MSc Mathematics

Track: Dynamical Systems

Master thesis

Methods for Solving Parametric Parabolic Equations

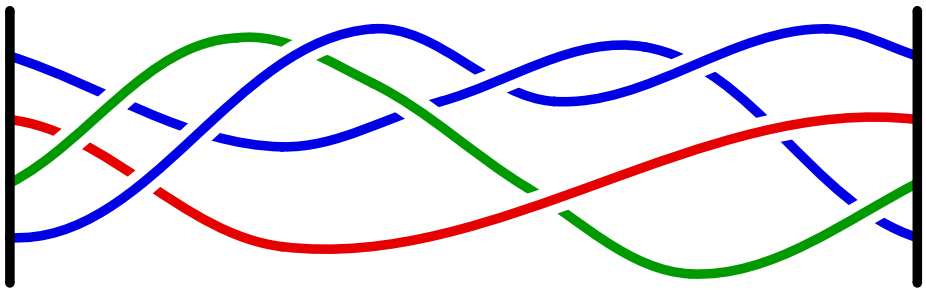
by

Harald Monsuur

February 5, 2021

Supervisor: prof.dr. Rob Stevenson

Second examiner: prof.dr. Joost Hulshof



Department of Mathematics

Faculty of Sciences

VU UNIVERSITY AMSTERDAM

Abstract

We discuss two methods for approximating solutions of parametric PDEs. These two methods are polynomial interpolation and the reduced basis method. Error estimates for both methods are given when we assume analyticity of the solution operators. Furthermore, we show how parabolic equations fit into the theoretical framework of the reduced basis method. This is achieved by providing a full-variational formulation for parabolic equations. Lastly, we test and compare both methods for an example problem.

Title: Methods for Solving Parametric Parabolic Equations
Author: Harald Monsuur, harald.monsuur@hotmail.com, 2584314
Supervisor: prof.dr. Rob Stevenson
Second examiner: prof.dr. Joost Hulshof
Date: February 5, 2021

Department of Mathematics
VU University Amsterdam
de Boelelaan 1081, 1081 HV Amsterdam
http://www.math.vu.nl/

Contents

Acknowledgements

I am thankful for the supervision of Rob Stevenson. The amount of time you have spent for me was more than sufficient. Being able to speak to you every week was outstanding, especially considering the current situation in the world. I enjoyed working with you. I must also thank my family and friends for supporting me during this time. I feel blessed because of all of you.

Introduction

In this thesis we describe two numerical methods for solving parametric PDEs using an online/offline decomposition. An offline/online decomposition is a decomposition of the workload for solving equations. The offline phase is typically performed on a supercomputer and the online phase on a computer or smartphone. Offline we are allowed to approximate solutions using high fidelity finite element methods and do other heavy computations, whereas online we are only allowed to do simple 'low-dimensional' calculations. The idea is that in the online phase we are able to find approximations of parametric PDEs very quickly for any parameter using the expensive work done in the offline phase.

Situations in which methods with an offline/online decomposition can be used arise very naturally. For example, in [20] the problem of supplying offshore platforms is mentioned. These offshore platforms need supplies that are provided by offshore supply vessels. The vessels must be able to maneuver and operate with a high precision; even in the worst conditions as conditions out at sea are often rough. In order for the vessels to keep their position and prevent drifting, difficult equations need to be solved. Obviously, the computations necessary might be too much to handle for a simple computer that is available on the vessel. Here, an offline/online decomposition, so that the heavy computations (offline) are done in advance, can prove to be useful.

In this thesis we consider the following setting. We assume that X is a Banach or Hilbert space, $a\colon X \times X \times \mathcal{P} \to \mathbb{R}$ a coercive parametric bilinear form, $f\colon X \times \mathcal{P} \to \mathbb{R}$ a parametric linear form and $\mathcal{P} \subset \mathbb{R}^p$ a compact parameter space. For $\mu \in \mathcal{P}$ we then have the problem of finding $u(\mu)$ such that

$$a(u(\mu), v; \mu) = f(v; \mu), \ v \in X. \tag{0.0.1}$$

Our focus in this thesis is on solution operators $\mu \mapsto u(\mu)$ that are holomorphic.

Online we are interested in finding approximations of $u(\mu)$ for any $\mu \in \mathcal{P}$. To accomplish this we assume the possibility of acquiring a set of approximations or 'snapshots' $u_H(\mu_i)$, $i = 1, \dots, n$. Here the μ_i, $i = 1, \dots, n$ can be chosen. In the online phase we use these solutions and parameters to obtain approximations of (0.0.1) for any parameter. Here we can expect the 'snapshots' to give enough information about the entire solution operator because we assume $\mu \mapsto u(\mu)$ to be holomorphic.

We describe two possible approaches of finding solutions using the 'snapshots' of u in the online phase. Firstly, we describe polynomial interpolation of the solution operator $\mu \mapsto u(\mu)$. This method tries to write $u(\mu)$ as a X-valued polynomial of μ that coincides with $u_H(\mu_i)$ for all $i = 1, \dots, n$. It turns out that the theory of interpolating holomorphic Banach valued operators is equivalent to the theory for interpolating real valued functions. Secondly, we describe the reduced basis method (RB-method). This method uses the 'snapshots' found in the offline phase to form an approximation space. We then use the Galerkin method to obtain approximations of $u(\mu)$ for all $\mu \in \mathcal{P}$. For this method it is also possible to find error bounds if we assume holomorphy of the solution operator.

Additionally, in this thesis we describe how the parabolic equation fits into this theoretical framework. To achieve this, we need to find an equation of the form (0.0.1) that describes the parabolic equation. This formulation of the parabolic equation also makes it possible to prove holomorphy of the solution map in many cases.

In the first chapter we will present the theory of polynomial interpolation in Banach spaces. In the second chapter we will present the reduced basis method. After that, we will provide a full-variational formulation for parabolic PDEs. Lastly, we will carry out some numerical experiments on two example problems.

1 Polynomial Interpolation

In this chapter we introduce ways of interpolating solution operators of PDEs. In fact, the theory we develop here is valid for any holomorphic Banach-valued operator. The reason for this is that the interpolation method is a black-box method. The interpolation method does not care about the operator, it only needs some of its evaluations. The fact that this method is a black box is an advantage and disadvantage at the same time. The advantage is that it is easy to understand and implement, the disadvantage is that it is expected to behave worse than other methods that do use the structure of operators.

It turns out that the theory of interpolating real-valued functions can be applied to Banach-valued functions. Hence, we mostly focus on interpolating real-valued functions in this chapter. In the last section we tie everything together using the Hahn-Banach theorem.

We first provide the theory for one-dimensional interpolation. After that we discuss ways of interpolating higher-dimensional functions. There we introduce the Smolyak algorithm and give an error estimate in two dimensions.

1.1 One-dimensional interpolation

In this section we review some of the standard facts of interpolating real functions of one variable. This topic has been studied extensively already, thus we only provide a summary in this thesis. We first introduce the Lagrange interpolation method, then we discuss the Lebesgue constant and appropriate choices for the nodes.

Definition 1.1. For any sequence of points $\overline{x} : a < x_0 < x_1 < \ldots < x_n < b$ in $[a, b]$ we define the Lagrange polynomials $l_i^n(x) = l_{i,\overline{x}}^n(x)$ as

$$l_i^n(x) := \prod_{j=0; i \neq j}^{n} \frac{x - x_j}{x_i - x_j}. \tag{1.1.1}$$

The Lagrange polynomial interpolation operator is then defined as $I_n = I_{n,\overline{x}} \colon C([a, b]) \to \mathbb{P}_n$

$$I_n g = \sum_{i=0}^{n} g(x_i) l_i^n. \tag{1.1.2}$$

For the error analysis we notice that the Lagrange polynomial interpolator is a projection operator on the space $\mathbb{P}_n$. From this we can deduce the following for $p \in \mathbb{P}_n$:

$$||g - I_n g||_\infty = ||g - p - (I_n g - p)||_\infty \leq ||g - p||_\infty + ||I_n(g - p)||_\infty \leq (1 + ||I_n||)||g - p||_\infty.$$

If we take the infimum over $p \in \mathbb{P}_n$ we find

$$||g - I_n g||_\infty \leq (1 + \Lambda_n(\overline{x}))E_n(g), \tag{1.1.3}$$

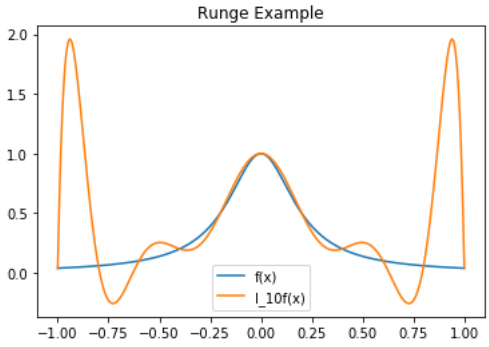


Figure 1.1: Runge's phenomenon

where $\Lambda_n(\overline{x}) := ||I_n||$ and $E_n(g) = \inf_{p\in\mathbb{P}_n} ||g-p||_\infty$. Hence, we see that the error analysis of an interpolation method is equivalent to finding the so-called Lebesgue constant Λ_n and the minimal polynomial error $E_n(g)$. Of course, the Lebesgue constant can be improved by picking appropriate nodes, whereas the minimal polynomial error is fixed for every function. Hence, the main task is finding suitable nodes $\overline{x} : a < x_0 < x_1 < \ldots < x_n < b$, such that the corresponding Lebesgue constant is small. If the Lebesgue constant grows to rapidly for $n \to \infty$ the error $||g - I_n g||_\infty$ might diverge. The Lebesgue constant can also be calculated (generally not easy) by maximizing the Lebesgue function, which is defined as

$$\lambda_n(x) = \sum_{i=0}^{n} |l_i^n(x)|. \tag{1.1.4}$$

Remark 1.2. In this thesis we try to approximate solution operators of an PDE. The 'evaluation' of this function is done using a finite element method which is not an exact method. This means that we are actually closer to interpolating the finite element solution operator F_H. Furthermore, computers make rounding errors. Hence, we are actually interpolating a perturbed operator $\tilde{F}$ satisfying $||\tilde{F} - F_H||_\infty \lesssim \epsilon_{machine}$. In the error analysis we can deal with this problem in two ways. Firstly, we can estimate the error in the following way:

$$||I_n\tilde{F} - F||_\infty = ||I_n\tilde{F} - I_nF + I_nF - F||_\infty \leq \Lambda_n||F - \tilde{F}||_\infty + (1+\Lambda_n)E_n(F).$$

The problem with this estimate is that we expect $||F - \tilde{F}||_\infty$ to be at least the error of the finite element method. Hence, for high values of n the term $\Lambda_n||F - \tilde{F}||_\infty$ can be reasonably large. Consequently $\Lambda_n||F - \tilde{F}||_\infty$ could be quite large also.

Another approach is to estimate the error in the following way:

$$\begin{aligned} ||I_n\tilde{F} - F||_\infty &\leq ||I_n\tilde{F} - F_H||_\infty + ||F_H - F||_\infty \\ &\leq \Lambda_n||F_H - \tilde{F}||_\infty + (1+\Lambda_n)E_n(F_H) + ||F_H - F||_\infty. \end{aligned}$$

Here $||\tilde{F} - F_H||_\infty$ is expected to be of the order machine precision. Hence, this error estimate is better provided $E_n(F_H)$ decays at the same rate as $E_n(F)$. This happens if both F_H and F are holomorphic, which is the case for the PDEs in Example 1.24 and Remark 3.10.

Example 1.3. A classical and illustrative example of the shortcomings of polynomial interpolation is Runge's phenomenon. Runge's phenomenon arises when we try to interpolate

$f(x) = \frac{1}{1+25x^2}$ on the interval $[-1, 1]$ with equidistant points. As Figure 1.1 indicates, there is no convergence of the interpolation polynomials to the function f. In fact, we notice divergence at the borders of the interval. Apparently, the Lebesgue constant increases faster than the error in the best polynomial approximation of f decays.

Below, there is an analysis from [6] which links the position of poles of functions with the convergence of interpolation methods. Notice that f has poles of order 1 at $\pm 0.2i$.

Let $K > 0$ be a real number such that f is holomorphic on $D = \{z \in \mathbb{C} : \text{dist}(z, [a, b]) < K\}$ and let $a \leq x_0 < x_1 \dots x_n \leq b$ be the interpolation nodes. Now, for some $x \in [a, b]$ define the function

$$g(z) = \frac{f(z)}{z - x} \prod_{j=0}^{n} \frac{x - x_j}{z - x_j},$$

A calculation shows that this function g has poles at the points x and $\{x_i\}_{i=0}^n$, with residuals $f(x)$ and $-f(x_i)l_i^n(x)$, $i = 0, \dots, n$. Hence, we have the following equality:

$$f(x) - I_n f(x) = \frac{1}{2\pi i} \int_{\partial D} \frac{f(z)}{z - x} \prod_{j=0}^{n} \frac{x - x_j}{z - x_j} dz.$$

This integral can be estimated by the length of ∂D ($|\partial D| = 2(b - a) + 2\pi K$) times an upper-bound of the integrand:

$$\begin{aligned} |f(x) - I_n f(x)| &= \left| \frac{1}{2\pi i} \int_{\partial D} \frac{f(z)}{z - x} \prod_{j=0}^{n} \frac{x - x_j}{z - x_j} dz \right| \\ &\leq \frac{(b - a + \pi K)M}{\pi} \left(\frac{b - a}{K} \right)^{n+1}. \end{aligned}$$

Here M is the maximum of f on ∂D. This rough inequality shows that there is convergence for all choices of nodes if $K > b - a$ when $n \to \infty$. This is clearly not the case for f at the interval $[-1, 1]$; this analysis only ensures convergence for intervals smaller than $(-0.1, -0.1)$.

Remark 1.4. Obviously, we can improve the estimation of the integral above by studying the maximum of the function $\pi(x) := \prod_{i=0}^{n}(x - x_i)$. The smaller this maximum, the smaller the interpolation error. For equidistant points this maximum is too large to interpolate the Runge function. If we choose the points differently, for example by using the roots of the Chebyshev polynomial, we get much better estimates.

1.1.1 Lebesgue constants for various interpolation nodes

In this section we provide some choices of nodes and the corresponding Lebesgue constants. The definitions and results of this section are from [7], [8], [9] and [10]. Proofs of these results are beyond the scope of this text. However, before we introduce some sets of nodes we first give an important property of the Lebesgue constant and then a universal lower bound for the Lebesgue constant. The important property of the Lebesgue constant is the invariance under linear transformations: if we have a set of nodes $\overline{x}$ for the interval $[-1, 1]$ and a set of nodes $a\overline{x}$ for the interval $[-a, a]$, the Lebesgue constants of the corresponding interpolation operators

are equal. This follows from the following calculation and noticing that the maximum of the Lebesgue functions are the same:

$$l^n_{i,a\overline{x}}(ax) = \prod_{j=0;i\neq j}^{n} \frac{ax - ax_j}{ax_i - ax_j} = \prod_{j=0;i\neq j}^{n} \frac{x - x_j}{x_i - x_j} = l^n_{i,\overline{x}}(x).$$

This means that we can 'forget' about the interval in the presentation of Lebesgue constants. We now give the universal lower bound:

Theorem 1.5. *For any choice of nodes $\overline{x}$ the following inequality for the Lebesgue constant holds:*

$$\Lambda_n(\overline{x}) > \frac{\log n}{8\sqrt{\pi}}.$$

A consequence of this theorem is that by the uniform boundedness principle for every set of nodes $\overline{x}$, because $||I_n|| = \Lambda_n(\overline{x}) \to \infty$, there exists a $f \in C[a,b]$ such that $||I_n f||_\infty \to \infty$. In other words, there does not exist a set of nodes that interpolates every continuous function accurately, when $n \to \infty$. Hence, we need more requirements on functions in order to prove convergence of interpolation methods.

Equidistant nodes

A 'bad' choice of nodes is the set of equidistant nodes. As we have seen earlier, the Runge Phenomenom shows that this set of nodes can fail for infinitely differentiable functions. The Lebesgue constant Λ_n grows exponentially with the asymptotic estimate

$$\Lambda_n \cdot \left(\frac{2^{n+1}}{e \cdot n(\log n + \gamma)} \right)^{-1} \to 1 \text{ for } n \to \infty,$$

where $\gamma \approx 0.577$ is Euler's constant. This growth rate is much worse than the universal lower bound given in Theorem 1.5. The biggest error in equidistant interpolation is often seen at the borders. We will not be using these nodes.

Chebyshev nodes

The Chebyshev polynomials on $[-1,1]$ are defined by $C_n(\cos(t)) = \cos(nt)$ or $C_n(x) = \cos(n \cos^{-1} x)$. The functions $C_n(x)$ are actual polynomials and satisfy the recursion relation $C_{n+1}(x) = 2xC_n(x) - C_{n-1}(x)$. In the following lemma we summarize some known properties of the Chebyshev polynomials

Lemma 1.6. *The Chebyshev polynomials C_n on $[-1,1]$ have the following properties:*

- *For all $x \in [-1,1]$ and $n \geq 1$ we have $|C_n(x)| \leq 1$ and $||C_n||_\infty = 1$.*
- *For all $n \geq 1$, $2^{-(n-1)}C_n$ is a monic polynomial. This polynomial has the smallest ∞-norm among all monic polynomials of degree n on $[-1,1]$.*
- *The best approximation of $x \mapsto x^{n+1}$ in $\mathbb{P}_n$ is given by $x^{n+1} - 2^{-n}T_{n+1}(x)$.*

- *The roots of $C_n(x)$ are*

$$x_j = \cos \frac{(2j-1)\pi}{2n},\ j = 1, \ldots, n.$$

- *The extrema of $C_n(x)$ are at*

$$x_j = \cos(\frac{j\pi}{n}),\ j = 0, \ldots, n$$

- *The polynomials C_n form a system of orthonormal polynomials on $[-1, 1]$ with weight function $w(x) := \frac{1}{\sqrt{1-x^2}}$ (so the inner product is $\int_{-1}^{1} f(x)g(x)w(x)dx$).*

Proof. These results can be found in [6]. □

Now, the roots of the Chebyshev polynomial C_{n+1} can be used as the interpolation nodes $\overline{x}$. The corresponding Lebesgue constant $\Lambda_n(\overline{x})$ has the following properties:

- The asymptotic growth of the Lebesgue constant is given by

$$\Lambda_n(\overline{x}) \simeq \frac{2}{\pi} \log(n+1) + \frac{2}{\pi}(\gamma + \log \frac{8}{\pi}) \text{ for } n \to \infty,$$

  where $\gamma \approx 0.577$ is Euler's constant.

- More precisely, we can bound the Lebesgue constant:

$$\frac{2}{\pi} \log(n+1) + 0.9625... < \Lambda_n(\overline{x}) < \frac{2}{\pi} \log(n+1) + 0.9734...$$

Comparing this bound with (1.5) we can see the Lebesgue constant of the Chebyshev nodes are actually almost optimal.

Another set of nodes arising from the Chebyshev polynomials are its extrema. The advantage here is that the set of $n+1$ extrema $\overline{y}$ of C_n is a subset of the set of extrema of C_{2n}. Hence, we can easily add more nodes to our polynomial approximation using the function evaluations we already have. Notice that these function evaluations can be expensive when we consider solution operators of PDEs. The Lebesgue constant of the nodes $\overline{y}$ satisfy

$$\Lambda_n(\overline{y}) = \begin{cases} \Lambda_{n-1}(\overline{x}) & n \text{ odd}, \\ \Lambda_{n-1}(\overline{x}) - \alpha_n & 0 < \alpha_n < \frac{1}{n^2}, n \text{ even}, \end{cases}$$

where $\overline{x}$ are the roots of C_{n+1}. As the Lebesgue constant of the extrema of C_n is quite comparable to that of the Chebyshev roots nodes, we seem to have a good alternative for the choice of nodes here.

Several other types of Chebyshev nodes can be found in [7] and [8].

Leja sequences

The Leja sequences are defined using the unit disk D in the complex plane and an initial point. Namely, one first picks an initial point e_0 on the unit disk and then defines inductively

$$e_j = \text{argmax}_{z \in D} \left| \prod_{i=0}^{j-1} (z - e_i) \right|.$$

Note that this sequence is not uniquely defined by the initial point e_0. After projection of this sequence onto the interval $[-1, 1]$ we obtain a sequence of points $\{x_i\}_{i=0}^{\infty}$ that can be used as interpolation nodes. The advantage of using this sequence of points is that the corresponding interpolation operators are nested. This will be useful later when defining a certain type of sparse interpolation operator. The Lebesgue constant Λ_{L_n} of the interpolation operator using the nodes $\{x_0, \ldots, x_n\}$ satisfies

$$\Lambda_{L_n} \leq 8\sqrt{2}n^2 \text{ for } n \geq 2.$$

From now on, we will call the sequence $\{x_i\}_{i=0}^{\infty}$ the Leja points.

1.1.2 Best polynomial approximation of analytic functions

In this thesis we have the hope that the solution of our PDEs depends analytically on the parameters. It turns out that analytic functions can be approximated really well by polynomials. In the Runge example we already saw that for a function that is holomorphic on a sufficiently large disk the interpolation error converges to zero $n \to \infty$. In this section we try to refine the argument used there by giving a better bound on the error in the best polynomial approximation of such functions. This bound can then be combined with the Lebesgue constant of an interpolator to give an estimate of the interpolation error.

First we introduce some notation:

Definition 1.7. Let $x \in \mathbb{C}$ and $R > r > 0$, then we define the following domains:

- The ball centered at x with radius r

$$B(x, r) := \{z \in \mathbb{C} \colon |z - x| < r\}.$$

- The circle $C(x, r)$ centered at x with radius r is the boundary of $B(x, r)$.
- The annulus centered at x with radii r and R

$$A(x, r, R) := \{z \in \mathbb{C} \colon |z - x| > r \text{ and } |x - r| < R\}.$$

- The open elliptic disk D_ρ around $[-1, 1]$

$$D_\rho := T(A(0, \rho^{-1}, \rho),$$

where $T(z) := \frac{z + z^{-1}}{2}$ and $\rho > 1$.

Remark 1.8. An elliptic disk can also, more intuitively, be defined as the domain enclosed by the curve $a\cos(t) + b\sin(t)i$, $0 \leq t < 2\pi$, for certain $a, b > 0$. To see that both definitions coincide, notice that we may write $T(\rho e^{it}) = \frac{1}{2}(\rho + \rho^{-1})\cos(t) + \frac{1}{2}(\rho - \rho^{-1})\sin(t)i$. Hence for $\rho > 1$ we have that $T(B(0,\rho) = T(B(0,\rho^{-1})$ is an ellipse around the interval $[-1,1]$, while for $\rho = 1$ we have $[-1,1] = T(B(0,1)$. By continuity, $T(A(0,\rho^{-1},\rho)$ is an ellipse enclosed by the curve $\frac{1}{2}(\rho + \rho^{-1})\cos(t) + \frac{1}{2}(\rho - \rho^{-1})\sin(t)i$.

We are now ready to state the main result of this section. This result combined with the bound of the Lebesgue constants can be used to prove convergence of interpolation methods.

Theorem 1.9. *Let $f\colon [-1,1] \to \mathbb{R}$ be a function that can be extended to a holomorphic function on D_ρ for some $\rho > 1$. Then we have that $E_n(f) \leq \frac{2M}{\rho - 1}\rho^{-n}$, where M is the maximum absolute value of f on D_ρ.*

Proof. From [11]. First notice that $f(\cos(t))$ has a Fourier series that consists of only cosine basisfunctions. From this Fourier series we then obtain the Chebyshev series

$$f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{\infty} a_k C_k(x),$$

with

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(\cos t)\cos k t dt.$$

Now, if we cut off the sum at $k = n$ we obtain a polynomial of degree n, hence

$$E_n(f) \leq ||f - \frac{1}{2}a_0 - \sum_{k=1}^{n} a_k C_k||_\infty = ||\sum_{k=n+1}^{\infty} a_k C_k||_\infty \leq \sum_{k=n+1}^{\infty} |a_k|.$$

Thus we want to estimate the Fourier constants $|a_k|$. This can be done using complex analyis. First we make the substition $z = e^{it}$ to get on the complex domain:

$$a_k = \frac{1}{\pi i}\int_{C(0,1)} f\left(\frac{z + z^{-1}}{2}\right)\frac{z^k + z^{-k}}{2}\frac{dz}{z}$$

Now, define $g(z) = f(\frac{z+z^{-1}}{2})$ and pick $1 < \rho_1 < \rho$. Since f is analytic on D_ρ we have that g is analytic on a neighborhood of $A(0,\rho_1^{-1},\rho_1)$. We can now deduce the following by splitting the integral and using Cauchy's integral formula

$$\begin{aligned} a_k &= \frac{1}{\pi i}\int_{C(0,1)} f\left(\frac{z + z^{-1}}{2}\right)\frac{z^k + z^{-k}}{2}\frac{dz}{z} \\ &= \frac{1}{\pi i}\int_{C(0,1)} g(z)\frac{z^{k-1}}{2}dz + \frac{1}{\pi i}\int_{C(0,1)} g(z)\frac{z^{-(k+1)}}{2}dz \\ &= \frac{1}{\pi i}\int_{C(0,\rho_1^{-1})} g(z)\frac{z^{k-1}}{2}dz + \frac{1}{\pi i}\int_{C(0,\rho_1)} g(z)\frac{z^{-(k+1)}}{2}dz \end{aligned}$$

After estimating both integrals we obtain $|a_k| \leq \frac{1}{2\pi} M[\rho_1^{-(k-1)} 2\pi\rho_1^{-1} + \rho_1^{-(k+1)} 2\pi\rho_1] = 2M\rho_1^{-k}$, where M is the maximum absolute value of f on D_ρ. Because $1 < \rho_1 < \rho$ was arbitrary we find that $|a_k| \leq 2M\rho^{-k}$ We can now finish the proof by calculating

$$\begin{aligned} E_n(f) &\leq \sum_{k=n+1}^{\infty} |a_k| \\ &\leq \sum_{k=n+1}^{\infty} 2M\rho^{-k} \\ &= \frac{2M}{\rho - 1}\rho^{-n}. \end{aligned}$$

□

Remark 1.10. In [11] it is proven that f is holomorphic on D_ρ if and only if $\limsup_{n\to\infty} \sqrt[n]{E_n(f)} \leq \rho^{-1}$. This implies that the bound proven above is asymptotically sharp.

1.2 Higher-Dimensional Interpolation

In this section we discuss the interpolation of functions that are dependent on multiple variables. We first look at full tensor product interpolation and discuss the 'curse of dimension'. The solution to this problem is sparse tensor product interpolation for which we give an analysis of the error.

1.2.1 Full Tensor Product

Before we are able to introduce the interpolation operator for higher dimensions, we need some kind of notation that suits this subject.

Definition 1.11. For $J \in \mathbb{N}$ we call $\mathbb{N}^J$ the set of J-dimensional indices. If for $\nu, \mu \in \mathbb{N}^J$ we have that $\nu_i \leq \mu_i$ for all $i = 1, \ldots, N$ then we say that $\nu \leq \mu$. This defines a partial order on $\mathbb{N}^J$.

We call a set $\Lambda \in \mathbb{N}^J$ lower if for all $\nu \in \Lambda$ and $\mu \leq \nu$ we have that $\mu \in \Lambda$.

Now, let $\{x_0, x_1, \ldots\}$ be an infinite sequence of mutually distinct interpolation nodes in $[-1, 1]$ and let $f\colon \mathbb{R}^J \to \mathbb{R}$ be a real-valued function dependent on J parameters. The full-tensor grid interpolator maps f to a J-variate polynomial in $\bigotimes_{j=1}^{J} \mathbb{P}_{\nu_j}$ that coincides with f on the grid $\Gamma_{\mathcal{B}_\nu} := \{x = (x_{\mu_1}, \ldots, x_{\mu_J}) \in \mathbb{R}^J : \mu \in \mathcal{B}_\nu\}$, where $\mathcal{B}_\nu := \{\mu \leq \nu\}$ for some $\nu \in \mathbb{N}^J$ is a lower set. This operator can conveniently be defined using tensor product. Namely, we can simply define $I_\nu = \otimes_{i=1}^{J} I_{\nu_n}$.

Remark 1.12. The tensor product between operators can be seen as a composition of operators. Namely, take two operators $I_n\colon C([a,b]) \to \mathbb{P}_n, I_m\colon C([a,b]) \to \mathbb{P}_m$ and a function we wish to interpolate $f\colon [a,b]^2 \to \mathbb{R}$, where $[a,b]$ is some interval. Then we have that

$I_n \otimes I_m f = I_n(x \mapsto I_m(y \mapsto f(x,y)))$. For example,

$$\begin{aligned}(I_1 \otimes I_1)f(x,y) &= I_1(x \mapsto I_1(y \mapsto f(x,y))) \\ &= I_1[x \mapsto \frac{y-y_1}{y_0-y_1}f(x,y_0) + \frac{y-y_0}{y_1-y_0}f(x,y_1)] \\ &= \frac{x-x_1}{x_0-x_1}[\frac{y-y_1}{y_0-y_1}f(x,y_0) + \frac{y-y_0}{y_1-y_0}f(x,y_1)]_{x=x_0} \\ &+ \frac{x-x_0}{x_1-x_0}[\frac{y-y_1}{y_0-y_1}f(x,y_0) + \frac{y-y_0}{y_1-y_0}f(x,y_1)]_{x=x_1}.\end{aligned}$$

This resulting function coincides with f in the points (x_i, y_j) for $i, j = 0, 1$.

Similarly, one can show that $I_n \otimes I_m$ coincides with f on the grid $\mathcal{B}_{(n,m)}$.

The error analysis for the full-tensor product is quite straigthforward. Again we have to deal with two things. Firstly, we need to bound the Lebesgue constant of the interpolator operator. Secondly we need to find an estimate for the best polynomial approximation of the function we wish to interpolate. These two things are done below, where we investigate the operator $I_\nu = \otimes_{i=1}^J I_{\nu_n}$ for some $\nu \in \mathbb{N}^J$.

- Let Λ_{ν_i} be the Lebesgue constant of I_{ν_i} for $i = 1, \ldots, J$. Then we can show that the Lebesgue constant Λ_ν of I_ν is equal to $\prod_{j=1}^J \Lambda_{\nu_j}$. First of all, for $i = 1, \ldots, J$, let λ_{ν_i} be the normalized Lebesgue functions of I_{ν_i}, so that $||I_{\nu_i}\lambda_{\nu_i}||_\infty = \Lambda_{\nu_i}$. Then we have the following

$$||I_\nu \left(\prod_{j=1}^J \lambda_{\nu_j} \right) ||_\infty = || \otimes_{j=1}^J I_{\nu_j}\lambda_{\nu_j}||_\infty = \prod_{j=1}^J ||I_{\nu_j}\lambda_{\nu_j}||_\infty = \prod_{j=1}^J \Lambda_{\nu_j},$$

  hence $\Lambda_\nu \geq \prod_{j=1}^J \Lambda_{\nu_j}$. Now, pick an a function $f\colon [a,b]^J \to \mathbb{R}$ with supremum-norm 1. Then we can make the following calculation, where $\vec{x} = (x_1, \ldots x_{J-1})$:

$$||I_\nu f||_\infty = || \otimes_{j=1}^{J-1} I_{\nu_j}(I_{\nu_J} f_{\vec{x}}(x_J))||_\infty \leq || \otimes_{j=1}^{J-1} I_{\nu_j}||_\infty \cdot ||I_{\nu_J} f_{\vec{x}}(x_J)||_\infty \leq || \otimes_{j=1}^{J-1} I_{\nu_j}||_\infty \Lambda_{\nu_J}.$$

  From this calculation it follows that $\Lambda_\nu \leq || \otimes_{j=1}^{J-1} I_{\nu_j}||\Lambda_{\nu_J}$. Then using induction we find that $\Lambda_\nu \leq \prod_{j=1}^J \Lambda_{\nu_j}$. Hence, $\Lambda_\nu = \prod_{j=1}^J \Lambda_{\nu_j}$.

  Now, if the interpolator operator uses the Chebyshev roots or extrema as its nodes, we can find the following estimate for $\Lambda(I_\nu)$:

$$\Lambda(I_\nu) \leq \prod_{j=1}^J \frac{2}{\pi}\log(\nu_j + 1) + 1.$$

- Before we can find an estimate for the best polynomial approximation of a function we need to identify range of the operator I_ν. Clearly, this is the polynomial space defined as follows:

$$\mathbb{P}_\nu := \bigotimes_{j=1}^J \mathbb{P}_{\nu_j} = \{p_1 \cdot p_2 \cdot \ldots \cdot p_J \colon p_j \in \mathbb{P}_{\nu_j} \text{ for all } j = 1, \ldots, J\}.$$

Now, we apply the one-dimensional result to the higher-dimensional case. Namely, assume that f defined on $[-1,1]^J$ can be extented to a holomorphic function on D_ρ^J for some $\rho > 1$. For simplicity we assume that $J = 2$. Because $f(\cos x, \cos y)$ is an even function it has a Fourier series from which we can obtain a two-dimensional Chebyshev series:

$$f(x,y) = \frac{1}{4}a_{00} + \frac{1}{2}\sum_{k=1}^{\nu_1} a_{0k}C_k(x) + \frac{1}{2}\sum_{k=1}^{\nu_2} a_{k0}C_k(y) + \sum_{n=1}^{\nu_1}\sum_{m=1}^{\nu_2} a_{nm}C_n(x)C_m(y),$$

where

$$a_{nm} = \frac{1}{(\pi i)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(\cos t_1, \cos t_2)\cos(nt_1)\cos(mt_2)dt_1dt_2.$$

Now, just like in one dimension, we want to bound the coefficients a_{nm}. This is again done using complex analysis. First of all we have the following identity:

$$a_{nm} = \frac{1}{\pi i}\int_{C(0,1)} \frac{1}{\pi i}\int_{C(0,1)} f\left(\frac{z_1+z_1^{-1}}{2}, \frac{z_2+z_2^{-1}}{2}\right)\frac{z^n+z^{-n}}{2}\frac{dz_1}{z_1}\frac{z^m+z^{-m}}{2}\frac{dz_2}{z_2}.$$

Now, using the strategy as in Theorem 1.9 we find that $\int_{C(0,1)} f\left(\frac{z_1+z_1^{-1}}{2}, \frac{z_2+z_2^{-1}}{2}\right)\frac{z^n+z^{-n}}{2}\frac{dz_1}{z_1}$ is a function of z_2 that can be bounded by $2M\rho^{-n}$. Additionally, this function is holomorphic on the annulus as in Theorem 1.9. Hence, we can again apply the strategy from the same theorem to find that $|a_{nm}| \leq 2(2M\rho^{-n})\rho^{-m} = 4M\rho^{-n-m}$. After summation of these coefficients we find that $E_{\nu_1,\nu_2}(f) \leq 4M\sum_{\substack{\mu\not\leq\nu\\ \mu\in\mathbb{N}_0^2}} \rho^{-\mu_1-\mu_2} = 4M\left(\frac{\rho^{-\nu_1}+\rho^{-\nu_2}-\rho^{-\nu_1-\nu_2}}{(\rho-1)^2} + \frac{\rho^{-\nu_1}+\rho^{-\nu_2}}{\rho-1}\right)$. Notice that we allow the value zero in the multi-indices here. Of course this result can be generalized for higher dimensions. We obtain the following estimate:

$$E_\nu(f) \leq 2^J M \sum_{\substack{\mu\not\leq\nu\\ \nu\in\mathbb{N}_0^J}} \rho^{-\sum_{j=1}^J \mu_j}.$$

From this estimate we can see that for the error to converge to zero, we need all ν_j to grow to infinity. Namely, pick ν with $\nu_1 = k$, then $(k+1,0,0,\ldots,0) \not\leq \nu$. Hence, the above upper bound for the best polynomial approximation, in this case, is bigger than $2^J M\rho^{-(k+1)}$. Hence, for the upper bound to converge to zero, ν_1 needs to grow to infinity.

The estimate of the Lebesgue constant and the best polynomial approximation can be put together to form the following theorem:

Theorem 1.13. *Let $f\colon [-1,1]^J \to \mathbb{R}$ be a function that can be extended holomorphically to the multi-dimensional ellipse D_ρ^J for some $\rho > 1$. Furthermore, let $\nu \in \mathbb{N}^J$ be a multi-index and I_ν be the corresponding interpolation operator, which uses Chebyshev points or Chebyshev roots as its nodes. Then we have the following error estimate:*

$$||f - I_\nu f||_\infty \leq 2^J M\left(1 + \prod_{j=1}^J\left[\frac{2}{\pi}\log(\nu_j+1)+1\right]\right)\sum_{\substack{\mu\not\leq\nu\\ \nu\in\mathbb{N}_0^J}} \rho^{-\sum_{j=1}^J \mu_j},$$

where M is the maximum of f on D_ρ^J. Consequently, for $\rho > 1$, if $\nu_j \to \infty$ for every $j = 1, \dots J$, the interpolation error converges to zero.

This theorem shows that interpolation using the full tensor product works. However, the number of interpolation points increases rapidly when J becomes larger. For example, if we use n points in each direction, we need to perform n^J function evaluations. For more expensive functions, like solution operators of PDEs, this is not desirable. This problem is called 'the curse of dimension'. We would like to find a way to still succesfully interpolate functions while using a moderate number of interpolation nodes. This is done in the next section.

Remark 1.14. If we choose $J = 2$ and $\nu_1 = \nu_2 = n$ then the number of function evaluations $|\Gamma_{\mathcal{B}_\nu}|$ is equal to n^2. The error estimate in terms of $|\Gamma_{\mathcal{B}_\nu}|$ in this case is

$$||f - I_\nu f||_\infty \leq 4M \left(1 + (\frac{2}{\pi}\log(|\Gamma_{\mathcal{B}_\nu}|^{1/2} + 1) + 1)^2\right)\left(\frac{2\rho^{-\sqrt{|\Gamma_{\mathcal{B}_\nu}|}} - \rho^{-2\sqrt{|\Gamma_{\mathcal{B}_\nu}|}}}{(\rho-1)^2} + \frac{2\rho^{-\sqrt{|\Gamma_{\mathcal{B}_\nu}|}}}{\rho - 1}\right).$$

In the next section, we provide an algorithm which can perform better in terms of the function evaluations $|\Gamma|$.

1.2.2 Sparse grid interpolation

In this section we introduce the Smolyak algorithm. This interpolation strategy was first introduced by Smolyak in 1963, see [16]. The idea behind the Smolyak algorithm is to use a sparse set of interpolation points to make the computational burden smaller. It turns out that the decay of the error in terms of the number of interpolation points can be faster than the decay of the error for the full tensor interpolator. The notation in this section is adopted from [17] and [18].

We assume that we have a strictly increasing function $m\colon \mathbb{N}_0 \to \mathbb{N}_0$ with $m_0 = 0$ and interpolation operators $\mathcal{U}^i := I_{m_i - 1}$, $i \in \mathbb{N}_0$ on m_i interpolation points, and $\mathcal{U}^0 = 0$. Furthermore, we assume that the interpolation points of these operators are nested. Hence, we have a sequence of interpolation nodes $\{x_i\}_{i=0}^\infty$ such that the operators $\mathcal{U}^i := I_{m_i - 1}$ are defined using these nodes and Definition 1.1. We also define difference operators by setting

$$\Delta_k := \mathcal{U}^k - \mathcal{U}^{k-1}$$

for $k \geq 1$ and

$$\Delta_\nu := \otimes_{j=1}^J \Delta_{\nu_j}$$

for $\nu \in \mathbb{N}^J$.

Using the notation above we define the Smolyak interpolation operator:

Definition 1.15. Let $m\colon \mathbb{N}_0 \to \mathbb{N}_0$ with $m_0 = 0$ be strictly increasing and let $q, J \in \mathbb{N}$ be integers such that $q \geq J$. Then we define $\mathcal{A}(q, J)$ by setting

$$\mathcal{A}(q, J)f := \sum_{|\nu| \leq q} \otimes_{j=1}^J \Delta_{\nu_j} f, \text{ for all continuous } f\colon \mathbb{R}^J \to \mathbb{R}.$$

Remark 1.16. Alternatively, using telescopic cancellation, we can define the Smolyak algorithm as

$$\mathcal{A}(q,J)f = \sum_{\substack{|\nu|\leq q-1 \\ \nu\in\mathbb{N}^{J-1}}} \left(\bigotimes_{i=1}^{J-1} \Delta_{\nu_i} \otimes \mathcal{U}^{q-|\nu|} \right).$$

Remark 1.17. The difference operators give us an alternative way of defining the full tensor interpolator. Namely, using telescopic cancellation and taking $m_i = i$, so that $\mathcal{U}^i$ is an interpolation operator on i points, we can find a different expression for the full tensor interpolation operator:

$$\sum_{\mu\leq\nu} \Delta_\nu f = \sum_{\mu\leq\nu} \otimes_{j=1}^{J} \Delta_{\nu_j} f = \otimes_{j=1}^{J} \sum_{\mu_j\leq\nu_j} \Delta_{\mu_j} f = \otimes_{j=1}^{J} I_{\mu_{j-1}} f.$$

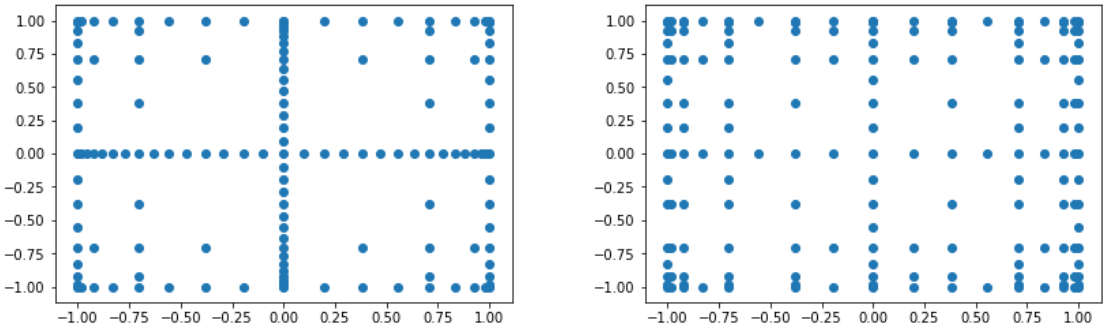
Remark 1.18. We can define the grid that is used in the sparse grid interpolation. The grid consists of all the points in $[-1,1]^J$ that f needs to be evaluated in. This grid is equal to

$$\Gamma = \bigsqcup_{|\nu|\leq q} \Gamma_\nu,$$

where Γ_ν is defined as

$$\Gamma_\nu = \{(x_{\mu_1},\ldots,x_{\mu_J}) | (m_{\nu_1-1}-1,\ldots,m_{\nu_J-1}-1) < \mu \leq (m_{\nu_1}-1,\ldots,m_{\nu_J}-1)\}.$$

Figure 1.2 shows two examples of a sparse grid. Both sparse grids will be analysed in more detail.



(a) Clenshaw-Curtis points for $q = 7$.

(b) Sparse Leja points for $q = 19$.

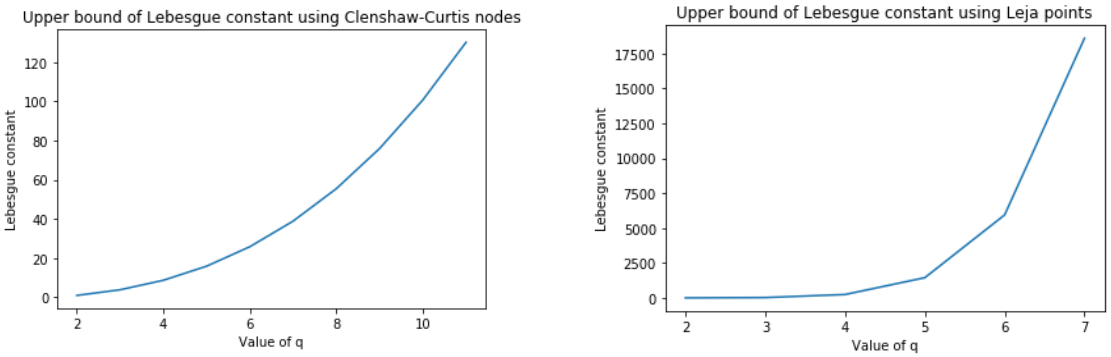
Figure 1.2: Two examples of an sparse grid.

Remark 1.19. According to Remark 1.2 we also need to estimate the Lebesgue constant of the sparse grid operator. For $J = 2$, the Lebesgue constant can be estimated using the

triangle inequality:

$$\begin{aligned}\Lambda &\leq \sum_{k=1}^{q-1} ||\Delta_k \otimes \mathcal{U}^{q-k}|| \\ &\leq \sum_{k=1}^{q-1} ||I_{m_k-1} - I_{m_{k-1}-1}|| \cdot ||I_{m_{q-k}-1}|| \\ &\leq \sum_{k=1}^{q-1} (\Lambda_{m_k-1} + \Lambda_{m_{k-1}-1})\Lambda_{m_{q-k}-1}.\end{aligned}$$

This estimate is displayed in Figure 1.3 for different Smolyak algorithms.



(a) Upper bound using Clenshaw-Curtis.

(b) Upper bound using sparse Leja points.

Figure 1.3: Upper bound for the Lebesgue constant of different Smolyak algorithms.

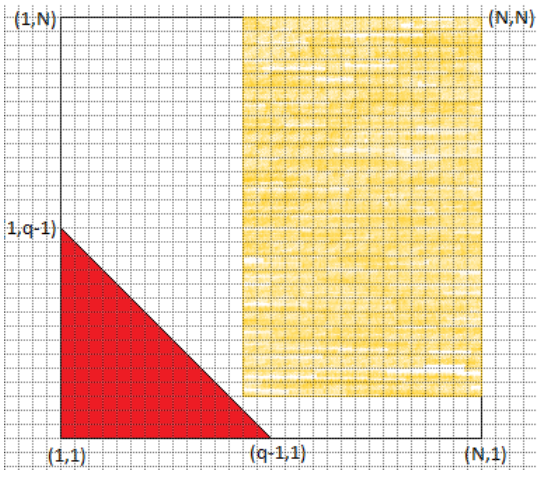


Figure 1.4: The intersections represent difference operators $\Delta_n \otimes \Delta_k$; e.g. $\Delta_1 \otimes \Delta_1$ is on the bottom left. The red triangle including its boundary consists of the difference operators used in the Smolyak algorithm.

Now, Definition 1.15 needs some explanation as it is not immediately clear why this interpolation strategy is sound. Therefore, we now derive an error estimate for the case $J = 2$.

Firstly, notice that the error of the Smolyak algorithm can be expressed in the following way:

$$\begin{aligned}||(\mathrm{Id}-\mathcal{A}(q,2))f||_\infty &= ||(\mathrm{Id}-\sum_{|\nu|\leq q}\Delta_{\nu_1}\otimes\Delta_{\nu_2})f||_\infty \\ &= ||\sum_{|\nu|>q}\Delta_{\nu_1}\otimes\Delta_{\nu_2}f|| \\ &= \lim_{N\to\infty}||\sum_{\substack{|\nu|>q \\ \nu\leq(N,N)}}\Delta_{\nu_1}\otimes\Delta_{\nu_2}f||\end{aligned}$$

Figure 1.4 illustrates the last expression. All the difference operator outside the red triangle contribute to the error of the Smolyak algorithm.

Now, consider the contribution to the total error of the difference operators inside and on the boundary of the yellow rectangle in Figure 1.4. This contribution is given by

$$||\sum_{\substack{\nu_1\geq q-c+1,\nu_2\geq c \\ \nu\leq(N,N)}}\Delta_{\nu_1}\otimes\Delta_{\nu_2}f||_\infty = ||(\mathcal{U}^N-\mathcal{U}^{q-c})\otimes(\mathcal{U}^N-\mathcal{U}^{c-1})f||_\infty$$

for some $c\geq 1$. Now, by adding the contribution to the error of all such rectangles we find that

$$\begin{aligned}||\sum_{|\nu|>q;\nu\leq(N,N)}\Delta_{\nu_1}\otimes\Delta_{\nu_2}f|| &\leq \sum_{c=1}^{q}||\sum_{\substack{\nu_1\geq q-c+1,\nu_2\geq c \\ \nu\leq(N,N)}}\Delta_{\nu_1}\otimes\Delta_{\nu_2}f||_\infty \\ &= \sum_{c=1}^{q}||(\mathcal{U}^N-\mathcal{U}^{q-c})\otimes(\mathcal{U}^N-\mathcal{U}^{c-1})f||_\infty.\end{aligned}$$

To continue we estimate

$$\begin{aligned}&||(\mathcal{U}^N-\mathcal{U}^{q-c})\otimes(\mathcal{U}^N-\mathcal{U}^{c-1})f||_\infty \\ &= |||[(\mathcal{U}^N-\mathrm{Id})+(\mathrm{Id}-\mathcal{U}^{q-c})]\otimes[(\mathcal{U}^N-\mathrm{Id})+(\mathrm{Id}-\mathcal{U}^{c-1})]f||_\infty \\ &\leq ||(\mathcal{U}^N-\mathrm{Id})\otimes(\mathcal{U}^N-\mathrm{Id})f||_\infty + ||(\mathcal{U}^N-\mathrm{Id})\otimes(\mathrm{Id}-\mathcal{U}^{c-1})f||_\infty \\ &+ ||(\mathrm{Id}-\mathcal{U}^{q-c})\otimes(\mathcal{U}^N-\mathrm{Id})f||_\infty + ||(\mathrm{Id}-\mathcal{U}^{q-c})\otimes(\mathrm{Id}-\mathcal{U}^{c-1})f||_\infty \\ &\leq 4M\Lambda_{m_N,m_N}\frac{\rho^{-m_N-m_N+2}}{(\rho-1)^2}+4M\Lambda_{m_N,m_{c-1}}\frac{\rho^{-m_N-m_{c-1}+2}}{(\rho-1)^2} \\ &+ 4M\Lambda_{m_{q-c},m_N}\frac{\rho^{-m_{q-c}-m_N+2}}{(\rho-1)^2}+4M\Lambda_{m_{q-c},m_{c-1}}\frac{\rho^{-m_{q-c}-m_{c-1}+2}}{(\rho-1)^2}.\end{aligned}$$

Now, if $\Lambda_{m_N}\rho^{-m_N}\to 0$ for $N\to\infty$, we obtain

$$\lim_{N\to\infty}||(\mathcal{U}^N-\mathcal{U}^{q-c})\otimes(\mathcal{U}^N-\mathcal{U}^{c-1})f||_\infty \leq 4M\Lambda_{m_{q-c},m_{c-1}}\frac{\rho^{-m_{q-c}-m_{c-1}+2}}{(\rho-1)^2}.$$

Hence, if we put everything together we find that the error of the Smolyak algorithm can be bounded in the following way:

$$||(\mathrm{Id}-\mathcal{A}(q,2))f||_\infty \leq \frac{4M}{(\rho-1)^2}\sum_{c=1}^{q}\Lambda_{m_{q-c},m_{c-1}}\rho^{-m_{q-c}-m_{c-1}+2}.$$

Depending on the function m and the choice of interpolation nodes this estimate can be simplified.

Clenshaw-Curtis operator

We can use the Chebyshev extrema as the interpolation nodes. For this, we choose $m_1 = 1$ with interpolation point $x_0 = 0$ and for $i > 1$ we choose $m_i = 2^{i-1} + 1$ with nodes $\{\cos(\frac{j\pi}{2^{i-1}}\}_{i=0}^{2^{i-1}}$. The corresponding Lebesgue constants satisfy $\Lambda_{m_i} \leq 2(i+1)$ and $\Lambda_{m_i-1} \leq 2i$. The interpolation nodes in the resulting sparse grid are called the Clenshaw-Curtis nodes.

The interpolation error can be estimated by

$$\begin{aligned} ||(\mathrm{Id} - \mathcal{A}(q,2))f||_\infty &\leq \frac{4M}{(\rho-1)^2} \sum_{c=1}^{q} \Lambda_{m_{q-c},m_{c-1}} \rho^{-m_{q-c}-m_{c-1}+2} \\ &= \frac{4M}{(\rho-1)^2} \sum_{c=1}^{q} 4(q-c+1)c\rho^{-m_{q-c}-m_{c-1}+2} \\ &\leq \frac{4M(q+1)^2}{(\rho-1)^2} q\rho^{-2^{\frac{q-1}{2}-1}} \end{aligned}$$

Here we used that $(q-c+1)c \leq (q+1)^2/4$ and $\rho^{-m_{q-c}-m_{c-1}+2} \leq \rho^{-2^{\frac{q-1}{2}-1}}$ for $c = 1, \ldots, q$.

Using the next lemma one can find the error of the Clenshaw-Curtis method with respect to the number of interpolation points.

Lemma 1.20. *The number of interpolation points $|\Gamma|$ used in the Clenshaw-Curtis method satisfies*

$$J(2^{q-J+1} - 1) \leq |\Gamma| \leq \left(\frac{e}{J-1}\right)^{J-1} \cdot (q-1)^{J-1} \cdot 2^{q-J+1},$$

where the Clenshaw-Curtis operator is given by $\mathcal{U}_q f := \sum_{|k|\leq q} \bigotimes_{i=1}^{J} \Delta_{k_i} f$. For $J = 2$ we have the sharper bound

$$|\Gamma| \leq (q-1)2^{q-1}$$

so that

$$q \geq \frac{\log(|\Gamma|)}{\log(2)+1} + 1.$$

Proof. Adapted from [19]. By counting the nodes level-wise we find that the number of points S is equal to

$$\sum_{|i|\leq q} \prod_{n=1}^{J} r(i_n),$$

where $r(1) := 1$, $r(2) = 2$ and $r(i) = 2^{i-2}$ for $i \geq 3$. Now, the function r conveniently satisfies $2^{i-2} \leq r(i) \leq 2^{i-1}$. Hence, we can estimate

$$\begin{aligned}
|\Gamma| &= \sum_{|i|\leq q} \prod_{n=1}^{J} r(i_n) \\
&\leq \sum_{|i|\leq q} 2^{|i-1|} \\
&= \sum_{n=J}^{q} \binom{n-1}{J-1} 2^{n-J} \\
&\leq \sum_{n=J}^{q} \left(\frac{(n-1)e}{J-1}\right)^{J-1} 2^{n-J} \\
&\leq \sum_{n=J}^{q} \left(\frac{(q-1)e}{J-1}\right)^{J-1} 2^{n-J} \\
&\leq \left(\frac{e}{J-1}\right)^{J-1} \cdot (q-1)^{J-1} \cdot 2^{q-J+1}.
\end{aligned}$$

Now, for the lower bound, let $\hat{q}$ range from 1 to $q - J + 1$ and consider the nodes $i^{\hat{q},n}$ defined by $i_k^{\hat{q},n} = 1$ for $k \neq n$ and $i_n^{\hat{q},n} = \hat{q}$. For each value of $\hat{q}$ there are J nodes $i^{\hat{q},n}$. Hence,

$$|\Gamma| \geq \sum_{\hat{q}=1}^{q-J+1} J 2^{\hat{q}-1} = J(2^{q-J+1} - 1).$$

□

In [19] one can find similar kind of estimates for the Clenshaw-Curtis operator for arbitrary $J \in \mathbb{N}$. In this article an induction argument is used.

Using Leja points

We can also use the Leja points as the interpolation points. This allows us to choose $m_i = i$ for $i \geq 0$ as the corresponding interpolation operators $\mathcal{U}^i$ are then defined using nested interpolation nodes.

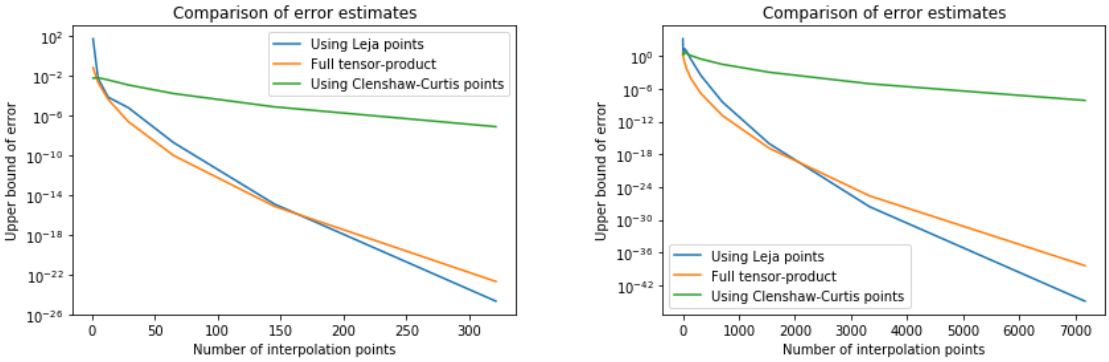
The interpolation error can be estimated by

$$\begin{aligned}
||(\mathrm{Id} - \mathcal{A}(q,2))f||_\infty &\leq \frac{4M}{(\rho-1)^2} \sum_{c=1}^{q} \Lambda_{m_{q-c},m_{c-1}} \rho^{-m_{q-c}-m_{c-1}+2} \\
&= \frac{4M}{(\rho-1)^2} \sum_{c=1}^{q} 128(q-c-1)^2(c-2)^2\rho^{-q+3} \\
&\leq \frac{512M}{(\rho-1)^2} (\frac{q-1}{2} - 2)^4 \rho^{-q+3}
\end{aligned}$$

The number of interpolation point $|\Gamma|$ is equal to $q(q-1)/2$.

A comparison of the error bounds for both Smolyak algorithms and the full tensor interpolator is shown in Figure 1.5. It might be possible to improve the upper bound of the error for

the Smolyak operator that uses Leja points by further investigation of the Lebesgue constants Λ_{L_n}.



(a) $M = 1$, $\rho = 20$.

(b) $M = 1$, $\rho = 3$.

Figure 1.5: Comparison of the upper bounds of the error for two values of ρ.

1.3 Interpolation in Banach spaces

In this section we discuss how we use the theory of interpolation of the real valued functions for the more general case of Banach valued functions. This topic is of great interest for us as it allows us to interpolate parametric PDEs. In this section we also show holomorphic dependence of the solution on the parameters for a few classes of PDEs. Holomorphy leads to similar error estimates as in the real case.

Definition 1.21. Let X be a Banach space and $U \subset \mathbb{C}$ be an open set. A function $f\colon U \to X$ is holomorphic on U, if and only if, for every $\varphi \in X'$ the function $\varphi \circ f$ is holomorphic on U.

Using the definition above we can immediately state a general theorem about the interpolation error in Banach spaces. Here we define $||f||_\infty = \sup\{||f(x)||_X | x \in U\}$ for $f\colon U \to X$.

Theorem 1.22. *Let X be a Banach space and $f\colon [-1,1]^J \to X$ be a function that is holomorphic on $U \subset \mathbb{C}^J$ and let $\mathcal{I}$ be an interpolation operator. Let $C > 0$ be a constant such that for all holomorphic $g\colon U \to \mathbb{R}$ we have that $||\mathcal{I}g - g||_\infty \leq C||g||_\infty$. Then we have the error estimate*

$$||\mathcal{I}f - f||_\infty \leq C||f||_\infty.$$

Proof. Because $\mathcal{I}$ is an interpolator operator, we have $\varphi(\mathcal{I}f) = \mathcal{I}(\varphi \circ f)$ for all $\varphi \in X'$. Because $\varphi \circ f$ is holomorphic on U we can estimate

$$\begin{aligned} ||\varphi(\mathcal{I}f - f)||_\infty &= ||\mathcal{I}(\varphi \circ f) - \varphi \circ f||_\infty \\ &\leq C||\varphi \circ f||_\infty \\ &\leq C||\varphi||_{X'}||f||_\infty. \end{aligned}$$

Now, pick any $x \in [-1,1]^J$ then by the above we have that

$$|\varphi((\mathcal{I}f - f)(x))| \leq C||\varphi||_{X'}||f||_\infty$$

for all $\varphi \in X'$. Because $(\mathcal{I}f - f)(x) \in X$ we have by Hahn-Banach that there is a $\varphi_x \in X'$ such that $||\varphi_x||_{X'} = 1$ and $\varphi_x((\mathcal{I}f - f)(x)) = ||(\mathcal{I}f - f)(x)||_X$. Hence,

$$||(\mathcal{I}f - f)(x)||_X = \varphi_x((\mathcal{I}f - f)(x)) \leq C||f||_\infty.$$

□

The following lemma can be used to prove holomorphic dependence on the parameters of PDEs.

Lemma 1.23. *Let $U \subset \mathbb{C}$ be an open set and $f\colon U \to X$ for X Banach. If the limit*

$$f'(z) := \lim_{w \to z} \frac{f(w) - f(z)}{w - z}$$

exists for all $z \in U$, then f is holomorphic on U.

Proof. We have

$$(\varphi \circ f)'(z) = \lim_{w \to z} \frac{\varphi \circ f(w) - \varphi \circ f(z)}{w - z} = \varphi\left(\lim_{w \to z} \frac{f(w) - f(z)}{w - z}\right)$$

□

Example 1.24. Consider the following parametric PDE:

$$\begin{cases} -u'' + g(\mu)u &= f \text{ on } [-1, 1] \\ u(-1) = u(1) &= 0, \end{cases}$$

where $g \geq 0$ is holomorphic on D_ρ for some $\rho > 1$. Solutions u_μ to this PDE satisfy the following relation:

$$\int_{-1}^{1} u'_\mu(t)v'(t) + g(\mu)u_\mu(t)v(t)dt = \int_{-1}^{1} f(t)v(t)dt.$$

Following the strategy in [18] we can prove that the solution operator u_μ is holomorphic. Firstly, we know that $||u_\mu||_{H^1} \leq \frac{1}{C}||f||_{H^{-1}}$, where $C > 0$ is a constant such that $|u|^2_{H^1} \geq C||u||^2_{H^1}$ (Poincaré inequality). To find the derivative with respect to μ of u_μ we can differentiate the variational formulation to obtain

$$\int_{-1}^{1} (\partial_\mu u_\mu)'(t)v'(t) + g(\mu)(\partial_\mu u_\mu)(t)v(t)dt = -\int_{-1}^{1} g'(\mu)u_\mu(t)v(t)dt \quad \text{for all } v \in H^1_0([-1, 1]). \tag{1.3.1}$$

The solution $\partial_\mu u_\mu$ to this variational formulation is the candidate derivative for u_μ. We just have to prove that this is actually the case. For this we first need to prove continuity of $\mu \mapsto u^\mu$. So, pick $z \in D_\rho$ then we can obtain the identity:

$$\int_{-1}^{1} g(\mu)(u_z(t) - u_\mu(t))v(t) + (u'_z(t) - u'_\mu(t))v'(t)dt = -\int_{-1}^{1} (g(z) - g(\mu))u_z(t)v(t)dt. \tag{1.3.2}$$

Now, if we substitute $v = u_z - u_\mu$ we see that the left-hand side can be bounded from below by $C_1||u_z - u_\mu||^2_{H^1}$ for some $C_1 > 0$ and that the right-hand side is smaller than

$C_2|g(z)-g(\mu)|\cdot||u_z-u_\mu||_{H^1}$ for some $C_2>0$. This proves continuity. Now, if we divide (1.3.2) by $z-\mu$ and subtract (1.3.1) we find

$$\int_{-1}^{1}\left(-\frac{g(z)-g(\mu)}{z-\mu}u_z(t)+g'(\mu)u_\mu(t)\right)v(t)dt=\int_{-1}^{1}g(\mu)\left(\frac{u_z(t)-u_\mu(t)}{z-\mu}-(\partial_\mu u_\mu)(t)\right)v(t)+\left(\frac{u_z'(t)-u_\mu'(t)}{z-\mu}-(\partial_\mu u_\mu')(t)\right)v'(t)dt.$$

After substituting $v=\frac{u_z-u_\mu}{z-\mu}-(\partial_\mu u_\mu)$ we obtain that the absolute value of the left-hand side is smaller than

$$\left[(||g'(\mu)u_\mu-\frac{g(z)-g(\mu)}{z-\mu}u_\mu||_{L^2}+\ ||\frac{g(z)-g(\mu)}{z-\mu}(u_z-u_\mu)||_{L^2})\cdot\frac{||f||}{C}\right]\cdot||\frac{u_z-u_\mu}{z-\mu}-(\partial_\mu u_\mu)||_{L^2}$$
$$=C_{z-\mu}||\frac{u_z-u_\mu}{z-\mu}-(\partial_\mu u_\mu)||_{H^1},$$

where $C_{z-\mu}\to 0$ for $z\to\mu$ because $u_z\to u_\mu$ in L^2, and that the right-hand side is larger than

$$C||\frac{u_z-u_\mu}{z-\mu}-(\partial_\mu u_\mu')||_{H^1}^2,$$

by the Poincaré inequality. Hence, we find that

$$C||\frac{u_z-u_\mu}{z-\mu}-(\partial_\mu u_\mu)||_{H^1}^2\leq C_{z-\mu}||\frac{u_z-u_\mu}{z-\mu}-(\partial_\mu u_\mu)||_{H^1}.$$

Then, after letting $z\to\mu$, we obtain

$$\lim_{z\to\mu}\frac{u_z-u_\mu}{z-\mu}=\partial_\mu u_\mu \text{ in } H_0^1([-1,1]).$$

Hence, the solution operator of this PDE is holomorphic on D_ρ. This means that we can apply the usual interpolation techniques to interpolate the solution map of this PDE.

2 Reduced Basis Methods

In this chapter we introduce the reduced basis method. We apply this method to the problem described in the introduction. Namely, let X be a Banach space, $a\colon X \times X \times \mathcal{P} \to \mathbb{R}$ a parametric bilinear form and $f\colon X \times \mathcal{P} \to \mathbb{R}$ a parametric linear functional and $\mathcal{P} \subset \mathbb{R}^p$ a compact parameter space. We are interested in approximating $u(\mu)$ for all $\mu \in \mathcal{P}$ where $u(\mu)$ solves

$$a(u(\mu), v; \mu) = f(v; \mu),\ v \in X. \tag{2.0.1}$$

In the reduced basis method we do this by using a set of approximations $u_H(\mu_i)$ of $u(\mu_i)$, $i = 1, \dots, N$ where the $\mu_i \in \mathcal{P}$ can be chosen. We use this set to form an approximation space $X_N := \operatorname{span}\{u_H(\mu_1), \dots, u_H(\mu_N)\}$ for the solution manifold $\mathcal{M}$. Here $\mathcal{M}$ is defined as

$$\mathcal{M} := \{u(\mu)\colon \mu \in \mathcal{P}\}.$$

This approximation space X_N is then used in a Galerkin method.

However, because the 'snapshots' $u_H(\mu_i)$ are not exact evaluations of u we are not approximating $\mathcal{M}$ directly. In this chapter we assume that we approximate $u(\mu_i)$ with $u_H(\mu_i)$ using some finite elements method with finite element space V_H. Hence, we are actually approximating the manifold

$$\mathcal{M}_H := \{u_H(\mu)\colon \mu \in \mathcal{P}\} \subset V_H.$$

In other words, the reduced basis method actually approximates the (finite-dimensional) operator u_H. This also means that a good error estimate will most likely be found using the triangle inequality $||\hat{u}(\mu) - u(\mu)|| \leq ||u_H(\mu) - u(\mu)|| + ||u_H(\mu) - \hat{u}(\mu)||$, where $\hat{u}(\mu)$ is the approximation found by the reduced basis method.

Following [13], we make some additional assumptions on the equation (2.0.1).

- We assume that X is a Hilbert space.
- We assume that the parametric bilinear form $a(\cdot, \cdot; \mu)$ is uniformly continuous and coercive. Namely, there is a constant $\overline{\gamma} < \infty$ such that

  $$\overline{\gamma} \geq \gamma(\mu) := \sup_{u,v \in X; ||u||=||v||=1} |a(u, v; \mu)|,$$

  for all $\mu \in \mathcal{P}$ and there is a $\overline{\alpha} > 0$ such that

  $$\overline{\alpha} \leq \alpha(\mu) := \inf_{u \in X; ||u||=1} a(u, u; \mu),$$

  for all $\mu \in \mathcal{P}$.
- We assume that the parametric linear functional $f(\cdot; \mu)$ is uniformly continuous, i.e. there is a $\overline{\gamma}_f$ such that

  $$||f(\cdot; \mu)||_{X'} \leq \overline{\gamma}_f$$

- We assume parameter-separability for both a and f. This means that for $q = 1, \dots, Q_a$, $Q_a \in \mathbb{N}$, there exist coefficient functions $\theta_q^a \colon \mathcal{P} \to \mathbb{R}$ and continuous bilinear forms $a_q(\cdot,\cdot) \colon X \times X \to \mathbb{R}$ such that

$$a(u, v; \mu) = \sum_{q=1}^{Q_a} \theta_q^a(\mu) a_q(u, v), \ \mu \in \mathcal{P}, u, v \in X. \tag{2.0.2}$$

  Similarly,

$$f(u; \mu) = \sum_{q=1}^{Q_f} \theta_q^f(\mu) f_q(u), \ \mu \in \mathcal{P}, u \in X, \tag{2.0.3}$$

  with coeffient functions $\theta_q^f \colon \mathcal{P} \to \mathbb{R}$ for $q = 1, \dots, Q_f$, $Q_f \in \mathbb{N}$. We further assume that the coefficient functions θ_q^a, θ_q^f can be evaluated rapidly. We refer to [14] for that case that either the bilinear form a or the linear form f is not parameter separable.

From the first two assumptions and the Lax-Milgram theorem we immediately infer that for every $\mu \in \mathcal{P}$ there is a unique solution $u(\mu)$ satisfying

$$||u(\mu)|| \leq \frac{||f(\mu)||_{X'}}{\alpha(\mu)} \leq \frac{\overline{\gamma}_f}{\overline{\alpha}}.$$

Hence, the solution space $\mathcal{M}$ is bounded by $\frac{\overline{\gamma}_f}{\overline{\alpha}}$.

2.1 Outline of Method

The reduced basis method consists of two main steps. In the first step we choose parameters $\mu_i \in \mathcal{P}$, $i = 1, \dots, N$ to obtain an parameter set $S_N = \{\mu_1, \dots, \mu_N\}$ in some suitable way. We then calculate so-called 'snapshots' $u_H(\mu_i)$ for each μ_i using a finite element method with finite element space V_H. Here we assume that $X_N = \text{span}\{u_H(\mu_1), \dots, u_H(\mu_N)\}$ is a N-dimensional space with basis $\Phi_N := \{\phi_1, \dots, \phi_N\}$. In the second step we use the basis Φ_N to approximate the solution manifold $\mathcal{M}_H$ using a Galerkin projection. Céa's lemma shows that the Galerkin projection finds good approximations if for each $u_H(\mu) \in \mathcal{M}_H$ there is a $u_N \in X_N$ with $||u_H - u_N||$ very small. The latter can be achieved if the solution operator is holomorphic when we use the so-called greedy method.

In the next section we discuss the second step and its computational aspects in detail. We show how this method can be decomposed into an offline and online part. The offline part consists of all the heavy computations, whereas the online part consists of very cheap computations that can be done on almost any device. In Section 2.2 we show how specific methods for choosing the $\mu_i' s$ result in nice convergence properties. Lastly, we give an error analysis of the reduced-basis method. We say RB instead of reduced basis from now on.

2.1.1 Galerkin Projection

The second step consist of the RB-formulation and is given by the following Galerkin method:

For $\mu \in \mathcal{P}$ find $u_N(\mu) \in \Phi_N$ such that

$$a(u_N(\mu), v, ; \mu) = f(v; \mu), \text{ for all } v \in X_N.$$

Because the space X_N is a closed linear subspace of X we find that $u_N(\mu)$ exists and satisfies $||u_N(\mu)||_X \leq \frac{\overline{\gamma}_f}{\overline{\alpha}}$. Using the basis Φ_N we can find a discrete formulation of the problem. Namely, we first define the matrix $A_N(\mu)$ and vector $f_N(\mu)$

$$\mathbf{A}_N(\mu) := (a(\phi_j, \phi_i; \mu))_{i,j=1}^N \in \mathbb{R}^{N\times N}$$
$$\mathbf{f}_N(\mu) := (f(\phi_i; \mu)_{i=1}^N \in \mathbb{R}^N.$$

Then the solution $\mathbf{u}_N(\mu) \in \mathbb{R}^N$ of the equation

$$\mathbf{A}_N(\mu)\mathbf{u}_N(\mu) = \mathbf{f}_N(\mu), \tag{2.1.1}$$

satisfies $u_N(\mu) = \sum_{i=1}^N (\mathbf{u}_N(\mu))_i \phi_i$.

The above is a very concise description of the second step in the RB-method. From this description it is not immediately clear how this method can be split into an offline/online decomposition. We will come back to this in Section 2.1.2. First, the stability of the equation above is of interest. It turns out that the condition number of the matrix $A_N(\mu)$ is bounded independent of N if the basis Φ_N is orthonormal:

Proposition 2.1. *Let $\mu \in \mathcal{P}$, if $a(\cdot, \cdot; \mu)$ is symmetric and Φ_N is an orthonormal basis, then we have the following bound for the condition number of $A_N(\mu)$:*

$$\text{cond}_2(A_N(\mu)) := ||A_N(\mu)||_2 ||A_N(\mu)^{-1}||_2 \leq \frac{\gamma(\mu)}{\alpha(\mu)} \leq \frac{\overline{\gamma}}{\overline{\alpha}}.$$

Proof. Slightly adapted from [13]. By assumption $A_N(\mu)$ is a symmetric and positive definite matrix, hence $\text{cond}_2(A_N(\mu))$ is bounded by the ratio between the biggest and smallest eigenvalue of $A_N(\mu)$. Firstly, because of orthogonality, for $u = \sum_{i=1}^N \mathbf{u}_i \phi_i$, we have that $||\mathbf{u}||_2 = ||u||_X$. Now, if $\mathbf{u}, \lambda$ is an eigenpair of $A_N(\mu)$ we get the following:

$$\lambda ||\mathbf{u}||^2 = \mathbf{u}^T A_N(\mu)\mathbf{u} = a(\sum_{i=1}^N \mathbf{u}_i\phi_i, \sum_{i=1}^N \mathbf{u}_i\phi_i; \mu).$$

The last term is smaller than $\gamma(\mu)||\mathbf{u}||^2$ by continuity and bigger than $\alpha(\mu)||\mathbf{u}||^2$ by coercivity. Hence, the eigenvalues lie in the interval $[\alpha(\mu), \gamma(\mu)]$, so that

$$\text{cond}_2(A_N(\mu)) \leq \frac{\gamma(\mu)}{\alpha(\mu)}.$$

□

Just like in the finite-elements method we can find some a posteriori error estimators for the RB-method. For this we first need a relation between the residual and the error.

Definition 2.2. For $\mu \in \mathcal{P}$ we define the residual $r(\cdot; \mu) \in V_H'$ as

$$r(v; \mu) := f(v; \mu) - a(u_N(\mu), v; \mu),\ v \in V_H$$

We define the error $e(\mu)$ as $e(\mu) := u_H(\mu) - u_N(\mu) \in V_H$.

Proposition 2.3. *We have the equality*

$$a(e, v; \mu) = r(v; \mu) \text{ for all } v \in V_H.$$

Proof.

$$a(e, v; \mu) = a(u_H(\mu), v; \mu) - a(u_N(\mu), v; \mu) = f(v; \mu) - a(u_N(\mu), v; \mu) = r(v; \mu).$$

□

Now, we can use this relation to find an a posteriori error bound. For this a lower bound $\alpha_{LB}(\mu)$ can also be used; if this is not available one can use $\overline{\alpha}$. This error bound can be used later in the construction of the parameter set S_N. More information about this and other a posteriori error estimators can be found in [13].

Proposition 2.4. *For $\mu \in \mathcal{P}$ we have the following error bound:*

$$||u_H(\mu) - u_N(\mu)||_X \leq \Delta_u(\mu) := \frac{||r(\cdot; \mu)||_{V_H'}}{\alpha_{LB}(\mu)}.$$

Proof. We can perform the following calculation:

$$\alpha_{LB}(\mu)||e||^2 \leq a(e, e; \mu) = r(e; \mu) \leq ||r(\cdot; \mu)||_{V_H'}||e||$$

□

2.1.2 Offline/Online Decomposition

In this section we show how the reduced basis method can be implemented. The goal is to find a proper offline/online decomposition to spread the workload associated to applying the RB-method. Obviously, we want the 'online part' to only involve computations with a complexity dependent on N. Hence, evaluating the snapshots $u_H(\mu_i)$, $i = 1, \ldots, N$ should be done offline. On the other hand, solving (2.1.1) with known $\mathbf{A}_N(\mu)$ and $\mathbf{f}_N(\mu)$ can be done in $\mathcal{O}(N^3)$ computations and therefore can and should be done online. However, the number of computations for constructing the matrix $\mathbf{A}_N(\mu)$ and the vector $\mathbf{f}_N(\mu)$ are not only dependent on N as they involve evaluations of $a(\cdot, \cdot; \mu)$ and $f(\cdot; \mu)$. Fortunately, using the parameter-separability we can find a online/offline decomposition for computing the matrix $\mathbf{A}_N(\mu)$ and the vector $\mathbf{f}_N(\mu)$. For this we need to compute the matrices of the next definition:

Definition 2.5. We define the following matrices which are to be computed in the offline phase:

$$\mathbf{A}_{N,q} := (a_q(\phi_j, \phi_i))_{i,j=1}^N \in \mathbb{R}^{N \times N}, q = 1, \ldots, Q_a \tag{2.1.2}$$

$$\mathbf{f}_{N,q} := (f_q(\phi_i))_{i=1}^N \in \mathbb{R}^N, q = 1 \ldots, Q_f \tag{2.1.3}$$

Using the definition above we now get the following equalities which follow from linearity and (2.0.2) and (2.0.3):

$$\mathbf{A}_N(\mu) = \sum_{i=1}^{Q_a} \theta_q^a(\mu)\mathbf{A}_{N,q}$$

$$\mathbf{f}_N(\mu) = \sum_{q=1}^{Q_f} \theta_q^f(\mu)\mathbf{f}_{N,q}.$$

Hence, we can outline the RB-method as follows:

Offline

- Choose μ_i, $i = 1, \dots, N$ according to some principle.
- Calculate the snapshots $u(\mu_i)$, $i = 1, \dots, N$ using a finite element method and form an orthogonal basis $\Phi_N := \{\phi_1, \dots, \phi_N\}$ with span equal to $X_N = \text{span}\{u(\mu_1), \dots, u(\mu_N)\}$.
- Compute the matrices:

$$\mathbf{A}_{N,q} := (a_q(\phi_j, \phi_i))_{i,j=1}^N, q = 1, \dots, Q_a$$
$$\mathbf{f}_{N,q} := (f_q(\phi_i))_{i=1}^N, q = 1, \dots, Q_f$$

Online

- For the desired $\mu \in \mathcal{P}$, evaluate

$$\mathbf{A}_N(\mu) = \sum_{i=1}^{Q_a} \theta_q^a(\mu) \mathbf{A}_{N,q}$$
$$\mathbf{f}_N(\mu) = \sum_{q=1}^{Q_f} \theta_q^f(\mu) \mathbf{f}_{N,q}.$$

- Solve $\mathbf{A}_N(\mu)\mathbf{u}_N(\mu) = \mathbf{f}_N(\mu)$.
- Evaluate $u_N(\mu) = \sum_{i=1}^N (\mathbf{u}_N(\mu))_i \phi_i$.

Remark 2.6. We can also find an offline/online decomposition for the a posteriori error $\frac{||r(\cdot;\mu)||_{V_H'}}{\alpha_{LB}(\mu)}$. For this the parameter-separability of a and f is used, the Riesz representation theorem and the fact that u_N is an element of X_N. Namely, we can write

$$r(v;\mu) = \sum_{q=q}^{Q_f} \theta_q^f(\mu) f_q(v) - \sum_{i=1}^{N} \sum_{q=1}^{Q_a} (\mathbf{u}_N(\mu))_i \theta_q^a(\mu) a_q(\phi_i, v),$$

where $\{\phi\}_{i=1}^N$ is the reduced basis. By taking the Riesz representatives of both sides we get a parameter-separable vector. The computation of the norm of this vector can then be decomposed into offline/online phases. For the details see [13].

2.2 Choice of parameters

In this section we adress the question of how to choose the parameters μ_i, $i = 0, \dots, N-1$ that are used for the construction of the space $X_N = \text{span}(u_H(\mu_0), \dots, u_H(\mu_{N-1}))$. The goal is to find parameters in such a way that the global error

$$E_N := \sup_{\mu \in \mathcal{P}} ||u(\mu) - u_N(\mu)||_X$$

is small. Céa's lemma shows that the global error is small if the distance between X_N and $\mathcal{M}_H$

$$d(\mathcal{M}_H, X_N) = \sup_{\mu \in \mathcal{P}} \inf_{v \in X_N} ||u_H(\mu) - v||_X.$$

is small:

Lemma 2.7 (Cea)**.** *For all $\mu \in \mathcal{P}$ we have*

$$||u_H(\mu) - u_N(\mu)||_X \leq \frac{\gamma(\mu)}{\alpha(\mu)} \inf_{v \in X_N} ||u_H(\mu) - v||.$$

Minimization of the error E_n or $d(\mathcal{M}, X_N)$ is a very complex problem, and is not done in practice [13]. Instead, the parameters are usually chosen by a greedy algorithm. In a greedy algorithm new parameters are iteratively added using some rule. This algorithm is described below, here $\Delta(X_N, \mu)$ is an error indicator that approximates $\inf_{v \in X_N} ||u(\mu) - v||_X$. For the algorithm to terminate the error indicator should satisfy $\Delta(Y, \mu) = 0$ if $u(\mu) \in Y$. We provide some examples of choices for $\Delta(\cdot, \cdot)$ below. The training set S_{train} is a finite subset of $\mathcal{P}$. This set should be carefully chosen as the theory supporting this algorithm supposes this training set to be the entire parameter set $\mathcal{P}$.

Assumption 2.8 (Greedy Algorithm; from [13])**.** Let $S_{train} \subset \mathcal{P}$ be a training set of parameters
and $\epsilon_{tol} > 0$ a given error tolerance. Set $X_0 = \{0\}$, $S_0 = \varnothing$, $\Phi_0 = \varnothing$, then we have the following algorithm:

---

1 $n = 1$
2 while $\epsilon_n := \max_{\mu \in S_{train}} \Delta(X_n, \mu) > \epsilon_{tol}$
3 $\mu_{n+1} := \operatorname{argmax}_{\mu \in S_{train}} \Delta(X_n, \mu)$
4 $S_{n+1} := S_n \cup \{\mu_{n+1}\}$
5 $\phi_{n+1} = u_H(\mu_{n+1})$
6 orthonormalize ϕ_{n+1} to Φ_n
7 $\Phi_{n+1} = \Phi_n \cup \{\phi_{n+1}\}$
8 $X_{n+1} = X_n + \operatorname{span}(\phi_{n+1})$
9 $n = n + 1$

Remark 2.9. Here is a list of possible choices for the error indicator $\Delta(X_N, \mu)$ from [13]:

- We can use the projection error defined as

  $$\Delta(X_N, \mu) := \inf_{v \in X_N} ||u_H(\mu) - v|| = ||u_H(\mu) - P_{X_N} u_H(\mu)||,$$

  with P_{X_N} the orthogonal projection operator onto X_N. An greedy algorithm using this error indicator is called a 'strong greedy' procedure. A downside is that this error indicator is expensive to use as we need to have $u_H(\mu)$ available for all $\mu \in S_{train}$.

- Secondly, we can use the true RB-error as an estimate. Namely, we can define

  $$\Delta(X_N, \mu) = ||u_H(\mu) - u_N(\mu)||.$$

  Using this error indicator we forcefully try to minimize the global error E_N. Again, this error indicator is expensive to use however as we need to have $u(\mu)$ available for all $\mu \in S_{train}$. We also need to have an RB-model available in each step.

- Lastly, we can use an a-posteriori estimator $\Delta_u(\mu)$. Namely, we define

  $$\Delta(X_N, \mu) = \Delta_u(\mu).$$

A greedy algorithm using this error estimate is called 'weak gready'. This algorithm can use a much larger training set S_{train} than the algorithms above as the error indicator is cheap to evaluate; we don't need to have the snapshots available for all $\mu \in S_{train}$.

One needs to be aware that the greedy algorithm is an learning algorithm, hence 'overfitting' or 'underfitting' may occur. Therefore, one always need check the results (for example using another training set $\hat{S}_{train}$).

Remark 2.10. Alternatively, we could choose the nodes of some interpolation operator to be the parameters for the RB-method. For these parameters we readily have an a priori error-analysis; see chapter 1.

2.3 Error Analysis

We now give an error analysis for the greedy method. This section is based on [15]. In this section we are working with the space X, the same results holds if we replace X with V_H in this section. Furthermore, we assume that X is a separable Hilbert space, so that, without loss of generality, $X = \ell^2$. Furthermore, we assume that we have a sequence of functions $\{u_0, u_1, \ldots\} \subset \ell^2$ generated by the greedy algorithm. Again, we denote $X_n := \text{span}(u_1, \ldots, u_{n-1}) = \text{span}(u_1^*, \ldots, u_{n-1}^*)$, where $\{u_0^*, u_1^*, \ldots\}$ is obtained from the first sequence by Gram-Schmidt orthogonalization. We may also assume that $u_i^* = e_i \in \ell^2$. Furthermore, we define the infinite matrix $A = (a)_{i,j=0}^{\infty}$ through the following relation:

$$u_i = \sum_{j=0}^{i} a_{i,j} u_j^*, \; a_{i,j} = \langle u_i, u_j^* \rangle$$

and $a_{i,j} = 0$ for $j > i$. We now define some notions of distance:

Definition 2.11. Let X_n be a subspace generated by a reduced basis method then for $u \in \mathcal{M}$ we define the projection error as

$$\sigma_n(u) := ||u - P_n u||_X,$$

where P_n is the orthogonal projection operator onto X_n. Additionally, we define the maximal projection error as

$$\sigma_n := \max_{u \in \mathcal{M}} \sigma_n(u).$$

Definition 2.12. If $\mathcal{M}$ is a compact solution space, then for $n \geq 1$ we define the Kolmogorov n-width as

$$d_n(\mathcal{M}) := \inf_{\dim(Y)=n} \sup_{u \in \mathcal{M}} \text{dist}(u, Y).$$

If $n = 0$ we define

$$d_0(\mathcal{M}) = \sup_{u \in \mathcal{M}} \sigma_0(u)$$

The space which attains the Kolmogorov width is essentially the best Lagrangian space we can hope for. Obviously, the greedy algorithm cannot be expected to produce the best possible space, but we can find a relation between the decay rate of the Kolmogorov width and the global error of the Lagrangian basis X_N produced by a greedy algorithm. Before we show this we make an assumption on the greedy algorithm used. Note that this assumption can only be valid if the training set S_{train} is sufficiently big or equal to $\mathcal{P}$.

Assumption 2.13. For building the Lagrangian space $X_N \subset \mathcal{M}$ we assume that there is a $\gamma \leq 1$ such that the greedy algorithm does the following:

---

1 pick $u_0 \in \mathcal{M}$ such that $||u_0|| \geq \gamma\sigma_0(\mathcal{M})$
2 for $n = 1, \ldots, N$
3 pick $u_n \in \mathcal{M}$ such that $\sigma_n(u_n) \geq \gamma\sigma_n$

For $\gamma = 1$, we have the 'strong' greedy algorithm.

We are now ready to prove two things about the matrix A:

Proposition 2.14. *The matrix A satisfies the following:*

- *for $n \geq 0$, we have $\gamma\sigma_n \leq |a_{n,n}| \leq \sigma_n$,*
- *for all $m \geq n$ we have $\sum_{j=n}^{m} a_{m,j}^2 \leq \sigma_n^2$*

Proof. For the first statement notice that $a_{n,n}^2 = ||u_n - \sum_{j=0}^{n-1} an, ju_j^*||^2 = \sigma_n(u_n)^2 \geq \gamma^2\sigma_n^2$ and $\sigma(u_n) \leq \sigma_n$. For the second statement we notice that $\sum_{j=n}^{m} a_{m,j}^2 = ||u_m - \sum_{j=1}^{n-1} a_{m,j}||^2 = \sigma_n(u_m) \leq \sigma_n^2$ □

With these two properties we can prove the following useful lemma. In this lemma it is shown that the Kolmogorov width is of some size if the error of the reduced basis method σ_n stagnates.

Lemma 2.15. *Let $0 < \theta < 1$ be some constant, and define $q := \lceil 2\gamma^{-1}\theta^{-1} \rceil^2$. If for some $n, n \in \mathbb{N}$ it holds that $\sigma_{n+qm} \geq \theta_n\sigma_n$ then we have $\sigma_n(\mathcal{M}) \leq q^{\frac{1}{2}} d_m(\mathcal{M})$*

Proof. Adapted from [15]. The idea of the proof is to show that the space $X_{n+qm} \subset \mathcal{M}$ contains a subspace with a Kolmogorov width that is of the order σ_n (note that q is a whole number). This would imply that $d_m(\mathcal{M})$ is of that same order. More specifically, the space that we consider is the space $\hat{X}$ spanned by $\{g_i \colon i = 0, \ldots qm\}$, where $g_i := \sum_{j=n}^{n+qm} a_{i,j}u_i^*$. Now, let Y be a m-dimensional subspace of X and. Let $\phi_1, \ldots, \phi_m$ be an orthogonal basis for Y_m. Since for all i, ϕ_i has norm one we infer that for some $k \in \{0, \ldots, qm\}$

$$\sum_{i=1}^{m} |\phi_i(n+k)|^2 \leq q^{-1}.$$

Now, let y_k be the projection of g_k then we find that

$$\begin{aligned}|y_k(n+k)| &= |\sum_{i=1}^{m}\langle g_k,\phi_i\rangle\phi_i(n+k)\\ &\leq \left(\sum_{i=1}^{m}|\langle g_k,\phi_i\rangle|^2\right)^{1/2}\left(\sum_{i=1}^{m}|\phi_i(n+k)|^2\right)^{1/2}\\ &\leq q^{-1/2}||g_k||\\ &\leq q^{-1/2}\sigma_n,\end{aligned}$$

here we used the second statement of Prop. 2.14. Additionally, we have the following inequality:

$$|g_k(n+k)| = |a_{n+k,n+k}| \geq \gamma\sigma_{n+k} \geq \theta\sigma_n \geq 2q^{-1/2}\sigma_n.$$

Hence,

$$||g_k - y_k||_X \geq |g_k(n+k) - y_k(n+k)| \geq q^{-1/2}\sigma_n.$$

Therefore, $\sup_{u\in\hat{X}}\operatorname{dist}(u,Y) \geq q^{-1/2}\sigma_n$, and because Y was arbitrary we also find $d_m(\hat{X}) \geq q^{-1/2}\sigma_n$. Now, because $\hat{X}$ is a subspace of $\mathcal{M}$ we see that $d_m(\mathcal{M}) \geq d_m(\hat{X}) \geq q^{-1/2}\sigma_n$. □

Using this lemma we can show exponential decay of the projection error of the RB-method whenever the Kolmogorov width decreases exponentially.

Theorem 2.16 (From [15])**.** *Let $0 < \gamma \leq 1$ be the parameter in Assumption 2.13. Suppose that*

$$d_n(\mathcal{M}) \leq Me^{-an^\alpha}, \;\; n \geq 0,$$

for some $M, a, \alpha > 0$. Then we have that

$$\sigma_n(\mathcal{M}) \leq CMe^{-cn^\beta}, \;\; n \geq 0, \tag{2.3.1}$$

where $\beta := \frac{\alpha}{\alpha+1}$, $C := \max\{e^{cN_0^\beta}, q^{1/2}\}$, $c := \min\{|\log\theta|, (4q)^{-\alpha}a\}$, $0 < \theta < 1$, $q := \lceil 2\gamma^{-1}\theta^{-1}\rceil^2$, and $N_0 := \lceil (8q)^{\frac{1}{1-\beta}}\rceil = \lceil (8q)^{\alpha+1}\rceil$.

Proof. From [15]. By definition of C we have that (2.3.1) holds for all $n \leq N_0$:

$$\sigma_n \leq \sigma_0 \leq \sigma_0 e^{cN_0^\beta}e^{-cn^\beta} \leq CMe^{-cn^\beta}.$$

Now, suppose that (2.3.1) doesn't hold for all $n \geq 0$, and let $N > N_0$ be the smallest integer such that

$$CMe^{-cN^\beta} < \sigma_N.$$

We now derive a contradiction. First let m be a positive integer such that

$$e^{c(N-qm)^\beta}e^{-cN^\beta} \geq \theta \tag{2.3.2}$$

Then because (2.3.1) holds for $N - qm$ we have

$$\sigma_{N-qm} \leq CMe^{-c(N-qm)^\beta} \leq \theta^{-1}e^{-cN^\beta} \leq \theta^{-1}\sigma_N.$$

Hence, by applying Lemma 2.15 we find that

$$\sigma_N \leq \sigma_{N-qm} \leq q^{1/2}d_m\mathcal{M} \leq q^{1/2}Me^{-am^\alpha}.$$

This almost finishes the proof. We only need to show that there exist an m satisfying (2.3.2) such that

$$q^{1/2}Me^{-am^\alpha} \leq CMe^{-cN^\beta}. \tag{2.3.3}$$

A possible choice for m is given by $m := \lfloor \frac{N^{1-\beta}}{2q} \rfloor$. For this choice it holds that $N - qm \geq N/2$, hence by the mean value theorem, we have for $\xi \in (N - qm, N)$

$$N^\beta - (N - qm)^\beta = \beta\xi^{\beta-1}qm \leq qm\beta\left(\frac{N}{2}\right)^{\beta-1} \leq \frac{N^{1-\beta}}{2}\beta\left(\frac{N}{2}\right)^{\beta-1} \leq 2^{-\beta}\beta \leq 1.$$

Hence, (2.3.2) holds for this m if we consider the definition of c.

To conclude, for (2.3.3), notice that $N > N_0$ so that $m \geq 4$ and therefore $m \geq N^{1-\beta}/4q$. Hence, from the definition of c we have

$$am^\alpha - cN^\beta \geq a\left(\frac{N^{1-\beta}}{4q}\right)^\alpha - cN^\beta = (a(4q)^{-\alpha} - c)N^\beta \geq 0 \geq \frac{1}{2}\log q - \log C.$$

The last inequality is equivalent to (2.3.3). □

In the theorem above we didn't take numerical errors into account. The conditions in Assumption 2.13 are not necessarily met. Namely, instead of 'finding' $u_n \in \mathcal{M}$ such that $\sigma_n(u_n) \geq \gamma\sigma_n$, we can only expect to find a noisy version of u_n, namely $\hat{u}_n$, such that $||u_n - \hat{u}_n|| \leq \epsilon$. Here ϵ is some bound on the numerical error. Additionally, we cannot guarantee that $\hat{u}_n \in \mathcal{M}$. More realistic assumptions on the greedy algorithm would be the following:

Assumption 2.17. For building the Lagrangian space $X_N \subset X$ we assume that there is a $\gamma \geq 1$ and $\epsilon > 0$ such that the (weak) greedy algorithm does the following:

---

1 pick $\hat{u}_0 \in \mathcal{M}$ such that $||u_0 - \hat{u}_0|| < \epsilon$, and $||u_0|| \geq \gamma\sigma_0(\mathcal{M})$
2 for $n = 1, \ldots, N$
3 pick $\hat{u}_n \in \mathcal{M}$ such that $||u_n - \hat{u}_n|| < \epsilon$, and $\sigma_n(u_n) \geq \gamma\sigma_n$

In [15] an alternative version of Theorem 2.16 is proven for this more realistic case. For the proof we refer to the same article.

Theorem 2.18. *Suppose that*

$$d_n(\mathcal{M}) \leq Me^{-an^\alpha}, \quad n \geq 0,$$

for some $M, a, \alpha > 0$. Then one has

$$\sigma_n \leq C\max\{Me^{-cn^\beta}, \epsilon\}, \quad n \geq 0,$$

where $\beta := \frac{\alpha}{1+\alpha}$ and c, C are constants that depend on a, α and on γ.

2.3.1 Kolmogorov width for holomorphic solution manifolds

In the previous section we found a relation between the projection error σ_n and the Kolmogorov n-width $d_n(\mathcal{M})$. For this relation to be useful we actually need to show that the Kolmogorov widths of $\mathcal{M}$ decrease fast enough. In this section we show that this is the case for the image $\mathcal{M}$ of a holomorphic mapping. For this we use the theory we developed earlier for polynomial interpolation.

Case $J = 1$

We first estimate the Kolmogorov width for a function $f\colon [-1,1] \to X$ where X is Banach space and f has an holomorphic extension to D_ρ with $\rho > 1$. For this we need to define the Chebyshev series like in Theorem 1.9, but now for Banach-valued functions.

Definition 2.19. Let $g\colon [-1,1] \to X$ be a function where X is Banach and assume that g has a holomorphic extension to D_ρ, for some $\rho > 1$. For $n \in \mathbb{N}$ we define the truncated Chebyshev series $p_n^g(x)$ of g as

$$p_n^g(x) := \frac{1}{2}a_0^g + \sum_{k=1}^{n} a_k^g C_k(x),$$

where

$$a_k^g = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\cos t)\cos kt dt.$$

Now, in the proof of Theorem 1.22 we used that $\phi(\mathcal{I}f) = \mathcal{I}(\phi \circ f)$ for all $\phi \in X'$ whenever $\mathcal{I}$ is an interpolation operator. We would like to see that this property also holds for the Chebyshev series. The following calculation indeed shows that $p_n^{\phi\circ f}(x) = \phi(p_n^f(x))$ for all $\phi \in X'$ and $x \in [-1,1]$:

$$\begin{aligned}
a_k^{\phi\circ f} &= \frac{1}{\pi}\int_{-\pi}^{\pi} \phi(f(\cos t))\cos kt dt \\
&= \frac{1}{\pi}\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\phi(f(\cos(\frac{2i\pi}{N})))\cos(\frac{2i\pi}{N}) \\
&= \frac{1}{\pi}\lim_{N\to\infty}\phi\left[\frac{1}{N}\sum_{i=1}^{N} f(\cos(\frac{2i\pi}{N}))\cos(\frac{2i\pi}{N})\right] \\
&= \frac{1}{\pi}\phi\left[\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} f(\cos(\frac{2i\pi}{N}))\cos(\frac{2i\pi}{N}))\right] \\
&= \phi\left[\frac{1}{\pi}\int_{-\pi}^{\pi} f(\cos t)\cos kt dt\right] \\
&= \phi(a_k^f)
\end{aligned}$$

The above allows us to use the proof of Theorem 1.22 and apply Theorem 1.9 to obtain the estimate

$$||f - p_n^f||_\infty \leq \frac{2||f||_\infty}{\rho - 1}\rho^{-n} = \frac{2||f||_\infty}{\rho(\rho-1)}\rho^{-(n-1)}. \tag{2.3.4}$$

Hence, if we define the subspace $Y_{n+1} = \text{span}\{a_0^f, \ldots, a_n^f\}$, then, for $\mathcal{M} := f([-1,1])$, we find that

$$d_n(\mathcal{M}) \leq \sup_{u \in \mathcal{M}} \text{dist}(u, Y_n) \leq ||f - p_n^f||_\infty \leq \frac{2||f||_\infty}{\rho(\rho-1)} \rho^{-n}. \tag{2.3.5}$$

We have proven exponential decay of the Kolmogorov width of $\mathcal{M}$.

Higher-dimensional case

We now assume that $J \geq 2$, so that $f\colon [-1,1]^J \to X$ has a holomorphic extension to D_ρ^J. In order to find an estimate for the Kolmogorov width of $\mathcal{M} := f([-1,1]^J)$ we use the same strategy as in the one-dimensional case. Namely, we define a multi-variate polynomial with coefficients in X that approximates f. We then use these coefficients to form a space that provides an estimate for the Kolmogorov width. We have already seen in the chapter about polynomial interpolation that the sparse grid interpolator is able to interpolate better than the full-tensor grid interpolator while using less interpolation nodes. Therefore, we define an operator using the Chebyshev series, and using the notation for these sparse grids. To be more specific, we use the same notation as in Section 1.2.2, but now we choose $\mathcal{U}^i f := p_{m_i - 1}^f = p_{m_i - 1}(f)$ and define

$$\mathcal{A}(q, J) f := \sum_{|\nu| \leq q} \bigotimes_{i=1}^{J} \Delta_{\nu_i} f.$$

We again choose $m(n) = n$, like we did for the Smolyak algorithm using Leja points.

First we need to know the number of coefficients in X the multi-variate polynomial $\mathcal{A}(q, J)f$ has.

Lemma 2.20. *The number of coefficients Q in $\mathcal{A}(q, J)f$ is equal to the $|\Gamma|$, which is the size of the grid of the corresponding sparse grid interpolator defined in Remark 1.18. This leads to the estimates*

$$q \cdot 2 \left(\frac{q + J - 2}{2J - 2} \right)^{J-1} \leq Q \leq \frac{q(q-1)^{J-1}}{(J-1)! \cdot 2}.$$

Proof. Notice that the $\mathcal{A}(q, J)f$ lies in the space

$$\sum_{|\nu| = q} X \otimes \left(\bigotimes_{k=1}^{J} \mathbb{P}_{m_{\nu_k} - 1} \right).$$

We count the coefficients index-wise. Hence, for each $|\nu| \leq q$ we count the terms of the type $x_1^{c_1} \cdot \ldots \cdot x_J^{c_J}$, $c_i \in \mathbb{N}_0$ that are in $\bigotimes_{k=1}^J \mathbb{P}_{m_{\nu_k} - 1}$ but not in $\bigotimes_{k=1}^J \mathbb{P}_{m_{\hat{\nu}_k} - 1}$ for all $\hat{\nu} \leq \nu$, $\hat{\nu} \neq \nu$ to prevent counting terms multiple times. The number of terms we need to count for each ν are clearly equal to $|\Gamma_\nu|$ as defined in Remark 1.18. Hence

$$Q = \sum_{|\nu| \leq q} |\Gamma_\nu| = |\Gamma|.$$

Now the size of the grid $|\Gamma|$ is equal to $\sum_{|\nu|\leq q} 1$. This can be estimated in the following way:

$$\sum_{|\nu|\leq q} 1 = \sum_{n=J}^{q} \binom{n-1}{J-1} \leq \sum_{n=1}^{q} \frac{(n-1)^{(J-1)!}}{J-1} \leq \frac{q(q-1)^{J-1}}{(J-1)!\cdot 2}.$$

The last inequality follows from Gauss' method for summation. For the lower bound we estimate in the following way:

$$\sum_{n=J}^{q} \binom{n-1}{J-1} \geq \sum_{n=J}^{q} \left(\frac{n-1}{J-1}\right)^{J-1} \geq q \cdot 2 \left(\frac{q+J-2}{2J-2}\right)^{J-1}$$

□

Next we provide an error estimate for $\mathcal{A}(q,J)f$ in terms of q.

Lemma 2.21. *Whenever $f\colon [-1,1]^J \to \mathbb{R}$ has an holomorphic extension to D_ρ^J we have the error estimate*

$$||(I_J - \mathcal{A}(q,J))f||_\infty \leq \frac{2^J M}{(\rho-1)^J} \rho^{2J} \binom{q}{J-1} \rho^{-q}$$

Proof. We can generalize the argumentation used in Section 1.2.2. Namely, we can estimate

$$\begin{aligned}
||(\mathrm{Id} - \mathcal{A}(q,J))f||_\infty &= \lim_{N\to\infty} || \sum_{\substack{|\nu|>q \\ \nu\leq(N,\dots,N)}} \Delta_\nu f||_\infty \\
&\leq \lim_{N\to\infty} \sum_{|\nu|=q+1} || \sum_{\substack{\mu_i\geq\nu_i \\ \nu\leq(N,\dots,N)}} \Delta_\mu f||_\infty \\
&= \lim_{N\to\infty} \sum_{|\nu|=q+1} || \bigotimes_{i=1}^{J} (\mathcal{U}^N - \mathcal{U}^{\nu_n-1}) f||_\infty \\
&\leq \sum_{|\nu|=q+1} M \prod_{i=1}^{J} \frac{2}{\rho-1} \rho^{-\nu_n+2} \\
&= \sum_{|\nu|=q+1} 2^J M \frac{\rho^{-q+2J}}{(\rho-1)^J} \\
&= \frac{2^J M}{(\rho-1)^J} \rho^{2J} \binom{q}{J-1} \rho^{-q}
\end{aligned}$$

□

Theorem 2.22. *Let $f\colon [-1,1]^J \to X$ be a Banach-valued function that can be extended holomorphically to D_ρ^J. Now, if Q is the number of coefficients of $\mathcal{A}(q,J)f$ for some q, then for $n \geq Q$ we have that the Kolmogorov n-width of $\mathcal{M} := f([-1,1]^J)$ satisfies*

$$d_n(\mathcal{M}) \leq \min\{||f||_\infty, C_J M \frac{\rho^{2J}}{(\rho-1)^J} Q \cdot \beta(J,\rho)^{-\sqrt[J]{Q}}\},$$

with $C_J > 0$ some constant dependent on J and $\beta(J,\rho) := \rho^{[(J-1)!\cdot 2]^{1/J}}$.

Proof. From Lemma 2.20 we get

$$((J-1)!2Q)^{1/J} \leq q \leq \frac{Q^{1/J}(2J-2)}{2^{1/J}}.$$

Hence, after substitution we find

$$||(I_J - \mathcal{A}(q,J))f|| \leq \frac{2^J M}{(\rho-1)^J}\rho^{2J}\frac{(\frac{Q^{1/J}(2J-2)}{2^{1/J}})^{J-1}}{J-1}\rho^{-((J-1)!2Q)^{1/J}}.$$

After simplification we obtain the theorem. □

Remark 2.23. Using the theorem above it is most likely possible to prove exponential decay of the Kolmogorov width. This exponential decay is the only condition needed in Theorems 2.16 and 2.18.

3 Space-time variational formulation for parabolic equations of second order

In this chapter we provide a space-time variational formulation for parabolic equations. Usually, parabolic equations are solved using a semi-variational formulation. The problem with this approach for solving parabolic PDEs is that the reduced basis method cannot be applied directly; although there are ways to circumvent this [13]. Furthermore, one cannot show quasi-optimality of numerical approximations when using this method. In this chapter we show that the full-variational formulation results in an equation with a coercive bilinear form defined on a well-understood space. This full-variational formulation allows an direct application of the reduced basis method described in chapter 2. Consequently, quasi-optimal numerical approximations also exist.

We first give an abstract definition of a parabolic PDE. We then work towards providing a variational formulation that has nice properties and is easy to work with. We also show holomorphy of the solution map for parabolic PDEs whenever the coefficients are holomorphic.

3.1 Definition parabolic PDE

Before we introduce a parabolic PDE we first define the Gelfand triple. Let V, H be separable Hilbert spaces such that $V \hookrightarrow H$ with dense embedding. This means that V is dense in H and

$$||f||_H \leq C||f||_V \text{ for all } f \in V,$$

for some $C > 0$.

In this setting we can identify elements in H' with elements in H via the Riesz Representation Theorem. If we do this, we cannot identify V' with V in the same way (notice that $V \subsetneq H = H' \subsetneq V' = V$ would lead to a contradiction). However, there is a representation of V' after identifying $H' = H$, namely via the inner product on H.

This is done in the following way. Let $\langle \cdot, \cdot \rangle_H$ be the inner product on H. Pick $h \in H$ then $f_h(\cdot) := \langle h, \cdot \rangle_H$ defines an element in $H' \subset V'$. This map $f \colon H \to H'$ is an isomorphism by the Riesz representation Theorem and identifies H with H'. Now, if H' is dense in V', then this map can be extended continuously to V'. Namely, if $f \in V'$ then there are $(h_n) \subset H$ such that $f_{h_n} \to f$ in V' for $n \to \infty$. The next lemma states that H' is indeed dense in V'.

Lemma 3.1. *If V is dense in H and V, H are Banach spaces, then H' is dense in V'.*

Proof. Thanks to Hahn-Banach it suffices to prove that for every $\psi \in V''$ such that $\psi(H') = \psi(H) = 0$ it follows that $\psi = 0$.

Now, first define the canonical map $c \colon V \to V''$ by $c(v)(v') = v'(v)$. Since V is reflexive, c is an isomorphism. Hence, $\psi = c(v)$ for some $v \in V$. From this it follows that for all $h \in H$ we have that $0 = \psi(h') = c(v)(h') = h'(v) = \langle h, v \rangle$. Hence, $v = 0$ and therefore $\psi = c(v) = 0$. □

For the rest of this thesis a Gelfand triple $V \hookrightarrow H \hookrightarrow V'$ is treated as above.

Example 3.2. Let $\Omega \subset \mathbb{R}^n$, because $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$ it follows that $H_0^1(\Omega)$ is also dense in $L^2(\Omega)$. Hence, the spaces $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ and $V' = H^{-1}(\Omega)$ form the Gelfand triple $V \hookrightarrow H \hookrightarrow V'$. This is the most relevant Gelfand triple for this thesis. Notice that the set of functions $v \in L^2(\Omega)$ with

$$v(g) := \int_\Omega v(x)g(x)dx,$$

form an dense set in $H^{-1}(\Omega)$.

We are now ready to give an abstract definition of a parabolic PDE.

Definition 3.3. Let V, H such that $V \hookrightarrow H \hookrightarrow V'$ form a Gelfand triple and let $I = [0, T] \subset \mathbb{R}$ be an interval. For a.e. $t \in I$, let $a(t; \cdot, \cdot)$ be a sesqui-linear form on $V \times V$ such that for any η, $\zeta \in V$ $a(\cdot; \eta, \zeta)$ is measurable, and such that for some constants M_a, $\alpha > 0$, $\lambda \in \mathbb{R}$ and for a.e. $t \in I$, we have

$$|a(t; \eta, \xi)| \leq M_a ||\eta||_V ||\zeta||_V \text{ for all } \eta, \zeta \in V \tag{3.1.1}$$

$$\Re a(t; \eta, \eta) + \lambda ||\eta||_H^2 \geq \alpha ||\eta||_V^2 \text{ for all } \eta \in V. \tag{3.1.2}$$

For a.e. $t \in I$, let $A(t) \in \mathcal{L}(V, V')$ be defined by

$$\langle A(t)\eta, \zeta \rangle_H = a(t; \eta, \zeta). \tag{3.1.3}$$

Then for $g \in L^2(I; V')$ and $h \in H$, the parabolic equation is formulated as, find $u \colon I \to V$, with $\frac{du}{dt} \colon I \to V'$ such that

$$\frac{du}{dt}(t) + A(t)u(t) = g(t) \text{ in } V', u(0) = h \text{ in } H. \tag{3.1.4}$$

Example 3.4. The definition above is quite abstract and is used for theoretical purposes. In this thesis we look at parabolic PDEs of the following form where we set $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$:

$$\begin{cases} \partial_t u - \operatorname{div} A\nabla_x u + b \cdot \nabla_x u + cu &= f \quad \text{on } I \times \Omega, \\ u &= 0 \quad \text{on } I \times \partial\Omega, \\ u(0, \cdot) &= u_0 \quad \text{on } \Omega. \end{cases} \tag{3.1.5}$$

Here $\Omega \subset \mathbb{R}^d$, $b \in L^\infty(I \times \Omega)^d$, $c \in L^\infty(I \times \Omega)$, $A = A^T \in L^\infty(I \times \Omega)^{d \times d}$ uniformly positive definite and $f \in L^2(I \times \Omega)$. Here $a(t, u, v)$ defined as

$$a(t; u, v) := \int_\Omega A(t, x)\nabla u(x) \cdot \nabla v(x) + (b(t, x) \cdot \nabla u(x) + c(t, x)u(x))v(x)dx,$$

satisfies (3.1.1) which follows from the Cauchy-Schwarz inequality and the Gårding inequality (3.1.2) which follows from [12], chapter 5.6.

3.2 Variational formulation

In order to apply the reduced basis method directly we need to be able to formulate a parabolic PDE in a well-posed variational way:

$$a(\mu; u, v) = f(\mu; v).$$

This is relatively easy for elliptic PDEs. In this section we show that this is also possible for time-dependent PDEs, albeit in a more complicated way. The spaces necessary are much more involved. These space are the search-space $\mathcal{X} = L^2(I;V) \cap H^1(I;V')$ with norm (notice that $||v||_{L^2(I;V')} \leq C||v||_{L^2(I;V)}$)

$$||v||_{\mathcal{X}} := (||v||^2_{L^2(I;V)} + ||\frac{dv}{dt}||^2_{L^2(I;V')})^{\frac{1}{2}}. \tag{3.2.1}$$

and the test-space $\mathcal{Y} = L^2(I;V) \times H$ with norm

$$||(v_1, v_2)||_{\mathcal{Y}} = (||v_1||^2_{L^2(I;V)} + ||v_2||^2_H)^{\frac{1}{2}}. \tag{3.2.2}$$

Before we can move on, we need a result about $\mathcal{X}$:

Lemma 3.5 ([4]; Ch.XVIII, §1, Th.1)**.** *We have that $\mathcal{X} \hookrightarrow C([0,T];H)$. Furthermore, the quantity*

$$M_e := \sup_{0 \neq w \in \mathcal{X}} \frac{||w(0)||_H}{||w||_{\mathcal{X}}} \tag{3.2.3}$$

is bounded uniformly in the choice of $V \hookrightarrow H$, and is only dependent on T when it tends to zero.

Now, just like in any other case, to find a variational formulation, we multiply both sides of (3.1.4) with a test function in $(v_1, v_2) \in \mathcal{Y}$ and integrate. We find

$$\int_I \langle \frac{du}{dt}(t), v_1(t) \rangle_H + a(t; u(t), v_1(t))dt + \langle u(0), v_2 \rangle_H = \int_I \langle g(t), v_1(t) \rangle_H dt + \langle h, v_2 \rangle_H$$

Hence, for the solution u this equation holds for all $v \in \mathcal{Y}$. The important question now is whether the converse is true. For this we need to prove the following theorem:

Theorem 3.6 (From [1])**.** *The functional $B \in \mathcal{L}(\mathcal{X}, \mathcal{Y}')$ defined by*

$$(Bw)(v) := b(w, v) := \int_I \langle \frac{dw}{dt}(t), v_1(t) \rangle_H + a(t; w(t), v_1(t))dt + \langle w(0), v_2 \rangle_H \tag{3.2.4}$$

is boundedly invertible.

Proof. We give an summary of what is done in [1]. To prove that B is boundedly invertible we have to prove three things about the bilinear form b:

$$M_b := \sup_{0 \neq w \in \mathcal{X}, 0 \neq w \in \mathcal{Y}} \frac{|b(w,v)|}{||w||_{\mathcal{X}}||v||_{\mathcal{Y}}} < \infty, \tag{3.2.5}$$

$$\beta := \inf_{0 \neq w \in \mathcal{X}} \sup_{0 \neq w \in \mathcal{Y}} \frac{|b(w,v)|}{||w||_{\mathcal{X}}||v||_{\mathcal{Y}}} > 0, \tag{3.2.6}$$

$$\forall 0 \neq v \in \mathcal{Y}, \sup_{0 \neq w \in \mathcal{X}} |b(w,v)| > 0. \tag{3.2.7}$$

It is easily seen (see [5]; Theorem 4.48) that first inequality shows continuity of B and the second inequality shows that B is injective and that $\mathrm{Im}(B)$ is closed. To understand the last inequality, notice that $\mathcal{Y}'$ is the dual of a Hilbert space and that $B(\mathcal{X})$ is a closed linear subspace. Now, after identification of $\mathcal{Y}$ and $\mathcal{Y}'$, if $B(\mathcal{X})$ were a proper subspace of $\mathcal{Y}$ then it would need to have a orthogonal complement. But (3.2.7) shows that this is not the case.

In order to prove these three statements it is shown in the proof that you may assume that $\lambda = 0$. Now we make a straightforward calculation to show the first statement:

$$\begin{aligned}
|b(w,v)| &\leq |\int_I \langle \frac{dw}{dt}(t), v_1(t)\rangle_H + a(t; w(t), v_1(t))dt| + |\langle w(0), v_2\rangle_H| \\
&\leq \int_I |\langle \frac{dw}{dt}(t), v_1(t)\rangle_H| + |a(t; w(t), v_1(t))|dt + ||w(0)||_H||v_2||_H \\
&\leq \int_I ||\frac{dw}{dt}(t)||_{V'}||v_1(t)||_V + M_a||w(t)||_V||v_1(t)||_V dt + M_e||w||_{\mathcal{X}}||v_2||_H \\
&\leq ||\frac{dw}{dt}||_{L^2(I;V')}||v_1||_{L^2(I;V)} + M_a||w||_{L^2(I;V)}||v_1||_{L^2(I;V)} + M_e||w||_{\mathcal{X}}||v_2||_H \\
&\leq \sqrt{2\max\{1, M_a^2\} + M_e^2}||w||_{\mathcal{X}}||v||_{\mathcal{Y}}.
\end{aligned}$$

For the second statement we pick a arbitrary $w \in \mathcal{X}$. We want to find an $v_w \in \mathcal{Y}$ such that $|b(w, v_w)| \geq C||w||_{\mathcal{X}}||v_w||_{\mathcal{Y}}$, where $C > 0$ is independent of w. In fact, we can define $v_w = (v_1, v_2)$ as $v_1(t) = z_w(t) + w(t)$ and $v_2 = w(0)$, where $z_w(t) := (A(t)')^{-1}\frac{dw}{dt}(t)$. Here the adjoint $A(t)'$ satisfies $\langle A(t)'\eta, \zeta\rangle = \overline{a(t; \zeta, \eta)}$. In [1] it is eventually found that

$$|b(w, v_w)| \geq \frac{\min(\frac{\alpha}{M-a^2}, \alpha)}{\sqrt{2\max(\alpha^{-2}, 1) + M_e^2}}||w||\mathcal{X}||v_w||_{\mathcal{Y}}.$$

For the third statement pick an $0 \neq v = (v_1, v_2) \in \mathcal{Y}$. We want to find a $w \in \mathcal{X}$ such that $b(w, v) > 0$. We define the basis $\{\phi_i \colon i \in \mathbb{N}\}$ for V. Let $V_n = \mathrm{span}\{\phi_i \colon i = 1, \ldots, n\}$. We let $z_n(t) = \sum_{i=1}^n z_i^{(n)}(t)\phi_i$ such that

$$\left\{ \begin{array}{rcl} \langle \frac{dz_n}{dt}(t), \zeta_n\rangle_H + a(t; z_n(t), \zeta_n) & = & a(t; v_1(t), \zeta_n), \\ z_n(0) & = & \sum_{j=1}^n v_{2,i}^{(n)}\phi_i, \end{array} \right. \tag{3.2.8}$$

for all $\zeta_n \in V_n$ and a.e. $t \in [0, T]$, and $\sum_{j=1}^n v_{2,i}^{(n)}\phi_i \to v_2$ in H for $n \to \infty$. This equation is actually a system of linear ODE's, which has a unique solution $z_n \in C([0, T]; V_n)$, with $\frac{dz_n}{dt} \in L^2(0, T; V_n)$. If we assume that $V = V_n$ for some n then $b(z_n, v) = \int_0^T a(t; v_1(t), v_1(t))dt + \langle v_2, v_2\rangle_H$ which is obviously bigger than zero, which would be enough for the statement. Obviously V is often infinite-dimensional, but in [1] it is shown that z_n converges weakly to some z in $\mathcal{X}$, which still satisfies the inequality above. □

The problem with this variational formulation is that b is obviously not a coercive bilinear form; the space $\mathcal{X}$ and $\mathcal{Y}$ do not coincide. Secondly, the space $\mathcal{Y}$ has a norm that is difficult to compute. One possible solution is to transform this second order PDE into a first order system of PDEs and then use a least squares formulation of the problem. In [2] this technique is shown for the parabolic PDE formulated like in Example 3.4. We follow this article, but with the assumption that $u_\Omega = 0$. Hence, in this case $V = H_0^1(\Omega)$ and $H = L^2(\Omega)$, with $\Omega \subset \mathbb{R}^d$. Also, $\mathcal{X} = L^2(I; H_0^1(\Omega)) \cap H^1(I; H_0^1(\Omega)')$ and $\mathcal{Y} = L^2(I; H_0^1(\Omega)) \times L^2(\Omega)$. We now

show what was done there. We define the following space, which is in fact a Hilbert space (see [2]):

$$\mathcal{U} := \{u = (u_1, u_2) \in L^2(I; H^1(\Omega)) \times L^2(I \times \Omega)^d : \operatorname{div} u \in L^2((I \times \Omega)\}$$

with norm

$$||u||^2_{\mathcal{U}} = ||u_1||^2_{L^2(I;H^1(\Omega))} + ||u_2||^2_{L^2(I;L^2(\Omega)^d)} + ||\operatorname{div} u||^2_{L^2(I\times\Omega)}.$$

Here $\operatorname{div} u(t,x) = \partial_t u_1(t,x) + \operatorname{div}_x u_2(t,x)$. We then define the closed subspace $\mathcal{U}_0$ of $\mathcal{U}$:

$$\mathcal{U}_0 := \{u \in L^2(I; H^1_0(\Omega)) \times L^2(I \times \Omega)^d \colon \operatorname{div} u \in L^2(I \times \Omega)\}.$$

It turns out that this space $\mathcal{U}_0$ is the same as another space with an equivalent norm as the lemma below shows:

$$\mathcal{U}_0 \simeq \overline{\mathcal{U}}_0 := \{u \in \mathcal{X} \times L^2(I \times \Omega)^d \colon \operatorname{div} u \in L^2(I \times \Omega)\},$$

with norm

$$||u||^2_{\overline{\mathcal{U}}_0} = ||\partial_t u_1||^2_{L^2(I;H^1_0(\Omega)')} + ||u_1||^2_{L^2(I;H^1(\Omega))} + ||u_2||^2_{L^2(I;L^2(\Omega)^d)} + ||\operatorname{div} u||^2_{L^2(I\times\Omega)}.$$

Lemma 3.7. *For $u = (u_1, u_2) \in H_0(\operatorname{div}, I \times \Omega)$, it holds that $\partial_t u_1 \in L^2(I; H^{-1}(\Omega))$ with*

$$||\partial_t u_1||_{L^2(I;H^1_0(\Omega)')} \leq ||u||_{H(\operatorname{div};I\times\Omega)}.$$

Proof. Firstly, using the triangle inequality we obtain

$$||\partial_t u_1||_{L^2(I;H^1_0(\Omega)')} \leq ||\operatorname{div} u||_{L^2(I;H^1_0(\Omega)')} + ||\operatorname{div}_x u_2||_{L^2(I;H^1_0(\Omega)')}.$$

Then, because of the inequalities below we obtain that $||\operatorname{div} u(t,\cdot)||_{H^{-1}(\Omega)} \leq ||\operatorname{div} u(t,\cdot)||_{L^2(\Omega)}$ and $||\operatorname{div}_x u_2||_{H^{-1}(\Omega)} \leq ||u_2(t,\cdot)||_{L^2(\Omega)^d}$:

$$|\int_\Omega \operatorname{div} u(t,x) \cdot v(x) dx| \leq ||\operatorname{div} u(t,\cdot)||_{L^2(\Omega)} ||v||_{H^1(\Omega)}$$

$$|\int_\Omega \operatorname{div}_x u_2(t,x) v(x) dx| \leq |\int_\Omega u_2(t,x) \nabla_x v(x) dx| \leq ||u_2(t,\cdot)||_{L^2(\Omega)^d} ||v||_{H^1(\Omega)}.$$

Hence,

$$\begin{aligned}
||\partial_t u_1||_{L^2(I;H^1_0(\Omega)')} &\leq \left(\int_I ||\operatorname{div} u||^2_{H^{-1}(\Omega)} dt\right)^{1/2} + \left(\int_I ||\operatorname{div}_x u_2||^2_{H^{-1}(\Omega)} dt\right)^{1/2} \\
&\leq \left(\int_I ||\operatorname{div} u(t,\cdot)||^2_{L^2(\Omega)} dt\right)^{1/2} + \left(\int_I ||u_2(t,\cdot)||^2_{L^2(\Omega)^d} dt\right)^{1/2} \\
&\leq ||\operatorname{div} u||_{H(\operatorname{div};I\times\Omega)}
\end{aligned}$$

□

The space $\overline{\mathcal{U}}_0$ is used as a link between $\mathcal{U}_0$ and Theorem 3.6. Namely, define the following linear operator $G\colon \mathcal{U}_0 \to L^2(I\times\Omega)^d \times L^2(I\times\Omega)\times L^2(\Omega)$ given by

$$G\colon (u_1,u_2)\mapsto (u_2 + A\nabla_x u_1, \operatorname{div} u - b\cdot A^{-1}u_2 + cu_1, u_1(0,\cdot)).$$

Notice that the solution to the parabolic equation in Example 3.4 satisfies $G(u,-A\nabla_x u) = (0,f,u_0)$. This G is another way of formulating the parabolic PDE, but now as a first order system of equations. This formulation is a well-posed variational formulation as the next theorem shows.

Theorem 3.8 (From [2])**.** *The linear operator G is boundedly invertible.*

Proof. Pick $u\in\mathcal{U}_0$, and notice that $||u||_{\mathcal{U}_0} \lesssim ||u||_{\overline{\mathcal{U}}_0}$. It is now a straightforward calculation to show that G is bounded. Now, for injectivity, notice that

$$\begin{aligned}
||u_1||_{L^2(I;H^1(\Omega))} &\le ||u_1||_{\mathcal{X}}\\
&\lesssim ||Bu_1||_{\mathcal{Y}'}\\
&\le ||\operatorname{div} u - b\cdot A^{-1}u_2 + cu_1||_{L^2(I\times\Omega)}\\
&\quad + (1+||b||_\infty||A^{-1}||_\infty)||u_2 + A\nabla_x u_1||_{L^2(I\times\Omega)^d} + ||u(0)||_{L^2(\Omega)},
\end{aligned}$$

where the last inequality follows from integration by parts:

$$\begin{aligned}
B(u_1,v) &= \int_I \langle \frac{du_1}{dt}(t,x), v_1(t,x)\rangle_{L^2(\Omega)} + a(t;u_1(t,x),v_1(t,x))dt + \langle u(0),v_2\rangle_{L^2(\Omega)}\\
&= \int_I\int_\Omega \frac{du_1}{dt}(t,x)v_1(t,x) + A(t,x)\nabla u(t,x)\cdot\nabla v_1(t,x) + (b(t,x)\cdot\nabla u(t,x)\\
&+ c(t,x)u(t,x))v_1(t,x)dxdt + \langle u(0),v_2\rangle_{L^2(\Omega)}\\
&= \int_I\int_\Omega \frac{du_1}{dt}(t,x)v_1(t,x) - u_2(t,x)\cdot\nabla v_1(t,x) + (u_2(t,x) + A\nabla u_1(t,x))\nabla v_1(t,x)\\
&+ (b(t,x)\cdot\nabla u(t,x) + c(t,x)u(t,x))v_1(t,x)dxdt + \langle u(0),v_2\rangle_{L^2(\Omega)}\\
&= \int_I\int_\Omega \left(\frac{du_1}{dt}(t,x) + \operatorname{div}_x u_2(t,x)\right)v_1(t,x) + (u_2(t,x) + A\nabla u_1(t,x))\nabla v_1(t,x)\\
&+ (b(t,x)\cdot\nabla u(t,x) + c(t,x)u(t,x))v_1(t,x)dxdt + \langle u(0),v_2\rangle_{L^2(\Omega)}\\
&\le \Big(||\operatorname{div} u + b\cdot\nabla_x u_1 + cu_1||_{L^2(I\times\Omega)} + ||u_2 + A\nabla_x u_1||_{L^2(I\times\Omega)^d} + ||u(0)||_{L^2(\Omega)}\Big)||v||_{\mathcal{Y}}\\
&\le \big(||\operatorname{div} u - b\cdot A^{-1}u_2 + cu_1||_{L^2(I\times\Omega)}\\
&+(1+||b||_\infty||A^{-1}||_\infty)||u_2 + A\nabla_x u_1||_{L^2(I\times\Omega)^d} + ||u(0)||_{L^2(\Omega)}\Big)||v||_{\mathcal{Y}},
\end{aligned}$$

for all $v\in\mathcal{Y}$. Furthermore, we have

$$\begin{aligned}
||u_2||_{L^2(I\times\Omega)^d} &\le ||u_2 + A\nabla_x u_1||_{L^2(I\times\Omega)^d} + ||A\nabla_x u_1||_{L^2(I\times\Omega)^d}\\
&\le ||u_2 + A\nabla_x u_1||_{L^2(I\times\Omega)^d} + ||A||_\infty||u_1||_{L^2(I;H^1(\Omega))}
\end{aligned}$$

and

$$\begin{aligned}
||\operatorname{div} u||_{L^2(I\times\Omega)} &\le ||\operatorname{div} u - b\cdot A^{-1}u_2 + cu_1||_{L^2(I\times\Omega)} + ||u||_{L^2(I\times\Omega)^{d+1}}\\
&\le ||\operatorname{div} u - b\cdot A^{-1}u_2 + cu_1||_{L^2(I\times\Omega)} + ||c||_\infty||u_1||_{L^2(I;H^1(\Omega))}\\
&+ ||b||_\infty||A^{-1}||_\infty||u_2||_{L^2(I\times\Omega)^d}.
\end{aligned}$$

Combining all this gives $||u||_{\mathcal{U}} \lesssim ||Gu||_{L^2(I\times\Omega)^d \times L^2(I\times\Omega)\times L^2(\Omega)}$. This shows injectivity of G. We now prove surjectivity. Pick $(q, h, u_0) \in L^2(I \times \Omega)^d \times L^2(I \times \Omega) \times L^2(\Omega)$ and let $u_1 \in \mathcal{X}$ be the solution of

$$Bu_1 = [v \mapsto \int_{I\times\Omega} (h + b \cdot A^{-1}q)v_1 + q \cdot \nabla_x v_1 dxdt + \langle u_0, v_2\rangle_H] \in \mathcal{Y}'.$$

Then, for $v \in L^2(I; H_0^1(\Omega))$ and $u_2 := q - A\nabla_x u_1 \in L^2(I \times \Omega)^d$ we get

$$\int_{I\times\Omega} \partial_t u_1 v - u_2 \cdot \nabla_x v dxdt = \int_{I\times\Omega} (h + b \cdot A^{-1}u_2 - cu_1)v dxdt.$$

Now, if $v \in H_0^1(I\times\Omega)$ is smooth then $\int_{I\times\Omega} \partial_t u_1 v dxdt = -\int_{I\times\Omega} u_1 \partial_t v dxdt$, hence $\int_{I\times\Omega} \partial_t u_1 v - u_2 \cdot \nabla_x v dxdt = -\int_{I\times\Omega} u \cdot \nabla v dxdt$, so that $\operatorname{div} u = h + b \cdot A^{-1}u_2 - cu_1$ as a weak derivative. Hence, for $u = (u_1, u_2)$ we find that

$$\begin{aligned} Gu &= (u_2 + A\nabla_x u_1, \operatorname{div} u - b \cdot A^{-1}u_2 + cu_1, u_1(0,\cdot)) \\ &= (q - A\nabla_x u_1 + A\nabla_x u_1, h + b \cdot A^{-1}u_2 - cu_1 - b \cdot A^{-1}u_2 + cu_1, u_0). \end{aligned}$$

This shows surjectivity of G. □

Because of the theorem above it suffices to solve $G(u) = G(u_1, u_2) = (0, f, u_0) = g$ to solve the parabolic PDE of Example 3.4. This is equivalent to the least-squares formulation:

$$u = \operatorname{argmin}_{w\in\mathcal{U}_0} ||Gw - g||^2$$

which is the same as the problem of searching $u \in \mathcal{U}_0$ such that

$$\langle Gu, Gw\rangle = \langle g, Gw\rangle \quad \text{for all } w \in \mathcal{U}_0. \tag{3.2.9}$$

The form $(u, v) \mapsto \langle Gu, Gv\rangle$ used here is a bounded and coercive bilinear form, and the functional $u \mapsto \langle g, Gu\rangle$ is bounded on $\mathcal{U}_0$. This means that the reduced basis method can be applied to the equation above. In addition, it is easily seen that the bilinear form and the functional used here are parameter separable whenever A, b, c and f are parameter separable.

Solving this equation numerically using finite elements is also possible in a natural way. Notice, that the inner product that needs to be evaluated is from an L^2 space, hence we can just use the integral definition of the inner-product. Furthermore, the test-space $\mathcal{U}_0$ contains the very the natural piece-wise polynomials. In [3] a convergent finite element method using finite elements was shown for the variational formulation above. The mesh $\mathcal{T}_h$ of the domain $I \times \Omega$ used in that article consists of squares. The finite element space used there is defined as $S_0^1 \times S^1$, where S^1 consist of continuous affine functions on $I \times \Omega$ and S_0^1 consist of continuous affine functions which are zero on $I \times \partial\Omega$.

Remark 3.9. We can estimate the coercivity constant of the bilinear form above. This can be useful for computing the error estimator in Proposition 2.4. Here we assume that $A = Id$,

$c = 0$ and $d = 1$.

$$\begin{aligned}
||u||_{\mathcal{U}}^2 &\leq ||u_1||_{L^2(I;H^1(\Omega))}^2 + ||u_2||_{L^2(I;L^2(\Omega))}^2 + ||\operatorname{div} u||_{L^2(I\times\Omega)}^2 \\
&\leq ||u_1||_{L^2(I;H^1(\Omega))}^2 + [||u_2 + \nabla_x u_1||_{L^2(I\times\Omega)^d} + ||u_1||_{L^2(I;H^1(\Omega))}]^2 \\
&+ [||\operatorname{div} u - b\cdot u_2||_{L^2(I\times\Omega)} + ||b||_\infty ||u_2||_{L^2(I\times\Omega)^d}]^2 \\
&\leq (3 + 4||b||_\infty^2)||u_1||_{L^2(I;H^1(\Omega))}^2 + 2||u_2 + \nabla_x u_1||_{L^2(I\times\Omega)^d}^2 \\
&+ (2 + 4||b||_\infty^2)||\operatorname{div} u - b\cdot u_2||_{L^2(I\times\Omega)}^2 \\
&\leq (3 + 4||b||_\infty^2)(1 + ||b||_\infty)^2(||B^{-1}||^2)||Gu||^2 + 2||u_2 + \nabla_x u_1||_{L^2(I\times\Omega)^d}^2 \\
&+ (2 + 4||b||_\infty^2)||\operatorname{div} u - b\cdot u_2||_{L^2(I\times\Omega)}^2 \\
&\leq \left((3 + 4||b||_\infty^2)(1 + ||b||_\infty)^2(||B^{-1}||^2) + (2 + 4||b||_\infty^2)\right)||Gu||^2.
\end{aligned}$$

By careful investigation of the proof in [1] one can see that
$||B^{-1}|| \leq e^{2||b||_\infty^2 T} 2(1 + ||b||_\infty)^2 \max(\sqrt{1 + ||b||_\infty^2 \xi^4}, \sqrt{2})\sqrt{8 + M_e^2}$, where $I = [0, T]$ and ξ is the Poincaré constant for Ω.

Remark 3.10. Using the strategy in Example 1.24 we can prove that the solution-operator $\mu \mapsto u^\mu$ of the following parabolic equation is holomorphic on U if b is holomorphic on U.

$$\begin{cases}
\partial_t u^\mu - \operatorname{div} A\nabla_x u^\mu + b(\mu)\cdot\nabla_x u^\mu + cu^\mu &= f \quad \text{on } I\times\Omega, \\
u^\mu &= 0 \quad \text{on } I\times\partial\Omega, \\
u^\mu(0,\cdot) &= 0 \quad \text{on } \Omega.
\end{cases}$$

We first need the variational formulation of this PDE. Namely, we have that u^μ satisfies the following for all $v \in \mathcal{U}_0$:

$$\begin{aligned}
\int_{I\times\Omega} (u_2^\mu + A\nabla_x u_1^\mu)(v_2 + A\nabla_x v_2) + (\operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu)(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1)dx \\
+ \int_\Omega u_1^\mu(0,\cdot)v_1(0,\cdot)dx = \int_{I\times\Omega} f(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1)dx.
\end{aligned}$$

Because b is bounded, the coercivity constant is bounded away from zero. Lax Milgram now states that $||u^\mu||_{\mathcal{U}}$ is bounded for all $\mu \in U$. To find a candidate derivative we differentiate both sides of the equation and obtain

$$\begin{aligned}
&\int_{I\times\Omega} (\partial_\mu u_2^\mu + A\nabla_x \partial_\mu u_1^\mu)(v_2 + A\nabla_x v_2) \\
&+ (\operatorname{div}\partial_\mu u^\mu - b(\mu)A^{-1}\partial_\mu u_2^\mu + c\partial_\mu u_1^\mu)(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1)dx \\
&+ \int_\Omega \partial_\mu u_1^\mu(0,\cdot)v_1(0,\cdot)dx \\
&- \int_{I\times\Omega} f(-b'(\mu)A^{-1}v_2) \\
&- \int_{I\times\Omega} (-b'(\mu)A^{-1}u_2^\mu)(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1) \\
&- \int_{I\times\Omega} (\operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu)(-b'(\mu)A^{-1}v_2)dx = 0
\end{aligned}$$

By rearranging the terms we see that $\partial_\mu u^\mu$ satisfies a parabolic equation. Hence, $\partial_\mu u^\mu$ exists. We now want to prove that $\partial_\mu u^\mu$ is in fact the derivative of u^μ. First we prove continuity of u^μ. For this, we subtract the variational formulations of u_μ and u_z to obtain

$$\begin{aligned}
&\int_{I\times\Omega}(u_2^\mu + A\nabla_x u_1^\mu - u_2^z - A\nabla_x u_1^z)(v_2 + A\nabla_x v_2)dx \\
&+ \int_{I\times\Omega}(\operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu - \operatorname{div} u^z + b(\mu)A^{-1}u_2^z - cu_1^z)(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1)dx \\
&+ \int_\Omega (u_1^\mu(0,\cdot) - u_1^z(0,\cdot))v_1(0,\cdot)dx \\
&= \int_{I\times\Omega} f(-b(\mu)A^{-1}v_2 + b(z)A^{-1}v_2)dx \\
&+ \int_{I\times\Omega}(-b(z)A^{-1}u_2^z + b(\mu)A^{-1}u_2^z)(\operatorname{div} v - b(z)A^{-1}v_2 + cv_1)dx \\
&- \int_{I\times\Omega}(\operatorname{div} u^z - b(\mu)A^{-1}u_2^z + cu_1^z)(-b(\mu)A^{-1}v_2 + b(z)A^{-1}v_2)dx.
\end{aligned}$$

If we substitute $v = u^\mu - u^z$ we find that the left-hand side is exactly $||Gv||^2$ which is larger than $||v||_{\mathcal{U}}^2$. The right-hand side is smaller than $C|\mu - z| \cdot ||v||_{\mathcal{U}}$. Hence, $||u_\mu - u_z||_{\mathcal{U}} \to 0$.

We can now prove differentiability. For this we substract the variational formulation of μ and z, then divide by $(z-\mu)$ and subtract the formulation of the candidate derivative to obtain

$$\begin{aligned}
&\int_{I\times\Omega}(\partial_\mu u_2^\mu + A\nabla_x \partial_\mu u_1^\mu - \frac{u_2^\mu + A\nabla_x u_1^\mu - u_2^z - A\nabla_x u_1^z}{z-\mu})(v_2 + A\nabla_x v_2) \\
&+ \int_{I\times\Omega}(\operatorname{div}\partial_\mu u^\mu - b(\mu)A^{-1}\partial_\mu u_2^\mu + c\partial_\mu u_1^\mu \\
&- \frac{\operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu - \operatorname{div} u^z + b(z)A^{-1}u_2^z - cu_1^z}{z-\mu})(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1)dx \\
&+ \int_\Omega (\partial_\mu u_1^\mu(0,\cdot) - \frac{u_1^\mu(0,\cdot) - u_1^z(0,\cdot)}{z-\mu})\cdot v_1(0,\cdot)dx \\
&= \int_{I\times\Omega} f(x)(-b'(\mu)A^{-1}v_2 - \frac{-b(\mu)A^{-1}v_2 + b(z)A^{-1}v_2}{z-\mu})dx \\
&+ \int_{I\times\Omega}(-b'(\mu)A^{-1}u_2^\mu - \frac{-b(z)A^{-1}u_2^\mu + b(\mu)A^{-1}u_2^\mu}{z-\mu})(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1)dx \\
&- \int_{I\times\Omega}(\frac{-b(z)A^{-1}u_2^z + b(\mu)A^{-1}u_2^z}{z-\mu} - \frac{-b(z)A^{-1}u_2^\mu + b(\mu)A^{-1}u_2^\mu}{z-\mu})(\operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1)dx \\
&- \int_{I\times\Omega}(\frac{-b(z)A^{-1}u_2^z + b(\mu)A^{-1}u_2^z}{z-\mu})(-b(z)A^{-1}v_2 + b(\mu)A^{-1}v_2)dx \\
&- \int_{I\times\Omega}(\operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu)(-b'(\mu)A^{-1}v_2 + \frac{-b(\mu)A^{-1}v_2 + b(z)A^{-1}v_2}{z-\mu})dx \\
&- \int_{I\times\Omega}(\operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu - \operatorname{div} u^z + b(\mu)A^{-1}u_2^z - cu_1^z)(\frac{-b(\mu)A^{-1}v_2 + b(z)A^{-1}v_2}{z-\mu})dx
\end{aligned}$$

If we now substitute $v = \partial_\mu u - \frac{u^z - u^\mu}{z-\mu}$ then we find that the left-hand side is equal to $||Gv||^2$

which is larger than $||v||_{\mathcal{U}}^2$. The right-hand side is smaller than

$$
\begin{aligned}
&||f||_{L^2} \cdot |b'(\mu) - \frac{b(z)-b(\mu)}{z-\mu}|||A^{-1}v_2||_{L^2} \\
&+ ||A^{-1}u_2^\mu||_{L^2}|b'(\mu) - \frac{b(z)-b(\mu)}{z-\mu}||| \operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1||_{L^2} \\
&+ |\frac{b(z)-b(\mu)}{z-\mu}|||A^{-1}(u_2^z - u_2^\mu)||_{L^2}|| \operatorname{div} v - b(\mu)A^{-1}v_2 + cv_1||_{L^2} \\
&+ |\frac{b(z)-b(\mu)}{z-\mu}|||A^{-1}u_2^z||_{L^2}|| - b(z)A^{-1}v_2 + b(\mu)A^{-1}v_2||_{L^2} \\
&+ || \operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu||_{L^2} \cdot |b'(\mu) - \frac{b(z)-b(\mu)}{z-\mu}| \cdot ||A^{-1}v_2||_{L^2} \\
&+ || \operatorname{div} u^\mu - b(\mu)A^{-1}u_2^\mu + cu_1^\mu - \operatorname{div} u^z + b(\mu)A^{-1}u_2^z - cu_1^z||_{L^2} \cdot |\frac{b(z)-b(\mu)}{z-\mu}| \cdot ||A^{-1}v_2||_{L^2}
\end{aligned}
$$

This can be bounded by $C|z-\mu| \cdot ||v||_{\mathcal{U}}$ for some $C > 0$. Hence, after letting $z \to \mu$ we find that $\partial_\mu u - \frac{u^z - u^\mu}{z-\mu} \to 0$.

4 Numerical Experiments

We have compared the method of interpolation and the RB-method for solving a parabolic equation based on two examples from [22].

Example 4.1. This first equation reads, find $u(\mu)$, $\mu \in [-1,1]^2$ such that

$$\begin{cases} \partial_t u^\mu - \partial_{xx} u^\mu + b(t, v_0, \tau) \cdot \partial_x u^\mu &= f \quad \text{on } I \times \Omega, \\ u^\mu &= 0 \quad \text{on } I \times \partial\Omega, \\ u^\mu(0, \cdot) &= 0 \quad \text{on } \Omega. \end{cases} \tag{4.0.1}$$

Here $\Omega = [0,1]$, $I = [0,1]$ and $b(t;\mu) = b(t; v_0, \tau) = v_0 + v_\infty \frac{b_0^2 t}{b_0^2 + \tau}$, with $v_\infty = -3$, $b_0 = 2$, $v_0 = (-3.28+3.94)/2\mu_1 + (-3.94-3.28)/2$ and $\tau = (0.40-0.13)/2\mu_2 + (0.13+0.40)/2$. Furthermore, if we choose $f(t,x) = \sin(2\pi x) + 4\pi^2 y \sin(2\pi x) + (-3.445 + v_\infty \frac{b0^2 t}{b_0^2 + 0.3325}) 2\pi t \cos(2\pi x)$, then $u(0.5, 0.5) = t \sin(2\pi x)$.

The solution operator $u(v_0, \tau)$ is holomorphic on the disk D_ρ^2 with $\rho = 24$ because of Remark 3.10. In fact, it follows from the same remark that the solution operator of any finite element method is holomorphic on the same disk D_ρ^2. Additionally, the bilinear form $\langle G\cdot, G\cdot\rangle_{\mathcal{U}}$ and the linear form $\langle g, G\cdot\rangle_{\mathcal{U}}$ are parameter separable. Hence, both methods can be applied, and are expected to do well in approximating the solution map of this equation.

Example 4.2. The second equation reads, find $u(\mu)$, $\mu \in [-1,1]^3$ such that

$$\begin{cases} \partial_t u^\mu - \partial_{xx} u^\mu + b(t, v_0, \tau) \cdot \partial_x u^\mu &= f \quad \text{on } I \times \Omega, \\ u^\mu &= 0 \quad \text{on } I \times \partial\Omega, \\ u^\mu(0, \cdot) &= 0 \quad \text{on } \Omega. \end{cases} \tag{4.0.2}$$

Here $\Omega = [0,1]$, $I = [0,1]$ and $b(t;\mu) = b(t; b_0, \alpha, \beta) = -b_0^2(\alpha + \beta x)$, with $b_0 = 1.25 + 0.75\mu_1$, $\alpha = 4\mu_2$ and $\beta = 8\mu_3$. Furthermore, if we choose $f(t,x) = \sin(2\pi x) + 4\pi^2 y \sin(2\pi x) + b(0.5, 0.5, 0.5) 2\pi t \cos(2\pi x)$, then $u(0.5, 0.5, 0.5) = t \sin(2\pi x)$.

Again, the solution operator $u(\mu)$ and the finite element method are holomorphic because of Remark 3.10. Additionally, the bilinear form $\langle G\cdot, G\cdot\rangle_{\mathcal{U}}$ and the linear form $\langle g, G\cdot\rangle_{\mathcal{U}}$ are parameter separable. Again, both methods can be applied, and be expected to do well in approximating the solution map of this equation.

All the numerical experiments were performed using NGSolve [21]. From here, we will denote $\overline{\mu}$ as $(0.5, 0.5)$ or $(0.5, 0.5, 0.5)$.

4.1 Results

In the offline phase of both methods, we used the full-variational formulation for the parabolic equations. Namely, find $u(\mu) = (u_1, u_2) \in \mathcal{U}$ such that

$$\langle Gu, Gv\rangle_{\mathcal{U}} = \langle g, Gv\rangle_{\mathcal{U}} \text{ for all } v \in \mathcal{U}. \tag{4.1.1}$$

Here $g = (0, f, 0)$. The above equation is the same as

$$\int_{I\times\Omega} (u_2 + \partial x u_1)(v_2 + \partial_x v_1) + (\operatorname{div} u - b u_2)(\operatorname{div} v - b v_2) dx dt$$
$$+ \int_\Omega u_1(0,\cdot) v_1(0,\cdot) dx = \int_{I\times\Omega} f \cdot (\operatorname{div} v - b v_2) dx dt \text{ for all } v \in \mathcal{U}.$$

This equation was solved using a finite element method. For this we used some triangulation of the full domain $I \times \Omega$ with some maximal width h. The test space we used is defined as $V_H := X_1 \times X_2 \subset \mathcal{U}_0$. Here, X_1 is a Lagrangian element space of order 2 on $I \times \Omega$ with elements that are zero on $I \times \partial\Omega$ and X_2 is a Lagrangian element space of order 2 on $I \times \Omega$. Although it is not a proven fact, we assumed that this finite element method converges for $h \to 0$. The table below indicates that this is true for the parabolic equation in Example 4.1 with $\mu = \hat{\mu}$. This table also shows the strength of the $\mathcal{U}$-norm.

h	$\|\|u(\mu) - u_H(\mu)\|\|_{L^2(I\times\Omega)}$	$\|\|u(\mu) - u_H(\mu)\|\|_{H^1(I\times\Omega)}$	$\|\|u(\mu) - u_H(\mu)\|\|_{\mathcal{U}}$
0.5	0.224	3.942	11.842
0.1	3.750e-06	0.003	0.080
0.05	2.811e-08	0.00014	0.0118
0.01	4.741e-13	5.731e-08	0.000199
0.006	1.512e-14	4.847e-09	5.924e-05

Table 4.1: Error of finite element method for $\mu = (0.5, 0.5, 0.5)$

RB-method

For the RB-method we needed to find a parameter space S_N and the corresponding space of solutions X_N. For the problem in Example 4.2 the RB-method was tested in two ways. Namely, via the greedy algorithm and by using the Clenshaw-Curtis nodes. Both ways of choosing parameters lead to an RB-solution that provably converges to the finite element solution for $N \to \infty$. The problem in Example 4.1 was only solved using the greedy algorithm.

In the greedy algorithm we used the norm of the residual $||r||_{V_H'}$ as the error indicator. One could also use the error estimator $||r||_{V_H'}/\alpha_{LB}(\mu)$ from Proposition 2.4. However, the fluctuation of the coercivity constant of these problems does not cause any problems. For both problems, the training set S_{train} consisted of the tensor product of 20 evenly spread points in all directions.

For the problem in Example 4.1 we chose $h = 0.1$. For most $N \leq 30$ we estimated the maximal error of the RB-method with respect to the finite element method. This error was estimated by computing $\max_{\mu\in S_{train}} ||u_H(\mu) - u_N(\mu)||_{\mathcal{U}}$. The results are shown in Figure 4.1

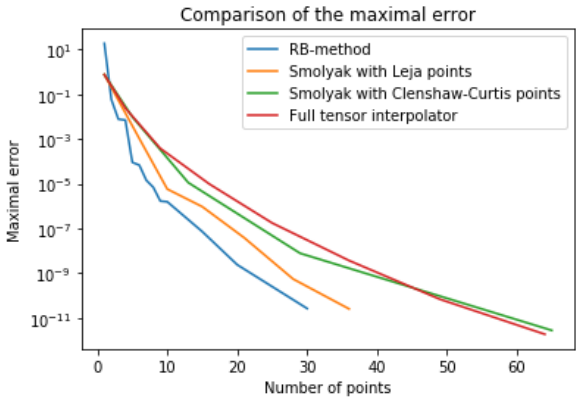
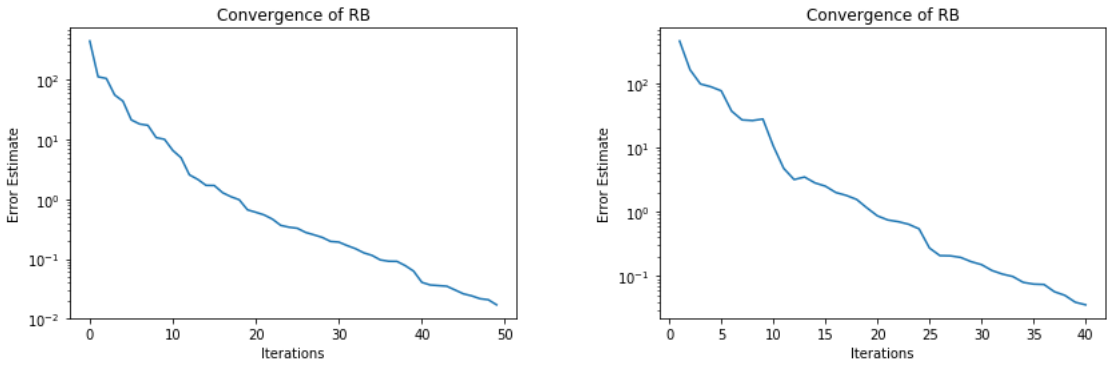


Figure 4.1: A comparison of the maximal errors for all methods considered in this thesis for Example 4.1, with $h = 0.1$.

For the problem in Example 4.2 the exact error in $\overline{\mu}$ after 40 iterations with $h = 0.01$ was equal to 0.0040 in the $\mathcal{U}$-norm and $1.78e-09$ in the $L^2(I \times \Omega)$-norm. In the online phase, for $N = 40$, the device used was able to compute solutions in 0.13 seconds on average. For comparison, the finite element method with $h = 0.04$, had an average runtime of 0.57 seconds and an error of 0.0075 in the $\mathcal{U}$-norm. The convergence of the greedy algorithm is plotted in Figure 4.2, where the error estimator $||r||_{V_H'}$ is shown on the y-axis.

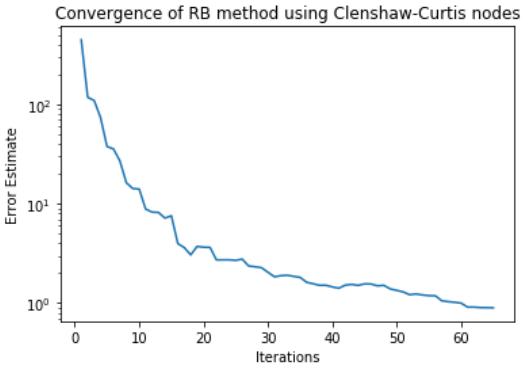
The performance of the reduced basis method using Chebyshev points is also shown in Figure 4.2. For $N = 69$, the error in the point $\overline{\mu}$ in the $\mathcal{U}$-norm of the reduced basis method with respect to the finite element solution was equal to 0.00089.



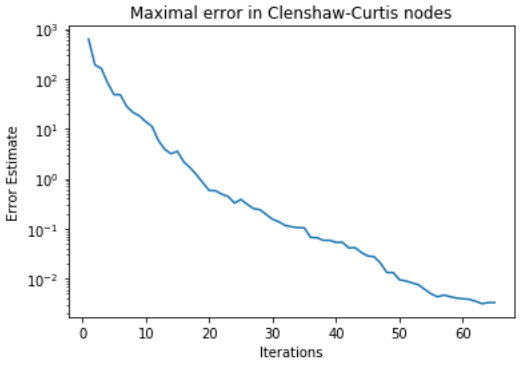
(a) $h = 0.1$.

(b) $h = 0.01$.

Figure 4.2: Performance of the reduced basis method for Example 4.2.



(a) Maximal value of the error estimator in S_{train}.



(b) Maximal value of the error estimator for all Clenshaw-Curtis nodes.

Figure 4.3: Performance of the reduced basis method for Example 4.2 when using Clenshaw Curtis nodes.

Interpolation

For the problem in Example 4.1 we tested both Smolyak algorithms and the full tensor interpolator. For each method we computed the maximal error with respect to the finite elements solution. The result is shown in Figure 4.1. For the problem in Example 4.2 we tested the Smolyak algorithm using Clenshaw-Curtis nodes and the full tensor interpolator. To test the performance of the Smolyak algorithm using Clenshaw-Curtis nodes we computed the error of $\mathcal{A}(q, J))u(\overline{\mu})$ with respect to the finite element solution for different values of q. The graph below shows this error on the y-axis and the number of interpolations points on the x-axis.

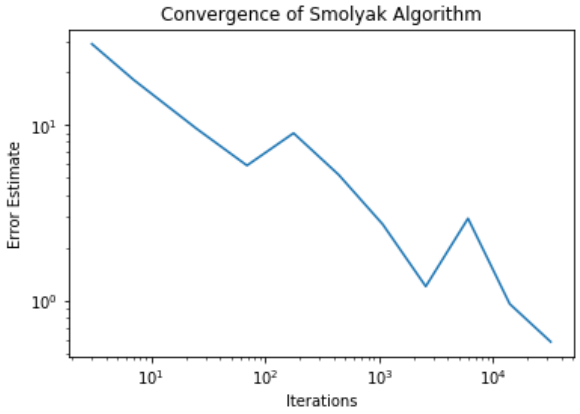


Figure 4.4: Performance of the Smolyak algorithm using Clenshaw-Curtis nodes, for Example 4.2, with $h = 0.1$.

For the full tensor interpolator we computed the maximum over all μ in S_{train} of $||I_{(N,N,N)}u(\mu) - u_H(\mu)||_{\mathcal{U}}$ for different values of N. The graph below shows this maximal error on the y-axis and the number of interpolations points on the x-axis.

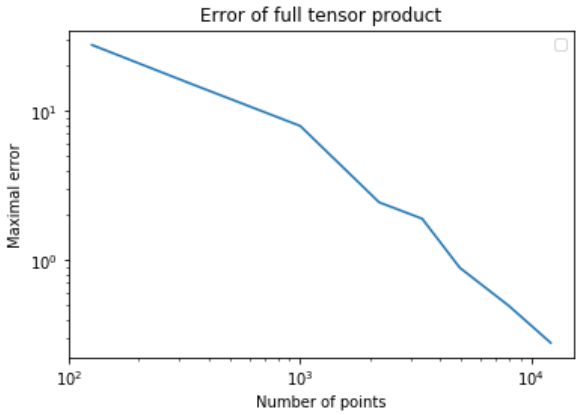


Figure 4.5: Maximal error of the full tensor interpolator, for Example 4.2, with $h = 0.1$.

5 Conclusion

In this thesis we investigated two methods for solving parametric parabolic equations using an offline/online decomposition. These methods are the reduced basis method and polynomial interpolation.

For both methods we managed to provide an error analysis for when either the solution operator or the solution of the finite element method is holomorphically dependent on the parameters. However, the error analysis for the reduced basis method was not fully completed for the higher dimensional case. In the error analysis polynomial interpolation it was remarkable that the full tensor product had a smaller upper bound of the error compared to the Smolyak algorithm using Clenshaw-Curtis points. The Smolyak algorithm using Leja points had the smallest upper bound of the error in most situations.

The numerical results suggest that the reduced basis method is superior to polynomial interpolation in the context of this thesis. This was visible in both two and three dimensions. In three dimensions we were not able to obtain reasonable results using interpolation (partially due to inadequate computational power of the device used). The reduced basis method did behave in an acceptable way. In fact, the goal of reducing the computing time for solving parabolic equations in the online phase was achieved.

There are still some things that need to be investigated. First of all, more numerical experiments are needed to understand the behaviour of the methods considered in this thesis. Furthermore, for the method of interpolation we saw that the Smolyak algorithm using Leja points had the smallest asymptotic upper bound of the error. However, the upper bounds of Lebesgue constants of the Leja points are still quite large. Investigation into the reduction of these upper bounds is needed. Additionally, the error bound for the Smolyak algorithm using Clenshaw-Curtis points was worse than expected beforehand. The numerical experiments suggest that the error bound for this interpolation method could perhaps be improved. Lastly, the error analysis for the reduced basis method still needs to be finished. Using the theory in this thesis it could be possible to show that the error of the reduced basis method decays exponentially when using the greedy algorithm.

Populaire Samenvatting (in dutch)

Deze scriptie beschrijft twee methodes om parabolische vergelijkingen op te lossen. De eerste methode is het makkelijkst te begrijpen, namelijk polynomiale interpolatie, de tweede methode is de gereduceerde basismethode. Het doel van beide methodes is dat we, met behulp van wat voorwerk in het 'offline' gedeelte, zeer snel oplossingen kunnen vinden van een parabolisch probleem gedurende het 'online' gedeelte voor alle mogelijke parameterwaarden. Dit kan heel handig zijn als de vergelijking te moeilijk is voor een computer op locatie; bijvoorbeeld op een schip. Deze methodes maken het mogelijk dat een supercomputer van te voren al wat werk verricht.

Een parabolische vergelijking is een vergelijking van de volgende vorm: zoek een functie $u\colon \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ zodat

$$\begin{cases} \partial_t u - \operatorname{div} A\nabla_x u + b \cdot \nabla_x u + cu &= f & \text{on } I \times \Omega, \\ u &= 0 & \text{on } I \times \partial\Omega, \\ u(0,\cdot) &= u_0 & \text{on } \Omega. \end{cases}$$

Hier is $I = [0,T]$ een tijdsinterval en $\Omega \subset \mathbb{R}^d$ een d-dimensionaal gebied. Bij een parameterische parabolische vergelijking hangt minstens één van de functies A, b, c of f van een parameter μ af. In deze scriptie hebben we ervoor gekozen dat μ in een n-dimensionale kubus leeft, dus $\mu \in [-1,1]^n$.

Zoals eerder gezegd, polynomiale interpolatie is het makkelijkst te begrijpen. Allereerst kunnen we opmerken dat een een N-dimensionaal polynoom bepaald wordt door $N+1$ functiewaarden. Bijvoorbeeld als $p(x) = ax + b$ dan ook $p(x) = (p(1) - p(0))x + p(0)$. Met dit gegeven kunnen we met behulp van $N+1$ oplossingen van een parabolische vergelijking een N-dimensionaal polynoom bepalen. Dit polynoom is dan alleen niet meer reëelwaardig, het beeld van zo'n polynoom bestaat uit functies die al dan niet het parabolisch probleem oplossen voor een zekere parameterwaarde. Een voorbeeld van zo'n polynoom is $p(\mu) = tx^2\mu + \cos(2\pi x)\sin(4\pi t)\mu^2 + te^{xt}$. Voor elke waarde van μ krijgen we een functie die afhangt van x en t terug. We hopen nu natuurlijk, dat voor alle parameterwaarden μ, dat $p(\mu)$ op de oplossing van de betreffende parameterische parabolische vergelijking lijkt in μ.

In deze scriptie hebben we gekeken hoe we deze interpolatie het beste kunnen uitvoeren. Door wat eisen te leggen op de afhankelijkheid van de vergelijking op de parameters is het ook mogelijk om de 'interpolatiefout' te schatten.

De tweede methode is een stuk complexer en heeft veel meer te maken met een bepaalde manier van kijken naar partiële differentiaalvergelijkingen (PDE). Over het algemeen is het mogelijk om een PDE te schrijven met behulp van een 'variationele formulering': zoek een functie $u\colon \mathbb{R}^d \to \mathbb{R}$ zodat

$$a(u,v) = f(v), \text{ voor alle functies } v.$$

Hier is a een bilineaire afbeelding en f een lineaire afbeelding. Een schrijfwijze als deze blijkt ook mogelijk te zijn voor een parabolische vergelijking. Dit is allerminst triviaal.

Deze variationele formulering is een belangrijk ingrediënt voor de gereduceerde basismethode. Het idee is als volgt: we kunnen een aantal parameterwaarden μ_i $i = 1, \ldots, N$ kiezen zodat de oplossingen $u(\mu_i)$ een lineaire ruimte X_N vormen. Deze lineaire ruimte bevat hopelijk genoeg informatie over de oplossingen in andere parameterwaarden. Als dat echt zo is dan kunnen we de oplossing in andere parameterwaarden benaderen door het volgende probleem op te lossen: zoek $u \in X_N$ zodat

$$a(u, v) = f(v), \text{ voor alle functies } v \in X_N.$$

We hebben dus de 'zoekruimte' beperkt, maar ook de 'testruimte', waar de v'tjes vandaan komen, beperkt. Op deze manier verkrijgen we een versimpelde parabolische vergelijking, die we kunnen omschrijven naar een matrix-vector vergelijking. Dit is dus op te lossen voor een computer.

Uit de numerieke experimenten en ook de wiskunde, volgt veelal dat de gereduceerde basismethode het wint van polynomiale interpolatie. Het is alleen lastiger om de gereduceerde basismethode te implementeren.

Bibliography

[1] Christoph Schwab, and Rob Stevenson. “Space-Time Adaptive Wavelet Methods for Parabolic Evolution Problems.” Mathematics of computation 78.267 (2009): 1293–1318.

[2] Gantner, Stevenson. “Further Results on a Space-Time FOSLS Formulation of Parabolic PDEs.” (2020)

[3] Führer, Karkulik. “Space-Time Least-Squares Finite Elements for Parabolic Equations.” (2019)

[4] Dautray, Robert. et al. Mathematical Analysis and Numerical Methods for Science and Technology / Vol. 5, Evolution Problems I / with the Collab. of Michel Artola, Michel Cessenat and Hélène Lanchon; Transl. from the French by Alan Craig; Transl. Ed.: Ian N. Sneddon. [2nd ed.]. Berlin, Springer-Verlag, 2000. Print.

[5] Rynne, Bryan., and M.A. Youngson. Linear Functional Analysis. Second edition. London: Springer London, 2008.

[6] Süli, Endre, and D. F. Mayers. An Introduction to Numerical Analysis. Cambridge; Cambridge University Press, 2003.

[7] Ibrahimoglu, Bayram. “Lebesgue Functions and Lebesgue Constants in Polynomial Interpolation.” Journal of inequalities and applications 2016.1 (2016): 1–15.

[8] Simon J. Smith. “Lebesgue constants in polynomial interpolation.” Annales Mathematicae et Informaticae 33 (2006) pp. 109–123

[9] Dzjadyk, Ivanov. “On Asymptotics and Estimates for the Uniform Norms of the Lagrange Interpolation Polynomials Corresponding to the Chebyshev Nodal Points.” Analysis mathematica (Budapest) 9.2 (1983): 85–97.

[10] Chkifa, Abdellah. “New Bounds on the Lebesgue Constants of Leja Sequences on the Unit Disc and Their Projections $\Re$-Leja Sequences.” (2015)

[11] R. A. Devore, G. G. Lorentz. “Constructive Approximation”, Grundlehren Math. Wiss. 303, Springer-Verlag, Berlin. 1993

[12] Brenner, Scott. The Mathematical Theory of Finite Element Methods. Vol. 15. New York, NY: Springer New York.

[13] Haasdonk, B.: Reduced Basis Methods for Parametrized PDEs – A Tutorial Introduction for Stationary and Instationary Problems. Chapter in P. Benner, A. Cohen, M. Ohlberger and K. Willcox (eds.): ”Model Reduction and Approximation: Theory and Algorithms”, SIAM, Philadelphia, 2017.

[14] Quarteroni, Alfio, Andrea Manzoni, and Federico Negri. Reduced Basis Methods for Partial Differential Equations: an Introduction . 1st ed. 2016. Cham, Switzerland: Springer, 2016.

[15] Binev, Cohen. "Convergence Rates for Greedy Algorithms in Reduced Basis Methods." SIAM journal on mathematical analysis 43.3 (2011): 1457–1472.

[16] S.A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Soviet Math. Dokl. 4 (1963) 240–243.

[17] Barthelmann, Novak. "High Dimensional Polynomial Interpolation on Sparse Grids." Advances in computational mathematics 12.4 (2000): 273–288.

[18] Moulay Abdellah Chkifa. Sparse polynomial methods in high dimension : application to parametric PDE. General Mathematics [math.GM]. Université Pierre et Marie Curie - Paris VI, 2014. English. ffNNT : 2014PA066218ff. fftel-01083620

[19] Nobile, Tempone. "A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data." SIAM journal on numerical analysis 46.5 (2008): 2309–2345.

[20] A.T. Patera, K. Urban "High performance computing on smartphones." Snapshots of modern mathematics from Oberwolfach No 6 2016

[21] NGSolve, https://ngsolve.org/

[22] Fokker-Planck. draft.