Research Article

Thomas Führer*, Norbert Heuer and Michael Karkulik

Analysis of Backward Euler Primal DPG Methods

<https://doi.org/10.1515/cmam-2021-0056>

Received March 23, 2021; revised May 27, 2021; accepted June 20, 2021

Abstract: We analyze backward Euler time stepping schemes for a primal DPG formulation of a class of parabolic problems. Optimal error estimates are shown in a natural norm and in the L^2 norm of the field variable. For the heat equation the solution of our primal DPG formulation equals the solution of a standard Galerkin scheme and, thus, optimal error bounds are found in the literature. In the presence of advection and reaction terms, however, the latter identity is not valid anymore and the analysis of optimal error bounds requires to resort to elliptic projection operators. It is essential that these operators be projections with respect to the spatial part of the PDE, as in standard Galerkin schemes, and not with respect to the full PDE at a time step, as done previously.

Keywords: Primal DPG, Time-Stepping, Heat Equation, Parabolic Problems, A Priori Analysis

MSC 2010: 65N30, 65N12

1 Introduction

In this work we analyze a backward Euler primal DPG time stepping scheme for the parabolic problem

$$\dot{u} - \operatorname{div} \mathbf{A} \nabla u + \boldsymbol{\beta} \cdot \nabla u + \gamma u = f \quad \text{in } (0, T) \times \Omega, \quad (1.1a)$$

$$u = 0 \quad \text{in } (0, T) \times \partial\Omega, \quad (1.1b)$$

$$u(0, \cdot) = u_0 \quad \text{in } \Omega. \quad (1.1c)$$

Here, $\Omega \subset \mathbb{R}^d$ ($d \in 2, 3$) denotes a bounded Lipschitz domain. We make the standing assumptions that the coefficients $\mathbf{A}$, $\boldsymbol{\beta}$, and γ are time-independent, that they are essentially bounded in space, that $\mathbf{A}$ is a symmetric uniformly positive matrix almost everywhere in Ω , and that $(\boldsymbol{\beta} \cdot \nabla v + \gamma v, v) \geq 0$ for all $v \in H_0^1(\Omega)$. The latter assumption plays a crucial role in our analysis (see Lemma 2 below). The $L^2(\Omega)$ inner product is denoted with $(\cdot, \cdot)$. We assume that the data satisfy $f(t, \cdot) \in L^2(\Omega)$ for all $t \in [0, T]$ and $u_0 \in L^2(\Omega)$.

The discontinuous Petrov–Galerkin method with optimal test functions (DPG) pertains to the class of minimum residual methods and was introduced in a series of papers [5, 6, 10]. It has successfully been applied to elliptic problems, see, e.g., [7, 8] for the Poisson problem and [16, 18] for fourth-order problems. Through the use of optimal test functions, the discrete problem inherits the stability of the continuous problem. This comes in advantageous for problems where robustness is one of the main challenges, e.g., singularly perturbed problems, see [24] for reaction-dominated diffusion problems and [3, 11] for the convection-dominated case. Space-time DPG methods have been studied previously, see, e.g., [9, 12, 13].

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

Norbert Heuer, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Santiago, Chile, e-mail: nheuer@mat.uc.cl. <https://orcid.org/0000-0002-5171-1457>

Michael Karkulik, Departamento de Matemática, Universidad Técnica Federico Santa María, Valparaíso, Chile, e-mail: michael.karkulik@usm.cl

For other space-time minimum residual methods we refer to [1, 21, 28, 30]. Approaches employing the DPG methodology for the time discretization of parabolic and hyperbolic initial value problems have recently been investigated, cf. [25, 26]. On the other hand, time-stepping methods for ODEs are frequently employed in combination with standard Galerkin finite element methods in space, cf. the monograph [29] for parabolic equations, but less so with DPG methods. To the best of our knowledge there exist only two works in this direction, dealing with time-stepping and spatial DPG methods for the heat equation, namely [19] and [27].

In [19], a backward Euler method is used to discretize in time, and then the DPG methodology is applied to the ultraweak variational formulation of the resulting equations. The a priori analysis given there employs the Galerkin projection with respect to these very equations, and hence a spatial discretization error has to be accounted for in every time step. This gives rise to a theoretical error bound of order $\mathcal{O}(\frac{h}{k}) + \mathcal{O}(k)$ for lowest-order discretizations with h and k being the spatial mesh-width and time step, respectively. As numerical experiments from [27] indicate, this asymptotic error bound is not optimal. The authors of [27] study general θ -schemes (including the backward Euler and the Crank–Nicolson time discretization) based on the primal DPG method [8] and the ultraweak DPG method [7], and provide an extensive numerical study and comparison of the different approaches, and it turns out that $\mathcal{O}(h + k)$ is the optimal error bound which can be expected for the method from [19].

Our motivation for the present work is to give a theoretically sound explanation of the optimal convergence rates seen in the numerical experiments from [27] for a backward Euler primal DPG method. A DPG approximation is attractive, e.g., for singularly perturbed problems not considered here, cf. [11, 24]. An application for a primal DPG method can be found in the recent work [23]. We will consider general second order linear elliptic spatial differential operators. For the heat equation ($\mathbf{A}$ is the identity, $\boldsymbol{\beta} = 0$, and $\gamma = 0$) optimal error estimates follow from the fact that the field solution component of the primal DPG method is identical to the solution of a standard Galerkin FEM (cf. Section 2.2.3 for details). Therefore, well-known results for time-stepping Galerkin FEMs, see e.g. [29], apply. However, for the general case $\boldsymbol{\beta} \neq 0$ this is not true anymore and a new analysis needs to be provided. The accumulation of the spatial discretization error in every time step, identified above as the reason for suboptimality of previous theoretical results, is usually avoided in standard FEMs by using the Galerkin projection with respect to the elliptic part of the parabolic equation only. This idea, nowadays being referred to as *elliptic projection operators*, was introduced in [31] and is by now one of the main tools in the analysis of time-stepping FEMs, as witnessed again in [29]. In the present work we prove optimal error estimates in the context of practical primal DPG methods by the use of an elliptic projection operator. We follow some ideas from [20] where an elliptic projection in the analysis of time-stepping first-order system least-squares finite element methods is used in the same spirit.

The remainder of this paper is organized as follows: In Section 2 we introduce the fully discrete method as well as the necessary notation. We prove stability of the method and provide quasi-optimality results for the elliptic projection operator. In Section 3 we use these results to show optimal error estimates in the $H^1(\Omega)$ norm. Section 4 is devoted to optimal error estimates in weaker norms, particularly in $L^2(\Omega)$.

2 Time-Stepping DPG Formulation

2.1 Notation

The notation $a \lesssim b$ means that there exists a constant $C > 0$ that (possibly) depends on Ω , $\mathbf{A}$, $\boldsymbol{\beta}$, γ , and T , but is independent of involved functions. We write $a \approx b$ if $a \lesssim b$ and $b \lesssim a$. Furthermore, $a \gtrsim b$ stands for $b \lesssim a$. If the dependence on T is not (at most) linear, we explicitly state the dependence. Moreover, we use the $\mathcal{O}(\cdot)$ notation, e.g., $g = \mathcal{O}(h^p)$ with $p > 0$ means that $|g| \leq h^p$ for $0 < h \rightarrow 0$.

We consider a time discretization $0 = t_0 < \dots < t_N = T$, and for notational simplicity we use a uniform time step size $k = t_n - t_{n-1}$ (but we stress that this is not necessary). By $(\cdot, \cdot)$ and $\|\cdot\|$ we denote the inner product and norm in $L^2(\Omega)$. We consider spatial discretizations based on a shape-regular conforming simplicial mesh $\mathcal{T}$ of Ω . To any partition $\mathcal{T}$ we associate its skeleton $\mathcal{S}$ consisting of the boundaries of all elements,

and the trace space

$$H^{-\frac{1}{2}}(\mathcal{S}) := \left\{ \widehat{\sigma} \in \prod_{K \in \mathcal{T}} H^{-\frac{1}{2}}(\partial K) \mid \text{there exists } \boldsymbol{\sigma} \in \mathbf{H}(\text{div}; \Omega) \text{ such that } \widehat{\sigma}|_{\partial K} = \boldsymbol{\sigma} \cdot \mathbf{n}_K|_{\partial K} \text{ for all } K \in \mathcal{T} \right\},$$

where $\mathbf{n}_K$ denotes the unit outward normal vector on ∂K . This is a Hilbert space with norm

$$\|\widehat{\sigma}\|_{-\frac{1}{2},k} := \inf \{ \|\boldsymbol{\sigma}\|^2 + k \|\text{div } \boldsymbol{\sigma}\|^2 \}^{\frac{1}{2}} \mid \boldsymbol{\sigma} \in \mathbf{H}(\text{div}; \Omega), \widehat{\sigma}|_{\partial K} = \boldsymbol{\sigma} \cdot \mathbf{n}_K|_{\partial K} \text{ for all } K \in \mathcal{T} \}.$$

The trial space of our method will be

$$U := H_0^1(\Omega) \times H^{-\frac{1}{2}}(\mathcal{S})$$

equipped with the norm

$$\|(u, \widehat{\sigma})\|_{U,k}^2 := \|\nabla u\|^2 + \|\widehat{\sigma}\|_{-\frac{1}{2},k}^2,$$

and the test space will be

$$V := H^1(\mathcal{T}) := \prod_{K \in \mathcal{T}} H^1(K)$$

equipped with the norm

$$\|v\|_{V,k}^2 := \frac{1}{k} \|v\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2.$$

Here, $\nabla_{\mathcal{T}} v$ is the $\mathcal{T}$ -piecewise gradient. For functions $\widehat{\sigma} \in H^{-\frac{1}{2}}(\mathcal{S})$ we define

$$\langle \widehat{\sigma}, v \rangle_{\mathcal{S}} := (\text{div } \boldsymbol{\sigma}, v) + (\boldsymbol{\sigma}, \nabla_{\mathcal{T}} v) \quad \text{for all } v \in H^1(\mathcal{T}),$$

where $\boldsymbol{\sigma} \in \mathbf{H}(\text{div}; \Omega)$ with $\boldsymbol{\sigma} \cdot \mathbf{n}_K|_{\partial K} = \widehat{\sigma}|_{\partial K}$ for all $K \in \mathcal{T}$. We note that the definition is independent of the choice of $\boldsymbol{\sigma}$ and recall the following result.

Lemma 1. *One has*

$$\|\widehat{\sigma}\|_{-\frac{1}{2},k} \approx \sup_{0 \neq v \in V} \frac{\langle \widehat{\sigma}, v \rangle_{\mathcal{S}}}{\|v\|_{V,k}}.$$

Proof. The proof follows by now well-established arguments, see [4] or [17] for problems where the norms depend on parameters. Note that if we use $\|\mathbf{A}^{-\frac{1}{2}} \boldsymbol{\sigma}\|$ instead of $\|\boldsymbol{\sigma}\|$ in the definition of $\|\cdot\|_{-\frac{1}{2},k}$, then equality holds. $\square$

Throughout this work we will use that $\langle \widehat{\sigma}, v \rangle_{\mathcal{S}} = 0$ for $v \in H_0^1(\Omega)$ which follows from the definition of this duality and integration by parts, i.e.,

$$\langle \widehat{\sigma}, v \rangle_{\mathcal{S}} = (\text{div } \boldsymbol{\sigma}, v) + (\boldsymbol{\sigma}, \nabla_{\mathcal{T}} v) = (\text{div } \boldsymbol{\sigma}, v) + (\boldsymbol{\sigma}, \nabla v) = \langle \boldsymbol{\sigma} \cdot \mathbf{n}_{\partial \Omega}, v \rangle_{\partial \Omega} = 0$$

for all $\widehat{\sigma} \in H^{-\frac{1}{2}}(\mathcal{S})$, $v \in H_0^1(\Omega)$, where $\boldsymbol{\sigma} \in \mathbf{H}(\text{div}; \Omega)$ is an extension of the trace $\widehat{\sigma}$.

2.2 Primal DPG Formulation

For a simpler notation, we use superindices to indicate time evaluations, i.e., for a time-dependent function $v = v(t; \cdot)$ we use the notation $v^n(\cdot) := v(t_n; \cdot)$. Moreover, we use similar notations for discrete quantities which however are not defined for all $t \in [0, T]$. For instance, later on u_h^n will denote an approximation of $u^n = u(t_n; \cdot)$, where u is the exact solution of the parabolic equation. We approximate $\dot{u}^n$ by a backward difference, i.e., $\dot{u}^n \approx (u^n - u^{n-1})/k$, this formally yields the elliptic PDE

$$\frac{1}{k} u^n - \text{div } \mathbf{A} \nabla u^n + \boldsymbol{\beta} \cdot \nabla u^n + \gamma u^n = f^n + \frac{1}{k} u^{n-1},$$

which admits a unique solution $u^n \in H_0^1(\Omega)$. By testing with $v \in V$ and introducing $\widehat{\sigma}^n$ as

$$\widehat{\sigma}^n|_{\partial K} := \mathbf{A} \nabla u^n \cdot \mathbf{n}_K|_{\partial K} \quad \text{for all } K \in \mathcal{T},$$

and integrating by parts one obtains the primal DPG formulation

$$\frac{1}{k} (u^n, v) + (\mathbf{A} \nabla u^n, \nabla_{\mathcal{T}} v) + (\boldsymbol{\beta} \cdot \nabla u^n, v) + (\gamma u^n, v) - \langle \widehat{\sigma}^n, v \rangle_{\mathcal{S}} = (f^n, v) + \frac{1}{k} (u^{n-1}, v).$$

We introduce some bilinear forms and the right-hand side functional: For $\mathbf{U} = (u, \widehat{\sigma}) \in U$, $v \in V$ and $w, g \in L^2(\Omega)$ set

$$\begin{aligned} b(\mathbf{U}, v) &:= (\mathbf{A} \nabla u, \nabla_{\mathcal{T}} v) + (\boldsymbol{\beta} \cdot \nabla u, v) + (\gamma u, v) - \langle \widehat{\sigma}, v \rangle_S, \\ a(\mathbf{U}, v) &:= \frac{1}{k}(u, v) + b(\mathbf{U}, v), \\ F(g, w; v) &:= (g, v) + \frac{1}{k}(w, v). \end{aligned}$$

Therefore, our formulation simply reads

$$a(\mathbf{U}^n, v) = F(f^n, u^{n-1}; v) \quad \text{for all } v \in V.$$

We note that the bilinear form $a(\cdot, \cdot)$ satisfies a continuous inf-sup condition using the norms $\|\cdot\|_{U,k}$ and $\|\cdot\|_{V,k}$. This can be seen by extending the analysis of standard works like [4] to general elliptic PDEs. We do not pursue this further, as we ultimately want to analyze the practical DPG method. Furthermore, we note that a is not bounded independently of k using the above norms, but rather

$$|a(\mathbf{U}, v)| \leq (k^{-\frac{1}{2}} \|u\| + \|\mathbf{U}\|_{U,k}) \|v\|_{V,k}, \quad (2.1)$$

where $\mathbf{U} = (u, \widehat{\sigma})$.

2.2.1 Fully Discrete Scheme

By $\mathcal{P}^p(K)$ we denote the space of polynomials of degree $p \in \mathbb{N}_0$, and

$$\mathcal{P}^p(\mathcal{T}) := \{v \in L^2(\Omega) : v|_K \in \mathcal{P}^p(K) \text{ for all } K \in \mathcal{T}\}$$

For the discretization of traces we use the space of facewise polynomials denoted by $\mathcal{P}^p(\mathcal{S})$. In particular, we note that $\mathcal{P}^p(\mathcal{S})$ is the (normal-)trace space of the p -th order Raviart–Thomas space. We consider the spaces

$$\begin{aligned} U_h &:= \mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega) \times \mathcal{P}^p(\mathcal{S}), \\ V_h &:= \mathcal{P}^{p+d}(\mathcal{T}). \end{aligned}$$

The (discrete) trial-to-test operator is given by

$$(\Theta_h \mathbf{U}_h, v_h)_{V,k} = a(\mathbf{U}_h, v_h) \quad \text{for all } v_h \in V_h.$$

We recall that the inner product is given by

$$(v, \delta v)_{V,k} = \frac{1}{k}(v, \delta v) + (\mathbf{A} \nabla_{\mathcal{T}} v, \nabla_{\mathcal{T}} \delta v).$$

The fully discrete scheme then reads: Given $u_h^0 \in L^2(\Omega)$, solve

$$a(\mathbf{U}_h^n, \Theta_h \mathbf{W}_h) = F(f^n, u_h^{n-1}; \Theta_h \mathbf{W}_h) \quad \text{for all } \mathbf{W}_h \in U_h, \quad n = 1, 2, \dots \quad (2.2)$$

The next lemma establishes coercivity of $a(\cdot, \Theta_h \cdot)$ on U_h .

Lemma 2. *Let $\mathbf{U}_h = (u_h, \widehat{\sigma}_h) \in U_h$. Then*

$$\frac{1}{k} \|u_h\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \leq \|\Theta_h \mathbf{U}_h\|_{V,k}^2.$$

If, additionally, $hk^{-\frac{1}{2}} \leq C_1$ for some constant $C_1 > 0$, then

$$C_2 \|\widehat{\sigma}_h\|_{-\frac{1}{2},k} \leq C_3 \|\Theta_h(0, \widehat{\sigma}_h)\|_{V,k} \leq \|\Theta_h \mathbf{U}_h\|_{V,k}.$$

Here, $C_2, C_3 > 0$ denote generic constants independent of h, k and $\mathbf{U}_h$.

Proof. For the first part we use that $\mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega) \subset V_h$ and $(\beta \cdot \nabla w + \gamma w, w) \geq 0$ for $w \in H_0^1(\Omega)$: Let $\mathbf{U}_h = (u_h, \hat{\sigma}_h) \in U_h$ be given. With the latter observation, the definition of the optimal test function and the fact that $\langle \hat{\sigma}_h, w \rangle_S = 0$ for $w \in H_0^1(\Omega)$ we get that

$$\begin{aligned} \frac{1}{k} \|u_h\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 &\leq \frac{1}{k} (u_h, u_h) + (\mathbf{A} \nabla u_h, \nabla u_h) + (\beta \cdot \nabla u_h, u_h) + (\gamma u_h, u_h) \\ &= a((u_h, 0), u_h) = (\Theta_h \mathbf{U}_h, u_h)_{V,k} \leq \|\Theta_h \mathbf{U}_h\|_{V,k} \|u_h\|_{V,k} \\ &= \|\Theta_h \mathbf{U}_h\|_{V,k} \left(\frac{1}{k} \|u_h\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

This finishes the proof of the first estimate.

For the second one we use some results established in the literature on fully discrete DPG formulations (practical DPG), e.g., [22]. Let $\Pi_h: V \rightarrow V_h$ be the Fortin operator defined in [22, Lemma 3.2] which in [22] is denoted by Π_r^{grad} with $r = p + d$. By Lemma 1 and the Fortin property we have that

$$\|\hat{\sigma}_h\|_{-\frac{1}{2},k} \approx \sup_{0 \neq v \in H^1(\mathcal{T})} \frac{\langle \hat{\sigma}_h, v \rangle_S}{\|v\|_{V,k}} = \sup_{0 \neq v \in H^1(\mathcal{T})} \frac{\langle \hat{\sigma}_h, \Pi_h v \rangle_S}{\|v\|_{V,k}}. \quad (2.3)$$

Moreover, from [22, proof of Lemma 3.2] we infer that

$$k^{-\frac{1}{2}} \|\Pi_h v\| + \|\nabla \Pi_h v\| \leq k^{-\frac{1}{2}} \|v\| + (1 + hk^{-\frac{1}{2}}) \|\nabla v\|.$$

Thus, using the assumption $hk^{-\frac{1}{2}} \leq 1$, we see that

$$\|\Pi_h v\|_{V,k} \leq \|v\|_{V,k} \quad \text{for all } v \in V.$$

Combining this with (2.3) we conclude that

$$\begin{aligned} \|\hat{\sigma}_h\|_{-\frac{1}{2},k} &\leq \sup_{0 \neq v \in H^1(\mathcal{T})} \frac{\langle \hat{\sigma}_h, \Pi_h v \rangle_S}{\|v\|_{V,k}} \leq \sup_{0 \neq v \in H^1(\mathcal{T})} \frac{\langle \hat{\sigma}_h, \Pi_h v \rangle_S}{\|\Pi_h v\|_{V,k}} \\ &\leq \sup_{0 \neq v_h \in V_h} \frac{\langle \hat{\sigma}_h, v_h \rangle_S}{\|v_h\|_{V,k}} = \|\Theta_h(0, \hat{\sigma}_h)\|_{V,k}. \end{aligned}$$

Then $\|\Theta_h(0, \hat{\sigma}_h)\|_{V,k} \leq \|\Theta_h(u_h, \hat{\sigma}_h)\|_{V,k} + \|\Theta_h(u_h, 0)\|_{V,k}$. It remains to estimate $\|\Theta_h(u_h, 0)\|_{V,k}$: We use the notation $\Theta_h(u_h, 0) =: v_h$,

$$\begin{aligned} \|\Theta_h(u_h, 0)\|_{V,k}^2 &= a((u_h, 0), \Theta_h(u_h, 0)) = \frac{1}{k} (u_h, v_h) + (\mathbf{A} \nabla u_h, \nabla v_h) + (\beta \cdot \nabla u_h + \gamma u_h, v_h) \\ &\leq \left(\frac{1}{k} \|u_h\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \right)^{\frac{1}{2}} \|v_h\|_{V,k} \leq \|\Theta_h \mathbf{U}_h\|_{V,k} \|\Theta_h(u_h, 0)\|_{V,k}. \end{aligned}$$

Note that the involved constant only depends on the end time T and the coefficients, but is otherwise independent of k or h . To see the possible dependence on T we note that the triangle and Poincaré inequalities as well as boundedness of the coefficients imply that

$$(\beta \cdot \nabla u_h + \gamma u_h, v_h) \leq \|\nabla u_h\| \|v_h\| = k^{\frac{1}{2}} \|\nabla u_h\| k^{-\frac{1}{2}} \|v_h\| \leq T^{\frac{1}{2}} \|\nabla u_h\| \|v_h\|_{V,k}.$$

This finishes the proof. $\square$

It should be noted that the condition $hk^{-\frac{1}{2}} \leq 1$ is only needed to get a k independent estimate for the trace norm. If this condition is not satisfied then the constant depends on k .

Theorem 3. Problem (2.2) is well posed. In particular, the solutions are stable in the sense that

$$\|u_h^n\| \leq (\|u_h^n\|^2 + k \|\mathbf{A}^{\frac{1}{2}} \nabla u_h^n\|^2)^{\frac{1}{2}} \leq \sum_{j=1}^n k \|f^j\| + \|u_h^0\|.$$

Proof. According to Lemma 2, Problem (2.2) admits unique solutions $\mathbf{U}_h^n \in U_h$, $n = 1, 2, \dots$. Then

$$\begin{aligned} \|\Theta_h \mathbf{U}_h^n\|_{V,k}^2 &= a(\mathbf{U}_h^n, \Theta_h \mathbf{U}_h^n) = F(f^n, u_h^{n-1}; \Theta_h \mathbf{U}_h^n) = (f^n + k^{-1} u_h^{n-1}, \Theta_h \mathbf{U}_h^n) \\ &\leq k^{\frac{1}{2}} \|f^n\| + k^{-1} u_h^{n-1} \|k^{-\frac{1}{2}} \|\Theta_h \mathbf{U}_h^n\| \leq k^{\frac{1}{2}} \|f^n\| + k^{-1} u_h^{n-1} \|\Theta_h \mathbf{U}_h^n\|_{V,k}. \end{aligned}$$

This together with the estimate from Lemma 2 shows that

$$\left(\|u_h^n\|^2 + k\|\mathbf{A}^{\frac{1}{2}} \nabla u_h^n\|^2 \right)^{\frac{1}{2}} \leq k^{\frac{1}{2}} \|\Theta_h \mathbf{U}_h^n\|_{V,k} \leq k\|f^n\| + \|u_h^{n-1}\|.$$

Iterating the arguments concludes the proof. $\square$

2.2.2 Remark: Known Results for a Bubnov–Galerkin Method for the Heat Equation

We recall results from [29, Section 1] for a standard Bubnov–Galerkin backward Euler method: Given an initial solution $u_h^0 \in L^2(\Omega)$, find $u_h^n \in \mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega)$ ($n = 1, 2, \dots$) such that

$$\frac{1}{k}(u_h^n, v_h) + (\nabla u_h^n, \nabla v_h) = (f^n, v_h) + \frac{1}{k}(u_h^{n-1}, v_h) \quad \text{for all } v_h \in \mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega). \quad (2.4)$$

Proposition 4 ([29, p. 16 and Theorem 1.5]). *Let u denote the solution of (1.1) and let $u_h^n \in \mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega)$ denote the solution of (2.4). Suppose that $u_0|_{\partial\Omega} = 0$. Then*

$$\|\nabla(u^n - u_h^n)\| \leq C_1 h^{p+1} \left(\|u_0\|_{H^{p+2}(\Omega)} + \int_0^{t_n} \|\dot{u}(t)\|_{H^{p+2}(\Omega)} dt \right) + C_2 k \int_0^{t_n} \|\ddot{u}(t)\| dt.$$

If Ω is convex then,

$$\|u^n - u_h^n\| \leq C_3 h^{p+2} \left(\|u_0\|_{H^{p+2}(\Omega)} + \int_0^{t_n} \|\dot{u}(t)\|_{H^{p+2}(\Omega)} dt \right) + k \int_0^{t_n} \|\ddot{u}(t)\| dt.$$

The constants C_j , $j = 1, 2, 3$, are independent of the solution u and T .

2.2.3 Remark: Primal DPG for the Heat Equation — The Trivial Case

Let us consider the simplest model which is the heat equation where $\mathbf{A}$ is the identity, $\beta = 0$, $\gamma = 0$.

It is straightforward to see that for $(u_h, 0) \in U_h$, $\Theta_h(u_h, 0) = u_h$ since

$$(u_h, v_h)_{V,k} = \frac{1}{k}(u_h, v_h) + (\nabla u_h, \nabla v_h) = a((u_h, 0), v_h) = (\Theta_h(u_h, 0), v_h)_{V,k} \quad \text{for all } v_h \in V_h.$$

For any $v_h \in \mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega)$ consider the test function $\Theta_h(v_h, 0)$ in (2.2). Using the fact that $\langle \hat{\sigma}_h, v_h \rangle_S = 0$, we see that

$$a(\mathbf{U}_h^n, \Theta_h(v_h, 0)) = \frac{1}{k}(u_h^n, v_h) + (\nabla u_h^n, \nabla v_h) = (f^n + k^{-1}u_h^{n-1}, v_h) = F(f^n, u_h^{n-1}; \Theta_h(v_h, 0)).$$

Let $\mathbf{U}_h^n = (u_h^n, \hat{\sigma}_h^n) \in U_h$ denote the solution to (2.2), then the solution component u_h^n satisfies

$$\frac{1}{k}(u_h^n, v_h) + (\nabla u_h^n, \nabla v_h) = (f^n, v_h) + \frac{1}{k}(u_h^{n-1}, v_h) \quad \text{for all } v_h \in \mathcal{P}^{p+1}(\mathcal{T}) \cap H_0^1(\Omega),$$

which is the standard Galerkin FEM, see [29, Section 1]. Thus the primal DPG solution component u_h is identical to the standard Galerkin FEM solution. In particular, optimal error estimates in $L^\infty(L^2)$ and $L^\infty(H_0^1)$, see [29], are valid for the primal DPG solution. In [27] optimal error estimates $L^\infty(H_0^1)$ have been observed in numerical experiments, which can be explained with the observation above.

Note that this remark is true only if we consider $\beta = 0$, $\gamma = 0$.

2.3 Elliptic Projection-Type Operator

To obtain optimal error estimates, we introduce an elliptic projection. The idea goes back to [31] to obtain optimal $L^2(\Omega)$ a priori error estimates and is extensively used for Galerkin methods, but has not been studied

for least-squares methods until only very recently in [20]. For DPG methods, an additional difficulty arises since test norms are mesh dependent. A main difference to Galerkin methods is that, although the elliptic part of the parabolic PDE might be symmetric, our elliptic projection operator always corresponds to a non-symmetric problem. The reason is that we need to use optimal test functions used in (2.2) in combination with the bilinear form $b(\cdot, \cdot)$, which models the elliptic part.

We define the elliptic projection operator $\mathcal{E}_h : U \rightarrow U_h$ by

$$b(\mathcal{E}_h \mathbf{U}, \Theta_h \mathbf{W}_h) = b(\mathbf{U}, \Theta_h \mathbf{W}_h) \quad \text{for all } \mathbf{W}_h \in U_h. \quad (2.5)$$

In the proofs below and in some results we will use the (semi-)norm

$$\|\mathbf{U}\|_k^2 := \|\mathbf{A}^{\frac{1}{2}} \nabla u\|^2 + \|\Theta_h(0, \hat{\sigma})\|_{V,k}^2.$$

Note that from boundedness of $\mathbf{A}$ and the definition of the discrete trial-to-test operator we infer that

$$\begin{aligned} \|\mathbf{U}\|_k &\approx \|\mathbf{A}^{\frac{1}{2}} \nabla u\| + \|\Theta_h(0, \hat{\sigma})\|_{V,k} = \|\mathbf{A}^{\frac{1}{2}} \nabla u\| + \sup_{0 \neq v \in V_h} \frac{a((0, \hat{\sigma}), v)}{\|v\|_{V,k}} \\ &\leq \|\nabla u\| + \sup_{0 \neq v \in V_h} \frac{\langle \hat{\sigma}, v \rangle_S}{\|v\|_{V,k}} \leq \|\nabla u\| + \sup_{0 \neq v \in V} \frac{\langle \hat{\sigma}, v \rangle_S}{\|v\|_{V,k}} \\ &\approx \|\nabla u\| + \|\hat{\sigma}\|_{-\frac{1}{2},k} \approx \|\mathbf{U}\|_{U,k}, \end{aligned}$$

i.e., $\|\mathbf{U}\|_k \lesssim \|\mathbf{U}\|_{U,k}$ for $\mathbf{U} \in U$. Lemma 2 proves

$$\|\hat{\sigma}_h\|_{-\frac{1}{2},k} \lesssim \|\Theta_h(0, \hat{\sigma}_h)\|_{V,k},$$

hence, uniform positive definiteness of $\mathbf{A}$ shows $\|\mathbf{U}_h\|_{U,k} \lesssim \|\mathbf{U}_h\|_k$ for $\mathbf{U}_h \in U_h$ and the involved constants are independent of h, k if $hk^{-\frac{1}{2}} = \mathcal{O}(1)$. In particular, the norm equivalence $\|\mathbf{U}_h\|_k \approx \|\mathbf{U}_h\|_{U,k}$ holds for $\mathbf{U}_h \in U_h$.

The next lemma establishes boundedness and an inf-sup condition of $b(\cdot, \Theta_h \cdot)$. Note that this is not as trivial as it seems since Θ_h is calculated using the bilinear form $a(\cdot, \cdot)$ and the inner product in V . Recall that $a(\cdot, \cdot)$ is not bounded independently of k and that $(\cdot, \cdot)_{V,k}$ includes terms weighted with negative powers of the time step size k .

Lemma 5. *The bilinear form $b(\cdot, \Theta_h \cdot)$ is bounded,*

$$\begin{aligned} |b(\mathbf{U}, \Theta_h \mathbf{W}_h)| &\lesssim \|\mathbf{U}\|_{U,k} \|\mathbf{W}_h\|_{U,k}, \\ |b(\mathbf{U}, \Theta_h \mathbf{W}_h)| &\lesssim \|\mathbf{U}\|_{U,k} \|\mathbf{W}_h\|_k \end{aligned}$$

for all $\mathbf{U} \in U$, $\mathbf{W}_h \in U_h$, and fulfills the inf-sup condition

$$\|\mathbf{U}_h\|_k \lesssim \sup_{0 \neq \mathbf{V}_h \in U_h} \frac{b(\mathbf{U}_h, \Theta_h \mathbf{V}_h)}{\|\mathbf{V}_h\|_k}$$

for all $\mathbf{U}_h \in U_h$. In particular, we conclude that Problem 2.5 admits a unique solution.

Proof. We show only the second boundedness estimate, as the first one follows from $\|\cdot\|_k \lesssim \|\cdot\|_{U,k}$. To that end, let $v := \Theta_h \mathbf{W}_h = \Theta_h(w_h, \hat{\chi}_h)$. We use the splitting $v = w_h + \bar{w}_h$. From the definition of the optimal test function we see that

$$\begin{aligned} &\frac{1}{k}(w_h, \delta v) + (\mathbf{A} \nabla w_h, \nabla_{\mathcal{T}} \delta v) + \frac{1}{k}(\bar{w}_h, \delta v) + (\mathbf{A} \nabla_{\mathcal{T}} \bar{w}_h, \nabla_{\mathcal{T}} \delta v) \\ &= \frac{1}{k}(w_h, \delta v) + (\mathbf{A} \nabla w_h, \nabla_{\mathcal{T}} \delta v) + (\beta \cdot \nabla w_h, \delta v) + (\gamma w_h, \delta v) - \langle \hat{\chi}_h, \delta v \rangle_S \quad \text{for all } \delta v \in V_h. \end{aligned}$$

Therefore the component $\bar{w}_h \in V_h$ satisfies

$$\frac{1}{k}(\bar{w}_h, \delta v) + (\mathbf{A} \nabla_{\mathcal{T}} \bar{w}_h, \nabla_{\mathcal{T}} \delta v) = (\beta \cdot \nabla w_h, \delta v) + (\gamma w_h, \delta v) - \langle \hat{\chi}_h, \delta v \rangle_S \quad \text{for all } \delta v \in V_h.$$

We note that $a((0, \hat{\chi}_h), \delta v) = -\langle \hat{\chi}_h, \delta v \rangle_S$ and therefore the definition of the optimal test function yields

$$|\langle \hat{\chi}_h, \delta v \rangle_S| \leq \|\Theta_h(0, \hat{\chi}_h)\|_{V,k} \|\delta v\|_{V,k}.$$

Standard estimates then show that

$$\|\tilde{w}_h\|_{V,k}^2 \leq k^{\frac{1}{2}} \|\beta \cdot \nabla w_h + \gamma w_h\| k^{-\frac{1}{2}} \|\tilde{w}_h\| + \|\Theta_h(0, \hat{\chi}_h)\|_{V,k} \|\tilde{w}_h\|_{V,k}.$$

Using that $k \leq T$, boundedness of the coefficients and a Poincaré inequality, we infer that

$$\|\tilde{w}_h\|_{V,k} \lesssim k^{\frac{1}{2}} \|\beta \cdot \nabla w_h + \gamma w_h\| + \|\Theta_h(0, \hat{\chi}_h)\|_{V,k} \lesssim \|\nabla w_h\| + \|\Theta_h(0, \hat{\chi}_h)\|_{V,k} \lesssim \|\mathbf{W}_h\|_k. \quad (2.6)$$

Observe that $\langle \hat{\sigma}_h, w_h \rangle_S = 0$ since $w_h \in H_0^1(\Omega)$. Finally, boundedness of the coefficients, a Poincaré inequality and the definition of $\|\cdot\|_k$ give

$$\begin{aligned} |b(\mathbf{U}, v)| &\leq |b(\mathbf{U}, w_h)| + |b(\mathbf{U}, \tilde{w}_h)| \\ &\leq |(\mathbf{A} \nabla u, \nabla w_h) + (\beta \cdot \nabla u + \gamma u, w_h)| + \|\mathbf{U}\|_{U,k} \|\tilde{w}_h\|_{V,k} \\ &\lesssim \|\nabla u\| \|\nabla w_h\| + \|\mathbf{U}\|_{U,k} \|\tilde{w}_h\|_{V,k} \lesssim \|\mathbf{U}\|_{U,k} \|\mathbf{W}_h\|_k. \end{aligned}$$

This shows boundedness. In order to show the inf-sup condition, we first establish the coercivity estimate

$$\|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \leq 2b(\mathbf{U}_h, \Theta_h \mathbf{U}_h). \quad (2.7)$$

Recall that the optimal test function is characterized by

$$\frac{1}{k} (\Theta_h \mathbf{U}_h, \delta v) + (\mathbf{A} \nabla_{\mathcal{T}} \Theta_h \mathbf{U}_h, \nabla_{\mathcal{T}} \delta v) = a(\mathbf{U}_h, \delta v) = \frac{1}{k} (u_h, \delta v) + b(\mathbf{U}_h, \delta v) \quad \text{for all } \delta v \in V_h.$$

Setting $\delta v = v = \Theta_h \mathbf{U}_h$, the last identity together with Young's inequality proves that

$$b(\mathbf{U}_h, v) = \frac{1}{k} \|v\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2 - \frac{1}{k} (u_h, v) \geq \frac{1}{k} \|v\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2 - \frac{1}{2k} \|u_h\|^2 - \frac{1}{2k} \|v\|^2. \quad (2.8)$$

We are going to estimate $k^{-\frac{1}{2}} \|u_h\|$: From Lemma 2 we get that $k^{-1} \|u_h\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \leq k^{-1} \|v\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2$, hence,

$$\|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \leq \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2 + \frac{1}{k} \|v\|^2 - \frac{1}{k} \|u_h\|^2.$$

Combining this with (2.8) yields

$$\begin{aligned} b(\mathbf{U}_h, v) &\geq \frac{1}{k} \|v\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2 - \frac{1}{2k} \|u_h\|^2 - \frac{1}{2k} \|v\|^2 \\ &= \left(\frac{1}{2} \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2 + \frac{1}{2k} \|v\|^2 - \frac{1}{2k} \|u_h\|^2 \right) + \frac{1}{2} \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2 \\ &\geq \frac{1}{2} \|\mathbf{A}^{\frac{1}{2}} \nabla_{\mathcal{T}} v\|^2 + \frac{1}{2} \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \geq \frac{1}{2} \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2. \end{aligned}$$

Now we are in a position to establish the inf-sup condition

$$\|\|\mathbf{U}_h\|\|_k \leq \sup_{0 \neq \mathbf{V}_h \in U_h} \frac{b(\mathbf{U}_h, \Theta_h \mathbf{V}_h)}{\|\|\mathbf{V}_h\|\|_k}.$$

With the boundedness estimates for $b(\cdot, \cdot)$ we have that

$$\begin{aligned} \|\Theta_h(0, \hat{\sigma}_h)\|_{V,k}^2 &= a((0, \hat{\sigma}_h), \Theta_h(0, \hat{\sigma}_h)) = b((0, \hat{\sigma}_h), \Theta_h(0, \hat{\sigma}_h)) \\ &= b(\mathbf{U}_h, \Theta_h(0, \hat{\sigma}_h)) - b((u_h, 0), \Theta_h(0, \hat{\sigma}_h)) \\ &\leq \sup_{0 \neq \mathbf{V}_h \in U_h} \frac{b(\mathbf{U}_h, \Theta_h \mathbf{V}_h)}{\|\|\mathbf{V}_h\|\|_k} \|\Theta_h(0, \hat{\sigma}_h)\|_{V,k} + \|\mathbf{A}^{\frac{1}{2}} \nabla u_h\| \|\Theta_h(0, \hat{\sigma}_h)\|_{V,k}. \end{aligned}$$

Applying the coercivity estimate (2.7) for the second term yields

$$\|\mathbf{A}^{\frac{1}{2}} \nabla u_h\|^2 \leq b(\mathbf{U}_h, \Theta_h \mathbf{U}_h) = \frac{b(\mathbf{U}_h, \Theta_h \mathbf{U}_h)}{\|\|\mathbf{U}_h\|\|_k} \|\|\mathbf{U}_h\|\|_k.$$

Combining the latter two estimates shows that

$$\|\|\mathbf{U}_h\|\|_k^2 \leq \left(\sup_{0 \neq \mathbf{V}_h \in U_h} \frac{b(\mathbf{U}_h, \Theta_h \mathbf{V}_h)}{\|\|\mathbf{V}_h\|\|_k} \right)^2 + \sup_{0 \neq \mathbf{V}_h \in U_h} \frac{b(\mathbf{U}_h, \Theta_h \mathbf{V}_h)}{\|\|\mathbf{V}_h\|\|_k} \|\|\mathbf{U}_h\|\|_k.$$

Young's inequality finishes the proof of the inf-sup condition. $\square$

Lemma 6. Let $\mathcal{E}_h \mathbf{U} := (u_h, \hat{\sigma}_h) \in U_h$ be the solution of Problem (2.5). Then it holds

$$\|\nabla(u - u_h)\| \lesssim \|\mathbf{U} - \mathbf{U}_h\|_k \lesssim \inf_{\mathbf{V}_h \in U_h} \|\mathbf{U} - \mathbf{V}_h\|_{U,k}.$$

If, additionally, $hk^{-\frac{1}{2}} = \mathcal{O}(1)$ (see Lemma 2), then

$$\|\hat{\sigma} - \hat{\sigma}_h\|_{-\frac{1}{2},k} \lesssim \inf_{\mathbf{V}_h \in U_h} \|\mathbf{U} - \mathbf{V}_h\|_{U,k}.$$

Proof. The best approximation properties follow from standard arguments (Babuška's theorem). We give the details only for sake of completeness: Let $\mathbf{W}_h \in U_h$ be arbitrary; then

$$\|\mathbf{U} - \mathbf{U}_h\|_k \leq \|\mathbf{U} - \mathbf{W}_h\|_k + \|\mathbf{U}_h - \mathbf{W}_h\|_k,$$

and using (2.5) and the preceding lemma,

$$\|\mathbf{U}_h - \mathbf{W}_h\|_k \lesssim \sup_{0 \neq \mathbf{V}_h \in U_h} \frac{b(\mathbf{U}_h - \mathbf{W}_h, \Theta_h \mathbf{V}_h)}{\|\mathbf{V}_h\|_k} \leq \sup_{0 \neq \mathbf{V}_h \in U_h} \frac{b(\mathbf{U} - \mathbf{W}_h, \Theta_h \mathbf{V}_h)}{\|\mathbf{V}_h\|_k} \lesssim \|\mathbf{U} - \mathbf{W}_h\|_{U,k}.$$

Since $\|\mathbf{U}\|_k \gtrsim \|\nabla u\|$ this shows the first of the claimed best-approximation estimates. The second follows by noting that $\|\hat{\sigma}_h\|_{-\frac{1}{2},k} \lesssim \|\Theta_h(0, \hat{\sigma}_h)\|_{V,k}$ if $hk^{-\frac{1}{2}} = \mathcal{O}(1)$ (see Lemma 2) and considering

$$\|\hat{\sigma} - \hat{\sigma}_h\|_{-\frac{1}{2},k} \leq \|\mathbf{U} - \mathbf{U}_h\|_{U,k} \leq \|\mathbf{U} - \mathbf{W}_h\|_{U,k} + \|\mathbf{U}_h - \mathbf{W}_h\|_{U,k} \lesssim \|\mathbf{U} - \mathbf{W}_h\|_{U,k} + \|\mathbf{U}_h - \mathbf{W}_h\|_k.$$

The last term is handled as before which concludes the proof. $\square$

Combining the latter quasi-best approximation result with standard approximation properties yields:

Corollary 7. Suppose that the components of $\mathbf{U} = (u, \hat{\sigma})$ are sufficiently smooth. Then

$$\|\mathbf{U} - \mathcal{E}_h \mathbf{U}\|_k = \mathcal{O}(h^{p+1}).$$

If, additionally, $hk^{-\frac{1}{2}} = \mathcal{O}(1)$, then

$$\|\mathbf{U} - \mathcal{E}_h \mathbf{U}\|_{U,k} = \mathcal{O}(h^{p+1}).$$

3 Optimal Error Estimate in Energy Norm

This section is devoted to proving optimal error estimates in the $H^1(\Omega)$ norm of the n -th solution of the backward Euler method (2.2). We make the same assumptions on the regularity of solutions as in [29, Section 1], cf. Proposition 4.

Theorem 8. Let $\mathbf{U}_h^n = (u_h^n, \hat{\sigma}_h^n) \in U_h$ denote the solution of (2.2). Suppose that the components of $\mathbf{U}(t; \cdot)$ are sufficiently regular. Under these assumptions there exists k_0 (independent of h) such that for $k < k_0$ the solution satisfies

$$\|\nabla u^n - \nabla u_h^n\| = \mathcal{O}(h^{p+1} + k) + \mathcal{O}(\|\nabla(u^0 - u_h^0)\|).$$

If, additionally, $hk^{-\frac{1}{2}} = \mathcal{O}(1)$, then

$$\|\hat{\sigma}^n - \hat{\sigma}_h^n\|_{-\frac{1}{2},k} = \mathcal{O}(h^{p+1} + k) + \mathcal{O}(\|\nabla(u^0 - u_h^0)\|).$$

In both estimates, the dependence on T is exponential.

Proof. With the elliptic projection operator $\mathcal{E}_h$ (see (2.5)) we consider the splitting

$$\mathbf{U}^n - \mathbf{U}_h^n = (\mathbf{U}^n - \mathcal{E}_h \mathbf{U}^n) + (\mathcal{E}_h \mathbf{U}^n - \mathbf{U}_h^n).$$

We may also use the norm $\|\cdot\|_k$ defined in Section 2.3.

Step 1. By Corollary 7 we have

$$\|\mathbf{U}^n - \mathcal{E}_h \mathbf{U}^n\|_k = \mathcal{O}(h^{p+1}).$$

Step 2. We derive error equations: First, write $\mathcal{E}_h \mathbf{U}^n = (\mathcal{E}_h^1 \mathbf{U}^n, \mathcal{E}_h^2 \mathbf{U}^n)$. Then, by (2.5),

$$\begin{aligned} a(\mathcal{E}_h \mathbf{U}^n, v_{\text{opt}}) &= \frac{1}{k}(\mathcal{E}_h^1 \mathbf{U}^n, v_{\text{opt}}) + b(\mathcal{E}_h \mathbf{U}^n, v_{\text{opt}}) = \frac{1}{k}(\mathcal{E}_h^1 \mathbf{U}^n, v_{\text{opt}}) + b(\mathbf{U}^n, v_{\text{opt}}) \\ &= (k^{-1} \mathcal{E}_h^1 \mathbf{U}^n + f^n - \dot{u}^n, v_{\text{opt}}) \quad \text{for all } v_{\text{opt}} \in \Theta_h(U_h). \end{aligned}$$

Second, by (2.2),

$$a(\mathbf{U}_h^n, v_{\text{opt}}) = (f^n + k^{-1} u_h^{n-1}, v_{\text{opt}}) \quad \text{for all } v_{\text{opt}} \in \Theta_h(U_h).$$

Third, combining both identities and writing $\mathbf{W}^n = (w^n, \hat{\chi}^n) := \mathcal{E}_h \mathbf{U}^n - \mathbf{U}_h^n$ yields

$$\frac{1}{k}(w^n, v_{\text{opt}}) + b(\mathbf{W}^n, v_{\text{opt}}) = \frac{1}{k}(e_h^n + w^{n-1}, v_{\text{opt}}) \quad \text{for all } v_{\text{opt}} \in \Theta_h(U_h), \quad (3.1)$$

where

$$e_h^n := \mathcal{E}_h^1(\mathbf{U}^n - \mathbf{U}^{n-1}) - k \dot{u}^n. \quad (3.2)$$

Putting the term with w^{n-1} on the left-hand side yields

$$\frac{1}{k}(w^n - w^{n-1}, v_{\text{opt}}) + b(\mathbf{W}^n, v_{\text{opt}}) = \frac{1}{k}(e_h^n, v_{\text{opt}}) \quad \text{for all } v_{\text{opt}} \in \Theta_h(U_h), \quad (3.3)$$

Step 3. We use the test function $v_{\text{opt}} = \Theta_h(v_h, 0) = v_h + \tilde{v}_h$ with $v_h = w^n - w^{n-1}$. Reordering the terms in the error equations (3.3),

$$\begin{aligned} \frac{1}{k}\|w^n - w^{n-1}\|^2 + \frac{1}{2}\|\mathbf{A}^{\frac{1}{2}} \nabla w^n\|^2 - \frac{1}{2}\|\mathbf{A}^{\frac{1}{2}} \nabla w^{n-1}\|^2 + \frac{1}{2}\|\mathbf{A}^{\frac{1}{2}} \nabla(w^n - w^{n-1})\|^2 \\ = \frac{1}{k}(e_h^n, v_h + \tilde{v}_h) - \frac{1}{k}(w^n - w^{n-1}, \tilde{v}_h) - b(\mathbf{W}^n, \tilde{v}_h) - (\boldsymbol{\beta} \cdot \nabla w^n, v_h) - (\gamma w^n, v_h) \\ = r_1 + r_2 + r_3 + r_4 + r_5. \end{aligned} \quad (3.4)$$

Step 4. We estimate the contributions r_j of the right-hand side of the error equation (3.4). Recall that $v_h = w^n - w^{n-1}$. Throughout we use Young's inequality with a parameter $\delta > 0$ and the estimate from (2.6), i.e., $k^{-\frac{1}{2}} \|\tilde{v}_h\| \leq \|\tilde{v}_h\|_{V,k} \leq k^{\frac{1}{2}} \|\nabla(w^n - w^{n-1})\|$. First,

$$\begin{aligned} |r_1| &= \frac{1}{k} |(e_h^n, v_h + \tilde{v}_h)| \leq k^{-\frac{1}{2}} \|e_h^n\| k^{-\frac{1}{2}} \|v_h\| + k^{-\frac{1}{2}} \|e_h^n\| k^{-\frac{1}{2}} \|\tilde{v}_h\| \\ &\leq \delta^{-1} k^{-1} \|e_h^n\|^2 + \delta k^{-1} \|w^n - w^{n-1}\|^2 + \delta k \|\mathbf{A}^{\frac{1}{2}} \nabla(w^n - w^{n-1})\|^2. \end{aligned}$$

Second,

$$|r_2| \leq k^{-\frac{1}{2}} \|w^n - w^{n-1}\| k^{-\frac{1}{2}} \|\tilde{v}_h\| \leq \delta k^{-1} \|w^n - w^{n-1}\|^2 + \delta^{-1} k \|\mathbf{A}^{\frac{1}{2}} \nabla(w^n - w^{n-1})\|^2.$$

Third, recall the definition of the (discrete) optimal test function, i.e.,

$$|\langle \hat{\chi}^n, \tilde{v}_h \rangle_S| = |a((0, \hat{\chi}^n), \tilde{v}_h)| = |(\Theta_h(0, \hat{\chi}^n), \tilde{v}_h)_{V,k}|$$

which together with the Cauchy–Schwarz inequality shows that

$$\begin{aligned} |r_3| &= |b(\mathbf{W}^n, \tilde{v}_h)| \leq (\|\nabla w^n\| + \|\Theta_h(0, \hat{\chi}^n)\|_{V,k}) k^{\frac{1}{2}} \|\nabla v_h\| \\ &\leq \|\nabla(w^n - w^{n-1})\| k^{\frac{1}{2}} \|\nabla v_h\| + k^{\frac{1}{2}} \|\nabla w^{n-1}\| \|\nabla v_h\| + k^{\frac{1}{2}} \|\Theta_h(0, \hat{\chi}^n)\|_{V,k} \|\nabla v_h\|. \end{aligned}$$

We need to estimate $\|\Theta_h(0, \hat{\chi}^n)\|_{V,k}$. To this end, using the error equation (3.1)

$$\frac{1}{k}(w^n, \Theta_h(0, \hat{\chi}^n)) + b(\mathbf{W}^n, \Theta_h(0, \hat{\chi}^n)) = \frac{1}{k}(e_h^n + w^{n-1}, \Theta_h(0, \hat{\chi}^n))$$

and the definition of Θ_h , we get that

$$\begin{aligned} \frac{1}{k}(\Theta_h(0, \hat{\chi}^n), w^n) + \underbrace{(\mathbf{A} \nabla_{\mathcal{T}} \Theta_h(0, \hat{\chi}^n), \nabla_{\mathcal{T}} w^n)}_{= -\langle \hat{\chi}^n, w^n \rangle_S = 0} + (\boldsymbol{\beta} \cdot \nabla w^n + \gamma w^n, \Theta_h(0, \hat{\chi}^n)) + \|\Theta_h(0, \hat{\chi}^n)\|_{V,k}^2 \\ = \frac{1}{k}(e_h^n + w^{n-1}, \Theta_h(0, \hat{\chi}^n)). \end{aligned}$$

Using the definition of Θ_h another time shows that

$$\frac{1}{k}(w^{n-1}, \Theta_h(0, \hat{\chi}^n)) = -(\mathbf{A} \nabla_{\mathcal{T}} \Theta_h(0, \hat{\chi}^n), \nabla w^{n-1}) - \underbrace{\langle \hat{\chi}^n, w^{n-1} \rangle_S}_{=0} \lesssim \|\Theta_h(0, \hat{\chi}^n)\|_{V,k} \|\nabla w^{n-1}\|.$$

Combining the latter two estimates together with standard estimates gives

$$\|\Theta_h(0, \hat{\chi}^n)\|_{V,k} \lesssim k^{-\frac{1}{2}} \|e_h^n\| + \|\nabla w^{n-1}\| + k^{\frac{1}{2}} \|\nabla w^n\|.$$

Using equations (3.1) with $v_{\text{opt}} = \Theta_h \mathbf{W}^n$, we see that

$$\|\Theta_h \mathbf{W}^n\|_{V,k}^2 = a(\mathbf{W}^n, \Theta_h \mathbf{W}^n) = k^{-1}(e_h^n + w^{n-1}, \Theta_h \mathbf{W}^n) \leq k^{-\frac{1}{2}} \|e_h^n + w^{n-1}\| \|\Theta_h \mathbf{W}^n\|_{V,k}.$$

Together with Lemma 2, multiplying with $k^{\frac{1}{2}}$ and a Poincaré inequality we get $k^{\frac{1}{2}} \|\nabla w^n\| \lesssim \|e_h^n\| + \|\nabla w^{n-1}\|$. Therefore,

$$\|\Theta_h(0, \hat{\chi}^n)\|_{V,k} \lesssim k^{-\frac{1}{2}} \|e_h^n\| + \|\nabla w^{n-1}\|. \quad (3.5)$$

For the term r_3 we thus get the estimate

$$|r_3| \lesssim (k^{\frac{1}{2}} + \delta) \|\mathbf{A}^{\frac{1}{2}} \nabla(w^n - w^{n-1})\|^2 + k\delta^{-1} \|\nabla w^{n-1}\|^2 + \delta^{-1} \|e_h^n\|^2.$$

Fourth,

$$|r_4| \lesssim k^{\frac{1}{2}} \|\nabla w^n\| k^{-\frac{1}{2}} \|v_h\| \lesssim k\delta^{-1} \|\mathbf{A}^{\frac{1}{2}} \nabla(w^n - w^{n-1})\| + k\delta^{-1} \|\mathbf{A}^{\frac{1}{2}} \nabla w^{n-1}\|^2 + \delta k^{-1} \|w^n - w^{n-1}\|^2.$$

Fifth, using Poincaré's inequality,

$$\begin{aligned} |r_5| &\lesssim \|w^n\| \|v_h\| \lesssim k\delta^{-1} \|\nabla w^n\|^2 + \delta k^{-1} \|v_h\|^2 \\ &\lesssim k\delta^{-1} \|\mathbf{A}^{\frac{1}{2}} \nabla(w^n - w^{n-1})\|^2 + k\delta^{-1} \|\mathbf{A}^{\frac{1}{2}} \nabla w^{n-1}\|^2 + \delta k^{-1} \|w^n - w^{n-1}\|^2. \end{aligned}$$

Step 5. Using the bounds for the r_j in the error equation (3.4), choosing first δ sufficiently small independently of k , and then $k \leq k_0$ sufficiently small in dependence of δ , and taking into account the norm equivalence $\|\mathbf{A}^{\frac{1}{2}} \nabla(\cdot)\| \approx \|\nabla(\cdot)\|$, we obtain

$$\|\mathbf{A}^{\frac{1}{2}} \nabla w^n\|^2 - \|\mathbf{A}^{\frac{1}{2}} \nabla w^{n-1}\|^2 \lesssim k^{-1} \|e_h^n\|^2 + k \|\nabla w^{n-1}\|^2.$$

Iterating this estimate, applying the discrete Gronwall inequality [29, Lemma 10.5], $\sum_{j=1}^n k \leq T$ and norm equivalence $\|\mathbf{A}^{\frac{1}{2}} \nabla(\cdot)\| \approx \|\nabla(\cdot)\|$ once again show that

$$\|\nabla w^n\|^2 \lesssim \sum_{j=1}^n \frac{1}{k} \|e_h^j\|^2 + \|\nabla w^0\|^2.$$

Step 6. The basic ideas to estimate the sum $\sum_{j=1}^n \frac{1}{k} \|e_h^j\|^2$ are the same as given in [29, Section 1] with some modifications, i.e., the use of elliptic projection operators and its properties defined in the present work. For the sake of completeness we recall the main steps where we closely follow the presentation of our own work [20]: We write the error e_h^j defined in (3.2) as

$$e_h^j := e_h^{j,1} + e_h^{j,2} := [\mathcal{E}_h^1(\mathbf{U}^j - \mathbf{U}^{j-1}) - (u^j - u^{j-1})] + [u^j - u^{j-1} - k\dot{u}^j].$$

We write

$$\mathbf{U}^j - \mathbf{U}^{j-1} = \int_{t_{j-1}}^{t_j} \dot{\mathbf{U}}(s) \, ds.$$

Then the first term is estimated with Corollary 7 and the Cauchy–Schwarz inequality in the time variable, yielding

$$\|e_h^{j,1}\| \lesssim \int_{t_{j-1}}^{t_j} h^{p+1} C(\dot{\mathbf{U}}(s)) \, ds \leq k^{\frac{1}{2}} \left(\int_{t_{j-1}}^{t_j} h^{2(p+1)} C(\dot{\mathbf{U}}(s))^2 \, ds \right)^{\frac{1}{2}}.$$

Here, $C(\dot{\mathbf{U}}(s))$ depends on higher-order Sobolev norms of $\dot{\mathbf{U}}(s)$. Similar arguments show for the second contribution $e_h^{j,2}$ that

$$\|e_h^{j,2}\| = \|u^j - u^{j-1} - k\dot{u}^j\| \leq k \int_{t_{j-1}}^{t_j} \|\ddot{u}(s)\| \, ds \leq k^{\frac{3}{2}} \left(\int_{t_{j-1}}^{t_j} \|\ddot{u}(s)\|^2 \, ds \right)^{\frac{1}{2}}.$$

Combining the latter two estimates, we conclude that

$$\frac{1}{k} \sum_{j=1}^n \|e_h^j\|^2 = \mathcal{O}(h^{2(p+1)} + k^2).$$

Step 7. The trace estimate follows from

$$\|\Theta(0, \hat{\chi}^n)\|_{V,k} \leq k^{-\frac{1}{2}} \|e_h^n\| + \|\nabla w^{n-1}\|,$$

cf. (3.5), by norm equivalence $\|\Theta(0, \hat{\chi}^n)\|_{V,k} \approx \|\hat{\chi}^n\|_{-\frac{1}{2},k}$ (Lemma 2) under the assumption $hk^{-\frac{1}{2}} = \mathcal{O}(1)$. $\square$

4 Optimal Error Estimate in Weaker Norms

Throughout this section we assume that the coefficients $\mathbf{A}, \boldsymbol{\beta}, \gamma$ and the domain Ω are such that for given data $f, g \in L^2(\Omega)$ the unique solutions $w, v \in H_0^1(\Omega)$ of the primal problem

$$-\operatorname{div} \mathbf{A} \nabla w + \boldsymbol{\beta} \cdot \nabla w + \gamma w = f \quad \text{in } \Omega, \quad (4.1a)$$

$$w|_{\partial\Omega} = 0 \quad (4.1b)$$

and the dual problem

$$-\operatorname{div}(\mathbf{A} \nabla v + \boldsymbol{\beta} v) + \gamma v = g \quad \text{in } \Omega, \quad (4.2a)$$

$$v|_{\partial\Omega} = 0 \quad (4.2b)$$

satisfy the regularity estimates

$$\|w\|_{H^2(\Omega)} + \|\mathbf{A} \nabla w\|_{H^1(\mathcal{T})} \lesssim \|f\|, \quad \|v\|_{H^2(\Omega)} \lesssim \|g\|. \quad (4.3)$$

4.1 Further Analysis of Elliptic Projection

We need to take a closer look at the elliptic projection operator in order to show higher rates in the $L^2(\Omega)$ norm. To that end, we will consider duality arguments (Aubin–Nitsche trick), see [2] for the primal DPG method for the Poisson problem. Duality arguments for the DPG method with ultra-weak formulation have been explored in [14, 15]. Contrary to the mentioned works, here we need to develop duality arguments for a non-symmetric bilinear form that defines the elliptic projection. Additionally, the norms in this work depend on the time step which can be arbitrarily small, making the analysis more delicate.

Recall the definition of the elliptic projection from (2.5): Given $\mathbf{U} \in U$, define $\mathcal{E}_h \mathbf{U} := \mathbf{U}_h \in U_h$ by

$$b(\mathbf{U}_h, \Theta_h \mathbf{W}_h) = b(\mathbf{U}, \Theta_h \mathbf{W}_h) \quad \text{for all } \mathbf{W}_h \in U_h, \quad (4.4)$$

where the discrete trial-to-test operator $\Theta_h: U_h \rightarrow V_h$ is defined through

$$(\Theta_h \mathbf{W}_h, \delta v)_{V,k} = a(\mathbf{W}_h, \delta v) \quad \text{for all } \delta v \in V_h.$$

For the analysis we will use an equivalent representation as a non-standard mixed system:

Lemma 9. *Problem (4.4) is equivalent to the mixed system: Find $(v_h, \mathbf{U}_h) \in V_h \times U_h$ such that*

$$(v_h, \delta v)_{V,k} + b(\mathbf{U}_h, \delta v) = b(\mathbf{U}, \delta v), \quad (4.5a)$$

$$a(\delta \mathbf{W}, v_h) = 0 \quad (4.5b)$$

for all $(\delta v, \delta \mathbf{W}) \in V_h \times U_h$. In particular,

$$\|v_h\|_{V,k} \leq \|\mathbf{U} - \mathbf{U}_h\|_k.$$

Proof. Let $\mathbf{U}_h \in U_h$ be the solution of (4.4). Define $v_h \in V_h$ by $(v_h, \delta v)_{V,k} = b(\mathbf{U} - \mathbf{U}_h, \delta v)$ for all $\delta v \in V_h$. Then

$$a(\delta \mathbf{W}, v_h) = (\Theta_h \delta \mathbf{W}, v_h)_{V,k} = b(\mathbf{U} - \mathbf{U}_h, \Theta_h \delta \mathbf{W}) = 0$$

for all $\delta \mathbf{W} \in U_h$.

To see the other direction, suppose that $(v_h, \mathbf{U}_h) \in V_h \times U_h$ solves (4.5). With the same consideration as above we get that

$$0 = a(\delta \mathbf{W}, v_h) = (\Theta_h \delta \mathbf{W}, v_h)_{V,k} = b(\mathbf{U} - \mathbf{U}_h, \Theta_h \delta \mathbf{W}),$$

for all $\delta \mathbf{W} \in U_h$ which means that $\mathbf{U}_h \in U_h$ satisfies (4.4).

The final estimate follows from (4.5a) and the definition of the norm $\|\cdot\|_k$,

$$\begin{aligned} \|v_h\|_{V,k}^2 &= b(\mathbf{U} - \mathbf{U}_h, v_h) = b((u - u_h, 0), v_h) - \langle \hat{\sigma} - \hat{\sigma}_h, v_h \rangle_S \\ &= b((u - u_h, 0), v_h) + a((0, \hat{\sigma} - \hat{\sigma}_h), v_h) \\ &= b((u - u_h, 0), v_h) + (\Theta_h(0, \hat{\sigma} - \hat{\sigma}_h), v_h)_{V,k} \lesssim \|\mathbf{U} - \mathbf{U}_h\|_k \|v_h\|_{V,k}. \end{aligned}$$

This concludes the proof. $\square$

Theorem 10. Suppose that problems (4.1) and (4.2) satisfy the regularity estimates in (4.3). Then, for given $\mathbf{U} = (u, \hat{\sigma}) \in U$, the elliptic projection $\mathbf{U}_h := (u_h, \hat{\sigma}_h) := \mathcal{E}_h \mathbf{U} \in U_h$ converges at a higher rate in the $L^2(\Omega)$ norm, i.e.,

$$\|u - u_h\| \lesssim h \|\mathbf{U} - \mathbf{U}_h\|_k.$$

Proof. We divide the proof into three steps.

Step 1. Let $v \in H_0^1(\Omega)$ denote the solution of the dual problem (4.2) with $g = u - u_h$. Integration by parts and the definition of the bilinear form $b(\cdot, \cdot)$ yield

$$\begin{aligned} \|u - u_h\|^2 &= (u - u_h, -\operatorname{div}(\mathbf{A} \nabla v + \boldsymbol{\beta} v) + \gamma v) \\ &= (\mathbf{A} \nabla(u - u_h), \nabla v) + (\boldsymbol{\beta} \cdot \nabla(u - u_h) + \gamma(u - u_h), v) = b(\mathbf{U} - \mathbf{U}_h, v). \end{aligned}$$

Step 2. We characterize $\mathbf{W}$ with $v = \Theta \mathbf{W}$, where $\Theta: U \rightarrow V$ is the continuous trial-to-test operator which is an isomorphism and defined through

$$(\Theta \mathbf{W}, \delta v)_{V,k} = a(\mathbf{W}, \delta v) \quad \text{for all } \delta v \in V.$$

From the latter definition we obtain

$$(v, \delta v)_{V,k} = \frac{1}{k}(v, \delta v) + (\mathbf{A} \nabla v, \nabla \delta v) = a((v, 0), \delta v) - (\boldsymbol{\beta} \cdot \nabla v + \gamma v, \delta v) \quad \text{for all } \delta v \in V.$$

It follows that $\mathbf{W} = (v, 0) + \mathbf{W}^*$, where $\mathbf{W}^* = (w^*, \hat{\chi}^*) \in U$. The component $w^* \in H_0^1(\Omega)$ solves

$$-\operatorname{div} \mathbf{A} \nabla w^* + \boldsymbol{\beta} \cdot \nabla w^* + \gamma w^* + \frac{1}{k} w^* = -\boldsymbol{\beta} \cdot \nabla v - \gamma v \quad (4.6)$$

and $\hat{\chi}^*$ is given by $\hat{\chi}^*|_{\partial T} = \mathbf{A} \nabla w^* \cdot \mathbf{n}_K|_{\partial K}$ for all $K \in \mathcal{T}$.

Step 3. Let $v_h \in V_h$ be such that $(v_h, \mathbf{U}_h) \in V_h \times U_h$ is the solution of the equivalent mixed system (4.5). Then we see with Step 1 and using (4.5) that

$$\begin{aligned} \|u - u_h\|^2 &= b(\mathbf{U} - \mathbf{U}_h, v) = b(\mathbf{U} - \mathbf{U}_h, v) - (v, v_h)_{V,k} + a(\mathbf{W}, v_h) \\ &= b(\mathbf{U} - \mathbf{U}_h, v - \delta v) - (v - \delta v, v_h)_{V,k} + a(\mathbf{W} - \delta \mathbf{W}, v_h) \end{aligned}$$

for all $(\delta v, \delta \mathbf{W}) \in V_h \times U_h$.

We choose $\delta v \in \mathcal{P}^{p+1}(\Omega) \cap H_0^1(\Omega)$ to be the best approximation of v in the $H_0^1(\Omega)$ norm and

$$\delta \mathbf{W} = (\delta v, 0) + \delta \mathbf{W}^*,$$

where $\delta \mathbf{W}^* \in U_h$ will be chosen below. The first term is estimated (note that $v - \delta v \in H_0^1(\Omega)$) using the regularity of v by

$$|b(\mathbf{U} - \mathbf{U}_h, v - \delta v)| \leq \|\nabla(u - u_h)\| \|\nabla(v - \delta v)\| \lesssim h \|v\|_{H^2(\Omega)} \|\nabla(u - u_h)\| \lesssim h \|u - u_h\| \|\mathbf{U} - \mathbf{U}_h\|_k.$$

For the remaining terms, note that

$$\mathbf{W} - \delta \mathbf{W} = (v - \delta v, 0) + (\mathbf{W}^* - \delta \mathbf{W}^*).$$

Then, using the estimate $\|v_h\|_{V,k} \leq \| \mathbf{U} - \mathbf{U}_h \|_k$ from Lemma 9 and boundedness (2.1) of a ,

$$\begin{aligned} -(v - \delta v, v_h)_{V,k} + a(\mathbf{W} - \delta \mathbf{W}, v_h) &= (\boldsymbol{\beta} \cdot \nabla(v - \delta v) + \gamma(v - \delta v), v_h) + a(\mathbf{W}^* - \delta \mathbf{W}^*, v_h) \\ &\leq k^{\frac{1}{2}} \|\nabla(v - \delta v)\| \|v_h\|_{V,k} + (k^{-\frac{1}{2}} \|\mathbf{w}^* - \delta \mathbf{w}^*\| + \|\mathbf{W}^* - \delta \mathbf{W}^*\|_{U,k}) \|v_h\|_{V,k} \\ &\leq h \|u - u_h\| \| \mathbf{U} - \mathbf{U}_h \|_k + (k^{-\frac{1}{2}} \|\mathbf{w}^* - \delta \mathbf{w}^*\| + \|\mathbf{W}^* - \delta \mathbf{W}^*\|_{U,k}) \| \mathbf{U} - \mathbf{U}_h \|_k. \end{aligned}$$

To finish the proof, it remains to show the estimate $(k^{-\frac{1}{2}} \|\mathbf{w}^* - \delta \mathbf{w}^*\| + \|\mathbf{W}^* - \delta \mathbf{W}^*\|_{U,k}) \lesssim h \|u - u_h\|$. Basically, this follows from the weak form of the PDE defined in Step 2: Let $\delta \mathbf{W}^* = (\delta \mathbf{w}^*, \delta \hat{\chi}^*) \in U_h$ with $\delta \mathbf{w}^*$ being the best approximation of $\mathbf{w}^*$ with respect to $\|\nabla(\cdot)\|$ and $\delta \hat{\chi}^*|_{\partial K} = \delta \chi^* \cdot \mathbf{n}_K|_{\partial K}$ for all $K \in \mathcal{T}$. Here, $\delta \chi^*$ denotes the Raviart–Thomas projection of $\chi^* = \mathbf{A} \nabla \mathbf{w}^*$ into the Raviart–Thomas space of order p .

The weak form of PDE (4.6) shows that

$$\begin{aligned} \frac{1}{k} \|\mathbf{w}^*\|^2 + \|\mathbf{A}^{\frac{1}{2}} \nabla \mathbf{w}^*\|^2 &\leq \frac{1}{k} (\mathbf{w}^*, \mathbf{w}^*) + (\mathbf{A} \nabla \mathbf{w}^*, \nabla \mathbf{w}^*) + (\boldsymbol{\beta} \cdot \nabla \mathbf{w}^* + \gamma \mathbf{w}^*, \mathbf{w}^*) \\ &= (-\boldsymbol{\beta} \cdot \nabla v - \gamma v, \mathbf{w}^*) \leq \|\nabla v\| \|\mathbf{w}^*\|, \end{aligned}$$

which implies first $k^{-1} \|\mathbf{w}^*\| \leq \|\nabla v\|$. Bootstrapping this estimate then shows

$$k^{-1} \|\mathbf{w}^*\| + k^{-\frac{1}{2}} \|\nabla \mathbf{w}^*\| \leq \|\nabla v\| \lesssim \|u - u_h\|.$$

Moreover, PDE (4.6) together with our regularity assumptions (4.3), and using the latter estimate show that

$$\|\mathbf{w}^*\|_{H^2(\Omega)} \lesssim \|-\boldsymbol{\beta} \cdot \nabla v - \gamma v - k^{-1} \mathbf{w}^*\| \lesssim \|\nabla v\| \lesssim \|u - u_h\|.$$

Therefore, standard approximation results and the foregoing stability analysis prove that

$$k^{-\frac{1}{2}} \|\mathbf{w}^* - \delta \mathbf{w}^*\| + \|\nabla(\mathbf{w}^* - \delta \mathbf{w}^*)\| \lesssim h k^{-\frac{1}{2}} \|\nabla \mathbf{w}^*\| + h \|\mathbf{w}^*\|_{H^2(\Omega)} \lesssim h \|u - u_h\|.$$

By our regularity assumptions we have that $\chi^* = \mathbf{A} \nabla \mathbf{w}^* \in H^1(\mathcal{T})^d \cap \mathbf{H}(\text{div}; \Omega)$ so that the trace terms can be estimated in a similar fashion as in [14, Theorem 5]: From Lemma 1 we know that

$$\|\hat{\chi}^* - \delta \hat{\chi}^*\|_{-\frac{1}{2},k} \approx \sup_{0 \neq \tilde{v} \in V} \frac{\langle \hat{\chi}^* - \delta \hat{\chi}^*, \tilde{v} \rangle_S}{\|\tilde{v}\|_{V,k}}.$$

Then integration by parts yields

$$\langle \hat{\chi}^* - \delta \hat{\chi}^*, \tilde{v} \rangle_S = (\chi^* - \delta \chi^*, \nabla_{\mathcal{T}} \tilde{v}) + (\text{div } \chi^*, (1 - \Pi^p) \tilde{v}) \lesssim (h \|\chi^*\|_{H^1(\mathcal{T})} + h \|\text{div } \chi^*\|) \|\nabla_{\mathcal{T}} \tilde{v}\|,$$

where Π^p is the L^2 projection on $\mathcal{P}^p(\mathcal{T})$. For details on the arguments used we refer to [14, proof of Theorem 5]. We stress that the estimates are independent of k . Using the bounds from Step 3, this shows that

$$\min_{\delta \hat{\chi} \in H^{-\frac{1}{2}}(\mathcal{S})} \|\hat{\chi}^* - \delta \hat{\chi}\|_{-\frac{1}{2},k} \lesssim h \|u - u_h\|,$$

which concludes the proof. $\square$

4.2 Error Analysis in the $L^2(\Omega)$ Norm

Theorem 11. Let $U_h^n \in U_h$ denote the solution of (2.2). Suppose that the components of $\mathbf{U}(t; \cdot)$ are sufficiently regular. Under the additional regularity assumptions (4.3) the solution satisfies

$$\|u^n - u_h^n\| = \mathcal{O}(h^{p+2} + k) + \mathcal{O}(\|u^0 - u_h^0\|).$$

Proof. With the elliptic projection operator $\mathcal{E}_h$ (see (2.5)) we consider the splitting

$$u^n - u_h^n = (u^n - \mathcal{E}_h^1 U^n) + (\mathcal{E}_h^1 U^n - u_h^n).$$

By Theorem 10 and Corollary 7 we get that

$$\|u^n - \mathcal{E}_h^1 U^n\| = \mathcal{O}(h^{p+2}).$$

Writing $\mathbf{W}^n = (w^n, \hat{\chi}^n) := \mathcal{E}_h U^n - U_h^n$, we recall the error equation from (3.3),

$$\frac{1}{k}(w^n, v_{\text{opt}}) + b(\mathbf{W}^n, v_{\text{opt}}) = \frac{1}{k}(e_h^n + w^{n-1}, v_{\text{opt}}) \quad \text{for all } v_{\text{opt}} \in \Theta_h(U_h).$$

We test with $v_{\text{opt}} = \Theta_h \mathbf{W}^n$ and using Lemma 2, we infer that

$$k^{-1}\|w^n\|^2 \leq a(\mathbf{W}^n, \Theta_h \mathbf{W}^n) = \frac{1}{k}(e_h^n + w^{n-1}, \Theta_h \mathbf{W}^n) \leq k^{-\frac{1}{2}}\|e_h^n + w^{n-1}\|a(\mathbf{W}^n, \Theta_h \mathbf{W}^n)^{\frac{1}{2}},$$

and further that

$$\|w^n\| \leq \|e_h^n\| + \|w^{n-1}\|.$$

Iterating the arguments yields

$$\|w^n\| \leq \sum_{j=1}^n \|e_h^j\| + \|w^0\|.$$

The last term is estimated with the triangle inequality together with Theorem 10 and Corollary 7 to obtain that

$$\|w^0\| \leq \|u^0 - u_h^0\| + \|u^0 - \mathcal{E}_h^1 U^0\| = \|u^0 - u_h^0\| + \mathcal{O}(h^{p+2}).$$

The estimate

$$\sum_{j=1}^n \|e_h^j\| = \mathcal{O}(h^{p+2} + k)$$

is shown by following the very same argumentation as in [29, Theorem 1.5]. For the sake of completeness we repeat the main steps which are also similar to the ones already presented in Step 7 of the proof of Theorem 8.

From Step 7 of the proof of Theorem 8 we recall the splitting

$$e_h^j := e_h^{j,1} + e_h^{j,2} := [\mathcal{E}_h^1(U^j - U^{j-1}) - (u^j - u^{j-1})] + [u^j - u^{j-1} - k\dot{u}^j].$$

With the same arguments (without using a Cauchy–Schwarz inequality in the time variable) we get that

$$\|e_h^{j,2}\| \leq k \int_{t_{j-1}}^{t_j} \|\ddot{u}(s)\| \, ds.$$

Similarly, for the first contribution we apply Corollary 7 and Theorem 10 to get that

$$\|e_h^{j,1}\| \leq h^{p+2} \int_{t_{j-1}}^{t_j} C(\dot{U}(s)) \, ds,$$

where $C(\dot{U}(s))$ depends on higher-order Sobolev norms of $\dot{U}(s)$. Summing over j finishes the proof. $\square$

Funding: This work was supported by ANID through FONDECYT projects 1210391 and 1190009 and 1210579.

References

- [1] R. Andreev, Stability of sparse space-time finite element discretizations of linear parabolic evolution equations, *IMA J. Numer. Anal.* **33** (2013), no. 1, 242–260.
- [2] T. Bouma, J. Gopalakrishnan and A. Harb, Convergence rates of the DPG method with reduced test space degree, *Comput. Math. Appl.* **68** (2014), no. 11, 1550–1561.
- [3] D. Broersen and R. Stevenson, A robust Petrov–Galerkin discretisation of convection-diffusion equations, *Comput. Math. Appl.* **68** (2014), no. 11, 1605–1618.

- [4] C. Carstensen, L. Demkowicz and J. Gopalakrishnan, Breaking spaces and forms for the DPG method and applications including Maxwell equations, *Comput. Math. Appl.* **72** (2016), no. 3, 494–522.
- [5] L. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. Part I: The transport equation, *Comput. Methods Appl. Mech. Engrg.* **199** (2010), no. 23–24, 1558–1572.
- [6] L. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. Part II. Optimal test functions, *Numer. Methods Partial Differential Equations* **27** (2011), no. 1, 70–105.
- [7] L. Demkowicz and J. Gopalakrishnan, Analysis of the DPG method for the Poisson equation, *SIAM J. Numer. Anal.* **49** (2011), no. 5, 1788–1809.
- [8] L. Demkowicz and J. Gopalakrishnan, A primal DPG method without a first-order reformulation, *Comput. Math. Appl.* **66** (2013), no. 6, 1058–1064.
- [9] L. Demkowicz, J. Gopalakrishnan, S. Nagaraj and P. Sepúlveda, A spacetime DPG method for the Schrödinger equation, *SIAM J. Numer. Anal.* **55** (2017), no. 4, 1740–1759.
- [10] L. Demkowicz, J. Gopalakrishnan and A. H. Niemi, A class of discontinuous Petrov–Galerkin methods. Part III: Adaptivity, *Appl. Numer. Math.* **62** (2012), no. 4, 396–427.
- [11] L. Demkowicz and N. Heuer, Robust DPG method for convection-dominated diffusion problems, *SIAM J. Numer. Anal.* **51** (2013), no. 5, 2514–2537.
- [12] L. Diening and J. Storn, A space-time DPG method for the heat equation, preprint (2020), <https://arxiv.org/abs/2012.13229>.
- [13] J. Ernesti and C. Wieners, Space-time discontinuous Petrov–Galerkin methods for linear wave equations in heterogeneous media, *Comput. Methods Appl. Math.* **19** (2019), no. 3, 465–481.
- [14] T. Führer, Superconvergence in a DPG method for an ultra-weak formulation, *Comput. Math. Appl.* **75** (2018), no. 5, 1705–1718.
- [15] T. Führer, Superconvergent DPG methods for second-order elliptic problems, *Comput. Methods Appl. Math.* **19** (2019), no. 3, 483–502.
- [16] T. Führer, A. Haberl and N. Heuer, Trace operators of the bi-Laplacian and applications, *IMA J. Numer. Anal.* **41** (2021), no. 2, 1031–1055.
- [17] T. Führer and N. Heuer, A robust DPG method for large domains, *Comput. Math. Appl.* **94** (2021), 15–27.
- [18] T. Führer, N. Heuer and A. H. Niemi, An ultraweak formulation of the Kirchhoff–Love plate bending model and DPG approximation, *Math. Comp.* **88** (2019), no. 318, 1587–1619.
- [19] T. Führer, N. Heuer and J. Sen Gupta, A time-stepping DPG scheme for the heat equation, *Comput. Methods Appl. Math.* **17** (2017), no. 2, 237–252.
- [20] T. Führer and M. Karkulik, New a priori analysis of first-order system least-squares finite element methods for parabolic problems, *Numer. Methods Partial Differential Equations* **35** (2019), no. 5, 1777–1800.
- [21] T. Führer and M. Karkulik, Space-time least-squares finite elements for parabolic equations, *Comput. Math. Appl.* **92** (2021), 27–36.
- [22] J. Gopalakrishnan and W. Qiu, An analysis of the practical DPG method, *Math. Comp.* **83** (2014), no. 286, 537–552.
- [23] S. Henneking, J. Grosek and L. Demkowicz, Model and computational advancements to full vectorial Maxwell model for studying fiber amplifiers, *Comput. Math. Appl.* **85** (2021), 30–41.
- [24] N. Heuer and M. Karkulik, A robust DPG method for singularly perturbed reaction-diffusion problems, *SIAM J. Numer. Anal.* **55** (2017), no. 3, 1218–1242.
- [25] J. Muñoz Matute, D. Pardo and L. Demkowicz, A DPG-based time-marching scheme for linear hyperbolic problems, *Comput. Methods Appl. Mech. Engrg.* **373** (2021), Article ID 113539.
- [26] J. Muñoz Matute, D. Pardo and L. Demkowicz, Equivalence between the DPG method and the exponential integrators for linear parabolic problems, *J. Comput. Phys.* **429** (2021), Article ID 110016.
- [27] N. V. Roberts and S. Henneking, Time-stepping DPG formulations for the heat equation, *Comput. Math. Appl.* **95** (2021), 242–255.
- [28] R. Stevenson and J. Westerdiep, Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations, *IMA J. Numer. Anal.* **41** (2021), no. 1, 28–47.
- [29] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer Ser. Comput. Math. 25, Springer, Berlin, 2006.
- [30] K. Voronin, C. S. Lee, M. Neumüller, P. Sepulveda and P. S. Vassilevski, Space-time discretizations using constrained first-order system least squares (CFOSLS), *J. Comput. Phys.* **373** (2018), 863–876.
- [31] M. F. Wheeler, A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations, *SIAM J. Numer. Anal.* **10** (1973), 723–759.