Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Consistency and order 1 convergence of cell-centered finite volume discretizations of degenerate elliptic problems in any space dimension

Martin Heida[1], Alexander Sikorski[2], Marcus Weber[2]

submitted: January 13, 2022

[1] Weierstraß-Institut
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: martin.heida@wias-berlin.de

[2] Zuse Institut Berlin
Takustraße 7
14195 Berlin
Germany
E-Mail: sikorski@zib.de
weber@zib.de

No. 2913
Berlin 2022

2020 *Mathematics Subject Classification.* 65N08, 65N12, 65N50, 80M12.

Key words and phrases. Elliptic, finite volume, Voronoi.

MH has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 “Scaling Cascades in Complex Systems”, Project C05 “Effective models for materials and interfaces with multiple scales”. AS has been funded by CRC 1114 Project B05 ”Origin of scaling cascades in protein dynamics”.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Consistency and order 1 convergence of cell-centered finite volume discretizations of degenerate elliptic problems in any space dimension

Martin Heida, Alexander Sikorski, Marcus Weber

Abstract

We study consistency of cell-centered finite difference methods for elliptic equations with degenerate coefficients in any space dimension $d \geq 2$. This results in order of convergence estimates in the natural weighted energy norm and in the weighted discrete L^2-norm on admissible meshes. The cells of meshes under consideration may be very irregular in size. We particularly allow the size of certain cells to remain bounded from below even in the asymptotic limit. For uniform meshes we show that the order of convergence is at least 1 in the energy semi-norm, provided the discrete and continuous solutions exist and the continuous solution has H^2 regularity.

1 Introduction

In [5, 8] the state of the art of the order of convergence for cell-centered finite difference methods on a family of admissible meshes (e.g. cell-centered Voronoi) has been established for $d \leq 3$. The analysis there is based on the fact that H^2-functions in dimension up to 3 allow for pointwise evaluation and on the boundedness of some integrals that appear there. A more recent result [7] treats higher space dimensions and unstructured grids for nonlinear problems, but does not provide an estimate for the order of convergence. Order of convergence is provided in [1] using corrective additional terms.

However, order of convergence estimates for finite volume approximations of elliptic PDE are important for applied sciences, in particular for conformation dynamics [12]. In conformation dynamics the analysis of rare transition events between different states of a molecular system are analyzed. In order to avoid long-term molecular simulations, the problem has been reformulated in terms of solving a partial differential equation and simple discretization schemes have been developed [10]. This approach makes it necessary to consider order of convergence of finite difference methods in very high dimensions (e.g. [4] where $d = 9$), with degenerate weights and on meshes that have non-uniform distributions of cell-sizes. In the followings, we show how $W^{2,p}$-regularity ($p \geq 2$) of the continuous solution can provide quantitative information on the quality of the approximation, even in high space dimensions (see Cor. 2.7).

The argument we use relies pretty much on the original argument in [5, 8], in particular on the formulas (3.2)–(3.3) below. As the major difference, [5, 8] use the pointwise evaluation operator to map H^2-function onto discrete functions, which is well defined in $d \leq 3$. However, the integrals showing up in this approach become unsolvable in high dimensions. This problem no longer exists when using an averaging operator to match H^2-functions with discrete functions. We will not go into further details of [5, 8] but encourage the reader to compare the two approaches directly from the source.

Aside from the missing $C(\overline{\Omega})$-regularity of solutions, a further problem in high dimensions is the number of control volumes which grows exponentially with the dimension for a prescribed resolution:

DOI 10.20347/WIAS.PREPRINT.2913 Berlin 2022

if for example $\Omega = [0,1]^d$ and if we chose as control volumes cubes of edge length h we need h^{-d} cubes. Already in dimension $d = 6$ with $h = 0.01$ this implies 10^{12} cubes. Furthermore, every cube has 12 neighbors and hence the matrix has 10^{13} entries even though the matrix is sparse.

However, there is hope in certain fields of application. For example, if we apply our results in conformation dynamics, then we can assume that the solution function is almost constant in large parts of Ω. These parts correspond to what is called "metastable conformation" in the field of conformation dynamics. We can make use of this assumption by choosing large control volumes in this area and by choosing small control volumes only in those regions where we expect significant changes of the solution function. We will account for the size of cells by an outer radius R_k and an inner radius r_K and our convergence estimates will weight the local H^2-norm of the limit solutions vs. a polynomial in R_K and eccentricity $\frac{R_K}{r_K}$. The H^2-norm is in particular interesting in applications, because the shape of the solution functions can be assumed to be almost linear (the 2nd derivative vanishes) in regions which would be named "transition region" in conformation dynamics. In this special field of application there exists a heuristic approach for choosing an adaptive cell volume (see p. 38 bottom in [14]). Large cell volumes are possible whenever the molecular system is "rapidly mixing" within the cells (metastable conformations), or whenever the cells are rapidly left (transition region, high energy regions). In this article, this heuristics is justified and refined on the basis of error estimates in Theorem 2.5.

The setting and the analytical results in a nutshell

In order to show that our main theorems 2.5 (convergence in energy norm) and 2.14 (discrete weighted Poincaré inequality) are consistent with existing results in low dimensions $d \leq 3$ [5, 8] and with uniform meshes, we provide an interpretation of our result in this particular case at the end of this introduction.

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$ be a polytopal connected domain. We furthermore assume we are given $\kappa \in C^1(\mathbb{R}^d)$ with $\kappa > 0$ almost everywhere on $\overline{\Omega}$, except on lower dimensional structures of dimension less or equal to $d-1$. On the set $\kappa > 0$ we introduce

$$V(x) := -2 \ln \kappa(x) \qquad \text{such that} \qquad \kappa = \exp(-\tfrac{1}{2}V)\,. \tag{1.1}$$

We then study the elliptic equation

$$-\mathrm{div}\,(\kappa \nabla u) = f \qquad \text{on } \Omega \tag{1.2}$$

equipped with Dirichlet, Neumann or periodic boundary conditions (the later only if Ω is rectangular), or a combination. The existence of a solution $u \in H^2_{\mathrm{BC}}(\Omega)$ to (1.2) is guaranteed for certain conditions on κ, on the boundary behavior and on the right hand side f which are provided in literature (e.g. the early work [3], the review [2] and references therein). In this work, we simply assume the existence of a solution $u \in H^2_{\mathrm{BC}}(\Omega)$ to (1.2) and ask for the rate of convergence of solutions $u_{\mathcal{T}} \in H_{\mathcal{T},\mathrm{BC}}$ of (1.3) to u in an appropriate sense. Here, $H_{\mathcal{T},\mathrm{BC}}$ is the discrete function space that reflects the discrete version of the respective boundary conditions, see below.

On an admissible mesh $\mathcal{T}$ in the sense of Definition 2.1 with polytopal control volumes $\mathcal{V} = \{K\}_K$ with facets $\mathcal{E}_K$ and discrete derivatives $\partial_{K,\sigma}$ given in (2.1) we consider the discretization

$$\forall K \in \mathcal{V}: \quad \sum_{\sigma \in \mathcal{E}_K} m_\sigma \kappa_\sigma \partial_{K,\sigma} u_{\mathcal{T}} = m_K f_K\,, \quad \text{where } f_K := \fint_K f \tag{1.3}$$

and $\kappa_\sigma := S(\kappa_K, \kappa_L)$ defined in (2.1) is an average of the neighbored values κ_L and κ_K given by $S \in C(\mathbb{R}^2)$ with

$$S(x,x) = x\,, \qquad S(x,y) = S(y,x)\,, \qquad \min\{x,y\} \leq S(x,y) \leq \max\{x,y\}\,.$$

Let us note we assume $\kappa > 0$ a.e. and hence we choose $\kappa_\sigma > 0$ implying the existence of a unique solution $u_\mathcal{T}$ of (1.3) is always guarantied for discrete Dirichlet, Neumann and periodic boundary conditions.

In order to quantify the quality of approximation we introduce the radii

$$\forall K \in \mathcal{V}: \qquad r_K := \frac{1}{2}\sup\{r|\, \mathbb{B}_r(x_k) \subset K\}\,, \quad R_K := \max\{R|\, K \subset \mathbb{B}_R(x_K)\}\,. \tag{1.4}$$

Definition 1.1 (Quasi uniform meshes)**.** A family of admissible meshes $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ (in the sense of Definition 2.1 below) indexed by $h > 0$, is called *quasi uniform* if there exists $C_{\rm uni} > 0$ such that

$$\forall h > 0,\, K \in \mathcal{V}_h: \qquad R_K < h\,,\ \frac{R_K}{r_K} \le C_{\rm uni}\,. \tag{1.5}$$

Given $L^2(\mathcal{T})$ the space of functions $\mathcal{V} \mapsto \mathbb{R}$, with norm $\|u\|^2_{L^2(\mathcal{T})} := \sum_K m_K u_K^2$, in (2.4)–(2.6) we introduce a discretization operator $\mathcal{R}_{\mathcal{T},\rm BC}$ with adjoint $\mathcal{R}^*_\mathcal{T}$ i.e.,

$$\mathcal{R}_{\mathcal{T},\rm BC}:\, L^2(\Omega) \to L^2(\mathcal{T})\,, \quad \mathcal{R}^*_\mathcal{T}:\, L^2(\mathcal{T}) \to L^2(\Omega)\,,$$

satisfying for every $u \in L^2(\mathcal{T})$ the relation $\mathcal{R}_{\mathcal{T}_h,\rm BC}\mathcal{R}^*_{\mathcal{T}_h} u = u$. Furthermore, Lemma 2.4 implies for a family of quasi uniform meshes that

$$\left\|\mathcal{R}^*_{\mathcal{T}_h}\mathcal{R}_{\mathcal{T}_h,\rm BC}u - u\right\|_{L^2(\Omega)} \le Ch\,. \tag{1.6}$$

This justifies to use $\mathcal{R}_{\mathcal{T}_h,\rm BC}$ in the formulation of our convergence results.

Theorem (Simplified version of our main Theorem 2.5)**.** *Let $u \in H^2(\Omega)$ be a solution to* (1.2) *and let $\mathcal{T}_h$ be a family of quasi uniform meshes satisfying* (1.5)*. There exists a family of discrete solution $u_{\mathcal{T}_h}$ of* (1.3) *satisfying*

$$\sum_{\sigma\in\mathcal{E}_h} m_\sigma h_\sigma \kappa_\sigma \left[\partial_\sigma\left(u_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h,\rm BC}u\right)\right]^2 \le C(\kappa, d, \Omega, \mathrm{BC}, C_{\rm uni})\, h^2\, \|\nabla u\|^2_{H^1(\Omega)}\,,$$

where $C(\kappa, d, \Omega, \mathrm{BC}, C_{\rm uni})$ does not depend on the discretization $\mathcal{T}_h$.

Proof. This follows from $\frac{R_K}{r_K} \le C_{\rm uni}$, $R_K \le C_{\rm uni} h$ and $\left|\frac{\kappa(x)}{\sqrt{\kappa_\sigma}} - \sqrt{\kappa_\sigma}\right| \le R_K \left\|\nabla\sqrt{\kappa}\right\|_\infty$ as well as $|\kappa_\sigma| \le \|\kappa\|_\infty$. □

Remark. A uniform positive lower bound on κ would turn the above result into the well known standard result for $d \le 3$ [5].

Finally, we introduce a weighted discrete Poincaré inequality. For this we introduce the concept of piecewise pseudo-monotone coefficient fields on an underlying tessellation of the original domain Ω. For such coefficient fields κ we are able to introduce on every admissible mesh $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ a function $\tilde\kappa_\mathcal{T}(x)$ constant on every $K \in \mathcal{V}$ and an average $\overline{u}^{\tilde\kappa}$ such that for some constant $C > 0$ independent from $\mathcal{T}$

$$\sum_K \tilde\kappa_{\mathcal{T},K} m_K \left(u_K - \overline{u}^{\tilde\kappa}\right)^2 \le C \sum_{\sigma\in\mathcal{E}} m_\sigma h_\sigma \kappa_\sigma \left(\partial_\sigma u\right)^2\,.$$

The last estimate links convergence in the energy norm to the convergence in a weighted discrete L^2-norm. Particularly given $O = \kappa^{-1}(0)$ and a quasi uniform family of meshes $\mathcal{T}_h$ the findings below yield for every $\varepsilon > 0$ the existence of a constant $C_\varepsilon > 0$ such that

$$\sum_K \tilde\kappa_{\mathcal{T}_h,K} \left|K \cap (\Omega\backslash\mathbb{B}_\varepsilon(O)\right| \left(u_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h,\rm BC}u\right)^2 \le C(\kappa, d, \Omega, \mathrm{BC}, C_{\rm uni}, C_\varepsilon)\, h^2\, \|\nabla u\|^2_{H^1(\Omega)}\,.$$

It is worth mentioning that the main theorem does not make explicit use of the form of the average S or its differentiability. While earlier [9] work was relying on the special form of derivatives of S, we emphasize that our proof does not at all use the specific form of S and hence our result always holds. In view of [9] this on the other hand implies the need of future research concerning the optimal choice of S.

Overdamped Lagenvin equation and SQRA

To see how the above theory is related to conformation dynamics, recall that the Kolmogorov forward equation for the overdamped Langevin equation

$$dX_t = -\nabla V(x)dt + \sigma dB_t \tag{1.7}$$

with potential V is given by the PDE

$$\beta U_t = \nabla \cdot \left(\nabla U + \beta U \nabla V\right) , \tag{1.8}$$

where $\beta = 2\sigma^{-2}$. Together with $\kappa \propto \exp\left(-\beta V\right)$ and $U = \frac{\tilde{u}}{\kappa}$ this takes the form of (1.2):

$$\kappa\beta\tilde{u}_t = \nabla \cdot \left(\kappa \nabla \tilde{u}\right) . \tag{1.9}$$

In this context, the use of the geometric mean $S(a,b) = \sqrt{ab}$ on the densities corresponds to the usual (arithmetic) average in the potentials and is henceforth a very natural choice for this problem [4]:

$$S(\kappa(x),\kappa(y)) = \sqrt{\kappa(x)\kappa(y)} = \exp\left(-\beta\frac{V(x)+V(y)}{2}\right) . \tag{1.10}$$

This is the canonical choice of κ_σ in the SQRA method.

Numerical calculations and discussion

In Section 5 we construct numerical examples which will confirm the error estimate from Theorem 2.5.

Since the numerical algorithm approximates the solution u to the PDE (1.2) with prescribed right hand side f, we start from an a priori known solution $u \in H^2(\mathbb{R}^4)$ with support in the unit ball which allows us to compute the corresponding right hand side f exactly by differentiation.

The spatial discretization will be achieved by a Voronoi tessellation based on 4 different families of sampling points. Furthermore, in order to simplify the calculation of the right hand side in (1.3) we assume $\kappa \equiv 1$. Since we developed the method and the theory based on the assumption that $u \equiv 0$ in those regions where κ is not degenerate and where $|\nabla\kappa|$ is small, the right hand side in the error estimate in Theorem 2.5 is dominated by the derivatives of u, anyway.

Within the above simplified setting, the simulations show the predicted behavior. We show this by comparing the difference between the discrete and the continuous solution in the energy norm and the hypothetical upper bound given by the error $I_{2,\mathcal{T}}$, see Fig. 2.

Furthermore, under the assumption that the cells are of approximate similar size as above, Fig. 3 shows that the error goes with $N^{-\frac{1}{d}}$, where $N \propto h^{-d}$ is the number of cells for a given average cell diameter h. Furthermore, Fig. 3 clearly shows that an adaptive sampling based on the expected distribution of $|\nabla u|$ or $|\nabla^2 u|$ (where u is the exact solution) is superior to a normal or uniform sampling.

2 Main analytical results

2.1 Preliminaries and notation

We recall the definition of an admissible mesh [5, 6] in order to fix notation.

Definition 2.1. An admissible mesh of Ω in the sense of [5, 6] is a triangulation $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ consisting of

1 a family of control volumes $\mathcal{V}$ which are mutually disjoint, convex polytopal open sets with $\bigcup_{K\in\mathcal{V}} \overline{K} = \overline{\Omega}$. Every K has mass m_K.

2 $\mathcal{E} = \mathcal{E}_{int} \cup \mathcal{E}_\partial$ is a finite family of disjoint subsets of $\overline{\Omega}$ such that every $\sigma \in \mathcal{E}$ is subset of a hyperplane and for every $K \in \mathcal{V}$ there exists $\mathcal{E}_K \subset \mathcal{E}$ with $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$. The $d-1$ dimensional measure of σ is m_σ. Furthermore, $\mathcal{E}_\partial = \{\sigma_i : i = 1, \dots, N_\partial\}$ and $\sigma \in \mathcal{E}$ satisfies $\sigma \in \mathcal{E}_\partial$ iff $\sigma \subset \partial\Omega$. For every K and $\sigma \in \mathcal{E}_K$ we write $\nu_{K,\sigma}$ for the outer orthonormal vector of K on σ. Furthermore, for every $\sigma \in \mathcal{E}_{int}$ the set $\mathcal{V}_\sigma := \{K \in \mathcal{V} : \sigma \in \mathcal{E}_K\}$ has exactly two elements and for $\sigma \in \mathcal{E}_\partial$ the set $\mathcal{V}_\sigma$ has exactly one element.

3 $\mathcal{P} = \mathcal{P}_{int} \cup \mathcal{P}_\partial$ is a family of points with $\mathcal{P}_{int} = \{x_K : K \in \mathcal{V}\}$, $\mathcal{P}_\partial = \{x_\sigma : \sigma \in \mathcal{E}_\partial\}$ such that $x_K \in K$, $x_\sigma \in \sigma$ are unique for each $K \in \mathcal{V}$, $\sigma \in \mathcal{E}_\sigma$. We assume for $\sigma \in \mathcal{E}_{int}$ and $\mathcal{V}_\sigma = \{K, L\}$ that $(x_K - x_L) \perp \sigma$ and for $\sigma \in \mathcal{E}_\partial$, $K \in \mathcal{V}_\sigma$ $(x_K - x_\sigma) \perp \sigma$.

Notation 2.2. Let $u : \mathcal{P} \to \mathbb{R}$. For $K \in \mathcal{V}$ and $\sigma \in \mathcal{E}_\partial$ we write $u_K := u(x_K)$ and $u_\sigma := u(x_\sigma)$. For $\kappa \in C^1(\mathbb{R}^d)$ we denote $\kappa_\mathcal{T} : \mathcal{P} \to \mathbb{R}$ the pointwise evaluation of κ in $\mathcal{P}$ with $\kappa_K := \kappa(x_K)$.

1 If $\sigma \in \mathcal{E}_{int}$ there exist $K, L \in \mathcal{V}$ such that $\{\sigma\} = \mathcal{E}_K \cap \mathcal{E}_L$ and $h_\sigma := |(x_K - x_L) \cdot \nu_{K,\sigma}|$. Furthermore, we write

$$\partial_{K,\sigma} u := \frac{1}{h_\sigma}(u_L - u_K), \quad \kappa_\sigma := S(\kappa_K, \kappa_L), \quad \partial_\sigma u := \nu_{K,\sigma}\partial_{K,\sigma} u, \tag{2.1}$$

where κ_σ and $\partial_\sigma u$ are invariant under permutation of K and L.

2 If $\sigma \in \mathcal{E}_\partial$ there exists a unique $K \in \mathcal{V}$ such that $\sigma \in \mathcal{E}_K$ and $h_\sigma := |(x_K - x_\sigma) \cdot \nu_{K,\sigma}|$. Furthermore, we write $\partial_{K,\sigma} u := \frac{1}{h_\sigma}(u_\sigma - u_K)$, $\kappa_\sigma := S(\kappa_K, \kappa(x_\sigma))$ and $\partial_\sigma u := \nu_{K,\sigma}\partial_{K,\sigma} u$.

Assumption 2.3. *The mesh $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ is such that for and every $K \in \mathcal{V}$ it holds $\kappa_K > 0$.*

In what follows, we write $H^2_0(\Omega) := H^2(\Omega) \cap H^1_0(\Omega)$ as well as $H^2_{\mathrm{per}}(\Omega)$ for periodic $H^2(\Omega)$ functions with mean value 0 and

$$H^2_{(0)}(\Omega) := \left\{ u \in H^2(\Omega) \,|\, \int_\Omega u = 0,\ \partial_\nu u = 0 \text{ on } \partial\Omega \right\}.$$

These spaces clearly correspond to homogeneous Dirichlet boundary conditions (BC), periodic boundary conditions and homogeneous Neumann boundary conditions. We introduce corresponding discrete spaces incorporating discrete boundary conditions (DBC) as follows:

- Dirichlet: $H_{\mathcal{T},0} := \{u : \mathcal{P} \to \mathbb{R} \mid \forall \sigma \in \mathcal{E}_\partial\ u_\sigma = 0\}$
- Neumann: $H_{\mathcal{T},(0)} := \{u : \mathcal{P} \to \mathbb{R} \mid \forall K \in \mathcal{V},\ \sigma \in \mathcal{E}_\partial \cap \mathcal{E}_K : \partial_{K,\sigma} u = 0,\ \sum_K m_k u_K = 0\}$

- Periodic: $\Omega = \bigotimes_{i=1}^{d}[0,\omega_i)$ is the domain and $(\boldsymbol{e}_i)_{i=1,\dots,d}$ the canonical basis. $\mathcal{P}_\partial = \varnothing$ and $\mathcal{P} = \mathcal{P}_{int}$ is extended adding $\tilde{\mathcal{P}} := \bigcup_{z\in\mathbb{Z}^d}\left(\mathrm{diag}(\omega_i)_{i=1,\dots,d}z + \mathcal{P}\right)$ and

$$H_{\mathcal{T},\mathrm{per}} := \left\{ u : \tilde{\mathcal{P}} \to \mathbb{R} \mid \sum_K m_k u_K = 0 \text{ and } \forall K :\ u_K = u_{K+\omega_i \boldsymbol{e}_i} \right\} .$$

In the following, we always match discrete with the corresponding continous BC. When there is no need to distinguish between the cases, we simply write $H^2_{\mathrm{BC}}(\Omega)$ and $H_{\mathcal{T},\mathrm{BC}}$ and use the index BC accordingly throughout this work. We study the discrete equation (1.3) i.e.,

$$\forall K \in \mathcal{V} : \quad \sum_{\sigma\in\mathcal{E}_K} m_\sigma \kappa_\sigma \partial_{K,\sigma} u_{\mathcal{T}} = m_K f_K ,$$

in either one of the spaces $H_{\mathcal{T},0}$, $H_{\mathcal{T},(0)}$ or $H_{\mathcal{T},\mathrm{per}}$ and with the additional condition $\int_\Omega \kappa u = 0$ in case of Neumann or periodic boundary conditions (BC) i.e. $\sum_K m_K u_{\mathcal{T},K} = 0$.

2.2 Convergence in the energy norm

Defining $L^2(\mathcal{T}) := \{v | \mathcal{P}_{int} \to \mathbb{R}\}$ and

$$\|v\|^2_{L^2(\mathcal{T})} := \sum_{K\in\mathcal{V}} m_K v_K^2 , \qquad \|v\|^2_{H_{\mathcal{T},\kappa}} := \sum_{\sigma\in\mathcal{E}} m_\sigma h_\sigma \kappa_\sigma \left|\partial_\sigma v\right|^2 , \tag{2.2}$$

we observe that $\|\cdot\|^2_{L^2(\mathcal{T})}$ evidently is a norm. Using $\kappa_\sigma > 0$ for all $\sigma \in \mathcal{E}$ and a classical discrete Poincaré inequality [5] which we recall after Theorem 2.5 below, $\|\cdot\|^2_{H_{\mathcal{T},\kappa}}$ is indeed a norm on $H_{\mathcal{T},\mathrm{BC}}$ and we refer to $H_{\mathcal{T},\mathrm{BC}}$ equipped with the norm $\|\cdot\|^2_{H_{\mathcal{T},\kappa}}$ as $H_{\mathcal{T},\kappa,\mathrm{BC}}$. Then solving (1.3) is equivalent with minimizing the functional

$$u_{\mathcal{T}} \mapsto \frac{1}{2}\|u_{\mathcal{T}}\|^2_{H_{\mathcal{T},\kappa}} - \sum_K m_K f_K u_{\mathcal{T},K} . \tag{2.3}$$

In order to match discrete with continuous functions, we introduce the pair of operators

$$\begin{aligned} &\tilde{\mathcal{R}}_{\mathcal{T}} : L^2(\Omega) \to L^2(\mathcal{T}) , \quad \forall i : \left(\tilde{\mathcal{R}}_{\mathcal{T}} u\right)_K := \fint_{\mathbb{B}_{r_K}(x_K)} u , \\ &\mathcal{R}^*_{\mathcal{T}} : L^2(\mathcal{T}) \to L^2(\Omega) , \qquad \left(\mathcal{R}^*_{\mathcal{T}} u\right)(x) := u_K \text{ if } x \in K \end{aligned} \tag{2.4}$$

with $\tilde{\mathcal{R}}_{\mathcal{T}}\mathcal{R}^*_{\mathcal{T}} u = u$. On the other hand, we find the following.

Lemma 2.4. *There exists a constant $C > 0$ depending only on d such that for every bounded polygonal domain Ω, every admissible grid $\mathcal{T} = (\mathcal{T},\mathcal{E},\mathcal{P})$ on Ω, every positive function $\alpha : \mathcal{V} \to \mathbb{R}$ and every $u \in H^1(\Omega)$ it holds*

$$\left\| \sqrt{\mathcal{R}^*_{\mathcal{T}}\alpha_{\mathcal{T}}} \left(\mathcal{R}^*_{\mathcal{T}}\tilde{\mathcal{R}}_{\mathcal{T}} u - u\right) \right\|_{L^2(\Omega)} < \left(\sum_K \alpha_{\mathcal{T},K} R_K^2 \left(\frac{r_K}{R_K}\right)^{1-d} \left(1 + \frac{r_K}{R_K}\right) \|\nabla u\|^2_{L^2(K)} \right)^{\frac{1}{2}} .$$

Proof. This follows from application of Lemma A.1 to $\mathcal{R}^*_{\mathcal{T}}\tilde{\mathcal{R}}_{\mathcal{T}} u - u$ on every cell $K \in \mathcal{V}$, where we $\int_{\mathbb{B}_{r_K}(x_k)} \mathcal{R}^*_{\mathcal{T}}\tilde{\mathcal{R}}_{\mathcal{T}} u - u = 0$ and thus we obtain for some $C > 0$ depending only on d that

$$\int_K \left|\mathcal{R}^*_{\mathcal{T}}\tilde{\mathcal{R}}_{\mathcal{T}} u - u\right|^2 \le C R_K^2 \left(\frac{r_K}{R_K}\right)^{1-d} \left(1 + \frac{r_K}{R_K}\right) \|\nabla u\|^2_{L^2(K)} .$$

□

We extend $\tilde{\mathcal{R}}_{\mathcal{T}}$ to account for discrete Dirichlet BC by $\left(\mathcal{R}_{\mathcal{T},0}u\right)_K := \left(\tilde{\mathcal{R}}_{\mathcal{T}}u\right)_K$ and

$$\forall \sigma \in \mathcal{E}_\partial : \quad \left(\mathcal{R}_{\mathcal{T},0}u\right)_\sigma := 0\,, \tag{2.5}$$

and for Neumann BC by$\mathcal{R}_{\mathcal{T},(0)}u := \tilde{\mathcal{R}}_{\mathcal{T}}u - \left(\sum_K m_K \left(\tilde{\mathcal{R}}_{\mathcal{T}_h}u\right)_K\right)$ and

$$\forall \sigma \in \mathcal{E}_\partial : \quad \left(\mathcal{R}_{\mathcal{T},(0)}u\right)_\sigma := \left(\mathcal{R}_{\mathcal{T}}u\right)_K\,, \quad K \in \mathcal{V}_\sigma\,. \tag{2.6}$$

For periodic BC, we set $\mathcal{R}_{\mathcal{T},\mathrm{per}}u := \tilde{\mathcal{R}}_{\mathcal{T}}u - \left(\sum_K m_K \left(\tilde{\mathcal{R}}_{\mathcal{T}_h}u\right)_K\right)$ and find the general relation $\mathcal{R}_{\mathcal{T},\mathrm{BC}} : H^2_{\mathrm{BC}}(\Omega) \to H_{\mathcal{T},\mathrm{BC}}$.

Theorem 2.5. *Given a polygonal bounded domain $\boldsymbol{Q} \subset \mathbb{R}^d$ and $u \in H^2(\boldsymbol{Q})$ a solution to* (1.2) *with $f \in L^2(\boldsymbol{Q})$ satisfying the boundary conditions* BC *then for every admissible mesh $\mathcal{T}$ it holds: there exists a unique solution $u_{\mathcal{T}}$ to* (1.3) *for $f_{\mathcal{T}}$ given by* (1.3) *satisfying the discrete boundary conditions* BC*. Furthermore*

$$\left\|u_{\mathcal{T}} - \mathcal{R}_{\mathcal{T},\mathrm{BC}}u\right\|_{H_{\mathcal{T},\kappa}} \le \left(I_{1,\mathcal{T}}(u) + I_{2,\mathcal{T}}(u)\right)\,, \tag{2.7}$$

$$I_{1,\mathcal{T}}(u) = \left(\sum_{\sigma\in\mathcal{E}} h_\sigma m_\sigma \kappa_\sigma^{-1} \left(⨍_\sigma |\kappa - \kappa_\sigma|\,|\nabla u|\right)^2\right)^{\frac{1}{2}}\,,$$

$$I_{2,\mathcal{T}}(u) = \left(\sum_{\sigma\in\mathcal{E}} m_\sigma \kappa_\sigma h_\sigma \left(⨍_\sigma \nabla u\cdot\nu_{\sigma,K} - \partial_{\sigma,K}\mathcal{R}_{\mathcal{T}}u\right)^2\right)^{\frac{1}{2}}\,.$$

Furthermore, there exists a constant $C > 0$ depending only on d such that for every $u \in H^2(\boldsymbol{Q}) \cap H^1_0(\boldsymbol{Q})$ the following holds:

$$|I_{1,\mathcal{T}}(u)|^2 \le C\left(\sum_K \frac{R_K^3}{r_K^3} R_K^2 \left\|\sqrt{\kappa}\nabla u\right\|^2_{H^1(K)} \left\|\nabla\kappa\right\|^2_{L^\infty(K_\sigma)} \sum_{\sigma\in\mathcal{E}_K} ⨍_\sigma \frac{1}{\kappa\kappa_\sigma}\right)\,, \tag{2.8}$$

$$|I_{1,\mathcal{T}}(u)|^2 \le C\left(\sum_K \frac{R_K^3}{r_K^3} R_K^2 \left\|\nabla u\right\|^2_{H^1(K)} \left\|\nabla\kappa\right\|^2_{L^\infty(K_\sigma)} \sum_{\sigma\in\mathcal{E}_K} \frac{1}{\kappa_\sigma}\right)\,, \tag{2.9}$$

$$|I_{2,\mathcal{T}}(u)|^2 \le C\left(\sum_K R_K^2 \left(\frac{R_K}{r_k}\right)^{d+1} \left\|\nabla^2 u\right\|^2_{L^2(K)} \sum_{\sigma\in\mathcal{E}_K} \kappa_\sigma\right)\,. \tag{2.10}$$

Remark 2.6. Theorem 2.5 shows that the error (2.7) can be split into two parts. The first part $I_{1,\mathcal{T}}(u)$ accounts for errors due to non-uniformity of κ. That said, if $\kappa = 1$ we find $I_{1,\mathcal{T}}(u) = 0$. It is less sensitive to the excentricity than the second error term $I_{2,\mathcal{T}}(u)$. To see this we have a look at (2.9), which combined with $\frac{\kappa(x)}{\sqrt{\kappa_\sigma}} - \sqrt{\kappa_\sigma} \approx \nabla\sqrt{\kappa}R_K$ shows that for bounded $\frac{R_K^2}{r_K^2}$ we have $|I_{1,\mathcal{T}}(u)|^2 \le C\sum_K R_K^2 \frac{R_K^2}{r_K^2}\left\|\nabla u\right\|^2_{H^1(K)}$. Provided $\left\|\nabla u\right\|^2_{H^1(K)}$ and $\left\|\nabla^2 u\right\|^2_{L^2(K)}$ are of comparable size, the geometric artefacts such as excentricity dominate the error estimate via $I_{2,\mathcal{T}}(u)$ as can be seen from the estimate (2.10).

The proof of Theorem 2.5 is given in Section 3. It partially relies on the well known uniform Poincaré inequality (Section 10.2 of [5]) polytopal meshes of the following form: there exists a constant C depending only on Ω such that

$$\forall v \in L^2(\mathcal{T}) : \qquad \left\|v\right\|^2_{L^2(\mathcal{T})} \le C\sum_{\sigma\in\mathcal{E}} m_\sigma h_\sigma \left|\partial_\sigma v\right|^2 + 2\left|\Omega\right|^{-1}\left|\overline{v}\right|^2\,, \qquad \overline{v} = \sum_K m_k v_K\,, \tag{2.11}$$

together with a corresponding improved Poincaré inequality on $H_{\mathcal{T},0}$:

$$\forall v \in H_{\mathcal{T},0} : \qquad \left\|v\right\|^2_{L^2(\mathcal{T})} \le C\sum_{\sigma\in\mathcal{E}} m_\sigma h_\sigma \left|\partial_\sigma v\right|^2\,.$$

The Poincaré inequality is typically used to obtain discrete L^2-convergence of the solutions from the convergence in the energy norm. However, in case κ is degenerate, the above Poincaré inequality is no longer applicable.

We close this section by the following observation which is a direct consequence of Theorem 2.5 and Hölder's inequality.

Corollary 2.7. *In the setting of of Theorem 2.5 let additionally $u \in W^{2,p}(\boldsymbol{Q})$, $p > 2$. Then*

$$\|u_{\mathcal{T}} - \mathcal{R}_{\mathcal{T},\mathrm{BC}} u\|_{H_{\mathcal{T},\kappa}} \leq C(\kappa) \left(\sum_K m_K \left(R_K^2 \left(\frac{R_K}{r_k} \right)^{d+1} \right)^{\frac{p}{p-2}} \right)^{\frac{p-2}{2p}} \|\nabla u\|_{W^{1,p}(\Omega)} .$$

2.3 Weighted Poincaré inequality and convergence in L^2

Given $x, y \in \mathbb{R}^d$ we denote by $[x, y]$ the closed straight line segment connecting x and y.

Definition 2.8. Let Ω be simply connected, let $\omega \subset \Omega$ be open convex and let $\kappa : \overline{\Omega} \to \mathbb{R}$ be a simple piecewise constant function. Given $\kappa_0 \geq \kappa_1 > 0$ and denoting $\omega(\kappa, \kappa_0) := \{x \in \omega | \, \kappa(x) \geq \kappa_0\}$ we say that κ is pseudo monotone on ω w.r.t κ_0, κ_1 and an open ball $\mathbb{B} \subset \omega(\kappa, \kappa_0)$ if for every $x \in \omega \backslash \omega(\kappa, \kappa_0)$ and every $y \in \mathbb{B}$ there exists $z \in \partial\omega(\kappa, \kappa_0)$ such that $t \mapsto \kappa(x + t(z - x))$ is monotone increasing on $[0, 1]$ and if κ restricted to the closed convex hull of $\omega(\kappa, \kappa_0)$ is bigger or equal to κ_1.

The concept of pseudo monotonicity can be defined also for piecewise continuous functions. However, pseudo monotonicity of κ does not imply pseudo monotonicity of $\mathcal{R}_{\mathcal{T}}^* \kappa_{\mathcal{T}}$ on a given admissible mesh $\mathcal{T}$. This is a severe problem for the formulation and the proof of discrete weighted Poincaré inequalities. We circumvent this problem by the following construction.

Assumption 2.9 (Pseudo-monotone κ)**.** *Ω is simply connected, and there exists a partition of Ω into open convex subsets $(\omega_i)_{i=1,\dots,N}$ such that $\overline{\Omega} = \bigcup_i \overline{\omega_i}$ and $\bigcup \omega_i$ is connected (but not necessarily $\bigcup_i \omega_i = \Omega$). Assume for some constants $\kappa_0 \geq \kappa_1 > 0$ that for every $i, j \in \{1, \dots, N\}$ with $\omega_i \cap \omega_j \neq \varnothing$ there exists a ball $\mathbb{B}_{ij} \subset \omega_i(\kappa, \kappa_0) \cap \omega_j(\kappa, \kappa_0)$ such that $\kappa|_{\mathbb{B}_{ij}} \geq \kappa_0$ and κ is pseudo monotone on ω_i and ω_j w.r.t κ_0 and κ_1 and $\mathbb{B}_{ij}$.*

Definition 2.10. A function $\kappa \in C(\overline{\Omega})$ such that Assumption 2.9 can be satisfied for some pair κ_0, κ_1 is called piecewise pseudo monotone.

The family of functions κ and sets Ω such that Assumption 2.9 can be satisfied is large.

Example 2.11. Consider the following situations

1 Let $f \in C(\mathbb{R})$, $f(0) = 0$ and f strictly monotone increasing. Let $\kappa(x) = f(|x|)$, $\Omega = [0, 1]^d$, $\kappa_0 = f(\frac{1}{2})$, $\kappa_1 := \inf_{x \in \mathrm{conv}\left(\Omega \backslash \mathbb{B}_{\frac{1}{2}}(0)\right)} f(|x|)$. Then κ is pseudo monotone w.r.t κ_0, κ_1 and $\mathbb{B} = \mathbb{B}_{0.1}(0.9\mathbb{I})$, $\mathbb{I} = (1, \dots, 1)$.

2 Let $f \in C(\mathbb{R})$, $f(0) = 0$ and f strictly monotone increasing. In this setting let $\kappa(x) = f(|x|)$ is piecewise pseudo monotone on $\Omega = [-1, 1]^d$.

3 Let $f \in C(\mathbb{R})$, $f(0) = 0$ and f strictly monotone increasing, let Ω be a bounded convex polygonal domain and let $\kappa(x) = f(\mathrm{dist}(x, \partial\Omega))$. Then κ is pseudo monotone for some $\kappa_0, \kappa_1 > 0$ and some ball $\mathbb{B} \subset \Omega$ with the center in the x_0 such that $\kappa(x_0)$ is maximal.

4 Let $f_1, f_2 \in C(\mathbb{R})$, $f_1(0) = f_2(0) = 0$ and f_1, f_2 strictly monotone increasing. In this setting let $\kappa(x) = f_1(|x|)\, f_2(\operatorname{dist}(x, \partial\Omega))$. Then κ is pseudo monotone for some $\kappa_0, \kappa_1 > 0$ and some ball $\mathbb{B} \subset \Omega$.

Proof. Point 1. is straight forward to prove. In the second statement we rely on 1. splitting Ω into $2d$ pairwise disjoint domains being rotated versions of $\omega_1 = (0,1)^d$ and connecting them by bars as in the following example: given ω_1 and $\omega_2 = \omega_1 - (2, 0, \dots, 0)$ with the balls $\mathbb{B}_1 = \mathbb{B}_{0.1}(0.9\mathbb{I})$ as well as $\mathbb{B}_2 = \mathbb{B}_{0.1}(0.9\mathbb{I} - (2, 0, \dots, 0))$ and $\omega_3 := \operatorname{conv}(\mathbb{B}_1 \cup \mathbb{B}_2)$. Then $\mathbb{B}_{1,3} = \mathbb{B}_1$ and $\mathbb{B}_{1,2} = \mathbb{B}_2$. Proceed accordingly for the rest of the domain. Point 3. is straight forward and 4. is a combination of 2. and 3. □

In the formulation of the following property, we rely on the set $\mathcal{E}_{\mathcal{T},x,y}$ defined by two points $x, y \in \Omega$ and a mesh $\mathcal{T} = (\mathcal{P}, \mathcal{V}, \mathcal{E})$:

$$\mathcal{E}_{\mathcal{T},x,y} := \{\sigma \in \mathcal{E} |\ [x,y] \cap \sigma \neq \varnothing\}\ .$$

Definition 2.12. Let Assumption 2.9 hold and let $\mathcal{T}$ be an admissible mesh on Ω. We define $\kappa_{\mathcal{T}}(x) := \left(\mathcal{R}^*_{\mathcal{T}}\kappa_{\mathcal{T}}\right)(x)$ and the following function for $x \in \omega_i$ and corresponding $\mathbb{B}_{ij} \subset \omega_i$:

$$a_{\kappa,\mathcal{T}}(x) := \min\left\{\left(\mathcal{R}^*_{\mathcal{T}}\kappa_{\mathcal{T}}\right)(x),\ \inf_{y\in\mathbb{B}_{ij}} \inf_{\sigma\in\mathcal{E}_{\mathcal{T},x,y}} \kappa_\sigma\right\},$$

$$\tilde{\kappa}_{\mathcal{T}}(x) := \begin{cases} \left(\mathcal{R}^*_{\mathcal{T}}\kappa_{\mathcal{T}}\right)(x) & \text{if } \left(\mathcal{R}^*_{\mathcal{T}}\kappa_{\mathcal{T}}\right)(x) \geq \kappa_0 \text{ and } a_{\kappa,\mathcal{T}}(x) \geq \kappa_1 \\ a_{\kappa,\mathcal{T}}(x) & \text{else} \end{cases}.$$

Corollary 2.13. *Let $O := \kappa^{-1}(0)$. For every $\varepsilon > 0$ there exists $C > 0$ such that the following holds: For every admissible mesh $\mathcal{T}$ and $K \in \mathcal{V}$:*

$$\left\|\mathcal{R}^*_{\mathcal{T}}\kappa_{\mathcal{T}} - \kappa\right\|_{C(\overline{K}\cap\overline{\Omega\setminus\mathbb{B}_\varepsilon(O)})} + \left\|\mathcal{R}^*_{\mathcal{T}}\tilde{\kappa}_{\mathcal{T}} - \kappa\right\|_{C(\overline{K}\cap\overline{\Omega\setminus\mathbb{B}_\varepsilon(O)})} \leq C R_K \left\|\kappa\right\|_{C(\overline{K})}.$$

Proof. This follows from $\kappa = \exp(-\frac{1}{2}V)$ and $\nabla\kappa = -\frac{1}{2}\kappa\nabla V$. □

Next, we introduce the notation $\tilde{\kappa}_{\mathcal{T},K} := m_K^{-1}\int_K \tilde{\kappa}_{\mathcal{T}}$. Based on this we write for $u \in L^2(\mathcal{T})$:

$$\overline{\kappa}^{\mathcal{V}} := \int_\Omega \tilde{\kappa}_{\mathcal{T}}(x)\,, \qquad \overline{u}^{\tilde{\kappa}} := \frac{1}{\overline{\kappa}^{\mathcal{V}}}\int_\Omega \tilde{\kappa}_{\mathcal{T}}\mathcal{R}^*_{\mathcal{T}}u\,.$$

Theorem 2.14. *Under the above assumptions on Ω and κ let Assumption 2.9 hold for $\tilde{\Omega} \subset \Omega$, for $\kappa_0 > 0$, $(\omega_i)_i$ and an admissible mesh $\mathcal{T}$. Then there exists a constant C depending only on d, $\tilde{\Omega}$, $C(\mathcal{T}, \kappa_0)$, κ_0 and $\|\kappa\|_\infty$ such that*

$$\sum_K \tilde{\kappa}_{\mathcal{T},K} m_K \left(u_K - \overline{u}^{\tilde{\kappa}}\right)^2 \leq C\left\|u\right\|^2_{H_{\mathcal{T},\kappa}}. \tag{2.12}$$

The proof of Theorem 2.14 is given in Section 4.

Remark 2.15. To understand the formulation of Theorem 2.14 in terms of $\tilde{\Omega} \subset \Omega$ consider $O := \kappa^{-1}(0)$ as in Corollary 2.13. Then we observe for the discrete solution $u_{\mathcal{T}} - \mathcal{R}_{\mathcal{T},\mathrm{BC}}u$ and the continuous solution u that

$$\int_{\Omega\setminus O} \tilde{\kappa}_{\mathcal{T}}\left(\mathcal{R}^*_{\mathcal{T}}u_{\mathcal{T}} - u\right)^2 \leq 2\sum_K |K\cap(\Omega\setminus O)|\,\tilde{\kappa}_{\mathcal{T},K}\left(u_{\mathcal{T}} - \mathcal{R}_{\mathcal{T},\mathrm{BC}}u\right)^2 + 2\int_{\Omega\setminus O}\tilde{\kappa}_{\mathcal{T}}\left(\mathcal{R}_{\mathcal{T},\mathrm{BC}}u - u\right)^2$$

and on behalf of Lemma 2.4 and Corollary 2.13

$$\int_{\Omega\setminus O} \tilde{\kappa}_{\mathcal{T}} \left(\mathcal{R}_{\mathcal{T},\mathrm{BC}} u - u\right)^2 \leq \sum_K \left(2 + CR_K\right) R_K^2 \left(\frac{r_K}{R_K}\right)^{1-d} \left\|\sqrt{\kappa}\nabla u\right\|^2_{L^2(K)} .$$

Hence, Theorem 2.14 combined with Theorem 2.5 yields

$$\begin{aligned}\int_{\Omega\setminus O} \tilde{\kappa}_{\mathcal{T}} \left(\mathcal{R}^*_{\mathcal{T}} u_{\mathcal{T}} - u\right)^2 \leq \sum_K \left(2 + CR_K + C\right) R_K^2 \left(\frac{r_K}{R_K}\right)^{1-d} \left\|\sqrt{\kappa}\nabla u\right\|^2_{H^1(K)} \\ + \sum_K R_K^2 \left(\frac{R_K}{r_k}\right)^{d+1} \left\|\nabla^2 u\right\|^2_{L^2(K)} \sum_{\sigma\in\mathcal{E}_K} \kappa_\sigma . \end{aligned} \tag{2.13}$$

3 Proof of Theorem 2.5

Lemma 3.1. *Let $K \subset \mathbb{R}^d$ be a bounded, convex and polytopal domain with $0 < r < R < +\infty$ such that $\mathbb{B}_r(0) \subset K \subset \mathbb{B}_R(0)$. Then there exists C depending only on the dimension d such that*

$$\forall u \in H^1(K): \qquad \|u\|^2_{L^2(\partial K)} \leq C\frac{R^2}{r^3} \|u\|^2_{H^1(K)} .$$

Proof. The surface ∂K is piecewise Lipschitz with a Lipschitz constant bounded by R/r. Furthermore, ∂K can be covered by balls of radius $\frac{r}{2}$ with the total number of balls covering any $x \in \partial K$ bounded by $d!$. The norm of the local trace operator in any of these small balls B is proportional to the Lipschitz constant R/r. The pre-factor r^{-1} comes from the partition of unity which is necessary to glue together the local trace operators. Alternatively, one may first consider $r = 1$ and apply a scaling argument. □

Proof of Theorem 2.5. Since $\kappa_\sigma > 0$ for every $\sigma \in \mathcal{E}$ by construction, the discrete Poincaré inequality (2.11) yields that the left hand side of (1.3) is invertible and there exists a unique solution $u_{\mathcal{T}} \in H_{\mathcal{T},\mathrm{BC}}$ of (1.3). Testing with $u_{\mathcal{T}}$ and using (2.11) yields the discrete apriori estimate

$$\|u_{\mathcal{T}}\|_{H_{\mathcal{T},\kappa}} \leq C \|f\|_{L^2(\mathcal{T})} , \tag{3.1}$$

where $C > 0$ depends on $\mathcal{T}$.

Equivalently the unique existence of $u_{\mathcal{T}}$ follows from considering the equivalent variational minimization problem (2.3).

Proof of (2.7)*:* For arbitrary $v \in H_{\mathcal{T}}$ we observe with help of (1.2) and (1.3)

$$\begin{aligned}\sum_{\sigma\in\mathcal{E}} \kappa_\sigma m_\sigma h_\sigma \partial_\sigma v \, \partial_\sigma \left(u_{\mathcal{T}} - \mathcal{R}_{\mathcal{T},\mathrm{BC}} u\right) &= \sum_K v_K \int_K f - \sum_{\sigma\in\mathcal{E}} \kappa_\sigma m_\sigma h_\sigma \partial_\sigma v \, \partial_\sigma \mathcal{R}_{\mathcal{T},\mathrm{BC}} u \\ &= \sum_K v_K \int_{\partial K} \kappa\nabla u \cdot \nu_{\sigma_K} - \sum_{\sigma\in\mathcal{E}} \kappa_\sigma m_\sigma h_\sigma \partial_\sigma v \, \partial_\sigma \mathcal{R}_{\mathcal{T},\mathrm{BC}} u .\end{aligned}$$

From here (2.7) follows with $\kappa\nabla u = \kappa_\sigma \nabla u + \left(\kappa - \kappa_\sigma\right)\nabla u$.

Proof of (2.8) *and* (2.9)*:* We only prove (2.8) as (2.9) can be proved in a similar way. Since κ is continuous, for every $\sigma \in \mathcal{E}$ there exists $K_\sigma \in \mathcal{V}_\sigma$ and $x_\sigma \in K_\sigma$ such that $\kappa(x_\sigma) = \kappa_\sigma$. But then we

can apply Hölders Theorem and Lemma 3.1 to find

$$\begin{aligned}|I_{1,\mathcal{T}}(u)|^2 &\le \sum_\sigma h_\sigma \fint_\sigma \frac{(\kappa-\kappa_\sigma)^2}{\kappa\kappa_\sigma}\int_\sigma \kappa|\nabla u|^2\\ &\le \sum_\sigma h_\sigma \left\|\nabla\kappa\right\|^2_{L^\infty(K_\sigma)} \fint_\sigma \frac{R^2_{K_\sigma}}{\kappa\kappa_\sigma}\frac{R^2_{K_\sigma}}{r^3_{K_\sigma}}\left\|\kappa|\nabla u|\right\|^2_{H^1(K_\sigma)}\end{aligned}$$

and from here we conclude with $h_\sigma \le 2R_{K_\sigma}$.

Proof of (2.10)*:*

We first consider some $\sigma \in \mathcal{E}_{int}$. To simplify calculations, let $\sigma \subset \{0\}\times\mathbb{R}^{d-1}$ and let $x_- < 0 < x_+$ with $r_+, r_- > 0$, $|x_\pm| \ge 2r_\pm$. The values $r_\pm$ correspond to the radii r_K and r_L for the neighboring cells K and L of σ. Denoting $\boldsymbol{Q}_\pm := \mathbb{B}_{r_\pm}(x_\pm \boldsymbol{e}_1)$ and exploiting Taylor's formula we have for every $y \in \sigma$, $z \in \mathbb{B}_{r_\pm}(0)$

$$\begin{aligned}u(x_\pm\boldsymbol{e}_1+z) &= u(y) + \nabla u(y)\cdot(x_\pm\boldsymbol{e}_1+z-y) + I(x_\pm\boldsymbol{e}_1+z-y,y)\,, \qquad (3.2)\\ I(\boldsymbol{x},y) &= \int_0^1 \boldsymbol{x}^\top D^2u(y+t\boldsymbol{x})\boldsymbol{x}\,(1-t)\,\mathrm{d}t\,,\end{aligned}$$

where D^2u is the Hessian of u and using

$$\fint_{\boldsymbol{Q}_\pm}\nabla u(y)\cdot(z-y)\,\mathrm{d}z = \fint_{\mathbb{B}_{r_\pm}(0)}\nabla u(y)\cdot(x_\pm\boldsymbol{e}_1+z-y)\,\mathrm{d}z = \nabla u(y)\cdot(x_\pm\boldsymbol{e}_1-y)$$

we have

$$\fint_{\boldsymbol{Q}_+}u - \fint_{\boldsymbol{Q}_-}u = \fint_\sigma\left[\nabla u(y)\cdot(x_+-x_-)\,\boldsymbol{e}_1 + \sum_\pm(\pm1)\fint_{\mathbb{B}_{r_\pm}(0)} I((x_\pm+z)\,x_\pm\boldsymbol{e}_1+z-y,y)\,\mathrm{d}z\right]\mathrm{d}y$$

and from there

$$\left||x_+-x_-|\int_\sigma\partial_1u - m_\sigma\left(\fint_{\boldsymbol{Q}_+}u-\fint_{\boldsymbol{Q}_-}u\right)\right| \le \int_\sigma\sum_\pm\fint_{\mathbb{B}_{r_\pm}(0)}|I(x_\pm\boldsymbol{e}_1+z-y,y)|\,\mathrm{d}z\,\mathrm{d}y\,. \qquad (3.3)$$

The last one implies particularly for $\sigma \in \mathcal{E}_{int}$

$$|I_{2,\mathcal{T}}(u)|^2 \le \sum_{\sigma\in\mathcal{E}} m_\sigma h_\sigma^{-1}\left|\fint_\sigma\sum_\pm\fint_{\mathbb{B}_{r_\pm}(0)}|I(x_\pm\boldsymbol{e}_1+z-y,y)|\,\mathrm{d}z\,\mathrm{d}y\right|^2.$$

Furthermore, we will need the following general observation: if $f \in L^1(\mathbb{R}^d)$ is a positive function and A_1, A_2 are measurable sets, it holds

$$\begin{aligned}\int_{A_1}\int_{A_2} f(x+y)\,\mathrm{d}x\,\mathrm{d}y &= \int_{\mathbb{R}^d} f(y)\int_{\mathbb{R}^d}\chi_{A_1}(x)\chi_{A_2}(y-x)\,\mathrm{d}x\,\mathrm{d}y\\ &\le \int_{\mathbb{R}^d}\chi_{\mathrm{conv}(A_1+A_2)}(x)\,|A_2|\;f(x)\,\mathrm{d}x \qquad (3.4)\end{aligned}$$

We introduce the "cone" $\mathfrak{C}_\pm := \{(1-t)y + tx_\pm\boldsymbol{e}_1 + tz : t\in[0,1],\, y\in\sigma,\, z\in\mathbb{B}_{r_\pm}(0)\}$ and the "slices" $\mathfrak{C}_{\pm,z_1} := \{(z_1,y)\in\mathfrak{C}_\pm :\ y\in\mathbb{R}^{d-1}\}$. For readability, we furtheron drop the $\pm$, write $f = |D^2u|^2$ and find

with the transformation $(z_1, \tilde{z}) = tz$, $z_1 \in \mathbb{R}$, $\tilde{z} \in \mathbb{R}^{d-1}$ and $\tilde{y} = (1-t)y$ and by Jensen's inequality that

$$
\begin{aligned}
&m_\sigma \left(\fint_\sigma \mathrm{d}y \fint_{\mathbb{B}_r(0)} \mathrm{d}z \, |I(x\boldsymbol{e}_1 + z - y, y)| \right)^2 \\
&\qquad \leq \left(\int_\sigma \mathrm{d}y \fint_{\mathbb{B}_r(0)} \mathrm{d}z \int_0^1 |x\boldsymbol{e}_1 + z - y|^4 \left| D^2 u \right|^2 (y + t\,(x\boldsymbol{e}_1 + z - y))\,(1-t)^2 \,\mathrm{d}t \right)^2 \\
&\leq \frac{1}{|\mathbb{S}^{d-1}|\, r^d} \int_0^1 \mathrm{d}t \int_{(1-t)\sigma} \mathrm{d}\tilde{y} \int_{-tr}^{tr} \mathrm{d}z_1 \int_{\mathbb{B}^{d-1}_{\sqrt{(tr)^2 - z_1^2}}(0)} \mathrm{d}\tilde{z}\, f(\tilde{y} + tx\boldsymbol{e}_1 + \tilde{z} + z_1\boldsymbol{e}_1) \frac{1}{t^d (t-1)^{d-3}} R^4 \,,
\end{aligned}
$$

where $R \geq \sup\{(x\boldsymbol{e}_1 + z - y) \,:\, y \in \sigma,\, z \in \mathbb{B}_r(0)\}$. Exploiting (3.4) in $\tilde{z}$ and $\tilde{y}$ we find for the case $x = x_+$ and $r(t, z_1) := \sqrt{(tr)^2 - z_1^2}$ that

$$
\begin{aligned}
&\int_{(1-t)\sigma} \mathrm{d}\tilde{y} \int_{\mathbb{B}^{d-1}_{\sqrt{(tr)^2 - z_1^2}}(0)} \mathrm{d}\tilde{z}\, f(\tilde{y} + tx\boldsymbol{e}_1 + \tilde{z} + z_1\boldsymbol{e}_1) \\
&\qquad \leq \int_{\mathbb{R}^{d-1}} f(\tilde{y} + tx\boldsymbol{e}_1 + z_1\boldsymbol{e}_1) \begin{cases} \int_{\mathbb{R}^{d-1}} \chi_{\mathbb{B}^{d-1}_{r(t,z_1)} + (1-t)\sigma}(\tilde{y} - \tilde{z})\, \chi_{(1-t)\sigma}(\tilde{z})\, \mathrm{d}\tilde{z}\, \mathrm{d}\tilde{y} \\ \int_{\mathbb{R}^{d-1}} \chi_{\mathbb{B}^{d-1}_{r(t,z_1)} + (1-t)\sigma}(\tilde{y} - \tilde{z})\, \chi_{\mathbb{B}^{d-1}_{r(t,z_1)}}(\tilde{z})\, \mathrm{d}\tilde{z}\, \mathrm{d}\tilde{y} \end{cases} \\
&\qquad \leq \int_{\mathfrak{C}_{+,z_1}} f(\tilde{y} + tx\boldsymbol{e}_1 + z_1\boldsymbol{e}_1) \begin{cases} (1-t)^{d-1} |\sigma| & t \geq \frac{1}{2} \\ (tr)^{d-1} |\mathbb{S}^{d-2}| & t < \frac{1}{2} \end{cases} .
\end{aligned}
$$

Writing $F(tx + z_1) = \int_{\mathfrak{C}_{tx+z_1}} f(y + tx\boldsymbol{e}_1 + z_1\boldsymbol{e}_1)$ we find

$$
\begin{aligned}
\int_0^1 \mathrm{d}t\, \frac{1}{t} \int_{-tr}^{tr} F(tx + z_1) &= \int_0^1 \mathrm{d}t\, t^{-1} \int_{\mathbb{R}} \mathrm{d}z\, \chi_{[-tr, tr]}(z - tx)\, F(z) \\
&= \int_{\mathbb{R}} \mathrm{d}z\, F(z) \int_{\frac{z}{x+r}}^{\frac{z}{x-r}} \mathrm{d}t\, t^{-1} \leq \frac{2r}{x-r} \int_{\mathbb{R}} \mathrm{d}z\, F(z) \,.
\end{aligned}
$$

Due to the definition of r_K and R_K we have $x \geq 2r$ and $R < R_K$. Furthermore, because $\mathbb{B}_{R_K}(x_K) \supset K$ it holds $|\sigma| \leq C R_K^{d-1}$ where C depends only on the dimension and also $h_\sigma > r_K$. From the above we then conclude the proof in case of Neumann and periodic boundary conditions.

In case $\sigma \in \mathcal{E}_\partial$ with $K \in \mathcal{V}_\sigma$ we consider some artificial $x_L := x_K + 2(x_\sigma - x_K)$ and proceed like above with $u_L := 0$.

Combining the above steps leads to (2.10). □

4 Proof of Theorem 2.14

We follow Section 10.2 of [5].

In this proof, it is beneficial to identify every function $u \in L^2(\mathcal{T})$ with $\mathcal{R}_\mathcal{T}^* u$ as an element of $L^2(\Omega)$. In this way we use the notation $u(x) := (\mathcal{R}_\mathcal{T}^* u)(x)$. We then observe that

$$
\overline{\kappa}^\mathcal{V} = \fint_\Omega \tilde{\kappa}_\mathcal{T}(x) \,, \qquad \overline{u}^{\kappa,\mathcal{V}} = \frac{1}{\overline{\kappa}^\mathcal{V}} \int_\Omega \tilde{\kappa}_\mathcal{T} \mathcal{R}_\mathcal{T}^* u \,,
$$

and because $\mathcal{R}^*_{\mathcal{T}} u$ is piecewise constant, inequality (2.12) can be recast into the form

$$\int_\Omega \tilde{\kappa}_{\mathcal{T}} \left(\mathcal{R}^*_{\mathcal{T}} u - \overline{u}^{\kappa,\mathcal{V}} \right)^2 \leq C \left\| u \right\|^2_{H_{\mathcal{T},\kappa}} .$$

Step 1: In this first step, let $\omega = \omega_i$ for some $i \in \{1, \dots, N\}$ and $\mathbb{B} = \mathbb{B}_{ij} \subset \omega$ for some j. Writing for any open subset $A \subset \Omega$

$$\overline{u}^{\tilde{\kappa}_{\mathcal{T}},A} := \frac{1}{\overline{\tilde{\kappa}}^A_{\mathcal{T}}} \int_A \tilde{\kappa}_{\mathcal{T}} u \,, \qquad \overline{\tilde{\kappa}}^A_{\mathcal{T}} := \int_A \tilde{\kappa}_{\mathcal{T}}$$

we show that

$$\int_\omega \tilde{\kappa}_{\mathcal{T}} \left(u - \overline{u}^{\tilde{\kappa}_{\mathcal{T}},\mathbb{B}} \right)^2 \leq \frac{\left\| \kappa \right\|^2_\infty}{\kappa_0 \kappa_1} \left(\mathrm{diam} \omega \right)^2 \left\| u \right\|^2_{H_{\mathcal{T},\kappa}} \tag{4.1}$$

holds independently from $\mathbb{B}$. It holds by Jensens inequality

$$\begin{aligned} \int_\omega \tilde{\kappa}_{\mathcal{T}} \left(u - \overline{u}^{\tilde{\kappa}_{\mathcal{T}},\mathbb{B}} \right)^2 &\leq \int_\omega \tilde{\kappa}_{\mathcal{T}} \left(u - \frac{1}{\int_{\mathbb{B}} \tilde{\kappa}_{\mathcal{T}}} \int_{\mathbb{B}} \tilde{\kappa}_{\mathcal{T}} u \right)^2 \\ &\leq \frac{1}{\overline{\tilde{\kappa}}^{\mathbb{B}}_{\mathcal{T}}} \int_\omega \int_{\mathbb{B}} \tilde{\kappa}_{\mathcal{T}}(x) \tilde{\kappa}_{\mathcal{T}}(y) \left(u(x) - u(y) \right)^2 . \end{aligned}$$

For $\sigma \in \mathcal{E}_{int}$ let $\chi_\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \{0,1\}$ be defined as

$$\chi_\sigma(x,y) = \begin{cases} 1 & \text{if } x, y \in \Omega,\ [x,y] \cap \sigma \neq \varnothing \\ 0 & \text{else} \end{cases} .$$

As explained in the proof of Lemma 10.2 in [5] it holds for every $x \in \omega$, $y \in \mathbb{B}$

$$\left(u(x) - u(y) \right)^2 \leq \mathrm{diam}(\omega) \sum_{\sigma \in \mathcal{E}_{int}} \frac{\left| \partial_\sigma u \right|^2}{h_\sigma c_{\sigma,x-y}} \chi_\sigma(x,y) \,,$$

where $c_{\sigma,x-y} := \left| \frac{y-x}{|x-y|} \cdot \nu_\sigma \right|$. By the definition of $\tilde{\kappa}_{\mathcal{T}}$ it holds for $x \in \omega$, $y \in \mathbb{B}$ either $\tilde{\kappa}_{\mathcal{T}}(x) \leq \inf_{\sigma \in \mathcal{E}_{\mathcal{T},x,y}} \kappa_\sigma$ or $\tilde{\kappa}_{\mathcal{T}}(x) \leq \left\| \kappa \right\|_\infty$ and $\kappa_\sigma \geq \kappa_1$ for every $\sigma \in \mathcal{E}_{\mathcal{T},x,y}$. In both cases it holds $\tilde{\kappa}_{\mathcal{T}}(x) \leq \frac{\|\kappa\|_\infty}{\kappa_1}$ and $\tilde{\kappa}_{\mathcal{T}}(y) \leq \left\| \kappa \right\|_\infty$. Hence wWe may multiply the last inequality with $\kappa(x)\kappa(y)$ and integrate over x and y to find

$$\int_\omega \tilde{\kappa}_{\mathcal{T}} \left(u - \overline{u}^{\kappa,\mathbb{B}} \right)^2 \leq \frac{\left\| \kappa \right\|^2_\infty}{\kappa_0 \kappa_1} \mathrm{diam}(\omega) \int_\omega \int_{\mathbb{B}} \sum_{\sigma \in \mathcal{E}_{int}} \kappa_\sigma \frac{\left| \partial_\sigma u \right|^2}{h_\sigma c_{\sigma,x-y}} \chi_\sigma(x,y) \, \mathrm{d}x \, \mathrm{d}y \,.$$

Now we observe with the transformation $z = x - y$ that

$$\int_\omega \chi_\sigma(x, z+x) \, \mathrm{d}x \leq m_\sigma \left| z \cdot \nu_\sigma \right| \leq m_\sigma \mathrm{diam}(\omega) \, c_{\sigma,z}$$

and hence

$$\int_\omega \int_\omega \frac{\chi_\sigma(x,y)}{c_{\sigma,x-y}} \mathrm{d}y \, \mathrm{d}x \leq m_\sigma \mathrm{diam}(\omega) \, |\mathbb{B}| \,.$$

This yields (4.1).

Step 2: We define for every ω_i the value $m_i(u) := \overline{u}^{\tilde{\kappa}_{\mathcal{T}},\omega_i}$ and for $i \neq j$ with the corresponding ball $\mathbb{B}_{ij} \subset \omega_i \cap \omega_j$

$$m_{\mathcal{T},ij}(u) := \overline{u}^{\tilde{\kappa}_{\mathcal{T}},\mathbb{B}_{ij}} .$$

Estimate (4.1) implies the existence of $C_{ij,i} > 0$ depending only on κ_0, κ_1, $\|\kappa\|_\infty$, Ω, the dimension d and $\mathbb{B}_{ij}$ such that

$$|m_{\mathcal{T},ij}(u) - m_{\mathcal{T},i}(u)| \leq C_{ij,i} \|u\|_{H_{\mathcal{T},\kappa}} . \tag{4.2}$$

From here, we conclude for $\omega_i \cap \omega_j \neq \varnothing$ the existence of $C_{i,j} > 0$ such that

$$|m_{\mathcal{T},i}(u) - m_{\mathcal{T},j}(u)| \leq C_{i,j} \|u\|_{H_{\mathcal{T},\kappa}} . \tag{4.3}$$

Because $\bigcup_i \omega_i$ is connected this implies by (4.2)–(4.3) and the iterated application of the triangle inequality the existence of $C > 0$ such that

$$\sup_{i\in\{1,\dots,N\}} |m_{\mathcal{T},i}(u) - m_{\mathcal{T},1}(u)| + \sup_{\substack{i,j\in\{1,\dots,N\}\\ \omega_j\cap\omega_i\neq\varnothing}} |m_{\mathcal{T},ij}(u) - m_{\mathcal{T},1}(u)| \leq C \|u\|^2_{H_{\mathcal{T},\kappa}} . \tag{4.4}$$

Step 3: Now assume the theorem was wrong, then for every $n \in \mathbb{N}$ there exists an admissible mesh $\mathcal{T}_n = (\mathcal{P}_n, \mathcal{V}_n, \mathcal{E}_n)$ and $u_n \in H_{\mathcal{T}_n,\kappa}$ such that

$$1 = \sum_{K\in\mathcal{V}_n} m_K \tilde{\kappa}_{\mathcal{T}_n,K} \left(u_n - \overline{u}_n^{\tilde{\kappa}_{\mathcal{T}_n}}\right)^2 \geq n \|u_n\|^2_{H_{\mathcal{T}_n,\kappa}} . \tag{4.5}$$

Without loss of generality, we may assume that $\overline{u}^{\kappa,\mathbb{B}_{1,2}} = 1$. Inequality (4.1) implies that

$$\int_{\omega_1} \tilde{\kappa}_{\mathcal{T}_n} (u_n - 1)^2 \to 0 ,$$

which implies for every $x \in \omega_1$ either $u_n \to 1$ or otherwise $\tilde{\kappa}_{\mathcal{T}_n}(x) \to 0$. Furthermore, since κ is essentially bounded we obtain from Hölders inequality

$$\left| \int_{\omega_1} u_n \tilde{\kappa}_{\mathcal{T}_n} - \int_{\omega_1} \tilde{\kappa}_{\mathcal{T}_n} \right| \to 0 ,$$

which implies $m_{\mathcal{T}_n,1} \to 1$. A combination of (4.4) and (4.5) yields $m_{\mathcal{T}_n,ij}(u_n) \to 1$ for every valid pair i, j. But then (4.1) implies for every i that $\int_{\omega_i} \tilde{\kappa}_{\mathcal{T}_n} (u_n - 1)^2 \to 0$ and hence

$$\sum_{K\in\mathcal{V}} m_K \tilde{\kappa}_{\mathcal{T}_n,K} (u_n - 1)^2 \to 0 .$$

By Hölders inequality this implies $\overline{u}_n^{\tilde{\kappa}_{\mathcal{T}_n}} \to 1$, resulting in a contradiction to (4.5).

5 Numerical experiments

As discussed in the introduction, we start by defining the exact solution $u : \mathbb{R}^4 \to \mathbb{R}$ a priori to be

$$u(x) = \begin{cases} p(\|x\|) \cdot x_1 & \text{if } \|x\| < 1 \\ 0 & \text{else} \end{cases} \tag{5.1}$$

with

$$p(x) = 2\left(x + \frac{1}{2}\right)(1 - x)^2 \tag{5.2}$$

being the polynomial of degree 3 with $p(0) = 1$, $p(1) = 0$ and $p'(0) = p'(1) = 0$. Hence u is anti-symmetric in the first and symmetric in all other components and indeed $u \in H^2(\mathbb{R}^4)$ with support $\Omega = \{\|x\| \leq 1\}$.

Choosing $\kappa(x) \equiv 1$ simplifies the study of equation (2.7) by computation of $I_{2,\mathcal{T}}$ alone since $I_{1,\mathcal{T}} = 0$ vanishes for constant κ. Computing the right hand side f from (1.2) with the discretization f_K from (1.3) we know that u by construction is the exact solution to the arising PDE problem.

Given a tesselation $\mathcal{T}$ we compute the nonconstant term in the right hand side of the error bound 2.10:

$$B(u,\mathcal{T}) := \sqrt{\left(\sum_K R_K^2 \left(\frac{R_K}{r_k}\right)^{d+1} \|\nabla^2 u\|^2_{L^2(K)} \sum_{\sigma\in\mathcal{E}_K} \kappa_\sigma\right)} \tag{5.3}$$

by means of Monte Carlo integration of the Hessian, computed by automatic differentiation, with 10 samples per ray and 100 rays per cell. According to Theorem 2.5, for $\kappa \equiv 1$, there should be a constant C such that

$$H(u,\mathcal{T}) := \|u_\mathcal{T} - \mathcal{R}_{\mathcal{T},\mathrm{BC}} u\|_{H_{\mathcal{T},\kappa}} < |I_{2,\mathcal{T}}(u)| \le C \cdot B(u,\mathcal{T})\,. \tag{5.4}$$

The last inequality together with the structure of $B(u,\mathcal{T})$ leads us to three conjectures:

1 For our given exact solution u, the left hand side $H(u,\mathcal{T})$ and the right hand side $B(u,\mathcal{T})$ are functions depending solely on $\mathcal{T}$ and in double logarithmic diagram, the pairs $(B(u,\mathcal{T}), H(u,\mathcal{T}))$ should all lie above a straight line of slope 1.

2 Figure 1 indicates that the meshes generated in our approach are quasi uniform. In view of (1.6) we expect that there exists a linear relation between the average cell diameter of order h, which is proportional to $N^{\frac{1}{d}}$, and the approximation error.

3 Discretization schemes that make use of the apriori knowledge of u and adjust the size of cells in an inverse correlation to $|\nabla u|$ or $|\nabla^2 u|$ should converge better than those discretization schemes that are ignorant to this information.

5.1 Implementation

We generate the required admissible mesh by a Voronoi tesselation $\mathcal{T}$ of N random points x_i, $i = 1, ..., N$, resulting in N cells with the x_i at their respective centers. Such a tessellation in admissible in the sense of Definition 2.1. Unlike regular grids, exhibiting the curse of dimensionality, the choice of Voronoi tesselations is advantageous in the high dimensional setting such as in Molecular Design where one can obtain a tesselation adapted in resolution to a target density by sampling.

However, the computation of the geometric properties of the tesselation, i.e. the cell volumes m_K and their boundary areas m_σ can be challenging. For their computation we developed the Julia package *VoronoiGraph.jl* [13], an open-source implementation of the raycasting Voronoi-tesselation algorithm [11] with corresponding methods to compute the volume, areas as well as monte carlo integrators over the cells, which we require for the evaluation of the error bound $I_{1,\mathcal{T}}$.

In order to be more close to the classical notation in numerical linear algebra, we use i, j as indices instead of K in the following. Once the cell volumes and their boundary areas are computed, the discretized equation (1.3) can be written as the linear system

$$f_\mathcal{T} = Q u_\mathcal{T} \quad \text{where} \quad Q_{ij} = \frac{A_{ij}}{V_i h_{ij}} S(\kappa_i, \kappa_j) \tag{5.5}$$

with areas, volumes and distances denoted by $A_{ij} = m_{\sigma(i,j)}$, $V_i = m_i$ and $h_{ij} = \|x_i - x_j\|$, for all $i \neq j$ such that $x_i \in \Omega$ and $Q_{ii} = -\sum_{j\neq i} Q_{ij}$ on the diagonal.

We conduct the computation for different distributions of cell centers x_i. As a baseline we sample x_i uniformly from the ball with radius 1.5 which properly contains the domain Ω, as well as from a multivariate normal with standard deviation $\frac{1}{2}$. Since the normal distribution gravitates toward the center, we expect better results from the increased resolution inside Ω. To showcase the benefits of adaptive sampling we also sample x_i from rejection sampling using densities proportional to $\|\nabla u\|$ and $\|\nabla^2 u\|$ by means of rejection sampling. Figure 1 depicts the different methods with 300 samples on an only two dimensional domain for illustrative purposes.

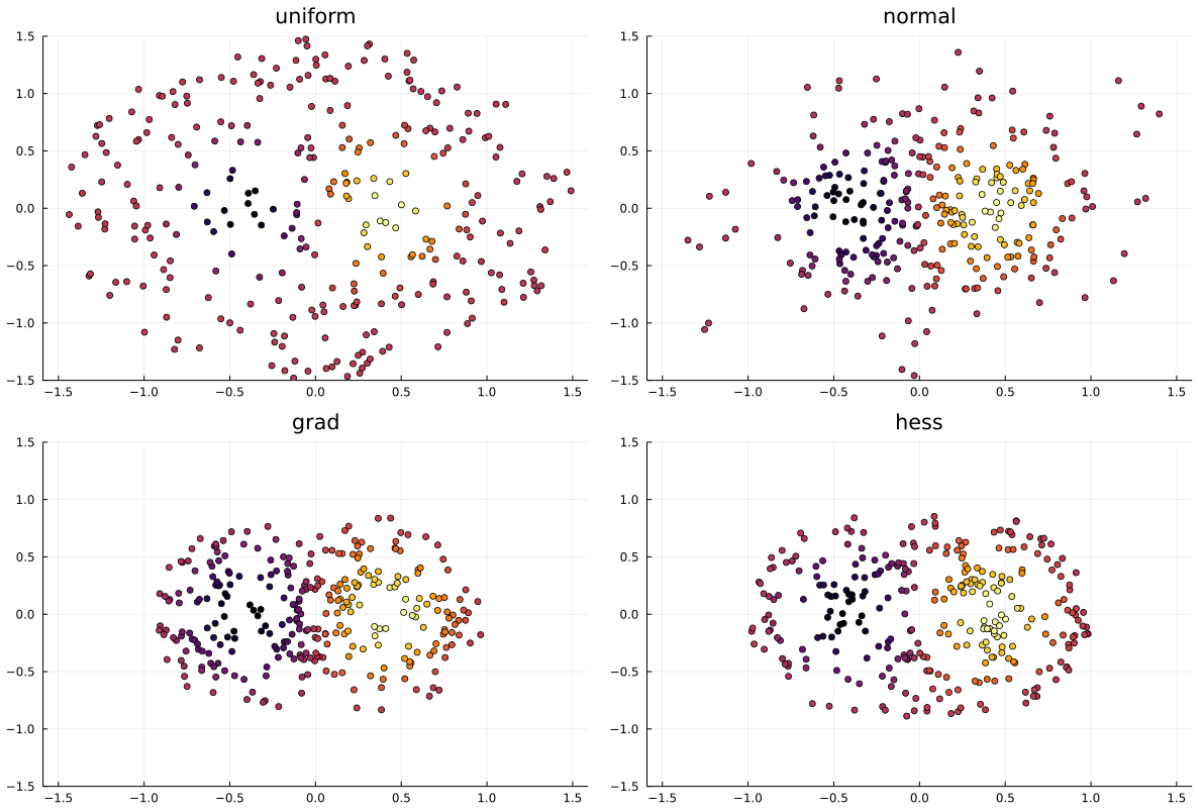

Figure 1: Illustration of the different sampling schemes for 300 points in 2D, corresponding to 90.000 points in 4D: uniform distribution (uniform), normal distribution (normal) and rejection sampling using $|\nabla u|$ (grad) and $|\nabla^2 u|$ (hess) as underlying density. It can be seen that for our maximal density of points, the distribution of points is still sparse. Hence, there is not a major difference between uniform, normal and adaptively sampled point clouds.

In order to enforce the boundary conditions we constrain the values of cells at the boundary to zero. However since the boundaries of the Voronoi meshes dont coincide with the original problem boundary (which is not even polygonal), we have to approximate the boundary by defining which cells we consider as part of the boundary. Cells which are part of the boundary or lie outside are not counted in the diagrams of Figures 2 and 3. The remaining points (those which are counted with $i = 1, \dots, N$) are called interior points.

In the case of the uniform and normal samples we have sampling points outside of the unit ball Ω, henceforth we set all those cells to the boundary value zero whose *center* lie outside the unit ball.

In the case of the adaptive samplings, with their densities supported completely inside Ω, we fix all those cells to the boundary value which are not completely supported inside Ω, i.e. who have a vertex lying outside of the unit ball. Since u can be extended by 0 to $\mathbb{R}^4$ still having H^2-regularity and f correspondingly expands by 0, this will not contribute an additional error, provided both $H(u, \mathcal{T})$ and $B(u, \mathcal{T})$ are calculated properly.

As in Theorem 2.5 we then compare the solution of this linear problem to the a priori given solution

of the PDE by means of their $H_{\mathcal{T},\kappa}$ distance (2.2), $\|u_{\mathcal{T}} - \mathcal{R}_{\mathcal{T},\mathrm{BC}} u\|_{H_{\mathcal{T},\kappa}}$. Herein we approximate the discretization operator $\mathcal{R}_{\mathcal{T},\mathrm{BC}}$ by the pointwise evaluation which in our setting only incurs an additional error of order one and therefore does not alter our conclusions. To see this, observe that $\|\nabla^2 u\|_\infty < \infty$ and thus

$$\left| u(x_K) - \fint_{\mathbb{B}_{r_K}(x_K)} u \right| \le \fint_{\mathbb{B}_{r_K}(x_K)} r_K^2 \int_0^1 \left| \nabla^2 u(t x_K (1-t) x) \right| \mathrm{d}t \mathrm{d}x \le r_K^2 \|\nabla^2 u\|_\infty .$$

5.2 Results and discussion

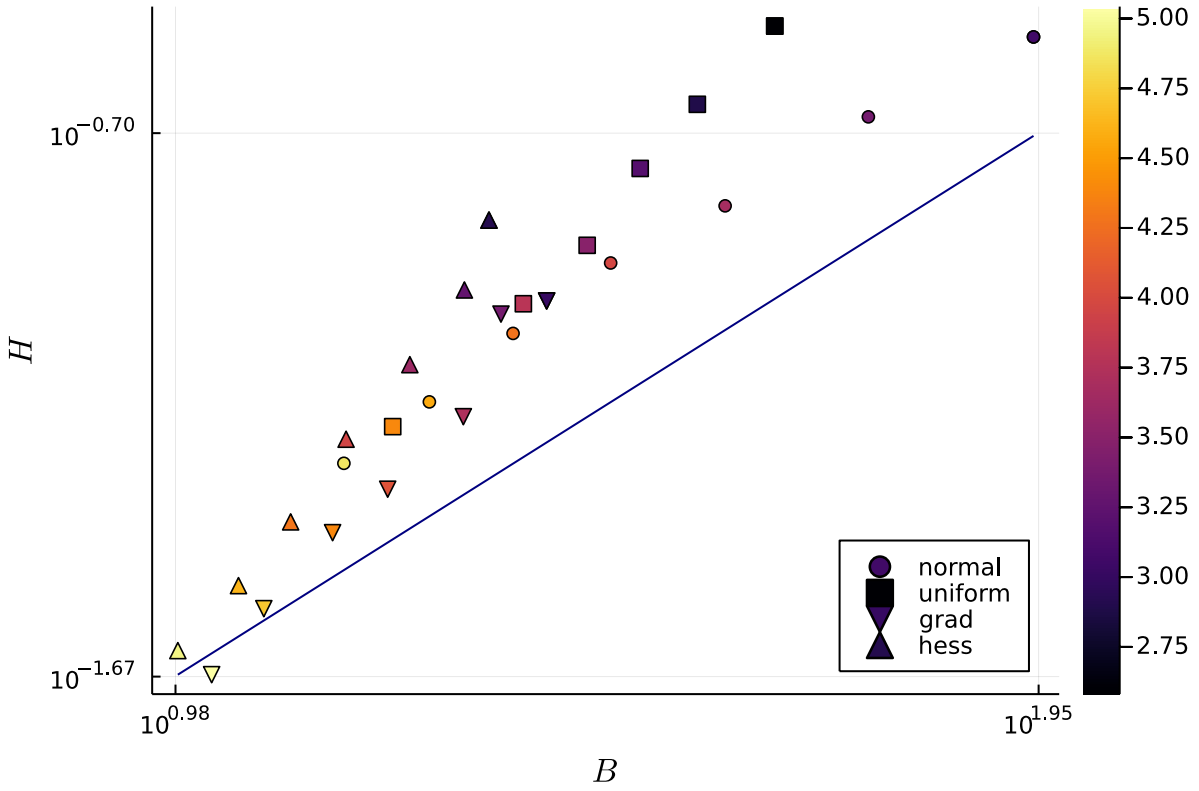

Figure 2: Log-log plot of the error bound $B(u,\mathcal{T})$ of I_2 against true error $H(u,\mathcal{T})$ for the four different sampling methods with varying numbers of interior cells ($\log_{10}$-color). The straight line is for reference indicating linear relations between B and H and thus has slope 1.

In figure 2 we plot both, $H(u,\mathcal{T})$ and $B(u,\mathcal{T})$ for varying numbers of samples $N = 2.000, 4.000, 8.000, \ldots, 128.000$ and the for sampling methods described above. The color represents the number of sampling cells *inside* the domain of u (on a $\log_{10}$ scale) which allows to compare the sampling methods accounting for the differing treatment of boundary values. A linear relation between the $H_{\mathcal{T},\kappa}$ norm and B would display in the log-log plot as a straight line with slope 1, which we plotted for reference. Indeed we see that for the normal sampling the data points exhibit this linear behaviour, indicating that in this setting the error bound is sharp.

Concerning the other three ways of sampling, we clearly observe a super-linear convergence rate of H vs. B. However, we note that the points are distributed sparsely, with $h \approx 0.1$. In such a regime, it is not unusual to have convergence at a rate faster than expected (see e.g. [9] with simulations in 1d). Another effect is that for a comparable amount of points, the value B is strongly reduced for adaptively sampled meshes compared to normal meshes. Together with the sparsety of the meshes this may combine to an initial quadratic behavior of H vs. B for our relatively "large"values of h.

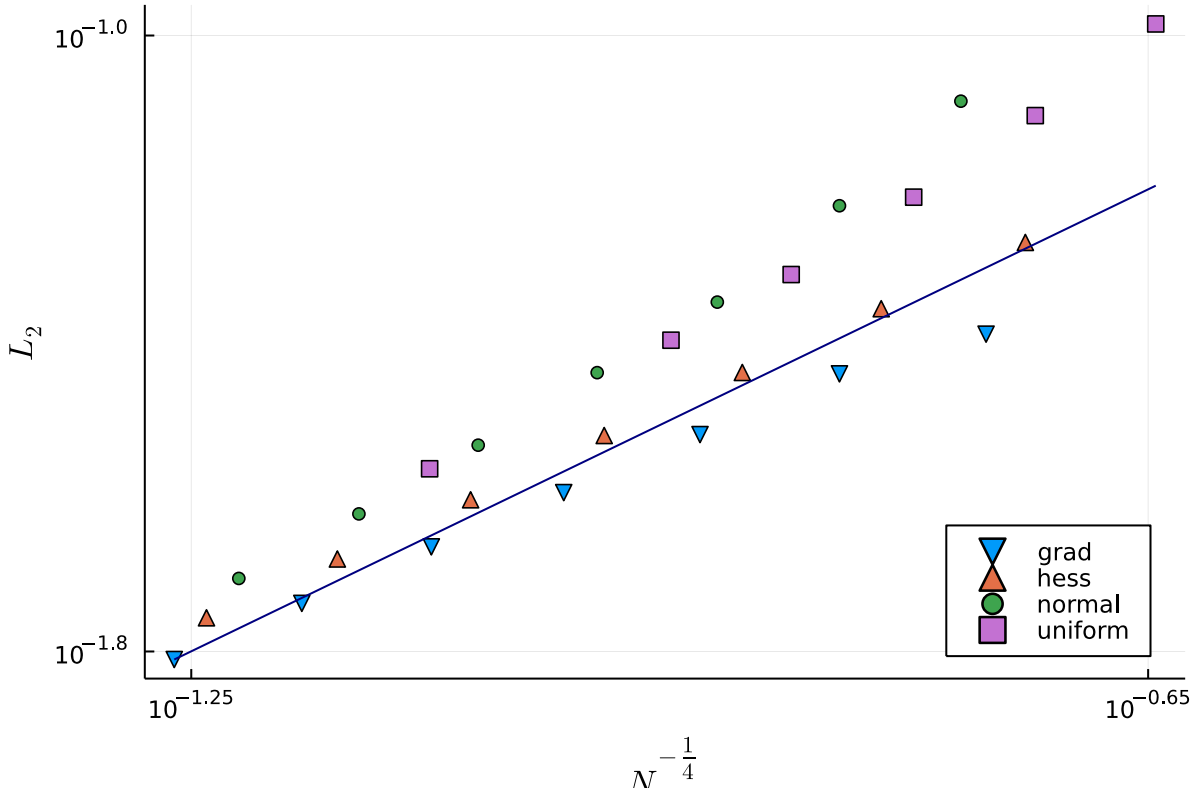

Figure 3: Approximate cell diameter $\left(h \tilde{\propto} N^{-\frac{1}{d}}\right)$ against error in the L_2 norm for varying mesh resolutions and sampling methods.

Another aspect is shown in figure 3 were we plot the approximate average cell diameter (scaling with $h \tilde{\propto} N^{-\frac{1}{d}}$, N the number of interior points) against the L_2 error $\|R^*_{\mathcal{T}} u_{\mathcal{T}} - u\|_{L^2(\Omega)}$ computed by Monte Carlo integration. Note that even though the number of initial samples is the same for all methods, the number of cells classified as interior differs, leading e.g. to the coarser resolution for the uniform sampling. Figure 3 thus illustrates an L_2 convergence of order 1 in the average cell diameter $N^{-\frac{1}{4}}$ for adaptive sampling. For uniform and normal sampling we see a super-linear convergence, but bigger error for a comparable amount of cells. In particular, we see that the adaptive methods lead to better errors then the uniformly and normal distributed cells. This supports our second and third conjecture of this section.

Acknowledgement. This article has been partially founded by the projects C05 and A05 in the German collaborative research center CRC-1114 “Scaling Cascades in Complex Systems”.

A A Poincaré lemma

Lemma A.1. *For every $p \in [1, \infty)$ there exists $C_p > 0$ such that the following holds: Let $0 < r < R$ and $x \in \mathbb{B}_R(0)$ and let Ω be a convex polygonal domain such that such that $\mathbb{B}_r(x) \subset \Omega \subset \mathbb{B}_R(0)$ then for every $u \in W^{1,p}(\mathbb{B}_R(0))$*

$$\|u\|^p_{L^p(\Omega)} \leq C_p \left(R^p \frac{R^{d-1}}{r^{d-1}} \|\nabla u\|^p_{L^p(\Omega)} + \frac{R^d}{r^d} \|u\|^p_{L^p(\mathbb{B}_r(x))} \right), \tag{A.1}$$

and for every u *with* $\int_{\mathbb{B}_r(x)} u = 0$ *it holds*

$$\|u\|^p_{L^p(\Omega)} \le C_p R^p \left(\frac{r}{R}\right)^{1-d} \left(1 + \left(\frac{r}{R}\right)^{p-1}\right) \|\nabla u\|^p_{L^p(\Omega)} \ . \tag{A.2}$$

Remark. In case $p \ge d$ we find that (A.2) holds iff $u(x) = 0$ for some $x \in \mathbb{B}_1(0)$.

Proof. In a first step, we assume $x = 0$ and $R = 1$. The underlying idea of the proof is to compare every $u(y)$, $y \in \mathbb{B}_1(0)\backslash\mathbb{B}_r(0)$ with $u(rx)$. In particular, we obtain for $y \in \mathbb{B}_1(0)\backslash\mathbb{B}_r(0)$ that

$$u(y) = u(ry) + \int_0^1 \nabla u(ry + t(1-r)y) \cdot (1-r)y \,\mathrm{d}t$$

and hence by Jensen's inequality

$$|u(y)|^p \le C \left(\int_0^1 |\nabla u(ry + t(1-r)y)|^p \, (1-r)^p \, |y|^p \, \mathrm{d}t + |u(ry)|^p \right) .$$

We integrate the last expression over $\mathbb{B}_1(0)\backslash\mathbb{B}_r(0)$ and find

$$\begin{aligned}
\int_{\mathbb{B}_1(0)\backslash\mathbb{B}_r(0)} |u(y)|^p \,\mathrm{d}y &\le \int_{S^{d-1}} \int_r^1 C \left(\int_0^1 |\nabla u(rs\nu + t(1-r)s\nu)|^p \, (1-r)^p s^p \,\mathrm{d}t \right) s^{d-1} \mathrm{d}s \mathrm{d}\nu \\
&\quad + \int_{\mathbb{B}_1(0)\backslash\mathbb{B}_r(0)} |u(ry)|^p \,\mathrm{d}y \\
&\le \int_{S^{d-1}} \int_r^1 C \left(\int_{rs}^s |\nabla u(t\nu)|^p \, (1-r)^{p-1} s^{p-1} \,\mathrm{d}t \right) s^{d-1} \mathrm{d}s \\
&\quad + \int_{\mathbb{B}_1(0)\backslash\mathbb{B}_r(0)} |u(ry)|^p \,\mathrm{d}y \\
&\le C \int_r^1 \mathrm{d}s \; s^{d-1} \frac{1}{(rs)^{d-1}} \int_{rs}^s \mathrm{d}t \; t^{d-1} \int_{S^{d-1}} |\nabla u(t\nu)|^p \, (1-r)^{p-1} s^{p-1} \\
&\quad + \int_{\mathbb{B}_1(0)\backslash\mathbb{B}_r(0)} |u(ry)|^p \,\mathrm{d}y \\
&\le C \frac{1}{r^{d-1}} \|\nabla u\|^p_{L^p(\mathbb{B}_1(0))} + \frac{1}{r^d} \|u\|^p_{L^p(\mathbb{B}_r(0))} \ .
\end{aligned}$$

Furthermore, since there holds $\|u\|^p_{L^p(\mathbb{B}_1(0))} \le C \|\nabla u\|^p_{L^p(\mathbb{B}_1(0))}$ for every $u \in W^{1,p}_{(0)}(\mathbb{B}_1(0))$, a scaling argument shows $\|u\|^p_{L^p(\mathbb{B}_r(0))} \le C r^p \|\nabla u\|^p_{L^p(\mathbb{B}_r(0))}$ for every $u \in W^{1,p}_{(0),r}(\mathbb{B}_1(0))$ and hence (A.2). For general $R > 0$ use a scaling argument. □

References

[1] L. Bonaventura and A. Della Rocca. Convergence analysis of a cell centered finite volume diffusion operator on non-orthogonal polyhedral meshes. *arXiv preprint arXiv:1806.09180*, 2018.

[2] A. C. Cavalheiro. Weighted sobolev spaces and degenerate elliptic equations. *Boletim da Sociedade Paranaense de Matemática*, 26(1-2):117–132, 2008.

[3] S. Chanillo and R. Wheeden. Existence and estimates of green's function for degenerate elliptic equations. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 15(2):309–340, 1988.

[4] L. Donati, M. Weber, and B. G. Keller. Markov models from the square root approximation of the fokker–planck equation: calculating the grid-dependent flux. *Journal of Physics: Condensed Matter*, 33(11):115902, 2021.

[5] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.

[6] R. Eymard, T. Gallouët, and R. Herbin. Finite volume approximation of elliptic problems and convergence of an approximate gradient. *Applied Numerical Mathematics*, 37(1):31 – 53, 2001.

[7] R. Eymard, T. Gallouët, and R. Herbin. A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension. *IMA Journal of Numerical Analysis*, 26(2):326, 2006.

[8] T. Gallouët, R. Herbin, and M. H. Vignal. Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions. *SIAM Journal on Numerical Analysis*, 37(6):1935–1972, 2000.

[9] M. Heida, M. Kantner, and A. Stephan. Consistency and convergence for a family of finite volume discretizations of the fokker–planck operator. *accepted by ESAIM M2AN*, 2020.

[10] H. C. Lie, K. Fackeldey, and M. Weber. A square root approximation of transition rates for a markov state model. *SIAM Journal on Matrix Analysis and Applications*, 34:738–756, 2013.

[11] V. Polianskii and F. T. Pokorny. Voronoi graph traversal in high dimensions with applications to topological data analysis and piecewise linear interpolation. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[12] C. Schütte, W. Huisinga, and P. Deuflhard. *Transfer Operator Approach to Conformational Dynamics in Biomolecular Systems*. Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems. Springer, Berlin, Heidelberg, 2001.

[13] A. Sikorski. VoronoiGraph.jl. https://github.com/axsk/VoronoiGraph.jl, 2021.

[14] M. Weber. *A Subspace Approach to Molecular Markov State Models via a New Infinitesimal Generator*. Habilitation Thesis, FU Berlin, 2011.