Approximation with matrix-valued kernels and highly effective error estimators for reduced basis approximations

Von der Fakultät Mathematik und Physik der Universität Stuttgart und dem Stuttgarter Zentrum für Simulationswissenschaften (SC SimTech) der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

vorgelegt von

Dominik Wittwar

aus Stuttgart

Hauptberichter:	Prof. Dr. Bernard Haasdonk
Mitberichter:	Prof. Dr. Stefano De Marchi
	Prof. Dr. Christian Rieger

Tag der mündlichen Prüfung: 18 November, 2021

Institut für Angewandte Analysis und Numerische Simulation,
Universität Stuttgart

2022

Abstract

This thesis can be summarized under the aspect of surrogate modelling for vector-valued functions and error quantification for those surrogate models. The thesis, in a broad sense, is split into two different parts. The first aspect deals with constructing surrogate models via matrix-valued kernels using both interpolation and regularization procedures. For this purpose, a new class of so called uncoupled separable matrix-valued kernels is introduced and heavy emphasis is placed on how suitable sample points for the construction of the surrogate can be chosen in such a way that quasi-optimal convergence rates can be achieved.

In the second part, the focus does not lie on the construction of the surrogate itself, but on how existing a-posteriori error estimation can be improved to result in highly efficient error bounds. This is done in the context of reduced basis methods, which similar to the kernel surrogates, construct surrogate models by using data acquired from samples of the desired target function.

Both parts are accompanied by numerical experiments which illustrate the effectiveness as well as verify the analytically derived properties of the presented methods.

Zusammenfassung

Diese Arbeit kann unter dem Gesichtspunkt der Ersatzmodellierung vektorwertige Funktionen und deren Fehlerquantifizierung zusammengefasst werden. Die Arbeit ist dabei in zwei Bereiche geteilt. Der erste Teil beschäftigt sich mit der Konstruktion von Ersatzmodellen mittels matrizenwertigen Kernen interpolatorischen und regularisierenden Approximationsmethoden. Für diesen Zweck präsentieren wir eine neue Klasse von so genannten ungekoppelten separablen matrizenwertigen Kernen und setzten einen besonderen Schwerpunkt darauf, auf welche Weise geeignete Auswertungspunkte gewählt werden können, sodass bei der Konstruktion des Ersatzmodells quasi-optimale Konvergenzraten erreicht werden.

Im zweiten Teil der Arbeit liegt das Hauptaugenmerk nicht mehr auf der Konstruktion der Ersatzmodelle, sondern darauf, wie bereits existierende a-posteriori Fehlerschranken verbessert werden können, um gute Effektivitäten zu erreichen. Dies geschieht im Zusammenhang mit reduzierten Basis Methoden, welche auf ähnliche Weise zu den Kernersatzmodellen, Ersatzmodelle aus Auswertungen einer Zielfunktion generieren.

In beiden Teilen wird dabei die Wirksamkeit der dargestellten Methoden durch numerische Experimente sowie durch analytisch hergeleitete Eigenschaften verdeutlicht.

Contents

1 Introduction

1.1 Motivation

Many of the processes in nature and engineering can be modelled via partial differential equations describing the underlying dynamics of a given problem. For most of these, no closed analytical solutions are known, and hence numerical solutions for these problems are sought. This can be done, by discretizing the possibly infinite dimensional space in which a true solution resides, into a finite dimensional one by posing certain restrictions. Most commonly, this is achieved by reformulating the problem into a weak formulation and then searching for weak solutions in smaller spaces. Common examples for this methodology are for finite element methods [16, 52] or finite volume schemes [56, 31]. Moreover, with the ever increasing complexity of problems in our modern society, these systems become increasingly large, as to improve the quality of the numerical solution. Furthermore, many of these problems are dependent on various input parameters, and hence many of these high dimensional problems have to be considered. While the computational performance which can be delivered by today's hardware seems to be ever increasing, it cannot keep up with today's demand. This might cause an issue with regards to the computational time when multiple solutions for varying parameters of the same problem are considered. This is referred to as a multi-query scenario and most widespread examples are parameter studies, parameter optimization and uncertainty quantification. Another aspect is the availability of high performing computational hardware and many systems might not even have the ability to execute the necessary computations due to a limit in storage ability. To circumvent these restrictions, the field of surrogate modelling is in ever increasing demand and is based on the following principle:

The high fidelity solutions described above can be summarized concisely via a function $f : \mathcal{P} \to \mathbb{R}^N$ mapping from some given parameter space $\mathcal{P}$ into a high dimensional vector space. This function is expensive to evaluate and hence surrogate models, i.e. alternate functions $\hat{f} : \mathcal{P} \to \mathbb{R}^N$ operating on the same parameter space are desired, which can be easily evaluated. A classical example for such surrogates are interpolation polynomials which derive a surrogate by posing interpolatory conditions at certain inputs. More sophisticated surrogate modelling approaches include kriging also known as

Gaussian process regression [77, 97], support vector machines [99, 23], kernel based approximations [108, 91], reduced order modelling [11, 7, 8, 9] and artificial neural networks [45, 51, 73]. The latter of which has been increasing in popularity and a wide variety of different approaches have been developed in the recent years. The commonality between all these methods is that they are data driven, i.e. the surrogate is constructed using samples $\{f(x)\,|\,x \in \mathcal{P}_{\text{sample}}\}$ of the so called target function for some sample set $\mathcal{P}_{\text{sample}} \subset \mathcal{P}$. The construction of the surrogate can then be divided into the three steps: 1. Selection of suitable sample parameters and subsequent evaluation of the target function, 2. Construction of the surrogate model and optimization of potential model parameters and 3. quantification of the quality of the approximation. In particular, step three is of utmost importance as the surrogate should reflect the target function f in a suitable fashion, i.e. we do not only want a function that is fast to evaluate, we also want sufficient accuracy and certification by quantifying this. Since it is a-priori unclear how many sample points should be selected in the first step in order to generate a sufficiently accurate surrogate, one might iterate over these three steps using the current surrogate to identify new parameters that should be included in the sampling process. This is commonly known as active learning [22].

The two methods we focus on in this thesis are kernel based approximation and reduced order modelling. For the kernel based surrogate, we mostly focus on step one and three, whereas for the reduced order modelling surrogate we focus on step three. The basic ideas behind these two methods can be roughly summarized as follows:

Kernel based surrogate

The kernel based surrogate takes the form

$$\hat{f}(x) = \sum_{i=1}^{n} K(x, x_i)\alpha_i,$$

where $\{x_1, \dots, x_n\}$ are the chosen sample parameters, $k(x, x_i) \in \mathbb{R}^{N \times N}$ is a matrix-valued function called “kernel” and $\{\alpha_1, \dots, \alpha_n\} \subset \mathbb{R}^N$ are coefficient vectors that depend on the evaluations of the target function at the sample parameters, i.e. $\alpha_i = \alpha_i(f(x_1), \dots, f(x_n))$. The kernel based surrogate does not care what processes and dynamics underlie the evaluation $\mathcal{P} \ni x \mapsto f(x) \in \mathbb{R}^N$ and thus only the evaluations themselves are required in the approximation process.

Reduced order modelling

Surrogates in the framework of reduced order modelling, on the other hand, often rely on the inner structure of the evaluation $x \mapsto f(x)$. Here, the approach is to use the information provided in the samples $\{f(x) \,|\, x \in \mathcal{P}_{\text{sample}}\}$ to derive a model of reduced order but of similar structure. This is commonly done by projecting the original system using a so called reduced basis $V \in \mathbb{R}^{N \times n}$ with $n \ll N$. The corresponding small system can now be solved fast resulting in a reduced solution $f_n(x) \in \mathbb{R}^n$ and the surrogate is computed by projecting this reduced solution back into the high dimensional space, i.e. $\hat{f}(x) = V f_n(x)$.

1.2 Structure of this thesis

The thesis is structured as follows.

In Chapter 2 we summarize basic concepts of matrix-valued kernels and extend properties known for scalar-valued kernels to matrix-valued kernels including approximation of functions via interpolation and a corresponding error analysis via the so called power function. We further introduce a new subclass of kernels, called *uncoupled separable kernels* and show how their structure can be used for efficient computation of the power function.

In Chapter 3 we study how data points should be selected in the approximation process by means of different variants of greedy algorithms. In particular, we focus on variants of the so called P–Greedy algorithm, which relies on the aforementioned power function and for which we were able to show that, under certain restrictions, quasi-optimal approximation rates can be achieved.

In Chapter 4 we shift our focus from interpolation based approximation to a regularization based one using matrix-valued weight functions. Similar to the previous chapter we introduce a novel greedy algorithm, the so called regularized P–Greedy algorithm, for the selection of suitable data points used in the construction. Likewise, we show that the point sets generated by the regularized P–Greedy algorithm result in quasi-optimal approximation rates.

In Chapter 5 we shift gears and switch from kernel based approximation to the topic of reduced order modelling. In this context we introduce a novel error estimation procedure based on the introduction of an auxiliary linear problem. Solving this additional problem allows us to reach highly effective error estimators while only slightly increasing the computational overhead. The effectivity of this error estimation procedure is affirmed both analytically as well as numerically for a variety of different problems.

1.3 Publications

During the course of my PhD studies, I was involved in the publication of following articles

1 D. Wittwar, G. Santin, and B. Haasdonk. Interpolation with uncoupled separable matrix-valued kernels. *Dolomites Res. Notes Approx.*, 11:23–29, 2018

2 D. Wittwar and B. Haasdonk. Greedy algorithms for matrix-valued kernels. In F. A. Radu, K. Kumar, I. Berre, J. M. Nordbotten, and I. S. Pop, editors, *Numerical Mathematics and Advanced Applications ENUMATH 2017*, pages 113–121, Cham, 2019. Springer International Publishing

3 D. Wittwar and B. Haasdonk. Convergence rates for matrix P-Greedy variants. In Cornelis Vuik Fred J. Vermolen, editor, *Numerical Mathematics and Advanced Applications ENUMATH 2019*, pages 1195–1203, Cham, 2021. Springer International Publishing

4 M. Köppel, F. Franzelin, I. Kröker, S. Oladyshkin, G. Santin, D. Wittwar, A. Barth, B. Haasdonk, W. Nowak, D. Pflüger, and C. Rohde. Comparison of data-driven uncertainty quantification methods for a carbon dioxide storage benchmark scenario. *Computational Geosciences*, 23(2):339–354, Apr 2019

5 A. Schmidt, D. Wittwar, and B. Haasdonk. Rigorous and effective a-posteriori error bounds for nonlinear problems – Application to RB methods. *Advances in Computational Mathematics*, 46(32), 2020

6 G. Santin B. Haasdonk, B. Hamzi and D. Wittwar. Greedy kernel methods for center manifold approximation. In J.Peiro P. E. Vincent S. J. Sherwin, D. Moxey and C. Schwab, editors, *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2018*, pages 95–106. Springer International Publishing, 2020

7 B. Haasdonk, B. Hamzi, G. Santin, and D. Wittwar. Kernel methods for center manifold approximation and a weak data-based version of the center manifold theorem. *Physica D: Nonlinear Phenomena*, 427:133007, 2021

Article 1. is the basis for Section 2.2 and Section 2.4–2.5 in Chapter 2. Articles 2.–3. provide the basis for Chapter 3, where parts of article 4. were used in Section 3.5. Finally, article 5. is the basis of Chapter 5. Articles 6. and 7. are not included in this thesis. However, they represent first applications of the surrogate scheme presented in Chapter 4.

2 Matrix-Valued Kernels

2.1 Motivation

Kernel methods are useful tools for dealing with a wide variety of different tasks ranging from machine learning e.g. via Support Vector Machines (SVMs) ([13, 86, 100]), function approximation from scattered data ([33, 62]) and many more. Especially the approximation aspect can be employed for generating surrogate models to speed up expensive function evaluation, see [112]. In cases where the given output data or the desired target function is vector-valued, simple approaches which build individual models for each function component can still be very costly, if the output is high dimensional and the component models rely on independent data sets such that the union of those results in overly large sets. Additionally, approximating a vector-valued function componentwise with identical ansatz spaces might be the wrong choice, e.g. in case of different frequencies. We thus propose the use of matrix-valued kernels which lead to surrogates that can deal with correlations between function components, respective structural properties of the target function, and therefore provide a more suitable model. For divergence-free kernels, matrix-valued kernel approximations have already been successfully applied, see e.g. ([32, 59, 69, 35]).

2.2 Basic Definitions and Properties

Definition 2.2.1 (Matrix-valued kernel)**.** Let Ω be a non empty set. A bivariate function $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$, $m \in \mathbb{N}$, is called a matrix-valued kernel if

$$K(x,y) = K(y,x)^T \qquad \text{for all } x, y \in \Omega \tag{2.1}$$

Definition 2.2.2 (Reproducing kernel Hilbert space (RKHS))**.** Let $\mathcal{H}$ denote a Hilbert space of $\mathbb{R}^m$-valued functions over a domain Ω with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced norm $\|\cdot\|_{\mathcal{H}}$. We call $\mathcal{H}$ a reproducing kernel Hilbert space, if for all $x \in \Omega$ and $\alpha \in \mathbb{R}^m$

the directional point evaluation functional $\delta_x^\alpha : \mathcal{H} \to \mathbb{R}$ defined by

$$\delta_x^\alpha(f) := f(x)^T \alpha, \qquad \text{for all } f \in \mathcal{H} \tag{2.2}$$

is bounded, i.e.

$$\|\delta_x^\alpha\|_{\mathcal{H}'} := \sup_{f \in \mathcal{H}\setminus\{0\}} \frac{\delta_x^\alpha(f)}{\|f\|_{\mathcal{H}}} < \infty.$$

We can immediately conclude that the property of being a reproducing kernel Hilbert space (RKHS) is inherited by any closed subspace of $\mathcal{H}$:

Corollary 2.2.3 (Closed subspaces are RKHS)**.** *Let $\mathcal{N} \subset \mathcal{H}$ be a closed subspace. Then $\mathcal{N}$ is a RKHS*

Proof. Since $\mathcal{N}$ is a closed subspace it is itself a Hilbert space, whose inner product and norm stem from the restriction of the inner product and norm onto $\mathcal{N}$, respectively. Therefore, we have for any directional point evaluation functional

$$\|\delta_x^\alpha\|_{\mathcal{N}'} = \sup_{f \in \mathcal{N}\setminus\{0\}} \frac{\delta_x^\alpha(f)}{\|f\|_{\mathcal{N}}} \leq \sup_{f \in \mathcal{H}\setminus\{0\}} \frac{\delta_x^\alpha(f)}{\|f\|_{\mathcal{H}}} < \infty.$$

□

Similar to the scalar-valued case which was first presented by Aronszajin in [4], there exists a one-to-one correspondence between RKHS of vector-valued functions and positive definite matrix-valued kernels. A necessary concept for this is the notion of positive definiteness which is a straightforward extension from the scalar-valued case and is given as follows:

Definition 2.2.4 (Definiteness)**.** Let Ω be non empty and $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ be a matrix-valued kernel. For any finite set $X := \{x_1, \ldots, x_n\} \subset \Omega$, $n \in N$, we define the Gramian matrix $\boldsymbol{K} \in \mathbb{R}^{mn \times mn}$ as the block matrix given by

$$\boldsymbol{K} := K(X, X) := (K(x_i, x_j))_{i,j=1}^n = \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{bmatrix}. \tag{2.3}$$

The kernel K is denoted as positive definite (p.d.), if for all $n \in \mathbb{N}$ and arbitrary $X = \{x_1, \ldots, x_N\} \subset \Omega$, the Gramian matrix $\boldsymbol{K}$ is positive semi-definite, i.e. it holds

$$\alpha^T \boldsymbol{K} \alpha \geq 0 \qquad \forall \, \alpha \in \mathbb{R}^{mn}. \tag{2.4}$$

The kernel is called strictly positive definite (s.p.d.), if for all $n \in \mathbb{N}$ and pairwise distinct $X = \{x_1, \dots, x_n\} \subset \Omega$ the Gramian matrix is positive definite, i.e. we have

$$\alpha^T \boldsymbol{K} \alpha > 0 \qquad \forall\, \alpha \in \mathbb{R}^{mn} \setminus \{0\}. \tag{2.5}$$

Definition 2.2.5. Partial ordering on the set of symmetric positive (semi-) definite matrices Let $A, B \in \mathbb{R}^{m \times m}$ be two symmetric matrices. We write

$$A \preceq B \qquad \text{or equivalently} \qquad B \succeq A$$

if the difference $B - A$ is positive semi-definite. In particular, we might denote positive semi-definite matrices simply via $A \succeq 0$. In a similar fashion, we write

$$A \prec B \qquad \text{or equivalently} \qquad B \succ A$$

if the difference $B - A$ is positive definite. Again, we might denote positive definite matrices via $A \succ 0$.

As we have mentioned before, each RKHS corresponds to a unique p.d. matrix-valued kernel and vice versa. This is a well-known result for scalar-valued kernels and extensions to the matrix-valued, and even operator valued, case have been made, c.f. [53]. Nonetheless, we summarize this correspondence in the following theorem including a proof as it proves insightful for the structure of RKHS

Theorem 2.2.6 (One-to-one correspondence)**.** *Let $\mathcal{H}$ be an RKHS consisting of $\mathbb{R}^m$ valued functions over a set Ω. Then there exists a unique positive definite matrix-valued kernel $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$, such that for all $x \in \Omega$ and $f \in \mathcal{H}$*

$$K(\cdot, x)\alpha \in \mathcal{H} \qquad \text{and} \qquad \langle f, K(\cdot, \alpha) \rangle_{\mathcal{H}} = f(x)^T \alpha. \tag{2.6}$$

Conversely, if $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ is a p.d. matrix-valued kernel, then there exists a unique Hilbert space $\mathcal{H}$ consisting of function $f : \Omega \to \mathbb{R}^m$ such that (2.6) *holds.*

Proof. We first prove the direct implication. To this end, we assume that $\mathcal{H}$ is a RKHS. By Definition 2.2.2 this means that the directional point evaluation functionals δ_x^α as defined in (2.2) are bounded linear functionals. Hence, each of these functionals has a Riesz representation in $\mathcal{H}$ which we shall denote as K_x^α. Since the directional point evaluation functionals are by definition linear with respect to their direction α, we know that δ_x^α is uniquely determined by $\delta_x^{e_i}$, where $e_i \in \mathbb{R}^m$, $i = 1, \dots, m$, denote the standard basis of

$\mathbb{R}^m$. It follows that the Riesz representation can be written via

$$K_x^\alpha = \begin{bmatrix} K_x^{e_1} & \cdots & K_x^{e_m} \end{bmatrix} \alpha =: K(\cdot, x)\alpha \tag{2.7}$$

where the matrix-valued function $K(\cdot, x) : \Omega \to \mathbb{R}^{m \times m}$ is defined by

$$K(\cdot, x) = \begin{bmatrix} K_x^{e_1} & \cdots & K_x^{e_m} \end{bmatrix}.$$

Accordingly, we can reinterpret K as a function over $\Omega \times \Omega$. It is easy to see that by construction of K the conditions in (2.6) are met. Furthermore for any $x, y \in \Omega$ and any distinct pair of standard basis vectors $e_i, e_j \in \mathbb{R}^m$ we have

$$e_i^T K(x, y) e_j = \langle K(\cdot, x)e_i, K(\cdot, y)e_j \rangle_{\mathcal{H}} = \langle K(\cdot, y)e_j, K(\cdot, x)e_i \rangle_{\mathcal{H}} = e_j^T K(y, x) e_i$$

and consequently $K(x, y) = K(y, x)^T$. Furthermore, for any set $X = \{x_1, \cdots, x_n\} \subset \Omega$ the matrix $\boldsymbol{K}$, as defined in (2.3), is the Gramian matrix for the elements $K(\cdot, x_1)e_1$, ..., $K(\cdot, x_n)e_m$ and therefore it is positive semi-definite. In total we have found a matrix-valued p.d. kernel, which satisfies (2.6). If $\hat{K}$ is a second kernel satisfying (2.6) then it holds by definition for arbitrary $x \in \Omega$ and all $\alpha \in \mathbb{R}^m$

$$\alpha^T \hat{K}(x, x)\alpha = \langle \hat{K}(\cdot, x)\alpha, K(\cdot, x)\alpha \rangle_{\mathcal{H}} = \langle K(\cdot, x)\alpha, \hat{K}(\cdot, x)\alpha \rangle_{\mathcal{H}} = \alpha^T K(x, x)\alpha$$

an thus $K = \hat{K}$, i.e. the kernel is unique. For the converse claim, let us assume that K is a p.d. matrix-valued kernel. Let $\mathcal{H}_0$ denote the space of functions spanned by

$$\mathcal{H}_0 := \operatorname{span}\{K(\cdot, x)\alpha \,|\, x \in \Omega,\ \alpha \in \mathbb{R}^m\}. \tag{2.8}$$

For any two functions $f, g \in \mathcal{H}_0$ given by

$$f = \sum_{i=1}^{p} K(\cdot, x_i)\alpha_i \qquad \text{and} \qquad g = \sum_{j=1}^{q} K(\cdot, y_j)\beta_j$$

we can define an inner product on $\mathcal{H}_0$ by

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^{p} \sum_{j=1}^{q} \alpha_i^T K(x_i, y_j)\beta_j,$$

where the positive definiteness of K is used to ensure the positive definiteness of the above inner product. With this, $\mathcal{H}_0$ becomes a pre-Hilbert space and the inner product induces a norm on $\mathcal{H}_0$ via

$$\|f\|_{\mathcal{H}_0}^2 = \langle f, f \rangle_{\mathcal{H}_0} .$$

Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy-Sequence in $\mathcal{H}_0$. Then we have for any $n, m \in \mathbb{N}$, $x \in \Omega$ and $\alpha \in \mathbb{R}^m$

$$\begin{aligned}|f_n(x)^T\alpha - f_m(x)^T\alpha| &= |\langle f_n, K(\cdot, x)\alpha\rangle - \langle f_m, K(\cdot, x)\alpha\rangle| \\ &\leq \|f_n - f_m\|_{\mathcal{H}_0} \|K(\cdot, x)\alpha\|_{\mathcal{H}_0} .\end{aligned}$$

Hence $(f_n(x)^T\alpha)_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$ for any $x \in \Omega$, $\alpha \in \mathbb{R}^m$ and $f(x) \coloneqq \lim\limits_{n\to\infty} f_n(x)$ is a well-defined function $f : \Omega \to \mathbb{R}^m$. We now denote

$$\mathcal{H} \coloneqq \{f : \Omega \to \mathbb{R}^m \,|\, f \text{ is the pointwise limit of some Cauchy sequence } (f_n)_{n\in\mathbb{N}} \subset \mathcal{H}_0\} .$$

Then $\mathcal{H}$ is a linear space and $\mathcal{H}_0 \subset \mathcal{H}$ since each constant sequence is a Cauchy sequence. We can now equip $\mathcal{H}$ with an inner product via

$$\langle f, g\rangle \coloneqq \lim_{n\to\infty} \lim_{m\to\infty} \langle f_n, g_m\rangle_{\mathcal{H}_0},$$

where $(f_n)_{n\in\mathbb{N}}$ and $(g_m)_{m\in\mathbb{N}}$ are Cauchy sequences in $\mathcal{H}_0$ such that $f(x) = \lim\limits_{n\to\infty} f_n(x)$ and $g(x) = \lim\limits_{m\to\infty} g_m(x)$. The bilinearity as well as the positive definiteness is directly inherited from the inner product $\langle\cdot,\cdot,\rangle_{\mathcal{H}_0}$. To see that it is also well-defined let us consider a second choice of a Cauchy sequence $(\hat{f}_n)_{n\in\mathbb{N}}$ with $f(x) = \lim\limits_{n\to\infty} \hat{f}_n(x)$. For $g_m = \sum\limits_{i=1}^{M_m} \beta_{i,m} K(\cdot, x_{i,m})\alpha_{i,m} \in \mathcal{H}_0$ we have

$$\begin{aligned}\langle f_n, g_m\rangle - \langle \hat{f}_n, g_m\rangle &= \langle f_n - \hat{f}_n, g_m\rangle \\ &= \sum_{i=1}^{M_m} \beta_i f_n(x_{i,m})^T\alpha_{i,m} - \hat{f}_n(x_{i,m})^T\alpha_{i,m} \to 0 \quad (n \to \infty).\end{aligned}$$

Similarly, the same holds for other Cauchy sequences $(\hat{g}_m)_{m\in\mathbb{N}}$ which converge pointwise towards g and thus $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is well-defined. By construction $\mathcal{H}$ is complete and by definition of the inner product the reproducing property (2.6)

$$\langle f, K(\cdot, x)\alpha\rangle_{\mathcal{H}} = \lim_{n\to\infty} f_n(x)^T\alpha = f(x)^T\alpha$$

is satisfied as well. To see that $\mathcal{H}$ is in fact unique, let $\hat{\mathcal{H}}$ denote an alternate Hilbert space, such that K satisfies (2.6). Then we have $\mathcal{H}_0 \subset \hat{\mathcal{H}}$ and the inner product on $\mathcal{H}_0$

and the restriction of $\langle \cdot, \cdot \rangle_{\hat{\mathcal{H}}}$ coincide on $\mathcal{H}_0$ and therefore $\mathcal{H} \subset \hat{\mathcal{H}}$ is a closed subspace by construction of $\mathcal{H}$. If $\mathcal{H} \neq \hat{\mathcal{H}}$ then there exists some function $f \in \hat{\mathcal{H}} \setminus \{0\}$ such that f is orthogonal to $\mathcal{H}$, i.e. $\langle f, g \rangle_{\hat{\mathcal{H}}} = 0$ for all $g \in \mathcal{H}$. In particular, we get

$$\langle f, K(\cdot, x)\alpha \rangle_{\hat{\mathcal{H}}} = f(x)^T \alpha = 0$$

for all $x \in \Omega$ and $\alpha \in \mathbb{R}^m$. However, this implies $f(x) = 0$ for all $x \in \Omega$ and therefore $f \equiv 0$. Ultimately, we get a contradiction and we have in fact $\mathcal{H} = \hat{\mathcal{H}}$, i.e. uniqueness. □

In the case of such a pairing, we call K the reproducing kernel of $\mathcal{H}$ and the conditions in (2.6) are refered to as the reproducing property of K. In this case, we also write $\mathcal{H} = \mathcal{H}_K$ to indicate this correspondence.

By Corollary 2.2.3 any subspace $\mathcal{N} \subset \mathcal{H}_K$ is again a RKHS. By the above, it has its own reproducing kernel which can be related to K as follows:

Corollary 2.2.7 (Reproducing kernel for closed subspaces)**.** *Let $\mathcal{N} \subset \mathcal{H}_K$ be a closed subspace and let $\Pi_{\mathcal{N}} : \mathcal{H}_K \to \mathcal{N}$ denote the orthogonal projection onto $\mathcal{N}$, then the reproducing kernel $K_{\mathcal{N}}$ of $\mathcal{N}$ can be defined via*

$$K_{\mathcal{N}}(\cdot, x)\alpha = \Pi_{\mathcal{N}} \left(K(\cdot, x)\alpha \right) \tag{2.9}$$

for any $x \in \Omega$ and all $\alpha \in \mathbb{R}^m$.

Proof. By Theorem 2.2.6 we know that the reproducing kernel of $\mathcal{N}$ is unique. It is therefore sufficient to show that the kernel defined by (2.9) does satisfy the reproducing property (2.6). By definition it immediately follows that

$$K_{\mathcal{N}}(\cdot, x)\alpha = \Pi_{\mathcal{N}} \left(K(\cdot, x)\alpha \right) \in \mathcal{N}$$

as $\Pi_{\mathcal{N}}$ is the projection onto $\mathcal{N}$. Furthermore, for any $f \in \mathcal{N}$ we have

$$\langle f, K_{\mathcal{N}}(\cdot, x)\alpha \rangle_{\mathcal{N}} = \langle f, \Pi_{\mathcal{N}} \left(K(\cdot, x)\alpha \right) \rangle_{\mathcal{H}_K} = \langle \Pi_{\mathcal{N}}(f), K(\cdot, x)\alpha \rangle_{\mathcal{H}_K} = f(x)^T \alpha,$$

where we made use of the fact that $\Pi_{\mathcal{N}}$ is self-adjoint and equal to the identity when restricted to $\mathcal{N}$ itself. □

As we have seen in (2.7), the evaluation $K(\cdot, x)\alpha$ can be seen as the Riesz representer of δ_x^α. At the same time, we can interpret $K(\cdot, x)\alpha$ as having applied the directional point evaluation functional to the rows of $K(\cdot, \cdot)$. This can be generalized when considering more general bounded linear functionals on $\mathcal{H}$:

Proposition 2.2.8 (Riesz representer for bounded functionals). *Let $\lambda : \mathcal{H}_K \to \mathbb{R}$ be a bounded linear functional and let $r_\lambda \in \mathcal{H}_K$ denote its Riesz representer. Then r_λ is given by*

$$r_\lambda = \sum_{i=1}^{m} \lambda^1(K(\cdot, x)e_i)e_i,$$

where the superscript 1 indicates that λ is applied to the function by evaluation in the first component.

Proof. Let $K_x^{e_i} := K(\cdot, x)e_i \in \mathcal{H}_K$, then

$$r_\lambda(x)^T e_i = \langle r_\lambda, K_x^{e_i} \rangle_{\mathcal{H}_K} = \lambda(K_x^{e_i}) = \lambda^1(K(\cdot, x)e_i).$$

Therefore, by summing the above for all standard basis vectors e_i we get

$$r_\lambda(x) = \sum_{i=1}^{m} \left(r_\lambda(x)^T e_i \right) e_i = \sum_{i=1}^{m} \lambda^1(K(\cdot, x)e_i)e_i.$$

□

As we have mentioned above, the superscript 1 indicates that we apply the functional to the function defined by evaluation in the first component. This can be seen as applying the functional to the columns of $K(\cdot, x)$. However, by Definition 2.2.1 we have $K(\cdot, x)e_i = \left(e_i^T K(x, \cdot)\right)^T$. Thus we can equally apply λ to the rows of $K(x, \cdot)$. This leads to the shorter notation

$$r_\lambda(x) = \lambda^2 K(x, \cdot),$$

where the superscript 2 denotes that the function is applied to the second component, i.e. the rows.

It immediately follows that the inner product of any two bounded linear functionals can be computed by applying the functionals to the rows and columns of K successively.

Corollary 2.2.9 (Inner product of functionals). *Let $\lambda, \mu : \mathcal{H}_K \to \mathbb{R}$ be two bounded linear functionals, then it holds*

$$\langle \lambda, \mu \rangle_{\mathcal{H}_K'} = \langle \lambda^2 K, \mu^2 K \rangle_{\mathcal{H}_K} = \lambda^1 \lambda^2 K.$$

Furthermore, for any finite collection of functionals $\{\lambda_1, \ldots, \lambda_p\} \subset \mathcal{H}_K'$, the matrix $C \in$

$\mathbb{R}^{p\times p}$ defined by

$$C_{ij} \coloneqq \lambda_i^1 \lambda_j^2 K \tag{2.10}$$

is positive semi-definite.

Unfortunately, not all operations on $\mathcal{H}_K$ can be expressed as a single functional. Even considering straightforward point evaluation, multiple functionals have to be "stacked" to achieve the desired results. This can be generalized in terms of the so called sampling operator:

Definition 2.2.10 (Sampling operator)**.** Let $\Lambda \coloneqq \{\lambda_1, \dots, \lambda_p\} \subset \mathcal{H}_K'$ be a set of bounded linear functionals on $\mathcal{H}_K$. Then the sampling operator $S_\Lambda : \mathcal{H}_K \to \mathbb{R}^p$ is given by

$$S_\Lambda(f) \coloneqq \begin{pmatrix} \lambda_1(f) & \cdots & \lambda_p(f) \end{pmatrix}^T \in \mathbb{R}^p \tag{2.11}$$

for any $f \in \mathcal{H}_K$

Similar to what we have seen in Proposition 2.2.8 we can likewise apply the sampling operator to the rows or colums of K. In these cases we write

$$S_\Lambda^1 K = \begin{bmatrix} S_\Lambda^1\left(K(\cdot,\cdot)e_1\right) & \dots & S_\Lambda^1\left(K(\cdot,\cdot)e_m\right) \end{bmatrix} \in \mathbb{R}^{p\times m} \tag{2.12}$$

when the sampling operator is applied to the columns, and

$$S_\Lambda^2 K = \left(S_\Lambda^1 K\right)^T \in \mathbb{R}^{m\times p}$$

when applied to the rows. In particular, the Gramian matrix C for the set Λ given in (2.10) now has the compact notation

$$C = S_\Lambda^1 S_\Lambda^2 K. \tag{2.13}$$

Furthermore, for the special case of $\Lambda_X \coloneqq \{\delta_x^{e_1}, \dots, \delta_x^{e_m} \mid x \in X\}$ we may write $S_\Lambda = S_X$ and

$$f(X) \coloneqq S_X(f), \quad K(X,\cdot) \coloneqq S_X^1 K, \quad K(\cdot,X) \coloneqq S_X^2 K, \quad K(X,X) \coloneqq S_X^1 S_X^2 K. \tag{2.14}$$

Using the above, we get the following alternative characterization of s.p.d. kernels.

Corollary 2.2.11. *Let K be a positive definite kernel with RKHS $\mathcal{H}_K$. Then K is strictly positive if and only if for any set of pairwise distinct points $X = \{x_1, \dots, x_n\}$ the set of directional point evaluation functionals $\{\delta_x^{e_1}, \dots, \delta_x^{e_m} \mid x \in X\}$ is linearly independent.*

Proof. The above results follow immediately from the definition of strict positive definiteness and the fact that the Gram matrix of a linearly independent set is positive definite:

$$K \text{ is s.p.d.} \quad \Leftrightarrow \quad K(X,X) \succ 0 \quad \Leftrightarrow \quad \Lambda_X \text{ is linearly independent.}$$

□

Using the sampling operator we can now give a succinct definition of the subspaces we will consider going forward.

Definition 2.2.12. Let $\Lambda = \{\lambda_1, \dots, \lambda_p\} \in \mathcal{H}_K'$. Then we define the subspace $\mathcal{N}(\Lambda) \subset \mathcal{H}_K$ via

$$\mathcal{N}(\Lambda) \coloneqq \operatorname{span}\{\lambda^2 K \mid \lambda \in \Lambda\}.$$

In particular, for $\Lambda = \Lambda_X$ for some $X \subset \Omega$ we may also write

$$\mathcal{N}(X) \coloneqq \mathcal{N}(\Lambda_X).$$

We can immediately conclude that any $f \in \mathcal{N}(\Lambda)$ can now be written as

$$f = S_\Lambda^2 K\alpha \tag{2.15}$$

for some $\alpha \in \mathbb{R}^p$. Furthermore, for the above defined subspaces, we obtain that the orthogonal projection operator onto $\mathcal{N}(\Lambda)$ coincides with the minimal norm interpolation operator with respect to the functional set Λ:

Theorem 2.2.13 (Orthogonal projection and minimal norm interpolation)**.** *Let $\Pi_{\mathcal{N}(\Lambda)} : \mathcal{H}_K \to \mathcal{N}(\Lambda)$ denote the orthogonal projection operator. Then for any $f \in \mathcal{H}_K$ we have*

$$S_\Lambda\left(\Pi_{\mathcal{N}(\Lambda)}(f)\right) = S_\Lambda(f) \tag{2.16}$$

and for all $g \in \mathcal{H}_K$ with $S_\Lambda(g) = S_\Lambda(f)$ we have $\left\|\Pi_{\mathcal{N}(\Lambda)}(f)\right\|_{\mathcal{H}_K} \leq \|g\|_{\mathcal{H}_K}$.

Proof. For all $\alpha \in \mathbb{R}^p$ we have $S_\Lambda^2 K\alpha \in \mathcal{N}(\Lambda)$ and since $\Pi_{\mathcal{N}(\Lambda)}$ is the orthogonal projection we have for all $f \in \mathcal{H}_K$

$$0 = \left\langle \left(\operatorname{id} - \Pi_{\mathcal{N}(\Lambda)}\right)(f), S_\Lambda^2 K\alpha \right\rangle_{\mathcal{H}_K} = \left(S_\Lambda(f) - S_\Lambda\left(\Pi_{\mathcal{N}(\Lambda)}(f)\right)\right)^T \alpha$$

and therefore (2.16) follows. Furthermore, for any $g \in \mathcal{H}_K$ satisfying $S_\Lambda(g) = S_\Lambda(f)$ we have

$$g = \Pi_{\mathcal{N}(\Lambda)}(f) + g^{\perp}$$

for some $g^{\perp}$ in $\mathcal{N}(\Lambda)^{\perp}$ and thus

$$\|g\|^2_{\mathcal{H}_K} = \left\|\Pi_{\mathcal{N}(\Lambda)}(f)\right\|^2_{\mathcal{H}_K} + \left\|g^{\perp}\right\|^2_{\mathcal{H}_K} \geq \left\|\Pi_{\mathcal{N}(\Lambda)}(f)\right\|^2_{\mathcal{H}_K}.$$

□

As an immediate consequence of the above observation, we have that the for expansion of $f \in \mathcal{N}(\Lambda)$ in the form of (2.15) the coefficient vector must solve the linear system

$$S_\Lambda(f) = S_\Lambda\left(\Pi_{\mathcal{N}(\Lambda)}(f)\right) = S^1_\Lambda S^2_\Lambda K\alpha.$$

This is indeed the case, as one can identify the range of the Gram matrix with the range of the sampling operator:

Lemma 2.2.14 (Range of sampling operator)**.** *For any finite collection $\Lambda \subset \mathcal{H}'_K$ we have*

$$\operatorname{range}(S_\Lambda) = \operatorname{range}\left(S^1_\Lambda S^2_\Lambda K\right)$$

Proof. "⊃" Let $\beta \in \operatorname{range}(S^1_\Lambda S^2_\Lambda K)$ and $\alpha \in \mathbb{R}^p$ such that $(S^1_\Lambda S^2_\Lambda K)\,\alpha = \beta$. Define $f \in \mathcal{H}_K$ via

$$f = S^2_\Lambda K\alpha.$$

It follows

$$S_\Lambda(f) = \left(S^1_\Lambda S^2_\Lambda K\right)\alpha$$

"⊂" Let $f \in \mathcal{H}_K$ and let $\Pi_{\mathcal{N}(\Lambda)}f$ be the orthogonal projection of f onto $\mathcal{N}(\Lambda)$. Then

$$\Pi_{\mathcal{N}(\Lambda)}f = S_\Lambda K\alpha$$

for some $\alpha \in \mathbb{R}^p$ and therefore

$$S_\Lambda(f) = S_\Lambda\left(\Pi_{\mathcal{N}(\Lambda)}(f)\right) = S^1_\Lambda S^2_\Lambda K\alpha.$$

□

The above provides us with a necessary condition for whether or not a function $f : \Omega \to \mathbb{R}^m$ is an element of the RKHS:

Corollary 2.2.15 (Necessary condition for $f \in \mathcal{H}_K$)**.** *If $f \in \mathcal{H}_K$, then it holds for any finite collection $\Lambda \in \mathcal{H}'_K$*

$$S_\Lambda(f) \in \operatorname{range}\left(S^1_\Lambda S^2_\Lambda K\right).$$

Likewise, the converse of the above is true. That is if $f : \Omega \to \mathbb{R}^m$ is a function such that $S_\Lambda(f)$ is well defined for any $\Lambda \subset \mathcal{H}'_K$ and $S_\Lambda(f) \in \operatorname{range}(S^1_\Lambda S^2_\Lambda K)$, then we have $f \in \mathcal{H}_K$. To see this we use the special case $\Lambda = \{\lambda\}$ for any $\lambda \in \mathcal{H}'_K$. By the above $\lambda(f)$ is well defined and hence we can identify f as an element of the bidual space of $\mathcal{H}_K$, i.e. $f \in (\mathcal{H}'_K)'$. However, since $\mathcal{H}_K$ is a Hilbert space itself, it can be identified as its bidual space and hence $f \in \mathcal{H}_K$.

Lemma 2.2.14 allows us to give a more compact form for the orthogonal projection onto $\mathcal{N}(\Lambda)$. However, as we have mentioned in Corollary 2.2.11, the Gram matrix $S^1_\Lambda S^2_\Lambda K$ is only positive definite, i.e. invertible, if the set Λ consists of linearly independent elements. In this case the coefficient vector α of the projection $\Pi_{\mathcal{N}(\Lambda)}(f) = S_\Lambda K\alpha$ is given by

$$\alpha = \left(S^1_\Lambda S^2_\Lambda K\right)^{-1} S_\Lambda(f).$$

If the set Λ is not linearly independent, we can replace the inverse by the so called Moore-Penrose-Pseudoinverse which is defined as follows, c.f. [74].

Definition 2.2.16 (Moore-Penrose-Pseudoinverse)**.** Let $A \in \mathbb{R}^{m\times n}$ be an arbitrary matrix. The Moore-Penrose-Pseudoinverse of A denoted by $A^+ \in \mathbb{R}^{n\times m}$ is the unique matrix that satisfies the four Moore-Penrose conditions:

(a) $AA^+A = A$

(b) $A^+AA^+ = A^+$

(c) $\left(AA^+\right)^T = AA^+$

(d) $\left(A^+A\right)^T = A^+A$

Remark 2.2.17. If A is invertible, the Moore-Penrose-Pseudoinverse is just given via $A^+ = A^{-1}$. In the cases where A has full column or row rank, the Moore-Penrose-Pseudoinverse is the left or right inverse of A, respectively. In particular, the conditions guarantee that AA^+ is the orthogonal projection onto range(A). Likewise, A^+A is the orthogonal projection onto range(A^+). Furthermore, if A is symmetric and positive (semi-) definite, then A^+ is also symmetric and positive (semi-) definite.. In practice A^+ can be computed by performing a singular value decomposition for A.

Ultimately, we get the following representation for the orthogonal projection for any $f \in \mathcal{H}_K$

$$\Pi_{\mathcal{N}(\Lambda)}(f) = S^2_\Lambda K \left(S^1_\Lambda S^2_\Lambda K\right)^+ S_\Lambda(f). \tag{2.17}$$

In particular, the reproducing kernel $K_{\mathcal{N}(\Lambda)}$ of the subspace $\mathcal{N}(\Lambda)$ takes the form

$$K_{\mathcal{N}(\Lambda)} = S_\Lambda^2 K \left(S_\Lambda^1 S_\Lambda^2 K\right)^+ S_\Lambda^1 K. \tag{2.18}$$

Remark 2.2.18. Both (2.17) and (2.18) enable us to easily work with p.d. kernels and general functionals. In the case of s.p.d. kernels, the Moore-Penrose-Pseudoinverse coincides with the standard inverse and for the special case of point evaluation, i.e. $S_\Lambda = S_X$ for some $X \subset \Omega$ we obtain the simplified expressions

$$\Pi_{\mathcal{N}(X)}(f) = K(\cdot, X)K(X,X)^{-1}f(X)$$

and

$$K_{\mathcal{N}(X)}(x,y) = K(x,X)K(X,X)^{-1}K(X,y).$$

We conclude this section by showing that the elements of $\mathcal{H}_K$ inherit certain properties of their reproducing kernel K. Namely, if the the Kernel is $2k$-times continuously differentiable over $\Omega \times \Omega$, then each element $f \in \mathcal{H}_K$ is at least k-times continuously differentiable. In other words, for $K \in C^{2k}(\Omega \times \Omega, \mathbb{R}^{m\times m})$ we get the inclusion $\mathcal{H}_K \subset C^k(\Omega, \mathbb{R}^m)$. In particular, we have that if K is continuous so are all functions in the RKHS $\mathcal{H}_K$. The following Theorem and proof are adapted from [108], in which an analogous statement for scalar-valued kernels was shown.

Theorem 2.2.19 (Embedding into the space of continuously differentiable functions)**.** *Let $K \in C^{2k}(\Omega \times \Omega, \mathbb{R}^{m\times m})$ and $\Omega \subset \mathbb{R}^d$, then any $f \in \mathcal{H}_K$ is at least k-times continuously differentiable. Furthermore, let $D_\beta = \partial_{x_1}^{\beta_1} \dots \partial_{x_d}^{\beta_d}$ be a partial differential operator for some multiindex $\beta \in \mathbb{N}_0^d$ with $|\beta| \leq k$. Then $\delta_x^\alpha \circ D_\beta \in \mathcal{H}'$ for all $x \in \Omega$ and $\alpha \in \mathbb{R}^m$.*

Proof. Due to its inductive nature, and symmetry in the multiindex β, we can restrict ourselves to the case $\beta = (1, 0, \dots, 0)^T \in \mathbb{N}_0^d$. We know that if a function f is continuously differentiable, then we have for all $x \in \Omega$

$$\left(\partial_{x_1} f\right)(x) = \lim_{n\to\infty} \frac{f(x + e_1/n) - f(x)}{1/n},$$

For any $x \in \Omega$ and $\alpha \in \mathbb{R}^m$ we can now define a sequence in $\mathcal{H}_K$ via

$$g_n := \frac{1}{n}\left(K(\cdot, x + e_1/n)\alpha - K(\cdot, x)\alpha\right),$$

which converges to $\partial_{x_1} K(\cdot, x)\alpha$ with respect to the norm on $\mathcal{H}_K$. Furthermore, by the reproducing property we have

$$\langle f, g_n \rangle_{\mathcal{H}_K} = \frac{f(x+e_1/n)^T\alpha - f(x)^\alpha}{1/n}$$

and therefore

$$\langle g_m, g_n \rangle_{\mathcal{H}_K} \to \alpha^T \left(\partial^1_{x_1} \partial^2_{x_1} K(x,x) \right) \alpha, \qquad n, m \to \infty.$$

We conclude that $(g_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, as

$$\begin{aligned} \|g_m - g_n\|^2_{\mathcal{H}_K} &= \langle g_m, g_m \rangle_{\mathcal{H}_K} + \langle g_n, g_n \rangle_{\mathcal{H}_K} - 2\langle g_m, g_n \rangle_{\mathcal{H}_K} \\ &\to \alpha^T \left(\partial^1_{x_1}\partial^2_{x_1} K(x,x) + \partial^1_{x_1}\partial^2_{x_1} K(x,x) - 2\partial^1_{x_1}\partial^2_{x_1} K(x,x) \right) \alpha = 0. \end{aligned}$$

Therefore, there exists a $g \in \mathcal{H}_K$ such that $g_n \to g$. For any $f \in \mathcal{H}_K$ it holds

$$(\partial_{x_1} f)(x)^T\alpha = \lim_{n\to\infty} \frac{f(x+e_1/n)^T\alpha - f(x)^\alpha}{1/n} = \lim_{n\to\infty} \langle f, g_n \rangle_{\mathcal{H}_K} = \langle f, g \rangle_{\mathcal{H}_K} \in \mathbb{R}.$$

Since $x \in \Omega$ and $\alpha \in \mathbb{R}^m$ were arbitrary, this show that $\partial_{x_1} f$ exists. In particular, this is the case for $K(\cdot, y)\beta$ and by the reproducing property we have

$$g(y)^T\beta = \langle g, K(\cdot,y)\beta \rangle_{\mathcal{H}_K} = \left(\partial^1_{x_1} K(x,y)\beta \right)^T \alpha = \left(\partial^2_{x_1} K(y,x)\alpha \right)^T \beta,$$

where we made use of (2.1) for the last equality. Hence, we can identify $g = \partial^2_{x_1} K(\cdot, x)\alpha$. To proove that f is differentiable it is sufficient to show that $\partial_{x_1} f$ is continuous. Let $x, y \in \Omega$ and $\alpha \in \mathbb{R}^m$, then we have by the above

$$\begin{aligned} |(\partial_{x_1} f(x) - \partial_{x_1} f(y))^T\alpha|^2 &= \left| \langle f, \partial^2_{x_1} K(\cdot,x)\alpha - \partial^2_{x_1} K(\cdot,y)\alpha \rangle_{\mathcal{H}_K} \right|^2 \\ &\le \|f\|^2_{\mathcal{H}_K} \left\| \partial^2_{x_1} K(\cdot,x)\alpha - \partial^2_{x_1} K(\cdot,y)\alpha \right\|^2_{\mathcal{H}_K} \\ &= \|f\|^2_{\mathcal{H}_K} \alpha^T \left(\partial^1_{x_1}\partial^2_{x_1} K(y,y) + \partial^1_{x_1}\partial^2_{x_1} K(x,x) - 2\partial^1_{x_1}\partial^2_{x_1} K(x,y) \right) \alpha \\ &\to 0 \qquad \text{as } y \to x, \end{aligned}$$

since $\partial^1_{x_1}\partial^2_{x_1} K$ is continuous by assumption. We conclude that $\partial_{x_1} f$ is continuous since $\alpha \in \mathbb{R}^m$ was arbitrary. □

2.3 Basic construction methods and invariant kernels

In this section we present different construction methods for matrix-valued kernels. Over the years a multitude of different approaches to this subject have been presented, see for example [78, 3, 6, 19, 64, 63, 66, 92, 53], many of which make use of the existing theory for scalar-valued kernels and extending these to the matrix-valued (or even operator valued) case. Likewise, many of the following constructions can be seen as extension of different methods for scalar-valued kernels. However, we focus primarly on basic construction methods as well as take a closer look at so called translational (and rotational) invariant kernels on $\mathbb{R}^d \times \mathbb{R}^d$, which can be characterized via their Fourier transformation.

However, before we start with the basic construction methods, we present some examples of matrix-valued kernels.

Example 2.3.1 (Matrix-valued kernels)**.** Let $\Omega \subset \mathbb{R}^d$. The following functions are examples for matrix-valued kernels.

(a) $K(x, y) = xy^T$ is a p.d. kernel and it spans the space of polynomials of degree 1 such that the i-th component is a polynomial in the i-th coordinate x_i of x. By definition it is clear that $K(y, x) = K(x, y)^T$. Furthermore, for any finite set of points $X = \{x_1, \ldots, x_n\}$ we have

$$K(X, X) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}^T \succeq 0$$

and thus the kernel is positive definite.

(b) If $\omega : \Omega \to \mathbb{R}^{m \times m}$ is a symmetric positive definite function, i.e. $\omega(x) = \omega(x)^T \succeq 0$ for all $x \in \Omega$, then $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ given by

$$K(x, y) = \omega(x)\delta_x(y) = \begin{cases} \omega(x), & \text{if } x = y \\ 0, & \text{else} \end{cases}$$

is a p.d. kernel. Again, we immediately notice that $K(y, x) = K(x, y)^T$, since the kernel is zero for $x \neq 0$. Likewise, for any set $X = \{x_1, \ldots, x_n\}$ we have

$$K(X, X) = \begin{pmatrix} \omega(x_1) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \omega(x_n) \end{pmatrix} \succeq 0$$

(c) The Gaussian kernel $K(x,y) = e^{-\varepsilon\|x-y\|^2}$ is a scalar-valued, i.e. $m=1$, s.p.d. kernel.

Matrix-valued kernels come as a natural extension of scalar-valued ones. Hence many properties and construction methods can be directly transferred. In the following we want to list some extensions.

Proposition 2.3.2 (Basic kernel construction)**.** *In the following let $K_1, K_2 : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ denote two (p.d.) matrix-valued kernels.*

(a) *$K = K_1 + K_2$ is a p.d. kernel. Furthermore, if K_1 or K_2 is s.p.d., then so is K.*

(b) *$K = c \cdot K_1$ is a p.d. kernel for any $c \geq 0$.*

(c) *$K = K_1 \odot K_2$, i.e. the Hadamard product (elementwise product) is a p.d. kernel.*

(d) *$K = K_1 \otimes K_2$ is a p.d. kernel. Here $\otimes$ denotes the Kronecker product.*

Proof. Let $X = \{x_1, \dots, x_n\}$ denote a set of (pairwise) distinct points.

(a) It holds

$$K(X,X) = K_1(X,X) + K_2(X,X) \succeq 0$$

and $K(X,X) \succ 0$ is achieved, if $K_1(X,X) \succ 0$ or $K_2(X,X) \succ 0$.

(b) We have $K(X,X) = cK_1(X,X) \succ 0$.

(c) By definition we have $K(X,X) = K_1(X,X) \odot K_2(X,X) \succeq 0$. Here we made use of the well known fact, see [118], that the Hadamard product of positive (semi-) definite matrices is again positive (semi-) definite.

(d) Since the Kronecker product conserves positive (semi-) definiteness, we know that $K_1(X,X) \otimes K_2(X,X)$ is positive definite. However, the matrix $K(X,X)$ is a minor of $K_1(X,X) \otimes K_2(X,X)$, i.e. there exists a matrix $P \in \mathbb{R}^{n^2m^2 \times nm^2}$ such that

$$K(X,X) = P^T \left(K_1(X,X) \otimes K_2(X,X)\right) P \succeq 0.$$

□

We first want to note that **(a)** and **(b)** show that the set of all p.d. matrix-valued kernels mapping into $\mathbb{R}^{m\times m}$ is a cone. Furthermore, we take note of the fact that in the scalar-valued case ($m=1$) the Hadamard and Kronecker product coincide with the regular multiplication in $\mathbb{R}$. Hence, the product of scalar-valued p.d. kernels is again p.d.. For matrix-valued kernels, this is no longer the case. To this end consider $K = K_1 \cdot K_2$.

Since both K_1 and K_2 are kernels, we have by Definition 2.2.1 $K_1(y,x) = K_1(x,y)^T$ and $K_2(y,x) = K_2(x,y)^T$ for any $x, y \in \Omega$. However if we require the same condition for K we obtain

$$\begin{aligned} K_1(y,x)K_2(y,x) = K(y,x) \overset{!}{=} K(x,y)^T &= (K_1(x,y)K_2(x,y))^T \\ &= K_2(x,y)^T K_1(x,y)^T = K_2(y,x)K_1(y,x). \end{aligned}$$

In other words K_1 and K_2 have to commute for all possible input pairs x, y. In general this is not even satisfied if we choose $K_1 = K_2$ and thus ordinary matrix multiplication does not result in a new matrix-valued kernel. As we will see in a later example, i.e. Example 2.4.12, even if the above is satisfied by a p.d. kernel, this does not guarantee that the positive definiteness is preserved.

While Proposition 2.3.2 gives an insight into the basic construction of matrix-valued kernels, it provides no information on the structure of the RKHS of the newly formed kernel or how it relates to the RKHS of the kernels K_1 and K_2. Thus, we consider the following construction procedure, which relies on a linear operator and enables us to directly link the RKHS to the newly constructed kernel to the RKHS of the original kernel used in the construction. The same concept has already been applied in the framework of scalar-valued kernels and the following theorem and proof are modified versions of a result, which can be found in [108] and were updated to the matrix-valued setting.

Theorem 2.3.3. *Let $\mathcal{H}_K$ be an RKHS with reproducing kernel $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$. Let $L : \mathcal{H}_K \to \{f \,|\, f : \Omega_L \to \mathbb{R}^M\}$ be a linear operator mapping from the RKHS into the space of $\mathbb{R}^M$ valued functions over $\Omega_L \subset \Omega$ such that $\delta_x^\alpha \circ L \in \mathcal{H}_K'$ for all $x \in \Omega_L$ and $\alpha \in \mathbb{R}^M$. Then*

$$\mathcal{H}_{K_L} := L(\mathcal{H}_K) = \{L(f) \,|\, f \in \mathcal{H}_K\}$$

is a RKHS with reproducing kernel $K_L : \Omega_L \to \Omega_L \to \mathbb{R}^{M \times M}$ given by

$$K_L := L^1 L^2 K \tag{2.19}$$

and for any $g \in \mathcal{H}_{K_L}$ its norm is given by

$$\|g\|_{\mathcal{H}_{K_L}} = \min\{\|f\|_{\mathcal{H}_K} \,|\, L(f) = g\}. \tag{2.20}$$

Proof. Since $\delta x^{e_i} \circ L$ is a bounded linear operator on $\mathcal{H}_K$ for every $x \in \Omega_L$ and for each standard basis vector $e_i, i = 1, \dots, M$, we have that

$$\mathcal{N}_x^{e_i} := \mathcal{N}(\delta_x^{e_i} \circ L)^{\perp} = \{f \in \mathcal{H}_K \,|\, L(f)(x)^T e_i = 0\}$$

is a closed subspace of of $\mathcal{H}_K$. Consequently, $\mathcal{N}_L \subset \mathcal{H}_K$ given by

$$\mathcal{N}_L := \bigcup_{x \in \Omega_L} \bigcup_{i=1,\dots,M} \mathcal{N}_x^{e_i} = \{f \in \mathcal{H}_K \,|\, L(f) \equiv 0\}$$

is a closed subspace. Therefore, $\mathcal{H}_K = \mathcal{N}_L \oplus \mathcal{N}_L^{\perp}$ and $T := L|_{\mathcal{N}_L^{\perp}} : \mathcal{N}_L^{\perp} \to \mathcal{H}_{K_L}$ is invertible. We can now equip $\mathcal{H}_{K_L}$ with an inner product via

$$\langle g_1, g_2 \rangle_{\mathcal{H}_{K_L}} := \langle T^{-1}(g_1), T^{-1}(g_2) \rangle_{\mathcal{H}_K}.$$

It remains to show, that K_L given by (2.19) satisfies the reproducing property (2.6) and that the above inner product induces the norm defined in (2.20). By definition of K_L we have

$$K_L(\cdot, x)\alpha = L^1 L^2 K(\cdot, x)\alpha = L(L^2 K(\cdot, x)\alpha) \in \mathcal{H}_{K_L}$$

since $L^2 K(\cdot, x)\alpha \in \mathcal{H}_K$ as it is the Riesz representer of $\delta_x^{\alpha} \circ L$ by Proposition 2.2.8. Let $h_x^{\alpha} := T^{-1}(K_L(\cdot, x)\alpha)$, then $h_x^{\alpha} \in \mathcal{N}_L$ and $h_x^{\alpha} - L^2 K(\cdot, x)\alpha \in \mathcal{N}_L^{\perp}$. It follows for any $g \in \mathcal{H}_{K_L}$

$$\begin{aligned}
\langle g, K_L(\cdot, x)\alpha \rangle_{\mathcal{H}_{K_L}} &= \langle T^{-1}(g), h \rangle_{\mathcal{H}_K} = \langle T^{-1}(g), h - L^2 K(\cdot, x)\alpha + L^2 K(\cdot, x)\alpha \rangle_{\mathcal{H}_K} \\
&= \langle T^{-1}(g), h - L^2 K(\cdot, x)\alpha \rangle_{\mathcal{H}_K} + \langle T^{-1}(g), L^2 K(\cdot, x)\alpha \rangle_{\mathcal{H}_K} \\
&= \langle T^{-1}(g), L^2 K(\cdot, x)\alpha \rangle_{\mathcal{H}_K} = L(T^{-1}(g))(x)^T \alpha = g(x)^T \alpha.
\end{aligned}$$

Hence K_L is the reproducing kernel. That the norm (2.20) is induced by the inner product follows from the observation that $L^{-1}(\{g\}) = T^{-1}(g) + \mathcal{N}_L$ and thus

$$\langle g, g \rangle_{\mathcal{H}_{K_L}} = \langle T^{-1}(g), T^{-1}(g) \rangle_{\mathcal{H}_K} = \Big(\min\{\|f\|_{\mathcal{H}_K} \,|\, L(f) = g\}\Big)^2.$$

□

As a direct consequence, we can relate the RKHS of various combinations or modifications of kernels to the individual RKHS.

Corollary 2.3.4.

(a) *Let $K_1, K_2 : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ be p.d. kernels with RKHS $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$, respectively. Then $\mathcal{H}_K := \mathcal{H}_{K_1} + \mathcal{H}_{K_2}$ is a RKHS with reproducing kernel $K = K_1 + K_2$ and with* *norm*

$$\|f\|_{\mathcal{H}_K}^2 = \min\left\{\|f_1\|_{\mathcal{H}_{K_1}}^2 + \|f_2\|_{\mathcal{H}_{K_2}}^2 \mid f = f_1 + f_2\right\}$$

(b) *Let K be a p.d. kernel with RKHS $\mathcal{H}_K$. For any matrix $S \in \mathbb{R}^{M\times m}$ the kernel $K_S : \Omega \times \Omega \to \mathbb{R}^{M\times M}$ defined by $K_S(x,y) = SK_1(x,y)S^T$ is the reproducing kernel of $\mathcal{H}_{K_S} \coloneqq S\mathcal{H}_K = \{Sf \mid f \in \mathcal{H}_K\}$, equipped with the norm*

$$\|f\|_{\mathcal{H}_{K_S}} = \min\left\{\|g\|_{\mathcal{H}_K} \mid Sg = f\right\}.$$

(c) *Let $K : \Omega \times \Omega \subset \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{m\times m}$ be a p.d. kernel with RKHS $\mathcal{H}_K$. If K is twice continuously differentiable, then $\mathcal{H}_{K_\nabla} \coloneqq \{\nabla f \mid f \in \mathcal{H}_K\}$ is an RKHS with reproducing Kernel $K_\nabla : \Omega \times \Omega \to \mathbb{R}^{dm\times dm}$ given by*

$$K_\nabla = \nabla^1\nabla^2 K$$

and norm

$$\|f\|_{\mathcal{H}_{K_\nabla}} = \min\left\{\|g\|_{\mathcal{H}_K} \mid \nabla g = f\right\}.$$

Here, ∇ denotes the stacked gradient operator, i.e.

$$\nabla(f) = \nabla((f_1,\dots,f_m)^T) = \begin{pmatrix}\partial_{x_1} f_1 & \cdots & \partial_{x_d} f_m\end{pmatrix}^T \in \mathbb{R}^{dm}.$$

(d) *Let $K : \Omega \times \Omega \subset \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{m\times m}$ be a p.d. kernel with RKHS $\mathcal{H}_K$. Let $\Omega_r \subset \Omega$ be a subset. Then the kernel $K_r : \Omega_r \times \Omega_r \to \mathbb{R}^{m\times m}$ given by $K_r = K|_{\Omega_r\times\Omega_r}$ is p.d. and its RKHS is given by $\mathcal{H}_{K_r} = \{f|_{\Omega_r} \mid f \in \mathcal{H}_K\}$ equipped with the norm*

$$\|f\|_{\mathcal{H}_{K_r}} = \min\left\{\|g\|_{\mathcal{H}_K} \mid g|_{\Omega_r} = f\right\}.$$

Proof. **(a)** $\mathcal{H} = \mathcal{H}_{K_1} \times \mathcal{H}_{K_2}$ is a RKHS when equipped with the inner product

$$\langle (f_1,f_2),(g_1,g_2)\rangle_{\mathcal{H}} = \langle f_1,g_1\rangle_{\mathcal{H}_{K_1}} + \langle f_2,g_2\rangle_{\mathcal{H}_{K_2}}$$

and reproducing kernel

$$K(x,y) = \begin{pmatrix} K_1(x,y) & 0 \\ 0 & K_2(x,y)\end{pmatrix} \in \mathbb{R}^{2m\times 2m}.$$

We define $L : \mathcal{H}_K \to \{f : \Omega \to \mathbb{R}^m\}$ via

$$L((f_1, f_2)) = f_1 + f_2.$$

It is easy to see that $\delta_x^{e_i} \circ L \in \mathcal{H}_K'$ for any $x \in \Omega$ and each standard basis vector $e_i \in \mathbb{R}^{2m}$. By Theorem 2.3.3 it holds that

$$\mathcal{H}_{K_L} = \{L(f) \,|\, f \in \mathcal{H}_K\} = \{f_1 + f_2 \,|\, f_1 \in \mathcal{H}_{K_1},\, f_2 \in \mathcal{H}_{K_2}\} = \mathcal{H}_{K_1} + \mathcal{H}_{K_2}$$

is a RKHS with reproducing kernel $K_L = L^1 L^2 K = K_1 + K_2$ and with norm

$$\|g\|^2_{\mathcal{H}_{K_L}} = \min\left\{\|f\|^2_{\mathcal{H}_K} \mid L(f) = g\right\} = \min\left\{\|f_1\|^2_{\mathcal{H}_{K_1}} + \|f_2\|^2_{\mathcal{H}_{K_2}} \mid f_1 + f_2 = g\right\}.$$

(b) The result follows by applying Theorem 2.3.3 to $\mathcal{H}_K$, where $L(f) = Sf$.

(c) Using Theorem 2.2.19, we know that $\delta_x^{e_i} \circ \nabla \in \mathcal{H}_K'$. Hence the result follows from Theorem 2.3.3 for the choice $L = \nabla$.

(d) This also follows from Theorem 2.3.3, where L is the restriction operator from Ω onto Ω_r.

□

2.3.1 Translational and rotational invariant kernels

Similar to what we have seen in Theorem 2.2.19, all elements of the RKHS inherit certain properties of their reproducing kernel. Conversely, properties that hold for any $f \in \mathcal{H}_K$ can be traced back to the reproducing kernel. One class of such a property is the invariance under certain transformations.

Lemma 2.3.5 (Behaviour of the reproducing kernel under isometric transformation)**.** *Let $\mathcal{H}_K$ be a RKHS with reproducing kernel $K : \Omega \times \Omega \to \mathbb{R}^{m\times m}$. Then $T : \mathcal{H}_K \to \mathcal{H}_K$ is an isometry if and only if $K = T^1 T^2 K$, where the superscript denote whether the T is applied to the colums or rows of the matrix-valued function K.*

Proof. By assumption we have that $\delta_x^\alpha \circ T \in \mathcal{H}_K'$ since

$$\|(\delta_x^\alpha \circ T)(f)\| \leq \|\delta_x^\alpha\|_{\mathcal{H}_K'} \|(T(f))\| = \|\delta_x^\alpha\|_{\mathcal{H}_K'} \|f\|_{\mathcal{H}_K}.$$

Since T is an isometry, we have $\langle Tf, Tg\rangle_{\mathcal{H}_K} = \langle f, g\rangle_{\mathcal{H}_K}$ for all $f, g \in \mathcal{H}_K$. In particular, this holds for $K_x^\alpha = K(\cdot, x)\alpha$ for any $x \in \Omega$ and $\alpha \in \mathbb{R}^m$. By the above and Propostion 2.2.8 we thus have

$$\alpha^T \left(T^1 T^2 K(x,y)\right) \beta = \left\langle T K_x^\alpha, T K_y^\beta \right\rangle_{\mathcal{H}_K} = \left\langle K_x^\alpha, K_y^\beta \right\rangle_{\mathcal{H}_K} = \alpha^T K(x,y)\beta,$$

which gives us the desired results since $x, y \in \Omega$ and $\alpha, \beta \in \mathbb{R}^m$ were arbitrary. Conversely if $T^1 T^2 K = K$, then T defines an isometry on the space $\mathcal{H}_0 = \operatorname{span}\{K(\cdot, x)\alpha \mid x \in \Omega, \alpha \in \mathbb{R}^m\}$. However, this space is dense in $\mathcal{H}_K$ by Theorem 2.2.6 and hence we can extend it to an isometry on all of $\mathcal{H}_K$. □

Corollary 2.3.6 (Translational invariance)**.** *Let $\mathcal{H}_K$ be an RKHS with reproducing kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{m\times m}$. If the translation operator $T_y : \mathbb{R}^d \to \mathbb{R}^d$, $T_y(x) = x + y$ induces an isometry on $\mathcal{H}_K$ via $T_y(f) := f(\cdot + y)$ for any $y \in \mathbb{R}^d$, then there exists a function $\Phi : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ such that $K(x,y) = \Phi(x-y)$ for all $x, y \in \mathbb{R}^d$. The converse is also true.*

Proof. Let $\mathcal{H}_K$ be an RKHS with reproducing kernel K, such that for any $y \in \mathbb{R}^d$ T_y induces an isometry on $\mathcal{H}_K$. By Lemma 2.3.5 we now have

$$K(x-y, 0) = T^1_{-y} T^2_{-y} K(x,y) = K(x,y)$$

for all $x \in \mathbb{R}^d$. Since $y \in \mathbb{R}^d$ was arbitrary, the above holds for any $x, y \in \mathbb{R}^d$, hence we have $K(x,y) = \Phi(x-y)$, where $\Phi : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ is given by $\Phi := K(\cdot, 0)$. Conversely, if $K(x,y) = \Phi(x-y)$, then we have $T^1_y T^2_y K = K$ and by Lemma 2.3.5 it holds that T_y is an isometry on $\mathcal{H}_K$ onto itself. □

Similar to the above, one can infer further structure on K if, in addition to the translational invariance, a rotational invariance is also stipulated in the sense that every orthogonal matrix $O \in \mathbb{R}^{m\times m}$ induces an isometry on $\mathcal{H}_K$ via $O(f)(x) := f(OxO^T)$. In this case the function Φ is only dependent on the norm of its argument. For more detail on this and on matrix-valued translational (rotational) invariant kernels we defer to [66]. We only summarize the above in the following definition

Definition 2.3.7 (Translational (rotational) invariant kernels)**.** A matrix-valued kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{m\times m}$ is called translational invariant, if there exists a matrix-valued function $\Phi : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ such that $K(x,y) = \Phi(x-y)$. It is further called translation rotational invariant, if Φ only depends on the norm of its argument, i.e. $K(x,y) = \Phi(x-y) = \phi(\|x-y\|)$ for some $\phi : [0,\infty) \to \mathbb{R}^{m\times m}$. In the latter, we call Φ a radial basis function (RBF) and K an RBF kernel. Moreover, the function Φ is called positive definite, if its corresponding matrix-valued kernel is positive definite.

From the above definitions we can immediately conclude property of the function Φ:

Corollary 2.3.8. *Let* $\Phi : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ *a matrix-valued function such that* $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{m\times m}$ *given by* $K(x,y) = \Phi(x-y)$*, then it holds for all* $x \in \mathbb{R}^d$

$$\Phi(x) = \Phi(-x)^T.$$

Proof. Since K is a matrix-valued kernel it holds for all $x, y \in \mathbb{R}^d$

$$\Phi(x-y) = K(x,y) = K(y,x)^T = \Phi(y-x)^T.$$

Choosing $y = 0$ gives the desired identity. □

2.3.2 Native space for RBF kernels

In the following, we give an alternate representation of the native space if the kernel is translational invariant. In this case, the native space $\mathcal{H}_K$ can be expressed via the Fourier transformation of the underlying Φ for which $K(x,y) = \Phi(x-y)$ holds. For scalar-valued kernels, this representation is well-known and we will extend these results to the matrix-valued case. For this purpose we follow the structure outlined in [108].

Definition 2.3.9 (Lebesgue spaces)**.** For $p \geq 1$ the set $\mathcal{L}_p(\Omega, \mathbb{R}^m)$ defined by

$$\mathcal{L}_p(\Omega,\mathbb{R}^m) := \left\{ f : \Omega \to \mathbb{R}^m \,|\, f_i \text{ is Lebesgue-measurable and } \int_\Omega \|f(x)\|^p \,\mathrm{d}x < \infty \right\}$$

is a vector space such that $\|\cdot\|_{\mathcal{L}_p(\Omega,\mathbb{R}^m)}$ given by

$$\|f\|_{\mathcal{L}_p(\Omega,\mathbb{R}^m)} = \left(\int_\Omega \|f(x)\| \,\mathrm{d}x \right)^{1/p}$$

defines a semi-norm. Here we use the Euclidean norm on $\mathbb{R}^m$ for the pointwise norm $\|f(x)\|$. Let $\mathcal{N} = \{f \in \mathcal{L}_p(\Omega,\mathbb{R}^m) \,|\, f \equiv 0 \text{ almost everywhere}\}$. Then the Lebesgue space $L_p(\Omega,\mathbb{R}^m)$ is given as the quotient space

$$L_p(\Omega,\mathbb{R}^m) = \mathcal{L}_p(\Omega,\mathbb{R}^m)/\mathcal{N}$$

and equipped with the norm

$$\|f\|_{L_p(\Omega,\mathbb{R}^m)} := \min\left\{ \|g\|_{\mathcal{L}_p(\Omega,\mathbb{R}^m)} \,|\, g \equiv f \text{ a.e.} \right\}.$$

it becomes a Banach space. Likewise, we can define $L_\infty(\Omega,\mathbb{R}^m)$ via $\mathcal{L}_\infty(\Omega,\mathbb{R}^m)$ where

$$\mathcal{L}_\infty = \left\{ f : \Omega \to \mathbb{R}^m \,|\, f_i \text{ is Lebesgue-measurable and } \|f\|_{\mathcal{L}_\infty(\Omega,\mathbb{R}^m)} = \operatorname*{ess\,sup}_{x\in\Omega} \|f(x)\| < \infty \right\}$$

Analogous definitions hold, if we allow complex valued functions, i.e. $f : \Omega \to \mathbb{C}^m$.

In the case of $p = 2$ the Lebesgue space is a Hilbert space. Furthermore, we make use of matrix-valued functions, hence we identify $L_p(\Omega, \mathbb{R}^{m \times m})$ with $L_p(\Omega, \mathbb{R}^{m^2})$. In particular, the Euclidean norm on $\mathbb{R}^{m^2}$ is then equivalent to the Forbenius norm on $\mathbb{R}^{m \times m}$.

Definition 2.3.10 (Fourier transformation)**.** For $f \in L_1(\mathbb{R}^d, \mathbb{C}^m)$ the Fourier transform $\hat{f}$ of f is given via

$$\mathcal{F}(f)(x) := \hat{f}(x) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(\omega) e^{-\mathrm{i}x^T\omega} \mathrm{d}\omega$$

and the inverse Fourier transform is given by

$$\mathcal{F}^{-1}(f)(x) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(\omega) e^{\mathrm{i}x^T\omega} \mathrm{d}\omega$$

In case of $m \geq 2$ the above integrals operate on each individual component.

Please note, that we defined the Fourier and inverse Fourier transform for complex valued functions. This is more practical, as the Fourier transform $\mathcal{F}(f)$ is in general not real-valued even if f itself was real-valued to begin with.

The following lemma provides us with basic properties of the Fourier and inverse Fourier transform. We will omit the proofs and refer to the literature, such as [96].

Lemma 2.3.11 (Properties of the Fourier transform)**.** *The Fourier transform and inverse Fourier transform as given in Definition 2.3.10 satisfy*

(a) $\int_{\mathbb{R}^d} \hat{f}(x)g(x)\mathrm{d}x = \int_{\mathbb{R}^d} f(x)\hat{g}(x)\mathrm{d}x$ *for all* $f \in L_1(\mathbb{R}^d, \mathbb{R}^m), g \in L_1(\mathbb{R}^d)$.

(b) $\mathcal{F}(f(\cdot - y))(\omega) = e^{-\mathrm{i}y^T\omega}\hat{f}(\omega)$, *for all* $f \in L_1(\mathbb{R}^d, \mathbb{R}^m)$

(c) $\mathcal{F}(f)$ *and* $\mathcal{F}^{-1}(f)$ *are continuous and bounded for all* $f \in L_1(\mathbb{R}^d, \mathbb{R}^m)$.

The next lemma also gives us the necessary tools to prove the main result of this subsection. A proof can be found in [108]:

Lemma 2.3.12. *There exists a sequence* $(g_n)_{n\in\mathbb{N}}$ *of positive functions in* $L_1(\mathbb{R}^d, \mathbb{R})$*, such that*

(a) $\int_{\mathbb{R}^d} g_n(x)\mathrm{d}x = 1$

(b) $\mathcal{F}(\mathcal{F}(g_n)) = g_n$

(c) $\lim_{n\to\infty} \hat{g}_n(x) = \frac{1}{(2\pi)^{d/2}}$

(d) $\Phi(x) = \lim_{n\to\infty} \int_{\mathbb{R}^d} \Phi(\omega) g_n(\omega - x) \mathrm{d}\omega$ *if* $\Phi : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ *is continuous and bounded.*

In [67] it was shown that any translational invariant positive definite matrix-valued kernel can be identified as the Fourier transform of a positive definite, matrix-valued Borel measure. For the sake of completeness, we will list this fact in the following Lemma.

Lemma 2.3.13 (Bochner characterization)**.** *A continuous function* $\Phi : \mathbb{R}^d \to \mathbb{R}^{m\times m}$ *is positive definite if and only if it is the Fourier transform of a finite positive semi-definite self-adjoint matrix-valued Borel measure* μ *on* $\mathbb{R}^d$*.*

As a direct consequence of the above, we can observe the following property for the Fourier transformation of any continuous p.d. Φ.

Lemma 2.3.14. *If* $\Phi \in L_1(\mathbb{R}^d, \mathbb{R}^{m\times m})$ *is continuous and positive definite, then* $\hat{\Phi}(\omega) \succeq 0$ *for all* $\omega \in \mathbb{R}^d$ *and the Fourier transform does not vanish, i.e. it is never identical to the zero matrix. Furthermore,* $\hat{\Phi} \in L_1(\mathbb{R}^d, \mathbb{C}^{m\times m})$ *is self-adjoint*

Proof. By Lemma 2.3.11 we now that $\hat{\Phi}$ is continuous and bounded. Thus by Lemma 2.3.12, Lemma 2.3.11 and Lemma 2.3.13 we have

$$\begin{aligned}
\hat{\Phi}(\omega) &= \lim_{n\to\infty} \int_{\mathbb{R}^d} \hat{\Phi}(x) g_n(x - \omega) \mathrm{d}x \\
&= \lim_{n\to\infty} \int_{\mathbb{R}^d} \Phi(x) \hat{g}_n(x) e^{-\mathrm{i}\omega^T x} \mathrm{d}x \\
&= \lim_{n\to\infty} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} e^{-\mathrm{i}x^T y} \mathrm{d}\mu(y) \hat{g}_n(x) e^{-\mathrm{i}\omega^T x} \mathrm{d}x \\
&= \lim_{n\to\infty} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \hat{g}_n(x) e^{-\mathrm{i}x^T(\omega+y)} \mathrm{d}x \mathrm{d}\mu(y) \\
&= \lim_{n\to\infty} \int_{R^d} g_n(\omega + y) \mathrm{d}\mu(y) \succeq 0
\end{aligned}$$

where we used the fact that μ is a positive semi-definite matrix-valued measure and g_n is positive for all $n \in \mathbb{N}$. Since μ is self-adjoint due to Lemma 2.3.13 and since g_n is real-valued, we immediately conclude that $\hat{\Phi}(\omega)$ is self-adjoint as well. To see that $\hat{\Phi} \in L_1(\mathbb{R}^d, \mathbb{C}^{m\times m})$ we again make use of the afforementioned Lemmata and get

$$\begin{aligned}
\Phi(0) &= \lim_{n\to\infty} \int_{\mathbb{R}^d} \Phi(x) g_n(x) \mathrm{d}x = \lim_{n\to\infty} \int_{\mathbb{R}^d} \hat{\Phi}(\omega) \hat{g}_n(\omega) \mathrm{d}\omega \\
&= \int_{\mathbb{R}^d} \hat{\Phi}(\omega) \lim_{n\to\infty} \hat{g}_n(\omega) \mathrm{d}\omega = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{\Phi}(\omega) \mathrm{d}\omega.
\end{aligned}$$

Here we can exchange taking the limit and integration, as $\hat{\Phi}$ is bounded. Since $\hat{\Phi}(\omega)$ is positive semi-definite by the above, we have

$$\left\| \hat{\Phi}(\omega) \right\| \leq m \operatorname{tr}(\hat{\Phi}(\omega))$$

and thus by the linearity of the trace operator $\operatorname{tr}(\cdot)$

$$\begin{aligned}\left\|\hat{\Phi}\right\|_{L_1(\mathbb{R}^d,\mathbb{R}^{m\times m})} &= \int_{\mathbb{R}^d} \left\|\hat{\Phi}(\omega)\right\| \mathrm{d}\omega \leq \int_{\mathbb{R}^d} \operatorname{tr}\left(\hat{\Phi}(\omega)\right) \mathrm{d}\omega \\ &=\leq \operatorname{tr}\left(\int_{\mathbb{R}^d} \hat{\Phi}(\omega)\mathrm{d}\omega\right) = (2\pi)^{d/2} \operatorname{tr}(\Phi(0)) < \infty.\end{aligned}$$

Furthermore, since the norm does not vanish neither does $\hat{\Phi}$. □

Finally, we are able to prove the main theorem of this subsection which gives us an alternative characterization of the native space $\mathcal{H}_K$ for $K(x,y) = \Phi(x-y)$.

Theorem 2.3.15 (Characterization of $\mathcal{H}_K$ via Fourier transform)**.** *Let $K(x,y) = \Phi(x-y)$ be a continuous, p.d. and translation invariant kernel. Furthermore for any $\omega \in \mathbb{R}^m$ let $\hat{\Phi}(\omega)^{+/2}$ denote a square root of the Moore-Penrose-Pseudoinverse of $\hat{\Phi}(\omega)$. We define the space $\mathcal{H}$ of functions mapping from $\mathbb{R}^d$ into $\mathbb{R}^m$ via*

$$\mathcal{H} := \left\{ f \in L_1(\mathbb{R}^d, \mathbb{R}^m) \,|\, \hat{\Phi}(\omega)^{+/2}\hat{f}(\omega) \in L_2(\mathbb{R}^d, \mathbb{R}^m) \text{ and } \hat{f}(\omega) \in \operatorname{range}(\hat{\Phi}(\omega)) a.e. \right\}$$

and equip it with the inner product

$$\langle f, g\rangle_{\mathcal{H}} := \int_{\mathbb{R}^d} \hat{f}(\omega)^* \hat{\Phi}(\omega)^+ \hat{g}(\omega)\mathrm{d}\omega,$$

where $\hat{f}(w)^*$ denotes the transposed and complex conjugate of $\hat{f}(\omega)$, i.e. $\hat{f}(\omega)^* = \overline{\hat{f}(\omega)}^T$. Then $\mathcal{H} = \mathcal{H}_K$ and the inner products coincide. In particular, K is the reproducing kernel of $\mathcal{H}$.

Proof. We start by showing that $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ does in fact denote an inner product on $\mathcal{H}$. From Lemma 2.3.14 we already now that $\hat{\Phi}(\omega)$ is self-adjoint and positive (semi-) definite. By definition of the Moore-Penrose-Pseudoinverse this also holds true for $\hat{\Phi}(\omega)^+$. Hence $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is at least positive semi-definite sesquilinear. Let $f \in \mathcal{H}$ such that $\langle f, f\rangle_{\mathcal{H}} = 0$ then

$$\hat{f}(\omega)^* \hat{\Phi}(\omega)^+ \hat{f}(\omega) = 0 \quad a.e.,$$

i.e. $\hat{f} \in \operatorname{null}(\hat{\Phi}^{+/2})$. Since $\hat{\Phi}$ is self-adjoint we have

$$\hat{f}(\omega) \in \operatorname{null}(\hat{\Phi}(\omega)^{+/2}) = \operatorname{null}(\hat{\Phi}(\omega)^{1/2}) \subset \operatorname{null}(\hat{\Phi}(\omega))$$

almost everywhere. However, by definition of $\mathcal{H}$ we also have $\hat{f}(\omega) \in \operatorname{range}(\hat{\Phi}(\omega))$ almost everywhere, and therefore $\hat{f}(\omega) = 0$ almost everywhere. Consequently we already have $f = 0$ and thus $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is an inner product. We now show that $\mathcal{H}$ is in fact a Hilbert space, to this end it is sufficient to show that $\mathcal{H}$ is closed under the norm induced by

$\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Let $(f_n)_{n\in\mathbb{N}}$ be a Cauchy-sequence in $\mathcal{H}$. Then $g_n(\omega) := \hat{\Phi}(\omega)^{+/2}\hat{f}_n(\omega)$ defines a Cauchy-sequence in $L_2(\mathbb{R}^d, \mathbb{R}^m)$ since

$$\|g_n - g_m\|^2_{L_2(\mathbb{R}^d,\mathbb{R}^m)} = \int_{\mathbb{R}^d} (\hat{f}_n(\omega) - \hat{f}_m(\omega))^T \hat{\Phi}(\omega)^+ (\hat{f}_n(\omega) - \hat{f}_m(\omega))\mathrm{d}\omega = \|f_n - f_m\|^2_{\mathcal{H}}\,.$$

Thus there exists a $g \in L_2(\mathbb{R}^d, \mathbb{R}^m)$ such that $g_n \to g$. Then $\hat{\Phi}^{1/2}g \in L_1(\mathbb{R}^d, \mathbb{R}^m) \cap L_2(\mathbb{R}^d, \mathbb{R}^m)$ since

$$\begin{aligned}\int\limits_{\mathbb{R}^d} \left\|\hat{\Phi}(\omega)^{1/2}g(\omega)\right\| \mathrm{d}\omega &\le \int_{\mathbb{R}^d} \left(\lambda_{\max}\left(\hat{\Phi}(\omega)\right)\right)^{1/2} \|g(\omega)\| \,\mathrm{d}\omega \\ &\le \left(\int_{\mathbb{R}^d} \lambda_{\max}\left(\hat{\Phi}(\omega)\right) \mathrm{d}\omega\right) \left(\int_{\mathbb{R}^d} \|g(\omega)\|^2 \,\mathrm{d}\omega\right) \\ &\le \left\|\hat{\Phi}\right\|_{L_1(\mathbb{R}^d,\mathbb{R}^{m\times m})} \|g\|_{L_2(\mathbb{R}^d,\mathbb{R}^m)}\,,\end{aligned}$$

where $\lambda_{\max}\left(\hat{\Phi}(\omega)\right)$ denotes the largest eigenvalue of $\hat{\Phi}(\omega)$, and

$$\begin{aligned}\int\limits_{\mathbb{R}^d} \left\|\hat{\Phi}(\omega)^{1/2}g(\omega)\right\|^2 \mathrm{d}\omega &\le \int_{\mathbb{R}^d} \left(\lambda_{\max}\left(\hat{\Phi}(\omega)\right)\right) \|g(\omega)\|^2 \,\mathrm{d}\omega \\ &\le \left\|\hat{\Phi}\right\|_{L_\infty(\mathbb{R}^d,\mathbb{R}^{m\times m})} \|g\|_{L_2(\mathbb{R}^d,\mathbb{R}^m)}\,.\end{aligned}$$

In the last inequality we made use of the fact that $\hat{\Phi}(\omega)$ is self-adjoint and therefore the largest eigenvalue of $\hat{\Phi}(\omega)$ is smaller than its Frobenius norm. Hence, we can apply the inverse Fourier transform and get $f := \mathcal{F}^{-1}(\hat{\Phi}^{1/2}g)$. It now follows with Lemma 2.3.11

$$\begin{aligned}\|f(x) - f_n(x)\|_{L_\infty(\mathbb{R}^d,\mathbb{R}^m)} &\le \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left\|\hat{\Phi}(\omega)^{1/2}g(\omega) - \hat{f}_n(\omega)\right\| \mathrm{d}\omega \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left\|\hat{\Phi}(\omega)^{1/2}g(\omega) - \hat{\Phi}(\omega)^{1/2}\hat{\Phi}(\omega)^{+/2}\hat{f}_n(\omega)\right\| \mathrm{d}\omega \\ &\le \frac{1}{(2\pi)^{d/2}} \left\|\hat{\Phi}\right\|_{L_1(\mathbb{R}^d,\mathbb{R}^{m\times m})} \left\|g - \hat{\Phi}(\omega)^{+/2}\hat{f}_n(\omega)\right\|_{L_2(\mathbb{R}^d,\mathbb{R}^m)}\,.\end{aligned}$$

The above tends to 0 as $n \to \infty$ and hence f is real-valued. Furthermore, we have by definition that $\hat{f}(\omega) = \hat{\Phi}(\omega)^{1/2}g(\omega) \in \mathrm{range}(\hat{\Phi}(\omega))$.

$$\begin{aligned}\|f\|^2_{\mathcal{H}} &= \int_{\mathbb{R}^d} \hat{f}(\omega)^* \hat{\Phi}(\omega)^+ \hat{f}(\omega)\mathrm{d}\omega = \int_{\mathbb{R}^d} g(\omega)^T \hat{\Phi}(\omega)^{1/2}\hat{\Phi}(\omega)^+\hat{\Phi}(\omega)^{1/2}g(\omega)\mathrm{d}\omega \\ &\le \int_{\mathbb{R}^d} \lambda_{\max}\left(\hat{\Phi}(\omega)^{1/2}\hat{\Phi}(\omega)^{+/2}\right) \|g(\omega)\|^2 \,\mathrm{d}\omega \le \|g\|_{L_2(\mathbb{R}^d,\mathbb{R}^m)}\,.\end{aligned}$$

Therefore, $f \in \mathcal{H}$. We now only have to show that $f_n \to f$ with respect to the norm on

$\mathcal{H}$:

$$
\begin{aligned}
\|f\|_{\mathcal{H}}^2 &= \int_{\mathbb{R}^d} \left\| \hat{\Phi}(\omega)^{+/2} \left(\hat{f}(\omega) - \hat{f}_n(\omega) \right) \right\|^2 \mathrm{d}\omega \\
&= \int_{\mathbb{R}^d} \left\| \hat{\Phi}(\omega)^{+/2} \hat{\Phi}(\omega)^{1/2} \left(g - \hat{\Phi}(\omega)^{+/2} \hat{f}_n(\omega) \right) \right\|^2 \mathrm{d}\omega \\
&= \int_{\mathbb{R}^d} \lambda_{\max} \left(\hat{\Phi}(\omega)^{+/2} \hat{\Phi}(\omega)^{1/2} \right)^2 \left\| g - \hat{\Phi}(\omega)^{+/2} \hat{f}_n(\omega) \right\|^2 \mathrm{d}\omega \\
&\leq \left\| g - \hat{\Phi}^{+/2} \hat{f}_n \right\| \to 0, \qquad \text{as } n \to \infty.
\end{aligned}
$$

We now show that K satisfies the reproducing property (2.6) for the above inner product. It then follows that $\mathcal{H} = \mathcal{H}_K$ by Theorem 2.2.6. Let $\alpha \in \mathbb{R}^m$ and $x \in \mathbb{R}^d$. Let $K_x^\alpha = K(\cdot, x)\alpha = \Phi(\cdot - x)\alpha$. Since the Fourier transform operates elementwise, we have with Lemma 2.3.11

$$
\hat{K}_x^\alpha(\omega) = \hat{\Phi}(\omega) e^{-\mathrm{i}x^T\omega}.
$$

Therefore $\hat{K}_x^\alpha(\omega) \in \operatorname{range}(\hat{\Phi}(\omega))$ and

$$
\begin{aligned}
\int_{\mathbb{R}^d} \hat{K}_x^\alpha(\omega)^* \hat{\Phi}(\omega)^+ \hat{K}_x^\alpha(\omega) \mathrm{d}\omega = \int_{\mathbb{R}^d} e^{\mathrm{i}x^T w} \alpha^T \hat{\Phi}(\omega) \alpha e^{-\mathrm{i}x^T w} \mathrm{d}\omega &= \int_{\mathbb{R}^d} e^{\mathrm{i}x^T w} \alpha^T \hat{K}_x^\alpha(\omega) \\
&= \alpha^T K_x^\alpha(x) = \alpha^T K(x, x,)\alpha.
\end{aligned}
$$

Consequently $K(\cdot, x)\alpha \in \mathcal{H}$. For any $f \in \mathcal{H}$ we have similar to the above

$$
\begin{aligned}
\langle f, K_x^\alpha \rangle_{\mathcal{H}} &= \int_{\mathbb{R}^d} \hat{f}(\omega)^* \hat{\Phi}(\omega)^+ \hat{\Phi}(\omega) \alpha e^{-\mathrm{i}x^T w} \mathrm{d}\omega \\
&= \int_{\mathbb{R}^d} \hat{f}(\omega)^* \alpha e^{-\mathrm{i}x^T w} \mathrm{d}\omega \\
&= f(x)^* \alpha = f(x)^T \alpha
\end{aligned}
$$

since $f(x)$ is real-valued. □

Depending on the choice of the translational invariant kernel K and its underlying function Φ, the native space can coincide with a Sobolev space which is given as follows.

Definition 2.3.16 (Sobolev spaces)**.** The Sobolev space of order $s \geq 0$ over $\mathbb{R}^d$ is defined by

$$
W^s(\mathbb{R}^d, \mathbb{R}^m) := \left\{ f \in L_2(\mathbb{R}^d, \mathbb{R}^m) \mid (1 + \|\cdot\|^2)^{s/2} \hat{f} \in L_2(\mathbb{R}^d, \mathbb{R}^m) \right\}.
$$

This is a Hilbert space when equipped with the inner product

$$\langle f, g\rangle_{W^s}^2 = \int_{\mathbb{R}^d} \left(1+\|\omega\|^2\right)^s \hat{f}(\omega)^* \hat{g}(\omega) \mathrm{d}\omega.$$

Likewise, we define $W^s(\mathbb{R}^d, V)$ for any subspace $V \subset \mathbb{R}^m$.

In the case that we restrict ourselves to the condition $s \in \mathbb{N}$ an alternative characterization of the Sobolev space can be made by requiring that all weak derivatives up to order s are contained in $L_2(\mathbb{R}^d, \mathbb{R}^m)$. For more details on the theory of Sobolev spaces we refer to [61]. We only want to remark that in the case $s \in \mathbb{N}$ the Sobolev space can be defined for subsets $\Omega \subset \mathbb{R}^d$ using the aforementioned weak derivatives. We can immediately see from the similarity of the definition of $W^s(\mathbb{R}^d, \mathbb{R}^m)$ and the alternate characterization of the native space, that the native space coincides with $W^s(\mathbb{R}^d, \mathbb{R}^m)$ if the Fourier transform of Φ meets certain conditions.

Corollary 2.3.17 (Sobolev spaces as RKHS)**.** *Let $\Phi \in L_1(\mathbb{R}^d, \mathbb{R}^{m\times m})$ be a continuous p.d. function such that there exists a symmetric positive semi-definite matrix $B \in \mathbb{R}^{m\times m}$ with*

$$c\left(1+\|\omega\|^2\right)^{-s} B \preceq \hat{\Phi}(\omega) \preceq C\left(1+\|\omega\|^2\right)^{-s} B \tag{2.21}$$

for some constants $C > c > 0$ and $s > d/2$. Then the RKHS for the kernel given by $K(x,y) = \Phi(x-y)$ is the Sobolev space $W^s(\mathbb{R}^d, \operatorname{range}(B))$

Proof. We first note that the chain of inequalities in (2.21) guarantees that $\operatorname{range}(\hat{\Phi}(\omega)) = \operatorname{range}(B)$ for all $\omega \in \mathbb{R}^d$. To see this, we note that since B and $\hat{\Phi}(\omega)$ are symmetric we have

$$\mathbb{R}^m = \operatorname{range}(B) \oplus \operatorname{range}(B)^\perp = \operatorname{range}(\hat{\Phi}(\omega)) \oplus \operatorname{range}(\hat{\Phi}(\omega))^\perp. \tag{2.22}$$

Let $b \in \operatorname{range}(B)^\perp$ then we have by (2.21)

$$0 = c\left(1+\|\omega\|^2\right)^{-s} b^T B b \preceq b^T \hat{\Phi}(\omega) b \preceq C\left(1+\|\omega\|^2\right)^{-s} b^T B b = 0$$

and hence $b \in \operatorname{range}(\hat{\Phi}(\omega))^\perp$, i.e. $\operatorname{range}(B) \subset \operatorname{range}(\hat{\Phi}(\omega))^\perp$. Analogously we get $\operatorname{range}(\hat{\Phi}(\omega))^\perp \subset \operatorname{range}(B)$ and therefore with (2.22) we have $\operatorname{range}(B) = \operatorname{range}(\hat{\Phi}(\omega))$. In particular both matrices have the same rank and therefore (cf. [116])

$$\frac{1}{C}\left(1+\|\omega\|^2\right)^s B^+ \preceq \hat{\Phi}(\omega)^+ \preceq \frac{1}{c}\left(1+\|\omega\|^2\right)^s B^+.$$

By definition of the Sobolev space $W^s(\mathbb{R}^d, \operatorname{range}(B))$ we have $f(x) \in \operatorname{range}(B)$ and therefore $\hat{f}(\omega) \in \operatorname{range}(B) = \operatorname{range}(\hat{\Phi}(\omega))$ for all $\omega \in \mathbb{R}^d$. Furthermore, we have for any $f \in W^s(\mathbb{R}^d, \operatorname{range}(B))$

$$\begin{aligned} \|f\|_{\mathcal{H}_K}^2 &= \int_{\mathbb{R}^d} \hat{f}(\omega)^* \hat{\Phi}(\omega)^+ \hat{f}(\omega)^* \mathrm{d}\omega \leq \int_{\mathbb{R}^d} \frac{1}{c} \left(1 + \|\omega\|^2\right)^s \hat{f}(\omega)^* B^+ \hat{f}(\omega)^* \mathrm{d}\omega \\ &\leq \frac{\lambda_{\max}(B^+)}{c} \|f\|_{W^s(\mathbb{R}^d, \operatorname{range}(B))}^2 \end{aligned}$$

And therefore $W^s(\mathbb{R}^d, \operatorname{range}(B)) \subset \mathcal{H}_K$ by Theorem 2.3.15. For the converse inclusion we first note that the assumption $\hat{f}(\omega) \in \operatorname{range}(\hat{\Phi}(\omega)) = \operatorname{range}(B)$ leads to

$$\hat{f}(\omega)^* B^+ \hat{f}(\omega) \geq \lambda_{\min}(B^+) \left\|\hat{f}(\omega)\right\|^2,$$

where $\lambda_{\min}(B^+)$ denotes the smallest non-zero eigenvalue of B^+. Therefore

$$\begin{aligned} \|f\|_{\mathcal{H}_K}^2 &= \int_{\mathbb{R}^d} \hat{f}(\omega)^* \hat{\Phi}(\omega)^+ \hat{f}(\omega)^* \mathrm{d}\omega \geq \int_{\mathbb{R}^d} \frac{1}{C} \left(1 + \|\omega\|^2\right)^s \hat{f}(\omega)^* B^+ \hat{f}(\omega)^* \mathrm{d}\omega \\ &\geq \frac{\lambda_{\min}(B^+)}{C} \|f\|_{W^s(\mathbb{R}^d, \operatorname{range}(B))}^2 \end{aligned}$$

and consequently $\mathcal{H}_K \subset W^s(\mathbb{R}^d, \operatorname{range}(B))$. □

The above equivalency only holds for kernels which are defined on the entire $\mathbb{R}^d$. However, using Theorem 2.3.3 we can determine conditions on the domain $\Omega \subset \mathbb{R}^d$ such that the equivalency carries over.

Corollary 2.3.18. *Let Φ satisfy the assumptions of Corollary 2.3.17 for some $s \in \mathbb{N}$ and some symmetric positive semi-definite $B \in \mathbb{R}^{m \times m}$. Let $\Omega \subset \mathbb{R}^d$ and $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ be given by $K(x, y) = \Phi(x - y)$. Then*

$$\mathcal{H}_K = W^s(\Omega, \operatorname{range}(B)) \tag{2.23}$$

and the norms are equivalent, if and only if there exists a continuous extension operator $E : W^s(\Omega, \operatorname{range}(B)) \to W^s(\mathbb{R}^d, \operatorname{range}(B))$, i.e. $E(f)|_\Omega = f$ and $\|E(f)\|_{W^s(\mathbb{R}^d, \operatorname{range}(B))} \leq C_E \|f\|_{W^s(\Omega, \operatorname{range}(B))}$.

Proof. Let $\mathcal{H}_\Phi$ denote the RKHS of Theorem 2.3.15. By Theorem 2.3.3 we get

$$\mathcal{H}_K = \mathcal{H}_\Phi|_\Omega = W^s(\mathbb{R}^d, \operatorname{range}(B))\Big|_\Omega \subset W^s(\Omega, \operatorname{range}(B)).$$

As we have mentioned before, in the case of $s \in \mathbb{N}$ the Sobolev spaces can alternatively be characterized by having weak derivatives up to order s which all have a finite $L_2(\Omega, \operatorname{range}(B))$ norm. Thus the restriction of $f \in W^s(\mathbb{R}^d, \operatorname{range}(B))$ results in

$f|_\Omega \in W^s(\Omega, \mathrm{range}(B))$. Unfortunately, depending on the domain Ω, not every element is given by this restriction operator. However, the existence of a continuous extension operator guarantees equality in the last case. Conversely, let (2.23) hold and the norms be equivalent. Then by Theorem 2.3.3 we have

$$\|f\|_{\mathcal{H}_K} = \min\left\{\|g\|_{\mathcal{H}_\Phi} \mid g|_\Omega = f,\, g \in \mathcal{H}_\Phi\right\}.$$

We now define $E : \mathcal{H}_K \to \mathcal{H}_\Phi$ via

$$E(f) = \arg\min\left\{\|g\|_{\mathcal{H}_\Phi} \mid g|_\Omega = f,\, g \in \mathcal{H}_\Phi\right\}$$

and obviously we have $E(f)|_\Omega = f$. Furthermore, due to the equivalency of norms it holds

$$\|E(f)\|_{W^s(\mathbb{R}^d,\mathrm{range}(B))} \leq c_1 \|E(f)\|_{\mathcal{H}_\Phi} = c_1 \|f\|_{\mathcal{H}_K} \leq c_1 c_2 \|f\|_{W^s(\Omega,\mathrm{range}(B))}$$

and hence E is a continuous extension operator. □

The existence of such an extension operator is given if the domain Ω is sufficiently nice, for example if it has a Lipschitz boundary, see [1].

One open question that remains is if there exist closed expressions for a function Φ whose Fourier transform satisfies (2.21). Fortunately, we can positively answer this question. In the scalar-valued case, Wendland gave construction formulas for compactly supported radial functions, so called Wendland functions, for any $d, k \in \mathbb{N}$, see [107], such that the corresponding native space is norm equivalent to the Sobolev space $W^k(\mathbb{R}^d, \mathbb{R})$. Using the fact that the Fourier transformation as given in Definition 2.3.10 operates componentwise, we can thus infer that this property carries over if we consider matrix-valued linear combinations of these Wendland functions. Kernels which take this form are called separable and will be the focus of the next section.

2.4 Uncoupled separable kernels

In the following we summarize, modify and extend our prior work on uncoupled separable kernels which were first introduced in our preliminary work [115]. In particular, we added further characterization of uncoupled separable kernels as well as sufficient conditions that guarantee the existence of these types of kernels.

Definition 2.4.1 (Separable kernels). A matrix-valued kernel $K : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ is called separable if there exist scalar-valued kernels $k_i : \Omega \times \Omega \to \mathbb{R}$ and symmetric matrices $Q_i \in \mathbb{R}^{m\times m}$, $i = 1, \dots, p$ such that

$$K(x,y) = \sum_{i=1}^{p} k_i(x,y) Q_i \qquad \text{for all } x, y \in \Omega. \tag{2.24}$$

In this case, the set of tuples $\{(k_i, Q_i)\}_{i=1}^{p}$ is called a decomposition of K and p is called the length. If p is minimal, i.e. there exists no decomposition of length $q < p$, we also refer to it as the order of K.

To guarantee the (strict) positive definiteness of the kernel K, further assumptions have to be made on both the scalar-valued kernels k_i and the symmetric matrices Q_i. Taking a closer look at the Gram matrix $K(X,X)$ for a point set $X = \{x_1, \dots, x_n\} \subset \Omega$ one sees that

$$K(X,X) = \sum_{i=1}^{p} k_i(X,X) \otimes Q_i,$$

where $\otimes$ denotes the Kronecker product. Since sums and Kronecker product of positive (semi-) definite matrices are again positive (semi-) definite, we can conclude that it is sufficient to assume that k_i are p.d. and that the Q_i are positive semi-definite to guarantee the positive definiteness of K. Alternatively, one can apply Corollary 2.3.4 to see that $K_i := k_i Q_i$ is a positive definite kernel, since

$$K_i = \sum_{j=1}^{m} q_j k_i q_j^T$$

where q_j are scaled eigenvectors of Q_i such that $\sum_{j=1}^{m} q_j q_j^T = Q_i$. In order to guarantee that K is also s.p.d. further assumptions have to be made. Obviously it is sufficient that all k_i are s.p.d. and all Q_i are positive definite. This is a rather tight restriction, however, we can loosen it a bit to still maintain strict positive definiteness.

Lemma 2.4.2. *Let K be a separable kernel and $\{(k_i, Q_i)\}_{i=1}^{p}$ a decomposition. If the kernels k_i are s.p.d. and $Q_i \succeq 0$ such that $\sum\limits_{i=1}^{p} Q_i \succ 0$, then K is s.p.d..*

Proof. Let $X = \{x_1, \dots, x_n\} \subset \Omega$ be a set of pairwise distinct points. Furthermore, let $\boldsymbol{K} = K(X,X)$ and $\boldsymbol{K}_i = k_i(X,X)$. As mentioned before, we have

$$\boldsymbol{K} = \sum_{i=1}^{p} \boldsymbol{K}_i \otimes Q_i.$$

Since each k_i is s.p.d. the matrices $\boldsymbol{K}_i$ are positive definite. Let $\lambda = \min\{\lambda_{\min}(\boldsymbol{K}_i) \,|\, i =$

$1, \dots, p\} > 0$. Then we have

$$\boldsymbol{K} = \sum_{i=1}^{p} \boldsymbol{K}_i \otimes Q_i \succeq \sum_{i=1}^{p} \lambda I_n \otimes Q_i = \lambda I_n \otimes \left(\sum_{i=1}^{p} Q_i \right) \succ 0,$$

where we made use of the fact that the Kronecker product commutes with matrix addition. □

Remark 2.4.3. We want to mention that the assumption $\sum_{i=1}^{p} Q_i \succ 0$ has the further benefit that it guarantees the universality of the Kernel K provided that the scalar-valued kernel k_i were universal to begin with. This means that for every compact subset $\Omega_c \subset \Omega$ the subspace

$$\mathcal{N} = \operatorname{span}\{K(\cdot, x)\alpha \,|\, x \in \Omega_c,\ \alpha \in \mathbb{R}^m\}$$

is dense in the set of continuous functions $C(\Omega_c, \mathbb{R}^m)$. For further details on the concept of universality as well as for a proof of the above assertion, we refer to [18, 65, 98].

Lemma 2.4.4 (Sufficient and necessary minimality condition)**.** *Let K be separable kernel such that there exists a decomposition of length p. The the following statements are equivalent*

- ***(a)*** *p is minimal, i.e. no decomposition of shorter length exist*
- ***(b)*** *For any decomposition $\{(k_i, Q_i)\}_{i=1}^{p}$ of length p both sets $\{k_1, \dots, k_p\}$ and $\{Q_1, \dots, Q_p\}$ are linearly independent, respectively.*

Proof. "⇒" Let $\{(k_i, Q_i)\}_{i=1}^{p}$ be a decomposition of length p. Assume that either $\{k_1, \dots, k_p\}$ or $\{Q_1, \dots, Q_p\}$ is linearly dependent, i.e. we can assume without loss of generality that

$$k_1 = \sum_{i=2}^{p} \alpha_i k_i \quad \text{or} \quad Q_1 = \sum_{i=2}^{p} \beta_i Q_i.$$

Therefore

$$k = \sum_{i=1}^{p} k_i Q_i = \sum_{i=2}^{p} k_i (Q_i + \alpha_i Q_1) \quad \text{or} \quad k = \sum_{i=1}^{p} k_i Q_i = \sum_{i=2}^{p} (k_i + \beta_i k_1) Q_i.$$

In either case we have found a smaller decomposition, which contradicts the minimality of p.

"⇐" Let $\{(k_i, Q_i)\}_{i=1}^{p}$ be a decomposition such that $\{k_1, \dots, k_p\}$ and $\{Q_1, \dots, Q_p\}$ are linearly independent. Assume there exists a second decomposition $\{(\hat{k}_i, \hat{Q}_i)\}_{i=1}^{q}$ with length $q < p$. Let $\operatorname{vec} : \mathbb{R}^{m \times m} \to \mathbb{R}^{m^2}$ be the vectorization operator stacking the columns of a matrix $A \in \mathbb{R}^{m \times m}$ on top of one another. By assumption it holds

$$\sum_{i=1}^{p} k_i Q_i = K = \sum_{j=1}^{q} \hat{k}_j \hat{Q}_j$$

and therefore

$$\sum_{i=1}^{p} k_i \operatorname{vec}(Q_i) = K = \sum_{j=1}^{q} \hat{k}_j \operatorname{vec}(\hat{Q}_j). \tag{2.25}$$

We define $Q \in \mathbb{R}^{m^2 \times p}$ and $\hat{Q} \in \mathbb{R}^{m^2 \times q}$ via

$$Q \coloneqq \begin{bmatrix} \operatorname{vec}(Q_1) & \cdots & \operatorname{vec}(Q_p) \end{bmatrix} \quad \text{and} \quad \hat{Q} \coloneqq \begin{bmatrix} \operatorname{vec}(\hat{Q}_1) & \cdots & \operatorname{vec}(\hat{Q}_q) \end{bmatrix}.$$

By assumption we have $\operatorname{rank}(Q) = p$ and hence there exists a left inverse of Q, i.e. a matrix $A \in \mathbb{R}^{p \times m^2}$ such that $AQ = I_p$. Multiplying both sides of (2.25) with A we get

$$\begin{pmatrix} k_1 \\ \vdots \\ k_p \end{pmatrix} = A\hat{Q} \begin{pmatrix} \hat{k}_1 \\ \vdots \\ \hat{k}_q \end{pmatrix}.$$

This shows that $\operatorname{span}\{k_1, \ldots, k_p\} \subset \operatorname{span}\{\hat{k}_1, \ldots, \hat{k}_q\}$ which contradicts the linear independence of the first set. □

Unfortunately, the minimality assumption on the length p of the decomposition is not sufficient to guarantee the uniqueness of the decomposition in the sense that for any two decompositions of length p coincide up to permutation of $\{1, \ldots, p\}$ and scaling of k_i and Q_i, respectively. We illustrate this by the following example:

Example 2.4.5. Let $k_1, k_2 : \Omega \times \Omega \to \mathbb{R}$ denote two linear independent scalar-valued kernels. We define $K : \Omega \times \Omega \to \mathbb{R}^{2 \times 2}$ via

$$K(x, y) = \begin{pmatrix} k_1(x, y) & 0 \\ 0 & k_2(x, y) \end{pmatrix}$$

for all $x, y \in \Omega$. However, this kernel has infinitely many decompositions of length 2. Let $\lambda \in [0, 1]$, then it holds

$$K(x, y) = k_1(x, y) Q_1(\lambda) + \left((1 - \lambda) k_1(x, y) + k_2(x, y)\right) Q_2$$

where

$$Q_1(\lambda) = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \quad \text{and} \quad Q_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

However, we notice that there exists just one decomposition such that the subspaces of $\mathbb{R}^2$ spanned by the columns of $Q_1(\lambda)$ and Q_2 only intersect in $\{0\}$. This leads us to the definition of a new subclass of separable kernels which we denote as uncoupled.

Definition 2.4.6 (Uncoupled separable kernels)**.** Let $K : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ be a separable kernel. If there exists at least one decomposition $\{(k_i, Q_i)\}_{i=1}^p$ such that

$$\operatorname{rank}\left(\sum_{i=1}^p Q_i\right) = \sum_{i=1}^p \operatorname{rank}(Q_i) \tag{2.26}$$

we say K is an uncoupled separable kernel.

While the above formulation differs slightly from the previous observation that the spaces spanned by the matrices Q_i should only intersect in $\{0\}$, we will see in the following lemma that this is in fact equivalent to (2.26).

Lemma 2.4.7. *Let $Q_1, \dots, Q_p \in \mathbb{R}^{m\times m}$ be symmetric matrices. The the following statements are equivalent.*

(a) $\operatorname{rank}\left(\sum_{i=1}^p Q_i\right) = \sum_{i=1}^p \operatorname{rank}(Q_i)$

(b) $\operatorname{range}(Q_i) \cap \operatorname{range}(Q_j) = \{0\}$ *for* $i \neq j$

(c) $\operatorname{range}\left(\sum_{i=1}^p Q_i\right) = \bigoplus_{i=1}^p \operatorname{range}(Q_i)$.

Proof. "**(a)** ⇒ **(b)**" It is sufficient to inspect the case $i = 1$, $j = 2$ as all other cases work analogously. We first note that by definition $\operatorname{rank}(Q) = \dim(\operatorname{range}(Q))$ for any matrix $Q \in \mathbb{R}^{m\times m}$ and furthermore assume, that **(b)** is not satisfied. Otherwise, there is nothing to show. In this case it follows that

$$\begin{aligned} \operatorname{rank}(Q_1 + Q_2) &= \dim(\operatorname{range}(Q_1 + Q_2)) \\ &= \dim(\operatorname{range}(Q_1)) + \dim(\operatorname{range}(Q_2)) - \dim(\operatorname{range}(Q_1) \cap \operatorname{range}(Q_2)) \\ &< \operatorname{rank}(Q_1) + \operatorname{rank}(Q_2). \end{aligned}$$

Consequently, we have

$$\operatorname{rank}\left(\sum_{i=1}^p Q_i\right) < \sum_{i=1}^p \operatorname{rank}(Q_i)$$

which is a contradiction.
"**(b)** ⇒ **(c)**" This follows from the definition of the direct sum of vector spaces, i.e. their intersection has to be the zero space $\{0\}$.
"**(c)** ⇒ **(a)**" Since the sum is direct we immediately get

$$\begin{aligned}\operatorname{rank}\left(\sum_{i=1}^{p} Q_i\right) &= \dim\left(\operatorname{range}\left(\sum_{i=1}^{p} Q_i\right)\right) = \dim\left(\bigoplus_{i=1}^{p}\operatorname{range}(Q_i)\right)\\ &= \sum_{i=1}^{p}\dim(\operatorname{range}(Q_i)) = \sum_{i=1}^{p}\operatorname{rank}(Q_i).\end{aligned}$$

□

Remark 2.4.8. Please note that if we consider a separable kernel K with a decomposition $\{(k_i, Q_i)\}_{i=1}^{p}$ then the positive definiteness of k_i and positive semi-definiteness of Q_i guarantee that the kernel K is itself p.d. (see Lemma 2.4.2). The converse, however, is no longer true as we could extend the decomposition by adding the pairs $(-k_1, Q_1), (k_1, Q_1)$. This still results in the same kernel, however $-k_1$ does not have to be p.d.. However, if we consider uncoupled kernels this still holds.

Theorem 2.4.9 (Positive definite uncoupled separable kernels)**.** *Let K be a p.d. uncoupled separable kernel. Furthermore, let $\{(k_i, Q_i)\}_{i=1}^{p}$ be an uncoupled decomposition. Then k_i is p.d. for any $i = 1, \ldots, p$ and $Q_i \succeq 0$ for all $i = 1, \ldots, p$. Moreover, if K is s.p.d., then so are the k_i and we have $\sum_{i=1}^{p}\operatorname{rank}(Q_i) = m$.*

Proof. Without loss of generality we can assume that k_1 is not positive definite. Thus there exists a set $X = \{x_1, \ldots, x_n\} \subset \Omega$ such that $k_1(X, X)$ has a negative eigenvalue. Let $v \in \mathbb{R}^n$ be an eigenvector for this negative eigenvalue. Furthermore, let $w \in \mathbb{R}^m$ be a vector such that $w^T Q_1 w = 1$ and $w^T Q_i w = 0$ for $i = 2, \ldots, p$. Such a vector exists due to Lemma 2.4.7. We can now define a vector $\alpha \in \mathbb{R}^{mn}$ via the Kronecker product $\alpha = v \otimes w$. For this vector we have

$$\begin{aligned}\alpha^T K(X, X)\alpha = \alpha^T\left(\sum_{i=1}^{p} k_i(X, X) \otimes Q_i\right)\alpha &= \sum_{i=1}^{p}(v \otimes w)^T\left(k_i(X, X) \otimes Q_i\right)(v \otimes w)\\ = \sum_{i=1}^{p}\left(v^T k_i(X, X)v\right)\left(w^T Q_i w\right) &= v^T k_1(X, X)v < 0\end{aligned}$$

and thus K is not p.d.. Analogously it follows that all k_i are s.p.d. if K is s.p.d.. Let us now assume without loss of generality that Q_1 is not positive semi-definite, i.e. there exists a $w \in \mathbb{R}^{p\times p}$ such that $w^T Q_1 w = -1$ and $w^T Q_i w = 0$ for $i = 2, \ldots, p$. This can again be concluded from Lemma 2.4.7 since the matrices are uncoupled. Let $x \in \Omega$ such that $k_1(x,x) > 0$. We conclude

$$w^T K(x,x)w = \sum_{i=1}^{p} k_i(x,x) w^T Q_i w = -k_1(x,x) < 0,$$

i.e. K is not p.d.. If $\sum_{i=1}^{p} \operatorname{rank}(Q_i) \neq m$, then there exists a vector $w \in \mathbb{R}^m$ that lies in the null space of all Q_i. Analogous to above this would lead to $K(x,x)w = 0$ for all $x \in \Omega$ and thus the kernel would not be s.p.d.. □

The RKHS have a particular structure. In in a certain sense they decompose into the RKHS of the scalar-valued kernels k_i, albeit projected into higher dimensions via the matrices Q_i. The previous theorem just guarantees the existence of the RKHS for the scalar-valued kernels.

Theorem 2.4.10 (RKHS for uncoupled separable kernels)**.** *Let K be an uncoupled separable p.d. kernel with decomposition $\{(k_i, Q_i)\}_{i=1}^{p}$. Let K_i denote the uncoupled separable kernel with decomposition $K_i = k_i Q_i$. Furthermore, let $q_1^i, \ldots, q_{r_i}^i$ denote a basis of* range(Q_i) *and $r_i =$* rank(Q_i)*. Then we have*

$$\mathcal{H}_{K_i} = \bigoplus_{j=1}^{r_i} \mathcal{H}_{k_i} q_j^i \tag{2.27}$$

for all $i = 1, \ldots, p$. Furthermore, it holds

$$\mathcal{H}_K = \bigoplus_{i=1}^{p} \mathcal{H}_{K_i}. \tag{2.28}$$

Proof. We first note that by assumption $q_1^i, \ldots, q_{r_i}^i$ are linearly independent and thus the sum in the right hand side of (2.27) is direct. It remains to show that it spans the entire space. One easily sees that $\mathcal{H}_{k_i} q_j^i \subset \mathcal{H}_{K_i}$ for each $j = 1, \ldots, r_i$ and thus the sum is a subspace. For the converse inclusion we simply note that $K_i(\cdot, x)\alpha \in \bigoplus_{j=1}^{r_i} \mathcal{H}_{k_i} q_j^i$ since the $\mathcal{H}_{k_i} q_j^i \subset \mathcal{H}_{K_i}$ form a basis of range(Q_i). Hence, the span and subsequent closure of all such element lies in the right hand side of (2.27). However, this is precisely $\mathcal{H}_{K_i}$ by Theorem 2.2.6. In an analogous fashion (2.28) can be concluded from the uncoupledness of the matrices Q_i. □

With this notion of uncoupledness we can now impose a sufficient condition for the uniqueness of a minimal decomposition.

Theorem 2.4.11 (Uniqueness of minimal uncoupled decomposition)**.** *Let K be an uncoupled separable kernel and let $\{(k_i, Q_i)\}_{i=1}^{p}$ be an uncoupled decomposition. If p is minimal, then the decomposition is unique up to permutations and scaling.*

Proof. Let $\{(k_i, Q_i)\}_{i=1}^p$ and $\{(\hat{k}_j, \hat{Q}_j)\}_{j=1}^p$ be two uncoupled decompositions. Let $Q = \sum_{i=1}^p Q_i$. Since the first decomposition is uncoupled we have by the previous Lemma

$$\operatorname{range}(Q) = \operatorname{range}(Q_i) + \operatorname{range}(Q - Q_i)$$

for all $i = 1, \dots, p$. Thus there exist vectors $c_i \in \mathbb{R}^m$ such that $Q_i c_i \neq 0$ and $Q_j c_i = 0$ for all $i \neq j$. Therefore,

$$k_i Q_i c_i = K c_i = \sum_{i=1}^p \hat{k}_j \hat{Q}_j c_i,$$

i.e. we can write each k_i as a linear combination of $\hat{k}_1, \dots, \hat{k}_p$. Let $A = (a_{i,j})_{i,j=1}^p$ be the coefficient matrix such that

$$\begin{pmatrix} k_1 \\ \vdots \\ k_p \end{pmatrix} = A \begin{pmatrix} \hat{k}_1 \\ \vdots \\ \hat{k}_p \end{pmatrix}.$$

Since p is minimal, the scalar-valued kernels of each decomposition are linearly independent by Lemma 2.4.4 and thus

$$\hat{Q}_j = \sum_{i=1}^p a_{ij} Q_i$$

holds as well. By Lemma 2.4.7 it now holds

$$\operatorname{range}(\hat{Q}_j) = \bigoplus_{i=1}^p a_{ij} \operatorname{range}(Q_i).$$

However, we have for any $j \neq j'$ that $\operatorname{range}(\hat{Q}_j) \cap \operatorname{range}(\hat{Q}_{j'}) = \{0\}$ and consequently a_{ij} or $a_{ij'}$ are equal to 0. Since this holds for all $i, j = 1, \dots, p$, we have that for any i there exists exactly one $j = j(i)$ such that $a_{ij(i)} \neq 0$. In other words $k_i = a_{i(ji)} \hat{k}_{j(i)}$. Furthermore, we have

$$\begin{aligned} 0 = K - K &= \sum_{i=1}^p k_i Q_i - \sum_{j=1}^p \hat{k}_j \hat{Q}_j \\ &= \sum_{i=1}^p k_i Q_i - \sum_{i=1}^p \hat{k}_{j(i)} \hat{Q}_{j(i)} \\ &= \sum_{i=1}^p \hat{k}_{j(i)} \left(a_{ij(i)} Q_i - \hat{Q}_{j(i)} \right) \end{aligned}$$

which leads to $a_{ij(i)}Q_i = \hat{Q}_{j(i)}$. Ultimately, the decompositions coincide up to the permutation $i \mapsto j(i)$ and the scalings described by the coefficients of A. □

In general, the existence of an uncoupled or even minimal uncoupled decomposition cannot be guaranteed, since (2.26) necessitates that the length of any uncoupled decomposition is at most m. Hence, any separable kernel of order $p \geq m+1$ cannot be uncoupled. In the following we present sufficient and necessary conditions for the existence of an uncoupled decomposition. For this purpose we want to recall that in the scalar-valued case the product of p.d. kernels is again a p.d. kernel. As we have seen in the matrix-valued case, this is only possible if the kernels commute for any input pairs $(x, y) \in \Omega \times \Omega$, i.e.

$$K_1(x,y)K_2(x,y) = K_2(x,y)K_1(x,y), \qquad \text{for all } x, y \in \Omega.$$

Unfortunately, the above is not sufficient to guarantee positive definiteness in the matrix-valued case which we illustrate with the following example.

Example 2.4.12. Let $k_1, k_2 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the kernels given by

$$k_1(x,y) = e^{-\frac{1}{10}(x-y)^2} \qquad \text{and} \qquad k_2(x,y)e^{-(x-y)^2}.$$

Furthermore let $Q_1, Q_2 \in \mathbb{R}^{2\times 2}$ be the symmetric matrices

$$Q_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad \text{and} \qquad Q_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let K be a separable kernel with decomposition $\{(k_1, Q_1), (k_2, Q_2)\}$ and $X = \{0, 1\}$. By Lemma 2.4.2 the kernel is s.p.d. and obviously the kernel commutes with itself. However the kernel K^2 is not even p.d. as

$$K^2(X,X) = \begin{pmatrix} 5 & 3 & 2e^{-\frac{1}{5}} + 2e^{-\frac{11}{10}} + e^{-2} & 2e^{-\frac{1}{5}} + e^{-\frac{11}{10}} \\ 3 & 2 & 2e^{-\frac{1}{5}} + e^{-\frac{11}{10}} & 2e^{-\frac{1}{5}} \\ 2e^{-\frac{1}{5}} + 2e^{-\frac{11}{10}} + e^{-2} & 2e^{-\frac{1}{5}} + e^{-\frac{11}{10}} & 5 & 3 \\ 2e^{-\frac{1}{5}} + e^{-\frac{11}{10}} & 2e^{-\frac{1}{5}} & 3 & 2 \end{pmatrix}$$

has a negative eigenvalue $\lambda \approx -0.044$.

Upon further inspection of the Gram matrix $K^2(X,X)$ we see that it can be written via the block-Hadamard product

$$K^2(X,X) = K(X,X) \,\square\, K(X,X) := (K(x_i,x_j)K(x_i,x_j))_{i,j}.$$

As it was shown in [37], the block Hadamard product of two positive (semi-)definite block matrices $A = (A_{ij})$ and $B = (B_{ij})$ is in general only positive definite if each block of A commutes with each block of B. If we apply this restriction to the Gram matrix $K(X, X)$ for any finite collection of points $X \subset \Omega$, this results in the condition

$$K(x,y)K(x',y') = K(x',y')K(x,y) \qquad \text{for all } x,y,x',y' \in \Omega.$$

If the above is satisfied by some p.d. kernel K we obtain the following characterization.

Theorem 2.4.13 (Orthogonally uncoupled kernels)**.** *Let $K : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ be a p.d. kernel such that $K(x,y) = K(y,x)$ for all $x, y \in \Omega$, then the following statements are equivalent*

(a) $K(x,y)K(x',y') = K(x',y')K(x,y)$ *for all $x,y,x',y' \in \Omega$.*

(b) *There exists an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ such that $Q^T K(x,y)Q$ is diagonal for all $x, y \in \Omega$.*

(c) *K is uncoupled separable and there exists an uncoupled decomposition $\{(k_i, Q_i)\}_{i=1}^p$ with length $p \le m$, symmetric Q_i and for which $Q_iQ_j = 0$ for $i \neq j$.*

Proof. "**(a)** ⇒ **(b)**" Let $A_1, \ldots, A_D$ denote a basis of $\operatorname{span}\{K(x,y)|x, y \in \Omega\}$. Then the A_i are symmetric and commute with one another. Hence they are simultaneously diagonalizable, i.e. there exists an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ such that $Q^T A_i Q$ is diagonal. It follows that $K(x,y) \in \operatorname{span}\{A_1, \ldots, A_d\}$ is diagonalizable via Q for any $x, y \in \Omega$.

"**(b)** ⇒ **(c)**" By assumption we have

$$Q^T K(x,y)Q = \begin{pmatrix} k_1(x,y) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & k_m(x,y) \end{pmatrix},$$

where $k_1, \ldots, k_m : \Omega \times \Omega \to \mathbb{R}$ are scalar-valued kernels. For $i = 1, \ldots, m$ let $J(i) := \{j \,|\, k_i = a_{ij}k_j \text{ for some } a_{ij} > 0\}$. Then there exist $i_1, \ldots, i_p$ with minimal p such that

$$\bigcup_{l=1}^{p} J(i_l) = \{1, \ldots, m\} \qquad \text{and} \qquad J(i_l) \cap J(i_k) = \emptyset \text{ for } i_l \neq i_k.$$

We now have

$$K = \sum_{i=1}^{m} k_i (Qe_i)(Qe_i)^T = \sum_{l=1}^{p} k_{i_l} \sum_{j \in J(i_l)} a_{i_l,j} (Qe_j)(Q_e j)^T = \sum_{l=1}^{p} k_{i_l} Q_{i_l},$$

where

$$Q_{i_l} = \sum_{j \in J(i_l)} a_{i_l,j} (Qe_j)(Q_e j)^T.$$

Since the columns of Q are orthogonal and the sets $J(i_l)$ have empty intersection, one easily verifies $Q_{i_l} Q_{i_k} = 0$ for $i_l \neq i_k$.
"**(c)** $\Rightarrow$ **(a)**" It holds

$$\begin{aligned} K(x,y)K(x',y') &= \left(\sum_{i=1}^{p} k_i(x,y) Q_i \right) \left(\sum_{j=1}^{p} k_j(x',y') Q_j \right) = \sum_{i,j=1}^{p} k_i(x,y) k_j(x',y') Q_i Q_j \\ &= \sum_{i,j=1}^{p} k_j(x',y') k_i(x,y) Q_j Q_i \\ &= \left(\sum_{j=1}^{p} k_j(x',y') Q_j \right) \left(\sum_{i=1}^{p} k_i(x,y) Q_i \right) = K(x',y')K(x,y). \end{aligned}$$

□

As an immediate consequence of the above, we may apply analytical functions to any K satisfying one of the above conditions.

Corollary 2.4.14. *Let $h : \mathbb{R}^{m \times m} \to \mathbb{R}^{m \times m}$ be an analytical function such that the coefficients in the analytical expansion are positive. If $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ satisfies any one of the conditions in Theorem 2.4.13, then $h \circ K$ is a p.d. kernel.*

Proof. By Theorem 2.4.13 K^n is a p.d. kernel for all $n \in \mathbb{N}$. Therefore $h \circ K$ is p.d. since h is analytical. □

In the case that none of the conditions of Theorem 2.4.13 are satisfied we can still pose sufficient and necessary conditions for a p.d. kernel to be uncoupled.

Theorem 2.4.15. *Let $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ be a p.d. kernel. Then the following statements are equivalent.*

- ***(a)*** *K is uncoupled separable.*
- ***(b)*** *There exists an invertible $P \in \mathbb{R}^{m \times m}$ such that $P^T K(x,y) P$ is diagonal for all $x, y \in \Omega$.*

Proof. "⇒" Let $\{(k_i, Q_i)\}_{i=1}^p$ be an uncoupled decomposition of K. By assumption K is p.d. and therefore the symmetric matrices Q_i are positive semi-definite. Hence we can write

$$Q_i = \sum_{j=1}^{r_i} q_j^i \left(q_j^i\right)^T$$

where $r_i = \operatorname{rank}(Q_i)$ and $q_1^i, \dots q_{r_i}^i$ are linearly independent scaled eigenvectors of Q_i such that $\left(q_j^i\right)^T Q_i q_j^i = 1$. Furthermore, we have $r = \sum_{i=1}^p r_i \leq m$ and therefore the matrix $P' \in \mathbb{R}^{m \times r}$ given by

$$P' = \begin{pmatrix} q_1^1 & \cdots & q_{r_p}^p \end{pmatrix}$$

has rank r. Hence we can extend P' to an invertible matrix $P \in \mathbb{R}^{m \times m}$, such that the first r columns coincide with P'. For this P we now have that $D_i := P^T Q_i P$ is diagonal such that $(D_i)_{jj} = 1$ if $\sum_{l=1}^i r_i + 1 \leq j \leq \sum_{l=1}^{i+1} r_i$ and zero otherwise. Consequently

$$P^T K(x,y) P = \sum_{i=1}^p k_i P^T Q_i P = \sum_{i=1}^p k_i D_i$$

is diagonal.
"⇐" The proof is analogous to the one of Theorem 2.4.13 implication "**(b)** ⇒ **(c)**". □

Unfortunately, there is no easy way to check if condition **(b)** of Theorem 2.4.15 is satisfied. Even when we consider a separable kernel K with decomposition $\{(k_i, Q_i)\}_{i=1}^p$ the condition still translates to the matrices Q_i being simultaneously diagonalizable albeit with a possibly non orthogonal matrix P. This is still an open problem if more than two matrices are considered, and the existence and computation of a respective diagonalization matrix P has many applications in the field of signal processing [117].

As an immediate consequence, we can deduce that any separable kernel of order 2 is uncoupled provided one of the matrices in its decomposition has full rank.

Corollary 2.4.16. *Let $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$ be a separable p.d. kernel of order 2 such that there exists a decomposition $\{(k_1, Q_1), (k_2, Q_2)\}$ with p.d. kernels k_i and positive semi-definite matrices Q_i such that $\operatorname{rank}(Q_2) = m$. Then there exists an uncoupled decomposition of at most length $p = \operatorname{rank}(Q_1) + 1$.*

Proof. Since Q_2 has full rank, it is positive definite. Hence there exists a Cholesky decomposition $LL^T = Q_2$, where L is an invertible lower triangular matrix. Let $A := L^{-1} Q_1 L^{-T}$, then A is symmetric and positive semi-definite and hence there exists an orthogonal matrix Q such that $D = Q^T A Q$ is diagonal such that $D_{ii} \neq 0$ for $i \leq \operatorname{rank}(Q_1)$ and zero otherwise. The matrix $P \coloneqq L^{-T}Q$ is invertible and guarantees that P^TKP is diagonal for all $x, y \in \Omega$. By Theorem 2.4.15 it follows that K is uncoupled. Likewise, P^TKP is an uncoupled kernel that has the decomposition

$$\begin{aligned} P^TKP = k_1D + k_2I &= \sum_{i=1}^{\operatorname{rank}(Q_1)} k_1 d_{ii} e_i e_i^T + k_2 I \\ &= \sum_{i=1}^{\operatorname{rank}(Q_1)} (d_{ii}k_1 + k_2) e_i e_i^T + k_2 \left(\sum_{i=\operatorname{rank}(Q_1)+1}^{m} e_i e_i^T \right) \end{aligned}$$

which is of length $\operatorname{rank}(Q_1) + 1$. Consequently $K = P^{-T}P^TKPP^{-1}$ is uncoupled with a decomposition of at most length $p = \operatorname{rank}(Q_1)+1$. This follows since for any matrices $Q, \hat{Q}$ with $\operatorname{range}(Q) \cap \operatorname{range}(\hat{Q}) = \{0\}$ we have $\operatorname{range}(P^{-T}QP^{-1}) \cap \operatorname{range}(P^{-T}\hat{Q}P^{-1}) = \{0\}$. Thus we can see that uncoupledness is preserved by using Lemma 2.4.7. □

We now describe a general procedure by which suitable uncoupled separable kernels can be constructed, if sufficient data on the inputs and outputs of the target function is provided. For this purpose, let $X = \{x_1, \dots, x_N\} \subset \Omega$ and $Y = \{y_1, \dots, y_N\} \subset \mathbb{R}^m$ be a sufficient amount of input and output data. We then compute the empirical covariance matrix $\operatorname{Cov}(Y_n) \in \mathbb{R}^{m \times m}$ for a smaller subset $Y_n = \{y_1, \dots, y_n\} \subset Y$ via

$$\operatorname{Cov}(Y_n) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T,$$

where $\mu \in \mathbb{R}^m$ denotes the arithmetic mean of the output data Y_n. By definition, the empirical covariance matrix $\operatorname{Cov}(Y_n)$ is symmetric and positive (semi-) definite. Hence there exists a spectral decomposition

$$\operatorname{Cov}(Y_n) = U(Y_n)^T \Sigma(Y_n) U(Y_n)$$

with orthogonal matrix $U(Y_n) = \begin{bmatrix} u_1(Y_n) & \dots & u_n(Y_n) \end{bmatrix} \in \mathbb{R}^{m \times m}$ and diagonal matrix $\sigma(Y_n) = \operatorname{diag}(\sigma_1(Y_n), \dots, \sigma_m(Y_n)) \in \mathbb{R}^{m \times m}$ sorted in descending order, i.e. $\sigma_1(Y_n) \geq \sigma_2(Y_n) \geq \dots \geq \sigma_m(Y_n)$. By choosing suitable kernels $k_1, \dots, k_p$, suitable index sets $\mathcal{I}_1, \dots, \mathcal{I}_p$, with

$$\mathcal{I}_i \cap \mathcal{I}_j = \emptyset \qquad \text{and} \qquad \bigcup_{i=1}^{p} \mathcal{I}_i = \{1, \dots, n\}$$

and computing their corresponding matrices

$$Q_i = \sum_{j \in \mathcal{I}_j} u_j(Y_n) u_j(Y_n)^T, \qquad i = 1, \dots, p$$

we obtain an uncoupled separable decomposition $\{(k_i, Q_i)\}_{i=1}^{p}$ via Theorem 2.4.13. Unfortunately, the selection of the scalar-valued kernels k_i and index sets $\mathcal{I}_i$ is extremely sensitive with respect to the provided data. Hence, trying to optimize the selection procedure is exceptionally expensive and in general practice infeasible. Nonetheless, one can try heuristic approaches based on the eigenvalues $\sigma_i(Y_n)$. Since large values $\sigma_i(Y_n) \gg 1$ corresponds to high variance in the given data in the direction $u_i(Y_n)$ and small values $\sigma_i(Y_n) \ll 1$ likewise correspond to small variance along the direction $u_i(Y_n)$, it is sensible to group indices for which the eigenvalues have similar magnitude. Furthermore, index sets representing small values can be handled using simpler scalar-valued kernels k_i, whereas the sets representing larger values should be coupled with more sophisticated kernels. This is advisable, as these indexes will have the most impact on the quality of the resulting kernel approximation. Both the index and kernel selection procedure can be described via corresponding selection methods, which depent on the eigenvalues $\Sigma(Y_n)$, directions $U(Y_n)$ and the previously unused output data $Y \setminus Y_n$. In the case of scalar-valued kernels, the latter is commonly required for validation processes that are used to fine tune certain parameters in the kernels [101]. We summarize the above procedure into the following algorithm

```
Algorithm 1: Generation of uncoupled separable kernels from functional data
  Data: Output data $Y_n \subset Y \subset \mathbb{R}^m$ generated by some unknown function
        $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$, suitable index set selection method $\text{ind}_{\text{select}}$, suitable
        kernel selection method $\text{kernel}_{\text{select}}$.
  Result: Uncoupled separable decomposition $\{(k_i, Q_i)\}_{i=1}^{p}$ for a matrix-valued
          kernel $K$.
1 Compute the empirical covariance matrix $\text{Cov}(Y_n)$
2 Compute the eigenvalues and eigenvectors of the empirical covariance matrix
    $\text{Cov}(Y_n) = U(Y_n)^T \Sigma(Y_n) U(Y_n)$
3 Determine the index sets via the chosen selection methods
    $(\mathcal{I}_1, \dots, \mathcal{I}_p) = \text{ind}_{\text{select}}(\Sigma(Y_n), U(Y_n))$
4 Determine the scalar kernel functions via the kernel selection method
    $(k_1, \dots, k_p) = \text{kernel}_{\text{select}}(\Sigma(Y_n), \mathcal{I}_1, \dots, \mathcal{I}_p, Y)$
```

Remark 2.4.17. Performing a spectral decomposition of the covariance matrix $\text{Cov}(Y_n)$ to extract information about the target function from the given data, is not the only way in which uncoupled kernels can be constructed. One might use other methods such as directly performing a singular value decomposition on the data Y_n to obtain matrices $U(Y_n)$ and $\Sigma(Y_n)$ and then continue with step 3 and 4 as outlined in Algorithm 1.

We conclude this section with a small numerical example, which illustrates how uncoupled separable kernels can be more beneficial than a componentwise approach.

3-dimensional Example

Let $\Omega = [-2, 2] \subset \mathbb{R}$. We consider the target function $f : \Omega \to \mathbb{R}^3$ given by

$$f(x) = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{\sqrt{2}}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{pmatrix} \begin{pmatrix} e^{-2.5(x-0.5)^2} + e^{-2.0(x+0.5)^2} \\ e^{-3.5(x-0.7)^2} \\ 1 \end{pmatrix}$$

Making use of Algorithm 1, we obtain the following two kernels $K_1, K_2 : \Omega \times \Omega \to \mathbb{R}^{3\times 3}$ given by

$$\begin{aligned} K_1(x, y) &= e^{-\varepsilon_{11}(x-y)^2} I_3, \\ K_2(x, y) &= e^{-\varepsilon_{21}(x-y)^2} u_1 u_1^T + e^{-\varepsilon_{22}(x-y)^2} \left(u_2 u_2^T + u_3 u_3^T \right) \end{aligned}$$

with parameters $\varepsilon_{11}, \varepsilon_{21}, \varepsilon_{22} \in (0, \infty)$. The vectors u_1, u_2, u_3 correspond to the eigenvalues of the empirical covariance matrix for output data Y generated by 401 random evaluations of f.

The kernel K_1 reflects a global approach, i.e. the index selection method results in a single index set containing all components. This is not the case for the kernel K_2. The approximation is computed as the interpolant/best approximation in the spaces $\mathcal{N}_1(X_N)$ and $\mathcal{N}_2(X_N)$, respectively, where $X_N \subset \Omega$ consists of $N = 35$ equidistantly spaced points.

The parameters are determined by minimizing the maximum pointwise error on a validation set $\Omega_M \subset \Omega$ consisting of 40 randomly chosen points and for 50 logarithmic equidistantly distributed parameters in $[0.1, 100]$. The parameters selected by the above are depicted in Table 2.1.

Parameter	ε_{11}	ε_{21}	ε_{22}
Value	1.931	0.244	3.393

Table 2.1: Results of the parameter selection for the different kernels.

The pointwise error measured in the Euclidean norm is depicted in Figure 2.1. We can see a maximum pointwise error in the order of magnitude 10^{-7} for K_1 and 10^{-9} for K_2. These occur near the boundary of the domain. However, even for the interior of the domain we obtain an improvement in the approximation quality by roughly a factor of 10, which showcases that a more sophisticated approach using (uncoupled) separable kernels with non diagonal matrices, i.e. no componentwise approach, might prove to be beneficial, but we want to note that since the kernel K_2 is dependent on two parameters, the problem of optimal parameter selection is exacerbated. For higher dimensional problems this can be a limiting factor if the length p of the separable decomposition is too large.

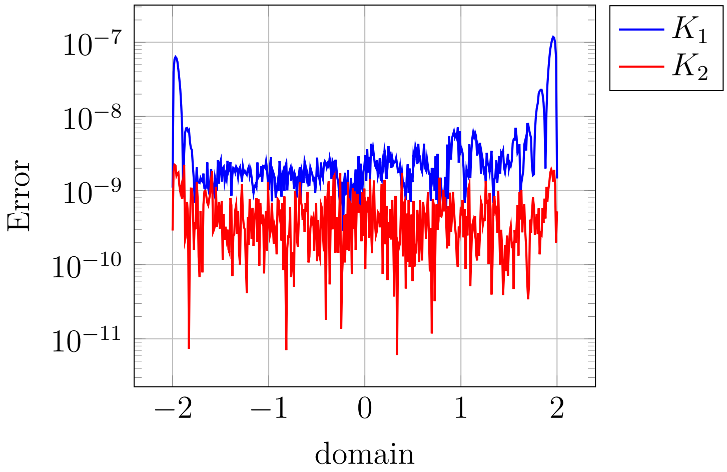

Figure 2.1: Pointwise error over the domain $\Omega = [-2, 2]$ measured in the Euclidean norm.

2.5 Error bounds on the best approximation

In this section we present bounds on the error between a function f and its best approximation in a closed subset $\mathcal{N}$. Similar to section 2.4, the core of the following error analysis was previously performed in [115]. However, this was strictly done in the setting of directional point evaluations and is hence generalized for the use of arbitrary bounded linear functionals that operate in the RKHS of the individual kernels. The error analysis relies heavily on the so called Power function that is given as follows.

Definition 2.5.1 (Power function)**.** Let $\mathcal{N} \subset \mathcal{H}_K$ be a closed subspace. Then the Power function $\mathcal{P}_{\mathcal{N},K} : \mathcal{H}'_K \to \mathbb{R}$ corresponding to $\mathcal{N}$ is given by

$$\mathcal{P}_{\mathcal{N},K}(\lambda) = \sup_{f \in \mathcal{H}_K \setminus \{0\}} \frac{\lambda\left(f - \Pi_{\mathcal{N}}(f)\right)}{\|f\|_{\mathcal{H}_K}} = \|\lambda \circ (\operatorname{id} - \Pi_{\mathcal{N}})\|_{\mathcal{H}'_K} . \tag{2.29}$$

In the last equality we made use of the fact that $\lambda \in \mathcal{H}'_K$ implies $\lambda \circ (\operatorname{id} - \Pi_{\mathcal{N}}) \in \mathcal{H}'_K$ which one can easily conclude as

$$|\lambda \circ (\operatorname{id} - \Pi_{\mathcal{N}})\,(f)| = |\lambda\left(f - \Pi_{\mathcal{N}}(f)\right)| \leq \|\lambda\|_{\mathcal{H}'_K} \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K} \leq \|\lambda\|_{\mathcal{H}'_K} \|f\|_{\mathcal{H}_K} \tag{2.30}$$

Due to (2.30) we can alternatively represent the Power function as follows.

Lemma 2.5.2. *Let $\mathcal{N} \subset \mathcal{H}_K$ be a closed subspace. Then for any $\lambda \in \mathcal{H}_K'$ we have*

$$\mathcal{P}_{\mathcal{N},K}(\lambda)^2 = \lambda^1\lambda^2 K_{\mathcal{N}^\perp}$$

where $K_{\mathcal{N}^\perp}$ is the reproducing kernel of $\mathcal{N}^\perp$. In particular, if $\mathcal{N} = \mathcal{N}(\Lambda)$ for some finite set $\Lambda \in \mathcal{H}_K'$ we have

$$\mathcal{P}_{\mathcal{N},K}(\lambda)^2 = \lambda^1\lambda^2 K - \lambda^1 S_\Lambda^2 K \left(S_\Lambda^1 S_\Lambda^2 K\right)^+ S_\Lambda^1 \lambda^2 K. \tag{2.31}$$

Proof. By Proposition 2.2.8 the Riesz representer of $\lambda \circ (\mathrm{id} - \Pi_{\mathcal{N}})$ is given by $(\lambda \circ (\mathrm{id} - \Pi_{\mathcal{N}}))^2 K$, i.e. the operator is applied to the second component. Furthermore, we have by Corollary 2.2.9

$$\mathcal{P}_{\mathcal{N},K}(\lambda)^2 = \|\lambda \circ (\mathrm{id} - \Pi_{\mathcal{N}})\|_{\mathcal{H}_K'}^2 = (\lambda \circ (\mathrm{id} - \Pi_{\mathcal{N}}))^1 (\lambda \circ (\mathrm{id} - \Pi_{\mathcal{N}}))^2 K = \lambda^1\lambda^2 K_{\mathcal{N}^\perp}, \tag{2.32}$$

where we made us of the identity $\mathrm{id} - \Pi_{\mathcal{N}} = \Pi_{\mathcal{N}^\perp}$ and Corollary 2.2.7. Let now $\mathcal{N} = \mathcal{N}(\Lambda)$ for some finite $\Lambda \in \mathcal{H}_K'$. We first note that due to the identity $\mathrm{id} - \Pi_{\mathcal{N}} = \Pi_{\mathcal{N}^\perp}$ we have

$$K_{\mathcal{N}^\perp} = K - K_{\mathcal{N}}$$

and thus by using the representation (2.18) for $K_{\mathcal{N}}$ we get

$$K_{\mathcal{N}^\perp} = K - S_\Lambda^2 K \left(S_\Lambda^1 S_\Lambda^2 K\right)^+ S_\Lambda^1 K.$$

Finally, the result follows by using (2.32).

□

By definition of the Power function it is clear that we can bound the error between any $f \in \mathcal{H}_K$ and its best approximation in the subspace $\mathcal{N}$ as follows.

Theorem 2.5.3 (General error bound on best approximation)**.** *Let $\mathcal{N} \subset \mathcal{H}_K$ be a closed subset and let $\mathcal{P}_{\mathcal{N},K}$ denote the Power function corresponding to $\mathcal{N}$. Then we have for any $\lambda \in \mathcal{H}_K'$*

$$|\lambda(f) - \lambda(\Pi_{\mathcal{N}}(f))| \leq \mathcal{P}_{\mathcal{N},K}(\lambda) \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K} \leq \mathcal{P}_{\mathcal{N},K}(\lambda) \|f\|_{\mathcal{H}_K}. \tag{2.33}$$

Proof. For any $f \in \mathcal{H}_K$ we have $\|f\|_{\mathcal{H}_K} \leq \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K} = \|\Pi_{\mathcal{N}^\perp}(f)\|_{\mathcal{H}_K}$, hence the supremum in (2.29) is actually realized over $\mathcal{N}^\perp$, i.e.

$$\mathcal{P}_{\mathcal{N},K}(\lambda) = \sup_{f\in\mathcal{H}_K\setminus\{0\}} \frac{\lambda\left(f-\Pi_{\mathcal{N}}(f)\right)}{\|f\|_{\mathcal{H}_K}} = \sup_{f\in\mathcal{H}_K\setminus\{0\}} \frac{\lambda\left(f-\Pi_{\mathcal{N}}(f)\right)}{\|f-\Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}}$$

which gives the desired result. □

As we have mentioned in previous sections, in the matrix-valued case even point-wise evaluation of elements in the RKHS can not be represented by a single functional. Hence, we want to introduce the concept of the so called Power function matrix, which better suits our setting.

Definition 2.5.4 (Power function Matrix)**.** Let $\Lambda \subset \mathcal{H}_K'$ and $\mathcal{N} \subset \mathcal{H}$ be a closed subspace. Then the Power function matrix corresponding to $\Lambda = \{\lambda_1, \dots, \lambda_p\}$ is given by

$$\boldsymbol{P}_{\mathcal{N},K}(\Lambda) = S_\Lambda^1 S_\Lambda^2 K_{\mathcal{N}^\perp} \in \mathbb{R}^{p\times p}$$

In the case where $\Lambda = \{\delta_x^{e_1}, \dots, \delta_x^{e_m}\}$ we may simply write

$$\boldsymbol{P}_{\mathcal{N},K}(x) = K_{\mathcal{N}^\perp}(x,x) = K(x,x) - K_{\mathcal{N}}(x,x)$$

instead.

It is easy to see that by definition the Power function matrix is positive (semi-) definite and for any functional that can be written as

$$\lambda = \begin{pmatrix}\lambda_1 & \cdots & \lambda_p\end{pmatrix}\alpha$$

for some $\alpha \in \mathbb{R}^p$ we have the following identity

$$\mathcal{P}_{\mathcal{N},K}(\lambda)^2 = \alpha^T \boldsymbol{P}_{\mathcal{N},K}(\Lambda)\alpha. \tag{2.34}$$

In particular for any directional point evaluation we have

$$\mathcal{P}_{\mathcal{N},K}(\delta_x^\alpha)^2 = \alpha^T \boldsymbol{P}_{\mathcal{N},K}(x)\alpha.$$

With the above and Theorem 2.5.3 we can now give estimates on the pointwise error $f(x) - \Pi_{\mathcal{N}}(f)(x)$ in various norms.

Corollary 2.5.5 (Bounds on the pointwise error)**.** *Let $\mathcal{N} \subset \mathcal{H}_K$ be a closed subset. Then the following bounds hold for all $x \in \Omega$.*

(a) $\|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_2^2 \le \lambda_{\max}(\boldsymbol{P}_{\mathcal{N},K}(x))\, \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2$

(b) $\|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_\infty^2 \leq \max \operatorname{diag}(\boldsymbol{P}_{\mathcal{N},K}(x)) \, \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2$

(c) $\|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_1^2 \leq \operatorname{tr}(\boldsymbol{P}_{\mathcal{N},K}(x)) \, \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2$

Proof. **(a)** For fixed $f \in \mathcal{H}_K$ and $x \in \Omega$ let

$$\alpha = \frac{f(x) - \Pi_{\mathcal{N}}(f)(x)}{\|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_2}.$$

By definition we have $\|\alpha\|_2 = 1$. For $\lambda := \delta_x^\alpha$ we thus have

$$\begin{aligned} \|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_2^2 = \lambda(f - \Pi_{\mathcal{N}}(f)) &\leq \alpha^T \boldsymbol{P}_{\mathcal{N},K}(x) \alpha \, \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2 \\ &\leq \lambda_{\max}(\boldsymbol{P}_{\mathcal{N},K}(x)) \, \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2 . \end{aligned}$$

(b) For the choice $\alpha = e_i$ and $\lambda := \delta_x^\alpha$ we get

$$(f(x) - \Pi_{\mathcal{N}}(f)(x))_i^2 = \lambda(f - \Pi_{\mathcal{N}}(f)) = (\boldsymbol{P}_{\mathcal{N},K}(x))_{ii} \, \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2 .$$

Hence the result directly follows from the definition of $\|\cdot\|_\infty$ on $\mathbb{R}^m$.

(c) This follows from the above results, since we take the sum over all $i = 1, \ldots, m$ in the above, resulting in the trace on the right hand side.

□

By definition $\mathcal{P}_{\mathcal{N},K}(\lambda)$ is the smallest value such that a bound of the form (2.33) holds for any $f \in \mathcal{H}_K$. Consequently any bound of the form

$$|\lambda(f) - \lambda(\Pi_{\mathcal{N}}(f))| \leq C(\lambda) \, \|f\|_{\mathcal{H}_K} ,$$

provides a bound $\mathcal{P}_{\mathcal{N},K}(\lambda) \leq C(\lambda)$ on the Power function. In the case of scalar-valued kernels, i.e. $m = 1$, and for functionals of the form $\delta_x \circ D_\alpha$ many such bounds have been derived over the years, [57, 59, 58, 34]. The most notably bounds where achieved for subspaces of the form $\mathcal{N} = \mathcal{N}(X)$ for some finite point set $X = \{x_1, \ldots, x_n\}$ and are for the most part formulated in terms of the so called fill distance defined as follows.

Definition 2.5.6 (Fill distance)**.** Let $\Omega \subset \mathbb{R}^d$ be bounded and let $X \subset \Omega$ be a finite set of pairwise distinct points. The fill distance $h_{X,\Omega}$ corresponding to $X \subset \Omega$ is given by

$$h_{X,\Omega} = \sup_{x \in \Omega} \min_{x' \in X} \|x - x'\| .$$

Depending on the smoothness of the scalar-valued kernel k a large selection of uniform bounds on $\mathcal{P}_{\mathcal{N}(X),k}$ in terms of $h_{X,\Omega}$ can be found in [108], all of which rely on estimates stemming from polynomial reproduction. For cases in which the kernel k stems from a scalar-valued radial basis function ϕ whose RKHS is isomorphic to Sobolev spaces, similar bounds were achieved using so called sampling inequalities [109, 81, 79, 80]. In the case of matrix-valued kernels, bounds are also available depending on the specific properties of the kernel. Under the assumption of sufficient smoothness similar bounds to the ones in [108] where achieved in [57, 59, 58] which also rely on the same polynomial reproduction. Extensions to divergence-free and curl-free kernels for $\Omega \subset \mathbb{R}^d$, with $d = 2, 3$ can be found in [34].

In the following we want to extend the results of [109] to the matrix-valued case. To this end, we make use of the sampling inequalities provided in [81], which rely on the fact that the domain $\Omega \subset \mathbb{R}^d$ satisfies the so called interior cone condition:

Definition 2.5.7 (Interior cone condition)**.** We say that $\Omega \subset \mathbb{R}^d$ satisfies an interior cone condition if there exists a radius $r > 0$ and an angle $\theta \in (0, \pi/2)$ such that for every $x \in \Omega$ one can find a unit vector $u \in \mathbb{R}^d$ such that the cone

$$C(x, u, \theta, r) := \{x + \lambda y \,|\, y \in \mathbb{R}^d, \|y\|_2 = 1,\, y^T u \geq \cos(\theta), \lambda \in [0, r]\}$$

is contained in Ω, i.e. we can always find a cone with radius r and angle θ and cone tip x which fits into Ω in its entirety.

Theorem 2.5.8. *Assume that $\Omega \subset \mathbb{R}^d$ is bounded and satisfies an interior cone condition. Let $\alpha \in \mathbb{N}_0^d$ be a multiindex and $k > |\alpha| + \frac{d}{2}$. Then there exists a constant C such that for every set $X \subset \Omega$ with sufficiently small fill distance $h := h_{X,\Omega}$ it holds*

$$\|D^\alpha u\|_{L_\infty(\Omega,\mathbb{R})} \leq C \left(h^{k-|\alpha|-d/2} \|f\|_{W^k(\Omega,\mathbb{R})} + h^{-|\alpha|} \|u(X)\|_\infty \right)$$

for all $u \in W^k(\Omega, \mathbb{R})$.

While the above is formulated for function mapping into the real numbers $\mathbb{R}$, the results still hold true when vector-valued outputs are considered. This follows immediately from the fact that by definition of the vector-valued Sobolev space, each individual function component lies in a corresponding real-valued Sobolev space and hence the above theorem is applicable. Therefore, we can combine the above result with Corollary 2.3.18 to obtain the following bounds on the pointwise error:

Theorem 2.5.9. *Suppose $\Omega \subset \mathbb{R}^d$ is bounded and satisfies an interior cone condition. Let $\Phi : \Omega \to \mathbb{R}^{m \times m}$ be a continuous p.d. function, such that the assumptions of Corollary 2.3.17 are satisfied for some $s > d/2$ and $B \in \mathbb{R}^{m \times m}$, and let $K(x, y) = \Phi(x - y)$ denote the p.d. kernel induced by Φ. Then there exists a constant $C > 0$ such that for any* *set $X \subset \Omega$ with sufficiently small fill distance $h := h_{X,\Omega}$ and for any multiindex $\alpha \in \mathbb{N}_0^d$ with $0 \leq |\alpha| < k - d/2$ we have*

$$\left\| D^\alpha \left(f(x) - \Pi_{\mathcal{N}(X)} f(x) \right) \right\|_2 \leq C \left\| f \right\|_{\mathcal{H}_K} h^{k-|\alpha|-d/2}$$

for all $f \in \mathcal{H}_K$.

Proof. By Corollary 2.3.18 we know that $\mathcal{H}_K = W^s(\Omega, \operatorname{range}(B))$ and the norms are equivalent, i.e. there exist constants c_1, c_2 such that

$$c_1 \left\| f \right\|_{W^s(\Omega, \operatorname{range}(B))} \leq \left\| f \right\|_{\mathcal{H}_K} \leq c_2 \left\| f \right\|_{W^s(\Omega, \operatorname{range}(B))}$$

for any $f \in \mathcal{H}_K$. Let $f \in \mathcal{H}_K$ be fixed and $u = f - \Pi_{\mathcal{N}(X)} f$. Then $u \in \mathcal{H}_K$ and in particular, $u \in W^s(\Omega, \operatorname{range}(B))$. Since the orthogonal projection operator coincides with the interpolation operator we further have $u(X) = 0$ and therefore it follows with Theorem 2.5.8 that

$$\begin{aligned} \left\| D^\alpha u \right\|_{L_\infty(\Omega, \operatorname{range}(B))} &\leq \tilde{C} h^{k-|\alpha|-d/2} \left\| u \right\|_{W^s(\Omega, \operatorname{range}(B))} \\ &\leq C h^{k-|\alpha|-d/2} \left\| u \right\|_{\mathcal{H}_K}, \end{aligned}$$

where $C = \tilde{C} c_2$. We note that the norm on $L_\infty(\Omega, \operatorname{range}(B))$ is given by

$$\left\| D^\alpha u \right\|_{L_\infty(\Omega, \operatorname{range}(B))} = \sup_{x \in \Omega} \left\| u(x) \right\|_2,$$

where $\left\| \cdot \right\|_2$ denotes the Euclidean norm. □

It now immediately follows that the above gives us an upper bound on the Power function (matrix):

Corollary 2.5.10. *Let the assumptions of Theorem 2.5.9 hold. Then for any $X \subset \Omega$ such that $h := h_{X,\Omega}$ is sufficiently small, we have for all functionals $D^\alpha \circ \delta_x^\beta$ with $\|\beta\| \leq 1$, $x \in \Omega$ and $\alpha \in \mathbb{N}_0^d, |\alpha| < k - d/2$*

$$\mathcal{P}(D^\alpha \circ \delta_x^\beta) \leq h^{k-|\alpha|-d/2} C.$$

Proof. By definition of the Power function it is the smallest value such that an inequality of the form

$$|D^\alpha \circ \delta_x^\beta (f - \Pi_{\mathcal{N}(X)} f)| \leq \mathcal{P}(D^\alpha \circ \delta_x^\beta) \left\| f \right\|_{\mathcal{H}_K}$$

holds for all $f \in \mathcal{H}_K$. And since

$$|D^\alpha \circ \delta_x^\beta (f - \Pi_{\mathcal{N}(X)} f)| = |D^\alpha (f - \Pi_{\mathcal{N}(X)} f)(x)^T \beta| \leq \left\| D^\alpha (f - \Pi_{\mathcal{N}(X)} f)(x) \right\|_2$$

the claim follows from Theorem 2.5.9. □

In [115] we have provided alternate bounds compared to the ones mentioned in the references above. These bounds apply to uncoupled kernels and function on the principle that due to the assumption of uncoupledness, the Power function matrix and hence the Power function as well, can be linked to the Power functions of the scalar-valued kernels of the uncoupled decomposition. This is possible since Theorem 2.4.9 guarantees that each scalar-valued kernel of the decomposition is p.d.. To this end, we first consider separable kernels of order 1 as these are uncoupled by definition.

Lemma 2.5.11 (Power function matrix for separable kernels of order 1)**.** *Let $K = kQ$ be an (uncoupled) separable p.d. kernel with decomposition $\{(k, Q)\}$. Let $\lambda \in \mathcal{H}_k'$, then $\alpha^T S_\lambda \in \mathcal{H}_K'$ for any $\alpha \in \mathbb{R}^m$, where S_λ denotes the Sampling operator applying λ componentwise. Let $\Lambda \subset \mathcal{H}_k$ and define the subspaces $\mathcal{N}_K(\Lambda)$, $\mathcal{N}_k(\Lambda)$ as in 2.2.12 and let $\mathcal{P}_{\mathcal{N}_K(\Lambda),K}$ and $\mathcal{P}_{\mathcal{N}_k(\Lambda),k}$ denote their corresponding Power functions. Then it holds*

$$\mathcal{P}_{\mathcal{N}_K(\Lambda),K}(\lambda \cdot \alpha)^2 = \mathcal{P}_{\mathcal{N}_k(\Lambda),k}(\lambda)^2 \alpha^T Q \alpha.$$

In particular this holds for any directional point evaluations δ_x^α.

Proof. By (2.34) this is equivalent to

$$\boldsymbol{P}_{\mathcal{N}_K(\Lambda),K}(S_\lambda) = \mathcal{P}_{\mathcal{N}_k(\Lambda),k}(\lambda)^2 Q,$$

which in turn is equivalent to

$$K_{\mathcal{N}_K(\Lambda)^\perp} = k_{\mathcal{N}_{k(\Lambda)^\perp}} Q$$

by definition of the Power function matrix. Using Corollary 2.2.7 and (2.18) we get

$$\begin{aligned} K_{\mathcal{N}_K(\Lambda)^\perp} &= K - K_{\mathcal{N}_K(\Lambda)} = K - S_\Lambda^2 K \left(S_\Lambda^1 S_\Lambda^2 K \right)^+ S_\Lambda^1 K \\ &= kQ - \left(S_\Lambda^1 k \otimes Q \right) \left(S_\Lambda^1 S_\Lambda^2 k \otimes Q \right)^+ \left(S_\Lambda^1 k \otimes Q \right) \\ &= kQ - \left(S_\Lambda^2 k \left(S_\Lambda^1 S_\Lambda^2 k \right)^+ S_\Lambda^1 k \right) \otimes Q \\ &= \left(k - S_\Lambda^2 k \left(S_\Lambda^1 S_\Lambda^2 k \right)^+ S_\Lambda^1 k \right) = k_{\mathcal{N}_{k(\Lambda)^\perp}} Q, \end{aligned}$$

where we made use of the fact that if one of the factors is scalar, the Kronecker product coincides with the regular (matrix) multiplication. □

This result can be extended in the case of uncoupled decompositions of higher order.

Lemma 2.5.12 (Power function matrix for uncoupled separable kernels of order p)**.** *Let K be an uncoupled separable p.d. kernel with decomposition $\{(k_i, Q_i)\}_{i=1}^p$. Let $K_i = k_i Q_i$. Furthermore let Λ be a finite collection of functionals such that $\mu \in \bigcap_{i=1}^p \mathcal{H}'_{K_i}$ for all $\mu \in \Lambda$. Let $\mathcal{N}_i(\Lambda) := \mathcal{N}_{K_i(\Lambda)}$ and $\mathcal{N}(\Lambda) = \mathcal{N}_K(\Lambda)$. Then it holds for any $\lambda \in \bigcap_{i=1}^p \mathcal{H}'_{K_i}$*

$$\mathcal{P}_{\mathcal{N}(\Lambda),K}(\lambda)^2 = \sum_{i=1}^p \mathcal{P}_{\mathcal{N}_i(\Lambda),K_i}(\lambda)^2.$$

Proof. By Theorem 2.4.10 we have $\mathcal{H}_K = \bigoplus_{i=1}^p \mathcal{H}_{K_i}$. Due to the assumption of uncoupledness it now also follows that $\mathcal{N}(\Lambda) = \bigoplus_{i=1}^p \mathcal{N}_i(\Lambda)$. In particular, both sides have the same reproducing kernel. The reproducing kernel of the left hand side is given via Corollary 2.2.7 as $K_{\mathcal{N}(\Lambda)}$. For the right hand side, we can combine this previous Corollary with the results of Corollary 2.3.4 to obtain

$$K_{\bigoplus_{i=1}^p \mathcal{N}_i(\Lambda)} = \sum_{i=1}^p K_{\mathcal{N}_i(\Lambda)}$$

which concludes the proof. □

Remark 2.5.13. If we consider functionals of the form $\lambda = \alpha^T S_{\Lambda'}$ for a finite set of functionals $\Lambda' \subset \mathcal{H}_K$ satisfying the assumptions of Lemma 2.5.12 we can generalize the above equality to the corresponding Power function matrices, i.e. we have

$$\boldsymbol{P}_{\mathcal{N}(\Lambda),K}(\Lambda') = \sum_{i=1}^p \boldsymbol{P}_{\mathcal{N}_i(\Lambda),K_i}(\Lambda').$$

If instead of the above we consider a separable kernel K with decomposition $\{(k_i, Q_i)\}_{i=1}^p$, it does not necessarily hold that each $\mathcal{N}_i(\Lambda) \subset \mathcal{N}(\Lambda)$ since we can recover $k_i Q_i \alpha$ from $K\beta$ for a suitable $\beta \in \mathbb{R}^m$. However, we still have $\mathcal{N}(\Lambda) \subset \bigoplus_{i=1}^p \mathcal{N}_i(\Lambda)$. Similar to the above we can make use of Corollary 2.2.7 to obtain

$$\sum_{i=1}^p \mathcal{P}_{\mathcal{N}_i(\Lambda),K_i}(\lambda)^2 \leq \mathcal{P}_{\mathcal{N}(\Lambda),K}(\lambda)^2,$$

i.e. we can only find a lower bound on the Power function in terms of the Power functions of the scalar-valued kernels. Similarly, we only have

$$\sum_{i=1}^{p} \boldsymbol{P}_{\mathcal{N}_i(\Lambda),K_i}(\lambda)^2 \preceq \boldsymbol{P}_{\mathcal{N}(\Lambda),K}(\lambda)^2,$$

for the Power function matrices.

The above does not make use of the special structure of the kernels $K_i = k_i Q_i$ or the separability of K in general. As such, the above inequality does hold for arbitrary matrix-valued kernels K_i and for any kernel K with $K = \sum_{i=1}^{p} K_i$. Furthermore, bounds of the form

$$\mathcal{P}_{\mathcal{N}(\Lambda),K}(\lambda)^2 \leq C \left(\sum_{i=1}^{p} \mathcal{P}_{\mathcal{N}_i(\Lambda),K_i}(\lambda)^2 \right)$$

for some constant $C \geq 1$ do not exist in general. To illustrate this, we consider the following example.

Example 2.5.14. Let $K_1, K_2 : \Omega \times \Omega \to \mathbb{R}$ be the polynomial kernels given by

$$K_1(x,y) = x^T y \qquad \text{and} \qquad K_2(x,y) = (x^T y)^2.$$

Then $\mathcal{H}_{K_1}$ is the space of multivariate polynomials of degree 1 and $\mathcal{H}_{K_2}$ is the space of multivariate polynomials of degree 2. In particular we have $\dim(\mathcal{H}_{K_1}) = d$ and $\dim(\mathcal{H}_{K_2}) = \frac{d(d+1)}{2}$. If we choose $X = \{x_i\}_{i=1}^{d(d+1)/2}$ such that $\{K_2(\cdot, x_i)\}_{i=1}^{d(d+1)/2}$ is linearly independent, then $\mathcal{N}_1(X)$ and $\mathcal{N}_2(X)$ are equal to the respective RKHS. Hence $\mathcal{P}_{\mathcal{N}_1(X),K_1}$ and $\mathcal{P}_{\mathcal{N}_2(X),K_2}$ vanish. However, for $K = K_1 + K_2$ the RKHS $\mathcal{H}_K$ contains the multivariate polynomials of both degree 1 and 2 and therefore $\dim(\mathcal{H}_K) = \frac{d(d+3)}{2}$. Consequently, $\mathcal{N}(X) \neq \mathcal{H}_K$ and therefore $\mathcal{P}_{\mathcal{N}(X),K}$ does not vanish.

3 Greedy Algorithms for Matrix-Valued Kernels

3.1 Greedy Kernel Algorithms

While we mostly dealt with a collection of (arbitrary) functionals $\lambda \in \mathcal{H}_K$ in the previous sections in order to construct an approximant of a function $f \in \mathcal{H}_K$, we now focus on (directional) point evaluation functionals and their respective sampling operators. Nonetheless, many of the following methods can be generalized to non point evaluation functionals as well.

As one might expect, the selection of interpolation points $X = \{x_1, \dots, x_n\} \in \Omega$ and consequently the choice if a suitable approximation subspace $\mathcal{N} = \mathcal{N}(X)$, is essential for the quality of the interpolant. Hence, the question arises how a suitable, or possibly optimal, choice can be made. Recalling standard result from approximation with polynomials in one dimension, it is known that the Chebyshev nodes provide the optimal set of interpolation points in the sense, that the Lebesgue constant has logarithmic growth, when intervals $[a, b]$ are considered. Similarly, in two dimensions, the minimal growth rate can be obtained by making use of the so called Padua points [14]. However, in higher dimension the problem of choosing optimal points for polynomial interpolation is in general not solved. Likewise, no optimal choice for the interpolation process using (matrix-valued) kernels is known to this author.

The method we present here, is to choose the interpolation points using a greedy algorithm. Its structure in the frame of kernel approximation are outlined in Algorithm 2 and works as follows: We assume to have a given finite sampling $\Omega_N \subset \Omega$ of the input space, an initial set of centers $X \subset \Omega$, this may be empty, a tolerance $\varepsilon > 0$ and an error indicator function E. Now, we iteratively select a point maximizing E, add it to the set of centers and compute the next approximant by interpolation on the small set of chosen points. This is repeated until the tolerance ε is reached. The following extends the known greedy selection rules for scalar-valued kernels [26, 85] and are named analogously. Analogous ideas are used in the case of polynomial interpolation to define the approximate Leja points and the approximate Fekete points [15].

```
Algorithm 2: General Kernel Greedy Algorithm
  Data: finite sampling of the input domain $\Omega_N \subset \Omega$, kernel $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$,
        target function $f : \Omega \to \mathbb{R}^m$, initial set of centers $X$, error indicator
        function $E$, tolerance $\varepsilon > 0$.
  Result: Set of interpolation points $X$.
  while $\max_{x \in \Omega_N} E(K, f, X, x) \geq \varepsilon$ do
      $x^* = \arg\max_{x \in \Omega_N} E(K, f, X, x)$;
      Extend set of interpolation points $X = X \cup \{x^*\}$;
  end
```

Depending on the choice of the error indicator function E we can divide Algorithm 2 into three major categories. To this end, we first recall the pointwise error bounds of Corollary 2.5.5: If $\mathcal{N} \subset \mathcal{H}_K$ is a closed subspace, then we have for any $x \in \Omega$

$$\|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_2^2 \leq \lambda_{\max}(\boldsymbol{P}_{\mathcal{N},K}(x)) \|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2 .$$

Using the greedy algorithm, we now try to minimize the pointwise error on the residual. Using the above bound, this can be achieved in atleast three different ways. First, by trying to minimize the left-hand side directly, second, by trying to minimize the error in the native space norm, and third, by trying to minimize the maximum eigenvalue of the Power function matrix. These three different types are called f–Greedy, f/P–Greedy and P–Greedy, respectively. The contents of the following subsections were previously published in [113] and were only slightly modified to better fit into the context of this thesis as well as to include more detailed proofs of mathematical statements.

3.1.1 f–Greedy

For the f–Greedy, the error indicator function E_f is given by

$$E_f(K, f, X, x) := \left\| f(x) - \Pi_{\mathcal{N}(X)}(f)(x) \right\|_2^2 . \tag{3.1}$$

One can see that the indicator relies heavily on the evaluation of the target function and hence $f(x)$ has to be available for all $x \in \Omega_N$, which might prove computationally expensive, if the target function f itself is expensive to evaluate. However, one can expect that the set of interpolation points selected with this error indicator should be well suited for approximating the target function. We note that as we iteratively progress through the steps in the greedy algorithm, the values of the indicator function E_f do not necessarily decrease, i.e. in general the inequality

$$E_f(K, f, X, x) \leq E_F(K, f, Y, x)$$

for $x \in \Omega_N$ and $Y \subset X$ is not satisfied. Nonetheless, we have $E_f(K, f, X, x) = 0$ for all $x \in X$ and hence, no point is selected twice. Furthermore, since $\Omega_N \subset \Omega$ is finite, the algorithm terminates after a finite number of steps.

3.1.2 f/P–Greedy

For the f/P–Greedy, the error indicator function is given by

$$E_{f/P}(K, f, X, x) = \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right)^T \boldsymbol{P}_{\mathcal{N}(X),K}(x)^+ \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right).$$

The motivation for the f/P–Greedy was the minimization of the error in the native space norm, i.e. the minimization of $\left\|f - \Pi_{\mathcal{N}(X)}(f)\right\|_{\mathcal{H}_K}^2$ in each iteration. The following lemma, which extends a result in [87] to the matrix-valued case, connects the chosen indicator to the error in the native space norm.

Lemma 3.1.1 (Local optimality of the f/P–Greedy selection rule)**.** *Let $K : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ be a positive definite matrix-valued kernel with native space $\mathcal{H}_K$. Furthermore, let $f \in \mathcal{H}_K$ and $X = \{x_1, \dots, x_n\} \subset \Omega$ be a finite set of pairwise distinct points. Then it holds for all $x \in \Omega$*

$$\begin{aligned}\left\|\Pi_{\mathcal{N}(X\cup\{x\})}(f)\right\|_{\mathcal{H}_K}^2 = &\left\|\Pi_{\mathcal{N}(X)}(f)\right\|_{\mathcal{H}_K}^2 \\ &+ \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right)^T \boldsymbol{P}_{\mathcal{N}(X),K}(x)^+ \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right).\end{aligned}$$

Proof. We first note that the columns of $K(\cdot, x)$ are functions in $\mathcal{H}_K$. Therefore, we have by Corollary 2.2.15 that the colums of $K(X, x)$ are in the range of $K(X, X)$. We now consider the block matrix

$$\begin{bmatrix} A & B \\ B^T & D \end{bmatrix} = \begin{bmatrix} K(x,x) & K(x,X) \\ K(X,x) & K(X,X) \end{bmatrix} = K(\{x\} \cup X, \{x\} \cup X)$$

This matrix now has a Schur-like decomposition

$$\begin{bmatrix} A & B \\ B^T & D \end{bmatrix} = \begin{bmatrix} I_m & BD^+ \\ 0 & I_{nm} \end{bmatrix} \begin{bmatrix} A - BD^+B^T & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I_m & 0 \\ D^+B^T & I_{nm} \end{bmatrix}$$

and therefore

$$K(\{x\}\cup X,\{x\}\cup X)^+ = \begin{bmatrix} I_m & 0 \\ -D^+B^T & I_{nm} \end{bmatrix} \begin{bmatrix} \left(A - BD^+B^T\right)^+ & 0 \\ 0 & D^+ \end{bmatrix} \begin{bmatrix} I_m & -BD^+ \\ 0 & I_{nm} \end{bmatrix}.$$

Please note, that in general, the Moore-Penrose-Pseudoinverse does not satisfy $(CE)^+ = E^+C^+$. However, in this particular case, the above holds since D and thus D^+ is symmetric and the outer matrices are invertible and satisfy

$$\begin{bmatrix} I_m & BD^+ \\ 0 & I_{nm} \end{bmatrix}^T = \begin{bmatrix} I_m & 0 \\ D^+B^T & I_{nm} \end{bmatrix}.$$

One can then easily verify the four conditions in the definition of the Moore-Penrose-Pseudoinverse. Due to Lemma 2.2.14 we have that

$$\Pi_{\mathcal{N}(X)}(f)(x) = K(x,X)K(X,X)^+f(X) = BD^+f(X)$$

and therefore

$$\begin{bmatrix} I_m & 0 \\ D^+B^T & I_{nm} \end{bmatrix} \begin{pmatrix} f(x) \\ f(X) \end{pmatrix} = \begin{pmatrix} f(x) - \Pi_{\mathcal{N}(X)}(f)(x) \\ f(X) \end{pmatrix}.$$

Furthermore, by definition of the Power function matrix we have

$$\boldsymbol{P}_{\mathcal{N}(X),K}(x) = K(x,x) - K(x,X)K(X,X)^+K(X,x) = A - BD^+B^T.$$

It now follows that

$$\begin{aligned} \left\|\Pi_{\mathcal{N}(X\cup\{x\})}(f)\right\|^2_{\mathcal{H}_K} &= \begin{pmatrix} f(x)^T & f(X)^T \end{pmatrix} K(\{x\}\cup X,\{x\}\cup X)^+ \begin{pmatrix} f(x) \\ f(X) \end{pmatrix} \\ &= \begin{pmatrix} f(x)^T & f(X)^T \end{pmatrix} \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}^+ \begin{pmatrix} f(x) \\ f(X) \end{pmatrix} \\ &= \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right)^T \boldsymbol{P}_{\mathcal{N}(X),K}(x)^+ \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right) \\ &\quad + f(X)^T D^+ f(X) \\ &= \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right)^T \boldsymbol{P}_{\mathcal{N}(X),K}(x)^+ \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right) \\ &\quad + \left\|\Pi_{\mathcal{N}(X)}(f)\right\|^2_{\mathcal{H}_K}. \end{aligned}$$

□

Using the previous result and the fact that the interpolant is the best approximation, see Theorem 2.2.13, we have

$$\begin{aligned}\left\|f - \Pi_{\mathcal{N}(X\cup\{x\})}(f)\right\|_{\mathcal{H}_K}^2 &= \|f\|_{\mathcal{H}_K}^2 - \left\|\Pi_{\mathcal{N}(X\cup\{x\})}(f)\right\|_{\mathcal{H}_K}^2 \\ &= \|f\|_{\mathcal{H}_K}^2 - \left\|\Pi_{\mathcal{N}(X)}(f)\right\|_{\mathcal{H}_K}^2 - E_{f/P}(K,X,x) \\ &= \left\|f - \Pi_{\mathcal{N}(X)}(f)\right\|_{\mathcal{H}_K}^2 - E_{f/P}(K,X,x).\end{aligned}$$

Therefore, maximizing $E_{f/P}(K,X,x)$ is equivalent to minimizing $\left\|f - \Pi_{\mathcal{N}(X\cup\{x\})}(f)\right\|_{\mathcal{H}_K}^2$.

Similar to the f–Greedy, the f/P–Greedy is dependent on the target function and generates a point set which is tailored to one specific target function. Moreover, the f/P–Greedy not only requires the evaluations of f on Ω_N but it furthermore requires the computation of

$$\left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right)^T \boldsymbol{P}_{\mathcal{N}(X),K}(x)^+ \left(f(x) - \Pi_{\mathcal{N}(X)}(f)(x)\right),$$

i.e. solving an m–dimensional linear system, for all $x \in \Omega_N$ in each iteration. Henceforth, it is more expensive than the f–Greedy. The indicator $E_{f/P}$ further satisfies that $E_{f/P}(K,f,X,x) = 0$ if $x \in X$ and it is also not monotonically decreasing in general.

The name f/P greedy stems from the scalar-valued case, where the indicator can equiavently be written as

$$E_{f/P}(K,f,X,x) = \frac{|f(x) - \Pi_{\mathcal{N}(X)}(f)(x)}{\boldsymbol{P}_{\mathcal{N}(X),K}(x)},$$

i.e. a fraction of function dependent quantity and the Power function (matrix).

In the case of separable kernels of order one, i.e. $K = k \cdot I$ for some scalar-valued kernel k, the above coincides withe the vectorial kernel orthognal greedy algorithm (VKOGA), which was introduced in [111].

3.1.3 P–Greedy

For the P–Greedy, the error indicator function E_P is given by

$$E_P(K,f,X,x) = E_P(K,X,x) := \lambda_{\max}(\boldsymbol{P}_{\mathcal{N}(X),K}(x)) = \left\|\boldsymbol{P}_{\mathcal{N}(X),K}(x)\right\|_2. \tag{3.2}$$

Unlike the f–Greedy and f/P–Greedy, the P–Greedy is independent of the target function f itself and therefore no (expensive) evaluation of the target function is required, which speeds up the point selection process for this indicator function. Furthermore, the selected points are not tailored towards a specific target function and thus one should expect a less accurate approximation of any target function, when compared to the f–Greedy and f/P–

Greedy. However, this instance of the greedy algorithm leads to point sets which provide good approximation for all functions in the native space $\mathcal{H}_K$. This generalizability cannot be seen in the f–Greedy and f/P-Greedy. Similar to the f–Greedy and f/P-Greedy, we have $E_P(K, X, x) = 0$ if $x \in X$. This follows immediately from Definition 2.5.4, since we have

$$\boldsymbol{P}_{\mathcal{N}(X),K}(x) = K(x, x) - K_{\mathcal{N}(X)}(x, x)$$

and K and $K_{\mathcal{N}(X)}$ coincide on $\mathcal{N}(X)$. Furthermore, as a direct consequence of Definition 2.5.4 and Corollary 2.2.7 we have $E_P(K, X, x) \leq E_P(K, Y, x)$ for all $x \in \Omega_N$ and $Y \subset X$, since in this case we have $\mathcal{N}(Y) \subset \mathcal{N}(X)$. In particular, the algorithm terminates after a finite number of steps and no point is chosen a second time.

3.1.4 Numerical investigation

We now investigate the different error indicators with respect to their effect on the quality of the approximation and the distribution of the selected points. To this end, we consider the unit disc segment $\Omega = \{x = (r\cos(\varphi), r\sin(\varphi)) \in \mathbb{R}^2 \mid (r, \varphi) \in \tilde{\Omega}\}$, where $\tilde{\Omega} = [0, 1] \times [\frac{1}{3}\pi, \frac{5}{3}\pi]$, and the target function $f = (f_i)_{i=1}^8 : \Omega \to \mathbb{R}^8$ given by

$$f_i(x) := \sum_{j=1}^{10} e^{-\lfloor (i+1)/2 \rfloor \|x - x_j\|^2}, \quad i = 1, \dots, 8,$$

with $x_1 = (0, 0)^T$ and $x_j = 0.1(\cos(\frac{j}{6}\pi), \sin(\frac{j}{6}\pi))^T, j = 2, \dots, 10$. For the kernel we use $K : \Omega \times \Omega \to \mathbb{R}^{8\times 8}$ given by a diagonal Gaussian with decaying widths

$$K_{i,j}(x, y) := \begin{cases} e^{-\lfloor (i+1)/2 \rfloor \|x - y\|^2}, & i = j \\ 0, & i \neq j. \end{cases}$$

One easily sees that $f(x) = K(x, Y)\mathbf{1}$ where $Y = \{x_1, \dots, x_{10}\}$ and $\mathbf{1} \in \mathbb{R}^{80}$ is the vector containing only ones. Therefore, we have $\|f\|_{\mathcal{H}_K} = \mathbf{1}^T K(Y, Y)\mathbf{1} \approx 768.295$. For the greedy algorithm we choose Ω_N by transforming 50×50 uniformly distributed points in $\tilde{\Omega}$ to Euclidean coordinates, which results in $|\Omega_N| = 2451$ points. For the tolerance we choose $\varepsilon = 10^{-7}$. The sets of interpolation points generated by different greedy algorithms are denoted by X_{type}, where type $\in \{f, f/P, P\}$ and each type corresponds to the respective error indicator E_{type}.

In Figure 3.1 the decay of the error indicators, i.e. the maximum of E_{type} over the training set Ω_N, and the maximum training error measured in the Euclidean norm are depicted for increasing size of the set X_{type}. As we can see, the P–Greedy algorithm

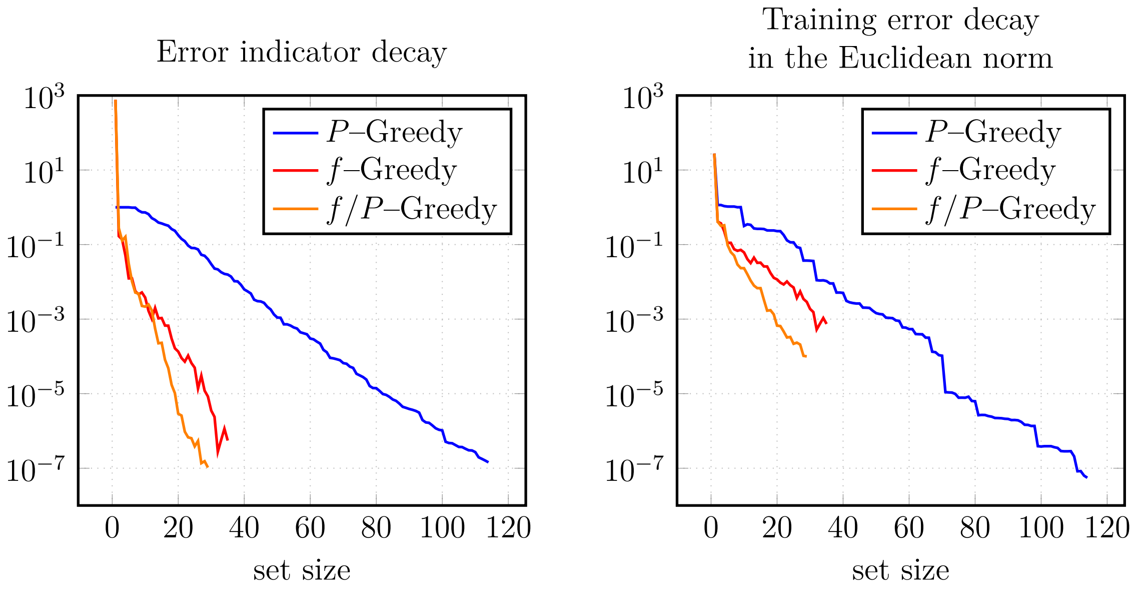

Figure 3.1: Error indicator decay (left) and maximum training error decay in the Euclidean norm (right) for increasing set size.

terminates after 114, whereas the f–Greedy terminates after 35 and the f/P–Greedy terminates after 29 iterations. This slower decay for E_P is likely caused by the narrow Gaussians which model the last target function components. In Figure 3.2 the maximum test error in the Euclidean norm on the test set Ω_{Test} generated by transforming 100×100 uniformly distributed points in $\tilde{\Omega}$ and the error in the native space norm are shown. As we have conjectured in the subsections 3.1.1 – 3.1.3 both f– and f/P–Greedy generate sets which are tailored towards the target function f which leads to a better approximation when compared to the sets of the same size generated by the P–Greedy. For example, to reach a test error of 10^{-3} in the Euclidean norm, the sets generated by f–, f/P– and P–Greedy have sizes 34, 19 and 55, respectively. However, this benefit in terms of approximation quality is counteracted by a poorer conditioning of the linear system

$$K(X_{\text{type}}, X_{\text{type}})\alpha = f(X_{\text{type}})$$

which has to be solved in order to construct the approximant. The condition of the linear system for the respective type is depicted in Figure 3.3. And we can see that the condition for the f/P–Greedy has the sharpest increase, where we already have a condition number of $\approx 1,4 \cdot 10^{14}$ for $|X_{f/P}| = 10$. In contrast the f– and P–Greedy only lead to condition numbers of $\approx 3,8 \cdot 10^{10}$ and $2,9 \cdot 10^{11}$, respectively. This rapid increase for the condition number when using the f/P–Greedy is tightly connected to the distribution of the selected points, which are depicted in Figure 3.4. We can see that the points selected by the f/P–Greedy algorithm are not well distributed over Ω and tend to cluster next to each other,

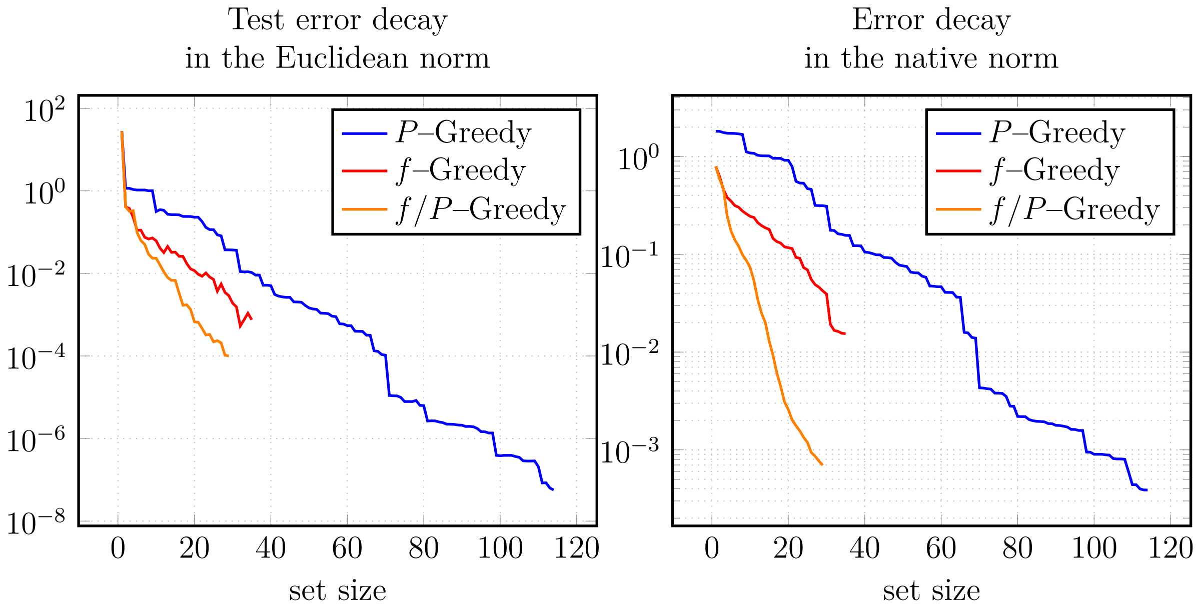

Figure 3.2: Test error decay in the Euclidean norm (left) and in the Hilbert space norm (right) for increasing set size

which leads to a bad conditioning of the respective linear system.

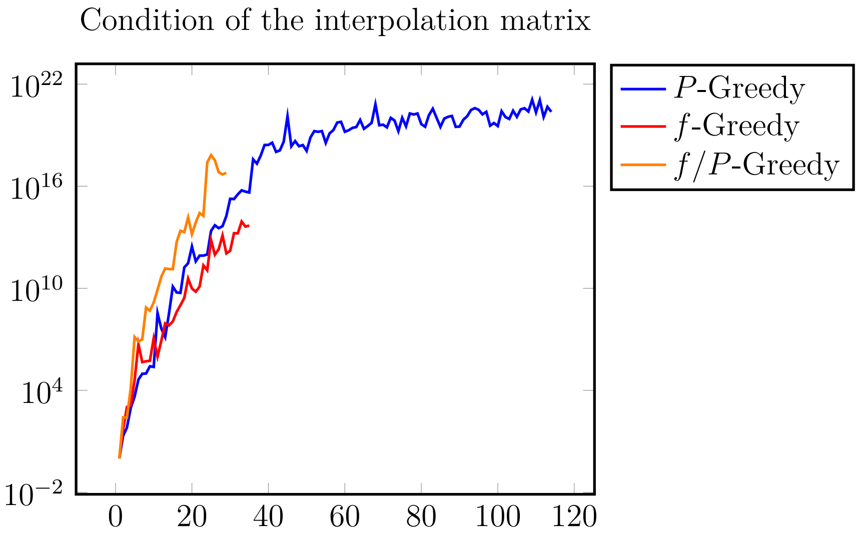

Figure 3.3: Condition number of the linear system required in the approximation process for increasing set size.

3.2 Matrix P–Greedy Variants

In this section we will take a closer look at the P–Greedy algorithm, as it is function independent and therefore applicable to a wider variety of problems. Furthermore, it is

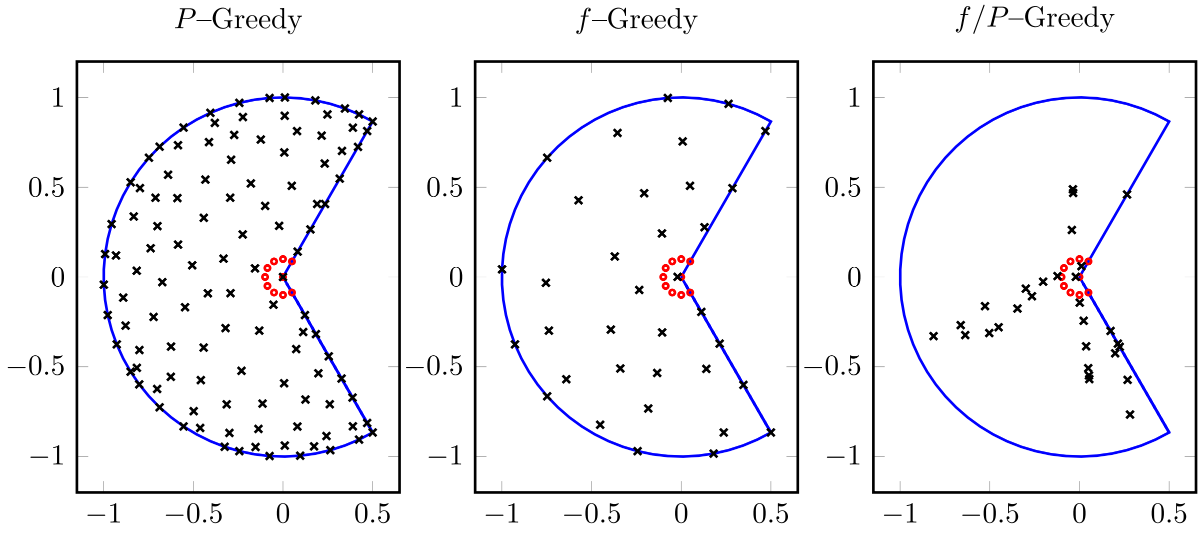

Figure 3.4: Point distribution of the sets X_P, X_f and $X_{f/P}$ (black) and the centers making up the target function (red).

computationally less demanding than the f– and f/P–Greedy, if the evaluation of f is expensive, or if the target function is high dimensional and hence expensive (pseudo) inverses have to be computed. For the same reason we present further variants of the P–Greedy, which rely on a different indicator function than the one presented in (3.2), as it requires the solution of an eigenvalue problem for each point in the training set Ω_N. Moreover, we take a closer look at different extension strategies for enriching the approximation space, where we do not necessarily include all columns of $K(\cdot, x)$ if the point x is selected during the greedy iteration. A first version of this and the two subsequent subsections where first published in [114]. However, here the contents were significantly extended by including new mathematical analysis, more detailed proofs and further numerical experiments.

The basic principle for these P–Greedy variants is outlined in Algorithm 3.

Algorithm 3: Matrix P-greedy Algorithm

```
Data: finite sampling of the input domain $\Omega_N \subset \Omega$, kernel $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$,
      initial approximation space $\mathcal{N}$, error indicator function $E$, tolerance $\varepsilon > 0$,
      space extension routine “extend”.
Result: Approximation space $\mathcal{N}$
while $\max_{x \in \Omega_N} E(\boldsymbol{P}_{\mathcal{N}}(x)) \geq \varepsilon$ do
    $x^* = \arg\max_{x \in \Omega_N} E(\boldsymbol{P}_{\mathcal{N},K}(x))$;
    $\mathcal{N} = \text{extend}(\mathcal{N}, K(\cdot, x^*))$;
end
```

Similar to the indicator functions that where introduced in Section 3.1, we again make use of Corollary 2.5.5 to derive further indicator functions. Recalling the results of Corollary 2.5.5, we have the bounds

$$\begin{aligned}\|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_2^2 &\leq \lambda_{\max}(\boldsymbol{P}_{\mathcal{N},K}(x))\,\|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2\\ \|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_\infty^2 &\leq \max\operatorname{diag}(\boldsymbol{P}_{\mathcal{N},K}(x))\,\|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2\\ \|f(x) - \Pi_{\mathcal{N}}(f)(x)\|_1^2 &\leq \operatorname{tr}(\boldsymbol{P}_{\mathcal{N},K}(x))\,\|f - \Pi_{\mathcal{N}}(f)\|_{\mathcal{H}_K}^2,\end{aligned}$$

which naturally lead to the following three indicator functions $E_i : \mathbb{R}^{m\times m} \to \mathbb{R}$, $i \in \{1, 2, \infty\}$ given by

$$E_1(B) \coloneqq \frac{1}{m}\operatorname{tr}(B), \qquad E_2(B) \coloneqq \lambda_{\max}(B), \qquad E_\infty(B) \coloneqq \max\operatorname{diag}(B). \tag{3.3}$$

Here E_2 is just the indicator E_P used in the previous section. However, we included it again with the change of notation to make the distinction between the three variants easier.

For the extension routine we propose

$$\begin{aligned}\text{extend}_{\text{full}}(\mathcal{N}, K(\cdot, x)) &:= \mathcal{N} + \operatorname{colspan}(K(\cdot, x)),\\ \text{extend}_{\text{eig}}(\mathcal{N}, K(\cdot, x)) &:= \mathcal{N} + \operatorname{span}(K(\cdot, x)\alpha_{\max}),\\ \text{extend}_{\text{diag}}(\mathcal{N}, K(\cdot, x)) &:= \mathcal{N} + \operatorname{span}(K(\cdot, x)e_{\max}),\end{aligned}$$

where $\alpha_{\max}$ denotes an eigenvector to the largest eigenvalue and $e_{\max}$ the standard basis vector to the largest diagonal value of $\boldsymbol{P}_{\mathcal{N},K}(x)$, respectively.

As the name suggests, extend$_{\text{full}}$ enriches the approximation space $\mathcal{N}$ with all columns of $K(\cdot, x)$. In other words, the dimension of $\mathcal{N}$ increases by m in every iteration, which might lead to a rapid increase in the overall dimension of the approximation space, if m is large. However, all components of the target function value $f(x)$ will be used in the subsequent approximation process. In contrast, extend$_{\text{eig}}$ and extend$_{\text{diag}}$ increase the approximation space dimension by 1 in every iteration. Therefore, one might expect the final approximation space $\mathcal{N}$ to be smaller, when compared to the full extension routine, as potentially unnecessary columns of $K(\cdot, x)$ are not included. Nevertheless, we might require a larger number of individual target function evaluations in the approximation process, which is in turn expensive if the evaluation of f is computationally demanding. We will consider all possible combinations of the above indicator and extension routines and shall denote them via greedy$_{i,\text{type}}$ with $i \in \{1, 2, \infty\}$ and type $\in \{\text{full}, \text{eig}, \text{diag}\}$.

Similar to what was outlined in Section 3.1, the above three methods coincide in the scalar-valued case, i.e. $m = 1$ and represent the well known P–Greedy algorithm [26]. An efficient way to implement the different P–Greedy variants is available by making use of the so called Newton basis [68]

Definition 3.2.1 (Newton basis)**.** Let $\Lambda = \{\lambda_1, \dots, \lambda_n\} \subset \mathcal{H}_K'$ be a set of linearly independent functionals. Furthermore, let $LL^T = S_\Lambda^1 S_\Lambda^2 K$ be a Cholesky factorization and denote as $l_1, \dots, l_n$ the columns of L^{-T}. Then the Newton basis $\{v_1, \dots, v_n\}$ of $\mathcal{N}(\Lambda)$ is given by

$$v_i \coloneqq S_\Lambda^2 K(\cdot, \cdot) l_i, \qquad \text{for } i = 1, \dots, n$$

The Newton basis now has the following properties

Proposition 3.2.2 (Properties of the Newton basis)**.** *Let $\Lambda = \{\lambda_1, \dots, \lambda_n\}$ be a set of linearly independent functionals and $\{v_1, \dots, v_n\}$ the corresponding Newton basis of $\mathcal{N}(\Lambda)$. Then it holds*

(a) $\lambda_j(v_i) = 0$ *for all $j < i$.*

(b) *The set $\{v_1, \dots, v_n\}$ is an orthonormal basis and can be generated by applying the Gram-Schmidt orthonormalization procedure to the set $\{\lambda_1^2 K, \dots, \lambda_n^2 K\}$.*

Proof. We only need to show the second property, as the first follows from the fact that $\lambda_j(v_i) = \langle \lambda_j^2 K, v_i \rangle_{\mathcal{H}_K}$. For the second property it is sufficient to note, that performing the Gram-Schmidt orthonormalization is equivalent to computing the Cholesky factorization of

$$LL^T = S_\Lambda^1 S_\Lambda^2 K = \left(\langle \lambda_j^2 K, \lambda_i^2 K \rangle_{\mathcal{H}_K} \right)_{i,j=1}^n .$$

The basis generated by the Gram-Schmidt procedure can then be expressed via

$$v_i = S_\Lambda^2 K L^{-T} e_i$$

which is exactly the definition of the Newton basis. □

The name Newton basis stems from the fact that the first property in Proposition 3.2.2 mimics the classical Newton basis used in polynomial interpolation. The second property allows us to efficiently update the Newton basis, if a further functional λ_{n+1} is added to Λ. This follows from the fact that the Cholesky factorization of $S_{\Lambda\cup\{\lambda\}}^1 S_{\Lambda\cup\{\lambda\}}^2 K$ can be computed by updating the Cholesky factorization of $S_\Lambda^1 S_\Lambda^2 K$. This updatability carries over to the Power function (matrix) as follows.

Theorem 3.2.3 (Representation of the Power function (matrix) via the Newton basis)**.** *Let $\Lambda = \{\lambda_1, \dots, \lambda_n\}$ be a collection of linearly independent functionals and $\{v_1, \dots, v_n\}$ be the corresponding Newton basis. Let $\Gamma = \{\gamma_1, \dots, \gamma_p\}$ be a second set of functionals. Then the Power function matrix is given by*

$$\boldsymbol{P}_{\mathcal{N}(\Lambda),K}(\Gamma) = S_\Gamma^1 S_\Gamma^2 K - \sum_{i=1}^n S_\Gamma(v_i) S_\Gamma(v_i)^T.$$

Proof. Let $K_{\mathcal{N}(\Lambda)}$ be the reproducing kernel of $\mathcal{N}(\Lambda)$. By definition of the Power function matrix we have

$$\boldsymbol{P}_{\mathcal{N}(\Lambda),K}(\Gamma) = S_\Gamma^1 S_\Gamma^2 K - S_\Gamma^1 S_\Gamma^2 K_{\mathcal{N}(\Lambda)}.$$

Hence, it is sufficient to show that

$$K_{\mathcal{N}(\Lambda)}(x,y) = \sum_{i=1}^n v_i(x) v_i(y)^T. \tag{3.4}$$

However, this is equivalent to showing that the right hand side of (3.4) satisfies the reproducing property (2.6) on $\mathcal{N}(\Lambda)$. Since $v_1, \dots, v_n$ are a basis of $\mathcal{N}(X)$ we have for any $x \in \Omega$ and $\alpha \in \mathbb{R}^m$

$$\sum_{i=1}^n v_i v_i(x)^T \alpha \in \mathcal{N}(\Lambda).$$

Let $f \in \mathcal{N}(\Lambda)$, then

$$f = \sum_{j=1}^n \beta_j v_j$$

for some $\beta \in \mathbb{R}^n$. Since the Newton basis is orthonormal it follows that

$$\left\langle f, \sum_{i=1}^n v_i v_i(x)^T \alpha \right\rangle = \sum_{i,j=1}^n \langle v_j, v_i \rangle \beta_j v_i(x)^T \alpha = \sum_{i=1}^n \beta_i v_i(x)^T \alpha = f(x)^T \alpha.$$

□

The Power function (matrix) is therefore updatable since by the above we have for any λ_{n+1} that

$$\boldsymbol{P}_{\mathcal{N}(\Lambda \cup \{\lambda_{n+1}\})}(\Gamma) = \boldsymbol{P}_{\mathcal{N}(\Lambda)}(\Gamma) - S_\Gamma(v_{n+1}) S_\Gamma(v_{n+1})^T. \tag{3.5}$$

Remark 3.2.4. In Theorem 3.2.3 we only used the fact that the Newton basis is orthonormal. Hence, the theorem can be generalized if we replace $\mathcal{N}(\Lambda)$ with an arbitrary subspace $\mathcal{N}$ which has an orthonormal basis.

As it was shown in [83], the P–Greedy algorithm generates a sequence of sets which provide quasi-optimal approximation rates in the case of scalar-valued kernels. This was further extended in [84], where a wider variety of functionals, not just point evaluations, where considered. Both of the two previous results are based on the work presented in [27, 12], in which the approximation quality of spaces selected by greedy algorithms in the context of reduced basis where considered. Likewise, we also make use of these foundations to show how quasi optimal convergence rates can be achieved when the greedy variants greedy$_{i,\text{type}}$ are used in the non scalar-valued case, i.e. $m > 1$. Since the work in [27] is not framed in the context of kernel based approximations, we similarly use a more general approach which also extends the result for multi-dimensional greedy extensions in different fields.

3.3 Convergence Rates for k–dimensional Greedy Space Extensions

In order to apply the techniques employed in [27] we first have to extend to notion of a weak greedy algorithm in the context a multi-dimensional space extension procedure. For this purpose, we assume for this section that $\mathcal{H}$ is a Hilbert space and $\mathcal{F} \subset \mathcal{H}$ is a compact subset, which we want to approximate by a suitable subspace spanned by elements $f \in \mathcal{F}$. A more abstract version of the greedy algorithm is outlined in Algorithm 4.

Algorithm 4: k–dimensional greedy space extension.

```
Data: $\mathcal{F} \subset \mathcal{H}$ compact set, initial approximation space $V_0 \subset \mathcal{H}$, indicator function
      $E$, maximum number of iterations $n_{\max} \in \mathbb{N}$ and greedy extension
      dimension $k \in \mathbb{N}$.
Result: Approximation space $V_{n_{\max}}$
Initialize $n = 1$;
while $n \leq n_{\max}$ do
    $W_n := \arg \max_{\substack{W \subset \text{span}\{\mathcal{F}\} \\ \dim(W)=k}} E(W, V_{n-1})$;
    $V_n := V_{n-1} + W_n$;
end
```

The only requirement we now have, is that the indicator function E produces a so called weak greedy algorithm:

Definition 3.3.1 (Weak Greedy Algorithm)**.** Let $\{W_n\}_{n=1}^{n_{\max}}$ and $\{V_n\}_{n=0}^{n_{\max}}$ be the sequence of subspaces chosen as outlined Algorithm 4. The algorithm is called a weak greedy algorithm if there exist a constant $0 < \gamma \leq 1$ such that

$$\max_{f\in W_n} \left\| f - \Pi_{V_{n-1}} f \right\|_{\mathcal{H}} = \max_{f\in W_n} \left\| \Pi_{V_{n-1}^{\perp}} f \right\|_{\mathcal{H}} \geq \gamma \max_{f\in \mathcal{F}} \left\| f - \Pi_{V_{n-1}} f \right\|_{\mathcal{H}} = \gamma \max_{f\in \mathcal{F}} \left\| \Pi_{V_{n-1}^{\perp}} f \right\|_{\mathcal{H}} \tag{3.6}$$

In other words, the algorithm is called a weak greedy algorithm if in every iteration, the subspace chosen to enrich the approximation space contains an element such that the ratio of the error of the best approximation in said subspace, and the maximum best approximation error over all of $\mathcal{F}$ is bounded by a constant strictly bigger than 0. It is clear that whether or not Algorithm 4 is a weak greedy algorithm, soley depends on the choice of the error indicator function E. For example the choice

$$E(W, V) = \max_{f\in W} \|f - \Pi_V f\|_{\mathcal{H}}$$

results in a weak greedy algorithm with $\gamma = 1$. Albeit, in the case of $\gamma = 1$, one usually refers to the algorithm as a so called strong greedy algorithm.

We now want to compare the quality of the approximation in the subspaces V_n to the best possible approximation spaces of the same dimension. This is quantified in terms of the Kolmogorov n-width $d_n(\mathcal{F})$ which is given by

$$d_n(\mathcal{F}) = \inf_{\substack{V\subset\mathcal{H} \\ \dim(V)=n}} \sup_{f\in\mathcal{F}} \|f - \Pi_V f\|_{\mathcal{H}} \,. \tag{3.7}$$

Moreover, let $\mathcal{H}_m$ denote an m-dimensional Kolmogorov subspace such that

$$\sup_{f\in\mathcal{F}} \|f - \Pi_{\mathcal{H}_m} f\|_{\mathcal{H}} = d_m(\mathcal{F}).$$

Analogous to (3.7) we can quantify the approximation quality of the subspaces V_n generated by a weak greedy algorithm via

$$\sigma_n(\mathcal{F}) \coloneqq \max_{f\in\mathcal{F}} \|f - \Pi_{V_n} f\|_{\mathcal{H}} \,.$$

In order to relate the above two quantities, we need the following lemma.

Lemma 3.3.2 (Lemma 2.1 from [27])**.** *Let $G \in \mathbb{R}^{K\times K}$ be a lower triangular matrix with rows $\mathbf{g}_1, \ldots, \mathbf{g}_K \in \mathbb{R}^K$. Let $W \subset \mathbb{R}^K$ be an m–dimensional subspace and Π_W the* *orthogonal projection from* $\mathbb{R}^K$ *onto* W*. Then it holds*

$$\prod_{i=1}^{K} \mathbf{g}_{i,i}^2 \leq \left(\frac{1}{m}\sum_{i=1}^{K} \|\Pi_W \mathbf{g}_i\|_2^2\right)^m \left(\frac{1}{K-m}\sum_{i=1}^{K} \|\mathbf{g}_i - \Pi_W \mathbf{g}_i\|_2^2\right)^{K-m}.$$

With this we are able to extend the results for the scalar-valued P–Greedy algorithm. However, we have to be careful, as $\sigma_n(\mathcal{F})$ represent the approximation quality of the space V_n which is $n \cdot k$–dimensional, whereas $d_n(\mathcal{F})$ represents the approximation quality of the best n–dimensional subspace. Here we implicitly assume that the initial subspace V_0 is 0–dimensional, i.e. $V_0 = \{0\}$.

Theorem 3.3.3. *Let* $\{W_n\}_{n\in\mathbb{N}}$ *and* $\{V_n\}_{n\in\mathbb{N}}$ *be the sequence of spaces generated by a weak greedy algorithm with constant* $0 < \gamma \leq 1$*. Furthermore, let* $\sigma_n = \sigma_n(\mathcal{F})$ *and* $d_n = d_n(\mathcal{F})$*, then we have for any* $N \geq 0$*,* $K \geq 1$ *and* $1 \leq k < K$*:*

$$\prod_{i=0}^{K-1} \sigma_{N+i}^2 \leq \gamma^{-2K} \left(\frac{K}{m}\right)^m \left(\frac{K}{K-m}\right)^{K-m} \sigma_N^{2m} d_m^{2K-2m}.$$

Proof. For $n \in \mathbb{N}$, let

$$g_n^1 := \arg\max_{f\in W_n} \left\| f - \Pi_{V_{n-1}} f \right\|_{\mathcal{H}}$$

and $g_n^2, \dots, g_n^k$ such that $g_n^1, \dots, g_n^k$ form a basis of W_n. Let now $\{\hat{g}_n^i \mid n \in \mathbb{N}, i = 1, \dots, k\}$ denote the orthonormal system generated by applying the Gram-Schmidt orthonormalization procedure to $\{g_1^1, g_1^2, \dots, g_1^k, g_2^1, \dots\}$. We now define the infinite lower-triangular matrix A by

$$A := (a_{ij})_{i,j=1}^{\infty}, m \qquad a_{ij} = \langle g_i^1, \hat{g}_j^1 \rangle_{\mathcal{H}}.$$

By construction we have

$$\hat{g}_n^1 = \frac{1}{\left\| g_n^1 - \Pi_{V_{n-1} g_n^1} \right\|_{\mathcal{H}}} \left(g_n^1 - \Pi_{V_{n-1}} g_n^1 \right)$$

and therefore

$$a_{nn} = \langle g_n^1, \hat{g}_n^1 \rangle_{\mathcal{H}} = \left\| g_n^1 - \Pi_{V_{n-1}} g_n^1 \right\|_{\mathcal{H}} = \max_{f\in W_n} \left\| g_n^1 - \Pi_{V_{n-1}} g_n^1 \right\|_{\mathcal{H}}.$$

In particular, we have $\gamma\sigma_{n-1} \leq a_{nn} \leq \sigma_{n-1}$. Furthermore, we have for any $M \geq n$ that

$$\begin{aligned}\sum_{j=n}^{M} a_{Mj}^2 &= \sum_{j=n}^{M} \langle g_M^1, \hat{g}_j^1 \rangle_{\mathcal{H}}^2 \leq \sum_{j=n}^{M} \sum_{i=1}^{m} \langle g_M^1, \hat{g}_j^i \rangle_{\mathcal{H}}^2 \\ &= \left\| g_M^1 - \Pi_{V_{n-1}} g_M^1 \right\|_{\mathcal{H}}^2 \leq \max_{f \in \mathcal{F}} \left\| f - \Pi_{V_{n-1}} f \right\|_{\mathcal{H}}^2 = \sigma_{n-1}^2. \qquad (3.8)\end{aligned}$$

Let $G \in \mathbb{R}^{K \times K}$ be the lower triangular matrix which is formed by taking the submatrix of A with column and row indices $N+1, \dots, N+K$, i.e. $G = (a_{ij})_{i,j=N+1}^{N+K}$. Each row $\mathbf{g}_i$ of G is now the restriction of g_{N+i}^1 to the coordinates $N+1, \cdot, N+K$. From (3.8) it now follows, that

$$\|\mathbf{g}_i\|^2 = \sum_{j=N+1}^{N+i} a_{N+i,j}^2 \leq \sigma_N^2,$$

where we used that the $i+1$–th to K–th component of $\mathbf{g}_i$ are zero, since G is lower triangular. Let $\mathcal{H}_m$ denote the m-dimensional Kolmogorov subspace of $\mathcal{H}$. Then we have $\operatorname{dist}(g_{N+i}^1, \mathcal{H}_m) \leq d_m$ for $i = 1, \dots, K$. Let $\tilde{W}$ be the linear space which is the restriction of $\mathcal{H}_m$ onto the coordinates $N+1, \dots, N+K$ and let $\tilde{W} \subset W$ be the space for the coordinates $N+1, \dots, N+K$. Then we have

$$\|\Pi_{\tilde{W}} \mathbf{g}_i\|_2 \leq \|\mathbf{g}_i\|_2 \leq \sigma_N$$

and

$$\|\mathbf{g}_i - \Pi_{\tilde{W}} \mathbf{g}_i\|_2 \leq \|\mathbf{g}_i - \Pi_W \mathbf{g}_i\|_2 = \operatorname{dist}(\mathbf{g}_i, \tilde{W}) \leq \operatorname{dist}(g_{N+i}^1, \mathcal{H}_m) \leq d_m, \quad i = 1, \dots, K.$$

Using Lemma 3.3.2 we now get

$$\gamma^{2K} \prod_{i=0}^{K-1} \sigma_{N+i}^2 \leq \left(\frac{1}{m} \sum_{i=1}^{K} \sigma_N \right)^m \left(\frac{1}{K-m} \sum_{i=1}^{K} d_m \right)^{K-m}$$

which is equivalent to

$$\prod_{i=0}^{K-1} \sigma_{N+i}^2 \leq \gamma^{-2K} \left(\frac{K}{m} \right)^m \left(\frac{K}{K-m} \right)^{K-m} \sigma_N^{2m} d_m^{2K-2m}.$$

□

From [27] we take the following corollary. However, we include a proof for the sake of completeness as some of the steps were unclear to us without further clarification.

Corollary 3.3.4. *Let $\mathcal{F} \subset \mathcal{H}$ be compact such that $d_n(\mathcal{F}) \leq C_0 n^{-\alpha}$, then*

$$\sigma_n(\mathcal{F}) \leq C_1 n^{-\alpha}, \qquad \textit{where } C_1 := 2^{5\alpha+1}\gamma^{-2}C_0.$$

Proof. First, we can assume without loss of generality that $\mathcal{F}$ is scaled in such a way that

$$\sup_{f\in\mathcal{F}} \|f\|_{\mathcal{H}} = \sigma_1 = 1.$$

Furthermore, by definition we have that $\{\sigma_n\}_{n\in\mathbb{N}}$ is a monotonically decreasing sequence. Using $N = n+1$, $K = n$ and any $1 \leq m < n$ in Theorem 3.3.3 we thus get

$$\sigma_{2n}^{2n} \leq \prod_{i=1}^{K} \sigma_{n+i}^2 \leq \gamma^{-2n} \left(\frac{n}{m}\right)^m \left(\frac{n}{n-m}\right)^{n-m} \sigma_{n+1}^{2m} d_m^{2n-2m}.$$

Making use of $\sigma_{n+1} \leq \sigma_n$ we can solve for σ_{2n} and get

$$\sigma_{2n} \leq \gamma^{-1} \left(\left(\frac{n}{m}\right)^{\frac{m}{n}} \left(\frac{n}{n-m}\right)^{\frac{n-m}{n}} \right)^{\frac{1}{2}} \sigma_n^{\frac{m}{n}} d_m^{\frac{n-m}{n}}. \tag{3.9}$$

Now, let $h : (0,1) \to \mathbb{R}$ be the function given by

$$h(x) = x^{-x}(1-x)^{x-1}.$$

Then $h(x)$ is positive and has the derivative

$$h'(x) = x^{-x}(1-x)^{x-1}\left(\ln(1-x) - \ln(x)\right).$$

And the derivative vanishes if $\ln(1-x) = \ln(x)$ which is the case for $x = \frac{1}{2}$. Furthermore, we have

$$\ln(1-x) - \ln(x) > 0 \text{ for } x < \frac{1}{2} \qquad \text{and} \qquad \ln(1-x) - \ln(x) < 0 \text{ for } x > \frac{1}{2}.$$

Hence h assumes its maximum value for $x = \frac{1}{2}$, i.e.

$$\max_{0<x<1} h(x) = h(1/2) = 2.$$

With the choice $x = \frac{m}{n}$ we get

$$\left(\frac{n}{m}\right)^{\frac{m}{n}} \left(\frac{n}{n-m}\right)^{\frac{n-m}{n}} = h(m/n) \leq 2.$$

Combining this with (3.9) results in

$$\sigma_{2n} \leq \sqrt{2}\gamma^{-1}\sigma_n^{\frac{m}{n}} d_m^{\frac{n-m}{n}}, \qquad \text{for } m = 1, \ldots, n. \tag{3.10}$$

For the special case $n = 2s$ and $m = s$ for some $s \in \mathbb{N}$, (3.10) simplifies to

$$\sigma_{4s} \leq \sqrt{2}\gamma^{-1}\sqrt{\sigma_{2s} d_s}. \tag{3.11}$$

Finally, we can proove our claim via contradiction. For this let $M \in \mathbb{N}$ be the first number such that $\sigma_M(\mathcal{F}) > C_1 M^{-\alpha}$. We first consider the case $M = 4s$ for some $s \in \mathbb{N}$. Using (3.11) it follows

$$\sigma_{4s} \leq \sqrt{2}\gamma^{-1}\sqrt{C_1(2s)^{-\alpha} C_0 s^{-\alpha}} = \sqrt{2^{1-\alpha} C_0 C_1}\gamma^{-1} s^{-\alpha}.$$

Here we used, that $\sigma_{2s} \leq C_1^{(}2s)^{-\alpha}$ by assumption on M. It follows that

$$C_1(4s)^{-\alpha} < \sigma_{4s} \leq \sqrt{2^{1-\alpha} C_0 C_1}\gamma^{-1} s^{-\alpha}.$$

Therefore by solving for C_1

$$C_1 < 2^{3\alpha+1}\gamma^{-2} C_0 < 2^{5\alpha+1}\gamma^{-2} C_0 = C_1$$

and we have reached a contradiction. Let now $M = 4s + q$ for some $q \in \{1, 2, 3\}$, then we have similarly

$$C_1 2^{-3\alpha} s^{-\alpha} = C_1 2^{-\alpha}(4s)^{-\alpha} \leq C_1(4s+q)^{-\alpha} < \sigma_{4s+q} \leq \sigma_{4s} \leq \sqrt{2^{1-\alpha} C_0 C_1}\gamma^{-1} s^{-\alpha}.$$

Solving for C_1 once again, we obtain

$$C_1 < 2^{5\alpha+1}\gamma^{-2} C_0 = C_1$$

which is the desired contradiction. □

The comparison between $d_n(\mathcal{F})$ and $\sigma_n(\mathcal{F})$ in Corollary 3.3.4 is somewhat ill suited, as $d_n(\mathcal{F})$ deals with n–dimensional subspaces, whereas $\sigma_n(\mathcal{F})$ pertains to $N = n \cdot k$–dimensional subspaces. In order to make the two quantities more comparable, we have to express the latter in terms of the dimension of its approximation space:

Corollary 3.3.5. *Let $\{V_n\}_{n\in\mathbb{N}}$ and $\{W_n\}_{n\in\mathbb{N}}$ be the sequence of spaces generated by a weak greedy algorithm with constant γ and equip each W_n with a basis $\{g_n^1, \ldots, g_n^k\}$ such* *that*

$$g_n^i = \max_{g \in W_n} \left\| g - \Pi_{V_n + \mathrm{span}\{g_n^1, \ldots, g_n^{i-1}\}} g \right\|_{\mathcal{H}}.$$

Let $N = k \cdot n + l$ for some $n \in \mathbb{N}$ and some $1 \leq l \leq k-1$ and denote as $U_N \subset \mathcal{H}$ the subspaces given by

$$U_N = \mathrm{span}\{g_1^1, \ldots, g_n^l\}.$$

Furthermore, let $\tilde{\sigma}_N(\mathcal{F}) = \sup_{f \in \mathcal{F}} \|f - \Pi_{U_N} f\|_{\mathcal{H}}$. Then it holds that if $d_n(\mathcal{F}) \leq C_0 n^{-\alpha}$, then

$$\tilde{\sigma}_N \leq C_2 N^{-\alpha}, \qquad \text{for all } N \geq 2k-2,$$

where $C_2 := 2^{6\alpha+1} k^\alpha \gamma^{-2} C_0$.

Proof. First let $N = n \cdot k$. Then we have $\tilde{\sigma}_N(\mathcal{F}) = \sigma_n(\mathcal{F})$ and therefore by Corollary 3.3.4

$$\tilde{\sigma}_N(\mathcal{F}) \leq 2^{5\alpha+1}\gamma^{-2} n^{-\alpha} = 2^{5\alpha+1}\gamma^{-2} k^\alpha N^{-\alpha} C_0 \leq 2^{6\alpha+1}\gamma^{-2} k^\alpha N^{-\alpha} C_0.$$

Let now $N = n \cdot k + l$ for some $1 \leq l \leq k-1$. Then we have due to the monotonicity of $\{\tilde{\sigma}_m\}_{m \in \mathbb{N}}$

$$\tilde{\sigma}_N(\mathcal{F}) \leq \tilde{\sigma}_{nk} \leq 2^{5\alpha+1}\gamma^{-2} n^{-\alpha}. \tag{3.12}$$

Solving $N = nk + l$ for n we get $n = \frac{N-l}{k}$ and therefore (3.12)

$$\begin{aligned} \tilde{\sigma}_N(\mathcal{F}) \leq 2^{5\alpha+1}\gamma^{-2} C_0 \left(\frac{N-l}{k}\right)^{-\alpha} &\leq 2^{5\alpha+1}\gamma^{-2} k^\alpha C_0 \left(\frac{N-l}{N}\right)^{-\alpha} N^{-\alpha} \\ \leq 2^{5\alpha+1}\gamma^{-2} k^\alpha C_0 \left(\frac{1}{2}\right)^{-\alpha} N^{-\alpha} &= 2^{6\alpha+1}\gamma^{-2} k^\alpha C_0 N^{-\alpha} \end{aligned}$$

□

As we can see, the bound deteriorates by a factor of 2^α if we also want to consider dimensions that are not a multiple of k. Nonetheless, we maintain the same asymptotic behaviour and thus the weak greedy algorithm generates a sequence of subspaces which provide a quasi optimal approximation rate.

3.4 Convergence Rates for the P–Greedy Variants

We now show how the P–Greedy variants that were introduced in Section 3.2 fit into the framework described in Section 3.3. In this case, the Hilbert space is just given as the RKHS of the kernel $K : \Omega \times \Omega \to \mathbb{R}^{m\times m}$, i.e. $\mathcal{H} = \mathcal{H}_K$. Furthermore, the set $\mathcal{F}$ is given by

$$\mathcal{F} = \{K(\cdot, x)\alpha \mid x \in \Omega,\, \alpha \in \mathbb{R}^m,\, \|\alpha\|_2 = 1\}.$$

In the following we will assume that the kernel K is continuous and $\Omega \subset \mathbb{R}^d$ is compact. We can now conclude that the set $\Omega \times \{\alpha \in \mathbb{R}^m \mid \|\alpha\|_2 = 1\}$ is compact as a product of compact sets. Therefore $\mathcal{F}$ is compact as the image of the continuous mapping

$$\Omega \times \{\alpha \in \mathbb{R}^m \mid \|\alpha\|_2 = 1\} \ni (x, \alpha) \mapsto K(\cdot, x)\alpha.$$

With this, we can now show that the different routines described in Algorithm 3 are weak greedy algorithms as defined in Definition 3.3.1.

Proposition 3.4.1 (P–Greedy Variants are weak greedy algorithms)**.** *The P–Greedy variants of Algorithm 3 are all weak (or strong) greedy algorithms and the weak greedy constant are given by Table 3.1.*

Indicator	E_1			E_2			E_∞		
extension	full	eig	diag	full	eig	diag	full	eig	diag
constant γ	$1/m$	$1/m$	$1/m$	1	1	$1/m$	$1/m$	$1/m$	$1/m$

Table 3.1: Weak Greedy constants for the P–Greedy variants

Proof. Let $\{V_n\}_{n\in\mathbb{N}}$ denote the sequence of spaces generated by any of the P–Greedy variants. Then it holds

$$\begin{aligned}\sigma_n(\mathcal{F})^2 &= \max_{f\in\mathcal{F}} \|f - \Pi_{V_n} f\|_{\mathcal{H}_K}^2 = \max_{\substack{(x,\alpha)\in\Omega\times\mathbb{R}^m\\ \|\alpha\|_2=1}} \|K(\cdot, x)\alpha - K_{V_n}(\cdot, x)\alpha\|_{\mathcal{H}_K}^2\\ &= \max_{\substack{(x,\alpha)\in\Omega\times\mathbb{R}^m\\ \|\alpha\|_2=1}} \alpha^T \left(K(x, x) - K_{V_n}(x, x)\right)\alpha\\ &= \max_{x\in\Omega} \lambda_{\max}(\boldsymbol{P}_{V_n,K}(x)).\end{aligned}$$

We can immediately conclude that both $\text{greedy}_{2,\text{full}}$ and $\text{greedy}_{2,\text{eig}}$ lead to a strong greedy, i.e. $\gamma = 1$. For the remaining combinations we use the following inequality for symmetric positive semi-definite matrices $A \in \mathbb{R}^{m\times m}$:

$$\max \operatorname{diag}(A) \leq \lambda_{\max}(A) \leq \operatorname{tr}(A) \leq m \max \operatorname{diag}(A) \leq m\lambda_{\max}(A). \tag{3.13}$$

To finish off the variants using the indicator E_2, we immediately conclude from (3.13) that greedy$_{2,\text{diag}}$ is a weak greedy with constant $\gamma = m^{-1}$. We now consider the indicator function E_1. For every $n \in \mathbb{N}$ let $x_n \in \Omega$ denote the point selected by E_1, i.e.

$$x_n = \arg\max_{x\in\Omega} E_1(\boldsymbol{P}_{V_{n-1},K}(x)).$$

Furthermore, let x_n^* denote the point chosen by the indicator E_2, i.e the point such that

$$E_2(\boldsymbol{P}_{V_{n-1},K}(x_n^*)) = \lambda_{\max}(\boldsymbol{P}_{V_{n-1},K}(x_n^*)) = \sigma_{n-1}(\mathcal{F}).$$

By definition of x_n, x_n^* and by (3.13) it now follows that

$$\begin{aligned} m\lambda_{\max}(\boldsymbol{P}_{V_{n-1},K}(x_n)) &\geq m \max \operatorname{diag}(\boldsymbol{P}_{V_{n-1},K}(x_n)) \\ &\geq \operatorname{tr}(\boldsymbol{P}_{V_{n-1},K}(x_n)) = E_1(\boldsymbol{P}_{V_{n-1},K}(x_n)) \\ &\geq E_1(\boldsymbol{P}_{V_{n-1},K}(x_n^*)) = \operatorname{tr}(\boldsymbol{P}_{V_{n-1},K}(x_n^*)) \\ &\geq \lambda_{\max}(\boldsymbol{P}_{V_{n-1},K}(x_n^*)) = \sigma_{n-1}(\mathcal{F}). \end{aligned}$$

Therefore, the spaces selected by either extension routine result in a weak greedy with constant $\gamma = m^{-1}$. Finally, we consider the indicator function E_∞. Let $\{y_n\}_{n\in\mathbb{N}}$ be the sequence of points selected by this indicator. Then we have by (3.13) and definition of y_n, x_n^*

$$\begin{aligned} \lambda_{\max}(\boldsymbol{P}_{V_{n-1},K}(y_n)) &\geq \max \operatorname{diag}(\boldsymbol{P}_{V_{n-1},K}(y_n)) \geq \max \operatorname{diag}(\boldsymbol{P}_{V_{n-1},K}(x_n^*)) \\ &\geq \frac{1}{m}\lambda_{\max}(\boldsymbol{P}_{V_{n-1},K}(x_n^*)) = \frac{1}{m}\sigma_{n-1}(\mathcal{F}). \end{aligned}$$

Consequently, all routines result in a weak greedy algorithm with constant $\gamma = m^{-1}$. □

Remark 3.4.2. As we have mentioned before in Section 3.3, the work in [12, 27] on the convergence rates of weak greedy algorithms was done with the application to reduced basis methods in mind. Likewise, we note some similarities between the results of Proposition 3.4.1 and existing work in the reduced basis context. In particular, the routine greedy$_{1,\text{eig}}$ is equivalent to the POD-Greedy introduced in [41] for which analogous rates were proven [38].

We now consider kernels which stem from p.d. RBF which satisfy the conditions of Corollary 2.3.17 for some $s > d/2$. As we have seen in Theorem 2.5.9 and Corollary 2.5.10

bounds on the Power function are available. These bounds translate into upper bounds on the Kolmogorov n–width.

Lemma 3.4.3 (Kolmogorov n–width for p.d. RBF kernels)**.** *Let Φ be a p.d. RBF that satisfies the assumptions of Theorem 2.5.9 for some $s > d/2$ and let K denote the kernel induced by Φ, then for sufficiently large $n \in \mathbb{N}$ we have*

$$d_n(\mathcal{F}) \leq Cn^{-\frac{s-d/2}{d}} \tag{3.14}$$

for some $C > 0$.

Proof. Let $\{X_n\}_{n\in\mathbb{N}}$ with $X_n \subset \Omega$ be a sequence of sets, such that the points in each set are asymptotically uniformly distributed in Ω, i.e. there exists a constant $c > 0$ such that the fill distance satisfies

$$h_n := h_{X_n,\Omega} \leq cn^{-\frac{1}{d}}.$$

For sufficiently large n the fill distance h_n can therefore be arbitrarily small. From Corollary 2.5.10 we can conclude, that

$$d_n(\mathcal{F})^2 \leq \sup_{f\in\mathcal{F}} \left\| f - \Pi_{\mathcal{N}(X_n)} f \right\|_{\mathcal{H}_K}^2 = \sup_{x\in\Omega} \lambda_{\max}(\boldsymbol{P}_{\mathcal{N}(X_n),K}(x)) \leq \tilde{C} h_n^{2s-d} \leq c\tilde{C}n^{-\frac{2s-d}{d}}$$

for some $\tilde{C} > 0$. Taking the root gives the desired result. for $C = \sqrt{c\tilde{C}}$. □

The above shows, that the Kolmogorov n–width asymptotically behaves like $n^{-\alpha}$ with $\alpha = \frac{2s-d}{2d}$. However, because $\mathcal{N}(X_n)$ is $n \cdot m$–dimensional, the constant in (3.14) scales with $m^{\frac{1}{d}}$. Combining Corollary 3.3.5, Proposition 3.4.1 and Lemma 3.4.3, we can conclude the following.

Corollary 3.4.4. *Let Φ be a p.d. RBF that satisfies the assumptions of Theorem 2.5.9 for some $s > d/2$ and let K denote the kernel induced by Φ. Then each of the P–Greedy variants described in Algorithm 3 generates a sequence of subspaces $\{V_N\}_{N\in\mathbb{N}}$ such that*

$$\sigma_N(\mathcal{F}) \leq CN^{-\frac{2s-d}{2d}}.$$

for some constant $C > 0$ and for N sufficiently large.

While the previous corollary guarantees that the pointwise error for the interpolation process behaves like $N^{-\frac{2s-d}{2d}}$ asymptotically, it does not give any insight to the pointwise error if derivatives are considered. Nonetheless, this can be remedied by the following observation, which for scalar-valued kernels can be found in [110, 25] and which carries over to the matrix-valued case.

Lemma 3.4.5. *Let K be a p.d. kernel, such that $\mathcal{H}_K$ is norm equivalent to $W^s(\Omega)$ for some $s > d/2$. If for $X \subset \Omega$ there exists a bound of the form*

$$\left\| f - \Pi_{\mathcal{N}(X)} f \right\|_{L_\infty(\Omega)} \leq \epsilon \left\| f \right\|_{\mathcal{H}_K}$$

for all $f \in \mathcal{H}_K$, then the fill distance is bounded by

$$h_{X,\Omega} \leq C \epsilon^{\frac{2}{2s-d}}.$$

If we combine this with the results of Corollary 3.4.4 we get the following result on the fill distance, if the full extension routines are used in the P–Greedy algorithm.

Lemma 3.4.6. *Let K be a p.d. Kernel, such that $\mathcal{H}_K$ is norm equivalent to $W^s(\Omega)$ for some $s > d/2$. Then the P–Greedy variants of Algorithm 3 using the full extension routine generate a sequence $\{X_n\}_{n\in\mathbb{N}}$ of sets such that their respective fill distance is bounded by*

$$h_{X_n,\Omega} \leq C n^{\frac{1}{d}}$$

for some constant $C > 0$ and for sufficiently large n.

Proof. By Corollary 3.4.4 we have for sufficiently large n

$$\left\| f - \Pi_{\mathcal{N}(X_n)} f \right\|_{L_\infty(\Omega)} \leq \sup_{x\in\Omega} \lambda_{\max}(\boldsymbol{P}_{\mathcal{N}(X_n),K})(x)) \left\| f \right\|_{\mathcal{H}_k} \leq C n^{-\frac{2s-d}{2d}} \left\| f \right\|_{\mathcal{H}_K}$$

and we conclude using the previous lemma, that

$$h_{X_n,\Omega} \leq c \left(C n^{-\frac{2s-d}{2d}} \right)^{\frac{2}{2s-d}} = c C^{\frac{2}{2s-d}} n^{\frac{1}{d}}. \tag{3.15}$$

□

Consequently, we can use (3.15) in Theorem 2.5.9 to obtain:

Corollary 3.4.7. *Let K be a p.d. kernel such that $\mathcal{H}_K$ is norm equivalent to $W^s(\Omega)$ for some $s > d/2$. Then the P–Greedy variants of Algorithm 3 using the full extension routine generate a sequence $\{X_n\}_{n\in\mathbb{N}}$ of sets such that for any multiindex $\alpha \in \mathbb{N}_0^d$ with $|\alpha| < s - d/2$ we have*

$$\left\| D^\alpha \left(f(x) - \Pi_{\mathcal{N}(X_n)} f(x) \right) \right\|_2 \leq C n^{-\frac{k-|\alpha|-d/2}{d}}.$$

for some $C > 0$ and n sufficiently large.

3.4.1 Numerical investigation

We now want to investigate the effect of the different P–Greedy variants on the quality of the approximation. For this purpose we consider the following two matrix-valued kernels. For the first, let $\Omega_1 = [-1,1]$ and $K_1 : \Omega_1 \times \Omega_1 \to \mathbb{R}^{10\times 10}$ be a separable kernel given by

$$K_1(x,y) = e^{-4\|x-y\|^2} A_1 + e^{-10\|x-y\|^2} A_2$$

for two random but fixed symmetric matrices $A_1, A_2 \in \mathbb{R}^{10\times 10}$. As a target function we consider

$$f_1(x) = \sum_{i=1}^{10} K_1(x, x_i)\alpha_i,$$

where $x_1, \dots, x_{10}$ and $\alpha_1, \dots, \alpha_{10}$ are randomly chosen in Ω_1 and $\mathbb{R}^{10}$, respectively. For the second kernel, let $\Omega_2 = [-1,1]^3$ and the kernel $K_2 : \Omega_2 \times \Omega_2 \to \mathbb{R}^{3\times 3}$ be given as $K_2(x,y) = (\nabla\nabla^T k)(x,y)$, with the scalar-valued kernel $k(x,y) = e^{-\|x-y\|^2}$. Here ∇ denotes the application to the columns, whereas ∇^T denotes the application to the rows. For a target function we also consider

$$f_2(x) = \sum_{i=1}^{10} K_2(x, x_i)\alpha_i,$$

where $x_1, \dots, x_{10}$ and $\alpha_1, \dots, \alpha_{10}$ are randomly chosen in Ω_2 and $\mathbb{R}^3$, respectively.

All of the following experiments were implemented in MATLAB 2018a and run on a machine with an Intel Core i7-7500U CPU with 16GB Ram.

For a first test, we run the P–Greedy algorithm with the different indicator functions and the "full" extension routine until we reach an approximation space dimension of 300. The decay of the maximum indicator function values are depicted in Figure 3.5 and we can see that the maximum values decay at similar rates for both examples. However, using E_2 is computationally more expensive, as we have to solve multiple eigenvalue problems in every iteration. Furthermore, this requires the evaluation of the entire Power function matrix, whereas for E_1 and E_∞ we only need the diagonal values. In Figure 3.6 the distribution of the selected point sets for the different variants is displayed. In the case of K_1 we can see that the initial three points are identical for all three variants. Afterwards, we can observe that the same points might still be chosen, but not necessarily at the same step during the algorithm. In the case of K_2 we can not display both the points as well as the step in which each point was selected due to a lack of dimensionality. Nonetheless, we can make similar observation as all variants start by selecting the eight corners of $\Omega_2 = [-1,1]^3$.

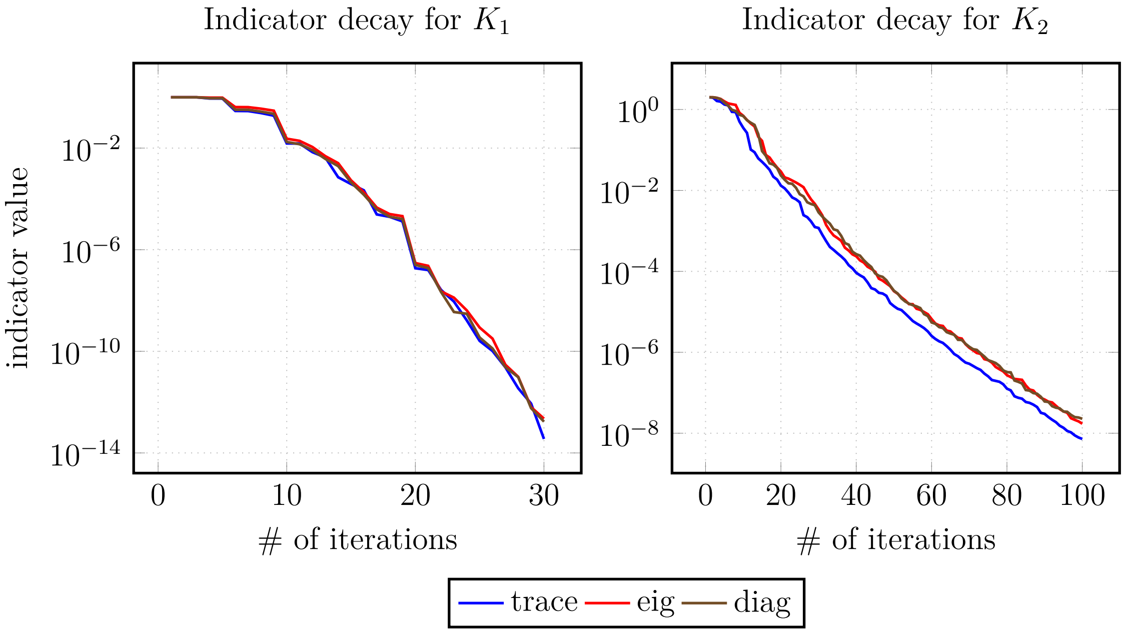

Figure 3.5: Decay of the maximum indicator function value with respect to the number of iterations for the three P–Greedy variants.

In the following we chose E_1 for further testing the different extension routines. For both examples we compute the sequence of interpolants, which we denote as

$$\left(s^n_{1,\text{full}}\right)_{1\leq n\leq 30}, \quad \left(s^n_{1,\text{eig}}\right)_{1\leq n\leq 300} \quad \text{and} \ \left(s^n_{1,\text{diag}}\right)_{1\leq n\leq 300}$$

for the kernel K_1 and

$$\left(s^n_{2,\text{full}}\right)_{1\leq n\leq 100}, \quad \left(s^n_{2,\text{eig}}\right)_{1\leq n\leq 300} \quad \text{and} \ \left(s^n_{2,\text{diag}}\right)_{1\leq n\leq 300}$$

for the kernel K_2 as well as the sequences of the error in the squared native space norm

$$\Delta^n_{i,\text{type}} := \left\| f - s^n_{i,\text{type}} \right\|^2_{\mathcal{H}_{K_i}}, \qquad i = 1, 2, \quad \text{type} \in \{\text{full}, \text{eig}, \text{diag}\}.$$

The sequences for the "full" extension routine are shorter than their respective counterpart, since the dimension of the approximation space increases faster. The decay in the error with respect to the dimension of the approximation space is depicted in Figure 3.7 and we can easily conclude that all P–Greedy extension routines result in approximations of the same quality. However, taking a closer look at the number of unique function evaluations which are required, we can see that the "full" extension routine clearly outperforms the other two. For our two examples this is shown in Figure 3.8 and we can infer larger dimensions, it seems to be the best choice.

As the final part of this subsection we want to verify the results of Lemma 3.4.6, i.e. if

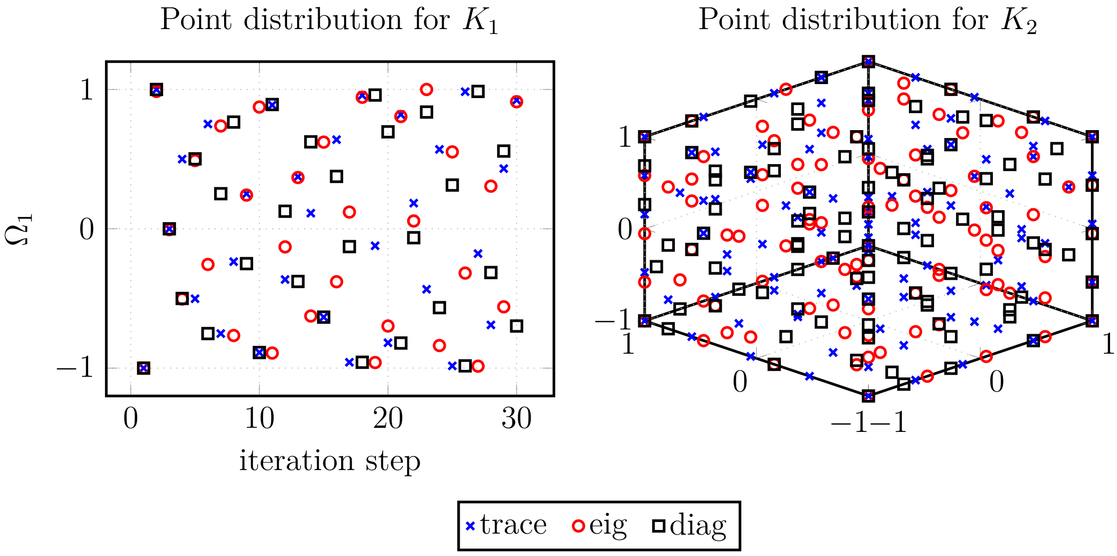

Figure 3.6: Point distribution for the different P–Greedy variants

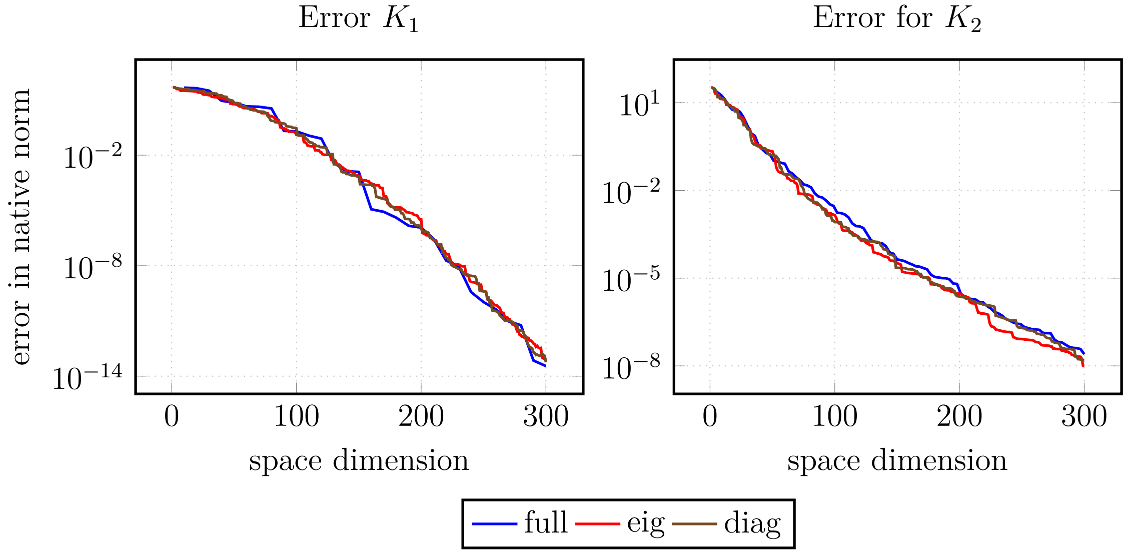

Figure 3.7: Error decay in the squared native norm with respect to the dimension of the approximation space

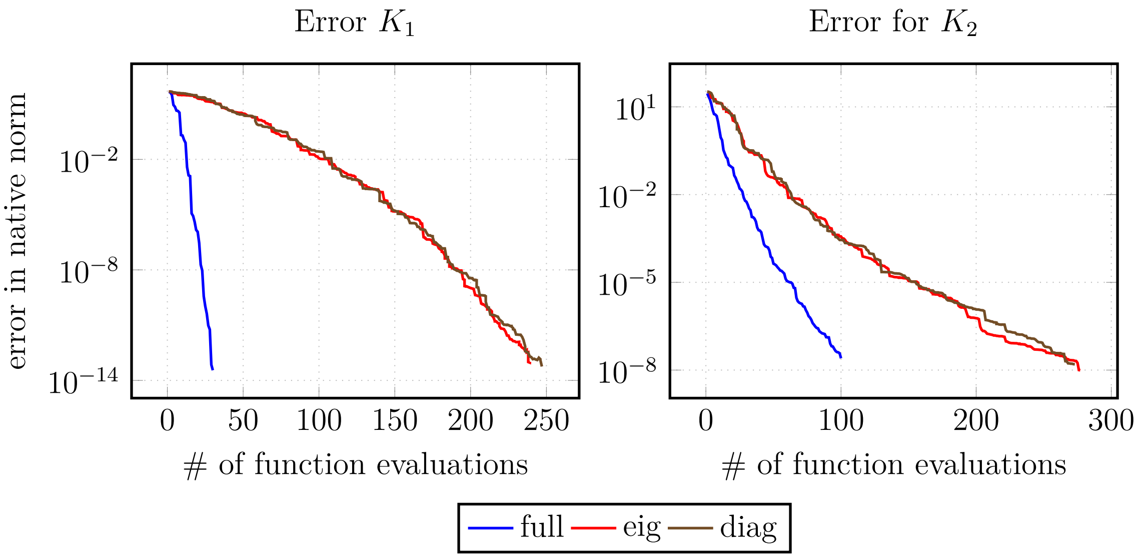

Figure 3.8: Error decay in the squared native norm with respect to the number of function evaluations

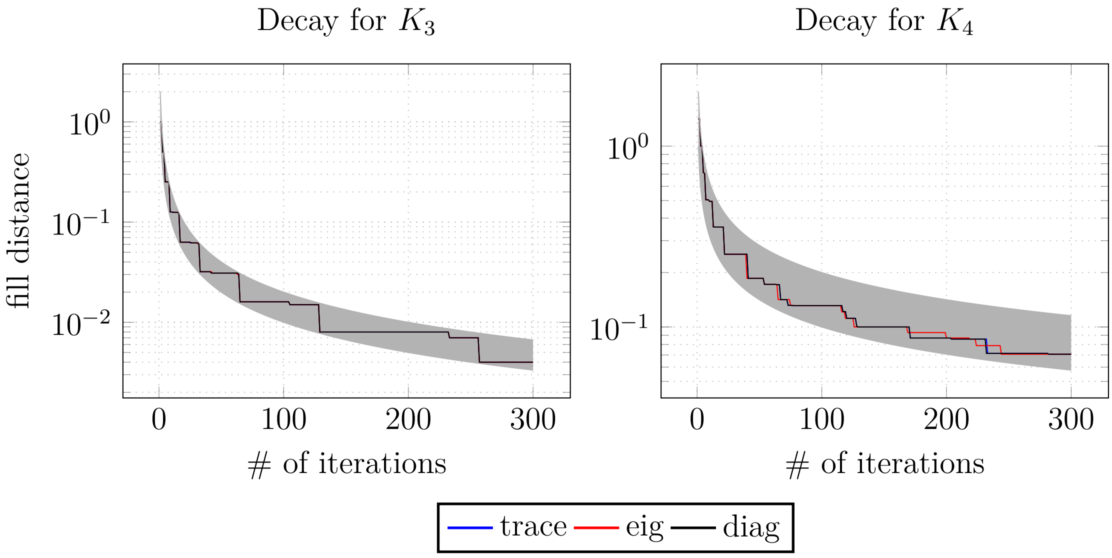

Figure 3.9: Decay of the fill distance in one and two dimensions.

the RKHS of the chosen kernel is norm equivalent to W^s for $s > d/2$, then the fill distance for the sets generated by all variants decay with a rate of $n^{1/d}$ if the full extension routine was used. To this end we consider the following four RBF

$$\phi_{w,1}(r) = (1+3r)(1-r)_+^3, \qquad \phi_{w,2}(r) = (1+18r+35r^2)(1-r)_+^6$$
$$\phi_{m,1}(r) = e^{-r}, \qquad \phi_{m,2}(r) = (1+r)e^{-r}$$

and the kernels $K_3 : [-1,1] \times [-1,1] \to \mathbb{R}^{3\times 3}$ and $K_4 : [0,1]^2 \times [0,1]^2 \to \mathbb{R}^{3\times 3}$ given by

$$K_3(x,y) = \phi_{w,1}(\|x-y\|_2)B_1 + \phi_{m,1}(\|x-y\|_2)B_2$$
$$K_4(x,y) = \phi_{w,2}(\|x-y\|_2)C_1 + \phi_{m,2}(\|x-y\|_2)C_2.$$

Here B_1, B_2, C_1, C_2 are randomly chosen but fixed symmetric positive definite matrices. It follows that both K_3 and K_4 are strictly positive definite kernels and by our choice of the Wendland functions $\phi_{w,1}, \phi_{w,2}$ and the Matérn kernels $\phi_{m,1}, \phi_{m,2}$ their RKHS are norm equivalent to $W^1([0,1])$ and $W^2([0,1]^2)$, respectively. In Figure 3.9 the decay of the fill distance for 300 greedy steps is displayed for both kernels and all three greedy variants. In grey, the area between $1/n$ and $2/n$ as well as $1/\sqrt{n}$ and $2/\sqrt{n}$ is plotted. As we can see, the decay for all variants verifies the results of Lemma 3.4.6.

3.5 Surrogate Modelling for uncertainty quantification for a carbon dioxide storage scenario

We conclude this chapter on approximation of vector-valued functions via interpolation by looking at a carbon dioxide storage scenario where multiple function evaluations are necessary for uncertainty quantification. The example stems from the collaborative work [54] in which multiple different data-driven surrogate modelling approaches were compared. For a more detailed look into the modelling of the carbon dioxide storage scenario as well as the remaining data-driven methods which were used, namely non-intrusive arbitrary polynomial chaos expansion [71], spatially adaptive sparse grids [75] and Hybrid stochastic Galerkin methods [55], we refer to the original paper as well as references therein. In this section, we will only focus on the surrogate modelling via matrix-valued kernels, as it comprises elements of all previous sections and chapters. In particular, we include further comparisons with uncoupled separable kernels which were constructed by the method described in Algorithm 1. All numerical computations were redone for this section and slightly modified to allow for better comparison between the kernel approximations themselves. For our purposes, it is enough to know that the desired quantity, namely the saturation $\boldsymbol{S}$ of carbon dioxide in the respective reservoir at final time, can be modelled via a function $\boldsymbol{S} : X \subset \mathbb{R}^3 \to \mathbb{R}^{250}$, where the three random input parameters relate to the injection rate, reservoir porosity and the relative permeability of the reservoir, respectively.

In the context of uncertainty quantification, it is now required to perform a large sampling of the saturation function $\boldsymbol{S}$ for different input parameter combinations. In order to quantify the performance of the surrogate model, the function $\boldsymbol{S}$ is evaluated on the set X_{eval} consisting of 10000 randomly selected parameters and the mean and variance are computed.

The evaluation of the target function $\boldsymbol{S}$ is implemented in C++, whereas the remaining implementations were done using MATLAB 2018a. All computations were run on a machine with an Intel Core i7-7500U CPU with 16GB RAM.

The construction of our kernel surrogate models are now achieved as follows. First, we split our set X_{eval} into the disjoint subsets $X_{\text{train}}, X_{\text{build}}, X_{\text{val}}, X_{\text{test}} \subset X_{\text{eval}}$ consisting of 3000, 1000, 1000 and 5000 points, respectively. As outlined in Algorithm 1 we then compute the empirical covariance matrix of $\boldsymbol{S}(X_{\text{build}})$. The index sets are then selected by grouping all indexes for which the corresponding eigenvalue in the spectral decomposition $\text{Cov}(\boldsymbol{S}(X_{\text{build}})) = U(\boldsymbol{S}(X_{\text{build}}))\Sigma(\boldsymbol{S}(X_{\text{build}}))U(\boldsymbol{S}(X_{\text{build}}))^T$ are of the same magnitude, up to a magnitude of 10^{-4}. This results in the following 6 index sets

$$\begin{aligned}
\mathcal{I}_1 &= \{i\,|10^0 \leq \sigma_i(\boldsymbol{S}(X_{\text{build}}))) < 10^1\} = \{1, \dots, 2\},\\
\mathcal{I}_2 &= \{i\,|10^{-1} \leq \sigma_i(\boldsymbol{S}(X_{\text{build}}))) < 10^0\} = \{3, \dots, 6\},\\
\mathcal{I}_3 &= \{i\,|10^{-2} \leq \sigma_i(\boldsymbol{S}(X_{\text{build}}))) < 10^{-1}\} = \{7, \dots, 20\},\\
\mathcal{I}_4 &= \{i\,|10^{-3} \leq \sigma_i(\boldsymbol{S}(X_{\text{build}}))) < 10^{-2}\} = \{21, \dots, 57\},\\
\mathcal{I}_5 &= \{i\,|10^{-4} \leq \sigma_i(\boldsymbol{S}(X_{\text{build}}))) < 10^{-3}\} = \{58, \dots, 127\},\\
\mathcal{I}_6 &= \{i\,|\sigma_i(\boldsymbol{S}(X_{\text{build}}))) < 10^{-4}\} = \{127, \dots, 250\}
\end{aligned}$$

and their corresponding matrices Q_i. The scalar-valued kernels are selected as follows. We start with the Wendland function

$$\phi_w(r) = \frac{1}{3}(3 + 18r + 35r^2)(1-r)_+^6$$

and then consider the three kernels

$$\begin{aligned}
&k_1(x,y) = \phi_w(0.2\,\|x-y\|_2), \qquad k_2(x,y) = \phi_w(0.35\,\|x-y\|_2),\\
&k_3(x,y) = \phi_w(0.5\,\|x-y\|_2).
\end{aligned}$$

We now construct all possible pairings of these three kernels with the first 5 index sets, which results in a total of $3^5 = 243$ combinations. For each of these combinations, the index set $\mathcal{I}_6$ is paired with the kernel k_3, as the influence on the approximation quality is deemed negligible. We then run the P–Greedy algorithm with the trace indicator function and the full extension routine on the set X_{train} until 500 points are selected. We then evaluate the corresponding 243 surrogate models in the validation set X_{val} and compute their respective mean. Likewise, we compute the mean of $\boldsymbol{S}(X_{\text{val}})$ and then select the two index combinations that maximize, i.e. $(1, 2, 1, 1, 1)$, and minimize, i.e. $(1, 1, 3, 3, 3)$, the error in the Euclidean norm. These kernels are then joint by the three matrix-valued kernels, for which all index sets are paired with the same scalar-valued kernels. This results in the following 5 uncoupled separable kernels

$$\begin{aligned}
K_1(x,y) &= k_1(x,y)(Q_1 + Q_3 + Q_4 + Q_5) + k_2(x,y)Q_2 + k_3(x,y)Q_6\\
K_2(x,y) &= k_1(x,y)(Q_1 + Q_2) + k_3(x,y)(Q_3 + Q_4 + Q_5 + Q_6),\\
K_3(x,y) &= k_1(x,y)I, \qquad K_4(x,y) = k_2(x,y)I, \qquad K_5(x,y) = k_3(x,y)I.
\end{aligned}$$

Finally, we generate a set X_{greedy} by intersecting the minimum box enclosing X_{eval} with a $50 \times 50 \times 50$ grid of uniformly spaced points. This results in $|X_{\text{greedy}}| = 86021$ points. We then run the P–Greedy algorithm for the five kernels on the set X_{greedy} generating five sets of 500 points each. The surrogate models are then built using $10, 20, 50, 100, 200$ and 500 of these points, respectively and are evaluated on the test set X_{test}. the Euclidean errors in the mean and the variance of the surrogate and the target function are depicted in Figure 3.10. We can notice no improvement in the test error for the mean of the function, when using the kernel K_2, compared to the kernels K_3 and K_4. In the variance, on the other hand we can notice a slight improvement in the quality of the approximation for K_2, compared to the other 4 kernels. However, compared to the additional number of function evaluations that were necessary to construct the surrogate models for the kernels K_1 and K_2, the benefit in approximation quality is negligible. Nonetheless, this does not mean that the ansatz using matrix-valued uncoupled separable kernels cannot result in better surrogate models, see the example in Section 2.4, but that more sophisticated methods for parameter validation are required. By definition of uncoupled separable kernels, both the choice of scalar-valued kernels k_i, as well as the choice of the matrices Q_i can bee seen internal parameters of the surrogate model, the latter of which scale quadratically with respect to the output dimension of the target function. In other words, the parameter validation becomes increasingly complex as the dimension of the problem increases.

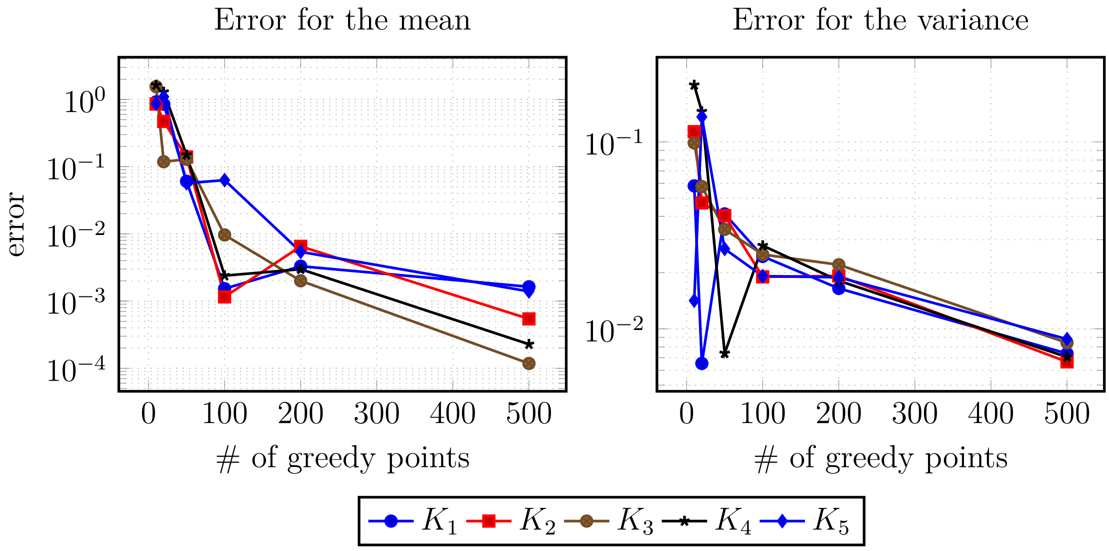

Figure 3.10: Error in the mean and variance on the test set X_{test} for the five different kernel models.

4 Weighted Regularized Interpolation

In the previous section we have only considered approximation via interpolation, i.e. best approximation in certain subspaces. In the following we want to focus on a regularized approach. This approach has some advantages over exact interpolation. Namely, as we have seen in Section 3.1, the condition of the linear systems which have to be solved become increasingly worse, as the number of interpolation points is increased. This problem can be alleviated using a regularized method. It is well suited for noisy data values, for which exact interpolation is not meaningful. Furthermore, in the case of positive definite kernels, the approximation procedure can be extended also to data, that does not correspond to a function in the RKHS, since the involved linear system is always solvable due to the influence of the regularization.

4.1 The regularized kernel interpolant

Definition 4.1.1 (Weighted loss function and cost functional)**.** Let $\omega : \Omega \to \mathbb{R}^{m\times m}$ be a strictly positive definite weight function, i.e. $\omega(x) \succ 0$ for all $x \in \Omega$. Furthermore, let $X = \{x_1, \dots, x_n\} \subset \Omega$ be a set of pairwise distinct points, $Y = \{y_1, \dots, y_n\} \subset \mathbb{R}^m$. Then the weighted loss functional $\mathcal{L}^{\omega}_{X,Y} : \mathcal{H}_K \to \mathbb{R}$ is given by

$$\mathcal{L}^{\omega}_{X,Y}(g) := \sum_{i=1}^{n} \left(g(x_i) - y_i\right)^T \omega(x_i)^{-1} \left(g(x_i) - y_i\right). \tag{4.1}$$

We further define a cost functional $\mathcal{J}^{\omega}_{X,Y} : \mathcal{H}_K \to \mathbb{R}$ via

$$\mathcal{J}^{\omega}_{X,Y}(g) := \mathcal{L}^{\omega}_{X,Y}(g) + \|g\|^2_{\mathcal{H}_K}. \tag{4.2}$$

In the following, we will just refer to $\omega : \Omega \to \mathbb{R}^{m\times m}$ as a weight function. If one neglects the second summand in the cost functional (4.2), the kernel interpolant corresponding to the data minimizes the loss functional. In a similar fashion, we will later on define the regularized interpolation as the minimizer of the cost functional. Here, the purpose of the second summand $\|\cdot\|_{\mathcal{H}_K}$ becomes apparent, as it punishes a potentially high norm in the interpolant. $\|\cdot\|_{\mathcal{H}_K}$ is thus referred to as a regularization functional. Of course, there are a wide variety of choices for the loss and regularization functional and one is not limited to the two above. In the case of scalar-valued kernels, the cost functional usually takes the form

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(g(x_i),y_i)+\omega_0\,\|g\|^2_{\mathcal{H}_K}$$

with the most common loss functions given by a least square loss function $\mathcal{L}(g(x_i),y_i) = |g(x_i)-y_i|^2$ or the ε–insensitive loss function $\mathcal{L}(g(x_i),y_i) = \max(0,|g(x_i)-y_i|-\varepsilon)$ and a constant scalar the so called penalization parameters $\omega_0 > 0$. In the case of the least square loss function this is also known as kernel ridge regression [106].

We can immediately see that compared to the scalar-valued ansatz, we shifted the constant penalization parameter ω_0 from the regularization functional to the loss functional, and further allow it to depend on the inputs x_i as well. While this provides us with more flexibility when constructing the surrogate, it also increases the number of parameters inside the surrogate model. Similar, to the scalar-valued case, our cost functional can be seen as a special case of the so called Tikhonov regularization [102], but tailored to the case of kernel based approximation.

For our specific choice a minimizer of the cost functional can be explicitly computed. In order to show this, the following lemma provides us with the necessary tool.

Lemma 4.1.2. *Let $A,B \in \mathbb{R}^{m\times m}$ be two symmetric and positive semi-definite matrices. Then it holds*

(i) $\mathrm{null}(A+B) = \mathrm{null}(A)\cap\mathrm{null}(B)$*,*

(ii) $\mathrm{range}(A+B) = \mathrm{range}(A)+\mathrm{range}(B)$*.*

Proof. **(a)** The inclusion $\mathrm{null}(A)\cap\mathrm{null}(B) \subset \mathrm{null}(A+B)$ is trivial. For the other direction let $v \in \mathrm{null}(A+B)$, then

$$0 = v^T(A+B)v = \underbrace{v^TAv}_{\geq 0}+\underbrace{v^TBv}_{\geq 0}$$

And therefore $Av = Bv = 0$.

(b) Since A,B and $A+B$ are symmetric, it holds $\mathrm{range}(A+B) = \mathrm{null}(A+B)^\perp$, $\mathrm{range}(A) = \mathrm{null}(A)^\perp$ and $\mathrm{range}(B) = \mathrm{null}(B)^\perp$. Furthermore it holds for general subspaces $U,V \subset \mathbb{R}^m$, that

$$(U+V)^\perp = U^\perp \cap V^\perp$$

and hence

$$(U^\perp + V^\perp)^\perp = (U^\perp)^\perp \cap (V^\perp)^\perp = U \cap V.$$

Thus, we have by (i)

$$\operatorname{range}(A+B) = \operatorname{null}(A+B)^\perp = (\operatorname{null}(A) \cap \operatorname{null}(B))^\perp = \operatorname{range}(A) + \operatorname{range}(B).$$

□

The existence of a unique minimizer can now be shown. As mentioned before, one can use different regularization functionals and it is well known [90] that unique minimizers exist, as long as the regularization functional is convex. We present the proof for our specific choice of functionals which also includes a direct way to compute the regularized approximant.

Theorem 4.1.3 (Representer Theorem). *Let $\omega : \Omega \to \mathbb{R}^{m\times m}$ be a weight function, $X = \{x_1, \dots, x_n\} \subset \Omega$ a set of pairwise distinct points, $Y = \{y_1, \dots, y_n\} \subset \mathbb{R}^m$, $\Lambda = \{\lambda_1, \dots, \lambda_p\} \subset \mathcal{H}_K'$ and $z \in \mathbb{R}^p$. If $z \in \operatorname{range}(S_\Lambda^1 S_\Lambda^2 K)$, this is equivalent to the condition that the set $\{g \in \mathcal{H}_K \mid S_\Lambda(g) = z\}$ is non-empty, then the minimization problem*

$$\min_{\substack{g\in\mathcal{H}_K\\ S_\Lambda(g)=z}} \mathcal{J}_{X,Y}^\omega(g) \tag{4.3}$$

has a uniqe solution $s \in \mathcal{N}(X) + \mathcal{N}(\Lambda)$. Furthermore, the solution is given by

$$s = K(\cdot, X)\alpha + S_\Lambda^2 K\beta \tag{4.4}$$

where the coefficient vectors $\alpha \in \mathbb{R}^{mn}$, $\beta \in \mathbb{R}^p$ solve the linear system

$$\underbrace{\begin{bmatrix} A+W & B \\ B^T & C \end{bmatrix}}_{=:M} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y \\ z \end{pmatrix} \tag{4.5}$$

with the quantities given as follows

$$\begin{aligned}
A &\coloneqq K(X,X) \in \mathbb{R}^{mn\times mn}\\
W &\coloneqq \operatorname{diag}(\omega(x_1),\dots,\omega(x_n)) \in \mathbb{R}^{mn\times mn}\\
B &\coloneqq S^2_\Lambda K(X,\cdot) \in \mathbb{R}^{mn\times p}\\
C &\coloneqq S^1_\Lambda S^2_\Lambda K \in \mathbb{R}^{p\times p}\\
y &\coloneqq (y_1^T,\dots,y_n^T)^T \in \mathbb{R}^{mn}.
\end{aligned}$$

Proof. Due to the assumption $z \in \operatorname{range}(C)$ the minimization problem is well posed by Corollary 2.2.15. Let

$$N_1 \coloneqq \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \qquad \text{and} \qquad N_2 \coloneqq \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}.$$

Then both N_1 and N_2 are symmetric and positive semi-definite. By Lemma 4.1.2 we have $\operatorname{range}(M) = \operatorname{range}(N_1) + \operatorname{range}(N_2)$. Thus, there exists a vector $u \in \mathbb{R}^{mn}$ such that $(u^T, z^T)^T \in \operatorname{range}(N_1)$ and since W is invertible, we have $\operatorname{range}(N_2) = \mathbb{R}^{mn} \times \{0\}^p$. We can conclude that $(y^T, z^T)^T \in \operatorname{range}(M)$, i.e. the linear system (4.5) has a solution. Furthermore, we have by Lemma 4.1.2 that $\operatorname{null}(M) = \operatorname{null}(N_1) \cap \operatorname{null}(N_2) = \{0\}^{mn} \times (C)$. Let $(0, \gamma^T)^T \in \operatorname{null}(M)$ and set

$$s_\gamma \coloneqq S^2_\Lambda K\gamma$$

and since $S_\Lambda(s_\gamma) = C\gamma = 0$, we have $s_y = 0$. In other words, every solution of (4.5) results in the same function s as given in (4.4). We now show that s is the unique minimizer of (4.3). To this end, let $g \in \mathcal{H}_K$, $g \neq 0$ with $S_\Lambda(g) = 0$. It follows

$$\begin{aligned}
\mathcal{J}^\omega_{X,Y}(s+g) &= \sum_{i=1}^n \left(s(x_i) + g(x_i) - y_i\right)^T \omega(x_i)^{-1} \left(s(x_i) + g(x_i) - y_i\right) + \|s+g\|^2_{\mathcal{H}_K}\\
&= \left(s(X) + g(X) - y\right)^T W^{-1} \left(s(X) + g(X) - y\right) + \|s+g\|^2_{\mathcal{H}_K}\\
&= \left(s(X) - y\right)^T W^{-1} \left(s(X) - y\right) + 2\left(s(X) - y\right)^T W^{-1} g(X)\\
&\quad + g(X)^T W^{-1} g(X) + \|s\|^2_{\mathcal{H}_K} + 2\langle s, g\rangle_{\mathcal{H}_K} + \|g\|^2_{\mathcal{H}_K}.\\
&= \mathcal{J}^\omega_{X,Y}(s) + 2\left(s(X) - y\right)^T W^{-1} g(X) + g(X)^T W^{-1} g(X)\\
&\quad + 2\langle s, g\rangle_{\mathcal{H}_K} + \|g\|^2_{\mathcal{H}_K}\\
&> \mathcal{J}^\omega_{X,Y}(s) + 2\left(s(X) - y\right)^T W^{-1} g(X) + 2\langle s, g\rangle_{\mathcal{H}_K}.
\end{aligned}$$

It is thus sufficient to show that

$$(s(X) - y)^T W^{-1} g(X) = -\langle s, g \rangle_{\mathcal{H}_K}.$$

On the one hand, we have

$$s(X) - y = A\alpha + B\beta - y = (A + W)\alpha + B\beta - Y - W\alpha = W\alpha$$

and consequently

$$(s(X) - y)^T W^{-1} g(X) = -\alpha^T g(X).$$

On the other hand, we have by the reproducing property that

$$\langle s, g \rangle_{\mathcal{H}_K} = \langle K(\cdot, X)\alpha + S_\Lambda^2 K\beta, g \rangle_{\mathcal{H}_K} = g(X)^T \alpha + S_\Lambda(g)^T \beta = g(X)^T \alpha.$$

□

Since the approximant s includes both an interpolation condition for the functionals in Λ, as well as a regularization via $\mathcal{J}^\omega_{X,Y}$, we shall denote it as a **regularized interpolant**. Furthermore, the uniqueness of s enables us to summarize the above procedure in terms of an operator, if the data in y stem from the evaluation of a target function $f \in \mathcal{H}_K$ on X:

Definition 4.1.4 (Regularized Interpolation Operator)**.** Let $X = \{x_1, \ldots, x_n\} \subset \Omega$, $\Lambda = \{\lambda_1, \ldots, \lambda_p\} \subset \mathcal{H}_K'$ and $\omega : \Omega \to \mathbb{R}^{m \times m}$ be a weight function. The regularized interpolation operator $\mathcal{I}^\omega_{X,\Lambda} : \mathcal{H}_K \to \mathcal{N}(X) + \mathcal{N}(\Lambda)$ is defined by

$$\mathcal{I}^\omega_{X,\Lambda}(f) \coloneqq \arg \min_{\substack{g \in \mathcal{H}_K \\ S_\Lambda(g) = S_\Lambda(f)}} \mathcal{J}^\omega_{X,f(X)}(g).$$

As we mentioned before, the regularized interpolant $\mathcal{I}^\omega_{X,\Lambda}(f)$ consists of an interpolation and a regularizing part. This becomes even more apparent when we consider the following split of $\mathcal{I}^\omega_{X,\Lambda}$:

Proposition 4.1.5 (Alternative representation of $\mathcal{I}^\omega_{X,\Lambda}$)**.** *We denote by $\mathcal{I}^\omega_X : \mathcal{N}(\Lambda)^\perp \to \mathcal{N}(X)$ a pure regularization operator, i.e. an empty interpolation set with respect to Definition 4.1.4, when applied to the kernel $K_{\mathcal{N}(\Lambda)^\perp}$. It now holds*

$$\mathcal{I}^\omega_{X,\Lambda}(f) = \Pi_{\mathcal{N}(\Lambda)}(f) + \mathcal{I}^\omega_X(\Pi_{\mathcal{N}(\Lambda)^\perp} f).$$

Proof. Using the same notations as in the proof of Theorem 4.1.3 we have by (2.17)

$$\Pi_{\mathcal{N}(\Lambda)} f = S_\Lambda^2 K C^+ S_\Lambda(f).$$

Likewise, applying Theorem 4.1.3 for $\mathcal{I}_X^\omega$ and using the representation of $K_{\mathcal{N}(\Lambda)^\perp}$ as given in (2.18) yields

$$\begin{aligned}\mathcal{I}_X^\omega(\Pi_{\mathcal{N}(\Lambda)^\perp} f) &= K_{\mathcal{N}(\Lambda)^\perp}(\cdot, X)\left(K_{\mathcal{N}(\Lambda)^\perp}(X,X) + W\right)^{-1} \Pi_{\mathcal{N}(\Lambda)^\perp} f(X)\\ &= K_{\mathcal{N}(\Lambda)^\perp}(\cdot, X)\left(A - BC^+B^T + W\right)^{-1}\left(f(X) - BC^+S_\Lambda(f)\right)\\ &= \left(K(\cdot, X) - S_\Lambda^2 K C^+ B^T\right)\left(A - BC^+B^T + W\right)^{-1}\left(f(X) - BC^+S_\Lambda(f)\right).\end{aligned}$$

Therefore

$$\begin{aligned}\Pi_{\mathcal{N}(\Lambda)} f + \mathcal{I}_X^\omega(\Pi_{\mathcal{N}(\Lambda)^\perp} f) = K(\cdot, X)&\left(A - BC^+B^T + W\right)^{-1}\left(f(X) - BC^+S_\Lambda(f)\right)\\ + S_\Lambda^2 K &\left(C^+S_\Lambda(f) - C^+B^T\left(A - BC^+B^T + W\right)^{-1}\left(f(X) - BC^+S_\Lambda(f)\right)\right)\end{aligned}$$

Thus, it is sufficient to show that

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} := \begin{pmatrix}\left(A - BC^+B^T + W\right)^{-1}(f(X) - BC^+S_\Lambda(f))\\ \left(C^+S_\Lambda(f) - C^+B^T\left(A - BC^+B^T + W\right)^{-1}(f(X) - BC^+S_\Lambda(f))\right)\end{pmatrix}$$

solve the linear system

$$\begin{bmatrix}A + W & B\\ B^T & C\end{bmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix}f(X)\\ S_\Lambda(f)\end{pmatrix}.$$

However, one easily verifies that

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{bmatrix}A + W & B\\ B^T & C\end{bmatrix}^+\begin{pmatrix}f(X)\\ S_\Lambda(f)\end{pmatrix}$$

by making use of the Schur-like decomposition which was also used in the proof of Lemma 3.1.1:

$$\begin{bmatrix}A + W & B\\ B^T & C\end{bmatrix}^+ = \begin{bmatrix}I_{nm} & 0\\ -C^+B^T & I_p\end{bmatrix}\begin{bmatrix}\left(A + W - BC^+B^T\right)^{-1} & 0\\ 0 & C^+\end{bmatrix}\begin{bmatrix}I_p & -BC^+\\ 0 & I_{nm}\end{bmatrix}.$$

□

Proposition 4.1.5 shows that the interpolatory constraints on the set Λ can be circumvented by computing the interpolant in $\mathcal{N}(\Lambda)$ and then performing the regularization part on the orthogonal complement $\mathcal{N}(\Lambda)^\perp$. Therefore, we only focus on the pure regularization part in the following, i.e. we assume that $\Lambda = \emptyset$. However, we want to remark that all subsequent results also hold for $\Lambda \neq \emptyset$ and one only has to replace RKHS $\mathcal{H}$ with $\mathcal{N}(\Lambda)^\perp$ and the the kernel K with the corresponding reproducing kernel $K_{\mathcal{N}(\Lambda)^\perp}$. As mentioned in the beginning of this chapter, in case of noisy data, one usually wants to avoid exact interpolation if possible. Nonetheless, interpolation at certain points or for certain functionals is still of interest, if certain properties of the function underlying the noisy data is known or required for the approximant. For example, one might know the position of maxima/minima or inflection points and hence prescribing zero values in the first or second order derivative guarantees these properties.

We conclude this subsection by introducing a pseudo Lagrangian basis and connecting it to a weighted version of the l_2–Lebesgue function:

Proposition 4.1.6 ((Pseudo) Lagrange Basis)**.** *Let $X = \{x_1, \dots, x_n\} \subset \Omega$, $\omega : \Omega \to \mathbb{R}^{m\times m}$ a weight function and denote $W = \operatorname{diag}(\omega(x_1), \dots, \omega(x_n))$. Then there exists a generating set $\{l^\omega_{X,1}, \dots, l^\omega_{X,nm}\}$ of $\mathcal{N}(X)$ such that*

$$\mathcal{I}^\omega_X(f) = L^\omega_X f(X) = \sum_{i=1}^{n}\sum_{j=1}^{m} l^\omega_{X,(i-1)n+j} f(x_i)^T e_j. \tag{4.6}$$

The generating set is given by the columns of

$$L^\omega_X = K(\cdot, X)\left(K(X,X) + W\right)^{-1}. \tag{4.7}$$

Furthermore, the weighted l_2–Lebesgue function Ψ^ω_X satisfies

$$\Psi^\omega_X(\lambda) \coloneqq \sup_{f\in\mathcal{H}_K} \frac{|\lambda(\mathcal{I}^\omega_X(f))|}{\|W^{-1/2}f(X)\|_2} = \left(\lambda(L^\omega_X)W\lambda(L^\omega_X)^T\right)^{1/2} \tag{4.8}$$

for any $\lambda \in \mathcal{H}'_K$ and $W^{-1/2}$ is the inverse of a symmetric positive definite root $W^{1/2}$ of W.

Proof. It follows from (4.7) that the columns of L^ω_X are a generating set of $\mathcal{N}(X)$, since the columns of $K(\cdot, X)$ are a generating set by definition. From (4.6) we can infer that

$$\mathcal{I}^\omega_X(f) = L^\omega_X f(X) = L^\omega_X W^{1/2} W^{-1/2} f(X)$$

and consequently by Cauchy-Schwarz inequality

$$|\lambda(\mathcal{I}^\omega_X(f))| \leq \left(\lambda(L^\omega_X)W\lambda(L^\omega_X)^T\right) \tag{4.9}$$

for any $f \in \mathcal{H}_K$. To reach equality, we just need to show that there exists a $f \in \mathcal{H}_K$ such that $W^{1/2}\lambda(L_X^\omega))^T$ and $W^{-1/2}f(X)$ are linearly dependent. This is equivalent to

$$W\lambda(L_X^\omega))^T = f(X)$$

for some $f \in \mathcal{H}_K$. By Lemma 2.2.14 we already know that $f(X) \in \text{range}(K(X,X))$ and we can further conclude for $A = K(X,X)$

$$\begin{aligned} W\lambda(L_X^\omega)^T &= W\lambda\left(K(\cdot,X)(A+W)^{-1}\right)^T = W\,(A+W)^{-1}\,\lambda^2 K(X,\cdot) \\ &= (A+W-A)(A+W)^{-1}\lambda^2 K(X,\cdot) \\ &= (\lambda^2 K)(X) - A(A+W)^{-1}(\lambda^2 K)(X) \in \text{range}(A). \end{aligned}$$

Therefore, there exists a function f such that equality holds in (4.9). □

In the case of $\omega = 0$ and for a strictly positive definite kernel K, the functions $l_{X,i}^0$ are in fact a basis of $\mathcal{N}(X)$ and satisfy

$$l_{X,(i-1)n+j}(x_k)^T e_p = \begin{cases} 1, & \text{if } (i,j) = (k,p) \\ 0, & \text{else} \end{cases}$$

which is the reason why the name "pseudo Lagrangian basis" was chosen in the case of nonzero ω.

4.1.1 Regularization and Interpolation with a modified Kernel

We again want to emphasize, that as a direct consequence of Proposition 4.1.5, we can absorb the interpolation part stemming from Λ into the regularization part by replacing the RKHS $\mathcal{H}_K$ with the orthogonal complement $\mathcal{N}(\Lambda)^\perp$ and changing the kernel accordingly. In this new space, the interpolation condition is always satisfied and thus we will proceed under the assumption that $\Lambda = \emptyset$ to avoid unnecessarily convoluted adaptions to the notation. Under these assumption, we can observe that, by Theorem 4.1.3, the coefficients $\alpha \in \mathbb{R}^{mn}$ in the expansion of $\mathcal{I}_X^\omega(f)$ for the generating system given by the columns of $K(\cdot,X)$ solve the linear system

$$(A+W)\alpha = \left(K(X,X) + \text{diag}(\omega(x_1),\ldots,\omega(x_n))\right)\alpha = f(X).$$

The matrix $A+W$ can be interpreted as the Gramian $K_\omega(X,X)$ of the modified kernel

$$K_\omega(x,y) := K(x,y) + \omega(x)\delta_x(y).$$

The kernel $\delta_\omega ::= \omega(x)\delta_x(y)$ is strictly positive definite, as was shown in Example 2.3.1. Consequently, the kernel K_ω is strictly positive definite as the sum of a p.d. and s.p.d. kernel, see Proposition 2.3.2. Furthermore, we have the following properties for the RKHS of K_ω:

Proposition 4.1.7. *Let $\omega : \Omega \to \mathbb{R}^{m\times m}$, $\omega \succ 0$ be a weight function and assume that K is continuous.*

(a) *Then the native space $\mathcal{H}_\omega \coloneqq \mathcal{H}_{K_\omega}$ of the modified kernel $K_\omega = K + \delta_\omega$ is given by*

$$\mathcal{H}_\omega = \mathcal{H}_K \oplus \mathcal{H}_{\delta_\omega}.$$

(b) *For all $f \in \mathcal{H}_\omega$ there exist unique $g \in \mathcal{H}_K$ and $h \in \mathcal{H}_{\delta_\omega}$ such that*

$$f = g + h \qquad \textit{and} \qquad \|f\|^2_{\mathcal{H}_\omega} = \|g\|^2_{\mathcal{H}_K} + \|h\|^2_{\mathcal{H}_{\delta_\omega}}.$$

Proof. By Corollary 2.3.4 we know that $\mathcal{H}_\omega = \mathcal{H}_K + \mathcal{H}_{\delta_\omega}$ and the norm on $\mathcal{H}_\delta$ is given by

$$\|f\|^2_{\mathcal{H}_\omega} = \min\{\|g\|^2_{\mathcal{H}_K} + \|h\|^2_{\mathcal{H}_{\delta_\omega}} \mid f = g + h\}.$$

The RKHS $\mathcal{H}_{\delta_\omega}$ consists of elements of the form

$$h = \sum_{i\in I} \omega(x_i)\alpha_i\delta_{x_i}(\cdot)$$

for some countable set $I \subset \mathbb{N}$. Since Ω is uncountable, it follows that $\mathcal{H}_{\delta_\omega}$ contains no continuous function except for $h = 0$. Since K is continuous by assumption, we have that all $g \in \mathcal{H}_K$ are continuous by Theorem 2.2.19. Therefore we have $\mathcal{H}_K \cap \mathcal{H}_{\delta_\omega} = \{0\}$. Hence, the sum of the two RKHS is direct and every element $f \in \mathcal{H}_\omega$ has a unique decomposition $f = g + h$ with $g \in \mathcal{H}_K$, $h \in \mathcal{H}_{\delta_\omega}$. □

For any finite set $X = \{x_1, \dots, x_n\} \subset \Omega$, we can now connect the regularized interpolant $\mathcal{I}^\omega_X(f)$ with the best approximation, i.e. interpolant, of $f \in \mathcal{H}_K \subset \mathcal{H}_\omega$ in the subspace $\mathcal{N}_\omega \coloneqq \mathcal{N}_{K_\omega}(X)$.

Corollary 4.1.8.

(a) *Let $\alpha \in \mathbb{R}^{mn}$ be given by $K_\omega(X,X)^{-1}f(X)$, then we have*

$$\mathcal{I}^\omega_X(f) = K(\cdot, X)\alpha \qquad \textit{and} \qquad \Pi_{\mathcal{N}_\omega(X)} f = K_\omega(\cdot, X)\alpha.$$

(b) *It holds for any $x \notin X$*

$$\mathcal{I}_X^\omega(f)(x) = \Pi_{\mathcal{N}_\omega(X)} f(x).$$

Proof. The first statement follows directly from the definition of the regularized interpolant $\mathcal{I}_X^\omega(f)$ and the best approximation $\Pi_{\mathcal{N}_\omega(X)} f$. The second statement can easily be concluded from

$$K(x,X) = K(x,X) + \underbrace{\delta_\omega(x,X)}_{=0} = K_\omega(x,X), \qquad \text{for all} x \notin X.$$

□

4.2 Error estimation for regularized interpolation

4.2.1 The regularized Power function

Analogous to the Power function (matrix) of Section 2.5, the error in the approximation via the regularized interpolation can be quantified by a modified version of the Power function.

Definition 4.2.1 (Regularized Power Function)**.** Let $X \subset \Omega$, $\omega \succ 0$ be a weight function and $\mathcal{I}_X^\omega : \mathcal{H}_K \to \mathcal{N}(X)$ the regularized interpolation operator. Then the regularized Power function $\mathcal{Q}^\omega_{\mathcal{N}(X),K} : \mathcal{H}_K' \to \mathbb{R}$ is given by

$$\mathcal{Q}^\omega_{\mathcal{N}(X),K}(\lambda) = \sup_{f \in \mathcal{H}_K} \frac{\lambda\left(f - \mathcal{I}_X^\omega(f)\right)}{\|f\|_{\mathcal{H}_K}} = \|\lambda \circ (\mathrm{id} - \mathcal{I}_X^\omega)\|_{\mathcal{H}_K'}. \tag{4.10}$$

For the Power function, we have seen in Lemma 2.5.2 that it can be expressed via K and the reproducing kernel $K_{\mathcal{N}(X)}$ of the subspace $\mathcal{N}(X)$. One essential ingredient for this observation is the fact that the orthogonal projection operator $\Pi_{\mathcal{N}(X)} : \mathcal{H}_K \to \mathcal{H}_K$ is self-adjoint. The same holds true for the regularized interpolation operator $\mathcal{I}_X^\omega$. To see this, let $f, g \in \mathcal{H}_K$. By Theorem 4.1.3 and the reproduction property we have

$$\begin{aligned}\langle \mathcal{I}_X^\omega(f), g\rangle_{\mathcal{H}_K} &= \langle K(\cdot,X)(A+W)^{-1} f(X), g\rangle_{\mathcal{H}_K} = g(X)^T (A+W)^{-1} f(X) \\ &= \langle f, K(\cdot,X)(A+W)^{-1} g(X)\rangle_{\mathcal{H}_K} = \langle f, \mathcal{I}_X^\omega(g)\rangle_{\mathcal{H}_K},\end{aligned}$$

where we used the abbreviations $A = K(X,X)$ and $W = \delta_\omega(X,X)$. This leads to an alternative representation of $\mathcal{Q}^\omega_{\mathcal{N}(X),K}$ involving the weighted l_2–Lebesgue function given in (4.8).

Lemma 4.2.2. *For the regularized Power function $\mathcal{Q}^{\omega}_{\mathcal{N}(X),K} : \mathcal{H}'_K \to \mathbb{R}$ it holds*

$$\mathcal{Q}^{\omega}_{\mathcal{N}(X),K}(\lambda)^2 = \lambda^1\lambda^2 K - \lambda^1 K(\cdot, X)(A+W)^{-1}\lambda^2 K(X,\cdot) - \Psi^{\omega}_X(\lambda)^2$$

where $A = K(X,X)$ and $W = \delta_\omega(X,X)$.

Proof. Due to the self-adjointness of $\mathcal{I}^{\omega}_X$ it follows that

$$\begin{aligned}
\mathcal{Q}^{\omega}_{\mathcal{N}(X),K}(\lambda)^2 &= \left\|\lambda^2 K - \mathcal{I}^{\omega}_X(\lambda^2 K)\right\|^2_{\mathcal{H}_K} \\
&= \lambda^1\lambda^2 K - 2\lambda^1 K(\cdot, X)(A+W)^{-1}\lambda^2 K(X,\cdot) \\
&\quad + \lambda^1 K(\cdot, X)(A+W)^{-1}A(A+W)^{-1}\lambda^2 K(X,\cdot) \\
&= \lambda^1\lambda^2 K - \lambda^1 K(\cdot, X)(A+W)^{-1}\lambda^2 K(X,\cdot) \\
&\quad - \lambda^1 K(\cdot, X)(A+W)^{-1}W(A+W)^{-1}\lambda^2 K(X,\cdot) \\
&= \lambda^1\lambda^2 K - \lambda^1 K(\cdot, X)(A+W)^{-1}\lambda^2 K(X,\cdot) - \lambda(L^{\omega}_X)W\lambda(L^{\omega}_X)^T \\
&= \lambda^1\lambda^2 K - \lambda^1 K(\cdot, X)(A+W)^{-1}\lambda^2 K(X,\cdot) - \Psi^{\omega}_X(\lambda)^2
\end{aligned}$$

□

Using Lemma 4.2.2, we can now define an extension analogous to the Power function matrix, which we likewise name regularized Power function matrix.

Definition 4.2.3 (Regularized Power Function Matrix and weighted Lebesgue Matrix)**.** Let $\Lambda \subset \mathcal{H}'_K$ be a finite collection of linear functionals, $|\Lambda| = p$. The weighted Lebesgue matrix $\boldsymbol{\Psi}^{\omega}_X(\Lambda) \in \mathbb{R}^{p\times p}$ corresponding to Λ is the unique symmetric matrix such that for all functionals of the form $\lambda = \alpha^T S_\Lambda$ with $\alpha \in \mathbb{R}^p$ we have

$$\Psi^{\omega}_X(\lambda)^2 = \alpha^T \boldsymbol{\Psi}^{\omega}_X(\Lambda)\alpha.$$

The regularized Power function matrix corresponding to Λ is then given by

$$\boldsymbol{Q}^{\omega}_{\mathcal{N}(X),K}(\Lambda) = S^1_\Lambda S^2_\Lambda K - S^1_\Lambda K(\cdot, X)(A+W)^{-1}S^2_\Lambda K(X,\cdot) - \boldsymbol{\Psi}^{\omega}_X(\Lambda).$$

In the case where the sampling operator S_Λ coincides with a point evaluation, i.e. $S_\Lambda(f) = f(x)$ for some $x \in \Omega$, we might also use the abbreviations

$$\boldsymbol{\Psi}^{\omega}_X(x) \coloneqq \boldsymbol{\Psi}^{\omega}_X(\Lambda) \quad \text{and} \quad \boldsymbol{Q}^{\omega}_{\mathcal{N}(X),K}(x) \coloneqq \boldsymbol{Q}^{\omega}_{\mathcal{N}(X),K}(\Lambda).$$

Analogously we can now express the error between any function f and its regularized interpolant by means of the regularized Power function (matrix) by following the same steps as in the proofs of Theorem 2.5.3 and Corollary 2.5.5.

Theorem 4.2.4 (Bounds on the regularized interpolation error)**.** *For any $\lambda \in \mathcal{H}_K'$ we have*

$$|\lambda(f - \mathcal{I}_X^\omega(f))| \leq \mathcal{Q}_{\mathcal{N}(X),K}^\omega(\lambda) \, \|f\|_{\mathcal{H}_K} . \tag{4.11}$$

Furthermore, for any $x \in \Omega$ it holds

(a) $\|f(x) - \mathcal{I}_X^\omega(f)(x)\|_2^2 \leq \lambda_{\max}(\boldsymbol{Q}_{\mathcal{N}(X),K}^\omega(x)) \, \|f\|_{\mathcal{H}_K}^2$

(b) $\|f(x) - \mathcal{I}_X^\omega(f)(x)\|_\infty^2 \leq \max \operatorname{diag}(\boldsymbol{Q}_{\mathcal{N}(X),K}^\omega(x)) \, \|f\|_{\mathcal{H}_K}^2$

(c) $\|f(x) - \mathcal{I}_X^\omega(f)(x)\|_1^2 \leq \operatorname{tr}(\boldsymbol{Q}_{\mathcal{N}(X),K}^\omega(x)) \, \|f\|_{\mathcal{H}_K}^2$

Remark 4.2.5. For scalar-valued kernels, constant weight function $\omega(x) = \omega_0$, i.e. kernel ridge regression, and point evaluation, the bound on the regularized Power function matrix reduces to

$$\mathcal{Q}_{\mathcal{N}(X),K}^\omega(x)^2 = K(x,x) - K(x,X)(A + \omega_0 I)^{-1} K(X,x) - \omega_0 K(x,X)(A + \omega_0)^{-2} K(X,x)$$

from which we recover the previously known bound presented in [29].

Next we want to study the relation between the Power function matrices $\boldsymbol{P}_{\mathcal{N}(X),K}$, $\boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}$ and the regularized Power function matrix $\boldsymbol{Q}_{\mathcal{N}(X),K}^\omega$. While both $\boldsymbol{P}_{\mathcal{N}(X),K}$ and $\boldsymbol{Q}_{\mathcal{N}(X),K}^\omega$ operate over the functional space $\mathcal{H}_K'$, for the Power function matrix $\boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}$ it is in general not obvious if arbitrary sets of functionals $\Lambda \subset \mathcal{H}_K'$ can be used. This stems from the fact that not all functionals are applicable to the delta kernel δ_ω, for example any functional involving derivation. Fortunately, by Proposition 4.1.7 we have $\mathcal{H}_\omega = \mathcal{H}_K \oplus \mathcal{H}_{\delta_\omega}$ and therefore

$$\mathcal{H}_\omega' = \mathcal{H}_K' \oplus \mathcal{H}_{\delta_\omega}'.$$

Consequently, we can identify each $\lambda \in \mathcal{H}_K'$ with $\lambda \notin \mathcal{H}_{\delta_\omega}$ as an element of $\mathcal{H}_\omega'$ as follows. Let $f = g + h \in \mathcal{H}_\omega$ with $g \in \mathcal{H}_K$ and $h \in \mathcal{H}_{\delta_\omega}$. By Proposition 4.1.7 this decomposition is unique and hence we can define an extension $\tilde{\lambda} \in \mathcal{H}_\omega'$ of λ via

$$\tilde{\lambda}(f) = \lambda(g).$$

In other words, we can apply any functional $\lambda \in \mathcal{H}_K'$ with $\lambda \notin \mathcal{H}_{\delta_\omega}'$ by setting

$$\lambda(h) = 0$$

for all $h \in \mathcal{H}_{\delta_\omega}$. With this we can order the different (regularized) Power function matrices as follows.

Theorem 4.2.6 (Ordering of (regularized) Power function matrices)**.**
Let $\Lambda = \{\lambda_1, \dots, \lambda_p\} \subset \mathcal{H}'_K$ be a finite collection of functionals such that for any $\alpha \in \mathbb{R}^p$, $\lambda = \alpha^T S_\Lambda$ the kernel satisfies

$$\lambda^2 K \notin \mathcal{N}(X). \tag{4.12}$$

Then it holds

$$\boldsymbol{P}_{\mathcal{N}(X),K}(\Lambda) \preceq \boldsymbol{Q}^\omega_{\mathcal{N}(X),K}(\Lambda) \preceq \boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}(\Lambda). \tag{4.13}$$

Furthermore, we have

$$\boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}(\Lambda) = \boldsymbol{Q}^\omega_{\mathcal{N}(X),K}(\Lambda) + S^1_\Lambda S^2_\Lambda \delta_\omega + \boldsymbol{\Psi}^\omega_X(\Lambda). \tag{4.14}$$

Proof. Let $\lambda = \alpha^T S_\Lambda$ for some $\alpha \in \mathbb{R}^p$. By definition of the Power function and the regularized Power function, and the self-adjointness of the respective approximation operators we have

$$\begin{aligned}
\mathcal{P}_{\mathcal{N}(X),K}(\lambda)^2 &= \left\|\lambda \circ (\mathrm{id} - \Pi_{\mathcal{N}(X)})\right\|_{\mathcal{H}'_K} \\
&= \left\|\left(\lambda \circ (\mathrm{id} - \Pi_{\mathcal{N}(X)})\right)^2 K\right\|^2_{\mathcal{H}_K} = \left\|(\mathrm{id} - \Pi_{\mathcal{N}(X)})\lambda^2 K\right\|^2_{\mathcal{H}_K} \\
&\leq \left\|(\mathrm{id} - \mathcal{I}^\omega_X)\lambda^2 K\right\|^2_{\mathcal{H}_K} = \left\|(\lambda \circ (\mathrm{id} - \mathcal{I}^\omega_X))^2 K\right\|^2_{\mathcal{H}_K} \\
&= \|\lambda \circ (\mathrm{id} - \mathcal{I}^\omega_X)\|_{\mathcal{H}'_K} = \mathcal{Q}^\omega_{\mathcal{N}(X),K}(\lambda).
\end{aligned}$$

This gives us the left inequality in (4.13). For the right hand side it is sufficient to show the identity given in (4.14), since

$$S^1_\Lambda S^2_\Lambda \delta_\omega + \boldsymbol{\Psi}^\omega_X(\Lambda) \succeq 0.$$

By assumption on Λ we have $S^2_\Lambda \delta_\omega(X, \cdot) = 0$ and therefore

$$S^2_\Lambda K(X, \cdot) = S^2_\Lambda K_\omega(X, \cdot).$$

Combining this with the result of Lemma 4.2.2 we have, for $A = K(X, X)$ and $W =$

$\delta_\omega(X, X)$,

$$\begin{aligned}\boldsymbol{Q}^\omega_{\mathcal{N}(X),K}(\Lambda) &= S^1_\Lambda S^2_\Lambda K - S^1_\Lambda K(\cdot, X)(A+W)^{-1} S^2_\Lambda K(X,\cdot) - \boldsymbol{\Psi}^\omega_X(\Lambda) \\ &= S^1_\Lambda S^2_\Lambda K + S^1_\Lambda S^2_\Lambda \delta_\omega - S^1_\Lambda K_\omega(\cdot, X)(A+W)^{-1} S^2_\Lambda K_\omega(X,\cdot) \\ &\quad - \left(S^1_\Lambda S^2_\Lambda \delta_\omega + \boldsymbol{\Psi}^\omega_X(\Lambda)\right) \\ &= \boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}(\Lambda) - \left(S^1_\Lambda S^2_\Lambda \delta_\omega + \boldsymbol{\Psi}^\omega_X(\Lambda)\right).\end{aligned}$$

Solving for $\boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}(\Lambda)$ gives the desired result. □

The condition (4.12) is necessary, since the Power functions $\mathcal{P}_{\mathcal{N}(X),K}$ and $\mathcal{P}_{\mathcal{N}(X),K_\omega}$ vanish for every $\lambda \in \mathcal{H}'_K$ such that $\lambda^2 K \in \mathcal{N}(X)$. However, this is in general not the case for the regularized Power function. In the case of function evaluation, i.e. $S_\Lambda(f) = f(x)$, the condition simplifies to $x \notin X$.

4.2.2 Bounds on the regularized Power function and weighted Lebesgue function

By definition, the regularized Power function gives the smallest value $C(\lambda)$ such that a bound of the form

$$|\lambda(f - \mathcal{I}^\omega_X(f))| \le C(\lambda)\, \|f\|_{\mathcal{H}_K} \tag{4.15}$$

is satisfied for all functions $f \in \mathcal{H}_K$. Therefore, any function $C(\lambda)$ which provides a bound of the form (4.15) results in an upper bound on $\mathcal{Q}^\omega_{\mathcal{N}(X),K}(\lambda)$. This mimics the same behaviour we have seen for the Power function and we will use similar ideas as in Section 2.5 to derive bounds on $\mathcal{Q}^\omega_{\mathcal{N}(X),K}(\lambda)$. For this purpose, we now again consider kernels for which the native space is norm equivalent to a Sobolev space $W^k(\Omega)$ for some $k > d/2$. We once again make use of sampling inequalities to derive bounds on the error for the regularized approximation. These bounds extend the result in [81, 109] in which scalar-valued kernels and constant weight functions were considered.

Theorem 4.2.7. *Let $X \subset \Omega$ and assume that Ω satisfies an interior cone condition. Then there exists a constant $C > 0$ such that for all $f \in \mathcal{H}_K$ and for any multiindex $\beta \in \mathbb{N}^d_0$ such that $k > |\beta| + d/2$ it holds*

$$\sup_{x\in\Omega} \left\| D^\beta (f - \mathcal{I}^\omega_X(f))(x) \right\|_2 \le C h_{X,\Omega}^{-|\beta|} \left(h_{X,\Omega}^{k-d/2} + \sqrt{\omega_{\max}} \right) \|f\|_{\mathcal{H}_K}$$

for sufficiently small fill distances $h_{X,\Omega}$, where $\omega_{\max} = \sup_{x\in\Omega} \lambda_{\max}(\omega(x))$.

Proof. Using Theorem 2.5.8 we get

$$\sup_{x\in\Omega}\left\|D^\beta(f-\mathcal{I}_X^\omega(f))(x)\right\|_2 \le Ch_{X,\Omega}^{-|\beta|}\left(h_{X,\Omega}^{k-d/2}\,\|f\|_{\mathcal{H}_K}\cdot+\|f(X)-\mathcal{I}_X^\omega(f)(X)\|_\infty\right)$$

It is therefore sufficient to show that

$$\|f(X)-\mathcal{I}_X^\omega(f)(X)\|_\infty \le \sqrt{\omega_{\max}}\,\|f\|_{\mathcal{H}_K}.$$

By definition of the regularized interpolation operator we have

$$\mathcal{J}_{X,f(X)}^\omega(\mathcal{I}_X^\omega(f)) = \min_{g\in\mathcal{H}_K}\mathcal{J}_{X,f(X)}^\omega(g) \le \mathcal{J}_{X,f(X)}^\omega(f) = \|f\|_{\mathcal{H}_K}^2.$$

and therefore

$$\begin{aligned}\|f(X)-\mathcal{I}_X^\omega(f)(X)\|_\infty^2 \le \|f(X)-\mathcal{I}_X^\omega(f)(X)\|_2^2 &= \left\|W^{1/2}W^{-1/2}(f(X)-\mathcal{I}_X^\omega(f)(X))\right\|_2^2\\ &\le \omega_{\max}\left\|W^{-1/2}(f(X)-\mathcal{I}_X^\omega(f)(X))\right\|_2^2\\ &\le \omega_{\max}\mathcal{J}_{X,f(X)}^\omega(\mathcal{I}_X^\omega) \le \omega_{\max}\,\|f\|_{\mathcal{H}_K}^2\end{aligned}$$

□

This immediately translates into a bound on the regularized Power function.

Corollary 4.2.8. *Let the assumptions of Theorem 4.2.7 hold, then we have for $\lambda=\delta_x^\alpha\circ D^\beta$*

$$\mathcal{Q}_{\mathcal{N}(X),K}(\lambda) \le Ch_{X,\Omega}^{-|\beta|}\left(h_{X,\Omega}^{k-d/2}+\sqrt{\omega_{\max}}\right)\|\alpha\|_2.$$

Using the identity (4.14), we can now bound the Power function $\mathcal{P}_{\mathcal{N}(X),K_\omega}$. However, since (4.14) involves the weighted Lebesgue function, we first derive a bound for it as in intermediate step.

Lemma 4.2.9. *Le the assumptions of Theorem 4.2.7 hold. For $\lambda=\delta_x^\alpha\circ D^\beta$ the bound*

$$\Psi_X^\omega(\lambda) \le Ch_{X,\Omega}^{-|\beta|}\left(h_{X,\Omega}^{\frac{2k-d}{2}}+2\sqrt{\omega_{\max}}\right)\|\alpha\|$$

holds.

Proof. Since $\mathcal{I}_X^\omega$ minimizes $\mathcal{J}_{X,f(X)}^\omega$ we have

$$\left\|W^{-1/2}\left(f(X)-\mathcal{I}_X^\omega(f)(X)\right)\right\|_2^2+\|\mathcal{I}_X^\omega(f)\|_{\mathcal{H}_K}^2 = \mathcal{J}_{X,f(X)}^\omega \le \mathcal{J}_{X,f(X)}^\omega(0) = \left\|W^{-1/2}f(X)\right\|_2^2.$$

In particular, both terms on the outer left hand side are bounded by the term on the outer right hand side. It follows that

$$\begin{aligned}\|\mathcal{I}_X^\omega(f)(X)\|_\infty &\leq \sqrt{\omega_{\max}}\left\|W^{-1/2}\mathcal{I}_X^\omega(f)(X)\right\|_2\\ &\leq \sqrt{\omega_{\max}}\left(\left\|W^{-1/2}(\mathcal{I}_X^\omega(f)(X)-f(X))\right\|_2+\left\|W^{-1/2}f(X)\right\|_2\right)\\ &\leq 2\sqrt{\omega_{\max}}\left\|W^{-1/2}f(X)\right\|_2.\end{aligned}$$

Using Theorem 2.5.8 we get

$$\begin{aligned}|D^\beta(\mathcal{I}_X^\omega(f))(x)^T\alpha| &\leq Ch_{X,\Omega}^{-|\beta|}\left(h_{X,\Omega}^{\frac{2k-d}{2}}\|\mathcal{I}_X^\omega(f)\|_{\mathcal{H}_K}+\|\mathcal{I}_X^\omega(f)(X)\|_\infty\right)\\ &\leq Ch_{X,\Omega}^{-|\beta|}\left(h_{X,\Omega}^{\frac{2k-d}{2}}+2\sqrt{\omega_{\max}}\right)\left\|W^{-1/2}f(X)\right\|_2.\end{aligned}$$

Since this holds true for all $f \in \mathcal{H}_K$, we get the desired result via the definition of the weighted Lebesgue function

$$\Psi_X^\omega(\lambda)=\sup_{f\in\mathcal{H}_K}\frac{|D^\beta(\mathcal{I}_X^\omega(f))(x)^T\alpha|}{\|W^{-1/2}f(X)\|_2}\leq Ch_{X,\Omega}^{-|\beta|}\left(h_{X,\Omega}^{\frac{2k-d}{2}}+2\sqrt{\omega_{\max}}\right).$$

□

Finally, we can conclude from Theorem 4.2.6, Corollary 4.2.8 and Lemma 4.2.9:

Corollary 4.2.10. *Let the assumptions of Theorem 4.2.7 hold. Then we have for any $x \in \Omega$:*

$$\left(\lambda_{\max}\left(\boldsymbol{P}_{\mathcal{N}(X),K_\omega}(x)\right)\right)^{1/2}\leq 2C\left(h_{X,\Omega}^{\frac{2k-d}{2}}+2\sqrt{\omega_{\max}}\right).$$

If ω is chosen such that $\omega_{\max} \leq h_{X,\Omega}^{2k-d}$ for all $x \in \Omega$, then we recover the same approximation rates as for a pure interpolation procedure, albeit with a slightly bigger constant.

4.3 Greedy Point Selection for Regularized Interpolation

Similar to the approximation based on interpolation on the data sites $X \subset \Omega$ which where considered in Chapter 3, one can ask the question how a suitable set of points X can be selected. While Theorem 4.2.7 provides us with a bound on the error with respect to the fill distance $h_{X,\Omega}$ of the set X, it is in general not clear how the points should be selected if the domain Ω is oddly shaped. Taking Section 3.4 into consideration, we have already seen how greedy algorithms can be used to select suitable sequences $\{X_n\}_{n\in\mathbb{N}}$ such that we can reach the same asymptotic rates as provided by the Kolmogorov n–width, see

(3.7), of the set $\mathcal{F} = \{K(\cdot, x)\alpha \mid x \in \Omega, \alpha \in \mathbb{R}^m, \|\alpha\|_2 = 1\}$. In the following we will use similar methods as in Section 3.2 to achieve slightly modified results.

Unfortunately, a direct approach, where we simply substitute the Power function matrix with the regularized Power function matrix in Algorithm 3 does not work. From Theorem 4.2.6 equation (4.14) we already know that the regularized Power function matrix satisfies

$$\boldsymbol{Q}_X^\omega(x) = \boldsymbol{P}_{\mathcal{N}(X),K_\omega}(x) - \omega(x) - \boldsymbol{\Psi}_X^\omega(x).$$

While $\omega(x)$ is easy to evaluate and $\boldsymbol{P}_{\mathcal{N}(X),K_\omega}(x)$ can be efficiently updated via the Newton basis, see Theorem 3.2.3, when enriching the set X, we were not able to find a way to easily update $\boldsymbol{\Psi}_X^\omega(x)$. Therefore, the regularized Power function matrix needs to be computed from scratch in every iteration of the greedy algorithm. The second reason as to why a straightforward substitution is not a good choice is the fact that we are unable to show that the resulting greedy algorithm is weak in the sense of Definition 3.3.1. Hence, we lack an underlying theoretical basis to indicate whether this greedy procedure would perform well or not. Nonetheless, Theorem 4.2.6 gives us a bound on $\boldsymbol{Q}_X^\omega(x)$ in terms of $\boldsymbol{P}_{\mathcal{N}(X),K_\omega}(x)$ and since K_ω is a s.p.d. kernel, we can formulate a regularized P–Greedy algorithm by making use of the Power function matrix for K_ω:

Algorithm 5: Regularized P-greedy Algorithm

```
Data: finite sampling of the input domain $\Omega_N \subset \Omega$, kernel $K : \Omega \times \Omega \to \mathbb{R}^{m \times m}$,
      empty initial set $X = \emptyset$, weight function $\omega : \Omega \to \mathbb{R}^{m \times m}$, error indicator
      function $E$, tolerance $\varepsilon > 0$.
Result: Point set $X$
while $\max_{x \in \Omega_N} E(\boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}(x)) \geq \varepsilon$ do
    $x^* = \arg\max_{x \in \Omega_N} E(\boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega}(x))$;
    $X = X \cup \{x^*\}$;
end
```

In other words, the regularized P–Greedy algorithm is just a P–Greedy algorithm for the modified kernel K_ω. However, since we add each maximizer x^* to the set X in each iteration, the Algorithm 5 employs the "full" extension routine. One might also consider the "eig" or "diag" extension routine, which would in general result in approximation space that do not have the form $\mathcal{N}(X)$ and the regularized interpolation operator can be modified to cover these cases. For the indicator function E each function $E_i, i \in \{1, 2, \infty\}$ of (3.3) can be used since they all result in a weak greedy, as was seen in Proposition 3.4.1.

We now consider sequences of sets $(X_n)_{n\in\mathbb{N}}$ which are generated by the regularized P–Greedy algorithm. Making use of the results in Theorem 3.3.3 and Corollary 3.3.4 we can obtain bounds on the decay of $\lambda_{\max}(\boldsymbol{P}_{\mathcal{N}_\omega,K_\omega}(x))$. To apply the afforementioned results, we first need to bound the Kolmogorov n–width of the set

$$\mathcal{F}_\omega := \{K_\omega(\cdot,x)\alpha \mid x\in\Omega, \alpha\in\mathbb{R}^m, \|\alpha\|_2 = 1\}.$$

Unlike in the setting of Section 3.4, the kernel K_ω is not continuous and therefore the set $\mathcal{F}_\omega$ is not compact. Consequently, the Kolmogorov n–width does not converge to 0 as n tends to infinity. Nonetheless, we can obtain bounds similar to Lemma 3.4.3 by adding a constant term which depends only on the weight function ω.

Lemma 4.3.1. *Let $K:\Omega\times\Omega\to\mathbb{R}^{m\times m}$ be the reproducing Kernel of a Hilbert space $\mathcal{H}_K$ which is norm equivalent to $W^k(\Omega)$ for some $k>d/2$ and let $\omega:\Omega\to\mathbb{R}^{m\times m}$ be a weight function. Then there exists a constant C such that the Kolmogorov n–width of $\mathcal{F}_\omega$ is bounded by*

$$d_n(\mathcal{F}_\omega) \leq C\left(n^{-\frac{2k-d}{2d}} + \sqrt{\omega_{\max}}\right).$$

Proof. Let $X=\{x_1,\dots,x_n\}\subset\Omega$ be a set of quasi uniformly distributed points, i.e. there exists a constant $c_1>0$ such that

$$h_{X,\Omega} \leq c_1 n^{-1/d}.$$

By Corollary 4.2.10 we have

$$\begin{aligned} d_{nm}(\mathcal{F}_\omega) &\leq \sup_{x\in\Omega}\left(\lambda_{\max}(\boldsymbol{P}_{\mathcal{N}_\omega(X),K_\omega})(x)\right) \leq c_2 h_{X,\Omega}^{(2k-d)/2} + c_2\sqrt{\omega_{\max}} \\ &\leq c_1c_2 n^{-\frac{2k-d}{2d}} + c_2\sqrt{\omega_{\max}}. \end{aligned}$$

Proceeding analogous as in the proof of Corollary 3.3.5 we get

$$d_n(\mathcal{F}_\omega) \leq c_1c_2(2m)^{\frac{2k-d}{2d}} n^{-\frac{2k-d}{2d}} + c_2\sqrt{\omega_{\max}} \leq C\left(n^{-\frac{2k-d}{2d}} + \sqrt{\omega_{\max}}\right).$$

□

In summary, we have bounds of the form

$$d_n(\mathcal{F}_\omega) \leq C(n^{-(2k-d)/2d} + \eta), \tag{4.16}$$

where η depends on the weight function and the constant C scales with the output space dimension via the factor $m^{(2k-d)/2d}$. A modified version of Corollary 3.3.4 shows that the same rates can be achieved, when using weak greedy algorithms. This gives us the following results on the sequence of sets which are selected by the regularized P–Greedy algorithms and consequently on the error for the regularized interpolation procedure.

Corollary 4.3.2. *The regularized P–Greedy algorithm generates a sequence of sets $(X_n)_{n\in\mathbb{N}}$ such that*

$$\sup_{x\in\Omega}\Big(\lambda_{\max}(\boldsymbol{Q}^{\omega}_{(X_n)}(x))\Big) \leq \sup_{x\in\Omega}\Big(\lambda_{\max}(\boldsymbol{P}_{\mathcal{N}_\omega(X_n),K_\omega})(x)\Big) \leq C\left(n^{-\frac{2k-d}{2d}} + \sqrt{\omega_{\max}}\right)$$

Proof. Let γ denote the weak greedy constant for the respective indicator function according to Proposition 3.4.1 and let

$$\sigma_n = \sigma_n(\mathcal{F}_\omega) = \Big(\lambda_{\max}(\boldsymbol{P}_{\mathcal{N}_\omega(X_n),K_\omega})(x)\Big).$$

Furthermore, let $\alpha, C_0, \eta > 0$ such that $d_n = d_n(\mathcal{F}) \leq C_0\left(n^{-\alpha}+\eta\right)$ and define $C = 2^{5\alpha+1}\gamma^{-2}C_0$. Proceeding analogously to Corollary 3.3.4 it is sufficient to show that

$$\sigma_n \leq C\left(n^{-\alpha}+\eta\right). \tag{4.17}$$

Let us now assume to the contrary that $M \in \mathbb{N}$ is the smallest natural number such that $\sigma_M > Cm^{\alpha}M^{-\alpha} + C\eta$. We first consider the case $M = 4s$. Following the proof of Corollary 3.3.4 we know from (3.11) that

$$\sigma_{4s} \leq \sqrt{2}\gamma^{-1}\sqrt{\sigma_{2s}d_s}$$

and therefore

$$C\left((4s)^{-\alpha}+\eta\right) < \sqrt{2}\sqrt{C\left((2s)^{-\alpha}+\eta\right)C_0\left(s^{-\alpha}+\eta\right)}.$$

Solving for C we obtain

$$C < \frac{2C_0\gamma^{-2}\left(2^{-\alpha}+\eta s^{\alpha}\right)\left(1+\eta s^{\alpha}\right)}{\left(4^{-\alpha}+\eta s^{\alpha}\right)^2} =: h(s).$$

However, since

$$h'(s) = -C_0\gamma^{-2}\frac{2^{3\alpha+1}(2^{\alpha}-1)\alpha\eta s^{\alpha-1}\left(4^{\alpha}\eta s^{\alpha}+2^{\alpha+1}\eta s^{\alpha}+2^{\alpha+1}+1\right)}{\left(4^{\alpha}\eta s^{\alpha}+1\right)^3} < 0$$

for $s \geq 0$ we have

$$C < h(0) = 2C_0\gamma^{-2}2^{-\alpha}4^{2\alpha} = 2^{3\alpha+1}\gamma^{-2}C_0 \leq C$$

and we reached a contradiction. In the case of $M = 4s + q$ with $q \in \{1, 2, 3\}$ one can analogously reach a contradiction. □

Theorem 4.2.7 gives us a bound in terms of the fill distance. Likewise, we can infer information about the fill distance from an error bound, similar to what was done in Lemma 3.4.5:

Lemma 4.3.3. *Let $X \subset \Omega$. If there exists a constant $C > 0$ such that*

$$\sup_{x\in\Omega} \|f(x) - \mathcal{I}_X^\omega(f)(x)\|_2 \leq \epsilon \|f\|_{\mathcal{H}_K}$$

for all $f \in \mathcal{H}_K$ and, then the fill distance is bounded by

$$h_{X,\Omega} \leq C\epsilon^{\frac{2}{2k-d}}.$$

Proof. The proof follows the one in [110] with slight modification to account for the altered approximation operator $\mathcal{I}_X^\omega$ and the constant η. From [25] we know, that there exist a bump function f with support in the unit ball and $\sup_{x\in\Omega} \|f(x)\|_2 = 1$, such that

$$\left\| f\left(\frac{\cdot}{h_{X,\Omega}}\right) \right\|_{\mathcal{H}_K} \leq ch_{X,\Omega}^{\frac{d-2k}{2}} \|f\|_{\mathcal{H}_K}.$$

Let $f_h := f\left(\frac{\cdot}{h_{X,\Omega}}\right)$, then we can place f_h between the points in X such that $f_h(X) = 0$ and consequently $\mathcal{I}_X^\omega(f) = 0$. We now conclude that

$$1 = \sup_{x\in\Omega} \|f_h(x)\|_2 = \sup_{x\in\Omega} \|f_h(x) - \mathcal{I}_X^\omega(f_h)(x)\|_2 \leq \epsilon \|f_h\|_{\mathcal{H}_K} \leq c\epsilon h_{X,\Omega}^{\frac{d-2k}{2}} \|f\|_{\mathcal{H}_K}.$$

Solving for the fill distance we get

$$h_{X,\Omega} \leq \left(\frac{c\|f\|_{\mathcal{H}_K}}{1-\eta}\right)^{\frac{2}{2k-d}} \epsilon^{\frac{2}{2k-d}} \leq \left(2c\|f\|_{\mathcal{H}_K}\right)^{\frac{2}{2k-d}} \epsilon^{\frac{2}{2k-d}} \leq C\epsilon^{\frac{2}{2k-d}}.$$

□

Combining Corollary 4.3.2 and Lemma 4.3.3 we ultimately get

Corollary 4.3.4. *The sequence of sets $(X_n)_{n\in\mathbb{N}}$ generated by the regularized P–Greedy* *algorithm satisfies for any* $\beta \in \mathbb{N}_0^d$ *such that* $|\beta| < k - d/2$

$$\sup_{x\in\Omega} \left\| D^\beta (f - \mathcal{I}_X^\omega(f))(x) \right\|_2 \leq C \left(n^{-\frac{2k-d}{2d}} + \sqrt{\omega_{\max}} \right)^{\frac{-2|\beta|}{2k-d}} \left(n^{-\frac{2k-d}{2d}} + \sqrt{\omega_{\max}} \right) \|f\|_{\mathcal{H}_K} .$$

Proof. Corollary 4.3.2 and Lemma 4.3.3 we have

$$h_{X_n,\Omega} \leq C \left(n^{-\frac{2k-d}{2d}} + \sqrt{\omega_{\max}} \right)^{\frac{2}{2k-d}} .$$

Using this fill distance in Theorem 4.2.7 we obtain the claim. □

4.4 Numerical Investigation

We consider the so called thermal block example, which is well known in the reduced basis community (cf. [39, 72]). It consist of a stationary heat transfer problem on the unit square $\mathcal{D} = (0,1)^2$ which is divided into a number of subblocks, here $\mathcal{D} = \bigcup_{i=1}^{4} B_i$ as illustrated in Figure 4.1, with possibly different heat conductivities on each subblock. In our case we consider the same conductivity $\mu_1 \in [1, 10]$ for B_1 and B_4 and $\mu_2 \in [1, 10]$ for B_2 and B_3, i.e. the square is divided into a 2×2 checkered grid. We prescribe a unit flux into the domain on the bottom boundary, which is denoted as $\Gamma_{N,1}$ with unit outward normal $n(\xi)$, where $\xi \in \mathcal{D}$ indicates the spatial variable. The left and right boundary part $\Gamma_{N,0}$ is insulated, which is modeled by a zero Neumann boundary condition and the top Dirichlet boundary Γ_D has constant 0 temperature.

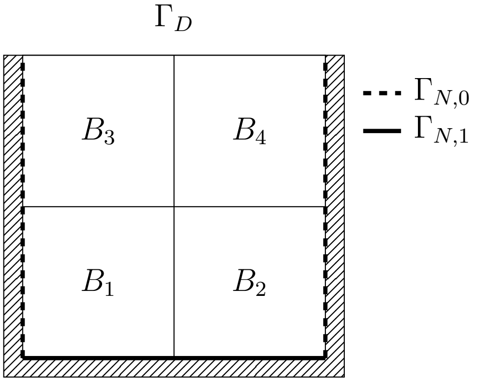

Figure 4.1: Illustration of the thermal block model.

The problem is then governed by the the equations

$$\begin{aligned} -\nabla \cdot (\kappa(\xi;\mu)\nabla u(\xi;\mu)) &= 0, && \xi \in \Omega, \\ u(\xi;\mu) &= 0, && \xi \in \Gamma_D, \\ (\kappa(\xi;\mu)\nabla u(\xi;\mu)) \cdot n(\xi) &= i, && \xi \in \Gamma_{N,i}, i = 0,1, \end{aligned}$$

where we define the heat conductivity function

$$\kappa(\cdot;\mu) : \mathcal{D} \to \mathbb{R}_+, \quad \kappa(\xi;\mu) := \sum_{i=1}^{B} \mu_i \chi_{B_i}(\xi),$$

using the indicator function χ_{B_i} for the subblocks $B_i \subset \mathcal{D}$. The above system is now discretized by linear finite elements, which results in a 420×420 dimensional system. We now want to approximate the target function $f : \Omega = [1,10]^2 \to \mathbb{R}^{420}$, which maps a pair of heat conductivities to its corresponding finite element solution.

All of the following experiments were implemented in MATLAB 2018a and run on a machine with an Intel Core i7-7500U CPU with 16GB Ram.

We then consider the two kernels

$$K_1(x,y) = \phi_{3,2}(0.125 \left\|x-y\right\|_2) I_{420} \qquad \text{and} \qquad K_2(x,y) = \phi_{3,2}(0.125 \left\|x-y\right\|_2) ZZ^T,$$

where the columns of $Z \in \mathbb{R}^{420 \times 8}$ contains the 8 singular vectors corresponding to the 8 largest singular values for a singular value decomposition of 1000 random random target function evaluations and $\phi_{3,2}$ denotes the Wendland function

$$\phi_{3,2}(r) = \frac{1}{3}(3 + 18r + 35r^2)(1-r)_x^6.$$

By Corollary 2.3.18, the RKHS can then be identified with $W^2(\Omega, \mathbb{R}^{420})$ and
$W^2(\Omega, \mathrm{range}(Z))$, respectively. We further consider the following three weight functions

$$\omega_i(x) = 10^{-7} \log\left(1 + \frac{5000}{x_1 x_2}\right) Q_i,$$

where $Q_1 = I_{420}$, and Q_2, Q_3 are the L^2 and H_0^1 gramian matrix, respectively. We further discretize our input domain into a 200×200 uniform grid resulting in a set $\Omega_N \subset \Omega$ consisting of 40000 points. We then run the P–Greedy algorithm with a trace indicator function and a full extension routine for the kernel K_1, as well as the regularized P–Greedy algorithm for the kernel $K_1 + \omega_1$ on the set Ω_N until 500 points are selected, resulting in the greedy point set X_{K_1} and X_{K_1,ω_1}. The decay of the respective indicator functions is depicted in Figure 4.2 and we can observe that they behave roughly in the same way with only slight differences. This is due to the fact, that the value of the indicator for the regularized P–Greedy is larger than the maximum value of the weight function at $\omega_{1.\max} \approx 8.5 \cdot 10^{-7}$.

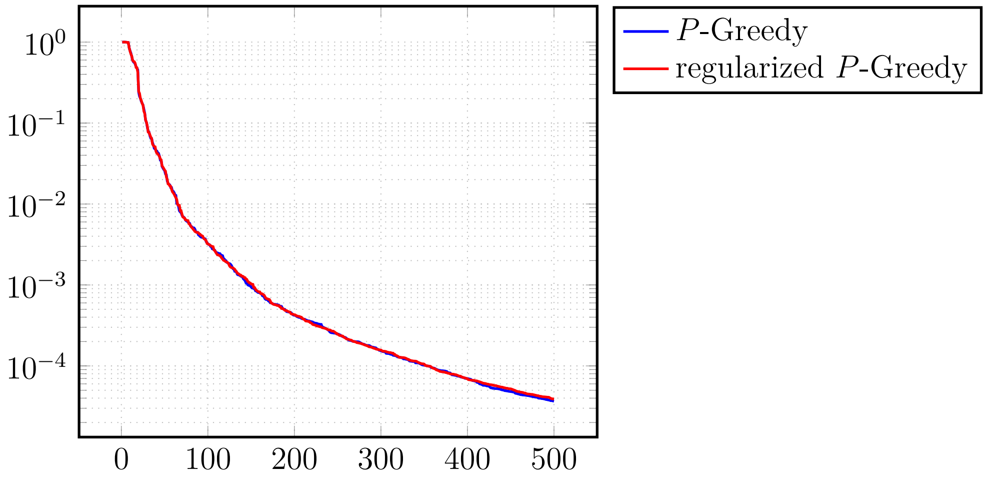

Figure 4.2: Decay of the indicator function for the P–Greedy and regularized P–Greedy.

Using the point set X_{K_1} and X_{K_1,ω_1}, we compute both the the interpolant as well as the regularized interpolant and evaluate the L^2 and H_0^1 errors on the discrete set Ω_N. The pointwise maximum errors are shown in Table 4.1. In both cases, we can see that greedy set X_{K_1,ω_1} generated by the regularized P–Greedy algorithm results in smaller errors. This is most likely caused by the subtle influences of the weight function which puts larger emphasis on smaller heat conductivities for which the stationary heat transfer problem is more difficult.

Point set	Approximation	L^2 error	H_0^1 error
X_{K_1}	Interpolation	2.89E-03	5.50E-03
	Reg. Interpolation	7.56E-03	1.36-02
X_{K_1,ω_1}	Interpolation	1.26E-03	3.51E-03
	Reg. Interpolation	6.79E-03	1.23E-02

Table 4.1: Maximum Pointwise L^2 and H_0^1 error for the different Greedy point sets and approximation schemes.

Similar to the above, we also perform the regularized P–Greedy algorithm on the set Ω_N for all combinations of the second kernel K_2 and the three weight functions $\omega_1, \omega_2, \omega_3$. This results in three point sets $X_{K_2,\omega_1}, X_{K_2,\omega_2}$ and X_{K_2,ω_3} of size 500, respectively. Please note, that by Corollary 2.4.16 there exists an uncoupled separable decomposition for the modified kernels $K_2(x, y) + \omega_i(x)\delta_x(y)$. Consequently, the regularized P–Greedy algorithm can be performed efficiently. The decay of the corresponding indicator functions is depicted in Figure 4.3 and we can observe a clear difference in the behaviour of the decay.

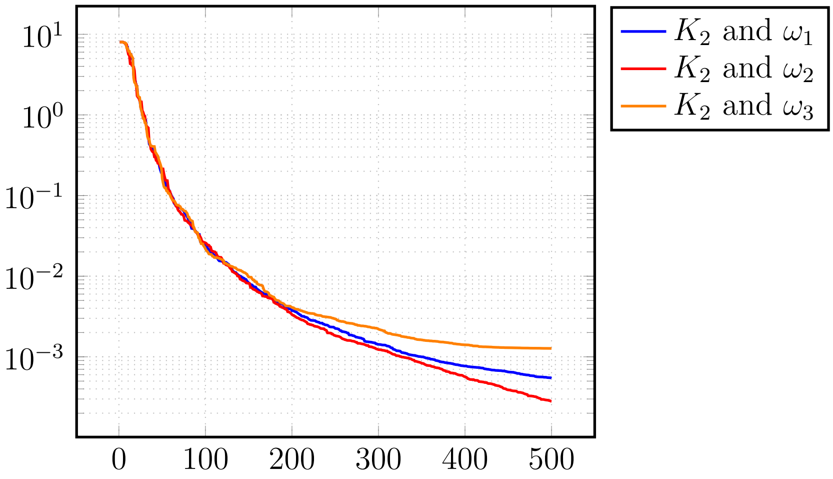

Figure 4.3: Decay of the indicator function for the regularized P–Greedy for the kernel K_2 in combination with the three weight functions

By Theorem 4.2.6, in particular equation (4.14), we know that each indicator function is bounded from below by $\operatorname{tr}(\omega_i)$ and thus, the indicator function will start to plateau at some point. For our specific choice of weight functions, we have

$$\begin{aligned} 1.6 \cdot 10^{-4} &\leq \operatorname{tr}(\omega_1(x)) \leq 3.6 \cdot 10^{-4} \\ 1.9 \cdot 10^{-7} &\leq \operatorname{tr}(\omega_2(x)) \leq 4.2 \cdot 10^{-7} \\ 6.1 \cdot 10^{-4} &\leq \operatorname{tr}(\omega_3(x)) \leq 1.4 \cdot 10^{-3}. \end{aligned}$$

We can observe the plateauing effect for the weight function ω_3. The closer the indicator value gets to the trace of the weight function, the larger the influence of the weight function on the selected points is. We can see this in Figure 4.6 as the points begin to cluster in the lower left corner, the region in which ω_3 takes its largest values. The distributions of the greedy points X_{K_2,ω_1} and X_{K_2,ω_2} are displayed in Figure 4.4 and Figure 4.5, respectively. In contrast to X_{K_2,ω_3}, no clustering occurs for the former two as the indicator function value is still sufficiently far away from the trace of the weight function.

By Lemma 3.4.6 we know, that the P–Greedy algorithm applied to the kernel K_1 results in greedy point sets, for which the fill distance decays with a rate of $n^{-1/2}$. Hence, in light of Theorem 4.2.7, using the points X_{K_1} for constructing the regularized interpolant should already provide quasi-optimal results. Thus, one may ask the question if using the regularized P–Greedy provides additional benefit compared to the P–Greedy. For this purpose, we compute the regularized interpolant for the point sets X_{K_1} and X_{K_2,ω_i} and evaluate the maximum error in both the L^2 and H_0^1 norm on the set Ω_N. The maximum pointwise errors are displayed in Table 4.2. On the one side, for the weight functions ω_1 and ω_2, we can observe an improvement in the quality of the approximation, with errors

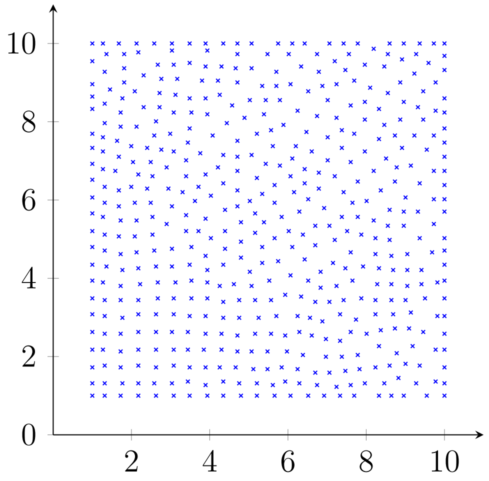

Figure 4.4: Distribution of the greedy point set X_{K_2,ω_1}

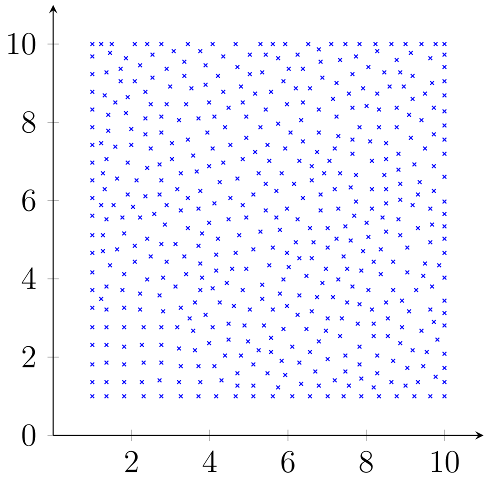

Figure 4.5: Distribution of the greedy point set X_{K_2,ω_2}

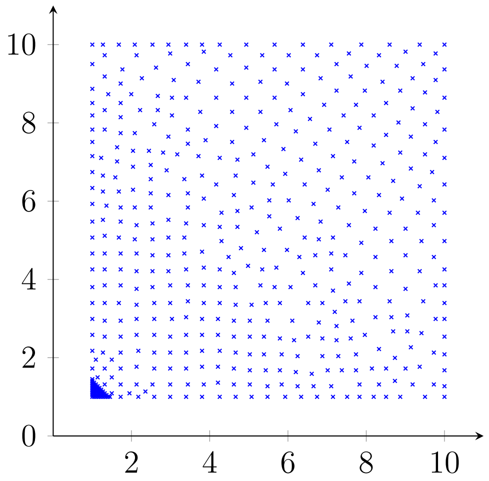

Figure 4.6: Distribution of the greedy point set X_{K_2,ω_3}

Weight	Points	L^2 error	H_0^1 error
ω_1	X_{K_2,ω_1}	1.02E-03	3.27E-03
	X_{K_1}	2.57E-03	4.67E-03
ω_2	X_{K_2,ω_2}	9.91E-04	3.22E-03
	X_{K_1}	3.50E-03	6.41E-03
ω_3	X_{K_2,ω_3}	9.91E-04	3.22E-03
	X_{K_1}	6.41E-03	2.22E-03

Table 4.2: Maximum L^2 and H_0^1 error for the regularized approximant with different weights and on different point sets.

improving upto a factor of 2 in both norms. On the other side, for the weight function ω_3, the quality of the approximant slightly deteriorates. However, this is reasonable, as we observed clustering of the points in X_{K_2,ω_3}.

5 Rigorous and effective a-posteriori error bounds for nonlinear problems

Most of this chapter has previously been published in [89]. However, we present further analysis of the presented bounds. In particular, we were able to find improved bounds on the effectivity of our estimators (see Lemma 5.3.6) without further requirements than the original ones. Consequently the following analytical results were modified to include this improved bound on the effectivity. Furthermore, we extend the methodology presented in the original work to linear time invariant systems.

5.1 Motivation

In many disciplines of applied mathematics, a-posteriori error estimates, i.e. error estimates which depend on the approximation itself, are important tools. Such estimates can be used to assess the quality of a numerical approximation scheme and to indicate whether or not the corresponding approximation is feasible for the respective problem. A common use of such a-posteriori error bounds are adaptive refinement strategies, where the error estimates are used to judge if further refinement, i.e. improvement of the approximation, is required. Examples include temporal or spatial discretization refinement when solving partial differential equation (PDE) with a numerical scheme [2, 70, 28]. In most cases such an estimator should satisfy two conditions: First, the estimator should be rigorous, i.e. it should be a valid upper bound on the error. In this case, we call it an error bound instead of an error estimator. This difference is visualised in Figure 5.1.

The second property is called effectivity. This notion stipulates that the factor of overestimation should be computable. In particular, an effective error bound detects if the error is zero, as otherwise, the quotient between error bound and actual error is not defined. In the context of the finite element method (FEM), these properties are also refered to as reliability and efficiency.

When dealing with numerical approximation schemes for PDEs, the corresponding numerical schemes quickly become computationally complex as they involve the solution of a high-dimensional system of equations. Several techniques have emerged which try to cir-

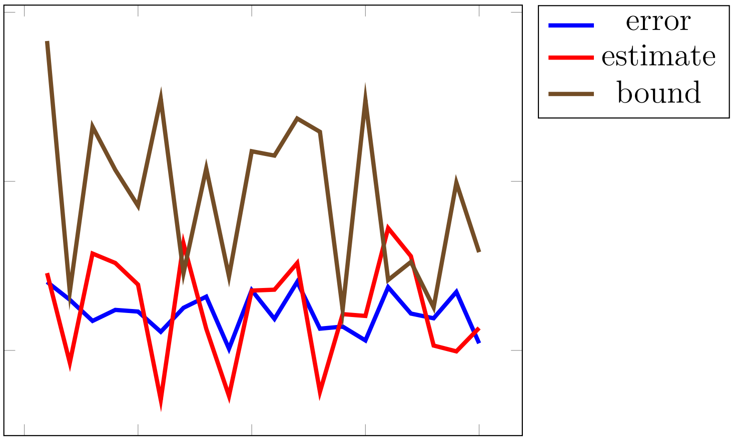

Figure 5.1: Illustration of the conceptual difference between an error estimate and an error bound.

cumvent the corresponding rise in computational cost for these high-dimensional system. The techniques we focus on in this part of the thesis can be encompassed by the framework of reduced-order modelling. Here the main idea is to replace the high-dimensional problem with a low-dimensional surrogate which should be computationally less expensive to solve. This is usually achieved by projecting the high-dimensional problem onto a low-dimensional subspace. The question then arises how the error introduced by this surrogate can be quantified. This is precisely where rigorous and effective a-posteriori error bounds are indispensable.

Giving a complete overview over the available methods and corresponding results for the error estimation is not the focus of this thesis. Instead, we refer to [10] and [11] for recent overviews of model (order) reduction in the parametric and non-parametric cases. Based on these techniques, approximate solutions can be calculated cheaply and in a computationally efficient manner. One framework that is particularly suitable for parametric problems is the reduced basis (RB) method. The essential idea of RB methods is to identify low-dimensional subspaces in the high-dimensional solution spaces by exploring the parameter domain which can for example be achieved by RB greedy algorithms. In this thesis we will demonstrate that classical error bounds, which are well-established within RB methods, can be significantly improved by introducing an auxiliary linear problem and subsequently computing its corresponding RB approximation. By the proposed procedure, we show that almost optimal effectivities, i.e. effectivities close to 1 and hence almost exact error prediction, can be reached both theoretically as well as verifiable in many numerical examples. Additionally, the necessity for good lower bounds on the inf–sup constant, which is commonly used in many of the standard error estimation theory for RB problems, can be alleviated, which allows the use of rougher (and thus computationally less expensive) lower bounds. Furthermore, the quality of the error bound can be tuned according to the application requirements.

In this work, we improve a residual-based error estimation technique, developed in [17], that has been used frequently during the last decades. We illustrate the essential idea of the improvement that enables highly accurate error estimates by the following simple example: Consider the system of linear equations

$$Ax = f, \qquad \text{for } A = \begin{pmatrix} 1 & 0 \\ 10 & 1 \end{pmatrix} \qquad \text{and} \qquad f = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The above equation has the unique solution $x^* = (0,1)^T$. Let us now assume that we have a numerical scheme that is provides us with an approximation $\hat{x}$ that satisfies $\hat{x} = 1.01x^*$. This results in an error of $\|\hat{x} - x^*\|_2 = 0.01$, when measured in the Euclidean norm. Generally, the true solution x^* is not available and hence we require an error bound. The standard approach for this is by defining a residual $r = A\hat{x} - f$ and realizing that the error $e = \hat{x} - x^*$ satisfies the equation

$$Ae = r.$$

We can now conclude that the error is bounded via

$$\|e\|_2 \le \left\|A^{-1}\right\|_2 \|r\|_2 \approx 10.1 \cdot 0.01 = 0.101.$$

Compared to the actual error this is an overestimation of around factor 10. To obtain a bound of higher quality we first note, that we can obtain the error by solving the equation $Ae = r$. Unfortunately, an exact computation of this error equation is as expensive as the original problem and therefore should be avoided. However, similar to the original problem, we assume that a numerical scheme is available which provides an approximation $\hat{e}$ that also satisfies $\hat{e} = 1.01e$. We can now bound the error in terms of this approximation by using the triangle inequality, i.e.

$$\|e\|_2 \le \|\hat{e}\|_2 + \|\hat{e} - e\|_2 .$$

The second term can now be bounded analogously to the first bound by introducing a second residual $R = A\hat{e} - r$. However, the norm of the second residual is much smaller than the norm of the original residual and we easily verfy that

$$\|e\|_2 \le \|\hat{e}\|_2 + \left\|A^{-1}\right\|_2 \|R\|_2 \le 0.0101 + 10.1 \cdot 0.0001 = 0.0111. \tag{5.1}$$

We can easily see that this second bound only overestimates the actual error by a factor of 1.111. Hence, the error bound was improved by a factor of almost 10.

In this chapter, we show how the key idea behind the above example can be generalized to a large class of linear, nonlinear as well as time-dependent problems. In particular, we study the applicability in the context of RB methods. The chapter is structured as follows:

In Section 5.2 we introduce the problem setting which we will use in the subsequent sections. In Section 5.3 we introduce a generic error bound for both linear and nonlinear problems, which arises as a refinement of results given in [17]. In Section 5.4 we show how highly effective error bounds can be reached by introducing an auxiliary linear problem that has to be solved or approximately solved, respectively. In Section 5.5 we investigate how our error bound fits into the RB framework and how our bound constitutes a significant improvement on previous results in this field. In Section 5.6 we validate our previous theoretical results for a well-known linear test problem, the thermal block model, as well as a nonlinear problem stemming from a nonlinear reaction-diffusion-advection equation. In Section 5.7 we extend our previous results to the case of linear time-invariant (LTI)-systems including a numerical analysis of the theoretical results.

5.2 Rigorous and effective error bounds

For this section we always assume that $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces with norm $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$, respectively. We further denote as $\mathcal{L}(\mathcal{X},\mathcal{Y})$ the space of all bounded linear operators mapping from $\mathcal{X}$ into $\mathcal{Y}$, and equip it with the norm given by

$$\|\mathcal{A}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} := \sup_{0\neq x\in\mathcal{X}} \frac{\|\mathcal{A}x\|_{\mathcal{Y}}}{\|x\|_{\mathcal{X}}}$$

for any $\mathcal{A} \in \mathcal{L}(\mathcal{X},\mathcal{Y})$.

In this part we are interested in problems which can be described as a root finding problem of some continuously differentiable operator $G : \mathcal{X} \to \mathcal{Y}$. Specifically, the problem takes the form

$$\text{Find } x \in X \text{ such that } G(x) = 0 \in \mathcal{Y}. \tag{P}$$

An element $x^* \in \mathcal{X}$ is now called a (true) solution of the problem (P), if $G(x^*) = 0$.

In the following, we always assume that at least one solution of (P) exists. We are interested in estimating the error $e := \hat{x} - x^*$ between the (true) solution $x^* \in \mathcal{X}$ and a fitting approximation $\hat{x} \in \mathcal{X}$ using reliable a-posteriori error bounds, which can be represented by functions $\Delta : \mathcal{X} \to \mathbb{R}$. The estimate is then given by evaluating the a-posteriori bound at the approximate solution, i.e.

$$\|e\|_{\mathcal{X}} = \|\hat{x} - x^*\|_{\mathcal{X}} \leq \Delta(\hat{x}). \tag{5.2}$$

The quality of an upper bound Δ can be assessed in term of its effectivity, which – for $\hat{x} \neq x^*$ – is given by

$$\mathrm{eff}_{\Delta}(\hat{x}) := \frac{\Delta(\hat{x})}{\|x^* - \hat{x}\|_{\mathcal{X}}}. \tag{5.3}$$

By definition it is clear that a reliable error bound satisfies $\mathrm{eff}(\hat{x}) \geq 1$. Since $\mathrm{eff}(\hat{x}) = 1$ is equivalent to exact error prediction we aim for error bounds whose effectivities are close to one.

A general framework for providing such error estimates for problems of the type (P) can be found in [17]. However, the presented results might lead to large overestimations of the actual error similar to what we have seen in the toy problem at the beginning of this section.

5.3 Rigorous, effective and computable a-posteriori error estimates with effectivity bounds

In this section we refine the results derived in [17] and show how significant improvements can be achieved. We want to emphasize that these derivations are independent of the specific approximation method used to compute an approximate solution $\hat{x}$. Due to the assumption $G \in C^1(\mathcal{X}, \mathcal{Y})$, G has a bounded Fréchet-derivative $\mathrm{D}G|_x \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ for any $x \in \mathcal{X}$. In particular, this is the case for $x = \hat{x}$. We further assume, that the derivate $\mathrm{D}G|_{\hat{x}}$ is invertible, which implies by means of the bounded inverse theorem that $\mathrm{D}G|_{\hat{x}}^{-1} \in \mathcal{L}(\mathcal{Y}, \mathcal{X})$. Therefore, we are able to define the following three quantities, where $\overline{B_\alpha}(\hat{x}) = \{x \in \mathcal{X} | \ \|x - \hat{x}\|_{\mathcal{X}} \leq \alpha\}$ denotes the closed ball in $\mathcal{X}$ with radius α around $\hat{x}$

$$\begin{aligned}
\epsilon(\hat{x}) &:= \left\| \mathrm{D}G|_{\hat{x}}^{-1} \left(G(\hat{x})\right) \right\|_{\mathcal{X}}, && \text{(non-split residual)}\\
\gamma(\hat{x}) &:= \left\| \mathrm{D}G|_{\hat{x}}^{-1} \right\|_{\mathcal{L}(Y,X)}, && \text{(stability constant)}\\
L(\alpha) &:= \sup_{x \in \overline{B_\alpha}(\hat{x})} \|\mathrm{D}G|_x - \mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}, && \text{(local nonlinearity indicator).}
\end{aligned}$$

Please note that $\epsilon(\hat{x})$ is not a residual of the problem (P) in the classical sense. Nonetheless, we call it the non-split residual, as it is the residual of a modified problem

$$\text{Find } x \in X \text{ such that } \left.\mathrm{D}G\right|_{\hat{x}}^{-1}(G(x)) = 0 \in \mathcal{X}$$

that has the same solution as the original problem.

These three quantities make the main ingredients for the following fundamental error estimate:

Theorem 5.3.1 (Rigorous a-posteriori error estimation)**.** *Let $\hat{x} \in \mathcal{X}$ be an approximate solution and assume that $\left.\mathrm{D}G\right|_{\hat{x}} : \mathcal{X} \to \mathcal{Y}$ is invertible.*

(a) *If the validity criterion*

$$\tau(\hat{x}) := 2\gamma(\hat{x})L(2\epsilon(\hat{x})) \leq 1$$

is met, then the problem $G(x) = 0$ has a unique solution $x^* \in \mathcal{X}$ in the closed ball $\overline{B_{2\epsilon(\hat{x})}}(\hat{x})$ and the following upper bound for the error $e = \hat{x} - x^* \in \mathcal{X}$ holds

$$\|e\|_{\mathcal{X}} = \|\hat{x} - x^*\|_{\mathcal{X}} \leq \Delta(\hat{x}) := \frac{1}{1 - \tau(\hat{x})/2}\epsilon(\hat{x}) \leq 2\epsilon(\hat{x}). \tag{5.4}$$

(b) *If $L(\alpha) \leq C\alpha$ for some $C > 0$ and if the modified validity criterion*

$$\hat{\tau}(\hat{x}) := 4\gamma(\hat{x})C\epsilon(\hat{x}) \leq 1$$

is satisfied, then the problem $G(x) = 0$ has a unique solution $x^* \in X$ in the closed ball $\overline{B_{2\epsilon(\hat{x})}}(\hat{x})$ and the error $e = \hat{x} - x^* \in \mathcal{X}$ is bounded by

$$\|e\|_{\mathcal{X}} = \|\hat{x} - x^*\|_{\mathcal{X}} \leq \hat{\Delta}(\hat{x}) := \frac{1 - \sqrt{1 - \hat{\tau}(\hat{x})}}{2\gamma(\hat{x})C} = \frac{2}{1 + \sqrt{1 - \hat{\tau}(\hat{x})}}\epsilon(\hat{x}). \tag{5.5}$$

Proof. We first note, that if $L(\alpha) \leq C\alpha$ for some $C > 0$, we have

$$\tau(\hat{x}) := 2\gamma(\hat{x})L(2\epsilon(\hat{x})) \leq 4\gamma(\hat{x})C\epsilon(\hat{x}) = \hat{\tau}(\hat{x}).$$

Therefore, if the modified validity criterion is met, the same holds for the validity criterion. In particular, we have

$$2\gamma(\hat{x})L(2\epsilon(\hat{x})) \leq 1$$

in both cases.

By applying the fundamental theorem of calculus we derive the identity

$$G(x) - G(x') = \int\limits_0^1 \mathrm{D}G|_{x'+t(x-x')}\,(x-x')\mathrm{d}t, \qquad x, x' \in \mathcal{X}. \tag{5.6}$$

Let $H : \mathcal{X} \to \mathcal{X}$ be defined via

$$H(x) := x - \mathrm{D}G|_{\hat{x}}^{-1}\,(G(x)).$$

It is easy to see that

$$G(x) = 0 \quad \Longleftrightarrow \quad H(x) = x$$

and thus it remains to show, that H has a fix-point in $M := \overline{B_{2\epsilon(\hat{x})}}(\hat{x})$. Hence, it is sufficient to show that $H(M) \subset M$ and that the restriction $H|_M$ is a contraction. To this end let $x \in \mathcal{X}$ and it now holds by (5.6) that

$$\begin{aligned}\|H(x_1) - H(x_2)\|_{\mathcal{X}} &= \left\|\mathrm{D}G|_{\hat{x}}^{-1}\,(\mathrm{D}G|_{\hat{x}}\,(x_1 - x_2) - (G(x_1) - G(x_2)))\right\|_{\mathcal{X}} \\ &= \left\|\mathrm{D}G|_{\hat{x}}^{-1}\left(\int_0^1 (\mathrm{D}G|_{\hat{x}} - \mathrm{D}G|_{x_1+t(x_2-x_1)})(x_1 - x_2)\mathrm{d}t\right)\right\|_{\mathcal{X}} \\ &\leq \gamma(\hat{x})L(2\epsilon(\hat{x}))\,\|x_1 - x_2\| \leq \frac{1}{2}\,\|x_1 - x_2\|\,,\end{aligned} \tag{5.7}$$

which proves the contraction property. Furthermore, it holds

$$\begin{aligned}\|H(x) - \hat{x}\|_{\mathcal{X}} &= \left\|x - \mathrm{D}G|_{\hat{x}}^{-1}\,(G(x)) - \hat{x}\right\|_X \\ &= \left\|\mathrm{D}G|_{\hat{x}}^{-1}\,[\mathrm{D}G|_{\hat{x}}\,(x - \hat{x}) - (G(x) - G(\hat{x}))] - \mathrm{D}G|_{\hat{x}}^{-1}\,(G(\hat{x}))\right\|_X \\ &= \left\|\mathrm{D}G|_{\hat{x}}^{-1}\left[\int_0^1 (\mathrm{D}G|_{\hat{x}} - \mathrm{D}G|_{\hat{x}+t(x-\hat{x})})(x - \hat{x})\mathrm{d}t\right] - \mathrm{D}G|_{\hat{x}}^{-1}\,(G(\hat{x}))\right\|_X\end{aligned}$$

and since $\hat{x} + t(x - \hat{x}) \in M$ for $t \in [0, 1]$ we get the estimate

$$\begin{aligned}\|H(x) - \hat{x}\|_{\mathcal{X}} &\leq \gamma(\hat{x}) \sup_{z\in M} \|\mathrm{D}G|_{\hat{x}} - \mathrm{D}G|_z\|_{\mathcal{L}(X,Y)}\,\|z - \hat{x}\|_{\mathcal{X}} + \epsilon(\hat{x}) \\ &\leq 2\gamma(\hat{x})L(2\epsilon(\hat{x}))\epsilon(\hat{x}) + \epsilon(\hat{x}).\end{aligned}$$

(a) If the validity criterion is met, we can immediately conclude that

$$\|H(x) - \hat{x}\|_{\mathcal{X}} \leq 2\gamma(\hat{x})L(2\epsilon(\hat{x}))\epsilon(\hat{x}) + \epsilon(\hat{x}) \leq 2\epsilon(\hat{x})$$

and $H|_M$ is a self-mapping. We can now apply Banach's fixed-point theorem which proves the existence of an $x^* \in M$ with $G(x^*) = 0$. Furthermore, for any $x \in M$ we get

$$\begin{aligned}\|x^* - x\|_{\mathcal{X}} &= \|H(x^*) - x\|_{\mathcal{X}} \\ &= \left\| \mathrm{D}G|_{\hat{x}}^{-1} \left(-G(x) + \int_0^1 (\mathrm{D}G|_{\hat{x}} - \mathrm{D}G|_{x^*+t(x-x^*)})(x - x^*)\,\mathrm{d}t \right) \right\|_{\mathcal{X}} . \\ &\leq \left\| \mathrm{D}G|_{\hat{x}}^{-1} (G(x)) \right\|_{\mathcal{X}} + \gamma(\hat{x}) L(2\epsilon(\hat{x})) \|x^* - x\|_{\mathcal{X}} .\end{aligned}$$

The choice of $x = \hat{x}$ and solving for x^* ultimately results in

$$\|x^* - \hat{x}\| \leq \frac{\epsilon(\hat{x})}{1 - \gamma(\hat{x})L(2\epsilon(\hat{x})} \leq 2\epsilon(\hat{x}).$$

(b) If the modified validity criterion is met, the set M can be replaced by the set $M_\alpha = \overline{B_\alpha}(\hat{x})$ and we can now try to find the minimal α, such that H is a self-mapping on M_α. Continuing the previous series of inequalities we get

$$\begin{aligned}\|H(x) - \hat{x}\| &\leq \gamma(\hat{x}) \sup_{z \in M_\alpha} \|\mathrm{D}G|_{\hat{x}} - \mathrm{D}G|_z\|_{\mathcal{L}(X,Y)} \|z - \hat{x}\|_{\mathcal{X}} + \epsilon(\hat{x}) \\ &\leq \gamma(\hat{x})L(\alpha)\alpha + \epsilon(\hat{x}) \leq \gamma(\hat{x})C\alpha^2 + \epsilon(\hat{x}) \overset{!}{\leq} \alpha.\end{aligned}$$

Solving the resulting quadratic inequality, we have that α is cointained in the interval $[\alpha_-, \alpha_+]$, where

$$\alpha_\pm = \frac{1 \pm \sqrt{1 - 4\gamma(\hat{x})C\epsilon(\hat{x})}}{2\gamma(\hat{x})C} = \frac{2}{1 \mp \sqrt{1 - 4\gamma(\hat{x})C\epsilon(\hat{x})}}\epsilon(\hat{x}) = \frac{2}{1 \mp \sqrt{1 - \hat{\tau}(\hat{x})}}\epsilon(\hat{x}).$$

Hence, the smallest α for which H is a self-mapping is given by α_-. Finally, it follows that

$$\|x^* - \hat{x}\|_{\mathcal{X}} \leq \alpha_- = \frac{2}{1 - \sqrt{1 + \hat{\tau}(\hat{x})}}\epsilon(\hat{x}).$$

□

Remark 5.3.2. Note that similar bounds have been derived by various authors ([88, 103, 95]). However, in the bounds in literature known to us, the non-split residual $\epsilon(\hat{x})$ is replaced by the upper bound

$$\epsilon_{\text{split}}(\hat{x}) := \gamma(\hat{x}) \left\| G(\hat{x}) \right\|_{\mathcal{Y}} \geq \left\| \mathrm{D}G|_{\hat{x}}^{-1} \left(G(\hat{x}) \right) \right\|_{\mathcal{X}} = \epsilon(\hat{x}), \tag{5.8}$$

which we call split residual for obvious reasons. As we have seen in the introduction and as we will see in the numerical results, this splitting can induce a very large overestimation. This is not the case when the quantity $\epsilon(\hat{x})$ or other, more accurate approximations to it, are used. Hence, the results in this thesis might improve many of the aforementioned existing results.

As the name suggests, the local nonlinearity indicator $L(\alpha)$ can be interpreted as a measure for the nonlinearity of the function G in a neighbourhood of the approximate solution $\hat{x} \in \mathcal{X}$. In particular, for affine linear problems $G(x) = Ax + b$ for some $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $b \in \mathcal{Y}$ we get $\mathrm{D}G|_{\hat{x}} = A$ and therefore

$$L(\alpha) = \sup_{x \in \overline{B_\alpha}(\hat{x})} \left\| \mathrm{D}G|_x - \mathrm{D}G|_{\hat{x}} \right\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} = \left\| A - A \right\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} = 0.$$

Hence, L detects the linearity of the problem. As a consequence, we get $\tau(\hat{x}) = \hat{\tau}(\hat{x}) = 0$, i.e. unconditional validity, which results in exact error prediction as stated in the following corollary.

Corollary 5.3.3 (Exact error prediction for affine linear problems)**.** *Let G be affine linear in x. Then it holds*

$$\left\| e \right\|_{\mathcal{X}} = \left\| \hat{x} - x^* \right\|_{\mathcal{X}} = \Delta(\hat{x}), \quad \textit{and} \quad \mathrm{eff}(\hat{x}) = 1.$$

Proof. Since G is affine linear in x it can be written as $G(x) = Ax+b$ for some $A \in \mathcal{L}(X, Y)$ and $g \in \mathcal{Y}$. We then obtain $G(\hat{x}) = G(\hat{x}) - G(x^*) = A(\hat{x} - x^*) = Ae$ or equivalently $e = A^{-1}(G(\hat{x}))$. We further infer

$$\left\| e \right\|_{\mathcal{X}} = \left\| A^{-1}(G(\hat{x})) \right\|_{\mathcal{X}} = \left\| \mathrm{D}G|_{\hat{x}}^{-1} \left(G(\hat{x}) \right) \right\|_{\mathcal{X}} = \epsilon(\hat{x}) = \Delta(\hat{x}),$$

since $\mathrm{D}G|_x = A$ for all $x \in \mathcal{X}$ and $\tau(\hat{x}) = 0$. □

As we have mentioned in the introduction, the key quantity that assesses the quality of the error bound is the effectivity $\mathrm{eff}(\hat{x})$. The assumptions made in Theorem 5.3.1 are already sufficient to obtain bounds on the effectivity of our approximates which depend on the condition number

$$\kappa(\hat{x}) := \left\| \mathrm{D}G|_{\hat{x}}^{-1} \right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \left\| \mathrm{D}G|_{\hat{x}} \right\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} = \gamma(\hat{x}) \left\| \mathrm{D}G|_{\hat{x}} \right\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$$

of the linearization around the approximate solution.

Theorem 5.3.4 (Generic effectivity bound)**.** *Let the assumptions of either Theorem 5.3.1* ***(a)*** *or Theorem 5.3.1* ***(b)*** *hold. Then the effectivity* $\operatorname{eff}(\hat{x})$ *of the error bound is bounded as follows*

(a) $\operatorname{eff}_\Delta(\hat{x}) \le \dfrac{\kappa(\hat{x})}{1-\tau(\hat{x})/2} + \dfrac{\tau(\hat{x})}{2-\tau(\hat{x})}$

(b) $\operatorname{eff}_{\hat{\Delta}}(\hat{x}) \le \dfrac{2\kappa(\hat{x})}{1+\sqrt{1-\hat{\tau}(\hat{x})}} + \dfrac{\hat{\tau}(\hat{x})}{2+2\sqrt{1-\hat{\tau}(\hat{x})}}$

Proof. Since G is continuously differentiable on X we have by the mean-value theorem for Fréchet-differentiable functions

$$\begin{aligned}
\epsilon(\hat{x}) &= \left\| \mathrm{D}G|_{\hat{x}}^{-1}\left(G(\hat{x})\right)\right\|_{\mathcal{X}} = \left\| \mathrm{D}G|_{\hat{x}}^{-1}\left(G(\hat{x}) - G(x^*)\right)\right\|_{\mathcal{X}} \\
&\le \left\| \mathrm{D}G|_{\hat{x}}^{-1}\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \|G(\hat{x}) - G(x^*)\|_{\mathcal{Y}} \le \gamma(\hat{x}) \sup_{x\in\overline{B_\alpha}(2\epsilon(\hat{x}))} \|\mathrm{D}G|_x\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \|\hat{x}-x^*\|_{\mathcal{X}} \\
&\le \gamma(\hat{x})\left(\|\mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + L(2\epsilon(\hat{x})\right)\|\hat{x}-x^*\|_{\mathcal{X}} \\
&= \left(\gamma(\hat{x})\|\mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + \gamma(\hat{x})L(2\epsilon(\hat{x}))\right)\|\hat{x}-x^*\|_{\mathcal{X}}.
\end{aligned}$$

We conclude

(a)

$$\begin{aligned}
\epsilon(\hat{x}) &\le \left(\gamma(\hat{x})\|\mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + \gamma(\hat{x})L(2\epsilon(\hat{x}))\right)\|\hat{x}-x^*\|_{\mathcal{X}} \\
&\le \left(\gamma(\hat{x})\|\mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + \frac{\tau(\hat{x})}{2}\right)\|\hat{x}-x^*\|_{\mathcal{X}}
\end{aligned}$$

and therefore

$$\begin{aligned}
\operatorname{eff}_\Delta(\hat{x}) &= \frac{\Delta(\hat{x})}{\|x^*-\hat{x}\|_{\mathcal{X}}} \\
&= \frac{1}{1-\tau(\hat{x})/2}\frac{\epsilon(\hat{x})}{\|x^*-\hat{x}\|_{\mathcal{X}}} \le \frac{1}{1-\tau(\hat{x})/2}\left(\gamma(\hat{x})\|\mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + \frac{\tau(\hat{x})}{2}\right) \\
&= \frac{\kappa(\hat{x})}{1-\tau(\hat{x})/2} + \frac{\tau(\hat{x})}{2-\tau(\hat{x})}.
\end{aligned}$$

(b)

$$\begin{aligned}
\epsilon(\hat{x}) &\le \left(\gamma(\hat{x})\|\mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + \gamma(\hat{x})L(2\epsilon(\hat{x}))\right)\|\hat{x}-x^*\|_{\mathcal{X}} \\
&\le \left(\gamma(\hat{x})\|\mathrm{D}G|_{\hat{x}}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + \frac{\hat{\tau}(\hat{x})}{2}\right)\|\hat{x}-x^*\|_{\mathcal{X}}
\end{aligned}$$

and therefore

$$\begin{aligned}
\mathrm{eff}_{\hat{\Delta}}(\hat{x}) &= \frac{\hat{\Delta}(\hat{x})}{\|x^* - \hat{x}\|_{\mathcal{X}}} \\
&= \frac{2}{1+\sqrt{1-\hat{\tau}(\hat{x})}} \frac{\epsilon(\hat{x})}{\|x^* - \hat{x}\|_{\mathcal{X}}} \leq \frac{2}{1+\sqrt{1-\hat{\tau}(\hat{x})}} \left(\gamma(\hat{x}) \left\| \mathrm{D}G|_{\hat{x}} \right\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} + \frac{\hat{\tau}(\hat{x})}{2} \right) \\
&= \frac{2\kappa(\hat{x})}{1+\sqrt{1-\hat{\tau}(\hat{x})}} + \frac{\hat{\tau}(\hat{x})}{2+2\sqrt{1-\hat{\tau}(\hat{x})}}.
\end{aligned}$$

□

Unfortunately, these generic bounds on the effectivity are rather coarse as one can easily see when once again inspecting the (affine) linear case. As was shown in Corollary 5.3.3 we have exact error prediction for the (affine) linear case which, in terms of the effectivity, can be expressed via $\mathrm{eff}(\hat{x}) = 1$. However, the bounds in Theorem 5.3.4 give the upper bound

$$\mathrm{eff}_{\Delta}(\hat{x}), \mathrm{eff}_{\hat{\Delta}}(\hat{x}) \leq \kappa(\hat{x})$$

which, even in the case of nonlinear problems, is often much larger than the actual value.

As we have seen in the proof of Theorem 5.3.4, the ability to bound the effectivity traces back to being able to bound the difference $\mathrm{D}G|_{\hat{x}}^{-1}(G(\hat{x})) - \mathrm{D}G|_{\hat{x}}^{-1}(G(x^*))$. For this purpose, we assume that the function $\mathrm{D}G|_{\hat{x}}^{-1} \circ G : \mathcal{X} \to \mathcal{X}$ is locally Lipschitz-continuous around $\hat{x}$. By this we mean that there exists a constant $C_G(\hat{x}) \geq 0$ such that

$$\left\| \mathrm{D}G|_{\hat{x}}^{-1}(G(x)) - \mathrm{D}G|_{\hat{x}}^{-1}(G(\hat{x})) \right\|_{\mathcal{X}} \leq C_G(\hat{x}) \left\| x - \hat{x} \right\|_{\mathcal{X}}, \quad \forall x \in \overline{B_{2\epsilon(\hat{x})}}(\hat{x}). \tag{5.9}$$

Please note that this assumption slightly weakens the conventional definition of local Lipschitz-continuity as we only require this property in a neighbourhood of the approximation. However, based on this property we can provide a (possibly) sharper estimate on the effectivity.

Lemma 5.3.5 (Lipschitz based effectivity estimate)**.** *Let $\mathrm{D}G|_{\hat{x}}^{-1} \circ G$ be locally Lipschitz-continuous around $\hat{x}$ with constant $C_G(\hat{x})$ and let the error estimate from Theorem 5.3.1* ***(a)*** *or* ***(b)*** *hold true. Then it holds*

(a) $\mathrm{eff}_{\Delta}(\hat{x}) \leq \dfrac{C_G(\hat{x})}{1-\tau(\hat{x})/2}$ ***(b)*** $\mathrm{eff}_{\hat{\Delta}}(\hat{x}) \leq \dfrac{2C_G(\hat{x})}{1+\sqrt{1-\hat{\tau}(\hat{x})}}$

Proof. **(a)** The proof follows directly from the fact that $G(x^*) = 0$ and

$$\begin{aligned}\Delta(\hat{x}) &= \frac{1}{1-\tau(\hat{x})/2} \left\| \mathrm{D}G|_{\hat{x}}^{-1} (G(\hat{x})) \right\|_{\mathcal{X}} \\ &= \frac{1}{1-\tau(\hat{x})/2} \left\| \mathrm{D}G|_{\hat{x}}^{-1} (G(\hat{x}) - G(x^*)) \right\|_{\mathcal{X}} \\ &\leq \frac{C_G(\hat{x})}{1-\tau(\hat{x})/2} \left\| \hat{x} - x^* \right\|_{\mathcal{X}}.\end{aligned}$$

(b) Likewise, using the fact that x^* is a solution of the problem (P), we get

$$\begin{aligned}\hat{\Delta}(\hat{x}) &= \frac{2}{1+\sqrt{1-\hat{\tau}(\hat{x})}} \left\| \mathrm{D}G|_{\hat{x}}^{-1} (G(\hat{x})) \right\|_{\mathcal{X}} \\ &= \frac{2}{1+\sqrt{1-\hat{\tau}(\hat{x})}} \left\| \mathrm{D}G|_{\hat{x}}^{-1} (G(\hat{x}) - G(x^*)) \right\|_{\mathcal{X}} \\ &\leq \frac{2C_G(\hat{x})}{1+\sqrt{1-\hat{\tau}(\hat{x})}} \left\| \hat{x} - x^* \right\|_{\mathcal{X}}.\end{aligned}$$

□

Please note that unlike the generic bounds of Theorem 5.3.4 the above maches with the result of Corollary 5.3.3, since for (affine) linear problems we have $\tau(\hat{x}) = \hat{\tau}(\hat{x}) = 0$ and $C_G(\hat{x}) = 1$, thus resulting in $\operatorname{eff}(\hat{x}) = 1$.

We furthermore note that the assumption of local Lipschitz-continuity around $\hat{x}$ is satisfied for a large class of problems. Taking a closer look at the proof of Corollary 5.3.3 one can see that it was in fact proven, that

$$\left\| \mathrm{D}G|_{\hat{x}}^{-1} (G(\hat{x})) - \mathrm{D}G|_{\hat{x}}^{-1} (G(x)) \right\|_{\mathcal{X}} \leq (\kappa(\hat{x}) + \gamma(\hat{x}) L(2\epsilon(\hat{x})) \left\| \hat{x} - x \right\|_{\mathcal{X}}.$$

In other words, under the assumption of Theorem 5.3.4 **(a)** or **(b)** we get the constants

(a) $C_G(\hat{x}) = \kappa(\hat{x}) + \dfrac{\tau(\hat{x})}{2}$ **(b)** $C_G(\hat{x}) = \kappa(\hat{x}) + \dfrac{\hat{\tau}(\hat{x})}{2}$

which, as mentioned above, are often a gross overestimate. However, we will see that under the same assumption a much sharper bound, i.e. a much smaller constant C_G can be achieved.

Lemma 5.3.6 (General local-Lipschitz continuity)**.** *Let the assumptions of Theorem 5.3.1 hold. Then the function* $\mathrm{D}G|_{\hat{x}}^{-1} \circ G : \mathcal{X} \to \mathcal{X}$ *is locally Lipschitz-continuous around* $\hat{x}$ *and* *the Lipschitz-constant is bounded by*

$$C_G(\hat{x}) \leq \frac{3}{2}. \tag{5.10}$$

Proof. Analogous to the proof of Theorem 5.3.1 let $H : \mathcal{X} \to \mathcal{X}$ be the function defined by

$$H(x) := x - \mathrm{D}G|_{\hat{x}}^{-1}(G(\hat{x})).$$

As it was shown in (5.7), H is a contraction in the Ball $\overline{B_{2\epsilon(\hat{x})}}(\hat{x})$ with constant $\frac{1}{2}$, which leads to

$$\|H(\hat{x}) - H(x)\|_{\mathcal{X}} \leq \frac{1}{2}\|\hat{x} - x\|_{\mathcal{X}} \qquad \text{for } x \in \overline{B_{2\epsilon(\hat{x})}}(\hat{x}).$$

Using the triangle inequality we now get

$$\begin{aligned} \left\|\mathrm{D}G|_{\hat{x}}^{-1}(G(\hat{x})) - \mathrm{D}G|_{\hat{x}}^{-1}(G(x))\right\|_{\mathcal{X}} &= \|H(\hat{x}) - H(x) - (\hat{x} - x)\|_{\mathcal{X}} \\ &\leq \|H(\hat{x}) - H(x)\|_{\mathcal{X}} + \|\hat{x} - x\|_{\mathcal{X}} \\ &\leq \frac{3}{2}\|\hat{x} - x\|_{\mathcal{X}}. \end{aligned}$$

□

Combining the results of Lemma 5.3.5 and Lemma 5.3.6 we achieve the following refined effectivity bound.

Corollary 5.3.7 (Refined generic effectivity bound)**.** *Let the assuptions of Theorem 5.3.1 hold. Then the effectivities for the error bounds $\Delta(\hat{x})$ and $\hat{\Delta}(\hat{x})$ are bounded by*

(a) $\mathrm{eff}_{\Delta}(\hat{x}) \leq \dfrac{3}{2 - \tau(\hat{x})} \leq 3$ ***(b)*** $\mathrm{eff}_{\hat{\Delta}}(\hat{x}) \leq \dfrac{3}{1 + \sqrt{1 - \hat{\tau}(\hat{x})}} \leq 3$

Proof. As mentioned above, using Lemma 5.3.5, Lemma 5.3.6 and the fact that the (modified) validity criterion is met, i.e. $\tau(\hat{x}), \hat{\tau}(\hat{x}) \leq 1$, we get

(a)

$$\mathrm{eff}_{\Delta}(x) \leq \frac{C_G(\hat{x})}{1 - \tau(\hat{x})/2} \leq \frac{3}{2 - \tau(\hat{x})} \leq 3$$

(b)

$$\operatorname{eff}_{\hat{\Delta}}(x) \leq \frac{2C_G(\hat{x})}{1+\sqrt{1-\hat{\tau}(\hat{x})}} \leq \frac{3}{1+\sqrt{1-\hat{\tau}(\hat{x})}} \leq 3$$

□

The bounds of Lemma 5.3.6 on the Lipschitz-constant and the resulting bounds on the effectivity display a vast improvement, especially since they are independent of the stability constant $\gamma(\hat{x})$. However, for problems of polynomial type we can further reduce the Lipschitz-constants. In particular, we will have a closer look at quadratic problems which encompass problems such as the Navier-Stokes equation, Burgers equation, nonlinear reaction-diffusion equations or the algebraic Riccati equation (ARE).

Proposition 5.3.8 (Local Lipschitz-continuity for quadratic problems)**.** *Let G be quadratic, i.e. there exists a $y_0 \in Y$, $A \in \mathcal{L}(\mathcal{X},\mathcal{Y})$ and a continuous bilinear mapping $B : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ such that*

$$G(x) = y_0 + Ax + \frac{1}{2}B(x,x).$$

Then $\mathrm{D}G|_{\hat{x}}^{-1} \circ G : \overline{B_{2\epsilon(\hat{x})}}(\hat{x}) :\to \mathcal{X}$ is locally Lipschitz-continuous around $\hat{x}$ with

$$C_G(\hat{x}) \leq 1 + \frac{1}{2}\gamma(\hat{x})c_B \left\|\hat{x} - x^*\right\|_{\mathcal{X}}$$

where $c_B := \sup\limits_{x,x' \in \mathcal{X}\setminus\{0\}} \frac{\|B(x,x')\|_{\mathcal{Y}}}{\|x\|_{\mathcal{X}}\|x'\|_{\mathcal{X}}}$ is the continuity constant of B.

Proof. It holds

$$\mathrm{D}G|_{\hat{x}}(x) = Ax + \frac{1}{2}\left(B(x,\hat{x}) + B(\hat{x},x)\right) \quad \text{and} \quad \mathrm{D}^2G\big|_{\hat{x}}(x,x') = \frac{1}{2}(B(x,x') + B(x',x)).$$

By direct computation we get the Taylor expansion

$$G(x) = G(\hat{x}) + \mathrm{D}G|_{\hat{x}}(x-\hat{x}) + \frac{1}{2}\,\mathrm{D}^2G\big|_{\hat{x}}(x-\hat{x},x-\hat{x})$$

and therefore

$$\begin{aligned}
\mathrm{D}G|_{\hat{x}}^{-1}(G(x)) - \mathrm{D}G|_{\hat{x}}^{-1}(G(\hat{x})) &= \mathrm{D}G|_{\hat{x}}^{-1}(G(x) - G(\hat{x})) \\
&= \mathrm{D}G|_{\hat{x}}^{-1}\Big(\mathrm{D}G|_{\hat{x}}(x-\hat{x}) + \frac{1}{2}\,\mathrm{D}^2G\Big|_{\hat{x}}(x-\hat{x}, x-\hat{x})\Big) \\
&= x - \hat{x} + \frac{1}{2}\,\mathrm{D}G|_{\hat{x}}^{-1}(\mathrm{D}^2G\Big|_{\hat{x}}(x-\hat{x}, x-\hat{x})) \\
&= x - \hat{x} + \frac{1}{2}\,\mathrm{D}G|_{\hat{x}}^{-1}(B(x-\hat{x}, x-\hat{x}).
\end{aligned}$$

Taking the norm on both sides, applying the definition of the continuity constant c_B and using the triangle inequality we get

$$\|\,\mathrm{D}G|_{\hat{x}}^{-1}(G(x)) - \mathrm{D}G|_{\hat{x}}^{-1}(G(\hat{x}))\|_{\mathcal{Y}} \leq \|x-\hat{x}\|_{\mathcal{X}} + \frac{1}{2}\gamma(\hat{x})c_B\|x-\hat{x}\|_{\mathcal{X}}^2.$$

The claim now follows for the choice $x = x^*$. □

Depending on the choice of error indicator, we can replace $\|\hat{x} - x^*\|$ by $\Delta(\hat{x})$ or $\hat{\Delta}(\hat{x})$, respectively. In all cases we get that $C_G(\hat{x}) \to 1$ as $\hat{x} \to x^*$. Since this is also accompanied by $\tau(\hat{x}), \hat{\tau}(\hat{x}) \to 0$ we can expect that

$$\mathrm{eff}_{\Delta}(\hat{x}), \mathrm{eff}_{\hat{\Delta}}(\hat{x}) \overset{5.3.5}{\to} 1 \qquad \text{as } \hat{x} \to x^*,$$

i.e. the quality of the effectivity bound improves with the quality of our approximation. On the contrary, the bounds of Corollary 5.3.7 only result in

$$\mathrm{eff}_{\Delta}(\hat{x}), \mathrm{eff}_{\hat{\Delta}}(\hat{x}) \leq \frac{3}{2} \qquad \text{as } \hat{x} \to x^*.$$

In the infinite-dimensional settings, i.e. $\dim(\mathcal{X}), \dim(\mathcal{Y}) = \infty$, the calculation of the necessary quantities is often impossible. Even in a finite-dimensional setting the computation of the quantities can be demanding or outright infeasible. In particular, this is true for very high-dimensional problems arising for example when dealing with semi-discretized PDEs. Nonetheless, one often has access to computable upper bounds, i.e.

$$\epsilon(\hat{x}) \leq \epsilon_{\mathrm{ub}}(\hat{x}), \quad \gamma(\hat{x}) \leq \gamma_{\mathrm{ub}}(\hat{x}), \quad L(\alpha) \leq L_{\mathrm{ub}}(\alpha). \tag{5.11}$$

In this case a slightly altered version of Theorem 5.3.1 can be formulated, where the quantities are replaced by their upper bounds:

Theorem 5.3.9 (Computable error bound)**.** *Let $\hat{x} \in \mathcal{X}$ be an approximate solution of the problem (P) and assume that $\mathrm{D}G|_{\hat{x}}$ is invertible.*

(a) *If the validity criterion*

$$\tau_{\mathrm{ub}}(\hat{x}) := 2\gamma_{\mathrm{ub}}(\hat{x}) L_{\mathrm{ub}}(2\epsilon_{\mathrm{ub}}(\hat{x})) \leq 1. \tag{5.12}$$

is met, then there exists a unique solution $x^* \in \mathcal{X}$ of (P) in the closed ball $\overline{B_{2\epsilon_{\mathrm{ub}}(\hat{x})}}(\hat{x})$. Furthermore, the error $e = \hat{x} - x^* \in \mathcal{X}$ is bounded by

$$\|e\|_{\mathcal{X}} = \|\hat{x} - x^*\|_{\mathcal{X}} \leq \Delta_{\mathrm{ub}}(\hat{x}) := \frac{1}{1 - \tau_{\mathrm{ub}}(\hat{x})/2} \epsilon_{\mathrm{ub}}(\hat{x}) \leq 2\epsilon_{\mathrm{ub}}(\hat{x}). \tag{5.13}$$

(b) *If $L_{\mathrm{ub}}(\alpha) \leq C_{\mathrm{ub}}\alpha$ for some $C_{\mathrm{ub}} > 0$ and if the modified validity criterion*

$$\hat{\tau}_{\mathrm{ub}}(\hat{x}) := 4\gamma_{\mathrm{ub}}(\hat{x}) C_{\mathrm{ub}} \epsilon_{\mathrm{ub}}(\hat{x}) \leq 1 \tag{5.14}$$

is satisfied, then there exists a unique solution $x^* \in \mathcal{X}$ of (P) in the closed ball $\overline{B_{2\epsilon_{\mathrm{ub}}(\hat{x})}}(\hat{x})$. Furthermore, the error $e = \hat{x} - x^* \in \mathcal{X}$ is bounded by

$$\|e\|_{\mathcal{X}} = \|\hat{x} - x^*\|_{\mathcal{X}} \leq \hat{\Delta}_{\mathrm{ub}}(\hat{x}) := \frac{2}{1 + \sqrt{1 - \hat{\tau}_{\mathrm{ub}}(\hat{x})}} \epsilon_{\mathrm{ub}}(\hat{x}). \tag{5.15}$$

Proof. The proof is identical to that of Theorem 5.3.1 if all of the key quantities are replaced by their corresponding upper bounds. □

Analogous to the results of Corollary 5.3.7, we can bound the effectivity of the computable error bounds in (5.13) and (5.15) if we further assume that the upper bound $\epsilon_{\mathrm{ub}}(\hat{x})$ is sufficiently well behaved.

Corollary 5.3.10 (Refined generic effectivity bound for computable quantities)**.** *Let the assumptions of Theorem 5.3.9 hold. If $\epsilon_{\mathrm{ub}}(\hat{x}) \geq \epsilon(\hat{x})$ is an upper bound such that there exists a constant C_ϵ with*

$$\epsilon_{\mathrm{ub}}(\hat{x}) \leq C_\epsilon \epsilon(\hat{x}),$$

then the effectivities of the error bounds $\Delta_{\mathrm{ub}}(\hat{x})$ and $\hat{\Delta}_{\mathrm{ub}}(\hat{x})$ are bounded as follows

(a) $\mathrm{eff}_{\Delta_{\mathrm{ub}}}(\hat{x}) \leq \dfrac{3}{2 - \tau_{\mathrm{ub}}(\hat{x})} C_\epsilon \leq 3C_\epsilon$

(b) $\mathrm{eff}_{\hat{\Delta}_{\mathrm{ub}}}(\hat{x}) \leq \dfrac{3}{1 + \sqrt{1 - \hat{\tau}_{\mathrm{ub}}(\hat{x})}} C_\epsilon \leq 3C_\epsilon$

Proof. Using Lemma 5.3.6 and the assumption on $\epsilon_{\mathrm{ub}}(\hat{x})$ we get

$$\epsilon_{\mathrm{ub}}(\hat{x}) \leq C_\epsilon \epsilon(\hat{x}) \leq \frac{3}{2} C_\epsilon \|\hat{x} - x^*\|_{\mathcal{X}} \tag{5.16}$$

Substituting (5.16) for $\epsilon_{\text{ub}}(\hat{x})$ in the error bounds $\Delta_{\text{ub}}(\hat{x})$ and $\hat{\Delta}_{\text{ub}}(\hat{x})$ we get the desired results. □

In the above corollary we made use of Lemma 5.3.6 which provides the Lipschitz-constant $C_G = \frac{3}{2}$. In cases where a better Lipschitz-constant is available, the upper bound can of course be refined resulting in

$$\operatorname{eff}_{\Delta_{\text{ub}}}(\hat{x}) \leq \frac{C_G(\hat{x})C_\epsilon}{1-\tau_{\text{ub}}(\hat{x})/2}$$

and

$$\operatorname{eff}_{\hat{\Delta}_{\text{ub}}}(\hat{x}) \leq \frac{2C_G(\hat{x})C_\epsilon}{1+\sqrt{1-\hat{\tau}_{\text{ub}}(\hat{x})}},$$

respectively.

5.4 Sharp effectivities through auxiliary linear problems

In this subsection, we show how a very sharp bound for $\epsilon(\hat{x})$ which satisfies the assumptions of Corollary 5.3.10 can be obtained with low additional computational overhead. As it was illustrated in the introduction by a simple two-dimensional linear problem, the effectivity of the a-posteriori error bound deteriorates heavily if the calculation of $\epsilon(\hat{x})$ is split according to equation (5.8). This can be seen when considering that

$$\epsilon_{\text{split}}(\hat{x}) := \gamma(\hat{x}) \left\|G(\hat{x})\right\|_{\mathcal{Y}} \geq \left\| \mathrm{D}G|_{\hat{x}}^{-1}\left(G(\hat{x})\right)\right\|_{\mathcal{X}} = \epsilon(\hat{x}),$$

constitutes an upper bound of $\epsilon(\hat{x})$ satisfying Corollary 5.3.10 with the constant $C_\epsilon = \kappa(\hat{x})$, as we have

$$\epsilon_{\text{split}}(\hat{x}) = \gamma(\hat{x}) \left\| \mathrm{D}G|_{\hat{x}}\, \mathrm{D}G|_{\hat{x}}^{-1}\left(G(\hat{x})\right)\right\|_{\mathcal{Y}} \leq \kappa(\hat{x}) \left\| \mathrm{D}G|_{\hat{x}}^{-1}\left(G(\hat{x})\right)\right\|_{\mathcal{X}} = \kappa(\hat{x})\epsilon(\hat{x}). \qquad (5.17)$$

Hence, the quality of the error bounds deteriorate by the factor $\kappa(\hat{x})$ which is usually much larger than one. Therefore, the key towards highly effective (i.e. $\operatorname{eff}(\hat{x}) \approx 1$) error bounds lies in finding highly effective approximations or bounds to $\epsilon(\hat{x})$.

For this reason, we shall first note, that $\epsilon(\hat{x})$ can be calculated by solving the following linear equation

$$\mathrm{D}G|_{\hat{x}}\left(E(\hat{x})\right) = G(\hat{x}) \qquad (5.18)$$

for $E(\hat{x}) \in \mathcal{X}$ and then taking the norm $\epsilon(\hat{x}) = \left\|E(\hat{x})\right\|_{\mathcal{X}}$. While the above equation computes the exact error for (affine) linear problems (P), i.e. for (affine) linear G, this no longer is the case for non-linear G. Hence, we shall refer to (5.18) as the auxiliary linear problem (ALP) and will consequently denote the here presented error bounds as ALP-based error bounds. While linear problems of the form (5.18) can often be solved relatively easy, it can be laborious in a high-dimensional or multi-query scenario. With the goal of a computationally efficient scheme in mind, we shall therefore replace $E(\hat{x})$ by an approximation $\hat{E}(\hat{x}) \in \mathcal{X}$, which is itself computed via a suitable approximation method.

Remark 5.4.1. Equation (5.18) can be interpreted as Newton-update as follows:

For the problem $G(x) = 0$, the Newton-iteration iteratively computes an approximation by starting from an initial guess $x_0 \in \mathcal{X}$ and setting

$$x_{n+1} = x_n + \Delta x_n, \qquad \text{with} \qquad \mathrm{D}G|_{x_n}(\Delta x_n) = -G(x_n), \quad n \geq 0.$$

In this sense, the computation of $E(\hat{x})$ is equal to computing the Newton-update Δx_n. Likewise, any approximation $\hat{E}(\hat{x})$ can be interpreted as a quasi Newton update. However, unlike for the Newton iteration where the additional computational effort is used for improved approximation by setting $\hat{\hat{x}} = \hat{x} - E(\hat{x})$ or $\hat{\hat{x}} = \hat{x} - \hat{E}(\hat{x})$, respectively, we use the addition effort for improving the error estimate.

This idea, to invest further computational resources in order to improve the quantification of the error is also explored in [43]. Instead of solving an auxiliary linear problem, the authors assume to have two approximations $\hat{x}_1$ and $\hat{x}_2$ available, where the latter is of higher quality. The error can then be bounded by

$$\|\hat{x}_1 - x^*\|_{\mathcal{X}} \leq \|\hat{x}_2 - \hat{x}_1\|_{\mathcal{X}} + \|\hat{x}_2 - x^*\|_{\mathcal{X}} \leq \|\hat{x}_2 - \hat{x}_1\|_{\mathcal{X}} + \gamma_{\mathrm{ub}}(\hat{x}_2)\, \|G(\hat{x}_2)\|_{\mathcal{Y}},$$

which should result in highly effective error bounds provided the second approximation $\hat{x}_2$ is of sufficient quality.

As we will see in the following lemma, the strength of the method proposed here lies in the simple fact that it enables us to easily derive a rigorous error bound for the quantity $\epsilon(\hat{x})$, when an approximation $\hat{E}(\hat{x})$ is available:

Lemma 5.4.2 (Upper bound for $\epsilon(\hat{x})$)**.** *Let $\hat{E}(\hat{x}) \in \mathcal{X}$ be an approximate solution to the ALP* (5.18) *and define the ALP residual $R(\hat{x}) := \mathrm{D}G|_{\hat{x}}(\hat{E}(\hat{x})) - G(\hat{x})$. Then the upper bound*

$$\epsilon(\hat{x}) \leq \epsilon_{\mathrm{ub}}(\hat{x}) := \left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} + \gamma_{\mathrm{ub}}(\hat{x})\, \|R(\hat{x})\|_{Y}. \tag{5.19}$$

holds true, where $\gamma(\hat{x}) \leq \gamma_{\mathrm{ub}}(\hat{x})$ is an upper bound of the stability constant.

Proof. The proof follows immediately from the triangle inequality as we have

$$\epsilon(\hat{x}) = \|E(\hat{x})\|_X = \left\|E(\hat{x}) + \hat{E}(\hat{x}) - \hat{E}(\hat{x})\right\|_X \leq \left\|\hat{E}(\hat{x})\right\|_X + \left\|\hat{E}(\hat{x}) - E(\hat{x})\right\|_X .$$

For the difference $\hat{E}(\hat{x}) - E(\hat{x})$, we make use of the linearity of the ALP and obtain the relation $\mathrm{D}G|_{\hat{x}}\left(\hat{E}(\hat{x}) - E(\hat{x})\right) = R(\hat{x})$, from which we get

$$\left\|\hat{E}(\hat{x}) - E(\hat{x})\right\|_{\mathcal{X}} = \left\|\mathrm{D}G|_{\hat{x}}^{-1}\left(R(\hat{x})\right)\right\|_{\mathcal{X}} \leq \gamma(\hat{x})\left\|R(\hat{x})\right\|_{\mathcal{Y}} \leq \gamma_{\mathrm{ub}}(\hat{x})\left\|R(\hat{x})\right\|_{\mathcal{Y}} . \qquad (5.20)$$

□

Under the assumption that an efficient approximation scheme for the computation of $\hat{E}$ is available, the computational overhead for the calculation of $\epsilon_{\mathrm{ub}}(\hat{x})$ is relatively small as it only requires the calculation of $\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}$ and $\|R(\hat{x})\|_{\mathcal{Y}}$. The latter of these two quantities is itself often used as an abortion criterion, when iterative solvers for large-scale linear systems are considered. Additionally, no further computation of quantities is required, as $\gamma_{\mathrm{ub}}(\hat{x})$ already has to be calculated to check the validity criterion.

At this point, we want to emphasize that the numerical results presented in the later parts of this thesis reveal very accurate error predictions when using $\Delta_{\mathrm{ub}}(\hat{x})$ from Theorem 5.3.9 with the choice $\epsilon_{\mathrm{ub}}(\hat{x})$ according to Lemma 5.4.2. One possible explanation for this observation can be deduced from

$$\|\hat{x} - x^*\|_{\mathcal{X}} \leq \Delta(\hat{x}) \leq 2\epsilon(\hat{x}) \leq 2\epsilon_{\mathrm{ub}}(\hat{x}),$$

and the fact that $\epsilon_{\mathrm{ub}}(\hat{x})$ is a very accurate estimate of $\epsilon(\hat{x})$. In contrast to the original splitting of $\epsilon(\hat{x})$ in equation (5.8), the splitting in (5.20) does not deteriorate the bound $\epsilon_{\mathrm{ub}}(\hat{x})$ significantly since $\|R(\hat{x})\|_{\mathcal{Y}}$ is often much smaller than $\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}$. To quantify this observation rigorously, we use the following lemma.

Lemma 5.4.3 (Relation of $\epsilon(\hat{x})$ and $\epsilon_{\mathrm{ub}}(\hat{x})$). *Assume*

$$2\gamma_{\mathrm{ub}}(\hat{x})\left\|R(\hat{x})\right\|_{\mathcal{Y}} \leq \left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} \qquad (5.21)$$

Then the following inequality holds true for $\epsilon_{\mathrm{ub}}(\hat{x})$ chosen as in (5.19).

$$\epsilon_{\mathrm{ub}}(\hat{x}) \leq C_\epsilon(\hat{x})\epsilon(\hat{x}), \quad \textit{with} \quad C_\epsilon(\hat{x}) := \left(1 + 4\frac{\gamma_{\mathrm{ub}}(\hat{x})\left\|R(\hat{x})\right\|_{\mathcal{X}}}{\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}}\right) \leq 3.$$

Proof. Note that the following proof is similar to a proof for the effectivity of relative RB error bounds [72]: The proof follows with $E(\hat{x}) = \mathrm{D}G|_{\hat{x}}^{-1}(G(\hat{x}))$ and $\|E(\hat{x})\|_{\mathcal{X}} \neq 0$

$$\begin{aligned}\epsilon_{\mathrm{ub}}(\hat{x}) &= \left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} + \gamma_{\mathrm{ub}}(\hat{x}) \|R(\hat{x})\|_{\mathcal{Y}} \\ &\leq \|E(\hat{x})\|_{\mathcal{X}} + \left\|\hat{E}(\hat{x}) - E(\hat{x})\right\|_{\mathcal{X}} + \gamma_{\mathrm{ub}}(\hat{x}) \|R(\hat{x})\|_{\mathcal{Y}} \\ &= \left(1 + \frac{\left\|\hat{E}(\hat{x}) - E(\hat{x})\right\|_{\mathcal{X}}}{\|E(\hat{x})\|_{\mathcal{X}}} + \frac{\gamma_{\mathrm{ub}}(\hat{x}) \|R(\hat{x})\|_{\mathcal{Y}}}{\|E(\hat{x})\|_{\mathcal{X}}}\right) \|E(\hat{x})\|_{\mathcal{X}} . \end{aligned} \tag{5.22}$$

From the triangle inequality and (5.20) we infer

$$\left| \frac{\|E(\hat{x})\|_{\mathcal{X}} - \left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}}{\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}} \right| \leq \frac{\left\|\hat{E}(\hat{x}) - E(\hat{x})\right\|_{\mathcal{X}}}{\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}} \leq \frac{\gamma_{\mathrm{ub}}(\hat{x}) \|R(\hat{x})\|_{\mathcal{Y}}}{\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}} \leq \frac{1}{2}.$$

We now have to consider the following two cases

(a) if $\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} > \|E(\hat{x})\|_{\mathcal{X}}$, we get

$$\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} - \|E(\hat{x})\|_{\mathcal{X}} \leq \frac{1}{2} \left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}$$

and hence

$$\frac{1}{2} \left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} \leq \|E(\hat{x})\|_{\mathcal{X}}$$

(b) if $\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} \leq \|E(\hat{x})\|_{\mathcal{X}}$, it immediately follows

$$\frac{1}{2} \left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}} \leq \|E(\hat{x})\|_{\mathcal{X}}$$

Ultimately, we obtain

$$\frac{\left\|\hat{E}(\hat{x}) - E(\hat{x})\right\|_{\mathcal{X}}}{\|E(\hat{x})\|_{\mathcal{X}}} \leq \frac{\gamma_{\mathrm{ub}}(\hat{x}) \|R(\hat{x})\|_{\mathcal{Y}}}{\|E(\hat{x})\|_{\mathcal{X}}} \leq 2 \frac{\gamma_{\mathrm{ub}}(\hat{x}) \|R(\hat{x})\|_{\mathcal{Y}}}{\left\|\hat{E}(\hat{x})\right\|_{\mathcal{X}}}.$$

Inserting this twice into (5.22) yields the final result. □

Combining the results of Theorem 5.3.9, Corollary 5.3.10 and Lemma 5.4.3 we conclude this subsection with the following corollary:

Corollary 5.4.4. *Let the assumptions of Theorem 5.3.9 and Lemma 5.4.3 hold. Then the effectivity of the error bounds $\Delta_{\mathrm{ub}}(\hat{x})$ and $\hat{\Delta}_{\mathrm{ub}}(\hat{x})$ is bounded by*

(a) $\mathrm{eff}_{\Delta_{\mathrm{ub}}}(\hat{x}) \leq \dfrac{9}{2 - \tau_{\mathrm{ub}}(\hat{x})} \leq 9$

(b) $\mathrm{eff}_{\hat{\Delta}_{\mathrm{ub}}}(\hat{x}) \leq \dfrac{9}{1+\sqrt{1-\hat{\tau}_{\mathrm{ub}}(\hat{x})}} \leq 9.$

Remark 5.4.5. Similar to the results of Proposition 5.3.8 we also have, that $C_{\epsilon(\hat{x})}$ tends towards 1 as the approximation $\hat{E}(\hat{x})$ tends towards the true solution $E(\hat{x})$. In other words, in practice we often experience effectivities much smaller than 9 once the approximation $\hat{E}(\hat{x})$ is good enough for inequality (5.21) to be satisfied.

5.5 Highly accurate error bounds in the reduced basis context

In this section, we apply the proposed error bound within the RB framework. In particular, we explain how the a-posteriori error bound derived in Section 5.2 can be applied to parametric and nonlinear problems within the RB context. Additionally, we will give a short summary of the basic ideas of RB methods, but we refer to literature, e.g. [39, 46, 72, 76, 82] for a more detailed introduction into the topic.

5.5.1 Parametric nonlinear problems and the reduced basis method

In the following, we study parametric problems. For this purpose, let $\mu \in \mathcal{P}$ be a parameter vector living in the compact set of admissible parameters $\mathcal{P} \subset \mathbb{R}^P$ for some $P \in \mathbb{N}$. We now consider problems which take the form

$$\text{For } \mu \in \mathcal{P} \text{ find } x^*(\mu) \in \mathcal{X} \,:\, G(x^*(\mu);\mu) = 0, \tag{$P(\mu)$}$$

for some parameter-dependent operator $G(\cdot;\mu) : \mathcal{X} \to \mathcal{Y}$. Furthermore, we shall assume that for every parameter $\mu \in \mathcal{P}$ at least one solution exists. The main idea behind RB methods is to determine a suitable low-dimensional subspace $\mathcal{X}_N \subset \mathcal{X}$ with $N = \dim(\mathcal{X}_N) \ll \dim(\mathcal{X}) = d \leq \infty$ and to find approximate solutions in this subspace by solving an N-dimensional so-called reduced problem. To this end, we equip the approximation space $\mathcal{X}_N$ with a so called reduced basis $\{\phi_1, \ldots, \phi_N\} \subset \mathcal{X}$, of linearly independent basis elements $\phi_i \in \mathcal{X}$. The approximation $\hat{x}(\mu) \in \mathcal{X}_N$ is then given as a linear combination

$$\hat{x}(\mu) := \sum_{i=1}^{N} x_{N,i}\phi_i = \Phi x_N(\mu),$$

where the coefficient functions $x_{N,i} : \mathcal{P} \to \mathbb{R}$ are called reduced coordinates of the reduced coordinate vector $x_N = (x_{N,i})_{i=1}^N \in \mathbb{R}^N$ and where we introduce $\Phi := (\phi_1, \ldots, \phi_N)$ as the row vector of basis functions. We now restrict the set of all possible solutions of the problem $(P(\mu))$ to the subspace $\mathcal{X}_N$ and by projecting the correpsonding residual $G(\hat{x}(\mu); \mu) \subset \mathcal{Y}$ onto another low dimensional subspace $\mathcal{Y}_N \subset \mathcal{Y}$ with $\dim(\mathcal{Y}_N) = N$, we ultimately arrive at the reduced problem

$$\text{For } \mu \in \mathcal{P} \text{ find } \hat{x}(\mu) = \Phi x_N(\mu) \in \mathcal{X}_N \ : \ G_N(\hat{x}(\mu); \mu) = 0, . \tag{$P_N(\mu)$}$$

Here the operator $G_N(\cdot; \mu) : \mathcal{X}_N \to \mathcal{Y}_N$ is given by

$$G_N(\cdot; \mu) = \Pi_{\mathcal{Y}_N} \left(G(\cdot; \mu)|_{\mathcal{X}_N} \right)$$

where $\Pi_{\mathcal{Y}_N} : \mathcal{Y} \to \mathcal{Y}_N$ denotes a projection onto the subspace $\mathcal{Y}_N$. This procedure is commonly referred to as Petrov-Galerkin projection and it is widely used for projection-based model order reduction (MOR) methods. The solvability of the reduced problem $(P_N(\mu))$ can typically be ensured by a careful construction of the spaces $\mathcal{X}_N$ and $\mathcal{Y}_N$. Once again, we refer to literature, [21, 24, 30, 47, 105] for a more detailed view into the different construction methods that have been proposed over the years. In the context of this work, we will henceforth assume that both, the reduced and the non-reduced problem, are solvable and, in particular, that we can compute true solutions $x^*(\mu) \in \mathcal{X}$ and approximations $\hat{x}(\mu) \in \mathcal{X}_N$ for any given parameter $\mu \in \mathcal{P}$. However, we still do not require uniqueness of the solutions.

5.5.2 Effective error prediction for the RB method

A fundamental question which arises when using RB methods is the following: given an approximation $\hat{x}(\mu) \in \mathcal{X}_N$ of a (true) solution $x^*(\mu) \in \mathcal{X}$, are we able to quantify the error $e(\mu) = \hat{x}(\mu) - x^*(\mu)$ rigorously and with good effectivity. With the previous work of Section 5.2, we can give a positive answer to this question. To this end, we will apply Theorem 5.3.1 or rather its refinement Theorem 5.3.9. In a parametric problem setting, such as described in $(P(\mu))$, we require an efficient way for calculating the quantities

$$\begin{aligned}
\gamma(\mu) &:= \gamma(\hat{x}(\mu)) =:= \left\| \mathrm{D}G(\cdot; \mu)|_{\hat{x}(\mu)}^{-1} \right\|_{\mathcal{L}(\mathcal{Y}, \mathcal{X})}, \\
\epsilon(\mu) &:= \epsilon(\hat{x}(\mu)) = \left\| [\mathrm{D}G(\cdot; \mu)|_{\hat{x}(\mu)}^{-1}](G(\hat{x}(\mu); \mu)) \right\|_{\mathcal{X}}, \\
L(\alpha; \mu) &:= \sup_{x \in \overline{B_\alpha}(\hat{x}(\mu))} \left\| \mathrm{D}G(\cdot; \mu)|_{\hat{x}(\mu)} - \mathrm{D}G(\cdot; \mu)|_x \right\|_{\mathcal{L}(\mathcal{X}, \mathcal{Y})}.
\end{aligned}$$

In the following, we omit the explicit dependency of $\hat{x}(\mu)$ for the sake of readability, however, we keep the dependency on the parameter $\mu \in \mathcal{P}$.

Since a direct computation of the above quantities is often to computationally expensive, we assume to have access to rapidly computable upper bounds analogous to the non-parametric case of (5.11)

$$\epsilon(\mu) \leq \epsilon_{\mathrm{ub}}(\mu), \quad \gamma(\mu) \leq \gamma_{\mathrm{ub}}(\mu), \quad L(\alpha;\mu) \leq L_{\mathrm{ub}}(\alpha;\mu).$$

While all of these quantities are important for the error quantification, we will primarily focus on the efficient calculation of $\epsilon_{\mathrm{ub}}(\mu)$. Efficient calculations for $L_{\mathrm{ub}}(\alpha,\mu)$ for most of the problems can be reduced to finding a suitable constant $C_{\mathrm{ub}}(\mu) > 0$ such that $L_{\mathrm{ub}}(\alpha;\mu) \leq C_{\mathrm{ub}}(\mu)\alpha$. However, such a bound is for the most part rather problem specific and we will not go into further detail. In contrast, the literature is rich with methods for bounding the stability constant $\gamma(\mu)$ for a wide variety of parametric problems, and henceforth we refer to the existing literature, e.g. [93, 49, 50].

We recall that $\epsilon(\mu)$ can be calculated explicitly by solving the (parametric) ALP for $E(\mu) \in \mathcal{X}$

$$\text{For } \mu \in \mathcal{P} \text{ find } \; E(\mu) \in \mathcal{X} \; : \; [\mathrm{D}G(\cdot;\mu)|_{\hat{x}(\mu)}](E(\mu)) = G(\hat{x}(\mu);\mu). \tag{$P^E(\mu)$}$$

and consequently taking the norm $\epsilon(\mu) = \|E(\mu)\|_{\mathcal{X}}$. In Lemma 5.4.2 we have already seen how a suitable upper bound for the non-split residual $\epsilon(\mu)$ can be calculated, provided that we have access to an approximation $\hat{E}(\mu) \in \mathcal{X}$ of the solution $E(\mu)$. In the context of RB methods, the idea to obtain such an approximation is to once again employ a Petrov-Galerkin projection of the parametric ALP $(P^E(\mu))$. For this purpose, we consider secondary reduced spaces $\mathcal{X}_M^E \subset \mathcal{X}$ and $\mathcal{Y}_M^E \subset \mathcal{Y}$ with $\dim(\mathcal{X}_M^E) = \dim(\mathcal{Y}_M^E) = M \ll d = \dim(X)$. We equip both subspaces with bases $\{\phi_1^E, \ldots, \phi_M^E\} \subset \mathcal{X}$ and $\{\psi_1^E, \ldots, \psi_M^E\} \subset \mathcal{Y}$ consisting of linear independent basis functions. The approximation $\hat{E}(\mu)$ is then given as a linear combination

$$\widehat{E}(\mu) \coloneqq \sum_{i=1}^{M} E_{M,i}(\mu)\, \phi_i^E \in \mathcal{X}_M^E, \quad \text{with} \quad E_M(\mu) \coloneqq [E_{M,1}(\mu), \ldots, E_{M,M}(\mu)]^T \in \mathbb{R}^M.$$

By projecting the parametric ALP $(P^E(\mu))$ in an analogous fashion to the original problem $(P(\mu))$

$$\Pi_{\mathcal{Y}_M^E}\left([\mathrm{D}G(\cdot;\mu)|_{\hat{x}(\mu)}](\widehat{E}(\mu))\right) = \Pi_{\mathcal{Y}_M^E}\left(G(\hat{x}(\mu);\mu)\right). \tag{$P_M^E(\mu)$}$$

we obtain a reduced parametric ALP, whose solution provides us with the reduced coordinate functions $E_{M,i} : \mathcal{P} \to \mathbb{R}$. Please note that analogously to the reduced problem for the approximation $\hat{x}(\mu)$ the above problem ist M-dimensional and therefore can be solved efficiently, provided M is chosen sufficiently small. To get a rigorous upper bound $\epsilon_{\text{ub}}(\mu) \geq \epsilon(\mu)$ we define the residual $R(\mu) \in \mathcal{Y}$ of the approximation to the ALP via

$$R(\mu) := [\mathrm{D}G(\cdot;\mu)|_{\hat{x}(\mu)}](\widehat{E}(\mu)) - G(\hat{x}(\mu);\mu).$$

Applying Lemma 5.4.2 we achieve the upper bound

$$\epsilon(\mu) \leq \epsilon_{\text{ub}}(\mu) = \left\|\widehat{E}(\mu)\right\|_{\mathcal{X}} + \gamma_{\text{ub}}(\mu) \left\|R(\mu)\right\|_{\mathcal{Y}}. \tag{5.23}$$

Remark 5.5.1. Analogously to the original parametric problem $(P(\mu))$, one requires a careful construction of the spaces $\mathcal{X}_M^E$ and $\mathcal{Y}_M^E$. While the choice $\mathcal{X}_M^E = \mathcal{X}_N$ seems reasonable to avoid the secondary construction of a reduced space, the same principle cannot be applied to the space $\mathcal{Y}_M^E$. Since, based on the definition of the approximation $\hat{x}(\mu)$, the choice $\mathcal{Y}_M^E = \mathcal{Y}_N$ would lead to the reduced ALP

$$\begin{aligned}\Pi_{\mathcal{Y}_M^E}\left([\mathrm{D}G(\cdot;\mu)|_{\hat{x}(\mu)}](\widehat{E}(\mu))\right) &= \Pi_{\mathcal{Y}_M^E}\left(G(\hat{x}(\mu);\mu)\right)\\ &= \Pi_{\mathcal{Y}_N}\left(G(\hat{x}(\mu);\mu)\right) = G_N(\hat{x}(\mu);\mu) = 0\end{aligned}$$

which results in the approximation $\hat{E}(\mu) = 0$ and consequently in the residual $R(\mu) = G(\hat{x}(\mu);\mu)$. Combined with Lemma 5.4.2, the upper bound

$$\epsilon_{\text{ub}}(\mu) = \left\|\hat{E}(\mu)\right\|_{\mathcal{X}} + \gamma_{\text{ub}}(\mu) \left\|R(\mu)\right\|_{\mathcal{Y}} = \gamma_{\text{ub}}(\mu) \left\|G(\hat{x}(\mu);\mu\right\|_{\mathcal{Y}} = \epsilon_{\text{split}}(\mu)$$

is reduced to the split-bound $\epsilon_{\text{split}}(\mu)$ which we tried to avoid in the first place as the resulting error bounds have poor effectivity in general.

In the following, we denote as $\Delta_{\text{ub}}(\mu)$ and $\hat{\Delta}_{\text{ub}}(\mu)$ the parametric computable error bounds stemming from Theorem 5.3.9 where we use $\epsilon_{\text{ub}}(\mu)$ given by equation (5.23). The rigorousness of the upper bound can then be guaranteed by stipulating a parametrized version of the condition in Lemma 5.4.3 equation (5.21). Consequently this leads to rigorous error bounds as summarized in the following Theorem.

Theorem 5.5.2 (Rigorous error bounds for the parametric problem $(P(\mu))$)**.** *For $\mu \in \mathcal{P}$, let $\gamma_{\text{ub}}(\mu)$ and $L_{\text{ub}}(\alpha;\mu)$ be upper bounds and let $\epsilon_{\text{ub}}(\mu)$ be given by equation* (5.23)*. Furthermore, let the assumptions of Theorem 5.3.9, in particular the (modified) validity criterion* (5.12) : $\tau_{\text{ub}}(\mu) := \tau_{\text{ub}}(\hat{x}(\mu))$ ((5.14) : $\hat{\tau}_{\text{ub}}(\mu)) := \hat{\tau}_{\text{ub}}(\hat{x}(\mu)))$*, be met for the above* *quantities and denote as* $\Delta_{\mathrm{ub}}(\mu)$ *and* $\hat{\Delta}_{\mathrm{ub}}(\mu)$ *the resulting error bounds. If the inequality*

$$2\gamma_{\mathrm{ub}}(\mu) \left\| R(\mu) \right\|_{\mathcal{Y}} \leq \left\| \hat{E}(\mu) \right\|_{\mathcal{X}} \tag{5.24}$$

holds than the error bounds $\Delta_{\mathrm{ub}}(\mu)$ *and* $\hat{\Delta}_{\mathrm{ub}}(\mu)$ *are rigorous and their effectivities are bounded by*

(a) $\mathrm{eff}_{\Delta_{\mathrm{ub}}}(\mu) \coloneqq \mathrm{eff}_{\Delta_{\mathrm{ub}}}(\hat{x}(\mu)) \leq \dfrac{9}{2 - \tau_{\mathrm{ub}}(\mu)} \leq 9$

(b) $\mathrm{eff}_{\hat{\Delta}_{\mathrm{ub}}}(\mu) \coloneqq \mathrm{eff}_{\hat{\Delta}_{\mathrm{ub}}}(\hat{x}(\mu)) \leq \dfrac{9}{1 + \sqrt{1 - \hat{\tau}_{\mathrm{ub}}(\mu)}} \leq 9$

Proof. The bounds follow immediately by applying Corollary 5.4.4 when considering the above upper bounds to the required quantities. □

We conclude this section by showing that the error bounds $\Delta_{\mathrm{ub}}(\mu)$ and $\hat{\Delta}_{\mathrm{ub}}(\mu)$ satisfy another desired property of error bounds in the RB context. Namely the identification of true solutions $x^*(\mu) \in \mathcal{X}_N$. While this is guaranteed to hold if the assumptions of Theorem 5.5.2 are met and the error bounds are rigorous, this is no longer immediately evident, if the inequality (5.24) does not hold.

Proposition 5.5.3 (Identification of solutions in the reduced space $\mathcal{X}_N$)**.** *Let the assumptions of Theorem 5.3.9 be met. If* $x^*(\mu) = \hat{x}(\mu) \in \mathcal{X}_N$ *for some parameter* $\mu \in \mathcal{P}$ *then it holds*

$$\Delta_{\mathrm{ub}}(\mu) = \hat{\Delta}_{\mathrm{ub}}(\mu) = 0.$$

The converse also holds, i.e. if $\Delta_{\mathrm{ub}}(\mu) = \hat{\Delta}_{\mathrm{ub}}(\mu) = 0$*, then* $x^*(\mu) = \hat{x}(\mu) \in \mathcal{X}_N$*.*

Proof. If $x^*(\mu) = \hat{x}(\mu) \in \mathcal{X}_N$, then $G(\hat{x}(\mu);\mu) = 0$ and therefore $\Pi_{\mathcal{Y}_M^E}(G(\hat{x}(\mu);\mu)) = 0$. Hence, the reduced ALP $(P_M^E(\mu))$ has the solution $\hat{E}(\mu) = 0$ which in turn leads to $R(\mu) = 0$ and $\epsilon_{\mathrm{ub}}(\mu) = 0$. In total, this leads to $\Delta_{\mathrm{ub}}(\mu) = \hat{\Delta}_{\mathrm{ub}}(\mu) = 0$. The converse follows immediately, as by assumption we have that $\Delta_{\mathrm{ub}}(\mu)$ and $\hat{\Delta}_{\mathrm{ub}}(\mu)$ are upper bounds on the error which therefore vanishes. □

5.5.3 Improvement of classical RB bounds for linear elliptic problems

Classically the RB method is applied in the context of parametric PDE. In this section we recall the standard RB error estimation results for linear elliptic problems and relate them to the bounds presented in the previous section.

For this purpose, we shall now assume that $\mathcal{X}$ is a suitable Hilbert. We now consider the following weak formulation of a parametrized PDE:

$$\text{For } \mu \in \mathcal{P} \text{ find } u(\mu) \in \mathcal{X} \,:\, a(u(\mu), v; \mu) = f(v; \mu), \quad \forall v \in \mathcal{X}. \tag{5.25}$$

We assume that $a(\cdot, \cdot; \mu) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous bilinear form and $f(\cdot; \mu) \in \mathcal{Y}$, where $\mathcal{Y} := \mathcal{X}'$ denotes the dual space of $\mathcal{X}$. Furthermore, we assume the following properties which ensure that the weak formulation (5.25) is well-posed for any $\mu \in \mathcal{P}$:

$$\begin{aligned}
&\sup_{u \in \mathcal{X}} \sup_{v \in \mathcal{X}} \frac{|a(u, v; \mu)|}{\|u\|_{\mathcal{X}} \|v\|_{\mathcal{X}}} =: c(\mu) \le c_{\text{ub}}(\mu) < \infty, \quad \text{(continuity)},\\
&\inf_{u \in \mathcal{X}} \sup_{v \in \mathcal{X}} \frac{|a(u, v; \mu)|}{\|u\|_{\mathcal{X}} \|v\|_{\mathcal{X}}} =: \beta(\mu) \ge \beta_{\text{lb}}(\mu) > 0, \quad \text{(inf-sup stability)},
\end{aligned}$$

and for each $v \in \mathcal{X} \setminus \{0\}$ there exists a $u \in \mathcal{X}$ such that $a(u, v; \mu)$ does not vanish. Under the aforementioned assumptions it is a well-know result that there exists a unique solution $x^*(\mu) \in \mathcal{X}$ to the problem (5.25) (cf. [16]).

This problem can be incorporated into the general framework (P) by defining the parameter dependent operator $G(\cdot; \mu) : \mathcal{X} \to \mathcal{Y}$ via

$$G(u; \mu)(v) := a(u, v; \mu) - f(v; \mu), \qquad \forall v \in \mathcal{X}.$$

Since $a(\cdot, \cdot; \mu)$ is bilinear, we easily conclude that $G(\cdot; \mu)$ is continuously differentiable and that its Fréchet-derivative $\mathrm{D}G|_x \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ at the point $x \in \mathcal{X}$ is given by

$$\mathrm{D}G|_x (v) = a(x, v; \mu).$$

The assumptions on $a(\cdot, \cdot; \mu)$ now guarantee, firstly, that $\mathrm{D}G|_x$ is bounded as

$$\|\mathrm{D}G|_x\|_{\mathcal{L}(\mathcal{X}, \mathcal{Y})} = c(\mu) \le c_{\text{ub}}(\mu) \tag{5.26}$$

and, secondly, that $\mathrm{D}G|_x$ has a bounded inverse for any $x \in \mathcal{X}$ for which the norm can be bounded by

$$\left\|\mathrm{D}G|_x^{-1}\right\|_{\mathcal{L}(\mathcal{Y}, \mathcal{X})} = \frac{1}{\beta(\mu)} \le \frac{1}{\beta_{\text{lb}}(\mu)}. \tag{5.27}$$

Furthermore, since $G(\cdot; \mu)$ is affine linear the (modified) validity criterion is always satisfied and the bounds of Theorem 5.3.1 and Theorem 5.3.9 hold, respectively.

Let us assume that we have a RB approximation $\hat{x}(\mu) \in \mathcal{X}_N$ for a suitable reduced basis space $\mathcal{X}_N \subset \mathcal{X}$ with $\dim(\mathcal{X}_N) = N \ll d = \dim(\mathcal{X})$. The classical RB error bound then establishes a relation between the error $e(\mu) = \hat{x}(\mu) - x^*(\mu) \in \mathcal{X}$ and the residual of the approximation via the following expression

$$\|e(\mu)\|_{\mathcal{X}} = \|\hat{x}(\mu) - x^*(\mu)\|_{\mathcal{X}} \leq \Delta_{\mathrm{RB}}(\mu) := \frac{\|G(\hat{x}(\mu);\mu)\|_{\mathcal{Y}}}{\beta(\mu)} \leq \frac{\|G(\hat{x}(\mu);\mu)\|_{\mathcal{Y}}}{\beta_{\mathrm{lb}}(\mu)}.$$

As we have seen in (5.27) the inf-sup constant corresponds to the stability constant in the abstract formulation of this thesis by taking the reciprocal value. Recalling that for linear problems we get that the split bound

$$\|e(\mu)\|_{\mathcal{X}} = \epsilon(\mu) \leq \epsilon_{\mathrm{split}}(\mu) = \gamma(\mu)\, \|G(\hat{x}(\mu);\mu)\|_{\mathcal{Y}} = \Delta_{\mathrm{RB}}(\mu) \tag{5.28}$$

coincides with the standard RB error bound. Furthermore, using (5.17) as well as (5.26) and (5.27) we obtain the bound

$$\mathrm{eff}_{\Delta_{\mathrm{RB}}}(\mu) \overset{(5.17)}{\leq} \left\| \mathrm{D}G|_x^{-1} \right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \|\mathrm{D}G|_x\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \overset{(5.26)}{=} \frac{c(\mu)}{\beta(\mu)} \overset{(5.27)}{\leq} \frac{c_{\mathrm{ub}}(\mu)}{\beta_{\mathrm{lb}}(\mu)} \tag{5.29}$$

which also coincides with the standard bound on the effectivity for the RB error bound Δ_{RB}, c.f. [76, 39]. Therefore, by not splitting the calculation of the residual and by directly applying an approximation scheme for finding a sharp, rigorous bound on $\epsilon(\mu)$, we expect to have more accurate error predictions, i.e. smaller effectivity.

To apply the improved error estimation technique we first setup the ALP, which takes the following weak form

$$a(e(\mu), v; \mu) = a(\hat{x}(\mu), v; \mu) - f(v;\mu), \quad \forall v \in \mathcal{X}. \tag{5.30}$$

Since the above equation is equally expensive to solve as the original problem, we perform the additional RB approximation for the ALP according to the method described in Section 5.5.2. Accordingly, we assume to have another reduced basis space $\mathcal{X}_M^E \subset \mathcal{X}$ with $\dim(\mathcal{X}_M^E) = M \ll \dim(\mathcal{X})$, which is used to reduce the ALP.

$$a(\widehat{e}(\mu), v_M; \mu) = a(\hat{x}(\mu), v_M; \mu) - f(v_M;\mu), \quad \forall v_M \in \mathcal{X}_M^E.$$

We emphasize once more, that this equation is M-dimensional and can be solved computationally efficient, similarly to how the RB approximation of the main problem (5.25) is calculated. Consequently, we get the improved error bound by applying Lemma 5.4.2

$$\Delta_{\mathrm{ub}}(\mu) = \|\widehat{e}(\mu)\|_{\mathcal{X}} + \frac{1}{\beta_{\mathrm{lb}}(\mu)} \|R(\mu)\|_{\mathcal{Y}},$$

where the residual $R(\mu) \in \mathcal{Y}$ is given by

$$R(\mu)(v) = a(\widehat{e}(\mu) - \hat{x}(\mu), v; \mu) + f(v; .\mu). \tag{5.31}$$

In practice, one computes the norm $\|R(\mu)\|_{\mathcal{Y}}$ via the Riesz-representer of the residual.

For many problems, the calculation of the inf-sup constant $\beta(\mu)$ poses difficulties when it comes to an efficient implementation. To circumvent this, one often employs pessimistic lower bounds $\beta_{\mathrm{lb}}(\mu) \leq \beta(\mu)$ which can be calculated rapidly. For certain problems, lower bounds can be computed by employing standard estimation techniques in the RB framework. Some examples are the min–θ scheme or the successive constraint method (SCM) (cf. [49, 72]). However, these methods are either not applicable, computationally involved or deliver highly imprecise results which render the classical RB error bounds useless. Once again, this is due to the fact that the error bound scales linearly with the inverse of the inf-sup constant (the stability constant in the general framework). To further highlight this, we consider the following scenario:

Let $\lambda \gg 1$ and set $\beta_{\mathrm{lb}}(\mu) := \frac{\beta(\mu)}{\lambda}$ as a lower bound of the inf-sup constant. We then get the following error bounds when using the classical RB error bound $\Delta_{\mathrm{RB}}(\mu)$ and the improved version $\Delta_{\mathrm{ub}}(\mu)$.

$$\Delta_{\mathrm{RB}}(\mu) = \lambda \frac{\|G(\hat{x}(\mu); \mu)\|_{\mathcal{Y}}}{\beta(\mu)}$$

and

$$\Delta_{\mathrm{ub}}(\mu) = \|\widehat{e}(\mu)\|_{\mathcal{X}} + \lambda \frac{\|R(\mu)\|_{\mathcal{Y}}}{\beta_{\mathrm{lb}}(\mu)}.$$

As mentioned before, the effectivity $\mathrm{eff}_{\Delta_{\mathrm{RB}}}(\mu)$ now directly scales linearly with the underestimation factor λ which further degrades the quality of the classical RB bound. On the contrary, for the improved error bound $\Delta_{\mathrm{ub}}(\mu)$, the scaling factor λ is counteracted by the residual norm $\|R(\mu)\|_{\mathcal{Y}}$. Recalling the results of Lemma 5.4.3, as long as the inequality

$$2\lambda \frac{\|R(\mu)\|_{\mathcal{Y}}}{\beta_{\mathrm{lb}}(\mu)} \leq \|\widehat{e}(\mu)\|_{\mathcal{X}}$$

still holds, the effectivity $\mathrm{eff}_{\Delta_{\mathrm{ub}}}(\mu)$ is still bounded by 3 and since we expect $\|R(\mu)\|_{\mathcal{Y}} \ll \|G(\hat{x}(\mu);\mu)\|_{\mathcal{Y}}$, even severe underestimations have a lesser impact. The assumption that the norm of the residual $R(\mu)$ is (much) smaller than the residual of the original problem is not unfounded. Recalling Remark 5.4.1 we can interpret $-\widehat{e}(\mu)$ as a quasi Newton-update. Hence $\hat{x}(\mu) - \widehat{e}(\mu)$ should provide a better estimate. Using (5.31) we can see that

$$R(\mu) = a(\hat{x}(\mu) - \widehat{e}(\mu), v; \mu) - f(v; .\mu) = G(\hat{x}(\mu) - \widehat{e}(\mu); \mu)$$

and consequently $\|R(\mu)\|_{\mathcal{Y}} \ll \|G(\hat{x}(\mu);\mu)\|_{\mathcal{Y}}$ seems reasonable, provided the approximation $\widehat{e}(\mu)$ is of sufficient quality.

In total, the better absorbability of underestimation in the inf-sup constant might be useful in cases for which a (possibly pessimistic) global lower bound $\beta(\mu) \geq \bar{\beta} > 0$ for all parameters $\mu \in \mathcal{P}$ is available, as expensive estimations techniques for pointwise lower bounds $\beta(\mu) \geq \beta_{\mathrm{lb}}(\mu)$ can be avoided. We will study the influence of underestimation in the inf-sup constant, i.e. overestimation of the stability constant, experimentally in Section 5.6.

5.5.4 Offline/Online efficient implementation

In this section, we address how an efficient implementation of the proposed error quantification can be realised. Similar to the structure of the approximation methods using matrix-valued kernels, which were discussed in the first part of this thesis, we apply the same principle here. Namely, we can split the computation into two phases:

(a) the offline phase, in which all quantities necessary for the computation of the approximation and the error quantification are precomputed

(b) an online phase, in which the reduced problem $(P_N(\mu))$ and the reduced ALP $(P_M^E(\mu))$ are solved, the approximation $\hat{x}(\mu)$ is constructed and the error bounds are evaluated.

Before going into further detail, as to what each of these phases contain, we first want to introduce the notion of parameter separability of the problem, which is a classical assumption in the RB framework.

Definition 5.5.4 (Parameter separability)**.** We say a parametric dependent function $G(\cdot;\cdot) : \mathcal{X} \times \mathcal{P} \to \mathcal{Y}$ is parameter separable, if there exist a $Q \in \mathbb{N}$ and parameter dependent coefficient functions $\Theta_q : \mathcal{P} \to \mathbb{R}$, $q = 1, \dots, Q$ and parameter independent operators $G_q : \mathcal{X} \to \mathcal{Y}$, $q = 1, \dots, Q$ such that

$$G(\cdot; \mu) = \sum_{q=1}^{Q} \Theta_q(\mu) G_q(\cdot). \tag{5.32}$$

In this case, we call the set of tuples $\{(\Theta_1, G_1), \dots, (\Theta_Q, G_Q)\}$ a parameter separable decomposition of G.

Obviously, such a decomposition does not necessarily exist for any given operator G and even if one does exist, this does not mean that we have ready access to it. In this case one can employ interpolation techniques, such as the (discrete) empirical interpolation method ((D)EIM) (cf. [60, 20]) to generate an approximation of G which is parameter separable. We further note, that the approximations provided by the DEIM result in (affine) linear operators $G_q : \mathcal{X} \to \mathcal{Y}$. Hence, we assume in the context of this section that G is parameter separable and that the operators G_q are (affine) linear. This property is now inherited by both its derivative $\mathrm{D}G$ as well as the reduced problem operator G_N.

Lemma 5.5.5. *If G is parameter-separable with a decomposition $\{(\Theta_1, G_1), \dots, (\Theta_Q, G_Q)\}$, then both $\mathrm{D}G$ and G_N are parameter separable with the same coefficient functions Θ_q and operators given by*

$$G_{N,q} \coloneqq \Pi_{\mathcal{Y}_N}(G_q|_{\mathcal{X}_N}) \quad \text{and} \quad (\mathrm{D}G)_q \coloneqq \mathrm{D}G_q.$$

Proof. For G_N the above follows directly from the definition of the reduced problem operator, the linearity of the restriction on $\mathcal{X}_N$ and the linearity of the projection onto $\mathcal{Y}_N$. For $\mathrm{D}G$ the result is a direct consequence of the linearity of the differentiation operator. □

In a similar fashion one can show that the residual $R(\mu)$ of the ALP and the projection $\Pi_{\mathcal{Y}_M^E}\left[\mathrm{D}G(\cdot;\mu)|_{\hat{x}(\mu)}\right]\Big|_{\mathcal{X}_M^E}$ possess a parameter separable decomposition.

In the context of the RB approximation procedure and our improved error quantification method, these results enable us to efficiently compute the relevant quantities. To this end we assume now that $\mathcal{X}$ is a finite but high-dimensional vector space. Hence we can identify $\mathcal{X} = \mathbb{R}^d$ and the elements of the reduced spaces $\mathcal{X}_N$ and $\mathcal{X}_M^E$ can be represented via

$$x = V_{\mathcal{X}_N} x_N \quad \text{and} \quad x = V_{\mathcal{X}_M^E} x_M^E,$$

where $x_N \in \mathbb{R}^N, x_M^E \in \mathbb{R}^M$ and matrices $V_{\mathcal{X}_N} \in \mathbb{R}^{d\times N}, V_{\mathcal{X}_M^E} \in \mathbb{R}^{d\times M}$ representing a basis of $\mathcal{X}_N$ and $\mathcal{X}_M^E$, respectively. Similarly, we can represent $\mathcal{Y}_N$ and $\mathcal{Y}_M^E$ as coefficient matrices $V_{\mathcal{Y}_N} \in \mathbb{R}^{d\times N}$ and $V_{\mathcal{X}_M^E} \in \mathbb{R}^{d\times M}$ pertaining to (orthogonal) basis of the respective spaces.

The operators in the parameter separable decomposition of G_N then take the form

$$G_{N,q} = V_{\mathcal{Y}_N}^T G_q V_{\mathcal{X}_N} \in \mathbb{R}^{N \times N}$$

and since they are parameter independent, and small, they can be precomputed in the offline phase and stored for later use during the online phase. In total, we can now provide a more detailed overview of how the different parts of the approximation and error quantification procedure can be split into an offline and online phase:

(a) offline phase:

- If necessary: Compute a parameter separable decomposition of G via DEIM
- Compute a reduced space $\mathcal{X}_N$ with suitable representation via $V_{\mathcal{X}_N}$ as well as $\mathcal{Y}_N$ and $V_{\mathcal{Y}_N}$
- Compute the low-dimensional matrix representations of the parameter independent operators in the decomposition for G_N
- Compute a reduced space $\mathcal{X}_M^E$ with representation $V_{\mathcal{X}_M^E}$ as well as $\mathcal{Y}_M^E$ and $V_{\mathcal{Y}_M^E}$
- Compute the low-dimensional matrix representations of the parameter independent operators for both the reisudal $R(\mu)$ and the problem operator $\Pi_{\mathcal{Y}_M^E} \left[\mathrm{D}G(\cdot;\mu)|_{\hat{x}(\mu)} \right] \Big|_{\mathcal{X}_M^E}$ of the reduced ALP
- If possible: compute global lower (upper) bounds for the inf-sup constant (stability constant)

(b) online phase (for a given parameter $\mu \in \mathcal{P}$):

- Assemble G_N via its parameter separable decomposition
- Solve the reduced problem and obtain the approximation $\hat{x}(\mu)$
- Assemble the reduced ALP for the current approximation $\hat{x}(\mu)$ using the parameter separable decomposition
- Solve the reduced ALP to obtain $\hat{E}(\mu)$.
- Compute $\epsilon_{\mathrm{ub}}(\mu)$ using the parameter separability of the residual $R(\mu)$
- Evaluate the error bound

5.5.5 Basis generation

In this section we take a closer look at how suitable reduced bases $\mathcal{X}_N$ can be constructed. Furthermore, we set the projection space $\mathcal{Y}_N$ to be qual to the reduced spaces, i.e. $\mathcal{Y}_N = \mathcal{X}_N$. We focus on snapshot based techniques [8]. In this case the subspace is contained in the span of several true solutions, i.e. $\mathcal{X}_N \subset \operatorname{span}(\{x^*(\mu_1), \dots, x^*(\mu_N)\})$, so called snapshots, for suitable parameters $\mu_1, \dots, \mu_N \in \mathcal{P}$. Two popular methods which make use of this principle are the proper orthogonal decomposition (POD) method [105] and those contained under the framework of greedy algorithms [104]. We will roughly summarize these two approaches, however, we refer to the existing literature for a more detailed introduction into the topic. The POD works as follows:

Given a set of snapshots $S = \{x^*(\mu_1), \dots, x^*(\mu_n)\}) \subset \mathcal{X}$ we can construct the corresponding empirical correlation operator $\mathcal{R} : \mathcal{X} \to \mathcal{X}$ via

$$\mathcal{R}(x) = \frac{1}{n} \sum_{i=1}^{n} x^*(\mu_i) \langle x^*(\mu_i), x \rangle_{\mathcal{X}}.$$

It immediately follows that $\mathcal{R}$ is self-adjoint and since it has a finite-dimensional range, it is also compact. Thus, by the spectral theorem, there exists an orthornomal basis $\{\phi_1, \dots, \phi_n\}$ for the range $\operatorname{range}(\mathcal{R}) = \operatorname{span}(S)$ of the empirical correlation operator, such that ϕ_i is an eigenvector of $\mathcal{R}$ to the real eigenvalue λ_i, where the eigenvalues are in descending order, i.e. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$. The subspaces $V_N := \operatorname{span}(\{\phi_1, \dots, \phi_N\})$ stemming from this procedure now possess a best-approximation property with respect to the squared error, i.e.

$$\inf_{V \subset \mathcal{X}, \dim(V) = N} \frac{1}{n} \sum_{i=1}^{n} \|x^*(\mu_i) - \Pi_V x^*(\mu_i)\|_{\mathcal{X}}^2 = \frac{1}{n} \sum_{i=1}^{n} \|x^*(\mu_i) - \Pi_{V_N} x^*(\mu_i)\|_{\mathcal{X}}^2 = \sum_{i=N+1}^{n} \lambda_i.$$

In other words, V_N is the subspace of dimension N, that provides an optimal approximation of the given data with regards to the sum of squared errors. Furthermore, the quality of the approximation is quantified by summing the remaining eigenvalues of eigenfunctions not included in the approximation space V_N. If we now assume $x^*(\mu) \in \operatorname{span}(S)$, then V_N is a suitable choice for an approximation space, provided that N is chosen sufficiently large such that $\sum_{i=N+1}^{n} \lambda_i$ is small. Unfortunately, the empirical correlation operator $\mathcal{R}$ acts on the high- or even infinite-dimensional space $\mathcal{X}$ and therefore, the calculation of the basis elements ϕ_i by solving a corresponding eigenvalue problem is infeasible. Fortunately, this can be circumvented as each eigenvector ϕ_i can be related to the eigenvectors of the gram-matrix corresponding to the snapshots in S, given by

$$A_S = (\langle x^*(\mu_i), x^*(\mu_j) \rangle_{\mathcal{X}})_{i,j=1}^{n} \in \mathbb{R}^{n \times n}.$$

The gram-matrix A_S now has the eigenvalues $\lambda_i \cdot n$ and let $v_i = (v_{i,1}, \dots, v_{i,n})^T \in \mathbb{R}^n$ denote the eigenvector for the eigenvalue $\lambda_i \cdot n$. Then we have the following representation of the eigenvector ϕ_i of $\mathcal{R}$:

$$\phi_i = \sum_{j=1}^{n} v_{i,j} x^*(\mu_j).$$

In contrast to the eigenvalue problem for $\mathcal{R}$, the eigenvalue problem for the gram-matrix A_S is only of dimension n and therefore more readily solvable. To further illustrate the above connection, let us consider that $\mathcal{X}$ is a finite (albeit high-dimensional) Hilbert-space, i.e., we can represent $\mathcal{X}$ via $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Combining the snapshots $x^*(\mu_1), \ldots, x^*(\mu_n)$ into a matrix

$$X_S = \begin{pmatrix} x^*(\mu_1) & \cdots & x^*(\mu_n) \end{pmatrix} \in \mathbb{R}^{d \times n},$$

the correlation operator $\mathcal{R}$ and the gram-matrix then take the form

$$\mathcal{R} = \frac{1}{n} X_S X_S^T \in \mathbb{R}^{d \times d} \quad \text{and} \quad A_S = X_S^T X_S \in \mathbb{R}^{n \times n}.$$

By the above, we can observe that the optimal approximation spaces V_N can be computed by performing a singular value decomposition for the snapshot-matrix X_S, where the singular values σ_i correspond to $\sigma_i^2 = \lambda_i n$.

Please note, that the procedure outlined above works quite well if $n \ll d$, i.e. if the snapshot-matrix X_S is skinny. In cases where much more snapshots are available one might have to consider more sofisticated methods such as the hierarchical approximate POD [48].

We summarize the POD procedure in the following algorithm

Algorithm 6: Proper orthogonal decomposition

```
Data: Snapshots $S = \{x^*(\mu_1), \ldots, x^*(\mu_n)\}$, tolerance $\rho > 0$.
Result: Subspace $\mathcal{X}_N$.
Compute singular values $\sigma_1, \ldots, \sigma_n$ and corresponding singular vectors $v_1, \ldots, v_n$
  of the snapshot matrix $X_S$ corresponding to the snapshot set $S$, initialize $N = 0$
  and $X_0 = \{\}$;
while $\sum_{i=N+1}^{n} \sigma_i^2 > \rho$ do
    Extend subspace $\mathcal{X}_{N+1} = \mathcal{X}_N \bigoplus \mathrm{span}(v_{N+1})$;
    Increment $N \coloneqq N + 1$;
end
```

Greedy procedures for the construction of a subspace $\mathcal{X}_N$ are based on the same principle as the greedy algorithms outlined in the first part of this dissertation. However, we shall recall the specific structure in the context of the RB framework: Starting from an initial subspace $\mathcal{X}_0 \subset \mathcal{X}$ and a finite set of training parameters $\mathcal{P}_{\text{train}} \subset \mathcal{P}$, the maximum approximation error is sought by evaluating an error indicator $\delta(\cdot;\mu) : \mathcal{X} \to \mathbb{R}_+$ for all approximation with parameters in the training set $\mathcal{P}_{\text{train}}$. The subspace is then incrementally extended with the element that maximizes the error indicator until a certain tolerance is met. The procedure is outlined in the following algorithm

Algorithm 7: Greedy algorithm($\mathcal{P}_{\text{train}}$,$\rho$,$\delta$,$\mathcal{X}_0$)

```
Data: Training set $\mathcal{P}_{\text{train}}$, greedy tolerance $\rho$, error indicator $\delta$, initial subspace $\mathcal{X}_0$.
Result: Subspace $\mathcal{X}_N$.
Initialize $N = 0$, solve the full problem for $x_0^* \in \mathcal{X}_0$ while
 $\max_{\mu \in \mathcal{P}_{train}} \delta(x_N^*(\mu);\mu) > \rho$ do
    Set $\mu^* := \arg\max_{\mu \in \mathcal{P}_{\text{train}}} \delta(x_N^*(\mu);\mu)$;
    Solve the full problem $G(x;\mu^*) = 0$ for $x_{N+1}^*(\mu^*) \in \mathcal{X}$;
    Extend subspace $\mathcal{X}_{N+1} := \mathcal{X}_N \bigoplus \text{span}(x_{N+1}^*(\mu^*))$;
    Increment $N := N + 1$;
end
```

Please note that the error indicator $\delta(\cdot;\mu)$ is just that; an indicator. This means that $\delta(\cdot;\mu)$ does not necessarily have to be a (rigorous) bound on the approximation error, however, it should still capture the behaviour of the (true) error. This relaxation is allowed, as the computation of high quality error bounds is often computationally expensive, i.e. time consuming, or infeasible. Nonetheless, it can be favourable to use a more expensive error indicator $\delta(\cdot;\mu)$, since they should lead to an approximation space $\mathcal{X}_N$ of superior quality.

For the approximation of the ALP, we have to identify another reduced space $\mathcal{X}_M^E$. We proceed analogously to the construction of $\mathcal{X}_N = \mathcal{Y}_N$ by first restricting ourselves to the case $\mathcal{X}_M^E = \mathcal{Y}_M^E$. The RB space is then constructed by another greedy algorithm, where the original problem $(P(\mu))$ is replaced by the ALP $(P^E(\mu))$. In a similar fashion, we have to chose a suitable tolerance and a suitable error indicator. In our case we use

$$\delta(\hat{E}(\mu);\mu) := \Delta_{\text{RB}}^E = \frac{\|R(\mu)\|_{\mathcal{Y}}}{\beta_{\text{lb}}(\mu)}$$

which corresponds to the standard RB error bound for the reduced basis generation of the ALP.

Preferably, we would like to use our improved error bounds $\Delta_{\text{ub}}(\mu)$ or $\hat{\Delta}_{\text{ub}}(\mu)$ as error indicators in the construction of the reduced space $\mathcal{X}_N$ via the greedy algorithm. However, the ALP $(P^E(\mu))$, and henceforth, the space $\mathcal{X}_M^E$ are conditional on the reduced space $\mathcal{X}_N$. Therefore, we would need to reconstruct individual spaces $\mathcal{X}_M^E$ in each iteration during the construction of $\mathcal{X}_N$, i.e. in each greedy-step a separate full greedy algorithm would have to be performed. Unfortunately, this proves to be highly computationally expensive and thus we defer to a sequential computation of the spaces which makes the construction of both spaces computationally feasible. The pseudocode for this sequential double greedy algorithm is outlined in Algorithm 8.

Remark 5.5.6. The computational efficiency of our a-posteriori estimator is directly influenced by $\dim(\mathcal{X}_M^E)$. Thus, if one wants to achieve a fast online phase including error quantification, one might have to make sacrifices in the quality of the error space such that the computational overhead is comparable to the computational demands required for solving the reduced problem $(P_N(\mu))$.

Algorithm 8: Sequential Double Greedy algorithm($\mathcal{P}_{\text{train}}$, $\mathcal{P}^E_{\text{train}}$, ρ, ρ_E, δ, δ_E, $\mathcal{X}_0$, $\mathcal{X}_0^E$).

```
Data: Training sets $\mathcal{P}_{\text{train}}, \mathcal{P}^E_{\text{train}}$, greedy tolerances $\rho, \rho^E$, error indicators $\delta, \delta_E$,
      initial subspaces $\mathcal{X}_0, \mathcal{X}_0^E$.
Result: Subspaces $\mathcal{X}_N$ and $\mathcal{X}_M^E$.
Initialize $N = 0, x_0^* \in \mathcal{X}_0$. while $\max_{\mu \in \mathcal{P}_{train}} \delta(x_N^*(\mu); \mu) > \rho$ do
    Set $\mu^* := \arg\max_{\mu \in \mathcal{P}_{\text{train}}} \delta(x_N^*(\mu); \mu)$;
    Solve full problem $G(x; \mu^*) = 0$ for $x_{N+1}^*(\mu^*) \in \mathcal{X}$;
    Extend subspace $\mathcal{X}_{N+1} := \mathcal{X}_N \bigoplus \text{span}(x_{N+1}^*(\mu^*))$;
    Increment $N := N + 1$;
end
Initialize $M = 0, E_0 \in \mathcal{X}_0^E$. while $\max_{\mu \in \mathcal{P}^E_{train}} \delta_E(E_M(\mu); \mu) > \rho_E$ do
    Set $\mu^* := \arg\max_{\mu \in \mathcal{P}^E_{\text{train}}} \delta_E(E_M(\mu); \mu)$;
    Compute the approximation $\hat{x}(\mu^*) \in \mathcal{X}_N$;
    Solve the full problem $\mathrm{D}G|_{\hat{x}(\mu^*)}(E; \mu^*) = G(\hat{x}(\mu^*); \mu^*)$ for $E_{N+1}(\mu^*) \in \mathcal{X}$;
    Extend subspace $\mathcal{X}_{M+1}^E := \mathcal{X}_M^E \bigoplus \text{span}(E_{N+1}(\mu^*))$;
    Increment $M := M + 1$;
end
```

5.6 Experimental validation for RB approximations

In this section we evaluate the proposed a-posteriori error estimation theory in the context of the RB method. The first example is a well-known thermal-block test case, modelling a

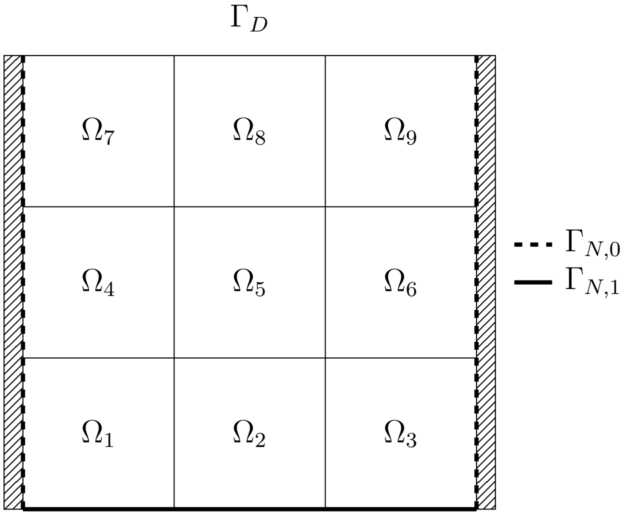

Figure 5.2: Illustration of the thermal block setting used in the examples.

parametric heat conduction problem on the unit square. Here we will see that by making use of the proposed method, we are able to reach effectivities arbitrarily close to one. The second example shows the application of the framework to a nonlinear finite-dimensional problem that stems from a semi-discretized parametric PDE with a non-variational finite difference (FD) discretization. All examples are implemented in the toolbox RBmatlab[1] and were run on a machine with an Intel Core i7-6700 CPU with 16GB RAM in MATLAB 2019a. In all experiments, the test parameters are distinct from the training parameters.

5.6.1 Linear test case: Thermal block model

We consider the thermal block example, which was previously used in Section 4.4. We recapitulate the problem description in the following and make slight alterations to describe it in a more general setting. An Illustration of the specific setting used in the numerical experiments is given in Figure 5.2. It consists of a steady linear heat equation on the unit square $\Omega = (0,1)^2$, which is divided into $B \coloneqq B_1 \cdot B_2$ subblocks, where $B_1, B_2 \in \mathbb{N}$ describe the number of subblocks per dimension. We denote the subblocks by Ω_i for $i = 1, \dots B$, counted row-wise starting from the left bottom. We prescribe a unit flux into the domain on the bottom boundary, which is denoted as $\Gamma_{N,1}$ with unit outward normal $n(\xi)$, where $\xi \in \Omega$ indicates the spatial variable. The left and right boundary part $\Gamma_{N,0}$ is insulated, which is modeled by a zero Neumann boundary condition and the top Dirichlet boundary Γ_D has constant 0 temperature. The parametric PDE for the temperature field

[1] http://www.morepas.org

$u(\cdot;\mu):\Omega\to\mathbb{R}$ for this example is given as

$$\begin{aligned}
-\nabla\cdot(\kappa(\xi;\mu)\nabla u(\xi;\mu)) &= 0, && \xi\in\Omega,\\
u(\xi;\mu) &= 0, && \xi\in\Gamma_D,\\
(\kappa(\xi;\mu)\nabla u(\xi;\mu))\cdot n(\xi) &= i, && \xi\in\Gamma_{N,i}, i=0,1,
\end{aligned}$$

where we define the heat conductivity function

$$\kappa(\cdot;\mu):\Omega\to\mathbb{R}_+,\quad \kappa(\xi;\mu):=\sum_{i=1}^{B}\mu_i\chi_{\Omega_i}(\xi),$$

using the indicator function χ_A for sets $A\subset\Omega$. The parametric domain for this problem is given as $\mathcal{P}\coloneqq[1/\mu_{\max},\mu_{\max}]^B$ for some $\mu_{\max}>1$. With the function space $\mathcal{X}=H^1_D(\Omega)\coloneqq\{v\in H^1(\Omega)\mid v|_{\Gamma_D}=0\}$ and its dual X', we can define a weak formulation of the above PDE via

$$\text{For } \mu\in\mathcal{P} \text{ find } u(\mu)\in\mathcal{X}\,:\,a(u(\mu),v;\mu)=f(v;\mu),\quad \forall v\in\mathcal{X},$$

where the bilinear form $a(\cdot,\cdot;\mu):\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ and the right-hand side $f:\mathcal{X}\to\mathbb{R}$ are given by

$$\begin{aligned}
a(u,v;\mu) &= \int_\Omega \kappa(\xi;\mu)\nabla u(\xi)\cdot\nabla v(\xi)\mathrm{d}\xi\\
f(v;\mu) &= \int_{\Gamma_{N,1}} v(\xi)\mathrm{d}\xi.
\end{aligned}$$

Following the results of Section 5.5.3, the parametrized problem operator $G(\cdot;\mu):\mathcal{X}\to\mathcal{X}'$ is now given by

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

Please note, that by definition, we have

$$\begin{aligned}
a(u,v;\mu) &= \int_\Omega \kappa(\xi;\mu)\nabla u(\xi)\cdot\nabla v(\xi)\mathrm{d}\xi = \int_\Omega\left(\sum_{i=1}^{B}\mu_i\chi_{\Omega_i}(\xi)\right)\nabla u(\xi)\cdot\nabla v(\xi)\mathrm{d}\xi\\
&= \sum_{i=1}^{B}\mu_i\int_{\Omega_i}\nabla u(\xi)\cdot\nabla v(\xi)\mathrm{d}\xi = \sum_{i=1}^{B}\mu_i a_i(u,v).
\end{aligned}$$

Hence a and consequently G are parameter-separable as defined in equation (5.32).

We further equip the space $\mathcal{X}$ with the norm

$$\|u\|_{\mathcal{X}} = \|u\|_{H^1_D(\Omega)} = \left(\int_{\Omega} \nabla u(\xi) \cdot \nabla u(\xi) \mathrm{d}\xi\right)^{1/2}.$$

It is a well-known fact that for every $\mu \in \mathcal{P}$ this problem possesses a unique solution $u^*(\mu) \in \mathcal{X}$. In the later analysis, we also consider the so-called energy norm which, for a fixed parameter $\bar{\mu} \in \mathcal{P}$, is given by

$$\|u\|_{\bar{\mu}} = (a(u, u; \bar{\mu}))^{1/2}.$$

We remark, that the above norm is well-defined as the bilinear form a is coercive for every $\mu \in \mathcal{P}$ with respect to the norm on $\mathcal{X}$ i.e.

$$a(u, u, \mu) \geq \beta(\mu) \|u\|^2_{\mathcal{X}} \geq \beta_{\mathrm{lb}}(\mu) \|u\|^2_{\mathcal{X}}$$

Recalling the results of Section 5.5.3, namely equation (5.27), this guarantees the invertibility of DG at every point and provides a bound on the stability constant via $\gamma_{\mathrm{ub}}(\mu) \leq \beta_{\mathrm{lb}}(\mu)^{-1}$. We further remark, that the standard norm we chose on $\mathcal{X}$ is just a special case of the energy norm, where $\bar{\mu}$ is a vector whose components are all 1.

For our first test we pick $B_1 = B_2 = 3$ which leads to a total of 9 different parameters. Furthermore, we choose $\mu_{\max} = 10$ resulting in $\mathcal{P} = [0.1, 10]$. For the truth-approximation we apply a finite-element approximation of the PDE using piecewise linear elements. This results in a $d = 3721$ dimensional problem. The basis $V_{\mathcal{X}_N}$ of the RB space $\mathcal{X}_N$ is then computed via the greedy algorithm (Algorithm 7), where we have chosen a tolerance $\rho = 10^{-3}$ and a training set $\mathcal{P}_{\mathrm{train}}$ consisting of 1000 randomly selected parameters. For the error indicator we use the standard RB error bound, i.e.

$$\delta(\hat{u}(\mu)) := \Delta_{\mathrm{RB}}(\mu) = \frac{\|a(\hat{u}(\mu), \cdot; \mu) - f(\cdot; \mu)\|_{\mathcal{X}'}}{\beta_{\mathrm{lb}}(\mu)} \leq \gamma_{\mathrm{ub}}(\mu) \|G(\hat{u}; \mu)\|_{\mathcal{X}'}$$

which yields a basis of size $N = 62$.

As described in Algorithm 8, we again employ a greedy algorithm for the construction of the reduced space $\mathcal{X}^E_M$ for the approximation of the ALP

$$\mathrm{D}G|_{\hat{u}(\mu)} (E(\mu)) = G(\hat{u}(\mu); \mu),$$

where $\hat{u}(\mu) \in \mathcal{X}_N$ denotes the RB-approximation using the RB space $\mathcal{X}_N$. For the greedy algorithm, we again choose the standard RB error bound for the ALP as an error indicator, i.e.

$$\delta_E(\mu) := \frac{\left\| \mathrm{D}G|_{\hat{u}(\mu)} (\hat{E}(\mu)) - G(\hat{u}(\mu)) \right\|_{\mathcal{X}'}}{\beta_{\mathrm{lb}}(\mu)}.$$

As a training set, we again choose $\mathcal{P}_{\mathrm{train}}$ consisting of 1000 randomly selected parameters. For the tolerance, we chose $\rho_E = 10^{-8}$. Such a small tolerance is required, as an empty initialization, i.e. $\hat{E}(\mu) = 0$ already leads to

$$\delta_E = \frac{\|G(\hat{u}(\mu))\|_{\mathcal{X}'}}{\beta_{\mathrm{lb}}(\mu)} = \delta(\mu) \leq \rho = 10^{-3}.$$

In other words, we can interpret the above as a continuation of the greedy for the original problem. We can observe this in the decay of the maximum error indicator for increasing basis size, which is depicted in Figure 5.3. This basis construction results in a subspace $\mathcal{X}_M^E$ of dimension $M = 118$

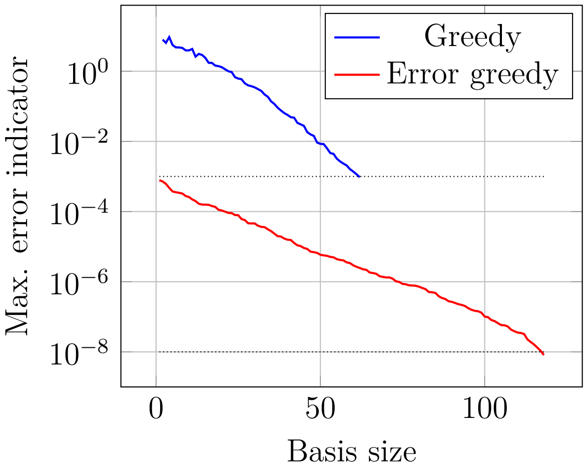

Figure 5.3: Decay of error indicator for the primal greedy and the greedy for the ALP.

In the following, we compare the improved error estimation techniques, that were presented in the previous sections of this thesis, to the standard error bounds, that are very widely used in the RB context. As a first test, we use the $H_D^1(\Omega)$-norm for the evaluation of the error bound and pick 20 random test parameters for the evaluation. For this test, we calculate the exact value of the stability constant $\gamma_{\mathrm{ub}}(\mu) := \gamma(\mu)$ by solving a d-dimensional eigenvalue problem. While this is obviously not online efficient, we still make use of the explicit calculation of the stability constants as the purpose of this first test is to solely demonstrate the improved quality of the error estimates. The results are presented in Figure 5.4, where we plotted the error, the standard RB bound, as well as our improved bounds for various sizes of the approximation space $\mathcal{X}_M^E$ corresponding to different thresholds in the greedy tolerance ρ_E.

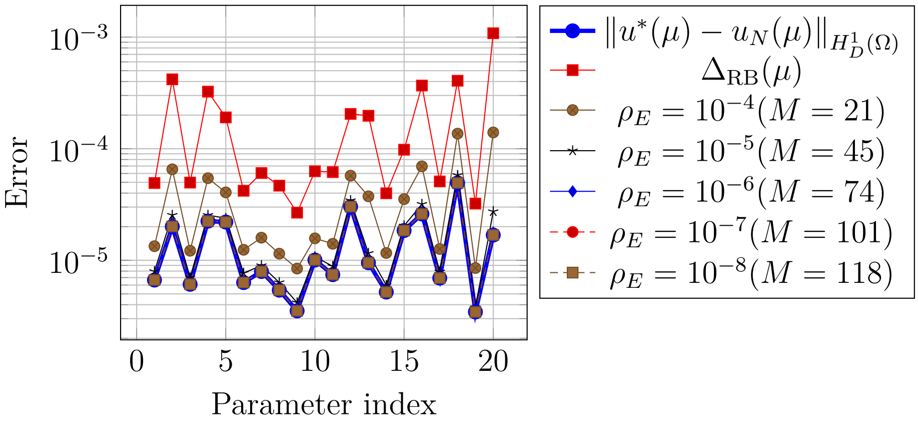

Figure 5.4: Test 1: Absolute error measured in the $H^1_D(\Omega)$-norm for 20 random test parameters.

We recall, as shown in (5.28), that the standard RB bounds corresponds to the split bound. The results in Figure 5.4 show a very large overestimation in the range of $10-100$ for all test parameters. Which confirms our theoretical findings in (5.29) where we saw that the effectivity of the standard RB bound is bounded by

$$\left\| \mathrm{D}G|^{-1}_{\hat{u}(\mu)} \right\|_{\mathcal{L}(\mathcal{X}',\mathcal{X})} \left\| \mathrm{D}G|_{\hat{u}(\mu)} \right\|_{\mathcal{L}(\mathcal{X},\mathcal{X}')} \leq \frac{\max(\mu)}{\min(\mu)} \leq 100.$$

For the above we made use of the fact, that

$$\inf_{u\in\mathcal{X}} \sup_{v\in\mathcal{X}} a(u,v;\mu) \geq \min(\mu) \geq \frac{1}{10}$$
$$\sup_{u\in\mathcal{X}} \sup_{v\in\mathcal{X}} a(u,v;\mu) \leq \max(\mu) \leq 10$$

which follows directly from the parameter-separable decomposition of the bilinear form.

In the same figure, we can see that the approximation of the residual $\epsilon(\mu) = \left\| \mathrm{D}G|^{-1}_{\hat{u}(\mu)} \left(G(\hat{u}(\mu))\right) \right\|_{\mathcal{X}}$, is improved for larger space dimensions M, which leads to an increased quality of the improved error bounds. In particular, for the tolerances $\rho_E \in \{10^{-6}, 10^{-7}, 10^{-8}\}$ ($M \in \{74, 101, 118\}$), we get an almost exact error prediction.

To verify the bounds on the effectivity derived in Lemma 5.4.3, we compute the bounds as well as the actual effectivites for the reduced spaces relating to the tolerances $\rho_E \in \{10^{-5}, 10^{-6}, 10^{-7}\}$. The results are plotted in Figure 5.5. We can see that for increasing space dimension, the quality of the effectivity bounds improves and all theoretical bounds are verified.

We recall that for the effectivity bound to hold, we need the inequality

$$2\gamma(\mu) \left\| R(\mu) \right\|_{\mathcal{X}'} = 2\gamma(\mu) \left\| \mathrm{D}G|_{\hat{u}(\mu)} (\hat{E}(\mu)) - G(\hat{u}(\mu)) \right\|_{\mathcal{X}'} \leq \left\| \hat{E}(\mu) \right\|_{\mathcal{X}}$$

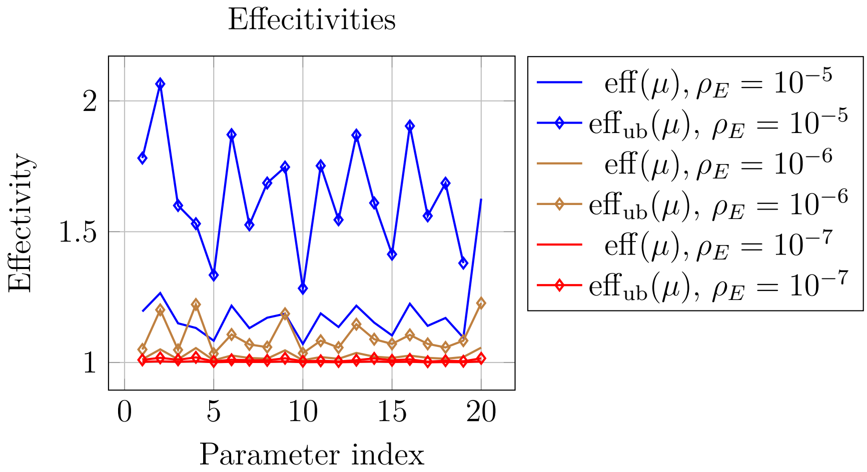

Figure 5.5: Test 1: Effectivity of the $H^1_D(\Omega)$-norm error bound for 20 random test parameters.

to be satisfied. The values of the left- and right-hand side for varying RB space dimensions of $\mathcal{X}^E_M$ are displayed in Figure 5.6. We can see, that once the space $\mathcal{X}^E_M$ is sufficiently rich, here $M \geq 40$ ($\rho_E \leq 10^{-5}$), the bounds of Lemma 5.4.3 apply. Furthermore, for increasing basis size, the quotient between the left- and right-hand side tends to 0, which in turn leads to effectivities, and bounds on the effectivities, that are close to 1.

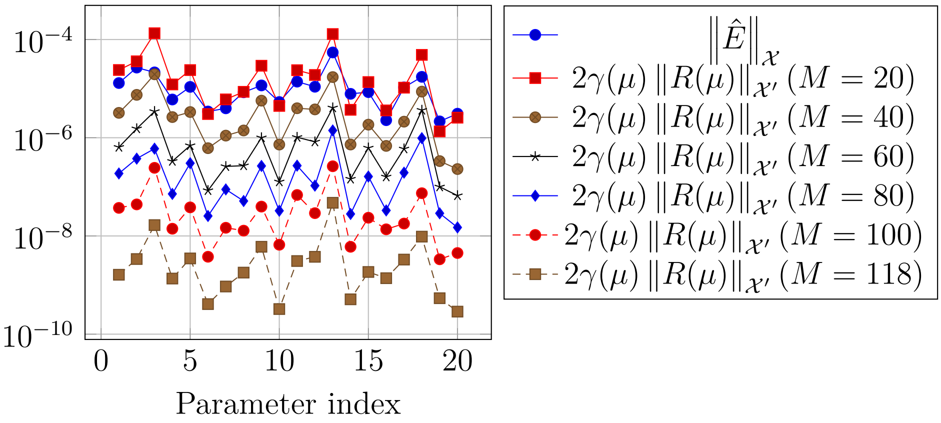

Figure 5.6: Comparison of the non-split residual approximation and the weighted error residual for approximation space size $N = 62$ and varying error space sizes M.

For the next test, we pick a larger test set $\mathcal{P}_{\text{test}} \subset \mathcal{P}$ consisting of 100 randomly selected parameters. We compare the error estimation for the $H^1_D(\Omega)$ and energy norm $\|\cdot\|_{\bar{\mu}}$, where we pick the parameter $\bar{\mu} = (1,2,1,2,1,2,1,2,1)^T \in \mathcal{P}$. Furthermore, we artificially worsen our stability constant by multiplying them with $\lambda \in \{1, 10, 100\}$. This represents an overestimation of the stability constant by the corresponding factor λ, i.e. $\gamma(\mu) \leq \gamma_{\text{ub}}(\mu) := \lambda \cdot \gamma(\mu)$, and is used to quantify the influence of possible overestimation on the quality of our error bounds. The mean and maximum value of the effectivities for varying space dimensions M are displayed in Table 5.1. The first row corresponds to the classic RB error bounds, which can be interpreted as using an approximation space $\mathcal{X}_M^E$ of dimension 0. In all cases we can observe a decay for decreasing tolerances ρ_E, i.e. richer subspaces $\mathcal{X}_M^E$ for the ALP. In particular, we obtain exact error prediction for the largest basis ($\rho = 10^{-8}$) and for $\lambda = 1$ for both, the $H_D^1(\Omega)$ and the energy norm. As expected, the influence of the overestimation factor λ has significantly less impact on the improved error bounds than on the classical RB error bounds. We recall that the RB error bounds scale linearly with λ, hence for $\lambda = 100$ the mean effectivity would increase from 64.27 to 6427 in the $H_D^1(\Omega)$ norm, whereas for the improved error bound (for $M = 118$), we only observe an increase from 1 to 1.22, i.e. a factor of 1.22.

			Maximum		Mean	
ρ_E	M	λ	$\lVert\cdot\rVert_{H_D^1(\Omega)}$	$\lVert\cdot\rVert_{\bar{\mu}}$	$\lVert\cdot\rVert_{H_D^1(\Omega)}$	$\lVert\cdot\rVert_{\bar{\mu}}$
		1	64.27	42.83	9.22	9.24
$1 \cdot 10^{-4}$	21	1	9.94	7.02	2.29	2.28
$1 \cdot 10^{-5}$	45	1	2.86	2.25	1.19	1.18
$1 \cdot 10^{-6}$	74	1	1.31	1.21	1.02	1.02
$1 \cdot 10^{-7}$	101	1	1.03	1.02	1	1
$1 \cdot 10^{-8}$	118	1	1	1	1	1
$1 \cdot 10^{-8}$	118	10	1.02	1.01	1	1
$1 \cdot 10^{-8}$	118	100	1.22	1.15	1.02	1.02

Table 5.1: Test 1: Maximum and mean effectivity of the error estimate for the thermal block example in three different norms. The first row shows the results for the standard RB bound Δ_{RB}, the remaining rows for the proposed improved error estimate.

Finally, we want to investigate the relation between the dimesnion of the RB spaces $\mathcal{X}_N$ and $\mathcal{X}_M$ when a certain effectivity is prescribed. To this end, we run the sequential double greedy algorithm (Algorithm 8) for $N \in \{10, 20, 30, 40, 50, 60\}$. We then select the basis size M of $\mathcal{X}_M$ in such a way that the effectivity of the error bound Δ_{ub} is smaller than $8, 4, 2$ or 1.1 on a test set of 50 randomly chosen parameters. The results are depicted in Figure 5.7. For the cases eff $\leq 4, 2, 1.1$, we notice an initial linear correlation between N, M. However, for larger values of N, a decay in the value of M can be observed. This can be attributed to the qualitatively better approximation $\widehat{E}$ of E for larger values of N. For the same pairs of N, M we average the computation time for the calculation of the approximation u_N and the classical error bound, and the computation time for the calculation of u_N combined with the improved error bound Δ_{ub}. The computation time is then averaged over 200 random parameters. The relative computational overhead, i.e., the

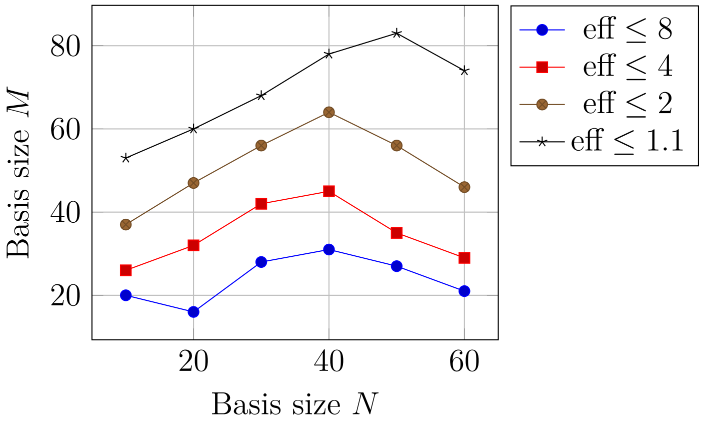

Figure 5.7: Basis size $M(N)$ required to achieve prescribed effectivities.

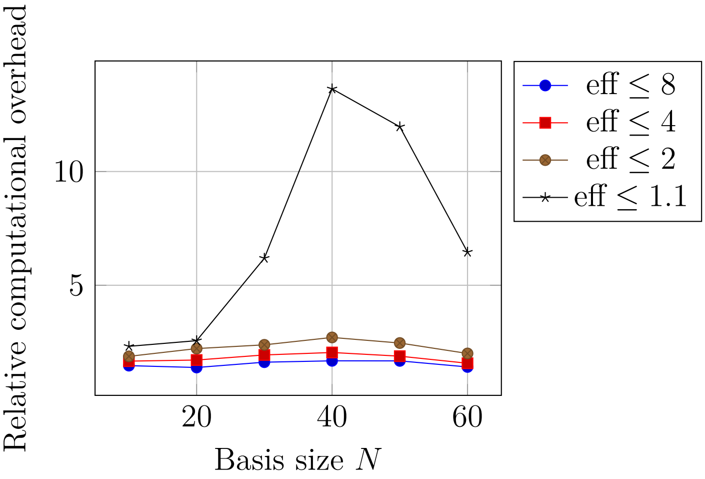

Figure 5.8: Relative overhead in the computation of the improved error bound compared to the standard RB bound.

quotient between the standard error quantification and the improved error quantification is plotted in Figure 5.8. In the cases eff $\leq 8, 4, 2$ we observe a relative overhead between 1.5 and 3 for all combinations of N and M. For the case eff ≤ 1.1, however, we observe a drastic increase in the relative overhead as both, N and M increases. This can be attributed to two reasons. First, as N increases the number of elements in the parameter-separable decomposition for the ALP increases, which makes the evaluation of the residual $R(\mu)$ more computationally expensive. Second, as both N and M increase, the majority of the computation time is spent on solving the reduced problem and the reduced ALP. Hence we can see a decline in the computational overhead after N, M cross a certain threshold.

5.6.2 Nonlinear finite-dimensional parametric problem

As a second example we consider the infinite-dimensional problem described by a nonlinear reaction-diffusion-advection equation in a one-dimensional domain $\Omega := (0, 1)$. The PDE is given by

$$
\begin{aligned}
-\mu_1 \partial_{\xi\xi} u(\xi;\mu) + \partial_\xi u(\xi;\mu) - \mu_2 u(\xi;\mu)^2 &= f(\xi), \quad \xi \in \Omega. \\
u(0;\mu) = u(1;\mu) &= 0,
\end{aligned}
$$

with parameters $\mu = (\mu_1, \mu_2)^T \in \mathcal{P} := [0.1, 1] \times [1, 10]$. Here μ_1 controls the diffusivity of the problem, whereas μ_2 changes the influence of the nonlinearity. The source term is given by $f(\xi) = \sin(\xi\pi)^2$ for $\xi \in \Omega$. To arrive at a finite-dimensional problem, we discretize the above PDE in space with a simple finite-difference scheme with upwind flux. This results in a $d = 400$ dimensional nonlinear problem of the form $G(x;\mu) = 0$ with $G(\cdot;\mu) : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$
G(x;\mu) := A(\mu_1)x + \mu_2 g(x) - f,
$$

with $A(\mu_1) \in \mathbb{R}^{d\times d}$, $g : \mathbb{R}^d \to \mathbb{R}^d$ and $f \in \mathbb{R}^d$. In this case, the Banach spaces are given by $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ which we equip with the standard Euclidean norm. We construct a subspace $\mathcal{X}_N$ of dimension $N = 6$ by applying the POD algorithm (Algorithm 6) to a collection of snapshots computed for 50 randomly selected parameters in the parameter domain $\mathcal{P}$. Note that this does not yield very accurate results when computing the RB-approximation. Nonetheless, it is sufficient to show the benefit of the improved error bound theory presented in Section 5.2. In Figure 5.9, the solutions and RB-approximations for three different parameters are depicted.

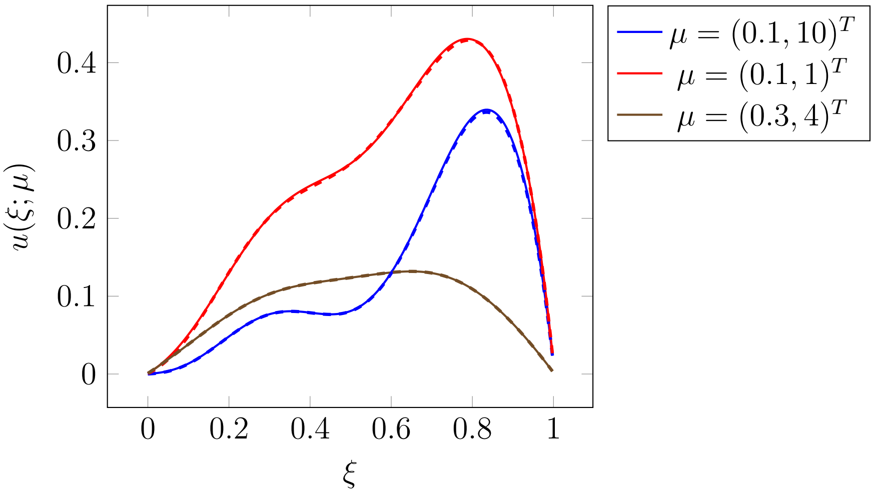

Figure 5.9: Example solutions for three different parameters.

	Error bound			Effectivity bound		
M	% valid	max	mean	% valid	max	mean
0	67	0.87	0.16			
5	67	0.83	0.15	0		
10	74	0.93	$9.24 \cdot 10^{-2}$	0		
15	82	0.27	$1.58 \cdot 10^{-2}$	59	2.93	2.23
20	83	0.59	$1.28 \cdot 10^{-2}$	100	1.86	1.02
25	83	0.59	$1.27 \cdot 10^{-2}$	100	1.84	1.01

Table 5.2: Percentage of parameters for which the error bound and effectivity bound are valid including their respective maximum and mean values.

For evaluating the error bound we have to calculate $\mathrm{D}G|_{\hat{x}(\mu)}$, which yields

$$\mathrm{D}G|_{\hat{x}(\mu)}(y) = A(\mu_1)y + 2\mu_2(\hat{x}(\mu) \circ y),$$

where $(a \circ b)_i := a_i b_i$ for $a, b \in \mathbb{R}^d$ denotes the component-wise product. From the explicit formula we immediately get $L(\alpha; \mu) \leq 2\mu_2\alpha =: L_{\mathrm{ub}}(\alpha)$. In all of the following computations the stability constant $\gamma(\mu) = \left\| \mathrm{D}G(\cdot; \mu)|_{\hat{x}}^{-1} \right\|$ is calculated exactly by solving a high-dimensional eigenvalue problem.

For the application of the error bound, we need to construct a RB space $\mathcal{X}_M^E$ for the ALP. For this, we compute snapshots of the ALP for 100 random parameters and subsequently use the POD algorithm.

We recall, that for the error bound to hold, a validity criterion of the form

$$\tau(\mu) = 2\gamma(\mu) L_{\mathrm{ub}}(\epsilon_{\mathrm{ub}}(\mu)) \leq 1$$

has to be satisfied. Hence, we will investigate the influence of the improved error estimation on the validity criterion. For this purpose, we evaluate the validity criterion for 200 random parameters and for varying dimensions $M \in \{0, 5, 10, 15, 20, 25\}$ of the reduced space $\mathcal{X}_M^E$. The choice $M = 0$ reflects the standard RB choice of $\epsilon_{\mathrm{ub}}(\mu) = \epsilon_{\mathrm{split}}(\mu)$, i.e. the standard RB bound. The percentage of parameters for which the validity criterion holds, as well as the maximum and mean value of $\tau(\mu)$ on these admissible parameters, is shown in Table 5.2. Furthermore, the combination of Corollary 5.3.10 and Lemma 5.4.3 provides us with a bound on the effectivity, if a second validity criterion (see equation (5.21)) is met. The percentage of parameters for which, on top of the error bound, this additional bound on the effectivity holds are displayed alongside the mean and maximum effectivity bound in Table 5.2

As we can see, the percentage of parameters for which the error bound is valid increases with the dimension of the RB space $\mathcal{X}_M^E$ as we generate tighter bounds on $\epsilon(\mu)$ which in turn leads to smaller values of $L(\epsilon_{\mathrm{ub}})$ and subsequently to smaller values of $\tau(\mu)$ as seen in the columns reflecting the mean and max value of $\tau(\mu)$. In total we were able to increase the number of admissible parameters by 16%. However, we are not able to reach 100% for all parameters as the space $\mathcal{X}_N$ is too coarse, thus we will eventually reach a bottleneck regardless of the dimension of $\mathcal{X}_M^E$. For the effectivity bound, we can see that the percentage of parameters for which the bound holds, increases with the dimension of $\mathcal{X}_M^E$ as well. Here the percentage does not represent the percentage of all 200 random parameters but only the percentage for which the respective error bound already holds. Similar to what we have seen in Section 5.6.1, we reach excellent effectivities for increasingly richer approximation spaces.

5.7 ALP Estimator for LTI-systems

In this section we apply the methodology developed in Section 5.2 to linear time invariant (LTI) systems. To this end let $I = [0, T] \subset \mathbb{R}$, $T > 0$ denote a time interval, let $A \in \mathbb{R}^{n \times n}$ be a system matrix and $b : I \to \mathbb{R}^n$ be an input. We now consider the following initial value problem (IVP)

$$x'(t) = Ax(t) + b(t), \qquad x(0) = x_0 = 0 \in \mathbb{R}^n \tag{5.33}$$

which can be reformulated as the integral equation (IE)

$$x(t) = \int_0^t Ax(s) + b(s)\mathrm{d}s. \tag{5.34}$$

Note that initial values $x_0 \neq 0$ are possible, however, we can transform any LTI-system into the above form simply by setting replacing $x(t)$ with $x(t) - x_0$. Furthermore, both the IVP and IE share the same unique solution.

In order to apply the method developed in Section 5.2, we first have to reformulate the IVP (5.33) or the IE (5.34) to a zero value problem

$$G(x) = 0$$

for a suitable operator $G : \mathcal{X} \to \mathcal{Y}$ and suitable Banach spaces $\mathcal{X}$, $\mathcal{Y}$. Since $x(t)$ needs to be continuously differentiable for (5.33) to be fulfilled, the choice

$$\mathcal{X}_1 \coloneqq C^1(I, \mathbb{R}^n) = \{f : I \to \mathbb{R}^n \mid f \text{ is continuosly differentiable and } f(0) = 0\}$$

paired with the norm

$$\|x\|_{\mathcal{X}_1} := \sup_{t\in I} \|x(t)\|_2 + \sup_{t\in I} \|x'(t)\|_2$$

seems to be a reasonable choice. Note that $\mathcal{X}_1$ paired with the above norm is in fact a Banach space. On the other hand, we can drop the assumption of differentiability for the IE (5.34) and therefore the choice

$$\mathcal{X}_2 := C(I, \mathbb{R}^n) = \{f : I \to \mathbb{R}^n \mid f \text{ is continuous}\}$$

paired with the norm

$$\|x\|_{\mathcal{X}_2} := \sup_{t\in I} \|x(t)\|_2$$

provides us with a suitable Banach-space. In both cases, $\|\cdot\|_2$ refers to the Euclidean-norm on $\mathbb{R}^n$. Based on the IVP and IE, we select the functions $G_1 : \mathcal{X}_1 \to \mathcal{Y}_1$ and $G_2 : \mathcal{X}_2 \to \mathcal{Y}_2$ defined by

$$G_1(x)(t) = x'(t) - Ax(t) - b(t) \in \mathcal{Y}_1 := C(I, \mathbb{R}^n).$$

and

$$G_2(x)(t) = x(t) - \int_0^t Ax(s) + b(s)\mathrm{d}s \in \mathcal{Y}_2 := \mathcal{X}_2 = C(I, \mathbb{R}^n).$$

In other words $\mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{X}_2$ and all spaces are equipped with the same norm $\|\cdot\|_{\mathcal{X}_2}$.

As mentioned before, any solution x^* of $G_1(x) = 0$ is a solution of $G_2(x) = 0$ and vice versa, provided b is a continuous function. However, $\mathcal{X}_1$ provides us with a stronger norm in the sense that for a given approximation $\hat{x}$ any bound on the error $e = \hat{x} - x^*$

$$\|e\|_{\mathcal{X}_1} \leq \Delta(\hat{x})$$

gives us information on both, the difference in the values as well as the derivatives. This is not the case for $\mathcal{X}_2$. Of course, this also requires that any approximation procedure that generates a suitable $\hat{x}$ has to guarantee that this is both, differentiable and satisfies $\hat{x}(0) = x_0$. In this sense, both problem settings described by G_1 and G_2 have their individual benefits and drawbacks In order to apply the error bound given in Theorem 5.3.1, we first need to show that both, $\mathrm{D}G_1|_{\hat{x}}$ and $\mathrm{D}G_2|_{\hat{x}}$, are invertible.

Lemma 5.7.1 (Regularity of $\mathrm{D}G_1$ and $\mathrm{D}G_2$)**.** *For any approximation $\hat{x}_1 \in \mathcal{X}_1$ and $\hat{x}_2 \in$*

$\mathcal{X}_2$, the Fréchet derivatives $\mathrm{D}G_1|_{\hat{x}_1}$ and $\mathrm{D}G_2|_{\hat{x}_2}$ are invertible and the respective inverse is given by

$$\mathrm{D}G_1|_{\hat{x}_1}^{-1}(y) = \int_0^t e^{A(t-s)} y(s)\mathrm{d}s \tag{5.35}$$

and

$$\mathrm{D}G_2|_{\hat{x}_2}^{-1}(y) = y(t) + \int_0^t e^{A(t-s)} Ay(s)\mathrm{d}s \tag{5.36}$$

where e^A denotes the matrix-exponential.

Proof. Since both G_1 and G_2 are (affine) linear, their derivatives are given by

$$\mathrm{D}G_1|_{\hat{x}_1}(x)(t) = x'(t) - Ax(t)$$

and

$$\mathrm{D}G_2|_{\hat{x}_2}(x)(t) = x(t) - \int_0^t Ax(s)\mathrm{d}s.$$

As mentioned before, the IVP and IE are equivalent formulations of the same problem. The same holds true for the equations for the problems $\mathrm{D}G_1|_{\hat{x}_1}(x) = 0$ and $\mathrm{D}G_2|_{\hat{x}_2}(x) = 0$. In particular, both share the same unique solution $x = 0$. Therefore, both derivatives are injective. For the surjectivity, we get by direct computation for any $y \in \mathcal{Y}_1, \mathcal{Y}_2$, respectively

$$\begin{aligned}\mathrm{D}G_1|_{\hat{x}}\left(\int_0^t e^{A(t-s)} y(s)\mathrm{d}s\right) &= A\int_0^t e^{A(t-s)} y(s)\mathrm{d}s + y(t) - A\left(\int_0^t e^{A(t-s)} y(s)\mathrm{d}s\right)\\ &= y(t)\end{aligned}$$

for $\mathrm{D}G_1$ and

$$\begin{aligned}\mathrm{D}G_2|_{\hat{x}}\left(y(t) + \int_0^t e^{A(t-s)} Ay(s)\mathrm{d}s\right) &= y(t) + \int_0^t e^{A(t-s)} Ay(s)\mathrm{d}s\\ &\quad - \int_0^t A\left(\int_0^u e^{A(s-u)} Ay(u)\mathrm{d}u\right)\mathrm{d}s\\ &= y(t)\end{aligned}$$

since

$$\int_0^t A\left(\int_0^u e^{(s-u)}Ay(u)\mathrm{d}u\right)\mathrm{d}s = \int_0^t\int_s^t Ae^{Au}\mathrm{d}uAy(s)\mathrm{d}s = \int_0^t e^{A(t-s)}Ay(s)\mathrm{d}s.$$

□

The functions G_1 and G_2 are both (affine) linear which is why the (modified) validity criterion of Theorem 5.3.1 is always satisfied. We still require (bounds on) the stability constants if we want to apply the improved and computable error bounds of Theorem 5.3.9, where the upper bound on the (non-split) residual is computed according to Lemma 5.4.2. For this purpose, we make use of the logarithmic norm ν_A (cf. [94, 44]) of the matrix A which satisfies

$$\left\|e^{At}\right\|_2 \leq e^{\nu_A t} \tag{5.37}$$

and which, in case of the spectral norm, can be computed by

$$\nu_A = \lambda_{\max}\left(\frac{A+A^T}{2}\right).$$

Lemma 5.7.2 (Bound on the stability constants for G_1 and G_2)**.** *The stability constant of G_1 and G_2 can be bounded as follows*

$$\begin{aligned}\gamma_1(\hat{x}_1) = \left\|\mathrm{D}G_1|_{\hat{x}_1}^{-1}\right\|_{\mathcal{L}(\mathcal{Y}_1,\mathcal{X}_1)} &\leq 1 + \sup_{t\in I}\int_0^t \|e^{A(t-s)}\|_2\mathrm{d}s \\ &\leq \gamma_{1,\mathrm{ub}}(\hat{x}) \coloneqq 1 + \frac{e^{\nu_A T}-1}{\nu_A}\end{aligned}$$

and

$$\begin{aligned}\gamma_2(\hat{x}_2) = \left\|\mathrm{D}G_1|_{\hat{x}_2}^{-1}\right\|_{\mathcal{L}(\mathcal{Y}_2,\mathcal{X}_2)} &\leq 1 + \sup_{t\in I}\int_0^t \|Ae^{A(t-s)}\|_2\mathrm{d}s \\ &\leq \gamma_{2,\mathrm{ub}}(\hat{x}) \coloneqq 1 + \|A\|_2\frac{e^{\nu_A T}-1}{\nu_A}.\end{aligned}$$

Proof. Using (5.35) and (5.37) we get

$$\begin{aligned}\left\| \mathrm{D}G_1|_{\hat{x}_1}^{-1}(y)\right\|_{\mathcal{X}_1} &= \sup_{t\in I}\left\|\int_0^t e^{A(t-s)}y(s)\mathrm{d}s\right\|_2 + \sup_{t\in I}\left\|\frac{\mathrm{d}}{\mathrm{d}t}\int_0^t e^{A(t-s)}y(s)\mathrm{d}s\right\|_2\\ &\leq \sup_{t\in I}\int_0^t \left\|e^{A(t-s)}\mathrm{d}s\right\|_2 \|y\|_{\mathcal{Y}_1} + \|y\|_{\mathcal{Y}_1} \leq \sup_{t\in I}\int_0^t e^{\nu_A(t-s)}\mathrm{d}s\,\|y\|_{\mathcal{Y}_1} + \|y\|_{\mathcal{Y}_1}\\ &= \|y\|_{\mathcal{Y}_1}\left(1+\frac{e^{\nu_A T}-1}{\nu_A}\right)\end{aligned}$$

Using (5.36) we get

$$\begin{aligned}\left\| \mathrm{D}G_2|_{\hat{x}_2}^{-1}(y)\right\|_{\mathcal{X}_2} &= \sup_{t\in I}\left\|y(t)+\int_0^t e^{A(t-s)}Ay(s)\mathrm{d}s\right\|_2 \leq \|y\|_{\mathcal{Y}_2} + \sup_{t\in I}\int_0^t \left\|e^{A(t-s)}A\right\|_2 \mathrm{d}s\,\|y\|_{\mathcal{Y}_2}\\ &\leq \|y\|_{\mathcal{Y}_2}\left(1+\|A\|_2\sup_{t\in I}\int_0^t e^{\nu_A(t-s)}\mathrm{d}s\right) = \|y\|_{\mathcal{Y}_2}\left(1+\|A\|_2\frac{e^{\nu_A T}-1}{\nu_A}\right)\end{aligned}$$

□

Depending on the system matrix A, the above bounds might be rather pessimistic. In particular, the second bound for $\gamma_2(\hat{x}_2)$ can have a devastating effect on the quality of the error bound as it scales linearly with the largest singular value of A. Furthermore, they are still expensive to evaluate as the computation of the logarithmic norm requires the computation of the largest eigenvalue of the symmetric part of A.

In practice, we expect that $\gamma_{1,\mathrm{ub}}(\hat{x}) \leq \gamma_{2,\mathrm{ub}}(\hat{x})$, and infer in light of Lemma 5.4.3, tighter error bounds for the problem setting G_1. We, however, stress again that this direct comparison might be unsuited, as the norm $\|\cdot\|_{\mathcal{X}_2}$ is weaker than the norm $\|\cdot\|_{\mathcal{X}_1}$. In specific, the approach using G_1 provides us with a bound on the sum of the error and its derivative, and hence we can not infer any bound on just the error $e = \hat{x} - x^*$, unless we compute the (potentially worse) bound provided by using G_2.

This is due to the fact that the setting of our original problem (P) is quite restrictive as we require the problem spaces $\mathcal{X}$ and $\mathcal{Y}$ to be Banach spaces. However, in view of Theorem 5.3.1 this assumption was only necessary to guarantee the existence of a solution $G(x^*) = 0$ in a neighbourhood of the approximation $\hat{x}$. In case of the IVP (5.33) or the equivalent IE (5.34), the existence of a (unique) solution x^* is already guaranteed by the Picard-Lindelöf-Theorem and consequently, we can achieve an error bound derived from the same principle as using an ALP. For this purpose, let $\hat{x}$ be such that $\hat{x}(0) = x_0$, then

$\hat{x}$ itself solves the IVP

$$\hat{x}'(t) = A\hat{x}(t) + b(t) + r(\hat{x})(t), \qquad \hat{x}(0) = x_0$$

where the residual is given by

$$r(\hat{x})(t) = \hat{x}'(t) - A\hat{x}(t) - b(t) \quad (= G_1(\hat{x})(t)).$$

The error $e = \hat{x} - x^*$ now satisfies the IVP

$$e'(t) = Ae(t) + r(\hat{x})(t), \qquad e(0) = 0. \tag{5.38}$$

Which has the unique solution

$$e(t) = \int_0^t e^{A(t-s)} r(\hat{x})(s) \mathrm{d}s.$$

The above is precisely the ALP we derived in the case of G_1. However, for this case we now applied the norm of e in the space $\mathcal{X}_1$. However, using the above representation we immediately get the following bound in the weaker norm on $\mathcal{X}_2$:

$$\|e(t)\|_{\mathcal{X}_2} \leq \left(\sup_{t \in I} \int_0^t \left\| e^{A(t-s)} \right\|_2 \mathrm{d}s \right) \|r(\hat{x})\|_{\mathcal{X}_2}$$

This bound now corresponds to the bound, where we used the split-residual, and can therefore be improved by yet again applying the key principle of the ALP framework to the IVP (5.38). To see this, let $\hat{e}$ be an approximation to (5.38) and let $r_e(\hat{e})$ denote the corresponding residual given by

$$r_e(\hat{e})(t) = e'(t) - Ae(t) - r(\hat{x})(t).$$

Then the error in the $\mathcal{X}_2$–norm is bounded by

$$\|e\|_{\mathcal{X}_2} \leq \|\hat{e}\|_{\mathcal{X}_2} + \left(\sup_{t \in I} \int_0^t \left\| e^{A(t-s)} \right\|_2 \mathrm{d}s \right) \|r_e(\hat{e})\|_{\mathcal{X}_2}.$$

Once again, the above bound should be sharper, since we expect $\|r_e(\hat{e})\|_{\mathcal{X}_2} \ll \|r(\hat{x})\|_{\mathcal{X}_2}$.

Please note that the process described above is analogous to the methodology we derived in Section 5.2. The only difference is that we already assume the solvability of the given problem and that the spaces $\mathcal{X}$ and $\mathcal{Y}$ are no longer Banach spaces but only normed spaces. Furthermore, many of the results of Section 5.2 have an analogue in the above setting. We will summarize this in the following theorem

Theorem 5.7.3 (Error Bound using a generalized ALP)**.** *Let $\mathcal{X}$, $\mathcal{Y}$ be two normed spaces and $G : \mathcal{X} \to \mathcal{Y}$ a (problem) operator, such that the problem*

$$\text{Find } x \in \mathcal{X} \text{ such that } G(x) = 0 \in \mathcal{Y}$$

has a (unique) solution $x^* \in \mathcal{X}$. Furthermore, let $\mathcal{A} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ be an invertible linear operator such that the error $e = x^* - \hat{x}$ for a given approximation $\hat{x} \in \mathcal{X}$ is bounded by

$$\|e\|_{\mathcal{X}} \leq \bar{\Delta}(\hat{x}) =: C(\hat{x}) \left\|\mathcal{A}^{-1}(r(\hat{x}))\right\|_{\mathcal{X}} \leq C(\hat{x}) \left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \|r(\hat{x})\|_{\mathcal{Y}} \tag{5.39}$$

where $C(\hat{x}) > 0$ is a positive constant and $r(\hat{x})$ is a residual depending on the approximation $\hat{x}$. If $\hat{E} \in \mathcal{X}$ is an approximation for the solution of the generalized ALP

$$\mathcal{A}(E) = r(\hat{x}),$$

with residual

$$R(\hat{E}) = \mathcal{A}(\hat{E}) - r(\hat{x})$$

such that

$$2\left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \left\|R(\hat{E})\right\|_{\mathcal{Y}} \leq \left\|\hat{E}\right\|_{\mathcal{X}},$$

then the upper bound given by

$$\bar{\Delta}_{\mathrm{ub}}(\hat{x}) = \left\|\hat{E}\right\|_{\mathcal{X}} + \left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \left\|R(\hat{E})\right\|_{\mathcal{Y}}$$

satisfies

$$\bar{\Delta}(\hat{x}) \leq \bar{\Delta}_{\mathrm{ub}}(\hat{x}) \leq 3\bar{\Delta}(\hat{x}).$$

Proof. The proof is analogous to the proof of Lemma 5.4.3, where the upper bound on the stability constant is replaced by the norm $\left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})}$. □

Remark 5.7.4. **(a)** We infer from Theorem 5.7.3 that the effectivity of the error bound $\bar{\Delta}_{\mathrm{ub}}(\hat{x})$ is at most worsened by a factor 3, if the approximation for the generalized ALP is of sufficient quality.

(b) Similar to the results in Section 5.2, we can replace $\left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})}$ by a suitable upper bound.

(c) In the context of Section 5.2, we have $\mathcal{A} = \mathrm{D}G|_{\hat{x}}$ and $r(\hat{x}) = G(\hat{x})$ and $\left\|\mathcal{A}^{-1}\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \leq \gamma_{1,\mathrm{ub}}(\hat{x}) - 1$. The above, however, shows that the principle behind the ALP is applicable and beneficial whenever one has a bound of the form (5.39) where the splitting step is executed.

(d) One can easily derive an extension of the above, if the bound takes the form

$$\bar{\Delta}_{\mathrm{ub}}(\hat{x}) = \sum_{i=1}^{m} C_i(\hat{x}) \left\|\mathcal{A}_i^{-1}(r_i(\hat{x}))\right\|_{\mathcal{X}_i},$$

i.e. the bound is in terms of multiple different linear operators $\mathcal{A}_i$ and residuals $r_i(\hat{x}) \in \mathcal{Y}_i$, where $\mathcal{X}_i \subset \mathcal{X}$ and $\mathcal{Y}_i \subset \mathcal{Y}$ are subspaces with (potentially different) norms $\|\cdot\|_{\mathcal{X}_i}$ and $\|\cdot\|_{\mathcal{Y}_i}$.

5.7.1 RB approximations for LTI-systems

In the subsequent numerical experiment, we again want to make use of the RB framework to derive suitable approximations to both, the problem, in its IVP or IE form, and their corresponding (generalized) ALP. For this purpose, we roughly cover the essential elements required for the approximation procedure, we will be using later. For a more general overview of the different techniques developed for time-dependent problems, we refer to [36, 42, 41].

For a parameter $\mu \in \mathcal{P}$ we consider the parametric IVP

$$x'(t;\mu) = A(\mu)x(t;\mu) + b(t;\mu), \qquad x(0;\mu) = x_0 \tag{5.40}$$

and its equivalent parametric IE

$$x(t;\mu) = x_0 + \int_0^t e^{A(\mu)(t-s)} b(s;\mu)\mathrm{d}s. \tag{5.41}$$

for $t \in I := [0,T]$, $A(\mu) \in \mathbb{R}^{n\times n}$, $b(t;\mu), x_0 \in \mathbb{R}^n$. Furthermore, we shift the problem into a time-discrete setting, i.e. we discretize the time-interval I into $K+1$ points and the solution is then represented by its discrete trajectory $x^* = \{x^*(k\cdot\Delta t)\}_{k=0}^{K}$, where $\Delta t = \frac{T}{K}$. In this case, the problem space can be represented by $\mathcal{X} = \mathbb{R}^{n\times(K+1)}$. Similar to the basis generation techniques presented in Section 5.5.5, we construct a reduced basis by applying the POD-algorithm (Algorithm 6) to a collection snapshots, i.e. a collection of discrete trajectories for different parameters. This results in a reduced basis represented by $V_{\mathcal{X}_N} \in \mathbb{R}^{n \times N}$ consisting of the orthogonal columns. The correpsonding reduced basis space is given by all discrete trajectories, whose elements can be represented as linear combinations of the columns of $V_{\mathcal{X}_N}$. In other words $\mathcal{X}_N = \operatorname{colspan}(V_{\mathcal{X}_N})$. If $x_0 \notin \operatorname{range}(V_{\mathcal{X}_N})$, we extend the basis accordingly. An approximation $\hat{x} = V_{\mathcal{X}_N} x_N$ is then computed by solving the reduced IVP

$$x_N(t;\mu) = V_{\mathcal{X}_N}^T A(\mu) V_{\mathcal{X}_N} x_N(t;\mu) + V_{\mathcal{X}_N}^T b(t;\mu), \quad x_N(0,\mu) = V_{\mathcal{X}_N}^T x_0$$

with the same (discrete-time) integrator which was used to obtain the snapshots for the basis construction. In an analogous fashion, we construct the reduced space for the ALP

$$e'(t;\mu) = A(\mu)e(t;\mu) + r(\hat{x}(\mu))(t), \qquad e(0) = 0.$$

Because of the equivalence of the parametric IVP (5.40) and the parametric IE (5.41), we can use the same construction method for both settings.

As was already mentioned in Section 5.2, if the system matrix $A(\mu)$ and the input $b(\cdot;\mu)$ are parameter-separable this is inherited by all occurring residuals and therefore efficient computations of the RB approximations and evaluation of the error bounds is possible.

5.7.2 Numerical Example

As an example we consider the parameter-dependent LTI system with parameter values $\mu \in \mathcal{P} = [0.01, 0.1]$, which is obtained by semi-discretizing the PDE

$$\begin{aligned} \partial_t u(\xi,t;\mu) &= \mu \partial_{\xi\xi} u(\xi,t;\mu) - \partial_\xi u(\xi,t;\mu) + \cos(2\pi\xi)t, \quad (\xi,t) \in I \times \Omega := (0,1) \times (0,1). \\ u(\xi,0;\mu) &= \sin(2\pi\xi)^2. \end{aligned}$$

with central finite differences for $\partial_{\xi\xi}$ and forward finite differences for ∂_ξ, where we discretize the spacial domain $\Omega = (0,1)$ with 100 equidistant points. The LTI system then takes the form

$$x'(t;\mu) = A(\mu)x(t,\mu) + b(t), \quad x(0;\mu) = x_0$$

with $A(\mu) = \mu A_1 + A_2 \in \mathbb{R}^{100\times 100}$ and $b(t), x_0 \in \mathbb{R}^{100}$, where the matrices $A_1, A_2 \in \mathbb{R}^{100\times 100}$ correspond to the discrete partial derivatives $\partial_{\xi\xi}$ and ∂_ξ, respectively. For the computation of the discrete-time trajectories, we employ the backward Euler method with a time-step size $\Delta t = \frac{1}{99}$, i.e. the trajectories consist of $K + 1 = 100$ individual elements. We construct a reduced basis space $\mathcal{X}_N$ for building the approximant by applying the POD algorithm to the trajectories of the discrete solution $x^*(\mu_0)$ for the parameter $\mu_0 = 0.04$.

The reduced basis is then represented by the first 19 left singular vectors of the matrix containing the trajectories which we further enrich via x_0 resulting in a 20-dimensional space, i.e. $V_{\mathcal{X}_N} \in R^{100\times 20}$. Similarly, we construct an approximation space for the ALP (in differential form)

$$e'(t;\mu) = A(\mu)e(t;\mu), \qquad e(0;\mu) = 0$$

by computing the trajectories for the parameters $\mu \in \{0.01, 0.025, 0.05, 0.075, 0.1\}$ and once again running the POD algorithm. We construct 4 different reduced basis spaces with this approach which are represented by the first 25, 30, 35 and 40 left singular vectors. We denote the corresponding reduced basis via $V_{\mathcal{X}^E_{25}}$, $V_{\mathcal{X}^E_{30}}$, $V_{\mathcal{X}^E_{35}}$ and $V_{\mathcal{X}^E_{40}}$, respectively. Furthermore, we denote by $\Delta^M_{1,\text{ub}}(\mu)$, $\Delta^M_{2,\text{ub}}(\mu)$ and $\bar{\Delta}^M_{\text{ub}}(\mu)$ the error bounds when using the the problem setting described by G_1, G_2 of Section 5.7, the generalized setting of Theorem 5.7.3 and when using the reduced basis $V_{\mathcal{X}^E_M}$ for the ALP. We recall that in this case, the error measured by $\Delta^M_{1,\text{ub}}$ entails the the maximum value of the pointwise error $e(t;\mu) = \hat{x}(t;\mu) - x^*(t;\mu)$ and the maximum value of its derivative $e'(t;\mu)$, whereas $\Delta^M_{2,\text{ub}}$ and $\bar{\Delta}^M_{\text{ub}}$ only measure the maximum pointwise error. Upper bounds for the stability constants $\gamma_{1,\text{ub}}(\mu)$ and $\gamma_{2,\text{ub}}(\mu)$ are computed following the results of Lemma 5.7.2 and the fact, that for the setting of $\bar{\Delta}_{\text{ub}}(\mu)$, the corresponding stability constant is given by

$$\bar{\gamma}(\mu) = \left\| \mathrm{D}G_1|^{-1}_{\hat{x}(\mu)} \right\|_{\mathcal{L}(\mathcal{Y}_2,\mathcal{X}_2)}$$

and therefore, it can, similar to the stability constant of G_1, be bounded by

$$\bar{\gamma}(\mu) \leq \bar{\gamma}_{\text{ub}}(\mu) := \gamma_{1,\text{ub}}(\mu) - 1.$$

Furthermore, we shall denote as $\text{eff}^M_1(\mu)$, $\text{eff}^M_2(\mu)$ and $\overline{\text{eff}}^M(\mu)$ the effectivities of the above bounds and as $\text{eff}^M_{1,\text{ub}}(\mu)$, $\text{eff}^M_{2,\text{ub}}(\mu)$ and $\overline{\text{eff}}^M_{\text{ub}}(\mu)$ the bounds on the effecitivites according to Corollary 5.3.10 combined with Lemma 5.4.3 and Theorem 5.7.3, respectively. The error and bounds measured in the $\mathcal{X}_1$–norm, i.e. the standard C^1–norm, for 20 random parameters in the parameter set $\mathcal{P}$ are displayed in Figure 5.10, while the error and bounds measured in the $\mathcal{X}_2$–norm , i.e. the standard C^0–norm, are shown in Figure 5.11. Similar to what we have seen in Section 5.6, the quality of the error bounds improves as the dimension M of the reduced basis for the ALP increases. This is further showcased in Figure 5.12, where the effectivities and the bounds on the effectivites for the richest approximation space (dimension $M = 40$), are displayed. Here we can see, that the theoretical results already guarantee effectivities of less than 1.3 and the actual effectivities are closer to 1. Taking a closer look at Figure 5.11 we can see that the bounds $\bar{\Delta}^M_{\text{ub}}$

are significantly better than the bounds $\Delta^M_{2,\mathrm{ub}}$, as the former matches the error visually already for basis size $M = 30$, whereas the latter requires a larger space of dimension $M = 40$. This is due to the fact, that the bound on stability constant of G_2 scales linearly with the norm of the system matrix $\|A(\mu)\|_2$. Computing the mean value of all bounds $\gamma_{2,\mathrm{ub}}(\mu)$ on the stability constant for the 20 random parameters, we obtain approximately 1680, whereas the mean for $\bar{\gamma}_{\mathrm{ub}}$ turns out to be only 0.77. In other words, on average $\gamma_{2,\mathrm{ub}}(\mu)$ is approximately 2180 times larger than $\bar{\gamma}_{\mathrm{ub}}(\mu)$ for a given parameter. Nonetheless, we can achieve effectivities close to one, if M is sufficiently large, as is highlighted in Table 5.3. We once again want to emphasize that this shows that the error bounds using the methodology presented in this work can counteract large stability constants, which in the case of classical RB approximation would lead to severe overestimation of the error.

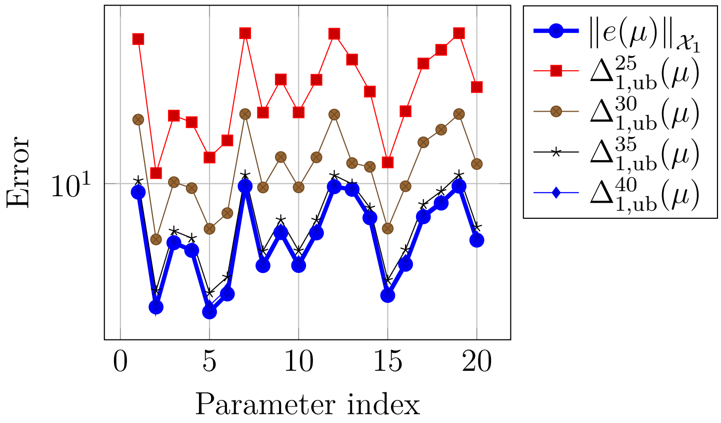

Figure 5.10: Error and error bounds for the C^1–norm for 20 random parameters.

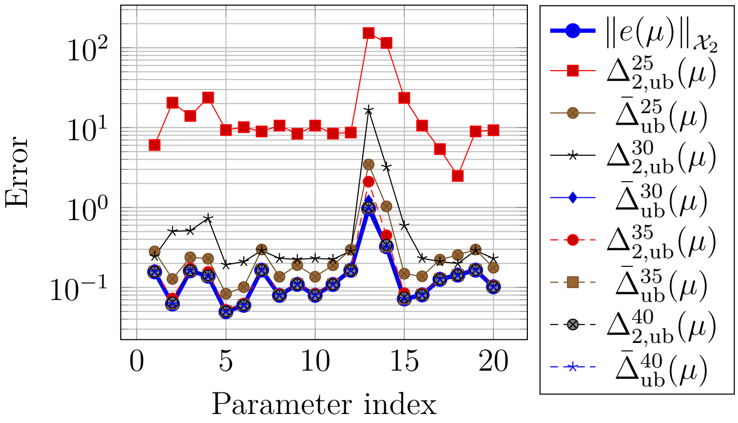

Figure 5.11: Error and error bounds for the C^0–norm for 20 random parameters.

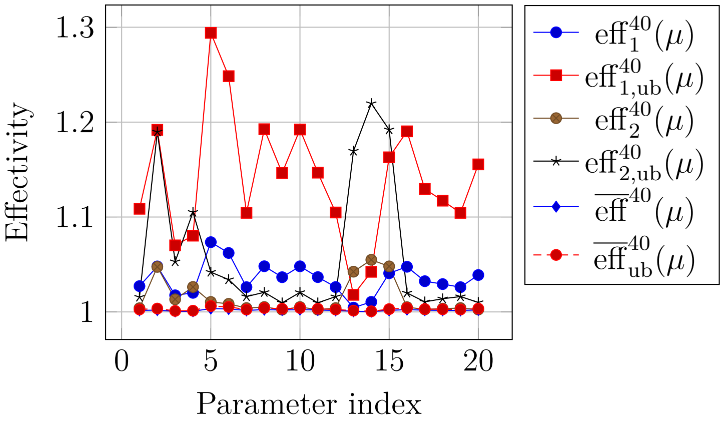

Figure 5.12: Effectivities and bound on the effectivities for the richest RB space $\mathcal{X}_{40}^{E}$ and for 20 random parameters.

	Mean effectivity			Maximum effectivity		
M	$\Delta_{1,\text{ub}}$	$\Delta_{2,\text{ub}}$	$\bar{\Delta}_{\text{ub}}$	$\Delta_{1,\text{ub}}$	$\Delta_{2,\text{ub}}$	$\bar{\Delta}_{\text{ub}}$
25	5.23	134.87	1.93	5.69	355.28	3.57
30	2.22	4.23	1.06	2.55	17.22	1.26
35	1.16	1.13	1.01	1.24	2.18	1.02
40	1.04	1.01	1	1.07	1.05	1

Table 5.3: Maximum and mean effectivity for the different error bounds and different dimension of the RB space for the ALP.

6 Conclusion

In the first part of this thesis, consisting of Chapters 2 through 4, we extended several known properties of scalar-valued kernels to the case of matrix-valued kernels. We introduced the new subclass of uncoupled separable kernels and were able to show that their specific structure can be exploited to efficiently evaluate the Power function associated with the kernel. Furthermore, we developed several variants of the P–Greedy algorithm all of which generate point sets that result in quasi-optimal approximation rates if RBF kernels whose native space can be identified with a Sobolev space, are used. This was done in a more generalized setting, in which the greedy space is enriched with a k-dimensional space in every iteration. This may be applied to other weak greedy scenarios not covered in this thesis. Moreover, we generalized the classical regularization ansatz for scalar-valued kernels, by replacing the more commonly used penalization parameter in front of the regularization functional with a positive definite weight function. While this allows for more flexibility in the approximation procedure, it also increases the number of parameters used and hence the selection of optimal parameters is made more difficult. Nonetheless, we were able to show that this increased flexibility can be exploited in order to improve the quality of the kernel approximation. In particular, we introduced a regularized greedy algorithm, which generates point sets that are better suited for the regularization approach and still maintain quasi-optimal rates.

Future work should investigate how to better train the internal parameters for the kernel approximations, as the number of parameters scale quadratically with the output dimension of the target function. We also want to further investigate the effect the matrix-valued weight function has on the point selection process during the regularized P–Greedy.

In the second part, consisting of Chapter 5, we presented a novel improvement of error bounding techniques for problems which can be expressed in terms of a root finding problem for some differentiable operator between two Banach spaces. We achieved this by introducing and solving an additional auxiliary linear problem (ALP) which counteracts the often severe overestimation that occurs when applying standard error bounding techniques. This resulted in the here presented ALP-based error bounds. These a-posteriori error bounds show significant improvements in their effectivity. In particular, we could show that if a certain validity criterion is met, the resulting error bound deviates from the actual error by at most a factor of 3. Furthermore, one can control the quality of the error prediction by tuning the quality of the approximation $\hat{E}$ of the ALP. We applied the technique in the context of RB methods, where we compared our new bounds to the existing standard error bounding methods. Numerical examples for both, linear and non-linear as well as (linear) time-dependent problems highlight the benefit of the presented technique and showcase that ALP-based error estimation enables us to reach excellent effectivities, i.e. effectivities close to 1.

In future work, we would like to extend the presented bounds to non-linear time dependent problems. We also want to study the quality of these bounds when approximation techniques, other than the reduced basis method, are considered.

Bibliography

[1] R. Adams and J. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics.* Elsevier, 2003.

[2] M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis.* Wiley, 2000.

[3] M. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[5] G. Santin B. Haasdonk, B. Hamzi and D. Wittwar. Greedy kernel methods for center manifold approximation. In J.Peiro P. E. Vincent S. J. Sherwin, D. Moxey and C. Schwab, editors, *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2018*, pages 95–106. Springer International Publishing, 2020.

[6] R. K. Beatson, W. zu Castell, and S. J. Schrödl. Kernel-based methods for vector-valued data with correlated components. *SIAM Journal on Scientific Computing*, 33(4):1975–1995, 2011.

[7] P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. H. A. Schilders, and L. M. Silveira, editors. *Model Order Reduction. Volume 1: System- and Data-Driven Methods and Algorithms.* De Gruyter, Berlin, 2020.

[8] P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. H. A. Schilders, and L. M. Silveira, editors. *Model Order Reduction. Volume 2: Snapshot-Based Methods and Algorithms.* De Gruyter, Berlin, 2020.

[9] P. Benner, S. Grivet-Talocia, A. Quarteroni, G. Rozza, W. H. A. Schilders, and L. M. Silveira, editors. *Model Order Reduction. Volume 3: Applications.* De Gruyter, Berlin, 2020.

[10] P. Benner, S. Gugercin, and K. Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Rev.*, 57(4):483–531, jan 2015.

[11] P. Benner, M. Ohlberger, A. Cohen, and K. Willcox. *Model Reduction and Approximation.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

[12] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.*, 43(3):1457–1472, 2011.

[13] C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 2. Springer, August 2006. ISBN 978-0-387-31073-2.

[14] L. Bos, M. Caliari, S. De Marchi, M. Vianello, and Y. Xu. Bivariate Lagrange interpolation at the Padua points: The generating curve approach. *Journal of Approximation Theory*, 143(1):15–25, 2006. Special Issue on Foundations of Computational Mathematics.

[15] L. Bos, S. De Marchi, A. Sommariva, and M. Vianello. Computing multivariate Fekete and Leja points by numerical linear algebra. *SIAM Journal on Numerical Analysis*, 48(5):1984–1999, 2010.

[16] D. Braess. *Finite Elements.* Cambridge University Press, 2007.

[17] G. Caloz and J. Rappaz. *Handbook of Numerical Analysis*, volume 5, chapter Numerical analysis for nonlinear and bifurcation problems, pages 487–637. 1997.

[18] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, August 2008.

[19] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Anal. Appl. (Singap.)*, 4(4):377–408, 2006.

[20] S. Chaturantabut and D. Sorensen. Nonlinear Model Reduction via Discrete Empirical Interpolation. *SIAM J. Sci. Comput.*, 32(5):2737–2764, 2010.

[21] Cohen, Albert, Dahmen, Wolfgang, DeVore, Ronald, and Nichols, James. Reduced basis greedy selection using random training sets. *ESAIM: M2AN*, 54(5):1509–1524, 2020.

[22] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, May 1994.

[23] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, 2000.

[24] W. Dahmen, C. Plesken, and G. Welper. Double greedy algorithms: Reduced basis methods for transport dominated problems. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 48(3):623–663, 2014.

[25] S. De Marchi and R. Schaback. Stability of kernel-based interpolation. *Adv. Comput. Math.*, 32(2):155–161, 2010.

[26] S. De Marchi, R. Schaback, and H. Wendland. Near-optimal data-independent point locations for radial basis function interpolation. *Adv. Comput. Math.*, 23(3):317–330, 2005.

[27] R. DeVore, G. Petrova, and P. Wojtaszczyk. Greedy algorithms for reduced bases in Banach spaces. *Constr. Approx.*, 37(3):455–466, 2013.

[28] V. Dolejší and M. Feistauer. *Discontinuous Galerkin Method.* Springer International Publishing, 2015.

[29] D. Driess, S. Schmitt, and M. Toussaint. Active inverse model learning with error and reachable set estimates. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1826–1833, 2019.

[30] M. Drohmann, B. Haasdonk, and M. Ohlberger. Adaptive reduced basis methods for nonlinear convection-diffusion equations. In *In Proc. FVCA6*, 2011.

[31] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. *Handbook of Numerical Analysis*, 7:713–1018, 2000.

[32] P. Farrell, K. Gillow, and H. Wendland. Multilevel interpolation of divergence-free vector fields. *IMA Journal of Numerical Analysis*, 37(1):332–353, 2017.

[33] R. Franke. *A critical comparison of some methods for interpolation of scattered data*, volume 253. Naval Postgraduate School Tech.Rep., March 1979.

[34] E. J. Fuselier. *Refined error estimates for matrix-valued radial basis functions.* PhD thesis, Texas A&M University, 2006.

[35] E. J. Fuselier and G. B. Wright. Stability and error estimates for vector field interpolation and decomposition on the sphere with RBFs. *SIAM Journal on Numerical Analysis*, 47(5):3213–3239, 2009.

[36] S. Glas, A. Mayerhofer, and K. Urban. *Two Ways to Treat Time in Reduced Basis Methods*, pages 1–16. Springer International Publishing, Cham, 2017.

[37] M. Günther and L. Klotz. Schur's theorem for a block Hadamard product. *Linear Algebra and its Applications*, 437:948–956, 2012.

[38] B. Haasdonk. Convergence rates of the POD–Greedy method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(3):859–873, 2013.

[39] B. Haasdonk. Reduced basis methods for parametrized PDEs – a tutorial introduction for stationary and instationary problems. In P. Benner, A. Cohen, M. Ohlberger, and K. Willcox, editors, *Model Reduction and Approximation: Theory and Algorithms*, pages 65–136. SIAM, Philadelphia, 2017.

[40] B. Haasdonk, B. Hamzi, G. Santin, and D. Wittwar. Kernel methods for center manifold approximation and a weak data-based version of the center manifold theorem. *Physica D: Nonlinear Phenomena*, 427:133007, 2021.

[41] B. Haasdonk and M. Ohlberger. Reduced basis method for finite volume approximations of parametrized linear evolution equations. *ESAIM: M2AN*, 42(2):277–302, March 2008.

[42] B. Haasdonk and M. Ohlberger. Efficient reduced models and a posteriori error estimation for parametrized dynamical systems by offline/online decomposition. *Math. Comput. Model. Dyn. Syst.*, 17(2):145–161, 2011.

[43] S. Hain, M. Ohlberger, M. Radic, and K. Urban. A hierarchical a posteriori error estimator for the reduced basis method. *Advances in Computational Mathematics*, Feb 2019.

[44] E. Heinz. Beiträge zur Störungstheorie der Spektralzerlegung. *Mathematische Annalen*, 123(1):415–438, 1951.

[45] J. Hertz, John, A. Krough, and R. G. Palmer. *Introduction to the Theory of Neural Computation*, volume 44. 1991.

[46] J. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations.* SpringerBriefs in Mathematics. Springer, 2016.

[47] J. S. Hesthaven, B. Stamm, and S. Zhang. Efficient greedy algorithms for high-dimensional parameter spaces with applications to empirical interpolation and reduced basis methods. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 48(1):259–283, 2014.

[48] C. Himpe, T. Leibner, and S. Rave. Hierarchical approximate proper orthogonal decomposition. *SIAM Journal on Scientific Computing*, 40(5):A3267–A3292, 2018.

[49] D. B. P. Huynh, D. J. Knezevic, Y. Chen, J. S. Hesthaven, and A. T. Patera. A natural-norm successive constraint method for inf-sup lower bounds. *Computer Methods in Applied Mechanics and Engineering*, 199:1963–1975, 2010.

[50] D. B. P. Huynh, G. Rozza, S. Sen, and A. T. Patera. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. *Comptes Rendus de l'Académie des Sciences, Series I*, 345:473–478, 2007.

[51] R. Andrade Flauzino L. H. B. Liboni S. F. dos Reis Alves I. Nunes Silva, D. Hernane Spatti. *Artificial Neural Networks.* Springer International Publishing, 1st edition, 2017.

[52] C. Johnson. *Numerical solution of partial differential equations by the finite element method.* Courier Corporation, 1989.

[53] H. Kadri, E. Duflos, P. Preux., S. Canu, A. Rakotomamonjy, and J. Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20), 2016.

[54] M. Köppel, F. Franzelin, I. Kröker, S. Oladyshkin, G. Santin, D. Wittwar, A. Barth, B. Haasdonk, W. Nowak, D. Pflüger, and C. Rohde. Comparison of data-driven uncertainty quantification methods for a carbon dioxide storage benchmark scenario. *Computational Geosciences*, 23(2):339–354, Apr 2019.

[55] M. Köppel, I. Kröker, and C. Rohde. Intrusive uncertainty quantification for hyperbolic-elliptic systems governing two-phase flow in heterogeneous porous media. *Computational Geosciences*, 21:1–26, 08 2017.

[56] Randall J. LeVeque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.

[57] S. Lowitzsch. A density theorem for matrix-valued radial basis functions. *Numerical Algorithms*, 39(1):253–256, 2005.

[58] S. Lowitzsch. Error estimates for matrix-valued radial basis function interpolation. *Journal of Approximation Theory*, 137(2):238–249, 2005.

[59] S. Lowitzsch. Matrix-valued radial basis functions: stability estimates and applications. *Advances in Computational Mathematics*, 23(3):299–315, Oct 2005.

[60] Y. Maday, N. C. Nguyen, A. T. Patera, and S. H. Pau. A general multipurpose interpolation procedure: the magic points. *Communications on Pure & Applied Analysis*, 8(1):383–404, 2009.

[61] V. Mazja. *Sobolev Spaces.* Springer, 1985.

[62] C. A. Micchelli. Interpolation of scattered data: Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[63] C. A. Micchelli and M. Pontil. Kernels for multi-task learning. *Advances in Neural Information Processing Systems*, 2004.

[64] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Comput.*, 17(1):177–204, 2005.

[65] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, June 2006.

[66] M. Michelo and J. A. Glaunes. Matrix-valued kernels for shape deformation analysis. *Geometry, Imaging and Computing*, 1(1):57–139, 2014.

[67] Ha Quang Minh. Operator-valued Bochner theorem, Fourier feature maps for operator-valued kernels, and vector-valued learning. arXiv 1608.05639, 2016.

[68] S. Müller and R. Schaback. A Newton basis for kernel spaces. *J. Approx. Theory*, 161(2):645–655, 2009.

[69] F. J. Narcowich and J. D. Ward. Generalized Hermite interpolation via matrix-valued conditionally positive definite functions. *Math. Comp.*, 63(208):661–687, 1994.

[70] M. Ohlberger. *A posteriori error estimates and adaptive methods for convection dominated transport processes.* PhD thesis, Albert-Ludwigs-Universität Freiburg, 2001.

[71] S. Oladyshkin and W. Nowak. Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliability Engineering & System Safety*, 106:179–190, 2012.

[72] A.T. Patera and G. Rozza. *Reduced Basis Approximation and a Posteriori Error Estimation for Parametrized Partial Differential Equations.* To appear in (tentative) MIT Pappalardo Graduate Monographs in Mechanical Engineering. MIT, 2007.

[73] M. Pazouki, S. S. Allaei, M. H. Pazouki, and D. P. F. Möller. Adaptive learning algorithm for RBF Neural Networks in kernel spaces. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4811–4818, July 2016.

[74] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

[75] D. Pflüger, B. Peherstorfer, and H.-J. Bungartz. Spatially adaptive sparse grids for high-dimensional data-driven problems. *Journal of Complexity*, 26(5):508–522, 2010. SI: HDA 2009.

[76] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced basis methods for partial differential equations: an introduction*, volume 92. Springer, 2015.

[77] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, 2006.

[78] M. Reisert and H. Burkhardt. Learning equivariant functions with matrix valued kernels. *J. Mach. Learn. Res.*, 8:385–408, May 2007.

[79] C. Rieger. *Sampling Inequalities and Applications.* PhD thesis, Fakultät für Mathematik und Informatik, Georg-August-Universität Göttingen, 2008.

[80] C. Rieger and H. Wendland. Sampling inequalities for anisotropic tensor product grids. *IMA Journal of Numerical Analysis*, 40(1):285–321, 01 2019.

[81] C. Rieger and B. Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Adv. Comput. Math.*, 32(1):103–129, 2008.

[82] G. Rozza. *Fundamentals of reduced basis method for problems governed by parametrized PDEs and applications*, pages 153–227. Springer Vienna, Vienna, 2014.

[83] G. Santin and B. Haasdonk. Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. *Dolomites Res. Notes Approx.*, 10:68–78, 2017.

[84] R. Schaback. A greedy method for solving classes of PDE problems. ArXiv 1903.11536, 2019.

[85] R. Schaback and H. Wendland. Adaptive greedy techniques for approximate solution of large RBF systems. *Numer. Algorithms*, 24(3):239–254, 2000.

[86] R. Schaback and H. Wendland. Kernel techniques: From machine learning to meshless methods. *Acta Numer.*, 15:543–639, May 2006.

[87] R. Schaback and J. Werner. Linearly constrained reconstruction of functions by kernes with applications to machine learning. *Advances in Computational Mathematics*, (25):237–258, 2006.

[88] A. Schmidt and B. Haasdonk. Reduced basis approximation of large scale parametric algebraic Riccati equations. *ESAIM: Control, Optimisation and Calculus of Variations*, 24(1):129–151, 2018.

[89] A. Schmidt, D. Wittwar, and B. Haasdonk. Rigorous and effective a-posteriori error bounds for nonlinear problems – Application to RB methods. *Advances in Computational Mathematics*, 46(32), 2020.

[90] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *Computational Learning Theory*, pages 416–426, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[91] B. Schölkopf and A.J. Smola. *Learning with Kernels.* The MIT Press, 2002.

[92] S.J. Schrödl. *Operator Valued Reproducing Kernels and Their Application in Approximation and Statistical Learning.* Berichte aus der Mathematik. Shaker, 2009.

[93] S. Sen, K. Veroy, D. B. P. Huynh, S. Deparis, N. C. Nguyen, and A. T. Patera. “Natural norm” a posteriori error estimators for reduced basis approximations. *Journal of Computational Physics*, 217:37–62, 2006.

[94] G. Söderlind. The logarithmic norm. history and modern theory. *BIT Numerical Mathematics*, 46(3):631–652, 2006.

[95] S. Steck and K. Urban. A reduced basis method for the Hamilton-Jacobi-Bellmann equation with application to the european union emission trading scheme. Preprint, University of Ulm, 2015.

[96] E. Stein and G. Weiss. *Introduction to Fourier Analysis on Euclidean Spaces.* Princeton University Press, 1971.

[97] M. L. Stein. *Interpolation of spatial data.* Springer Series in Statistics. Springer-Verlag, New York, 1999. Some theory for Kriging.

[98] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[99] I. Steinwart and A. Christmann. *Support Vector Machines.* Science + Business Media. Springer, 2008.

[100] I. Steinwart, D. Hush, and C. Scovel. Training SVMs without Offset. *J. Mach. Learn. Res.*, 12:141–202, February 2011.

[101] J. Friedman T. Hastie, R. Tibshirani. *The Elements of Statistical Learning.* Springer, New York, NY, 2 edition, 2009.

[102] A.N. Tikhonov and V.Y. Arsenin. *Solution of Ill-Posed Problems.* Winston, 1977.

[103] K. Veroy and A.T. Patera. Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations: Rigorous reduced-basis a posteriori error bounds. *International Journal for Numerical Methods in Fluids*, 47:773–788, 2005.

[104] K. Veroy, C. Prud'homme, D.V. Rovas, and A.T. Patera. A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *16th AIAA Computational Fluid Dynamics Conference.* American Institute of Aeronautics and Astronautics, 2003. Paper 2003-3847.

[105] S. Volkwein. Proper Orthogonal Decomposition: Theory and Reduced-Order Modelling. Lecture notes, Universität Konstanz, 2013.

[106] V. Vovk. Kernel ridge regression. In B. Schölkopf, Z. Luo, and V. Vovk, editors, *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 105–116. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[107] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Adv. Comput. Math.*, 4(1):389–396, 1995.

[108] H. Wendland. *Scattered Data Approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics.* Cambridge University Press, Cambridge, 2005.

[109] H. Wendland and C. Rieger. Approximate interpolation with applications to selecting smoothing parameters. *Numerische Mathematik*, 101(4):729–748, 2005.

[110] T. Wenzel, G. Santin, and B. Haasdonk. A novel class of stabilized greedy kernel approximation algorithms: Convergence, stability and uniform point distribution. *Journal of Approximation Theory*, 262:105508, 2021.

[111] D. Wirtz and B. Haasdonk. A vectorial kernel orthogonal greedy algorithm. *Dolomites Res. Notes Approx.*, 6:83–100, 2013.

[112] D. Wirtz, N. Karajan, and B. Haasdonk. Surrogate modelling of multiscale models using kernel methods. *International Journal of Numerical Methods in Engineering*, 101(1):1–28, 2015.

[113] D. Wittwar and B. Haasdonk. Greedy algorithms for matrix-valued kernels. In F. A. Radu, K. Kumar, I. Berre, J. M. Nordbotten, and I. S. Pop, editors, *Numerical Mathematics and Advanced Applications ENUMATH 2017*, pages 113–121, Cham, 2019. Springer International Publishing.

[114] D. Wittwar and B. Haasdonk. Convergence rates for matrix P-Greedy variants. In Cornelis Vuik Fred J. Vermolen, editor, *Numerical Mathematics and Advanced Applications ENUMATH 2019*, pages 1195–1203, Cham, 2021. Springer International Publishing.

[115] D. Wittwar, G. Santin, and B. Haasdonk. Interpolation with uncoupled separable matrix-valued kernels. *Dolomites Res. Notes Approx.*, 11:23–29, 2018.

[116] C. F. Wu. On some ordering properties of the generalized inverses of nonnegative definite matrices. *Linear Algebra and its Applications*, 32:49–60, 1980.

[117] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. *IEEE Transactions on Signal Processing*, 50(7):1545–1553, 2002.

[118] F. Zhang. *The Schur Complement and its Applications*, volume 4 of *Numerical Methods and Algorithms*. Springer, New York, 2005.