A DPG FRAMEWORK FOR STRONGLY MONOTONE OPERATORS
Pierre Cantin, Norbert Heuer

HAL Id: hal-01690281
https://hal.archives-ouvertes.fr/hal-01690281

Submitted on 22 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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.
singly 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 $A : U \rightarrow V'$ with Banach space $U$, Hilbert space $V$, and dual $V'$. A mixed (or saddle-point) form of $A u = F$ is

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

with solution $(u,v) = (u,0)$. Here, $R : V \rightarrow V'$ is the Riesz operator and $A^* : V \rightarrow U'$ 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

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

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 $u$), it is not necessary to consider the variable $v$. 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,

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

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

$$
- \nabla \cdot (\rho + \beta u) = f \quad \text{in} \quad \Omega, \quad u = 0 \quad \text{on} \quad \partial \Omega
$$
with the nonlinear flux \( \rho = \rho(\nabla u) \) 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 \( \rho = \lambda(\nabla u)\nabla u \) 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 \( U \) and \( V \) be two real Hilbert spaces with topological duals \( U' \) and \( V' \), respectively. We consider a bounded linear operator \( B : U \to V' \), an isomorphism \( R : V \to V' \), and a continuous nonlinear operator \( C : U \to U' \). (Later, \( C \) will be assumed to be Lipschitz continuous and strongly monotone in a certain sense.) Then we define the block operator \( T_\kappa : U \times V \to U' \times V' \), with \( \kappa > 0 \), as

\[
T_\kappa = \begin{pmatrix} C & B^* \\ B & -\frac{1}{\kappa} R \end{pmatrix}.
\]

Here, \( B^* : V \to U' \) is the adjoint operator of \( B \), i.e., \( \langle Bu, v \rangle = \langle B^* v, u \rangle \) for all \( u \in U \) and \( v \in V \). 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 \( P_B : U \to \mathcal{N}(B) \) the
that the operator \( T \) needed for the bijectivity of \( U \).

The following theorem provides sufficient conditions on the continuous operators \( B, C \) and \( R \) such that the operator \( T \) 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 \( \|u - P_B(u)\|^2 \leq c_B \|Bu\|^2_V \), for all \( u \in U \),

(iii) there exist \( c_U, c_V > 0 \) such that

\[
  c_U \|P_B(u_1 - u_2)\|^2_U \leq \langle C(u_1) - C(u_2), u_1 - u_2 \rangle + c_V \|B(u_1 - u_2)\|^2_V,
\]

for all \( u_1, u_2 \in U \).

Then, for all \( \kappa \geq \kappa_0 := (c_V + c_U c_B) c_R \), \( T_\kappa \) is bijective with Lipschitz continuous inverse. In particular,

\[
  (3) \quad c_U \|u_1 - u_2\|^2_U \leq \langle C(u_1) - C(u_2), u_1 - u_2 \rangle + \kappa \|R^{-1}B(u_1 - u_2), B(u_1 - u_2)\|
\]

for all \( u_1, u_2 \in U \).

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

\[
  (4) \quad \begin{pmatrix} C + \kappa B^*R^{-1} & 0 \\ 0 & -\kappa^{-1} R \end{pmatrix} = S^*_\kappa T_\kappa S_\kappa^*.
\]

Here, \( S_\kappa^* : U \times V \rightarrow U \times V \) and \( S_\kappa^* : U' \times V' \rightarrow U' \times V' \) are the respective involutory operators (i.e., equal to their inverses) defined as

\[
  S_\kappa^* = \begin{pmatrix} \text{Id}_U & 0 \\ \kappa R^{-1} B & -\text{Id}_V \end{pmatrix}, \quad \text{and} \quad S_\kappa^* = \begin{pmatrix} \text{Id}_U & \kappa B^* R^{-1} \\ 0 & -\text{Id}_V \end{pmatrix}.
\]

It then follows that \( T_\kappa \) is invertible if and only if the operator \( D_\kappa = C + \kappa B^* R^{-1} B : U \rightarrow U' \) is invertible. Since \( B^* R^{-1} B \) defines a linear bounded operator we infer that \( D_\kappa \) is continuous. Let us now prove that \( D_\kappa \) is strongly monotone, specifically that (3) holds. Let \( u_1, u_2 \in U \) and start observing that

\[
  \|u_1 - u_2\|^2_U = \|P_B(u_1 - u_2)\|^2_U + \|u_1 - u_2 - P_B(u_1 - u_2)\|^2_U,
\]

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

\[
  \|B(u_1 - u_2)\|^2_V \leq c_R \|R^{-1}B(u_1 - u_2), B(u_1 - u_2)\|.
\]

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

\[
  c_U \|u_1 - u_2\|^2_U \leq \langle C(u_1) - C(u_2), u_1 - u_2 \rangle + (c_V + c_B c_U) c_R \|R^{-1}B(u_1 - u_2), B(u_1 - u_2)\|.
\]

As a result, \( D_\kappa \) is continuous and strongly monotone for all \( \kappa \geq \kappa_0 := (c_V + c_B c_U) c_R \), i.e., \( D_\kappa \) is invertible for all \( \kappa \geq \kappa_0 \) 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 \( T_\kappa \). As in the linear case, here the invertibility of \( T_\kappa \) does not require the surjectivity of \( B \), but only that the range \( \mathcal{R}(B) \) is closed in \( V' \) (see, e.g., [22, Remark 4.3]).

(ii) In Proposition 2.1, if we assume in addition that \( C \) is monotone, i.e., \( \langle C(u_1) - C(u_2), u_1 - u_2 \rangle \geq 0 \) for all \( u_1, u_2 \in U \), then we infer from (3) that \( T_\kappa \) is bijective with Lipschitz continuous inverse for all \( \kappa \geq 1 \).
Now, given \( F \in V' \) and \( G \in U' \), we consider two variational formulations, of penalized mixed form
\[
(u, v) \in U \times V : \quad T_\kappa(u, v) = (G, F) \quad \text{in } U' \times V',
\]
and the reduced Schur variant
\[
u \in U : \quad \langle D_\kappa u, w \rangle = \kappa \langle R^{-1}F, Bw \rangle + \langle G, w \rangle \quad \forall w \in U,
\]
with
\[
\langle D_\kappa u, w \rangle := \kappa \langle R^{-1}Bu, Bw \rangle + \langle C(u), w \rangle \quad (u, w \in U).
\]
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 \( \kappa_0 \) as before, problems (5) and (6) are uniquely solvable and equivalent. Specifically, if \( (u, v) \in U \times V \) solves (5), then \( u \) solves (6). Conversely, if \( u \in U \) solves (6), then \( (u, \kappa R^{-1}(Bu - F)) \in U \times V \) 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) \( (C(u) - G, v) = 0 \) for all \( v \in N(B) \) implies \( C(u) = G \) in \( U' \),
(ii) \( B : U \rightarrow V' \) is surjective.
Then, \( u \in U \) solves (6) if and only if \( C(u) = G \) in \( U' \) and \( Bu = F \) in \( V' \).

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

### 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 \( T_\kappa \) or \( D_\kappa \) 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 \( U_h \) and \( V_h \) be two (families of) finite-dimensional spaces such that \( U_h \subset U \) and \( V_h \subset V \). Of course, later \( h \) will be a mesh parameter. We denote the canonical injection maps by \( i_h : U_h \rightarrow U \) and \( j_h : V_h \rightarrow V \), with \( i_h^* : U' \rightarrow U_h^* \) and \( j_h^* : V' \rightarrow V_h^* \) the respective adjoints. The discrete spaces \( U_h \) and \( V_h \) are provided with the induced norms \( \| \cdot \|_{U_h} := \| i_h \cdot \|_U \) and \( \| \cdot \|_{V_h} := \| j_h \cdot \|_V \).

**2.2.1. Semi-discrete scheme.** One possibility is to discretize the operator \( D_\kappa \) in the standard way, that is, considering \( D_{\kappa,h}^* : U_{h} \rightarrow U_{h}^* \) defined as
\[
D_{\kappa,h}^* = i_h^* D_\kappa i_h, \quad \text{with } D_\kappa = C + \kappa B^* R^{-1}B.
\]
This discretization still requires to calculate \( R^{-1} \), which is not feasible in practice. In DPG discretizations, \( R \) is the Riesz operator \( R_V : V \rightarrow V' \) and such a semi-discrete scheme is sometimes called ideal DPG method. It is distinguished from the practical variant which includes a discretization of \( R_V^{-1} \), cf. [23].
The operator $D_{\kappa,h}^*$ induces the problem
\begin{equation}
\label{eq:7}
\begin{aligned}
\mathbf{u}_h \in U_h : \quad \langle D_{\kappa,h}^* \mathbf{u}_h, \mathbf{w}_h \rangle = \kappa \langle R^{-1} \mathbf{F}, B \mathbf{u}_h \rangle + \langle \mathbf{G}, i_h \mathbf{w}_h \rangle \quad \forall \mathbf{w}_h \in U_h.
\end{aligned}
\end{equation}

**Theorem 2.5.** Assume that the assumptions from Proposition 2.1 hold true with constants $\kappa_0$ and $c_U$ specified there. Then, for all $\kappa \geq \kappa_0$, $D_{\kappa,h}^*$ is invertible with uniformly Lipschitz continuous inverse, and problem (7) is well posed. In addition, assuming that $C$ is Lipschitz continuous with constant $c_{\text{Lip}}$, we have the quasi-optimal error estimate
\begin{equation*}
\| \mathbf{u} - i_h(\mathbf{u}_h) \|_U \leq (1 + c_U^{-1}(c_{\text{Lip}} + \kappa |B^* R^{-1} B\|_{L^2(U,U')})) \inf_{\mathbf{w}_h \in U_h} \| \mathbf{u} - i_h(\mathbf{w}_h) \|_U.
\end{equation*}

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

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

### 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 $D_{\kappa}$, the full operator $T_\kappa$ as $T_{\kappa,h}$.

\begin{equation}
\label{eq:8}
\begin{aligned}
T_{\kappa,h} = \begin{pmatrix}
C_h & B_{\kappa}^* \\
B_h & -\frac{1}{\kappa} R_{h} 
\end{pmatrix} = \begin{pmatrix}
i_h \\
0 
\end{pmatrix} \begin{pmatrix}T_{\kappa} & i_h \\
0 & j_h \end{pmatrix}.
\end{aligned}
\end{equation}

Applying the Schur factorization to this discrete operator we obtain the fully discrete problem
\begin{equation}
\label{eq:9}
\begin{aligned}
\mathbf{u}_h \in U_h : \quad \langle D_{\kappa,h} \mathbf{u}_h, \mathbf{w}_h \rangle = \kappa \langle R_{h}^{-1} j_h^* \mathbf{F}, B_{h} \mathbf{w}_h \rangle + \langle i_h^* \mathbf{G}, \mathbf{w}_h \rangle \quad \forall \mathbf{w}_h \in U_h,
\end{aligned}
\end{equation}

with
\begin{equation}
\label{eq:10}
\langle D_{\kappa,h} \mathbf{u}_h, \mathbf{w}_h \rangle := \langle C_h(\mathbf{u}_h), \mathbf{w}_h \rangle + \kappa \langle R_{h}^{-1} B_{h} \mathbf{u}_h, B_{h} \mathbf{w}_h \rangle.
\end{equation}

The well-posedness of (9) follows similarly as in Proposition 2.1, by using the existence of a Fortin operator $\Pi : V \to V_h$ satisfying, uniformly in $h$,
\begin{equation}
\label{eq:11a}
\langle B i_h \mathbf{u}_h, \mathbf{v} - j_h \Pi \mathbf{v}_h \rangle = 0 \quad \forall \mathbf{u}_h \in U_h, \mathbf{v} \in V,
\end{equation}
\begin{equation}
\label{eq:11b}
\exists c_{\text{II}} > 0, \quad |j_h \Pi \mathbf{v}|_V^2 \leq c_{\text{II}} |\mathbf{v}|_V^2 \quad \forall \mathbf{v} \in V.
\end{equation}

In the context of DPG methods, Gopalakrishnan and Qiu \cite{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. \cite[Proof of Theorem 2.1]{23}.

**Lemma 2.6.** Assume that (11) holds. Then, $|B i_h \mathbf{u}_h|_V^2 \leq c_{\text{II}} |B \mathbf{u}_h|_{V_h}^2$, for all $\mathbf{u}_h \in U_h$.

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

**Proposition 2.7.** Assume that,
\begin{itemize}
\item [(i')] $R_{h}^{-1}$ is coercive on $\mathcal{R}(B_{h})$ with coercivity constant $c_{R_{h}} > 0$,
\item [(ii)] there exists $c_B > 0$ such that $|\mathbf{u} - P_{B}(\mathbf{u})|^2 \leq c_B |\mathbf{B} \mathbf{u}|_V^2$, for all $\mathbf{u} \in U$,
\item [(iii)] there exist $c_U, c_V > 0$ such that
\begin{equation*}
\begin{aligned}
c_U |P_{B}(\mathbf{u}_1 - \mathbf{u}_2)|_U^2 \leq \langle C(\mathbf{u}_1) - C(\mathbf{u}_2), \mathbf{u}_1 - \mathbf{u}_2 \rangle + c_V |\mathbf{B}(\mathbf{u}_1 - \mathbf{u}_2)|_V^2,
\end{aligned}
\end{equation*}
\end{itemize}

for all $\mathbf{u}_1, \mathbf{u}_2 \in U$, \[\quad \]
(iv) there exists a Fortin operator $\Pi : V \to V_h$ satisfying (11). Then, for all $\kappa \geq \kappa_{0,h} := (c_V + c_B c_U) c_{\Pi} c_{R_h}$, $T_{\kappa,h}$ is invertible with Lipschitz continuous inverse. In particular,

$$
(12) \quad c_V \|u_h - v_h\|^2_{L_h} \leq \langle C_h(u_h) - C_h(v_h), u_h - v_h \rangle + \kappa \langle R^{-1}_h B_h(u_h - v_h), B_h(u_h - v_h) \rangle
$$

for all $u_h, v_h \in U_h$.

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

Let $u_h, v_h \in U_h$. By assumptions (ii), (iv) and Lemma 2.6 we have

$$
\|i_h(u_h - v_h) - P_B i_h(u_h - v_h)\|^2_{U} \leq c_B \|B i_h(u_h - v_h)\|^2_{V'} \leq c_B c_{\Pi} \|B_h(u_h - v_h)\|^2_{V'},
$$

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

$$
\|i_h(u_h - v_h) - P_B i_h(u_h - v_h)\|^2_{U} \leq c_B c_{\Pi} c_{R_h} \langle R^{-1}_h B_h(u_h - v_h), B_h(u_h - v_h) \rangle.
$$

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

$$
c_V \|P_B i_h(u_h - v_h)\|^2_{U} \leq \langle C_h(u_h) - C_h(v_h), u_h - v_h \rangle + c_V \|B i_h(u_h - v_h)\|^2_{V'}.
$$

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

$$
c_V \|u_h - v_h\|^2_{L_h} \leq \langle C_h(u_h) - C_h(v_h), u_h - v_h \rangle + (c_V + c_B c_U) c_{\Pi} c_{R_h} \langle B_h(u_h - v_h), R^{-1}_h B_h(u_h - v_h) \rangle.
$$

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

**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

$$
|u - i_h(u_h)|_U \leq (1 + c_U^{-1}(c_{Lip} + \kappa \|B^* j_h R^{-1}_h j_h^* B\|_{L(U, U')}) \inf_{w_h \in U_h} \|u - i_h(w_h)\|_U.
$$

Here, $u \in U$ and $u_h \in U_h$ are the unique solutions of (6) and (9), respectively.

**Proof.** Let $w_h \in U_h$ and denote $\xi_h = u_h - w_h$. By Proposition 2.7 we have

$$
c_U \|\xi_h\|^2_{L_h} \leq \langle C_h(u_h) - C_h(w_h), \xi_h \rangle + \kappa \langle R^{-1}_h B_h(u_h - w_h), B_h \xi_h \rangle
$$

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

$$
\langle R^{-1}_h j_h^* B u, B_h \xi_h \rangle = \langle R^{-1}_h j_h^* F, B_h \xi_h \rangle \quad \text{and} \quad \langle i_h^* C(u), \xi_h \rangle = \langle i_h^* G, \xi_h \rangle.
$$

hold. Hence, we conclude that

$$
\kappa \langle R^{-1}_h j_h^* B(u - i_h u_h), B_h \xi_h \rangle + \langle C(u) - C(i_h u_h), j_h \xi_h \rangle = 0,
$$
yielding
\[ c_U \|\xi_h\|_{L_h}^2 \leq \langle C(u) - C(i_h w_h), i_h \xi_h \rangle + \kappa \langle B^* j_h R_h^{-1} j_h^* B(u - i_h w_h), i_h \xi_h \rangle. \]

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 \( u : \Omega \rightarrow \mathbb{R} \), with \( \Omega \subset \mathbb{R}^d (d \geq 2) \) a connected Lipschitz domain, the solution of

\[
\begin{align*}
-\nabla \cdot (\lambda(x, |\nabla u(x)|)\nabla u(x) + \beta(x)u(x)) &= f(x) \quad \text{for a.a. } x \in \Omega, \\
 u(x) &= 0 \quad \text{for a.a. } x \in \partial \Omega.
\end{align*}
\]

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

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

- \( \beta \) is Lipschitz continuous on \( \Omega \) and satisfies
  \[
  \text{ess inf}_\Omega (-\nabla \cdot \beta) \geq 0.
  \]
- There exist constants \( 0 < \lambda_0 \leq \lambda_2 \) such that, for all \( \sigma, \theta \in L^2(\Omega) \),
  \[
  \lambda_0 |\sigma - \theta|_{L^2(\Omega)}^2 \leq \langle \lambda(|\sigma|)|\sigma - \lambda(|\theta|)\theta, \sigma - \theta \rangle_{L^2(\Omega)},
  \]
  \[
  \lambda_2 |\sigma - \theta|_{L^2(\Omega)} \geq |\lambda(|\sigma|)|\sigma - \lambda(|\theta|)\theta|_{L^2(\Omega)}.
  \]

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 \( L^2(\Omega) \) be the standard Lebesgue spaces collecting \( \mathbb{R} \)-valued and \( \mathbb{R}^d \)-valued functions, respectively, satisfying

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

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

\[
(v, w)_{H^1(\Omega)} := (v, w)_\Omega + \ell^2_\Omega(\nabla v, \nabla w)_\Omega \quad \forall v, w \in H^1(\Omega),
\]

\[
(\tau, \eta)_{H(\text{div}; \Omega)} := (\tau, \eta)_\Omega + \ell^2_\Omega(\nabla \tau, \nabla \eta)_\Omega \quad \forall \tau, \eta \in H(\text{div}; \Omega).
\]

We also denote by \( H^1_0(\Omega) \) the closure of the space collecting infinitely differentiable functions with compact support in \( \Omega \) 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 \( \ell_\Omega \) is chosen such that

\[
|\beta|_{L^\infty(\Omega)} \ell_\Omega \lambda_2^{-1} = 1.
\]
3.2. Mesh partition and product spaces. Let $\Omega_h$ be a non-overlapping partition of $\Omega$ composed of open elements $T \in \Omega_h$ with Lipschitz boundary $\partial T$ and outwardly oriented by $n_T$. Let $H^1(\Omega_h)$ and $H(\text{div}; \Omega_h)$ be the product or “broken” Sobolev spaces equipped with inner products

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

Here, $(\cdot, \cdot)_{\Omega_h}$ 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 $H(\text{div}; T)$, respectively. They are dual to each other. Traces on the mesh skeleton $\partial \Omega_h$ are defined with the trace maps $\gamma : H^1(\Omega_h) \rightarrow \times_{T \in \Omega_h} H^{1/2}(\partial T)$ and $\gamma_n : H(\text{div}; \Omega_h) \rightarrow \times_{T \in \Omega_h} H^{-1/2}(\partial T)$, defined as

$$\gamma(u)_{\partial T} = u_{|\partial T} \quad \forall T \in \Omega_h \quad \forall u \in H^1(\Omega_h),$$
$$\gamma_n(\rho)_{\partial T} = (\rho n_T)_{|\partial T} \quad \forall T \in \Omega_h \quad \forall \rho \in H(\text{div}; \Omega_h).$$

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

We also introduce the trace spaces

$$H^{1/2}_{\text{div}}(\partial \Omega_h) := \left\{ \hat{\rho} \in \times_{T \in \Omega_h} H^{1/2}(\partial T) \mid \exists \rho \in H(\text{div}; \Omega), \hat{\rho} = \gamma_n(\rho) \right\},$$

$$H^{-1/2}(\partial \Omega_h) := \left\{ \hat{v} \in \times_{T \in \Omega_h} H^{-1/2}(\partial T) \mid \exists v \in H^1(\Omega), \hat{v} = \gamma(v) \right\},$$

equipped with their respective quotient norms,

$$\|\hat{v}\|_{H^{1/2}(\partial \Omega_h)} := \inf_{w \in H^1_0(\Omega)} \left\{ \|w\|_{H^1(\Omega)} ; \gamma(w) = \hat{v} \right\} \quad \forall \hat{v} \in H^{1/2}_{\text{div}}(\partial \Omega_h),$$

$$\|\hat{\rho}\|_{H^{-1/2}(\partial \Omega_h)} := \inf_{\rho \in H(\text{div}; \Omega)} \left\{ \|\rho\|_{H(\text{div}; \Omega)} ; \gamma_n(\rho) = \hat{\rho} \right\} \quad \forall \hat{\rho} \in H^{-1/2}(\partial \Omega_h).$$

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 $\ell_\Omega$.

**Lemma 3.1.** We have

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

with $c_{PS,\Omega}^{-1} = 2(1 + c_{PS,0})$ and $c_{PS,0} > 0$ the Poincaré-Steklov constant in $H^1_0(\Omega)$ satisfying $\|\xi\|^2_{L^2(\Omega)} \leq c_{PS,0} \ell^2_\Omega \|\nabla \xi\|^2_{L^2(\Omega)}$ 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 $-\ell^2_\Omega \Delta \xi = v$. It holds

$$\|v\|^2_{L^2(\Omega)} = \ell^2_\Omega (\nabla v, \nabla \xi)_{\Omega_h} - \ell^2_\Omega (\gamma_n(\nabla \xi), \gamma(v))_{\partial \Omega_h}$$

$$\leq \ell^2_\Omega \|\nabla v\|_{L^2(\Omega_h)} \|\nabla \xi\|_{L^2(\Omega)} + \ell^2_\Omega \|\gamma_n(\nabla \xi)\|_{H^{-1/2}(\partial \Omega_h)} \sup_{\tau \in H^{-1/2}(\partial \Omega_h)} \frac{\langle \tau, \gamma(v) \rangle_{\partial \Omega_h}}{\|\tau\|^2_{H^{-1/2}(\partial \Omega_h)}}.$$
Furthermore, \( \ell_1^2 \| \nabla \xi \|^2_{L^2(\Omega)} \leq c_{PS,0} \| v \|^2_{L^2(\Omega)} \) by the standard Poincaré-Steklov inequality in \( H_0^1(\Omega) \).

Hence, observing that \( \| \gamma_n(\nabla \xi) \|^2_{H^{-1/2}(\partial \Omega_h)} \leq (1 + c_{PS,0}) \ell_2^{-2} \| v \|^2_{L^2(\Omega)} \), the statement follows. \( \Box \)

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

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

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 \( \rho = \lambda(\sigma) \sigma. \)

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 \( u \in L^2(\Omega), \sigma \in L^2(\Omega), \rho \in L^2(\Omega), \hat{u} \in H^{1/2}(\partial \Omega_h), \hat{\rho} \in H^{-1/2}(\partial \Omega_h) \) such that

\[
\begin{align}
\text{(17a)} \quad (u, \beta \nabla v)_{\Omega_h} + (\rho, \nabla v)_{\Omega_h} - \langle \hat{\rho}, \gamma(v) \rangle_{\partial \Omega_h} &= (f, v)_{\Omega_h} \quad \forall v \in H^1(\Omega_h), \\
\text{(17b)} \quad (\sigma, \tau)_{\Omega_h} + (u, \nabla \cdot \tau)_{\Omega_h} - \gamma_n(\tau), \hat{u})_{\partial \Omega_h} &= 0 \quad \forall \tau \in H(\text{div}; \Omega_h),
\end{align}
\]

and denote the spaces

\[
\begin{align}
U &:= L^2(\Omega) \times L^2(\Omega) \times L^2(\Omega) \times H^{1/2}(\partial \Omega_h) \times H^{-1/2}(\partial \Omega_h), \\
V &:= H^1(\Omega_h) \times H(\text{div}; \Omega_h).
\end{align}
\]

Define the operator \( B : U \to V' \) as

\[
\langle Bu, v \rangle := (u, \beta \nabla v + \nabla \cdot \tau)_{\Omega_h} + (\sigma, \tau)_{\Omega_h} + (\rho, \nabla v)_{\Omega_h} - \langle \hat{\rho}, \gamma(v) \rangle_{\partial \Omega_h} - \langle \gamma_n(\tau), \hat{u} \rangle_{\partial \Omega_h}
\]

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

\[
\text{(19)} \quad u \in U : \quad Bu = F \in V',
\]

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

**Lemma 4.1.** Let \( u = (u, \sigma, \rho, \hat{u}, \hat{\rho}) \in U \) be a solution of (19). Then, \( u \in H_0^1(\Omega), \rho \in H(\text{div}; \Omega), \) and \( \sigma = \nabla u, -\nabla \cdot (\rho + \beta u) = f, \hat{u} = \gamma(u) \) and \( \hat{\rho} = \gamma_n(\rho + \beta u). \)

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

\[
\| u \|^2_U := \| u \|^2_{L^2(\Omega)} + \ell_1^2 \| \sigma \|^2_{L^2(\Omega)} + \ell_2^2 \lambda_n^{-2} \| \rho \|^2_{L^2(\Omega)} + \| \hat{u} \|^2_{H^{1/2}(\partial \Omega_h)} + \ell_2^2 \lambda_n^{-2} \| \hat{\rho} \|^2_{H^{-1/2}(\partial \Omega_h)}.
\]
Similarly, the space $V$ is equipped with the inner product $(\cdot, \cdot)_V$ defined for $v = (v, \tau), w = (w, \eta) \in V$ as

$$(v, w)_V := \lambda_0^2 \ell_{\Omega}^{-4} (v, w)_{H^1(\Omega_h)} + \ell_{\Omega}^{-2} (\tau, \eta)_{H(\text{div}; \Omega_h)}.$$  

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

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

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

$$|B^* v|_{U'}^2 \geq \|\beta \nabla v + \nabla \cdot \tau\|_{L^2(\Omega_h)}^2 + \ell_{\Omega}^{-2} \|\tau\|_{L^2(\Omega)}^2 + \ell_{\Omega}^{-2} \lambda_0^2 \|\nabla v\|_{L^2(\Omega_h)}^2$$  

$$+ \frac{1}{2} \ell_{\Omega}^{-2} \lambda_0^2 \left( \sup_{\tau \in H^{1/2}(\partial \Omega_h)} \frac{\langle \tau, \nabla v \rangle_{\partial \Omega_h}}{\|\tau\|_{H^{1/2}(\partial \Omega_h)}} \right)^2.$$  

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

$$|B^* v|_{U'}^2 \geq \|\beta \nabla v + \nabla \cdot \tau\|_{L^2(\Omega_h)}^2 + \ell_{\Omega}^{-2} \|\tau\|_{L^2(\Omega)}^2 + \frac{1}{2} \ell_{\Omega}^{-2} \lambda_0^2 \|\nabla v\|_{L^2(\Omega_h)}^2 + \frac{1}{2} c_{PS} \ell_{\Omega}^{-4} \lambda_0^2 \|v\|_{H^1(\Omega_h)}^2.$$  

Hence, observing that $\|\tau\|_{H(\text{div}; \Omega_h)}^2 \leq \|\tau\|_{L^2(\Omega)}^2 + \ell_{\Omega}^{-2} \|\nabla \cdot \tau + \beta \nabla v\|_{L^2(\Omega_h)}^2 + \lambda_0^2 \|\nabla v\|_{L^2(\Omega_h)}^2$ by the triangle inequality and assumption (15), it follows that

$$|B^* v|_{U'}^2 \geq \frac{1}{2} \ell_{\Omega}^{-2} \|\tau\|_{H(\text{div}; \Omega_h)}^2 + \frac{1}{2} c_{PS} \ell_{\Omega}^{-4} \lambda_0^2 \|v\|_{H^1(\Omega_h)}^2.$$  

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

Denoting by $P_B : U \to \mathcal{N}(B)$ the projector of $U$ 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.** $|u - P_B u|_{U'}^2 \leq c_B \|Bu\|_V^2$, for all $u \in U$, 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 $B : U \to V'$ and the boundedness below of $B^*, B : \mathcal{N}(B)^\perp \to V'$ is injective with closed range, i.e., $B^* : V \to (\mathcal{N}(B)^\perp)'$ is surjective, so that

$$\|Bu\|_{V'}^2 = \sup_{v \in V} \frac{\langle B^* v, u - P_B u \rangle_{V'}}{\|v\|_V} \geq c_B^{1/2} \sup_{v \in V} \frac{\langle B^* v, u - P_B u \rangle_{V'}}{\|B^* v\|_{U'}} = c_B^{1/2} \|u - P_B u\|_U$$  

for any $u \in U$.  

The non-trivial kernel $\mathcal{N}(B)$ can be represented as the image of the map $E : H^1_0(\Omega) \times H \to U$ defined, for all $\psi \in H^1_0(\Omega)$ and $\eta \in H = \{v \in H(\text{div}; \Omega) \mid \nabla \cdot v = 0\}$, as

$$E(\psi, \eta) = \left( \psi, \nabla \psi, \eta - \beta \psi, \gamma(\psi), \gamma_n(\eta) \right).$$  

(20)

**Lemma 4.4.** The map $E : H^1_0(\Omega) \times H \to U$ satisfies

$$c_E \|E(\psi, \eta)\|_U^2 \leq \ell_{\Omega}^2 \|\nabla \psi\|_{L^2(\Omega)}^2 + \ell_{\Omega}^{-2} \lambda_0^{-2} \|\eta - \beta \psi\|_{L^2(\Omega)}^2 \quad \forall \psi \in H^1_0(\Omega), \eta \in H,$$  

where $c_E$ is the Poincaré-Steklov constant from Lemma 3.1.
with $c_E^{-1} = 4c_{PS,0} + 3$ and $c_{PS,0}$ the Poincaré-Steklov constant defined in Lemma 3.1. In addition, $E(H^1_0(\Omega) \times H) = N(B)$.

Proof. Estimate (21) is consequence of the Poincaré-Steklov inequality in $H^1_0(\Omega)$ and the triangle inequality. Let us prove that $|\gamma| > \lambda$. Observing that the Cauchy–Schwarz inequality and assumption (14c) yield:

$$E \leq 4 - \gamma(\psi) \eta \|\alpha\lambda \eta - \eta\|_{\Omega}.$$

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

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

so that $\gamma(u) = \bar{u} \in H^{1/2}(\partial \Omega_h)$, we infer that $u \in H^1(\Omega)$. Proceeding similarly by testing with $v = (v, 0)$, we infer that $\beta u + \rho \in H(\mathbf{div}; \Omega_h)$ so that $\beta u + \rho \in H(\mathbf{div}; \Omega)$, and then $\beta u + \rho \in H$. As a result, there exists $\eta \in H$ such that $\rho = -\beta u + \eta$. To conclude, if $u \in N(B)$, then $u \in H^1(\Omega_h)$, $\rho + \beta u \in H$ and $u = E(u, \rho + \beta u)$.

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

$$\langle C(u), v \rangle := \ell_2 \lambda_2^{-2} \left( \lambda(|\mathbf{\sigma}|)\mathbf{\sigma} - \rho, \alpha \lambda \theta - \eta \right)_{\Omega} \quad \forall u, v \in U,$$

with $\pi(u) = (\mathbf{\sigma}, \rho)$ and $\pi(v) = (\theta, \eta)$. Here, $\alpha > 0$ denotes a stability parameter that will be chosen greater than $\lambda_{2}/\lambda_{3}$, with $\lambda_{3}/\lambda_{2}$ the diffusive anisotropy ratio.

We start by establishing the Lipschitz continuity of $C$.

Lemma 4.5. $|C(u_1) - C(u_2)| \leq c_{lip}|u_1 - u_2|$ for all $u_1, u_2 \in U$, with $c_{lip} = 2\max(\alpha \lambda^{-1}, 1)$.

Proof. Let $u_1, u_2, v \in U$ such that $\pi(u_1) = (\sigma_1, \rho_1), \pi(u_2) = (\sigma_2, \rho_2)$ and $\pi(v) = (\theta, \eta)$. Owing to (22), it follows that

$$|\langle C(u_1) - C(u_2), v \rangle| \leq \ell_2 \lambda_2^{-2} \left| \left( \lambda(|\sigma_1|)\sigma_1 - \lambda(|\sigma_2|)\sigma_2, \alpha \lambda \theta - \eta \right)_{\Omega} \right| + \ell_2 \lambda_2^{-2} \left| \left( \rho_1 - \rho_2, \alpha \lambda \theta - \eta \right)_{\Omega} \right|.$$

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

$$|\langle C(u_1) - C(u_2), v \rangle| \leq \ell_2 \lambda_2^{-1} \left( \|\sigma_1 - \sigma_2\|_{L^2(\Omega)} + \lambda_2^{-1}\|\rho_1 - \rho_2\|_{L^2(\Omega)} \right) \|\alpha \lambda \theta - \eta\|_{L^2(\Omega)}.$$

Observing that $\|\alpha \lambda \theta - \eta\|_{L^2(\Omega)} \leq 2\lambda \lambda^{-1} \max(\lambda_2^{-1}, 1)\|v\|_U$, the desired estimate follows.

The next lemma verifies assumption (iii) of Proposition 2.1 for the model problem.
Lemma 4.6. Assume that $\alpha > \lambda_{l/p}^2$. Then

$$c_U |P_B(u_1 - u_2)|^2_U \leq \langle C(u_1) - C(u_2), u_1 - u_2 \rangle + c_V |B(u_1 - u_2)|^2_V, \quad \forall u_1, u_2 \in U,$$

where $c_U$ and $c_V$ are given by

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

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

Proof. Let $u_1, u_2 \in U$ be such that $\pi(u_1) = (\sigma_1, \rho_1)$ and $\pi(u_2) = (\sigma_2, \rho_2)$. By Lemma 4.4 there exist $\psi_1, \psi_2 \in H_0^1(\Omega)$ and $\eta_1, \eta_2 \in H$ such that

$$P_B(u_1) = E(\psi_1, \eta_1) = (\psi_1, \nabla \psi_1, \eta_1 - \beta \psi_1, \gamma(\psi_1), \gamma_n(\eta_1)), \quad \text{and} \quad P_B(u_2) = E(\psi_1, \eta_1) = (\psi_2, \nabla \psi_2, \eta_2 - \beta \psi_2, \gamma(\psi_2), \gamma_n(\eta_2)).$$

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

$$c_E \|P_B(u_1 - u_2)\|^2_U \leq \ell_{\Omega}^2 \|\nabla \psi_1 - \nabla \psi_2\|^2_{L^2(\Omega)} + \ell_{\Omega}^2 \lambda_2^{-2} \|\eta_1 - \eta_2\|^2_{L^2(\Omega)},$$

where we have denoted $\eta_i = \eta_i - \beta \psi_i$ for $i = 1, 2$. By the triangle inequality we have

$$c_E \|P_B(u_1 - u_2)\|^2_U \leq \ell_{\Omega}^2 \|\nabla \psi_1 - \nabla \psi_2\|^2_{L^2(\Omega)} + \ell_{\Omega}^2 \lambda_2^{-2} \|\eta_1 - \eta_2\|^2_{L^2(\Omega)} + \ell_{\Omega}^2 \lambda_2^{-2} \|\rho_1 - \rho_2\|^2_{L^2(\Omega)}.$$

Observe that $\|\sigma_1 - \sigma_2 - (\nabla \psi_1 - \nabla \psi_2)\|^2_{L^2(\Omega)} \leq c_B \ell_{\Omega}^{-2} \|B(u_1 - u_2)\|^2_V$, and $\|\rho_1 - \rho_2 - (\eta_1 - \eta_2)\|^2_{L^2(\Omega)} \leq c_B \lambda_2^4 \ell_{\Omega}^{-2} \|B(u_1 - u_2)\|^2_V$, owing to Lemma 4.3, we obtain

$$c_E \|P_B(u_1 - u_2)\|^2_U \leq c_B \|B(u_1 - u_2)\|^2_V + \ell_{\Omega}^2 \|\sigma_1 - \sigma_2\|^2_{L^2(\Omega)} + \ell_{\Omega}^2 \lambda_2^{-2} \|\rho_1 - \rho_2\|^2_{L^2(\Omega)}.$$

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

$$c \left( \ell_{\Omega}^2 \|\sigma_1 - \sigma_2\|^2_{L^2(\Omega)} + \ell_{\Omega}^2 \lambda_2^{-2} \|\rho_1 - \rho_2\|^2_{L^2(\Omega)} \right) \leq \langle C(u_1) - C(u_2), u_1 - u_2 \rangle + c' \|B(u_1 - u_2)\|^2_V.$$

Let $F(u_1, u_2) = \ell_{\Omega}^{-2} \lambda_2^2 (C(u_1) - C(u_2), u_1 - u_2)$ and observe that

$$F(u_1, u_2) = \alpha \lambda_2 \langle \lambda(|\sigma_1|) \sigma_1 - \lambda(|\sigma_2|) \sigma_2, \sigma_1 - \sigma_2 \rangle_{\Omega} + \langle \rho_1 - \rho_2, \rho_1 - \rho_2 \rangle_{\Omega}$$

$$- \langle \lambda(|\sigma_1|) \sigma_1 - \lambda(|\sigma_2|) \sigma_2, \rho_1 - \rho_2 \rangle_{\Omega} - \alpha \lambda_2 \langle \rho_1 - \rho_2, \sigma_1 - \sigma_2 \rangle_{\Omega}.$$

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

$$F(u_1, u_2) \geq \alpha \lambda_2^2 \|\sigma_1 - \sigma_2\|^2_{L^2(\Omega)} + \|\rho_1 - \rho_2\|^2_{L^2(\Omega)}$$

$$- \lambda_2 \|\sigma_1 - \sigma_2\|_{L^2(\Omega)} \|\rho_1 - \rho_2\|_{L^2(\Omega)} - \alpha \lambda_2 \langle \rho_1 - \rho_2, \sigma_1 - \sigma_2 \rangle_{\Omega}.$$
Hence, applying Young’s inequality, we conclude that

\[ F(u_1, u_2) \geq \left( \alpha \lambda^2 - \frac{\lambda^2}{2} \right) |\sigma_1 - \sigma_2|^2_{L^2(\Omega)} + \frac{1}{2} |\rho_1 - \rho_2|^2_{L^2(\Omega)} - \alpha \lambda \langle \rho_1 - \rho_2, \sigma_1 - \sigma_2 \rangle_{\Omega}. \]

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

\[ (\rho_1 - \rho_2, \sigma_1 - \sigma_2)_{\Omega} = (\rho_1 - \rho_2, \sigma_1 - \sigma_2 - (\nabla \psi_1 - \nabla \psi_2)) \Omega + (\eta_1 - \eta_2, \nabla \psi_1 - \nabla \psi_2)_{\Omega} + (\rho_1 - \rho_2 - (\eta_1 - \eta_2), \nabla \psi_1 - \nabla \psi_2)_{\Omega} =: T_1 + T_2 + T_3. \]

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

\[ |T_1| \leq c_B^{1/2} |\rho_1 - \rho_2|_{\Omega} \|\nabla \psi_1 - \nabla \psi_2\|_{L^2(\Omega)}, \]

and

\[ |T_3| \leq c_B^{1/2} \lambda^2 \eta_{\Omega}^{-1} \|B(u_1 - u_2)\|_{\Omega} \|\nabla \psi_1 - \nabla \psi_2\|_{L^2(\Omega)} \leq c_B^{1/2} \lambda^2 \eta_{\Omega}^{-1} \|B(u_1 - u_2)\|_{\Omega} \|\psi_1 - \psi_2\|_{L^2(\Omega)} + c_B \lambda \eta_{\Omega}^{-1} \|B(u_1 - u_2)\|_{\Omega}. \]

Recalling the definition \( \eta_i = -\beta \psi_{i} + \eta_i, i = 1, 2, \) with \( \eta_i \in H = (\nabla H^1(\Omega))^{\perp} \), we infer that

\[ T_2 = -(\beta \psi_1 - \beta \psi_2, \nabla \psi_1 - \nabla \psi_2)_{\Omega} \leq 0 \]

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

\[
\left( \alpha \lambda^2 - \frac{\lambda^2}{2} \right) |\sigma_1 - \sigma_2|^2_{L^2(\Omega)} + \frac{1}{2} |\rho_1 - \rho_2|^2 \leq F(u_1, u_2) + c_B \alpha \lambda^2 \lambda_{\Omega}^{-2} \|B(u_1 - u_2)\|_{\Omega} \|\nabla \psi_1 - \nabla \psi_2\|_{L^2(\Omega)} + c_B \lambda \eta_{\Omega}^{-1} \|B(u_1 - u_2)\|_{\Omega} \|\psi_1 - \psi_2\|_{L^2(\Omega)}. \]

Young’s inequality and assumption \( \alpha > \lambda^2_{\Omega} \) yield

\[
\frac{1}{2} \left( \alpha \lambda^2 - \frac{\lambda^2}{2} \right) |\sigma_1 - \sigma_2|^2_{L^2(\Omega)} + \frac{1}{2} |\rho_1 - \rho_2|^2 \leq F(u_1, u_2) + \frac{5}{2} c_B \lambda^2 \lambda_{\Omega}^{-2} \alpha^2 \|B(u_1 - u_2)\|_{\Omega}^2 \|\psi_1 - \psi_2\|_{L^2(\Omega)}. \]

As a result, multiplying by \( \lambda_{\Omega}^{-2} \), we obtain

\[
\frac{1}{2} \min \left( \frac{1}{2}, \alpha \lambda_{\Omega}^{-2} - 1 \right) \left( |\sigma_1 - \sigma_2|^2_{L^2(\Omega)} + \lambda_{\Omega}^{-2} |\rho_1 - \rho_2|^2_{L^2(\Omega)} \right) \leq \lambda_{\Omega}^{-2} F(u_1, u_2) + \frac{5}{2} c_B \lambda^2 \lambda_{\Omega}^{-2} \alpha^2 \|B(u_1 - u_2)\|_{\Omega}^2 \|\psi_1 - \psi_2\|_{L^2(\Omega)}. \]

The statement follows multiplying (24) by \( \frac{1}{2} \min \left( \frac{1}{2}, \alpha \lambda_{\Omega}^{-2} - 1 \right) \).
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 \( R_V : V \to V' \) be the Riesz operator, i.e., \( \langle R_V v, w \rangle = (v, w)_V \) for \( v, w \in V \), and, for \( \kappa \geq 0 \), let \( T_\kappa : U \times V \to U' \times V' \) be the nonlinear operator

\[
T_\kappa := \begin{pmatrix} C & B^* \\ B & -\frac{1}{\kappa} R_V \end{pmatrix}.
\]

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

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

\[
(27) \quad (u, v) \in U \times V: \quad T_\kappa(u, v) = (0, F) \quad \text{in} \quad U' \times V'
\]

and

\[
(28) \quad u \in U: \quad \kappa(R_V^{-1} Bu, Bv) + \langle C(u), v \rangle = \kappa(R_V^{-1} F, Bv) \quad \forall v \in U.
\]

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

\[
(29) \quad u \in U: \quad \kappa(Bu, \Theta v) + \langle C(u), v \rangle = \kappa(F, \Theta v) \quad \forall v \in U.
\]

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

Selecting \( \kappa \) sufficiently large, both problems are well posed and equivalent.

**Theorem 4.7.** Assume that \( \alpha > \lambda_{2/3}^2 \) 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) \( T_\kappa : U \times V \to U' \times V' \) defined by (26) is invertible with Lipschitz continuous inverse,

(c) the solution \( u \in U \) of (29) satisfies \( Bu = F \) in \( V' \) and \( C(u) = 0 \) in \( U' \).

**Proof.** Statement (a) is a direct consequence of Theorem 2.3 and the definition of \( \Theta \). The second statement (b) follows from Proposition 2.1. Indeed, \( \langle Bv, R_V^{-1} Bv \rangle = \|Bv\|_{V'}^2 \) for all \( v \in V \), 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

\[
\langle C(u), v \rangle = 0 \quad \forall v \in \mathcal{N}(B) \quad \implies \quad C(u) = 0 \quad \text{in} \quad U'.
\]

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

\[
\langle C(u), v \rangle_{U' \times U} = -\epsilon_\Omega^2 \lambda_-^2 (\lambda(|\sigma|) \sigma - \rho, \eta)_\Omega = 0,
\]

since \( \Theta v = R_V^{-1} Bv = 0 \). Therefore, \( \lambda(|\sigma|) \sigma - \rho \in H^1_0 = \nabla H_0^1(\Omega) \) since \( \nabla H_0^1(\Omega) \) is closed in \( L^2(\Omega) \). That is, there exists \( \psi \in H_0^1(\Omega) \) such that \( \nabla \psi = \lambda(|\sigma|) \sigma - \rho \). Next, choosing \( v = (\varphi, \nabla \varphi, -\beta \varphi, \gamma(\varphi), 0) \in \mathcal{N}(B) \) with \( \varphi \in H^1_0(\Omega) \), it follows that \( (\nabla \psi, \alpha \lambda \nabla \varphi + \beta \varphi)_\Omega = 0 \) for all \( \varphi \in H^1_0(\Omega) \). As a result, owing to assumption (14a), we infer that \( \psi = 0 \), and then \( \lambda(|\sigma|) \sigma = \rho \), so that \( C(u) = 0 \) in \( U' \). \( \Box \)
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 $U_h$ and $V_h$ with $U_h \subset U$ and $V_h \subset V$. Our discrete problem is

\begin{equation}
(30) \quad \mathbf{u}_h \in U_h : \langle C_h(\mathbf{u}_h), \mathbf{w}_h \rangle + \kappa \langle B_h \mathbf{u}_h, \Theta_h \mathbf{w}_h \rangle = \kappa \langle j_h^* F, \Theta_h \mathbf{w}_h \rangle \quad \forall \mathbf{v}_h \in V_h,
\end{equation}

with

\[ C_h := i_h^* C i_h, \quad B_h := j_h^* B i_h, \quad R_{V_h} = j_h^* R_{V} j_h, \quad \text{and} \quad \Theta_h = R_{V_h}^{-1} B_h. \]

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

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

\[ |u - i_h(u_h)|_U \leq (1 + c_{\text{Lip}}^{-1}(c_B + \kappa \|B^* j_h R_{V}^{-1} j_h^* B\|_{L(U,V)})) \inf_{w_h \in U_h} \|u - i_h(w_h)\|_U, \]

with $c_{\text{Lip}} > 0$ defined in Lemma 4.5. Here, $u \in U$ and $u_h \in U_h$ 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 $R_{V_h}$ satisfies $\|B_h \mathbf{v}_h\|_{V_h} = \langle B_h \mathbf{v}_h, R_{V_h}^{-1} B_h \mathbf{v}_h \rangle$ for all $\mathbf{v}_h \in U_h$. 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. $\square$

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 $\beta$. 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 $U_h$, $V_h$ defined by

\begin{align}
(31a) \quad U_h &= P_0(\Omega_h; \mathbb{R}) \times P_0(\Omega_h; \mathbb{R}^2) \times P_0(\Omega_h; \mathbb{R}^2) \times P_0(\partial \Omega_h; \mathbb{R}) \times P_0(\partial \Omega_h; \mathbb{R}), \\
(31b) \quad V_h &= P_2(\Omega_h; \mathbb{R}) \times P_2(\Omega_h; \mathbb{R}^2).
\end{align}

Here, $P_k(\Omega_h; \mathbb{R}^d)$ denotes the spaces of $\Omega_h$-piecewise $d$-variate polynomials of degree $k$ (meshes are defined below) and $P_0^1(\partial \Omega_h; \mathbb{R}) \subset P_1(\partial \Omega_h; \mathbb{R})$ is the largest subspace of “continuous” functions satisfying the homogeneous Dirichlet condition, that is, $\gamma(\partial \Omega_h) \subset H^{1/2}_{00}(\partial \Omega_h)$. The test space $V_h$ is chosen such that there exists a Fortin operator $\Pi : V \to V_h$ satisfying (11), ensuring that (30) is well posed by Theorem 5.1.

**Lemma 6.1.** Let $U_h$ and $V_h$ be defined by (31), and consider $\beta \in P_0(\Omega_h; \mathbb{R}^2)$. Then, there exists a Fortin operator $\Pi : V \to V_h$ satisfying (11).
Proof. The statement is consequence of [23, Lemmata 3.2, 3.3]. Indeed, as proved in [23], there are continuous operators \( \Pi^g : H^1(\Omega_h) \to P_2(\Omega_h; \mathbb{R}) \) and \( \Pi^d : H(\text{div}; \Omega_h) \to P_2(\Omega_h; \mathbb{R}^2) \) such that

\[
\mathbf{B}_h \mathbf{u}_h, \mathbf{j}_h \Pi v - v = (u_h, \beta \nabla (\Pi^g(v) - v) + \nabla \cdot (\Pi^d(\tau) - \tau))_{\Omega_h} + (\sigma_h, \Pi^d \tau - \tau)_{\Omega_h} + \beta \nabla (\Pi^g(v) - v)_{\Omega_h} = 0,
\]

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 \( \lambda = 3 \). We consider the nonlinear example without advection (6.2). Comparing the two plots of Figure 1, the presence of the advective field \( \beta \) 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].

REFERENCES
\[
\| u - u_h \|_{L^2(\Omega)} \quad \| \sigma - \sigma_h \|_{L^2(\Omega)} \quad \| \rho - \rho_h \|_{L^2(\Omega)} \quad \| u - T^\rho_h(\hat{u}_h) \|_{H^1(\Omega)} \quad \| \rho + \beta u - T^\rho_h(\hat{\rho}_h) \|_{H(\text{div};\Omega)}
\]

Fig. 1. Left panel: numerical errors with zero advective field. Right panel: numerical error with the advective field \( \beta(x, y) = (y, -x) \).


