Research Article

Thomas Führer*

Superconvergent DPG Methods for Second-Order Elliptic Problems

https://doi.org/10.1515/cmam-2018-0250
Received September 25, 2018; revised January 10, 2019; accepted March 5, 2019

Abstract: We consider DPG methods with optimal test functions and broken test spaces based on ultra-weak formulations of general second-order elliptic problems. Under some assumptions on the regularity of solutions of the model problem and its adjoint, superconvergence for the scalar field variable is achieved by either increasing the polynomial degree in the corresponding approximation space by one or by a local postprocessing. We provide a uniform analysis that allows the treatment of different test norms. Particularly, we show that in the presence of convection only the quasi-optimal test norm leads to higher convergence rates, whereas other norms considered do not. Moreover, we also prove that our DPG method delivers the best L^2 approximation of the scalar field variable up to higher-order terms, which is the first theoretical explanation of an observation made previously by different authors. Numerical studies that support our theoretical findings are presented.

Keywords: DPG Method, Ultra-Weak Formulation, Best Approximation, Duality Arguments, Postprocessing, Superconvergence

MSC 2010: 65N30, 65N12

1 Introduction

In this work we investigate convergence rates of DPG methods based on an ultra-weak formulation of second-order elliptic problems stated in the form of the general first-order system

$$\nabla u - \boldsymbol{\beta} u + \boldsymbol{C}\boldsymbol{\sigma} = \boldsymbol{C}\boldsymbol{f} \quad \text{in } \Omega, \tag{1.1a}$$

$$\operatorname{div} \boldsymbol{\sigma} + \gamma u = f \quad \text{in } \Omega, \tag{1.1b}$$

$$u = 0 \quad \text{on } \Gamma := \partial\Omega, \tag{1.1c}$$

where $\Omega \subseteq \mathbb{R}^d$, $d \geq 2$, is a polyhedral domain and $\boldsymbol{C} \in L^\infty(\Omega)^{d\times d}$ denotes a symmetric, uniformly positive definite matrix valued function, $\boldsymbol{\beta} \in L^\infty(\Omega)^d$, $\gamma \in L^\infty(\Omega)$. Throughout we suppose that the coefficients additionally satisfy

$$L^\infty(\Omega) \ni \tfrac{1}{2}\operatorname{div}(\boldsymbol{C}^{-1}\boldsymbol{\beta}) + \gamma \geq 0, \tag{1.2}$$

which implies that for $f \in L^2(\Omega)$, $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega) := L^2(\Omega)^d$ our model problem (1.1) admits a unique solution $(u, \boldsymbol{\sigma})$ with $u \in H^1_0(\Omega)$, $\boldsymbol{\sigma} \in \boldsymbol{H}(\operatorname{div}; \Omega) := \{\boldsymbol{\tau} \in \boldsymbol{L}^2(\Omega) : \operatorname{div} \boldsymbol{\tau} \in L^2(\Omega)\}$. To see this, use (1.1a) in (1.1b) which results in a second-order elliptic problem. Testing with $v \in H^1_0(\Omega)$ gives a bilinear form that is, using (1.2), coercive and, thus, solvability can be obtained by classical arguments.

In this work we consider DPG methods with optimal test functions and broken test spaces, which have been introduced by Demkowicz and Gopalakrishnan, see [5, 6] and also [8, 22]. For a unified stability analysis which also covers our model problem we refer to [3]. We analyze ultra-weak formulations of (1.1), which are

***Corresponding author: Thomas Führer,** Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile, e-mail: tofuhrer@mat.uc.cl. http://orcid.org/0000-0001-5034-6593

obtained by multiplying with locally supported functions and integration by parts, see, e.g., [7] for a Poisson model problem. On the one hand, this has the advantage that the field variables can be sought in $L^2(\Omega)$, since no derivative operator is applied to these unknowns after integration by parts. On the other hand, this requires the introduction of trace variables $\widehat{u}$ and $\widehat{\sigma}$ that live on the skeleton (these unknowns impose weak continuity conditions). However, as analyzed in the recent work [21] the use of ultra-weak formulations also allows to define conforming finite element spaces on polygonal meshes.

The motivation of this work is to analyze superconvergence properties for approximations of the scalar field variable u that have been observed in our recent work [10] for a simple reaction-diffusion problem, where $\boldsymbol{C}$ is the identity matrix, $\boldsymbol{\beta} = 0$, $\gamma = 1$, and $\boldsymbol{f} = 0$. Here we generalize and extend [10] to the model problem (1.1) and introduce new ideas that allow the treatment of different test norms. As in [10], the proofs rely on duality arguments and regularity theory for elliptic PDEs. Such arguments are common when proving higher convergence rates, e.g., the classical Aubin-Nitsche trick, or more recently in variants of DG methods, e.g., [4]. Some early works on convergence in mixed finite element methods include [9, 11, 20].

Let us also mention the recent works [14, 15] that deal with dual problems in the context of DPG methods (the DPG* method and goal-oriented problems). Particularly, we point out the reference [1]. There the authors consider a primal DPG method (without the first-order reformulation) for the Poisson problem and analyze convergence rates (with reduced degrees in test spaces). Moreover, they develop duality arguments and prove that the error in the primal variable u converges at a higher rate when measured in a weaker norm.

1.1 Summary of Results

We seek approximations $u_h \in \mathcal{P}^p(\mathcal{T})$, $\boldsymbol{\sigma}_h \in \mathcal{P}^p(\mathcal{T})^d$ of the field variables u, $\boldsymbol{\sigma}$, where $\mathcal{T}$ is a mesh of simplices and $\mathcal{P}^p(\mathcal{T})$ denotes the space of $\mathcal{T}$-piecewise polynomials of degree less than or equal to $p \in \mathbb{N}_0$, and approximations $\widehat{u}_h$, $\widehat{\sigma}_h$ of the traces $\widehat{u}$, $\widehat{\sigma}$ in spaces that will be defined later on. For sufficient regular solutions basic a priori analysis arguments give the estimate

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \simeq \|u - u_h\| + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\| + \|(\widehat{u} - \widehat{u}_h, \widehat{\sigma} - \widehat{\sigma}_h)\|_S = \mathcal{O}(h^{p+1}),$$

where $\|\cdot\|$ denotes the $L^2(\Omega)$ norm and $\|\cdot\|_S$ is some appropriate norm for the traces. This estimate is optimal, since we seek approximations of u and $\boldsymbol{\sigma}$ in polynomial spaces of the same order and their errors are measured in $L^2(\Omega)$ norms. Nevertheless, it is unsatisfactory to some extent. Consider $\boldsymbol{C}$ the identity, $\boldsymbol{\beta} = 0$, $\boldsymbol{f} = 0$ in (1.1). Then $\boldsymbol{\sigma} = \nabla u$ and we seek approximations of u and its gradient $\boldsymbol{\sigma}$ in polynomial spaces of the same order, which seems to be suboptimal. Fortunately, there exist at least two possibilities to achieve higher convergence rates under some assumptions on the regularity of solutions of (1.1) and its adjoint problem:

- *Augmenting the trial space:* Instead of seeking approximations $u_h \in \mathcal{P}^p(\mathcal{T})$ we seek approximations $u_h^+ \in \mathcal{P}^{p+1}(\mathcal{T})$ and show that
$$\|u - u_h^+\| = \mathcal{O}(h^{p+2}).$$
- *Postprocessing:* We use a common local postprocessing technique (see, e.g., the early works [11, 20]) to obtain an approximation $\widetilde{u}_h \in \mathcal{P}^{p+1}(\mathcal{T})$ and prove that
$$\|u - \widetilde{u}_h\| = \mathcal{O}(h^{p+2}).$$

Based on similar techniques we also provide a proof of the following:

- *DPG for ultra-weak formulations delivers the $L^2(\Omega)$ best approximation up to a higher-order term*, i.e., for the approximation $u_h \in \mathcal{P}^p(\mathcal{T})$ it holds that
$$\|u - u_h\| \leq \|u - \Pi^p u\| + \mathcal{O}(h^{p+2}),$$
where Π^p denotes the $L^2(\Omega)$ projection to $\mathcal{P}^p(\mathcal{T})$.

The latter observation is quite interesting, because it shows that even though we do not aim for higher convergence rates (by increasing the polynomial degree in the trial space or by postprocessing) we get highly accurate approximations. We stress that this result has been observed in various numerical experiments, particularly also for more complex model problems like Stokes [17], but up to now a rigorous proof has not been given.

If $\boldsymbol{\beta} = 0$, we show that these results hold true when using different test norms (one of them is the so-called quasi-optimal test norm or graph norm). Surprisingly (at this point), for $\boldsymbol{\beta} \neq 0$ the results are only valid if the quasi-optimal test norm is used, although all test norms under consideration are equivalent. This is also observed in our numerical studies.

1.2 Basic Ideas

For the proofs of the main results, we develop duality arguments and show approximation results (Lemma 8 and Lemma 10). To get the essential idea, consider the abstract formulation: Find $\boldsymbol{u} \in U$ such that

$$b(\boldsymbol{u}, \boldsymbol{v}) = F(\boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in V,$$

where U denotes the trial space and V the test space. With the trial-to-test operator $\Theta : U \to V$,

$$(\Theta \boldsymbol{w}, \boldsymbol{v})_V = b(\boldsymbol{w}, \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in V,$$

the ideal DPG method reads: Find $\boldsymbol{u}_h \in U_h \subset U$ such that

$$b(\boldsymbol{u}_h, \Theta \boldsymbol{w}_h) = F(\Theta \boldsymbol{w}_h) \quad \text{for all } \boldsymbol{w}_h \in U_h.$$

Then we solve a dual problem: For some given $g \in L^2(\Omega)$, we determine $\boldsymbol{v} \in V$ and $\boldsymbol{w} = \Theta^{-1}\boldsymbol{v} \in U$, both unique, and employ Galerkin orthogonality to obtain

$$(u - u_h, g) = b(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{v}) = b(\boldsymbol{u} - \boldsymbol{u}_h, \Theta \boldsymbol{w}) = b(\boldsymbol{u} - \boldsymbol{u}_h, \Theta(\boldsymbol{w} - \boldsymbol{w}_h)) \lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U \|\boldsymbol{w} - \boldsymbol{w}_h\|_U$$

for arbitrary $\boldsymbol{w}_h \in U_h$.

For the case where we want to show that the approximation $u_h \in \mathcal{P}^p(\mathcal{T})$ is nearly the $L^2(\Omega)$ best approximation, we have $g = \Pi^p(u - u_h) = \Pi^p u - u_h$. Therefore,

$$\|g\|^2 = (u - u_h, g) \lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U \|\boldsymbol{w} - \boldsymbol{w}_h\|_U \lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U h \|g\|.$$

The latter estimate is what we have to show. Suppose that it holds. With the estimate for $\|\boldsymbol{u} - \boldsymbol{u}_h\|_U$ from above, it is straightforward to see that

$$\|u - u_h\| \leq \|u - \Pi^p u\| + \|\Pi^p u - u_h\| = \|u - \Pi^p u\| + \|g\| = \|u - \Pi^p u\| + \mathcal{O}(h^{p+2}).$$

Let us come back to the essential estimate

$$\|\boldsymbol{w} - \boldsymbol{w}_h\|_U \lesssim h \|g\|.$$

It holds if we would know that the higher derivatives of $\boldsymbol{w}$ exist (in some sense) and can be bounded by the norm of g, so that, formally,

$$\|\boldsymbol{w} - \boldsymbol{w}_h\|_U \lesssim h \|D^{\text{higher}} \boldsymbol{w}\| \lesssim h \|g\|$$

by some standard arguments. In our case we have that $\boldsymbol{v} \in H_0^1(\Omega) \times \boldsymbol{H}(\text{div}; \Omega) \subset V$ is the solution to the adjoint problem of (1.1) and under some assumptions has the higher regularity $\boldsymbol{v} \in H^2(\Omega) \times \boldsymbol{H}^1(\mathcal{T}) \cap \boldsymbol{H}(\text{div}; \Omega)$, where $\boldsymbol{H}^1(\mathcal{T})$ denotes $\mathcal{T}$-piecewise Sobolev functions. Recall that $\boldsymbol{w} = \Theta^{-1}\boldsymbol{v}$. One difficulty is that the inverse of the trial-to-test operator does not map regular functions back to regular functions. However, it turns out (Lemma 8) that $\boldsymbol{w}$ can be written as

$$\boldsymbol{w} = (g, 0, 0, 0) + \widetilde{\boldsymbol{w}} + \boldsymbol{w}^\star,$$

where components of $\widetilde{\boldsymbol{w}} \in U$ are related to the dual solution $\boldsymbol{v}$, which is sufficiently regular and $\boldsymbol{w}^\star$ is the solution of the (primal) problem (1.1) with data f and $\boldsymbol{f}$ depending on the dual solution $\boldsymbol{v}$ so that $\boldsymbol{w}^\star$ has sufficient regularity as well. Let us point out that this idea used in the proofs is new and allows to treat different test norms. In [10], which deals with a simple reaction-diffusion problem and one specific test norm only, the representation of $\boldsymbol{w}$ is obtained by integration by parts using the dual solution $\boldsymbol{v}$ and it is not clear if

that approach can be generalized to the present setting. Here, in the general case we have to consider the regularity of the dual solution $\boldsymbol{v}$ and the regularity of the solution $\boldsymbol{w}^\star$ of the primal problem. For the proofs it is also necessary that g is a function in the finite element space, so that we can choose $\boldsymbol{w}_h = (g, 0, 0, 0) + \overline{\boldsymbol{w}}_h$, where $\overline{\boldsymbol{w}}_h$ is the best approximation of $\widetilde{\boldsymbol{w}} + \boldsymbol{w}^\star$. Then we show that the above estimates hold true.

Let us note that Θ is defined through the inner product in the test space. Thus, the representation of $\boldsymbol{w} = \Theta^{-1}\boldsymbol{v}$ from above strongly depends on the choice of the test norm and has to be analyzed for each norm individually (this is done in Lemma 8). Earlier Keith, Vaziri Astaneh and Demkowicz [15] considered the optimal test norm. In our notation this would yield $\boldsymbol{w} = (g, 0, 0, 0)$, i.e., $\widetilde{\boldsymbol{w}} = 0 = \boldsymbol{w}^\star$.

Moreover, the ideas so far dealt with the ideal DPG method. In this paper we work out all results for the *practical DPG method* under standard assumptions, i.e., the existence of Fortin operators. This implies that we have to deal with additional discretization errors.

Finally, we note that higher convergence rates for the dual variable $\boldsymbol{\sigma}$ can not be obtained with the same arguments since the components of $\boldsymbol{v} = (v, \boldsymbol{\tau})$ with

$$(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \boldsymbol{g}) = b(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{v})$$

are less regular than in the case described above.

1.3 Outline

The remainder of the paper is organized as follows: Section 2 introduces basic notations, states the assumptions, and presents the main results (Theorem 3–5). The proofs of these theorems are postponed to Section 3, which also includes an a priori convergence estimate (Theorem 6) and the important auxiliary results Lemma 8, 10. In Section 4 we present two numerical experiments. The final Section 5 concludes this work with some remarks.

2 Main Results

2.1 Notation

We make use of the notation $\lesssim$, i.e., $A \lesssim B$ means that there exists a constant $C > 0$, which is independent of relevant quantities, such that $A \le CB$. Moreover, $A \simeq B$ means that both directions hold, i.e., $A \lesssim B$ and $B \lesssim A$.

2.2 Mesh

Let $\mathcal{T}$ denote a regular mesh of Ω consisting of simplices T and let $\mathcal{S} := \{\partial T : T \in \mathcal{T}\}$ denote the skeleton. We suppose that $\mathcal{T}$ is shape-regular, i.e., there exists a constant $\kappa_{\mathcal{T}} > 0$ such that

$$\max_{T\in\mathcal{T}} \frac{\operatorname{diam}(T)^d}{|T|} \le \kappa_{\mathcal{T}},$$

where $|T|$ denotes the volume measure of $T \in \mathcal{T}$. As usual, $h := h_{\mathcal{T}} := \max_{T\in\mathcal{T}} \operatorname{diam}(T)$ denotes the mesh-size.

2.3 Ultra-Weak Formulation

Before we derive the ultra-weak formulation of (1.1) in this subsection, we introduce some notation. Let $T \in \mathcal{T}$. We denote by $(\cdot,\cdot)_T$ the $L^2(T)$ scalar product and with $\|\cdot\|_T$ the induced norm. On boundaries ∂T, the $L^2(\partial T)$ scalar product is denoted by $\langle\cdot,\cdot\rangle_{\partial T}$ and extended to the duality between the spaces $H^{1/2}(\partial T)$ and $H^{-1/2}(\partial T)$.

Furthermore, we define the piecewise trace operators

$$\gamma_{0,\mathcal{S}} : H^1(\Omega) \to \prod_{T\in\mathcal{T}} H^{1/2}(\partial T), \qquad (\gamma_{0,\mathcal{S}} v)|_{\partial T} = v|_{\partial T},$$
$$\gamma_{\boldsymbol{n},\mathcal{S}} : \boldsymbol{H}(\operatorname{div};\Omega) \to \prod_{T\in\mathcal{T}} H^{-1/2}(\partial T), \quad (\gamma_{\boldsymbol{n},\mathcal{S}} \boldsymbol{\tau})|_{\partial T} = \boldsymbol{\tau}\cdot\boldsymbol{n}_T|_{\partial T},$$

where $\boldsymbol{n}_T$ denotes the normal on ∂T pointing from T to its complement. With these operators we define the trace spaces

$$H_0^{1/2}(\mathcal{S}) := \gamma_{0,\mathcal{S}}(H_0^1(\Omega)),$$
$$H^{-1/2}(\mathcal{S}) := \gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{H}(\operatorname{div};\Omega)).$$

These Hilbert spaces are equipped with minimum energy extension norms

$$\|\widehat{u}\|_{1/2,\mathcal{S}} := \inf\{\|u\|_{H^1(\Omega)} : \gamma_{0,\mathcal{S}} u = \widehat{u}\},$$
$$\|\widehat{\sigma}\|_{-1/2,\mathcal{S}} := \inf\{\|\boldsymbol{\sigma}\|_{\boldsymbol{H}(\operatorname{div};\Omega)} : \gamma_{\boldsymbol{n},\mathcal{S}} \boldsymbol{\sigma} = \widehat{\sigma}\}.$$

We use the broken test spaces

$$H^1(\mathcal{T}) := \{v \in L^2(\Omega) : v|_T \in H^1(T) \text{ for all } T \in \mathcal{T}\},$$
$$\boldsymbol{H}(\operatorname{div};\mathcal{T}) := \{\boldsymbol{\tau} \in \boldsymbol{L}^2(\Omega) : \boldsymbol{\tau}|_T \in \boldsymbol{H}(\operatorname{div};T) \text{ for all } T \in \mathcal{T}\}$$

and define the piecewise differential operators $\nabla_{\mathcal{T}} : H^1(\mathcal{T}) \to \boldsymbol{L}^2(\Omega)$, $\operatorname{div}_{\mathcal{T}} : \boldsymbol{H}(\operatorname{div};\mathcal{T}) \to L^2(\Omega)$ on each $T \in \mathcal{T}$ by

$$\nabla_{\mathcal{T}} v|_T := \nabla(v|_T),$$
$$\operatorname{div}_{\mathcal{T}} \boldsymbol{\tau}|_T := \operatorname{div}(\boldsymbol{\tau}|_T).$$

Moreover, we define the following dualities for all $\widehat{u} \in H_0^{1/2}(\mathcal{S})$, $\boldsymbol{\tau} \in \boldsymbol{H}(\operatorname{div};\mathcal{T})$, $\widehat{\sigma} \in H^{-1/2}(\mathcal{S})$, $v \in H^1(\mathcal{T})$:

$$\langle \widehat{u}, \boldsymbol{\tau}\cdot\boldsymbol{n}\rangle_{\mathcal{S}} := \sum_{T\in\mathcal{T}} \langle \widehat{u}|_{\partial T}, \boldsymbol{\tau}\cdot\boldsymbol{n}_T|_{\partial T}\rangle_{\partial T},$$
$$\langle \widehat{\sigma}, v\rangle_{\mathcal{S}} := \sum_{T\in\mathcal{T}} \langle \widehat{\sigma}|_{\partial T}, v|_{\partial T}\rangle_{\partial T}.$$

These dualities measure the jumps of $\boldsymbol{v} = (v,\boldsymbol{\tau}) \in H^1(\mathcal{T}) \times \boldsymbol{H}(\operatorname{div};\mathcal{T})$, i.e.,

$$v \in H_0^1(\Omega) \iff \langle \widehat{\sigma}, v\rangle_{\mathcal{S}} = 0 \qquad \text{for all } \widehat{\sigma} \in H^{-1/2}(\mathcal{S}), \tag{2.1a}$$
$$\boldsymbol{\tau} \in \boldsymbol{H}(\operatorname{div};\Omega) \iff \langle \widehat{u}, \boldsymbol{\tau}\cdot\boldsymbol{n}\rangle_{\mathcal{S}} = 0 \quad \text{for all } \widehat{u} \in H_0^{1/2}(\mathcal{S}), \tag{2.1b}$$

see, e.g., [3, Theorem 2.3].

The ultra-weak formulation is then derived from (1.1) by testing (1.1a) with $\boldsymbol{\tau} \in \boldsymbol{H}(\operatorname{div};\mathcal{T})$, (1.1b) with $v \in H^1(\mathcal{T})$, and piecewise integration by parts, i.e.,

$$-(u, \operatorname{div}_{\mathcal{T}} \boldsymbol{\tau}) + \langle \gamma_{0,\mathcal{S}} u, \boldsymbol{\tau}\cdot\boldsymbol{n}\rangle_{\mathcal{S}} - (\boldsymbol{\beta} u, \boldsymbol{\tau}) + (\boldsymbol{C}\boldsymbol{\sigma}, \boldsymbol{\tau}) = (\boldsymbol{C}\boldsymbol{f}, \boldsymbol{\tau}),$$
$$-(\boldsymbol{\sigma}, \nabla_{\mathcal{T}} v) + \langle \gamma_{\boldsymbol{n},\mathcal{S}}\boldsymbol{\sigma}, v\rangle_{\mathcal{S}} + (\gamma u, v) = (f, v).$$

Here, $(\cdot,\cdot) := (\cdot,\cdot)_\Omega$ is the $L^2(\Omega)$ scalar product with norm $\|\cdot\|$. Set

$$U := L^2(\Omega) \times \boldsymbol{L}^2(\Omega) \times H_0^{1/2}(\mathcal{S}) \times H^{-1/2}(\mathcal{S}),$$
$$V := H^1(\mathcal{T}) \times \boldsymbol{H}(\operatorname{div};\mathcal{T})$$

and define $F : V \to \mathbb{R}$ and $b : U \times V \to \mathbb{R}$ by

$$F(\boldsymbol{v}) := (f, v) + (\boldsymbol{f}, \boldsymbol{C}\boldsymbol{\tau}),$$
$$b(\boldsymbol{u}, \boldsymbol{v}) := (u, -\operatorname{div}_{\mathcal{T}} \boldsymbol{\tau} - \boldsymbol{\beta}\cdot\boldsymbol{\tau} + \gamma v) + (\boldsymbol{\sigma}, \boldsymbol{C}\boldsymbol{\tau} - \nabla_{\mathcal{T}} v) + \langle \widehat{u}, \boldsymbol{\tau}\cdot\boldsymbol{n}\rangle_{\mathcal{S}} + \langle \widehat{\sigma}, v\rangle_{\mathcal{S}}$$

for all $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \widehat{u}, \widehat{\sigma}) \in U$, $\boldsymbol{v} = (v, \boldsymbol{\tau}) \in V$. The ultra-weak formulation then reads: Find $\boldsymbol{u} \in U$ such that

$$b(\boldsymbol{u}, \boldsymbol{v}) = F(\boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in V. \tag{2.2}$$

2.4 DPG Method and Approximation

In U we use the canonical norm,

$$\|\boldsymbol{u}\|_U^2 := \|u\|^2 + \|\boldsymbol{\sigma}\|^2 + \|\widehat{u}\|_{1/2,\mathcal{S}}^2 + \|\widehat{\sigma}\|_{-1/2,\mathcal{S}}^2 \quad \text{for } \boldsymbol{u} = (u, \boldsymbol{\sigma}, \widehat{u}, \widehat{\sigma}) \in U.$$

For the test space V we define the three different norms

$$\|\boldsymbol{v}\|_{\mathrm{V,qopt}}^2 := \|-\mathrm{div}_{\mathcal{T}}\,\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\boldsymbol{\tau} + \gamma v\|^2 + \|\boldsymbol{C}^{1/2}\boldsymbol{\tau} - \boldsymbol{C}^{-1/2}\nabla_{\mathcal{T}} v\|^2 + \|\boldsymbol{C}^{1/2}\boldsymbol{\tau}\|^2 + \|v\|^2, \tag{2.3a}$$

$$\|\boldsymbol{v}\|_{\mathrm{V},1}^2 := \|\boldsymbol{C}^{-1/2}\nabla_{\mathcal{T}} v\|^2 + \|v\|^2 + \|\mathrm{div}_{\mathcal{T}}\,\boldsymbol{\tau}\|^2 + \|\boldsymbol{C}^{1/2}\boldsymbol{\tau}\|^2, \tag{2.3b}$$

$$\|\boldsymbol{v}\|_{\mathrm{V},2}^2 := \|\nabla_{\mathcal{T}} v\|^2 + \|v\|^2 + \|\mathrm{div}_{\mathcal{T}}\,\boldsymbol{\tau}\|^2 + \|\boldsymbol{\tau}\|^2 \tag{2.3c}$$

for $\boldsymbol{v} = (v, \boldsymbol{\tau}) \in V$ and denote by $(\cdot,\cdot)_{V,\star}$ the corresponding scalar products. Note that all norms in (2.3) are equivalent with equivalence constants depending on the coefficients $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ. However, our main results hold for the quasi-optimal test norm $\|\cdot\|_{\mathrm{V,qopt}}$ under mild assumptions on the coefficient $\boldsymbol{\beta}$, whereas they hold for $\|\cdot\|_{\mathrm{V},1}$, $\|\cdot\|_{\mathrm{V},2}$ only if $\boldsymbol{\beta} = 0$, i.e., for symmetric problems.

We stress that $b : U \times V \to \mathbb{R}$ is a bounded bilinear form and satisfies the inf–sup conditions with mesh independent constant. This can be proved with the theory developed in [3]. For our model problem we explicitly refer to [3, Example 3.7] for the details. There it is assumed that $\mathrm{div}(\boldsymbol{C}^{-1}\boldsymbol{\beta}) = 0$ and $\gamma \geq 0$. We note that their analysis can also be done with our more general assumption (1.2).

The DPG method seeks an approximation $\boldsymbol{u}_h \in U_h \subset U$ of the solution $\boldsymbol{u} \in U$ using the optimal test space $\Theta(U_h)$, where $\Theta : U \to V$ is defined by

$$(\Theta\boldsymbol{w}, \boldsymbol{v})_V = b(\boldsymbol{w}, \boldsymbol{v}) \quad \text{for all } \boldsymbol{w} \in U,\ \boldsymbol{v} \in V. \tag{2.4}$$

Then $\boldsymbol{u}_h \in U_h$ is the solution of

$$b(\boldsymbol{u}_h, \boldsymbol{v}_h) = F(\boldsymbol{v}_h) \quad \text{for all } \boldsymbol{v}_h \in \Theta(U_h).$$

An essential feature of DPG is that inf–sup stability directly transfers to the discrete problem. However, in practice we replace Θ by a discrete version $\Theta_h : U_h \to V_h \subset V$ defined by

$$(\Theta_h\boldsymbol{w}_h, \boldsymbol{v}_h)_V = b(\boldsymbol{w}_h, \boldsymbol{v}_h) \quad \text{for all } \boldsymbol{w}_h \in U_h,\ \boldsymbol{v}_h \in V_h.$$

Then the *practical DPG method* reads: Find $\boldsymbol{u}_h \in U_h$ such that

$$b(\boldsymbol{u}_h, \Theta_h\boldsymbol{w}_h) = F(\Theta_h\boldsymbol{w}_h) \quad \text{for all } \boldsymbol{w}_h \in U_h. \tag{2.5}$$

In this work we deal with the piecewise polynomial trial spaces

$$U_{hp} := \mathcal{P}^p(\mathcal{T}) \times \mathcal{P}^p(\mathcal{T})^d \times \mathcal{P}^{p+1}_{c,0}(\mathcal{S}) \times \mathcal{P}^p(\mathcal{S}),$$
$$U_{hp}^+ := \mathcal{P}^{p+1}(\mathcal{T}) \times \mathcal{P}^p(\mathcal{T})^d \times \mathcal{P}^{p+1}_{c,0}(\mathcal{S}) \times \mathcal{P}^p(\mathcal{S})$$

and the piecewise polynomial test spaces

$$V_{hk} := \mathcal{P}^{k_1}(\mathcal{T}) \times \mathcal{P}^{k_2}(\mathcal{T})^d.$$

Here, we set

$$\mathcal{P}^p(T) := \{v \in L^2(T) : v \text{ is polynomial of degree} \leq p\},$$
$$\mathcal{P}^p(\mathcal{T}) := \{v \in L^2(\Omega) : v|_T \in \mathcal{P}^p(T),\ T \in \mathcal{T}\}, \qquad \mathcal{P}^{p+1}_{c,0}(\mathcal{T}) := \mathcal{P}^{p+1}(\mathcal{T}) \cap H^1_0(\Omega),$$
$$\mathcal{P}^{p+1}_{c,0}(\mathcal{S}) := \gamma_{0,\mathcal{S}}(\mathcal{P}^{p+1}_{c,0}(\mathcal{T})), \qquad \mathcal{P}^p(\mathcal{S}) := \gamma_{\boldsymbol{n},\mathcal{S}}(\mathcal{RT}^p(\mathcal{T})),$$

where

$$\mathcal{RT}^p(\mathcal{T}) = \{\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div};\Omega) : \boldsymbol{\tau}|_T(\boldsymbol{x}) = \boldsymbol{a} + \boldsymbol{x}b,\ \boldsymbol{a} \in \mathcal{P}^p(T)^d,\ b \in \widetilde{\mathcal{P}}^p(T),\ T \in \mathcal{T}\}$$

is the space of Raviart–Thomas functions (here $\widetilde{\mathcal{P}}^p(T)$ denotes the space of homogeneous polynomials of degree p).

We also use the space $C^1(\mathcal{T}) := \{v \in L^\infty(\Omega) : v|_T \in C^1(\overline{T}),\ T \in \mathcal{T}\}$.

2.5 Fortin Operators

It is well known, see, e.g., [12], that (2.5) satisfies inf–sup conditions (and therefore admits a unique solution) if there exists a Fortin operator $\Pi_F : V \to V_h$ such that

$$\|\Pi_F \boldsymbol{v}\|_V \leq C_F \|\boldsymbol{v}\|_V \quad \text{and} \quad b(\boldsymbol{u}_h, \boldsymbol{v}) = b(\boldsymbol{u}_h, \Pi_F \boldsymbol{v}) \quad \text{for all } \boldsymbol{v} \in V,\ \boldsymbol{u}_h \in U_h. \tag{2.6}$$

Throughout, we suppose that a Fortin operator exists for the discrete polynomial trial and test spaces under consideration and that C_F depends only on $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ, $p \in \mathbb{N}_0$, and the shape-regularity of $\mathcal{T}$. Let us note that for general coefficients $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ the existence of such operators is not known, except for some special cases, i.e., the Poisson model problem where $\boldsymbol{C}$ is the identity and $\boldsymbol{\beta} = 0 = \gamma$. Fortin operators for the latter problem on simplicial meshes have been constructed and analyzed in [3, 12]. We refer also to [16] for the construction and analysis of Fortin operators for second-order problems.

Supposing the existence of an Fortin operator, i.e., (2.6), we have:

Proposition 1. *Problems* (2.2), (2.5) *admit unique solutions* $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \widehat{u}, \widehat{\sigma}) \in U$, $\boldsymbol{u}_h \in U_h$ *and*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \leq C_{\mathrm{opt}} \min_{\boldsymbol{w}_h \in U_h} \|\boldsymbol{u} - \boldsymbol{w}_h\|_U.$$

The constant $C_{\mathrm{opt}} > 0$ *depends only on* Ω, $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ, $p \in \mathbb{N}_0$, *and the shape-regularity of* $\mathcal{T}$.

2.6 Adjoint Problem and Regularity Assumptions

We define the adjoint problem (in the sense of $L^2(\Omega)$-adjoints) of (1.1) as

$$-\operatorname{div} \boldsymbol{\tau} - \boldsymbol{\beta} \cdot \boldsymbol{\tau} + \gamma v = g \qquad \text{in } \Omega, \tag{2.7a}$$

$$\boldsymbol{C}\boldsymbol{\tau} - \nabla v = \boldsymbol{C}\boldsymbol{g} \quad \text{in } \Omega, \tag{2.7b}$$

$$u = 0 \qquad \text{on } \Gamma. \tag{2.7c}$$

Supposing (1.2) this problem admits a unique solution $(v, \boldsymbol{\tau}) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div}; \Omega)$ for $g \in L^2(\Omega)$, $\boldsymbol{g} \in \boldsymbol{L}^2(\Omega)$.

For our results we make use of the following assumptions:

Assumption. We suppose that the coefficients $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ and the domain Ω are such that for $f, g \in L^2(\Omega)$, $\boldsymbol{f}, \boldsymbol{g} \in \boldsymbol{H}^1(\mathcal{T}) \cap \boldsymbol{H}(\operatorname{div}; \Omega)$ the unique solutions $(u, \boldsymbol{\sigma}) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div}; \Omega)$ resp. $(v, \boldsymbol{\tau}) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div}; \Omega)$ of (1.1) resp. (2.7) satisfy

$$\|u\|_{H^2(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^1(\mathcal{T})} \leq C(\|f\| + \|\boldsymbol{f}\|_{\boldsymbol{H}^1(\mathcal{T})}), \tag{2.8a}$$

$$\|v\|_{H^2(\Omega)} + \|\boldsymbol{\tau}\|_{\boldsymbol{H}^1(\mathcal{T})} \leq C(\|g\| + \|\boldsymbol{g}\|_{\boldsymbol{H}^1(\mathcal{T})}). \tag{2.8b}$$

Here, $\|\cdot\|_{H^s(\Omega)}$ is the usual notation for norms in the Sobolev space $H^s(\Omega)$ ($s > 0$), and $\|\cdot\|$ is the $L^2(\Omega)$ norm and $\|\cdot\|_{\boldsymbol{H}^s(\mathcal{T})}$ the broken Sobolev norm for vector valued functions. We note that the constant $C > 0$ strongly depends on the material data, i.e., the coefficients $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ. For example, a small constant diffusion implies a blow-up of C.

Remark 2. The regularity estimates (2.8) are satisfied if $d = 2$, $\boldsymbol{C}$ is the identity matrix, $\boldsymbol{\beta} \in C^1(\mathcal{T})^d \cap \boldsymbol{H}(\operatorname{div}; \Omega)$ and Ω is convex. This can be seen as follows: The first component $u \in H^1_0(\Omega)$ of the solution of (1.1) satisfies

$$-\Delta u = f - \operatorname{div} \boldsymbol{f} - (\operatorname{div} \boldsymbol{\beta})u + \boldsymbol{\beta} \cdot \nabla u - \gamma u \in L^2(\Omega).$$

Then $u \in H^2(\Omega)$ and $\|u\|_{H^2(\Omega)}$ is bounded by the $L^2(\Omega)$ norm of the right-hand side, since Ω is a convex polyhedral domain, see [13]. Finally, the second equation of the model problem (1.1) shows

$$\|\boldsymbol{\sigma}\|_{\boldsymbol{H}^1(\mathcal{T})} = \|\boldsymbol{f} - \nabla u + \boldsymbol{\beta} u\|_{\boldsymbol{H}^1(\mathcal{T})} \lesssim \|\boldsymbol{f}\|_{\boldsymbol{H}^1(\mathcal{T})} + \|u\|_{H^2(\Omega)} \lesssim \|f\| + \|\boldsymbol{f}\|_{\boldsymbol{H}^1(\mathcal{T})}.$$

Similarly, one shows (2.8b) (even a less regular coefficient $\boldsymbol{\beta}$ suffices for the adjoint problem).

2.7 Assumptions on Coefficients and Test Norms

Besides the assumptions on the coefficients and the domain to ensure unique solvability of problems (1.1) and (2.7) and estimates (2.8), we also need some additional assumptions on the coefficients that are listed in Table 1. We emphasize that $\boldsymbol{\beta} = 0$ in Cases (b) and (c) is also necessary in general. In particular, in Section 4 we provide a simple example where $\boldsymbol{\beta} \neq 0$ and the choice $\|\boldsymbol{v}\|_V = \|\boldsymbol{v}\|_{\mathrm{V},1}$ or $\|\boldsymbol{v}\|_V = \|\boldsymbol{v}\|_{\mathrm{V},2}$ does not lead to higher convergence rates, whereas $\|\boldsymbol{v}\|_V = \|\boldsymbol{v}\|_{\mathrm{V},\mathrm{qopt}}$ does.

Case	Test norm $\|\cdot\|_V$	C	β	γ
(a)	$\|\cdot\|_{\mathrm{V},\mathrm{qopt}}$	$C^1(\mathcal{T})^{d\times d}$	$C^1(\mathcal{T})^d$	$C^1(\mathcal{T})$
(b)	$\|\cdot\|_{\mathrm{V},1}$	$C^1(\mathcal{T})^{d\times d}$	0	$C^1(\mathcal{T})$
(c)	$\|\cdot\|_{\mathrm{V},2}$	$C^{0,1}(\overline{\Omega})^{d\times d} \cap C^1(\mathcal{T})^{d\times d}$	0	$C^1(\mathcal{T})$

Table 1: Additional assumptions (besides (1.2) and (2.8)) on the coefficients for the three test norms under consideration.

2.8 $L^2(\Omega)$ Projection

Our first main result shows that the DPG method with ultra-weak formulation delivers up to a higher-order term the $L^2(\Omega)$ best approximation for the scalar field variable. To that end let $\Pi^p : L^2(\Omega) \to \mathcal{P}^p(\mathcal{T})$ denote the $L^2(\Omega)$ projector.

Theorem 3. *Consider one of Cases* (a), (b), *or* (c). *Let* $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \widehat{u}, \widehat{\sigma}) \in U$ *be the solution of* (2.2) *for some given* $f \in L^2(\Omega)$, $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ *and suppose* $u \in H^{p+2}(\Omega)$, $\boldsymbol{\sigma} \in \boldsymbol{H}^{p+1}(\mathcal{T})$. *Let* $\boldsymbol{u}_h = (u_h, \boldsymbol{\sigma}_h, \widehat{u}_h, \widehat{\sigma}_h) \in U_h := U_{hp}$ *be the solution of the practical DPG method* (2.5). *Suppose* $\mathcal{P}^1_{c,0}(\mathcal{T}) \times \mathcal{RT}^p(\mathcal{T}) \subseteq V_{hk}$. *It holds that*

$$\|u - \Pi^p u\| \le \|u - u_h\| \le \|u - \Pi^p u\| + Ch^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}).$$

The constant $C > 0$ *depends only on* Ω, $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ, $p \in \mathbb{N}_0$, *and shape-regularity of* $\mathcal{T}$.

2.9 Higher Convergence Rate by Increasing Polynomial Degree

Our second main result shows that higher convergence rates for the scalar field variable are obtained by increasing the polynomial degree in the approximation space.

Theorem 4. *Consider one of Cases* (a), (b), *or* (c). *Let* $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \widehat{u}, \widehat{\sigma}) \in U$ *be the solution of* (2.2) *for some given* $f \in L^2(\Omega)$, $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ *and suppose* $u \in H^{p+2}(\Omega)$, $\boldsymbol{\sigma} \in \boldsymbol{H}^{p+1}(\mathcal{T})$. *Let* $\boldsymbol{u}_h^+ = (u_h^+, \boldsymbol{\sigma}_h, \widehat{u}_h, \widehat{\sigma}_h) \in U_h := U_{hp}^+$ *be the solution of the practical DPG method* (2.5). *Suppose* $\mathcal{P}^1_{c,0}(\mathcal{T}) \times \mathcal{RT}^{p+1}(\mathcal{T}) \subseteq V_{hk}$. *It holds that*

$$\|u - u_h^+\| \le Ch^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}).$$

The constant $C > 0$ *depends only on* Ω, $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ, $p \in \mathbb{N}_0$, *and shape-regularity of* $\mathcal{T}$.

2.10 Higher Convergence Rate by Postprocessing

Our third and final main result shows that higher convergence rates for the scalar field variable are obtained by postprocessing the solution: Let $\boldsymbol{u}_h = (u_h, \boldsymbol{\sigma}_h, \widehat{u}_h, \widehat{\sigma}_h) \in U_h := U_{hp}$ be the solution of (2.5). We define $\widetilde{u}_h \in \mathcal{P}^{p+1}(\mathcal{T})$ on each element $T \in \mathcal{T}$ as the solution of the local Neumann problem

$$(\nabla\widetilde{u}_h, \nabla v_h)_T = (\boldsymbol{Cf} - \boldsymbol{C\sigma}_h + \boldsymbol{\beta}u_h, \nabla v_h)_T \quad \text{for all } v_h \in \mathcal{P}^{p+1}(T), \tag{2.9a}$$

$$(\widetilde{u}_h, 1)_T = (u_h, 1)_T. \tag{2.9b}$$

Let us note that this type of postprocessing is common in literature and can already be found in the early works [11, 20].

Theorem 5. *Consider one of Cases* (a), (b), *or* (c). *Let* $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \widehat{u}, \widehat{\sigma}) \in U$ *be the solution of* (2.2) *for some given* $f \in L^2(\Omega)$, $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ *and suppose* $u \in H^{p+2}(\Omega)$, $\boldsymbol{\sigma} \in \boldsymbol{H}^{p+1}(\mathcal{T})$. *Let* $\boldsymbol{u}_h = (u_h, \boldsymbol{\sigma}_h, \widehat{u}_h, \widehat{\sigma}_h) \in U_h := U_{hp}$ *be the solution of the practical DPG method* (2.5) *and define* $\widetilde{u}_h \in \mathcal{P}^{p+1}(\mathcal{T})$ *by* (2.9). *Suppose* $\mathcal{P}^1_{c,0}(\mathcal{T}) \times \mathcal{RT}^p(\mathcal{T}) \subseteq V_{hk}$. *It holds that*

$$\|u - \widetilde{u}_h\| \le Ch^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}).$$

The constant $C > 0$ *depends only on* Ω, $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ, $p \in \mathbb{N}_0$, *and shape-regularity of* $\mathcal{T}$.

3 Proofs

In this section we prove the results stated in Theorems 3, 4, and 5. First, in Section 3.1 we collect some standard results on projection operators and consider approximation results with respect to $\|\cdot\|_U$. Second, Section 3.2 recalls the equivalent mixed formulation of the practical DPG method. Then Section 3.3 provides auxiliary results that allow to prove the main results in a uniform fashion. Finally, in Sections 3.4, 3.5, 3.6 we give the proofs of our main results.

3.1 Projection Operators and Approximation Results

Throughout let $p \in \mathbb{N}_0$. Let $\Pi^p : L^2(\Omega) \to \mathcal{P}^p(\mathcal{T})$ denote the $L^2(\Omega)$ projector. For $\boldsymbol{\tau} \in \boldsymbol{L}^2(\Omega)$ the term $\Pi^p\boldsymbol{\tau}$ is understood as the application of Π^p to each component. We have the (local) approximation properties

$$\|u - \Pi^p u\| \le C_p h^{p+1}|u|_{H^{p+1}(\mathcal{T})} \quad \text{and} \quad \|\boldsymbol{\sigma} - \Pi^p\boldsymbol{\sigma}\| \le C_p h^{p+1}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{p+1}(\mathcal{T})}, \tag{3.1a}$$

where $|\cdot|_{H^n(\mathcal{T})} := \|D^n_{\mathcal{T}}\cdot\|$ with $D^n_{\mathcal{T}}$ denoting the $\mathcal{T}$-elementwise n-th derivative operator. Let

$$\Pi^{p+1}_\nabla : H^1_0(\Omega) \to \mathcal{P}^{p+1}_{c,0}(\mathcal{T})$$

denote the Scott–Zhang projection operator [19] or any other operator with the property

$$\|u - \Pi^{p+1}_\nabla u\|_{H^1(\Omega)} \le C_p h^{p+1}\|u\|_{H^{p+2}(\Omega)}. \tag{3.1b}$$

Moreover, let $\Pi^p_{\mathrm{div}} : \boldsymbol{H}(\mathrm{div};\Omega) \cap \boldsymbol{H}^1(\mathcal{T}) \to \mathcal{RT}^p(\mathcal{T})$ denote the Raviart–Thomas operator, which satisfies

$$\|\boldsymbol{\sigma} - \Pi^p_{\mathrm{div}}\boldsymbol{\sigma}\| \le C_p h^{k+1}|\boldsymbol{\sigma}|_{\boldsymbol{H}^{k+1}(\mathcal{T})} \quad \text{for } k \in [0,p], \tag{3.1c}$$

and the commutativity property

$$\mathrm{div}\,\Pi^p_{\mathrm{div}}\boldsymbol{\sigma} = \Pi^p\,\mathrm{div}\,\boldsymbol{\sigma}.$$

Note that Π^p_{div} is well defined for functions $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div};\Omega) \cap \boldsymbol{H}^1(\mathcal{T})$: First, normal traces of $\boldsymbol{\sigma} \in \boldsymbol{H}^1(\mathcal{T})$ are well defined on each facet of ∂T, $T \in \mathcal{T}$, in the sense of $L^2(\partial T)$, i.e., $\boldsymbol{\sigma}\cdot\boldsymbol{n}_T \in L^2(\partial T)$ and, second, $\boldsymbol{\sigma} \in \boldsymbol{H}(\mathrm{div};\Omega)$ implies unisolvency of normal traces. The constant $C_p > 0$ in (3.1) depends only on $p \in \mathbb{N}_0$ and shape-regularity of $\mathcal{T}$.

The following result is an adaptation of [10, Theorem 5 and Corollary 6].

Theorem 6. *Let* $p \in \mathbb{N}_0$ *and let* $w \in H^{p+2}(\Omega)$, $\boldsymbol{\chi} \in \boldsymbol{H}^{p+1}(\mathcal{T}) \cap \boldsymbol{H}(\mathrm{div};\Omega)$. *Define* $\boldsymbol{w} := (w, \boldsymbol{\chi}, \gamma_{0,\mathbb{S}}w, \gamma_{\boldsymbol{n},\mathbb{S}}\boldsymbol{\chi}) \in U$. *If* $U_h \in \{U_{hp}, U^+_{hp}\}$, *then*

$$\min_{\boldsymbol{w}_h \in U_h} \|\boldsymbol{w} - \boldsymbol{w}_h\|_U \le Ch^{p+1}(\|w\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\chi}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}).$$

The constant $C > 0$ *depends only on* p *and shape-regularity of* $\mathcal{T}$.

Proof. Define

$$\boldsymbol{w}_h := (\Pi^p w, \Pi^p\boldsymbol{\chi}, \gamma_{0,\mathbb{S}}\Pi^{p+1}_\nabla w, \gamma_{\boldsymbol{n},\mathbb{S}}\Pi^p_{\mathrm{div}}\boldsymbol{\chi}) \in U_h.$$

We estimate the terms in

$$\|\boldsymbol{w} - \boldsymbol{w}_h\|^2_U = \|w - \Pi^p w\|^2 + \|\boldsymbol{\chi} - \Pi^p\boldsymbol{\chi}\|^2 + \|\gamma_{0,\mathbb{S}}(w - \Pi^{p+1}_\nabla w)\|^2_{1/2,\mathbb{S}} + \|\gamma_{\boldsymbol{n},\mathbb{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\|^2_{-1/2,\mathbb{S}}.$$

First, we follow [10, Proof of Theorem 5] to estimate $\|\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\|_{-1/2,\mathcal{S}}$: To this end, we start with the identity from [3, Theorem 2.3], i.e.,

$$\|\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\|_{-1/2,\mathcal{S}} = \sup_{0\neq v\in H^1(\mathcal{T})} \frac{\langle \gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}), v\rangle_{\mathcal{S}}}{\|v\|_{H^1(\mathcal{T})}}.$$

Then elementwise integration by parts and the commutativity property yield

$$\begin{aligned}\langle \gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}), v\rangle_{\mathcal{S}} &= (\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}, \nabla_{\mathcal{T}} v) + (\mathrm{div}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}), v)\\ &= (\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}, \nabla_{\mathcal{T}} v) + ((1-\Pi^p)\,\mathrm{div}\,\boldsymbol{\chi}, v).\end{aligned}$$

Using the L^2 projection property and the approximation properties (3.1a) and (3.1c), we estimate the last two terms by

$$\begin{aligned}|(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}, \nabla_{\mathcal{T}} v)| + |((1-\Pi^p)\,\mathrm{div}\,\boldsymbol{\chi}, v)| &= |(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}, \nabla_{\mathcal{T}} v)| + |((1-\Pi^p)\,\mathrm{div}\,\boldsymbol{\chi}, (1-\Pi^p)v)|\\ &\lesssim h^{p+1}|\boldsymbol{\chi}|_{\boldsymbol{H}^{p+1}(\mathcal{T})}\|\nabla_{\mathcal{T}} v\| + h^p|\mathrm{div}\,\boldsymbol{\chi}|_{H^p(\mathcal{T})}h\|\nabla_{\mathcal{T}} v\|\\ &\lesssim h^{p+1}|\boldsymbol{\chi}|_{\boldsymbol{H}^{p+1}(\mathcal{T})}\|\nabla_{\mathcal{T}} v\|.\end{aligned}$$

Putting the last estimates together this shows that

$$\|\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\|_{-1/2,\mathcal{S}} \lesssim h^{p+1}\|\boldsymbol{\chi}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}.$$

Next, observe that $\|\gamma_{0,\mathcal{S}}(\cdot)\|_{1/2,\mathcal{S}} \leq \|\cdot\|_{H^1(\Omega)}$ by definition of the norms. Finally, applying the approximation properties (3.1a)–(3.1b) and putting altogether finishes the proof. □

Remark 7. As pointed out in [10] the estimate $\|\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\|_{-1/2,\mathcal{S}} \lesssim h^{p+1}\|\boldsymbol{\chi}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}$ in the proof of Theorem 6 is non-trivial. A direct application of the trace theorem gives

$$\|\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\|_{-1/2,\mathcal{S}} \leq \|\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}\| + \|\mathrm{div}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\|.$$

Using the commutativity property and the approximation properties (3.1a), (3.1c) we get

$$\|\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi}\| + \|\mathrm{div}(\boldsymbol{\chi} - \Pi^p_{\mathrm{div}}\boldsymbol{\chi})\| \lesssim h^{p+1}|\boldsymbol{\chi}|_{H^{p+1}(\mathcal{T})} + h^{p+1}|\mathrm{div}\,\boldsymbol{\chi}|_{H^{p+1}(\mathcal{T})}.$$

Thus, in order to get the same convergence rate as in Theorem 6 we have to assume the higher regularity $\mathrm{div}\,\boldsymbol{\chi} \in H^{p+1}(\mathcal{T})$.

3.2 Mixed Formulation of the Practical DPG Method

The practical DPG method (2.5) can be reformulated as a mixed problem, see, e.g., [1]. Recall that we made the assumption of the existence of a Fortin operator (2.6). The mixed DPG formulation then reads: Find $(\boldsymbol{u}_h, \boldsymbol{\varepsilon}_{hk}) \in U_h \times V_{hk}$ such that

$$(\boldsymbol{\varepsilon}_{hk}, \boldsymbol{v}_{hk})_V + b(\boldsymbol{u}_h, \boldsymbol{v}_{hk}) = F(\boldsymbol{v}_{hk}) \quad \text{for all } \boldsymbol{v}_{hk} \in V_{hk}, \tag{3.2a}$$

$$b(\boldsymbol{w}_h, \boldsymbol{\varepsilon}_{hk}) = 0 \quad \text{for all } \boldsymbol{w}_h \in U_h. \tag{3.2b}$$

The Riesz representation $\boldsymbol{\varepsilon}_{hk} \in V_{hk}$ of the residual (sometimes also called the error representation function) satisfies

$$\|\boldsymbol{\varepsilon}_{hk}\|_V \lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U,$$

under assumption (2.6), see [2, Theorem 2.1]. Note that the solution $\boldsymbol{u}_h$ in (3.2) is identical to the solution of (2.5). Recall that the residual on the continuous level vanishes and, therefore, the Riesz representation is zero. Setting $\boldsymbol{\varepsilon} := 0$, we have that $(\boldsymbol{u}, \boldsymbol{\varepsilon}) \in U \times V$ satisfies the mixed formulation for all test functions $(\boldsymbol{w}, \boldsymbol{v}) \in U \times V$. In particular, we have Galerkin orthogonality

$$a((\boldsymbol{u} - \boldsymbol{u}_h), (\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_{hk}), (\boldsymbol{w}_h, \boldsymbol{v}_{hk})) = 0 \quad \text{for all } (\boldsymbol{w}_h, \boldsymbol{v}_{hk}) \in U_h \times V_{hk},$$

where $a((\boldsymbol{w}, \boldsymbol{v}), (\delta\boldsymbol{w}, \delta\boldsymbol{v})) := b(\boldsymbol{w}, \delta\boldsymbol{v}) + (\boldsymbol{v}, \delta\boldsymbol{v})_V - b(\delta\boldsymbol{w}, \boldsymbol{v})$ for all $\boldsymbol{w}, \delta\boldsymbol{w} \in U$, $\boldsymbol{v}, \delta\boldsymbol{v} \in V$.

3.3 Auxiliary Results

Recall the adjoint problem (2.7) with $g \in L^2(\Omega)$, $\boldsymbol{g} = 0$,

$$\begin{aligned} -\operatorname{div}\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\boldsymbol{\tau} + \gamma v &= g, \\ \nabla v - \boldsymbol{C}\boldsymbol{\tau} &= 0, \\ v|_\Gamma &= 0. \end{aligned} \tag{3.3}$$

Note that $\boldsymbol{v} = (v, \boldsymbol{\tau}) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div};\Omega) \subset V$. In particular, there exists a unique $\boldsymbol{w} \in U$ with $\Theta\boldsymbol{w} = \boldsymbol{v}$, since $\Theta : U \to V$ is an isomorphism. Note that by the definition of the trial-to-test operator (2.4), the element $\boldsymbol{w}$ depends on the choice of scalar products in V. This is investigated in the following result.

Lemma 8. *Let $g \in L^2(\Omega)$ and let $\boldsymbol{v} := (v, \boldsymbol{\tau}) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div};\Omega)$ denote the solution of (3.3). The unique element $\boldsymbol{w} \in U$ with $\Theta\boldsymbol{w} = \boldsymbol{v}$ has the following representation depending on the cases from Section 2.7:*

- Case (a) ($\|\cdot\|_V = \|\cdot\|_{\mathrm{V,qopt}}$):

$$\boldsymbol{w} = (g, 0, 0, 0) + (u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathbb{S}} u^*, \gamma_{\boldsymbol{n},\mathbb{S}} \boldsymbol{\sigma}^*),$$

  where $(u^*, \boldsymbol{\sigma}^*) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div};\Omega)$ solves (1.1) with $f = v$ and $\boldsymbol{f} = \boldsymbol{\tau}$.
- Case (b) ($\|\cdot\|_V = \|\cdot\|_{\mathrm{V},1}$):

$$\boldsymbol{w} = (g - \gamma v, 0, 0, \gamma_{\boldsymbol{n},\mathbb{S}}\boldsymbol{\tau}) + (u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathbb{S}} u^*, \gamma_{\boldsymbol{n},\mathbb{S}} \boldsymbol{\sigma}^*),$$

  where $(u^*, \boldsymbol{\sigma}^*) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div};\Omega)$ solves (1.1) with $f = \gamma(\gamma v - g) - \operatorname{div}\boldsymbol{\tau} + v$ and $\boldsymbol{f} = \boldsymbol{\tau}$.
- Case (c) ($\|\cdot\|_V = \|\cdot\|_{\mathrm{V},2}$):

$$\boldsymbol{w} = (g - \gamma v, 0, 0, \gamma_{\boldsymbol{n},\mathbb{S}}(\boldsymbol{C}\boldsymbol{\tau})) + (u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathbb{S}} u^*, \gamma_{\boldsymbol{n},\mathbb{S}} \boldsymbol{\sigma}^*),$$

  where $(u^*, \boldsymbol{\sigma}^*) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div};\Omega)$ solves (1.1) with $f = \gamma(\gamma v - g) - \operatorname{div}(\boldsymbol{C}\boldsymbol{\tau}) + v$ and $\boldsymbol{f} = \boldsymbol{C}^{-1}\boldsymbol{\tau}$.

Moreover,

$$\|v\|_{H^2(\Omega)} + \|\boldsymbol{\tau}\|_{\boldsymbol{H}^1(\mathcal{T})} + \|u^*\|_{H^2(\Omega)} + \|\boldsymbol{\sigma}^*\|_{\boldsymbol{H}^1(\mathcal{T})} \le C\|g\|. \tag{3.4}$$

For Case (c) it also holds that $\boldsymbol{\tau}, \boldsymbol{\sigma}^* \in \boldsymbol{H}^1(\Omega)$.

Proof. We consider the three cases.

Case (a). Recall that

$$(\Theta\boldsymbol{w}, (\mu, \boldsymbol{\lambda}))_V = b(\boldsymbol{w}, (\mu, \boldsymbol{\lambda})) \quad \text{for all } (\mu, \boldsymbol{\lambda}) \in V.$$

With the inner product in V and $\operatorname{div}_{\mathcal{T}}\boldsymbol{\tau} = \operatorname{div}\boldsymbol{\tau}$, $\nabla_{\mathcal{T}} v = \nabla v$ we have for $(\mu, \boldsymbol{\lambda}) \in V$ that

$$\begin{aligned} (\boldsymbol{v}, (\mu, \boldsymbol{\lambda}))_V &= (-\operatorname{div}\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\boldsymbol{\tau} + \gamma v, -\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda} - \boldsymbol{\beta}\cdot\boldsymbol{\lambda} + \gamma\mu) \\ &\qquad + (\boldsymbol{C}^{1/2}\boldsymbol{\tau} - \boldsymbol{C}^{-1/2}\nabla v, \boldsymbol{C}^{1/2}\boldsymbol{\lambda} - \boldsymbol{C}^{-1/2}\nabla_{\mathcal{T}}\mu) + (\boldsymbol{C}\boldsymbol{\tau}, \boldsymbol{\lambda}) + (v, \mu) \\ &= (g, -\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda} - \boldsymbol{\beta}\cdot\boldsymbol{\lambda} + \gamma\mu) + (\boldsymbol{C}\boldsymbol{\tau}, \boldsymbol{\lambda}) + (v, \mu) \\ &= b((g, 0, 0, 0), (\mu, \boldsymbol{\lambda})) + (\boldsymbol{C}\boldsymbol{\tau}, \boldsymbol{\lambda}) + (v, \mu). \end{aligned}$$

Let $(u^*, \boldsymbol{\sigma}^*) \in H^1_0(\Omega) \times \boldsymbol{H}(\operatorname{div};\Omega)$ solve the (primal) problem (1.1) with $f = v \in L^2(\Omega)$ and $\boldsymbol{f} = \boldsymbol{\tau} \in \boldsymbol{H}(\operatorname{div};\Omega)$. In particular, $(u^*, \boldsymbol{\sigma}^*)$ solves the ultra-weak formulation (2.2), i.e.,

$$b((u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathbb{S}} u^*, \gamma_{\boldsymbol{n},\mathbb{S}} \boldsymbol{\sigma}^*), (\mu, \boldsymbol{\lambda})) = (\boldsymbol{C}\boldsymbol{\tau}, \boldsymbol{\lambda}) + (v, \mu) \quad \text{for all } (\mu, \boldsymbol{\lambda}) \in V.$$

Defining $\boldsymbol{w} := (g, 0, 0, 0) + (u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathbb{S}} u^*, \gamma_{\boldsymbol{n},\mathbb{S}} \boldsymbol{\sigma}^*)$ and putting altogether shows

$$\begin{aligned} (\boldsymbol{v}, (\mu, \boldsymbol{\lambda}))_V &= b((g, 0, 0, 0), (\mu, \boldsymbol{\lambda})) + b((u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathbb{S}} u^*, \gamma_{\boldsymbol{n},\mathbb{S}} \boldsymbol{\sigma}^*), (\mu, \boldsymbol{\lambda})) \\ &= b(\boldsymbol{w}, (\mu, \boldsymbol{\lambda})). \end{aligned}$$

Thus, $\Theta\boldsymbol{w} = \boldsymbol{v}$.

Case (b). The scalar product in this case is given by

$$((v,\boldsymbol{\tau}),(\mu,\boldsymbol{\lambda}))_V = (\operatorname{div}\boldsymbol{\tau},\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda}) + (\boldsymbol{C}\boldsymbol{\tau},\boldsymbol{\lambda}) + (\boldsymbol{C}^{-1}\nabla v,\nabla_{\mathcal{T}}\mu) + (v,\mu).$$

Recall that $\boldsymbol{\beta} = 0$ and note that $\operatorname{div}\boldsymbol{\tau} = -g + \gamma v$ by (3.3). Therefore,

$$\begin{aligned}(\operatorname{div}\boldsymbol{\tau},\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda}) = (g-\gamma v,-\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda}) &= (g-\gamma v,-\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda}+\gamma\mu) + (\gamma(\gamma v-g),\mu)\\ &= b((g-\gamma v,0,0,0),(\mu,\boldsymbol{\lambda})) + (\gamma(\gamma v-g),\mu).\end{aligned}$$

With $\boldsymbol{C}\boldsymbol{\tau} = \nabla v$ and piecewise integration by parts we obtain

$$\begin{aligned}(\boldsymbol{C}^{-1}\nabla v,\nabla_{\mathcal{T}}\mu) = (\boldsymbol{\tau},\nabla_{\mathcal{T}}\mu) &= \langle\gamma_{\boldsymbol{n},\mathcal{S}}\boldsymbol{\tau},\mu\rangle_{\mathcal{S}} + (-\operatorname{div}\boldsymbol{\tau},\mu)\\ &= b((0,0,0,\gamma_{\boldsymbol{n},\mathcal{S}}\boldsymbol{\tau}),(\mu,\boldsymbol{\lambda})) + (-\operatorname{div}\boldsymbol{\tau},\mu).\end{aligned}$$

Thus,

$$\begin{aligned}((v,\boldsymbol{\tau}),(\mu,\boldsymbol{\lambda}))_V &= (\operatorname{div}\boldsymbol{\tau},\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda}) + (\boldsymbol{C}\boldsymbol{\tau},\boldsymbol{\lambda}) + (\boldsymbol{C}^{-1}\nabla v,\nabla_{\mathcal{T}}\mu) + (v,\mu)\\ &= b((g-\gamma v,0,0,\gamma_{\boldsymbol{n},\mathcal{S}}\boldsymbol{\tau}),(\mu,\boldsymbol{\lambda})) + (\gamma(\gamma v-g)-\operatorname{div}\boldsymbol{\tau}+v,\mu) + (\boldsymbol{C}\boldsymbol{\tau},\boldsymbol{\lambda}).\end{aligned}$$

Defining

$$\boldsymbol{w} := (g-\gamma v,0,0,\gamma_{\boldsymbol{n},\mathcal{S}}\boldsymbol{\tau}) + (u^*,\boldsymbol{\sigma}^*,\gamma_{0,\mathcal{S}}u^*,\gamma_{\boldsymbol{n},\mathcal{S}}\boldsymbol{\sigma}^*),$$

where $(u^*,\boldsymbol{\sigma}^*)$ solves (1.1) with data $f = \gamma(\gamma v - g) - \operatorname{div}\boldsymbol{\tau} + v$, $\boldsymbol{f} = \boldsymbol{\tau}$, shows

$$((v,\boldsymbol{\tau}),(\mu,\boldsymbol{\lambda}))_V = b(\boldsymbol{w},(\mu,\boldsymbol{\lambda})) \quad \text{for all } (\mu,\boldsymbol{\lambda}) \in V.$$

Case (c). The proof is similar as for Case (b). Thus, we only give details on the important differences. We have to take care of the terms involving the matrix $\boldsymbol{C}$. Note that by the assumptions on $\boldsymbol{C}$ it holds $\boldsymbol{C}^{-1}\boldsymbol{\tau} \in \boldsymbol{H}(\operatorname{div};\Omega)$ and $\boldsymbol{C}\boldsymbol{\tau} \in \boldsymbol{H}(\operatorname{div};\Omega)$ as well. We have

$$(\boldsymbol{\tau},\boldsymbol{\lambda}) = (\boldsymbol{C}\boldsymbol{C}^{-1}\boldsymbol{\tau},\boldsymbol{\lambda}),$$

and using $\boldsymbol{C}\boldsymbol{\tau} = \nabla v$ and integration by parts,

$$(\nabla v,\nabla_{\mathcal{T}}\mu) = (\boldsymbol{C}\boldsymbol{\tau},\nabla_{\mathcal{T}}\mu) = \langle\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{C}\boldsymbol{\tau}),\mu\rangle_{\mathcal{S}} - (\operatorname{div}(\boldsymbol{C}\boldsymbol{\tau}),\mu).$$

Defining

$$\boldsymbol{w} := (g-\gamma v,0,0,\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{C}\boldsymbol{\tau})) + (u^*,\boldsymbol{\sigma}^*,\gamma_{0,\mathcal{S}}u^*,\gamma_{\boldsymbol{n},\mathcal{S}}\boldsymbol{\sigma}^*),$$

where $(u^*,\boldsymbol{\sigma}^*)$ solves (1.1) with data $f = \gamma(\gamma v - g) - \operatorname{div}(\boldsymbol{C}\boldsymbol{\tau}) + v$, $\boldsymbol{f} = \boldsymbol{C}^{-1}\boldsymbol{\tau}$, shows

$$\begin{aligned}((v,\boldsymbol{\tau}),(\mu,\boldsymbol{\lambda}))_V &= (\operatorname{div}\boldsymbol{\tau},\operatorname{div}_{\mathcal{T}}\boldsymbol{\lambda}) + (\boldsymbol{\tau},\boldsymbol{\lambda}) + (\nabla v,\nabla_{\mathcal{T}}\mu) + (v,\mu)\\ &= b(\boldsymbol{w},(\mu,\boldsymbol{\lambda})) \quad \text{for all } (\mu,\boldsymbol{\lambda}) \in V.\end{aligned}$$

Finally, note that for all three cases it is straightforward to prove

$$\|f\| + \|\boldsymbol{f}\|_{\boldsymbol{H}^1(\mathcal{T})} \lesssim \|g\|.$$

Then (2.8) shows estimate (3.4). Moreover, in Case (c) we have $\boldsymbol{\tau} = \boldsymbol{C}^{-1}\nabla v \in \boldsymbol{H}^1(\Omega)$ and $\boldsymbol{f} = \boldsymbol{C}^{-1}\boldsymbol{\tau} \in \boldsymbol{H}^1(\Omega)$, thus, $\boldsymbol{\sigma}^* = \boldsymbol{C}^{-1}\boldsymbol{\tau} - \boldsymbol{C}^{-1}\nabla u^* \in \boldsymbol{H}^1(\Omega)$. This finishes the proof. □

Remark 9. As already mentioned in the introduction, in the recent work [15] the optimal test norm

$$\|\boldsymbol{v}\|_{V,\mathrm{opt}} := \sup_{0\neq\boldsymbol{u}\in U}\frac{b(\boldsymbol{u},\boldsymbol{v})}{\|\boldsymbol{u}\|_U}$$

is considered and the problem of finding $\boldsymbol{w} \in U$ such that $G(\boldsymbol{u}) = b(\boldsymbol{u},\Theta\boldsymbol{w})$, where $G \in U'$, is analyzed. In view of our previous results we have $G(\boldsymbol{u}) = (g,u)$ and following [15] one would find $\boldsymbol{w} = (g,0,0,0)$. Since the optimal test norm is not feasible in computations, we only consider the test norms from Lemma 8 in the remainder of the work.

Lemma 10. *Consider one of Cases* (a)–(c). *Let* $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \widehat{u}, \widehat{\sigma}) \in U$ *denote the solution of problem* (2.2) *and let* $\boldsymbol{u}_h = (u_h, \boldsymbol{\sigma}_h, \widehat{u}_h, \widehat{\sigma}_h) \in U_h \in \{U_{hp}, U_{hp}^+\}$ *denote the solution of* (2.5). *Suppose* $(g, 0, 0, 0) \in U_h$, *i.e.*, $g \in \mathcal{P}^p(\mathcal{T})$ *if* $U_h = U_{hp}$ *resp.* $g \in \mathcal{P}^{p+1}(\mathcal{T})$ *if* $U_h = U_{hp}^+$. *Moreover, suppose that*

- $\mathcal{P}^1_{c,0}(\mathcal{T}) \times \mathcal{RT}^p(\mathcal{T}) \subset V_{hk}$ *if* $U_h = U_{hp}$,
- $\mathcal{P}^1_{c,0}(\mathcal{T}) \times \mathcal{RT}^{p+1}(\mathcal{T}) \subset V_{hk}$ *if* $U_h = U_{hp}^+$.

It holds that

$$|(u - u_h, g)| \le Ch\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \|g\|.$$

The constant $C > 0$ *only depends on* Ω, $\boldsymbol{C}$, $\boldsymbol{\beta}$, γ, $p \in \mathbb{N}_0$, *and shape-regularity of* $\mathcal{T}$.

Proof. Let $\boldsymbol{v} = (v, \boldsymbol{\tau}) \in V$ denote the solution of the adjoint problem (3.3) with the given $g \in L^2(\Omega)$. Let $\boldsymbol{w} = \Theta^{-1}\boldsymbol{v} \in U$ denote the element from Lemma 8. Since $(v, \boldsymbol{\tau}) \in H^1_0(\Omega) \times \boldsymbol{H}(\mathrm{div}; \Omega)$, the identities in (2.1) and the adjoint problem (3.3) imply that

$$(u - u_h, g) = b(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{v}).$$

With the bilinear form $a(\cdot, \cdot)$ of the mixed formulation of DPG (Section 3.2) and the fact that

$$b(\boldsymbol{w}, \delta\boldsymbol{v}) = (\boldsymbol{v}, \delta\boldsymbol{v})_V = (\delta\boldsymbol{v}, \boldsymbol{v})_V \quad \text{for all } \delta\boldsymbol{v} \in V,$$

we infer

$$(u - u_h, g) = b(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{v}) = a((\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_h), (\boldsymbol{w}, \boldsymbol{v})).$$

Here, $\boldsymbol{\varepsilon} = 0$ and $\boldsymbol{\varepsilon}_h \in V_{hk}$ is the error function which satisfies $\|\boldsymbol{\varepsilon}_h\|_V \lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U$ (see Section 3.2). This, Galerkin orthogonality and boundedness of the bilinear form $a(\cdot, \cdot)$ show for arbitrary $(\boldsymbol{w}_h, \boldsymbol{v}_h) \in (U_h, V_{hk})$ that

$$\begin{aligned}(u - u_h, g) &= a((\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_h), (\boldsymbol{w}, \boldsymbol{v})) \\ &= a((\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_h), (\boldsymbol{w} - \boldsymbol{w}_h, \boldsymbol{v} - \boldsymbol{v}_h)) \\ &\lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U (\|\boldsymbol{w} - \boldsymbol{w}_h\|_U + \|\boldsymbol{v} - \boldsymbol{v}_h\|_V).\end{aligned}$$

It remains to prove $\|\boldsymbol{w} - \boldsymbol{w}_h\|_U + \|\boldsymbol{v} - \boldsymbol{v}_h\|_V \lesssim h\|g\|$. We estimate $\|\boldsymbol{v} - \boldsymbol{v}_h\|_V$ for all three cases simultaneously and handle the estimation of $\|\boldsymbol{w} - \boldsymbol{w}_h\|_U$ for the three cases separately, since the representation of $\boldsymbol{w}$ by Lemma 8 depends on the choice of norms in V.

We start with the estimation of $\|\boldsymbol{v} - \boldsymbol{v}_h\|_V$: We first consider $U_h = U_{hp}$. Note that $\mathcal{P}^1_{c,0}(\mathcal{T}) \times \mathcal{RT}^p(\mathcal{T}) \subset V_{hk}$. Choose $\boldsymbol{v}_h = (\Pi^1_{\mathrm{V}} v, \Pi^p_{\mathrm{div}} \boldsymbol{\tau}) \in V_{hk}$. Recall that all norms under consideration are equivalent, i.e.,

$$\|\cdot\|_{\mathrm{V,qopt}} \eqsim \|\cdot\|_{\mathrm{V},1} \eqsim \|\cdot\|_{\mathrm{V},2}.$$

Then, using the approximation properties (3.1) together with (3.4), we get

$$\begin{aligned}\|\boldsymbol{v} - \boldsymbol{v}_h\|_V \eqsim \|\boldsymbol{v} - \boldsymbol{v}_h\|_{\mathrm{V},2} &\le \|v - \Pi^1_{\mathrm{V}} v\|_{H^1(\Omega)} + \|\boldsymbol{\tau} - \Pi^p_{\mathrm{div}} \boldsymbol{\tau}\|_{\boldsymbol{H}(\mathrm{div};\Omega)} \\ &\lesssim h\|g\| + \|\mathrm{div}(\boldsymbol{\tau} - \Pi^p_{\mathrm{div}} \boldsymbol{\tau})\|.\end{aligned}$$

Then, for the remaining term the commutativity property of the Raviart–Thomas projection, the adjoint problem (3.3) and $g \in \mathcal{P}^p(\mathcal{T})$ yield

$$\begin{aligned}\|\mathrm{div}(\boldsymbol{\tau} - \Pi^p_{\mathrm{div}} \boldsymbol{\tau})\| &= \|(1 - \Pi^p) \,\mathrm{div}\, \boldsymbol{\tau}\| \\ &= \|(1 - \Pi^p)(-g - \boldsymbol{\beta} \cdot \boldsymbol{\tau} + \gamma v)\| \\ &= \|(1 - \Pi^p)(\gamma v - \boldsymbol{\beta} \cdot \boldsymbol{\tau})\|.\end{aligned}$$

Using the approximation properties of Π^0, $\gamma \in C^1(\mathcal{T})$, $\boldsymbol{\beta} \in C^1(\mathcal{T})^d$, and (3.4) shows

$$\begin{aligned}\|(1 - \Pi^p)(\gamma v - \boldsymbol{\beta} \cdot \boldsymbol{\tau})\| &\le \|(1 - \Pi^0)(\gamma v - \boldsymbol{\beta} \cdot \boldsymbol{\tau})\| \\ &\lesssim h\|\nabla_{\mathcal{T}}(\gamma v - \boldsymbol{\beta} \cdot \boldsymbol{\tau})\| \\ &\lesssim h\|g\|.\end{aligned}$$

Therefore, we obtain $\|\boldsymbol{v} - \boldsymbol{v}_h\|_V \lesssim h\|g\|$. If $U_h = U_{hp}^+$, then we choose $\boldsymbol{v}_h = (\Pi^1_{\mathrm{V}}, \Pi^{p+1}_{\mathrm{div}} \boldsymbol{\tau}) \in V_{hk}$. With the same lines of proof we also infer $\|\boldsymbol{v} - \boldsymbol{v}_h\|_V \lesssim h\|g\|$.

It only remains to estimate $\|\boldsymbol{w} - \boldsymbol{w}_h\|_U$. We distinguish between the three different cases.

Case (a). By Lemma 8 we have $\boldsymbol{w} = (g, 0, 0, 0) + \widetilde{\boldsymbol{w}}$, where $\widetilde{\boldsymbol{w}} = (u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathcal{S}} u^*, \gamma_{\boldsymbol{n},\mathcal{S}} \boldsymbol{\sigma}^*)$. We choose

$$\boldsymbol{w}_h = (g, 0, 0, 0) + \widetilde{\boldsymbol{w}}_h,$$

where $\widetilde{\boldsymbol{w}}_h \in U_{h0} \subseteq U_h$ is the best-approximation of $(u^*, \boldsymbol{\sigma}^*, \gamma_{0,\mathcal{S}} u^*, \gamma_{\boldsymbol{n},\mathcal{S}} \boldsymbol{\sigma}^*)$ with respect to $\|\cdot\|_U$. From Theorem 6 and (3.4) it follows that

$$\|\boldsymbol{w} - \boldsymbol{w}_h\|_U = \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}_h\|_U \lesssim h\|g\|.$$

Case (b). By Lemma 8 we have $\boldsymbol{w} = (g - \gamma v, 0, 0, \gamma_{\boldsymbol{n},\mathcal{S}} \boldsymbol{\tau}) + \widetilde{\boldsymbol{w}}$ and choose

$$\boldsymbol{w}_h = (g - \Pi^0 \gamma v, 0, 0, \gamma_{\boldsymbol{n},\mathcal{S}} \Pi^0_{\mathrm{div}} \boldsymbol{\tau}) + \widetilde{\boldsymbol{w}}_h,$$

where $\widetilde{\boldsymbol{w}}_h \in U_{h0}$ is the best approximation of $\widetilde{\boldsymbol{w}}$ with respect to $\|\cdot\|_U$. Note that the same arguments as before lead to $\|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}_h\|_U \lesssim h\|g\|$. Therefore,

$$\|\boldsymbol{w} - \boldsymbol{w}_h\|_U \le \|(1 - \Pi^0)\gamma v\| + \|\gamma_{\boldsymbol{n},\mathcal{S}}(\boldsymbol{\tau} - \Pi^0_{\mathrm{div}} \boldsymbol{\tau})\|_{-1/2,\mathcal{S}} + \|\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{w}}_h\|_U \lesssim h\|g\|,$$

where we used (3.1) and the approximation property of $\gamma_{\boldsymbol{n},\mathcal{S}} \Pi^p_{\mathrm{div}}$ in the $H^{-1/2}(\mathcal{S})$ norm (see the proof of Theorem 6) together with (3.4).

Case (c). The proof follows as for Case (b). Therefore, we omit the details. □

3.4 Proof of Theorem 3

The best approximation property of Π^p and the triangle inequality show that

$$\|u - \Pi^p u\| \le \|u - u_h\| \le \|u - \Pi^p u\| + \|\Pi^p(u - u_h)\|.$$

With $g := \Pi^p u - u_h \in \mathcal{P}^p(\mathcal{T})$ observe that

$$\|g\|^2 = (g, g) = (\Pi^p(u - u_h), g) = (u - u_h, g).$$

We apply Lemma 10, and the approximation result from Theorem 6 to see

$$\|g\|^2 = (u - u_h, g) \lesssim h\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \|g\| \lesssim h h^{p+1}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})})\|g\|.$$

Dividing by $\|g\|$ we infer

$$\|u - u_h\| \le \|u - \Pi^p u\| + \|g\| \le \|u - \Pi^p u\| + Ch^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}),$$

which finishes the proof.

3.5 Proof of Theorem 4

The proof is similar to the one for Theorem 3. We consider

$$\|u - u_h^+\| \le \|u - \Pi^{p+1} u\| + \|\Pi^{p+1} u - u_h^+\|.$$

Define $g := \Pi^{p+1} u - u_h^+ \in \mathcal{P}^{p+1}(\mathcal{T})$. To estimate the second term, we argue as in the proof of Theorem 3 to obtain $\|g\| \lesssim h^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})})$. The first term is estimated with the approximation property (3.1a) of the L^2 projection, i.e.,

$$\|u - \Pi^{p+1} u\| \lesssim h^{p+2}\|u\|_{H^{p+2}(\Omega)}.$$

This finishes the proof.

3.6 Proof of Theorem 5

Note that (2.9b) is equivalent to $\Pi^0 \widetilde{u}_h = \Pi^0 u_h$. This yields

$$\begin{aligned}\|u - \widetilde{u}_h\| &\le \|(1 - \Pi^0)(u - \widetilde{u}_h)\| + \|\Pi^0(u - \widetilde{u}_h)\| \\ &\lesssim h\|\nabla_{\mathcal{T}}(u - \widetilde{u}_h)\| + \|\Pi^0(u - u_h)\|,\end{aligned}$$

where we have used the local approximation property of Π^0. We define $g := \Pi^0(u - u_h)$. Applying Lemma 10 and Theorem 6 shows

$$\begin{aligned}\|g\|^2 &= (\Pi^0(u - u_h), g) \\ &= (u - u_h, g) \\ &\lesssim h\|\boldsymbol{u} - \boldsymbol{u}_h\|_U \|g\| \\ &\lesssim h^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})})\|g\|.\end{aligned}$$

It remains to estimate $\|\nabla_{\mathcal{T}}(u - \widetilde{u}_h)\|$. The proof follows standard arguments from finite element analysis and is included for completeness. To that end define $\overline{u}_h \in \mathcal{P}^{p+1}(\mathcal{T})$ as the solution of the auxiliary Neumann problem

$$\begin{aligned}(\nabla \overline{u}_h, \nabla v_h)_T &= (\boldsymbol{C}\boldsymbol{f} - \boldsymbol{C}\boldsymbol{\sigma} + \boldsymbol{\beta}u, \nabla v_h)_T \quad \text{for all } v_h \in \mathcal{P}^{p+1}(T), \\ (\overline{u}_h, 1)_T &= 0\end{aligned}$$

for all $T \in \mathcal{T}$. Then

$$\begin{aligned}\|\nabla_{\mathcal{T}}(\overline{u}_h - \widetilde{u}_h)\|^2 &= (-\boldsymbol{C}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h) + \boldsymbol{\beta}(u - u_h), \nabla_{\mathcal{T}}(\overline{u}_h - \widetilde{u}_h)) \\ &\lesssim \|\boldsymbol{u} - \boldsymbol{u}_h\|_U \|\nabla_{\mathcal{T}}(\overline{u}_h - \widetilde{u}_h)\| \\ &\lesssim h^{p+1}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})})\|\nabla_{\mathcal{T}}(\overline{u}_h - \widetilde{u}_h)\|.\end{aligned}$$

To estimate $\|\nabla_{\mathcal{T}}(u - \overline{u}_h)\|$, note that there holds Galerkin orthogonality

$$(\nabla_{\mathcal{T}}(u - \overline{u}_h), \nabla_{\mathcal{T}} v_h) = 0 \quad \text{for all } v_h \in \mathcal{P}^{p+1}(\mathcal{T}).$$

Hence, standard approximation results show

$$\|\nabla_{\mathcal{T}}(u - \overline{u}_h)\| = \min_{v_h \in \mathcal{P}^{p+1}(\mathcal{T})} \|\nabla_{\mathcal{T}}(u - v_h)\| \lesssim h^{p+1}\|u\|_{H^{p+2}(\Omega)}.$$

Putting altogether gives

$$\begin{aligned}\|u - \widetilde{u}_h\| &\lesssim h\|\nabla_{\mathcal{T}}(u - \widetilde{u}_h)\| + \|g\| \\ &\lesssim h(\|\nabla_{\mathcal{T}}(u - \overline{u}_h)\| + \|\nabla_{\mathcal{T}}(\overline{u}_h - \widetilde{u}_h)\|) + h^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}) \\ &\lesssim h^{p+2}(\|u\|_{H^{p+2}(\Omega)} + \|\boldsymbol{\sigma}\|_{\boldsymbol{H}^{p+1}(\mathcal{T})}),\end{aligned}$$

which finishes the proof.

4 Numerical Studies

In this section we present results of two numerical examples. Let $\Omega = (0, 1)^2$ be a square domain. Throughout we consider the manufactured solution

$$u(x, y) = \sin(\pi x)\sin(\pi y), \quad (x, y) \in \Omega,$$

which is smooth and satisfies $u|_\Gamma = 0$.

Let $\boldsymbol{u}_h = (u_h, \boldsymbol{\sigma}_h, \widehat{u}_h, \widehat{\sigma}_h) \in U_{hp}$ and $\boldsymbol{u}_h^+ = (u_h^+, \boldsymbol{\sigma}_h^+, \widehat{u}_h^+, \widehat{\sigma}_h^+) \in U_{hp}^+$ denote the solutions of the practical DPG method (2.5) and let $\widetilde{u}_h \in \mathcal{P}^{p+1}(\mathcal{T})$ be the postprocessed solution of $\boldsymbol{u}_h$, see Section 2.10. We present results for $p = 0, 1, 2, 3$, where we use the test space

$$V_{hk} := \mathcal{P}^{p+2}(\mathcal{T}) \times \mathcal{P}^{p+2}(\mathcal{T})^d.$$

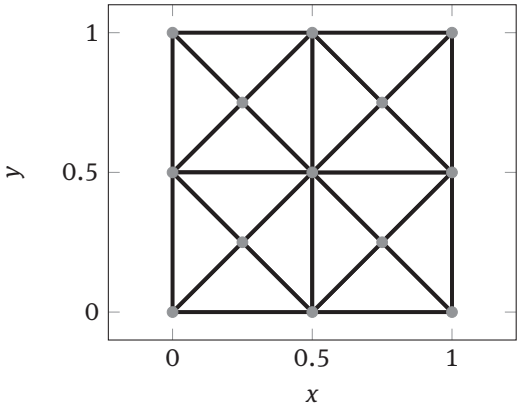

Figure 1: Initial triangulation $\mathcal{T}_1$ of domain $\Omega = (0,1)^2$.

To verify our main results (Theorem 3, Theorem 4, and Theorem 5), we check the convergence rates of the L^2 errors $\|\Pi^p u - u_h\|$, $\|u - u_h^+\|$, and $\|u - \widetilde{u}_h\|$. In all examples below we choose $\boldsymbol{C}$ to be the identity matrix. Thus, $\|\cdot\|_{\mathrm{V},1} = \|\cdot\|_{\mathrm{V},2}$ and Cases (b), (c) are identical. The other coefficients are chosen such that the regularity assumptions (2.8) are satisfied.

All computations start with the initial triangulation $\mathcal{T}_1$ visualized in Figure 1.

4.1 Example 1

Define $T_1 := \operatorname{conv}\{(0,0),(1,0),(\frac{1}{2},\frac{1}{2})\}$, $T_2 := \operatorname{conv}\{(1,1),(0,1),(\frac{1}{2},\frac{1}{2})\}$. In the first example we set $\boldsymbol{\beta} = 0$ and

$$\gamma(x,y) := \begin{cases} 1 & (x,y) \in T_1, \\ \frac{1}{2} & (x,y) \in T_2, \\ 0 & (x,y) \in \Omega \setminus (T_1 \cup T_2). \end{cases}$$

Moreover, we choose

$$\boldsymbol{f}(x,y) := \begin{cases} \begin{pmatrix} 1 \\ 1 \end{pmatrix} & x < \frac{1}{2}, \\ \begin{pmatrix} 1 \\ -1 \end{pmatrix} & x \geq \frac{1}{2}. \end{cases}$$

Note that $\operatorname{div} \boldsymbol{f} = 0$ and $\boldsymbol{f} \in \boldsymbol{H}(\operatorname{div};\Omega) \cap \boldsymbol{H}^1(\mathcal{T}_1)$. With the coefficients, $\boldsymbol{f}$ and the exact solution at hand, we calculate the right-hand side f and $\boldsymbol{\sigma}$ through (1.1). Table 2 resp. Table 3 show errors and convergence rates when using the test norm $\|\cdot\|_{\mathrm{V,qopt}}$ resp. $\|\cdot\|_{\mathrm{V},1} = \|\cdot\|_{\mathrm{V},2}$.

4.2 Example 2

For this example we choose

$$\boldsymbol{f} = 0, \quad \gamma = 0, \quad \boldsymbol{\beta}(x,y) = (1,1)^T \quad \text{for } (x,y) \in \Omega\,.$$

Note that $\boldsymbol{\beta}$ is smooth. Again we calculate f and $\boldsymbol{\sigma}$ through (1.1). Table 4 resp. Table 5 show the results for Case (a) ($\|\cdot\|_V = \|\cdot\|_{\mathrm{V,qopt}}$) resp. Case (b), (c) ($\|\cdot\|_V = \|\cdot\|_{\mathrm{V},1} = \|\cdot\|_{\mathrm{V},2}$). Observe from Table 5 that we do not get higher convergence rates neither for solutions from the augmented space U_{hp}^+ nor for the postprocessed solution. Even for the L^2 error of $\Pi^p u - u_h$ we do not get higher rates, whereas with the use of the quasi-optimal test norm $\|\cdot\|_{\mathrm{V,qopt}}$ higher rates are obtained. This demonstrates that the assumption $\boldsymbol{\beta} = 0$ in Section 2.7 for the Cases (b)–(c) is not an artefact used in the proofs but in general is also necessary to obtain superconvergence results with the norms $\|\cdot\|_{\mathrm{V},1}$, $\|\cdot\|_{\mathrm{V},2}$.

p	$\#\mathcal{T}$	$\|u-u_h\|$	rate	$\|\Pi^p u-u_h\|$	rate	$\|u-u_h^+\|$	rate	$\|u-\widetilde{u}_h\|$	rate
0	16	1.94e−01	–	7.41e−02	–	8.37e−02	–	1.23e−01	–
	64	9.37e−02	1.05	1.85e−02	2.00	2.09e−02	2.01	3.21e−02	1.94
	256	4.64e−02	1.01	4.63e−03	2.00	5.20e−03	2.00	8.12e−03	1.98
	1024	2.32e−02	1.00	1.16e−03	2.00	1.30e−03	2.00	2.04e−03	2.00
	4096	1.16e−02	1.00	2.90e−04	2.00	3.25e−04	2.00	5.09e−04	2.00
	16384	5.79e−03	1.00	7.24e−05	2.00	8.13e−05	2.00	1.27e−04	2.00
	65536	2.89e−03	1.00	1.81e−05	2.00	2.03e−05	2.00	3.18e−05	2.00
1	16	3.47e−02	–	3.02e−03	–	5.96e−03	–	7.89e−03	–
	64	8.86e−03	1.97	5.58e−04	2.44	8.72e−04	2.77	9.53e−04	3.05
	256	2.22e−03	1.99	7.92e−05	2.82	1.16e−04	2.92	1.18e−04	3.01
	1024	5.56e−04	2.00	1.02e−05	2.95	1.47e−05	2.98	1.48e−05	3.00
	4096	1.39e−04	2.00	1.29e−06	2.99	1.84e−06	2.99	1.84e−06	3.00
	16384	3.48e−05	2.00	1.62e−07	3.00	2.31e−07	3.00	2.30e−07	3.00
2	16	4.51e−03	–	2.55e−04	–	3.51e−04	–	6.14e−04	–
	64	5.74e−04	2.98	1.30e−05	4.29	1.98e−05	4.14	4.18e−05	3.88
	256	7.20e−05	2.99	7.67e−07	4.09	1.21e−06	4.04	2.68e−06	3.96
	1024	9.01e−06	3.00	4.72e−08	4.02	7.50e−08	4.01	1.68e−07	3.99
	4096	1.13e−06	3.00	2.97e−09	3.99	4.69e−09	4.00	1.05e−08	4.00
3	16	2.20e−04	–	2.08e−05	–	2.01e−05	–	5.48e−05	–
	64	1.39e−05	3.98	8.34e−07	4.64	8.38e−07	4.58	1.67e−06	5.03
	256	8.70e−07	4.00	2.82e−08	4.89	2.86e−08	4.87	5.18e−08	5.01
	1024	5.44e−08	4.00	9.08e−10	4.96	9.24e−10	4.95	1.62e−09	5.00

Table 2: Errors and rates for the problem from Section 4.1 with test norm $\|\cdot\|_V = \|\cdot\|_{V,\mathrm{qopt}}$.

p	$\#\mathcal{T}$	$\|u-u_h\|$	rate	$\|\Pi^p u-u_h\|$	rate	$\|u-u_h^+\|$	rate	$\|u-\widetilde{u}_h\|$	rate
0	16	1.92e−01	–	6.88e−02	–	7.86e−02	–	8.48e−02	–
	64	9.35e−02	1.04	1.73e−02	1.99	1.97e−02	1.99	2.17e−02	1.97
	256	4.64e−02	1.01	4.33e−03	2.00	4.94e−03	2.00	5.44e−03	1.99
	1024	2.32e−02	1.00	1.08e−03	2.00	1.23e−03	2.00	1.36e−03	2.00
	4096	1.16e−02	1.00	2.71e−04	2.00	3.09e−04	2.00	3.41e−04	2.00
	16384	5.79e−03	1.00	6.77e−05	2.00	7.71e−05	2.00	8.51e−05	2.00
	65536	2.89e−03	1.00	1.69e−05	2.00	1.93e−05	2.00	2.13e−05	2.00
1	16	3.49e−02	–	4.81e−03	–	6.96e−03	–	6.79e−03	–
	64	8.87e−03	1.98	7.36e−04	2.71	9.71e−04	2.84	8.82e−04	2.95
	256	2.22e−03	2.00	9.82e−05	2.91	1.26e−04	2.95	1.12e−04	2.98
	1024	5.56e−04	2.00	1.25e−05	2.97	1.59e−05	2.99	1.41e−05	2.99
	4096	1.39e−04	2.00	1.57e−06	2.99	1.99e−06	3.00	1.76e−06	3.00
	16384	3.48e−05	2.00	1.96e−07	3.00	2.49e−07	3.00	2.20e−07	3.00
2	16	4.53e−03	–	4.38e−04	–	5.07e−04	–	5.22e−04	–
	64	5.74e−04	2.98	2.53e−05	4.11	3.00e−05	4.08	3.25e−05	4.01
	256	7.20e−05	2.99	1.54e−06	4.04	1.85e−06	4.02	2.03e−06	4.00
	1024	9.01e−06	3.00	9.58e−08	4.01	1.15e−07	4.01	1.27e−07	4.00
	4096	1.13e−06	3.00	6.03e−09	3.99	7.22e−09	3.99	7.94e−09	4.00
3	16	2.25e−04	–	5.14e−05	–	5.06e−05	–	6.01e−05	–
	64	1.40e−05	4.01	1.75e−06	4.88	1.73e−06	4.87	1.96e−06	4.94
	256	8.71e−07	4.00	5.62e−08	4.96	5.55e−08	4.96	6.20e−08	4.98
	1024	5.44e−08	4.00	1.80e−09	4.96	1.78e−09	4.96	1.96e−09	4.98

Table 3: Errors and rates for the problem from Section 4.1 with test norm $\|\cdot\|_V = \|\cdot\|_{V,2}$.

p	$\#\mathcal{T}$	$\|u-u_h\|$	rate	$\|\Pi^p u-u_h\|$	rate	$\|u-u_h^+\|$	rate	$\|u-\widetilde{u}_h\|$	rate
0	16	1.96e−01	–	7.95e−02	–	8.85e−02	–	1.27e−01	–
	64	9.41e−02	1.06	2.04e−02	1.96	2.25e−02	1.98	3.36e−02	1.92
	256	4.65e−02	1.02	5.14e−03	1.99	5.64e−03	1.99	8.51e−03	1.98
	1024	2.32e−02	1.00	1.29e−03	2.00	1.41e−03	2.00	2.13e−03	1.99
	4096	1.16e−02	1.00	3.22e−04	2.00	3.53e−04	2.00	5.34e−04	2.00
	16384	5.79e−03	1.00	8.05e−05	2.00	8.82e−05	2.00	1.34e−04	2.00
	65536	2.89e−03	1.00	2.01e−05	2.00	2.21e−05	2.00	3.34e−05	2.00
1	16	3.47e−02	–	2.77e−03	–	5.91e−03	–	8.02e−03	–
	64	8.85e−03	1.97	5.22e−04	2.40	8.59e−04	2.78	9.73e−04	3.04
	256	2.22e−03	1.99	7.47e−05	2.80	1.14e−04	2.92	1.21e−04	3.01
	1024	5.56e−04	2.00	9.69e−06	2.95	1.44e−05	2.98	1.51e−05	3.00
	4096	1.39e−04	2.00	1.22e−06	2.99	1.81e−06	2.99	1.89e−06	3.00
	16384	3.48e−05	2.00	1.53e−07	3.00	2.27e−07	3.00	2.36e−07	3.00
2	16	4.51e−03	–	2.37e−04	–	3.44e−04	–	6.25e−04	–
	64	5.73e−04	2.98	1.19e−05	4.32	1.95e−05	4.14	4.24e−05	3.88
	256	7.20e−05	2.99	6.97e−07	4.09	1.19e−06	4.04	2.72e−06	3.97
	1024	9.01e−06	3.00	4.28e−08	4.02	7.37e−08	4.01	1.71e−07	3.99
	4096	1.13e−06	3.00	2.68e−09	4.00	4.60e−09	4.00	1.07e−08	4.00
3	16	2.20e−04	–	1.95e−05	–	1.98e−05	–	5.51e−05	–
	64	1.39e−05	3.98	7.80e−07	4.64	8.14e−07	4.61	1.68e−06	5.04
	256	8.70e−07	4.00	2.65e−08	4.88	2.78e−08	4.87	5.21e−08	5.01
	1024	5.44e−08	4.00	8.73e−10	4.92	9.16e−10	4.92	1.63e−09	5.00

Table 4: Errors and rates for the problem from Section 4.2 with test norm $\|\cdot\|_V = \|\cdot\|_{\mathrm{V,qopt}}$.

p	$\#\mathcal{T}$	$\|u-u_h\|$	rate	$\|\Pi^p u-u_h\|$	rate	$\|u-u_h^+\|$	rate	$\|u-\widetilde{u}_h\|$	rate
0	16	4.37e−01	–	3.98e−01	–	4.15e−01	–	4.00e−01	–
	64	2.25e−01	0.96	2.06e−01	0.95	2.14e−01	0.96	2.06e−01	0.96
	256	1.14e−01	0.99	1.04e−01	0.98	1.08e−01	0.99	1.04e−01	0.99
	1024	5.70e−02	1.00	5.21e−02	1.00	5.41e−02	1.00	5.21e−02	1.00
	4096	2.85e−02	1.00	2.61e−02	1.00	2.71e−02	1.00	2.61e−02	1.00
	16384	1.43e−02	1.00	1.30e−02	1.00	1.35e−02	1.00	1.30e−02	1.00
	65536	7.13e−03	1.00	6.52e−03	1.00	6.77e−03	1.00	6.52e−03	1.00
1	16	6.23e−02	–	5.18e−02	–	5.69e−02	–	1.64e−02	–
	64	1.63e−02	1.93	1.37e−02	1.92	1.49e−02	1.93	5.70e−03	1.52
	256	4.12e−03	1.98	3.47e−03	1.98	3.77e−03	1.98	1.61e−03	1.83
	1024	1.03e−03	2.00	8.67e−04	2.00	9.42e−04	2.00	4.19e−04	1.94
	4096	2.57e−04	2.00	2.17e−04	2.00	2.35e−04	2.00	1.06e−04	1.98
	16384	6.43e−05	2.00	5.41e−05	2.00	5.88e−05	2.00	2.68e−05	1.99
2	16	7.46e−03	–	5.95e−03	–	6.56e−03	–	9.34e−04	–
	64	9.32e−04	3.00	7.35e−04	3.02	8.17e−04	3.00	6.23e−05	3.91
	256	1.17e−04	3.00	9.17e−05	3.00	1.02e−04	3.00	4.01e−06	3.96
	1024	1.46e−05	3.00	1.15e−05	3.00	1.28e−05	3.00	2.57e−07	3.96
	4096	1.82e−06	3.00	1.43e−06	3.00	1.59e−06	3.00	1.66e−08	3.95
3	16	6.03e−04	–	5.62e−04	–	5.59e−04	–	7.46e−05	–
	64	3.88e−05	3.96	3.63e−05	3.95	3.64e−05	3.94	3.93e−06	4.25
	256	2.44e−06	3.99	2.28e−06	3.99	2.29e−06	3.99	2.41e−07	4.03
	1024	1.52e−07	4.00	1.42e−07	4.00	1.43e−07	4.00	1.52e−08	3.99

Table 5: Errors and rates for the problem from Section 4.2 with test norm $\|\cdot\|_V = \|\cdot\|_{\mathrm{V},2}$.

5 Concluding Remarks

We conclude this work with some remarks. The results and their proofs are presented in a systematic way that allow to extend and transfer them to other types of meshes and different model problems. In principle, the crucial results Lemma 8 and Lemma 10 have to be verified. Consider for instance that $\mathcal{T}$ is a mesh with polygonal elements. Lemma 8 still holds true in that case since it is independent of the underlying mesh so that only the assertion of Lemma 10 has to be shown. To be more precise: Analyzing the proof one finds out that it only remains to provide the estimate

$$\min_{\boldsymbol{w}_h \in U_h} \|\boldsymbol{w} - \boldsymbol{w}_h\|_U + \min_{\boldsymbol{v}_k \in V_{hk}} \|\boldsymbol{v} - \boldsymbol{v}_k\|_V \lesssim h\|g\|,$$

which is an optimal a priori error bound for sufficient regular functions (see Lemma 10 for details on the definition of the functions $\boldsymbol{w}$ and $\boldsymbol{v}$). In the case of triangular meshes we have proven the estimate by using basic properties of well-known interpolation operators. If operators with the same properties can be defined on meshes with polygonal elements, then, clearly, the estimate holds true as well. We note that the analysis of DPG methods for ultra-weak formulations on general (polygonal) meshes is an ongoing research. For an overview we refer to the recent work [21].

Future research will include other model problems, e.g., linear elasticity. Another possible application of the developed ideas could be to the Stokes problem. Consider its velocity-gradient-pressure formulation: Find $(\boldsymbol{u}_S, \boldsymbol{\sigma}_S, p_S)$ such that

$$\begin{aligned} -\nabla p_S + \operatorname{div} \boldsymbol{\sigma}_S &= \boldsymbol{f} && \text{in } \Omega, \\ \boldsymbol{\sigma}_S - \nabla \boldsymbol{u}_S &= 0 && \text{in } \Omega, \\ \operatorname{div} \boldsymbol{u}_S &= 0 && \text{in } \Omega, \\ \boldsymbol{u}_S &= 0 && \text{on } \partial\Omega. \end{aligned}$$

DPG methods based on ultra-weak formulations are known and thoroughly analyzed [17]. Since regularity theory is also known, our main results (Theorem 3–5) should carry over (for the velocity variable $\boldsymbol{u}_S$ instead of u) to the Stokes problem following the same lines in the proofs. In particular, the assertion of Theorem 3 has been already observed in numerical experiments [17, Section 3] even for different test norms. We refer also to [18, Section 3] for numerical evidence in the case of incompressible Navier Stokes problems.

Another point we like to mention is that the principal ideas of the proofs and, thus, our main results carry over to the low regularity case, i.e., when we do not have the "full" regularity $u \in H^2(\Omega)$, $v \in H^2(\Omega)$ for solutions of (1.1) and (2.7) but rather $u \in H^{1+s}(\Omega)$, $v \in H^{1+s}(\Omega)$ for some $s \in (\frac{1}{2}, 1)$. This is usually the case when Ω is a nonconvex polygonal domain. Nevertheless, we stress that our main results (Theorem 3–5) hold true with h^{p+2} replaced by h^{p+1+s}. Therefore, one still obtains higher convergence rates than the overall error $\|\boldsymbol{u} - \boldsymbol{u}_h\| = \mathcal{O}(h^{p+1})$. For the particular case of a reaction-diffusion model problem ($\boldsymbol{C}$ is the identity, $\boldsymbol{\beta} = 0$, and $\gamma = 1$) Theorem 4 and 5 are analyzed in [10] for $\|\cdot\|_V = \|\cdot\|_{\mathrm{V},1} = \|\cdot\|_{\mathrm{V},2}$.

Finally, let us remark the importance of the choice of norms in the test space. Although all test norms under consideration are equivalent and, thus, the corresponding DPG methods have the same stability properties (i.e., the inf–sup constants resp. boundedness constants are equivalent), only one of the norms under consideration (the quasi-optimal norm $\|\cdot\|_{\mathrm{V,qopt}}$) yields higher convergence rates for general model problems with $\boldsymbol{\beta} \neq 0$. This has to be taken into account in the design of DPG methods.

Funding: This work was supported by FONDECYT project 11170050.

References

[1] T. Bouma, J. Gopalakrishnan and A. Harb, Convergence rates of the DPG method with reduced test space degree, *Comput. Math. Appl.* **68** (2014), no. 11, 1550–1561.

[2] C. Carstensen, L. Demkowicz and J. Gopalakrishnan, A posteriori error control for DPG methods, *SIAM J. Numer. Anal.* **52** (2014), no. 3, 1335–1353.

[3] C. Carstensen, L. Demkowicz and J. Gopalakrishnan, Breaking spaces and forms for the DPG method and applications including Maxwell equations, *Comput. Math. Appl.* **72** (2016), no. 3, 494–522.
[4] B. Cockburn, B. Dong and J. Guzmán, A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems, *Math. Comp.* **77** (2008), no. 264, 1887–1916.
[5] L. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. Part I: The transport equation, *Comput. Methods Appl. Mech. Engrg.* **199** (2010), no. 23–24, 1558–1572.
[6] L. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions, *Numer. Methods Partial Differential Equations* **27** (2011), no. 1, 70–105.
[7] L. Demkowicz and J. Gopalakrishnan, Analysis of the DPG method for the Poisson equation, *SIAM J. Numer. Anal.* **49** (2011), no. 5, 1788–1809.
[8] L. Demkowicz, J. Gopalakrishnan and A. H. Niemi, A class of discontinuous Petrov–Galerkin methods. Part III: Adaptivity, *Appl. Numer. Math.* **62** (2012), no. 4, 396–427.
[9] A. Demlow, Suboptimal and optimal convergence in mixed finite element methods, *SIAM J. Numer. Anal.* **39** (2002), no. 6, 1938–1953.
[10] T. Führer, Superconvergence in a DPG method for an ultra-weak formulation, *Comput. Math. Appl.* **75** (2018), no. 5, 1705–1718.
[11] L. Gastaldi and R. H. Nochetto, Sharp maximum norm error estimates for general mixed finite element approximations to second order elliptic equations, *RAIRO Modél. Math. Anal. Numér.* **23** (1989), no. 1, 103–128.
[12] J. Gopalakrishnan and W. Qiu, An analysis of the practical DPG method, *Math. Comp.* **83** (2014), no. 286, 537–552.
[13] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.
[14] B. Keith, L. Demkowicz and J. Gopalakrishnan, DPG* method, preprint (2017), https://arxiv.org/abs/1710.05223.
[15] B. Keith, A. Vaziri Astaneh and L. Demkowicz, Goal-oriented adaptive mesh refinement for non-symmetric functional settings, preprint (2017), https://arxiv.org/abs/1711.01996.
[16] S. Nagaraj, S. Petrides and L. F. Demkowicz, Construction of DPG Fortin operators for second order problems, *Comput. Math. Appl.* **74** (2017), no. 8, 1964–1980.
[17] N. V. Roberts, T. Bui-Thanh and L. Demkowicz, The DPG method for the Stokes problem, *Comput. Math. Appl.* **67** (2014), no. 4, 966–995.
[18] N. V. Roberts, L. Demkowicz and R. Moser, A discontinuous Petrov–Galerkin methodology for adaptive solutions to the incompressible Navier–Stokes equations, *J. Comput. Phys.* **301** (2015), 456–483.
[19] L. R. Scott and S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions, *Math. Comp.* **54** (1990), no. 190, 483–493.
[20] R. Stenberg, Postprocessing schemes for some mixed finite elements, *RAIRO Modél. Math. Anal. Numér.* **25** (1991), no. 1, 151–167.
[21] A. Vaziri Astaneh, F. Fuentes, J. Mora and L. Demkowicz, High-order polygonal discontinuous Petrov–Galerkin (PolyDPG) methods using ultraweak formulations, *Comput. Methods Appl. Mech. Engrg.* **332** (2018), 686–711.
[22] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo and V. M. Calo, A class of discontinuous Petrov–Galerkin methods. Part IV: the optimal test norm and time-harmonic wave propagation in 1D, *J. Comput. Phys.* **230** (2011), no. 7, 2406–2432.