Error estimation for second-order PDEs in non-variational form

Jan Blechschmidt∗ Roland Herzog† Max Winkler‡

August 13, 2020

Second-order partial differential equations in non-divergence form are considered. Equations of this kind typically arise as subproblems for the solution of Hamilton-Jacobi-Bellman equations in the context of stochastic optimal control, or as the linearization of fully nonlinear second-order PDEs. The non-divergence form in these problems is natural. If the coefficients of the diffusion matrix are not differentiable, the problem can not be transformed into the more convenient variational form.

We investigate tailored non-conforming finite element approximations of second-order PDEs in non-divergence form, utilizing finite-element Hessian recovery strategies to approximate second derivatives in the equation. We study both approximations with continuous and discontinuous trial functions. Of particular interest are a priori and a posteriori error estimates as well as adaptive finite element methods. In numerical experiments our method is compared with other approaches known from the literature.

1 Introduction

Many boundary value problems feature linear, second-order partial differential equations in divergence form. That is, the differential operator may be written as

\[ \tilde{L}u := \text{div}(\tilde{A}\nabla u) + \tilde{b}^\top \nabla u + \tilde{c}u \] (1)

with coefficients \( \tilde{A} : \Omega \to \mathbb{R}^{d \times d} \), \( \tilde{b} : \Omega \to \mathbb{R}^d \), \( \tilde{c} : \Omega \to \mathbb{R} \). Here and in the following \( \Omega \subset \mathbb{R}^d \), \( d \in \mathbb{N} \), is a bounded domain. Although formulation (1) covers a wide range of applications, there are some linear problems involving operators in non-divergence form

\[ Lu := A : \nabla^2 u + b^\top \nabla u + c u. \] (2)

∗Technische Universität Chemnitz, Faculty of Mathematics, Professorship Numerical Mathematics (Partial Differential Equations), 09107 Chemnitz, Germany, jan.blechschmidt@mathematik.tu-chemnitz.de, https://www.tu-chemnitz.de/mathematik/part_dgl/people/blechschmidt

†Technische Universität Chemnitz, Faculty of Mathematics, Professorship Numerical Mathematics (Partial Differential Equations), 09107 Chemnitz, Germany, roland.herzog@mathematik.tu-chemnitz.de, https://www.tu-chemnitz.de/mathematik/part_dgl/people/herzog

‡Technische Universität Chemnitz, Faculty of Mathematics, Professorship Numerical Mathematics (Partial Differential Equations), 09107 Chemnitz, Germany, max.winkler@mathematik.tu-chemnitz.de, https://www.tu-chemnitz.de/mathematik/part_dgl/people/winkler
Here, $A : B$ denotes the Frobenius inner product $\sum_{i,j=1}^{d} a_{ij} b_{ij}$ of two matrices $A, B \in \mathbb{R}^{d \times d}$, and $A : \Omega \to \mathbb{R}^{d \times d}, b : \Omega \to \mathbb{R}^{d}$ and $c : \Omega \to \mathbb{R}$ are given coefficients. The matrix $A$ is assumed to be almost everywhere positive definite and symmetric.

Classical and strong solutions of problems in non-divergence form with Hölder-regular or continuous coefficients, respectively, have been analyzed in [23, Ch. 6, 9]. In the applications of interest here, however, coefficients are only bounded and measurable. Under even higher smoothness assumptions on the coefficient $A$, a non-divergence form operator (2) can be transformed into an operator in divergence form (1) with $\tilde{A} = A$ and $\tilde{b} = b - \text{Div} A$, where $\text{Div} A$ denotes the row-wise divergence of the matrix $A$. Even if $A$ is smooth, however, this transformation may lead to convection dominated problems which induce further challenges.

Our aim in this paper is to investigate the boundary value problem

$$\mathcal{L}u = f \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } \Gamma := \partial \Omega, \quad (3)$$

for some source term $f \in L^2(\Omega)$. Let us briefly mention some applications where problems of this kind are of interest. Naturally, linear problems with operators in non-divergence form arise in the context of stochastic differential equations due to the Itô formula, see [2, 4, 19, 36]. Such problems play a central role in financial mathematics, e.g., the valuation of financial products.

For the solution of problems in the form (3) several different approaches avoiding the transformation into a divergence form PDE have recently been studied. Many approaches aim at approximating strong solutions, i.e.,

$$\text{Find } u \in H^2(\Omega) \cap H^1_0(\Omega) : \int_{\Omega} \mathcal{L}u \, v \, dx = \int_{\Omega} f \, v \, dx \quad \forall v \in L^2(\Omega). \quad (4)$$

A discrete approximation of the solution $u \in H^2(\Omega) \cap H^1_0(\Omega)$ of (4) in the case $\mathcal{L} = A : \nabla^2$ is usually obtained by solving a problem of the form

$$\text{Find } u_h \in V_{h,0} : \int_{\Omega} A : \mathbb{H}(u_h) \, \tau_h(v_h) \, dx = \int_{\Omega} f \, \tau_h(v_h) \, dx \quad \forall v_h \in V_{h,0}, \quad (5)$$

where $V_{h,0} := \mathbb{V}_h \cap H^1_0(\Omega)$ is a finite-dimensional trial and test space and $\mathbb{H}(u_h)$ is an approximation of the Hessian $\nabla^2 u_h$, also sought in a finite-dimensional space $\mathcal{V}_h(\mathbb{R}^{d \times d})$ with discretization parameter $h > 0$. Several approaches have been studied in the literature and most discretization strategies differ in the choice of the discrete spaces $\mathcal{V}_h$, $\mathbb{W}_h$, the approximation $\mathbb{H}$ of the Hessian and the realization of the test function $\tau_h : \mathbb{V}_h \to L^2(\Omega)$. 

2
Let us briefly summarize the most prominent approaches. The first article discussing a direct treatment of a non-variational problem, to the best of the authors’ knowledge, is Lakkis and Pryer [29]. Therein, \( V_h \) and \( W_h \) consist of continuous Lagrange finite element functions of order \( p \geq 1 \), the choice \( \tau_h = \text{id} \) is used and the finite-element Hessian \( \mathbb{H}(u_h) \in W_h(\mathbb{R}^{d \times d}) \) is obtained by a discrete version of the integration-by-parts formula, i.e.,

\[
\int_{\Omega} \mathbb{H}(u_h) : v_h \, dx = - \int_{\Omega} \nabla u_h \cdot \text{Div} v_h \, dx + \int_{\Gamma} \nabla u_h \cdot (v_h n_\Gamma) \, ds \quad \forall v_h \in W_h(\mathbb{R}^{d \times d}).
\]  

(6)

Here, \( n_\Gamma : \Gamma \to \mathbb{R}^d \) denotes the outer normal vector on \( \Gamma := \partial \Omega \). A closely related approach using a discontinuous Galerkin approximation for the finite-element Hessian \( \mathbb{H}(u_h) \) is studied by Neilan [31].

There are other approaches that avoid the coupling with an additional variational formulation used for the computation of a Hessian approximation. This is possible when using the cell-wise exact Hessian \( \mathbb{H} := \nabla^2_h \) but additional jump penalty terms over the interior cell edges/faces have to be added to the bilinear form. This idea is first studied by Smears and Süli [39], under the weak assumption that \( A \) belongs to \( L^\infty(\Omega; \mathbb{R}^{d \times d}) \) and fulfills a so-called Cordes condition. In that work, discontinuous Galerkin approximations and the choice \( \tau_h(v_h) = \Delta_h v_h \) are used and appropriate jump penalty terms are added to the bilinear form so that discrete coercivity is guaranteed. Quite similar is the approach of Neilan, Salgado, and Zhang [33] who use continuous Lagrange elements. In both approaches the coercivity is shown via a discrete Miranda-Talenti estimate. In a related line of research, Feng, Hennings, and Neilan [14] employ continuous Lagrange finite elements using the choice \( \tau_h = \text{id} \). They show well-posedness of the discrete scheme ensuing via a discrete inf-sup condition. For this approach, at least continuity of the coefficients of \( A \) has to be assumed as a localization argument by freezing the coefficients of \( A \) is applied in the proofs. Analogous results are presented in Feng, Neilan, and Schnake [17] for a discontinuous Galerkin approximation. Finally, an extension of the technique of Smears and Süli [39] to curved domains can be found in [27].

Before continuing, it is worth pointing out that the respective discrete linear systems and the techniques employed to prove their well-posedness differ in the references above and have far-reaching implications for computational practice. In particular, [14, 17] rely on a discrete Calderon-Zygmund estimate and therefore on a continuity assumption for the leading coefficients of the differential operator, as well as sufficiently fine meshes. Unfortunately, the former is typically not satisfied for HJB equations, which we have in mind as future applications. Moreover, the requirement of sufficiently fine initial meshes obstructs the utility of an adaptive mesh refinement strategy, which we develop here. Such limitations are not present in discretization approaches relying on the Cordes condition, including [39, 33, 27] and the present work.

A further method, which is proposed by Gallistl [21], is based on a stabilized mixed finite element discretization involving an approximation of the gradient \( w_h \in W_h(\mathbb{R}^d) \) by \( \int_{\Omega} (\nabla u_h - w_h) : \nabla v_h \, dx = 0 \) for all \( v_h \in V_h \). This is, to the best of our knowledge, the first contribution proving also a posteriori error estimates and the convergence of an adaptive finite element method for the solution of non-divergence form PDEs.

In order to complete our survey, we want to mention that there are many further approaches that do not directly fit into the framework (5). This includes for instance regularization approaches like the vanishing moment method studied in [16] and the references therein, the primal-dual weak Galerkin method [43], the Alexandroff-Bakelman-Pucci (ABP) method [35], or certain finite element schemes based on a very weak formulation of the model problem [20].

In the present paper we discuss a new method combining multiple ideas of the previously outlined approaches. To be more precise, we consider the discrete formulation (5) with a finite-
element Hessian obtained either by continuous finite elements as in (6) or by a discontinuous ansatz that we specify later. For the test functions we use $\tau_{h}(u_{h}) = \mathbb{H}_{\Delta}(u_{h}) := \text{trace}\mathbb{H}(u_{h})$.

The main results of this article include a rigorous proof of the well-posedness of the discrete scheme, which is also based on a discrete Miranda-Talenti estimate following from a Cordes condition. Moreover, preconditioning strategies for the resulting system of linear equations are studied and we observe in experiments that the preconditioner is robust with respect to the mesh parameter. Furthermore, we study a priori and reliable a posteriori error estimates in the energy norm. Based on the a posteriori error estimates we implement an adaptive finite element method and confirm by experiments that the convergence rate is optimal in all test cases, even for less smooth solutions.

Our method combines several advantages of the previously mentioned approaches. First, it is applicable to problems with discontinuous coefficients $A$ and hence allows an extension to HJB equations. Among the approaches presented in our survey, only [39] and [33] possess this property as well. Second, under additional assumptions, our discretization can be realized without the addition of stabilization terms, which would involve jump penalties at the cell interfaces. In numerical experiments we observed that all approaches which do use stabilization terms do not converge with an optimal rate in the $L^{2}(\Omega)$-norm. This surprising observation deserves further investigation. In addition to our approach, only the methods from [31] and [28] likewise exhibit optimal $L^{2}(\Omega)$ rates. It should be noted that the computational cost for the approaches using a Hessian recovery strategy, including the proposed scheme, is naturally higher than the cost for schemes relying on the broken exact Hessian. However, the advantages mentioned above may justify this additional effort.

2 The continuous problem

Throughout this article $\Omega \subset \mathbb{R}^{d}$, $d \in \mathbb{N}$, is a bounded and convex domain. We consider the boundary value problem with a second-order differential operator in non-divergence form

\begin{align}
A : \nabla^{2}u = f & \quad \text{in } \Omega, \\
u = 0 & \quad \text{on } \Gamma \tag{7}
\end{align}

with $f \in L^{2}(\Omega)$. The coefficient matrix $A$ is assumed to belong to $L^{\infty}(\Omega; \mathbb{R}^{d \times d})$, to be symmetric and uniformly positive definite, i.e., there exists a constant $\lambda_{E} > 0$ such that

\begin{equation}
\xi^{T} A \xi \geq \lambda_{E} |\xi|^{2} \quad \text{for all } \xi \in \mathbb{R}^{d}, \tag{8}
\end{equation}

almost everywhere in $\Omega$.

As the coefficient matrix $A$ is not necessarily differentiable, one can at most ask for strong solutions of (7), i.e., functions $u \in X := H^{2}(\Omega) \cap H_{0}^{1}(\Omega)$ solving

\begin{equation}
\int_{\Omega} A : \nabla^{2}u \nabla v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in L^{2}(\Omega). \tag{9}
\end{equation}

Since the Laplacian $\Delta : X \to L^{2}(\Omega)$ is bijective due to the convexity of $\Omega \subset \mathbb{R}^{d}$, the latter equation is equivalent to

\begin{equation}
\int_{\Omega} A : \nabla^{2}u \Delta v \, dx = \int_{\Omega} f \Delta v \, dx \quad \forall v \in X. \tag{10}
\end{equation}
Existence of strong solutions follow for instance under the slightly stronger assumption $A \in C(\Omega; \mathbb{R}^{d \times d})$ and when $\Gamma := \partial \Omega$ is of class $C^{1,1}$, even for non-convex domains, see [23, Theorem 9.15].

Another idea, which implies well-posedness even for general convex domains and which allows for discontinuous coefficients, is to impose a Cordes condition, i.e., the existence of a constant $\varepsilon \in (0, 1]$ such that
\[
\|A\|_{F}^{2} = \frac{\sum_{i,j=1}^{d} (A_{ij})^{2}}{\left(\sum_{i=1}^{d} A_{ii}\right)^{2}} \leq \frac{1}{d-1+\varepsilon} \text{ a.e. in } \Omega. \tag{11}
\]

In the two-dimensional case, this assumption follows from (8). As has been discussed in the recent literature, e.g. [39], a rescaling of the equation (10) with the normalization coefficient $\gamma := \frac{\sum_{i=1}^{d} A_{ii}}{\sum_{i,j=1}^{d} (A_{ij})^{2}} \in L^{\infty}(\Omega)$ becomes advantageous in the analysis of the problem. This can be explained with the following result, whose proof is stated in [39, Lemma 1].

**Lemma 1.** Assume that $A$ belongs to $L^{\infty}(\Omega; \mathbb{R}^{d \times d})$ and satisfies (11). Then the inequality
\[
\|\gamma A - I\|_{F} \leq \sqrt{1 - \varepsilon} \text{ a.e. in } \Omega
\]
defines.

Obviously, (11) guarantees that the rescaled matrix $\gamma A$ is close to the identity matrix, and consequently, the differential operator $\gamma A : \nabla^{2}$ is close to the elliptic Laplace operator. Thus, if the Cordes condition is fulfilled one can consider instead of (10) a variational problem with the bilinear form $a : X \times X \rightarrow \mathbb{R}$ defined by
\[
a(u, v) := \int_{\Omega} \gamma A : \nabla^{2} u \Delta v \, dx,
\]
and the linear form $F \in X'$ defined by
\[
F(v) := \int_{\Omega} \gamma f \Delta v \, dx.
\]
The variational problem we are going to study in this article is defined by
\[
\text{Find } u \in X \text{ such that } a(u, v) = F(v) \quad \forall v \in X. \tag{12}
\]
Under the assumption (11) the bilinear form $a$ is elliptic in $X = H^{2}(\Omega) \cap H^{1}_{0}(\Omega)$ and with the Lax-Milgram Lemma one can immediately prove the following result.

**Lemma 2.** Assume that the coefficient matrix $A$ belongs to $L^{\infty}(\Omega; \mathbb{R}^{d \times d})$ and fulfills the Cordes condition (11) with $\varepsilon \in (0, 1]$. Then, the problem (12) possesses a unique solution $u \in H^{2}(\Omega) \cap H^{1}_{0}(\Omega)$. Moreover, the a priori estimate
\[
\|u\|_{H^{2}(\Omega)} \leq C_{a} \|f\|_{L^{2}(\Omega)}
\]

is fulfilled with some constant $C_{a} = C_{a}(d, \text{diam}(\Omega), \lambda_{E}, \|A\|_{L^{\infty}(\Omega)}, \varepsilon)$.

**Proof.** See [39, Theorem 3].

Notice that we restrict our discussion to problem (7) mainly for notational simplicity. For related investigations of equations involving also drift and potential terms, i.e., the differential operator of the PDE is of the form (2) with $b \neq 0$ and/or $c \neq 0$, we refer the reader, e.g., to [38].
3 Discretization

We decompose our domain $\Omega$ into a family of feasible triangulations $\mathcal{T}_h$ (triangular for $d = 2$, tetrahedral for $d = 3$) with discretization parameter $h = \max_{T \in \mathcal{T}_h} h_T$, $h_T := \text{diam}(T)$. The diameter of the largest inscribed ball in a cell $T \in \mathcal{T}_h$ is denoted by $\rho_T$. Throughout this article we assume that $\{\mathcal{T}_h\}_{h>0}$ is shape-regular, i.e., there holds

$$\kappa_T := \frac{h_T}{\rho_T} \leq \kappa \quad \forall T \in \mathcal{T}_h,$$

where the maximal aspect ratio $\kappa$ is independent of $h$. Moreover, meshes are considered which have a limited variation in the element size of neighboring elements, i.e., there is a constant $C_v > 0$ with $h_{T'} \leq C_v h_T$ for each $T, T' \in \mathcal{T}_h$, $T \cap T' \neq \emptyset$.

By $\mathcal{F}_h$ we denote the set of facets of $\mathcal{T}_h$ and by $n_F$ a unit normal vector on $F \in \mathcal{F}_h$. The normal vectors $n_F$ are chosen to point outwards if $F$ is a boundary facet and it has arbitrary but fixed orientation for interior facets. The diameter of a facet $F \in \mathcal{F}_h$ is denoted by $h_F$. Moreover, we denote the set of facets in the interior by $\mathcal{F}_h^I$. This includes all facets in the intersection of two elements in $\mathcal{T}_h$. Entities on either side of an interior facet are denoted by $\cdot^+$ and $\cdot^-$, respectively, chosen in such a way that for $F = \partial T^+ \cap \partial T^-$, $n_F$ points towards $T^+$.

By $\mathcal{P}_p(T)$, $T \in \mathcal{T}_h$, we denote the set of polynomials on $T$ of degree not larger than $p \in \mathbb{N}$. Throughout this article generic constants are denoted by $c_{v_1,v_2,...}$ where $v_1,v_2,...$, are the quantities they depend on.

Furthermore, we introduce the following average and jump operators. The average operators are defined by

$$\{\{u\}\}_F := \frac{1}{2}(u^+ + u^-) \quad \text{if} \quad F \in \mathcal{F}_h^I, \quad \{\{u\}\}_F := u \quad \text{if} \quad F \in \mathcal{F}_h \setminus \mathcal{F}_h^I.$$

In a similar way, we define the jump operators for matrix-valued functions $u \in H^1(\mathcal{T}_h; \mathbb{R}^{d \times d})$ and for vector-valued functions $v \in H^1(\mathcal{T}_h; \mathbb{R}^d)$ by

$$\llbracket u \rrbracket_F := u^+ n^+ + u^- n^-, \quad \llbracket v \rrbracket_F := v^+ \otimes n^+ + v^- \otimes n^-,$$

$$\llbracket u \rrbracket_F := u n, \quad \llbracket v \rrbracket_F := v \otimes n,$$

with $n^+$ and $n^-$ the outward unit normal vectors on $\partial T^+$ and $\partial T^-$. For scalar-valued functions we simply set $\llbracket u \rrbracket_F = u^+ - u^-$ for $F \in \mathcal{F}_h^I$ and $\llbracket u \rrbracket_F = u$ for $F \in \mathcal{F}_h \setminus \mathcal{F}_h^I$.

We will frequently use inverse inequalities and trace theorems in our analysis. These results are summarized in the following lemma.

**Lemma 3.** The following inequalities hold:

a) For given $0 \leq \ell \leq k \leq 1$ and $s, t \in [1, \infty]$ there exists some $C_{tr} > 0$ depending on $k, \ell, s, t, p, d$ and $\kappa$ such that the inequality

$$|v_h|_{W^{k,s}(T)} \leq C_{tr} h_T^{k-\ell} |T|^{1/s-1/t} |v_h|_{W^{\ell,t}(T)}, \quad \forall T \in \mathcal{T}_h,$$

is fulfilled for all $v_h \in \mathcal{P}_p(T)$.

b) For given $s \in [1, \infty]$ there exists some $C_{tr} > 0$ depending on $s, p, d$ and $\kappa$ such that the inequality

$$\|v_h\|_{L^s(F)} \leq C_{tr} h_T^{-1/s} \|v_h\|_{L^s(T)}, \quad \forall F \ni v \in \mathcal{T}_h,$$

is fulfilled for all $v_h \in \mathcal{P}_p(T)$.
Note that we use the same notation for both constants in the previous lemma as they depend on the same quantities.

For our analysis we need the following broken Sobolev spaces

\[ H^2_h(\Omega) := \{ v \in L^2(\Omega): v|_T \in H^2(T) \quad \forall T \in \mathcal{T}_h \}. \]

Moreover, we introduce a mesh-dependent norm for the space \( H^2_h(\Omega) \cap H^1_0(\Omega) \)

\[ \|v\|_{H^2_h(\Omega)}^2 := \sum_{T \in \mathcal{T}_h} \|\nabla^2 v\|_{L^2(T)}^2 + \sum_{F \in \mathcal{F}^+_h} h_F^{-1} \|\nabla v \cdot n_F\|_{L^2(F)}^2, \tag{13} \]

### 3.1 Approximation of the Hessian

Our discretization approach relies on a finite element approximation of the Hessian of \( u \) also referred to as Hessian recovery. For related ideas we refer to [24] and the references therein. In this article we study two different approaches. The first approach uses an approximation with \( C^0 \)-conforming finite elements. To illustrate the idea of the construction, consider the integration-by-parts formula for the second derivatives, i.e.,

\[ \int_\Omega \partial_{ij} u w \, dx = - \int_\Omega \partial_i u \partial_j w \, dx + \int_\Gamma \partial_i u w n_j \, ds \quad \forall w \in H^1(\Omega), \]

which is valid for all \( u \in H^2(\Omega) \) and \( i, j = 1, \ldots, d \). Here, \( n(x) = (n_1(x), \ldots, n_d(x))^\top \) denotes the outer unit normal vector on \( \Gamma \). Alternatively, one can use the more compact equivalent formulation

\[ \int_\Omega \nabla^2 u : w \, dx = - \int_\Omega \nabla u \cdot \text{Div} \, w \, dx + \int_\Gamma \nabla u \cdot (w n) \, ds \quad \forall w \in H^1(\Omega; \mathbb{R}^{d \times d}). \]

The Hessian approximation is sought in the finite-dimensional space

\[ \mathcal{W}^{CG}(\mathbb{R}^{d \times d}) := \{ w_h \in C(\overline{\Omega}; \mathbb{R}^{d \times d}): w_h|_T \in \mathcal{P}^{d \times d} \quad \forall T \in \mathcal{T}_h \}, \]

with polynomial degree \( p \in \mathbb{N} \). To shorten the notation we will omit the superscript \( (p) \), except when a different polynomial degree is used. The previous integral identity motivates the following definition.

**Definition 1** (Continuous Galerkin Hessian). *For each \( u \in H^2(\Omega) \), the discrete Hessian \( \mathbb{H}^{CG}(u) \in \mathcal{W}^{CG}(\mathbb{R}^{d \times d}) \) is defined by the variational problem*

\[ \int_\Omega \mathbb{H}^{CG}(u) : w_h \, dx = - \int_\Omega \nabla u \cdot \text{Div} \, w_h \, dx + \int_\Gamma \nabla u \cdot (w_h n) \, ds \quad \forall w_h \in \mathcal{W}^{CG}(\mathbb{R}^{d \times d}). \tag{14} \]

A further strategy is an approximation by piecewise polynomial but discontinuous functions. To this end, we define the space

\[ \mathcal{W}^{DG}(\mathbb{R}^{d \times d}) := \{ w_h \in L^\infty(\Omega; \mathbb{R}^{d \times d}): w_h|_T \in \mathcal{P}^{d \times d} \quad \forall T \in \mathcal{T}_h \}. \]

We obtain a Hessian approximation by discretizing the element-wise integration-by-parts formula

\[ \int_\Omega \partial_{ij} u w \, dx = \sum_{T \in \mathcal{T}_h} \left[ - \int_T \partial_i u \partial_j w \, dx + \int_{\partial T} \partial_i u w n_{\partial T,j} \, ds \right] \]

\[ = - \sum_{T \in \mathcal{T}_h} \int_T \partial_i u \partial_j w \, dx + \sum_{F \in \mathcal{F}^+_h} \int_F \{\partial_i u\} \|w\| n_{F,j} \, ds + \sum_{F \in \mathcal{F}_h \setminus \mathcal{F}^+_h} \int_F \partial_i u w n_j \, ds \]

which is valid for all \( u \in H^2(\Omega) \) and \( w \in H^1_0(\Omega) \). This motivates the following definition:
Lemma 4. Let this case we drop the superscript and simply write $H$ for all

\[ \int_{\Omega} H(u) : w_h \, dx = - \sum_{T \in T_h} \int_{T} \nabla u \cdot \text{Div} \, w_h \, dx + \sum_{F \in \mathcal{F}_h} \int_{F} \{ \nabla u \} \cdot [w_h] \, ds \quad (15) \]

for all $w_h \in W_h^{DG}(\mathbb{R}^{d \times d})$.

Many results in this article are independent of the choice of the Hessian approximation. In this case we drop the superscript and simply write $\mathbb{H}$ and $W_h$ which means either $\mathbb{H}^{CG}$ and $W_h^{CG}$ or $H_h^{DG}$ and $W_h^{DG}$.

We conclude this section with the following approximation result:

Lemma 4. Let $u \in H^2(\Omega)$ be given. The approximate Hessian $H(u)$ (either $\mathbb{H}^{CG}(u)$ or $H_h^{DG}(u)$) coincides with the $L^2(\Omega)$-projection of $\nabla^2 u$ onto $W_h(\mathbb{R}^{d \times d})$, i.e.,

\[ \| \nabla^2 u - H(u) \|_{L^2(\Omega)} = \inf_{W \in W_h(\mathbb{R}^{d \times d})} \| \nabla^2 u - W \|_{L^2(\Omega)}. \]

Moreover, there holds the stability estimate

\[ \| H(u) \|_{L^2(\Omega)} \leq c_d C_{tr} \| u \|_{H^2(\Omega)} \quad \forall u \in H^2(\Omega). \quad (16) \]

Proof. The desired result follows from the definition (14) and the integration-by-parts formula which yields

\[ \int_{\Omega} H(u) : w_h \, dx = - \int_{\Omega} \nabla u \cdot \text{Div} \, w_h \, dx + \int_{\Gamma} \nabla u \cdot (w_h \, n) \, ds = \int_{\Omega} \nabla^2 u : w_h \, dx \]

for all $w_h \in W_h^{CG}(\mathbb{R}^{d \times d})$. This implies that $H^{CG}(u)$ is the $L^2(\Omega)$-projection of $\nabla^2 u$ onto $W_h^{CG}(\mathbb{R}^{d \times d})$. To show the stability result we exploit the cell-wise integration-by-parts formula, taking into account the equality $\nabla u \cdot (w_h \, n) = (\nabla u \otimes n) : w_h$, the Cauchy-Schwarz inequality and the discrete trace Lemma 3. We obtain for each $w_h \in W_h^{CG}(\mathbb{R}^{d \times d})$

\[ \int_{\Omega} H(u) : w_h \, dx = - \int_{\Omega} \nabla u \cdot \text{Div} \, (w_h) \, dx + \int_{\Gamma} \nabla u \cdot (w_h \, n) \, ds \]

\[ = \sum_{T \in T_h} \int_{T} \nabla^2 u : w_h \, dx - \sum_{F \in \mathcal{F}_h} \int_{F} \| \nabla u \| : w_h \, ds \]

\[ \leq c_d \left( \sum_{T \in T_h} \| \nabla^2 u \|_{L^2(T)}^2 + C_{tr} \sum_{F \in \mathcal{F}_h} h_F^{-1} \| \nabla u \|_{L^2(F)}^2 \right)^{1/2} \| w_h \|_{L^2(\Omega)}. \]

With a simple computation taking into account that $u_h$ is continuous at the facets we deduce

\[ \| \nabla u \|_{L^2(F)} = \| \nabla u \cdot n_F \|_{L^2(F)}. \]

Finally, we test the previous inequality with $w_h = H^{CG}(u)$ and divide the left- and right-hand side by $\| H^{CG}(u) \|_{L^2(\Omega)}$ to conclude (16).

With similar arguments one can conclude the same results for the DG Hessian $H^{DG}$. The proof can be found in [31, Lemma 2.1].
3.2 A finite element scheme

The finite element approximations of our problem (10) are sought in the space of continuous Lagrange finite elements of order $p \in \mathbb{N}$, i.e.,

$$ \mathcal{V}_h := \{ v_h \in C(\Omega) : v_h|_T \in \mathcal{P}_p \quad \forall T \in \mathcal{T}_h \}, $$

and moreover, we define $\mathcal{V}_{h,0} = \mathcal{V}_h \cap H^1_0(\Omega)$ to incorporate essential boundary conditions. The polynomial degree $p \geq 2$ is the same as for the space $\mathcal{W}_h$. Later, we will see that this choice leads to an optimal balance of the approximation errors for the Hessian $\nabla^2 u$ and the solution $u$. Motivated by the strong formulation of the continuous problem (10) we test the discrete equations with the finite element Laplacian

$$ \mathbb{H}_\Delta(v) := \text{trace } \mathbb{H}(v) = \sum_{i=1}^d \mathbb{H}_{ii}(v) \in \mathcal{W}_h(\mathbb{R}). $$

The bilinear and linear forms we are going to use in the discrete scheme are defined by

$$ a_h(u_h, v_h) := \int_{\Omega} \gamma A : \mathbb{H}(u_h) \mathbb{H}_\Delta(v_h) \, dx, $$

$$ F_h(v_h) := \int_{\Omega} \gamma f \mathbb{H}_\Delta(v_h) \, dx. $$

The discrete problem reads

$$ \text{Find } u_h \in \mathcal{V}_{h,0} \text{ such that } a_h(u_h, v_h) + J_h(u_h, v_h) = F_h(v_h) \quad \forall v_h \in \mathcal{V}_{h,0}. \quad (17) $$

The bilinear form $J_h : \mathcal{V}_{h,0} \times \mathcal{V}_{h,0} \to \mathbb{R}$ may contain several stabilization terms in order to guarantee discrete coercivity. The specific form of the stabilization terms will be introduced later.

The nodal basis functions of $\mathcal{V}_{h,0}$ and $\mathcal{W}_h$ are denoted by

$$ V_{h,0} = \text{span}\{ \varphi_\ell \}_{\ell=1}^{N_V}, \quad \mathcal{W}_h = \text{span}\{ \psi_\ell \}_{\ell=1}^{N_W}. $$

For some function $u_h \in \mathcal{V}_{h,0}$ we denote by $u = (u_1, \ldots, u_{N_V})^\top$ the coefficient vector satisfying $u_h = \sum_{\ell=1}^{N_V} u_\ell \varphi_\ell$. Analogously, we use the vector representation $h_{ij} = (h_{ij,1}, \ldots, h_{ij,N_W})^\top$ for the Hessian approximations $\mathbb{H}(u_h)$, i.e., $H_{ij}(u_h) = \sum_{\ell=1}^{N_W} h_{ij,\ell} \psi_\ell$ for $i, j = 1, \ldots, d$.

To realize our algorithm with $\mathbb{H} = \mathbb{H}^{\text{CG}}$ we first assemble the matrices and load vector

$$ M_W = \left( \int_{\Omega} \psi_\ell \psi_k \right)_{k,\ell} \in \mathbb{R}^{N_W \times N_W}, \quad \text{(mass matrix in } \mathcal{W}_h) $$

$$ C_{ij} = \left( -\int_{\Omega} \frac{\partial \varphi_\ell}{\partial x_i} \frac{\partial \psi_k}{\partial x_j} + \int_{\Gamma} \frac{\partial \varphi_\ell}{\partial x_i} \frac{\partial \psi_k}{\partial n_j} \right)_{k,\ell} \in \mathbb{R}^{N_W \times N_V}, \quad \text{(partial mixed stiffness matrix)} $$

$$ B_{ij} = \left( \int_{\Omega} \gamma A_{ij} \psi_\ell \psi_k \right)_{k,\ell} \in \mathbb{R}^{N_W \times N_W}, \quad \text{(weighted mass matrix in } \mathcal{W}_h) $$

$$ S = J_h(\varphi_\ell, \varphi_\ell)_{k,\ell} \in \mathbb{R}^{N_V \times N_V}, \quad \text{(stabilization matrix)} $$

$$ f_W = \left( \int_{\Omega} \gamma f \psi_k \right)_k \in \mathbb{R}^{N_W}. \quad \text{(load vector w.r.t. } \mathcal{W}_h) $$

In the case $\mathbb{H} = \mathbb{H}^{\text{DG}}$ the matrices $C_{ij}$ have to be modified according to the right-hand side of (15). Moreover, the dimension of the matrices increases as the number of degrees of freedom $N_W$
is higher for the function space $\mathcal{W}_h^{DG}$. Obviously, the equations (14) or (15) with $u$ replaced by $u_h$ can be expressed by means of

$$M_W h_{ij} = C_{ij} u.$$ \hfill (18)

The application of $H_\Delta$ to the test function $v_h \in \mathcal{V}_h$ represented by the coefficient vector $v \in \mathbb{R}^{N_V \times N_V}$ leads to a new function $w_h := H_\Delta(v_h)$ with coefficient vector $w \in \mathbb{R}^{N_W}$ and can be computed by means of

$$M_W w = \sum_{i=1}^{d} C_{ii} v.$$ \hfill (19)

The right-hand side of (17) can be evaluated by means of

$$F_h(v_h) = \int_\Omega \gamma f H_\Delta(v_h) \, dx = v^\top \left( \sum_{i=1}^{d} C_{ii} \right)^\top M_W^{-1} f_W.$$ \hfill (20)

A representation for the left-hand side follows after insertion of (18)–(20) into (17). This yields

$$a_h(u_h, v_h) + J_h(u_h, v_h) = v^\top \left( \sum_{i=1}^{d} C_{ii} \right)^\top M_W^{-1} \left( \sum_{i,j=1}^{d} B_{ij} M_W^{-1} C_{ij} \right) u + v^\top S u.$$ \hfill (21)

Consequently, problem (17) is equivalent to

$$\left( \sum_{i=1}^{d} C_{ii} \right)^\top M_W^{-1} \left( \sum_{i,j=1}^{d} B_{ij} M_W^{-1} C_{ij} \right) + S \right) u = \left( \sum_{i=1}^{d} C_{ii} \right)^\top M_W^{-1} f_W =: f_V.$$ \hfill (22)

Although the system matrix cannot be assembled explicitly, one can compute matrix-vector products, each of which requires the solution of $d^2 + 1$ linear equation systems for the mass matrix $M_W$. In our numerical experiments we precomputed an LU factorization of $M_W$. Each evaluation of $M_W^{-1}$ then corresponds to an inexpensive forward-backward substitution. The nonsymmetric system (22) can be efficiently solved by a preconditioned GMRES algorithm. As a preconditioner we utilize the matrix

$$P = \left( \sum_{i=1}^{d} C_{ii} \right)^\top \hat{M}_W^{-1} \sum_{i,j=1}^{d} \hat{B}_{ij} \hat{M}_W^{-1} C_{ij} + S,$$ \hfill (23)

where $\hat{M}_W^{-1}$ is the inverse of the main diagonal of $M_W$ and $\hat{B}_{ij}$ is the main diagonal of $B_{ij}$. This allows us to assemble $P$ explicitly. Employing $\hat{B}_{ij}$ instead of $B_{ij}$ yields a sparser preconditioner and in case of a problem with vanishing off-diagonal entries of $A$, i.e., $A_{ij} = 0$ for $i \neq j$, a symmetric preconditioner $P$. A direct solver is then used to solve the systems of linear equations involving $P$. Note that it is not appropriate to use the lumped mass matrix as this might yield a singular matrix whenever the polynomial degree of the space $W_h$ is larger than one. The numerical experiments conducted in Section 4 indicate that the preconditioned GMRES method for (22) is robust with respect to mesh refinement. For a more sophisticated preconditioning strategy for non-divergence form PDEs we refer to [37], where a domain decomposition preconditioner is studied.

An alternative viable strategy is the solution of a block system equivalent to (22). This becomes particularly useful if, in addition to the solution vector $u$, one is interested in the finite-element

\[ \text{Page 10} \]
Hessian, e.g., for the solution of HJB equations. To this end, we use the substitution from (18) as well as

\[ M_W p = \sum_{i,j=1}^{d} B_{ij} h_{ij} \]

and arrive (in case \( d = 2 \)) at the equation system

\[
\begin{pmatrix}
M_W - C_{11} & -C_{12} \\
-C_{12} & C_{22} \\
-B_{11} & -B_{12} & -B_{21} & -B_{22} \\
C_{11} & C_{22} & S & \end{pmatrix}
\begin{pmatrix}
h_{11} \\
h_{12} \\
h_{21} \\
h_{22} \\
p \\
u \\
f_V \\
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
0 \\
0 \\
f_V \\
\end{pmatrix}
\]

equivalent to (22). The modification for the three-dimensional case is obvious.

### 3.3 Well-posedness of the discrete scheme

The scheme (17) can be interpreted as a non-conforming discretization of the variational problem (12) as the usage of approximate Hessians and Laplacians implies \( a \neq a_h \) and \( F \neq F_h \), and there also holds \( \mathcal{V}_{h,0} \not\subset X = H^2(\Omega) \cap H_0^1(\Omega) \).

**Lemma 5.** The inequality

\[ \| v \|^2_{H^2_h(\Omega)} \leq c_{\kappa,d} \sum_{T \in T_h} \left( h_T^{-2} \| \nabla v \|^2_{L^2(T)} + \| \nabla^2 v \|^2_{L^2(T)} \right) \]  

is valid for all \( v \in H^2_h(\Omega) \). Furthermore, \( \| \cdot \|_{H^2_h(\Omega)} \) is a norm in \( X = H^2(\Omega) \cap H_0^1(\Omega) \).

**Proof.** To show (24) we merely have to discuss the jump terms in the definition (13). To this end, we apply the triangle inequality

\[ \sum_{F \in \mathcal{F}_h^T} h_F^{-1} \| \nabla v \cdot n_F \|^2_{L^2(F)} \leq \sum_{F \in \mathcal{F}_h^T} \sum_{T \subset F} h_F^{-1} \| \nabla v \|_{L^2(F)}^2 \]  

and a trace theorem on a reference setting

\[ \| \nabla v \|_{L^2(F)} \leq c_\kappa \left( h_F^{-1/2} \| \nabla v \|_{L^2(T)} + h_F^{1/2} \| \nabla^2 v \|_{L^2(T)} \right). \]

Using also the assumed shape regularity, which implies \( c_\kappa^{-1} h_T \leq h_F \leq c_\kappa h_T \) for \( F \subset T \), we infer (24). The fact that \( \| \cdot \|_{H^2_h(\Omega)} \) is a norm in \( H^2(\Omega) \cap H^1_0(\Omega) \) follows from standard arguments.

The main ingredient for the proof of the existence result for strong solutions (Lemma 2) is a Miranda-Talenti estimate of the form \( \| u \|_{H^2(\Omega)} \leq \| \Delta u \|_{L^2(\Omega)} \) which is valid, e.g., if the underlying domain \( \Omega \) is convex. To show well-posedness of our discrete scheme we first have to prove a discrete Miranda-Talenti estimate. A similar result, but for a discretization using the element-wise exact Hessian and Laplacian, is proved in [34, Theorem 1]. We begin with the following auxiliary result.
Lemma 6. For each polynomial degree $p \in \mathbb{N}$, there exists a lifting operator $\mathbb{E}_{\mathbb{W}_{CG}^h} : \mathbb{W}_{DG}^h \to \mathbb{W}_{CG}^h$ satisfying the estimate
\[
\|v_h - \mathbb{E}_{\mathbb{W}_{CG}^h} (v_h)\|_{L^2(\Omega)} \leq C \left( \sum_{F \in \mathcal{F}_h^T} h_F \|v_h\|_{L^2(F)}^2 \right)^{1/2}
\]
for all $v_h \in \mathbb{W}_{DG}^h$, where $C > 0$ depends on $d, p, C_v, C_{tr}$ and $\kappa$.

Proof. The proof is similar to the proof of [26, Theorem 2.2] and [10, Theorem 2.2], but in both articles a slightly different setting is considered. For the convenience of the reader we repeat the proof.

We denote by $x_{T,i}$ the Lagrange points of the local finite element $(T, \mathcal{P}_p, \Sigma)$. That is, the functionals $\sigma_{T,i} \in \Sigma, i = 1, \ldots, s$, with $s \doteq \frac{1}{2} (p+1)(p+2)$ if $d = 2$ and $s \doteq \frac{1}{6} (p+1)(p+2)(p+3)$ if $d = 3$, have the form $\sigma_{T,i}(v) = v(x_{T,i})$. The sets $\mathcal{T}_h := \{T \in \mathcal{T}_h : x_{T,i} \in T\}$ contain all elements of $\mathcal{T}_h$ sharing the Lagrange point $x_{T,i}$ and we denote the number of cells in $\mathcal{T}_h$ by $\# \mathcal{T}_h$. In a similar way we define the patch of facets $\mathcal{F}_{T,i} := \{F \in \mathcal{F}_h^T : x_{T,i} \in F\}$ having $x_{T,i}$ as a vertex. Moreover, $\{\varphi_{T,i}\}_{i=1}^s$ is the nodal basis of $(T, \mathcal{P}_p, \Sigma)$, i.e., $\sigma_{T,i}(\varphi_{T,i}) = \delta_{ij}$ holds for all $i,j = 1, \ldots, s$. The precise definition of our lifting operator is
\[
\mathbb{E}_{\mathbb{W}_{CG}^h} (v_h)|_T := \sum_{i=1}^s \left( \frac{1}{\# \mathcal{T}_{T,i}} \sum_{T' \in \mathcal{T}_{T,i}} (v_h|_{T'}) (x_{T,i}) \right) \varphi_{T,i}.
\]
Next, we derive local estimates for the lifting error on a single element $T \in \mathcal{T}_h$. From the definition of $\mathbb{E}_{\mathbb{W}_{CG}^h}$ and the triangle inequality we conclude
\[
\|v_h - \mathbb{E}_{\mathbb{W}_{CG}^h} (v_h)\|_{L^2(T)}^2 = \int_T \left( \sum_{i=1}^s \left( (v_h|_T)(x_{T,i}) - \frac{1}{\# \mathcal{T}_{T,i}} \sum_{T' \in \mathcal{T}_{T,i}} (v_h|_{T'})(x_{T,i}) \right) \varphi_{T,i} \right)^2 dx \leq \int_T \left( \sum_{i=1}^s \frac{1}{\# \mathcal{T}_{T,i}} \sum_{T' \in \mathcal{T}_{T,i}} \|v_h|_{T'} - v_h|_{T'}(x_{T,i}) \varphi_{T,i} \right)^2 dx. \tag{26}
\]
We distinguish several cases: if $x_{T,i}$ is a Lagrange point in the interior of $T$ or in the interior of a boundary facet $F \in \mathcal{F}_h$ with $F \subset \Gamma$, then $\mathcal{T}_{T,i} = \{T\}$ holds and consequently
\[
\sum_{T' \in \mathcal{T}_{T,i}} \|v_h|_{T'} - v_h|_{T'}(x_{T,i}) = 0.
\]
If $x_{T,i}$ is located in the interior of an inner facet $F \in \mathcal{F}_h^T$, there holds $\mathcal{T}_{T,i} = \{T,T'\}$ with $F = T \cap T'$ and we obtain together with the inverse inequality from Lemma 3
\[
\|v_h|_T - v_h|_{T'}(x_{T,i}) \leq \|v_h\|_{L^\infty(F)} \leq C_{tr} h_F^{-(d-1)/2} \|v_h\|_{L^2(F)}.
\]
If $x_{T,i}$ coincides with a vertex of $T$ or, in the case $d = 3$, is located at an edge of $T$, we choose a sequence of simplices $T = T_1, T_2, \ldots, T_\ell = T'$ such that $T_j$ and $T_{j+1}, j = 1, \ldots, \ell - 1$, share a common facet $F_j \in \mathcal{F}_{T,j}$. With the triangle inequality and similar arguments like in the previous case we deduce
\[
\|v_h|_T - v_h|_{T'}(x_{T,i}) \leq \sum_{j=1}^{\ell-1} \|v_h|_{T_j} - v_h|_{T_{j+1}}(x_{T,i}) \leq C_{tr} \sum_{F \in \mathcal{F}_{T,i}} h_F^{-(d-1)/2} \|v_h\|_{L^2(F)}.
\]
12
We summarize the previous cases and infer
\[
\frac{1}{\# F_{T,i}} \sum_{T \in F_{T,i}} |v_h|_T - v_h|_T^T (x_{T,i}) \leq C_{tr} \sum_{F \in F_{T,i}} h_F^{- (d - 1)/2} \|v_h\|_{L^2(F)}.
\]
Insertion into (26) yields together with the discrete Cauchy-Schwarz inequality
\[
\|v_h - E_{W^h} \chi(v_h)\|_{L^2(T)}^2 \leq C_F s C_{tr}^2 \sum_{i=1}^s \sum_{F \in F_{T,i}} h_F^{- (d - 1)} \|\|v_h\|_{L^2(F)} \int_T \varphi_{T,i}^2 \, dx
\]
\[
\leq c_{d,k} s C_{tr}^2 c_v^{d} \sum_{i=1}^s \sum_{F \in F_{T,i}} h_F \|\|v_h\|_{L^2(F)}^2,
\]
with \( C_F := \max_{T \in \mathcal{T}_h} \max_{i=1,...,s} \# F_{T,i} \leq c_{d,k} \). The last step follows from \( \int_T \varphi_{T,i}^2 \, dx \leq |T| \leq c_{\kappa} h_T^2 \leq c_{\kappa} C_v^d h_T^2 \). Summation over all \( T \in \mathcal{T}_h \) leads to the assertion. \( \square \)

**Lemma 7** (Discrete Miranda-Talenti estimate). Let \( \Omega \subset \mathbb{R}^d \) be a bounded and convex domain. The polynomial degree of \( V_h \) and \( W_h \) is \( p \geq 2 \). There exist constants \( C_1, C_2 > 0 \) depending on \( C_{tr}, C_v, \kappa \) and \( d \) and, if \( d = 3 \), also on \( \Omega \), such that for each \( u_h \in V_h \) the inequalities
\[
\|\mathbb{H}_r^{CG}(u_h)\|_{L^2(\Omega)}^2 \leq \|\mathbb{H}_r^{CG}(u_h)\|_{L^2(\Omega)}^2 + C_1 \sum_{F \in F_h} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2
\]
\[
+ C_2 \sum_{F \in F_h} h_F \|\nabla^2 u_h\|_{L^2(F)}^2,
\]
\[
\|\mathbb{H}_r^{DG}(u_h)\|_{L^2(\Omega)}^2 \leq \|\mathbb{H}_r^{DG}(u_h)\|_{L^2(\Omega)}^2 + C_1 \sum_{F \in F_h} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2.
\]
are fulfilled.

**Proof.** We first introduce a further lifting operator \( \mathbb{E}_h : V_h \rightarrow V_{h,\text{conf}} \) which maps \( u_h \) into an \( H^2 \)-conforming finite element space \( V_{h,\text{conf}} \). In the case \( d = 2 \), we will make use of the space generated by the Hsieh-Clough-Tocher (HCT) element [11] or some higher-order analogue. The lifting operator \( \mathbb{E}_h \) fulfills the estimate
\[
\|\nabla_k^k (u_h - \mathbb{E}_h(u_h))\|_{L^2(\Omega)} \leq C_E \left( \sum_{F \in F_h} h_F^{2k-2}\|\nabla u_h \cdot n_F\|_{L^2(F)}^2 \right)^{1/2}, \quad k = 0, 1, 2,
\]
with a constant \( C_E > 0 \) depending on \( C_v, C_{tr}, \kappa \) and \( d \), but for the case \( d = 3 \) also on the structure of \( \Omega \). In particular, if an opening angle at a sharp edge \( \alpha \) of \( \Omega \) tends to \( \pi \), then \( C_E \rightarrow \infty \).

A proof of (29) in the two-dimensional case can be found in [42, Section 4.11.3], [5, Equation (2.9)] for the case \( p = 2 \) and in [22, Lemma 3.1] for \( p \geq 2 \). For the three-dimensional case we refer to [34], where a 3D HCT element for polynomial degrees \( p \in \{2, 3\} \) is studied and to [9], where a different function space based on virtual elements of arbitrary order is used.

We set \( \widetilde{u}_h := \mathbb{E}_h(u_h) \) and obtain with the triangle inequality
\[
\|\mathbb{H}(u_h)\|_{L^2(\Omega)} \leq \|\nabla^2 \widetilde{u}_h\|_{L^2(\Omega)} + \|\nabla_k^k \widetilde{u}_h - \nabla_k^k u_h\|_{L^2(\Omega)} + \|\nabla_k^k u_h - \mathbb{H}(u_h)\|_{L^2(\Omega)}.
\]
For the first term on the right-hand side we can directly apply the continuous Miranda-Talenti estimate from [39, Theorem 2]. After insertion of further intermediate functions we obtain
\[
\|\nabla^2 \tilde{u}_h\|_{L^2(\Omega)} \leq \|\Delta \tilde{u}_h\|_{L^2(\Omega)} \\
\leq \|H(\tilde{u}_h)\|_{L^2(\Omega)} + \|H_{\Delta}(u_h) - \Delta u_h\|_{L^2(\Omega)} + \|\Delta u_h - \Delta_h \tilde{u}_h\|_{L^2(\Omega)}.
\]
It remains to bound the two last terms on the right-hand sides of (30) and (31). From the error estimate (29) we infer
\[
\|\Delta u_h - \Delta_h \tilde{u}_h\|_{L^2(\Omega)} + \|\nabla^2 \tilde{u}_h - \nabla^2 H u_h\|_{L^2(\Omega)} \leq C_E \left( \sum_{F \in \mathcal{T}^2_h} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2 \right)^{1/2}.
\]
In order to prove a bound for the approximations $H(\tilde{u}_h)$ of $\Delta u_h$ and $H(u_h)$ of $\nabla^2 u_h$ we introduce the $L^2(\Omega)$-projection $P_{\mathcal{V}_h}$ onto $\mathcal{W}_h(\mathbb{R}^{d \times d})$ and obtain
\[
\|H(u_h) - \nabla^2 H u_h\|_{L^2(\Omega)} \leq \|H(u_h) - P_{\mathcal{V}_h}(\nabla^2 u_h)\|_{L^2(\Omega)} + \|P_{\mathcal{V}_h}(\nabla^2 u_h) - \nabla^2 H u_h\|_{L^2(\Omega)}.
\]
To bound the first term on the right-hand side of (33) in case of $H = H^{CG}$ we test (14) with the function $w_h := H^{CG}(u_h) - P_{\mathcal{V}_h}(\nabla^2 u_h) \in \mathcal{W}_h^{CG}(\mathbb{R}^{d \times d})$, apply the orthogonality of $P_{\mathcal{V}_h}$, the definition of $H^{CG}$, the integration-by-parts formula and Lemma 3 to arrive at
\[
\|H^{CG}(u_h) - P_{\mathcal{V}_h}(\nabla^2 H u_h)\|_{L^2(\Omega)}^2 = \int_{\Omega} (H^{CG}(u_h) - \nabla^2 H u_h) : w_h \, dx \\
= \int_{\Gamma} \nabla u_h \cdot (w_h n) \, ds - \sum_{T \in \mathcal{T}_h} \int_{\partial T} \nabla u_h \cdot (w_h n_{\partial T}) \, ds \\
= - \sum_{F \in \mathcal{T}^2_h} \int_F \|\nabla u_h\| : w_h \, ds \leq C_E \left( \sum_{F \in \mathcal{T}^2_h} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2 \right)^{1/2} \|w_h\|_{L^2(\Omega)}.
\]
Note that we used the relation $w_h : (n_F \otimes \nabla u_h) = \nabla u_h \cdot (w_h n_F)$ as well as the fact that the jumps in tangential direction $\|\nabla u_h \cdot t_F\|$ vanish as $u_h$ is continuous along the facets $F$.

In the case $H = H^{DG}$ we use similar arguments, in particular the integration-by-parts formula and (15), to obtain
\[
\|H^{DG}(u_h) - P_{\mathcal{V}_h}(\nabla^2 H u_h)\|_{L^2(\Omega)}^2 = \int_{\Omega} (H^{DG}(u_h) - \nabla^2 H u_h) : w_h \, dx \\
= \sum_{F \in \mathcal{T}^2_h} \int_F \{\nabla u_h\} : \{w_h\} \, ds - \sum_{T \in \mathcal{T}_h} \int_{\partial T} \nabla u_h \cdot (w_h n_{\partial T}) \, ds \\
= - \sum_{F \in \mathcal{T}^2_h} \int_F \|\nabla u_h\| : \{w_h\} \, ds \leq C_E \left( \sum_{F \in \mathcal{T}^2_h} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2 \right)^{1/2} \|w_h\|_{L^2(\Omega)}.
\]
an estimate from Lemma 6 and the property \( \nabla_h^2 u_h \in \mathcal{W}_h^{DG(p-2)}(\mathbb{R}^{d \times d}) \subset \mathcal{W}_h^{DG}(\mathbb{R}^{d \times d}) \), i.e.,

\[
\| P_{\mathcal{W}_h^{DG}}(\nabla_h^2 u_h) - \nabla_h^2 u_h \|_{L^2(\Omega)} \leq \| E_{\mathcal{W}_h^{DG}}(\nabla_h^2 u_h) - \nabla_h^2 u_h \|_{L^2(\Omega)} \\
\leq C \left( \sum_{F \in \mathcal{F}_h^2} h_F \sum_{i,j=1}^d \| \partial_{ij} u_h \|_{L^2(F)}^2 \right)^{1/2} \\
\leq C \left( \sum_{F \in \mathcal{F}_h^2} h_F \| \nabla^2 u_h \|_{L^2(F)}^2 \right)^{1/2}. \tag{36}
\]

Note that the jump operator for matrix-valued functions involves only jumps in normal direction. In order to confirm the last step in the previous estimate, one just has to take into account that \( u_h \) is continuous along the element facets so that the tangential components of the jumps vanish.

Finally, one observes that the second term on the right-hand side of (33) vanishes in case of \( H_0 = H_{DG} \), i.e.,

\[
\| P_{\mathcal{W}_h^{DG}}(\nabla_h^2 u_h) - \nabla_h^2 u_h \|_{L^2(\Omega)} = 0,
\tag{37}
\]

which is due to the fact that \( \nabla_h^2 u_h \in \mathcal{W}_h^{DG}(\mathbb{R}^{d \times d}) \) holds.

The discrete Miranda-Talenti estimates follow after inserting (34) and (36) in case of \( H_0 = H^{CG} \), or (35) and (37) in case of \( H_0 = H_{DG} \), into (33), and combining the resulting estimates with (30) and (31).

The following result is needed in order to treat the fact that we work with different norms for the spaces \( \mathcal{V}_h \) which have to be bounded by each other.

**Lemma 8.** The following estimates are valid for arbitrary \( u_h \in \mathcal{V}_h \):

\[
\| H_{DG}^\Delta(u_h) \|_{L^2(\Omega)}^2 \geq \frac{1}{3} \| \Delta_h u_h \|_{L^2(\Omega)}^2 - \frac{d+1}{3} C_{tr}^2 \sum_{F \in \mathcal{F}_h^2} h_F^{-1} \| \nabla u_h \cdot n_F \|_{L^2(F)}^2 - \frac{d+1}{6} C_{E}^2 \sum_{F \in \mathcal{F}_h^2} h_F \| \Delta u_h \|_{L^2(F)}^2,
\]

\[
\| H_{DG}^\nabla(u_h) \|_{L^2(\Omega)}^2 \geq \frac{1}{3} \| \Delta_h u_h \|_{L^2(\Omega)}^2 - \frac{d+1}{3} C_{tr}^2 \sum_{F \in \mathcal{F}_h^2} h_F^{-1} \| \nabla u_h \cdot n_F \|_{L^2(F)}^2.
\]

**Proof.** We start with the case \( H_0 = H_{DG} \). We apply the definition (15) and the integration-by-parts formula and obtain for arbitrary \( w_h \in \mathcal{W}_h^{DG} \)

\[
(H_{DG}^\Delta(u_h), w_h)_{L^2(\Omega)} = - (\nabla_h u_h, \nabla_h w_h)_{L^2(\Omega)} + \sum_{F \in \mathcal{F}_h} \int_F \{\nabla u_h \cdot n_F\} \{w_h\} \, ds \\
= (\Delta_h u_h, w_h)_{L^2(\Omega)} - \sum_{F \in \mathcal{F}_h} \int_F [\nabla u_h \cdot n_F] \{w_h\} \, ds.
\]

Testing this equation with \( w_h = H_{DG}^\Delta(u_h) \in \mathcal{W}_h^{DG} \) yields together with a further application of
the integration-by-parts formula

\[
\| \mathbb{H}_\Delta^{DG}(u_h) \|_{L^2(\Omega)}^2 = (\Delta_h u_h, \mathbb{H}_\Delta^{DG}(u_h))_{L^2(\Omega)} - \sum_{F \in \mathcal{F}_h^2} \int_F \| \nabla u_h \cdot n_F \| \{ \mathbb{H}_\Delta^{DG}(u_h) \} \, ds
\]

\[
= - (\nabla_h \Delta_h u_h, \nabla_h u_h)_{L^2(\Omega)} - \sum_{F \in \mathcal{F}_h^2} \int_F (\| \nabla u_h \cdot n_F \| \{ \mathbb{H}_\Delta^{DG}(u_h) \} - \{ \nabla u_h \cdot n_F \} \{ \Delta u_h \} ) \, ds
\]

\[
+ \sum_{F \in \mathcal{F}_h \setminus \mathcal{F}_h^2} \int_F \nabla u_h \cdot n_F \Delta u_h \, ds
\]

\[
= (\Delta_h u_h, \Delta_h u_h)_{L^2(\Omega)} - \sum_{F \in \mathcal{F}_h^2} \int_F (\| \nabla u_h \cdot n_F \| \{ \mathbb{H}_\Delta^{DG}(u_h) \} - \{ \nabla u_h \cdot n_F \} \{ \Delta u_h \} ) \, ds
\]

\[
+ \sum_{F \in \mathcal{F}_h \setminus \mathcal{F}_h^2} \int_F \nabla u_h \cdot n_F \Delta u_h \, ds - \sum_{F \in \mathcal{F}_h} \int_F \| \nabla u_h \cdot n_F \| \{ \Delta u_h \} \, ds.
\]

Together with the identity

\[
\| \nabla u_h \cdot n_F \| \{ \Delta u_h \} = \{ \nabla u_h \cdot n_F \} \| \Delta u_h \| + \| \nabla u_h \cdot n_F \| \{ \Delta u_h \}
\]

for all \( F \in \mathcal{F}_h^2 \) we arrive at

\[
\| \mathbb{H}_\Delta^{DG}(u_h) \|_{L^2(\Omega)}^2 = \| \Delta_h u_h \|_{L^2(\Omega)}^2 - \sum_{F \in \mathcal{F}_h^2} \int_F \| \nabla u_h \cdot n_F \| \{ \mathbb{H}_\Delta^{DG}(u_h) \} + \{ \Delta u_h \} \) \, ds.
\]

With the Cauchy-Schwarz and the Young inequality using also the discrete trace inequality from Lemma 3, we then deduce

\[
\| \mathbb{H}_\Delta^{DG}(u_h) \|_{L^2(\Omega)}^2 \geq \| \Delta_h u_h \|_{L^2(\Omega)}^2 - \xi \sum_{F \in \mathcal{F}_h^2} h_F^{-1} \| \| \nabla u_h \cdot n_F \| \|_{L^2(F)}^2
\]

\[
- \frac{C^2 tr}{d+1} \xi \left( \| \mathbb{H}_\Delta^{DG}(u_h) \|_{L^2(\Omega)}^2 + \| \Delta u_h \|_{L^2(\Omega)}^2 \right)
\]

for arbitrary \( \xi > 0 \). We use the choice \( \xi = C^2 tr / (d+1)/2 \) and after rearrangement of the above inequality we arrive at

\[
\| \mathbb{H}_\Delta^{DG}(u_h) \|_{L^2(\Omega)}^2 \geq \frac{1}{3} \| \Delta_h u_h \|_{L^2(\Omega)}^2 - \frac{1}{3} (d+1) C^2 tr \sum_{F \in \mathcal{F}_h^2} h_F^{-1} \| \nabla u_h \cdot n_F \|_{L^2(F)}^2.
\]

(38)

In a similar way we derive the estimate for \( \mathbb{H}_\Delta = \mathbb{H}_C^{CG} \). First, we use the definition (6) and the integration-by-parts formula and obtain for each \( w_h \in \mathcal{V}_h^{CG} \)

\[
(\mathbb{H}_\Delta^{CG}(u_h), w_h)_{L^2(\Omega)} = -(\nabla u_h, \nabla w_h)_{L^2(\Omega)} + \int_F \nabla u_h \cdot n_F w_h \, ds
\]

\[
= (\Delta_h u_h, w_h)_{L^2(\Omega)} - \sum_{F \in \mathcal{F}_h} \int_F \| \nabla u_h \cdot n_F \| \{ w_h \} \, ds.
\]
From this we infer with the properties of $P$ further application of the definition (6) and the integration-by-parts formula of the estimates (27) and (28) the proof of the coercivity of the bilinear form $a$ well-posedness of our discrete scheme. However, due to the jump terms on the right-hand sides where we exploited that $w_h$ is continuous across the element facets. We choose the test function $w_h = \mathbb{H}^{CG}(u_h)$, use the orthogonality of the $L^2(\Omega)$-projection $P_{W_k}^{CG}$ onto $W^{CG}$ and get with a further application of the definition (6) and the integration-by-parts formula

$$\|H^{\Delta}(u_h)\|^2_{L^2(\Omega)} = (P_{W_k}^{CG}(\Delta_h u_h), H^{\Delta}(u_h))_{L^2(\Omega)} - \sum_{F \in F_h} \int_F \|\nabla u_h \cdot n_F\| H^{\Delta}(u_h) \, ds$$

$$= - (\nabla P_{W_k}^{CG}(\Delta_h u_h), \nabla u_h)_{L^2(\Omega)} - \sum_{F \in F_h} \int_F \|\nabla u_h \cdot n_F\| H^{\Delta}(u_h) \, ds$$

$$+ \int_{\Omega} \nabla u_h \cdot n_F P_{W_k}^{CC}(\Delta_h u_h) \, ds$$

$$= (P_{W_k}^{CG}(\Delta_h u_h), \Delta_h u_h)_{L^2(\Omega)} - \sum_{F \in F_h} \|\nabla u_h \cdot n_F\| (H^{\Delta}(u_h) + P_{W_k}^{CG}(\Delta_h u_h)) \, ds.$$ From this we infer with the properties of $P_{W_k}^{CG}$ and similar arguments as in (38)

$$\|H^{\Delta}(u_h)\|^2_{L^2(\Omega)} \geq \frac{1}{3} \|P_{W_k}^{CG}(\Delta_h u_h)\|^2_{L^2(\Omega)} - \frac{d+1}{3} C_{tr} \sum_{F \in F_h} h_F^{-1} \|\nabla u_h \cdot n_F\|^2_{L^2(F)}.$$ Furthermore, with the property $\|P_{W_k}^{CG} v\|^2_{L^2(\Omega)} = \|v\|^2_{L^2(\Omega)} + \|v - P_{W_k}^{CG} v\|^2_{L^2(\Omega)}$ and the estimate (36) we get

$$\|H^{\Delta}(u_h)\|^2_{L^2(\Omega)} \geq \frac{1}{3} \|\Delta_h u_h\|^2_{L^2(\Omega)} - \frac{d+1}{3} C_{tr} \sum_{F \in F_h} h_F^{-1} \|\nabla u_h \cdot n_F\|^2_{L^2(F)}$$

$$- \frac{d+1}{6} C_E \sum_{F \in F_h} h_F \|\Delta u_h\|^2_{L^2(F)}.$$}

Next, we want to mimic the proof of Lemma 2 for the continuous setting in order to show well-posedness of our discrete scheme. However, due to the jump terms on the right-hand sides of the estimates (27) and (28) the proof of the coercivity of the bilinear form $a_h$ will fail if $\varepsilon$ is too small, see (11). To this end, stabilization terms in the discrete scheme are needed and we define

$$J_h(u_h, v_h) := \eta_1 \sum_{F \in F_h} h_F^{-1} \int_F \|\nabla u_h \cdot n_F\| \|\nabla v_h \cdot n_F\| \, ds + \eta_2 \sum_{F \in F_h} h_F \int_F \|\nabla^2 u_h\| \|\nabla^2 v_h\| \, ds \quad (39)$$

to be inserted into (17). The penalty parameters $\eta_1, \eta_2 \geq 0$ have to be chosen appropriately to guarantee the coercivity of $a_h + J_h$. For the stabilized scheme one can show the following well-posedness result.

**Lemma 9.** Let $\Omega \subset \mathbb{R}^d$ be a bounded and convex domain. Let $C_1$ and $C_2$ be the constants from Lemma 7, where we set $C_2 = 0$ in case of $\mathbb{H} = \mathbb{H}^{DG}$. Assume that $A$ fulfills the Cordes condition (11) with a constant $\varepsilon \in (0, 1]$ and that the polynomial degree of $V_h$ and $W_h$ is $p \geq 2$. Then the
bilinear form $a_h + J_h$ is bounded and uniformly elliptic, i.e., there exist constants $\alpha_0, \beta_0$ such that the inequalities
\begin{align*}
a_h(u_h, v_h) + J_h(u_h, v_h) &\leq \beta_0 \| u_h \|_{H^2_h(\Omega)} \| v_h \|_{H^2_h(\Omega)} \quad \forall u_h, v_h \in \mathcal{V}_h, \tag{40} \\
a_h(u_h, u_h) + J_h(u_h, u_h) &\geq \varepsilon \alpha_0 \| u_h \|_{H^2_h(\Omega)}^2 \quad \forall u_h \in \mathcal{V}_h, \tag{41}
\end{align*}
are fulfilled, provided that the penalty parameters in $J_h$ fulfill the inequalities
\begin{equation}
\eta_1 \geq \frac{(1 - \varepsilon) C_1}{2} + \frac{\varepsilon}{6} (1 + (d + 1) C^2_\text{tr}) \quad \eta_2 \geq \frac{(1 - \varepsilon) C_2}{2} + \frac{d + 1}{12} \varepsilon C^2_\text{tr}. \tag{42}
\end{equation}
The constants $\alpha_0$ and $\beta_0$ depend on $\kappa, d, p, C_{\text{tr}}, C_v$ and in the case $d = 3$ on the geometry of $\Omega$, but not on $\varepsilon$ and $h$.

As a consequence, problem (17) possesses a unique solution $u_h \in \mathcal{V}_h$ for each $f \in L^2(\Omega)$.

**Proof.** First, we show the boundedness of $a_h + J_h$. With the Cauchy-Schwarz inequality and (16) we obtain
\begin{equation}
a_h(u_h, v_h) \leq c_d C^2_\text{tr} \| \gamma A \|_{L^\infty(\Omega)} \| u_h \|_{H^2_h(\Omega)} \| v_h \|_{H^2_h(\Omega)}. \tag{43}
\end{equation}
With Lemma 1 and the assumed Cordes condition (11) we moreover conclude $\| \gamma A \|_{L^\infty(\Omega)} \leq 1 + \sqrt{1 - \varepsilon} \leq 2$. To derive a similar estimate for the stabilization term $J_h$ we apply the Cauchy-Schwarz inequality on each inner facet $F \in \mathcal{F}^2_h$, and for the second term in $J_h$ we additionally employ the discrete trace theorem from Lemma 3 $\| \nabla^2 u_h \|_{L^2(F)} \leq C_{\text{tr}} h^{-1/2}_F \| \nabla^2 u_h \|_{L^2(T_F)}$, where $T_F \in \mathcal{T}_h$ is an arbitrary element with $F \subset T_F$. This yields
\begin{equation}
\sum_{F \in \mathcal{F}^2_h} h^{-1}_F \int_F \| \nabla u_h \cdot n_F \| \| \nabla v_h \cdot n_F \| \, ds \leq \| u_h \|_{H^2_h(\Omega)} \| v_h \|_{H^2_h(\Omega)}, \tag{44}
\end{equation}
from which we deduce
\begin{equation}
J(u_h, v_h) \leq (\eta_1 + \eta_2 C^2_\text{tr} (d + 1)) \| u_h \|_{H^2_h(\Omega)} \| v_h \|_{H^2_h(\Omega)}. \tag{45}
\end{equation}
The inequalities (43) and (45) lead to (40).

The coercivity follows from Lemma 7, taking into account the Cordes condition (11) with the estimate from Lemma 1 and Young’s inequality with weight $\xi > 0$, i.e.,
\begin{align*}
\int_{\Omega} & \gamma A : \mathbb{H}(u_h) \mathbb{H}(u_h) \, dx = \| \mathbb{H}(u_h) \|_{L^2(\Omega)}^2 + \int_{\Omega} (\gamma A - I) : \mathbb{H}(u_h) \mathbb{H}(u_h) \, dx \\
&\geq \| \mathbb{H}(u_h) \|_{L^2(\Omega)}^2 - \sqrt{1 - \varepsilon} \| \mathbb{H}(u_h) \|_{L^2(\Omega)} \\
&\times \left( \| \mathbb{H}(u_h) \|_{L^2(\Omega)}^2 + C_1 \sum_{F \in \mathcal{F}^2_h} h^{-1}_F \| \nabla u_h \cdot n_F \|_{L^2(F)}^2 + C_2 \sum_{F \in \mathcal{F}^2_h} h_F \| \nabla^2 u_h \|_{L^2(F)}^2 \right)^{1/2} \\
&\geq \left( 1 - \frac{\xi (1 - \varepsilon)}{2} \right) \| \mathbb{H}(u_h) \|_{L^2(\Omega)}^2 \\
&\quad - \frac{1}{2\xi} \left( C_1 \sum_{F \in \mathcal{F}^2_h} h^{-1}_F \| \nabla u_h \cdot n_F \|_{L^2(F)}^2 + C_2 \sum_{F \in \mathcal{F}^2_h} h_F \| \nabla^2 u_h \|_{L^2(F)}^2 \right). \tag{46}
\end{align*}
The jump terms in (46) can be canceled by the stabilization terms from (39). We use the choice \( \xi = (1 - \varepsilon)^{-1} \) and insert the estimate from Lemma 8 to arrive at

\[
\begin{align*}
  a_h(u_h, u_h) + J_h(u_h, u_h) &\geq \frac{\varepsilon}{2} \|H(u_h)\|^2_{L^2(\Omega)} + \left( \eta_1 - \frac{(1 - \varepsilon)}{2} \right) \sum_{F \in F_h^T} h_F^{-1} \|\nabla u_h \cdot n_F\|^2_{L^2(F)} \\
  &+ \left( \eta_2 - \frac{(1 - \varepsilon)}{2} \right) \sum_{F \in F_h^T} h_F \|\nabla^2 u_h\|^2_{L^2(F)} \\
  &\geq \frac{\varepsilon}{6} \|\Delta_h u_h\|^2_{L^2(\Omega)} + \left( \eta_1 - \frac{(1 - \varepsilon)}{2} - \frac{d + 1}{6} \varepsilon C_{\text{tr}}^2 \right) \sum_{F \in F_h^T} h_F^{-1} \|\nabla u_h \cdot n_F\|^2_{L^2(F)} \\
  &+ \left( \eta_2 - \frac{(1 - \varepsilon)}{2} - \frac{d + 1}{12} \varepsilon C_{\text{tr}}^4 \right) \sum_{F \in F_h^T} h_F \|\nabla^2 u_h\|^2_{L^2(F)}.
\end{align*}
\]

Taking into account the assumptions (42) we may further estimate

\[
a_h(u_h, u_h) + J_h(u_h, u_h) \geq \frac{\varepsilon}{6} \left( \|\Delta_h u_h\|^2_{L^2(\Omega)} + \sum_{F \in F_h^T} h_F^{-1} \|\nabla u_h \cdot n_F\|^2_{L^2(F)} \right).
\]

The right-hand side forms a norm on \( H^1_0(\Omega) \cap H_0^1(\Omega) \) which is equivalent to the norm defined in (13). This is a consequence of a Miranda-Talenti estimate for the broken Hessian, see [34].

The Lax-Milgram Lemma finally implies the existence and uniqueness of a discrete solution \( u_h \in V_{h,0} \) of (17). \( \square \)

**Remark 1.** The assumption (42) can be relaxed such that the choice \( \eta_1 = \eta_2 = 0 \) is also feasible. This requires the following modification in the proof of the previous theorem. As \( \|H(\cdot)\|^2_{L^2(\Omega)} \) is also a norm in the finite-dimensional space \( V_{h,0} \), there exists a constant \( C_{\text{tr}} \) independent of \( \varepsilon \) such that the estimate \( \|H(\cdot)\|^2_{L^2(\Omega)} \geq C_{\text{tr}} \|u_h\|^2_{H^2(\Omega)} \) is valid for all \( u_h \in V_{h,0} \). Using this estimate and (44) we can modify the last step in (47) to arrive at

\[
\begin{align*}
  a_h(u_h, u_h) + J_h(u_h, u_h) &\geq \frac{\varepsilon}{2} \|H(u_h)\|^2_{L^2(\Omega)} + \left( \eta_1 - \frac{(1 - \varepsilon)}{2} \right) \sum_{F \in F_h^T} h_F^{-1} \|\nabla u_h \cdot n_F\|^2_{L^2(F)} \\
  &+ \left( \eta_2 - \frac{(1 - \varepsilon)}{2} \right) \sum_{F \in F_h^T} h_F \|\nabla^2 u_h\|^2_{L^2(F)} \\
  &\geq \frac{\varepsilon}{4} \|H(u_h)\|^2_{L^2(\Omega)} + \left( \eta_1 - \frac{(1 - \varepsilon)}{2} + \frac{\varepsilon}{8} C_{\text{tr}} \right) \sum_{F \in F_h^T} h_F^{-1} \|\nabla u_h \cdot n_F\|^2_{L^2(F)} \\
  &+ \left( \eta_2 - \frac{(1 - \varepsilon)}{2} + \frac{\varepsilon}{8} C_{\text{tr}} \right) \sum_{F \in F_h^T} h_F \|\nabla^2 u_h\|^2_{L^2(F)}.
\end{align*}
\]

One observes that coercivity of \( a_h + J_h \) can be guaranteed without the presence of the penalty terms, provided that \( (1 - \varepsilon) C_1 \leq \varepsilon C_{\text{tr}} \) and \( C_{\text{tr}} \) \((d + 1) (1 - \varepsilon) C_2 \leq \varepsilon C_{\text{tr}} \) are fulfilled. Note that \( C_1, C_2, C_{\text{tr}} \) and \( C_{\text{tr}} \) are independent of \( \varepsilon \). Thus, these inequalities are valid when \( \varepsilon \) is sufficiently close to 1.

In the numerical experiments we observed that neglecting the penalty terms \( J_h \) is in most situations feasible, but has negative influence on the robustness of preconditioned iterative solvers. However, the experimental convergence rates are better when the penalty terms are omitted, see Section 4.1.
3.4 A priori and a posteriori error estimates

This section is devoted to the a priori and a posteriori error analysis of the finite element approximation (17).

**Theorem 1** (A priori error estimate). Let $\Omega \subset \mathbb{R}^d$ be a bounded and convex domain. Assume that the penalty parameters $\eta_1, \eta_2$ in $J_h$ satisfy (42) and that the solution $u$ of (7) belongs to $H^s(\Omega)$ with

$$s \leq p + 1 \quad \text{and} \quad s > \begin{cases} 5/2, & \text{if } \mathbb{H} = \mathbb{H}^{CG}, \\ 2, & \text{if } \mathbb{H} = \mathbb{H}^{DG}. \end{cases} \quad (48)$$

The approximate solutions $u_h \in V_{h,0}$ of (17) with $p \geq 2$ fulfill the a priori error estimate

$$\|u - u_h\|_{H^s_h(\Omega)} \leq c_{\kappa,p,d} \left( 1 + \frac{C_{\kappa}(1 + \eta_1 + \eta_2)}{\varepsilon \alpha_0} \right) h^{s-2} \|u\|_{H^s(\Omega)}.$$

**Proof.** We introduce the nodal interpolant $I_{V_h}(u)$ as an intermediate function and deduce with Lemma 5 and standard interpolation error estimates

$$\|u - I_{V_h}(u)\|_{H^s_h(\Omega)} \leq c_{\kappa,d} \left( h^{-1} \|\nabla(u - I_{V_h}(u))\|_{L^2(\Omega)} + \|\nabla_h^2(u - I_{V_h}(u))\|_{L^2(\Omega)} \right) \leq c_{\kappa,p,d} h^{s-2} \|u\|_{H^s(\Omega)}. \quad (49)$$

Next, we derive an estimate for the norm of the discrete function $w_h := u_h - I_{V_h}(u)$. Therefore, we apply the discrete ellipticity (41), the definition of $u_h$ and the strong formulation (9) taking into account $\mathbb{H}_\Delta(w_h) \in W_h \subset L^2(\Omega)$ as well as $J_h(u, w_h) = 0$ which holds under the assumption (48). These arguments imply

$$\varepsilon \alpha_0 \|u_h - I_{V_h}(u)\|^2_{H^s_h(\Omega)} \leq a_h(u_h - I_{V_h}(u), w_h) + J_h(u_h - I_{V_h}(u), w_h)$$

$$= \int_\Omega \gamma f \mathbb{H}_\Delta(w_h) \, dx - a_h(I_{V_h}(u), w_h) - J_h(I_{V_h}(u), w_h)$$

$$= \int_\Omega \gamma A : (\nabla^2 u - \mathbb{H}(I_{V_h}(u))) \mathbb{H}_\Delta(w_h) \, dx + J_h(u - I_{V_h}(u), w_h). \quad (50)$$

With the triangle inequality, Lemmas 4 and 5 and standard interpolation error estimates we conclude

$$\int_\Omega \gamma A : (\nabla^2 u - \mathbb{H}(I_{V_h}(u))) \mathbb{H}_\Delta(w_h) \, dx$$

$$\leq \|\gamma A\|_{L^\infty(\Omega)} \left( \|\nabla^2 u - \mathbb{H}(u)\|_{L^2(\Omega)} + \|\mathbb{H}(u - I_{V_h}(u))\|_{L^2(\Omega)} \right) \|\mathbb{H}_\Delta(w_h)\|_{L^2(\Omega)}$$

$$\leq c_{\kappa,p,d} C_{\kappa}^2 h^{s-2} \|\gamma A\|_{L^\infty(\Omega)} \|u\|_{H^s(\Omega)} \|w_h\|_{H^s_h(\Omega)}. \quad$$

Note again that $\|\gamma A\|_{L^\infty(\Omega)} \leq 2$ holds due to (11). Analogously, we derive the following estimate for the jump terms in (50)

$$J_h(u - I_{V_h}(u), w_h) \leq C_{\kappa} \left( \sum_{F \in F_h^x} \left( \eta_1^2 h_F^{-1} \|\nabla(u - I_{V_h}(u)) \cdot n_F\|_{L^2(F)}^2 \right) + \eta_2^2 h_F \|\nabla^2(u - I_{V_h}(u))\|_{L^2(F)}^2 \right)^{1/2} \|w_h\|_{H^s_h(\Omega)}$$

$$\leq c_{\kappa,p,d} C_{\kappa} (\eta_1 + \eta_2) h^{s-2} \|u\|_{H^s(\Omega)} \|w_h\|_{H^s_h(\Omega)}.$$
Theorem 2 (A posteriori error estimate). Under the assumptions of Theorem 1 the solutions \( u_h \) of (17) fulfill the a posteriori error estimate

\[
\|u - u_h\|_{H^2(\Omega)}^2 \leq \sum_{T \in \mathcal{T}_h} \left( c_1 \|\gamma f - \gamma A : \nabla^2 u_h\|_{L^2(T)}^2 + c_2 \sum_{F \in \mathcal{F}_T \cap \mathcal{F}_h^F} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2 \right),
\]

with \( c_1 := 4 C_a^2 \) and \( c_2 := \frac{1}{2} (1 + (2 + 16 C_a^2) C_a^2) \), where \( \mathcal{F}_T \) denotes the set of facets of the element \( T \in \mathcal{T}_h \). The constants \( C_a \) and \( C_\Omega \) are defined in Lemma 2 and (29).

Proof. As in the proof of Lemma 7 we introduce the lifting operator \( \mathcal{E}_h \) which maps functions from \( \mathcal{V}_h \) into the \( H^2(\Omega) \)-conforming HCT or virtual finite element space \( \mathcal{V}_{h,\text{conf}} \). With this operator at hand we introduce a further approximation of the finite element solution \( u_h \), namely \( \tilde{u}_h = \mathcal{E}_h(u_h) \in \mathcal{V}_{h,\text{conf}} \).

With the triangle inequality, the definition of the norm in \( H^2_h(\Omega) \cap H^1_0(\Omega) \) and the fact that the jump terms vanish for \( u \in H^2(\Omega) \) we may represent the error term under consideration by

\[
\|u - u_h\|_{H^2_h(\Omega)}^2 \leq 2 \left( \|\nabla^2 (u - \tilde{u}_h)\|_{L^2(\Omega)}^2 + \sum_{T \in \mathcal{T}_h} \|\nabla^2 (\tilde{u}_h - u_h)\|_{L^2(T)}^2 \right)
+ \sum_{F \in \mathcal{F}_h^F} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2.
\]

We start by proving an estimate for the first term on the right-hand side of (52). We define the error functional \( J \in X' \) (recall that \( X = H^2(\Omega) \cap H^1_0(\Omega) \))

\[
J(\varphi) := \int_{\Omega} \nabla^2 (u - \tilde{u}_h) : \nabla^2 \varphi \, dx
\]

and easily confirm

\[
\|J\|_{X'} = \sup_{\varphi \in X \setminus \{0\}} \frac{|J(\varphi)|}{\|\varphi\|_X} \leq 1.
\]

This functional forms the right-hand side of a dual equation

\[
a(\varphi, z) = J(\varphi) \quad \forall \varphi \in X
\]

and from Lemma 2 we conclude the existence of a unique solution \( z \in X \) satisfying

\[
\|z\|_{H^2(\Omega)} \leq C_a \|J\|_{X'} \leq C_a.
\]

The definition of the lifting operator \( \mathcal{E}_h \) guarantees \( u - \tilde{u}_h \in X \) and thus,

\[
\|\nabla^2 (u - \tilde{u}_h)\|_{L^2(\Omega)} = J(u - \tilde{u}_h) = a(u - \tilde{u}_h, z).
\]

21
The right-hand side of the previous equation is treated as follows. We apply (12), insert the intermediate function $\gamma A : \nabla^2 u_h$, apply the Cauchy-Schwarz inequality as well as (54) to obtain
\[
\|\nabla^2 (u - \bar{u}_h)\|_{L^2(\Omega)} = a(u - \bar{u}_h, z)
= \int_{\Omega} (\gamma f - \gamma A : \nabla^2 \bar{u}_h) \Delta z \, dx
= \sum_{T \in \mathcal{T}_h} \int_T \left( (\gamma f - \gamma A : \nabla^2 u_h) + \gamma A : \nabla^2 (u_h - \bar{u}_h) \right) \Delta z \, dx
\leq C_n \sqrt{2} \left( \sum_{T \in \mathcal{T}_h} \left( \|\gamma f - \gamma A : \nabla^2 u_h\|_{L^2(T)}^2 + \|\gamma A\|_{L^\infty(\Omega)}^2 \|\nabla^2 (u_h - \bar{u}_h)\|_{L^2(T)}^2 \right) \right)^{1/2}. \tag{55}
\]
Finally, using $\|\gamma A\|_{L^\infty(\Omega)} \leq 2$, insertion of (55) into (52) and applying the estimate (32) for the lifting error terms leads to the desired result. \hfill \Box

The error estimate from the previous lemma provides a local a posteriori error estimator, namely
\[
\eta^2_F(u_h) := \|\gamma f - \gamma A : \nabla^2 u_h\|_{L^2(T)}^2 + \sum_{F \in F_T \cap F^2_h} h_F^{-1} \|\nabla u_h \cdot n_F\|_{L^2(F)}^2, \tag{56}
\]
and a global estimator
\[
\eta^2(u_h) := \sum_{T \in \mathcal{T}_h} \eta^2_F(u_h) \tag{57}
\]
which is a reliable bound for the error $\|u - u_h\|_{H^1_0(\Omega)}$.

**Theorem 3.** Let the assumptions of Theorem 1 be fulfilled. The a posteriori error estimate (51) is sharp in the sense that
\[
\frac{1}{2} \eta^2_F(u_h) \leq \|u - u_h\|_{H^1_0(T)} := \left( \|\nabla^2 (u - u_h)\|_{L^2(T)}^2 + \sum_{F \in F_T \cap F^2_h} h_F^{-1} \|\nabla (u - u_h) \cdot n_F\|_{L^2(F)}^2 \right)^{1/2}. \tag{58}
\]

**Proof.** The jump terms from the left-hand side of the desired estimate appear also in the norm of the right-hand side. We merely have to take into account that $\|\nabla u \cdot n_F\| = 0$ a.e. on all interior facets $F \in F^2_h$. The volume residuals are bounded by the element-wise $H^2(\Omega)$-seminorm due to
\[
\|\gamma f - \gamma A : \nabla^2 u_h\|_{L^2(T)} = \|\gamma A : (\nabla^2 u - \nabla^2 u_h)\|_{L^2(T)} \leq \|\gamma A\|_{L^\infty(\Omega)} \|\nabla^2 (u - u_h)\|_{L^2(T)}
\]
and $\|\gamma A\|_{L^\infty(\Omega)} \leq 2$. \hfill \Box

### 3.5 A method using the piecewise Hessian

Instead of using Hessian recovery techniques for the realization of our method, as investigated in the previous sections, it is also possible to use the cellwise exact Hessian, i.e., $\mathbb{H} := \nabla^2 h$. This idea is proposed in [33]. As the resulting bilinear form is not coercive additional jump penalty terms have to be added. The resulting equation reads

Find $u_h \in V_{h,0}$ s.t.
\[
a_h(u_h, v_h) := \int_\Omega \gamma A : \nabla^2_h u_h \Delta_h v_h \, dx + \eta_1 \sum_{F \in F^2_h} h_F^{-1} \int_F \|\nabla u_h \cdot n_F\| \|\nabla v_h \cdot n_F\| \, ds = \int_\Omega \gamma f \Delta v_h \, dx \tag{58}
\]
for all $v_h \in V_{h,0}$. Under the assumption that $\eta_1 > 0$ is sufficiently large ($\eta_1 = 0$ is not allowed here) and that the Cordes condition (11) is fulfilled with some $\varepsilon \in (0, 1]$, it has been proved in [33, Lemma 4.3] that the bilinear form $a_h(\cdot, \cdot)$ is uniformly coercive on $V_{h,0}$ and hence, (58) possesses a unique solution $u_h \in V_{h,0}$. This is a direct consequence of a discrete Miranda-Talenti estimate similar to Lemma 7 and the techniques applied in the proof of Lemma 9.

Due to the consistency of this scheme, one can easily conclude the a priori estimate

$$\|u - u_h\|_{H^2_0(\Omega)} \leq C h^{s-2} |u|_{H^s(\Omega)}, \quad s \in [2, p + 1],$$

provided that $u$ belongs to $H^s(\Omega)$.

A posteriori error estimates can be derived with the same argument as in Theorem 2. To be more precise, one can show by a slight modification of the proofs from the previous section that the estimator from (57) is a reliable and sharp bound for $\|u - u_h\|_{H^2_0(\Omega)}$.

An advantage of the direct scheme (58) is that the computational effort is less than for our system (17) since no additional equations for the computation of the Hessian approximation are needed. As we will observe in our numerical experiments, the approximation properties for the error $u - u_h$ in the $H^2_0(\Omega)$-norm as well as in the $H^1(\Omega)$-norm will be the same for both approaches. However, it turns out that the convergence rate in the $L^2(\Omega)$-norm is higher for the approach studied in the previous sections.

4 Numerical experiments

In this section, we perform different numerical experiments. All implementations were done in Python using the finite element library FEniCS 2019.1 [1, 30]. Our code is residing in a GitHub repository and it will be made publicly available upon acceptance of the manuscript.

It is our purpose to compare four discretization approaches, i.e.,

- the method using a finite-element Hessian with continuous and discontinuous trial functions (denoted by CG and DG in the following) discussed in the present article (Section 3.2),

- the Petrov-Galerkin scheme (N) proposed by Neilan [31], which likewise utilizes a DG finite-element Hessian but with $\tau_h = \text{id}$, i.e., there is no Laplacian acting on the test function,

- and the method using the piecewise Hessian proposed by Neilan, Salgado and Zhang (NSZ) [33] that we discussed briefly in Section 3.5.

4.1 A problem with almost violated Cordes condition

We choose a problem on the unit square with matrix

$$A = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

and determine the source term $f$ such that the smooth, exact solution of (7) is given by

$$u(x) = \sin(2 \pi x_1) \sin(2 \pi x_2).$$

The matrix $A$ fulfills the Cordes condition if $\kappa \in (-1, 1)$. If $\kappa$ is sent to 1, $\varepsilon$ and hence the coercivity constant from Lemma 9 will tend to zero so that the problem is harder to solve with an
Table 1: Iteration number for GMRES to achieve an absolute and relative tolerance of $10^{-8}$ for the method using a finite element Hessian with continuous trial functions, varying $\eta_1$ and fixed $\eta_2 = 0$.

| $h$ | $\eta_1 = 0$ | $\eta_1 = 1$ | $\eta_1 = 0$ | $\eta_1 = 1$ | $\eta_1 = 0$ | $\eta_1 = 1$ |
|-----|--------------|--------------|--------------|--------------|--------------|--------------|
| $2^{-3}$ | 23 | 13 | 25 | 13 | 26 | 13 |
| $2^{-4}$ | 27 | 17 | 25 | 17 | 25 | 17 |
| $2^{-5}$ | 23 | 19 | 25 | 19 | 30 | 19 |
| $2^{-6}$ | 24 | 19 | 26 | 19 | 27 | 19 |
| $2^{-7}$ | 23 | 19 | 25 | 20 | 27 | 20 |
| $2^{-8}$ | 23 | 20 | 25 | 27 | 26 | 20 |

iterative method like GMRES. This behavior is also observed in our numerical experiments. The iteration numbers required to realize our method with a CG Hessian for piecewise quadratic trial functions ($p = 2$) for different stabilization parameters in $J_h$ and different values of $\kappa$ are reported in Table 1. Obviously, with the pre-conditioner proposed in (23) and the stabilization term $J_h$, we observe that the iteration numbers mildly increase when the mesh parameter decreases or when $\kappa$ approaches 1. The incorporation of an additional jump term for the second derivatives in $J_h$, i.e., the choice $\eta_2 > 0$ in (39), did not lead to an improvement of the computational results.

![Figure 1: Comparison of absolute errors for different polynomial degrees for the example from Section 4.1 with $\kappa = 0.5$.](image)

In a further numerical test, we computed the discretization error for different polynomial degrees. Here, we used the choice $\kappa = 1/2$. As the Cordes condition for this example is fulfilled with a sufficiently large $\varepsilon$ we dropped the stabilization terms, i.e., we set $\eta_1 = \eta_2 = 0$. For comparison, we also present computational results for the piecewise Hessian approach (NSZ). The error plots in different norms and for varying polynomial degrees are shown in Figure 1. All convergence rates in the $H^2_h(\Omega)$-norm coincide with the ones predicted by Theorem 1. It is also observed that both approaches behave quite similarly. In the $H^2_h(\Omega)$-norm the errors decay almost identically. However we observe two advantages for our approach using a Hessian recovery.
strategy. First, it even converges in the $L^2(\Omega)$- and $H^1(\Omega)$-norm if the polynomial degree $p = 1$ is used. This coincides with the observations from [28], where the case $p = 1$ is allowed as well. Second, the convergence rate in the $L^2(\Omega)$-norm is higher for the Hessian recovery approach in case of quadratic elements. This is caused by the fact that a stabilization term is not needed in the present situation.

\[
\text{(a) } \|u - u_h\|_{L^2(\Omega)} \text{ for } p = 2 \\
\text{(b) } \|u - u_h\|_{H^1(\Omega)} \text{ for } p = 2 \\
\text{(c) } \|u - u_h\|_{H^2(\Omega)} \text{ for } p = 2 \\
\text{(d) } \|u - u_h\|_{L^2(\Omega)} \text{ for } p = 3 \\
\text{(e) } \|u - u_h\|_{H^1(\Omega)} \text{ for } p = 3 \\
\text{(f) } \|u - u_h\|_{H^2(\Omega)} \text{ for } p = 3
\]

Figure 2: This figure shows the performance of the algorithms for the example from Section 4.1 for $p = 2$ and $p = 3$ for different discretization strategies.

In a last test for this example we check how the methods studied in the present article compare with the approaches (N) and (NSZ) mentioned at the beginning of this section. The error curves for different norms and different polynomial degrees can be found in Figure 2. Although all approaches behave quite similarly, we observe a difference in the convergence rates in $L^2(\Omega)$ for quadratic elements. Obviously, the approaches taking into account stabilization terms (these are our approaches with $\eta_1 \neq 0$ and (NSZ)) converge only with order 2, while the remaining approaches (these are our approach with $\eta_1 = \eta_2 = 0$ and (N)) converge with order 3. A proof of this conjecture is subject of future research.

### 4.2 A problem with singular solution

In this example we consider the Poisson problem, i.e., the diffusion matrix is chosen as $A = I_{2 \times 2}$, in the domain $\Omega = (0, 1)^2$. Emphasis is put on problems whose solutions have reduced regularity.
To this end, we construct the right-hand side $f$ in such a way that

$$u(x) = r(x)^\alpha \sin(2\varphi(x)) (1 - x_1)(1 - x_2)$$

is the exact solution. Here, $(r(x), \varphi(x))$ are polar coordinates centered in the origin. A simple computation shows that $u \in H^s(\Omega)$ holds for all $s < 1 + \alpha$. In the present experiment we choose the value $\alpha = 3/2$ and expect the regularity of almost $H^{5/2}(\Omega)$, and thus, as predicted by Theorem 1, the convergence rate in the $H^2_h(\Omega)$-norm should be $1/2 - \varepsilon$ for arbitrary $\varepsilon > 0$. We would also expect that an adaptive finite element method will retain the optimal convergence rate.

The adaptive strategy we implemented uses the local error estimator (56), the Dörfler marking strategy in such a way that those elements contributing 90% to the globally estimated error are marked, and the bisection refinement strategy provided by the FENiCS library. The results shown in Figure 3 confirm the optimality of the adaptively generated finite element meshes. It is also observed that the convergence rates in the $H^1(\Omega)$- and $L^2(\Omega)$-norm are optimal.

![Finite element mesh](image1)

![Numerical solution](image2)

(a) Finite element mesh  
(b) Numerical solution $u_h$

![Errors for the Hessian recovery approach](image3)

![Errors for the piecewise Hessian approach (NSZ)](image4)

(c) Errors for the Hessian recovery approach (CG$_{\eta_1=0,\eta_2=0}$)  
(d) Errors for the piecewise Hessian approach (NSZ)

\[
\begin{align*}
\|u - u_h\|_{L^2(\Omega)} &\quad \|u - u_h\|_{H^1(\Omega)} &\quad \|u - u_h\|_{H^2_h(\Omega)} \\
\end{align*}
\]

Figure 3: Comparison between uniform (dashed lines) and adaptive refinement (solid lines) for the example from Section 4.2.
4.3 A problem with discontinuous coefficient matrix

This example illustrates the capability of the method to handle discontinuous diffusion coefficients. Problems of this type are of particular interest as a transformation to a PDE in divergence form is not possible. The coefficient matrix in the present example is

\[ A(x) = \begin{pmatrix} 2 & \text{sgn}(x_1 x_2) \\ \text{sgn}(x_1 x_2) & 2 \end{pmatrix}, \]

and \( f \) is chosen in such a way that \( u(x_1, x_2) = x_1 x_2 (1 - e^{1-|x_1|})(1 - e^{1-|x_2|}) \) is the exact solution. The computational domain is \( \Omega := (-1,1)^2 \). This example is also used in the numerical experiments from [17, 38, 43], where different discretization approaches are studied. Here, we apply our finite element scheme from Section 3.2 and investigate the behavior of an adaptive finite element method based on the error estimator derived in Theorem 2. The adaptively generated mesh as well as the error curves can be found in Figure 4.

4.4 A problem with anisotropic and discontinuous coefficient matrix

In this example we consider a problem in \( \Omega := (-1,1)^2 \) with the input data

\[ f = -1 \quad \text{and} \quad A = \begin{pmatrix} 0.02 & 0.01 \\ 0.01 & 1 + 1_{(x^3 - y^3 > 0)} \end{pmatrix}. \]
Figure 5: Comparison of absolute errors for different polynomial degrees $p$ for the example from Section 4.3 with discontinuous coefficient matrix.

The diffusion in $x_1$-direction is very small so that the solution exhibits a boundary layer at the boundary edges $x_2 = -1$ and $x_2 = 1$. Moreover, the coefficient $A_{22}$ is discontinuous. The computational results for our adaptive finite element method are illustrated in Figure 6. We observe that the discontinuity and the boundary layer are both resolved by the mesh. Furthermore, the propagation of the error is illustrated and one observes that the adaptive refinement retains the optimal convergence rate. Note that we used the value of the global estimator $\eta$ as an error measure since an explicit solution is not available for this example.

4.5 A three-dimensional problem with reduced regularity

In this numerical experiment we show the applicability of our procedure to the three-dimensional case. We choose the diffusion matrix to be $A = I_{3 \times 3}$ in the domain $\Omega = (0,1)^3$. Similar as in Section 4.2, the solution

$$u(x) = r(x)^{\alpha}$$

with $r(x) = \sqrt{\sum_{i=1}^{3} (x_i - 0.5)^2}$ possesses a reduced regularity, i.e., $u \in H^s(\Omega)$ for $s < 1.5 + \alpha$. The results for various choices of $\alpha$ are shown in Figure 7 and confirm the expected behavior. For example, the selection $\alpha = 1$ results in a regularity of almost $H^{5/2}(\Omega)$, which in turn yields an expected convergence rate in the $H^1(\Omega)$- and $L^2(\Omega)$-norms, respectively. The corresponding error plot in Figure 7 confirms this. An adaptive refinement strategy using the local error estimator (56) with a refinement threshold of 95% is capable of confirming the convergence rate of the errors in the $L^2(\Omega)$-norm and improving the convergence rates in the $H^1(\Omega)$- and $H^2(\Omega)$-norm.

5 Conclusion and outlook

The proposed method can be extended to parabolic problems. Given a regular solution and an appropriate time stepping scheme one can observe the same convergence rates as in the elliptic case. Numerical tests have been performed to confirm this and they are included in the GitHub repository accompanying the paper. A detailed analysis as in [40] is left to future
research. Another subject, and this was the authors’ original motivation to study this topic, is the application of the proposed discretization to Hamilton-Jacobi-Bellman equations. A preliminary implementation is also available in the repository and the related theoretical foundation will be examined in further publications.

Besides these two extensions there are further interesting questions left. An obvious question is the proof for error estimates in lower-order norms. Note that one advantage of the approach proposed in this article is that optimal convergence in $L^2(\Omega)$ is observed. However, a proof of this observation is still missing. To the best of the authors’ knowledge the only article dealing with estimates in lower-order norms is [18], where an $H^1(\Omega)$-norm estimate for the Petrov-Galerkin approach using $\tau_h = \text{id}$ is shown. Related studies for methods exploiting the Cordes condition are not available in the literature. A proof based on the usual duality argument is likely not expedient as, for instance, the dual equation of a non-divergence form PDE is a PDE in double-divergence form whose solutions possess insufficient regularity. In the special case $A = \text{id}$ and for the method proposed in Section 3.5, the discretization coincides with a $C^0$-interior penalty discretization of
the biharmonic equation and error estimates in lower-order norms can be directly concluded from [8]. However, an extension to approaches using recovered Hessians is not straightforward and requires further investigations.

**Acknowledgments**

The authors thank Martin Stoll for discussions on preconditioners for our equation system.

This work was partially supported by DFG grant HE 6077/7–1 Impulse Control Problems and Adaptive Numerical Solution of Quasi-Variational Inequalities in Markovian Factor Models. Funding is gratefully acknowledged.

**References**

[1] Martin Alnæs et al. “The FEniCS project version 1.5”. In: *Archive of Numerical Software* 3.100 (2015), pp. 9–23. DOI: 10.11588/ans.2015.100.20553.

[2] Guy Barles and Espen Robstad Jakobsen. “On the convergence rate of approximation schemes for Hamilton-Jacobi-Bellman equations”. In: *M2AN. Mathematical Modelling and Numerical Analysis* 36.1 (2002), pp. 33–54. DOI: 10.1051/m2an:2002002.

[3] Jean-David Benamou, Brittany D. Froese, and Adam M. Oberman. “Two numerical methods for the elliptic Monge-Ampère equation”. In: *M2AN. Mathematical Modelling and Numerical Analysis* 44.4 (2010), pp. 737–758. DOI: 10.1051/m2an/2010017.

[4] Jan Blechschmidt and Roland Herzog. “Improving policies for Hamilton-Jacobi-Bellman equations by postprocessing”. In: *Hamilton-Jacobi-Bellman equations*. Vol. 21. Radon Series on Computational and Applied Mathematics. De Gruyter, Berlin, 2018, pp. 25–42. DOI: 10.1515/9783110543599-002.
[5] Susanne C. Brenner, Thirupathi Gudi, and Li-Yeng Sung. “An a posteriori error estimator for a quadratic $C^0$-interior penalty method for the biharmonic problem.” In: IMA Journal of Numerical Analysis 30.3 (2010), pp. 777–798. doi: 10.1093/imanum/drn057.

[6] Susanne C. Brenner and Ellya L. Kawecki. Adaptive $C^0$ interior penalty methods for Hamilton-Jacobi-Bellman equations with Cordes coefficients. 2019. arXiv: 1911.05407.

[7] Susanne C. Brenner and Michael Neilan. “Finite element approximations of the three-dimensional Monge-Ampère equation”. In: ESAIM. Mathematical Modelling and Numerical Analysis 46.5 (2012), pp. 979–1001. doi: 10.1051/m2an/2011067.

[8] Susanne C. Brenner and Li-Yeng Sung. “$C^0$ interior penalty methods for fourth order elliptic boundary value problems on polygonal domains”. In: Journal of Scientific Computing 22/23 (2005), pp. 83–118. doi: 10.1007/s10915-004-4135-7.

[9] Susanne C. Brenner and Li-yeng Sung. Virtual Enriching Operators. 2019. arXiv: 1903.08476.

[10] Erik Burman. “A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty”. In: SIAM Journal on Numerical Analysis 43.5 (2005), pp. 2012–2033. doi: 10.1137/S0036142903437374.

[11] Philippe G. Ciarlet. “Interpolation error estimates for the reduced Hsieh-Clough-Tocher triangle”. In: Mathematics of Computation 32.142 (1978), pp. 335–335. doi: 10.1090/s0025-5718-1978-0482249-1.

[12] Edward J. Dean and Roland Glowinski. “Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type”. In: Computer Methods in Applied Mechanics and Engineering 195.13-16 (2006), pp. 1344–1386. doi: 10.1016/j.cma.2005.05.023.

[13] Todd Dupont and Ridgway Scott. “Polynomial approximation of functions in Sobolev spaces”. In: Mathematics of Computation 34.150 (1980), pp. 441–463. doi: 10.2307/2006095.

[14] Xiaobing Feng, Lauren Hennings, and Michael Neilan. “Finite element methods for second order linear elliptic partial differential equations in non-divergence form”. In: Mathematics of Computation 86.307 (2017), pp. 2025–2051. doi: 10.1090/mcom/3168.

[15] Xiaobing Feng and Michael Neilan. “Mixed finite element methods for the fully nonlinear Monge-Ampère equation based on the vanishing moment method”. In: SIAM Journal on Numerical Analysis 47.2 (2009), pp. 1226–1250. doi: 10.1137/070710378.

[16] Xiaobing Feng and Michael Neilan. “Vanishing moment method and moment solutions for fully nonlinear second order partial differential equations”. In: Journal of Scientific Computing 38.1 (2008), pp. 74–98. doi: 10.1007/s10915-008-9221-9.

[17] Xiaobing Feng, Michael Neilan, and Stefan Schnake. “Interior penalty discontinuous Galerkin methods for second order linear non-divergence form elliptic PDEs”. In: Journal of Scientific Computing 74.3 (2018), pp. 1651–1676. doi: 10.1007/s10915-017-0519-3.

[18] Xiaobing Feng and Stefan Schnake. $H^1$-norm error estimate for a nonstandard finite element approximation of second-order linear elliptic PDEs in non-divergence form. 2019. arXiv: 1909.13803.

[19] Wendell H. Fleming and Raymond W. Rishel. Deterministic and stochastic optimal control. Applications of Mathematics, No. 1. Springer-Verlag, Berlin-New York, 1975.
[20] Thomas Führer. Ultraweak formulation of linear PDEs in nondivergence form and DPG approximation. 2019. arXiv: 1908.08981.

[21] Dietmar Gallistl. “Variational formulation and numerical analysis of linear elliptic equations in nondivergence form with Cordès coefficients”. In: SIAM Journal on Numerical Analysis 55.2 (2017), pp. 737–757. DOI: 10.1137/16m1080495.

[22] Emmanuel H. Georgoulis, Paul Houston, and Juha Virtanen. “An a posteriori error indicator for discontinuous Galerkin approximations of fourth-order elliptic problems”. In: IMA Journal of Numerical Analysis 31.1 (2011), pp. 281–298. DOI: 10.1093/imanum/drp023.

[23] David Gilbarg and Neil S. Trudinger. Elliptic partial differential equations of second order. Classics in Mathematics. Reprint of the 1998 edition. Springer-Verlag, Berlin, 2001.

[24] Hailong Guo, Zhimin Zhang, and Ren Zhao. “Hessian recovery for finite element methods”. In: Mathematics of Computation 86.306 (2016), pp. 1671–1692. DOI: 10.1090/mcom/3186.

[25] Cristian E. Gutiérrez. The Monge-Ampère equation. Vol. 44. Progress in Nonlinear Differential Equations and their Applications. Birkhäuser Boston, Inc., Boston, MA, 2001. DOI: 10.1007/978-1-4612-0195-3.

[26] Ohannes A. Karakashian and Frederic Pascal. “A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems”. In: SIAM Journal on Numerical Analysis 41 (2003), pp. 2374–2399.

[27] Ellya L. Kawecki. “A DGFEM for nondivergence form elliptic equations with Cordes coefficients on curved domains”. In: Numerical Methods for Partial Differential Equations 35.5 (2019), pp. 1717–1744. DOI: 10.1002/num.22372.

[28] Omar Lakkis and Tristan Pryer. “A finite element method for nonlinear elliptic problems”. In: SIAM Journal on Scientific Computing 35.4 (2013), A2025–A2045. DOI: 10.1137/120887655.

[29] Omar Lakkis and Tristan Pryer. “A finite element method for second order nonvariational elliptic problems”. In: SIAM Journal on Scientific Computing 33.2 (2011), pp. 786–801. DOI: 10.1137/100787672.

[30] Anders Logg, Garth N. Wells, and Johan Hake. “DOLFIN: a C++/Python Finite Element Library”. In: Automated Solution of Differential Equations by the Finite Element Method, Volume 84 of Lecture Notes in Computational Science and Engineering. Ed. by Anders Logg, Kent-Andre Mardal, and Garth N. Wells. Springer, 2012. Chap. 10. DOI: 10.1007/978-3-642-23099-8_10.

[31] Michael Neilan. “Convergence analysis of a finite element method for second order nonvariational elliptic problems”. In: Journal of Numerical Mathematics 25.3 (2017), pp. 169–184. DOI: 10.1515/jnma-2016-1017.

[32] Michael Neilan. “Finite element methods for fully nonlinear second order PDEs based on a discrete Hessian with applications to the Monge–Ampère equation”. In: Journal of Computational and Applied Mathematics 263 (2014), pp. 351–369. DOI: 10.1016/j.cam.2013.12.027.

[33] Michael Neilan, Abner J. Salgado, and Wujun Zhang. “Numerical analysis of strongly nonlinear PDEs”. In: Acta Numerica 26 (2017), pp. 137–303. DOI: 10.1017/S0962492917000071.

[34] Michael Neilan and Mohan Wu. “Discrete Miranda-Talenti estimates and applications to linear and nonlinear PDEs”. In: Journal of Computational and Applied Mathematics 356 (2019), pp. 358–376. DOI: 10.1016/j.cam.2019.01.032.
[35] Ricardo H. Nochetto and Wujun Zhang. “Discrete ABP estimate and convergence rates for linear elliptic equations in non-divergence form”. In: Foundations of Computational Mathematics. The Journal of the Society for the Foundations of Computational Mathematics 18.3 (2018), pp. 537–593. DOI: 10.1007/s10208-017-9347-y.

[36] Huyên Pham. Continuous-time stochastic control and optimization with financial applications. Vol. 61. Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2009. DOI: 10.1007/978-3-540-89500-8.

[37] Iain Smears. “Nonoverlapping domain decomposition preconditioners for discontinuous Galerkin approximations of Hamilton-Jacobi-Bellman equations”. In: Journal of Scientific Computing 74.1 (2018), pp. 145–174. DOI: 10.1007/s10915-017-0428-5.

[38] Iain Smears and Endre Süli. “Discontinuous Galerkin finite element approximation of Hamilton-Jacobi-Bellman equations with Cordes coefficients”. In: SIAM Journal on Numerical Analysis 52.2 (2014), pp. 993–1016. DOI: 10.1137/130909536.

[39] Iain Smears and Endre Süli. “Discontinuous Galerkin finite element approximation of nondivergence form elliptic equations with Cordès coefficients”. In: SIAM Journal on Numerical Analysis 51.4 (2013), pp. 2088–2106. DOI: 10.1137/120899613.

[40] Iain Smears and Endre Süli. “Discontinuous Galerkin finite element methods for time-dependent Hamilton-Jacobi-Bellman equations with Cordes coefficients”. In: Numerische Mathematik 133.1 (2015), pp. 141–176. DOI: 10.1007/s00211-015-0741-6.

[41] Neil S. Trudinger and Xu-Jia Wang. “The Monge-Ampère equation and its geometric applications”. In: Handbook of geometric analysis. No. 1. Vol. 7. Advanced Lectures in Mathematics. International Press of Boston, Somerville, MA, 2008, pp. 467–524.

[42] Rüdiger Verfürth. A posteriori error estimation techniques for finite element methods. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013. DOI: 10.1093/acprof:oso/9780199679423.001.0001.

[43] Chunmei Wang and Junping Wang. “A primal-dual weak Galerkin finite element method for second order elliptic equations in non-divergence form”. In: Mathematics of Computation 87.310 (2017), pp. 515–545. DOI: 10.1090/mcom/3220.