Tensor Methods for the Numerical Solution of High-Dimensional Parametric Partial Differential Equations

Vorgelegt von
M.Sc. in Mathematik
Max Pfeffer
geboren in Berlin

von der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
Dr. rer. nat.

genehmigte Dissertation

Promotionsausschuss:
Vorsitzender: Prof. Dr. rer. nat. Peter Bank
Gutachter: Prof. Dr. rer. nat. Reinhold Schneider
Gutachter: Prof. Dr. rer. nat. Hermann G. Matthies
Gutachter: Dr. rer. nat. André Uschmajew

Tag der wissenschaftlichen Aussprache: 28. März 2018

Berlin 2018

In loving memory of my father

Preface

Some of the main results of this thesis have been previously published in

Martin Eigel, Max Pfeffer, and Reinhold Schneider.
Adaptive stochastic Galerkin FEM with hierarchical tensor representations.
Numer. Math., 136(3):765–803, 2017.

The section on tensor decomposition formats is an edited version of my own contribution to the review article

Szilárd Szalay, Max Pfeffer, Valentin Murg, Gergely Barcza, Frank Verstraete, Reinhold Schneider, and Örs Legeza.
Tensor product methods and entanglement optimization for ab initio quantum chemistry.
International Journal of Quantum Chemistry, pages 1342–1391, 2015.

Additionally, some results have not yet been published. They are to appear under the working title

Martin Eigel, Manuel Marshall, Max Pfeffer, and Reinhold Schneider.
Adaptive stochastic Galerkin FEM with hierarchical tensor representations.
work in progress.

or later publications.

I would like to thank my supervisor Prof. Dr. Reinhold Schneider for his mentoring, his help, and especially for the great time that I had as his student both in- and outside of work. I would also like to thank Prof. Dr. Hermann Matthies for taking his time to assessing this thesis, and similarly Dr. André Uschmajew, who has put me under the right amount of pressure and helped a great deal in the final stages of this work. I thank my colleagues for fruitful discussions, Alexandra Schulte for keeping us all sane, and Szilárd Szalay for allowing me to use the pictures that are mostly his work. In particular, I also thank Manuel Marshall for patiently enduring my many requests and rendering the numerical part of this thesis possible.

Finally, I want to thank my friends and family for the good times away from work, for friendship and love, and for believing in me.

Berlin, April 2018 *Max Pfeffer*

Notational Conventions

$u(\cdot), (\cdot,\cdot), \ldots$	Placeholder for variable that is not fixed
$\langle \cdot, \cdot \rangle$	Dual pairing
$\mathcal{N}_\sigma$	Gaussian distribution with mean 0 and standard deviation σ
$\equiv$	constantly equal to
$\lesssim$	smaller than up to a constant
$\mathfrak{B}(\Gamma)$	Borel set over some set Γ
id	Identity operator
I_{q_m}	Identity matrix in the Euclidean space $\mathbb{R}^{q_m}$
$e^{(m)}_{\mu_m}$	μ_m-th unit vector of the space $\mathbb{R}^{q_m}$ (the dimension q_m is implied)
δ_{mn}	Kronecker delta: $\delta_{mn} = 1$ if $m = n$, $\delta_{mn} = 0$ otherwise
$\operatorname{diam}(T)$	Largest width of a simplicial element T
$u_{\vert D}$	Restriction of the function u to the domain D
$\mathcal{L}(\mathcal{V}, \mathcal{V}^*)$	Space of linear operators on the Hilbert space $\mathcal{V}$ to its dual
$\mathcal{C}^0(D)$	Space of continuous functions on the domain D
$\mathcal{P}_p(D)$	Space of polynomials of degree at most p on the domain D
$\operatorname{cond}(\mathbf{A})$	Condition number of the discrete linear operator $\mathbf{A}$

Contents

1 Introduction

Partial differential equations (PDEs) are used to describe and model many physical phenomena that are encountered in nature. Famous examples include the *Maxwell equations* that model the propagation of electro-magnetic waves, the *Navier-Stokes equation* that describes turbulent flows, and the *Schrödinger equation*, which is used to predict the wave functions of quantum objects. Elliptic PDEs are an important subset of these fundamental equations and amongst them, the leading example is the diffusion equation

$$-\nabla \cdot (a \nabla u) = f.$$

This is a rather simple but nonetheless powerful differential equation. It describes such different phenomena as the distribution of heat in a room or any other material, the diffusion of chemical liquids, the tension in a membrane, etc. For this reason, and for its simplicity, the diffusion equation has been studied extensively for many decades. Its solution theory is abundant and an uncountable number of algorithms for its solution have been studied, implemented, and used.

However, most processes in real-life can only be modelled approximately and thus this simple form of the equation is often not more than a model for mechanisms that are in fact more complicated. For instance, in many cases, not all properties of a material may be known. And almost always, measurements are disturbed by random influences that are hard to predict. For this reason, the introduction of randomness to PDEs has become popular. This results in stochastic partial differential equations, i.e. PDEs that depend in some way on a random variable or a random field. For example, the coefficient a in the diffusion equation may be treated as a random field.

In this thesis, we simplify this notion and investigate problems with *parametric* dependence, i.e. equations where the coefficient a depends on parameters y with known and simple distributions, instead of random variables whose distribution is given intrinsically with the coefficient. Such a model problem is a simplification that allows for the development of an in-depth theory before it may be generalised to more prevalent and applicable tasks. However, it still covers several general cases that pose more or less diverse theoretical problems.

The simplest example for a parametric diffusion coefficient is the so called *affine* case, in which a admits a series expansion

$$a(y) = a_0 + \sum_{m=1}^{\infty} a_m y_m$$

with uniformly distributed $y = (y_m)_{m \in \mathbb{N}}$ around some mean function a_0. This representation may be the result of a *Karhunen-Loève-transform* [38, 53] or other expansions that are akin to a *separation of variables*. Essentially, one "trades randomness for high-dimensionality" [75]: The coefficient now depends on an infinite number of parameters $y_1, y_2, \ldots$ and even if we truncate at some $M \in \mathbb{N}$, the problem remains high-dimensional. An application of the Galerkin method immediately leads to a Galerkin operator with tensor product structure [14] and the Galerkin problem suffers from the *curse of dimensionality*, i.e. the Galerkin spaces grow exponentially with the truncation parameter M. The numerical solution of such systems quickly becomes prohibitive due to their size. Therefore, the aim of this thesis is to use the machinery of low rank tensor calculus for the numerical solution of parametric PDEs.

In addition to the affine diffusion coefficient, we investigate the so called *log-normal* coefficient, where $\log(a)$ has the above series expansion but with normally distributed parameters y_m. While this problem is more realistic, it is also much harder to tackle because the coefficient cannot be uniformly bounded away from zero and infinity. Instead, the bounds are again random variables. Several approaches to overcoming this obstacle have been developed [75, 23, 60]. Essentially, an additional weight has to be introduced in order to obtain the necessary bounds. This can be done by either shrinking the test space only or by adding the weight into the

measure, thus producing an intermediate space both for the solution and the test functions. We follow the latter approach and we will give an overview of the theory without adding any theoretical results.

As many authors before us, for the numerical calculation of the solution, we employ a Galerkin method, which relies on the orthogonal projection onto a discrete and finite dimensional subspace of the problem space. This is opposed to Monte Carlo (MC) methods, where certain moments or other properties of the solution are sampled and the resulting problem is deterministic. Galerkin discretisations of parametric PDEs have been done by many authors before us [4, 15, 16, 75, 40, 41]. In the deterministic domain, we choose a finite element (FEM) approach using piecewise polynomial functions [28, 15, 16, 17]. Instead, one could employ a wavelet basis, as for example done in [75]. The parameter space is discretised by a generalised polynomial chaos (GPC) basis. This was introduced by Wiener [91] and generalised by Cameron and Martin [10]. Extensive research has since been done [24, 25, 80, 59] and we combine several results for an efficient numerical treatment of the equations. For the affine case, we apply a basis of Legendre polynomials, while in the log-normal problem the original Hermite basis has proved more fitting. We exploit their respective properties in order to obtain explicit Galerkin operators and error estimators.

Recent debate has been around the question of the applicability of tensor methods to the problem. The affine coefficient immediately yields a Galerkin operator that has a natural low rank tensor structure. This motivates the application of recently popularised tensor methods [30, 5]. The alternate approach of best N-term approximation has been more widely applied in the past [15, 16, 3]. This however relies heavily on sparsity features both of the operator and the approximate discrete solution. If we allow different deterministic basis sets for each GPC basis function in the parameters, and under certain assumptions, one can establish that a direct choice of basis functions is superior to a tensor approach. This is a result of the fact that the Legendre basis is already optimal for uniformly distributed parameters.

However, we believe that the low rank tensor approach has a claim for recognition as well due to its elegance and simplicity. Additionally, for the log-normal problem, the optimal basis set is not a priori known and reduced basis functions can be advantageous. It is so far not known how the ranks of the solution behave and even their relation to the regularity of the solution is not fully understood [11]. However, it is possible to estimate the error that originates in the low rank approach and we employ an equilibration of errors. Since the FEM discretisation and the GPC discretisation generate a certain error, an excessive rank increase is not necessary. Our results show that the ranks grow moderately when balanced with the other errors.

As a new contribution, we develop error estimators for the deterministic as well as the parametric and the discrete errors. This is a continuation of the publications by Eigel et al. [15, 16]. We show that these estimators can be calculated efficiently using tensor techniques. For the first time, we also state error estimators for the log-normal parametric diffusion equation and the tensor format also allows for their efficient computation. We develop an adaptive algorithm that refines the respective discretisation that exhibits the largest error. A Monte Carlo error sampling is performed in order to approximate the mean error of our solution, therefore checking the reliability of our estimators.

Our main contribution to the state of the art is the application of hierarchical tensor representations to the problem of affine and log-normal parametric PDEs and the development of error estimators in these tensor formats. The tensor approach presented in our publications and in this thesis can be seen as a reduced basis ansatz [7]. We discretise with respect to a certain basis that is chosen a priori and justified by previous research. The low rank tensor format is used to find more optimal subspaces in the space spanned by these finite basis functions. Since the affine problem has a natural tensor structure, some research in this direction has been done [41, 39, 12]. Both the tensor train (TT) format used in this thesis and the hierarchical tensor (HT) format have been applied [6]. However, in these publications, this has always been done by a sampling based approach. Khoromski et al. apply a cross approximation that samples the solution for some parameters $y \in \Gamma$ and then completes the tensor using existing techniques [40]. This is called *stochastic collocation* [12]. Alternatively, if one wants to know only certain aspects of the solution, one could apply a similar approach brought forward by

Grasedyck et al. [6]. Here, the desired quantity of interest is sampled and then calculated using tensor completion for HT. Our work is the first to solve a full Galerkin linear system in TT format [17].

Regarding the treatment of the log-normal case using tensor techniques, previous research is sparse. This could be attributed to the fact that the coefficient does not directly reveal a tensor structure of the Galerkin operator. Using tensor based methods on this problem is nevertheless tempting. We show that the coefficient has a composition that is evocative of the Tucker tensor format. However, other difficulties, including the ones mentioned above, might have led to a certain cautiousness when tackling this problem numerically. A sampling based approach for the log-normal parametric diffusion equation in TT format has been developed by Oseledets [63]. A decomposition of the coefficient in the canonical tensor format and the HT format has been advocated by Matthies et al. and they even establish certain error bounds for the operator [20]. However, their approach assumes the knowledge of the eigenfunctions of the covariance operator or an approximation of them, something which is not always known. Additionally, the error estimation is done in the Frobenius norm. While this is certainly insightful, it does not yield any predictions on the ellipticity of the resulting operator. This is a problem that has so far not been solved. Our approach uses a tensor approximation of the log-normal coefficient function that does not require the a priori knowledge of the basis of the deterministic space. Furthermore, it is possible to retain the continuity of the approximated coefficient. This is a novel approach that combines methods for the calculation of continuous reduced basis functions with approximations using quadrature of integrals. As a result, the operator can be approximated in the offline phase of the algorithm and does not need to be calculated again. However, the required L^∞-error estimation of the coefficient is still a matter of future research. We therefore only state our method and assume that the result is sufficiently close to the original problem to retain the hard earned ellipticity of the operator.

A big chunk of this work is the discussion and analysis of tensor decomposition formats [43]. This serves as a good overview of current research and we hope to have added value by tying this all together and unifying some theory that has so far been considered independently. First, we briefly discuss the canonical tensor format that was introduced by Hitchcock almost a century ago [35]. This is the most straightforward generalisation of the singular value decomposition in the matrix case in the sense that it consists of a sum of R *elementary tensor*, i.e. tensors of rank one. The canonical representation of a tensor $V \in \mathbb{R}^{q^M}$ of order M reads

$$V = \sum_{k=1}^{R} v_k^{(1)} \otimes \ldots \otimes v_k^{(M)}.$$

However, despite its simplicity, this has a number of downsides, which is why different formats have been developed, most of them very recently. We state a very general way of representing tensors and combine this vision with insights from quantum physics, where these formats have been known under different names for some time [74, 87, 67]. Tensor formats that exhibit a tree structure all more or less share the same advantages. In particular, we explore the older Tucker format and show its similarities with a reduced basis ansatz [79]. Most of the theory that applies to all tensor tree formats is established using the example of the Tensor Train (TT) format that has been introduced by Oseledets and Tyrtyshnikov [64]. Here, the tensor V is given as a product of tensors of order 3, here stated element-wise:

$$V(\mu_1, \ldots, \mu_M) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_{M-1}=1}^{r_{M-1}} V_1(\mu_1, k_1) V_2(k_1, \mu_2, k_2) \cdots V_M(k_{M-1}, \mu_M).$$

Even though the discovery was independent, TT tensors can be understood as a special case of the more general hierarchical tensors (HT) that were first stated by Hackbusch and Kühn [29].

An important property and the basis of most algorithms using tensor decompositions is the fact that tree tensors of a fixed rank form a manifold that can be embedded in the tensor space [37, 84]. This allows for Riemannian optimisation. A fundamental work for Riemannian

optimisation in the matrix case is the book by Absil [1]. Several generalisations to tensors have been established [70, 44, 76]. This field of research is especially flourishing and many results are currently being investigated [42, 57, 55, 58]. Convergence theory of these algorithms is more or less established, although some open questions remain [73].

Perhaps the most widely used algorithm is the Alternating Least Squares (ALS) [36] algorithm - which is why it is dubbed the "workhorse" algorithm in the review by Kolda and Bader [43]. This is akin to the Density Matrix Renormalisation Group (DMRG), which is a similar algorithm mostly employed in quantum chemistry for the solution of spin systems or the stationary electronic Schrödinger equation [90, 78]. However, although it is seen as the benchmark algorithm for most tensor methods, convergence theory is sparse. To a certain degree, the ALS can be seen as a Riemannian method and some results have been published over time [82, 52]. Unfortunately, examples can be constructed that observe sub-linear convergence rates of the ALS [19]. The DMRG algorithm overcomes some of the limitations. Most notably, it is rank adaptive. Nevertheless, even here it is possible to construct examples that lead to undesired results. We give one of these examples explicitly.

Since the ALS algorithm is used for solving the linear system resulting from the Galerkin method in this thesis, we give a thorough overview over its approach. Additionally, we propose an efficient way of applying a preconditioner only for the large deterministic component. The preconditioning is done using the mean of the respective operator and we show that this property is easily accessible in the TT format.

Finally, we include numerical experiments to underline the validity of our approach. We show calculations for the approximation of the log-normal diffusion coefficient, the application of our log-normal preconditioner, and convergence results of the adaptive algorithm. In particular, we will observe the applicability of the newly developed error estimators.

Structure of the Thesis

This thesis is organised in an introductory part, which includes this chapter and Chapter 2; a part on parametric PDEs and their numerical treatment, namely Chapters 3 and 4; a Chapter on tensor decompositions; and finally a part that combines these notions and contains the main results in Chapter 6, Chapter 7, and Chapter 8.

More specifically, Chapter 2 will deal with all the necessary theory that is required to introduce parametric PDEs and tensor decompositions. We introduce tensor products of Hilbert spaces and general elliptic PDEs as well as the Galerkin method.

Chapter 3 first introduces parametric PDEs in its most general form. We then state the leading example of the parametric diffusion equation and give a short introduction into stochastic problems. After that, we deal with the solution theory both of the affine and the log-normal problem. This is meant to be a concise overview over the theory and especially for the log-normal case, we cover the main issues but go into detail only when absolutely necessary. Furthermore, as a small example for a non-decaying coefficient, the localised uncertainty problem is stated.

In Chapter 4, we investigate the different discretisations of the deterministic and the parametric space. A short introduction is given and some alternatives are discussed briefly. The bulk of this chapter deals with the treatment of the polynomial chaos bases for the different problems. This leads to a Galerkin system that is stated here.

Chapter 5 is a very broad overview over the state of the art of numerical tensor methods. First, we give a very general formulation for tensor representations before we define certain criteria that make them more manageable and ultimately lead to tensor trees. We discuss different tensor formats as well as their advantages and disadvantages. An introduction of Riemannian optimisation in tensor formats is given including some brief discussion on convergence theory. We then review other minimisation techniques, most notably the ALS and the DMRG algorithm.

Chapter 6 is an interim discussion of some numerical aspects of the tensor treatment of parametric PDEs. This contains mainly our own work that is necessary for efficient computation. It explores a decomposition of the log-normal Galerkin operator as well as an efficient implementation of a preconditioner for both problems.

In Chapter 7, we present the error estimators for the parametric diffusion equation in tensor formats. We state error estimators for the affine problem and prove their validity. Additionally, we derive similar estimators also for the log-normal diffusion equation. We show that these estimators can be calculated efficiently in the tensor format and introduce an adaptive algorithm for the solution of the parametric diffusion equation. We also present a way to measure the mean error that is made in our computations in order to show the validity of our estimators and to compare with other established methods.

Chapter 8 shows numerical experiments for a given coefficient both in the affine and the log-normal case. This chapter serves to underline the feasibility and applicability of our method.

2 Basic Theory

Before it is possible to explore the theory of parametric partial differential equations and their tensor treatment, it is necessary to clear up some basic definitions, conventions and mechanisms of both tensor analysis and differential equations. In this first chapter of the thesis, we first recapitulate basic theory about Hilbert spaces and tensor products thereof. The interested reader is referred to the book by Hackbusch [30], which studies this topic exhaustively. Additionally, we rely on the much more accessible work of Reed and Simon [68] and the dissertation of Uschmajew [83] that have introduced tensors in a way that we consider most fitting.

Secondly, we introduce some very basic prerequisites for partial differential equations, most notably the Galerkin method for discretisation. For this, we resort to the book by Grossmann and Roos [28]. The purpose of this chapter is not be exhaustive, but to lay out the foundations for the coming chapters.

2.1 Hilbert Space Theory

In this thesis, we exclusively deal with *separable* Hilbert spaces $(\mathcal{V}, (\cdot,\cdot)_\mathcal{V})$, i.e. there exists an at most countably infinite subset that is dense in $\mathcal{V}$. This implies that $\mathcal{V}$ admits a countable orthonormal basis $\mathcal{B} \subset \mathcal{V}$, meaning that for all $\phi, \psi \in \mathcal{B}$ with $\phi \neq \psi$ it holds

$$(\phi, \psi)_\mathcal{V} = 0, \|\phi\|_\mathcal{V} = 1$$

and

$$\overline{\operatorname{span}(\mathcal{B})} = \mathcal{V}.$$

Without the use of the axiom of choice, it can be shown that these two properties are equivalent.

Theorem 2.1. *A Hilbert space is separable if and only if it admits a countable orthonormal basis.*

The realm of this thesis will be confined to vector spaces over the real numbers, i.e. any Hilbert space $\mathcal{V}$ is an $\mathbb{R}$-vector space. Many results can be generalised to complex vector spaces or more general vector spaces over some field $\mathbb{K}$ if one adjusts the usual definitions. However, this is not always the case. We denote the dual space as $(\mathcal{V}^*, (\cdot,\cdot)_{\mathcal{V}^*})$ and we know that any Hilbert space is isomorphic to its dual. The dual pairing will be denoted by $\langle\cdot,\cdot\rangle : \mathcal{V}^* \times \mathcal{V} \to \mathbb{R}$.

An important result is the *Riesz representation theorem*:

Theorem 2.2. *For every $f \in \mathcal{V}^*$ there exists exactly one $u \in \mathcal{V}$ such that*

$$\langle f, v\rangle = (u, v)_\mathcal{V} \quad \forall v \in \mathcal{V}.$$

Furthermore, it holds

$$\|u\|_\mathcal{V} = \|f\|_{\mathcal{V}^*}.$$

Examples of Hilbert Spaces

Almost all Hilbert spaces used in this thesis are one of the following:

- the real coordinate spaces $\mathbb{R}^d$ of finite dimension $d < \infty$,
- the Lebesgue space

$$L^2(D) = \left\{ v : D \to \mathbb{R} : \int_D |v(x)|^2 \,\mathrm{d}x < \infty \right\}$$

  of square-integrable functions over some domain D and with the Lebesgue measure,

- the Lebesgue-Bochner space,

$$L^2_\pi(\Gamma, \mathcal{X}) := \left\{ v : \Gamma \to \mathcal{X} : \int_\Gamma \|v(y)\|^2_{\mathcal{X}} \, \mathrm{d}\pi(y) < \infty \right\}$$

  of square-integrable functions over some space Γ with values in another Hilbert space $(\mathcal{X}, (\cdot,\cdot)_{\mathcal{X}})$ and the measure π,

- the Sobolev spaces $H^k(D) = W^{k,2}(D)$ of k-times weakly differentiable, square-integrable functions over a domain $D \subset \mathbb{R}^2, \mathbb{R}^3$ with the Lebesgue measure.

Remark 2.3. We will also use the Lebesgue spaces $L^p(D)$ and the Lebesgue-Bochner spaces $L^p_\pi(\Gamma, \mathcal{X})$ for $p \neq 2$ and they are defined analogously. However, these do not have a scalar product and are therefore not Hilbert spaces. The elements of any Lebesgue space are equivalence classes of functions. It is conventional to denote these L^p-functions in the same way as classical functions by a representative.

The Sobolev Spaces H^1 and H^1_0

For the Hilbert space $H^1(D)$, we define the H^1-norm for $v \in H^1(D)$

$$\|v\|_{H^1(D)} := \|v\|_{L^2(D)} + |v|_{H^1(D)},$$

where

$$\|v\|^2_{L^2(D)} = \int_D v^2(x) \, \mathrm{d}x$$

is the usual L^2-norm and

$$|v|^2_{H^1(D)} = \int_D \nabla v(x) \cdot \nabla v(x) \, \mathrm{d}x = \|\nabla v\|^2_{L^2(D)}$$

is the H^1-seminorm.

We often use the space of H^1-functions with Dirichlet boundary $v_{|\partial D} = 0$, i.e.

$$H^1_0(D) = \{ v \in H^1(D) : \overline{\operatorname{supp}(v)} \subset D \}.$$

On this space, the H^1-seminorm is actually a norm and we set

$$\|v\|_{H^1_0(D)} := |v|_{H^1(D)} \quad \forall v \in H^1_0(D).$$

The dual space $H^{-1}(D)$ of $H^1_0(D)$ has the regular dual norm

$$\|f\|_{H^{-1}(D)} = \sup_{v \in H^1_0(D)} \frac{|\langle f, v \rangle|}{\|v\|_{H^1_0(D)}}.$$

Tensor Products of Hilbert Spaces

The tensor product of Hilbert spaces is crucial for this thesis. We give a very straightforward introduction.

Let $(\mathcal{X}, (\cdot,\cdot)_{\mathcal{X}})$ and $(\mathcal{Y}, (\cdot,\cdot)_{\mathcal{Y}})$ be separable Hilbert spaces of dimension at least 2 and with countable basis sets $\mathcal{B} = \{\phi^{(1)}, \phi^{(2)}, \ldots\}$ and $\mathcal{B}' = \{\psi^{(1)}, \psi^{(2)}, \ldots\}$. For any $v_x \in \mathcal{X}$ and $v_y \in \mathcal{Y}$, we define the bilinear form $v_x \otimes v_y$ via

$$\begin{aligned} v_x \otimes v_y &: \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \\ (v_x \otimes v_y)(w_x, w_y) &:= (v_x, w_x)_{\mathcal{X}} (v_y, w_y)_{\mathcal{Y}}. \end{aligned}$$

Then $v_x \otimes v_y$ is called the *tensor product* of v_x and v_y and the space

$$\mathcal{X} \otimes_a \mathcal{Y} := \operatorname*{span}_{i,j \in \mathbb{N}} (\phi^{(i)} \otimes \psi^{(j)})$$

is called the *algebraic tensor product* of $\mathcal{X}$ and $\mathcal{Y}$. On this space, we can define the *induced scalar product*

$$(v_x \otimes v_y, w_x \otimes w_y)_{\mathcal{X}\otimes\mathcal{Y}} := (v_x, w_x)_{\mathcal{X}}(v_y, w_y)_{\mathcal{Y}},$$

which extends by linearity, and we also define the induced norm $\|\cdot\|_{\mathcal{X}\otimes\mathcal{Y}}$ [68].

In the case where $\mathcal{X}$ and $\mathcal{Y}$ are finite dimensional, this space is complete as it is also finite dimensional. In general, we define the *tensor product* of $\mathcal{X}$ and $\mathcal{Y}$ as the closure of the algebraic tensor product under the induced norm

$$\mathcal{X} \otimes \mathcal{Y} := \overline{\operatorname*{span}_{i,j\in\mathbb{N}}(\phi^{(i)} \otimes \psi^{(j)})}^{\|\cdot\|_{\mathcal{X}\otimes\mathcal{Y}}}.$$

The space $(\mathcal{X} \otimes \mathcal{Y}, (\cdot,\cdot)_{\mathcal{X}\otimes\mathcal{Y}})$ is thus again a Hilbert space. Elements of this space are called *tensors* and all elements that can be written as a single tensor product $v_x \otimes v_y$ are called *elementary tensors*. The induced norm is a *crossnorm*, i.e. for elementary tensors, it fulfils

$$\|v_x \otimes v_y\|_{\mathcal{X}\otimes\mathcal{Y}} = \|v_x\|_{\mathcal{X}}\|v_y\|_{\mathcal{Y}}.$$

However, not all norms that can be defined on $\mathcal{X} \otimes \mathcal{Y}$ need to be crossnorms and we will encounter a few later in this thesis.

By definition of the basis, we know that

$$\mathcal{X} = \overline{\operatorname*{span}_{i\in\mathbb{N}}(\phi^{(i)})}^{\|\cdot\|_{\mathcal{X}}} \quad \text{and} \quad \mathcal{Y} = \overline{\operatorname*{span}_{j\in\mathbb{N}}(\psi^{(j)})}^{\|\cdot\|_{\mathcal{Y}}}.$$

Since the tensor product is already defined as the closure of the algebraic tensor product, it also holds

$$\mathcal{X} \otimes \mathcal{Y} = \overline{\operatorname*{span}_{i\in\mathbb{N}}(\phi^{(i)}) \otimes_a \operatorname*{span}_{j\in\mathbb{N}}(\psi^{(j)})}^{\|\cdot\|_{\mathcal{X}\otimes\mathcal{Y}}} = \operatorname*{span}_{i\in\mathbb{N}}(\phi^{(i)}) \otimes \operatorname*{span}_{j\in\mathbb{N}}(\psi^{(j)})$$

and the set $\mathcal{B} \otimes \mathcal{B}' := \{\phi^{(i)} \otimes \psi^{(j)} : i, j \in \mathbb{N}\}$ is an orthonormal basis of $\mathcal{X} \otimes \mathcal{Y}$.

An element of $\mathcal{X} \otimes \mathcal{Y}^*$ can be interpreted as a linear operator from $\mathcal{Y}$ to $\mathcal{X}$, i.e.

$$\mathcal{X} \otimes \mathcal{Y}^* \subset \mathcal{L}(\mathcal{Y}, \mathcal{X})$$

and we know that $\mathcal{X} \otimes \mathcal{Y}^*$ is isometrically isomorphic to $\mathcal{X} \otimes \mathcal{Y}$. These operators are called *Hilbert-Schmidt* operators and they are in particular compact, see [30]. Then the Hilbert-Schmidt theorem for tensor products holds:

Theorem & Definition 2.1. *Any $f \in \mathcal{X} \otimes \mathcal{Y}$ allows the* Schmidt-decomposition

$$f = \sum_{k=1}^{\infty} \varsigma_k (v_x^{(k)} \otimes v_y^{(k)}), \quad \varsigma_1 \geq \varsigma_2 \geq \ldots, \quad \varsigma_k \searrow 0$$

and it holds $\|f\|^2_{\mathcal{X}\otimes\mathcal{Y}} = \sum_{k=1}^{\infty} \varsigma_k^2$. The sets $\{v_x^{(k)} : k \in \mathbb{N}\}$ and $\{v_y^{(k)} : k \in \mathbb{N}\}$ are orthonormal systems but not necessarily basis sets of $\mathcal{X}$ and $\mathcal{Y}$. The ς_k are called singular values *of f and in the case where $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional, the sum is finite and we also call the Schmidt-decomposition the* singular value decomposition*.*

Example 2.4. For the Euclidean spaces $\mathcal{X} = \mathbb{R}^{q_1}$ and $\mathcal{Y} = \mathbb{R}^{q_2}$, the tensor product space is the space of matrices

$$A \in \mathbb{R}^{q_1} \otimes \mathbb{R}^{q_2} = \mathbb{R}^{q_1 \times q_2}$$

and its singular value decomposition is denoted element-wise by

$$A(i_1, i_2) = \sum_{k=1}^{r} \varsigma_r X(i_1, k) Y(i_2, k) \tag{2.1}$$
$$A = X\Sigma Y^{\mathsf{T}}$$

with $\Sigma = \operatorname{diag}(\varsigma_k)$ and $X \in \mathbb{R}^{q_1 \times r}, Y \in \mathbb{R}^{q_2 \times r}$ left orthogonal, i.e. $X^\mathsf{T} X = Y^\mathsf{T} Y = I_r$. The number of positive singular values is called the *rank* of A, $\operatorname{rank}(A) := r$, and it holds

$$\|A\|^2_{\ell^2(\mathbb{R}^{q_1 \times q_2})} = \sum_{k=1}^{r} \varsigma_k^2.$$

This is also called the *Frobenius norm* of A.

It is possible to approximate the tensor $f \in \mathcal{X} \otimes \mathcal{Y}$ by truncating all summands with $k > \tilde{r}$ for some $\tilde{r}$. The discovery that this already yields the optimal result is mostly accredited to Eckart and Young in mathematics, while most physics articles recognise the fact that it had been proven by Schmidt long before for the more complicated case of integral operators [13, 71]. We give the result only for finite dimensional Hilbert spaces, i.e. for matrices, although it holds true for all Hilbert-Schmidt operators.

Theorem 2.5 (Eckart/Young, Schmidt)**.** *Let $A \in \mathbb{R}^{q_1 \times q_2}$ with an SVD as in* (2.1) *and* $\operatorname{rank}(A) = r$*. Then the best approximation $\tilde{A} \in \mathbb{R}^{q_1 \times q_2}$ with $\operatorname{rank}(\tilde{A}) = \tilde{r} \leq r$ in the Frobenius norm is given by*

$$\tilde{A}(i_1, i_2) = \sum_{k=1}^{\tilde{r}} \varsigma_r X(i_1, k) Y(i_2, k)$$

and it holds

$$\|A - \tilde{A}\|_{\ell^2(\mathbb{R}^{q_1 \times q_2})} = \min_{\operatorname{rank}(Q)=\tilde{r}} \|A - Q\|_{\ell^2(\mathbb{R}^{q_1 \times q_2})} = \sqrt{\sum_{k=\tilde{r}+1}^{r} \varsigma_k^2}.$$

The tensor product is associative and the construction of a *tensor product of order d* is possible,

$$\bigotimes_{i=1}^{d} \mathcal{V} = \mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_d.$$

Its elements are called *tensors of order d*, which are of course multilinear forms. Elementary tensors of order d have the form $v_1 \otimes \cdots \otimes v_d$. Tensor products of orthonormal basis sets form again an orthonormal basis set of tensors of order d. However, a generalisation of the Schmidt decomposition is non-trivial and the Eckart/Young theorem does not hold without a multiplicative constant - the discussion of this will take up a big part of Chapter 5 in this thesis.

The Tensor Product Representation of Lebesgue-Bochner Spaces

An example of a tensor product of Hilbert spaces are the Lebesgue-Bochner spaces $L^p_\pi(\Gamma, \mathcal{X})$ introduced above. The Fubini theorem for vector valued functions yields the tensor product structure of Lebesgue-Bochner spaces [8].

Theorem 2.6. *Let $(\mathcal{X}, (\cdot,\cdot)_\mathcal{X})$ be a separable Hilbert space and $(\Gamma, \mathfrak{B}(\Gamma), \pi)$ some measure space. Then for $1 \leq p \leq \infty$, the Lebesgue-Bochner space $L^p_\pi(\Gamma, \mathcal{X})$ is isometrically isomorphic to the tensor product space $\mathcal{X} \otimes L^p_\pi(\Gamma)$.*

These spaces will play a central role in the process of this thesis. According to the theorem, any element $v \in L^p_\pi(\Gamma, \mathcal{X})$ has a basis representation

$$v = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} V(i,j) \phi_i \otimes \psi_j$$

with

$$\mathcal{X} = \overline{\operatorname{span}_{i \in N}(\phi_i)} \quad \text{and} \quad L^p_\pi(\Gamma) = \overline{\operatorname{span}_{j \in \mathbb{N}}(\psi_j)}.$$

The norm

$$\begin{aligned}\|v\|^p_{L^p_\pi(\Gamma,\mathcal{X})} &= \int_\Gamma \|v(y)\|^p_{\mathcal{X}} \,\mathrm{d}\pi(y)\\ &= \sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}} V(i,j)\|\phi_i\|_{\mathcal{X}}\|\psi_j\|_{L^p_\pi(\Gamma,\mathcal{X})}\end{aligned}$$

is therefore the induced crossnorm.

2.2 Elliptic PDEs

In the most general form, any elliptic second order partial differential equations (PDE) on a separable Hilbert space $\mathcal{V}$ can be characterised by the operator equation

$$\mathcal{A}u = f, \tag{2.2}$$

where $\mathcal{A} : \mathcal{V} \to \mathcal{V}^*$ is linear and $f \in \mathcal{V}^*$. The operator $\mathcal{A}$ induces a bilinear form

$$B(u,v) := \langle \mathcal{A}u, v\rangle$$

for $u, v \in \mathcal{V}$. We assume that B is

(i) bounded:
$$\exists c_1 < \infty : |B(v,w)| \le c_1\|v\|_{\mathcal{V}}\|w\|_{\mathcal{V}} \quad \forall v,w \in \mathcal{V}, \tag{2.3}$$

(ii) coercive:
$$\exists c_2 > 0 : B(v,v) \ge c_2\|v\|^2_{\mathcal{V}} \quad \forall v \in \mathcal{V}. \tag{2.4}$$

This implies that $\mathcal{A}$ is elliptic and boundedly invertible and the Lax-Milgram theorem yields a unique *weak* solution $u \in \mathcal{V}$ [49] of the variational problem

$$B(u,v) = \langle f, v\rangle \quad \forall v \in \mathcal{V},$$

which is also the solution of (2.2) under the given assumptions.

Galerkin Discretisation

Even though existence and uniqueness of the weak solution u are ensured, in most cases, this solution can only be calculated approximately. One way to calculate the approximate solution is to discretise the infinite dimensional Hilbert space $\mathcal{V}$. This *Galerkin Discretisation* is given by a finite dimensional subspace $\mathcal{V}_{\text{disc}} \subset \mathcal{V}, \dim(\mathcal{V}_{\text{disc}}) = N$ [9]. The approximate solution $u_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ can be found by solving the now finite linear equation

$$B(u_{\text{disc}}, v_{\text{disc}}) = \langle f, v_{\text{disc}}\rangle = f(v_{\text{disc}}), \quad \forall v_{\text{disc}} \in \mathcal{V}_{\text{disc}}.$$

This is done by choosing a finite basis set $\mathcal{B} = \{\phi^{(1)}, \dots, \phi^{(N)}\}, \operatorname{span}(\mathcal{B}) = \mathcal{V}_{\text{disc}}$ and solving the discrete linear system

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{2.5}$$

where

$$\mathbf{A}(i,j) = B(\phi^{(i)}, \phi^{(j)}), \quad \mathbf{f}(i) = f(\phi^{(i)})$$

and thus

$$u_{\text{disc}} = \sum_{j=1}^{N} \mathbf{u}(j)\phi^{(j)}.$$

A straightforward but important result is the Galerkin orthogonality of the solution u_{disc} with respect to the scalar product $B(\cdot,\cdot)$:

$$B(u - u_{\text{disc}}, v_{\text{disc}}) = 0 \quad \forall v_{\text{disc}} \in \mathcal{V}_{\text{disc}}. \tag{2.6}$$

Céa's Lemma then states that, since B is symmetric, it holds for every $v_{\text{disc}} \in \mathcal{V}$

$$\|u - u_{\text{disc}}\|_{\mathcal{V}}^2 \leq c_2 B(u - u_{\text{disc}}, u - u_{\text{disc}}) \leq c_2 B(u - v_{\text{disc}}, u - v_{\text{disc}}) \leq \frac{c_2}{c_1} \|u - v_{\text{disc}}\|_{\mathcal{V}}^2$$

and we thus obtain a *quasi-best* Galerkin approximation

$$\|u - u_{\text{disc}}\|_{\mathcal{V}} \leq \sqrt{\frac{c_2}{c_1}} \inf_{v_{\text{disc}} \in \mathcal{V}_{\text{disc}}} \|u - v_{\text{disc}}\|_{\mathcal{V}}. \tag{2.7}$$

In general, if we have a sequence of Galerkin spaces $(\mathcal{V}_n)_{n\in\mathbb{N}}$, $\mathcal{V}_n \subset \mathcal{V}\ \forall n \in \mathbb{N}$, that fulfils

$$\lim_{n\to\infty} \inf_{v_n \in \mathcal{V}_n} \|v_n - v\| = 0 \quad \forall v \in \mathcal{V},$$

then the Galerkin solutions u_n will clearly converge to the exact solution u. In particular, this holds if $\mathcal{V}_n$ is defined by a *basis* of $\mathcal{V}$, i.e. an at most countable infinite sequence $(\phi_n)_{n\in\mathbb{N}} \subset \mathcal{V}$, any finite combination of which has to be linearly independent and satisfy

$$\mathcal{V}_n = \operatorname{span}(\phi_1, \ldots, \phi_n), \quad \mathcal{V} = \overline{\bigcup_{n\in\mathbb{N}} \mathcal{V}_n}.$$

3 Parametric PDEs

In this thesis, we consider *parametric* operator equations of the form

$$\mathcal{A}(y)u(y) = f(y) \quad \forall y \in \Gamma, \tag{3.1}$$

where for each $y \in \Gamma$, $\mathcal{A}(y) : \mathcal{X} \to \mathcal{X}^*$ is a bounded linear operator. Thus, the Hilbert space $\mathcal{V}$ is given as the Lebesgue-Bochner space

$$\mathcal{V} = L^2_\pi(\Gamma, \mathcal{X}), \quad \mathcal{V}^* = L^2_\pi(\Gamma, \mathcal{X}^*).$$

We typically consider

$$\mathcal{X} = H^1_0(D), \quad \mathcal{X}^* = H^{-1}(D).$$

This means, we have a *deterministic* domain $D \subset \mathbb{R}^2, \mathbb{R}^3$ and a parameter space Γ with some measure π, that could for example be derived from a probability space $(\Omega, \Sigma, \mathbb{P})$. Even if there is no stochastic dependence, we will refer to the dependence on $x \in D$ as *deterministic* for simplicity.

For each $y \in \Gamma$, there is a corresponding bilinear form

$$B(y; u(y), v(y)) := \langle \mathcal{A}(y)u(y), v(y) \rangle \tag{3.2}$$

and right side

$$f(y; v(y)) = \langle f(y), v(y) \rangle. \tag{3.3}$$

The full operator in (2.2) is thus given as

$$\mathcal{A} : \mathcal{V} \to \mathcal{V}^*, \quad v \mapsto [y \mapsto \mathcal{A}(y)v(y)] \tag{3.4}$$

and the equation corresponds to the variational formulation

$$B(u, v) := \int_\Gamma B(y; u(y), v(y)) \, \mathrm{d}\pi(y) = \int_\Gamma f(y; v(y)) \, \mathrm{d}\pi(y) =: f(v). \tag{3.5}$$

These definitions are given under reservation of integrability over the parameter space.

In order to obtain a weak solution $u \in \mathcal{V}$ using the Lax-Milgram theorem, it is necessary to ensure *uniform* boundedness and coercivity of the bilinear form, i.e. the constants in (2.3) and (2.4) need to hold *for all* $y \in \Gamma$. We will see later that this condition can be somewhat weakened but for now we require

(i) uniform boundedness:

$$\exists c_1 < \infty : |B(y; v(y), w(y))| \leq c_1 \|v(y)\|_{\mathcal{X}} \|w(y)\|_{\mathcal{X}} \quad \forall v, w \in L^2_\pi(\Gamma, \mathcal{X}), y \in \Gamma, \tag{3.6}$$

(ii) uniform coercivity:

$$\exists c_2 > 0 : B(y; v(y), v(y)) \geq c_2 \|v(y)\|^2_{\mathcal{X}} \quad \forall v \in L^2_\pi(\Gamma, \mathcal{X}), y \in \Gamma. \tag{3.7}$$

In addition to the regular norms of the Hilbert spaces $\mathcal{V}$ and $\mathcal{X}$, we define the *energy norm* $\|\cdot\|_{\mathcal{A}}$ on $\mathcal{V}$ as the induced norm of the bilinear form in (3.5)

$$(v, w)_{\mathcal{A}} = \langle \mathcal{A}v, w \rangle = B(v, w). \tag{3.8}$$

3.1 The Parametric Diffusion Equation

The leading example in thesis is the boundary value problem that is given by the parametric diffusion equation

$$\begin{aligned} -\nabla \cdot (a(x,y)\nabla u(x,y)) &= f(x,y), \quad x \in D, y \in \Gamma, \\ u(x,y) &= 0, \quad x \in \partial D, y \in \Gamma, \end{aligned} \tag{3.9}$$

where the parametric dependence on y is either the result of a transformation of a random field on a probability space $(\Omega, \Sigma, \mathbb{P})$, or it is given on the parameter space Γ directly. In this thesis, the coefficient a will always be defined by some series expansion or a function thereof, in order to separate the deterministic variable x from the parametric variable y. Additionally, we restrict to homogenous Dirichlet boundary conditions, though other conditions are often easily convertible to the present case.

3.1.1 Transformation of Stochastic PDEs

The *stochastic* diffusion equation is given similarly as above

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

but with random dependence on a probability space $(\Omega, \Sigma, \mathbb{P})$. Hence, the coefficient

$$a : D \times \Omega \to \mathbb{R}$$

is a random field, with Ω being the sample space, Σ a σ-algebra on Ω and $\mathbb{P}$ a probability measure on Σ. At each point $x \in D$, a is a random variable on this probability space and - vice-versa - for each realisation $\omega \in \Omega$, a is a deterministic coefficient function on D. For the treatment of random fields, see [14].

We denote the mean of the coefficient as

$$a_0(x) := \mathbb{E}(a(x,\cdot)) = \int_\Omega a(x,\omega)\, \mathrm{d}\mathbb{P}(\omega),$$

the variance as

$$\mathbb{V}(a(x,\cdot)) = \int_\Omega (a(x,\omega))^2 \, \mathrm{d}\mathbb{P}(\omega),$$

and the covariance as

$$\mathrm{Cov}_a(x,x') = \mathbb{E}\big((a(x,\cdot) - a_0(x))(a(x',\cdot) - a_0(x'))\big).$$

Therefore, we assume that a has finite variance and is thus a square-integrable random field. Furthermore, we assume that the covariance of a is continuous.

The Karhunen-Loève Expansion

As explained above, we aim at separating the deterministic dependence from the stochastic dependence in order to treat them individually. The most common way of achieving this is via the Karhunen-Loève expansion. There are other possibilities, a more general approach is explained in [75], and we refer the interested reader there.

Theorem 3.1 (Karhunen-Loève)**.** *Let $a : D \times \Omega \to \mathbb{R}$ be a square-integrable random field over a domain D and a probability space $(\Omega, \Sigma, \mathbb{P})$ with continuous covariance $\mathrm{Cov}_a(x,x')$, i.e. $a \in L^\infty(D) \otimes L^2_\mathbb{P}(\Omega)$. Then it admits the representation*

$$a(x,\omega) = a_0(x) + \sum_{m=1}^{\infty} a_m(x) y_m(\omega), \tag{3.10}$$

where the series converges and $(y_m)_{m\in\mathbb{N}}$ is a series of mutually uncorrelated square-integrable random variables with zero mean and unit variance.

Idea of proof. The complete proof can be found in [38, 53]. It relies heavily on the Schmidt-decomposition that was introduced above. In principle, the spacial functions $a_m : D \to \mathbb{R}$ are the eigenfunctions of the *covariance operator*

$$v \mapsto \int_D \mathrm{Cov}_a(x, \cdot)v(x)\,\mathrm{d}x$$

multiplied with the square root of their respective eigenvalues. The random variables are given by the projection of the random field a onto the these functions:

$$y_m(\omega) = \int_D a(x, \omega)a_m(x)\,\mathrm{d}x.$$

This means that they are also orthogonal with respect to the probability measure $\mathbb{P}$ and the expansion is therefore bi-orthogonal. The original proof requires that a has zero mean but this can be overcome by applying the theorem to the random field $a - a_0$. Additionally, we generalise the theorem to any domain D instead of a closed interval in $\mathbb{R}$. □

Since the eigenfunctions of the covariance operator are orthogonal in $L^2(D)$ and the eigenvalues form a descending sequence of non-negative numbers converging to zero, truncating (3.10) at some value M gives an approximation of the coefficient

$$a(x, \omega) \approx a_0(x) + \sum_{m=1}^{M} a_m(x)y_m(\omega).$$

The random variables $y_m : \Omega \to \mathbb{R}$ inherit their probability distribution from the coefficient and they can thus be seen as parameters

$$y = (y_m)_{m\in\mathbb{N}} \in \Gamma = \bigtimes_{m\in\mathbb{N}} \Gamma_m,$$

with $\Gamma_m \subseteq \mathbb{R}$ for all $m \in \mathbb{N}$. They are defined via their density, which is some (probability) measure π with

$$\mathrm{d}\pi(y) = \prod_{m=1}^{\infty} \mathrm{d}\pi_m(y_m).$$

Therefore, in this thesis, we assume the a priori knowledge of the series representation of the coefficient a as a parametric coefficient in the above sense.

We discuss several model cases that are widely discussed either for their simplicity and elegance or because they capture realistic applications well.

3.1.2 The Affine Problem

In the affine problem, the diffusion coefficient in (3.9) is given as

$$a(x, y) = a_0(x) + \sum_{m=1}^{\infty} a_m(x)y_m \tag{3.11}$$

with independently and identically, uniformly distributed $y_m \in \Gamma_m = [-1, 1]$ for all $m \in \mathbb{N}$. This means that the product measure $\pi = \bigotimes_{m\in\mathbb{N}} \pi_m$ is given via the Lebesgue measure

$$\mathrm{d}\pi_m(y_m) = \frac{1}{2}\mathrm{d}y_m.$$

In order to fulfil the uniform boundedness and coercivity conditions (3.6) and (3.7), we will assume that a is *uniformly elliptic*, i.e. there exist constants $0 < a_{\min} \leq a_{\max} < \infty$ such that

$$a_{\min} \leq a(x, y) \leq a_{\max} \quad \forall (x, y) \in D \times \Gamma.$$

This is for example the case if the mean is constant, $a_0(x) \equiv a_0$, and

$$\left| \sum_{m=1}^{\infty} a_m(x) y_m \right| \leq \lambda a_0 \quad \text{with} \quad \lambda < 1, \tag{3.12}$$

which implies $a(x,y) \geq (1-\lambda) > 0$ and our previous assumption $|y_m| \leq 1$. Here, we assume the more general condition

$$\operatorname*{ess\,inf}_{x \in D} a_0(x) > 0, \quad \operatorname*{ess\,sup}_{x \in D} \sum_{m=1}^{\infty} \left| \frac{a_m(x)}{a_0(x)} \right| \leq \lambda < 1 \tag{3.13}$$

as in [15]. Additionally, we assume that the sequence $(\|a_m/a_0\|_{L^\infty(D)})_{m \in \mathbb{N}}$ is arranged in descending order and in ℓ^2. With this, we obtain

$$\mathcal{A}(y) = \mathcal{A}_0 + \sum_{m=1}^{\infty} \mathcal{A}_m y_m$$

with

$$\mathcal{A}_m = \mathcal{X} \to \mathcal{X}^*, \quad v_x \mapsto -\nabla \cdot (a_m \nabla v_x)$$

for all $m \in \mathbb{N}_0$. The series converges in $\mathcal{L}(\mathcal{X}, \mathcal{X}^*)$.

Note that the energy norm (3.8) is not a crossnorm. However, for the tensor treatment of the problem, this is necessary. Thus, in accordance with [15, 16, 17], we endow $\mathcal{V}$ with a different scalar product, the mean energy scalar product

$$\begin{aligned}(v,w)_{\mathcal{V}} &:= \int_\Gamma \langle \mathcal{A}_0 v(y), w(y) \rangle \, \mathrm{d}\pi(y) \\ &= \int_\Gamma \int_D a_0(x) \nabla v(x,y) \cdot \nabla w(x,y) \, \mathrm{d}x \mathrm{d}\pi(y)\end{aligned}$$

and the induced mean energy norm $\|v\|_{\mathcal{V}} = \sqrt{(v,v)_{\mathcal{V}}}$. This is clearly a crossnorm, as the mean a_0 only depends on the deterministic variable x.

As a direct consequence of (3.13), we get that the two norms are equivalent [25]:

Lemma 3.2. *It holds*

$$(1-\lambda)\|v\|_{\mathcal{A}_0} \leq \|v\|_{\mathcal{A}} \leq (1+\lambda)\|v\|_{\mathcal{A}_0}.$$

Therefore, in the affine case, this will be the scalar product of the Hilbert space $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ and the regular scalar product on $L^2_\pi(\Gamma, H^1_0(D))$ will be denoted explicitly if necessary. Similarly, the norm on $\mathcal{X}$ will be given by the scalar product

$$(v_x, w_x)_{\mathcal{X}} := \int_D a_0(x) \nabla v_x(x) \cdot \nabla w_x(x) \, \mathrm{d}x$$

in the affine case.

Variational Formulation and Tensor Structure

We obtain the variational formulation (3.5)

$$B(u,v) = (u,v)_{\mathcal{A}} = \langle f, v \rangle \quad \forall v \in \mathcal{V} \tag{3.14}$$

for which one can show the following [25]:

Theorem 3.3. *For any $f \in \mathcal{V}^*$, the solution u of* (3.1) *is the unique solution to the variational problem* (3.14) *in $\mathcal{V}$ and it holds*

$$\|u\|_{\mathcal{V}} \lesssim \|f\|_{\mathcal{V}^*}.$$

Since $\mathcal{X}$ is separable, Theorem 2.6 yields that the Lebesgue-Bochner space $\mathcal{V} = L^2_\pi(\Gamma, \mathcal{X})$ has tensor product structure

$$\mathcal{V} = \mathcal{X} \otimes \mathcal{Y} \quad \text{with} \quad \mathcal{Y} = \bigotimes_{m=1}^{\infty} L^2_{\pi_m}(\Gamma_m).$$

We define the multiplication operators

$$\mathcal{K}_m : \mathcal{Y} \to \mathcal{Y}, \quad [y \mapsto v_y(y)] \mapsto [y \mapsto y_m v_y(y)].$$

Then the full operator has the form

$$\mathcal{A} : \mathcal{V} \to \mathcal{V}^*, \quad \mathcal{A} = \mathcal{A}_0 \otimes \mathrm{id}_{\mathcal{Y}} + \sum_{m=1}^{\infty} \mathcal{A}_m \otimes \mathcal{K}_m \tag{3.15}$$

and it is boundedly invertible by Theorem 3.3, i.e. it is bounded and admits a bounded inverse.

3.1.3 The Log-Normal Problem

We consider again the parametric diffusion equation (3.9). However, for the log-normal problem, the logarithm of the coefficient a is expanded as a series

$$a(x, y) = \exp\left(a_0(x) + \sum_{m=1}^{\infty} b_m(x) y_m\right) = \exp(a_0(x)) \prod_{m=1}^{\infty} \exp(b_m(x) y_m) \tag{3.16}$$

and the parameters $y = (y_m)_{m \in N} \in \mathbb{R}^\infty$ are independent standard Gaussian random variables, i.e. they are distributed by the product measure $\gamma = \bigotimes_{m \in \mathbb{N}} \mathcal{N}_1$,

$$\mathrm{d}\gamma(y) = \prod_{m=1}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y_m^2\right) \mathrm{d}y_m.$$

Since dealing with the series $b_0(x) + \sum_{m=1}^{\infty} b_m(x) y_m$ directly would result in negative values and thus destroy ellipticity of the operator, it is necessary and customary to consider the exponential of the expansion. If $\log(a)$ is Gaussian, then the series would be the result of a Karhunen-Loève expansion as explained above. For this reason, the log-normal diffusion equation is a more realistic example of a parametric PDE.

Note that because of the nature of the exponential function, the mean b_0 of the exponent simply results in an extra factor that is independent of the parameter. Therefore, for simplicity of notation, we will assume

$$b_0(x) \equiv 0, \quad \exp(b_0(x)) \equiv 1$$

and omit this factor in the following.

The first step to solving the log-normal problem would be to ensure the uniform boundedness and coercivity conditions (3.6) and (3.7) as done in Section 3.1.2 in order to ensure that the problem is well-posed. However, since the parameters y_m can be arbitrarily small, the coefficient cannot be bounded away from zero. As mentioned before, this problem can be overcome in different ways, all of which rely more or less on the introduction of an additional weight function that counterbalances the vanishing coefficient [60, 61]. This weight can either be introduced explicitly, as done in [23], which leads to a Petrov-Galerkin ansatz where the solution space and the test space are different, or it can be incorporated in the Gauss measure, thus altering both the solution space and the test space in the same way. The latter approach has been formulated most notably in [75] and we will follow their reasoning very closely. Wherever proofs of the lemmata and theorems in this section are left out, we refer the reader to said article.

First of all, in the abstract case presented here, the series in (3.16) may not converge and we therefore assume

$$b_m \in L^\infty(D), \quad \|b_m\|_{L^\infty(D)} =: \alpha_m$$

for all $m \in \mathbb{N}$ and

$$\sum_{m=1}^{\infty} \alpha_m < \infty, \tag{3.17}$$

i.e. $(\alpha_m)_{m\in\mathbb{N}} \subset \ell^1(\mathbb{N})$. Then the series converges in $L^\infty(D)$ for all y in the set

$$\Gamma := \left\{ y \in \mathbb{R}^\infty : \sum_{m=1}^{\infty} \alpha_m |y_m| < \infty \right\} \tag{3.18}$$

for which the following two lemmata hold.

Lemma 3.4. *$\Gamma \in \mathfrak{B}(\mathbb{R}^\infty)$ and $\gamma(\Gamma) = 1$.*

Lemma 3.5. *For all $y \in \Gamma$, the diffusion coefficient* (3.16) *is well-defined and satisfies*

$$0 < a_{\min}(y) =: \operatorname*{ess\,inf}_{x\in D} a(x,y) \le \operatorname*{ess\,sup}_{x\in D} a(x,y) =: a_{\max}(y) < \infty. \tag{3.19}$$

Lemma 3.4 implies that all parameters y, for which the diffusion coefficient does not converge, lie in a null set and will therefore be attained with probability zero. For this reason, even though the parameter space Γ is not a product domain, we can define the product Gauss measure γ on this set by restriction. Lemma 3.5 shows that a cannot be uniformly bounded away from zero but it has a lower bound $a_{\min}$ that is itself a random variable. This fact will be exploited below. In the same way, the coefficient can be bounded from above by a random variable $a_{\max}$.

The Lax-Milgram theorem will therefore only yield the following result:

Theorem 3.6. *For all $y \in \Gamma$, the variational problem with bilinear form* (3.2) *and right side* (3.3)

$$B(y; u(y), v) = f(y; v(y)) \quad \forall v \in \mathcal{X} \tag{3.20}$$

has a unique solution $u(y) \in \mathcal{X}$ that satisfies

$$\|u(y)\|_{\mathcal{X}} \le \frac{1}{a_{\min}(y)} \|f(y)\|_{\mathcal{X}^*} \quad \forall y \in \Gamma. \tag{3.21}$$

This solution is only given point-wise. It is clear that the bound is useless if $a_{\min}$ approaches zero. This means that $u(y)$ is not necessarily integrable over the parameters and a numerical approach might fail. For this reason, we will introduce auxiliary Gaussian measures that counterbalance the vanishing lower bound.

Auxiliary Gaussian Measures

For $\rho \in \mathbb{R}$ we define the sequence

$$\sigma = (\sigma_m)_{m\in\mathbb{N}} := (\exp(\rho\alpha_m))_{m\in\mathbb{N}} \in \ell^1(\mathbb{N}),$$

which is clearly convergent because of (3.17). Furthermore, we define the product Gaussian measure

$$\gamma_\rho := \bigotimes_{m\in\mathbb{N}} \mathcal{N}_{\sigma_m^2}$$

on $(\mathbb{R}^\infty, \mathfrak{B}(\mathbb{R}^\infty))$ and therefore also on the parameter space Γ as a product of centred Gaussian measures on $\mathbb{R}$ with standard deviation σ_m. This is a generalisation of the standard Gaussian measure and in accordance with the above, we omit the subscript if $\rho = 0$. Then in relation with the measures defined before, we obtain the functions

$$\zeta_\rho(y) := \prod_{m=1}^{\infty} \zeta_{\rho,m}(y_m),$$

$$\zeta_{\rho,m}(y_m) := \frac{1}{\sigma_m} \exp\left(-\frac{1}{2}(\sigma_m^{-2} - 1) y_m^2 \right),$$

such that

$$\mathrm{d}\gamma_\rho(y) = \zeta_\rho(y)\mathrm{d}\gamma(y)$$

and consequently

$$\mathrm{d}\gamma_\rho(y) = \prod_{m=1}^{\infty} \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{1}{2\sigma_m^2} y_m^2\right) \mathrm{d}y_m.$$

For $\vartheta < \rho$, γ_ρ is a *stronger* measure than γ_ϑ in the sense of the following lemma.

Lemma 3.7. *Let $\vartheta < \rho$. Then for all $y \in \Gamma$*

$$\frac{\zeta_\vartheta(y)}{\zeta_\rho(y)} a_{\max}(y) \leq \hat{c}_{\vartheta,\rho} \tag{3.22}$$

and

$$\frac{\zeta_\rho(y)}{\zeta_\vartheta(y)} a_{\min}(y) \geq \check{c}_{\vartheta,\rho}$$

for some constants $\hat{c}_{\vartheta,\rho}, \check{c}_{\vartheta,\rho} > 0$.

Proof. With (3.19) it holds for all $y \in \Gamma$

$$\exp\left(-\sum_{m=1}^{\infty} \alpha_m |y_m|\right) \leq a_{\min}(y) \leq a_{\max}(y) \leq \exp\left(\sum_{m=1}^{\infty} \alpha_m |y_m|\right).$$

Then for all $y \in \Gamma$, using the inequality

$$\exp(-2\rho\alpha_m) - \exp(-2\vartheta\alpha_m) = \exp(-2\rho\alpha_m)\big(1 - \exp(2(\rho-\vartheta)\alpha_m)\big) \leq -2\exp(-2\rho\alpha_m)\big((\rho-\vartheta)\alpha_m\big),$$

it holds

$$\begin{aligned}
\frac{\zeta_\vartheta(y)}{\zeta_\rho(y)} a_{\max}(y) &\leq \exp\left(\sum_{m=1}^{\infty}\Big((\rho-\vartheta)\alpha_m + \frac{1}{2}(\exp(-2\rho\alpha_m) - \exp(-2\vartheta\alpha_m))y_m^2 + \alpha_m|y_m|\Big)\right) \\
&\leq \exp\left(\sum_{m=1}^{\infty}\Big((\rho-\vartheta)\alpha_m - \exp(-2\rho\alpha_m)(\rho-\vartheta)\alpha_m y_m^2 + \alpha_m|y_m|\Big)\right) \\
&= \exp\Bigg(\sum_{m=1}^{\infty}\alpha_m\Big((\rho-\vartheta) - \big(\sqrt{\rho-\vartheta}\exp(-\rho\alpha_m)|y_m| - \frac{\exp(\rho\alpha_m)}{2\sqrt{\rho-\vartheta}}\big)^2 \\
&\qquad\qquad + \frac{\exp(2\rho\alpha_m)}{4(\rho-\vartheta)}\Big)\Bigg) \\
&\leq \exp\left(\sum_{m=1}^{\infty}\alpha_m\Big((\rho-\vartheta) + \frac{\exp(2\rho\alpha_m)}{4(\rho-\vartheta)}\Big)\right) =: \hat{c}_{\vartheta,\rho},
\end{aligned}$$

and therefore also

$$\frac{\zeta_\rho(y)}{\zeta_\vartheta(y)} a_{\min}(y) \geq \frac{\zeta_\rho(y)}{\zeta_\vartheta(y)} \exp\left(-\sum_{m=1}^{\infty}\alpha_m|y_m|\right) \geq (\hat{c}_{\vartheta,\rho})^{-1} =: \check{c}_{\vartheta,\rho}.$$

□

Using similar arguments, we can show the embeddings

Proposition 3.8. *Let $\vartheta < \rho$. Then*

$$L^2_{\gamma_\rho}(\Gamma,\mathcal{X}) \subset L^2_{\gamma_\vartheta}(\Gamma,\mathcal{X}),$$

i.e.

$$\|v\|_{L^2_{\gamma_\vartheta}(\Gamma,\mathcal{X})} \lesssim \|v\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})} \quad \forall v \in L^2_{\gamma_\rho}(\Gamma,\mathcal{X}).$$

Here, in addition to the norm on $\mathcal{V} = L^2_\gamma(\Gamma, \mathcal{X})$, whose notation remains unchanged, we have used the norms on $L^2_{\gamma_\rho}(\Gamma, \mathcal{X})$ that we will denote explicitly. These also entail a different dual pairing that we will denote with $\langle f, v\rangle_\rho$ for $f \in L^2_{\gamma_\rho}(\Gamma, \mathcal{X}^*)$ and $v \in L^2_{\gamma_\rho}(\Gamma, \mathcal{X})$.

Before we define the variational problem, we can give a statement about the integrability of the point-wise solution:

Proposition 3.9. *Let $\rho > 0$ and $f \in L^2_{\gamma_\rho}(\Gamma, \mathcal{X}^*)$. Then the solution u in Theorem 3.6 is in $L^2_\gamma(\Gamma, \mathcal{X})$ and it holds*

$$\|u\|_{L^2_\gamma(\Gamma,\mathcal{X})} \lesssim \|f\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X}^*)}.$$

Proof. Borel-measurability of the map $y \mapsto u(y)$ is shown in [24] under the assumption that the right hand side is measurable. By (3.21), it holds

$$\begin{aligned}\int_\Gamma \|u(y)\|^2_{\mathcal{X}} \,\mathrm{d}\gamma(y) &\le \int_\Gamma \frac{1}{a_{\min}(y)^2}\|f(y)\|^2_{\mathcal{X}^*}\,\zeta_\rho^{-1}(y)\mathrm{d}\gamma_\rho(y)\\ &\le \big(\operatorname*{ess\,inf}_{y\in\Gamma} \zeta_\rho^{-1}(y) a_{\min}(y)^{-2}\big)\int_\Gamma \|f(y)\|^2_{\mathcal{X}^*}\,\mathrm{d}\gamma_\rho(y).\end{aligned}$$

The claim follows with a calculation similar to the proof of Lemma 3.7. □

Variational Formulation and Solution

The diffusion coefficient a is neither uniformly bounded nor uniformly coercive on Γ, which means that simply integrating the variational problem as in (3.5) does not yield a well-posed linear variational problem on $L^2_\gamma(\Gamma, \mathcal{X})$. We will therefore integrate with respect to a measure that is stronger than γ but not as strong as γ_ρ.

From now on we set $0 \le \vartheta < 1$ and $\rho > 0$ and we define

$$\begin{aligned}B_{\vartheta\rho}(v, w) &:= \int_\Gamma B(y; v(y), w(y))\,\mathrm{d}\gamma_{\vartheta\rho}(y)\\ &= \int_\Gamma \int_D a(x, y)\nabla v(x, y)\cdot\nabla w(x, y)\,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y)\end{aligned}$$

and for $f \in L^2_{\gamma_\rho}(\Gamma, \mathcal{X}^*)$

$$\begin{aligned}f_{\vartheta\rho}(v) &:= \int_\Gamma f(y; v(y))\,\mathrm{d}\gamma_{\vartheta\rho}(y)\\ &= \int_\Gamma \int_D f(x, y)v(x, y)\,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y).\end{aligned}$$

Since $\vartheta\rho < \rho$, the right hand side $f_{\vartheta\rho}$ is well defined using Proposition 3.8.

Then we can show properties similar to uniform boundedness and uniform coercivity:

Lemma 3.10. *For all $v, w \in L^2_{\gamma_\rho}(\Gamma, \mathcal{X})$, it holds*

$$|B_{\vartheta\rho}(v, w)| \le \hat{c}_{\vartheta\rho,\rho}\|v\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})}\|w\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})}.$$

For all $v \in L^2_\gamma(\Gamma, \mathcal{X})$ with $B_{\vartheta\rho}(v, v) < \infty$, it holds

$$B_{\vartheta\rho}(v, v) \ge \check{c}_{0,\vartheta\rho}\|v\|^2_{L^2_\gamma(\Gamma,\mathcal{X})}. \tag{3.23}$$

Proof. Let $v, w \in L^2_{\gamma_\rho}(\Gamma, \mathcal{X})$. Using the Cauchy-Schwarz inequality and (3.22), we get

$$\begin{aligned}|B_{\vartheta\rho}(v, w)| &= \left|\int_\Gamma \int_D a(x, y)\nabla v(x, y)\cdot\nabla w(x, y)\,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y)\right|\\ &= \left|\int_\Gamma \int_D \left(\frac{\zeta_{\vartheta\rho}(y)}{\zeta_\rho(y)}a(x, y)\right)\nabla v(x, y)\cdot\nabla w(x, y)\,\mathrm{d}x\mathrm{d}\gamma_\rho(y)\right|\\ &\le \hat{c}_{\vartheta\rho,\rho}\|v\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})}\|w\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})}.\end{aligned}$$

Let $v \in L^2_\gamma(\Gamma, \mathcal{X})$ with $B_{\vartheta\rho}(v,v) < \infty$. Similar to above, we get

$$\begin{aligned} B_{\vartheta\rho}(v,v) &= \int_\Gamma \int_D a(x,y)\nabla v(x,y)\cdot\nabla v(x,y)\,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y) \\ &= \int_\Gamma \int_D \big(\zeta_{\vartheta\rho}(y)a(x,y)\big)\nabla v(x,y)\cdot\nabla v(x,y)\,\mathrm{d}x\mathrm{d}\gamma(y) \\ &\geq \check{c}_{0,\vartheta\rho}\|v\|^2_{L^2_\gamma(\Gamma,\mathcal{X})}. \end{aligned}$$

□

We define the vector space

$$\mathcal{V}_{\vartheta\rho} := \{v : \Gamma \to \mathcal{X} \text{measurable} : B_{\vartheta\rho}(v,v) < \infty\}.$$

The full operator (3.4) is given on this space as $\mathcal{A} : \mathcal{V}_{\vartheta\rho} \to \mathcal{V}^*_{\vartheta\rho}$. As opposed to the affine case and (3.8), we define the energy norm for $v \in \mathcal{V}_{\vartheta\rho}$ as

$$\|v\|_{\mathcal{A}} = \sqrt{B_{\vartheta\rho}(v,v)}.$$

Then we can state the following:

Proposition 3.11. *The space $\mathcal{V}_{\vartheta\rho}$ endowed with the inner product $B_{\vartheta\rho}(\cdot,\cdot)$ is a Hilbert space. It holds*

$$L^2_{\gamma_\rho}(\Gamma,\mathcal{X}) \subset \mathcal{V}_{\vartheta\rho} \subset L^2_\gamma(\Gamma,\mathcal{X}).$$

Proof. The embeddings follow immediately from Lemma 3.10. The lemma also implies that $B_{\vartheta\rho}(\cdot,\cdot)$ is an inner product. It remains to show that $\mathcal{V}_{\vartheta\rho}$ is complete with respect to the energy norm. This is shown analogously to the completeness of regular Lebesgue spaces and can be found in [24]. □

We required above that $f \in L^2_{\gamma_\rho}(\Gamma,\mathcal{X}^*)$. This is reasonable, as then $f_{\vartheta\rho} \in \mathcal{V}^*_{\vartheta\rho}$ according to Proposition 3.11. With this, we can finally state the theorem from [75].

Theorem 3.12. *Let $f \in L^2_{\gamma_\rho}(\Gamma,\mathcal{X}^*)$ and therefore $f_{\vartheta\rho} \in \mathcal{V}^*_{\vartheta\rho}$. Then the solution u of the point-wise problem* (3.20) *is the unique solution in $\mathcal{V}_{\vartheta\rho}$ of the linear variational problem*

$$B_{\vartheta\rho}(u,v) = f_{\vartheta\rho}(v) \quad \forall v \in \mathcal{V}_{\vartheta\rho}. \tag{3.24}$$

In summary, the log-normal diffusion coefficient is not uniformly bounded or coercive in $L^2_\gamma(\Gamma,\mathcal{X})$. This means that the variational problem on this space is ill-posed. We overcame this by integrating with respect to a stronger measure $\gamma_{\vartheta\rho}$. Then we obtain a weaker boundedness with respect to the norm of the smaller space $L^2_{\gamma_\rho}(\Gamma,\mathcal{X})$ and a weaker coercivity with respect to the norm of the bigger space $L^2_\gamma(\Gamma,\mathcal{X})$. However, this is enough to show that a unique solution exists in the Hilbert space $\mathcal{V}_{\vartheta\rho}$ and that this coincides with the solution of the point-wise problem that is in $L^2_\gamma(\Gamma,\mathcal{X})$ if the right hand side is sufficiently integrable.

Later we will show what is required in order to ensure $u \in L^2_{\gamma_\rho}(\Gamma,\mathcal{X})$ as this will be necessary to obtain a tensor product representation of u. As opposed to the affine case, we do not have a nice tensor product representation of the operator $\mathcal{A}$. This can only be done by approximating it, which we will discuss in Chapter 6.

3.1.4 The Localised Uncertainty Problem

Let us briefly introduce a third example of parametric PDEs without presenting any solution theory. Here, the coefficient is vastly different on several distinct areas $D_m \subset D$ of the domain with

$$\bigcup_{m=1}^{M} D_m = D, \quad D_m \cup D_n = \emptyset, \quad m \neq n.$$

This is not necessarily stochastic but it can be. The resulting problem therefore has localised independent uncertainty. The coefficient is given as

$$a(x,y) = \sum_{m=1}^{M} \chi_m(x) a_m(y_m),$$
$$\chi_m = \begin{cases} 1, & x \in D_m, \\ 0, & x \notin D_m, \end{cases}$$

and the local functions a_m could for example be material constants depending on some random perturbation. One could assume that their dependence on y_m can be modelled with a simple probability distribution, e.g., if the noise is just additive:

$$a_m(y_m) = \bar{a}_m + y_m.$$

Otherwise, it would be possible to perform a Karhunen-Loève decomposition on each of these functions, which has been done in [32].

As opposed to a coefficient expansion that results from a Karhunen-Loève decomposition, the characteristic functions χ_m in the localised problem are all a priori equally important. Therefore, there is no decay in the coefficient and this example can serve as a motivation to use other tensor formats on some parametric problems, as we will briefly discuss in Chapter 6.

4 Numerical Solutions of Parametric PDEs

It is usually impossible to find an analytic solution of a partial differential equation. This would require that the solution can be expressed as a finite combination of previously known functions. However, the solution spaces in this thesis are infinite-dimensional Hilbert spaces and thus functions that are expressible by a finite combination of unrelated functions are uncommon in these spaces. This means that, while a problem can be constructed to have such a function as a solution, this is not the case in general.

Hence, the exact solution u of the operator equation (2.2) can in general not be expressed in any other more meaningful way. Finding the *numerical solution* means that we want to express u as a finite combination of other functions that are more intuitive to us, e.g. polynomials, wavelets or trigonometric functions. Such a *representation* will inherently be only an approximation of the exact solution and it can be a good or a bad approximation depending on how small the error is. We will discuss error estimation in Chapter 7. Additionally, it makes sense to call an approximation *good* if it requires few combinations of other functions to express the solution with a fixed error. Thus, some functions can be better at approximating a certain problem than others and it is often not trivial to know which basis set to use. However, in many cases, this has been studied extensively and the optimal basis set is known.

Often, we do not need to know a good approximation of the full solution u but it suffices to know some of its properties. It then makes sense to invest the computing power only into finding these properties instead of constructing a good approximation of the full solution. For stochastic PDEs, this could be the case if we are interested only in finding the mean and/or the variance of the solution.

Before we introduce some techniques for finding numerical solutions of parametric PDEs, let us repeat that in our case, the solution space and the test space of the variational problem (3.5) are the Lebesgue-Bochner spaces

$$\mathcal{V} = L^2_\pi(\Gamma, \mathcal{X})$$

on some parameter space Γ and with a measure π. As we have discussed before, these spaces are isometrically isomorphic to the tensor product space

$$\mathcal{V} = \mathcal{X} \otimes \mathcal{Y} \quad \text{with} \quad \mathcal{Y} = L^2_\pi(\Gamma).$$

Monte Carlo Methods

In the case of stochastic PDEs, the parametric aspect of the PDE is given by random perturbations of an otherwise deterministic setting. This means the parameter space Γ is a transformation of the probability space $(\Omega, \Sigma, \mathbb{P})$ with some measurable set and probability distribution. In many cases, only the mean $u_0 := \mathbb{E}(u) \in \mathcal{X}$ of the solution is of interest. However, this cannot simply be calculated by solving

$$\mathcal{A}_0 u = f$$

even if $f(y) \equiv f(0)$, i.e. if the right hand side is completely deterministic, because in general, at least the operator also depends on the parameter y.

A common approach to approximate the mean u_0 is Monte Carlo (MC) sampling, which is done by calculating the deterministic solution of

$$\mathcal{A}(y_{[j]})u(y_{[j]}) = f(y_{[j]})$$

at many samples $y_{[j]}, j = 1, \dots, P_{\text{MC}}$, and then taking the average of these solutions. If the samples are for example chosen randomly (as opposed to other approaches), then the law of large numbers ensures convergence of the MC method. Furthermore, the steps can be parallelised, which can improve computation times significantly [75].

Many variations of this method have been proposed (QuasiMC, Multi-level MC, etc.), but they all ultimately suffer from similar disadvantages. First of all, convergence rates of MC are usually slow. In general, the method only converges with $\mathcal{O}(\sqrt{P_{\mathrm{MC}}})$, thus being computationally expensive. More importantly however, this method can only compute certain moments of the solution, rather than the full solution. This is often enough, but there are many applications in which more properties are needed.

In this thesis, we are interested in approximating the full solution u with all its moments. We will therefore also discretise the parametric part of the solution. This means that we will apply a Galerkin discretisation of both the deterministic and the parametric part, as seen in the following.

4.1 Discretisation of the Lebesgue-Bochner Space

In Monte Carlo methods, solving the deterministic part means finding a numerical solution of the deterministic problem (3.1) for some fixed $y \in \Gamma$. This is done by using some kind of discretisation of the deterministic solution space $\mathcal{X} = H_0^1(D)$ and thus also of the test space, since they are the same. The full operator equation (2.2) is defined on the Lebesgue-Bochner space $\mathcal{V} = L^2_\pi(\Gamma, \mathcal{X})$ which has tensor product structure. This means that the treatment of the deterministic and the parametric part can be done independently.

4.1.1 Discretisation of the Deterministic Space

The discretisation of the deterministic space $\mathcal{X}$ is done in the same way as for a fully deterministic problem and the theory is therefore well known and basic. In this thesis, we employ a finite element discretisation.

The Finite Element method

The finite element (FEM) discretisation of the deterministic space uses a simplicial triangulation $\mathcal{T}$ of the deterministic domain $D \subset \mathbb{R}^2$(or $\mathbb{R}^3$), i.e. a decomposition into triangles (or tetrahedra) $T \in \mathcal{T}$ using a homeomorphism $\mathcal{T} \leftrightarrow D$. For the sake of simplicity, we will assume that D is a polygon that is partitioned by $\mathcal{T}$, but more general cases can be considered with slight modifications. The set of edges S is denoted by $\mathcal{S}$ and thus the edges of one triangle T are given by $\mathcal{S} \cap \partial T$. Additionally, we (informally) define the distances

$$h_T := \operatorname{diam}(T), \quad h_S := \operatorname{diam}(S)$$

as the element size and edge size respectively.

We define the finite element space

$$\mathcal{X}_p(\mathcal{T}) = \{v_x \in \mathcal{C}^0(\overline{D}) : v_{x|T} \in \mathcal{P}_p(T)\ \forall T \in \mathcal{T}\} \tag{4.1}$$

of continuous piecewise polynomial functions of degree p on the triangulation $\mathcal{T}$. Moreover, we require *conformity*, i.e. $\mathcal{X}_p(\mathcal{T}) \subset \mathcal{X}$, in order to obtain a Galerkin discretisation. Then there exists a nodal basis $(\varphi_i)_{i=0}^{N-1}$, $\operatorname{span}(\varphi_1, \dots, \varphi_N) = \mathcal{X}_p(\mathcal{T})$ with $N = \dim(\mathcal{X}_p(\mathcal{T}))$. The functions φ_i are supported on all faces containing the i-th node and thus they fulfil the *small support* condition of finite element functions [75].

It will be necessary to define the Clément interpolation operator $Q : \mathcal{X} \to \mathcal{X}_p(\mathcal{T})$ that fulfils

$$\begin{aligned} \|v_x - Qv_x\|_{L^2(T)} &\le c_{\mathcal{T}} h_T |v_x|_{\mathcal{X},T} \quad \forall T \in \mathcal{T}, \\ \|v_x - Qv_x\|_{L^2(S)} &\le c_{\mathcal{S}} h_S^{1/2} |v_x|_{\mathcal{X},S} \quad \forall S \in \mathcal{S}, \end{aligned}$$

where $|\cdot|_{\mathcal{X},T}$ and $|\cdot|_{\mathcal{X},S}$ denote the seminorm that is the restriction of $\|\cdot\|_{\mathcal{X}}$ to the union of all elements of $\mathcal{T}$ sharing at least a vertex with T or S, respectively. This is a standard result that can be found for example in [9].

Then every function $v \in \mathcal{V}$ can be approximated in the semi-discretised form

$$v(x,y) \approx v_N(x,y) := \sum_{i=0}^{N-1} v_{N,i}(y)\varphi_i(x) \tag{4.2}$$

with $v_{N,i} \in \mathcal{Y}$ for all $0 \leq i < N$. It holds $v_N \in \mathcal{X}_p(\mathcal{T}) \otimes \mathcal{Y}$ and thus for fixed $y \in \Gamma$, $v_N(y) \in \mathcal{X}_p(\mathcal{T})$.

4.1.2 Discretisation of the Parametric Space

We discretise the parametric space $\mathcal{Y} = L^2_\pi(\Gamma)$ using a *generalised polynomial chaos* (GPC) approach. This has been first proposed by Wiener [91] and Cameron and Martin [10] but has since undergone some changes. We refer to the works [24, 75, 25] for more insight.

Polynomial Chaos

For now we will assume that the parameter space has a cartesian structure $\Gamma = \bigtimes_{m\in\mathbb{N}} \Gamma_m$. We further assume that each Γ_m is a Borel subset of $\mathbb{R}$ and that the measure π on Γ is a product measure $\pi = \bigotimes_{m\in\mathbb{N}} \pi_m$ such that every π_m is *determinate*, i.e. it is uniquely characterised by its (finite) moments.

Then we obtain a tensor product structure

$$L^2_\pi(\Gamma) = \bigotimes_{m\in\mathbb{N}} L^2_{\pi_m}(\Gamma_m)$$

and we define $\mathcal{Y}_m := L^2_{\pi_m}(\Gamma_m)$ for convenience.

For $y_m \in \Gamma_m$ we construct polynomials $(P_{\mu_m})_{\mu_m\in\mathbb{N}}$ by the three-term recursion

$$\beta_{\mu_m+1}P_{\mu_m+1}(y_m) = y_m P_{\mu_m}(y_m) - \beta_{\mu_m}P_{\mu_m-1}(y_m) \quad \forall \mu_m \in \mathbb{N} \tag{4.3}$$

with the initialisation $P_{-1} \equiv 0$ and $\beta_0 := 0$. The coefficients β_{μ_m} are uniquely determined by the requirement that the series $(P_{\mu_m})_{\mu_m\in\mathbb{N}}$ is an orthonormal polynomial basis of $\mathcal{Y}_m$ of degree $\deg(P_{\mu_m}) = \mu_m$. This property can be ensured since the measure π_m is determinate [75].

We further define the space of finitely supported multi-indices

$$\mathcal{F} = \{\mu = (\mu_m)_{m\in\mathbb{N}} \in \mathbb{N}_0^\infty : |\operatorname{supp}\mu| < \infty\},$$

where $\operatorname{supp}\mu := \{m \in \mathbb{N} : \mu_m \neq 0\}$. Then any function $v \in \mathcal{V}$ admits a semi-discretised representation in the parametric domain

$$v(x,y) = \sum_{\mu\in\mathcal{F}} v_\mu(x)P_\mu(y) \tag{4.4}$$

where $P_\mu := \bigotimes_{m\in\mathbb{N}} P_{\mu_m}$. This is called the *polynomial chaos expansion* and it is not an approximation but a full representation of v, because the parametric space $\mathcal{Y}$ is discretised but still infinite-dimensional.

In order to get a finite dimensional Galerkin space, we define the finite subset

$$\Lambda \subset \mathcal{F}, \quad |\Lambda| < \infty.$$

This yields a Galerkin space

$$\mathcal{Y}(\Lambda) := \left\{ v \in \mathcal{Y} : v(y) = \sum_{\mu\in\Lambda} v_\mu P_\mu(y) \right\}$$

and for $v \in \mathcal{V}$, we get the semi-discretised approximation

$$v(x,y) \approx v_\Lambda(x,y) := \sum_{\mu\in\Lambda} v_{\Lambda,\mu}(x)P_\mu(y) \tag{4.5}$$

with $v_{\Lambda,\mu} \in \mathcal{X}$ for all $\mu \in \Lambda$. As above, it holds $v_\Lambda \in \mathcal{X} \otimes \mathcal{Y}(\Lambda)$.

In particular, since Λ is finite, this entails that there is a number $M \in \mathbb{N}$ such that for all $\mu \in \mathcal{F}$ we get

$$\mu_m = 0 \quad \forall m > M.$$

We call the smallest such M the *truncation parameter* of the discretised parametric solution space.

Tensor Set vs. Best N-term Approximation

There are several possibilities to choose the subset $\Lambda \subset \mathcal{F}$ that contains the indices for polynomials in the Galerkin basis. It could be possible to select a fixed subset using *a priori* information and sticking with it for finding a numerical solution. However, this information is hard to come by and other approaches seem more practical. In this thesis, we follow the widely used approach of *a posteriori* adaptivity. This means that we begin the approximation in some small finite dimensional Galerkin space, for instance only polynomials of degree one, and then adaptively refine it in order to improve the solution. For this, we estimate the error and adapt the Galerkin space accordingly. For the deterministic part, this means that we try to estimate how to refine the mesh. For the parametric part, we want to adaptively enlarge the space Λ and thus possibly also the truncation parameter M which depends on it.

Error estimation and adaptivity will be discussed in Chapter 7. For now we want to remark on a substantial decision that has to be made at this point. There are basically two competing approaches for choosing the index set Λ and they are discussed in more detail in [4]. An intuitive approach is to estimate which indices $\mu \in \mathcal{F} \setminus \Lambda$ would decrease the error the most when included in the set Λ and then include a fixed number of them in each iteration until the solution is approximated sufficiently well or Λ reaches a certain maximum size. This is called *best N-term approximation* as we only include the best terms in the approximation. Bachmayr and Dahmen have shown that this approach is preferable under certain assumptions. In fact, if the basis polynomials from the GPC are optimally chosen, then the semi-discrete approximation v_Λ in (4.5) is also optimally sparse in $\mathcal{V} = \ mathcalX \otimes \mathcal{Y}$ because the parametric discretisation is spanned by optimal basis functions

$$\mathcal{Y}(\Lambda) = \operatorname*{span}_{\mu \in \Lambda}(P_\mu).$$

However, there are two disadvantages to this approach. Firstly, we also want to discretise the deterministic space $\mathcal{X}$. Optimality can only be retained if we allow a different discretisation for every deterministic coefficient function $v_\mu \in \mathcal{X}$ in (4.5), i.e. if we can choose a different triangulation and possibly even a different degree p for every element of the basis set $\{P_\mu : \mu \in \Lambda\}$. This is allowed for example in [15] but it becomes very tedious and even notation suffers. Secondly, the best N-term approximation requires a treatment with sparse matrices as it is itself an approach akin to sparsity. Sparse matrices do not trivially allow a low rank treatment and other calculations become more difficult as well.

We do not outright reject this approach and we acknowledge its validity. We do not make any statement on which approach is preferable in general. For our purposes and to enable further research in this direction, we will choose the competing approach of low rank tensor decompositions. For this, we define Λ to be a full tensor set, i.e.

$$\Lambda = \{(\mu_1, \dots, \mu_M, 0, \dots) \in \mathcal{F} : 0 \leq \mu_m < q_m \ \forall 1 \leq m \leq M\}$$

where for all $1 \leq m \leq M$, $q_m \in \mathbb{N}$ are called the *GPC-degrees*.

The full tensor set Λ has $\prod_{m=1}^{M} q_m$ elements and the cardinality grows exponentially with M. This results in a significant enlargement of the space as opposed to best N-term approximation where Λ was only assumed to be monotone or *downward closed*, which means that if $\mu \in \Lambda$, then every combination ν with $\nu_m \leq \mu_m$ for all $m \leq M$ needs to also be an element of Λ. This yields only a linear or polynomial dependence on the order M. However, since the full tensor set is so large, it also yields a much more accurate approximation of the solution and we hope

to retain this property when we break the *curse of dimensionality* by applying low rank tensor techniques.

Using this index space, the discretised parametric space $\mathcal{Y}(\Lambda)$ is the tensor product space

$$\mathcal{Y}(\Lambda) = \bigotimes_{m=1}^{M} \mathcal{Y}_m(q_m)$$

with $\mathcal{Y}_m(q_m) := \operatorname{span}_{0 \leq \mu_m < q_m}(P_{\mu_m}) \subset L^2_\pi(\Gamma_m)$.

4.1.3 Full Discretisation and Galerkin Projection

We combine the discretised spaces $\mathcal{X}_p(\mathcal{T})$ and $\mathcal{Y}(\Lambda)$ in order to get the fully discretised space

$$\mathcal{V}(p, \mathcal{T}, \Lambda) := \mathcal{X}_p(\mathcal{T}) \otimes \mathcal{Y}(\Lambda)$$

that we will abbreviate by $\mathcal{V}_{\text{disc}} := \mathcal{V}(p, \mathcal{T}, \Lambda) \subset \mathcal{V}$. Any function $v \in \mathcal{V}$ can be approximated by a function $v_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ as

$$v(x, y) \approx v_{\text{disc}}(x, y) = \sum_{i=0}^{N-1} \sum_{\mu \in \Lambda} V(i, \mu) \varphi_i(x) P_\mu(y). \tag{4.6}$$

The tensor product of the two discretised spaces means that each discretisation is independent of the other. Therefore, every deterministic function $v_{\Lambda,\mu} \in \mathcal{X}$ in (4.5) is approximated in the same discretised space $\mathcal{X}_p(\mathcal{T})$

$$v_{\Lambda,\mu}(x) \approx \sum_{i=0}^{N-1} V(i, \mu) \varphi_i(x)$$

and vice-versa for all $v_{N,i}(y) \in \mathcal{Y}$ in (4.2) we have an approximation in $\mathcal{Y}(\Lambda)$

$$v_{N,i}(y) \approx \sum_{\mu \in \Lambda} V(i, \mu) P_\mu(y).$$

The coefficients $V(i, \mu)$ are in the space

$$V \in \mathbb{R}^{N \times q_1 \times \cdots \times q_M}$$

and this space is isomorphic to the discrete, finite dimensional Hilbert space $\mathcal{V}_{\text{disc}}$:

$$\mathbb{R}^{N \times q_1 \times \cdots \times q_M} \cong \mathcal{X}_p(\mathcal{T}) \otimes \bigotimes_{m=1}^{M} \mathcal{Y}_m(q_m).$$

On this Galerkin space $\mathcal{V}_{\text{disc}}$, the variational problem (3.5) becomes the discrete Galerkin problem (2.5) with the linear operator

$$\mathbf{A} : \mathbb{R}^{N \times q_1 \times \cdots \times q_M} \to \mathbb{R}^{N \times q_1 \times \cdots \times q_M} \tag{4.7}$$

and the Galerkin projection of u onto $\mathcal{V}_{\text{disc}}$ is the unique element $u_{\text{disc}} \in V_{\text{disc}}$ that satisfies the variational problem (3.5):

$$B(u_{\text{disc}}, v_{\text{disc}}) = f(v_{\text{disc}}) \quad \forall v_{\text{disc}} \in \mathcal{V}_{\text{disc}}. \tag{4.8}$$

Since $\mathcal{V}_{\text{disc}}$ is a finite-dimensional subspace of $\mathcal{V}$, it is also a Hilbert space with the scalar product $B(\cdot, \cdot)$, and therefore u_{disc} is well-defined.

4.2 Tensor Product Structure of the Discretised Parametric Diffusion Equation

We state the representations of the Galerkin operator for the examples of the affine and the log-normal parametric diffusion equation. In its general form, it can be given element-wise as

$$\mathbf{A}(i,\mu;j,\mu') = \int_\Gamma \int_D a(x,y)\nabla\varphi_i(x)P_\mu(y)\cdot\nabla\varphi_j(x)P_{\mu'}(y)\,\mathrm{d}x\mathrm{d}\pi(y).$$

In the space $\mathcal{V}_{\vartheta\rho}$, the right hand side is given in its discretised form

$$\begin{aligned} F &\in \mathbb{R}^{N\times q_1\times\cdots\times q_M}, \\ F(j,\mu') &= \int_\Gamma\int_D f(x)\varphi_j(x)P_{\mu'}(y)\,\mathrm{d}x\mathrm{d}\pi(y) \\ &= \int_D f(x)\varphi_j(x)\,\mathrm{d}x\prod_{m=1}^{M}\int_\Gamma P_{\mu_m}(y_m)\,\mathrm{d}\pi_m(y_m), \end{aligned}$$

We will from now on assume that the right hand side f is independent of the parameters, i.e.

$$f(x,y)\equiv f(x)\quad\forall y\in\Gamma.$$

This simplifies notation and numerical treatment, but most results can be stated without loss of generality. We will remark on the integrability in the parameter domain where it becomes necessary. Therefore, F is the elementary tensor

$$F = f_0\otimes e_1^{(1)}\otimes\cdots\otimes e_1^{(M)}, \tag{4.9}$$

where the length is given by the truncation parameter M and

$$f_0(j) = \int_D f(x)\varphi_j(x)\,\mathrm{d}x$$

is a vector $f_0\in\mathbb{R}^N$ and the first unitary vectors $e_1^{(m)}\in\mathbb{R}^{q_m}$ result from the fact that

$$\int_\Gamma P_{\mu_m}(y_m)\,\mathrm{d}\pi_m(y_m) = \delta_{0\mu_m}.$$

This allows for an efficient tensor treatment that we will introduce in the following.

The Affine Problem

The discretisation of the affine problem with the operator $\mathcal{A}: L^2_\pi(\Gamma,\mathcal{X})\to L^2_\pi(\Gamma,\mathcal{X}^*)$ and parameter space $\Gamma=[-1,1]^\infty$ with uniform product measure π, as in Section 3.1.2 will be done using *Legendre polynomials* $(L_{\mu_m})_{\mu_m\in\mathbb{N}}$. With the above three-term recursion (4.3), we obtain

$$\left(4-(\mu_m+1)^{-2}\right)^{-1/2}L_{\mu_m+1} = y_mL_{\mu_m} - \left(4-\mu_m^{-2}\right)^{-1/2}L_{\mu_m-1},$$

i.e. $\beta_{\mu_m} = \left(4-\mu_m^{-2}\right)^{-1/2}$ for $\mu_m\in\mathbb{N}$. This can also be written explicitly when evaluated at $y_m\in\Gamma_m$ as

$$L_{\mu_m}(y_m) = \frac{\sqrt{2\mu_m+1}}{2^{\mu_m}\mu_m!}\frac{\partial^{\mu_m}}{\partial y_m^{\mu_m}}(y_m^2-1)^{\mu_m}\quad\forall\mu_m\in\mathbb{N}_0.$$

We allow this ambivalence in notation as it will be clear from context whether we denote Legendre polynomials or Lebesgue spaces and we define as above

$$L_\mu = \bigotimes_{m\in\mathbb{N}} L_{\mu_m}$$

for all finitely supported $\mu\in\mathcal{F}$.

The following proposition is shown in [75].

Proposition 4.1. *The tensor product Legendre polynomials $(L_\mu)_{\mu\in\mathcal{F}}$ form an orthonormal basis of $\mathcal{Y} = L^2_\pi(\Gamma)$.*

This will be referred to as the *Legendre chaos basis*. The deterministic space $\mathcal{X}$ is discretised as above using a finite element method and we choose a tensor subset $\Lambda \subset \mathcal{F}$. Then the functions in the Galerkin space $v_{\text{disc}} \in \mathcal{V}_{\text{disc}} = \mathcal{X}_p(\mathcal{T}) \otimes \mathcal{Y}(\Lambda)$ are given as

$$v_{\text{disc}}(x,y) = \sum_{i=0}^{N-1} \sum_{\mu\in\Lambda} V(i,\mu)\varphi_i(x)L_\mu(y).$$

As in Chapter 4.1.1, we can state the Clément interpolation operator $Q : \mathcal{X} \to \mathcal{X}_p(\mathcal{T})$ but with slightly different estimators

$$\begin{aligned}\|a_0^{1/2}(v_x - Qv_x)\|_{L^2(T)} &\le c_{\mathcal{T}} h_T |v_x|_{\mathcal{X},T} \quad \forall T \in \mathcal{T},\\ \|a_0^{1/2}(v_x - Qv_x)\|_{L^2(S)} &\le c_{\mathcal{S}} h_S^{1/2} |v_x|_{\mathcal{X},S} \quad \forall S \in \mathcal{S}.\end{aligned}$$

The extra weight $a_0^{1/2}$ has no effect as it is bounded from below and above [15]. Additionally, we define the tensor product interpolation operator

$$\begin{aligned}\mathcal{Q} &: \mathcal{V} \to \mathcal{V}_{\text{disc}},\\ \mathcal{Q}v &= \sum_{\mu\in\Lambda} (Qv_\mu)L_\mu,\end{aligned} \tag{4.10}$$

where $v \in \mathcal{V}$ is expanded exactly as in (4.4). This means that $\mathcal{V} \setminus \mathcal{V}(\Lambda) \subseteq \ker \mathcal{Q}$, i.e. for all $\mu \in \mathcal{F} \setminus \Lambda$, the semi-discrete functions v_μ in (4.4) are interpolated by zero. As a direct consequence, we have similar estimates for the tensor product interpolation operator, if $v \in \mathcal{V}(\Lambda)$:

$$\|a_0^{1/2}(v - \mathcal{Q}v)\|_{L^2_\pi(\Gamma,L^2(T))} \le c_{\mathcal{T}} h_T |v|_{\mathcal{V},T} \quad \forall T \in \mathcal{T} \tag{4.11}$$

$$\|a_0^{1/2}(v - \mathcal{Q}v)\|_{L^2_\pi(\Gamma,L^2(S))} \le c_{\mathcal{S}} h_S^{1/2} |v|_{\mathcal{V},S} \quad \forall S \in \mathcal{S}, \tag{4.12}$$

with

$$|v|^2_{\mathcal{V},T} := \int_\Gamma |v(y)|^2_{\mathcal{X},T}\,\mathrm{d}\pi(y), \quad |v|^2_{\mathcal{V},S} := \int_\Gamma |v(y)|^2_{\mathcal{X},S}\,\mathrm{d}\pi(y).$$

Because of the expansion of the affine diffusion coefficient a, the Galerkin operator in (4.7) can be further simplified:

$$\begin{aligned}\mathbf{A}(i,\mu;j,\mu') &= \int_\Gamma \int_D a(x,y)\nabla\varphi_i(x)L_\mu(y)\cdot\nabla\varphi_j(x)L_{\mu'}(y)\,\mathrm{d}x\mathrm{d}\pi(y)\\ &= \sum_{m=1}^{M} \int_D a_m(x)\nabla\varphi_i(x)\cdot\nabla\varphi_j(x)\,\mathrm{d}x \prod_{\substack{m'=1\\ m'\neq m}}^{M} \int_{-1}^{1} L_{\mu_{m'}}(y_{m'})L_{\mu'_{m'}}(y_{m'})\,\mathrm{d}\pi_{m'}(y_{m'})\\ &\quad\times \int_{-1}^{1} y_m L_{\mu_m}(y_m)L_{\mu'_m}(y_m)\,\mathrm{d}\pi_m(y_m),\end{aligned}$$

and therefore

$$\mathbf{A} = \sum_{m=1}^{M} \mathbf{A}_m, \tag{4.13}$$

$$\mathbf{A}_m := \sum_{m=1}^{M} A_m \otimes I_{q_1} \otimes \cdots \otimes I_{q_{m-1}} \otimes K_m \otimes I_{q_{m+1}} \otimes \cdots \otimes I_{q_M},$$

with a finite M and where

$$\begin{aligned}A_m(i,j) &:= \int_D a_m(x)\nabla\varphi_i(x)\cdot\nabla\varphi_j(x)\,\mathrm{d}x,\\ K_m(\mu_m,\mu'_m) &:= \int_{-1}^{1} y_m L_{\mu_m}(y_m)L_{\mu'_m}(y_m)\,\mathrm{d}\pi_m(y_m).\end{aligned}$$

The matrices $A_m \in \mathbb{R}^{N^2}$ are sparse due to the small overlap of the polynomials in the FEM discretisation and because of orthonormality of the Legendre basis, most other matrices are the identity. Because of (4.3) the matrices $K_m \in \mathbb{R}^{q_m^2}$ form a *triple-product*

$$K_m(\mu_m, \mu'_m) = \int_D \left(\beta_{\mu_m+1} L_{\mu_m+1}(y_m) + \beta_{\mu_m} L_{\mu_m-1}(y_m)\right) L_{\mu'_m}(y_m)\, \mathrm{d}\pi_m(y_m) \tag{4.14}$$
$$= \begin{cases} \beta_{\mu_m+1}, & \mu_m + 1 = \mu'_m, \\ \beta_{\mu_m}, & \mu_m - 1 = \mu'_m, \\ 0, & \text{otherwise}, \end{cases}$$

that is non-zero only for $|\mu_m - \mu'_m| = 1$. Therefore, only the entries next to the diagonal are non-zero. The right hand side $F \in \mathbb{R}^{N \times q_1 \times \cdots \times q_M}$ is given as in (4.9).

This yields the discrete Galerkin problem (2.5). Here we will write $U = \mathbf{u}, F = \mathbf{f}$, then this yields

$$\mathbf{A}U = F$$

as this is an operator equation for the unknown tensor $U \in \mathbb{R}^{N \times q_1 \times \cdots \times q_M}$. Céa's Lemma applies as in (2.7)

$$\|u - u_{\mathrm{disc}}\|_{\mathcal{V}} \leq \sqrt{\frac{c_2}{c_1}} \inf_{v_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}} \|u - v_{\mathrm{disc}}\|_{\mathcal{V}}.$$

with the constants from the uniform boundedness (3.6) and coercivity (3.7) for the discrete solution

$$u_{\mathrm{disc}} = \sum_{i=0}^{N-1} \sum_{\mu \in \Lambda} U(i, \mu) \varphi_i L_\mu$$

of (4.8). Numerical methods on how to approximate and compute the Galerkin solution u_{disc} via the tensor U will be investigated after we discuss the discretisation of the log-normal problem.

The Log-Normal Problem

For the log-normal diffusion equation we had restricted the parameter space Γ as in (3.18) and with different Gaussian product measures. In Proposition 3.9, we have established that the solution u is in the space $L^2_\gamma(\Gamma, \mathcal{X})$. Since $\Gamma \subset \mathbb{R}^\infty$, it makes sense to define the *Hermite polynomials* $(H_{\mu_m})_{\mu_m \in \mathbb{N}}$ that form the natural orthonormal basis on the space $L^2_\gamma(\mathbb{R})$ and restrict to parameters on Γ. Thus, for $y_m \in \mathbb{R}$, we define via the three-term recursion (4.3)

$$\sqrt{\mu_m + 1} H_{\mu_m+1}(y_m) = y_m H_{\mu_m}(y_m) + \sqrt{\mu_m} H_{\mu_m-1}(y_m) \quad \forall \mu_m \in \mathbb{N}$$

i.e. $\beta_{\mu_m} = \sqrt{\mu_m}$ for all $\mu_m \in \mathbb{N}$. Again, this can be given explicitly as

$$H_{\mu_m}(y_m) = \frac{(-1)^{\mu_m}}{\sqrt{\mu_m!}} \exp\left(\frac{1}{2} y_m^2\right) \frac{\partial^{\mu_m}}{\partial y_m^{\mu_m}} \exp\left(-\frac{1}{2} y_m^2\right).$$

As before, we define the *Hermite chaos polynomials*

$$H_\mu = \bigotimes_{m \in \mathbb{N}} H_{\mu_m}$$

for all finitely supported $\mu \in \mathcal{F}$ and we present a proposition from [75].

Proposition 4.2. *The tensor product Hermite polynomials $(H_\mu)_{\mu \in \mathcal{F}}$ form an orthonormal basis of $L^2_\gamma(\mathbb{R})$ and by restriction also of $L^2_\gamma(\Gamma)$.*

Remark 4.3. In most publications, the Hermite polynomials are not normalised and orthogonality is defined for two different measures. We would like to indicate that in this thesis, the Hermite polynomial are normalised and orthogonal with respect to the Gaussian measure γ defined above.

Obviously, the product of two polynomials of degree μ_m and ν_m has degree $\mu_m + \nu_m$ and therefore its expansion is also finite. This can be explicitly calculated [80]:

Proposition 4.4. *For two Hermite polynomials H_{μ_m}, H_{ν_m}, their product has the finite hermite expansion*

$$H_{\mu_m} H_{\nu_m} = \sum_{\eta_m=0}^{\min(\mu_m,\nu_m)} \frac{\sqrt{(\mu_m+\nu_m-2\eta_m)!\mu_m!\nu_m!}}{\eta_m!(\mu_m-\eta_m)!(\nu_m-\eta_m)!} H_{\mu_m+\nu_m-2\eta_m}.$$

Moreover, the integral over three Hermite polynomials $H_{\mu_m}, H_{\nu_m}, H_{\eta_m}$ takes the value

$$\int_{-\infty}^{\infty} H_{\mu_m}(y_m) H_{\nu_m}(y_m) H_{\eta_m}(y_m)\,\mathrm{d}\gamma_m(y_m) = \kappa_{\mu_m,\nu_m,\eta_m}$$

with

$$\kappa_{\mu_m,\nu_m,\eta_m} = \begin{cases} \frac{\sqrt{\eta_m!\mu_m!\nu_m!}}{\frac{\mu_m+\nu_m-\eta_m}{2}!\frac{\mu_m-\nu_m+\eta_m}{2}!\frac{-\mu_m+\nu_m+\eta_m}{2}!}, & \mu_m+\nu_m-\eta_m \text{ is even} \\ & \text{and } |\mu_m-\nu_m| \le \eta_m \le \mu_m+\nu_m, \\ 0, & \text{otherwise.} \end{cases} \quad (4.15)$$

In Section 3.1.3, we have shown that for $0 < \vartheta < 1$ and $\rho > 0$ it holds

$$L^2_{\gamma_\rho}(\Gamma,\mathcal{X}) \subset \mathcal{V}_{\vartheta\rho} \subset L^2_{\gamma}(\Gamma,\mathcal{X}),$$

and since the solution u of the variational problem (3.24) is in $\mathcal{V}_{\vartheta\rho}$ we need to discretise the smaller space $\mathcal{V}_{\vartheta\rho}$. This space is not accessible and so we have to ensure $u \in L^2_{\gamma_\rho}(\Gamma,\mathcal{X})$:

Theorem 4.5. *Let $2 < q < \infty$ and $\rho > 0$. If $f \in L^q_{\gamma_\rho}(\Gamma,\mathcal{X})$, then the solution u of (3.5) is in $L^2_{\gamma_\rho}(\Gamma,\mathcal{X})$ and it holds*

$$\|u\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})} \le c_{\rho,q}\|f\|_{L^q_{\gamma_\rho}(\Gamma,\mathcal{X})}.$$

The arguments in the proof are very similar to the one used for the inequalities in Lemma (3.7) and they can be found in [75]. Since we assumed $f(x,y) \equiv f(x)$ for all $y \in \Gamma$, $f \in L^q_{\gamma_\rho}(\Gamma)$ is always ensured for a $q > 2$.

The discretisation of the space $L^2_{\gamma_\rho}(\Gamma,\mathcal{X})$ is straightforward: We define the function

$$\tau_\rho : \mathbb{R}^\infty \to \mathbb{R}^\infty, \quad \tau_\rho = \bigotimes_{m=1}^{\infty} \tau_{m,\rho},$$
$$\tau_{m,\rho} : \mathbb{R} \to \mathbb{R}, \quad y_m \mapsto \exp(-\rho\alpha_m) y_m \quad \forall m \in \mathbb{N}.$$

This is a bijection and it also maps Γ bijectively onto itself. For all $\mu \in \mathcal{F}$, we define the generalised Hermite chaos polynomials

$$H_{\rho,\mu}(y) = \bigotimes_{m\in\mathbb{N}} H_{\rho,\mu_m}(y_m),$$
$$H_{\rho,\mu_m}(y_m) := H_{\mu_m}(\tau_\rho(y_m)) \quad \forall m \in \mathbb{N}.$$

This is again a basis of the space $L^2_{\gamma_\rho}(\Gamma)$ [75]:

Lemma 4.6. *The map*

$$L^2_{\gamma}(\Gamma) \to L^2_{\gamma_\rho}(\Gamma), \quad v_y \mapsto v_y \circ \tau_\rho$$

is a unitary isomorphism of Hilbert spaces and it holds

$$\int_\Gamma v_y(y)\,\mathrm{d}\gamma(y) = \int_\Gamma v_y(\tau_\rho(y))\,\mathrm{d}\gamma_\rho(y) \quad \forall v_y \in L^2_\gamma(\Gamma).$$

In particular, the shifted Hermite chaos basis $(H_{\rho,\mu})_{\mu\in\mathcal{F}}$ is an orthonormal basis of $L^2_{\gamma_\rho}(\Gamma)$.

A trivial consequence of this Lemma is that Proposition 4.4 also holds for the shifted Hermite polynomials $H_{\rho,\mu}$ and it turns out that the coefficients (4.15), i.e. the results of the triple-product integral, are the same for any $\rho \in \mathbb{R}$.

As a result of the above, the solution $u \in L^2_{\gamma_\rho}(\Gamma, \mathcal{X})$ is of the form

$$u(x,y) = \sum_{\mu \in \mathcal{F}} u_\mu(x) H_{\rho,\mu}(y)$$

as in (4.4). The full discretisation is done by choosing a FEM space $\mathcal{X}_p(\mathcal{T}) \subset \mathcal{X}$ and a tensor index set $\Lambda \subset \mathcal{F}$. Then we define the Galerkin space

$$\mathcal{V}_{\text{disc}} := \mathcal{V}(p, \mathcal{T}, \Lambda) = \{v_{\text{disc}} \in \mathcal{V}_{\vartheta\rho} : v_{\text{disc}} = \sum_{i=0}^{N-1} \sum_{\mu \in \Lambda} V(i,\mu) \varphi_i H_{\rho,\mu}\}$$

and it holds $\mathcal{Y}(\Lambda) = \text{span}_{\mu\in\Lambda}(H_{\rho,\mu})$. Since this space is finite dimensional and represented by polynomials of finite degree, it is possible to express every function $v_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ also with respect to a Hermite chaos basis of the same degree that is orthonormal with respect to the auxiliary measure $\gamma_{\vartheta\rho}$. This is gathered in the following result:

Proposition 4.7. *The generalised Hermite chaos polynomials for the Gaussian measure γ_ϑ can be represented in another generalised Hermite chaos basis for the measure γ_ρ, $\vartheta \neq \rho$, using the formula*

$$H_{\vartheta,\mu_m}(y_m) = \sum_{\nu_m=0}^{\lfloor \mu_m/2 \rfloor} \kappa^{\vartheta\to\rho}_{\mu_m,\nu_m} H_{\rho,\mu_m-2\nu_m}(y_m)$$

for every $m \in \mathbb{N}, \mu_m \in \mathbb{N}$, with

$$\kappa^{\vartheta\to\rho}_{\mu_m,\nu_m} = \frac{\sqrt{\mu_m!}}{\nu_m!\sqrt{\mu_m - 2\nu_m!}} \left(\frac{\tilde\sigma_m}{\sigma_m}\right)^{\mu_m} \left(\frac{\tilde\sigma_m^2 - \sigma_m^2}{2\tilde\sigma_m^2}\right)^{\nu_m},$$

where

$$\sigma_m = \exp(\vartheta\alpha_m) \quad \textit{and} \quad \tilde\sigma_m = \exp(\rho\alpha_m)$$

as before.

Proof. The proof works similarly to the one for Proposition 4.4. The Hermite polynomials are generated by

$$\sum_{\mu_m=0}^{\infty} H_{\mu_m}(y_m) \frac{(\sqrt{2}t)^{\mu_m}}{\sqrt{\mu_m!}} = \exp(\sqrt{2}ty_m - t^2),$$

see for example [80]. Then for the Hermite polynomials, using $H_{\vartheta,\mu_m}(y_m) = H_{\mu_m}(\frac{1}{\sigma_m} y_m)$ and substituting $s := \frac{\sqrt{2}t}{\sigma_m}$, we get

$$\begin{aligned}
\sum_{\mu_m=0}^{\infty} H_{\vartheta,\mu_m}(y_m) \frac{(\sigma_m s)^{\mu_m}}{\sqrt{\mu_m!}} &= \exp\left(sy_m - \frac{1}{2}\sigma_m^2 s^2\right) \\
&= \exp\left(sy_m - \frac{1}{2}\tilde\sigma_m^2 s^2 + \frac{1}{2}\tilde\sigma_m^2 s^2 - \frac{1}{2}\sigma_m^2 s^2\right) \\
&= \left(\sum_{\eta_m=0}^{\infty} H_{\rho,\eta_m}(y_m) \frac{(\tilde\sigma_m s)^{\eta_m}}{\sqrt{\eta_m!}}\right) \exp\left(\frac{1}{2}(\tilde\sigma_m^2 - \sigma_m^2)s^2\right) \\
&= \sum_{\eta_m=0}^{\infty} \sum_{\nu_m=0}^{\infty} H_{\rho,\eta_m}(y_m) \frac{\tilde\sigma_m^{\eta_m} (\tilde\sigma_m^2 - \sigma_m^2)^{\nu_m}}{2^{\nu_m}\nu_m!\sqrt{\eta_m!}} s^{\eta_m + 2\nu_m}.
\end{aligned}$$

Then, matching coefficients such that $\mu_m = \eta_m + 2\nu_m$, it follows that

$$H_{\vartheta,\mu_m}(y_m) \frac{(\sigma_m s)^{\mu_m}}{\sqrt{\mu_m!}} = \sum_{\eta_m=0}^{\infty} \sum_{\nu_m=0}^{\infty} H_{\rho,\eta_m}(y_m) \frac{\tilde\sigma_m^{\eta_m} (\tilde\sigma_m^2 - \sigma_m^2)^{\nu_m}}{2^{\nu_m}\nu_m!\sqrt{\eta_m!}}$$

and setting $\eta_m = \mu_m - 2\nu_m$ yields the desired result. □

This means in particular that it also holds

$$\mathcal{Y}(\Lambda) := \operatorname*{span}_{\mu\in\Lambda}(H_{\vartheta\rho,\mu}),$$

i.e. we can represent $v_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ alternatively as

$$v_{\text{disc}} = \sum_{i=0}^{N-1}\sum_{\mu\in\Lambda} V(i,\mu)\varphi_i H_{\vartheta\rho,\mu}$$

with some other coefficients $V(i,\mu)$. This will be used in the following, because it simplifies calculations. Since $\gamma_{\vartheta\rho}$ is the measure employed in the scalar product $B_{\vartheta\rho}(\cdot,\cdot)$, the Hermite polynomials $H_{\vartheta\rho,\mu}$ are orthonormal under this measure. When endowed with this scalar product, $\mathcal{V}_{\text{disc}}$ is again a Hilbert space as it is a finite dimensional subset of $\mathcal{V}_{\vartheta\rho}$.

As before, we obtain the Clément interpolation operator $Q : \mathcal{X} \to \mathcal{X}_p(\mathcal{T})$ as in Chapter 4.1.1 without any extra weights. We define the tensor product interpolation operator in the log-normal case as

$$\begin{aligned} \mathcal{Q} &: \mathcal{V} \to \mathcal{V}_{\text{disc}} \\ \mathcal{Q}v &= \sum_{\mu\in\Lambda}(Qv_\mu)H_\mu \end{aligned} \tag{4.16}$$

and for $v \in \mathcal{V}(\Lambda)$, we again have the estimates

$$\begin{aligned} \|v - \mathcal{Q}v\|_{L^2_\gamma(\Gamma,L^2(T))} &\le c_{\mathcal{T}} h_T |v|_{\mathcal{V},T} \quad \forall T \in \mathcal{T}, \\ \|v - \mathcal{Q}v\|_{L^2_\gamma(\Gamma,L^2(S))} &\le c_{\mathcal{S}} h_S^{1/2} |v|_{\mathcal{V},S} \quad \forall S \in \mathcal{S}. \end{aligned}$$

The discrete Galerkin problem is

$$\mathbf{A}U = F \tag{4.17}$$

acting on the tensor $U \in \mathbb{R}^{N\times q_1\times\cdots\times q_M}$. Céa's Lemma is a bit different here, as it can only be given with respect to the norm in $L^2_{\gamma_\rho}(\Gamma,\mathcal{X})$:

Theorem 4.8. *For $p > 2$ let $f \in L^p_{\gamma_\rho}(\Gamma,\mathcal{X})$. Then the Galerkin projection $u_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ satisfies*

$$\|u - u_{\text{disc}}\|_{L^2_\gamma(\Gamma,\mathcal{X})} \le \sqrt{\frac{\hat{c}_{\vartheta,\rho}}{\check{c}_{\vartheta,\rho}}} \inf_{v_{\text{disc}}\in\mathcal{V}_{\text{disc}}} \|u - v_{\text{disc}}\|_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})},$$

where $u \in \mathcal{V}$ is the exact solution.

Proof. Theorem 4.5 implies $u \in L^2_{\gamma_\rho}(\Gamma,\mathcal{X})$. By Galerkin orthogonality with respect to $B_{\vartheta\rho}(\cdot,\cdot)$ and using Lemma 3.10, we get

$$\begin{aligned} \check{c}_{\vartheta,\rho}\|u - u_{\text{disc}}\|^2_{L^2_\gamma(\Gamma,\mathcal{X})} &\le B_{\vartheta\rho}(u - u_{\text{disc}}, u - u_{\text{disc}}) \\ &= \inf_{v_{\text{disc}}\in\mathcal{V}_{\text{disc}}} B_{\vartheta\rho}(u - v_{\text{disc}}, u - v_{\text{disc}}) \\ &\le \hat{c}_{\vartheta,\rho} \inf_{v_{\text{disc}}\in\mathcal{V}_{\text{disc}}} \|u - v_{\text{disc}}\|^2_{L^2_{\gamma_\rho}(\Gamma,\mathcal{X})}. \end{aligned}$$

□

Since the errors are measured in different norms, this is only *almost* quasi-optimal. The constant depends on the choice of ρ and ϑ and it tends to ∞ as ρ approaches 0 or ∞ or if ϑ approaches 0. This is consistent with the bound given in Theorem 3.6.

As opposed to the affine case, the operator $\mathbf{A}$ of the log-normal equation does not have a natural tensor structure because only the logarithm of the diffusion coefficient has a series expansion. It is however possible to approximate the coefficient on a tensor product space and thus obtain a similar representation of the Galerkin operator $\mathbf{A}$. This will be done in Section 6 after we have introduced several tensor decomposition formats and tensor product approximation techniques.

5 Tensor Decomposition Formats

We generalise (4.6) by considering tensor products of Hilbert spaces of order M

$$\mathcal{V} = \bigotimes_{m=1}^{M} \mathcal{V}_m$$

with an orthonormal basis set $\mathcal{B} = \{\phi_{\mu_1}^{(1)} \otimes \cdots \otimes \phi_{\mu_M}^{(M)} : \mu_m \in \mathbb{N}, m = 1, \ldots, M\}$. Any tensor $v \in \mathcal{V}$ has the basis expansion

$$v = \sum_{\mu_1,\ldots,\mu_M=1}^{\infty} V(\mu_1, \ldots, \mu_M)\phi_{\mu_1}^{(1)} \otimes \cdots \otimes \phi_{\mu_M}^{(M)}. \tag{5.1}$$

For the remainder of this chapter, we will work only with finite dimensional Hilbert spaces unless stated otherwise, i.e. we have some finite approximation

$$v \approx v_{\text{disc}} = \sum_{\mu_1=1}^{q_1} \cdots \sum_{\mu_M=1}^{q_M} V(\mu_1, \ldots, \mu_M)\phi_{\mu_1}^{(1)} \otimes \cdots \otimes \phi_{\mu_M}^{(M)}. \tag{5.2}$$

The tensor $v_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ is equivalent to the multi-indexed array

$$\begin{aligned} &V \in \mathbb{R}^{q_1 \times \ldots \times q_M}, \\ &V(\mu_1, \ldots, \mu_M), \quad \mu_m \in \{1, \ldots, q_m\}, \quad m = 1, \ldots, M. \end{aligned}$$

Remark 5.1. Since we are dealing with arrays, we begin counting at 1 in this chapter. This is opposed to the previous chapter, in which a tensor serves as a coefficent for a basis expansion, but the two cases are clearly equivalent.

We define the inner product for tensors $V, W \in \mathbb{R}^{q_1 \times \ldots \times q_M}$ as the Frobenius inner product

$$\langle V, W \rangle = (V, W)_{\ell^2(\mathbb{R}^{q_1 \times \ldots \times q_M})}$$

and treat tensors and their duals equally. In the following, as long as it is not confusing, we simplify notation and leave out the subscript to denote Frobenius norm by

$$\|V\| = \|V\|_{\ell^2(\mathbb{R}^{q_1 \times \ldots \times q_M})}$$

for tensors $V \in \mathbb{R}^{q_1 \times \ldots \times q_M}$.

Computation with tensors suffer from the *curse of dimensionality* [31], since the storage of the complete array grows exponentially with the order M. We seek to reduce computational costs by parametrising the tensors in some data-sparse representation.

5.1 The Canonical Tensor Decomposition

We adhere to the *separation of variables*, a classical approach which traces back to Bernoulli and Fourier among others. In principle, we want to represent or approximate tensors as multivariate functions by a sum of products of univariate functions. This concept is well established for tensors of order $d = 2$ where it leads to fundamental results known as the singular value decomposition (SVD) or Schmidt decomposition, proper orthogonal decomposition, the Karhunen-Loève transform and so on, see Chapters 2 and 3. In the discrete case discussed here, i.e. in matrix theory, this is known as *low rank approximation*. However, the generalisation of the concept of ranks to higher order tensors is not as straightforward as one may expect [30]. There are many possible and a priori equally justifiable tensor decompositions that all yield different definitions of a tensor rank.

The *canonical tensor representation* separates the variables as a sum of elementary tensors:

$$V(\mu_1, \dots, \mu_M) = \sum_{k=1}^{R} v_k^{(1)}(\mu_1) \cdots v_k^{(M)}(\mu_M).$$

The canonical tensor rank of V is the smallest R such that this representation is exact. This is then called the *canonical decomposition* of the tensor [30].

However, while this is a beautiful and rather simplistic tensor representation, it has several severe drawbacks. First of all, finding the canonical rank and thus also its decomposition is NP-hard [43]. Additionally, the set of tensors with rank smaller or equal R is not closed, i.e. it is possible to find a sequence of rank-R tensors that converges to a tensor with rank greater than R, see the *border rank problem* [46, 30]. While the former property obviously poses problems in computation, the latter can be very undesirable as well when it comes to optimisation algorithms. Altogether, the canonical format has not only led to deep and difficult mathematical problems [47, 46], but also computational experience has often been disappointing, by observing slow convergence, low accuracy and the like. It is not clear how to circumvent these problems while still retaining its outstanding complexity scaling. In recent years, the canonical format has therefore been put into question, albeit not completely disqualified, and we are looking for alternatives with favourable properties.

5.2 Tensor Representations

We parametrise a tensor in a very general form to define a *tensor representation* via

$$V(\mu_1, \dots, \mu_M) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_J=1}^{r_J} \prod_{\ell=1}^{L} V_\ell(\underline{\mu_\ell}; \underline{k_\ell}), \tag{5.3}$$

where $\underline{\mu_\ell} \subseteq \{\mu_1, \dots, \mu_M\}$ and $\underline{k_\ell} \subseteq \{k_1, \dots, k_J\}$ and

$$\begin{aligned} k &= \bigcup_{\ell=1}^{L} \underline{k_\ell}, \\ k_j &= 1, \dots, r_j \quad \forall j = 1, \dots, J. \end{aligned}$$

Since the ordering of the indices is irrelevant in this context, we maintain the slight abuse of notation and interpret multi-indices as sets of natural numbers.

The tensor representation (5.3) is parametrised by L *component tensors* $V_1, \dots, V_L$. The μ_m are called *physical* indices and the k_j are called *virtual*. A component V_ℓ is called virtual, if it does not have any physical indices, i.e. $\underline{\mu_\ell} = \emptyset$. Otherwise it is called physical. Summation over a common virtual index k_j is called the *contraction* over k_j.

We can demand a number of further properties that allow for simpler treatment of the tensor [78]. First of all, since a tensor is multi-linear, it is conventional to only deal with multi-linear representations:

Criterion 5.2. For each $m \in \{1, \dots, M\}$ there exists *exactly* one $\ell \in \{1, \dots, L\}$ such that $\mu_m \in \underline{\mu_\ell}$.

This means that no two components can depend on the same physical index.

It is our central aim to reduce complexity of the tensor and we therefore need to choose the representation carefully. Firstly, the number of components should not be exceedingly high, as this makes the representation more complicated. But more importantly, we try to minimise the dimensions r of the multi-index k over all possible representations (5.3). If these dimensions r_j are minimal for the given parametrisation, the tuple r is called the *rank*, or better *multi-linear rank* of the representation and the representation is called a *decomposition*. However, as mentioned for the canonical format above, this notion of rank leads to extreme difficulties even for the simplest forms.

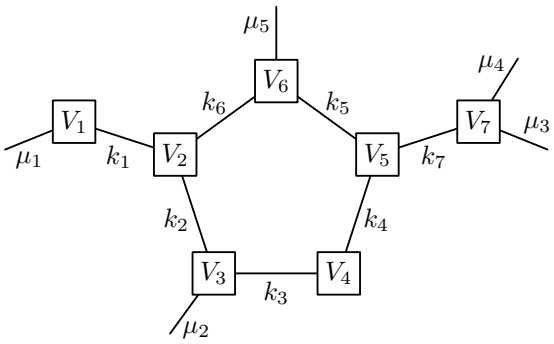

Figure 1: A general tensor network representation of a tensor of order 5.

5.3 Tensor Networks

For a proper definition of multi-linear ranks, we consider subclasses of tensor representations and introduce a further restriction that each $k_j \in k$ appears exactly twice:

Criterion 5.3. For each virtual index $k_j \in k$ there exist *exactly* two component tensors V_{ℓ_1}, V_{ℓ_2} with k_j as an index, i.e.

$$k_j \in \underline{k_{\ell_1}}, k_j \in \underline{k_{\ell_2}},$$
$$k_j \notin \underline{k_\ell}, \quad \forall \ell \neq \ell_1, \ell_2.$$

Any parametrisation satisfying Criterion 5.2 and 5.3 can be expressed as a simple undirected weighted graph with half-edges, see Figure 1 for an example. We obtain what is called a *tensor network* or a *tensor network state* in quantum physics. The component tensors give the vertices of the graph, the contractions are represented by the edges between the vertices and the physical indices yield half edges. Therefore, we get a graph $TNS(V) := (\mathtt{V}, \mathtt{E}, \mathtt{H})$, that we denote in typewriter font

$$\mathtt{V} = \{V_\ell : \ell = 1, \dots, L\}, \quad \mathtt{E} = k = \{k_1, \dots, k_J\}, \quad \mathtt{H} = \{\mu_1, \dots, \mu_M\}.$$

Because of Criterion 5.2, each half-edge has exactly one incident vertex, and because of 5.3, each edge has exactly two incident vertices. Thus, this is well-defined. The weight of the half-edge μ_m is given by its dimension q_m and the weight of the edge k_j is given by its dimension r_j in (5.3). In accordance with the tensor decompositions, we call the vector r the *rank* of the tensor network if it is point-wise minimal. The number $q_1 \cdots q_M$ is naturally the dimension of the tensor network.

Since a contraction over an index with dimension 1 is trivial, we can choose to either omit this index or even to introduce extra indices. In general, we require the tensor network graph to be connected and if it is not we invent an arbitrary index of dimension 1 to make it so. Apart from that, any index of dimension 1 that is not necessary for connectedness will usually be omitted.

Although heavily used in physics, this general concept still suffers from some instabilities. Recently, it has been shown that tensor networks which contain closed loops are not necessarily Zariski closed [47], i.e. tensors represented by them do not form algebraic varieties without further restrictions. This is closely related to the border rank problem for the canonical format. While we will not go into these details here, we highlight that all these difficulties can be avoided if we restrict ourselves to tensors fulfilling the following criterion [47]:

Criterion 5.4. The tensor network $TNS(V)$ is cycle-free.

Since we have the trivial connectedness mentioned above, any tensor network that fulfils Criterion 5.4 is a tree. It is thus called a *Tree Tensor Network* or, in accordance with nomenclature from Quantum Physics, a *Tree Tensor Network State (TTNS)*. See Figure 2 for an example.

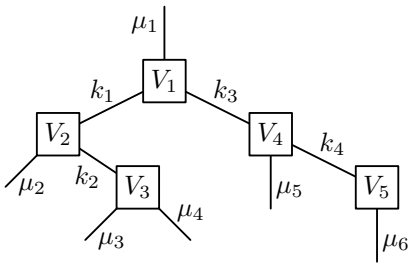

Figure 2: An arbitrary example of a tensor tree.

While general tensor network representations, like the canonical format, might still be very useful and shall not be disqualified, we presently only consider the special case of non-circular graphs that prevents these fundamental difficulties.

5.4 Subspace Formats

5.4.1 Reduced Basis Functions

We consider again the tensor product space $\mathcal{V} = \bigotimes_{m=1}^{M} \mathcal{V}_m$ with tensor product basis

$$\mathcal{B} = \{\phi_{\mu_1}^{(1)} \otimes \cdots \otimes \phi_{\mu_M}^{(M)} : \mu_m \in \mathbb{N}, m = 1, \ldots, M\}.$$

The following results also hold for infinite dimensional spaces $\mathcal{V}_m$. For each m, the sets $\mathcal{B}_m := \{\phi_{\mu_m}^{(m)} : \mu_m \in \mathbb{N}\}$ form an orthonormal basis of $\mathcal{V}_m$.

In Chapter 4.1, we have chosen finite dimensional Galerkin spaces in order to approximate a function $v \in \mathcal{V}$, i.e. we discretised using a fixed basis set that we obtained heuristically or with some a priori knowledge of the Hilbert space. Another approach would be to find better or even the best basis sets by minimising

$$v_\epsilon = \operatorname*{argmin}_{w_\epsilon \in \mathcal{W}} \|v - w_\epsilon\|, \tag{5.4}$$

with $\mathcal{W} = \bigotimes_{m=1}^{M} \mathcal{W}_m$, $\dim \mathcal{W}_m \leq r_m$. We then optimise over the tensor product of all univariate subspaces $\mathcal{W}_m \subset \mathcal{V}_m$ of dimension at most r_m [30, 31], which is a well-posed problem, see [21]. This is called *reduced basis optimisation*, as for each $m = 1, \ldots, M$, we aim at finding an optimal basis set $\{\psi_{k_m}^{(m)} : k_m = 1, \ldots, r_m\}$ of a finite-dimensional subspace $\mathcal{W}_m \subset \mathcal{V}_m$. Conceptually, this approach has to be distinguished from the Galerkin approximation.

We get a different approximation similar to (5.2) that reads

$$v \approx v_\epsilon = \sum_{k_1=1}^{r_1} \cdots \sum_{k_M=1}^{r_M} C(k_1, \ldots, k_M) \psi_{k_1}^{(1)} \otimes \cdots \otimes \psi_{k_M}^{(M)}.$$

Due to the optimal choice of basis, this approximation will be better than the Galerkin approximation if we allow the same number of basis functions, i.e. $r_m = q_m$. However, we have not stated a way of obtaining the reduced basis functions $\psi_{k_m}^{(m)}$ and this is highly non-trivial. Therefore, we use the Galerkin basis introduced above and perform a reduced basis method on the finite dimensional coefficient tensor V. This leads to the widely used *Tucker decomposition format* of a tensor, the TT format, and other so called *subspace formats*.

5.4.2 The Tucker Format

In our discrete, finite-dimensional setting, the tensor space and the tensor product space are equivalent, that is

$$\mathbb{R}^{q_1 \times \ldots \times q_M} \cong \bigotimes_{m=1}^{M} \mathbb{R}^{q_m} \cong \mathcal{V}_{\text{disc}}$$

via the trivial formula

$$V(\mu_1, \dots, \mu_M) = \sum_{\mu_1=1}^{q_1} \cdots \sum_{\mu_M=1}^{q_M} V(\mu_1, \dots, \mu_M) e_{\mu_1}^{(1)} \otimes \cdots \otimes e_{\mu_M}^{(M)}, \tag{5.5}$$

where $e_{\mu_m}^{(m)}$ denote the standard Euclidean basis vectors of $\mathbb{R}^{q_m}$ for each m.

The reduced basis approximation (5.6) of V then becomes

$$V_\epsilon = \operatorname{argmin} \left\{ \|V - W_\epsilon\| : W_\epsilon \in \bigotimes_{m=1}^{M} \mathcal{W}_m, \dim \mathcal{W}_m \leq r_m \right\}, \tag{5.6}$$

this time with optimal subspaces $\mathcal{W}_m \subset \mathbb{R}^{q_m}, m = 1, \dots, M$, that are spanned by a reduced basis $\{b_{k_m}^{(m)} : k_m = 1, \dots, r_m\}$, and we get the approximation

$$V \approx V_\epsilon = \sum_{k_1=1}^{r_1} \dots \sum_{k_M=1}^{r_M} C(k_1, \dots, k_M) b_{k_1}^{(1)} \otimes \cdots \otimes b_{k_M}^{(M)}.$$

If we can recover the tensor V exactly, i.e. $\|V - V_\epsilon\| = 0$, we call V_ϵ the *Tucker representation* of V. In accordance with the above, a Tucker representation is called a *Tucker decomposition* if the dimensions r_m are the ranks $r_{\text{Tucker}} := r = (r_1, \dots, r_M)$, i.e. they are the smallest numbers such that the tensor can still be recovered exactly. Here, $C \in \mathbb{R}^{r_1 \times \dots \times r_M}$ is a reduced core tensor, that is hopefully much smaller than the original coefficient tensor $V \in \mathbb{R}^{q_1 \times \dots \times q_M}$, due to the optimal choice of basis.

For exact recovery, obtaining the basis vectors in the discrete setting is relatively straightforward. It can be achieved by applying a singular value decomposition in every mode - thus called Higher Order SVD (HOSVD) - of the tensor: For the m-th mode, we compute the SVD of the *m-mode matricisation* as in Theorem 2.1

$$\begin{aligned} [V]^m_{1,\dots,\cancel{m},\dots,M} &\in \mathbb{R}^{(q_1 \cdots \cancel{q_m} \cdots q_M) \times q_m}, \\ [V]^m_{1,\dots,\cancel{m},\dots,M} &= \sum_{k_m=1}^{r_m} X_m \Sigma_m Y_m^\mathsf{T} \end{aligned} \tag{5.7}$$

and obtain the basis vectors

$$b_{k_m}^{(m)} = Y(\cdot, k_m)$$

which span the optimal subspace of $\mathbb{R}^{q_m}$, see [48, 30, 54].

In many applications, we want to approximate the tensor with lower rank $\tilde{r} \leq r$. According to Theorem 2.5, in the matrix case $d = 2$, this can be done by truncating the above SVD and omitting the basis vectors that belong to the smallest $r - \tilde{r}$ singular values.

Unfortunately, this result cannot be generalised to tensors with $d > 2$. It has been shown in [34] that even finding the best rank one, i.e. $r = (1, 1, \dots, 1)$, can be NP-hard if $d > 2$. Nevertheless, truncating the HOSVD in every mode yields a *quasi-optimal approximation* with respect to the ℓ_2-norm [48]. In many cases, this is satisfactory.

The Tucker format is a subspace decomposition as the tensor is expressed in the basis of a subspace of the tensor space. At the same time, it yields a tensor tree, i.e. its representation fulfils Criterion 5.2, 5.3 and 5.4. The core tensor $C \in \mathbb{R}^{r_1 \times \dots \times r_M}$ is the only virtual component and

$$V_m(\mu_m, k_m) = (e_{\mu_m}^{(m)})^\mathsf{T} b_{k_m}^{(m)}$$

yields the M physical components, see Figure 3.

The HOSVD gives us a constructive algorithm that computes the Tucker decomposition, i.e. a representation of the form (6.3) with minimal rank r, in polynomial time. Additionally, the set of tensors with Tucker rank at most r is known to be Zariski-closed [47]. Therefore, it is closed and we overcome the border rank problem. In terms of storage complexity however, this format is far from being optimal. It still scales exponentially in r_m, i.e. for $R := \max\{r_m\}$ the scaling is in $\mathcal{O}(qMR + R^M)$. Especially for small q_m, where we do not have $r_m \ll q_m$, we cannot hope for much reduction of complexity. In particular, for $q_m = 2$ we do not gain any nontrivial progress.

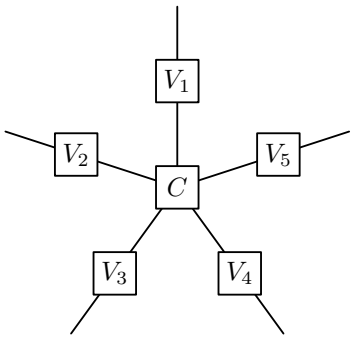

Figure 3: A Tucker tensor of order 5.

5.4.3 Matricisation and Tensor Multiplication

Let us generalise the concept of matricisation in (5.7). Let $t \subseteq \{1, \dots, M\}$ be a collection of physical dimensions and $t^c := \{1, \dots, M\} \setminus t$ its complement. Then

$$[V]_t^{t^c} \in \mathbb{R}^{\underline{q_t}} \otimes \mathbb{R}^{\underline{q_{t^c}}}$$

is the matricisation with the dimensions $\underline{q_t} = \{q_m \in \{q_1, \dots, q_M\} : m \in t\}$ as row dimensions and $\underline{q_{t^c}} = \{q_m \in \{q_1, \dots, q_M\} \setminus q_t\}$ as column dimensions. A special case is the m-th matricisation

$$[V]_{1,\dots,m}^{m+1,\dots,M} \in \mathbb{R}^{q_1 \cdots q_m} \otimes \mathbb{R}^{q_{m+1} \cdots q_M} \tag{5.8}$$

utilised further down that is casting the first m-variables into the row index, and the remaining $M - m$ in the column index. Furthermore, we define the first and last matricisation as *right and left unfolding* respectively

$$V^{\mathsf{R}} := [V]_1^{2,\dots,M}, \tag{5.9}$$

$$V^{\mathsf{L}} := [V]_{1,\dots,M-1}^{M}. \tag{5.10}$$

If we understand the matricisation as a linear operator, it makes sense to investigate it in the context of dual spaces [46]. Thus, we have $[V]_t^{t^c} \in \mathbb{R}^{\underline{q_t}} \otimes \mathbb{R}^{\underline{q_{t^c}}} \cong \mathbb{R}^{\underline{q_t}} \otimes (\mathbb{R}^{\underline{q_{t^c}}})^*$. Hence, a matricisation is not necessarily a matrix, but can be seen as a tensor product of vector spaces and dual spaces. For the m-th matricisation for example, we have

$$\mathbb{R}^{q_1 \cdots q_m} \otimes \mathbb{R}^{q_{m+1} \cdots q_M} \cong \mathbb{R}^{q_1} \otimes \cdots \otimes \mathbb{R}^{q_m} \otimes (\mathbb{R}^{q_{m+1}})^* \otimes \cdots \otimes (\mathbb{R}^{q_M})^*,$$

and with a trivial basis expansion similar to (5.5),

$$[V]_{1,\dots,m}^{m+1,\dots,M} = \sum_{\mu_1=1}^{q_1} \cdots \sum_{\mu_M=1}^{q_M} V(\mu_1, \dots, \mu_M) e_{\mu_1}^{(1)} \otimes \cdots \otimes e_{\mu_m}^{(m)} \otimes (e_{\mu_{m+1}}^{(m+1)})^* \otimes \cdots \otimes (e_{\mu_M}^{(M)})^*.$$

Here, $(e_{\mu_m}^{(m)})^*$ denotes the dual unit vector and it holds $(e_{\mu_m}^{(m)})^*(e_{\nu_m}^{(m)}) = \delta_{\mu_m,\nu_m}$.

This Einstein-like notation allows us to introduce a *tensor multiplication*. Let $V \in \mathbb{R}^{q_1 \times \dots \times q_{M_1}}$ and $W \in \mathbb{R}^{d_1 \times \dots \times d_{M_2}}$. Then if for two matricisations $t_1 \in \{1, \dots, M_1\}, t_2 \in \{1, \dots, M_2\}$ it holds $\underline{q_{t_1^c}} = \underline{d_{t_2}}$, i.e. the dual dimensions of V are the same as the regular dimensions of W, we get

$$[V]_{t_1}^{t_1^c}[W]_{t_2}^{t_2^c} = \sum_{\mu_1=1}^{q_1} \cdots \sum_{\mu_{M_1}=1}^{q_{M_1}} \sum_{\nu_1=1}^{d_1} \cdots \sum_{\nu_{M_2}=1}^{d_{M_2}} \delta_{\underline{\mu_{t_1^c}},\underline{\nu_{t_2}}} V(\underline{\mu_{t_1}}, \underline{\mu_{t_1^c}}) W(\underline{\mu_{t_1^c}}, \underline{\nu_{t_2^c}}) e_{\underline{\mu_{t_1}}}^{(t_1)} \otimes (e_{\underline{\nu_{t_2^c}}}^{(t_2^c)})^*.$$

This is exactly the matrix multiplication of the matricisations and it is the contraction over the common indices $\underline{\mu_{t_1^c}}$. In the case where no dual space is involved, i.e. no contraction is performed, we obtain the tensor product

$$[V]_{1,\dots,M_1}[W]_{1,\dots,M_2} = V \otimes W.$$

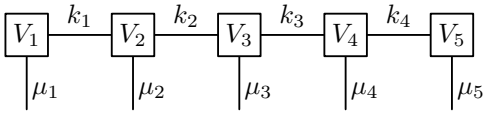

Figure 4: A tensor of order 5 in TT representation.

Remark 5.5. Note that in the case of complex vector spaces, it is inadvisable to understand the matricisation in the context of duality. Instead, it should only be seen as the reordering and grouping of indices here. This is due to the fact that it is impossible to take the complex conjugate only in a few indices of V, which would be required for this concept [77]. Switching the ordering of the indices gives only the transpose and not the hermitian of the original matricisation:

$$[V]_t^{t^c} = \left([V]_{t^c}^t\right)^{\mathsf{T}} = \overline{\left([V]_{t^c}^t\right)^{\mathsf{H}}}.$$

Finally, we want to simplify the notation for the unambiguous case where we multiply over *all* common indices. This will be denoted with a circle, since it can be seen as a composition of two linear operators:

$$V \circ W := [V]_{t_1}^{t_1^c}[W]_{t_2}^{t_2^c}, \quad q_{m_1} \neq d_{m_2} \quad \forall \quad m_1 \in t_1, m_2 \in t_2^c.$$

5.4.4 Matrix Product States or the Tensor Train Format

Another example of a tensor network is the *Tensor Train (TT)* decomposition of a tensor. The tensor V is given element-wise as

$$V(\mu_1, \ldots, \mu_M) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_{M-1}=1}^{r_{M-1}} V_1(\mu_1, k_1) V_2(k_1, \mu_2, k_2) \cdots V_M(k_{M-1}, \mu_M).$$

We get M component tensors of order 2 or 3. Their graph has the structure of a chain or train, hence the name. For the sake of notation, we often set $r_0 = r_M = 1$ and thus

$$r_{\mathrm{TT}} := r = (1, r_1, \ldots, r_{M-1}, 1).$$

Figure (4) illustrates the TT decomposition.

The TT format maintains the positive characteristics of the Tucker format and overcomes most of the disadvantages of the canonical format. However, the complexity now scales only quadratically in the ranks, or with $\mathcal{O}(qMR^2)$, for $R = \max\{r_m\}$. While the Tensor Train decomposition is not the only format that has this advantage, it is one of the most widely used ones and it will also be the standard format in this thesis.

This format has been introduced to the mathematical realm by Oseledets and Tyrtyshnikov [64]. While it was developed independently, it can be seen as a special case of the *Hierarchical Tucker (HT)* decomposition developed by Hackbusch and Kühn [29]. However, we will restrict ourselves to the TT format and deal with the HT format only briefly further down. As stated above, nearly everything of the following can be generalised to a general tensor tree format without much effort, but notation becomes more complex.

In physics, the Tensor Train decomposition has been known as Matrix Product States (MPS) since the late nineties and many results can be taken directly from there. The name Matrix Product States is justified if we fix the physical indices. This yields a chain of matrix products:

$$V(\mu_1, \ldots, \mu_M) = V_1(\mu_1, \cdot) V_2(\cdot, \mu_2, \cdot) \cdots V_{M-1}(\cdot, \mu_{M-1}, \cdot) V_M(\cdot, \mu_M)$$

with $V_m(\cdot, \mu_m, \cdot) \in \mathbb{R}^{r_{m-1} \times r_m}$.

Let it be noted that an important modification of the Tensor Train format follows if we introduce a contraction of rank greater than 1 between the first and last component, also called

periodic boundary conditions,

$$V(\mu_1,\dots,\mu_M) = \sum_{k_1=1}^{r_1} \dots \sum_{k_M=1}^{r_M} V_1(k_M,\mu_1,k_1)V_2(k_1,\mu_2,k_2)\cdots V_M(k_{M-1},\mu_M,k_M).$$

These *uniform Matrix Product States (uMPS)* are especially significant in physics. Verstraete et al. deal with uMPS that are also translation invariant, i.e. all components are equal $V_1 = \ldots = V_d$ [67]. The graph of this decomposition is circular and therefore does not suffice Criterion 5.4. As mentioned above, this poses a number of problems [46] that are - in a nutshell - similar to those of the canonical format. Verstraete developed an *injectivity condition* that aims at overcoming that problem for uniform MPS with periodic boundary conditions [33]. However, we will only deal with non-circular Matrix Product States from now on.

Subspace Notation of the TT Format

The TT format can be considered as a multi-layered subspace representation, as it relies on a successive reduced basis ansatz [30]. This is achieved in a hierarchical way. For the tensor $V \in \mathbb{R}^{q_1\times\dots\times q_M}$, we consider the optimal subspace $\mathcal{W}_1 \subset \mathbb{R}^{q_1}$ given by the basis set $\{b^{(1)}_{k_1} : k_1 = 1,\dots,r_1\}$, where

$$b^{(1)}_{k_1} := \sum_{\mu_1=1}^{q_1} V_1(\mu_1,k_1)e^{(1)}_{\mu_1}.$$

This basis set of the first component is the same as in the Tucker case, and it can be computed using an SVD of the first matricisation (5.8), which is the same as the 1-mode matricisation (5.7). Note that for now, we use *exact* representations, i.e. it holds

$$V \in \mathcal{W} \otimes \mathbb{R}^{q_2\times\dots\times q_M}.$$

We proceed with a subspace of the partial tensor product space $\mathcal{W}_{\{1,2\}} \subset \mathbb{R}^{q_1\times q_2}$ of dimension $r_{\{1,2\}} \le q_1q_2$. Indeed $\mathcal{W}_{\{1,2\}}$ is defined via a new basis set $\{b^{(1,2)}_{k_{\{1,2\}}} : k_{\{1,2\}} = 1,\dots,r_{\{1,2\}}\}$, where the new basis vectors are given in the form

$$b^{(1,2)}_{k_{\{1,2\}}} = \sum_{\mu_1=1}^{q_1}\sum_{\mu_2=1}^{q_2} V_{\{1,2\}}(\mu_1,\mu_2,k_{\{1,2\}})e^{(1)}_{\mu_1}\otimes e^{(2)}_{\mu_2}.$$

We observe that, since the representation was exact, $\mathcal{W}_{\{1,2\}} \subset \mathcal{W}_1 \otimes \mathbb{R}^{q_2}$ with

$$b^{(1,2)}_{k_{\{1,2\}}} = \sum_{k_1=1}^{r_1}\sum_{\mu_2=1}^{q_2} V_2(k_1,\mu_2,k_{\{1,2\}})b^{(1)}_{k_1}\otimes e^{(2)}_{\mu_2}$$

and thus

$$V_{\{1,2\}}(\mu_1,\mu_2,k_{\{1,2\}}) = \sum_{k_1=1}^{r_1} V_1(\mu_1,k_1)V_2(k_1,\mu_2,k_{\{1,2\}}).$$

For this reason, when dealing with TT tensors, we simplify the notation and often set $\{1,2\} \simeq 2$, and in general $\{1,2,\dots,m\} \simeq m$.

The tensor is recursively defined by the component tensors V_m and by the reduced basis functions

$$\begin{aligned} b^{(m)}_{k_m} &= \sum_{k_{m-1}=1}^{r_m}\sum_{\mu_m=1}^{q_m} V_m(k_{m-1},\mu_m,k_m)e^{(m)}_{\mu_m} \\ &= \sum_{k_1=1}^{r_1}\cdots\sum_{k_{m-1}=1}^{r_m}\sum_{\mu_1=1}^{q_1}\cdots\sum_{\mu_m=1}^{q_m} V_1(\mu_1,k_1)\cdots V_m(k_{m-1},\mu_m,k_m)e^{(1)}_{\mu_1}\otimes\cdots\otimes e^{(m)}_{\mu_m} \end{aligned}$$

by taking $\mathcal{W}_{\{1,\dots,m\}} \subset \mathcal{W}_{\{1,\dots,m-1\}} \otimes \mathbb{R}^{q_m}$.

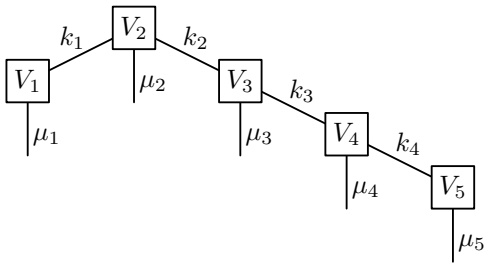

Figure 5: Hierarchical picture of a Tensor Train with V_2 as the root.

We may also proceed differently, e.g. $\mathcal{W}_{\{1,2,3,4,\dots\}} \subset \mathcal{W}_{\{1,2\}} \otimes \mathcal{W}_{\{3,4\}} \otimes \dots \otimes \mathcal{W}_{\{M-1,M\}}$. Especially, it can be advantageous to start from the right hand side, i.e. taking $\mathbb{R}^{q_m} \otimes \mathcal{W}_{\{m+1,\dots,M\}}$ etc., obtaining basis vectors

$$\bar{b}^{(m)}_{k_{m-1}} \in \mathcal{W}_{\{m,\dots M\}}.$$

Let us fix some $m \in \{1, \dots, M\}$ and call it the *root*. This gives a hierarchical picture (see Figure 5).

We consider the spaces $\mathcal{L}_m := \mathcal{W}_{\{1,\dots,m-1\}}$ and $\mathcal{R}_m := \mathcal{W}_{\{m+1,\dots,d\}}$. Their dimensions are given by

$$\dim \mathcal{L}_m = r_{m-1}, \dim \mathcal{R}_m = r_m$$

and hence, the full tensor V is contained in the $r_{m-1}q_m r_m$-dimensional subspace [31, 27]

$$\begin{aligned} &V \in \mathcal{L}_m \otimes \mathbb{R}^{q_m} \otimes \mathcal{R}_m, \qquad (5.11)\\ &V = \sum_{k_{m-1}=1}^{r_{m-1}} \sum_{k_m=1}^{r_m} \sum_{\mu_m=1}^{q_m} V_m(k_{m-1}, \mu_m, k_m) b^{(m-1)}_{k_{m-1}} \otimes e^{(m)}_{\mu_m} \otimes \bar{b}^{(m+1)}_{k_m}. \end{aligned}$$

A canonical but not necessary choice is that the left basis vectors $b^{(m-1)}_1, \dots, b^{(m-1)}_{r_{m-1}}$ and the right basis vectors $\bar{b}^{(m+1)}_1, \dots, \bar{b}^{(m+1)}_{r_m}$ are orthogonal and normalised. This is equivalent to requiring

$$\begin{aligned} (V^{\mathsf{L}}_\ell)^{\mathsf{T}} V^{\mathsf{L}}_\ell &= I_{r_{m-1}}, \quad \ell = 1, \dots, m-1,\\ V^{\mathsf{R}}_\ell (V^{\mathsf{R}}_\ell)^{\mathsf{T}} &= I_{r_m}, \quad \ell = m+1, \dots, M, \end{aligned}$$

i.e. that the first $m-1$ components are left orthogonal and the last components are right orthogonal. We will see in the following that this hierarchical or multi-layered subspace approximation constitutes the mechanism behind the *renormalisation group* formalism in the one-site DMRG (density matrix renormalisation group).

Orthogonality and Quasi-Optimal Approximations

An obvious observation following from the above will be that the minimal dimension r_m is the rank of the m-th matricisation (5.8), see [36] for a formal proof:

Theorem 5.6 (Separation Theorem)**.** *For any tensor $V \in \mathbb{R}^{q_1 \times \dots \times q_M}$, there exists a minimal TT (MPS) representation, thus called TT decomposition $TT(U)$, such that for $m = 1, \dots, M-1$ the dimensions r_m of the contractions $k_m = 1, \dots, r_m$ are minimal and given by*

$$r_m = \operatorname{rank}([V]^{m+1,\dots,M}_{1,\dots,m}).$$

We can change the hierarchy, e.g., by choosing the next component V_{m+1} as the root. In most applications, it will then become necessary to shift the orthogonalisation such that $\{b^{(m)}_{k_m} : k_m = 1, \dots, r_m\}$ and $\{\bar{b}^{(m+2)}_{k_{m+1}} : k_{m+1} = 1, \dots, r_{m+1}\}$ are orthonormal sets. This can

be done by applying the singular value decomposition to the left unfolding (5.10) of the m-th component

$$V_m^{\mathsf{L}} = \tilde{V}_m^{\mathsf{L}} \Sigma_m Y_m^{\mathsf{T}}$$

and shifting $\Sigma_m, Y_m^{\mathsf{T}} \in \mathbb{R}^{r_m \times r_m}$ to the right unfolding (5.9) of the next component

$$\tilde{V}_{m+1}^{\mathsf{R}} = \Sigma_m Y_m^{\mathsf{T}} V_{m+1}^{\mathsf{R}}.$$

For V we obtain

$$\begin{aligned} V &= \sum_{k_{m-1}=1}^{r_{m-1}} \sum_{k_m=1}^{r_m} \sum_{\mu_m=1}^{q_m} V_m(k_{m-1}, \mu_m, k_m) b_{k_{m-1}}^{(m-1)} \otimes e_{\mu_m}^{(m)} \otimes \bar{b}_{k_m}^{(m+1)} \\ &= \sum_{k_m=1}^{r_m} \sum_{k_{m+1}=1}^{r_{m+1}} \sum_{\mu_{m+1}=1}^{q_{m+1}} \tilde{V}_{m+1}(k_m, \mu_{m+1}, k_{m+1}) b_{k_m}^{(m)} \otimes e_{\mu_{m+1}}^{(m+1)} \otimes \bar{b}_{k_{m+1}}^{(m+2)}. \end{aligned} \tag{5.12}$$

Alternatively one may use QR factorisation for the orthogonalisation, but it is often advantageous to keep the small diagonal matrix $\Sigma_m \in \mathbb{R}^{r_m \times r_m}$ containing the singular values in between to adjacent component tensors.

In fact this provides a *standard representation* or *HSVD representation* of V, see Figure 6

$$V = X_1 \circ \Sigma_1 \circ X_2 \circ \Sigma_2 \circ \cdots \circ \Sigma_{M-1} \circ X_M.$$

This representation has been developed independently by different authors [88, 64, 26]. In physics, it is accredited to Vidal and is hence also known as the *Vidal representation*. Very beneficial is the criterion that

$$X_1^{\mathsf{T}} X_1 = I_{r_1}, \quad X_M X_M^{\mathsf{T}} = I_{r_{M-1}}$$

and for all $1 < m < M$

$$\begin{aligned} \Big((\Sigma_{m-1} \circ X_m)^{\mathsf{L}} \Big)^{\mathsf{T}} (\Sigma_{m-1} \circ X_m)^{\mathsf{L}} &= I_{r_{m-1}}, \\ (X_m \circ \Sigma_m)^{\mathsf{R}} \Big((X_m \circ \Sigma_m)^{\mathsf{R}} \Big)^{\mathsf{T}} &= I_{r_m}. \end{aligned}$$

This means, we can shift the root, and thus the orthogonality, by simply shifting the *density matrices* Σ_m, see Figure 7.

This representation can be computed by applying a sequence of singular value decompositions and storing the singular values. The procedure is called *Hierarchical SVD (HSVD)*. It recovers the tensor exactly. However, as mentioned for the Tucker format and the HOSVD, the HSVD can be used for approximation by thresholding the singular values. For density matrices $\Sigma = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_r)$ we define two thresholding operators

$$\begin{aligned} H_{\tilde{r}}(\Sigma) &= \operatorname*{diag}_{1 \le k \le r}(\varsigma_k), \quad \tilde{r} \le r, \\ H_{\epsilon}(\Sigma) &= \operatorname*{diag}_{\varsigma_k \ge \epsilon}(\varsigma_k), \quad \epsilon > 0, \end{aligned}$$

and for TT tensors

$$\begin{aligned} H_{\tilde{r}}(U) &= X_1 \circ H_{\tilde{r}_1}(\Sigma_1) \circ X_2 \circ H_{\tilde{r}_2}(\Sigma_2) \circ \cdots \circ H_{\tilde{r}_{M-1}}(\Sigma_{M-1}) \circ X_M, \\ H_{\epsilon}(U) &= X_1 \circ H_{\epsilon_1}(\Sigma_1) \circ X_2 \circ H_{\epsilon_2}(\Sigma_2) \circ \cdots \circ H_{\epsilon_{M-1}}(\Sigma_{M-1}) \circ X_M. \end{aligned}$$

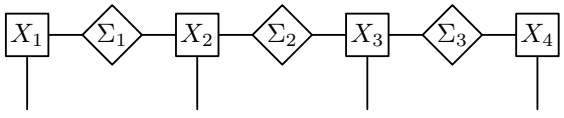

Figure 6: A TT tensor of order 4 in standard representation.

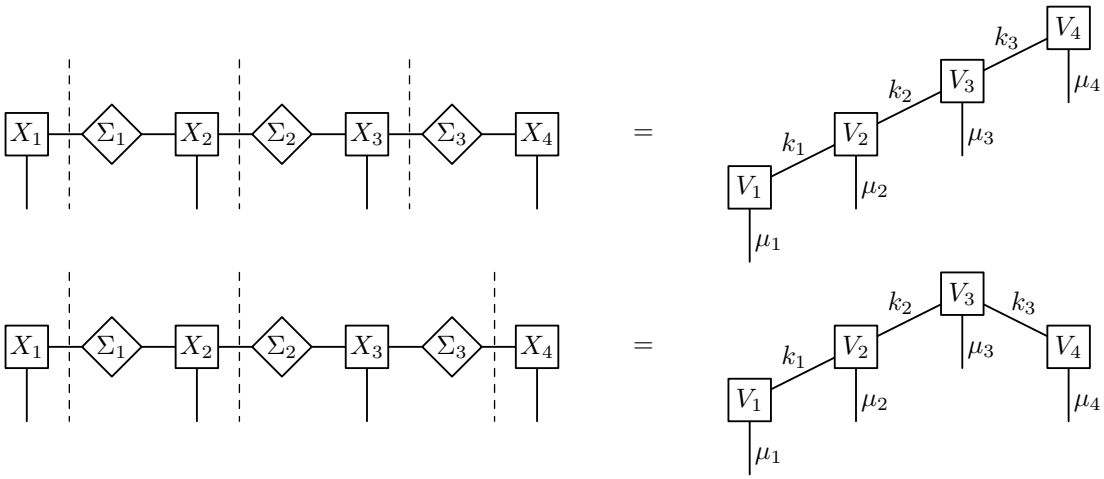

Figure 7: A shift of orthogonality in the standard representation. The root denotes the only component that is not orthogonalised. A shift of the density matrix Σ_3 to the left shifts the root to the component V_3 and V_4 is now right orthogonalised.

Again, this will not yield the best approximation of the tensor, as it does in the matrix case. As with Tucker tensors, we maintain a so called *quasi-optimality*, see [64, 26, 30, 31].

Theorem 5.7 (Quasi-Optimality)**.** *The truncation of the HSVD can be estimated by*

$$\|V - H_{\tilde{r}}(V)\| \leq \sqrt{M-1} \inf_{W \in \mathcal{M}_{\leq \tilde{r}}} \|V - W\|,$$

where $\mathcal{M}_{\leq \tilde{r}}$ is the space of all tensors with TT rank not exceeding $\tilde{r}$.

As most other results, the separation theorem and the quasi optimality can be readily generalised to all tree tensor networks. It is also possible to formulate a standard representation for other trees. In contrast to the parametrisation (5.3), the subspace representation provides further essential information about minimal representability and approximability.

Basic Operations in TT Format

Lastly, before we discuss the more general HT format, let us briefly comment on the efficiency of basic operations in the TT format. These notions are again similar for all tensor trees. Let $V, W \in \mathbb{R}^{q_1 \times \dots \times q_M}$ be two TT tensors with TT ranks r and r' respectively.

The sum $V + W \in \mathbb{R}^{q_1 \times \dots \times q_M}$ is again a TT tensor with ranks at most $r + r'$. This can be stated component-wise and element-wise for $m = 1, \dots, M$:

$$(V+W)_m(k''_{m-1}, \mu_m, k''_m) = \begin{cases} V_m(k_{m-1}, \mu_m, k_m), & 1 \leq k''_{m-1} \leq r_{m-1} \text{ and } 1 \leq k''_m \leq r_m, \\ W_m(k'_{m-1}, \mu_m, k'_m), & \begin{array}{l} r_{m-1}+1 \leq k''_{m-1} \leq r_{m-1} + r'_{m-1} \\ \text{and } r_m + 1 \leq k''_m \leq r_m + r'_m, \end{array} \\ 0, & \text{otherwise.} \end{cases}$$

Note that this definition yields a representation of the sum $V + W$, but not necessarily a decomposition. We can compute the decomposition with the exact ranks $r'' \leq r + r'$ if we perform a HSVD on the above representation.

The scalar product of the TT tensors V and W can also be computed efficiently:

$$\langle V, W \rangle = \sum_{k_1=1}^{r_1} \cdots \sum_{k_{M-1}=1}^{r_{M-1}} \sum_{k'_1=1}^{r'_1} \cdots \sum_{k'_{M-1}=1}^{r'_{M-1}} \left(\prod_{m=1}^{M} \sum_{\mu_m=1}^{q_m} V_m(k_{m-1}, \mu_m, k_m) W_m(k'_{m-1}, \mu_m, k'_m) \right). \tag{5.13}$$

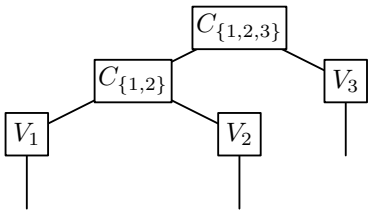

Figure 8: A tensor of order 3 in HT format.

This means, for all $m = 1, \dots, M$, we sum over the physical indices μ_m for the components V_m and W_m and all that remains is a number of matrix-vector multiplications. Both the addition and the scalar product of TT tensors can of course only be performed on tensors of the same physical dimensions, as otherwise they would be elements of different tensor spaces and these operations are not generally defined then.

There are numerous other operations that are efficient in the TT format and in all other tree formats. Possibly the most notable operation is the multiplication of TT matrices and TT tensors, see [64]. All these operations have been defined in many articles and the reader is referred elsewhere for this reason, e.g., the survey of Bachmayr et al. [5].

5.4.5 Dimension Trees and the Hierarchical Tensor Decomposition

We briefly discuss the *Hierarchical Tucker* or *Hierarchical Tensor* (HT) representation that has been introduced by Hackbusch and Kühn [29] in 2009 and has since received a lot of attention. This is also due to the fact that it is a reasonable generalisation of the TT format.

The HT representation is defined by a *dimension tree*, usually a binary tree, where the leafs $V_1, \dots, V_M$ constitute the physical components and the inner vertices C_t are virtual. Hackbusch gives the following comprehensive notation in [30]: The vertices of the tree tensor network $TTNS(V) = (\mathtt{V}, \mathtt{E}, \mathtt{H})$ are labelled

(i) $t_M = \{1, \dots, M\}$ for the root,

(ii) $t \in \mathtt{L} := \big\{\{1\}, \dots, \{M\}\big\}$ for the leafs and

(iii) $t \in \mathtt{V} \setminus \mathtt{L}$ for inner vertices, which have sons $t_1, \dots, t_P$ that are an ordered partition of t, i.e.

$$\bigcup_{p=1}^{P} t_p = t \text{ and } \mu < \nu \quad \forall \mu \in t_p, \nu \in t_q, p < q.$$

For an inner vertex $t \subset \mathtt{V} \setminus \mathtt{L}$, with sons $t_1, \dots, t_P$ (usually $P = 2$), there is a subspace $\mathcal{W}_t$ defined by its basis set $\{b_{k_t}^{(t)} : k_t = 1, \dots, r_t\}$ given by

$$b_{k_t}^{(t)} = \sum_{k_{t_1}=1}^{r_{t_1}} \cdots \sum_{k_{t_P}=1}^{r_{t_P}} V_t(k_{t_1}, \dots, k_{t_P}, k_t) b_{k_{t_1}}^{(t_1)} \otimes \cdots \otimes b_{k_{t_P}}^{(t_P)},$$

see [21]. The root $t_M = \{1, \dots, M\}$, with sons $t_1, \dots, t_P$, is to reconstruct the tensor

$$V = \sum_{k_{t_1}=1}^{r_{t_1}} \cdots \sum_{k_{t_P}=1}^{r_{t_P}} V_{t_M}(k_{t_1}, \dots, k_{t_p}) b_{k_{t_1}}^{(t_1)} \otimes \cdots \otimes b_{k_{t_p}}^{(t_p)}.$$

Therefore the tensor V is defined completely by the component tensors V_t, using the above representations recursively, see Figure 8. There are at most $\mathcal{O}(M)$ vertices and consequently the complexity is $\mathcal{O}(qMR + mR^{P+1})$. For $P = 2$ we obtain $\mathcal{O}(qMR + MR^3)$ [30, 31].

As with the Tucker and the TT format, obtaining the HT format can be done by applying the singular value decomposition successively, in a hierarchical fashion. Again, we maintain a

well-defined rank through a separation theorem, a quasi optimality of a truncated HSVD and so on.

In fact, the Tensor Train decomposition can be seen as a special case of the Hierarchical Tucker decomposition, where we use an unbalanced tree and omit the optimal subspaces in the leafs. In the HT format, the leafs $V_1, \dots, V_M$ form exactly the optimal subspaces already observed in the Tucker decomposition and the tree is just an additional decomposition of the core tensor. Arguably, this could make the HT format superior to the TT format. However, the notation becomes very messy and all notable theoretical results are valid for any tree tensor network. Hence, we refrain from dealing with the Hierarchical format and proceed with the Tensor Train format, keeping the similarities in mind. Additionally, in the case of parametric PDEs, the sequential nature of the TT format is intuitively more reasonable, because the coefficient functions naturally decay, see Chapter 3. However, in some cases, the binary tree structure can be advantageous [89]. This will be briefly addressed in Chapter 6.

5.4.6 Fixed Rank Manifolds and Varieties

In order to reduce complexity and overcome the curse of dimensionality, we consider the set of tensors of fixed (low) TT rank

$$\mathcal{M}_r := \{V \in \mathbb{R}^{q_1 \times \dots \times q_M} : r_{\mathrm{TT}} = r\}.$$

If r is not a feasible TT rank, then this set is empty. In the following, we assume that $\mathcal{M}_r$ is non-empty, i.e. there is at least one tensor $V \in \mathbb{R}^{q_1 \times \dots \times q_M}$ that has TT rank r. The set $\mathcal{M}_r$ is no longer a linear space nor is it convex and therefore numerical approximation on this space is not trivial. However, the following holds, see [84]:

Theorem 5.8 (Fixed-rank manifold)**.** *The space $\mathcal{M}_r$ forms a manifold of dimension*

$$\dim \mathcal{M}_r = \sum_{m=1}^{M} r_{m-1} q_m r_m - \sum_{m=1}^{M-1} r_m^2.$$

This manifold can be globally embedded in the tensor space $\mathcal{M}_r \subset \mathbb{R}^{q_1 \times \dots \times q_M}$ and we call it the TT manifold*.*

The local embedding property was first shown in [36], alongside a parametrisation of the tangent space $\mathrm{T}_V \mathcal{M}_r$ in $V \in \mathcal{M}_r$. For this, we fix the orthogonalisation of the tensor such that all but the last component are left orthogonal $(V_m^{\mathsf{L}})^{\mathsf{T}} V_m^{\mathsf{L}} = I_{r_m}, m = 1, \dots, M$, i.e. using the standard representation,

$$V_1 := X_1, \tag{5.14}$$

$$V_m := \Sigma_m \circ X_m \quad \forall 1 < m \leq M. \tag{5.15}$$

Additionally, we introduce the *insertion operator*

$$\begin{aligned} E_m : \mathbb{R}^{r_{m-1} \times q_m \times r_m} &\to \mathbb{R}^{q_1 \times \dots \times q_M}, \\ W_m &\mapsto E_m W_m = V_1 \circ \dots \circ V_{m-1} \circ W_m \circ V_{m+1} \circ \dots \circ V_M. \end{aligned}$$

We remark that for left orthogonal tensors that fulfil (5.14) and (5.15), it holds

$$E_m E_m^{\mathsf{T}} = I_{r_{m-1}} \otimes I_{q_m} \otimes \Sigma_m^2,$$

i.e. the pseudo-inverse of the insertion operator is

$$\begin{aligned} E_m^{\dagger} : \mathbb{R}^{q_1 \times \dots \times q_M} &\to \mathbb{R}^{r_{m-1} \times q_m \times r_m}, \\ W &\mapsto E_m^{\dagger} W = (E_m E_m^{\mathsf{T}})^{-1} E_m^{\mathsf{T}} W = (E_m^{\mathsf{T}} W) \circ \Sigma_m^{-2}, \end{aligned}$$

for $m = 1, \dots, M-1$ and

$$\begin{aligned} E_M^{\dagger} : \mathbb{R}^{q_1 \times \dots \times q_M} &\to \mathbb{R}^{r_{M-1} \times q_m}, \\ W &\mapsto E_M^{\dagger} W = E_M^{\mathsf{T}} W. \end{aligned}$$

With this, we can state the lemma:

Lemma 5.9. *For $V \in \mathcal{M}_r$, any element $\xi_V \in \mathrm{T}_V\mathcal{M}_r$ can be represented uniquely by the product rule as*

$$\xi_V = \sum_{m=1}^{M} V_1 \circ \cdots \circ \xi_{V,m} \circ \cdots \circ V_M = \sum_{m=1}^{M} E_m \xi_{V,m},$$

where $\xi_{V,m}$ satisfies the gauge conditions $(\xi_{V,m}^{\mathsf{L}})^{\mathsf{T}} V_m^{\mathsf{L}} = 0$ for $m = 1, \ldots, M-1$.

Furthermore, [57] give a formula for the orthogonal projection onto the tangent space:

$$\begin{aligned} P_V &: \mathbb{R}^{q_1 \times \cdots \times q_M} \to \mathrm{T}_V\mathcal{M}_r, \\ P_V(F) &= \sum_{m=1}^{M-1} V_1 \circ \cdots \circ \Big(I_{r_{m-1}q_m} - V_m^{\mathsf{L}} (V_m^{\mathsf{L}})^{\mathsf{T}} \Big) \circ E_m^{\dagger} F \circ \cdots \circ V_M \\ &\quad + V_1 \circ \cdots \circ V_{M-1} \circ E_M^{\mathsf{T}} F. \end{aligned} \tag{5.16}$$

The manifold $\mathcal{M}_r$ is *Riemannian*, i.e. it is be equipped with a Riemannian metric

$$g_V : \mathrm{T}_V\mathcal{M}_r \times \mathrm{T}_V\mathcal{M}_r \to \mathbb{R}$$

that is symmetric, bilinear and positive definite as well as smooth in $V \in \mathcal{M}_r$. Since the manifold is embedded, a possible metric is inherited from the ambient space $\mathbb{R}^{q_1 \times \cdots \times q_M}$, i.e.

$$\begin{aligned} g_V(\xi_V, \eta_V) &= \left\langle \sum_{m=1}^{M} V_1 \circ \cdots \circ \xi_{V,m} \circ \cdots \circ V_M, \sum_{m=1}^{M} V_1 \circ \cdots \circ \eta_{V,m} \circ \cdots \circ V_M \right\rangle \\ &= \mathrm{Tr}\left(\sum_{m=1}^{M-1} \left((\xi_{V,m} \circ \Sigma_m^2)^{\mathsf{L}} \right)^{\mathsf{T}} (\eta_{V,m})^{\mathsf{L}} \right) + \mathrm{Tr}\left(\left((\xi_{V,M})^{\mathsf{L}} \right)^{\mathsf{T}} (\eta_{V,M})^{\mathsf{L}} \right). \end{aligned} \tag{5.17}$$

With this, we can also define a norm $\|\cdot\|$ on $\mathcal{M}_r$ that is independent of V because g is independent of V.

The manifold $\mathcal{M}_r$ can be parametrised by the component space

$$\begin{aligned} \mathcal{C} &= \{\underline{V} = (V_1, \ldots, V_M) : V_m \in \mathbb{R}_*^{r_{m-1} \times q_m \times r_m}\} \\ &\cong \bigtimes_{m=1}^{M} \mathbb{R}_*^{r_{m-1} \times q_m \times r_m}. \end{aligned}$$

For each m, $\mathbb{R}_*^{r_{m-1} \times q_m \times r_m}$ is the space of all elements with full multilinear rank

$$\begin{aligned} \mathbb{R}_*^{r_{m-1} \times q_m \times r_m} = \{V_m \in \mathbb{R}^{r_{m-1} \times q_m \times r_m} &: \mathrm{rank}(V_m^{\mathsf{R}}) = \max(r_{m-1}, q_m r_m), \\ &\mathrm{rank}(V_m^{\mathsf{L}}) = \max(r_{m-1} q_m, r_m)\}. \end{aligned}$$

Let it be noted that this space is a smooth manifold [1] and thus their cartesian product $\mathcal{C}$ is also a smooth manifold. This relation is helpful in many numerical applications and it is shown in [84].

The Algebraic Variety of Low Rank Tensors

The manifold $\mathcal{M}_r$ is not closed. However, in finite dimensions, its closure is given by

$$\overline{\mathcal{M}_r} = \mathcal{M}_{\leq r}.$$

This is based on the observation that the matrix rank is an upper semi-continuous function [21, 30]. The singular points are exactly those where the actual rank is not maximal.

As mentioned above, the set $\mathcal{M}_{\leq r}$ is Zariski-closed and thus forms a *real algebraic variety*, i.e. it is the set of common zeros of real polynomials. This is easy to see: We know from the Separation Theorem 5.6, that $\mathcal{M}_{\leq r}$ is the intersection of all tensors where the corresponding

matricisations $[V]_{1,\dots,m}^{m+1,\dots,M}$ have at most rank r_m. The sets of matrices with rank at most r_m are known to be real algebraic varieties [72], each some zero-set of polynomials [46]. Then, trivially, the intersection is the zero-set of the union of all such polynomials. Again, this property generalises to all tensor trees.

Recently, a method for the desingularisation of the low rank tensor variety has been proposed [42]. This consists of tracking the tangential directions in addition to the tensor, thus "unfolding" the variety and obtaining a manifold structure. While this method is subject to current research, it promises to yield new numerical techniques and simplify the analysis of the old algorithms that will be introduced next.

5.5 Numerical Techniques for Tensors

Numerical optimisation with tensors mostly aims at approximating a problem by finding a low-rank solution $U \in \mathcal{M}_r$ or $U \in \mathcal{M}_{\leq r}$. It is also possible to develop numerical algorithms on the quotient manifold $\mathcal{C}/\mathcal{G}$ using similar techniques for quotient manifolds. We will focus on the TT manifold instead, but keep in mind that all Riemannian methods also work on the quotient manifold with slight modifications.

In general, we want to minimise a continuously differentiable functional

$$j : \mathbb{R}^{q_1 \times \dots \times q_M} \to \mathbb{R}$$

over the set of low-rank tensors, i.e. we want to obtain a solution

$$U = \operatorname*{argmin}_{V \in \mathcal{M}_r} j(V)$$

or similarly for $U \in \mathcal{M}_{\leq r}$. Standard examples for such functionals are

- Low-rank approximation:
$$j(V) = \|V - F\|,$$
- Linear systems:
$$j(V) = \frac{1}{2}\langle \mathbf{A}V, V\rangle - \langle F, V\rangle,$$
- Eigenvalue problems:
$$j(V) = \frac{\langle \mathbf{A}V, V\rangle}{\langle V, V\rangle}, \quad V \neq 0.$$

The most common optimisation algorithms will be discussed in the following, as well as parts of their convergence analysis.

5.5.1 Riemannian Optimisation

Having shown that the low rank tensors form an embedded manifold in the tensor space $\mathbb{R}^{q_1 \times \dots \times q_M}$, the most natural optimisation technique to apply is Riemannian optimisation. This essentially consists of steps in the general direction of the negative gradient using first or second-order information, and a line search procedure to determine the step length. However, on manifolds, we have to proceed with care, as certain generalisations of the Euclidean case are necessary.

First of all, we define a *retraction* from the tangent bundle to the manifold. Here, we follow [2, 44].

Definition 5.10 (Retraction)**.** A smooth mapping R from the tangent bundle $\mathrm{T}\mathcal{M}_r$ onto the manifold $\mathcal{M}_r$ is called a *retraction on* $\mathcal{M}_r$, if for any $V \in \mathcal{M}_r$ there exists a neighbourhood $\mathcal{U} \subset \mathrm{T}\mathcal{M}_r$ around $(V, 0_V)$ such that

(i) R is defined and smooth on $\mathcal{U}$,

(ii) $R(W, 0_W) = W$ for all $(W, 0_W) \in \mathcal{U}$,

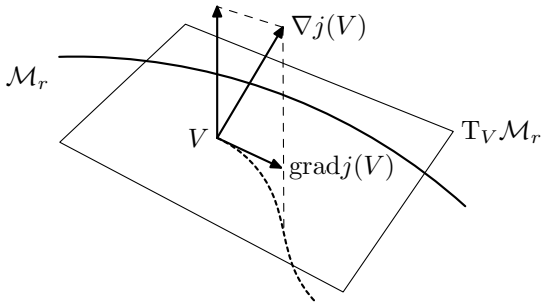

Figure 9: An illustration of the gradient descend on the manifold $\mathcal{M}_r$.

(iii) $\frac{\mathrm{d}}{\mathrm{d}t} R(W, t\xi_W)\Big|_{t=0} = \xi_W$ for all $(W, \xi_W) \in \mathcal{U}$.

Here, 0_V denotes the zero element of $\mathrm{T}_V\mathcal{M}_r$ and for the *local rigidity property* (iii), we used the canonical identification $\mathrm{T}_{0_W}\mathrm{T}_W\mathcal{M}_r \cong \mathrm{T}_W\mathcal{M}_r$.

Remark 5.11. This definition varies from the one in [1] because it is defined and smooth only on small neighbourhoods in the tangent bundle. This is necessary as the manifold $\mathcal{M}_r$ is not necessarily connected and a global retraction would not make much sense here. Additionally, as opposed to [44], we defined the retraction for all $V \in \mathcal{M}_r$, but this is straightforward.

Given the Riemannian metric g, the *Riemannian gradient* $\mathrm{grad}\, j(V)$ at $V \in \mathcal{M}_r$ is the unique element that satisfies

$$g_V(\mathrm{grad}\, j(V), \xi_V) = \langle \nabla j(V), \xi_V \rangle \quad \forall \xi_V \in \mathrm{T}_V\mathcal{M}_r.$$

Since $\mathcal{M}_r$ is an embedded manifold and j can be defined on the whole ambient space, this is well defined and it simply yields

$$\mathrm{grad}\, j(V) = P_V(\nabla j(V)),$$

if the metric is the ambient metric (5.17). As mentioned above, this metric is independent of V.

A point on the manifold $V^* \in \mathcal{M}_r$ is a *critical point* if

$$\mathrm{grad}\, j(V^*) = 0_V.$$

This could be a global or a local minimum on $\mathcal{M}_r$, but it could also be a saddle point or even a maximum. On embedded manifolds, we can define the *normal space* $\mathrm{N}_V\mathcal{M}_r$ at $V \in \mathcal{M}_r$, that is the orthogonal complement of the tangent space. Since the Riemannian gradient is defined via the orthogonal projection P_V, we know that for any critical point $V^* \in \mathcal{M}_r$ it holds

$$\nabla j(V^*) \in \mathrm{N}_{V^*}\mathcal{M}_r.$$

We denote a series of tensors by $(V_{[i]})_{i\in\mathbb{N}}$ in order to avoid confusion with the tensor components V_m. The most basic Riemannian optimisation algorithm is *Riemannian gradient descend*. For this, we choose a random starting point $V_{[0]} \in \mathcal{M}_r$ and iterate with

$$V_{[i+1]} = R(V_{[i]}, -\alpha_i \,\mathrm{grad}\, j(V_{[i]})).$$

The step sizes α_i are chosen such that they satisfy the *Armijo condition* [62]

$$j(R(V_{[i]}, \alpha_i \xi_{V_{[i]}})) \leq j(V_{[i]}) + c_1 \alpha_i \langle \mathrm{grad}\, j(V_{[i]}), \xi_{V_{[i]}} \rangle,$$

with the search directions $\xi_{V_{[i]}} = -\,\mathrm{grad}\, j(V_{[i]})$ and paired with an adequate backtracking procedure, see [62]. Figure 9 illustrates this simple method.

This method can be improved using a better line search procedure than the simple backtracking, possibly by finding the analytic minimum of $j(R(V_{[i]}, \alpha_i \xi_{V_{[i]}}))$ with respect to α_i, but also by improving the search direction $\xi_{V_{[i]}}$. This can be done using a heuristic or otherwise justified preconditioner, most notably the Hessian or some approximation of it, which results in Riemannian Newton or Quasi-Newton methods. The definition of second derivatives on general Riemannian manifolds requires the notion of *Riemannian connections*, which exceeds the scope of this thesis. The interested reader is referred to the book by Absil et al. [1] or, for more fundamental differential geometry, to the book by Lee [50].

For convergence analysis of Riemannian optimisation methods in tensor formats, we refer to the book by Absil et al. [1] and to the works by Kressner et al. [86, 44, 76]. Essentially, if the sequence of search directions $(\xi_{V_{[i]}})_{i\in\mathbb{N}}$ is *gradient related*, i.e. their angle with the Riemannian gradient is almost always bounded away from zero, it is possible to show that every accumulation point of the series $(V_{[i]})_{i\in\mathbb{N}}$ is a critical point of j. If the *level set* of j for the starting point $V_{[0]}$,

$$\mathrm{Level}(j, V_{[0]}) = \{V \in \mathcal{M}_r : j(V) \leq j(V_{[0]})\}$$

is compact, then we clearly get

$$\lim_{i\to\infty} \|\operatorname{grad} j(V_{[i]})\| = 0.$$

However, this is not generally the case. The manifold $\mathcal{M}_r$ is itself not compact and highly non-convex. Its curvature tends to infinity close to the points where singular values vanish. It is possible to construct a regularised functional $h : \mathbb{R}^{q_1\times\cdots\times q_M} \to \mathbb{R}$ that ensures a compact level set and keeps the iteration away from the singularities, see [86, 76] and the references therein. This means that local minima can only be found on the connected component of $\mathcal{M}_r$ that one has started on.

Another technique for optimisation on the TT manifold $\mathcal{M}_r$ is *Dynamical Low Rank Approximation*, see [57]. It turns out that this procedure defines a retraction onto the manifold and it is a Riemannian optimisation method. Therefore, the convergence analysis is the same. Recently, a projector splitting procedure for Dynamical Low Rank Approximation has been introduced in [55, 56] that allows to perform the projection of the gradient successively and optimising each component V_m independently. This is akin to one of the most widely used methods in tensor approximation that we will introduce next.

5.5.2 The Alternating Least Squares Algorithm

We consider the functional

$$\begin{aligned} J : \mathcal{C} &\to \mathbb{R} \\ (V_1, \ldots, V_M) &\mapsto J(V_1, \ldots, V_M) := j(V_1 \circ \cdots \circ V_M)) \end{aligned}$$

which is defined on the components instead of the full tensor. For $m \in \{1, \ldots, M\}$ we fix $V_1, \ldots, V_{m-1}$ and $V_{m+1}, \ldots, V_M$ and solve the subproblem

$$V_m^+ := \operatorname*{argmin}_{W_m \in \mathbb{R}^{r_{m-1}\times q_m \times r_m}} J(V_1, \ldots, W_m, \ldots, V_M).$$

This is done in a successive manner and with alternating directions, which - for the best least squares fit $j(V) = \|V - F\|$ - justifies the name *Alternating Least Squares* (ALS) algorithm. This is an instance of the well-known Gauß-Seidel iteration.

The TT format allows for a special formulation of this algorithm, sometimes dubbed the *Alternating Linear Scheme* to maintain the abbreviation. In this case, we can give a closed form for each subproblem and they can be solved using standard tools from linear algebra and numerical optimisation.

In every step, one has to solve a small problem in order to achieve the minimum. Note that we allow $V_m \in \mathbb{R}^{r_{m-1}\times q_m \times r_m}$, i.e. the ranks can decrease in each step. This automatically restricts j to the variety $\mathcal{M}_{\leq r}$ since the components can have full rank or less, but obviously not more than that.

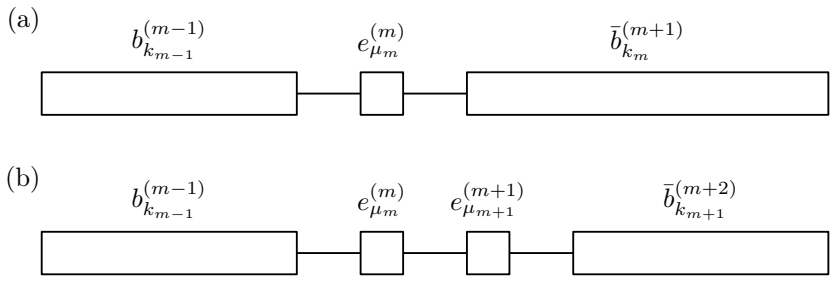

Figure 10: Reduced basis representation for (a) the ALS algorithm, and (b) the two-site DMRG. The reduced basis vectors span a subspace of the large spaces $\mathbb{R}^{q_1 \times \cdots \times q_{m-1}}$ and $\mathbb{R}^{q_{m+1} \times \cdots \times q_M}$ or $\mathbb{R}^{q_{m+2} \times \cdots \times q_M}$ for the ALS and the DMRG respectively.

The small subproblems will be of the same kind as the original problem, i.e. linear equations will be turned into small linear equations and eigenvalue problems give rise to relatively small (generalised) eigenvalue problems. In physics this supports the renormalisation picture, where an originally large system is reduced to a small system with the same ground state energy and possibly further physical quantities.

As we have observed before, this simple approach should be realised with some care. Since the representation is redundant, minimising over the full parameter space $\mathbb{R}^{r_{m-1} \times q_m \times r_m}$ yields unstable results. Rather, one should optimise over some non-linear quotient space and it would become necessary to introduce gauge conditions like above. However, this can be avoided if we choose to minimise only the root of the tensor as there is no redundancy in this part. After the minimisation, it would then be crucial to restructure the hierarchy of the tensor and consider the next component as the root. This can be done by shifting the orthogonality as explained in (5.12). The extension to general hierarchical trees is straightforward.

Conforming with the earlier notation (5.11), each subproblem becomes a problem over a small subset that constitutes a subspace

$$\mathcal{L}_m \otimes \mathbb{R}^{q_m} \otimes \mathcal{R}_m \subseteq \mathbb{R}^{q_1 \times \cdots \times q_M}.$$

We define the orthogonal projector onto this space

$$P_m : \mathbb{R}^{q_1 \times \cdots \times q_M} \to \mathcal{L}_m \otimes \mathbb{R}^{q_m} \otimes \mathcal{R}_m.$$

If we choose orthogonal bases $b^{(m-1)}_1, \ldots, b^{(m-1)}_{r_{m-1}}$ and $\bar{b}^{(m+1)}_1, \ldots, \bar{b}^{(m+1)}_{r_m}$, we obtain

$$P_m = E_m E_m^{\mathsf{T}}.$$

This can easily be seen, as for $W \in \mathbb{R}^{q_1 \times \cdots \times q_M}$ it holds

$$\begin{aligned} P_m W &= \sum_{k_{m-1}=1}^{r_{m-1}} \sum_{k_m=1}^{r_m} \sum_{\mu_m=1}^{q_m} \tilde{W}_m(k_{m-1}, \mu_m, k_m) b^{(m-1)}_{k_{m-1}} \otimes e_{\mu_m} \otimes \bar{b}^{(m+1)}_{k_m} \\ &= \sum_{\mu_1=1}^{q_1} \cdots \sum_{\mu_M=1}^{q_M} E_m \tilde{W}_m(\mu_1, \ldots, \mu_M) e^{(1)}_{\mu_1} \otimes \ldots \otimes e^{(M)}_{\mu_M} \end{aligned}$$

and

$$E_m^{\mathsf{T}} W = \tilde{W}_m.$$

Note that E_m is a bijection onto its image, and since it is also orthogonal, its transpose is well-defined as its inverse. See Figure 10(a) for an illustration of the reduced basis.

To formulate the procedure explicitly, we consider a linear system

$$\mathbf{A}U = F, \tag{5.18}$$

which is equivalent to minimising

$$j(V) = \frac{1}{2}\langle \mathbf{A}V, V\rangle - \langle F, V\rangle,$$

where $\mathbf{A} \in \mathcal{L}(\mathbb{R}^{q_1\times\dots\times q_M}, \mathbb{R}^{q_1\times\dots\times q_M})$ is a linear operator that is symmetric and positive definite (SPD). This operator can be stored and viewed in a canonical-like format, i.e. as a sum of rank-one tensor products

$$\mathbf{A} = \sum_{k=1}^{S} A_{1,k} \otimes \cdots \otimes A_{M,k}$$

or even in a TT-Matrix or Matrix Product Operator (MPO) format [63]. The Galerkin operator (4.13) of the parametric problem with affine diffusion coefficient is of the former kind and the log-normal Galerkin operator (4.17) can be approximated in the latter form as we will see in the following.

A single subproblem can then be expressed as

$$\begin{aligned} V_m^+ &= \operatorname*{argmin}_{W_m\in\mathbb{R}^{r_{m-1}\times q_m\times r_m}} j(V_1,\dots,W_m,\dots,V_M) \\ &= \operatorname*{argmin}_{W_m\in\mathbb{R}^{r_{m-1}\times q_m\times r_m}} \big(\frac{1}{2}\langle \mathbf{A}E_mW_m, E_mW_m\rangle - \langle F, E_mW_m\rangle\big) \\ &= \operatorname*{argmin}_{W_m\in\mathbb{R}^{r_{m-1}\times q_m\times r_m}} \big(\frac{1}{2}\langle E_m^\mathsf{T}\mathbf{A}E_mW_m, W_m\rangle - \langle E_m^\mathsf{T}F, W_m\rangle\big). \end{aligned}$$

At stationary points W_m^* of the functional $j\circ E_m$, there holds the first order condition

$$\nabla(j\circ E_m)(V_m) = E_m^\mathsf{T}\mathbf{A}E_mW_m^* - E_m^\mathsf{T}F = 0.$$

As such, one micro-iteration of the ALS algorithm can be defined as

$$\begin{aligned} V^+ &:= V_1\circ\cdots\circ V_m^+\circ\cdots\circ V_M, \\ V_m^+ &= \big(E_m^\mathsf{T}\mathbf{A}E_m\big)^{-1}E_m^\mathsf{T}F. \end{aligned} \tag{5.19}$$

See Figure 11(a) for an illustration.

In the subspace notation, we get

$$V^+ = \operatorname*{argmin}_{W\in\mathcal{L}_m\otimes\mathbb{R}^{q_m}\otimes\mathcal{R}_m} \big(\frac{1}{2}\langle \mathbf{A}W, W\rangle - \langle F, W\rangle\big) = P_m\mathbf{A}^{-1}P_mF.$$

Algorithm 1 summarises the ALS method for solving a linear system (5.18).

For this to work, $\mathbf{A}$ does not necessarily have to be invertible on the whole tensor space but only on the small subspaces $\mathcal{L}_m\otimes\mathbb{R}^{q_m}\otimes\mathcal{R}_m$. This is of course guaranteed if $\mathbf{A}$ is invertible as a whole. Additionally, one can see that the eigenvalues of $\mathbf{A}$ on $\mathcal{L}_m\otimes\mathbb{R}^{q_m}\otimes\mathcal{R}_m$ are bounded by the eigenvalues on the whole space and in particular it holds $\operatorname{cond}_m(\mathbf{A}) \le \operatorname{cond}(\mathbf{A})$ [37].

General convergence theory of the ALS is subject to research [69]. One cannot assume that the update directions in each component are related to the full gradient. However, it is possible to understand the ALS as a Riemannian-like method. For linear systems, we can write the global update in each subproblem as follows:

$$\begin{aligned} V^+ &= E_m\big(E_m^\mathsf{T}\mathbf{A}E_m\big)^{-1}E_m^\mathsf{T}F \\ &= E_mV_m - E_m\big(E_m^\mathsf{T}\mathbf{A}E_m\big)^{-1}\Big(\big(E_m^\mathsf{T}\mathbf{A}E_m\big)V_m - E_m^\mathsf{T}F\Big) \\ &= V - E_m\big(E_m^\mathsf{T}\mathbf{A}E_m\big)^{-1}E_m^\mathsf{T}\Big(E_mE_m^\mathsf{T}\big(\mathbf{A}V - F\big)\Big) \end{aligned}$$

and if we define the preconditioner

$$\mathbf{P}_{m,V} := E_m\big(E_m^\mathsf{T}\mathbf{A}E_m\big)^{-1}E_m^\mathsf{T}$$

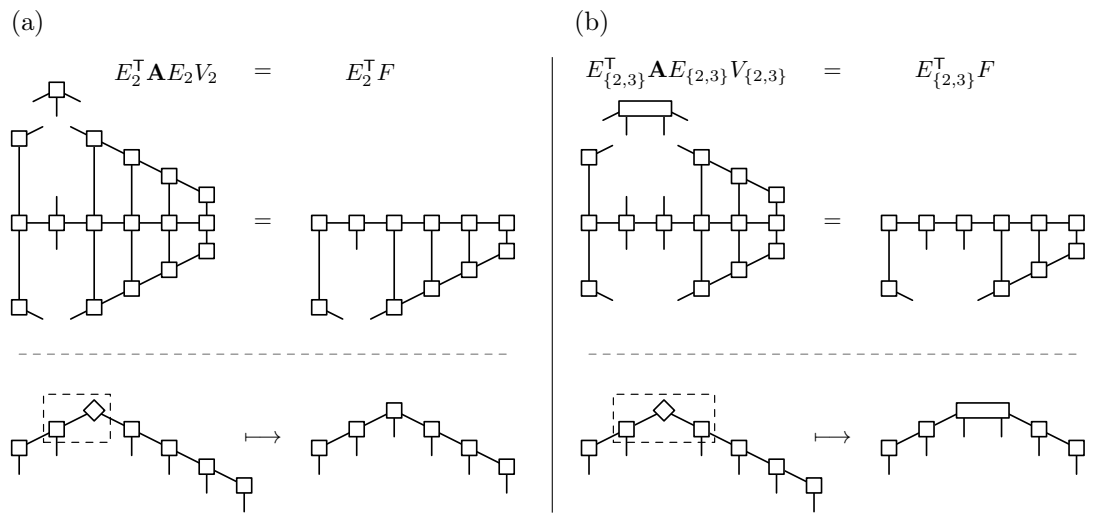

Figure 11: A micro-iteration of (a) the ALS algorithm, and (b) the two-site DMRG. The operator $\mathbf{A}$ is in TT matrix format and the tensor V is left orthogonal in the first and right orthogonal in the third to last component. After the subproblem is solved, an SVD is performed and the orthogonality is shifted to the right.

we get the update

$$V^+ = V - \mathbf{P}_{m,V} \operatorname{grad}_m j(V), \tag{5.20}$$

where $\operatorname{grad}_m j(V)$ is the part of the gradient $\operatorname{grad} j(V)$ in (5.16) that corresponds to the root in the m-th component. It is orthogonal to the other summands and as (5.20) implies, it points into a direction in which the manifold $\mathcal{M}_r$ is linear. Therefore, no retraction is necessary for this update. The manifold is always linear in its root and this is exactly what the ALS algorithms exploits: We shift the root in each step and optimise in the linear direction successively.

Since the summands in the Riemannian gradient are orthogonal, we get

$$\langle \operatorname{grad} j(V), -\operatorname{grad}_m j(V) \rangle = -\| \operatorname{grad}_m j(V)\|^2.$$

While this is always smaller than zero, this is not necessarily true for its accumulation points. This means, the norm of the subgradient $\operatorname{grad}_m j(V)$ might converge to zero before the full gradient does. This would entail that the sequence of search directions is not gradient related. One way of avoiding this is to always optimise the component with the largest residual, thus ensuring

$$\| \operatorname{grad}_m j(V)\| \geq \frac{1}{M} \| \operatorname{grad} j(V)\|,$$

i.e. to proceed in a Gauß-Southwell like fashion [72, 52]. This is often called *pivoted ALS*. While this is possible, it requires the calculation of every subgradient for one substep, which is impractical as this has the same complexity as an entire sweep of the traditional ALS.

Many tests show satisfactory convergence behaviour of the ALS, but one can construct examples in which it converges sublinearly [19]. Using similar techniques, it is possible to show general convergence of the ALS under certain conditions [19, 82].

The Density Matrix Renormalisation Group

The subspace notation suggests that the ALS is closely related to the DMRG algorithm, as we will see in the following. In fact, it is often called the *one-site* DMRG as it can be seen as a simple modification of that algorithm. In comparison, the ALS has the advantage that it optimises the tensor on very small subspaces. On the other hand, the ranks $r = (r_1, \dots, r_{M-1})$ remain fixed and have to be guessed at the beginning. In order to introduce higher ranks, one has to do this in a greedy fashion, e.g., by adding a rank-one approximation of the residual, see [85].

Algorithm 1: One back-and-forth sweep of the ALS algorithm.

input : SPD matrix $\mathbf{A}$, right side F;
start tensor V with TT ranks $r_1, \dots, r_{M-1}$.
output: TT Tensor solution U;

Left orthogonalise V;
for $m \leftarrow 1$ **upto** $M-1$ **do**
  Set
$$V_m \leftarrow (E_m^{\mathsf{T}} \mathbf{A} E_m)^{-1} E_m^{\mathsf{T}} F;$$
  Perform SVD:
$$V_m^{\mathsf{L}} = X_m^{\mathsf{L}} \Sigma_m Y_m;$$
  Shift orthogonality:
$$V_m \leftarrow X_m, \quad V_{m+1}^{\mathsf{R}} \leftarrow \Sigma_m Y_m V_{m+1}^{\mathsf{R}};$$
end
for $m \leftarrow M$ **downto** 2 **do**
  Set
$$V_m \leftarrow (E_m^{\mathsf{T}} \mathbf{A} E_m)^{-1} E_m^{\mathsf{T}} F;$$
  Perform SVD:
$$V_m^{\mathsf{R}} = X_m \Sigma_m Y_m^{\mathsf{R}};$$
  Shift orthogonality:
$$V_m \leftarrow Y_m, \quad V_{m-1}^{\mathsf{L}} \leftarrow V_{m-1}^{\mathsf{L}} X_m \Sigma_m;$$
end
U := V;

The classical two-site DMRG is a clever modification. Here, we minimise over the bigger subspace $\mathcal{L}_m \otimes \mathbb{R}^{q_m \times q_{m+1}} \otimes R_{m+1}$, with the basis representation as in Figure 10(b),

$$V = \sum_{k_{m-1}=1}^{r_{m-1}} \sum_{k_{m+1}=1}^{r_{m+1}} \sum_{\mu_m=1}^{q_m} \sum_{\mu_{m+1}=1}^{q_{m+1}} V_{\{m,m+1\}}(k_{m-1}, \mu_m, \mu_{m+1}, k_{m+1}) \times b_{k_{m-1}}^{(m-1)} \otimes e_{\mu_m}^{(m)} \otimes e_{\mu_{m+1}}^{(m+1)} \otimes \bar{b}_{k_{m+1}}^{(m+2)}.$$

This means we optimise two components at the same time, see Figure 11(b). The insertion operator $E_{\{m,m+1\}}$ is defined analogously to the one-site case. The advantage is that a subsequent SVD after the optimisation step in order to separate the two components yields a new - and possibly higher - rank. The DMRG algorithm is summarised in Algorithm 2. To control the size of these new ranks, a further truncation is often required. Several strategies for *dynamical rank selection* can be implemented by considering the error in different norms [37, 51].

Convergence theory of the DMRG is somewhat different to the ALS and other Riemannian methods because the DMRG does not act only on the manifold or the variety of fixed rank. If we restrict the ranks, we will eventually run into similar problems as with the ALS. However, if the rank is allowed to increase uninhibited, one would expect convergence even to the global minimum that is unique in the full tensor space if, e.g., j is strictly convex. Unfortunately, this is not always the case. We give the following counterexample:

Example 5.12. Consider the TT variety $\mathcal{M}_{\leq(2,2)} \subset \mathbb{R}^{3\times3\times3}$ of TT tensors with rank at most 2 everywhere. We want to approximate the tensor

$$F = \begin{pmatrix}1\\0\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\\0\end{pmatrix} + \begin{pmatrix}0\\1\\0\end{pmatrix} \otimes \begin{pmatrix}0\\1\\0\end{pmatrix} \otimes \begin{pmatrix}0\\1\\0\end{pmatrix}$$

Algorithm 2: One back-and-forth sweep of the DMRG algorithm.

input : SPD matrix $\mathbf{A}$, right side F;
start tensor V, truncation accuracy ϵ_{DMRG}.
output: TT Tensor solution U;

Left orthogonalise V;
for $m \leftarrow 1$ **upto** $M-1$ **do**
  Set
$$V_{\{m,m+1\}} \leftarrow (E_{\{m,m+1\}}^{\mathsf{T}} \mathbf{A} E_{\{m,m+1\}})^{-1} E_{\{m,m+1\}}^{\mathsf{T}} F;$$
  Perform SVD:
$$[V_{\{m,m+1\}}]_{1,2}^{3,4} = X_m^{\mathsf{L}} \Sigma_m Y_m^{\mathsf{R}};$$
  Truncate $\tilde{\Sigma}_m = \text{diag}_{k=1,\dots,r}(\varsigma_{m,k})$ at $\varsigma_{m,k+1} < \epsilon_{\text{DMRG}}$;
  Shift orthogonality:
$$V_m \leftarrow X_m, \quad V_{\{m+1,m+2\}} \leftarrow \tilde{\Sigma}_m \circ Y_m \circ V_{m+2};$$
end
for $m \leftarrow M$ **downto** 2 **do**
  Set
$$V_{\{m-1,m\}} \leftarrow (E_{\{m-1,m\}}^{\mathsf{T}} \mathbf{A} E_{\{m-1,m\}})^{-1} E_{\{m-1,m\}}^{\mathsf{T}} F;$$
  Perform SVD:
$$[V_{\{m-1,m\}}]_{1,2}^{3,4} = X_{m-1}^{\mathsf{L}} \Sigma_{m-1} Y_{m-1}^{\mathsf{R}};$$
  Truncate $\tilde{\Sigma}_{m-1} = \text{diag}_{k=1,\dots,r}(\varsigma_{m-1,k})$ at $\varsigma_{m-1,k+1} < \epsilon_{\text{DMRG}}$;
  Shift orthogonality:
$$V_m \leftarrow Y_m, \quad V_{\{m-2,m-1\}} \leftarrow V_{m-2} \circ X_{m-1} \circ \tilde{\Sigma}_{m-1};$$
end
U := V;

and we start with the elementary tensor $V \in \mathcal{M}_{(1,1)}$

$$V = \begin{pmatrix}1\\0\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\\0\end{pmatrix}.$$

Then, if we update the first two components, we get

$$V_{\{1,2\}}^{+} = E_{\{1,2\}}^{\mathsf{T}} F = \begin{pmatrix}1\\0\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\\0\end{pmatrix} = V_{\{1,2\}},$$

i.e. the tensor does not change. Analogously, this would happen if we started with the last two components. Therefore, even though we allowed ranks at most 2, the DMRG algorithm stagnates at rank 1. This happens, because the DMRG algorithm can only optimise on the variety $\mathcal{M}_{\leq(2,1)}$ or $\mathcal{M}_{\leq(1,2)}$ but not simultaneously on both of them. In these varieties, V is a stationary point and the algorithm stagnates.

Remark 5.13. A problem that is related to the above example is the following: As shown in [73], for the matrix case it holds that if the projection of the gradient onto the tangent cone at a point $V \in \mathcal{M}_{\leq r}$ is zero, then either V has rank r, or it is the global minimum of j in the full space $\mathbb{R}^{q_1 \times q_2}$. This is intuitive, as the rank would have to increase if there is still some information in the gradient. Example 5.12 shows that this is not the case for tensors: A point can be stationary on a TT variety without having full rank or being the global minimum. This has been discussed in [45].

The example relies on the choice of a starting point for the DMRG algorithm. The choice is very unfortunate, as it is orthogonal to one part of the tensor we want to approximate. If we choose a generic starting point, this is (almost) impossible, as the orthogonal complement of a subspace is a null set in the ambient space. We conjecture that this unfortunate choice of a starting point is the only reason that results in a stagnating DMRG algorithm, i.e. the unrestricted DMRG algorithm converges to the global minimum of j for almost every starting point V_0.

6 Technicalities of the Tensor Methods

Applying the above theory on tensor decompositions to the problem of parametric PDEs requires a few technical steps that will be dealt with in this chapter. These are a main contribution of this thesis. In the following we focus on the TT format and we use the ALS algorithm as a solver. We recall from Chapter 4 that the Galerkin solution $u_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ is represented as

$$u_{\text{disc}} = \sum_{i=0}^{N-1} \sum_{\mu \in \Lambda} U(i,\mu)\varphi_i P_\mu \tag{6.1}$$

for a FEM basis $(\varphi_i)_{0\le i\le N-1}$ and general chaos polynomials $(P_\mu)_{\mu\in\Lambda}$. We assume that U is a TT tensor of order $M+1$ with TT ranks $r = (r_1, \dots, r_M)$

$$U(i,\mu) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_M=1}^{r_M} U_0(i,k_1)U_1(k_1,\mu_1,k_2)\cdots U_M(k_M,\mu_M),$$

where we have denoted the first component as U_0 in order to distinguish the deterministic part from the parametric part. Inserting this in (6.1) we get

$$u_{\text{disc}} = \sum_{k_1=1}^{r_1} \cdots \sum_{k_M=1}^{r_M} \left(\sum_{i=0}^{N-1} U_0(i,k_1)\varphi_i\right)\left(\sum_{\mu_1=0}^{q_1-1} U_1(k_1,\mu_1,k_2)P_{\mu_1}\right)\cdots\left(\sum_{\mu_M=0}^{q_M-1} U_M(k_M,\mu_1)P_{\mu_M}\right).$$

The Galerkin solution (2.5) therefore lies in the low rank tensor manifold $\mathcal{M}_r$ and calculations are efficient if the discrete operator $\mathbf{A}$ is in a tensor format as well. This is the case for the affine diffusion equation introduced in Chapter 3, where we have seen that $\mathbf{A}$ is a sum of tensor products of operators

$$\mathbf{A} = \sum_{m=0}^{M} \mathbf{A}_m.$$

For the log-normal diffusion equation, this requires some further work and we will eventually only be able to approximate the operator in the tensor format, as shown in the following.

Furthermore, in order to improve convergence of the ALS algorithm, we employ a preconditioning with the mean of the operator $\mathcal{A}(y)$ in (3.1). This can be done efficiently in the TT format and it yields significant improvement of the speed of convergence.

6.1 Tensor Decomposition of the Log-Normal Operator

We start by deriving a tensor structured approximation of the log-normal diffusion coefficient. The log-normal diffusion coefficient is

$$a(x,y) = \prod_{\ell=1}^{\infty} \exp(b_\ell(x)y_\ell)$$

as in chapter 3.1.3. Our aim is to isolate the dependence on the spatial variable x. First of all, we cut off the infinite product at some maximum length L and we get

$$a(x,y) \approx a_L(x,y) := \prod_{\ell=1}^{L} \exp(b_\ell(x)y_\ell).$$

For fixed $x \in D$, every factor can be expanded in a Hermite basis for the measure $\gamma_{\vartheta\rho}$ since these form a basis of $L^2_{\gamma_{\vartheta\rho}}(\Gamma)$, see Chapter 4. This yields

$$\exp(b_\ell(x)y_\ell) = \sum_{\nu_\ell=1}^{\infty} c_{\nu_\ell}^{(\ell)}(x) H_{\vartheta\rho,\nu_\ell}(y_\ell)$$

and with $\sigma = (\sigma_\ell)_{1\le\ell\le L} = (\exp(\vartheta\rho\alpha_\ell))_{1\le\ell\le L}$ we get

$$c_{\nu_\ell}^{(\ell)}(x) = \frac{(\sigma_\ell b_\ell(x))^{\nu_\ell}}{\sqrt{\nu_\ell!}} \exp\big((\sigma_\ell b_\ell(x))^2/2\big), \tag{6.2}$$

see [59]. The convergence of this series in $L^2_{\gamma_{\vartheta\rho}}(\Gamma)$ is point-wise in x but one can clearly see that the coefficient functions are continuous. We approximate this by cutting off this series for every ℓ at a finite d_ℓ, obtaining a finite discretisation in y_ℓ:

$$\exp(b_\ell(x)y_\ell) \approx \sum_{\nu_\ell=0}^{d_\ell-1} c_{\nu_\ell}^{(\ell)}(x) H_{\nu_\ell}(y_\ell).$$

In accordance with chapter 4.1.2, we define the full tensor set

$$\Delta = \{(\nu_1,\dots,\nu_L,0,\dots) \in \mathcal{F} : 0 \le \nu_\ell < d_\ell \quad \forall 0 \le \ell \le L\}$$

The dependence on x can then be isolated by defining a core function

$$c_{\nu_1,\dots,\nu_L}(x) := c_{\nu_1}^{(1)}(x)\cdots c_{\nu_L}^{(L)}(x) \tag{6.3}$$

with dependence only on the spatial variable x, which yields the further approximation of the coefficient

$$a(x,y) \approx a_\Delta(x,y) = \sum_{\nu\in\Delta} c_{\nu_1,\dots,\nu_L}(x) H_\nu(y).$$

Note that as in the definition of the tensor set Λ in Chapter 4, the cut-off at L is included in the definition of the set Δ.

This is akin to a Tucker decomposition for discrete tensors as in chapter 5.4.2. The main disadvantage of this approach is that the core function now suffers from the *curse of dimensionality*, i.e. it grows exponentially with the tensor order L. Even with the knowledge of all coefficient functions $c_{\nu_\ell}^{(\ell)}$, this is infeasible. Furthermore, in general, the right-hand side in (6.3) cannot be collapsed into a simple equation and usually all coefficient functions need to be remembered.

Therefore, we will employ a Tensor Train (TT) decomposition of the Tucker core. The choice of this tensor format is justified by the fact that the coefficient functions b_ℓ are of decreasing norm and the components corresponding to these smaller coefficients are less important. Thus, we expect the approximation to become only slightly less accurate with moderate TT ranks and the error to remain manageable. Further down, we will briefly address the localised problem, where each component is of the same importance and a TT decomposition would therefore be inappropriate.

A straightforward approach to attain the decomposition would be to discretise the spatial variable x in some points $x_p, p = 1,\dots,P$. For each coefficient function, this yields a discrete matrix

$$G_\ell(p,\nu_\ell) := c_{\nu_\ell}^{(\ell)}(x_p).$$

Starting at $\ell = L$, we decompose this matrix using a singular value decomposition or a QR decomposition:

$$G_L(p,\nu_L) = \sum_{k_L=1}^{s_L} Q_L(p,k_L) A_L(k_L,\nu_L),$$

where A_L is orthogonal (for example by shifting the singular values into Q_L or by a transposed QR decomposition). The dependence on the discrete p can then be shifted to the left to create a new matrix

$$\tilde{G}_{L-1}(p,\nu_{L-1},k_L) := G_{L-1}(p,\nu_{L-1}) Q_L(p,k_L)$$

and this can again be decomposed with a QR decomposition between p and the multi-index (ν_{L-1},k_L)

$$\tilde{G}_{L-1}(p,\mu_{L-1},k_m) = \sum_{k_{L-1}=1}^{s_{L-1}} Q_{L-1}(p,k_{L-1}) A_{L-1}(k_{L-1},\nu_{L-1},k_L).$$

This is done successively, always shifting the component Q_ℓ, until we obtain the last component $A_0(p, k_1) := Q_1(p, k_1)$ that now encodes all dependence on the discretised spatial variable x_p:

$$a(x_p, y) \approx a_{\Delta,s}(x_p, y) = \sum_{\nu\in\Delta} \left(\sum_{k_1=1}^{s_1} \cdots \sum_{k_L=1}^{s_L} A_0(p, k_1) A_1(k_1, \nu_1, k_2) \cdots A_L(k_L, \nu_L) \right) H_\nu(y). \tag{6.4}$$

It holds that $a_\Delta(x_p) = a_{\Delta,s}(x_p)$ at all points x_p only if we allowed the ranks $s = (s_1, \ldots, s_L)$ to be full. This will usually be infeasible, as in general the ranks would grow exponentially with L. Therefore, when s is restricted, we see $a_{\Delta,s}$ as another approximation of the coefficient. The accuracy of this approximation can be evaluated by a Monte Carlo (MC) sampling procedure for some samples $y_{[1]}, \ldots, y_{[P_{\mathrm{MC}}]}$ and at all points $x_p, p = 1, \ldots, P$.

This straightforward TT decomposition has the obvious disadvantage that it can only be evaluated at predefined points x_p and if we want to obtain a value at any other point, we would have to either interpolate, which could be very inaccurate, or calculate the entire decomposition again for a new set of points, which would be computationally expensive. Additionally, for fixed ranks s_ℓ, a finer discretisation of D, i.e. in more points x_p, could actually *decrease* the accuracy in each point, because higher ranks would be needed for larger sets of points.

We therefore explore a different approach. We aim at preserving the continuity of the spatial variable x. This can be done by defining some kind of reduced basis as in chapter 5.4.1. Here, we first use a correlation matrix $C_L \in \mathbb{R}^{d_L \times d_L}$ on the right-most side

$$C_L(\nu_L, \nu'_L) := \int_D c^{(L)}_{\nu_L}(x) c^{(L)}_{\nu'_L}(x) \,\mathrm{d}x.$$

This matrix is obviously symmetric, i.e. we can perform an eigenvalue decomposition

$$C_L(\nu_L, \nu'_L) = \sum_{k_L=1}^{s_L} \varsigma^2_{k_L} A_L(k_L, \nu_L) A_L(k_L, \nu'_L)$$

and define basis functions

$$\tilde{c}^{(L)}_{k_L}(x) := \sum_{\nu_L=0}^{d_L-1} A_L(k_L, \nu_L) c^{(L)}_{\nu_L}(x), \quad k_L = 1, \ldots, s_L.$$

If s_L is not the full rank, this will yield a reduced basis set. As before, these functions can be grouped together with the coefficient functions to the left to form the next correlation matrix $C_{L-1} \in \mathbb{R}^{d_{L-1} \times s_L \times d_{L-1} \times s_L}$:

$$C_{L-1}(\nu_{L-1}, k_L, \nu'_{L-1}, k'_L) := \int_D c^{(L-1)}_{\nu_{L-1}}(x) \tilde{c}^{(L)}_{k_L}(x) c^{(L-1)}_{\nu'_{L-1}}(x) \tilde{c}^{(L)}_{k'_L}(x) \,\mathrm{d}x.$$

We perform another eigenvalue decomposition

$$C_{L-1}(\nu_{L-1}, k_L, \nu'_{L-1}, k'_L) = \sum_{k_{L-1}=1}^{s_{L-1}} \varsigma^2_{k_{L-1}} A_{L-1}(k_{L-1}, \nu_{L-1}, k_L) A_{L-1}(k_{L-1}, \nu'_{L-1}, k'_L)$$

and obtain reduced basis functions

$$\tilde{c}^{(L-1)}_{k_{L-1}}(x) := \sum_{\nu_{L-1}=0}^{d_{L-1}-1} \sum_{k_L=1}^{s_L} A_{L-1}(k_{L-1}, \nu_{L-1}, k_L) c^{(L-1)}_{\nu_{L-1}}(x) \tilde{c}^{(L)}_{k_L}(x)$$

for $k_{L-1} = 1, \ldots, s_{L-1}$.

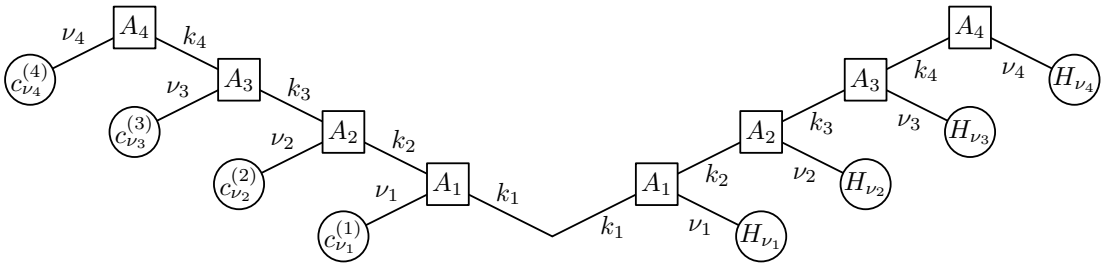

Figure 12: A coefficient splitting for $L = 4$.

This will be done successively until we obtain a final reduced basis

$$\begin{aligned}
\tilde{c}^{(1)}_{k_1}(x) &= \sum_{\nu_1=0}^{d_1-1} \sum_{k_2=1}^{s_2} A_1(k_1,\nu_1,k_2) c^{(1)}_{\nu_1}(x) \tilde{c}^{(2)}_{k_2}(x) \\
&= \sum_{\nu_1=0}^{d_1-1} \sum_{\nu_2=0}^{d_2-1} \sum_{k_2=1}^{s_2} \sum_{k_3=1}^{s_3} A_1(k_1,\nu_1,k_2) A_2(k_2,\nu_2,k_3) c^{(1)}_{\nu_1}(x) c^{(2)}_{\nu_2}(x) \tilde{c}^{(3)}_{k_3}(x) \\
&= \dots \\
&= \sum_{\nu\in\Delta} \sum_{k_2=1}^{s_2} \cdots \sum_{k_L=1}^{s_L} A_1(k_1,\nu_1,k_2) \cdots A_L(k_L,\nu_L) c^{(1)}_{\nu_1}(x) \cdots c^{(L)}_{\nu_L}(x).
\end{aligned} \tag{6.5}$$

This is basically a TT decomposition of the core function $c_{\nu_1,\dots,\nu_L}$ used above.

The final reduced basis functions are denoted by

$$a_0[k_1](x) := \tilde{c}^{(1)}_{k_1}(x) \quad \forall k_1 = 1,\dots,s_1$$

as they are still continuous in x. They can be evaluated at any point x by evaluating the known coefficient functions $c^{(\ell)}_{\nu_\ell}$ from (6.2) at that point and collapsing the TT expansion (6.5). This is done via the recursive formula

$$\begin{aligned}
\tilde{c}^{(L)}_{k_L}(x) &= \sum_{\nu_L=0}^{d_L-1} A_L(k_L,\nu_L) c^{(L)}_{\nu_L}(x), \\
\tilde{c}^{(\ell)}_{k_\ell}(x) &:= \sum_{\nu_\ell=0}^{d_\ell-1} \sum_{k_{\ell+1}=1}^{s_{\ell+1}} A_\ell(k_\ell,\nu_\ell,k_{\ell+1}) c^{(\ell)}_{\nu_\ell}(x) \tilde{c}^{(\ell+1)}_{k_{\ell+1}}(x) \quad \forall \ell = 1,\dots,L-1.
\end{aligned}$$

As a result, the coefficient is approximated by a TT tensor that is continuous in the first component

$$a_{\Delta,s}(x,y) = \sum_{\nu\in\Delta} \left(\sum_{k_1=1}^{s_1} \cdots \sum_{k_L=1}^{s_L} a_0[k_1](x) A_1(k_1,\nu_1,k_2) \cdots A_L(k_L,\nu_L) \right) H_\nu(y). \tag{6.6}$$

Evaluating this coefficient at all grid points $x_p, p = 1,\dots,P$ yields (6.4) with the first component

$$A_0(p,k_1) := \tilde{c}^{(1)}_{k_1}(x_p).$$

This means that the TT components $A_1,\dots,A_L$ yield the coefficients both for the Hermite decomposition and the continuous deterministic functions $a_0[k_1]$, see Figure 12.

The attentive reader will have noticed that for this approach, it is necessary to calculate the quadruple integral

$$C_\ell(\nu_\ell,k_{\ell+1},\nu'_\ell,k'_{\ell+1}) := \int_D c^{(\ell)}_{\nu_\ell}(x) \tilde{c}^{(\ell+1)}_{k_{\ell+1}}(x) c^{(\ell)}_{\nu'_\ell}(x) \tilde{c}^{(\ell+1)}_{k'_{\ell+1}}(x)\,\mathrm{d}x, \tag{6.7}$$

and since the reduced basis functions $\tilde{c}^{(\ell)}_{k_\ell}$ are only given as a linear combination of the coefficient functions $c^{(\ell)}_{\nu_\ell}$, they cannot be saved explicitly. This means, the reduced functions have to be "unfolded" in order to calculate the quadruple integral in each step. For this, we would need to eventually calculate the full integral in

$$C_1(\nu_1,k_2,\nu_1',k_2') := \sum_{\nu_2,\nu_2'=0}^{d_2-1} \cdots \sum_{\nu_M,\nu_M'=0}^{d_M-1} \sum_{k_2,k_2'=1}^{s_2} \cdots \sum_{k_L,k_L'=1}^{s_L} A_2(k_2,\nu_2,k_3)\cdots A_L(k_L,\nu_L) \\ \times A_2(k_2,\nu_2',k_3)\cdots A_L(k_L,\nu_L') \int_D c^{(1)}_{\nu_1}(x)\cdots c^{(L)}_{\nu_L}(x) c^{(1)}_{\nu_1'}(x)\cdots c^{(L)}_{\nu_L'}(x)\,\mathrm{d}x,$$

which is clearly infeasible as it suffers from the curse of dimensionality. However, since we are calculating numerical integrals, these will follow a certain quadrature rule and thus the full integral is only an approximation either way. This means that for (6.7), we use a quadrature at points $\chi_q, q = 1,\dots,P_{\mathrm{quad}}$ with weights ω_q

$$C_\ell(\nu_\ell,k_{\ell+1},\nu_\ell',k_{\ell+1}') \approx \sum_{q=1}^{Q} c^{(\ell)}_{\nu_\ell}(\chi_q)\tilde{c}^{(\ell+1)}_{k_{\ell+1}}(\chi_q) c^{(\ell)}_{\nu_\ell'}(\chi_q)\tilde{c}^{(\ell+1)}_{k_{\ell+1}'}(\chi_q)\omega_q. \tag{6.8}$$

If we save the reduced functions only at the quadrature points χ_q, it is possible to store them explicitly as vectors of the quadrature points

$$\tilde{c}^{(\ell)}_{k_\ell}(\chi_q) := \sum_{\nu_\ell=0}^{d_\ell-1} \sum_{k_{\ell+1}=1}^{s_{\ell+1}} A_\ell(k_\ell,\nu_\ell,k_{\ell+1}) c^{(\ell)}_{\nu_\ell}(\chi_q)\tilde{c}^{(\ell+1)}_{k_{\ell+1}}(\chi_q)$$

and we can calculate all integrals (6.7) explicitly without the curse of dimensionality as in (6.8). This will be an approximation, but the approximation follows from the fact that we integrate numerically and not from some predetermined discretisation of the functions. The procedure is summarised in Algorithm 3.

It is interesting to note that if we use a simple sum over the grid points $x_p, p = 1,\dots,P$ as quadrature, i.e. $P = P_{\mathrm{quad}}, x_p = \chi_p, \omega_p = \frac{1}{P}$ for all $p = 1,\dots,P$, we obtain the same TT decomposition as in the straightforward approach presented at the beginning:

$$C_L(\nu_L,\nu_L') := \int_D c^{(L)}_{\nu_L}(x) c^{(L)}_{\nu_L'}(x)\,\mathrm{d}x \approx \frac{1}{P}\sum_{p=1}^{P} G_L(p,\nu_L) G_L(p,\nu_L').$$

This results in the same components A_L, as the weight $\omega_p = \frac{1}{P}$ will only influence the Eigenvalues $\varsigma^2_{k_L}$ that play no role here. The difference is that in our approach, the function can be evaluated at any point $x \in D$! However, we propose using a more accurate quadrature that should be good enough for solving the integrals, and not on some grid that is motivated by the application. This approximation needs to be done only once with high enough accuracy.

Remark 6.1. A similar approach to decomposing the coefficient has been chosen in [20]. Here, the knowledge of the eigenfunctions of the covariance operator has been assumed a priori. This means that one has an orthogonal basis also for the deterministic part in x and all that remains to do is to decompose the coefficient tensor for this basis representation. This is also done using some quadrature and the error of this approximation can be estimated.

In the following, we fix Δ and the ranks s and we set $a_{\mathrm{disc}} := a_{\Delta,s}$. Then we define the operators $\mathcal{A}_+ : \mathcal{V}_{\vartheta\rho} \to \mathcal{V}^*_{\vartheta\rho}$ and $\mathcal{A}_- : \mathcal{V}_{\vartheta\rho} \to \mathcal{V}^*_{\vartheta\rho}$ by

$$\begin{aligned} \mathcal{A}(v) &= \mathcal{A}_+(v) + \mathcal{A}_-(v), \\ \mathcal{A}_+(v) &:= -\nabla\cdot(a_{\mathrm{disc}}\nabla v), \\ \mathcal{A}_-(v) &:= -\nabla\cdot\Big((a - a_{\mathrm{disc}})\nabla v\Big) \end{aligned} \tag{6.9}$$

Algorithm 3: Algorithm for coefficient splitting.

input : Coefficient functions $c^{(1)}_{\nu_1}, \dots, c^{(L)}_{\nu_L}, \quad \nu_\ell = 0, \dots, d_\ell - 1; \quad \ell = 1, \dots, L;$
ranks $s_1, \dots, s_L$;
quadrature rule $(\chi_q, \omega_q), q = 1, \dots, P_{\text{quad}}$.
output: TT Tensor components $A_1, \dots, A_L$;

Set $s_{L+1} = 1, \tilde{c}^{(L+1)}_1 \equiv 1$;
for $\ell \leftarrow L$ **downto** 1 **do**
Arrange correlation matrix C_ℓ:

$$C_\ell(\nu_\ell, k_{\ell+1}, \nu'_\ell, k'_{\ell+1}) := \sum_{q=1}^{P_{\text{quad}}} c^{(\ell)}_{\nu_\ell}(\chi_q)\tilde{c}^{(\ell+1)}_{k_{\ell+1}}(\chi_q)c^{(\ell)}_{\nu'_\ell}(\chi_q)\tilde{c}^{(\ell+1)}_{k'_{\ell+1}}(\chi_q)\omega_q;$$

Compute eigenvalue decomposition:

$$C_\ell(\nu_\ell, k_{\ell+1}, \nu'_\ell, k'_{\ell+1}) = \sum_{k_\ell=1}^{s_\ell} \varsigma^2_{k_\ell} A_\ell(k_\ell, \nu_\ell, k_{\ell+1})A_\ell(k_\ell, \nu'_\ell, k'_{\ell+1});$$

Set reduced basis functions:

$$\tilde{c}^\ell_{k_\ell}(\chi_q) = \sum_{\nu_\ell=0}^{d_\ell-1} \sum_{k_{\ell+1}=1}^{s_{\ell+1}} A_\ell(k_\ell, \nu_\ell, k_{\ell+1})c^\ell_{\nu_\ell}(\chi_q)\tilde{c}^{\ell+1}_{k_{\ell+1}}(\chi_q);$$

end

for $v \in \mathcal{V}_{\vartheta\rho}$. We will assume that the error $a - a_{\text{disc}}$ is small enough in the L^∞-sense, such that the bounds in Lemma 3.7 still hold with weaker constants. This means that the variational problem corresponding to the discrete coefficient a_{disc}, i.e.

$$\int_\Gamma \int_D a_{\text{disc}}(x, y)\nabla u(x, y) \cdot \nabla v(x, y) \,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y) = \int_\Gamma \int_D f(x)v(x, y) \,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y), \tag{6.10}$$

is well-posed for all $v, w \in \mathcal{V}_{\vartheta\rho}$ and the bilinear form

$$B_+(u, v) := \int_\Gamma \int_D a_{\text{disc}}(x, y)\nabla u(x, y) \cdot \nabla v(x, y) \,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y) \tag{6.11}$$

is uniformly bounded and coercive at least in the sense of Lemma 3.10. Then the above solution theory of Chapter 3 remains valid. However, this cannot be guaranteed a priori. Neither the approximation in the Hermite expansion nor the low rank reduced basis approximation yield sufficient estimators. Therefore, the fact that the L^∞-error of $a - a_{\text{disc}}$ is small has to remain an assumption.

TT Matrix Galerkin Operator

The above considerations yield that the approximative operator $\mathcal{A}_+$ has the following tensor product structure

$$\mathcal{A}_+ = \sum_{k_1=1}^{s_1} \cdots \sum_{k_\ell=1}^{s_L} \mathcal{A}_0[k_1] \otimes \mathcal{A}_1[k_1, k_2] \otimes \cdots \otimes \mathcal{A}_L[k_L]$$

with the operator components

$$\mathcal{A}_0[k_1] : \mathcal{X} \to \mathcal{X}^*$$
$$\mathcal{A}_0[k_1](v_x) = -\nabla \cdot (a_0[k_1]\nabla v_x)$$

and for all $\ell = 1, \dots, L$:

$$\mathcal{A}_\ell[k_\ell, k_{\ell+1}] : L^2_{\gamma_{\vartheta\rho,m}}(\mathbb{R}) \to L^2_{\gamma_{\vartheta\rho,m}}(\mathbb{R})$$
$$\mathcal{A}_\ell[k_\ell, k_{\ell+1}](v_y) = \sum_{\nu_\ell=0}^{d_\ell-1} A_\ell(k_\ell, \nu_\ell, k_{\ell+1}) H_{\vartheta\rho,\nu_\ell} v_y.$$

With the preceding coefficient approximation, the variational formulation of the lognormal diffusion problem is given for all $v \in \mathcal{V}_{\text{disc}}$ by (6.10) and the Galerkin solution $u_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ has the form

$$u_{\text{disc}}(x, y) = \sum_{i=0}^{N-1} \sum_{\mu\in\Lambda} U(i, \mu)\varphi_i(x) H_{\vartheta\rho,\mu}(y)$$

and is determined by the linear system

$$\mathbf{A}U = F.$$

The Galerkin operator $\mathbf{A}$ is a tensor structured approximation of the differential operator as in chapter 4.2. It is given by

$$\begin{aligned}
\mathbf{A}(i, \mu, i', \mu') &= \int_\Gamma \int_D a_{\text{disc}}(x, y)\nabla\varphi_i(x) H_{\vartheta\rho,\mu}(y) \cdot \nabla\varphi_{i'}(x) H_{\vartheta\rho,\mu'}(x)\,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y) \\
&= \sum_{k_1=1}^{s_1} \cdots \sum_{k_L=1}^{s_L} \int_D a_0[k_1](x)\nabla\varphi_i(x) \cdot \nabla\varphi_{i'}(x)\,\mathrm{d}x \\
&\quad \times \prod_{m=1}^{M} \sum_{\nu_m=0}^{d_m-1} A_m(k_m, \nu_m, k_{m+1}) \int_\Gamma H_{\vartheta\rho,\nu_m}(y_m) H_{\vartheta\rho,\mu_m}(y_m) H_{\vartheta\rho,\mu'_m}(y_m)\,\mathrm{d}\gamma_{\vartheta\rho}(y_m) \\
&\quad \times \prod_{\ell=M+1}^{L} \sum_{\nu_\ell=0}^{d_\ell-1} A_\ell(k_\ell, \nu_\ell, k_{\ell+1}) \int_\Gamma H_{\vartheta\rho,\nu_\ell}(y_\ell)\,\mathrm{d}\gamma_{\theta\rho}(y_\ell) \\
&= \sum_{k_1=1}^{s_1} \cdots \sum_{k_M=1}^{s_M} \mathbf{A}_0(i, i', k_1)\mathbf{A}_1(k_1, \mu_1, \mu'_1, k_2) \cdots \mathbf{A}_M(k_M, \mu_M, \mu'_M).
\end{aligned}$$

Here,

$$\mathbf{A}_0(i, i', k_1) := \int_D a_0[k_1](x)\nabla\varphi_i(x) \cdot \nabla\varphi_{i'}(x)\,\mathrm{d}x$$

is obtained by quadrature on $\mathcal{T}$ and for $0 < m \le M \le L$,

$$\begin{aligned}
\mathbf{A}_m(k_m, \mu_m, \mu'_m, k_{m+1}) &:= \sum_{\nu_m=0}^{d_m-1} A_m(k_m, \nu_m, k_{m+1}) \\
&\quad \times \int_\Gamma H_{\vartheta\rho,\nu_m}(y_m) H_{\vartheta\rho,\mu_m}(y_m) H_{\vartheta\rho,\mu'_m}(y_m)\,\mathrm{d}\gamma_{\vartheta\rho}(y_m) \\
&= \sum_{\nu_m=0}^{d_m-1} A_m(k_m, \nu_m, k_{m+1})\kappa_{\mu_m,\mu'_m,\nu_m}
\end{aligned} \tag{6.12}$$

can be calculated efficiently with (4.15) in Proposition 4.4. If $M < L$, the last component is given by

$$\begin{aligned}
\mathbf{A}_M(k_M, \mu_M, \mu'_M) &= \sum_{k_{M+1}=1}^{s_{M+1}} \cdots \sum_{k_L=1}^{s_L} \sum_{\nu_M=0}^{d_M-1} A_M(k_M, \nu_M, k_{M+1})\kappa_{\mu_M\mu'_M,\nu_M} \\
&\quad \times \prod_{\ell=M+1}^{L} \sum_{\nu_\ell=0}^{d_\ell-1} A_\ell(k_\ell, 0, k_{\ell+1}).
\end{aligned}$$

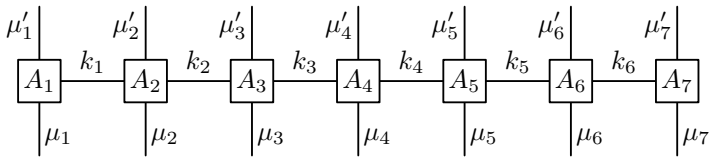

Figure 13: A TT matrix operator for $M = 7$.

By $L^2_{\gamma_{\vartheta_\rho}}$-orthonormality of the Hermite polynomials only matrix-vector multiplications have to be carried out for this. This means that $\mathbf{A}$ is a TT matrix operator [65], see Figure 13 for an illustration.

For fixed polynomial degree of the solution $q_m, m = 1, \dots, M$, the integral of the triple product in (6.12) is equal to zero for all $\nu_m \geq 2q_m - 1$ and therefore the coefficient can be truncated at $d_m = 2q_m - 2, m = 1, \dots, M$, without loss of accuracy on the discretised space. This means, for full ranks s of the coefficient a_{disc}, the operator is actually exact on the Galerkin space $\mathcal{V}_{\text{disc}}$, allowing for a cutoff at L. Full ranks are however infeasible for large M and therefore assumption of a small error $a - a_{\text{disc}}$ is necessary to ensure that the Galerkin operator $\mathbf{A}$ is positive definite. However, this can be tested a priori.

Remark 6.2. We remember the problem of localised uncertainty in Chapter 3.1.4. Here, the deterministic parts of the coefficient are of equal importance and an approximation of that operator in the TT format seems inappropriate. As stated above, the splitting of the coefficient can be seen as a decomposition of the Tucker core that depends on x. In the localised case, it could therefore be possible to decompose this core using a Hierarchical Tucker decomposition, see Chapter 5.4.5.

For this, one could perform the above algorithm in a similar fashion but by pairing the components rather than approximating them successively. The algorithm is one-to-one to a regular HT decomposition algorithm. This remark should justify the use of HT decompositions even for parametric problems like the ones in this thesis. However, since we do have decaying coefficients, we adhere to the TT format.

6.2 Preconditioning

The ALS algorithm described in Chapter 5.5.2 is used to minimise the functional

$$j(V) = \frac{1}{2}\langle \mathbf{A}V, V\rangle - \langle F, V\rangle.$$

on the fixed rank manifold $\mathcal{M}_r$. Preconditioning is key to the convergence behaviour of the solver. The preconditioner can be chosen as an approximation of the Hessian of the functional j. In the case of Galerkin approximation and the resulting linear system, the Hessian is exactly the Galerkin operator $\mathbf{A}$. Therefore the preconditioner should be an approximation of the inverse operator $\mathbf{A}^{-1}$. Often, second order information is not available or hard to compute and therefore another choice has to be made. In the following, we describe our choice of preconditioner for the affine and the log-normal case as well as an efficient method to include it in the ALS algorithm that we will use for minimising the functional.

The Affine Case

A well known choice of preconditioner in the affine case is the mean value of the coefficient such that $\mathbf{A}_0 = A_0 \otimes I_{q_1} \otimes \cdots \otimes I_{q_M}$ as for instance discussed in [81]. We denote the unique Cholesky decomposition of A_0^{-1} with C and

$$\mathbf{C} := C \otimes I_{q_1} \otimes \cdots \otimes I_{q_M}.$$

Thus, it holds

$$\mathbf{A}_0^{-1} = A_0^{-1} \otimes I_{q_1} \otimes \cdots \otimes I_{q_M} = CC^{\mathsf{T}} \otimes I_{q_1} \otimes \cdots \otimes I_{q_M} = \mathbf{C}\mathbf{C}^{\dagger}.$$

This is done only for notational purposes, as we will see in the following that an explicit computation of the inverse A_0^{-1} as well as its Cholesky decomposition is not necessary.

With this preconditioner, for each m, we are minimising

$$j(V_0, \dots, W_m, \dots, V_M) = \frac{1}{2}\langle E_m^\dagger \mathbf{C}\mathbf{A}\mathbf{C}^\dagger E_m W_m, W_m\rangle - \langle E_m^\dagger \mathbf{C} F, W_m\rangle. \tag{6.13}$$

For $m = 1, \dots, M$, this can be incorporated into the insertion operators,

$$\begin{aligned} \tilde{V}_0 &= C^\mathsf{T} V_0, \\ \tilde{E}_m &= \mathbf{C}^\dagger E_m \end{aligned} \tag{6.14}$$

and we can rewrite (6.13) as

$$j(V_0, \dots, W_m, \dots, V_M) = \frac{1}{2}\langle \tilde{E}_m^\dagger \mathbf{A} \tilde{E}_m W_m, W_m\rangle - \langle \tilde{E}_m^\dagger F, W_m\rangle.$$

We can circumvent the calculation of C by exploiting that for left orthogonal V_0, (6.14) is equivalent to

$$(\tilde{V}_0^\mathsf{L})^\mathsf{T} A_0 \tilde{V}_0^\mathsf{L} = I_{r_1}.$$

This means, if one requires the first component to be orthogonal with respect to the A_0-inner product instead of the Euclidean inner product, the preconditioner is automatically included in the insertion operator $\tilde{E}_m$ for all $m = 1, \dots, M$. In the reduced basis notation of Chapter 5.4.4, this yields orthogonality of the first basis functions $b_1^{(0)}, \dots, b_{r_1}^{(0)}$ in the A_0-inner product, i.e.

$$\langle b_{k_1}^{(0)}, A_0 b_{k_1'}^{(0)}\rangle = \delta_{k_1 k_1'}.$$

In other words, $b_{k_1}^{(0)}$ defined as the reduced basis function

$$b_{k_1}^{(0)} = \sum_{i=0}^{N-1} \tilde{V}_0(i, k_1) e_i^{(0)}.$$

This orthogonalisation can be carried out in a similar fashion as in (5.12) and does not require the explicit calculation of C.

As a consequence, by ensuring orthogonality of the first component with respect to the mean inner product on $\mathcal{X}$, we can incorporate the preconditioner effectively into the ALS-step. An explicit preconditioning only has to be carried out for the first subproblem, where it is part of an efficient inner CG subroutine.

The Log-Normal Case

In the log-normal case, we proceed similarly. However, the mean of the operator $\mathcal{A}$ is not readily available. This requires the calculation of the mean of the log-normal diffusion coefficient

$$\begin{aligned} \mathbb{E}(a(x,\cdot)) &= \int_\Gamma a(x,y)\,\mathrm{d}\gamma_{\vartheta\rho}(y) \\ &= \prod_{\ell=1}^{L} c_0^{(\ell)}(x) \end{aligned}$$

Since we use the approximate coefficient (6.6) in the linear system, we calculate the mean of $a_{\Delta,s}$ instead [17]. It holds

$$\begin{aligned} \mathbb{E}(a_{\mathrm{disc}}(x,\cdot)) &= \int_\Gamma y a_{\Delta,s}(x,y)\,\mathrm{d}\gamma_{\vartheta\rho}(y) \\ &= \sum_{k_1=1}^{s_1} \cdots \sum_{k_L=1}^{s_L} a_0[k_1](x) \sum_{\nu\in\Delta} A_1(k_1,\nu_1,k_2)\cdots A_L(k_L,\nu_L) \int_\Gamma H_\nu(y)\,\mathrm{d}\gamma_{\vartheta\rho}(y) \\ &= \sum_{k_1=1}^{s_1} \cdots \sum_{k_L=1}^{s_L} a_0[k_1](x) A_1(k_1,0,k_2)\cdots A_L(k_L,0), \end{aligned}$$

since

$$\int_{-\infty}^{\infty} H_{\nu_\ell}(y_\ell)\,\mathrm{d}\gamma_{\vartheta\rho}(y_\ell) = \delta_{\nu_\ell 0}$$

for all $\ell = 1, \ldots, L$. This is easy to compute in TT format as it consists only of L matrix multiplications. After a FEM-discretisation, we obtain the preconditioner

$$\bar{\mathbf{A}} = \bar{A} \otimes I_{q_1} \otimes \cdots \otimes I_{q_M}$$

with

$$\bar{A}(i,j) = \int_D \mathbb{E}(a_{\mathrm{disc}}(x,\cdot))\nabla\varphi_i(x) \cdot \nabla\varphi_j(x)\,\mathrm{d}x.$$

This can be applied in the same way as in the affine case.

7 Error Estimation and Adaptivity

After we have found an approximate solution $u_{\text{disc}} \in \mathcal{V}_{\text{disc}}$, it is desirable to estimate the error that we have made. This can be done in many different ways. First of all, one has to distinguish between *a priori* and *a posteriori* error estimation. Put simply, the former is the abstract *prediction* of the error that results from the discretisation space that we used while the latter is the approximate estimation of the distance between the approximate solution w_{disc} and the exact solution $u \in \mathcal{V}$ in a desired norm *after* w_{disc} has been computed. In both cases, a difficulty is that the exact solution is not accessible and an exact calculation of the error is therefore impossible. For this reason, it is often advantageous to compare the computed solution to a solution that has been found using a different scheme where the error is known or easily predictable. For parametric problems, a commonly used method is to estimate the expected value of the error using a Monte-Carlo sampling.

Once the error is estimated, we can decide on this basis if we are content with the result or if we need to further refine the solution. This means that either the solver needs to reach a higher accuracy or - most commonly - the Galerkin space needs to be enlarged. In our case, the dimension of the Galerkin space and therefore the accuracy of the approximation depends on several factors: the degree of the polynomials and the refinement of the mesh in the deterministic domain; the size of the index space in the parametric domain; and ultimately the rank of the coefficient tensor. It is possible to decompose the estimated error in order to estimate the influence of each of these factors on the accuracy of the solution. Based on this, it is then reasonable to refine the respective part, keeping the errors equilibrated.

In this thesis, all error estimation is done a posteriori because this allows us to refine the different aspects of the approximation independently. We will first present a way to estimate the error both for the affine and the log-normal diffusion equation before we develop a broad adaptive scheme as briefly discussed in Chapter 4.

7.1 A Posteriori Error Estimation

A posteriori error estimation for the parametric diffusion equation with different diffusion coefficients is based on - but also varies from - regular a posteriori estimators of elliptic second order problems. An important point is that we have to distinguish the deterministic error from the parametric error and they can be and have to be estimated in a different way.

To begin, we define the *residual* of any discrete function $w_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ as

$$\mathcal{R}(w_{\text{disc}}) = f - \mathcal{A}(w_{\text{disc}}) \in \mathcal{V}^*.$$

We utilise the fact that

$$\|v\|_{\mathcal{V}} \lesssim \|v\|_{\mathcal{A}} \tag{7.1}$$

for all $v \in \mathcal{V}$, as shown in Lemma 3.2 and Proposition 3.11. Then we know that the error satisfies

$$\text{err}(w_{\text{disc}}) := \|u - w_{\text{disc}}\|_{\mathcal{V}} \lesssim \|u - w_{\text{disc}}\|_{\mathcal{A}}$$

for the true solution $u \in \mathcal{V}$ of

$$(u, v)_{\mathcal{A}} = \langle f, v \rangle \quad \forall v \in \mathcal{V}. \tag{7.2}$$

We can state the following estimate similar to [15]:

Theorem 7.1. *Let $u_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ be the Galerkin solution of (7.2). Then for any $w_{\text{disc}} \in \mathcal{V}_{\text{disc}}$ and any bounded, linear map $\mathcal{Q} : \mathcal{V} \to \mathcal{V}_{\text{disc}}$, it holds*

$$\text{err}(w_{\text{disc}})^2 \lesssim \left(\sup_{v \in \mathcal{V}} \frac{|\langle \mathcal{R}(w_{\text{disc}}), v - \mathcal{Q}v \rangle|}{\|v\|_{\mathcal{V}}} + c_{\mathcal{Q}} \|w_{\text{disc}} - u_{\text{disc}}\|_{\mathcal{A}} \right)^2 + \|w_{\text{disc}} - u_{\text{disc}}\|_{\mathcal{A}}^2.$$

Proof. The Galerkin solution $u_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ satisfies the Galerkin orthogonality as in (2.6)

$$(u - u_{\mathrm{disc}}, v_{\mathrm{disc}})_{\mathcal{A}} = 0 \quad \forall v_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}.$$

Hence, setting $v_{\mathrm{disc}} = u_{\mathrm{disc}} - w_{\mathrm{disc}}$, we get

$$\mathrm{err}(w_{\mathrm{disc}})^2 = \|u - u_{\mathrm{disc}}\|_{\mathcal{A}}^2 + \|u_{\mathrm{disc}} - w_{\mathrm{disc}}\|_{\mathcal{A}}^2. \tag{7.3}$$

By Cauchy-Schwarz on the first part it is clear that

$$\|u - u_{\mathrm{disc}}\|_{\mathcal{A}} = \sup_{v \in \mathcal{V}} \frac{|(u - u_{\mathrm{disc}}, v)_{\mathcal{A}}|}{\|v\|_{\mathcal{A}}}.$$

We now utilise the Galerkin orthogonality and introduce the bounded, linear map $\mathcal{Q} : \mathcal{V} \to \mathcal{V}_{\mathrm{disc}}$, obtaining

$$\|u - u_{\mathrm{disc}}\|_{\mathcal{A}} = \sup_{v \in \mathcal{V}} \frac{|(u - u_{\mathrm{disc}}, v - \mathcal{Q}v)_{\mathcal{A}}|}{\|v\|_{\mathcal{A}}}$$

and since we do not have access to the Galerkin solution u_{disc}, we reintroduce w_{disc}

$$\begin{aligned}
\|u - u_{\mathrm{disc}}\|_{\mathcal{A}} &= \sup_{v \in \mathcal{V}} \frac{|(u - w_{\mathrm{disc}} + w_{\mathrm{disc}} - u_{\mathrm{disc}}, v - \mathcal{Q}v)_{\mathcal{A}}|}{\|v\|_{\mathcal{A}}} \\
&\leq \sup_{v \in \mathcal{V}} \frac{|(u - w_{\mathrm{disc}}, v - \mathcal{Q}v)_{\mathcal{A}}|}{\|v\|_{\mathcal{A}}} + \sup_{v \in \mathcal{V}} \frac{|(w_{\mathrm{disc}} - u_{\mathrm{disc}}, v - \mathcal{Q}v)_{\mathcal{A}}|}{\|v\|_{\mathcal{A}}} \\
&\leq \sup_{v \in \mathcal{V}} \frac{|(u - w_{\mathrm{disc}}, v - \mathcal{Q}v)_{\mathcal{A}}|}{\|v\|_{\mathcal{A}}} + c_{\mathcal{Q}} \|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{A}}.
\end{aligned}$$

Here, we used the boundedness of $\mathcal{Q}$ in the energy norm by defining the constant as the operator norm

$$c_{\mathcal{Q}} := \sup_{v \in \mathcal{V}} \frac{\|(\mathrm{id} - \mathcal{Q})v\|_{\mathcal{A}}}{\|v\|_{\mathcal{A}}}.$$

We apply (7.1) again to the denominator, which yields

$$\sup_{v \in \mathcal{V}} \frac{|(u - w_{\mathrm{disc}}, v - \mathcal{Q}v)_{\mathcal{A}}|}{\|v\|_{\mathcal{A}}} \lesssim \sup_{v \in \mathcal{V}} \frac{|\langle \mathcal{A}(u) - \mathcal{A}(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle|}{\|v\|_{\mathcal{V}}} = \sup_{v \in \mathcal{V}} \frac{|\langle \mathcal{R}(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle|}{\|v\|_{\mathcal{V}}}..$$

Inserting this into (7.3) yields the desired estimate. □

Further estimates will focus on the first part. We motivate the following by noting that the above is the dual norm of the residual if $\mathcal{Q}v = 0$:

$$\|\mathcal{R}(w_{\mathrm{disc}})\|_{\mathcal{V}^*}^2 = \sup_{v \in \mathcal{V}} \frac{|\langle \mathcal{R}(w_{\mathrm{disc}}), v \rangle|}{\|v\|_{\mathcal{V}}}.$$

Since $\mathcal{V} = \mathcal{X} \otimes \mathcal{Y}$ has tensor product structure and $\|\cdot\|_{\mathcal{V}}$ is a crossnorm, we can decompose the residual by defining the *active* residual as $\mathcal{R}_\Lambda(w_{\mathrm{disc}}) \in \mathcal{X}^* \otimes \mathcal{Y}(\Lambda)$ such that

$$\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v_\Lambda \rangle = \langle \mathcal{R}(w_{\mathrm{disc}}), v_\Lambda \rangle \quad \forall v_\Lambda \in \mathcal{V}(\Lambda) = \mathcal{X} \otimes \mathcal{Y}(\Lambda).$$

This is the part of the residual that can be tested on the space that has finite dimension in the parametric domain. By the triangle inequality, the discretisation error of the discrete solution $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ can be decomposed into

$$\begin{aligned}
\mathrm{err}(w_{\mathrm{disc}}) &\leq \mathrm{err}_{\mathrm{disc}}(w_{\mathrm{disc}}) + \mathrm{err}(u_{\mathrm{disc}}), \\
\mathrm{err}(u_{\mathrm{disc}}) &\leq \mathrm{err}_{\mathrm{det}}(u_{\mathrm{disc}}) + \mathrm{err}(u_\Lambda),
\end{aligned}$$

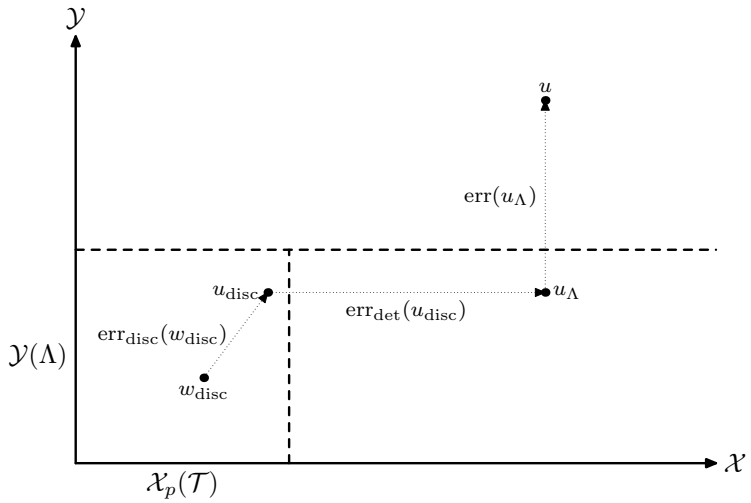

Figure 14: An illustration of the different errors.

where

$$\begin{aligned}\mathrm{err}_{\mathrm{disc}}(w_{\mathrm{disc}}) &:= \|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{V}},\\ \mathrm{err}_{\mathrm{det}}(u_{\mathrm{disc}}) &:= \|u_{\mathrm{disc}} - u_\Lambda\|_{\mathcal{V}}\end{aligned}$$

for the semi-discrete Galerkin solution $u_\Lambda \in \mathcal{X} \otimes \mathcal{Y}(\Lambda)$ and the fully discrete Galerkin solution $u_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$. This means that, assuming we could find $u_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$, its error consists firstly on the parametric semi-discretisation and then on the deterministic discretisation of the semi-discrete space $\mathcal{V}(\Lambda)$, see Figure 14.

It turns out that the deterministic error can be described by the active residual

$$\begin{aligned}\mathrm{err}_{\mathrm{det}}(u_{\mathrm{disc}}) &= \sup_{v_\Lambda \in \mathcal{V}(\Lambda)} \frac{|(u_{\mathrm{disc}} - u_\Lambda, v_\Lambda)_{\mathcal{A}}|}{\|v_\Lambda\|_{\mathcal{A}}}\\ &\lesssim \sup_{v_\Lambda \in \mathcal{V}(\Lambda)} \frac{|\langle \mathcal{A}(u_{\mathrm{disc}}) - f, v_\Lambda\rangle|}{\|v_\Lambda\|_{\mathcal{V}}}\\ &= \|\mathcal{R}_\Lambda(u_{\mathrm{disc}})\|_{\mathcal{V}^*}.\end{aligned}$$

For these reasons, it makes sense to decompose the residual into an active and an inactive part, which will be done explicitly for each problem in the following.

7.1.1 Error Estimation for the Affine Problem

We have seen in (3.15) that the operator $\mathcal{A} : \mathcal{V} \to \mathcal{V}^*$ has tensor product structure. When applied to a discretised function $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ with

$$w_{\mathrm{disc}} = \sum_{i=0}^{N-1} \sum_{\mu \in \Lambda} W(i,\mu)\varphi_i L_\mu,$$

it yields

$$\mathcal{A}(w_{\mathrm{disc}}) = \sum_{i=0}^{N-1} \sum_{\mu \in \Lambda} W(i,\mu) \left(\mathcal{A}_0(\varphi_i)L_\mu + \sum_{m=1}^{M} \mathcal{A}_m(\varphi_i)\mathcal{K}_m(L_\mu) \right),$$

using the tensorised form of the operator (3.15). This entails that $\mathcal{A}(w_{\mathrm{disc}})$ has an expansion in $\mathcal{V}^* = \mathcal{X}^* \otimes \mathcal{Y}$ and since the multiplication operator $\mathcal{K}_m$ will increase the degree of the chaos polynomials L_μ by at most one, we can define the *inactive* or boundary index set

$$\partial\Lambda = \{\mu \in \mathcal{F} \setminus \Lambda \mid \exists m \in \mathbb{N} : \mu - \epsilon_m \in \Lambda\},$$

where $\epsilon_m := (\delta_{mn})_{n\in\mathbb{N}}$ denotes the Kronecker sequence. Thus, we know

$$\mathcal{A}(w_{\text{disc}}) \in \mathcal{X}^* \otimes \mathcal{Y}(\Lambda \cup \partial\Lambda)$$

and because the right side f is independent of the parameters, we also get

$$\mathcal{R}(w_{\text{disc}}) = f - \mathcal{A}(w_{\text{disc}}) \in \mathcal{X}^* \otimes \mathcal{Y}(\Lambda \cup \partial\Lambda).$$

Using the recursive formula for the polynomials (4.3), we can deduce for $\mu \in \Lambda$ that

$$\mathcal{K}_m(L_\mu) = \beta_{\mu_m} L_{\mu-\epsilon_m} + \beta_{\mu_m+1} L_{\mu+\epsilon_m}$$

and an explicit decomposition of the residual is possible:

$$\mathcal{R}(w_{\text{disc}}) = f + \nabla \cdot \sum_{i=0}^{N-1} \sum_{\mu\in\Lambda} W(i,\mu) \left(a_0 \nabla\varphi_i L_\mu + \sum_{m=1}^{\infty} a_m \nabla\varphi_i \big(\beta_{\mu_m} L_{\mu-\epsilon_m} + \beta_{\mu_m+1} L_{\mu+\epsilon_m}\big) \right).$$

In the space $\mathcal{X}^* \otimes \mathcal{Y}(\Lambda \cup \partial\Lambda)$, the residual can be split into an active and an inactive part

$$\mathcal{R}(w_{\text{disc}}) = \mathcal{R}_\Lambda(w_{\text{disc}}) + \mathcal{R}_{\partial\Lambda}(w_{\text{disc}}) \tag{7.4}$$

that have the decompositions

$$\begin{aligned} \mathcal{R}_\Lambda(w_{\text{disc}}) &= f + \sum_{\mu\in\Lambda} (\nabla \cdot r_\mu) L_\mu, \\ \mathcal{R}_{\partial\Lambda}(w_{\text{disc}}) &= \sum_{\mu\in\partial\Lambda} (\nabla \cdot r_\mu) L_\mu \end{aligned} \tag{7.5}$$

with $\nabla \cdot r_\mu \in \mathcal{X}^*$ for all $\mu \in \Lambda \cup \partial\Lambda$, for which it holds

$$r_\mu = a_0 \nabla w_{\text{disc},\mu} + \sum_{m=1}^{\infty} a_m \Big(\beta_{\mu_m} \nabla w_{\text{disc},\mu-\epsilon_m} + \beta_{\mu_m+1} \nabla w_{\text{disc},\mu+\epsilon_m} \Big).$$

Here, we used

$$w_{\text{disc}} = \sum_{\mu\in\Lambda} w_{\text{disc},\mu} L_\mu$$

as in (4.5). Because of orthogonality of the Legendre polynomials, we get

$$\|\mathcal{R}(w_{\text{disc}})\|^2_{\mathcal{V}^*} = \|\mathcal{R}_\Lambda(w_{\text{disc}})\|^2_{\mathcal{V}^*} + \|\mathcal{R}_{\partial\Lambda}(w_{\text{disc}})\|^2_{\mathcal{V}^*}.$$

With the help of the discrete tensor operators

$$\begin{aligned} \mathbf{K}_0 &:= I_N \otimes I_{q_1} \otimes \cdots \otimes I_{q_M}, \\ \mathbf{K}_m &:= I_N \otimes I_{q_1} \otimes \cdots \otimes K_m \otimes \cdots \otimes I_{q_M}, \end{aligned}$$

with K_m as in (4.14), the active residual can be explicitly given as

$$\mathcal{R}_\Lambda(w_{\text{disc}}) = f + \sum_{m=0}^{M} \left(\sum_{i=0}^{N-1} \sum_{\mu\in\Lambda} (\mathbf{K}_m W)(i,\mu) \nabla \cdot (a_m \nabla\varphi_i) L_\mu \right).$$

This allows us to define the estimators.

Definition 7.2. For semi-discrete $w_\Lambda \in \mathcal{V}(\Lambda)$ as given in (4.5) and fully discrete $w_{\text{disc}} \in \mathcal{V}_{\text{disc}}$, we define:

1. The affine *parametric* or *tail estimator*:

$$\mathrm{est}_{\mathrm{param}}(w_\Lambda) := \left(\sum_{m=1}^{\infty} \mathrm{est}^2_{\mathrm{param},m}(w_\Lambda) \right)^{1/2},$$

where

$$\mathrm{est}_{\mathrm{param},m}(w_\Lambda) := \beta_{q_m} \left\| \frac{a_m}{a_0} \right\|_{L^\infty(D)} \left(\sum_{\mu \in \Lambda_m} \|w_{\Lambda,\mu-\epsilon_m}\|^2_{\mathcal{X}} \right)^{1/2} \qquad \forall m \in \mathbb{N};$$

Here, for all $m \in \mathbb{N}$, we used the index sets

$$\Lambda_m := \{\mu \in \partial\Lambda : \mu - \epsilon_m \in \Lambda\}; \tag{7.6}$$

2. The affine *deterministic error estimator*:

$$\begin{aligned}
\mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) &:= \left(\sum_{T \in \mathcal{T}} \mathrm{est}^2_{\mathrm{det},T}(w_{\mathrm{disc}}) + \sum_{S \in \mathcal{S}} \mathrm{est}^2_{\mathrm{det},S}(w_{\mathrm{disc}}) \right)^{1/2} \\
\mathrm{est}_{\mathrm{det},T}(w_{\mathrm{disc}}) &:= h_T \|a_0^{-1/2} \mathcal{R}_\Lambda(w_{\mathrm{disc}})\|_{L^2_\pi(\Gamma,L^2(T))} \\
\mathrm{est}_{\mathrm{det},S}(w_{\mathrm{disc}}) &:= h_S^{1/2} \|a_0^{-1/2} [\![\mathcal{R}_\Lambda(w_{\mathrm{disc}})]\!]\|_{L^2_\pi(\Gamma,L^2(S))},
\end{aligned}$$

where $[\![\cdot]\!]$ denotes the normal jump over S;

3. The affine *discrete error estimator*:

$$\mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}}) := \|\mathbf{A}_0^{-1/2}(\mathbf{A}W - F)\|_{\ell^2(\mathbb{R}^{N \times q_1 \times \cdots \times q_M})}.$$

Remark 7.3. The tail estimator is the geometric mean of the estimated errors that occur in each component because the polynomial degree is finite, and of the error that is caused by the truncation of the coefficient. We remember that the truncation can be seen as an approximation with polynomial degree equal to zero up to a factor. For the affine case, we utilise an L^∞-estimator. Since this sometimes inaccurate, we will employ a different estimator in the log-normal case, where more summands have to be taken into account

Remark 7.4. For the deterministic error estimator, we calculate the L^2-norm of the residual $\mathcal{R}_\Lambda \in \mathcal{V}^*$. This is in general not possible, as the residual is not necessarily continuous and also not always square-integrable. However, since the discrete function w_{disc} is piecewise polynomial according to (4.1), it is in particular continuously differentiable on each element T and thus its second derivative is bounded and infinitely often integrable on T.

Following [15, 17], we state the main theorem

Theorem 7.5. *Let $\mathcal{Q} : \mathcal{V} \to \mathcal{V}_{\mathrm{disc}}$ be the tensor product interpolation operator as in* (4.16)*, which is bounded and linear by definition. Then for any $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ it holds*

$$\sup_{v \in \mathcal{V}} \frac{|\langle \mathcal{R}(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle|}{\|v\|_{\mathcal{V}}} \lesssim \mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) + \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}).$$

Proof. For any $v \in \mathcal{V}$, we know

$$|\langle \mathcal{R}(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle| \leq |\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle| + |\langle \mathcal{R}_{\partial\Lambda}(w_{\mathrm{disc}}), v \rangle|.$$

This holds because $\langle \mathcal{R}_{\partial\Lambda}(w_{\mathrm{disc}}), \mathcal{Q}v \rangle = 0$ since $\mathcal{Q}v \in \mathcal{V}_{\mathrm{disc}}$.

For the first part we decompose

$$\begin{aligned}\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v - \mathcal{Q}v\rangle &= \int_\Gamma \sum_{T\in\mathcal{T}} \int_T f(v-\mathcal{Q}v) - a\nabla w_{\mathrm{disc}} \cdot \nabla(v-\mathcal{Q}v)\,\mathrm{d}x\mathrm{d}\pi(y)\\ &= \sum_{T\in\mathcal{T}} \int_\Gamma \int_T \mathcal{R}_\Lambda(w_{\mathrm{disc}})(v-Qv)\,\mathrm{d}x\mathrm{d}\pi(y)\\ &\quad + \sum_{S\in\mathcal{S}} \int_\Gamma \int_S [\![\mathcal{R}_\Lambda(w_{\mathrm{disc}})]\!](v-Qv)\,\mathrm{d}S\mathrm{d}\pi(y)\end{aligned}$$

and thus, using the Cauchy-Schwarz inequality

$$\begin{aligned}|\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v-\mathcal{Q}v\rangle| \le \sum_{T\in\mathcal{T}} &\|a_0^{-1/2}\mathcal{R}_\Lambda(w_{\mathrm{disc}})\|_{L^2_\pi(\Gamma,L^2(T))}\|a_0^{1/2}(v-\mathcal{Q}v)\|_{L^2_\pi(\Gamma,L^2(T))}\\ &+ \sum_{S\in\mathcal{S}} \|a_0^{-1/2}[\![\mathcal{R}_\Lambda(w_{\mathrm{disc}})]\!]\|_{L^2_\pi(\Gamma,L^2(S))}\|a_0^{1/2}(v-\mathcal{Q}v)\|_{L^2_\pi(\Gamma,L^2(S))}.\end{aligned}$$

The interpolation properties (4.11) and (4.12) yield

$$\begin{aligned}|\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v-\mathcal{Q}v\rangle| \lesssim \sum_{T\in\mathcal{T}} &h_T\|a_0^{-1/2}\mathcal{R}_\Lambda(w_{\mathrm{disc}})\|_{L^2_\pi(\Gamma,L^2(T))}|v|_{\mathcal{V},T}\\ &+ \sum_{S\in\mathcal{S}} h_S^{1/2}\|a_0^{-1/2}[\![\mathcal{R}_\Lambda(w_{\mathrm{disc}})]\!]\|_{L^2_\pi(\Gamma,L^2(S))}|v|_{\mathcal{V},S}\end{aligned}$$

and, since the overlaps of faces and edges is uniformly bounded and using the Hölder inequality,

$$|\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v-\mathcal{Q}v\rangle| \lesssim \mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}})\|v\|_{\mathcal{V}}.$$

For the second part of the proof, we begin with a much simpler estimation

$$|\langle \mathcal{R}_{\partial\Lambda}(w_{\mathrm{disc}}), v\rangle| \le \|\mathcal{R}_{\partial\Lambda}(w_{\mathrm{disc}})\|_{\mathcal{V}^*}\|v\|_{\mathcal{V}},$$

and with (7.5) we obtain for the first factor

$$\|\mathcal{R}_{\partial\Lambda}(w_{\mathrm{disc}})\|^2_{\mathcal{V}^*} = \sum_{\mu\in\partial\Lambda} \|\nabla\cdot r_\mu\|^2_{\mathcal{X}^*},$$

where

$$\begin{aligned}\|\nabla\cdot r_\mu\|_{\mathcal{X}^*} = \sup_{v_x\in\mathcal{X}} \frac{1}{\|v_x\|_{\mathcal{X}}} \int_D \Bigg(&a_0\nabla w_{\mathrm{disc},\mu}\\ &+ \sum_{m=1}^{\infty} a_m\Big(\beta_{\mu_m}\nabla w_{\mathrm{disc},\mu-\epsilon_m} + \beta_{\mu_m+1}\nabla w_{\mathrm{disc},\mu+\epsilon_m}\Big)\Bigg)\cdot\nabla v_x\,\mathrm{d}x.\end{aligned}$$

For $\mu\in\partial\Lambda$, we know that $w_{\mathrm{disc},\mu}=0$ and since Λ is a tensor set also $w_{\mathrm{disc},\mu+\epsilon_m}=0$ for all $m\in\mathbb{N}$. In fact, even $w_{\mathrm{disc},\mu-\epsilon_m}\neq 0$ only if $\mu\in\Lambda_m$ as defined in (7.6). Thus, for $\mu\in\Lambda_m, m\in\mathbb{N}$, this yields with Cauchy-Schwarz

$$\begin{aligned}\|\nabla\cdot r_\mu\|_{\mathcal{X}^*} &= \beta_{q_m}\sup_{v_x\in\mathcal{X}}\frac{1}{\|v_x\|_{\mathcal{X}}}\int_D a_m\nabla w_{\mathrm{disc},\mu-\epsilon_m}\cdot\nabla v_x\,\mathrm{d}x\\ &\le \beta_{q_m}\left\|\frac{a_m}{a_0}\right\|_{L^\infty(D)}\|w_{\mathrm{disc},\mu-\epsilon_m}\|_{\mathcal{X}}\end{aligned}$$

and thus

$$\begin{aligned}\|\mathcal{R}_{\partial\Lambda}(w_{\mathrm{disc}})\|^2_{\mathcal{V}^*} &\le \sum_{m=1}^{\infty}\sum_{\mu\in\Lambda_m}\left(\beta_{q_m}\left\|\frac{a_m}{a_0}\right\|_{L^\infty(D)}\|w_{\mathrm{disc},\mu-\epsilon_m}\|_{\mathcal{X}}\right)^2\\ &= \sum_{m=1}^{\infty}\mathrm{est}^2_{\mathrm{param},m}(w_{\mathrm{disc}}).\end{aligned}$$

Putting together we get

$$\begin{aligned}\sup_{v\in\mathcal{V}}\frac{|\langle\mathcal{R}(w_{\mathrm{disc}}),v-\mathcal{Q}v\rangle|}{\|v\|_{\mathcal{V}}}&\leq\sup_{v\in\mathcal{V}}\frac{|\langle\mathcal{R}_{\Lambda}(w_{\mathrm{disc}}),v-\mathcal{Q}v\rangle|+|\langle\mathcal{R}_{\partial\Lambda}(w_{\mathrm{disc}}),v\rangle|}{\|v\|_{\mathcal{V}}}\\&\lesssim\mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}})+\mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}).\end{aligned}$$

The constants that are omitted here derive from the interpolation properties. This completes the proof. □

All in all, we get the following error estimates:

Corollary 7.6. *For any $w_{\mathrm{disc}}\in\mathcal{V}_{\mathrm{disc}}$, the overall error can be estimated by*

$$\mathrm{err}(w_{\mathrm{disc}})^2\lesssim\left(\mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}})+\mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}})+\mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}})\right)^2+\mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}})^2.$$

Proof. The discrete Galerkin solution $u_{\mathrm{disc}}\in\mathcal{V}_{\mathrm{disc}}$ has the abstract form

$$u_{\mathrm{disc}}=\sum_{i=0}^{N-1}\sum_{\mu\in\Lambda}U(i,\mu)\varphi_iL_\mu$$

and therefore

$$\begin{aligned}\|u_{\mathrm{disc}}-w_{\mathrm{disc}}\|_{\mathcal{A}}^2&=\langle\mathcal{A}(u_{\mathrm{disc}}-w_{\mathrm{disc}}),u_{\mathrm{disc}}-w_{\mathrm{disc}}\rangle\\&=\langle\mathbf{A}W-\mathbf{A}U,W-U\rangle\\&\leq\|\mathbf{A}_0^{-1/2}(\mathbf{A}W-\mathbf{A}U)\|_{\ell^2(\mathbb{R}^{N\times q_1\times\cdots\times q_M})}\|\mathbf{A}_0^{1/2}(W-U)\|_{\ell^2(\mathbb{R}^{N\times q_1\times\cdots\times q_M})}\\&=\|\mathbf{A}_0^{-1/2}(\mathbf{A}W-F)\|_{\ell^2(\mathbb{R}^{N\times q_1\times\cdots\times q_M})}\|u_{\mathrm{disc}}-w_{\mathrm{disc}}\|_{\mathcal{V}}\\&\lesssim\mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}})\|u_{\mathrm{disc}}-w_{\mathrm{disc}}\|_{\mathcal{A}},\end{aligned}$$

where we used the fact that $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\mathcal{V}}$ are equivalent norms. The rest follows from Theorem 7.1 and Theorem 7.5 with constants deriving from several factors, see e.g. [15]. □

Efficient Computation of the Affine Estimators

We will now briefly show that these error estimators are computable in polynomial time because of the tensor representation. For the tail estimator $\mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}})$ we first look at $m\in\mathbb{N}\setminus\mathrm{supp}(\Lambda)$:

$$\begin{aligned}\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}})&=\beta_{q_m}\left\|\frac{a_m}{a_0}\right\|_{L^\infty(D)}\left(\sum_{\mu\in\Lambda_m}\|w_{\mathrm{disc},\mu-\epsilon_m}\|_{\mathcal{X}}^2\right)^{1/2}\\&=\beta_1\left\|\frac{a_m}{a_0}\right\|_{L^\infty(D)}\left(\sum_{\mu\in\Lambda}\|w_{\mathrm{disc},\mu}\|_{\mathcal{X}}^2\right)^{1/2}\\&=\beta_1\left\|\frac{a_m}{a_0}\right\|_{L^\infty(D)}\|w_{\mathrm{disc}}\|_{\mathcal{V}}.\end{aligned}$$

The norm $\|w_{\mathrm{disc}}\|_{\mathcal{V}}$ can be calculated in advance. Note that we required in (3.12)

$$\sum_{m=1}^{\infty}\left\|\frac{a_m}{a_0}\right\|_{L^\infty(D)}<\lambda<1$$

and thus the infinite series in the parametric estimator $\mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}})$ converges.

For $m\in\mathrm{supp}(\Lambda)$, we use the discretisation

$$\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}})=\beta_{q_m}\left\|\frac{a_m}{a_0}\right\|_{L^\infty(D)}\left(\sum_{\mu\in\Lambda_m}\sum_{i,j=0}^{N-1}W(i,\mu-\epsilon_m)W(j,\mu-\epsilon_m)(\varphi_i,\varphi_j)_{\mathcal{X}}\right)^{1/2}$$

and in tensor notation, we can compute

$$\sum_{\mu\in\Lambda_m} W(i,\mu-\epsilon_m)W(j,\mu-\epsilon_m) = \sum_{\mu_1=0}^{q_1-1}\cdots\cancel{\sum_{\mu_m=0}^{q_m-1}}\cdots\sum_{\mu_M=0}^{q_M-1} W(i,\mu_1,\dots,q_m-1,\dots,\mu_M)W(j,\mu_1,\dots,q_m-1,\dots,\mu_M)$$

and since W is in TT format this is computable in polynomial time as it is just a series of matrix and vector products and the result is a low rank matrix [17]. The computation of $\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}})$ only requires an additional Hadamard product of this matrix with the Gram matrix

$$G_1(i,j) := (\varphi_i,\varphi_j)_{\mathcal{X}}.$$

For the calculation of the deterministic estimator, we first isolate the right side

$$\begin{aligned}\mathrm{est}^2_{\mathrm{det},T}(w_{\mathrm{disc}}) &= h_T^2\|a_0^{-1/2}\mathcal{R}_\Lambda(w_{\mathrm{disc}})\|^2_{L^2_\pi(\Gamma,L^2(T))}\\ &= h_T^2\|a_0^{-1/2}f\|^2_{L^2(T)} + 2h_T^2(a_0^{-1}f,\nabla\cdot r_0(w_{\mathrm{disc}}))_{L^2(T)}\\ &\quad + h_T^2\|a_0^{-1/2}\sum_{\mu\in\Lambda}\nabla\cdot r_\mu(w_{\mathrm{disc}})\|^2_{L^2_\pi(\Gamma,L^2(T))}\end{aligned}$$

and then we expand

$$\|a_0^{-1/2}\sum_{\mu\in\Lambda}\nabla\cdot r_\mu(w_{\mathrm{disc}})\|^2_{L^2_\pi(\Gamma,L^2(T))} = \sum_{m,m'=0}^{M}\sum_{i,j=0}^{N-1}\sum_{\mu\in\Lambda}(\mathbf{K}_mW)(i,\mu)(\mathbf{K}_{m'}W)(j,\mu)(a_0^{-1}\nabla\cdot a_m\nabla\varphi_i,\nabla\cdot a_{m'}\nabla\varphi_j)_{L^2(T)}.$$

Again, the product $\sum_{\mu\in\Lambda}(\mathbf{K}_mW)(i,\mu)(\mathbf{K}_{m'}W)(j,\mu)$ can be calculated efficiently in TT format, see (5.13), and what remains is a sum of $2M$ Hadamard products with the Gram matrix

$$G_2[m,m'](i,j) = (a_0^{-1}\nabla\cdot a_m\nabla\varphi_i,\nabla\cdot a_{m'}\nabla\varphi_i)_{L^2(T)},$$

which can be computed beforehand.

Using a similar reordering for the edges S, one obtains

$$\mathrm{est}^2_{\mathrm{det,S}}(w_{\mathrm{disc}}) := h_S\sum_{m,m'=0}^{M}\sum_{i,j=0}^{N-1}\sum_{\mu\in\Lambda}(\mathbf{K}_mW)(i,\mu)(\mathbf{K}_{m'}W)(j,\mu)(a_0^{-1}[\![a_m\nabla\varphi_i]\!],[\![a_{m'}\nabla\varphi_j]\!])_{L^2(S)}.$$

We can also efficiently evaluate the discrete error since it just requires taking the Frobenius norm on the tensor space $\mathbb{R}^{N\times q_1\times\cdots\times q_M}$, with $w_{\mathrm{disc}} = \sum_{i=0}^{N-1}\sum_{\mu\in\Lambda}W(i,\mu)\varphi_iL_\mu$, which is efficient in the TT format.

7.1.2 Error Estimators for the Log-Normal Problem

Error estimation in the log-normal case of the parametric diffusion equation poses a few difficulties that do not occur in the affine problem. Firstly, the operator has to be approximated in order to obtain the discrete Galerkin operator $\mathbf{A}$. Secondly, the expansion of this discrete operator is mostly independent of the discretisation of the solution, in the sense that we have to choose a rank and a degree for the Hermite expansion in each component that is not dictated by the rank of the solution or its Hermite approximation. And lastly, we have to deal with the extra weights $\zeta_{\vartheta\rho}$ that enter the integrals.

To address the first problem, we define the operators as in (6.9). This leads to a decomposition of the residual

$$\begin{aligned}\mathcal{R}(v) &= \mathcal{R}_+(v) + \mathcal{R}_-(v),\\ \mathcal{R}_+(v) &:= f - \mathcal{A}_+(v),\\ \mathcal{R}_-(v) &:= -\mathcal{A}_-(v).\end{aligned}$$

In this thesis, we assume that the operator is given in its approximate semi-discrete form $\mathcal{A}_+$ and only estimate the error

$$\operatorname{err}(w_{\mathrm{disc}})^2 := \|w_{\mathrm{disc}} - u\|_{\mathcal{A}_+}^2 = \int_\Gamma \int_D a_{\mathrm{disc}} \nabla(w_{\mathrm{disc}} - u) \cdot \nabla(w_{\mathrm{disc}} - u) \,\mathrm{d}x \mathrm{d}\gamma_{\vartheta\rho}(y),$$

which is equivalent to the dual norm of the semi-discrete residual $\mathcal{R}_+(v)$. This is of course a great simplification. Estimation of the error that results from approximating the operator is subject to future research. Again, we decompose

$$\begin{aligned} \operatorname{err}(w_{\mathrm{disc}}) &\leq \operatorname{err}_{\mathrm{disc}}(w_{\mathrm{disc}}) + \operatorname{err}_{\mathrm{det}}(u_{\mathrm{disc}}) + \operatorname{err}(u_\Lambda), \\ \operatorname{err}_{\mathrm{disc}}(w_{\mathrm{disc}}) &= \|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{A}_+}, \\ \operatorname{err}_{\mathrm{det}}(u_{\mathrm{disc}}) &= \|u_{\mathrm{disc}} - u_\Lambda\|_{\mathcal{A}_+} \end{aligned}$$

with Galerkin solutions $u_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ and $u_\Lambda \in \mathcal{V}_{\vartheta\rho}(\Lambda)$.

If one applies the tensorised operator $\mathcal{A}_+$ from (6.9) to the discrete solution $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ that reads

$$w_{\mathrm{disc}} = \sum_{i=0}^{N-1} \sum_{\mu \in \Lambda} W(i,\mu) \varphi_i H_{\vartheta\rho,\mu},$$

one obtains

$$\begin{aligned} \mathcal{A}_+(w_{\mathrm{disc}}) = \sum_{i=0}^{N-1} \sum_{\mu\in\Lambda} W(i,\mu) \Bigg(\sum_{k_1=1}^{s_1} \cdots \sum_{k_\ell=1}^{s_L} \mathcal{A}_0[k_1](\varphi_i) \prod_{m=1}^{M} \mathcal{A}_m[k_m,k_{m+1}](H_{\vartheta\rho,\mu_m}) \\ \times \prod_{\ell=M+1}^{L} \mathcal{A}_\ell[k_\ell,k_{\ell+1}](H_{\vartheta\rho,0}) \Bigg). \end{aligned}$$

For $m \leq M$ this results in a product of two Hermite polynomials that can again be stated in a Hermite basis using Proposition 4.4 and the coefficients in (4.15)

$$\begin{aligned} \mathcal{A}_m[k_m,k_{m+1}](H_{\vartheta\rho,\mu_m}) &= \sum_{\nu_m=0}^{d_m-1} A_m(k_m,\nu_m,k_{m+1}) H_{\vartheta\rho,\nu_m} H_{\vartheta\rho,\mu_m} \\ &= \sum_{\eta_m=0}^{d_m+q_m-2} \left(\sum_{\nu_m=0}^{d_m-1} A_m(k_m,\nu_m,k_{m+1}) \kappa_{\mu_m,\nu_m,\eta_m} \right) H_{\vartheta\rho,\eta_m}. \end{aligned} \tag{7.7}$$

Our aim is to decompose the residual into an active and an inactive part using the TT decomposition of the discrete function

$$\begin{aligned} w_{\mathrm{disc}} &= \sum_{i=0}^{N-1} \sum_{\mu\in\Lambda} W(i,\mu) \varphi_i H_{\vartheta\rho,\mu} \\ &= \sum_{k_1=1}^{r_1} \cdots \sum_{k_M=1}^{r_M} \left(\sum_{i=0}^{N-1} W_0(i,k_1) \varphi_i \right) \left(\prod_{m=1}^{M} \sum_{\mu_m=0}^{q_m-1} W_m(k_m,\mu_m,k_{m+1}) H_{\vartheta\rho,\mu_m} \right). \end{aligned}$$

Since the product of the Hermite polynomials for each $m = 1,\dots,M$ has degree at most $d_m + q_m - 2$, it will be useful to define the following index set:

$$\begin{aligned} \Xi := \Delta + \Lambda := \big\{ \eta = (\eta_1,\dots,\eta_L,0,\dots) \colon \eta_m = 0,\dots,d_m+q_m-2, \quad m = 1,\dots,M; \\ \eta_\ell = 0,\dots,d_\ell - 1, \quad \ell = M+1,\dots,L \big\}. \end{aligned}$$

As a consequence, it is possible to write the residual as a TT representation:

$$\begin{aligned}\mathcal{R}_+(w_{\mathrm{disc}}) &= f - \mathcal{A}_+(w_{\mathrm{disc}})\\ &= f + \sum_{k_1=1}^{r_1}\cdots\sum_{k_M=1}^{r_M}\sum_{k'_1=1}^{s_1}\cdots\sum_{k'_L=1}^{s_L}\nabla\cdot r_0[k_1,k'_1]\\ &\quad\times\sum_{\eta\in\Xi}\left(\prod_{m=1}^{M}R_m(k_m,k'_m,\eta_m,k_{m+1},k'_{m+1})\prod_{\ell=M+1}^{L}A_\ell(k'_\ell,\eta_\ell,k'_{\ell+1})\right)H_{\vartheta\rho,\eta}.\end{aligned}$$

with continuous first component

$$r_0[k_1,k'_1](x) = \sum_{i=0}^{N-1} a_0[k'_1](x)W_0(i,k_1)\nabla\varphi_i(x)$$

and stochastic components for $m = 1,\dots,M$:

$$R_m(k_m,k'_m,\eta_m,k_{m+1},k'_{m+1}) = \sum_{\mu_m=0}^{q_m-1}\sum_{\nu_m=0}^{d_m-1}A(k'_m,\nu_m,k'_{m+1})W_m(k_m,\mu_m,k_{m+1})\kappa_{\mu_m,\nu_m,\eta_m}.$$

Altogether, for all $\eta\in\Xi$, we obtain the tensorised function similar to (6.6)

$$\begin{aligned}r(x,\eta) &= \sum_{k_1=1}^{r_1}\cdots\sum_{k_M=1}^{r_M}\sum_{k'_1=1}^{s_1}\cdots\sum_{k'_L=1}^{s_L}r_0[k_1,k'_1](x)\\ &\quad\times\left(\prod_{m=1}^{M}R_m(k_m,k'_m,\eta_m,k_{m+1},k'_{m+1})\prod_{\ell=M+1}^{L}A_\ell(k'_\ell,\eta_\ell,k'_{\ell+1})\right).\end{aligned}$$

This is again a TT tensor with ranks $(r_1s_1,\dots,r_Ms_M,s_{M+1},\dots,s_L)$ and continuous first component. The physical dimensions are d_m+q_m-2 for all $m=1,\dots,M$, which follows from (7.7), and $d_\ell-1$ for $\ell=M+1,\dots,L$. This allows for a representation of the residual as an active and an inactive residual similar to (7.4)

$$\begin{aligned}\mathcal{R}_+(w_{\mathrm{disc}}) &= f - \mathcal{A}_+(w_{\mathrm{disc}})\\ &= f + \sum_{\eta\in\Xi}\nabla\cdot r(\cdot,\eta)H_{\vartheta\rho,\eta}\\ &= \mathcal{R}_{+,\Lambda}(w_{\mathrm{disc}}) + \mathcal{R}_{+,\Xi\setminus\Lambda}(w_{\mathrm{disc}})\end{aligned}$$

with

$$\begin{aligned}\mathcal{R}_{+,\Lambda}(w_{\mathrm{disc}}) &= f + \sum_{\eta\in\Lambda}\nabla\cdot r(\cdot,\eta)H_{\vartheta\rho,\eta},\\ \mathcal{R}_{+,\Xi\setminus\Lambda}(w_{\mathrm{disc}}) &= \sum_{\eta\in\Xi\setminus\Lambda}\nabla\cdot r(\cdot,\eta)H_{\vartheta\rho,\eta},\end{aligned}$$

where $\nabla\cdot r(\cdot,\eta)\in\mathcal{X}^*$ for all $\eta\in\Xi$.

Finally, we define the linear basis change operator that translates integrals over two Hermite polynomials in the measure $\gamma_{\vartheta\rho}$ to the measure γ:

$$\begin{aligned}\mathbf{H}_{\vartheta\rho\to0} : \mathbb{R}^{N\times q_1\times\cdots\times q_M} &\to \mathbb{R}^{N\times q_1\times\cdots\times q_M}\\ \mathbf{H}_{\vartheta\rho\to0} &:= Z_0\otimes Z_1\otimes\cdots\otimes Z_M\\ Z_0(i,j) &:= \int_D \nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}x\\ Z_m(\mu_m,\mu'_m) &:= \int_{-\infty}^{\infty}H_{\vartheta\rho,\mu_m}(y_m)H_{\vartheta\rho,\mu'_m}(y_m)\,\mathrm{d}\gamma_m(y_m)\\ &= \sum_{\nu_m=0}^{\lfloor\mu_m/2\rfloor}(\kappa^{\vartheta\rho\to0}_{\mu_m,\nu_m})^2\quad\forall m=1,\dots,M.\end{aligned}$$

The log-normal error estimators are a bit different from the affine case:

Definition 7.7. For fully discrete $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ the log-normal error estimators are defined as follows:

1. The log-normal *parametric or tail estimator*:

$$\mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}) := \left(\int_\Gamma \int_D \Big(\sum_{\eta \in \Xi \setminus \Lambda} r(x,\eta) H_{\vartheta\rho,\eta}(y)\, \zeta_{\vartheta\rho}(y) \Big)^2 \mathrm{d}x \mathrm{d}\gamma(y) \right)^{1/2};$$

2. The log-normal *deterministic error estimator*:

$$\begin{aligned} \mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) &:= \left(\sum_{T \in \mathcal{T}} \mathrm{est}_{\mathrm{det},T}^2(w_{\mathrm{disc}}) + \sum_{S \in \mathcal{S}} \mathrm{est}_{\mathrm{det},S}^2(w_{\mathrm{disc}}) \right)^{1/2}, \\ \mathrm{est}_{\mathrm{det},T}(w_{\mathrm{disc}}) &:= h_T \| \mathcal{R}_\Lambda(w_{\mathrm{disc}}) \zeta_{\vartheta\rho} \|_{L^2_\gamma(\Gamma, L^2(T))}, \\ \mathrm{est}_{\mathrm{det},S}(w_{\mathrm{disc}}) &:= h_S^{1/2} \| [\![\mathcal{R}_\Lambda(w_{\mathrm{disc}})]\!] \zeta_{\vartheta\rho} \|_{L^2_\gamma(\Gamma, L^2(S))}; \end{aligned}$$

3. The log-normal *discrete error estimator*:

$$\mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}}) := \| (\mathbf{A}(W) - F) \mathbf{H}_{\vartheta\rho \to 0}^{-1/2} \|_{\ell^2(\mathbb{R}^{N \times q_1 \times \cdots \times q_M})}. \tag{7.8}$$

Theorem 7.8. *The error of the discrete function $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ for the discretised operator $\mathcal{A}_+$ can be estimated with*

$$\mathrm{err}(w_{\mathrm{disc}})^2 \lesssim \big(\mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) + \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}) + \mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}}) \big)^2 + \mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}})^2.$$

Proof. In the log-normal case, the equation from Theorem 7.1 is

$$\mathrm{err}(w_{\mathrm{disc}})^2 \lesssim \left(\sup_{v \in \mathcal{V}_{\vartheta\rho}} \frac{|\langle \mathcal{R}_+(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle_{\vartheta\rho}|}{\|v\|_{\mathcal{V}}} + c_{\mathcal{Q}} \|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{A}_+} \right)^2 + \|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{A}_+}^2.$$

Note that here, we have the dual pairing $\langle \cdot, \cdot \rangle_{\vartheta\rho}$ and the energy norms of the discrete operator $\mathcal{A}_+$ that are defined via the bilinear form (6.11). As before, we need to show

$$\sup_{v \in \mathcal{V}_{\vartheta\rho}} \frac{|\langle \mathcal{R}_+(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle_{\vartheta\rho}|}{\|v\|_{\mathcal{V}}} \lesssim \mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) + \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}})$$

and

$$\|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{A}_+} \lesssim \mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}}). \tag{7.9}$$

For the first inequality, we decompose

$$|\langle \mathcal{R}_+(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle_{\vartheta\rho}| \leq |\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle_{\vartheta\rho}| + |\langle \mathcal{R}_{\Xi\setminus\Lambda}(w_{\mathrm{disc}}), v \rangle_{\vartheta\rho}|$$

and analogously to the proof of Theorem 7.5, we get

$$|\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}), v - \mathcal{Q}v \rangle_{\vartheta\rho}| = |\langle \mathcal{R}_\Lambda(w_{\mathrm{disc}}) \zeta_{\vartheta\rho}, v - \mathcal{Q}v \rangle| \lesssim \mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) \|v\|_{\mathcal{V}}.$$

The stochastic error is estimated in a different way here. Instead of isolating the L^∞-norm of the diffusion coefficient, we use the Cauchy-Schwarz inequality:

$$\begin{aligned} \langle \mathcal{R}_{\Xi\setminus\Lambda}(w_{\mathrm{disc}}), v \rangle_{\vartheta\rho} &= \int_\Gamma \int_D \Big(\sum_{\eta \in \Xi \setminus \Lambda} r(x,\eta) H_{\vartheta\rho,\eta}(y) \Big) \cdot \nabla v(x,y)\, \zeta_{\vartheta\rho}(y) \mathrm{d}x \mathrm{d}\gamma(y) \\ &\leq \int_\Gamma \int_D \Big(\sum_{\eta \in \Xi \setminus \Lambda} r(x,\eta) H_{\vartheta\rho,\eta}(y)\, \zeta_{\vartheta\rho}(y) \Big)^2 \mathrm{d}x \mathrm{d}\gamma(y) \|v\|_{\mathcal{V}} \\ &= \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}) \|v\|_{\mathcal{V}}. \end{aligned}$$

The second inequality (7.9) is the estimation of the discrete error, which is obtained in a similar fashion as in the proof of Corollary 7.6. For $v_{\mathrm{disc}} = \sum_{i=0}^{N-1} \sum_{\mu\in\Lambda} V(i,\mu)\varphi_i H_{\vartheta\rho,\mu} \in \mathcal{V}_{\mathrm{disc}}$, it holds

$$\int_\Gamma \int_D \nabla v_{\mathrm{disc}} \cdot \nabla v_{\mathrm{disc}} \,\mathrm{d}x\mathrm{d}\gamma(y) = \langle V\mathbf{H}_{\vartheta\rho\to 0}, V\rangle = \|V\mathbf{H}^{1/2}_{\vartheta\rho\to 0}\|_{\ell^2(\mathbb{R}^{N\times q_1\times\cdots\times q_M})}.$$

With this and using (3.23), we can see that

$$\begin{aligned}
\|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|^2_{\mathcal{A}_+} &= \int_\Gamma \int_D \mathcal{A}_+(w_{\mathrm{disc}} - u_{\mathrm{disc}}) \cdot \nabla(w_{\mathrm{disc}} - u_{\mathrm{disc}}) \,\mathrm{d}x\mathrm{d}\gamma_{\vartheta\rho}(y) \\
&= \langle \mathbf{A}W - F, W - U\rangle \\
&= \langle (\mathbf{A}W - F)\mathbf{H}^{-1/2}_{\vartheta\rho\to 0}, (W-U)\mathbf{H}^{1/2}_{\vartheta\rho\to 0}\rangle \\
&\le \|(\mathbf{A}W - F)\mathbf{H}^{-1/2}_{\vartheta\rho\to 0}\|_{\ell^2(\mathbb{R}^{N\times q_1\times\cdots\times q_M})} \|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{L^2_\gamma(\Gamma,\mathcal{X})} \\
&\lesssim \|(\mathbf{A}W - F)\mathbf{H}^{-1/2}_{\vartheta\rho\to 0}\|_{\ell^2(\mathbb{R}^{N\times q_1\times\cdots\times q_M})} \|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{A}_+}
\end{aligned}$$

and thus

$$\|w_{\mathrm{disc}} - u_{\mathrm{disc}}\|_{\mathcal{A}_+} \lesssim \mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}}).$$

This yields the desired result. □

Remark 7.9. In order to get suitable measures for the estimation in Definition 7.7, the squared density $\zeta^2_{\vartheta\rho}$ appears, which upon scaling again is a Gaussian measure. We show this for $L < \infty$: With $\sigma = (\sigma_\ell)_{1\le\ell\le L} = (\exp(\vartheta\rho\alpha_\ell))_{1\le\ell\le L}$, it holds

$$\begin{aligned}
\zeta_{\vartheta\rho}(y)^2 &= \left(\prod_{\ell=1}^L \frac{1}{\sigma_\ell^2}\right) \exp\left(-\sum_{\ell=1}^L (\sigma_\ell^{-2} - 1) y_\ell^2\right) \\
&= \left(\prod_{\ell=1}^L \frac{1}{\sigma_\ell\sqrt{2-\sigma_\ell^2}}\right) \left(\prod_{\ell=1}^L \frac{\sqrt{2-\sigma_\ell^2}}{\sigma_\ell}\right) \exp\left(-\frac{1}{2}\sum_{\ell=1}^L \Big(\frac{2-\sigma_\ell^2}{\sigma_\ell^2} - 1\Big) y_\ell^2\right) \\
&= c_\sigma \tilde{\zeta}_{\sigma'}(y).
\end{aligned} \tag{7.10}$$

The weight $\tilde{\zeta}_{\sigma'}$ is again a density with respect to γ. It corresponds to a mean zero Gaussian measure

$$\tilde{\gamma}_{\sigma'} = \bigotimes_{\ell=1}^L \mathcal{N}_{(\sigma'_\ell)^2}$$

with standard deviation $\sigma' = (\sigma'_\ell)_{1\le\ell\le L}$, where

$$\sigma'_\ell = \frac{\sigma_\ell}{\sqrt{2-\sigma_\ell^2}}.$$

The constant factor c_σ in (7.10) is given by

$$c_\sigma := \prod_{\ell=1}^L \frac{1}{\sigma_\ell\sqrt{2-\sigma_\ell^2}}.$$

Note that it is necessary to impose a restriction on ϑ such that $\sigma_\ell^2 = \exp(2\vartheta\rho\alpha_\ell) < 2$ for all $\ell = 1,\dots,L$ in order to ensure that the argument in the square root is positive.

It is also important to check whether the new measure is *weaker* or *stronger* than, e.g.,$\gamma_{\vartheta\rho}$ in the sense of Lemma 3.7. This is easy to see:

$$\sigma'_\ell = \frac{\sigma_\ell}{\sqrt{2-\sigma_\ell^2}} = \frac{\exp(\vartheta\rho\alpha_\ell)}{\sqrt{2-\exp(2\vartheta\rho\alpha_\ell)}} \ge \exp(\vartheta\rho\alpha_\ell) = \sigma_\ell.$$

Therefore, it holds

$$L^2_{\tilde{\gamma}_{\sigma'}}(\Gamma,\mathcal{X}^*) \subset L^2_{\gamma_{\vartheta\rho}}(\Gamma,\mathcal{X}^*).$$

This means that functions that are integrable with respect to the measure $\gamma_{\vartheta\rho}$ are not necessarily integrable with respect to the squared measure. However, since f is independent of the parameters and $\mathcal{A}_+(w_{\mathrm{disc}}) \in \mathcal{X}^* \otimes \mathcal{Y}(\Xi)$ has a polynomial chaos expansion of finite degree, the residual $\mathcal{R}_+(w_{\mathrm{disc}})$ is integrable over the parameters for any Gaussian measure and therefore also $\mathcal{R}_+(w_{\mathrm{disc}}) \in L^2_{\tilde{\gamma}_{\sigma'}}(\Gamma, \mathcal{X}^*)$. If f also depends on the parameters, we recall Theorem 4.5 and that we have to require

$$f \in L^q_{\gamma_\rho}(\Gamma, \mathcal{X}^*)$$

for $q > 2$. With the above arguments it is also necessary to ensure

$$f \in L^q_{\tilde{\gamma}_{\sigma'}}(\Gamma, \mathcal{X}^*),$$

which could be done by setting ϑ very small.

Efficient Computation of the Log-Normal Estimators

As before, the error estimators in Definition 7.7 can be calculated efficiently in the TT format. For each element $T \in \mathcal{T}$ of the triangulation, the residual estimator is

$$\begin{aligned}
\mathrm{est}_{\mathrm{det},T}(w_{\mathrm{disc}})^2 &= h_T^2 \|\mathcal{R}_\Lambda(w_{\mathrm{disc}})\zeta_{\vartheta\rho}\|^2_{L^2_\gamma(\Gamma, L^2(T))} \\
&= h_T^2 \int_\Gamma \int_T \left(f + \sum_{\eta\in\Lambda} \nabla \cdot r(x,\eta) H_{\vartheta\rho,\eta} \right)^2 \zeta^2_{\vartheta\rho} \mathrm{d}x \mathrm{d}\gamma(y) \\
&= h_T^2 (f,f)_{L^2(T)} \int_\Gamma \zeta^2_{\vartheta\rho}\, \mathrm{d}\gamma(y) + 2h_T^2 \sum_{\eta\in\Lambda} (f, \nabla\cdot r(x,\eta))_{L^2(T)} \int_\Gamma H_{\vartheta\rho,\eta}\, \zeta^2_{\vartheta\rho} \mathrm{d}\gamma(y) \\
&\quad + \sum_{\eta\in\Lambda}\sum_{\eta'\in\Lambda} (\nabla\cdot r(x,\eta), \nabla\cdot r(x,\eta'))_{L^2(T)} \int_\Gamma H_{\vartheta\rho,\eta} H_{\vartheta\rho,\eta'}\, \zeta^2_{\vartheta\rho} \mathrm{d}\gamma(y).
\end{aligned}$$

A downside of the change of the measure to γ and the involved weight $\zeta^2_{\vartheta\rho}$ is the fact that the shifted Hermite polynomials are not orthogonal with respect to this measure. However, this property can be restored easily by calculating the basis change integrals beforehand. This results in a another tensor product operator that is defined element-wise for $\eta, \eta' \in \Xi$:

$$\begin{aligned}
\tilde{\mathbf{H}}(\eta,\eta') &:= \tilde{Z}_1(\eta_1,\eta_1') \cdots \tilde{Z}_L(\eta_L,\eta_L') \\
\tilde{Z}_\ell(\eta_\ell,\eta_\ell') &= \int_\Gamma H_{\vartheta\rho,\eta_\ell} H_{\vartheta\rho,\eta_\ell'}\, \zeta^2_{\vartheta\rho,\ell} \mathrm{d}\gamma_\ell(y_\ell),
\end{aligned}$$

This operator encodes the basis change to the measure $\tilde{\zeta}_{\sigma'}$ and can be inserted in order to calculate the scalar product. With this, the estimator takes the form

$$\begin{aligned}
\mathrm{est}_{\mathrm{det},T}(w_{\mathrm{disc}})^2 &= h_T^2 (f,f)_{L^2(T)} \int_\Gamma \zeta^2_{\vartheta\rho}\, \mathrm{d}\gamma(y) + 2h_T^2 \sum_{\eta\in\Lambda} \tilde{\mathbf{H}}(\eta,0)(f, \nabla\cdot r(x,\eta))_{L^2(T)} \\
&\quad + \sum_{\eta\in\Lambda}\sum_{\eta'\in\Lambda} \tilde{\mathbf{H}}(\eta,\eta')(\nabla\cdot r(x,\eta), \nabla\cdot r(x,\eta'))_{L^2(T)}.
\end{aligned}$$

Since $\tilde{\mathbf{H}}$ is a tensor product operator, this summation can be done component-wise, i.e. performing a matrix-vector multiplication of every component of the operator $\tilde{\mathbf{H}}$ with the corresponding component of the tensor function r. The first two summands are easily computable, as they are independent of the high-dimensional sum over $\eta \in \Xi$.

Similarly, for the jump over the edge S we obtain the estimator

$$\mathrm{est}_{\mathrm{det},S}(w_{\mathrm{disc}})^2 = h_S \sum_{\eta\in\Lambda}\sum_{\eta'\in\Lambda} \tilde{\mathbf{H}}(\eta,\eta')([\![\nabla\cdot r(x,\eta)]\!], [\![\nabla\cdot r(x,\eta')]\!])_{L^2(S)}.$$

Analogously to the affine case, both of these estimators can then be computed efficiently in the TT format.

The parametric error estimator $\mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}})$ can be estimated in a similar way. To gain additional information about the residual influence of certain stochastic dimensions, we sum over specific index sets. Let

$$\Xi_m := \{(\eta_1, \dots, q_m, \dots, \eta_M, 0, \dots) \in \mathcal{F} : \eta_\ell = 0, \dots, q_\ell - 1;\ \ell = 1, \dots, \cancel{m}, \dots, M\},$$

where the strike through means that ℓ takes all values but m. For every $m = 1, 2, \dots$ and $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ we define

$$\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}})^2 := \int_\Gamma \int_D \zeta_{\vartheta\rho}^2 \Big(\sum_{\eta \in \Xi_m} r(x,\eta) H_{\vartheta\rho,\eta} \Big)^2 \mathrm{d}x \mathrm{d}\gamma(y).$$

Using the same arguments and notation as above, we can simplify

$$\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}})^2 = \int_D \sum_{\eta \in \Xi_n} \Big(r(x,\eta) \cdot \sum_{\eta' \in \Xi_m} \tilde{\mathbf{H}}(\eta,\eta') r(x,\eta') \Big) \mathrm{d}x.$$

These operations, including the calculation of the discrete error estimator (7.8), can be executed efficiently in the TT format, see (5.13).

7.2 An Adaptive Algorithm

After having found a way to estimate the error for every aspect of the discretisation, it is possible to refine these adaptively according to the estimators. As discussed before, the deterministic estimator assesses the error that arises from the finite element method. The discrete error estimator evaluates the error made by a low rank approximation. The rest of the error is estimated by the parametric error estimator.

The adaptive algorithm described in this section is similar to the algorithms presented in [15, 16, 17]. Given some mesh $\mathcal{T}$, a fixed polynomial degree p, a finite tensor set $\Lambda \subset \mathcal{F}$, and a start tensor W with TT rank r, we assume that a numerical approximation $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ is obtained by a function

$$w_{\mathrm{disc}}^+ \leftarrow \mathrm{Solve}[\Lambda, \mathcal{T}, r, W].$$

In our implementation, we used the ALS algorithm for this, but all other algorithms are feasible as well. The affine and log-normal error indicators in Definitions 7.2 and 7.7 and thus the overall upper bound

$$\mathrm{est}_{\mathrm{all}}(w_{\mathrm{disc}}) := \sqrt{\big(\mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) + \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}) + \mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}})\big)^2 + \mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}})^2}$$

in Corollary 7.6 and Theorem 7.8 are computed by the methods

$$\begin{aligned} (\mathrm{est}_{\mathrm{det},T}(w_{\mathrm{disc}}))_{T\in\mathcal{T}}, \mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}) &\leftarrow \mathrm{Estimate}_x[w_{\mathrm{disc}}, \Lambda, \mathcal{T}, p], \\ (\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}}))_{m\in\mathbb{N}}, \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}) &\leftarrow \mathrm{Estimate}_y[w_{\mathrm{disc}}, \Lambda], \\ \mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}}) &\leftarrow \mathrm{Estimate}_{\mathrm{ALS}}[w_{\mathrm{disc}}]. \end{aligned}$$

Depending on which error is largest, either the mesh is refined, or the index set Λ is enlarged, or the rank r of the solution is increased. This is done as follows:

- If the deterministic error $\mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}})$ outweighs the others, we mark the elements $T \in \mathcal{T}$ that have the largest error $\mathrm{est}_{\mathrm{det},T}(w_{\mathrm{disc}})$ until the sum of errors on the elements in this marked subset $\mathcal{T}_{\mathrm{mark}} \subset \mathcal{T}$ exceeds a certain ratio $0 < \theta < 1$. This is called the *Dörfler property*

$$\sum_{T \in \mathcal{T}_{\mathrm{mark}}} \mathrm{est}_{\mathrm{det},T}(w_{\mathrm{disc}}) \geq \theta\, \mathrm{est}_{\mathrm{det}}(w_{\mathrm{disc}}).$$

  We denote this by

$$\mathcal{T}_{\mathrm{mark}} \leftarrow \mathrm{Mark}_x[\theta, (\mathrm{est}_{\mathrm{det},T}(w_N, \Lambda))_{T\in\mathcal{T}}, \mathrm{est}_{\mathrm{det}}(w_N, \Lambda, \mathcal{T})].$$

  The elements in this subset are then refined by splitting them into two smaller elements

$$\mathcal{T}^+ \leftarrow \mathrm{Refine}_x[\mathcal{T}, \mathcal{T}_{\mathrm{mark}}];$$

Algorithm 4: The TTASGFEM algorithm.

input : Old solution w_{disc} with solution tensor W and rank r;
mesh $\mathcal{T}$ with degrees p;
index set Λ; accuracy $\epsilon_{\mathrm{TTASGFEM}}$.
output: New solution w_{disc}^{+} with new solution tensor W^{+};
new mesh $\mathcal{T}^{+}$, or new index set Λ^{+}, or new rank $r+1$.

$w_{\mathrm{disc}}^{+} \leftarrow \mathrm{Solve}[\Lambda, \mathcal{T}, r, W]$;
$(\mathrm{est}_{\mathrm{det},T})_{T\in\mathcal{T}}, \mathrm{est}_{\mathrm{det}} \leftarrow \mathrm{Estimate}_x[w_{\mathrm{disc}}, \Lambda, \mathcal{T}, p]$;
$(\mathrm{est}_{\mathrm{param},m})_{m\in\mathbb{N}}, \mathrm{est}_{\mathrm{param}} \leftarrow \mathrm{Estimate}_y[w_{\mathrm{disc}}, \Lambda]$;
$\mathrm{est}_{\mathrm{disc}} \leftarrow \mathrm{Estimate}_{\mathrm{ALS}}[w_{\mathrm{disc}}]$;
while $\mathrm{est}_{\mathrm{all}} > \epsilon_{\mathrm{TTASGFEM}}$ **do**
  if $\mathrm{est}_{\mathrm{det}} = \max\{\mathrm{est}_{\mathrm{det}}, \mathrm{est}_{\mathrm{param}}, \mathrm{est}_{\mathrm{disc}}\}$ **then**
    $\mathcal{T}_{\mathrm{mark}} \leftarrow \mathrm{Mark}_x[\theta, (\mathrm{est}_{\mathrm{det,T}})_{T\in\mathcal{T}}, \mathrm{est}_{\mathrm{det}}]$;
    $\mathcal{T}^{+} \leftarrow \mathrm{Refine}_x[\mathcal{T}, \mathcal{T}_{\mathrm{mark}}]$;
  end
  else if $\mathrm{est}_{\mathrm{param}} = \max\{\mathrm{est}_{\mathrm{det}}, \mathrm{est}_{\mathrm{param}}, \mathrm{est}_{\mathrm{disc}}\}$ **then**
    $\mathcal{I}_{\mathrm{mark}} \leftarrow \mathrm{Mark}_y[(\mathrm{est}_{\mathrm{param},m})_{m\in\mathbb{N}}, \mathrm{est}_{\mathrm{param}}]$;
    $\Lambda^{+} \leftarrow \mathrm{Refine}_y[\Lambda, \mathcal{I}_{\mathrm{mark}}]$;
  end
  else
    $W^{+} \leftarrow \mathrm{Refine}_{\mathrm{TT}}[W]$;
  end
  $w_{\mathrm{disc}}^{+} \leftarrow \mathrm{Solve}[\Lambda, \mathcal{T}, r, W^{+}]$;
  $(\mathrm{est}_{\mathrm{det},T})_{T\in\mathcal{T}}, \mathrm{est}_{\mathrm{det}} \leftarrow \mathrm{Estimate}_x[w_{\mathrm{disc}}, \Lambda, \mathcal{T}, p]$;
  $(\mathrm{est}_{\mathrm{param},m})_{m\in\mathbb{N}}, \mathrm{est}_{\mathrm{param}} \leftarrow \mathrm{Estimate}_y[w_{\mathrm{disc}}, \Lambda]$;
  $\mathrm{est}_{\mathrm{disc}} \leftarrow \mathrm{Estimate}_{\mathrm{ALS}}[w_{\mathrm{disc}}]$;
end

- In case the parametric error $\mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}})$ dominates, we use estimators $\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}})$ in order to determine which components need to be refined. Here, we also mark until the Dörfler property is satisfied, that is, we obtain a subset $\mathcal{I}_{\mathrm{mark}} \subset \mathbb{N}$ such that

$$\sum_{m\in\mathcal{I}_{\mathrm{mark}}} \mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}}) \geq \theta\, \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}}).$$

  This is the marking

$$\mathcal{I}_{\mathrm{mark}} \leftarrow \mathrm{Mark}_y[(\mathrm{est}_{\mathrm{param},m}(w_{\mathrm{disc}}))_{m\in\mathbb{N}}, \mathrm{est}_{\mathrm{param}}(w_{\mathrm{disc}})]$$

  and we refine by increasing $q_m^{+} \leftarrow q_m + 1$ for $m \in \mathcal{I}_{\mathrm{mark}}$

$$\Lambda^{+} \leftarrow \mathrm{Refine}_y[\Lambda, \mathcal{I}_{\mathrm{mark}}];$$

- Finally, if $\mathrm{est}_{\mathrm{disc}}(w_{\mathrm{disc}})$ exceeds the other errors, we simply add a tensor of rank 1 to the solution tensor W. This increases all TT ranks of W by 1 almost surely, as one can easily see. It would also be possible to add an approximation of the discrete residual as this is also in TT format. However, since the ALS algorithm will be performed after the refinement, the advantage of this rather costly improvement has shown to be negligible [85]. Thus we get

$$W^{+} \leftarrow \mathrm{Refine}_{\mathrm{TT}}[W].$$

A single iteration step of the adaptive algorithm returns either a refined $\mathcal{T}^{+}$ or Λ^{+} or the tensor format solution with increased rank W^{+} and then solves the problem with these properties. This is done repeatedly until the overall error estimator $\mathrm{err}_{\mathrm{all}}(w_{\mathrm{disc}})$ is sufficiently small,

i.e. defined error bound $\epsilon_{\text{TTASGFEM}}$ or a maximum problem size is reached. This procedure is given by the function TTASGFEM, displayed in Algorithm 4. The upper error bounds directly follow from Corollary 7.6 and Theorem 7.8. Numerical simulations are performed in the next and final chapter.

8 Numerical Experiments

In this final chapter of the thesis, we present numerical experiments for several of the discussed algorithms in the previous chapters. They are implemented with the open source framework *ALEA* [18] which has already been used for the ASGFEM presented in [15, 16, 17], the *FEniCS* FE framework [22] and *ttpy* [66], an open source toolbox for the TT format.

8.1 Error Sampling

We have explained in Chapter 4 that Monte-Carlo sampling can be used to approximate some moments of the solution u. In the same way, it may be employed for experimental verification of the reliability of the error estimator, by approximating the mean error of the parametric solution. For this, a set of P_{MC} independent realisations $y_{[j]} \in \Gamma, j = 1, \dots, P_{\mathrm{MC}}$ of the stochastic parameters is determined. The $y_{[j],m}$ are sampled according to the probability measure π of the (vector-valued) random variable $y \in \Gamma$. Note that in the log-normal case, we have shown that a random draw $y_{[j]}$ is in Γ almost surely, see Lemma 3.4.

The mean-square error of the (inexact) discrete solution $w_{\mathrm{disc}} \in \mathcal{V}_{\mathrm{disc}}$ is approximated by a Monte Carlo sample average

$$\begin{aligned} \mathbb{E}(\|u(\cdot) - w_{\mathrm{disc}}(\cdot)\|_{\mathcal{X}}^2) &= \int_\Gamma \|u(y) - w_{\mathrm{disc}}(y)\|_{\mathcal{X}}^2 \,\mathrm{d}\pi(y) \\ &\approx \frac{1}{P_{\mathrm{MC}}} \sum_{i=1}^{P_{\mathrm{MC}}} \|\tilde{u}(y_{[j]}) - w_{\mathrm{disc}}(y_{[j]})\|_{\mathcal{X}}^2. \end{aligned} \tag{8.1}$$

The sampled solutions $\tilde{u}(y_{[j]})$ are approximations of the exact $u(y_{[j]}) = \mathcal{A}(y_{[j]})^{-1} f$ since the differential operator is discretised on a fine reference mesh. This is obtained by another uniform refinement of the adapted mesh generated from the FEM discretisation of the final iteration. Moreover, the truncated expansion in the random field $a(x, y)$ is expanded by the trailing largest 200 terms which are not considered by the best approximate parametric solution, i.e. we truncate the coefficient at $M + 200$ when we sample. We choose $P_{\mathrm{MC}} = 150$ for the Monte Carlo sampling of the reference error (8.1) which proved to be sufficient to assess the reliability of the error estimator.

Note that for the log-normal case, our solution w_{disc} corresponds to the discretised operator $\mathcal{A}_+$, but we compare it to the sample solutions of the deterministic operator $\mathcal{A}(y_{[j]})$. Thus, the error that results from the approximation of the operator is included in the mean error that we sample here.

8.2 Experiments

First, we compare Algorithm 3 to the naïve approach of calculating the discretisation of the coefficient for each refinement. Next, we observe that the preconditioner of Chapter 6.2 in the log-normal case does indeed lead to faster convergence of the ALS algorithm. Finally, we discuss convergence of the TTASGFEM algorithm 4 for specific coefficients.

Coefficient Splitting

We perform a splitting of the coefficient

$$a(x, y) = \prod_{m=1}^{\infty} \exp(b_m(x) y_m)$$

with

$$b_m(x) := m^{-2} \cos(2\pi\kappa_1(m)x_1) \cos(2\pi\kappa_2(m)x_2)$$

and

$$\kappa_1(m) = m - k(m)(k(m)+1)/2 \quad \text{and} \quad \kappa_2(m) = k(m) - \kappa_1(m) \tag{8.2}$$

with $k(m) = \lfloor -1/2 + \sqrt{1/4 + 2m} \rfloor$, i.e. the coefficient functions b_m enumerate all planar Fourier sine modes in increasing total order.

We then calculate the discretised coefficient accoring to Algorithm 3. We set the length $L = 10$ and the GPC degree to 8. Figure 15 shows the development of the sampled error

$$\mathbb{E}(\|a - a_{\Delta,s}\|_{\mathcal{X}}^2) \approx \frac{1}{P_{\mathrm{MC}}} \sum_{i=1}^{P_{\mathrm{MC}}} \sum_{p=1}^{P_{\mathcal{X}}} \big(a(x_p, y_{[j]}) - a_{\Delta,s}(x_p, y_{[j]})\big)^2.$$

for P_{MC} random samples $y_{[j]} \in \Gamma, j = 1, \ldots, P_{\mathrm{MC}}$ and $P_{\mathcal{X}}$ grid points $x_p \in D, p = 1, \ldots, P_{\mathcal{X}}$.

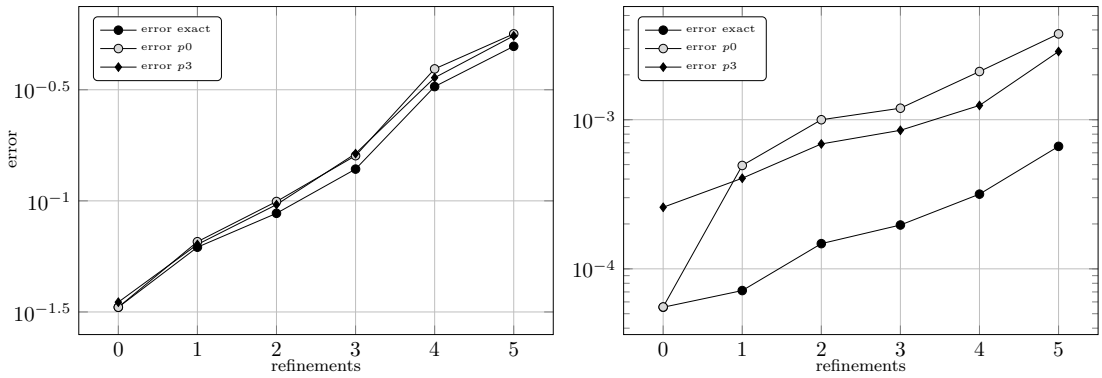

Figure 15: Coefficient splitting for polynomial quadrature of degrees 0 and 3 compared with the exact method. Ranks $s \equiv 20$ (left) and $s \equiv 80$ (right).

The figure on the left shows the error for a low rank $s \equiv 20$. We calculate the error both for a high and a low polynomial degree in the quadrature ($p3$ for degree 3 and $p0$ for degree 0) and compare this to a direct splitting at the points x_p. For each grid refinement, the number of grid points $P_{\mathcal{X}}$ is increased and the direct splitting is performed again. As expected, the direct splitting is always the most accurate, followed by the splitting with finer quadrature. The error increases over the refinements because the rank s is kept fixed while the grid gets finer. One observes that the rank is the limiting factor for all three splitting methods.

Therefore, we perform the same refinement strategy for higher ranks $s \equiv 80$ in the right figure. The other variables are kept the same. The error is much smaller and the difference between the errors is more pronounced. The direct splitting in every step is now an order of magnitude more accurate than the other errors, while the finer quadrature outperforms the coarse quadrature as expected. We can see that the coarse quadrature with constant polynomials is exactly the direct strategy when the quadrature points are the grid points, i.e. at refinement 0. Again, the rank s is the limiting factor over the refinements and one expects a better approximation for higher ranks.

Preconditioning

Here, we only look at one application of the ALS algorithm and we show the effect of the preconditioner. This is done only for the log-normal case, as preconditioning of the affine case is well known. We use the discretised log-normal coefficient as in (6.6) resulting from the above experiment for coefficient splitting.

Figure 16 shows the number of micro-iterations necessary for solving the first subproblem

$$V_1^+ = \left(E_1^{\mathsf{T}} \mathbf{A} E_1\right)^{-1} E_1^{\mathsf{T}} F$$

as in (5.19). This is done using a CG-method, once using the preconditioner explained in 6.2 and once without using it. The number of micro-iterations is significantly higher without the use

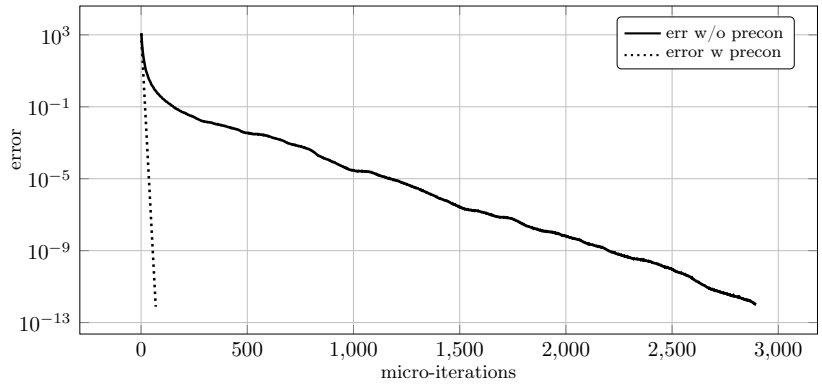

Figure 16: Test for the log-normal preconditioner for a tensor of sizes $M = 15$, ranks $r \equiv 5$ and GPC-degrees $q_m = 5$ for all $m = 1, \dots, M$. Error against number of micro-iterations in the first (deterministic) component of the first ALS sweep.

of our preconditioner. For the other subproblems, this effect was not as visible, as the system sizes are too small (the components are $V_m \in \mathbb{R}^{5\times5\times5}$ for all $m = 1, \dots, M-1$). The overall number of ALS sweeps was the same both with and without preconditioning. However, since the first subproblem is by far the largest and most time consuming, an efficient implementation of the preconditioning improves the performance of the overall algorithm significantly.

The Affine Diffusion Equation

We examine numerical simulations for the stationary diffusion problem (3.9) in a plane, polygonal domain $D \subset \mathbb{R}^2$. As in [17], the expansion coefficients of the affine parametric field (3.11) are given by

$$a_m(x) := \alpha_m \cos(2\pi\kappa_1(m)x_1)\cos(2\pi\kappa_2(m)x_2) \tag{8.3}$$

where $\alpha_m = \alpha_0 m^{-\beta}$ with $\beta > 1$ and some $0 < \alpha_0 < 1/\zeta(\beta)$ with the Riemann zeta function ζ. Then, (3.12) holds with $\lambda = \alpha_0\zeta(\beta)$. Moreover, κ_1 and κ_2 are defined as above in (8.2). To illustrate the influence which the stochastic coefficient plays in the adaptive algorithm, we examine the expansion with slow and fast decay of α_m, setting β in (8.3) to either 2 or 4. The computations are carried out with conforming FEM spaces of polynomial degrees 1 and 3. For the adaptive algorithm TTASGFEM of Chapter 7.2, the marking parameters are $\theta_x = \theta_y = 1/2$.

We solve the stationary diffusion equation (3.9) both on the unit square $D = (0,1)^2$ and on the L-shaped domain $D = (-1,1)^2 \setminus (0,1) \times (-1,0)$ with homogeneous Dirichlet boundary conditions and with right-hand side $f \equiv 1$. It is well-known that in the latter case, the solution exhibits a singularity at the reentrant corner at $(0,0)$ which has to be resolved by a pronounced mesh refinement in its vicinity in order to achieve optimal convergence rates. The refinement of the meshes for the square and the L-shaped domain is shown in Figure 17.

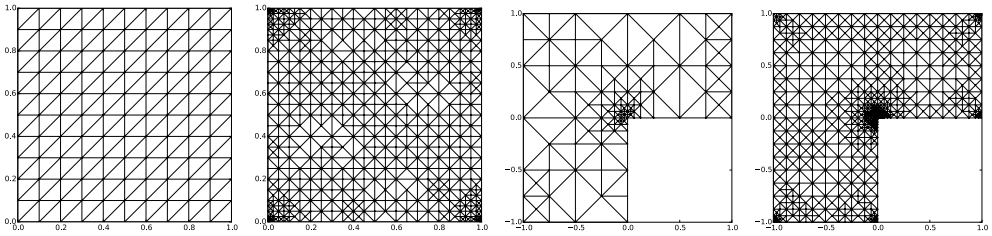

Figure 17: Adaptively refined meshes for square and L-shaped domains with $\beta = 4$. Iterations 2 and 47 (square), and 12 and 43 (L-shaped).

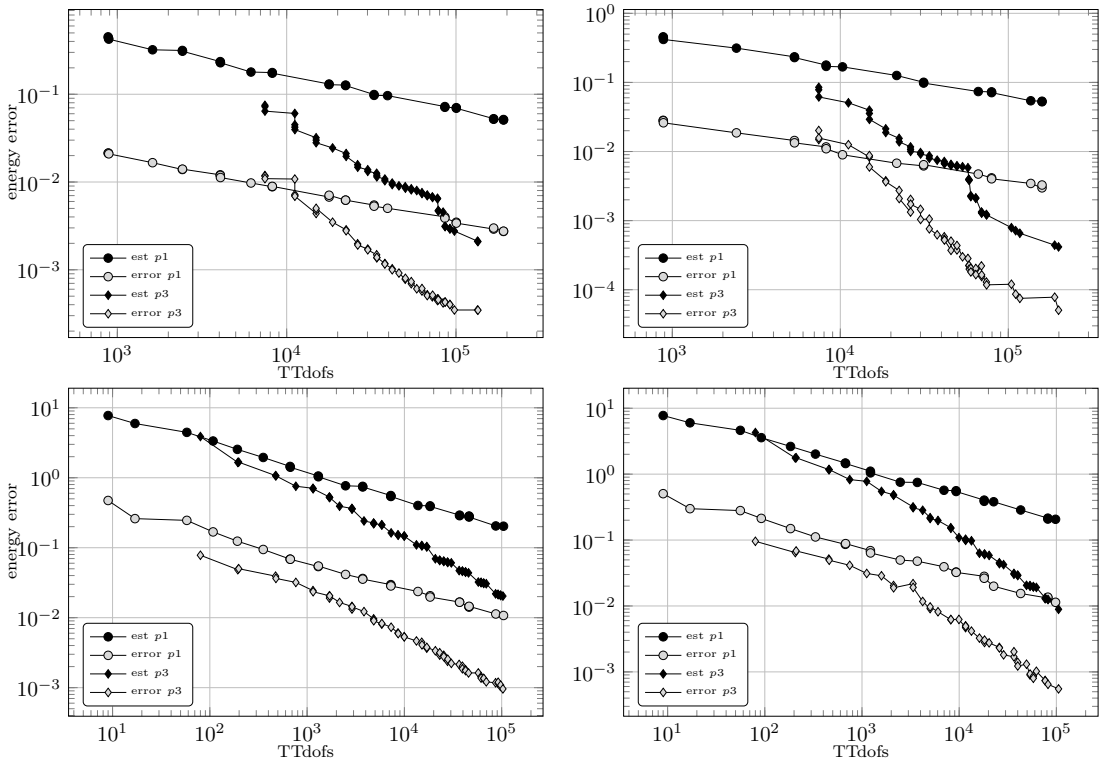

Figure 18: Convergence of the error estimator in the energy norm with FEM of degree $p = 1, 3$ for the stationary diffusion problem on the square (top) and the L-shaped (bottom) domain with homogeneous Dirichlet boundary conditions for slow ($\beta = 2$, left) and fast ($\beta = 4$, right) decay. Error estimator and sampled error for TT degrees of freedom.

The results of the adaptive algorithm *TTASGFEM* for a slow decay of the coefficients with $\beta = 2$ and a fast decay with $\beta = 4$ are depicted in Figure 18. The amplitude α_0 in (8.3) was chosen as $\lambda/\zeta(\tilde{\sigma})$ with $\lambda = 0.9$, resulting in $\alpha_0 \approx 0.547$ for $\beta = 2$ and $\alpha_0 \approx 0.832$ for $\beta = 4$. The figure shows the error estimator and the reference error obtained by Monte Carlo sampling as described in Section 8.1 for square (top) and L-shaped (bottom) domain against the TT degrees of freedom, i.e. the number of entries in the TT solution tensor W. These are calculated as follows:

$$\text{TTdofs} = Nr_0 + \sum_{m=1}^{M} r_{m-1} q_m r_m.$$

In all cases, the error estimator exhibits the same decay rate as the actual energy error. This is the largest for $p = 3$, as expected. In particular, the obtained accuracy for $p = 3$ is significantly higher than for the first order FE approximation $p = 1$. With faster decay of the coefficient (right column), the attained error levels are better than for slower decay (left column).

The degree of the Legendre chaos polynomials for the first 30 stochastic dimensions is depicted in Figure 19. Since the results are similar, this is done only for the L-shaped domain. As expected, for $p = 1$ the maximal polynomial degree is relatively small and only few dimensions are active. Opposite to this, for $p = 3$ the stochastic error dominates and the stochastic dimensions are thus increased drastically. In case of fast decay, the first stochastic dimension is discretised with polynomials up to degree 10. This is smaller in case of slow decay, namely degree 5. However, the number of active dimensions then is about 130.

We perform another example, in which we investigate the influence the TT rank has on the accuracy of the discrete solution. We use a square domain and determine the discrete solution u_{disc} with $p = 1, 2, 3$ FEM, $M = 30$ stochastic dimensions and a slow decay rate $\beta = 2$ in the

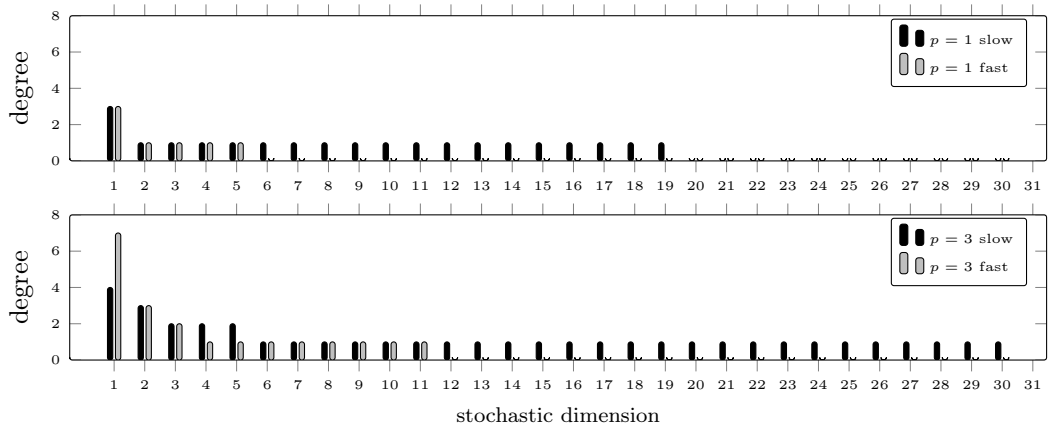

Figure 19: Polynomial degree for each stochastic mode with FEM of degree $p = 1, 3$ (top and bottom) for the stationary diffusion problem on the L-shaped domain with homogeneous Dirichlet boundary conditions for slow ($\beta = 2$) and fast ($\beta = 4$) decay.

coefficient representation. The stochastic variables are discretised with a uniform polynomial degree of 3. Figure 20 depicts the mean square energy error of the discrete solution subject to the rank of the low-rank TT representation. With increasing rank, the number of degrees of freedoms (TTdofs) continuously increased, starting with rank 1 and ending at about 10^5 dofs in the shown graphs. We compare the error progression with increasing ranks for each polynomial FE degree for a coarser and a finer grid.

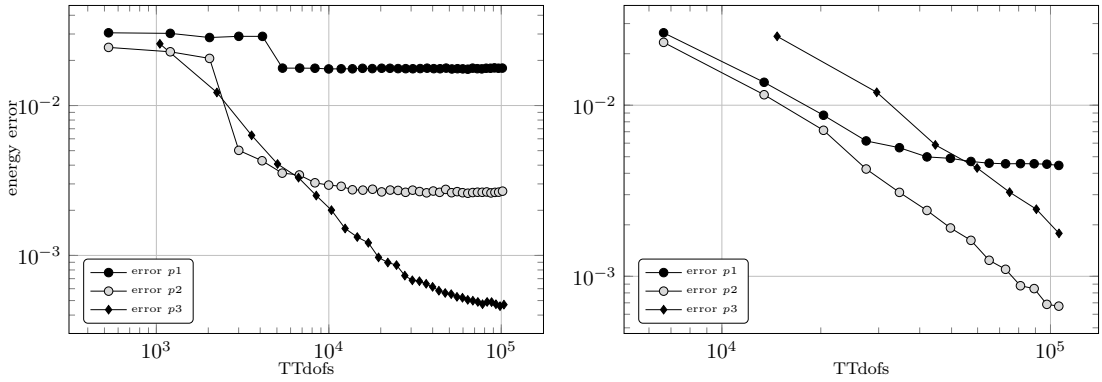

Figure 20: Convergence of energy error for sole rank increase of $p = 1, 2, 3$ solution on square domain with slow decay ($\beta = 2$), isotropic stochastic degree 3 and $M = 30$ stochastic dimensions. Computation on coarse grid (left) and finer grid (right) with starting rank 1 and subsequent increase by 1 in each iteration step until the degrees of freedom exceed 10^5.

The left graph in Figure 20 depicts the experiments with the coarser grid, the right graph shows the results on the finer grid. It can clearly be observed that the error crucially depends on the rank of the representation. With the coarser grid, a constant error decay can only be observed for $p = 3$. The solutions with $p = 1, 2$ show degraded error reduction for increasing ranks which is due to the prevailing FE approximation error which in these cases is dominant on the coarse mesh (left graph). This behaviour changes on a finer grid (right graph). Since the FE approximation error is reduced in for all FE polynomial degrees, the error can be decreased for all $p = 1, 2, 3$. While $p = 2$ now exhibits a constant decay rate, $p = 1$ levels off once the FE error dominates again. This point is reached with rank 4.

A consequence we can infer from these observations is that, in order to obtain a fully adaptivity approximation scheme, also the tensor rank has to be updated according to the

required accuracy of the numerical solution. Hence, this is included in the adaptive algorithm presented in Section 7.2. However, this experiment shows that the discrete error is not dominant under all circumstances and an equilibration of the errors is required. Then, the ranks remain relatively moderate.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. With a foreword by Paul Van Dooren.

[2] P.-A. Absil and I. V. Oseledets. Low-rank retractions: a survey and new results. *Comput. Optim. Appl.*, 62(1):5–29, 2015.

[3] Markus Bachmayr, Albert Cohen, Dinh Dũng, and Christoph Schwab. Fully discrete approximation of parametric and stochastic elliptic PDEs. *SIAM J. Numer. Anal.*, 55(5):2151–2186, 2017.

[4] Markus Bachmayr, Albert Cohen, and Wolfgang Dahmen. Parametric PDEs: Sparse or low-rank approximations? *IMA Journal of Numerical Analysis*, 2017. To appear.

[5] Markus Bachmayr, Reinhold Schneider, and André Uschmajew. Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations. *Found. Comput. Math.*, 16(6):1423–1472, 2016.

[6] Jonas Ballani and Lars Grasedyck. Hierarchical tensor approximation of output quantities of parameter-dependent PDEs. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):852–872, 2015.

[7] Jonas Ballani and Daniel Kressner. Reduced basis methods: from low-rank matrices to low-rank tensors. *SIAM J. Sci. Comput.*, 38(4):A2045–A2067, 2016.

[8] S. Bochner. Integration von funktionen, deren werte die elemente eines vektorraumes sind. *Fundamenta Mathematicae*, 20(1):262–176, 1933.

[9] 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.

[10] R. H. Cameron and W. T. Martin. The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals. *Ann. of Math. (2)*, 48:385–392, 1947.

[11] Albert Cohen, Ronald Devore, and Christoph Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl. (Singap.)*, 9(1):11–47, 2011.

[12] Sergey Dolgov, Boris N. Khoromskij, Alexander Litvinenko, and Hermann G. Matthies. Polynomial chaos expansion of random coefficients and the solution of stochastic partial differential equations in the tensor train format. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):1109–1135, 2015.

[13] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[14] Michael Eiermann, Oliver G. Ernst, and Elisabeth Ullmann. Computational aspects of the stochastic finite element method. *Comput. Vis. Sci.*, 10(1):3–15, 2007.

[15] Martin Eigel, Claude Jeffrey Gittelson, Christoph Schwab, and Elmar Zander. Adaptive stochastic Galerkin FEM. *Comput. Methods Appl. Mech. Engrg.*, 270:247–269, 2014.

[16] Martin Eigel, Claude Jeffrey Gittelson, Christoph Schwab, and Elmar Zander. A convergent adaptive stochastic Galerkin finite element method with quasi-optimal spatial meshes. *ESAIM Math. Model. Numer. Anal.*, 49(5):1367–1398, 2015.

[17] Martin Eigel, Max Pfeffer, and Reinhold Schneider. Adaptive stochastic Galerkin FEM with hierarchical tensor representations. *Numer. Math.*, 136(3):765–803, 2017.

[18] Martin Eigel and Elmar Zander. `alea` - A Python Framework for Spectral Methods and Low-Rank Approximations in Uncertainty Quantification, https://bitbucket.org/aleadev/alea.

[19] Mike Espig, Wolfgang Hackbusch, and Aram Khachatryan. On the convergence of alternating least squares optimisation in tensor format representations. 05 2015.

[20] Mike Espig, Wolfgang Hackbusch, Alexander Litvinenko, Hermann G. Matthies, and Philipp Wähnert. Efficient low-rank approximation of the stochastic Galerkin matrix in tensor formats. *Comput. Math. Appl.*, 67(4):818–829, 2014.

[21] Antonio Falcó, Wolfgang Hackbusch, and Anthony Nouy. Geometric structures in tensor representations (release 2). Preprint, Max Planck Institute for Mathematics in the Sciences, 2014.

[22] FEniCS Project - Automated solution of Differential Equations by the Finite Element Method, http://fenicsproject.org.

[23] J. Galvis and M. Sarkis. Approximating infinity-dimensional stochastic Darcy's equations without uniform ellipticity. *SIAM J. Numer. Anal.*, 47(5):3624–3651, 2009.

[24] C. J. Gittelson. Stochastic Galerkin discretization of the log-normal isotropic diffusion problem. *Math. Models Methods Appl. Sci.*, 20(2):237–263, 2010.

[25] Claude Jeffrey Gittelson. Stochastic Galerkin approximation of operator equations with infinite dimensional noise. Technical Report 2011-10, Seminar for Applied Mathematics, ETH Zürich, 2011.

[26] Lars Grasedyck. Hierarchical singular value decomposition of tensors. *SIAM J. Matrix Anal. Appl.*, 31(4):2029–2054, 2009/10.

[27] Lars Grasedyck and Wolfgang Hackbusch. An introduction to hierarchical h-rank and TT-rank of tensors with examples. *Comput. Methods Appl. Math.*, 11(3):291–304, 2011.

[28] Christian Grossmann and Hans-Görg Roos. *Numerical treatment of partial differential equations*. Universitext. Springer, Berlin, 2007. Translated and revised from the 3rd (2005) German edition by Martin Stynes.

[29] W. Hackbusch and S. Kühn. A new scheme for the tensor representation. *J. Fourier Anal. Appl.*, 15(5):706–722, 2009.

[30] Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42 of *Springer series in computational mathematics*. Springer, Heidelberg, 2012.

[31] Wolfgang Hackbusch and Reinhold Schneider. Tensor spaces and hierarchical tensor representations. In *Extraction of quantifiable information from complex systems*, volume 102 of *Lect. Notes Comput. Sci. Eng.*, pages 237–261. Springer, Cham, 2014.

[32] Mohammad Hadigol, Alireza Doostan, Hermann G. Matthies, and Rainer Niekamp. Partitioned treatment of uncertainty in coupled domain problems: a separated representation approach. *Comput. Methods Appl. Mech. Engrg.*, 274:103–124, 2014.

[33] Jutho Haegeman, Michaël Mariën, Tobias J Osborne, and Frank Verstraete. Geometry of matrix product states: metric, parallel transport, and curvature. *Journal of Mathematical Physics*, 55(2):50, 2014.

[34] Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *J. ACM*, 60(6):Art. 45, 39, 2013.

[35] Frank L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[36] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. The alternating linear scheme for tensor optimization in the tensor train format. *SIAM J. Sci. Comput.*, 34(2):A683–A713, 2012.

[37] Sebastian Holtz, Thorsten Rohwedder, and Reinhold Schneider. On manifolds of tensors of fixed TT-rank. *Numer. Math.*, 120(4):701–731, 2012.

[38] Kari Karhunen. über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 1947(37):79, 1947.

[39] Vladimir Kazeev, Oleg Reichmann, and Christoph Schwab. Low-rank tensor structure of linear diffusion operators in the TT and QTT formats. *Linear Algebra Appl.*, 438(11):4204–4221, 2013.

[40] B. N. Khoromskij and I. Oseledets. Quantics-TT collocation approximation of parameter-dependent and stochastic elliptic PDEs. *Comput. Methods Appl. Math.*, 10(4):376–394, 2010.

[41] Boris N. Khoromskij and Christoph Schwab. Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs. *SIAM J. Sci. Comput.*, 33(1):364–385, 2011.

[42] Valentin Khrulkov and Ivan Oseledets. Desingularization of bounded-rank matrix sets. arXiv preprint 1612.03973, 2016.

[43] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009.

[44] Daniel Kressner, Michael Steinlechner, and Bart Vandereycken. Low-rank tensor completion by Riemannian optimization. *BIT*, 54(2):447–468, 2014.

[45] Benjamin Kutschan. Tangent cones to TT varieties. *CoRR*, abs/1705.10152, 2017.

[46] J. M. Landsberg. *Tensors: geometry and applications*, volume 128 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[47] Joseph M. Landsberg, Yang Qi, and Ke Ye. On the geometry of tensor network states. *Quantum Inf. Comput.*, 12(3-4):346–354, 2012.

[48] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, March 2000.

[49] P. D. Lax and A. N. Milgram. Parabolic equations. In *Contributions to the theory of partial differential equations*, Annals of Mathematics Studies, no. 33, pages 167–190. Princeton University Press, Princeton, N. J., 1954.

[50] Jeffrey M. Lee. *Manifolds and differential geometry*, volume 107 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2009.

[51] Ö. Legeza, J. Röder, and B. A. Hess. Controlling the accuracy of the density-matrix renormalization-group method: The dynamical block state selection approach. *Phys. Rev. B*, 67:125114, Mar 2003.

[52] Zhening Li, André Uschmajew, and Shuzhong Zhang. On convergence of the maximum block improvement method. *SIAM J. Optim.*, 25(1):210–233, 2015.

[53] Michel Loève. *Probability theory. II.* Springer-Verlag, New York-Heidelberg, fourth edition, 1978. Graduate Texts in Mathematics, Vol. 46.

[54] Christian Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, 2008.

[55] Christian Lubich and Ivan V. Oseledets. A projector-splitting integrator for dynamical low-rank approximation. *BIT*, 54(1):171–188, 2014.

[56] Christian Lubich, Ivan V. Oseledets, and Bart Vandereycken. Time integration of tensor trains. *SIAM J. Numer. Anal.*, 53(2):917–941, 2015.

[57] Christian Lubich, Thorsten Rohwedder, Reinhold Schneider, and Bart Vandereycken. Dynamical approximation by hierarchical Tucker and tensor-train tensors. *SIAM J. Matrix Anal. Appl.*, 34(2):470–494, 2013.

[58] Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Comput. Statist.*, 29(3-4):591–621, 2014.

[59] Antje Mugler. *Verallgemeinertes polynomielles Chaos zur Lösung stationärer Diffusionsprobleme mit zufälligen Koeffizienten*. PhD thesis, BTU Cottbus, 2013.

[60] Antje Mugler and Hans-Jörg Starkloff. On elliptic partial differential equations with random coefficients. *Stud. Univ. Babeş-Bolyai Math.*, 56(2):473–487, 2011.

[61] Antje Mugler and Hans-Jörg Starkloff. On the convergence of the stochastic Galerkin method for random elliptic partial differential equations. *ESAIM Math. Model. Numer. Anal.*, 47(5):1237–1263, 2013.

[62] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[63] I. V. Oseledets. Tensor-train decomposition. *SIAM J. Sci. Comput.*, 33(5):2295–2317, 2011.

[64] I. V. Oseledets and E. E. Tyrtyshnikov. Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM J. Sci. Comput.*, 31(5):3744–3759, 2009.

[65] Ivan Oseledets, Eugene Tyrtyshnikov, and Nickolai Zamarashkin. Tensor-train ranks for matrices and their inverses. *Comput. Methods Appl. Math.*, 11(3):394–403, 2011.

[66] Ivan V. Oseledets. ttpy - A Python Implementation of the TT-Toolbox, https://github.com/oseledets/ttpy.

[67] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac. Matrix product state representations. *Quantum Info. Comput.*, 7(5):401–430, jul 2007.

[68] Michael Reed and Barry Simon. *Methods of modern mathematical physics. I.* Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, second edition, 1980. Functional analysis.

[69] Thorsten Rohwedder and André Uschmajew. On local convergence of alternating schemes for optimization of convex problems in the tensor train format. *SIAM J. Numer. Anal.*, 51(2):1134–1162, 2013.

[70] Berkant Savas and Lek-Heng Lim. Quasi-Newton methods on Grassmannians and multilinear approximations of tensors. *SIAM J. Sci. Comput.*, 32(6):3352–3393, 2010.

[71] Erhard Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. *Math. Ann.*, 63(4):433–476, 1907.

[72] Reinhold Schneider and André Uschmajew. Approximation rates for the hierarchical tensor format in periodic Sobolev spaces. *J. Complexity*, 30(2):56–71, 2014.

[73] Reinhold Schneider and André Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via łojasiewicz inequality. *SIAM J. Optim.*, 25(1):622–646, 2015.

[74] Ulrich Schollwöck. The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1):96 – 192, 2011. January 2011 Special Issue.

[75] Christoph Schwab and Claude Jeffrey Gittelson. Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs. *Acta Numer.*, 20:291–467, 2011.

[76] Michael Steinlechner. Riemannian optimization for high-dimensional tensor completion. *SIAM J. Sci. Comput.*, 38(5):S461–S484, 2016.

[77] Szilárd Szalay. Quantum entanglement in finite-dimensional hilbert spaces. *arXiv [quant-ph]*, page 1302.4654, 2013.

[78] Szilárd Szalay, Max Pfeffer, Valentin Murg, Gergely Barcza, Frank Verstraete, Reinhold Schneider, and Örs Legeza. Tensor product methods and entanglement optimization for ab initio quantum chemistry. *International Journal of Quantum Chemistry*, pages 1342–1391, 2015.

[79] Ledyard R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.

[80] Elisabeth Ullmann. *Solution Strategies for Stochastic Finite Element Discretizations*. PhD thesis, TU Bergakademie Freiberg, 2008.

[81] Elisabeth Ullmann. A Kronecker product preconditioner for stochastic Galerkin finite element discretizations. *SIAM J. Sci. Comput.*, 32(2):923–946, 2010.

[82] André Uschmajew. Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM J. Matrix Anal. Appl.*, 33(2):639–652, 2012.

[83] André Uschmajew. *Zur Theorie der Niedrigrangapproximation in Tensorprodukten von Hilberträumen*. PhD thesis, TU Berlin, 2013.

[84] André Uschmajew and Bart Vandereycken. The geometry of algorithms using hierarchical tensors. *Linear Algebra Appl.*, 439(1):133–166, 2013.

[85] André Uschmajew and Bart Vandereycken. Line-search methods and rank increase on low-rank matrix varieties. *preprint*, juli 2014. Accepted for NOLTA2014.

[86] Bart Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013.

[87] F. Verstraete, D. Porras, and J. I. Cirac. Density matrix renormalization group and periodic boundary conditions: A quantum information perspective. *Phys. Rev. Lett.*, 93:227205, Nov 2004.

[88] Guifré Vidal. Efficient classical simulation of slightly entangled quantum computations. *Phys. Rev. Lett.*, 91:147902, Oct 2003.

[89] Haobin Wang and Michael Thoss. Multilayer formulation of the multiconfiguration time-dependent Hartree theory. *The Journal of Chemical Physics*, 119(3):1289–1299, 2003.

[90] Steven R. White. Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.*, 69:2863–2866, Nov 1992.

[91] Norbert Wiener. The Homogeneous Chaos. *Amer. J. Math.*, 60(4):897–936, 1938.

Abstract

This thesis deals with tensor methods for the numerical solution of parametric partial differential equations (PDEs). Because of their parametric dependence, these differential equations exhibit a high dimensionality. Numerical methods like the Galerkin method quickly reach their limits, since the resulting linear system grows exponentially with the number of parameters. This is known as the *curse of dimensionality*.

The work can be roughly divided into three parts. The first part covers the theory of parametric PDEs using the leading example of a diffusion equation with parametric coefficients. We consider both a so called *affine* coefficient as well as a *log-normal* one. The necessary theoretical foundations for the numerical treatment of these problems are laid out.

The second part is a review of the state of the art of tensor methods. These have gained a lot of recognition in recent years. We introduce a number of tensor decomposition formats and discuss their advantages and disadvantages. In particular, we devise a general tensor representation and unite some concepts from mathematics and quantum physics where these formats have been known for a while. We highlight especially the methods for low rank approximation with tensors, because these will be used for the numerical solution of the parametric diffusion equation.

In the last part, we present our own results. We developed a number of numerical methods in order to facilitate the use of tensor methods especially in the log-normal case. Additionally, we present error estimators, both for the affine and the log-normal case, that allow for the comparison of the errors resulting from the different discretisations. These enable us to formulate an adaptive algorithm for the numerical solution of these equations. Finally, we illustrate the applicability of our methods with some numerical experiments.

Zusammenfassung

Diese Arbeit behandelt Tensormethoden für die numerische Lösung von parametrischen partiellen Differentialgleichungen (PDEs). Aufgrund ihrer parametrischen Abhängigkeit weisen diese Differentialgleichungen eine hohe Dimensionalität auf. Numerische Lösungsverfahren wie die Galkerkin Methode stoßen schnell an ihre Grenzen, da das resultierende Gleichungssystem mit der Anzahl der Parameter exponentiell wächst. Dies ist der sogenannte *Fluch der Dimension*.

Die Arbeit kann grob in drei Teile unterteilt werden. Der erste Teil behandelt die Theorie parametrischer PDEs am Beispiel einer Diffusionsgleichung mit parametrischem Koeffizienten. Wir betrachten den sogenannten *affinen* Fall eines Koeffizienten, sowie den *log-normalen* Fall. Die nötigen theoretischen Grundlagen für die numerische Lösung dieser Probleme werden gelegt.

Der zweite Teil ist eine Übersicht über den State of the Art der Tensormethoden. Diese haben in den letzten Jahren sehr an Zuspruch gewonnen. Wir führen einige Tensorzerlegungsformate ein und behandeln deren Vor- und Nachteile. Insbesondere formulieren wir eine allgemeine Form der Tensordarstellung und vereinigen Konzepte aus der Mathematik und der Quantenphysik, wo diese Zerlegungen schon länger bekannt sind. Wir gehen dann besonders auf Methoden zur Niedrigrangapproximation mit Tensoren ein, da diese zur Annäherung der Lösung der parametrischen Diffusionsgleichungen eingesetzt werden.

Im letzten Teil stellen wir unsere eigenen Ergebnisse vor. Es wurden einige numerische Verfahren entwickelt, um den Einsatz von Tensormethoden besonders im log-normalen Fall zu ermöglichen. Außerdem präsentieren wir Fehlerschätzer, sowohl für die affine als auch für die log-normale Diffusionsgleichung, mit deren Hilfe die unterschiedlichen Diskretisierungen verglichen werden können. Dies erlaubt die Formulierung eines adaptiven Algortithmus' für die numerischen Lösung der Gleichungen. Zuletzt verdeutlichen wir die Anwendbarkeit unserer Methoden mit Hilfe von numerischen Experimenten.