\mathbf{u}_{\text{bubb.}}^{\text{opt}}$  and we suggest static condensation routines for DLS problems of the form (LS) be implemented using it.

#### 4. ASSEMBLY

In this section, we describe the construction of the global linear systems for DPG methods. As it was outlined in Section 3.3, we are primarily interested in two different procedures: (1) directly assembling the normal equation; and (2) assembling the overdetermined system. To date, forming the normal equation has been the primary assembly procedure for DPG linear systems. The main advantages of this approach is that the assembly algorithm is identical to all traditional conforming finite element methods and it will involve the least storage. Moreover, many efficient direct and iterative solvers specialized for Hermitian/symmetric positive definite systems can be employed. Nonetheless, there is an important disadvantage: the condition number of this global stiffness matrix  $A$  is the square of the condition number of the alternative, which is the (global) enriched stiffness matrix  $\tilde{B}$ . The enriched stiffness matrix is, however, rectangular, and so other solvers—generally more expensive solvers—have to be used to solve the overdetermined linear system it is involved with. Be that as it may, this second approach can be applied to ill-conditioned problems, where forming the normal equation becomes an unsatisfactory option.

We proceed by giving a brief description of the two different assembly procedures.

**4.1. The normal equation.** The assembly of the normal equation for the DPG method can easily be incorporated into any finite element code supporting exact sequence conforming shape functions [3, 36]. Recall that the DPG Gram matrix  $G$  is block diagonal. Therefore, it can be inverted element-wise and, therefore,  $G^{-1}$  is also block diagonal. Let  $B_K$  denote the enriched stiffness matrix for element  $K$  and  $\mathbf{l}_K$  the corresponding load vector. Additionally, let  $G_K$  be the element Gram matrix. Then, the DPG element stiffness matrix  $A_K$  and load vector  $\mathbf{f}_K$  are given by  $A_K = B_K^* G_K^{-1} B_K$  and  $\mathbf{f}_K = B_K^* G_K^{-1} \mathbf{l}_K$ , respectively. Using the Cholesky factorization of  $G_K = L_K L_K^*$ , we have:

$$A_K = B_K^* G_K^{-1} B_K = (L_K^{-1} B_K)^* (L_K^{-1} B_K) \quad \text{and} \quad \mathbf{f}_K = B_K^* G_K^{-1} \mathbf{l}_K = (L_K^{-1} B_K)^* (L_K^{-1} \mathbf{l}_K).$$

We note that one may wish to precondition before the above operations, as in (3.16).

The computation of the element DPG stiffness matrix and load vector is given by Algorithm 1. The assembly

---

**Algorithm 1** Element stiffness matrix and load vector for the DPG normal equation

---

- |   |                                 |
|---|---------------------------------|
| 1: $L_K \leftarrow \text{Cholesky}(G_K)$  |                                 |
| 2: $\tilde{B}_K \leftarrow \text{Triangular solve}(L_K \tilde{B}_K = B_K)$                          |                                 |
| 3: $\tilde{\mathbf{l}}_K \leftarrow \text{Triangular solve}(L_K \tilde{\mathbf{l}}_K = \mathbf{l})$ |                                 |
| 4: $A_K \leftarrow \tilde{B}_K^* \tilde{B}_K$   | // DPG element stiffness matrix |
| 5: $\mathbf{f}_K \leftarrow \tilde{B}_K^* \tilde{\mathbf{l}}_K$                                     | // DPG element load vector      |
- 

of the global DPG stiffness matrix and load vector can be implemented by following the common algorithm of any standard finite element code [26, 31]. Note that there are two modifications that one should make to the element stiffness matrices before the global assembly: account for Dirichlet boundary conditions; and (optional) accommodate degrees of freedom associated to constrained nodes for adaptive mesh refinement (hanging nodes are possibly created after adaptive  $h$ -refinements). We refer the reader to [26, 31] for detailed discussion on both of these modifications.

After pre-processing the element stiffness matrices,  $A_K \mapsto A_K^{\text{mod}}$ , one should proceed with static condensation,  $A_K^{\text{mod}} \mapsto A_K^c$ , to reduce the complexity of the global system. For these square and symmetric matrices, this operation is described in Appendix B.

Additionally, as in standard FEM, the assembly is driven by the so-called “local-to-global connectivity maps”. These maps assign to the local degrees of freedom their corresponding global degrees of freedom. The construction of these maps is based on the “donor strategy” and is implemented as in [26].

A description of the assembly procedure is given in Algorithm 2.

---

**Algorithm 2** Assembly of DPG normal equation

---

```

1: Initialize global stiffness matrix and load vector A and f.
2: for  $K \leftarrow 1$  to  $N_K$  do                                // for each element in the mesh
3:   Compute  $A_K$  and  $f_K$                                     // element stiffness matrix and load vector
4:   Compute  $A_K^{\text{mod}}$  and  $f_K^{\text{mod}}$                             // modified element matrix and load vector
5:   Compute  $A_K^c$  and  $f_K^c$                                 // condensed element matrix and load vector
6:   Get  $\text{Con}_K$                                             // local-to-global connectivity map
7:   for  $k_1 \leftarrow 1$  to  $\text{ndof}_K$  do                        // for each element degree of freedom (DOF)
8:      $i \leftarrow \text{Con}_K(k_1)$                             // global index for local DOF
9:      $f(i) \leftarrow f(i) + f_K(k_1)$                     // accumulate for the global load vector
10:    for  $k_2 \leftarrow 1$  to  $\text{ndof}_K$  do                    // for each element DOF
11:       $j \leftarrow \text{Con}_K(k_2)$                         // global index for local DOF
12:       $A(i, j) \leftarrow A(i, j) + A_K(k_1, k_2)$         // accumulate for the global stiffness matrix

```

---

**4.2. The overdetermined system.** Constructing the global overdetermined system requires some modifications to the assembly algorithms above. First, in order to deliver *rectangular* element stiffness matrices  $\tilde{B}$  and load vectors  $\tilde{l}$ , one should only perform the first three steps of Algorithm 1. Note that the column size of the element stiffness matrix  $\tilde{B}$  corresponds to the number of trial degrees of freedom and the row size to the number of test degrees of freedom. Similarly, the size of the load vector  $\tilde{l}$  corresponds to the number of the test degrees of freedom.

As with square stiffness matrices, after the rectangular element matrices and load vectors have been computed, they need to be modified in order to accommodate Dirichlet boundary conditions and constrained nodes. The Dirichlet boundary conditions can be accounted for,  $\tilde{B} \mapsto \tilde{B}_{\text{hom}}$ , and  $\tilde{l} \mapsto \tilde{l}_{\text{lift}}$ , through the operation described in Section 3.2. This, however, is performed locally, like in the assembly algorithm for the normal equation, so that  $\tilde{B}_K \mapsto \tilde{B}_K^{\text{mod}}$  and  $\tilde{l}_K \mapsto \tilde{l}_K^{\text{mod}}$ .

For constrained nodes, the procedure is similar to the one for the normal equation, with the difference being that now the modifications are performed only on the trial space because the test space is broken. Note that there are no modifications needed for the load vector.

The final local step is static condensation,  $\tilde{B}_K^{\text{mod}} \mapsto \tilde{B}_K^c$  and  $\tilde{l}_K^{\text{mod}} \mapsto \tilde{l}_K^c$  (see Section 3.4).

The global assembly algorithm then proceeds in a similar manner as for the normal equation. However, there is one important difference because the test space is broken: the need for accumulation of the contributions from different elements in both the (global) enriched stiffness matrix  $\tilde{B}$  and the (global) enriched load vector  $\tilde{l}$  has been eliminated. Therefore, in the global stiffness matrix, every row is independent. Note that this allows for a fully parallel assembly algorithm.

This entire global assembly procedure is summarized in Algorithm 3.

Note that, in practice, both Algorithms 2 and 3 are modified to exploit sparsity of the stiffness matrices,  $A$  and  $\tilde{B}$ , depending on what solution package is used for the linear solve.

**4.3. Comparison.** For the one-dimensional DPG ultraweak formulation of Poisson equation (see Section 5), Figures 4.1 and 4.2 depict the global DPG stiffness matrix for the normal equation and the overdetermined system, respectively. The mesh used consisted of ten quadratic elements and the order of approximation for the

**Algorithm 3** Assembly of DPG overdetermined system

---

```

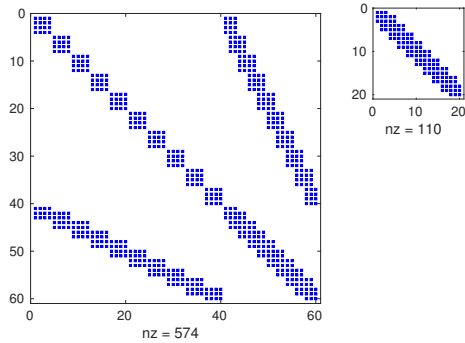
1: Initialize global stiffness matrix and load vector  $\tilde{\mathbf{B}}$  and  $\tilde{\mathbf{f}}$ .
2: Initialize global test DOF counter  $i$ .
3: for  $K \leftarrow 1$  to  $N_K$  do                                // for each element in the mesh
4:   Compute  $\tilde{\mathbf{B}}_K$  and  $\tilde{\mathbf{l}}_K$                                 // element stiffness matrix and load vector
5:   Compute  $\tilde{\mathbf{B}}_K^{\text{mod}}$  and  $\tilde{\mathbf{l}}_K^{\text{mod}}$                         // modified element matrix and load vector
6:   Compute  $\tilde{\mathbf{B}}_K^c$  and  $\tilde{\mathbf{l}}_K^c$                                 // condensed element matrix and load vector
7:   Get  $\text{Con}_K$                                                 // local-to-global connectivity map
8:   for  $k_1 \leftarrow 1$  to  $\text{ndof}_T K$  do                        // for each element test degree of freedom
9:      $i \leftarrow i + 1$                                         // global test DOF counter
10:     $\tilde{\mathbf{l}}(i) \leftarrow \tilde{\mathbf{l}}_K(k_1)$                             // fill in the global load vector
11:    for  $k_2 \leftarrow 1$  to  $\text{ndof}_E K$  do                        // for each element DOF
12:       $j \leftarrow \text{Con}_K(k_2)$                                 // global index for local DOF
13:       $\tilde{\mathbf{B}}(i, j) \leftarrow \tilde{\mathbf{B}}_K(k_1, k_2)$                 // fill in the global stiffness matrix

```

---

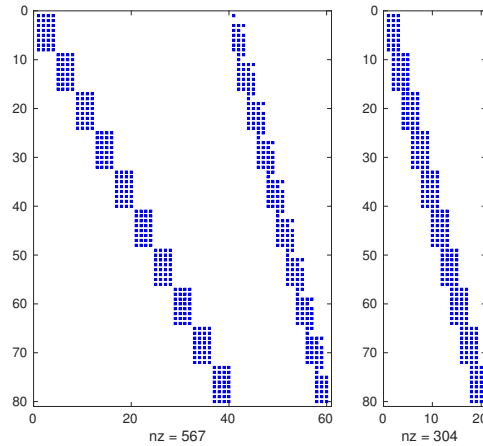
enriched test space was three. In a broken ultraweak formulation, continuity is enforced with the introduction of new interface unknowns (Lagrange multipliers), usually called fluxes and traces, that are traces of dual energy spaces on the mesh skeleton. This explains the structure of the matrix for the total system seen in Figure 4.1 (A). After static condensation of the interior degrees of freedom, the resulting linear system involved only the interface unknowns, and, therefore, the matrix in Figure 4.1 (B) consists of only one band of overlapping blocks.

The situation is slightly different in the case of the overdetermined system (see Figure 4.2 (A)). For instance, because the test space is broken, there is no overlap between rows. Similar to the case of the normal equation, static condensation led to a linear system involving only the interfaces unknowns. However, here, the size of test space remained the same and therefore only the column dimension was reduced (see Figure 4.2 (B)).



(A) Total system

(B) Condensed system

FIGURE 4.1. DPG stiffness matrix  $\mathbf{A}$ .

(A) Total system

(B) Condensed system

FIGURE 4.2. DPG stiffness matrix  $\tilde{\mathbf{B}}$ .

## 5. CONDITIONING

**5.1. Robustness.** Observe that if  $\mathcal{B}(\mathcal{U}_h) \subset \mathcal{R}_{\mathcal{V}} \mathcal{V}_r$  then  $\mathcal{B}(\mathcal{U}_h) \perp \mathcal{R}_{\mathcal{V}} \mathcal{V}_r^\perp$  for  $\mathcal{V} = \mathcal{V}_r \oplus \mathcal{V}_r^\perp$ . Therefore,

$$\mathcal{R}_{\mathcal{V}}^{-1} \mathcal{B}(\mathcal{U}_h) = (\mathcal{R}_{\mathcal{V}_r}^{-1} \oplus \mathcal{R}_{\mathcal{V}_r^\perp}^{-1}) \mathcal{B}(\mathcal{U}_h) = \mathcal{R}_{\mathcal{V}_r}^{-1} \mathcal{B}(\mathcal{U}_h) \oplus 0 = \mathcal{R}_{\mathcal{V}_r}^{-1} \mathcal{B}(\mathcal{U}_h).$$

This indicates that, in special circumstances, no error is introduced from discretizing the Riesz map  $\mathcal{R}_V$ . Using this result, rewriting (2.6), and proceeding similarly with (3.1), yields

$$(5.1) \quad \langle \mathcal{B}\mathbf{u}_{h,r}^{\text{opt}} - \ell, \mathcal{R}_{V_r}^{-1} \mathcal{B}\delta\mathbf{u}_h \rangle = 0 = \langle \mathcal{B}\mathbf{u}_h^{\text{opt}} - \ell, \mathcal{R}_V^{-1} \mathcal{B}\delta\mathbf{u}_h \rangle = \langle \mathcal{B}\mathbf{u}_h^{\text{opt}} - \ell, \mathcal{R}_{V_r}^{-1} \mathcal{B}\delta\mathbf{u}_h \rangle.$$

Thus, if  $\mathcal{B}(\mathcal{U}_h) \subset \mathcal{R}_V \mathcal{V}_r$ , then  $\mathbf{u}_h^{\text{opt}} = \mathbf{u}_{h,r}^{\text{opt}}$ ; meaning that the discrete solution is actually the same solution which minimizes the continuous residual. In the particular case of least-squares methods, as in (2.8), this implies

$$(5.2) \quad (\mathbf{A}_{\text{LS}})_{ij} = (\mathcal{L}\mathbf{u}_j, \mathcal{L}\mathbf{u}_i)_{L^2(\Omega)} = \langle \mathcal{B}' \mathcal{R}_{L^2(\Omega)}^{-1} \mathcal{B}\mathbf{u}_j, \mathbf{u}_i \rangle = \langle \mathcal{B}' \mathcal{R}_{V_r}^{-1} \mathcal{B}\mathbf{u}_j, \mathbf{u}_i \rangle = \mathbf{A}_{ij}.$$

Hence, even the monolithic coefficient matrix of a least-squares method,  $\mathbf{A}_{\text{LS}}$ , is exactly the same as that of a DLS discretization in the normal equation setting, provided  $\mathcal{B}(\mathcal{U}_h) \subset \mathcal{R}_V \mathcal{V}_r$ . However, only with the DLS discretization is the factorization of the matrix  $\mathbf{A} = \tilde{\mathbf{B}}^* \tilde{\mathbf{B}}$  readily available. One advantage being that with a QR algorithm, as discussed in Section 3.3, one could solve for exactly the same solution, but with less numerical sensitivity.

We will now compare the condition number behavior, numerical sensitivity, discretization error, and round-off error incurred in DPG methods. As in [63], in each of our experiments, the stiffness matrix was diagonally preconditioned:

$$\mathbf{A} \mapsto \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}, \quad \mathbf{f} \mapsto \mathbf{D}^{-1/2} \mathbf{f},$$

or, equivalently,

$$\tilde{\mathbf{B}} \mapsto \tilde{\mathbf{B}} \mathbf{D}^{-1/2}, \quad \tilde{\mathbf{l}} \mapsto \tilde{\mathbf{l}},$$

where  $\mathbf{D} = \text{diag}(\mathbf{A})$ . The cost of this procedure is computationally negligible and it is common practice to scale the matrix in this way before iterative solution methods. Meanwhile, it is performed implicitly in most direct solvers. Therefore, we presume no offense in this action. For additional perspective on several topics we do not cover, related to the condition number of DPG stiffness matrices, but with a focus on Stokes equation, we refer the interested reader to [63, Chapter 9].

*First-order system least squares.* For illustration, consider Poisson's equation in  $\mathbb{R}^2$  with body force  $f$  and Dirichlet boundary condition  $u|_{\partial\Omega} = \hat{f}$ . Here, the aim is to find  $(u, \boldsymbol{\sigma}) \in (u^{\text{lift}} + H_0^1(\Omega)) \times \mathbf{H}(\text{div}, \Omega)$  such that

$$(5.3) \quad \begin{aligned} -\text{div } \boldsymbol{\sigma} + \alpha u &= f, \\ \boldsymbol{\sigma} - \text{grad } u &= 0, \end{aligned}$$

where,  $\alpha = 0$ ,  $f \in L^2$ , and  $u^{\text{lift}} \in H^1(\Omega)$  is an extension to  $\Omega$  where  $\text{tr}_{H^1(\Omega)} u^{\text{lift}} = \hat{f} \in H^{1/2}(\partial\Omega)$ . These equations can be solved using the first-order system least-squares (FOSLS) method [16, 17] by directly discretizing the trial space  $\mathcal{U}^{\text{hom.}} = H_0^1(\Omega) \times \mathbf{H}(\text{div}, \Omega)$ , as in (2.8). The appropriate discretization  $\mathcal{U}_h^{\text{hom.}} \subset \mathcal{U}^{\text{hom.}}$  should be locally derived from an exact sequence of spaces of polynomial order  $p$ , such as the ones found in [36] for triangles and quadrilaterals. Indeed, for a master square,  $(0, 1)^2$ , consider the spaces  $W^p = \mathcal{Q}^{p,p}$ ,  $\mathbf{V}^p = \mathcal{Q}^{p,p-1} \times \mathcal{Q}^{p-1,p}$  and  $Y^p = \mathcal{Q}^{p-1,p-1}$ , where  $\mathcal{Q}^{p,q} = \mathcal{P}^p(x) \otimes \mathcal{P}^q(y)$ . Thus, for a quadrilateral mesh, a compatible basis  $\mathcal{U}_h^{\text{hom.}}$  of  $\mathcal{U}_h^{\text{hom.}} \subset \mathcal{U}^{\text{hom.}}$  can be constructed, locally, at each element  $K \in \mathcal{T}$ , as the pullback of  $W^p \times \mathbf{V}^p$  from the master square to  $K$ .

On the other hand, solving (5.3) using a DLS discretization requires defining a bilinear form with a test space  $\mathcal{V}$  and its corresponding discretization  $\mathcal{V}_r \subset \mathcal{V}$ . Thus, dropping the superscript-hom. from now on, the goal is that of finding  $(u, \boldsymbol{\sigma}) \in H_0^1(\Omega) \times \mathbf{H}(\text{div}, \Omega) = \mathcal{U}$  such that

$$b((u, \boldsymbol{\sigma}), (v, \boldsymbol{\tau})) = \ell((v, \boldsymbol{\tau})), \quad \text{for all } (v, \boldsymbol{\tau}) \in L^2(\Omega) \times \mathbf{L}^2(\Omega) = \mathcal{V},$$

where

$$(5.4) \quad \begin{aligned} b((u, \boldsymbol{\sigma}), (v, \boldsymbol{\tau})) &= -(\text{div } \boldsymbol{\sigma}, v)_{\Omega} + (\alpha u, v)_{\Omega} + (\boldsymbol{\sigma}, \boldsymbol{\tau})_{\Omega} - (\text{grad } u, \boldsymbol{\tau})_{\Omega}, \\ \ell((v, \boldsymbol{\tau})) &= (f, v)_{\Omega} - (\alpha u^{\text{lift}}, v)_{\Omega} + (\text{grad } u^{\text{lift}}, \boldsymbol{\tau})_{\Omega}. \end{aligned}$$

And, where  $(\cdot, \cdot)_{\Omega} = (\cdot, \cdot)_{L^2(\Omega)}$ ,  $\mathbf{L}^2(\Omega) = L^2(\Omega) \times L^2(\Omega)$  (with the standard norm  $\langle \mathcal{R}_V \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|_{\mathcal{V}}^2 = \|\mathbf{v}\|_{L^2(\Omega)}^2$  for all  $\mathbf{v} = (v, \boldsymbol{\tau}) \in \mathcal{V}$ ), and  $(u + u^{\text{lift}}, \boldsymbol{\sigma})$  is the solution to (5.3). This is known as the (first-order)

“strong variational formulation” of the Poisson equation [19]. A DLS discretization can be found by using the same finite-dimensional spaces  $\mathcal{U}_h$  as in the FOSLS method above, as well as an  $L^2$ -conforming test space  $\mathcal{V}_r$  derived from an exact sequence of order  $p + \Delta p$ . Here, locally,  $\mathcal{V}_r$  would be a pullback of  $(Y^{p+\Delta p})^3$ . Notice that, at the master element level,  $\operatorname{div} \boldsymbol{\sigma} \in Y^p$ , while  $\boldsymbol{\sigma} - \operatorname{grad} u \in (Y^{p+1})^2$ . Thus, it follows that  $\mathcal{B}(\mathcal{U}_h) \subset \mathcal{R}_{\mathcal{V}} \mathcal{V}_r$  whenever  $\Delta p \geq 1$ .

Let  $\Omega = (0, 1)^2$  be partitioned by a uniform quadrilateral mesh of side length  $h$ , and assume  $u(x, y) = \sin(\pi x) \sin(\pi y)$ . Define  $\hat{f} = 0$  and  $f = -\operatorname{div}(\operatorname{grad} u)$ . With the lift  $u^{\operatorname{lift}} = 0$ , we separately solved (5.3) using both least-squares (i.e. FOSLS) and DLS methods (with  $\Delta p \geq 1$ ). Moreover, because of the uniformity of the mesh and from using  $L^2$ -orthogonal shape functions at the master element level [36], we produced a diagonal Gram matrix,  $G_{ij} = (\mathbf{v}_i, \mathbf{v}_j)_{L^2(\Omega)}$ .

The condition number we computed for each of the three possible stiffness matrices is presented in Figure 5.1. Note that the condition number of the FOSLS stiffness matrix  $A_{LS}$  was verified to grow as  $\mathcal{O}(h^{-2})$  (same as with a DLS stiffness matrix  $A$ ), but the condition number of the DLS enriched stiffness matrix  $\tilde{B}$  was verified to grow only as  $\mathcal{O}(h^{-1})$ .

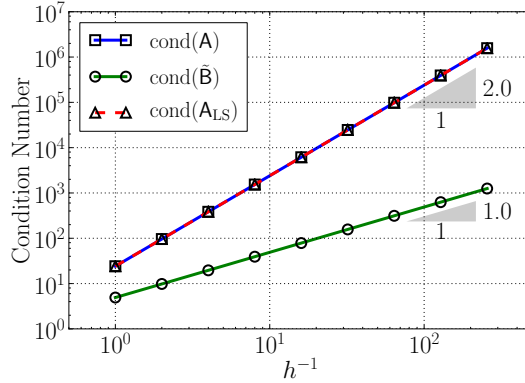


FIGURE 5.1. Comparison with Poisson’s equation of the condition number of the FOSLS stiffness matrix  $A_{LS}$  to the condition number of the DLS stiffness matrices  $A$  and  $\tilde{B}$  coming from the strong formulation, (5.4). Exact sequence polynomial order  $p = 2$  for  $\mathcal{U}_h$  and, in the DLS setting,  $\Delta p = 1$ . Observe that  $\operatorname{cond}(A_{LS}) = \operatorname{cond}(A) = \operatorname{cond}(\tilde{B})^2$ . All reported results are for the statically condensed and diagonally preconditioned matrices.

Recall (2.4) and (3.1). The FOSLS solution would be  $\mathbf{u}_h^{\operatorname{opt}}$ , while that coming from the DLS discretization would be  $\mathbf{u}_{h,r}^{\operatorname{opt}}$ . As expected from (5.2), numerical results also confirmed that when  $\Delta p \geq 1$ ,  $\mathbf{u}_h^{\operatorname{opt}} = \mathbf{u}_{h,r}^{\operatorname{opt}}$  and  $A_{LS} = A$ , up to floating-point precision.

When  $\mathcal{B}(\mathcal{U}_h) \not\subset \mathcal{R}_{\mathcal{V}} \mathcal{V}_r$ , the solutions  $\mathbf{u}_h^{\operatorname{opt}}$  and  $\mathbf{u}_{h,r}^{\operatorname{opt}}$ , and matrices  $A_{LS}$  and  $A$  will no longer be equal. Naturally, however, the distance between  $\mathbf{u}_{h,r}^{\operatorname{opt}}$  to  $\mathbf{u}_h^{\operatorname{opt}}$  is expected to decrease as  $\mathcal{V}_r$  is enriched (i.e. as  $\Delta p$  is increased). Specifically, if the enriched test spaces are nested,  $\mathcal{V}_{r_1} \subsetneq \mathcal{V}_{r_2} \subsetneq \mathcal{V}_{r_3} \subsetneq \dots$ , we expect  $\|\mathbf{u}_h^{\operatorname{opt}} - \mathbf{u}_{h,r_k}^{\operatorname{opt}}\|_{\mathcal{U}} \rightarrow 0$  as  $k \in \mathbb{N}$  increases. Indeed, this was observed when considering  $\alpha(x, y) = \sin(\pi x) \sin(\pi y)$  in (5.3) and (5.4) and comparing the FOSLS solution,  $\mathbf{u}_h^{\operatorname{opt}}$ , to the DLS normal equation solution,  $\mathbf{u}_{h,\Delta p}^{\operatorname{opt}}$ , for increasing values of  $\Delta p$ . The results in Figure 5.2 (A) show that the rate of  $h$ -convergence between the two discrete solutions grows with  $\Delta p$ . Moreover, Figure 5.2 (B) shows that the matrix  $A$  converges to  $A_{LS}$ . These numerical results suggest that the error incurred in discretizing  $\mathcal{V}$  to  $\mathcal{V}_r$  can be made very small, and we expect this to be true even with non-trivial variational formulations (i.e. after integrating by parts).

*Primal DPG and Bubnov-Galerkin methods.* In order to explore less trivial variational formulations, consider the broken primal formulation of the Poisson equation in (5.3) (with  $\alpha = 0$ ) [28]. In this setting, we seek a

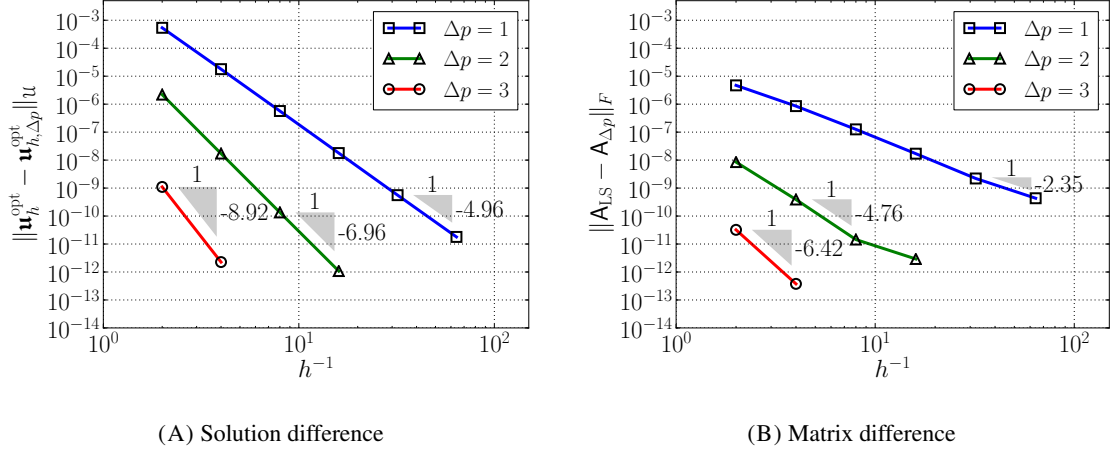


FIGURE 5.2. Distance between the FOSLS and DLS solutions,  $u_h^{\text{opt}}$  and  $u_{h,\Delta p}^{\text{opt}}$ , and the stiffness matrices  $A_{LS}$  and  $A$ . The solution  $u_h^{\text{opt}}$  was computed, in each experiment, by way of the normal equation (NE). Exact sequence polynomial order  $p = 2$  was used for  $\mathcal{U}_h$ , for all computations. Note that  $\|(u, \sigma)\|_{\mathcal{U}}^2 = \|u\|_{H^1}^2 + \|\sigma\|_{H(\text{div})}^2$  and  $\|\cdot\|_F^2$  is the Frobenius norm.

solution  $(u, \hat{\sigma}_n) \in H_0^1(\Omega) \times H^{-1/2}(\partial\mathcal{T}) = \mathcal{U}$  (see [19, 48] for definitions of mesh-trace Sobolev spaces) such that

$$b((u, \hat{\sigma}_n), v) = \ell(v), \quad \text{for all } v \in H^1(\mathcal{T}) = \mathcal{V},$$

where

$$(5.5) \quad \begin{aligned} b((u, \hat{\sigma}_n), v) &= (\text{grad } u, \text{grad } v)_{\mathcal{T}} - \langle \hat{\sigma}_n, v \rangle_{\partial\mathcal{T}}, \\ \ell(v) &= (f, v)_{\mathcal{T}} - (\text{grad } u^{\text{lift}}, \text{grad } v)_{\mathcal{T}}. \end{aligned}$$

Here,  $H^1(\mathcal{T})$  is a *broken* Sobolev space with norm  $\|v\|_{H^1(\mathcal{T})}^2$  [19], and the restriction of each member of this space to any single  $K \in \mathcal{T}$  is in  $H^1(K)$ . Likewise,  $(\cdot, \cdot)_{\mathcal{T}} = \sum_{K \in \mathcal{T}} (\cdot, \cdot)_K$  and similarly with  $\langle \cdot, \cdot \rangle_{\partial\mathcal{T}}$ . This second pairing,  $\langle \cdot, \cdot \rangle_{\partial\mathcal{T}}$ , can be understood, intuitively, as a mesh-boundary integral, however, the inquisitive reader may wish to examine [19, 48] for further detail. The lift  $u^{\text{lift}} \in H^1(\Omega)$  in (5.5) is identical to that from the least-squares setting presented before.

For discretization, let the trial space  $\mathcal{U}_h$  be derived from an exact sequence of order  $p$ . This implies that, at each element  $K \in \mathcal{T}$ , it is the pullback of  $W^p \times V^p|_{\partial\hat{K}} \cdot \hat{n}_{\hat{K}}$  from the master square  $\hat{K}$  to the physical element  $K$ . For the test space  $\mathcal{V}_r$ , it is sufficient to be from a  $p$ -enriched sequence of order  $p + \Delta p$ , locally. Therefore, at each  $K$  the test space is homeomorphic to  $W^{p+\Delta p}$ . Here, however, conformity requirements are not imposed across  $\partial K$ . That is,  $\mathcal{V}_r$  is a space of *discontinuous* piecewise-defined polynomials.

Observe that (5.5) is similar to the standard Bubnov-Galerkin problem; in which, the aim is to find  $u \in H_0^1(\Omega) = \mathcal{U}^{\text{BG}}$  such that

$$b^{\text{BG}}(u, v) = \ell^{\text{BG}}(v), \quad \text{for all } v \in H_0^1(\Omega) = \mathcal{V}^{\text{BG}} = \mathcal{U}^{\text{BG}},$$

where

$$(5.6) \quad \begin{aligned} b^{\text{BG}}(u, v) &= (\text{grad } u, \text{grad } v)_{\Omega}, \\ \ell^{\text{BG}}(v) &= (f, v)_{\Omega} - (\text{grad } u^{\text{lift}}, \text{grad } v)_{\Omega}. \end{aligned}$$

Here, the polynomial space  $W^p$  is commonly used for both  $\mathcal{U}_h^{\text{BG}} = \mathcal{V}_r^{\text{BG}}$ , locally, and continuity is required along the entirety of every element interface.

As far as we are aware, the condition number of the DPG (or DLS) stiffness matrix  $A$  coming from (5.5) has never been derived analytically. We leave that work to another researcher and so do not derive it here,

either. Nevertheless, we expect it to grow like  $h^{-2}$ , as the Bubnov-Galerkin stiffness matrix,  $A_{BG}$ , does [13]. This hypothesis was confirmed with experiments on a square domain  $\Omega = (0, 1)^2$  starting from a uniform mesh of four square elements and exact solution  $u(x, y) = \sin(10\pi x) \sin(10\pi y)$  (see Figure 5.3). Likewise, similar to the least-squares scenario,  $\text{cond}(\tilde{B}) = \text{cond}(A)^{1/2}$  was confirmed to grow only as  $h^{-1}$ , and so it eventually became less than the condition number of  $A_{BG}$ , which started out the smallest. Notably, in contrast to least-squares finite element methods, this shows that a first-order system formulation is not required to achieve  $\mathcal{O}(h^{-2})$  (or even  $\mathcal{O}(h^{-1})$ ) condition number growth with a DLS method.

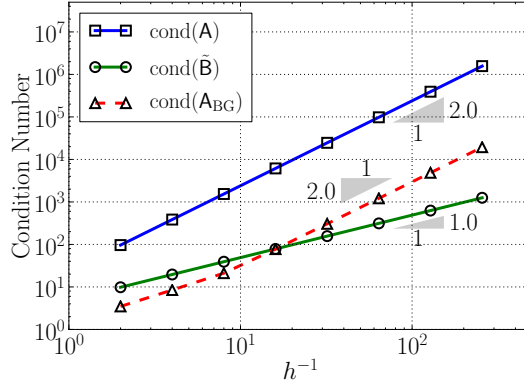


FIGURE 5.3. Comparison with Poisson’s equation of the condition number of the Bubnov-Galerkin stiffness matrix  $A_{BG}$  to the condition number of the DPG stiffness matrices  $A$  and  $\tilde{B}$  coming from the broken primal formulation, (5.5). Exact sequence polynomial order  $p = 2$  for  $\mathcal{U}_h$  and  $\mathcal{U}_h^{BG}$ , and, in the DPG setting,  $\Delta p = 1$ . Observe that  $\text{cond}(A_{BG}) \neq \text{cond}(A)$  and that  $\text{cond}(\tilde{B}) < \text{cond}(A_{BG})$ , eventually, for small enough  $h$ . All reported results are for the statically condensed and diagonally preconditioned matrices.

The optimal solution of the discrete minimum residual problem (2.6) coming from the variational formulation (5.5) would be  $\mathbf{u}_{h,\Delta p}^{\text{opt}} = (u_{h,\Delta p}^{\text{opt}}, (\hat{\sigma}_n^{\text{opt}})_{h,\Delta p}^{\text{opt}})$ . If we define the Bubnov-Galerkin problem’s solution to be  $u^{BG}$ , and the solution of (5.5) to be  $(u^{\text{opt}}, \hat{\sigma}_n^{\text{opt}})$  then, it can be shown that  $u^{BG} = u^{\text{opt}}$  [19]. However, for any given  $\Delta p$ , there is no reason to expect  $u_h^{BG}$  to be equal to  $u_{h,\Delta p}^{\text{opt}}$ . Indeed, these two solutions did not always agree as is demonstrated in Figure 5.4. Nevertheless, the two different solutions clearly converged to each other, rapidly.

Lastly, as suggested in Section 2.3, it is worth mentioning that, if  $f = 0$ , the Bubnov-Galerkin solution  $u_h^{BG}$  can also be derived from a minimization principle like (2.11):

$$(5.7) \quad u_h^{BG} = \arg \min_{u_h \in \mathcal{U}_h^{BG}} \|\text{grad } u_h - \mathbf{g}\|_{L^2(\Omega)}^2,$$

where  $\mathbf{g} = -\text{grad } u^{\text{lift}}$ . Likewise, with the DLS theory, it is not difficult to alternatively construct the Bubnov-Galerkin stiffness matrix through a factorization  $A_{BG} = \tilde{B}_{BG}^* \tilde{B}_{BG}$  [73]. Since this factorization is independent of  $f$ , this obviously appears appealing because of the reduced condition number of  $\tilde{B}_{BG}$ . Indeed, had  $\text{cond}(\tilde{B}_{BG})$  been computed and added to Figure 5.3 we would certainly verify that  $\text{cond}(\tilde{B}_{BG}) = \mathcal{O}(h^{-1})$ . Unfortunately, only in the scenario of (5.7) (i.e. when  $f = 0$ ) can orthogonal decomposition methods be used to solve for the solution  $u_h^{BG}$  more accurately and therefore to reduce its sensitivity to round-off error. Furthermore, Bubnov-Galerkin stiffness matrices for most difficult problems will not even be Hermitian/symmetric, so they can not admit *any* such factorization,  $A_{BG} \neq M^*M$  for any matrix  $M$ . Therefore, this observation is of extremely limited practical interest.



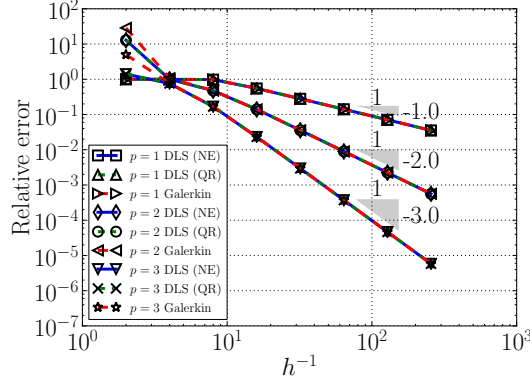


FIGURE 5.4. Relative error in the discrete solutions  $u_{h,\Delta p}^{\text{opt}}$  and  $u_h^{\text{BG}}$ . In the DPG solution,  $\Delta p = 1$ . Observe that, up to the smallest mesh size considered, the solutions  $u_{h,\Delta p}^{\text{opt}}$  obtained by the normal equation (NE) do not differ noticeably from those computed with QR factorization of the least-squares problem (LS).

*Ultraweak DPG methods.* Lastly, consider the broken ultraweak formulation of the Poisson equation in (5.3) (with  $\alpha = 0$ ). This seeks a solution  $(u, \sigma, \hat{u}, \hat{\sigma}_n) \in L^2(\Omega) \times \mathbf{L}^2(\Omega) \times H_0^{1/2}(\partial\mathcal{T}) \times H^{-1/2}(\partial\mathcal{T}) = \mathcal{U}$  (again, see [19, 48] for definitions of mesh-trace Sobolev spaces) such that

$$b((u, \sigma, \hat{u}, \hat{\sigma}_n), (v, \tau)) = \ell((v, \tau)), \quad \text{for all } (v, \tau) \in H^1(\mathcal{T}) \times \mathbf{H}(\text{div}, \mathcal{T}) = \mathcal{V},$$

where

$$(5.8) \quad \begin{aligned} b((u, \sigma, \hat{u}, \hat{\sigma}_n), (v, \tau)) &= (\sigma, \text{grad } v + \tau)_{\mathcal{T}} + (u, \text{div } \tau)_{\mathcal{T}} - \langle \hat{\sigma}_n, v \rangle_{\partial\mathcal{T}} - \langle \hat{u}, \tau \cdot \hat{n} \rangle_{\partial\mathcal{T}}, \\ \ell((v, \tau)) &= (f, v)_{\mathcal{T}} + \langle \hat{u}^{\text{lift}}, \tau \cdot \hat{n} \rangle_{\partial\mathcal{T}}. \end{aligned}$$

Again,  $H^1(\mathcal{T})$  and  $\mathbf{H}(\text{div}, \mathcal{T})$  are broken Sobolev spaces which essentially means that the restriction of each member to any single  $K \in \mathcal{T}$  is in  $H^1(K)$  and  $\mathbf{H}(\text{div}, K)$ , respectively. We assume that the norm for  $\mathcal{V}$  is then  $\|\mathbf{v}\|_{\mathcal{V}}^2 = \|v\|_{H^1(\mathcal{T})}^2 + \|\sigma\|_{\mathbf{H}(\text{div}, \mathcal{T})}^2$ , although other choices are possible [21]. Lastly,  $\hat{u}^{\text{lift}} \in H^{1/2}(\partial\mathcal{T})$  is an extension of  $\hat{f} \in H^{1/2}(\partial\Omega)$  to  $\partial\mathcal{T}$ .

As before, consider a discrete space  $\mathcal{U}_h$  derived from a sequence of order  $p$ , which would imply that it is (locally) pullbacks of  $Y^p \times (Y^p)^2 \times W^p|_{\partial\hat{K}} \times \mathbf{V}^p|_{\partial\hat{K}} \cdot \hat{n}_{\hat{K}}$  from the master element  $\hat{K}$  to the physical element  $K \in \mathcal{T}$ . Here, obtaining a discrete solution  $\mathbf{u}_h^{\text{opt}}$  minimizing the continuous residual (2.4) is not possible and so we have no reference discrete solution to compare different  $\mathbf{u}_{h,r}^{\text{opt}}$  with.

Let the enriched test space  $\mathcal{V}_r$  come from a sequence of order  $p + \Delta p$ , so that locally at each  $K$  the space is homeomorphic to  $W^{p+\Delta p} \times \mathbf{V}^{p+\Delta p}$ . In this situation, it was proven in [45] that, provided  $\Delta p \geq 2$ , the condition number of  $A$  would grow quadratically with  $h^{-1}$ . However, this was shown only with triangles, where  $W^p = \mathcal{P}^p$ ,  $\mathbf{V}^p = (\mathcal{P}^{p-1})^2 + x\mathcal{P}^{p-1}$  and  $Y^p = \mathcal{P}^{p-1}$ , and we have considered quadrilateral elements. Nevertheless, this  $\mathcal{O}(h^{-2})$  growth was indeed confirmed with computations as can be observed in Figure 5.5. More importantly, the enriched stiffness matrix  $\tilde{B}$  was, therefore, verified to have a much improved condition number growth of  $\mathcal{O}(h^{-1})$ .

**5.2. Failure study.** In some circumstances, finite element stiffness matrices can be so poorly conditioned that the round-off error in solving the discrete equations will compete with, or even surpass, the truncation error coming from the method itself and interpolation spaces being used. In such scenarios, exploiting the overdetermined system of equations with QR is very attractive.

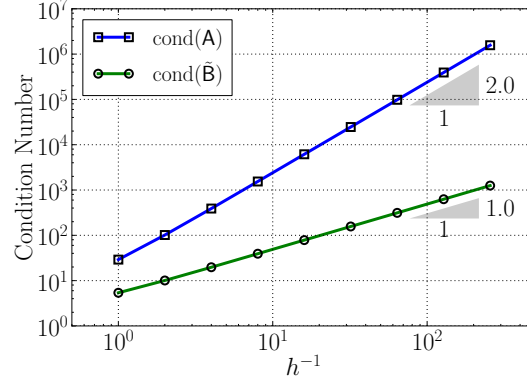
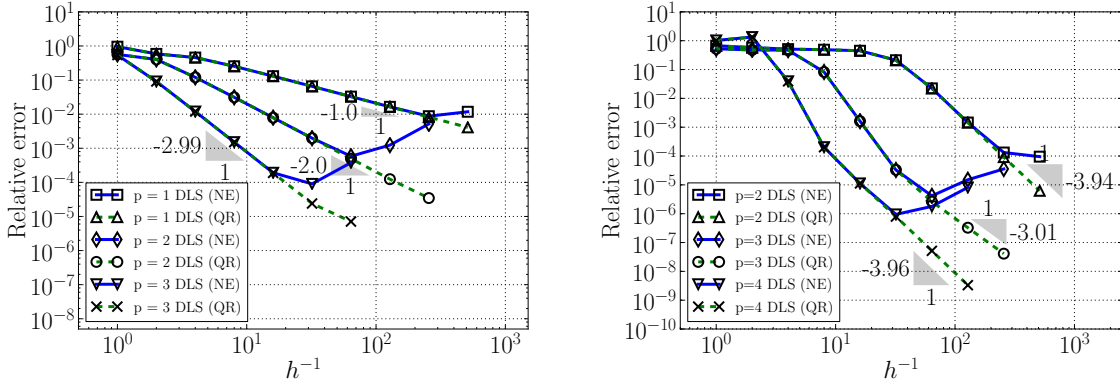


FIGURE 5.5. Condition number growth of the DPG stiffness matrices  $A$  and  $\tilde{B}$  coming from broken ultraweak formulation (5.8). Here,  $p = 2$  and  $\Delta p = 1$ . All reported results are for the statically condensed and diagonally preconditioned matrices.

Due to time and space limitation, we will illustrate this behavior only for ultraweak DPG methods for problems of the form (5.3). We have chosen the ultraweak setting because it is the most actively researched variational setting for DPG methods, at this time. For the normal equation, we solved the system with MUMPS 5.0.1 [52, 1]; and for the overdetermined system, we used qr\_mumps 1.2 [15]. Our results are reported in Figure 5.6.



(A) Poisson's equation (single precision).

(B) Linear acoustics (double precision).

FIGURE 5.6. Divergence of the discrete solution is observed for various polynomial orders  $p$  in two standard ultraweak DPG methods when the normal equation (NE) is constructed, statically condensed, and then solved. Notice that, instead, when QR factorization was used to directly solve the (statically condensed) least-squares problem (LS), the convergence of the discrete solution was maintained, at least for longer. In the  $p = 2$  run for (B), had more refinements been performed, we expect that the anticipated rate of convergence (i.e.  $2 \neq 3.94$ ) would have been recovered.  $\Delta p = 1$  in all experiments.

First, we performed a *single-precision* floating-point computation with Poisson's equation (5.3) to verify that round-off error would eventually overwhelm truncation error, even in the most well-behaved of problems. Here, we chose  $\Omega = (0, 1)^2$  and exact solution  $u(x, y) = x^2(1 - x)^2y^2(1 - y)^2$ , and imposed (homogeneous)

Dirichlet boundary conditions around the entire boundary  $\partial\Omega$ . Uniform  $h$ -refinements of quadrilateral elements, starting from a single element, were then performed until our computer ran out of memory. In Figure 5.6 (A), we report the loss of convergence to this polynomial exact solution for  $p = 1, 2$ , and 3 encountered after several mesh refinements when solving the normal equation. Notice that, however, for each polynomial order, `qr_mumps` applied to the overdetermined system continued to produce the expected rates of convergence after the normal equation approach failed.

In the  $p = 2$  case, we can use Figure 5.5 to corroborate this outcome. For example, note that, the two solutions began to visibly diverge at the 7<sup>th</sup> mesh. Here, the relative error was just below  $10^{-3}$ . Also, note that, from inspecting Figure 5.5, the corresponding condition number of the stiffness matrix  $A$  was approximately  $10^5$ . Since machine single precision is approximately  $10^{-7}$ , the round-off error would have been, at most, approximately  $10^{-2}$ . It was, therefore, large enough to compete with, or surpass, the truncation error—and this is indeed what we began to see.

Our second example is with the linear acoustics problem in  $\Omega = (0, 1)^2$ , which we chose to solve with the ultraweak formulation with broken test spaces, *near resonance*, in *double precision*. The equations of linear acoustics are:

$$(5.9) \quad \begin{aligned} i\omega p + \operatorname{div} \mathbf{u} &= f, \\ i\omega \mathbf{u} + \operatorname{grad} p &= 0. \end{aligned}$$

where  $p$  is the pressure and  $\mathbf{u}$  is the velocity. We imposed a hard boundary for this problem. That is Neumann-type boundary conditions  $\mathbf{u}|_{\partial\Omega} \cdot \hat{\mathbf{n}} = \hat{g}$  were specified on the entire boundary.

Similar to (5.8), we can derive the corresponding ultraweak formulation. In it, we seek a solution  $(p, \mathbf{u}, \hat{p}, \hat{u}_n) \in L^2(\Omega) \times \mathbf{L}^2(\Omega) \times H^{1/2}(\partial\mathcal{T}) \times H_0^{-1/2}(\partial\mathcal{T}) = \mathcal{U}$  such that

$$b((p, \mathbf{u}, \hat{p}, \hat{u}_n), (q, \mathbf{v})) = \ell((q, \mathbf{v})), \quad \text{for all } (q, \mathbf{v}) \in H^1(\mathcal{T}) \times \mathbf{H}(\operatorname{div}, \mathcal{T}) = \mathcal{V},$$

where

$$(5.10) \quad \begin{aligned} b((p, \mathbf{u}, \hat{p}, \hat{u}_n), (q, \mathbf{v})) &= -(p, i\omega q + \operatorname{div} \mathbf{v}) - (\mathbf{u}, i\omega \mathbf{v} + \operatorname{grad} q) + \langle \hat{u}_n, q \rangle_{\partial\mathcal{T}} + \langle \hat{p}, \mathbf{v} \cdot \hat{\mathbf{n}} \rangle_{\partial\mathcal{T}}, \\ \ell((q, \mathbf{v})) &= (f, q)_{\mathcal{T}} - \langle \hat{u}_n^{\text{lift}}, q \rangle_{\partial\mathcal{T}}. \end{aligned}$$

And, where  $\hat{u}_n^{\text{lift}} \in H^{-1/2}(\partial\mathcal{T})$  is an extension of  $\hat{g} \in H^{-1/2}(\partial\Omega)$  to  $\partial\mathcal{T}$ . The discretization is also similar to (5.8)—except in  $\mathbb{F} = \mathbb{C}$  instead of  $\mathbb{R}$ —and, again, uses pullbacks of  $Y^p \times (Y^p)^2 \times W^p|_{\partial\hat{K}} \times \mathbf{V}^p|_{\partial\hat{K}} \cdot \hat{\mathbf{n}}_{\hat{K}}$ .

We chose to solve the problem for  $\omega = 0.5001 \cdot 2\pi$  which is very close to a resonance frequency. Note that, the first eigenvalue of the Laplacian in this setting is  $\omega_1 = \pi$ . Therefore, we can expect that the stiffness matrix will become very badly-conditioned as the mesh is refined. Using a Gaussian beam for the exact solution (see [61] for equation) and a corresponding discrete lift of  $\hat{u}_n^{\text{lift}}$  constructed using projection-based interpolation [27], we clearly received more robust convergence with the QR approach. The results are presented in Figure 5.6 (B).

## 6. FUTURE DIRECTIONS

Many different research pathways for the DLS procedure described here are immediately possible, not the least of which is the exploration of iterative solvers. Another instinctive direction is towards residual problems,

$$\mathbf{u}_h^{\text{opt}} = \arg \min_{\mathbf{u}_h \in \mathcal{U}_h} \|\mathcal{F}(\mathbf{u}_h)\|_{\mathcal{V}'},$$

where  $\mathcal{F} : \mathcal{U} \rightarrow \mathcal{V}'$  is an appropriate nonlinear operator. In the analogous setting for the statistics community, Gauss-Newton algorithms,

$$\mathbf{u}_h^{\text{opt}} \mapsto \mathbf{u}_h^{\text{opt}} + \arg \min_{\Delta \mathbf{u}_h \in \mathcal{U}_h} \|D\mathcal{F}(\mathbf{u}_h)\Delta \mathbf{u}_h + \mathcal{F}(\mathbf{u}_h)\|_{\mathcal{V}'} \quad (\text{repeat}),$$

are often used to find local minima. And as has been noted in [22], similar techniques are very useful for approximating solutions of nonlinear PDE in the minimum residual fashion. As before, solving the associated normal equation will precede a less satisfactory error bound in each search direction,  $\Delta \mathbf{u}_h$ , than would occur from handling each least-squares problem directly. This is crucial when the discretization of  $D\mathcal{F}(\mathbf{u}_h)$  becomes

singular or nearly-singular because the corresponding generalized least-squares problem will always permit a solution, although uniqueness of the solution may be lost.

Moving to more general functional settings than Hilbert spaces—such as featured in Banach space minimum residual problems like those proposed in [55]—a natural extension may exist in applying iteratively reweighted least-squares techniques for a (iteratively updated) DLS discretization. What is more, for problems with parameter uncertainty a total least-squares [53] approach to DLS may have many benefits.

## 7. CONCLUSION

In this article we have presented a general framework for discrete least-squares finite element methods and illustrated features of this special class of methods. We have demonstrated, both analytically and discretely, how to incorporate essential boundary conditions and equality constraints into this framework and discussed different direct methods to solve the discrete equations. Due to space and time restrictions, our treatment of topics such as preconditioners and iterative solvers is extremely limited. Indeed, these topics are also somewhat open-ended and remain subjects of future research.

We have discussed assembly algorithms for two different formulations of DLS problems which can be solved with separate techniques. In our experiments with one construction, the growth of the condition number of the associated stiffness matrix, after only diagonal preconditioning, was demonstrated to be  $\mathcal{O}(h^{-1})$  when solving Poisson's equation. In other experiments which compared sensitivity to round-off error, we demonstrated the associated QR-factorization approach is particularly well-suited for ill-conditioned problems. We have also discussed some techniques which may further improve the conditioning of these matrices during assembly as well as static condensation which can improve the efficiency of the associated solvers.

Although the key results we present are directly towards DPG methods, we have maintained a comprehensive and broad perspective in our presentation which has allowed us to make several connections with other methods in the literature.

**Acknowledgements.** This work was partially supported with grants by NSF (DMS-1418822), AFOSR (FA9550-12-1-0484), and ONR (N00014-15-1-2496).

## APPENDIX A. CONSTRAINED MINIMIZATION

In this appendix, we demonstrate the necessary and sufficient conditions for existence and uniqueness of the general minimum residual principle over convex sets referred to in Section 2.4. We do not present *a priori* estimates for the rate of converge of the discrete solution. For further inquiring in the DPG setting, see [40].

Let  $\mathcal{K}_h \subset \mathcal{U}_h \subset \mathcal{U}$  be a non-empty convex set, let  $\|\mathbf{u}\|_{\mathcal{U}}$  be a semi-norm on  $\mathcal{U}$ , and assume that the continuous bilinear form  $b$  satisfies the continuous inf-sup condition

$$(A.1) \quad \sup_{\mathbf{v} \in \mathcal{V}} \frac{b(\mathbf{u}, \mathbf{v})}{\|\mathbf{v}\|_{\mathcal{V}}} = \|\mathcal{B}\mathbf{u}\|_{\mathcal{V}'} \geq \gamma \|\mathbf{u}\|_{\mathcal{U}},$$

for some constant  $\gamma > 0$  and all  $\mathbf{u} \in \mathcal{U}$ . The corresponding minimum residual problem on  $\mathcal{K}_h$  is

$$(A.2) \quad \min_{\mathbf{u}_h \in \mathcal{K}_h} \|\mathcal{B}\mathbf{u}_h - \ell\|_{\mathcal{V}'},$$

whose solution satisfies

$$(\mathcal{B}\mathbf{u}_h - \ell, \mathcal{B}(\delta\mathbf{u}_h - \mathbf{u}_h))_{\mathcal{V}'} \geq 0 \text{ for all } \delta\mathbf{u}_h \in \mathcal{K}_h.$$

Or, equivalently,

$$a(\mathbf{u}_h, \delta\mathbf{u}_h - \mathbf{u}_h) \geq f(\delta\mathbf{u}_h - \mathbf{u}_h) \quad \forall \delta\mathbf{u}_h \in \mathcal{K}_h,$$

where  $a : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{F}$  and  $f : \mathcal{U} \rightarrow \mathbb{F}$  are defined in Section 2.2.

**Theorem A.1.** *The operator  $\mathcal{A} = \mathcal{B}' \mathcal{R}_{\mathcal{V}}^{-1} \mathcal{B}$  is monotone and coercive on  $\mathcal{K}_h$  with discrete coercivity constant equal to  $\gamma^2$ , coming from (A.1).*

*Proof.* Observe that for every  $\mathbf{u}_h, \delta \mathbf{u}_h \in \mathcal{K}_h$ ,

$$\begin{aligned}
\langle \mathcal{A} \delta \mathbf{u}_h - \mathcal{A} \mathbf{u}_h, \delta \mathbf{u}_h - \mathbf{u}_h \rangle &= a(\delta \mathbf{u}_h - \mathbf{u}_h, \delta \mathbf{u}_h - \mathbf{u}_h) \\
&= b(\delta \mathbf{u}_h - \mathbf{u}_h, \mathcal{R}_V^{-1} \mathcal{B}(\delta \mathbf{u}_h - \mathbf{u}_h)) \\
&= \langle \mathcal{B}(\delta \mathbf{u}_h - \mathbf{u}_h), \mathcal{R}_V^{-1} \mathcal{B}(\delta \mathbf{u}_h - \mathbf{u}_h) \rangle \\
&= \| \mathcal{R}_V^{-1} \mathcal{B}(\delta \mathbf{u}_h - \mathbf{u}_h) \|_V^2 \\
&= \| \mathcal{B}(\delta \mathbf{u}_h - \mathbf{u}_h) \|_{V'}^2 \\
&\geq \gamma^2 \| \delta \mathbf{u}_h - \mathbf{u}_h \|_{\mathcal{U}}^2,
\end{aligned}$$

from which all results follow.  $\square$

**Corollary A.1.** *Solutions to (A.2) exist.*

**Corollary A.2.** *If  $\| \cdot \|_{\mathcal{U}}$  separates points in  $\mathcal{K}_h$  then  $\mathcal{A}$  is strictly monotone in  $\mathcal{K}_h$  and the solution of (A.2) is unique.*

The proof of each corollary is classic; see [51].

## APPENDIX B. STATIC CONDENSATION

The purpose of this appendix is to show that the static condensation derived in Section 3.4 coincides with the usual static condensation of finite element matrices via Schur complements. That is, that it coincides with the Schur complement of the normal equation, (NE). With this in mind consider the matrix  $A = \tilde{B}^* \tilde{B}$ , where  $\tilde{B} = [\tilde{B}_{\text{bubb.}} \ \tilde{B}_{\text{interf.}}]$ , so that the normal equation is rewritten as

$$(B.1) \quad \begin{bmatrix} \tilde{B}_{\text{bubb.}}^* & \tilde{B}_{\text{bubb.}} \\ \tilde{B}_{\text{interf.}}^* & \tilde{B}_{\text{interf.}} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\text{bubb.}} \\ \mathbf{u}_{\text{interf.}} \end{bmatrix} = A \mathbf{u} = \tilde{B}^* \tilde{\mathbf{l}} = \begin{bmatrix} \tilde{B}_{\text{bubb.}}^* \\ \tilde{B}_{\text{interf.}}^* \end{bmatrix} \tilde{\mathbf{l}}.$$

Then writing the Schur complement for  $\mathbf{u}_{\text{interf.}}$  yields,

$$(B.2) \quad \begin{aligned} \tilde{B}_{\text{interf.}}^* (I - P_{\text{bubb.}}) \tilde{B}_{\text{interf.}} \mathbf{u}_{\text{interf.}} &= (\tilde{B}_{\text{interf.}}^* \tilde{B}_{\text{interf.}} - \tilde{B}_{\text{interf.}}^* \tilde{B}_{\text{bubb.}} (\tilde{B}_{\text{bubb.}}^* \tilde{B}_{\text{bubb.}})^{-1} \tilde{B}_{\text{bubb.}}^* \tilde{B}_{\text{interf.}}) \mathbf{u}_{\text{interf.}} \\ &= \tilde{B}_{\text{interf.}}^* \tilde{\mathbf{l}} - \tilde{B}_{\text{interf.}}^* \tilde{B}_{\text{bubb.}} (\tilde{B}_{\text{bubb.}}^* \tilde{B}_{\text{bubb.}})^{-1} \tilde{B}_{\text{bubb.}}^* \tilde{\mathbf{l}} = \tilde{B}_{\text{interf.}}^* (I - P_{\text{bubb.}}) \tilde{\mathbf{l}}, \end{aligned}$$

where  $P_{\text{bubb.}} = \tilde{B}_{\text{bubb.}} (\tilde{B}_{\text{bubb.}}^* \tilde{B}_{\text{bubb.}})^{-1} \tilde{B}_{\text{bubb.}}^*$ . Then, the only observation that remains is that  $P_{\text{bubb.}} = P_{\text{bubb.}}^*$  is an orthogonal projection and it can easily be verified that  $I - P_{\text{bubb.}} = (I - P_{\text{bubb.}})^2$ , so that

$$(B.3) \quad ((I - P_{\text{bubb.}}) \tilde{B}_{\text{interf.}})^* (I - P_{\text{bubb.}}) \tilde{B}_{\text{interf.}} \mathbf{u}_{\text{interf.}} = ((I - P_{\text{bubb.}}) \tilde{B}_{\text{interf.}})^* (I - P_{\text{bubb.}}) \tilde{\mathbf{l}}.$$

This can easily be seen to be equivalent to an overdetermined system exactly the same as (3.20), and this is the desired result.

## REFERENCES

- [1] P. R. AMESTOY, I. S. DUFF, J.-Y. L'EXCELLENT, AND J. KOSTER, *A fully asynchronous multifrontal solver using distributed dynamic scheduling*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 15–41.
- [2] E. ANDERSON, Z. BAI, AND J. DONGARRA, *Generalized QR factorization and its applications*, Linear Algebra Appl., 162 (1992), pp. 243–271.
- [3] D. N. ARNOLD, R. S. FALK, AND R. WINTHER, *Finite element exterior calculus, homological techniques, and applications*, Acta Numer., 15 (2006), pp. 1–155.
- [4] A. T. BARKER, V. DOBREV, J. GOPALAKRISHNAN, AND T. KOLEV, *A scalable preconditioner for a DPG method*, ArXiv e-prints, (2016).
- [5] M. BENZI, G. H. GOLUB, AND J. LIESEN, *Numerical solution of saddle point problems*, Acta Numer., 14 (2005), pp. 1–137.
- [6] M. BERNDT, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Local error estimates and adaptive refinement for first-order system least squares (FOSLS)*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43.

- [7] S. BEUCHLER, V. PILLWEIN, J. SCHÖBERL, AND S. ZAGLMAYR, *Sparsity optimized high order finite element functions on simplices*, in Numerical and Symbolic Scientific Computing, U. Langer and P. Paule, eds., Texts & Monographs in Symbolic Computation, Springer, Vienna, 2012, pp. 21–44.
- [8] Å. BJÖRCK, *Numerical methods for least squares problems*, SIAM, 1996.
- [9] Å. BJÖRCK AND I. S. DUFF, *A direct method for the solution of sparse linear least squares problems*, Linear Algebra Appl., 34 (1980), pp. 43–67.
- [10] P. BOCHEV, *Least-squares finite element methods for first-order elliptic systems*, Int. J. Numer. Anal. Model, 1 (2004), pp. 49–64.
- [11] P. B. BOCHEV AND M. D. GUNZBURGER, *Least-squares finite element methods*, vol. 166, Springer Science & Business Media, 2009.
- [12] D. BOFFI, M. FORTIN, AND F. BREZZI, *Mixed finite element methods and applications*, Springer series in computational mathematics, Springer, Berlin, Heidelberg, 2013.
- [13] S. BRENNER AND R. SCOTT, *The mathematical theory of finite element methods*, vol. 15, Springer Science & Business Media, 2007.
- [14] T. BUI-THANH AND O. GHATTAS, *A PDE-constrained optimization approach to the discontinuous Petrov-Galerkin method with a trust region inexact Newton-CG solver*, Comput. Methods Appl. Mech. Engrg., 278 (2014), pp. 20–40.
- [15] A. BUTTARI, *Fine-grained multithreading for the multifrontal QR factorization of sparse matrices.*, SIAM J. Sci. Comput., 35 (2013).
- [16] Z. CAI, R. LAZAROV, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *First-order system least squares for second-order partial differential equations: Part I*, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.
- [17] Z. CAI, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *First-order system least squares for second-order partial differential equations: Part II*, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.
- [18] C. CARSTENSEN, L. DEMKOWICZ, AND J. GOPALAKRISHNAN, *A posteriori error control for DPG methods*, SIAM J. Numer. Anal., 52 (2014), pp. 1335–1353.
- [19] C. CARSTENSEN, L. DEMKOWICZ, AND J. GOPALAKRISHNAN, *Breaking spaces and forms for the DPG method and applications including Maxwell equations*, Comput. Math. Appl., 72 (2016), pp. 494 – 522.
- [20] J. CHAN, J. A. EVANS, AND W. QIU, *A dual Petrov-Galerkin finite element method for the convectiondiffusion equation*, Comput. Math. Appl., 68 (2014), pp. 1513–1529.
- [21] 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 (2014), pp. 771–795.
- [22] J. L. CHAN, *A DPG method for convection-diffusion problems*, PhD thesis, The University of Texas at Austin, Austin, Texas, U.S.A., 2013.
- [23] P. G. CIARLET, *Linear and Nonlinear Functional Analysis with Applications*, SIAM, 2013.
- [24] A. COHEN, W. DAHMEN, AND G. WELPER, *Adaptivity and variational stabilization for convection-diffusion equations*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 1247–1273.
- [25] W. DAHMEN, C. HUANG, C. SCHWAB, AND G. WELPER, *Adaptive Petrov-Galerkin methods for first order transport equations*, SIAM J. Numer. Anal., 50 (2012), pp. 2420–2445.
- [26] L. DEMKOWICZ, *Computing with hp Finite Elements. I. One and Two Dimensional Elliptic and Maxwell Problems*, Chapman & Hall/CRC Press, New York, October 2006.
- [27] L. DEMKOWICZ, *Polynomial exact sequences and projection-based interpolation with application to Maxwell equations*, in Mixed Finite Elements, Compatibility Conditions, and Applications, D. Boffi and L. Gastaldi, eds., vol. 1939 of Lecture Notes in Mathematics, Springer, Berlin, 2008, pp. 101–158.
- [28] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *A primal DPG method without a first-order reformulation*, Comput. Math. Appl., 66 (2013), pp. 1058–1064.
- [29] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *Discontinuous Petrov-Galerkin (DPG) method*, ICES Report 15-20, The University of Texas at Austin, 2015.
- [30] L. DEMKOWICZ, J. GOPALAKRISHNAN, S. NAGARAJ, AND P. SEPULVEDA, *A spacetime DPG method for the Schrödinger equation*, arXiv preprint arXiv:1610.04678, (2016).
- [31] L. DEMKOWICZ, J. KURTZ, D. PARDO, M. PASZYŃSKI, W. RACHOWICZ, AND A. ZDUNEK, *Computing with hp Finite Elements. II. Frontiers: Three Dimensional Elliptic and Maxwell Problems with Applications*, Chapman & Hall/CRC, New York, October 2007.
- [32] I. EKELAND AND R. TÉMAM, *Convex Analysis and Variational Problems*, vol. 28 of Classics in Applied Mathematics, SIAM, 1999.
- [33] T. ELLIS, J. CHAN, AND L. DEMKOWICZ, *Robust DPG methods for transient convection-diffusion*, in Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations, Springer, 2016, pp. 179–203.
- [34] T. ELLIS, L. DEMKOWICZ, AND J. CHAN, *Locally conservative discontinuous Petrov–Galerkin finite elements for fluid problems*, Comput. Math. Appl., 68 (2014), pp. 1530–1549.
- [35] F. FUENTES, L. DEMKOWICZ, AND A. WILDER, *Using a DPG method to validate DMA experimental calibration of viscoelastic materials*, arXiv preprint arXiv:1703.10998, (2017).
- [36] F. FUENTES, B. KEITH, L. DEMKOWICZ, AND S. NAGARAJ, *Orientation embedded high order shape functions for the exact sequence elements of all shapes*, Comput. Math. Appl., 70 (2015), pp. 353–458.
- [37] F. FUENTES, B. KEITH, L. DEMKOWICZ, AND P. L. TALLEC, *Coupled variational formulations of linear elasticity and the DPG methodology*, arXiv preprint arXiv:1609.08180, (2016).
- [38] T. FÜHRER AND N. HEUER, *Robust coupling of DPG and BEM for a singularly perturbed transmission problem*, Comput. Math. Appl., (2016).

- [39] T. FÜHRER, N. HEUER, AND J. SEN GUPTA, *A time-stepping DPG scheme for the heat equation*, Comput. Methods Appl. Math., 17 (2016), pp. 237–252.
- [40] T. FÜHRER, N. HEUER, AND E. P. STEPHAN, *On the DPG method for Signorini problems*, arXiv preprint arXiv:1609.00765, (2016).
- [41] G. GOLUB, *Numerical methods for solving linear least squares problems*, Numer. Math, 7 (1965), pp. 206–216.
- [42] G. H. GOLUB AND C. GREIF, *On solving block-structured indefinite linear systems*, SIAM J. Sci. Comput., 24 (2003), pp. 2076–2092.
- [43] G. H. GOLUB, C. GREIF, AND J. M. VARAH, *An algebraic analysis of a block diagonal preconditioner for saddle point systems*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 779–792.
- [44] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, vol. 4, The Johns Hopkins University Press, 2013.
- [45] J. GOPALAKRISHNAN AND W. QIU, *An analysis of the practical DPG method*, Math. Comp., 83 (2014), pp. 537–552.
- [46] M. GULLIKSSON AND P.-Å. WEDIN, *Modifying the QR-decomposition to constrained and weighted linear least squares*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1298–1313.
- [47] P. D. HOUGH AND S. A. VAVASIS, *Complete orthogonal decomposition for weighted least squares*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 369–392.
- [48] B. KEITH, F. FUENTES, AND L. DEMKOWICZ, *The DPG methodology applied to different variational formulations of linear elasticity*, Comput. Methods Appl. Mech. Engrg., 309 (2016), pp. 579 – 609.
- [49] B. KEITH, P. KNECHTGES, N. ROBERTS, S. ELGETI, M. BEHR, AND L. DEMKOWICZ, *An ultraweak DPG method for viscoelastic fluids*, ICES Report 16-26, The University of Texas at Austin, 2016.
- [50] A. KESSY, A. LEWIN, AND K. STRIMMER, *Optimal whitening and decorrelation*, Am. Stat., (2017).
- [51] D. KINDERLEHRER AND G. STAMPACCHIA, *An introduction to variational inequalities and their applications*, SIAM, 2000.
- [52] J. W. H. LIU, *The multifrontal method for sparse matrix solution: Theory and practice*, SIAM Review, 34 (1992), pp. 82–109.
- [53] I. MARKOVSKY AND S. VAN HUFFEL, *Overview of total least-squares methods*, Signal processing, 87 (2007), pp. 2283–2302.
- [54] D. MORO, N. NGUYEN, AND J. PERAIRE, *A hybridized discontinuous Petrov–Galerkin scheme for scalar conservation laws*, Int. J. Numer. Meth. Eng., 91 (2012), pp. 950–970.
- [55] I. MUGA AND K. G. VAN DER ZEE, *Discretization of linear problems in banach spaces: Residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods*, arXiv preprint arXiv:1511.04400, (2015).
- [56] S. NAGARAJ, S. PETRIDES, AND L. DEMKOWICZ, *Construction of a DPG Fortin operator*, ICES Report 15-22, The University of Texas at Austin, 2015.
- [57] J. T. ODEN AND L. F. DEMKOWICZ, *Applied functional analysis*, CRC Press Incorporated, 1996.
- [58] C. PAIGE, *Computer solution and perturbation analysis of generalized linear least squares problems*, Math. Comp., 33 (1979), pp. 171–183.
- [59] C. C. PAIGE, *Fast numerically stable computations for generalized linear least squares problems*, SIAM J. Numer. Anal., 16 (1979), pp. 165–171.
- [60] ———, *Some aspects of generalized QR factorizations*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Clarendon Press, Oxford, UK, 1990, pp. 71–91.
- [61] R. PASCHOTTA, *Article on ‘Gaussian beams’*, in Encyclopedia of Laser Physics and Technology, Wiley-VCH, ISBN 978-3-527-40828-3, 2008.
- [62] S. PETRIDES AND L. F. DEMKOWICZ, *An adaptive DPG method for high frequency time-harmonic wave propagation problems*, ICES Report 16-20, The University of Texas at Austin, 2016.
- [63] N. V. ROBERTS, *A discontinuous Petrov–Galerkin methodology for incompressible flow problems*, PhD thesis, The University of Texas at Austin, Austin, Texas, U.S.A., 2013.
- [64] ———, *Camellia v1.0 manual: Part I*, Tech. Rep. ANL/ALCF-16/3, Argonne National Laboratory, Argonne, Illinois, 2016.
- [65] N. V. ROBERTS, T. BUI-THANH, AND L. DEMKOWICZ, *The DPG method for the Stokes problem*, Comput. Math. Appl., 67 (2014), pp. 966–995.
- [66] N. V. ROBERTS AND J. CHAN, *A geometric multigrid preconditioning strategy for DPG system matrices*, arXiv preprint arXiv:1608.02567, (2016).
- [67] N. V. ROBERTS, L. DEMKOWICZ, AND R. MOSER, *A discontinuous Petrov–Galerkin methodology for adaptive solutions to the incompressible Navier–Stokes equations*, J. Comput. Phys., 301 (2015), pp. 456–483.
- [68] G. STRANG, *A framework for equilibrium equations*, SIAM Rev., 30 (1988), pp. 283–297.
- [69] ———, *Fast transforms: Banded matrices with banded inverses*, Proc. Natl. Acad. Sci. U.S.A., 107 (2010), pp. 12413–12416.
- [70] L. N. TREFETHEN AND D. BAU III, *Numerical linear algebra*, vol. 50, SIAM, 1997.
- [71] C. VAN LOAN, *On the method of weighting for equality-constrained least-squares problems*, SIAM J. Numer. Anal., 22 (1985), pp. 851–864.
- [72] S. A. VAVASIS, *Stable numerical algorithms for equilibrium systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1108–1131.
- [73] ———, *Stable finite elements for problems with wild coefficients*, SIAM J. Numer. Anal., 33 (1996), pp. 890–916.
- [74] J. Y. YUAN, *Numerical methods for generalized least squares problems*, J. Comput. Appl. Math., 66 (1996), pp. 571–584.