A DPG FRAMEWORK FOR STRONGLY MONOTONE OPERATORS

Pierre Cantin, Norbert Heuer

A DPG FRAMEWORK FOR STRONGLY MONOTONE OPERATORS*

PIERRE CANTIN† AND NORBERT HEUER‡

Abstract. We present and analyze a hybrid technique to numerically solve strongly monotone nonlinear problems by the discontinuous Petrov–Galerkin method with optimal test functions (DPG). Our strategy is to relax the nonlinear problem to a linear one with additional unknown and to consider the nonlinear relation as a constraint. We propose to use optimal test functions only for the linear problem and to enforce the nonlinear constraint by penalization. In fact, our scheme can be seen as a minimum residual method with nonlinear penalty term. We develop an abstract framework of the relaxed DPG scheme and prove under appropriate assumptions the well-posedness of the continuous formulation and the quasi-optimal convergence of its discretization. As an application we consider an advection-diffusion problem with nonlinear diffusion of strongly monotone type. Some numerical results in the lowest-order setting are presented to illustrate the predicted convergence.

Key words. Discontinuous Petrov–Galerkin method, optimal test functions, strongly monotone operator, advection-diffusion, nonlinear penalty.

AMS subject classifications. 65N30, 65J15, 65N12, 47H05

1. Introduction.

In recent years, the discontinuous Petrov–Galerkin method with optimal test functions ("DPG method" in the following) has proved to be an attractive strategy to produce inf-sup stable approximations for a wide class of problems. The basic setting stems from Demkowicz and Gopalakrishnan [14, 13] and has been extended, e.g., to linear elasticity [1, 18], the Stokes and Maxwell equations [28, 7], the Schrödinger equation [15], boundary integral and fractional equations [24, 17]. Another promising application area is singularly perturbed problems [16, 9, 3, 4, 25].

All the cited references, however, deal with linear problems. An extension of the DPG technology to nonlinear problems, on the other hand, is a delicate issue. Principal problem is that the calculation (or approximation) of optimal test functions involves an application of the underlying operator (the DPG method is a minimum residual method). For nonlinear problems this step thus becomes nonlinear, i.e., expensive. One way to circumvent the nonlinearity is, of course, to linearize the underlying problem. This has been the approach in [8, 29]. A different idea is to apply the minimum residual technique in product or "broken" spaces to the nonlinear problem. Bui-Thanh and Ghattas [5] did this by considering the entire nonlinear problem as a constraint, and Carstensen *et al.* [6] developed a representation of the DPG scheme by a nonlinear mixed form and analyzed the case of lowest order approximations. DPG for contact problems has been studied in [21], though in this case the nonlinearity is due to the contact condition which is treated by a variational inequality. We also note that Muga and van der Zee [27] study problems posed in Banach spaces. In those cases the calculation of optimal test functions becomes nonlinear even though the underlying PDE is linear.

In this paper we propose a combined scheme that employs the DPG technique to a *linear relaxation* of the nonlinear problem and where the nonlinearity is added as a constraint. Specifically, we relax the nonlinear problem by introducing an additional variable which then has a nonlinear relation with the original variables. This nonlinear relation is dealt with outside the DPG framework which can therefore develop its full potential, e.g., for singularly perturbed problems. In fact, although here we consider continuous and strongly monotone operators, we claim that our technique is applicable to

*Fundings :This work has been supported by CONICYT-Chile through FONDECYT projects 3170170 and 1150056.

†Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Santiago, Chile. pierre.cantin@mat.uc.cl.

‡Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Santiago, Chile. nheuer@mat.uc.cl.

singularly perturbed strongly monotone operators. This, and extensions to more general nonlinear problems, is ongoing research. In contrast, we do not see an obvious extension of [5] or [6] to singularly perturbed nonlinear problems, except for introducing a linear relaxation as we propose.

In the context of the nonlinear DPG scheme from [6] we mentioned their extension to a nonlinear mixed form. In fact, in the linear case it is well known that there is a mixed form of the DPG scheme, and this is precisely the method proposed (for a specific model problem) by Cohen *et al.* [11]. As we will see, our scheme can also be viewed as an extension of this mixed form. To be specific at this point, let us consider a (linear) continuously invertible operator $A : U \to V'$ with Banach space U, Hilbert space V, and dual V'. A mixed (or saddle-point) form of $A\boldsymbol{u} = F$ is

$$\begin{pmatrix} 0 & A^* \\ A & -R \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} = \begin{pmatrix} 0 \\ F \end{pmatrix} \qquad \text{in} \quad U' \times V'$$

with solution $(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}, 0)$. Here, $R : V \to V'$ is the Riesz operator and $A^* : V \to U'$ the adjoint of A. The (practical) DPG method [23] can be seen as a conforming discretization of this saddle point problem. In the nonlinear case, our method is equivalent to replacing the operator A by a linear relaxation B and the zero block by a nonlinear operator C (and, of course, redefining spaces and variables). This yields an operator of the form

$$\begin{pmatrix} C & B^* \\ B & -R \end{pmatrix} : U \times V \to U' \times V', \tag{1}$$

whose stability relies (among other properties) on the boundedness below of B. In our case, the nonlinearity is outsourced to C so that all the (linear) DPG strategies, aiming precisely at the boundedness-below property, can be employed. In our analysis it will be necessary to weight the operator R, though in specific applications there are precise bounds for this weighting parameter. We stress the fact that, since both B and C are acting on the unknowns of interest (represented by $\boldsymbol{u}$), it is not necessary to consider the variable $\boldsymbol{v}$. Indeed, in the numerical scheme we will be dealing with the Schur complement of $-R$ only. Standard DPG feature is to using product (broken) spaces V so that the numerical inversion of a discretization of R can be done locally and is, thus, cheap.

In the linear context, similar ideas have been used to deal with boundary, transmission and contact conditions outside the DPG framework, [19, 20, 21]. In fact, our abstract framework includes the analysis of linear boundary and transmission problems presented in [19, 20] as special cases. Though, differently from before, we decompose (or extend) nonlinear operators and develop an analysis based on the saddle point structure. Furthermore, we present an analysis that includes the approximation of optimal test functions whereas in [19, 20], these functions were assumed to be known exactly.

In this paper we consider, as a model, an advection-diffusion problem with nonlinear strongly monotone diffusivity,

$$-\nabla\cdot(\boldsymbol{\lambda}(|\nabla u|)\nabla u + \boldsymbol{\beta} u) = f \ \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega \tag{2}$$

with connected Lipschitz domain $\Omega \subset \mathbb{R}^d$ and $d \geq 2$. Of course, there is extensive literature on the numerical analysis of advection-diffusion problems, going back at least to Ciarlet, Schultz and Varga [10] when considering monotone operators. We do not start to discuss all the options as there are too many and since, more importantly, we use this problem only as a model to illustrate our idea and to show its applicability.

Considering the model problem (2), our relaxed linearized problem will be

$$-\nabla\cdot(\boldsymbol{\rho} + \boldsymbol{\beta} u) = f \ \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$

with the nonlinear flux $\boldsymbol{\rho} = \boldsymbol{\rho}(\nabla u)$ as additional variable. The solution of this problem is not unique so that the associated operator, denoted by B, has a non-trivial kernel. The missing nonlinear closure relation $\boldsymbol{\rho} = \boldsymbol{\lambda}(|\nabla u|)\nabla u$ will be represented by the kernel of a nonlinear operator C.

The rest of this paper is organized as follows. In the next section we present an abstract framework for operators of the form (1). To the best of our knowledge, such kind a of operator has not been analyzed before, although the particular case of $R = 0$ has the typical structure of a mixed method for nonlinear problems. Under appropriate assumptions we prove its invertibility by using the Schur complement (Theorem 2.3 in §2.1). In §2.2 we present our discrete scheme in abstract form and prove its quasi-optimal convergence (Theorem 2.8). The remainder of the paper is devoted to applying the abstract framework to the model problem (2). In Section 3, we precisely define the model problem, state necessary assumptions, and introduce spaces and norms. The introduction of meshes and corresponding product spaces is necessary for the DPG approximation, i.e., to localize the calculation of optimal test functions. In Section 4 we develop a variational formulation of our model problem with resulting operator of the type (1). The relaxed linear part is developed in §4.1. Here, any well-posed variational formulation of the linear problem will do, but for illustration we focus on an ultra-weak variant. This is by no means mandatory and, indeed, different formulations are equivalent, cf. [7]. Though in complicated cases like singular perturbations ultra-weak formulations are easier to analyze (current state of the art) and give the option of direct access to field variables, cf. [16, 25]. The nonlinear closure relation is studied in §4.2 and afterwards, in §4.3, the combined variational formulation is presented and its well-posedness is proved (Theorem 4.7). With all these preparations at hand, the presentation of our *relaxed DPG scheme* for the model problem is brief and a proof of its quasi-optimal convergence (Theorem 5.1) is immediate. This is the contents of Section 5. In Section 6 we present a numerical realization of our relaxed DPG scheme for the model problem and report on results for the cases with and without advection.

To alleviate notation, the expression $|\cdot|$ is context-dependent and denotes either the Lebesgue measure of a set, the absolute value of a real number or the Euclidean norm of a vector. We use boldface letters for vector and tensor valued quantities. In the calculation of norms via duality, suprema are taken over non-zero elements without further notice.

2. Abstract nonlinear penalized mixed problem. In this section we present the abstract framework of our DPG scheme. We first discuss specific continuous formulations, as an operator system similar to a saddle point problem and its Schur complement. In the second part we present two discretizations. The first is a conforming scheme based on the Schur complement and amounts to a DPG method with exactly optimal test functions. The second discretization uses the operator system and amounts to approximating the optimal test functions. Under appropriate assumptions we prove the quasi-optimal convergence of both methods (Theorems 2.5 and 2.8).

2.1. Continuous setting. Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be two real Hilbert spaces with topological duals $\boldsymbol{U}'$ and $\boldsymbol{V}'$, respectively. We consider a bounded linear operator $B : \boldsymbol{U} \to \boldsymbol{V}'$, an isomorphism $R : \boldsymbol{V} \to \boldsymbol{V}'$, and a continuous nonlinear operator $C : \boldsymbol{U} \to \boldsymbol{U}'$. (Later, C will be assumed to be Lipschitz continuous and strongly monotone in a certain sense.) Then we define the block operator $\boldsymbol{T}_\kappa : \boldsymbol{U}\times\boldsymbol{V} \to \boldsymbol{U}'\times\boldsymbol{V}'$, with $\kappa > 0$, as

$$\boldsymbol{T}_\kappa = \begin{pmatrix} C & B^* \\ B & -\frac{1}{\kappa}R \end{pmatrix}.$$

Here, $B^* : \boldsymbol{V} \to \boldsymbol{U}'$ is the adjoint operator of B, i.e., $\langle B\boldsymbol{u}, \boldsymbol{v}\rangle = \langle B^*\boldsymbol{v}, \boldsymbol{u}\rangle$ for all $\boldsymbol{u} \in \boldsymbol{U}$ and $\boldsymbol{v} \in \boldsymbol{V}$. To alleviate the notation, $\langle\cdot,\cdot\rangle$ is used generically to denote the duality between two arbitrary dual spaces. Denoting $\mathcal{N}(B)$ and $\mathcal{R}(B)$ the kernel and the range of B, respectively, we define $P_B : \boldsymbol{U} \to \mathcal{N}(B)$ the

$\boldsymbol{U}$-orthogonal projector of $\boldsymbol{U}$ on $\mathcal{N}(B)$.

The following theorem provides sufficient conditions on the continuous operators B, C and R such that the operator $\boldsymbol{T}_\kappa$ is bijective, with Lipschitz continuous inverse.

PROPOSITION 2.1. *Additionally to the assumptions made above, assume that*
(i) R^{-1} *is coercive on* $\mathcal{R}(B)$ *with coercivity constant* $c_R^{-1} > 0$,
(ii) there exists $c_B > 0$ *such that* $\|\boldsymbol{u} - P_B(\boldsymbol{u})\|^2 \le c_B \|B\boldsymbol{u}\|^2_{\boldsymbol{V}'}$ *for all* $\boldsymbol{u} \in \boldsymbol{U}$,
(iii) there exist $c_U, c_V > 0$ *such that*

$$c_U \|P_B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\|^2_{\boldsymbol{U}} \le \langle C(\boldsymbol{u}_1) - C(\boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2\rangle + c_V \|B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\|^2_{\boldsymbol{V}'}$$

for all $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \boldsymbol{U}$.
Then, for all $\kappa \ge \kappa_0 := (c_V + c_U c_B)c_R$, $\boldsymbol{T}_\kappa$ *is bijective with Lipschitz continuous inverse. In particular,*

$$c_U \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2_{\boldsymbol{U}} \le \langle C(\boldsymbol{u}_1) - C(\boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2\rangle + \kappa \langle R^{-1}B(\boldsymbol{u}_1 - \boldsymbol{u}_2), B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\rangle \tag{3}$$

for all $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \boldsymbol{U}$.

Proof. Since R is bijective we can consider the Schur complement of $-\frac{1}{\kappa}R$ to obtain

$$\begin{pmatrix} C + \kappa B^* R^{-1} B & 0 \\ 0 & -\kappa^{-1} R \end{pmatrix} = \boldsymbol{S}_\kappa^\ddagger \boldsymbol{T}_\kappa \boldsymbol{S}_\kappa^\dagger. \tag{4}$$

Here, $\boldsymbol{S}_\kappa^\dagger : \boldsymbol{U}\times\boldsymbol{V} \to \boldsymbol{U}\times\boldsymbol{V}$ and $\boldsymbol{S}_\kappa^\ddagger : \boldsymbol{U}'\times\boldsymbol{V}' \to \boldsymbol{U}'\times\boldsymbol{V}'$ are the respective involutory operators (i.e., equal to their inverses) defined as

$$\boldsymbol{S}_\kappa^\dagger = \begin{pmatrix} \mathrm{Id}_{\boldsymbol{U}} & 0 \\ \kappa R^{-1} B & -\mathrm{Id}_{\boldsymbol{V}} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{S}_\kappa^\ddagger = \begin{pmatrix} \mathrm{Id}_{\boldsymbol{U}'} & \kappa B^* R^{-1} \\ 0 & -\mathrm{Id}_{\boldsymbol{V}'} \end{pmatrix}.$$

It then follows that $\boldsymbol{T}_\kappa$ is invertible if and only if the operator $\boldsymbol{D}_\kappa = C + \kappa B^* R^{-1} B : \boldsymbol{U} \to \boldsymbol{U}'$ is invertible. Since $B^* R^{-1} B$ defines a linear bounded operator we infer that $\boldsymbol{D}_\kappa$ is continuous. Let us now prove that $\boldsymbol{D}_\kappa$ is strongly monotone, specifically that (3) holds. Let $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \boldsymbol{U}$ and start observing that

$$\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2_{\boldsymbol{U}} = \|P_B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\|^2_{\boldsymbol{U}} + \|\boldsymbol{u}_1 - \boldsymbol{u}_2 - P_B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\|^2_{\boldsymbol{U}},$$

owing to the definition of P_B. By assumption (i) we can bound

$$\|B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\|^2_{\boldsymbol{V}'} \le c_R \langle R^{-1}B(\boldsymbol{u}_1 - \boldsymbol{u}_2), B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\rangle.$$

Combining this bound with assumptions (ii) and (iii), it follows that

$$c_U \|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2_{\boldsymbol{U}} \le \langle C(\boldsymbol{u}_1) - C(\boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2\rangle + (c_V + c_B c_U)c_R \langle R^{-1}B(\boldsymbol{u}_1 - \boldsymbol{u}_2), B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\rangle.$$

As a result, $\boldsymbol{D}_\kappa$ is continuous and strongly monotone for all $\kappa \ge \kappa_0 := (c_V + c_B c_U)c_R$, i.e., $\boldsymbol{D}_\kappa$ is invertible for all $\kappa \ge \kappa_0$ with Lipschitz continuous inverse, see, e.g., [26, Chap. 2]) ∎

Remark 2.2. (i) In the case of a linear operator C, $-\frac{1}{\kappa}R$ plays the role of a regularizing operator needed for the bijectivity of $\boldsymbol{T}_\kappa$. As in the linear case, here the invertibility of $\boldsymbol{T}_\kappa$ does not require the surjectivity of B, but only that the range $\mathcal{R}(B)$ is closed in $\boldsymbol{V}'$ (see, e.g., [22, Remark 4.3]).

(ii) In Proposition 2.1, if we assume in addition that C is monotone, i.e., $\langle C(\boldsymbol{u}_1) - C(\boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2\rangle \ge 0$ for all $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \boldsymbol{U}$, then we infer from (3) that $\boldsymbol{T}_\kappa$ is bijective with Lipschitz continuous inverse for all $\kappa \ge 1$.

Now, given $F \in \boldsymbol{V}'$ and $G \in \boldsymbol{U}'$, we consider two variational formulations, of penalized mixed form

$$(\boldsymbol{u},\boldsymbol{v}) \in \boldsymbol{U}\times\boldsymbol{V}: \qquad \boldsymbol{T}_\kappa(\boldsymbol{u},\boldsymbol{v}) = (G,F) \quad \text{in } \boldsymbol{U}'\times\boldsymbol{V}', \tag{5}$$

and the reduced Schur variant

$$\boldsymbol{u} \in \boldsymbol{U}: \qquad \langle \boldsymbol{D}_\kappa \boldsymbol{u}, \boldsymbol{w}\rangle = \kappa\langle R^{-1}F, B\boldsymbol{w}\rangle + \langle G, \boldsymbol{w}\rangle \quad \forall \boldsymbol{w} \in \boldsymbol{U}, \tag{6}$$

with

$$\langle \boldsymbol{D}_\kappa \boldsymbol{u}, \boldsymbol{w}\rangle := \kappa\langle R^{-1}B\boldsymbol{u}, B\boldsymbol{w}\rangle + \langle C(\boldsymbol{u}), \boldsymbol{w}\rangle \qquad (\boldsymbol{u},\boldsymbol{w} \in \boldsymbol{U}).$$

The following result is a consequence of the Schur factorization (4) and Proposition 2.1.

THEOREM 2.3. *Let the assumptions of Proposition 2.1 hold true. Then, for all $\kappa \geq \kappa_0$ with κ_0 as before, problems (5) and (6) are uniquely solvable and equivalent. Specifically, if $(\boldsymbol{u},\boldsymbol{v}) \in \boldsymbol{U}\times\boldsymbol{V}$ solves (5), then $\boldsymbol{u}$ solves (6). Conversely, if $\boldsymbol{u} \in \boldsymbol{U}$ solves (6), then $(\boldsymbol{u}, \kappa R^{-1}(B\boldsymbol{u} - F)) \in \boldsymbol{U}\times\boldsymbol{V}$ solves (5).*

Under additional assumptions on the operators B and C, we have a strong characterization of the solution of (6).

PROPOSITION 2.4. *Assume that*
(i) $\langle C(\boldsymbol{u}) - G, \boldsymbol{v}\rangle = 0$ for all $\boldsymbol{v} \in \mathcal{N}(B)$ implies $C(\boldsymbol{u}) = G$ in $\boldsymbol{U}'$,
(ii) $B: \boldsymbol{U} \to \boldsymbol{V}'$ is surjective.
Then, $\boldsymbol{u} \in \boldsymbol{U}$ solves (6) if and only if $C(\boldsymbol{u}) = G$ in $\boldsymbol{U}'$ and $B\boldsymbol{u} = F$ in $\boldsymbol{V}'$.

Proof. Let $\boldsymbol{u} \in \boldsymbol{U}$ be solving (6). It follows that $\langle C(\boldsymbol{u}) - G, \boldsymbol{w}\rangle = 0$ for all $\boldsymbol{w} \in \mathcal{N}(B)$. By assumption (i) we infer that $C(\boldsymbol{u}) = G$ in $\boldsymbol{U}'$. Therefore, $\boldsymbol{u}$ satisfies $\langle R^{-1}B\boldsymbol{u}, B\boldsymbol{w}\rangle = \langle R^{-1}F, B\boldsymbol{w}\rangle$ for all $\boldsymbol{w} \in \boldsymbol{U}$, i.e., $B^*R^{-1}(B\boldsymbol{u} - F) = 0$ in $\boldsymbol{U}'$. Using assumption (ii) and recalling that $R: \boldsymbol{V} \to \boldsymbol{V}'$ is an isomorphism, we conclude that $B^*R^{-1}: \boldsymbol{V}' \to \boldsymbol{U}'$ is injective, so that $B\boldsymbol{u} = F$ in $\boldsymbol{V}'$. The other direction is immediate. ∎

2.2. Discretization. We analyze approximations of the continuous problems (5) and (6). At the continuous level, these two problems are equivalent by Theorem 2.3 so that considering $\boldsymbol{T}_\kappa$ or $\boldsymbol{D}_\kappa$ with their respective right-hand side yields the same problem. However, considering one operator or the other at the discrete level is no longer equivalent.

For an index parameter $h > 0$, let $\boldsymbol{U}_h$ and $\boldsymbol{V}_h$ be two (families of) finite-dimensional spaces such that $\boldsymbol{U}_h \subset \boldsymbol{U}$ and $\boldsymbol{V}_h \subset \boldsymbol{V}$. Of course, later h will be a mesh parameter. We denote the canonical injection maps by $\boldsymbol{i}_h: \boldsymbol{U}_h \to \boldsymbol{U}$ and $\boldsymbol{j}_h: \boldsymbol{V}_h \to \boldsymbol{V}$, with $\boldsymbol{i}_h^*: \boldsymbol{U}' \to \boldsymbol{U}_h'$ and $\boldsymbol{j}_h^*: \boldsymbol{V}' \to \boldsymbol{V}_h'$ the respective adjoints. The discrete spaces $\boldsymbol{U}_h$ and $\boldsymbol{V}_h$ are provided with the induced norms $\|\cdot\|_{\boldsymbol{U}_h} := \|\boldsymbol{i}_h \cdot\|_{\boldsymbol{U}}$ and $\|\cdot\|_{\boldsymbol{V}_h} := \|\boldsymbol{j}_h \cdot\|_{\boldsymbol{V}}$.

2.2.1. Semi-discrete scheme. One possibility is to discretize the operator $\boldsymbol{D}_\kappa$ in the standard way, that is, considering $\boldsymbol{D}^\star_{\kappa,h}: \boldsymbol{U}_h \to \boldsymbol{U}_h'$ defined as

$$\boldsymbol{D}^\star_{\kappa,h} = \boldsymbol{i}_h^* \boldsymbol{D}_\kappa \boldsymbol{i}_h, \quad \text{with} \quad \boldsymbol{D}_\kappa = C + \kappa B^*R^{-1}B.$$

This discretization still requires to calculate R^{-1}, which is not feasible in practice. In DPG discretizations, R is the Riesz operator $R_{\boldsymbol{V}}: \boldsymbol{V} \to \boldsymbol{V}'$ and such a semi-discrete scheme is sometimes called *ideal* DPG method. It is distinguished from the *practical* variant which includes a discretization of $R_{\boldsymbol{V}}^{-1}$, cf. [23].

The operator $\boldsymbol{D}^\star_{\kappa,h}$ induces the problem

$$\boldsymbol{u}_h \in \boldsymbol{U}_h \ : \quad \langle \boldsymbol{D}^\star_{\kappa,h}\boldsymbol{u}_h, \boldsymbol{w}_h\rangle = \kappa\langle R^{-1}F, B\boldsymbol{i}_h\boldsymbol{w}_h\rangle + \langle G, \boldsymbol{i}_h\boldsymbol{w}_h\rangle \quad \forall \boldsymbol{w}_h \in \boldsymbol{U}_h. \tag{7}$$

THEOREM 2.5. *Assume that the assumptions from Proposition 2.1 hold true with constants κ_0 and c_U specified there. Then, for all $\kappa \geq \kappa_0$, $\boldsymbol{D}^\star_{\kappa,h}$ is invertible with uniformly Lipschitz continuous inverse, and problem (7) is well posed. In addition, assuming that C is Lipschitz continuous with constant c_{Lip}, we have the quasi-optimal error estimate*

$$\|\boldsymbol{u} - \boldsymbol{i}_h(\boldsymbol{u}_h)\|_{\boldsymbol{U}} \leq \big(1 + c_U^{-1}\big(c_{\mathrm{Lip}} + \kappa\|B^*R^{-1}B\|_{\mathcal{L}(\boldsymbol{U},\boldsymbol{U}')}\big)\big) \inf_{\boldsymbol{w}_h\in\boldsymbol{U}_h} \|\boldsymbol{u} - \boldsymbol{i}_h(\boldsymbol{w}_h)\|_{\boldsymbol{U}}.$$

Here, $\boldsymbol{u} \in \boldsymbol{U}$ and $\boldsymbol{u}_h \in \boldsymbol{U}_h$ are the unique solutions of (6) and (7), respectively.

Proof. The discrete operator $\boldsymbol{D}^\star_{\kappa,h}$ defines a conforming approximation of the continuous problem (6). Therefore, its uniform Lipschitz continuous invertibility follows from the Lipschitz continuous invertibility of $\boldsymbol{D}_\kappa$, cf. Theorem 2.3. Again, by the conformity of the approximation, the a priori error estimate follows by standard arguments using the monotonicity (3). ∎

2.2.2. Fully discrete scheme. In order to avoid the inversion of the operator R, present in (7), we discretize, instead of the Schur complement $\boldsymbol{D}_\kappa$, the full operator $\boldsymbol{T}_\kappa$ as $\boldsymbol{T}_{\kappa,h} : \boldsymbol{U}_h\times\boldsymbol{V}_h \to \boldsymbol{U}_h'\times\boldsymbol{V}_h'$ defined by

$$\boldsymbol{T}_{\kappa,h} = \begin{pmatrix} C_h & B_h^* \\ B_h & -\frac{1}{\kappa}R_h \end{pmatrix} = \begin{pmatrix} \boldsymbol{i}_h^* & 0 \\ 0 & \boldsymbol{j}_h^* \end{pmatrix} \boldsymbol{T}_\kappa \begin{pmatrix} \boldsymbol{i}_h & 0 \\ 0 & \boldsymbol{j}_h \end{pmatrix}. \tag{8}$$

Applying the Schur factorization to this discrete operator we obtain the fully discrete problem

$$\boldsymbol{u}_h \in \boldsymbol{U}_h \ : \quad \langle \boldsymbol{D}_{\kappa,h}\boldsymbol{u}_h, \boldsymbol{w}_h\rangle = \kappa\langle R_h^{-1}\boldsymbol{j}_h^*F, B_h\boldsymbol{w}_h\rangle + \langle \boldsymbol{i}_h^*G, \boldsymbol{w}_h\rangle \quad \forall \boldsymbol{w}_h \in \boldsymbol{U}_h, \tag{9}$$

with

$$\langle \boldsymbol{D}_{\kappa,h}\boldsymbol{u}_h, \boldsymbol{w}_h\rangle := \langle C_h(\boldsymbol{u}_h), \boldsymbol{w}_h\rangle + \kappa\langle R_h^{-1}B_h\boldsymbol{u}_h, B_h\boldsymbol{w}_h\rangle. \tag{10}$$

The well-posedness of (9) follows similarly as in Proposition 2.1, by using the existence of a Fortin operator $\Pi : \boldsymbol{V} \to \boldsymbol{V}_h$ satisfying, uniformly in h,

$$\langle B\boldsymbol{i}_h\boldsymbol{u}_h, \boldsymbol{v} - \boldsymbol{j}_h\Pi\boldsymbol{v}_h\rangle = 0 \qquad \forall \boldsymbol{u}_h \in \boldsymbol{U}_h,\ \boldsymbol{v} \in \boldsymbol{V}, \tag{11a}$$

$$\exists c_\Pi > 0, \quad \|\boldsymbol{j}_h\Pi\boldsymbol{v}\|_{\boldsymbol{V}}^2 \leq c_\Pi\|\boldsymbol{v}\|_{\boldsymbol{V}}^2 \quad \forall \boldsymbol{v} \in \boldsymbol{V}. \tag{11b}$$

In the context of DPG methods, Gopalakrishnan and Qiu [23] have employed such an operator to analyze the approximation of optimal test functions. This is precisely our motivation. Let us recall this result, cf. [23, Proof of Theorem 2.1].

LEMMA 2.6. *Assume that (11) holds. Then, $\|B\boldsymbol{i}_h\boldsymbol{u}_h\|_{\boldsymbol{V}'}^2 \leq c_\Pi\|B_h\boldsymbol{u}_h\|_{\boldsymbol{V}_h'}^2$ for all $\boldsymbol{u}_h \in \boldsymbol{U}_h$.*

The following proposition extends the statements of Proposition 2.1 to the discrete level.

PROPOSITION 2.7. *Assume that,*
(i') R_h^{-1} is coercive on $\mathcal{R}(B_h)$ with coercivity constant $c_{R_h}^{-1} > 0$,
(ii) there exists $c_B > 0$ such that $\|\boldsymbol{u} - P_B(\boldsymbol{u})\|^2 \leq c_B\|B\boldsymbol{u}\|_{\boldsymbol{V}'}^2$ for all $\boldsymbol{u} \in \boldsymbol{U}$,
(iii) there exist $c_U, c_V > 0$ such that

$$c_U\|P_B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\|_{\boldsymbol{U}}^2 \leq \langle C(\boldsymbol{u}_1) - C(\boldsymbol{u}_2), \boldsymbol{u}_1 - \boldsymbol{u}_2\rangle + c_V\|B(\boldsymbol{u}_1 - \boldsymbol{u}_2)\|_{\boldsymbol{V}'}^2$$

for all $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \boldsymbol{U}$,

(iv) there exists a Fortin operator $\Pi : \boldsymbol{V} \to \boldsymbol{V}_h$ *satisfying* (11).
Then, for all $\kappa \geq \kappa_{0,h} := (c_V + c_B c_U) c_\Pi c_{R_h}$, $\boldsymbol{T}_{\kappa,h}$ *is invertible with Lipschitz continuous inverse. In particular,*

$$c_U \|\boldsymbol{u}_h - \boldsymbol{v}_h\|_{\boldsymbol{U}_h}^2 \leq \langle C_h(\boldsymbol{u}_h) - C_h(\boldsymbol{v}_h), \boldsymbol{u}_h - \boldsymbol{v}_h\rangle + \kappa \langle R_h^{-1} B_h(\boldsymbol{u}_h - \boldsymbol{v}_h), B_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\rangle \tag{12}$$

for all $\boldsymbol{u}_h, \boldsymbol{v}_h \in \boldsymbol{U}_h$.

Proof. We follow the same route as in the proof of Proposition 2.1. Specifically, the discrete operator $\boldsymbol{T}_{\kappa,h}$ is invertible if and only if $\boldsymbol{D}_{\kappa,h}$ is invertible. We prove that $\boldsymbol{D}_{\kappa,h}$ is continuous and strongly monotone.

Let $\boldsymbol{u}_h, \boldsymbol{v}_h \in \boldsymbol{U}_h$. By assumptions (ii), (iv) and Lemma 2.6 we have

$$\|\boldsymbol{i}_h(\boldsymbol{u}_h - \boldsymbol{v}_h) - P_B \boldsymbol{i}_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\|_{\boldsymbol{U}}^2 \leq c_B \|B \boldsymbol{i}_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\|_{\boldsymbol{V}'}^2 \leq c_B c_\Pi \|B_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\|_{\boldsymbol{V}_h'}^2,$$

so that, using assumption (i'), it follows that

$$\|\boldsymbol{i}_h(\boldsymbol{u}_h - \boldsymbol{v}_h) - P_B \boldsymbol{i}_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\|_{\boldsymbol{U}}^2 \leq c_B c_\Pi c_{R_h} \langle R_h^{-1} B_h(\boldsymbol{u}_h - \boldsymbol{v}_h), B_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\rangle.$$

Owing to assumption (iii) and the definition of C_h, we also have

$$c_U \|P_B \boldsymbol{i}_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\|_{\boldsymbol{U}}^2 \leq \langle C_h(\boldsymbol{u}_h) - C_h(\boldsymbol{v}_h), \boldsymbol{u}_h - \boldsymbol{v}_h\rangle + c_V \|B \boldsymbol{i}_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\|_{\boldsymbol{V}'}^2.$$

Again applying Lemma 2.6 and combining the two last inequalities, it follows that

$$\begin{aligned} c_U \|\boldsymbol{u}_h - \boldsymbol{v}_h\|_{\boldsymbol{U}_h}^2 \leq{}& \langle C_h(\boldsymbol{u}_h) - C_h(\boldsymbol{v}_h), \boldsymbol{u}_h - \boldsymbol{v}_h\rangle \\ &+ (c_V + c_B c_U) c_\Pi c_{R_h} \langle B_h(\boldsymbol{u}_h - \boldsymbol{v}_h), R_h^{-1} B_h(\boldsymbol{u}_h - \boldsymbol{v}_h)\rangle. \end{aligned}$$

In particular, (12) holds for $\kappa \geq \kappa_{0,h} := (c_V + c_B c_U) c_\Pi c_{R_h}$. We conclude that $\boldsymbol{D}_{\kappa,h}$ is continuous and strongly monotone for all $\kappa \geq \kappa_{0,h}$ with Lipschitz continuous inverse. ∎

THEOREM 2.8. *Assume that the assumptions from Propositions 2.1, 2.7 hold true, and that* $\kappa \geq \max(\kappa_0; \kappa_{0,h})$, *with the constants* $\kappa_0, \kappa_{0,h}$ *and* c_U *from before. Then, problems* (6) *and* (9) *are well posed. In addition, assuming that the assumptions from Proposition 2.4 hold true and that* C *is Lipschitz continuous with Lipschitz constant* c_{Lip}, *we have the quasi-optimal error estimate*

$$\|\boldsymbol{u} - \boldsymbol{i}_h(\boldsymbol{u}_h)\|_{\boldsymbol{U}} \leq \big(1 + c_U^{-1}\big(c_{\mathrm{Lip}} + \kappa \|B^* \boldsymbol{j}_h R_h^{-1} \boldsymbol{j}_h^* B\|_{\mathcal{L}(\boldsymbol{U}, \boldsymbol{U}')}\big)\big) \inf_{\boldsymbol{w}_h \in \boldsymbol{U}_h} \|\boldsymbol{u} - \boldsymbol{i}_h(\boldsymbol{w}_h)\|_{\boldsymbol{U}}.$$

Here, $\boldsymbol{u} \in \boldsymbol{U}$ *and* $\boldsymbol{u}_h \in \boldsymbol{U}_h$ *are the unique solutions of* (6) *and* (9), *respectively.*

Proof. Let $\boldsymbol{w}_h \in \boldsymbol{U}_h$ and denote $\boldsymbol{\xi}_h = \boldsymbol{u}_h - \boldsymbol{w}_h$. By Proposition 2.7 we have

$$c_U \|\boldsymbol{\xi}_h\|_{\boldsymbol{U}_h}^2 \leq \langle C_h(\boldsymbol{u}_h) - C_h(\boldsymbol{w}_h), \boldsymbol{\xi}_h\rangle + \kappa \langle R_h^{-1} B_h(\boldsymbol{u}_h - \boldsymbol{w}_h), B_h \boldsymbol{\xi}_h\rangle$$

for all $\kappa \geq \kappa_{0,h}$. In addition, since $B\boldsymbol{u} = F$ and $C(\boldsymbol{u}) = G$ according to Proposition 2.4, the relations

$$\langle R_h^{-1} \boldsymbol{j}_h^* B \boldsymbol{u}, B_h \boldsymbol{\xi}_h\rangle = \langle R_h^{-1} \boldsymbol{j}_h^* F, B_h \boldsymbol{\xi}_h\rangle \;\text{ and }\; \langle \boldsymbol{i}_h^* C(\boldsymbol{u}), \boldsymbol{\xi}_h\rangle = \langle \boldsymbol{i}_h^* G, \boldsymbol{\xi}_h\rangle.$$

hold. Hence, we conclude that

$$\kappa \langle R_h^{-1} \boldsymbol{j}_h^* B(\boldsymbol{u} - \boldsymbol{i}_h \boldsymbol{u}_h), B_h \boldsymbol{\xi}_h\rangle + \langle C(\boldsymbol{u}) - C(\boldsymbol{i}_h \boldsymbol{u}_h), \boldsymbol{j}_h \boldsymbol{\xi}_h\rangle = 0,$$

yielding

$$c_U\|\boldsymbol{\xi}_h\|_{U_h}^2 \le \langle C(\boldsymbol{u}) - C(\boldsymbol{i}_h\boldsymbol{w}_h), \boldsymbol{i}_h\boldsymbol{\xi}_h\rangle + \kappa\langle B^*\boldsymbol{j}_h R_h^{-1}\boldsymbol{j}_h^* B(\boldsymbol{u} - \boldsymbol{i}_h\boldsymbol{w}_h), \boldsymbol{i}_h\boldsymbol{\xi}_h\rangle.$$

The triangle inequality and the Lipschitz continuity of C then prove the statement. ∎

3. Nonlinear model problem and functional setting. In the remainder of this paper we show how our abstract DPG framework applies to an advection-diffusion model problem with nonlinear diffusion. In this section we specify the model problem and consider its continuous formulation.

Given a source term f, let us consider $u:\ \Omega \to \mathbb{R}$, with $\Omega \subset \mathbb{R}^d$ ($d \ge 2$) a connected Lipschitz domain, the solution of

$$-\nabla\cdot\big(\boldsymbol{\lambda}(\boldsymbol{x}, |\nabla u(\boldsymbol{x})|)\nabla u(\boldsymbol{x}) + \boldsymbol{\beta}(\boldsymbol{x})u(\boldsymbol{x})\big) = f(\boldsymbol{x}) \quad \text{for a.a. } \boldsymbol{x} \in \Omega, \tag{13a}$$

$$u(\boldsymbol{x}) = 0 \qquad \text{for a.a. } \boldsymbol{x} \in \partial\Omega. \tag{13b}$$

Here, $\boldsymbol{\beta}$ denotes an $\mathbb{R}^d$-valued advection field and $\boldsymbol{\lambda}$ an $\mathbb{R}^{d\times d}$-valued diffusion tensor. By $\partial\Omega$ we denote the boundary of Ω, with outwardly oriented unit normal vector $\boldsymbol{n}$. For simplicity, we write $\boldsymbol{\lambda}(|\nabla u|)$ for $\boldsymbol{\lambda}(\boldsymbol{x}, |\nabla u(\boldsymbol{x})|)$ for almost all $\boldsymbol{x} \in \Omega$.

Owing to the theory of continuous strongly monotone operators (see, e.g., [26, Chap. 2]), this model problem admits a unique solution $u \in H^1_0(\Omega)$ for any source term $f \in L^2(\Omega)$ if the physical parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ satisfy the following assumptions:

- $\boldsymbol{\beta}$ is Lipschitz continuous on Ω and satisfies

$$\operatorname{ess\,inf}_\Omega\ (-\nabla\cdot\boldsymbol{\beta}) \ge 0. \tag{14a}$$

- There exist constants $0 < \lambda_\flat \le \lambda_\sharp$ such that, for all $\boldsymbol{\sigma}, \boldsymbol{\theta} \in \boldsymbol{L}^2(\Omega)$,

$$\lambda_\flat\|\boldsymbol{\sigma} - \boldsymbol{\theta}\|^2_{\boldsymbol{L}^2(\Omega)} \le (\boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma} - \boldsymbol{\lambda}(|\boldsymbol{\theta}|)\boldsymbol{\theta}, \boldsymbol{\sigma} - \boldsymbol{\theta})_{\boldsymbol{L}^2(\Omega)}, \tag{14b}$$

$$\lambda_\sharp\|\boldsymbol{\sigma} - \boldsymbol{\theta}\|_{\boldsymbol{L}^2(\Omega)} \ge \|\boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma} - \boldsymbol{\lambda}(|\boldsymbol{\theta}|)\boldsymbol{\theta}\|_{\boldsymbol{L}^2(\Omega)}. \tag{14c}$$

Throughout the remainder of this paper we assume that all these conditions are satisfied, specifically that $f \in L^2(\Omega)$.

3.1. Standard Sobolev spaces and Péclet number. Let $L^2(\Omega)$ and $\boldsymbol{L}^2(\Omega)$ be the standard Lebesgue spaces collecting $\mathbb{R}$-valued and $\mathbb{R}^d$-valued functions, respectively, satisfying

$$(v, v)_\Omega = \|v\|^2_{L^2(\Omega)} = \int_\Omega |v|^2 < +\infty \text{ and } (\boldsymbol{v}, \boldsymbol{v})_\Omega = \|\boldsymbol{v}\|^2_{\boldsymbol{L}^2(\Omega)} = \int_\Omega |\boldsymbol{v}|^2 < +\infty.$$

We denote by $H^1(\Omega)$ and $\boldsymbol{H}(\mathrm{div};\Omega)$ the classical Sobolev spaces equipped with the scaled inner products

$$(v, w)_{H^1(\Omega)} := (v, w)_\Omega + \ell_\Omega^2(\nabla v, \nabla w)_\Omega \qquad \forall v, w \in H^1(\Omega),$$

$$(\boldsymbol{\tau}, \boldsymbol{\eta})_{\boldsymbol{H}(\mathrm{div};\Omega)} := (\boldsymbol{\tau}, \boldsymbol{\eta})_\Omega + \ell_\Omega^2(\nabla\cdot\boldsymbol{\tau}, \nabla\cdot\boldsymbol{\eta})_\Omega \quad \forall \boldsymbol{\tau}, \boldsymbol{\eta} \in \boldsymbol{H}(\mathrm{div};\Omega).$$

We also denote by $H^1_0(\Omega)$ the closure of the space collecting infinitely differentiable functions with compact support in Ω with the norm $\|\cdot\|_{H^1(\Omega)}$.

The characteristic length $\ell_\Omega > 0$ is introduced so that the above inner products are dimensionally coherent. Its definition is arbitrary, but fixed once and for all. To avoid the proliferation of constants, the reference length ℓ_Ω is chosen such that

$$\|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(\Omega)}\ell_\Omega\lambda_\sharp^{-1} = 1. \tag{15}$$

This means that the global Péclet number is of order 1, i.e., the magnitude of the advective and diffusive effects are comparable.

3.2. Mesh partition and product spaces. Let Ω_h be a non-overlapping partition of Ω composed open elements $T \in \Omega_h$ with Lipschitz boundary ∂T and outwardly oriented by $\boldsymbol{n}_T$. Let $H^1(\Omega_h)$ and $\boldsymbol{H}(\mathrm{div};\Omega_h)$ be the product or "broken" Sobolev spaces equipped with inner products

$$(v,w)_{H^1(\Omega_h)} := (v,w)_{\Omega_h} + \ell_\Omega^2(\nabla v, \nabla w)_{\Omega_h} \qquad \forall v, w \in H^1(\Omega_h),$$
$$(\boldsymbol{\tau},\boldsymbol{\eta})_{\boldsymbol{H}(\mathrm{div};\Omega_h)} := (\boldsymbol{\tau},\boldsymbol{\eta})_{\Omega_h} + \ell_\Omega^2(\nabla\cdot\boldsymbol{\tau}, \nabla\cdot\boldsymbol{\eta})_{\Omega_h} \quad \forall \boldsymbol{\tau}, \boldsymbol{\eta} \in \boldsymbol{H}(\mathrm{div};\Omega_h).$$

Here, $(\cdot,\cdot)_{\Omega_h} = \sum_{T\in\Omega_h}(\cdot,\cdot)_T$ denotes the element-wise L^2-inner product, that is, appearing differential operators are taken in a piecewise form. For all $T \in \Omega_h$ we denote by $H^{1/2}(\partial T)$ and $H^{-1/2}(\partial T)$ the trace spaces of $H^1(T)$ and $\boldsymbol{H}(\mathrm{div};T)$, respectively. They are dual to each other. Traces on the mesh skeleton $\partial\Omega_h$ are defined with the trace maps $\gamma : H^1(\Omega_h) \to \bigtimes_{T\in\Omega_h} H^{1/2}(\partial T)$ and $\gamma_n : \boldsymbol{H}(\mathrm{div};\Omega_h) \to \bigtimes_{T\in\Omega_h} H^{-1/2}(\partial T)$, defined as

$$\gamma(u)_{|\partial T} = u_{|\partial T} \qquad \forall T \in \Omega_h \quad \forall u \in H^1(\Omega_h),$$
$$\gamma_n(\boldsymbol{\rho})_{|\partial T} = (\boldsymbol{\rho}\cdot\boldsymbol{n}_T)_{|\partial T} \quad \forall T \in \Omega_h \quad \forall \boldsymbol{\rho} \in \boldsymbol{H}(\mathrm{div};\Omega_h).$$

The duality product between $\bigtimes_{T\in\Omega_h} H^{-1/2}(\partial T)$ and $\bigtimes_{T\in\Omega_h} H^{1/2}(\partial T)$ is denoted by $\langle\cdot,\cdot\rangle_{\partial\Omega_h} = \sum_{T\in\Omega_h}\langle\cdot,\cdot\rangle_{\partial T}$ with duality $\langle\cdot,\cdot\rangle_{\partial T}$ between $H^{-1/2}(\partial T)$ and $H^{1/2}(\partial T)$ $(T \in \Omega_h)$.

We also introduce the trace spaces

$$H^{1/2}_{00}(\partial\Omega_h) := \Big\{\hat{u} \in \bigtimes_{T\in\Omega_h} H^{1/2}(\partial T) \,|\, \exists w \in H^1_0(\Omega),\ \hat{u} = \gamma(w)\Big\},$$
$$H^{-1/2}(\partial\Omega_h) := \Big\{\hat{\rho} \in \bigtimes_{T\in\Omega_h} H^{-1/2}(\partial T) \,|\, \exists \boldsymbol{\rho} \in \boldsymbol{H}(\mathrm{div};\Omega),\ \hat{\rho} = \gamma_n(\boldsymbol{\rho})\Big\},$$

equipped with their respective quotient norms,

$$\|\hat{v}\|_{H^{1/2}_{00}(\partial\Omega_h)} := \inf_{w\in H^1_0(\Omega)} \big\{\|w\|_{H^1(\Omega)};\ \gamma(w) = \hat{v}\big\} \quad \forall \hat{v} \in H^{1/2}_{00}(\partial\Omega_h), \tag{16a}$$

$$\|\hat{\rho}\|_{H^{-1/2}(\partial\Omega_h)} := \inf_{\boldsymbol{\rho}\in\boldsymbol{H}(\mathrm{div};\Omega)} \big\{\|\boldsymbol{\rho}\|_{\boldsymbol{H}(\mathrm{div};\Omega)};\ \gamma_n(\boldsymbol{\rho}) = \hat{\rho}\big\} \quad \forall \hat{\rho} \in H^{-1/2}(\partial\Omega_h). \tag{16b}$$

Finally, we close this section by recalling the Poincaré-Steklov inequality in the product space $H^1(\Omega_h)$, cf. [2]. The proof is given for completeness, here including the length scale parameter ℓ_Ω.

LEMMA 3.1. *We have*

$$c_{PS}\ell_\Omega^{-2}\|v\|^2_{H^1(\Omega_h)} \le \|\nabla v\|^2_{L^2(\Omega_h)} + \left(\sup_{\hat{\tau}\in H^{-1/2}(\partial\Omega_h)} \frac{\langle\hat{\tau},\gamma(v)\rangle_{\partial\Omega_h}}{\|\hat{\tau}\|_{H^{-1/2}(\partial\Omega_h)}}\right)^2 \quad \forall v \in H^1(\Omega_h),$$

with $c_{PS}^{-1} = 2(1 + c_{PS,0})$ *and* $c_{PS,0} > 0$ *the Poincaré-Steklov constant in* $H^1_0(\Omega)$ *satisfying* $\|\xi\|^2_{L^2(\Omega)} \le c_{PS,0}\ell_\Omega^2\|\nabla\xi\|^2_{\boldsymbol{L}^2(\Omega)}$ *for all* $\xi \in H^1_0(\Omega)$.

Proof. Let $v \in H^1(\Omega_h)$ and let $\xi \in H^1_0(\Omega)$ such that $-\ell_\Omega^2\Delta\xi = v$. It holds

$$\begin{aligned}\|v\|^2_{L^2(\Omega)} &= \ell_\Omega^2\,(\nabla v, \nabla\xi)_{\Omega_h} - \ell_\Omega^2\langle\gamma_n(\nabla\xi),\gamma(v)\rangle_{\partial\Omega_h}\\ &\le \ell_\Omega^2\|\nabla v\|_{\boldsymbol{L}^2(\Omega_h)}\|\nabla\xi\|_{\boldsymbol{L}^2(\Omega)} + \ell_\Omega^2\|\gamma_n(\nabla\xi)\|_{H^{-1/2}(\partial\Omega_h)} \sup_{\hat{\tau}\in H^{-1/2}(\partial\Omega_h)} \frac{\langle\hat{\tau},\gamma(v)\rangle_{\partial\Omega_h}}{\|\hat{\tau}\|_{H^{-1/2}(\partial\Omega_h)}}.\end{aligned}$$

Furthermore, $\ell_\Omega^2\|\nabla\xi\|^2_{\boldsymbol{L}^2(\Omega)} \leq c_{PS,0}\|v\|^2_{L^2(\Omega)}$ by the standard Poincaré-Steklov inequality in $H^1_0(\Omega)$. Hence, observing that $\|\gamma_n(\nabla\xi)\|^2_{H^{-1/2}(\partial\Omega_h)} \leq (1+c_{PS,0})\ell_\Omega^{-2}\|v\|^2_{L^2(\Omega)}$, the statement follows. ∎

4. Penalized variational formulation of the model problem. In this section we apply the results of Section 2 to devise and approximate a penalized formulation of (13). The objective of our non-standard formulation is to separate the linear and the nonlinear parts of our problem, namely rewriting formally (13a) as

$$-\nabla\cdot(\boldsymbol{\rho}+\boldsymbol{\beta}u) = f, \quad \boldsymbol{\sigma} = \nabla u \ \text{ and } \ \boldsymbol{\rho} = \boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma}.$$

In the following, the operator B stands for the representation of the ultra-weak formulation of the first two linear equations, and the nonlinear operator C is used to enforce the nonlinear closure relation $\boldsymbol{\rho} = \boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma}$.

4.1. Ultra-weak formulation of the linear part. We start by specifying a variational formulation of the linear part of the model problem. In this case we select an ultra-weak variant. In some cases like singularly perturbed problems the ultra-weak form has its advantages but for our model problem this selection is not essential.

We consider the following linear problem. Find $u \in L^2(\Omega)$, $\boldsymbol{\sigma} \in \boldsymbol{L}^2(\Omega)$, $\boldsymbol{\rho} \in \boldsymbol{L}^2(\Omega)$, $\hat{u} \in H^{1/2}_{00}(\partial\Omega_h)$, $\hat{\rho} \in H^{-1/2}(\partial\Omega_h)$ such that

$$(u, \boldsymbol{\beta}\cdot\nabla v)_{\Omega_h} + (\boldsymbol{\rho}, \nabla v)_{\Omega_h} - \langle\hat{\rho}, \gamma(v)\rangle_{\partial\Omega_h} = (f, v)_\Omega \quad \forall v \in H^1(\Omega_h), \tag{17a}$$

$$(\boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega + (u, \nabla\cdot\boldsymbol{\tau})_{\Omega_h} - \langle\gamma_n(\boldsymbol{\tau}), \hat{u}\rangle_{\partial\Omega_h} = 0 \qquad \forall \boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div};\Omega_h), \tag{17b}$$

and denote the spaces

$$\boldsymbol{U} := L^2(\Omega)\times\boldsymbol{L}^2(\Omega)\times\boldsymbol{L}^2(\Omega)\times H^{1/2}_{00}(\partial\Omega_h)\times H^{-1/2}(\partial\Omega_h),$$
$$\boldsymbol{V} := H^1(\Omega_h)\times\boldsymbol{H}(\mathrm{div};\Omega_h).$$

Define the operator $B : \boldsymbol{U} \to \boldsymbol{V}'$ as

$$\langle B\boldsymbol{u}, \boldsymbol{v}\rangle := (u, \boldsymbol{\beta}\cdot\nabla v + \nabla\cdot\boldsymbol{\tau})_{\Omega_h} + (\boldsymbol{\sigma}, \boldsymbol{\tau})_\Omega + (\boldsymbol{\rho}, \nabla v)_{\Omega_h} - \langle\hat{\rho}, \gamma(v)\rangle_{\partial\Omega_h} - \langle\gamma_n(\boldsymbol{\tau}), \hat{u}\rangle_{\partial\Omega_h} \tag{18}$$

for $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \boldsymbol{\rho}, \hat{u}, \hat{\rho}) \in \boldsymbol{U}$ and $\boldsymbol{v} = (v, \boldsymbol{\tau}) \in \boldsymbol{V}$. Problem (17) is then reformulated as

$$\boldsymbol{u} \in \boldsymbol{U} \ : \quad B\boldsymbol{u} = F \in \boldsymbol{V}', \tag{19}$$

with $F \in \boldsymbol{V}'$ such that $F : \boldsymbol{v} \mapsto (f, v)_\Omega$ for all $\boldsymbol{v} = (v, \boldsymbol{\tau}) \in \boldsymbol{V}$. The following lemma gives a strong characterization of $\boldsymbol{u} \in \boldsymbol{U}$ solving (19).

LEMMA 4.1. *Let $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \boldsymbol{\rho}, \hat{u}, \hat{\rho}) \in \boldsymbol{U}$ be a solution of (19). Then, $u \in H^1_0(\Omega)$, $\boldsymbol{\rho} \in \boldsymbol{H}(\mathrm{div};\Omega)$, and $\boldsymbol{\sigma} = \nabla u$, $-\nabla\cdot(\boldsymbol{\rho}+\boldsymbol{\beta}u) = f$, $\hat{u} = \gamma(u)$ and $\hat{\rho} = \gamma_n(\boldsymbol{\rho}+\boldsymbol{\beta}u)$.*

In the following analysis, the Cartesian space $\boldsymbol{U}$ is equipped with the scaled norm $\|\cdot\|_{\boldsymbol{U}}$ defined for all $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \boldsymbol{\rho}, \hat{u}, \hat{\rho}) \in \boldsymbol{U}$ as

$$\|\boldsymbol{u}\|^2_{\boldsymbol{U}} := \|u\|^2_{L^2(\Omega)} + \ell_\Omega^2\|\boldsymbol{\sigma}\|^2_{\boldsymbol{L}^2(\Omega)} + \ell_\Omega^2\lambda_\sharp^{-2}\|\boldsymbol{\rho}\|^2_{\boldsymbol{L}^2(\Omega)} + \|\hat{u}\|^2_{H^{1/2}_{00}(\partial\Omega_h)} + \ell_\Omega^2\lambda_\sharp^{-2}\|\hat{\rho}\|^2_{H^{-1/2}(\partial\Omega_h)}.$$

Similarly, the space $\boldsymbol{V}$ is equipped with the inner product $(\cdot,\cdot)_{\boldsymbol{V}}$ defined for $\boldsymbol{v}=(v,\boldsymbol{\tau}),\boldsymbol{w}=(w,\boldsymbol{\eta})\in\boldsymbol{V}$ as

$$(\boldsymbol{v},\boldsymbol{w})_{\boldsymbol{V}} := \lambda_\sharp^2\ell_\Omega^{-4}\,(v,w)_{H^1(\Omega_h)} + \ell_\Omega^{-2}\,(\boldsymbol{\tau},\boldsymbol{\eta})_{\boldsymbol{H}(\mathrm{div};\Omega_h)}\,.$$

With these norms in $\boldsymbol{U}$ and $\boldsymbol{V}$, B is uniformly bounded. Owing to the following lemma, B is surjective or, equivalently, B^* is injective with closed range in $\boldsymbol{U}'$.

LEMMA 4.2. *$\|B^*\boldsymbol{v}\|_{\boldsymbol{U}'}^2 \geq c_B\|\boldsymbol{v}\|_{\boldsymbol{V}}^2$ for all $\boldsymbol{v}\in\boldsymbol{V}$ with $c_B = c_{PS}/2$ and c_{PS} the Poincaré-Steklov constant from Lemma 3.1.*

Proof. Let $\boldsymbol{v}\in\boldsymbol{V}$ with $\boldsymbol{v}=(v,\boldsymbol{\tau})$ and note that

$$\|B^*\boldsymbol{v}\|_{\boldsymbol{U}'}^2 \geq \|\boldsymbol{\beta}\cdot\nabla v + \nabla\cdot\boldsymbol{\tau}\|_{L^2(\Omega_h)}^2 + \ell_\Omega^{-2}\|\boldsymbol{\tau}\|_{L^2(\Omega)}^2 + \ell_\Omega^{-2}\lambda_\sharp^2\|\nabla v\|_{L^2(\Omega_h)}^2 + \ell_\Omega^{-2}\lambda_\sharp^2\left(\sup_{\hat\tau\in H^{-1/2}(\partial\Omega_h)}\frac{|\langle\hat\tau,\gamma(v)\rangle_{\partial\Omega_h}|}{\|\hat\tau\|_{H^{-1/2}(\partial\Omega_h)}}\right)^2.$$

Then, using the Poincaré-Steklov inequality in $H^1(\Omega_h)$ from Lemma 3.1, it follows

$$\|B^*\boldsymbol{v}\|_{\boldsymbol{U}'}^2 \geq \|\boldsymbol{\beta}\cdot\nabla v + \nabla\cdot\boldsymbol{\tau}\|_{L^2(\Omega_h)}^2 + \ell_\Omega^{-2}\|\boldsymbol{\tau}\|_{L^2(\Omega)}^2 + \frac{1}{2}\ell_\Omega^{-2}\lambda_\sharp^2\|\nabla v\|_{L^2(\Omega_h)}^2 + \frac{1}{2}c_{PS}\ell_\Omega^{-4}\lambda_\sharp^2\|v\|_{H^1(\Omega_h)}^2.$$

Hence, observing that $\|\boldsymbol{\tau}\|_{\boldsymbol{H}(\mathrm{div};\Omega_h)}^2 \leq \|\boldsymbol{\tau}\|_{\boldsymbol{L}^2(\Omega)}^2 + \ell_\Omega^2\|\nabla\cdot\boldsymbol{\tau}+\boldsymbol{\beta}\cdot\nabla v\|_{\boldsymbol{L}^2(\Omega_h)}^2 + \lambda_\sharp^2\|\nabla v\|_{\boldsymbol{L}^2(\Omega_h)}^2$ by the triangle inequality and assumption (15), it follows that

$$\|B^*\boldsymbol{v}\|_{\boldsymbol{U}'}^2 \geq \frac{1}{2}\ell_\Omega^{-2}\|\boldsymbol{\tau}\|_{\boldsymbol{H}(\mathrm{div};\Omega_h))}^2 + \frac{1}{2}c_{PS}\ell_\Omega^{-4}\lambda_\sharp^2\|v\|_{H^1(\Omega_h)}^2.$$

Observing that $c_{PS}\leq 1$ in Lemma 3.1, the statement follows by definition of the norm in $\boldsymbol{V}$. ∎

Denoting by $P_B:\ \boldsymbol{U}\to\mathcal{N}(B)$ the projector of $\boldsymbol{U}$ on $\mathcal{N}(B)$, the following lemma is consequence of the boundedness of B and Lemma 4.2. In particular, assumption (ii) of Proposition 2.1 is satisfied.

LEMMA 4.3. *$\|\boldsymbol{u}-P_B\boldsymbol{u}\|_{\boldsymbol{U}}^2 \leq c_B\|B\boldsymbol{u}\|_{\boldsymbol{V}'}^2$ for all $\boldsymbol{u}\in\boldsymbol{U}$, with the constant c_B defined in Lemma 4.2.*

Proof. By standard arguments the statement is equivalent to the statement of Lemma 4.2. Indeed, by the continuity of $B:\ \boldsymbol{U}\to\boldsymbol{V}'$ and the boundedness below of B^*, $B:\ \mathcal{N}(B)^\perp\to\boldsymbol{V}'$ is injective with closed range, i.e., $B^*:\ \boldsymbol{V}\to(\mathcal{N}(B)^\perp)'$ is surjective, so that

$$\|B\boldsymbol{u}\|_{\boldsymbol{V}'} = \sup_{\boldsymbol{v}\in\boldsymbol{V}}\frac{\langle B^*\boldsymbol{v},\boldsymbol{u}-P_B(\boldsymbol{u})\rangle}{\|\boldsymbol{v}\|_{\boldsymbol{V}}} \geq c_B^{1/2}\sup_{\boldsymbol{v}\in\boldsymbol{V}}\frac{\langle B^*\boldsymbol{v},\boldsymbol{u}-P_B(\boldsymbol{u})\rangle}{\|B^*\boldsymbol{v}\|_{\boldsymbol{U}'}} = c_B^{1/2}\|\boldsymbol{u}-P_B(\boldsymbol{u})\|_{\boldsymbol{U}}$$

for any $\boldsymbol{u}\in\boldsymbol{U}$. ∎

The non-trivial kernel $\mathcal{N}(B)$ can be represented as the image of the map $E: H_0^1(\Omega)\times\boldsymbol{H}\to\boldsymbol{U}$ defined, for all $\psi\in H_0^1(\Omega)$ and $\boldsymbol{\eta}\in\boldsymbol{H}=\{\boldsymbol{v}\in\boldsymbol{H}(\mathrm{div};\Omega)\,|\,\nabla\cdot\boldsymbol{v}=0\}$, as

$$E(\psi,\boldsymbol{\eta}) = \big(\psi,\nabla\psi,\boldsymbol{\eta}-\boldsymbol{\beta}\psi,\gamma(\psi),\gamma_n(\boldsymbol{\eta})\big). \tag{20}$$

LEMMA 4.4. *The map $E: H_0^1(\Omega)\times\boldsymbol{H}\to\boldsymbol{U}$ satisfies*

$$c_E\|E(\psi,\boldsymbol{\eta})\|_{\boldsymbol{U}}^2 \leq \ell_\Omega^2\|\nabla\psi\|_{\boldsymbol{L}^2(\Omega)}^2 + \ell_\Omega^2\lambda_\sharp^{-2}\|\boldsymbol{\eta}-\boldsymbol{\beta}\psi\|_{\boldsymbol{L}^2(\Omega)}^2 \quad \forall\psi\in H_0^1(\Omega),\,\boldsymbol{\eta}\in\boldsymbol{H}, \tag{21}$$

with $c_E^{-1} = 4c_{PS,0} + 3$ *and* $c_{PS,0}$ *the Poincaré-Steklov constant defined in Lemma 3.1. In addition,* $E\left(H_0^1(\Omega)\times\boldsymbol{H}\right) = \mathcal{N}(B)$.

Proof. Estimate (21) is consequence of the Poincaré-Steklov inequality in $H_0^1(\Omega)$ and the triangle inequality. Let us prove that $E\left(H_0^1(\Omega)\times\boldsymbol{H}\right) = \mathcal{N}(B)$. Consider $\psi \in H_0^1(\Omega)$, $\boldsymbol{\eta} \in \boldsymbol{H}$ and let $\boldsymbol{w} = E(\psi,\boldsymbol{\eta})$. For any $\boldsymbol{v} = (v,\boldsymbol{\tau}) \in \boldsymbol{V}$ it follows that

$$\langle B\boldsymbol{w},\boldsymbol{v}\rangle = (\psi,\nabla\cdot\boldsymbol{\tau})_{\Omega_h} + (\nabla\psi,\boldsymbol{\tau})_{\Omega} + (\boldsymbol{\eta},\nabla v)_{\Omega_h} \\ - \langle\gamma_n(\boldsymbol{\eta}),\gamma(v)\rangle_{\partial\Omega_h} - \langle\gamma_n(\boldsymbol{\tau}),\gamma(\psi)\rangle_{\partial\Omega_h}.$$

Hence, integrating by parts we obtain $\langle B\boldsymbol{w},\boldsymbol{v}\rangle = 0$ so that $\boldsymbol{w} \in \mathcal{N}(B)$. Conversely, assume that $\boldsymbol{u} = (u,\boldsymbol{\sigma},\boldsymbol{\rho},\hat{u},\hat{\rho}) \in \mathcal{N}(B)$. Then, computing $\langle B\boldsymbol{u},(0,\boldsymbol{\tau})\rangle$ we deduce that $u \in H^1(\Omega_h)$ and that $\boldsymbol{\sigma}_{|T} = (\nabla u)_{|T}$ for all $T \in \Omega_h$. We conclude that

$$0 = (u,\nabla\cdot\boldsymbol{\tau})_{\Omega_h} + (\boldsymbol{\sigma},\boldsymbol{\tau})_{\Omega} - \langle\gamma_n(\boldsymbol{\tau}),\hat{u}\rangle_{\partial\Omega_h} = \langle\gamma_n(\boldsymbol{\tau}),\gamma(u)\rangle_{\partial\Omega_h} - \langle\gamma_n(\boldsymbol{\tau}),\hat{u}\rangle_{\partial\Omega_h},$$

so that $\gamma(u) = \hat{u} \in H_{00}^{1/2}(\partial\Omega_h)$. Then, by definition of $H_{00}^{1/2}(\partial\Omega_h)$, we infer that $u \in H_0^1(\Omega)$. Proceeding similarly by testing with $\boldsymbol{v} = (v,\boldsymbol{0})$, we infer that $\boldsymbol{\beta}u + \boldsymbol{\rho} \in \boldsymbol{H}(\mathrm{div};\Omega_h)$ with $(\nabla\cdot(\boldsymbol{\beta}u+\boldsymbol{\rho}))_{|T} = 0$ for all $T \in \Omega_h$. It follows that $\hat{\rho} = \gamma_n(\boldsymbol{\beta}u+\boldsymbol{\rho}) \in H^{-1/2}(\partial\Omega_h)$ so that $\boldsymbol{\beta}u+\boldsymbol{\rho} \in \boldsymbol{H}(\mathrm{div};\Omega)$, and then $\boldsymbol{\beta}u+\boldsymbol{\rho} \in \boldsymbol{H}$. As a result, there exists $\boldsymbol{\eta} \in \boldsymbol{H}$ such that $\boldsymbol{\rho} = -\boldsymbol{\beta}u+\boldsymbol{\eta}$. To conclude, if $\boldsymbol{u} \in \mathcal{N}(B)$, then $u \in H_0^1(\Omega)$, $\boldsymbol{\rho}+\boldsymbol{\beta}u \in \boldsymbol{H}$ and $\boldsymbol{u} = E(u,\boldsymbol{\rho}+\boldsymbol{\beta}u)$. ∎

4.2. Nonlinear penalty term. In this section we devise a nonlinear penalty form to enforce the closure relation $\boldsymbol{\rho} = \boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma}$ and to control $\mathcal{N}(B)$. To simplify the presentation let us introduce $\pi : \boldsymbol{U} \to \boldsymbol{L}^2(\Omega)\times\boldsymbol{L}^2(\Omega)$ by defining $\pi(\boldsymbol{u}) := (\boldsymbol{\sigma},\boldsymbol{\rho})$ for $\boldsymbol{u} = (u,\boldsymbol{\sigma},\boldsymbol{\rho},\hat{u},\hat{\rho}) \in \boldsymbol{U}$. Then we define a nonlinear operator $C : \boldsymbol{U} \to \boldsymbol{U}'$ by

$$\langle C(\boldsymbol{u}),\boldsymbol{v}\rangle := \ell_\Omega^2\lambda_\sharp^{-2}\big(\boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma}-\boldsymbol{\rho},\alpha\lambda_\flat\boldsymbol{\theta}-\boldsymbol{\eta}\big)_\Omega \quad \forall\boldsymbol{u},\boldsymbol{v} \in \boldsymbol{U}, \tag{22}$$

with $\pi(\boldsymbol{u}) = (\boldsymbol{\sigma},\boldsymbol{\rho})$ and $\pi(\boldsymbol{v}) = (\boldsymbol{\theta},\boldsymbol{\eta})$. Here, $\alpha > 0$ denotes a stability parameter that will be chosen greater than $\lambda_{\sharp/\flat}^2$, with $\lambda_{\sharp/\flat} := \lambda_\sharp/\lambda_\flat$ the diffusive anisotropy ratio.

We start by establishing the Lipschitz continuity of C.

LEMMA 4.5. $\|C(\boldsymbol{u}_1) - C(\boldsymbol{u}_2)\|_{\boldsymbol{U}'} \le c_{\mathrm{Lip}}\|\boldsymbol{u}_1-\boldsymbol{u}_2\|_{\boldsymbol{U}}$ *for all* $\boldsymbol{u}_1,\boldsymbol{u}_2 \in \boldsymbol{U}$, *with* $c_{\mathrm{Lip}} = 2\max(\alpha\lambda_{\sharp/\flat}^{-1},1)$.

Proof. Let $\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{v} \in \boldsymbol{U}$ such that $\pi(\boldsymbol{u}_1) = (\boldsymbol{\sigma}_1,\boldsymbol{\rho}_1)$, $\pi(\boldsymbol{u}_2) = (\boldsymbol{\sigma}_2,\boldsymbol{\rho}_2)$ and $\pi(\boldsymbol{v}) = (\boldsymbol{\theta},\boldsymbol{\eta})$. Owing to (22), it follows that

$$|\langle C(\boldsymbol{u}_1)-C(\boldsymbol{u}_2),\boldsymbol{v})\rangle| \le \ell_\Omega^2\lambda_\sharp^{-2}\left|\big(\boldsymbol{\lambda}(|\boldsymbol{\sigma}_1|)\boldsymbol{\sigma}_1-\boldsymbol{\lambda}(|\boldsymbol{\sigma}_2|)\boldsymbol{\sigma}_2,\alpha\lambda_\flat\boldsymbol{\theta}-\boldsymbol{\eta}\big)_\Omega\right| \\ + \ell_\Omega^2\lambda_\sharp^{-2}\left|\big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2,\alpha\lambda_\flat\boldsymbol{\theta}-\boldsymbol{\eta}\big)_\Omega\right|.$$

The Cauchy–Schwarz inequality and assumption (14c) yield

$$|\langle C(\boldsymbol{u}_1)-C(\boldsymbol{u}_2),\boldsymbol{v}\rangle| \\ \le \ell_\Omega^2\lambda_\sharp^{-1}\left(\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)} + \lambda_\sharp^{-1}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)}\right)\|\alpha\lambda_\flat\boldsymbol{\theta}-\boldsymbol{\eta}\|_{\boldsymbol{L}^2(\Omega)}.$$

Observing that $\|\alpha\lambda_\flat\boldsymbol{\theta}-\boldsymbol{\eta}\|_{\boldsymbol{L}^2(\Omega)} \le 2\lambda_\sharp\ell_\Omega^{-1}\max(\alpha\lambda_{\sharp/\flat}^{-1},1)\|\boldsymbol{v}\|_{\boldsymbol{U}}$, the desired estimate follows. ∎

The next lemma verifies assumption (iii) of Proposition 2.1 for the model problem.

LEMMA 4.6. *Assume that* $\alpha > \lambda_{\sharp/\flat}^2$. *Then*

$$c_U\|P_B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{U}}^2 \le \langle C(\boldsymbol{u}_1)-C(\boldsymbol{u}_2),\boldsymbol{u}_1-\boldsymbol{u}_2\rangle + c_V\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}^2 \quad \forall \boldsymbol{u}_1,\boldsymbol{u}_2\in\boldsymbol{U},$$

where c_U *and* c_V *are given by*

$$c_U = \frac{c_E}{2}\min\left(\frac{1}{2},\alpha\lambda_{\sharp/\flat}^{-2}-1\right), \quad c_V = \frac{c_B}{2}\min\left(\frac{1}{2},\alpha\lambda_{\sharp/\flat}^{-2}-1\right)+\frac{5}{2}c_B\lambda_{\sharp/\flat}^{-2}\alpha^2,$$

with c_B *and* c_E *the constants defined in Lemmata 4.2 and 4.4, respectively.*

Proof. Let $\boldsymbol{u}_1,\boldsymbol{u}_2\in\boldsymbol{U}$ be such that $\pi(\boldsymbol{u}_1)=(\boldsymbol{\sigma}_1,\boldsymbol{\rho}_1)$ and $\pi(\boldsymbol{u}_2)=(\boldsymbol{\sigma}_2,\boldsymbol{\rho}_2)$. By Lemma 4.4 there exist $\psi_1,\psi_2\in H_0^1(\Omega)$ and $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2\in\boldsymbol{H}$ such that

$$P_B(\boldsymbol{u}_1) = E(\psi_1,\boldsymbol{\eta}_1) = \big(\psi_1,\nabla\psi_1,\boldsymbol{\eta}_1-\boldsymbol{\beta}\psi_1,\gamma(\psi_1),\gamma_n(\boldsymbol{\eta}_1)\big), \tag{23a}$$

$$P_B(\boldsymbol{u}_2) = E(\psi_1,\boldsymbol{\eta}_1) = \big(\psi_2,\nabla\psi_2,\boldsymbol{\eta}_2-\boldsymbol{\beta}\psi_2,\gamma(\psi_2),\gamma_n(\boldsymbol{\eta}_2)\big). \tag{23b}$$

Owing to the continuity estimate (21) from Lemma 4.4, it follows that

$$c_E\|P_B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{U}}^2 \le \ell_\Omega^2\|\nabla\psi_1-\nabla\psi_2\|_{\boldsymbol{L}^2(\Omega)}^2 + \ell_\Omega^2\lambda_\sharp^{-2}\|\overline{\boldsymbol{\eta}}_1-\overline{\boldsymbol{\eta}}_2\|_{\boldsymbol{L}^2(\Omega)}^2,$$

where we have denoted $\overline{\boldsymbol{\eta}}_i=\boldsymbol{\eta}_i-\boldsymbol{\beta}\psi_i$ for $i=1,2$. By the triangle inequality we have

$$\begin{aligned} c_E\|P_B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{U}}^2 \le{}& \ell_\Omega^2\|\nabla\psi_1-\nabla\psi_1-(\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2)\|_{\boldsymbol{L}^2(\Omega)}^2 + \ell_\Omega^2\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)}^2 \\ &+ \ell_\Omega^2\lambda_\sharp^{-2}\|\overline{\boldsymbol{\eta}}_1-\overline{\boldsymbol{\eta}}_2-(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2)\|_{\boldsymbol{L}^2(\Omega)}^2 + \ell_\Omega^2\lambda_\sharp^{-2}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)}^2. \end{aligned}$$

Observing that $\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2-(\nabla\psi_1-\nabla\psi_2)\|_{\boldsymbol{L}^2(\Omega)}^2 \le c_B\ell_\Omega^{-2}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}^2$ and $\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2-(\overline{\boldsymbol{\eta}}_1-\overline{\boldsymbol{\eta}}_2)\|_{\boldsymbol{L}^2(\Omega)}^2 \le c_B\lambda_\sharp^2\ell_\Omega^{-2}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}^2$ owing to Lemma 4.3, we obtain

$$c_E\|P_B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{U}}^2 \le c_B\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}^2 + \ell_\Omega^2\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)}^2 + \ell_\Omega^2\lambda_\sharp^{-2}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)}^2. \tag{24}$$

It remains to prove that there exist constants $c,c'>0$ such that

$$\begin{aligned} &c\left(\ell_\Omega^2\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)}^2 + \ell_\Omega^2\lambda_\sharp^{-2}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)}^2\right) \\ &\qquad\le \langle C(\boldsymbol{u}_1)-C(\boldsymbol{u}_2),\boldsymbol{u}_1-\boldsymbol{u}_2\rangle + c'\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}^2. \end{aligned} \tag{25}$$

Let $F(\boldsymbol{u}_1,\boldsymbol{u}_2)=\ell_\Omega^{-2}\lambda_\sharp^2\langle C(\boldsymbol{u}_1)-C(\boldsymbol{u}_2),\boldsymbol{u}_1-\boldsymbol{u}_2\rangle$ and observe that

$$\begin{aligned} F(\boldsymbol{u}_1,\boldsymbol{u}_2) ={}& \alpha\lambda_\flat\big(\boldsymbol{\lambda}(|\boldsymbol{\sigma}_1|)\boldsymbol{\sigma}_1-\boldsymbol{\lambda}(|\boldsymbol{\sigma}_2|)\boldsymbol{\sigma}_2,\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\big)_\Omega + \big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2,\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\big)_\Omega \\ &- \big(\boldsymbol{\lambda}(|\boldsymbol{\sigma}_1|)\boldsymbol{\sigma}_1-\boldsymbol{\lambda}(|\boldsymbol{\sigma}_2|)\boldsymbol{\sigma}_2,\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\big)_\Omega - \alpha\lambda_\flat\big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2,\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\big)_\Omega. \end{aligned}$$

Owing to assumptions (14b) and (14c), it follows that

$$\begin{aligned} F(\boldsymbol{u}_1,\boldsymbol{u}_2) \ge{}& \alpha\lambda_\flat^2\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)}^2 + \|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)}^2 \\ &- \lambda_\sharp\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)} - \alpha\lambda_\flat\big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2,\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\big)_\Omega. \end{aligned}$$

Hence, applying Young's inequality, we conclude that

$$F(\boldsymbol{u}_1,\boldsymbol{u}_2) \geq \left(\alpha\lambda_\flat^2 - \frac{\lambda_\sharp^2}{2}\right)\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|^2_{\boldsymbol{L}^2(\Omega)} + \frac{1}{2}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|^2_{\boldsymbol{L}^2(\Omega)} - \alpha\lambda_\flat\big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2,\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\big)_\Omega.$$

The last term above can be bounded from below by using the representation (23) of $\mathcal{N}(B)$. First, we observe that

$$\begin{aligned}\big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2,\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\big)_\Omega &= \big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2,\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2-(\nabla\psi_1-\nabla\psi_2)\big)_\Omega + \big(\overline{\boldsymbol{\eta}}_1-\overline{\boldsymbol{\eta}}_2,\nabla\psi_1-\nabla\psi_2\big)_\Omega\\ &\quad + \big(\boldsymbol{\rho}_1-\boldsymbol{\rho}_2-(\overline{\boldsymbol{\eta}}_1-\overline{\boldsymbol{\eta}}_2),\nabla\psi_1-\nabla\psi_2\big)_\Omega =: T_1+T_2+T_3.\end{aligned}$$

Applying the Cauchy–Schwarz and triangle inequalities, and Lemma 4.3 for T_1 and T_3, it follows

$$|T_1| \leq c_B^{1/2}\ell_\Omega^{-1}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)},$$

and

$$\begin{aligned}|T_3| &\leq c_B^{1/2}\lambda_\sharp\ell_\Omega^{-1}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}\|\nabla\psi_1-\nabla\psi_2\|_{\boldsymbol{L}^2(\Omega)}\\ &\leq c_B^{1/2}\lambda_\sharp\ell_\Omega^{-1}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)} + c_B\lambda_\sharp\ell_\Omega^{-2}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|^2_{\boldsymbol{V}'}.\end{aligned}$$

Recalling the definition $\overline{\boldsymbol{\eta}}_i = -\boldsymbol{\beta}\psi_i+\boldsymbol{\eta}_i$, $i=1,2$, with $\boldsymbol{\eta}_i\in\boldsymbol{H}=(\nabla H_0^1(\Omega))^\perp$, we infer that

$$T_2 = -\big(\boldsymbol{\beta}\psi_1-\boldsymbol{\beta}\psi_2,\nabla\psi_1-\nabla\psi_2\big)_\Omega \leq 0$$

owing to assumption (14a). Collecting the previous estimates we obtain

$$\begin{aligned}\left(\alpha\lambda_\flat^2-\frac{\lambda_\sharp^2}{2}\right)\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|^2_{\boldsymbol{L}^2(\Omega)} + \frac{1}{2}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|^2 &\leq F(\boldsymbol{u}_1,\boldsymbol{u}_2)\\ &\quad + c_B\alpha\lambda_\flat^2\lambda_{\sharp/\flat}\ell_\Omega^{-2}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|^2_{\boldsymbol{V}'}\\ &\quad + c_B^{1/2}\alpha\lambda_\flat\ell_\Omega^{-1}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|_{\boldsymbol{L}^2(\Omega)}\\ &\quad + c_B^{1/2}\alpha\lambda_\flat^2\lambda_{\sharp/\flat}\ell_\Omega^{-1}\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|_{\boldsymbol{V}'}\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|_{\boldsymbol{L}^2(\Omega)}.\end{aligned}$$

Young's inequality and assumption $\alpha > \lambda_{\sharp/\flat}^2$ yield

$$\frac{1}{2}\left(\alpha\lambda_\flat^2-\lambda_\sharp^2\right)\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|^2_{\boldsymbol{L}^2(\Omega)} + \frac{1}{4}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|^2 \leq F(\boldsymbol{u}_1,\boldsymbol{u}_2) + \frac{5}{2}c_B\lambda_\flat^2\ell_\Omega^{-2}\alpha^2\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|^2_{\boldsymbol{V}'}.$$

As a result, multiplying by $\lambda_\sharp^{-2}$, we obtain

$$\begin{aligned}\frac{1}{2}\min\left(\frac{1}{2},\alpha\lambda_{\sharp/\flat}^{-2}-1\right)\left(\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|^2_{\boldsymbol{L}^2(\Omega)} + \lambda_\sharp^{-2}\|\boldsymbol{\rho}_1-\boldsymbol{\rho}_2\|^2_{\boldsymbol{L}^2(\Omega)}\right) &\leq \lambda_\sharp^{-2}F(\boldsymbol{u}_1,\boldsymbol{u}_2)\\ &\quad + \frac{5}{2}c_B\ell_\Omega^{-2}\lambda_{\sharp/\flat}^{-2}\alpha^2\|B(\boldsymbol{u}_1-\boldsymbol{u}_2)\|^2_{\boldsymbol{V}'}.\end{aligned}$$

The statement follows multiplying (24) by $\frac{1}{2}\min\left(\frac{1}{2},\alpha\lambda_{\sharp/\flat}^{-2}-1\right)$. ∎

4.3. Penalized variational formulation and well-posedness. We are now in a position to present our penalized variational formulation of the model problem and to prove its well-posedness.

Let $R_{\boldsymbol{V}} : \boldsymbol{V} \to \boldsymbol{V}'$ be the Riesz operator, i.e., $\langle R_{\boldsymbol{V}} \boldsymbol{v}, \boldsymbol{w}\rangle = (\boldsymbol{v}, \boldsymbol{w})_{\boldsymbol{V}}$ for $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{V}$, and, for $\kappa \geq 0$, let $\boldsymbol{T}_\kappa : \boldsymbol{U}\times\boldsymbol{V} \to \boldsymbol{U}'\times\boldsymbol{V}'$ be the nonlinear operator

$$\boldsymbol{T}_\kappa := \begin{pmatrix} C & B^* \\ B & -\frac{1}{\kappa} R_{\boldsymbol{V}} \end{pmatrix}. \tag{26}$$

Here, $B : \boldsymbol{U} \to \boldsymbol{V}'$ and $C : \boldsymbol{U} \to \boldsymbol{U}'$ are the linear and nonlinear operators defined by (18) and (22), respectively.

Following Section 2, problem (13) has the variational formulations

$$(\boldsymbol{u}, \boldsymbol{v}) \in \boldsymbol{U}\times\boldsymbol{V} : \qquad \boldsymbol{T}_\kappa(\boldsymbol{u}, \boldsymbol{v}) = (0, F) \quad \text{in } \boldsymbol{U}'\times\boldsymbol{V}' \tag{27}$$

and

$$\boldsymbol{u} \in \boldsymbol{U} : \qquad \kappa\langle R_{\boldsymbol{V}}^{-1} B\boldsymbol{u}, B\boldsymbol{v}\rangle + \langle C(\boldsymbol{u}), \boldsymbol{v}\rangle = \kappa\langle R_{\boldsymbol{V}}^{-1} F, B\boldsymbol{v}\rangle \quad \forall \boldsymbol{v} \in \boldsymbol{U}. \tag{28}$$

In addition, since $R_{\boldsymbol{V}}^{-1}$ is self-adjoint, (28) can be reformulated as

$$\boldsymbol{u} \in \boldsymbol{U} : \qquad \kappa\langle B\boldsymbol{u}, \Theta\boldsymbol{v}\rangle + \langle C(\boldsymbol{u}), \boldsymbol{v}\rangle = \kappa\langle F, \Theta\boldsymbol{v}\rangle_\Omega \quad \forall \boldsymbol{v} \in \boldsymbol{U}. \tag{29}$$

Here, we have used the so-called *trial-to-test operator* $\Theta := R_{\boldsymbol{V}}^{-1} B : \boldsymbol{U} \to \boldsymbol{V}$, cf., e.g., [12].

Selecting κ sufficiently large, both problems are well posed and equivalent.

THEOREM 4.7. *Assume that $\alpha > \lambda_{\sharp/\flat}^2$ and $\kappa \geq c_V + c_B c_U$ with c_B, c_U, c_V defined in Lemmata 4.2 and 4.6. Then*

(a) problems (27) and (29) are equivalent,

(b) $\boldsymbol{T}_\kappa : \boldsymbol{U}\times\boldsymbol{V} \to \boldsymbol{U}'\times\boldsymbol{V}'$ defined by (26) is invertible with Lipschitz continuous inverse,

(c) the solution $\boldsymbol{u} \in \boldsymbol{U}$ of (29) satisfies $B\boldsymbol{u} = F$ in $\boldsymbol{V}'$ and $C(\boldsymbol{u}) = 0$ in $\boldsymbol{U}'$.

Proof. Statement (a) is a direct consequence of Theorem 2.3 and the definition of Θ. The second statement (b) follows from Proposition 2.1. Indeed, $\langle B\boldsymbol{v}, R_{\boldsymbol{V}}^{-1} B\boldsymbol{v}\rangle = \|B\boldsymbol{v}\|_{\boldsymbol{V}'}^2$ for all $\boldsymbol{v} \in \boldsymbol{V}$, so that assumption (i) of Proposition 2.1 is satisfied with coercivity constant $c_R = 1$. Since assumptions (ii) and (iii) of Proposition 2.1 hold by Lemmata 4.3 and 4.6, we infer statement (b). Finally, statement (c) is a consequence of Proposition 2.4. The operator B being surjective by Lemma 4.2, it remains to prove that

$$\langle C(\boldsymbol{u}), \boldsymbol{v}\rangle = 0 \quad \forall \boldsymbol{v} \in \mathcal{N}(B) \quad \Longrightarrow \quad C(\boldsymbol{u}) = 0 \text{ in } \boldsymbol{U}'.$$

In (28) we select $\boldsymbol{v} = (0, \boldsymbol{0}, \boldsymbol{\eta}, 0, \gamma_n(\boldsymbol{\eta})) \in \mathcal{N}(B)$ with $\boldsymbol{\eta} \in \boldsymbol{H}$, giving

$$\langle C(\boldsymbol{u}), \boldsymbol{v}\rangle_{\boldsymbol{U}',\boldsymbol{U}} = -\ell_\Omega^2 \lambda_\sharp^{-2} \left(\boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma} - \boldsymbol{\rho}, \boldsymbol{\eta}\right)_\Omega = 0,$$

since $\Theta\boldsymbol{v} = R_{\boldsymbol{V}}^{-1} B\boldsymbol{v} = 0$. Therefore, $\boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma} - \boldsymbol{\rho} \in \boldsymbol{H}^\perp = \nabla H_0^1(\Omega)$ since $\nabla H_0^1(\Omega)$ is closed in $\boldsymbol{L}^2(\Omega)$. That is, there exists $\psi \in H_0^1(\Omega)$ such that $\nabla\psi = \boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma} - \boldsymbol{\rho}$. Next, choosing $\boldsymbol{v} = (\varphi, \nabla\varphi, -\boldsymbol{\beta}\varphi, \gamma(\varphi), 0) \in \mathcal{N}(B)$ with $\varphi \in H_0^1(\Omega)$, it follows that $(\nabla\psi, \alpha\lambda_\flat\nabla\varphi + \boldsymbol{\beta}\varphi)_\Omega = 0$ for all $\varphi \in H_0^1(\Omega)$. As a result, owing to assumption (14a), we infer that $\psi = 0$, and then $\boldsymbol{\lambda}(|\boldsymbol{\sigma}|)\boldsymbol{\sigma} = \boldsymbol{\rho}$, so that $C(\boldsymbol{u}) = 0$ in $\boldsymbol{U}'$. ∎

5. Relaxed DPG scheme. With all the preparations at hand, the formulation of our DPG scheme for the model problem is immediate and the proof of its quasi-optimal convergence is straightforward.

Considering the continuous problem (27), which is equivalent to (29) in the sense of Theorem 2.3, we follow the presentation of Section 2 and consider two finite-dimensional spaces $\boldsymbol{U}_h$ and $\boldsymbol{V}_h$ with $\boldsymbol{U}_h \subset \boldsymbol{U}$ and $\boldsymbol{V}_h \subset \boldsymbol{V}$. Our discrete problem is

$$\boldsymbol{u}_h \in \boldsymbol{U}_h \ : \quad \langle C_h(\boldsymbol{u}_h), \boldsymbol{w}_h\rangle + \kappa\langle B_h\boldsymbol{u}_h, \Theta_h\boldsymbol{w}_h\rangle = \kappa\langle \boldsymbol{j}_h^* F, \Theta_h\boldsymbol{w}_h\rangle \quad \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \tag{30}$$

with

$$C_h := \boldsymbol{i}_h^* C \boldsymbol{i}_h, \quad B_h := \boldsymbol{j}_h^* B \boldsymbol{i}_h, \quad R_{\boldsymbol{V}_h} = \boldsymbol{j}_h^* R_{\boldsymbol{V}} \boldsymbol{j}_h, \ \text{ and } \ \Theta_h = R_{\boldsymbol{V}_h}^{-1} B_h.$$

The well-posedness of this problem follows from Proposition 2.7.

THEOREM 5.1. *Assume that there exists a Fortin operator $\Pi : \boldsymbol{V} \to \boldsymbol{V}_h$ satisfying (11). Then, for all $\kappa \geq (c_V + c_B c_U)c_\Pi$, with c_B, c_U, c_V defined in Lemmata 4.2 and 4.6, (30) is well-posed and converges quasi-optimally,*

$$\|\boldsymbol{u} - \boldsymbol{i}_h(\boldsymbol{u}_h)\|_{\boldsymbol{U}} \leq \big(1 + c_U^{-1}\big(c_{\mathrm{Lip}} + \kappa\|B^* \boldsymbol{j}_h R_{\boldsymbol{V}_h}^{-1} \boldsymbol{j}_h^* B\|_{\mathcal{L}(\boldsymbol{U},\boldsymbol{U}')}\big)\big) \inf_{\boldsymbol{w}_h \in \boldsymbol{U}_h} \|\boldsymbol{u} - \boldsymbol{i}_h(\boldsymbol{w}_h)\|_{\boldsymbol{U}},$$

with $c_{\mathrm{Lip}} > 0$ defined in Lemma 4.5. Here, $\boldsymbol{u} \in \boldsymbol{U}$ and $\boldsymbol{u}_h \in \boldsymbol{U}_h$ are the unique solution of (28) and (30), respectively.

Proof. The well-posedness of (30) is a consequence of Proposition 2.7. Indeed, assumptions (ii) and (iii) hold by Lemmata 4.3 and 4.6, respectively, and (i') holds with $c_{R_h} = 1$ since $R_{\boldsymbol{V}_h}$ satisfies $\|B_h\boldsymbol{v}_h\|_{\boldsymbol{V}_h'}^2 = \langle B_h\boldsymbol{v}_h, R_{\boldsymbol{V}_h}^{-1} B_h\boldsymbol{v}_h\rangle$ for all $\boldsymbol{v}_h \in \boldsymbol{U}_h$. The error estimate is finally a consequence of Theorem 2.8 using the statement (c) from Theorem 4.7 and the Lipschitz continuity of C from Lemma 4.5. ∎

6. Numerical example. In this section, we present some numerical results of a lowest-order implementation of our nonlinear DPG scheme (30) for a model problem with and without advective field β. The specific discretization including Fortin operator is presented in the first subsection, and numerical results are reported afterwards.

6.1. Discrete setting and Fortin operator. We use lowest-order test and trial spaces $\boldsymbol{U}_h$, $\boldsymbol{V}_h$ defined by

$$\boldsymbol{U}_h = \mathbb{P}_0(\Omega_h;\mathbb{R}) \times \mathbb{P}_0(\Omega_h;\mathbb{R}^2) \times \mathbb{P}_0(\Omega_h;\mathbb{R}^2) \times \mathbb{P}_1^0(\partial\Omega_h;\mathbb{R}) \times \mathbb{P}_0(\partial\Omega_h;\mathbb{R}), \tag{31a}$$

$$\boldsymbol{V}_h = \mathbb{P}_2(\Omega_h;\mathbb{R}) \times \mathbb{P}_2(\Omega_h;\mathbb{R}^2). \tag{31b}$$

Here, $\mathbb{P}_k(\Omega_h;\mathbb{R}^d)$ denotes the spaces of Ω_h-piecewise d-variate polynomials of degree k (meshes are defined below) and $\mathbb{P}_1^0(\partial\Omega_h;\mathbb{R}) \subset \mathbb{P}_1(\partial\Omega_h;\mathbb{R})$ is the largest subspace of "continuous" functions satisfying the homogeneous Dirichlet condition, that is, $\gamma\big(\mathbb{P}_1^0(\partial\Omega_h;\mathbb{R})\big) \subset H_{00}^{1/2}(\partial\Omega_h)$. The test space $\boldsymbol{V}_h$ is chosen such that there exists a Fortin operator $\Pi : \boldsymbol{V} \to \boldsymbol{V}_h$ satisfying (11), ensuring that (30) is well posed by Theorem 5.1.

LEMMA 6.1. *Let $\boldsymbol{U}_h$ and $\boldsymbol{V}_h$ be defined by (31), and consider $\boldsymbol{\beta} \in \mathbb{P}_0(\Omega_h;\mathbb{R}^2)$. Then, there exists a Fortin operator $\Pi : \boldsymbol{V} \to \boldsymbol{V}_h$ satisfying (11).*

Proof. The statement is consequence of [23, Lemmata 3.2, 3.3]. Indeed, as proved in [23], there are continuous operators $\Pi^g : H^1(\Omega_h) \to \mathbb{P}_2(\Omega_h;\mathbb{R})$ and $\Pi^d : \boldsymbol{H}(\mathrm{div};\Omega_h) \to \mathbb{P}_2(\Omega_h;\mathbb{R}^2)$ such that

$$\langle \gamma(\Pi^g v - v)_{|\partial T}, q\rangle_{\partial T} = 0 \quad \forall q \in \mathbb{P}_0(\{F\}_{F\in\mathcal{F}_T};\mathbb{R}), \tag{32a}$$

$$\big((\Pi^d \boldsymbol{\tau} - \boldsymbol{\tau})_{|T}, \boldsymbol{q}\big)_{\boldsymbol{L}^2(T)} = 0 \quad \forall \boldsymbol{q} \in \mathbb{P}_0(T;\mathbb{R}^2), \tag{32b}$$

$$\langle \gamma_n(\Pi^d_T \boldsymbol{\tau} - \boldsymbol{\tau})_{|\partial T}, \hat{q}\rangle_{\partial T} = 0 \quad \forall \hat{q} \in \mathbb{P}_1(\{F\}_{F\in\mathcal{F}_T};\mathbb{R}) \cap C^0(\partial T) \tag{32c}$$

for all $v \in H^1(\Omega_h)$, $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div};\Omega_h)$, and $T \in \Omega_h$. Here, $\mathcal{F}_T$ denotes the set of faces of an element $T \in \Omega_h$. The operator $\Pi = (\Pi^g, \Pi^d) : \boldsymbol{V} \to \boldsymbol{V}_h$ is continuous and satisfies assumption (11a). In fact, considering $\boldsymbol{u}_h = (u_h, \boldsymbol{\sigma}_h, \boldsymbol{\rho}_h, \hat{u}_h, \hat{\rho}_h) \in \boldsymbol{U}_h$, we have

$$\begin{aligned}\langle B\boldsymbol{i}_h\boldsymbol{u}_h, \boldsymbol{j}_h\Pi\boldsymbol{v} - \boldsymbol{v}\rangle = \big(u_h, \boldsymbol{\beta}\cdot\nabla(\Pi^g(v) - v) + \nabla\cdot(\Pi^d(\boldsymbol{\tau}) - \boldsymbol{\tau})\big)_{\Omega_h} + \big(\boldsymbol{\sigma}_h, \Pi^d\boldsymbol{\tau} - \boldsymbol{\tau}\big)_{\Omega_h}\\ + (\boldsymbol{\rho}_h, \nabla(\Pi^g v - v))_{\Omega_h} - \langle\hat{\rho}_h, \gamma(\Pi^g(v) - v)\rangle_{\partial\Omega_h} - \langle\gamma_n(\Pi^d(\boldsymbol{\tau}) - \boldsymbol{\tau}), \hat{u}_h\rangle_{\partial\Omega_h}.\end{aligned}$$

Since $\boldsymbol{\beta} \in \mathbb{P}_0(\Omega_h;\mathbb{R}^2)$, it follows by integration by parts that

$$\begin{aligned}(u_h, \boldsymbol{\beta}\cdot\nabla(\Pi^g(v) - v))_{\Omega_h} &= -(\nabla\cdot(\boldsymbol{\beta}u_h), \Pi^g(v) - v)_{\Omega_h} + \langle\gamma_n(\boldsymbol{\beta}u_h), \gamma(\Pi^g(v) - v)\rangle_{\partial\Omega_h}\\ &= 0,\end{aligned}$$

owing to (32a). Proceeding similarly for the other terms, we infer (11a). ∎

The nonlinear discrete problem (30) is solved by using the standard Newton method. The initial iterate is defined as the solution of (13) with $\boldsymbol{\lambda}(|\nabla\boldsymbol{u}|) = \boldsymbol{\lambda}(0)$ (it is different from 0 in our experiment). Numerical experiments are performed on a uniform refinement of a two-dimensional triangular mesh sequence $\{\Omega_h\}_h$, indexed by the level of refinement $h := \max_{T\in\Omega_h}|T|$, and satisfying $\mathcal{O}(h) = \mathcal{O}(\dim(\boldsymbol{U}_h)^{-\frac{1}{2}})$. The numerical parameter κ is chosen equal to 1 and the stability parameter $\alpha := \lambda^2_{\sharp/\flat}$. Denoting by $\boldsymbol{u} = (u, \boldsymbol{\sigma}, \boldsymbol{\rho}, \hat{u}, \hat{\rho}) \in \boldsymbol{U}$ and $\boldsymbol{u}_h = (u_h, \boldsymbol{\sigma}_h, \boldsymbol{\rho}_h, \hat{u}_h, \hat{\rho}_h) \in \boldsymbol{U}_h$ the exact and discrete solutions, respectively, we depict the three errors $\|u - u_h\|_{L^2(\Omega)}$, $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\boldsymbol{L}^2(\Omega)}$ and $\|\boldsymbol{\rho} - \boldsymbol{\rho}_h\|_{\boldsymbol{L}^2(\Omega)}$. In addition, from definitions (16) of the trace norms and Lemma 4.1, we depict, instead, the corresponding upper bounds

$$\|\hat{u} - \hat{u}_h\|_{H^{1/2}_{00}(\partial\Omega_h)} \le \|u - \mathcal{I}^g_h(\hat{u}_h)\|_{H^1(\Omega)}, \ \|\hat{\rho} - \hat{\rho}_h\|_{H^{-1/2}(\partial\Omega_h)} \le \|\boldsymbol{\rho} + \boldsymbol{\beta}u - \mathcal{I}^d_h(\hat{\rho}_h)\|_{\boldsymbol{H}(\mathrm{div};\Omega)}.$$

Here, $\mathcal{I}^g_h : \mathbb{P}^0_1(\partial\Omega_h;\mathbb{R}) \to \mathbb{P}_1(\Omega_h;\mathbb{R}) \cap C^0(\Omega)$ denotes the $\mathbb{P}_1$-Lagrange interpolation operator, and $\mathcal{I}^d_h : \mathbb{P}_0(\partial\Omega_h;\mathbb{R}^2) \to \mathbb{RT}_0(\Omega_h)$ is the lowest-order Raviart–Thomas interpolation operator.

6.2. Example with and without advection. We consider the nonlinear example without advection from [6] and a corresponding example with non-zero advection. The exact solution is given by $u(x,y) = \cos(\pi x/2)\cos(\pi y/2)$ on the domain $(-1,1)^2$. The nonlinear diffusive tensor is $\mathbb{R}$-valued and defined by $\lambda(s) = 2 - (1+s)^{-2}$ for $s \ge 0$. Assumptions (14b), (14c) are satisfied with $\lambda_\flat = 1$ and $\lambda_\sharp = 3$. We also consider the case when the advective field $\boldsymbol{\beta}$ is different from zero, and given by $\boldsymbol{\beta}(x,y) = (y, -x)$, satisfying assumption (14a). The approximation errors are depicted in Figure 1.

As expected, all the errors behave like $\mathcal{O}(h)$. Comparing the two plots of Figure 1, the presence of the advective field does not impact the accuracy of our method for this test case. In addition, our method delivers very similar result to those obtained with the approach proposed in [6].

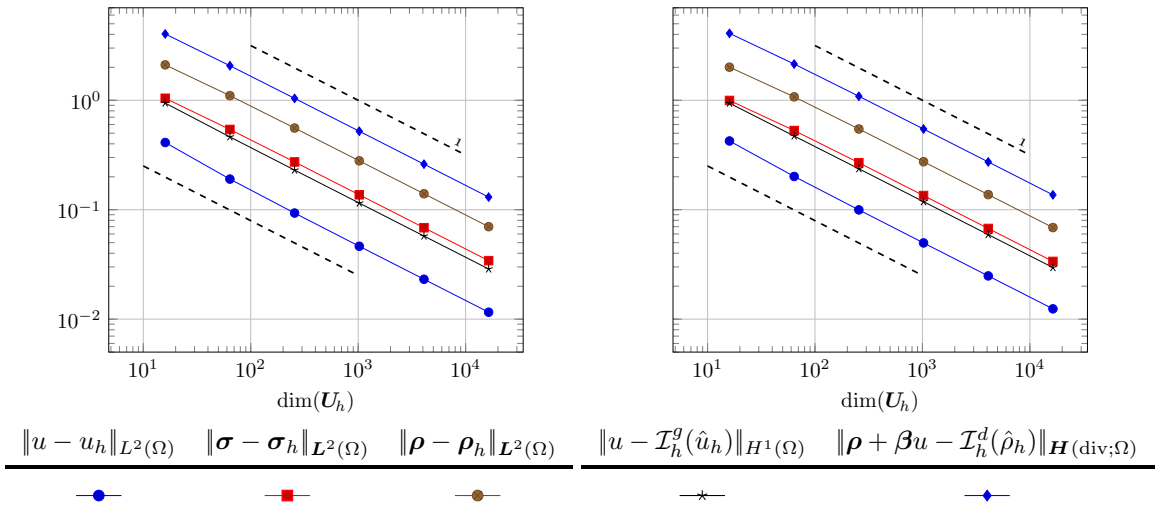

FIG. 1. *Left panel: numerical errors with zero advective field. Right panel: numerical error with the advective field* $\boldsymbol{\beta}(x,y)=(y,-x)$.

[1] J. BRAMWELL, L. DEMKOWICZ, J. GOPALAKRISHNAN, AND W. QIU, *A locking-free hp DPG method for linear elasticity with symmetric stresses*, Numerische Mathematik, 122 (2012), pp. 671–707.

[2] S. C. BRENNER, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM Journal on Numerical Analysis, 41 (2003), pp. 306–324.

[3] D. BROERSEN AND R. STEVENSON, *A robust Petrov-Galerkin discretisation of convection-diffusion equations*, Computers & Mathematics with Applications, 68 (2014), pp. 1605–1618.

[4] D. BROERSEN AND R. STEVENSON, *A Petrov-Galerkin discretization with optimal test space of a mild-weak formulation of convection-diffusion equations in mixed form*, IMA Journal of Numerical Analysis, 35 (2015), pp. 39–73.

[5] T. BUI-THANH AND O. GHATTAS, *A PDE-constrained optimization approach to the discontinuous Petrov-Galerkin method with a trust region inexact Newton-CG solver*, Computer Methods in Applied Mechanics and Engineering, 278 (2014), pp. 20–40.

[6] C. CARSTENSEN, P. BRINGMANN, F. HELLWIG, AND P. WRIGGERS, *Nonlinear discontinuous Petrov–Galerkin methods*, arXiv: 1710.00529, 2017.

[7] C. CARSTENSEN, L. DEMKOWICZ, AND J. GOPALAKRISHNAN, *Breaking spaces and forms for the DPG method and applications including Maxwell equations*, Computers & Mathematics with Applications, 72 (2016), pp. 494–522.

[8] J. CHAN, L. DEMKOWICZ, AND R. MOSER, *A DPG method for steady viscous compressible flow*, Computers & Fluids, 98 (2014), pp. 69–90.

[9] J. CHAN, N. HEUER, T. BUI-THANH, AND L. DEMKOWICZ, *Robust DPG method for convection-dominated diffusion problems II: Adjoint boundary conditions and mesh-dependent test norms*, Computers & Mathematics with Applications, 67 (2014), pp. 771–795.

[10] P. G. CIARLET, M. H. SCHULTZ, AND R. S. VARGA, *Numerical methods of high-order accuracy for nonlinear boundary value problems*, Numerische Mathematik, 13 (1969), pp. 51–77.

[11] A. COHEN, W. DAHMEN, AND G. WELPER, *Adaptivity and variational stabilization for convection-diffusion equations*, ESAIM. Mathematical Modelling and Numerical Analysis, 46 (2012), pp. 1247–1273.

[12] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *A class of discontinuous Petrov-Galerkin methods. Part I: the transport equation*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 1558–1572.

[13] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *Analysis of the DPG method for the Poisson problem*, SIAM Journal on Numererical Analysis, 49 (2011), pp. 1788–1809.

[14] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions*, Numerical Methods for Partial Differential Equations, 27 (2011), pp. 70–105.

[15] L. DEMKOWICZ, J. GOPALAKRISHNAN, S. NAGARAJ, AND P. SEPÚLVEDA, *A spacetime DPG method for the Schrödinger equation*, arXiv: 1610.04678, 2016.

[16] L. Demkowicz and N. Heuer, *Robust DPG method for convection-dominated diffusion problems*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 2514–2537.
[17] V. J. Ervin, T. Führer, N. Heuer, and M. Karkulik, *DPG method with optimal test functions for a fractional advection diffusion equation*, Journal of Scientific Computing, 72 (2017), pp. 568–585.
[18] F. Fuentes, B. Keith, L. Demkowicz, and P. Le Tallec, *Coupled variational formulations of linear elasticity and the DPG methodology*, arXiv: 1609.08160, 2016.
[19] T. Führer, N. Heuer, and M. Karkulik, *On the coupling of DPG and BEM*, Mathematics of Computation, 86 (2017), pp. 2261–2284.
[20] T. Führer, N. Heuer, M. Karkulik, and R. Rodríguez, *Combining the DPG method with finite elements*, Computational Methods in Applied Mathematics. (in press, DOI:10.1515/cmam-2017-0041).
[21] T. Führer, N. Heuer, and E. P. Stephan, *On the DPG method for Signorini problems*, IMA Journal of Numerical Analysis. (in press, DOI:10.1093/imanum/drx048).
[22] V. Girault and P.-A. Raviart, *Finite element methods for Navier-Stokes equations*, vol. 5 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1986. Theory and algorithms.
[23] J. Gopalakrishnan and W. Qiu, *An analysis of the practical DPG method*, Mathematics of Computation, 83 (2014), pp. 537–552.
[24] N. Heuer and M. Karkulik, *Discontinuous Petrov-Galerkin boundary elements*, Numerische Mathematik, 135 (2017), pp. 1011–1043.
[25] N. Heuer and M. Karkulik, *A robust DPG method for singularly perturbed reaction-diffusion problems*, SIAM Journal on Numerical Analysis, 55 (2017), pp. 1218–1242.
[26] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Études mathématiques, Dunod; Gauthier-Villars, Paris, 1969.
[27] I. Muga and K. G. van der Zee, *Discretization of linear problems in Banach spaces: Residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods*, arXiv: 1511.04400, 2015.
[28] N. V. Roberts, T. Bui-Thanh, and L. Demkowicz, *The DPG method for the Stokes problem*, Computers & Mathematics with Applications, 67 (2014), pp. 966–995.
[29] N. V. Roberts, L. Demkowicz, and R. Moser, *A discontinuous Petrov-Galerkin methodology for adaptive solutions to the incompressible Navier-Stokes equations*, Journal of Computational Physics, 301 (2015), pp. 456–483.