Numer. Math. (2018) 139:529–561
https://doi.org/10.1007/s00211-018-0947-5

Numerische Mathematik

Nonlinear discontinuous Petrov–Galerkin methods

C. Carstensen[1] · P. Bringmann[1] · F. Hellwig[1] · P. Wriggers[2]

Received: 20 April 2017 / Revised: 11 October 2017 / Published online: 6 March 2018

Abstract The discontinuous Petrov–Galerkin method is a minimal residual method with broken test spaces and is introduced for a nonlinear model problem in this paper. Its lowest-order version applies to a nonlinear uniformly convex model example and is equivalently characterized as a mixed formulation, a reduced formulation, and a weighted nonlinear least-squares method. Quasi-optimal a priori and reliable and efficient a posteriori estimates are obtained for the abstract nonlinear dPG framework for the approximation of a regular solution. The variational model example allows

The work has been written while the first author enjoyed the hospitality of the Hausdorff Research Institute of Mathematics in Bonn, Germany, during the Hausdorff Trimester Program 'Multiscale Problems: Algorithms, Numerical Analysis and Computation'. The second and third author were supported by the Berlin Mathematical School. The research of all four authors has been supported by the Deutsche Forschungsgemeinschaft in the Priority Program 1748 'Reliable simulation techniques in solid mechanics. Development of non-standard discretization methods, mechanical and mathematical analysis' under the project 'Foundation and application of generalized mixed FEM towards nonlinear problems in solid mechanics' (CA 151/22-1 and WR 19/51-1).

✉ C. Carstensen
cc@math.hu-berlin.de

P. Bringmann
bringman@math.hu-berlin.de

F. Hellwig
hellwigf@math.hu-berlin.de

P. Wriggers
wriggers@ikm.uni-hannover.de

[1] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[2] Institut für Kontinuumsmechanik, Gottfried Wilhelm Leibniz Universität Hannover, Appelstraße 11, 30167 Hannover, Germany

for a built-in guaranteed error control despite inexact solve. The subtle uniqueness of discrete minimizers is monitored in numerical examples.

Mathematics Subject Classification 47H05 · 49M15 · 65N12 · 65N15 · 65N30

1 Introduction

The discontinuous Petrov–Galerkin methodology (dPG) has recently been introduced with the intention to design the optimal test spaces in a Petrov–Galerkin scheme for maximal stability. On the continuous level, the weak form of a PDE may assume the general form $b(u, \cdot) = F$ with a unique solution u in some real Banach space X and some bilinear form $b : X \times Y \to \mathbb{R}$ for some real Hilbert space Y with scalar product $a : Y \times Y \to \mathbb{R}$ and a given right-hand side $F \in Y^*$, the dual to Y. Well-posedness is understood to lead to an inf-sup condition on the continuous level. Given some discrete trial space $X_h \subset X$, the restriction $b|_{X_h \times Y}$ clearly satisfies the inf-sup condition (even with a possibly slightly better inf-sup constant) but it is less clear how to choose the best trial space M_h, i.e. some subspace, $M_h \subset Y$ such that

$$0 < \beta(X_h, M_h) := \inf_{x_h \in X_h} \sup_{y_h \in M_h} \frac{b(x_h, y_h)}{\|x_h\|_X \|y_h\|_Y} \tag{1.1}$$

is maximal under the condition that $\dim(X_h) = \dim(M_h)$ is fixed. The idealized dPG method computes the optimal test space utilizing some Riesz representations in the infinite-dimensional Hilbert space Y [18]. The practical realization utilizes, first, a test-search space $Y_h \subset Y$ with dimension $n = \dim(Y_h)$ much larger than the dimension $m = \dim(X_h)$ of the trial space X_h and, second, a minimal residual method to compute the discrete solution as a minimizer

$$x_h \in \operatorname*{argmin}_{\xi_h \in X_h} \|F - b(\xi_h, \bullet)\|_{Y_h^*}. \tag{1.2}$$

The method is in fact equivalent to a Petrov–Galerkin scheme with the bilinear form restricted to $X_h \times M_h$ for an appropriate subspace $M_h \subset Y_h$ of dimension m as pointed out in [8, Thm. 3.3]. Therefore, the large discrete space Y_h (which is an input of the dPG scheme) is called test-search space [17] and the (implicit) test space M_h is not visible in (1.2).

The computation of x_h in (1.2) is equivalent to solving the normal equations and so possibly expensive. This guided Demkowicz and Gopalakrishnan [19] to break the norms in the test (and ansatz) spaces [6]. This allows a parallel computation of the dual norm separately for each individual element domain. As it stands today, the term dPG abbreviates "discontinuous Petrov–Galerkin" and stands for a *minimal residual method with broken test or ansatz functions* and solely outlines a paradigm. The dPG methodology allows various weak and ultra-weak formulations, where X and Y are completely different and b is not at all symmetric. The least-squares finite element methods can be seen as a (degenerated) subset of (an idealized) dPG with a degenerated test space in which the Lebesgue norm can be evaluated exactly.

To the best knowledge of the authors, not much is known about nonlinear versions of the methodology. One first choice is to linearize the problem and then apply the dPG schemes to the linear equations to generalize the Gauss-Newton method. There exist already suggestions for nonlinear applications, in which there are constraints plus a linear problem, e.g., for the contact problem in [21]. Concepts of nonlinear dPG in fluid mechanics have been discussed in [16]. Another usage of the term *nonlinear* is in nonlinear approximation theory and there is the contribution [22] on linear problems with an attempt to replace the Hilbert space Y by some uniformly convex Banach space.

This paper introduces a direct *nonlinear dPG methodology* and replaces the above bilinear form b by some nonlinear mapping $b : X \times Y \to \mathbb{R}$, which is linear and bounded in the second component to allow the computation of the dual norm in the minimal residual method. To stress the nonlinear dependence in the first component in X, the notation in this papers follows [24] and separates the linear components by a semi-colon so that the nonlinear dPG method replaces $b(\xi_h, \bullet)$ in (1.2) by $b(\xi_h; \bullet)$.

The simplest case study for the nonlinear dPG methodology is an energy minimization problem with some Hilbert space setting and a nonlinearity with quadratic growth in the gradient. The scalar model example of this paper stands for a larger class of Hencky materials [28, Sect. 62.8] and is the first model problem in line towards real-life applications with a matrix-valued stress $\sigma(F)$ given as a nonlinear function of some deformation gradient F (such as the gradient ∇u of the displacement u) and the remaining equilibration equation

$$f + \operatorname{div} \sigma(\nabla u) = 0 \quad \text{a.e. in } \Omega \tag{1.3}$$

for some prescribed source term f in the domain Ω. Although the existence of discrete solutions x_h to (1.2) follows almost immediately, the closeness of x_h to some continuous solution x is completely open (cf. Remark 2.11 below for a brief discussion).

One critical point is the role of the stability condition (1.1) in the nonlinear setting for a regular solution and its low-order discretizations (as the most natural first choice for nonlinear problems, partly because of limited known regularity properties). In the situation of the model scenario (1.3), the discrete stability follows from the stability of the continuous form for piecewise constant ∇u_h and so the local discrete stability simply follows from the linearization.

The overall structure of the nonlinear dPG of the type (1.2) but for a nonlinear map b with derivative b' with respect to the first variable is also characterized as a nonlinear mixed formulation with solution $(x_h, y_h) \in X_h \times Y_h$ to

$$\begin{aligned} a(y_h, \eta_h) + b(x_h; \eta_h) &= F(\eta_h) \quad \text{for all } \eta_h \in Y_h, \\ b'(x_h; \xi_h, y_h) &= 0 \quad \text{for all } \xi_h \in X_h. \end{aligned} \tag{M}$$

Another characterization in the lowest-order cases under consideration is that as a weighted least-squares functional on Courant finite element functions $S^1_0(\mathcal{T})$ with homogeneous Dirichlet boundary values and the Raviart–Thomas finite element functions $RT_0(\mathcal{T})$ with some mesh-dependent piecewise constant weight $S_0 \in P_0(\mathcal{T}; \mathbb{R}^{n\times n})$

$$(u_{\mathrm{C}}, p_{\mathrm{RT}}) \in \operatorname*{argmin}_{(v_{\mathrm{C}}, q_{\mathrm{RT}}) \in S^1_0(\mathcal{T}) \times RT_0(\mathcal{T})} \Big(\|\Pi_0 f + \operatorname{div} q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|(I_{n\times n} + S_0)^{-1/2}\big(\Pi_0 q_{\mathrm{RT}} - \sigma(\nabla v_{\mathrm{C}}) + \Pi_0(f(\mathrm{id} - \mathrm{mid}(\mathcal{T})))\big)\|^2_{L^2(\Omega)}\Big).$$

This is already a new result even for the linear cases in [9,13] and opens the door to a convergence analysis of adaptive algorithms via a generalization of [11,14].

This paper contributes the aforementioned equivalent characterizations and a first convergence analysis in the natural norms. The a priori result is local quasi-optimal convergence for the simple model problem in that any discrete solution $x_h \in X_h$, sufficiently close to the exact regular solution $x \in X$, satisfies

$$\|x - x_h\|_X \lesssim \inf_{\xi_h \in X_h} \|x - \xi_h\|_X.$$

It has been discussed in [5,9,13] that the norm of the computed residual $\|y_h\|_Y = \|F - b(v_C, q_{\mathrm{RT}}; \bullet)\|_{Y_h^*}$ is almost a computable error estimator for linear problems and this paper extends it to the a posteriori error estimate

$$\|p - q_{\mathrm{RT}}\|^2_{H(\operatorname{div},\Omega)} + |||u - v_C|||^2 \approx \|F - b(v_C, q_{\mathrm{RT}}; \bullet)\|^2_{Y_h^*} + \|(1 - \Pi_0) f\|^2_{L^2(\Omega)} + \|(1 - \Pi_0) q_{\mathrm{RT}}\|^2_{L^2(\Omega)} \tag{1.4}$$

for the nonlinear model problem (1.3). Since $\|F - b(v_C, q_{\mathrm{RT}}; \bullet)\|_{Y_h^*}$ is the computable residual, this leads to built-in error control despite inexact solve: The discrete quantities (v_C, q_{RT}) in (1.4) do *not* need to solve the nonlinear dPG discrete problem.

The analysis is given for the primal version of the nonlinear dPG for brevity but applies to the other formulations of Sect. 4.4 as well. The results of this paper can be generalized, e.g., to the Hencky material [28, Sect. 62.8], and then applied to more complicated real-life computational challenges where the advantages of the dPG methodology are more striking.

The remaining parts of of this paper are organised as follows. Section 2 discusses an abstract framework for different equivalent formulations of a dPG method for nonlinear problems and develops an abstract a priori estimate. Section 3 presents a model problem with a dPG discretization. Section 4 analyses this discretization and gives proofs of the existence of a solution and an a posteriori error estimate. Some numerical examples in Sect. 5 conclude the paper.

This paper employs standard notation of Sobolev and Lebesgue spaces $H^k(\Omega)$, $H(\operatorname{div}, \Omega)$, $L^2(\Omega)$, and $L^\infty(\Omega)$ and the corresponding spaces of vector- or matrix-valued functions $H^k(\Omega; \mathbb{R}^n)$, $L^2(\Omega; \mathbb{R}^n)$, $L^\infty(\Omega; \mathbb{R}^n)$, $H^k(\Omega; \mathbb{R}^{n\times n})$, $H(\operatorname{div}, \Omega; \mathbb{R}^{n\times n})$, $L^2(\Omega; \mathbb{R}^{n\times n})$, and $L^\infty(\Omega; \mathbb{R}^{n\times n})$. For any regular triangulation $\mathcal{T}$ of Ω, let $H^k(\mathcal{T}) := \prod_{T\in\mathcal{T}} H^k(T) := \{v \in L^2(\Omega) \,|\, \forall T \in \mathcal{T},\, v|_T \in H^k(T)\}$ denote the piecewise (or broken) Sobolev spaces and $(\nabla_{\mathrm{NC}} v)\big|_T = \nabla(v\big|_T)$ on $T \in \mathcal{T}$ the piecewise gradient for $v \in H^1(\mathcal{T})$. Let $|||\bullet||| := |\bullet|_{H^1(\Omega)} = \|\nabla \bullet\|_{L^2(\Omega)}$ abbreviate the energy norm. For every Hilbert space X, let $(\bullet, \bullet)_X$ denote the associated

inner product and, for every normed space $(X, \|\bullet\|_X)$, $\mathcal{S}(X) := \{x \in X \mid \|x\|_X = 1\}$ the sphere in X. The measure $|\bullet|$ is context-dependent and refers to the number of elements of some finite set or the length $|E|$ of an edge E or the area $|T|$ of some triangle T and not just the modulus of a real number or the Euclidean length of a vector.

Throughout the paper, $A \lesssim B$ abbreviates the relation $A \leq C\,B$ with a generic constant $0 < C$, which does not depend on the mesh-size of the underlying triangulation $\mathcal{T}$ but solely on the initial triangulation $\mathcal{T}_0$; $A \approx B$ abbreviates $A \lesssim B \lesssim A$, e.g., in (1.4).

2 Abstract framework

This section analyses an abstract nonlinear dPG method and presents an a priori error estimate.

2.1 Abstract nonlinear dPG

For an open set $D \neq \emptyset$ in a real Banach space X and a real Hilbert space Y with scalar product $a : Y \times Y \to \mathbb{R}$, let $B \in C^1(D; Y^*)$ be a differentiable nonlinear map with Fréchet derivative $\mathrm{D}B(x) \in L(X; Y^*)$ at $x \in D$. With the duality bracket $\langle \bullet, \bullet \rangle$ in Y, associate the nonlinear map $b : X \times Y \to \mathbb{R}$, $b(x; \bullet) := \langle B(x), \bullet \rangle$, which is linear and bounded in the second component. Let $b'(x; \bullet)$ abbreviate the derivative $\mathrm{D}B(x) \in L(X; Y^*)$ with $b'(x; \xi, \eta) := \langle \mathrm{D}B(x; \xi), \eta \rangle$ for $x \in D, \xi \in X, \eta \in Y$.

Given $F \in Y^*$, let $x \in D$ be a *regular solution* to the problem $B(x) = F$ in Y^*. That means that x solves $B(x) = F$ and the Fréchet derivative $\mathrm{D}B$ at x is a bijection from X to Y^*. The latter implies the inf-sup condition for the Fréchet derivative at the regular solution x, namely,

$$0 < \beta(x) := \inf_{\xi \in \mathcal{S}(X)} \sup_{\eta \in \mathcal{S}(Y)} b'(x; \xi, \eta). \tag{2.1}$$

The minimal residual formulation of the continuous problem seeks $x \in X$ with

$$x \in \operatorname*{argmin}_{\xi \in D} \|F - B(\xi)\|_{Y^*}. \tag{2.2}$$

The existence of a solution x to (2.2) is immediate from the assumption $B(x) = F$. In particular, the minimum is zero and any minimizer x in (2.2) solves $B(x) = F$. The situation is (in general) different on the discrete level with some discrete subspaces $X_h \subset X$ and $Y_h \subset Y$, the dPG scheme seeks a minimizer $x_h \in D_h := X_h \cap D$ of the residual $F - B(\xi_h)$ in the norm of Y_h^*,

$$x_h \in \operatorname*{argmin}_{\xi_h \in D_h} \|F - B(\xi_h)\|_{Y_h^*}. \tag{dPG}$$

The existence of a solution to (dPG) requires further assumptions and follows in Proposition 4.3 for a model problem.

2.2 Derivation of nonlinear dPG

A formal Lagrange ansatz leads to the minimization of the Lagrange functional $\mathcal{L} : D_h \times Y_h \times \mathbb{R} \to \mathbb{R}$ defined for $(x_h, y_h, \lambda) \in X_h \times Y_h \times \mathbb{R}$ by

$$\mathcal{L}(x_h, y_h, \lambda) := F(y_h) - b(x_h; y_h) - \frac{\lambda}{2}\big(a(y_h, y_h) - 1\big).$$

The stationary points $x_h \in D_h$, $y_h \in Y_h$, and $\lambda \in \mathbb{R}$ of $\mathcal{L}$ are characterized by the first derivatives of $\mathcal{L}$ with respect to each argument in the sense that, for all $\eta_h \in Y_h$ and $\xi_h \in X_h$,

$$\lambda a(y_h, \eta_h) + b(x_h; \eta_h) = F(\eta_h), \quad b'(x_h; \xi_h, y_h) = 0, \quad a(y_h, y_h) = 1.$$

For $\eta_h = y_h$, this implies $\lambda = F(y_h) - b(x_h; y_h)$. The substitution of y_h by λy_h leads to a modified system of equations. The resulting mixed formulation of the nonlinear dPG method seeks $x_h \in X_h$ and $y_h \in Y_h$ with

$$\begin{aligned} a(y_h, \eta_h) + b(x_h; \eta_h) &= F(\eta_h) \quad \text{for all } \eta_h \in Y_h, \\ b'(x_h; \xi_h, y_h) &= 0 \quad \text{for all } \xi_h \in X_h. \end{aligned}$$

Notice that this is known for linear problems (there, $b = b'(x_h; \bullet)$) [17, Sect. 2.3].

2.3 Equivalent mixed formulation

It is known in linear problems that the dPG method is equivalent to the mixed problem (M) and this is generalized in this subsection to the nonlinear problem $B(x) = F$ at hand. Any local (or global) minimizer of $\Phi(\xi_h) := \|F - B\xi_h\|^2_{Y_h^*}/2$ is a stationary point of Φ.

Definition 2.1 (*stationary point*) Any $x_h \in D_h := D \cap X_h$ is a stationary point of the dPG discretization (dPG) if any directional derivative of $\Phi(\xi_h) := \|F - B\xi_h\|^2_{Y_h^*}/2$ vanishes at x_h, i.e., $\lim_{\delta\to 0}(\Phi(x_h + \delta\xi_h) - \Phi(x_h))/\delta = 0$ for all $\xi_h \in X_h$.

Stationary points are exactly the solutions to (M).

Theorem 2.2 *((dPG) $\Leftrightarrow$ (M))*

(a) *Suppose x_h is a stationary point of* (dPG) *and y_h is the residual's Riesz representation (i.e. $a(y_h, \bullet) = F - b(x_h; \bullet)$) in Y_h. Then (x_h, y_h) solves* (M).
(b) *Suppose that (x_h, y_h) solves* (M)*, then x_h is a stationary point of* (dPG).

Proof (a) For any $\xi_h \in D_h$, the unique Riesz representation $\varrho_h(\xi_h) \in Y_h$ of the residual $F - b(\xi_h; \bullet) \in Y_h^*$ satisfies

$$\Phi(\xi_h) = \frac{1}{2}\|\varrho_h(\xi_h)\|_Y^2.$$

Given the stationary point $x_h \in D_h$ to (dPG) and $\xi_h \in X_h$, consider $\Phi(x_h + t\xi_h)$ as a scalar function of the real parameter t with a derivative zero at $t = 0$. For $|t|$ small such that $x_h(t) := x_h + t\xi_h \in D_h$ and $y_h(t) := \varrho_h(x_h(t))$, it follows

$$a(y_h(t), \bullet) + b(x_h(t); \bullet) = F \quad \text{in } Y_h^*.$$

A differentiation with respect to t shows for $\dot{y}_h := \partial y_h(0)/\partial t$ and $\dot{x}_h := \partial x_h(0)/\partial t = \xi_h$ that $\dot{y}_h$ exists and is the Riesz representation of $-b'(x_h; \xi_h, \bullet) = a(\dot{y}_h, \bullet)$ in Y_h. Therefore, $\Phi(x_h(t)) = a(y_h(t), y_h(t))/2$ is differentiable and the derivative vanishes at $t = 0$, which leads to

$$0 = a(\dot{y}_h, y_h) \quad \text{for } y_h := y_h(0).$$

It follows that

$$b'(x_h; \xi_h, y_h) = 0 \quad \text{for all } \xi_h \in X_h.$$

Since $y_h = y_h(0) = \varrho_h(x_h)$, (x_h, y_h) solves (M).

(b) Conversely, if (x_h, y_h) solves (M) then, for any $\xi_h \in D_h$ and the above notation for the Riesz representation $y_h(t)$ of $F - Bx_h(t)$ in Y_h,

$$\|F - Bx_h(t)\|_{Y_h^*}^2 = a(y_h(t), y_h(t)) = F(y_h(t)) - b(x_h(t); y_h(t))$$

has a derivative with respect to t at $t = 0$, namely, for $y_h := y_h(0)$

$$2\,a(\dot{y}_h, y_h) = F(\dot{y}_h) - b'(x_h; \xi_h, y_h) - b(x_h; \dot{y}_h).$$

Since $b'(x_h; \xi_h, y_h) = 0$ and $F(\dot{y}_h) - b(x_h; \dot{y}_h) = a(y_h, \dot{y}_h)$, this implies $a(\dot{y}_h, y_h) = 0$. Recall $\partial\Phi(x(t))/\partial t|_{t=0} = a(\dot{y}_h, y_h) = 0$, and so x_h is a stationary point of Φ. □

Proposition 2.3 *(necessary and sufficient second-order condition)*
Assume that Φ is twice differentiable. (a) *If x_h solves* (dPG)*, then*

$$b''(x_h; \xi_h, \xi_h, y_h) \le \|b'(x_h; \xi_h, \bullet)\|_{Y_h^*}^2 \quad \textit{for all } \xi_h \in X_h. \tag{2.3}$$

(b) *If, in addition,*

$$b''(x_h; \xi_h, \xi_h, y_h) < \|b'(x_h; \xi_h, \bullet)\|_{Y_h^*}^2 \quad \textit{for all } \xi_h \in X_h \setminus \{0\}, \tag{2.4}$$

then x_h is locally unique.

Proof The second derivative of $\Phi(x_h(t))$ reads $a(\partial^2 y_h/\partial t^2, y_h) + \|\partial y_h/\partial t\|_Y^2$. Recall from the proof of Theorem 2.2 for $t = 0$, that the Riesz representation $\dot{y}_h = \partial y_h(0)/\partial t$ satisfies

$$a(\dot{y}_h, \bullet) = -b'(x_h; \xi_h, \bullet) \quad \text{in } Y_h \quad \text{and} \quad \|\dot{y}_h\|_Y = \|b'(x_h; \xi_h, \bullet)\|_{Y_h^*}.$$

Another differentiation with respect to t shows that $\ddot{y}_h := \partial^2 y_h(0)/\partial t^2$ satisfies

$$a(\ddot{y}_h, \bullet) = -b''(x_h; \xi_h, \xi_h, \bullet) \quad \text{in } Y_h.$$

Consequently, the second derivative of $\Phi(x_h(t))$ at $t = 0$ is

$$-b''(x_h; \xi_h, \xi_h, y_h) + \|b'(x_h; \xi_h, \bullet)\|_{Y_h^*}^2. \tag{2.5}$$

The assertion follows from this and standard arguments in the calculus of stationary and minimal points. □

Remark 2.4 (linear problems) For a linear problem, $b''(x_h; \bullet)$ vanishes and (2.4) holds. This implies local uniqueness in the linear situation (which is a global one).

The uniqueness of the discrete solution is observed in numerical examples; cf. Theorem 4.4 for a sufficient condition in the model example below.

2.4 Abstract a priori error analysis

This section presents a best-approximation result based on a discrete inf-sup condition and the existence of a Fortin operator.

Hypothesis 2.5 Throughout this paper, assume that there exists a linear bounded projection $\Pi_h : Y \to Y_h$ with $\Pi_h|_{Y_h} = \mathrm{id}|_{Y_h}$ and

$$b'(D_h; X_h, (1 - \Pi_h)Y) = 0, \tag{2.6}$$

i.e., for all $x_h \in D_h$ and all $y \in Y$, $\Pi_h y \in Y_h$ satisfies $b'(x_h; \xi_h, y - \Pi_h y) = 0$ for all $\xi_h \in X_h$. Let $\|\Pi_h\|$ denote the bound of Π_h in $L(Y; Y)$.

The following theorem generalizes [1, Prop. 5.4.2] to the nonlinear problem at hand. A sufficiently fine initial triangulation guarantees that $B(x, \varepsilon) \cap X_h \subset D_h$ is nonempty.

Theorem 2.6 *(discrete inf-sup condition) Given a regular solution x to $B(x) = F$, there exists an open ball $B(x, \varepsilon) := \{\tilde{x} \in X \mid \|x - \tilde{x}\|_X < \varepsilon\}$ of radius $\varepsilon > 0$ around x such that, for all $\tilde{x}_h \in B(x, \varepsilon) \cap X_h \subset D_h$, the following discrete inf-sup condition holds*

$$0 < \frac{\beta(x; X_h, Y_h)}{2\|\Pi_h\|} \le \beta(\tilde{x}_h; X_h, Y_h) := \inf_{\xi_h \in \mathcal{S}(X_h)} \sup_{\eta_h \in \mathcal{S}(Y_h)} b'(\tilde{x}_h; \xi_h, \eta_h).$$

Proof The continuous inf-sup condition (2.1) and the continuity of $\mathrm{D}B$ in D lead to some ε such that

$$B(x,\varepsilon) \subset D, \tag{2.7}$$

$$\beta(x)/2 \le \inf_{\xi\in B(x,\varepsilon)} \beta(\xi). \tag{2.8}$$

Then $\tilde{x}_h \in B(x,\varepsilon)\cap X_h$ and (2.6) imply

$$\begin{aligned}
\beta(x)/2 \le \beta(\tilde{x}_h) &\le \inf_{\xi_h\in\mathcal{S}(X_h)} \sup_{\eta\in\mathcal{S}(Y)} b'(\tilde{x}_h;\xi_h,\eta)\\
&= \inf_{\xi_h\in\mathcal{S}(X_h)} \sup_{\eta\in\mathcal{S}(Y)} b'(\tilde{x}_h;\xi_h,\Pi_h\eta)\\
&= \inf_{\xi_h\in\mathcal{S}(X_h)} \sup_{\eta\in\mathcal{S}(Y)} \|\Pi_h\eta\|_Y b'\big(\tilde{x}_h;\xi_h,\Pi_h\eta/\|\Pi_h\eta\|_Y\big)\\
&\le \|\Pi_h\| \inf_{\xi_h\in\mathcal{S}(X_h)} \sup_{\eta\in\mathcal{S}(Y)} b'\big(\tilde{x}_h;\xi_h,\Pi_h\eta/\|\Pi_h\eta\|_Y\big)\\
&= \|\Pi_h\| \inf_{\xi_h\in\mathcal{S}(X_h)} \sup_{\eta_h\in\mathcal{S}(Y_h)} b'(\tilde{x}_h;\xi_h,\eta_h).
\end{aligned}$$

Hence, any $\tilde{x}_h \in B(x,\varepsilon)$ satisfies $0 < \frac{\beta(x)}{2\|\Pi_h\|} \le \beta(\tilde{x}_h;X_h,Y_h)$. □

Remark 2.7 (converse of Theorem 2.6) Given the discrete inf-sup condition

$$0 < \inf_{\xi_h\in\mathcal{S}(X_h)} \sup_{\eta_h\in\mathcal{S}(Y_h)} b'(\tilde{x}_h;\xi_h,\eta_h) \tag{2.9}$$

at some point $\widetilde{x}_h \in D_h$, the techniques of [9, Lemma 10] guarantee the existence of a linear bounded projection $\Pi_h(\widetilde{x}_h) : Y \to Y_h$ with (2.6), which depends on $\widetilde{x}_h$. The above proof shows that the existence of $\Pi_h(\widetilde{x}_h)$ is also sufficient for (2.9). The class of model examples allows for the simple more uniform Hypothesis 2.5 with $\Pi_h(\widetilde{x}_h) = \Pi_h$ independent of $\widetilde{x}_h \in D_h$.

Theorem 2.8 *(local best-approximation) Given a regular solution x to $B(x) = F$, there exist positive constants $\varepsilon > 0$ and $C(x,\varepsilon) > 0$ such that any solution (x_h, y_h) to* (M) *with $\|x - x_h\|_X < \varepsilon$ satisfies*

$$\|x - x_h\|_X + \|y_h\|_Y \le C(x,\varepsilon) \inf_{\xi_h\in X_h} \|x - \xi_h\|_X.$$

The proof of the theorem requires the following lemma.

Lemma 2.9 *Any $\varepsilon > 0$ and $x_h \in B(x,\varepsilon) \subset D$ satisfy*

$$\begin{aligned}
&\|b'(x_h; x - x_h, \bullet) - b(x;\bullet) + b(x_h;\bullet)\|_{Y^*}\\
&\quad\le 2 \sup_{\xi\in B(x,\varepsilon)} \|\mathrm{D}B(x) - \mathrm{D}B(\xi)\|_{L(X;Y^*)}\|x - x_h\|_X
\end{aligned} \tag{2.10}$$

$$\|B(x) - B(x_h)\|_{Y^*} \leq \sup_{\xi \in B(x,\varepsilon)} \|\mathrm{D}B(\xi)\|_{L(X;Y^*)} \|x - x_h\|_X. \tag{2.11}$$

Proof Given any $\eta \in \mathcal{S}(Y)$, the Taylor's formula of b at x_h with remainder reads

$$\begin{aligned} &b'(x_h; x - x_h, \eta) - b(x; \eta) + b(x_h; \eta) \\ &\quad = \int_0^1 \Big(b'(x_h; x - x_h, \eta) - b'(x_h + s(x - x_h); x - x_h, \eta) \Big) \, \mathrm{d}s. \end{aligned}$$

Since $\|x - x_h\|_X < \varepsilon$ implies $\|x - (x_h + s(x - x_h))\|_X < \varepsilon$ for $0 \leq s \leq 1$, the triangle inequality proves

$$\begin{aligned} &b'(x_h; x - x_h, \eta) + b(x_h; \eta) - b(x; \eta) \\ &\quad \leq 2 \sup_{\xi \in B(x,\varepsilon)} |b'(x; x - x_h, \eta) - b'(\xi; x - x_h, \eta)| \\ &\quad \leq 2 \sup_{\xi \in B(x,\varepsilon)} \|\mathrm{D}B(x) - \mathrm{D}B(\xi)\|_{L(X;Y^*)} \|x - x_h\|_X. \end{aligned}$$

Since $\eta \in \mathcal{S}(Y)$ is arbitrary, this implies (2.10). The assertion (2.11) follows from the same arguments without the term $b'(x_h; x - x_h, \eta)$. □

Proof (of Theorem 2.8) Let $\tilde{x}_h$ be the best-approximation to x in X_h, i.e.,

$$\|x - \tilde{x}_h\|_X = \inf_{\xi_h \in D_h} \|x - \xi_h\|_X \leq \|x - x_h\|_X < \varepsilon.$$

Suppose $\varepsilon > 0$ satisfies (2.7)–(2.8) and, with the continuity of $\mathrm{D}B$ at x,

$$\sup_{\xi \in B(x,\varepsilon)} \|\mathrm{D}B(\xi)\|_{L(X;Y^*)} \leq 2 \|\mathrm{D}B(x)\|_{L(X;Y^*)}. \tag{2.12}$$

The discrete inf-sup condition from Theorem 2.6 plus the Brezzi splitting lemma [3, Thm. 4.3 in Ch. III] with inf-sup constants, $\beta(x)/2$ and 1, and continuity constants, $2\|\mathrm{D}B(x)\|_{L(X;Y^*)}$ and 1, for the bilinear form $b'(x_h; \bullet, \bullet)$ and scalar product a prove the global inf-sup condition $0 < \gamma \leq \beta(x_h; X_h, Y_h)$ for

$$\gamma := \inf_{(\tilde{\xi}_h, \tilde{\eta}_h) \in \mathcal{S}(X_h \times Y_h)} \sup_{(\xi_h, \eta_h) \in \mathcal{S}(X_h \times Y_h)} \Big(b'(x_h; \tilde{\xi}_h, \eta_h) + b'(x_h; \xi_h, \tilde{\eta}_h) + a(\tilde{\eta}_h, \eta_h) \Big).$$

independent of ε with (2.7)–(2.8) and (2.12). Given $\gamma > 0$ and $\beta(x) > 0$ suppose, for some smaller $\varepsilon > 0$ if necessary, that $\varepsilon > 0$ satisfies (2.7)–(2.8), (2.12), and, from the continuity of $\mathrm{D}B$ at x,

$$\sup_{\xi \in B(x,\varepsilon)} \|\mathrm{D}B(x) - \mathrm{D}B(\xi)\|_{L(X;Y^*)} \leq \min\{\gamma/4, \beta(x)/8\}. \tag{2.13}$$

For the best-approximation $\tilde{y}_h = 0$ to $y = 0$ in Y_h and $(\tilde{\xi}_h, \tilde{\eta}_h) = (\tilde{x}_h - x_h, \tilde{y}_h - y_h)$, this implies the existence of $(\xi_h, \eta_h) \in \mathcal{S}(X_h \times Y_h)$ with

$$\gamma\big(\|\tilde{x}_h - x_h\|_X + \|y_h\|_Y\big) \le b'(x_h; \tilde{x}_h - x_h, \eta_h) - b'(x_h; \xi_h, y_h) - a(y_h, \eta_h).$$

Since (x_h, y_h) solves (M) and $\tilde{y}_h = 0$, this leads to

$$\begin{aligned}\gamma\big(\|\tilde{x}_h - x_h\|_X + \|y_h\|_Y\big) \le\ & b'(x_h; \tilde{x}_h - x, \eta_h) \\ & + b'(x_h; x - x_h, \eta_h) + b(x_h; \eta_h) - b(x; \eta_h).\end{aligned}$$

Lemma 2.9 and (2.13) imply

$$b'(x_h; x - x_h, \eta_h) + b(x_h; \eta_h) - b(x; \eta_h) \le \gamma\|x - x_h\|_X/2.$$

The combination of the preceding two displayed formulae reads

$$\frac{\gamma}{2}\|\tilde{x}_h - x_h\|_X + \gamma\|y_h\|_Y \le b'(x_h; \tilde{x}_h - x_h, \eta_h).$$

With (2.12), this is bounded from above by

$$\|\,\mathrm{D}B(x_h)\|_{L(X;Y^*)}\|x - \tilde{x}_h\|_X \le 2\|\,\mathrm{D}B(x_h)\|_{L(X;Y^*)}\|x - \tilde{x}_h\|_X.$$

The triangle inequality concludes the proof. □

Remark 2.10 Under further smoothness conditions on the nonlinear mapping b' the local existence and uniqueness of a discrete solution, e.g., follows from [25, Thm. 2].

Remark 2.11 The Newton–Kantorovich theorem [27, Section 5.2] is another tool for the proof of the existence of discrete solutions close to the regular solution. In the model problem of Sect. 3, the higher Fréchet derivatives for this argument do *not* exist, cf. Remark 3.3 for details.

2.5 Abstract a posteriori error analysis

This subsection is devoted to a brief abstract a posteriori error analysis of the nonlinear dPG. Given a discrete approximation x_h close to the regular solution x to $B(x) = F$, the residual $F - B(x_h) \in Y^*$ has a norm $\|F - B(x_h)\|_{Y^*}$ that, in principle, is accessible in the sense that lower and upper bounds may be computable. The latter issue is a typical general task in the a posteriori error analysis and will be adressed in Sect. 3 for a model example.

Theorem 2.12 *(local a posteriori analysis) Let x be a regular solution to $B(x) = F$ with inf-sup constant $\beta(x)$ from (2.1). Then there exists some $\varepsilon > 0$ such that any $x_h \in B(x, \varepsilon) \subset D$ satisfies*

$$\frac{\beta(x)}{4}\|x - x_h\|_X \le \|F - B(x_h)\|_{Y^*} \le 2\|\,\mathrm{D}B(x)\|_{L(X;Y^*)}\|x - x_h\|_X.$$

Proof With the choice of $\varepsilon > 0$ from the proof of Theorem 2.8 it follows (2.7)–(2.8) and (2.12)–(2.13). The continuous inf-sup condition (2.1) implies the existence of $\eta \in \mathcal{S}(Y)$ with

$$\begin{aligned}\frac{\beta(x)}{2}\|x - x_h\|_X &\le b'(x_h; x - x_h, \eta)\\ &\le b(x;\eta) - b(x_h;\eta) + |b'(x_h; x - x_h, \eta) - b(x;\eta) + b(x_h;\eta)|.\end{aligned}$$

Lemma 2.9 for the last term, $b(x;\eta) = F(\eta)$, and (2.13) show

$$\frac{\beta(x)}{2}\|x - x_h\|_X \le F(\eta) - b(x_h;\eta) + \frac{\beta(x)}{4}\|x - x_h\|_X.$$

This proves the asserted reliability

$$\frac{\beta(x)}{4}\|x - x_h\|_X \le \|F(\eta) - b(x_h;\eta)\|_{Y^*}.$$

To prove the efficiency, utilize $F = B(x)$, Lemma 2.9, and (2.12) to verify

$$\begin{aligned}\|F - B(x_h)\|_{Y^*} &= \|B(x) - B(x_h)\|_{Y^*}\\ &\le 2\|\mathrm{D}B(x)\|_{L(X;Y^*)}\|x - x_h\|_X.\end{aligned}$$

□

Remark 2.13 Since $y = 0$ and y_h is computed, the a posteriori error $\|y - y_h\|_Y = \|y_h\|_Y$ is already an error estimator and can be added on both sides of the reliability (resp. efficiency) a posteriori error estimate. This justifies the usage of the extended residual $\|F - a(y_h, \bullet) - b(x_h; \bullet)\|_{Y^*} + \|y_h\|_Y$ of the system (M).

Remark 2.14 The constants $\beta(x)/4$ (resp. $2\|\mathrm{D}B(x)\|_{L(X;Y^*)}$) in Theorem 2.12 follow from the choice of ε in the a priori error analysis in the proof of Theorem 2.8. For smaller and smaller values of ε, those constants could be replaced by any number $< \beta(x)$ (resp. $> \|\mathrm{D}B(x)\|_{L(X;Y^*)}$) in the following sense. For any $0 < \lambda < 1$ there exists some $\varepsilon > 0$ such that any $x_h \in B(x, \varepsilon)$ satisfies $\lambda\beta(x) \le \|F - B(x_h)\|_{Y^*} \le (1 + \lambda)\|\mathrm{D}B(x)\|_{L(X;Y^*)}$.

3 Model problem

This section introduces a nonlinear model problem and a low-order dPG discretization and establishes two further equivalent characterizations of the nonlinear dPG method: reduced discretization and weighted least-squares.

3.1 Convex energy minimization

The nonlinear model problem involves a nonlinear function $\phi \in C^2(0, \infty)$ with $0 < \gamma_1 \le \phi(t) \le \gamma_2$ and $0 < \gamma_1 \le \phi(t) + t\phi'(t) \le \gamma_2$ for all $t \ge 0$ and universal positive

constants γ_1, γ_2. Given $f \in L^2(\Omega)$ and the convex function φ, $\varphi(t) := \int_0^t s\,\phi(s)\,\mathrm{d}s$ for $t \geq 0$, the model problem minimizes the energy functional

$$E(v) := \int_\Omega \varphi(|\nabla v(x)|)\,\mathrm{d}x - \int_\Omega f v\,\mathrm{d}x \quad \text{among all } v \in H_0^1(\Omega).$$

The convexity of φ and the above assumptions on ϕ lead to growth-conditions and sequential weak lower semicontinuity of E and guarantee the unique existence of a minimizer u of E in $H_0^1(\Omega)$ [29, Thm. 25.D]. The equivalent Euler-Lagrange equation reads

$$\int_\Omega \phi(|\nabla u|)\,\nabla u \cdot \nabla v\,\mathrm{d}x = \int_\Omega f v\,\mathrm{d}x \quad \text{for all } v \in H_0^1(\Omega) \tag{3.1}$$

and has the unique solution u in $H_0^1(\Omega)$. The stress variable $\sigma(A) := \phi(|A|)A$ defines a function $\sigma \in C^1(\mathbb{R}^n; \mathbb{R}^n)$ with Fréchet derivative

$$\mathrm{D}\sigma(A) = \phi(|A|)I_{n\times n} + \phi'(|A|)|A|\,\mathrm{sign}(A) \otimes \mathrm{sign}(A) \tag{3.2}$$

with the sign function $\mathrm{sign}(A) := A/|A|$ for $A \in \mathbb{R}^n \setminus \{0\}$ and the closed unit ball $\mathrm{sign}(0) := \overline{B(0,1)}$ in $\mathbb{R}^n$. The prefactor $\phi'(|A|)|A|$ makes $\mathrm{D}\sigma$ a continuous function in $\mathbb{R}^n$. In fact $\mathrm{D}\sigma \in C^0(\mathbb{R}^{n\times n}_{\mathrm{sym}})$ is bounded with eigenvalues in the compact interval $[\gamma_1, \gamma_2] \subset (0, \infty)$.

Remark 3.1 (Lip(σ) $\leq \gamma_2$) For $A, B \in \mathbb{R}^n$, the argument $\sigma(A) - \sigma(B) = \int_0^1 \mathrm{D}\sigma(sA + (1-s)B)(A-B)\,\mathrm{d}s$ and (3.2) imply the global Lipschitz continuity of σ,

$$|\sigma(A) - \sigma(B)| \leq \int_0^1 |\mathrm{D}\sigma(sA + (1-s)B)(A-B)|\,\mathrm{d}s \leq \gamma_2|A-B|.$$

Example 3.2 In the following examples, $0 \leq \phi'' \leq 2$ is bounded as well as ϕ' and $\mathrm{D}\sigma$ from (3.2) is globally Lipschitz continuous. (a) $\phi(t) := 2 + (1+t)^{-2}$ with $\gamma_1 = 1 < \gamma_2 = 3$ [15] and $\mathrm{Lip}(\mathrm{D}\sigma) \leq 4$ and (b) $\phi(t) := 2 - (1+t^2)^{-1}$ with $\gamma_1 = 1 < \gamma_2 = 4$ and $\mathrm{Lip}(\mathrm{D}\sigma) \leq 2$.

Remark 3.3 (second derivative) A formal calculation with $s(j) := (\mathrm{sign}\,A)_j$, $s(j,k) := (\mathrm{sign}\,A)_j(\mathrm{sign}\,A)_k$ etc. and the Kronecker symbol δ_{jk} for $j, k, \ell = 1, \ldots, n$ leads at any $A \in \mathbb{R}^n$ to

$$\mathrm{D}^2\sigma(A)_{j,k,\ell} = \phi'(|A|)(\delta_{jk}s(\ell) + \delta_{j\ell}s(k) + \delta_{k\ell}s(j)) + (\phi''(|A|)|A| - \phi'(|A|))s(j,k,\ell).$$

Although $\mathrm{D}^2\sigma(A)$ may be bounded (at least in the Example 3.2.a and b), it may be discontinuous for $A \to 0$. In Example 3.2.b, $\phi'(0) = 0$ and $\mathrm{D}^2\sigma$ is continuous with $\mathrm{D}^2\sigma(0) = 0$. The associated trilinear form $b''(x; \bullet)$, however, is not well-defined on $X \times Y \times Y$ because the product of three Lebesgue functions in $L^2(\Omega)$ is, in general, not in $L^1(\Omega)$.

3.2 Breaking the test spaces

Let $\Omega \subseteq \mathbb{R}^n$ be a bounded Lipschitz domain with polyhedral boundary $\partial\Omega$. Let $\mathcal{T}$ denote a regular triangulation of the domain Ω into n-simplices and let $\mathcal{E}$ (resp. $\mathcal{E}(T)$) denote the set of all sides in the triangulation (resp. of an n-simplex $T \in \mathcal{T}$).

The unit normal vector ν_T along the boundary ∂T of an n-simplex $T \in \mathcal{T}$ (is constant along each side of T and) points outwards. For any side $E = \partial T_+ \cap \partial T_- \in \mathcal{E}$ shared by two simplices, the enumeration of the neighbouring simplices $T_\pm$ is globally fixed and so defines a unique orientation of the unit normal $\nu_E = \nu_{T_+}|_E$. Let h_T denote the diameter of $T \in \mathcal{T}$, $h_{\max} := \max_{T\in\mathcal{T}} h_T \leq \operatorname{diam}(\Omega)$ and $h_{\mathcal{T}}|_K = h_K$ for any $K \in \mathcal{T}$. The barycenter $\operatorname{mid}(T)$ of $T \in \mathcal{T}$ defines the piecewise constant function $\operatorname{mid}(\mathcal{T}) \in P_0(\mathcal{T};\mathbb{R}^n)$ by $\operatorname{mid}(\mathcal{T})|_K := \operatorname{mid}(K)$ for any $K \in \mathcal{T}$ and $\operatorname{mid}(E)$ is the barycenter of $E \in \mathcal{E}$. The piecewise affine function $\bullet - \operatorname{mid}(\mathcal{T}) \in P_1(\mathcal{T};\mathbb{R}^n)$ equals $x - \operatorname{mid}(T)$ at $x \in T \in \mathcal{T}$.

Recall that $H^k(\mathcal{T}) := \prod_{T\in\mathcal{T}} H^k(T) := \{v \in L^2(\Omega) \,|\, \forall T \in \mathcal{T},\ v|_T \in H^k(T)\}$ denotes the piecewise Sobolev space. Define the discrete spaces

$$
\begin{aligned}
P_k(T) &:= \{v_k \in L^\infty(T) \ \mid\ v_k \text{ is polynomial on } T \text{ of degree } \leq k\},\\
P_k(\mathcal{T}) &:= \{v_k \in L^\infty(\Omega) \ \mid\ \forall T \in \mathcal{T},\ v_k|_T \in P_k(T)\},\\
P_k(\mathcal{T};\mathbb{R}^n) &\equiv P_k(\mathcal{T})^n,\\
S^k_0(\mathcal{T}) &:= P_k(\mathcal{T}) \cap H^1_0(\Omega),\\
RT_k(\mathcal{T}) &:= \{q_k \in H(\operatorname{div},\Omega) \,|\, \exists A \in P_k(\mathcal{T};\mathbb{R}^n), \exists b \in P_k(\mathcal{T}),\\
&\qquad\qquad q_k = A + b(\bullet - \operatorname{mid}(\mathcal{T}))\},\\
CR^1(\mathcal{T}) &:= \{v_{\mathrm{CR}} \in P_1(\mathcal{T}) \,|\, \forall E \in \mathcal{E}(\Omega),\ v_{\mathrm{CR}} \text{ continuous at } \operatorname{mid}(E)\},\\
CR^1_0(\mathcal{T}) &:= \{v_{\mathrm{CR}} \in CR^1(\mathcal{T}) \,|\, \forall E \in \mathcal{E}(\partial\Omega),\ v_{\mathrm{CR}}(\operatorname{mid}(E)) = 0\},\\
P_k(\mathcal{E}) &:= \{t_k \in L^2(\partial\mathcal{T}) \,|\, t_k|_E \in P_k(E) \text{ for any } E \in \mathcal{E}\}.
\end{aligned}
$$

Definition 3.4 For a triangulation $\mathcal{T}$ with skeleton $\partial\mathcal{T} := \bigcup_{T\in\mathcal{T}} \bigcup_{E\in\mathcal{E}(T)} E$ and $T \in \mathcal{T}$, recall the local trace spaces $H^{1/2}(\partial T)$ and $H^{-1/2}(\partial T) = (H^{1/2}(\partial T))^\star$ and

$$
\begin{aligned}
H^{-1/2}(\partial\mathcal{T}) := \{t = (t_T)_{T\in\mathcal{T}} \in \textstyle\prod_{T\in\mathcal{T}} H^{-1/2}(\partial T) \,|\\
\exists q \in H(\operatorname{div},\Omega), \forall T \in \mathcal{T}, t_T = (q|_T)|_{\partial T} \cdot \nu_T\}
\end{aligned}
$$

endowed with the minimal extension norm, for $t \in H^{-1/2}(\partial\mathcal{T})$,

$$
\|t\|_{H^{-1/2}(\partial\mathcal{T})} := \min\{\|q\|_{H(\operatorname{div},\Omega)} \,|\, q \in H(\operatorname{div},\Omega), \forall T \in \mathcal{T}, t_T = (q|_T)|_{\partial T} \cdot \nu_T\}.
$$

The duality brackets $\langle \bullet, \bullet \rangle_{\partial T}$ in $H^{-1/2}(\partial T) \times H^{1/2}(\partial T)$ extend the L^2 scalar product in $L^2(\partial T)$ and lead to the duality bracket on the skeleton for any $t = (t_T)_{T\in\mathcal{T}} \in \prod_{T\in\mathcal{T}} H^{-1/2}(\partial T)$ and $s = (s_T)_{T\in\mathcal{T}} \in \prod_{T\in\mathcal{T}} H^{1/2}(\partial T)$ defined by

$$
\langle t, s \rangle_{\partial\mathcal{T}} := \sum_{T\in\mathcal{T}} \langle t_T, s_T \rangle_{\partial T}.
$$

Remark 3.5 ($RT_0(\mathcal{T}) \equiv P_0(\mathcal{E})$)

The spaces $RT_0(\mathcal{T})$ and $P_0(\mathcal{E})$ are isomorphic [8, Lemma 3.2] in the sense that any $q_{\mathrm{RT}} \in RT_0(\mathcal{T})$ and $E \in \mathcal{E}$ with fixed unit normal vector ν_E satisfies $q_{\mathrm{RT}}|_E \cdot \nu_E \in P_0(E)$. Conversely, for any $t_0 \in P_0(\mathcal{E})$, there exists a unique $q_{\mathrm{RT}} \in RT_0(\mathcal{T})$ with $q_{\mathrm{RT}}|_E \cdot \nu_E = t_0|_E$ for any $E \in \mathcal{E}$, in short notation $q_{\mathrm{RT}} \cdot \nu = t_0$ in $\partial\mathcal{T}$. Since $\|t_0\|_{H^{-1/2}(\partial\mathcal{T})} \approx \|q_{\mathrm{RT}}\|_{H(\mathrm{div},\Omega)}$, this identification justifies the embedding $P_0(\mathcal{E}) \subseteq H^{-1/2}(\partial\mathcal{T})$, where any $T \in \mathcal{T}$ and $E \in \mathcal{E}(T)$ satisfy $(q_{\mathrm{RT}} \cdot \nu_T)|_E = \pm t_0|_E$ with the sign $\pm = \nu_T \cdot \nu_E$ depending on the (globally fixed) choice of the orientation of the unit normal $\nu_E \in \{\nu_{T_\pm}|_E\}$.

Definition 3.6 Define $S_0 \in P_0(\mathcal{T}; \mathbb{R}^{n\times n})$ and $H_0 : L^2(\Omega) \to P_0(\mathcal{T}; \mathbb{R}^n)$ for $T \in \mathcal{T}$ and $f \in L^2(\Omega)$ by

$$\begin{aligned} S_0|_T &:= \Pi_0((\bullet - \mathrm{mid}(T)) \otimes (\bullet - \mathrm{mid}(T))), \\ H_0 f &:= \Pi_0(f(\bullet - \mathrm{mid}(\mathcal{T}))) \in P_0(\mathcal{T}; \mathbb{R}^n). \end{aligned} \tag{3.3}$$

Remark 3.7 An analysis of the eigenvalues of the piecewise symmetric positive semi-definite matrix S_0 shows that any $T \in \mathcal{T}$ and $v \in \mathbb{R}^n$ satisfies

$$|v| \le |(I_{n\times n} + S_0|_T)v| \le (1 + h_T^2)|v| \text{ and } |v| \le |(I_{n\times n} + S_0|_T)^{1/2}v| \le (1 + h_T)|v|.$$

Furthermore, $\|H_0 f\|_{L^2(\Omega)} \le h_{\max}\|(1 - \Pi_0) f\|_{L^2(\Omega)}$ for the maximal mesh-size $h_{\max} = \max h_{\mathcal{T}}$ in $\mathcal{T}$.

3.3 Lowest-order dPG discretization

The nonlinear model problem of this paper concerns the nonlinear map $\sigma : \mathbb{R}^n \to \mathbb{R}^n$ of Sect. 3.1. A piecewise integration by parts in (3.1) and the introduction of the new variable $t := \sigma(\nabla u) \cdot \nu$ on $\partial\mathcal{T}$ leads to the nonlinear primal dPG method with $F(v) := \int_\Omega f v \,\mathrm{d}x$ and $b : X \times Y \to \mathbb{R}$ for $X := H_0^1(\Omega) \times H^{-1/2}(\partial\mathcal{T})$ and $Y := H^1(\mathcal{T})$ defined by

$$b(u, t; v) := \int_\Omega \sigma(\nabla u) \cdot \nabla_{\mathrm{NC}} v \,\mathrm{d}x - \langle t, v\rangle_{\partial\mathcal{T}} =: \langle B(u, t), y\rangle_Y. \tag{3.4}$$

for all $x = (u, t) \in X := H_0^1(\Omega) \times H^{-1/2}(\partial\mathcal{T})$ and $y = v \in Y = H^1(\mathcal{T})$ with associated norms and the scalar product a in Y. Given the subspaces $X_h := S_0^1(\mathcal{T}) \times P_0(\mathcal{E})$ and $Y_h := P_1(\mathcal{T})$, the discrete problem minimizes the residual norm and seeks $(u_h, t_h) = x_h \in X_h$ with

$$\|F - B(x_h)\|_{Y_h^*} = \min_{\xi_h \in X_h} \|F - B(\xi_h)\|_{Y_h^*}. \tag{3.5}$$

The derivative $\mathrm{D}\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ gives rise to the map

$$b'(u, t; w, s, v) := \int_\Omega \nabla w \cdot (\mathrm{D}\sigma(\nabla u)\, \nabla_{\mathrm{NC}} v) \,\mathrm{d}x - \langle s, v\rangle_{\partial\mathcal{T}}. \tag{3.6}$$

This defines a bounded bilinear form $b'(u,t;\bullet) : X \times Y \to \mathbb{R}$ for any $x = (u,t) \in X$ and the operator B associated with b belongs to $C^1(X; Y^*)$. Recall the equivalent mixed formulation from (M) for the model problem at hand, which seeks $(u_h, t_h) \in X_h$ and $v_h \in Y_h$ with

$$\begin{aligned} a(v_h,\eta_h) + b(u_h,t_h;\eta_h) &= F(\eta_h) \quad \text{for all } \eta_h \in Y_h,\\ b'(u_h,t_h;w_h,s_h,v_h) &= 0 \quad \text{for all } (w_h,s_h) \in X_h. \end{aligned} \tag{3.7}$$

Remark 3.8 (regular solution) Since $\mathrm{D}\sigma(\nabla u) \in L^\infty(\Omega; \mathbb{R}^{n\times n}_{\mathrm{sym}})$ uniformly positive definite, the splitting lemma from the linear theory [6, Thm. 3.3] implies the inf-sup condition (2.1) for the nondegenerate bilinear form $b'(x;\bullet,\bullet) : X \times Y \to \mathbb{R}$. Hence, the solution $x \in X$ to $B(x) = F$ is regular.

3.4 Reduced discretization

The dPG discretization (3.5) can be simplified to a modified problem that seeks $(u_h, v_h) \in S^1_0(\mathcal{T}) \times CR^1_0(\mathcal{T})$ with

$$\begin{aligned} a(v_h, w_{\mathrm{CR}}) + \int_\Omega \sigma(\nabla u_h)\cdot\nabla_{\mathrm{NC}} w_{\mathrm{CR}}\,\mathrm{d}x &= \int_\Omega f\, w_{\mathrm{CR}}\,\mathrm{d}x \text{ for all } w_{\mathrm{CR}} \in CR^1_0(\mathcal{T}),\\ \int_\Omega \nabla w_{\mathrm{C}}\cdot\big(\mathrm{D}\sigma(\nabla u_h)\,\nabla_{\mathrm{NC}} v_h\big)\,\mathrm{d}x &= 0 \quad \text{for all } w_{\mathrm{C}} \in S^1_0(\mathcal{T}). \end{aligned} \tag{R}$$

Theorem 3.9 *((3.7) ⇔ (R))*
(a) *If $(u_h, t_h; v_h) \in X_h \times Y_h$ solves (3.7), then $v_h \in CR^1_0(\mathcal{T})$ and $(u_h, v_h) \in S^1_0(\mathcal{T}) \times CR^1_0(\mathcal{T})$ solves (R).*
(b) *For any solution $(u_h, v_h) \in S^1_0(\mathcal{T}) \times CR^1_0(\mathcal{T})$ to (R), there exists a unique $t_h \in P_0(\mathcal{E})$ such that $(u_h, t_h; v_h)$ solves (3.7).*

The proof utilizes the following discrete inf-sup condition of a *linear* primal dPG method [19]. Let the bilinear forms $a_{\mathrm{NC}} : H^1(\mathcal{T}) \times H^1(\mathcal{T}) \to \mathbb{R}$ and $\widetilde{b} : X \times Y \to \mathbb{R}$ be defined by

$$\begin{aligned} a_{\mathrm{NC}}(v_1, v_2) &:= \int_\Omega \nabla_{\mathrm{NC}} v_1 \cdot \nabla_{\mathrm{NC}} v_2\,\mathrm{d}x \quad \text{for } v_1, v_2 \in H^1(\mathcal{T}),\\ \widetilde{b}(x, y) &:= a_{\mathrm{NC}}(u, w) - \langle t, w\rangle_{\partial\mathcal{T}} \quad \text{for } x = (u,t) \in X,\ y = w \in Y. \end{aligned}$$

Lemma 3.10 *The bilinear form $\widetilde{b} : X_h \times Y_h \to \mathbb{R}$ satisfies the discrete inf-sup condition*

$$0 < \widetilde{\beta}_h := \inf_{\xi_h \in \mathcal{S}(X_h)} \sup_{\eta_h \in \mathcal{S}(Y_h)} \widetilde{b}(\xi_h, \eta_h). \tag{3.8}$$

Proof The proof follows the arguments from [8, Thm. 3.5] for the bilinear form $\widetilde{b}$ in the lowest-order scheme at hand. □

Proof (of Theorem 3.9) (a) Since $b'(x_h; 0, s_h, v_h) = -\langle s_h, v_h\rangle_{\partial\mathcal{T}} = 0$ for all $s_h \in P_0(\mathcal{E})$, $v_h \in CR_0^1(\mathcal{T})$. Then, (3.7) reduces to (R).

(b) Conversely, suppose (u_h, v_h) solves (R), then the second equation in (3.7) follows from the second equation in (R) and $v_h \in CR_0^1(\mathcal{T})$. The first equation in (R) leads to the first equation in (3.7) for any $t_h \in P_0(\mathcal{E})$ and test functions in $CR_0^1(\mathcal{T})$. In other words, the linear functional

$$\Lambda_h := a(v_h, \bullet) + \int_\Omega \sigma(\nabla u_h) \cdot \nabla_{\mathrm{NC}} \bullet \,\mathrm{d}x - F \in Y_h^*$$

vanishes on $CR_0^1(\mathcal{T}) \subset \ker \Lambda_h$. It remains to show that there exists $t_h \in P_0(\mathcal{E})$ with $\langle t_h, \bullet\rangle_{\partial\mathcal{T}} = \Lambda_h$, because then (u_h, t_h, v_h) solves (3.7). To prove the existence of such a t_h for $\Lambda_h \in Y_h^*$ with $CR_0^1(\mathcal{T}) \subset \ker \Lambda_h$, recall the bilinear form $\widetilde{b}$ from Lemma 3.10 with discrete inf-sup condition (3.8) and consider the linear problem that seeks $(u_h, t_h, v_h) \in X_h \times Y_h$ with

$$\begin{aligned} a_{\mathrm{NC}}(v_h, w_1) + \widetilde{b}(u_h, t_h, w_1) &= -\Lambda_h(w_1) \quad \text{for all } w_1 \in Y_h, \\ \widetilde{b}(w_C, s_0, v_h) &= 0 \quad \text{for all } (w_C, s_0) \in X_h. \end{aligned} \tag{L}$$

Since $\langle r_0, v_1\rangle_{\partial\mathcal{T}} = 0$ for $v_1 \in P_1(\mathcal{T})$ and for all $r_0 \in P_0(\mathcal{E})$ implies $v_1 \in CR_0^1(\mathcal{T}) \subset P_1(\mathcal{T})$, the kernel

$$Z_h := \{v_1 \in P_1(\mathcal{T}) \,|\, \widetilde{b}(x_h, v_1) = 0 \text{ for all } x_h \in X_h\}$$

of $\widetilde{b}$ consists of particular Crouzeix–Raviart functions, $Z_h \subset CR_0^1(\mathcal{T})$, and the discrete Friedrichs inequality [4, p. 301] shows that a_{NC} is Z_h-elliptic.

Hence, the Brezzi splitting lemma [3, Thm. 4.3 in Ch. III] applies to the linear system (L) and (L) has a unique solution $(u_h, t_h, v_h) \in X_h \times Y_h$. The test of the first equation in (L) with $w_1 \in CR_0^1(\mathcal{T}) \subset \ker \Lambda_h$ shows $a_{\mathrm{NC}}(v_h + u_h, \bullet) = 0$ in $CR_0^1(\mathcal{T})$. The second equation in (L) implies $v_h \in CR_0^1(\mathcal{T})$ and this proves $v_h = -u_h$. This leads to $\langle t_h, \bullet\rangle_{\partial\mathcal{T}} = \Lambda_h$ in $P_1(\mathcal{T})$. The uniqueness of t_h follows from the fact that $\langle t_h, \bullet\rangle_{\partial\mathcal{T}} = 0$ in $P_1(\mathcal{T})$ implies $t_h = 0$. □

3.5 Least-squares formulation

Recall $S_0 \in P_0(\mathcal{T}; \mathbb{R}^{n\times n})$ and $H_0 : L^2(\Omega) \to P_0(\mathcal{T}; \mathbb{R}^n)$ from (3.3) to define an equivalent least-squares formulation.

Theorem 3.11 *(dPG is LS) Any $x_h = (u_C, t_0) \in X_h$ and $p_{\mathrm{RT}} \in RT_0(\mathcal{T})$ with $p_{\mathrm{RT}} \cdot \nu = t_0$ in $\partial\mathcal{T}$ satisfy*

$$\begin{aligned} \|F - b(x_h; \bullet)\|_{Y_h^*}^2 = \|(I_{n\times n} + S_0)^{-1/2}\big(\Pi_0 p_{\mathrm{RT}} - \sigma(\nabla u_C) + H_0 f\big)\|_{L^2(\Omega)}^2 \\ + \|\Pi_0 f + \operatorname{div} p_{\mathrm{RT}}\|_{L^2(\Omega)}^2. \end{aligned} \tag{3.9}$$

Consequently, any solution $x_h = (u_C, t_0) \in X_h$ to (3.7) and $p_{\mathrm{RT}} \cdot \nu = t_0$ in $\partial\mathcal{T}$ from Remark 3.5 minimizes the weighted least-squares functional (3.9).

Proof Let $v_1 \in P_1(\mathcal{T}) \equiv Y_h$ be the Riesz representation of $b(x_h; \bullet) - F \in Y_h^*$, i.e., any $w_1 \in P_1(\mathcal{T})$ satisfies

$$a(v_1, w_1) = b(x_h; w_1) - F(w_1).$$

The substitution of $t_0 = p_{\mathrm{RT}} \cdot \nu$ based on the isometry in Remark 3.5 and an integration by parts lead to

$$b(x_h; w_1) - F(w_1) = \int_\Omega (\sigma(\nabla u_C) - p_{\mathrm{RT}}) \cdot \nabla_{\mathrm{NC}} w_1 \,\mathrm{d}x - \int_\Omega (f + \operatorname{div} p_{\mathrm{RT}}) w_1 \,\mathrm{d}x.$$

With $w_1 = \Pi_0 w_1 + \nabla_{\mathrm{NC}} w_1 \cdot (\bullet - \mathrm{mid}(\mathcal{T}))$, this results in

$$\begin{aligned}\int_\Omega \Pi_0 v_1 \Pi_0 w_1 \,\mathrm{d}x + \int_\Omega (I_{n\times n} + S_0) \nabla_{\mathrm{NC}} v_1 \cdot \nabla_{\mathrm{NC}} w_1 \,\mathrm{d}x \\ = \int_\Omega (\sigma(\nabla u_C) - \Pi_0 p_{\mathrm{RT}}) \cdot \nabla_{\mathrm{NC}} w_1 \,\mathrm{d}x - \int_\Omega (\Pi_0 f + \operatorname{div} p_{\mathrm{RT}}) \Pi_0 w_1 \,\mathrm{d}x \\ - \int_\Omega H_0 f \cdot \nabla_{\mathrm{NC}} w_1 \,\mathrm{d}x.\end{aligned}$$

For any $T \in \mathcal{T}$, the choices $w_1 = \chi_T$ and $w_1 = \chi_T e_k \cdot (\bullet - \mathrm{mid}(\mathcal{T}))$, $k = 1, \ldots, n$, show

$$\begin{aligned}\Pi_0 v_1 &= -(\operatorname{div} p_{\mathrm{RT}} + \Pi_0 f), \\ (I_{n\times n} + S_0) \nabla_{\mathrm{NC}} v_1 &= \sigma(\nabla u_C) - \Pi_0 p_{\mathrm{RT}} - H_0 f.\end{aligned}$$

The Riesz isometry and $\|v_1\|_{H^1(\Omega)}^2 = \|\Pi_0 v_1\|_{L^2(\Omega)}^2 + \|(I_{n\times n} + S_0)^{1/2} \nabla_{\mathrm{NC}} v_1\|_{L^2(\Omega)}^2$ for any $v_1 \in P_1(\mathcal{T})$ conclude the proof. □

4 Mathematical analysis of dPG for the model problem

This section analyses the low-order dPG method presented in Sect. 3 and proves an a posteriori result next to the existence of a solution and applies the abstract framework from Sect. 2. Recall the discrete spaces $X_h := S_0^1(\mathcal{T}) \times P_0(\mathcal{E})$, $Y_h := P_1(\mathcal{T})$, and the nonlinear map from (3.4).

4.1 Well-posedness

This subsection is devoted to the equivalence of the dPG residuals and the errors. For $q_{\mathrm{RT}} \in RT_0(\mathcal{T})$ and $v_C \in S_0^1(\mathcal{T})$, the isomorphism between $RT_0(\mathcal{T})$ and $P_0(\mathcal{E})$ from Remark 3.5 leads to the abbreviation $b(v_C, q_{\mathrm{RT}}; \bullet) := b((v_C, (q_{\mathrm{RT}} \cdot \nu_T)_{T\in\mathcal{T}}); \bullet)$. Recall the energy norm $|||\bullet||| = \|\nabla \bullet\|_{L^2(\Omega)}$ in $H_0^1(\Omega)$.

Theorem 4.1 *The exact solution $u \in H^1_0(\Omega)$ to the model problem (3.1) with stress $p := \sigma(\nabla u) \in H(\mathrm{div}, \Omega)$ and any discrete $(v_C, q_{\mathrm{RT}}) \in S^1_0(\mathcal{T}) \times RT_0(\mathcal{T})$ satisfy the equivalence*

$$\|p - q_{\mathrm{RT}}\|^2_{H(\mathrm{div},\Omega)} + |||u - v_C|||^2 \approx \|F - b(v_C, q_{\mathrm{RT}}; \bullet)\|^2_{Y_h^*} + \|(1 - \Pi_0) f\|^2_{L^2(\Omega)} + \|(1 - \Pi_0) q_{\mathrm{RT}}\|^2_{L^2(\Omega)}.$$

The proof is based on a lemma on the nonlinear least-squares formulation. The related least-squares formulation is associated with the nonlinear residual $\mathcal{R}(f; \bullet) : H(\mathrm{div}, \Omega) \times H^1_0(\Omega) \to L^2(\Omega) \times L^2(\Omega; \mathbb{R}^n)$ for the first-order system of (3.1) and defined, for $(p, u) \in H(\mathrm{div}, \Omega) \times H^1_0(\Omega)$, by

$$\mathcal{R}(f; p, u) := (f + \mathrm{div}\, p, p - \sigma(\nabla u)).$$

Lemma 4.2 *Any $(p, u), (q, v) \in H(\mathrm{div}, \Omega) \times H^1_0(\Omega)$ satisfy*

$$\|\mathcal{R}(f; p, u) - \mathcal{R}(f; q, v)\|^2_{L^2(\Omega)} \approx \|p - q\|^2_{H(\mathrm{div},\Omega)} + |||u - v|||^2.$$

Proof Following [23, Thm. 4.4], the fundamental theorem of calculus shows

$$\begin{aligned}\mathcal{R}(f; q, v) - \mathcal{R}(f; p, u) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} \mathcal{R}(f; p + s(q - p), u + s(v - u))\, \mathrm{d}s \\ &= \int_0^1 \mathcal{R}'(p + s(q - p), u + s(v - u); q - p, v - u)\, \mathrm{d}s.\end{aligned}$$

For $x \in \Omega$ and $0 \le s \le 1$, define $F(s) := \nabla u(x) + s\, \nabla(v - u)(x)$ and

$$M(x) := \int_0^1 \Big(\phi(|F(s)|) I_{n\times n} + \phi'(|F(s)|) \frac{F(s) \otimes F(s)}{|F(s)|}\Big)\, \mathrm{d}s.$$

Then

$$\|\mathcal{R}(f; q, v) - \mathcal{R}(f; p, u)\|^2_{L^2(\Omega)} = \|\mathrm{div}(q - p)\|^2_{L^2(\Omega)} + \|q - p - M\, \nabla(v - u)\|^2_{L^2(\Omega)}.$$

Since the assumptions on ϕ show that $M \in L^2(\Omega; \mathbb{R}^{n\times n})$ is pointwise symmetric and positive definite with eigenvalues in the real compact interval $[\gamma_1, \gamma_2] \subset (0, \infty)$, the triangle inequality shows

$$\|\mathcal{R}(f; q, v) - \mathcal{R}(f; p, u)\|^2_{L^2(\Omega)} \le 2 \max\{1, \gamma_2^2\}\big(\|q - p\|_{H(\mathrm{div},\Omega)} + |||v - u|||^2\big).$$

For the reverse estimate, the positive definiteness of M provides the unique existence of a solution $\alpha \in H^1_0(\Omega)$ to the weighted problem

$$\int_\Omega M\, \nabla \alpha \cdot \nabla \gamma\, \mathrm{d}x = \int_\Omega (q - p) \cdot \nabla \gamma\, \mathrm{d}x \text{ for any } \gamma \in H^1_0(\Omega).$$

An integration by parts shows $r := q - p - M\nabla\alpha \in H(\mathrm{div}, \Omega)$ with $\mathrm{div}\, r = 0$. The Friedrichs inequality with constant C_{F} (i.e. $\|\alpha\|_{L^2(\Omega)} \le C_{\mathrm{F}}|||\alpha|||$) implies

$$\|M^{1/2}\nabla\alpha\|^2_{L^2(\Omega)} = \int_\Omega \mathrm{div}(p-q)\alpha\,\mathrm{d}x \le C_{\mathrm{F}}/\gamma_1 \|\mathrm{div}(q-p)\|_{L^2(\Omega)} \|M^{1/2}\nabla\alpha\|_{L^2(\Omega)}.$$

The orthogonality of $\nabla H^1_0(\Omega)$ and $H(\mathrm{div}, \Omega) \cap \{\mathrm{div} = 0\}$ in $L^2(\Omega)$ shows

$$\begin{aligned}\|q-p\|^2_{L^2(\Omega)} &= \|M\nabla\alpha + r\|^2_{L^2(\Omega)} \le \gamma_2 \|M^{1/2}\nabla\alpha + M^{-1/2}r\|^2_{L^2(\Omega)} \\ &= \gamma_2\|M^{1/2}\nabla\alpha\|^2_{L^2(\Omega)} + \gamma_2\|M^{-1/2}r\|^2_{L^2(\Omega)}.\end{aligned}$$

The two previous displayed inequalities, the triangle inequality, and the abbreviation $e := v - u$ yield

$$\begin{aligned}\|q-p\|^2_{H(\mathrm{div},\Omega)} + |||e|||^2 &\le (2+\gamma_2)\|M^{1/2}\nabla\alpha\|^2_{L^2(\Omega)} + \|\mathrm{div}(q-p)\|^2_{L^2(\Omega)} \\ &\quad + \max\{2,\gamma_2\}\big(\|M^{1/2}\nabla(\alpha-e)\|^2_{L^2(\Omega)} + \|M^{-1/2}r\|^2_{L^2(\Omega)}\big) \\ &\le ((2+\gamma_2)C_{\mathrm{F}}^2/\gamma_1^2 + 1)\|\mathrm{div}(q-p)\|^2_{L^2(\Omega)} \\ &\quad + \max\{2,\gamma_2\}/\gamma_1 \|q-p-M\nabla e\|^2_{L^2(\Omega)}.\end{aligned}$$

□

Proof (of Theorem 4.1) Since $\mathcal{R}(f; p, u) = 0$, Lemma 4.2 with $(q, v) := (q_{\mathrm{RT}}, v_C)$ shows

$$\|p - q_{\mathrm{RT}}\|^2_{H(\mathrm{div},\Omega)} + |||u - v_C|||^2 \approx \|f + \mathrm{div}\, q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|q_{\mathrm{RT}} - \sigma(\nabla v_C)\|^2_{L^2(\Omega)}.$$

The L^2-orthogonality of $(1-\Pi_0)q_{\mathrm{RT}}$ and $(1-\Pi_0)f$ onto piecewise constants implies

$$\begin{aligned}&\|f + \mathrm{div}\, q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|q_{\mathrm{RT}} - \sigma(\nabla v_C)\|^2_{L^2(\Omega)} \\ &\quad = \|\Pi_0 f + \mathrm{div}\, q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|(1-\Pi_0)q_{\mathrm{RT}}\|^2_{L^2(\Omega)} \\ &\qquad + \|(1-\Pi_0)f\|^2_{L^2(\Omega)} + \|\Pi_0 q_{\mathrm{RT}} - \sigma(\nabla v_C)\|^2_{L^2(\Omega)}.\end{aligned}$$

The triangle inequality and the estimates of Remark 3.7 result in

$$\begin{aligned}&\|p - q_{\mathrm{RT}}\|^2_{H(\mathrm{div},\Omega)} + |||u - v_C|||^2 \\ &\quad \lesssim \|\Pi_0 f + \mathrm{div}\, q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|(1-\Pi_0)q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|(1-\Pi_0)f\|^2_{L^2(\Omega)} \\ &\qquad + \|H_0 f\|^2_{L^2(\Omega)} + \|\Pi_0 q_{\mathrm{RT}} - \sigma(\nabla v_C) + H_0 f\|^2_{L^2(\Omega)} \\ &\quad \lesssim \|\Pi_0 f + \mathrm{div}\, q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|(1-\Pi_0)q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|(1-\Pi_0)f\|^2_{L^2(\Omega)} \\ &\qquad + \|\Pi_0 q_{\mathrm{RT}} - \sigma(\nabla v_C) + H_0 f\|^2_{L^2(\Omega)}.\end{aligned}$$

Recall that S_0 is pointwise positive semi-definite, hence $I_{n\times n}+S_0$ is positive definite and Remark 3.7 also proves

$$\|\Pi_0 q_{\mathrm{RT}}-\sigma(\nabla v_C)+H_0 f\|^2_{L^2(\Omega)} \approx \|(I_{n\times n}+S_0)^{-1/2}(\Pi_0 q_{\mathrm{RT}}-\sigma(\nabla v_C)+H_0 f)\|^2_{L^2(\Omega)}.$$

The proof of the converse estimate utilizes the last estimate and the triangle inequality to show

$$\begin{aligned}\|(I_{n\times n}+S_0)^{-1/2}(\Pi_0 q_{\mathrm{RT}}-\sigma(\nabla v_C)+H_0 f)\|^2_{L^2(\Omega)} &+ \|(1-\Pi_0)f\|^2_{L^2(\Omega)}\\ &\lesssim \|\Pi_0 q_{\mathrm{RT}}-\sigma(\nabla v_C)\|^2_{L^2(\Omega)} + \|H_0 f\|^2_{L^2(\Omega)} + \|(1-\Pi_0)f\|^2_{L^2(\Omega)}\\ &\lesssim \|\Pi_0 q_{\mathrm{RT}}-\sigma(\nabla v_C)\|^2_{L^2(\Omega)} + \|(1-\Pi_0)f\|^2_{L^2(\Omega)}.\end{aligned}$$

This and the aforementioned orthogonalities imply

$$\begin{aligned}\|(I_{n\times n}+S_0)^{-1/2}(\Pi_0 q_{\mathrm{RT}}-\sigma(\nabla v_C)+H_0 f)\|^2_{L^2(\Omega)} &+ \|(1-\Pi_0)f\|^2_{L^2(\Omega)}\\ &+ \|(1-\Pi_0)q_{\mathrm{RT}}\|^2_{L^2(\Omega)} + \|\Pi_0 f+\operatorname{div} q_{\mathrm{RT}}\|^2_{L^2(\Omega)}\\ &\lesssim \|q_{\mathrm{RT}}-\sigma(\nabla v_C)\|^2_{L^2(\Omega)} + \|f+\operatorname{div} q_{\mathrm{RT}}\|^2_{L^2(\Omega)}\\ &\lesssim \|p-q_{\mathrm{RT}}\|^2_{H(\mathrm{div},\Omega)} + |||u-v_C|||^2.\end{aligned}$$

□

4.2 Existence and uniqueness of discrete solutions

The existence of discrete solutions follows from variational arguments, while their uniqueness is fairly open.

Proposition 4.3 *The discrete problem* (3.5) *has a solution.*

Proof The proof follows with the direct method in the calculus of variations and, in the present case of finite dimensions, from the global minimum of a continuous functional on a compact set from the growth condition

$$\lim_{\|\xi_h\|_X\to\infty} \|F-B\xi_h\|_{Y_h^*} = \infty. \tag{4.1}$$

The latter property follows from Theorem 4.1 up to some perturbation terms. Theorem 3.11 shows

$$\begin{aligned}\|(1-\Pi_0)q_{\mathrm{RT}}\|_{L^2(\Omega)} &\le \|h_{\mathcal{T}}\Pi_0 f\|_{L^2(\Omega)} + h_{\max}\|\Pi_0 f+\operatorname{div} q_{\mathrm{RT}}\|_{L^2(\Omega)}\\ &\le \|h_{\mathcal{T}}\Pi_0 f\|_{L^2(\Omega)} + h_{\max}\|F-B\xi_h\|_{Y_h^*}.\end{aligned}$$

The combination with Theorem 4.1 shows that the right-hand side of Theorem 4.1 is bounded from above by

$$\begin{aligned}(1+2h_{\max}^2)\|F-B\xi_h\|_{Y_h^*}^2+2h_{\max}^2\|\Pi_0 f\|_{L^2(\Omega)}^2+\|(1-\Pi_0)f\|_{L^2(\Omega)}^2\\ \leq (1+2h_{\max}^2)(\|F-B\xi_h\|_{Y_h^*}+\|f\|_{L^2(\Omega)}).\end{aligned}$$

Hence the left-hand side in Theorem 4.1 is controlled by this and so

$$\|x-\xi_h\|_X \lesssim \|F-B\xi_h\|_{Y_h^*}+\|f\|_{L^2(\Omega)}.$$

Since f and x are fixed, this implies (4.1) and concludes the proof. □

The uniqueness of the exact solution (u,t) on the continuous level does not imply the uniqueness of discrete solutions. There is, however, a sufficient condition for a global unique discrete solution. Notice that $v_h = v = 0$ on the continuous level $h = 0$ satisfies (4.2).

Theorem 4.4 *(a posteriori uniqueness) Suppose that* $(u_h, v_h) \in S_0^1(\mathcal{T}) \times CR_0^1(\mathcal{T})$ *solves* (R) *with* $\mathrm{D}\sigma \in C(\mathbb{R}^n; \mathbb{R}^{n\times n}_{sym})$ *globally Lipschitz continuous and*

$$Lip(\mathrm{D}\sigma)(1+C_{\mathrm{F}}^2)/\gamma_1^2\,\|\nabla_{\mathrm{NC}}\, v_h\|_{L^\infty(\Omega)} < 1 \tag{4.2}$$

with the Friedrichs constant C_{F} *from* $\|\bullet\|_{L^2(\Omega)} \leq C_{\mathrm{F}}|||\bullet|||$ *in* $H_0^1(\Omega)$. *Then* (R) *has exactly one solution* $(u_h, v_h) \in S_0^1(\mathcal{T}) \times CR_0^1(\mathcal{T})$.

Proof Suppose that $(\widetilde{u}_h, \widetilde{v}_h) \in S_0^1(\mathcal{T}) \times CR_0^1(\mathcal{T})$ solves (R) as well and so

$$\begin{aligned}a(v_h, v_h) &= F(v_h) - \int_\Omega \sigma(\nabla u_h)\cdot\nabla_{\mathrm{NC}}\, v_h\,\mathrm{d}x\\ &= a(\widetilde{v}_h, v_h) + \int_\Omega (\sigma(\nabla\widetilde{u}_h)-\sigma(\nabla u_h))\cdot\nabla_{\mathrm{NC}}\, v_h\,\mathrm{d}x.\end{aligned}$$

This and the second equation of (R) imply

$$a(v_h-\widetilde{v}_h, v_h) = \int_\Omega \nabla_{\mathrm{NC}}\, v_h\cdot(\sigma(\nabla\widetilde{u}_h)-\sigma(\nabla u_h)-\mathrm{D}\sigma(\nabla u_h)\nabla(\widetilde{u}_h-u_h))\,\mathrm{d}x.$$

Since $\sigma \in C^1(\mathbb{R}^n)$ is bounded and $\mathrm{D}\sigma$ Lipschitz continuous, any $A, B \in \mathbb{R}^n$ with $F(s) := (1-s)A + sB$ for $0 \leq s \leq 1$ satisfy

$$\begin{aligned}|\sigma(B)-\sigma(A)-\mathrm{D}\sigma(A)(B-A)| &= \Big|\int_0^1 (\mathrm{D}\sigma(F(s))-\mathrm{D}\sigma(A))(B-A)\,\mathrm{d}s\Big|\\ &\leq \mathrm{Lip}(\mathrm{D}\sigma)|B-A|\int_0^1 |F(s)-A|\,\mathrm{d}s = \frac{1}{2}\mathrm{Lip}(\mathrm{D}\sigma)|B-A|^2.\end{aligned}$$

With $A = \nabla u_h(x)$ and $B = \nabla \widetilde{u}_h(x)$ for a.e. x and an integration over Ω, this leads in the preceding identity to

$$a(v_h - \widetilde{v}_h, v_h) \leq \frac{1}{2}\mathrm{Lip}(\mathrm{D}\sigma) \| \nabla_{\mathrm{NC}} v_h \|_{L^\infty(\Omega)} |||u_h - \widetilde{u}_h|||^2. \tag{4.3}$$

The discrete solutions of (R) lead to the same minimal discrete residual norm and hence

$$\|v_h\|_{Y_h} = \|F - b(u_h, t_h; \bullet)\|_{Y_h^*} = \|F - b(\widetilde{u}_h, \widetilde{t}_h; \bullet)\|_{Y_h^*} = \|\widetilde{v}_h\|_{Y_h}.$$

This shows $a(v_h - \widetilde{v}_h, v_h + \widetilde{v}_h) = 0$ and the combination with (4.3) is

$$\begin{aligned} \|v_h - \widetilde{v}_h\|_{Y_h}^2 &= 2a(v_h - \widetilde{v}_h, v_h) - a(v_h - \widetilde{v}_h, v_h + \widetilde{v}_h) \\ &\leq \mathrm{Lip}(\mathrm{D}\sigma) \| \nabla_{\mathrm{NC}} v_h \|_{L^\infty(\Omega)} |||u_h - \widetilde{u}_h|||^2. \end{aligned} \tag{4.4}$$

On the other hand, $\mathrm{D}\sigma(A) \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ has eigenvalues in the compact interval $[\gamma_1, \gamma_2] \subset (0, \infty)$ and so, for all $A, B \in \mathbb{R}^n$,

$$\begin{aligned} \gamma_1 |A - B|^2 &\leq \int_0^1 (A - B) \cdot \mathrm{D}\sigma(B + s(A - B))(A - B)\, \mathrm{d}s \\ &= (\sigma(A) - \sigma(B)) \cdot (A - B) \leq \gamma_2 |A - B|^2. \end{aligned} \tag{4.5}$$

With $A = \nabla u_h(x)$ and $B = \nabla \widetilde{u}_h(x)$, and an integration over a.e. $x \in \Omega$, this shows

$$\gamma_1 |||u_h - \widetilde{u}_h|||^2 \leq \int_\Omega (\sigma(\nabla u_h) - \sigma(\nabla \widetilde{u}_h)) \cdot \nabla(u_h - \widetilde{u}_h)\, \mathrm{d}x.$$

The first identity in (R) for (u_h, v_h) and $(\widetilde{u}_h, \widetilde{v}_h)$, respectively, results in

$$\begin{aligned} \int_\Omega (\sigma(\nabla u_h) - \sigma(\nabla \widetilde{u}_h)) \cdot \nabla(u_h - \widetilde{u}_h)\, \mathrm{d}x &= a(\widetilde{v}_h - v_h, u_h - \widetilde{u}_h) \\ &\leq \|v_h - \widetilde{v}_h\|_{Y_h} \sqrt{1 + C_{\mathrm{F}}^2}\, |||u_h - \widetilde{u}_h|||. \end{aligned}$$

The combination with the previous inequality shows

$$\gamma_1 |||u_h - \widetilde{u}_h||| \leq \sqrt{1 + C_{\mathrm{F}}^2}\, \|v_h - \widetilde{v}_h\|_{Y_h}. \tag{4.6}$$

The substitution in (4.4) results in

$$\|v_h - \widetilde{v}_h\|_{Y_h}^2 \leq \mathrm{Lip}(\mathrm{D}\sigma)(1 + C_{\mathrm{F}}^2)/\gamma_1^2 \| \nabla_{\mathrm{NC}} v_h \|_{L^\infty(\Omega)} \|v_h - \widetilde{v}_h\|_{Y_h}^2.$$

This and (4.2) show $v_h = \widetilde{v}_h$. Then (4.6) implies $u_h = \widetilde{u}_h$. □

4.3 Best-approximation

For any $v \in H^1(\mathcal{T})$, the *nonconforming interpolation* $I_{\mathrm{NC}}^{\mathrm{loc}} v \in P_1(\mathcal{T})$ is defined, on each triangle $T \in \mathcal{T}$, by piecewise linear interpolation of the values

$$(I_{\mathrm{NC}}^{\mathrm{loc}} v)(\mathrm{mid}(E)) := \fint_E v\big|_T \, \mathrm{d}s \tag{4.7}$$

at the midpoints of the sides $E \in \mathcal{E}(T)$.

Proposition 4.5 *The operator $\Pi := I_{\mathrm{NC}}^{\mathrm{loc}}$ satisfies Hypothesis 2.5.*

Proof Given $v \in H^1(\mathcal{T})$, set $v_h := I_{\mathrm{NC}}^{\mathrm{loc}} v \in P_1(\mathcal{T})$. For every $K \in \mathcal{T}$, an integration by parts leads to

$$\nabla \left(v_h\big|_K \right) = \int_{\partial K} v_h \cdot \nu_K \, \mathrm{d}s = \int_{\partial K} v \cdot \nu_K \, \mathrm{d}s = \int_K \nabla v \, \mathrm{d}x \Big/ |K|.$$

Since $\nabla w_C \in P_0(\mathcal{T}; \mathbb{R}^n)$ and (3.2) shows $\mathrm{D}\sigma(\nabla u_C) \in P_0(\mathcal{T}; \mathbb{R}^{n\times n})$ for all $w_C, u_C \in S_0^1(\mathcal{T})$, this implies

$$\int_\Omega \nabla w_C \cdot (\mathrm{D}\sigma(\nabla u_C) \, \nabla_{\mathrm{NC}}(v - v_h)) \, \mathrm{d}x = 0.$$

Moreover, (4.7) guarantees that any $s_0 \in P_0(\mathcal{E})$ satisfies

$$\langle s_0, v - v_h \rangle_{\partial\mathcal{T}} = 0.$$

Consequently, any $x_h = (u_C, t_0) \in X_h$ and $\xi_h = (w_C, s_0) \in X_h$ satisfy

$$b'(x_h; \xi_h, v - v_h) = \int_\Omega \nabla w_C \cdot (\mathrm{D}\sigma(\nabla u_C) \, \nabla_{\mathrm{NC}}(v - v_h)) \, \mathrm{d}x - \langle s_0, v - v_h \rangle_{\partial\mathcal{T}} = 0.$$

□

The estimates for the function $\mathrm{D}\sigma$ from Sect. 3.1 lead to an explicit generic constant for the best-approximation estimate from Theorem 2.8 without any local hypothesis.

Theorem 4.6 *(best-approximation)*
Let $x = (u, t) \in X$ be the unique solution to $B(x) = F$ for the nonlinear map B from (3.4) in Sect. 3.3. Any discrete solution $(u_h, t_h; v_h) \in X_h \times Y_h$ to (3.5) satisfies

$$|||u - u_h||| + \|v_h\|_Y \lesssim \inf_{u_C \in S_0^1(\mathcal{T})} |||u - u_C||| + \inf_{t_0 \in P_0(\mathcal{E})} \|t - t_0\|_{H^{-1/2}(\partial\mathcal{T})}.$$

Proof Given the best-approximation $x_h^* = (u_h^*, t_h^*) \in X_h$ to (u,t) in X_h and let $A = \nabla u_h^*(x)$ and $B = \nabla u_h(x)$ in (4.5) and integrate over a.e. $x \in \Omega$. Then

$$\begin{aligned}\gamma_1 |||u_h^* - u_h|||^2 &\le \int_\Omega \big(\sigma(\nabla u_h^*) - \sigma(\nabla u_h) \cdot \nabla(u_h^* - u_h)\big)\,\mathrm{d}x \\ &= b(x_h^*; u_h^* - u_h) - b(x_h; u_h^* - u_h).\end{aligned}$$

Since $b(u,t;\bullet) = F$, the last term is equal to

$$\begin{aligned}&F(u_h^* - u_h) - b(x_h; u_h^* - u_h) + b(x_h^*; u_h^* - u_h) - b(u; u_h^* - u_h) \\ &\quad = a(v_h, u_h^* - u_h) + \int_\Omega \big(\sigma(\nabla u_h^*) - \sigma(\nabla u)\big) \cdot \nabla(u_h^* - u_h)\,\mathrm{d}x.\end{aligned} \tag{4.8}$$

The Lipschitz continuity from Remark 3.1 leads to

$$\|\sigma(\nabla u) - \sigma(\nabla u_h^*)\|_{L^2(\Omega)} \le \gamma_2 |||u - u_h^*||| \tag{4.9}$$

and the last term in (4.8) is controlled by

$$\int_\Omega \big(\sigma(\nabla u_h^*) - \sigma(\nabla u)\big) \cdot \nabla(u_h^* - u_h)\,\mathrm{d}x \le \gamma_2 |||u - u_h^*|||\, |||u_h^* - u_h|||. \tag{4.10}$$

Since x_h is a global discrete minimizer,

$$\|v_h\|_Y = \|F - b(x_h;\bullet)\|_{Y_h^*} \le \|F - b(x_h^*;\bullet)\|_{Y_h^*} = \|b(x;\bullet) - b(x_h^*;\bullet)\|_{Y_h^*}.$$

The Lipschitz continuity (4.9) of σ and the structure of the map b from (3.4) show that the last term is $\le \gamma_2 |||u - u_h^*||| + \|t - t_h^*\|_{H^{-1/2}(\partial\mathcal{T})}$. The combination of $\|v_h\|_Y \le \gamma_2 |||u - u_h^*||| + \|t - t_h^*\|_{H^{-1/2}(\partial\mathcal{T})}$ with (4.8) and (4.10) shows

$$\gamma_1 |||u_h^* - u_h||| \le \sqrt{1 + C_{\mathrm{F}}^2}\|v_h\|_Y + \gamma_2 |||u - u_h^*|||.$$

A triangle inequality concludes the proof with explicit constants

$$\begin{aligned}|||u - u_h||| &\le (1 + \gamma_2(1 + \sqrt{1 + C_{\mathrm{F}}^2})/\gamma_1) \inf_{u_{\mathrm{C}} \in S_0^1(\mathcal{T})} |||u - u_{\mathrm{C}}||| \\ &\quad + \gamma_2\sqrt{1 + C_{\mathrm{F}}^2}/\gamma_1 \inf_{t_0 \in P_0(\mathcal{E})} \|t - t_0\|_{H^{-1/2}(\partial\mathcal{T})}.\end{aligned}$$

□

The following a posteriori error estimate holds for any discrete approximation, and even for inexact solve, and generalizes the built-in error control despite inexact solve of [5, Thm. 2.1] to the nonlinear model problem at hand.

Theorem 4.7 *(a posteriori) There exist universal constants* $\kappa \approx 1 \approx C_{\mathrm{dF}}$ *such that the exact solution* $(u, t) \in X$ *of* $B(x) = F$ *and any discrete* $(v_C, s_0) \in S^1_0(\mathcal{T}) \times P_1(\mathcal{T})$ *satisfy*

$$\gamma_1^2 |||u - v_C|||^2 \leq (1 + C_{\mathrm{dF}}^2) \|F - b(v_C, s_0; \bullet)\|^2_{CR^1_0(\mathcal{T})^*} + \kappa^2 \|h_{\mathcal{T}} f\|^2_{L^2(\Omega)}.$$

Remark 4.8 The proof reveals that C_{dF} is the constant in the discrete Friedrichs inequality [4, p. 301] $\|\bullet\| \leq C_{\mathrm{dF}} \|\nabla_{\mathrm{NC}} \bullet\|$ in $CR^1_0(\mathcal{T})$. The explicit bounds of C_{dF} in [10] allow quantitative estimates in 2D and show in particular $C_{\mathrm{dF}} \leq 6.24$ for a convex domain with $\mathrm{diam}(\Omega) \leq 1$ and a triangulation with right isosceles triangles.

Remark 4.9 The proof reveals that κ is the constant in interpolation error estimate for the nonconforming interpolation operator $\|h_{\mathcal{T}}^{-1}(1 - I_{\mathrm{NC}})v\| \leq \kappa \|\nabla_{\mathrm{NC}}(1 - I_{\mathrm{NC}})v\|$ for $v \in H^1(\Omega)$. An estimate with the first positive root $j_{1,1}$ of the Bessel function of the first kind in [7, Thm. 4] in 2D reads $\kappa = (1/48 + 1/j^2_{1,1})^{1/2} = 0.29823$.

Proof (of Theorem 4.7) The estimate (4.5) with $A = \nabla u(x)$, $B = \nabla u_h(x)$, $e := u - u_h$, and an integration over a.e. $x \in \Omega$ leads to

$$\gamma_1 |||e|||^2 \leq \int_\Omega \big(\sigma(\nabla u) - \sigma(\nabla u_h)\big) \cdot \nabla e \, \mathrm{d}x. \tag{4.11}$$

Since $(u, t) \in X$ solves $b(u, t; \bullet) = F$ in Y^* and with the nonconforming interpolation operator (4.7), this is equal to

$$\begin{aligned} F(e) - \int_\Omega \sigma(\nabla u_h) \cdot \nabla e \, \mathrm{d}x &= F((1 - I_{\mathrm{NC}})e) + F(I_{\mathrm{NC}}e) - \int_\Omega \sigma(\nabla u_h) \cdot \nabla_{\mathrm{NC}} I_{\mathrm{NC}} e \, \mathrm{d}x \\ &= F((1 - I_{\mathrm{NC}})e) + F(I_{\mathrm{NC}}e) - b(u_h, t_h; I_{\mathrm{NC}}e) \\ &\leq F((1 - I_{\mathrm{NC}})e) + \|F - b(u_h, t_h; \bullet)\|_{CR^1_0(\mathcal{T})^*} \|I_{\mathrm{NC}}e\|_{Y_h}. \end{aligned}$$

The interpolation error estimate for the nonconforming interpolation operator with constant κ [7, Thm. 4] yields

$$F((1 - I_{\mathrm{NC}})e) \leq \kappa \|h_{\mathcal{T}} f\|_{L^2(\Omega)} |||e - I_{\mathrm{NC}}e|||_{\mathrm{NC}}.$$

The discrete Friedrichs inequality [4, p. 301] and $I_{\mathrm{NC}}e \in CR^1_0(\mathcal{T})$ prove

$$\|I_{\mathrm{NC}}e\|_{Y_h} \leq \sqrt{1 + C_{\mathrm{dF}}^2} |||I_{\mathrm{NC}}e|||_{\mathrm{NC}}.$$

The Cauchy inequality in $\mathbb{R}^2$ and the theorem of Pythagoras imply

$$\gamma_1 |||e|||^2 \leq \big(\kappa^2 \|h_{\mathcal{T}} f\|^2_{L^2(\Omega)} + (1 + C_{\mathrm{dF}}^2) \|F - b(u_h, t_h; \bullet)\|^2_{CR^1_0(\mathcal{T})^*}\big)^{1/2} |||e|||_{\mathrm{NC}}. \qquad \square$$

4.4 Other nonlinear dPG methods

This section illustrates the plethora of dPG methodology by introducing the primal mixed, the dual, and the ultraweak dPG method for the nonlinear model problem. All three methods concern the first-order system of (3.1) with the convex function φ and $\sigma = \mathrm{D}(\varphi \circ |\bullet|)$ and its dual φ^* so that the relation $p = \sigma(\nabla u)$ is equivalent to $\nabla u = \mathrm{D}\varphi^*(|p|)$ sign p on the continuous level. Recall the space of functions with piecewise divergence $H(\mathrm{div}; \mathcal{T}) := \prod_{T\in\mathcal{T}} H(\mathrm{div}; T)$ from [8] as well as the piecewise version $RT_k^{\mathrm{NC}}(\mathcal{T}) \subset H(\mathrm{div}; \mathcal{T})$ of $RT_k(\mathcal{T})$, and the subspace $S_0^k(\mathcal{E}) \equiv S_0^k(\mathcal{T})\big|_{\partial\mathcal{T}}$ of

$$H_0^{1/2}(\partial\mathcal{T}) := \{s = (s_T)_{T\in\mathcal{T}} \in \textstyle\prod_{T\in\mathcal{T}} H^{1/2}(\partial T) \,|\, \exists v \in H_0^1(\Omega), \\ \forall T \in \mathcal{T}, s_T = (v|_T)|_{\partial T}\}.$$

Recall the **primal nonlinear dPG method** (dPG) in Sect. 3.3 with b from (3.4) and general polynomial degree $k \geq 0$ and $m \geq k$ in the discrete spaces

$$X_h := S_0^{k+1}(\mathcal{T}) \times P_k(\mathcal{E}) \text{ and } Y_h := P_{m+1}(\mathcal{T}).$$

The **primal mixed nonlinear dPG method** departs from a piecewise integration by parts and employs the spaces and discrete subspaces

$$X := L^2(\Omega; \mathbb{R}^n) \times H_0^1(\mathcal{T}) \times H^{1/2}(\partial\mathcal{T}) \text{ and } Y := L^2(\Omega; \mathbb{R}^n) \times H^1(\mathcal{T}),$$
$$X_h := P_k(\mathcal{T}; \mathbb{R}^n) \times S_0^{k+1}(\mathcal{T}) \times P_k(\mathcal{E}) \text{ and } Y_h := P_m(\mathcal{T}; \mathbb{R}^n) \times P_{m+1}(\mathcal{T}).$$

For $(p, u, t) \in X$ and $(q, v) \in Y$, define (dPG) with $F(q, v) := (f, v)_{L^2(\Omega)}$ and

$$b(p, u, t; q, v) := (p - \sigma(\nabla u), q)_{L^2(\Omega)} + (p, \nabla_{\mathrm{NC}} v)_{L^2(\Omega)} - \langle t, v\rangle_{\partial\mathcal{T}}.$$

The **dual nonlinear dPG method** utilizes the spaces and discrete subspaces

$$X := H(\mathrm{div}; \Omega) \times L^2(\Omega) \times H_0^{1/2}(\partial\mathcal{T}) \text{ and } Y := H(\mathrm{div}; \mathcal{T}) \times L^2(\Omega),$$
$$X_h := RT_k(\mathcal{T}) \times P_k(\mathcal{T}) \times S_0^{k+1}(\mathcal{E}) \text{ and } Y_h := RT_m^{\mathrm{NC}}(\mathcal{T}) \times P_m(\mathcal{T}).$$

For F as before and $(p, u, s) \in X$ and $(q, v) \in Y$, define (dPG) with $b(p, u, s; q, v) :=$

$$(\mathrm{D}\varphi^*(|p|) \text{ sign } p, q)_{L^2(\Omega)} + (u, \mathrm{div}_{\mathrm{NC}}\, q)_{L^2(\Omega)} - (\mathrm{div}\, p, v)_{L^2(\Omega)} - \langle q \cdot \nu, s\rangle_{\partial\mathcal{T}}.$$

The **ultraweak nonlinear dPG method** utilizes a piecewise integration by parts in both equations of the first-order system and the spaces

$$X := L^2(\Omega; \mathbb{R}^n) \times L^2(\Omega) \times H^{1/2}(\partial\mathcal{T}) \times H_0^{1/2}(\partial\mathcal{T}) \text{ and } Y := H(\mathrm{div}; \mathcal{T}) \times H^1(\mathcal{T}),$$
$$X_h := P_k(\mathcal{T}; \mathbb{R}^n) \times P_k(\mathcal{T}) \times P_k(\mathcal{E}) \times S_0^{k+1}(\mathcal{E}) \text{ and } Y_h := RT_m^{\mathrm{NC}}(\mathcal{T}) \times P_{m+1}(\mathcal{T}).$$

For F from the above primal mixed method and $(p, u, t, s) \in X$ and $(q, v) \in Y$, define (dPG) with

$$b(p, u, t, s; q, v) := (\mathrm{D}\varphi^*(|p|)\,\mathrm{sign}\,p, q)_{L^2(\Omega)} + (p, \nabla_{\mathrm{NC}}\, v)_{L^2(\Omega)} + (u, \mathrm{div}_{\mathrm{NC}}\, q)_{L^2(\Omega)} - \langle q \cdot \nu, s\rangle_{\partial\mathcal{T}} - \langle t, v\rangle_{\partial\mathcal{T}}.$$

The linear version is analysed in [2,6,8,18]. The four nonlinear dPG methods may be further analysed in the spirit of this section.

5 Numerical experiments

This section presents numerical experiments with the LS-FEM of Sect. 3.5.

5.1 Computational realization

Given $f \in L^2(\Omega)$, the discrete solution of (3.7) is computed by a Newton scheme with an initial iterate from the solution of the scaled linear Poisson model problem. Let $S^1_0(\mathcal{T})$ be endowed with the energy norm $|||\bullet|||$ and $RT_0(\mathcal{T})$ with $\|\bullet\|_{H(\mathrm{div},\Omega)}$ and let $|||\bullet|||_*$ denote the norm of the dual space of $S^1_0(\mathcal{T}) \times RT_0(\mathcal{T})$. The first Fréchet derivative $\mathrm{D}LS(f; u_{\mathrm{C}}, p_{\mathrm{RT}})$ of $LS(f; \bullet)$ belongs to the dual space of $S^1_0(\mathcal{T}) \times RT_0(\mathcal{T})$. After at most 5 Newton iterations, every displayed discrete solution (u_h, p_h) in the following subsections satisfy $|||\,\mathrm{D}LS(f; u_h, p_h, \bullet)|||_* = 0$ up to machine precision. In the case of successive mesh-refinement, the iteration starts with the prolongated solution from the coarser triangulation and terminates in at most 3 or 4 iterations.

Table 1 presents the errors $|||\,\mathrm{D}LS(f; u_h^{(j)}, p_h^{(j)}, \bullet)|||_*$ of the Newton iterate $(u_h^{(j)}, p_h^{(j)})$ for $j = 0, 1, \ldots, 5$ on fixed triangulations of the square domain from Sect. 5.2 and the L-shaped domain from Sect. 5.4 with the convex function ϕ from Example 3.2.a. The iterations (A) and (B) utilize a uniform triangulation of the square domain with 4 096 triangles (`ndof` = 8 193) and an initial iterate from a Poisson model problem for (A) and a weighted Poisson problem with constant weight 2.5 in (B). The adaptive mesh of the L-shaped domain with 3 450 triangles (`ndof` = 6.901) in (C) and (D) has been generated by the algorithm from Sect. 5.3 below at level $\ell = 12$. Iteration (C) starts with a weighted Poisson solution with constant factor 2.5 and (D) with the prolongated solution from the previous mesh.

From the very beginning of the Newton iteration, all values in Table 1 provide numerical evidence for Q-quadratic convergence.

In order to investigate the uniqueness of discrete solutions, the minimal and the maximal eigenvalue $\lambda_{\min}$ and $\lambda_{\max}$ of the Hessian matrix $\mathrm{D}^2LS(f; u_h, p_h; \bullet, \bullet)$ of the least-squares functional is computed, where $(u_h, q_h) \in X_h$ and $\lambda \in \mathbb{R}$ satisfy, for all $(\widetilde{v}_h, \widetilde{q}_h) \in X_h$,

$$\mathrm{D}^2LS(f; u_h, p_h; v_h, q_h, \widetilde{v}_h, \widetilde{q}_h) = \lambda\big(a_{\mathrm{NC}}(v_h, \widetilde{v}_h) + (q_h, \widetilde{q}_h)_{H(\mathrm{div},\Omega)}\big). \tag{5.1}$$

Table 1 Convergence history of Newton iteration for 4 representative examples

n_{iter}	(A)	(B)	(C)	(D)
0	1.67431×10^{1}	8.73230×10^{0}	6.431245×10^{0}	1.16987×10^{-1}
1	2.20124×10^{0}	7.69274×10^{-2}	4.194389×10^{-2}	1.94092×10^{-3}
2	1.59872×10^{-1}	2.13909×10^{-4}	8.04263×10^{-5}	3.30072×10^{-6}
3	9.61529×10^{-4}	1.74101×10^{-9}	3.74273×10^{-10}	1.17441×10^{-11}
4	4.11730×10^{-8}	1.12689×10^{-14}	6.111556×10^{-15}	6.26485×10^{-15}
5	1.13667×10^{-14}	1.09142×10^{-14}	5.70819×10^{-15}	5.93587×10^{-15}
6	1.11131×10^{-14}	1.10560×10^{-14}	5.92760×10^{-15}	5.82990×10^{-15}
7	1.09108×10^{-14}			6.03978×10^{-15}
8	1.14493×10^{-14}			

The value $\lambda_{\min}$ is uniformly bounded from zero for the examples in the following subsections, so that every computed discrete solution (u_h, p_h) is a local minimizer.

For any discrete approximation (u_h, p_h), Theorems 3.11 and 4.7 verify the a posteriori error estimator $\eta^2(\mathcal{T}) := \mathrm{LS}(f; u_h, p_h) + \|h_{\mathcal{T}} f\|^2_{L^2(\Omega)}$ even for *inexact solve* in its computation. In view of a lacking proof in Sect. 5.4 below that the computed discrete solution is in fact a *global discrete minimizer* (at least up to machine precision), it is only by this universal a posteriori error control that we know that the computed approximations converge to the exact solution.

5.2 Numerical example on square domain

This subsection considers the nonlinear model problem for the exact solution

$$u(x) := \cos(\pi x_1/2)\cos(\pi x_2/2) \quad \text{for } x \in \Omega := (-1, 1)^2$$

with homogeneous Dirichlet boundary condition, $f := -\operatorname{div}(\sigma(\nabla u))$, and ϕ from Example 3.2.a. This defines the exact stress function $p := \sigma(\nabla u) \in H(\operatorname{div}, \Omega)$.

Figure 1 displays the error estimator $\eta_\ell := \eta(\mathcal{T}_\ell)$ at the discrete solutions (u_ℓ, p_ℓ) on each level ℓ of a sequence of uniform triangulations as well as the error to the exact solution (u, p). The reference energy $E(u) = -5.774337908509$ in the energy difference $E(u_\ell) - E(u) \geq \gamma_1 |||u - u_\ell|||^2/2$ has been approximated by the energies of P_1-conforming finite element solution with an Aitken Δ^2 extrapolation. The eigenvalues of (5.1) in all experiments of Fig. 1 satisfy $1.597 \leq \lambda_{\min} \leq 1.722$ and $9.943 \leq \lambda_{\max} \leq 16.128$ and so prove that the discrete solutions are local minimizers. The parallel graphs confirm the equivalence of the built-in error estimator $\|y_\ell\|_Y = (LS(f; u_\ell, p_\ell))^{1/2}$ with the exact error from Theorem 4.1.

With the Friedrichs constant $C_{\mathrm{F}} = \sqrt{2}/\pi$ of the square domain, the criterion (4.2) is equivalent to $\|\nabla_{\mathrm{NC}} v_h\|_{L^\infty(\Omega)} < \gamma_1^2 \mathrm{Lip}(\mathrm{D}\sigma)^{-1}(1 + C_{\mathrm{F}}^2)^{-1} = 0.17239892$ and so Fig. 1 shows that the criterion (4.2) holds for each level $\ell \geq 6$ and Theorem 4.4

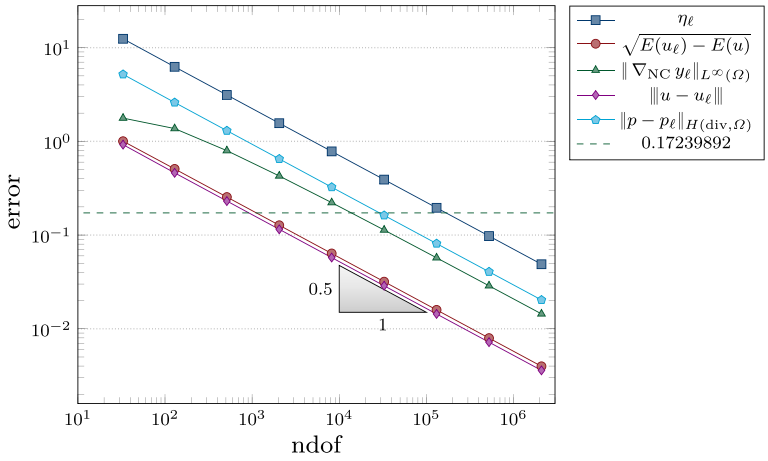

Fig. 1 Convergence history for a sequence of uniform triangulations of the square domain with exact solution u from Sect. 5.2

implies global uniqueness of the computed (u_ℓ, p_ℓ). This proves that there exists only one local minimizer in the discrete problem (dPG).

5.3 Adaptive mesh-refinement

The natural adaptive algorithm with collective Dörfler marking [20] utilizes the local error estimator $\eta^2(\mathcal{T}, T) := \|(I_{n\times n} + S_0)^{-1/2}\big(\Pi_0 p_{\mathrm{RT}} - \sigma(\nabla u_C) + H_0 f\big)\|^2_{L^2(T)} + \|\Pi_0 f + \operatorname{div} p_{\mathrm{RT}}\|^2_{L^2(T)} + \|h_T f\|^2_{L^2(T)}$ for any $(u_C, p_{\mathrm{RT}}) \in S^1_0(\mathcal{T}) \times RT_0(\mathcal{T})$ and $T \in \mathcal{T}$ as follows.

Input: Regular triangulation $\mathcal{T}_0$ of the polygonal domain Ωinto simplices.
for any level $\ell = 0, 1, 2, \ldots$ **do**
Solve generalized LS-FEM with respect to triangulation $\mathcal{T}_\ell$ and solution (u_ℓ, p_ℓ).
Compute error estimator $\eta_\ell := \eta(\mathcal{T}_\ell)$.
Mark a subset $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell$ of (almost) minimal cardinality $|\mathcal{M}_\ell|$ with

$$0.3\,\eta_\ell^2 \le \eta_\ell^2(\mathcal{M}_\ell) := \sum_{T\in\mathcal{M}_\ell} \eta^2(\mathcal{T}_\ell, T)$$

Compute smallest regular refinement $\mathcal{T}_{\ell+1}$ of $\mathcal{T}_\ell$ with $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell \setminus \mathcal{T}_{\ell+1}$ by newest-vertex bisection (NVB). **od**
Output: Sequence of discrete solutions $(u_\ell, p_\ell)_{\ell\in\mathbb{N}_0}$ and triangulations $(\mathcal{T}_\ell)_{\ell\in\mathbb{N}_0}$.

See [26] for details on adaptive mesh-refinement and NVB.

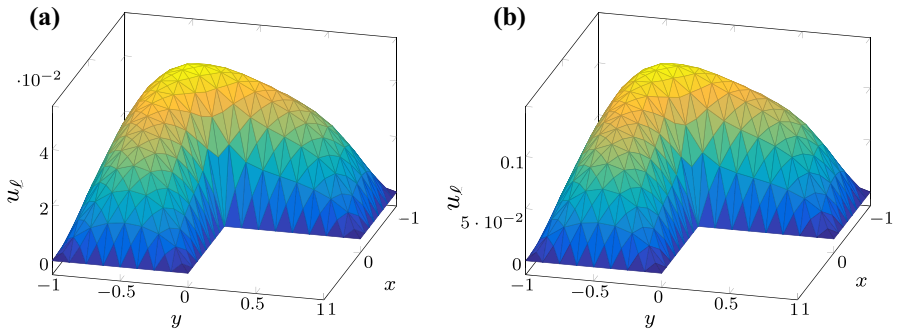

Fig. 2 Solution u_h for different functions ϕ on a uniform triangulation of the L-shaped domain into 768 (ndof = 1 537). **a** ϕ from Example 3.2.a. **b** ϕ from Example 3.2.b

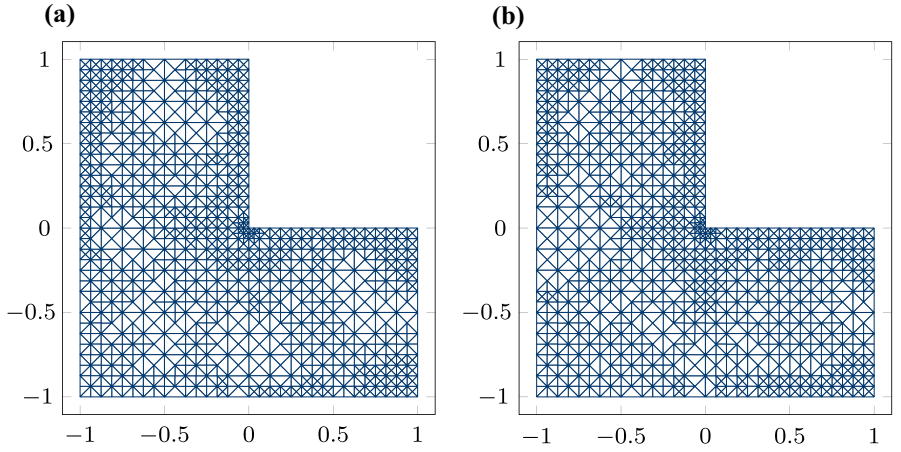

Fig. 3 Adaptively refined triangulation $\mathcal{T}_\ell$ for different functions ϕ. **a** ϕ from Example 3.2.a, mesh with 1 783 triangles (ndof = 3 567). **b** ϕ from Example 3.2.b, mesh with 1 838 triangles (ndof = 3 677)

5.4 Numerical example on L-shaped domain

This subsection considers $f \equiv 1$ on the L-shaped domain $\Omega := (-1, 1)^2 \setminus [0, 1]^2 \subset \mathbb{R}^2$ with homogeneous Dirichlet boundary condition $u|_{\partial\Omega} \equiv 0$ and unknown exact solution u. Figure 2 displays the corresponding discrete solutions u_h on a uniform triangulation of Ω for the different functions ϕ from Example 3.2.a and b.

Figure 3 shows two typical adaptively generated triangulations with considerable refinement at the re-entrant corner for different functions ϕ. At first glance, the meshes appear similar and resemble the undisplayed adaptive triangulation from the Poisson model problem.

For ϕ from Example 3.2.a, Fig. 4 shows the convergence history plot of the natural least-squares error estimator $\eta_\ell = \eta(\mathcal{T}_\ell)$ and the difference of the energy $E(u_\ell)$ of the solution u_ℓ and a reference energy $E(u) = -3.657423002939 \times 10^{-2}$ (com-

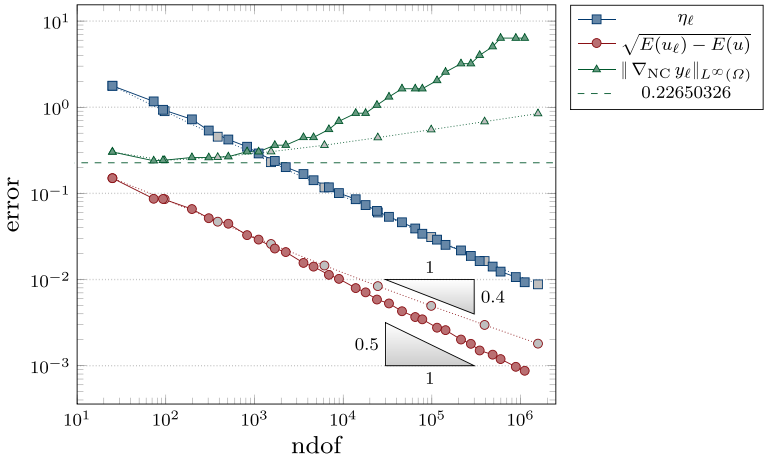

Fig. 4 Convergence history for adaptive mesh-refinement (solid lines) and uniform mesh-refinement (dotted lines) with ϕ from Example 3.2.a

puted by the energies of P_1-conforming finite element solutions with an Aitken Δ^2 extrapolation).

The eigenvalues of (5.1) in all experiments satisfy $1.787 \leq \lambda_{\min} \leq 1.914$ and $16.682 \leq \lambda_{\max} \leq 17.932$ and so prove that all the discrete solutions are local minimizers. The function ϕ from Example 3.2.b leads to (undisplayed) similar results.

For the L-shaped domain, the smallest eigenvalue $\lambda_1 = 9.6397238$ of the Laplacian with homogeneous Dirichlet boundary conditions yields the Friedrichs constant $C_{\mathrm{F}} = 1/\sqrt{\lambda_1} = 0.32208293$. Since $\| \nabla_{\mathrm{NC}} v_\ell \|_{L^\infty(\Omega)} \geq \gamma_1^2 \mathrm{Lip}(\mathrm{D}\sigma)^{-1}(1 + C_{\mathrm{F}}^2)^{-1} = 0.22650326$ for all level $\ell \in \mathbb{N}_0$ in Fig. 4, Theorem 4.4 is *not* applicable to any triangulation $\mathcal{T}_\ell$ of the computation at hand.

To guarantee optimal convergence rates for least-squares FEMs with an alternative a posteriori error estimator, the choice of a sufficiently small bulk parameter is crucial [11,14]. However, for the natural error estimator with the values of the least-squares functional, the plain convergence proof of [12] requires the bulk parameter sufficiently close to 1. For the nonlinear model problem at hand, the convergence history plot in Fig. 4 provides numerical evidence for optimal convergence rates for adaptive mesh-refinement of Sect. 5.3 and suboptimal convergence for uniform refinement.

References

1. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications, Springer Series in Computational Mathematics, vol. 44. Springer, Heidelberg (2013)
2. Bouma, T., Gopalakrishnan, J., Harb, A.: Convergence rates of the DPG method with reduced test space degree. Comput. Math. Appl. **68**(11), 1550–1561 (2014)
3. Braess, D.: Finite Elements, 3rd edn. Cambridge University Press, Cambridge (2007)
4. Brenner, S.C., Scott, L.R.: The mathematical Theory of Finite Element Methods, Texts in Applied Mathematics, 3rd edn. Springer, New York (2008)
5. Carstensen, C., Demkowicz, L., Gopalakrishnan, J.: A posteriori error control for DPG methods. SIAM J. Numer. Anal. **52**(3), 1335–1353 (2014)

6. Carstensen, C., Demkowicz, L., Gopalakrishnan, J.: Breaking spaces and forms for the DPG method and applications including Maxwell equations. Comput. Math. Appl. **72**(3), 494–522 (2016)
7. Carstensen, C., Gallistl, D.: Guaranteed lower eigenvalue bounds for the biharmonic equation. Numer. Math. **126**(1), 33–51 (2014)
8. Carstensen, C., Gallistl, D., Hellwig, F., Weggler, L.: Low-order dPG-FEM for an elliptic PDE. Comput. Math. Appl. **68**(11), 1503–1512 (2014)
9. Carstensen, C., Hellwig, F.: Low-order discontinuous Petrov–Galerkin finite element methods for linear elasticity. SIAM J. Numer. Anal. **54**(6), 3388–3410 (2016)
10. Carstensen, C, Hellwig, F.: Some constants in discrete Poincaré and Friedrichs inequalities and application to discrete quasiinterpolation. CMAM, 1–27 (2018). arXiv:1709.00577. https://www.degruyter.com/view/j/cmam.ahead-of-print/cmam-2017-0044/cmam-2017-0044.xml?format=INT
11. Carstensen, C., Park, E.: Convergence and optimality of adaptive least squares finite element methods. SIAM J. Numer. Anal. **53**(1), 43–62 (2015)
12. Carstensen, C., Park, E., Bringmann, P.: Convergence of natural adaptive least squares finite element methods. Numer. Math. **136**(4), 1097–1115 (2017)
13. Carstensen, C., Puttkammer, S.: A low-order discontinuous Petrov-Galerkin method for the Stokes equations, submitted (2016)
14. Carstensen, C., Rabus, H.: Axioms of adaptivity for separate marking. SIAM J. Numer. Anal. **55**(6), 2644–2665 (2017). https://doi.org/10.1137/16M1068050
15. Carstensen, C., Stephan, E.P.: Adaptive coupling of boundary elements and finite elements. RAIRO Modél. Math. Anal. Numér. **29**(7), 779–817 (1995)
16. Chan, J., Demkowicz, L., Moser, R.: A DPG method for steady viscous compressible flow. Comput. Fluids **98**, 69–90 (2014)
17. Cohen, A., Dahmen, W., Welper, G.: Adaptivity and variational stabilization for convection–diffusion equations. ESAIM Math. Model. Numer. Anal. **46**(5), 1247–1273 (2012)
18. Demkowicz, L., Gopalakrishnan, J.: Analysis of the DPG method for the Poisson equation. SIAM J. Numer. Anal. **49**(5), 1788–1809 (2011)
19. Demkowicz, L., Gopalakrishnan, J.: A primal DPG method without a first-order reformulation. Comput. Math. Appl. **66**(6), 1058–1064 (2013)
20. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**(3), 1106–1124 (1996)
21. Führer, T., Heuer, N., Stephan, E.P.: On the DPG method for Signorini problems. IMA J. Numer. Anal. (2018). https://doi.org/10.1093/imanum/drx048
22. Muga, I., van der Zee, K.G.: Discretization of linear problems in Banach spaces: Residual minimization, nonlinear Petrov-Galerkin, and monotone mixed methods. arXiv:1511.04400, 1–29, submitted (2015)
23. Müller, B., Starke, G., Schwarz, A., Schröder, J.: A first-order system least squares method for hyperelasticity. SIAM J. Sci. Comput. **36**(5), B795–B816 (2014)
24. Nečas, J.: Introduction to the Theory of Nonlinear Elliptic Equations, vol. 52. BSB B. G. Teubner Verlagsgesellschaft, Leipzig (1983)
25. Pousin, J., Rappaz, J.: Consistency, stability, a priori and a posteriori errors for Petrov–Galerkin methods applied to nonlinear problems. Numer. Math. **69**(2), 213–231 (1994)
26. Stevenson, R.: The completion of locally refined simplicial partitions created by bisection. Math. Comp. **77**, 227–241 (2008)
27. Zeidler, E.: Nonlinear Functional Analysis and Its Applications, vol. I. Springer-Verlag, New York (1986)
28. Zeidler, E.: Nonlinear Functional Analysis and Its Applications, vol. IV. Springer-Verlag, New York (1988)
29. Zeidler, E.: Nonlinear Functional Analysis and Its Applications, vol. II/B. Springer-Verlag, New York (1990)