Numerical Methods for Covariance Functions of Elliptic Problems under Uncertainty

Von der Fakultät für Mathematik und Naturwissenschaften der Carl von Ossietzky Universität Oldenburg zur Erlangung des Grades und Titels eines

Doctor rerum naturalium (Dr. rer. nat.)

angenommene Dissertation

von Herrn Erik Marc Schetzke

geboren am 14.10.1984 in Siegburg.

Gutachter: Prof. Dr. Alexey Chernov
Weitere Gutachter: Prof. Dr. Helmut Harbrecht
Tag der Disputation: 11.09.2020

Zusammenfassung

Die vorliegende Dissertation ist der Entwicklung von Methoden zur Approximation von Kovarianzfunktionen von elliptischen partiellen Differentialgleichungen gewidmet, wenn diese Unsicherheiten unterworfen sind. Wir beginnen mit einer kurzen Fixierung der nötigen Notation, des funktionalanalytischen Rahmens für die Analysis und einer rudimentären Einführung in die Ideen von Finite Element Methoden und des adaptiven Prozesses.

Um Kovarianzfunktionen zu approximieren, entwickeln wir adaptive Fehlerkontrolle für die deterministischen zweiten Momentengleichungen eines elliptischen Modellproblems im eindimensionalen Setting mithilfe von a posteriori Fehlerschätzern, die dazu genutzt werden die adaptive Gitterverfeinerung zu steuern. In diesem Setting untersuchen wir einen a posteriori Fehlerschätzer basierend auf einer L^2-Representation des Residuums und einen hierarchischen a posteriori Fehlerschätzer mittels sogenannter Element-Bubble und Kanten-Bubble Funktionen. Diese werden hergeleitet und von ihnen wird gezeigt, dass sie *zuverlässig* (eng.: reliable) sind. Da im Gegensatz zu Fehlerschätzern des zugrundeliegenden Modellproblems für diese a posteriori Fehlerschätzer für das zweite Momentproblem allerdings nur *schwache Effizienz* nachgewiesen werden kann, liefern diese keine *asymptotisch exakten* Fehlerschätzer. Um dennoch einen Fehlerschätzer ohne dieses Defizit anzugeben, analysieren wir einen Schätzer, der auf einer Mittelungsprozedur beruht und zeigen, dass dieser *asymptotisch exakt* ist, d.h. *zuverlässig* und *effizient* (eng.: reliable and efficient).

Darüberhinaus wird dann ein zweidimensionales Modellproblem analysiert und zugehörige residuale und hierarchische Fehlerschätzer für das zweite Moment der Lösung in diesem Setting entwickelt. Hier gilt wieder lediglich *Zuverlässigkeit* und *schwache Effizienz* für die residualen und hierarchischen Fehlerschätzer, weshalb wir abermals auf einen Fehlerschätzer basierend auf einer Mittelungsprozedur zurückgreifen, von welchem ebenfalls in der vierdimensionalen Situation gezeigt werden kann, dass er *asymptotisch exakt* ist.

Am Ende des Kapitels untersuchen wir dann den Effekt der *schwachen Effizienz* der residualen und hierarchischen Fehlerschätzer auf die Konvergenz des adaptiven Prozesses. Wir bemerken, dass aufgrund der abgeschwächten unteren Abschätzung dieser Schätzer der Fehlerreduktionsfaktor sehr nahe bei 1 liegen mag and als solches eine Art Konvergenzabflachung verursachen kann. Dieses Verhalten ist hiernach ebenfalls bei den numerischen Experimenten für die Modellprobleme für eine gewisse Klasse von Kovarianzfunktionen zu beobachten und wird kurz diskutiert. Es ist auch zu sehen, dass die adaptive Methode durchaus eine bessere Approximation an das zweite Moment liefern kann als die Finite Element Methode auf uniform verfeinerten Gittern. Dies ist in typischen Fällen wie im Falle einer Lösung, die starke Anstiege oder im Allgemeinen abrupte Änderungen im Verhalten aufweist, zu beobachten.

Danach schreiten wir voran zur Approximation von Kovarianzfunktionen mithilfe von Monte Carlo (MC) und multilevel Monte Carlo (MLMC) Methoden. Zuerst präsentieren wir einen theoretischen Standpunkt der Approximation in Tensorprodukten von Hilberträumen und führen den Begriff der Voll-Tensorprodukt- und Dünn-Tensorprodukt-Gitter-Approximation ein. Die präsentierten Ideen haben Ähnlichkeiten mit der sogenannten *Kombinationstechnik* (cf. [34]). Danach wird die Fehleranalysis für die Voll-Tensorprodukt- und Dünn-Tensorprodukt-Approximation durchgeführt. Darüberhinaus zeigen wir asymptotische Schranken für die Kosten- und Speicheranforderungen der zugehörigen MC und MLMC Methoden. Da die analysierten Methoden ein unterschiedliches Verhalten unter unterschiedlichen Rahmenbedingungen zeigen, welche von der Qualität und der Effizienz des benutzten Lösers abhängen, wird ein Vergleich für die unter-

schiedlichen Rahmenbedingungen gegeben und ein kurzes zusammenfassendes Ranking präsentiert, welches eine kurze Entscheidungshilfe bereitstellen soll.

Da beide Methodentypen, das heißt die adpative Approximation von Kovarianzfunktionen mithilfe deterministischer Momentengleichungen und der Approximation mithilfe von Voll-Tensorprodukt- und Dünn-Tensorprodukt-Gitter MLMC Methoden, dazu verwendet werden können, um Kovarianzfunktionen zu approximieren, bietet das letzte Kapitel einen Vergleich beider Methodentypen, der die unterschiedlichen Voraussetzungen, Vorteile und Nachteile, z.B. im Blick auf das Abklingverhalten der Eigenwerte und der Korrelationslänge des zugrundeliegenden Kovarianzoperators des angenommenen stochastischen Modells, beleuchtet.

Abstract

The present dissertation is dedicated to the development of methods for the approximation of covariance functions of elliptic partial differential equations when these are subject to uncertainty. We start out by briefly fixing the necessary notation, the functional analytic framework for the analysis and a rudimentary introduction into Finite Elements and the adaptive process.

In order to approximate covariance functions, we develop adaptive error control for the deterministic second moment equations of an elliptic model problem in the one-dimensional setting by means of a posteriori error estimators that are used to guide the adaptive refinement of the mesh. In this setting we investigate an a posteriori error estimator based on an L^2-representation of the residual and a hierarchical a posteriori error estimator using element bubble and edge bubble functions. These estimators are derived and proven to be reliable. However, since these a posteriori error estimators for the deterministic second moment problem are shown to be only *weakly efficient*, in contrast to the estimators of the underlying model problem, those do not provide asymptotically exact estimators. In order to provide an error estimator without this drawback, an estimator by means of an averaging procedure is analyzed and shown to be asymptotically exact, i.e. reliable *and* efficient.

Moreover, a two-dimensional model problem is then analyzed and corresponding residual and hierarchical error estimators for the second moment of the solution are developed in this setting. Again only reliability holds for the residual and hierarchical estimators and as such we again resort to an error estimator based on averaging, which is proven to be an asymptotically exact estimator in the four dimensional situation.

At the end of the chapter we look into the effect of the *weak efficiency* of the residual and hierarchical error estimators on the convergence of the adaptive process. We note that due to the weakened lower bound of these estimators the error reduction factors may become very close to 1 and as such may create a type of convergence shelf. This behavior is thereafter also observed in the numerical experiments for the model problems for a certain class of covariance functions and briefly discussed. It is also seen that the adaptive method may yield a better approximation to the second moment than the corresponding Finite Element Method on uniformly refined meshes in typical cases, i.e. for instance when the solution features steep gradients or, in general, abrupt changes of behavior.

We then move on with the approximation of covariance functions by means of the Monte Carlo (MC) and multilevel Monte Carlo (MLMC) methods. First we present a theoretical point of view of the approximation in tensor products of Hilbert spaces and introduce the notion of full tensor product and sparse tensor product approximation. The presented ideas have similarities to the so-called *combination technique* (cf. [34]). Error analysis is then conducted with respect to full tensor product and sparse tensor product approximations. Moreover, we show asymptotic bounds for the cost and memory requirements of the corresponding MC and MLMC methods. As the analyzed methods exhibit different asymptotic behavior in different regimes which depend on the quality and efficiency of the used solver, a comparison in the different regimes is given and a short summarizing ranking is presented, which aims to give a quick guideline on which method to use.

As both methods, i.e. the adaptive approximation of covariance functions by means of deterministic moment equations and the approximation by means of full and sparse tensor product MLMC methods, can be used to approximate the covariance function, the last chapter offers a comparison of both methods considering the different requirements,

advantages and disadvantages, e.g. in lieu of the rate of decay of eigenvalues and the correlation length of the underlying covariance operator of the assumed stochastic model.

Introduction

The field of Uncertainty Quantification (UQ) is a relatively young field of research, which has reached a certain degree of maturity in the last two decades, which is due to the rapid development of methods to assess random behavior in systems of all kind. Typically in problems related to natural phenomena we are not equipped with complete information, but have some random fluctuation imbedded into the model itself, e.g. it is impossible to know the temperature at each point in space or the velocity of each particle of air in space for a given volume and/or all times.

With the advent of more and more computing power made available through technological advancements the possibility of numerical analysis of this kind of uncertain systems has grown enormously in recent decades. In general we are interested in finding a solution u to a model that is subject to some uncertain behavior which is modeled by a *forward solution operator* S in combination with some given data, such as boundary conditions or source terms, which themselves might depend on uncertain parameters as well, such that one can codify the problem for some data f via

$$u(\omega) = S(f, \omega)$$

in the sense that S realizes the *forward propagation* of the uncertain parameters to the solution u for a given ω which is an instance of the set of all plausible events. In this thesis we will understand S as the solution operator of an elliptic partial differential equation which has either a random diffusion coefficient or a source term that is subject to uncertain behavior. In this way the uncertainty of the data propagates to the solution u which is now also a *stochastic* quantity and is called a *random variable* if it has values in a field (e.g. $\mathbb{R}$) or a *random field* if it has values in other general spaces, e.g. Banach spaces.

Usually the events ω are modeled by means of a complete probability space $(\Omega, \Sigma, \mathbb{P})$, where Ω is the set of events, Σ the σ-algebra of subsets of Ω and $\mathbb{P}$ a probability measure. As the underlying probability space is usually also infinitely dimensional the question arises of how to accurately assess the validity of a model or its solution. Because of this fact the problems under consideration are intrinsically high dimensional in nature and special methods have to be employed to make approximation and assessment feasible and quantifiable. In order to make sense of the random behavior of a solution of such a system one is typically interested in solving for the so-called *moments* of u, for e.g. the expectation $\mathbb{E}[u]$, the variance $\mathbb{V}[u]$ of u or its (auto)covariance function $\mathbb{C}\mathrm{ov}\,[u, u]$. Moreover, the covariance or covariance function of u is an important tool in the assessment of the random behavior of u as it gives a measure of the variability and mutual dependence with respect to the expectation of u. Since one usually wants to assess the quality of a model with respect to reality or the spread of variability in a certain system, the expectation and variance of the system are key quantities to observe and analyze.

Since usually the exact evaluation of probabilistic models is either impractical or intractable, one has to deal with certain discretizations of the underlying model as well as of its stochastic behavior. This can be done in many different ways, which usually also depend on the problem structure or the quantity of interest one wishes to approximate. In particular, considering stochastic processes one has to deal with discretizations in space or time and additionally with discretizations of the stochastic aspects.

For the approximation of the moments of stochastic solutions to uncertain PDEs different approaches have been applied. The stochastic Galerkin method, which can be thought of as a Galerkin discretization of the underlying space and probability space as

well, leads to a high dimensional deterministic problem (cf. e.g. [25] and the references therein).

Another important tool is the Karhunen-Loéve expansion, which by truncation after a reasonable number of terms of its series allows to approximate a second order process of a random field. This expansion can be used in conjunction with a large class of methods that deal with the quantification of random behavior of solutions to uncertain PDEs, the so-called Monte Carlo-type (MC-type) methods. Here we mention only a few references, e.g. [5, 20, 7, 9] and the references therein. Monte Carlo (MC) and multilevel Monte Carlo (MLMC) methods are extensively used in UQ. This is due to the simplicity, generality and accessability of Monte Carlo methods for a vast number of problems as one typically only has to make sense of the sampling procedure in the context in question to be in business. Another great flexibility of these methods is that one has complete freedom of the underlying solver for the PDE which does not change the overall algorithm and is thus relatively easy implemented. The great flexibility and quick applicability, however, is counterweighted by the necessity of having to solve the underlying uncertain PDE for a certain number M of *samples* to achieve a certain error tolerance ε. In the case of MC methods this is usually a very expensive procedure which is the reason for the development of the MLMC method. The MLMC method balances the work needed against the required accuracy by sampling less on finer discretization levels and introducing the notion of level corrections. Moreover, for the MLMC method on each level of discretization ℓ a different amount of samples M_ℓ is chosen to achieve the aforementioned balancing.

Under certain assumptions on the underlying operator of the elliptic PDE another approach to find the moments of u leads to a formulation where the moments of u fulfill a *deterministic* variational equation, the so-called *deterministic moment equations* (DME) (cf. e.g. [55]). In this thesis the concern is the deterministic second moment equation for e.g. the model problem

$$-\nabla \cdot (\kappa(x,\omega)\nabla u(x,\omega)) = f(x,\omega), \qquad \text{in } D,$$

where $D \subset \mathbb{R}^d$, $d = 1, 2$ is an open bounded polygonal domain and the diffusion coefficient κ as well as the source term f may depend randomly on ω. In general, the k-fold tensorization of this equation and subsequently integrating with respect to the stochastic variable, i.e. taking expectation of both sides of the tensorized equation, leads to a fully deterministic variational formulation after we mulitply by a deterministic test function on D^k and perform the usual integration by parts to arrive at a weak formulation. The advantage of this formulation lies in the fact that one does *not* have to know the exact probability densities of f and of the probability measure $\mathbb{P}$ to recover the k-th order statistics of u.

As the solution of the second moment equation may feature singularities (cf. [43]) on the diagonal of the domain and/or on the boundary at certain orders of weak derivatives, it is desireable to have a solution procedure which adapts the underlying mesh, i.e. the spatial discretization, in the best way possible to the given data. As the computational domain for the k-th deterministic moment equation is given by D^k and a naïve uniform FEM thus has a number of degrees of freedom which is proportional to N_L^k on discretization level L, when all elements are refined uniformly from one discretization level to the next. For this reason we have chosen to develop adaptive Finite Element Methods (AFEM) with a posteriori error estimators to guide the discretization procedure and thereby speed up the solution process and profit from the efficiency of the so generated meshes. Another way of accelerating the convergence is to employ higher order methods in this setting (cf. [43]) or even more sophisticated ideas in the line of spectral hyperbolic cross approximation as in [18].

In contrast to this approach the class of Monte Carlo methods as mentioned above is another important tool in the field for the approximation of higher order moments of a random field u (cf. e.g. [9]). Here, in order to estimate the moments of u, *samples* have to be drawn with respect to a number of events $\omega^i \in \Omega, i = 1, \ldots, M$, which only give *sample solutions* $u^i \equiv u(\cdot, \omega^i)$ which are averaged to quantify the quantity of interest. This usually involves the discretization of the underlying spatial geometry and as such one has not only to deal with the usual discretization error but also with the error that is committed by means of the sampling procedure itself, i.e. the *sampling error*. The popularity of these methods has soared in accordance with the available computational power and has sparked improvements and specialization such as the multilevel Monte Carlo method. In order to approximate second order moments by the MC and MLMC methods we choose to optimize the dimension of the underlying tensor product discretization spaces by virtue of sparse tensor product approximation techniques which facilitate the solution procedure.

This thesis is structured as follows. In Chapter 1 we fix the necessary notation, vocabulary and give a necessary frame that is used in the subsequent chapters, such as function spaces, Hilbert space valued random fields and a brief outline of the Finite Element Method and the corresponding self-adaptive process guided by a posteriori error estimators.

In Chapter 2 we are then concerned with deriving adaptive methods for the deterministic second moment equation of the model problems in one and two space dimensions, which lead to adaptive methods in two and four dimensions, respectively. In particular, we develop residual and hierarchical *a posteriori* error estimators and show their *reliability*. *Weak efficiency* of these error estimators is also defined and established. Due to that fact these estimators are not *asymptotically exact* and in order to provide an asymptotically exact error estimator we adapt an error indicator based on a popular averaging procedure to our situation. At the end of the chapter we shall then have a look at numerical experiments which validate the theoretical findings and exhibit the strengths and weaknesses of the methods.

In Chapter 3 we take another perspective on the second moment problem to approximate covariance functions by help of Monte Carlo methods. Here we are concerned with the approximation of covariance functions in tensor products of Hilbert spaces. After investigating how to approximate elementary algebraic tensor products by means of full and sparse tensor products we analyze full and sparse tensor product Monte Carlo (MC) and multilevel Monte Carlo (MLMC) methods for the approximation of covariance functions in the context of tensor products of Hilbert spaces. We show convergence and give asymptotic bounds for the cost and memory requirements of these methods. The chapter is then concluded by numerical experiments verifying the theoretical analysis.

Since the methods in Chapter 2 and Chapter 3 are competing in the approximation of covariance functions, it is natural to ask the question whether one should prefer one over the other and if so in what situation that may be advantageous. For this purpose Chapter 4 conducts a comparison in this direction of the two approaches which we have chosen for approximation.

Contents

Chapter 1

Preliminaries

Before we go at the heart of the matter we shall in this chapter prepare the necessary theoretical background that enables us to present the research conducted in a concise and efficient manner. In order to do so, we introduce the necessary function spaces, functional analytic framework and fix most of the commonly used notation of this thesis.

1.1 Function spaces

In this section we briefly give definitions and properties of the most used spaces and objects used in this work.

1.1.1 Sobolev spaces

A more in depth introduction of the spaces presented here can be found in [1]. We also refer to [29] and [13] for further details.

Let $D \subset \mathbb{R}^d$ with $d = 1, 2, 3, 4$ an open and bounded domain with boundary $\Gamma = \partial D$.

Denote by $C(D)$ the space of continous functions on D. The partial derivative associated with a multi-index $\mathbf{s} \in \mathbb{N}_0^d$ is denoted by

$$\partial^{\mathbf{s}} := \frac{\partial^{|\mathbf{s}|}}{\partial_{x_1}^{s_1} \cdots \partial_{x_d}^{s_d}},$$

where $|\mathbf{s}| = \sum_i |s_i|$ is the order of the derivative. Also we introduce a shorthand notation for the sum of partial derivatives of the same order k by

$$D^k := \sum_{|s|=k} \frac{\partial^{|\mathbf{s}|}}{\partial_{x_1}^{s_1} \cdots \partial_{x_d}^{s_d}}.$$

The space $C^k(D)$ of k-times continously differentiable functions on D equipped with the usual supremum norm, i.e.

$$\|v\|_{C^k(D)} := \sum_{|\mathbf{s}|\leq k} \sup_{x\in D} |\partial^{\mathbf{s}} v(x)|,$$

is defined in the usual manner via

$$C^k(D) := \{v \in C(D) : \partial^{\mathbf{s}} v \in C(D),\ |\mathbf{s}| \leq k\}.$$

Moreover, instead of $C^0(D)$ we shall usually write $C(D)$.

For any $p \in [1, \infty]$ we denote by $L^p(D)$ the space of p-Lebesgue integrable functions on a domain D equipped with the norm

$$\|f\|_{L^p(D)} = \begin{cases} \left(\int_D |f(x)|^p \, \mathrm{d}x \right)^{1/p}, & p < \infty, \\ \operatorname*{ess\,sup}_{x \in D} |f(x)|, & p = \infty. \end{cases}$$

Functions are considered equal as long as they differ only on a set of measure zero and also the essential supremum has to be understood in this way.

The weak derivative of order k associated with a multiindex $\mathbf{s}$ with $|\mathbf{s}| = k$ of a function $v \in L^p(D)$ is defined as a function $w \in L^p(D)$, such that for all infinitely differentiable functions φ with compact support in D, i.e. $\varphi \in C_0^\infty(D)$, there holds

$$\int_D \partial^{\mathbf{s}} v \, \varphi \, \mathrm{d}x = (-1)^{|\mathbf{s}|} \int_D w \, \partial^{\mathbf{s}} \varphi \, \mathrm{d}x,$$

or in symbols $\partial^{(\mathbf{s})} v = w$, if we want to emphasize that w is only a *weak derivative* of v. Then we say w is a weak derivative of v of order $k = |\mathbf{s}|$.

As usual, we denote by $W^{m,p}(D)$ the space of functions in $L^p(D)$, such that all weak derivatives of order up to $m \in \mathbb{N}_0$ are again contained in $L^p(D)$, namely

$$W^{m,p}(D) := \{v \in L^p(D) : \partial^{\mathbf{s}} v \in L^p(D), |\mathbf{s}| \leq m\},$$

equipped with the norm

$$\|f\|_{W^{m,p}(D)} = \begin{cases} \left(\sum_{k=1}^{m} \int_D |D^k f(x)|^p \, \mathrm{d}x \right)^{1/p}, & p < \infty, \\ \max\limits_{|\mathbf{s}| \leq m} \operatorname*{ess\,sup}_{x \in D} |\partial^{\mathbf{s}} f(x)|, & p = \infty. \end{cases}$$

We shall also frequently make use of the corresponding seminorms given by the follwing expressions

$$|f|_{W^{m,p}(D)} = \begin{cases} \left(\int_D |D^m f(x)|^p \, \mathrm{d}x \right)^{1/p}, & p < \infty, \\ \max\limits_{|\mathbf{s}| = m} \operatorname*{ess\,sup}_{x \in D} |\partial^{\mathbf{s}} f(x)|, & p = \infty. \end{cases}$$

Moreover, we adopt the convention to denote $H^m(D) := W^{m,2}(D)$ and note that for $m \geq 0$ the spaces $H^m(D)$ are Hilbert spaces with associated inner product

$$\langle u, v \rangle_{H^m(D)} = \sum_{k=0}^{m} \int_D D^k u \, D^k v \, \mathrm{d}x.$$

The dual space $W^{-s,p}(D)$ of $W_0^{s,q}$, where $1 \leq p, q \leq \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$, with the convention that $\frac{1}{\infty} = 0$, we define for any $s > 0$

$$\|u\|_{W^{-s,p}(D)} := \sup_{v \in W_0^{s,q}(D) : v \neq 0} \frac{\langle u, v \rangle_D}{\|v\|_{W_0^{s,q}(D)}},$$

where we have denoted the duality pairing between $W^{-s,p}(D)$ and $W_0^{s,q}(D)$ by

$$\langle u, v \rangle_D = \int_D uv \, \mathrm{d}x$$

and

$$W_0^{s,q}(D) := \{v \in L^q(D) : \partial^{\mathbf{s}} v \in L^q(D), v|_\Gamma = 0, |\mathbf{s}| \leq s\}.$$

1.1.2 Tensor product spaces

More details on the presentation of this section can be found in [47]. For further details and topics we also refer to [40] for approximation in tensor product spaces as well as [35] for more details on tensor calculus in the context of numerical analysis. For the functional analysis aspects we refer to [60] and [58].

Let H and $\tilde{H}$ denote two separable Hilbert spaces over $\mathbb{R}$ with their associated inner products $\langle\cdot,\cdot\rangle_H$ and $\langle\cdot,\cdot\rangle_{\tilde{H}}$, respectively. We note that the space of algebraic tensor products $H \otimes \tilde{H}$ of all formal finite sums of the form

$$\sum_{k=1}^{M} h_k \otimes \tilde{h}_k, \qquad \forall h_k \in H, \forall \tilde{h}_k \in \tilde{H},$$

endowed with the usual algebraic operations is a vector space. Defining a norm on this vector space there are several options which are briefly presented in the following. The completion of the tensor product with respect to any of these norms gives rise to different spaces (cf. [47, Chapter 2–4]). For our purposes it is sufficient to present the three most prominent choices. These are the projective tensor product, the injective tensor product and the Hilbert tensor product space, which is given through the canonical tensor product Hilbert space norm as we will see later on.

The *projective* tensor product space $H \otimes_\pi \tilde{H}$ is defined as the closure of the algebraic tensor product space $H \otimes \tilde{H}$ where the norm is defined by

$$\|U\|_{H\otimes_\pi \tilde{H}} := \inf\left\{\sum_{k=1}^{M} \|u_k\|_H \|\tilde{u}_k\|_{\tilde{H}} : U = \sum_{k=1}^{M} u_k \otimes \tilde{u}_k\right\}.$$

As a second choice the *injective* tensor product space $H \otimes_\varepsilon \tilde{H}$ is obtained when taking the closure of the algebraic tensor product space with respect to the norm

$$\|U\|_{H\otimes_\varepsilon \tilde{H}} := \sup\left\{\left|\sum_{k=1}^{M} \phi(u_k)\psi(\tilde{u}_k)\right| : \phi \in H', \psi \in \tilde{H}', \|\phi\|_{H'} = \|\psi\|_{\tilde{H}'} = 1\right\},$$

where H' and $\tilde{H}'$ denote the dual spaces of H and $\tilde{H}$, respectively. We will mostly work with the following choice of completion of $H \otimes \tilde{H}$. We define an inner product on $H \otimes \tilde{H}$ via

$$(U, V)_{H\otimes\tilde{H}} := \sum_{k'=1}^{M'} \sum_{k=1}^{M} \langle u_k, \tilde{u}_{k'}\rangle_H \langle v_k, \tilde{v}_{k'}\rangle_{\tilde{H}} \tag{1.1}$$

and then consider the closure of $H \otimes \tilde{H}$ by the norm induced by the aforementioned inner product and denote it by $H \otimes_2 \tilde{H}$. As we will mostly be concerned with tensor products of Hilbert spaces, we shall also frequently abuse notation and write $H \otimes \tilde{H}$ for $H \otimes_2 \tilde{H}$.

For ease of notation we denote the k-fold tensor product of a Hilbert space H by

$$H^{(k)} := \bigotimes_{i=1}^{k} H$$

and understand it as the completion with respect to the canonical Hilbert tensor product norm $\|\cdot\|_{H^{(k)}}$, which is induced by the inner product on $H^{(k)}$. For any $U = u_1 \otimes \cdots \otimes u_k$ the inner product on $H^{(k)}$ is defined by

$$\|U\|^2_{H^{(k)}} := \langle U, U\rangle_{H^{(k)}} := \langle u_1, u_1\rangle_H \cdots \langle u_k, u_k\rangle_H.$$

In particular, we note that by definition there holds the so-called *crossnorm property*

$$\|U\|_{H^{(k)}} = \|u_1 \otimes \cdots \otimes u_k\|_{H^{(k)}} = \|u_1\|_H \cdots \|u_k\|_H.$$

We will finish this section with some theoretical remarks about tensor product spaces. For the three constructions of tensor product spaces we have the following chain of embeddings, e.g. cf. [47],

$$H \otimes_\pi \tilde{H} \hookrightarrow H \otimes_2 \tilde{H} \hookrightarrow H \otimes_\varepsilon \tilde{H},$$

where the constants of the associated embeddings are equal to one (are of unit norm). Moreover, when dealing with operators on tensor product spaces, the following fact is useful. Letting $S \in \mathcal{L}(H, X)$ and $T \in \mathcal{L}(\tilde{H}, \tilde{X})$ and defining

$$(S \otimes T)(h \otimes \tilde{h}) := (Sh) \otimes (T\tilde{h}), \quad h \in H, \tilde{h} \in \tilde{H} \tag{1.2}$$

yields a well-defined linear operator $S \otimes T : H \otimes \tilde{H} \to X \otimes \tilde{X}$, when one extends this definition by linearity to all elements of $H \otimes \tilde{H}$. The space $\mathcal{L}(H, X)$ denotes the space of all continuous and bounded linear operators from H to X, which endowed with the operator norm

$$\|S\| := \sup_{h \in H} \frac{\|Sh\|_X}{\|h\|_H}$$

is a Banach space. Furthermore, there exists a unique extension to a continuous linear operator $S \otimes_2 T : H \otimes \tilde{H} \to X \otimes \tilde{X}$ and its operator norm is given by

$$\|S \otimes T\|_{\mathcal{L}(H \otimes \tilde{H}, X \otimes \tilde{X})} = \|S\|_{\mathcal{L}(H,X)} \|T\|_{\mathcal{L}(\tilde{H}, \tilde{X})}.$$

Moreover, let us note that the projective tensor product in general does not respect subspaces, i.e. if W is a subspace of H then the completion of $W \otimes_\pi \tilde{H}$ need not be a subspace of the completion $H \otimes_\pi \tilde{H}$ (cf. [47, Chapter 2, p.21]). Also refer to [47, Proposition 2.11,Corollary 2.12] for the requirements when $W \otimes_\pi Z$ with $W \subset H$ and $Z \subset \tilde{H}$ is a subspace of $H \otimes_\pi \tilde{H}$. Although, interestingly enough, the projective tensor product space does not respect subspaces, it does respect quotients.

In contrast to that we see (cf. [47, Chapter 3]) that the injective tensor product respects subspaces in the sense that if W is a closed subspace of H and Z is a closed subspace of $\tilde{H}$ then the completion of $W \otimes_\varepsilon Z$ is a closed subspace in the completion of $H \otimes_\varepsilon \tilde{H}$. But now the injective tensor product does not respect quotients, i.e. if W is a quotient space of H then the completion of $W \otimes_\varepsilon \tilde{H}$ does not need to be a quotient of the completion of $H \otimes_\varepsilon \tilde{H}$.

Let us introduce tensor products of Sobolev spaces which will be used extensively throughout this thesis. To this end let us denote by

$$H^{\mathbf{s}}(D^k) := H^{s_1}(D) \otimes H^{s_2}(D) \otimes \cdots \otimes H^{s_k}(D)$$

the anisotropic Sobolev space of order $|\mathbf{s}|$, where $\mathbf{s} \in \mathbb{N}_0^k$ is a multiindex. Since the spaces $H^s(D)$ are also Hilbert spaces, their canonical inner product for any $u, v \in H^{\mathbf{s}}(D^k)$ with $u = u_1 \otimes \cdots \otimes u_k$ and $v = v_1 \otimes \cdots \otimes v_k$ is given by

$$\begin{aligned}
\langle u, v \rangle_{H^{\mathbf{s}}(D^k)} =& \langle u_1, v_1 \rangle_{H^{s_1}(D)} \langle u_2, v_2 \rangle_{H^{s_2}(D)} \cdots \langle u_k, v_k \rangle_{H^{s_k}(D)} \\
&= \left(\sum_{\ell_1=0}^{s_1} \int_D D_{x_1}^{\ell_1} u_1 D_{x_1}^{\ell_1} v_1 \, \mathrm{d}x_1 \right) \cdots \left(\sum_{\ell_k=0}^{s_k} \int_D D_{x_k}^{\ell_k} u_k D_{x_k}^{\ell_k} v_k \, \mathrm{d}x_k \right) \\
&= \sum_{\mathbf{0} \le \mathbf{t} \le \mathbf{s}} \int_{D^k} D^{\mathbf{t}} u D^{\mathbf{t}} v \, \mathrm{d}x,
\end{aligned}$$

where $D^{\mathbf{t}} = D^{t_1} \otimes \cdots \otimes D^{t_k}$, $\mathbf{t}, \mathbf{s} \in \mathbb{N}_0^k$ and with the convention that

$$\mathbf{t} \leq \mathbf{s} :\Leftrightarrow t_i \leq s_i, \quad \forall i = 1, \ldots, k.$$

The corresponding seminorm of order $|\mathbf{s}|$ for any $u \in H^{\mathbf{s}}(D^k)$ can be written as follows

$$|u|^2_{H^{\mathbf{s}}(D^k)} := \|D^{\mathbf{s}} u\|^2_{L^2(D^k)}$$

and finally the norm in $H^{\mathbf{s}}(D^k)$ is then given as

$$\|u\|_{H^{\mathbf{s}}(D^k)} := \left(\sum_{\mathbf{0} \leq \mathbf{t} \leq \mathbf{s}} |u|^2_{H^{\mathbf{t}}(D^k)} \right)^{1/2}.$$

In particular, since we will be dealing with second moment problems in order to approximate covariance functions, we encounter the space $H^{1,1}(D \times D)$, where D is a one or two dimensional bounded domain. Moreover, the spaces $H^{t,s}(D \times D)$ for $t \neq s$ will also be of importance. For example consider the anisotropic spaces $H^{1,0}(D \times D) \simeq H^1(D) \otimes L^2(D)$ and $H^{0,1}(D \times D) \simeq L^2(D) \otimes H^1(D)$ with the norms

$$\begin{aligned} \|u\|^2_{H^{1,0}(D\times D)} &:= \|u\|^2_{L^2(D\times D)} + \|(\nabla \otimes \mathrm{id})u\|^2_{L^2(D\times D)}, \\ \|u\|^2_{H^{0,1}(D\times D)} &:= \|u\|^2_{L^2(D\times D)} + \|(\mathrm{id} \otimes \nabla)u\|^2_{L^2(D\times D)}. \end{aligned}$$

Here the map id is the identity operator and the maps $(\nabla \otimes \mathrm{id})$ and $(\mathrm{id} \otimes \nabla)$ can be interpreted using (1.2).

1.2 Random fields and their statistics

Here we keep the presentation to a minimum and refer the reader to the extensive literature for more details. For topics on probability we refer the reader to the book [38] which gives a thorough and comprehensive introduction. For more information about Bochner spaces and moments of random fields we refer the reader to [52]. For applications in this context we refer to [9, 30, 50, 49, 53, 44, 45] to name only a very few.

Let $(\Omega, \Sigma, \mathbb{P})$ a complete probability space, where $\mathbb{P}$ is a probability measure on a measurable space (Ω, Σ) with underlying σ-algebra Σ. We note the following definitions.

Definition 1.1. *A strongly $\mathbb{P}$-measurable function X defined on a probability space $(\Omega, \Sigma, \mathbb{P})$ with values in H, i.e. $X : \Omega \to H$, is called an H-valued random field. If $H = \mathbb{R}$, then X is often called a random variable.*

Definition 1.2. *The distribution of an H-valued random field X is the Borel probability measure μ_X on H defined by*

$$\mu_X(B) := \mathbb{P}\{X \in B\}, \quad B \in \mathcal{B}(H),$$

where $\mathcal{B}(H)$ denotes the Borel σ-algebra of H which is induced by the set of all open sets in H. Random fields that have the same distribution are said to be identically distributed.

In order to make sense of random fields we encounter during this thesis we introduce the notion of *Bochner spaces*.

Definition 1.3 (Bochner spaces)**.** *Let $k \in \mathbb{N} \cup \{\infty\}$ and H a separable Hilbert space. Then we define the* Bochner *space $L^k(\Omega, \mathbb{P}; H)$ of Hilbert space valued mappings $u : \Omega \to H$ for which the following norm is finite*

$$\|u\|_{L^k(\Omega,H)} := \left(\int_\Omega \|u(\omega)\|_H^k \, \mathrm{d}\mathbb{P}(\omega) \right)^{1/k},$$

with the obvious modification if $k = \infty$.

We shall frequently suppress the probability measure from the notation and write $L^k(\Omega; H)$ instead of $L^k(\Omega, \mathbb{P}; H)$. Moreover, if $f \in L^k(\Omega; H)$ then f can be approximated by simple functions, namely by H-valued step functions. Similar to the Lebesgue spaces L^p we have the following

Theorem 1.4 (Bochner)**.** *A function f belongs to $L^k(\Omega; H)$ if and only if there exists a sequence of H-valued step functions $(f_j)_{j\in\mathbb{N}}$ with $f_j \to f$ $\mathbb{P}$-a.e. on Ω and for $j \to \infty$ there holds*

$$\int_\Omega \|f_j - f\|_H^k \, \mathrm{d}\mathbb{P}(\omega) \to 0.$$

Definition 1.5 (Statistical moments)**.** *For $k \in \mathbb{N}$, we introduce the k-th order statistical moment for $u \in L^k(\Omega; H)$ as defined by*

$$\mathcal{M}^k u := \int_\Omega \bigotimes_{i=1}^k u(\omega) \, \mathrm{d}\mathbb{P}(\omega) = \int_\Omega u(\omega)^{(k)} \, \mathrm{d}\mathbb{P}(\omega) \in H^{(k)},$$

where we abbreviate $u(\omega)^{(k)} := \bigotimes_{i=1}^k u(\omega)$, where no notational confusion should occur with the otherwise usual notation of derivatives in the one-dimensional setting.

In the following we abbreviate by $\mathbb{E}[u] := \mathcal{M}^1 u$ the first moment (the expectation) and the second moment $\mathrm{Cor}_u := \mathcal{M}^2 u$ (the two-point correlation). By the previous definition it is easy to see that with $H^{(k)} = H^{\mathbf{s}}(D^k)$ there holds the following representation for the k-th moment, i.e. the k-point correlation function, of u

$$\mathcal{M}^k u(x_1, \ldots, x_k) := \int_\Omega u(x_1, \omega) \otimes \cdots \otimes u(x_k, \omega) \, \mathrm{d}\mathbb{P}(\omega)$$

for any $(x_1, \ldots, x_k) \in D^k$. Since for the covariance function there holds the representation

$$\begin{aligned} \mathbb{C}\mathrm{ov}\,[u, u]\,(x, x') =& \mathbb{E}[(u(x, \omega) - \mathbb{E}[u(x, \omega)]) \otimes (u(x', \omega) - \mathbb{E}[u(x', \omega)])] \\ =& \mathcal{M}^2 u(x, x') - \mathbb{E}[u](x) \otimes \mathbb{E}[u](x'), \end{aligned}$$

and one is usually able to assume without loss of generality that $\mathbb{E}[u] = 0$, approximating the second moment offers a possibility to approximate the covariance function.

1.3 Operator equations and abstract Galerkin methods

In this section we briefly present some selected results from functional analysis, introduce the notion of operator equations and describe the Galerkin method in an abstract setting, e.g. cf. [51].

Let H be a Hilbert space with inner product $(\cdot\,,\cdot)_H$ endowed with the induced norm $\|\cdot\|_H = \sqrt{(\cdot\,,\cdot)_H}$ and let H' denote the dual space of H. Moreover, by ${}_{H'}\langle \cdot\,, \cdot \rangle_H$ we denote the duality pairing between H and H'. Note that subscripts of spaces in duality

pairings will usually be omitted if no misunderstanding can occur. Then for all $f \in H'$ there holds

$$\|f\|_{H'} := \sup_{0 \neq v \in H} \frac{|{}_{H'}\langle f\,,\, v\rangle_H|}{\|v\|_H}, \quad \forall f \in H'.$$

Let $A : H \to H'$ be a bounded linear and self adjoint operator with

$$\|Av\|_{H'} \leq c_2^A \|v\|_H, \quad \forall v \in H.$$

First we consider the *deterministic* operator equation

$$Au = f, \qquad \text{for any given } f \in H', \tag{1.3}$$

that we would like to solve for $u \in H$. Alternatively, we may consider the equivalent variational problem: Find $u \in H$, such that

$${}_{H'}\langle Au\,,\, v\rangle_H = {}_{H'}\langle f\,,\, v\rangle_H \qquad \forall v \in H. \tag{1.4}$$

It can be shown that the solution u of the operator equation also satisfies the variational problem and vice versa. In particular, the operator A induces a bilinear form via duality pairing

$$a(u\,,\, v) := {}_{H'}\langle Au\,,\, v\rangle_H \qquad \forall u, v \in H. \tag{1.5}$$

Conversely, any bilinear form (1.5) defines an operator $A : H \to H'$. In order to ensure unique solvability of the operator equation (1.3) and the variational formulation (1.4), we introduce another property of the operator A and its associated bilinear form $a(\cdot\,,\,\cdot)$.

Definition 1.6 (H-ellipticity)**.** *An operator $A : H \to H'$ is called H-elliptic, if there holds*

$$\langle Av\,,\, v\rangle \geq c_{ell}^A \|v\|_H^2, \quad \forall v \in H$$

with the ellipticity constant $c_{ell}^A > 0$.

For later and further reference let us state the Lax-Milgram Lemma.

Theorem 1.7 (Lax-Milgram Lemma)**.** *Let $A : H \to H'$ be bounded and H-elliptic. Then, for any $f \in H'$, there exists a unique solution $u \in H$ of the operator equation (1.3) with*

$$\|u\|_H \leq \frac{1}{c_{ell}^A} \|f\|_{H'}.$$

In the setting of tensor product spaces we consider the operator $A^{(k)} := \bigotimes_{i=1}^k A$ from the space $H^{(k)} := \bigotimes_{i=1}^k H$ to $(H^{(k)})' := \bigotimes_{i=1}^k H'$. We can formulate a tensorized version of (1.3) via

$$A^{(k)} U = F \qquad \text{in } (H^{(k)})', \tag{1.6}$$

where we have set $U := \bigotimes_{i=1}^k u$ and $F := \bigotimes_{i=1}^k f$.

Since the tensor product of Hilbert spaces, when taking the completion with respect to the canonical Hilbert norm, is again a Hilbert space, the Lax-Milgram Lemma follows immediately for Hilbert tensor product spaces by the properties of the induced norm, i.e.

$$\|U\|_{H^{(k)}} \leq \left(\frac{1}{c_{\text{ell}}^A}\right)^k \|F\|_{(H^{(k)})'}.$$

We proceed by giving a description of the Galerkin method in an abstract setting. From now on we assume that the operator A is bounded and H-elliptic. Considering the

variational problem (1.4) there exists a unique solution $u \in H$ of the variational problem by virtue of the Lax-Milgram Lemma. For $M \in \mathbb{N}$ let

$$H_M := \operatorname{span}\{\phi_k\}_{k=1}^M$$

be a sequence of conforming trial spaces, i.e. $H_M \subset H$ for some $\phi_k \in H$ and for all M. By means of the spaces H_M we have constructed a *discretisation*, i.e. a finite dimensional approximation, of the (possibly) infinite dimensional variational problem. The approximate solution

$$u_M := \sum_{k=1}^M \mathbf{u}_k \phi_k \in H_M,$$

where the components of $\mathbf{u} \in \mathbb{R}^M$ denote the coefficients of the basis functions ϕ_k, is then defined as the solution of the variational problem

$$\langle Au_M \, , \, v_M \rangle = \langle f \, , \, v_M \rangle \qquad \forall v_M \in H_M. \tag{1.7}$$

The functions ϕ_k, which yield a representation for u_M are called *trial functions* and the functions $v_M := \sum_k \mathbf{v}_k \psi_k$ and their constituents ψ_k are called *test functions*. Note that in the previous presentation the same trial and test functions have been used, i.e. $\phi_k = \psi_k$. It remains to investigate the unique solvability of (1.7), the stability of solutions u_M and the convergence to the unique solution $u \in H$ as $M \to \infty$. Since $H_M \subset H$, we can select $v = v_M \in H_M$ in (1.4). Subtracting (1.7) from the continuous variational formulation (1.4) we obtain the so-called Galerkin orthogonality

$$\langle A(u - u_M) \, , \, v_M \rangle = 0, \qquad \forall v_M \in H_M. \tag{1.8}$$

Upon inserting the approximate solution into the Galerkin formulation (1.7) we obtain the following finite dimensional problem due to the linearity of A

$$\sum_{k=1}^M \mathbf{u}_k \langle A\phi_k \, , \, \phi_\ell \rangle = \langle f \, , \, \phi_\ell \rangle \qquad \text{for } k = 1, \ldots, M.$$

Redefining the left- and right-hand sides more compactly as a matrix A_M and a load vector $\underline{\mathbf{f}}$ we write

$$A_M[\ell, k] := \langle A\phi_k \, , \, \phi_\ell \rangle \, , \quad f_\ell := \langle f \, , \, \phi_\ell \rangle$$

for $k, \ell = 1, \ldots, M$. Thus, we obtain the following system of linear equations

$$A_M \underline{\mathbf{u}} = \underline{\mathbf{f}},$$

where we have to find the coefficient vector $\underline{\mathbf{u}} \in \mathbb{R}^M$. Note that we have a one-to-one correspondence of a vector $\underline{\mathbf{v}} \in \mathbb{R}^M$ with the function

$$v_M = \sum_{k=1}^M \mathbf{v}_k \phi_k \in H_M.$$

We also note for arbitrary vectors $\underline{\mathbf{u}}, \underline{\mathbf{v}} \in \mathbb{R}^M$ that

$$\begin{aligned}(A_M \underline{\mathbf{u}} \, , \, \underline{\mathbf{v}})_H &= \sum_{k=1}^M \sum_{\ell=1}^M A_M[\ell, k] \mathbf{u}_k \mathbf{v}_\ell = \sum_{k=1}^M \sum_{\ell=1}^M \langle A\phi_k \, , \, \phi_\ell \rangle \mathbf{u}_k \mathbf{v}_l \\ &= \left\langle A \sum_{k=1}^M \mathbf{u}_k \phi_k \, , \, \sum_{\ell=1}^M \mathbf{v}_\ell \phi_\ell \right\rangle_H = \langle Au_M \, , \, v_M \rangle.\end{aligned}$$

By construction all properties of the operator A are inherited by the *stiffness matrix* $A_M \in \mathbb{R}^{M\times M}$, i.e. A_M is symmetric and positive definite, since A is self-adjoint and H-elliptic. Truely,

$$(A_M \underline{\mathbf{v}}\,,\, \underline{\mathbf{v}}) = (Av_M\,,\, v_M) \geq c_{\mathrm{ell}}^A \|v_M\|_H^2$$

for all $\underline{\mathbf{v}} \in \mathbb{R}^M$ and the corresponding function $v_M \in H_M$ implies that A_M is positive definite.

Theorem 1.8 (Cea's Lemma)**.** *Let $A : H \to H'$ be a bounded and H-elliptic linear operator. For the unique solution $u_M \in H_M$ of the variational problem (1.7) there holds*

$$\|u_M\|_H \leq \frac{1}{c_{ell}^A}\|f\|_{H'}$$

as well as the error estimate

$$\|u - u_M\|_H \leq \frac{c_2^A}{c_{ell}^A} \inf_{v_M \in H_M} \|u - v_M\|_H.$$

The convergence of $u_M \to u$ as $M \to \infty$ then follows from the approximation property of the trial space H_M,

$$\lim_{M\to\infty} \inf_{v_M \in H_M} \|v - v_M\|_H = 0 \qquad \forall v \in H.$$

This means that it is necessary to construct the sequence $\{H_M\}_{M\in\mathbb{N}}$ in such a way that the approximation property can be ascertained.

In the case of tensor products of Hilbert spaces, as long as $A^{(k)} : H^{(k)} \to (H^{(k)})'$ is boundedly invertible and $H^{(k)}$-elliptic, the previous presentation still applies when we set $A := A^{(k)}$ as well as $H := H^{(k)}$ and $H' := (H^{(k)})'$. Hence, we also have the Lax-Milgram lemma and Céa's lemma in the case of tensor products of Hilbert spaces (cf. [55]).

1.4 The Finite Element method and the adaptive process

Here we shall give a brief discussion of the basic notions and definitions with respect to the Finite Element Method. To give only a few references out of the vast literature available on this topic we refer the reader to [12, 13, 46, 48, 56, 22, 23, 51, 27] for more details.

As we will employ the Finite Element Method (FEM) as our method of choice for solving partial differential equations (PDEs), we shall give here a brief presentation of it in order to not overencumber the presentation in later chapters.

Let $D \subset \mathbb{R}^d$, $d = 1, 2, 3, 4$ be a bounded, polygonal/polyhedral domain and denote by Γ its boundary, i.e. $\Gamma := \partial D$. Furthermore, in general suppose the boundary Γ consists of a nonempty Dirichlet boundary part Γ_{D} and a Neumann boundary part Γ_{N}, where $|\Gamma_{\mathrm{D}}| > 0$ and $|\cdot|$ denotes the measure of the given set. Hence, $\overline{\Gamma} = \overline{\Gamma_{\mathrm{D}}} \cup \overline{\Gamma_{\mathrm{N}}}$, where possibly $\Gamma_N = \emptyset$.

A finite partition $\mathcal{T}$ of bounded open and non-overlapping sets $K \subset D$, subsequently called *elements*, such that

$$\bigcup_{K\in\mathcal{T}} \overline{K} = \overline{D}$$

and $\mathcal{T} := \{K_i\}$, $i \in \mathcal{I}$, where $\mathcal{I}$ is an index set that realises an enumeration of all elements, is called an FE *mesh*. In this thesis we shall consider meshes that are *regular* in the sense

of Ciarlet (cf. [19]), i.e. the intersection of two elements K and K' is either empty, a vertex or an edge. Furthermore, we consider 1-irregular meshes, which allow one hanging node per geometric entity of the mesh to be present. This notion will be made precise later. Usually the elements K are thought to be intervals in one space dimension, simplices or quadrilaterals in two dimensions, simplices, hexahedra, pyramids or prisms in three dimensions and so on. In this thesis we will restrict the presentation to tensor products of intervals and rectangles, i.e. we will be concerned with intervals ($d = 1$), quadrilaterals ($d = 2$), cubes ($d = 3$) and hypercubes ($d = 4$). Moreover, let us denote by $\mathcal{N}, \mathcal{E}, \mathcal{F}$ and $\mathcal{Q}$ the sets of *vertices*, *edges*, *faces* and *cubes* in $\mathcal{T}$, respectively.

In two dimensions the geometry of D, i.e. the collection of elements, is represented using $\mathcal{N}$ and $\mathcal{E}$, i.e. the set of vertices and the set of edges, by subdiving D into a finite set of elements. In three dimensions the elements are represented via the set $\mathcal{N}$ of vertices, the set of edges $\mathcal{E}$ and the set of faces $\mathcal{F}$, and in four dimensions we additionally have three dimensional boundary surfaces of hypercubes which are listed in $\mathcal{Q}$. A subscript K to $\mathcal{N}$ or $\mathcal{E}$ or $\mathcal{Q}$ indicates that only those vertices or edges or cubes, respectively, are considered, which are contained in the corresponding set in the subscript. The union of all edges in $\mathcal{E}$ is denoted by $\mathfrak{S}$ and is called the skeleton of $\mathcal{T}$.

Moreover, we shall consider in this thesis only *affine meshes*, i.e. any $K \in \mathcal{T}$ is assumed to be an *affine* image of the reference d-cube

$$\hat{K}^d := (-1, 1)^d,$$

with respect to the *element map* $F_K : \hat{K}^d \to K$ as defined by

$$F_K(\mathbf{x}) = B_K \mathbf{x} + a_K,$$

where $B_K \in \mathbb{R}^{d \times d}$ and $a_K \in \mathbb{R}^d$. We will often drop the superscript d, if the dimension is clear from the context. In this fashion we have that

$$K = F_K(\hat{K}).$$

Furthermore, with every element K we associate a polynomial degree vector $p_K = (p_1, \ldots, p_d)^\top$, where p_i is the partial degree in the i-th coordinate direction, and collect them in the set

$$\mathbf{p}_\mathcal{T} := \{p_K : K \in \mathcal{T}\}.$$

For each element $K \in \mathcal{T}$ we denote the element diameter by

$$h_K := \operatorname{diam}(K)$$

and similarly collect them in the set $\mathbf{h}_\mathcal{T} = \{h_K : K \in \mathcal{T}\}$. The diameter of the biggest incircle of an element K is denoted by

$$\rho_K := \sup\{\operatorname{diam}(B) : B = B_r(x_0) \subset K, r > 0, x_0 \in K\},$$

where $B_r(x) = \{x \in \mathbb{R}^d : \|x - y\|_2 < r\}$ is an open ball with radius $r \in \mathbb{R}_{>0}$ and center in x. The *mesh width* of a partition $\mathcal{T}$ is given as

$$h := \sup_{K \in \mathcal{T}} h_K.$$

In this thesis we will work with *shape regular* meshes $\mathcal{T}$ in the following sense.

Definition 1.9 (Shape regularity)**.** *A family of meshes $\mathfrak{F} = \{\mathcal{T}_i\}_{i\in\mathcal{I}}$ is called* shape regular*, if there exists a constant $\sigma > 0$, independent of i, such that for all $i \in \mathcal{I}$ there holds*

$$\sigma \leq \min_{K\in\mathcal{T}_i} \frac{h_K}{\rho_K} \leq \max_{K\in\mathcal{T}_i} \frac{h_K}{\rho_K} \leq \sigma^{-1}.$$

For any function v on D we introduce the interelemental jump on an edge E with $E = K^+ \cap K^-$ via

$$[\![v]\!] := v^+ - v^- := v|_{K^+} - v|_{K^-}, \quad E = K^+ \cap K^-, E \not\subset \partial D. \tag{1.9}$$

For an edge E of K on Γ, we set $[\![v]\!] = v$.

Note that expressions of the form $[\![\nabla v \cdot \mathbf{n}]\!]$ do not depend on the orientation of the normal vector $\mathbf{n}$ on E.

In the four dimensional setting we use the same notation for jumps on boundary cubes Q, i.e. $Q = K \cap K'$ and $Q \not\subset \partial D$. We will frequently make use of patches around certain geometric entities. Let us start with the two dimensional setting. Denote by ω_K the set of all elements that share an edge with the element K, i.e.

$$\omega_K := \bigcup_{K':K'\cap K=E\in\mathcal{E}} K',$$

and by $\tilde{\omega}_K$ the set of all elements that share at least one vertex with K, i.e.

$$\tilde{\omega}_K := \bigcup_{K':\mathcal{N}_{K'}\cap\mathcal{N}_K\neq\emptyset} K'.$$

In four dimensions in the definition of ω_K we replace $E \in \mathcal{E}$ with $Q \in \mathcal{Q}$. Analogously we understand the sets $\omega_E, \tilde{\omega}_E, \omega_F, \tilde{\omega}_F, \omega_Q$ and $\tilde{\omega}_Q$ for $E \in \mathcal{E}, F \in \mathcal{F}$ and $Q \in \mathcal{Q}$. We may occasionally term these sets *edge*, *face*, *cube* or *element patches*.

For adaptive methods to come we still need to define the notion of a hanging node.

Definition 1.10 (Hanging node)**.** *Let $\mathcal{T}$ a mesh and a node $x \in \mathcal{N}$. If x lies on the boundary of an element $K \in \mathcal{T}$, i.e. $x \in \partial K$, but is not a vertex of K, then x is called a* hanging node*.*

This definition is *not complete* on purpose. As in two space dimensions for our construction there cannot be hanging nodes situated on the boundary of $D \times D$, the situation in three and four space dimensions is different. There hanging nodes on edges and faces on the boundary of the domain D are possible. Hence, the aforementioned definition is kept a little more open to encompass all cases considered in this thesis.

Our refinement procedure of choice will involve 1-irregular meshes. These meshes are characterized by the fact that on each geometric entity, there is at most one hanging node. We formalize this in the following definition.

Definition 1.11 (1-irregularity)**.** *If for all elements K of a mesh $\mathcal{T}$, every edge $E \in \mathcal{E}$, every face $F \in \mathcal{F}$, and every cube $Q \in \mathcal{Q}$ of K, respectively, has at most one associated hanging node, then the mesh $\mathcal{T}$ is said to be* 1-irregular*.*

1.4.1 General Finite Element spaces

Let us now define the Finite Element spaces which will be used in the subsequent analysis. Since we will be operating in $H_0^{1,1}(\mathcal{D})$ it is mandatory that the global shape functions on the FE spaces are continuous across the whole mesh. Let $\mathcal{I} = [a, b]$.

Definition 1.12 (1D finite element space)**.** *Let $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ a partition of $\mathcal{I}$ into subintervals $K_i := I_i := [x_i, x_{i+1}], i = 0, \ldots, n-1$ and $\mathcal{T} := \{K_i : i = 1, \ldots, n-1\}$ a mesh on $\mathcal{I}$. For given degree vector $p \equiv p_{\mathcal{T}}$, we define the space*

$$\mathcal{S}^{p,k}(\mathcal{T}) := \{v \in C^k(\mathcal{I}) : v|_K \in \mathcal{P}_{p_K}(K), K \in \mathcal{T}\},$$

the space of k times continuously differentiable functions that restricted to an element $K \in \mathcal{T}$ are polynomials of degree p_K with

$$\mathcal{P}_{p_K}(K) = \operatorname{span}\{x^n : 0 \leq n \leq p_K\}.$$

Moreover, let

$$\mathcal{S}_0^{p,k}(\mathcal{T}) := \{v \in \mathcal{S}^{p,k}(\mathcal{T}) : v(x_0) = v(x_n) = 0\}.$$

Definition 1.13 (Tensor finite element spaces)**.** *Consider a partition of $\mathcal{I}^d$ into d-hypercubes of the form*

$$K_{\mathbf{i}} := \mathcal{I}^d_{i_1,\ldots,i_d} := [x_{1,i_1}, x_{1,i_1+1}] \times \cdots \times [x_{d,i_d}, x_{d,i_d+1}], \quad \mathbf{i} = (i_1, \ldots, i_d) \in \mathbf{I},$$

where $\mathbf{I}$ is a suitable index set realizing an enumeration of the elements.
Thus, $\mathcal{T} := \{K_{\mathbf{i}} : \mathbf{i} \in \mathbf{I}\}$ is a mesh on $\mathcal{I}^d$. We set

$$\mathcal{S}^{p,k}(\mathcal{T}) := \{v \in C^k(\mathcal{I}^d) : v|_K \in \mathcal{P}_{p_K}(K), K \in \mathcal{T}\}$$

with

$$\mathcal{P}_{p_K}(K) = \operatorname{span}\{\mathbf{x}^{\mathbf{n}} : \mathbf{0} \leq \mathbf{n} \leq p_K\}.$$

Analogously, we define as in the one dimensional case

$$\mathcal{S}_0^{p,k}(\mathcal{T}) := \{v \in \mathcal{S}^{p,k}(\mathcal{T}) : v = 0 \text{ on } \partial\mathcal{I}^d\}.$$

Moreover, we denote by

$$\mathcal{S}^{p,-1}(\mathcal{T}) := \{v : v|_K \in \mathcal{P}_{p_K}(K), K \in \mathcal{T}\}$$

the discontinuous and broken Finite Element space. We shall at some points abuse notation and denote by $\mathcal{P}_k(\mathcal{T}) \equiv \mathcal{S}^{k,-1}(\mathcal{T})$ the space of elementwise polynomials of partial degree at most k on the mesh $\mathcal{T}$.

1.4.2 Adaptive Finite Elements

The general idea of adaptive methods is to refine areas of the mesh where the error is larger than its surrounding. Usually the true error is not available and so another way of estimating the error has to be used. To this end, *a posteriori* error estimators are developed and shown to fulfill certain properties that help guide the adaptive process to guarantee convergence. In general the adaptive process consists of the following four steps:

(1) Solve, **(2) Estimate**, **(3) Mark**, **(4) Refine**.

This procedure is formalized in Algorithm 1. We start with an initial triangulation $\mathcal{T}_0$ and solve a discrete problem on $\mathcal{T}_0$. The discrete solution $u_{\mathcal{T}_0}$ is then used to compute an error estimator or error indicator η which is an approximation of the true error $\|u - u_{\mathcal{T}_0}\|$ for a certain norm $\|\cdot\|$.

Algorithm 1 Adaptive Finite Element Algorithm

Input: data f and tolerance ε

Output: A numerical solution $u_{\mathcal{T}}$ with an error less than ε

1: Create an initial regular coarse mesh $\mathcal{T}_0$ and set $k = 0$.
2: **repeat**
3: **Solve** discrete problem on mesh $\mathcal{T}_k$
4: **Estimate** error using estimator/indicator η_K for every element $K \in \mathcal{T}$
5: **Mark** elements of $\mathcal{T}$ subject to a rule and the magnitude of η_K
6: **Refine** marked elements of $\mathcal{T}_k$ to construct $\mathcal{T}_{k+1}$ and set $k = k + 1$
7: **until** $(\sum_K \eta_K)^{1/2} < \varepsilon$

Definition 1.14 (Error estimator η)**.** *An approximation η of $\|u - u_{\mathcal{T}}\|$ is called an* a posteriori error estimator*, or just estimator for brevity, if it is a computable function of known quantities such as the right-hand side f, the domain $\mathcal{D}$, the boundary $\partial\mathcal{D}$ as well as $u_{\mathcal{T}}$ itself or certain derivatives of the numerical solution $u_{\mathcal{T}}$.*

Now after the first and second step of the kth iteration of the algorithm we are equipped with a numerical solution $u_{\mathcal{T}_k}$ and a corresponding elementwise error indicator η_K with the global estimator given by $\eta := (\sum \eta_K^2)^{1/2}$, which we suppose to be at least reliable (see below). The next step, **Mark**, can be done in many ways and the reader is referred to [54] the literature for all types of marking strategies. For example, Dörfler marking consists of selecting a subset $\tilde{\mathcal{T}}_k \subset \mathcal{T}_k$, such that with $\vartheta \in (0, 1)$ there holds

$$\sum_{K \in \tilde{\mathcal{T}}_k} \eta_K^2 \geq \vartheta \sum_{K \in \mathcal{T}_k} \eta_K^2.$$

In this thesis, for reasons of simplicity we have chosen to implement the so-called *maximum strategy*. Here we determine the maximal error indicator $\eta_{\max} := \max_{K \in \mathcal{T}_k} \eta_K$ and mark every element that fulfills the condition $\eta_K \geq \vartheta \eta_{\max}$. The so marked elements will then be refined. The refinement procedure for the different settings will be discussed later.

Moreover, error indicators can be developed in many different ways, depending on which aspect of the error has been put under scrutiny. For example, one could ask to minimise the local residual with respect to the discrete solution or for a certain functional that depends on the error to be minimal or maximal. After a certain indicator is computed the information is then processed to find the elements with the biggest error and mark those for refinement. The refinement process is then executed and the adaptive process begins anew.

The above definition of an error estimator η is not useful *per se*, but only if η provides computable upper and lower bounds of the true error.

Definition 1.15 (Reliability of η)**.** *Let η be an error estimator. Then η is called a* reliable *a posteriori error estimator if there exists a constant C_{rel}, such that there holds an upper bound of the form*

$$\|u - u_{\mathcal{T}}\| \leq C_{rel}\eta + H_{rel},$$

for a certain norm of the error $\|\cdot\|$ and such that there holds $H_{rel} = o(\|u - u_{\mathcal{T}}\|)$, where the function H_{rel} denotes a generic higher order error term.

Definition 1.16 (Efficiency of η)**.** *Let η be an error estimator. Then η is called an* efficient *a posteriori error estimator if there exists a constant C_{eff}, such that there holds a lower bound of the form*

$$\eta \leq C_{\mathit{eff}}\|u - u_{\mathcal{T}}\| + H_{\mathit{eff}}$$

for a certain norm of the error $\|\cdot\|$ *and such that the generic higher order error term* H_{eff} *satisfies* $H_{\mathit{eff}} = o(\|u - u_{\mathcal{T}}\|)$.

Definition 1.17 (Asymptotic Exactness)**.** *If* η *is reliable and efficient, then* η *is called* asymptotically exact.

In this thesis we are concerned with the development of a posteriori error indication techniques to guide the adaptive refinement and approximation for the deterministic second moment problem in two and four space dimensions. To this end we have a look at residual error estimation, a hierarchical error estimation approach with higher order basis functions, and a so-called averaging error estimation approach. For each of these error estimators reliability and efficiency are investigated and the corresponding error analysis is performed.

Chapter 2

Deterministic Second Moment Equations

This chapter is dedicated to the development of adaptive methods for the approximation of second moments of elliptic partial differential equations under uncertainty by means of deterministic moment equations.

We begin with the general definition of deterministic moment equations for a chosen simple model problem, have a brief look at the regularity, and give a short summary of the necessary ideas of already existing literature. In passing we also discuss the functional analytic setting and the solution procedure briefly. Then we develop and analyze *a posteriori* error estimators for selected one and two dimensional ellitpic model problems. At the end of the chapter we investigate the convergence of the adaptive process when guided by the developed error estimation techniques and look at numerical experiments which validate our theoretical findings.

2.1 Model problem and moment equations

In the following we are concerned with the *deterministic* model problem as given by

$$\begin{aligned} -\nabla \cdot (a(x)\nabla u(x)) &= f(x), \quad \text{in } D, \\ u &= 0, \quad \text{on } \partial D, \end{aligned} \tag{2.1}$$

as well as its *stochastic* counterpart for almost every $\omega \in \Omega$

$$\begin{aligned} -\nabla \cdot (a(x,\omega)\nabla u(x,\omega)) &= f(x,\omega), \quad \text{in } D, \\ u &= 0, \quad \text{on } \partial D, \end{aligned} \tag{2.2}$$

where $(\Omega, \Sigma, \mathbb{P})$ is a probability space, and $D \subset \mathbb{R}^d$, $d = 1, 2$ is an open and bounded Lipschitz domain. Note that either a or f or both may depend randomly on ω, where this dependence is modeled by means of the probability space $(\Omega, \Sigma, \mathbb{P})$, and as such we term the solution of the elliptic model problem u to be *stochastic*.

From now on we consider the diffusion coefficient a to be deterministic, i.e. it only depends on the spatial variable $x \in D$. Furthermore, we assume a to fulfill the following ellipticity condition

$$0 < a_- \leq a(x) \leq a_+ < \infty, \quad \text{a.e. in } D.$$

Let $V := H_0^1(D), V' = H^{-1}(D)$. Defining the operator $A : V \to V'$ by $Av := -\nabla \cdot (a(x)\nabla v)$, we can write the deterministic model problem as an operator equation in V'

$$Au = f,$$

where we consider $u \in V$. Since the expectation commutes with the operator A, taking the expectation of this equation we arrive at the deterministic first moment problem in $L^1(\Omega; V')$

$$A(\mathbb{E}[u]) = \mathbb{E}[f],$$

where $u \in L^1(\Omega; V)$.

In general considering the k-fold tensorization of the operator A, i.e.

$$A^{(k)} := \underbrace{A \otimes \cdots \otimes A}_{k \text{ times}},$$

which maps the space $V^{(k)}$ to $(V')^{(k)}$, taking the expectation of the tensorized operator equation yields for $u^{(k)} \in L^1(\Omega; V^{(k)})$ and $f^{(k)} \in L^1(\Omega; (V')^{(k)})$

$$A^{(k)} \mathcal{M}^k u = \mathcal{M}^k f. \tag{2.3}$$

As this equation provides a *deterministic* way for the computation of the statistical moments of u, it is called the *deterministic k-th moment problem* or the *deterministic k-th moment equation*.

In [55, 50, 49, 30] it is shown that as long as $A^{(k)}$ is injective, continuous and strongly elliptic, i.e. satisfies a Gårding inequality, as a map from $V^{(k)}$ to $(V')^{(k)}$, then (2.3) has a unique solution for every $f \in L^k(\Omega; (V')^{(k)})$ which coincides with the k-th moment $\mathcal{M}^k u$.

Moreover, by [55, Thm. 2.6], also cf. [18], there holds a shift theorem, if the corresponding operator allows a shift theorem.

If $\mathcal{M}^k f \in H^{s,\dots,s}(D^k)$ for some $s \geq -1$ then $\mathcal{M}^k u \in H^{s+2,\dots,s+2}(D^k) \cap H^{1,\dots,1}(D^k)$ and there holds the a priori estimate

$$\|\mathcal{M}^k u\|_{H^{s+2,\dots,s+2}(D^k)} \leq C \|\mathcal{M}^k f\|_{H^{s,\dots,s}(D^k)}, \qquad s \geq -1.$$

Our main interest is the deterministic 2nd moment equation, which is the foundation for the adaptive algorithms we shall develop in this chapter.

2.2 1D model problem

2.2.1 Problem formulation

Letting $a \equiv 1$, we consider the 1D stochastic model problem, cf. (2.2), in $D = [-1, 1]$ of finding $u \in H_0^1(D)$, such that

$$\begin{aligned} -\partial_x^2 u(\cdot, \omega) &= f(\cdot, \omega) && \text{in } H^{-1}(D), \\ u &= 0 && \text{on } \partial D, \end{aligned} \tag{2.4}$$

for almost all $\omega \in \Omega$.

With $V := H_0^1(D), V' = H^{-1}(D)$, and $A := -\partial_x^2$ we find that the variational problem for the associated operator equation (cf. (1.3)) reads:

Problem 2.1. *Given $f(\omega) \in H^{-1}(D)$, find $u(\omega) \in H_0^1(D)$, such that*

$$\int_D u'(x, \omega) v'(x) \, \mathrm{d}x = \int_D f(x, \omega) v(x) \, \mathrm{d}x, \qquad \forall v \in H_0^1(D), \text{ for } \mathbb{P}\text{-a.a. } \omega \in \Omega.$$

The corresponding deterministic k-th moment problem, cf. (2.3), for the one dimensional model problem then takes the following form:

Problem 2.2. *Given* $\mathcal{M}^k f \in H^{-1,\dots,-1}(D^k)$, *find* $\mathcal{M}^k u \in H_0^{1,\dots,1}(D^k)$, *such that*

$$\int_{D^k} \partial^{\mathbf{1}}(\mathcal{M}^k u)\, \partial^{\mathbf{1}} v \,\mathrm{d}\mathbf{x} = \int_{D^k} \mathcal{M}^k f v \,\mathrm{d}\mathbf{x}, \quad \forall v \in H_0^{1,\dots,1}(D^k), \tag{2.5}$$

where $\mathbf{1} = (1, 1, \dots, 1)$.

In the particular case $k = 2$ we set $\mathcal{D} = D \times D$ and are concerned with solving the deterministic second moment problem corresponding to (2.3):

Given $\mathcal{C}_f \in H^{-1,-1}(\mathcal{D})$, find $\mathcal{C}_u \in H_0^{1,1}(\mathcal{D})$, such that

$$\int_{\mathcal{D}} \partial_x \partial_y \mathcal{C}_u \partial_x \partial_y v \,\mathrm{d}x\,\mathrm{d}y = \int_{\mathcal{D}} \mathcal{C}_f \, v \,\mathrm{d}x\,\mathrm{d}y, \quad \forall v \in H_0^{1,1}(\mathcal{D}). \tag{2.6}$$

The discrete version of (2.6) then takes the form:

Given $\mathcal{C}_f \in H^{-1,-1}(\mathcal{D})$, find $u_{\mathcal{T}} \in \mathcal{S}_0^{p,0}(\mathcal{T})$, such that

$$\int_{\mathcal{D}} \partial_x \partial_y u_{\mathcal{T}} \partial_x \partial_y v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y = \int_{\mathcal{D}} \mathcal{C}_f v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y, \quad \forall v_{\mathcal{T}} \in \mathcal{S}_0^{p,0}(\mathcal{T}), \tag{2.7}$$

where we write $u_{\mathcal{T}}$ instead of $\mathcal{C}_{u,\mathcal{T}}$ to shorten the notation and keep subscript indices hopefully minimal. Note that, since $\mathcal{S}_0^{p,0}(\mathcal{T}) \subset H_0^{1,1}(\mathcal{D})$, the approximation is conforming and we have Galerkin orthogonality

$$\mathcal{B}(u - u_{\mathcal{T}}, v_{\mathcal{T}}) = 0, \quad \forall v_{\mathcal{T}} \in \mathcal{S}_0^{p,0}(\mathcal{T})$$

with the associated bilinear form

$$\mathcal{B}(u, v) = \int_{\mathcal{D}} \partial_x \partial_y u \partial_x \partial_y v \,\mathrm{d}x\,\mathrm{d}y.$$

Also note that the energy norm of this problem is the $|\cdot|_{H^{1,1}(\mathcal{D})}$-seminorm. The latter is a norm on $H_0^{1,1}(\mathcal{D})$ by the crossnorm property and corresponding one dimensional Friedrichs inequalities. This, in combination with the symmetry of the bilinear form yields not only the unique solvability of the problem by the Lax-Milgram Lemma, but also the best approximation property by means of Céa's Lemma.

2.2.2 Approximation and auxiliary results

In order to accomplish a concise presentation, we firstly define a suitable two dimensional approximation operator $\Pi_{p_K}^{1,1}$ by tensorization of one dimensional operators, present the error analysis and collect various auxiliary results that are used in subsequent sections.

Regarding the presentation in this section see also [48, 43, 54].

Definition 2.3 (Legendre polynomials and their antiderivatives)**.** *The Legendre polynomials* $L_n, n = 0, 1, 2, \dots$ *are defined as solutions to the* Legendre differential equation *on the interval* $[-1, 1]$

$$((1 - x^2)L_n'(x))' + n(n + 1)L_n(x) = 0$$

and may be expressed by Rodrigues' formula

$$L_n(x) = \frac{1}{2^n n!} \frac{\mathrm{d}^n}{\mathrm{d}x^n}((1 - x^2)^n).$$

The antiderivatives of the Legendre polynomials $\hat{L}_n$ are then defined as follows

$$\hat{L}_n(\xi) := \begin{cases} 1, & n = 0, \\ \int_{-1}^{\xi} L_{n-1}(t)\,\mathrm{d}t, & n \geq 1. \end{cases}$$

Moreover, there holds

$$\hat{L}_n = \frac{L_{n+1} - L_{n-1}}{2n+3}$$

and $\hat{L}_n(\pm 1) = 0$ for $n \geq 2$.

Definition 2.4 (1D projection operators)**.** *If $u \in L^2([-1,1])$, then u can be expanded into a* Legendre series

$$u(\xi) = \sum_{i=0}^{\infty} b_i L_i(\xi),$$

where the L_i are the Legendre polynomials with respect to $[-1,1]$. Then the operator

$$\hat{\pi}^p : L^2([-1,1]) \longrightarrow \mathcal{P}_p([-1,1]),$$

termed the $L^2([-1,1])$-projection, is defined by truncation of the Legendre series after $p+1$ terms, i.e.

$$\hat{\pi}^p u(\xi) = \sum_{i=0}^{p} b_i L_i(\xi).$$

For any $u \in H^1([-1,1])$ we can then define the following operator

$$\pi^p : H^1([-1,1]) \longrightarrow \mathcal{P}_p([-1,1])$$

for any $\xi \in [-1,1]$ by

$$\begin{aligned} (\pi^p u)(\xi) &= u(-1) + \int_{-1}^{\xi} \hat{\pi}^{p-1} u'(t)\,\mathrm{d}t \\ &= u(-1) + \sum_{i=0}^{p-1} b_i \int_{-1}^{\xi} L_i(t)\,\mathrm{d}t \\ &= u(-1)\hat{L}_0 + \sum_{i=1}^{p} b_i \hat{L}_i, \end{aligned}$$

where the coefficients b_i are given by

$$b_i = \frac{2i+1}{2} \int_{-1}^{1} u'(t) L_i(t)\,\mathrm{d}t.$$

It is easily seen that there holds $\pi^p u(\pm 1) = u(\pm 1)$ (cf. [48, Theorem 3.14]).

Definition 2.5. *By tensorization of π^p we can now define for any $\hat{u} \in H^{1,1}(\hat{K})$ and $p_{\hat{K}} = (\hat{p}_1, \hat{p}_2)$ the operator*

$$\hat{\Pi}^{1,1}_{p_{\hat{K}}} : H^{1,1}(\hat{K}) \to \mathcal{P}_{p_{\hat{K}}}(\hat{K})$$

via

$$\hat{\Pi}^{1,1}_{p_{\hat{K}}} \hat{u}(\xi,\eta) := (\pi^{\hat{p}_1}_{\xi} \otimes \pi^{\hat{p}_2}_{\eta}) \hat{u}(\xi,\eta),$$

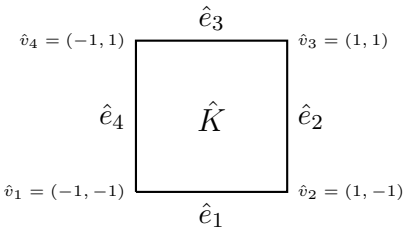

Figure 2.1: The reference element $\hat{K} = [-1, 1]^2$

where the subscripts ξ and η indicate the coordinate direction, which the operators act upon. By the properties of the one dimensional operator π^p, we have the following properties for $\hat{\Pi}^{1,1}_{p_{\hat{K}}}$ (cf. [48, Lemma 4.67])

$$\begin{aligned} \hat{\Pi}^{1,1}_{p_{\hat{K}}} \hat{u}(\pm 1, \pm 1) &= \hat{u}(\pm 1, \pm 1), \\ (\hat{\Pi}^{1,1}_{p_{\hat{K}}} \hat{u})|_{\hat{e}_i} &= \begin{cases} \pi^{\hat{p}_1}_{\xi}(\hat{u}|_{\hat{e}_i}), & \text{if } i = 1, 3, \\ \pi^{\hat{p}_2}_{\eta}(\hat{u}|_{\hat{e}_i}), & \text{if } i = 2, 4. \end{cases} \end{aligned}$$

For an application on Maxwell's equations involving this operator see also [21].

Definition 2.6 (Affine element map)**.** *We consider an affine map $F_K : \hat{K} \to K$ that maps the reference element $\hat{K}$ to the element $K \in \mathcal{T}$. For a rectangular and axially parallel element K, using the notation from Figure 2.2, F_K can be written explicitly as*

$$F_K(\hat{x}) = \frac{1}{4}(v_1^\top + v_2^\top + v_3^\top + v_4^\top) + \frac{1}{2} \begin{pmatrix} h_x & 0 \\ 0 & h_y \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix},$$

where $h_x = x_2 - x_1$ and $h_y = y_2 - y_1$ denote the edge length in the x- and y-direction, respectively. Note that for square elements K, this representation holds with $h_x = h_y$.

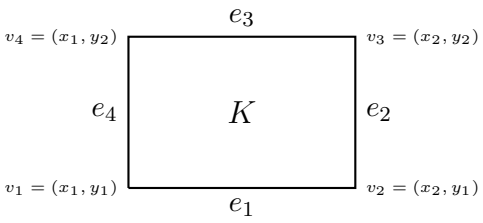

Figure 2.2: An element $K = [x_1, x_2] \times [y_1, y_2] \in \mathcal{T}$

In order to enable the error analysis in the two dimensional setting we shall now give local and global approximation results that involve the operator $\hat{\Pi}^{1,1}_{p_K}$.

Lemma 2.7 (Trace inequality)**.** *For any $\hat{u} \in H^{1,1}(\hat{K})$ there holds*

$$\|\hat{u} - \hat{\Pi}^{1,1}_{p_{\hat{K}}} \hat{u}\|^2_{L^2(\hat{e})} \leq C \frac{1}{p(p+1)} \left(\|\partial_\tau \hat{u}\|^2_{L^2(\hat{K})} + \|\partial_\xi \partial_\eta \hat{u}\|^2_{L^2(\hat{K})} \right), \tag{2.8}$$

where $\hat{e}$ being any edge of the reference element $\hat{K}$, $p \in \mathbb{N}$ with $p_{\hat{K}} = (p, p)$ and

$$\partial_\tau = \begin{cases} \partial_\xi, & \text{if } \hat{e} = \hat{e}_1 \text{ or } \hat{e}_3, \\ \partial_\eta, & \text{if } \hat{e} = \hat{e}_2 \text{ or } \hat{e}_4. \end{cases} \tag{2.9}$$

Proof. Without loss of generality, consider $\hat{e} = \hat{e}_1$. The other cases are then obtained by a rotation of coordinates. Considering u represented as an infinite series of antiderivatives of the Legendre polynomials as $\hat{u}(\xi,\eta) = \sum_{i,j=0}^{\infty} c_{ij}\hat{L}_i(\xi)\hat{L}_j(\eta)$ with $(\xi,\eta)^\top \in \hat{K}$ and using properties of the projector $\hat{\Pi}^{1,1}_{p_{\hat{K}}}$ in addition to [48, Theorem 3.14], we get

$$\begin{aligned}
\|\hat{u} - \hat{\Pi}^{1,1}_{p_{\hat{K}}}\hat{u}\|^2_{L^2(\hat{e}_1)} &= \|(\hat{u} - \hat{\Pi}^{1,1}_{p_{\hat{K}}}\hat{u})(\cdot,-1)\|^2_{L^2([-1,1])} = \|(\hat{u} - \pi^{\hat{p}_1}_{\xi}\hat{u})(\cdot,-1)\|^2_{L^2([-1,1])} \\
&\leq \left\| \frac{(\hat{u} - \pi^{\hat{p}_1}_{\xi}\hat{u})(\cdot,-1)}{\sqrt{1-\xi^2}} \right\|^2_{L^2([-1,1])} = \sum_{i=p}^{\infty} \frac{2}{i(i+1)(2i+1)} |c_{i0}|^2 \\
&\leq \frac{1}{p(p+1)} \|\partial_\xi \hat{u}(\cdot,-1)\|^2_{L^2([-1,1])} \\
&\leq C \frac{1}{p(p+1)} \left(\|\partial_\xi \hat{u}\|^2_{L^2(\hat{K})} + \|\partial_\xi \partial_\eta \hat{u}\|^2_{L^2(\hat{K})} \right)
\end{aligned}$$

where in the last step invoking a trace inequality on $\hat{K}$ completes the proof. □

Lemma 2.8 ([43, Lemma 5.24,5.25,5.26])**.** *Let $p \in \mathbb{N}$, $\hat{v} \in H^{k+1}(\hat{K})$ and $p_{\hat{K}} = (p,p)$ with $p \in \mathbb{N}$. Then there holds*

(i) $$\begin{aligned}\|\hat{v} - \hat{\Pi}^{1,1}_{p_{\hat{K}}}\hat{v}\|^2_{L^2(\hat{K})} \leq & \frac{1}{p(p+1)} \left(2\frac{(p-s)!}{(p+s)!} \|\partial^{s+1}_\xi \hat{v}\|^2_{L^2(\hat{K})} + 4\frac{(p-s)!}{(p+s)!} \|\partial^{s+1}_\eta \hat{v}\|^2_{L^2(\hat{K})} \right. \\ & \left. + \frac{4}{p(p+1)} \frac{(p-s+1)!}{(p+s-1)!} \|\partial_\xi \partial^s_\eta \hat{v}\|^2_{L^2(\hat{K})} \right)\end{aligned}$$

for any integer $1 \leq s \leq \min\{p,k\}$.

(ii) $$\begin{aligned}\|\partial_\xi \left(\hat{v} - \hat{\Pi}^{1,1}_{p_{\hat{K}}}\hat{v} \right)\|^2_{L^2(\hat{K})} \leq & 2\frac{(p-s)!}{(p+s)!} \|\partial^{s+1}_\xi \hat{v}\|^2_{L^2(\hat{K})} \\ & + \frac{2}{p(p+1)} \frac{(p-s+1)!}{(p+s-1)!} \|\partial_\xi \partial^s_\eta \hat{v}\|^2_{L^2(\hat{K})}\end{aligned}$$

for any integer $1 \leq s \leq \min\{p,k\}$ and analagously for $\|\partial_\eta \left(\hat{v} - \hat{\Pi}^{1,1}_{p_{\hat{K}}}\hat{v} \right)\|^2_{L^2(\hat{K})}$.

(iii) $$\begin{aligned}\|\partial_\xi \partial_\eta \left(\hat{v} - \hat{\Pi}^{1,1}_{p_{\hat{K}}}\hat{v} \right)\|^2_{L^2(\hat{K})} \leq & 2\frac{(p-s)!}{(p+s)!} \|\partial^{s+1}_\xi \partial_\eta \hat{v}\|^2_{L^2(\hat{K})} \\ & + 2\frac{(p-s+1)!}{(p+s-1)!} \|\partial_\xi \partial^{s+1}_\eta \hat{v}\|^2_{L^2(\hat{K})}\end{aligned}$$

for any integer $0 \leq s \leq \min\{p,k-1\}$.

In particular for a function $u \in H^{1,1}(\hat{K})$ we have the following result.

Lemma 2.9. *Let $\hat{u} \in H^{1,1}(\hat{K})$ and $p_{\hat{K}} = (p,p)$. Then the projector $\hat{\Pi}^{1,1}_{p_{\hat{K}}}$ satisfies the bound*

$$\|\hat{u} - \hat{\Pi}^{1,1}_{p_{\hat{K}}}\hat{u}\|^2_{L^2(\hat{K})} \leq \frac{1}{p(p+1)} \left(2\|\partial_\xi \hat{u}\|^2_{L^2(\hat{K})} + 4\|\partial_\eta \hat{u}\|^2_{L^2(\hat{K})} + \frac{4}{p(p+1)} \|\partial_\xi \partial_\eta \hat{u}\|^2_{L^2(\hat{K})} \right)$$

Proof. See [43, Lemma 5.24] for $t_1 = t_2 = t_3 = 0$. □

Remark 2.10. *By the above estimate, we also have for $u \in H^{1,1}(\hat{K})$ and $p_{\hat{K}} = (p,p)$ with $p \in \mathbb{N}$ the trivial estimate*

$$\|\hat{u} - \hat{\Pi}^{1,1}_{\mathbf{p}} \hat{u}\|^2_{L^2(\hat{K})} \leq \frac{4}{p(p+1)} \|u\|^2_{H^{1,1}(\hat{K})}$$

as well as for $u \in H^{1,1}_0(\hat{K})$ the estimate

$$\|\hat{u} - \hat{\Pi}^{1,1}_{\mathbf{p}} \hat{u}\|^2_{L^2(\hat{K})} \leq \frac{C}{p(p+1)} |u|^2_{H^{1,1}(\hat{K})},$$

where the latter follows by corresponding one dimensional Friedrichs' inequalities.

In the following analysis we will make extensive use of the shape regularity of the mesh $\mathcal{T}$. Let h_x and h_y denote the local mesh width in the x- or y-direction, respectively, and let $h_E := \operatorname{diam}(E)$. That means in particular that for some constants $c, C > 0$ we have the following bounds

$$h_x h_y \leq C h_K^2, \quad c h_K \leq h_E \leq C h_K, \quad h_x \leq C h_E, \quad h_y \leq C h_E.$$

Investigating efficiency of the proposed a posteriori error estimators will involve polynomial inverse inequalities on elements for which the following lemma is a crucial building block. A proof can be found in [54, Lemma 3.42,Proposition 3.44], where we note the typo in the second inequality, i.e. the missing square root with respect to the term $(1-x^2)$ in [54, Lemma 3.42].

Lemma 2.11. *Let $q \in \mathcal{P}_p([-1,1])$, $p \in \mathbb{N}_0$. Then there holds*

$$\begin{aligned}
\|(1-x^2)^{1/2} q'\|_{L^2([-1,1])} &\leq p(p+1) \|q\|_{L^2([-1,1])}, \\
\|q\|_{L^2([-1,1])} &\leq (p+2) \|(1-x^2)^{1/2} q\|_{L^2([-1,1])}, \\
\| \left((1-x^2)^{1/2} q \right)' \|_{L^2([-1,1])} &\leq (4 + 2\sqrt{p(p+1)}) \|q\|_{L^2([-1,1])}.
\end{aligned}$$

We are now ready to give local estimates and trace inequalities for elements $K \in \mathcal{T}$ and edges $E \in \mathcal{E}$. For this purpose, denoting the polynomial degree vector on K by p_K, we define the projection on K in the usual way by

$$(\Pi^{1,1}_{p_K} u)(x) = (\hat{\Pi}^{1,1}_{p_K} \hat{u})(F_K^{-1}(x))$$

where $\hat{u}(\hat{x}) = u(x) = u(F_K(\hat{x}))$.

Lemma 2.12. *For $u \in H^{1,1}(K)$ and a polynomial degree vector $p_K = (p,p)$ for the element K with $p \in \mathbb{N}$, there holds*

$$\|u - \Pi^{1,1}_{p_K} u\|^2_{L^2(E)} \leq C \frac{h_E}{p(p+1)} \|\partial_\tau u\|^2_{L^2(K)} + \frac{h_E^3}{p(p+1)} \|\partial_x \partial_y u\|^2_{L^2(K)},$$

$$\|u - \Pi^{1,1}_{p_K} u\|^2_{L^2(K)} \leq C \frac{h_K^2}{p(p+1)} \left(\|\partial_x u\|^2_{L^2(K)} + \|\partial_y u\|^2_{L^2(K)} + |u|^2_{H^{1,1}(K)} \right),$$

where the constant C only depends on the shape regularity of the element K.

Proof. Let us denote by $\vec{x} = (x, y)^\top$ and $\hat{x} = (\xi, \eta)^\top$ the coordinates in K and $\hat{K}$, respectively. We find that

$$\begin{aligned}
\|u - \Pi^{1,1}_{p_K} u\|^2_{L^2(E)} &= \int_E (u - \Pi^{1,1}_{p_K} u)^2 \,\mathrm{d}s_{\vec{x}} \\
&= \int_{\hat{E}} \left(u(F_K(\hat{x})) - \Pi^{1,1}_{p_K} u(F_K(\hat{x}))\right)^2 \frac{h_E}{2} \,\mathrm{d}s_{\hat{x}} \\
&= \frac{h_E}{2} \int_{\hat{E}} (\hat{u}(\hat{x}) - \hat{\Pi}^{1,1}_{p_K} \hat{u}(\hat{x}))^2 \,\mathrm{d}s_{\hat{x}} \\
&= \frac{h_E}{2} \|\hat{u} - \hat{\Pi}^{1,1}_{p_K} \hat{u}\|^2_{L^2(\hat{E})}.
\end{aligned}$$

Applying (2.8), we conclude with a scaling argument

$$\begin{aligned}
\|u - \Pi^{1,1}_{p_K} u\|^2_{L^2(E)} &\leq \frac{1}{p(p+1)} \frac{h_E}{2} \left(\|\partial_\tau u\|^2_{L^2(\hat{K})} + \|\partial_\xi \partial_\eta u\|^2_{L^2(\hat{K})} \right) \\
&\leq C \frac{h_E}{p(p+1)} \|\partial_\tau u\|^2_{L^2(K)} + \frac{h_E^3}{p(p+1)} \|\partial_x \partial_y u\|^2_{L^2(K)}.
\end{aligned}$$

For the last estimate we proceed similarly, now using Lemma 2.9 and a scaling argument, to find

$$\begin{aligned}
\|u - \Pi^{1,1}_{p_K} u\|^2_{L^2(K)} &= \int_K (u - \Pi^{1,1}_{p_K} u)^2 \,\mathrm{d}\vec{x} = \int_{\hat{K}} \left(u(F_K(\hat{x})) - \Pi^{1,1}_{p_K} u(F_K(\hat{x}))\right)^2 \frac{h_x h_y}{4} \,\mathrm{d}\hat{x} \\
&= \frac{h_x h_y}{4} \|\hat{u} - \hat{\Pi}^{1,1}_{p_K} \hat{u}\|^2_{L^2(K)} \\
&\leq C \frac{h_x h_y}{p(p+1)} \left\{ 2\|\partial_\xi \hat{u}\|^2_{L^2(\hat{K})} + 2\|\partial_\eta \hat{u}\|^2_{L^2(\hat{K})} + \frac{4}{p(p+1)} \|\partial_\xi \partial_\eta \hat{u}\|^2_{L^2(\hat{K})} \right\} \\
&\leq C \frac{h_K^2}{p(p+1)} \left\{ \|\partial_x u\|^2_{L^2(K)} + \|\partial_y u\|^2_{L^2(K)} + \frac{h_K^2}{p(p+1)} \|\partial_x \partial_y u\|^2_{L^2(K)} \right\} \\
&\leq C \frac{h_K^2}{p(p+1)} \left\{ \|\partial_x u\|^2_{L^2(K)} + \|\partial_y u\|^2_{L^2(K)} + \|\partial_x \partial_y u\|^2_{L^2(K)} \right\}
\end{aligned}$$

where we have used that $h_x h_y \leq C h_K^2$. □

In the following we will often make use of the so-called edge and element bubbles ψ_E and ψ_K, respectively, where ψ_K is the minimal polynomial, such that ψ_K vanishes on the boundary of K and attains its unit maximum at the barycenter of K and ψ_E that has support in the union of the two elements that share $E = K \cap K'$ and attains its unit maximum on the edge E it is associated with.

On the reference element $\hat{K}$ these functions are defined as

$$\begin{aligned}
\psi_{\hat{K}} &:= (1 - \xi^2)(1 - \eta^2), \\
\psi_{\hat{E}} &:= \begin{cases} \frac{(1-\hat{y})}{2}(1 - \hat{x}^2), & \hat{E} = \hat{e}_1, \\ \frac{(1+\hat{x})}{2}(1 - \hat{y}^2), & \hat{E} = \hat{e}_2, \\ \frac{(1+\hat{y})}{2}(1 - \hat{x}^2), & \hat{E} = \hat{e}_3, \\ \frac{(1-\hat{x})}{2}(1 - \hat{y}^2), & \hat{E} = \hat{e}_4. \end{cases}
\end{aligned} \tag{2.10}$$

These are then transformed to elements $K \in \mathcal{T}$ by means of the element map F_K, also cf. (2.36) and (2.37) for a representation using integrated Legendre polynomials.

Lemma 2.13. *Let $w \in \mathcal{P}_p(E)$ and $v \in \mathcal{P}_p(K)$ polynomials of degree p on E and K, respectively. Moreover, let ψ_K and ψ_E denote the element and edge bubble functions from above. Then there holds*

$$\|w\psi_E\|_{L^2(K)} \leq Ch_E^{1/2}\|w\|_{L^2(E)}, \tag{2.11}$$

$$\|\partial_x\partial_y(w\psi_E)\|_{L^2(K)} \leq C(4+2\sqrt{p(p+1)})h_E^{-3/2}\|w\|_{L^2(E)}, \tag{2.12}$$

$$\|\partial_x\partial_y(v\psi_K)\|_{L^2(K)} \leq C(32+8p(p+1))^{1/2}(4+2\sqrt{p(p+1)})h_K^{-2}\|v\|_{L^2(K)}, \tag{2.13}$$

where C depends on the shape regularity of the mesh.

Proof. All inequalities are proved by transforming to the reference element, estimating there and transforming back to the physical element. W.l.o.g. let $\hat{E}$ denote the edge of the reference element with $\eta = -1$. The other cases follow by a rotation of coordinates.

Note that $\hat{w}$ does not depend on η and $\psi_{\hat{E}} = (1-\eta)(1-\xi^2)/2$. Hence,

$$\partial_\eta(\psi_{\hat{E}}\hat{w}) = -\frac{1}{2}(1-\xi^2)\hat{w}.$$

For the first inequality, transforming to $\hat{K}$ and integrating with respect to η then yields

$$\|w\psi_E\|^2_{L^2(K)} = \frac{h_xh_y}{4}\int_{\hat{K}}(\hat{w}\psi_{\hat{E}})^2\,\mathrm{d}\xi\,\mathrm{d}\eta = \frac{h_xh_y}{6}\int_{\hat{E}}\left((1-\xi^2)\hat{w}\right)^2\,\mathrm{d}\xi \leq Ch_E\|w\|^2_{L^2(E)}.$$

Further we observe that using Lemma 2.11 there holds

$$\begin{aligned}
\|\partial_x\partial_y(w\psi_E)\|^2_{L^2(K)} =& \frac{4}{h_xh_y}\int_{\hat{K}}(\partial_\xi\partial_\eta(\hat{w}\psi_{\hat{E}}))^2\,\mathrm{d}\xi\,\mathrm{d}\eta = \frac{2}{h_xh_y}\int_{\hat{E}}\left(\partial_\xi((1-\xi^2)\hat{w})\right)^2\,\mathrm{d}\hat{s}\\
\leq& C\frac{2}{h_xh_y}(4+2\sqrt{p(p+1)})^2\|\hat{w}\|^2_{L^2(\hat{E})}\\
\leq& Ch_E^{-3}(4+2\sqrt{p(p+1)})^2\|w\|^2_{L^2(E)},
\end{aligned}$$

where in the last step we have transformed back to the physical edge E and used shape regularity to estimate $h_x \leq Ch_E$, $h_y \leq Ch_E$. Similarly, for the last inequality we obtain

$$\begin{aligned}
\|\partial_x\partial_y(v\psi_K)\|^2_{L^2(K)} =& \frac{4}{h_xh_y}\int_{\hat{K}}(\partial_\xi\partial_\eta(\hat{v}\psi_{\hat{K}}))^2\,\mathrm{d}\xi\,\mathrm{d}\eta\\
=& \frac{4}{h_xh_y}\int_{\hat{K}}\left\{\partial_\xi\left[(1-\xi^2)\partial_\eta(\hat{v}(1-\eta^2))\right]\right\}^2\,\mathrm{d}\xi\,\mathrm{d}\eta\\
\leq& \frac{8}{h_xh_y}\int_{\hat{K}}\left\{\partial_\xi(1-\xi^2)\partial_\eta(\hat{v}(1-\eta^2))\right\}^2\\
&\qquad+\left\{(1-\xi^2)\partial_\eta(\partial_\xi\hat{v}(1-\eta^2))\right\}^2\,\mathrm{d}\xi\,\mathrm{d}\eta\\
\leq& \frac{32}{h_xh_y}(4+2\sqrt{p(p+1)})^2\int_{\hat{K}}\hat{v}^2\,\mathrm{d}\xi\,\mathrm{d}\eta\\
&\qquad+\frac{8}{h_xh_y}(4+2\sqrt{p(p+1)})^2\int_{\hat{K}}\left\{(1-\xi^2)\partial_\xi\hat{v}\right\}^2\,\mathrm{d}\xi\,\mathrm{d}\eta\\
\leq& \frac{32}{h_xh_y}(4+2\sqrt{p(p+1)})^2\|\hat{v}\|^2_{L^2(\hat{K})}+\\
&\qquad+\frac{8}{h_xh_y}(4+2\sqrt{p(p+1)})^2(p(p+1))\|\hat{v}\|^2_{L^2(\hat{K})}\\
\leq& C(32+8p(p+1))(4+2\sqrt{p(p+1)})^2h_K^{-4}\|v\|^2_{L^2(K)},
\end{aligned}$$

where we have used Young's inequality and transformed back to element K. □

The following Lemma quantifies the local approximation quality of the operator $\Pi^{1,1}_{p_K}$ which is needed to show convergence of the associated Galerkin FEM. The proof uses Lemma 2.14 and standard scaling arguments. The proof is therefore omitted for brevity. For more details the reader may confer the given references.

Lemma 2.14 ([43, Lemma 5.37–5.40])**.** *Let $k \in \mathbb{N}$, $\hat{v} \in H^{k+1}(\hat{K})$, and $p_K = (p, p)$. Then there holds*

(i)
$$\begin{aligned}\|v - \Pi^{1,1}_{p_K} v\|^2_{L^2(K)} \leq & C \frac{1}{p(p+1)} \left(2 \frac{(p-s)!}{(p+s)!} \left(\frac{h_K}{2} \right)^{2s+2} \|\partial_x^{s+1} v\|^2_{L^2(K)} \right. \\ & + 4 \frac{(p-s)!}{(p+s)!} \left(\frac{h_K}{2} \right)^{2s+2} \|\partial_y^{s+1} v\|^2_{L^2(K)} \\ & \left. + \frac{4}{p(p+1)} \frac{(p-s+1)!}{(p+s-1)!} \left(\frac{h_K}{2} \right)^{2s+2} \|\partial_x \partial_y^{s} v\|^2_{L^2(K)} \right)\end{aligned}$$
for any integer $1 \leq s \leq \min\{p, k\}$.

(ii)
$$\begin{aligned}\|\partial_x \left(v - \Pi^{1,1}_{p_K} v \right)\|^2_{L^2(K)} \leq & C \left(\frac{(p-s)!}{(p+s)!} \left(\frac{h_K}{2} \right)^{2s} \|\partial_x^{s+1} v\|^2_{L^2(K)} \right. \\ & \left. + \frac{2}{p(p+1)} \frac{(p-s+1)!}{(p+s-1)!} \left(\frac{h_K}{2} \right)^{2s} \|\partial_x \partial_y^{s} v\|^2_{L^2(K)} \right)\end{aligned}$$
for any integer $1 \leq s \leq \min\{p, k\}$ and analagously for $\|\partial_\eta \left(v - \Pi^{1,1}_{p_K} v \right)\|^2_{L^2(K)}$.

(iii)
$$\begin{aligned}\|\partial_x \partial_y \left(v - \Pi^{1,1}_{p_K} v \right)\|^2_{L^2(K)} \leq & C \left(2 \frac{(p-s)!}{(p+s)!} \left(\frac{h_K}{2} \right)^{2s} \|\partial_x^{s+1} \partial_y v\|^2_{L^2(K)} \right. \\ & \left. + 2 \frac{(p-s+1)!}{(p+s-1)!} \left(\frac{h_K}{2} \right)^{2s} \|\partial_x \partial_y^{s+1} v\|^2_{L^2(K)} \right)\end{aligned}$$
for any integer $0 \leq s \leq \min\{p, k-1\}$.

(iv)
$$\begin{aligned}\|v - \Pi^{1,1}_{p_K} v\|^2_{H^{1,1}(K)} \leq & C \frac{(p-s)!}{(p+s)!} \left(\frac{h_K}{2} \right)^{2s} \left(\|\partial_x^{s+2} v\|^2_{L^2(K)} + \|\partial_y^{s+2} v\|^2_{L^2(K)} \right. \\ & \left. + \|\partial_x^{s+1} \partial_y v\|^2_{L^2(K)} + \|\partial_x \partial_y^{s+1} v\|^2_{L^2(K)} \right)\end{aligned}$$
for any integer $0 \leq s \leq p-1$.

Remark 2.15. *The previous lemma allows to give a bound of the $H^{1,1}$-norm on an element K where only derivatives of the same order are present. On the one hand unless $u \in H^{k+1}(K)$ with $k \geq 2$ we are unable to recover a power of h_K in the local error estimate for $p_K = 1$. If on the other hand however $u \in H^3(K)$, then letting $s = 1$ in Lemma 2.14 (iv) we find*

$$\begin{aligned}\|v - \Pi^{1,1}_{p_K} v\|_{H^{1,1}(K)} \leq & C h_K \left(\|\partial_x^3 v\|^2_{L^2(K)} + \|\partial_y^3 v\|^2_{L^2(K)} \right. \\ & \left. + \|\partial_x^2 \partial_y v\|^2_{L^2(K)} + \|\partial_x \partial_y^2 v\|^2_{L^2(K)} \right)^{1/2}.\end{aligned}$$

Proceeding in the usual way by defining a global interpolation operator via

$$(\Pi^{1,1}_{\mathbf{p}_\mathcal{T}} u)\big|_K := (\Pi^{1,1}_{p_K} u)\big|_K$$

we can show convergence of the FEM for $u \in H^3(\mathcal{D}) \cap H^{1,1}_0(\mathcal{D})$.

Lemma 2.16. *Let $u \in H^3(\mathcal{D}) \cap H^{1,1}_0(\mathcal{D})$. Then there holds*

$$\|u - \Pi^{1,1}_{\mathbf{p}_\mathcal{T}} u\|_{H^{1,1}(\mathcal{D})} \leq Ch|u|_{H^3(\mathcal{D})},$$

where the polynomial degree vector $\mathbf{p}_\mathcal{T}$ is set to $p_K = 1$ for every element $K \in \mathcal{T}$.

Proof. By Lemma 2.14 we find

$$\begin{aligned}\|u - \Pi^{1,1}_{\mathbf{p}_\mathcal{T}} u\|^2_{H^{1,1}(\mathcal{D})} &= \sum_{K\in\mathcal{T}} \|u - \Pi^{1,1}_{p_K} u\|^2_{H^{1,1}(K)} \\ &\leq \sum_{K\in\mathcal{T}} Ch_K^2 |u|^2_{H^3(K)} \\ &\leq Ch^2 |u|^2_{H^3(\mathcal{D})},\end{aligned}$$

where with $h = \sup_{K\in\mathcal{T}} h_K$ we have the assertion. □

Remark 2.17. *By the previous lemma we see that for $u \in H^3(\mathcal{D}) \cap H^{1,1}_0(\mathcal{D})$, since the associated bilinear form is symmetric, continuous and coercive that by Céa's lemma we have the best approximation property for the discrete Galerkin solution $u_\mathcal{T} \in \mathcal{S}^{1,0}_0(\mathcal{T})$ and therefore*

$$\begin{aligned}|u - u_\mathcal{T}|_{H^{1,1}(\mathcal{D})} &\leq c \inf_{v_\mathcal{T}\in\mathcal{S}^{1,0}_0(\mathcal{T})} |u - v_\mathcal{T}|_{H^{1,1}(\mathcal{D})} \\ &\leq c\|u - \Pi^{1,1}_{\mathbf{p}_\mathcal{T}} u\|_{H^{1,1}(\mathcal{D})} \\ &\leq Ch|u|_{H^3(\mathcal{D})}.\end{aligned}$$

For the hierarchical a posteriori error estimators we will need the following inequality, which is interesting in its own right. A proof can be found in [26], also cf. [3, Thm. 3.1].

Lemma 2.18 (strengthened Cauchy-Buniakowski-Schwarz inequality)**.** *Suppose H is a Hilbert space with inner product $\langle\cdot,\cdot\rangle_H$. Furthermore, let $V \subset H$ a finite dimensional subspace and $U \subset H$ closed. If there holds $U \cap V = \{0\}$, then there exists a constant $\gamma = \gamma(U,V) \in [0,1)$, such that for every $v \in V$ and $u \in U$*

$$|\langle u, v\rangle_H| \leq \gamma \|u\|_H \|v\|_H. \tag{2.14}$$

2.2.3 Discretization and constrained approximation in 2D

For our adaptive mesh refinement procedure we shall adopt the following rule:

An element $K \in \mathcal{T}$ may be refined if and only if all vertices of K are regular.

Otherwise, we have to refine certain irregular neighbors of K *first*, such that the rule applies and we are subsequently allowed to refine the originally marked element K. To make this precise we shall adopt the viewpoint that a hanging node is associated with an *irregular element* that has the hanging node situated on its boundary.

Algorithm 2 1-irregular mesh refinement in 2D (REFINE)

Input: Mesh $\mathcal{T}$, list M of marked elements $K \in \mathcal{T}$
Output: Refined mesh $\tilde{\mathcal{T}}$ which is 1-irregular

```
while M ≠ ∅ do
  Let K be the first element of M
  if K has hanging nodes e_i as vertex then
    Find irregular neighbors K' w.r.t. e_i
    Append all elements K' and K to M
  else
    Refine K by subdivision into 4 smaller squares K̃_s, s = 1,...4
    Remove K from M and T̃, and add four children K_s to T̃
  end if
end while
```

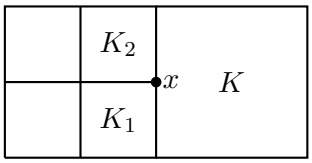

Figure 2.3: A hanging node x (•) in two dimensions. We say K is an irregular element with respect to x and an irregular neighbor of K_1 and K_2.

Definition 2.19 (Irregular Elements and Neighbors (2D))**.** *An element K is called* irregular *(with respect to a node x) if there exists a hanging node x, such that $x \in \partial K$, but x is not a vertex of K. Moreover, we shall call K an* irregular neighbor *to any element K' (with respect to x), if K' has x as a vertex (see Figure 2.3).*

The easiest treatment of hanging nodes consists in making no difference between regular and hanging nodes in the assembly procedure of the system matrix. Then in a second step, certain constraints for the hanging nodes are introduced. This is commonly known as constrained approximation, cf. e.g. [54, 56, 22, 23] to name just a very few and for the idea of system reduction after assembly see e.g. [31].

For the sake of completeness, we shall outline the procedure very briefly. Let us assume that we arrive at the linear system of equations

$$Ac = b$$

by means of a Finite Element method on a given 1-irregular mesh after assembly, ignoring at first that the global shape functions with respect to hanging nodes are non-conforming, since they are still discontinuous at this stage. The missing continuity is then recovered by enforcing constraints on the hanging nodes.

In two dimensions, for 1-irregular meshes with $p = 1$, we let the set of nodes of $\mathcal{T}$ be given in the following splitting $\mathcal{N} = \mathcal{N}_r \cup \mathcal{N}_h$, where $\mathcal{N}_r$ is the set of unconstrained and henceforth called *regular* nodes and $\mathcal{N}_h$ the set of all constrained or hanging nodes. Here we only have to deal with hanging nodes on edges. For convenience, let us define the set O_c for any hanging node in $\mathcal{N}_h$, which consists of the constraining neighboring regular nodes of c, i.e.

$$O_c := \{v \in \mathcal{N}_r : \exists e \in \mathcal{E} \text{ with } e = (v, c) \vee e = (c, v)\}.$$

In a 1-irregular mesh each hanging node $c \in \mathcal{N}_h$ is constrained by two regular vertices, i.e. the coefficient u_c of the global nodal function of the hanging node c is constrained by

the coefficients of its regular node neighbors $i, j \in O_c$, by the relation

$$u_c = \frac{1}{2}u_i + \frac{1}{2}u_j.$$

In general we can define a global connectivity matrix $P \in \mathbb{R}^{N,N}$ in the following way, where $N := |\mathcal{N}|$ denotes the total number of nodes in the mesh. If $i \in \mathcal{N}_r$, then the corresponding column in P is zero except for a 1 at the i-th position. If, on the other hand, $c \in \mathcal{N}_h$, then the corresponding column is zero and features the value $1/2$ at the positions $i, j \in \mathcal{N}_r$ of its regular node neighbors. Note that proceeding in this way that the rows corresponding to hanging nodes in P are zero. In order to reduce the system and compute the coefficients of the regular nodes of the mesh we define $P_r \in \mathbb{R}^{N_r,N}$ as the matrix which results by omitting all zero rows of P, where $N_r = |\mathcal{N}_r|$. Then we can reduce the original system of equations using P_r to

$$P_r A P_r^\top c_r = P_r b,$$

where we now only have to solve for the coefficients of regular vertices. The complete solution is then obtained by another multiplication with $P_r^\top$

$$c = P_r^\top c_r.$$

Remark 2.20. *Usually in applications the matrices P and P_r are* not *assembled explicitly, but rather local connectivity is exploited in the assembly process. However, this procedure is convenient as it provides a simple way of system reduction and does not require the restructuring of existing Finite Element codes that are readily available. This is at the expense of more work when solving the global system in this manner.*

Remark 2.21. *For higher polynomial degrees p the procedure is similar and the reader is confered to e.g. [56].*

2.2.4 A residual a posteriori error estimator

With the necessary preparations in place we can embark on the development of a *residual* a posteriori error estimator. Recall that $\mathcal{D} = D \times D = [-1, 1]^2$. Deriving an *a posteriori* residual error estimator for the deterministic second moment problem of the one dimensional model problem we consider for any $v \in H_0^{1,1}(\mathcal{D})$ the equation

$$\int_{\mathcal{D}} \partial_x\partial_y(u - u_{\mathcal{T}})\partial_x\partial_y v \,\mathrm{d}x\,\mathrm{d}y = \int_{\mathcal{D}} f v \,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{D}} \partial_x\partial_y u_{\mathcal{T}}\partial_x\partial_y v \,\mathrm{d}x\,\mathrm{d}y, \tag{2.15}$$

which implicitly defines the residual $\mathcal{R}$ with respect to $u_{\mathcal{T}}$ as an element of the dual space of $H_0^{1,1}(\mathcal{D})$ via

$$\langle \mathcal{R}, v\rangle = \int_{\mathcal{D}} f v \,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{D}} \partial_x\partial_y u_{\mathcal{T}}\partial_x\partial_y v \,\mathrm{d}x\,\mathrm{d}y, \tag{2.16}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing. The Cauchy-Schwarz inequality yields for a fixed $w \in H_0^{1,1}(\mathcal{D})$ and arbitrary $v \in H_0^{1,1}(\mathcal{D})$ with $\partial_x\partial_y v \neq 0$ that

$$\|\partial_x\partial_y w\|_{L^2(\mathcal{D})} = \sup_{v \in H_0^{1,1}(\mathcal{D})\setminus\{0\}} \frac{\langle w, v\rangle}{\|\partial_x\partial_y v\|_{L^2(\mathcal{D})}}.$$

This further implies a similar identity for the error

$$\|\partial_x\partial_y(u - u_{\mathcal{T}})\|_{L^2(\mathcal{D})} = \sup_{v \in H_0^{1,1}(\mathcal{D})\setminus\{0\}} \frac{\langle \mathcal{R}, v\rangle}{\|\partial_x\partial_y v\|_{L^2(\mathcal{D})}}.$$

Starting from the representation of the residual

$$\langle \mathcal{R}, v\rangle = \int_{\mathcal{D}} fv \,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{D}} \partial_x\partial_y u_{\mathcal{T}} \partial_x\partial_y v \,\mathrm{d}x\,\mathrm{d}y \tag{2.17}$$

we consider a mesh $\mathcal{T}$ on $\mathcal{D}$ with elements K being rectangular and of the form $K = [x_1, x_2] \times [y_1, y_2] \in \mathcal{T}$ for certain values of $x_i \equiv x_i(K), y_i \equiv y_i(K) \in [-1, 1]$, $i = 1, 2$. Noting that the operators ∂_x and ∂_y commute, integrating by parts with respect to the x-variable yields

$$\begin{aligned}
\langle \mathcal{R}, v\rangle &= \int_{\mathcal{D}} fv \,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{D}} \partial_x\partial_y u_{\mathcal{T}}\, \partial_x\partial_y v \,\mathrm{d}x\,\mathrm{d}y \\
&= \sum_{K\in\mathcal{T}} \left\{ \int_K fv - \int_{y_1}^{y_2}\int_{x_1}^{x_2} \partial_x\partial_y u_{\mathcal{T}}\, \partial_x\partial_y v \,\mathrm{d}x\,\mathrm{d}y \right\} \\
&= \sum_{K\in\mathcal{T}} \left\{ \int_K fv \,\mathrm{d}x\,\mathrm{d}y - \int_{y_1}^{y_2} \left([\partial_x\partial_y u_{\mathcal{T}}\, \partial_y v]_{x_1}^{x_2} - \int_{x_1}^{x_2} \partial_x^2\partial_y u_{\mathcal{T}} \partial_y v \,\mathrm{d}x \right) \mathrm{d}y \right\},
\end{aligned}$$

where we have used the notation

$$[f(x)g(x)]_a^b = f(b)g(b) - f(a)g(a)$$

and the one dimensional integration by parts formula

$$\int_a^b f(x)g'(x)\,\mathrm{d}x = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)\,\mathrm{d}x.$$

Using Fubini's theorem to interchange the integrals and integrating by parts, now with respect to the y-variable, we get

$$\begin{aligned}
&\sum_{K\in\mathcal{T}} \int_K fv \,\mathrm{d}x\,\mathrm{d}y - \sum_{K\in\mathcal{T}} \left(\int_{y_1}^{y_2} [\partial_x\partial_y u_{\mathcal{T}} \partial_y v]_{x_1}^{x_2} \,\mathrm{d}y - \int_{y_1}^{y_2}\int_{x_1}^{x_2} \partial_x^2\partial_y u_{\mathcal{T}} \partial_y v \,\mathrm{d}x\,\mathrm{d}y \right) \\
&= \sum_{K\in\mathcal{T}} \left\{ \int_K fv - \int_{y_1}^{y_2} [\partial_x\partial_y u_{\mathcal{T}} \partial_y v]_{x_1}^{x_2} \,\mathrm{d}y + \int_{x_1}^{x_2} \left(\left[\partial_x^2\partial_y u_{\mathcal{T}} v\right]_{y_1}^{y_2} - \int_{y_1}^{y_2} \partial_x^2\partial_y^2 u_{\mathcal{T}} v \,\mathrm{d}y \right) \mathrm{d}x \right\} \\
&= \sum_{K\in\mathcal{T}} \left\{ \int_K fv - \int_{y_1}^{y_2} \partial_x\partial_y u_{\mathcal{T}}(x_2, y)\, \partial_y v(x_2, y) - \partial_x\partial_y u_{\mathcal{T}}(x_1, y)\, \partial_y v(x_1, y) \,\mathrm{d}y \right. \\
&\qquad \left. + \int_{x_1}^{x_2} \left[\partial_x^2\partial_y u_{\mathcal{T}} v\right]_{y_1}^{y_2} \,\mathrm{d}x - \int_{x_1}^{x_2}\int_{y_1}^{y_1} \partial_x^2\partial_y^2 u_{\mathcal{T}} v \,\mathrm{d}y\,\mathrm{d}x \right\}.
\end{aligned}$$

Integrating by parts once more, since $\partial_y v$ need not have an L^2-trace on the element boundary, we find

$$\sum_{K\in\mathcal{T}}\left\{\int_K fv\,\mathrm{d}x\,\mathrm{d}y-\left(\int_{y_1}^{y_2}\partial_x\partial_y u_\mathcal{T}(x_2,y)\,\partial_y v(x_2,y)-\partial_x\partial_y u_\mathcal{T}(x_1,y)\,\partial_y v(x_1,y)\,\mathrm{d}y\right.\right.$$
$$\left.\left.-\int_{x_1}^{x_2}\left[\partial_x^2\partial_y u_\mathcal{T} v\right]_{y_1}^{y_2}\,\mathrm{d}x+\int_{x_1}^{x_2}\int_{y_1}^{y_1}\partial_x^2\partial_y^2 u_\mathcal{T} v\,\mathrm{d}x\,\mathrm{d}y\right)\right\}$$
$$=\sum_{K\in\mathcal{T}}\left\{\int_K fv\,\mathrm{d}x\,\mathrm{d}y-\left(\left[\partial_x\partial_y u_\mathcal{T}(x_2,y)v(x_2,y)\right]_{y_1}^{y_2}-\left[\partial_x\partial_y u_\mathcal{T}(x_1,y)v(x_1,y)\right]_{y_1}^{y_2}\right.\right.$$
$$-\int_{y_1}^{y_2}\partial_x\partial_y^2 u_\mathcal{T}(x_2,y)\,v(x_2,y)\,\mathrm{d}y+\int_{y_1}^{y_2}\partial_x\partial_y^2 u_\mathcal{T}(x_1,y)\,v(x_1,y)\,\mathrm{d}y$$
$$-\int_{x_1}^{x_2}\partial_x^2\partial_y u_\mathcal{T}(x,y_2)\,v(x,y_2)\,\mathrm{d}x+\int_{x_1}^{x_2}\partial_x^2\partial_y u_\mathcal{T}(x,y_1)\,v(x,y_1)\,\mathrm{d}x$$
$$\left.\left.+\int_{x_1}^{x_2}\int_{y_1}^{y_1}\partial_x^2\partial_y^2 u_\mathcal{T} v\,\mathrm{d}x\,\mathrm{d}y\right)\right\}.$$

After combining suitable terms, we arrive at the representation

$$\begin{aligned}\langle\mathcal{R},v\rangle=\sum_{K\in\mathcal{T}}\left(\int_K(f-\partial_x^2\partial_y^2 u_\mathcal{T})v\,\mathrm{d}x\,\mathrm{d}y\right.&\\+\left[\partial_x\partial_y u_\mathcal{T}(x_1,y)v(x_1,y)\right]_{y_1}^{y_2}-\left[\partial_x\partial_y u_\mathcal{T}(x_2,y)v(x_2,y)\right]_{y_1}^{y_2}&\\+\int_{y_1}^{y_2}\nabla(\partial_y^2 u_\mathcal{T}(x_2,y))\cdot\vec{\mathbf{n}}\,v(x_2,y)\,\mathrm{d}y+\int_{y_1}^{y_2}\nabla(\partial_y^2 u_\mathcal{T}(x_1,y))\cdot\vec{\mathbf{n}}\,v(x_1,y)\,\mathrm{d}y&\\\left.+\int_{x_1}^{x_2}\nabla(\partial_x^2 u_\mathcal{T}(x,y_2))\cdot\vec{\mathbf{n}}\,v(x,y_2)\,\mathrm{d}x+\int_{x_1}^{x_2}\nabla(\partial_x^2 u_\mathcal{T}(x,y_1))\cdot\vec{\mathbf{n}}\,v(x,y_1)\,\mathrm{d}x\right),&\end{aligned}\tag{2.18}$$

where $\vec{\mathbf{n}}$ denotes the outer normal vector to K on the current edge.

Using the notation for interelemental jumps (1.9) the representation of the residual (2.18) can be rewritten in the form

$$\begin{aligned}\langle\mathcal{R},v\rangle=\sum_{K\in\mathcal{T}}\int_K rv\,\mathrm{d}x\,\mathrm{d}y+\sum_{E\in\mathcal{E}}\int_E jv\,\mathrm{d}s&\\+\sum_{K\in\mathcal{T}}\left(\left[\partial_x\partial_y u_\mathcal{T}(x_1,y)v(x_1,y)\right]_{y_1}^{y_2}-\left[\partial_x\partial_y u_\mathcal{T}(x_2,y)v(x_2,y)\right]_{y_1}^{y_2}\right),&\end{aligned}\tag{2.19}$$

where

$$r|_K = f - \partial_x^2 \partial_y^2 u_{\mathcal{T}},$$

$$j|_E = \begin{cases} [\![\nabla(\partial_y^2 u_{\mathcal{T}}) \cdot n]\!], & \text{for a vertical edge } E \subset K \in \mathcal{T}, E \not\subset \partial\mathcal{D}, \\ [\![\nabla(\partial_x^2 u_{\mathcal{T}}) \cdot n]\!], & \text{for a horizontal edge } E \subset K \in \mathcal{T}, E \not\subset \partial\mathcal{D}, \\ 0, & E \subset \partial\mathcal{D}. \end{cases}$$

In the following we show that the above representation of the residual does in fact allow us to define an *a posteriori* error estimator.

An upper bound on the error

In this section we concern ourselves with proving that (2.19) guarantees an upper bound on the error and thus a *reliable* residual a posteriori error estimator can be defined. From (2.19) we draw motivation for the following definition.

Definition 2.22. *For the deterministic second moment problem (2.6) we define the residual a posteriori error estimator $\eta_{\mathcal{R},K}$ element-wise by*

$$\eta_{\mathcal{R},K}^2 := h_K^2 \|r\|_{L^2(K)}^2 + \frac{1}{2} \sum_{E \subset \partial K} h_E \|j\|_{L^2(E)}^2. \tag{2.20}$$

The global error estimator $\eta_{\mathcal{R}}$ is then given as

$$\eta_{\mathcal{R}} := \left(\sum_{K \in \mathcal{T}} \eta_{\mathcal{R},K}^2 \right)^{1/2}. \tag{2.21}$$

That this definition makes sense is shown in the following lemma.

Lemma 2.23 (Reliability of $\eta_{\mathcal{R}}$)**.** *Let $\mathcal{D} = [-1,1]^2$. Furthermore, u be the exact solution of (2.6) and $u_{\mathcal{T}}$ the solution of the corresponding discrete variational formulation (2.7). Then there exists a positive constant c^* which only depends on the shape regularity of the mesh $\mathcal{T}$, the polynomial degree $\mathbf{p}_{\mathcal{T}}$, and $\mathcal{D}$, such that*

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq c^* \left\{ \sum_{K \in \mathcal{T}} h_K^2 \|r\|_{L^2(K)}^2 + \sum_{E \in \mathcal{E}} h_E \|j\|_{L^2(E)}^2 \right\}^{1/2}. \tag{2.22}$$

Proof. In order to show that the error $|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})}$ is bounded from above globally by $\eta_{\mathcal{R},K}$, we start out from the expression (2.19) and insert the function $v - \Pi_{\mathbf{p}_{\mathcal{T}}}^{1,1} v$ instead of v, where we assume $p_K = (p,p)$ for all K. Thus, by the properties of $\Pi_{p_K}^{1,1}$ the point evaluations in the representation (2.19) vanish, since $\Pi_{p_K}^{1,1} v$ interpolates v at the nodes of K. By use of the Cauchy-Schwarz inequality for sums and integrals in conjunction with approximation results from the previous section and Galerkin orthogonality, we obtain

for any $v \in H_0^{1,1}(\mathcal{D})$ that there holds

$$\begin{aligned}
\langle \mathcal{R}, v\rangle =& \langle \mathcal{R}, v - \Pi^{1,1}_{\mathbf{p}_\mathcal{T}} v\rangle \\
=& \sum_{K\in\mathcal{T}} \int_K r(v - \Pi^{1,1}_{p_K} v)\,\mathrm{d}x\,\mathrm{d}y + \sum_{E\in\mathcal{E}} \int_E j(v - \Pi^{1,1}_{p_K} v)\,\mathrm{d}s \\
\leq& \sum_{K\in\mathcal{T}} \|r\|_{L^2(K)} \|v - \Pi^{1,1}_{p_K} v\|_{L^2(K)} + \sum_{E\in\mathcal{E}} \|j\|_{L^2(E)} \|v - \Pi^{1,1}_{p_K} v\|_{L^2(E)} \\
\leq& \sum_{K\in\mathcal{T}} \|r\|_{L^2(K)} C_1 \frac{h_K}{\sqrt{p(p+1)}} \left(\|\partial_x v\|^2_{L^2(K)} + \|\partial_y v\|^2_{L^2(K)} + |v|^2_{H^{1,1}(K)} \right)^{1/2} \\
&+ \sum_{E\in\mathcal{E}} \|j\|_{L^2(E)} \left(C_2 \frac{h_E}{p(p+1)} \|\partial_\tau v\|^2_{L^2(K)} + \frac{h_E^3}{p(p+1)} \|\partial_\xi \partial_\eta v\|^2_{L^2(K)} \right)^{1/2} \\
\leq& \max\{C_1, C_2\} \left\{ \sum_{K\in\mathcal{T}} \frac{h_K^2}{p(p+1)} \|r\|^2_{L^2(K)} + \sum_{E\in\mathcal{E}} \frac{h_E}{p(p+1)} \|j\|^2_{L^2(E)} \right\}^{1/2} \\
&\times \left\{ \sum_{K\in\mathcal{T}} \left(\|\partial_x v\|^2_{L^2(K)} + \|\partial_y v\|^2_{L^2(K)} + |v|^2_{H^{1,1}(K)} \right) \right. \\
&\qquad \left. + \sum_{E\in\mathcal{E}} \sum_{K\subset\omega_E} \left(\|\partial_\tau v\|^2_{L^2(K)} + |v|^2_{H^{1,1}(K)} \right) \right\}^{1/2},
\end{aligned}$$

where in the last step the Cauchy-Schwarz inequality for sums has been used and we recall that ω_E denotes the patch of elements K that have E as a common edge. Furthermore, we have made use of (2.9) for ∂_τ. Since the Lebesgue integral is additive we have

$$\sum_{K\in\mathcal{T}} \|\partial_x v\|^2_{L^2(K)} = \|\partial_x v\|^2_{L^2(\mathcal{D})}, \qquad \sum_{K\in\mathcal{T}} \|\partial_y v\|^2_{L^2(K)} = \|\partial_y v\|^2_{L^2(\mathcal{D})},$$

from which we infer by applying Friedrichs' inequality in the y- or x-direction, respectively, that

$$\|\partial_x v\|^2_{L^2(\mathcal{D})} \leq C_{F,y} \|\partial_x \partial_y v\|^2_{L^2(\mathcal{D})}, \qquad \|\partial_y v\|^2_{L^2(\mathcal{D})} \leq C_{F,x} \|\partial_x \partial_y v\|^2_{L^2(\mathcal{D})}.$$

Moreover,

$$\sum_{E\in\mathcal{E}} \sum_{K\subset\omega_E} \left(\|\partial_\tau v\|^2_{L^2(K)} + |v|^2_{H^{1,1}(K)} \right) \leq c |v|^2_{H^{1,1}(\mathcal{D})},$$

by additivity of the Lebesgue integral, the fact that every element is counted only finitely many times in the second sum on the left-hand side by shape regularity, and that as earlier we can bound $\sum_{E\in\mathcal{E}} \sum_{K\subset\omega_E} \|\partial_\tau v\|^2_{L^2(K)} \leq C \|\partial_x \partial_y v\|^2_{L^2(\mathcal{D})}$. This implies

$$\begin{aligned}
&\left\{ \sum_{K\in\mathcal{T}} \left(\|\partial_x v\|^2_{L^2(K)} + \|\partial_y v\|^2_{L^2(K)} + |v|^2_{H^{1,1}(K)} \right) \right. \\
&\qquad \left. + \sum_{E\in\mathcal{E}} \sum_{K\subset\omega_E} \left(\|\partial_\tau v\|^2_{L^2(K)} + |v|^2_{H^{1,1}(K)} \right) \right\}^{1/2} \leq \tilde{c} |v|_{H^{1,1}(\mathcal{D})},
\end{aligned}$$

where $\tilde{c}$ only depends on the shape regularity of the mesh $\mathcal{T}$, takes into consideration that elements are counted multiple times on the left-hand side and incorporates the constants of Friedrichs' inequality in the x- and y-direction. Thus,

$$\langle \mathcal{R}, v \rangle \leq C(p(p+1))^{-1} |v|_{H^{1,1}(\mathcal{D})} \left\{ \sum_{K \in \mathcal{T}} h_K^2 \|r\|_{L^2(K)}^2 + \sum_{E \in \mathcal{E}} h_E \|j\|_{L^2(E)}^2 \right\}^{1/2}$$

with $C = \tilde{c} \max\{C_1, C_2\}$. Since $v \in H_0^{1,1}(\mathcal{D})$ was arbitrary, it follows that there exists a constant $c^* > 0$, such that we have

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq c^* \left\{ \sum_{K \in \mathcal{T}} h_K^2 \|r\|_{L^2(K)}^2 + \sum_{E \in \mathcal{E}} h_E \|j\|_{L^2(E)}^2 \right\}^{1/2}.$$

□

Remark 2.24. *In general evaluating the integrals on the right-hand side of (2.22) for arbitrary functions f might be prohibitively expensive. Therefore integrals are usually evaluated by suitable quadrature formulae. Alternatively, the function f may be approximated by polynomials in an element-wise fashion, thus enabling exact evaluation.*

As a corollary of the previous lemma we find the following, if we assume that f is replaced by an approximation $f_{\mathcal{T}}$.

Corollary 2.25. *Let u be the exact solution of (2.6) and $u_{\mathcal{T}}$ the solution of the corresponding discrete variational formulation (2.7). Then there exists a constant c^* which only depends on the shape regularity of the mesh $\mathcal{T}$, the polynomial degree $\mathbf{p}_{\mathcal{T}}$, and $\mathcal{D}$, such that*

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq c^* \left(\eta_{\mathcal{R}}^2 + \mathsf{osc}_f \right)^{1/2}. \tag{2.23}$$

where $\mathsf{osc}_f := \sum_{K \in \mathcal{T}} h_K^2 \|f - f_{\mathcal{T}}\|_{L^2(K)}^2$

Proof. Splitting the integral of the element residual into

$$\begin{aligned} &\sum_{K \in \mathcal{T}} \int_K r(v - \Pi_{p_K}^{1,1} v) \, \mathrm{d}x \, \mathrm{d}y \\ &= \sum_{K \in \mathcal{T}} \int_K \tilde{r}(v - \Pi_{p_K}^{1,1} v) \, \mathrm{d}x \, \mathrm{d}y + \sum_{K \in \mathcal{T}} \int_K (f - f_{\mathcal{T}})(v - \Pi_{p_K}^{1,1} v) \, \mathrm{d}x \, \mathrm{d}y \end{aligned}$$

with $\tilde{r} = f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}}$ and then repeating the steps of the proof of Lemma 2.23 gives the assertion. □

Remark 2.26. *Note that the factor $\frac{1}{2}$ in front of the edge residuals takes care of the fact that each inner edge is counted twice when summing over all $\eta_{\mathcal{R},K}^2$ to get $\eta_{\mathcal{R}}$.*

Remark 2.27. *By the definition of $\eta_{\mathcal{R},K}$ we see that for $p = 1$ that $\eta_{\mathcal{R},K}$ only depends on the right-hand side f as the other terms vanish. Because of this fact we will consider different extensions of the local Finite Element space, when we examine the numerical experiments.*

A lower bound on the error

We will now proceed to show a lower bound on the error, which is done in an element-wise fashion. Our focus is on bounding the element and edge resduals. Let us start by fixing an element K and write

$$f_{\mathcal{T}}|_K = \mathbf{P} f|_K$$

where $\mathbf{P}$ is a suitable projection of f into the finite element space $\mathcal{S}^{p,-1}(\mathcal{T})$. For the sake of simplicity let us define $f_{\mathcal{T}}|_K$ to be the L^2-projection on $\mathcal{P}_0(K)$.

Moreover, let $w_{\mathcal{T}}$ be the function that restricted to K has the form

$$w_{\mathcal{T}} = (f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}})\psi_K,$$

where ψ_K denotes the element bubble function, i.e. ψ_K is the minimal degree polynomial that is zero on the boundary of K and attains its unit maximum in the barycenter of K (cf. (2.10)).

Lemma 2.28 (Element residuals)**.** *Let $f_{\mathcal{T}} \in \mathcal{S}^{p,-1}(\mathcal{T})$ any approximation of the right-hand side f. Let u be the exact solution of (2.6) and $u_{\mathcal{T}}$ the exact solution of the discrete problem (2.7). Then for any element $K \in \mathcal{T}$, there exist constants C_1 and C_2 such that there holds the estimate*

$$h_K^2 \|f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}}\|_{L^2(K)} \leq C_1 \|\partial_x \partial_y (u - u_{\mathcal{T}})\|_{L^2(K)} + C_2 h_K^2 \|f_{\mathcal{T}} - f\|_{L^2(K)}. \tag{2.24}$$

Proof. With notations as above we have,

$$\begin{aligned}
\int_K (f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}})^2 \psi_K \,\mathrm{d}x\,\mathrm{d}y &= \int_K (f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}}) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\
&= \int_K r w_{\mathcal{T}} \,\mathrm{d}\vec{x} + \int_K (f_{\mathcal{T}} - f) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\
&= \int_K \partial_x \partial_y (u - u_{\mathcal{T}}) \partial_x \partial_y w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y + \int_K (f_{\mathcal{T}} - f) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y,
\end{aligned} \tag{2.25}$$

where we note that, since ψ_K is zero on the boundary of K, the edge residuals as well as the evaluation terms in (2.19) in the nodes of K vanish. There holds

$$\int_K (f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}})^2 \psi_K \,\mathrm{d}x\,\mathrm{d}y \geq (p+2)^{-4} \|f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}}\|_{L^2(K)}^2,$$

which is due to Lemma 2.11. Estimating the right hand side of (2.25) by means of approximation results from section 2.2.2 yields

$$\begin{aligned}
\int_K \partial_x \partial_y (u - u_{\mathcal{T}}) \partial_x \partial_y w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y &\leq \|\partial_x \partial_y (u - u_{\mathcal{T}})\|_{L^2(K)} \|\partial_x \partial_y w_{\mathcal{T}}\|_{L^2(K)} \\
&\leq \|\partial_x \partial_y (u - u_{\mathcal{T}})\|_{L^2(K)} C h_K^{-2} \|f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}}\|_{L^2(K)}, \\
\int_K (f_{\mathcal{T}} - f) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y &\leq \|f_{\mathcal{T}} - f\|_{L^2(K)} \|w_{\mathcal{T}}\|_{L^2(K)} \\
&\leq \|f_{\mathcal{T}} - f\|_{L^2(K)} \|f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}}\|_{L^2(K)}.
\end{aligned}$$

Combining these estimates we find the assertion

$$h_K^2 \|f_{\mathcal{T}} - \partial_x^2 \partial_y^2 u_{\mathcal{T}}\|_{L^2(K)} \leq C_1 \|\partial_x \partial_y (u - u_{\mathcal{T}})\|_{L^2(K)} + C_2 h_K^2 \|f_{\mathcal{T}} - f\|_{L^2(K)}. \tag{2.26}$$

□

We now turn to estimating the edge residuals. Consider an edge $E \in \mathcal{E}$ and insert the function

$$w_E = j\psi_E$$

into (2.19), where ψ_E is the minimal polynomial such that ψ_E attains its unit maximum in the barycenter of $E = K \cap K'$ and is zero on all other edges E' of K and K'. The support of ψ_E is $\omega_E = K \cup K'$.

Lemma 2.29 (Edge residuals)**.** *With the notation as above and u the exact solution of (2.6) and $u_\mathcal{T}$ the solution of the corresponding discrete variational formulation (2.7). Then there exist constants C_1 and C_2, such that*

$$h_E^{3/2}\|j\|_{L^2(E)} \leq C_1\|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_E)} + C_2 \sum_{K\subset\omega_E} h_K^2\|f_\mathcal{T} - f\|_{L^2(K)}. \tag{2.27}$$

Proof. Inserting w_E into representation of the residual yields

$$\begin{aligned}\int_E (j^2\psi_E)\,\mathrm{d}s &= \int_E jw_E\,\mathrm{d}s \\ &= \int_{\omega_E} \partial_x\partial_y(u-u_\mathcal{T})\partial_x\partial_y w_E\,\mathrm{d}x\,\mathrm{d}y - \int_{\omega_E}\tilde{r}w_E\,\mathrm{d}x\,\mathrm{d}y - \int_{\omega_E}(f-f_\mathcal{T})w_E\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Then by Lemma 2.11 there exists a constant such that the left-hand side is bounded from below as

$$\int_E jw_E\,\mathrm{d}s \geq (p+2)^{-2}\|j\|_{L^2(E)}^2.$$

Now we bound the terms on the right-hand side using estimates from section 2.2.2 and find

$$\begin{aligned}\int_{\omega_E} \partial_x\partial_y(u-u_\mathcal{T})\partial_x\partial_y w_E\,\mathrm{d}x\,\mathrm{d}y &\leq \|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_E)}\|\partial_x\partial_y w_E\|_{L^2(\omega_E)} \\ &\leq \|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_E)} C h_E^{-3/2}\|j\|_{L^2(E)},\end{aligned}$$

$$\begin{aligned}\sum_{K\subset\omega_E}\int_K \tilde{r}w_E\,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{K\subset\omega_E}\|\tilde{r}\|_{L^2(K)}\|w_E\|_{L^2(K)} \\ &\leq \sum_{K\subset\omega_E}\|\tilde{r}\|_{L^2(K)} C h_E^{1/2}\|j\|_{L^2(E)},\end{aligned}$$

$$\begin{aligned}\sum_{K\subset\omega_E}\int_K (f-f_\mathcal{T})w_E\,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{K\subset\omega_E}\|(f-f_\mathcal{T})\|_{L^2(K)}\|w_E\|_{L^2(K)} \\ &\leq \sum_{K\subset\omega_E}\|(f-f_\mathcal{T})\|_{L^2(K)} C h_E^{1/2}\|j\|_{L^2(E)}.\end{aligned}$$

Combining the aforementioned estimates we get

$$\begin{aligned}C\|j\|_{L^2(E)}^2 \leq{}& \|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_E)} C h_E^{-3/2}\|j\|_{L^2(E)} + \sum_{K\subset\omega_E}\|\tilde{r}\|_{L^2(K)} C h_E^{1/2}\|j\|_{L^2(E)} \\ &+ \sum_{K\subset\omega_E}\|(f-f_\mathcal{T})\|_{L^2(K)} C h_E^{1/2}\|j\|_{L^2(E)}\end{aligned}$$

and further upon dividing the above inequality by $\|j\|_{L^2(E)}$

$$\begin{aligned} C\|j\|_{L^2(E)} \leq Ch_E^{-3/2}\|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_E)} \\ + \sum_{K\subset\omega_E} Ch_E^{1/2}\|\tilde{r}\|_{L^2(K)} + \sum_{K\subset\omega_E} Ch_E^{1/2}\|(f-f_\mathcal{T})\|_{L^2(K)}. \end{aligned}$$

Furthermore, inserting the bound for $\|\tilde{r}\|_{L^2(K)}$ from (2.26) yields

$$\begin{aligned} h_E^{3/2}\|j\|_{L^2(E)} &\leq C\|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_E)} \\ &\quad + \sum_{K\subset\omega_E} Ch_E^2\|\tilde{r}\|_{L^2(K)} + \sum_{K\subset\omega_E} Ch_E^2\|(f-f_\mathcal{T})\|_{L^2(K)} \\ &\leq C\|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_E)} + C\sum_{K\subset\omega_E} h_K^2\|f-f_\mathcal{T}\|_{L^2(K)}. \end{aligned}$$

□

This concludes the inspection of a lower bound of the residual error estimator $\eta_\mathcal{R}$. In total we have shown the

Theorem 2.30. *Let u and $u_\mathcal{T}$ denote the solutions of the variational problems (2.6) and (2.7), respectively. Let the residual error estimator $\eta_{\mathcal{R},K}$ be given as in Definition 2.22. Moreover, let $f_\mathcal{T}$ denote an approximation of f on the mesh $\mathcal{T}$. There exists constants c^* and c_* that only depend on the shape regularity of the given mesh, the polynomial degree $\mathbf{p}_\mathcal{T}$, and the domain $\mathcal{D}$, such that*

$$|u-u_\mathcal{T}|_{H^{1,1}(\mathcal{D})} \leq c^*\left\{\sum_{K\in\mathcal{T}}\eta_{\mathcal{R},K}^2 + \sum_{K\in\mathcal{T}} h_K^2\|f-f_\mathcal{T}\|_{L^2(K)}^2\right\}^{1/2} \tag{2.28}$$

and for all $K\in\mathcal{T}$ there holds

$$h_K\eta_{\mathcal{R},K} \leq c_*\left\{|u-u_\mathcal{T}|_{H^{1,1}(\omega_K)}^2 + \sum_{K'\subset\omega_K} h_K^4\|f-f_\mathcal{T}\|_{L^2(K')}^2\right\}^{1/2}, \tag{2.29}$$

where ω_K denotes the union of all elements K' that share an edge with K, i.e. $\{K'\in\mathcal{T} : \overline{K}\cap\overline{K'} = E, E\in\mathcal{E}\}$.

Proof. The upper bound is proven in Lemma 2.23 and Corollary 2.25. The lower bound follows from the combination of Lemma 2.28 and Lemma 2.29 with the shape regularity of the mesh $\mathcal{T}$, which immediately yields

$$\begin{aligned} h_K^2\eta_{\mathcal{R},K}^2 &\leq c\left(h_K^4\|f_\mathcal{T} - \partial_x^2\partial_y^2 u_\mathcal{T}\|_{L^2(K)}^2 + \sum_{e\in\mathcal{E}_K} h_E^3\|j\|_{L^2(E)}^2\right) \\ &\leq C\left(\|\partial_x\partial_y(u-u_\mathcal{T})\|_{L^2(\omega_K)}^2 + \sum_{K'\subset\omega_K} h_K^4\|f_\mathcal{T}-f\|_{L^2(K')}^2\right) \end{aligned}$$

and upon taking the square root concludes the proof. □

Remark 2.31. *Note that there is gap of h between the upper and lower bounds of the error. This is due to the fact that the a posteriori estimator has to take care of the worst case scenarios in the upper and lower bound of the approximation on $\mathcal{D}$. Since on the one hand it might happen, that u factors as $g(x)h(y)$ on an element K, in which case the error analysis yields an extra power of h_K. If on the other hand u does not allow a factorization on K, i.e. $u \neq g(x)h(y)$ with appropriate functions g and h, then we only recover a single power of h_K in the error estimate.*

Motivated by the preceding theorem and to have a short terminology we state the following definition.

Definition 2.32 (Weak Efficiency)**.** *Let η be an error estimator. Then we call η a* weakly efficient *estimator if there exists a constant C_{eff} and there holds an upper bound of the form*

$$h_K \eta \leq C_{\mathit{weff}} \|u - u_{\mathcal{T}}\| + H_{\mathit{weff}}$$

for a certain norm of the error $\|\cdot\|$ and such that the generic higher order error term H_{weff} satisfies $H_{\mathit{weff}} = o(\|u - u_{\mathcal{T}}\|)$.

Remark 2.33. *The lower bound of $\eta_{\mathcal{R}}$ can lead to a deterioration of the convergence of the adaptive process. This circumstance is investigated in a later section.*

2.2.5 A hierarchical a posteriori error estimator

The presentation here follows the ideas of [54], also see [27]. Let $Y_{\mathcal{T}} = \mathcal{S}_0^{1,0}(\mathcal{T})$ and consider a finite dimensional Finite Element space $X_{\mathcal{T}}$ for which holds

$$Y_{\mathcal{T}} \subset X_{\mathcal{T}} \subset H_0^{1,1}(\mathcal{D}).$$

The space $X_{\mathcal{T}}$ may be induced by a uniform refinement of the mesh $\mathcal{T}$ or consist of higher order elements. We shall adopt the latter idea. Furthermore, let us denote by $x_{\mathcal{T}} \in X_{\mathcal{T}}$ the solution of

$$\int_{\mathcal{D}} \partial_x\partial_y x_{\mathcal{T}} \partial_x\partial_y v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y = \int_{\mathcal{D}} f v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y, \quad \forall v_{\mathcal{T}} \in X_{\mathcal{T}}.$$

In order to compare the solution $u_{\mathcal{T}} \in \mathcal{S}_0^{1,0}(\mathcal{T})$ of (2.7) with $x_{\mathcal{T}}$, we subtract

$$\int_{\mathcal{D}} \partial_x\partial_y u_{\mathcal{T}} \partial_x\partial_y v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y$$

from both sides of the equation characterizing $x_{\mathcal{T}}$, yielding for all $v_{\mathcal{T}} \in X_{\mathcal{T}}$

$$\begin{aligned} \int_{\mathcal{D}} \partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}}) \partial_x\partial_y v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y &= \int_{\mathcal{D}} f v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{D}} \partial_x\partial_y u_{\mathcal{T}} \partial_x\partial_y v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathcal{D}} \partial_x\partial_y (u - u_{\mathcal{T}}) \partial_x\partial_y v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y, \end{aligned} \tag{2.30}$$

where $u \in H_0^{1,1}(\mathcal{D})$ denotes the unique solution of the variational formulation (2.6). As $\mathcal{S}_0^{1,0}(\mathcal{T}) \subset X_{\mathcal{T}}$, we may write $v_{\mathcal{T}} = x_{\mathcal{T}} - u_{\mathcal{T}}$ and thus, by the Cauchy-Schwarz inequality, we have

$$\|\partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \|\partial_x\partial_y (u - u_{\mathcal{T}})\|_{L^2(\mathcal{D})}.$$

In order to prove the converse estimate, we introduce a condition on the space $X_{\mathcal{T}}$.

Definition 2.34 (Saturation Property)**.** *The space $X_\mathcal{T}$ is said to satisfy a* saturation property *(with respect to a subspace, here: $Y_\mathcal{T} = \mathcal{S}_0^{1,0}(\mathcal{T})$), if there exists $\beta \in [0,1)$ such that*

$$\|\partial_x\partial_y(u - x_\mathcal{T})\|_{L^2(\mathcal{D})} \leq \beta\|\partial_x\partial_y(u - u_\mathcal{T})\|_{L^2(\mathcal{D})}.$$

If the larger space $X_\mathcal{T}$ provides approximations of higher order, it seems intuitive that one could expect that $u_\mathcal{T} \neq x_\mathcal{T}$ and that $x_\mathcal{T}$ is a better approximation of u as long as u is sufficiently regular. In [10, Prop. 2.6] the authors have shown that there are non-trivial right-hand sides f, such that $u_\mathcal{T} = x_\mathcal{T}$ for the Poisson problem. The proof is also valid in our situation for the splitting $X_\mathcal{T} = \mathcal{S}_0^{1,0}(\mathcal{T}) \oplus Z_\mathcal{T}$. Although if the mesh width is small enough and as in our situation we have a direct hierarchical extension, the saturation assumption is clearly satisfied, if the solution is regular enough. Moreover, in [15] the authors have shown that in most cases the saturation assumption holds except for a few pathological examples where the starting partition $\mathcal{T}_0$ only has one internal degree of freedom. For the hierarchical error estimator we will derive another proof, but the saturation assumption will also play a role when we have a look at the a posteriori error estimator that is built via an averaging procedure.

Now if $X_\mathcal{T}$ satisfies a saturation property, we conclude with the triangle inequality that

$$\begin{aligned}\|\partial_x\partial_y(u - u_\mathcal{T})\|_{L^2(\mathcal{D})} &\leq \|\partial_x\partial_y(u - x_\mathcal{T})\|_{L^2(\mathcal{D})} + \|\partial_x\partial_y(x_\mathcal{T} - u_\mathcal{T})\|_{L^2(\mathcal{D})} \\ &\leq \beta\|\partial_x\partial_y(u - u_\mathcal{T})\|_{L^2(\mathcal{D})} + \|\partial_x\partial_y(x_\mathcal{T} - u_\mathcal{T})\|_{L^2(\mathcal{D})}\end{aligned}$$

and hence

$$\|\partial_x\partial_y(u - u_\mathcal{T})\|_{L^2(\mathcal{D})} \leq \frac{1}{1-\beta}\|\partial_x\partial_y(x_\mathcal{T} - u_\mathcal{T})\|_{L^2(\mathcal{D})}.$$

Overall we have the two-sided bound

$$\|\partial_x\partial_y(x_\mathcal{T} - u_\mathcal{T})\|_{L^2(\mathcal{D})} \leq \|\partial_x\partial_y(u - u_\mathcal{T})\|_{L^2(\mathcal{D})} \leq \frac{1}{1-\beta}\|\partial_x\partial_y(x_\mathcal{T} - u_\mathcal{T})\|_{L^2(\mathcal{D})}.$$

This shows that we may use $\|\partial_x\partial_y(x_\mathcal{T} - u_\mathcal{T})\|_{L^2(\mathcal{D})}$ as an a posteriori error indicator. However, since the computation of the solution $x_\mathcal{T}$ is at least as costly as $u_\mathcal{T}$, this approach is not at all cost-efficient.

In order to remedy the situation, we consider the space $X_\mathcal{T}$ to admit a hierarchical splitting in the form

$$X_\mathcal{T} = \mathcal{S}_0^{1,0}(\mathcal{T}) \oplus Z_\mathcal{T}.$$

If now the spaces $\mathcal{S}_0^{1,0}(\mathcal{T})$ and $Z_\mathcal{T}$ satisfy a strengthened Cauchy-Schwarz inequality (cf. 2.14), i.e. the spaces are in a sense (nearly) orthogonal, we can exploit this fact to formulate a more efficient tool. In this context the aforementioned inequality takes the form

$$\left|\int_\mathcal{D} \partial_x\partial_y z_\mathcal{T}\partial_x\partial_y u_\mathcal{T}\,\mathrm{d}x\,\mathrm{d}y\right| \leq \gamma\|\partial_x\partial_y z_\mathcal{T}\|_{L^2(\mathcal{D})}\|\partial_x\partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})},$$

where $z_\mathcal{T} \in Z_\mathcal{T}$ and $u_\mathcal{T} \in \mathcal{S}_0^{1,0}(\mathcal{T})$. Hence, the idea is to replace $\|\partial_x\partial_y(x_\mathcal{T} - u_\mathcal{T})\|_{L^2(\mathcal{D})}$ by $\|\partial_x\partial_y z_\mathcal{T}\|_{L^2(\mathcal{D})}$ with a suitable $z_\mathcal{T} \in Z_\mathcal{T}$, which is hopefully easier and foremost less costly to compute than $x_\mathcal{T}$. To this end let $z_\mathcal{T}$ be defined as the unique solution in $Z_\mathcal{T}$ of the defect problem

$$\int_\mathcal{D} \partial_x\partial_y z_\mathcal{T}\partial_x\partial_y\zeta_\mathcal{T}\,\mathrm{d}x\,\mathrm{d}y = \int_\mathcal{D} f\zeta_\mathcal{T}\,\mathrm{d}x\,\mathrm{d}y - \int_\mathcal{D} \partial_x\partial_y u_\mathcal{T}\partial_x\partial_y\zeta_\mathcal{T}\,\mathrm{d}x\,\mathrm{d}y \tag{2.31}$$

for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$. By the preceding considerations (cf. (2.30),(2.15)) we note for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$ the identity

$$\begin{aligned}\int_{\mathcal{D}} \partial_x\partial_y z_{\mathcal{T}} \partial_x\partial_y \zeta_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y &= \int_{\mathcal{D}} \partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}}) \partial_x\partial_y \zeta_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathcal{D}} \partial_x\partial_y (u - u_{\mathcal{T}}) \partial_x\partial_y \zeta_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y.\end{aligned} \tag{2.32}$$

In particular, for $x_{\mathcal{T}}$ we have

$$\int_{\mathcal{D}} \partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}}) \partial_x\partial_y v_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y = 0, \quad \forall v_{\mathcal{T}} \in \mathcal{S}_0^{1,0}(\mathcal{T}).$$

Upon inserting $\zeta_{\mathcal{T}} = z_{\mathcal{T}}$ into (2.32) and applying the Cauchy-Schwarz inequality, we get

$$\|\partial_x\partial_y z_{\mathcal{T}}\|_{L^2(\mathcal{D})} \leq \|\partial_x\partial_y (u - u_{\mathcal{T}})\|_{L^2(\mathcal{D})}.$$

Furthermore, writing $x_{\mathcal{T}} - u_{\mathcal{T}} = \tilde{v}_{\mathcal{T}} + \tilde{z}_{\mathcal{T}}$ with $\tilde{v}_{\mathcal{T}} \in \mathcal{S}_0^{p,0}(\mathcal{T})$ and $\tilde{z}_{\mathcal{T}} \in Z_{\mathcal{T}}$, we have by the strengthened Cauchy-Schwarz inequality and Young's inequality $ab \leq \frac{1}{2}(a^2 + b^2)$ that

$$\begin{aligned}\|\partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}})\|^2_{L^2(\mathcal{D})} =& \|\partial_x\partial_y (\tilde{v}_{\mathcal{T}} + \tilde{z}_{\mathcal{T}})\|^2_{L^2(\mathcal{D})} \\ \leq& \|\partial_x\partial_y \tilde{v}_{\mathcal{T}}\|^2_{L^2(\mathcal{D})} + 2\gamma \|\partial_x\partial_y \tilde{v}_{\mathcal{T}}\|_{L^2(\mathcal{D})} \|\partial_x\partial_y \tilde{z}_{\mathcal{T}}\|_{L^2(\mathcal{D})} \\ &+ \|\partial_x\partial_y \tilde{z}_{\mathcal{T}}\|^2_{L^2(\mathcal{D})} \\ \leq& (1+\gamma)(\|\partial_x\partial_y \tilde{v}_{\mathcal{T}}\|^2_{L^2(\mathcal{D})} + \|\partial_x\partial_y \tilde{z}_{\mathcal{T}}\|^2_{L^2(\mathcal{D})})\end{aligned}$$

and similarly

$$(1-\gamma)\left(\|\partial_x\partial_y \tilde{v}_{\mathcal{T}}\|^2_{L^2(\mathcal{D})} + \|\partial_x\partial_y \tilde{z}_{\mathcal{T}}\|^2_{L^2(\mathcal{D})}\right) \leq \|\partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}})\|^2_{L^2(\mathcal{D})}.$$

This further implies that we also have

$$\|\partial_x\partial_y \tilde{z}_{\mathcal{T}}\|_{L^2(\mathcal{D})} \leq \frac{1}{(1-\gamma)^{1/2}} \|\partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}})\|_{L^2(\mathcal{D})}.$$

Exploiting the Galerkin orthogonality in combination with (2.32), we conclude with $\zeta_{\mathcal{T}} = \tilde{z}_{\mathcal{T}}$ that

$$\begin{aligned}\|\partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}})\|^2_{L^2(\mathcal{D})} &= \int_{\mathcal{D}} \partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}}) \partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}}) \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathcal{D}} \partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}}) \partial_x\partial_y (\tilde{v}_{\mathcal{T}} + \tilde{z}_{\mathcal{T}}) \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathcal{D}} \partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}}) \partial_x\partial_y \tilde{z}_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathcal{D}} \partial_x\partial_y z_{\mathcal{T}} \partial_x\partial_y \tilde{z}_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\ &\leq \|\partial_x\partial_y z_{\mathcal{T}}\|_{L^2(\mathcal{D})} \|\partial_x\partial_y \tilde{z}_{\mathcal{T}}\|_{L^2(\mathcal{D})} \\ &\leq \frac{1}{(1-\gamma)^{1/2}} \|\partial_x\partial_y z_{\mathcal{T}}\|_{L^2(\mathcal{D})} \|\partial_x\partial_y (x_{\mathcal{T}} - u_{\mathcal{T}})\|_{L^2(\mathcal{D})}\end{aligned}$$

and thereby

$$\|\partial_x\partial_y(u-u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \frac{1}{(1-\beta)}\|\partial_x\partial_y(x_{\mathcal{T}}-u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \frac{1}{(1-\beta)(1-\gamma)^{1/2}}\|\partial_x\partial_y z_{\mathcal{T}}\|_{L^2(\mathcal{D})}.$$

Combining these results we arrive at the following two-sided bound for the error

$$\|\partial_x\partial_y z_{\mathcal{T}}\|_{L^2(\mathcal{D})} \leq \|\partial_x\partial_y(u-u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \frac{1}{(1-\beta)(1-\gamma)^{1/2}}\|\partial_x\partial_y z_{\mathcal{T}}\|_{L^2(\mathcal{D})} \tag{2.33}$$

and are thus able to use $\|\partial_x\partial_y z_{\mathcal{T}}\|$ as an a posteriori error indicator.

In the following we will be concerned with the efficient computation of $z_{\mathcal{T}}$. At first sight it might seem cheaper to compute $z_{\mathcal{T}}$, because the dimension of $Z_{\mathcal{T}}$ is smaller than that of $X_{\mathcal{T}}$. Although that is the case, in order to compute $z_{\mathcal{T}}$ we still have to solve a global system of equations over $\mathcal{D}$ and hence the computation of $z_{\mathcal{T}}$ might in the worst case be as expensive as that of $u_{\mathcal{T}}$. In most applications in the literature, see e.g. [10, 11], the functions in $Z_{\mathcal{T}}$ are chosen such that they vanish at the nodes $\mathcal{N}$ of the mesh as $Z_{\mathcal{T}}$ is the hierarchical complement of $\mathcal{S}_0^{1,0}(\mathcal{T})$ in $X_{\mathcal{T}}$. Moreover, this implies that the stiffness matrix with respect to $Z_{\mathcal{T}}$ is spectrally equivalent to a suitably lumped mass matrix by means of usual inverse inequalities for piecewise polynomial functions in $Z_{\mathcal{T}}$. Then we can replace $z_{\mathcal{T}}$ by a certain $z_{\mathcal{T}}^*$ which is computable by solving a diagonal linear system of equations. This is equivalent to assuming that there exists a bilinear form $\mathcal{B}^* : Z_{\mathcal{T}} \times Z_{\mathcal{T}} \to \mathbb{R}$ which exhibits a diagonal stiffness matrix and gives rise to an equivalent norm to $\|\partial_x\partial_y \cdot\|_{L^2(\mathcal{D})}$ on $Z_{\mathcal{T}}$, i.e. there exist $0 < \lambda \leq \Lambda$ such that

$$\lambda\|\partial_x\partial_y\zeta_{\mathcal{T}}\|^2_{L^2(\mathcal{D})} \leq \mathcal{B}^*(\zeta_{\mathcal{T}},\zeta_{\mathcal{T}}) \leq \Lambda\|\partial_x\partial_y\zeta_{\mathcal{T}}\|^2_{L^2(\mathcal{D})}, \quad \forall\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}.$$

This in turn implies that we can find a unique $z_{\mathcal{T}}^* \in Z_{\mathcal{T}}$ satisfying

$$\mathcal{B}^*(z_{\mathcal{T}}^*,\zeta_{\mathcal{T}}) = \int_{\mathcal{D}} f\zeta_{\mathcal{T}}\,\mathrm{d}x\,\mathrm{d}y - \int_{\mathcal{D}} \partial_x\partial_y u_{\mathcal{T}}\partial_x\partial_y\zeta_{\mathcal{T}}\,\mathrm{d}x\,\mathrm{d}y, \quad \forall\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}. \tag{2.34}$$

Similarly as before we conclude for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$

$$\mathcal{B}^*(z_{\mathcal{T}}^*,\zeta_{\mathcal{T}}) = \int_{\mathcal{D}} \partial_x\partial_y(u-u_{\mathcal{T}})\partial_x\partial_y\zeta_{\mathcal{T}}\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathcal{D}} \partial_x\partial_y z_{\mathcal{T}}\partial_x\partial_y\zeta_{\mathcal{T}}\,\mathrm{d}x\,\mathrm{d}y.$$

Using the equivalence of $\mathcal{B}^*$ with the norm $\|\partial_x\partial_y \cdot\|_{L^2(\mathcal{D})}$ and inserting $\zeta_{\mathcal{T}} = z_{\mathcal{T}}^*$ as well as $\zeta_{\mathcal{T}} = z_{\mathcal{T}}$, we find

$$\begin{aligned}
\mathcal{B}^*(z_{\mathcal{T}}^*,z_{\mathcal{T}}^*) &= \int_{\mathcal{D}} \partial_x\partial_y(u-u_{\mathcal{T}})\partial_x\partial_y z_{\mathcal{T}}^* \\
&\leq \|\partial_x\partial_y(u-u_{\mathcal{T}})\|_{L^2(\mathcal{D})}\|\partial_x\partial_y z_{\mathcal{T}}^*\|_{L^2(\mathcal{D})} \\
&\leq \|\partial_x\partial_y(u-u_{\mathcal{T}})\|_{L^2(\mathcal{D})}\frac{1}{\sqrt{\lambda}}\mathcal{B}^*(z_{\mathcal{T}}^*,z_{\mathcal{T}}^*)^{1/2}
\end{aligned}$$

and

$$\begin{aligned}
\|\partial_x\partial_y z_{\mathcal{T}}\|^2_{L^2(\mathcal{D})} &= \mathcal{B}^*(z_{\mathcal{T}}^*,z_{\mathcal{T}}) \\
&\leq \mathcal{B}^*(z_{\mathcal{T}}^*,z_{\mathcal{T}}^*)^{1/2}\mathcal{B}^*(z_{\mathcal{T}},z_{\mathcal{T}})^{1/2} \\
&\leq \mathcal{B}^*(z_{\mathcal{T}}^*,z_{\mathcal{T}}^*)^{1/2}\sqrt{\Lambda}\|\partial_x\partial_y z_{\mathcal{T}}\|_{L^2(\mathcal{D})}.
\end{aligned}$$

This proves the two-sided error bound

$$\sqrt{\lambda}\mathcal{B}^*(z_{\mathcal{T}}^*, z_{\mathcal{T}}^*)^{1/2} \leq \|\partial_x\partial_y(u-u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \frac{\sqrt{\Lambda}}{(1-\beta)(1-\gamma)^{1/2}}\mathcal{B}^*(z_{\mathcal{T}}^*, z_{\mathcal{T}}^*)^{1/2}. \tag{2.35}$$

Our main obstacle to define a hierarchical a posteriori estimator now lies with the specific choice of $Z_{\mathcal{T}}$ and a suitable choice for $\mathcal{B}^*$. Let us first consider $\mathcal{T}$ to be regular and let

$$X_{\mathcal{T}} = \mathcal{S}_0^{1,0}(\mathcal{T}) \oplus Z_{\mathcal{T}},$$

which defines the hierarchical complement $Z_{\mathcal{T}} = \mathcal{S}_0^{2,0}(\mathcal{T})\backslash\mathcal{S}_0^{1,0}(\mathcal{T})$ as the space of piecewise continuous polynomials of degree 2 on $\mathcal{T}$ which vanish at the nodes $\mathcal{N}$ of $\mathcal{T}$. We associate with every edge $E \in \mathcal{E}$ and element $K \in \mathcal{T}$ the functions ψ_E and ψ_K, which are called *edge* and *element bubble* functions, respectively, and split $Z_{\mathcal{T}}$ as follows

$$Z_{\mathcal{T}} = Z_{\mathcal{T}}^E \oplus Z_{\mathcal{T}}^K,$$

where now $Z_{\mathcal{T}}^E = \operatorname{span}\{\psi_E : E \in \mathcal{E}\}$ is the space of edge bubbles and $Z_{\mathcal{T}}^K = \operatorname{span}\{\psi_K : K \in \mathcal{T}\}$ the space of element bubbles.

Then $\psi_K \in \mathcal{P}_2(K)$ and $0 \leq \psi_K \leq C_K$ on K as well as $\psi_K|_{\partial K} = 0$. For an interior edge E we have $\psi_E \in \mathcal{P}_2(K \cup K')$ with $E = \overline{K} \cap \overline{K'}$ in addition to $0 \leq \psi_E \leq C_E$ and $\psi_E|_{\partial(K\cup K')} = 0$. The constants depend on the specific definition of $\mathcal{S}_0^{2,0}(\mathcal{T})$ in that $C_K = \max_{x\in K}|\psi_K(x)|$ and $C_E = \max_{x\in(K\cup K')}|\psi_E(x)|$. In particular, frequently ψ_E and ψ_K are chosen such that $C_K = C_E = 1$. On $\hat{K}$ we can explicitly define

$$\begin{aligned}
\psi_{\hat{K}}(\hat{x},\hat{y}) &:= C_{\hat{K}}(1-\hat{x}^2)(1-\hat{y}^2),\\
\psi_{\hat{E}}(\hat{x},\hat{y}) &:= C_{\hat{E}}\cdot\begin{cases}
\frac{(1-\hat{y})}{2}(1-\hat{x}^2), & \hat{E} = [-1,1]\times\{-1\},\\
\frac{(1+\hat{x})}{2}(1-\hat{y}^2), & \hat{E} = \{1\}\times[-1,1],\\
\frac{(1+\hat{y})}{2}(1-\hat{x}^2), & \hat{E} = [-1,1]\times\{1\},\\
\frac{(1-\hat{x})}{2}(1-\hat{y}^2), & \hat{E} = \{-1\}\times[-1,1].
\end{cases}
\end{aligned} \tag{2.36}$$

Choosing $\mathcal{B}^*$ as the restriction of the $H^{1,1}(\mathcal{D})$ scalar product, namely

$$\mathcal{B}^*\left(\sum_S \alpha_S\psi_S, \sum_{S'}\alpha_{S'}\psi_{S'}\right) = \sum_{S,S'}\alpha_S\alpha_{S'}\int_{\mathcal{D}}\partial_x\partial_y\psi_S\partial_x\partial_y\psi_{S'}\,\mathrm{d}x\,\mathrm{d}y,$$

with $S, S' \in \mathcal{T}\cup\mathcal{E}$ shows that the strengthened Cauchy-Schwarz inequality holds with $\gamma = 0$. More precisely, since $\psi_S, S \in \mathcal{T}\cup\mathcal{E}$ can be written in terms of tensor products of antiderivatives of the Legendre polynomials $\hat{L}_0(\xi) \equiv 1, \hat{L}_1(\xi) = \xi + 1, \hat{L}_2(\cdot) = \frac{1}{2}(\xi^2 - 1)$ as

$$\begin{aligned}
\psi_{\hat{K}}(\hat{x},\hat{y}) =& \tilde{c}\hat{L}_2(\hat{x})\hat{L}_2(\hat{y}),\\
\psi_{\hat{E}}(\hat{x},\hat{y}) =& \tilde{c}\cdot\begin{cases}
(\hat{L}_0(\hat{y}) - \frac{1}{2}\hat{L}_1(\hat{y}))\hat{L}_2(\hat{x}), & \hat{E} = [-1,1]\times\{-1\},\\
\frac{1}{2}\hat{L}_1(\hat{x})\hat{L}_2(\hat{y}), & \hat{E} = \{1\}\times[-1,1],\\
\frac{1}{2}\hat{L}_1(\hat{y})\hat{L}_2(\hat{x}), & \hat{E} = [-1,1]\times\{1\},\\
(\hat{L}_0(\hat{x}) - \frac{1}{2}\hat{L}_1(\hat{x}))\hat{L}_2(\hat{y}), & \hat{E} = \{-1\}\times[-1,1],
\end{cases}
\end{aligned} \tag{2.37}$$

where $\tilde{c}$ in each instance is a certain normalization constant. The aforementioned orthogonality of $S_0^{1,0}(\mathcal{T})$ and $Z_\mathcal{T}$ is then seen as a result of orthogonality properties of the Legendre polynomials. Moreover, we note that in this way the element and edge bubble functions are also mutually orthogonal.

For convenience let us denote the inner product on $H^{1,1}(\mathcal{D})$ by

$$\langle u, v\rangle_{H^{1,1}(\mathcal{D})} := \int_\mathcal{D} \partial_x\partial_y u \partial_x\partial_y v \,\mathrm{d}x\,\mathrm{d}y$$

and note that $\langle u, u\rangle_{H^{1,1}(\mathcal{D})} = |u|^2_{H^{1,1}(\mathcal{D})}$. Using the established orthogonality between local shape functions in $Y_\mathcal{T}$ and $Z_\mathcal{T}$ shows that (cf. (2.34)) we only have to solve the simpler problem:

Problem 2.35. *Find $z^*_\mathcal{T} \in Z_\mathcal{T}$, such that*

$$\mathcal{B}^*(z^*_\mathcal{T}, \zeta_\mathcal{T}) = \int_\mathcal{D} f\zeta_\mathcal{T} \,\mathrm{d}x\,\mathrm{d}y, \quad \forall \zeta_\mathcal{T} \in Z_\mathcal{T}.$$

Furthermore, note that for this choice of $\mathcal{B}^*$, namely as the restriction of $\mathcal{B}$ to $Z_\mathcal{T}$, the spectral equivalence with $|\cdot|_{H^{1,1}(\mathcal{D})}$ on $Z_\mathcal{T}$ is trivial. Considering the functions ψ_S of $Z^K_\mathcal{T}$ with $S \in \mathcal{T}$ a straightforward calculation to find the coefficients of $z^*_\mathcal{T} = \sum_S \alpha_S\psi_S$ where $S \in \mathcal{T}\cup\mathcal{E}$ by testing with a certain $\psi_{S'}$ with $S' \in \mathcal{T}$ then leads to

$$\begin{aligned}\langle z^*_\mathcal{T}, \psi_{S'}\rangle_{H^{1,1}(\mathcal{D})} &= \int_\mathcal{D} \partial_x\partial_y z^*_\mathcal{T} \partial_x\partial_y\psi_{S'} \,\mathrm{d}x\,\mathrm{d}y = \sum_{S\in\mathcal{T}} \alpha_S \int_\mathcal{D} \partial_x\partial_y\psi_S\partial_x\partial_y\psi_{S'} \,\mathrm{d}x\,\mathrm{d}y \\ &= \alpha_{S'}\|\partial_x\partial_y\psi_{S'}\|^2_{L^2(\mathcal{D})} \\ &= \int_\mathcal{D} f\psi_{S'} \,\mathrm{d}x\,\mathrm{d}y,\end{aligned}$$

where we have used the fact that the element bubbles do not have overlapping support. This shows that for all $K \in \mathcal{T}$

$$\alpha_K = \frac{(f, \psi_K)_{L^2(K)}}{\|\partial_x\partial_y\psi_K\|^2_{L^2(K)}}. \tag{2.38}$$

To find the remaining coefficients of the edge bubble functions in $Z^E_\mathcal{T}$ we note that bubble functions for edges in x- and y-direction are orthogonal. Thus, for any vertical edge $E \in \mathcal{E}$

$$\begin{aligned}\langle z^*_\mathcal{T}, \psi_E\rangle_{H^{1,1}(\mathcal{D})} &= \int_\mathcal{D} \partial_x\partial_y z^*_\mathcal{T} \partial_x\partial_y\psi_E \,\mathrm{d}x\,\mathrm{d}y \\ &= \sum_{S\in\mathcal{E}} \alpha_S \int_\mathcal{D} \partial_x\partial_y\psi_S\partial_x\partial_y\psi_E \,\mathrm{d}x\,\mathrm{d}y \\ &= \alpha_E\|\partial_x\partial_y\psi_E\|^2_{L^2(\omega_E)} + \alpha_{E^-}\langle\psi_E, \psi_{E^-}\rangle_{H^{1,1}(\omega_E\cap\omega_{E^-})} \\ &\qquad + \alpha_{E^+}\langle\psi_E, \psi_{E^+}\rangle_{H^{1,1}(\omega_E\cap\omega_{E^+})} \\ &= \int_\mathcal{D} f\psi_E \,\mathrm{d}x\,\mathrm{d}y,\end{aligned}$$

and an analogous argument for the horizontal edges leads to

$$\alpha_E = \frac{(f, \psi_E)_{L^2(\mathrm{supp}(\psi_E))} - \alpha_{E^-}\langle\psi_E, \psi_{E^-}\rangle_{H^{1,1}(\omega_E\cap\omega_{E^-})} - \alpha_{E^+}\langle\psi_E, \psi_{E^+}\rangle_{H^{1,1}(\omega_E\cap\omega_{E^+})}}{\|\partial_x\partial_y\psi_E\|^2_{L^2(\mathrm{supp}(\psi_E))}}, \tag{2.39}$$

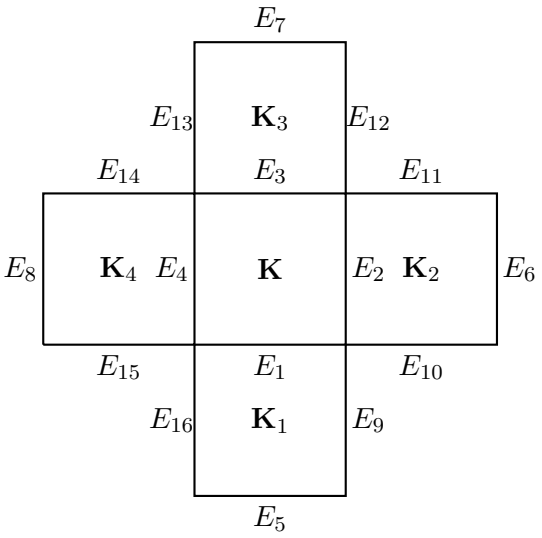

Figure 2.4: Interaction of basis functions on ω_K

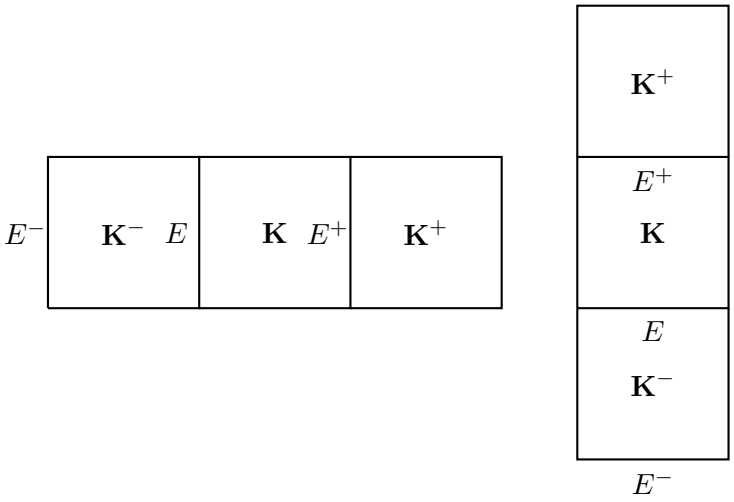

Figure 2.5: Interaction of edge bubble functions across a horizontal/vertical patch of elements

where the notation is as in Figure 2.5.

Taking the L^2-norm of $\partial_x\partial_y z^*_{\mathcal{T}}$ on ω_K, with notations as in Figure 2.4 and abbreviating $\langle\cdot,\cdot\rangle_{H^{1,1}(\cdot)}$ to just $\langle\cdot,\cdot\rangle$, we find the following explicit expression

$$\begin{aligned}\|\partial_x\partial_y z^*_{\mathcal{T}}\|^2_{L^2(\omega_K)} = &\sum_{K'\subset\omega_K}\alpha^2_{K'}|\psi_{K'}|^2_{H^{1,1}(K')} + \sum_{E\in\mathcal{E}_{\omega_K}}\alpha^2_E|\psi_E|^2_{H^{1,1}(\omega_E\cap\omega_K)}\\ &+\alpha_{E_1}\alpha_{E_5}\langle\psi_{E_1},\psi_{E_5}\rangle + \alpha_{E_1}\alpha_{E_3}\langle\psi_{E_1},\psi_{E_3}\rangle\\ &+\alpha_{E_2}\alpha_{E_6}\langle\psi_{E_2},\psi_{E_6}\rangle + \alpha_{E_2}\alpha_{E_4}\langle\psi_{E_2},\psi_{E_4}\rangle\\ &+\alpha_{E_3}\alpha_{E_1}\langle\psi_{E_3},\psi_{E_1}\rangle + \alpha_{E_3}\alpha_{E_7}\langle\psi_{E_3},\psi_{E_7}\rangle\\ &+\alpha_{E_4}\alpha_{E_8}\langle\psi_{E_4},\psi_{E_8}\rangle + \alpha_{E_4}\alpha_{E_2}\langle\psi_{E_4},\psi_{E_2}\rangle\\ &+\alpha_{E_{16}}\alpha_{E_9}\langle\psi_{E_{16}},\psi_{E_9}\rangle + \alpha_{E_{10}}\alpha_{E_{11}}\langle\psi_{E_{10}},\psi_{E_{11}}\rangle\\ &+\alpha_{E_{12}}\alpha_{E_{13}}\langle\psi_{E_{12}},\psi_{E_{13}}\rangle + \alpha_{E_{14}}\alpha_{E_{15}}\langle\psi_{E_{14}},\psi_{E_{15}}\rangle.\end{aligned} \tag{2.40}$$

The equations for the coefficients α_K and α_E constitute the part of the global stiffness matrix with respect to the functions in $Z_{\mathcal{T}}$, which does not interact with the degrees of freedom of the numerical solution $u_{\mathcal{T}} \in \mathcal{S}_0^{1,0}(\mathcal{T})$. If we denote by $N_{\mathcal{E}}$ the number of edges in $\mathcal{E}$ and by $N_{\mathcal{T}}$ the number of elements, then the linear system we have to solve in order to compute our hierarchical error estimator has size $N_{\mathcal{E}} + N_{\mathcal{T}}$ and is sparse. But as the inverse of a sparse matrix may fail to be sparse, a direct solution of the linear system may be as costly as the computation of $u_{\mathcal{T}}$. This and the fact that we are using 1-irregular partitions is the departure point for the following considerations.

Remark 2.36. *In the situation, when $p = 2$ and we are dealing with 1-irregular meshes and perform a system reduction similiar to the procedure in section 2.2.3 many edge bubbles will necessarily be coupled with nodal functions and hence the structure of system is more complicated. Solving exactly would mean that we have to compute the global system matrix for $p = 2$ although we only use the shape functions of order two for error estimation. This is of course ineffective and undesireable.*

Instead of solving the global system directly, we may resort to an approximation by a single Jacobi iteration step or we may choose $Z_{\mathcal{T}} \equiv Z_{\mathcal{T}}^K$, i.e. the extension of the bilinear FE space on an element by its associated element bubble function ψ_K.

Before we have another look at these questions we shall define the hierarchical error estimator η_H and show reliability and weak efficiency.

Definition 2.37. *We define the local hierarchical error estimator by means of the sum over the relevant basis functions in $Z_{\mathcal{T}}$ that are associated with the element patch ω_K, i.e.*

$$\eta_{H,K}^2 := h_K^{-2} \|\partial_x \partial_y z_{\mathcal{T}}^*\|_{L^2(\omega_K)}^2 \tag{2.41}$$

and moreover set the global error estimator to

$$\eta_H := \left(\sum_{K \in \mathcal{T}} \eta_{H,K}^2 \right)^{1/2}. \tag{2.42}$$

Theorem 2.38. *Let u be the exact solution of (2.6) and $u_{\mathcal{T}}$ the corresponding solution of the discrete problem (2.7). Let $Z_{\mathcal{T}}$ as above and let the* hierarchical a posteriori *error estimator η_H be given as in (2.42). Then there exist constants $c^*, c_* > 0$, such that η_H is* reliable

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq c^* \left\{ \eta_H^2 + \sum_{K \in \mathcal{T}} h_K^2 \|f_{\mathcal{T}} - f\|_{L^2(K)}^2 \right\}^{1/2} \tag{2.43}$$

and there holds the lower bound

$$h_K \eta_{H,K} \leq c_* \sum_{K' \subset \omega_K} \left(|u - u_{\mathcal{T}}|_{H^{1,1}(\omega_{K'})}^2 + \sum_{K'' \subset \omega_{K'}} h_{K''}^4 \|f - f_{\mathcal{T}}\|_{L^2(K'')}^2 \right)^{1/2}. \tag{2.44}$$

Proof. We will firstly show the reliability. To this end we use arguments as for the efficiency of the residual error estimator (cf. (2.29)). Note that there holds for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$

$$\int_{\mathcal{D}} \partial_x \partial_y (u - u_{\mathcal{T}}) \partial_x \partial_y \zeta_{\mathcal{T}} = \int_{\mathcal{D}} f \zeta_{\mathcal{T}} - \int_{\mathcal{D}} \partial_x \partial_y u_{\mathcal{T}} \partial_x \partial_y \zeta_{\mathcal{T}} = \int_{\mathcal{D}} \partial_x \partial_y z_{\mathcal{T}}^* \partial_x \partial_y \zeta_{\mathcal{T}}.$$

Then as in Lemma 2.28 with $\tilde{r} = f_{\mathcal{T}} - \partial_x^2\partial_y^2 u_{\mathcal{T}}$ and $w_K = \tilde{r}\psi_K \in Z_{\mathcal{T}}$ we have

$$\begin{aligned}
c\|\tilde{r}\|^2_{L^2(K)} &\le \int (f_{\mathcal{T}} - \partial_x^2\partial_y^2 u_{\mathcal{T}})^2\psi_K \\
&= \int_K r w_K + \int_K (f_{\mathcal{T}} - f) w_K \\
&= \int_K \partial_x\partial_y(u - u_{\mathcal{T}})\partial_x\partial_y w_K + \int_K (f_{\mathcal{T}} - f) w_K \\
&= \int_K \partial_x\partial_y z^*_{\mathcal{T}}\partial_x\partial_y w_K + \int_K (f_{\mathcal{T}} - f) w_K \\
&\le |z^*_{\mathcal{T}}|_{H^{1,1}(K)}|w_K|_{H^{1,1}(K)} + \|f - f_{\mathcal{T}}\|_{L^2(K)}\|w_K\|_{L^2(K)} \\
&\le C h_K^{-2}|z^*_{\mathcal{T}}|_{H^{1,1}(K)}\|\tilde{r}\|_{L^2(K)} + \|f - f_{\mathcal{T}}\|_{L^2(K)}\|\tilde{r}\|_{L^2(K)}
\end{aligned}$$

and hence $h_K^2\|\tilde{r}\|_K \le c|z^*_{\mathcal{T}}|_{H^{1,1}(K)} + h_K^2\|f - f_{\mathcal{T}}\|_{L^2(K)}$. Similarly, with $w_E = j\psi_E \in Z_{\mathcal{T}}$ as in Lemma 2.29 we find

$$\begin{aligned}
c\|j\|^2_{L^2(E)} &\le \int_E j w_E = \int_{\omega_E} \partial_x\partial_y(u - u_{\mathcal{T}})\partial_x\partial_y w_E - \int_{\omega_E} r w_E \\
&= \sum_{K\subset\omega_E} \Big\{ \int_K \partial_x\partial_y(u - u_{\mathcal{T}})\partial_x\partial_y w_E - \int_K (f_{\mathcal{T}} - \partial_x^2\partial_y^2 u_{\mathcal{T}}) w_E \\
&\quad - \int_K (f - f_{\mathcal{T}}) w_E \Big\} \\
&= \sum_{K\subset\omega_E} \Big\{ \int_K \partial_x\partial_y z^*_{\mathcal{T}}\partial_x\partial_y w_E - \int_K (f_{\mathcal{T}} - \partial_x^2\partial_y^2 u_{\mathcal{T}}) w_E \\
&\quad - \int_K (f - f_{\mathcal{T}}) w_E \Big\} \\
&\le \sum_{K\subset\omega_E} \big\{ \|\partial_x\partial_y z^*_{\mathcal{T}}\|_{L^2(K)}\|\partial_x\partial_y w_E\|_{L^2(K)} + \|\tilde{r}\|_{L^2(K)}\|w_E\|_{L^2(K)} \\
&\quad + \|(f - f_{\mathcal{T}})\|_{L^2(K)}\|w_E\|_{L^2(K)} \big\} \\
&\le \sum_{K\subset\omega_E} \Big\{ C_1\|\partial_x\partial_y z^*_{\mathcal{T}}\|_{L^2(K)} h_E^{-3/2}\|j\|_{L^2(E)} + C_2\|\tilde{r}\|_{L^2(K)} h_E^{1/2}\|j\|_{L^2(E)} \\
&\quad + C_3\|(f - f_{\mathcal{T}})\|_{L^2(K)} h_E^{1/2}\|j\|_{L^2(E)} \Big\}.
\end{aligned}$$

Hence, with the bound for $h_K^2\|\tilde{r}\|_{L^2(K)}$ we have

$$h_E^{3/2}\|j\| \le C \sum_{K\subset\omega_E} \left(\|\partial_x\partial_y z^*_{\mathcal{T}}\|_{L^2(K)} + h_K^2\|(f - f_{\mathcal{T}})\|_{L^2(K)} \right),$$

where we have used shape regularity to estimate $h_E^2 \le c h_K^2$. Thus, since

$$\|r\|_{L^2(K)} \le \|\tilde{r}\|_{L^2(K)} + \|f - f_{\mathcal{T}}\|_{L^2(K)}$$

we find

$$h_K^4\|\tilde{r}\|^2_{L^2(K)} + \sum_{E\in\mathcal{E}_K} h_E^3\|j\|^2_{L^2(E)} \le C \sum_{K'\subset\omega_K} \left(\|\partial_x\partial_y z^*_{\mathcal{T}}\|^2_{L^2(K')} + h_{K'}^4\|f - f_{\mathcal{T}}\|^2_{L^2(K')} \right) \tag{2.45}$$

which in turn with shape regularity implies that

$$\eta_{\mathcal{R},K}^2 \leq Ch_K^{-2}\|\partial_x\partial_y z_{\mathcal{T}}^*\|_{L^2(\omega_K)}^2 + \sum_{K'\subset\omega_K} h_{K'}^2\|f - f_{\mathcal{T}}\|_{L^2(K')}^2.$$

Summing over all $K \in \mathcal{T}$ and taking the square root shows the reliability of η_H, since $\eta_{\mathcal{R}}$ is reliable by Theorem 2.30. Let us now show the weak efficiency. First we note that there holds

$$\|\partial_x\partial_y z_{\mathcal{T}}^*\|_{L^2(\omega_K)}^2 = \sum_{K'\subset\omega_K} \alpha_{K'}\langle z_{\mathcal{T}}^*, \psi_{K'}\rangle_{H^{1,1}(K')} + \sum_{E\in\mathcal{E}_{\omega_K}} \alpha_E\langle z_{\mathcal{T}}^*, \psi_E\rangle_{H^{1,1}(\omega_E\cap\omega_K)}.$$

Furthermore, by Lemma 2.13 for any $K \subset \omega_K$ and the residual $\mathcal{R}$ expressed by (2.16) we find the bound

$$\begin{aligned}
\langle z_{\mathcal{T}}^*, \psi_K\rangle_{H^{1,1}(K)} &= \langle \mathcal{R}, \psi_K\rangle \\
&= \int_K f\psi_K - \int_K \partial_x\partial_y u_{\mathcal{T}}\partial_x\partial_y\psi_K \\
&\leq \|f - \partial_x^2\partial_y^2 u_{\mathcal{T}}\|_{L^2(K)}\|\psi_K\|_{L^2(K)} \\
&\leq Ch_K^2\|r\|_{L^2(K)}\|\partial_x\partial_y\psi_K\|_{L^2(K)}
\end{aligned}$$

as well as for any $E \in \mathcal{E}_{\omega_K}$ there holds

$$\begin{aligned}
\langle z_{\mathcal{T}}^*, \psi_E\rangle_{H^{1,1}(\omega_E\cap\omega_K)} &= \langle \mathcal{R}, \psi_E\rangle \\
&= \int_{\omega_E\cap\omega_K} (f - \partial_x^2\partial_y^2 u_{\mathcal{T}})\psi_E + \int_E [\![\nabla\partial_\tau^2 u_{\mathcal{T}}\cdot\mathbf{n}]\!]\psi_E \\
&\leq Ch_K^2\|r\|_{L^2(\omega_E\cap\omega_K)}\|\partial_x\partial_y\psi_E\|_{L^2(\omega_E\cap\omega_K)} \\
&\quad + Ch_E^{3/2}\|j\|_{L^2(E)}\|\partial_x\partial_y\psi_E\|_{L^2(\omega_E\cap\omega_K)}.
\end{aligned}$$

Combining these estimates yields

$$\begin{aligned}
\|\partial_x\partial_y z_{\mathcal{T}}^*\|_{L^2(\omega_K)}^2 &= \sum_{K'\subset\omega_K} \alpha_{K'}\langle z_{\mathcal{T}}^*, \psi_{K'}\rangle_{H^{1,1}(K')} + \sum_{E\in\mathcal{E}_{\omega_K}} \alpha_E\langle z_{\mathcal{T}}^*, \psi_E\rangle_{H^{1,1}(\omega_E\cap\omega_K)} \\
&\leq \sum_{K'\subset\omega_K} \alpha_{K'}\left(Ch_{K'}^2\|r\|_{L^2(K)}\|\partial_x\partial_y\psi_{K'}\|_{L^2(K')}\right) \\
&\quad + \sum_{E\in\mathcal{E}_{\omega_K}} \alpha_E\left(Ch_K^2\|r\|_{L^2(\omega_E\cap\omega_K)}\|\partial_x\partial_y\psi_E\|_{L^2(\omega_E\cap\omega_K)}\right. \\
&\qquad \left. + Ch_E^{3/2}\|j\|_{L^2(E)}\|\partial_x\partial_y\psi_E\|_{L^2(\omega_E\cap\omega_K)}\right) \\
&\leq C\left(\sum_{K'\subset\omega_K} h_K^4\|r\|_{L^2(K')}^2 + \sum_{E\in\mathcal{E}_{\omega_K}} h_E^3\|j\|_{L^2(E)}^2\right)^{1/2} \\
&\quad \times \left(\sum_{K'\subset\omega_K} \alpha_{K'}^2\|\partial_x\partial_y\psi_{K'}\|_{L^2(K')}^2 + \sum_{E\in\mathcal{E}_{\omega_K}} \alpha_E^2\|\partial_x\partial_y\psi_E\|_{L^2(\omega_E\cap\omega_K)}^2\right)^{1/2} \\
&\leq C\left(\sum_{K'\subset\omega_K} h_{K'}^4\|r\|_{L^2(K')}^2 + \sum_{E\in\mathcal{E}_{\omega_K}} h_E^3\|j\|_{L^2(E)}^2\right)^{1/2} \\
&\quad \times \left(\sum_{K'\subset\omega_K} \alpha_{K'}\langle z_{\mathcal{T}}^*, \psi_{K'}\rangle_{H^{1,1}(K')} + \sum_{E\in\mathcal{E}_{\omega_K}} \alpha_E\langle z_{\mathcal{T}}^*, \psi_E\rangle_{H^{1,1}(\omega_E\cap\omega_K)}\right)^{1/2}
\end{aligned}$$

and therefore with $h_K^2 \leq ch_E^2$

$$h_K^{-2}\|\partial_x\partial_y z_{\mathcal{T}}^*\|^2_{L^2(\omega_K)} \leq C\left(\sum_{K'\subset\omega_K} h_{K'}^2\|r\|^2_{L^2(K')} + \sum_{E\in\mathcal{E}_{\omega_K}} h_E\|j\|^2_{L^2(E)}\right) = C\sum_{K'\subset\omega_K}\eta^2_{\mathcal{R},K'}. \tag{2.46}$$

This implies that

$$\begin{aligned} h_K\eta_{H,K} \leq & C\sum_{K'\subset\omega_K} h_{K'}\eta_{\mathcal{R},K'} \\ \leq & C\sum_{K'\subset\omega_K}\left(|u-u_{\mathcal{T}}|^2_{H^{1,1}(\omega_{K'})} + \sum_{K''\subset\omega_{K'}} h_{K''}^4\|f-f_{\mathcal{T}}\|^2_{L^2(K'')}\right)^{1/2} \end{aligned}$$

from which we find the assertion. □

Remark 2.39. *Similar arguments show that we can also take*

$$\tilde{\eta}^2_{H,K} = h_K^{-2}\|\partial_x\partial_y z_{\mathcal{T}}^*\|^2_K$$

as a more cost effective a posteriori *error estimator. Then we also set*

$$\tilde{\eta}_H = \left(\sum_{K\in\mathcal{T}}\tilde{\eta}^2_{H,K}\right)^{1/2}.$$

A closer inspection of (2.40) shows that

$$|\psi_E|^2_{H^{1,1}(K)} \leq C|\psi_K|^2_{H^{1,1}(K)}, \forall E\in\mathcal{E}_K$$

as well as

$$|\langle\psi_E,\psi_{E'}\rangle_{H^{1,1}(K)}| \leq C|\psi_K|^2_{H^{1,1}(K)}, \forall E\in\mathcal{E}_K$$

and so there exists a finite constant C and a constant $\tilde{c}$ such that

$$\alpha_K^2\|\partial_x\partial_y\psi_K\|^2_{L^2(K)} \leq \|\partial_x\partial_y z_{\mathcal{T}}\|^2_{L^2(K)} \leq C(\alpha_K^2+\tilde{c})\|\partial_x\partial_y\psi_K\|^2_{L^2(K)}, \tag{2.47}$$

where

$$\tilde{c} = \sum_{E\in\mathcal{E}_K}\alpha_E^2 + \alpha_{E_1}\alpha_{E_3} + \alpha_{E_2}\alpha_{E_4}$$

Moreover, this implies that there exists a uniformly bounded constant C, such that

$$\|\partial_x\partial_y z_{\mathcal{T}}\|^2_{L^2(\omega_K)} \leq C\alpha_K^2\|\partial_x\partial_y\psi_K\|^2_{L^2(K)},$$

Of course, the coefficients $\alpha_E, E\in\mathcal{E}$ depend on the right-hand side f, but since the functions in $Z_{\mathcal{T}}$ represent a hierarchical extension of $\mathcal{S}^{1,0}(\mathcal{T})$ we expect the coefficients α_E to become small as $h\to 0$. This is a tempting device for error estimation, since we only have to compute the coefficients with respect to the element bubbles which is equivalent to solving a diagonal system of equations.

This suggests that using

$$\hat{\eta}_{H,K} = \alpha_K^2 h_K^{-2}\|\partial_x\partial_y\psi_K\|^2_{L^2(K)} \tag{2.48}$$

as an error estimator is a cost-effective solution.

Remark 2.40. *By the preceding heuristics we define*

$$\hat{\eta}_{H,K}^2 := \alpha_K^2 h_K^{-2} \|\partial_x \partial_y \psi_K\|_K^2$$

as a cost-effective alternative to η_H and $\tilde{\eta}_H$ and denote the global error estimator by

$$\hat{\eta}_H = \left(\sum_{K \in \mathcal{T}} \hat{\eta}_{H,K}^2 \right)^{1/2}.$$

2.2.6 An a posteriori estimator based on averaging

In the following we want to construct a variant of an a posteriori estimator based on a popular averaging technique, which was first introduced by Zhu and Zienkiewicz in [61]. We follow the presentation of [54] and the ideas of [16, 17] to develop an asymptotically exact a posteriori error estimator for our situation of the second moment problem.

To this end, suppose u solves the variational formulation

$$\int_{\mathcal{D}} \partial_x \partial_y u \partial_x \partial_y v = \int_{\mathcal{D}} f v, \quad \forall v \in H_0^{1,1}(\mathcal{D})$$

and $u_{\mathcal{T}}$ is the solution of the corresponding discrete formulation with $p = 1$. We are interested in finding a higher order approximation q with respect to $\partial_x \partial_y u_{\mathcal{T}}$ of $\partial_x \partial_y u$ which is easily computable and for which we can expect that

$$\|\partial_x \partial_y u - \partial_x \partial_y u_{\mathcal{T}}\|_{L^2(\mathcal{D})} \leq \|q - \partial_x \partial_y u_{\mathcal{T}}\|_{L^2(\mathcal{D})}.$$

This motivates the following general definition.

Definition 2.41. *The elementwise error indicator is given by*

$$\eta_{Z,K} := \min_{q \in \mathcal{S}^{1,0}(\mathcal{T})} \|\partial_x \partial_y u_{\mathcal{T}} - q\|_{L^2(K)}$$

and the global error estimator as

$$\eta_Z = \left(\sum_{K \in \mathcal{T}} \eta_{Z,K}^2 \right)^{1/2}.$$

Suppose that η_Z is reliable and that we have at our disposal an easily computable function $\mathcal{A}(\partial_x \partial_y u_{\mathcal{T}}) \in \mathcal{S}^{1,0}(\mathcal{T})$ with an operator $\mathcal{A} : \mathcal{P}_1(\mathcal{T}) \to \mathcal{S}^{1,0}(\mathcal{T})$ which yields an approximation to $\partial_x \partial_y u$. Then the previous Definition shows reliability immediately by setting $q = \mathcal{A}(\partial_x \partial_y u_{\mathcal{T}})$, since

$$\eta_{Z,K} \leq \eta_{\mathcal{A},K} := \|\partial_x \partial_y u_{\mathcal{T}} - \mathcal{A}(\partial_x \partial_y u_{\mathcal{T}})\|_{L^2(\mathcal{D})}.$$

If $\mathcal{A}(\partial_x \partial_y u_{\mathcal{T}})$ is indeed a higher order approximation, one might expect that $\mathcal{A}(\partial_x \partial_y u_{\mathcal{T}})$ fulfills a saturation property with respect to $\mathcal{S}^{1,0}(\mathcal{T})$, i.e.

$$\|\partial_x \partial_y u - \mathcal{A}(\partial_x \partial_y u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \beta \|\partial_x \partial_y u - \partial_x \partial_y u_{\mathcal{T}}\|_{L^2(\mathcal{D})} \tag{2.49}$$

for some $\beta \in [0, 1)$. Then, on the one hand by the triangle inequality we have that

$$\begin{aligned}
\|\mathcal{A}(\partial_x \partial_y u_{\mathcal{T}}) - \partial_x \partial_y u_{\mathcal{T}}\|_{L^2(\mathcal{D})} &= \|\mathcal{A}(\partial_x \partial_y u_{\mathcal{T}}) - \partial_x \partial_y u + \partial_x \partial_y u - \partial_x \partial_y u_{\mathcal{T}}\|_{L^2(\mathcal{D})} \\
&\leq \|\mathcal{A}(\partial_x \partial_y u_{\mathcal{T}}) - \partial_x \partial_y u\|_{L^2(\mathcal{D})} + \|\partial_x \partial_y (u - u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \\
&\leq (\beta + 1) \|\partial_x \partial_y (u - u_{\mathcal{T}})\|_{L^2(\mathcal{D})}
\end{aligned}$$

as well as on the other hand again by triangle inequality

$$\begin{aligned}\|\partial_x\partial_y u - \partial_x\partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})} &= \|\partial_x\partial_y u - \mathcal{A}(\partial_x\partial_y u_\mathcal{T}) + \mathcal{A}(\partial_x\partial_y u_\mathcal{T}) - \partial_x\partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})} \\ &\leq \|\partial_x\partial_y u - \mathcal{A}(\partial_x\partial_y u_\mathcal{T})\|_{L^2(\mathcal{D})} + \|\mathcal{A}(\partial_x\partial_y u_\mathcal{T}) - \partial_x\partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})} \\ &\leq \beta\|\partial_x\partial_y(u - u_\mathcal{T})\|_{L^2(\mathcal{D})} + \|\mathcal{A}(\partial_x\partial_y u_\mathcal{T}) - \partial_x\partial_y u_\mathcal{T})\|_{L^2(\mathcal{D})}.\end{aligned}$$

Therefore, in total we have

$$\begin{aligned}\frac{1}{1+\beta}\|\mathcal{A}(\partial_x\partial_y u_\mathcal{T}) - \partial_x\partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})} \leq& \|\partial_x\partial_y(u - u_\mathcal{T})\|_{L^2(\mathcal{D})} \\ \leq& \frac{1}{1-\beta}\|\mathcal{A}(\partial_x\partial_y u_\mathcal{T}) - \partial_x\partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})}.\end{aligned}$$

Hence, the quantity $\|\mathcal{A}(\partial_x\partial_y u_\mathcal{T}) - \partial_x\partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})}$ can be used as an a posteriori error estimator. For $p = 1$ and $\mathcal{T}$ consisting of squares or rectangles, we observe that $\partial_x\partial_y u_\mathcal{T}$ is piecewise constant on $\mathcal{T}$. Our hope is now that L^2-projecting $\partial_x\partial_y u_\mathcal{T}$ into the space of piecewise continuous linear functions on $\mathcal{T}$, i.e. defining $\mathcal{A}(\partial_x\partial_y u_\mathcal{T})$ via

$$\int_\mathcal{D} \mathcal{A}(\partial_x\partial_y u_\mathcal{T})v = \int_\mathcal{D} \partial_x\partial_y u_\mathcal{T} v, \quad \forall v \in \mathcal{S}^{1,0}(\mathcal{T}),$$

satisfies (2.49) for some $0 \leq \beta < 1$. However, computing this projection is as costly as computing the discrete solution itself and is thus not a viable option. To remedy the situation we resort to an approximation of the $L^2(\mathcal{D})$-inner product.

Denote by $\mathcal{W}_\mathcal{T}$ the space of all piecewise linear functions on $\mathcal{T}$, i.e. $\mathcal{W}_\mathcal{T} = \mathcal{P}_1(\mathcal{T}) = \{v : v|_K \in \mathcal{P}_1(K), K \in \mathcal{T}\}$, and set $\mathcal{V}_\mathcal{T} = \mathcal{W}_\mathcal{T} \cap C(\mathcal{D})$. Note that $\partial_x\partial_y\mathcal{P}_1(\mathcal{T}) \subset \mathcal{W}_\mathcal{T}$ and $\mathcal{V}_\mathcal{T} = \mathcal{S}^{1,0}(\mathcal{T})$. As the aforementioned approximation of the inner product, we define a mesh-dependent inner product $(\cdot,\cdot)_\mathcal{T}$ on $\mathcal{W}_\mathcal{T}$ via

$$(v, w)_\mathcal{T} = \sum_{K\in\mathcal{T}} \frac{|K|}{4}\left(\sum_{z\in\mathcal{N}_K} v|_K(z)w|_K(z)\right),$$

where $|K|$ denotes the two dimensional Lebesgue measure of K. This inner product is given by the tensorized version of the trapezoidal rule on an interval applied to each element $K \in \mathcal{T}$. Note, that for either v or w being a piecewise constant function we have

$$(v, w)_\mathcal{T} = \int_\mathcal{D} vw \,\mathrm{d}x.$$

With this in place, we let $\mathcal{A}(\partial_x\partial_y u_\mathcal{T})$ be the $(\cdot,\cdot)_\mathcal{T}$-projection of $\partial_x\partial_y u_\mathcal{T}$ onto $\mathcal{V}_\mathcal{T}$, i.e.

$$(\mathcal{A}(\partial_x\partial_y u_\mathcal{T}), v_\mathcal{T})_\mathcal{T} = (\partial_x\partial_y u_\mathcal{T}, v_\mathcal{T})_\mathcal{T}, \qquad \forall v_\mathcal{T} \in \mathcal{V}_\mathcal{T} \tag{2.50}$$

and define the elementwise error indicator and the global error estimator as follows.

Definition 2.42. *We define the elementwise error indicator as*

$$\eta_{\mathcal{A},K} := \|\partial_x\partial_y u_\mathcal{T} - \mathcal{A}(\partial_x\partial_y u_\mathcal{T})\|_{L^2(K)}$$

and the global error estimator as

$$\eta_\mathcal{A} = \left(\sum_{K\in\mathcal{T}} \eta_{\mathcal{A},K}^2\right)^{1/2}.$$

Note that for all $v, w \in \mathcal{V}_\mathcal{T}$ there holds

$$(v, w)_\mathcal{T} = \frac{1}{4} \sum_{z \in \mathcal{N}} |\omega_z| v(z) w(z) \tag{2.51}$$

and in lieu of (2.50), by inserting the nodal function for z in place of $v_\mathcal{T}$ we readily find for all $z \in \mathcal{N}$ the representation

$$\mathcal{A}(\partial_x \partial_y u_\mathcal{T})(z) = \sum_{K \subset \omega_z} \frac{|K|}{|\omega_z|} \partial_x \partial_y u_\mathcal{T}|_K, \tag{2.52}$$

which is the coefficient of the nodal basis function λ_z that is associated with the node z.

For any $q \in \mathcal{S}^{1,0}(\mathcal{T})$ that constitutes an approximation of $\partial_x \partial_y u_\mathcal{T}$ the triangle inequality readily shows efficiency of η_Z by

$$\eta_Z \leq \|\partial_x \partial_y u - \partial_x \partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})} + \|\partial_x \partial_y u - q\|_{L^2(\mathcal{D})} \tag{2.53}$$

and since q was arbitrary and $\mathcal{S}^{1,0}(\mathcal{T})$ is a finite dimensional subspace of $H^{1,1}(\mathcal{D})$ there even holds

$$\eta_Z \leq \|\partial_x \partial_y u - \partial_x \partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})} + \min_{q \in \mathcal{S}^{1,0}(\mathcal{T})} \|\partial_x \partial_y u - q\|_{L^2(\mathcal{D})}. \tag{2.54}$$

If one can now prove that indeed $\|\partial_x \partial_y u - q\|_{L^2(\mathcal{D})}$ is of higher order, the efficiency follows immediately. First we concentrate on proving the reliability. As we are emplyoing an averaging technique to $\partial_x \partial_y u_\mathcal{T}$ and compute the upper bound $\|\partial_x \partial_y u_\mathcal{T} - \mathcal{A}(\partial_x \partial_y u_\mathcal{T})\|_{L^2(\mathcal{D})}$ of η_Z we want to show that there holds

$$\|\partial_x \partial_y u - \partial_x \partial_y u_\mathcal{T}\|_{L^2(\mathcal{D})} \leq C\eta_Z + \mathrm{H} \leq C\|\partial_x \partial_y u_\mathcal{T} - \mathcal{A}(\partial_x \partial_y u_\mathcal{T})\|_{L^2(\mathcal{D})} + \mathrm{H}, \tag{2.55}$$

where H denotes a generic higher order error term.

Lemma 2.43. *Let $u, q \in H^{2,2}(\mathcal{D}) \cap H_0^{1,1}(\mathcal{D}) =: V(\mathcal{D})$ and $\partial_x \partial_y u_\mathcal{T} \in \mathcal{P}_0(\mathcal{T})$ the solution of the discrete problem (2.7) with*

$$\int_\mathcal{D} \partial_x \partial_y (u - u_\mathcal{T}) \partial_x \partial_y w_\mathcal{T} \, \mathrm{d}x = 0, \quad \forall w_\mathcal{T} \in \mathcal{S}^{1,0}(\mathcal{T}).$$

Furthermore, denote by $f_\mathcal{T}$ an elementwise approximation of f. Then there holds

$$\begin{aligned} |u - u_\mathcal{T}|_{H^{1,1}(\mathcal{D})} \leq & C \inf_{q \in V(\mathcal{D})} \left\{ \|\partial_x \partial_y (u_\mathcal{T} - q)\|_{L^2(\mathcal{D})} + \left(\sum_{K \in \mathcal{T}} h_K^2 \|f - \partial_x^2 \partial_y^2 q\|_{L^2(K)}^2 \right)^{1/2} \right\} \\ & + C \left(\sum_{K \in \mathcal{T}} h_K^2 \|f - f_\mathcal{T}\|_{L^2(K)}^2 \right)^{1/2}. \end{aligned} \tag{2.56}$$

Proof. Let $0 \neq w \in H_0^{1,1}(\mathcal{D})$ arbitrary. By Galerkin orthogonality, an application of integration by parts, the Cauchy-Schwarz inequality and approximation results from section

2.2.2 we find the assertion

$$
\begin{aligned}
&\int_{\mathcal{D}} \partial_x\partial_y(u-u_{\mathcal{T}})\partial_x\partial_y w\,\mathrm{d}x = \int_{\mathcal{D}} \partial_x\partial_y(u-u_{\mathcal{T}})\partial_x\partial_y(w-\Pi^{1,1}w)\,\mathrm{d}x\\
=&\int_{\mathcal{D}} \partial_x\partial_y(u-q)\partial_x\partial_y(w-\Pi^{1,1}w)\,\mathrm{d}x + \int_{\mathcal{D}} \partial_x\partial_y(q-u_{\mathcal{T}})\partial_x\partial_y(w-\Pi^{1,1}w)\,\mathrm{d}x\\
=&\sum_{K\in\mathcal{T}}\int_K (\partial_x^2\partial_y^2 u - f_{\mathcal{T}} - f + f_{\mathcal{T}} + f - \partial_x^2\partial_y^2 q)(w-\Pi^{1,1}w)\,\mathrm{d}x\\
&\qquad + \int_{\mathcal{D}} \partial_x\partial_y(q-u_{\mathcal{T}})\partial_x\partial_y(w-\Pi^{1,1}w)\,\mathrm{d}x\\
\leq& 2\sum_{K\in\mathcal{T}} \|f-f_{\mathcal{T}}\|_{L^2(K)}\|w-\Pi^{1,1}_{p_K}w\|_{L^2(K)} + \sum_{K\in\mathcal{T}} \|f-\partial_x^2\partial_y^2 q\|_{L^2(K)}\|w-\Pi^{1,1}_{p_K}w\|_{L^2(K)}\\
&\qquad + \|\partial_x\partial_y(u_{\mathcal{T}}-q)\|_{L^2(\mathcal{D})}\|\partial_x\partial_y(w-\Pi^{1,1}w)\|_{L^2(\mathcal{D})}\\
\leq& C\|\partial_x\partial_y(u_{\mathcal{T}}-q)\|_{L^2(\mathcal{D})}\|\partial_x\partial_y w\|_{L^2(\mathcal{D})}\\
&\qquad + \left\{ C\left(\sum_{K\in\mathcal{T}} h_K^2\|f-f_{\mathcal{T}}\|^2_{L^2(K)}\right)^{1/2} + C\left(\sum_{K\in\mathcal{T}} h_K^2\|f-\partial_x^2\partial_y^2 q\|^2_{L^2(K)}\right)^{1/2}\right\}\\
&\qquad\quad \times \underbrace{\left(\sum_{K\in\mathcal{T}}\left(\|\partial_x w\|^2_{L^2(K)} + \|\partial_y w\|^2_{L^2(K)} + |w|^2_{H^{1,1}(K)}\right)\right)^{1/2}}_{=\left(\|\partial_x w\|^2_{L^2(\mathcal{D})}+\|\partial_x w\|^2_{L^2(\mathcal{D})}+\|\partial_x\partial_y w\|^2_{L^2(\mathcal{D})}\right)^{1/2}\leq c\|\partial_x\partial_y w\|_{L^2(\mathcal{D})}}\\
\leq& C\inf_{q\in V(\mathcal{D})}\left\{\|\partial_x\partial_y(u_{\mathcal{T}}-q)\|_{L^2(\mathcal{D})} + \left(\sum_{K\in\mathcal{T}} h_K^2\|f-\partial_x^2\partial_y^2 q\|^2_{L^2(K)}\right)^{1/2}\right\}\|\partial_x\partial_y w\|_{L^2(\mathcal{D})}\\
&\qquad + C\left(\sum_{K\in\mathcal{T}} h_K^2\|f-f_{\mathcal{T}}\|^2_{L^2(\mathcal{D})}\right)^{1/2}\|\partial_x\partial_y w\|_{L^2(\mathcal{D})}.
\end{aligned}
$$

□

Theorem 2.44. *Let $\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}}) \in \mathcal{S}^{1,0}(\mathcal{T})$ be the average of $\partial_x\partial_y u_{\mathcal{T}}$ defined by (2.50)–(2.52). Furthermore, let u denote the solution of (2.6) and $u_{\mathcal{T}}$ the solution of (2.7). Then the error estimator $\eta_{\mathcal{A}}$ defined above is* reliable, *i.e. there holds*

$$
\begin{aligned}
|u-u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq& C\eta_{\mathcal{A}} + C\left(\sum_{K\in\mathcal{T}} h_K^2\|f-\partial_x\partial_y(\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}}))\|^2_{L^2(K)}\right)^{1/2}\\
&+ C\left(\sum_{K\in\mathcal{T}} h_K^2\|f-f_{\mathcal{T}}\|^2_{L^2(K)}\right)^{1/2}.
\end{aligned}
\tag{2.57}
$$

Proof. Considering that q is a polynomial of higher order than $u_{\mathcal{T}}$ which fulfills the Dirichlet boundary conditions and whose first mixed derivative is *globally* continuous by setting $\partial_x\partial_y q := \mathcal{A}(\partial_x\partial_y u_{\mathcal{T}})$, we note that $q \in H^{1,1}_0(\mathcal{D}) \cap H^{2,2}(\mathcal{D})$ and so we infer by

Lemma 2.43 that there holds

$$\begin{aligned}|u-u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq & C\|\partial_x\partial_y u_{\mathcal{T}} - \mathcal{A}(\partial_x\partial_y u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \\ & + C\left(\sum_{K\in\mathcal{T}} h_K^2 \|f - \partial_x\partial_y(\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}}))\|^2_{L^2(K)}\right)^{1/2} \\ & + C\left(\sum_{K\in\mathcal{T}} h_K^2 \|f - f_{\mathcal{T}}\|^2_{L^2(K)}\right)^{1/2} \\ = & C\eta_{\mathcal{A}} + C\left(\sum_{K\in\mathcal{T}} h_K^2 \|f - \partial_x\partial_y(\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}}))\|^2_{L^2(K)}\right)^{1/2} \\ & + C\left(\sum_{K\in\mathcal{T}} h_K^2 \|f - f_{\mathcal{T}}\|^2_{L^2(K)}\right)^{1/2}\end{aligned}$$

which yields the claim. □

Remark 2.45. *Note that the polynomial q is a purely theoretical tool and does not have to be determined explicitly.*

Remark 2.46. *Note that the term*

$$h_K^2 \|f - \partial_x\partial_y(\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}}))\|^2_{L^2(K)}$$

is akin to the data oscillation terms $h_K^2\|f - f_{\mathcal{T}}\|^2_{L^2(K)}$ as $\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}})$ is an approximation of $\partial_x\partial_y u$, which means that in some sense $\partial_x\partial_y(\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}}))$ is an approximation of f. Of course, the magnitude of this error is closely related to the averaging procedure used by the operator $\mathcal{A}$. Averaging operators of this type and of higher order are considered in [28].

Let us now turn our focus to the efficiency of $\eta_{\mathcal{A}}$, which we achieve by proving equivalence of $\eta_{\mathcal{A}}$ with η_Z.

Lemma 2.47. *There exists a uniform constant $b > 0$ such that*

$$\eta_Z \leq \eta_{\mathcal{A}} \leq \left(1 + \frac{4}{3}b\right)\eta_Z.$$

Proof. Since $\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}}) \in \mathcal{S}^{1,0}(\mathcal{T})$, it clearly holds that

$$\eta_Z = \min_{q\in\mathcal{S}^{1,0}(\mathcal{T})} \|\partial_x\partial_y u_{\mathcal{T}} - q\|_{L^2(\mathcal{D})} \leq \|\partial_x\partial_y u_{\mathcal{T}} - \mathcal{A}(\partial_x\partial_y u_{\mathcal{T}})\|_{L^2(\mathcal{D})} = \eta_{\mathcal{A}}.$$

In order to prove the upper bound we have a look at the L^2-stability of the averaging operator. First we note that by Cauchy-Schwarz for sums there holds

$$\|\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}})\|^2_{L^2(\mathcal{D})} = \int_{\mathcal{D}} \left|\sum_{z\in\mathcal{N}} \mathcal{A}(\partial_x\partial_y u_{\mathcal{T}})(z)\lambda_z\right|^2 \mathrm{d}x \leq 4\int_{\mathcal{D}} \sum_{z\in\mathcal{N}} |\mathcal{A}(\partial_x\partial_y u_{\mathcal{T}})(z)|^2|\lambda_z|^2 \,\mathrm{d}x.$$

If we now denote by M the element mass-matrix on K and by $\tilde{M}$ the element mass-matrix on K scaled by $|K|^{-1}$, i.e.

$$\tilde{M} = (\tilde{m}_{ij})_{i,j=1}^n \quad \text{with} \quad \tilde{m}_{ij} = |K|^{-1}\int_K \lambda_{z_i}\lambda_{z_j}\,\mathrm{d}x,$$

where $n = |\mathcal{N}_K|$ is the number of nodes of an element K and the indices realize a certain fixed enumeration of the nodes of K. Noting that for any $p = \sum_{z\in\mathcal{N}_K} p_z\lambda_z \in \mathcal{P}_1(K)$, where $p_z = p|_K(z)$, there holds

$$\|p\|^2_{L^2(K)} = \int_K \left(\sum_{z\in\mathcal{N}_K} p_z\lambda_z\right)^2 \mathrm{d}x = \sum_{i,j=1}^{n} p_{z_i}\cdot\int_K \lambda_{z_i}\lambda_{z_j}\,\mathrm{d}x\cdot p_{z_j} = \underline{p}^\top\cdot M\underline{p},$$

where by $\underline{p}$ we denote the vector of coefficients of p on K, i.e. $\underline{p} = (p_{z_1}, p_{z_2}, ..., p_{z_n})^\top$. Hence, by a Rayleigh quotient argument we find by letting $\tilde{\lambda}_1$ to be the smallest positive eigenvalue of $\tilde{M}$ that

$$\tilde{\lambda}_1|p_z|^2 \le \tilde{\lambda}_1\sum_{z\in\mathcal{N}_K}|p_z|^2 = \tilde{\lambda}_1\underline{p}^\top\cdot\underline{p} \le \underline{p}^\top\cdot\tilde{M}\underline{p} = |K|^{-1}\|p\|^2_{L^2(K)}.$$

In the two dimensional situation for parallelograms (cf. also [14]) we have that $\tilde{\lambda}_1 = \frac{1}{36}$ and hence, for any $v\in\mathcal{P}_1(K)$ there holds

$$|v|_K(z)|^2 \le \frac{36}{|K|}\|v\|^2_{L^2(K)} \tag{2.58}$$

and furthermore we have the existance of a uniform constant $b > 0$, such that

$$|v|_K(z)| \le \frac{b}{|\omega_z|^{1/2}}\|v\|_{L^2(\omega_z)}.$$

Thus, we find

$$\begin{aligned}
\|\mathcal{A}(\partial_x\partial_y u_\mathcal{T})\|^2_{L^2(\mathcal{D})} \le& 4\int_\mathcal{D}\sum_{z\in\mathcal{N}}|\mathcal{A}(\partial_x\partial_y u_\mathcal{T})(z)|^2|\lambda_z|^2\,\mathrm{d}x\\
\le& 4\sum_{z\in\mathcal{N}}\frac{b^2\|\partial_x\partial_y u_\mathcal{T}\|^2_{L^2(\omega_z)}}{|\omega_z|}\int_{\omega_z}|\lambda_z|^2\,\mathrm{d}x\\
=& 4\sum_{z\in\mathcal{N}}\frac{b^2\|\partial_x\partial_y u_\mathcal{T}\|^2_{L^2(\omega_z)}}{|\omega_z|}\frac{|\omega_z|}{9}\\
=& \frac{16b^2}{9}\|\partial_x\partial_y u_\mathcal{T}\|^2_{L^2(\mathcal{D})},
\end{aligned} \tag{2.59}$$

since $\|\lambda_z\|^2_{L^2(\omega_z)} = |\omega_z|/9$ and every element K appears four times.

Let $\partial_x\partial_y u_\mathcal{T}\in\mathcal{P}_0(\mathcal{T})$. Then there is a unique decomposition $\partial_x\partial_y u_\mathcal{T} = u_c + u_d$ with a continuous component $u_c\in\mathcal{S}^{1,0}(\mathcal{T})$ and a component u_d of the orthogonal complement of $\mathcal{S}^{1,0}(\mathcal{T})$ in $L^2(\mathcal{D})$. Note that for any $p\in\mathcal{S}^{1,0}(\mathcal{T})$ that $\mathcal{A}(p)(z) = p(z)$ and thus averaging is the identity on $\mathcal{S}^{1,0}(\mathcal{T})$. Hence,

$$\begin{aligned}
\|\partial_x\partial_y u_\mathcal{T} - \mathcal{A}(\partial_x\partial_y u_\mathcal{T})\|_{L^2(\mathcal{D})} =& \|(u_c+u_d) - \mathcal{A}(u_c+u_d)\|_{L^2(\mathcal{D})}\\
=& \|u_d - \mathcal{A}(u_d)\|_{L^2(\mathcal{D})}\\
\le& \|u_d\|_{L^2(\mathcal{D})} + \|\mathcal{A}(u_d)\|_{L^2(\mathcal{D})}\\
=& \left(1+\frac{4}{3}b\right)\|u_d\|_{L^2(\mathcal{D})}\\
=& \left(1+\frac{4}{3}b\right)\min_{q\in\mathcal{S}^{1,0}(\mathcal{T})}\|\partial_x\partial_y u_\mathcal{T} - q\|_{L^2(\mathcal{D})}\\
=& \left(1+\frac{4}{3}b\right)\eta_Z,
\end{aligned}$$

since $u_c \in \mathcal{S}^{1,0}(\mathcal{T})$. □

Remark 2.48 (Concerning b)**.** *For all $v \in \mathcal{P}_1(\mathcal{T}_z)$, where $\mathcal{T}_z := \{K : K \subset \omega_z\}$, by (2.58) there holds*

$$\begin{aligned}|\omega_z|^{1/2}|\mathcal{A}(\partial_x\partial_y u_\mathcal{T})(z)| \leq & 6 \sum_{K\in\mathcal{T}_z} \frac{|\omega_z|^{1/2}}{|K|^{1/2}} \frac{|K|}{|\omega_z|} \|\partial_x\partial_y u_\mathcal{T}\|_{L^2(K)} \\ \leq & 6 \left(\sum_{K\in\mathcal{T}_z} \frac{|\omega_z|}{|K|} \frac{|K|^2}{|\omega_z|^2} \right)^{1/2} \|\partial_x\partial_y u_\mathcal{T}\|_{L^2(\omega_z)}\end{aligned}$$

and thusly the constant $b > 0$ in (2.59) is

$$b = \max_{z\in\mathcal{N}} 6 \left(\sum_{K\in\mathcal{T}_z} \frac{|K|}{|\omega_z|} \right)^{1/2} = 6.$$

Theorem 2.49. *Let $\mathcal{A}(\partial_x\partial_y u_\mathcal{T}) \in \mathcal{S}^{1,0}(\mathcal{T})$ be the average of $\partial_x\partial_y u_\mathcal{T}$ defined by (2.50)–(2.52), where $u_\mathcal{T}$ is the solution of (2.7). Then the error estimator from Definition 2.42, i.e.*

$$\eta_\mathcal{A} = \|\partial_x\partial_y u_\mathcal{T} - \mathcal{A}(\partial_x\partial_y u_\mathcal{T})\|_{L^2(K)},$$

is asymptotically exact*.*

Proof. A combination of Lemma 2.43 and Lemma 2.47 readily implies that $\eta_\mathcal{A}$ is *asymptotically exact.* □

2.3 2D model problem

We now turn our attention to a two dimensional stochastic elliptic model problem and develop adaptive Finite Element methods for the second moment in the four dimensional setting.

2.3.1 Problem formulation

Letting again $a \equiv 1$, we consider now the 2D stochastic elliptic model problem, cf. (2.2), in $D = [-1, 1]^2$ of finding $u \in H_0^1(D)$, such that

$$\begin{aligned} -\Delta_x u(x,\omega) &= f(x,\omega) && \text{in } H^{-1}(D), \\ u &= 0 && \text{on } \partial D, \end{aligned} \tag{2.60}$$

for almost all $\omega \in \Omega$ and where Δ_x denotes the Laplace operator $\Delta_x = \partial_{x_1}^2 + \partial_{x_2}^2$.

With $V := H_0^1(D), V' = H^{-1}(D)$, as well as $A := -\Delta_x$ we find that the corresponding variational problem (cf.(1.4)) reads:

Problem 2.50. *Given $f(x,\omega) \in H^{-1}(D)$, find $u(x,\omega) \in H_0^1(D)$, such that*

$$\int_D \nabla_x u(x,\omega)\nabla_x v(x)\,\mathrm{d}x = \int_D f(x,\omega)v(x)\,\mathrm{d}x, \qquad \forall v \in H_0^1(D), \text{ for } \mathbb{P}\text{-a.a. } \omega \in \Omega.$$

The corresponding deterministic k-th moment problem, cf. (2.3), for the two dimensional model problem then takes the following form:

Problem 2.51. *Given $\mathcal{M}^k f \in H^{-1,\dots,-1}(D^k)$, find $\mathcal{M}^k u \in H_0^{1,\dots,1}(D^k)$, such that*

$$\int_{D^k} \nabla^{(k)}(\mathcal{M}^k u)\nabla^{(k)} v \,\mathrm{d}\mathbf{x} = \int_{D^k} \mathcal{M}^k f v \,\mathrm{d}\mathbf{x}, \quad \forall v \in H_0^{1,\dots,1}(D^k), \tag{2.61}$$

where $\mathbf{x} = (x_1, \dots, x_k)$ and

$$\nabla^{(k)} := \nabla_{x_1} \otimes \nabla_{x_2} \otimes \cdots \otimes \nabla_{x_k}$$

and by x_i denote the two dimensional coordinates of the i-th copy of D.

As before, for $k = 2$ we set $\mathcal{D} = D \times D$ and are interested in solving the deterministic second moment problem corresponding to (2.3):

Given $\mathcal{C}_f \in H^{-1,-1}(\mathcal{D})$, find $\mathcal{C}_u \in H_0^{1,1}(\mathcal{D})$, such that

$$\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y)\mathcal{C}_u(\nabla_x \otimes \nabla_y) v \,\mathrm{d}x \,\mathrm{d}y = \int_{\mathcal{D}} \mathcal{C}_f v \,\mathrm{d}x \,\mathrm{d}y, \quad \forall v \in H_0^{1,1}(\mathcal{D}). \tag{2.62}$$

The discrete version of (2.62) takes the form:

Given $\mathcal{C}_f \in H^{-1,-1}(\mathcal{D})$, find $u_\mathcal{T} \in \mathcal{S}_0^{p,0}(\mathcal{T})$, such that

$$\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) u_\mathcal{T} (\nabla_x \otimes \nabla_y) v_\mathcal{T} \,\mathrm{d}x \,\mathrm{d}y = \int_{\mathcal{D}} \mathcal{C}_f v_\mathcal{T} \,\mathrm{d}x \,\mathrm{d}y, \quad \forall v_\mathcal{T} \in \mathcal{S}_0^{p,0}(\mathcal{T}). \tag{2.63}$$

We shall again write $u_\mathcal{T}$ instead of $\mathcal{C}_{u,\mathcal{T}}$ to alleviate the notation. Note for $p \in \mathbb{N}$, since $\mathcal{S}_0^{p,0}(\mathcal{T}) \subset H_0^{1,1}(\mathcal{D})$, that the approximation is conforming and we have Galerkin orthogonality

$$\mathcal{B}(u - u_\mathcal{T}, v_\mathcal{T}) = 0, \quad \forall v_\mathcal{T} \in \mathcal{S}_0^{p,0}(\mathcal{T})$$

with the associated bilinear form

$$\mathcal{B}(u, v) = \int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) u (\nabla_x \otimes \nabla_y) v \,\mathrm{d}x \,\mathrm{d}y.$$

We also note that the energy norm of this problem is as for the one dimensional model problem the $|\cdot|_{H^{1,1}(\mathcal{D})}$-seminorm. The latter is a norm on $H_0^{1,1}(\mathcal{D})$ by the crossnorm property and corresponding two dimensional Friedrichs' inequalities.

2.3.2 Approximation and auxiliary results

In the four dimensional setting, since for $D \subset \mathbb{R}^2$ there holds

$$H^1(D) \not\subset C(D),$$

we have to use another operator to analyze our adaptive Finite Element procedure. We make use of the following tensorized quasi-interpolation operator $\mathbb{I}_\mathcal{T} : L^1(\mathcal{D}) \to S_0^{1,0}(\mathcal{T})$ for any $v \in L^1(\mathcal{D})$ defined by

$$\mathbb{I}_\mathcal{T} := \mathcal{I}_x \otimes \mathcal{I}_y, \tag{2.64}$$

where the two dimensional quasi-interpolation operators $\mathcal{I}_x$ and $\mathcal{I}_y$ are given by the expression

$$\mathcal{I}_\nu v := \sum_{z_\nu \in \mathcal{N}_\nu} \overline{v}_{\omega_{z_\nu}} \lambda_{z_\nu}, \tag{2.65}$$

where $\nu \in \{x, y\}$, λ_{z_ν} denotes the nodal shape function with respect to the node z_ν, and ω_{z_ν} is the node patch around z_ν in the mesh restricted to the ν-coordinates. Furthermore, $\mathcal{N}_\nu$ denotes the set of nodes in the ν-coordinates and $\mathcal{N}_{\nu,\partial D}$ denotes the nodes on the boundary of the copy of D in the ν-coordinates.

Here we choose $\overline{v}_{\omega_{z_\nu}}$ in the following way (cf. also [54, Section 3.5] or [16])

$$\overline{v}_{\omega_{z_\nu}} = \begin{cases} \dfrac{\int_{\omega_{z_\nu}} \lambda_{z_\nu} v \,\mathrm{d}\nu}{\int_{\omega_{z_\nu}} \lambda_{z_\nu} \,\mathrm{d}\nu}, & z_\nu \in \mathcal{N}_\nu, \\ 0, & z_\nu \in \mathcal{N}_{\nu,\partial D}. \end{cases}$$

Let us check that $\mathbb{I}_\mathcal{T}$ as such is well defined. Let $v \in L^1(\mathcal{D})$, then

$$\begin{aligned}
\mathbb{I}_\mathcal{T} v =& (\mathcal{I}_x \otimes \mathcal{I}_y)v = (\mathcal{I}_x \otimes \mathrm{id})((\mathrm{id} \otimes \mathcal{I}_y)v) \\
=& (\mathcal{I}_x \otimes \mathrm{id}) \left\{ \sum_{z_y \in \mathcal{N}_y} \left(\frac{\int_{\omega_{z_y}} \lambda_{z_y} v(x,y) \,\mathrm{d}y}{\int_{\omega_{z_y}} \lambda_{z_y} \,\mathrm{d}y} \right) \cdot \lambda_{z_y} \right\} \\
=& \sum_{z_y \in \mathcal{N}_y} \left\{ \frac{\int_{\omega_{z_y}} \lambda_{z_y} \left(\sum_{z_x \in \mathcal{N}_x} \left[\frac{\int_{\omega_{z_x}} \lambda_{z_x} v(x,y) \,\mathrm{d}x}{\int_{\omega_{z_x}} \lambda_{z_x} \,\mathrm{d}x} \right] \cdot \lambda_{z_x} \right) \,\mathrm{d}y}{\int_{\omega_{z_y}} \lambda_{z_y} \,\mathrm{d}y} \right\} \cdot \lambda_{z_y} \\
=& \sum_{z_x \in \mathcal{N}_x} \sum_{z_y \in \mathcal{N}_y} \left(\frac{\int_{\omega_{z_x}} \int_{\omega_{z_y}} (\lambda_{z_x} \otimes \lambda_{z_y}) v(x,y) \,\mathrm{d}y \,\mathrm{d}x}{\int_{\omega_{z_x}} \int_{\omega_{z_y}} \lambda_{z_x} \otimes \lambda_{z_y} \,\mathrm{d}y \,\mathrm{d}x} \right) \cdot \lambda_{z_x} \otimes \lambda_{z_y} \\
=& \sum_{z \in \mathcal{N}} \overline{v}_{\omega_z} \lambda_z,
\end{aligned}$$

where now $z = (z_x, z_y)$, $\omega_z = \omega_{z_x} \times \omega_{z_y}$ and $\lambda_z = \lambda_{z_x} \otimes \lambda_{z_y}$. The previous considerations show that $\mathbb{I}_\mathcal{T}$ is well defined and that the neighborhood of a node z is the product of the respective neighborhoods of nodes of the tensor factors.

The next lemma gives local estimates for the two dimensional quasi-interpolation operators $\mathcal{I}_x$ and $\mathcal{I}_y$ for which a proof can be found in [54, Section 3.5].

Lemma 2.52 (Two dimensional quasi-interpolation)**.** *Let $p \in [1, \infty]$. Denote by $\mathcal{I}_\mathcal{T}$ the quasi-interpolation operator according to (2.65) on the two dimensional domain D. For any element $K \in \mathcal{T}$, edge $E \in \mathcal{E}_K$, and $v \in W_0^{1,p}(D)$ there hold the estimates*

$$\begin{aligned}
\|v - \mathcal{I}_\mathcal{T} v\|_{L^p(K)} \leq& C \|v\|_{L^p(\tilde{\omega}_K)}, \\
\|v - \mathcal{I}_\mathcal{T} v\|_{L^p(K)} \leq& C h_K \|\nabla v\|_{L^p(\tilde{\omega}_K)}, \\
\|\nabla(v - \mathcal{I}_\mathcal{T} v)\|_{L^p(K)} \leq& C \|\nabla v\|_{L^p(\tilde{\omega}_K)}, \\
\|v - \mathcal{I}_\mathcal{T} v\|_{L^p(E)} \leq& C h_E^{1-1/p} \|\nabla v\|_{L^p(\tilde{\omega}_E)}.
\end{aligned}$$

Equipped with these estimates we derive corresponding estimates for the tensorized quasi-interpolation operator $\mathbb{I}_\mathcal{T}$.

Lemma 2.53 (Quasi-Interpolation Error Estimates)**.** *Let K an element in $\mathcal{T}$ and $v \in H_0^{1,1}(K)$. Then there holds*

$$\begin{aligned}
\|v - \mathbb{I}_\mathcal{T} v\|_{L^2(K)} \leq& C h_K \left(\|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_K)} + \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} + |v|_{H^{1,1}(\tilde{\omega}_K)} \right), \\
\|v - \mathbb{I}_\mathcal{T} v\|_{L^2(Q)} \leq& C h_Q^{1/2} \left(\|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_K)} + \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} + |v|_{H^{1,1}(\tilde{\omega}_K)} \right), \\
\|v - \mathbb{I}_\mathcal{T} v\|_{L^2(F)} \leq& C h_F^{1/2} \left(\|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_K)} + \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} + |v|_{H^{1,1}(\tilde{\omega}_K)} \right).
\end{aligned}$$

Proof. First we note that $K = K_x \times K_y$. By letting $w := v - (\mathrm{id} \otimes \mathcal{I}_y)v$ and $w' := v - (\mathcal{I}_x \otimes \mathrm{id})v$ we find the expressions

$$\begin{aligned} v - \mathbb{I}_{\mathcal{T}} v =& v - (\mathcal{I}_x \otimes \mathrm{id})v + (\mathcal{I}_x \otimes \mathrm{id})(v - (\mathrm{id} \otimes \mathcal{I}_y)v) \\ =& v - (\mathcal{I}_x \otimes \mathrm{id})v - w + w - (\mathcal{I}_x \otimes \mathrm{id})w \\ =& v - (\mathrm{id} \otimes \mathcal{I}_y)v - w' + w' - (\mathrm{id} \otimes \mathcal{I}_y)w'. \end{aligned}$$

Thus, we have

$$\begin{aligned} \|v - \mathbb{I}_{\mathcal{T}} v\|_{L^2(K)} \leq& \|v - (\mathcal{I}_x \otimes \mathrm{id})v\|_{L^2(K)} + \|v - (\mathrm{id} \otimes \mathcal{I}_y)v\|_{L^2(K)} + \|w - (\mathcal{I}_x \otimes \mathrm{id})w\|_{L^2(K)} \\ \leq& Ch_K \|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_{K_x} \times K_y)} + Ch_K \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(K_x \times \tilde{\omega}_{K_y})} \\ &+ Ch_K \|(\nabla_x \otimes \mathrm{id})v - (\mathrm{id} \otimes \mathcal{I}_y)((\nabla_x \otimes \mathrm{id})v)\|_{L^2(\tilde{\omega}_{K_x} \times K_y)} \\ \leq& Ch_K \|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_K)} + Ch_K \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} \\ &+ Ch_K^2 \|(\nabla_x \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)}, \end{aligned}$$

where we have invoked the triangle inequality twice and made use of the properties of the two dimensional operators $\mathcal{I}_x$ and $\mathcal{I}_y$ from Lemma 2.52. Without loss of generality let $Q = K_x \times E_y$, where E_y is supposed to be an edge of K_y (cf. 2.2), w as above, and by applying the triangle inequality and Lemma 2.52, we arrive at

$$\begin{aligned} \|v - \mathbb{I}_{\mathcal{T}} v\|_{L^2(Q)} \leq& \|v - (\mathcal{I}_x \otimes \mathrm{id})v\|_{L^2(Q)} + \|v - (\mathrm{id} \otimes \mathcal{I}_y)v\|_{L^2(Q)} + \|w - (\mathcal{I}_x \otimes \mathrm{id})w\|_{L^2(Q)} \\ \leq& Ch_{K_x} \|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_{K_x} \times E_y)} + Ch_{E_y}^{1/2} \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(K_x \times \tilde{\omega}_{E_y})} \\ &+ Ch_{K_x} \|(\nabla_x \otimes \mathrm{id})w\|_{L^2(\tilde{\omega}_{K_x} \times E_y)}. \end{aligned}$$

Applying the two dimensional trace inequality

$$\|u\|_{L^2(E)} \leq C \left(h_K^{-1/2} \|u\|_{L^2(K)} + h_K^{1/2} \|\nabla u\|_{L^2(K)} \right)$$

in the y-direction, which is valid for any $u \in W^{1,2}(K)$, we find by invoking shape regularity $h_{K_y} \leq ch_Q$, $h_{K_x} \leq ch_Q$ and $h_{E_y} \leq ch_Q$

$$\begin{aligned} \|v - \mathbb{I}_{\mathcal{T}} v\|_{L^2(Q)} \leq& Ch_{K_x} \left\{ ch_{K_y}^{-1/2} \|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_{K_x} \times K_y)} + h_{K_y}^{1/2} \|(\nabla_x \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_{K_x} \times K_y)} \right\} \\ &+ Ch_Q^{1/2} \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(K_x \times \tilde{\omega}_{E_y})} + Ch_{K_x} \|(\nabla_x \otimes \mathrm{id})w\|_{L^2(\tilde{\omega}_{K_x} \times E_y)} \\ \leq& Ch_Q^{1/2} \left\{ \|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_K)} + \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} \right\} \\ &+ Ch_Q^{3/2} \|(\nabla_x \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)}, \end{aligned}$$

where we have used Lemma 2.52 again in the last step. Finally, we note for $F = E_x \times E_y$ by the same arguments as above that

$$\begin{aligned} \|v - \mathbb{I}_{\mathcal{T}} v\|_{L^2(F)} \leq& Ch_{E_x}^{1/2} \|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_{E_x} \times E_y)} + Ch_{E_y}^{1/2} \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(E_x \times \tilde{\omega}_{E_y})} \\ &+ Ch_{E_x}^{1/2} h_{E_y}^{1/2} \|(\nabla_x \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} \\ \leq& Ch_F^{1/2} \left\{ \|(\nabla_x \otimes \mathrm{id})v\|_{L^2(\tilde{\omega}_K)} + \|(\nabla_x \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} \right\} \\ &+ Ch_F^{1/2} \left\{ \|(\mathrm{id} \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} + \|(\nabla_x \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)} \right\} \\ &+ Ch_F \|(\nabla_x \otimes \nabla_y)v\|_{L^2(\tilde{\omega}_K)}, \end{aligned}$$

where we have used again shape regularity and the trace inequality $\|v\|_{L^2(\partial K)} \leq C\|v\|_{H^1(K)}$, where C only depends on the shape of the element but not on its size, in the x- and y-direction, respectively. This concludes the proof. □

Remark 2.54. *As the operators $\mathcal{I}_\nu, \nu \in \{x, y\}$ do only allow to bound the H^1-seminorm by a constant times the H^2-seminorm without a factor of h_K, and thus we cannot prove convergence of the FEM with the operator $\mathbb{I}_\mathcal{T}$ directly. We therefore resort to the operator*

$$\mathbf{I}_\mathcal{T} := \mathcal{I}^{\mathrm{av}}_{h0} \otimes \mathcal{I}^{\mathrm{av}}_{h0}$$

with $\mathcal{I}^{\mathrm{av}}_{h0}$ from [28, Chapter 6]. This operator respects the homogeneous Dirichlet boundary conditions, satisfies bounds in the form of those of Lemma 2.52 and additionally admits the following bound for $v \in H^2(\tilde{\omega}_K)$ and polynomial degree $p = 1$ (cf. [28, Thm. 6.4])

$$|v - \mathcal{I}^{\mathrm{av}}_{h0} v|_{H^1(K)} \leq c h_K |v|_{H^2(\tilde{\omega}_K)}$$

as well as

$$|v - \mathcal{I}^{\mathrm{av}}_{h0} v|_{H^1(K)} \leq c |v|_{H^1(\tilde{\omega}_K)}.$$

In short words, this operator is constructed as the concatenation of a quasi-interpolation operator on $L^1(K)$ and an averaging operator, for more details see [28].

Lemma 2.55. *Let $u \in H^3(\mathcal{D}) \cap H^{1,1}_0(\mathcal{D})$*

$$\|(\nabla_x \otimes \nabla_y)(u - \boldsymbol{I}_\mathcal{T} u)\|_{L^2(\mathcal{D})} \leq c h |u|_{H^3(\mathcal{D})}.$$

Proof. Abusing the notation let us denote for brevity

$$\mathbf{I}_\mathcal{T} := \mathcal{I}^{\mathrm{av}}_x \otimes \mathcal{I}^{\mathrm{av}}_y := \mathcal{I}^{\mathrm{av}}_{h0} \otimes \mathcal{I}^{\mathrm{av}}_{h0}.$$

Then with $w := u - (\mathrm{id} \otimes \mathcal{I}^{\mathrm{av}}_y)u$ we see that there holds

$$\begin{aligned}
&\|(\nabla_x \otimes \nabla_y)(u - \mathbf{I}_\mathcal{T} u)\|_{L^2(K)} \\
=&\|(\nabla_x \otimes \nabla_y)(u - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})u - w + w - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})w)\|_{L^2(K)} \\
\leq&\|(\nabla_x \otimes \nabla_y)(u - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})u)\|_{L^2(K)} + \|(\nabla_x \otimes \nabla_y)(u - (\mathrm{id} \otimes \mathcal{I}^{\mathrm{av}}_y)u)\|_{L^2(K)} \\
&\quad + \|(\nabla_x \otimes \nabla_y)(w - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})w)\|_{L^2(K)} \\
=&\|(\nabla_x \otimes \mathrm{id})\left((\mathrm{id} \otimes \nabla_y)u - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})(\mathrm{id} \otimes \nabla_y)u\right)\|_{L^2(K)} \\
&\quad + \|(\mathrm{id} \otimes \nabla_y)((\nabla_x \otimes \mathrm{id})u - (\mathrm{id} \otimes \mathcal{I}^{\mathrm{av}}_y)(\nabla_x \otimes \mathrm{id})u)\|_{L^2(K)} \\
&\quad + \|(\nabla_x \otimes \mathrm{id})((\mathrm{id} \otimes \nabla_y)w - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})(\mathrm{id} \otimes \nabla_y)w)\|_{L^2(K)}.
\end{aligned}$$

By means of Remark 2.54 we are allowed to bound the H^1-seminorms of the last equality, where we set for brevity $u_x := (\nabla_x \otimes \mathrm{id})u$ and analogously $u_y := (\mathrm{id} \otimes \nabla_y)u$,

$$\|(\nabla_x \otimes \mathrm{id})\left(u_y - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})u_y\right)\|_{L^2(K)} \leq c h_{K_x} |u_y|_{H^{2,0}(\tilde{\omega}_{K_x} \times K_y)}$$

and similarly for the remaining terms

$$\begin{aligned}
\|(\mathrm{id} \otimes \nabla_y)\left(u_x - (\mathrm{id} \otimes \mathcal{I}^{\mathrm{av}}_y)u_x\right)\|_{L^2(K)} \leq& c h_{K_y} |u_x|_{H^{0,2}(K_x \times \tilde{\omega}_{K_y})}, \\
\|(\nabla_x \otimes \mathrm{id})\left(w_y - (\mathcal{I}^{\mathrm{av}}_x \otimes \mathrm{id})w_y\right)\|_{L^2(K)} \leq& c h_{K_y} |w_y|_{H^{2,0}(\tilde{\omega}_{K_x} \times K_y)}.
\end{aligned}$$

Using the second inequality of Remark 2.54 for the last term gives

$$|w_y|_{H^{2,0}(\tilde{\omega}_{K_x} \times K_y)} \leq C \|u\|_{H^{2,1}(\tilde{\omega}_{K_x} \times \tilde{\omega}_{K_y})}$$

whence we have by shape regularity

$$\begin{aligned}\|(\nabla_x \otimes \nabla_y)(u - \mathbf{I}_{\mathcal{T}} u)\|_{L^2(K)} \leq & ch_{K_x} |u|_{H^{2,1}(\tilde{\omega}_{K_x} \times K_y)} + ch_{K_y} |u|_{H^{1,2}(K_x \times \tilde{\omega}_{K_y})} \\ & + ch_{K_y} |u|_{H^{2,1}(\tilde{\omega}_{K_x} \times \tilde{\omega}_{K_y})} \\ \leq & ch_K |u|_{H^3(\tilde{\omega}_K)}.\end{aligned}$$

Thus, we find that

$$\begin{aligned}\|(\nabla_x \otimes \nabla_y)(u - \mathbf{I}_{\mathcal{T}} u)\|^2_{L^2(\mathcal{D})} &= \sum_{K \in \mathcal{T}} \|(\nabla_x \otimes \nabla_y)(u - \mathbf{I}_{\mathcal{T}} u)\|^2_{L^2(K)} \\ &\leq \sum_{K \in \mathcal{T}} ch_K^2 |u|^2_{H^3(\tilde{\omega}_K)} \\ &\leq ch^2 \sum_{K \in \mathcal{T}} |u|^2_{H^3(\tilde{\omega}_K)} = ch^2 |u|^2_{H^3(\mathcal{D})}\end{aligned}$$

which with $h := \sup_{K \in \mathcal{T}} h_K$ upon taking square roots yields the claim. □

Combining the previous remark and lemma we find the following

Lemma 2.56. *Let $u \in H_0^{1,1}(\mathcal{D}) \cap H^3(\mathcal{D})$ be the unique solution of the deterministic second moment problem (2.62) and denote by $u_{\mathcal{T}}$ the corresponding discrete solution of (2.63). Then there holds*

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq ch |u|_{H^3(\mathcal{D})},$$

where $h := \sup_{K \in \mathcal{T}} h_K$.

Proof. As the bilinear form of the deterministic second moment problem is symmetric, continuous and coercive, we have by the Lax-Milgram lemma the unique solvability of the problem and Céa's lemma guarantees that our solution fulfills the best approximation property

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq c \inf_{v_{\mathcal{T}} \in \mathcal{S}_0^{1,0}(\mathcal{T})} |u - v_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})}.$$

Making the particular choice $v_{\mathcal{T}} := \mathbf{I}_{\mathcal{T}} u$ and using Lemma 2.55 shows that

$$\begin{aligned}|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq & c \inf_{v_{\mathcal{T}} \in \mathcal{S}_0^{1,0}(\mathcal{T})} |u - v_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \\ \leq & c |u - \mathbf{I}_{\mathcal{T}} u|_{H^{1,1}(\mathcal{D})} \\ \leq & ch |u|_{H^3(\mathcal{D})}\end{aligned}$$

which is the claim. □

For the analysis of the a posteriori error estimators to come we need, as in the two dimensional situation, certain norm equivalences over spaces of polynomials on the geometric entities present in the mesh. To this end the next lemma is necessary.

Lemma 2.57 (Polynomial inverse inequalities)**.** *Let $u \in \mathcal{P}_p(F), w \in \mathcal{P}_p(Q)$ and $v \in \mathcal{P}_p(K)$ be polynomials of order p, where F, Q are a boundary face and cube, respectively, of the element K. Denote by ψ_F, ψ_Q, and ψ_K the face-, cube- or element bubble functions,*

respectively. Then there holds

$$\begin{aligned}
\|u\psi_F\|^2_{L^2(Q)} &\leq Ch_F\|u\|^2_{L^2(F)},\\
\|u\psi_F\|^2_{L^2(K)} &\leq Ch_F^2\|u\|^2_{L^2(F)},\\
\|w\psi_Q\|^2_{L^2(K)} &\leq Ch_Q\|w\|^2_{L^2(Q)},\\
\|v\psi_K\|^2_{L^2(K)} &\leq C\|v\|^2_{L^2(K)},\\
\|(\nabla_x\otimes\nabla_y)(u\psi_F)\|^2_{L^2(K)} &\leq Ch_F^{-2}\|u\|^2_{L^2(F)},\\
\|(\nabla_x\otimes\nabla_y)(w\psi_Q)\|^2_{L^2(K)} &\leq Ch_Q^{-3}\|w\|^2_{L^2(Q)},\\
\|(\nabla_x\otimes\nabla_y)(v\psi_K)\|^2_{L^2(K)} &\leq Ch_K^{-4}\|v\|^2_{L^2(K)}.
\end{aligned}$$

Proof. Let us denote $\hat{x} = (\hat{x}_1, \hat{x}_2)^\top, \hat{y} = (\hat{y}_1, \hat{y}_2)^\top$ and $\hat{\xi} = (\hat{x}, \hat{y})^\top$ as well as $\xi = (x, y)^\top = (x_1, x_2, y_1, y_2)^\top$. Without loss of generality let F be the image under F_K of the face $\hat{F} = [-1, 1] \times [-1, 1] \times \{-1\} \times \{-1\}$. The result follows for the other faces F of K by a rotation of coordinates. Since for such $\hat{F}$ we have $\psi_{\hat{F}} = \frac{(1-\hat{y}_1)}{2}\frac{(1-\hat{y}_2)}{2}(1-\hat{x}_1^2)(1-\hat{x}_2^2)$ and noting that $u \in \mathcal{P}_k(F)$, we have for any $Q \subset \partial K$ with $F \subset Q$ that

$$\|u\psi_F\|^2_{L^2(Q)} = \frac{2}{3}\frac{h_xh_yh_z}{8}\int_{\hat{F}}(\hat{u}\psi_{\hat{F}})^2\,\mathrm{d}\hat{x} \leq Ch_F\|u\|^2_{L^2(F)}.$$

as well as any $F \subset Q \subset \partial K$ that

$$\|u\psi_F\|^2_{L^2(K)} = \frac{4}{9}\frac{h_xh_yh_zh_w}{16}\int_{\hat{F}}(\hat{u}\psi_{\hat{F}})^2\,\mathrm{d}\hat{x} \leq Ch_F^2\|u\|^2_{L^2(F)}.$$

Similarly, we find

$$\|w\psi_Q\|^2_{L^2(K)} = \frac{2}{3}\frac{h_xh_yh_zh_w}{16}\int_{\hat{Q}}(\hat{w}\psi_{\hat{Q}})^2\,\mathrm{d}\hat{x} \leq Ch_Q\|w\|^2_{L^2(Q)}$$

and since $\|\psi^2_{\hat{K}}\|_{L^\infty(\hat{K})} = 1$,

$$\|v\psi_K\|^2_{L^2(K)} \leq \|v\|^2_{L^2(K)}.$$

For the remaining estimates we note that for any $u \in H^{1,1}(K)$ there holds

$$\begin{aligned}
\|(\nabla_x\otimes\nabla_y)u\|^2_{L^2(K)} &= \frac{h_xh_yh_zh_w}{16}\int_{\hat{K}}((\nabla_x\otimes\nabla_y)u(F_K(\hat{\xi}))^2\,\mathrm{d}\hat{\xi}\\
&= \sum_{i,j=1}^{2}\frac{h_xh_yh_zh_w}{16}\int_{\hat{K}}\left(\frac{2}{h_{x_i}}\frac{2}{h_{y_j}}\partial_{\hat{x}_i}\partial_{\hat{y}_j}\hat{u}\right)^2\,\mathrm{d}\hat{\xi}.
\end{aligned}$$

Then abbreviating $c_K^{i,j} = \frac{h_xh_yh_zh_w}{16}\frac{4}{h_{x_i}^2}\frac{4}{h_{y_j}^2}$ and considering $v \in \mathcal{P}_p(K)$ we find

$$\|(\nabla_x\otimes\nabla_y)(v\psi_K)\|^2_{L^2(K)} = \sum_{i,j=1}^{2}c_K^{i,j}\int_{\hat{K}}\left(\partial_{\hat{x}_i}\partial_{\hat{y}_j}(\hat{v}\psi_{\hat{K}})\right)^2\,\mathrm{d}\hat{x}\,\mathrm{d}\hat{y}$$

$$\begin{aligned}
= \sum_{i,j=1}^{2}c_K^{i,j}\int_{\hat{K}}\big[&\partial_{\hat{x}_i}((1-\hat{x}_1^2)(1-\hat{x}_2^2))\partial_{\hat{y}_j}(\hat{v}(1-\hat{y}_1^2)(1-\hat{y}_2^2))\\
&+ (1-\hat{x}_1^2)(1-\hat{x}_2^2)\partial_{\hat{x}_i}\partial_{\hat{y}_j}(\hat{v}(1-\hat{y}_1^2)(1-\hat{y}_2^2))\big]^2\,\mathrm{d}\hat{x}\,\mathrm{d}\hat{y}
\end{aligned}$$

$$
\begin{aligned}
\leq & 2\sum_{i,j=1}^{2} c_K^{i,j}\left\{\int_{\hat{K}}\left[\partial_{\hat{x}_i}((1-\hat{x}_1^2)(1-\hat{x}_2^2))\partial_{\hat{y}_j}(\hat{v}(1-\hat{y}_1^2)(1-\hat{y}_2^2))\right]^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y}\right. \\
& \left. + \int_{\hat{K}}\left[(1-\hat{x}_1^2)(1-\hat{x}_2^2)\partial_{\hat{x}_i}\partial_{\hat{y}_j}(\hat{v}(1-\hat{y}_1^2)(1-\hat{y}_2^2))\right]^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y}\right\} \\
= & 2\sum_{i,j=1}^{2} c_K^{i,j}(J_{i,j}^1 + J_{i,j}^2).
\end{aligned}
$$

We proceed by bounding the terms $J_{i,j}^{\ell}, \ell = 1,2$ on the right-hand side. Using the one dimensional inequality involving $\psi_{[-1,1]}$ from Lemma 2.11 we find for $i = 1,2$ and s the complementary index, i.e. letting $s = 1$ if $i = 2$ and $s = 2$ if $i = 1$,

$$
\begin{aligned}
J_{i,j}^1 &= \int_{\hat{K}} \left(\partial_{\hat{x}_i}((1-\hat{x}_i^2)(1-\hat{x}_s^2))\right)^2 \left(\partial_{\hat{y}_j}(\hat{v}(1-\hat{y}_1^2)(1-\hat{y}_2^2))\right)^2 \\
&= \int_{\hat{x}_s=-1}^{1}(1-\hat{x}_s^2)^2 \int_{\hat{x}_i=-1}^{1}(\partial_{\hat{x}_i}(1-\hat{x}_i^2))^2 \underbrace{\int_{\hat{y}_1=-1}^{1}\int_{\hat{y}_2=-1}^{1}\left(\partial_{\hat{y}_j}(\hat{v}(1-\hat{y}_1^2)(1-\hat{y}_2^2))\right)^2}_{\leq C(4+2\sqrt{p(p+1)})^2\|\hat{v}(\hat{x}_1,\hat{x}_2,\cdot,\cdot)\|^2_{L^2([-1,1]^2)}} \\
&\leq C\left(4+2\sqrt{p(p+1)}\right)^2 \|\hat{v}\|^2_{L^2(\hat{K})}
\end{aligned}
$$

and using arguments similar to Lemma 2.13 and letting s now denote the complementary index to j

$$
\begin{aligned}
J_{i,j}^2 &= \int_{\hat{K}}(1-\hat{x}_1^2)^2(1-\hat{x}_2^2)^2\left(\partial_{\hat{y}_j}\left[(\partial_{\hat{x}_i}\hat{v})(1-\hat{y}_1^2)(1-\hat{y}_2^2)\right]\right)^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y} \\
&\leq \left(4+2\sqrt{p(p+1)}\right)^2 \int_{\hat{K}}(1-\hat{x}_1^2)^2(1-\hat{x}_2^2)^2(1-\hat{y}_s^2)^2(\partial_{\hat{x}_i}\hat{v})^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y} \\
&\leq \left(4+2\sqrt{p(p+1)}\right)^2 p(p+1)\|\hat{v}\|^2_{L^2(\hat{K})}.
\end{aligned}
$$

Transforming back to the element K and adding up the estimates for $J_{i,j}^1$ and $J_{i,j}^2$ yields the claim

$$
\|(\nabla_x \otimes \nabla_y)(v\psi_K)\|^2_{L^2(K)} \leq C h_K^{-4}\|v\|^2_{L^2(K)}.
$$

For the remaining inequalities we proceed analogously. Letting $\hat{Q} = [-1,1]^3 \times \{-1\}$ there holds

$$
\begin{aligned}
\|(\nabla_x \otimes \nabla_y)(w\psi_Q)\|^2_{L^2(K)} &= \sum_{i,j=1}^{2} c_K^{i,j}\int_{\hat{K}}\left(\partial_{\hat{x}_i}\partial_{\hat{y}_j}(\hat{w}\psi_{\hat{Q}})\right)^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y} \\
= & \sum_{i,j=1}^{2} c_K^{i,j}\int_{\hat{K}}\left[\partial_{\hat{x}_i}((1-\hat{x}_1^2)(1-\hat{x}_2^2))\partial_{\hat{y}_j}\left(\hat{w}(1-\hat{y}_1^2)\frac{(1-\hat{y}_2)}{2}\right)\right. \\
& \left. + (1-\hat{x}_1^2)(1-\hat{x}_2^2)\partial_{\hat{x}_i}\partial_{\hat{y}_j}\left(\hat{v}(1-\hat{y}_1^2)\frac{(1-\hat{y}_2)}{2}\right)\right]^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y} \\
\leq & 2\sum_{i,j=1}^{2} c_K^{i,j}\left\{\int_{\hat{K}}\left[\partial_{\hat{x}_i}((1-\hat{x}_1^2)(1-\hat{x}_2^2))\partial_{\hat{y}_j}\left(\hat{w}(1-\hat{y}_1^2)\frac{(1-\hat{y}_2)}{2}\right)\right]^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y}\right. \\
& \left. + \int_{\hat{K}}\left[(1-\hat{x}_1^2)(1-\hat{x}_2^2)\partial_{\hat{x}_i}\partial_{\hat{y}_j}\left(\hat{v}(1-\hat{y}_1^2)\frac{(1-\hat{y}_2)}{2}\right)\right]^2 \mathrm{d}\hat{x}\,\mathrm{d}\hat{y}\right\}
\end{aligned}
$$

$$=2\sum_{i,j=1}^{2} c_K^{i,j}(J_{i,j}^1 + J_{i,j}^2).$$

Noting that $\hat{w}$ is independent of $\hat{y}_2$ we find for $j = 1, 2$ in the first term on the right-hand side

$$\begin{aligned}
J_{i,1}^1 =& \int_{\hat{K}} \left[\partial_{\hat{x}_i}((1-\hat{x}_1^2)(1-\hat{x}_2^2))\frac{(1-\hat{y}_2)}{2}\partial_{\hat{y}_1}(\hat{w}(1-\hat{y}_1^2))\right]^2 \\
\leq& C\left(4+2\sqrt{p(p+1)}\right)^2 h_Q^{-3}\|w\|_{L^2(Q)}^2, \\
J_{i,2}^1 =& \int_{\hat{K}} \left[\partial_{\hat{x}_i}((1-\hat{x}_1^2)(1-\hat{x}_2^2))\frac{-1}{2}\hat{w}(1-\hat{y}_1^2)\right]^2 \\
\leq& C h_Q^{-3}\|w\|_{L^2(Q)}^2,
\end{aligned}$$

where we have used the one dimensional estimate of Lemma 2.11 and have integrated out $\hat{y}_2$ before transforming back to the physical element. For the second term we find

$$\begin{aligned}
J_{i,1}^2 =& \int_{\hat{K}} \left[(1-\hat{x}_1^2)(1-\hat{x}_2^2)\frac{(1-\hat{y}_2)}{2}\partial_{\hat{y}_1}\left((\partial_{\hat{x}_i}\hat{w})(1-\hat{y}_1^2)\right)\right]^2 \\
\leq& C(4+2\sqrt{p(p+1)})^2 p(p+1) h_Q^{-3}\|w\|_{L^2(Q)}^2, \\
J_{i,2}^2 =& \int_{\hat{K}} \left[(1-\hat{x}_1^2)(1-\hat{x}_2^2)\frac{-1}{2}(\partial_{\hat{x}_i}\hat{w})(1-\hat{y}_1^2)\right]^2 \\
\leq& C p(p+1) h_Q^{-3}\|w\|_{L^2(Q)}^2.
\end{aligned}$$

Hence, there holds

$$\|(\nabla_x \otimes \nabla_y)(w\psi_Q)\|_{L^2(K)}^2 \leq C h_Q^{-3}\|w\|_{L^2(Q)}^2. \tag{2.66}$$

For the last estimate we proceed analogously. Since, letting $\hat{F} = [-1,1]\times\{-1\}\times[-1,1]\times\{-1\}$, using arguments as above and the fact that $\hat{u}$ is constant in the $\hat{x}_2$- and $\hat{y}_2$-direction, we find

$$\begin{aligned}
\|(\nabla_x \otimes \nabla_y)(u\psi_F)\|_{L^2(K)}^2 =& \sum_{i,j=1}^{2} c_K^{i,j}\int_{\hat{K}} \left(\partial_{\hat{x}_i}\partial_{\hat{y}_j}(\hat{u}\psi_{\hat{F}})\right)^2 \,\mathrm{d}\hat{x}\,\mathrm{d}\hat{y} \\
=& \sum_{i,j=1}^{2} c_K^{i,j}\int_{\hat{K}} \left[\partial_{\hat{x}_i}\left((1-\hat{x}_1^2)\frac{(1-\hat{x}_2)}{2}\partial_{\hat{y}_j}\left(\hat{u}(1-\hat{y}_1^2)\frac{(1-\hat{y}_2)}{2}\right)\right)\right]^2 \,\mathrm{d}\hat{x}\,\mathrm{d}\hat{y} \\
\leq& C(4+2\sqrt{p(p+1)})^2 p(p+1)\|\hat{u}\|_{L^2(\hat{F})}^2.
\end{aligned}$$

Transforming back to the physical element shows that

$$\|(\nabla_x \otimes \nabla_y)(u\psi_F)\|_{L^2(K)}^2 \leq C h_F^{-2}\|u\|_{L^2(F)}^2, \tag{2.67}$$

which concludes the proof. □

2.3.3 Discretization and constrained approximation in 4D

In the following we shall describe the discretization and lay out the procedure which is used to provide the necessary 1-irregular mesh refinement for our adaptive process in four space dimensions. To make the presentation of the algorithm more palpable, let us denote by

$e_i \in \mathcal{N}_{h,\mathcal{E}}$ hanging nodes on edges, by $f_j \in \mathcal{N}_{h,\mathcal{F}}$ hanging nodes on faces, and by $q_k \in \mathcal{N}_{h,\mathcal{Q}}$ hanging nodes on cubes, where the subscripts i, j, k realize an enumeration of the hanging nodes of $\mathcal{N}_{h,\mathcal{E}}$,$\mathcal{N}_{h,\mathcal{F}}$, and $\mathcal{N}_{h,\mathcal{Q}}$, respectively. Hence, as in the two dimensional setting we have a splitting of the nodes of the mesh $\mathcal{T}$ as $\mathcal{N} = \mathcal{N}_r \cup \mathcal{N}_{h,\mathcal{E}} \cup \mathcal{N}_{h,\mathcal{F}} \cup \mathcal{N}_{h,\mathcal{Q}} = \mathcal{N}_r \cup \mathcal{N}_h$ with $\mathcal{N}_r$ the set of regular nodes in $\mathcal{T}$.

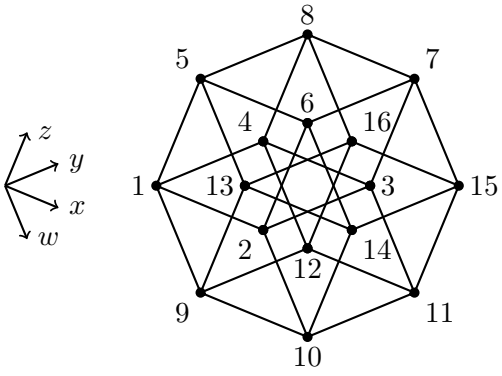

Figure 2.6: The reference 4D cube $\hat{K} = [-1, 1]^4$.

The reference four dimensional cube $\hat{K} := [-1, 1]^4$ is depicted in Figure 2.6. As a means for orientation we have prescribed an *a priori* ordering of the nodes. By pulling the three dimensional cube Q with the nodes $\{1, 2, 3, 4, 5, 6, 7, 8\}$ into the fourth dimension and connecting all vertices of Q with the vertices of its "duplicate" Q'. For simplicity the vertices of the "duplicate" Q' are enumerated as $\{9, 10, 11, 12, 13, 14, 15, 16\}$. In this sense the "original" cube $Q = [-1, 1]^3 \times \{-1\}$ is located at $w = -1$ and the "duplicate" $Q' = [-1, 1]^3 \times \{1\}$ at $w = 1$. Note that Q and Q' lie opposite to each other in $\hat{K}$ and do not share any vertices, edges nor faces.

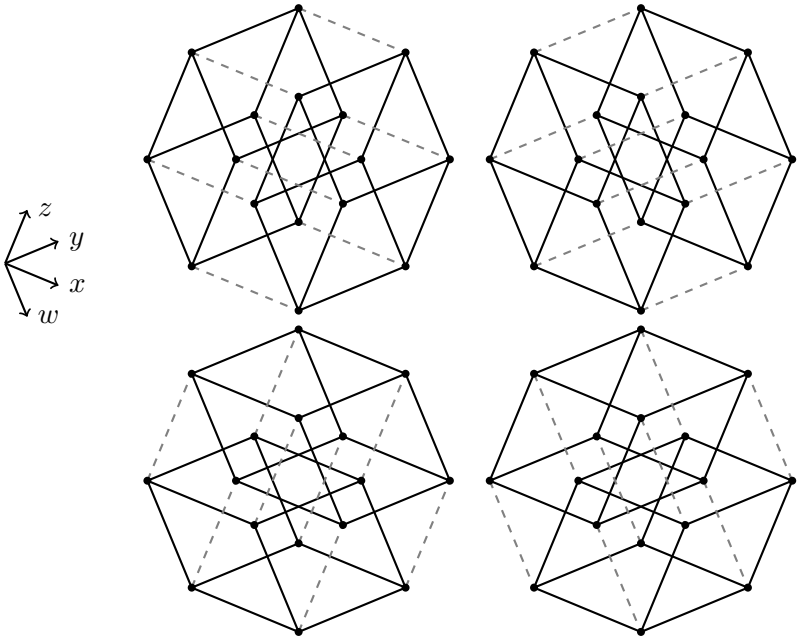

Table 2.1: Top left: x-cubes; top right: y-cubes; botom left: z-cubes; bottom right: w-cubes. The gray dashed line symbolizes the constant direction.

For this reason when visualizing the boundary of the four dimensional cube, it is

helpful to depict the three dimensional boundary cube pairs, which lie opposite to each other, such that one coordinate is kept fixed at -1 or 1. This situation is shown in the four pictures of Table 2.1. Note that all of these boundary cube pairs do not share any faces, edges nor vertices.

The situation of two four dimensional cubes sharing a three dimensional boundary cube, where one 4-cube is one level more refined than the other is depicted in Figure 2.7. With any of these types of hanging nodes we associate their corresponding *irregular elements* that have the hanging node situated on an edge, face, or cube, respectively, according to the following definition, which is similar to the situation in two space dimensions.

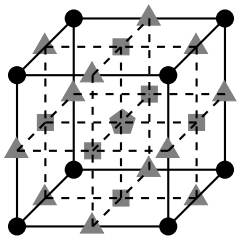

Figure 2.7: Hanging nodes on a three dimensional boundary cube that is shared by two four dimensional hypercubes K and K'. K is one level less refined than K'. The cube with the nodes marked by circles is a boundary cube of K and by the dashed lines we have indicated the refined boundary cube of K'. All nodes marked by a circle are *regular* nodes whereas the hanging nodes on edges are marked as gray triangles, hanging nodes on faces as gray squares and the hanging node on the boundary cube depicted is marked as a gray pentagon.

Definition 2.58 (Irregular Elements and Neighbors (4D))**.** *An element K is called* irregular *(with respect to a node x) if there exists a hanging node x, such that $x \in \partial K$ and x is either a hanging node on an edge E, on a face F, or on a cube Q of K, respectively, but x is* not *a vertex of K. Moreover, we shall call K an* irregular neighbor *to any element K' (with respect to x), if K' has x as a vertex.*

Note that the situation in the four dimensional setting is much more involved than in two dimensions as hanging nodes on edges and faces can be associated with *more than one* irregular element. In the case of a hanging node on an edge, there can be up to seven irregular neighbors and for hanging nodes on a face, there can be up to three irregular neighbors. The situation for hanging nodes on cubes mirrors the case of edges in the two dimensional constrained approximation. Moreover, note that there can be hanging nodes on faces and edges on the boundary $\partial\mathcal{D}$, which makes dealing with all types of hanging nodes a non-trivial endeavor.

For the adaptive mesh refinement procedure we shall adopt again the rule, that

an element $K \in \mathcal{T}$ may be refined if and only if all vertices of K are regular.

If any vertex of K is a hanging node of any type, we have to refine all irregular neighbors of K with respect to that hanging node *first*, such that the rule applies and we are subsequently allowed to refine the originally marked element K. This is formalized in Algorithm 3.

Another reason for adopting the aforementioned simplified refinement rule is that there is already a plethora of possible refinements in three space dimensions (cf. [23]) pertaining to different anisotropic refinements. The situation in four space dimensions is then even

more involved and the reason for restricting the refinement to the rule above. Note that in this we are somewhat neglecting the otherwise valuable anisotropic refinement information that an error estimator may carry and also give up the freedoms that different anisotropic refinements may deliver.

Algorithm 3 1-irregular mesh refinement in 4D (REFINE)

Input: Mesh $\mathcal{T}$, list M of marked elements $K \in \mathcal{T}$
Output: Refined mesh $\tilde{\mathcal{T}}$ which is 1-irregular

```
while M ≠ ∅ do
  Let K be the first element of M
  if K has any hanging nodes e_i or f_i or q_i as vertex then
    Find all irregular neighbors K' w.r.t. e_i, f_i and q_i
    Append all elements K' and K to M
  else
    Refine K by subdivision into 16 smaller 4-cubes K_s, s = 1, ..., 16
    Remove K from M and T̃, and add 16 children K_s to T̃
  end if
end while
```

Let us now discuss the refinement of the four dimensional cube. In Table 2.2 we have depicted the isotropic refinement of the reference cubes in their respective dimension for $d = 1, 2, 3$. This is to be seen as a preparation for the refinement for $d = 4$. We note that the two dimensional refined square can be viewd as having three refined copies of the refined interval at the slices $y = -1, y = 0$, and $y = 1$ which are connected to each other. For the refined cube the situation is analogous as each slice at $z = -1, z = 0$, and $z = 1$ contains a refined square which are then connected to each other. We use this observation to give a depiction of the refined hypercube in four dimensions. Now we have three copies of the refined cube at the three 3D slices $w = -1, w = 0$ and $w = 1$, which cover all vertices of the refined hypercube and is shown in Figure 2.8. We note that the refined 4-cube is given by 16 smaller 4-cubes which are represented using a total of 81 vertices.

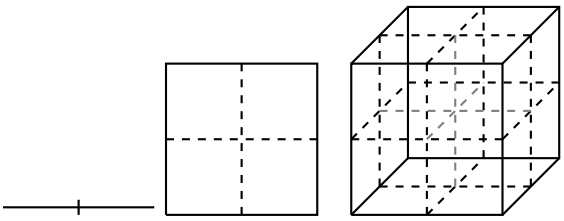

Table 2.2: Isotropic refinement of the one, two, and three dimensional cubes.

As in two dimensions we shall now briefly describe the treatment of hanging nodes in the four dimensional setting. We assume again that by a standard assembly process we arrive at the linear system of equations

$$Ac = b.$$

In the four dimensional setting we have to be more careful with the presentation. Here, we have to distinguish between the different types of hanging nodes. On edges, each hanging node $e \in \mathcal{N}_{h,\mathcal{E}}$ is constrained by its two regular neighbor vertices $i, j \in \mathcal{N}_r$ and

$i, j \in O_e$ by the relation

$$u_e = \frac{1}{2}u_i + \frac{1}{2}u_j,$$

as in two space dimensions.

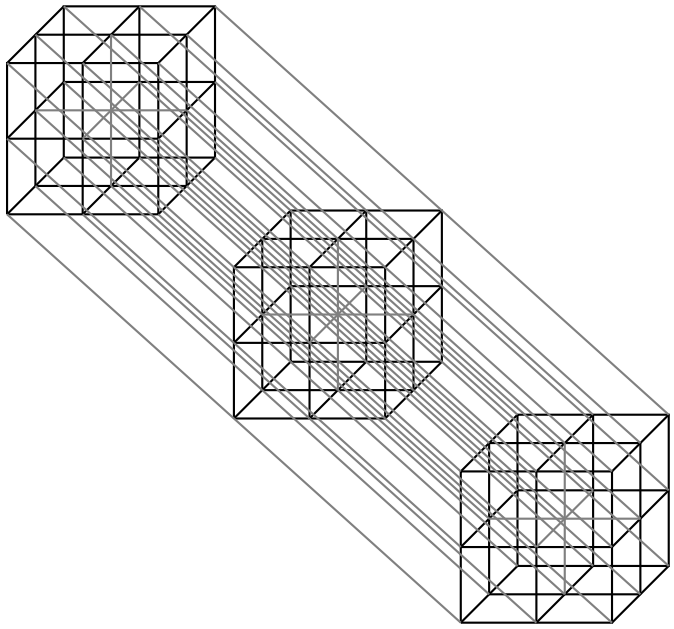

Figure 2.8: Isotropic refinement of the four dimensional cube (a simplified view).

On faces, each hanging node $f \in \mathcal{N}_{h,\mathcal{F}}$ is constrained by its four regular neighbor vertices $i, j, k, \ell \in \mathcal{N}_r$ via

$$u_f = \frac{1}{4}\left(u_i + u_j + u_k + u_\ell\right). \tag{2.68}$$

And on cubes, each hanging node $q \in \mathcal{N}_{h,\mathcal{Q}}$ is constrained by its eight regular neighbor vertices $v_1, \ldots, v_8$ by the expression

$$u_q = \frac{1}{8}\sum_{i=1}^{8} u_{v_i}. \tag{2.69}$$

The global connectivity matrix $P \in \mathbb{R}^{N,N}$, where $N := |\mathcal{N}|$, then takes the following form. Columns of regular vertices are as in the two dimensional situation. If, on the other hand, $c \in \mathcal{N}_h$, then the corresponding column is filled with values according to the type of the hanging node c.

If $c \in \mathcal{N}_{h,\mathcal{E}}$, then as in the two dimensional setting the corresponding column in P is zero and features the value $1/2$ at positions $i, j \in \mathcal{N}_r$ of its regular vertex neighbors. If $c \in \mathcal{N}_{h,\mathcal{F}}$, then the corresponding column in P is populated with values according to (2.68), i.e. denoting the column of c by p_c we have

$$p_c = (0 \ldots 0 \underbrace{1/4}_{i} 0 \ldots 0 \underbrace{1/4}_{j} 0 \ldots 0 \underbrace{1/4}_{k} 0 \ldots 0 \underbrace{1/4}_{\ell} 0 \ldots 0)^\top.$$

Finally, if $c \in \mathcal{N}_{h,\mathcal{Q}}$, the corresponding column p_c is given similarly as for face hanging nodes, but now at positions that correspond to the vertices v_i, $i = 1, \ldots, 8$ of the regular vertex neighbors the value $1/8$ is inserted.

With this in hand we can reduce the system as in two space dimensions, by letting P_r the matrix that arises when cutting all zero rows from P,

$$P_r A P_r^\top c_r = P_r b,$$

where only the coefficients of regular vertices are computed and subsequently all coefficients are recovered via

$$c = P_r^\top c_r.$$

2.3.4 A residual a posteriori error estimator

In order to derive a residual *a posteriori* error estimator in this situation, we proceed similarly as in the case of the one dimensional model problem. In the following denote by $\partial_{n,\nu} u = \nabla_\nu u \cdot \mathbf{n}_\nu$ for $\nu \in \{x, y\}$ the normal derivative with respect to a certain variable. Firstly, we recall that $\mathcal{D} = D_x \times D_y$ and that the weak formulation for the deterministic second moment problem for (2.62) is given by

$$\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) \mathcal{C}_u (\nabla_x \otimes \nabla_y) v \,\mathrm{d}y \,\mathrm{d}x = \int_{\mathcal{D}} \mathcal{C}_f v \,\mathrm{d}y \,\mathrm{d}x.$$

Considering the residual $\mathcal{R}$ which is implicitly defined as an element of the dual space of $H_0^{1,1}(\mathcal{D})$ by the left-hand side of the expression

$$\begin{aligned}
\langle \mathcal{R}, v \rangle &= \int_{D_x} \int_{D_y} (\nabla_x \otimes \nabla_y)(\mathcal{C}_u - u_{\mathcal{T}})(\nabla_x \otimes \nabla_y) v \,\mathrm{d}y \,\mathrm{d}x \\
&= \int_{D_x} \int_{D_y} \mathcal{C}_f v \,\mathrm{d}y \,\mathrm{d}x - \int_{D_x} \int_{D_y} (\nabla_x \otimes \nabla_y) u_{\mathcal{T}} (\nabla_x \otimes \nabla_y) v \,\mathrm{d}y \,\mathrm{d}x,
\end{aligned}$$

we localize the integrals to elements $K = K_x \times K_y \in \mathcal{T}$ and proceed with integration by parts to arrive at

$$\begin{aligned}
\langle \mathcal{R}, v \rangle &= \sum_{K \in \mathcal{T}} \int_{K_x} \int_{K_y} \mathcal{C}_f v \,\mathrm{d}y \,\mathrm{d}x - \sum_{K \in \mathcal{T}} \int_{K_x} \int_{K_y} (\nabla_x \otimes \nabla_y) u_{\mathcal{T}} (\nabla_x \otimes \nabla_y) v \,\mathrm{d}y \,\mathrm{d}x \\
&= \sum_{K \in \mathcal{T}} \int_K \mathcal{C}_f v \,\mathrm{d}y \,\mathrm{d}x - \sum_{K \in \mathcal{T}} \int_{K_x} \Bigg\{ \int_{K_y} -(\nabla_x \otimes \Delta_y) u_{\mathcal{T}} (\nabla_x \otimes \mathrm{id}) v \,\mathrm{d}y \\
&\quad + \int_{\partial K_y} (\nabla_x \otimes \partial_{n,y}) u_{\mathcal{T}} (\nabla_x \otimes \mathrm{id}) v \,\mathrm{d}s_y \Bigg\} \,\mathrm{d}x.
\end{aligned}$$

Rearranging the terms and applying integration by parts twice more, we end up with the following expression of the residual:

$$\begin{aligned}
\langle \mathcal{R}, v \rangle &= \sum_{K \in \mathcal{T}} \int_K \mathcal{C}_f v \,\mathrm{d}y \,\mathrm{d}x - \sum_{K \in \mathcal{T}} \Bigg\{ \int_{K_x} \int_{K_y} -(\nabla_x \otimes \Delta_y) u_{\mathcal{T}} (\nabla_x \otimes \mathrm{id}) v \,\mathrm{d}y \,\mathrm{d}x \\
&\quad + \int_{K_x} \int_{\partial K_y} (\nabla_x \otimes \partial_{n,y}) u_{\mathcal{T}} (\nabla_x \otimes \mathrm{id}) v \,\mathrm{d}s_y \,\mathrm{d}x \Bigg\} \\
&= \sum_{K \in \mathcal{T}} \int_K \mathcal{C}_f v \,\mathrm{d}y \,\mathrm{d}x - \sum_{K \in \mathcal{T}} \Bigg\{ \int_{K_y} \Bigg\{ \int_{K_x} (\Delta_x \otimes \Delta_y) u_{\mathcal{T}} v \,\mathrm{d}x \\
&\quad + \int_{\partial K_x} -(\partial_{n,x} \otimes \Delta_y) u_{\mathcal{T}} v \,\mathrm{d}s_x \Bigg\} \,\mathrm{d}y
\end{aligned}$$

$$
\begin{aligned}
&\quad + \int_{K_x}\int_{\partial K_y} (\nabla_x \otimes \partial_{n,y}) u_{\mathcal{T}} (\nabla_x \otimes \mathrm{id}) v \,\mathrm{d}s_y \,\mathrm{d}x \Bigg\} \\
&= \sum_{K\in\mathcal{T}} \int_K \mathcal{C}_f v \,\mathrm{d}y \,\mathrm{d}x - \sum_{K\in\mathcal{T}} \Bigg\{ \int_{K_y}\int_{K_x} (\Delta_x \otimes \Delta_y) u_{\mathcal{T}} v \,\mathrm{d}x \,\mathrm{d}y \\
&\quad + \int_{K_y}\int_{\partial K_x} -(\partial_{n,x} \otimes \Delta_y) u_{\mathcal{T}} v \,\mathrm{d}s_x \,\mathrm{d}y \\
&\quad + \int_{K_x}\int_{\partial K_y} (\nabla_x \otimes \partial_{n,y}) u_{\mathcal{T}} (\nabla_x \otimes \mathrm{id}) v \,\mathrm{d}s_y \,\mathrm{d}x \Bigg\} \\
&= \sum_{K\in\mathcal{T}} \Bigg\{ \int_K (\mathcal{C}_f - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}}) v \,\mathrm{d}y \,\mathrm{d}x + \int_{K_y}\int_{\partial K_x} (\partial_{n,x} \otimes \Delta_y) u_{\mathcal{T}} v \,\mathrm{d}s_x \,\mathrm{d}y \\
&\quad + \int_{K_x}\int_{\partial K_y} (\Delta_x \otimes \partial_{n,y}) u_{\mathcal{T}} v \,\mathrm{d}s_y \,\mathrm{d}x - \int_{\partial K_x}\int_{\partial K_y} (\partial_{n,x} \otimes \partial_{n,y}) u_{\mathcal{T}} v \,\mathrm{d}s_x \,\mathrm{d}s_y \Bigg\}.
\end{aligned}
$$

A closer inspection of the last term reveals that x_1x_2- and y_1y_2-faces do *not* contribute to the residual. In other words, only the "mixed" faces, i.e. that are given as $F = e_x \times e_y$ with an edge in x-coordinates and an edge in y-coordinates, yield contributions to the error estimator. In order to simplify the notation we shall denote these faces by $F_{xy} \subset Q$. Hence, we can represent the residual by the following expression

$$
\langle \mathcal{R}, v \rangle = \sum_{K\in\mathcal{T}} \Bigg\{ \int_K r v \,\mathrm{d}x + \sum_{Q\subset\partial K} \int_Q j_1 v \,\mathrm{d}s + \sum_{F_{xy}\subset Q\subset\partial K} \int_{F_{xy}} j_2 v \,\mathrm{d}s \Bigg\}, \tag{2.70}
$$

where for later reference and brevity of notation we set

$$
\begin{aligned}
r &:= \mathcal{C}_f - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}}, \quad \text{on every } K \in \mathcal{T}, \\
j_1 &:= \begin{cases} [\![(\partial_{n,x} \otimes \Delta_y) u_{\mathcal{T}}]\!], & Q \text{ is a } x_1\text{-}y_1\text{-}y_2 \text{ or } x_2\text{-}y_1\text{-}y_2 \text{ cube}, \\ [\![(\Delta_x \otimes \partial_{n,y}) u_{\mathcal{T}}]\!], & Q \text{ is a } x_1\text{-}x_2\text{-}y_1 \text{ or } x_1\text{-}x_2\text{-}y_2 \text{ cube}, \end{cases} \\
j_2 &:= \begin{cases} -\{\{ (\partial_{n,x} \otimes \partial_{n,y}) u_{\mathcal{T}} \}\}, & \text{on every face } F_{xy} \subset Q \subset \partial K, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}
$$

For the generalized jump over faces, i.e. for the term j_2, we introduce the following notation

$$
\{\{ (\partial_{n,x} \otimes \partial_{n,y}) u_{\mathcal{T}} \}\} := \sum_{K\in\mathcal{T}: F_{xy}\subset\partial K} ((\partial_{n,x} \otimes \partial_{n,y}) u_{\mathcal{T}})|_K. \tag{2.71}
$$

The previous representation of the residual motivates the following definition.

Definition 2.59. *For the deterministic second moment problem (2.62) we define the residual a posteriori error estimator $\eta_{\mathcal{R},K}$ element-wise by the expression*

$$
\eta_{\mathcal{R},K}^2 := h_K^2 \|r\|_{L^2(K)}^2 + \frac{1}{2} \sum_{Q\subset\partial K} h_Q \|j_1\|_{L^2(Q)}^2 + \frac{1}{4} \sum_{\substack{F_{xy}\subset Q: \\ Q\subset\partial K}} h_{F_{xy}} \|j_2\|_{L^2(F_{xy})}^2. \tag{2.72}
$$

The global error estimator $\eta_{\mathcal{R}}$ is then given as

$$
\eta_{\mathcal{R}} := \left(\sum_{K\in\mathcal{T}} \eta_{\mathcal{R},K}^2 \right)^{1/2}. \tag{2.73}
$$

The appearing tensorized differential operators $\Delta_x \otimes \Delta_y, \partial_{n,x} \otimes \Delta_y$, and $\Delta_x \otimes \partial_{n,y}$ can be interpreted straighforwardly as $\Delta_x \Delta_y, \partial_{n,x} \Delta_y$, and $\Delta_x \partial_{n,y}$, respectively, as the factors act on different variables, whence they commute and their composition is unambiguous. The operator $\partial_{n,x} \otimes \partial_{n,y}$ can be understood in the following way for computations. By definition of the tensor product we have

$$\begin{aligned}(\partial_{n,x} \otimes \partial_{n,y}) u_{\mathcal{T}} &= \nabla_x (\nabla_y u_{\mathcal{T}} \cdot \mathbf{n}_y) \cdot \mathbf{n}_x = \nabla_y (\nabla_x u_{\mathcal{T}} \cdot \mathbf{n}_x) \cdot \mathbf{n}_y \\ &= \mathbf{n}_x^\top \cdot \begin{pmatrix} \partial_{x_1} \partial_{y_1} u_{\mathcal{T}} & \partial_{x_1} \partial_{y_2} u_{\mathcal{T}} \\ \partial_{x_2} \partial_{y_1} u_{\mathcal{T}} & \partial_{x_2} \partial_{y_2} u_{\mathcal{T}} \end{pmatrix} \cdot \mathbf{n}_y \\ &= \sum_{i,j=1}^{2} \partial_{x_i} \partial_{y_j} u \cdot \mathbf{n}_x^i \mathbf{n}_y^j.\end{aligned}$$

Furthermore, let us state that by definition of the canonical tensor product Hilbert space norm we have with $u = u_1 \otimes u_2 \in H^{1,1}(\mathcal{D})$ the following representation

$$\begin{aligned}\|(\nabla_x \otimes \nabla_y) u\|_{L^2(\mathcal{D})}^2 &= \langle (\nabla_x \otimes \nabla_y) u, (\nabla_x \otimes \nabla_y) u \rangle_{L^2(\mathcal{D})} \\ &= \langle \nabla_x u_1, \nabla_x u_1 \rangle_{L^2(D_x)} \langle \nabla_y u_2, \nabla_y u_2 \rangle_{L^2(D_y)} \\ &= \int_{D_x} (\nabla_x u_1)^2 \,\mathrm{d}x \int_{D_y} (\nabla_y u_2)^2 \,\mathrm{d}y \\ &= \sum_{i,j=1}^{2} \int_{\mathcal{D}} (\partial_{x_i} \partial_{y_j} u)^2 \,\mathrm{d}y \,\mathrm{d}x\end{aligned}$$

and with $v = v_1 \otimes v_2 \in H^{1,1}(\mathcal{D})$

$$\begin{aligned}\langle (\nabla_x \otimes \nabla_y) u, (\nabla_x \otimes \nabla_y) v \rangle_{L^2(\mathcal{D})} &= \langle \nabla_x u_1, \nabla_x v_1 \rangle_{(L^2(D_x)} \langle \nabla_y u_2, \nabla_y v_2 \rangle_{(L^2(D_y)} \\ &= \int_{D_x} (\nabla_x u_1)(\nabla_x v_1) \,\mathrm{d}x \int_{D_y} (\nabla_y u_2)(\nabla_y v_2) \,\mathrm{d}y \\ &= \sum_{i,j=1}^{2} \int_{\mathcal{D}} \left(\partial_{x_i} \partial_{y_j} u\right) \left(\partial_{x_i} \partial_{y_j} v\right) \,\mathrm{d}y \,\mathrm{d}x.\end{aligned}$$

An upper bound on the error

As for the deterministic second moment problem of the one dimensional model problem we proceed by proving the reliability of the derived residual a posteriori error estimator $\eta_{\mathcal{R}}$.

Lemma 2.60 (Reliability)**.** *Let $\mathcal{D} = D_x \times D_y = [-1,1]^4$. Furthermore, u be the exact solution of (2.62) and $u_{\mathcal{T}}$ the solution of the corresponding discrete variational formulation (2.63). Then there exists a positive constant c^* which only depends on the shape regularity of the mesh $\mathcal{T}$, the polynomial degree $\mathbf{p}_{\mathcal{T}}$, and $\mathcal{D}$, such that*

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})}^2 \leq c^* \sum_{K \in \mathcal{T}} \left(h_K^2 \|r\|_{L^2(K)}^2 + \sum_{Q \subset \partial K} h_Q \|j_1\|_{L^2(Q)}^2 + \sum_{\substack{F_{xy} \subset Q: \\ Q \subset \partial K}} h_{F_{xy}} \|j_2\|_{L^2(F_{xy})}^2 \right). \tag{2.74}$$

Proof. Let $v \in H_0^{1,1}(\mathcal{D})$ be arbitrary. In order to show that the error $|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})}$ is bounded from above globally, we start out from the expression (2.70) and insert the

function $v - \mathbb{I}_{\mathcal{T}} v$ instead of v. Using the Cauchy-Schwarz inequality for sums and integrals in conjunction with approximation results from the previous section and Galerkin orthogonality, we obtain

$$
\begin{aligned}
\langle \mathcal{R}, v \rangle =& \langle \mathcal{R}, v - \mathbb{I}_{\mathcal{T}} v \rangle \\
=& \sum_{K \in \mathcal{T}} \left\{ \int_K r(v - \mathbb{I}_{\mathcal{T}} v) \,\mathrm{d}x \,\mathrm{d}y + \sum_{Q \subset \partial K} \int_Q j_1 (v - \mathbb{I}_{\mathcal{T}} v) \,\mathrm{d}s \right. \\
& \left. + \sum_{\substack{F_{xy} \subset Q: \\ Q \subset \partial K}} \int_{F_{xy}} j_2 (v - \mathbb{I}_{\mathcal{T}} v) \,\mathrm{d}s \right\} \\
\leq & \sum_{K \in \mathcal{T}} \left\{ \|r\|_{L^2(K)} \|v - \mathbb{I}_{\mathcal{T}} v\|_{L^2(K)} + \sum_{Q \subset \partial K} \|j_1\|_{L^2(Q)} \|v - \mathbb{I}_{\mathcal{T}} v\|_{L^2(Q)} \right. \\
& \left. + \sum_{\substack{F_{xy} \subset Q: \\ Q \subset \partial K}} \|j_2\|_{L^2(F_{xy})} \|v - \mathbb{I}_{\mathcal{T}} v\|_{L^2(F_{xy})} \right\} \\
\leq & \sum_{K \in \mathcal{T}} \left\{ \|r\|_{L^2(K)} C_1 h_K \left(\|\nabla_x v\|_{L^2(\tilde{\omega}_K)} + \|\nabla_y v\|_{L^2(\tilde{\omega}_K)} + |v|_{H^{1,1}(\tilde{\omega}_K)} \right) \right. \\
& + \sum_{Q \subset \partial K} \|j_1\|_{L^2(Q)} C_2 h_Q^{1/2} \left(\|\nabla_x v\|_{L^2(\tilde{\omega}_K)} + \|\nabla_y v\|_{L^2(\tilde{\omega}_K)} + |v|_{H^{1,1}(\tilde{\omega}_K)} \right) \\
& \left. + \sum_{\substack{F_{xy} \subset Q: \\ Q \subset \partial K}} \|j_2\|_{L^2(F_{xy})} C_3 h_{F_{xy}}^{1/2} \left(\|\nabla_x v\|_{L^2(\tilde{\omega}_K)} + \|\nabla_y v\|_{L^2(\tilde{\omega}_K)} + |v|_{H^{1,1}(\tilde{\omega}_K)} \right) \right\} \\
\leq & \max\{C_1, C_2, C_3\} \times \\
& \left\{ \sum_{K \in \mathcal{T}} \left[h_K^2 \|r\|_{L^2(K)}^2 + \sum_{Q \subset K} h_Q \|j_1\|_{L^2(Q)}^2 + \sum_{\substack{F_{xy} \subset Q: \\ Q \subset \partial K}} h_{F_{xy}} \|j_2\|_{L^2(F_{xy})}^2 \right] \right\}^{1/2} \\
& \times \left\{ \sum_{K \in \mathcal{T}} \left[\|\nabla_x v\|_{L^2(\tilde{\omega}_K)}^2 + \|\nabla_y v\|_{L^2(\tilde{\omega}_K)}^2 + |v|_{H^{1,1}(\tilde{\omega}_K)}^2 \right. \right. \\
& + \sum_{Q \subset K} \left(\|\nabla_x v\|_{L^2(\tilde{\omega}_K)}^2 + \|\nabla_y v\|_{L^2(\tilde{\omega}_K)}^2 + |v|_{H^{1,1}(\tilde{\omega}_K)}^2 \right) \\
& \left. \left. + \sum_{\substack{F_{xy} \subset Q: \\ Q \subset \partial K}} \left(\|\nabla_x v\|_{L^2(\tilde{\omega}_K)}^2 + \|\nabla_y v\|_{L^2(\tilde{\omega}_K)}^2 + |v|_{H^{1,1}(\tilde{\omega}_K)}^2 \right) \right] \right\}^{1/2},
\end{aligned}
$$

where we have abbreviated $(\nabla_x \otimes \mathrm{id})$ to ∇_x as well as $(\mathrm{id} \otimes \nabla_y)$ to ∇_y and $\tilde{\omega}_K$ denotes the patch of elements K' that have at least one vertex in common with K and in the last step the Cauchy-Schwarz inequality for sums has been used. Now, by additivity of the

Lebesgue integral and Friedrichs' inequalities in x- and y-direction, respectively, we have

$$\begin{aligned}\Bigg\{ &\sum_{K\in\mathcal{T}} \left(\|\nabla_x v\|^2_{L^2(\tilde{\omega}_K)} + \|\nabla_y v\|^2_{L^2(\tilde{\omega}_K)} + |v|^2_{H^{1,1}(\tilde{\omega}_K)} \right) \\ &+ \sum_{K\in\mathcal{T}} \sum_{Q\subset K} \left(\|\nabla_x v\|^2_{L^2(\tilde{\omega}_K)} + \|\nabla_y v\|^2_{L^2(\tilde{\omega}_K)} + |v|^2_{H^{1,1}(\tilde{\omega}_K)} \right) \\ &+ \sum_{K\in\mathcal{T}} \sum_{\substack{F_{xy}\subset Q: \\ Q\subset\partial K}} \left(\|\nabla_x v\|^2_{L^2(\tilde{\omega}_K)} + \|\nabla_y v\|^2_{L^2(\tilde{\omega}_K)} + |v|^2_{H^{1,1}(\tilde{\omega}_K)} \right) \Bigg\}^{1/2} \leq c_{\mathcal{T}} |v|_{H^{1,1}(\mathcal{D})},\end{aligned}$$

where $c_{\mathcal{T}}$ depends on the shape regularity of the mesh $\mathcal{T}$, takes into consideration that elements are counted multiple times on the left-hand side and incorporates the constants of Friedrichs' inequality in x- and y-direction, respectively. Thus,

$$\frac{\langle \mathcal{R}, v\rangle}{|v|_{H^{1,1}(\mathcal{D})}} \leq C \left\{ \sum_{K\in\mathcal{T}} \left(h_K^2 \|r\|^2_{L^2(K)} + \sum_{Q\subset\partial K} h_Q \|j_1\|^2_{L^2(Q)} + \sum_{\substack{F_{xy}\subset Q: \\ Q\subset\partial K}} h_{F_{xy}} \|j_2\|^2_{L^2(F_{xy})} \right) \right\}^{1/2}$$

with $C = c_{\mathcal{T}} \max\{C_1, C_2, C_3\}$ and taking the supremum over all $v \in H_0^{1,1}(\mathcal{D})$ shows that there exists a constant $c^* > 0$, such that

$$|u - u_{\mathcal{T}}|^2_{H^{1,1}(\mathcal{D})} \leq c^* \sum_{K\in\mathcal{T}} \left(h_K^2 \|r\|^2_{L^2(K)} + \sum_{E\in\mathcal{E}} h_Q \|j_1\|^2_{L^2(Q)} + \sum_{\substack{F_{xy}\subset Q: \\ Q\subset\partial K}} h_{F_{xy}} \|j_2\|^2_{L^2(F_{xy})} \right).$$

□

Assuming that f is replaced by an approximation $f_{\mathcal{T}}$ we have the following result.

Corollary 2.61. *Let u be the exact solution of (2.62) and $u_{\mathcal{T}}$ the solution of the corresponding discrete variational formulation (2.63). Then there exists a constant c^* which only depends on the shape regularity of the mesh $\mathcal{T}$, the polynomial degree $\mathbf{p}_{\mathcal{T}}$, and $\mathcal{D}$, such that there holds*

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{Q})} \leq c^* \left(\eta_{\mathcal{R}}^2 + \mathsf{osc}_f \right)^{1/2}. \tag{2.75}$$

where $\mathsf{osc}_f := \sum_{K\in\mathcal{T}} h_K^2 \|f - f_{\mathcal{T}}\|^2_{L^2(K)}$

Proof. The proof is completely analogous to Corollary 2.25 and is therefore omitted. □

Remark 2.62. *Note that the factors $1/4$ and $1/2$ in front of the cube and face residuals in the definition of $\eta_{\mathcal{R}}$ merely take care of the fact that inner cubes are counted twice and inner faces up to four times when summing over all element contributions $\eta^2_{\mathcal{R},K}$ to get $\eta_{\mathcal{R}}$.*

A lower bound on the error

We will now proceed to attempt to show a lower bound on the error. This is done in an element-wise fashion. We start by bounding the element residuals involving r. Then

we turn to the cube and face residuals. To this end fix an element K and let $w_{\mathcal{T}}$ be the function that restricted to K is of the form

$$w_{\mathcal{T}} = (f_{\mathcal{T}} - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}}) \psi_K,$$

where ψ_K again denotes the element bubble function and $f_{\mathcal{T}}$ is a suitable approximation of f.

Lemma 2.63 (Element residuals)**.** *Let $f_{\mathcal{T}}$ any suitable approximation of the right-hand side f. Let u be the exact solution of (2.62) and $u_{\mathcal{T}}$ the exact solution of the discrete problem (2.63). Then for any element $K \in \mathcal{T}$, there exist positive constants c_1 and c_2, such that there holds*

$$h_K^2 \|f_{\mathcal{T}} - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}}\|_{L^2(K)} \leq c_1 \|(\nabla_x \otimes \nabla_y)(u - u_{\mathcal{T}})\|_{L^2(K)} + c_2 h_K^2 \|f_{\mathcal{T}} - f\|_{L^2(K)}. \quad (2.76)$$

Proof. With notations as above we have,

$$\begin{aligned} \int_K (f_{\mathcal{T}} - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}})^2 \psi_K \,\mathrm{d}x\,\mathrm{d}y &= \int_K (f_{\mathcal{T}} - \Delta_x \otimes \Delta_y u_{\mathcal{T}}) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_K r w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y + \int_K (f_{\mathcal{T}} - f) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_K (\nabla_x \otimes \nabla_y)(u - u_{\mathcal{T}})(\nabla_x \otimes \nabla_y) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y \\ &\qquad + \int_K (f_{\mathcal{T}} - f) w_{\mathcal{T}} \,\mathrm{d}x\,\mathrm{d}y. \end{aligned} \quad (2.77)$$

Using properties of the element bubble functions (cf. Lemma 2.11) in each coordinate direction we conclude that

$$\int_K (f_{\mathcal{T}} - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}})^2 \psi_K \,\mathrm{d}x\,\mathrm{d}y \geq (p+2)^{-8} \|f_{\mathcal{T}} - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}}\|_{L^2(K)}^2.$$

Estimating the right hand side of (2.77) yields

$$\begin{aligned} \int_K (\nabla_x \otimes \nabla_y)(u - u_{\mathcal{T}})(\nabla_x \otimes \nabla_y) w_{\mathcal{T}} &\leq |u - u_{\mathcal{T}}|_{H^{1,1}(K)} |w_{\mathcal{T}}|_{H^{1,1}(K)} \\ &\leq |u - u_{\mathcal{T}}|_{H^{1,1}(K)} C h_K^{-2} \|f_{\mathcal{T}} - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}}\|_{L^2(K)}, \\ \int_K (f_{\mathcal{T}} - f) w_{\mathcal{T}} &\leq \|f_{\mathcal{T}} - f\|_{L^2(K)} \|w_{\mathcal{T}}\|_{L^2(K)} \\ &\leq \|f_{\mathcal{T}} - f\|_{L^2(K)} \|f_{\mathcal{T}} - (\Delta_x \otimes \Delta_y) u_{\mathcal{T}}\|_{L^2(K)}. \end{aligned}$$

Combining these estimates we find

$$h_K^2 \|f_{\mathcal{T}} - \Delta_x \otimes \Delta_y u_{\mathcal{T}}\|_{L^2(K)} \leq c_1 \|(\nabla_x \otimes \nabla_y)(u - u_{\mathcal{T}})\|_{L^2(K)} + c_2 h_K^2 \|f_{\mathcal{T}} - f\|_{L^2(K)}. \quad (2.78)$$

□

We now turn to estimating the element boundary residuals. First we consider a boundary cube $Q \subset \partial K$ and insert the function

$$w_Q = j_1 \psi_Q$$

into (2.70), where ψ_Q denotes the minimal polynomial such that ψ_Q attains its unit maximum in the barycenter of $Q = K \cap K'$ and is zero on all other boundary cubes Q' of K and K'. The support of ψ_Q evidently is given by $\omega_Q = K \cup K'$.

Lemma 2.64 (Boundary Cube Residual Estimate)**.** *With the notation as above and u the exact solution of (2.62) and $u_{\mathcal{T}}$ the solution of the corresponding discrete variational formulation (2.63). Then there exist positive constants c_1 and c_2, such that*

$$h^{3/2}\|j_1\|_{L^2(Q)} \leq c_1\|(\nabla_x \otimes \nabla_y)(u-u_{\mathcal{T}})\|_{L^2(\omega_Q)} + c_2 \sum_{K\subset\omega_Q} h_K^2\|f_{\mathcal{T}} - f\|_{L^2(K)}.$$

Proof. Inserting w_Q into the representation of the residual yields

$$\begin{aligned}\int_Q j_1^2\psi_Q \,\mathrm{d}s &= \int_Q j_1 w_Q \,\mathrm{d}s \\ &= \int_{\omega_Q} (\nabla_x \otimes \nabla_y)(u-u_{\mathcal{T}})(\nabla_x \otimes \nabla_y) w_Q \,\mathrm{d}x\,\mathrm{d}y - \int_{\omega_Q} r w_Q \,\mathrm{d}x\,\mathrm{d}y \\ &\quad - \int_{\omega_Q} (f - f_{\mathcal{T}}) w_Q \,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

We note that by construction w_Q vanishes in all vertices of K, on all remaining boundary cubes and on all two dimensional faces F of K. Then similarly to the element residual by Lemma 2.11 we have

$$\int_Q j_1 w_Q \,\mathrm{d}s \geq (p+2)^{-6}\|j_1\|^2_{L^2(Q)}.$$

Now we bound the terms on the right-hand side using estimates from section 2.3.2 and find

$$\begin{aligned}\int_{\omega_Q} (\nabla_x \otimes \nabla_y)(u-u_{\mathcal{T}})(\nabla_x \otimes \nabla_y) w_Q \,\mathrm{d}x\,\mathrm{d}y &\leq |u-u_{\mathcal{T}}|_{H^{1,1}(\omega_Q)} |w_Q|_{H^{1,1}(\omega_Q)} \\ &\leq |u-u_{\mathcal{T}}|_{H^{1,1}(\omega_Q)} C h_Q^{-3/2}\|j_1\|_{L^2(Q)},\end{aligned}$$

$$\begin{aligned}\sum_{K\subset\omega_Q}\int_K r w_Q \,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{K\subset\omega_Q} \|r\|_{L^2(K)}\|w_Q\|_{L^2(K)} \\ &\leq \sum_{K\subset\omega_Q} \|r\|_{L^2(K)} C h_Q^{1/2}\|j_1\|_{L^2(Q)},\end{aligned}$$

$$\begin{aligned}\sum_{K\subset\omega_Q}\int_K (f-f_{\mathcal{T}}) w_Q \,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{K\subset\omega_Q} \|(f-f_{\mathcal{T}})\|_{L^2(K)}\|w_Q\|_{L^2(K)} \\ &\leq \sum_{K\subset\omega_Q} \|(f-f_{\mathcal{T}})\|_{L^2(K)} C h_Q^{1/2}\|j_1\|_{L^2(Q)}.\end{aligned}$$

Combining the aforementioned estimates and dividing by $h_Q^{-3/2}\|j_1\|_{L^2(E)}$, we get

$$\begin{aligned}h_Q^{3/2}\|j_1\|_{L^2(Q)} &\leq C\|(\nabla_x \otimes \nabla_y)(u-u_{\mathcal{T}})\|_{L^2(\omega_Q)} \\ &\quad + \sum_{K\subset\omega_Q} C h_Q^2\|r\|_{L^2(K)} + \sum_{K\subset\omega_Q} C h_Q^2\|(f-f_{\mathcal{T}})\|_{L^2(K)} \\ &\leq c_1\|(\nabla_x \otimes \nabla_y)(u-u_{\mathcal{T}})\|_{L^2(\omega_Q)} + c_2 \sum_{K\subset\omega_Q} h_K^2\|f-f_{\mathcal{T}}\|_{L^2(K)},\end{aligned}$$

where in the last step we have used shape regularity and the estimate for the element residual (2.78). □

We still need to investigate the influence of face residuals. As before we set

$$w_F = j_2 \psi_F.$$

Moreover, we recall that only "mixed" faces F_{xy} are of importance to us.

Lemma 2.65 (Boundary Face Residual Estimate)**.** *With the notation as above and u the exact solution of (2.62) and $u_\mathcal{T}$ the solution of the corresponding discrete variational formulation (2.63). Then there exist positive constants c_1 and c_2, such that*

$$h_F \|j_2\|_{L^2(F)} \leq c_1 \|(\nabla_x \otimes \nabla_y)(u - u_\mathcal{T})\|_{L^2(\omega_F)} + c_2 \sum_{K \subset \omega_F} h_K^2 \|f_\mathcal{T} - f\|_{L^2(K)}.$$

Proof. Inserting w_F into the representation of the residual yields

$$\begin{aligned}
\int_F (j_2^2 \psi_F)\,\mathrm{d}s &= \int_F j_2 w_F \,\mathrm{d}s \\
&= \int_{\tilde{\omega}_F} (\nabla_x \otimes \nabla_y)(u - u_\mathcal{T})(\nabla_x \otimes \nabla_y) w_F \,\mathrm{d}x\,\mathrm{d}y - \int_{\tilde{\omega}_F} r w_F \,\mathrm{d}x\,\mathrm{d}y \\
&\quad - \int_{\tilde{\omega}_F} (f - f_\mathcal{T}) w_F \,\mathrm{d}x\,\mathrm{d}y - \int_{\widehat{\omega}_F} j_1 w_F \,\mathrm{d}s,
\end{aligned}$$

where $\tilde{\omega}_F = \{K \in \mathcal{T} : \mathcal{N}_K \cap \mathcal{N}_F \neq \emptyset\}$ and $\widehat{\omega}_F = \{Q : K \in \mathcal{T}, Q \subset \partial K, F \subset Q\}$. Now, analogous to previous results, we find

$$\int_F j_2 w_F \geq (p+2)^{-4} \|j_2\|_{L^2(F)}^2$$

and for the right-hand side

$$\begin{aligned}
\sum_{K \in \tilde{\omega}_F} \int_K (\nabla_x \otimes \nabla_y)(u - u_\mathcal{T})(\nabla_x \otimes \nabla_y) w_F \,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{K \in \tilde{\omega}_F} |u - u_\mathcal{T}|_{H^{1,1}(K)} |w_F|_{H^{1,1}(K)} \\
&\leq \sum_{K \in \tilde{\omega}_F} |u - u_\mathcal{T}|_{H^{1,1}(K)} C h_F^{-1} \|j_2\|_{L^2(F)}, \\
\sum_{K \in \tilde{\omega}_F} \int_K r w_F \,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{K \in \tilde{\omega}_F} \|r\|_{L^2(K)} \|w_F\|_{L^2(K)} \\
&\leq \sum_{K \in \tilde{\omega}_F} \|r\|_{L^2(K)} C h_F \|j_2\|_{L^2(F)}, \\
\sum_{K \in \tilde{\omega}_F} \int_K (f - f_\mathcal{T}) w_F \,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{K \in \tilde{\omega}_F} \|f - f_\mathcal{T}\|_{L^2(K)} \|w_F\|_{L^2(K)} \\
&\leq \sum_{K \in \tilde{\omega}_F} \|f - f_\mathcal{T}\|_{L^2(K)} C h_F \|j_2\|_{L^2(F)}, \\
\sum_{Q \in \widehat{\omega}_F} \int_Q j_1 w_F \,\mathrm{d}x\,\mathrm{d}y &\leq \sum_{Q \in \widehat{\omega}_F} \|j_1\|_{L^2(Q)} \|w_F\|_{L^2(Q)} \\
&\leq \sum_{Q \in \widehat{\omega}_F} \|j_1\|_{L^2(Q)} C h_F^{1/2} \|j_2\|_{L^2(F)}.
\end{aligned}$$

This shows that

$$\begin{aligned}
C \|j_2\|_{L^2(F)}^2 \leq \sum_{K \in \tilde{\omega}_F} |u - u_\mathcal{T}|_{H^{1,1}(K)} C h_F^{-1} \|j_2\|_{L^2(F)} + \sum_{K \in \tilde{\omega}_F} \|r\|_{L^2(K)} C h_F \|j_2\|_{L^2(F)} \\
+ \sum_{K \in \tilde{\omega}_F} \|f - f_\mathcal{T}\|_{L^2(K)} C h_F \|j_2\|_{L^2(F)} + \sum_{Q \in \widehat{\omega}_F} \|j_1\|_{L^2(Q)} C h_F^{1/2} \|j_2\|_{L^2(F)}
\end{aligned}$$

and thus, by shape regularity, Lemma 2.63 and Lemma 2.64, we conclude

$$\begin{aligned} h_F\|j_2\|_{L^2(F)} \leq & C_1 \sum_{K\in\tilde{\omega}_F} |u-u_{\mathcal{T}}|_{H^{1,1}(K)} + \sum_{K\in\tilde{\omega}_F} Ch_K^2\|r\|_{L^2(K)} \\ & + \sum_{K\in\tilde{\omega}_F} Ch_K^2\|f-f_{\mathcal{T}}\|_{L^2(K)} + \sum_{Q\in\widehat{\omega}_F} Ch_Q^{3/2}\|j_1\|_{L^2(Q)} \\ \leq & c_1 \sum_{K\in\tilde{\omega}_F} |u-u_{\mathcal{T}}|_{H^{1,1}(K)} + c_2 \sum_{K\in\tilde{\omega}_F} h_K^2\|f-f_{\mathcal{T}}\|_{L^2(K)}. \end{aligned}$$

□

This concludes the inspection of a lower bound for the residual error estimator $\eta_{\mathcal{R}}$. In total we have shown the

Theorem 2.66. *Let u and $u_{\mathcal{T}}$ denote the solutions of the variational problems (2.62) and (2.63), respectively. Let the residual error estimator $\eta_{\mathcal{R},K}$ be given as in Definition 2.59. Moreover, let $f_{\mathcal{T}}$ denote an approximation of f on the mesh $\mathcal{T}$. There exists constants c^* and c_* that only depend on the shape regularity of the given mesh, the polynomial degree $\mathbf{p}_{\mathcal{T}}$, and $\mathcal{D}$, such that*

$$|u-u_{\mathcal{T}}|_{H^{1,1}(\mathcal{Q})} \leq c^* \left\{ \sum_{K\in\mathcal{T}} \eta_{\mathcal{R},K}^2 + \sum_{K\in\mathcal{T}} h_K^2\|f-f_{\mathcal{T}}\|_{L^2(K)}^2 \right\}^{1/2}$$

and for all $K\in\mathcal{T}$ there holds

$$h_K\eta_{\mathcal{R},K} \leq c_* \left\{ |u-u_{\mathcal{T}}|_{H^{1,1}(\tilde{\omega}_K)}^2 + \sum_{K'\subset\tilde{\omega}_K} h_{K'}^4\|f-f_{\mathcal{T}}\|_{L^2(K')}^2 \right\}^{1/2},$$

where $\tilde{\omega}_K$ again denotes the union of all elements K' that share at least one vertex with K.

Proof. The upper bound is proven in Lemma 2.60 and Corollary 2.61. The lower bound follows from the combination of Lemma 2.63, Lemma 2.64, and Lemma 2.65 which yields with shape regularity that

$$\begin{aligned} h_K^2\eta_{\mathcal{R},K}^2 \leq & c \left(h_K^4\|r\|_{L^2(K)}^2 + \frac{1}{2}\sum_{Q\subset\partial K} h_Q^3\|j_1\|_{L^2(Q)}^2 + \frac{1}{4}\sum_{\substack{F_{xy}\subset Q:\\ Q\subset\partial K}} h_{F_{xy}}^3\|j_2\|_{L^2(F_{xy})}^2 \right) \\ \leq & c \left(h_K^4\|r\|_{L^2(K)}^2 + \frac{1}{2}\sum_{Q\subset\partial K} h_Q^3\|j_1\|_{L^2(Q)}^2 + \frac{1}{4}\sum_{\substack{F_{xy}\subset Q:\\ Q\subset\partial K}} h_{F_{xy}}^2\|j_2\|_{L^2(F_{xy})}^2 \right) \\ \leq & c \left(\|(\nabla_x\otimes\nabla_y)(u-u_{\mathcal{T}})\|_{L^2(\tilde{\omega}_K)}^2 + \sum_{K'\subset\tilde{\omega}_K} h_{K'}^4\|f_{\mathcal{T}}-f\|_{L^2(K')}^2 \right) \end{aligned}$$

and so we conclude the proof. □

2.3.5 A hierarchical error estimator

As for the two dimensional situation we consider a conforming finite dimensional Finite Element space $X_{\mathcal{T}}$ that contains $\mathcal{S}_0^{1,0}(\mathcal{T}), p \in \mathbb{N}$, i.e.

$$\mathcal{S}_0^{p,0}(\mathcal{T}) \subset X_{\mathcal{T}} \subset H_0^{1,1}(\mathcal{D}).$$

As earlier (cf. section 2.2.5) let the space $X_{\mathcal{T}}$ be induced by either a uniform refinement of the whole mesh $\mathcal{T}$ or be of higher order. Furthermore, let us denote by $x_{\mathcal{T}} \in X_{\mathcal{T}}$ the solution of

$$\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) x_{\mathcal{T}} (\nabla_x \otimes \nabla_y) v_{\mathcal{T}} \,\mathrm{d}x \,\mathrm{d}y = \int_{\mathcal{D}} f v_{\mathcal{T}} \,\mathrm{d}x \,\mathrm{d}y, \quad \forall v_{\mathcal{T}} \in X_{\mathcal{T}}. \tag{2.79}$$

In order to compare the solution $u_{\mathcal{T}} \in \mathcal{S}_0^{1,0}(\mathcal{T})$ of (2.63) with $x_{\mathcal{T}}$, we subtract

$$\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) u_{\mathcal{T}} (\nabla_x \otimes \nabla_y) v_{\mathcal{T}} \,\mathrm{d}x \,\mathrm{d}y$$

from both sides of equation (2.79). This yields for all $v_{\mathcal{T}} \in X_{\mathcal{T}}$

$$\begin{aligned}
&\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y)(x_{\mathcal{T}} - u_{\mathcal{T}})(\nabla_x \otimes \nabla_y) v_{\mathcal{T}} \,\mathrm{d}x \,\mathrm{d}y \\
&= \int_{\mathcal{D}} f v_{\mathcal{T}} \,\mathrm{d}x \,\mathrm{d}y - \int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) u_{\mathcal{T}} (\nabla_x \otimes \nabla_y) v_{\mathcal{T}} \,\mathrm{d}x \,\mathrm{d}y \\
&= \int_{\mathcal{D}} (\nabla_x \otimes \nabla_y)(u - u_{\mathcal{T}})(\nabla_x \otimes \nabla_y) v_{\mathcal{T}} \,\mathrm{d}x \,\mathrm{d}y,
\end{aligned}$$

where $u \in H_0^{1,1}(\mathcal{D})$ denotes the unique solution of the variational formulation (2.62). Arguing as in the two dimensional setting assuming a saturation property for $X_{\mathcal{T}}$ we have the two sided bound

$$|x_{\mathcal{T}} - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq |u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq \frac{1}{1-\beta} |x_{\mathcal{T}} - u_{\mathcal{T}})|_{H^{1,1}(\mathcal{D})}.$$

This shows that we may use $|x_{\mathcal{T}} - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})}$ as an a posteriori error indicator. As for the one dimensional model problem we thusly consider the space $X_{\mathcal{T}}$ to admit a hierarchical splitting in the form

$$X_{\mathcal{T}} = \mathcal{S}_0^{1,0}(\mathcal{T}) \oplus Z_{\mathcal{T}}.$$

As long as the spaces $\mathcal{S}_0^{1,0}(\mathcal{T})$ and $Z_{\mathcal{T}}$ satisfy a strengthened Cauchy-Schwarz inequality (cf. Lemma 2.18), we can build a more efficient device. For the sake of completenes we shall proceed as in section 2.2.5 and let $z_{\mathcal{T}}$ be defined as the unique solution in $Z_{\mathcal{T}}$ of the variational defect problem

$$\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) z_{\mathcal{T}} (\nabla_x \otimes \nabla_y) \zeta_{\mathcal{T}} = \int_{\mathcal{D}} f \zeta_h - \int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) u_{\mathcal{T}} (\nabla_x \otimes \nabla_y) \zeta_{\mathcal{T}}$$

for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$. Under the given assumptions, repeating the arguments of Section 2.2.5 we arrive at the following two-sided bound for the error

$$|z_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq |u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq \frac{1}{(1-\beta)(1-\gamma)^{1/2}} |z_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})}$$

and want to use $|z_{\mathcal{T}}|_{H_0^{1,1}(\mathcal{D})}$ as an a posteriori error estimator.

We shall assume again the existence of a bilinear form $\mathcal{B}^* : Z_\mathcal{T} \times Z_\mathcal{T} \to \mathbb{R}$ which exhibits a diagonal stiffness matrix and additionally defines an equivalent norm on $Z_\mathcal{T}$ to $|\cdot|_{H^{1,1}(\mathcal{D})}$. This leads to an analogous two-sided error bound in the four dimensional setting

$$\sqrt{\lambda}\mathcal{B}^*(z_\mathcal{T}^*, z_\mathcal{T}^*)^{1/2} \leq |u - u_\mathcal{T}|_{H^{1,1}(\mathcal{D})} \leq \frac{\sqrt{\Lambda}}{(1-\beta)(1-\gamma)^{1/2}}\mathcal{B}^*(z_\mathcal{T}^*, z_\mathcal{T}^*)^{1/2}.$$

Let us now give an explicit choice for $Z_\mathcal{T}$ and $\mathcal{B}^*$. Setting

$$X_\mathcal{T} = \mathcal{S}_0^{1,0}(\mathcal{T}) \oplus (\mathcal{S}_0^{2,0}(\mathcal{T}) \backslash \mathcal{S}_0^{1,0}(\mathcal{T})),$$

which defines the hierarchical complement $Z_\mathcal{T} = \mathcal{S}_0^{2,0}(\mathcal{T}) \backslash \mathcal{S}_0^{1,0}(\mathcal{T})$ as the space of piecewise quadratic continuous polynomials in each component on $\mathcal{T}$ which vanish at the nodes $\mathcal{N}$ of $\mathcal{T}$.

To make this precise, we associate with every edge $E \in \mathcal{E}$, face $F \in \mathcal{F}$, cube $Q \in \mathcal{Q}$, and element $K \in \mathcal{T}$ the functions ψ_E, ψ_F, ψ_Q and ψ_K, respectively. These are the associated edge, face, cube and element bubbles, respectively. For the sake of completeness we shall state them shortly on the reference element $\hat{K}$. For simplicity of notation let us denote the coordinates on $\hat{K}$ by $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_4)$. Then letting $\hat{E}$ any edge, $\hat{F}$ any face, $\hat{Q}$ any cube of the reference element $\hat{K}$, we have the following representations:

$$\begin{aligned}
\hat{\psi}_{\hat{E}}(\hat{x}) &:= C \cdot (1-\hat{x}_i^2) \prod_{i \neq j} \tfrac{1 \pm \hat{x}_j}{2},\ i = 1, \ldots, 4, \\
\hat{\psi}_{\hat{F}}(\hat{x}) &:= C \cdot (1-\hat{x}_i^2)(1-\hat{x}_j^2) \prod_{k \neq i \wedge k \neq j} \tfrac{1 \pm \hat{x}_k}{2},\ i, j = 1, \ldots, 4, i \neq j, \\
\hat{\psi}_{\hat{Q}}(\hat{x}) &:= C \cdot \tfrac{1 \pm \hat{x}_i}{2} \prod_{i \neq j} (1-\hat{x}_j^2),\ i = 1, \ldots, 4, \\
\hat{\psi}_{\hat{K}}(\hat{x}) &:= C \prod_{i=1}^{4} (1-\hat{x}_i^2),
\end{aligned}$$

where C denotes a certain constant, which is usually used to normalize the function value of $\hat{\psi}_S$, $S \in \{\hat{E}, \hat{F}, \hat{Q}, \hat{K}\}$, at the barycenter of S. These functions represent all basis functions in $Z_\mathcal{T}$ on $\hat{K}$.

When defining $\mathcal{S}_0^{1,0}(\mathcal{T})$ with the help of tensor products of antiderivatives of the Legendre polynomials, we make a particular choice for $\mathcal{B}^*$, namely

$$\mathcal{B}^*(z_\mathcal{T}, \zeta_\mathcal{T}) = \mathcal{B}^*\left(\sum_S \alpha_S \psi_S, \sum_{S'} \alpha_{S'} \psi_{S'}\right) = \sum_{S,S'} \alpha_S \alpha_{S'} \int_\mathcal{D} (\nabla_x \otimes \nabla_y)\psi_S (\nabla_x \otimes \nabla_y)\psi_{S'},$$

where $S, S' \in \mathcal{E} \cup \mathcal{F} \cup \mathcal{Q} \cup \mathcal{T}$. For our choice of spaces Lemma 2.18 guarantees that the strengthened Cauchy-Schwarz inequality holds for a $\gamma \in [0, 1)$. Furthermore, $\gamma \neq 0$, since there holds for any $z \in \mathcal{N}_K$ that

$$\int_K (\nabla_x \otimes \nabla_y)\lambda_z \left(\sum_{E \in \mathcal{E}_K} (\nabla_x \otimes \nabla_y)\psi_E\right) \neq 0.$$

This is seen by a careful and rather tedious but straightforward computation. Moreover, this shows that in general we have to solve the defect problem:

Problem 2.67. *Find $z_\mathcal{T}^* \in Z_\mathcal{T}$, such that*

$$\mathcal{B}^*(z_\mathcal{T}^*, \zeta_\mathcal{T}) = \int_\mathcal{D} f\zeta_\mathcal{T} - \int_\mathcal{D} (\nabla_x \otimes \nabla_y)u_\mathcal{T}(\nabla_x \otimes \nabla_y)\zeta_\mathcal{T}, \quad \forall \zeta_\mathcal{T} \in Z_\mathcal{T}.$$

As there are 81 shape functions on an element for $p = 2$, the computation of the hierarchical error estimator is more costly than that of the presented residual error estimator for $p = 1$, since there are only 16 shape functions per element contributing to the error estimator. For this reason we are inclined to look for a more easily computable device.

If we make again the simplification of only enriching the Finite Element space with element bubble functions ψ_K, then the strengthened Cauchy-Schwarz inequality again holds with $\gamma = 0$. This again follows by a simple calculation using the orthogonality properties of tensorized Legendre polynomials. Since ψ_K's do not have overlapping support this leads to solving a diagonal system of equations which is of $O(N)$ complexity, where N now briefly denotes the number of elements in $\mathcal{T}$.

A straightforward calculation to find the coefficients for $z^*_{\mathcal{T}} = \sum_S \alpha_S \psi_S$ by testing with a certain $\psi_{S'}$ yields for all $\zeta_{\mathcal{T}} \in Z_{\mathcal{T}}$

$$\begin{aligned}\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) z^*_{\mathcal{T}} (\nabla_x \otimes \nabla_y)\psi_{S'} \,\mathrm{d}x &= \sum_S \alpha_S \int_{\mathcal{D}} (\nabla_x \otimes \nabla_y)\psi_S (\nabla_x \otimes \nabla_y)\psi_{S'} \,\mathrm{d}x \\ &= \alpha_{S'} \|(\nabla_x \otimes \nabla_y)\psi_{S'}\|^2_{L^2(\mathcal{D})} \\ &= \int_{\mathcal{D}} f\psi_{S'} \,\mathrm{d}x,\end{aligned}$$

which shows that

$$\alpha_S = \frac{\int_{\mathcal{D}} f\psi_S \,\mathrm{d}\vec{x}}{\|(\nabla_x \otimes \nabla_y)\psi_S\|^2_{L^2(\mathcal{D})}}.$$

Hence,

$$\begin{aligned}\mathcal{B}^*(z^*_{\mathcal{T}}, z^*_{\mathcal{T}}) &= \sum_S \alpha_S^2 \|(\nabla_x \otimes \nabla_y)\psi_S\|^2_{L^2(\mathcal{D})} \\ &= \sum_S \left(\frac{\int_{\mathcal{D}} f\psi_S \,\mathrm{d}\vec{x}}{\|(\nabla_x \otimes \nabla_y)\psi_S\|^2_{L^2(\mathcal{D})}} \right)^2 \|(\nabla_x \otimes \nabla_y)\psi_S\|^2_{L^2(\mathcal{D})} \\ &= \sum_S \frac{\left(\langle f, \psi_S\rangle_{L^2(\mathcal{D})}\right)^2}{\|(\nabla_x \otimes \nabla_y)\psi_S\|^2_{L^2(\mathcal{D})}}.\end{aligned}$$

For completeness' sake we procede as in the two dimensional setting. We define and state reliability as well as weak efficiency for the full hierarchical error estimator η_H. Since the proof is analogous to the two dimensional situation only bulkier, we omit most of it.

Definition 2.68. *We define the local hierarchical error estimator by means of the sum over the relevant basis functions in $Z_{\mathcal{T}}$ that are associated with the element patch ω_K, i.e.*

$$\eta^2_{H,K} := h_K^{-2} \|(\nabla_x \otimes \nabla_y) z^*_{\mathcal{T}}\|^2_{L^2(\omega_K)} \tag{2.80}$$

and moreover set the global error estimator to

$$\eta_H := \left(\sum_{K \in \mathcal{T}} \eta^2_{H,K} \right)^{1/2}. \tag{2.81}$$

Theorem 2.69. *Let u be the exact solution of (2.62) and $u_{\mathcal{T}}$ the corresponding solution of the discrete problem (2.63). Let $Z_{\mathcal{T}}$ as above and let the* hierarchical a posteriori *error*

estimator η_H *be given as in 2.81. Then there exist constants* $c^*, c_* > 0$*, such that* η_H *is* reliable

$$|u - u_{\mathcal{T}}|_{H^{1,1}(\mathcal{D})} \leq c^* \left\{ \eta_H^2 + \sum_{K \in \mathcal{T}} h_K^2 \|f_{\mathcal{T}} - f\|_{L^2(K)}^2 \right\}^{1/2} \tag{2.82}$$

and weakly efficient

$$h_K \eta_{H,K} \leq c_* \sum_{K' \subset \omega_K} \left(|u - u_{\mathcal{T}}|_{H^{1,1}(\omega_{K'})}^2 + \sum_{K'' \subset \omega_{K'}} h_{K''}^4 \|f - f_{\mathcal{T}}\|_{L^2(K'')}^2 \right)^{1/2}. \tag{2.83}$$

Proof. As we have a similar representation as in the two dimensional situation, i.e.

$$\begin{aligned} \|(\nabla_x \otimes \nabla_y) z_{\mathcal{T}}^*\|_{L^2(\omega_K)}^2 = &\sum_{K' \subset \omega_K} \alpha_{K'} \langle z_{\mathcal{T}}^*, \psi_{K'} \rangle_{H^{1,1}(K')} + \sum_{Q \in \mathcal{Q}_{\omega_K}} \alpha_Q \langle z_{\mathcal{T}}^*, \psi_Q \rangle_{H^{1,1}(\omega_Q \cap \omega_K)} \\ &+ \sum_{F \in \mathcal{F}_{\omega_K}} \alpha_F \langle z_{\mathcal{T}}^*, \psi_F \rangle_{H^{1,1}(\omega_F \cap \omega_K)} + \sum_{E \in \mathcal{E}_{\omega_K}} \alpha_E \langle z_{\mathcal{T}}^*, \psi_E \rangle_{H^{1,1}(\omega_E \cap \omega_K)}, \end{aligned}$$

using approximation results from previous sections and with analogous arguments as in the proof for the two dimensional hierarchical estimator the assertion follows. □

Remark 2.70. *As we shall not implement the full hierarchical error estimator, we define the local hierarchical "bubble" indicator as in the two dimensional setting by*

$$\hat{\eta}_{H,K}^2 := \alpha_K^2 h_K^{-2} \|(\nabla_x \otimes \nabla_y) \psi_K\|_K^2$$

and denote the global error estimator by

$$\hat{\eta}_H = \left(\sum_{K \in \mathcal{T}} \hat{\eta}_{H,K}^2 \right)^{1/2}.$$

2.3.6 An a posteriori estimator based on averaging

In the following we want to construct an averaging technique in the four dimensional setting. As a lot of the arguments and the approach are very similar to the situation in two dimensions we keep the presentation as short as possible.

To this end, suppose u solves the variational formulation

$$\int_{\mathcal{D}} (\nabla_x \otimes \nabla_y) u (\nabla_x \otimes \nabla_y) v = \int_{\mathcal{D}} f v, \quad \forall v \in H_0^{1,1}(\mathcal{D})$$

and $u_{\mathcal{T}}$ denotes the solution of the corresponding discrete formulation with $p = 1$. Similarly to the two dimensional situation we state the following definition of η_Z.

Definition 2.71. *The elementwise error indicator* η_Z *is given by*

$$\eta_{Z,K} := \min_{q \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}} \|(\nabla_x \otimes \nabla_y) u_{\mathcal{T}} - q\|_{L^2(K)}$$

and the global error estimator as

$$\eta_Z = \left(\sum_{K \in \mathcal{T}} \eta_{Z,K}^2 \right)^{1/2}.$$

Suppose that η_Z is reliable. Note that $(\nabla_x \otimes \nabla_y) u_\mathcal{T} \in (\mathcal{P}_1(\mathcal{T}))^{2\times 2}$ is a matrix valued function with in general discontinuous components in $\mathcal{P}_1(\mathcal{T})$. Furthermore, assume that we have at our disposal an easily computable function $\mathcal{A}((\nabla_x \otimes \nabla_y) u_\mathcal{T}) \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$ with an operator $\mathcal{A} : (\mathcal{P}_1(\mathcal{T}))^{2\times 2} \to \big((\mathcal{S}^{1,0}(\mathcal{T}))\big)^{2\times 2}$, where we fix the notation

$$\mathcal{A}((\nabla_x \otimes \nabla_y) u_\mathcal{T}) := \begin{pmatrix} \mathcal{A}_{1,1}(\partial_{x_1}\partial_{y_1} u_\mathcal{T}) & \mathcal{A}_{1,2}(\partial_{x_1}\partial_{y_2} u_\mathcal{T}) \\ \mathcal{A}_{2,1}(\partial_{x_2}\partial_{y_1} u_\mathcal{T}) & \mathcal{A}_{2,2}(\partial_{x_2}\partial_{y_2} u_\mathcal{T}) \end{pmatrix}$$

with operators $\mathcal{A}_{i,j} : \mathcal{P}_1(\mathcal{T}) \to \mathcal{S}^{1,0}(\mathcal{T})$, $i,j = 1,2$, which yields an approximation to $(\nabla_x \otimes \nabla_y) u$. Reliability immediately follows by setting $q = \mathcal{A}(\partial_x \partial_y u_\mathcal{T})$, since

$$\eta_{Z,K} \le \eta_{\mathcal{A},K} := \|(\nabla_x \otimes \nabla_y) u - \mathcal{A}((\nabla_x \otimes \nabla_y) u_\mathcal{T})\|_{L^2(K)}.$$

In this situation we opt for L^2-projecting the components of $(\nabla_x \otimes \nabla_y) u_\mathcal{T}$ into the space of piecewise continuous four-linear functions on $\mathcal{T}$, i.e.

$$\int_\mathcal{D} \mathcal{A}_{i,j}((\partial_{x_i} \otimes \partial_{y_j}) u_\mathcal{T}) v_\mathcal{T} = \int_\mathcal{D} (\partial_{x_i} \otimes \partial_{y_j}) u_\mathcal{T} v_\mathcal{T}, \quad \forall v_\mathcal{T} \in \mathcal{S}^{1,0}(\mathcal{T}), i,j = 1,2.$$

However, computing these projections is as costly as computing the discrete solution itself and is thus not a viable option. We resort again to an approximation of the $L^2(\mathcal{D})$-inner product.

Denote by $\mathcal{W}_\mathcal{T}$ the space of all piecewise quad-linear functions on $\mathcal{T}$, i.e. $\mathcal{W}_\mathcal{T} = \mathcal{P}_1(\mathcal{T}) = \{v : v|_K \in \mathcal{P}_1(K), K \in \mathcal{T}\}$, and set $\mathcal{V}_\mathcal{T} = \mathcal{W}_\mathcal{T} \cap C(\mathcal{D})$. Note that $(\partial_{x_i} \otimes \partial_{y_j})\mathcal{P}_1(\mathcal{T}) \subset \mathcal{W}_\mathcal{T}, i,j = 1,2$ and $\mathcal{V}_\mathcal{T} = \mathcal{S}^{1,0}(\mathcal{T})$. As approximation of the inner product, we define a mesh-dependent inner product $(\cdot,\cdot)_\mathcal{T}$ on $\mathcal{W}_\mathcal{T}$ via tensorization of the trapezoidal rule in every coordinate direction, which leads to

$$(v,w)_\mathcal{T} := \sum_{K\in\mathcal{T}} \frac{|K|}{16} \left(\sum_{z\in\mathcal{N}_K} v|_K(z) w|_K(z) \right),$$

where $|K|$ denotes the four dimensional Lebesgue measure of K. Note, that for either v or w being a piecewise constant function we have

$$(v,w)_\mathcal{T} = \int_\mathcal{D} vw \,\mathrm{d}x.$$

With this in place, we define $\mathcal{A}_{i,j}((\partial_{x_i} \otimes \partial_{y_j}) u_\mathcal{T})$ to be the $(\cdot,\cdot)_\mathcal{T}$-projection of $(\partial_{x_i} \otimes \partial_{y_j}) u_\mathcal{T}$ onto $\mathcal{V}_\mathcal{T}$ for $i,j = 1,2$, i.e.

$$(\mathcal{A}_{i,j}((\partial_{x_i} \otimes \partial_{y_j}) u_\mathcal{T}), v_\mathcal{T})_\mathcal{T} = ((\partial_{x_i} \otimes \partial_{y_j}) u_\mathcal{T}, v_\mathcal{T})_\mathcal{T}, \qquad \forall v_\mathcal{T} \in \mathcal{V}_\mathcal{T} \tag{2.84}$$

and define the associated elementwise error indicator and the global error estimator as follows.

Definition 2.72. *We define the elementwise error indicator as*

$$\eta_{\mathcal{A},K} := \|(\nabla_x \otimes \nabla_y) u_\mathcal{T} - \mathcal{A}((\nabla_x \otimes \nabla_y) u_\mathcal{T})\|_{L^2(K)}$$

and the global error estimator as

$$\eta_\mathcal{A} = \left(\sum_{K\in\mathcal{T}} \eta_{\mathcal{A},K}^2 \right)^{1/2}.$$

Note that for all $v, w \in \mathcal{V}_\mathcal{T}$ there holds

$$(v, w)_\mathcal{T} = \frac{1}{16} \sum_{z \in \mathcal{N}} |\omega_z| v(z) w(z) \tag{2.85}$$

and by inserting the nodal function for z instead of $v_\mathcal{T}$ into (2.84) we find for all $z \in \mathcal{N}$ and $i, j = 1, 2$ that

$$\mathcal{A}_{i,j}(\partial_{x_i}\partial_{y_j} u_\mathcal{T})(z) = \sum_{K \subset \omega_z} \frac{|K|}{|\omega_z|} \left(\partial_{x_i}\partial_{y_j} u_\mathcal{T}\big|_K\right)(z). \tag{2.86}$$

Thus, for $x \in \mathcal{D}$ we have the representation

$$\mathcal{A}_{i,j}(\partial_{x_i}\partial_{y_j} u_\mathcal{T})(x) = \sum_{z \in \mathcal{N}} \mathcal{A}_{i,j}(\partial_{x_i}\partial_{y_j} u_\mathcal{T})(z) \lambda_z(x) = \sum_{z \in \mathcal{N}} \alpha_z^{i,j} \lambda_z(x) \in \mathcal{S}^{1,0}(\mathcal{T}),$$

where for ease of notation we have set

$$\alpha_z^{i,j} := \mathcal{A}_{i,j}(\partial_{x_i}\partial_{y_j} u_\mathcal{T})(z).$$

We shall now procede to show reliability as we have done in the two dimensional situation. Here we note, since

$$\|(\nabla_x \otimes \nabla_y) u_\mathcal{T} - \mathcal{A}((\nabla_x \otimes \nabla_y) u_\mathcal{T})\|^2_{L^2(K)} = \sum_{i,j=1}^{2} \|(\partial_{x_i}\partial_{y_j}) u_\mathcal{T} - \mathcal{A}_{i,j}((\partial_{x_i}\partial_{y_j}) u_\mathcal{T})\|^2_{L^2(K)},$$

that it suffices to adapt Lemma 2.43 to the four dimensional situation.

Lemma 2.73. *Let $u, q \in H^{2,2}(\mathcal{D}) \cap H_0^{1,1}(\mathcal{D})$ and $u_\mathcal{T}$ denote the solution of the corresponding discrete problem (2.63) with*

$$\int_\mathcal{D} (\nabla_x \otimes \nabla_y)(u - u_\mathcal{T})(\nabla_x \otimes \nabla_y) w_\mathcal{T} \,\mathrm{d}x = 0, \quad \forall w_\mathcal{T} \in \mathcal{S}^{1,0}(\mathcal{T}).$$

Furthermore, denote by $f_\mathcal{T}$ an elementwise approximation of f. Then there holds

$$\begin{aligned} |u - u_\mathcal{T}|_{H^{1,1}(\mathcal{D})} \leq C \inf_{q \in H^{2,2}(\mathcal{D})} \Bigg\{ & \|(\nabla_x \otimes \nabla_y)(u_\mathcal{T} - q)\|_{L^2(\mathcal{D})} \\ & + \left(\sum_{K \in \mathcal{T}} h_K^2 \|f - (\Delta_x \otimes \Delta_y) q\|^2_{L^2(K)} \right)^{1/2} \Bigg\} \\ & + C \left(\sum_{K \in \mathcal{T}} h_K^2 \|f - f_\mathcal{T}\|^2_{L^2(K)} \right)^{1/2}. \end{aligned}$$

Proof. The proof carries over verbatim when using the quasi-interpolation operator $\mathbb{I}_\mathcal{T}$ instead of $\Pi^{1,1}$ and is therefore omitted for brevity. □

Theorem 2.74. *Let $\mathcal{A}((\nabla_x \otimes \nabla_y) u_\mathcal{T}) \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$ be the average of $(\nabla_x \otimes \nabla_y) u_\mathcal{T}$ defined via (2.84)–(2.86). Furthermore, let u denote the solution of (2.62) and $u_\mathcal{T}$ the solution of (2.63). Then the error estimator $\eta_\mathcal{A}$ defined above is* reliable, *i.e. there holds*

$$\begin{aligned} |u - u_\mathcal{T}|_{H^{1,1}(\mathcal{D})} \leq C\eta_\mathcal{A} + C & \left(\sum_{K \in \mathcal{T}} h_K^2 \|f - (\nabla_x \otimes \nabla_y) \cdot (\mathcal{A}((\nabla_x \otimes \nabla_y) u_\mathcal{T}))\|^2_{L^2(K)} \right)^{1/2} \\ & + C \left(\sum_{K \in \mathcal{T}} h_K^2 \|f - f_\mathcal{T}\|^2_{L^2(K)} \right)^{1/2}. \end{aligned}$$

Let us now turn our focus to the efficiency of $\eta_{\mathcal{A}}$, which we achieve by proving equivalence of $\eta_{\mathcal{A}}$ with η_Z. The following lemma is analogous to the two dimensional situation, but for the sake of clarity and completeness the proof is not omitted.

Lemma 2.75. *There exists a uniform constant $b > 0$ such that*

$$\eta_Z \leq \eta_{\mathcal{A}} \leq \left(1 + \frac{16}{9}b\right)\eta_Z.$$

Proof. Since $\mathcal{A}(\nabla_x \otimes \nabla_y u_{\mathcal{T}}) \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$, it clearly holds that

$$\begin{aligned}\eta_Z =& \min_{q\in(\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}} \|(\nabla_x \otimes \nabla_y)(u_{\mathcal{T}} - q)\|_{L^2(\mathcal{D})} \\ \leq& \|(\nabla_x \otimes \nabla_y)u_{\mathcal{T}} - \mathcal{A}((\nabla_x \otimes \nabla_y)u_{\mathcal{T}})\|_{L^2(\mathcal{D})} = \eta_{\mathcal{A}}.\end{aligned}$$

In order to prove the upper bound we have a look at the L^2-stability of the averaging operator. First we note that by a Cauchy-Schwarz for sums there holds

$$\begin{aligned}\|\mathcal{A}((\nabla_x \otimes \nabla_y)u_{\mathcal{T}})\|^2_{L^2(\mathcal{D})} &= \sum_{i,j=1}^{2} \|\mathcal{A}_{i,j}((\partial_{x_i}\partial_{y_j})u_{\mathcal{T}})\|^2_{L^2(\mathcal{D})} \\ &= \sum_{i,j=1}^{2} \int_{\mathcal{D}} \left|\sum_{z\in\mathcal{N}} \alpha_z^{i,j}\lambda_z\right|^2 \mathrm{d}x \\ &\leq 16 \sum_{i,j=1}^{2} \int_{\mathcal{D}} \sum_{z\in\mathcal{N}} |\alpha_z^{i,j}|^2|\lambda_z|^2 \,\mathrm{d}x.\end{aligned}$$

We proceed as in the two dimensional situation. If we denote by M the element mass-matrix on K and by $\tilde{M}$ the element mass-matrix on K scaled by $|K|^{-1}$, i.e.

$$\tilde{M} = (\tilde{m}_{ij})_{i,j=1}^{n} \quad \text{with} \quad \tilde{m}_{ij} = |K|^{-1}\int_K \lambda_{z_i}\lambda_{z_j}\,\mathrm{d}x,$$

where $n = |\mathcal{N}_K|$ is the number of nodes of an element K and the indices realize a certain fixed enumeration of the nodes of K. Noting that for any $p = \sum_{z\in\mathcal{N}_K} p_z\lambda_z \in \mathcal{P}_1(K)$, where $p_z = p|_K(z)$, there holds

$$\|p\|^2_{L^2(K)} = \int_K \left(\sum_{z\in\mathcal{N}_K} p_z\lambda_z\right)^2 \mathrm{d}x = \sum_{i,j=1}^{n} p_{z_i}\cdot\int_K \lambda_{z_i}\lambda_{z_j}\,\mathrm{d}x\cdot p_{z_j} = \underline{p}^\top\cdot M\underline{p},$$

where by $\underline{p}$ we denote the vector of coefficients of p on K, i.e. $\underline{p} = (p_{z_1}, p_{z_2}, ..., p_{z_n})^\top$. Hence, by a Rayleigh quotient argument we find by letting $\tilde{\lambda}_1$ to be the smallest positive eigenvalue of $\tilde{M}$ that

$$\tilde{\lambda}_1|p_z|^2 \leq \tilde{\lambda}_1\sum_{z\in\mathcal{N}_K}|p_z|^2 = \tilde{\lambda}_1\underline{p}^\top\cdot\underline{p} \leq \underline{p}^\top\cdot\tilde{M}\underline{p} = |K|^{-1}\|p\|^2_{L^2(K)}.$$

In the four dimensional situation for hypercubes K, a computation of the mass on $\hat{K}$ and multiplying the eigenvalues by $|\hat{K}|^{-1} = \frac{1}{16}$, we find that $\tilde{\lambda}_1 = \frac{1}{1296}$ and so, for any $v \in \mathcal{P}_1(K)$ there holds

$$|v|_K(z)|^2 \leq \frac{1296}{|K|}\|v\|^2_{L^2(K)}, \tag{2.87}$$

which furthermore implies that there exists a uniform constant $b > 0$ such that

$$|v|_K(z)| \leq \frac{b}{|\omega_z|^{1/2}} \|v\|_{L^2(\omega_z)}.$$

we have

$$\begin{aligned}
\|\mathcal{A}((\nabla_x \otimes \nabla_y) u_{\mathcal{T}})\|^2_{L^2(\mathcal{D})} \leq & 16 \sum_{i,j=1}^{2} \int_{\mathcal{D}} \sum_{z \in \mathcal{N}} |\alpha_z^{i,j}|^2 |\lambda_z|^2 \, \mathrm{d}x \\
\leq & 16 \sum_{i,j=1}^{2} \int_{\mathcal{D}} \sum_{z \in \mathcal{N}} |\alpha_z^{i,j}|^2 |\lambda_z|^2 \, \mathrm{d}x \\
\leq & 16 \sum_{i,j=1}^{2} \sum_{z \in \mathcal{N}} \frac{b^2 \|\partial_{x_i} \partial_{y_j} u_{\mathcal{T}}\|^2_{L^2(\omega_z)}}{|\omega_z|} \int_{\omega_z} |\lambda_z|^2 \, \mathrm{d}x \\
= & 16 \sum_{i,j=1}^{2} \sum_{z \in \mathcal{N}} \frac{b^2 \|\partial_{x_i} \partial_{y_j} u_{\mathcal{T}}\|^2_{L^2(\omega_z)}}{|\omega_z|} \frac{|\omega_z|}{81} \\
= & \frac{256}{81} b^2 \sum_{i,j=1}^{2} \|\partial_{x_i} \partial_{y_j} u_{\mathcal{T}}\|^2_{L^2(\mathcal{D})} \\
= & \frac{256}{81} b^2 \|(\nabla_x \otimes \nabla_y) u_{\mathcal{T}}\|^2_{L^2(\mathcal{D})},
\end{aligned} \tag{2.88}$$

since $\|\lambda_z\|^2_{L^2(\omega_z)} = \frac{|\omega_z|}{81}$ and each element K is counted sixteen times.

Let $(\nabla_x \otimes \nabla_y) u_{\mathcal{T}} \in (\mathcal{P}_1(\mathcal{T}))^{2\times 2}$. Then there is a unique decomposition $(\nabla_x \otimes \nabla_y) u_{\mathcal{T}} = u_c + u_d$ with a continuous component $u_c \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$ and a component u_d of the orthogonal complement of $(\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$ in $(L^2(\mathcal{D}))^{2\times 2}$. Note that for any $p \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$ that $\mathcal{A}(p)(z) = p(z)$ and thus $\mathcal{A}$ is the identity on $(\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$. Hence,

$$\begin{aligned}
& \|(\nabla_x \otimes \nabla_y) u_{\mathcal{T}} - \mathcal{A}((\nabla_x \otimes \nabla_y) u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \\
= & \|(u_c + u_d) - \mathcal{A}(u_c + u_d)\|_{L^2(\mathcal{D})} \\
= & \|u_d - \mathcal{A}(u_d)\|_{L^2(\mathcal{D})} \\
\leq & \|u_d\|_{L^2(\mathcal{D})} + \|\mathcal{A}(u_d)\|_{L^2(\mathcal{D})} \\
= & \left(1 + \frac{16}{9} b\right) \|u_d\|_{L^2(\mathcal{D})} \\
= & \left(1 + \frac{16}{9} b\right) \min_{q \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}} \|(\nabla_x \otimes \nabla_y) u_{\mathcal{T}} - q\|_{L^2(\mathcal{D})} \\
= & \left(1 + \frac{16}{9} b\right) \eta_Z,
\end{aligned}$$

since $u_c \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$. □

Remark 2.76 (Concerning b). *For all $v \in \mathcal{P}_1(\mathcal{T}_z)$, where $\mathcal{T}_z := \{K : K \subset \omega_z\}$, by (2.87) there holds*

$$\begin{aligned}
|\omega_z|^{1/2} |\mathcal{A}((\nabla_x \otimes \nabla_y) u_{\mathcal{T}})(z)| \leq & 36 \sum_{K \in \mathcal{T}_z} \frac{|\omega_z|^{1/2}}{|K|^{1/2}} \frac{|K|}{|\omega_z|} \|(\nabla_x \otimes \nabla_y) u_{\mathcal{T}}\|_{L^2(K)} \\
\leq & 36 \left(\sum_{K \in \mathcal{T}_z} \frac{|\omega_z|}{|K|} \frac{|K|^2}{|\omega_z|^2} \right)^{1/2} \|(\nabla_x \otimes \nabla_y) u_{\mathcal{T}}\|_{L^2(\omega_z)}
\end{aligned}$$

and thusly the constant $b > 0$ *in* (2.88) *is*

$$b = \max_{z \in \mathcal{N}} 36 \left(\sum_{K \in \mathcal{T}_z} \frac{|K|}{|\omega_z|} \right)^{1/2} = 36.$$

Theorem 2.77. *Let* $\mathcal{A}((\nabla_x \otimes \nabla_y) u_{\mathcal{T}}) \in (\mathcal{S}^{1,0}(\mathcal{T}))^{2\times 2}$ *be the average of* $(\nabla_x \otimes \nabla_y) u_{\mathcal{T}}$ *defined in a componentwise fashion via* (2.84)–(2.86). *Furthermore, let* u *denote the solution of* (2.62) *and* $u_{\mathcal{T}}$ *the solution of* (2.63). *Then the error estimator* $\eta_{\mathcal{A}}$ *defined above is* asymptotically exact.

Proof. A combination of Lemma 2.73 and Lemma 2.75 readily implies that $\eta_{\mathcal{A}}$ is *asymptotically exact.* □

2.4 On the convergence of the adaptive process

In order to understand the effect that *weak efficiency*, i.e. the effect of the lower bound of the residual a posteriori error estimator $\eta_{\mathcal{R}}$ and the hierarchical error estimator η_H, has on the convergence of the adaptive process, we shall investigate the convergence of the adaptive process for these cases. As the structure of the lower bounds is the same for $\eta_{\mathcal{R}}$ and η_H in the two and four dimensional situation, we carry out the following analysis only for the residual error estimator $\eta_{\mathcal{R}}$ in two dimensions. The arguments for the other cases are completely analogous. The presentation closely follows [54].

Consider $\mathcal{T}_1$ to be a mesh on $\mathcal{D}$ and $\mathcal{T}_2$ be a refinement of $\mathcal{T}_1$, such that the associated Finite Element spaces are nested, i.e.

$$\mathcal{S}_0^{1,0}(\mathcal{T}_1) \subset \mathcal{S}_0^{1,0}(\mathcal{T}_2).$$

Moreover, let u denote the exact solution to the exact variational problem and let u_1 as well as u_2 denote solutions to the discrete problems on the meshes $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. Then by Galerkin orthogonality we have the relation

$$\|\partial_x \partial_y (u - u_2)\|^2_{L^2(\mathcal{D})} = \|\partial_x \partial_y (u - u_1)\|^2_{L^2(\mathcal{D})} - \|\partial_x \partial_y (u_1 - u_2)\|^2_{L^2(\mathcal{D})}. \tag{2.89}$$

Let us first consider the case in which the right-hand side f is piecewise constant on $\mathcal{T}_1$, i.e. for example we might exchange it with its L^2-projection on the piecewise constant functions on $\mathcal{T}_1$. We will remove this restriction later on. Then by reliability of the residual error estimator we have

$$\|\partial_x \partial_y (u - u_1)\|^2_{L^2(\mathcal{D})} \leq c^{*2} \sum_{K \in \mathcal{T}_1} \eta^2_{\mathcal{R},K}, \tag{2.90}$$

because the data oscillation term vanishes in this case. Now, let $\vartheta \in (0,1)$ and find $\tilde{\mathcal{T}}_1 \subset \mathcal{T}_1$ with the property that

$$\sum_{K \in \tilde{\mathcal{T}}_1} \eta^2_{\mathcal{R},K} \geq \vartheta \sum_{K \in \mathcal{T}_1} \eta^2_{\mathcal{R},K}. \tag{2.91}$$

Then by (2.90) and (2.91) there holds

$$\|\partial_x \partial_y (u - u_1)\|^2 \leq \frac{c^{*2}}{\vartheta} \sum_{K \in \tilde{\mathcal{T}}_1} \eta^2_{\mathcal{R},K}. \tag{2.92}$$

Since we want to make use of the same arguments as in the proofs on reliability and (weak) efficiency, we make the following assumption.

Assumption 2.78. *Let $\mathcal{T}_2$ be a refined partition of $\tilde{\mathcal{T}}_1$ that satisfies the following conditions:*

(i) the midpoint of every edge in $\tilde{\mathcal{T}}_1$ is a vertex of an element in $\mathcal{T}_2$;

(ii) there exists a point x in the interior of every element in $\tilde{\mathcal{T}}_1$, such that x is a vertex of an element in $\mathcal{T}_2$.

The foregoing assumption can be fulfilled by isotropically subdividing every element in $\tilde{\mathcal{T}}_1$, i.e. into 4 smaller squares/rectangles. Then the assumption implies that we are allowed exchange the edge and face bubble functions in the proofs of reliability and efficiency with certain global nodal functions and hence the functions w_K and w_E do belong to the space $\mathcal{S}_0^{1,0}(\mathcal{T}_2)$. This allows us to replace u with u_2 in the estimates for the efficiency of $\eta^2_{\mathcal{R},K}$. Thus,

$$\sum_{K\in\tilde{\mathcal{T}}_1} h_K^2\eta^2_{\mathcal{R},K} \leq c_*^2\|\partial_x\partial_y(u_2-u_1)\|^2_{L^2(\mathcal{D})}. \tag{2.93}$$

Moreover, a combination of (2.92) and (2.93) reveals that there holds

$$\begin{aligned}
-\|\partial_x\partial_y(u_2-u_1)\|^2_{L^2(\mathcal{D})} &\leq -\frac{1}{c_*^2}\sum_{K\in\tilde{\mathcal{T}}_1} h_K^2\eta^2_{\mathcal{R},K}\\
&\leq -\frac{1}{c_*^2}\min_{K\in\tilde{\mathcal{T}}_1} h_K^2 \sum_{K\in\tilde{\mathcal{T}}_1}\eta^2_{\mathcal{R},K}\\
&\leq -\frac{\vartheta}{c_*^2c^{*2}}\min_{K\in\tilde{\mathcal{T}}_1} h_K^2\|\partial_x\partial_y(u-u_1)\|^2_{L^2(\mathcal{D})}.
\end{aligned}$$

In lieu of (2.89) we thus find

$$\begin{aligned}
\|\partial_x\partial_y(u-u_2)\|^2_{L^2(\mathcal{D})} &= \|\partial_x\partial_y(u-u_1)\|^2_{L^2(\mathcal{D})} - \|\partial_x\partial_y(u_1-u_2)\|^2_{L^2(\mathcal{D})}\\
&\leq \left(1-\min_{K\in\tilde{\mathcal{T}}_1} h_K^2\frac{\vartheta}{c_*^2c^{*2}}\right)\|\partial_x\partial_y(u-u_1)\|^2_{L^2(\mathcal{D})}.
\end{aligned}$$

From this we immediately see that the error reduction factor is given by

$$\gamma := \left(1-\min_{K\in\tilde{\mathcal{T}}_1} h_K^2\tfrac{\vartheta}{c_*^2c^{*2}}\right).$$

Since the minimum element diameter of the marked elements enters γ for $\eta_{\mathcal{R}}$ and η_H, this implies that γ might be very close to one if the minimal element diameter of $\tilde{\mathcal{T}}_1$ is small and is then responsible for a deterioration of the convergence. This can lead to a convergence shelf behavior for the residual and hierarchical error estimators, since in many steps the error reduction may become very small. The adaptive process will still converge, but is obviously seriously hampered by the element diameter of the smallest marked element. This is a structural drawback of the estimates for the residual and hierarchical error estimators, which cannot be overcome as such.

Remark 2.79. *Note that since $\eta_{\mathcal{A}}$ is asymptotically exact this drawback does not affect the a posteriori error estimator based on averaging, because the meshwidth of $\tilde{\mathcal{T}}_1$ does not enter the error reduction factor γ, which then reads*

$$\gamma = \left(1-\tfrac{\vartheta}{c_*^2c^{*2}}\right).$$

Remark 2.80. *One idea to alleviate the convergence $\eta_{\mathcal{R}}$ and η_H under these circumstances, when the adaptive method hits a convergence shelf, is to monitor the convergence rate between adaptive steps and to force a mandatory uniform refinement when the convergence drops below a given threshold and then to proceed with the adaptive process.*

In order to overcome the restriction that f be piecewise constant we consider $f \in L^2(\mathcal{D})$ and $\mathcal{T}$ an arbitrary mesh on which the piecewise constant function $f_{\mathcal{T}}$ is defined by means of the weighted integral average of f on an element $K \in \mathcal{T}$, i.e.

$$f_{\mathcal{T}} := \frac{1}{|K|}\int_K f.$$

Let u the exact solution with right-hand side f and denote by $\tilde{u}$ and $\tilde{u}_{\mathcal{T}}$ the exact and discrete solution with right-hand side $f_{\mathcal{T}}$, respectively. Moreover, denote by $u_{\mathcal{T}}$ the solution of the discrete problem on $\mathcal{T}$ with right-hand side f. Because $u_{\mathcal{T}}$ is the best approximation of u in $S_0^{1,0}(\mathcal{T})$ with respect to the energy norm $\|\partial_x\partial_y(\cdot)\|_{L^2(\mathcal{D})}$ there holds

$$\|\partial_x\partial_y(u-u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \|\partial_x\partial_y(u-\tilde{u}_{\mathcal{T}})\|_{L^2(\mathcal{D})}$$

and furthermore the triangle inequality provides

$$\|\partial_x\partial_y(u-u_{\mathcal{T}})\|_{L^2(\mathcal{D})} \leq \|\partial_x\partial_y(u-\tilde{u})\|_{L^2(\mathcal{D})} + \|\partial_x\partial_y(\tilde{u}-\tilde{u}_{\mathcal{T}})\|_{L^2(\mathcal{D})}.$$

The second term of the previous inequality can be dealt with as above. We proceed with inspection of the first term.

Since u and $\tilde{u}$ solve the variational problem with right-hand side f and $f_{\mathcal{T}}$, respectively, we find that for all $w \in H_0^{1,1}(\mathcal{D})$ there holds

$$\int_{\mathcal{D}} \partial_x\partial_y(u-\tilde{u})\partial_x\partial_y w = \int_{\mathcal{D}} (f-f_{\mathcal{T}})w.$$

When we consider $f_{\mathcal{T}}$ to be the piecewise constant L^2-projection on $\mathcal{T}$, we see by Galerkin orthogonality that

$$\int_{\mathcal{D}} (f-f_{\mathcal{T}})w = \int_{\mathcal{D}} (f-f_{\mathcal{T}})(w-w_{\mathcal{T}}) = \sum_{K\in\mathcal{T}} \int_K (f-f_K)(w-w_K),$$

where $f_{\mathcal{T}}|_K = f_K$ and $w_{\mathcal{T}}|_K = w_K$ with

$$f_K = \frac{1}{|K|}\int_K f, \qquad w_K = \frac{1}{|K|}\int_K w.$$

Then the Cauchy-Schwarz inequality implies that

$$\int_{\mathcal{D}} \partial_x\partial_y(u-\tilde{u})\partial_x\partial_y w \leq \sum_{K\in\mathcal{T}} \|f-f_K\|_{L^2(K)}\|w-w_K\|_{L^2(K)}$$

and since every element is convex, a Poincaré inequality and the equivalence of the energy norm with the $H^{1,1}$-norm yield

$$\|w-w_K\|_{L^2(K)} \leq \frac{h_K}{\pi}\|\nabla w\|_{L^2(K)} \leq c\frac{h_K}{\pi}\|\partial_x\partial_y w\|_{L^2(K)}.$$

Summing over all $K \in \mathcal{T}$ we find with the Cauchy-Schwarz inequality for sums

$$\|\partial_x\partial_y(u-\tilde{u})\|^2_{L^2(\mathcal{D})} \leq \frac{C}{\pi}\sum_{K\in\mathcal{T}} h_K^2\|f-f_K\|^2_{L^2(K)}.$$

This shows that for general right-hand sides with $f \in L^2(\mathcal{D})$ we have to control the data oscillation to obtain overall convergence. This can be achieved as follows.

Consider again a mesh $\mathcal{T}_2$ which is a refinement of $\mathcal{T}_1$. Given parameters $\vartheta \in (0,1)$, we apply the four steps of the adaptive algorithm, i.e. (1) Solve, (2) Estimate, (3) Mark, (4) Refine, with the following modifications. Step (1) is omitted and in (2) the generic error estimator η_K^2 is replaced by the data oscillation term $h_K^2\|f - f_K\|_{L^2(K)}^2$ to determine a subset $\tilde{\mathcal{T}}_1$ of $\mathcal{T}_1$ with the property

$$\sum_{K\in\tilde{\mathcal{T}}_1} h_K^2\|f - f_K\|_{L^2(K)}^2 \geq \vartheta \sum_{K\in\mathcal{T}_1} h_K^2\|f - f_K\|_{L^2(K)}^2.$$

We make another assumption on $\mathcal{T}_2$.

Assumption 2.81. *Each element $K \in \tilde{\mathcal{T}}_1$ is the union of elements in $\mathcal{T}_2$ of which all have an element diameter of at most $h_K/2$.*

This assumption again can be ascertained by applying one step of isotropic refinement to every element in $\tilde{\mathcal{T}}_1$. Now, we split the mesh $\mathcal{T}_2$ into two disjoint subsets $\mathcal{T}_{2,R}$ and $\mathcal{T}_{2,U}$ with $\cup_{K\in\mathcal{T}_{2,R}} K = \cup_{K\in\tilde{\mathcal{T}}_1} K$. So, there holds

$$\begin{aligned}\sum_{K\in\mathcal{T}_{2,U}} h_K^2\|f - f_K\|_{L^2(K)}^2 &\leq \sum_{K\in\mathcal{T}_1\setminus\tilde{\mathcal{T}}_1} h_K^2\|f - f_K\|_{L^2(K)}^2 \\ &= \sum_{K\in\mathcal{T}_1} h_K^2\|f - f_K\|_{L^2(K)}^2 - \sum_{K\in\tilde{\mathcal{T}}_1} h_K^2\|f - f_K\|_{L^2(K)}^2\end{aligned}$$

as well as

$$\sum_{K\in\mathcal{T}_{2,R}} h_K^2\|f - f_K\|_{L^2(K)}^2 \leq \frac{1}{4}\sum_{K\in\tilde{\mathcal{T}}_1} h_K^2\|f - f_K\|_{L^2(K)}^2.$$

This leads to

$$\begin{aligned}\sum_{K\in\mathcal{T}_2} h_K^2\|f - f_K\|_{L^2(K)}^2 &= \sum_{K\in\mathcal{T}_{2,R}} h_K^2\|f - f_K\|_{L^2(K)}^2 + \sum_{K\in\mathcal{T}_{2,U}} h_K^2\|f - f_K\|_{L^2(K)}^2 \\ &\leq \sum_{K\in\mathcal{T}_1} h_K^2\|f - f_K\|_{L^2(K)}^2 - \frac{3}{4}\sum_{K\in\tilde{\mathcal{T}}_1} h_K^2\|f - f_K\|_{L^2(K)}^2 \\ &\leq \left(1 - \vartheta\frac{3}{4}\right)\sum_{K\in\mathcal{T}_1} h_K^2\|f - f_K\|_{L^2(K)}^2.\end{aligned}$$

Given any fixed error tolerance ε the adaptive algorithm finds a mesh $\mathcal{T}$ such that $\|\partial_x\partial_y(u-\tilde{u})\|_{L^2(\mathcal{D})} \leq \varepsilon/2$ after finitely many steps of the algorithm. This partition is used as starting point for the adaptive algorithm for the term $\|\partial_x\partial_y(\tilde{u}-\tilde{u}_{\mathcal{T}})\|_{L^2(\mathcal{D})}$, which also after finitely many steps will provide a mesh $\mathcal{T}'$ that guarantees $\|\partial_x\partial_y(\tilde{u}-\tilde{u}_{\mathcal{T}'})\|_{L^2(\mathcal{D})} \leq \varepsilon/2$. Therefore we find $\|\partial_x\partial_y(u-u_{\mathcal{T}'})\|_{L^2(\mathcal{D})} \leq \varepsilon$.

2.5 Numerical experiments for the deterministic moment equations

The numerical experiments have been implemented in MATLAB and have been run on a laptop with an Intel i7-4720HQ with 16GB of RAM.

2.5.1 Experiments for the 1D model problem

A smooth right-hand side We start our presentation of the numerical experiments by showing results for each of the proposed error estimators with a smooth right-hand side, which is given by

$$f = \frac{\pi^4}{16} \cos\left(\frac{\pi}{2}x\right) \cos\left(\frac{\pi}{2}y\right).$$

The tests have been performed with $\vartheta = 0.5$ and a stopping criterion of 200'000 elements. The meshes shown are picked from a selected iteration to ensure comparability of around 10'000 elements. We begin with the residual error estimator. Since for $p = 1$ the error estimator $\eta_{\mathcal{R}}$ does not relate to the numerical solution at all, we enrich the Finite Element space by only a bubble function and then have a look at using the full set of shape functions with $p = 2$ for the residual error estimator. The corresponding error curves can be seen in Figure 2.9.

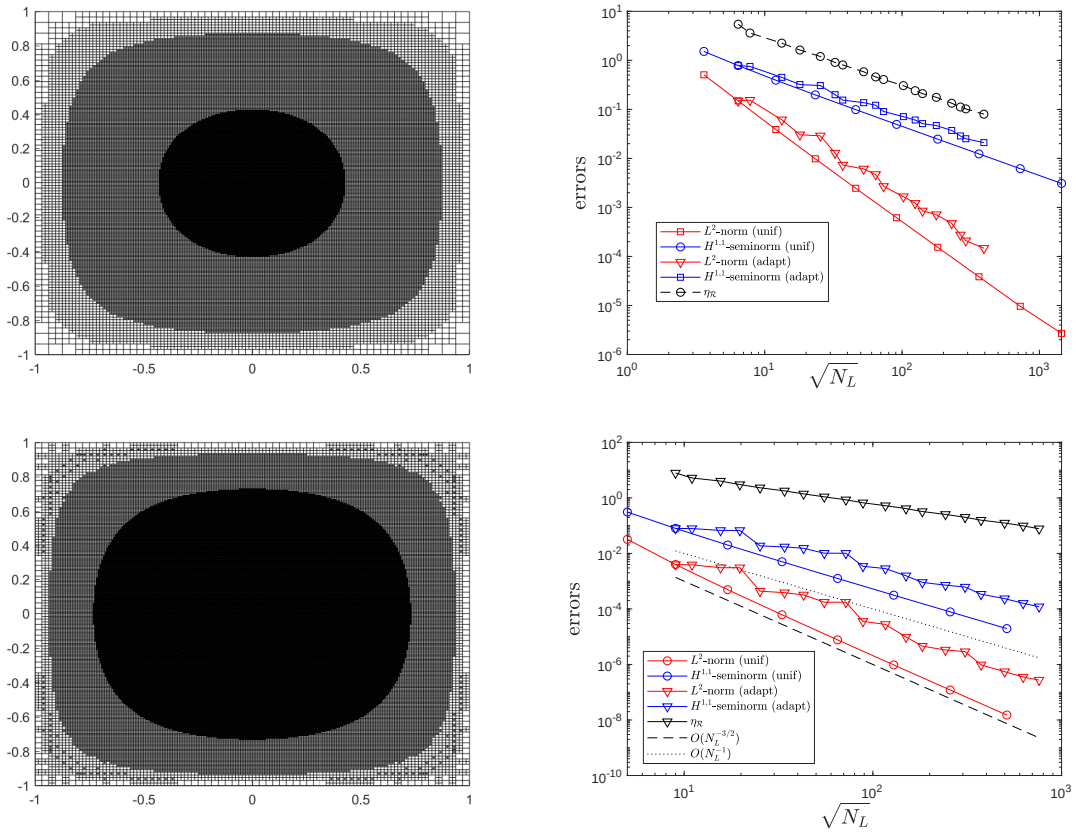

Figure 2.9: Numerical results for $\eta_{\mathcal{R}}$ with bubble function (top) and $\eta_{\mathcal{R}}$ with the full set of shape functions $p = 2$ (bottom).

We see that the meshes are qualitatively similar for both of the variants. We observe in both experiments that the error estimator $\eta_{\mathcal{R}}$, converges in the $H^{1,1}$-seminorm at rate $\frac{1}{2}$ for the FE space enriched with a bubble and at rate 1 for the full set of shape functions with $p = 2$. As the solution does not feature any abrupt changes the adaptive algorithm has a hard time against the uniform refinement, which is clearly seen from the convergence graphs. It can further be seen that the estimator clearly overestimates the error in both instances.

We now turn our attention to the numerical results for the variants of the hierarchical error estimator η_H. Here we take a look at the hierarchical error estimator $\tilde{\eta}_H$, which uses the full set of bubble functions, and $\hat{\eta}_H$, which only uses the element bubble functions. Meshes and corresponding error curves can be found in Figure 2.10. We observe that the meshes are qualitatively very similar and that the error estimators $\tilde{\eta}_H$ and $\hat{\eta}_H$ both converge at a rate of $\frac{1}{2}$ as the error in the $H^{1,1}$-seminorm. Futhermore, it can be seen from the convergence graphs that the constant in the reliability estimate behaves more favorably for $\tilde{\eta}_H$, which is on the one hand due to the fact that information for edge jumps is incorporated in $\tilde{\eta}_H$. On the other hand we expect $\hat{\eta}_H$ to be worse, since we are omitting the information of the edge bubble functions of the error estimator, and we pay for this by a larger constant in the reliability estimate for $\hat{\eta}_H$. Moreover, we see that $\hat{\eta}_H$ significantly underestimates the true error as well as does $\tilde{\eta}_H$, albeit by not that much.

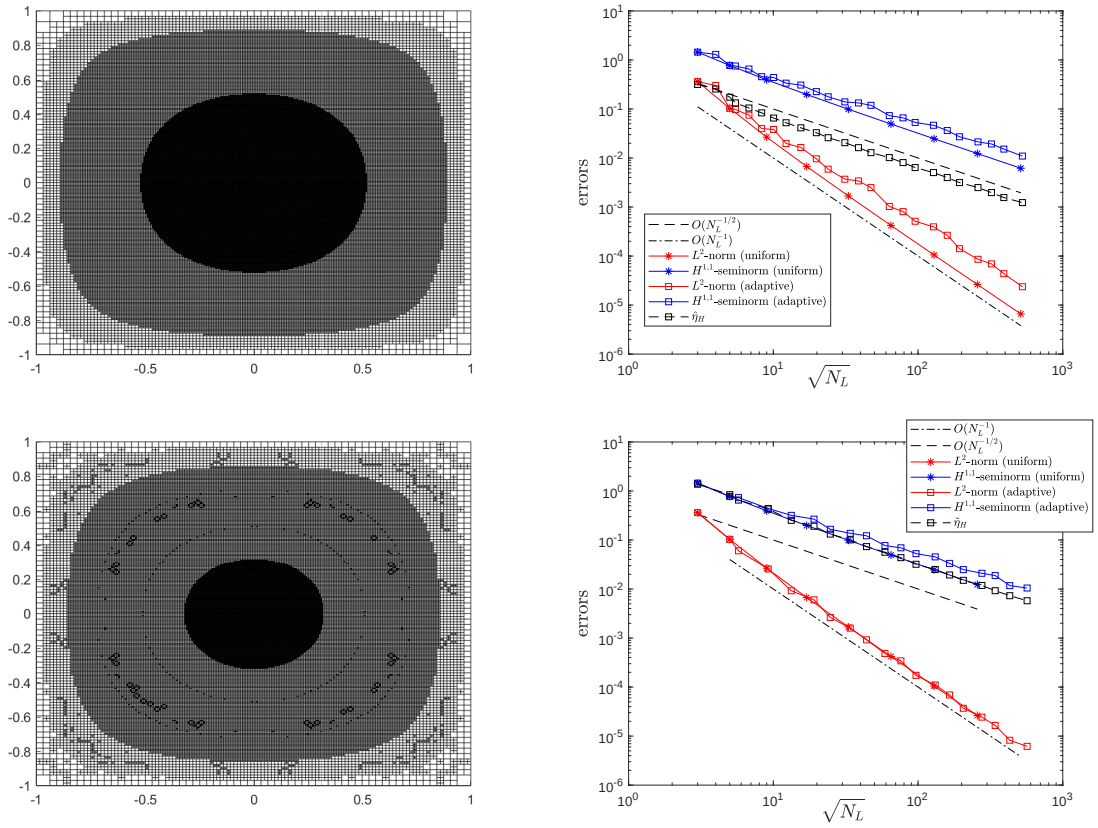

Figure 2.10: Numerical results for $\hat{\eta}_H$ (top) and $\tilde{\eta}_H$ (bottom).

At last we have a look at the numerical results for the a posteriori error estimator based on averaging. Here we do not have to resort to an enrichment of the Finite Element space and present the results for $p = 1$. It is clearly seen from the convergence graph in Figure 2.11 that the error estimator $\eta_{\mathcal{A}}$ converges at a rate of $\frac{1}{2}$ as the error in the $H^{1,1}$-seminorm.

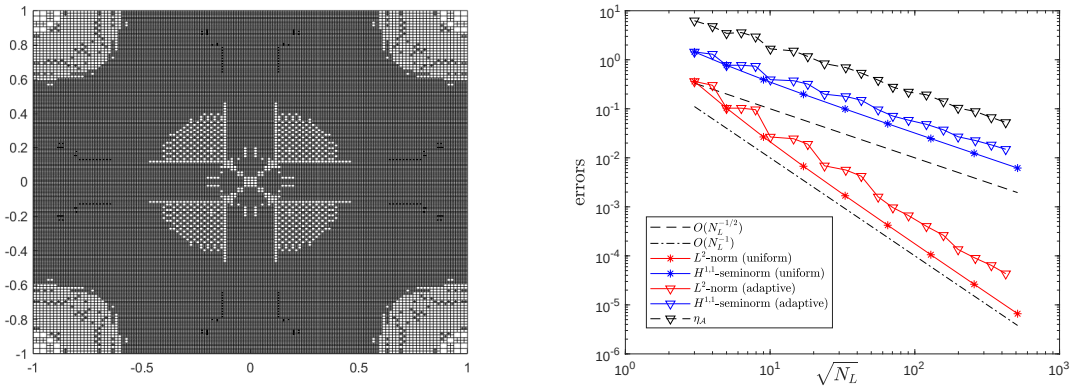

Figure 2.11: Numerical results for $\eta_{\mathcal{A}}$.

Analytic solution with large gradients We now have a look at the adaptive process with exact solution chosen as

$$u(x,y) = \exp\left(-\frac{|x-y|^2}{\lambda}\right)(1-x^2)(1-y^2),$$

with $\lambda = \frac{1}{100}$. Albeit rather artificial, u has been chosen in a way to illustrate the advantages of the adaptive methods in a situation like this. The tests have been performed with $\vartheta = 0.5$ with a stopping criterion of 200'000 elements.

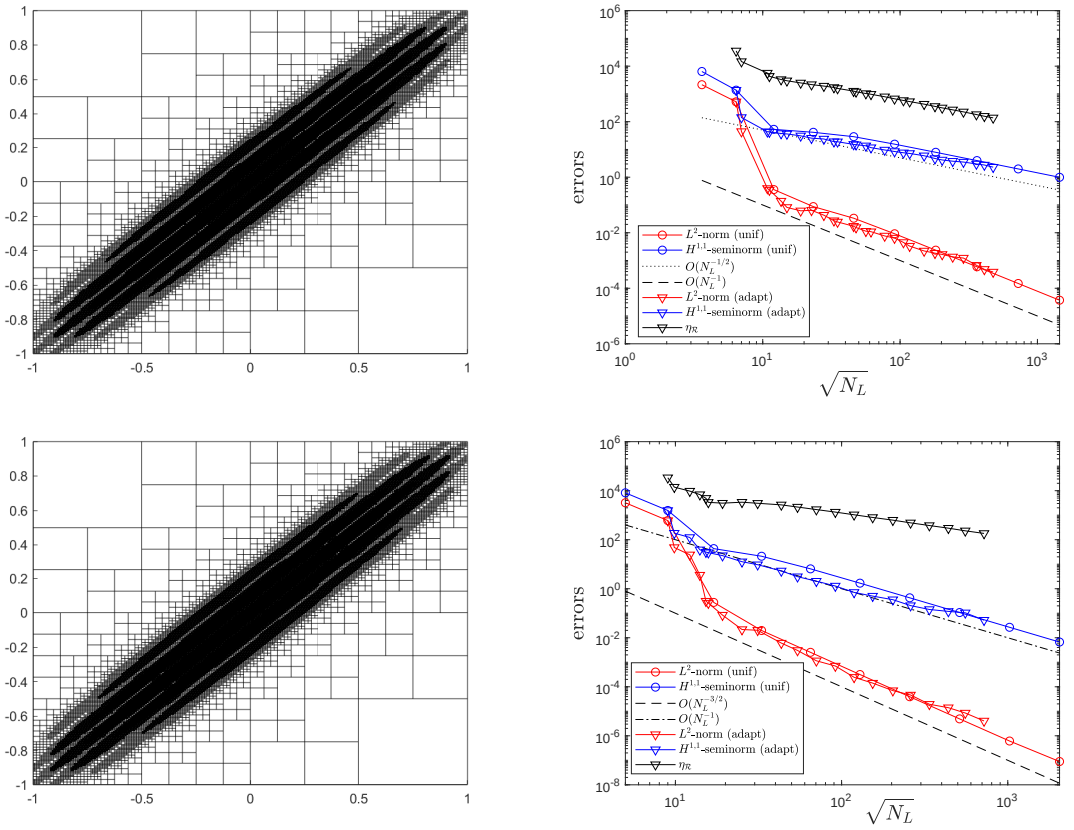

Figure 2.12: Numerical results for $\eta_{\mathcal{R}}$ with a bubble per element (top row) and $\eta_{\mathcal{R}}$ for $p = 2$ (bottom row) .

This solution features large gradients in the vicinity of the diagonal which is a typical feature of covariance functions. We begin the discussion with the numerical results for the residual error estimators and the corresponding error curves in Figure 2.12. We again observe that error estimator $\eta_{\mathcal{R}}$ converges at rate $\frac{1}{2}$ and 1 in correspondence with the error in the $H^{1,1}$-seminorm, respectively. Moreover, we observe that the adaptive method is able to outperform the uniform FEM in the sense that it needs a significant amount of degrees of freedom less to achieve a comparable accuracy.

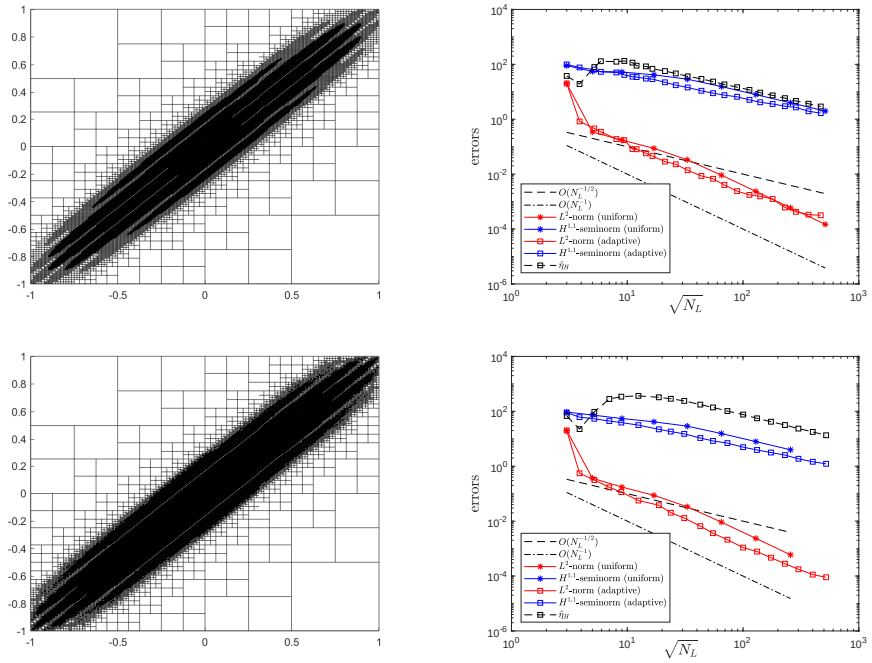

Figure 2.13: Numerical results for $\hat{\eta}_H$(top) and $\tilde{\eta}_H$ (bottom).

Let us now turn our attention to the numerical results for the hierarchical error estimators $\tilde{\eta}_H$ and $\hat{\eta}_H$. Corresponding meshes and error curves can be found in Figure 2.13. We observe that the adaptive method is again able to beat the uniform FEM and that the errors converge at rate $\frac{1}{2}$ in the $H^{1,1}$-seminorm which is expected from the theory. The produced meshes are very similar and we see that using $\tilde{\eta}_H$ is not necessarily more advantageous than the error estimator $\hat{\eta}_H$, although $\hat{\eta}_H$ incorporates less information than $\tilde{\eta}_H$.

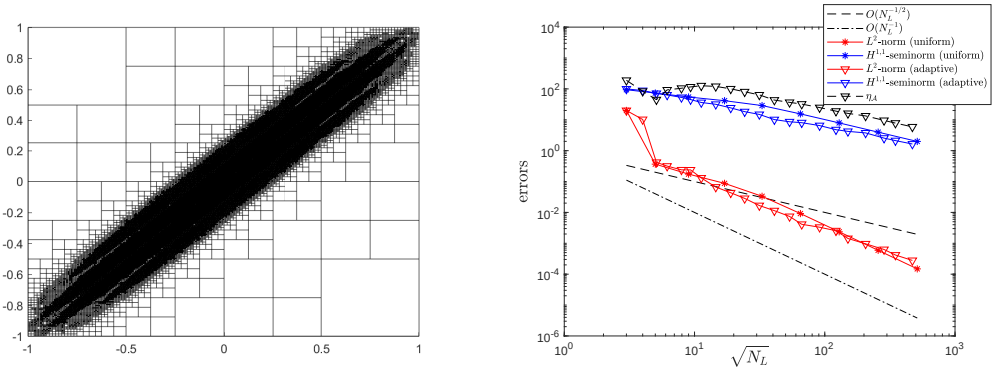

Figure 2.14: Numerical results for $\eta_{\mathcal{A}}$.

At last we have a look at the numerical results for the a posteriori error estimator based on averaging $\eta_{\mathcal{A}}$ for this problem. We observe that $\eta_{\mathcal{A}}$ is competetive against the uniform FEM for this problem. Moreover, as can be seen from the convergence graph the error estimator $\eta_{\mathcal{A}}$ slightly overestimates the exact error.

Exponential covariance As another experiment we have chosen the right-hand side

$$f(x,y) = \exp\left(-\frac{|x-y|}{\rho}\right)$$

with $\rho = \frac{1}{10}$. This covariance function is a representant of the family of so-called Matérn covariance functions of the form

$$M_\nu(x,y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\rho} d\right)^\nu K_\nu \left(\frac{\sqrt{2\nu}}{\rho} d\right),$$

for $\nu = 1/2$ and $\sigma = 1$. σ denotes the variance and ν is referred to as the smoothness parameter of the family, whereas $d = |x-y|$.

In order to provide convergence graphs we have computed the analytical solution u (cf. Section 4.1) and then computed the $H^{1,1}$-seminorm by numerically integrating $(\partial_x \partial_y u)^2$ with $u = \mathcal{C}_u$ from (4.3) for $5 \cdot 10^4$ terms with a high order Gauß-Legendre quadrature rule. This yielded

$$|u|^2_{H^{1,1}(\mathcal{D})} \approx 0.006711721115088.$$

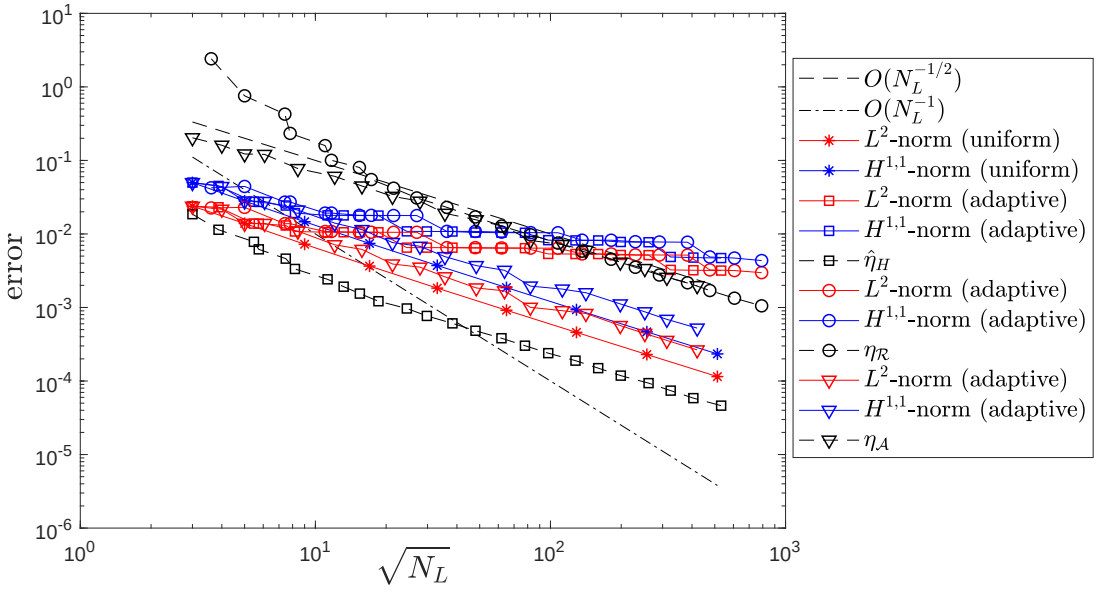

Figure 2.15: Convergence comparison for the exponential covariance.

In Figure 2.15 we have depicted the convergence history for the three types of a posteriori error estimators $\eta_{\mathcal{R}}, \hat{\eta}_H$ and $\eta_{\mathcal{A}}$ as well as the corresponding convergence graph of the uniform FEM for comparison. First and foremost, we observe that the residual and hierarchical error estimators hit a proverbial wall in terms of convergence due to the effect of the weak efficiency. They are still converging as can be seen from the graph, but at a seriously reduced rate. As the averaging estimator $\eta_{\mathcal{A}}$ is asymptotically exact it can

be seen to not having to suffer from this deficiency and converges at the expected rate. Moreover, we can infer from the graphs that the solution $\mathcal{C}_u$ is rather tame and thus not even the averaging estimator can beat the uniform approximation.

Sinusoidal covariance We have chosen the right-hand side

$$f(x,y) = \frac{\rho}{|x-y|} \sin\left(\frac{|x-y|}{\rho}\right)$$

with $\rho = \frac{1}{10}$. In order to provide convergence graphs we have computed $H^{1,1}$-seminorm via a substitute reference solution u_{ref} which was chosen in this case as the solution of the uniform FEM on level 10.

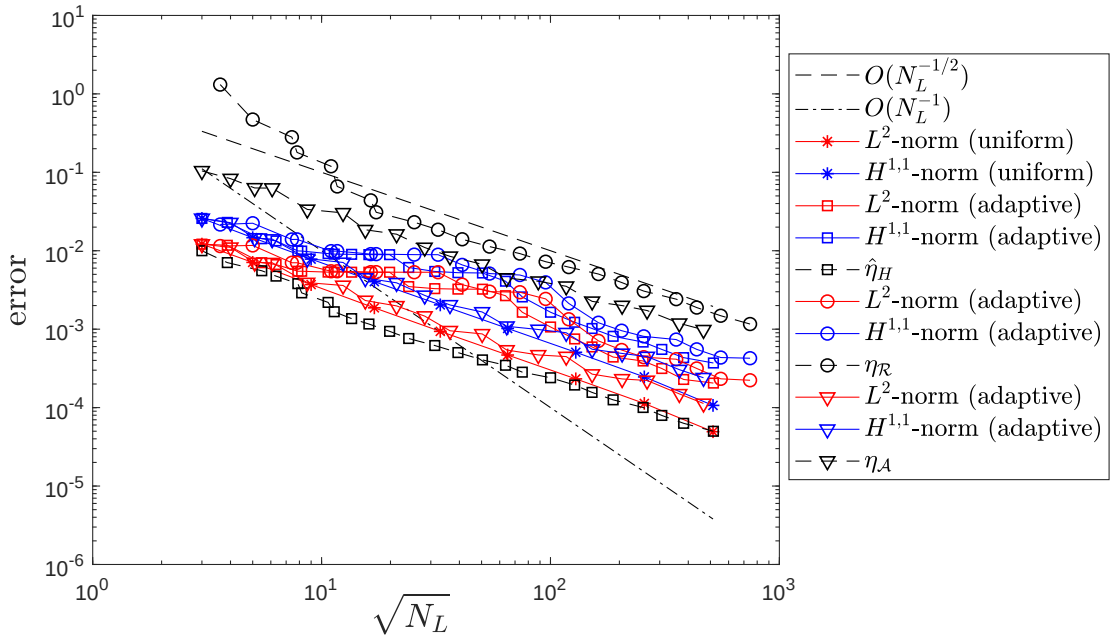

Figure 2.16: Convergence comparison for the sinusoidal covariance.

For this experiment we again noticeably experience the effect of the weak efficiency of $\eta_{\mathcal{R}}$ and $\hat{\eta}_H$ (cf. Figure 2.16). Here however the shelf is not as pronounced as for the exponential covariance and it seems that the error estimators start to recover at a threshold of roughly 10^5 degrees of freedom and converge at the expected and desired rate. For this experiment it can also be observed that $\eta_{\mathcal{A}}$ behaves quite competetively against the uniform FEM, but still loses out as the solution is again very tame.

2.5.2 Experiments for the 2D model problem

Smooth solution We start out again with a validation experiment where we take a look at the adaptive algorithm for the tensorizable right-hand side

$$f(x,y,z,w) = \frac{\pi^4}{4}\cos\left(\frac{\pi}{2}x\right)\cos\left(\frac{\pi}{2}y\right)\cos\left(\frac{\pi}{2}z\right)\cos\left(\frac{\pi}{2}w\right).$$

As the solution is a product of cosine functions with zero boundary conditions and does not feature any significant abrupt changes we expect the adaptive process not to outperform the uniform FEM in this case, but rather still be competetive. One has to keep in mind that in four space dimensions going from one uniform refinement level to the next is effectively mulitplying the number of elements by 16. Thus, even though the adaptive procedure might not be able to outperform the uniform FEM, it enables the possibility of aborting the computation earlier for a reasonable approximation when the next step of uniform refinement would be too computationally expensive, *days* say.

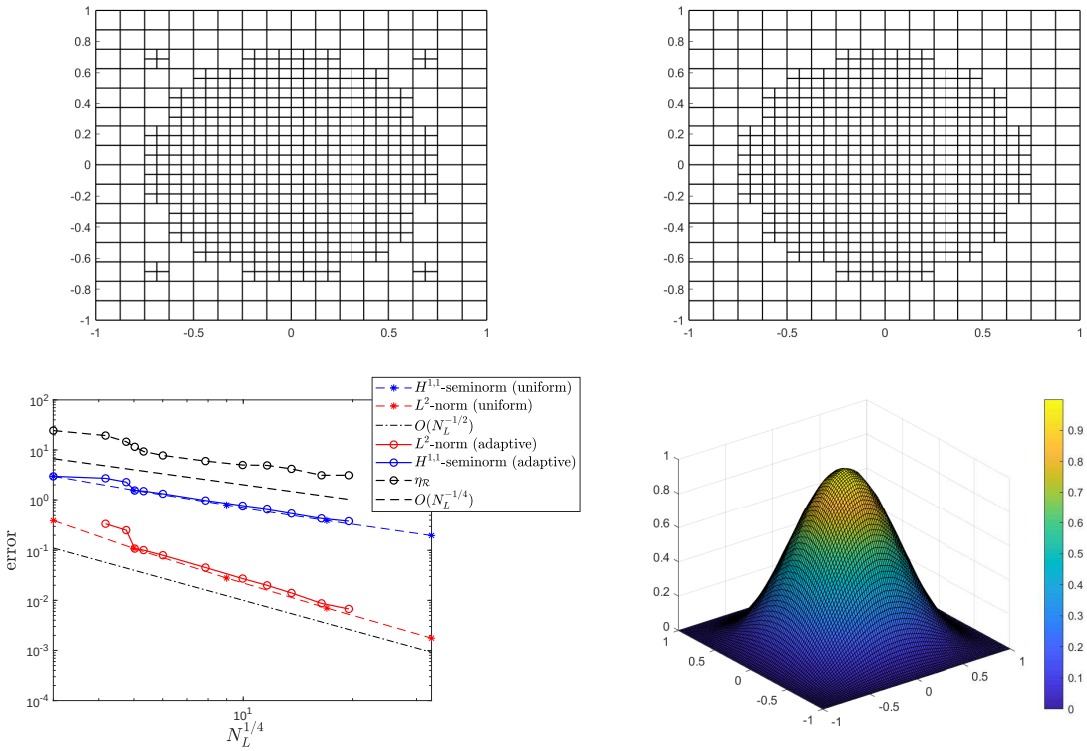

Figure 2.17: Top row: Mesh slices for the numerical solution at iteration 12 of the adaptive process. From left to right the slices are as follows: xy-slice (left) is identical to the yw-slice, and the xw-slice (right) is identical the yz-slice,xw-slice and zw-slice. Bottom row: Convergence history (left) for the uniform FEM for 5 levels and the adaptive method with $\eta_{\mathcal{R}}$ for 12 levels. Diagonal slice (right)

Moreover, depicted mesh slices are generated as follows. For example, consider the shown xw-mesh slice of the xw-plane of $\mathcal{T}$ in Figure 2.17 (top right). By setting $z = 0$ and $y = 0$ and sieving through the nodes of the mesh we obtain the depicted slice. For the other slices we procede analogously. We have opted to show only relevant slices and omit otherwise superfluous repetition. Diagonal slices are evaluated on a uniformly distributed grid of 10^4 points in $[-1,1]^2 = \{(x,y) = (z,w)\}$. Furthermore, all numerical

experiments shown have been run with $\vartheta = 0.5$ and a maximal element number limit of 100'000 elements.

In Figure 2.17 the reader may find selected mesh slices of the four dimensional mesh $\mathcal{T}$, a convergence history for the residual error estimator $\eta_{\mathcal{R}}$, which does not need to be amended with a bubble function in this case, and the diagonal slice of the finest refinement level, which is an approximation of the variance. Note that the diagonal is supposed to be an approximation to $\cos^2(\pi x/2)\cos^2(\pi y/2)$. The shown depiction is in close agreement with this albeit the mesh size is still very large considering we are in four space dimensions. Further, we see from the convergence graph that the optimal rate of convergence $\frac{1}{4}$ is attained and in particular that it is quite competetive against the uniform FEM. Note also that the estimator significantly overestimates the error.

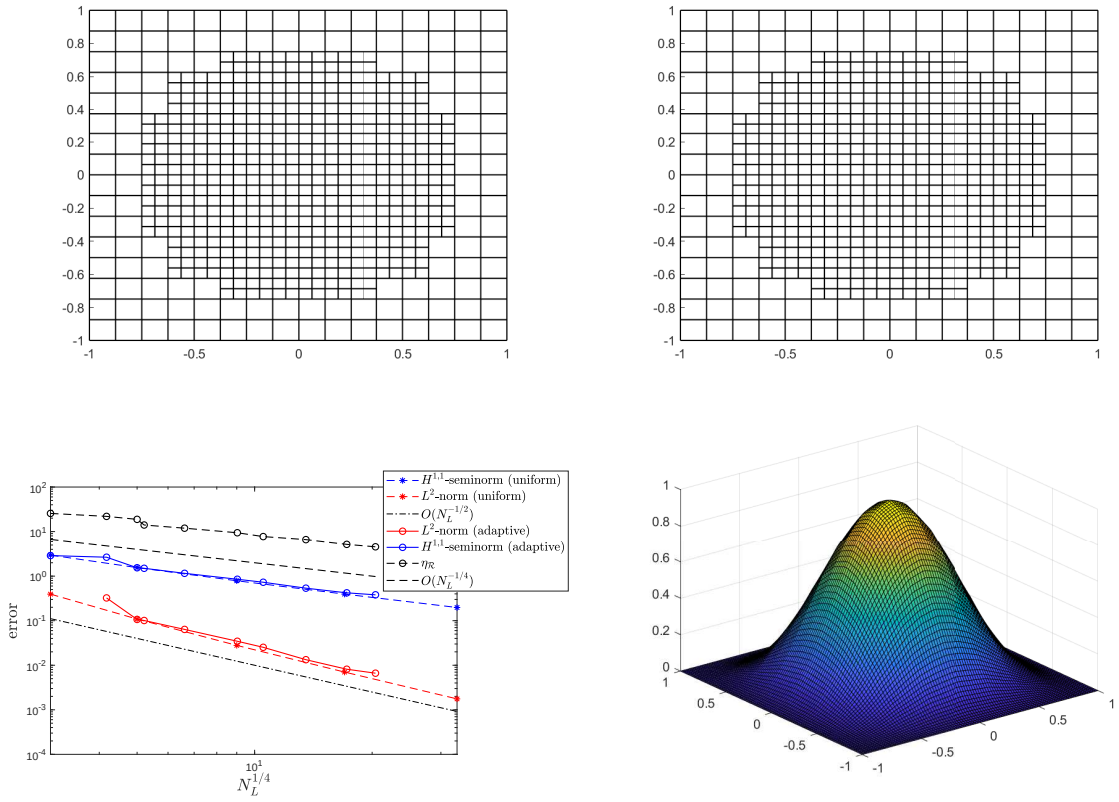

Figure 2.18: Top row: Mesh slices for the numerical solution at iteration 10 of the adaptive process. All slices are identical in this instance. Bottom row: Convergence history (left) for the uniform FEM for 5 levels and the adaptive method with $\eta_{\mathcal{R}}$ with element bubble function for 10 levels.

In Figure 2.18 we have now depicted selected mesh slices, a convergence history and the diagonal slice for the residual error estimator $\eta_{\mathcal{R}}$ with added element bubble function. This test was conducted in order to have the full residual error estimator contribute to the estimation. The meshes and the convergence are almost identical to the estimator without added bubble function and thus, the additional work does not seem to be justified. From the convergence graph we find however that the optimal rate of convergence is attained and is nonetheless quite competetive against the uniform FEM. Furthermore, this variant also quite overestimates the true error.

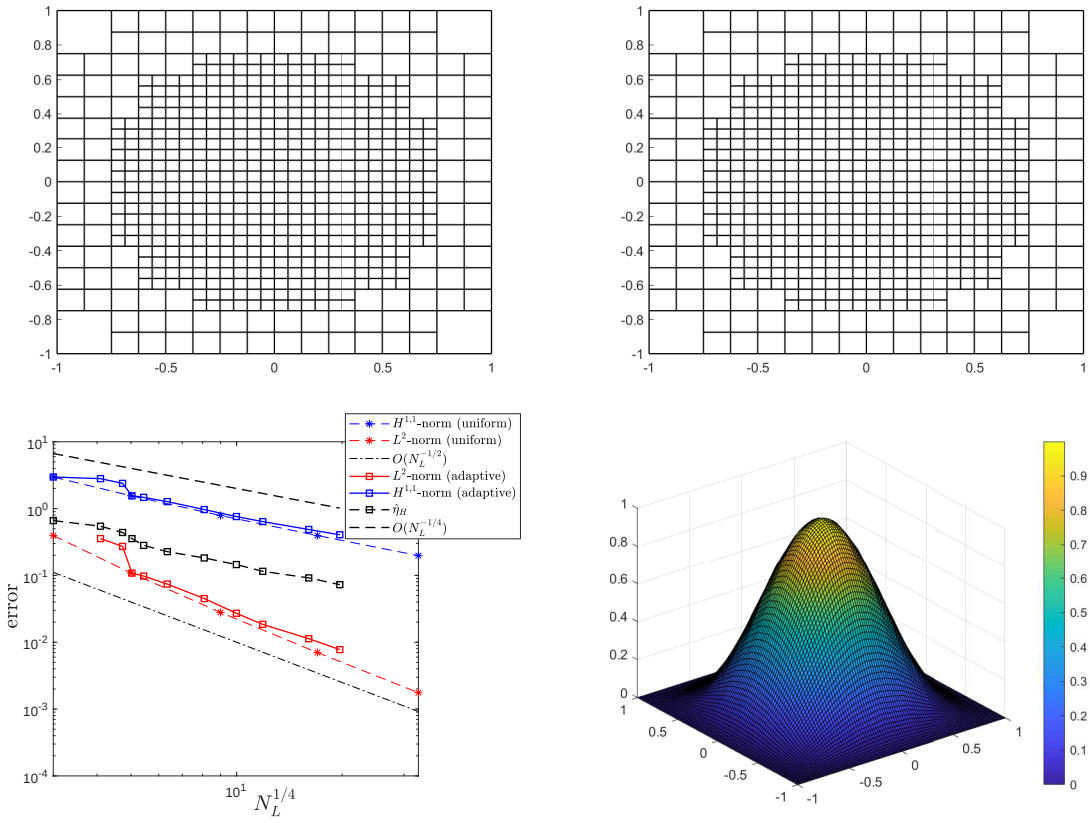

Figure 2.19: Top row: Mesh slices for the numerical solution at iteration 11. xy-slice (left), yw-slice(right). Bottom row: Convergence graph for $\hat{\eta}_H$(left) and approximate variance (right).

For the hierarchical error estimator we have depicted in Figure 2.19 another display of selected mesh slices, a convergence history and the diagonal slice. From the convergence graph we infer again that the optimal rate of convergence is attained and is also quite competetive against the uniform FEM.

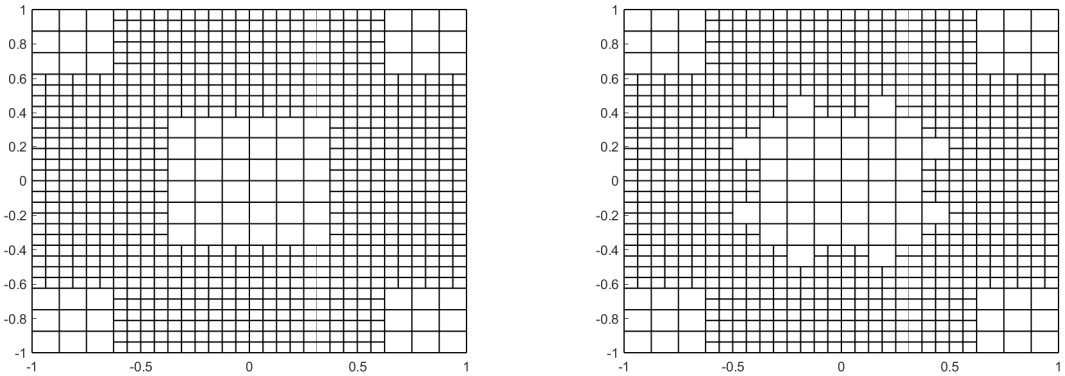

Figure 2.20: Top row: Mesh slices for the numerical solution at iteration 8. xy-slice (left), xz-slice(right). Bottom row: Convergence graph for $\eta_{\mathcal{A}}$ (left) and approximate variance on the diagonal slice (right)

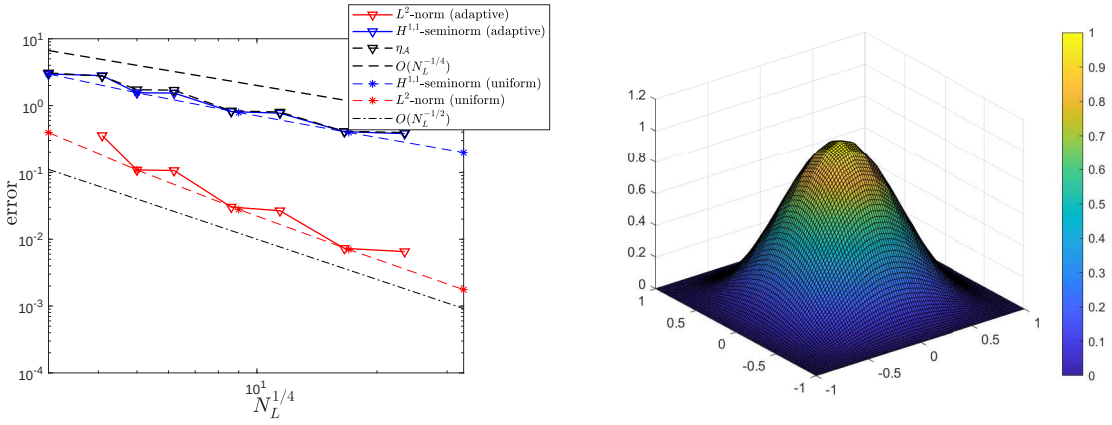

Figure 2.21: Top row: Mesh slices for the numerical solution at iteration 8. xy-slice (left), xz-slice(right). Bottom row: Convergence graph for $\eta_{\mathcal{A}}$ (left) and approximate variance on the diagonal slice (right)

Furthermore, it can be seen that $\hat{\eta}_H$ underestimates the true error by a significant amount which was already observed for the two dimensional counterpart, since we are only incorporating information via the element bubble and omit the rest.

At last we have a look at the averaging a posteriori error estimator $\eta_{\mathcal{A}}$. In Figure 2.20 we show plots of mesh slices and in Figure 2.21 a convergence history and the approximate variance is shown which is in good agreement with the exact solution. From the convergence graph we can see that $\eta_{\mathcal{A}}$ is in fact very competetive against the uniform FEM in the four dimensional setting.

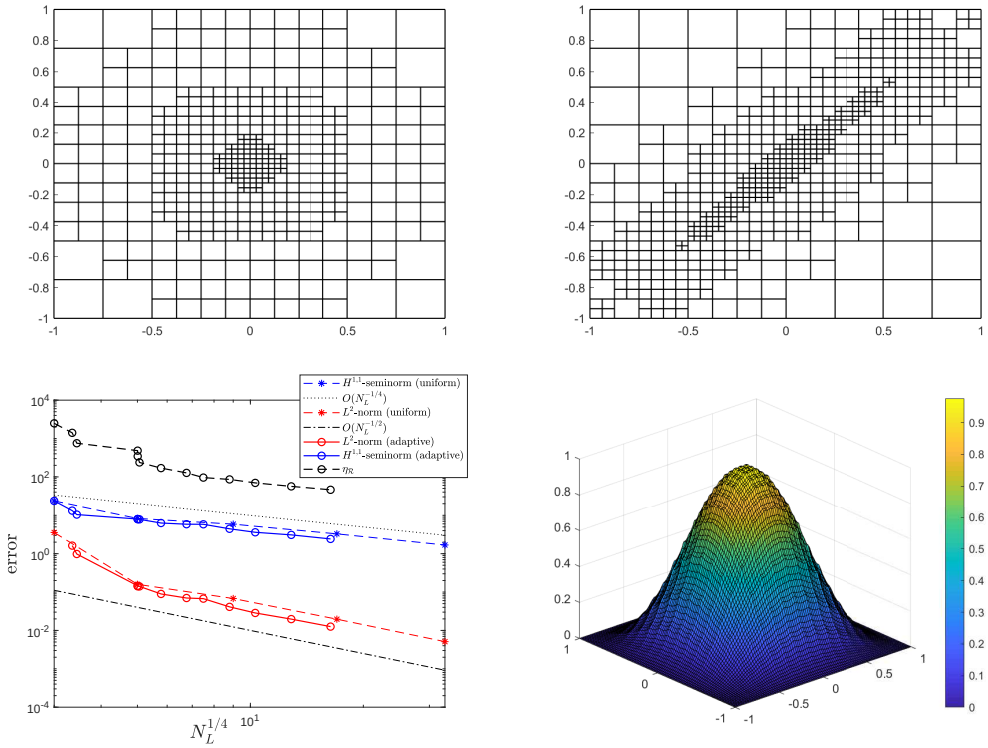

Figure 2.22: Mesh slices for the numerical solution at iteration 13 of the adaptive process. xy-slice(left), xz-slice(right). Convergence history for $\eta_{\mathcal{R}}$ and diagonal slice.

Solution with large gradients Here we have a look at the adaptive process for the exact solution

$$u(x,y,z,w) = \exp\left(-\frac{(x-z)^2+(y-w)^2}{\lambda}\right)(1-x^2)(1-y^2)(1-z^2)(1-w^2),$$

where $\lambda = \frac{1}{10}$. Here the right-hand side was computed as $f = (\Delta_x \otimes \Delta_y)u$ using Wolfram Mathematica. As the resulting terms are rather bulky we omit the explicit representation of f. Since the right-hand side exhibits steep gradients in a neighborhood of the diagonal of $[-1,1]^4$ we can see that all of the derived a posteriori error indicators are able to outperform uniform FEM as the solution is very smooth, too. Corresponding mesh slices, convergence histories and diagonal slices for $\eta_{\mathcal{R}}$ with and without element bubble,$\hat{\eta}_H$ and $\eta_{\mathcal{A}}$ can be found in Figures 2.22, 2.23, 2.24 and 2.25, respectively. Moreover, it is not sure if the residual or hierarchical error estimators might hit a convergence shelf at some point.

As a last numerical experiment for the four dimensional adaptive process we show the numerical results for tensorized exponential covariance function to demonstrate the different behaviour of the estimators in the next paragraph. The results are in line with the discussion in the two dimensional situation and are therefore left to the reader to observe.

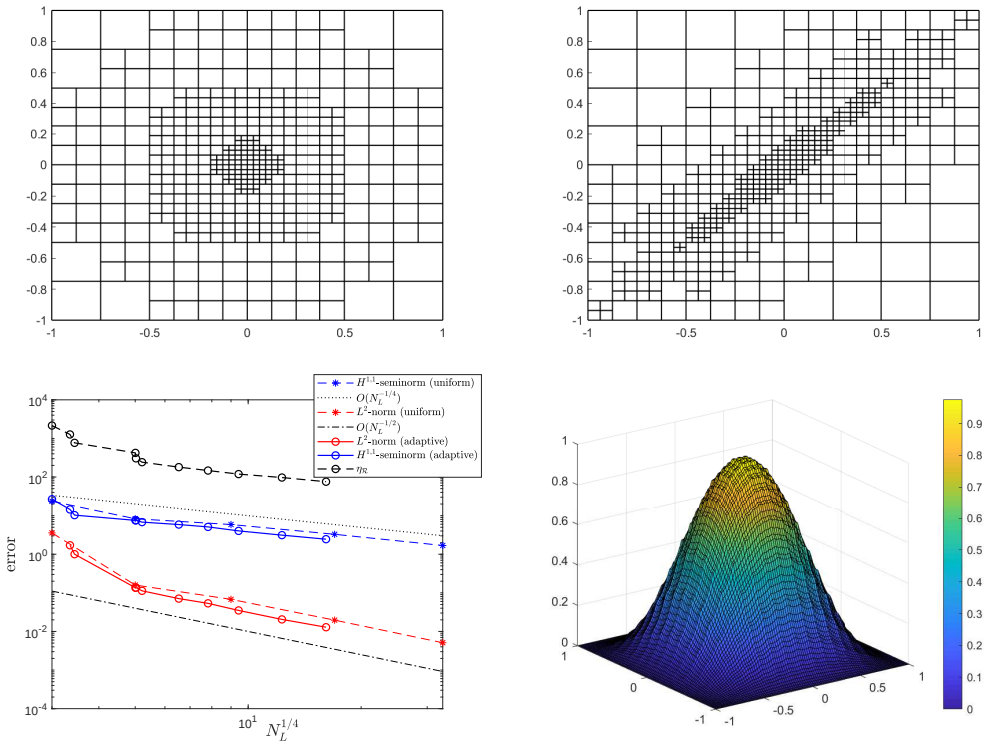

Figure 2.23: Mesh slices for the numerical solution at iteration 11 of the adaptive process. From left to right the slices are as follows: xy-slice, xz-slice. Convergence history for $\eta_{\mathcal{R}}$ with added element bubble and corresponding diagonal slice.

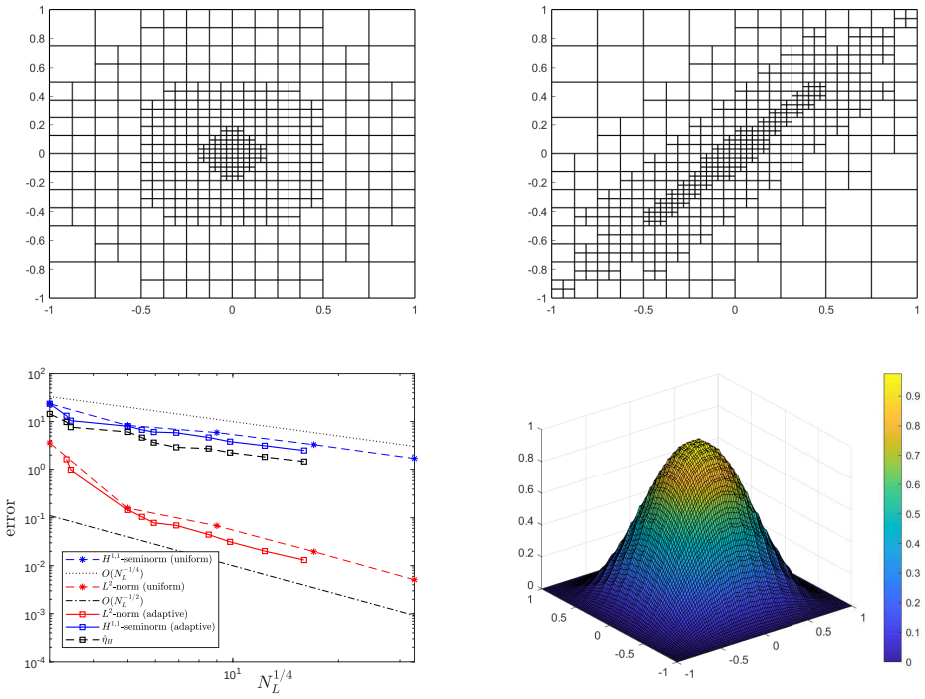

Figure 2.24: Mesh slices for the numerical solution at iteration 11 of the adaptive process. From left to right from top to bottom the slices are as follows: xy-slice, xz-slice. Convergence history for $\hat{\eta}_H$ and corresponding diagonal slice

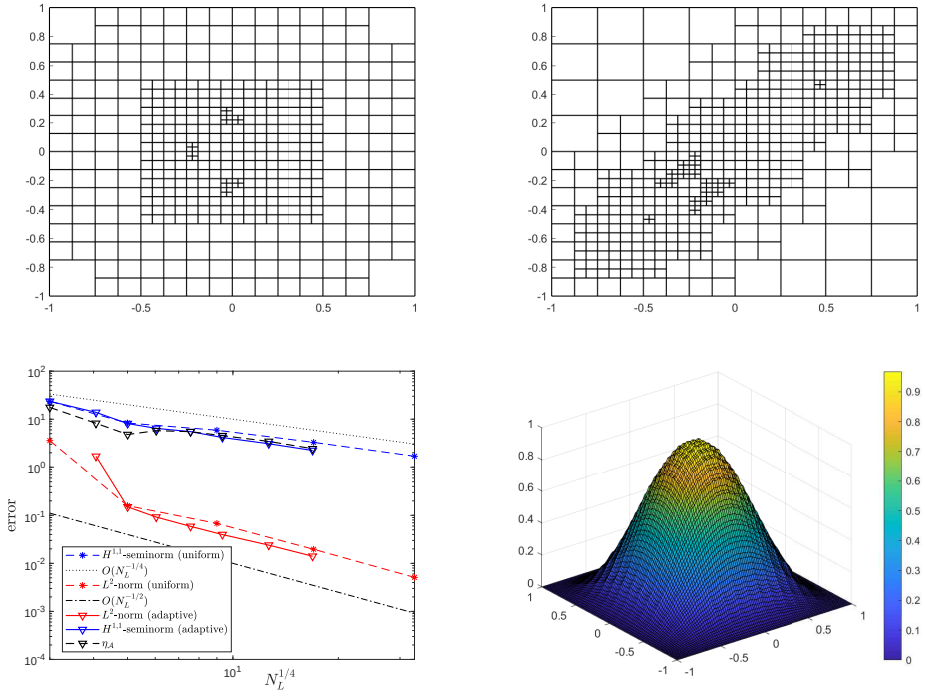

Figure 2.25: Mesh slices for the numerical solution at iteration 8 of the adaptive process. From left to right from top to bottom the slices are as follows: xy-slice, xz-slice. Convergence history for $\eta_{\mathcal{A}}$ and corresponding diagonal slice.

Product of exponential covariances Here we have a look at the adaptive process for the right-hand side

$$f(x,y,z,w) = \exp\left(-\left(\frac{|x-z|}{\rho_1} + \frac{|y-w|}{\rho_2}\right)\right)$$

with $\lambda_1 = \frac{1}{10}$ and $\lambda_2 = \frac{1}{2}$.

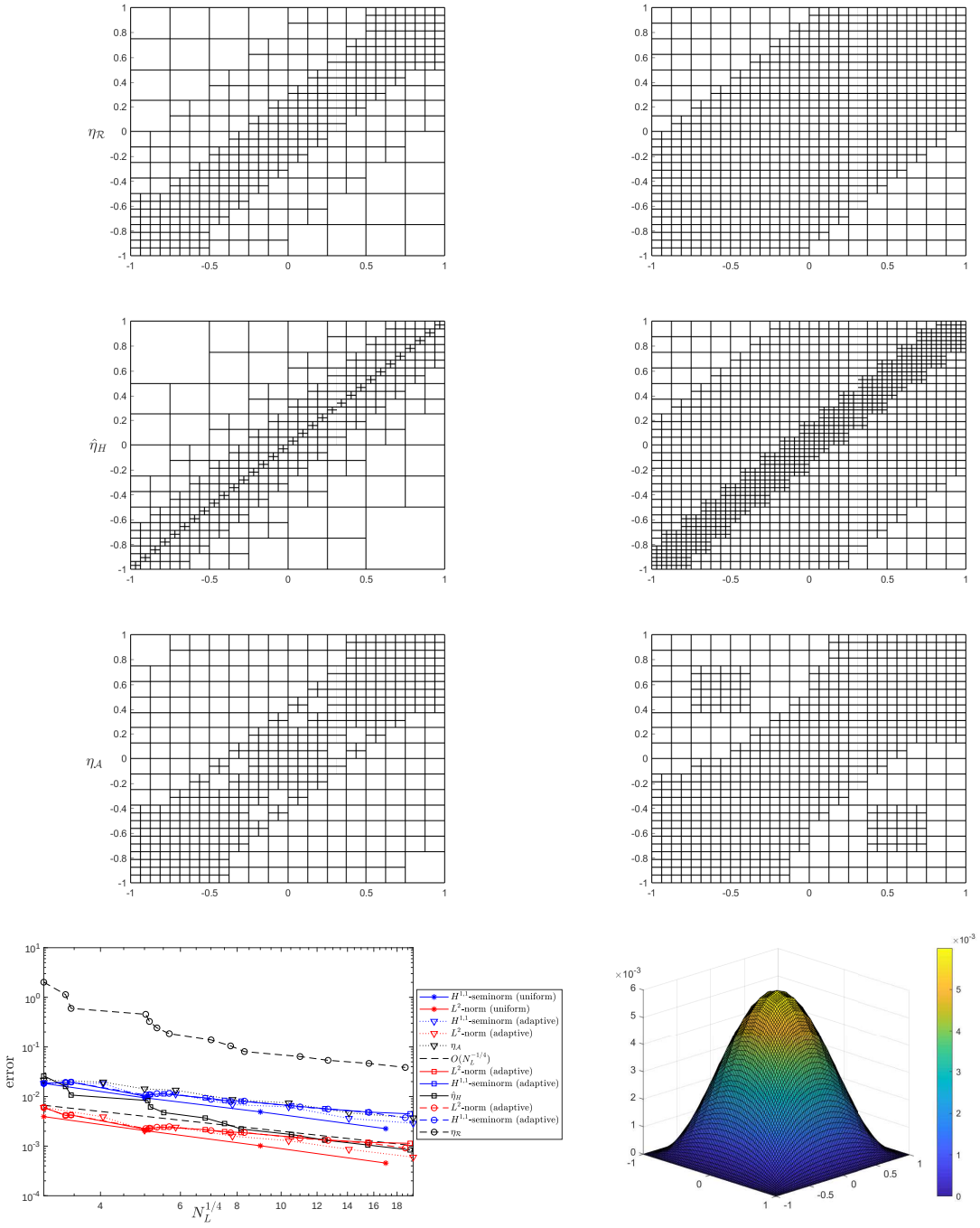

Figure 2.26: Mesh slices for the numerical solution at the last iteration of the adaptive process with $\eta_{\mathcal{R}}, \hat{\eta}_H$ and $\eta_{\mathcal{A}}$. xy-slices(left), xz-slices(right). Convergence history of uniform and adaptive FEM and the diagonal slice of the last iteration the adaptive process with $\eta_{\mathcal{A}}$.

Chapter 3

Monte Carlo Methods for the Approximation of Covariance Functions

3.1 Problem setting

Let us recall that the domain $D \subset \mathbb{R}^d, d = 1, 2$ is an open bounded polyhedral domain with Lipschitz boundary, that we set for abbreviation $\mathcal{D} = D \times D$, and that by $(\Omega, \Sigma, \mathbb{P})$ we denote a complete probability space with associated σ-algebra Σ and probability measure $\mathbb{P}$. We again consider for *all* $\omega \in \Omega$ the stochastic elliptic boundary value problem

$$\begin{cases} -\nabla \cdot (\kappa(x,\omega)\nabla u(x,\omega)) &= f(x,\omega), \quad \text{in } D, \\ u(x,\omega) &= 0, \qquad\quad \text{on } \partial D, \end{cases} \tag{3.1}$$

where $f(x,\omega)$ is a given stochastic source term and $\kappa(x,\omega)$ is the stochastic diffusion coefficent.

Again we are interested in solving for the second moment of u. More precisely, we want to solve the associated second moment problem (cf. also (2.3)):

Problem 3.1. *Find $u \in L^p(\Omega; H)$, such that for all $v \in H$ there holds*

$$\begin{aligned} \mathbb{E}\left[\int_{\mathcal{D}} (\kappa(x,\omega) \otimes \kappa(y,\omega))(\nabla_x \otimes \nabla_y)(u(x,\omega) \otimes u(y,\omega))(\nabla_x \otimes \nabla_y)v(x,y)\,\mathrm{d}x\,\mathrm{d}y\right] \\ = \mathbb{E}\left[\int_{\mathcal{D}} (f(x,\omega) \otimes f(y,\omega))v(x,y)\,\mathrm{d}x\,\mathrm{d}y\right]. \end{aligned} \tag{3.2}$$

If it was possible to solve the previous problem exactly, we would have access to the exact solution $\mathcal{C}_u(x,y) := \mathbb{E}[u(x,\omega) \otimes u(y,\omega)]$. There are multiple difficulties with this approach. First of all, solving an infinite dimensional problem is usually out of question, except in cases where an explicit solution is available. Another thorn in our sight is that usually we also do not have access to the exact solution $u(x,\omega)$, which is a random field. This means that we will have to be content with certain approximations u_ℓ of u, where $u_\ell \to u$ in some sense is required. Moreover, most often taking the exact expectation is also neither practical nor applicable, for which reason we resort to computing solutions with respect to certain events and try to quantify the error we commit by computing certain sample averages.

More specifically, to find an approximation to the two-point correlation function or the covariance function of u, we compute first a set of solutions with respect to certain events $\omega^i \in \Omega, i = 1, \dots, M$ of the following variational equation

$$\int_{\mathcal{D}} \kappa(x, \omega^i) \nabla_x u(x, \omega^i) \nabla_x v(x) \, \mathrm{d}x = \int_{\mathcal{D}} f(x, \omega^i) v(x) \, \mathrm{d}x.$$

Suppose we have access to the exact solution of the previous variational problem, then we are able to compute the sample average to approximate the expectation of u by

$$\mathrm{E}_M[u](x) := \frac{1}{M} \sum_{i=1}^{M} u(x, \omega^i),$$

the two point correlation function of u by

$$\mathrm{E}_M[u \otimes u](x, y) := \frac{1}{M} \sum_{i=1}^{M} (u(x, \omega^i) \otimes u(y, \omega^i)),$$

as well as the corresponding covariance function using the sample covariance via

$$\mathrm{Cov}_M\,[u, u]\,(x, y) := \frac{1}{M-1} \sum_{i=1}^{M} (u(x, \omega^i) - \mathrm{E}_M[u](x)) \otimes (u(y, \omega^i) - \mathrm{E}_M[u](y)),$$

where now a solution $u(x, \omega^i)$ is called a *sample*.

The two point correlation of a function is closely related to the variance and covariance of u via the equation

$$\begin{aligned} \mathbb{C}\mathrm{ov}\,[u, u]\,(x, y) = \mathbb{C}\mathrm{ov}\,[u]\,(x, y) =& \mathbb{E}[(u(x, \omega) - \mathbb{E}[u](x)) \otimes (u(y, \omega) - \mathbb{E}[u](y))] \\ =& \mathbb{E}[u(x, \omega) \otimes u(y, \omega)] - \mathbb{E}[u](x) \otimes \mathbb{E}[u](y), \\ \mathrm{Var}[u](x) =& \mathbb{C}\mathrm{ov}\,[u, u]\,(x, x). \end{aligned}$$

From this we see that the two point correlation function for a centered random field u coincides with its covariance function. With this in mind we will now set out to approximate covariance functions in a more general context. We want to study under which circumstances we can construct approximations to general covariance functions of the form

$$\mathbb{C}\mathrm{ov}\,[X, Y] := \mathbb{E}[\overline{X} \otimes \overline{Y}] = \mathbb{E}[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])]$$

where X and Y are now considered to be random fields in the Bochner space $L^p(\Omega; H), p \in [1, \infty]$ (cf. section 1.2) with a given Hilbert space H and $\overline{X} := X - \mathbb{E}[X]$. In the following, H shall always denote a separable Hilbert space, unless specified otherwise.

In order to present the theory in a succinct manner, we shall start by explaining and constructing full tensor product approximations of elementary tensor products in the deterministic setting. We then shall additionally derive results for the approximation in the stochastic setting. This will serve as a basis for the more advanced topics to follow, when we discuss the principles of a sparse approximation of covariance functions.

We shall look at convergence in the case of sampling in exact and discrete circumstances and describe the amount of work needed to find approximations that fulfill a certain error tolerance requirement ε.

3.2 Approximation in tensor product spaces

In this section we shall take a look at the approximation of elementary tensor products $X \otimes Y \in H \otimes H$ as well as briefly describe the so-called sparse tensor product approximation in the deterministic setting. The reason for this approach is to keep the presentation as simple as possible and to make it more accessible. Moreover, a clear understanding of the convergence of the approximations is needed for a sensible comparison at the end. These results can then be interpreted in a stochastic context under certain assumptions on the random fields.

3.2.1 Full tensor product approximation and convergence

Since we would like to understand how an economical and accurate approximation of covariance functions of the form $\mathbb{C}\mathrm{ov}\,[X] = \mathbb{E}[(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])]$ can be achieved, we study in this section different approaches of approximation of elementary deterministic tensor products of the form $X \otimes Y$. For this, let H be a Hilbert space and let $X, Y \in H$ be two arbitrary but fixed elements with associated sequences $\mathcal{X} = \{X_\ell\} \subset H$ and $\mathcal{Y} = \{Y_\ell\} \subset H$ converging to X and Y, respectively. Moreover, we consider that the convergence takes place at a given rate $\delta > 0$ with respect to a possibly smaller space $W \subset H$. More precisely, if $X, Y \in W \subset H$, then we assume that there holds

$$\exists c > 0 : \qquad \|X - X_\ell\|_H \leq c \cdot N_\ell^{-\delta} \|X\|_W, \qquad \|Y - Y_\ell\|_H \leq c \cdot N_\ell^{-\delta} \|Y\|_W. \tag{3.3}$$

Typically, X_ℓ and Y_ℓ are elements of finite dimensional subspaces V_ℓ of H and $N_\ell := \dim(V_\ell)$ stands for its dimension. As we see later, X and Y will play roles of two independent realizations of the same random field. Hence, on the one hand, the possibility of having $X \neq Y$ will be necessary; on the other hand, this justifies assumption (3.3) for two sequences X_ℓ and Y_ℓ having *identical* asymptotic convergence rates.

Moreover, we shall from now on require that $\{N_\ell\}$ is an approximately exponentially increasing sequence, i.e. we assume that for some $a > 1$ there exists a constant $R \geq 1$, such that

$$R^{-1} \leq \frac{N_k}{a^k N_0} \leq R, \qquad \forall k \in \mathbb{N}. \tag{3.4}$$

Concerning (3.3), W is a subspace of H having a *stronger* norm, i.e.

$$\exists \nu > 0 : \qquad \|Z\|_H \leq \nu \|Z\|_W \qquad \forall Z \in W. \tag{3.5}$$

Lemma 3.2. *Let $X, Y \in H$ with associated sequences $\mathcal{X}$ and $\mathcal{Y}$, respectively, such that (3.3) and (3.5) hold. Then there holds*

$$\|X \otimes Y - X_L \otimes Y_L\|_{H \otimes H} \leq C N_L^{-\delta} \|X\|_W \|Y\|_W \tag{3.6}$$

where the constant C is independent of L and N_L.

Proof. Note the identity

$$X \otimes Y - X_L \otimes Y_L = X \otimes (Y - Y_L) + (X - X_L) \otimes Y - (X - X_L) \otimes (Y - Y_L).$$

Then the triangle inequality in $H \otimes H$, the crossnorm property of the canonical norm on $H \otimes H$, (3.3) and (3.5) imply

$$\begin{aligned} \|X \otimes Y - X_L \otimes Y_L\|_{H \otimes H} &\leq \|X \otimes (Y - Y_L)\|_{H \otimes H} + \|(X - X_L) \otimes Y\|_{H \otimes H} \\ &\quad + \|(X - X_L) \otimes (Y - Y_L)\|_{H \otimes H} \\ &\leq (2\nu c N_L^{-\delta} + c^2 N_L^{-2\delta}) \|X\|_W \|Y\|_W, \end{aligned}$$

which finishes the proof. □

Corollary 3.3. *Let $q \in [1,\infty]$ and $X, Y \in L^{2q}(\Omega; H)$ with associated sequences $\mathcal{X}$ and $\mathcal{Y}$, respectively, such that (3.3) and (3.5) hold. Then there holds*

$$\|X \otimes Y - X_L \otimes Y_L\|_{L^q(\Omega; H\otimes H)} \leq C N_L^{-\delta} \|X\|_{L^{2q}(\Omega;W)} \|Y\|_{L^{2q}(\Omega;W)} \tag{3.7}$$

where the constant C is independent of L and N_L.

Proof. The proof is completely analogous to that of the previous lemma. Using the triangle inequality for the $L^q(\Omega; H \otimes H)$-norm in combination with the crossnorm property of the $H \otimes H$-norm, the result follows by an application of the Cauchy-Schwarz inequality and the requirements on X and Y. □

3.2.2 Sparse tensor product approximation and convergence

Let us proceed by introducing the *sparse tensor product* approximation to a product $X \otimes Y$. First, for a sequence $\mathcal{X} = \{X_\ell\}$ we define increments by virtue of the difference operator Δ_ℓ by

$$\Delta_\ell \mathcal{X} := \begin{cases} X_\ell - X_{\ell-1}, & \ell \geq 1, \\ X_0, & \ell = 0, \end{cases} \tag{3.8}$$

which acts on suitable sequences $\mathcal{X}$ and analogously on $\mathcal{Y}$. This yields

$$X_n - X_k = \sum_{\ell=k+1}^{n} \Delta_\ell \mathcal{X}$$

and moreover we note the following expressions for later use

$$X_n = \sum_{\ell=0}^{n} \Delta_\ell \mathcal{X}, \qquad X - X_k = \sum_{\ell=k+1}^{\infty} \Delta_\ell \mathcal{X}, \qquad X = \sum_{\ell=0}^{\infty} \Delta_\ell \mathcal{X}. \tag{3.9}$$

These summation rules immediately imply the following observation. For a tensor product of the form $X_L \otimes X_L$ there holds

$$X_L \otimes X_L = \left(\sum_{k=0}^{L} \Delta_k \mathcal{X} \right) \otimes \left(\sum_{\ell=0}^{L} \Delta_\ell \mathcal{X} \right) = \sum_{k,\ell=0}^{L} (\Delta_k \mathcal{X} \otimes \Delta_\ell \mathcal{X}) . \tag{3.10}$$

In the following we show that it makes sense to consider

$$\sum_{k+\ell \leq L} (\Delta_k \mathcal{X} \otimes \Delta_\ell \mathcal{X}) \approx X_L \otimes X_L$$

as an approximation and quantify the error we make by restricting the summation range. This is a simple and fast but crude way of describing the notion of sparse tensor approximation techniques. For a proper introduction and more details we refer the reader to the excellent articles [33, 34].

Following this idea we define the sparse tensor product approximation operator $\hat{P}_L(\mathcal{X}, \mathcal{Y})$, which we define to act on sequences $\mathcal{X}$ and $\mathcal{Y}$ according to the following definition.

Definition 3.4. *Let $\mathcal{X} = \{X_\ell\}$ and $\mathcal{Y} = \{Y_\ell\}$ be sequences that converge to X and Y at a certain rate (cf. 3.3), respectively. Then we define the sparse tensor product approximation operator by*

$$\hat{P}_L(\mathcal{X}, \mathcal{Y}) := \sum_{k+j \leq L} \Delta_k \mathcal{X} \otimes \Delta_j \mathcal{Y}, \tag{3.11}$$

where here and in what follows the summation is understood over nonnegative indices $k, j \geq 0$.

We now show the approximation property of $\hat{P}_L(\cdot,\cdot)$, which plays a key role in the sparse approximation methods we will propose and analyze. To this end we note that assumption (3.4) particularly implies that

$$\log_a(N_k) - \log_a(RN_0) \leq k \leq \log_a(N_k) + \log_a(R/N_0) \qquad \forall k \in \mathbb{N} \tag{3.12}$$

and

$$R^{-3} N_0 N_{k+j} \leq N_k N_j \leq R^3 N_0 N_{k+j} \qquad \forall k, j \in \mathbb{N}. \tag{3.13}$$

The uniform bounds (3.12) and (3.13) will be central in the forthcoming analysis.

Lemma 3.5. *Let $X, Y \in H$ and $\mathcal{X}, \mathcal{Y}$ denote sequences that converge to X and Y, respectively, at a given rate specified by (3.3) with respect to a subspace W of H with a stronger norm (cf. (3.5)). Moreover, assume that (3.4) holds. Then*

$$\|X \otimes Y - \hat{P}_L(\mathcal{X}, \mathcal{Y})\|_{H\otimes H} \leq (C_1 + C_2 \log N_L) \cdot N_L^{-\delta} \cdot \|X\|_W \|Y\|_W \tag{3.14}$$

where the constants C_1 and C_2 are independent of L and $N_0, \ldots, N_L$. Moreover, if N_0 is sufficiently large, the constant C_2 may be forced to satisfy $C_2 \leq \epsilon$ for any fixed $\epsilon > 0$.

Proof. The summation properties (3.9) imply $X \otimes Y = \sum_{k,j} \Delta_k \mathcal{X} \otimes \Delta_j \mathcal{Y}$. Hence, by (3.11) and the bilinearity of the tensor product we obtain

$$\begin{aligned} X \otimes Y - \hat{P}_L(\mathcal{X}, \mathcal{Y}) &= \sum_{k+j>L} \Delta_k \mathcal{X} \otimes \Delta_j \mathcal{Y} \\ &= \sum_{k=0}^{L} \left(\Delta_k \mathcal{X} \otimes \sum_{j=L+1-k}^{\infty} \Delta_j \mathcal{Y} \right) + \sum_{k=L+1}^{\infty} \left(\Delta_k \mathcal{X} \otimes \sum_{j=0}^{\infty} \Delta_j \mathcal{Y} \right) \\ &= X_0 \otimes (Y - Y_L) + \sum_{k=1}^{L} \Delta_k \mathcal{X} \otimes (Y - Y_{L-k}) + (X - X_L) \otimes Y, \end{aligned}$$

where (3.9) has been used in the last step. Regrouping the terms we arrive at a symmetric representation

$$\begin{aligned} X \otimes Y - \hat{P}_L(\mathcal{X}, \mathcal{Y}) =& X \otimes (Y - Y_L) + (X - X_L) \otimes Y \\ &+ \sum_{k+j=L-1} (X - X_k) \otimes (Y - Y_j) - \sum_{k+j=L} (X - X_k) \otimes (Y - Y_j). \end{aligned}$$

The triangle inequality, (3.3) and the crossnorm property of the $\|\cdot\|_{H\otimes H}$-norm imply

$$\begin{aligned} \|X \otimes Y - \hat{P}_L(\mathcal{X}, \mathcal{Y})\|_{H\otimes H} &\leq \|X \otimes (Y - Y_L)\|_{H\otimes H} + \|(X - X_L) \otimes Y\|_{H\otimes H} \\ &+ \sum_{k+j=L-1} \|(X - X_k) \otimes (Y - Y_j)\|_{H\otimes H} + \sum_{k+j=L} \|(X - X_k) \otimes (Y - Y_j)\|_{H\otimes H} \\ &\leq c^2 \|X\|_W \|Y\|_W \left(\frac{2\gamma}{c} N_L^{-\delta} + \sum_{k+j=L-1} (N_k N_j)^{-\delta} + \sum_{k+j=L} (N_k N_j)^{-\delta} \right) \end{aligned}$$

From (3.4) and (3.13) we find

$$\sum_{k+j=L} \left(\frac{N_0 N_L}{N_k N_j} \right)^{\delta} \leq R^{3\delta}(L+1), \qquad \frac{N_L}{N_{L-1}} \leq aR^2$$

and therefore

$$\begin{aligned}&\|X \otimes Y - \hat{P}_L(\mathcal{X}, \mathcal{Y})\|_{H\otimes H} \\ &\qquad \leq c^2 N_L^{-\delta} \|X\|_W \|Y\|_W \left(\frac{2\gamma}{c} + 2 \left(\frac{aR^5}{N_0} \right)^{\delta} (L+1) \right).\end{aligned} \tag{3.15}$$

The assertion now follows by the upper bound in (3.12). □

Corollary 3.6. *Let $q \in [1, \infty]$, $X, Y \in L^{2q}(\Omega; H)$ and $\mathcal{X}, \mathcal{Y}$ denote sequences that converge to X and Y, respectively, at a given rate specified by (3.3) with respect to a subspace W of H with a stronger norm (cf. (3.5)). Moreover, assume that (3.4) holds. Then*

$$\|X \otimes Y - \hat{P}_L(\mathcal{X}, \mathcal{Y})\|_{L^q(\Omega; H\otimes H)} \leq (C_1 + C_2 \log N_L) \cdot N_L^{-\delta} \cdot \|X\|_{L^{2q}(\Omega;W)} \|Y\|_{L^{2q}(\Omega;W)} \tag{3.16}$$

where the constants C_1 and C_2 are independent of L and $N_0, \ldots, N_L$. Moreover, if N_0 is sufficiently large, the constant C_2 may be forced to satisfy $C_2 \leq \epsilon$ for any fixed $\epsilon > 0$.

Proof. Now using the triangle inequality for the $L^q(\Omega; H \otimes H)$-norm in combination with the crossnorm property of the norm on $H \otimes H$, the result follows by an application of the Cauchy-Schwarz inequality and the requirements on X and Y. □

3.3 Monte Carlo and multilevel Monte Carlo methods

In order to construct Monte Carlo methods for Problem 3.1, we will first recall well-known results from the literature and lay the necessary groundwork for the more advanced topics. We start by recalling the approximation of the mean by Monte Carlo (MC) and multilevel Monte Carlo (MLMC) methods in a general context, i.e. in the context of Hilbert space valued random fields. In a similar fashion, approximation of covariance functions in the single level and multilevel context will then be presented. For more details on the presentation we refer the reader to [7, 9, 8, 20] and the references therein.

3.3.1 Monte Carlo approximation of the mean

In preparation for the subsequent analysis we suppose that $X \in L^2(\Omega; H)$ and recall some standard notations and results of Monte Carlo (MC) methods for the approximation of the mean of a Hilbert space valued random field X.

It is a well-known result that the expectation of a Hilbert space valued random field $X \in L^2(\Omega; H)$ can be approximated by its sample mean. Assuming $\{X^i\}_{i=1}^M$ are *independent identically distributed* (iid) samples of X, the sample mean is given by

$$\mathrm{E}_M[X] = \frac{1}{M} \sum_{i=1}^{M} X^i,$$

where $M \in \mathbb{N}$ is the size of the underlying sample ensemble. Moreover, there holds the following representation of the mean square error (cf. [7])

$$\mathrm{MSE} = \|\mathbb{E}[X] - \mathrm{E}_M[X]\|^2_{L^2(\Omega;H)} = \frac{1}{M} \|X - \mathbb{E}[X]\|^2_{L^2(\Omega;H)}. \tag{3.17}$$

Since the exact random field X is commonly not available for sampling, a suitable approximation X_ℓ of X is chosen for sampling instead. Here we consider X_ℓ as an element of a sequence $\mathcal{X} = \{X_\ell\}$ converging to X, where the convergence is now understood to

take place in the space $L^2(\Omega; H)$. Hence, the exactness of the sample will also depend on the convergence rate of the sequence and we denote the set of samples of X on level ℓ of the discretisation by $\{X_\ell^i\}_{i=1}^M$. Since for iid samples X_ℓ^i there holds the identity $\mathbb{E}[\mathrm{E}_M[X_\ell]] = \mathbb{E}[X_\ell]$, i.e. the sample mean is an unbiased estimator, and by the linearity of the inner product in $L^2(\Omega; H)$ in the first argument, we have the following convenient splitting of the mean square error (cf. also [7])

$$\|\mathbb{E}[X] - \mathrm{E}_M[X_\ell]\|^2_{L^2(\Omega;H)} = \|\mathbb{E}[X - X_\ell]\|_H^2 + \frac{1}{M}\|X_\ell - \mathbb{E}[X_\ell]\|^2_{L^2(\Omega;H)}. \tag{3.18}$$

The first term in this splitting characterises the *discretisation error*, which we shall subsequently call the *bias*, whereas the second term provides information on the *sampling error* and resembles a variance-like operator that we shall frequently abbreviate by the expression

$$\mathcal{V}(X) \equiv \mathcal{V}(X, H) = \|X - \mathbb{E}[X]\|^2_{L^2(\Omega;H)}, \tag{3.19}$$

where the Hilbert space H will be omitted from the notation, if it is clear from the context.

In order to quantify the work needed to compute the sample mean we proceed as follows. Let ℓ be fixed and suppose the cost C_ℓ to obtain a single sample X_ℓ^i is bounded as $C_\ell \lesssim N_\ell^\gamma$, where here and in the following $\lesssim$ shall denote $\leq C$ with a certain but unspecified constant C, i.e. smaller up to a constant factor C. Then for a given error tolerance ε_ℓ and

$$\|\mathbb{E}[X] - \mathrm{E}_M[X_\ell]\|_{L^2(\Omega;H)} \leq \varepsilon_\ell$$

the ensued total work is given as $\mathrm{Work}(\mathrm{E}_M[X_\ell]) = MC_\ell$. In light of (3.17) we infer that the optimal sample size M_ℓ is proportional to ε_ℓ^{-2}. By Jensen's inequality and (3.3) we find

$$\|\mathbb{E}[X - X_\ell]\|_H \lesssim N_\ell^{-\alpha}, \qquad \alpha \geq \delta,$$

which in turn implies that the cost of evaluation of a single sample scales as $C_\ell \sim \varepsilon_\ell^{-\gamma/\alpha}$. This tells us that in the worst case we are restricted by the deterministic convergence rate δ and α is considered to be the weak convergence rate, which might exceed δ. This yields a total work estimate of

$$\mathrm{Work}(\mathrm{E}_{M_\ell}[X_\ell]) \lesssim \varepsilon^{-2-\frac{\gamma}{\alpha}}.$$

This shows that the MC method behaves rather unfavourably as one does not only have to reduce the discretisation error but also has to enlarge the size of the sample ensemble. These results are summarized in the following Theorem for which a proof can be found in [5], also confer [20].

Theorem 3.7. *Suppose C_ℓ is the cost of evaluation of one sample X_ℓ^i and $\{N_\ell\}_{\ell=0}^\infty$ is a sequence satisfying $\frac{N_\ell}{N_{\ell-1}} \geq a$ for some fixed $a > 1$. Assume there exist $\alpha, \gamma > 0$, such that there holds*

$$\|\mathbb{E}[X - X_\ell]\|_H \lesssim N_\ell^{-\alpha}, \qquad C_\ell \lesssim N_\ell^\gamma.$$

Then for any mean square error (MSE) tolerance ε_ℓ^2, i.e. $\|\mathbb{E}[X] - \mathrm{E}_M[X_\ell]\|^2_{L^2(\Omega;H)} < \varepsilon_\ell^2$, there exists $M = M(\varepsilon)$, such that the cost of evaluation of $\mathrm{E}_M[X_\ell]$ satisfies the asymptotic bound

$$\mathrm{Cost}(\mathrm{E}_M[X_\ell]) \lesssim \varepsilon^{-2-\gamma/\alpha}.$$

The mulitplicative cost estimate for the single level Monte Carlo method can be overcome by means of the so-called multilevel Monte Carlo method with which we will concern ourselves next.

3.3.2 Multilevel Monte Carlo approximation of the mean

The main idea of the multilevel Monte Carlo (MLMC) method is that for given L we can write X_L by means of a telescoping sum according to (3.9) as

$$X_L = \sum_{\ell=0}^{L} \Upsilon_\ell, \text{ with } \Upsilon_\ell = \Delta_\ell \mathcal{X} \text{ and } X_{-1} := 0,$$

where the terms Υ_ℓ are commonly referred to as *level corrections*. The expectation $\mathbb{E}[X]$ can then be approximated by the multilevel sample mean

$$\mathrm{E}^{\mathrm{ML}}[X] := \sum_{\ell=0}^{L} \mathrm{E}_{M_\ell}[\Upsilon_\ell]. \tag{3.20}$$

It can be immediately seen from the definition of the multilevel sample mean, that now for different values of ℓ the sample mean is taken with respect to a (possibly) varying size of samples M_ℓ.

In order to have an optimal cost to accuracy relation, the size of the sample ensembles on each level, i.e. for each level correction Υ_ℓ, is chosen in such a way that level corrections composed of coarser approximations are sampled more in comparison to level corrections of finer approximations.

We quote the following results from [7] which give the splitting of the mean square error and the entailed total cost for the MLMC approximation of the mean.

Theorem 3.8. *Suppose H is a separable Hilbert space and let $X, \{X_\ell\}_{\ell=1}^{L} \in L^2(\Omega; H)$ be random fields. Then*

$$\mathrm{MSE} = \|\mathbb{E}[X] - \mathrm{E}^{\mathrm{ML}}[X]\|^2_{L^2(\Omega;H)} = \|\mathbb{E}[X - X_L\|^2_H + \sum_{\ell=1}^{L} \frac{1}{M_\ell} \mathcal{V}(X_\ell - X_{\ell-1}). \tag{3.21}$$

Theorem 3.9. *Let $C_\ell := \mathrm{Cost}(Y_\ell^i)$ denote the cost of the evaluation of a single sample of the level correction $Y_\ell^i = X_\ell^i - X_{\ell-1}^i$, and $\{N_\ell\}_{\ell=1}^{\infty}$ be an exponentially increasing sequence with the property $N_\ell / N_{\ell-1} \geq a$ for some fixed $a > 1$. Moreover, let $\alpha, \beta, \gamma > 0$ assume that the following asymptotic bounds are valid*

$$1)\ \|\mathbb{E}[X - X_\ell]\|_H \lesssim N_\ell^{-\alpha}, \quad 2)\ \mathcal{V}(Y_\ell) \lesssim N_\ell^{-\beta}, \quad 3)\ C_\ell \lesssim N_\ell^{\gamma}. \tag{3.22}$$

Then for any error tolerance $\varepsilon > 0$ and $\|\mathbb{E}[X] - \mathrm{E}^{\mathrm{ML}}[X]\|_{L^2(\Omega;H)} < \varepsilon$ there exists $L \in \mathbb{N}$ and a sequence $M_1, \ldots, M_L \in \mathbb{N}$, such that the total cost of the multilevel estimator admits the asymptotic bound

$$\mathrm{Cost}(\mathrm{E}^{\mathrm{ML}}[X]) \lesssim \varepsilon^{-\gamma/\alpha} + \begin{cases} \varepsilon^{-2}, & \beta > \gamma, \\ \varepsilon^{-2} |\log(\varepsilon)|^2, & \beta = \gamma, \\ \varepsilon^{-2 - \frac{\gamma - \beta}{\alpha}}, & \beta < \gamma. \end{cases} \tag{3.23}$$

Remark 3.10. *For convenience of our analysis we start counting the levels in the multilevel Monte Carlo Method from zero to be more in tune with the summation rules of (3.9). Note that this index shift is merely a matter of convenience.*

3.3.3 Monte Carlo approximation of covariance functions

Let us recall the covariance function of two elements X, Y of the space $L^2(\Omega; H)$, which takes the form

$$\mathbb{C}\mathrm{ov}\left[X, Y\right] = \mathbb{E}\left[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])\right]. \tag{3.24}$$

With this in hand we adopt the covariance estimator from [42] and adapt it to the situation of tensor products of Hilbert spaces.

Definition 3.11. *Let $X, Y \in L^2(\Omega; H)$ and $M \in \mathbb{N}$ a number of samples, then we define the sample covariance estimator by the expression*

$$\begin{aligned} \mathrm{Cov}_M\left[X, Y\right] :=& \frac{1}{M-1} \sum_{i=1}^{M} (X^i - \mathrm{E}_M[X]) \otimes (Y^i - \mathrm{E}_M[Y]) \\ =& \frac{M}{M-1} \left(\mathrm{E}_M[X \otimes Y] - \mathrm{E}_M[X] \otimes \mathrm{E}_M[Y]\right). \end{aligned} \tag{3.25}$$

We note that, since there holds

$$\mathbb{E}[\mathrm{Cov}_M\left[X, Y\right]] = \frac{M}{M-1} \left(\mathbb{E}[X \otimes Y] - \frac{1}{M^2} \sum_{i,j} \mathbb{E}[X^i \otimes Y^j] \right) = \mathbb{C}\mathrm{ov}\left[X, Y\right] \tag{3.26}$$

that $\mathrm{Cov}_M\left[\cdot, \cdot\right]$ is an unbiased estimator. Moreover, note that if X and Y are centered, i.e. $\mathbb{E}[X] = \mathbb{E}[Y] = 0$, that there holds $\mathbb{C}\mathrm{ov}\left[X, Y\right] = \mathbb{C}\mathrm{ov}\left[\overline{X}, \overline{Y}\right]$.

Remark 3.12. *In order to compute the covariance estimator we will use the following update formula (cf. [42, Remark 3.1]). Denote by*

$$C = \sum_{i=1}^{M} (X^i - \mathrm{E}_M[X]) \otimes (Y^i - \mathrm{E}_M[Y])$$

and draw one new sample of X and Y w.r.t. ω^j. Then C can be updated according to

$$C = C + \frac{M-1}{M} (X^j - \mathrm{E}_M[X]) \otimes (Y^j - \mathrm{E}_M[Y]). \tag{3.27}$$

Dividing C afterwards by $(\tilde{M} - 1)$, where $\tilde{M} = M + 1$ is the total amount of samples after updating, then C constitutes an unbiased covariance estimator (cf. (3.26)).

In the forthcoming analysis we study the convergence of the mean square error, which here takes the form

$$\mathrm{MSE}\left[\mathrm{Cov}_M\left[X, Y\right]\right] := \mathbb{E}\left[\|\mathrm{Cov}_M\left[X, Y\right] - \mathbb{C}\mathrm{ov}\left[X, Y\right]\|_{H \otimes H}^2\right]. \tag{3.28}$$

There are multiple reasons for this approach. In particular, MSE is a standard risk measure and the associated space $L^2(\Omega; H \otimes H)$ is also a Hilbert space. This enables the use of the Hilbertian structure in the forthcoming analysis which results in simple and sharp estimates.

Before we prepare the analysis of the Monte Carlo methods, let us define a *covariance-like* operator for the sake of simplicity of notation by

$$\mathcal{C}[A, B] := \mathbb{E}\left[\langle A - \mathbb{E}[A], B - \mathbb{E}[B] \rangle_H\right], \tag{3.29}$$

where $A, B \in L^2(\Omega; H)$.

The following technical lemma will enable the analysis of the proposed single level and multilevel Monte Carlo methods.

Lemma 3.13. *Let H be a Hilbert space and $A, B, S, T : \Omega \to H$ centered random fields, i.e. $\mathbb{E}[A] = \mathbb{E}[B] = \mathbb{E}[S] = \mathbb{E}[T] = 0$. Then the following identity holds*

$$\begin{aligned}\mathbb{E}[\langle \mathrm{Cov}_M[A,B], \mathrm{Cov}_M[S,T]\rangle_{H\otimes H}] - \langle \mathbb{C}\mathrm{ov}[A,B], \mathbb{C}\mathrm{ov}[S,T]\rangle_{H\otimes H} \\ = \frac{1}{M}\mathcal{C}[A\otimes B, S\otimes T] + \frac{1}{M(M-1)}F(A,B,S,T),\end{aligned} \tag{3.30}$$

where the higher order terms are expressed by

$$F(A,B,S,T) = \mathcal{C}[A,S]\cdot\mathcal{C}[B,T] + \langle \mathbb{C}\mathrm{ov}[A,T], \mathbb{C}\mathrm{ov}[S,B]\rangle. \tag{3.31}$$

Proof. In view of (3.25) the left-hand side of (3.30) can be written as

$$\begin{aligned}e^2 := \frac{M^2}{(M-1)^2}\mathbb{E}\Big[&\langle E_M[A\otimes B], E_M[S\otimes T]\rangle - \langle E_M[A\otimes B], E_M[S]\otimes E_M[T]\rangle \\ &- \langle E_M[A]\otimes E_M[B], E_M[S\otimes T]\rangle + \langle E_M[A]\otimes E_M[B], E_M[S]\otimes E_M[T]\rangle\Big] \\ &- \langle \mathbb{E}[A\otimes B], \mathbb{E}[S\otimes T]\rangle.\end{aligned} \tag{3.32}$$

For the first term we obtain

$$f := \mathbb{E}\langle E_M[A\otimes B], E_M[S\otimes T]\rangle = \frac{1}{M^2}\sum_{ij}\mathbb{E}\langle A^i\otimes B^i, S^j\otimes T^j\rangle.$$

Splitting the sum into two parts—with $i = j$ and $i \neq j$—we observe that

$$\begin{aligned}f &= \frac{1}{M}\Big(\mathbb{E}\langle A\otimes B, S\otimes T\rangle + (M-1)\langle \mathbb{E}[A\otimes B], \mathbb{E}[S\otimes T]\rangle\Big) \\ &= \frac{1}{M}\mathcal{C}[A\otimes B, S\otimes T] + \langle \mathbb{E}[A\otimes B], \mathbb{E}[S\otimes T]\rangle.\end{aligned} \tag{3.33}$$

For the second term there holds

$$\begin{aligned}\mathbb{E}\langle E_M[A\otimes B], E_M[S]\otimes E_M[T]\rangle &= \frac{1}{M^3}\sum_{ijk}\mathbb{E}\langle A^k\otimes B^k, S^i\otimes T^j\rangle \\ &= \frac{1}{M^3}\sum_{ik}\mathbb{E}\langle A^k\otimes B^k, S^i\otimes T^i\rangle = \frac{1}{M}f,\end{aligned}$$

where the second identity holds, because the summands with $i \neq j$ are zero. Indeed, in this case either i or j are different from k. Assume w.l.o.g. that $j \neq k$, then

$$\mathbb{E}\langle A^k\otimes B^k, S^i\otimes T^j\rangle = \mathbb{E}\langle\langle A^k, S^i\rangle B^k, T^j\rangle = \langle \mathbb{E}[\langle A^k, S^i\rangle B^k], \mathbb{E}[T^j]\rangle = 0,$$

since T is centered, i.e. $\mathbb{E}[T] = 0$. By symmetry, the same representation holds for the third term in (3.32). The fourth term takes the form

$$\mathbb{E}\langle E_M[A]\otimes E_M[B], E_M[S]\otimes E_M[T]\rangle = \frac{1}{M^4}\sum_{ijk\ell}\mathbb{E}\langle A^i\otimes B^j, S^k\otimes T^\ell\rangle.$$

We split the sum into three parts: 1) $i = j \wedge k = \ell$, 2) $i = k \neq j = \ell$ and 3) $i = \ell \neq j = k$. The first sum is proportional to f and the remaining part to $F(A, B, S, T)$:

$$\frac{1}{M^4} \sum_{ik} \mathbb{E}\langle A^i \otimes B^i, S^k \otimes T^k \rangle = \frac{1}{M^2} f,$$

$$\begin{aligned} \frac{1}{M^4} \sum_{i \neq j} \mathbb{E}\langle A^i \otimes B^j, S^i \otimes T^j \rangle &= \frac{1}{M^4} \sum_{i \neq j} \mathbb{E}[\langle A^i, S^i \rangle] \cdot \mathbb{E}[\langle B^j, T^j \rangle] \\ &= \frac{M-1}{M^3} \mathcal{C}[A, S] \cdot \mathcal{C}[B, T], \end{aligned} \tag{3.34}$$

$$\begin{aligned} \frac{1}{M^4} \sum_{i \neq j} \mathbb{E}\langle A^i \otimes B^j, S^j \otimes T^i \rangle &= \frac{1}{M^4} \sum_{i \neq j} \mathbb{E}\langle A^i \otimes T^i, S^j \otimes B^j \rangle \\ &= \frac{M-1}{M^3} \langle \mathbb{C}\mathrm{ov}\left[A, T\right], \mathbb{C}\mathrm{ov}\left[S, B\right] \rangle. \end{aligned} \tag{3.35}$$

Hence,

$$\langle E_M[A] \otimes E_M[B], E_M[S] \otimes E_M[T] \rangle = \frac{1}{M^2} f + \frac{M-1}{M^3} F(A, B, S, T).$$

Collecting above representations for the terms in (3.32) we obtain

$$\begin{aligned} e^2 &= \frac{M^2}{(M-1)^2} \left(f \left(1 - \frac{2}{M} + \frac{1}{M^2} \right) + \frac{M-1}{M^3} F(A, B, S, T) \right) - \langle \mathbb{C}\mathrm{ov}\left[A, B\right], \mathbb{C}\mathrm{ov}\left[S, T\right] \rangle \\ &= \frac{1}{M} \mathcal{C}[A \otimes B, S \otimes T] + \frac{1}{M(M-1)} F(A, B, S, T). \end{aligned}$$

This finishes the proof. □

Remark 3.14. *When setting $A = S$ and $B = T$, it is readily seen that the aforementioned lemma characterises the* MSE *with respect to the covariance estimator of Definition 3.11.*

The following lemma characterises the sampling error with respect to the covariance estimator from Definition 3.11, when the exact random fields X and Y are available for sampling.

Lemma 3.15. *Suppose $X, Y \in L^4(\Omega; H)$ and recall definitions (3.24) and (3.25). Then for $\mathbb{C} := \mathbb{C}\mathrm{ov}\left[X, Y\right]$ and $C_M := \mathrm{Cov}_M\left[X, Y\right]$ there holds*

$$\|C_M - \mathbb{C}\|^2_{L^2(\Omega; H \otimes H)} \leq \frac{M+1}{M(M-1)} \|X - \mathbb{E}[X]\|^2_{L^4(\Omega; H)} \|Y - \mathbb{E}[Y]\|^2_{L^4(\Omega; H)}. \tag{3.36}$$

Proof. Assume without loss of generality that $\mathbb{E}[X] = 0 = \mathbb{E}[Y]$, which readily implies

$$\begin{aligned} \|C_M - \mathbb{C}\|^2_{L^2(\Omega; H \otimes H)} &= \|C_M\|^2_{L^2(\Omega; H \otimes H)} - 2\mathbb{E}\left[\langle C_M, \mathbb{C} \rangle_{H \otimes H}\right] + \|\mathbb{C}\|^2_{L^2(\Omega; H \otimes H)} \\ &= \|C_M\|^2_{L^2(\Omega; H \otimes H)} - \|\mathbb{C}\|^2_{L^2(\Omega; H \otimes H)}. \end{aligned} \tag{3.37}$$

Then by Lemma 3.13 we find the splitting

$$\begin{aligned} \|C_M - \mathbb{C}\|^2_{L^2(\Omega; H \otimes H)} = \frac{1}{M} \mathcal{C}[X \otimes Y, X \otimes Y] \\ + \frac{1}{M(M-1)} \left(\mathcal{C}[X, X] \cdot \mathcal{C}[Y, Y] + \|\mathbb{C}\|^2_{H \otimes H} \right). \end{aligned}$$

Now, since

$$\mathcal{C}[X \otimes Y, X \otimes Y] = \|\overline{X \otimes Y}\|^2_{L^2(\Omega;H\otimes H)} \leq \|X\|^2_{L^4(\Omega;H)}\|Y\|^2_{L^4(\Omega;H)}$$

by virtue of the Cauchy-Schwarz inequality as well as

$$\mathcal{C}[X,X] \cdot \mathcal{C}[Y,Y] = \|X\|^2_{L^2(\Omega;H)}\|Y\|^2_{L^2(\Omega;H)},$$

we find by another application of the Cauchy-Schwarz and Jensen's inequality that

$$\|\mathbb{C}\|^2_{H\otimes H} \leq \left(\mathbb{E}\left[\|X\|_H\|Y\|_H\right]\right)^2 \leq \|X\|^2_{L^2(\Omega;H)}\|Y\|^2_{L^2(\Omega;H)},$$

which leads us to verify the assertion with the bound $\|X\|_{L^2(\Omega;H)} \leq \|X\|_{L^4(\Omega;H)}$. □

Since the exact random fields X and Y are in general *not* available, we have to make due with certain approximations to the exact random fields. The error this entails is quantified by means of the following

Lemma 3.16. *Suppose $X, Y \in L^4(\Omega;H)$ and let X_L, Y_L satisfy*

$$\|X_L - \mathbb{E}[X_L]\|_{L^4(\Omega;H)} \leq \eta\|X\|_{L^4(\Omega;W)}, \qquad \|Y_L - \mathbb{E}[Y_L]\|_{L^4(\Omega;H)} \leq \eta\|Y\|_{L^4(\Omega;W)}. \tag{3.38}$$

Then for $\mathbb{C} := \mathbb{C}\mathrm{ov}\,[X,Y]$ and $C_{L;M} := \mathrm{Cov}_M\,[X_L, Y_L]$ there holds

$$\|C_{L;M} - \mathbb{C}\|^2_{L^2(\Omega;H\otimes H)} \leq \left(c^2 \cdot N_L^{-2\delta} + \eta^2 \frac{M+1}{M(M-1)}\right)\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}. \tag{3.39}$$

Proof. Without loss of generality we assume that X and Y are centered, i.e. $\mathbb{E}[X] = 0 = \mathbb{E}[Y]$, and we abbreviate $\mathbb{C}_L := \mathbb{E}[C_{L;M}]$. Then by the unbiasedness of the estimator $C_{L;M}$ we have that

$$\|C_{L;M} - \mathbb{C}\|^2_{L^2(\Omega;H\otimes H)} = \|\mathbb{C}_L - \mathbb{C}\|^2_{H\otimes H} + \|C_{L;M} - \mathbb{C}_L\|^2_{L^2(\Omega;H\otimes H)}. \tag{3.40}$$

By Jensen's inequality in combination with (3.6) we have that

$$\|\mathbb{C} - \mathbb{C}_L\|_{H\otimes H} \leq cN_L^{-\delta}\|X\|_{L^2(\Omega;W)}\|Y\|_{L^2(\Omega;W)} \leq cN_L^{-\delta}\|X\|_{L^4(\Omega;W)}\|Y\|_{L^4(\Omega;W)}. \tag{3.41}$$

Moreover, by Lemma 3.15 with X replaced by X_L and Y by Y_L, respectively, there holds

$$\|C_{L;M} - \mathbb{C}_L\|^2_{L^2(\Omega;H\otimes H)} \leq \frac{M+1}{M(M-1)}\|X_L\|^2_{L^4(\Omega;H)}\|Y_L\|^2_{L^4(\Omega;H)}.$$

We arrive at the claim by means of stability expressed by(3.38). □

We are now in a position to state the following result about the cost and accuracy of the full tensor product covariance single level Monte Carlo estimator .

Theorem 3.17. *Suppose that the evaluation cost of a sample of X_L (and likewise for Y_L) satisfies the bound*

$$\mathrm{Cost}(X_L) \lesssim N_L^{\gamma}.$$

Then under assumptions of Lemma 3.16 the accuracy

$$\|C_{L;M} - \mathbb{C}\|_{L^2(\Omega,H\otimes H)} < \varepsilon$$

can be achieved for any $\varepsilon > 0$ at the computational cost and memory requirements

$$\mathrm{Cost}(C_{L;M}) \lesssim \varepsilon^{-2-\frac{\max\{\gamma,2\}}{\delta}} \tag{3.42}$$

and memory

$$\mathrm{Memory}(C_{L;M}) \sim N_L^2. \tag{3.43}$$

Proof. By Lemma 3.16 the mean square error of the full tensor product Monte Carlo approximation of the covariance function admits the asymptotic bound

$$\|C_{L;M} - \mathbb{C}\|^2_{L^2(\Omega, H\otimes H)} \lesssim N_L^{-2\delta} + M^{-1}.$$

This implies $N_L^{-2\delta} \sim \varepsilon^2 \sim M^{-1}$ for an optimal balancing of cost versus accuracy. Since we have to compute M samples, where each sample is of cost N_L^γ and per sample we have to compute a tensor product to update the covariance estimator, which takes N_L^2 operations, the preceding observations and assumptions imply that

$$\mathrm{Cost}(C_{L;M}) \lesssim M \cdot (N_L^\gamma + N_L^2) = \varepsilon^{-2} \cdot (\varepsilon^{-\gamma/\delta} + \varepsilon^{-2/\delta}) \lesssim \varepsilon^{-2-\frac{\max\{\gamma,2\}}{\delta}}.$$

The definition of $C_{L;M}$ readily shows that on a given level L there are N_L^2 coefficients to represent the approximate covariance function, which we update according to (3.27). To evaluate the computed approximation of the covariance function all N_L^2 coefficients are necessary whence the assertion follows. □

Remark 3.18. *Note that due to the maximum in the exponent of ε in (3.42) there is no benefit of using any solver of sub-quadratic complexity, i.e. for $\gamma \in [1, 2)$, as the update of the coefficients of the covariance will dominate the overall costs anyway.*

As a remedy for the unfavorable growth of cost and memory requirements we propose a sparse tensor single level covariance estimator as follows.

Definition 3.19. *Let $X, Y \in L^2(\Omega; H)$ and $\mathcal{X}, \mathcal{Y}$ sequences converging to X and Y in H, respectively. The sparse tensor covariance estimator $\widehat{\mathrm{Cov}}_M[X_L, Y_L]$ at level L for the approximation of $\mathbb{C}\mathrm{ov}[X, Y]$ is then defined as*

$$\begin{aligned}\widehat{\mathrm{Cov}}_M[X_L, Y_L] &:= \sum_{k+j\le L} \mathrm{Cov}_M[\Delta_k\mathcal{X}, \Delta_j\mathcal{Y}] \\ &= \frac{1}{M-1}\sum_{i=1}^{M}\sum_{k+j\le L} (\Delta_k\mathcal{X}^i - \mathrm{E}_M[\Delta_k\mathcal{X}]) \otimes (\Delta_j\mathcal{Y}^i - \mathrm{E}_M[\Delta_j\mathcal{Y}]).\end{aligned} \tag{3.44}$$

Theorem 3.20. *Let $\{N_\ell\}$ be an approximately exponentially increasing sequence with respect to some $a > 1$, such that (3.4),(3.12) and (3.13) hold. Furthermore, assume there exist $\gamma, \delta > 0$, such that*

$$\begin{cases} \|X - X_L\|_H \lesssim N_L^{-\delta}\|X\|_W, \\ \|Y - Y_L\|_H \lesssim N_L^{-\delta}\|Y\|_W, \\ \mathrm{Cost}(X_L) \lesssim N_L^\gamma \text{ and } \mathrm{Cost}(Y_L) \lesssim N_L^\gamma, \end{cases}$$

where $X, Y \in L^4(\Omega; H)$ with sequences $\mathcal{X} = \{X_\ell\}$ and $\mathcal{Y} = \{Y_\ell\}$ converging to X and Y, respectively. Then for all $\varepsilon > 0$ and

$$\|\mathbb{C}\mathrm{ov}[X, Y] - \widehat{\mathrm{Cov}}_M[X_L, Y_L]\|_{L^2(\Omega; H\otimes H)} < \varepsilon,$$

there holds

$$\mathrm{Cost}(\widehat{\mathrm{Cov}}_M[X_L, Y_L]) \lesssim \varepsilon^{-2-\gamma/\delta} \cdot \begin{cases} |\log(\varepsilon)|^{1+1/\delta}, & \gamma = 1, \\ |\log(\varepsilon)|^{\gamma/\delta}, & \gamma > 1, \end{cases}$$

and we require

$$\mathrm{Memory}(\widehat{\mathrm{Cov}}_M[X_L, Y_L]) \sim N_L \log(N_L).$$

Proof. Without loss of generality let us assume that X and Y are centered. For $\mu := \mathbb{C}\mathrm{ov}[X,Y]$, $m_L := \widehat{\mathrm{Cov}}_M[X_L,Y_L]$ and $\mu_L := \mathbb{E}[m_L]$ we note

$$\begin{aligned}\mathrm{MSE} &= \|\mathbb{C}\mathrm{ov}[X,Y] - \widehat{\mathrm{Cov}}_M[X_L,Y_L]\|^2_{L^2(\Omega;H\otimes H)}\\ &= \|\mu-\mu_L\|^2_{L^2(\Omega;H\otimes H)} + \|\mu_L - m_L\|^2_{L^2(\Omega;H\otimes H)} \equiv I_1 + I_2.\end{aligned}$$

Considering I_1, we observe by means of Jensen's inequality and Lemma 3.5 that

$$\sqrt{I_1} \le \left\| \sum\nolimits_{k+j>L} \overline{\Delta_k \mathcal{X}} \otimes \overline{\Delta_j \mathcal{Y}} \right\|_{L^1(\Omega;H\otimes H)} \lesssim N_L^{-\delta}\log(N_L)\|X\|_{L^2(\Omega;W)}\|Y\|_{L^2(\Omega;W)}.$$

Clearly, by (3.29), Definition 3.11 and Lemma 3.13

$$\begin{aligned}I_2 &= \sum_{k_1+j_1\le L}\sum_{k_2+j_2\le L}\Big\{\mathbb{E}[\langle \mathrm{Cov}_M\left[\overline{\Delta_{k_1}\mathcal{X}},\overline{\Delta_{j_1}\mathcal{Y}}\right], \mathrm{Cov}_M\left[\overline{\Delta_{k_2}\mathcal{X}},\overline{\Delta_{j_2}\mathcal{Y}}\right]\rangle_{H\otimes H}]\\ &\qquad - \langle \mathbb{C}\mathrm{ov}\left[\overline{\Delta_{k_1}\mathcal{X}},\overline{\Delta_{j_1}\mathcal{Y}}\right], \mathbb{C}\mathrm{ov}\left[\overline{\Delta_{k_2}\mathcal{X}},\overline{\Delta_{j_2}\mathcal{Y}}\right]\rangle_{H\otimes H}\Big\}\\ &= \sum_{k_1+j_1\le L}\sum_{k_2+j_2\le L}\Big\{\frac{1}{M}\mathcal{C}[\overline{\Delta_{k_1}\mathcal{X}}\otimes\overline{\Delta_{j_1}\mathcal{Y}}, \overline{\Delta_{k_2}\mathcal{X}}\otimes\overline{\Delta_{j_2}\mathcal{Y}}]\\ &\qquad + \frac{1}{M(M-1)}F(\overline{\Delta_{k_1}\mathcal{X}},\overline{\Delta_{j_1}\mathcal{Y}},\overline{\Delta_{k_2}\mathcal{X}},\overline{\Delta_{j_2}\mathcal{Y}})\Big\}\\ &\equiv J_1 + J_2.\end{aligned}$$

Rearranging J_1, we find by virtue of the triangle inequality, Corollary 3.6 and the stronger norm on W that

$$\begin{aligned}J_1 =& \frac{1}{M}\mathcal{C}[\hat{P}_L(\overline{\mathcal{X}},\overline{\mathcal{Y}}),\hat{P}_L(\overline{\mathcal{X}},\overline{\mathcal{Y}})] = \frac{1}{M}\mathcal{V}(\hat{P}_L(\overline{\mathcal{X}},\overline{\mathcal{Y}}))\\ \le& \frac{2}{M}\left(\|\overline{X}\otimes\overline{Y}\|^2_{L^2(\Omega;H\otimes H)} + \|\overline{X}\otimes\overline{Y} - \hat{P}_L(\overline{\mathcal{X}},\overline{\mathcal{Y}})\|^2_{L^2(\Omega;H\otimes H)}\right)\\ \le& \frac{2}{M}\left(\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)} + cN_L^{-2\delta}\log(N_L)^2\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}\right)\\ \lesssim& \frac{1}{M}\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}.\end{aligned}$$

Proceeding, we note that by (3.31) the higher order term J_2 splits into two summands, i.e. $J_2 = J_2^a + J_2^b$. Assuming independence of the sequences $\mathcal{X}$ and $\tilde{\mathcal{Y}}$, rearranging and taking the sums into the inner product we find that

$$\begin{aligned}J_2^a =& \frac{1}{M(M-1)}\sum_{k_1+j_1\le L}\sum_{k_2+j_2\le L}\mathcal{C}[\overline{\Delta_{k_1}\mathcal{X}},\overline{\Delta_{k_2}\mathcal{X}}]\cdot\mathcal{C}[\overline{\Delta_{j_1}\tilde{\mathcal{Y}}},\overline{\Delta_{j_2}\tilde{\mathcal{Y}}}]\\ =& \frac{1}{M(M-1)}\mathbb{E}[\|\hat{P}_L(\overline{\mathcal{X}},\overline{\tilde{\mathcal{Y}}})\|^2_{H\otimes H}] = \frac{1}{M(M-1)}\mathcal{V}(\hat{P}_L(\overline{\mathcal{X}},\overline{\tilde{\mathcal{Y}}}))\\ \lesssim& \frac{1}{M(M-1)}\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}\end{aligned} \tag{3.45}$$

in analogy to J_1, where we have used again Lemma 3.5, Jensen's inequality and the triangle inequality. Similarly, assuming mutual independence of the sequences $\mathcal{X},\tilde{\mathcal{X}},\mathcal{Y}$

and $\tilde{\mathcal{Y}}$, we arrive at

$$\begin{aligned} J_2^b &= \frac{1}{M(M-1)}\mathbb{E}[\langle \hat{P}_L(\overline{\tilde{\mathcal{X}}}, \overline{\mathcal{Y}}), \hat{P}_L(\overline{\mathcal{X}}, \overline{\tilde{\mathcal{Y}}})\rangle_{H\otimes H}] \\ &\leq \frac{1}{M(M-1)}(\mathbb{E}[\|\hat{P}_L(\overline{\mathcal{X}}, \overline{\tilde{\mathcal{Y}}})\|^2_{H\otimes H}])^{1/2}(\mathbb{E}[\|\hat{P}_L(\overline{\tilde{\mathcal{X}}}, \overline{\mathcal{Y}})\|^2_{H\otimes H}])^{1/2} \\ &\lesssim \frac{1}{M(M-1)}\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}, \end{aligned} \tag{3.46}$$

where after rearranging and taking the sums into the inner product we have used symmetry of the scalar product on H and the Cauchy-Schwarz inequality.

By collecting the previous estimates we deduce the following asymptotic bound for the mean square error

$$\mathrm{MSE} \lesssim N_L^{-2\delta}(\log(N_L))^2\|X\|^2_{L^2(\Omega;W)}\|Y\|^2_{L^2(\Omega;W)} + \frac{M+1}{M(M-1)}\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}.$$

For the optimal balancing of the MSE we note, since $(M+1)/(M-1) \searrow 1$, that asymptotically

$$\mathrm{MSE} \sim N_L^{-2\delta}(\log(N_L))^2 + \frac{1}{M} \sim \varepsilon^2,$$

whence $M \sim \varepsilon^{-2}$. Further, since for L large enough $N_L^{-\delta}\log(N_L) \sim \varepsilon$ implies $\log(N_L) \sim |\log(\varepsilon)|$, we find $N_L^{-\delta} \sim \varepsilon/|\log(\varepsilon)|$. Since we need $O(N_L\log(N_L))$ coefficients, our memory requirements are also bounded by a constant factor times $N_L\log(N_L)$. For the total cost, considering that it takes $O(N_L\log(N_L))$ work to build the sparse tensor approximant and $O(N_L^\gamma)$ work to compute the samples, we conclude

$$\begin{aligned} \mathrm{Cost}(\widehat{\mathrm{Cov}}_M[X,Y]) &\lesssim M(N_L^\gamma + N_L\log(N_L)) \sim \varepsilon^{-2}\cdot\begin{cases}\varepsilon N_L^{\delta+1}, & \gamma = 1,\\ N_L^\gamma, & \gamma > 1,\end{cases} \\ &\sim \varepsilon^{-2}\begin{cases}\varepsilon\left(\frac{\varepsilon}{|\log(\varepsilon)|}\right)^{-\frac{\delta+1}{\delta}}, & \gamma = 1,\\ \left(\frac{\varepsilon}{|\log(\varepsilon)|}\right)^{-\frac{\gamma}{\delta}}, & \gamma > 1,\end{cases} \\ &\sim \varepsilon^{-2-\gamma/\delta}\cdot\begin{cases}|\log(\varepsilon)|^{1+1/\delta}, & \gamma = 1,\\ |\log(\varepsilon)|^{\frac{\gamma}{\delta}}, & \gamma > 1.\end{cases} \end{aligned}$$

□

Remark 3.21. *To see that the construction of the sparse tensor product covariance estimator* $\widehat{\mathrm{Cov}}_M[\cdot,\cdot]$ *on discretisation level* L *takes* $O(N_L\log(N_L))$ *work, we consider* (3.44) *and* (3.9) *and write*

$$\begin{aligned} \widehat{\mathrm{Cov}}_M[X_L,Y_L] &:= \sum_{k+j\leq L}\mathrm{Cov}_M[\Delta_k\mathcal{X}, \Delta_j\mathcal{Y}] = \sum_{k=0}^{L}\mathrm{Cov}_M\left[\Delta_k\mathcal{X}, \sum_{j=0}^{L-k}\Delta_j\mathcal{Y}\right] \\ &= \sum_{k=0}^{L}\mathrm{Cov}_M[\Delta_k\mathcal{X}, Y_{L-k}] \\ &= \sum_{k=0}^{L}\mathrm{Cov}_M[X_k, Y_{L-k}] - \sum_{k=0}^{L-1}\mathrm{Cov}_M[X_k, Y_{L-k-1}]. \end{aligned}$$

Then by (3.13) we find the asserted bound for the work of constructing the sparse tensor product covariance estimator. This representation is similar to that of the combination technique of sparse tensor product approximation (cf. [34, 36]).

3.3.4 Multilevel Monte Carlo approximation of covariance functions

Another approach of improving the full tensor product approximation of covariance functions is realized by the multilevel variant of the full tensor product single level approximation Monte Carlo method.

As before we shall firstly consider full tensor product approximations and then turn to sparse tensor product approximations.

Definition 3.22. *Let $X, Y \in L^2(\Omega; H)$ and $\mathcal{X}, \mathcal{Y}$ sequences converging to X and Y, respectively. The full tensor product covariance estimator $\mathrm{Cov}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]$ at level L for the approximation of $\mathbb{C}\mathrm{ov}[X, Y]$ is then defined as*

$$\begin{aligned}\mathrm{Cov}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}] &:= \sum_{\ell=0}^{L} \mathrm{Cov}_{M_\ell}[X_\ell, Y_\ell] - \mathrm{Cov}_{M_\ell}[X_{\ell-1}, Y_{\ell-1}] \\ &= \sum_{\ell=0}^{L} C_{\ell;M_\ell} - C_{\ell-1;M_\ell}.\end{aligned} \tag{3.47}$$

Theorem 3.23. *Let $\{N_\ell\}$ be an approximately exponentially increasing sequence with respect to some $a > 1$, such that (3.4),(3.12) and (3.13) hold. Furthermore, assume there exist $\gamma, \delta > 0$, such that*

$$\begin{cases} \|X - X_L\|_H \lesssim N_L^{-\delta} \|X\|_W, \\ \|Y - Y_L\|_H \lesssim N_L^{-\delta} \|Y\|_W, \\ \mathrm{Cost}(X_L) \lesssim N_L^{\gamma} \text{ and } \mathrm{Cost}(Y_L) \lesssim N_L^{\gamma}. \end{cases}$$

Suppose that the evaluation of a sample of the level correction $\Upsilon_\ell = C_{\ell;M_\ell} - C_{\ell-1;M_\ell}$ satisfies $C_\ell := \mathrm{Cost}(\Upsilon_\ell) \lesssim N_\ell^{\max\{\gamma,2\}}$. Under assumptions of Lemma 3.16 and letting for brevity

$$C^{\mathrm{ML}} := \mathrm{Cov}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}],$$

then for any accuracy $\varepsilon > 0$

$$\|C^{\mathrm{ML}} - \mathbb{C}\|_{L^2(\Omega, H\otimes H)} < \varepsilon$$

can be achieved at the computational cost

$$\mathrm{Cost}(C^{\mathrm{ML}}) \lesssim \begin{cases} \varepsilon^{-2}, & 2\delta > \max\{\gamma, 2\}, \\ \varepsilon^{-2} |\log(\varepsilon)|^2, & 2\delta = \max\{\gamma, 2\}, \\ \varepsilon^{-\frac{\max\{\gamma,2\}}{\delta}}, & 2\delta < \max\{\gamma, 2\}, \end{cases}$$

and memory requirements

$$\mathrm{Memory}(C^{\mathrm{ML}}) \sim N_L^2.$$

Proof. Using notation as in the full tensor single level case, analogously to (3.40), we note

$$\|C^{\mathrm{ML}} - \mathbb{C}\|^2_{L^2(\Omega; H\otimes H)} = \|\mathbb{C}_L - \mathbb{C}\|^2_{H\otimes H} + \|C^{\mathrm{ML}} - \mathbb{C}_L\|^2_{L^2(\Omega; H\otimes H)} \tag{3.48}$$

and $\|\mathbb{C}_L - \mathbb{C}\|^2_{H\otimes H} \lesssim N_L^{-2\delta}$. For the second summand there holds

$$\|C^{\mathrm{ML}} - \mathbb{C}_L\|^2_{L^2(\Omega;H\otimes H)} = \|C^{\mathrm{ML}}\|^2_{L^2(\Omega;H\otimes H)} - \|\mathbb{C}_L\|^2_{H\otimes H}.$$

Moreover, we observe for $k \neq \ell$ that $C_{k;M_k}$ and $C_{\ell;M_\ell}$ are built using independent sequences of realizations $\{X_k\}, \{Y_k\}, \{X_\ell\}$ and $\{Y_\ell\}$. Let us denote for brevity

$$C_{\ell;M_\ell} - C_{\ell-1;M_\ell} = \sum_{k\vee j=\ell} \mathrm{Cov}_{M_\ell}\left[\Delta_k\mathcal{X}, \Delta_j\mathcal{Y}\right] =: \sum_{k,j}{}' \mathrm{Cov}_{M_\ell}\left[\Delta_k\mathcal{X}, \Delta_j\mathcal{Y}\right].$$

Hence, rearranging the terms, taking advantage of indepedence and Lemma 3.13 we get

$$\begin{aligned}
&\|C^{\mathrm{ML}}\|^2_{L^2(\Omega;H\otimes H)} - \|\mathbb{C}_L\|^2_{H\otimes H} \\
&= \sum_{\ell=0}^{L} \sum_{k,j}{}' \sum_{k',j'}{}' \Big\{ \mathbb{E}[\langle \mathrm{Cov}_{M_\ell}\left[\Delta_k\overline{\mathcal{X}}, \Delta_j\overline{\mathcal{Y}}\right], \mathrm{Cov}_{M_\ell}\left[\Delta_{k'}\overline{\mathcal{X}}, \Delta_{j'}\overline{\mathcal{Y}}\right]\rangle_{H\otimes H}] \\
&\qquad\qquad - \langle \mathbb{C}\mathrm{ov}\left[\Delta_k\overline{\mathcal{X}}, \Delta_j\overline{\mathcal{Y}}\right], \mathbb{C}\mathrm{ov}\left[\Delta_{k'}\overline{\mathcal{X}}, \Delta_{j'}\overline{\mathcal{Y}}\right]\rangle_{H\otimes H} \Big\} \\
&= \sum_{\ell=0}^{L} \sum_{k,j}{}' \sum_{k',j'}{}' \Big\{ \frac{1}{M_\ell}\mathcal{C}[\Delta_k\overline{\mathcal{X}}\otimes\Delta_j\overline{\mathcal{Y}}, \Delta_{k'}\overline{\mathcal{X}}\otimes\Delta_{j'}\overline{\mathcal{Y}}] \\
&\qquad\qquad + \frac{1}{M_\ell(M_\ell-1)} F(\Delta_k\overline{\mathcal{X}}, \Delta_j\overline{\mathcal{Y}}, \Delta_{k'}\overline{\mathcal{X}}, \Delta_{j'}\overline{\mathcal{Y}}) \Big\} \\
&= \sum_{\ell=0}^{L} (J_{1,\ell} + J_{2,\ell}) \equiv J_1 + J_2.
\end{aligned}$$

Let $Q_\ell = \overline{X}_\ell \otimes \overline{Y}_\ell - \overline{X}_{\ell-1} \otimes \overline{Y}_{\ell-1}$. Then with $Q_{-1} \equiv 0$

$$J_1 = \sum_{\ell=0}^{L} J_{1,\ell} = \sum_{\ell=0}^{L} \frac{1}{M_\ell}\mathcal{C}[Q_\ell, Q_\ell] = \sum_{\ell=0}^{L} \frac{1}{M_\ell}\mathcal{V}(Q_\ell).$$

Now, since $\mathcal{V}(Q_\ell) = \mathbb{E}\left[\|Q_\ell - \mathbb{E}[Q_\ell]\|^2_{H\otimes H}\right] \leq \|Q_\ell\|^2_{L^2(\Omega;H\otimes H)}$, and so for $\ell \geq 1$ there holds

$$\begin{aligned}
\|Q_\ell\|_{L^2(\Omega;H\otimes H)} &\leq \|\overline{X}_\ell \otimes \overline{Y}_\ell - \overline{X} \otimes \overline{Y}\|_{L^2(\Omega;H\otimes H)} + \|\overline{X} \otimes \overline{Y} - \overline{X}_{\ell-1} \otimes \overline{Y}_{\ell-1}\|_{L^2(\Omega;H\otimes H)} \\
&\leq cN_\ell^{-\delta}(1 + (aR^2)^\delta)\|X\|_{L^4(\Omega;W)}\|Y\|_{L^4(\Omega;W)}.
\end{aligned}$$

Obviously, for $\ell = 0$ there holds

$$\begin{aligned}
\|Q_0\|_{L^2(\Omega;H\otimes H)} &\leq \|\overline{X} \otimes \overline{Y}\|_{L^2(\Omega;H\otimes H)} + \|\overline{X_0} \otimes \overline{Y_0} - \overline{X} \otimes \overline{Y}\|_{L^2(\Omega;H\otimes H)} \\
&\leq C(1 + N_0^{-\delta})\|X\|_{L^4(\Omega;W)}\|Y\|_{L^4(\Omega;W)}
\end{aligned}$$

and hence,

$$J_1 \leq C\frac{1}{M_0}\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)} + \tilde{C}\sum_{\ell=0}^{L} \frac{1}{M_\ell}N_\ell^{-2\delta}\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}.$$

For the higher order terms we proceed as follows. Let $J_2 = J_2^a + J_2^b$ and estimate both terms separately. Then for J_2^a by (3.34) we find by assuming independence of the se-

quences $\mathcal{X}$ and $\tilde{\mathcal{Y}}$

$$\begin{aligned} J_2^a &= \sum_{\ell=0}^{L} \sum_{k,j}{}' \sum_{k',j'}{}' \frac{1}{M_\ell(M_\ell-1)} \mathcal{C}[\Delta_k \overline{\mathcal{X}}, \Delta_{k'} \overline{\mathcal{X}}] \cdot \mathcal{C}[\Delta_j \overline{\tilde{\mathcal{Y}}}, \Delta_{j'} \overline{\tilde{\mathcal{Y}}}] \\ &= \sum_{\ell=0}^{L} \sum_{k,j}{}' \sum_{k',j'}{}' \frac{1}{M_\ell(M_\ell-1)} \mathcal{C}[\Delta_k \overline{\mathcal{X}} \otimes \Delta_j \overline{\tilde{\mathcal{Y}}}, \Delta_{k'} \overline{\mathcal{X}} \otimes \Delta_{j'} \overline{\tilde{\mathcal{Y}}}] \\ &\lesssim \frac{1}{M_0(M_0-1)} \|X\|^2_{L^4(\Omega;W)} \|Y\|^2_{L^4(\Omega;W)} + \sum_{\ell=0}^{L} \frac{N_\ell^{-2\delta}}{M_\ell(M_\ell-1)} \|X\|^2_{L^4(\Omega;W)} \|Y\|^2_{L^4(\Omega;W)}. \end{aligned}$$

Similarly, for J_2^b by (3.35), by assuming independence of the sequences $\mathcal{X}, \tilde{\mathcal{X}}, \mathcal{Y}$ and $\tilde{\mathcal{Y}}$ in addition to the Cauchy-Schwarz inequality and the symmetry of the inner product on H there holds

$$\begin{aligned} J_2^b &= \sum_{\ell=0}^{L} \sum_{k,j}{}' \sum_{k',j'}{}' \frac{1}{M_\ell(M_\ell-1)} \langle \mathbb{C}\mathrm{ov}\left[\Delta_k \overline{\mathcal{X}}, \Delta_{j'} \overline{\mathcal{Y}}\right], \mathbb{C}\mathrm{ov}\left[\Delta_{k'} \overline{\tilde{\mathcal{X}}}, \Delta_j \overline{\tilde{\mathcal{Y}}}\right] \rangle_{H \otimes H} \\ &= \sum_{\ell=0}^{L} \sum_{k,j}{}' \sum_{k',j'}{}' \frac{1}{M_\ell(M_\ell-1)} \mathbb{E}\left[\langle \Delta_k \overline{\mathcal{X}} \otimes \Delta_j \overline{\tilde{\mathcal{Y}}}, \Delta_{k'} \overline{\tilde{\mathcal{X}}} \otimes \Delta_{j'} \overline{\mathcal{Y}} \rangle_{H \otimes H}\right] \\ &\lesssim \frac{1}{M_0(M_0-1)} \|X\|^2_{L^4(\Omega;W)} \|Y\|^2_{L^4(\Omega;W)} + \sum_{\ell=0}^{L} \frac{N_\ell^{-2\delta}}{M_\ell(M_\ell-1)} \|X\|^2_{L^4(\Omega;W)} \|Y\|^2_{L^4(\Omega;W)}. \end{aligned}$$

Collecting the estimates for J_1, J_2^a and J_2^b, and noting that $(M_\ell+1)/(M_\ell-1) \to 1$ from above shows that the mean square error admits the asymptotic bound

$$\mathrm{MSE} \lesssim N_L^{-2\delta} + M_0^{-1} + \sum_{\ell=0}^{L} M_\ell^{-1} N_\ell^{-2\delta}.$$

Here, we have $\varepsilon \sim N_L^{-\delta}$ and so the optimal choice of samples for balancing cost versus accuracy, which we compute using a Lagrange multiplier method with respect to the cost functional $F(M_0, M_1, \dots, M_L) = \sum_{\ell=0}^{L} C_\ell M_\ell$ with a fixed variance at $\varepsilon^2 \overset{!}{=} \sum_{\ell=0}^{L} M_\ell^{-1} N_\ell^{-2\delta}$, is given by

$$M_k \sim N_k^{-\delta} (N_k^\gamma + N_k^2)^{-1/2} \cdot \begin{cases} N_L^{2\delta}, & 2\delta > \max\{\gamma, 2\}, \\ L N_L^{2\delta}, & 2\delta = \max\{\gamma, 2\}, \\ (N_L^\gamma + N_L^2) N_L^\delta, & 2\delta < \max\{\gamma, 2\}. \end{cases} \tag{3.49}$$

Since on each level we approximate the covariance using N_ℓ^2 coefficients, the memory

requirements are bounded by $O(N_L^2)$. For the total cost we find

$$\begin{aligned}
&\mathrm{Cost}(\mathrm{C}^{\mathrm{ML}}) \sim \sum_{\ell=0}^{L} C_\ell M_\ell \lesssim \sum_{\ell=0}^{L} C_\ell \lceil M_\ell \rceil \lesssim \sum_{\ell=0}^{L} C_\ell M_\ell + \sum_{\ell=0}^{L} C_\ell \\
&\lesssim \sum_{\ell=0}^{L} (N_\ell^\gamma + N_\ell^2) \cdot \left(N_L^{2\delta} N_\ell^{-\delta} (N_\ell^\gamma + N_\ell^2)^{-1/2} \left(\sum_{k=0}^{L} (N_k^\gamma + N_k^2)^{1/2} N_k^{-\delta} \right) \right) + N_L^{\max\{\gamma,2\}} \\
&\sim N_L^{2\delta} \left(\sum_{\ell=0}^{L} (N_\ell^\gamma + N_\ell^2)^{1/2} N_\ell^{-\delta} \right)^2 + N_L^{\max\{\gamma,2\}} \\
&\lesssim \left\{ \begin{matrix} N_L^{2\delta}, & 2\delta > \max\{\gamma,2\} \\ L^2 N_L^{2\delta}, & 2\delta = \max\{\gamma,2\} \\ N_L^{\max\{\gamma,2\}}, & 2\delta < \max\{\gamma,2\} \end{matrix} \right\} + N_L^{\max\{\gamma,2\}} \\
&\lesssim \left\{ \begin{matrix} \varepsilon^{-2}, & 2\delta > \max\{\gamma,2\} \\ \varepsilon^{-2} |\log(\varepsilon)|^2, & 2\delta = \max\{\gamma,2\} \\ \varepsilon^{-\frac{\max\{\gamma,2\}}{\delta}}, & 2\delta < \max\{\gamma,2\} \end{matrix} \right\} + \varepsilon^{-\frac{\max\{\gamma,2\}}{\delta}},
\end{aligned}$$

which upon noting the cases yields the claim. □

Definition 3.24. *Let $X, Y \in L^2(\Omega; H)$ and $\mathcal{X}, \mathcal{Y}$ sequences converging to X and Y, respectively. The sparse tensor product covariance estimator $\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]$ at level L for the approximation of $\mathbb{C}\mathrm{ov}[X, Y]$ is then defined as*

$$\begin{aligned}
\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}] &:= \sum_{\ell=0}^{L} \widehat{\mathrm{Cov}}_{M_\ell}[X_\ell, Y_\ell] - \widehat{\mathrm{Cov}}_{M_\ell}[X_{\ell-1}, Y_{\ell-1}] \\
&= \sum_{\ell=0}^{L} \hat{C}_{\ell;M_\ell} - \hat{C}_{\ell-1;M_\ell}.
\end{aligned} \tag{3.50}$$

We now turn our attention to the analysis of the sparse tensor product multilevel covariance estimator $\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]$.

Theorem 3.25. *Let $\{N_\ell\}$ be an approximately exponentially increasing sequence with respect to some $a > 1$, such that (3.4), (3.12) and (3.13) hold. Let $C_\ell := \mathrm{Cost}(\Upsilon_\ell)$ denote the cost of one level correction on level ℓ, where the level corrections are given by*

$$\Upsilon_\ell = \widehat{\mathrm{Cov}}_{M_\ell}[X_\ell, Y_\ell] - \widehat{\mathrm{Cov}}_{M_\ell}[X_{\ell-1}, Y_{\ell-1}] =: \hat{C}_{\ell;M_\ell} - \hat{C}_{\ell-1;M_\ell}.$$

Assume there exist $\delta > 0$ and $\gamma \geq 1$, such that

$$\left\{ \begin{matrix} \|X - X_\ell\|_H \lesssim N_\ell^{-\delta} \|X\|_W, \\ \|Y - Y_\ell\|_H \lesssim N_\ell^{-\delta} \|Y\|_W, \\ C_\ell \lesssim N_\ell^\gamma + N_\ell \log(N_\ell), \end{matrix} \right.$$

where $X, Y \in L^4(\Omega; H)$. Then for any $\varepsilon > 0$ and

$$\|\mathbb{C}\mathrm{ov}[X, Y] - \widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]\|_{L^2(\Omega; H \otimes H)} < \varepsilon,$$

there exist $L \in \mathbb{N}$ *and a sequence* $M_0, \dots, M_L \in \mathbb{N}$*, such that for* $\gamma = 1$

$$\mathrm{Cost}(\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]) \lesssim \begin{cases} \varepsilon^{-2}, & \delta > 1/2, \\ \varepsilon^{-2}|\log(\varepsilon)|^5, & \delta = 1/2, \\ \varepsilon^{-1/\delta}|\log(\varepsilon)|^{1+1/\delta}, & \delta < 1/2, \end{cases} \tag{3.51}$$

and for $\gamma > 1$

$$\mathrm{Cost}(\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]) \lesssim \begin{cases} \varepsilon^{-2}, & \delta > \gamma/2, \\ \varepsilon^{-2}|\log(\varepsilon)|^4, & \delta = \gamma/2, \\ \varepsilon^{-\gamma/\delta}|\log(\varepsilon)|^{\gamma/\delta}, & \delta < \gamma/2, \end{cases} \tag{3.52}$$

and memory

$$\mathrm{Memory}(\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]) \sim N_L \log(N_L).$$

Proof. W.l.o.g. let $\mathbb{E}[X] = 0 = \mathbb{E}[Y]$. With notation similar to the single level case, namely $\mu = \mathbb{C}\mathrm{ov}[X, Y]$, $m_L = \widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]$ and $\mu_L = \mathbb{E}[m_L]$, we have

$$\mathrm{MSE} = \|\mu - \mu_L\|^2_{L^2(\Omega; H\otimes H)} + \|\mu_L - m_L\|^2_{L^2(\Omega; H\otimes H)} \equiv I_1 + I_2.$$

As in the single level case we note, since $\mathbb{E}[\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X}, \mathcal{Y}]] = \mathbb{E}[\hat{P}_L(\overline{\mathcal{X}}, \overline{\mathcal{Y}})]$, that

$$\sqrt{I_1} \le \left\| \sum\nolimits_{k+j>L} \overline{\Delta_k \mathcal{X}} \otimes \overline{\Delta_j \mathcal{Y}} \right\|_{L^1(\Omega; H\otimes H)} \lesssim N_L^{-\delta} \log(N_L) \|X\|_{L^2(\Omega;W)} \|Y\|_{L^2(\Omega;W)}.$$

Moreover, we observe for $k \neq \ell$ that $\hat{P}_k(\mathcal{X}, \mathcal{Y}), \hat{P}_{k-1}(\mathcal{X}, \mathcal{Y}), \hat{P}_\ell(\mathcal{X}, \mathcal{Y}), \hat{P}_{\ell-1}(\mathcal{X}, \mathcal{Y})$ are built using independent sequences of realizations of sequences $\{X_k\}, \{Y_k\}, \{X_\ell\}$ and $\{Y_\ell\}$. Hence, rearranging the terms in I_2, taking advantage of indepedence and Lemma 3.13 there holds

$$\begin{aligned} I_2 &= \sum_{\ell=0}^{L} \sum_{m_1+n_1=\ell} \sum_{m_2+n_2=\ell} \Big\{ \mathbb{E}[\langle \mathrm{Cov}_{M_\ell}\left[\overline{\Delta_{m_1}\mathcal{X}}, \overline{\Delta_{n_1}\mathcal{Y}}\right], \mathrm{Cov}_{M_\ell}\left[\overline{\Delta_{m_2}\mathcal{X}}, \overline{\Delta_{n_2}\mathcal{Y}}\right]\rangle_{H\otimes H}] \\ &\qquad - \langle \mathbb{C}\mathrm{ov}\left[\overline{\Delta_{m_1}\mathcal{X}}, \overline{\Delta_{n_1}\mathcal{Y}}\right], \mathbb{C}\mathrm{ov}\left[\overline{\Delta_{m_2}\mathcal{X}}, \overline{\Delta_{n_2}\mathcal{Y}}\right]\rangle_{H\otimes H} \Big\} \\ &= \sum_{\ell=0}^{L} \sum_{m_1+n_1=\ell} \sum_{m_2+n_2=\ell} \Big\{ \frac{1}{M_\ell} \mathcal{C}[\overline{\Delta_{m_1}\mathcal{X}} \otimes \overline{\Delta_{n_1}\mathcal{Y}}, \overline{\Delta_{m_2}\mathcal{X}} \otimes \overline{\Delta_{n_2}\mathcal{Y}}] \\ &\qquad + \frac{1}{M_\ell(M_\ell - 1)} F(\overline{\Delta_{m_1}\mathcal{X}}, \overline{\Delta_{n_1}\mathcal{Y}}, \overline{\Delta_{m_2}\mathcal{X}}, \overline{\Delta_{n_2}\mathcal{Y}}) \Big\} \\ &= \sum_{\ell=0}^{L} (J_{1,\ell} + J_{2,\ell}) \equiv J_1 + J_2. \end{aligned}$$

Letting $\zeta_\ell = \hat{P}_\ell(\overline{\mathcal{X}}, \overline{\mathcal{Y}}) - \hat{P}_{\ell-1}(\overline{\mathcal{X}}, \overline{\mathcal{Y}})$, $\ell \ge 0$, with $\hat{P}_{-1} \equiv 0$ we obtain by taking the sums inside

$$J_1 = \sum_{\ell=0}^{L} J_{1,\ell} = \sum_{\ell=0}^{L} \frac{1}{M_\ell} \mathcal{C}[\zeta_\ell, \zeta_\ell] = \sum_{\ell=0}^{L} \frac{1}{M_\ell} \mathcal{V}(\zeta_\ell).$$

Now, since $\mathcal{V}(\zeta_\ell)^{1/2} = (\mathbb{E}[\|\zeta_\ell - \mathbb{E}[\zeta_\ell]\|^2_{H\otimes H}])^{1/2} \leq \|\zeta_\ell\|_{L^2(\Omega;H\otimes H)}$, for $\ell > 1$ we have by the triangle inequality and Corollary 3.6

$$\begin{aligned}\|\zeta_\ell\|_{L^2(\Omega;H\otimes H)} &\leq \|\overline{X}\otimes\overline{Y} - \hat{P}_\ell(\overline{\mathcal{X}},\overline{\mathcal{Y}})\|_{L^2(\Omega;H\otimes H)} + \|\overline{X}\otimes\overline{Y} - \hat{P}_{\ell-1}(\overline{\mathcal{X}},\overline{\mathcal{Y}})\|_{L^2(\Omega;H\otimes H)}\\ &\leq (cN_\ell^{-\delta}\log(N_\ell) + cN_{\ell-1}^{-\delta}\log(N_{\ell-1}))\|X\|_{L^4(\Omega;W)}\|Y\|_{L^4(\Omega;W)}\\ &\leq cN_\ell^{-\delta}\left(1+\left(\frac{N_\ell}{N_{\ell-1}}\right)^\delta\right)\log(N_\ell)\|X\|_{L^4(\Omega;W)}\|Y\|_{L^4(\Omega;W)}\\ &\leq \tilde{c}N_\ell^{-\delta}\log(N_\ell)\|X\|_{L^4(\Omega;W)}\|Y\|_{L^4(\Omega;W)}\end{aligned}$$

as well as $\|\zeta_0\|_{L^2(\Omega;H\otimes H)} \leq \tilde{c}(1+N_0^{-\delta}\log(N_0))\|X\|_{L^4(\Omega;W)}\|Y\|_{L^4(\Omega;W)}$. Thus, we obtain

$$J_1 \lesssim \left(\frac{1}{M_0} + \sum_{\ell=0}^{L}\frac{1}{M_\ell}N_\ell^{-2\delta}\log(N_\ell)^2\right)\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}$$

where the hidden constant depends only on δ, a and N_0. Dealing with J_2, firstly we note that analogously to the single level case that the higher order term J_2 splits into two parts, i.e. $J_2 = J_2^a + J_2^b$. By assuming independence of the sequences of realisations $\mathcal{X}$ and $\tilde{\mathcal{Y}}$, we observe for each summand in J_2^a and $\ell \geq 1$ as in the single level case, cf. (3.45), that

$$\begin{aligned}J_{2,\ell}^a &= \frac{1}{M_\ell(M_\ell-1)}\left(\sum_{m_1+n_1=\ell}\sum_{m_2+n_2=\ell}\mathcal{C}[\overline{\Delta_{m_1}\mathcal{X}},\overline{\Delta_{m_2}\mathcal{X}}]\cdot\mathcal{C}[\overline{\Delta_{n_1}\tilde{\mathcal{Y}}},\overline{\Delta_{n_2}\tilde{\mathcal{Y}}}]\right)\\ &\lesssim \frac{1}{M_\ell(M_\ell-1)}N_\ell^{-2\delta}\log(N_\ell)^2\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}\end{aligned}$$

Similarly, for $J_{2,\ell}^b$ we conclude by virtue of mutual independence of the sequences $\mathcal{X}, \tilde{\mathcal{X}}, \mathcal{Y}$ and $\tilde{\mathcal{Y}}$, the Cauchy-Schwarz inequality and the symmetry of the inner product on H that, cf. (3.46),

$$\begin{aligned}J_{2,\ell}^b &= \frac{1}{M_\ell(M_\ell-1)}\left(\sum_{m_1+n_1=\ell}\sum_{m_2+n_2=\ell}\langle\mathbb{C}\mathrm{ov}\left[\overline{\Delta_{m_1}\mathcal{X}},\overline{\Delta_{n_2}\mathcal{Y}}\right],\mathbb{C}\mathrm{ov}\left[\overline{\Delta_{m_2}\tilde{\mathcal{X}}},\overline{\Delta_{n_1}\tilde{\mathcal{Y}}}\right]\rangle_{H\otimes H}\right)\\ &\lesssim \frac{1}{M_\ell(M_\ell-1)}N_\ell^{-2\delta}\log(N_\ell)^2\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}\end{aligned}$$

Thus, noting that $J_{2,0}^a \lesssim (M_0+1)/(M_0(M_0+1))$ as well as $J_{2,0}^b \lesssim (M_0+1)/(M_0(M_0+1))$, summing over the estimates for $J_1, J_{2,\ell}^a$ and $J_{2,\ell}^b$ yields

$$\begin{aligned}\mathrm{MSE} &\lesssim N_L^{-2\delta}\log(N_L)^2\|X\|^2_{L^2(\Omega;W)}\|Y\|^2_{L^2(\Omega;W)}\\ &\quad + \left(\frac{M_0+1}{M_0(M_0-1)} + \sum_{\ell=0}^{L}\frac{M_\ell+1}{M_\ell(M_\ell-1)}N_\ell^{-2\delta}\log(N_\ell)^2\right)\|X\|^2_{L^4(\Omega;W)}\|Y\|^2_{L^4(\Omega;W)}.\end{aligned}$$

Now, for any $\varepsilon > 0$ and $\|\mu - \mu_L\|_{H\otimes H} \lesssim \varepsilon$ there exists L large enough with $N_L^{-\delta}\log(N_L) \sim \varepsilon$. The cost C_ℓ on each level ℓ is proportional to $N_\ell^\gamma + N_\ell\log(N_\ell)$, since we have to compute samples $X_0^i, \ldots, X_\ell^i$ and $Y_0^i, \ldots, Y_\ell^i$, respectively, which accounts for $O(N_\ell^\gamma)$ work and assembling the sparse tensor approximant requires $O(N_\ell\log(N_\ell))$ work. We distinguish the two cases $\gamma = 1$ and $\gamma > 1$, but we choose the number of samples on the coarsest level as $M_0 = L^{-2}N_L^{2\delta}$ in either case. Moreover, since the fraction $(M_\ell+1)/(M_\ell-1) \leq 3$ for

$M_\ell \geq 2$ and converges to 1 from above, the factor can be neglected asymptotically. Then for $\gamma = 1$ selecting M_ℓ for $\ell = 1, \dots, L$ as

$$M_\ell \sim N_\ell^{-(\delta+\frac{1}{2})} \log(N_\ell)^{1/2} \cdot \begin{cases} L^{-2} N_L^{2\delta}, & \delta > 1/2, \\ L^{1/2} N_L^{2\delta}, & \delta = 1/2, \\ L^{-1/2} N_L^{\delta+\frac{1}{2}}, & \delta < 1/2, \end{cases} \tag{3.53}$$

we find by a straightforward computation for the MSE that

$$\begin{aligned} \sum_{\ell=1}^{L} \frac{1}{M_\ell} N_\ell^{-2\delta} \log(N_\ell)^2 &\sim \left(\sum_{\ell=1}^{L} \ell^{3/2} N_\ell^{-\delta+1/2} \right) \cdot \begin{Bmatrix} L^2 N_L^{-2\delta}, & \delta > 1/2 \\ L^{1/2} N_L^{-2\delta}, & \delta = 1/2 \\ L^{-1/2} N_L^{-(\delta+1/2)}, & \delta < 1/2 \end{Bmatrix} \\ &\lesssim \begin{Bmatrix} 1, & \delta > 1/2 \\ L^{5/2}, & \delta = 1/2 \\ L^{3/2} N_L^{-(\delta-1/2)}, & \delta < 1/2 \end{Bmatrix} \cdot \begin{Bmatrix} L^2 N_L^{-2\delta}, & \delta > 1/2 \\ L^{-1/2} N_L^{-2\delta}, & \delta = 1/2 \\ L^{1/2} N_L^{-(\delta+1/2)}, & \delta < 1/2 \end{Bmatrix} \\ &\sim \begin{Bmatrix} L^2 N_L^{-2\delta}, & \delta > 1/2 \\ L^2 N_L^{-2\delta}, & \delta = 1/2 \\ L^2 N_L^{-2\delta}, & \delta < 1/2 \end{Bmatrix} \sim \varepsilon^2 \end{aligned}$$

In the case of a non-optimal solver, i.e. $\gamma > 1$, selecting M_ℓ for $\ell = 1, \dots, L$ as

$$M_\ell \sim N_\ell^{-(\frac{2\delta+\gamma}{2})} \log(N_\ell) \cdot \begin{cases} L^{-2} N_L^{2\delta}, & \delta > \gamma/2, \\ N_L^{2\delta}, & \delta = \gamma/2, \\ L^{-1} N_L^{\frac{2\delta+\gamma}{2}}, & \delta < \gamma/2, \end{cases} \tag{3.54}$$

leads to the desired balancing, since

$$\begin{aligned} \sum_{\ell=1}^{L} \frac{1}{M_\ell} N_\ell^{-2\delta} \log(N_\ell)^2 &\sim \left(\sum_{\ell=1}^{L} \ell N_\ell^{\frac{\gamma-2\delta}{2}} \right) \cdot \begin{Bmatrix} L^2 N_L^{-2\delta}, & \delta > \gamma/2 \\ N_L^{-2\delta}, & \delta = \gamma/2 \\ N_L^{-\frac{\gamma+2\delta}{2}}, & \delta < \gamma/2 \end{Bmatrix} \\ &\lesssim \begin{Bmatrix} 1, & \delta > \gamma/2 \\ L^2, & \delta = \gamma/2 \\ L N_L^{\frac{\gamma-2\delta}{2}}, & \delta < \gamma/2 \end{Bmatrix} \cdot \begin{Bmatrix} L^2 N_L^{-2\delta}, & \delta > \gamma/2 \\ N_L^{-2\delta}, & \delta = \gamma/2 \\ L N_L^{-\frac{\gamma+2\delta}{2}}, & \delta < \gamma/2 \end{Bmatrix} \\ &\sim \begin{Bmatrix} L^2 N_L^{-2\delta}, & \delta > \gamma/2 \\ L^2 N_L^{-2\delta}, & \delta = \gamma/2 \\ L^2 N_L^{-2\delta}, & \delta < \gamma/2 \end{Bmatrix} \sim \varepsilon^2. \end{aligned}$$

Finally, we consider the work ensued in both cases. First off we note that the cost of a sample on the coarsest grid is constant, i.e. $C_0 \sim N_0^\gamma + N_0 \log(N_0)$, and so $C_0 M_0 \lesssim L^{-2} N_L^{2\delta}$ in any case. We start with the case $\gamma = 1$ selecting M_ℓ as above to find by assumption on the cost C_ℓ that

$$\begin{aligned} \sum_{\ell=0}^{L} C_\ell M_\ell &\lesssim \sum_{\ell=0}^{L} N_\ell \log(N_\ell) \cdot \left(N_\ell^{-\left(\delta+\frac{1}{2}\right)} \log(N_\ell)^{1/2} \cdot \{\cdot\}_L \right) + N_L \log(N_L) \\ &\lesssim \begin{Bmatrix} L^{-2} N_L^{2\delta}, & \delta > 1/2, \\ L^3 N_L^{2\delta}, & \delta = 1/2, \\ L N_L, & \delta < 1/2. \end{Bmatrix} + N_L \log(N_L) \end{aligned}$$

Thus, considering that $N_L \sim \varepsilon^{-1/\delta}|\log(\varepsilon)|^{1/\delta}$ and $L \sim \log(N_L) \sim |\log(\varepsilon)|$ the cost for $\gamma = 1$ is given by

$$\mathrm{Cost}(\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X},\mathcal{Y}]) \lesssim \begin{Bmatrix} \varepsilon^{-2}, & \delta > 1/2, \\ \varepsilon^{-2}|\log(\varepsilon)|^5, & \delta = 1/2, \\ \varepsilon^{-1/\delta}|\log(\varepsilon)|^{1+1/\delta}, & \delta < 1/2. \end{Bmatrix} + \varepsilon^{-\frac{1}{\delta}}|\log(\varepsilon)|^{1+\frac{1}{\delta}}.$$

In the case that $\gamma > 1$, we find

$$\sum_{\ell=0}^{L} C_\ell M_\ell \lesssim \begin{Bmatrix} L^{-2}N_L^{2\delta}, & \delta > \gamma/2, \\ L^2 N_L^{2\delta}, & \delta = \gamma/2, \\ N_L^\gamma, & \delta < \gamma/2. \end{Bmatrix} + N_L^\gamma$$

and hence

$$\mathrm{Cost}(\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{X},\mathcal{Y}]) \lesssim \begin{Bmatrix} \varepsilon^{-2}, & \delta > \gamma/2, \\ \varepsilon^{-2}|\log(\varepsilon)|^4, & \delta = \gamma/2, \\ \varepsilon^{-\gamma/\delta}|\log(\varepsilon)|^{\gamma/\delta}, & \delta < \gamma/2. \end{Bmatrix} + \varepsilon^{-\frac{\gamma}{\delta}}|\log(\varepsilon)|^{\frac{\gamma}{\delta}}$$

By noticing that the cost on the coarsest level is dominated in every case by the cost on the finer levels, we arrive at the assertion. □

3.4 Theoretical comparison of the proposed methods

In this section we shall give a theoretical comparison of the proposed methods and discuss the parameters that lead us to choose one method over the other.

Let us first collect the results for cost and memory requirements of the analyzed methods into the following tables. We shall abbreviate the methods as follows. The single level full tensor product Monte Carlo method will be termed *FTP-MC* and the single level sparse tensor product Monte Carlo method by *STP-MC*. For the multilevel variants we choose the abbreviations *FTP-MLMC* and *STP-MLMC* for the full tensor and sparse tensor product MLMC, respectively.

FTP-MC	$\varepsilon^{-2-\frac{\max\{\gamma,2\}}{\delta}}$
STP-MC	$\varepsilon^{-2-\frac{\gamma}{\delta}} \cdot \begin{cases} \lvert\log(\varepsilon)\rvert^{1+1/\delta}, & \gamma = 1 \\ \lvert\log(\varepsilon)\rvert^{\gamma/\delta}, & \gamma > 1 \end{cases}$

Table 3.1: Cost of the proposed Monte Carlo methods

Let us consider the case of $\gamma \geq 2$ first. Then we see from Table 3.1 that the cost of FTP-MC is $\varepsilon^{-2-\gamma/\delta}$ and is thus better than that of the STP-MC, since the sparse tensor product approximation features another factor of $|\log(\varepsilon)|^{\gamma/\delta}$, which is not present for the FTP-MC. Comparing the FTP-MC with the multilevel variants shows that both methods are better than the FTP-MC, but that the STP-MLMC loses against the FTP-MLMC, because of the additional log factors. This gives the following ranking in terms of cost

$$\text{STP-MC} \gtrsim \text{FTP-MC} \gtrsim \text{STP-MLMC} \gtrsim \text{FTP-MLMC}, \tag{3.55}$$

when the solver has the complexity $\gamma \geq 2$.

In the case of better solvers, i.e. $\gamma \in [1, 2)$, for e.g. multigrid methods, the situation is different. Here, since the full tensor product Monte Carlo method does not benefit

FTP-MLMC	$\begin{cases} \varepsilon^{-2}, & 2\delta > \max\{\gamma, 2\} \\ \varepsilon^{-2}\lvert\log(\varepsilon)\rvert^2, & 2\delta = \max\{\gamma, 2\} \\ \varepsilon^{-\frac{\max\{\gamma,2\}}{\delta}}, & 2\delta < \max\{\gamma, 2\} \end{cases} + \varepsilon^{-\max\{\gamma,2\}/\delta}$
STP-MLMC	$\underline{\gamma = 1}:$ $\begin{cases} \varepsilon^{-2}, & 2\delta > 1 \\ \varepsilon^{-2}\lvert\log(\varepsilon)\rvert^5, & 2\delta = 1 \\ \varepsilon^{-1/\delta}\lvert\log(\varepsilon)\rvert^{1+1/\delta}, & 2\delta < 1 \end{cases} + \varepsilon^{-1/\delta}\lvert\log(\varepsilon)\rvert^{1/\delta}$
	$\underline{\gamma > 1}:$ $\begin{cases} \varepsilon^{-2}, & 2\delta > \gamma \\ \varepsilon^{-2}\lvert\log(\varepsilon)\rvert^4, & 2\delta = \gamma \\ \varepsilon^{-\gamma/\delta}\lvert\log(\varepsilon)\rvert^{\gamma/\delta}, & 2\delta < \gamma \end{cases} + \varepsilon^{-\gamma/\delta}\lvert\log(\varepsilon)\rvert^{\gamma/\delta}$

Table 3.2: Cost of the proposed multilevel Monte Carlo methods

at all from a better solver, we see that the sparse methods will win in any case. Since for the multilevel variant we experience a multilevel acceleration, which is due to not having to sample too much on the finer levels, it is obvious that STP-MC is more costly than the STP-MLMC. The question is now where the FTP-MLMC fits into the picture. Comparing the different estimates we see that the FTP-MLMC behaves worse than the STP-MLMC, since $\gamma \in [1, 2)$ and we are stuck with a 2 in the maximum of the exponent. If there holds

$$\varepsilon^{-2}\lvert\log(\varepsilon)\rvert^2 + \varepsilon^{-2/\delta} \leq \varepsilon^{-2}\lvert\log(\varepsilon)\rvert^4 + \varepsilon^{-\gamma/\delta}\lvert\log(\varepsilon)\rvert^{\gamma/\delta}$$

then the FTP-MLMC has a better cost effectiveness. This happens when the solver becomes worse, i.e. if $\gamma \approx 2$. Then it can happen, that in the computationally attractive range the FTP MLMC method might be more cost effective. This effect can e.g. be due to the work that has to be put in for highly resolving a source term f naïvely for the computation of the right-hand side of the linear system. Nevertheless, the STP-MLMC method will win in the long run, i.e. will have a better asymptotical cost to accuracy ratio, anyway. From this perspective it is the outright best choice of the methods presented. Finally, we note that in the case that $\gamma \in [1, 2)$ we have the following ranking of the proposed methods with respect to computational cost:

$$\text{FTP-MC} \gtrsim \text{STP-MC} \gtrsim \text{FTP-MLMC} \gtrsim \text{STP-MLMC}. \tag{3.56}$$

FTP-MC	N_L^2
STP-MC	$N_L \log(N_L)$
FTP-MLMC	N_L^2
STP-MLMC	$N_L \log(N_L)$

Table 3.3: Memory requirements for the proposed methods

With respect to the memory requirements we see in any case that the sparse tensor product methods are much better as they grow with an *essentially* linear rate, i.e. up to logarithmic factors. This means that in situations where the data of the solution needs to be archived the sparse tensor product methods are the obvious choice.

3.5 Numerical experiments

In order to illustrate the theoretical construction and convergence of the proposed methods, we recall in the following the stochastic elliptic model problem in a Lipschitz domain $D \subset \mathbb{R}^d$ with $d = 1, 2$, which is used in the numerical experiments:

$$\begin{cases} -\nabla \cdot (\kappa(x,\omega)\nabla u(x,\omega)) &= f(x,\omega), \quad \text{in } D, \\ u(x,\omega) &= 0, \qquad\quad \text{on } \Gamma = \partial D. \end{cases} \tag{3.57}$$

for *all* $\omega \in \Omega$. Furthermore, we will make the following assumption on the data f and the random diffusion coefficient κ.

Assumption 3.26. *For $2 \leq k \leq \infty$, assume that $f \in L^k(\Omega; L^2(D))$ and that for* every *$\omega \in \Omega$ we have $\kappa(\cdot,\omega) \in W^{1,\infty}(D)$, such that there exist two finite and positive constants $\kappa_-, \kappa_+ \in \mathbb{R}$ with*

$$0 < \kappa_- \leq \operatorname{ess\,inf}_{x\in D}\kappa(x,\omega) \leq \|\kappa(x,\omega)\|_{L^\infty(D)} \leq \kappa_+ < \infty, \quad \forall \omega \in \Omega,$$

and hence that the random diffusion coefficient $\kappa(x,\omega)$ is uniformly bounded for all $\omega \in \Omega$ on D. Additionally, assume that κ and f are independent and strongly measurable as mappings taking values in $W^{1,\infty}(D)$ and $L^2(D)$, respectively.

In the following discussion let $k = 2$ and set $H = H_0^1(D)$, whence the weak formulation of (3.57) reads: Find $u \in L^2(\Omega; H)$, such that for all $v \in H$ there holds

$$a(u,v) := \mathbb{E}\left[\int_D \kappa(x,\omega)\nabla u(x,\omega)\nabla v(x)\,\mathrm{d}x\right] = \mathbb{E}\left[\int_D f(x,\omega)v(x)\,\mathrm{d}x\right] =: L(v). \tag{3.58}$$

Under Assumption 3.26 it is easy to show that the stochastic elliptic boundary value problem (3.57) admits a unique solution $u \in L^2(\Omega; H)$ for every given $f \in L^2(\Omega; L^2(D))$. Moreover, Assumption 3.26 guarantees that the solution $u \in H^2_{\mathrm{loc}}(D)$ $\mathbb{P}$-almost surely. More precisely, $u \in L^k(\Omega; W)$ for $W := \{w \in H : \Delta w \in L^2(D), w = 0 \text{ on } \Gamma\}$ endowed with the norm $\|w\|_W = \|\Delta w\|_{L^2(D)} + \|w\|_{L^2(D)}$ and for $2 \leq k \leq \infty$ there holds the *a priori* estimate

$$\|u\|_{L^k(\Omega;W)} \leq C(\kappa)\|f\|_{L^k(\Omega;L^2(D))}.$$

For more details we refer the reader to [5].

3.5.1 Discretization

We will employ the Finite Element Method (FEM) to discretise the model problem. Therefore, let $\{\mathcal{T}\}_{\ell=1}^\infty$ denote a sequence of regular meshes on the polygonal domain D of quasi-uniform intervals for $d = 1$ and triangles for $d = 2$, respectively. As usual a mesh $\mathcal{T}$ is called regular, if the intersection of two elements K and K' is either empty, a vertex, or an entire edge. Moreover, denote the meshwidth of $\mathcal{T}_\ell$ by $h_\ell = \max_{K\in\mathcal{T}_\ell} \operatorname{diam}(K)$ and assume that $\mathcal{T}_\ell$ is σ-shape regular for all ℓ, i.e. there exists a finite positive constant σ, such that $\sigma = \sup_\ell \max_{K\in\mathcal{T}_\ell} \frac{h_K}{\rho_K}$, where ρ_K denotes the maximal radius of the element incircle. Uniform mesh refinement is achieved by regular subdivision of the elements of $\mathcal{T}_\ell$. Thus, $\mathcal{T}_{\ell+1}$ is obtained by uniform refinement of $\mathcal{T}_\ell$. We will use the spaces $V_\ell = \mathcal{S}_0^{1,0}(\mathcal{T}_\ell), \ell \geq 1$. Since

$$V_1 \subset V_2 \subset \cdots \subset V_\ell \subset \cdots \subset H$$

the corresponding Galerkin formulation is conforming and the discrete problem reads: Find $u_\ell \in L^2(\Omega; V_\ell)$, such that

$$a(u_\ell, v_\ell) = L(v_\ell), \quad \forall v_\ell \in L^2(\Omega; V_\ell).$$

By the assumption on the random diffusion coefficient κ and the conformity of the method there exists a unique FE solution $u_\ell \in L^2(\Omega; V_\ell)$. Moreover, it is well-known that the solution u_ℓ admits the following quasi-optimality property

$$\|u - u_\ell\|_{L^2(\Omega;H)} \leq C \inf_{v_\ell \in V_\ell} \|u - v_\ell\|_{L^2(\Omega;H)}.$$

Moreover, standard FEM theory yields that

$$\inf_{v_\ell \in V_\ell} \|u - v_\ell\|_V \leq C N_\ell^{-1/d} \|u\|_W,$$

where $N_\ell = \dim(V_\ell)$ and $N_\ell \sim h_\ell^{-d}$. We recall the unbiased covariance estimator $\mathrm{Cov}_M[X, Y]$ from Definition 3.11 which is given by

$$\mathrm{Cov}_M[X, Y] := \frac{1}{M-1} \sum_{i=1}^{M} (X^i - \mathrm{E}_M[X]) \otimes (Y^i - \mathrm{E}_M[Y]).$$

For the approximation of the covariance function of u we introduce more convenient representatitons for the implementation of the covariance estimators $\widehat{\mathrm{Cov}}_M[u_L, u_L]$ and $\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{U}, \mathcal{U}]$, respectively, where $\mathcal{U} = \{u_\ell\}_{\ell=0}^L$ is the sequence of FE solutions. Then we can write

$$\begin{aligned}
\widehat{\mathrm{Cov}}_M[u_L, u_L] &= \frac{1}{M-1} \sum_{i=1}^{M} \sum_{k+j \leq L} (\Delta_k \mathcal{U}^i - \mathrm{E}_M[\Delta_k \mathcal{U}]) \otimes (\Delta_j \mathcal{U}^i - \mathrm{E}_M[\Delta_j \mathcal{U}]) \\
&= \sum_{k=0}^{L} \mathrm{Cov}_M[u_k, u_{L-k}] - \sum_{k=0}^{L-1} \mathrm{Cov}_M[u_k, u_{L-k-1}], \qquad (3.59) \\
\widehat{\mathrm{Cov}}^{\mathrm{ML}}[\mathcal{U}, \mathcal{U}] &= \sum_{\ell=0}^{L} \left\{ \widehat{\mathrm{Cov}}_{M_\ell}[u_\ell, u_\ell] - \widehat{\mathrm{Cov}}_{M_\ell}[u_{\ell-1}, u_{\ell-1}] \right\} \\
&= \sum_{\ell=0}^{L} \left\{ \sum_{k=0}^{\ell} \mathrm{Cov}_{M_\ell}[u_k, u_{\ell-k}] - 2 \sum_{k=0}^{\ell-1} \mathrm{Cov}_{M_\ell}[u_k, u_{\ell-k-1}] \right. \\
&\qquad \left. + \sum_{k=0}^{\ell-2} \mathrm{Cov}_{M_\ell}[u_k, u_{\ell-k-2}] \right\}. \qquad (3.60)
\end{aligned}$$

Remark 3.27. *By symmetry not all blocks* $\mathrm{Cov}_M[u_k, u_{L-k}]$ *have to be stored as their transposed coefficient matrices represent those of* $\mathrm{Cov}_M[u_{L-k}, u_k]$. *The same holds for the level corrections in the multilevel variant, there not all* $\mathrm{Cov}_{M_\ell}[u_k, u_{\ell-k}], \mathrm{Cov}_{M_\ell}[u_k, u_{\ell-k-1}]$ *nor all* $\mathrm{Cov}_{M_\ell}[u_k, u_{\ell-k-2}]$ *have to be stored for same reason. This saving in memory requirements is on top of the already favourable scaling of memory requirements as* $O(N_L \log(N_L))$.

3.5.2 Computation of $H^{1,1}(D \times D)$ norms and errors

In this section we present how to compute norms of errors and norms of discrete functions in $H^{1,1}(D \times D)$. For example, consider a covariance function $\mathrm{Cov}_M[X_k, X_\ell] \in$

$L^2(\Omega; V_k \otimes V_\ell) \subset L^2(\Omega; H \otimes H)$ for certain finite dimensional subspaces $V_k, V_\ell \subset H$ subject to a number M of samples of X_k, X_ℓ. Let us denote the dimension of any finite dimensional space V_j by N_j, i.e. $\dim(V_j) = N_j$, and the corresponding bases of V_k and V_ℓ by $\{\phi_s^k\}_{s=1}^{N_k}$ and $\{\phi_t^\ell\}_{t=1}^{N_\ell}$, respectively. There holds the *universal* representation

$$\mathrm{Cov}_M\left[X_k, X_\ell\right](x_1, x_2) = \sum_{s=1}^{N_k} \sum_{t=1}^{N_\ell} c_{s,t}^{k,\ell} \phi_s^k(x_1) \phi_t^\ell(x_2),$$

where we have suppressed the random dependence of the coefficients $c_{s,t}^{k,\ell}$ on $\omega \in \Omega$. This representation is *universal* in the sense that the specific basis chosen for the spaces V_k and V_ℓ is *a priori* arbitrary. Let $\eta, \psi \in H \otimes H$, more precisely, let $\eta \in V_{k_1} \otimes V_{k_2} =: V_{k_1,k_2}$ and $\psi \in V_{\ell_1} \otimes V_{\ell_2} =: V_{\ell_1,\ell_2}$ with the following representations

$$\eta(x_1, x_2) = \sum_{i_1=1}^{N_{k_1}} \sum_{i_2=1}^{N_{k_2}} c_{i_1,i_2}^{k_1,k_2} \phi_{i_1}^{k_1}(x_1) \phi_{i_2}^{k_2}(x_2),$$
$$\psi(x_1, x_2) = \sum_{j_1=1}^{N_{\ell_1}} \sum_{j_2=1}^{N_{\ell_2}} d_{j_1,j_2}^{\ell_1,\ell_2} \phi_{j_1}^{\ell_1}(x_1) \phi_{j_2}^{\ell_2}(x_2).$$

Then

$$\begin{aligned}
(\eta, \psi)_{H^{1,1}(D\times D)} &= \int_D \int_D (\nabla_{x_1} \otimes \nabla_{x_2}) \left(\sum_{i_1=1}^{N_{k_1}} \sum_{i_2=1}^{N_{k_2}} c_{i_1,i_2}^{k_1,k_2} \phi_{i_1}^{k_1}(x_1) \phi_{i_2}^{k_2}(x_2) \right) \\
&\qquad \cdot (\nabla_{x_1} \otimes \nabla_{x_2}) \left(\sum_{j_1=1}^{N_{\ell_1}} \sum_{j_2=1}^{N_{\ell_2}} d_{j_1,j_2}^{\ell_1,\ell_2} \phi_{j_1}^{\ell_1}(x_1) \phi_{j_2}^{\ell_2}(x_2) \right) \mathrm{d}x_1 \, \mathrm{d}x_2 \\
&= \sum_{i_1=1}^{N_{k_1}} \sum_{i_2=1}^{N_{k_2}} \sum_{j_1=1}^{N_{\ell_1}} \sum_{j_2=1}^{N_{\ell_2}} \left(c_{i_1,i_2}^{k_1,k_2} d_{j_1,j_2}^{\ell_1,\ell_2} \int_D \nabla \phi_{i_1}^{k_1}(x_1) \nabla \phi_{j_1}^{\ell_1}(x_1) \, \mathrm{d}x_1 \right. \\
&\qquad \left. \times \int_D \nabla \phi_{i_2}^{k_2}(x_2) \nabla \phi_{j_2}^{\ell_2}(x_2) \, \mathrm{d}x_2 \right) \\
&= \sum_{i_1=1}^{N_{k_1}} \sum_{i_2=1}^{N_{k_2}} \sum_{j_1=1}^{N_{\ell_1}} \sum_{j_2=1}^{N_{\ell_2}} c_{i_1,i_2}^{k_1,k_2} d_{j_1,j_2}^{\ell_1,\ell_2} a_{i_1,j_1}^{k_1,\ell_1} a_{i_2,j_2}^{k_2,\ell_2}.
\end{aligned}$$

Now, since the coefficient matrices $C^{k_1,k_2} = (c_{i_1,i_2}^{k_1,k_2}) \in \mathbb{R}^{N_{k_1},N_{k_2}}$, $D^{\ell_1,\ell_2} = (d_{j_1,j_2}^{\ell_1,\ell_2}) \in \mathbb{R}^{N_{j_1},N_{j_2}}$ and the "mixed" stiffness matrices $A^{k_r,\ell_r} = (a_{i_r,j_r}^{k_r,\ell_r}) \in \mathbb{R}^{N_{k_r},N_{\ell_r}}, r = 1, 2$, upon contracting indices by suitable matrix multiplications we arrive at

$$\begin{aligned}
(\eta, \psi)_{H^{1,1}(D\times D)} &= \sum_{i_1=1}^{N_{k_1}} \sum_{i_2=1}^{N_{k_2}} \sum_{j_1=1}^{N_{\ell_1}} \sum_{j_2=1}^{N_{\ell_2}} c_{i_1,i_2}^{k_1,k_2} d_{j_1,j_2}^{\ell_1,\ell_2} a_{i_1,j_1}^{k_1,\ell_1} a_{i_2,j_2}^{k_2,\ell_2} \\
&= \sum_{i_1=1}^{N_{k_1}} \sum_{j_2=1}^{N_{\ell_2}} \underbrace{\left\{ \sum_{i_2=1}^{N_{k_2}} c_{i_1,i_2}^{k_1,k_2} a_{i_2,j_2}^{k_2,\ell_2} \right\}}_{=: f_{i_1,j_2}} \cdot \underbrace{\left\{ \sum_{j_1=1}^{N_{\ell_1}} a_{i_1,j_1}^{k_1,\ell_1} d_{j_1,j_2}^{\ell_1,\ell_2} \right\}}_{=: g_{i_1,j_2}} \\
&= \sum_{i_1=1}^{N_{k_1}} \sum_{j_2=1}^{N_{\ell_2}} f_{i_1,j_2} g_{i_1,j_2} = \langle F, G \rangle_{\mathrm{Frob}} = \langle C^{k_1,k_2} A^{k_2,\ell_2}, A^{k_1,\ell_1} D^{\ell_1,\ell_2} \rangle_{\mathrm{Frob}}.
\end{aligned}$$

Hence the exact computation of the inner product can be realised by computing certain Frobenius norms of matrix products of stiffness matrices and corresponding coefficient matrices. The mixed stiffness matrices $A^{k_r,\ell_r}, r = 1, 2$, can be computed once in the postprocessing step, such that the Frobenius inner product can then be readily computed.

Computation of norms of errors using a reference solution

When computing the errors for the presented full tensor product and sparse tensor product Monte Carlo or multilevel Monte Carlo methods, the question arises in which way a comparison is most suitable. As the MLMC method is usually superior in its cost to accuracy relation, we propose taking as reference solution u_{ref} the sparse tensor MLMC approximation computed on the finest discretisation level available.

By the previous observation and universal representation of approximations involving tensor products, we can characterise the error simply by the $H^{1,1}(D \times D)$ inner product, namely

$$|u_{\text{ref}} - \eta|^2_{H^{1,1}(D\times D)} = |u_{\text{ref}}|^2_{H^{1,1}(D\times D)} - 2(u_{\text{ref}}, \eta)_{H^{1,1}(D\times D)} + |\eta|^2_{H^{1,1}(D\times D)}.$$

Obviously, $|u_{\text{ref}}|^2_{H^{1,1}(D\times D)}$ has to be computed only once and is analagous to Section 3.5.2. Hence, the subsequent discussion deals mostly with the computation of the two remaining terms in this expression.

Here we have denoted by η any solution obtained by either full tensor MC, sparse tensor MC or sparse tensor MLMC.

Remark 3.28. *Note that we have not listed how to proceed for the full tensor product MLMC solution. This is due to the fact that the computation in that case is analogous to the computations needed for the full tensor product MC case.*

In what follows we briefly describe how to compute the errors for the three methods involved. In order to keep the presentation simple, we state the format of solutions for each of the cases explicitly. As a shorthand notation let us also define $\eta_{k,\ell} \in V_{k,\ell}$ in terms of the basis functions in $V_{k,\ell}$ by

$$\eta_{k,\ell}(x_1, x_2) = \sum_{i_1=1}^{N_k} \sum_{i_2=1}^{N_\ell} c^{k,\ell}_{i_1,i_2} \phi^k_{i_1}(x_1)\phi^\ell_{i_2}(x_2).$$

Using this expression we can write solutions of the three methods in question symbolically as follows:

(1) full tensor MC on level L: $\eta_{L,L}(x_1, x_2)$,

(2) sparse tensor MC on level L: $\eta_L(x_1, x_2) = \sum_{i=0}^{L} \eta_{i,L-i}(x_1, x_2) - \sum_{i=0}^{L-1} \eta_{i,L-1-i}(x_1, x_2)$,

(3) sparse tensor MLMC on level L:
$$\eta^L(x_1, x_2) = \sum_{\lambda=0}^{L} \left\{ \sum_{i=0}^{\lambda} \eta_{i,\lambda-i}(x_1, x_2) - 2 \sum_{i=0}^{\lambda-1} \eta_{i,\lambda-1-i}(x_1, x_2) + \sum_{i=0}^{\lambda-2} \eta_{i,\lambda-2-i}(x_1, x_2) \right\}.$$

It is noteworthy that, while the explicit expressions for single level and multilevel MC look rather bulky, we will by symmetry only have to store roughly half of the corresponding diagonals of (small) full tensor products, cf. also Remark 3.21.

The subsequent formulae are presented as a quick look-up reference for the computation.

η as full tensor product MC solution

We are concerned with the computation of $|\eta_{L,L}|^2_{H^{1,1}(D\times D)}$ and $(u_{\mathrm{ref}}, \eta_{L,L})_{H^{1,1}(D\times D)}$. There holds

$$\begin{aligned}
|\eta_{L,L}|^2_{H^{1,1}(D\times D)} &= \int_D \int_D ((\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2))^2 \,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \sum_{i_1=1}^{N_L}\sum_{i_2=1}^{N_L}\sum_{j_1=1}^{N_L}\sum_{j_2=1}^{N_L} \left(d^{L,L}_{i_1,i_2} d^{L,L}_{j_1,j_2} \int_D \nabla\phi^L_{i_1}(x_1)\nabla\phi^L_{j_1}(x_1)\,\mathrm{d}x_1 \right. \\
&\qquad \left. \int_D \nabla\phi^L_{i_2}(x_2)\nabla\phi^L_{j_2}(x_2)\,\mathrm{d}x_2 \right) \\
&= \sum_{i_1=1}^{N_L}\sum_{i_2=1}^{N_L}\sum_{j_1=1}^{N_L}\sum_{j_2=1}^{N_L} d^{L,L}_{i_1,i_2} d^{L,L}_{j_1,j_2} a^{L,L}_{i_1,j_1} a^{L,L}_{i_2,j_2} \\
&= \langle D^{L,L} A^{L,L}, A^{L,L} D^{L,L} \rangle_{\mathrm{Frob}}.
\end{aligned}$$

Concerning $(u_{\mathrm{ref}}, \eta_{L,L})_{H^{1,1}(D\times D)}$, we proceed with u_{ref} given by

$$u_{\mathrm{ref}} = \sum_{\lambda=0}^{L_{\mathrm{ref}}} \left\{ \sum_{i=0}^{\lambda} u_{i,\lambda-i} - 2\sum_{i=0}^{\lambda-1} u_{i,\lambda-1-i} + \sum_{i=0}^{\lambda-2} u_{i,\lambda-2-i} \right\}$$

as follows

$$\begin{aligned}
(u_{\mathrm{ref}}, \eta_{L,L}) &= \int_D \int_D ((\nabla_{x_1} \otimes \nabla_{x_2})u_{\mathrm{ref}}(x_1,x_2))((\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \int_D \int_D (\nabla_{x_1} \otimes \nabla_{x_2}) \left(\sum_{\lambda=0}^{L_{\mathrm{ref}}} \left\{ \sum_{i=0}^{\lambda} u_{i,\lambda-i} - 2\sum_{i=0}^{\lambda-1} u_{i,\lambda-1-i} + \sum_{i=0}^{\lambda-2} u_{i,\lambda-2-i} \right\} \right) \\
&\qquad \times (\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \int_D \int_D (\nabla_{x_1} \otimes \nabla_{x_2}) \left(\sum_{\lambda=0}^{L_{\mathrm{ref}}} \sum_{i=0}^{\lambda} u_{i,\lambda-i}(x_1,x_2) \right) (\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&\quad - 2\int_D \int_D (\nabla_{x_1} \otimes \nabla_{x_2}) \left(\sum_{\lambda=0}^{L_{\mathrm{ref}}} \sum_{i=0}^{\lambda-1} u_{i,\lambda-1-i}(x_1,x_2) \right) (\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&\quad + \int_D \int_D (\nabla_{x_1} \otimes \nabla_{x_2}) \left(\sum_{\lambda=0}^{L_{\mathrm{ref}}} \sum_{i=0}^{\lambda-2} u_{i,\lambda-2-i}(x_1,x_2) \right) (\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \sum_{\lambda=0}^{L_{\mathrm{ref}}} \sum_{i=0}^{\lambda} \int_D \int_D ((\nabla_{x_1} \otimes \nabla_{x_2})u_{i,\lambda-i}(x_1,x_2)) ((\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&\quad - 2\sum_{\lambda=0}^{L_{\mathrm{ref}}} \sum_{i=0}^{\lambda-1} \int_D \int_D ((\nabla_{x_1} \otimes \nabla_{x_2})u_{i,\lambda-1-i}(x_1,x_2)) ((\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&\quad + \sum_{\lambda=0}^{L_{\mathrm{ref}}} \sum_{i=0}^{\lambda-2} \int_D \int_D ((\nabla_{x_1} \otimes \nabla_{x_2})u_{i,\lambda-2-i}(x_1,x_2)) ((\nabla_{x_1} \otimes \nabla_{x_2})\eta_{L,L}(x_1,x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2 \\
&= \sum_{\lambda=0}^{L_{\mathrm{ref}}} \sum_{i=0}^{\lambda} \left(\sum_{i_1=1}^{N_i}\sum_{j_2=1}^{N_L} \left\{ \sum_{i_2=1}^{N_{\lambda-i}} c^{i,\lambda-i}_{i_1,i_2} a^{\lambda-i,L}_{i_2,j_2} \right\} \cdot \left\{ \sum_{j_1=1}^{N_L} a^{i,L}_{i_1,j_1} d^{L,L}_{j_1,j_2} \right\} \right)
\end{aligned}$$

$$
\begin{aligned}
&- 2\sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-1}\left(\sum_{i_1=1}^{N_i}\sum_{j_2=1}^{N_L}\left\{\sum_{i_2=1}^{N_{\lambda-1-i}} c_{i_1,i_2}^{i,\lambda-1-i} a_{i_2,j_2}^{\lambda-1-i,L}\right\}\cdot\left\{\sum_{j_1=1}^{N_L} a_{i_1,j_1}^{i,L} d_{j_1,j_2}^{L,L}\right\}\right)\\
&+\sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-2}\left(\sum_{i_1=1}^{N_i}\sum_{j_2=1}^{N_L}\left\{\sum_{i_2=1}^{N_{\lambda-2-i}} c_{i_1,i_2}^{i,\lambda-2-i} a_{i_2,j_2}^{\lambda-2-i,L}\right\}\cdot\left\{\sum_{j_1=1}^{N_L} a_{i_1,j_1}^{i,L} d_{j_1,j_2}^{L,L}\right\}\right)\\
&= \sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda}\langle C^{i,\lambda-i}A^{\lambda-i,L}, A^{i,L}D^{L,L}\rangle_{\mathrm{Frob}} - 2\sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-1}\langle C^{i,\lambda-1-i}A^{\lambda-1-i,L}, A^{i,L}D^{L,L}\rangle_{\mathrm{Frob}}\\
&+\sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-2}\langle C^{i,\lambda-2-i}A^{\lambda-2-i,L}, A^{i,L}D^{L,L}\rangle_{\mathrm{Frob}}.
\end{aligned}
$$

η as sparse tensor product MC solution

We are concerned with the computation of $|\eta_L|^2_{H^{1,1}(D\times D)}$ and $(u_{\mathrm{ref}},\eta_L)_{H^{1,1}(D\times D)}$. There holds

$$
\begin{aligned}
|\eta_L|^2_{H^{1,1}(D\times D)} &= \int_D\int_D((\nabla_{x_1}\otimes\nabla_{x_2})\eta_L(x_1,x_2))^2\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&= \int_D\int_D\left((\nabla_{x_1}\otimes\nabla_{x_2})\left(\sum_{i=0}^{L}\eta_{i,L-i}(x_1,x_2)-\sum_{i=0}^{L-1}\eta_{i,L-1-i}(x_1,x_2)\right)\right)^2\mathrm{d}x_1\,\mathrm{d}x_2\\
&= \int_D\int_D\left((\nabla_{x_1}\otimes\nabla_{x_2})\left(\sum_{i=0}^{L}\eta_{i,L-i}(x_1,x_2)\right)\right)^2\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad - 2\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})\left(\sum_{i=0}^{L}\eta_{i,L-i}(x_1,x_2)\right)(\nabla_{x_1}\otimes\nabla_{x_2})\left(\sum_{i=0}^{L-1}\eta_{i,L-1-i}(x_1,x_2)\right)\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad + \int_D\int_D\left((\nabla_{x_1}\otimes\nabla_{x_2})\left(\sum_{i=0}^{L-1}\eta_{i,L-1-i}(x_1,x_2)\right)\right)^2\mathrm{d}x_1\,\mathrm{d}x_2\\
&= \sum_{s=0}^{L}\sum_{t=0}^{L}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-s}(x_1,x_2)(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{t,L-t}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad - 2\sum_{s=0}^{L}\sum_{t=0}^{L-1}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-s}(x_1,x_2)(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{t,L-1-t}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad + \sum_{s=0}^{L-1}\sum_{t=0}^{L-1}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-1-s}(x_1,x_2)(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{t,L-1-t}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&= \sum_{s=0}^{L}\sum_{t=0}^{L}\left(\sum_{i_1=1}^{N_s}\sum_{j_2=1}^{N_{L-t}}\left\{\sum_{i_2=1}^{N_{L-s}} c_{i_1,i_2}^{s,L-s} a_{i_2,j_2}^{L-s,L-t}\right\}\cdot\left\{\sum_{j_1=1}^{N_t} a_{i_1,j_1}^{s,t} c_{j_1,j_2}^{t,L-t}\right\}\right)\\
&\quad - 2\sum_{s=0}^{L}\sum_{t=0}^{L-1}\left(\sum_{i_1=1}^{N_s}\sum_{j_2=1}^{N_{L-1-t}}\left\{\sum_{i_2=1}^{N_{L-s}} c_{i_1,i_2}^{s,L-s} a_{i_2,j_2}^{L-s,L-1-t}\right\}\cdot\left\{\sum_{j_1=1}^{N_t} a_{i_1,j_1}^{s,t} c_{j_1,j_2}^{t,L-1-t}\right\}\right)\\
&\quad + \sum_{s=0}^{L-1}\sum_{t=0}^{L-1}\left(\sum_{i_1=1}^{N_s}\sum_{j_2=1}^{N_{L-1-t}}\left\{\sum_{i_2=1}^{N_{L-1-s}} c_{i_1,i_2}^{s,L-1-s} a_{i_2,j_2}^{L-1-s,L-1-t}\right\}\cdot\left\{\sum_{j_1=1}^{N_t} a_{i_1,j_1}^{s,t} c_{j_1,j_2}^{t,L-1-t}\right\}\right)
\end{aligned}
$$

$$= \sum_{s=0}^{L}\sum_{t=0}^{L}\langle C^{s,L-s}A^{L-s,L-t}, A^{s,t}C^{t,L-t}\rangle_{\mathrm{Frob}} - 2\sum_{s=0}^{L}\sum_{t=0}^{L-1}\langle C^{s,L-s}A^{L-s,L-1-t}, A^{s,t}C^{t,L-1-t}\rangle_{\mathrm{Frob}}$$
$$+ \sum_{s=0}^{L-1}\sum_{t=0}^{L-1}\langle C^{s,L-1-s}A^{L-1-s,L-1-t}, A^{s,t}C^{t,L-1-t}\rangle_{\mathrm{Frob}}.$$

Concerning $(u_{\mathrm{ref}}, \eta_L)_{H^{1,1}(D\times D)}$, we proceed with u_{ref} given as before and find

$$\begin{aligned}
(u_{\mathrm{ref}}, \eta_L) &= \int_D\int_D ((\nabla_{x_1}\otimes\nabla_{x_2})u_{\mathrm{ref}}(x_1,x_2))((\nabla_{x_1}\otimes\nabla_{x_2})\eta_L(x_1,x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&= \int_D\int_D (\nabla_{x_1}\otimes\nabla_{x_2})\left(\sum_{\lambda=0}^{L_{\mathrm{ref}}}\left\{\sum_{i=0}^{\lambda}u_{i,\lambda-i} - 2\sum_{i=0}^{\lambda-1}u_{i,\lambda-1-i} + \sum_{i=0}^{\lambda-2}u_{i,\lambda-2-i}\right\}\right)\\
&\qquad\times(\nabla_{x_1}\otimes\nabla_{x_2})\left(\sum_{i=0}^{L}\eta_{i,L-i}(x_1,x_2) - \sum_{i=0}^{L-1}\eta_{i,L-1-i}(x_1,x_2)\right)\mathrm{d}x_1\,\mathrm{d}x_2\\
&= \sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda}\sum_{s=0}^{L}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})u_{i,\lambda-i}(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-s}\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad - 2\sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-1}\sum_{s=0}^{L}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})u_{i,\lambda-1-i}(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-s}\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad + \sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-2}\sum_{s=0}^{L}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})u_{i,\lambda-2-i}(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-s}\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad - \sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda}\sum_{s=0}^{L-1}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})u_{i,\lambda-i}(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-1-s}\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad + 2\sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-1}\sum_{s=0}^{L-1}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})u_{i,\lambda-1-i}(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-1-s}\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&\quad - \sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{i=0}^{\lambda-2}\sum_{s=0}^{L-1}\int_D\int_D(\nabla_{x_1}\otimes\nabla_{x_2})u_{i,\lambda-2-i}(\nabla_{x_1}\otimes\nabla_{x_2})\eta_{s,L-1-s}\,\mathrm{d}x_1\,\mathrm{d}x_2\\
&= \sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{s=0}^{L}\left\{\sum_{i=0}^{\lambda}\langle D^{i,\lambda-i}A^{\lambda-i,L-s}, A^{i,s}C^{s,L-s}\rangle_{\mathrm{Frob}}\right.\\
&\quad - 2\sum_{i=0}^{\lambda-1}\langle D^{i,\lambda-1-i}A^{\lambda-1-i,L-s}, A^{i,s}C^{s,L-s}\rangle_{\mathrm{Frob}}\\
&\quad \left. + \sum_{i=0}^{\lambda-2}\langle D^{i,\lambda-2-i}A^{\lambda-2-i,L-s}, A^{i,s}C^{s,L-s}\rangle_{\mathrm{Frob}}\right\}\\
&\quad - \sum_{\lambda=0}^{L_{\mathrm{ref}}}\sum_{s=0}^{L-1}\left\{\sum_{i=0}^{\lambda}\langle D^{i,\lambda-i}A^{\lambda-i,L-1-s}, A^{i,s}C^{s,L-1-s}\rangle_{\mathrm{Frob}}\right.\\
&\qquad - 2\sum_{i=0}^{\lambda-1}\langle D^{i,\lambda-1-i}A^{\lambda-1-i,L-1-s}, A^{i,s}C^{s,L-1-s}\rangle_{\mathrm{Frob}}\\
&\qquad \left. + \sum_{i=0}^{\lambda-2}\langle D^{i,\lambda-2-i}A^{\lambda-2-i,L-1-s}, A^{i,s}C^{s,L-1-s}\rangle_{\mathrm{Frob}}\right\}.
\end{aligned}$$

η as sparse tensor product MLMC solution

In this situation the reference solution is of the same form as the solution η^L on level L. We are concerned with the computation of $|\eta^L|^2_{H^{1,1}(D\times D)}$ and $(u_{\mathrm{ref}}, \eta^L)_{H^{1,1}(D\times D)}$. There holds

$$
\begin{aligned}
|\eta^L|^2_{H^{1,1}(D\times D)} &= \int_D \int_D ((\nabla_{x_1} \otimes \nabla_{x_2})\eta^L(x_1,x_2))^2 \,\mathrm{d}x_1 \,\mathrm{d}x_2 \\
&= \sum_{\lambda_1=0}^{L}\sum_{\lambda_2=0}^{L} \int_D\int_D (\nabla_{x_1}\otimes\nabla_{x_2}) \left\{ \sum_{i=0}^{\lambda_1} \eta_{i,\lambda_1-i} - 2\sum_{i=0}^{\lambda_1-1}\eta_{i,\lambda_1-1-i} + \sum_{i=0}^{\lambda_1-2}\eta_{i,\lambda_1-2-i}\right\} \\
&\qquad \times (\nabla_{x_1}\otimes\nabla_{x_2}) \left\{ \sum_{i=0}^{\lambda_2} \eta_{i,\lambda_2-i} - 2\sum_{i=0}^{\lambda_2-1}\eta_{i,\lambda_2-1-i} + \sum_{i=0}^{\lambda_2-2}\eta_{i,\lambda_2-2-i}\right\} \mathrm{d}x_1 \,\mathrm{d}x_2 \\
&= \sum_{\lambda_1=0}^{L}\sum_{\lambda_2=0}^{L} \int_D\int_D \left\{ \left(\sum_{i_1=0}^{\lambda_1} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_1,\lambda_1-i_1}\right)\left(\sum_{i_2=0}^{\lambda_2} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_2,\lambda_2-i_2}\right)\right\} \\
&\qquad + 4\left\{ \left(\sum_{i_1=0}^{\lambda_1-1} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_1,\lambda_1-1-i_1}\right)\left(\sum_{i_2=0}^{\lambda_2-1} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_2,\lambda_2-1-i_2}\right)\right\} \\
&\qquad + \left\{ \left(\sum_{i_1=0}^{\lambda_1-2} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_1,\lambda_1-2-i_1}\right)\left(\sum_{i_2=0}^{\lambda_2-2} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_2,\lambda_2-2-i_2}\right)\right\} \\
&\qquad + 2\left\{ \left(\sum_{i_1=0}^{\lambda_1} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_1,\lambda_1-i_1}\right)\left(\sum_{i_2=0}^{\lambda_2-2} ((\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_2,\lambda_2-2-i_2}\right)\right\} \\
&\qquad - 4\left\{ \left(\sum_{i_1=0}^{\lambda_1} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_1,\lambda_1-i_1}\right)\left(\sum_{i_2=0}^{\lambda_2-1} ((\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_2,\lambda_2-1-i_2}\right)\right\} \\
&\qquad - 4\left\{ \left(\sum_{i_1=0}^{\lambda_1-1} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_1,\lambda_1-1-i_1}\right)\left(\sum_{i_2=0}^{\lambda_2-2} (\nabla_{x_1}\otimes\nabla_{x_2})\eta_{i_2,\lambda_2-2-i_2}\right)\right\} \mathrm{d}x_1 \,\mathrm{d}x_2 \\
&= \sum_{\lambda_1=0}^{L}\sum_{\lambda_2=0}^{L} \left\{ \sum_{i_1=0}^{\lambda_1}\sum_{i_2=0}^{\lambda_2} \langle C^{i_1,\lambda_1-i_1} A^{\lambda_1-i_1,\lambda_2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-i_2}\rangle_{\mathrm{Frob}} \right. \\
&\qquad + 4\sum_{i_1=0}^{\lambda_1-1}\sum_{i_2=0}^{\lambda_2-1} \langle C^{i_1,\lambda_1-1-i_1} A^{\lambda_1-1-i_1,\lambda_2-1-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-1-i_2}\rangle_{\mathrm{Frob}} \\
&\qquad + \sum_{i_1=0}^{\lambda_1-2}\sum_{i_2=0}^{\lambda_2-2} \langle C^{i_1,\lambda_1-2-i_1} A^{\lambda_1-2-i_1,\lambda_2-2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-2-i_2}\rangle_{\mathrm{Frob}} \\
&\qquad + 2\sum_{i_1=0}^{\lambda_1}\sum_{i_2=0}^{\lambda_2-2} \langle C^{i_1,\lambda_1-i_1} A^{\lambda_1-i_1,\lambda_2-2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-2-i_2}\rangle_{\mathrm{Frob}} \\
&\qquad - 4\sum_{i_1=0}^{\lambda_1}\sum_{i_2=0}^{\lambda_2-1} \langle C^{i_1,\lambda_1-i_1} A^{\lambda_1-i_1,\lambda_2-1-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-1-i_2}\rangle_{\mathrm{Frob}} \\
&\qquad \left. -4\sum_{i_1=0}^{\lambda_1-1}\sum_{i_2=0}^{\lambda_2-2} \langle C^{i_1,\lambda_1-1-i_1} A^{\lambda_1-1-i_1,\lambda_2-2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-2-i_2}\rangle_{\mathrm{Frob}} \right\}.
\end{aligned}
$$

Analogously for $(u_{\mathrm{ref}}, \eta^L)_{H^{1,1}(D\times D)}$, we obtain

$$
\begin{aligned}
(u_{\mathrm{ref}}, \eta^L) = \sum_{\lambda_1=0}^{L_{\mathrm{ref}}} \sum_{\lambda_2=0}^{L} \Bigg\{ & \sum_{i_1=0}^{\lambda_1} \sum_{i_2=0}^{\lambda_2} \langle C^{i_1,\lambda_1-i_1} A^{\lambda_1-i_1,\lambda_2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-i_2} \rangle_{\mathrm{Frob}} \\
& + 4 \sum_{i_1=0}^{\lambda_1-1} \sum_{i_2=0}^{\lambda_2-1} \langle C^{i_1,\lambda_1-1-i_1} A^{\lambda_1-1-i_1,\lambda_2-1-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-1-i_2} \rangle_{\mathrm{Frob}} \\
& + \sum_{i_1=0}^{\lambda_1-2} \sum_{i_2=0}^{\lambda_2-2} \langle C^{i_1,\lambda_1-2-i_1} A^{\lambda_1-2-i_1,\lambda_2-2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-2-i_2} \rangle_{\mathrm{Frob}} \\
& + \sum_{i_1=0}^{\lambda_1} \sum_{i_2=0}^{\lambda_2-2} \langle C^{i_1,\lambda_1-i_1} A^{\lambda_1-i_1,\lambda_2-2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-2-i_2} \rangle_{\mathrm{Frob}} \\
& + \sum_{i_1=0}^{\lambda_1-2} \sum_{i_2=0}^{\lambda_2} \langle C^{i_1,\lambda_1-2-i_1} A^{\lambda_1-2-i_1,\lambda_2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-i_2} \rangle_{\mathrm{Frob}} \\
& - 2 \sum_{i_1=0}^{\lambda_1-1} \sum_{i_2=0}^{\lambda_2} \langle C^{i_1,\lambda_1-1-i_1} A^{\lambda_1-1-i_1,\lambda_2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-i_2} \rangle_{\mathrm{Frob}} \\
& - 2 \sum_{i_1=0}^{\lambda_1} \sum_{i_2=0}^{\lambda_2-1} \langle C^{i_1,\lambda_1-i_1} A^{\lambda_1-i_1,\lambda_2-1-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-1-i_2} \rangle_{\mathrm{Frob}} \\
& - 2 \sum_{i_1=0}^{\lambda_1-1} \sum_{i_2=0}^{\lambda_2-2} \langle C^{i_1,\lambda_1-1-i_1} A^{\lambda_1-1-i_1,\lambda_2-2-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-2-i_2} \rangle_{\mathrm{Frob}} \\
& - 2 \sum_{i_1=0}^{\lambda_1-2} \sum_{i_2=0}^{\lambda_2-1} \langle C^{i_1,\lambda_1-2-i_1} A^{\lambda_1-2-i_1,\lambda_2-1-i_2}, A^{i_1,i_2} C^{i_2,\lambda_2-1-i_2} \rangle_{\mathrm{Frob}} \Bigg\}.
\end{aligned}
$$

3.5.3 Implementation and numerical experiments

In the following we give a description of the implementation and present numerical experiments that complement the foregoing analysis and verify the theoretical results. We start by discussing the approximation of the random diffusion coefficient κ for $\omega \in \Omega$. In order to be able to conduct the numerical experiments, the random field $\kappa(x, \omega)$ has to be represented parametrically.

Karhunen-Loève expansion of κ

To this end, we suppose the random field

$$\kappa(x, \omega) \in L^2(\Omega; W^{1,\infty}(D))$$

admits a Karhunen-Loève expansion in terms of eigenpairs $(\lambda_k, \varphi_k)_{k=1}^{\infty}$ with respect to the underlying covariance operator. The covariance operator is a self-adjoint and compact integral operator with kernel C_κ given by

$$C_\kappa := \mathbb{E}[(\kappa - \mathbb{E}[\kappa]) \otimes (\kappa - \mathbb{E}[\kappa])]$$

or formally represented pointwise as

$$C_\kappa(x, x') := \mathbb{E}[(\kappa(x, \omega) - \mathbb{E}[\kappa](x)) \otimes (\kappa(x', \omega) - \mathbb{E}[\kappa](x'))].$$

We assume that the eigenfunctions φ_k are normalized in $L^2(D)$ and that the λ_k are ordered by decreasing magnitude. Then the Karhunen-Loève expansion of κ can be written as

$$\kappa(x,\omega) = \mathbb{E}[\kappa(x,\omega)] + \sum_{i=1}^{\infty} \sqrt{\lambda_i} Y_i(\omega)\varphi_i(x),$$

where the random coefficients $Y_i(\omega)$, $i = 1, 2, \dots$ are defined by the expression

$$Y_i(\omega) = \begin{cases} \dfrac{1}{\sqrt{\lambda_i}} \displaystyle\int_D (\kappa(x,\omega) - \mathbb{E}[\kappa](x))\varphi_i(x)\,\mathrm{d}x, & \lambda_i > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The given expansion converges in $L^2(\Omega; L^2(D))$ and emphasis is put on the fact, that in order to determine the Karhunen-Loève expansion of κ, explicit knowledge of the covariance kernel C_κ is mandatory.

The numerical experiments have been implemented in MATLAB and have been run on a laptop with an Intel i7-4720HQ at 2.6GHz with 16GB of RAM.

Numerical experiment for $d = 1$ on $D = [-1, 1]$

In this case the family of meshes $\mathcal{T}_\ell$ consists of intervals of the form

$$[-1 + (i-1)2^{-\ell}, -1 + i2^{-\ell}]$$

for $i = 1, \dots, 2^{\ell+1}$. The meshwidth is thus given by $h_\ell = 2^{-\ell}h_0 = 2^{-\ell}$ with a total number of $2^{\ell+1}$ elements on level ℓ. Consequently, we have a nested family of meshes $\{\mathcal{T}_\ell\}_{\ell=0}^{\infty}$ and the meshes are 1-shape regular. Since we only consider Dirichlet boundary conditions, there are no degrees of freedom on the boundary and we use the FE spaces $V_\ell = \mathcal{S}_0^{1,0}(\mathcal{T}_\ell)$ as specified earlier. For every ℓ the space V_ℓ is spanned by the usual basis of hat functions $\phi_i^\ell, i = 1, \dots, 2^{\ell+1} - 1$ and $\dim(V_\ell) = 2^{\ell+1} - 1$. The basis $(\phi_i^\ell)_{i=1}^{2^{\ell+1}-1}$ is explicitly given by

$$\phi_i^\ell(x) := \frac{1}{h_\ell} \begin{cases} x - (-1 + (i-1)h_\ell), & x \in [-1 + (i-1)h_\ell, -1 + ih_\ell], \\ (-1 + (i+1)h_\ell) - x, & x \in [-1 + ih_\ell, -1 + (i+1)h_\ell], \\ 0, & \text{otherwise.} \end{cases}$$

We adopt the following numerical example from [6, Example 5.1] and refer the reader to this article and references therein for more details. We set $f \equiv 1$ and specify the random diffusion coefficient κ as follows. Let $\mathbb{E}[\kappa](x) = 5 + x$ and the corresponding covariance kernel C_κ is chosen as

$$C_\kappa(x,y) = \frac{\min\{x,y\} + 1}{2} \in H^1(D) \otimes H^1(D).$$

It can be shown that the eigenpairs in the Karhunen-Loève expansion are given by

$$\lambda_k = \frac{8}{\pi^2(2k-1)^2}, \qquad \varphi_k(x) = \sin\left(\frac{x+1}{\sqrt{2\lambda_k}}\right), \qquad k \geq 1. \tag{3.61}$$

We truncate the Karhunen-Loève series for the random diffusion coefficient κ after the first term and redefine this truncated representation as the *exact* random diffusion coefficient κ_1. This is done to avoid an additional error in the simulation. Then with $\Upsilon_1 \sim \mathcal{N}(0,1)$,

i.e. Υ_1 is normally distributed with expectation zero and unit variance, we can get realizations of the stochastic diffusion coefficient by the following expression

$$\kappa_1(x,\omega) := 5 + x + \frac{2\sqrt{2}}{\pi}\Upsilon_1(\omega)\sin\left(\frac{\pi(x+1)}{4}\right).$$

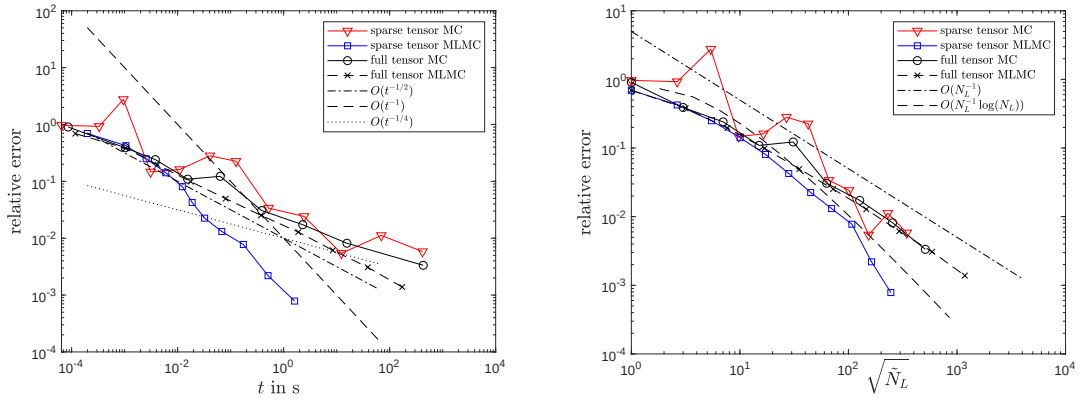

Figure 3.1: Plots of the relative error with respect to the reference solution on level $L = 14$ for the sparse tensor MLMC and the reference solution on level $L = 11$ for the full tensor MLMC against the elapsed time (left) and degrees of freedom (right), where $\tilde{N}_L$ denotes the amount of DOFs that are needed to represent the covariance function in the different cases.

To compute the covariance function with respect to u we proceed as follows. We compute solutions u_ℓ^i, $\ell = 1,\dots,L, i = 1,\dots,M$ for given realizations $\kappa(\cdot,\omega_i)$ of the random diffusion coefficient at level ℓ for the Monte Carlo methods and compute $M_\ell, \ell = 1,\dots,L$ samples of the level corrections for the MLMC methods. For each method we have chosen the optimal number of samples, i.e. for the Monte Carlo variants we have chosen $M \sim \varepsilon^{-2}$ and for the multilevel Monte Carlo methods we chosen the number of samples in a way to balance the computational cost against the accuracy of the method (cf. (3.49), (3.53) and (3.54)). As a solver we have used the MATLAB built-in backslash operator and we assume that since the resulting system matrix is tridiagonal that the complexity of the solver is given by $\gamma = 1$. In Figure 3.1 we have given two graphs, where one is showing the relative error against the runtime of the computation and the other shows the convergence rate of the methods by plotting the relative error against the number of degrees of freedom.

From the convergence graph it can be seen that the full tensor product methods clearly show a convergence rate of N_L^{-1}, which is expected from the theory as the FEM in 1D yields a convergence rate of one. In tune with the theory are also the error curves of the sparse tensor product MC and MLMC methods, which show the theoretically found convergence rate of $N_L^{-1}\log(N_L)$. For the full tensor product methods as well as the sparse tensor product methods we have chosen as a reference solution the solution of the FTP MLMC and STP MLMC, respectively. This means that the FTP MC is compared against the reference solution of the FTP MLMC method and the STP MC solutions are compared against the reference solution of the finest refinement level of the STP MLMC method.

For the parameters of the asymptotic cost bounds we note that $\delta = 1$ and as such we can see in the other graph of Figure 3.1 that the full tensor product MC is in accordance

with the theory. Here the predicted asymptotic cost is $\varepsilon^{-2-\max\{\gamma,2\}/\delta}$, which amounts to a ε^{-4} in this context and can be seen from the graph as the error curve becomes parallel to the $t^{-1/4}$ curve. The sparse tensor product MC method has a predicted asymptotic cost of $\varepsilon^{-2-\gamma/\delta} \cdot |\log(\varepsilon)|^{1+1/\delta}$ which amounts to $\varepsilon^{-3}|\log(\varepsilon)|^2$ in this situation. This rate is better than that of the FTP-MC and can be seen from the figure. Note that the methods have only been run once and as such the MC methods exhibit a fair amount of fluctuation. But this is no hindrance to show the expected convergence, also cf. [5]. Considering the full tensor product MLMC method as there holds $2\delta = \max\{\gamma, 2\}$ we have an expected asymptotic cost of $\varepsilon^{-2}|\log \varepsilon|^2 + \varepsilon^{-2}$ which is also seen in the graph. More precisely, we see that up to a logarithmic factor that the full tensor product MLMC method becomes parallel to the curve for $t^{-1/2}$. The last data point is a little bit misleading and can be explained by the use of a reference solution to compute the errors. For the sparse tensor product MLMC method we have an expected asymptotic cost of $\varepsilon^{-2} + \varepsilon^{-1}|\log(\varepsilon)|$ which is clearly visible as the blue curve becomes parallel with the curve for $t^{-1/2}$ before it shows even better behavior for the last two datapoints. This may still be an artefact of the fact that we have only performed one run as well as the use of a reference solution for the convergence history. As seen from the graph as well as from the ranking of section 3.4 is the fact that the sparse tensor product MLMC method outperforms the other methods clearly in terms of cost. Moreover, keep in mind that the sparse tensor product MLMC also has a more adavantageous asymptotic memory requirement as it is essentially linear and still achieves comparable accuracy.

Numerical experiment for $d = 2$ on $D = [-1, 1]^2$

Here, we consider the unit square and define the mesh on level $\ell = 0$ to be the set of triangles specified as $P_1P_2P_4$ and $P_2P_3P_4$ with the vertices $P_1 = (-1, -1), P_2 = (1, -1), P_3 = (1, 1), P_4 = (-1, 1)$.

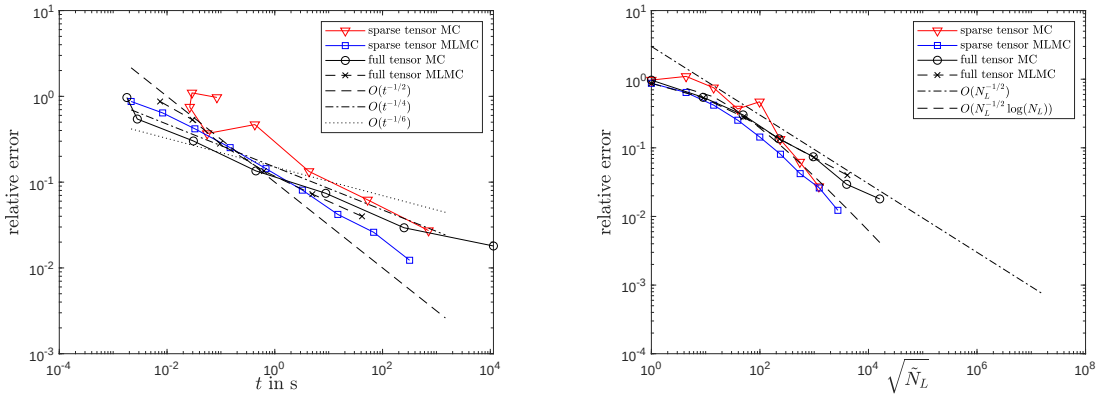

Figure 3.2: Plots of the relative error with respect to the reference solution on level $L = 10$ for the sparse tensor MLMC and the reference solution on level $L = 7$ for the full tensor MLMC against the elapsed time (left) and degrees of freedom (right), where $\tilde{N}_L$ denotes the amount of DOFs that are needed to represent the covariance function in the different cases.

Then we build $\mathcal{T}_1$ by uniform refinement of the two triangles in $\mathcal{T}_0$, such that $\mathcal{T}_1$ has one internal degree of freedom and consists of eight triangles. Solving on the mesh $\mathcal{T}_0$ is rendered superfluous by the homogeneous Dirichlet boundary conditions. The meshes $\mathcal{T}_\ell$ for $\ell > 1$ are constructed in the same way, such that $\mathcal{T}_{\ell+1}$ is obtained by uniform

refinement of all triangles in $\mathcal{T}_\ell$. Furthermore, we define the random diffusion coefficient κ by tensorization of the one dimensional random diffusion coefficient κ_1 of the previous section

$$\kappa(x, y, \omega) = \kappa_1(x, \omega)\kappa_1(y, \omega).$$

For the computation of the covariance function in this setting we proceed as for the one dimensional experiment, i.e. we compute the optimal amount of samples for each method in question. We have again used the solution on the finest grind of the FTP MLMC method as a reference soution for the full tensor product methods and the solution of the STP MLMC method on finest refinement level as a reference solution for the sparse tensor product methods.

The convergence graph (cf. Figure 3.2) shows that the theoretically predicted convergence rates of the methods are verified. In particular, since we have $\delta = 1/2$ and assume $\gamma = 1$, although the backslash operator might exhibit slightly non-optimal behaviour, the convergence rate of the full tensor product MC and MLMC method on level L is given by $N_L^{-1/2}$, whereas the sparse tensor product methods show a convergence of $N_L^{-1/2}\log(N_L)$.

Regarding the asymptotic costs of the method we find according to Figure 3.2 and theoretical prediction that the full tensor product MC method has an asymptotic cost of ε^{-6}. For the full tensor product MLMC method, since $2\delta < \max\{\gamma, 2\}$, we find that the theoretical asymptotic cost is given by ε^{-4}. This can also be seen from the graph as the curve becomes almost parallel with $t^{-1/4}$ curve. For the sparse tensor product MC the theory predicts an asymptotic cost of $\varepsilon^{-4}|\log(\varepsilon)|^3$, when assuming $\gamma = 1$, which is also visible from the graph. In case of the sparse tensor product MLMC method, because $2\delta = 1$, we expect an asymptotic cost of $\varepsilon^{-2}|\log(\varepsilon)|^5 + \varepsilon^{-2}|\log(\varepsilon)| \sim \varepsilon^{-2}|\log(\varepsilon)|^5$, which is clearly seen from Figure 3.2 as well as the logarithmic factor.

In the two dimensional situation we see that again the sparse tensor product MLMC method noticeably outperforms the other methods, which is already apparent from the asymptotic costs, but is much more pronounced than for the one dimensional model problem.

Chapter 4

Comparison of DME and MLMC approximation

4.1 1D model problem

We are interested in solving the *deterministic moment equation* (DME)

$$\begin{aligned} \partial_x^2 \partial_y^2 \mathcal{C}_u &= \mathcal{C}_f, \quad \text{in } \mathcal{D}, \\ \mathcal{C}_u &= 0, \quad \text{on } \partial\mathcal{D}, \end{aligned}$$

on $\mathcal{D} = D \times D$ with $D = [-1, 1]$, where $\mathcal{C}_u(x, y) = \mathbb{E}[u(x, \omega) \otimes u(y, \omega)]$ and $\mathcal{C}_f(x, y) = \mathbb{E}[f(x, \omega) \otimes f(y, \omega)]$. This problem follows from the tensorization of the following stochastic elliptic model problem: Find $u(x, \omega) \in L^2(\Omega; H^1(D))$, such that for a.e. $\omega \in \Omega$

$$\begin{aligned} -\partial_x^2 u(x, \omega) &= f(x, \omega), \quad \text{in } D, \\ u &= 0, \quad \text{on } \partial D. \end{aligned}$$

Suppose that f is given as a Karhunen-Loève expansion

$$f(x, \omega) = \overline{f}(x) + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \Upsilon_i(\omega) \varphi_i(x)$$

where $\overline{f}(x) = \mathbb{E}[f(x, \omega)]$, $\mathbb{E}[\Upsilon_i] = 0$, $\mathbb{E}[\Upsilon_i^2] = 1, \forall i$, and $(\lambda_i, \varphi_i(x))$ are the eigenpairs of the associated covariance kernel $\mathcal{C}_f$, i.e.

$$\int_D \mathcal{C}_f(x, y) \varphi_i(x) \, \mathrm{d}x = \lambda_i \varphi_i(y),$$

where it is assumed that $\int_D \varphi_i \varphi_j = \delta_{ij}$, i.e. the φ_i are mutually orthonormal.

In order to compare the methods we have analyzed in this thesis, we would like to find a representation of the exact solution $\mathcal{C}_u(x, y)$. To this end we suppose that the random field $u(x, \omega)$ has a representation of the form

$$u(x, \omega) = \overline{u}(x) + \sum_{i=1}^{\infty} \Upsilon_i(\omega) \psi_i(x)$$

with certain functions ψ_i and random variables Υ_i. To this end, we see that there must

hold

$$\begin{aligned} -\partial_x^2 u(x,\omega) =& -\partial_x^2 \overline{u} - \sum_{i=1}^{\infty} \Upsilon_i(\omega)\partial_x^2\psi_i(x) \\ \overset{!}{=}& \overline{f}(x) + \sum_{i=1}^{\infty} \sqrt{\lambda_i}\Upsilon_i(\omega)\varphi_i(x) \\ =& f(x,\omega) \end{aligned}$$

and therefore that

$$\begin{aligned} -\partial_x^2\overline{u}(x) &= \overline{f}(x), \quad \text{in } D, \\ \overline{u} &= 0, \quad \text{on } \partial D, \end{aligned}$$

and for all i the ψ_i's have to statisfy the following boundary value problem

$$\begin{aligned} -\partial_x^2\psi_i(x) &= \sqrt{\lambda_i}\varphi_i(x), \quad \text{in } D, \\ \psi_i &= 0, \quad \text{on } \partial D. \end{aligned} \tag{4.1}$$

If moreover $\overline{f} = 0$, then we find that $\overline{u} = 0$ and therefore the covariance functions $\mathcal{C}_u$ and $\mathcal{C}_f$ can be written as follows:

$$\begin{aligned} \mathcal{C}_u(x,y) &= \sum_{i=1}^{\infty} \psi_i(x)\psi_i(y), \\ \mathcal{C}_f(x,y) &= \sum_{i=1}^{\infty} \lambda_i\varphi_i(x)\varphi_i(y). \end{aligned}$$

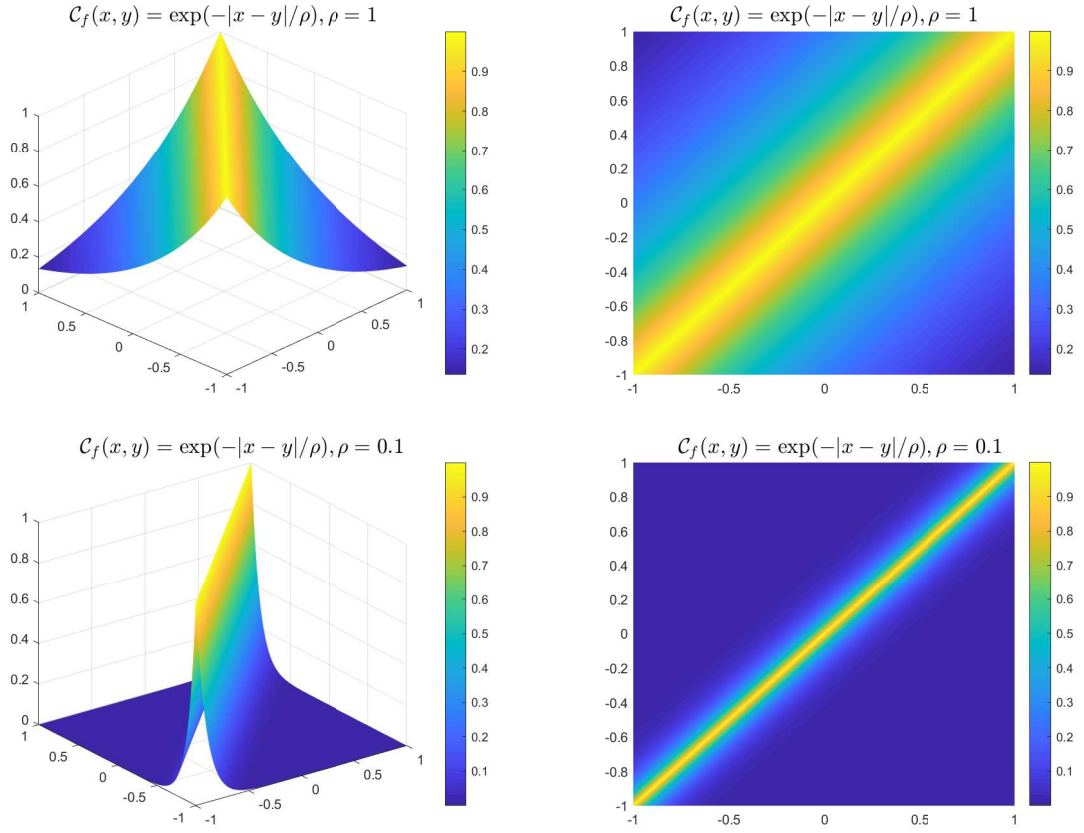

Figure 4.1: Exponential covariance function for diferent values of ρ.

In order to illustrate the advantages and disadvantages of the methods investigated in this thesis, we choose a representative example. Let us suppose that $\mathcal{C}_f$ is given as the exponential covariance function expressed by

$$\mathcal{C}_f = \exp(-|x-y|/\rho),$$

where ρ again denotes the correlation length. A plot of this covariance function for $\rho \in \{1, 0.1\}$ can be found in Figure 4.1. It can be easily seen that this covariance is strongly concentrated towards the diagonal of the domain for small values of ρ. For this case of $\mathcal{C}_f$ with $D = [-a, a]$ the eigenpairs (λ_i, φ_i) are given by (cf. [32])

$$\begin{aligned} \lambda_i &= \frac{2\rho}{1+\rho^2\omega_i^2}, & \lambda_i^* &= \frac{2\rho}{1+\rho^2(\omega_i^*)^2}, \\ \varphi_i(x) &= \frac{\cos(\omega_i x)}{\sqrt{a+\frac{\sin(2\omega_i a)}{2\omega_i}}}, & \varphi_i^*(x) &= \frac{\sin(\omega_i^* x)}{\sqrt{a-\frac{\sin(2\omega_i^* a)}{2\omega_i^*}}}. \end{aligned} \tag{4.2}$$

In Figure 4.2 we have depicted the decay of eigenvalues for different values of ρ. For this example the eigenvalues decay quadratically after a possible "shelf" of eigenvalues of the same magnitude. This is especially clear for the smaller values of ρ.

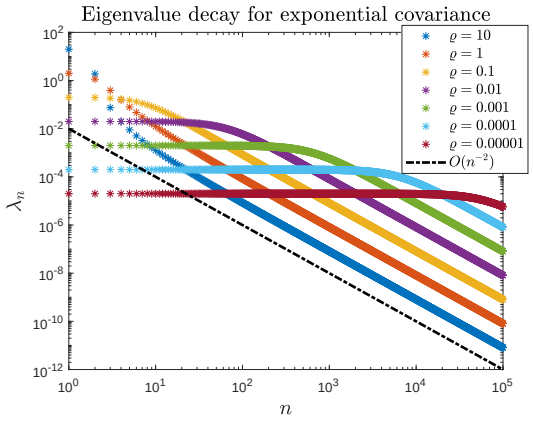

Figure 4.2: Eigenvalue decay for different values of ρ

The associated eigenfunctions multiplied by their respective eigenvalues have been plotted in Figure 4.3 for $\rho = 1$. We see that the φ_i's are even and that the φ_i^*'s are odd functions. Equipped with these eigenpairs we can write

$$\begin{aligned} \mathcal{C}_f(x,y) = &\sum_{i=1}^{\infty} \frac{2\rho}{1+\rho^2\omega_i^2} \frac{\cos(\omega_i x)}{\sqrt{a+\frac{\sin(2\omega_i a)}{2\omega_i}}} \\ &+ \sum_{i=1}^{\infty} \frac{2\rho}{1+\rho^2(\omega_i^*)^2} \frac{\sin(\omega_i^* x)}{\sqrt{a-\frac{\sin(2\omega_i^* a)}{2\omega_i^*}}}, \end{aligned}$$

where the ω_i and ω_i^* are given by the two following transcendental equations for ω and ω^*, respectively:

$$\frac{1}{\rho} - \omega\tan(\omega a) = 0, \quad \omega^* + \frac{1}{\rho}\tan(\omega^* a) = 0.$$

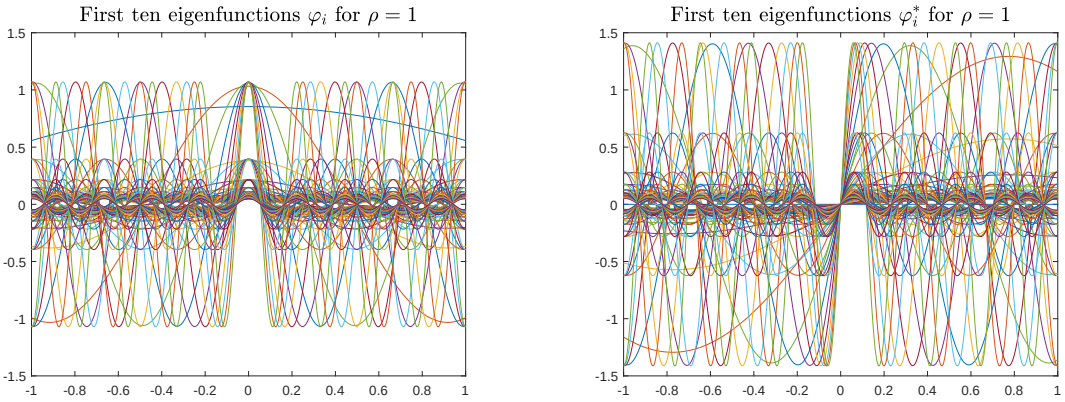

Figure 4.3: Eigenfunctions of the exponential covariance for $\rho = 1$ multiplied with their respective eigenvalues.

Solving the ordinary differential equation (4.1) for the ψ_i's and ψ_i^*'s, respectively, leads to the following representations

$$\psi_i = \frac{\sqrt{\lambda_i}}{\omega_n^2 \sqrt{a + \frac{\sin(2\omega_n a)}{2\omega_n}}} \left(\cos(\omega_n x) - \cos(\omega_n a)\right),$$

$$\psi_i^* = \frac{\sqrt{\lambda_i}}{(\omega_n^*)^2 \sqrt{a - \frac{\sin(2\omega_n^* a)}{2\omega_n^*}}} \left(\sin(\omega_n^* x) - \frac{\sin(\omega_n^* a))}{a} x\right).$$

One readily verfies that the ψ_i's fulfill the boundary value problem (4.1) in D and have zero boundary values on ∂D. This implies the following representation of $\mathcal{C}_u$ as

$$\mathcal{C}_u(x, y) = \sum_{i=1}^{\infty} \left(\psi_i(x)\psi_i(y) + \psi_i^*(x)\psi_i^*(y)\right). \tag{4.3}$$

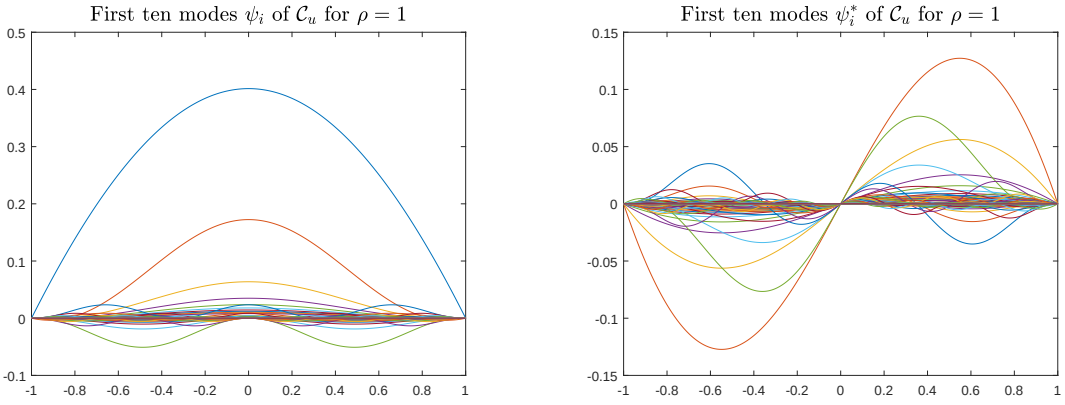

Figure 4.4: Modes of $\mathcal{C}_u$

So far we only have one parameter to manipulate the behavior of the covariance function $\mathcal{C}_f$, which is through the correlation length ρ. Moreover, we can control the decay of the eigenvalues in the Karhunen-Loève expansion by replacing λ_n and λ_n^*, respectively, with

$$\tilde{\lambda}_n = \lambda_n^\gamma, \quad \tilde{\lambda}_n^* = (\lambda_n^*)^\gamma, \quad \forall n = 1, 2, \ldots$$

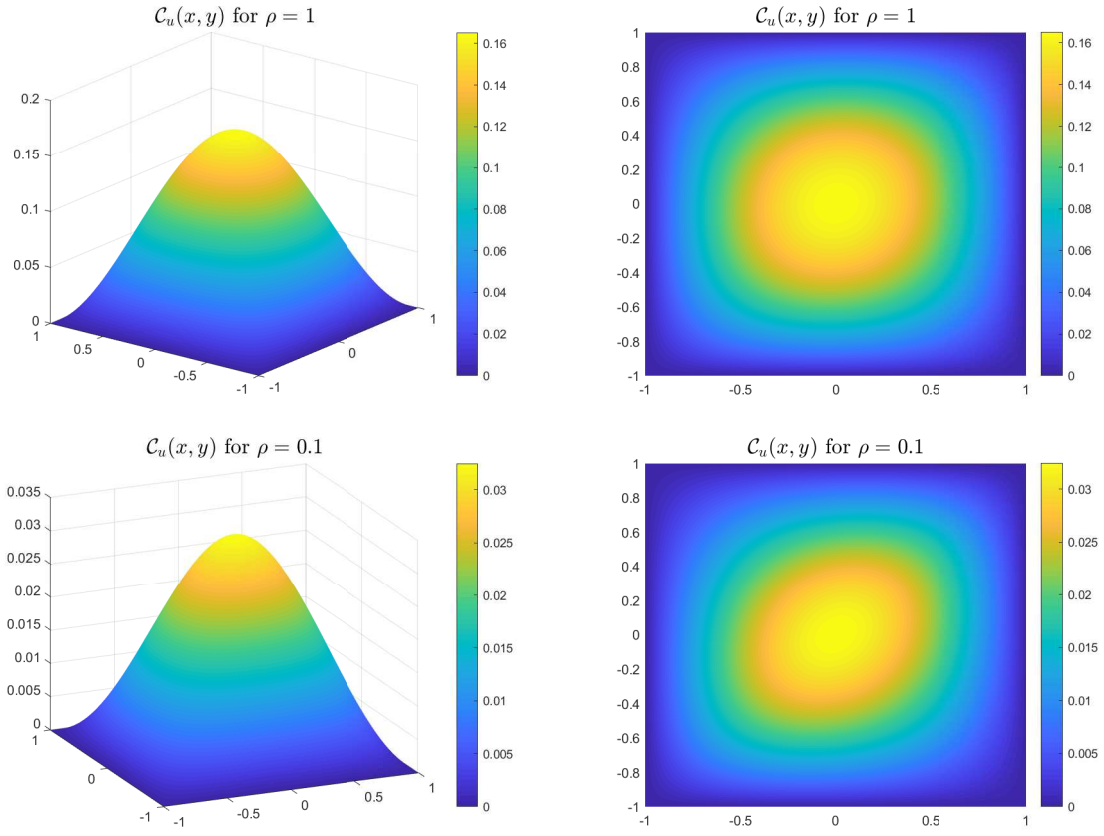

Figure 4.5: Plots of the exact solution $\mathcal{C}_u$ of the DME for different values of ρ.

By defintion of the λ_n and λ_n^* (cf. (4.2)) this then leads to an algebraic decay of the eigenvalues as $O(n^{-2\gamma})$.

Test suite 1 ($\gamma = 1, \rho \in \{1, 0.1\}$) As the title suggests here we let $\gamma = 1$, such that the decay of the eigenvalues for the covariance function of f is equal to two. Furthermore, we choose $\rho = 1$ and $\rho = 0.1$ and run all methods for the problem with exponential covariance function and the given parameters. We expect the error of the adaptive Finite Element Methods (AFEM) to exhibit a convergence rate of $\tilde{N}_L^{-1/2}$, where $\tilde{N}_L$ denotes the dimension of the global FE space, which is used to represent the approximate covariance function. For the full and sparse tensor product methods we expect a convergence rate of the error as N_L^{-1} and $N_L^{-1}\log(N_L)$, respectively, where N_L denotes the dimension of the approximation space for the samples of X_L and Y_L. The degrees of freedom to represent the covariance function for the FTP and STP MC and MLMC methods is also denoted as $\tilde{N}_L$ and is of $O(N_L^2)$ and $O(N_L \log(N_L))$ magnitude, respectively. The amount of samples taken for the MC and MLMC methods has not been adjusted to the correlation length of the covariance kernels of the given experiments. Moreover, insofar the given comparison is meant to be more of an "out of the box"-investigation of the competetiveness of the presented methods.

For the MC and MLMC method we have chosen to truncate the Karhunen-Loève expansion (KLE) after a hundred terms (i.e. $M_{\mathrm{KL}} = 100$) as this is sufficient for the range in which we are comparing the methods. Also this has been done to avoid unnecessary errors introduced due to the high oscillatory nature of the eigenfunctions in the KLE with respect to the mesh width. If one wants to take more terms, one has to adjust for better

quadrature in the solver. This leads to a longer runtime proportional to the amount of work added, either via an additional amount of quadrature or the need to compute more terms in the KLE. For the comparison we have chosen for the STP MC and MLMC as reference solution the solution on the finest grid of STP MLMC method, which is not shown in the graphs, and analogously taken the solution on the finest refinement level of the FTP MLMC method as reference solution for the FTP MC and MLMC methods.

In Figure 4.6 we find the convergence history of the numerical experiments for $\rho \in \{0.1, 1\}$. For $\rho = 1$ we can clearly see that the $H^{1,1}$-seminrom error for the FTP and STP

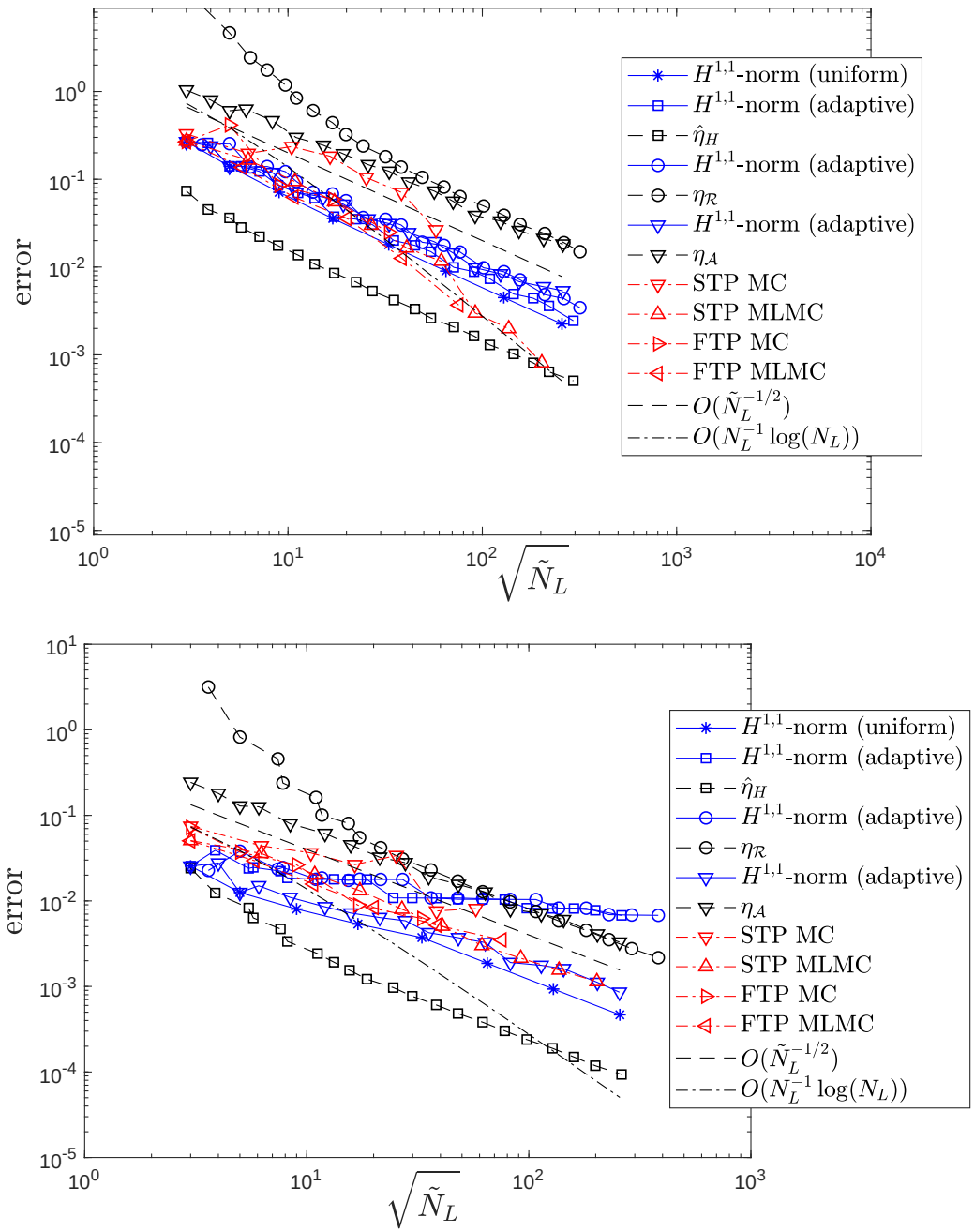

Figure 4.6: Convergence history for all methods and varying ρ. top: $\rho = 1$. bottom: $\rho = 0.1$.

MC and MLMC methods converge at the expected rate of $O(N_L^{-1})$ and $O(N_L^{-1}\log(N_L))$, respectively. This clearly outperforms the AFEM for any configuration and even the uniform FEM as their rate is clearly seen to converge at $O(\tilde{N}_L^{-1/2})$. This shows that Monte Carlo methods are very well suited to problems where the underlying random process features a high degree of correlation.

For $\rho = 0.1$ we observe the convergence shelf of the AFEM for the residual and hierarchical estimator, such that these cannot attain a convergence rate of the error of $O(\tilde{N}_L^{-1/2})$. The AFEM guided by $\eta_{\mathcal{A}}$ performs at the optimal rate, albeit it cannot beat the uniform FEM here. Furthermore, it can be observed that the error in the $H^{1,1}$-seminorm fails to converge at the expected rate for the FTP and STP MC and MLMC methods. This behavior is observed with excessively high quadrature, or even adaptive Gauß-Kronrod type quadrature and is also not alleviated by taking more terms in the KLE. This effect of the correlation length versus decay of eigenvalue decay for the effectiveness of the methods can also be observed in the next test suite.

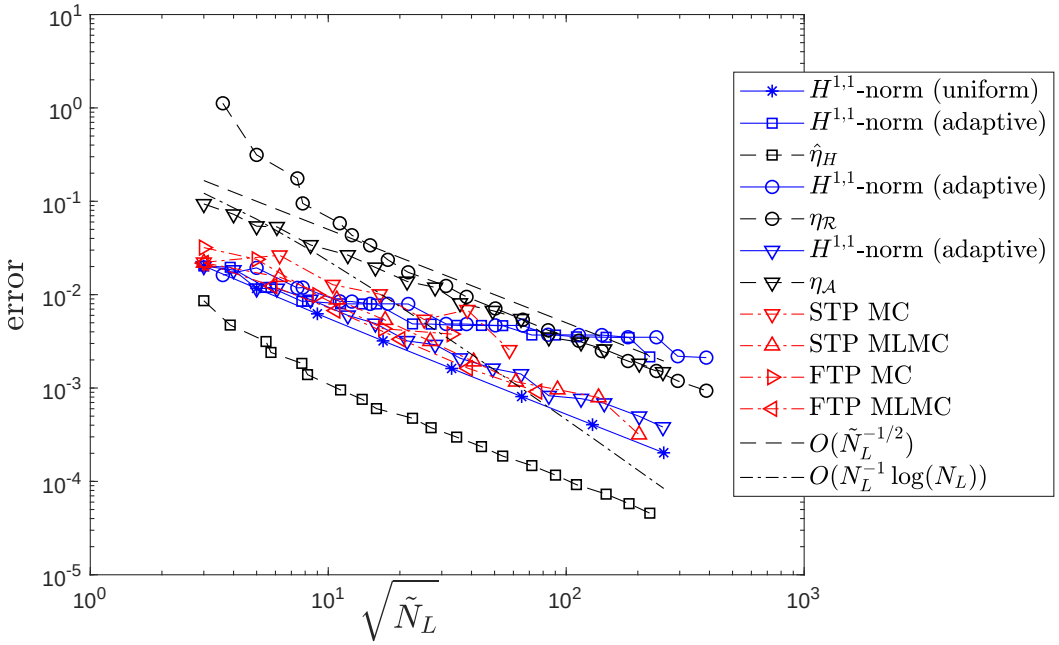

Figure 4.7: Convergence history for all methods and $\gamma = 3/2$.

Test suite 2 ($\gamma \in \{3/2, 5/2\}, \rho = 0.1$) In order to find out if the effect of the correlation is connected to the decay of the eigenvalue we choose $\gamma \in \{3/2, 5/2\}$ for $\rho = 0.1$.

For the MC and MLMC method we have again chosen to truncate the Karhunen-Loève expansions after a hundred terms (i.e. $M_{\mathrm{KL}} = 100$).

For $\gamma = 3/2$ and as such for a rate of decay of 3 of the eigenvalues, we see in Figure 4.7 that the full tensor and sparse tensor MC and MLMC methods cannot attain their theoretically predicted convergence rates in the range of comparison. Here it is again seen that the AFEM and FTP and STP MC and MLMC methods are very competetive. The AFEM seems to perform better for problems with a low degree of correlation.

In the graph of Figure 4.8 we see that for $\gamma = 5/2$ the FTP and STP MC and MLMC are able to beat the AFEM again in the range of comparison and attain their theoretically predicted error convergence rate of $O(N_L^{-1})$ and $O(N_L^{-1}\log(N_L))$, respectively. Here the rate of decay of the eigenvalues is 5 which indicates a highly correlated process. Moreover,

we observe that the AFEM behaves almost identically in both cases for all types of error estimators for both choices of γ. Furthermore, it seems in both tests that the correlation length has more influence on the MC methods than on the MLMC methods.

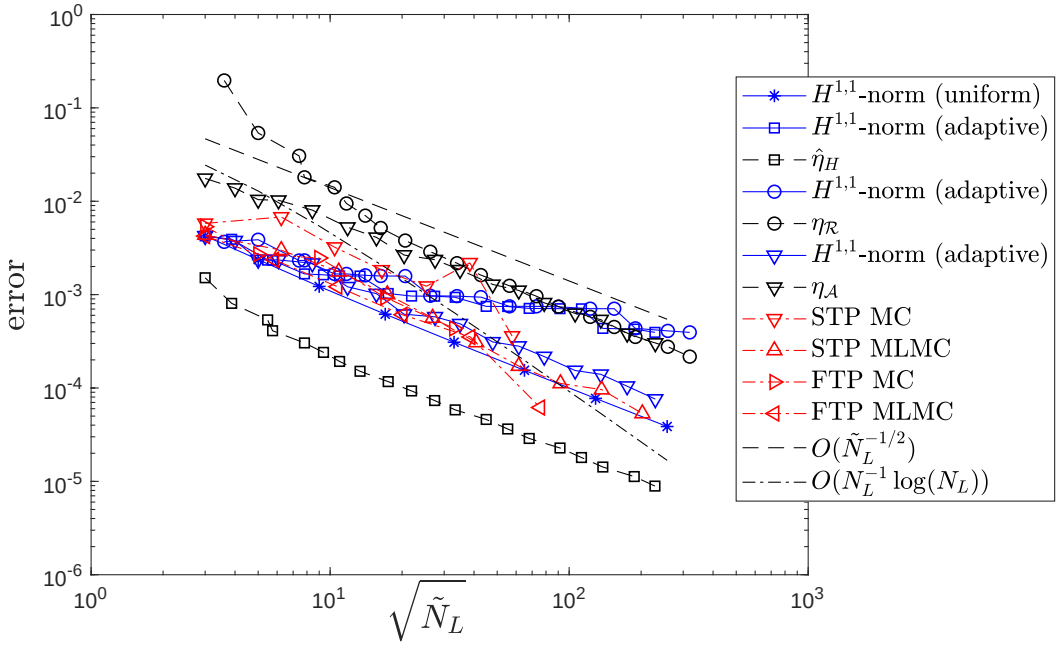

Figure 4.8: Convergence history for all methods and $\gamma = 5/2$.

4.2 2D model problem

Here we are interested in solving the DME

$$\begin{aligned}(\Delta_x \otimes \Delta_y)\mathcal{C}_u &= \mathcal{C}_f, \quad \text{in } \mathcal{D},\\ \mathcal{C}_u &= 0, \quad \text{on } \partial\mathcal{D},\end{aligned}$$

on $\mathcal{D} = D \times D$ with $D = [-1,1]^2$ to compare the analyzed methods. Let us first make the following observation. Let us assume that

$$\mathcal{C}_f(x,y) := \mathcal{C}_{f_1}(x_1,y_1)\mathcal{C}_{f_2}(x_2,y_2)$$

is given as a tensorized covariance kernel, where $x = (x_1,x_2)$ and $y = (y_1,y_2)$. If we consider

$$\mathcal{C}_{f_k}(x_k,y_k) := \exp(-|x_k - y_k|/\rho_k), \quad k = 1,2$$

then

$$\mathcal{C}_{f_k}(x_k,y_k) = \sum_{i=1}^{\infty} \left\{\lambda_{k,i}\varphi_{k,i}(x_k)\varphi_{k,i}(y_k) + \lambda_{k,i}^*\varphi_{k,i}^*(x_k)\varphi_{k,i}^*(y_k)\right\}.$$

This leads to the following representation of $\mathcal{C}_f$ as a double sum via

$$\begin{aligned}
&\mathcal{C}_f(\vec{x},\vec{y}) = \mathcal{C}_{f_1}(x_1,y_1)\mathcal{C}_{f_2}(x_2,y_2)\\
&= \sum_{i,j=1}^{\infty} \Big\{ \lambda_{1,i}\lambda_{2,j}\varphi_{1,i}(x_1)\varphi_{1,i}(y_1)\varphi_{2,j}(x_2)\varphi_{2,j}(y_2) + \lambda_{1,i}\lambda_{2,j}^*\varphi_{1,i}(x_1)\varphi_{1,i}(y_1)\varphi_{2,j}^*(x_2)\varphi_{2,j}^*(y_2)\\
&\qquad + \lambda_{1,i}^*\lambda_{2,j}\varphi_{1,i}^*(x_1)\varphi_{1,i}^*(y_1)\varphi_{2,j}(x_2)\varphi_{2,j}(y_2) + \lambda_{1,i}^*\lambda_{2,j}^*\varphi_{1,i}^*(x_1)\varphi_{1,i}^*(y_1)\varphi_{2,j}^*(x_2)\varphi_{2,j}^*(y_2)\Big\}\\
&= \sum_{i,j=1}^{\infty} \Big\{ \lambda_{1,i}\lambda_{2,j}\varphi_{1,i}(x_1)\varphi_{2,j}(x_2)\varphi_{1,i}(y_1)\varphi_{2,j}(y_2) + \lambda_{1,i}\lambda_{2,j}^*\varphi_{1,i}(x_1)\varphi_{2,j}^*(x_2)\varphi_{1,i}(y_1)\varphi_{2,j}^*(y_2)\\
&\qquad + \lambda_{1,i}^*\lambda_{2,j}\varphi_{1,i}^*(x_1)\varphi_{2,j}(x_2)\varphi_{1,i}^*(y_1)\varphi_{2,j}(y_2) + \lambda_{1,i}^*\lambda_{2,j}^*\varphi_{1,i}^*(x_1)\varphi_{2,j}^*(x_2)\varphi_{1,i}^*(y_1)\varphi_{2,j}^*(y_2)\Big\}.
\end{aligned} \tag{4.4}$$

If we assume, without loss of generality, that $\mathbb{E}[f(x_1,x_2,\omega)] = 0$, then f can be written as

$$\begin{aligned}
f(x_1,x_2,\omega) &= \left(\sum_{i=1}^{\infty} \sqrt{\lambda_{1,i}}\Upsilon_{1,i}(\omega)\varphi_{1,i}(x_1) + \sqrt{\lambda_{1,i}^*}\Upsilon_{1,i}^*(\omega)\varphi_{1,i}^*(x_1)\right)\\
&\qquad \times \left(\sum_{j=1}^{\infty} \sqrt{\lambda_{2,j}}\Upsilon_{2,j}(\omega)\varphi_{2,j}(x_2) + \sqrt{\lambda_{2,j}^*}\Upsilon_{2,j}^*(\omega)\varphi_{2,j}^*(x_2)\right)\\
&= \sum_{i,j=1}^{\infty} \Bigg\{ \sqrt{\lambda_{1,i}\lambda_{2,j}}\Upsilon_{1,i}(\omega)\Upsilon_{2,j}(\omega)\varphi_{1,i}(x_1)\varphi_{2,j}(x_2)\\
&\qquad + \sqrt{\lambda_{1,i}\lambda_{2,j}^*}\Upsilon_{1,i}(\omega)\Upsilon_{2,j}^*(\omega)\varphi_{1,i}(x_1)\varphi_{2,j}^*(x_2)\\
&\qquad + \sqrt{\lambda_{1,i}^*\lambda_{2,j}}\Upsilon_{1,i}^*(\omega)\Upsilon_{2,j}(\omega)\varphi_{1,i}^*(x_1)\varphi_{2,j}(x_2)\\
&\qquad + \sqrt{\lambda_{1,i}^*\lambda_{2,j}^*}\Upsilon_{1,i}^*(\omega)\Upsilon_{2,j}^*(\omega)\varphi_{1,i}^*(x_1)\varphi_{2,j}^*(x_2)\Bigg\}
\end{aligned}$$

where the $\Upsilon_{1,i}, \Upsilon_{1,i}^*, \Upsilon_{2,j}, \Upsilon_{2,j}^*, \forall i,j$ are assumed to be independent identically distributed random variables with zero mean and unit variance. Comparing $\mathbb{E}[f(x_1,x_2,\omega)\otimes f(y_1,y_2,\omega)]$ with (4.4) shows that f has $\mathcal{C}_f(x,y)$ as covariance function. This enables us to draw samples from the previous representation by assuming, *exempla gratia*

$$\Upsilon_{1,i}, \Upsilon_{1,i}^*, \Upsilon_{2,j}, \Upsilon_{2,j}^* \sim \mathcal{N}(0,1)$$

or by setting

$$\Upsilon_{k,i} = \sqrt{3}\Theta_{k,i}^*, \Upsilon_{k,i}^* = \sqrt{3}\Theta_{k,i}^*, \qquad k = 1,2,\forall i$$

where $\Theta_{k,\ell} \sim \mathcal{U}([-1,1])$ and $\Theta_{k,\ell}^* \sim \mathcal{U}([-1,1]), k = 1,2,\forall i$ are uniformly distributed on the interval $[-1,1]$ and thus the $\Upsilon_{k,\ell}$'s and $\Upsilon_{k,\ell}^*$'s have mean zero and unit variance.

As $\mathcal{C}_f$ has been constructed as the tensor product of two exponential covariance functions one might feel the urge to presume that maybe $\mathcal{C}_u$ can be written as a tensor product as well. Assuming $\mathcal{C}_u(x,y) = \mathcal{C}_{u_1}(x_1,y_1)\mathcal{C}_{u_2}(x_2,y_2)$ leads to

$$\begin{aligned}
(\Delta_x \otimes \Delta_y)\mathcal{C}_u(x,y) &= \partial_{x_1}^2\partial_{y_1}^2\mathcal{C}_{u_1}(x_1,y_1)\mathcal{C}_{u_2}(x_2,y_2) + \partial_{x_1}^2\mathcal{C}_{u_1}(x_1,y_1)\partial_{y_2}^2\mathcal{C}_{u_2}(x_2,y_2)\\
&\quad + \partial_{y_1}^2\mathcal{C}_{u_1}(x_1,y_1)\partial_{x_2}^2\mathcal{C}_{u_2}(x_2,y_2) + \mathcal{C}_{u_1}(x_1,y_1)\partial_{x_2}^2\partial_{y_2}^2\mathcal{C}_{u_2}(x_2,y_2)\\
&\neq \mathcal{C}_{f_1}(x_1,y_1)\mathcal{C}_{f_2}(x_2,y_2)
\end{aligned}$$

and assuming $\mathcal{C}_u(x,y) = \mathcal{C}_{u_1}(x_1,x_2)\mathcal{C}_{u_2}(y_1,y_2)$ yields

$$\begin{aligned}(\Delta_x \otimes \Delta_y)\mathcal{C}_u(x,y) &= \Delta_x \mathcal{C}_{u_1}(x_1,x_2)\Delta_y \mathcal{C}_{u_2}(y_1,y_2)\\ &= \partial^2_{x_1}\mathcal{C}_{u_1}(x_1,x_2)\partial^2_{y_1}\mathcal{C}_{u_2}(y_1,y_2) + \partial^2_{x_1}\mathcal{C}_{u_1}(x_1,x_2)\partial^2_{y_2}\mathcal{C}_{u_2}(y_1,y_2)\\ &\quad + \partial^2_{x_2}\mathcal{C}_{u_1}(x_1,x_2)\partial^2_{y_1}\mathcal{C}_{u_2}(y_1,y_2) + \partial^2_{x_2}\mathcal{C}_{u_1}(x_1,x_2)\partial^2_{y_2}\mathcal{C}_{u_2}(y_1,y_2)\\ &\neq \mathcal{C}_{f_1}(x_1,y_1)\mathcal{C}_{f_2}(x_2,y_2)\end{aligned}$$

where $\mathcal{C}_{u_1}$ and $\mathcal{C}_{u_2}$ are defined by (4.3) for given $\mathcal{C}_{f_1}$ and $\mathcal{C}_{f_2}$. Thus, even for this simple example in two dimensions of a tensorized covariance function $\mathcal{C}_f$ leads to a non-tensorizable covariance function $\mathcal{C}_u$ for u.

Here by construction we have a rate of decay of the eigenvalues of the corresponding covariance operator as 4 and as such a quite highly correlated process. We expect the FTP and STP MC and MLMC methods to perform at their theoretical convergence rates. This is why we have again chosen to investigate the effect of the correlation length on the behaviour of all methods. For this we choose the values for ρ_1 and ρ_2 independently.

For the MC and MLMC method we have again chosen to truncate the Karhunen-Loève expansion after a hundred terms (i.e. $M_{\mathrm{KL}} = 100$) as it is sufficient for the range in which we want to compare the methods in question. Also this has been done to avoid unnecessary errors introduced due to the high oscillatory nature of the eigenfunctions in the KLE with respect to the mesh width. These are now two dimensional and the problems of quadrature can also otherwise dominate the overall runtime. We have chosen 4096 quadrature points on each element for the AFEM and the uniform FEM in four dimensions as otherwise the method was not stable(!). If one wants to take more terms, one must adjust for better quadrature in the solver.

For the comparison we have chosen for the STP MC and MLMC as reference solution the solution on the finest grid of STP MLMC method, which is not shown in the graphs, and analogously taken the solution on the finest refinement level of the FTP MLMC method as reference solution for the FTP MC and MLMC methods. As such only the error level may be compared. The accuracy is better of course for the deterministic methods as the MC and MLMC methods can only deliver more accurate results when the results are averaged for a growing number of runs.

The AFEM has been run with $\vartheta = 0.5$ and the maximum strategy for marking the elements has been used. Moreover, we have implemented a stopping criterion of 50'000 elements, because of the high amount of quadrature points per element (4096 points), which are the result of a tensorized Gauß-Legendre qudrature rule with 8 points in each direction. As a reference solution for the AFEM and the uniform FEM we used the uniform FEM solution on a grid of 16^5 four dimensional cubes in all cases.

Test 1($\gamma = 1, \rho_1 = 1, \rho_2 = 1$) For this experiment we see the same situation as for the 1D model problem. The FTP and STP MLMC are able to outperform the adaptive methods in the range of comparison, which can be seen in Figure 4.9. The STP and FTP MC suffer from fluctuation and are in general more susceptible to outliers and thus the convergence in this early range may rather be chaotic. The FTP MC shows a rather preasymptotic behavior at first whereas the STP MC may just have had some "bad samples". Ignoring the last two data points of the STP MC curve the predicted behavior is visible. These fluctuations of course would be flattened if one would only do enough runs for this comparison. But since for this comparison one run of the MC and MLMC methods for the tests 1,2 and 3 takes a total of roughly 12 and a half hours, we have not pursued this as the tests are indicative as they are already. The AFEM then still has to

be run for the comparison, but in comparison is negligible although it also takes in total roughly 4 hours to run.

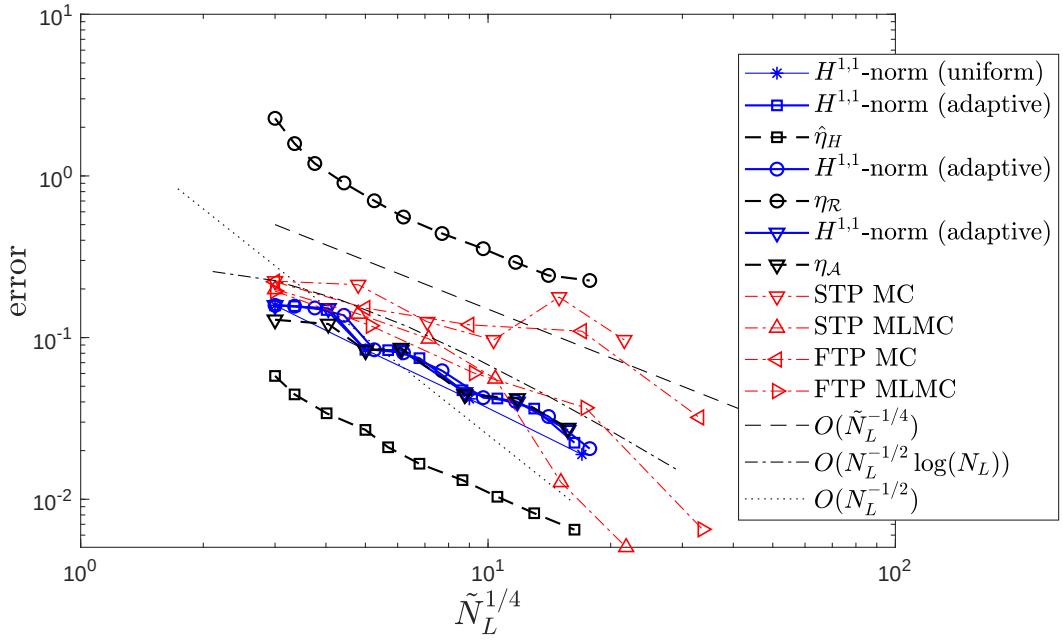

Figure 4.9: Convergence history for tensorized exponential covariance with $\rho_1 = 1$ and $\rho_2 = 1$.

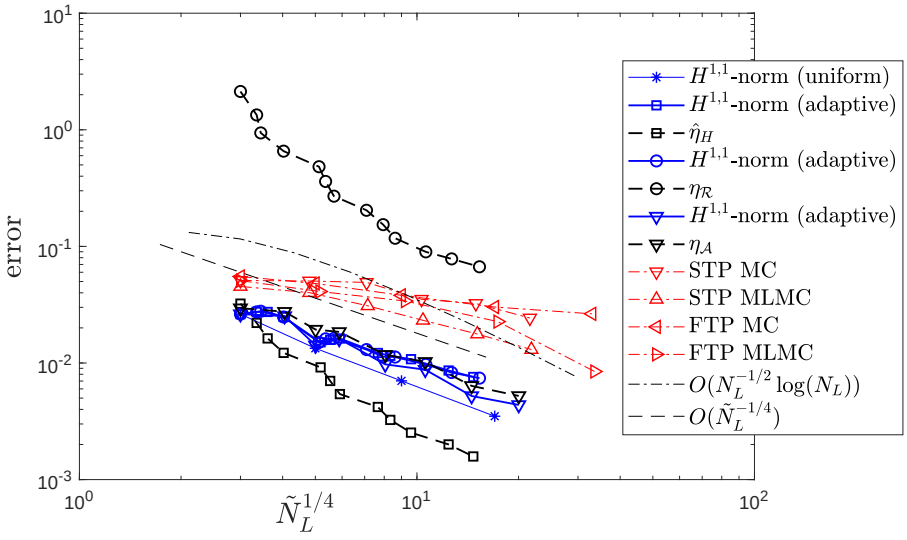

Figure 4.10: Convergence history for tensorized exponential covariance with $\rho_1 = 1$ and $\rho_2 = 0.1$.

Test 2($\gamma = 1, \rho_1 = 1, \rho_2 = 0.1$) In this experiment (cf. Figure 4.10) we again find that lowering the correlation length, this time just in one of the factors, results in a deterioration of the convergence rate of the MC and MLMC methods. Here the AFEM

and uniform FEM for the deterministic moment equations (DME) are very competetive and beat the MC/MLMC mehtods outright. As observed before the AFEM behaves more favorably if covariance kernels with a low degree of correlation are involved.

Test 3($\gamma = 1, \rho_1 = 0.1, \rho_2 = 0.1$) As a last experiment in this direction we have lowered both correlation lengths. From Figure 4.11 we infer that the convergence in the range of comparison is further deteriorated for the MC and MLMC methods. The AFEM and uniform FEM for the DME converge nonetheless, although it seems that the AFEM guided by the error indicators $\hat{\eta}_H$ and $\eta_{\mathcal{R}}$ are experiencing a convergence shelf phenomenon.

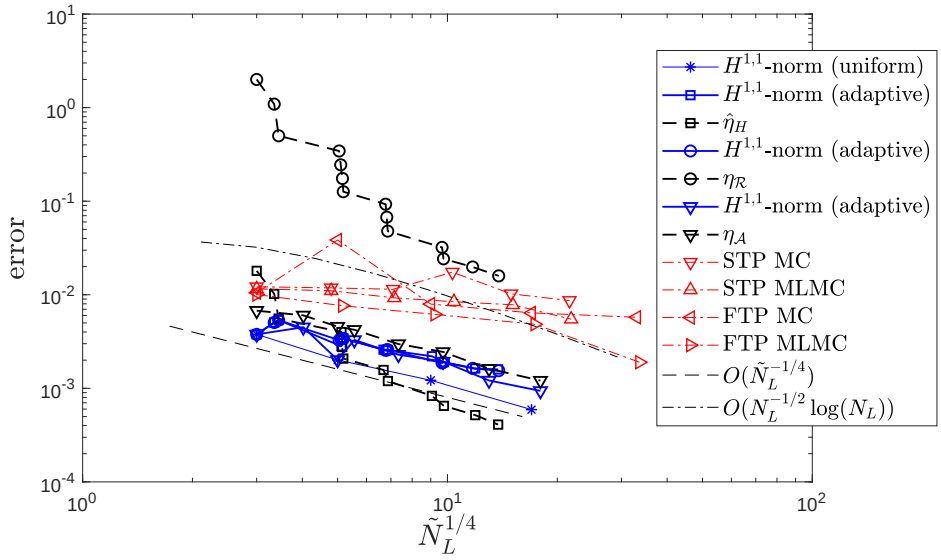

Figure 4.11: Convergence history for tensorized exponential covariance with $\rho_1 = 0.1$ and $\rho_2 = 0.1$.

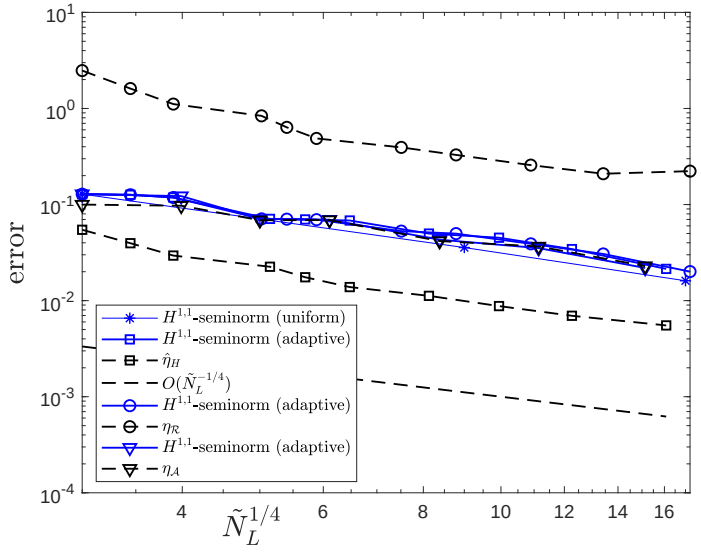

Figure 4.12: Convergence history for a non-tensorizable exponential type covariance with $\rho_1 = 10$ and $\rho_2 = 0.1$.

Test 4 ($\rho_1 = 10, \rho_2 = 0.1$) Here we have chosen $\mathcal{C}_f$ as

$$\mathcal{C}_f(x,y) = \exp\left(-\sqrt{\tfrac{(x_1-y_1)^2}{10} + \tfrac{(x_2-y_2)^2}{0.1}}\right)$$

to illustrate that we are not bound by a Karhunen-Loève expansion to find the covariance function $\mathcal{C}_u$ to such a hand-picked covariance function. The convergence history for this experiment can be seen in Figure 4.12.

The numerical experiments have been implemented in MATLAB and have been run on a laptop with an Intel i7-4720HQ with 16GB of RAM.

Remark 4.1. *Another possibility of approximating $\mathcal{C}_u$ can be employed if an expansion like (4.4) is known. Since then there holds*

$$\mathcal{C}_u(x,y) = \sum_{i,j=1}^{\infty} \Psi_{i,j}(x,y)$$

where for all pairs of indices $i,j = 1,\dots,\infty$ the $\Psi_{i,j}$'s satisfy the PDE

$$\begin{aligned} (\Delta_x \otimes \Delta_y)\Psi_{i,j} &= \Phi_{i,j}, && \text{in } \mathcal{D}, \\ \Psi_{i,j} &= 0, && \text{on } \partial\mathcal{D}, \end{aligned}$$

with

$$\begin{aligned} \Phi_{i,j}(x,y) := {} & (\lambda_{1,i}\lambda_{2,j})^{1/2}\varphi_{1,i}(x_1)\varphi_{1,i}(y_1)\varphi_{2,j}(x_2)\varphi_{2,j}(y_2) \\ & + (\lambda_{1,i}\lambda_{2,j}^*)^{1/2}\varphi_{1,i}(x_1)\varphi_{1,i}(y_1)\varphi_{2,j}^*(x_2)\varphi_{2,j}^*(y_2) \\ & + (\lambda_{1,i}^*\lambda_{2,j})^{1/2}\varphi_{1,i}^*(x_1)\varphi_{1,i}^*(y_1)\varphi_{2,j}(x_2)\varphi_{2,j}(y_2) \\ & + (\lambda_{1,i}^*\lambda_{2,j}^*)^{1/2}\varphi_{1,i}^*(x_1)\varphi_{1,i}^*(y_1)\varphi_{2,j}^*(x_2)\varphi_{2,j}^*(y_2). \end{aligned}$$

Thus, instead of solving for $\mathcal{C}_u$ "in total" one might resort to approximating the most contributing $\Psi_{i,j}$'s and build a solution by adding up the partial solutions. In particular, this can be an attractive idea in terms of parallelization.

Conclusion

In this thesis we have developed adaptive FEM for the approximation of covariance functions and second moments of elliptic partial differential equations. We could show that the developed residual and hierarchical error estimators, $\eta_{\mathcal{R}}$ and η_H, are *reliable* but only *weakly efficient*. As a remedy for this deficiency we have additionally developed an error estimator based on averaging and shown it to be *asymptotically exact*, i.e. *reliable* and *efficient*. It is seen that in typical situations, such as solutions with steep gradients, that the adaptive Finite Element Method (AFEM) for the approximation of covariance functions is very competetive. Since the polynomial degree is kept fixed at $p = 1$, the quality of approximation is of a low order, which is a drawback of the presented approach, if the data are of high regularity. By adopting existing *hp*-error estimators (cf. [41]) to the situation of deterministic moment equations the presented methods may be extended to (potentially) yield higher if not exponential convergence rates (cf. e.g. [44, 45] for a related problem) and alleviate the difficulties when approximating singular derivatives of solutions.

Moreover, the 1-irregularity might be viewed as an inflexible restriction on the adaptive procedure. The 1-irregularity condition on the mesh $\mathcal{T}$ is not well suited to the four dimensional situation, as there can be a large magnitude of "unnecessary" refinements. These could be avoided if one considers the extension of the given theory to the situation of k-irregular meshes (cf. [24, 57, 39]), which can give a better handle on the refinement procedure in four space dimensions as well as decrease the need for too many implied refinements because of hanging nodes. This idea also fits well together with *hp*-refinement ideas.

We have also presented in this thesis the error analysis of Monte Carlo and multilevel Monte Carlo approximations for covariance functions by way of the second moment problem. We have shown that we can improve upon a full tensor product approximation by virtue of sparse tensor approximation techniques and have shown that in all regimes, which are given by the quality of the solver used, the analyzed sparse tensor product multilevel Monte Carlo method is the best method in terms of cost versus accuracy and in terms of asymptotic memory requirements. These methods might further be improved by considering different approximating sequences $\mathcal{X}$ and $\mathcal{Y}$ that are used in combination with the sparse tensor product operator $\hat{P}_L(\mathcal{X}, \mathcal{Y})$. For example, if the sequences $\mathcal{X}$ and $\mathcal{Y}$ are based on an adaptive solution procedure, the dimensions of the corresponding finite dimensional spaces may be optimized to yield an overall better asymptotic cost requirement as well as a potentially better convergence rate in the presence of singularities may be experienced.

Finally, we have presented numerical experiments comparing the Monte Carlo and multilevel Monte Carlo methods with the AFEM as well as its uniform variant. It can be seen that in the case of a highly correlated random process and in the presence of a Karhunen-Loève expansion that the Monte Carlo and multilevel Monte Carlo methods are the preferrable methods, when one wants to approximate the covariance function. If on the other hand the underlying process has a short correlation length, then it is also shown in the experiments that the adaptive and uniform Finite Element Method are the preferrable methods.

Acknowledgement

I would like to extend my utmost gratitude towards Prof. Dr. Alexey Chernov for without his help this thesis would not have been possible in its entirety. Moreover, I would like to express my appreciation for the freedom I was awarded in conducting the present research and the excellent guidance I was imparted when it was necessary. Furthermore, I would like to cordially thank Prof. Dr. Helmut Harbrecht for accepting the task of a co-examiner of this thesis.

I would also like to thank my office colleagues Claudio, Lorenzo, Marlon, Nick and Tùng for an always helpful and friendly atmosphere, which has been an essential ingredient in the process of the development of this thesis. Last but not least, I want to thank my family for the unwaivering support throughout the years.

Eidesstattliche Erklärung

Hiermit erkläre ich, Erik Marc Schetzke, an Eides statt, dass ich die Dissertation, **Numerical Methods for Covariance Functions of Elliptic Problems under Uncertainty**, selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Oldenburg,

Unterschrift

Bibliography

[1] Robert A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[2] Mark Ainsworth and J. Tinsley Oden. A posteriori error estimation in finite element analysis. *Computer Methods in Applied Mechanics and Engineering*, 142, 1997.

[3] J. M. Aldaz. Strengthened Cauchy-Schwarz and Hölder inequalities. *JIPAM. J. Inequal. Pure Appl. Math.*, 10, 2009.

[4] Kendall Atkinson and Weimin Han. *Theoretical numerical analysis*, volume 39 of *Texts in Applied Mathematics*. Springer, Dordrecht, third edition, 2009. A functional analysis framework.

[5] A. Barth, Ch. Schwab, and N. Zollinger. Multi-level monte carlo finite element method for elliptic pdes with stochastic coefficients. *Numerische Mathematik*, 119(1):123–161, Sep 2011.

[6] Marcel Bieri, Roman Andreev, and Christoph Schwab. Sparse tensor discretization of elliptic spdes. *SIAM J. on Sci. Comput.*, 31(6):4281–4304, 2009/10.

[7] C. Bierig and A Chernov. Convergence analysis of multilevel monte carlo variance estimators and application for random obstacle problems. *Numerische Mathematik*, 130(4):579–613, August 2015.

[8] Claudio Bierig and Alexey Chernov. Approximation of probability density functions by the multilevel Monte Carlo maximum entropy method. *J. Comput. Phys.*, 314:661–681, 2016.

[9] Claudio Bierig and Alexey Chernov. Estimation of arbitrary order central statistical moments by the multilevel Monte Carlo method. *Stoch. Partial Differ. Equ. Anal. Comput.*, 4(1):3–40, 2016.

[10] Folkmar A. Bornemann, Bodo Erdmann, and Ralf Kornhuber. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM Journal on Numerical Analysis*, 33(3):1188–1204, 1996.

[11] Malte Braack and Nico Taschenberger. Hierarchical a posteriori residual based error estimators for bilinear finite elements. 10(2):466–480, 2013.

[12] Dietrich Braess. *Finite elemente*. Springer, fifth edition, 2013. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.

[13] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[14] C. Carstensen. All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable. 73, 2004.

[15] C. Carstensen, D. Gallistl, and J. Gedicke. Justification of the saturation assumption. 134(1):1–25, 2016.

[16] Carsten Carstensen and Sören Bartels. Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. I. Low order conforming, nonconforming, and mixed FEM. *Mathematics of Computation*, 71(239):945–969, 2002.

[17] Carsten Carstensen and Max Jensen. Averaging techniques for reliable and efficient a posteriori finite element error control: Analysis and applications. *Recent advances in adaptive computation. Providence, RI: American Mathematical Society, pp. 15-34. Contemporary mathematics. (383).*, 2005.

[18] Alexey Chernov. Sparse polynomial approximation in positive order Sobolev spaces with bounded mixed derivatives and applications to elliptic problems with random loading. 62, 2012.

[19] Phillipe G. Ciarlet. *The Finite Element Method for Elliptic Problems.* The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, Mathematics and its Applications, 1978.

[20] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15, 2011.

[21] Dauge M. Schwab C. Costabel, M. Exponential convergence of *hp*-FEM for Maxwell equations with weighted regularization in polygonal domains. *Math. Models Methods Appl. Sci.*, 15:575–622, 2005.

[22] Leszek Demkowicz. *Computing with hp-adaptive finite elements. Vol. 1.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2007. One and two dimensional elliptic and Maxwell problems, With 1 CD-ROM (UNIX).

[23] Leszek Demkowicz, Jason Kurtz, David Pardo, Maciej Paszyński, Waldemar Rachowicz, and Adam Zdunek. *Computing with hp-adaptive finite elements. Vol. 2.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2008. Frontiers: three dimensional elliptic and Maxwell problems with applications.

[24] Paolo Di Stolfo, Andreas Schröder, Nils Zander, and Stefan Kollmannsberger. An easy treatment of hanging nodes in *hp*-finite elements. *Finite Elements in Analysis and Design*, 121, 2016.

[25] Martin Eigel, Claude Jeffrey Gittelson, Christoph Schwab, and Elmar Zander. Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mech. Engrg.*, 270, 2014.

[26] Victor Eijkhout and Panayot Vassilevski. The role of the strengthened Cauchy-Buniakowski ĭ-Schwarz inequality in multilevel methods. 33(3):405–419, 1991.

[27] Alexandre Ern and Jean-Luc Guermond. *Theory and Practice of Finite Elements*. Applied Mathematical Sciences. Springer-Verlag, New York, 2004.

[28] Alexandre Ern and Jean-Luc Guermond. Finite element quasi-interpolation and best approximation. 51(4):1367–1385, 2017.

[29] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[30] Philipp Frauenfelder, Christoph Schwab, and Radu Alexandru Todor. Finite elements for elliptic problems with stochastic coefficients. *Comput. Methods Appl. Mech. Engrg.*, 194, 2005.

[31] Thomas-Peter Fries, Andreas Byfut, Alaskar Alizada, Kwok Wah Cheng, and Andreas Schröder. Hanging nodes and XFEM. *International Journal for Numerical Methods in Engineering*, 86, 2011.

[32] R. Ghanem and P. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag New York, 1991.

[33] Michael Griebel and Helmut Harbrecht. On the construction of sparse tensor product spaces. *Mathematics of Computation*, 82(282):975–994, 2013.

[34] Michael Griebel and Helmut Harbrecht. On the convergence of the combination technique. In *Sparse grids and applications—Munich 2012*, volume 97 of *Lect. Notes Comput. Sci. Eng.*, pages 55–74. Springer, Cham, 2014.

[35] Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2012.

[36] Helmut Harbrecht, Michael Peters, and Markus Siebenmorgen. Combination technique based k-th moment analysis of elliptic problems with random diffusion. 252, 2013.

[37] Paul Houston, Christoph Schwab, and Endre Süli. Discontinuous hp-finite element methods for advection-diffusion-reaction problems. *SIAM Journal on Numerical Analysis*, 39(6).

[38] Achim Klenke. *Probability theory*. Universitext. Springer, London, second edition, 2014. A comprehensive course.

[39] Pavel Kus, Pavel Solin, and David Andrs. Arbitrary-level hanging nodes for adaptive *hp*-FEM approximations in 3D. *Journal of Computational and Applied Mathematics*, 270, 2014.

[40] W. A. Light and E. W. Cheney. *Approximation theory in tensor product spaces*, volume 1169 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1985.

[41] J. M. Melenk and B. I. Wohlmuth. On residual-based a posteriori error estimation in *hp*-FEM. volume 15. 2001. A posteriori error estimation and adaptive computational methods.

[42] P. Pébay. Formulas for robust, one-pass parallel computation of covariances and arbitrary-order statistical moments. *Sandia Report SAND2008-6212, Sandia National Laboratories*, 2008.

[43] B. Pentenrieder. *Exponential convergence of hp-FEM for second moments of elliptic PDEs with stochastic data*. PhD thesis, ETH Zürich, 2009.

[44] Bastian Pentenrieder and Christoph Schwab. *hp*-FEM for second moments of elliptic PDEs with stochastic data. I. Analytic regularity. *Numer. Methods Partial Differential Equations*, 28(5):1497–1526, 2012.

[45] Bastian Pentenrieder and Christoph Schwab. *hp*-FEM for second moments of elliptic PDEs with stochastic data. II: Exponential convergence for stationary singular covariance functions. *Numer. Methods Partial Differential Equations*, 28(5):1527–1557, 2012.

[46] Alfio Quarteroni and Alberto Valli. *Numerical approximation of partial differential equations*, volume 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1994.

[47] Raymond A. Ryan. *Introduction to tensor products of Banach spaces*. Springer Monographs in Mathematics. Springer-Verlag London, Ltd., London, 2002.

[48] Ch. Schwab. *p- and hp-finite element methods*. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 1998. Theory and applications in solid and fluid mechanics.

[49] Ch. Schwab and R. A. Todor. Sparse finite elements for stochastic elliptic problems—higher order moments. *Computing. Archives for Scientific Computing*, 71, 2003.

[50] Christoph Schwab and Radu-Alexandru Todor. Sparse finite elements for elliptic problems with stochastic loading. *Numerische Mathematik*, 95, 2003.

[51] Olaf Steinbach. *Numerical approximation methods for elliptic boundary value problems*. Springer, New York, 2008. Finite and boundary elements, Translated from the 2003 German original.

[52] T. J. Sullivan. *Introduction to uncertainty quantification*, volume 63 of *Texts in Applied Mathematics*. Springer, Cham, 2015.

[53] Radu Alexandru Todor. *Sparse perturbation algorithms for elliptic PDE's with stochastic data*. PhD thesis, ETH Zürich, 2005.

[54] R. Verfürth. *A Posteriori Error Estimation Techniques for Finite Element Methods*. Numerical Mathematics and Scientific Computation. Oxford Science Publications, 2013.

[55] Tobias von Petersdorff and Christoph Schwab. Sparse finite element methods for operator equations with stochastic data. *Applications of Mathematics*, 51, 2006.

[56] Pavel Šolín, Karel Segeth, and Ivo Doležel. *Higher-order finite element methods*. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL, 2004. With 1 CD-ROM (Windows, Macintosh, UNIX and LINUX).

[57] Pavel Šolín, Jakub Červený, and Ivo Doležel. Arbitrary-level hanging nodes and automatic adaptivity in the *hp*-FEM. *Mathematics and Computers in Simulation*, 77, 2008.

[58] Dirk Werner. *Funktionalanalysis*. Springer-Verlag, Berlin, 2007.

[59] A.M. Yaglom. *An Introduction to the Theory of Stationary Random Functions*. Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[60] K. Yosida. *Functional Analysis*. Springer Verlag, 1995.

[61] O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *International Journal for Numerical Methods in Engineering*, 24(2):337–357, 1987.