Acta Numerica (2025), pp. 293–384
doi:10.1017/S0962492924000102

The discontinuous Petrov–Galerkin method

Leszek Demkowicz
Oden Institute, The University of Texas at Austin,
Austin, TX 78712-1229, USA
E-mail: leszek@oden.utexas.edu

Jay Gopalakrishnan
PO Box 751 (MTH), Portland State University,
Portland, OR 97207-0751, USA
E-mail: gjay@pdx.edu

The discontinuous Petrov–Galerkin (DPG) method is a Petrov–Galerkin finite element method with test functions designed for obtaining stability. These test functions are computable locally, element by element, and are motivated by optimal test functions which attain the supremum in an inf-sup condition. A profound consequence of the use of nearly optimal test functions is that the DPG method can inherit the stability of the (undiscretized) variational formulation, be it coercive or not. This paper combines a presentation of the fundamentals of the DPG ideas with a review of the ongoing research on theory and applications of the DPG methodology. The scope of the presented theory is restricted to linear problems on Hilbert spaces, but pointers to extensions are provided. Multiple viewpoints to the basic theory are provided. They show that the DPG method is equivalent to a method which minimizes a residual in a dual norm, as well as to a mixed method where one solution component is an approximate error representation function. Being a residual minimization method, the DPG method yields Hermitian positive definite stiffness matrix systems even for non-self-adjoint boundary value problems. Having a built-in error representation, the method has the out-of-the-box feature that it can immediately be used in automatic adaptive algorithms. Contrary to standard Galerkin methods, which are uninformed about test and trial norms, the DPG method must be equipped with a concrete test norm which enters the computations. Of particular interest are variational formulations in which one can tailor the norm to obtain robust stability. Key techniques to rigorously prove convergence of DPG schemes, including construction of Fortin operators, which in the DPG case can be done element by element, are discussed in detail. Pointers to open frontiers are presented.

2020 Mathematics Subject Classification: Primary 65M60, 65N12
Secondary 35F45

CONTENTS

1. Introduction

In variational methods, approximate solutions are sought in 'trial' spaces, while equations are enforced using 'test' spaces. Methods with different trial and test spaces are referred to as Petrov–Galerkin (PG) formulations. A classical result on such methods, restated below in Theorem 1.1, provides the following useful insight for designing Petrov–Galerkin methods: while one must choose trial spaces with good approximation properties, test spaces may be chosen solely for stability. Leveraging this insight, discontinuous Petrov–Galerkin (DPG) methods were originally conceived (Demkowicz and Gopalakrishnan 2010, 2011*b*) as Petrov–Galerkin methods that obtain stability automatically by local test space design using discontinuous functions. The goal of this review is to provide an introduction to these methods and present selected recent advances. We review established DPG techniques, give a few new avenues to existing results, and also present a few new results. We describe the mathematical foundations for the popular features that make the DPG method a powerful tool for solving boundary value problems, including the ease with which automatic adaptivity can be enabled and stable solvers for complex problems can be built.

Let us begin by describing a standard difficulty in PG formulations of boundary value problems that the DPG method addresses. The 'wellposedness' (or continuous dependence of solutions on data) of PG formulations need not automatically yield stability of their discretizations, unlike coercive Galerkin formulations. In simpler formulations with equal trial and test spaces, once wellposedness of the variational formulation (usually set in infinite-dimensional Sobolev spaces) is established through coercivity, stability of the computable Galerkin discretization (using finite-dimensional subspaces) follows. The situation for non-coercive PG methods is more complicated.

To make it precise, we introduce the following *setting we shall use throughout this review* (and a few departures from it will be announced with ample warning). Let X and Y denote two Hilbert spaces over the complex field $\mathbb{C}$. Let Y^* denote the space of continuous antilinear (or conjugate-linear) functionals on Y and let $b(\cdot,\cdot)\colon X \times Y \to \mathbb{C}$ denote a continuous sesquilinear form. A wellposedness formulation is such that

$$\begin{gathered}\text{for any } \ell \in Y^*, \text{ there is a unique } x \in X \text{ satisfying}\\ b(x,y) = \ell(y) \quad \text{for all } y \in Y. \end{gathered} \tag{1.1}$$

By the well-known theory of mixed systems (Babuška 1971, Brezzi 1974, Ern and Guermond 2021, Nečas 1962), we know that (1.1) holds if and only if there is a $\gamma > 0$ such that

$$\inf_{0\neq z\in X} \sup_{0\neq y\in Y} \frac{|b(z,y)|}{\|z\|_X\|y\|_Y} \geq \gamma, \quad \text{and} \tag{1.2a}$$

$$\{y \in Y\colon b(z,y) = 0 \text{ for all } z \in X\} = \{0\}, \tag{1.2b}$$

or equivalently

$$\inf_{0\neq y\in Y} \sup_{0\neq z\in X} \frac{|b(z,y)|}{\|y\|_Y\|z\|_X} \geq \gamma, \quad \text{and} \tag{1.3a}$$

$$\{z \in X\colon b(z,y) = 0 \text{ for all } y \in Y\} = \{0\}. \tag{1.3b}$$

For a computationally realizable Petrov–Galerkin method, one uses finite-dimensional subspaces $X_h \subset X$ and $Y_h \subset Y$ of equal dimension, $\dim(X_h) = \dim(Y_h)$. Here h is a subscript related to the finite dimension. Letting

$$\|b\| = \sup_{0\neq x\in X} \sup_{0\neq y\in Y} \frac{|b(x,y)|}{\|x\|_X\|y\|_Y},$$

a classical result of Babuška (see Babuška 1971, Babuška, Aziz, Fix and Kellogg 1972 or Xu and Zikatanov 2003) can be stated as follows.

Theorem 1.1. In the above setting of Hilbert spaces X, Y and finite-dimensional subspaces $X_h \subset X$, $Y_h \subset Y$ satisfying $\dim(X_h) = \dim(Y_h)$, suppose (1.1), (1.2) or (1.3) hold. If, in addition, there exists a $\gamma_h > 0$ such that

$$\inf_{0\neq z_h\in X_h} \sup_{0\neq y_h\in Y_h} \frac{|b(z_h,y_h)|}{\|y_h\|_Y} \geq \gamma_h, \tag{1.4}$$

then there is a unique $x_h \in X_h$ satisfying

$$b(x_h,y_h) = \ell(y_h) \quad \text{for all } \ y_h \in Y_h \tag{1.5}$$

and

$$\|x - x_h\|_X \leq \frac{\|b\|}{\gamma_h} \inf_{z_h \in X_h} \|x - z_h\|_X. \tag{1.6}$$

The theorem clarifies the above-mentioned difficulty in inheriting discrete stability from the wellposedness of the variational problem. Specifically, the standard difficulty in the analysis of Petrov–Galerkin discretizations of the form (1.5) is that the *inf-sup condition* (1.2a) does not generally imply the *discrete inf-sup condition* (1.4). Hence, unlike coercive forms $b(\cdot, \cdot)$, it is easy to obtain unstable PG methods even when the variational equation (1.1) is wellposed. A second important observation is that in Theorem 1.1, the test space Y_h is absent from the error estimate (1.6). It only appears in the inf-sup condition (1.4), which is responsible for stability. The approximation rates are determined by the trial space X_h in (1.6). Letting the trial space carry the burden of achieving good approximation properties liberates the test space from it. The takeaway is that we can focus solely on stability when designing test spaces. Hence techniques to design discrete subspaces Y_h that guarantee the discrete inf-sup condition (1.4), with γ_h independent of the finite dimension, are useful.

The next section (Section 2) provides such a technique through the concept of optimal test functions which attain the supremum in an inf-sup condition. In Section 3 we shall see that DPG methods can equivalently be thought of as methods that minimize a residual in a non-standard norm, as well as a non-standard mixed method with an approximate error representation as a discrete solution component. One of the key steps that enable *local* computation of the optimal test functions is a reformulation of the boundary value problem using a test space of functions which have no continuity constraints across elements. Such spaces are often referred to as 'broken' spaces and the process is akin to 'hybridization'. We will see these terms again when the test space localization technique is introduced precisely in Section 4. Next, we address the usual practice of computing the optimal test functions inexactly. The loss of optimality in the stability constant and techniques to regain discrete stability despite the inexact computation are topics covered in Section 5. The key ingredient there is a local Fortin operator. The built-in error estimator contained in all DPG methods is then described in Section 6. The wide scope of applicability of DPG methods becomes clearer in Section 7, where we show how to accomplish the above-mentioned localization through a reformulation in broken graph spaces of very general partial differential equations (PDE). Application of optimal test functions to create enhanced time integrators is the subject of Section 8. Duality arguments, certain drawbacks in applying them to DPG methods, a dual DPG* method, and application to estimating error in goal functionals are briefly discussed in Section 9. Section 10 contains a collection of remarks on ongoing efforts to incorporate DPG techniques into nonlinear boundary value problems, a research area where the last word seems yet to be written. We conclude in Section 11 with pointers to further works whose details could not be included in this paper.

2. Optimal test spaces

Is it possible to find a test space Y_h for which the exact inf-sup condition (1.2a) implies the discrete inf-sup condition (1.4)? We begin with a simple and affirmative answer in Proposition 2.2 below. This then gives rise to an 'ideal Petrov–Galerkin method', a precursor to the DPG method.

Given any trial space X_h, we define its *optimal test space* for the continuous sesquilinear form $b(\cdot,\cdot)\colon X \times Y \to \mathbb{C}$ by

$$Y_h^{\text{opt}} = T(X_h), \tag{2.1}$$

where $T\colon X \to Y$, a 'trial-to-test operator', is defined by

$$(Tz, y)_Y = b(z, y) \quad \text{for all } y \in Y,\ z \in X. \tag{2.2}$$

Equation (2.2) uniquely defines a Tz for any given $z \in X$ by the Riesz representation theorem. We shall call Tz an *optimal test function* because it solves the optimization problem stated next.

Proposition 2.1 (Optimizer). For any $z \in X$, the maximum of

$$f_z(y) = |b(z, y)|/\|y\|_Y$$

over all non-zero $y \in Y$ is attained at $y = Tz$.

Proof. Rewriting f_z using (2.2), duality in Hilbert spaces implies

$$\sup_{0\neq y\in Y} f_z(y) = \sup_{0\neq y\in Y} \frac{|(Tz, y)_Y|}{\|y\|_Y} = \|Tz\|_Y.$$

Moreover, $f_z(Tz) = \|Tz\|_Y$. □

Proposition 2.2 (Exact inf-sup condition $\Longrightarrow$ Discrete inf-sup condition). If the inf-sup condition (1.2a) holds with some $\gamma > 0$, then the discrete inf-sup condition (1.4) holds with some $\gamma_h \geq \gamma > 0$ when we set $Y_h = Y_h^{\text{opt}}$.

Proof. For any $z_h \in X_h$, letting

$$s_1 = \sup_{0\neq y\in Y} \frac{|b(z_h, y)|}{\|y\|_Y}, \qquad s_2 = \sup_{0\neq y_h\in Y_h^{\text{opt}}} \frac{|b(z_h, y_h)|}{\|y_h\|_Y},$$

it is obvious that $s_1 \geq s_2$. To prove that $s_1 \leq s_2$, since $s_1 = \|Tz_h\|_Y$ by Proposition 2.1,

$$s_1 = \|Tz_h\|_Y = \frac{|(Tz_h, Tz_h)_Y|}{\|Tz_h\|_Y} \leq \sup_{y_h\in Y_h^{\text{opt}}} \frac{|(Tz_h, y_h)_Y|}{\|y_h\|_Y} = \sup_{y_h\in Y_h^{\text{opt}}} \frac{|b(z_h, y_h)|}{\|y_h\|_Y} = s_2,$$

so $s_1 = s_2$. Hence the discrete inf-sup condition (1.4) holds with $\gamma_h \geq \gamma$. □

Definition 2.3. For any trial subspace $X_h \subset X$, the *ideal Petrov Galerkin (IPG) method* finds $x_h \in X_h$ solving

$$b(x_h, y_h) = \ell(y_h) \quad \text{for all } y_h \in Y_h^{\text{opt}}, \tag{2.3}$$

where $Y_h^{\text{opt}} \subset Y$ is computed using the Y-inner product by (2.1)–(2.2).

Theorem 2.4 (Quasioptimality). If (1.3) holds, then the IPG method (2.3) is uniquely solvable for x_h and

$$\|x - x_h\|_X \le \frac{\|b\|}{\gamma} \inf_{z_h \in X_h} \|x - z_h\|_X, \tag{2.4}$$

where x is the unique exact solution of (1.1).

Proof. We apply Theorem 1.1. To verify that T is injective, suppose $Tz = 0$. Then, by (2.2), we have $b(z, y) = 0$ for all $y \in Y$, so (1.3b) implies that $z = 0$. Thus $\dim(X_h) = \dim(Y_h^{\text{opt}})$.

Next, since (1.3) implies (1.2), the other inf-sup condition,

$$\gamma \|z\|_Y \le \sup_{0 \ne y \in Y} \frac{|b(z, y)|}{\|y\|_Y} \quad \text{for all } z \in X, \tag{2.5}$$

holds with the same constant γ. Hence Proposition 2.2 shows that the discrete inf-sup condition (1.4) holds with the same constant, so Theorem 1.1 gives the result. □

Example 2.5 (L^2 least-squares). Suppose $\Omega \subset \mathbb{R}^N$, $N \ge 1$, is an open set and $A\colon X \to L^2(\Omega)^m$ is a continuous bijective linear operator (where, as before, X is some Hilbert space). Then, setting $Y = L^2(\Omega)^m$, the problem of finding a $u \in X$ such that $Au = f$, for any given $f \in Y$, can be put into a variational formulation by setting

$$b(u, v) = (Au, v)_Y, \quad \ell(v) = (f, v)_Y. \tag{2.6}$$

Then (2.2) implies that $Tu = Au$, so $Y_h^{\text{opt}} = AX_h$. Hence (2.3) reduces to

$$(Ax_h, Az_h)_Y = (f, Az_h)_Y \quad \text{for all } z_h \in X_h,$$

that is, for this example, the IPG method of (2.3) coincides with the standard $L^2(\Omega)$-based least-squares method, which has been well studied (Bochev and Gunzburger 2009, Cai, Lazarov, Manteuffel and McCormick 1994). Not all DPG methods are L^2 least-squares methods. But as we will see in the next section, all DPG methods minimize a residual in some norm, not necessarily the L^2-norm.

Note that Theorem 2.4 gives an error estimate for the L^2 least-squares method (2.6), since its assumptions are readily verified. Indeed, by the injectivity of A, any $z \in X$ satisfying $b(z, y) = 0$ for all $y \in L^2(\Omega)$ yields $Az = 0$, which implies that $z = 0$. Also,

$$\sup_{0 \ne z \in X} \frac{|(Az, y)_Y|}{\|z\|_X} \ge \frac{|(y, y)_Y|}{\|A^{-1}y\|_X} \ge \gamma \|y\|_Y,$$

with $\gamma = \|A^{-1}\|^{-1}$, so (1.3) holds. Hence the error estimate (2.4) holds.

Example 2.6 (A one-dimensional boundary value problem). Letting $\Omega = (0, 1)$, the unit interval in $\mathbb{R}$, and $f \in L^2(\Omega)$, consider the boundary value problem to find $u(x)$ satisfying

$$u' = f \quad \text{in } \Omega, \tag{2.7a}$$

$$u(0) = 0, \tag{2.7b}$$

for some given $f \in L^2(\Omega)$. This fits into the framework of Example 2.5 and yields the variational form (2.6) with $Au = du/dx \equiv u'$, $X = \{u \in H^1(\Omega) \colon u(0) = 0\}$, $m = 1$ and $Y = L^2(\Omega)$ (and it is easy to prove that $A \colon X \to Y$ is a bijection).

A different variational formulation for (2.7) can be obtained if we integrate by parts after multiplying (2.7a) by a test function. Then, using (2.7b) and letting the unknown value $u(1)$ to be a separate variable $\hat{u}_1$, to be determined, we have

$$-\int_0^1 uv' + \hat{u}_1 v(1) = \int_0^1 fv.$$

Grouping the trial variable into $z = (u, \hat{u}_1)$, set

$$b(z, v) \equiv b((u, \hat{u}_1), v) = \hat{u}_1 v(1) - \int_0^1 uv', \quad \ell(v) = \int_0^1 fv. \tag{2.8}$$

Set the spaces and their norms by

$$X = L^2(\Omega) \times \mathbb{R}, \quad Y = H^1(\Omega),$$

$$\|z\|_X^2 \equiv \|(u, \hat{u}_1)\|_X^2 = \|u\|_{L^2(\Omega)}^2 + |\hat{u}_1|^2, \quad \|v\|_Y^2 = \|v'\|_{L^2(\Omega)}^2 + |v(1)|^2.$$

By Sobolev inequality, $\|v\|_Y$ is equivalent to the standard $H^1(\Omega)$-norm. With these settings, it is easy to prove that (1.3) holds with

$$\gamma = \|b\| = 1. \tag{2.9}$$

One can also easily calculate the trial-to-test operator by analytically solving (2.2) for this example: for any $z = (u, \hat{u}_1) \in X$,

$$Tz \equiv T(u, \hat{u}_1) = \hat{u}_1 + \int_x^1 u(s)\, ds. \tag{2.10}$$

This implies that letting $P_p(\Omega)$ denote the space of polynomials of degree at most p, restricted to Ω, and setting the discrete trial space to $X_h = P_p(\Omega) \times \mathbb{R}$, we have

$$Y_h^{\text{opt}} = P_{p+1}(\Omega).$$

The solution $x_h = (u_h, \hat{u}_{1,h})$ of the resulting IPG method, in view of Theorem 2.4 and (2.9), is interesting in that u_h equals the $L^2(\Omega)$-projection of u onto $P_p(\Omega)$. In the general case, although one cannot expect the method to deliver the best L^2-approximation from the trial space, the solution is the best approximation in some norm, as will be proved in the next section.

Bibliographical notes. The material of this section is based on Demkowicz and Gopalakrishnan (2011*b*). Sequels by Demkowicz, Gopalakrishnan and Niemi (2012*a*) and Zitelli *et al.* (2011) developed the theme further. A prequel by Demkowicz and Gopalakrishnan (2010) focused solely on the transport equation (not discussed in the review) and used a test space consisting of parts of analytically solvable optimal test functions.

3. Minimization and other viewpoints

The ideal Petrov–Galerkin method (2.3) admits two equivalent reformulations, one as a least-squares method that minimizes the residual in a non-standard dual norm, and another as a mixed Galerkin method (with the same trial and test spaces) solving an associated min-max for a saddle point.

Let $R_Y : Y \to Y^*$ denote the standard *Riesz map* defined by $(R_Y y)(v) = (y, v)_Y$, for all y and v in Y. Here and throughout, the inner product of Y is denoted by $(\cdot,\cdot)_Y$. It is well known to be invertible and isometric:

$$\|R_Y y\|_{Y^*} = \|y\|_Y. \tag{3.1}$$

Let $B : X \to Y^*$ be the operator generated by the form $b(\cdot,\cdot)$, i.e. $Bx(y) = b(x, y)$ for all $x \in X$ and $y \in Y$. Since the defining equation of T, namely (2.2), can be rewritten using this notation as $R_Y T z = Bz$, we see that

$$T = R_Y^{-1} \circ B. \tag{3.2}$$

3.1. Equivalent characterization as a residual minimizer

Definition 3.1. On the trial space, we define an *energy norm* of $z \in X$ by $|||z|||_X := \|Tz\|_Y$. Clearly, by Proposition 2.1,

$$|||z|||_X = \|Tz\|_Y = \sup_{0 \neq y \in Y} \frac{|b(z, y)|}{\|y\|_Y}.$$

This is indeed a norm if (1.2a) holds due to easily seen norm equivalence

$$\gamma \|z\|_X \leq |||z|||_X \leq \|b\| \|z\|_X \quad \text{for all } z \in X.$$

Theorem 3.2 (Residual minimization). Suppose (1.1) holds. Then the following are equivalent statements for any given x_h in X_h.

(a) The x_h is the unique solution of the IPG method (2.3).

(b) The x_h is the best approximation to x from X_h in the sense that

$$|||x - x_h|||_X = \inf_{z_h \in X_h} |||x - z_h|||_X.$$

(c) The x_h minimizes the residual in the following sense:

$$x_h = \arg \min_{z_h \in X_h} \|\ell - B z_h\|_{Y^*}.$$

Proof. (a) $\Longleftrightarrow$ (b) By definition of the IPG method, x_h solves (2.3) if and only if $b(x - x_h, y_h) = 0$ for all $y_h \in Y_h^{\text{opt}}$. By the definition of the optimal test space, this is equivalent to

$$b(x - x_h, Tz_h) = 0 \quad \text{for all } z_h \in X_h,$$

which, in turn, is equivalent to

$$(T(x - x_h), Tz_h)_Y = 0 \quad \text{for all } z_h \in X_h,$$

due to (2.2). The result follows since $(T\cdot, T\cdot)_Y$ is the inner product generating the $|||\cdot|||_X$-norm.

(b) $\Longleftrightarrow$ (c) In view of (3.2),

$$|||x - z_h|||_X = \|T(x - z_h)\|_Y = \|R_Y^{-1} B(x - z_h)\|_Y.$$

Hence, by the isometry of the Riesz map (3.1), item (b) holds if and only if

$$\|B(x - x_h)\|_{Y^*} = \inf_{z_h \in X_h} \|B(x - z_h)\|_{Y^*},$$

which, since $\ell = Bx$, is the same as (c). □

3.2. *Equivalent characterization as a mixed formulation*

Definition 3.3. Let x be as in (1.1) and let x_h solve (2.3). Following earlier terminology, the *Riesz representation of the residual*, namely $\varepsilon = R_Y^{-1}(\ell - Bx_h)$, is often called the *error representation* (function). Clearly,

$$\|\varepsilon\|_Y = \|R_Y^{-1} B(x - x_h)\|_Y = \|T(x - x_h)\|_Y = |||x - x_h|||_X,$$

that is, the Y-norm of ε measures the error in the energy norm. Note that ε is the unique element of Y satisfying

$$(\varepsilon, y)_Y = \ell(y) - b(x_h, y) \quad \text{for all } y \in Y. \tag{3.3}$$

Theorem 3.4 (Mixed Galerkin reformulation). The following are equivalent statements.

(a) $x_h \in X_h$ solves the IPG method (2.3).

(b) x_h and ε solve the mixed formulation

$$(\varepsilon, y)_Y + b(x_h, y) = \ell(y) \quad \text{for all } y \in Y, \tag{3.4a}$$

$$b(z_h, \varepsilon) = 0 \quad \text{for all } z_h \in X_h. \tag{3.4b}$$

(c) ε and x_h form the saddle point of $L(y, z) = \frac{1}{2}\|y\|_Y^2 - \ell(y) + b(z, y)$ on $Y \times X_h$,

$$L(\varepsilon, x_h) = \min_{y \in Y} \max_{z \in X_h} L(y, z).$$

Proof. (a) $\Longrightarrow$ (b) Equation (3.4a) is the same as (3.3), so we only need to prove (3.4b). To this end,

$$b(z_h, \varepsilon) = (Tz_h, \varepsilon)_Y = (Tz_h, R_Y^{-1}(\ell - Bx_h))_Y = (Tz_h, T(x - x_h))_Y,$$

which, being the conjugate of $b(x - x_h, Tz_h)$, vanishes.

(b) $\Longrightarrow$ (a) Since (3.4a) implies $b(x_h, y_h) = \ell(y_h) - (\varepsilon, y_h)_Y$ for all $y_h \in Y_h^{\text{opt}}$, it suffices to prove that $(\varepsilon, y_h)_Y = 0$ for all $y_h \in Y_h^{\text{opt}}$. Any $y_h \in Y_h^{\text{opt}}$ is of the form $y_h = Tz_h$ for some $z_h \in X_h$, so

$$(\varepsilon, y_h)_Y = (Tz_h, \varepsilon)_Y = b(z_h, \varepsilon) = 0$$

by (3.4b).

(b) $\Longleftrightarrow$ (c) This follows from classical results on mixed methods (see e.g. Brezzi and Fortin 1991, Ch. II) or duality theory (see e.g. Ekeland and Témam 1999, Ch. VI). □

3.3. Optimal test norm and another trial-to-test operator

We have seen in Theorem 3.2(b) that the ideal PG method produces the best approximation in the energy norm $|||\cdot|||_X$ (defined in Definition 3.1). In practice, one may want the best approximation in a given trial space norm, say $\|\cdot\|_X$. Is it possible to engineer a test space norm such that the solution is the best approximation in a wanted trial space norm? The simple answer in the affirmative is provided by the optimal test norm, introduced below in the context of a generalized duality pairing.

We write the duality pairing in any Hilbert space Y as either

$$f(y) \quad \text{or} \quad \langle f, y\rangle_Y. \tag{3.5a}$$

Both denote the action of some $f \in Y^*$ on a $y \in Y$. The duality pairing satisfies

$$\|f\|_{Y^*} = \sup_{0\neq y\in Y} \frac{|\langle f, y\rangle_Y|}{\|y\|_Y} \quad \text{and} \quad \|y\|_Y = \sup_{0\neq f\in Y^*} \frac{|\langle f, y\rangle_Y|}{\|f\|_{Y^*}}. \tag{3.5b}$$

Definition 3.5. Analogous to the energy norm $|||\cdot|||_X$ in Definition 3.1, we define the *optimal test norm* $||||y||||_Y$ of any y in the test space Y by

$$||||y||||_Y = \sup_{0\neq z\in X} \frac{|b(z, y)|}{\|z\|_X}. \tag{3.6}$$

This is obviously a norm when (1.3) holds. We shall refer to a generic sesquilinear form $b(\cdot, \cdot)\colon X \times Y \to \mathbb{C}$ as a *generalized duality pairing* if

$$|||z|||_X = \|z\|_X \quad \text{and} \quad ||||y||||_Y = \|y\|_Y \tag{3.7}$$

hold for all $z \in X$ and $y \in Y$. This terminology is motivated by the standard duality pairing $b(\cdot, \cdot) = \langle \cdot, \cdot\rangle_Y$ in the case $X = Y^*$, where (3.5) implies

$$\|z\|_X = \sup_{0\neq y\in Y} \frac{|b(z, y)|}{\|y\|_Y} \quad \text{and} \quad \|y\|_Y = \sup_{0\neq z\in X} \frac{|b(z, y)|}{\|z\|_X}, \tag{3.8}$$

a pair of identities equivalent to (3.7). One of the pair implies the other, as shown shortly.

Let $B\colon X \to Y^*$ denote the operator generated by $b(\cdot,\cdot)$ by $\langle Bz, y\rangle_Y = b(z,y)$ for all $z \in X$, $y \in Y$. Identifying the bidual Y^{**} with Y, the adjoint $B^*\colon Y \to X^*$ of B satisfies

$$(B^*y)(z) = \overline{b(z,y)} \quad \text{for all } z \in X,\ y \in Y. \tag{3.9}$$

Using the Riesz maps in X and Y, we then immediately have

$$b(z,y) = \big(R_Y^{-1}Bz, y\big)_Y = \big(z, R_X^{-1}B^*y\big)_X, \quad z \in X,\ y \in Y. \tag{3.10}$$

This readily implies the twin identities

$$|||z|||_X = \|R_Y^{-1}Bz\|_Y, \quad |||\!|y|\!|||_Y = \|R_X^{-1}B^*y\|_X \tag{3.11}$$

for all $z \in X$ and $y \in Y$, by the definitions of energy norm and optimal test norm. Now we show that one may equivalently shorten the definition of the generalized duality pairing by omitting one of the two equalities in (3.7).

Proposition 3.6. The identity $|||z|||_X = \|z\|_X$ holds for all $z \in X$ if and only if $|||\!|y|\!|||_Y = \|y\|_Y$ for all $y \in Y$. Therefore, whenever either equality holds, we have $\|b\| = 1$ and $\gamma = 1$.

Proof. If $|||z|||_X = \|z\|_X$ for all $z \in X$, then (3.6) and (3.11) imply

$$|||\!|y|\!|||_Y = \sup_{0\neq z\in X} \frac{|b(z,y)|}{\|z\|_X} = \sup_{0\neq z\in X} \frac{|(R_Y^{-1}Bz, y)_Y|}{\|R_Y^{-1}Bz\|_Y}.$$

The last supremum equals $\|y\|_Y$ since $R_Y^{-1}B\colon X \to Y$ is a bijection. The converse is proved similarly using the other identity in (3.11). The last assertion on $\|b\|$ and γ immediately follows from (3.8). □

Clearly one direction of Proposition 3.6 answers the question posed at the beginning of this subsection. If we use the optimal test norm for Y, then the energy norm coincides with the given $\|\cdot\|_X$-norm, and Theorem 3.2(b) shows that the solution of the IPG method is guaranteed to be the best approximation in the given X-norm. However, as we shall see later, the optimal test norm is often not practically computable easily in the multi-dimensional examples we have in mind.

Next, we contrast the previously introduced trial-to-test operator which produces optimal test functions with an earlier trial-to-test operator given in Barrett and Morton (1984). To this end, let us introduce an adaptation of their ideas to our current Petrov–Galerkin setting. (They used equal trial and test spaces.) We define the 'Barrett–Morton trial-to-test operator' $T^{\mathrm{BM}}\colon X \to Y$ by

$$b(w, T^{\mathrm{BM}}z) = (w,z)_X \quad \text{for all } w, z \in X. \tag{3.12}$$

Using the inverse of B^*, an equivalent characterization of T^{BM} is

$$T^{\mathrm{BM}} = (B^*)^{-1} \circ R_X.$$

Comparing (3.12) with (2.2), we find two different trial-to-test mappings. The difference between our $T = R_Y^{-1} \circ B$ (see (3.2)) and T^{BM} is illustrated in the following diagram:

$$\begin{array}{ccc} X & \xrightarrow{B} & Y^* \\ \downarrow{\scriptstyle R_X} & & {\scriptstyle R_Y}\uparrow \\ X^* & \xleftarrow[B^*]{} & Y \end{array}$$

which is not commutative in general, and which further clarifies that $T \neq T^{\mathrm{BM}}$ in general.

Analogous to the IPG method of Definition 2.3, we can now consider a similar method using T^{BM} in place of T. Using any given trial subspace $X_h \subset X$, consider finding $x_h^{\mathrm{BM}} \in X_h$ that solves

$$b\big(x_h^{\mathrm{BM}}, y_h\big) = \ell(y_h) \quad \text{for all } y_h = T^{\mathrm{BM}} z_h,\ z_h \in X_h. \tag{3.13}$$

Subtracting this equation from (1.1) and substituting $w = x - x_h$ and $z = z_h$ in (3.12), we learn that

$$0 = b\big(x - x_h^{\mathrm{BM}}, T^{\mathrm{BM}} z_h\big) = \big(x - x_h^{\mathrm{BM}}, w_h\big)_X \quad \text{for all } w_h \in X_h.$$

This implies the remarkable property that the solution $x_h^{\mathrm{BM}} \in X_h$ of the method (3.13) equals the X-orthogonal projection of the exact solution x and explains the potential interest in the method (3.13). However, inverting B^* to compute test space basis functions is generally too expensive. In contrast, we will show in later sections that the inversion of R_Y to compute T can be realized locally if the problem is reformulated adequately.

Nevertheless, at this point it is useful to note one scenario where T and T^{BM} coincide. This occurs when b is a generalized duality pairing.

Proposition 3.7. If $\|z\|_X = |||z|||_X$ for all $z \in X$, then $T = T^{\mathrm{BM}}$.

Proof. By (3.11), $|||z|||_X = \|R_Y^{-1} B z\|_Y = \|Tz\|_Y$. Hence, whenever $\|z\|_X = |||z|||_X$ for all $z \in X$, by polarization, we have

$$(w, z)_X = (Tw, Tz)_Y = b(w, Tz) \quad \text{for all } z, w \in X,$$

where we have used (2.2) in the last equality. Comparing this with (3.12), we find that $b(w, Tz) = b(w, T^{\mathrm{BM}} z)$ for all $w, z \in X$. Hence $T = T^{\mathrm{BM}}$ by (1.2b). □

Thus, when b is a generalized duality pairing, the IPG method coincides with the method (3.13) and the discrete solution equals the X-orthogonal projection of the exact solution.

Example 3.8. It can be easily seen that the bilinear form b in (2.8) of Example 2.6 is a generalized duality pairing. Hence the analytically solved expression for T, given there in (2.10), coincides with T^{BM}.

Bibliographical notes. The interpretation of the IPG method as a residual minimization method was pointed out in Demkowicz and Gopalakrishnan (2011*b*, eq. (2.13)). The minimization of residual in dual norms was also the theme in many previous works such as Bramble, Lazarov and Pasciak (1997) and Bramble and Pasciak (2004), where the dual norm was replaced by a preconditioner action. Where the DPG methods depart from these works, as will be clear from the next section, is in the localization of the dual-norm computation through hybridization. The interpretation of the DPG method as a mixed Galerkin method has parallels in Cohen, Dahmen and Welper (2012). Theorems 3.2 and 3.4 can be seen in Gopalakrishnan (2013), Bouma, Gopalakrishnan and Harb (2014) and Demkowicz and Gopalakrishnan (2017). More recently, a substantial generalization of such theorems to a Banach space setting was achieved by Muga and van der Zee (2020). Optimal test norms were introduced in Zitelli *et al.* (2011). Generalized duality pairings and non-trivial examples of them in the context of certain trace spaces can be found in Demkowicz (2018). Proposition 3.7 connects our optimal test function idea to the old concepts of Barrett and Morton (1984). There is a considerable literature in pursuit of making their idea more computationally feasible, e.g. Barbone and Harari (2001), Celia, Russell, Ismael and Ewing (1990), Demkowicz and Oden (1986*b*,*a*), Loula, Hughes and Franca (1987) and Loula and Fernandes (2009). We instead switch course in the next section to pursue localization of the computation of our trial-to-test operator T.

4. Ideal DPG methods

In this and the next section, we define DPG methods. Throughout, the boundary value problems considered are posed on an open bounded domain $\Omega \subset \mathbb{R}^N$ with Lipschitz boundary. We further assume that Ω is partitioned into disjoint open subsets K (called elements), forming the collection Ω_h (called mesh), such that the union of $\bar{K}$ for all $K \in \Omega_h$ is $\bar{\Omega}$. We assume that the element boundaries ∂K are Lipschitz so we can apply trace theorems on them in specific applications. The shape of the elements is unimportant in this section. Let $Y(K)$ denote a Hilbert space of some space of functions on an element K, with inner product $(\cdot,\cdot)_{Y(K)}$.

Definition 4.1. An *ideal DPG method* is an IPG method (as in Definition 2.3) where Y is set to the Cartesian product of Hilbert spaces $Y(K)$, that is,

$$Y = \prod_{K\in\Omega_h} Y(K), \tag{4.1}$$

endowed with the inner product

$$(y, v)_Y = \sum_{K\in\Omega_h} (y_K, v_K)_{Y(K)} \quad \text{for all } y, v \in K, \tag{4.2}$$

where y_K denotes the $Y(K)$-component of any y in the Cartesian product (4.1).

Our interest in using such a product space for the test variable is the resulting *localization* of the trial-to-test operator T. Note that to compute a basis for the optimal test space, we must solve (2.2) to compute Tz for each z in a basis of X_h. That equation, $(Tz, y)_Y = b(z, y)$, decouples into independent equations on each element, if Y has the form (4.1). Localization of T refers to the fact that the part of Tz on an element K, namely $(Tz)_K$, can be computed, independently of other elements, by solving

$$((Tz)_K, y_K)_{Y(K)} = b(z, y_K) \quad \text{for all } y_K \in Y(K). \tag{4.3}$$

The adjective *discontinuous* in the name 'DPG' refers to the fact that test functions in Y of the form (4.1) admit discontinuous functions with no continuity constraints across element interfaces. For example, in many applications, we set Y to

$$H^1(\Omega_h) := \{v \in L^2(\Omega) \colon v|_K \in H^1(K) \text{ for all } K \in \Omega_h\},$$

which can be identified with the Cartesian product

$$H^1(\Omega_h) \equiv \prod_{K \in \Omega_h} H^1(K), \tag{4.4}$$

and contains functions that are discontinuous across element interfaces. Colloquially, we say that $H^1(\Omega_h)$ is a *broken Sobolev space*, obtained by breaking the inter-element continuity constraints of $H^1(\Omega)$. DPG methods are built using broken Sobolev spaces as test spaces.

Example 4.2 (Laplace equation: primal DPG formulation). Let $f \in L^2(\Omega)$ and u satisfy

$$-\Delta u = f \quad \text{in } \Omega, \tag{4.5a}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{4.5b}$$

The standard variational formulation for this problem finds u in $\mathring{H}^1(\Omega)$ such that

$$(\operatorname{grad} u, \operatorname{grad} v)_\Omega = (f, v)_\Omega \quad \text{for all } v \in \mathring{H}^1(\Omega). \tag{4.6}$$

A different variational formulation is obtained if we multiply (4.5a) by a possibly discontinuous test function $y \in H^1(\Omega_h)$ (defined in (4.4)) and integrate by parts, element by element. On a single element $K \in \Omega_h$, we have

$$\int_K \operatorname{grad} u \cdot \operatorname{grad} y - \int_{\partial K} (n \cdot \operatorname{grad} u) y = \int_K f y. \tag{4.7}$$

The integral over ∂K must be interpreted as a duality pairing in $H^{1/2}(\partial K)$ if u is not sufficiently regular. Recalling our notation for duality pairing in (3.5) and letting $n \cdot \operatorname{grad} u$ be an independent unknown denoted by $\hat{q}_n$, we now derive a Petrov–Galerkin formulation. To state it precisely, we use the following notation:

$$(r, s)_h = \sum_{K \in \Omega_h} (r, s)_K, \quad \langle \ell, w \rangle_h = \sum_{K \in \Omega_h} \langle \ell, w \rangle_{H^{1/2}(\partial K)}, \tag{4.8}$$

where $(\cdot,\cdot)_D$, for any domain D, denotes the $L^2(D)$-inner product and $\langle \ell, \cdot \rangle_{H^{1/2}(\partial K)}$ denotes the action of a conjugate linear functional $\ell \in H^{-1/2}(\partial K)$ on a function in $H^{1/2}(\partial K)$. Define the element-by-element trace operator

$$\mathrm{tr}_n \colon H(\mathrm{div}, \Omega) \to \prod_{K \in \Omega_h} H^{-1/2}(\partial K), \quad \mathrm{tr}_n\, r|_{\partial K} = r \cdot n|_{\partial K}. \tag{4.9}$$

Here and throughout, n denotes the *unit outward normal* vector of a domain under consideration, which is usually clear from the context, e.g. above n is the outward unit normal on each element boundary ∂K. (On an interior interface shared by two elements $K_\pm$, the n from $K_\pm$ will have opposite signs.) We endow the image of the trace map with a quotient norm,

$$\begin{aligned} H^{-1/2}(\partial\Omega_h) &= \mathrm{range}(\mathrm{tr}_n), \\ \|\hat r_n\|_{H^{-1/2}(\partial\Omega_h)} &= \inf_{q \in \mathrm{tr}_n^{-1}\{\hat r_n\}} \|q\|_{H(\mathrm{div},\Omega)}, \end{aligned} \tag{4.10}$$

where the infimum is over the preimage

$$\mathrm{tr}_n^{-1}\{\hat r_n\} = \{q \in H(\mathrm{div}, \Omega) \colon \mathrm{tr}_n(q) = \hat r_n\}.$$

Since the element boundary traces of $n \cdot \mathrm{grad}\, u$ appearing in (4.7) are in $H^{-1/2}(\partial\Omega_h)$, we now have a trial space to place the interface variables. Given a $\hat r_n \in H^{-1/2}(\partial\Omega_h)$, note that for any $v \in H^1(\Omega_h)$,

$$\langle \hat r_n, v\rangle_h = \langle n \cdot r, v\rangle_h$$

for all $r \in H(\mathrm{div}, \Omega)$ with $\mathrm{tr}_n(r) = \hat r_n$. The interior values of r are not seen by the right-hand side. When $r \cdot n$ is sufficiently smooth on each element interface, one can give an intrinsic characterization by orienting each interface; see (5.16) of Example 5.5.

With this notation, we can now give the Petrov–Galerkin formulation obtained by summing up (4.7) over all $K \in \Omega_h$. Set

$$X = \mathring{H}^1(\Omega) \times H^{-1/2}(\partial\Omega_h), \quad Y = H^1(\Omega_h).$$

Then the PG formulation finds $(u, \hat q_n) \in X$ satisfying

$$(\mathrm{grad}\, u, \mathrm{grad}\, v)_h - \langle \hat q_n, v\rangle_h = (f, v)_\Omega \quad \text{for all } v \in Y. \tag{4.11}$$

This is a 'hybrid' form of the standard formulation (4.6). Although it is different from the primal hybrid formulation of Raviart and Thomas (1977*b*), there are a number of common features, including the use of the quotient norm of the type (4.10). Since the test space $Y = H^1(\Omega_h)$ is a product space as in Definition 4.1, this formulation, provided we verify its wellposedness (which is done below), admits the construction of an ideal DPG method with localized optimal test space, known as the *primal DPG method* for the Laplace equation.

Processes that arrive at reformulations of a problem using spaces of discontinuous functions and new interface variables have traditionally been referred to as

hybridization; see e.g. Raviart and Thomas (1977*b*), Brezzi and Fortin (1991) or Cockburn and Gopalakrishnan (2004). We have used the adjective 'hybrid' in the above example following this tradition. To analyse hybrid formulations like those of Example 4.2, we formulate a result in a general setting. Let X_0, $\hat{X}$, and Y denote Hilbert spaces over $\mathbb{C}$, put $X = X_0 \times \hat{X}$, and let $b_0 \colon X_0 \times Y \to \mathbb{C}$ and $\hat{b} \colon \hat{X} \times Y \to \mathbb{C}$ denote continuous sesquilinear forms. Then

$$Y_0 = \{y \in Y \colon \hat{b}(\hat{x}, y) = 0 \text{ for all } \hat{x} \in \hat{X}\} \tag{4.12a}$$

is a closed subspace of Y. Suppose there are positive constants γ_0 and $\hat{\gamma}$ such that

$$\gamma_0 \|x\|_{X_0} \le \sup_{0 \neq y \in Y_0} \frac{|b_0(x, y)|}{\|y\|_Y} \quad \text{for all } x \in X_0, \quad \text{and} \tag{4.12b}$$

$$\hat{\gamma}\, \|\hat{x}\|_{\hat{X}} \le \sup_{0 \neq y \in Y} \frac{|\hat{b}(\hat{x}, y)|}{\|y\|_Y} \quad \text{for all } \hat{x} \in \hat{X}. \tag{4.12c}$$

Our abstraction of a hybrid formulation is based on the continuous sesquilinear form

$$b(\,(x, \hat{x}),\ y) = b_0(x, y) + \hat{b}(\hat{x}, y),$$

over $X = X_0 \times \hat{X}$ and Y. In examples, $\hat{X}$ will be a space of interface variables (on element boundaries) and Y will be a space admitting functions with no continuity constraints across element boundaries. Given an $\ell \in Y^*$, we are interested in the wellposedness of the hybrid problem to find $x \in X_0$ and $\hat{x} \in \hat{X}$ satisfying

$$b(\,(x, \hat{x}),\ y) = \ell(y) \quad \text{for all } y \in Y, \tag{4.13}$$

in relation to the problem of finding $x \in X_0$ satisfying

$$b_0(x, y) = \ell(y) \quad \text{for all } y \in Y_0. \tag{4.14}$$

Theorem 4.3 (Wellposedness of hybrid Petrov–Galerkin formulations). In the setting of (4.12), we have

$$\gamma_1 \|(x, \hat{x})\|_X \le \sup_{0 \neq y \in Y} \frac{|b(\,(x, \hat{x}),\ y)|}{\|y\|_Y}, \tag{4.15}$$

where γ_1 is given by

$$\frac{1}{\gamma_1^2} = \frac{1}{\gamma_0^2} + \frac{1}{\hat{\gamma}^2} \left(\frac{\|b_0\|}{\gamma_0} + 1 \right)^2.$$

Moreover,

$$Z = \{y \in Y \colon b(\,(x, \hat{x}), y) = 0 \text{ for all } x \in X_0 \text{ and } \hat{x} \text{ in } \hat{X}\} \quad \text{and}$$
$$Z_0 = \{y \in Y_0 \colon b_0(x, y) = 0 \text{ for all } x \in X_0\}$$

are equal:

$$Z = Z_0. \tag{4.16}$$

Consequently, if $Z_0 = \{0\}$, then (4.13) is uniquely solvable, and furthermore, the solution component x from (4.13) coincides with the solution of (4.14).

Proof. To prove (4.15), noting that

$$\|(x,\hat{x})\|_X^2 = \|x\|_{X_0}^2 + \|\hat{x}\|_{\hat{X}}^2,$$

we start by bounding $\|x\|_{X_0}$ as follows:

$$\begin{aligned}
\gamma_0\|x\|_{X_0} &\le \sup_{0\neq y_0\in Y_0} \frac{|b_0(x,y)|}{\|y\|_Y} && \text{by (4.12b)}\\
&\le \sup_{0\neq y_0\in Y_0} \frac{|b_0(x,y)+\hat{b}(\hat{x},y)|}{\|y\|_Y} && \text{by (4.12a)}\\
&\le \sup_{0\neq y\in Y} \frac{|b((x,\hat{x}),\,y)|}{\|y\|_Y} && \text{as } Y_0\subseteq Y.
\end{aligned}$$

Next, to bound $\|\hat{x}\|_{\hat{X}}$, using (4.12c),

$$\begin{aligned}
\hat{\gamma}\,\|\hat{x}\|_{\hat{X}} &\le \sup_{0\neq y\in Y} \frac{|\hat{b}(\hat{x},y)|}{\|y\|_Y} = \sup_{0\neq y\in Y} \frac{|b((x,\hat{x}),\,y)-b_0(x,y)|}{\|y\|_Y}\\
&\le \|b_0\|\,\|x\|_{X_0} + \sup_{0\neq y\in Y} \frac{|b((x,\hat{x}),\,y)|}{\|y\|_Y}.
\end{aligned}$$

Combining these bounds for $\|x\|_{X_0}$ and $\|\hat{x}\|_{\hat{X}}$, we obtain (4.15).

To prove the remaining claims, note that since we may choose x and $\hat{x}$ independently in the definition of Z, a $y \in Y$ is in Z if and only if $b_0(x,y) = 0$ for all $x \in X_0$ and $\hat{b}(\hat{x},y) = 0$ for all $\hat{x} \in \hat{X}$. The latter holds if and only if $y \in Y_0$ due to (4.12a). Hence (4.16) follows. The unique solvability of both (4.13) and (4.14) then follows from the equivalence of (1.2) and (1.1). Finally, restricting the test space in (4.13) to Y_0 and using (4.12a), we find that its solution component x must also solve (4.14). □

In particular examples, to apply the theorem, the main work is in verifying its assumptions. We shall now see examples of how to do so. A one-dimensional example is provided by a hybrid version of the formulation of Example 2.6 (using a natural broken space). Its analysis can be found in Demkowicz and Gopalakrishnan (2011*b*, § III). Here we proceed directly to the more interesting multi-dimensional examples of Laplace and Maxwell equations.

Example 4.4 (Laplace equation: wellposedness of primal DPG formulation). Continuing Example 4.2, we fit it into the framework above by setting

$$X_0 = \mathring{H}^1(\Omega), \qquad Y_0 = \mathring{H}^1(\Omega), \tag{4.17a}$$

$$\hat{X} = H^{-1/2}(\partial\Omega_h), \qquad Y = H^1(\Omega_h), \tag{4.17b}$$

$$b_0(u,y) = (\operatorname{grad} u, \operatorname{grad} y)_h, \quad \hat{b}(\hat{q}_n, y) = \langle \hat{q}_n, y\rangle_h. \tag{4.17c}$$

To verify (4.12a), letting

$$\tilde{Y} = \{y \in H^1(\Omega_h) \colon \langle \hat{q}_n, y\rangle_h = 0 \text{ for all } \hat{q}_n \in H^{-1/2}(\partial\Omega_h)\},$$

we must prove that $\mathring{H}^1(\Omega) = \tilde{Y}$. Recall that $\hat{q}_n$ above is always of the form $n \cdot q$ for some $q \in H(\mathrm{div}, \Omega)$. Let $y \in \mathring{H}^1(\Omega)$. Its zero trace implies that $\langle y, q \cdot n\rangle_{H^{1/2}(\partial\Omega)} = 0$ for any $q \in H(\mathrm{div}, \Omega)$. Hence, two integrations by parts, one element by element and the other over Ω, give

$$\langle y, q \cdot n\rangle_h = (y, \mathrm{div}\, q)_\Omega + (\mathrm{grad}\, y, q)_\Omega = \langle y, q \cdot n\rangle_{H^{1/2}(\partial\Omega)}, \tag{4.18}$$

and since the last term vanishes, $y \in \tilde{Y}$, i.e. $\mathring{H}^1(\Omega) \subseteq \tilde{Y}$.

For the reverse inclusion, consider a $y \in \tilde{Y}$. Note that the distributional gradient of a $y \in H^1(\Omega_h)$ acting on a ϕ in the Schwartz test space $\mathcal{D}(\Omega)^N$ satisfies

$$(\mathrm{grad}\, y)(\phi) = -(y, \mathrm{div}\, \phi)_\Omega = (\mathrm{grad}\, y, \phi)_h - \langle y, n \cdot \phi\rangle_h$$

where we have integrated by parts, element by element. The last term vanishes by the given condition on y. Hence $\mathrm{grad}\, y \in L^2(\Omega)^N$. Now integrating by parts again, but this time over Ω, we find, by (4.18), that $\langle y, q \cdot n\rangle_{H^{1/2}(\partial\Omega)} = \langle y, q \cdot n\rangle_h = 0$ for all $q \in H(\mathrm{div}, \Omega)$. Hence $y|_{\partial\Omega} = 0$ and $y \in \mathring{H}^1(\Omega)$. Thus

$$\mathring{H}^1(\Omega) = \tilde{Y} \tag{4.19}$$

and (4.12a) holds.

Condition (4.12b) obviously holds, since the $b_0(\cdot,\cdot)$ set in (4.17) is coercive on Y_0 by the Friedrichs inequality. It only remains to verify (4.12c). To do so, given a $\hat{q}_n \in H^{-1/2}(\partial\Omega_h)$, consider $q \in H(\mathrm{div}, K)$ and $w \in H^1(K)$ solving

$$-\mathrm{grad}(\mathrm{div}\, q) + q = 0 \quad \text{in } K, \quad n \cdot q = \hat{q}_n \quad \text{on } \partial K, \tag{4.20}$$

$$-\mathrm{div}(\mathrm{grad}\, w) + w = 0 \quad \text{in } K, \quad \frac{\partial w}{\partial n} = \hat{q}_n \quad \text{on } \partial K. \tag{4.21}$$

The boundary value problems (4.21) and (4.20) are equivalent in the sense that w solves (4.21) if and only if $q = \mathrm{grad}\, w$ solves (4.20) and moreover $\|w\|_{H^1(K)} = \|q\|_{H(\mathrm{div},K)}$. It is also obvious that from among all $H(\mathrm{div}, K)$-extensions of $\hat{q}_n$, the solution of (4.20) has the minimal $H(\mathrm{div}, K)$-norm, so

$$\begin{aligned} \|\hat{q}_n\|_{H^{-1/2}(\partial K)} &= \|q\|_{H(\mathrm{div},K)} = \|w\|_{H^1(K)} \\ &= \sup_{0\neq v\in H^1(K)} \frac{|(\mathrm{grad}\, w, \mathrm{grad}\, v)_K + (w, v)_K|}{\|v\|_{H^1(K)}} \\ &= \sup_{0\neq v\in H^1(K)} \frac{|\langle \hat{q}_n, v\rangle_{H^{1/2}(\partial K)}|}{\|v\|_{H^1(K)}}, \end{aligned} \tag{4.22}$$

where we used the variational form of (4.21) in the last step. Squaring and summing over all $K \in \Omega_h$, we find that (4.12c) holds with $\hat{\gamma} = 1$. (More identities similar to (4.22) appear in Theorem 4.6 below.)

Having verified the assumptions, Theorem 4.3 now gives the inf-sup condition

$$\|(u,\hat{q}_n)\|_{\mathring{H}^1(\Omega)\times H^{-1/2}(\partial\Omega_h)} \lesssim \sup_{0\neq y\in H^1(\Omega_h)} \frac{(\operatorname{grad} u,\operatorname{grad} y)_h + \langle \hat{q}_n, y\rangle_h}{\|y\|_{H^1(\Omega_h)}}, \tag{4.23}$$

thus proving the wellposedness of the primal DPG formulation for the Laplace equation.

Example 4.5 (Maxwell equations). We now develop and analyse a primal DPG method for the cavity problem in electromagnetics. Let the cavity Ω be an open bounded contractible domain in $\mathbb{R}^3$, on the boundary of which the so-called perfect electric conducting boundary condition is placed. Assuming that all time variations are harmonic of frequency $\omega > 0$, Maxwell equations in the cavity are

$$-\hat{\imath}\omega\mu H + \operatorname{curl} E = 0 \quad \text{in } \Omega, \tag{4.24a}$$

$$-\hat{\imath}\omega\epsilon E - \operatorname{curl} H = -J \quad \text{in } \Omega, \tag{4.24b}$$

$$n\times E = 0 \quad \text{on } \partial\Omega. \tag{4.24c}$$

The functions $E, H, J\colon \Omega \to \mathbb{C}^3$ represent electric field, magnetic field and imposed current, respectively, and $\hat{\imath}$ denotes the imaginary unit. We assume that the electromagnetic material properties ϵ and μ are bounded uniformly positive functions on Ω. The number $\omega > 0$ denotes a fixed wavenumber. Eliminating H from (4.24a) and (4.24b), we obtain the second-order (non-elliptic) equation

$$\operatorname{curl}\mu^{-1}\operatorname{curl} E - \omega^2\epsilon E = f, \tag{4.25}$$

where $f = \hat{\imath}\omega J$. Let

$$H(\operatorname{curl},\Omega) = \{F\in L^2(\Omega)^3\colon \operatorname{curl} F\in L^2(\Omega)^3\}$$

and let $\mathring{H}(\operatorname{curl},\Omega)$ denote the subspace of vector fields in $H(\operatorname{curl},\Omega)$ with zero tangential trace on $\partial\Omega$. A standard variational formulation for this problem is obtained by multiplying (4.25) by a test function $F\in\mathring{H}(\operatorname{curl},\Omega)$, integrating by parts and using the boundary condition (4.24c): find $E\in\mathring{H}(\operatorname{curl},\Omega)$ satisfying

$$(\mu^{-1}\operatorname{curl} E,\operatorname{curl} F)_\Omega - \omega^2(\epsilon E,F)_\Omega = \langle f,F\rangle \tag{4.26}$$

for any given $f\in\mathring{H}(\operatorname{curl},\Omega)'$. It is well known (see Monk 2003) that (4.26) has a unique solution for every $f\in\mathring{H}(\operatorname{curl},\Omega)'$ whenever ω is not a resonance of the cavity Ω, an assumption we place throughout this example.

The primal DPG method for the electric cavity problem is obtained by multiplying (4.25) by a test function F in the 'broken' space

$$H(\operatorname{curl},\Omega_h) = \prod_{K\in\Omega_h} H(\operatorname{curl},K)$$

and integrating by parts, element by element. On a single element $K\in\Omega_h$, we get

$$(\mu^{-1}\operatorname{curl} E,\operatorname{curl} F)_K + (n\times\mu^{-1}\operatorname{curl} E,F)_{\partial K} - \omega^2(\varepsilon E,F)_K = (f,F)_K. \tag{4.27}$$

To set the element boundary term in the right space, a space akin to (4.10), let us recall a few pertinent results on tangential traces on Lipschitz boundaries.

The tangential trace maps

$$E \mapsto \mathrm{tr}^K_{n\times} E := (n \times E)|_{\partial K}, \quad E \mapsto \mathrm{tr}^K_{\tau} E := n \times (E \times n)|_{\partial K} \equiv E_{\tau}|_{\partial K},$$

both well-defined for smooth vectors fields E on any mesh element $K \in \Omega_h$, can be extended to continuous linear maps

$$\mathrm{tr}^K_{n\times} : H(\mathrm{curl}, K) \to H^{-1/2}(\mathrm{div_F}, \partial K), \quad \mathrm{tr}^K_{\tau} : H(\mathrm{curl}, K) \to H^{-1/2}(\mathrm{curl_F}, \partial K)$$

by the work of Buffa, Costabel and Sheen (2002), which contains the definitions of the codomain spaces and the surface derivatives ($\mathrm{div_F}$ and $\mathrm{curl_F}$) above. Moreover, their results imply that the integration-by-parts formula

$$(\mathrm{curl}\, E, F)_K - (E, \mathrm{curl}\, F)_K = (n \times E, F)_{\partial K}$$

for smooth vector fields E and F on K can be extended to $E, F \in H(\mathrm{curl}, \Omega)$, with the understanding that the right-hand side above becomes a duality pairing $\langle \mathrm{tr}^K_{n\times} E, \mathrm{tr}^K_{\tau} F\rangle_{H^{-1/2}(\mathrm{curl_F}, \partial K)}$ between $H^{-1/2}(\mathrm{div_F}, \partial K)$ and $H^{-1/2}(\mathrm{curl_F}, \partial K)$. Re-using the notation of $(\cdot,\cdot)_h$ and $\langle\cdot,\cdot\rangle_h$ in (4.8) by extending inner products to vector fields in the obvious way and letting

$$\langle n \times E, F\rangle_h = \sum_{K\in\Omega_h} \langle \mathrm{tr}^K_{n\times} E, \mathrm{tr}^K_{\tau} F\rangle_{H^{-1/2}(\mathrm{curl_F}, \partial K)},$$

we sum (4.27) over all $K \in \Omega_h$ to obtain

$$(\mu^{-1}\,\mathrm{curl}\, E, \mathrm{curl}\, F)_h + \langle n \times \mu^{-1}\,\mathrm{curl}\, E, F\rangle_h - \omega^2(\varepsilon E, F)_h = (f, F)_h. \tag{4.28}$$

To set the interface term in the right space, we need some more machinery.

Applying the trace operators $\mathrm{tr}^K_{n\times}$ and tr^K_{τ} element by element, we define

$$\begin{aligned} &\mathrm{tr}_{n\times} : H(\mathrm{curl}, \Omega) \to \prod_{K\in\Omega_h} H^{-1/2}(\mathrm{div_F}, \partial K), \\ &(\mathrm{tr}_{n\times} E)|_{\partial K} = \mathrm{tr}^K_{n\times} E \equiv (n \times E)|_{\partial K}, \end{aligned} \tag{4.29}$$

and

$$\begin{aligned} &\mathrm{tr}_{\tau} : H(\mathrm{curl}, \Omega) \to \prod_{K\in\Omega_h} H^{-1/2}(\mathrm{curl_F}, \partial K), \\ &(\mathrm{tr}_{\tau} E)|_{\partial K} = \mathrm{tr}^K_{\tau} E \equiv (n \times (E \times n))|_{\partial K}. \end{aligned} \tag{4.30}$$

Next, analogous to what we did in (4.10), we define interface spaces as the ranges of the above trace operators, endowed with a quotient norm, namely

$$\begin{aligned} H^{-1/2}(\mathrm{div_F}, \partial\Omega_h) &:= \mathrm{range}(\mathrm{tr}_{n\times}), \\ \|n \times \hat{E}\|_{H^{-1/2}(\mathrm{div_F}, \partial\Omega_h)} &:= \inf_{E\in \mathrm{tr}^{-1}_{n\times}\{n\times\hat{E}\}} \|E\|_{H(\mathrm{curl},\Omega)}, \end{aligned}$$

and

$$\begin{aligned} H^{-1/2}(\operatorname{curl}_{\mathrm{F}}, \partial\Omega_h) &:= \operatorname{range}(\operatorname{tr}_{\tau}), \\ \|\hat{E}_{\tau}\|_{H^{-1/2}(\operatorname{curl}_{\mathrm{F}}, \partial\Omega_h)} &:= \inf_{E \in \operatorname{tr}_{\tau}^{-1}\{E_{\tau}\}} \|E\|_{H(\operatorname{curl}, \Omega)}, \end{aligned} \tag{4.31}$$

where the preimage sets are

$$\operatorname{tr}_{n\times}^{-1}\{n \times \hat{E}\} = \{E \in H(\operatorname{curl}, \Omega) \colon \operatorname{tr}_{n\times}^{K} E = (n \times \hat{E})|_{\partial K} \text{ on each } K \in \Omega_h\}$$

and

$$\operatorname{tr}_{\tau}^{-1}\{\hat{E}_{\tau}\} = \{E \in H(\operatorname{curl}, \Omega) \colon \operatorname{tr}_{\tau}^{K} E = \hat{E}_{\tau}|_{\partial K} \text{ on each } K \in \Omega_h\}.$$

Returning to (4.28), we now set $n\times\hat{H} = (\hat{\imath}\omega)^{-1} n\times\mu^{-1} \operatorname{curl} E$ to be an independent unknown on element boundaries, to be found in $H^{-1/2}(\operatorname{div}_{\mathrm{F}}, \Omega_h)$. Then (4.27) leads to the variational problem (4.13) with the following spaces and forms:

$$X_0 = \mathring{H}(\operatorname{curl}, \Omega), \qquad Y = H(\operatorname{curl}, \Omega_h), \tag{4.32a}$$

$$\hat{X} = H^{-1/2}(\operatorname{div}_{\mathrm{F}}, \partial\Omega_h), \qquad Y_0 = \mathring{H}(\operatorname{curl}, \Omega), \tag{4.32b}$$

$$b_0(E, F) = (\mu^{-1} \operatorname{curl} E, \operatorname{curl} F)_h - \omega^2(\varepsilon E, F)_h, \tag{4.32c}$$

$$b(n \times \hat{H}, F) = \hat{\imath}\omega\langle n \times \hat{H}, F\rangle_h. \tag{4.32d}$$

This is the *primal DPG formulation for the Maxwell cavity* problem.

To prove that this formulation is wellposed, we verify the conditions of Theorem 4.3. It is easy to verify (4.12a) by extending the same technique we used to prove it in Example 4.4. Condition (4.12b) follows from the previously mentioned unique solvability of (4.26) and the equivalence of (1.1) and (1.2). Finally, condition (4.12c) follows from (4.35c) of the next theorem (Theorem 4.6) below. Hence Theorem 4.3 gives wellposedness of the formulation (4.32).

The next result shows that the argument we used to prove (4.22) in Example 4.4 can be generalized to get other similar identities for quotient norms. Define the broken version of $H(\operatorname{div}, \Omega)$ by

$$H(\operatorname{div}, \Omega_h) = \prod_{K \in \Omega_h} H(\operatorname{div}, K). \tag{4.33}$$

Complementing already defined trace operators tr_n, $\operatorname{tr}_{n\times}$ and tr_{τ} (in (4.9), (4.29) and (4.30), respectively), define standard H^1 trace operator, applied elementwise, by

$$\operatorname{tr} \colon H^1(\Omega) \to \prod_{K \in \Omega_h} H^{1/2}(\partial K), \quad (\operatorname{tr} u)|_{\partial K} = u|_{\partial K}, \tag{4.34}$$

and let

$$H^{1/2}(\partial\Omega_h) := \operatorname{range}(\operatorname{tr}), \quad \|\hat{u}\|_{H^{1/2}(\partial\Omega_h)} := \inf_{u \in \operatorname{tr}^{-1}\{\hat{u}\}} \|u\|_{H^1(\Omega)},$$

where the quotient norm is a standard norm obtained by a 'minimal energy extension' as an infimum of the norm of all extensions of $\hat{u}$ in the preimage set

$\mathrm{tr}^{-1}\{\hat{u}\} = \{u \in H^1(\Omega)\colon u|_{\partial K} = \hat{u}|_{\partial K}\}$. The identity (4.35b) below shows that this infimum equals a supremum; in fact all identities of (4.35) are of a similar 'inf = sup' type.

Theorem 4.6 (Interface duality). The following identities hold for any $\hat{\sigma}_n$ in $H^{-1/2}(\partial\Omega_h)$, $\hat{u}$ in $H^{1/2}(\partial\Omega_h)$, $n\times\hat{E}$ in $H^{-1/2}(\mathrm{div_F}, \partial K)$, and $\hat{F}_\tau$ in $H^{-1/2}(\mathrm{curl}, \partial K)$:

$$\|\hat{\sigma}_n\|_{H^{-1/2}(\partial\Omega_h)} = \sup_{0\neq u\in H^1(\Omega_h)} \frac{|\langle\hat{\sigma}_n, u\rangle_h|}{\|u\|_{H^1(\Omega_h)}}, \tag{4.35a}$$

$$\|\hat{u}\|_{H^{1/2}(\partial\Omega_h)} = \sup_{0\neq\sigma\in H(\mathrm{div},\Omega_h)} \frac{|\langle n\cdot\sigma, \hat{u}\rangle_h|}{\|\sigma\|_{H(\mathrm{div},\Omega_h)}}, \tag{4.35b}$$

$$\|n\times\hat{E}\|_{H^{-1/2}(\mathrm{div_F},\partial\Omega_h)} = \sup_{0\neq F\in H(\mathrm{curl},\Omega_h)} \frac{|\langle n\times\hat{E}, F\rangle_h|}{\|F\|_{H(\mathrm{curl},\Omega_h)}}, \tag{4.35c}$$

$$\|\hat{F}_\tau\|_{H^{-1/2}(\mathrm{curl},\partial\Omega_h)} = \sup_{0\neq E\in H(\mathrm{curl},\Omega_h)} \frac{|\langle n\times E, \hat{F}_\tau\rangle_h|}{\|E\|_{H(\mathrm{curl},\Omega_h)}}. \tag{4.35d}$$

Furthermore,

(a) $v \in \mathring{H}^1(\Omega)$ if and only if $\langle\hat{\sigma}_n, v\rangle_h = 0$ for all $\hat{\sigma}_n \in H^{-1/2}(\partial\Omega_h)$,

(b) $\tau \in \mathring{H}(\mathrm{div}, \Omega)$ if and only if $\langle\tau\cdot n, \hat{u}\rangle_h = 0$ for all $\hat{u} \in H^{1/2}(\partial\Omega_h)$, and

(c) $F \in \mathring{H}(\mathrm{curl}, \Omega)$ if and only if $\langle n\times\hat{E}, F\rangle_h = 0$ for all $n\times\hat{E} \in H^{-1/2}(\mathrm{div_F}, \Omega_h)$.

Proof. The first equality was already proved in (4.22) and the argument is similar for all identities of (4.35). So we outline the proof of only one more, namely (4.35c).

Given $n\times\hat{E}$ in $H^{-1/2}(\mathrm{div_F}, \partial\Omega_h)$, its norm equals the norm of the following minimum energy extension $E \in H(\mathrm{curl}, K)$ satisfying

$$\mathrm{curl}(\mathrm{curl}\, E) + E = 0 \quad \text{in } K, \quad n\times E = n\times\hat{E} \quad \text{on } \partial K. \tag{4.36}$$

We compare this with the inverse of a Riesz map applied to a functional generated by $n\times\hat{E}$, namely

$$\mathrm{curl}(\mathrm{curl}\, F) + F = 0 \quad \text{in } K, \quad n\times(\mathrm{curl}\, F) = n\times\hat{E} \quad \text{on } \partial K. \tag{4.37}$$

Note that F solves (4.37) if and only if $E = \mathrm{curl}\, F$ solves (4.36). Moreover, (4.37) implies that $\mathrm{curl}\, E = -F$. Therefore $\|E\|_{H(\mathrm{curl},K)} = \|F\|_{H(\mathrm{curl},K)}$ and

$$\begin{aligned}\|n\times\hat{E}\|_{H^{-1/2}(\mathrm{div_F},\partial K)} &= \|E\|_{H(\mathrm{curl},K)} = \|F\|_{H(\mathrm{curl},K)} \\ &= \sup_{0\neq G\in H(\mathrm{curl},K)} \frac{|(\mathrm{curl}\, F, \mathrm{curl}\, G)_K + (F, G)_K|}{\|G\|_{H(\mathrm{curl},K)}} \\ &= \sup_{0\neq G\in H(\mathrm{curl},K)} \frac{|\langle n\times\hat{E}, G_\tau\rangle_{H^{-1/2}(\mathrm{curl_F},\partial K)}|}{\|G\|_{H(\mathrm{curl},K)}}.\end{aligned} \tag{4.38}$$

Summing over all elements, and using

$$\left(\sup_{0\neq G\in H(\mathrm{curl},\Omega_h)} \frac{|\langle n\times \hat{E}, G\rangle_h|}{\|G\|_{H(\mathrm{curl},\Omega_h)}} \right)^2 = \sum_{K\in\Omega_h} \left(\sup_{0\neq G\in H(\mathrm{curl},K)} \frac{|\langle n\times \hat{E}, G\rangle_{H^{-1/2}(\mathrm{curl_F},\partial K)}|}{\|G\|_{V(K)}} \right)^2,$$

the identity (4.35c) is proved.

Proofs of all items (a)–(c) are similar to the previously detailed proof of (4.19), so we omit them. □

It is interesting to observe that the norm of the dual space $H^{-1/2}(\mathrm{curl_F},\partial K)^*$ occurs in the above proof implicitly. Indeed, since the $H^{-1/2}(\mathrm{curl_F},\partial K)$-norm in (4.31) is the infimum of extension norms, its dual norm equals

$$\begin{aligned} \|n\times\hat{E}\|_{H^{-1/2}(\mathrm{curl_F},\partial K)^*} &= \sup_{0\neq G_\tau\in H^{-1/2}(\mathrm{curl_F},\partial K)} \frac{|\langle n\times\hat{E}, G_\tau\rangle_{H^{-1/2}(\mathrm{curl_F},\partial K)}|}{\|G_\tau\|_{H^{-1/2}(\mathrm{curl_F},\partial K)}} \\ &= \sup_{0\neq G\in H(\mathrm{curl},K)} \frac{|\langle n\times\hat{E}, G_\tau\rangle_{H^{-1/2}(\mathrm{curl_F},\partial K)}|}{\|G\|_{H(\mathrm{curl},K)}}, \end{aligned}$$

which is the supremum in (4.38). Thus the short argument in the previous proof also shows that

$$\|n\times\hat{E}\|_{H^{-1/2}(\mathrm{div_F},\partial K)} = \|n\times\hat{E}\|_{H^{-1/2}(\mathrm{curl_F},\partial K)^*}, \tag{4.39}$$

that is, the norms of $H^{-1/2}(\mathrm{div_F},\partial K)$ and $H^{-1/2}(\mathrm{curl_F},\partial K)^*$ are equal.

Bibliographical notes. An alternative and longer proof of the wellposedness of the DPG formulation for the Laplace equation in Example 4.4 first appeared in Demkowicz and Gopalakrishnan (2013), using techniques developed for a slightly different formulation for the same equation from Demkowicz and Gopalakrishnan (2011*a*). There, the adjoint inf-sup condition

$$\|y\|_{H^1(\Omega_h)} \lesssim \sup_{0\neq u\in \mathring{H}^1(\Omega),\, \hat{q}_n\in H^{-1/2}(\partial\Omega_h)} \frac{(\mathrm{grad}\, u, \mathrm{grad}\, y)_h + \langle \hat{q}_n, y\rangle_h}{\|(u,\hat{q}_n)\|_{\mathring{H}^1(\Omega)\times H^{-1/2}(\partial\Omega_h)}}$$

is proved instead of (4.23). Proving the adjoint inf-sup condition is an alternative path to wellposedness, in view of the equivalence between (1.2) and (1.3). The shorter approach we presented in Example 4.4 is facilitated by the simple result of Theorem 4.3. Similar results can be found in early works such as Brezzi and Fortin (1991, p. 40), and even in other recent works, e.g. Garg, Prudhomme, van der Zee and Carey (2014). Our discussions of Theorems 4.3, 4.6 and the Maxwell case in Example 4.5 are drawn from Carstensen, Demkowicz and Gopalakrishnan (2016), where further details can be found; see also Demkowicz (2018, § 4.2). Further properties of the norms in (4.39), including intrinsic characterizations, can be found in Buffa *et al.* (2002).

5. Practical DPG methods

Even if the construction of the test space is localized in the ideal DPG method of the previous section, a practical issue still remains. The computation of $(Tz)_K$ in (4.3) may require solving an infinite-dimensional problem on an element K if $Y(K)$ is of infinite dimension. To obtain a practical method, we must trade $Y(K)$ for a finite-dimensional space. This section does so, provides a key general tool (Theorem 5.2) involving Fortin operators to analyse the effect of this replacement on stability and error estimates, and details error analyses of the practical DPG methods for Laplace and Maxwell equations. Multiple subsections on Fortin operators show various techniques to construct Fortin operators that satisfy certain moment conditions needed for DPG analyses.

We start by introducing Y^r, a finite-dimensional subspace of Y, where r is related to its finite dimension. To retain the localization, when Y is a Cartesian product as in (4.1), the subspace Y^r is assumed to be of a similar form,

$$Y^r = \prod_{K\in\Omega_h} Y^r(K), \quad Y^r(K) \subseteq Y(K). \tag{5.1}$$

In analogy with (2.2), let $T^r : X \to Y^r$ be defined by

$$(T^r w, y)_Y = b(w, y) \quad \text{for all } y \in Y^r. \tag{5.2}$$

Then $(T^r w)_K \in Y^r(K)$, the component of $T^r w$ in element K, is computed locally within K by

$$((T^r w)_K, y_K)_{Y(K)} = b(w, y_K) \quad \text{for all } y_K \in Y^r(K). \tag{5.3}$$

A practical method is obtained using

$$Y_h^r := T^r(X_h)$$

in place of the exactly optimal test space Y_h^{opt} of (2.1).

Definition 5.1. A *(practical) DPG method* is a method that finds $x_h \in X_h$ satisfying

$$b(x_h, y) = \ell(y) \quad \text{for all } y \in Y_h^r, \tag{5.4}$$

where Y_h^r is computed locally using T^r by (5.3) in a finite-dimensional Y^r of the product form (5.1).

5.1. A general DPG convergence theorem

In general, $T^r \neq T$, and the test space in the practical DPG method, $Y_h^r \neq Y_h^{\text{opt}}$, is only an inexact version of the optimal test space. Hence, to obtain an error estimate for the practical DPG method, we cannot rely on the prior theory for the ideal DPG method. However, imposing an extra condition (see (5.5) below) gives a simple error analysis, as shown next. The condition involves an operator Π, which we

shall refer to as a 'Fortin operator', based on similar such operators in the study of mixed methods (Brezzi and Fortin 1991).

Theorem 5.2 (Fortin operator gives DPG convergence). Suppose (1.3) holds, $X_h \subset X$ and $Y^r \subset Y$. Assume that there is a bounded linear operator $\Pi : Y \to Y^r$, of operator norm $\|\Pi\|$, such that for all $w_h \in X_h$ and all $v \in Y$,

$$b(w_h, v - \Pi v) = 0. \tag{5.5}$$

Then the DPG method (5.4) is uniquely solvable for x_h and

$$\|x - x_h\|_X \leq \frac{\|b\| \, \|\Pi\|}{\gamma} \inf_{z_h \in X_h} \|x - z_h\|_X, \tag{5.6}$$

where x is the unique exact solution of (1.1).

Proof. The proof proceeds by verifying the assumptions of Theorem 1.1. Let us first prove that (5.5) implies that

$$T^r : X_h \to Y^r \text{ is injective.} \tag{5.7}$$

Indeed, if $T^r w_h = 0$ for some $w_h \in X_h$, then by (5.2), $b(w_h, y^r) = 0$ for all $y^r \in Y^r$, which implies that $b(w_h, \Pi y) = 0$ for all $y \in Y$. But (5.5) then shows that $b(w_h, y) = 0$ for all $y \in Y$, so by (1.3), $w_h = 0$. Therefore we have verified that

$$\dim(Y_h^r) = \dim(X_h).$$

To verify the inf-sup condition, fix an arbitrary $z_h \in X_h$, and let

$$s_0 = \sup_{0 \neq y \in Y} \frac{|b(z_h, y)|}{\|y\|_Y}, \quad s_1 = \sup_{0 \neq y \in Y^r} \frac{|b(z_h, y^r)|}{\|y^r\|_Y}, \quad s_2 = \sup_{0 \neq y \in Y_h^r} \frac{|b(z_h, y_h^r)|}{\|y_h^r\|_Y}.$$

The result will follow from Theorem 1.1 once we prove the discrete inf-sup condition

$$\gamma \, \|\Pi\|^{-1} \|z_h\|_X \leq s_2. \tag{5.8}$$

We proceed to bound $\|z_h\|_X$ using s_0, then s_1, and finally s_2. Since (1.3) is equivalent to (1.2), the inf-sup condition $\gamma \|z_h\|_X \leq s_0$ holds. Hence (5.5) implies

$$\begin{aligned} \gamma \|z_h\|_X &\leq \sup_{0 \neq y \in Y} \frac{|b(z_h, y)|}{\|y\|_Y} = \sup_{0 \neq y \in Y} \frac{|b(z_h, \Pi y)|}{\|y\|_Y} \\ &\leq \sup_{0 \neq y \in Y} \frac{|b(z_h, \Pi y)|}{\|\Pi\|^{-1} \|\Pi y\|_Y} \leq \sup_{0 \neq y \in Y^r} \frac{|b(z_h, y^r)|}{\|\Pi\|^{-1} \|y^r\|_Y}, \end{aligned}$$

that is, $\gamma \, \|\Pi\|^{-1} \|z_h\|_X \leq s_1$. To finish the proof of (5.8), it suffices to show that $s_1 \leq s_2$. The argument of Proposition 2.1 shows that the supremum s_1 is attained at $T^r z_h$, so

$$s_1 = \frac{(T^r z_h, T^r z_h)_Y}{\|T^r z_h\|_Y} \leq \sup_{0 \neq y_h^r \in Y_h^r} \frac{(T^r z_h, y_h^r)_Y}{\|y_h^r\|_Y} = s_2.$$

This shows (5.8) and finishes the proof. □

Example 5.3 (Test spaces containing the optimal test functions). Consider the DPG method obtained using

$$Y^r \supseteq Y_h^{\text{opt}}. \tag{5.9}$$

Then, setting Π to Π_{Y^r}, the Y-orthogonal projection into Y^r, observe that for any $z_h \in X_h$ and any $y \in Y$,

$$\begin{aligned} b(z_h, y - \Pi_{Y^r} y) &= (Tz_h, y - \Pi_{Y^r} y)_Y && \text{by (2.2)} \\ &= 0 && \text{since } Tz_h \in Y^r. \end{aligned}$$

Hence Theorem 5.2 applies, and moreover, in (5.6) we may set $\|\Pi\| = 1$ since Π is an orthogonal projection. Thus the DPG solution, in this case, satisfies exactly the same error estimate (2.4) as the ideal Petrov–Galerkin method.

More can be said by noting that (5.9) implies

$$T^r w = Tw, \quad w \in X. \tag{5.10}$$

Indeed, restricting the test functions y in defining equation (2.2) for Tw to $y = y^r \in Y^r$, we find that $Tw \in Y^r$ also satisfies equation (5.2) defining $T^r w$, thus proving the equality of Tw and $T^r w$ stated in (5.10). It then immediately implies that the solution of the IPG method with the optimal test space Y_h^{opt} and the solution of the DPG method with a test space Y^r satisfying (5.9) must coincide.

Since (5.9) seldom holds in practical multi-dimensional examples, typical applications of Theorem 5.2 involve more complex Fortin operators Π, as we shall see next. Nonetheless, this discussion shows that enlarging the test space beyond the optimal test space does not degrade stability or error estimates.

Bibliographical notes. The result of Theorem 5.2 and several of its applications, including Fortin operators useful for analysing DPG methods for the Poisson equation (see Example 5.5 below) and the elasticity equation (not discussed in this review), were presented first in Gopalakrishnan and Qiu (2014). Fortin operators for DPG analysis of plate-bending problems were given in Führer and Heuer (2019).

5.2. *First example of a non-trivial DPG Fortin operator*

As previously mentioned, DPG methods use test spaces of the product form (4.1), usually obtained as broken Sobolev spaces. Construction of Fortin operators on broken Sobolev spaces can therefore be done focusing only on one element. We proceed to construct a local Fortin operator on the broken H^1 space and use it to analyse the primal DPG method for the Laplace example.

From now on, we assume that the mesh Ω_h is a geometrically conforming finite element mesh of *simplicial* elements. Let $\triangle_j K$ denote the set of j-dimensional subsimplices of an N-simplex K. The set of mesh *facets*, denoted by $\mathcal{F}_h$, is the union of $\triangle_{N-1} K$ for all $K \in \Omega_h$. Let $P_p(D)$ denote the space of polynomials of

degree at most p restricted to a domain D, and let

$$R_p(D) = P_{p-1}(D)^N + xP_{p-1}(D) \tag{5.11}$$

denote the Raviart–Thomas element (Raviart and Thomas 1977*a*) which generates the finite element space

$$R_p^h = \{r \in H(\mathrm{div}, \Omega) \colon r|_K \in R_p(K)\}. \tag{5.12}$$

We shall also use the space

$$P_p(\Omega_h) = \prod_{K \in \Omega_h} P_p(K),$$

often used in discontinuous Galerkin (DG) methods. Let $h_K = \mathrm{diam}(K)$. We write

$$A \lesssim B$$

to indicate that the inequality $A \leq CB$ holds with some constant C (whose value at different occurrences may differ) independent of h_K but possibly dependent on the shape regularity of K and the polynomial degree p. We will prove the following theorem shortly after indicating how it is applied in a DPG method.

Theorem 5.4 (A local Fortin operator for H^1 in N dimensions). Let $v \in H^1(K)$ on an N-simplex K and let $r = p + N$. There is a locally constructible continuous linear operator $\Pi_r^{\mathrm{grad}} \colon H^1(K) \to P_r(K)$ satisfying

$$(\Pi_r^{\mathrm{grad}} v - v, q)_K = 0 \qquad \text{for all } q \in P_{p-1}(K), \tag{5.13a}$$

$$(\Pi_r^{\mathrm{grad}} v - v, \mu)_F = 0 \qquad \text{for all } \mu \in P_p(F),\ F \in \triangle_{N-1} K, \tag{5.13b}$$

$$\|\Pi_r^{\mathrm{grad}} v\|_{H^1(K)} \lesssim \|v\|_{H^1(K)} \qquad \text{for all } v \in H^1(K), \tag{5.13c}$$

$$\Pi_r^{\mathrm{grad}} c = c \qquad \text{for all constant functions } c. \tag{5.13d}$$

This result holds for all integers $p \geq 0$ with the understanding that when $p = 0$, condition (5.13a) is vacuous.

Example 5.5 (Laplace equation: discrete stability and error estimates). We now analyse a discretization of the primal DPG formulation of Example 4.4, making critical use of Theorem 5.4. In particular, we shall see that the *moment conditions* (5.13a)–(5.13b) help us verify the Fortin property (5.5).

Recall the trial and test spaces set in (4.17). The sesquilinear form of the problem on $X \times Y$ with $X = X_0 \times \hat{X} = \mathring{H}^1(\Omega) \times H^{-1/2}(\partial\Omega_h)$ and $Y = H^1(\Omega_h)$ is

$$b((u, \hat{q}_n), y) = (\mathrm{grad}\, u, \mathrm{grad}\, y)_h + \langle \hat{q}_n, y \rangle_h. \tag{5.14}$$

Recalling the Raviart–Thomas space defined in (5.12) and the element-by-element trace operator tr_n defined in (4.10), set the discrete trial space by $X_h = X_{0,h} \times \hat{X}_h$,

where

$$X_{0,h} = \{w \in \mathring{H}^1(\Omega)\colon w|_K \in P_{p+1}(K) \text{ for all } K \in \Omega_h\}, \tag{5.15a}$$

$$\hat{X}_h = \mathrm{tr}_n\big(R^h_{p+1}\big), \tag{5.15b}$$

$$Y_h = P_{p+N}(\Omega_h) \tag{5.15c}$$

for some degree $p \geq 0$. Clearly, the above set $X_{0,h}$ is a standard Lagrange finite element subspace of $\mathring{H}^1(\Omega)$ and Y_h is a standard DG space. Also, the space $\hat{X}_h$ set above is a subspace of $\hat{X} = H^{-1/2}(\partial\Omega_h)$ since $R^h_{p+1} \subseteq H(\mathrm{div}, \Omega)$. An alternative characterization of $\hat{X}_h = \mathrm{tr}_n\big(R^h_{p+1}\big)$ can be given assuming that every $F \in \mathcal{F}_h$ is provided a fixed unit normal n_F, which equals the outward pointing unit normal n if $F \subset \partial\Omega$, and equals either n or $-n$ on an interior facet F shared by an element K with unit outward normal n. Then it is easy to see from the well-known degrees of freedom of the Raviart–Thomas space that

$$\begin{aligned}\hat{X}_h = \{\hat{r}_n\colon &\text{on every } K \in \Omega_h \text{ and each } F \in \triangle_{N-1}K,\\ &\text{there is a } \mu \in P_p(F) \text{ such that } \hat{r}_n|_{\partial K} = (\mu n_F)\cdot n|_{\partial K}\}.\end{aligned} \tag{5.16}$$

Consequently, one may choose to implement $\hat{X}_h$ without using R^h_{p+1}. An implementation of (5.16) can proceed using only the standard polynomial space $P_p(F)$ on each facet $F \in \mathcal{F}_h$ together with some fixed facet orientation given by n_F.

Let us examine what the Fortin condition (5.5) entails for this discrete setting. Let $(w_h, \hat{r}_h) \in X_h$. Since $\hat{r}_h = r_h \cdot n$ for some $r_h \in R^h_{p+1}$, using the $b(\cdot,\cdot)$ in (5.14), condition (5.5) reads as follows:

$$(\mathrm{grad}\, w_h, \mathrm{grad}(y - \Pi y))_h + \langle r_h \cdot n, y - \Pi y\rangle_h = 0 \tag{5.17}$$

for all $w_h \in X_{h,0}$ and $r_h \in R^h_{p+1}$. By integration by parts, we see that (5.17) is implied by

$$\begin{aligned}(y - \Pi y, \Delta w_h)_K = 0 \quad &\text{and}\\ (y - \Pi y, (\mathrm{grad}\, w_h - r_h)\cdot n)_{\partial K} = 0&\end{aligned}$$

on every $K \in \Omega_h$. Once Π is set to Π^{grad}_r of Theorem 5.4, these two identities follow from (5.13a) and (5.13b), respectively, and we are ready to apply Theorem 5.2. Let $(u, \hat{q}_n) \in X$ be the exact solution, and let $q = \mathrm{grad}\, u$, so that $\hat{q}_n = \mathrm{tr}_n(q)$. If $u_h \in X_{0,h}$ and $\hat{q}_{n,h} \in \hat{X}_h$ together form the solution of the practical DPG method with discrete spaces set by (5.15), then Theorem 5.2 yields

$$\begin{aligned}&\|u - u_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)}\\ &\quad\lesssim \inf_{(w_h, r_{n,h})\in X_h}\big(\|u - w_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{r}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)}\big).\end{aligned}$$

To obtain convergence rates, the standard approximation rates for the Lagrange finite element space $X_{h,0}$,

$$\inf_{w_h \in X_{h,0}} \|u - w_h\|_{H^1(\Omega)} \lesssim h^s |u|_{H^{1+s}(\Omega)}, \quad 0 \leq s \leq p+1, \tag{5.18}$$

may be used to bound the first term in the infimum. For the other term we use the definition of the $H^{-1/2}(\partial\Omega_h)$-norm via the minimal extension norm, that is,

$$\inf_{r_{n,h}\in\hat{X}_h} \|\hat{q}_n - \hat{r}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} = \inf_{r_h\in R^h_{p+1}} \|q - r_h\|_{H(\mathrm{div},\Omega)}. \tag{5.19}$$

To obtain convergence rates from this, we use the standard Raviart–Thomas interpolant $\Pi^R q \in R^h_{p+1}$, which is well-defined when $q \in H^s(\Omega)^N \cap H(\mathrm{div},\Omega)$ with $s > 1/2$, together with its commutativity property

$$\mathrm{div}\,\Pi^R q = \Pi_p \,\mathrm{div}\, q, \tag{5.20}$$

where Π_p denotes the $L^2(\Omega)$-orthogonal projection into $P_k(\Omega_h)$, as follows:

$$\begin{aligned}\inf_{r_h\in R^h_{p+1}} \|q - r_h\|^2_{H(\mathrm{div},\Omega)} &\le \|q - \Pi^R q\|^2_\Omega + \|\,\mathrm{div}(q - \Pi^R q)\|^2_\Omega \\ &\le \|q - \Pi^R q\|^2_\Omega + \|(I - \Pi_p)\,\mathrm{div}\, q\|^2_\Omega \\ &\lesssim h^{2s}|q|^2_{H^s(\Omega)} + h^{2s}|\,\mathrm{div}\, q|^2_{H^s(\Omega)},\end{aligned} \tag{5.21}$$

by the usual Bramble–Hilbert argument. Combining (5.18) and (5.21), we obtain

$$\|u - u_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} \lesssim h^s|u|_{H^{1+s}(\Omega)} + h^s|\Delta u|_{H^s(\Omega)}, \tag{5.22}$$

for $1/2 < s \le p+1$.

Although the convergence rate with respect to h in (5.22) is optimal, the last term demands too much regularity. In the remainder of this example, we show how to improve the argument using (4.35a) of Theorem 4.6. Instead of (5.19), we start by applying (4.35a),

$$\begin{aligned}\inf_{r_{n,h}\in\hat{X}_h} \|\hat{q}_n - \hat{r}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)} &\le \|\,\mathrm{tr}_n(q - \Pi^R q)\|_{H^{-1/2}(\partial\Omega_h)} \\ &= \sup_{0\ne y\in H^1(\Omega_h)} \frac{\langle \mathrm{tr}_n(q - \Pi^R q), y\rangle_h}{\|y\|_{H^1(\Omega_h)}}.\end{aligned} \tag{5.23}$$

The numerator above, for any $y \in H^1(\Omega_h)$, satisfies

$$\begin{aligned}\langle \mathrm{tr}_n(q - \Pi^R q), y\rangle_h &= (q - \Pi^R q, \mathrm{grad}\, y)_h + (\mathrm{div}(q - \Pi^R q), y)_h \\ &= (q - \Pi^R q, \mathrm{grad}\, y)_h + ((I - \Pi_p)\,\mathrm{div}\, q, (I - \Pi_p)y)_h,\end{aligned}$$

where we have again used (5.20). Hence

$$\frac{\langle \mathrm{tr}_n(q - \Pi^R q), y\rangle_h}{\|y\|_{H^1(\Omega_h)}} \lesssim h^s|q|_{H^s(\Omega)} + h\|(I - \Pi_p)\,\mathrm{div}\, q\|_\Omega. \tag{5.24}$$

When $1/2 < s < 1$ we can estimate the last term simply by using the fact that the norm of the orthogonal projection $I - \Pi_p$ equals one:

$$h\|(I - \Pi_p)\,\mathrm{div}\, q\| \le h\|\,\mathrm{div}\, q\|, \quad 1/2 < s < 1.$$

When $1 \le s \le k+1$, we use the standard approximation estimate for L^2-projection:

$$h\|(I - \Pi_p)\,\mathrm{div}\, q\|_\Omega \lesssim hh^{s-1}\|\,\mathrm{div}\, q\|_{H^{s-1}(\Omega)}, \quad 1 \le s \le k+1.$$

Using these two estimates to bound the last term in (5.24) and returning to (5.23),

$$\begin{aligned}&\|u-u_h\|_{H^1(\Omega)}+\|\hat{q}_n-\hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)}\\&\quad\lesssim\begin{cases}h^s|u|_{H^{1+s}(\Omega)}+h\|\Delta u\|_{\Omega}, & 1/2<s<1,\\ h^s|u|_{H^{1+s}(\Omega)}+h^s|\Delta u|_{H^{s-1}(\Omega)}, & 1\le s\le k+1.\end{cases}\end{aligned}\tag{5.25}$$

This gives optimal rates at reduced regularity requirements compared to (5.22). For example, in the lowest-order case, if linear Lagrange elements are used for approximating u, and if $u\in H^2(\Omega)$, then the DPG error in u is $O(h)$ (a convergence rate and regularity requirement comparable to the standard finite element method), and additionally, at the price of including a piecewise constant flux $\hat{q}_n$, the DPG method gives a flux error that is also $O(h)$.

Having shown a typical application of a Fortin operator to analyse a DPG method in the above example, let us now proceed to detail the construction of the needed operator Π_r^{grad}. Let K be an N-simplex and let

$$\begin{aligned}\mathring{P}_r(K)&=\{u\in P_r(K)\colon u|_{\partial K}=0\},\\ B_r^0(K)&=\{u\in P_r(K)\colon u|_E=0 \text{ for all } E\in\triangle_{N-2}K\}.\end{aligned}$$

Let $\lambda_0,\dots,\lambda_N$ denote the standard linear barycentric coordinate functions of an N-simplex K, let F_i be the facet in $\triangle_{N-1}K$ where λ_i vanishes, and let

$$b_K=\prod_{j=0}^{N}\lambda_j,\quad b_{F_i}=\frac{b_K}{\lambda_i}=\prod_{j\neq i}\lambda_j.\tag{5.26}$$

Clearly $b_K\in P_{N+1}(K)$ is the element bubble of the simplex K and $b_F\in P_N(K)$ is the facet bubble of any facet F in $\triangle_{N-1}K$.

Lemma 5.6. Let $r=p+N$. Then

$$\dim B_r^0(K)=\dim P_{p-1}(K)+\sum_{F\in\triangle_{N-1}K}\dim P_p(F),\tag{5.27}$$

and for every $v\in H^1(K)$, there is a unique $\Pi_r^0 v\in B_r^0(K)$ satisfying

$$(\Pi_r^0 v-v,q)_K=0\quad\text{for all } q\in P_{p-1}(K),\tag{5.28a}$$

$$(\Pi_r^0 v-v,\mu)_F=0\quad\text{for all } \mu\in P_p(F),\ F\in\triangle_{N-1}K,\tag{5.28b}$$

$$\|\Pi_r^0 v\|_{L^2(K)}+h_K\|\operatorname{grad}\Pi_r^0 v\|_{L^2(K)}\lesssim\|v\|_{L^2(K)}+h_K\|\operatorname{grad} v\|_{L^2(K)}.\tag{5.28c}$$

Proof. Using the space of polynomials of vanishing trace, we can count the dimensions of $B_r^0(K)$. Indeed, $\dim B_r^0(K)=\dim\mathring{P}_r(K)+\sum_{F\in\triangle_{N-1}K}\dim\mathring{P}_r(F)$. Note that $\mathring{P}_r(K)=b_K P_{r-N-1}(K)$ and $\mathring{P}_r(F)=b_F P_{r-N}(F)$ for any $F\in\triangle_{N-1}K$. Therefore, by choosing $r=p+N$, we have

$$\dim\mathring{P}_r(K)=\dim P_{p-1}(K)\quad\text{and}\quad\dim\mathring{P}_r(F)=\dim P_p(F),$$

and consequently (5.28a)–(5.28b) is a square system for $\Pi_r^0 v$.

Now the existence of the stated $\Pi_r^0 v$ will follow from uniqueness, that is, it suffices to prove that if $v = 0$, then $\Pi_r^0 v = 0$. Since $\Pi_r^0 v \in B_r^0(K)$, on any face $F \in \triangle_{N-1}K$, we may write $(\Pi_r^0 v)|_F = b_F w_p$ for some $w_p \in P_p(F)$. But then (5.28b) implies that $\Pi_r^0 v$ must vanish on ∂K, so $\Pi_r^0 v = b_K z_{p-1}$ for some $z_{p-1} \in P_{p-1}(K)$. Then (5.28a) implies that $\Pi_r^0 v = 0$ on K. The estimate (5.28c) now follows by a standard scaling argument using a reference N-simplex. □

Proof of Theorem 5.4. Let $v \in H^1(K)$ on an N-simplex K. Define $\Pi_r^{\mathrm{grad}} v = \Pi_r^0(v - \overline{v}) + \overline{v}$, where $\overline{v}$ denotes the mean value of v on K:

$$\overline{v} = \frac{1}{|K|} \int_K v.$$

Equations (5.13a) and (5.13b) immediately follow from (5.28a) and (5.28b) of Lemma 5.6 since

$$\Pi_r^{\mathrm{grad}} v - v = (\Pi_r^0 - I)(v - \overline{v}).$$

To prove (5.13c), we use (5.28c) and the Poincaré inequality as follows:

$$\begin{aligned} \|\Pi_r^{\mathrm{grad}} v\|_{L^2(K)} &\le \|\overline{v}\|_{L^2(K)} + \|\Pi_r^0(v - \overline{v})\|_{L^2(K)} \\ &\lesssim \|\overline{v}\|_{L^2(K)} + \|v - \overline{v}\|_{L^2(K)} + h_K \|\operatorname{grad}(v - \overline{v})\|_{L^2(K)} \\ &\lesssim \|v\|_{L^2(K)} + h_K \|\operatorname{grad} v\|_{L^2(K)} \end{aligned}$$

and

$$\begin{aligned} h_K \|\operatorname{grad} \Pi_r^{\mathrm{grad}} v\|_{L^2(K)} &= h_K \|\operatorname{grad} \Pi_r^0(v - \overline{v})\|_{L^2(K)} \\ &\lesssim \|v - \overline{v}\|_{L^2(K)} + h_K \|\operatorname{grad}(v - \overline{v})\|_{L^2(K)} \\ &\lesssim h_K \|\operatorname{grad} v\|_{L^2(K)}. \end{aligned}$$

These estimates together imply (5.13c). □

When a lower-order reaction term, say $(u, y)_\Omega$, is added to the Laplace formulation (5.14), skimming through the analysis of Example 5.5, we immediately see that we would need another Fortin operator where the moment condition (5.13a) is strengthened to $q \in P_p(K)$ in place of $q \in P_{p-1}(K)$. To perform such modifications easily, and also to better understand the structure of the Fortin operator we have presented, it is useful to know an explicit representation of the prior Fortin operator and its generalization, which we describe now.

Denote the set of $(N+1)$-term multi-indices of length m by

$$\mathcal{I}_m^{N+1} = \left\{ \beta \equiv (\beta_1, \ldots, \beta_{N+1}) \colon \beta_i \ge 0 \text{ are integers and } |\beta| \equiv \sum_{i=1}^{N+1} \beta_i = m \right\}. \tag{5.29}$$

On an N-simplex, recall that a basis for $P_m(K)$ is given by $\lambda^\alpha = \lambda_1^{\alpha_1} \lambda_2^{\alpha_2} \cdots \lambda_{N+1}^{\alpha_{N+1}}$ for all multi-indices $\alpha \in \mathcal{I}_m^{N+1}$. Let $\eta_\alpha^K = b_K \lambda^\alpha$ and let $\chi_\beta^K \in P_m(K)$ denote the

dual basis of $\{\lambda^\alpha : \alpha \in \mathcal{I}_m^{N+1}\}$ in the $(b_K\cdot,\cdot)_K$ inner product, that is,

$$\left(\eta_\alpha^K, \chi_\beta^K\right)_K = \delta_\alpha^\beta, \quad \alpha,\beta \in \mathcal{I}_m^{N+1}, \tag{5.30}$$

where δ_α^β equals one or zero depending on whether α equals β or not. Let

$$\Pi_m^K v = \sum_{\alpha\in\mathcal{I}_m^{N+1}} \left(v, \chi_\alpha^K\right)_K \eta_\alpha^K, \tag{5.31a}$$

a polynomial in $P_{m+N+1}(K)$. We see from (5.30) that

$$\left(\Pi_m^K v, \chi_\beta^K\right)_K = \left(v, \chi_\beta^K\right)_K \quad \text{for any } \beta \in \mathcal{I}_m^{N+1},$$

so we obtain

$$\left(\Pi_m^K v - v, q\right)_K = 0 \quad \text{for all } q \in P_m(K),\ v \in L^2(K), \tag{5.31b}$$

after expanding q in the χ_β^K-basis.

Since this construction works on a simplex of any dimension, we can repeat it on any subsimplex of K. The barycentric coordinates of a subsimplex $F \in \triangle_{N-1}K$ are simply the restrictions of those of K to F, omitting the one that vanishes on F. Using them, we repeat the construction, now with shorter multi-indices $\alpha,\beta \in \mathcal{I}_k^N$. Namely, let $\eta_\alpha^F = b_F\lambda^\alpha$ and let $\chi_\beta^F \in P_k(F)$ form the dual basis of λ^α in the $(b_F\cdot,\cdot)_F$ inner product. Then set

$$\Pi_k^F v = \sum_{\alpha\in\mathcal{I}_k^N} \left(v, \chi_\alpha^F\right)_F \eta_\alpha^F. \tag{5.32a}$$

This is a polynomial in $P_{k+N}(K)$ since each η_α^F is a product of $k+N$ barycentric coordinates on F that have a natural polynomial extension into K. Furthermore, this polynomial vanishes on all facets in $\triangle_{N-1}K$ different from F. As in (5.31b), the analogue of (5.30) on F now gives

$$\left(\Pi_k^F v - v, q\right)_F = 0 \quad \text{for all } q \in P_k(F),\ v \in H^1(K). \tag{5.32b}$$

Let Π_0 denote the $L^2(K)$-orthogonal projection to constants on K. We put these ingredients together to construct the polynomial

$$\Pi_{k,m}^{\text{grad}} v = \Pi_0 v + \sum_{F\in\triangle_{N-1}K} \left(\Pi_k^F + \Pi_m^K\left(I - \Pi_k^F\right)\right)(I - \Pi_0)v. \tag{5.33}$$

Note that its trace on a facet F is determined solely by the Π_k^F-contribution since the traces after an application of Π_m^K or $\Pi_k^{\tilde F}$ are zero on F for any $\tilde F \neq F$. Note that when $m = p-1$ and $k = p$, we recover the operator of Theorem 5.4.

Theorem 5.7 (A more general $H^1(K)$ Fortin operator). The above-defined operator

$$\Pi_{k,m}^{\text{grad}} : H^1(K) \to P_r(K) \quad \text{for } r = \max(m+N+1, k+N)$$

satisfies, for every $v \in H^1(K)$ on an N-simplex K, the moment conditions

$$\left(\Pi^{\mathrm{grad}}_{k,m} v - v, q\right)_K = 0 \quad \text{for all } q \in P_m(K), \tag{5.34a}$$

$$\left(\Pi^{\mathrm{grad}}_{k,m} v - v, \mu\right)_F = 0 \quad \text{for all } \mu \in P_k(F),\ F \in \triangle_{N-1}K, \tag{5.34b}$$

$$\Pi^{\mathrm{grad}}_{k,m} c = c \quad \text{for all constant functions } c \text{ on } K, \tag{5.34c}$$

and the norm estimates

$$\|\Pi^{\mathrm{grad}}_{k,m} v\|_{L^2(K)} \lesssim \|v\|_{L^2(K)} + h_K \|\operatorname{grad} v\|_{L^2(K)}, \tag{5.34d}$$

$$\|\operatorname{grad} \Pi^{\mathrm{grad}}_{k,m} v\|_{L^2(K)} \lesssim \|\operatorname{grad} v\|_{L^2(K)}. \tag{5.34e}$$

Proof. It is easy to see that using (5.31) and (5.32) that (5.34a) and (5.34b) hold. Property (5.34c) is immediate from (5.33). The norm estimate of (5.34d) is proved along the same lines as (5.13) by scaling arguments, and the estimate (5.34e) follows from (5.34d) and (5.34b). □

Bibliographical notes. Theorem 5.4 and the construction of the Fortin operator $\Pi^{\mathrm{grad}}_{p+3}$ are taken from Gopalakrishnan and Qiu (2014). Its generalization in (5.33) is based on the recent work of Führer and Heuer (2024). They further show that generalizing such polynomial expressions to certain exponential ones, estimates like (5.13) but with $\|\cdot\|_{H^1(K)}$ replaced by

$$\|v\|_a = \left(\|v\|_K^2 + a\|\operatorname{grad} v\|_K^2\right)^{1/2}$$

for some small parameter a, can be obtained robustly in the parameter a as $a \to 0$. The discrete stability of the primal DPG method for the Laplace equation, discussed in Example 5.5, was first considered in Demkowicz and Gopalakrishnan (2013). There, and in earlier DPG analyses such as that of Demkowicz and Gopalakrishnan (2011*a*), error estimates comparable to (5.22) that demand extra regularity can be found. The discussion in Example 5.5, leading to (5.25) with better regularity requirements, is taken from the more recent work of Führer (2018, Theorem 5).

5.3. *A Fortin operator for divergence in N dimensions*

In this subsection we construct a continuous linear operator Π^{div}_{p+3} on $H(\mathrm{div}, K)$ satisfying certain moment conditions that are useful in analysis of DPG methods where $H(\mathrm{div}, \Omega_h)$ features in the test space.

We will perform our construction on the reference unit N-simplex $\hat{K}$ and map it to a general N-simplex K. Let $S_K : \hat{K} \to K$ be a one-to-one affine map that maps $\hat{K}$ onto a general tetrahedron K, and let $[S'_K]$ denote the Jacobian derivative matrix of S_K. Given vector fields q and E on K, we use the following pullback to map them to $\hat{K}$:

$$\Psi(q) = (\det[S'_K])\,[S'_K]^{-1}(q \circ S_K). \tag{5.35}$$

It is easy to see that

$$\operatorname{div}(\Psi(q)) = (\det[S_K'])\,(\operatorname{div} q)\circ S_K. \tag{5.36}$$

Recall the Raviart–Thomas element $R_p(K)$, $p \geq 1$, previously considered in (5.11).

Theorem 5.8 (A Fortin operator on $H(\operatorname{div}, K)$). On any N-simplex K, for any integer $k \geq 0$, an operator $\Pi^{\operatorname{div}}_{k+1}\colon H(\operatorname{div}, K) \to R_{k+1}(K)$ can be constructed such that for all $\tau \in H(\operatorname{div}, K)$, we have the commutativity property

$$\operatorname{div} \Pi^{\operatorname{div}}_{k+1}\tau = \Pi_k \operatorname{div}\tau, \tag{5.37}$$

the moment conditions

$$(\Pi^{\operatorname{div}}_{k+1}\tau - \tau, q)_K = 0 \quad \text{for all } q \in P_{k-1}(K)^N, \tag{5.38a}$$

$$(n\cdot(\Pi^{\operatorname{div}}_{k+1}\tau - \tau), \mu)_{\partial K} = 0 \quad \text{for all } \mu \in P_k(K) \tag{5.38b}$$

and the norm bounds

$$\|\Pi^{\operatorname{div}}_{k+1}\tau\|_K \lesssim \|\tau\|_K + h_K\|\operatorname{div}\tau\|_K, \tag{5.38c}$$

$$\|\operatorname{div}\Pi^{\operatorname{div}}_{k+1}\tau\|_K \leq \|\operatorname{div}\tau\|_K. \tag{5.38d}$$

Proof. We will first construct the operator on the reference unit N-simplex $\hat K$. Let

$$P_k^{\perp}(\partial\hat K) = \{\mu \in L^2(\partial K)\colon \mu|_F \in P_k(F) \text{ for all } F \in \triangle_{N-1}K \text{ and } (\mu, q)_{\partial K} = 0 \text{ for all } q \in P_k(K)\}.$$

It is the $L^2(\partial\hat K)$-orthogonal complement of $\operatorname{tr}(P_k(\hat K))$ in the space of piecewise polynomials on ∂K. Let

$$B^{\operatorname{div}}_{k+1}(\hat K) = \{\hat\tau \in R_{k+1}(\hat K)\colon (\hat p_\perp, \hat\tau\cdot\hat n)_{\partial\hat K} = 0 \text{ for all } \hat p_\perp \in P^{\perp}_{k+1}(\partial\hat K)\},$$

where $\hat n$ is the unit outward normal on $\partial\hat K$ and $R_{k+1}(\hat K)$ is as in (5.11). We claim that the equations

$$\big(\hat\Pi^{\operatorname{div}}_{k+1}\hat\tau, \hat q\big)_{\hat K} = (\hat\tau, \hat q)_{\hat K} \qquad \text{for all } \hat q \in P_{k-1}(\hat K)^N, \tag{5.39a}$$

$$\big(\hat\Pi^{\operatorname{div}}_{k+1}\hat\tau\cdot\hat n, \hat w\big)_{\partial\hat K} = \langle\hat\tau\cdot\hat n, \hat w\rangle_{H^{1/2}(\partial K)} \quad \text{for all } \hat w \in P_k(\hat K) \tag{5.39b}$$

uniquely determine $\hat\Pi^{\operatorname{div}}_{k+1}\hat\tau \in B^{\operatorname{div}}_{k+1}(\hat K)$ and thus define a linear continuous operator

$$\hat\Pi^{\operatorname{div}}_{k+1}\colon H(\operatorname{div}, \hat K) \to B^{\operatorname{div}}_{k+1}(\hat K).$$

Indeed, if the right-hand sides of (5.39) vanish, then since $\hat\Pi^{\operatorname{div}}_{k+1}\hat\tau$ is in $B^{\operatorname{div}}_{k+1}(\hat K) \subset R_{k+1}(\hat K)$, we find that $\hat\Pi^{\operatorname{div}}_{k+1}$ is a function in the Raviart–Thomas space all of whose canonical degrees of freedom vanish (see e.g. Arnold, Falk and Winther 2006 or Nédélec 1980, Definition 5), so

$$\hat\Pi^{\operatorname{div}}_{k+1}\hat\tau = 0.$$

Since the system (5.39) is square, we conclude that it uniquely defines $\hat\Pi^{\operatorname{div}}_{k+1}\hat\tau$.

Next, we map this operator to a general simplex K using the Piola transform Ψ in (5.35):

$$\Pi^{\mathrm{div}}_{k+1} = \Psi^{-1} \circ \hat{\Pi}^{\mathrm{div}}_{k+1} \circ \Psi.$$

By standard mapping arguments, the stated moment conditions of Π^{div}_{k+1} now follow from (5.39). The moment conditions also imply that for any $\omega \in P_k(K)$,

$$\begin{aligned}\big(\mathrm{div}\big(\Pi^{\mathrm{div}}_{k+1}\tau\big), \omega\big)_K &= -\big(\Pi^{\mathrm{div}}_{k+1}\tau, \mathrm{grad}\,\omega\big)_K + \big(\big(\Pi^{\mathrm{div}}_{k+1}\tau\big)\cdot n, \omega\big)_{\partial K} \\ &= -(\tau, \mathrm{grad}\,\omega)_K + (\omega, \tau\cdot n)_{\partial K} \\ &= (\mathrm{div}\,\tau, \omega)_K,\end{aligned}$$

thus proving the commutativity property (5.37). Also, since $\hat{\Pi}^{\mathrm{div}}_{k+1}$ is a continuous operator on $H(\mathrm{div}, \hat{K})$, standard scaling arguments prove

$$\|\Pi^{\mathrm{div}}_{k+1}\tau\|_K + h_K \|\,\mathrm{div}\,\Pi^{\mathrm{div}}_{k+1}\tau\|_K \lesssim \|\tau\|_K + h_K\|\,\mathrm{div}\,\tau\|_K.$$

Combined with the better bound on the divergence term,

$$\|\,\mathrm{div}\,\Pi^{\mathrm{div}}_{k+1}\tau\|_K \le \|\,\mathrm{div}\,\tau\|_K,$$

which obviously follows from (5.37), the stated norm bound is also proved. □

Bibliographical notes. Theorem 5.8 and its proof are from Gopalakrishnan and Qiu (2014). For a construction in $H(\mathrm{div}, K)$ in the same spirit as Theorem 5.7, see Führer and Heuer (2024).

5.4. *Commuting Fortin operators in three dimensions*

Having completed the previous discussions of a Fortin operator Π^{grad}_r on the broken H^1 space, as well an $H(\mathrm{div}, K)$ Fortin operator Π^{div}_{k+1}, we now proceed to show that they are part of a family of local commuting Fortin operators. We restrict ourselves to the three-dimensional (3D) $N = 3$ case. Let

$$N_p(D) = P_{p-1}(D)^3 + x \times P_{p-1}(D)^3$$

denote the Nédélec element (Nédélec 1980). Together with the Raviart–Thomas element $R_{p+1}(K)$ in (5.11), it forms the following well-known (see e.g. Arnold *et al.* 2006) exact complex:

$$0 \longrightarrow P_{p+1}(K)/\mathbb{R} \xrightarrow{\mathrm{grad}} N_{p+1}(K) \xrightarrow{\mathrm{curl}} R_{p+1}(K) \xrightarrow{\mathrm{div}} P_p(K) \longrightarrow 0. \quad (5.40)$$

We will prove the following result in this subsection. There Π_p denotes the L^2-orthogonal projection onto $P_p(K)$. In order to verify condition (5.5) in various DPG convergence analyses, the moment conditions (5.43)–(5.48) listed below are helpful.

Theorem 5.9 (Commuting 3D Fortin operators satisfying moment conditions). Let $p \geq 0$ be an integer. On any tetrahedron K, there are operators

$$\begin{aligned}\Pi^{\mathrm{grad}}_{p+3} &: H^1(K) \to P_{p+3}(K),\\ \Pi^{\mathrm{curl}}_{p+3} &: H(\mathrm{curl}, K) \to N_{p+3}(K),\\ \Pi^{\mathrm{div}}_{p+3} &: H(\mathrm{div}, K) \to R_{p+3}(K),\end{aligned}$$

such that for any $v \in H^1(K)$, $E \in H(\mathrm{curl}, K)$ and $\tau \in H(\mathrm{div}, K)$, the norm estimates

$$\|\Pi^{\mathrm{grad}}_{p+3} v\|_{H^1(K)} \lesssim \|v\|_{H^1(K)}, \tag{5.41a}$$

$$\|\Pi^{\mathrm{curl}}_{p+3} E\|_{H(\mathrm{curl},K)} \lesssim \|E\|_{H(\mathrm{curl},K)}, \tag{5.41b}$$

$$\|\Pi^{\mathrm{div}}_{p+3} \tau\|_{H(\mathrm{div},K)} \lesssim \|\tau\|_{H(\mathrm{div},K)} \tag{5.41c}$$

hold, the diagram

$$\begin{array}{ccccccc} H^1(K)/\mathbb{R} & \xrightarrow{\mathrm{grad}} & H(\mathrm{curl}, K) & \xrightarrow{\mathrm{curl}} & H(\mathrm{div}, K) & \xrightarrow{\mathrm{div}} & L^2(K) \\ \downarrow \Pi^{\mathrm{grad}}_{p+3} & & \downarrow \Pi^{\mathrm{curl}}_{p+3} & & \downarrow \Pi^{\mathrm{div}}_{p+3} & & \downarrow \Pi_{p+2} \\ P_{p+3}(K)/\mathbb{R} & \xrightarrow{\mathrm{grad}} & N_{p+3}(K) & \xrightarrow{\mathrm{curl}} & R_{p+3}(K) & \xrightarrow{\mathrm{div}} & P_{p+2}(K) \end{array} \tag{5.42}$$

commutes, and the following identities hold:

$$\left(\Pi^{\mathrm{grad}}_{p+3} v - v, q\right)_K = 0 \quad \text{for all } q \in P_{p-1}(K), \tag{5.43}$$

$$\left(\Pi^{\mathrm{grad}}_{p+3} v - v,\ n \cdot \sigma\right)_{\partial K} = 0 \quad \text{for all } \sigma \in R_{p+1}(K), \tag{5.44}$$

$$\left(\Pi^{\mathrm{curl}}_{p+3} E - E, v\right)_K = 0 \quad \text{for all } v \in P_p(K)^3, \tag{5.45}$$

$$\left(n \times \left(\Pi^{\mathrm{curl}}_{p+3} E - E\right), w\right)_{\partial K} = 0 \quad \text{for all } w \in P_{p+1}(K)^3, \tag{5.46}$$

$$\left(\Pi^{\mathrm{div}}_{p+3} \tau - \tau, q\right)_K = 0 \quad \text{for all } q \in P_{p+1}(K)^3, \tag{5.47}$$

$$\left(n \cdot \left(\Pi^{\mathrm{div}}_{p+3} \tau - \tau\right), \mu\right)_{\partial K} = 0 \quad \text{for all } \mu \in P_{p+2}(K). \tag{5.48}$$

Before proceeding to prove this theorem, we note that the operator $\Pi^{\mathrm{grad}}_{p+3}$ stated in the theorem is the same as Π^{grad}_r with $r = p + N$ constructed in Theorem 5.4, restricted to $N = 3$ dimensions (since (5.43)–(5.44) is the same as (5.28a)–(5.28b)). However, we are yet to prove the relevant commutativity property. Note also that the bound (5.41c) and the moment conditions (5.47) and (5.48) hold after putting $N = 3$ and replacing k with $p + 2$ in Theorem 5.8. It also gives the commutativity property of Π^{div}_{p+3} stated in the last part of (5.42).

It remains to construct $\Pi^{\mathrm{curl}}_{p+3}$. We will perform our construction on the reference unit N-simplex $\hat{K}$ and map it to a general N-simplex K. As before, let $S_K : \hat{K} \to K$ be an affine homeomorphism from $\hat{K}$ to a general tetrahedron K. Given a vector

fields E on K, we use the following pullback to map it to $\hat{K}$:

$$\Phi(E) = [S_K']^t (E \circ S_K). \tag{5.49}$$

It is easy to see that

$$\operatorname{curl}(\Phi(E)) = \Psi(\operatorname{curl} E). \tag{5.50}$$

We begin with a preliminary lemma whose relevance will be clear soon. Let

$$\mathring{N}_p(K) = \{q \in N_p(K) \colon n \times q|_{\partial K} = 0\}.$$

Lemma 5.10. For any $p \geq 0$, if $F \in H(\operatorname{curl}, K)$ satisfies

$$(\operatorname{curl} F, w)_K = 0 \quad \text{for all } w \in \mathring{N}_{p+1}(K), \tag{5.51}$$

then there is a $\phi \in P_{p+3}(K)$ such that

$$(F + \operatorname{grad} \phi, v)_K = 0 \quad \text{for all } v \in P_p(K)^3. \tag{5.52}$$

Proof. Proving (5.52) amounts to proving that there is a $\phi \in P_{p+3}(K)$ solving

$$A\phi = b, \tag{5.53}$$

where A and b are defined using the $L^2(K)^3$-orthogonal projection Π_p into $P_p(K)^3$ by

$$A = \Pi_p \operatorname{grad} \colon P_{p+3}(K) \to P_p(K)^3, \quad b = -\Pi_p F.$$

In other words, it suffices to show that $b \in \operatorname{range}(A) = \ker(A^*)^\perp$, where A^* is the L^2-adjoint of A.

Any $q \in P_p(K)^3$ is in $\ker(A^*)$ if and only if

$$\begin{aligned}(u, A^* q) &= (Au, q)_K = (\operatorname{grad} u, q)_K \\ &= -(u, \operatorname{div} q)_K + (u, q \cdot n)_{\partial K} = 0\end{aligned} \tag{5.54}$$

for all $u \in P_{p+3}(K)$. Recall the bubble functions b_K and b_F from (5.26). Choosing $u = b_K \operatorname{div} q$ in (5.54), we find that $\operatorname{div} q = 0$. Then, removing the term containing $\operatorname{div} q$ and choosing $u = (q \cdot n_F) b_F$ in (5.54), we find that $(q \cdot n_F)|_F = 0$, an argument that can repeated on every facet F. Including the obvious converse as well, we have proved that $q \in \ker(A^*)$ if and only if $\operatorname{div} q = 0$ and $q \cdot n|_{\partial K} = 0$, that is,

$$\ker(A^*) = \operatorname{curl} \mathring{N}_{p+1}(K).$$

Note that for any $w \in \mathring{N}_{p+1}(K)$, by the given condition (5.51),

$$(b, \operatorname{curl} w) = -(F, \operatorname{curl} w) = -(\operatorname{curl} F, w) = 0,$$

so $b \in \ker(A^*)^\perp = \operatorname{range}(A)$ and (5.53) has a solution. □

Before constructing the required $\Pi^{\operatorname{curl}}_{p+3}$, we need an intermediate operator $\hat{\Pi}^c_{p+3}$ on a reference unit tetrahedron $\hat{K}$. Let

$$D_{p+2}(\hat{K}) = \operatorname{curl} N_{p+3}(\hat{K}) = \{r \in R_{p+3}(\hat{K}) \colon \operatorname{div} r = 0\}$$

and

$$C = \operatorname{curl}\colon N_{p+3}(\hat{K}) \to D_{p+2}(\hat{K}).$$

Using $\hat{\Pi}_p$, the $L^2(\hat{K})^3$-orthogonal projection into $P_p(\hat{K})^3$, define $B = \hat{\Pi}_p\colon N_{p+3} \to P_p(\hat{K})^3$. The codomains of B and C are endowed with the L^2-norm, which then naturally define their L^2-adjoints B^* and C^*. Note that one of the commutativity properties in (5.42) and one of the moment conditions (5.45) read, respectively, as follows:

$$CF = \Pi^{\mathrm{div}}_{p+3} \operatorname{curl} E, \quad BF = \hat{\Pi}_p E, \tag{5.55}$$

with $F = \Pi^{\mathrm{curl}}_{p+3} E$. Accordingly, we seek the result of the application of the Fortin operator in the set

$$S(E) = \{F \in N_{p+3}(\hat{K})\colon F \text{ satisfies (5.55)}\}. \tag{5.56}$$

For defining the intermediate operator $\hat{\Pi}^c_{p+3}$, consider the problem of finding $\hat{\Pi}^c_{p+3} E \in N_{p+3}(\hat{K})$, $\lambda \in P_p(\hat{K})^3$ and $\mu \in D_{p+2}(\hat{K})$ satisfying

$$\hat{\Pi}^c_{p+3} E + B^*\lambda + C^*\mu = 0, \tag{5.57a}$$

$$B\hat{\Pi}^c_{p+3} E = \hat{\Pi}_p E, \tag{5.57b}$$

$$C\hat{\Pi}^c_{p+3} E = \hat{\Pi}^{\mathrm{div}}_{p+3} \operatorname{curl} E, \tag{5.57c}$$

where $\hat{\Pi}^{\mathrm{div}}_{p+3}$ is as defined in (5.39). One may view λ and μ above as Lagrange multipliers for the constrained minimization problem

$$\hat{\Pi}^c_{p+3} E = \arg \min_{F \in S(E)} \|F\|^2_{\hat{K}}, \tag{5.58}$$

where the minimization is over the affine set in (5.56).

We claim that there exists a unique $\hat{\Pi}^c_{p+3} E \in N_{p+3}(\hat{K})$ satisfying (5.57). First observe that (5.37) implies $\operatorname{div} \hat{\Pi}^{\mathrm{div}}_{p+3} \operatorname{curl} E = 0$, that is, by the exactness of (5.40), there is a $\hat{E}_{p+3} \in N_{p+3}(K)$ such that $\operatorname{curl} \hat{E}_{p+3} = \hat{\Pi}^{\mathrm{div}}_{p+3} \operatorname{curl} E$. By the moment condition (5.47) of Π^{div}_{p+3}, $F = \hat{E}_{p+3} - E$ satisfies $(\operatorname{curl} F, w) = 0$ for all $w \in P_{p+1}(K)$, and in particular, the condition (5.51) of Lemma 5.10. The lemma then gives the existence of $\phi \in P_{p+3}(K)$ such that $G = \hat{E}_{p+3} + \operatorname{grad} \phi \in N_{p+3}(\hat{K})$ satisfies

$$\begin{bmatrix} B \\ C \end{bmatrix} G = \begin{bmatrix} \hat{\Pi}_p E \\ \hat{\Pi}^{\mathrm{div}}_{p+3} \operatorname{curl} E \end{bmatrix},$$

that is, the right-hand side of (5.57) is in the range of $\left[\begin{smallmatrix} B \\ C \end{smallmatrix}\right]$ (or equivalently $S(E)$ is a non-empty feasible set of constraints). Hence, by standard arguments (see Brezzi and Fortin 1991, Proposition 1.1, p. 38), there *exists* a solution to (5.57) and moreover the $\hat{\Pi}^c_{p+3} E$ component of the solution is *unique*. The linearity of $\hat{\Pi}^c_{p+3} E$ with respect to the right-hand sides of (5.57) is obvious and the right-hand sides in turn depend linearly on E. Furthermore, since the ranges of B and C are closed finite-dimensional spaces, there is a $c_{\hat{K}} > 0$ such that (see Brezzi and Fortin 1991,

Proposition 1.2, p. 39) the linear operator $\hat{\Pi}^c_{p+3}\colon H(\mathrm{curl},\hat{K}) \to N_{p+3}(\hat{K})$ satisfies

$$\|\hat{\Pi}^c_{p+3}E\|_{H(\mathrm{curl},\hat{K})} \le c_{\hat{K}}\|E\|_{H(\mathrm{curl},\hat{K})}. \tag{5.59}$$

Next we map $\hat{\Pi}^c_{p+3}$ from $\hat{K}$ to an operator on any shape-regular tetrahedron K using the covariant pullback Φ in (5.50):

$$\Pi^c_{p+3} = \Phi^{-1} \circ \hat{\Pi}^c_{p+3} \circ \Phi.$$

Lemma 5.11. On any tetrahedron K, the operator $\Pi^c_{p+3}\colon H(\mathrm{curl},K) \to N_{p+3}(K)$ satisfies

$$\mathrm{curl}\,\Pi^c_{p+3}E = \Pi^{\mathrm{div}}_{p+3}\,\mathrm{curl}\,E \qquad \text{for all } E \in H(\mathrm{curl},K), \tag{5.60}$$

$$\big(\Pi^c_{p+3}E - E, v\big)_K = 0 \qquad \text{for all } v \in P_p(K)^3, \tag{5.61}$$

$$\big(n \times \big(\Pi^c_{p+3}E - E\big), w\big)_{\partial K} = 0 \qquad \text{for all } w \in P_{p+1}(K)^3, \tag{5.62}$$

$$\|\Pi^c_{p+3}E\|_{H(\mathrm{curl},K)} \lesssim \|E\|_{H(\mathrm{curl},K)}. \tag{5.63}$$

Proof. Mapping the equation (5.57b) from $\hat{K}$ to K, we obtain (5.61). It is easy to see that $\mathrm{curl}(\Phi(E)) = \Psi(\mathrm{curl}\,E)$ for any $E \in H(\mathrm{curl},K)$. Using it, we note that mapping (5.57c) from $\hat{K}$ to K we obtain the commutativity property (5.60) on K. To prove the extra boundary moment condition (5.62), we substitute $v = \mathrm{curl}\,w$ for some $w \in P_{p+1}(K)^3$ into (5.61) and integrate by parts to get

$$\begin{aligned} 0 &= \big(\Pi^c_{p+3}E - E, \mathrm{curl}\,w\big)_K \\ &= -\big(n \times \big(\Pi^c_{p+3}E - E\big), w\big)_K + \big(\mathrm{curl}\big(\Pi^c_{p+3}E - E\big), w\big)_K, \end{aligned}$$

and the last term vanishes by (5.60) and the moment condition (5.47) of Π^{div}_{p+3}.

To prove (5.63), note that the bound (5.59) and standard scaling arguments (detailed in Carstensen *et al.* 2016, eq. (52)) imply

$$\|\Pi^c_{p+3}E\|^2_K \lesssim \|E\|^2_K + h^2_K\|\mathrm{curl}\,E\|^2_K.$$

Additionally, since (5.57c) and (5.41b) imply that $\|\mathrm{curl}\,\Pi^c_{p+3}E\|_K \lesssim \|\mathrm{curl}\,E\|_K$, the estimate (5.63) follows. □

Recall that any $E \in H(\mathrm{curl},K)$ admits the unique orthogonal Helmholtz decomposition

$$E = \tilde{E} + \mathrm{grad}\,\psi, \tag{5.64}$$

where $\psi \in H^1(K)$ has zero mean value $\bar{\psi} = |K|^{-1}\int_K \psi = 0$, and $\tilde{E} \in H(\mathrm{curl},K)$ is such that $(\tilde{E}, \mathrm{grad}\,\varphi)_K = 0$ for all $\varphi \in H^1(K)$. Using the Helmholtz decomposition (5.64) of E, define

$$\Pi^{\mathrm{curl}}_{p+3}E = \Pi^c_{p+3}\tilde{E} + \mathrm{grad}\,\Pi^{\mathrm{grad}}_{p+3}\psi. \tag{5.65}$$

We proceed to prove that this operator has all the required properties.

Proof of Theorem 5.9. We have already shown that the N-dimensional operator Π^{div}_{p+3} in Theorem 5.8, restricted to $N = 3$ case, satisfies all the properties stated in the theorem. We have also shown that the operator in Theorem 5.4, restricted to $N = 3$, satisfies all the stated properties of $\Pi^{\mathrm{grad}}_{p+3}$ except for its commutativity property involving $\Pi^{\mathrm{curl}}_{p+3}$. Hence, to finish the proof, we now proceed to prove that the $\Pi^{\mathrm{curl}}_{p+3} E$ defined in (5.65) satisfies the norm bound (5.41b), the moment conditions (5.45)–(5.46), as well as the commutativity properties

$$\operatorname{curl} \Pi^{\mathrm{curl}}_{p+3} E = \Pi^{\mathrm{div}}_{p+3} \operatorname{curl} E, \tag{5.66}$$

$$\Pi^{\mathrm{curl}}_{p+3} \operatorname{grad} \phi = \operatorname{grad} \Pi^{\mathrm{grad}}_{p+3} \phi \tag{5.67}$$

for all $\phi \in H^1(K)$ and $E \in H(\mathrm{curl}, K)$.

First, we use, in succession, the definition of $\Pi^{\mathrm{curl}}_{p+3}$ in (5.65), the norm bound (5.63) of the intermediate operator Π^{c}_{p+3} and that of $\Pi^{\mathrm{grad}}_{p+3}$ in (5.41a), the orthogonality of the Helmholtz decomposition which implies

$$\|\tilde{E}\|^2_K + \|\operatorname{grad} \psi\|^2_K = \|E\|^2_K,$$

and the Poincaré inequality $\|\psi\|_K \lesssim h_K \|\operatorname{grad} \psi\|^2_K$, to get

$$\begin{aligned} \|\Pi^{\mathrm{curl}}_{p+3} E\|_{H(\mathrm{curl},K)} &\leq \|\Pi^{c}_{p+3} \tilde{E}\|_{H(\mathrm{curl},K)} + \|\operatorname{grad} \Pi^{\mathrm{grad}}_{p+3} \psi\|_K \\ &\lesssim \|\tilde{E}\|_{H(\mathrm{curl},K)} + \|\psi\|_{H^1(K)} \lesssim \|E\|_{H(\mathrm{curl},K)}, \end{aligned}$$

thus proving the required bound (5.41b).

To prove the interior moment condition (5.45), we again start by applying the definition (5.65):

$$\left(\Pi^{\mathrm{curl}}_{p+3} E - E, v\right)_K = \left(\Pi^{c}_{p+3} \tilde{E} - \tilde{E}, v\right)_K + \left(\operatorname{grad}\left(\Pi^{\mathrm{grad}}_{p+3} \psi - \psi\right), v\right)_K.$$

Note that the first term on the right-hand side vanishes since $\Pi^{c}_{p+3} \tilde{E}$ satisfies the moment condition (5.61) of Lemma 5.11. The last term vanishes after integrating by parts and using the moment conditions (5.43)–(5.44) of $\Pi^{\mathrm{grad}}_{p+3}$.

Next, to prove the element boundary moment condition (5.46), starting with

$$\begin{aligned} \left(n \times \left(\Pi^{\mathrm{curl}}_{p+3} E - E\right), w\right)_{\partial K} &= \left(n \times \left(\Pi^{c}_{p+3} \tilde{E} - \tilde{E}\right), w\right)_{\partial K} \\ &\quad + \left(n \times \operatorname{grad}\left(\Pi^{\mathrm{grad}}_{p+3} \psi - \psi\right), w\right)_{\partial K}, \end{aligned} \tag{5.68}$$

note that the first term on the right-hand side vanishes due to the moment condition (5.62) of $\Pi^{c}_{p+3} \tilde{E}$. To see that the last term also vanishes, letting $e = \Pi^{\mathrm{grad}}_{p+3} \psi - \psi$, observe that the equalities

$$(e, \operatorname{div} q)_K = 0 = -(q, \operatorname{grad} e)_K$$

can be seen to hold for any $q \in P_p(K)^3$ due to the moment conditions (5.43)–(5.44) of $\Pi^{\mathrm{grad}}_{p+3}$ and integration by parts. Putting $q = \operatorname{curl} w$ for any $w \in P_{p+1}(K)^3$, the

last equality implies that

$$0 = (\operatorname{curl} w, \operatorname{grad} e)_K = -(n \times \operatorname{grad} e, w)_{\partial K},$$

which shows that the last term in (5.68) vanishes.

The proof of the commutativity property (5.66) is straightforward:

$$\begin{aligned} \operatorname{curl} \Pi^{\operatorname{curl}}_{p+3} E &= \operatorname{curl} \Pi^{c}_{p+3} \tilde{E} \quad \text{by (5.65)} \\ &= \Pi^{\operatorname{div}}_{p+3} \operatorname{curl} \tilde{E} \quad \text{by (5.60)} \\ &= \Pi^{\operatorname{div}}_{p+3} \operatorname{curl} E \quad \text{by (5.64).} \end{aligned}$$

Finally, to prove the remaining commutativity property (5.67), observe that the Helmholtz decomposition (5.64) of $E = \operatorname{grad} \phi$ gives a vanishing $\tilde{E}$-component and a ψ-component that equals $\phi - \bar{\phi}$ for any $\phi \in H^1(K)$. Hence

$$\Pi^{\operatorname{curl}}_{p+3} \operatorname{grad} \phi = \operatorname{grad} \Pi^{\operatorname{grad}}_{p+3}(\phi - \bar{\phi}) = \operatorname{grad} \Pi^{\operatorname{grad}}_{p+3} \phi - \operatorname{grad} \bar{\phi} = \operatorname{grad} \Pi^{\operatorname{grad}}_{p+3} \phi,$$

where we have used (5.13d). □

Example 5.12 (Maxwell equations). We continue Example 4.5, where the infinite-dimensional spaces and forms for the primal DPG formulation of the Maxwell cavity problem were set by (4.32), and the formulation was proved to be wellposed. Now we focus on its discretization using subspaces $X_{0,h} \subset X_h$, $\hat{X}_h \subset \hat{X}$ and $Y_h \subset Y$ set by

$$X_{0,h} = \{E_h \in \mathring{H}(\operatorname{curl}, \Omega) \colon E_h|_K \in P_p(K)^3 \text{ for all } K \in \Omega_h\}, \tag{5.69a}$$

$$\hat{X}_h = \{n \times \hat{H}_h \in H^{-1/2}(\operatorname{div_F}, \partial\Omega_h) \colon n \times \hat{H}_h|_{\partial K} \in \operatorname{tr}^K_{n\times} P_{p+1}(K)^3 \text{ for all } K \in \Omega_h\},$$

$$Y_h = \{F_h \in H(\operatorname{curl}, \Omega_h) \colon F_h|_K \in N_{p+3}(K) \text{ for all } K \in \Omega_h\}. \tag{5.69b}$$

To obtain error estimates, we apply Theorem 5.2, under the additional assumption that the material coefficients μ, ε are constant on each mesh element $K \in \Omega_h$. Then (5.41b), (5.45) and (5.46) of Theorem 5.9 verify condition (5.5) with $\Pi = \Pi^{\operatorname{curl}}_{p+3}$. Hence we conclude that

$$\begin{aligned} &\|E - E_h\|^2_{H(\operatorname{curl},\Omega)} + \|n \times (\hat{H} - \hat{H}_h)\|^2_{H^{-1/2}(\operatorname{div_F},\partial\Omega_h)} \\ &\quad \lesssim \inf_{G_h \in X_{h,0},\ n\times\hat{R}_h \in \hat{X}_h} \Big[\|E - G_h\|^2_{H(\operatorname{curl},\Omega)} + \|n \times \hat{H} - n \times \hat{R}_h\|^2_{H^{-1/2}(\operatorname{div_F},\partial\Omega_h)} \Big]. \end{aligned}$$

Thus the method is quasioptimal. Convergence rates can be derived by bounding the right-hand side (as illustrated in Example 5.5). Curiously, unlike the standard finite element method for the cavity problem, for the DPG method there appears to be no need for h to be sufficiently small to obtain quasioptimality. However, the discrete stability of the DPG method, inherited from the wellposedness, can deteriorate when the exact inf-sup constant γ is poor (which is to be expected as ω approaches a cavity resonance).

Bibliographical notes. The construction of the $H(\mathrm{curl}, K)$ Fortin operator for DPG methods presented here is new, but is related to existing constructions. The first $H(\mathrm{curl}, K)$ Fortin operator was given in Carstensen *et al.* (2016). The construction there uses an appropriate bubble space and is similar in spirit to our constructions of $\Pi^{\mathrm{grad}}_{p+3}$ in Theorem 5.4 and Π^{div}_{p+3} in Theorem 5.8. Another natural method for construction of $\Pi^{\mathrm{curl}}_{p+3}$ is through the constrained minimization (5.58), where the required moment conditions are put as constraints. Such minimizations were used to construct Fortin operators in Demkowicz (2024) and Demkowicz and Zanotti (2020). The construction we have presented here is close but not identical to these, because in proving Theorem 5.9 we needed to establish commutativity between differently constructed Fortin operators. These techniques clearly show there are multiple avenues to construct Fortin operators for DPG schemes.

6. *A posteriori* error control

The DPG method comes with a built-in error estimator. The estimator naturally appears either from a residual minimization standpoint or through a characterization of the method as a mixed method, as revealed in this section. The estimator can be thought of as a hierarchical type error estimator obtained by exploiting test functions that do not contribute to the inexact optimal test space.

6.1. Discrete residual minimization, error estimators, and mixed formulation

Let x be as in (1.1). Consider the DPG method (5.4) for approximating x, obtained using some finite-dimensional spaces X_h and Y^r. Recall that following prior notation, $R_{Y^r} : Y^r \to (Y^r)^*$ denotes the Riesz map defined by $(R_{Y^r} y)(v) = (y, v)_Y$ for all y and v in Y^r. From the definition of the computable trial-to-test operator T^r in (5.2), it is easy to see that

$$T^r w_h = R_{Y^r}^{-1} B w_h, \quad w_h \in X_h. \tag{6.1}$$

Note that $R_{Y^r}^{-1}$ can be applied to Bw_h since it is in $Y^* \subset (Y^r)^*$. For any $x, w \in X$, let

$$(z, w)_r = (T^r z, T^r w)_Y, \quad |z|_r = \|T^r s\|_Y. \tag{6.2}$$

By (5.7), T^r is injective on X_h when a Fortin operator exists, so $|\cdot|_r$ is a norm on X_h. In general, $|\cdot|_r$ is only a seminorm on X. Even so, whenever $|\cdot|_r$ is a norm on the finite-dimensional space X_h, it is easy to see that there exists a unique minimizer x_h in X_h solving

$$x_h = \arg \min_{z_h \in X_h} |x - z_h|_r, \tag{6.3a}$$

which is characterized by

$$(x - x_h, z_h)_r = 0 \quad \text{for all } z_h \in X_h. \tag{6.3b}$$

This minimizer also minimizes a residual in a discrete dual norm and equals the solution of the (practical) DPG method, as stated next.

Theorem 6.1 (Inexact residual minimization). Under the assumptions of Theorem 5.2, the following are equivalent statements.

(a) $x_h \in X_h$ is the unique solution of the DPG method (5.4).

(b) x_h is the unique element of X_h satisfying

$$|x - x_h|_r = \inf_{z_h \in X_h} |x - z_h|_r.$$

(c) x_h minimizes the residual in the following sense:

$$x_h = \arg \min_{z_h \in X_h} \|\ell - Bz_h\|_{(Y^r)^*}.$$

Proof. Follow along the lines of proof of Theorem 3.2 but using (6.1) instead of (3.2) and noting (6.3). □

Definition 6.2. Let ℓ be as in (1.1), let $\tilde{x}_h$ be any element of X_h, and let Y^r be as in (5.1). The element of Y^r defined by

$$\tilde{\varepsilon}^r = R_{Y^r}^{-1}(\ell - B\tilde{x}_h) \tag{6.4}$$

is called the *inexact error representation* of $\tilde{x}_h$ (see Definition 3.3). When $\tilde{x}_h$ is set to the solution x_h of the DPG method (5.4), then its inexact error representation is denoted (omitting the tilde) by $\varepsilon^r = R_{Y^r}^{-1}(\ell - Bx_h)$.

It is easy to see that $\tilde{\varepsilon}^r$ in (6.4) is the unique element of Y^r satisfying

$$(\tilde{\varepsilon}^r, y)_Y = \ell(y) - b(\tilde{x}_h, y) \quad \text{for all } y \in Y^r. \tag{6.5}$$

This shows that the inexact error representation of the DPG solution, namely ε^r, is Y-orthogonal to the entire inexact optimal test space Y_h^r due to (5.4). Let

$$\tilde{\eta} = \|\tilde{\varepsilon}^r\|_Y, \quad \eta = \|\varepsilon^r\|_Y. \tag{6.6}$$

Clearly, (6.1), (6.5) and (6.2) imply

$$\tilde{\eta} = \|R_{Y^r}^{-1} B(x - \tilde{x}_h)\|_Y = \|T^r(x - \tilde{x}_h)\|_Y.$$

When Y^r is of the product form (5.1), the norm in (6.6) can be written in terms of local element contributions, each of which acts as a practically computable element-wise error indicator. It is useful to note the following analogue of Theorems 3.4.

Theorem 6.3 (Inexact error representation as a mixed solution component). Let ε^r denote the inexact error representation of Definition 6.2. Then the following are equivalent statements.

(a) $x_h \in X_h$ solves the DPG method (5.4).

(b) $x_h \in X_h$ and $\varepsilon^r \in Y^r$ solve the mixed formulation

$$(\varepsilon^r, y)_Y + b(x_h, y) = \ell(y) \quad \text{for all } y \in Y^r, \tag{6.7a}$$
$$b(z, \varepsilon^r) = 0 \quad \text{for all } z \in X_h. \tag{6.7b}$$

(c) ε^r and x_h form the saddle point of

$$L(y, z) = \frac{1}{2}\|y\|_Y^2 - \ell(y) + b(z, y)$$

on $Y^r \times X_h$, that is,

$$L(\varepsilon, x_h) = \min_{y \in Y^r} \max_{z \in X_h} L(y, z).$$

Proof. Follow along the lines of the proof of Theorem 3.4. □

The mixed reformulation (6.7) of the DPG method in Theorem 6.3 gives further insight into the stability of the method. In a typical two-equation mixed system, enlarging the test space in the first equation, while often helpful to prove the inf-sup condition by increasing the supremum, is fraught with the danger of losing the coercivity of the first term. However, in the DPG system (6.7), the first term $(\cdot, \cdot)_Y$, being an inner product, will never lose coercivity, no matter how liberally we enrich Y^r. This explains any perceived ease in proving stability of DPG formulations.

6.2. *Reliability and efficiency*

The basis for *a posteriori* error control in DPG methods using η is the following result, proved under the same prior assumption on the existence of a continuous Fortin operator Π.

Theorem 6.4 (Global reliability and efficiency for any approximation). Under the assumptions of Theorem 5.2, we have the following inequalities for the difference between the exact solution x and any $\tilde{x}_h \in X_h$ in terms of the corresponding computable error estimator $\tilde{\eta}$ of (6.6):

$$\gamma\|x - \tilde{x}_h\|_X \le \|\Pi\|\,\tilde{\eta} + \operatorname{osc}(\ell) \quad \text{(reliability)}, \tag{6.8a}$$
$$\tilde{\eta} \le \|b\|\,\|x - \tilde{x}_h\|_X \quad \text{(efficiency)}. \tag{6.8b}$$

Here

$$\operatorname{osc}(\ell) = \|\ell \circ (1 - \Pi)\|_{Y^*} \tag{6.8c}$$

represents a term akin to data-approximation error and it admits the following bound:

$$\operatorname{osc}(\ell) \le \|b\|\,\|1 - \Pi\| \min_{z_h \in X_h} \|x - z_h\|_X. \tag{6.8d}$$

Proof. To prove (6.8a), observe that

$$\begin{aligned} b(x-\tilde{x}_h, y) &= \ell(y) - b(\tilde{x}_h, y) \\ &= \ell(y - \Pi y) - b(\tilde{x}_h, y - \Pi y) + \ell(\Pi y) - b(\tilde{x}_h, \Pi y) \\ &= \ell(y - \Pi y) + (\tilde{\varepsilon}^r, \Pi y)_Y \end{aligned}$$

due to (5.5) and (6.5). Hence (1.2) implies

$$\gamma \|x - \tilde{x}_h\|_X \le \sup_{0 \ne y \in Y} \frac{|b(x-\tilde{x}_h, y)|}{\|y\|_Y} = \sup_{0 \ne y \in Y} \frac{|\ell \circ (1-\Pi)(y) + (\tilde{\varepsilon}^r, \Pi y)_Y|}{\|y\|_Y},$$

from which (6.8a) follows.

The global efficiency estimate is immediate from (6.5):

$$\tilde{\eta}^2 = b(x - \tilde{x}_h, \tilde{\varepsilon}^r) \le \|b\| \, \|x - \tilde{x}_h\|_X \, \tilde{\eta}.$$

Finally, to prove (6.8d),

$$\begin{aligned} (\ell \circ (1-\Pi))(y) &= \ell(y) - \ell(\Pi y) = b(x, y - \Pi y) \\ &= b(x - z_h, y - \Pi y) \le \|b\| \, \|x - z_h\|_X \, \|y - \Pi y\|_Y \end{aligned}$$

for any $z_h \in X_h$, where we used (5.5) to get the last equality. Taking the infimum over $z_h \in X_h$ and supremum over $0 \ne y \in Y$, we obtain (6.8d). □

Example 6.5 (Case of $Y^r \supseteq Y_h^{\mathrm{opt}}$). Reconsider the setting of Example 5.3 for some $Y^r \supseteq Y_h^{\mathrm{opt}}$. As shown there, the solution $x_h \in X_h$ of the IPG method and the practical DPG method are identical. We also showed there that the Fortin condition holds with Π set to the Y-orthogonal projection of Π_{Y^r} into Y^r. Hence Theorem 6.4 applies with $\|\Pi\| = \|\Pi_{Y^r}\| = 1$, so

$$\gamma \|x - x_h\|_X \le \eta + \operatorname{osc}(\ell), \quad \eta \le \|b\| \|x - x_h\|_X. \tag{6.9}$$

Here

$$\operatorname{osc}(\ell) = \|\ell \circ (I - \Pi_{Y^r})\|_{Y^*} \tag{6.10}$$

following its definition in (6.8c).

It is interesting to compare the exact and inexact error representations, $\varepsilon \in Y$ and $\varepsilon^r \in Y^r$ respectively, in this example. Consider the mixed method reformulations of the IPG and DPG methods, namely (3.4) and (6.7) respectively. Since x_h is the same in both cases, choosing y of (3.4a) in Y^r and subtracting (6.7a), we find that

$$(\varepsilon - \varepsilon^r, y)_Y = 0 \quad \text{for all } y \in Y^r, \tag{6.11}$$

that is, $\varepsilon^r = \Pi \varepsilon$ is the Y-orthogonal projection of the exact error representation. We claim that

$$\|\varepsilon^r\|_Y \le \|\varepsilon\|_Y \le \|\varepsilon^r\|_Y + \operatorname{osc}(\ell)^2. \tag{6.12}$$

The first inequality above is immediate from $\varepsilon^r = \Pi \varepsilon$. The second inequality

follows from the Pythagoras theorem $\|\varepsilon\|_Y^2 = \|\varepsilon^r\|_Y^2 + \|\varepsilon - \varepsilon^r\|_Y^2$ and

$$\begin{aligned} \|\varepsilon - \varepsilon^r\|_Y^2 &= (\varepsilon, \varepsilon - \varepsilon^r)_Y && \text{by (6.12)} \\ &= \ell(\varepsilon - \varepsilon^r) - b(x_h, \varepsilon - \varepsilon^r) && \text{by (3.4a)} \\ &= \ell(\varepsilon - \varepsilon^r) && \text{by (6.7b) and (3.4b).} \end{aligned}$$

Clearly

$$\ell(\varepsilon - \varepsilon^r) = \ell(\varepsilon - \Pi\varepsilon) \le \|\ell \circ (1 - \Pi)\|_{Y^*} \|\varepsilon - \varepsilon^r\|_Y = \operatorname{osc}(\ell)\|\varepsilon - \varepsilon^r\|_Y,$$

and the upper inequality of (6.12) follows.

Bibliographical notes. The proof given for Theorem 6.4 is a slightly simplified version of the original presented in Carstensen, Demkowicz and Gopalakrishnan (2014*a*, Theorem 2.1), and produces an improved reliability constant as in Carstensen, Gallistl, Hellwig and Weggler (2014*b*, Lemma 3.6). If in addition Π is idempotent, then (6.8a) can be further improved to

$$\gamma^2 \|x - x_h\|_X^2 \le \eta^2 + \left(\eta\sqrt{\|\Pi\|^2 - 1} + \operatorname{osc}(\ell)\right)^2, \tag{6.13}$$

as shown in Keith, Astaneh and Demkowicz (2019, Theorem 6.4). Note the relationship between (6.12) and (6.13) when Π is an orthogonal projection. It is easy to construct adaptive algorithms with marking strategies based on the DPG error estimator following the usual 'Solve → Estimate → Mark → Refine' paradigm. In all reports of practical performance (Demkowicz *et al.* 2012*a*, Petrides and Demkowicz 2017), such DPG algorithms work very well, but to the best of our knowledge their convergence and optimality are yet to be rigorously proved.

7. Ultraweak formulations

A rich set of examples to apply the DPG ideas is offered by the so-called 'ultraweak' formulations of boundary value problems seeking $u \in V$ satisfying $Au = f$ for an $f \in L^2(\Omega)^m$. Here V is a space where homogeneous boundary conditions are imposed, and A is a general partial differential operator (specified below). *In ultraweak formulations all derivatives in A are moved to test functions by integration by parts, element by element.* In order to use the previously developed DPG ideas, it is important to obtain a reformulation where the trial-to-test operator T is localized, i.e. a formulation where the test space has the form (4.1). Such a reformulation, set in a broken graph space, is derived and studied in this section. We prove its wellposedness by general arguments that cover many examples at once. The first main result (Theorem 7.6) of this section identifies conditions under which the wellposedness of ultraweak formulations in broken graph spaces can be obtained as soon as $A\colon V \to L^2(\Omega)^m$ is a bijection, no matter how complex the spaces of interface variables are. Another result of this section (Theorem 7.9), which has not appeared in previous literature, exhibits norms in which the best possible stability of ultraweak formulations can be obtained.

Let $k, m, N \geq 1$ be integers, let $\Omega \subseteq \mathbb{R}^N$ and Ω_h be as in Section 4, let e_i denote the standard Euclidean unit basis in $\mathbb{R}^N$, $w = \sum_{i=1}^m w_i e_i : \Omega \to \mathbb{C}^m$ for some smooth functions w_i, and let A be the partial differential operator

$$Aw = \sum_{i,j=1}^{m} \sum_{\alpha \in \mathcal{I}_k^N} e_i \, \partial^\alpha (a_{ij\alpha} w_j) \tag{7.1}$$

for some functions $a_{ij\alpha} : \Omega \to \mathbb{C}$ indexed by $i, j = 1, \ldots, m$, and multi-indices $\alpha \in \mathcal{I}_k^N$ (defined in (5.29)). As usual, $\partial^\alpha = \partial_1^{\alpha_1} \cdots \partial_N^{\alpha_N}$. We view A as an unbounded operator $A : \operatorname{dom}(A) \subseteq L^2(\Omega)^m \to L^2(\Omega)^m$. Given an $f \in L^2(\Omega)^m$, we want to

$$\text{find } u \in \operatorname{dom}(A) \text{ such that } Au = f, \tag{7.2}$$

where homogeneous boundary conditions we wish to impose are incorporated into functions in the subspace $\operatorname{dom}(A)$. At this point, the coefficients $a_{ij\alpha}$ are allowed to be general so long as the result of applying A is a Schwartz distribution, that is, we assume that

$$Au \in \mathcal{D}'(\Omega)^m \quad \text{for any } u \in L^2(\Omega)^m. \tag{7.3}$$

Of course, when u is in $\operatorname{dom}(A)$, Au is not merely a distribution, but is in $L^2(\Omega)^m$.

7.1. Graph spaces, boundary operators and their broken versions

Assume that the Schwartz space $\mathcal{D}(\Omega)^m$ of smooth compactly supported test functions in Ω is contained in the domain of A:

$$\mathcal{D}(\Omega)^m \subseteq \operatorname{dom}(A). \tag{7.4}$$

A consequence of (7.4) is that A is densely defined, so its (maximal) adjoint A^* is uniquely defined as follows (see e.g. Brezis 2011 or Kato 1995). First define the set $\operatorname{dom}(A^*) \subseteq L^2(\Omega)^m$ by

$$\begin{aligned} \operatorname{dom}(A^*) = \{ g \in L^2(\Omega)^m : \text{there is an } f \in L^2(\Omega)^m \text{ such that} \\ (g, Au)_\Omega = (f, u)_\Omega \text{ for all } u \in \operatorname{dom}(A) \}. \end{aligned} \tag{7.5}$$

Then define $A^* : \operatorname{dom}(A^*) \subseteq L^2(\Omega)^m \to L^2(\Omega)^m$ by

$$(A^* g, u)_\Omega = (g, Au)_\Omega \quad \text{for all } u \in \operatorname{dom}(A) \text{ and } g \in \operatorname{dom}(A^*). \tag{7.6}$$

By virtue of assumption (7.3), for any $u \in L^2(\Omega)$, the distribution Au is such that its action on a $\tilde{\varphi} \in \mathcal{D}(\Omega)^m$ takes the form

$$(Au)(\tilde{\varphi}) = \sum_{i=1}^{m} \sum_{\alpha \in \mathcal{I}_k^N} (-1)^{|\alpha|} (u, \, a_{ji\alpha} \, \partial^\alpha \tilde{\varphi}_j)_\Omega. \tag{7.7}$$

For any $u \in \operatorname{dom}(A)$, since Au is in $L^2(\Omega)^m$, the left-hand side above equals

$(Au, \tilde{\varphi})_\Omega$. Hence the condition in (7.5) is verified with $\tilde{\varphi}$ in place of g, that is,

$$\mathcal{D}(\Omega)^m \subseteq \mathrm{dom}(A^*). \tag{7.8}$$

In view of (7.4), (7.8) and (7.6), we can now identify A^* with the formal adjoint partial differential operator

$$A^*\tilde{\varphi} = \sum_{i,j=1}^{m} \sum_{\alpha \in \mathcal{I}_k^N} e_i \, (-1)^{|\alpha|} \overline{a_{ji\alpha}} \, \partial^\alpha \tilde{\varphi}_j. \tag{7.9}$$

To circumvent issues concerning products of distributions and non-smooth functions, we assume that the application of this formal adjoint satisfies

$$A^*\tilde{u} \in \mathcal{D}'(\Omega)^m \quad \text{for any } \tilde{u} \in L^2(\Omega)^m, \tag{7.10}$$

analogous to (7.3).

Note that (7.3) and (7.10) imply that we may restrict Au and $A^*\tilde{u}$ to any non-empty open subset $S \subseteq \Omega$ to get distributions in $\mathcal{D}'(S)$. This allows us to define the following *graph spaces* on S:

$$\begin{aligned} W(S) &= \{u \in L^2(S)^m : Au \in L^2(S)^m\}, \qquad \|w\|^2_{W(S)} = \|w\|^2_S + \|Aw\|^2_S, \\ \tilde{W}(S) &= \{u \in L^2(S)^m : A^*u \in L^2(S)^m\}, \qquad \|\tilde{w}\|^2_{\tilde{W}(S)} = \|\tilde{w}\|^2_S + \|A^*\tilde{w}\|^2_S. \end{aligned}$$

Here $\|\cdot\|_S$ denotes the norm of $L^2(S)^m$; the corresponding inner product is denoted $(\cdot,\cdot)_S$. Our assumptions imply that these inner product spaces are complete.

Lemma 7.1. The spaces $W(S)$ and $\tilde{W}(S)$ are Hilbert spaces.

Proof. Consider a Cauchy sequence u_n in $W(S)$. Clearly, u_n is Cauchy in $L^2(S)^m$ and Au_n is Cauchy in $L^2(S)^m$. Hence there is a $u \in L^2(S)^m$ and $f \in L^2(S)^m$ such that $\|u - u_n\|_S \to 0$ and $\|f - Au_n\|_S \to 0$. To show that u is in $W(S)$, we use (7.7), a consequence of assumption (7.3), to get that for any $\tilde{\varphi} \in \mathcal{D}(S)^m$, $(Au_n)(\tilde{\varphi}) = (u_n, A^*\tilde{\varphi})_S \to (u, A^*\tilde{\varphi})_S = (Au)(\tilde{\varphi})$ as $n \to \infty$. Since $Au_n \to f$ in $L^2(\Omega)^m$, this implies that the distribution Au must equal f in $L^2(\Omega)^m$. This proves the completeness of $W(S)$. The completeness of $\tilde{W}(S)$ is similarly proved by using assumption (7.10) in place of (7.3). □

Next, we need *boundary operators*, which are bounded linear operators

$$D_S : W(S) \to \tilde{W}(S)^* \quad \text{and} \quad \tilde{D}_S : \tilde{W}(S) \to W(S)^*$$

defined by

$$\begin{aligned} \langle D_S w, \tilde{w} \rangle_{\tilde{W}(S)} &= (Aw, \tilde{w})_S - (w, A^*\tilde{w})_S, \\ \langle \tilde{D}_S \tilde{w}, w \rangle_{W(S)} &= (A^*\tilde{w}, w)_S - (\tilde{w}, Aw)_S, \end{aligned} \tag{7.11}$$

for all $w \in W(S)$ and $\tilde{w} \in \tilde{W}(S)$. Obviously $\langle D_S w, \tilde{w} \rangle_{\tilde{W}(S)}$ is obtained by conjugating $\langle \tilde{D}_S \tilde{w}, w \rangle_{W(S)}$ and changing sign, but note that the domains and codomains of D_S and $\tilde{D}_S$ are different.

When $S = \Omega$, we abbreviate $W(S)$, $\tilde{W}(S)$, D_S and $\tilde{D}_S$ to W, $\tilde{W}$, D and $\tilde{D}$ respectively. Furthermore, let V and $\tilde{V}$ denote the linear subspaces $\mathrm{dom}(A)$ and $\mathrm{dom}(A^*)$ made into normed spaces using $\|\cdot\|_W$ and $\|\cdot\|_{\tilde{W}}$ respectively, that is,

$$V = (\mathrm{dom}(A),\ \|\cdot\|_W), \quad \tilde{V} = (\mathrm{dom}(A^*),\ \|\cdot\|_{\tilde{W}}). \tag{7.12}$$

Clearly $V \subset W$ and $\tilde{V} \subset \tilde{W}$. Given any subspace R of the dual space X^*, we denote its left annihilator by ${}^{\perp}R = \{w \in X\colon \langle s', w\rangle_X = 0 \text{ for all } s' \in R\}$.

Lemma 7.2. The space $DV = \{\tilde{v} \in \tilde{W}^*\colon \tilde{v} = Dv \text{ for some } v \in V\}$ satisfies

$$\tilde{V} = {}^{\perp}DV\,. \tag{7.13}$$

Proof. If $\tilde{v} \in \tilde{V} = \mathrm{dom}(A^*)$, then for any $v \in V = \mathrm{dom}(A)$, by the definition of the adjoint (7.6), we have $\langle Dv, \tilde{v}\rangle_{\tilde{W}} = (Av, \tilde{v})_\Omega - (v, A^*\tilde{v})_\Omega = 0$, so $\tilde{V} \subseteq {}^{\perp}DV$.

For the reverse inclusion, let

$$\tilde{w} \in {}^{\perp}DV = \{\tilde{w} \in \tilde{W}\colon \langle Dv, \tilde{w}\rangle_{\tilde{W}} = 0 \text{ for all } v \in V\}.$$

Then $f = A^*\tilde{w} \in L^2(\Omega)^m$ satisfies $(v, f)_\Omega - (Av, \tilde{w})_\Omega = -\langle Dv, \tilde{w}\rangle_{\tilde{W}} = 0$ for all $v \in V = \mathrm{dom}(A)$, so given the definition of $\mathrm{dom}(A^*)$ in (7.5), $\tilde{w}$ must be in $\tilde{V} = \mathrm{dom}(A^*)$. □

For our wellposedness theorems later, we need to place an assumption which represents an equality analogous to Lemma 7.2 but with the roles of $\tilde{V}$ and V reversed, namely

$$V = {}^{\perp}\tilde{D}\tilde{V}\,. \tag{7.14}$$

In applications, (7.14) being a constraint on $V = \mathrm{dom}(A)$ restricts admissible boundary conditions in (7.2), as in the theory of Friedrichs systems. Note that (7.14) implies that V is a closed subspace of W. Hence, for (7.14) to hold it is necessary for A to be a closed operator in $L^2(\Omega)^m$. Similarly, (7.13) of Lemma 7.2 implies, in particular, that $\tilde{V}$ is closed in $\tilde{W}$, in agreement with the fact that the adjoint A^* is a closed operator.

Next we use the mesh Ω_h to define *broken graph spaces* (which are generally infinite-dimensional) by

$$W_h = \prod_{K\in\Omega_h} W(K) \quad \text{and} \quad \tilde{W}_h = \prod_{K\in\Omega_h} \tilde{W}(K). \tag{7.15}$$

For any $w \in W_h$, as in (4.2), letting $w|_K \equiv w_K$ denote the component of the product function w on K, we recall that $D_K w_K$ is in $\tilde{W}(K)^*$. Let $D_h\colon W_h \to \tilde{W}_h^*$ be the continuous linear operator defined by

$$\langle D_h w, \tilde{w}\rangle_{\tilde{W}_h} = \sum_{K\in\Omega_h} \langle D_K w_K, \tilde{w}_K\rangle_{\tilde{W}_K} \quad \text{for all } w \in W_h, \tilde{w} \in \tilde{W}_h$$

and let $\tilde{D}_h\colon \tilde{W}_h \to W_h^*$ be defined by

$$\langle \tilde{D}_h\tilde{w}, w\rangle_{W_h} = \overline{\langle D_h w, \tilde{w}\rangle}_{\tilde{W}_h}.$$

For any $w \in W_h$, we let $A_h w$ denote the function obtained by applying A to w_K, element by element, for all $K \in \Omega_h$. The resulting function $A_h w$ may be viewed as an element of the Cartesian product $\Pi_{K\in\Omega_h} L^2(K)^m$, which can obviously be embedded in $L^2(\Omega)^m$. This defines the map $A_h : W_h \to L^2(\Omega)^m$. The operator $A_h^* : \tilde{W}_h \to L^2(\Omega)^m$ is defined similarly by evaluating the action of A^* (instead of A) element by element. Clearly

$$\langle D_h w, \tilde{w}\rangle_{\tilde{W}_h} = (A_h w, \tilde{w})_\Omega - (w, A_h^* \tilde{w})_\Omega \quad \text{for all } w \in W_h,\ \tilde{w} \in \tilde{W}_h. \tag{7.16}$$

The obvious norm of the Cartesian products defining W_h and $\tilde{W}_h$ can now be equivalently written in terms of A_h and A_h^*:

$$\|w\|_{\tilde{W}}^2 = \|w\|_\Omega^2 + \|A_h w\|_\Omega^2, \quad \|\tilde{w}\|_{\tilde{W}_h}^2 = \|\tilde{w}\|_\Omega^2 + \|A_h^* \tilde{w}\|_\Omega^2. \tag{7.17}$$

Lemma 7.3. For all $w \in W$ and $\tilde{w} \in \tilde{W}$, we have

$$\langle D_h w, \tilde{w}\rangle_{\tilde{W}_h} = \langle Dw, \tilde{w}\rangle_{\tilde{W}}.$$

Proof. Since piecewise differential operators coincide with the global ones when applied to functions in the unbroken spaces, $A_h w = Aw$ and $A_h^* \tilde{w} = A^* \tilde{w}$. Therefore (7.16) implies

$$\langle D_h w, \tilde{w}\rangle_{\tilde{W}_h} = (Aw, \tilde{w})_\Omega - (w, A^* \tilde{w})_\Omega = \langle D_\Omega w, \tilde{w}\rangle_{\tilde{W}},$$

where the last equality followed from (7.11). □

Lemma 7.4. The equality (7.14) implies that any $w \in W_h$ satisfying $D_h w = 0$ is in V. Similarly, any $\tilde{w} \in \tilde{W}_h$ satisfying $\tilde{D}_h \tilde{w} = 0$ is in $\tilde{V}$. In fact,

$$\tilde{V} = \{\tilde{w} \in \tilde{W}_h : \langle D_h z, \tilde{w}\rangle_{\tilde{W}_h} = 0 \text{ for all } z \in V\}. \tag{7.18}$$

Proof. Let us prove (7.18) first. If $\tilde{v} \in \tilde{V}$, then for any $z \in V$, using Lemmas 7.3 and 7.2, we have

$$\langle D_h z, \tilde{v}\rangle_{\tilde{W}_h} = \langle Dz, \tilde{v}\rangle_{\tilde{W}} = 0.$$

Thus $\tilde{V}$ is contained in the set on the right-hand side of (7.18).

To prove the reverse inclusion, consider a $\tilde{w} \in \tilde{W}_h$ satisfying $\langle D_h z, \tilde{w}\rangle_{\tilde{W}_h} = 0$ for all $z \in V$. Then

$$0 = \langle D_h z, \tilde{w}\rangle_{\tilde{W}_h} = (A_h z, \tilde{w})_\Omega - (z, A_h^* \tilde{w}) \quad \text{for all } z \in V. \tag{7.19}$$

Since $z \in V \subset W$, we can replace $A_h z$ by Az above. In fact (7.19) implies that $A_h^* \tilde{w}$ also equals $A^* \tilde{w}$, because $\tilde{w}$ is in $\tilde{W}$, as we now show. The action of the distribution $A^* \tilde{w}$ on any φ in $\mathcal{D}(\Omega)^m$ satisfies

$$\overline{(A^* \tilde{w})(\varphi)} = (A\varphi, \tilde{w})_\Omega = (A_h \varphi, \tilde{w})_\Omega = (\varphi, A_h^* \tilde{w})_\Omega + \langle D_h \varphi, \tilde{w}\rangle_{\tilde{W}_h}.$$

The last term must vanish because of the first equality of (7.19) and because $\varphi \in \mathcal{D}(\Omega)^m \subseteq \mathrm{dom}(A) = V$ by our assumption (7.4). Hence

$$|(A^* \tilde{w})(\varphi)| \le \|\varphi\|_\Omega \|A_h^* \tilde{w}\|_\Omega \quad \text{for all } \varphi \in \mathcal{D}(\Omega)^m,$$

which implies that the distribution $A^*\tilde{w}$ is in $L^2(\Omega)^m$. Therefore $\tilde{w} \in \tilde{W}$. Returning to (7.19), we find that

$$\begin{aligned} 0 &= (A_h z, \tilde{w})_\Omega - (z, A_h^* \tilde{w}) \\ &= (Az, \tilde{w})_\Omega - (z, A^*\tilde{w}) = \langle Dz, \tilde{w} \rangle_{\tilde{W}} \end{aligned}$$

for all $z \in V$. Hence $\tilde{w} \in {}^\perp DV$. By (7.13) of Lemma 7.2, we then conclude that $\tilde{w} \in \tilde{V}$, thus proving (7.18).

The proof of the second statement immediately follows from (7.18). Indeed, any $\tilde{w} \in \tilde{W}_h$ that is in the null space of $\tilde{D}_h$ satisfies $\langle D_h w, \tilde{w} \rangle_{\tilde{W}_h} = 0$ for all $w \in W_h$ due to the relationship between $\tilde{D}_h$ and D_h, so in particular (7.19) holds for all $z \in V$. Hence (7.18) implies that $\tilde{w}$ is in $\tilde{V}$.

The proof of the first statement proceeds similarly, but using (7.14) in place of (7.13). □

Bibliographical notes. Graph spaces of first-order differential operators are a classical ingredient in the theory of Friedrichs systems (Friedrichs 1958). More recently they were studied in Sheen (1992), Jensen (2004) and Ern, Guermond and Caplain (2007). Completeness and density results were proved in Jensen (2004), where one also finds the term 'broken graph space' in the context of DG methods. Analogues of our twin equalities of (7.13) and (7.14), namely $\tilde{V} = {}^\perp DV$ and $V = {}^\perp \tilde{D}\tilde{V}$, prominently feature as abstract conditions in modern takes on the theory of first-order Friedrichs systems (Ern *et al.* 2007). Our presentation here, which is not restricted to first-order operators, is based on Demkowicz, Gopalakrishnan, Nagaraj and Sepúlveda (2017).

7.2. *Hybrid ultraweak formulation suitable for DPG method*

Now that we have broken graph spaces $W_h, \tilde{W}_h$ and elementwise boundary operators $D_h, \tilde{D}_h$, we can perform elementwise operations analogous to performing integration by parts and moving all derivatives to the test functions. Namely, we derive an ultraweak formulation by multiplying the equation $Au = f$ by a test function $\tilde{w} \in \tilde{W}_h$, applying the definition of D_K on each element K, and summing over all $K \in \Omega_h$. Then we obtain

$$(u, A_h^* \tilde{w})_\Omega + \langle D_h u, \tilde{w} \rangle_{\tilde{W}_h} = \langle f, \tilde{w} \rangle_{\tilde{W}_h} \tag{7.20}$$

for any $\tilde{w}$ in $\tilde{W}_h$. Now, since D_h is not an injective operator in general, we define $\hat{q} = D_h u$, an unknown that we want to uniquely solve for. Note that W is contained in W_h, so V can be viewed as a subspace of the broken graph space W_h. Let $D_{h,V} = D_h|_V : V \to \tilde{W}_h^*$ denote the restriction of D_h from W_h to V. Analogous to the quotient norms that appeared earlier (such as (4.10) and (4.31)), we define

$$Q = \operatorname{range}(D_{h,V}), \quad \|\hat{q}\|_Q = \inf_{v \in D_{h,V}^{-1}\{\hat{q}\}} \|v\|_W, \tag{7.21}$$

that is, the space

$$Q = \{\hat{q} \in \tilde{W}_h^* \colon \text{there is a } v \in V \text{ satisfying } \hat{q} = D_h v\}$$

is endowed with the minimal norm of elements in the preimage set

$$D_{h,V}^{-1}\{\rho\} = \{v \in V \colon D_h v = \rho\},$$

a quotient norm that makes Q into a Hilbert space.

Using $\hat{q}$ in (7.20), we have completed the derivation of the following *(hybrid) ultraweak formulation* of (7.2): find $u \in L^2(\Omega)^m$ and $\hat{q} \in Q$ such that

$$(u, A_h^* \tilde{w})_\Omega + \langle \hat{q}, \tilde{w} \rangle_{\tilde{W}_h} = F(\tilde{w}) \quad \text{for all } \tilde{w} \in \tilde{W}_h, \tag{7.22}$$

where $F \in \tilde{W}_h^*$ is set by $F(\tilde{w}) = (f, \tilde{w})_\Omega$. This formulation can be viewed as a hybridized version of another with the unbroken graph space $\tilde{W}$ as the test space. Indeed, multiplying $Au = f$ with $\tilde{w}_0 \in \tilde{W}$ and using the definition of $D = D_\Omega$ (see (7.11)), we find that $u \in V$ solves $(u, A^* \tilde{w}_0)_\Omega + \langle Du, \tilde{w}_0 \rangle_{\tilde{W}} = F(\tilde{w}_0)$ for all $\tilde{w}_0 \in \tilde{W}$. By Lemma 7.2, $\langle Du, \tilde{w}_0 \rangle_{\tilde{W}} = 0$ is $\tilde{w}_0$ is in $\tilde{V}$. Hence restricting to test functions $\tilde{v} \in \tilde{V} \subset \tilde{W}$, we obtain an 'unbroken ultraweak formulation' that finds $u \in V$ such that

$$(u, A^* \tilde{v})_\Omega = F(\tilde{v}) \quad \text{for all } \tilde{v} \in \tilde{V}. \tag{7.23}$$

Comparing with the formulation in (7.22), we see the hybrid ultraweak formulation with broken graph spaces in (7.22) as a hybrid version of the unbroken ultraweak formulation (7.23) obtained by introducing an 'interface variable', which has now taken the abstract form of $\hat{q} \in \tilde{W}_h^*$.

This suggests that the stability of hybrid ultraweak formulation may follow from that of the unbroken ultraweak formulation (7.23) if we can verify the conditions of (4.12) and apply Theorem 4.3. The work to make this rigorous is completed in Theorem 7.6 below, whose proof contains a proof of the stability of the unbroken ultraweak formulation (7.23), as well techniques to handle the element interface terms to conclude the stability of the hybrid version (7.22). To fit into the setting of (4.12), put

$$\begin{aligned} &X_0 = V, && Y_0 = \tilde{V}, \\ &\hat{X} = Q, && Y = \tilde{W}_h, \\ &b_0(u, \tilde{w}) = (u, A_h^* \tilde{w})_\Omega, && \hat{b}(\hat{q}, \tilde{w}) = \langle \hat{q}, \tilde{w} \rangle_{\tilde{W}_h}. \end{aligned}$$

The sum of the above forms,

$$b((u, \hat{q}), \tilde{w}) = b_0(u, \tilde{w}) + \hat{b}(\hat{q}, \tilde{w}) = (u, A_h^* \tilde{w})_\Omega + \langle \hat{q}, \tilde{w} \rangle_{\tilde{W}_h},$$

is the ultraweak form in (7.22). We proceed to verify the conditions of (4.12) with the above choices of spaces and forms. The next lemma generalizes an argument we previously used to prove the 'inf = sup'-type interface duality identities (of Theorem 4.6) to the present scenario.

Lemma 7.5. Assumption (7.14) implies that for all $\hat{q} \in Q$,

$$\inf_{v \in D_{h,V}^{-1}\{\hat{q}\}} \|v\|_W = \sup_{0 \neq \tilde{w} \in \tilde{W}_h} \frac{|\langle \hat{q}, \tilde{w} \rangle_{\tilde{W}_h}|}{\|\tilde{w}\|_{\tilde{W}_h}}.$$

Proof. We use two functions, $\tilde{u}_{\hat{q}}$ and $u_{\hat{q}}$, both obtained from $\hat{q}$, but by solving two distinct boundary value problems. First, the supremum of the lemma, which we denote by s, is attained by the function $\tilde{u}_{\hat{q}}$ in $\tilde{W}_h$ satisfying

$$(A_h^* \tilde{u}_{\hat{q}}, A_h^* \tilde{w})_\Omega + (\tilde{u}_{\hat{q}}, \tilde{w})_\Omega = -\langle \hat{q}, \tilde{w} \rangle_{\tilde{W}_h} \tag{7.25}$$

for all $\tilde{w} \in \tilde{W}_h$, and moreover, it equals

$$s = \|\tilde{u}_{\hat{q}}\|_{\tilde{W}_h}. \tag{7.26}$$

Note that $\hat{q} = D_h z$ for some $z \in V$. Hence

$$-\langle \hat{q}, \tilde{\varphi} \rangle_{\tilde{W}_h} = -\langle D_h z, \tilde{\varphi} \rangle = -(A_h z, \tilde{\varphi}) + (z, A_h^* \tilde{\varphi}) = -(Az, \tilde{\varphi}) + (z, A^* \tilde{\varphi}) = 0$$

due to (7.6), since $z \in \mathrm{dom}(A)$ and $\tilde{\varphi} \in \mathrm{dom}(A^*)$ by assumption (7.8). Therefore, choosing $y = \tilde{\varphi} \in \mathcal{D}(\Omega)^m$ in (7.25), the right-hand side vanishes, and we conclude that the distribution $A(A_h^* \tilde{u}_{\hat{q}})$ is in $L^2(\Omega)^m$ and equals $-\tilde{u}_{\hat{q}}$. Hence (7.16) is applicable with $w = A_h^* \tilde{u}_{\hat{q}}$ and we obtain

$$A A_h^* \tilde{u}_{\hat{q}} + \tilde{u}_{\hat{q}} = 0, \tag{7.27a}$$

$$D_h A_h^* \tilde{u}_{\hat{q}} = \hat{q}. \tag{7.27b}$$

Let $u_{\hat{q}} = A_h^* \tilde{u}_{\hat{q}}$. Then (7.27a) implies $A u_{\hat{q}} = -\tilde{u}_{\hat{q}}$, which implies $A_h^* A u_{\hat{q}} = -A_h^* \tilde{u}_{\hat{q}} = -u_{\hat{q}}$. Combining with (7.27b), we conclude that $u_{\hat{q}}$ solves

$$A_h^* A u_{\hat{q}} + u_{\hat{q}} = 0, \tag{7.28a}$$

$$D_h u_{\hat{q}} = \hat{q}. \tag{7.28b}$$

Since $A u_{\hat{q}}$ is in $L^2(\Omega)$ by (7.27a), we know that $u_{\hat{q}}$ is in W. Let us now prove that $u_{\hat{q}}$ is actually in V. By assumption (7.14), it suffices to prove that $u_{\hat{q}}$ is in ${}^\perp \tilde{D} \tilde{V}$. For any $\tilde{v}$ in $\tilde{V}$, Lemma 7.3 implies

$$\langle D u_{\hat{q}}, \tilde{v} \rangle_{\tilde{W}} = \langle D_h u_{\hat{q}}, \tilde{v} \rangle_{\tilde{W}_h} = \langle \hat{q}, \tilde{v} \rangle_{\tilde{W}_h} = \langle D_h z, \tilde{v} \rangle_{\tilde{W}_h} = \langle D z, \tilde{v} \rangle_{\tilde{W}}.$$

The last term is zero by Lemma 7.2. Hence $u_{\hat{q}}$ is in ${}^\perp \tilde{D} \tilde{V} = V$.

Thus $u_{\hat{q}}$ is in the set $D_{h,V}^{-1}\{q\}$ over which the infimum of the lemma is taken. We claim that the infimum of the lemma is achieved by $u_{\hat{q}}$. Standard variational arguments show that the infimum is attained by a unique minimizer $v_{\hat{q}} \in V$ satisfying $D_h v_{\hat{q}} = \hat{q}$ and $(A_h v_{\hat{q}}, A_h v)_\Omega + (v_{\hat{q}}, v)_\Omega = 0$ for all $v \in D_{h,V}^{-1}\{0\}$. Choosing a v in $\mathcal{D}(K)^m$, whose extension by zero is in $D_{h,V}^{-1}\{0\}$, we conclude that distribution $A^*(A_h v_{\hat{q}})|_K$ is in $L^2(K)^m$ for any $K \in \Omega_h$. Therefore $A_h^* A_h v_{\hat{q}}$ is in $L^2(\Omega)^m$ and satisfies (7.28), so $v_{\hat{q}} = u_{\hat{q}}$.

To complete the proof, it now suffices to show that

$$\|u_{\hat{q}}\|_W = \|\tilde{u}_{\hat{q}}\|_{\tilde{W}_h}, \tag{7.29}$$

since the left-hand side equals the infimum, as we just established, and the right-hand side equals the supremum by (7.26). But (7.29) is obvious from $u_{\hat{q}} = A_h^* \tilde{u}_{\hat{q}}$ and $A u_{\hat{q}} = -\tilde{u}_{\hat{q}}$. □

Theorem 7.6 (Wellposedness of the hybrid ultraweak formulation). Let A be the partial differential operator in (7.1) satisfying (7.3), (7.4) and (7.10). If, in addition, (7.14) holds and

$$A : V \to L^2(\Omega)^m \text{ is a bijection,} \tag{7.30}$$

then the ultraweak formulation (7.22) is wellposed. Moreover, if $F(v) = (f, v)_\Omega$ for some $f \in L^2(\Omega)^m$, then the unique u and $\hat{q}$ satisfying (7.22) are such that u solves (7.2), u is in V, and $\hat{q}$ satisfies $\hat{q} = D_h u$.

Proof. To apply Theorem 4.3 for the current setting (7.24), we need to verify its conditions (4.12b) and (4.12c). In the present case, this task requires us to prove that there are positive constants $c_0, \hat{c}$ such that

$$c_0 \|u\|_\Omega \le \sup_{0 \neq y_0 \in Y_0} \frac{|(u, A_h^* y_0)_\Omega|}{\|y_0\|_{\tilde{W}_h}} \quad \text{for all } u \in L^2(\Omega)^m, \tag{7.31}$$

$$\hat{c}\, \|q\|_Q \le \sup_{0 \neq y \in \tilde{W}_h} \frac{|\langle q, \tilde{w} \rangle_{\tilde{W}_h}|}{\|\tilde{w}\|_{\tilde{W}_h}} \quad \text{for all } q \in Q, \tag{7.32}$$

where

$$Y_0 = \{ y \in \tilde{W}_h : \langle \hat{r}, y \rangle_{\tilde{W}_h} = 0 \text{ for all } \hat{r} \in Q \}.$$

By (7.18) of Lemma 7.4 and the definition of Q, we have

$$\tilde{V} = \{ \tilde{w} \in \tilde{W}_h : \langle \hat{r}, \tilde{w} \rangle_{\tilde{W}_h} = 0 \text{ for all } \hat{r} \in Q \},$$

that is,

$$Y_0 = \tilde{V}. \tag{7.33}$$

Since (7.32) follows with $\hat{c} = 1$ from Lemma 7.5, we focus on (7.31), which amounts to proving the stability of the unbroken ultraweak formulation (7.23).

As already noted, (7.14) implies that A is a closed operator. By the Closed Range Theorem for closed operators, if $A : \operatorname{dom} A \to L^2(\Omega)^m$ is a bijection, then $A^* : \operatorname{dom}(A^*) \to L^2(\Omega)^m$ is also a bijection. Hence, assumption (7.30) and (7.33) imply that $A^* : Y_0 \to L^2(\Omega)^m$ is a bijection. Thus there is a constant $c > 0$ such that $c\|y\|_\Omega \le \|A^* y\|_\Omega$ for all $y \in Y_0$. Moreover, given any $u \in L^2(\Omega)^m$, there is a $y_u \in Y_0$ such that $A^* y_u = u$. Hence the supremum in (7.31), for any given u, admits the bound

$$\sup_{0 \neq y_0 \in Y_0} \frac{|(u, A_h^* y_0)_\Omega|}{\|y_0\|_{\tilde{W}_h}} \ge \frac{\|A^* y_u\|_\Omega^2}{(c^2+1)^{1/2} \|A^* y_u\|_\Omega},$$

and (7.31) follows with $c_0 = 1/(c^2+1)^{1/2}$. Hence, applying Theorem 4.3, the proof is complete. □

Example 7.7 (Schrödinger equation). This is an example of a *second*-order operator for which the previous theory applies. Let ∂_{xx} denote the Laplacian with respect to a spatial variable $0 < x < L$, let ∂_t denote the derivative $\partial/\partial t$ with respect to $0 < t < T$ (where both L and T are finite), let $\Omega = (0,L) \times (0,T)$, and let $f \in L^2(\Omega)$. The classical form of the Schrödinger initial boundary value problem is

$$\hat{\imath}\partial_t u - \partial_{xx} u = f, \quad 0 < x < L,\ 0 < t < T, \tag{7.34a}$$

$$u(x,t) = 0, \quad x = 0 \text{ or } x = L,\ 0 < t < T, \tag{7.34b}$$

$$u(x,0) = 0, \quad 0 < x < L. \tag{7.34c}$$

Here f is any given function in $L^2(\Omega)$. Viewing Ω as a rectangle with time as the vertical axis, let Γ denote the union of vertical boundary walls and the bottom initial time slice, and let $\tilde{\Gamma}$ denote the union of vertical boundary walls and the top final time slice. Then the initial and boundary conditions together can be written as $u|_\Gamma = 0$.

To fit into the previous framework, set

$$k = 2,\ m = 1, \quad A = \hat{\imath}\partial_t - \partial_{xx}.$$

Then the formal adjoint expression in (7.9) reads $A^* = \hat{\imath}\partial_t - \partial_{xx} = A$. Hence the graph spaces are

$$W = \tilde{W} = \{u \in L^2(\Omega) \colon i\partial_t u - \Delta_x u \in L^2(\Omega)\},$$

and the boundary operator $\tilde{D} = D \colon W \to W^*$ is set by $\langle Dw, v\rangle_W = (Aw, v)_\Omega - (w, Av)_\Omega$ for all $w, v \in W$. As usual, let $\mathcal{D}(\bar{\Omega})$ denote the restrictions of functions from $\mathcal{D}(\mathbb{R}^N)$ to Ω. Integration by parts shows that

$$\langle D\phi, \psi\rangle_W = \int_{\partial\Omega} \hat{\imath} n_t \phi\bar{\psi} + \int_{\partial\Omega} \phi n_x \partial_x \bar{\psi} - \int_{\partial\Omega} n_x \partial_x \phi\bar{\psi} \tag{7.35}$$

for all $\phi, \psi \in \mathcal{D}(\bar{\Omega})$, where we have used the spatial and temporal components n_x, n_t of the outward unit normal n on $\partial\Omega$.

To incorporate the boundary and initial conditions into $\mathrm{dom}(A)$, circumventing the development of a full trace theory for the graph space, we first set

$$\tilde{\mathcal{V}} = \{\varphi \in \mathcal{D}(\bar{\Omega}) \colon \varphi|_{\tilde{\Gamma}} = 0\},$$

and use it to set

$$\mathrm{dom}(A) = \{u \in W \colon \langle Dv, u\rangle_W = 0 \text{ for all } v \in \tilde{\mathcal{V}}\}, \tag{7.36}$$

or equivalently

$$V = {}^{\perp}D\tilde{\mathcal{V}}. \tag{7.37}$$

For smooth $u \in \mathcal{D}(\bar{\Omega}) \cap \mathrm{dom}(A)$, the integration-by-parts formula (7.35) shows that

$u|_\Gamma$ must vanish. Note that assumptions (7.3), (7.4) and (7.10) are immediately verified. By Lemma 7.2, the domain of the maximal adjoint is given by

$$\tilde{V} = {}^{\perp}DV\,. \tag{7.38}$$

It is shown in Demkowicz *et al.* (2017, Theorem 3.1) that $\tilde{\mathcal{V}}$ is dense in $\tilde{V}$. Hence (7.37) implies $V = {}^{\perp}D\tilde{V}$ and assumption (7.14) is also verified.

To conclude wellposedness of the ultraweak formulation (7.22) for this Schrödinger problem, the only remaining assumption we need to verify is the bijectivity stated in (7.30). Let $\phi_k(x)$ in $H^1_0(0,L)$ and let $\lambda_k > 0$ be a Laplace eigenpair satisfying $-\partial_{xx}\phi_k = \lambda_k\phi_k$ normalized so that $\|\phi_k\|_{(0,L)} = 1$ for all natural numbers $k \geq 1$. Suppose $f \in L^2(\Omega)$ and

$$f_k(t) = \int_0^L f(x,t)\bar{\phi}_k(x)\,dx, \qquad u_k(t) = -\hat{\imath}\int_0^t e^{\hat{\imath}\lambda_k(t-s)} f_k(s)\,ds, \tag{7.39a}$$

$$F_M(x,t) = \sum_{k=1}^{M} f_k(t)\phi_k(x), \qquad U_M(x,t) = \sum_{k=1}^{M} u_k(t)\phi_k(x). \tag{7.39b}$$

It is immediately verified that $AU_M = F_M$. Since U_M and any $\varphi \in \tilde{\mathcal{V}}$ are smooth enough for integration by parts using $\varphi|_{\Gamma^*} = 0$ and $U_M|_\Gamma = 0$, we have

$$(\hat{\imath}\partial_t U_M, \varphi)_\Omega = (U_M, \hat{\imath}\partial_t\varphi)_\Omega,$$
$$(\Delta U_M, \varphi)_\Omega = (U_M, \Delta\varphi)_\Omega.$$

Hence $\langle D\varphi, U_M\rangle_W = (A\varphi, U_M)_\Omega - (\varphi, AU_M)_\Omega = 0$ for all $\varphi \in \tilde{\mathcal{V}}$. By (7.37), this implies that U_M is in V.

To prove that A is surjective, it now suffices to show that the limit u of U_M exists in V and solves $Au = f$. Note that U_M is a Cauchy sequence in V. Indeed, for any $N > M$, by (7.39),

$$\|U_M - U_N\|^2_\Omega = \sum_{k=M+1}^{N}\int_0^T |u_k(t)|^2\,dt \leq \frac{1}{2}T^2\sum_{k=M+1}^{\infty}\int_0^T |f_k(t)|^2\,dt,$$

$$\|A(U_M - U_N)\|^2_\Omega = \|F_M - F_N\|^2_\Omega \leq \sum_{k=M+1}^{\infty}\int_0^T |f_k(t)|^2\,dt,$$

both of which converge to 0 as $M \to \infty$, because $f \in L^2(\Omega)$. Thus, having shown that U_M is a Cauchy sequence in V, we conclude that it must have an accumulation point u in V. Moreover, since Au and f are $L^2(\Omega)$-limits of the same sequence $F_M = AU_M$, we have $Au = f$. Thus $A\colon V \to L^2(\Omega)$ is surjective.

That A is in fact a bijection can be shown in many ways. For example, we can use an argument, completely analogous to the above, but now using (7.38) and with u_k defined by integrals from T to t, to show that $A = A^*\colon V^* \to L^2(\Omega)$ is also surjective. Since $\ker(A) = {}^{\perp}\mathrm{range}(A^*)$ this implies that $A\colon V \to L^2(\Omega)$ is injective, thus completing the verification of (7.30).

Example 7.8 (Poisson equation in first-order form). Reconsidering the Dirichlet boundary value problem (4.5) of Example 4.2, we now develop a different variational formulation for it. Reformulating $-\Delta u = f$ into a first-order system by introducing the flux $q = -\operatorname{grad} u$,

$$\begin{aligned} q + \operatorname{grad} u &= 0 \quad \text{in } \Omega, \\ \operatorname{div} q &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Using a group variable $v = (q, u) \in L^2(\Omega)^N \times L^2(\Omega)$, consider the unbounded operator

$$\begin{aligned} Av \equiv A(q,u) &= (q + \operatorname{grad} u, \operatorname{div} q), \quad k = 1, m = N+1, \\ \operatorname{dom}(A) &= H(\operatorname{div}, \Omega) \times \mathring{H}^1(\Omega). \end{aligned}$$

We easily see that the adjoint operator, acting on $\tilde{v} = (\tilde{q}, \tilde{u}) \in L^2(\Omega)^N \times L^2(\Omega)$, is

$$\begin{aligned} A^*\tilde{v} \equiv A^*(\tilde{q}, \tilde{u}) &= (\tilde{q} - \operatorname{grad} \tilde{u}, -\operatorname{div} \tilde{q}), \\ \operatorname{dom}(A^*) &= \operatorname{dom}(A). \end{aligned}$$

Clearly assumptions (7.3), (7.4) and (7.10) hold for this example. Since $V = \tilde{V}$ and $D = \tilde{D}$, the assumption (7.14) also holds since it is the same as the conclusion (7.13) of Lemma 7.2.

Note that if (q, u) and $A(q, u)$ are both in $L^2(\Omega)^m$, then obviously $\operatorname{div} q \in L^2(\Omega)$ and $\operatorname{grad} u \in L^2(\Omega)^N$, so

$$\begin{aligned} W = \tilde{W} &= H(\operatorname{div}, \Omega) \times H^1(\Omega), \\ W_h = \tilde{W}_h &= H(\operatorname{div}, \Omega_h) \times H^1(\Omega_h), \end{aligned}$$

using the broken Sobolev spaces defined in (4.4) and (4.33). Hence, for any $(q, u) \in W_h$ and $(\tilde{q}, \tilde{u}) \in \tilde{W}_h$,

$$\begin{aligned} \langle D_h(q,u), (\tilde{q},\tilde{u})\rangle_{\tilde{W}} &= \sum_{K \in \Omega_h} \langle q \cdot n, \tilde{u}\rangle_{H^{1/2}(\partial K)} + \langle u, \tilde{q} \cdot n\rangle_{H^{-1/2}(\partial K)} \\ &\equiv \langle q \cdot n, \tilde{u}\rangle_h + \langle u, \tilde{q} \cdot n\rangle_h, \end{aligned} \tag{7.40}$$

where in the last step we have extended the previous notation of (4.8) $\langle \cdot, \cdot \rangle_h$ to include sums of duality pairings in both $H^{1/2}(\partial K)$ $H^{-1/2}(\partial K)$. Since any $(q, u) \in V = \operatorname{dom}(A)$ satisfies $u|_{\partial\Omega} = 0$ on the global boundary, its element-by-element trace $\operatorname{tr}(u)$, as defined in (4.34), lies in

$$\mathring{H}^{1/2}(\partial\Omega_h) = \{\hat{w} \in H^{1/2}(\partial\Omega_h) \colon w|_{\partial\Omega} = 0\} = \operatorname{tr} \mathring{H}^1(\Omega).$$

The (hybrid) ultraweak formulation (7.22) now takes the form (1.1) with the following forms:

$$\begin{aligned} b(\,(q,u,\hat{u},\hat{q}_n),\ (r,v)\,) &= (q,r)_h - (u,\operatorname{div} r)_h + \langle \hat{u}, r\cdot n\rangle_h \\ &\quad - (q,\operatorname{grad} v)_h + \langle v,\hat{q}_n\rangle_h, \\ \ell(r,v) &= (f,v)_\Omega, \end{aligned}$$

where $(q,u)\in W_h$, $(r,v)\in \tilde{W}_h$, $\hat{u}\in \mathring{H}^{1/2}(\partial\Omega_h)$ and $\hat{q}_n\in H^{-1/2}(\partial\Omega_h)$.

By Theorem 7.6, this ultraweak formulation is wellposed if $A\colon V\to L^2(\Omega)^{N+1}$ is a bijection, that is, if there is a unique $q\in H(\operatorname{div},\Omega)$ and $u\in \mathring{H}^1(\Omega)$ satisfying

$$q+\operatorname{grad} u = G \quad \text{on } \Omega, \tag{7.41a}$$

$$\operatorname{div} q = F \quad \text{on } \Omega \tag{7.41b}$$

for any given $F\in L^2(\Omega)$ and $G\in L^2(\Omega)^N$. To verify this condition, it is sufficient to note that q and u satisfy (7.41) if and only if they form the unique solution of the well-known mixed weak problem (see e.g. Brezzi and Fortin 1991, Ch. II, Prop. 1.3) to find q in $H(\operatorname{div},\Omega)$ and u in $L^2(\Omega)$ such that

$$(q,r)_\Omega - (u,\operatorname{div} r)_\Omega = (G,r)_\Omega \quad \text{for all } r\in H(\operatorname{div},\Omega), \tag{7.42a}$$

$$(\operatorname{div} q,w)_\Omega = (F,w)_\Omega \quad \text{for all } w\in L^2(\Omega). \tag{7.42b}$$

It is easy to see that (7.42a) also implies that $u\in \mathring{H}^1(\Omega)$ and (7.41a) holds. Hence the unique solution (q,u) of (7.42) is in V and solves $A(q,u)=(G,F)$, thus verifying assumption (7.30).

Bibliographical notes. Theorem 7.6 and the treatment of the Schrödinger equation (Example 7.7) by DPG methods appeared first in Demkowicz *et al.* (2017). There it is also pointed out why it is not advisable to split the Schrödinger equation into a first-order system. This is the reason for staying with the original second-order form of the Schrödinger equation while deriving the ultraweak formulation in Example 7.7. The wellposedness result in Example 7.8 was first proved in Demkowicz and Gopalakrishnan (2011*a*), but using different techniques. An application of Theorem 7.6 to the spacetime wave equation can be found in Gopalakrishnan and Sepúlveda (2019). That paper also notes how the spacetime wave operator produces a D_h operator with a non-trivial null space (even though $\hat{q}=D_h u$ can be uniquely determined) and how one overcomes the consequent difficulties in practically solving an ultraweak discretization.

7.3. Analysis with scaled and optimal norms

In practice, it is often useful to introduce a scaling parameter to tune the norm in which the residual is minimized. In this subsection we consider the case where the terms in the test space norm we have been working with (see (7.17)) are differently weighted. We continue to use the notation from the previous subsections, e.g. $\tilde{W}_h$

is as in (7.15) and Q is as in (7.21), but now consider a new test norm on $\tilde{W}_h$, defined for any $0 < s < \infty$, by

$$\|\tilde{w}\|_{Y,s}^2 = s^{-2}\|\tilde{w}\|_{\Omega}^2 + \|A_h^*\tilde{w}\|_{\Omega}^2, \quad \tilde{w} \in Y = \tilde{W}_h,$$

and a new s-dependent norm on the trial space by

$$\|(w,\hat{r})\|_{X,s}^2 = \inf_{v \in D_{h,V}^{-1}\{\hat{r}\}} \left(\|w - v\|_{\Omega}^2 + s^2\|Av\|_{\Omega}^2\right), \quad (w,\hat{r}) \in X = L^2(\Omega)^m \times Q.$$

For any fixed $s > 0$, the test norm $\|\tilde{w}\|_{Y,s}$ is obviously equivalent to $\|\tilde{w}\|_{\tilde{W}_h}$. The fact that $\|(w,\hat{r})\|_{X,s}$ is a norm follows from the next result. In these norms, the ultraweak form

$$b((w,\hat{r}),\ \tilde{w}) = \left(w, A_h^*\tilde{w}\right)_{\Omega} + \langle \hat{r}, \tilde{w}\rangle_{\tilde{W}_h}$$

becomes a generalized duality pairing (of Definition 3.5), as shown next. Consequently, the energy norm (of Definition 3.1) on X for the ultraweak formulation with the test norm $\|\cdot\|_{Y,s}$ is $\|(\cdot,\cdot)\|_{X,s}$, and simultaneously, the optimal test norm (of Definition 3.5) on Y corresponding to the trial norm $\|(\cdot,\cdot)\|_{X,s}$ is $\|\cdot\|_{Y,s}$.

Theorem 7.9 (Optimal norms for ultraweak formulations). Adopt the setting and assumptions of Theorem 7.6 and let $X = L^2(\Omega)^m \times Q$ and $Y = \tilde{W}_h$. Then, for all $(v,\hat{v}) \in X$ and $\tilde{w} \in Y$,

$$\|(v,\hat{v})\|_{X,s} = \sup_{0 \neq \tilde{w} \in Y} \frac{|b((v,\hat{v}),\ \tilde{w})|}{\|\tilde{w}\|_{Y,s}}, \quad \|\tilde{w}\|_{Y,s} = \sup_{0 \neq (v,\hat{v}) \in X} \frac{|b((v,\hat{v}),\ \tilde{w})|}{\|(v,\hat{v})\|_{X,s}}. \tag{7.43}$$

In these norms, both $\|b\|$ and the inf-sup constant γ are one. The approximation $(u_h,\hat{q}_h) \in X_h$ from the ideal DPG method to the ultraweak solution $(u,\hat{q})$ using any $X_h \subset X$ is the best in the sense that

$$\|(u - u_h, \hat{q} - \hat{q}_h)\|_{X,s} = \inf_{(w_h,\hat{r}_h) \in X_h} \|(u - w_h, \hat{q} - \hat{r}_h)\|_{X,s}. \tag{7.44}$$

Proof. We need only prove the second equality in (7.43). The first equality of (7.43) then follows from the second by Proposition 3.6, and moreover, (7.44) then follows from Theorem 3.2(b).

Let $\tilde{w} \in \tilde{W}_h$. We will produce a $(w,\hat{r}) \in V \times Q$ satisfying

$$b((w,\hat{r}),\tilde{w}) = \|\tilde{w}\|_{Y,s}^2, \quad \|(w,\hat{r})\|_{X,s} \leq \|\tilde{w}\|_{Y,s}. \tag{7.45}$$

By virtue of (7.30), there is a $z \in V$ such that

$$Az = \tilde{w}. \tag{7.46}$$

Then

$$\begin{aligned}
\|\tilde{w}\|_{Y,s}^2 &= (s^{-2}\tilde{w},\tilde{w})_{\Omega} + \left(A_h^*\tilde{w}, A_h^*\tilde{w}\right)_{\Omega} \\
&= (A(s^{-2}z),\tilde{w})_{\Omega} + \left(A_h^*\tilde{w}, A_h^*\tilde{w}\right)_{\Omega} \\
&= \left(s^{-2}z, A_h^*\tilde{w}\right)_{\Omega} + \langle D_h(s^{-2}z),\tilde{w}\rangle_{\tilde{W}_h} + \left(A_h^*\tilde{w}, A_h^*\tilde{w}\right)_{\Omega} \\
&= b((w,\hat{r}),\tilde{w})
\end{aligned}$$

with

$$w = s^{-2}z + A_h^*\tilde{w} \in L^2(\Omega)^m, \quad \hat{r} = D_h(s^2 z) \in Q. \tag{7.47}$$

Moreover,

$$\begin{aligned}\|(w,\hat{r})\|_{X,s}^2 &= \inf_{v \in D_{h,V}^{-1}\{\hat{q}\}} \left(\|w-v\|_\Omega^2 + s^2\|Av\|_\Omega^2\right) \\ &\le \|w - (s^{-2}z)\|_\Omega^2 + s^2\|A(s^{-2}z)\|_\Omega^2 \\ &= \|A_h^*\tilde{w}\|_\Omega^2 + s^{-2}\|\tilde{w}\|_\Omega^2 = \|\tilde{w}\|_{Y,s}^2,\end{aligned}$$

where we have used the formulas for w and z, from (7.47) and (7.46) respectively, in the last step. This proves (7.45), from which it readily follows that

$$\sup_{0\neq(v,\hat{v})\in X} \frac{|b((v,\hat{v}),\ \tilde{w})|}{\|(v,\hat{v})\|_{X,s}} \ge \frac{|b((w,\hat{r}),\ \tilde{w})|}{\|(w,\hat{r})\|_{X,s}} \ge \|\tilde{w}\|_{Y,s}. \tag{7.48}$$

In fact the supremum equals $\|\tilde{w}\|_{Y,s}$ because the reverse inequality also holds, as we now show. Letting $(v,\hat{v})$ be any element in X and choosing any $z \in V$ such that $\hat{v} = D_h z$,

$$\begin{aligned}b((v,\hat{v}),\tilde{w}) &= (v, A_h^*\tilde{w})_\Omega + \langle D_h z, \tilde{w}\rangle_{\tilde{W}_h} \\ &= (v, A_h^*\tilde{w})_\Omega + (Az,\tilde{w})_\Omega - (z, A_h^*\tilde{w})_\Omega \\ &= (v-z, A_h^*\tilde{w})_\Omega + (Az,\tilde{w})_\Omega \\ &\le \left(\|v-z\|_\Omega^2 + s^2\|Az\|_\Omega^2\right)^{1/2}\|\tilde{w}\|_{Y,s}.\end{aligned}$$

Taking the infimum over all $z \in D_{h,V}^{-1}\{\hat{v}\}$, we obtain

$$b((v,\hat{v}),\tilde{w}) \le \|(v,\hat{v})\|_{X,s}\|\tilde{w}\|_{Y,s},$$

which together with (7.48) proves the second equality of (7.43). □

The trial norm $\|\cdot\|_{X,s}$ of Theorem 7.9 is related to Peetre's K-functional (Bergh and Löfström 1976). To see this, suppose there is a $C_V > 0$ such that

$$\|v\|_\Omega \le C_V\|Av\|_\Omega \quad \text{for all } v \in V. \tag{7.49}$$

Assumption (7.30) certainly implies the existence of such a C_V (by the Closed Range Theorem). Let $V_0 = \{w_h \in W_h : D_h w_h = 0\}$ be the kernel of D_h, which is a closed subspace by the continuity of D_h. By Lemma 7.4, V_0 is a subspace of V. Hence, for any $\hat{u} \in Q$, the set $D_{h,V}^{-1}\{\hat{u}\}$ equals the affine translate $v_{\hat{u}} + V_0$ for any $v_{\hat{u}} \in V$ with the property $D_h v_{\hat{u}} = \hat{u}$. Minimization over this closed coset gives a minimal extension $E\hat{u}$ of $\hat{u}$ defined by

$$E\hat{u} = \arg\min_{v \in D_{h,V}^{-1}\{\hat{u}\}} \|Av\|_\Omega.$$

Note that by (7.49), $\|Av\|_\Omega$ is a norm on $D_{h,V}^{-1}\{\hat{u}\} \subset V$. Since E is defined through minimization over a translate of V_0, we see that

$$(AE\hat{u}, Av_0)_\Omega = 0 \quad \text{for all } v_0 \in V_0. \tag{7.50}$$

Define the K-functional for the scale of spaces between V_0 and $L^2(\Omega)^m$ by

$$K(s, w) = \inf_{v_0 \in V_0} \left(\|w - v_0\|_\Omega^2 + s^2 \|Av_0\|_\Omega^2 \right) \tag{7.51}$$

for any $w \in L^2(\Omega)^m$. The next two results help us better understand the norm $\|\cdot\|_{X,s}$ in Theorem 7.9.

Proposition 7.10. Suppose (7.49) holds. Then, for any $(u, \hat{u}) \in X$,

$$\|(u, \hat{u})\|_{X,s}^2 = s^2 \|AE\hat{u}\|_\Omega^2 + K(s, u - E\hat{u}).$$

Proof. Writing any $v \in D_{h,V}^{-1}\{\hat{u}\}$ as $v = E\hat{u} + v_0$ for a $v_0 \in V_0$,

$$\begin{aligned} \|u - v\|_\Omega^2 + s^2 \|Av\|_\Omega^2 &= \|u - E\hat{u} - v_0\|_\Omega^2 + s^2 \|A(E\hat{u} + v_0)\|_\Omega^2 \\ &= \|u - E\hat{u} - v_0\|_\Omega^2 + s^2 \|AE\hat{u}\|_\Omega^2 + s^2 \|Av_0\|_\Omega^2, \end{aligned}$$

where the last equality is due to (7.50). Hence the result follows by minimizing over $v_0 \in V_0$. □

Proposition 7.11. Let C_V be as in (7.49), $c_s = C_V^2/s^2$, and $k_s = \frac{1}{2}(c_s + \sqrt{c_s^2 + 4c_s})$. Then, for all $(u, \hat{u}) \in X$ and $s > 0$, we have these two-sided bounds:

$$(1 + k_s)^{-1} \|(u, \hat{u})\|_{X,s}^2 \le \|u\|_\Omega^2 + s^2 \|AE\hat{u}\|_\Omega^2 \le (1 + k_s) \|(u, \hat{u})\|_{X,s}^2. \tag{7.52}$$

Proof. By the triangle inequality,

$$\begin{aligned} \|u\|_\Omega^2 &\le (\|u - E\hat{u} - v_0\|_\Omega + \|E\hat{u} + v_0\|_\Omega)^2 \\ &\le (1 + \alpha^{-2}) \|u - E\hat{u} - v_0\|_\Omega^2 + (1 + \alpha^2) C_V^2 \|A(E\hat{u} + v_0)\|_\Omega^2, \end{aligned}$$

where we have used (7.49) and the inequality $(a + b)^2 \le (1 + \alpha^{-2})a^2 + (1 + \alpha^2)b^2$ for numbers a, b and $\alpha > 0$. Using (7.50),

$$\begin{aligned} \|u\|_\Omega^2 + s^2 \|AE\hat{u}\|_\Omega^2 \le{}& (1 + \alpha^{-2}) \|u - E\hat{u} - v_0\|_\Omega^2 \\ &+ [(1 + \alpha^2)c_s + 1] \left(s^2 \|AE\hat{u}\|_\Omega^2 + s^2 \|Av_0\|_\Omega^2 \right) \end{aligned}$$

with $c_s = C_V^2/s^2$. Now set $\alpha^2 = \frac{1}{2}(-c_s + \sqrt{c_s^2 + 4c_s})/c_s$ so that $(1 + \alpha^2)c_s = \alpha^{-2}$. Then $1 + \alpha^{-2} = 1 + (1 + \alpha^2)c_s = 1 + k_s$ and the last inequality of (7.52) follows after taking the infimum over all $v_0 \in V_0$ and applying Proposition 7.10.

For the first inequality of (7.52), we begin by noting that the choice of $v_0 = 0$ in (7.51) gives $K(s, w) \le \|w\|_\Omega^2$. Together with Proposition 7.10, we then have

$$\|(u, \hat{u})\|_{X,s} \le s^2 \|AE\hat{u}\|_\Omega^2 + \|u - E\hat{u}\|_\Omega^2.$$

By the triangle inequality and (7.49),

$$\|(u, \hat{u})\|_{X,s} \le (1 + \alpha^{-2}) \|u\|_\Omega^2 + [(1 + \alpha^{-2})c_s + 1] s^2 \|AE\hat{u}\|_\Omega^2.$$

Choosing exactly the same α as before, $1 + \alpha^{-2} = 1 + (1 + \alpha^2)c_s = 1 + k_s$, and the first inequality of (7.52) is proved. □

Note that when Proposition 7.11 is combined with (7.44) of Theorem 7.9, we obtain

$$\|u - u_h\|_{\Omega}^2 + s^2\|AE(\hat{u} - \hat{u}_h)\|_{\Omega}^2 \\ \leq (1 + k_s)^2\left[\inf_{w_h \in X_{h,0}} \|u - w_h\|_{\Omega}^2 + \inf_{\hat{r}_h \in \hat{X}_h} s^2\|AE(\hat{u} - \hat{r}_h)\|_{\Omega}^2\right], \tag{7.53}$$

where the constant k_s is as in Proposition 7.11. At the price of increasing the quasi-optimality constant from the optimal one, this estimate gives a simpler implication of (7.44) in easier norms.

Example 7.12 (Helmholtz equation for time-harmonic waves). The Helmholtz equation arises in varied applications, including electromagnetics and acoustics. For example, in the latter, the physics of acoustical disturbances (Courant and Friedrichs 1948) show that by linearizing the isentropic Euler equations around a hydrostatic solution and assuming harmonic time variations, we obtain

$$\hat{\imath}\omega v + \operatorname{grad}\phi = G \quad \text{in } \Omega, \tag{7.54a}$$

$$\hat{\imath}\omega\phi + \operatorname{div} v = F \quad \text{in } \Omega, \tag{7.54b}$$

for some given $\omega > 0$, $F \in L^2(\Omega)$ and $G \in L^2(\Omega)^N$. Here $v\colon \Omega \to \mathbb{C}^N$ and $\phi\colon \Omega \to \mathbb{C}$ are velocity and pressure variables, respectively, associated to the acoustic perturbations from equilibrium, complexified under the standard time-harmonic assumption. These equations must be supplemented by a boundary condition. Let us consider the impedance boundary condition

$$v \cdot n - \phi = 0 \quad \text{on } \partial\Omega. \tag{7.54c}$$

Other Dirichlet, Neumann or mixed-type boundary conditions can equally well be considered. Note that taking the divergence of (7.54a) and substituting the value of $\operatorname{div} v$ from (7.54b), we recover the popular second-order form of the Helmholtz equation for ϕ (which we shall not use here).

The first-order system (7.54) can be written as $Au = f$ using the group variable $u = (v, \phi) \in L^2(\Omega)^N \times L^2(\Omega)$, the unbounded operator

$$Au = (\hat{\imath}\omega v + \operatorname{grad}\phi,\ \hat{\imath}\omega\phi + \operatorname{div} v)$$

and $f = (G, F) \in L^2(\Omega)^N \times L^2(\Omega)$. Clearly (7.54) is in the setting of (7.2) with $m = N + 1$ and $\operatorname{dom} A$ equal to

$$V = \{(z, \mu) \in H(\operatorname{div}, \Omega) \times H^1(\Omega)\colon z \cdot n = \mu \text{ on } \partial\Omega\}.$$

Its adjoint is

$$A^*\tilde{u} = (-\hat{\imath}\omega\tilde{v} - \operatorname{grad}\tilde{\phi},\ -\hat{\imath}\omega\tilde{\phi} - \operatorname{div}\tilde{v})$$

for any $\tilde{u} = (\tilde{v}, \tilde{\phi})$ in dom A^*, which equals

$$\tilde{V} = \{(z, \mu) \in H(\mathrm{div}, \Omega) \times H^1(\Omega) \colon z \cdot n = -\mu \text{ on } \partial\Omega\},$$

a space analogous to V but with a change of sign in the boundary condition. It is easy to verify that (7.3), (7.4) and (7.10) hold. Using the standard trace theory of $H(\mathrm{div}, \Omega)$ and $H^1(\Omega)$, it is also easy to verify that (7.14) hold.

To apply Theorems 7.6 and 7.9, it therefore suffices to verify the bijectivity in (7.30). Injectivity follows from uniqueness of Helmholtz solutions, so (7.30) follows from stability results of the form

$$\|v\|_\Omega \le C(\omega)\|Av\|_\Omega, \quad v \in V, \tag{7.55}$$

which is the same as (7.49) with $C_V = C(\omega)$. The inequality (7.55) was proved in Demkowicz, Gopalakrishnan, Muga and Zitelli (2012*b*, Lemmas 4.2 and 4.3), using a result of Melenk (1995), with a $C(\omega)$ independent of ω on a convex domain Ω for the present case of impedance boundary conditions. For other boundary conditions or on trapping domains, we generally expect (7.55) to hold with an ω-dependent constant. Hence we proceed assuming that (7.55) holds, and consider the DPG ultraweak formulation to find $u \in L^2(\Omega)^{N+1}$ and $\hat{u} \in Q = \mathrm{range}(D_{h,V})$ satisfying

$$(u, A_h^* \tilde{w})_\Omega + \langle \hat{u}, \tilde{w} \rangle_{\tilde{W}_h} = F(\tilde{w}) \quad \text{for all } \tilde{w} \in \tilde{W}_h, \tag{7.56}$$

with the broken adjoint Helmholtz operator A_h^*. We conclude that this is a wellposed formulation by Theorem 7.6.

Next let us apply Theorem 7.9 with $s = 1/\delta$ for some small $0 < \delta \ll 1$ for an ideal DPG approximation $(u_h, \hat{u}_h) \in X_{0,h} \times \hat{X}_h \subset X_h$ of (7.56). Recall that the combination of Proposition 7.11 and Theorem 7.9 yields (7.53) with $k_s = 1 + c\delta$ for a constant $c > 0$ that depends only on $C(\omega)$. Then (7.53) implies

$$\begin{aligned} &\|u - u_h\|_\Omega^2 + \frac{1}{\delta^2}\|AE(\hat{u} - \hat{u}_h)\|_\Omega^2 \\ &\quad \le (1 + c\delta)^2 \left[\inf_{w_h \in X_{h,0}} \|u - w_h\|_\Omega^2 + \inf_{\hat{r}_h \in \hat{X}_h} \frac{1}{\delta^2}\|AE(\hat{u} - \hat{r}_h)\|_\Omega^2 \right]. \end{aligned}$$

This estimate was arrived at by other means (without using Theorem 7.9) in Gopalakrishnan, Muga and Olivares (2014). There it was offered as a justification for the practically visible marked improvement in DPG Helmholtz solutions as δ is made smaller. The improvement in solutions was further justified there via a dispersion analysis of the DPG method for the Helmholtz equation on an infinite uniform stencil. Since the test norm

$$\|\tilde{w}\|_{Y,1/\delta}^2 = \delta^2\|\tilde{w}\|_\Omega^2 + \|A_h^*\tilde{w}\|_\Omega^2$$

becomes smaller as δ is made smaller, a takeaway from such observations is that it pays to use a weaker norm on the test space when computing wave solutions.

8. Optimal test functions in time integrators

In this section we present an application of the DPG ideas to design an exponential integrator for initial value problems. The resulting method yields a discrete solution not only at the time steps but also between the time steps. In fact, in each time interval, the discrete solution is the best (in L^2-norm) possible approximation of the exact solution from a polynomial space. We also show how the DPG error representation can be used for *a posteriori* error control within the time integrators.

8.1. An initial value system

Let $K \in \mathbb{C}^{m\times m}$ be a non-singular matrix and $\Omega = (0,1) \subset \mathbb{R}$. Given $u_0 \in \mathbb{C}^m$ and $f \in L^2(\Omega)$, consider the initial value problem for $u\colon \Omega \to \mathbb{C}^m$ satisfying

$$\begin{aligned} \frac{du}{dt} + Ku &= f, \quad 0 < t < 1, \\ u(0) &= u_0. \end{aligned} \tag{8.1}$$

This can be viewed as a generalization of Example 2.6. We may proceed similarly to get the weak problem to find $u \in L^2(\Omega)^m$ and $\hat{u} \in \mathbb{C}^m$ satisfying

$$(u, A^*v)_\Omega + \hat{u}\,\overline{v(1)} = (f, v)_\Omega + u_0\,\overline{v(0)} \quad \text{for all } v \in H^1(\Omega)^m, \tag{8.2}$$

with

$$A^*v = -\frac{dv}{dt} + K^*v.$$

This fits into our framework with

$$b((w,\hat{w}), y) = (w, A^*y)_\Omega + \hat{w}\,\overline{y(1)}, \tag{8.3a}$$

$$\ell(v) = (f, v)_\Omega + u_0\,\overline{v(0)}. \tag{8.3b}$$

An alternative avenue to arrive at the same weak formulation is the approach of Section 7, that is, one would set $\Omega = (0,1)$, an unbounded operator $Au = du/dt + Ku$, on $L^2(\Omega)^m$ with $\operatorname{dom}(A) = \{u \in H^1(\Omega)^m\colon u(0) = 0\}$, and develop the following formulation for homogeneous initial conditions: $(u, A^*v) = (f, v)$ for all $v \in V^* = \operatorname{dom} A^* = \{u \in H^1(\Omega)^m\colon u(1) = 0\}$. One would then extend it to cover the non-homogeneous initial condition u_0 by a process similar to going from unbroken to broken graph spaces, employing a larger class of test functions that do not vanish at $t = 1$. This would then result in the additional unknown, the trace $\hat{u}$, and one would obtain (8.2) again. Of course, for regular solutions, $\hat{u} = u(1)$.

Returning to (8.3), we endow the trial space $X = L^2(\Omega)^m \times \mathbb{C}^m$ with the norm

$$\|(w,\hat{w})\|_X^2 = \|w\|_\Omega^2 + |\hat{w}|_2^2,$$

where $|\hat{w}|_2^2 = |\hat{w}_1|^2 + \cdots + |\hat{w}_m|^2$ denotes the square of the ℓ^2-norm of $\hat{w} \in \mathbb{C}^m$. On the test space $Y = H^1(\Omega)^m$, the corresponding optimal test norm of (3.6) can then

be computed easily. Namely, with b as in (8.3), we see that

$$|\!|\!|y|\!|\!|_Y = \sup_{0\neq(w,\hat{w})\in X} \frac{|b((w,\hat{w}),y)|}{\|(w,\hat{w})\|_X} = \left(\|A^*y\|_\Omega^2 + |y(1)|_2^2\right)^{1/2}. \tag{8.4a}$$

We then set the norm on $Y = H^1(\Omega)^m$ to be the optimal test norm, that is,

$$\|y\|_Y = |\!|\!|y|\!|\!|_Y. \tag{8.4b}$$

Clearly, for any $v, y \in Y$,

$$\begin{aligned}(v,y)_Y &= (A^*v, A^*y)_\Omega + v(1)\cdot\overline{y(1)} \\ &= (-\dot{v}+K^*v, -\dot{y}+K^*y)_\Omega + v(1)\cdot\overline{y(1)}\end{aligned} \tag{8.4c}$$

is the inner product that generates the optimal test norm above. Here $\dot{v} = dv/dt$. This is one of the rare cases where the optimal test norm is so readily calculable.

The ideal Petrov Galerkin method (i.e. the IPG method of Definition 2.3) uses the optimal test space, which we now examine. Using the inner product in (8.4c), the variational problem for the optimal test function v corresponding to any $(w,\hat{w}) \in X$ reads as follows for any $y \in H^1(\Omega)^m$:

$$(-\dot{v}+K^*v, -\dot{y}+K^*y)_\Omega + v(1)\cdot\overline{y(1)} = (w, -\dot{y}+K^*y)_\Omega + \hat{w}\,\overline{y(1)}. \tag{8.5}$$

For any $g \in L^2(\Omega)$, since the initial value problem $\dot{y} - K^*y = -g$ with initial condition $y(1) = 0$ is solvable, we find that (8.5) implies $(-\dot{v}+K^*v, g)_\Omega = (w,g)_\Omega$ for all g in $L^2(\Omega)$, i.e. $-\dot{v}+K^*v = w$. Using this in (8.5), we then conclude that $v(1) = \hat{w}$. Thus the optimal test function v of any $(w,\hat{w}) \in X$ is the solution of

$$-\frac{dv}{dt} + Kv = w \quad \text{in } \Omega, \tag{8.6a}$$

$$v(1) = \hat{w}. \tag{8.6b}$$

Recall the matrix exponential, defined by

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k. \tag{8.7}$$

Using it, the solution of (8.6) can be written down in closed form by the *variation of constants* method. Namely, the optimal test function v and the trial-to-test operator T are given by

$$v(t) = T(w,\hat{w}) = e^{K^*(t-1)}\hat{w} + e^{K^*t}\int_1^t e^{-K^*\tau} w(\tau)\,d\tau. \tag{8.8}$$

Throughout this section, we fix T to be this operator. Because we have chosen the optimal test norm, we are now able to prove that the resulting Petrov–Galerkin method produces the best possible L^2-approximation within $(0,1)$, and furthermore, the numerical flux approximation at the endpoint $t = 1$ has zero error.

Proposition 8.1 (Optimality of interior solution and endpoint exactness). Let U_h be any finite-dimensional subspace of $L^2(0,1)^m$ and let the interior solution $u_h \in U_h$ together with the endpoint solution $\hat{u}_h \in \mathbb{C}^m$ satisfy

$$(u_h, A^* v)_\Omega + \hat{u}_h \overline{v(1)} = (f, v)_\Omega + u_0 \overline{v(0)} \quad \text{for all } v \in Y_h^{\text{opt}}, \tag{8.9}$$

where $Y_h^{\text{opt}} = T(U_h \times \mathbb{C}^m)$ for the T given by (8.8). Let u be the exact solution of (8.1). Then

$$u_h = \Pi_U u \quad \text{and} \quad \hat{u}_h = \hat{u} = u(1), \tag{8.10}$$

where Π_U is the L^2-orthogonal projection into U_h.

Proof. The norm choice in (8.4b) makes the form $b((w,\hat{w}), y)$ into a generalized duality pairing (as in Definition 3.5), so by Proposition 3.6, the energy norm is the same as the $\|\cdot\|_X$-norm. Hence, by Theorem 3.2(b), solution of the IPG method for this formulation equals the best approximation, that is, the given u_h and $\hat{u}_h$ satisfy

$$\|u - u_h\|_\Omega^2 + |\hat{u} - \hat{u}_h|_2^2 \;=\; \inf_{w_h \in U_h \;\; \hat{w}_h \in \mathbb{C}^m} \big(\|u - w_h\|_\Omega^2 + |\hat{u} - \hat{w}_h|_2^2 \big).$$

It is easy to see that the infimum equals $\|u - \Pi_U u\|_\Omega$. Hence the identities of (8.10) follow. □

8.2. *The discrete system*

Consider the basis for the set of vector polynomials $P_p(\Omega)^m$ given by monomials $t^j e_i$ for $t \in \Omega = (0,1)$, $j = 0, \dots, p$, and $i = 1, \dots, m$ (where e_i are the standard unit vectors). Let us set U_h in Proposition 8.1 by

$$U_h = P_p(\Omega)^m = \operatorname{span}\{t^j e_i : \; j = 0, \dots, p, \; i = 1, \dots, m\}$$

and examine how to solve for u_h and $\hat{u}_h$ in (8.9). Then we introduce the following functions that emerge from the previous formula in (8.8) for the trial-to-test operator:

$$\begin{aligned}
\hat{v}_i &:= T(0, e_i) = e^{K^*(t-1)} e_i \\
v_{0,i} &:= T(e_i, 0) = K^{-*}[I - e^{K^*(t-1)}] e_i \\
v_{p,i} &:= T(t^p e_i, 0) = e^{K^* t} \int_t^1 e^{-K^* \tau} \tau^p e_i \, d\tau \\
&= K^{-*}(t^p e_i + p v_{p-1,i} - \hat{v}_i)
\end{aligned}$$

for $p = 1, 2, \dots$, where we have integrated by parts to get the last identity. Given any $M \in \mathbb{C}^{m \times m}$ and $t > 0$, define the matrix-valued functions

$$R_p(M,t) := \sum_{j=0}^{p} \frac{p!}{j!} (Mt)^j \quad \text{and} \quad \hat{v}(M,t) := e^{M(t-1)}.$$

Then let

$$v_r(M,t) = M^{-r-1}(R_r(M,t) - R_r(M,1)\,\hat{v}(M,t)). \tag{8.11}$$

Using this notation, we can express the previously given optimal test functions as

$$v_{r,i} = v_r(K^*,t)e_i \quad \text{for all } r = 0,1,\dots,p.$$

When these optimal test function expressions are substituted into (8.9), we obtain a system for the discrete solution

$$u_h = \sum_{j=0}^{p} u_{h,j}t^j, \quad u_{h,j} \in \mathbb{C}^m, \tag{8.12a}$$

which couples the solution coefficients $u_{h,j}$ by

$$\sum_{j=0}^{p} a_{rj}\,u_{h,j} = v_r(K,0)\,u_0 + \int_0^1 v_r(K,t)\,f(\tau)\,dt, \quad r = 0,\dots,p, \tag{8.12b}$$

where

$$a_{rj} := \int_0^1 t^{j+r}\,dt.$$

This is a system of $p+1$ equations for the (vector-valued) unknowns $u_{h,j}$, $j = 0,\dots,p$. The endpoint trace, which equals the exact solution by Proposition 8.1, is given by

$$\hat{u}_h = \hat{v}(K,0)\,u_0 + \int_0^1 \hat{v}(K,t)\,f(\tau)\,d\tau. \tag{8.12c}$$

Thus, to compute $\hat{u}_h$ and $u_{h,j}$, we need techniques to compute the integrals involving matrix exponentials.

8.3. Exponential quadrature rules

To proceed, as seen above, we must digress to review standard exponential integrators, which, for the system (8.1), are based on the formula for the exact solution obtained by the method of variation of constants, namely

$$u(t) = e^{-Kt}u_0 + e^{-Kt}\int_0^t e^{K\tau}f(\tau)\,d\tau. \tag{8.13}$$

Applying the formula recursively to intervals $[t_{k-1},t_k]$, we have

$$u(t_k) = e^{-h_kK}u(t_{k-1}) + \int_{t_{k-1}}^{t_k} e^{(\tau-t_k)K}f(\tau)\,d\tau, \quad h_k := t_k - t_{k-1}.$$

The integral above can be approximated using standard exponential quadrature rules that we now describe.

Selecting s arbitrary quadrature points $c_i \in [0,1]$, $i = 1, \ldots, s$, we approximate the right-hand side function $f(s)$ in the time interval $[0,1]$ by

$$f(\tau) \approx \sum_{i=1}^{s} f(t_{k-1} + c_i h_k)\, \tilde{l}_i(\tau),$$

where $l_i, i = 1, \ldots, s$ are the Lagrange polynomials of order $s-1$ on the unit interval $I = [0,1]$

$$l_i(\theta) = \prod_{j=1, j\neq i}^{s} \frac{\theta - c_j}{c_i - c_j}, \quad j = 1, \ldots, s,$$

and $\tilde{l}_i(\tau)$ are the corresponding mapped Lagrange polynomials on interval $[t_{k-1}, t_k]$. Substituting the approximation for $f(\tau)$ into the formula (8.13) from variation of constants, we obtain a time-marching scheme,

$$u^k = e^{-h_k K} u^{k-1} + h_k \sum_{i=1}^{s} b_i(-h_k K) f_i,$$

where $u^k \approx u(t_k)$, $f_i = f(t_{k-1} + c_i h_k)$, and the weights are defined by

$$b_i(z) := \int_0^1 e^{(1-\theta)z} l_i(\theta)\, d\theta, \quad i = 1, \ldots, s.$$

It is standard to compute the weights using the so-called 'ϕ-functions' (see e.g. Al-Mohy and Higham 2011 or Niesen and Wright 2012), defined as follows:

$$\begin{aligned} \phi_0(z) &:= e^z, \\ \phi_p(z) &:= \int_0^1 e^{(1-\theta)z} \frac{\theta^{p-1}}{(p-1)!}\, d\theta \\ &= \frac{1}{z}\left(\phi_{p-1}(z) - \frac{1}{(p-1)!}\right), \quad p = 1, 2, \ldots. \end{aligned}$$

The two simple examples below show how they are used.

Example 8.2 (A standard one-point integrator). Selecting a single point $c_1 \in [0,1]$, we have $l_1(\theta) = 1$, $b_1(z) = \phi_1(z)$, $e^z = z\phi_1(z) + 1$, which gives

$$u^k = u^{k-1} + h_k \phi_1(-h_k K)(f_1 - K u^{k-1}),$$

an integrator formula for the $s = 1$ case.

Example 8.3 (A standard two-point integrator). Selecting $c_1, c_2 \in [0,1]$, we have

$$\begin{aligned} b_1(z) &= \int_0^1 e^{(1-\theta)z} \frac{\theta - c_2}{c_1 - c_2}\, d\theta \\ &= \frac{1}{c_1 - c_2} \int_0^1 e^{(1-\theta)z} \theta \, d\theta - \frac{c_2}{c_1 - c_2} \int_0^1 e^{(1-\theta)z}\, d\theta \\ &= \frac{1}{c_1 - c_2} \phi_2(z) - \frac{c_2}{c_2 - c_1} \phi_1(z). \end{aligned}$$

Similarly,

$$b_2(z) = \frac{1}{c_2 - c_1} \phi_2(z) - \frac{c_1}{c_2 - c_1} \phi_1(z).$$

Thus we obtain

$$\begin{aligned} u^k = u^{k-1} &- h_k K \phi_1(-h_k K) u^{k-1} \\ &+ h_k \left(\frac{1}{c_1 - c_2} \phi_2(-h_k K) - \frac{c_2}{c_1 - c_2} \phi_1(-h_k K) \right) f_1 \\ &+ h_k \left(\frac{1}{c_2 - c_1} \phi_2(-h_k K) - \frac{c_1}{c_2 - c_1} \phi_1(-h_k K) \right) f_2, \end{aligned}$$

a standard two-point exponential integrator.

To connect these existing results to the IPG scheme, first note that (8.12c) is exactly the variation of constants formula (8.13) (which is also as expected from the endpoint exactness result of Proposition 8.1). Hence the above-described standard exponential integrator formulas can be used to compute the IPG fluxes $\hat{u}_h$ at the time steps t_k. It remains to discuss how to compute the solution u_h in between.

8.4. An exponential integrator for interior solution in between time steps

Going beyond the classical exponential integration schemes, we now discuss a new feature arising from the DPG method, namely the capability to also compute an interior solution field that represents the L^2-projection of the solution onto the polynomial spaces within the intervals $[t_{k-1}, t_k]$. To this end, we obtain a discrete scheme from the system (8.12b) using the following result. Its proof is an elementary but lengthy calculation which can be found in Muñoz-Matute, Pardo and Demkowicz (2021), and is omitted here. The result allows us to compute the optimal test functions using the standard ϕ functions.

Proposition 8.4. The following relations between optimal test functions (8.11) and the ϕ functions hold:

$$\begin{aligned} v_r(M,0) &= \sum_{j=0}^{r} \frac{r!}{j!}(-1)^{r-j}\,\phi_{r-j+1}(-M), \\ \int_0^1 v_r(M,t)\,t^q\,dt &= q!\sum_{j=0}^{r}\frac{r!}{j!}(-1)^{r-j}\,\phi_{r-j+q+2}(-M) \end{aligned} \tag{8.14}$$

for any M in $\mathbb{C}^{m\times m}$ (including the scalar case $m=1$).

Example 8.5 (A one-point integrator for the interior IPG solution). Utilizing Proposition 8.4, the system (8.12b) for $p=0$ reduces to the following scheme:

$$u^k_{h,0} = \phi_1(-h_kK)\hat{u}^{k-1}_h + h_k\phi_2(-h_kK)f_1.$$

It computes a constant interior solution given from $\hat{u}^{k-1}_h$ and f_1 that is guaranteed to equal the mean of the exact solution.

Example 8.6 (A two-point integrator for the interior IPG solution). For $p=2$, the system (8.12b) for the two coefficients of the interior solution reduces to the following system of two equations after applying Proposition 8.4:

$$u^k_{h,0} + \frac{1}{2}u^k_{h,1} = g_1\big(\hat{u}^{k-1}_h, f_1, f_2\big), \tag{8.15a}$$

$$\frac{1}{2}u^k_{h,0} + \frac{1}{3}u^k_{h,1} = g_2\big(\hat{u}^{k-1}_h, f_1, f_2\big), \tag{8.15b}$$

where

$$\begin{aligned} g_1\big(\hat{u}^{k-1}_h, f_1, f_2\big) &= \phi_1(-h_kK)\hat{u}^{k-1}_h \\ &+ h_k\left(\frac{1}{c_1-c_2}\phi_3(-h_kK) - \frac{c_2}{c_1-c_2}\phi_2(-h_kK)\right)f_1 \\ &+ h_k\left(\frac{1}{c_2-c_1}\phi_3(-h_kK) - \frac{c_1}{c_2-c_1}\phi_2(-h_kK)\right)f_2, \end{aligned}$$

and

$$\begin{aligned} g_2\big(\hat{u}^{k-1}_h, f_1, f_2\big) &= \phi_1(-h_kK)\hat{u}^{k-1}_h - \phi_2(-h_kK)\hat{u}^{k-1}_h \\ &+ h_k\left(\frac{1}{c_1-c_2}(\phi_3(-h_kK)-\phi_4(-h_kK)) - \frac{c_2}{c_1-c_2}(\phi_2(-h_kK)-\phi_3(-h_kK))\right)f_1 \\ &+ h_k\left(\frac{1}{c_2-c_1}(\phi_3(-h_kK)-\phi_4(-h_kK)) - \frac{c_1}{c_2-c_1}(\phi_2(-h_kK)-\phi_3(-h_kK))\right)f_2. \end{aligned}$$

The equations of (8.15) give the best L^2-approximation of the interior solution in the space of linear functions.

8.5. A posteriori *error estimation*

Now that we have an interior solution, it is possible to get an error representation through the DPG residual. Indeed, equation (3.4a) for the error representation ε now takes the form

$$\begin{aligned}(\varepsilon, v)_Y &= \ell(v) - b((u_h, \hat{u}_h), v)\\ &= (f, v)_\Omega + u_0\,\overline{v(0)} - \left[(u_h, A^* v)_\Omega + \hat{u}_h\,\overline{v(1)}\right]\end{aligned}$$

for all $v \in Y$. For this example, it is possible to derive an explicit formula for function ε, as shown in Muñoz-Matute, Demkowicz and Pardo (2022). However, applying the formula requires coming up with special quadrature rules for matrix-valued functions and it is cumbersome to use. Instead, it is recommended to compute an inexact error representation ε^r using (6.7a), namely

$$(\varepsilon^r, v)_Y = \ell(v) - b((u_h, \hat{u}_h), v) \quad \text{for all } v \in Y^r, \tag{8.16}$$

with a Y^r obtained by enlarging Y_h^{opt} by at least one linearly independent function. (One may, for example, set $Y^r = T(P_{p+1}(\Omega) \times \mathbb{C}^m) \supset Y_h^{\text{opt}}$.) Solving for ε^r from (8.16) then only involves solving a small linear system after the computation of u_h and $\hat{u}_h$. Moreover, ε^r is almost as good an error estimator as ε because of the following result.

Proposition 8.7 (Error estimator for time integrator). Set b and ℓ as in (8.3), $X_h = P_p(\Omega) \times \mathbb{C}^m$, $\Omega = (0, 1)$, and using any $Y^r \supset Y_h^{\text{opt}}$, solve the practical DPG method (6.7) for $x_h \in X_h$ and $\varepsilon^r \in Y^r$. Then x_h coincides with the solution $u_h, \hat{u}_h$ of the IPG method (8.9). Moreover,

$$\|\varepsilon^r\|_Y^2 \le \|\varepsilon\|_Y^2 \le \|\varepsilon^r\|_Y^2 + \operatorname{osc}(\ell)^2, \tag{8.17}$$

$$\|\varepsilon^r\|_Y \le \|u - u_h\|_\Omega \le \|\varepsilon^r\|_Y + \operatorname{osc}(\ell), \tag{8.18}$$

where $\operatorname{osc}(\ell) = \|\ell \circ (I - P_{Y^r})\|_{Y^*}$ and P_{Y^r} denotes the Y-orthogonal projection onto Y^r.

Proof. The stated results follow from discussions in Examples 5.3 and 6.5. Estimate (8.17) is immediate from (6.12). Furthermore, since the norm choice in (8.4b) makes $b((w, \hat{w}), y)$ into a generalized duality pairing, by Proposition 3.6, we know that $\|b\| = 1$ and $\gamma = 1$. Hence (6.9) implies

$$\|x - x_h\|_X \le \|\varepsilon^r\|_Y + \operatorname{osc}(\ell), \quad \|\varepsilon^r\|_Y \le \|x - x_h\|_X.$$

By the endpoint exactness of Proposition 8.1, $\|x - x_h\|_X = \|u - u_h\|_\Omega$, so (8.18) also follows. □

Adapting Proposition 8.7 to each time interval $[t_k, t_{k+1}]$, we obtain a practical strategy for adaptive step size control. The *unit constants* in (8.17)–(8.18) are notable and point to the effectiveness of the strategy. Note, however, that we have not stated any guarantee for $\operatorname{osc}(\ell)$ to be small. We would need to ensure that Y^r

contains enough functions to provide some approximation properties before we can quantitatively characterize the smallness of $\operatorname{osc}(\ell)$.

Bibliographical notes. The main ideas of this section are taken from Muñoz-Matute *et al.* (2021) and Muñoz-Matute *et al.* (2022). Our presentation here is slightly different and shorter. The DPG exponential integrator has also been recently extended to nonlinear problems in Muñoz-Matute and Demkowicz (2024).

9. Duality in DPG formulations

This section is devoted to formulations that are dual in a certain sense to the hybrid DPG formulations. We motivate the construction of the dual formulation using overdetermined and underdetermined systems, and provide typical applications of the dual problem, including the Aubin–Nitsche duality argument for estimating error in weaker norms, and error bounds for goal functionals. In the DPG context, the regularity of dual solutions can be a limiting factor. Even when all solutions of the DPG formulation are highly regular, the dual solutions may have very limited regularity.

9.1. Overdetermined and underdetermined equations

We have been occupied with the solution of the operator equation

$$Bx = \ell, \tag{9.1}$$

given $\ell \in Y^*$ and given $B\colon X \to Y^*$, the operator generated by the form $b(\cdot,\cdot)$ introduced and used in Section 3 (see e.g. (3.10)). Also using the adjoint B^* and the Riesz operators R_X and R_Y introduced there, consider the following two systems of operator equations. The first seeks $x \in X$ and $\zeta \in Y$ solving

$$\begin{aligned} R_Y\zeta + Bx &= \ell, \\ B^*\zeta \qquad &= 0. \end{aligned} \tag{9.2}$$

The second seeks $x \in X$ and $\lambda \in Y$ solving

$$\begin{aligned} R_X x + B^*\lambda &= 0, \\ Bx \qquad &= \ell. \end{aligned} \tag{9.3}$$

The system (9.3) is related to (9.1) since its second equation is identical to (9.1). The system (9.2) is also related to (9.1), since whenever x solves (9.1), it also solves (9.2) with $\zeta = 0$. Let us begin by studying in what sense these formulations are twin relatives of the same problem (9.1).

Suppose the inf-sup condition (1.2a) holds, but we do not know if the uniqueness condition (1.2b) holds. The inf-sup condition (1.2a) is the same as

$$\|Bz\|_{Y^*} \geq \gamma \|z\|_X \quad \text{for all } z \in X, \tag{9.4}$$

which is also equivalent to asserting that B is injective and that the range of B is

closed. But we do not know if B is surjective. Therefore we can only expect $Bx = \ell$ to be solvable if $\ell \in \operatorname{range}(B)$. Since range B equals the annihilator of the null space of B^*, a necessary compatibility condition for solvability of $Bx = \ell$ is that

$$\ell(y) = 0 \quad \text{for all } y \in \ker(B^*). \tag{9.5}$$

For general ℓ, the equation $Bx = \ell$ represents an *overdetermined* system.

Nonetheless, the inf-sup condition (9.4) immediately implies that (9.2) is uniquely solvable, by the standard theory of mixed methods; see e.g. Brezzi and Fortin (1991) or Ern and Guermond (2021). Since (9.2) uniquely solves for x even when $Bx = \ell$ is not solvable, we may interpret (9.2) as a regularized version of $Bx = \ell$. Indeed, (9.2) solves for x satisfying

$$B^* R_Y^{-1} Bx = B^* R_Y^{-1} \ell, \tag{9.6}$$

as can be seen by eliminating ζ from (9.2). When (9.4) holds, (9.6) can be solved for x even when $Bx = \ell$ cannot be solved.

Now suppose the adjoint inf-sup condition (1.3a) holds, but we do not know if the adjoint uniqueness condition (1.3b) holds. Note that (1.3a) is the same as

$$\|B^* y\|_{X^*} \geq \gamma \|y\|_Y \quad \text{for all } y \in Y. \tag{9.7}$$

By the Closed Range Theorem, (9.7) implies that B is surjective, but we do not know that B is injective. In other words, $Bx = \ell$ is solvable for any $\ell \in Y^*$, but its solution need not be unique in general. Hence, in this case, $Bx = \ell$ represents an *underdetermined* system.

As in the previous case, the inf-sup condition (9.7) immediately implies, by standard mixed method theory, that (9.3) is uniquely solvable. Eliminating λ, we find that the unique x it solves for is given by

$$x = -R_X^{-1} B^* \left(B R_X^{-1} B^* \right)^{-1} \ell.$$

This solution is orthogonal to $\ker B$ and has the least norm among all possible solutions of $Bx = \ell$.

9.2. *Relationship to the DPG method and a dual DPG* method*

Define $a \colon (X \times Y) \times (X \times Y) \to \mathbb{C}$ by

$$a((x, \zeta), (z, y)) = (\zeta, y)_Y + b(x, y) + \overline{b(z, \zeta)}$$

for all $x, z \in X$ and $\zeta, y \in Y$ and suppose $F \in (X \times Y)^*$ is given. Equation (9.2) can then be written as

$$a((x, \zeta), (z, y)) = F(z, y) \quad \text{for all } z \in X,\ y \in Y, \tag{9.8}$$

with $F(z, y) = \ell(y)$. Using subspaces $X_h \subset X$ and $Y^r \subset Y$ satisfying the discrete version of the inf-sup condition (1.2a),

$$1 \lesssim \inf_{0 \neq z \in X_h} \sup_{0 \neq y \in Y^r} \frac{|b(z, y)|}{\|z\|_X \|y\|_Y}, \tag{9.9}$$

consider the discrete problem to find $x_h \in X_h$ and $\zeta^r \in Y^r$ satisfying

$$a((x_h, \zeta^r), (z, y)) = F(z, y) \quad \text{for all } z \in X_h,\ y \in Y^r. \tag{9.10}$$

From the standard theory of mixed methods in Brezzi and Fortin (1991), we obtain quasioptimality of the method (9.10). To summarize, *suppose the exact inf-sup condition* (1.2a) *and the discrete inf-sup condition* (9.9) *hold. Then* (9.8) *and* (9.10) *are uniquely solvable for any* $F \in (X \times Y)^*$*, and their solutions satisfy*

$$\|x - x_h\|_X + \|\zeta - \zeta^r\|_Y \lesssim \inf_{z_h \in X_h, y^r \in Y^r} \big[\|x - z_h\|_X + \|\zeta - y^r\|_Y\big]. \tag{9.11}$$

This has implications for both the DPG method and a dual DPG* method defined shortly. First, the DPG method, in the form of the mixed system in Theorem 6.3(b), is a discretization of (9.2), or its equivalent form (9.8), with

$$F(z, y) = \ell(y). \tag{9.12}$$

Hence, once the inf-sup conditions (1.2a) and (9.9) are verified, *the DPG method in the mixed form* (6.7) *can be used to regularize and solve overdetermined systems*, even when it is not possible to verify the uniqueness assumption (1.2b) or the compatibility condition (9.5). Moreover, if B is a continuous bijection (so that the system is no longer overdetermined) and F is as in (9.12), then it is easy to see that $\zeta = 0$, and that $\zeta^r = \varepsilon^r$ together with x_h solves the DPG method (6.7). Then (9.11) reduces to

$$\|x - x_h\|_X + \|\varepsilon^r\|_Y \lesssim \inf_{z_h \in X_h, y^r \in Y^r} \|x - z_h\|_X, \tag{9.13}$$

an error estimate we can also conclude from the theory in prior sections.

Next, consider dual formulations of (9.8). Since the operator generated by the form $a(\cdot, \cdot)$ is self-adjoint, 'dual problems' of (9.8) take the same form as (9.8). By a *DPG* method* we mean the method (9.10) for the case

$$F(z, y) = g(z),$$

where $g \in X^*$. To distinguish from the previous case, let us now rename ζ^r as ξ^r and x_h as λ_h. We can then rewrite (9.10) to express the DPG* method as the method that finds $\xi^r \in Y^r$ and $\lambda_h \in X_h$ satisfying

$$(\xi^r, y)_Y + b(\lambda_h, y) = 0 \qquad \text{for all } y \in Y^r, \tag{9.14a}$$

$$\overline{b(z, \xi^r)} = g(z) \quad \text{for all } z \in X_h. \tag{9.14b}$$

Now it is evident that this is a discretization of (9.3) with the roles of X and Y reversed, B^* in place of B, ξ in place of x, and g in place of ℓ, that is,

$$(\xi, y)_Y + b(\lambda, y) = 0 \qquad \text{for all } y \in Y, \tag{9.15a}$$

$$\overline{b(z, \xi)} = g(z) \quad \text{for all } z \in X, \tag{9.15b}$$

thus revealing the connection with underdetermined systems. By verifying the exact same inf-sup conditions as for the DPG method, namely (1.2a) and (9.9),

the estimate (9.11) then gives that the DPG* method is uniquely solvable and the solution satisfies

$$\|\lambda - \lambda_h\|_X + \|\xi - \xi^r\|_Y \lesssim \inf_{z_h \in X_h, y^r \in Y^r} \big[\|\lambda - z_h\|_X + \|\xi - y^r\|_Y\big]. \tag{9.16}$$

An important difference between the DPG* estimate (9.16) and the DPG estimate (9.13) is that convergence in (9.16) depends on the regularity of an extraneous Lagrange multiplier λ.

9.3. Error in goal functionals

A typical application of duality is in characterizing the error in a goal functional or in goal-oriented adaptivity. Let G be a continuous linear functional on X such that $G(x)$ represents a goal of interest that depends on the solution x. After computing x_h by the DPG method, we obtain an approximate goal $G(x_h)$. We are interested in bounding the error $G(x) - G(x_h)$. The dual formulation of DPG* method is useful in this context.

Theorem 9.1 (Error in goal functional). Let $x \in X$ solve (1.1) and $x_h \in X_h$ solve the DPG discretization (5.4). Let $\xi \in Y$ and $\xi^r \in Y^r$ be as in the DPG* formulations (9.15) and (9.14) with $g(z) = \overline{G(z)}$. Then the error in the goal functional is given by

$$G(x) - G(x_h) = b(x - x_h, \xi - \xi^r). \tag{9.17}$$

Proof. First note that

$$\begin{aligned} b(x - x_h, \xi^r) &= -(\varepsilon^r, \xi^r)_Y && \text{by subtracting (6.7a) from (1.1)} \\ &= -\overline{b(\lambda_h, \varepsilon^r)} && \text{by (9.14a)} \\ &= 0 && \text{by (6.7b).} \end{aligned} \tag{9.18}$$

Hence

$$\begin{aligned} G(x - x_h) &= b(x - x_h, \xi) && \text{by (9.15b)} \\ &= b(x - x_h, \xi - \xi^r), \end{aligned} \tag{9.19}$$

and the result follows. □

An identity analogous to (9.17) holds for the error in the goal when using the standard Galerkin method, where we have the additional freedom to choose one of x_h or ξ^r arbitrarily from the corresponding finite element space. An analogous freedom exists in the DPG case as well. Since subtracting (5.4) from (1.1) gives $b(x - x_h, y_h) = 0$ for all $y_h \in T^r(X_h)$, we may combine it with (9.19) to obtain

$$G(x) - G(x_h) = b(x - x_h, \xi - T^r w_h), \tag{9.20}$$

an identity that holds for any w_h in X_h. Nonetheless, while obtaining convergence rates from either (9.20) or (9.17), the limiting factor is usually the regularity of the dual solution.

9.4. *Aubin–Nitsche argument for DPG methods*

Aubin–Nitsche duality arguments are typically used in finite element methods to prove higher rates of convergence in weaker norms. We present such an argument, adopting the general hybrid setting of (4.12) and Theorem 4.3, where $X = X_0 \times \hat{X}$ and the solution takes the form $(x, \hat{x})$ with $x \in X_0$ and $\hat{x} \in \hat{X}$. We return to our standard setting where one of (1.1), (1.2) or (1.3) holds, that is, B is a bijection (so we are no longer considering overdetermined or underdetermined systems). Limiting ourselves to showing how a duality argument can potentially yield higher rates of convergence for the solution component x in X_0.

Recall the equivalent mixed form of the DPG method given in (6.7). We rewrite it using the composite sesquilinear form $a(\cdot,\cdot)$, which in the hybrid case takes the form

$$a((x,\hat{x},\varepsilon),(z,\hat{z},y)) = (\varepsilon, y)_Y + b((x,\hat{x}),y) + \overline{b((z,\hat{z}),\varepsilon)}$$

for all $x, z \in X$, $\hat{x}, \hat{z} \in \hat{X}$ and $\varepsilon, y \in Y$. The system (6.7) can be reformulated as the problem of finding $x_h \in X_{h,0} \subset X_0$, $\hat{x}_h \in \hat{X}_h \subset \hat{X}$ and $\varepsilon^r \in Y^r$ satisfying

$$a((x_h,\hat{x}_h,\varepsilon^r),(z_h,\hat{z}_h,y^r)) = \ell(y^r) \tag{9.21}$$

for all $z_h \in X_h$, $\hat{z}_h \in \hat{X}_{h,0}$, $y^r \in Y^r$. The undiscretized version of this equation is to find $x \in X$ such that

$$a((x,\hat{x},0),(z,\hat{z},y)) = \ell(y) \tag{9.22}$$

for all $z \in X_0$, $\hat{z} \in \hat{X}$, $y \in Y$. Recall that ζ in (9.8) equals zero when B is a bijection. Obviously, (9.22) is equivalent to (4.13). Since the operator generated by the form $a(\cdot,\cdot)$ is self-adjoint, dual problems takes the same form, with the roles of test and trial functions reversed.

To detail a specific dual problem of interest, suppose L and Z are Hilbert spaces such that the embeddings

$$Z \subseteq X \times Y \text{ and } X_0 \subseteq L \text{ are continuous.} \tag{9.23a}$$

For any $g \in L$, we consider the 'dual problem' for

$$\xi_g = (x_g, \hat{x}_g, \varepsilon_g) \in X_0 \times \hat{X} \times Y$$

satisfying

$$a((z,\hat{z},y),\ \xi_g) = (z,g)_L \quad \text{for all } z \in X_0,\ \hat{z} \in \hat{X},\ y \in Y. \tag{9.23b}$$

The right-hand side is a continuous linear functional on X by (9.23a). Suppose there is a $c(h) > 0$ such that for any $g \in L$, there is an $x_g \in L$ and $\varepsilon_g \in Y$ satisfying (9.23b) and

$$\inf_{\varsigma_h \in X_h \times Y^r} \|\xi_g - \varsigma_h\|_{X\times Y} \le c(h)\|g\|_L. \tag{9.23c}$$

In examples, one would want to leverage regularity of the dual solution, if available, to verify (9.23c) and obtain some $c(h)$ that goes to zero as h decreases.

Theorem 9.2 (Duality argument for DPG formulations). Assume the setting of (9.23) and (4.12). Then

$$\|x - x_h\|_L \le c(h)\,\|a\|\,\|(x, \hat{x}, 0) - (x_h, \hat{x}_h, \varepsilon^r)\|_{X_0 \times \hat{X} \times Y}. \tag{9.24}$$

Proof. Subtracting (9.22) from (9.21),

$$a((x - x_h, \hat{x} - \hat{x}_h, 0 - \varepsilon^r), (z_h, \hat{z}_h, y^r)) = 0 \tag{9.25}$$

for all $z_h \in X_h, \hat{z}_h \in \hat{X}_{h,0}, y^r \in Y^r$. Next, we use (9.23b) with $g = x - x_h \in X_0 \subseteq L$, $z = x - x_h$, $\hat{z} = \hat{x} - \hat{x}_h$ and $y = -\varepsilon^r$, to get

$$\begin{aligned}
\|x - x_h\|_L^2 &= a((x - x_h, \hat{x} - \hat{x}_h, -\varepsilon^r), \xi_g) \\
&= a((x - x_h, \hat{x} - \hat{x}_h, -\varepsilon^r), \xi_g - \varsigma_h) && \text{by (9.25)} \\
&\le \|a\|\,\|(x - x_h, \hat{x} - \hat{x}_h, \varepsilon^r)\|_{X_0 \times \hat{X} \times Y} \|\xi_g - \varsigma_h\|_{X_0 \times \hat{X} \times Y}
\end{aligned}$$

for any $\varsigma_h \in X_{0,h} \times \hat{X}_h \times Y^r$. Hence the result follows from (9.23c). □

Example 9.3 (The dual of a primal DPG formulation on a convex domain). The primal DPG method for the Laplace equation of Example 5.5 offers a simple example of how one can determine the regularity of the dual solutions, assuming that the domain Ω is convex. Recall that there we have set $X_0 = \mathring{H}^1(\Omega)$, $\hat{X} = H^{-1/2}(\partial\Omega_h)$ and $Y = H^1(\Omega_h)$. Additionally, set

$$L = L^2(\Omega), \quad Z = (H^2(\Omega) \cap X_0) \times \hat{X} \times (H^2(\Omega) \cap Y).$$

Then (9.23a) is obvious. The dual problem (9.23b) for $\xi_g = (x_g, \hat{x}_g, \varepsilon_g) \in \mathring{H}^1(\Omega) \times H^{-1/2}(\partial\Omega_h) \times H^1(\Omega_h)$, after complex conjugations as needed, reads as follows:

$$(\varepsilon_g, y)_Y + (\operatorname{grad} x_g, \operatorname{grad} y)_h - \langle \hat{x}_g, y \rangle_h = 0, \tag{9.26a}$$

$$(\operatorname{grad} \varepsilon_g, \operatorname{grad} z)_h = (g, z)_\Omega, \tag{9.26b}$$

$$\langle \hat{z}, \varepsilon_g \rangle_h = 0 \tag{9.26c}$$

for all $y \in H^1(\Omega_h), w \in \mathring{H}^1(\Omega)$ and $\hat{z} \in H^{-1/2}(\partial\Omega_h)$.

We need to understand the regularity of solutions of (9.26). First, note that the ε_g component in $H^1(\Omega_h)$ is actually in $\mathring{H}^1(\Omega)$, as seen from (9.26c) after applying Theorem 4.6(a). Together with (9.26b), we conclude that

$$-\Delta\varepsilon_g = g \quad \text{on } \Omega, \tag{9.27a}$$

$$\varepsilon_g = 0 \quad \text{on } \partial\Omega. \tag{9.27b}$$

Next, observe that equation (9.26a) with $y \in \mathring{H}^1(\Omega)$ yields

$$(\operatorname{grad} x_g, \operatorname{grad} y) = -(\varepsilon_g, y)_h - (\operatorname{grad} \varepsilon_g, \operatorname{grad} y)_h = -(\varepsilon_g, y)_h + (\Delta\varepsilon_g, y)_h,$$

which implies $\Delta x_g = \varepsilon_g + g$. Finally, using the equations for x_g and ε_g in (9.26a)

and integrating by parts, we find $\langle \hat{x}_g, y\rangle_h = \langle n \cdot \operatorname{grad}(\varepsilon_g + x_g), y\rangle_h$. Hence

$$\Delta x_g = \varepsilon_g + g \qquad \text{on } \Omega, \tag{9.27c}$$
$$x_g = 0 \qquad \text{on } \partial\Omega, \tag{9.27d}$$
$$\hat{x}_g = n \cdot \operatorname{grad}(\varepsilon_g + x_g) \qquad \text{on } \partial K, \text{ for all } K \in \Omega_h. \tag{9.27e}$$

At this point we are able to use the well-known full regularity of the Dirichlet problem on a convex domain (see e.g. Grisvard 1985), to conclude that

$$\|\varepsilon_g\|_{H^2(\Omega)} \lesssim \|g\|_\Omega,$$
$$\|x_g\|_{H^2(\Omega)} \lesssim \|\varepsilon_g\|_\Omega + \|g\|_\Omega \lesssim \|g\|_\Omega,$$

which in turn also implies that the interface variable satisfies

$$\begin{aligned}\|\hat{x}_g\|_{H^{-1/2}(\partial\Omega_h)} &\le \|\operatorname{grad}(\varepsilon_g + x_g)\|_{H(\mathrm{div},\Omega)} \\ &= \|\operatorname{grad}(\varepsilon_g + x_g)\|_\Omega + \|\Delta(\varepsilon_g + x_g)\|_\Omega \lesssim \|g\|_\Omega.\end{aligned}$$

Hence we have shown the regularity estimate

$$\|\xi_g\|_Z = \|(x_g, \hat{x}_g, \varepsilon_g)\|_Z \lesssim \|g\|_\Omega. \tag{9.28}$$

To complete the verification of (9.23c), we now only need to bound approximation errors. By an application of the Bramble–Hilbert lemma as in Example 5.5, it is easy to show that there is an interpolant $\xi_{g,h} \equiv (x_{g,h}, \hat{x}_{g,h}, \varepsilon_{g,h})$ of $\xi_g = (x_g, \hat{x}_g, \varepsilon_g)$ such that

$$\|\xi_g - \xi_{g,h}\|_{X_0\times\hat{X}\times Y} \lesssim h\big(\|\varepsilon_g\|_{H^2(\Omega)} + \|x_g\|_{H^2(\Omega)}\big) \lesssim h\|g\|_\Omega,$$

where the last inequality followed from (9.28). This verifies (9.23c) with $c(h) = h$. Applying Theorem 9.2, we obtain

$$\|u - u_h\|_\Omega \le Ch\big(\|u - u_h\|_{H^1(\Omega)} + \|\hat{q}_n - \hat{q}_{n,h}\|_{H^{-1/2}(\partial\Omega_h)}\big),$$

which shows that on a convex domain we expect to obtain an L^2-convergence rate of one higher order for u_h than the H^1-rate we proved earlier in (5.25).

Bibliographical notes. The DPG* method was introduced in Demkowicz, Gopalakrishnan and Keith (2020), motivated by the $\mathcal{L}\mathcal{L}^*$ method (or the 'FOSLL* method') of Cai, Manteuffel, McCormick and Ruge (2001). Numerical experiments in Demkowicz *et al.* (2020, § 5.3) include a case where the λ in (9.16) is in $H^3(\Omega)$ while x is much more regular. It confirms that both the DPG* and the $\mathcal{L}\mathcal{L}^*$ methods have convergence rates that are limited by the regularity of λ. The argument of Theorem 9.1 can be found in Keith (2018). Such arguments are leveraged for goal-oriented adaptivity in Keith *et al.* (2019). Theorem 9.2 and its application to the primal DPG formulation in Example 9.3 are from Bouma *et al.* (2014). A further example applying the duality argument to an ultraweak DPG formulation can be found in Führer (2018).

10. Pointers to DPG techniques for nonlinear problems

Exploitation of DPG ideas to nonlinear problems is an active area of current research. In this section we discuss a DPG extension to nonlinear problems by the steepest descent method. First, however, we quickly give pointers to existing literature containing various other ways of utilizing DPG ideas in nonlinear problems.

10.1. Prior literature

A natural avenue for dealing with nonlinearities is the use of Newton–Raphson iterations that linearizes the nonlinear problem and applies the prior DPG ideas to the linearized problem. Many have adopted this avenue (Chan, Demkowicz and Moser 2014*a*, Roberts, Demkowicz and Moser 2015), and this approach is related to the classical Gauss–Newton method mentioned below in Section 10.3. A PDE-constrained residual minimization problem for solving nonlinear systems was formulated in Bui-Thanh and Ghattas (2014). They combined it with a trust-region inexact Newton conjugate gradient iteration to solve two-dimensional Burgers and Euler equations.

A nonlinear mixed problem cast in the residual minimization DPG framework can be found in Carstensen, Bringmann, Hellwig and Wriggers (2018) for a model nonlinear diffusion problem. They studied it in the context of the primal formulation and lowest-order approximations; their work includes *a priori* and *a posteriori* error estimates as well as an equivalent least-squares formulation, and is illustrated with two-dimensional examples involving adaptivity. A different approach for the same model problem was taken by Cantin and Heuer (2018), who, by introducing additional unknowns, reformulated the nonlinear problem as a linear one with a nonlinear algebraic constraint. The DPG technology is then used only for the linear problem, and the nonlinear constraint is enforced by penalization. The resulting system is an extension of the mixed form (6.7) of the DPG method to a saddle-point formulation with a strongly monotone diagonal block that is wellposed under appropriate conditions. Nonlinearities in the same block also arise in the work of Muga and van der Zee (2020) on residual minimization without a Hilbert structure through Banach duality maps.

Other lines of DPG research involving nonlinearities include problems characterized by variational inequalities such as contact problems in elasticity. Führer, Heuer and Stephan (2018*a*) have developed a DPG theory for (scalar) Signorini-type problems, where optimal test functions are used for discretizing the partial differential operator of the problem and duality terms are added to incorporate the nonlinear boundary conditions. This yields a variational inequality of the first kind. By using an ultraweak formulation they have direct access to normal derivatives through one of the trace variables (unlike standard weak formulations). They also derived reliable error estimators consisting of an error representation as in previous sections, plus a duality term measuring the violation of the complementarity condition.

10.2. Steepest descent iteration

We now discuss how to extend the residual minimization methodology to general nonlinear problems in the framework of the steepest descent method, borrowed from Li (2024). The same technique was applied much earlier by Bristeau *et al.* (1979, 1985) to solve the challenging transonic flow problem. Although nonlinear problems deserve to be set in Banach spaces, for simplicity we limit ourselves to the Hilbert space setting. Unlike the remainder of this paper, here we assume that all spaces are over $\mathbb{R}$, and that $B\colon X \to Y^*$ is a *nonlinear* operator generated by a form $b(x; y)$ which is nonlinear in x and linear in y via

$$B(x)(y) = b(x; y) \tag{10.1}$$

for any $x \in X$ and $y \in Y$.

We are interested in solving the nonlinear analogue of (1.1) using the b in (10.1), which we now recast as the problem of approximating a minimizer of

$$\min_{x\in X} \frac{1}{2}\|\ell - B(x)\|^2_{Y^*}, \tag{10.2}$$

given some $\ell \in Y^*$. Define nonlinear maps $C\colon X \to Y$ and $J\colon X \to \mathbb{R}$ by $C(x) = R_Y^{-1}(\ell - B(x))$ and $J(x) = \frac{1}{2}\|C(x)\|^2$. Then finding a minimizer in (10.2) is the same as finding

$$x = \arg\min_{w\in X} J(w). \tag{10.3}$$

Here we have used the isometry of the Riesz map R_Y in (3.1).

To compute (Gateaux) derivatives of these nonlinear maps, we use the following notation. For any normed linear spaces U, V and $F\colon U \to V$, we write

$$dF_u(z) \equiv dF_u z = \lim_{t\to 0} \frac{F(u + tz) - F(u)}{t},$$

for any u and z in U, if the limit exists in the topology of V and results in a continuous linear operator $dF_u\colon U \to V$. We proceed assuming that for the previously introduced maps J, C and B, the derivatives $dJ_x\colon X \to \mathbb{R}$, $dB_x\colon X \to Y^*$ and $dC_x\colon X \to Y$ exist at any $x \in X$. Note that by definition dJ_x is in X^* (which consists of continuous linear, not antilinear, functionals since X is now over $\mathbb{R}$).

The steepest descent iteration to approximate (10.3) uses the gradient of J, which is an endomorphism $\nabla J\colon X \to X$ defined by

$$(\nabla J)(x) = R_X^{-1} dJ_x. \tag{10.4}$$

Given an initial iterate $x_0 \in X$, the iteration produces x_n by

$$x_{n+1} = x_n - \alpha\,(\nabla J)(x_n), \quad n = 0, 1, 2, \ldots, \tag{10.5}$$

where $0 < \alpha \le 1$ is the step size. Let us compute ∇J. Recall our notation for the adjoint of a linear operator $M\colon X \to Y^*$, namely $M^*\colon Y \to X^*$, obtained after identifying the bidual of a Hilbert space with itself, as already mentioned in (3.9),

namely

$$(M^* y)(z) = (Mz)(y) \tag{10.6}$$

for any $y \in Y$ and $x \in X$ (with no conjugation since the spaces are now over $\mathbb{R}$).

Proposition 10.1. In the above setting, for any $x \in X$,

$$(\nabla J)(x) = -R_X^{-1}(dB_x)^* R_Y^{-1}(\ell - B(x)).$$

Proof. For any $x, z \in X$, since $dC_x z = (dR_Y^{-1}(\ell - B(x)))(z) = -R_Y^{-1} dB_x z$ and $J(x) = \frac{1}{2}(C(x), C(x))_Y$,

$$\begin{aligned} dJ_x z &= (dC_x z, C(x))_Y = (-R_Y^{-1} dB_x z, C(x))_Y \\ &= -(dB_x z)(C(x)) = -((dB_x)^* C(x))(z) \end{aligned}$$

by (10.6). Now the result follows from (10.4). □

By Proposition 10.1, the steepest descent iteration becomes

$$x_{n+1} = x_n + \alpha R_X^{-1}(dB_{x_n})^* R_Y^{-1}(\ell - B(x_n)), \quad n = 0, 1, \ldots. \tag{10.7}$$

It is applicable for DPG formulations when the inverse of the Gram matrices of both the X and the Y inner products can be efficiently applied; see the discussion in Section 10.4.

Example 10.2 (Specialization to the linear case). Suppose the operator B in (10.1) is a linear continuous bijection. Then $dB_x = B$ is independent of x. By Proposition 10.1, the steepest descent iteration (10.5) then becomes

$$x_{n+1} = x_n + \alpha R_X^{-1} B^* R_Y^{-1}(\ell - Bx_n).$$

If x is the exact solution of (1.1), then this can be rewritten as

$$x - x_{n+1} = \left(I - \alpha R_X^{-1} B^* R_Y^{-1} B\right)(x - x_n).$$

Consequently, if the error-reducing operator satisfies

$$\|I - \alpha R_X^{-1} B^* R_Y^{-1} B\| \le q < 1, \tag{10.8}$$

where the norm is the induced operator norm in X, we have a contraction which guarantees the convergence of the iterations. Note that the operator $R_X^{-1} B^* R_Y^{-1} B$ is self-adjoint. Indeed, we have for any $z, w \in X$,

$$\left(R_X^{-1} B^* R_Y^{-1} B z, w\right)_X = \left\langle B^* R_Y^{-1} B z, w \right\rangle_X = \left\langle Bw, R_Y^{-1} B z \right\rangle_Y = \left(R_Y^{-1} B w, R_Y^{-1} B z\right)_Y.$$

Since the last expression is symmetric in z and w, the self-adjointness follows. Additionally, we find that the operator norm in (10.8) is the same as

$$\begin{aligned} \left\|I - \alpha R_X^{-1} B^* R_Y^{-1} B\right\| &= \sup_{w \in X,\ \|w\|_X = 1} \left|\left(w - \alpha R_X^{-1} B^* R_Y^{-1} B w, w\right)_X\right| \\ &= \sup_{w \in X,\ \|w\|_X = 1} \left|\|w\|_X^2 - \alpha \|Bw\|_{Y^*}^2\right|. \end{aligned}$$

Using $\|b\|$ and the inf-sup constant γ introduced in Section 1, we know that for any $w \in X$,

$$\gamma\|w\|_X \le \|Bw\|_{Y^*} \le \|b\|\|w\|_X,$$

which implies

$$(\alpha\gamma^2 - 1)\|w\|^2 \le \alpha\|Bw\|_{Y^*}^2 - \|w\|_X^2 \le (\alpha\|b\|^2 - 1)\|w\|_X^2.$$

This shows that the sufficient condition (10.8) for convergence can be met if

$$-q \le \alpha\gamma^2 - 1 \quad \text{and} \quad \alpha\|b\|^2 - 1 \le q,$$

or equivalently

$$\frac{1-q}{\gamma^2} \le \alpha \le \frac{1+q}{\|b\|^2}.$$

The smallest possible contraction constant

$$q = \frac{\|b\|^2 - \gamma^2}{\|b\|^2 + \gamma^2} = \frac{C^2 - 1}{C^2 + 1}, \quad \text{where } C = \frac{\|b\|}{\gamma},$$

is achieved when the lower and upper bounds for α coincide. To summarize, *selecting any q such that $(C^2 - 1)/(C^2 + 1) \le q < 1$ and setting any α satisfying $(1 - q)/\gamma^2 \le \alpha \le (1 + q)/\|b\|^2$, the steepest descent iterations with such an α converge,* and the error-reducing operator is a contractive map with a contraction constant q or higher.

If X and Y are endowed with optimal norms that make $b(\cdot, \cdot)$ into a generalized duality pairing as in Definition 3.5, then $\gamma = \|b\| = C = 1$. Hence $q = 0$ and the steepest descent method with $\alpha = 1$ delivers the DPG solution in just one step, independently of the initial iterate.

10.3. Relation with the Gauss–Newton method

It is useful to compare (10.7) with the Gauss–Newton iterations (see e.g. Nocedal and Wright 2006), which are obtained by linearizing $B(x) = \ell$ around the current iterate. Namely, if x_n is a current iterate, then $\Delta x = x_{n+1} - x_n$ is obtained from the approximation $B(x_n + \Delta x) \approx B(x_n) + dB_{x_n}\Delta x$ by requiring

$$B(x_n) + dB_{x_n}\Delta x = \ell.$$

One can bring in DPG techniques to solve for the increment by minimizing the residual, that is, by finding

$$\Delta x = \arg\min_{w \in X} \frac{1}{2}\|B(x_n) + dB_{x_n} w - \ell\|_{Y^*}. \tag{10.9}$$

As we have shown previously (see e.g. (9.6)), an equivalent way to compute this minimizer is by solving

$$(dB_{x_n})^* R_Y^{-1} dB_{x_n}\, \Delta x = (dB_{x_n})^* R_Y^{-1}(\ell - B(x_n)). \tag{10.10}$$

This is solvable when dB_{x_n} satisfies the inf-sup condition. Comparing (10.10) with the increment $\widetilde{\Delta x}$ of the steepest descent step iteration (10.7) with unit step size $\alpha = 1$, which solves

$$R_X \widetilde{\Delta x} = (dB_{x_n})^* R_Y^{-1}(\ell - B(x_n)),$$

we see that the two iterations coincide with each other provided

$$R_X = (dB_{x_n})^* R_Y^{-1} dB_{x_n},$$

that is, provided we use the *step-dependent energy norm*

$$|||\Delta u|||_X = \|dB_{x_n} \Delta u\|_{Y^*}$$

for the space X (which, of course, is a norm when the linearized operator dB_{x_n} satisfies the inf-sup condition).

10.4. Trade-offs

The steepest descent iteration (10.7) requires the application of the inverse of the Gram matrices of both the X and Y inner products due to the presence of R_X^{-1} and R_Y^{-1} there. In Section 4 we discussed at length the DPG localization techniques to make R_Y^{-1} easy. However, we also need R_X^{-1} to implement (10.5). This is a drawback of the descent approach. Nonetheless, R_X is a linear Hermitian positive definite operator. Moreover, the component spaces of X have norms that are either standard norms or quotient trace norms, which may be treated as norms arising from Schur complements of Gram matrices of standard norms. As shown in Barker, Dobrev, Gopalakrishnan and Kolev (2018), such Schur complements and the corresponding R_X can be efficiently preconditioned using off-the-shelf preconditioners.

Consider the alternative of the Newton–Raphson iterations which, in the context of the minimum residual methodology, translates into the Gauss–Newton method. Here we require the linearization dB_x to satisfy the inf-sup condition. This requirement is absent for the steepest descent methodology and is another reason to opt for it. For example, in nonlinear elasticity, the linearized problem may be singular (such as in buckling) which, in numerics, manifests as bad conditioning. In the steepest descent approach, we need only invert well-conditioned Riesz operators. In both methodologies we need dB_x, but in steepest descent we use it only to compute the load, whereas in the Gauss–Newton approach we also use it to compute and invert the stiffness matrix. Steepest descent naturally provides the possibility of incorporating additional constraints by solving the minimum residual problem with additional constraints. Incorporating the constraints through a penalty method results in a minimal modification of the algorithm and, in particular, allows the use of a penalty term that is only once differentiable; see Bristeau *et al.* (1985). This is a common situation for inequality constraints. Thus the steepest descent methodology appears to be much more robust than Gauss–Newton.

On the negative side, the steepest descent iterations deliver only linear convergence compared with the quadratic convergence of Newton methods. It may be

advantageous to combine the two methodologies into a single algorithm. For example, one may start with the more robust steepest descent iterations and finish with Gauss–Newton iterations once the increments become small enough.

11. Further pointers and conclusion

We conclude this review by giving brief pointers to topics we have not covered. To keep this review manageable, we have omitted details of many DPG-based techniques, including DPG-style residual minimization in non-Hilbert norms (Muga and van der Zee 2020), polygonal elements (Vaziri Astaneh, Fuentes, Mora and Demkowicz 2018), fractional norms (Bacuta, Demkowicz, Mora and Xenophontos 2021*a*), regularization of rough functionals (Millar, Muga, Rojas and Van der Zee 2022), dispersion analysis (Gopalakrishnan *et al.* 2014), eigensolvers using contour integrals (Gopalakrishnan, Grubišić, Ovall and Parker 2020), DPG eigenvalue error indicators (Bertrand, Boffi and Schneider 2023) and connections with nonconforming methods developed in Carstensen *et al.* (2014*b*). Works on residual minimization under constraints include those of Ellis, Demkowicz and Chan (2014) and Ellis, Chan and Demkowicz (2016), who consider elementwise conservation, and Li and Demkowicz (2024), with circulation constraints around holes in the domain.

Discussion of coupling of DPG methods with other methods was also omitted: a variational formulation applying the DPG methodology to coupling boundary integral operators was developed in Heuer and Karkulik (2015), but it led to nonlocal optimal test functions boundary element degrees of freedom. This difficulty was later overcome by Führer, Heuer and Karkulik (2017), who provided a framework to efficiently couple the DPG method to Galerkin boundary element method (BEM) or other numerical methods. Specifics on coupling of DPG and standard finite element methods can be found in Führer, Heuer, Karkulik and Rodríguez (2018*b*), and an application to a singularly perturbed transmission problem can be found in Führer and Heuer (2017). A DPG BEM for hypersingular boundary integral operators in three dimensions can be found in Heuer and Karkulik (2017*a*).

The DPG method has been applied to many applications, including incompressible flows (Roberts, Bui-Thanh and Demkowicz 2014, Roberts *et al.* 2015), compressible flows (Chan *et al.* 2014*a*), the Cahn–Hilliard equation (Valseth, Romkes and Kaul 2021) and shallow water equations (Valseth and Dawson 2022). Applications to elasticity that we have not had a chance to detail include the work on the Kirchhoff–Love model (Führer, Heuer and Niemi 2019) for thin-structure deformation. They discuss conformity for bending moments in $H(\operatorname{div}\operatorname{div})$, the space of symmetric L_2-tensors τ with $\operatorname{div}\operatorname{div}(\tau)$ in L_2, appropriate for problems with non-convex corners. Their analysis and discretization is motivated by the DPG approach for ultraweak formulations; specifically, a conforming discretization of bending moments in Führer *et al.* (2019) was achievable by the restriction to traces, possible by the ultraweak DPG setting. Other works applying DPG ideas to plates

and shells include those of Calo, Collier and Niemi (2014), Führer, Heuer and Niemi (2022, 2023) and Führer, Heuer and Sayas (2020).

DPG ideas are playing an important role in development of parameter-robust methods. The work of Broersen, Dahmen and Stevenson (2018) gave stability estimates for the DPG method applied to the linear transport equation that are uniform in the relative orientation of the local mesh and the advection direction. For singularly perturbed problems, specifically for advection-dominated diffusion, parameter-robust stability was confirmed in Demkowicz and Heuer (2013) and Chan, Heuer, Bui-Thanh and Demkowicz (2014*b*). The case of reaction-dominated diffusion was studied in Heuer and Karkulik (2017*b*).

One of the attractive features of the DPG method is that it only requires the solution of a symmetric positive definite system, even when the original boundary value problem is non-self-adjoint. This can be leveraged in the design of iterative solvers and high performance computing. In Barker *et al.* (2018), one can find specifics on how to combine off-the-shelf algebraic preconditioners effectively to develop highly scalable DPG solvers. A DPG solver for harmonic wave propagation, integrated within an adaptive procedure, through a two-grid-like preconditioner for the conjugate gradient method, was developed in Petrides and Demkowicz (2017) as well as in Badger, Henneking, Petrides and Demkowicz (2023); it exhibits excellent practical efficiency. Connecting DPG with other similar saddle-point least-squares systems, certain solvers are suggested in Bacuta, Hayes and Jacavage (2021*b*). Parameter robustness in DPG solvers is still highly sought after in specific applications and remains an active area of research.

The design of stable spacetime formulations by DPG techniques is another topic we have not detailed in this review, except for the early work in Demkowicz *et al.* (2017), which is applicable beyond spacetime problems. Since then, spacetime DPG formulations for transient waves have been studied in Gopalakrishnan and Sepúlveda (2019), Sepúlveda (2018) and Ernesti and Wieners (2019), and a spacetime DPG method for the heat equation was developed in Diening and Storn (2022). The approach of Demkowicz *et al.* (2017) to prove wellposedness in graph spaces (along the lines of the theory of Friedrichs systems) was found to be difficult for various spacetime problems. A new approach has been proposed in the recent work of Führer, González and Karkulik (2024) using Bochner spaces. This shows promise for reducing the technicalities in proving convergence of DPG and residual minimization methods for spacetime methods.

An exciting new frontier is the use of DPG ideas for variationally correct machine learning approaches. The very recent work of Rojas *et al.* (2024) defines a quadratic loss functional, motivated by DPG-type formulations, within a physics-informed neural network to solve a boundary value problem (and earlier developments in physics-informed neural networks can be found in Kharazmi, Zhang and Karniadakis 2021). The recent work of Bachmayr, Dahmen and Oster (2024) centres around learning the parameter-to-solution map for systems of partial differential equations that depend on a potentially large number of parameters. These works show

emerging techniques based on DPG formulations with a variationally correct residual (measured in a dual norm like the ones we have seen in earlier sections) forming the basis for loss functionals in machine learning. The tools we have developed here can be used to establish that a loss function based on such dual residuals is uniformly proportional to the squared solution error in a mathematically correct norm. Such results show potential to augment machine learning predictions with rigorous *a posteriori* accuracy control. Other recent works that combine machine learning with DPG and residual minimization ideas include Brevis *et al.* (2024) and Brevis, Muga and van der Zee (2022).

These references show the wide variety of topics that the DPG ideas have impacted. The essential theoretical underpinnings of the DPG methodology, discussed earlier, should be enough preparation to delve into the above-mentioned works for further studies. We limited the scope of earlier sections by selecting topics for discussion that have potential applicability to a large variety of boundary value problems. Discussions of specific boundary value problems have been delineated as brief examples throughout, but the cited original sources are recommended for a complete picture of each case.

Acknowledgements

The ideas behind DPG methods were developed over the years, utilizing support from the Air Force Office of Scientific Research (AFOSR) and the National Science Foundation (NSF). Both authors are immensely grateful for this support. The preparation of this review was supported in part by AFOSR grant FA9550-23-1-0103 and NSF grant 2245077. This work also benefited from activities organized under the auspices of NSF RTG grant 2136228 as well as a decadal series of workshops on DPG and residual minimization methods held in Austin (USA, 2013), Delft (The Netherlands, 2015), Portland (USA, 2017), Berlin (Germany, 2019), Santiago (Chile, 2022) and Bilbao (Spain, 2024). The authors gratefully acknowledge feedback from several participants of these workshops which shaped this review.

References

A. H. Al-Mohy and N. J. Higham (2011), Computing the action of the matrix exponential, with an application to exponential integrators, *SIAM J. Sci. Comput.* **33**, 488–511.

D. N. Arnold, R. S. Falk and R. Winther (2006), Finite element exterior calculus, homological techniques, and applications, *Acta Numer.* **15**, 1–155.

I. Babuška (1971), Error-bounds for finite element method, *Numer. Math.* **16**, 322–333.

I. Babuška, A. K. Aziz, G. Fix and R. B. Kellogg (1972), Survey lectures on the mathematical foundations of the finite element method, in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (A. K. Aziz, ed.), Academic Press, pp. 1–359.

M. Bachmayr, W. Dahmen and M. Oster (2024), Variationally correct neural residual regression for parametric PDEs: On the viability of controlled accuracy. Available at arXiv:2405.20065.

C. Bacuta, L. Demkowicz, J. Mora and C. Xenophontos (2021*a*), Analysis of non-conforming DPG methods on polyhedral meshes using fractional Sobolev norms, *Comput. Math. Appl.* **95**, 215–241.

C. Bacuta, D. Hayes and J. Jacavage (2021*b*), Notes on a saddle point reformulation of mixed variational problems, *Comput. Math. Appl.* **95**, 4–18.

J. Badger, S. Henneking, S. Petrides and L. Demkowicz (2023), Scalable DPG multigrid solver for Helmholtz problems: A study on convergence, *Comput. Math. Appl.* **148**, 81–92.

P. E. Barbone and I. Harari (2001), Nearly H^1-optimal finite element methods, *Comput. Methods Appl. Mech. Engrg* **190**, 5679–5690.

A. T. Barker, V. Dobrev, J. Gopalakrishnan and T. Kolev (2018), A scalable preconditioner for a DPG method, *SIAM J. Sci. Comput.* **40**, A1187–A1203.

J. W. Barrett and K. W. Morton (1984), Approximate symmetrization and Petrov–Galerkin methods for diffusion-convection problems, *Comput. Methods Appl. Mech. Engrg* **45**, 97–122.

J. Bergh and J. Löfström (1976), *Interpolation Spaces: An Introduction*, Springer.

F. Bertrand, D. Boffi and H. Schneider (2023), Discontinuous Petrov–Galerkin approximation of eigenvalue problems, *Comput. Methods Appl. Math.* **23**, 1–17.

P. B. Bochev and M. D. Gunzburger (2009), *Least-Squares Finite Element Methods*, Vol. 166 of Applied Mathematical Sciences, Springer.

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

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

J. H. Bramble, R. D. Lazarov and J. E. Pasciak (1997), A least-squares approach based on a discrete minus one inner product for first order systems, *Math. Comp.* **66**, 935–955.

I. Brevis, I. Muga and K. G. van der Zee (2022), Neural control of discrete weak formulations: Galerkin, least squares & minimal-residual methods with quasi-optimal weights, *Comput. Methods Appl. Mech. Engrg* **402**, art. 115716.

I. Brevis, I. Muga, D. Pardo, O. Rodriguez and K. G. van der Zee (2024), Learning quantities of interest from parametric PDEs: An efficient neural-weighted minimal residual approach, *Comput. Math. Appl.* **164**, 139–149.

H. Brezis (2011), *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer.

F. Brezzi (1974), On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge* **8**, 129–151.

F. Brezzi and M. Fortin (1991), *Mixed and Hybrid Finite Element Methods*, Vol. 15 of Springer Series in Computational Mathematics, Springer.

M. O. Bristeau, O. Pironneau, R. Glowinski, J. Périaux, J. P. Perrier and G. Poirier (1979), On the numerical solution of nonlinear problems in fluid dynamics by least squares and finite element methods, I: Least square formulations and conjugate gradient solution of the continuous problems, *Comput. Methods Appl. Mech. Engrg* **17-18**, 619–657.

M. O. Bristeau, O. Pironneau, R. Glowinski, J. Périaux, J. P. Perrier and G. Poirier (1985), On the numerical solution of nonlinear problems in fluid dynamics by least squares and finite element methods, II: Application to transonic flow simulations, *Comput. Methods Appl. Mech. Engrg* **51**, 363–394.

D. Broersen, W. Dahmen and R. P. Stevenson (2018), On the stability of DPG formulations of transport equations, *Math. Comp.* **87**, 1051–1082.
A. Buffa, M. Costabel and D. Sheen (2002), On traces for $\mathbf{H}(\mathbf{curl}, \Omega)$ in Lipschitz domains, *J. Math. Anal. Appl.* **276**, 845–867.
T. Bui-Thanh and O. Ghattas (2014), 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**, 20–40.
Z. Cai, R. Lazarov, T. A. Manteuffel and S. F. McCormick (1994), First-order system least squares for second-order partial differential equations I, *SIAM J. Numer. Anal.* **31**, 1785–1799.
Z. Cai, T. A. Manteuffel, S. F. McCormick and J. Ruge (2001), First-order system $\mathcal{LL}^*$ (FOSLL*): Scalar elliptic partial differential equations, *SIAM J. Numer. Anal.* **39**, 1418–1445.
V. M. Calo, N. O. Collier and A. H. Niemi (2014), Analysis of the discontinuous Petrov–Galerkin method with optimal test functions for the Reissner–Mindlin plate bending model, *Comput. Math. Appl.* **66**, 2570–2586.
P. Cantin and N. Heuer (2018), A DPG framework for strongly monotone operators, *SIAM J. Numer. Anal.* **56**, 2731–2750.
C. Carstensen, P. Bringmann, F. Hellwig and P. Wriggers (2018), Nonlinear discontinuous Petrov–Galerkin methods, *Numer. Math.* **139**, 529–561.
C. Carstensen, L. Demkowicz and J. Gopalakrishnan (2014*a*), *A posteriori* error control for DPG methods, *SIAM J. Numer. Anal.* **52**, 1335–1353.
C. Carstensen, L. Demkowicz and J. Gopalakrishnan (2016), Breaking spaces and forms for the DPG method and applications including Maxwell equations, *Comput. Math. Appl.* **72**, 494–522.
C. Carstensen, D. Gallistl, F. Hellwig and L. Weggler (2014*b*), Low-order dPG-FEM for an elliptic PDE, *Comput. Math. Appl.* **68**, 1503–1512.
A. Celia, T. F. Russell, H. Ismael and R. E. Ewing (1990), An Eulerian–Lagrangian localized adjoint method for the advection–diffusion equation, *Adv. Water Resources* **13**, 187–206.
J. Chan, L. Demkowicz and R. Moser (2014*a*), A DPG method for steady viscous compressible flow, *Comput. Fluids* **98**, 69–90.
J. Chan, N. Heuer, T. Bui-Thanh and L. Demkowicz (2014*b*), Robust DPG method for convection-dominated diffusion problems II: Adjoint boundary conditions and mesh-dependent test norms, *Comput. Math. Appl.* **67**, 771–795.
B. Cockburn and J. Gopalakrishnan (2004), A characterization of hybridized mixed methods for the Dirichlet problem, *SIAM J. Numer. Anal.* **42**, 283–301.
A. Cohen, W. Dahmen and G. Welper (2012), Adaptivity and variational stabilization for convection–diffusion equations, *ESAIM Math. Model. Numer. Anal.* **46**, 1247–1273.
R. Courant and K. O. Friedrichs (1948), *Supersonic Flow and Shock Waves*, Interscience.
L. Demkowicz (2018), Lecture notes on energy spaces. Report 18-13, Oden Institute, The University of Texas at Austin.
L. Demkowicz (2024), *Mathematical Theory of Finite Elements*, Computational Science and Engineering, SIAM.
L. Demkowicz and J. Gopalakrishnan (2010), A class of discontinuous Petrov–Galerkin methods, Part I: The transport equation, *Comput. Methods Appl. Mech. Engrg* **199**, 1558–1572.

L. Demkowicz and J. Gopalakrishnan (2011*a*), Analysis of the DPG method for the Poisson equation, *SIAM J. Numer. Anal.* **49**, 1788–1809.
L. Demkowicz and J. Gopalakrishnan (2011*b*), A class of discontinuous Petrov–Galerkin methods, Part II: Optimal test functions, *Numer. Methods Partial Differential Equations* **27**, 70–105.
L. Demkowicz and J. Gopalakrishnan (2013), A primal DPG method without a first-order reformulation, *Comput. Math. Appl.* **66**, 1058–1064.
L. Demkowicz and J. Gopalakrishnan (2017), Discontinuous Petrov Galerkin (DPG) method, in *Encyclopedia of Computational Mechanics*, second edition, Wiley Computational Mechanics Online.
L. Demkowicz and N. Heuer (2013), Robust DPG method for convection-dominated diffusion problems, *SIAM J. Numer. Anal.* **51**, 2514–2537.
L. Demkowicz and J. T. Oden (1986*a*), An adaptive characteristic Petrov–Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in one space variable, *J. Comput. Phys.* **67**, 188–213.
L. Demkowicz and J. T. Oden (1986*b*), An adaptive characteristic Petrov–Galerkin finite element method for convection-dominated linear and nonlinear parabolic problems in two space variables, *Comput. Methods Appl. Mech. Engrg* **55**, 63–87.
L. Demkowicz and P. Zanotti (2020), Construction of DPG Fortin operators revisited, *Comput. Math. Appl.* **80**, 2261–2271.
L. Demkowicz, J. Gopalakrishnan and B. Keith (2020), The DPG-star method, *Comput. Math. Appl.* **79**, 3092–3116.
L. Demkowicz, J. Gopalakrishnan and A. Niemi (2012*a*), A class of discontinuous Petrov–Galerkin methods, Part III: Adaptivity, *Appl. Numer. Math.* **62**, 396–427.
L. Demkowicz, J. Gopalakrishnan, I. Muga and J. Zitelli (2012*b*), Wavenumber explicit analysis for a DPG method for the multidimensional Helmholtz equation, *Comput. Methods Appl. Mech. Engrg* **213/216**, 126–138.
L. Demkowicz, J. Gopalakrishnan, S. Nagaraj and P. Sepúlveda (2017), A spacetime DPG method for the Schrödinger equation, *SIAM J. Numer. Anal.* **55**, 1740–1759.
L. Diening and J. Storn (2022), A space-time DPG method for the heat equation, *Comput. Math. Appl.* **105**, 41–53.
I. Ekeland and R. Témam (1999), *Convex Analysis and Variational Problems*, Vol. 28 of Classics in Applied Mathematics, SIAM.
T. Ellis, J. Chan and L. Demkowicz (2016), Robust DPG methods for transient convection–diffusion, in *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations* (G. R. Barrenechea *et al.*, eds), Vol. 114 of Lecture Notes in Computational Science and Engineering, Springer.
T. Ellis, L. Demkowicz and J. Chan (2014), Locally conservative discontinuous Petrov–Galerkin finite elements for fluid problems, *Comput. Math. Appl.* **68**, 1530–1549.
A. Ern and J.-L. Guermond (2021), *Finite Elements II*, Springer.
A. Ern, J.-L. Guermond and G. Caplain (2007), An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems, *Commun. Partial Differential Equations* **32**, 317–341.
J. Ernesti and C. Wieners (2019), A space-time discontinuous Petrov–Galerkin method for acoustic waves, in *Space-Time Methods: Applications to Partial Differential Equations* (U. Langer and O. Steinbach, eds), De Gruyter, pp. 89–116.

K. O. Friedrichs (1958), Symmetric positive linear differential equations, *Commun. Pure Appl. Math.* **11**, 333–418.
T. Führer (2018), Superconvergence in a DPG method for an ultra-weak formulation, *Comput. Math. Appl.* **75**, 1705–1718.
T. Führer and N. Heuer (2017), Robust coupling of DPG and BEM for a singularly perturbed transmission problem, *Comput. Math. Appl.* **74**, 1940–1954.
T. Führer and N. Heuer (2019), Fully discrete DPG methods for the Kirchhoff–Love plate bending model, *Comput. Methods Appl. Mech. Engrg* **343**, 550–571.
T. Führer and N. Heuer (2024), Robust DPG test spaces and Fortin operators: The H^1 and $H(\mathrm{div})$ cases, *SIAM J. Numer. Anal.* **62**, 718–748.
T. Führer, R. González and M. Karkulik (2024), Well-posedness of first-order acoustic wave equations and space-time finite element approximation. Available at arXiv:2311.10536.
T. Führer, N. Heuer and M. Karkulik (2017), On the coupling of DPG and BEM, *Math. Comp.* **86**, 2261–2284.
T. Führer, N. Heuer and A. H. Niemi (2019), An ultraweak formulation of the Kirchhoff–Love plate bending model and DPG approximation, *Math. Comp.* **88**, 1587–1619.
T. Führer, N. Heuer and A. H. Niemi (2022), A DPG method for shallow shells, *Numer. Math.* **152**, 76–99.
T. Führer, N. Heuer and A. H. Niemi (2023), A DPG method for Reissner–Mindlin plates, *SIAM J. Numer. Anal.* **61**, 995–1017.
T. Führer, N. Heuer and F.-J. Sayas (2020), An ultraweak formulation of the Reissner–Mindlin plate bending model and DPG approximation, *Numer. Math.* **145**, 313–344.
T. Führer, N. Heuer and E. P. Stephan (2018*a*), On the DPG method for Signorini problems, *IMA J. Numer. Anal.* **38**, 1893–1926.
T. Führer, N. Heuer, M. Karkulik and R. Rodríguez (2018*b*), Combining the DPG method with finite elements, *Comput. Methods Appl. Math.* **18**, 639–652.
V. V. Garg, S. Prudhomme, K. G. van der Zee and G. F. Carey (2014), Adjoint-consistent formulations of slip models for coupled electroosmotic flow systems, *Adv. Model. Simul. Engrg* **2**, art. 15.
J. Gopalakrishnan (2013), Five lectures on DPG methods. Available at arXiv:1306.0557.
J. Gopalakrishnan and W. Qiu (2014), An analysis of the practical DPG method, *Math. Comput.* **83**, 537–552.
J. Gopalakrishnan and P. Sepúlveda (2019), A spacetime DPG method for acoustic waves, in *Space-Time Methods: Applications to Partial Differential Equations* (U. Langer and O. Steinbach, eds), Radon Series on Computational and Applied Mathematics, De Gruyter, pp. 129–154.
J. Gopalakrishnan, L. Grubišić, J. Ovall and B. Q. Parker (2020), Analysis of FEAST spectral approximations using the DPG discretization, *Comput. Methods Appl. Math.* **89**, 203–228.
J. Gopalakrishnan, I. Muga and N. Olivares (2014), Dispersive and dissipative errors in the DPG method with scaled norms for the Helmholtz equation, *SIAM J. Sci. Comput.* **36**, A20–A39.
P. Grisvard (1985), *Elliptic Problems in Nonsmooth Domains*, Vol. 24 of Monographs and Studies in Mathematics, Pitman Advanced Publishing Program.
N. Heuer and M. Karkulik (2015), DPG method with optimal test functions for a transmission problem, *Comput. Math. Appl.* **70**, 1504–1518.

N. Heuer and M. Karkulik (2017*a*), Discontinuous Petrov–Galerkin boundary elements, *Numer. Math.* **135**, 1011–1043.
N. Heuer and M. Karkulik (2017*b*), A robust DPG method for singularly perturbed reaction–diffusion problems, *SIAM J. Numer. Anal.* **55**, 1218–1242.
M. Jensen (2004), Discontinuous Galerkin methods for Friedrichs systems with irregular solutions. PhD thesis, University of Oxford.
T. Kato (1995), *Perturbation Theory for Linear Operators*, Classics in Mathematics, Springer.
B. Keith (2018), New ideas in adjoint methods for PDEs: A saddle-point paradigm for finite element analysis and its role in the DPG methodology. PhD thesis, The University of Texas at Austin.
B. Keith, A. V. Astaneh and L. Demkowicz (2019), Goal-oriented adaptive mesh refinement for discontinuous Petrov–Galerkin methods, *SIAM J. Numer. Anal.* **57**, 1649–1676.
E. Kharazmi, Z. Zhang and G. E. M. Karniadakis (2021), hp-VPINNs: Variational physics-informed neural networks with domain decomposition, *Comput. Methods Appl. Mech. Engrg* **374**, art. 113547.
J. Li (2024), A nonlinear mixed problem framework for discontinuous Petrov Galerkin (DPG) methods. PhD thesis, The University of Texas at Austin.
J. Li and L. Demkowicz (2024), A DPG method for planar div-curl problems, *Comput. Math. Appl.* **159**, 31–43.
A. F. D. Loula and D. T. Fernandes (2009), A quasi optimal Petrov–Galerkin method for Helmholtz problem, *Internat. J. Numer. Methods Engrg* **80**, 1595–1622.
A. F. D. Loula, T. J. R. Hughes and L. P. Franca (1987), Petrov–Galerkin formulations of the Timoshenko beam problem, *Comput. Methods Appl. Mech. Engrg* **63**, 115–132.
J. M. Melenk (1995), On generalized finite element methods. PhD thesis, University of Maryland.
F. Millar, I. Muga, S. Rojas and K. G. Van der Zee (2022), Projection in negative norms and the regularization of rough linear functionals, *Numer. Math.* **150**, 1087–1121.
P. Monk (2003), *Finite Element Methods for Maxwell's Equations*, Numerical Mathematics and Scientific Computation, Oxford University Press.
I. Muga and K. G. van der Zee (2020), Discretization of linear problems in Banach spaces: Residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods, *SIAM J. Numer. Anal.* **58**, 3406–3426.
J. Muñoz-Matute and L. Demkowicz (2024), Multistage discontinuous Petrov–Galerkin time-marching scheme for nonlinear problems, *SIAM J. Numer. Anal.* **62**, 1956–1978.
J. Muñoz-Matute, L. Demkowicz and D. Pardo (2022), Error representation of the time-marching DPG scheme, *Comput. Methods Appl. Mech. Engrg* **391**, art. 114480.
J. Muñoz-Matute, D. Pardo and L. Demkowicz (2021), Equivalence between the DPG method and the exponential integrators for linear parabolic problems, *J. Comput. Phys.* **429**, art. 110016.
J. Nečas (1962), Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle, *Ann. Sc. Norm. Super. Pisa Cl. Sci.* **16**, 305–326.
J.-C. Nédélec (1980), Mixed finite elements in $\mathbb{R}^3$, *Numer. Math.* **35**, 315–341.
J. Niesen and W. M. Wright (2012), Algorithm 919: A Krylov subspace algorithm for evaluating the ϕ-functions appearing in exponential integrators, *ACM Trans. Math. Softw.*

J. Nocedal and S. J. Wright (2006), *Numerical Optimization*, second edition, Springer.
S. Petrides and L. Demkowicz (2017), An adaptive DPG method for high frequency time-harmonic wave propagation problems, *Comput. Math. Appl.* **74**, 1999–2017.
P.-A. Raviart and J. M. Thomas (1977*a*), A mixed finite element method for 2-nd order elliptic problems, in *Mathematical Aspects of Finite Element Methods*, Vol. 606 of Lecture Notes in Mathematics, Springer, pp. 292–315.
P.-A. Raviart and J. M. Thomas (1977*b*), Primal hybrid finite element methods for 2nd order elliptic equations, *Math. Comp.* **31**, 391–413.
N. V. Roberts, T. Bui-Thanh and L. Demkowicz (2014), The DPG method for the Stokes problem, *Comput. Math. Appl.* **67**, 966–995.
N. V. Roberts, L. Demkowicz and R. Moser (2015), A discontinuous Petrov–Galerkin methodology for adaptive solutions to the incompressible Navier–Stokes equations, *J. Comput. Phys.* **301**, 456–483.
S. Rojas, P. Maczuga, J. Muñoz-Matute, D. Pardo and M. Paszyński (2024), Robust variational physics-informed neural networks, *Comput. Methods Appl. Mech. Engrg* **425**, art. 116904.
P. Sepúlveda (2018), Spacetime numerical techniques for the wave and Schrödinger equations. PhD thesis, Portland State University.
D. Sheen (1992), A generalized Green's theorem, *Appl. Math. Lett.* **5**, 95–98.
E. Valseth and C. Dawson (2022), A stable space-time FE method for the shallow water equations, *Comput. Geosci.* pp. 1–18.
E. Valseth, A. Romkes and A. R. Kaul (2021), A stable FE method for the space-time solution of the Cahn–Hilliard equation, *J. Comput. Phys.* **441**, art. 110426.
A. Vaziri Astaneh, F. Fuentes, J. Mora and L. Demkowicz (2018), High-order polygonal discontinuous Petrov–Galerkin (PolyDPG) methods using ultraweak formulations, *Comput. Methods Appl. Mech. Engrg* **332**, 686–711.
J. Xu and L. Zikatanov (2003), Some observations on Babuška and Brezzi theories, *Numer. Math.* **94**, 195–202.
J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo and V. Calo (2011), A class of discontinuous Petrov–Galerkin methods, Part IV: Wave propagation, *J. Comput. Phys.* **230**, 2406–2432.