ANALYSIS OF MIXED DISCONTINUOUS GALERKIN FORMULATIONS FOR QUASILINEAR ELLIPTIC PROBLEMS

MOHAMMAD ZAKERZADEH[†] AND GEORG MAY[†*]

Abstract. In this manuscript we present an approach to analyze the discontinuous Galerkin solution for general quasilinear elliptic problems. This approach is sufficiently general to extend most of the well-known discretization schemes, including BR1, BR2, SIPG and LDG, to nonlinear cases in a canonical way, and to establish the stability of their solution. Furthermore, in case of monotone and globally Lipschitz problems, we prove the existence and uniqueness of the approximated solution and the h-optimality of the error estimate in the energy norm as well as in the L_2 norm.

Key words. discontinuous Galerkin methods, nonlinear elliptic problems, optimal estimates

AMS subject classifications. 65N12, 65N15, 65N30

1. Introduction. A great amount of research has been devoted to the analysis of discontinuous Galerkin (DG) schemes for (non)linear elliptic and parabolic problems. This goes back to work of Babuska [4], Wheeler [41], Arnold [2] and Dupont et al. [29] for interior penalty (IP) methods. In later work, e.g. [9, 37], a non-symmetric IP scheme was presented and analyzed. Other classes of discretization techniques have also been introduced, e.g., the first and second method of Bassi–Rebay [7, 8], or Shu and Cockburn's LDG method [16]. For a comprehensive literature study of these problems we refer to [15] and [3].

In their seminal paper [3], Arnold et al. provided a unified framework for obtaining different classes of DG methods for the linear Poisson problem. For nonlinear problems the picture is more complicated, and obtaining different classes of methods usually follows very different approaches; while IP methods often have been introduced and analyzed in the primal form [17, 23, 24, 34, 28], for LDG methods the usual approach is to start from a mixed formulation, e.g., [11, 22, 12] or [42]. Originating from this, different techniques are usually employed in the analysis of each family. The LDG methods [22, 11, 12, 42] have been formulated by using three unknowns in the mixed form as proposed in [14], which yields an inconsistent primal formulation as analyzed, e.g., in [11] (and [35] for the simpler linear case). In contrast, IP methods like [24, 23, 28, 34, 17] usually enjoy a consistent primal form.

On the other hand, the nonlinear versions of the Bassi-Rebay methods are often obtained by ad-hoc extension, mimicking the linear counter part, e.g. [6], and to the best knowledge of the authors, there does not exist rigorous analysis of the second Bassi–Rebay method for nonlinear problems, although they are widely used for nonlinear problems like the Navier–Stokes equations.

This manuscript aims to extend the discussion presented in [11, 24, 22, 23, 28, 34] in two ways: Firstly, we try to treat different types of discretization in a canonical way, starting from a mixed formulation as [3]. We will show that our formulation leads to a slightly modified version of previously proposed nonlinear formulations for symmetric IP (SIPG) [23], and Bassi–Rebay [6], while we recover the LDG formulation of [11]. Here we are interested in these methods since, among the methods investigated in [3], these are the only ones that are stable and consistent for primal and

adjoint solutions, and we might hope for extending these properties to the corresponding nonlinear discretization. Also we prove that the approximate solution of the discontinuous Galerkin problem is well-posed; i.e, unique and stable, provided that the diffusion operator is strongly monotone and globally Lipschitz continuous. This is comparable to analysis of [11] for LDG and somewhat similar to [17] for IP; however our results cover different formulations. It is plausible that the same kind of analysis is applicable to cases which do not satisfy either strong monotonicity or global Lipschitz continuity as in [34, 24, 23]; nevertheless, we do not address this extension here and restrict our analysis to these two assumptions.

Secondly, we are going to provide an optimal error estimate for SIPG and Bassi–Rebay methods in terms of mesh size, both in energy norm and in the L^2-norm. Such estimates have previously been derived for LDG methods, in [11] for monotone and globally Lipschtitz continuous problem, and in [22] for cases which neither of these two assumptions hold. For the SIPG method, error estimates have been presented in [23, 24] and for incomplete penalty (IIPG) in [34], for non-monotone and not globally Lipschitz operators. Also we have the results of [17] for IIPG in case of monotone and globally Lipschitz operators. Let us remark that since the formulations presented in [34, 17] are adjoint inconsistent, they derived the error estimate only in the energy norm and did not deal with the corresponding adjoint problem.

Besides the canonical approach of discretization we present here, the rigorous analysis of the second method of Bassi–Rebay is novel in the literature, as well as the different approach we adopted in the L_2 error estimate for an asymptotically adjoint consistent formulation. Furthermore, we present explicit conditions of stability for all formulations. To the best knowledge of the authors this has not been done before for a nonlinear version of Bassi–Rebay methods and is similar to the explicit bounds for the SIPG method in [39, 18] and [31]. It is worth mentioning that for the SIPG method, unlike the cited literature, our stability analysis remains valid for degenerate diffusion.

The structure of this paper is as follows: In §2 we provide a short review on the quasi-linear elliptic problems and the properties of the diffusion operator. In §3 we review some approximation results as well as the properties of triangulation of the computational domain. Section 4 deals with our canonical approach for obtaining different DG formulation in the primal form, and later in §5, we analyze the consistency and adjoint consistency of the presented formulations. The sections 6 and 7 are devoted to the stability and uniqueness of the DG approximate solution, respectively. Note that before reaching section 7 we do not require the monotonicity of the operator and the stability result is valid for more general problems. Section 8 presents the optimal error convergence estimate in both energy and L_2 norms.

2. Quasi-linear elliptic problems. We consider the following quasilinear elliptic problem

$$(2.1) \qquad\qquad -\nabla \cdot \boldsymbol{a}(x, u, \nabla u) = f, \qquad\qquad \text{in } \Omega,$$

$$(2.2) \qquad\qquad\qquad\qquad u = u_D, \qquad\qquad \text{on } \partial\Omega,$$

where $f \in L_2(\Omega)$ and $u_D \in H^{1/2}(\partial\Omega)$. For simplicity we will set $u_D \equiv 0$ in later analysis. Also Ω is a bounded and simply connected domain in $\mathbb{R}^2$. For brevity one might use the notation $\boldsymbol{a}(\cdot, \boldsymbol{\zeta}) = \boldsymbol{a}(\cdot, u, \nabla u)$ where $\boldsymbol{\zeta} \in \mathbb{R}^3$, such that $\zeta_0 \equiv u$ and $\zeta_i \equiv u_{x_i}, i = 1, 2$. Moreover, by $\boldsymbol{a}_u$ and $\boldsymbol{a_z} := \left[\frac{\partial a_i(\cdot, \boldsymbol{\zeta})}{\partial \zeta_j}\right]_{i,j=1,2}$, we denote the derivative of $\boldsymbol{a}(x, u, \nabla u)$ with respect to its second and third arguments.

In the general theory of nonlinear elliptic problems (see [32, 45]), it is usually assumed that the function $\boldsymbol{a}(\cdot, \boldsymbol{\zeta}) = (a_1(\cdot, \boldsymbol{\zeta}), a_2(\cdot, \boldsymbol{\zeta}))$ satisfies some conditions:

A(i) The functions $a_i(x, \boldsymbol{\zeta}), i = 1, 2$ are continuous in $\Omega \times \mathbb{R}^3$ and satisfy the following growth condition;

$$(2.3) \qquad |a_i(x, \boldsymbol{\zeta})| \leq c_1 \Big(1 + \sum_{i=0}^{2} |\zeta_i|\Big) + |\phi_i(x)|, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^3, \forall x \in \Omega,$$

where $c_1 > 0$ and $\phi_i \in L_2(\Omega)$, for $i = 1, 2$.

A(ii) There exist two constants $0 < \lambda \leq \Lambda < \infty$ such that for all $\boldsymbol{\zeta}, \boldsymbol{\psi} \in \mathbb{R}^3$

$$(2.4) \qquad 0 < \lambda \sum_{k=1}^{2} \psi_k^2 \leq \sum_{\substack{j=1 \\ k=0}}^{2} \frac{\partial a_j(\cdot, \boldsymbol{\zeta})}{\partial \zeta_k} \psi_j \psi_k \leq \Lambda \sum_{k=1}^{2} \psi_k^2.$$

We also might consider a relaxed version of (2.4) as the following

$$(2.5) \qquad 0 < \lambda \sum_{k=1}^{2} \psi_k^2 \leq \sum_{\substack{j=1 \\ k=1}}^{2} \frac{\partial a_j(\cdot, \boldsymbol{\zeta})}{\partial \zeta_k} \psi_j \psi_k \leq \Lambda \sum_{k=1}^{2} \psi_k^2.$$

For simplicity in the later analysis we assume that $\boldsymbol{a_z}$ is symmetric. Then we can interpret (2.5) as imposition of lower and upper bound on the eigenvalues of $\boldsymbol{a_z}$; λ and Λ, respectively.

A(iii) The functions $a_i(x, \boldsymbol{\zeta}), i = 1, 2$ are in $\mathcal{C}_b^2(\Omega \times \mathbb{R}^3)$, i.e., they are twice continuously differentiable functions with all the derivatives through second order being bounded.

Note that A(i) is required to guarantee the meaningfulness of the of the definition of the discrete problem. On the other hand the assumptions A(ii) and A(iii) need to be valid only for the arguments provided in sections 7 and 8 for uniqueness of the discrete solution and error estimate, respectively. Furthermore, the relaxed version of assumption A(ii) in (2.5) is required in the stability analysis in section 6.

Let us note the following important definition for nonlinear elliptic PDEs, the so-called *strong monotonicity* and *global Lipschitz continuity* property of the diffusion operator (as [19], [44, p. 476]):

DEFINITION 2.1. *We say a diffusion operator $\boldsymbol{a}(x, u, \nabla u)$ satisfies strong monotonicity property, if there exists a constant $C_{sm} > 0$ such that*

$$(2.6) \qquad \big(\boldsymbol{a}(x, \xi', \boldsymbol{\eta}') - \boldsymbol{a}(x, \xi, \boldsymbol{\eta})\big) \cdot (\boldsymbol{\eta}' - \boldsymbol{\eta}) \geq C_{sm} |\boldsymbol{\eta} - \boldsymbol{\eta}'|^2$$

for all $\xi, \xi' \in \mathbb{R}$ and $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathbb{R}^2$. Moreover we call it a globally Lipschitz operator if there exists a constant $C_{lc} < \infty$ such that

$$(2.7) \qquad \boldsymbol{a}(x, \xi', \boldsymbol{\eta}') - \boldsymbol{a}(x, \xi, \boldsymbol{\eta}) \leq C_{lc} \big[|\xi - \xi'|^2 + |\boldsymbol{\eta} - \boldsymbol{\eta}'|^2\big]^{1/2},$$

for all $\xi, \xi' \in \mathbb{R}$ and $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathbb{R}^2$.

One can show that if the left and right inequalities in (2.4) are satisfied, the diffusion operator $\boldsymbol{a}$ is strongly monotone and globally Lipschitz continuous in the sense of Definition 2.1. We skip the proof here and refer to [5, Lemma 2.1] or [17, Lemma 4].

As examples of problems settled in the category of (2.1) we have

(a) Newtonian flow model

$$(2.8) \qquad \boldsymbol{a}(x, u, \nabla u) = \boldsymbol{a}(x, u) \nabla u$$

(b) Non-Newtonian flow model

$$\boldsymbol{a}(x, u, \nabla u) = \boldsymbol{a}(x, |\nabla u|)\nabla u \tag{2.9}$$

with an specific case of mean curvature flow

$$\boldsymbol{a}(x, u, \nabla u) = \frac{\nabla u}{(1 + |\nabla u|)^{1/2}} \tag{2.10}$$

We remark that for the case (b), following [11, 28], we typically assume the monotonicity of the diffusion operator (see Definition 2.1), while the diffusion operator in case (a) is not usually a monotone operator (only under very restrictive assumptions, see [1]). Since the presentation of the primal form of DG formulations in section 4 and stability result in section 6 are not affected by the monotonicity property, we do not exclude this case now. But in obtaining a priori error estimate in section 8, we will only consider monotone cases and one can refer to [34], [24] or [23] for a priori estimate for non-monotone cases. One may check that the mean curvature flow example satisfies the assumption A(ii); hence it is strongly monotone and globally Lipschitz continuous, see [23, section 5].

3. Preliminaries. First let us set a notation convention and suppress the dependence of the diffusion operator on the space variable x; henceforth we write $\boldsymbol{a}(v, \boldsymbol{z})$ instead of $\boldsymbol{a}(x, v, \boldsymbol{z})$, while we still allow explicit dependence on x.

Now, as a tool that we will use later in sections 6 and 8, let us consider the following integral form of Taylor's formula; for $u, v \in \mathbb{R}$ and $\boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^2$ one has

$$\boldsymbol{a}(v, \boldsymbol{z}) - \boldsymbol{a}(u, \boldsymbol{w}) = \boldsymbol{a}_u(u, \boldsymbol{w})(v - u) + \boldsymbol{a}_{\boldsymbol{z}}(u, \boldsymbol{w})(\boldsymbol{z} - \boldsymbol{w}) + R_{\boldsymbol{a}}(u - v, \boldsymbol{w} - \boldsymbol{z})$$
$$\tag{3.1} = \tilde{\boldsymbol{a}}_u(u, \boldsymbol{w})(v - u) + \tilde{\boldsymbol{a}}_{\boldsymbol{z}}(u, \boldsymbol{w})(\boldsymbol{z} - \boldsymbol{w})$$

where $R_{\boldsymbol{a}}(u - v, \boldsymbol{w} - \boldsymbol{z}) = (R_{\boldsymbol{a}_1}(u - v, \boldsymbol{w} - \boldsymbol{z}), R_{\boldsymbol{a}_2}(u - v, \boldsymbol{w} - \boldsymbol{z}))$ is defined as

$$R_{\boldsymbol{a}_i} = \tilde{\boldsymbol{a}}_{uu}(u - v)^2 + (\boldsymbol{w} - \boldsymbol{z})^t \tilde{\boldsymbol{a}}_{\boldsymbol{z}\boldsymbol{z}}(u, \boldsymbol{w})(\boldsymbol{w} - \boldsymbol{z}) + 2\tilde{\boldsymbol{a}}_{u\boldsymbol{z}}(u, \boldsymbol{w})(\boldsymbol{z} - \boldsymbol{w})(u - v). \tag{3.2}$$

for $i = 1, 2$. Moreover we define $\tilde{\boldsymbol{a}}_u, \tilde{\boldsymbol{a}}_{\boldsymbol{z}}, \tilde{\boldsymbol{a}}_{uu}, \tilde{\boldsymbol{a}}_{u\boldsymbol{z}}, \tilde{\boldsymbol{a}}_{\boldsymbol{z}\boldsymbol{z}}$ as

$$\tilde{\boldsymbol{a}}_u(u, \boldsymbol{w}) = \int_0^1 \boldsymbol{a}_u(v(t), \boldsymbol{z}(t)) \, dt, \qquad \tilde{\boldsymbol{a}}_{\boldsymbol{z}}(u, \boldsymbol{w}) = \int_0^1 \boldsymbol{a}_{\boldsymbol{z}}(v(t), \boldsymbol{z}(t)) \, dt$$

$$\tilde{\boldsymbol{a}}_{uu}(u, \boldsymbol{w}) = \int_0^1 (1 - t)\boldsymbol{a}_u(v(t), \boldsymbol{z}(t)) \, dt, \qquad \tilde{\boldsymbol{a}}_{u\boldsymbol{z}}(u, \boldsymbol{w}) = \int_0^1 (1 - t)\boldsymbol{a}_u(v(t), \boldsymbol{z}(t)) \, dt,$$

$$\tilde{\boldsymbol{a}}_{\boldsymbol{z}\boldsymbol{z}}(u, \boldsymbol{w}) = \int_0^1 (1 - t)\boldsymbol{a}_u(v(t), \boldsymbol{z}(t)) \, dt$$

where $v(t) = u + t(v - u)$ and $\boldsymbol{z}(t) = \boldsymbol{w} + t(\boldsymbol{z} - \boldsymbol{w})$.

Seeking clarity, sometimes we denote $\tilde{\boldsymbol{a}}_u, \tilde{\boldsymbol{a}}_{\boldsymbol{z}}, \tilde{\boldsymbol{a}}_{uu}, \tilde{\boldsymbol{a}}_{u\boldsymbol{z}}, \tilde{\boldsymbol{a}}_{\boldsymbol{z}\boldsymbol{z}}$ by all four arguments, e.g. $\tilde{\boldsymbol{a}}_u(u, \boldsymbol{w}, v, \boldsymbol{z})$ instead of $\tilde{\boldsymbol{a}}_u(u, \boldsymbol{w})$.

3.1. Triangulation and finite element space. Here we are going to consider the boundary of the domain Ω sufficiently smooth (in order to apply the duality argument, see section 5.2), e.g. a convex polygon. Then we consider a shape-regular triangulation on Ω as $\mathcal{T}_h = \{\kappa\}$ composed of (non-overlapping) triangular or rectangular elements (with possible hanging nodes) and h_κ is the

diameter of each $\kappa \in \mathcal{T}_h$. Also we define $h := \max_{\kappa \in \mathcal{T}_h} h_\kappa$ and ν_κ is the outward normal to $\partial\kappa$. In the following we assume that $\mathcal{T}_h$ is of *bounded variation*, that is, there exists a constant $l > 1$ such that

$$(3.3) \qquad l^{-1} \leq \frac{h_\kappa}{h_{\kappa'}} < l,$$

where $\kappa, \kappa' \in \mathcal{T}_h$ share an edge. This bounded variation property means that there is an upper bound for the number of neighboring elements of each $\kappa \in \mathcal{T}_h$, denoted by N_l. In case that $\mathcal{T}_h$ has no hanging nodes $N_l = 3$ and $N_l = 4$ for triangular and rectangular elements, respectively. We also need the following *quasi-uniformity* property of the mesh in the L_2 error analysis in section 8.2, that is

$$(3.4) \qquad h \leq C h_\kappa, \qquad \forall \kappa \in \mathcal{T}_h.$$

We denote the *skeleton* of the triangulation $\mathcal{T}_h$, i.e. the set of all edges of $\kappa \in \mathcal{T}_h$, by $\mathcal{E}_h$. Also we denote the set of boundary and interior edges of $\mathcal{T}_h$ by $\mathcal{E}_{h,\partial}$ and $\mathcal{E}_{h,I}$, respectively, and the length of edge e by h_e. Following [3], let us fix some definitions for the jumps and average of the discontinuous functions on the skeleton $\mathcal{E}_h$. Let us set the trace values as $w_{\kappa,e} = w_\kappa|_e$. For any interior edge $e \in \mathcal{E}_{h,I}$, where e is the common edge of $\kappa, \kappa' \in \mathcal{T}_h$, and for all $w \in \prod_{\kappa \in \mathcal{T}_h} L_2(\partial\kappa)$, we define

$$(3.5) \qquad \{\!\{w\}\!\} = \frac{1}{2}(w_{\kappa,e} + w_{\kappa',e}), \qquad [\![w]\!] = w_{\kappa,e}\nu_\kappa + w_{\kappa',e}\nu_{\kappa'}$$

and similarly for all $\boldsymbol{\tau} \in \prod_{\kappa \in \mathcal{T}_h}[L_2(\partial\kappa)]^2$

$$(3.6) \qquad \{\!\{\boldsymbol{\tau}\}\!\} = \frac{1}{2}(\boldsymbol{\tau}_{\kappa,e} + \boldsymbol{\tau}_{\kappa',e}), \qquad [\![\boldsymbol{\tau}]\!] = \boldsymbol{\tau}_{\kappa,e} \cdot \nu_\kappa + \boldsymbol{\tau}_{\kappa',e} \cdot \nu_{\kappa'}.$$

For any boundary edge $e \in \mathcal{E}_{h,\partial}$ we define

$$(3.7) \qquad [\![w]\!] = w_{\kappa,e}\nu_\kappa, \qquad \{\!\{\boldsymbol{\tau}\}\!\} = \boldsymbol{\tau}_{\kappa,e},$$

for all $w \in \prod_{\kappa \in \mathcal{T}_h} L_2(\partial\kappa)$ and $\boldsymbol{\tau} \in \prod_{\kappa \in \mathcal{T}_h}[L_2(\partial\kappa)]^2$.

Moreover, let us consider the following broken Sobolev space on the triangulation $\mathcal{T}_h$; for $1 \leq r < \infty$

$$(3.8) \qquad W_r^s(\Omega, \mathcal{T}_h) = \{v \in L_r : v|_\kappa \in W_r^s, \forall \kappa \in \mathcal{T}_h\},$$

with the corresponding norm and seminorm

$$(3.9) \qquad \|v\|_{W_r^s(\Omega, \mathcal{T}_h)} = \Big(\sum_{\kappa \in \mathcal{T}_h} \|v\|_{W_r^s(\kappa)}^r \Big)^{1/r}, \qquad |v|_{W_r^s(\Omega, \mathcal{T}_h)} = \Big(\sum_{\kappa \in \mathcal{T}_h} |v|_{W_r^s(\kappa)}^r \Big)^{1/r},$$

and for the case $r = \infty$ the associated norm and seminorm are defined as

$$(3.10) \qquad \|v\|_{W_\infty^s(\Omega, \mathcal{T}_h)} = \max_{\kappa \in \mathcal{T}_h} \|v\|_{W_\infty^s(\kappa)}, \qquad |v|_{W_\infty^s(\Omega, \mathcal{T}_h)} = \max_{\kappa \in \mathcal{T}_h} |v|_{W_\infty^s(\kappa)},$$

when $\|\cdot\|_{W_r^s(\kappa)}$ and $|\cdot|_{W_r^s(\kappa)}$ are the standard Sobolev norms on κ. Also we denote W_2^s as H^s by tradition.

Now let define the following finite dimensional approximation spaces

(3.11) $$V_{h,q} := \{u_h \in L_2(\Omega) : u_h|_\kappa \in \mathcal{P}^q(\kappa), \quad \forall \kappa \in \mathcal{T}_h\},$$

(3.12) $$\Sigma_{h,p} := \{\boldsymbol{\theta}_h \in [L_2(\Omega)]^2 : \boldsymbol{\theta}_h|_\kappa \in [\mathcal{P}^p(\kappa)]^2, \quad \forall \kappa \in \mathcal{T}_h\},$$

with $q \geq 1$ and $p = q$ or $p = q - 1$ (To satisfy the inclusion property $\nabla V_{h,q} \subset \Sigma_{h,p}$ as [3].) Here by $\mathcal{P}^q(\kappa)$ we denote the space of the polynomials of total degree q on $\mathbb{R}^2$ and restricted to κ.

Moreover, let us define two lifting operator $r : [L_2(\mathcal{E}_h)]^2 \to \Sigma_{h,p}$ and $l : L_2(\mathcal{E}_{h,I}) \to \Sigma_{h,p}$ as

(3.13) $$\int_\Omega r(\varphi) \cdot \boldsymbol{\tau} \, \mathrm{d}x = -\sum_{e \in \mathcal{E}_h} \int_e \varphi \cdot \{\!\{\boldsymbol{\tau}\}\!\} \, \mathrm{d}s, \qquad \int_\Omega l(\varphi) \cdot \boldsymbol{\tau} \, \mathrm{d}x = -\sum_{e \in \mathcal{E}_{h,I}} \int_e \varphi [\![\boldsymbol{\tau}]\!] \, \mathrm{d}s,$$

for all $\boldsymbol{\tau} \in \Sigma_{h,p}$. Using the Riesz representation theorem one can prove the existence and uniqueness of the lifting operators introduced by (3.13) (see e.g., [11, Lemma 3.3]). Also we define an edge-wise version of right and left lifting operators as $r^e : [L_2(e)]^d \to \Sigma_{h,p}$ and $l^e : L_2(e) \to \Sigma_{h,p}$, such that

(3.14) $$\int_\Omega r^e(\varphi) \cdot \boldsymbol{\tau} \, \mathrm{d}x = -\int_e \varphi \cdot \{\!\{\boldsymbol{\tau}\}\!\} \, \mathrm{d}s, \qquad \forall \boldsymbol{\tau} \in \Sigma_{h,p}, \forall e \in \mathcal{E}_h,$$

(3.15) $$\int_\Omega l^e(\varphi) \cdot \boldsymbol{\tau} \, \mathrm{d}x = -\int_e \varphi [\![\boldsymbol{\tau}]\!] \, \mathrm{d}s, \qquad \forall \boldsymbol{\tau} \in \Sigma_{h,p}, \forall e \in \mathcal{E}_{h,I}$$

for all edges $e \in \mathcal{E}_h$. Also by noting that $r(\varphi) = \sum_{e \in \mathcal{E}_h} r^e(\varphi)$ and, by applying Cauchy-Schwarz inequality and taking the norm over $\mathcal{T}_h$, one has

(3.16) $$\|r(\varphi)\|^2_{L_2(\Omega)} = \|\sum_{e \in \mathcal{E}_h} r^e(\varphi)\|^2_{L_2(\Omega)} \leq N_l \sum_{e \in \mathcal{E}_h} \|r^e(\varphi)\|^2_{L_2(\Omega)}.$$

Furthermore, one might note that for any $e \in \mathcal{E}_{h,I}$ and $e \subset \partial\kappa, \kappa \in \mathcal{T}_h$ it holds $l^e(\varphi) = 2r^e(\varphi\nu_\kappa)$ and consequently

(3.17) $$\|l(\varphi)\|^2_{L_2(\Omega)} = \|\sum_{e \in \mathcal{E}_h} l^e(\varphi)\|^2_{L_2(\Omega)} \leq N_l \sum_{e \in \mathcal{E}_h} \|l^e(\varphi)\|^2_{L_2(\Omega)} \leq 4N_l \sum_{e \in \mathcal{E}_h} \|r^e(\varphi\nu_\kappa)\|^2_{L_2(\Omega)}.$$

using Cauchy-Schwarz inequality.

Also let us introduce the space $V(h) := V_{h,q} + H^2(\Omega) \bigcap H^1_0(\Omega)$ and the corresponding energy norm $\|\cdot\|_h : V(h) \to \mathbb{R}$ as

(3.18) $$\|v\|^2_h := |v|^2_{1,\Omega} + \sum_{e \in \mathcal{E}_h} \|r^e([\![v_h]\!])\|^2_{L_2(\Omega)}, \qquad \forall v \in V(h)$$

as well as the following seminorm $|\cdot|_{*,h} : V(h) \to \mathbb{R}$ as

(3.19) $$|v|^2_{*,h} := \sum_{e \in \mathcal{E}_h} \|r^e([\![v]\!])\|^2_{L_2(\Omega)}, \qquad \forall v \in V(h).$$

From the structure of (3.19), using (3.16) and (3.17) reads, there exists $C_s < \infty$ such that

(3.20) $$\|r(v)\|^2_{L_2(\Omega)} + \|l(v)\|^2_{L_2(\Omega)} \leq C_s |v|^2_{*,h}$$

for all $v \in V(h)$. Moreover, we have the following estimate from [10, Lemma 2]

LEMMA 3.1. *There exist two positive constants $C_r, C_R > 0$, such that for all $e \in \mathcal{E}_h$ we have*

$$(3.21) \qquad C_r h_e^{-1/2} \|[\![w]\!]\|_{L_2(e)} \leq \|r^e([\![w]\!])\|_{L_2(\Omega)} \leq C_R h_e^{-1/2} \|[\![w]\!]\|_{L_2(e)}, \qquad \forall w \in V(h),$$

where the constants are h independent and only depend on the minimum angle of the triangles and the polynomial degree q.

Proof. The original proof in [10] was presented for $w_h \in V_{h,q}$, while in [3] the extended version to $V(h)$ was given, exploiting the Sobolev embedding to deduce $V(h) \setminus V_{h,q} \subset H^2 \subset \mathcal{C}(\Omega)$ in $\mathbb{R}^2$. Then (3.21) trivially holds for all $w \in V(h) \setminus V_{h,q}$. For the details on the explicit value of C_r and C_r we refer to [10] and [40]. ◻

In order to see the relation between $\|\!|\cdot|\!\|_h$ and $\|\cdot\|_{L_2(\Omega)}$, let us remark the following relation from [2, Lemma 2.1], for all $v \in H^1(\Omega, \mathcal{T}_h)$

$$(3.22) \qquad \|v\|_{L_2(\Omega)}^2 \leq C(\Omega, \mathcal{T}_h) \Big[\|\nabla v\|_{L_2(\Omega)}^2 + \sum_{e \in \mathcal{E}_h} h_e^{-1} \|[\![v]\!]\|_{L_2(e)}^2\Big],$$

which holds when Ω is convex. For a similar result on non-convex domains we refer to [11, 21, 38].

Now, restricted to $v \in V(h)$ and by using (3.21), (3.22) reduces to the following Poincaré-type inequality

$$(3.23) \qquad \|v_h\|_{L_2(\Omega)}^2 \leq C_{\text{en}} \|\!|v_h|\!\|_h^2,$$

with some $C_{\text{en}} < \infty$, and by the definition of the energy norm (3.18) one can write

$$(3.24) \qquad \|v_h\|_{H^1(\Omega, \mathcal{T}_h)}^2 \leq (C_{\text{en}} + 1) \|\!|v_h|\!\|_h^2.$$

Also we will need the following inverse inequality

LEMMA 3.2. *Let consider $v_h \in V_{h,q}$, then for $r \geq 2$ there exists a constant $C_{inv} > 0$ such that*

$$(3.25) \qquad \|v_h\|_{L_r(\kappa)} \leq C_{inv} h_\kappa^{2/r-1} \|v_h\|_{L_2(\kappa)}.$$

The proof of this lemma can be found, e.g., in [13, p. 140] and we skip it.

3.2. Approximation properties. First, let state some approximation properties in the next two lemmas

LEMMA 3.3. *[23, Lemma 2.1] For $\phi \in H^s(\kappa)$, there exists a positive constant C_A, depending on s and q, but independent of ϕ and h_κ, and a sequence $\phi_{h_\kappa} \in \mathcal{P}^q(\kappa)$, $q = 1, 2, \cdots$ such that*

(i) for any $s \geq l + \frac{1}{2}$

$$(3.26) \qquad \|\phi - \phi_{h_\kappa}\|_{H^l(e)} \leq C_A h_\kappa^{\mu-l-1/2} \|\phi\|_{H^s(\kappa)}$$

(ii) for any $0 \leq l \leq s - 1 + \frac{2}{r}$

$$(3.27) \qquad \|\phi - \phi_{h_\kappa}\|_{W_r^l(\kappa)} \leq C_A h_\kappa^{\mu-l-1+2/r} \|\phi\|_{H^s(\kappa)}$$

where $\mu = \min(q+1, s)$.

For the proof of this lemma we refer to [23, Lemma 2.1] and the references cited therein.

Using Lemma 3.3 and the properties of the energy norm (3.18), we present the following approximation result

LEMMA 3.4. *For $\forall \phi \in V(h)$ there exists a constant $C'_A > 0$ independent of h and ϕ, and a mapping $\pi_h : V(h) \to V_{h,q}$ such that*

$$(3.28) \qquad \|\phi - \pi_h \phi\|_h \leq C'_A \Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)} |\phi|^2_{H^s(\kappa)} \Big)^{1/2},$$

where $\mu = \min(q+1, s)$.

Proof. The proof exploits Lemmas 3.1 and 3.3 and the definition of $V(h)$ which provides continuity of $w \in V(h) \setminus V_{h,q}$. Then the proof follows the same lines as [3, section 4.3.]. □

Note that in our analysis in this work, we do not need to specify the explicit type of projection π_h. Hence we leave it undefined with merely assuming that it satisfies the approximation property described in Lemmas 3.3 and 3.4. We refer to [11] and [23] for explicit examples of such projections. Let us remark that the Galerkin $[L_2]^2$ projection that we introduce in (4.9) also satisfies these approximation properties.

4. Discontinuous Galerkin formulation. Here we follow the formulation presented in [11] (also see [14] and [35]). Let us start with writing the nonlinear elliptic problem (2.1) as a system of first order nonlinear PDEs in terms of new variables $(\boldsymbol{\sigma}, \boldsymbol{\theta}, u)$:

$$(4.1) \qquad -\nabla \cdot \boldsymbol{\sigma} = f, \qquad\qquad \text{in } \Omega,$$

$$(4.2) \qquad \boldsymbol{\sigma} = \boldsymbol{a}(u, \boldsymbol{\theta}), \qquad\qquad \text{in } \Omega,$$

$$(4.3) \qquad \boldsymbol{\theta} = \nabla u, \qquad\qquad \text{in } \Omega,$$

$$(4.4) \qquad u = 0, \qquad\qquad \text{on } \partial\Omega.$$

Our goal is to approximate the exact solution $(\boldsymbol{\sigma}, \boldsymbol{\theta}, u)$ by discrete functions $(\boldsymbol{\sigma}_h, \boldsymbol{\theta}_h, u_h)$ in the finite element space $\Sigma_{h,p} \times \Sigma_{h,p} \times V_{h,q}$. The weak formulation can be written as

$$(4.5) \qquad \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h) \cdot \boldsymbol{\zeta}_h \, \mathrm{d}x = \int_\Omega \boldsymbol{\sigma}_h \cdot \boldsymbol{\zeta}_h \, \mathrm{d}x, \qquad\qquad \forall \boldsymbol{\zeta}_h \in \Sigma_{h,p},$$

$$(4.6) \qquad \int_\Omega \boldsymbol{\theta}_h \cdot \boldsymbol{\tau}_h \, \mathrm{d}x + \int_\Omega u_h (\nabla_h \cdot \boldsymbol{\tau}_h) \, \mathrm{d}x = \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} \hat{u} \boldsymbol{\tau}_h \cdot \nu \, \mathrm{d}s, \qquad \forall \boldsymbol{\tau}_h \in \Sigma_{h,p},$$

$$(4.7) \qquad \int_\Omega \boldsymbol{\sigma}_h \cdot \nabla_h v_h \, \mathrm{d}x - \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} v_h (\hat{\boldsymbol{\sigma}} \cdot \nu) \, \mathrm{d}s = \int_\Omega f v_h \, \mathrm{d}x, \qquad \forall v_h \in V_{h,q}.$$

Here ∇_h and $\nabla_h \cdot$ are the element-wise version of the gradient and divergence operator, respectively.

This *flux* formulation is complete but the definition of the numerical fluxes $\hat{u}$ and $\hat{\boldsymbol{\sigma}}$ which depends on $(\boldsymbol{\sigma}_h, \boldsymbol{\theta}_h, u_h)$, and needs to be designed carefully. We postpone the explicit definition of these numerical fluxes till the next section, where we present the equivalent *primal* formulation; i.e., a formulation which has u_h as its only unknown.

The only requirement we impose here is that the $\hat{u}$ should be independent of $\boldsymbol{\theta}_h$ and $\boldsymbol{\sigma}_h$, i.e., $\hat{u} = \hat{u}(u_h)$, which provides the availability of the primal formulation. All the formulations we are going to present have this local property. For examples of other forms of non-local formulation and their analysis we refer to [42], [22] and [20].

4.1. Primal formulation. In order to obtain the primal formulation we need to solve the unknowns $\boldsymbol{\sigma}_h$ and $\boldsymbol{\theta}_h$ in terms of u_h. In the first step, by using (4.5), one can solve for $\boldsymbol{\sigma}_h$ as the

Galerkin $[L_2(\Omega)]^2$ projection,

$$(4.8) \qquad \boldsymbol{\sigma}_h = \mathcal{G}_h\big(\boldsymbol{a}(u_h, \boldsymbol{\theta}_h)\big),$$

where $\mathcal{G}_h : [L_2(\Omega)]^2 \to \Sigma_{h,p}$ has the following property; for all $\boldsymbol{\xi} \in [L_2(\Omega)]^2$

$$(4.9) \qquad \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \boldsymbol{\xi} \cdot \boldsymbol{\tau} \, \mathrm{d}x = \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \mathcal{G}_h(\boldsymbol{\xi}) \cdot \boldsymbol{\tau} \, \mathrm{d}x, \qquad \forall \boldsymbol{\tau} \in \Sigma_{h,p}.$$

Moreover, let us remark the following identity; for any $v \in \prod_{\kappa \in \mathcal{T}_h} L_2(\partial\kappa)$ and $\boldsymbol{\xi} \in \prod_{\kappa \in \mathcal{T}_h}[L_2(\partial\kappa)]^2$ the following holds

$$(4.10) \qquad \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa} v\boldsymbol{\xi} \cdot \nu \, \mathrm{d}s = \sum_{e \in \mathcal{E}_{h,I}} \int_e \{\!\{v\}\!\} [\![\boldsymbol{\xi}]\!] \, \mathrm{d}s + \sum_{e \in \mathcal{E}_h} \int_e [\![v]\!] \cdot \{\!\{\boldsymbol{\xi}\}\!\} \, \mathrm{d}s.$$

Applying (4.10) in (4.6) and (4.7), one can write

$$\int_\Omega \boldsymbol{\theta}_h \cdot \boldsymbol{\tau}_h \, \mathrm{d}x - \int_\Omega \nabla_h u_h \cdot \boldsymbol{\tau}_h \, \mathrm{d}x + \sum_{e \in \mathcal{E}_{h,I}} \int_e \{\!\{u_h - \hat{u}\}\!\} [\![\boldsymbol{\tau}_h]\!] \, \mathrm{d}s + \sum_{e \in \mathcal{E}_h} \int_e [\![u_h - \hat{u}]\!] \cdot \{\!\{\boldsymbol{\tau}_h\}\!\} \, \mathrm{d}s = 0,$$

$$\int_\Omega \boldsymbol{\sigma}_h \cdot \nabla_h v_h \, \mathrm{d}x - \sum_{e \in \mathcal{E}_h} \int_e \{\!\{\hat{\boldsymbol{\sigma}}\}\!\} \cdot [\![v_h]\!] \, \mathrm{d}s - \sum_{e \in \mathcal{E}_{h,I}} \int_e [\![\hat{\boldsymbol{\sigma}}]\!] \{\!\{v_h\}\!\} \, \mathrm{d}s = \int_\Omega f v_h \, \mathrm{d}x.$$

Now let us require the numerical fluxes $\hat{u}$ and $\hat{\boldsymbol{\sigma}}$ to be conservative, which leads to

$$(4.11) \qquad [\![\hat{u}]\!] = 0, \qquad \{\!\{\hat{u}\}\!\} = \hat{u}, \qquad [\![\hat{\boldsymbol{\sigma}}]\!] = 0, \qquad \{\!\{\hat{\boldsymbol{\sigma}}\}\!\} = \hat{\boldsymbol{\sigma}},$$

on any $e \in \mathcal{E}_{h,I}$. Also in accordance with the Dirichlet boundary condition (4.4), we set $\hat{u} = 0$ on $e \in \mathcal{E}_{h,\partial}$. Using (4.11), one can simplify the weak formulation as

$$(4.12) \qquad \int_\Omega \boldsymbol{\theta}_h \cdot \boldsymbol{\tau}_h \, \mathrm{d}x - \int_\Omega \nabla_h u_h \cdot \boldsymbol{\tau}_h \, \mathrm{d}x + \sum_{e \in \mathcal{E}_{h,I}} \int_e (\{\!\{u_h\}\!\} - \hat{u})[\![\boldsymbol{\tau}_h]\!] \, \mathrm{d}s + \sum_{e \in \mathcal{E}_h} \int_e [\![u_h]\!] \cdot \{\!\{\boldsymbol{\tau}_h\}\!\} \, \mathrm{d}s = 0,$$

$$(4.13) \qquad \int_\Omega \boldsymbol{\sigma}_h \cdot \nabla_h v_h \, \mathrm{d}x - \sum_{e \in \mathcal{E}_h} \int_e \hat{\boldsymbol{\sigma}} \cdot [\![v_h]\!] \, \mathrm{d}s = \int_\Omega f v_h \, \mathrm{d}x.$$

Using the definition of lifting operator (3.13) and (4.8), we arrive at

$$(4.14) \qquad \boldsymbol{\theta}_h = \nabla_h u_h + r([\![u_h]\!]) + l(\{\!\{u_h\}\!\} - \hat{u}),$$

$$(4.15) \qquad \boldsymbol{\sigma}_h = \mathcal{G}_h\big(\boldsymbol{a}(u_h, \nabla_h u_h + r([\![u_h]\!]) + l(\{\!\{u_h\}\!\} - \hat{u}))\big).$$

Now $(\boldsymbol{\theta}_h, \boldsymbol{\sigma}_h)$ can be solved locally in terms of u_h by inserting (4.14) and (4.15) in (4.13). Then one can obtain the primal formulation as the following; find $u_h \in V_{h,q}$ such that

$$(4.16) \qquad \mathcal{B}(u_h, v_h) = \mathcal{F}(v_h), \qquad \forall v_h \in V_{h,q},$$

where $\mathcal{F}(v_h) = \int_\Omega f v_h \, \mathrm{d}x$, and

$$(4.17) \qquad \mathcal{B}(u_h, v_h) = \int_\Omega \boldsymbol{a}\big(u_h, \nabla_h u_h + r([\![u_h]\!]) + l(\{\!\{u_h\}\!\} - \hat{u})\big) \cdot \nabla_h v_h \, \mathrm{d}x - \sum_{e \in \mathcal{E}_h} \int_e \hat{\boldsymbol{\sigma}} \cdot [\![v_h]\!] \, \mathrm{d}s.$$

Note that in the rest of the paper we might abuse the notation $\boldsymbol{\theta}_h$ defined in (4.14) in the operator form as $\boldsymbol{\theta}_h(\cdot) := \nabla_h(\cdot) + r(\llbracket \cdot \rrbracket) + l(\{\!\!\{ \cdot \}\!\!\} - \hat{u}(\cdot))$.

The only undefined part in the primal formulation (4.17) is the explicit definition of the numerical fluxes. We are going to consider four different formulations by adopting different $\hat{u}$ and $\hat{\boldsymbol{\sigma}}$:

(i) **BR1.** The BR1 formulation is defined as in [7],

$$\hat{u} = \begin{cases} \{\!\!\{ u_h \}\!\!\} & \text{on } \mathcal{E}_{h,I} \\ 0 & \text{on } \mathcal{E}_{h,\partial} \end{cases}, \qquad \hat{\boldsymbol{\sigma}} = \{\!\!\{ \boldsymbol{\sigma}_h \}\!\!\} \quad \text{on } \mathcal{E}_h.$$

Hence, $\boldsymbol{\theta}_h = \nabla_h u_h + r(\llbracket u_h \rrbracket)$. This leads to the following quasi-linear formulation

$$(4.18) \qquad \mathcal{B}(u_h, v_h) = \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h(u_h)) \cdot \boldsymbol{\theta}_h(v_h) \, \mathrm{d}x.$$

(ii) **BR2.** The BR2 formulation, inherited from the original definition [8], is defined as

$$\hat{u} = \begin{cases} \{\!\!\{ u_h \}\!\!\} & \text{on } \mathcal{E}_{h,I} \\ 0 & \text{on } \mathcal{E}_{h,\partial} \end{cases}, \qquad \hat{\boldsymbol{\sigma}} = \{\!\!\{ \mathcal{G}_h\big(\boldsymbol{a}(u_h, \nabla u_h) + \eta_e \boldsymbol{a}(u_h, r^e(\llbracket u_h \rrbracket))\big) \}\!\!\} \quad \text{on } \mathcal{E}_h$$

with some $\eta_e > 0$ as the stabilization parameter. Here, $\boldsymbol{\theta}_h$ is the same as in BR1, and the primal formulation reads

$$(4.19) \qquad B(u_h, v_h) = \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h) \cdot \nabla_h v_h + \boldsymbol{a}(u_h, \nabla_h u_h) \cdot r(\llbracket v_h \rrbracket) \, \mathrm{d}x$$
$$+ \sum_{e \in \mathcal{E}_h} \eta_e \int_\Omega \boldsymbol{a}(u_h, r^e(\llbracket u_h \rrbracket)) \cdot r^e(\llbracket v_h \rrbracket) \, \mathrm{d}x.$$

(iii) **SIPG.** Similar to [3], we choose the fluxes as

$$\hat{u} = \begin{cases} \{\!\!\{ u_h \}\!\!\} & \text{on } \mathcal{E}_{h,I} \\ 0 & \text{on } \mathcal{E}_{h,\partial} \end{cases}, \qquad \hat{\boldsymbol{\sigma}} = \{\!\!\{ \mathcal{G}_h(\boldsymbol{a}(u_h, \nabla u_h) \}\!\!\} - \frac{\mu_e}{h_e} \llbracket u_h \rrbracket \quad \text{on } \mathcal{E}_h$$

with some penalty parameter $\mu_e > 0$. Similarly to BR2, the primal formulation is given as

$$(4.20)$$
$$\mathcal{B}(u_h, v_h) = \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h) \cdot \nabla_h v_h + \boldsymbol{a}(u_h, \nabla_h u_h) \cdot r(\llbracket v_h \rrbracket) \, \mathrm{d}x + \sum_{e \in \mathcal{E}_h} \frac{\mu_e}{h_e} \int_e \llbracket u_h \rrbracket \cdot \llbracket v_h \rrbracket \, \mathrm{d}s.$$

(iv) **LDG.** The LDG formulation, inherited from the original version in [16], can be obtained by setting

$$\hat{u} = \begin{cases} \{\!\!\{ u_h \}\!\!\} - \boldsymbol{\beta} \cdot \llbracket u_h \rrbracket & \text{on } \mathcal{E}_{h,I} \\ 0 & \text{on } \mathcal{E}_{h,\partial} \end{cases}, \qquad \hat{\boldsymbol{\sigma}} = \begin{cases} \{\!\!\{ \boldsymbol{\sigma}_h \}\!\!\} + \boldsymbol{\beta} \llbracket \boldsymbol{\sigma}_h \rrbracket - \frac{\mu_e}{h_e} \llbracket u_h \rrbracket & \text{on } \mathcal{E}_{h,I} \\ \{\!\!\{ \boldsymbol{\sigma}_h \}\!\!\} - \frac{\mu_e}{h_e} \llbracket u_h \rrbracket & \text{on } \mathcal{E}_{h,\partial} \end{cases}$$

where $\boldsymbol{\beta} \in [L_2(\mathcal{E}_{h,I})]^2$ is some mesh-dependent parameter and constant on each edge. Also like SIPG, we have the penalty parameter $\mu_e > 0$. Note that in LDG formulation $\boldsymbol{\theta}_h = \nabla_h u_h + r(\llbracket u_h \rrbracket) + l(\boldsymbol{\beta} \cdot \llbracket u_h \rrbracket)$. Finally, the primal formulation reads

$$(4.21) \qquad \mathcal{B}(u_h, v_h) = \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h(u_h)) \cdot \boldsymbol{\theta}_h(v_h) \, \mathrm{d}x + \sum_{e \in \mathcal{E}_h} \frac{\mu_e}{h_e} \int_e \llbracket u_h \rrbracket \cdot \llbracket v_h \rrbracket \, \mathrm{d}s.$$

Remark 1. *Seeking comparison of the current formulations with the already proposed versions in the literature, let us record a few observations. Firstly, it is worth pointing out that those fluxes which depend only on u_h and $\boldsymbol{\sigma}_h$ (BR1, LDG) do not require any formal modification compared to the linear case. Furthermore, our LDG formulation is similar to [11] and the discretization proposed in [22] for a simpler form of nonlinearity in the diffusion.*

For the SIPG formulation, looking at (4.20) and exploiting Taylor's expansion (3.1), we can write

$$\mathcal{B}(u_h, v_h) = \int_\Omega \boldsymbol{a}(u_h, \nabla_h u_h) \cdot \nabla_h v_h \, \mathrm{d}x + \sum_{e \in \mathcal{E}_h} \int_e \frac{\mu_e}{h_e} [\![u_h]\!] \cdot [\![v_h]\!] \, \mathrm{d}s$$

$$+ \int_\Omega \tilde{\boldsymbol{a}}_{\boldsymbol{z}}(u_h, \nabla_h u_h, \boldsymbol{\theta}_h) r([\![u_h]\!]) \cdot \nabla_h v_h + \boldsymbol{a}(u_h, \nabla_h u_h) \cdot r([\![v_h]\!]) \, \mathrm{d}x$$

$$= \int_\Omega \boldsymbol{a}(u_h, \nabla_h u_h) \cdot \nabla_h v_h \, \mathrm{d}x + \sum_{e \in \mathcal{E}_h} \int_e \frac{\mu_e}{h_e} [\![u_h]\!] \cdot [\![v_h]\!] \, \mathrm{d}s$$

$$- \sum_{e \in \mathcal{E}_h} \int_e [\![u_h]\!] \cdot \{\!\{\mathcal{G}_h(\tilde{\boldsymbol{a}}_{\boldsymbol{z}}(u_h, \nabla u_h, \boldsymbol{\theta}_h) \nabla v_h)\}\!\} + \{\!\{\mathcal{G}_h(\boldsymbol{a}(u_h, \nabla u_h))\}\!\} \cdot [\![v_h]\!] \, \mathrm{d}s.$$

Ignoring the boundary terms, the primal formulation proposed in [23] (or [27] in simpler case) reads as

$$\mathcal{B}(u_h, v_h) = \int_\Omega \boldsymbol{a}(u_h, \nabla_h u_h) \cdot \nabla_h v_h \, \mathrm{d}x + \sum_{e \in \mathcal{E}_h} \int_e \frac{\mu_e}{h_e} [\![u_h]\!] \cdot [\![v_h]\!] \, \mathrm{d}s$$

$$- \sum_{e \in \mathcal{E}_h} \int_e [\![u_h]\!] \cdot \{\!\{\boldsymbol{a}_{\boldsymbol{z}}(u_h, \nabla_h u_h, \boldsymbol{\theta}_h) \nabla v_h\}\!\} + \{\!\{\boldsymbol{a}(u_h, \nabla_h u_h)\}\!\} \cdot [\![v_h]\!] \, \mathrm{d}s$$

which shows two differences in the second term: in [23] there is $\boldsymbol{a}_{\boldsymbol{z}}$ instead of the average $\tilde{\boldsymbol{a}}_{\boldsymbol{z}}$, which is a consequence of the direct consideration of the primal form. Also in our formulation we have an additional Galerkin projection, which seems essential to later proofs in the paper. Also one might notice that while $\boldsymbol{a}_{\boldsymbol{z}}$ needs to be computed explicitly in [23], in our formulation $\tilde{\boldsymbol{a}}_{\boldsymbol{z}}$ only appears in the analysis and there is no need to compute and implement this term.

Furthermore, comparing the current primal BR2 formulation with the version proposed in [6] for the case of nonlinear Helmholtz equation, and by following similar arguments to those shown for SIPG, one observes that the only difference of two formulations is in the application of the additional Galerkin projection in our formulation.

5. Consistency and adjoint consistency. It is clear that due to the discrete nature of the lifting operator and the Galerkin projection, the primal formulation (4.16) is inconsistent with the exact solution (for all four presented methods). This is also the case for the formulations presented in [11] and [35]. Moreover, the scheme is inconsistent with the smooth solution of its corresponding adjoint problem constructed by the linearization in the neighborhood of the exact solution.

In order to investigate these consistency errors, we consider a smooth solution to (2.1). In sections 5.1 and 5.2 we prove that, despite of the primal and adjoint inconsistency of the formulations, and in case of sufficiently regular solutions, one obtains asymptotic consistency in the mesh refining limit, for primal as well as for the adjoint problem.

In the following of this paper, due to the required regularity in the adjoint analysis, we assume that the exact solution of the problem (2.1), denoted by u, belongs to $W^2_\infty(\Omega)$.

5.1. Consistency. Due to the assumption A(i), the fact that $f \in L_2(\Omega)$ and $-\nabla \cdot \boldsymbol{a}(\cdot, u, \nabla u) = f$ on Ω for the exact solution u, we know that $\boldsymbol{a}(u, \nabla u) \in H(\mathrm{div}; \Omega)$. Hence, $[\![\boldsymbol{a}(u, \nabla u)]\!] = 0$ on each $e \in \mathcal{E}_{h,I}$. Also note that we have $\boldsymbol{\theta}_h(u) = \nabla u$ for all schemes. Let us first consider the schemes BR1, BR2, and SIPG. For these schemes, and for any $v \in V(h)$, one can rewrite the left hand side of (4.17) as

$$
\begin{aligned}
\mathcal{B}(u,v) &= \int_\Omega \boldsymbol{a}(u, \nabla u) \cdot \nabla_h v \, \mathrm{d}x + \boldsymbol{a}(u, \nabla u) \cdot r([\![v]\!]) \, \mathrm{d}x \\
&= \int_\Omega \boldsymbol{a}(u, \nabla u) \cdot \nabla_h v \, \mathrm{d}x - \sum_{e \in \mathcal{E}_h} \int_e \{\!\{\boldsymbol{a}(u, \nabla u)\}\!\} \cdot [\![v]\!] \, \mathrm{d}s + \sum_{e \in \mathcal{E}_h} \int_e \{\!\{(I - \mathcal{G}_h)(\boldsymbol{a}(u, \nabla u))\}\!\} \cdot [\![v]\!] \, \mathrm{d}s.
\end{aligned}
$$

Applying the divergence theorem on the first two terms on the right hand side, and using the continuity of $\boldsymbol{a}(u, \nabla u) \cdot \nu_\kappa$ on the interfaces yield

$$
\mathcal{B}(u,v) = -\int_\Omega \nabla \cdot \boldsymbol{a}(u, \nabla u) v \, \mathrm{d}x + \sum_{e \in \mathcal{E}_h} \int_e \{\!\{(I - \mathcal{G}_h)(\boldsymbol{a}(u, \nabla u))\}\!\} \cdot [\![v]\!] \, \mathrm{d}s.
$$

Noting (2.1), the consistency error for the primal formulation is the same for BR1, BR2 and SIPG; that is

$$
(5.1) \qquad \mathcal{E}_p(u,v) := \mathcal{B}(u,v) - \mathcal{F}(v) = \sum_{e \in \mathcal{E}_h} \int_e \{\!\{(I - \mathcal{G}_h)(\boldsymbol{a}(u, \nabla u))\}\!\} \cdot [\![v]\!] \, \mathrm{d}s, \qquad \forall v \in V(h).
$$

Similarly, for LDG scheme, following the same lines as [11] we obtain

$$
(5.2) \quad \mathcal{E}_p(u,v) := \sum_{e \in \mathcal{E}_h} \int_e \{\!\{(I - \mathcal{G}_h)(\boldsymbol{a}(u, \nabla u))\}\!\} \cdot [\![v]\!] \, \mathrm{d}s - \sum_{e \in \mathcal{E}_{h,I}} \int_e \boldsymbol{\beta}[\![(I - \mathcal{G}_h)(\boldsymbol{a}(u, \nabla u))]\!] \cdot [\![v]\!] \, \mathrm{d}s.
$$

In general, even for very regular solutions, or in case of linear diffusion $\boldsymbol{a}(u, \nabla u) = \nabla u$, this consistency error is not zero and in fact is equal to the Galerkin projection error into the polynomial space $\Sigma_{h,p}$. However, provided that the diffusion $\boldsymbol{a}$ is regular enough, the formulations are asymptotically consistent; i.e., $\|\mathcal{E}_p(u, \cdot)\|_{V'(h)} \to 0$ as $h \to 0$. Note that $\|\cdot\|_{V'(h)}$ is the dual norm on space $V(h)$ defined as

$$
(5.3) \qquad \qquad \|A(\cdot)\|_{V'(h)} := \sup_{0 \neq w \in V(h)} \frac{|A(w)|}{\|w\|_h},
$$

where $A : V(h) \to V'(h)$ is a linear continuous operator on $V(h)$.

In order to prove this asymptotic consistency, one might find an upper bound for $\mathcal{E}_p(u,v)$ which vanishes as h goes to zero. Here we present a generalized form of [11, Lemma 5.2] for different discretizations, which provides us with such an estimate:

LEMMA 5.1. *Assume* $\boldsymbol{a}(u, \nabla u) \in H^{s_*}(\Omega, \mathcal{T}_h)$ *with some non-negative integer* s_*. *Then, there exists* $C_{con} > 0$, *independent of h but dependent on q and s_*, such that for the schemes presented in section 4 the following holds*

$$
(5.4) \qquad |\mathcal{B}(u,w) - \mathcal{F}(w)| \leq C_{con} \Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)^{1/2} \|w\|_h,
$$

for all $w \in V(h)$, *where* $\mu_* = \min(s_*, q + 1)$.

Proof. The proof is provided in [11] for the LDG scheme, which in case $\boldsymbol{\beta} = 0$ gives the desired result for BR1, BR2, and SIPG. □

5.2. Adjoint consistency. Adjoint consistency is an important property in obtaining the optimal L_2 convergence rate for the solution, and super-convergence of the target functionals (cf. [26] or [36]). In order to apply Aubin-Nitsche duality argument, let us consider the following auxiliary dual problem

$$(5.5) \qquad -\nabla \cdot (\boldsymbol{a_z}(u,\nabla u)\nabla\psi) + \boldsymbol{a_u}(u,\nabla u) \cdot \nabla\psi = u - u_h, \qquad \text{in } \Omega,$$

$$(5.6) \qquad \psi = 0, \qquad \text{on } \partial\Omega,$$

where u is the exact solution of (2.1). From assumption A(iii) and $u \in W^2_\infty(\Omega)$, and provided that $u_h \in L_2(\Omega)$ (cf. section 6) one can check that

$$(5.7) \qquad \boldsymbol{a_u}(u,\nabla u) \in W^1_\infty(\Omega), \quad \boldsymbol{a_z}(u,\nabla u) \in W^1_\infty(\Omega), \quad u - u_h \in L_2(\Omega).$$

Using (5.7) combined with the convexity of Ω, it is a classical result that the unique solution of the adjoint problem, $\psi \in H^2 \cap H^1_0$ satisfies the following elliptic regularity [25, Theorem 9.1.22]

$$(5.8) \qquad \|\psi\|_{H^2(\Omega)} \le C\|u - u_h\|_{L_2(\Omega)}.$$

Moreover, from (5.7) and the structure of (5.5), one can conclude that $\boldsymbol{a_z}(u,\nabla u)\nabla\psi \in H(\text{div},\Omega)$. Therefore one gets $[\![\boldsymbol{a_z}(u,\nabla u)\nabla\psi]\!] = 0$, on any interior edge $e \in \mathcal{E}_{h,I}$.

In order to do the linearization, we take the Fréchet derivative of $\mathcal{B}(u,v)$ around the exact solution u. For the BR1 formulation one might get

$$(5.9) \qquad \mathcal{B}'[u](w,v) = \int_\Omega \big(\boldsymbol{a_u}(u,\nabla u)w + \boldsymbol{a_z}(u,\nabla u)(\nabla_h w + r([\![w]\!]))\big) \cdot \big(\nabla_h v + r([\![v]\!])\big) \, \mathrm{d}x.$$

Similarly, for BR2 and SIPG one has

$$\mathcal{B}'[u](w,v) = \int_\Omega \big(\boldsymbol{a_u}(u,\nabla u)w + \boldsymbol{a_z}(u,\nabla u)(\nabla_h w + r([\![w]\!]))\big) \cdot \big(\nabla_h v + r([\![v]\!])\big) \, \mathrm{d}x$$

$$(5.10) \qquad + \int_\Omega \boldsymbol{a_z}(u,\nabla u)r([\![w]\!]) \cdot r([\![v]\!]) \, \mathrm{d}x + F(u,w,v),$$

where the term F is the Fréchet derivative of the penalty term; there holds

$$(5.11) \qquad F(u,w,v) = \sum_{e \in \mathcal{E}_h} \int_e \frac{\mu_e}{h_e}[\![w]\!] \cdot [\![v]\!] \, \mathrm{d}s, \qquad F(u,w,v) = \sum_{e \in \mathcal{E}_h} \eta_e \int_\Omega r^e([\![w]\!]) \cdot r^e([\![v]\!]) \, \mathrm{d}x$$

for SIPG and BR2, respectively. For the LDG formulation we have

$$\mathcal{B}'[u](w,v) = \int_\Omega \big(\boldsymbol{a_u}(u,\nabla u)w + \boldsymbol{a_u}(u,\nabla u)(\nabla w + r([\![w]\!]) + l(\boldsymbol{\beta} \cdot [\![w]\!]))\big) \cdot \big(\nabla_h v + r([\![v]\!]) + l(\boldsymbol{\beta} \cdot [\![v]\!])\big) \, \mathrm{d}x$$

$$+ \sum_{e \in \mathcal{E}_h} \int_e \frac{\mu_e}{h_e}[\![w]\!] \cdot [\![v]\!] \, \mathrm{d}s.$$

Now, let us consider $\mathcal{B}'[u](w,\psi)$ for the smooth exact solution of the dual problem, $\psi \in H^2(\Omega)$; one

might write for BR1, BR2 and SIPG,

$$
\begin{aligned}
\mathcal{B}'[u](v, \psi) &= \int_\Omega \big(\boldsymbol{a}_u(u, \nabla u)v + \boldsymbol{a}_z(u, \nabla u)(\nabla v + r(\llbracket v \rrbracket))\big) \cdot \nabla_h \psi \, \mathrm{d}x \\
&= \int_\Omega \boldsymbol{a}_u(u, \nabla u)v \cdot \nabla_h \psi + (\boldsymbol{a}_z(u, \nabla u)\nabla_h \psi) \cdot \nabla_h v \, \mathrm{d}x + \int_\Omega (\boldsymbol{a}_z(u, \nabla u)\nabla_h \psi) \cdot r(\llbracket v \rrbracket) \, \mathrm{d}x \\
&= \int_\Omega \big(\boldsymbol{a}_u(u, \nabla u) \cdot \nabla_h \psi - \nabla_h \cdot (\boldsymbol{a}_z(u, \nabla u)\nabla_h \psi)\big)v \, \mathrm{d}x \\
&\quad + \sum_{e \in \mathcal{E}_h} \int_e \{\!\{(I - \mathcal{G}_h)(\boldsymbol{a}_z(u, \nabla u)\nabla \psi)\}\!\} \cdot \llbracket v \rrbracket \, \mathrm{d}s \\
&= \int_\Omega (u - u_h)v \, \mathrm{d}x + \mathcal{E}_d(u, v, \psi)
\end{aligned}
$$

for all $v \in V(h)$, with the following definition of the consistency error for the dual problem (5.5)

$$
(5.12) \qquad \mathcal{E}_d(u, v, \psi) := \sum_{e \in \mathcal{E}_h} \int_e \{\!\{(I - \mathcal{G}_h)(\boldsymbol{a}_z(u, \nabla u)\nabla \psi)\}\!\} \cdot \llbracket v \rrbracket \, \mathrm{d}s.
$$

Similarly, for the LDG formulation one can show

$$
\mathcal{E}_d(u, v, \psi) = \sum_{e \in \mathcal{E}_h} \int_e \{\!\{(I - \mathcal{G}_h)(\boldsymbol{a}_z(u, \nabla u)\nabla \psi)\}\!\} \cdot \llbracket v \rrbracket \, \mathrm{d}s - \sum_{e \in \mathcal{E}_{h,I}} \int_e \boldsymbol{\beta} \llbracket (I - \mathcal{G}_h)(\boldsymbol{a}_z(u, \nabla u)\nabla \psi) \rrbracket \cdot \llbracket v \rrbracket \, \mathrm{d}s.
$$

Presence of a non-zero consistency error for the dual problem reveals the dual inconsistency of the proposed scheme, but as we are going to show that one can obtain the *asymptotic dual consistency* of the scheme as proposed in [30]; the quasilinear form $\mathcal{B}$ is called asymptotic dual consistent with the target functional $\mathcal{J} : \mathbb{R} \to \mathbb{R}$ if the following holds

$$
(5.13) \qquad \lim_{h \to 0} \|\mathcal{B}'[u](\cdot, \psi) - \mathcal{J}'[u](\cdot)\|_{V'(h)} = 0,
$$

where u and ψ are the exact solutions of the primal and dual problems, (2.1) and (5.5) respectively. Here $\mathcal{J}'[u]$ is the Fréchet derivative of $\mathcal{J}$, which in our analysis and according to (5.5) is,

$$
(5.14) \quad \mathcal{J}(u) = \sum_{\kappa \in \mathcal{T}_h} \int_\kappa \frac{1}{2}(u - u_h)^2 \, \mathrm{d}x, \qquad \mathcal{J}'[u](w) = \sum_{\kappa \in \mathcal{T}_h} \int_\kappa (u - u_h)w \, \mathrm{d}x, \qquad \forall u, w \in V(h).
$$

As we will prove later in section 8.2, such an asymptotic adjoint consistency leads to the optimal convergence rate in L_2 norm. This property has been already investigated e.g., in [33] by numerical tests. Let us present the following lemma for the adjoint consistency error

LEMMA 5.2. *Assume that the exact solution of the adjoint problem* (5.5) *has elliptic regularity, i.e.,* (5.8) *holds. Then, there exists $\tilde{C}_{con} > 0$, independent of h but dependent on q, such that*

$$
(5.15) \qquad \mathcal{E}_d(u, v, \psi) \leq \tilde{C}_{con}h\|\boldsymbol{a}_z(u, \nabla u)\|_{W^1_\infty(\Omega)}\|\psi\|_{H^2(\Omega)}\|v\|_h,
$$

for all $v \in V(h)$.

Proof. Using assumption A(iii) and Lemmas 3.1 and 3.3, for BR2 and SIPG we have

$$
\begin{aligned}
|\mathcal{E}_d(u,v,\psi)| &= \Big| \sum_{e\in\mathcal{E}_h} \int_e \{\!\{(I-\mathcal{G}_h)(\boldsymbol{a_z}(u,\nabla u)\nabla\psi)\}\!\} \cdot [\![v]\!]\,\mathrm{d}s \Big| \\
&\le \Big(\sum_{e\in\mathcal{E}_h} \|(I-\mathcal{G}_h)(\boldsymbol{a_z}(u,\nabla u)\nabla\psi)\|_{L_2(e)}^2 \Big)^{1/2} \Big(\sum_{e\in\mathcal{E}_h} \|[\![v]\!]\|_{L_2(e)}^2 \Big)^{1/2} \\
&\le \Big(C_A \sum_{\kappa\in\mathcal{T}_h} h_\kappa \|(\boldsymbol{a_z}(u,\nabla u)\nabla\psi)\|_{H^1(\kappa)}^2 \Big)^{1/2} \Big(C_r^{-1} \sum_{e\in\mathcal{E}_h} h_e \|r^e([\![v]\!])\|_{L_2(\Omega)}^2 \Big)^{1/2} \\
&\le Ch\|(\boldsymbol{a_z}(u,\nabla u)\nabla\psi)\|_{H^1(\Omega,\mathcal{T}_h)} |v|_{*,h}.
\end{aligned}
$$

For LDG there exists additional term which can be bounded similarly as

$$
\Big| \sum_{e\in\mathcal{E}_{h,I}} \int_e \boldsymbol{\beta}[\![(I-\mathcal{G}_h)(\boldsymbol{a_z}(u,\nabla u)\nabla\psi)]\!] \cdot [\![v]\!]\,\mathrm{d}s \Big| \le Ch\|\boldsymbol{\beta}\|_{[L_\infty(\mathcal{E}_{h,I})]^2}\|\boldsymbol{a_z}(u,\nabla u)\nabla\psi\|_{H^1(\Omega,\mathcal{T}_h)}|v|_{*,h}.
$$

Using the boundedness of parameter $\boldsymbol{\beta}$ and the fact that

$$
\|\boldsymbol{a_z}(u,\nabla u)\nabla\psi\|_{H^1(\Omega,\mathcal{T}_h)} \le \|\boldsymbol{a_z}(u,\nabla u)\|_{W^1_\infty(\Omega,\mathcal{T}_h)}\|\psi\|_{H^2(\Omega,\mathcal{T}_h)} \le C,
$$

completes the proof of the lemma with sufficiently large $\tilde{C}_{\mathrm{con}}$. $\square$

Using the result of the Lemma 5.2, smoothness of $\boldsymbol{a_z}$ and elliptic regularity of ψ yields

$$
\tag{5.16}
\lim_{h\to 0} \|\mathcal{E}_d(u,\cdot,\psi)\|_{V'(h)} = 0,
$$

which shows the asymptotic adjoint consistency of the proposed formulations.

6. Stability. In this section we prove the stability of the approximated solution of (4.16). Let us remark that in this section we do not assume neither the monotonicity nor the Lipschitz continuity of the diffusion operator. The only additional assumption we need is mentioned as the following remark:

REMARK 2. *Based on the examples in section 1, and our main interest to interpret these problems as nonlinear diffusion phenomenon, it looks natural to assume the following*

$$
\tag{6.1}
\boldsymbol{a}(x,\eta,0) \equiv 0, \qquad \forall x \in \mathbb{R}^2, \eta \in \mathbb{R},
$$

which tells that there is no diffusive behavior in the absence of the gradient.

Using (6.1) *and the Taylor's expansion* (3.1), *as well as assumption A(iii) in form of* (2.5), *one may write*

$$
\tag{6.2}
\boldsymbol{a}(x,\eta,\boldsymbol{\xi}) \cdot \boldsymbol{\xi} = \tilde{\boldsymbol{a}}_{\boldsymbol{z}}(x,\eta,\boldsymbol{\xi}) \cdot \boldsymbol{\xi} \ge \lambda|\boldsymbol{\xi}|^2, \qquad \forall x \in \mathbb{R}^2, \eta \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^2.
$$

We will exploit this result in the rest of this section.

It is well-known that the BR1 method is only weakly-stable (cf. [3]), so we only present the stability result for BR2, SIPG and LDG as the following lemma:

LEMMA 6.1. *Let us assume $u_h \in V_{h,q}$ is the solution of the primal formulation* (4.16). *Then for LDG, BR2 and SIPG methods the following stability result holds*

$$
\tag{6.3}
\|u_h\|_h \le \frac{C_F}{C_{co}}\|f\|_{L_2(\Omega)}.
$$

Here, C_F is the continuity constant of linear form $\mathcal{F}(\cdot)$, that is

$$
(6.4) \qquad |\mathcal{F}(v)| \leq C_F \|f\|_{L_2(\Omega)} \|v\|_h, \qquad \forall v \in V(h),
$$

while $C_{co} > 0$ is the following coercivity constant of the operator $\mathcal{B}$ as

$$
(6.5) \qquad C_{co} \|v_h\|_h^2 \leq \mathcal{B}(v_h, v_h), \qquad \forall v_h \in V_{h,q}.
$$

Proof. For different formulations, we insert the test function $v_h = u_h$ and investigate if (6.5) actually holds. Using the fact that $\mathcal{B}(u_h, u_h) = \mathcal{F}(u_h)$, the rest of the proof is complete by the following line

$$
(6.6) \qquad \|u_h\|_h^2 \leq \frac{1}{C_{co}} \mathcal{F}(u_h) \leq \frac{C_F}{C_{co}} \|f\|_{L_2(\Omega)} \|u_h\|_h.
$$

So the only missing step is the coercivity proof, for which we check $\mathcal{B}(u_h, u_h)$ for different formulations:

(i) LDG: From (4.21) and (4.14) one has

$$
\mathcal{B}(u_h, u_h) = \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h) \cdot \boldsymbol{\theta}_h \, \mathrm{d}x + \sum_{e \in \mathcal{E}_h} \frac{\mu_e}{h_e} \int_e [\![u_h]\!] \cdot [\![u_h]\!] \, \mathrm{d}s.
$$

From the property (6.1) we have

$$
\int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h) \cdot \boldsymbol{\theta}_h \, \mathrm{d}x \geq \lambda \| \nabla_h u_h + r([\![u_h]\!]) + l(\boldsymbol{\beta} \cdot [\![u_h]\!]) \|_{L_2(\Omega)}^2
$$

$$
(6.7) \qquad \geq \lambda \Big[(1 - \delta) \| \nabla_h u_h \|_{L_2(\Omega)}^2 + (1 - \frac{1}{\delta}) \| r([\![u_h]\!]) + l(\boldsymbol{\beta} \cdot [\![u_h]\!]) \|_{L_2(\Omega)}^2 \Big]
$$

using Young's inequality with $\delta \in (0, 1)$. Now using (3.20) and the boundedness of $\boldsymbol{\beta}$, one has

$$
(6.8) \qquad \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h) \cdot \boldsymbol{\theta}_h \, \mathrm{d}x \geq \lambda \Big[(1 - \delta) \| \nabla_h u_h \|_{L_2(\Omega)}^2 + C(1 - \frac{1}{\delta}) |u_h|_{*,h}^2 \Big],
$$

for some constant C. Then (3.21) gives the following

$$
(6.9) \qquad \mathcal{B}(u_h, u_h) \geq \lambda(1 - \delta) \| \nabla_h u_h \|_{L_2(\Omega)}^2 + \big(\frac{\mu_e}{C_R^2} + \lambda C(1 - \frac{1}{\delta}) \big) |u_h|_{*,h}^2,
$$

which indicates the condition $\mu_e > 0$ as the coercivity requirement of LDG formulation. This coincides with the result for the Poisson problems (cf. [3]) and the version of LDG discussed in [31].

(ii) BR2, SIPG: We combine the proof for these two methods by writing them as a variation of BR1 method. This treatment will be used once more in section 7.1. Using (4.18), (4.19) and (4.20), one may write the following decomposition of primal formulation; for all $v, w \in V(h)$

$$
(6.10) \qquad \mathcal{B}_{\text{SIPG,BR2}}(v, w) = \mathcal{B}_{\text{BR1}}(v, w) + T^{(1)}(v, w) + T_{\text{SIPG,BR2}}^{(2)}(v, w),
$$

where $\mathcal{B}_{\mathrm{BR1}}$ is the primal formulation of BR1 scheme and the next two terms are defined as

$$(6.11) \qquad T^{(1)}(v, w) = -\int_\Omega \left(\boldsymbol{a}(v, \boldsymbol{\theta}_h(v)) - \boldsymbol{a}(v, \nabla_h v) \right) \cdot r(\llbracket w \rrbracket) \, \mathrm{d}x,$$

$$(6.12) \qquad T^{(2)}_{\mathrm{SIPG}}(w, v) = \sum_{e \in \mathcal{E}_h} \frac{\mu_e}{h_e} \int_e \llbracket v \rrbracket \cdot \llbracket w \rrbracket \, \mathrm{d}s,$$

$$(6.13) \qquad T^{(2)}_{\mathrm{BR2}}(w, v) = \sum_{e \in \mathcal{E}_h} \eta_e \int_\Omega \boldsymbol{a}(v, r^e(\llbracket v \rrbracket)) \cdot r^e(\llbracket w \rrbracket) \, \mathrm{d}x.$$

Using (6.2) one can easily show the positive semi-definiteness of $\mathcal{B}_{\mathrm{BR1}}(u_h, u_h)$ as (4.18). For $T^{(1)}(u_h, u_h)$ using the Taylor's expansion (3.1) we have

$$T^{(1)}(u_h, u_h) = -\int_\Omega \tilde{\boldsymbol{a}}_{\boldsymbol{z}}(u_h, \nabla_h u_h, \boldsymbol{\theta}_h) r(\llbracket u_h \rrbracket) \cdot r(\llbracket u_h \rrbracket) \, \mathrm{d}x,$$

and for the remaining term $T^{(2)}_{\mathrm{BR2,SIPG}}$ one can obtain

$$T^{(2)}_{\mathrm{SIPG}}(u_h, u_h) = \sum_{e \in \mathcal{E}_h} \frac{\mu_e}{h_e} \|\llbracket u_h \rrbracket\|^2_{L_2(e)} \geq \sum_{e \in \mathcal{E}_h} \frac{\mu_e}{C_R^2} \|r^e(\llbracket u_h \rrbracket)\|^2_{L_2(\Omega)},$$

$$T^{(2)}_{\mathrm{BR2}}(u_h, u_h) \geq \lambda \eta_e \sum_{e \in \mathcal{E}_h} \|r^e(\llbracket u_h \rrbracket))\|^2_{L_2(\Omega)}.$$

Here we have used (3.21) and (6.2) in the first and second equation, respectively. Henceforth, we adopt the following notation for a constant $C_\mathrm{T} > 0$

$$(6.14) \qquad \mathrm{SIPG} : C_\mathrm{T} := \min_{e \in \mathcal{E}_h} \{\mu_e / C_R^2\}, \qquad \mathrm{BR2} : C_\mathrm{T} := \min_{e \in \mathcal{E}_h} \{\lambda \eta_e\}.$$

Hence, using (6.10), (6.2) and (6.14), one might write

$$\mathcal{B}_{\mathrm{SIPG,BR2}}(u_h, u_h) \geq \int_\Omega \boldsymbol{a}(u_h, \boldsymbol{\theta}_h) \cdot \boldsymbol{\theta}_h \, \mathrm{d}x - \int_\Omega \tilde{\boldsymbol{a}}_{\boldsymbol{z}}(u_h, \nabla_h u_h, \boldsymbol{\theta}_h) r(\llbracket u_h \rrbracket) \cdot r(\llbracket u_h \rrbracket) \, \mathrm{d}x$$

$$+ C_\mathrm{T} \sum_{e \in \mathcal{E}_h} \|r^e(\llbracket u_h \rrbracket))\|^2_{L_2(\Omega)}$$

$$\geq \lambda \|\boldsymbol{\theta}_h(u_h))\|^2_{L_2(\Omega)} - \Lambda \|r(\llbracket u_h \rrbracket)\|^2_{L_2(\Omega)} + C_\mathrm{T} |u_h|^2_{*,h}.$$

Applying Young inequality with $0 < \delta < 1$ and (3.16), we arrive at

$$\mathcal{B}_{\mathrm{SIPG,BR2}}(u_h, u_h) \geq \lambda(1 - \delta) \|\nabla_h u_h\|^2_{L_2(\Omega)} - N_l \left(\Lambda + \lambda(\frac{1}{\delta} - 1) \right) |u_h|^2_{*,h} + C_\mathrm{T} |u_h|^2_{*,h}$$

This leads to the following criterion $N_l(\Lambda + \lambda(\frac{1}{\delta} - 1)) \leq C_\mathrm{T}$, and since δ can be arbitrary chosen close to 1, the margin for stability is $C_\mathrm{T} > N_l \Lambda$ which can be written separately for BR2 and SIPG as

$$(6.15) \qquad \mathrm{SIPG} : \mu_e > C_R^2 \Lambda N_l, \qquad \mathrm{BR2} : \eta_e > N_l \frac{\Lambda}{\lambda}.$$

This shows that by choosing sufficiently large penalty parameter, as (6.15) suggests, one can ensure the stability of the solution of SIPG and BR2 formulations.

This completes the proof of coercivity as well the proof of the lemma. ∎

REMARK 3. *The result of Lemma 6.1 shows that the stability estimate is available in the degenerate case, i.e. where $\lambda = 0$, for SIPG method while for BR2 the uniform lower bound λ should be uniformly larger than zero. In the special case when $\boldsymbol{a}(u, \nabla u) = a(u)\nabla u$, by the same lines of argument as in [43], one can check that the stability criteria for SIPG remains unchanged while the BR2 one reduces to $\eta_e \geq N_l$, which is consistent with the result of Poisson problem (cf. [3]).*

7. Existence and uniqueness of the discrete solution. In the following sections of this paper, we only consider the BR2 and SIPG formulations. The reason for doing this is the fact that our discussions on the LDG method will follow basically the arguments presented in [11]. Moreover, the BR1 method is not so common due to non-compact stencil and instability [3].

We first prove the strong monotonicity and Lipschitz continuity of the nonlinear operator corresponding to the primal formulation in sections 7.1 and 7.2, respectively. Then we prove the uniqueness of the discrete solution and present a Strang type error estimate.

7.1. Strong monotonicity. Let us start with the following lemma

LEMMA 7.1. *If the diffusion operator $\boldsymbol{a}$ satisfies the strong monotonicity property as (2.6), then there exists $C_{SM} > 0$ such that*

$$(7.1) \qquad \mathcal{B}(v, v - w) - \mathcal{B}(w, v - w) \geq C_{SM} \|v - w\|_h^2$$

for all $w, v \in V(h)$.

Before stating the proof, let us remark the following estimate which later will be used in the analysis:

LEMMA 7.2. *For all $w \in V(h)$, the following holds*

$$(7.2) \qquad \|\boldsymbol{\theta}_h(w)\|_{L_2(\Omega)}^2 \geq \frac{1}{2}\|w\|_h^2 - \eta|w|_{*,h}^2,$$

for $\eta \geq N_l + \frac{1}{2}$.

Proof. Using Young's inequality with $0 < \delta < 1$, (4.14) and (3.16), one can write

$$\|\boldsymbol{\theta}_h(w)\|_{L_2(\Omega)}^2 = \|\nabla_h w\|_{L_2(\Omega)}^2 + \|r(\llbracket w \rrbracket)\|_{L_2(\Omega)}^2 - \int_\Omega 2r(\llbracket w \rrbracket) \cdot \nabla_h w \, \mathrm{d}x$$

$$\geq (1 - \delta)\|\nabla_h w\|_{L_2(\Omega)}^2 + N_l(1 - \frac{1}{\delta}) \sum_{e \in \mathcal{E}_h} \|r^e(\llbracket w \rrbracket)\|_{L_2(\Omega)}^2.$$

Hence, by setting $\delta = 1/2$ we arrive at

$$\|\boldsymbol{\theta}_h(w)\|_{L_2(\Omega)}^2 \geq \frac{1}{2}\|\nabla_h w\|_{L_2(\kappa)}^2 - N_l|w|_{*,h}^2.$$

Using the definition of energy norm (3.18), one can add $\pm 1/2|w|_{*,h}^2$ to the right hand side and the result is obtained. ∎

Now we are ready to present the proof of Lemma 7.1:

Proof. [of Lemma 7.1] Using the decomposition (6.10), and setting $\xi_h = v_h - w_h$ one gets

$$\begin{aligned}
\mathcal{B}(v_h, \xi_h) - \mathcal{B}(w_h, \xi_h) &= \big[\mathcal{B}_{\mathrm{BR1}}(v_h, \xi_h) - \mathcal{B}_{\mathrm{BR1}}(w_h, \xi_h)\big] \\
&\quad + \big[T^{(1)}(v_h, \xi_h) - T^{(1)}(w_h, \xi_h)\big] + \big[T^{(2)}(v_h, \xi_h) - T^{(2)}(w_h, \xi_h)\big] \\
(7.3) \qquad &= \big[\mathcal{B}_{\mathrm{BR1}}(v_h, \xi_h) - \mathcal{B}_{\mathrm{BR1}}(w_h, \xi_h)\big] + F_1 + F_2.
\end{aligned}$$

Strong monotonicity (2.6) and (4.18) readily imply

$$(7.4) \qquad \mathcal{B}_{\mathrm{BR1}}(v_h, \xi_h) - \mathcal{B}_{\mathrm{BR1}}(w_h, \xi_h) \geq C_{\mathrm{sm}} \| \boldsymbol{\theta}_h(v_h) - \boldsymbol{\theta}_h(w_h) \|^2_{L_2(\Omega)}.$$

For the term F_1, using (2.7), (6.11) and a Young inequality with $\delta > 0$, one has

$$\begin{aligned} F_1 &\geq -\frac{1}{\delta}\Big(\| \boldsymbol{a}(v_h, \boldsymbol{\theta}_h(v_h)) - \boldsymbol{a}(w_h, \boldsymbol{\theta}_h(w_h)) \|^2_{L_2(\Omega)} \Big) \\ &\quad - \frac{1}{\delta}\Big(\| \boldsymbol{a}(v_h, \nabla_h v_h) - \boldsymbol{a}(w_h, \nabla_h w_h) \|^2_{L_2(\Omega)} \Big) - \frac{\delta}{2} \| r(\llbracket \xi_h \rrbracket) \|^2_{L_2(\Omega)} \\ &\geq -\frac{2C_{\mathrm{lc}}^2}{\delta} \| \xi_h \|^2_{L_2(\Omega)} - \frac{C_{\mathrm{lc}}^2}{\delta} \| \boldsymbol{\theta}_h(\xi_h) \|^2_{L_2(\Omega)} - \frac{C_{\mathrm{lc}}^2}{\delta} \| \nabla_h \xi_h \|^2_{L_2(\Omega)} - \frac{\delta}{2} \| r(\llbracket \xi_h \rrbracket) \|^2_{L_2(\Omega)} \end{aligned}$$

Using the fact that $-|\nabla_h \xi_h|^2 \geq -2|\boldsymbol{\theta}_h(\xi_h)|^2 - 2|r(\llbracket \xi_h \rrbracket)|^2$ and (3.16), we arrive at

$$F_1 \geq -\frac{2C_{\mathrm{lc}}^2}{\delta} \| \xi_h \|^2_{L_2(\Omega)} - \frac{3C_{\mathrm{lc}}^2}{\delta} \| \boldsymbol{\theta}_h(\xi_h) \|^2_{L_2(\Omega)} - N_l\Big(\frac{2C_{\mathrm{lc}}^2}{\delta} + \frac{\delta}{2} \Big) \sum_{e \in \mathcal{E}_h} \| r^e(\llbracket \xi_h \rrbracket) \|^2_{L_2(\Omega)}.$$

For the remaining term F_2, using the definitions (6.12) and (6.13), as well (3.21) and (2.6), one has

$$F_2 \geq C_R^{-2} \sum_{e \in \mathcal{E}_h} \mu_e \| r^e(\llbracket \xi_h \rrbracket) \|^2_{L_2(\Omega)}, \qquad F_2 \geq C_{\mathrm{sm}} \sum_{e \in \mathcal{E}_h} \eta_e \| r^e(\llbracket \xi_h \rrbracket) \|^2_{L_2(\Omega)}$$

for SIPG and BR2 methods, respectively. Similar to (6.14), we define C'_{T} such that, for both methods

$$(7.5) \qquad F_2(w_h, v_h) \geq C'_{\mathrm{T}} \sum_{e \in \mathcal{E}_h} \| r^e(\llbracket \xi_h \rrbracket) \|^2_{L_2(\Omega)}.$$

Hence, combining all terms one can write

$$\begin{aligned} \mathcal{B}(v_h, \xi_h) - \mathcal{B}(w_h, \xi_h) &\geq C_{\mathrm{sm}} \| \boldsymbol{\theta}_h(\xi_h) \|^2_{L_2(\Omega)} + C'_{\mathrm{T}} |\xi_h|^2_{*,h} \\ &\quad - \frac{2C_{\mathrm{lc}}^2}{\delta} \| \xi_h \|^2_{L_2(\Omega)} - \frac{3C_{\mathrm{lc}}^2}{\delta} \| \boldsymbol{\theta}_h(\xi_h) \|^2_{L_2(\Omega)} - N_l\Big(\frac{2C_{\mathrm{lc}}^2}{\delta} + \frac{\delta}{2} \Big) |\xi_h|^2_{*,h} \end{aligned}$$

Let us set η to be a constant larger than $N_l + 1/2$; then using Lemma 7.2 and (3.23) give

$$\begin{aligned} \mathcal{B}(v_h, \xi_h) - \mathcal{B}(w_h, \xi_h) &\geq \frac{1}{2} \Big[C_{\mathrm{sm}} - \frac{3C_{\mathrm{lc}}^2}{\delta} - \frac{4C_{\mathrm{lc}}^2 C_{\mathrm{en}}}{\delta} \Big] \| \xi_h \|^2_h \\ &\quad + \Big[C'_{\mathrm{T}} - \big(C_{\mathrm{sm}} - \frac{3C_{\mathrm{lc}}^2}{\delta} \big) \eta - N_l \big(\frac{2C_{\mathrm{lc}}^2}{\delta} + \frac{\delta}{2} \big) \Big] |\xi_h|^2_{*,h}. \end{aligned}$$

Now consider arbitrary $0 < C_{\mathrm{SM}} < \frac{C_{\mathrm{sm}}}{2}$ and set δ large enough such that the following holds

$$C_{\mathrm{sm}} - \frac{C_{\mathrm{lc}}^2(3 + 4C_{\mathrm{en}})}{\delta} > 2C_{\mathrm{SM}}, \quad \text{or} \quad \delta > \frac{C_{\mathrm{sm}} - 2C_{\mathrm{SM}}}{C_{\mathrm{lc}}^2(3 + 4C_{\mathrm{en}})}.$$

The only remaining free parameter is the stabilization parameter C'_{T} and one can choose it sufficiently large, such that

$$C'_{\mathrm{T}} \geq \big(C_{\mathrm{sm}} - \frac{3C_{\mathrm{lc}}^2}{\delta} \big) \eta + N_l \big(\frac{2C_{\mathrm{lc}}^2}{\delta} + \frac{\delta}{2} \big).$$

Finally we arrive at $\mathcal{B}(v_h, \xi_h) - \mathcal{B}(w_h, \xi_h) \geq C_{\mathrm{SM}} \| \xi_h \|^2_h$, for both BR2 and SIPG and the proof completes. $\square$

7.2. Lipschitz continuity. For the Lipschitz continuity we have the following lemma

LEMMA 7.3. *If the diffusion operator $\boldsymbol{a}$ satisfies the Lipschitz continuity property as (2.6), there exists $C_{LC} < \infty$ independent of the mesh size such that*

(7.6) $$|\mathcal{B}(z, w) - \mathcal{B}(v, w)| \leq C_{LC}\|z - v\|_h \|w\|_h,$$

for all $z, v, w \in V(h)$.

Proof. Using the decomposition (6.10) and similar to Lemma 7.1 one gets

(7.7) $$\mathcal{B}(z, w) - \mathcal{B}(v, w) = \big[\mathcal{B}_{\mathrm{BR1}}(z, w) - \mathcal{B}_{\mathrm{BR1}}(v, w)\big] + F_1 + F_2.$$

where

(7.8) $$F_1 = T^{(1)}(z, w) - T^{(1)}(v, w), \qquad F_2 = T^{(2)}(z, w) - T^{(2)}(v, w).$$

It is straightforward to show the Lipschitz continuity of $\mathcal{B}_{\mathrm{BR1}}(v, w)$ using (2.7). For F_1, noting the fact that one may write

$$\begin{aligned}
|\boldsymbol{a}(z, \nabla_h z) - \boldsymbol{a}(v, \nabla_h v) - \boldsymbol{a}(z, \boldsymbol{\theta}_h(z)) + \boldsymbol{a}(v, \boldsymbol{\theta}_h(v))| &\leq C_{\mathrm{lc}}\Big(|z - v|^2 + |\boldsymbol{\theta}_h(z) - \boldsymbol{\theta}_h(v)|^2\Big)^{1/2} \\
&\quad + C_{\mathrm{lc}}\Big(|z - v|^2 + |\nabla_h z - \nabla_h v|^2\Big)^{1/2} \\
&\leq C\Big(|z - v| + |\nabla_h(z - v)| + r(\llbracket z - v \rrbracket)|\Big),
\end{aligned}$$

and readily

$$F_1 \leq C \int_\Omega \Big(|z - v| + |\nabla_h(z - v)| + r(\llbracket z - v \rrbracket)|\Big)|r(\llbracket w \rrbracket)|\,\mathrm{d}x \leq C_{\mathrm{LC}}\|z - v\|_h \|w\|_h.$$

For BR2 scheme one may note, for any $e \in \mathcal{E}_h$

$$|\boldsymbol{a}(z, r^e(\llbracket z \rrbracket)) - \boldsymbol{a}(v, r^e(\llbracket v \rrbracket))| \leq C_{\mathrm{lc}}\Big(|z - v|^2 + |r^e(\llbracket z - v \rrbracket)|^2\Big)^{1/2} \leq C_{\mathrm{lc}}\Big(|z - v| + |r^e(\llbracket z - v \rrbracket)|\Big).$$

On the other hand, using the fact the r^e vanishes outside two neighbor elements, for the corresponding F_2 term we have,

$$F_2 \leq \sum_{e \in \mathcal{E}_h} \eta_e \int_\Omega C_{\mathrm{lc}}\Big(|z - v| + |r^e(\llbracket z - v \rrbracket)|\Big)|r^e(\llbracket w \rrbracket)|\,\mathrm{d}x \leq C_{\mathrm{LC}}\|z - v\|_h \|w\|_h.$$

Finally, for SIPG method one can easily write

$$F_2 = \sum_{e \in \mathcal{E}_h} \frac{\mu_e}{h_e} \int_e \llbracket v - z \rrbracket \cdot \llbracket w \rrbracket \,\mathrm{d}s \leq C_{\mathrm{LC}}\|z - v\|_h \|w\|_h.$$

Combining the results for $\mathcal{B}_{\mathrm{BR1}}$, F_1 and F_2 concludes the proof. $\square$

Now we present the existence, uniqueness and stability result for the approximated solution u_h as well as a Strang type error estimate. Though we discussed the stability of the solution in section 6, note that the result of Lemma 7.4 requires stronger condition on the diffusion operator

(strong monotonicity and global Lipschitz continuity) than the more general result already discussed in section 6.

LEMMA 7.4. *There exists a unique $u_h \in V_{h,q}$ solution of (4.16), which satisfies*

$$(7.9) \qquad \|u_h\|_h \leq \frac{1}{C_{SM}}\Big[C_F\|f\|_{L_2(\Omega)} + \|\mathcal{B}(0,\cdot)\|_{V'(h)}\Big].$$

Moreover, the following error estimate holds

$$(7.10) \qquad \|u - u_h\|_h \leq \big(1 + \frac{C_{LC}}{C_{SM}}\big) \inf_{v_h \in V_{h,q}} \|u - v_h\|_h + \frac{1}{C_{SM}} \sup_{0 \neq w_h \in V_{h,q}} \frac{|\mathcal{B}(u,w_h) - \mathcal{F}(w_h)|}{\|w_h\|_h},$$

where u is the exact solution of (2.1).

Proof. Using the strong monotonicity and Lipschitz continuity (Lemmas 7.1 and 7.3) the unique solvability of (4.16) can be proved by a well-known result as [32, Theorem 3.2.23] or [45, Theorem 35.4], also see [28]. For the stability proof we refer to [11, Theorem 4.5] and here we only present the proof of Strang type error estimate due to its application in the rest of our analysis.

Consider the error $e = u - u_h$ and decompose it as $e = \eta + \xi$, where $\eta = u - v_h$ and $\xi = v_h - u_h$ where $v_h \in V_{h,q}$. From Lemma 7.1 one has

$$C_{\mathrm{SM}}\|u_h - v_h\|_h^2 \leq \mathcal{B}(u_h, \xi) - \mathcal{B}(v_h, \xi) = \big[\mathcal{B}(u_h, \xi) - \mathcal{B}(u, \xi)\big] + \big[\mathcal{B}(u, \xi) - \mathcal{B}(v_h, \xi)\big]$$

While the first group of terms is the consistency error, applying Lemma 7.3 gives

$$(7.11) \qquad \|u_h - v_h\|_h \leq \frac{C_{\mathrm{LC}}}{C_{\mathrm{SM}}}\|u - v_h\|_h + \frac{1}{C_{\mathrm{SM}}} \sup_{0 \neq w_h \in V_{h,q}} \frac{|\mathcal{B}(u,w_h) - \mathcal{F}(w_h)|}{\|w_h\|_h}$$

Applying a triangle inequality completes the proof. □

8. A priori error estimates. In this section, we provide the error estimate of the BR2 and SIPG methods illustrated in section 4. In section 8.1, we prove the optimal error estimate in the energy norm $\|\cdot\|_h$ using the Strang type error estimate in Lemma 7.4 and the asymptotic consistency result in section 5.1. In section 8.2, we prove the optimal error estimate in L_2 norm exploiting the result on asymptotic adjoint consistency already provided in section 5.2.

8.1. Energy norm error estimate. Combining the result of Lemmas 7.4 and 5.1, gives the following corollary for the error estimate in the energy norm:

COROLLARY 8.1. *Let u_h and u be the solution of (4.16) and the exact solution of (2.1), respectively. Also assume that $u \in H^s(\Omega, \mathcal{T}_h)$ and $\mathbf{a}(u, \nabla u) \in H^{s_*}(\Omega, \mathcal{T}_h)$. Then the following holds*

$$(8.1) \qquad \|u - u_h\|_h^2 \leq C_{err}\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)}\|u\|_{H^s(\kappa)}^2 + \sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*}\|\mathbf{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)$$

with some $C_{err} > 0$, $\mu = \min(s, q+1)$ and $\mu_* = \min(s_*, q+1)$.

Proof. Using Lemmas 7.4 and 5.1 one can easily write

$$(8.2) \qquad \|u - u_h\|_h \leq \big(1 + \frac{C_{\mathrm{LC}}}{C_{\mathrm{SM}}}\big) \inf_{v_h \in V_{h,q}} \|u - v_h\|_h + \frac{C_{\mathrm{con}}}{C_{\mathrm{SM}}}\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*}\|\mathbf{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)^{1/2}.$$

Choosing $v_h = \pi_h u$ and applying the approximation result in Lemma 3.4 completes the proof with choosing a sufficiently large C_{err}. □

Using the error decomposition $e = \eta + \xi$ (as in Lemma 7.4), we also are interested in obtaining an estimate for $\xi = \pi_h u - u_h$. By setting $v_h = \pi_h u$ in (7.11) and similar to the proof of Corollary 8.1 one might get

$$(8.3) \qquad \|u_h - \pi_h u\|_h \leq \frac{C_{\mathrm{LC}}}{C_{\mathrm{SM}}} \|u - \pi_h u\|_h + \frac{C_{\mathrm{con}}}{C_{\mathrm{SM}}} \Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s*}(\kappa)}^2 \Big)^{1/2},$$

which yields, for some $\tilde{C}_{\mathrm{err}} > 0$

$$(8.4) \qquad \|u_h - \pi_h u\|_h \leq \tilde{C}_{\mathrm{err}} \Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)} \|u\|_{H^s(\kappa)}^2 + \sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s*}(\kappa)}^2 \Big)^{1/2}.$$

Moreover, using the definition of $\boldsymbol{\sigma}_h$ and $\boldsymbol{\theta}_h$ as (4.15) and (4.14), one can prove the corresponding lemma for their error estimate

LEMMA 8.1 (Theorem 5.5 in [11]). *Under the same assumptions as those of Corollary 8.1, there exists $\tilde{C}_{err} > 0$ independent of the mesh size, such that*

$$(8.5) \qquad \|\boldsymbol{\theta} - \boldsymbol{\theta}_h\|_{[L_2(\Omega)]^2} \leq \tilde{C}_{err} \Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)} \|u\|_{H^s(\kappa)}^2 + \sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s*}(\kappa)}^2 \Big)^{1/2},$$

and, not necessarily with the same $\tilde{C}_{err}$,

$$(8.6) \qquad \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{[L_2(\Omega)]^2} \leq \tilde{C}_{err} \Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)} \|u\|_{H^s(\kappa)}^2 + \sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s*}(\kappa)}^2 \Big)^{1/2},$$

where $\mu = \min(s, q+1)$ and $\mu_* = \min(s_*, q+1)$.

The proof follows the same lines as [11] and we skip it here.

8.2. L_2 norm error estimate. In this section, we present the error estimate of the solution u_h in the L_2 norm. Recalling the adjoint problem (5.5), and by inserting $w = u - u_h$ as the (infinite dimensional) test function in (5.10) and (5.14), to obtain

$$\|u_h - u\|_{L_2(\Omega)}^2 = \mathcal{B}'[u](u_h - u, \psi) - \Big(B'[u](u_h - u, \psi) - J'[u](u_h - u) \Big)$$

$$= \mathcal{B}(u_h, \psi) - \mathcal{B}(u, \psi) - \mathcal{N}(u, u_h, \psi) \pm \big(\mathcal{B}(u, \psi_h) - \mathcal{B}(u_h, \psi_h) \big) + \mathcal{E}_d(u, e, \psi)$$

$$(8.7) \qquad = \mathcal{B}(u_h, \psi - \psi_h) - \mathcal{B}(u, \psi - \psi_h) - \mathcal{E}_p(u, \psi_h) + \mathcal{E}_d(u, e, \psi) - \mathcal{N}(u, u_h, \psi)$$

where $\mathcal{E}_p$ and $\mathcal{E}_d$ are defined in (5.1) and (5.12) as the consistency errors of the primal and the dual problem, respectively, while $\mathcal{N}(u, u_h, \psi)$ is the second order linearization error. Here, $\psi_h = \pi_h \psi \in V_{h,q}$. By the Taylor's formula in section 3.2 we have (note that $\boldsymbol{\theta}_h(e) = \nabla_h e + r(\llbracket e \rrbracket)$ as already defined)

$$(8.8) \quad \mathcal{N}(u, u_h, \psi) = \sum_{\kappa \in \mathcal{T}_h} \int_\kappa R_a(u - u_h, \nabla u - \nabla u_h) \cdot \nabla \psi \, \mathrm{d}x$$

$$= \int_\Omega \Big[\tilde{\boldsymbol{a}}_{uu}(u, \nabla u) e^2 + \boldsymbol{\theta}_h(e)^t \tilde{\boldsymbol{a}}_{\boldsymbol{zz}}(u, \nabla u) \boldsymbol{\theta}_h(e) + 2\tilde{\boldsymbol{a}}_{uz}(u, \nabla u) \cdot \boldsymbol{\theta}_h(e) e \Big] \cdot \nabla_h \psi \, \mathrm{d}x.$$

First let us handle the error estimate of the last three terms on the right hand side of (8.7), $\mathcal{N}(u, u_h, \psi)$, $\mathcal{E}_p(u, \psi_h)$ and $\mathcal{E}_d(u, u_h, \psi)$, in the following steps:

(i) Using (8.8) and very similar arguments as [23, Lemma 3.10] we have

$$|\mathcal{N}(u, u_h, \psi)| \leq C\Big[\|e\|_{L_4(\Omega)} + \|\boldsymbol{\theta}_h(e)\|_{L_4(\Omega)}\Big]\Big[\|e\|_{L_2(\Omega)} + \|\boldsymbol{\theta}_h(e)\|_{L_2(\Omega)}\Big]\|\psi\|_{W_4^1(\Omega, \mathcal{T}_h)}$$

$$\leq C\Big[\|e\|_{W_4^1(\Omega, \mathcal{T}_h)} + \|r([\![e]\!])\|_{L_4(\Omega)}\Big]\Big[\|e\|_{W_2^1(\Omega, \mathcal{T}_h)} + \|r([\![e]\!])\|_{L_2(\Omega)}\Big]\|\psi\|_{W_4^1(\Omega, \mathcal{T}_h)}.$$

From the embedding theorem, the last term is bounded since we know $H^2(\Omega) \subset W_4^1(\kappa)$, i.e., $\|\psi\|_{W_4^1(\Omega, \mathcal{T}_h)} \leq C\|\psi\|_{H^2(\Omega, \mathcal{T}_h)}$ by a uniform constant C. Using (3.24) gives, since $u - u_h \in V(h)$

$$(8.9) \qquad \|e\|_{W_2^1(\Omega, \mathcal{T}_h)} + \|r([\![e]\!])\|_{L_2(\Omega)} \leq C\|u - u_h\|_h.$$

Now, the only term to handle (to obtain an additional order of h) is $\|u - u_h\|_{W_4^1(\Omega, \mathcal{T}_h)}$ and $\|r([\![e]\!])\|_{L_4(\Omega)}$. By writing $u - u_h = u - \pi_h u + \pi_h u - u_h = \eta + \xi$, one gets

$$\|u - u_h\|_{W_4^1(\Omega, \mathcal{T}_h)} + \|r([\![e]\!])\|_{L_4(\Omega)} \leq \|\eta\|_{W_4^1(\Omega, \mathcal{T}_h)} + \|\xi\|_{W_4^1(\Omega, \mathcal{T}_h)} + \|r([\![e]\!])\|_{L_4(\Omega)}.$$

In case of $q \geq 2$ and $u \in H^{5/2}(\Omega)$ (note that $q + 1 > 5/2$), application of Lemma 3.3 gives

$$\|\eta\|_{W_4^1(\Omega, \mathcal{T}_h)} \leq C\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^4 \|u\|_{H^{5/2}(\kappa)}^4\Big)^{1/4} \leq Ch\Big(\sum_{\kappa \in \mathcal{T}_h} \|u\|_{H^{5/2}(\kappa)}^2\Big)^{1/2} \leq Ch\|u\|_{H^{5/2}(\Omega)}.$$

On the other hand using the inverse inequality (3.25) one has

$$\|\xi\|_{W_4^1(\Omega, \mathcal{T}_h)} \leq \Big(\sum_{\kappa \in \mathcal{T}_h} C_{\text{inv}}^2 h_\kappa^{-1} \|\xi\|_{W_2^1(\kappa)}^2\Big)^{1/2}.$$

Then, application to (8.4) for term $\|\xi\|_{W_2^1(\kappa)}$ with $\mu = \min\{s, q+1\}$ and $\mu_* = \min\{s_*, q+1\}$, and quasi-uniformity condition (3.4) yield

$$\|\xi\|_{W_4^1(\Omega, \mathcal{T}_h)} \leq C\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)-1} \|u\|_{H^s(\kappa)}^2 + \sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*-1} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2\Big)^{1/2}$$

$$(8.10) \qquad \leq Ch^{\mu-3/2}\|u\|_{H^s(\Omega, \mathcal{T}_h)} + Ch^{\mu_*-1/2}\|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\Omega, \mathcal{T}_h)}.$$

Hence, $\|\xi\|_{W_4^1(\Omega, \mathcal{T}_h)} \leq Ch$ for $u \in H^{5/2}(\Omega, \mathcal{T}_h)$ and $q \geq 2$, provided that $\boldsymbol{a}(u, \nabla u) \in H^{3/2}(\Omega, \mathcal{T}_h)$. Similarly, since $r([\![e]\!]) \in \Sigma_{h,p}$ one can employ (3.25) to write

$$\|r([\![e]\!])\|_{L_4(\Omega)} \leq \Big(\sum_{\kappa \in \mathcal{T}_h} C_{\text{inv}}^2 h_\kappa^{-1} \|r([\![e]\!])\|_{L_2(\kappa)}^2\Big)^{1/2}.$$

Using (3.16), (3.4), Corollary 8.1, and with similar arguments as (8.10) yield

$$(8.11) \qquad \|r([\![e]\!])\|_{L_4(\Omega)} \leq Ch^{\mu-3/2}\|u\|_{H^s(\Omega, \mathcal{T}_h)} + Ch^{\mu_*-1/2}\|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\Omega, \mathcal{T}_h)} \leq Ch.$$

Combining all terms we have

$$(8.12) \qquad |\mathcal{N}(u, u_h, \psi)| \leq Ch\|u - u_h\|_h\|\psi\|_{H^2(\Omega)}$$

(ii) Using Lemma 5.1 and noticing that $\mathcal{E}_p(u, \psi) = 0$ (see (5.1) for smooth ψ as well as the boundary condition of (5.5)), one has the following upper bound for $\mathcal{E}_p(u, \psi_h)$

$$(8.13) \quad |\mathcal{E}_p(u, \psi_h)| = |\mathcal{E}_p(u, \psi - \psi_h)| \leq C_{con}\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)^{1/2} \|\psi - \psi_h\|_h.$$

The approximation result of Lemma 3.3 and H^2-regularity of ψ give

$$(8.14) \qquad |\mathcal{E}_p(u, \psi_h)| \leq Ch\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)^{1/2} \|\psi\|_{H^2(\Omega)}.$$

(iii) For the adjoint consistency error $\mathcal{E}_d(u, e, \psi)$, from Lemma 5.2 one has

$$|\mathcal{E}_d(u, e, \psi)| \leq \tilde{C}_{con} h \|\boldsymbol{a}_z(u, \nabla u)\|_{W_\infty^1(\Omega)} \|\psi\|_{H^2(\Omega)} \|u - u_h\|_h,$$

Combining steps (i)-(iii), with the energy error estimate in Corollary 8.1, one can write

$$|\mathcal{N}(u, u_h, \psi)| + |\mathcal{E}_p(u, \psi_h)| + |\mathcal{E}_d(u, e, \psi)| \leq Ch \|\psi\|_{H^2(\Omega)} \Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)} \|u\|_{H^s(\kappa)}^2 + h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)^{1/2}$$ ∎

The remaining term in (8.7) can be bounded using Lipschitz continuity (2.7) and Lemma 3.4

$$(8.15) \quad \mathcal{B}(u_h, \psi - \psi_h) - \mathcal{B}(u, \psi - \psi_h) \leq C_{LC} \|u_h - u\|_h \|\psi - \psi_h\|_h \leq C_{LC} C_A' h \|u_h - u\|_h \|\psi\|_{H^2(\Omega)}.$$

Application to the elliptic regularity of ψ and Corollary 8.1 gives

$$\|u_h - u\|_{L_2(\Omega)} \leq Ch\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu-1)} \|u\|_{H^s(\kappa)}^2 + h_\kappa^{2\mu_*} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)^{1/2}$$

which can be summarized as the following lemma for the optimal L_2 error estimate

LEMMA 8.2. *Assume that u_h and u are the solutions of (4.16) and the exact solution of (2.1), respectively. Also assume that $u \in H^s(\Omega, \mathcal{T}_h)$ and $\boldsymbol{a}(u, \nabla u) \in H^{s_*}(\Omega, \mathcal{T}_h)$ with $s \geq \frac{5}{2}$ and $s_* \geq \frac{3}{2}$. Then, there exists $C_{err}' > 0$ such that the following holds*

$$(8.16) \qquad \|u - u_h\|_{L_2(\Omega)} \leq C_{err}'\Big(\sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2\mu} \|u\|_{H^s(\kappa)}^2 + \sum_{\kappa \in \mathcal{T}_h} h_\kappa^{2(\mu_*+1)} \|\boldsymbol{a}(u, \nabla u)\|_{H^{s_*}(\kappa)}^2 \Big)^{1/2},$$

where $\mu = \min(s, q+1)$ and $\mu_* = \min(s_*, q+1)$ and $q \geq 2$.

9. Conclusion. In this work, we have analyzed different DG formulations of a quasilinear elliptic problem by introducing appropriate numerical flux functions inspired by their original version in linear problems. We showed that in spite of the fact that all of these formulations are inconsistent, they have the asymptotic consistency property for both primal and dual problem. Moreover, we also proved the stability of the solution in L_2 norm under mild assumptions on the problem.

Furthermore, for BR2 and SIPG discretizations, we proved the existence and uniqueness of the discrete solution in case of monotone and globally Lipschitz diffusion operator. Afterwards, under regularity assumptions for the exact solution, we proved the optimal convergence rate in energy norm as well as L_2 norm.

REFERENCES

[1] Assyr Abdulle and Gilles Vilmart, *A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems*, Numerische Mathematik, 121 (2012), pp. 397–431.

[2] Douglas N Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM journal on numerical analysis, 19 (1982), pp. 742–760.

[3] Douglas N Arnold, Franco Brezzi, Bernardo Cockburn, and L Donatella Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM journal on numerical analysis, 39 (2002), pp. 1749–1779.

[4] Ivo Babuška and Miloš Zlámal, *Nonconforming elements in the finite element method with penalty*, SIAM Journal on Numerical Analysis, 10 (1973), pp. 863–875.

[5] John W Barrett and WB Liu, *Quasi-norm error bounds for the finite element approximation of a non-newtonian flow*, Numerische Mathematik, 68 (1994), pp. 437–456.

[6] Francesco Bassi, Andrea Crivellini, Stefano Rebay, and Marco Savini, *Discontinuous Galerkin solution of the reynolds-averaged navier–stokes and k–ω turbulence model equations*, Computers & Fluids, 34 (2005), pp. 507–540.

[7] Francesco Bassi and Stefano Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible navier–stokes equations*, Journal of computational physics, 131 (1997), pp. 267–279.

[8] F Bassi, S Rebay, G Mariotti, S Pedinotti, and M Savini, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, in Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics, Technologisch Instituut, Antwerpen, Belgium, 1997, pp. 99–109.

[9] Carlos Erik Baumann and J Tinsley Oden, *A discontinuous hp finite element method for convection—diffusion problems*, Computer Methods in Applied Mechanics and Engineering, 175 (1999), pp. 311–341.

[10] Franco Brezzi, Gianmarco Manzini, Donatella Marini, Paola Pietra, and Alessandro Russo, *Discontinuous Galerkin approximations for elliptic problems*.

[11] Rommel Bustinza and Gabriel N Gatica, *A local discontinuous Galerkin method for nonlinear diffusion problems with mixed boundary conditions*, SIAM Journal on Scientific Computing, 26 (2004), pp. 152–177.

[12] ——, *A mixed local discontinuous galerkin method for a class of nonlinear problems in fluid mechanics*, Journal of Computational Physics, 207 (2005), pp. 427–456.

[13] Philippe G Ciarlet, *The finite element method for elliptic problems*, Elsevier, 1978.

[14] Bernardo Cockburn and Clint Dawson, *Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions*, (1999).

[15] Bernardo Cockburn, George E Karniadakis, and Chi-Wang Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods, Springer, 2000, pp. 3–50.

[16] Bernardo Cockburn and Chi-Wang Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM Journal on Numerical Analysis, 35 (1998), pp. 2440–2463.

[17] Vít Dolejší, *Analysis and application of the iipg method to quasilinear nonstationary convection–diffusion problems*, Journal of Computational and Applied Mathematics, 222 (2008), pp. 251–273.

[18] Yekaterina Epshteyn and Béatrice Rivière, *Estimation of penalty parameters for symmetric interior penalty galerkin methods*, Journal of Computational and Applied Mathematics, 206 (2007), pp. 843–872.

[19] Miloslav Feistauer and Alexander Ženíšek, *Finite element solution of nonlinear elliptic problems*, Numerische Mathematik, 50 (1986), pp. 451–475.

[20] Gabriel N Gatica, *Solvability and Galerkin approximations of a class of nonlinear operator equations*, Zeitschrift fur Analysis und ihre Anwendungen, 21 (2002), pp. 761–782.

[21] Jay Gopalakrishnan and Guido Kanschat, *A multilevel discontinuous Galerkin method*, Numerische Mathematik, 95 (2003), pp. 527–550.

[22] Thirupathi Gudi, Neela Nataraj, and Amiya Pani, *An hp-local discontinuous Galerkin method for some quasilinear elliptic boundary value problems of nonmonotone type*, Mathematics of Computation, 77 (2008), pp. 731–756.

[23] Thirupathi Gudi, Neela Nataraj, and Amiya K Pani, *hp-discontinuous Galerkin methods for strongly nonlinear elliptic boundary value problems*, Numerische Mathematik, 109 (2008), pp. 233–268.

[24] Thirupathi Gudi and Amiya K Pani, *Discontinuous Galerkin methods for quasi-linear elliptic problems of nonmonotone type*, SIAM Journal on Numerical Analysis, 45 (2007), pp. 163–192.

[25] Wolfgang Hackbusch, Regine Fadiman, and Patrick D. F. Ion, *Elliptic differential equations : theory and*

numerical treatment, Springer series in computational mathematics, Springer, Berlin, New York, Paris, 1992.

[26] RALF HARTMANN, *Adjoint consistency analysis of discontinuous Galerkin discretizations*, SIAM Journal on Numerical Analysis, 45 (2007), pp. 2671–2696.

[27] RALF HARTMANN AND PAUL HOUSTON, *Symmetric interior penalty dg methods for the compressible navier-stokes equations i: Method formulation*, (2005).

[28] PAUL HOUSTON, JANICE ROBSON, AND ENDRE SÜLI, *Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems i: The scalar case*, IMA journal of numerical analysis, 25 (2005), pp. 726–749.

[29] TODD DUPONT JIM DOUGLAS, *A Galerkin method for a nonlinear dirichlet problem*, Mathematics of Computation, 29 (1975), pp. 689–696.

[30] JAMES CHING-CHIEH LU, *An a posteriori error control framework for adaptive precision optimization using discontinuous Galerkin finite element method*, PhD thesis, Massachusetts Institute of Technology, 2005.

[31] SANDRA MAY, *Spacetime discontinuous Galerkin methods for convection-diffusion equations*, Bulletin of the Brazilian Mathematical Society, New Series, 47 (2016), pp. 561–573.

[32] JINDŘICH NEČAS, *Introduction to the theory of nonlinear elliptic equations*, vol. 52, Teubner, 1983.

[33] TODD A OLIVER AND DAVID L DARMOFAL, *Analysis of dual consistency for discontinuous Galerkin discretizations of source terms*, SIAM Journal on Numerical Analysis, 47 (2009), pp. 3507–3525.

[34] CHRISTOPH ORTNER AND ENDRE SÜLI, *Discontinuous Galerkin finite element approximation of nonlinear second-order elliptic and hyperbolic systems*, SIAM Journal on Numerical Analysis, 45 (2007), pp. 1370–1397.

[35] ILARIA PERUGIA AND DOMINIK SCHÖTZAU, *An hp-analysis of the local discontinuous Galerkin method for diffusion problems*, Journal of Scientific Computing, 17 (2002), pp. 561–571.

[36] NILES A PIERCE AND MICHAEL B GILES, *Adjoint recovery of superconvergent functionals from pde approximations*, SIAM review, 42 (2000), pp. 247–264.

[37] BÉATRICE RIVIÈRE, MARY F WHEELER, AND VIVETTE GIRAULT, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM Journal on Numerical Analysis, 39 (2001), pp. 902–931.

[38] TORGEIR RUSTEN, PANAYOT VASSILEVSKI, AND RAGNAR WINTHER, *Interior penalty preconditioners for mixed finite element approximations of elliptic problems*, Mathematics of Computation of the American Mathematical Society, 65 (1996), pp. 447–466.

[39] KHOSRO SHAHBAZI, *An explicit expression for the penalty parameter of the interior penalty method*, Journal of Computational Physics, 205 (2005), pp. 401–407.

[40] T WARBURTON AND JAN S HESTHAVEN, *On the constants in hp-finite element trace inverse inequalities*, Computer methods in applied mechanics and engineering, 192 (2003), pp. 2765–2773.

[41] MARY FANETT WHEELER, *An elliptic collocation-finite element method with interior penalties*, SIAM Journal on Numerical Analysis, 15 (1978), pp. 152–161.

[42] SANGITA YADAV, AMIYA PANI, AND EUN-JAE PARK, *Superconvergent discontinuous Galerkin methods for nonlinear elliptic equations*, Mathematics of Computation, 82 (2013), pp. 1297–1335.

[43] MOHAMMAD ZAKERZADEH AND GEORG MAY, *Entropy stable discontinuous Galerkin scheme for the compressible Navier–Stokes equations*, in 55th AIAA Aerospace Sciences Meeting, 2017, p. 0084.

[44] EBERHARD ZEIDLER, *Nonlinear Functional Analysis and Its Applications: III: Variational Methods and Optimization*, Springer Science & Business Media, 2013.

[45] ALEXANDER ŽENÍŠEK, *Nonlinear elliptic and evolution problems and their finite element approximations*, Academic Pr, 1990.