FUNCTION APPROXIMATION WITH KERNEL METHODS

BY

XUAN ZHOU

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Applied Mathematics
in the Graduate College of the
Illinois Institute of Technology

Approved 
Advisor

Chicago, Illinois
December 2015

ACKNOWLEDGMENT

First I would like to thank my advisor, Dr. Fred Hickernell, who offered valuable academic advice and guidance throughout my entire academic program. Not only did he teach me mathematical knowledge and skills, but he also helped me build every aspects that were essential to a researcher. I remember I could merely implement the ideas that he sketched at the beginning. I have now become more comfortable to formulate a problem and construct solutions independently. I have learned quite a lot from him and yet I have discovered that there is much more to learn every time we talked about research or just something in ordinary life. He is always a figure I look up to.

My special words of thanks also go to Dr. Greg Fasshauer. His course on meshfree methods has opened a gate for me to this interesting topic which has wide applications in so many real world problems. His impressive teaching skill helped me understand the material quickly and intrigued me to unleash deep thinking on related problems. Several research papers he co-authored with Dr. Hickernell have become the foundation this dissertation has built on.

I would also like to thank Dr. Lulu Kang, who enhanced my background in statistics, especially in machine learning. I am also glad I had the chance to work as her teaching assistant in my final year as a PhD student.

My gratitude also goes to Dr. Gady Agam, from the Department of Computer Science, for his time, insightful questions and rich comments.

Many faculty members and peer students at Illinois Institute of Technology have offered help towards my research. I thank them as well as other researchers I had fruitful discussion with at various conferences.

Finally I owe my deepest gratitude to my family. My parents gave me every-

thing I need to achieve my accomplishment. My wife, Lan Jiang, always offers unconditional support when I am pursuing my goal. My appreciation for her is beyond words. Last but not least, I thank my kids, Audrey and Alvin, for the joy they have brought to this family.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

This dissertation studies the problem of approximating functions of d variables in a separable Banach space $\mathcal{F}_d$. In particular we are interested in convergence and tractability results in the worst case setting and in the average case setting. The symmetric positive definite kernel in both settings is of a product form

$$\widetilde{K}_d(\boldsymbol{x},\boldsymbol{t}) := \prod_{\ell=1}^{d} \left(1-\alpha_\ell^2+\alpha_\ell^2 K_{\gamma_\ell}(x_\ell,t_\ell)\right) \quad \text{for all} \quad \boldsymbol{x},\boldsymbol{t}\in\mathbb{R}^d.$$

The kernel $\widetilde{K}_d$ generalizes the anisotropic Gaussian kernel, whose tractability properties have been established in the literature.

For a fixed d, we study rates of convergence, which indicate how quickly approximation errors decay. Since rates of convergence can deteriorate quickly as d increases, it is desirable to have dimension-independent convergence rates, which corresponds to the concept of strong polynomial tractability. We present sufficient conditions on $\{\alpha_\ell\}_{\ell=1}^{\infty}$ and $\{\gamma_\ell\}_{\ell=1}^{\infty}$ under which strong polynomial tractability holds for function approximation problems in $\mathcal{F}_d$. Numerical examples are presented to support the theory and guaranteed automatic algorithms are provided to solve the function approximation problem in a straightforward and efficient way.

CHAPTER 1

INTRODUCTION

The use of kernel methods has grown rapidly in function approximation and interpolation, in numerical integration, and in solving partial differential equations. This dissertation addresses the problem of function approximation with kernel methods, where the kernels are symmetric positive definite and of a product form.

We study the relation between convergence rates and the dimension of the approximation problem d. In particular we aim to identify conditions that ensure the existence of dimension-independent convergence rates.

Applications of kernel methods are diverse and include machine learning, kriging, finance, 3D reconstruction, handwriting recognition. Algorithms for function approximation based on symmetric positive definite kernels have arisen in both the numerical computation literature [3, 8, 11, 30, 39] and the statistical learning literature [1, 7, 12, 28, 31, 33, 34, 37]. These algorithms go by a variety of names, including radial basis function methods [3], scattered data approximation [39], meshfree methods [8], (smoothing) splines [37], kriging [33], Gaussian process models [28] and support vector machines [34].

The functions we want to approximate are assumed to be in a separable Banach space $\mathcal{F}_d$ of real-valued functions $f : \mathbb{R}^d \to \mathbb{R}$, associated with some kernel function $\widetilde{K}_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We consider both the worst case where $\mathcal{F}_d$ is a reproducing kernel Hilbert space associated with the reproducing kernel $\widetilde{K}_d$ and the average case where $\mathcal{F}_d$ is a sample space of Gaussian processes with the covariance kernel $\widetilde{K}_d$. The functions f are like black boxes in the sense that we obtain the outputs for given inputs without any knowledge of their internal workings. In a typical application we are given data of the form $y_i = f(\boldsymbol{x}_i)$ or $y_i = L_i(f)$ for $i = 1, \ldots, n$. That is, a function

f is sampled at the locations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, usually referred to as the *data sites* or the *design*, or more generally we know the values of n linear functionals $L_1, \ldots, L_n$ applied to f. Here we assume that the domain of f is a subset of $\mathbb{R}^d$. The goal is to construct $A_n f$, a good approximation to f that is inexpensive to evaluate, where A_n is an approximation algorithm that uses n pieces of data information.

For a fixed dimension d, convergence rates are often used to gauge the performance of an algorithm in the numerical analysis literature. However, as the dimension of the function approximation increases, we see in many problems that the convergence rates deteriorate exponentially as the dimension increases. This phenomenon is known as the *curse of dimensionality*. Since the number of applications dealing with high-dimensional data have exploded with the popularity of the concept of *big data*, it is desirable to have dimension-independent polynomial convergence rates of the form Cn^{-p}, where p denotes the smoothness of the functions, which is related to the smoothness of the kernel, and the positive constant C does not depend on d or n. This corresponds to *strong polynomial tractability*. When strong polynomial tractability does not hold, it is still desirable to have convergence rates of the form $Cd^q n^{-p}$, where $q > 0$, which corresponds to *polynomial tractability*.

The actual convergence rates depend on how the approximation error is defined. In this dissertation we study both the worst-case error and the average-case error. In the worst case setting we assume that the function to be approximated lies in a reproducing kernel Hilbert space associated with a reproducing kernel. The worst-case error is defined as the worst error of all approximations to a subset of functions in the space given by the best algorithm. In the average case setting the function to be approximated is assumed to belong to a separable Banach space equipped with a zero-mean Gaussian measure and a covariance kernel. The average-case error is the average error of all approximations to functions in the whole space given by the best

algorithm. In either case the kernel is assumed to be symmetric and positive definite. This means that for all $n \in \mathbb{N}$, $\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, and $\boldsymbol{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$, the following properties hold:

$$\widetilde{K}_d(\boldsymbol{x}, \boldsymbol{t}) = \widetilde{K}_d(\boldsymbol{t}, \boldsymbol{x}), \qquad \sum_{i=1}^{n} \sum_{j=1}^{n} \widetilde{K}_d(\boldsymbol{x}_i, \boldsymbol{x}_j) c_i c_j \geq 0. \tag{1.1}$$

We will defer the exact definitions of these errors to the next two chapters, nevertheless it is clear that the convergence rates depend on the spaces containing the functions to be approximated. Since these spaces are defined in terms of their kernels, whether (strong) polynomial tractability holds depends on which kernel we choose.

Many kernels used in practice are associated with a sequence of *shape parameters* $\{\gamma_\ell\}_{\ell=1}^d$, which allows more flexibility in the function approximation problem. Examples of such kernels include the Matérn, the multiquadrics, the inverse multiquadrics, and the extensively studied Gaussian kernel (also known as the squared exponential kernel). The anisotropic stationary Gaussian kernel used in the commercial software JMP [29] is given by

$$\widetilde{K}_d^{\mathrm{Gau}}(\boldsymbol{x}, \boldsymbol{t}) := \prod_{\ell=1}^{d} \mathrm{e}^{-\gamma_\ell^2 (x_\ell - t_\ell)^2}, \tag{1.2}$$

where γ_ℓ is a positive shape parameter for each variable x_ℓ. Choosing a small γ_ℓ has a beneficial effect on the rate of decay of the eigenvalues which appear in the eigendecomposition of the Gaussian kernel, but it also tends to cause conditioning problems for the function approximation algorithm. The optimal choice of γ_ℓ is application dependent and much work has been spent on the quest for the optimal shape parameter. Note that taking $\gamma_\ell = \gamma$ for all ℓ will recover the isotropic Gaussian kernel.

Convergence and tractability can be investigated by studying the eigendecomposition of the reproducing kernel in the worst case setting or the covariance kernel

in the average case setting. We define the integral operator associated with $\widetilde{K}_d$ by

$$Wf := W_{\widetilde{K}_d} f = \int_{\mathbb{R}^d} f(\boldsymbol{t}) \widetilde{K}_d(\cdot, \boldsymbol{t}) \varrho_d(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t} \quad \text{for all} \quad f \in \mathcal{F}_d. \tag{1.3}$$

According to Mercer's Theorem, there is an orthonormal basis $\{\varphi_{d,j}\}_{j=1}^{\infty}$ in $\mathcal{L}_2(\mathbb{R}^d, \varrho_d)$ consisting of eigenfunctions of W such that the corresponding eigenvalues $\nu_{d,j}$ are non-negative. Furthermore we can write the eigendecomposition of $\widetilde{K}_d$ as

$$\widetilde{K}_d(\boldsymbol{x}, \boldsymbol{t}) = \sum_{j=1}^{\infty} \nu_{d,j} \, \varphi_{d,j}(\boldsymbol{x}) \, \varphi_{d,j}(\boldsymbol{t}) \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{t} \in \mathbb{R}^d, \tag{1.4}$$

It is stated in [25, Theorem 5.1] that the function approximation problem is polynomially tractable in the worst case setting if there exists $\tau > 0$ such that

$$\sum_{j=1}^{\infty} \nu_{d,j}^{\tau} < \infty. \tag{1.5}$$

The counterpart of [25, Theorem 5.1] in the average case setting is [9, Lemma 1], which states that the function approximation problem is polynomially tractable in the average case setting if there exists $\tau > 0$ such that (1.5) holds. For the Gaussian kernel (1.2), convergence rates with polynomial tractability results in the worst case setting and in the average case setting are established in [10] and [9] respectively. In essence we can find the eigendecomposition of the univariate Gaussian kernel $\bar{K}_1(x,t) = \mathrm{e}^{-\gamma^2(x-t)^2}$ explicitly. The eigenvalues $\nu_{1,\gamma,j} = \tilde{\lambda}_{\gamma,j}$ are given by

$$\tilde{\lambda}_{\gamma,j} = (1 - \omega_\gamma)\omega_\gamma^{j-1},$$

where

$$\omega_\gamma = \frac{\gamma^2}{\frac{1}{2}\left(1 + \sqrt{1 + 4\gamma^2}\right) + \gamma^2}.$$

For $d > 1$, the multivariate anisotropic Gaussian kernel is a product of univariate Gaussian kernels, therefore the eigenvalues for the multivariate case are products of those for the univariate case. Specifically, let $\boldsymbol{j} = (j_1, j_2, \ldots, j_d) \in \mathbb{N}^d$. Then the eigenvalues $\nu_{d,\boldsymbol{\gamma},\boldsymbol{j}}$ are given by the products

$$\nu_{d,\boldsymbol{\gamma},\boldsymbol{j}} = \prod_{\ell=1}^{d} \tilde{\lambda}_{\gamma_\ell, j_\ell} = \prod_{\ell=1}^{d} (1 - \omega_{\gamma_\ell}) \omega_{\gamma_\ell}^{j_\ell - 1}.$$

We can sort $\nu_{d,\boldsymbol{\gamma},\boldsymbol{j}}$ in a non-increasing order to obtain $\nu_{d,j}$, and it is shown in [10] that there exists $\tau > 0$ such that (1.5) holds.

This dissertation studies kernels of a more general product form,

$$\widetilde{K}_d(\boldsymbol{x},\boldsymbol{t}) = \widetilde{K}_{d,\boldsymbol{\alpha},\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{t}) := \prod_{\ell=1}^{d} \widehat{K}_{\alpha_\ell,\gamma_\ell}(x_\ell,t_\ell) := \prod_{\ell=1}^{d} \left(1-\alpha_\ell^2+\alpha_\ell^2 K_{\gamma_\ell}(x_\ell,t_\ell)\right), \quad (1.6)$$

where $0 \le \alpha_\ell \le 1,\ \gamma_\ell > 0$. Many kernels in the machine learning literature take the form of (1.6). The Bernoulli distribution kernel that appears in [18], for example, is

$$\widetilde{K}_d^{\mathrm{Ber}}(\boldsymbol{x},\boldsymbol{t}) = \prod_{\ell=1}^{d} \left(1-x_\ell-t_\ell+2x_\ell t_\ell\right).$$

The *scale parameters* α_ℓ in (1.6) govern the vertical scale of the kernel across the ℓth dimension. In particular, taking $\alpha_\ell = 1$ for all ℓ and $K_\gamma(x,t) = \exp(-\gamma^2(x-t)^2)$ recovers the anisotropic Gaussian kernel (1.2). In many statistical learning applications the shape parameters are estimated using the training data and the user has no control of their magnitude. In such cases, as we shall see, the introduction of the scale parameters α_ℓ can help achieve dimension-independent convergence rates.

We assume that we know the eigendecomposition of the univariate kernel K_γ. This is true for certain kernels such as the Gaussians. In other cases one can specify customized eigenvalues and eigenfunctions upfront and then use (1.4) to construct kernels. The actual form of the kernel is not important if this approach is used as long as one ensures the positive definiteness of the kernel. Although knowing the eigendecomposition of K_γ does not generally provide us with explicit formulae for the eigendecomposition of the kernel $\widehat{K}_{\alpha,\gamma}$, the product form and the structure of a convex combination across each dimension indicate that we can derive upper and lower bounds of the eigenvalues of $\widehat{K}_{\alpha,\gamma}$ by approximating the corresponding integral operators by finite rank operators and applying some inequalities for eigenvalues of positive definite matrices. Lemma 1 summarizes this technique. Under certain constraints on the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, these bounds imply that there exists $\tau > 0$ such that (1.5)

holds, and consequently establish polynomial tractability for the product kernels. Convergence results are summarized in Theorem 1 and Theorem 2. Theorem 3 to Theorem 6 summarize tractability results in the worst case setting for different information classes and error criteria. These tractability results were first obtained in [41]. Tractability results in the average case setting are summarized in Theorem 7. It is worth noting that tractability results are the same for different information classes and error criteria in the average case setting under our assumptions.

These theorems give the best dimension-independent convergence rate one can achieve with product kernels. However the algorithms that yield such convergence rates are not known except in the average case setting when we are given arbitrary linear functional information, which is rarely the case in practice. Nevertheless we provide numerical examples both in the worst case setting and in the average case setting when function evaluation information is given. We apply the spline algorithm to solve the function approximation problem and observe strong polynomial tractability. Moreover the dimension-independent convergence rates appear to be consistent with those given in the theorems. These numerical examples indicate that the theoretical dimension-independent convergence rates are attainable even if some conditions in those theorems are somewhat loosened as long as we use a reasonably good algorithm.

There are numerous algorithms based on symmetric positive definite kernels which typically involving approximating functions by a linear combination of these kernels centered at a number of data sites, see e.g. [32, 23, 8, 11]. These algorithms do not require building a mesh grid upfront as other numerical methods like finite difference and finite volume methods, and are therefore suitable for problems with a complicated domain. Many existing algorithms require complicated user inputs and often lack guaranteed bounds for the approximation error. We propose guaranteed automatic algorithms that only require the user to input a black box function and a

desired error tolerance. These algorithms guarantee that the error tolerance will be met provided the input function satisfies a cone condition, characterized by the relative magnitude of a strong norm and a weak norm of the input function. Algorithm 1 works for any positive definite kernels, while Algorithm 2 gives a better convergence rate, but is only valid for Gaussian kernels. The correctness of these algorithms are summarized in Theorem 9 and Theorem 11.

This dissertation is organized as follows. Chapter 2 formulates the function approximation framework in detail, states some assumptions, and derives a key lemma (Lemma 1). Chapter 3 illustrates how the approximation error vanishes as n increases for a fixed dimension d in both the worst and average case settings. Chapter 4 and Chapter 5 present sufficient conditions under which dimension-independent convergence rates can be achieved in the worst and average case settings, respectively. Chapter 6 discusses some practical issues arisen from the kernel methods, presents some numerical examples, and introduces some easy-to-use guaranteed automatic algorithms to solve the function approximation problem. Chapter 7 proposes several open problems that can be studied in the future.

CHAPTER 2

BACKGROUND AND ASSUMPTIONS

2.1 Function Spaces

2.1.1 Reproducing Kernel Hilbert Spaces and Reproducing Kernels. For the function approximation problem in the worst case setting, we assume that the function to be approximated belongs to $\mathcal{F}_d = \mathcal{H}(\widetilde{K}_d)$. $\mathcal{F}_d$ is assumed to be continuously embedded in the space $\mathcal{L}_2 = \mathcal{L}_2(\mathbb{R}^d, \varrho_d)$ of square Lebesgue integrable functions, where ϱ_d is defined by (2.5) and the $\mathcal{L}_2$ inner product is defined by (2.6). $\mathcal{H}(K)$ denotes a reproducing kernel Hilbert space of real functions defined on $\mathbb{R}^d$ associated with the reproducing kernel K. Recall that the reproducing kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is symmetric positive definite if (1.1) holds for all $n \in \mathbb{N}$, $\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, $\boldsymbol{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$, and $f \in \mathcal{H}(K)$. The reproducing kernel also reproduces the function value, i.e., for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$f(\boldsymbol{x}) = \left\langle f, \widetilde{K}_d(\cdot, \boldsymbol{x}) \right\rangle_{\mathcal{H}(K)}.$$

Reproducing Kernel Hilbert spaces have wide applications, including complex analysis, harmonic analysis, and quantum mechanics. The representer theorem states that every function in a reproducing kernel Hilber space can be written as a linear combination of the kernel function centered at some data sites. This constitutes the foundation of function approximation algorithms using kernel methods.

2.1.2 Separable Banach Spaces and Covariance Kernels. For the function approximation problem in the average case setting, the functions to be approximated are assumed to lie in a separable Banach space, $\mathcal{F}_d$, which is also assumed to be continuously embedded in the space $\mathcal{L}_2(\mathbb{R}^d, \varrho_d)$. The space $\mathcal{F}_d$ is equipped with a

zero-mean Gaussian measure, μ_d, i.e.,

$$\int_{\mathcal{F}_d} L(f)\,\mu_d(\mathrm{d}f) = 0 \quad \text{for all} \quad L \in \mathcal{F}_d^*, \tag{2.1}$$

where $\mathcal{F}_d^*$ denotes the continuous dual space of $\mathcal{F}_d$.

Since separable Banach spaces are not necessarily Hilbert spaces, they may not admit reproducing kernels introduced in the last subsection. Nevertheless it is assumed that function evaluation at any point is a continuous linear functional on $\mathcal{F}_d$. This implies the existence of the covariance kernel, $\widetilde{K}_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined by

$$\widetilde{K}_d(\boldsymbol{x}, \boldsymbol{t}) := \int_{\mathcal{F}_d} f(\boldsymbol{x}) f(\boldsymbol{t})\,\mu_d(\mathrm{d}f). \tag{2.2}$$

2.1.3 Product Kernels. Suppose the univariate kernel K_γ is symmetric positive definite for any positive shape parameter γ and satisfies

$$\int_{\mathbb{R}} K_\gamma(t,t)\,\varrho_1(t)\,\mathrm{d}t = 1, \tag{2.3}$$

where ϱ_1 is a probability density function. The equation (2.3) is called the unit trace condition because the trace of W_{K_γ} is given by

$$\begin{aligned}\sum_{j=1}^{\infty} \tilde{\lambda}_{\gamma,j} &= \sum_{j=1}^{\infty} \tilde{\lambda}_{\gamma,j} \left\langle \tilde{\varphi}_{\gamma,j}, \tilde{\varphi}_{\gamma,j} \right\rangle_{\mathcal{L}_2(\mathbb{R},\varrho_1)} = \sum_{j=1}^{\infty} \tilde{\lambda}_{\gamma,j} \int_{\mathbb{R}} \tilde{\varphi}_{\gamma,j}^2(t)\varrho_1(t)\,\mathrm{d}t \\ &= \int_{\mathbb{R}} \sum_{j=1}^{\infty} \tilde{\lambda}_{\gamma,j} \tilde{\varphi}_{\gamma,j}^2(t)\varrho_1(t)\,\mathrm{d}t = \int_{\mathbb{R}} K_\gamma(t,t)\,\varrho_1(t)\,\mathrm{d}t,\end{aligned}$$

where $(\tilde{\lambda}_{\gamma,j}, \tilde{\varphi}_{\gamma,j})$ are eigenpairs of W_{K_γ}, and the last equality follows from (1.4). Given the sequences of scale parameters $\boldsymbol{\alpha} = \{\alpha_\ell\}_{\ell=1}^{\infty}$ and shape parameters $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell=1}^{\infty}$, where $0 < \alpha_\ell \leq 1$ and $\gamma_\ell > 0$ for all $\ell \in \mathbb{N}$, we can now define the multivariate kernel $\widetilde{K}_d$ by (1.6).

In the worst case setting, we work with the reproducing kernel Hilbert space $\mathcal{F}_d = \mathcal{H}(\widetilde{K}_d)$ associated with the reproducing kernel $\widetilde{K}_d$, while in the average case

setting we work with the separable Banach space associated with the covariance kernel $\widetilde{K}_d$. It is straightforward to see that $\widetilde{K}_d$ is also symmetric positive definite and satisfies the unit trace condition

$$\int_{\mathbb{R}^d} \widetilde{K}_d(\boldsymbol{t},\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t} = 1, \tag{2.4}$$

where ϱ_d is the probability density function with independent identical marginals, namely

$$\varrho_d(\boldsymbol{t}) = \varrho_1(t_1)\varrho_1(t_2)\cdots\varrho_1(t_d). \tag{2.5}$$

Note that assumption (2.3) is true for any product kernel (1.6) with $K_\gamma(x,x) = 1$ for all $x \in \mathbb{R}$. The unit trace condition implies that $\mathcal{F}_d$ is continuously embedded in the space $\mathcal{L}_2 = \mathcal{L}_2(\mathbb{R}^d, \varrho_d)$ of square Lebesgue integrable functions, where the $\mathcal{L}_2$ inner product is defined by

$$\langle f, g\rangle_{\mathcal{L}_2} := \int_{\mathbb{R}^d} f(\boldsymbol{t})g(\boldsymbol{t})\varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}. \tag{2.6}$$

Continuous embedding means that

$$\|I_d f\|_{\mathcal{L}_2} = \|f\|_{\mathcal{L}_2} \leq \|I_d\|\, \|f\|_{\mathcal{F}_d}, \quad \text{for all} \quad f \in \mathcal{L}_2,$$

where $I_d : \mathcal{F}_d \to \mathcal{L}_2$ is the embedding operator defined by $I_d f = f$.

2.2 Function Approximation Algorithms

Functions in $\mathcal{F}_d$ are approximated by linear algorithms of the form

$$(A_n f)(\boldsymbol{x}) := \sum_{i=1}^{n} L_i(f) a_i(\boldsymbol{x}) \quad \text{for all} \quad f \in \mathcal{F}_d, \quad \boldsymbol{x} \in \mathbb{R}^d \quad \text{for all} \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{2.7}$$

for some continuous linear functionals $L_i \in \mathcal{F}_d^*$, and functions $a_i \in \mathcal{L}_2$. Note that for known functions a_i, the cost of computing $A_n f$ is $\mathcal{O}(n)$, if we do not consider the cost of generating the data samples $L_i(f)$. The linear functionals, L_i, used by an algorithm A_n may either come from the class of arbitrary bounded linear functionals, $\Lambda^{\mathrm{all}} = \mathcal{F}_d^*$, or from the class of function evaluations, Λ^{std}.

The approximation, $A_n f$, considered here uses partial information about f, namely, n continuous linear functional evaluations denoted $L_1(f), L_2(f), \ldots, L_n(f)$. It is known that nonlinear algorithms and adaptive algorithms do not essentially help for the $\mathcal{L}_2$ approximation problem, see [36, 38]. Therefore we can restrict our attention to linear algorithms (2.7).

When one is able to draw data from the class Λ^{all}, the optimal design or sampling scheme, $\{L_i\}_{i=1}^n$, is known, see [26], [36], [38] and [25, Section 4.2]. In both the worst case setting and the average case setting the optimal sampling scheme is to choose

$$L_i(f) = \langle f, \varphi_{d,i} \rangle_{\mathcal{L}_2}, \tag{2.8}$$

for which the linear algorithm that minimizes the average case error corresponds to projecting the function f into the linear space spanned by the eigenfunctions corresponding to the n largest eigenvalues. This algorithm is of the form

$$A_n f = \sum_{i=1}^{n} \langle f, \varphi_{d,i} \rangle_{\mathcal{L}_2} \varphi_{d,i}, \tag{2.9}$$

where $\varphi_{d,i}$ are the eigenfunctions appearing in (1.4).

2.3 Convergence and Tractability

We address two problems of function approximation: convergence rate and tractability. The former considers how fast the error decays as n increases. This is the typical point of view taken in numerical analysis and many results in the kernel literature are summarized in, e.g., [8, 39]. However, studying convergence rates alone does not take into consideration the effects of d. The study of tractability arises in information-based complexity and it considers how the error depends on both d and n.

We now illustrate how errors are precisely defined. We introduce the setting variable Ξ and set $\Xi = \Xi^{\text{wor}}$ in the worst case setting and $\Xi = \Xi^{\text{avg}}$ in the average

case setting. The $\mathcal{L}_2$ function approximation error of an algorithm A is given by

$$\epsilon_d(A;\Xi) := \begin{cases} \sup_{\|f\|_{\mathcal{F}_d}\le 1} \|I_d f - Af\|_{\mathcal{L}_2}, & \Xi = \Xi^{\mathrm{wor}}, \\ \left(\int_{\mathcal{F}_d} \|I_d f - Af\|_{\mathcal{L}_2}^2 \,\mu(\mathrm{d}f)\right)^{1/2}, & \Xi = \Xi^{\mathrm{avg}}. \end{cases}$$

Note that $\|I_d f - Af\|_{\mathcal{L}_2} \le \epsilon_d(A;\Xi^{\mathrm{wor}})\|f\|_{\mathcal{F}_d}$ because I_d and A are linear operators. Let $\mathcal{A}_n$ denote the set of all algorithms that use at most n pieces of function information. The *nth minimal error* over all possible algorithms is defined as

$$e_d(n;\Xi,\Lambda) := \inf_{\{L_i\}_{i=1}^n \subset \Lambda} \inf_{A\in\mathcal{A}_n} \epsilon_d(A;\Xi), \quad \Lambda \in \{\Lambda^{\mathrm{all}}, \Lambda^{\mathrm{std}}\}. \tag{2.10}$$

2.3.1 Convergence. For a fixed d, we would like to know how fast $e_d(n;\Xi,\Lambda)$ vanishes as n goes to infinity. Let $R(\boldsymbol{z})$, the *rate of convergence* of any sequence $\boldsymbol{z} = \{z_n\}_{n\in\mathbb{N}}$, be defined by

$$r(\boldsymbol{z}) := \sup\left\{\beta > 0 : \sum_{n=1}^{\infty} z_n^{1/\beta} < \infty\right\}$$

with the convention that the supremum of the empty set is taken to be zero. In particular, we study the rate of convergence of the sequence $e_d(n;\Xi,\Lambda)$. Since the numbers $e_d(n;\Xi,\Lambda)$ are ordered, we have

$$R_d(\Xi,\Lambda) := r\left(\{e_d(n;\Xi,\Lambda)\}_{n\in\mathbb{N}}\right) = \sup\left\{\beta \ge 0 : \lim_{n\to\infty} e_d(n;\Xi,\Lambda) n^{\beta} = 0\right\}.$$

Roughly speaking, $R(\Xi,\Lambda)$ is the largest β for which the nth minimal errors $e_d(n;\Xi,\Lambda)$ behave like $n^{-\beta}$. Obviously, $R(\Lambda^{\mathrm{all}},\Xi) \ge R(\Lambda^{\mathrm{std}},\Xi)$. We would like to know both rates for the product kernels (1.6).

2.3.2 Tractability. While typical numerical analysis focuses on the rate of convergence, it does not take into consideration the effects of d. The study of tractability arises in information-based complexity and it considers how the error depends on the dimension, d, as well as the number of data, n.

In particular, we would like to know how $e_d(n;\Xi,\Lambda)$ depends on n and d under the *absolute* and *normalized* error criteria. For a given positive $\varepsilon \in (0,1)$ we want to find an algorithm A_n with the smallest n for which the error does not exceed ε under the absolute error criterion, and does not exceed $\varepsilon e_d(0;\Xi,\Lambda) = \varepsilon\|I_d\|$ under the normalized error criterion. That is,

$$n_d(\varepsilon;\Xi,\Lambda,\Psi) = \min\left\{ n : e_d(n;\Xi,\Lambda) \le \begin{cases} \varepsilon, & \Psi = \Psi^{\mathrm{abs}}, \\ \varepsilon\,\|I_d\|, & \Psi = \Psi^{\mathrm{nor}}, \end{cases} \right\}$$

where we introduce the error criterion variable Ψ and set $\Psi = \Psi^{\mathrm{abs}}$ under the absolute error criterion and $\Psi = \Psi^{\mathrm{nor}}$ under the normalized error criterion.

Let $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ denote the sequence of function approximation problems. We say that $\mathcal{I}$ is *polynomially tractable* if and only if there exist constants $C > 0$, $p, q \ge 0$ such that

$$n_d(\varepsilon;\Xi,\Lambda,\Psi) \le C\, d^{\,q}\, \varepsilon^{-p} \quad \text{for all} \quad d \in \mathbb{N} \quad \text{and} \quad \varepsilon \in (0,1).$$

If $q = 0$ above then we say that $\mathcal{I}$ is *strongly polynomially tractable* and the infimum of p satisfying the bound above is the *exponent* of strong polynomial tractability. Precisely, the exponent of strong polynomial tractability $p(\Xi,\Lambda,\Psi)$ is given by

$$p(\Xi,\Lambda,\Psi) = \inf\left\{ p \ge 0 : \sup_{d\in\mathbb{N},\varepsilon\in(0,1)} \varepsilon^p n_d(\varepsilon;\Xi,\Lambda,\Psi) < \infty \right\}.$$

The essence of polynomial tractability is to guarantee that a polynomial number of linear functionals or function values is enough to satisfy the function approximation problem to within ε. Strong polynomial tractability is especially challenging but desirable at the same time since then the number of linear functionals needed to achieve a tolerance within ε is independent of d. The problem becomes even more complicated when the eigendecomposition of the kernel is not explicitly known. Nevertheless, we provide here positive results on strong polynomial tractability for the product kernels under certain conditions.

Before delving into the detailed derivations of $p(\Xi, \Lambda, \Psi)$, we stress that they do not depend on n, d, or ε. They obviously depend on whether we choose the worst case setting or the average case setting. They also depend on the class of available function information Λ, and whether the absolute or normalized error criteria is used. Moreover they may also depend on the spaces $\mathcal{F}_d$ and the probability density function ϱ_d. The dependence on $\mathcal{F}_d$ in the worst case setting is through the reproducing kernel $\widetilde{K}_d$, while the dependence in the average case setting is through the measure μ_d, and consequently, the covariance kernel $\widetilde{K}_d$.

2.4 The Eigendecomposition of the Kernel

It can be clearly seen from (2.10) that we would like to find the optimal sampling scheme and the best linear algorithm. This proves to be a non-trivial task. Nevertheless for the class Λ^{all}, the optimal design and the best algorithm are given by (2.8) and (2.9). The nth minimal error in the worst case setting is given by

$$e_d(n; \Xi^{\text{wor}}, \Lambda^{\text{all}}) = \sqrt{\nu_{d,n+1}}, \quad \text{for all} \quad n \in \mathbb{N}, \tag{2.11}$$

and the nth minimal error in the average case setting is given by

$$e_d(n; \Xi^{\text{avg}}, \Lambda^{\text{all}}) = \left(\sum_{j=n+1}^{\infty} \nu_{d,j} \right)^{1/2}, \quad \text{for all} \quad n \in \mathbb{N}. \tag{2.12}$$

Equations (2.11) and (2.12) suggest convergence and tractability of the function approximation be closely related to the eigenvalues $\nu_{d,j}$. We provide below a detailed discussion of the eigendecomposition of a reproducing kernel $\widetilde{K}_d$ associated with a Hilbert space $\mathcal{F}_d = \mathcal{H}(\widetilde{K}_d)$ for completeness.

Let $W = I_d^* I_d : \mathcal{F}_d \to \mathcal{F}_d$, where $I_d^* : \mathcal{L}_2 \to \mathcal{F}_d$ denotes the adjoint of the embedding operator, i.e., the operator that satisfies $\langle f, I_d^* h \rangle_{\mathcal{F}_d} = \langle I_d f, h \rangle_{\mathcal{L}_2}$ for all $f \in \mathcal{F}_d$ and $h \in \mathcal{L}_2$. As a consequence, W is a self-adjoint and positive definite linear operator given by (1.3), because $\widetilde{K}_d$ is symmetric positive definite, In fact, W is a

Hilbert-Schmidt operator, see e.g., [17], and is hence a compact operator. Clearly,

$$\langle f, g\rangle_{\mathcal{L}_2} = \langle I_d f, I_d g\rangle_{\mathcal{L}_2} = \langle Wf, g\rangle_{\mathcal{F}_d} = \langle f, Wg\rangle_{\mathcal{F}_d} \quad \text{for all} \quad f, g \in \mathcal{F}_d.$$

It is known that $\lim_{n\to\infty} e(n; \Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}) = 0$ if and only if W is compact (see, e.g., [25, Section 4.2]).

The integral operator W can be seen as a generalization of matrix operators in a finite dimensional space. Let us denote the eigenpairs of W by $(\nu_{d,j}, \eta_{d,j})$, where the eigenvalues are ordered, $\nu_{d,1} \geq \nu_{d,2} \geq \cdots$, and

$$W\eta_{d,j} = \nu_{d,j}\,\eta_{d,j} \quad \text{with} \quad \langle \eta_{d,j}, \eta_{d,i}\rangle_{\mathcal{F}_d} = \delta_{i,j} \ \text{ for all } \ i, j \in \mathbb{N}. \tag{2.13}$$

Note also that for any $f \in \mathcal{F}_d$ we have

$$\langle f, \eta_{d,j}\rangle_{\mathcal{L}_2} = \nu_{d,j}\,\langle f, \eta_{d,j}\rangle_{\mathcal{F}_d}.$$

Taking $f = \eta_{d,i}$ we see that $\{\eta_{d,j}\}$ is a set of orthogonal functions in $\mathcal{L}_2$. Letting

$$\varphi_{d,j} = \nu_{d,j}^{-1/2}\eta_{d,j} \quad \text{for all} \quad j \in \mathbb{N},$$

we obtain an orthonormal sequence $\{\varphi_{d,j}\}$ in $\mathcal{L}_2$. Since $\{\eta_{d,j}\}$ is a complete orthonormal basis of $\mathcal{F}_d$, we have

$$\widetilde{K}_d(\boldsymbol{x}, \boldsymbol{t}) = \sum_{j=1}^{\infty} \eta_{d,j}(\boldsymbol{x})\,\eta_{d,j}(\boldsymbol{t}) = \sum_{j=1}^{\infty} \nu_{d,j}\,\varphi_{d,j}(\boldsymbol{x})\,\varphi_{d,j}(\boldsymbol{t}) \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{t} \in \mathbb{R}^d.$$

This is the expansion given by Mercer's theorem and we stress that the same expansion holds when $\widetilde{K}_d$ is a covariance kernel associated with a separable Banach space. The unit trace condition (2.4) implies that W is a Hilbert-Schmidt operator:

$$\sum_{j=1}^{\infty} \nu_{d,j} = \int_{\mathbb{R}^d} \widetilde{K}_d(\boldsymbol{t}, \boldsymbol{t})\varrho_d(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} = 1.$$

Similarly to (1.3), the integral operators corresponding to the one-dimensional kernels $\widehat{K}_{\alpha,\gamma}$ and K_γ are given by

$$W_{\widehat{K}_{\alpha,\gamma}} f = \int_{\mathbb{R}} f(t)\widehat{K}_{\alpha,\gamma}(\cdot, t)\varrho_1(t)\,\mathrm{d}t \quad \text{for all} \quad f \in \mathcal{H}(\widehat{K}_{\alpha,\gamma}).$$

$$W_{K_\gamma} g = \int_{\mathbb{R}} g(t)K_\gamma(\cdot, t)\varrho_1(t)\,\mathrm{d}t \quad \text{for all} \quad g \in \mathcal{H}(K_\gamma).$$

To standardize the notation, we drop the dependency on the dimension d to denote the eigenvalues of $W_{\widehat{K}_{\alpha,\gamma}}$ by $\tilde{\nu}_{\alpha,\gamma,1} \geq \tilde{\nu}_{\alpha,\gamma,2} \geq \cdots$. Similarly the eigenvalues of W_{K_γ} are denoted by $\tilde{\lambda}_{\gamma,1} \geq \tilde{\lambda}_{\gamma,2} \geq \cdots$. Since $\widetilde{K}_d(\boldsymbol{x}, \boldsymbol{t})$ is the product of $\widehat{K}(x_\ell, t_\ell)$, the eigenvalues $\nu_{d,j}$ are products of the eigenvalues $\tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}$. To be specific, for $\boldsymbol{j} = (j_1, j_2, \ldots, j_d) \in \mathbb{N}^d$, the eigenvalues $\nu_{d,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{j}}$ are given by

$$\nu_{d,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{j}} = \prod_{\ell=1}^{d} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j_\ell}.$$

Note that while the eigenvalues $\tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}$ are ordered in non-increasing magnitude, $\nu_{d,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{j}}$ are not. Yet we can sort $\nu_{d,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{j}}$ in a non-increasing order and relabel them to obtain $\nu_{d,j}$ that appear in (2.13). A useful relation between $\nu_{d,j}$ and $\tilde{\nu}_{\alpha,\gamma,j}$ is given by [10, Lemma 3.1]:

$$\sum_{j=1}^{\infty} \nu_{d,j}^{\tau} = \prod_{\ell=1}^{d} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}^{\tau} \right), \quad \tau > 0. \tag{2.15}$$

We want to investigate whether (strong) polynomial tractability holds when the dimension d is large. According to [25], strong polynomial tractability holds if $\sum_{j=1}^{\infty} \nu_{d,j}^{\tau}$ can be bounded for some $\tau > 0$. We assume that we know the eigenvalues $\lambda_{\gamma,j}$ of W_{K_γ} in terms of the shape parameter γ. Even though knowing $\lambda_{\gamma,j}$ does not give us immediate knowledge of the eigenvalues $\tilde{\nu}_{\alpha,\gamma}$, the following lemma provides us with some useful inequalities on eigenvalues of the integral operators corresponding to the kernels.

Lemma 1. *Let K_A, K_B and K_C be univariate symmetric positive definite kernels such that*

$$\int_{\mathbb{R}} K(t,t)\varrho_1(t)\,\mathrm{d}t < \infty, \quad K \in \{K_A, K_B, K_C\},$$

and $K_C = aK_A + bK_B$, $a, b \geq 0$. Assume that the Banach spaces they are associated with, $\mathcal{F}_1(K)$, $K \in \{K_A, K_B, K_C\}$, are continuously embedded in $\mathcal{L}_2(\mathbb{R}, \varrho_1)$. Define the integral operators W_{K_A}, W_{K_B}, and W_{K_C} by

$$W_K f = \int_{\mathbb{R}} f(t)K(\cdot,t)\varrho_1(t)\,\mathrm{d}t, \quad \textit{for all} \quad f \in \mathcal{F}_1(K), \quad K \in \{K_A, K_B, K_C\}.$$

Let the eigenvalues of the operators be sorted in a weakly decreasing order, i.e. $\lambda_{K,1} \geq \lambda_{K,2} \geq \cdots$. *Then these eigenvalues satisfy*

$$\lambda_{K_C,i+j+1} \leq a\lambda_{K_A,i+1} + b\lambda_{K_B,j+1}, \quad i,j = 1,2,\ldots \tag{2.16}$$

$$\lambda_{K_C,i} \geq \max(a\lambda_{K_A,i}, b\lambda_{K_B,i}), \quad i = 1,2,\ldots \tag{2.17}$$

Proof. Let $\{u_j\}_{j\in\mathbb{N}}$ be any orthonormal basis in $\mathcal{L}_2(\mathbb{R},\varrho_1)$. We assign the orthogonal projections P_n given by

$$P_n g = \sum_{j=1}^{n} \langle g, u_j\rangle\, u_j, \quad g \in \mathcal{L}_2(\mathbb{R},\varrho_1).$$

Since W_{K_A} is compact due to (1), it can be shown that $\|(I-P_n)W_{K_A}\| \to 0$ as $n\to\infty$, where the operator norm

$$\|(I-P_n)W_{K_A}\| := \sup_{\|g\|_{\mathcal{L}_2(\mathbb{R},\varrho_1)}\leq 1} \|(I-P_n)W_{K_A}g\|_{\mathcal{L}_2(\mathbb{R},\varrho_1)}.$$

Furthermore [27, Lemma 11.1 (OS_2)] states that for every pair $T_1, T_2 : \mathcal{L}_2(\mathbb{R},\varrho_1) \to \mathcal{L}_2(\mathbb{R},\varrho_1)$ of compact operators we have $|s_j(T_1) - s_j(T_2)| \leq \|T_1 - T_2\|,\ j\in\mathbb{N}$, where the singular values $s_j(T_k), k = 1,2$ are the square roots of the eigenvalues $\lambda_j(T_k^*T_k)$ arranged in a weakly decreasing order, thus $s_j(T_k) = \sqrt{\lambda_j(T_k^*T_k)}$. Now we can bound

$$\begin{aligned}
|s_j(W_{K_A}) - s_j(P_nW_{K_A}P_n)| &\leq |s_j(W_{K_A}) - s_j(P_nW_{K_A})| + |s_j(P_nW_{K_A}) - s_j(P_nW_{K_A}P_n)| \\
&\leq \|W_{K_A} - P_nW_{K_A}\| + \|P_nW_{K_A} - P_nW_{K_A}P_n\| \\
&\leq \|(I-P_n)W_{K_A}\| + \|W_{K_A}(I-P_n)\| \to 0
\end{aligned}$$

as $n\to\infty$. Thus the eigenvalues $\lambda_{P_nW_{K_A}P_n,j} \to \lambda_{W_{K_A},j}$ for all j as $n\to\infty$. Similarly this applies to the operators W_{K_B} and W_{K_C}. Note that we have

$$P_nW_{K_C}P_n = aP_nW_{K_A}P_n + bP_nW_{K_B}P_n$$

and these finite rank operators correspond to self-adjoint matrices. These matrices are symmetric positive definite because the kernels are symmetric positive definite.

The inequalities (2.16) are found by Weyl (see [21]) and (2.17) are a direct result of [2, Fact 8.19.4]. Since (2.16) and (2.17) hold for the eigenvalues of symmetric positive definite matrices, they also hold for the eigenvalues of the integral operators corresponding to symmetric positive definite kernels. □

We are ready to derive convergence and tractability results for product kernels now.

CHAPTER 3

RATES OF CONVERGENCE

In this chapter we consider the function approximation problem with product kernels (1.6). The dimension d is arbitrary, but fixed, throughout this chapter. As we shall see, rates of convergence, as well as tractability which is studied in Chapter 4 and Chapter 5, depend on the rate of decay of the products of scale and shape parameters $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ given by

$$\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) := r\left(\{\alpha_\ell \gamma_\ell\}_{\ell \in \mathbb{N}}\right) = \sup\left\{\beta > 0 : \sum_{\ell=1}^{\infty} (\alpha_\ell \gamma_\ell)^{1/\beta} < \infty\right\}, \tag{3.1}$$

with the convention that the supremum of the empty set is taken to be zero.

3.1 Rates of Convergence in the Worst Case Setting

Under appropriate assumptions on the eigenvalues, arbitrarily large rates of convergence can be achieved in the worst case setting. We summarize this result in the following theorem.

Theorem 1. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d \in \mathbb{N}}$ for Hilbert spaces with the symmetric positive definite product kernels* (1.6) *satisfying* (2.3) *in the worst case setting. If $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 0$ and there exist constants $C_1, C_2 > 0$ such that*

$$\int_{\mathbb{R}^2} K_\gamma(x,t)\varrho_1(x)\varrho_1(t)\,\mathrm{d}x\mathrm{d}t \geq 1 - C_1\gamma^2, \tag{3.2}$$

$$\sum_{j=2}^{\infty} \left(\frac{\tilde{\lambda}_{\gamma,j}}{\gamma^2}\right)^{\frac{1}{2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})}} \leq C_2 \tag{3.3}$$

hold for all $0 < \gamma < \sup\{\gamma_\ell : \ell \in \mathbb{N}\}$, then

$$R(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}) = \tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}),$$
$$R(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{std}}) = \frac{2\tilde{r}^2(\boldsymbol{\alpha}, \boldsymbol{\gamma})}{2\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) + 1}.$$

Proof. Consider first the case $d = 1$ for simplicity. Then the kernel $\widetilde{K}_{d,\boldsymbol{\alpha},\boldsymbol{\gamma}}$ in (1.6) becomes $\widehat{K}_{\alpha,\gamma}$. We will show that for $\tau = 1/(2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}))$, the eigenvalues of $\widehat{K}_{\alpha,\gamma}$ satisfy

$$\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha,\gamma,j}^{\tau} \leq 1 + C_{\mathrm{U}}(\alpha\gamma)^{2\tau}, \tag{3.4}$$

where the constant C_{U} does not depend on α or γ. Since all the eigenvalues of K_γ are non-negative due to the positive definiteness of K_γ, we clearly have for the first eigenvalue of K_γ,

$$\tilde{\nu}_{\alpha,\gamma,1} \leq 1. \tag{3.5}$$

On the other hand, (3.2) gives the lower bound of the first eigenvalue of $\widehat{K}_{\alpha,\gamma}$

$$\begin{aligned}\tilde{\nu}_{\alpha,\gamma,1} &\geq \int_{\mathbb{R}^2} \widehat{K}_{\alpha,\gamma}(x,t)\varrho_1(x)\varrho_1(t)\,\mathrm{d}t\mathrm{d}x = \int_{\mathbb{R}^2} \left(1 - \alpha^2 + \alpha^2 K_\gamma(x,t)\right) \varrho_1(x)\varrho_1(t)\,\mathrm{d}t\mathrm{d}x \\ &= 1 - \alpha^2 + \alpha^2 \int_{\mathbb{R}^2} K_\gamma(x,t)\varrho_1(x)\varrho_1(t)\,\mathrm{d}t\mathrm{d}x \geq 1 - C_1(\alpha\gamma)^2. \end{aligned} \tag{3.6}$$

The unit trace condition (2.4) implied that $\sum_{j=1}^{\infty} \lambda_{\gamma,j} = 1$. It follows that

$$\tilde{\nu}_{\alpha,\gamma,2} \leq C_1(\alpha\gamma)^2. \tag{3.7}$$

For $j \geq 3$, the upper bound of $\tilde{\nu}_{\alpha,\gamma,j}$ is given by (2.16) with $i = 1$:

$$\tilde{\nu}_{\alpha,\gamma,j} \leq \alpha^2 \tilde{\lambda}_{\gamma,j-1}, \tag{3.8}$$

which in turn yields

$$\sum_{j=3}^{\infty} \tilde{\nu}_{\alpha,\gamma,j}^{\tau} \leq \alpha^{2\tau} \sum_{j=3}^{\infty} \tilde{\lambda}_{\gamma,j-1}^{\tau} \leq C_2(\alpha\gamma)^{2\tau} \tag{3.9}$$

by (3.3). Combining (3.5), (3.7) and (3.9) gives (3.4), where the constant $C_{\mathrm{U}} = C_1^{\tau} + C_2$.

For the multivariate case $d > 1$, the sum of the τth power of the eigenvalues is bounded from above for any $\tau > 1/(2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}))$ because

$$\begin{aligned}\sum_{j=1}^{\infty} \nu_{d,j}^{\tau} &= \prod_{\ell=1}^{d} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}^{\tau} \right) \leq \prod_{\ell=1}^{\infty} \left(1 + C_{\mathrm{U}}(\alpha_\ell\gamma_\ell)^{2\tau}\right) \\ &= \exp\left(\sum_{\ell=1}^{\infty} \ln\left(1 + C_{\mathrm{U}}(\alpha_\ell\gamma_\ell)^{2\tau}\right) \right) \leq \exp\left(C_{\mathrm{U}} \sum_{\ell=1}^{\infty} (\alpha_\ell\gamma_\ell)^{2\tau} \right) < \infty. \end{aligned} \tag{3.10}$$

For the class Λ^{all} we know that $e_d(n;\Xi^{\mathrm{wor}},\Lambda^{\mathrm{all}}) = \sqrt{\nu_{d,n+1}}$. The ordering of $\nu_{d,j}$ implies that

$$\nu_{d,n+1} \le \left(\frac{1}{n+1}\sum_{j=1}^{n+1}\nu_{d,j}^{\tau}\right)^{1/\tau} \le \left(\frac{1}{n+1}\sum_{j=1}^{\infty}\nu_{d,j}^{\tau}\right)^{1/\tau} = \frac{1}{(n+1)^{1/\tau}}\left(\sum_{j=1}^{\infty}\nu_{d,j}^{\tau}\right)^{1/\tau}. \tag{3.11}$$

Since we have shown that $\left(\sum_{j=1}^{\infty}\nu_{d,j}^{\tau}\right)^{1/\tau} < \infty$ if $\tau > 1/(2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}))$, $\nu_{d,n+1}$ is proportional to $(n+1)^{-1/\tau}$ times a number that does not depend on n. This implies that $r(\Xi^{\mathrm{wor}},\Lambda^{\mathrm{all}}) \ge 1/(2\tau)$ if $\tau > 1/(2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}))$, and we conclude that $r(\Xi^{\mathrm{wor}},\Lambda^{\mathrm{all}}) \ge \tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})$.

Consider now the class Λ^{std}. We use [22, Theorem 5], which states that if there exist numbers $p > 1$ and $B > 0$ such that

$$\nu_{d,n} \le Bn^{-p} \quad \text{for all} \quad n \in \mathbb{N} \tag{3.12}$$

then for all $\delta \in (0,1)$ and $n \in \mathbb{N}$ there exists a linear algorithm A_n that uses at most n values and its worst case error is bounded by

$$\epsilon_d(A_n;\Xi^{\mathrm{wor}}) \le BC_{\delta,p}(n+1)^{-(1-\delta)p^2/(2p+2)}.$$

Here, $C_{\delta,p}$ is independent of n and d but may depend on δ and p.

Note that the assumption (3.12) holds in our case for $p < 2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})$ with B that can depend on d. Hence, $R(\Xi^{\mathrm{wor}},\Lambda^{\mathrm{std}}) \ge (1-\delta)p^2/(2p+2)$, and since δ can be arbitrarily small we conclude that $R(\Xi^{\mathrm{wor}},\Lambda^{\mathrm{std}}) \ge 2\tilde{r}^2(\boldsymbol{\alpha},\boldsymbol{\gamma})/(2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})+1)$. □

We stress that the algorithm A_n that was used in the proof for the class Λ^{std} is non-constructive. Nonetheless, given a design, it is known that the *spline algorithm* in [10, Section 2.3] is the best way to use the function data given via the design.

Theorem 1 states that the rate of convergence for a fixed d is at least the same order as the rate at which eigenvalues of the integral operator corresponding to the

reproducing kernel decays. The rate of convergence for the class Λ^{all} is slightly better than that for the class Λ^{std}.

3.2 Rates of Convergence in the Average Case Setting

The following theorem summarizes rates of convergence in the average case setting.

Theorem 2. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with the symmetric positive definite product kernels* (1.6) *satisfying* (2.3) *in the average case setting. If $\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}) > 1/2$ and there exist constants $C_1, C_2 > 0$ such that* (3.2) *and* (3.3) *hold for all $0 < \gamma < \sup\{\gamma_\ell : \ell \in \mathbb{N}\}$, then*

$$R(\Xi^{\mathrm{avg}}, \Lambda^{\mathrm{all}}) = R(\Xi^{\mathrm{avg}}, \Lambda^{\mathrm{std}}) = \tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}) - 1/2.$$

Proof. Consider first the class Λ^{all}. If $\tau > 1/(2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}))$, we have

$$e_d^2(n; \Xi^{\mathrm{avg}}, \Lambda^{\mathrm{all}}) = \sum_{k=n+1}^{\infty} \nu_{d,k} \le \left(\sum_{j=1}^{\infty} \nu_{d,j}^{\tau}\right)^{1/\tau} \sum_{k=n+1}^{\infty} \frac{1}{k^{1/\tau}} \le C n^{1-1/\tau},$$

where the constant C does not depend on n. We have used (3.11) in the first inequality, and the second inequality follows from the upper bound of the tail sum of the Riemann zeta function. This implies that $R(\Xi^{\mathrm{avg}}, \Lambda^{\mathrm{all}}) \ge 1/(2\tau) - 1/2$ if $\tau > 1/(2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}))$ and we conclude that $R(\Xi^{\mathrm{avg}}, \Lambda^{\mathrm{all}}) = \tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}) - 1/2$.

Consider now the class Λ^{std}. It is shown in [15] that in the average case setting there exist algorithms for the class Λ^{std} that achieve convergence rates of the same order as the optimal algorithm for the class Λ^{all} does. This implies that $R(\Xi^{\mathrm{avg}}, \Lambda^{\mathrm{all}}) = R(\Xi^{\mathrm{avg}}, \Lambda^{\mathrm{std}})$ and completes the proof. □

We remark that for kernels which have large decay rates for the eigenvalues such as the Gaussians, arbitrarily large rates of convergence can be achieved. However, the rate of convergence tells us nothing about the dependence on d. We are especially

concerned about a possible exponential dependence of d when d is large. The result may also depend on α_ℓ and γ_ℓ. This is the subject of the next several chapters.

CHAPTER 4

TRACTABILITY IN THE WORST CASE SETTING

4.1 Tractability Under the Absolute Error Criterion

We now consider the function approximation problem for Hilbert spaces $\mathcal{F}_d$ with product kernels (1.6) under the absolute error criterion.

4.1.1 Arbitrary Linear Functionals. We first analyse (strong) polynomial tractability for the class Λ^{all}.

Theorem 3. *Consider under the absolute error criterion the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces associated with the product kernels* (1.6) *satisfying* (2.3) *for the class Λ^{all} in the worst case setting. If $\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}) = 0$ or there exist constants $C_1, C_2, C_3 > 0$ such that* (3.2) *and*

$$C_2 \le \sum_{j=2}^{\infty} \left(\frac{\tilde{\lambda}_{\gamma,j}}{\gamma^2}\right)^{\frac{1}{2\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})}} \le C_3 \tag{4.1}$$

hold for all $0 < \gamma < \sup\{\gamma_\ell : \ell \in \mathbb{N}\}$, then

- *$\mathcal{I}$ is strongly polynomially tractable with exponent*

$$p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}}) = \min\left(2, \frac{1}{\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})}\right).$$

- *For all $d \in \mathbb{N}$ we have*

$$e_d(n; \Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}) \preceq n^{-1/p(\Xi^{\mathrm{wor}},\Lambda^{\mathrm{all}},\Psi^{\mathrm{abs}})} = n^{-\max(\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}),1/2)} \quad \text{as} \quad n \to \infty,$$
$$n_d(\varepsilon; \Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}}) \preceq \varepsilon^{-p(\Xi^{\mathrm{wor}},\Lambda^{\mathrm{all}},\Psi^{\mathrm{abs}})} \quad \text{as} \quad \varepsilon \to 0,$$

  where the notation $\preceq n^q$ means that for all $\delta > 0$ the quantity is bounded above *by $C_\delta n^{q+\delta}$ for all $n > 0$, where C_δ is some positive constant that is independent of the sample size n and the dimension d, but may depend on δ. $\preceq \varepsilon^q$ with $\varepsilon \to 0$ is analogous to $\preceq (1/\varepsilon)^{-q}$ with $1/\varepsilon \to \infty$.*

- *For the isotropic kernel with $\alpha_\ell = \alpha$ and $\gamma_\ell = \gamma$ for all ℓ, the exponent of strong tractability is 2. Furthermore strong polynomial tractability is equivalent to polynomial tractability.*

Proof. From [25, Theorem 5.1] it follows that $\mathcal{I}$ is strongly polynomially tractable if and only if there exist two positive numbers c_1 and τ such that

$$c_2 := \sup_{d\in\mathbb{N}} \left(\sum_{j=\lceil c_1 \rceil}^{\infty} \nu_{d,j}^{\tau} \right)^{1/\tau} < \infty, \tag{4.2}$$

Furthermore, the exponent $p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}})$ of strong polynomial tractability is the infimum of 2τ for which this condition holds. Obviously (4.2) holds for $c_1 = 1$ and $\tau = 1$ because

$$\begin{aligned} \sum_{j=1}^{\infty} \nu_{d,j} &= \prod_{\ell=1}^{d} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j} \right) = \prod_{\ell=1}^{d} \left(\int_{\mathbb{R}} (1 - \alpha_\ell^2 + \alpha_\ell^2 K_{\gamma_\ell}(t,t)) \varrho_1(t) dt \right) \\ &= \prod_{\ell=1}^{d} \left(1 - \alpha_\ell^2 + \alpha_\ell^2\right) = 1. \end{aligned}$$

This shows that $p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}}) \le 2$.

The case $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$ is trivial. Take now $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 0$. The finiteness of $\sum_{j=1}^{\infty} \nu_{d,j}^{\tau}$ shown in (3.10) implies that $p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}}) \le 1/\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$.

Consider now the lower bound of the sum of the τth powers of the eigenvalues in the univariate case $d = 1$. The lower bound we want to establish is that for $\tau < 1/(2\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}))$,

$$\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha,\gamma,j}^{\tau} \ge 1 + C_{\mathrm{L}} (\alpha\gamma)^{2\tau} \quad \text{if} \quad \alpha\gamma < \left(\frac{C_2}{2C_1} \right)^{1/[2(1-\tau)]}, \tag{4.3}$$

where $C_{\mathrm{L}} := C_2/2$. It follows from (3.6) that

$$\tilde{\nu}_{\alpha,\gamma,1}^{\tau} \ge \tilde{\nu}_{\alpha,\gamma,1} \ge 1 - C_1 (\alpha\gamma)^2. \tag{4.4}$$

In addition we apply the eigenvalue inequality (2.16) to obtain

$$\tilde{\nu}_{\alpha,\gamma,j} \ge \alpha^2 \tilde{\lambda}_{\gamma,j}, \quad j = 2, 3, \ldots$$

which in turn gives

$$\sum_{j=2}^{\infty} \tilde{\nu}_{\alpha,\gamma,j}^{\tau} \geq \alpha^{2\tau} \sum_{j=2}^{\infty} \tilde{\lambda}_{\gamma,j}^{\tau} \geq C_2 (\alpha\gamma)^{2\tau}, \tag{4.5}$$

where the last inequality follows from (4.1). Inequalities (4.4) and (4.5) combined give

$$\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha,\gamma,j}^{\tau} \geq 1 - C_1(\alpha\gamma)^2 + C_2(\alpha\gamma)^{2\tau} \geq 1 + (C_2/2)(\alpha\gamma)^{2\tau}$$

under the condition in (4.3) on small enough $\alpha\gamma$.

Consider the lower bound in the multivariate case $d > 1$ and define the set A by

$$A = \left\{ \ell : \alpha_\ell \gamma_\ell < \left(\frac{C_2}{2C_1} \right)^{\frac{1}{2(1-\tau)}} \right\}.$$

Then

$$\sup_{d\in\mathbb{N}} \left(\sum_{j=1}^{\infty} \nu_{d,j}^{\tau} \right) = \prod_{\ell=1}^{\infty} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}^{\tau} \right) = \prod_{\ell\in A} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}^{\tau} \right) \prod_{\ell\in\mathbb{N}\setminus A} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}^{\tau} \right).$$

We want to show that this supremum is infinite for $\tau < 1/(2\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}))$. We do this by proving that the first product on the right is infinite. Indeed for $\tau < 1/(2\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}))$,

$$\prod_{\ell\in A} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha_\ell,\gamma_\ell,j}^{\tau} \right) \geq \prod_{\ell\in A} \left(1 + C_{\mathrm{L}} (\alpha_\ell \gamma_\ell)^{2\tau} \right) \geq 1 + C_{\mathrm{L}} \sum_{\ell\in A} (\alpha_\ell \gamma_\ell)^{2\tau} = \infty.$$

This shows that $p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}}) \geq 1/\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$, which establishes the formula for $p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}})$. The estimates on $e_d(n; \Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}})$ and $n_d(\varepsilon; \Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{abs}})$ follow from the definition of strong polynomial tractability.

Finally, the exponent of strong polynomial tractability is 2 for the isotropic kernel because $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$ in this case. To prove that strong polynomial tractability is equivalent to polynomial tractability, it is enough to show that polynomial tractability implies strong polynomial tractability. From [25, Theorem 5.1] we know that polynomial tractability holds if and only if there exist numbers $c_1 > 0$, $q_1 \geq 0$, $q_2 \geq 0$

and $\tau > 0$ such that

$$c_2 := \sup_{d\in\mathbb{N}} \left\{ d^{-q_2} \left(\sum_{j=\lceil C_1\, d^{\,q_1}\rceil}^{\infty} \lambda_{d,j}^{\tau} \right)^{1/\tau} \right\} < \infty.$$

If so, then

$$n_d(\varepsilon; \Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm abs}) \le (c_1 + c_2^{\tau})\, d^{\max(q_1, q_2\tau)}\, \varepsilon^{-2\tau}$$

for all $\varepsilon \in (0,1)$ and $d \in \mathbb{N}$. Note that for all d we have

$$d^{-q_2\tau} \left(\sum_{j=1}^{\infty} \tilde{\nu}_{\alpha,\gamma,j}^{\tau} \right)^{d} - d^{-q_2\tau}(\lceil c_1 \rceil - 1)\tilde{\nu}_{\alpha,\gamma,1}^{\tau d} \le c_2^{\tau} < \infty.$$

This implies that $\tau \ge 1$. On the other hand, for $\tau = 1$ we can take $q_1 = q_2 = 0$ and arbitrarily small C_1, and obtain strong polynomial tractability. This completes the proof. □

Theorem 3 states that the exponent of strong polynomial tractability is at most 2, while for all shape parameters for which $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 1/2$ the exponent is smaller than 2. Again, although the rate of convergence of $e_d(n; \Xi^{\rm wor}, \Lambda^{\rm all})$ is always excellent, the dependence on d is eliminated only at the expense of the exponent which must be roughly $1/p(\Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm abs})$. Of course, if we take an exponentially decaying sequence of the products of scale parameters and shape parameters, say, $\alpha_\ell \gamma_\ell = q^\ell$ for some $q \in (0,1)$, then $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \infty$ and $p(\Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm abs}) = 0$. In this case, we have an excellent rate of convergence without any dependence on d.

4.1.2 Only Function Values. The tractability results for the class $\Lambda^{\rm std}$ are stated in the following theorem.

Theorem 4. *Consider under the absolute error criterion the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces associated with the symmetric positive definite product kernels* (1.6) *satisfying* (2.3) *for the class $\Lambda^{\rm std}$ in the worst case setting. If $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = 0$ or there exist constants $C_1, C_2, C_3 > 0$ such that* (3.2) *and* (4.1) *hold for all $0 < \gamma < \sup\{\gamma_\ell : \ell \in \mathbb{N}\}$, then*

- *$\mathcal{I}$ is strongly polynomially tractable with exponent of strong polynomial tractability at most 4. For all $d \in \mathbb{N}$ and $\varepsilon \in (0,1)$ we have*

$$e_d(n; \Xi^{\rm wor}, \Lambda^{\rm std}) \le \frac{\sqrt{2}}{n^{1/4}} \left(1 + \frac{1}{2\sqrt{n}}\right)^{1/2},$$
$$n_d(\varepsilon; \Xi^{\rm wor}, \Lambda^{\rm std}, \Psi^{\rm abs}) \le \left\lceil \frac{(1+\sqrt{1+\varepsilon^2})^2}{\varepsilon^4} \right\rceil.$$

- *For the isotropic kernel with $\alpha_\ell = \alpha$ and $\gamma_\ell = \gamma$ for all ℓ, the exponent of strong tractability is at least 2 and strong polynomial tractability is equivalent to polynomial tractability.*

Furthermore if $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 1/2$, then

- *$\mathcal{I}$ is strongly polynomially tractable with exponent of strong polynomial tractability*

$$\begin{aligned} p(\Xi^{\rm wor}, \Lambda^{\rm std}, \Psi^{\rm abs}) &\le \frac{1}{\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})} + \frac{1}{2\tilde{r}^2(\boldsymbol{\alpha}, \boldsymbol{\gamma})} \\ &= p(\Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm abs}) + \frac{1}{2} p^2(\Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm abs}) < 4. \end{aligned}$$

- *For all $d \in \mathbb{N}$ we have*

$$e_d(n; \Xi^{\rm wor}, \Lambda^{\rm std}) \preceq n^{-1/p(\Xi^{\rm wor}, \Lambda^{\rm std}, \Psi^{\rm abs})} = n^{-\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})/[1+1/(2\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}))]} \quad \text{as} \quad n \to \infty,$$
$$n_d(\varepsilon; \Xi^{\rm wor}, \Lambda^{\rm std}, \Psi^{\rm abs}) \preceq \varepsilon^{-p(\Xi^{\rm wor}, \Lambda^{\rm std}, \Psi^{\rm abs})} \quad \text{as} \quad \varepsilon \to 0.$$

Proof. The same proofs as for [10, Theorem 5.3 and Theorem 5.4] can be used. □

This theorem implies that for large $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$, the exponents of strong polynomial tractability are nearly the same for both classes $\Lambda^{\rm all}$ and $\Lambda^{\rm std}$. For an exponentially decaying sequence of shape parameters, say, $\alpha_\ell \gamma_\ell = q^\ell$ for some $q \in (0,1)$, we have $p(\Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm abs}) = p(\Xi^{\rm wor}, \Lambda^{\rm std}, \Psi^{\rm abs}) = 0$, and the rates of convergence are excellent and independent of d.

4.2 Tractability Under the Normalized Error Criterion

We now consider the function approximation problem for Hilbert spaces $\mathcal{F}_d$ with product kernels (1.6) under the normalized error criterion. That is, we want to find the smallest n for which

$$e_d(n;\Xi^{\rm wor},\Lambda) \le \varepsilon \, \|I_d\|, \quad \Lambda \in \{\Lambda^{\rm all},\Lambda^{\rm std}\}.$$

Note that $\|I_d\| = \sqrt{\nu_{d,1}} \le 1$ and it can be exponentially small in d. Therefore the normalized error criterion may be much harder than the absolute error criterion. It follows from [10, Theorem 6.1] that for the normalized error criterion, lack of polynomial tractability holds for the isotropic kernel for the class $\Lambda^{\rm all}$ and hence for the class $\Lambda^{\rm std}$.

4.2.1 Arbitrary Linear Functionals. We do not know whether polynomial tractability holds for kernels if $0 \le \tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}) < 1/2$. If $\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}) \ge 1/2$, we have the following theorem.

Theorem 5. *Consider under the normalized error criterion the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces associated with the symmetric positive definite product kernels* (1.6) *satisfying* (2.3) *for the class $\Lambda^{\rm all}$ in the worst case setting. If $\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma}) \ge 1/2$ and there exist constants $C_1, C_2, C_3 > 0$ such that* (3.2) *and* (4.1) *hold for all $0 < \gamma < \sup\{\gamma_\ell : \ell \in \mathbb{N}\}$, then*

- *$\mathcal{I}$ is strongly polynomially tractable with exponent of strong polynomial tractability*
$$p(\Xi^{\rm wor},\Lambda^{\rm all},\Psi^{\rm nor}) = \frac{1}{\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})}.$$
- *For all $d \in \mathbb{N}$ we have*
$$e_d(n;\Xi^{\rm wor},\Lambda^{\rm all}) \preceq \|I_d\| n^{-1/p(\Xi^{\rm wor},\Lambda^{\rm all},\Psi^{\rm nor})} = n^{-\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})} \quad \text{as} \quad n \to \infty,$$
$$n_d(\varepsilon;\Xi^{\rm wor},\Lambda^{\rm all},\Psi^{\rm nor}) \preceq \varepsilon^{-p(\Xi^{\rm wor},\Lambda^{\rm all},\Psi^{\rm nor})} \quad \text{as} \quad \varepsilon \to 0.$$

Proof. From [25, Theorem 5.2] we know that strong polynomial tractability holds if and only if there exits a positive number τ such that

$$c_2 := \sup_d \sum_{j=1}^{\infty} \left(\frac{\nu_{d,j}}{\nu_{d,1}}\right)^{\tau} = \sup_d \left\{\frac{1}{\nu_{d,1}^{\tau}} \sum_{j=1}^{\infty} \nu_{d,j}^{\tau}\right\} < \infty.$$

If so, then $n_d(\varepsilon; \Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm nor}) \le c_2\, \varepsilon^{-2\tau}$ for all $\varepsilon \in (0,1)$ and $d \in \mathbb{N}$, and the exponent of strong polynomial tractability is the infimum of 2τ for which $c_2 < \infty$.

For all $d \in \mathbb{N}$, we have $\sum_{j=1}^{\infty} \nu_{d,j}^{\tau} < \infty$ for $\tau = 1/(2\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}))$ from (3.10). It remains to note that $\sup_d\{1/\nu_{d,1}^{\tau}\} < \infty$ if and only if $\sup_d\{1/\nu_{d,1}\} < \infty$. Furthermore note that (3.6) implies that

$$\sup_d \left\{\frac{1}{\nu_{d,1}}\right\} \le \prod_{\ell=1}^{\infty} \frac{1}{1 - C_1(\alpha_\ell \gamma_\ell)^2}.$$

Clearly, $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \ge 1/2$ implies that $\sum_{\ell=1}^{\infty} (\alpha_\ell \gamma_\ell)^2 < \infty$, which yields $c_2 < \infty$. This also proves that $p(\Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm nor}) \le 1/\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$. The estimates on $e_d(n; \Xi^{\rm wor}, \Lambda^{\rm all})$ and $n_d(\varepsilon; \Xi^{\rm wor}, \Lambda^{\rm all}, \Psi^{\rm nor})$ follow from the definition of strong polynomial tractability. □

4.2.2 Only Function Values. We now turn to the class $\Lambda^{\rm std}$. We do not know if polynomial tractability holds for the class $\Lambda^{\rm std}$ if $0 \le \tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \le 1/2$. If $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 1/2$, we have the following theorem.

Theorem 6. *Consider under the normalized error criterion the function approximation problem $\mathcal{I} = \{I_d\}_{d \in \mathbb{N}}$ for Hilbert spaces associated with the symmetric positive definite product kernels* (1.6) *for the class $\Lambda^{\rm std}$ in the worst case setting. If $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 1/2$ and there exist constants $C_1, C_2, C_3 > 0$ such that* (3.2) *and* (4.1) *hold for all $0 < \gamma < \sup\{\gamma_\ell : \ell \in \mathbb{N}\}$, then*

- *$\mathcal{I}$ is strongly polynomially tractable with exponent of strong polynomial tract-*

ability

$$p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{std}}, \Psi^{\mathrm{nor}}) = \frac{1}{\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})} + \frac{1}{2\tilde{r}^2(\boldsymbol{\alpha}, \boldsymbol{\gamma})}$$
$$= p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{nor}}) + \frac{1}{2}p^2(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{all}}, \Psi^{\mathrm{nor}}) < 4.$$

For all $d \in \mathbb{N}$ we have

$$\begin{aligned} e_d(n; \Xi^{\mathrm{wor}}, \Lambda^{\mathrm{std}}) &\preceq n^{-1/p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{std}}, \Psi^{\mathrm{nor}})} \quad n \to \infty, \\ n_d(\varepsilon; \Xi^{\mathrm{wor}}, \Lambda^{\mathrm{std}}, \Psi^{\mathrm{nor}}) &\preceq \varepsilon^{-p(\Xi^{\mathrm{wor}}, \Lambda^{\mathrm{std}}, \Psi^{\mathrm{nor}})} \quad \varepsilon \to 0. \end{aligned}$$

Proof. The initial error is

$$\|I_d\| \geq \prod_{\ell=1}^{d} (1 - C_1(\alpha_\ell \gamma_\ell)^2)^{1/2} = \exp\left(\mathcal{O}(1) - \frac{1}{2} \sum_{\ell=1}^{d} (\alpha_\ell \gamma_\ell)^2 \right).$$

$\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 1/2$ implies that $\|I_d\|$ is uniformly bounded from below by a positive number. This shows that there is no difference between the absolute and normalized error criteria. This means that we can apply Theorem 4 for the class Λ^{std} with ε replaced by $\varepsilon\|I_d\| = \Theta(\varepsilon)$. This completes the proof. □

CHAPTER 5

TRACTABILITY IN THE AVERAGE CASE SETTING

We now present tractability results for the product kernels (1.6) in the average case setting. Note that when the unit trace condition (2.4) holds, the absolute and normalized error criteria are the same.

Although the tail sum of the ordered eigenvalues in (2.12) is often not directly accessible, it is shown in [9] that it can be bounded above by $CM_{d,\tau}$, where C is some positive constant, $0 < \tau \leq 1$ and

$$M_{d,\tau} := \left(\sum_{j=1}^{\infty} \nu_{d,j}^{\tau} \right)^{1/\tau}.$$

Note that $M_{d,1} = 1$ and by Jensen's inequality $M_{d,\tau} \geq M_{d,1} = 1$. Furthermore, $M_{d,\tau} = 1$ for $0 < \tau < 1$ if and only if $\nu_{d,1} = 1$ and $\nu_{d,j} = 0$ for all $j \geq 2$. Some lower bounds of $M_{d,\tau}$ are given in [16] and [15]. Thus the convergence rate in the average case setting depends on the finiteness of $M_{d,\tau}$, and strong tractability or dimension-independent convergence rates depends on the boundedness of $M_{d,\tau}$ over all d. This is summarized in the following lemma, which is a special case of [25, Theorem 6.1 and Theorem 6.2] for $\Lambda = \Lambda^{\mathrm{all}}$ and uses [15] for $\Lambda = \Lambda^{\mathrm{std}}$.

Lemma 2. *Consider the function approximation problem for Banach spaces associated with the symmetric positive definite covariance kernel* (1.6) *satisfying* (2.3) *for the class* Λ^{all} *and* Λ^{std} *in the average case setting. The function approximation problem has a dimension-independent convergence rate of* $\mathcal{O}(n^{-p})$ *provided that* $M_{d,2p+1}$ *is finite. There is dimension-independent convergence and strong polynomial tractability if and only if there exists* $\tau > 0$ *such that*

$$\sup_{d \in \mathbb{N}} M_{d,2\tau+1} < \infty.$$

In this case the exponents are

$$p(\Xi^{\mathrm{avg}}, \Lambda, \Psi) = \inf\left\{\tau > 0 : \sup_{d\in\mathbb{N}} M_{d,2\tau+1} < \infty\right\}, \quad \Lambda \in \{\Lambda^{\mathrm{all}}, \Lambda^{\mathrm{std}}\}, \quad \Psi \in \{\Psi^{\mathrm{abs}}, \Psi^{\mathrm{nor}}\}.$$

It is easy to see that dimension-independent convergence or strong polynomial tractability does not hold for $\widetilde{K}_d$ in the isotropic case with $\alpha_\ell = \alpha$ and $\gamma_\ell = \gamma$ for all $1 \le \ell \le d$, unless $\tilde{\lambda}_{\gamma,1} = 1$ and $\tilde{\lambda}_{\gamma,j} = 0$ for all $j \ge 2$. In fact one can show that the minimal number $n_d(\varepsilon; \Xi^{\mathrm{avg}}, \Lambda^{\mathrm{all}}, \Psi)$ depends exponentially on d, which is known as the curse of dimensionality.

Nevertheless when $\alpha_\ell\gamma_\ell$ decays quickly enough as ℓ increases, it is possible to achieve dimension-independent convergence and strong polynomial tractability under certain conditions. This is illustrated in the following theorem.

Theorem 7. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for separable Banach spaces associated with the symmetric positive definite covariance kernel* (1.6) *satisfying* (2.3) *for the class Λ^{all} and Λ^{std} in the average case setting. If $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 0$ and there exist constants $C_1, C_2, C_3 > 0$ such that* (3.2) *and* (4.1) *are satisfied for all $0 < \gamma < \sup\{\gamma_\ell : \ell \in \mathbb{N}\}$, then for $\Lambda \in \{\Lambda^{\mathrm{all}}, \Lambda^{\mathrm{std}}\}$ and $\Psi \in \{\Psi^{\mathrm{abs}}, \Psi^{\mathrm{nor}}\}$,*

- *$\mathcal{I}$ is strongly polynomially tractable if and only if $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 1/2$. In this case the exponent of strong polynomial tractability is*

$$p(\Xi^{\mathrm{avg}}, \Lambda, \Psi) = 1/\left(\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) - 1/2\right).$$

- *For all $d \in \mathbb{N}$ we have*

$$e_d(n; \Xi^{\mathrm{avg}}, \Lambda) \preceq n^{-1/p(\Xi^{\mathrm{avg}},\Lambda,\Psi)} = n^{1/2-\tilde{r}(\boldsymbol{\alpha},\boldsymbol{\gamma})} \quad as \quad n \to \infty,$$
$$n_d(\varepsilon; \Xi^{\mathrm{avg}}, \Lambda, \Psi) \preceq \varepsilon^{-p(\Xi^{\mathrm{avg}},\Lambda,\Psi)} \quad as \quad \varepsilon \to 0.$$

For the class Λ^{all}, the algorithm (2.9)*, which attains these exponents is given by projecting the function into the linear space spanned by the first n eigenfunctions. For the class Λ^{std}, the algorithm that attains these exponents is not known explicitly.*

Proof. If $\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) > 1/2$, then it is shown in the proof of Theorem 3 that

$$\sup_{d \in \mathbb{N}} M_{d,2\tau+1} < \infty$$

for any $0 < \tau < \tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) - 1/2$.

If $0 < \tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \leq 1/2$, then

$$\frac{1}{1+2\tau} < 1 \leq \frac{1}{2\tilde{r}(\boldsymbol{\alpha}, \boldsymbol{\gamma})}$$

for any $\tau > 0$. In this case it is shown in the proof of Theorem 3 that

$$\sup_{d \in \mathbb{N}} M_{d,2\tau+1} = \infty.$$

Applying Lemma 2 completes the proof. □

Theorem 7 states that dimension-independent convergence and strong polynomial tractability can be achieved for the product kernels and the exponents are the same for the class Λ^{all} and Λ^{std}. While the algorithm that attains these exponents is known for the class Λ^{all}, it is not yet known explicitly for the class Λ^{std}.

CHAPTER 6

PRACTICAL ISSUES OF FUNCTION APPROXIMATION WITH KERNEL METHODS

6.1 Introduction

In this chapter we discuss some practical issues of function approximation with kernel methods.

The tractability theory developed in the previous two chapters relies heavily on the assumptions made on the underlying function spaces, on the reproducing kernels or the covariance kernels, and definitions of various error criteria. In many practical applications one or more assumptions may be violated. However this does not mean that the tractability theory is useless under such circumstances. We present in Section 6.2 two numerical examples which show that strong polynomial tractability still holds even if some assumptions are relaxed to some extent.

To make the kernel approximation methods more accessible for users without detailed knowledge of the mathematics behind them, we also provide in Section 6.3 automatic algorithms that only require users to input the function to be approximated and the error tolerance. The algorithms we design have the merit that they guarantee that the user-specified error tolerance is met, every time, as long as the function to be approximated satisfies the cone condition.

6.2 Numerical Examples of Tractability With Product Kernels

We have seen in the previous two chapters that exponents of strong polynomial tractability for the class Λ^{all} are typically larger than those for the class Λ^{std}. Moreover the optimal algorithm that attains these exponents is only known for the class Λ^{all} in the average case setting. Unfortunately we do not have access to the information from the class Λ^{all} in most real-world applications. Nevertheless in this section we

present numerical evidence that indicates that dimension-independent convergence and strong polynomial tractability can be achieved with a product Gaussian kernel using function evaluation information.

We stress that the following numerical example serves only as evidence to support the theory developed in the previous chapters. Since not all assumptions of the theorems are satisfied and the errors used to compute the convergence rates are not exactly calculated according to the worst case or average case error definitions, the numerical example does not guarantee that the algorithm used in the example actually yields a dimension-independent convergence rate.

In this example we interpolate functions in a unit cube $[0,1]^d$ with the following product kernel

$$\widetilde{K}_d(\boldsymbol{x},\boldsymbol{t}) = \prod_{\ell=1}^{d} \left(1 - \alpha_\ell^2 + \alpha_\ell^2 \mathrm{e}^{-\gamma_\ell^2 (x_\ell - t_\ell)^2}\right),$$

where $\alpha_\ell = \gamma_\ell = \ell^{-1}$. We are given function value information $\boldsymbol{y} = \{y_i\}_{i=1}^n$ at the first n points of a d-dimensional scrambled Sobol sequence $\{\boldsymbol{x}_i\}_{i=1}^n$, where $y_i = f_d(\boldsymbol{x}_i)$. The interpolation problem is a special case of the approximation problem.

The test functions are constructed as a linear combination of the kernel functions centered at the first 7 points of another d-dimensional scrambled Sobol sequence $\{\boldsymbol{t}_i\}_{i=1}^7$, i.e.,

$$f_d(\boldsymbol{x}) = \sum_{i=1}^{7} a_i \widetilde{K}_d(\boldsymbol{x},\boldsymbol{t}_i).$$

Note that this Sobol sequence is independent from the one from which the function data are drawn. To properly choose the coefficients $\boldsymbol{a} := \{a_i\}_{i=1}^7$, we require f to satisfy the discrete versions of (2.1) and (2.2), namely

$$\mathbb{E}\{f_d(\boldsymbol{t}_i)\} = 0, \quad \text{and} \quad \mathbb{E}\{f_d(\boldsymbol{t}_i), f_d(\boldsymbol{t}_j)\} = \widetilde{K}_d(\boldsymbol{t}_i,\boldsymbol{t}_j), \quad i,j = 1,2,\ldots,7. \tag{6.1}$$

Let $\widehat{\mathsf{K}} := \left(\widetilde{K}_d(\boldsymbol{t}_i,\boldsymbol{t}_j)\right)_{i,j=1}^7$ and $\widehat{\mathsf{K}}^{-1} = LL^T$ be the Cholesky decomposition of $\widehat{\mathsf{K}}^{-1}$. We can choose $\boldsymbol{a} = L\boldsymbol{z}$, where $\boldsymbol{z} = \{z_i\}_{i=1}^7$ is a vector of 7 independent identically

distributed standard normal random variables. It can be easily verified that (6.1) is satisfied. Indeed,

$$\begin{aligned}\mathbb{E}\{f_d(\boldsymbol{t}_i)\} &= \sum_{j=1}^{7} \widetilde{K}_d(\boldsymbol{t}_i, \boldsymbol{t}_j)\mathbb{E}(a_j) = 0,\\ \mathbb{E}\{f_d(\boldsymbol{t}_i), f_d(\boldsymbol{t}_j)\} &= \widetilde{K}_d(\boldsymbol{t}_i, \boldsymbol{t}_j) = \sum_{k,l=1}^{7} \widetilde{K}_d(\boldsymbol{t}_i, \boldsymbol{t}_k)\widetilde{K}_d(\boldsymbol{t}_j, \boldsymbol{t}_l)\mathrm{cov}(a_k, a_l)\\ &= \sum_{k,l=1}^{7} \widehat{\mathsf{K}}_{ik}\widehat{\mathsf{K}}_{jl}(\widehat{\mathsf{K}}^{-1})_{kl} = \widehat{\mathsf{K}}_{ij} = \widetilde{K}_d(\boldsymbol{t}_i, \boldsymbol{t}_j),\end{aligned}$$

where we have used the fact that

$$\mathrm{cov}(\boldsymbol{a}) = \mathrm{cov}(L\boldsymbol{z}) = L\mathrm{cov}(\boldsymbol{z})L^T = \widehat{\mathsf{K}}^{-1}.$$

Let $\widetilde{\mathsf{K}} := \left(\widetilde{K}_d(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)_{i,j=1}^{n}$. The interpolant $\tilde{f}$ is a linear combination of the kernel functions centered at those n data points,

$$\tilde{f}_d(\boldsymbol{x}) = \sum_{i=1}^{n} c_i \widetilde{K}_d(\boldsymbol{x}, \boldsymbol{x}_i),$$

where $\boldsymbol{c} := \{c_i\}_{i=1}^{n}$ is the solution of the linear system

$$\widetilde{\mathsf{K}}\boldsymbol{c} = \boldsymbol{y}.$$

We construct 10 test functions from 10 independent random vectors that have the same distribution as $\boldsymbol{a}$. We obtain the interpolants for $d = 1, 2, \ldots, 10$ and compute the root mean squared errors (RMSE) at the first 1000 points of a third scrambled Sobol sequence for each of the test functions. For the worst case interpolation, we take the maximum of the RMSE and plot them against the number of data points for each dimension in Figure 6.1. It appears that the convergence rate approaches the theoretical dimension-independent convergence rate as d increases. Similarly for the average case interpolation, we take the average from the these RMSE and plot them against the number of data points for each dimension in Figure 6.2. We obtain

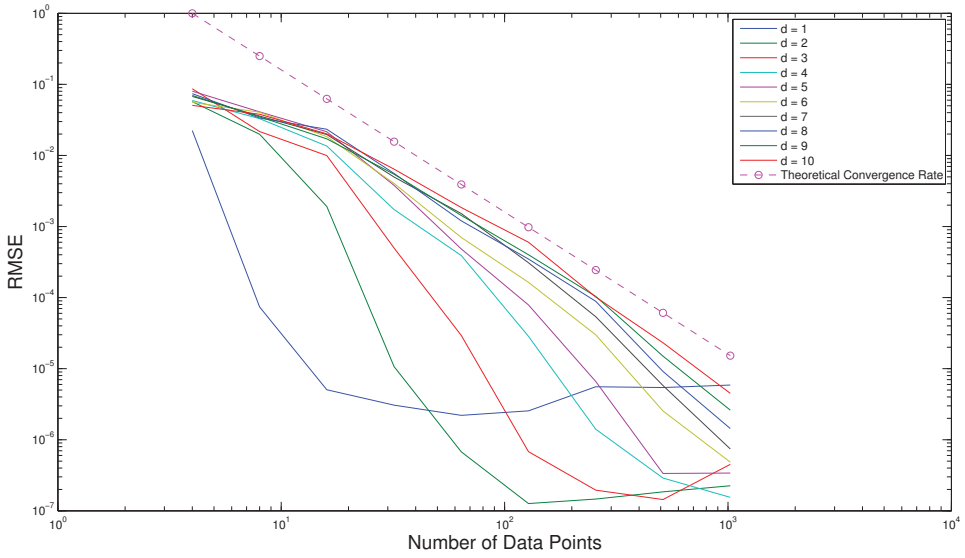

Figure 6.1. Worst case interpolation with the product Gaussian kernel

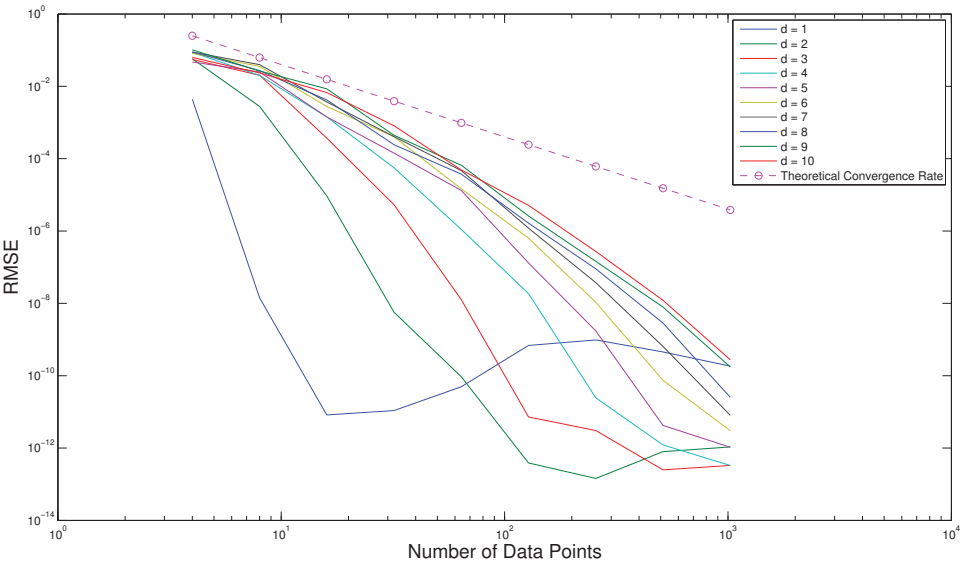

Figure 6.2. Average case interpolation with the product Gaussian kernel

a slightly better convergence rate as expected. Note that numerical instability are observed as n increases in both figures especially when d is small.

6.3 Algorithms That Guarantee Errors Within a Tolerance

Many numerical algorithms attempt to provide approximate solutions that differ from exact solutions by no more than a user-specified error tolerance. Although such algorithms are widely used in practice, most lack guarantees, i.e., conditions on inputs that ensure that the error tolerance is met.

One approach to construct guaranteed automatic algorithms is based on assumptions that the input functions lie in a cone. The Guaranteed Automatic Integration Library [5] contains guaranteed automatic adaptive algorithms for numerical integration (see [6]), for function approximation (see [6]), for optimization (see [35]), and for Monte Carlo sampling (see [14, 20, 13, 19]). Following this approach, we introduce a guaranteed automatic algorithm with kernel methods that ensures that the user-specified error tolerance is met.

The problem studied in this section is the scattered data interpolation problem in which we want to interpolate a function f that lies in a reproducing kernel Hilbert space $\mathcal{H}$ associated with a symmetric positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is assumed to be continuously embedded in $\mathcal{L}_2 = \mathcal{L}_2(\mathcal{X}, \varrho)$, where ϱ is some probability density function. We are given the function value information $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, where $y_i = f(\boldsymbol{x}_i)$.

6.3.1 Spline Algorithm. The spline algorithm to solve the scattered data interpolation problem is to construct the interpolant as

$$(A_n f)(\boldsymbol{x}) = \sum_{i=1}^{n} c_i K(\boldsymbol{x}, \boldsymbol{x}_i) = \boldsymbol{c}^T \boldsymbol{k}(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x})\boldsymbol{c}, \tag{6.2}$$

where $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^T$, $\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{x}, \boldsymbol{x}_1), K(\boldsymbol{x}, \boldsymbol{x}_2), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n))^T$. The inter-

polation condition, i.e.,

$$(A_n f)(\boldsymbol{x}_i) = y_i, \quad i = 1, 2, \ldots, n,$$

requires choosing $\boldsymbol{c}$ as

$$\boldsymbol{c} = \mathsf{K}^{-1}\boldsymbol{y}, \tag{6.3}$$

where $\mathsf{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1}^n$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. Note that K is a symmetric positive definite matrix because K is a symmetric positive definite kernel.

An equivalent way to write this is

$$(A_n f)(\boldsymbol{x}) = \boldsymbol{w}^T(\boldsymbol{x})\boldsymbol{y},$$

where the *cardinal functions* w_i are given by

$$\boldsymbol{w}(\boldsymbol{x}) = \mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}).$$

Note that in this form, it is obvious that $A_n f$ is linear in the function data $\boldsymbol{y}$. Thus, the interpolant may also be written as

$$(A_n f)(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x})\mathsf{K}^{-1}\boldsymbol{y}.$$

The spline algorithm depends on the function data, the design, and the kernel function.

It is recognized that this is a simplified version of spline algorithms. There are often low order polynomials included besides the nonparametric kernel part. However, the simplified framework here is sufficient for posing questions of interest.

6.3.2 Optimality From a Deterministic Perspective. We illustrate a couple of optimality properties of $A_n f$ in this subsection (see [8] and [39] for a detailed discussion of optimality using kernel methods). The spline, $A_n f$, has the property that

$$A_n f = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \{\|g\|_{\mathcal{H}} : g(\boldsymbol{x}_i) = y_i,\ i = 1, 2, \ldots, n\}.$$

Furthermore, if $L : \mathcal{H} \to \mathbb{R}$ is any bounded linear functional, then

$$L(A_n f) = \operatorname*{argmin}_{z \in \mathbb{R}} \sup_{g \in \mathcal{H}} \{|L(g) - z| : g(\boldsymbol{x}_i) = y_i,\ i = 1, 2, \ldots, n, \text{ and } \|g\|_{\mathcal{H}} \leq V\},$$

where the constant $V \geq \|A_n f\|_{\mathcal{H}}$. This is an interesting property for, say, $L(f) = f(\boldsymbol{x})$.

We are ready to derive the first error bound by expressing $L(f)$ in terms of the reproducing kernel and applying the Cauchy-Schwartz inequality. The following lemma is a special case of [24, Lemma 5].

Lemma 3. *Let the spline algorithm A_n be given by* (6.2) *with $\boldsymbol{c}$ chosen as* (6.3)*. Then the following error bound holds for all $f \in \mathcal{H}$ and all continuous linear functionals L:*

$$|L(f) - L(A_n f)| \leq \Phi_L \|f\|_{\mathcal{H}},$$

where

$$\Phi_L = \left(L^{\cdot} L^{\cdot\cdot} K(\cdot, \cdot\cdot) - \boldsymbol{\eta}^T \mathsf{K}^{-1} \boldsymbol{\eta}\right)^{1/2}$$

and $\boldsymbol{\eta} = (L(K(\cdot, \boldsymbol{x}_1)), L(K(\cdot, \boldsymbol{x}_2)), \ldots, L(K(\cdot, \boldsymbol{x}_n))^T$. Here $L^{\cdot} L^{\cdot\cdot} K(\cdot, \cdot\cdot)$ means that we first apply L to the second argument of K first to obtain a function of the first argument of K, and then apply L again to this function.

Proof. Let $\eta \in \mathcal{H}$ be the representer of L, i.e.,

$$L(f) = \langle \eta, f \rangle_{\mathcal{H}} \quad \text{for all} \quad f \in \mathcal{H}.$$

It then follows that

$$\eta(\boldsymbol{x}) = \langle K(\cdot, \boldsymbol{x}), \eta \rangle_{\mathcal{H}} = L(K(\cdot, \boldsymbol{x})).$$

We can bound the error from above by

$$
\begin{aligned}
&|L(f) - L(A_n f)|^2 \\
&= \left| \left\langle \eta - \sum_{i=1}^{n} L(w_i) K(\cdot, \boldsymbol{x}_i), f \right\rangle_{\mathcal{H}} \right|^2 \leq \left\| \eta - \sum_{i=1}^{n} L(w_i) K(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}}^2 \|f\|_{\mathcal{H}}^2. \\
&= \left(\|\eta\|_{\mathcal{H}}^2 - 2 \left\langle \eta, \sum_{i=1}^{n} L(w_i) K(\cdot, \boldsymbol{x}_i) \right\rangle_{\mathcal{H}} + \left\| \sum_{i=1}^{n} L(w_i) K(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}}^2 \right) \|f\|_{\mathcal{H}}^2. \qquad (6.4)
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
\|\eta\|_{\mathcal{H}}^2 &= L(\eta) = L^{\cdot} L^{\cdot\cdot} K(\cdot, \cdot\cdot), \qquad (6.5) \\
\left\langle \eta, \sum_{i=1}^{n} L(w_i) K(\cdot, \boldsymbol{x}_i) \right\rangle_{\mathcal{H}} &= \sum_{i=1}^{n} L(w_i) \left\langle \eta, K(\cdot, \boldsymbol{x}_i) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{n} L(w_i) \eta(\boldsymbol{x}_i) \\
&= \sum_{i=1}^{n} L(w_i) L(K(\cdot, \boldsymbol{x}_i) = \boldsymbol{\eta}^T \mathsf{K}^{-1} \boldsymbol{\eta}, \qquad (6.6) \\
\left\| \sum_{i=1}^{n} L(w_i) K(\cdot, \boldsymbol{x}_i) \right\|_{\mathcal{H}}^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle L(w_i) K(\cdot, \boldsymbol{x}_i), L(w_j) K(\cdot, \boldsymbol{x}_j) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} L(w_i) K(\boldsymbol{x}_i, \boldsymbol{x}_j) L(w_j) = \boldsymbol{\eta}^T \mathsf{K}^{-1} \boldsymbol{\eta}. \qquad (6.7)
\end{aligned}
$$

Substituting (6.5), (6.6) and (6.7) into (6.4) completes the proof. □

6.3.3 Error Estimation via Cones. Although Lemma (3) gives an error bound, we cannot compute $\|f\|_{\mathcal{H}}$ from n function values. To construct a guaranteed automatic algorithm, we need an error bound that can be computed from the data.

If we take $L(f) = f(\boldsymbol{x})$ and apply Lemma 3, we obtain a bound of the $\mathcal{L}_2$ error.

$$
\begin{aligned}
\|f - A_n f\|_{\mathcal{L}_2}^2 &= \int_{\mathcal{X}} |f(\boldsymbol{x}) - (A_n f)(\boldsymbol{x})|^2 \varrho(\boldsymbol{x}) \\
&\leq \|f\|_{\mathcal{H}}^2 \int_{\mathcal{X}} (K(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}^T(\boldsymbol{x}) \mathsf{K}^{-1} \boldsymbol{k}(\boldsymbol{x})) \varrho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = h^2(n) \|f\|_{\mathcal{H}}^2, \qquad (6.8)
\end{aligned}
$$

where

$$
h(n) := \left(\int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}) \varrho(\boldsymbol{x}) - \operatorname{tr} \left(\mathsf{K}^{-1} \widetilde{\mathsf{K}} \right) \right)^{1/2} \qquad (6.9)
$$

and $\widetilde{\mathsf{K}} := \int_{\mathcal{X}} \boldsymbol{k}(\boldsymbol{x})\boldsymbol{k}^T(\boldsymbol{x})\varrho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Now suppose that $T : \mathcal{H} \to \mathcal{L}_2$ is some linear operator. We assume that the linear functional $T_{\boldsymbol{x}}$, given by

$$T_{\boldsymbol{x}} f = (Tf)(\boldsymbol{x}), \tag{6.10}$$

is bounded for all $\boldsymbol{x} \in \mathcal{X}$. This implies that T is a bounded linear operator.

Let $\|\cdot\|_{\widetilde{\mathcal{H}}}$ be a weaker norm defined by $\|f\|_{\widetilde{\mathcal{H}}} := \|Tf\|_{\mathcal{L}_2}$. The algorithm for estimating the weaker norm is given by $\|A_n f\|_{\widetilde{\mathcal{H}}} = \|T(A_n f)\|_{\mathcal{L}_2}$. It follows that $(T(A_n f))(\boldsymbol{x}) = \boldsymbol{c}^T\boldsymbol{\xi}(\boldsymbol{x})$, where

$$\begin{aligned}\boldsymbol{\xi}(\boldsymbol{x}) =&(\xi(\boldsymbol{x},\boldsymbol{x}_1),\xi(\boldsymbol{x},\boldsymbol{x}_2),\ldots,\xi(\boldsymbol{x},\boldsymbol{x}_n))^T\\ :=&(T_{\boldsymbol{x}}K(\cdot,\boldsymbol{x}_1),T_{\boldsymbol{x}}K(\cdot,\boldsymbol{x}_2),\ldots,T_{\boldsymbol{x}}K(\cdot,\boldsymbol{x}_n))^T.\end{aligned}$$

Consequently, we have

$$\|A_n f\|^2_{\widetilde{\mathcal{H}}} = \|\boldsymbol{c}^T\boldsymbol{\xi}(\cdot)\|^2_{\mathcal{L}_2} = \boldsymbol{c}^T\widetilde{\mathsf{H}}\boldsymbol{c},$$

where $\widetilde{\mathsf{H}} := \int_{\mathcal{X}} \boldsymbol{\xi}(\boldsymbol{x})\boldsymbol{\xi}^T(\boldsymbol{x})\varrho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Consider the *cone* space $\mathcal{C}_\tau = \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} \le \tau\|g\|_{\widetilde{\mathcal{H}}}\}$ for $\tau > 0$. Note that if $f \in \mathcal{C}_\tau$, then $af \in \mathcal{C}_\tau$ for any $a > 0$, hence the term cone space. For functions in the space, the following theorem gives an error bound that can be computed from the function evaluation data.

Theorem 8. *Let the spline algorithm A_n be given by* (6.2) *with $\boldsymbol{c}$ chosen as* (6.3)*. Define $h(n)$ by* (6.9) *and let*

$$\tilde{h}(n) := \left(\int_{\mathcal{X}} \zeta(\boldsymbol{x},\boldsymbol{x})\varrho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \mathrm{tr}\left(\mathsf{K}^{-1}\widetilde{\mathsf{H}}\right)\right)^{1/2}$$

and $\zeta(\boldsymbol{x},\boldsymbol{t}) := T^{\cdot}_{\boldsymbol{x}}T^{\cdot\cdot}_{\boldsymbol{t}}K(\cdot,\cdot\cdot)$. If $T : \mathcal{H} \to \mathcal{L}_2$ is a linear operator such that the linear function $T_{\boldsymbol{x}}$ defined by (6.10) *is bounded for all $\boldsymbol{x} \in \mathcal{X}$, then the following error bound*

holds for all $f \in \mathcal{C}_\tau = \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} \leq \tau \|g\|_{\widetilde{\mathcal{H}}}\}$*:*

$$\|f - A_n f\|_{\mathcal{L}_2} \leq \frac{\tau h(n) \|A_n f\|_{\widetilde{\mathcal{H}}}}{1 - \tau \tilde{h}(n)},$$

provided that $\tilde{h}(n) < 1/\tau$*.*

Proof. Using (6.8) gives

$$\|f - A_n f\|_{\mathcal{L}_2}^2 = \int_{\mathcal{X}} |f(\boldsymbol{x}) - (A_n f)(\boldsymbol{x})|^2 \varrho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \leq \tilde{h}^2(n) \|f\|_{\mathcal{H}}^2.$$

It remains to show that

$$\|f\|_{\mathcal{H}} \leq \frac{\tau \|A_n f\|_{\widetilde{\mathcal{H}}}}{1 - \tau \tilde{h}(n)}. \tag{6.11}$$

We have

$$\begin{aligned} \left| \|f\|_{\widetilde{\mathcal{H}}} - \|A_n f\|_{\widetilde{\mathcal{H}}} \right|^2 &= \left| \|Tf\|_{\mathcal{L}_2} - \|T(A_n f)\|_{\mathcal{L}_2} \right|^2 \leq \|Tf - T(A_n f)\|_{\mathcal{L}_2}^2 \\ &= \int_{\mathcal{X}} ((Tf)(\boldsymbol{x}) - (T(A_n f))(\boldsymbol{x}))^2 \varrho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Applying Lemma 3 with $L = T_{\boldsymbol{x}}$ yields

$$\begin{aligned} \left| \|f\|_{\widetilde{\mathcal{H}}} - \|A_n f\|_{\widetilde{\mathcal{H}}} \right|^2 &\leq \|f\|_{\mathcal{H}}^2 \int_{\mathcal{X}} \Phi_{T_{\boldsymbol{x}}}^2 \varrho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \\ &= \|f\|_{\mathcal{H}}^2 \int_{\mathcal{X}} (\zeta(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{\xi}^T(\boldsymbol{x}) \mathsf{K}^{-1} \boldsymbol{\xi}(\boldsymbol{x})) \varrho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \tilde{h}^2(n) \|f\|_{\mathcal{H}}^2. \end{aligned}$$

Since $f \in \mathcal{C}_\tau$, it satisfies the cone condition $\|f\|_{\mathcal{H}} \leq \tau \|f\|_{\widetilde{\mathcal{H}}}$ and hence

$$\left| \|f\|_{\widetilde{\mathcal{H}}} - \|A_n f\|_{\widetilde{\mathcal{H}}} \right| \leq \tau \tilde{h}(n) \|f\|_{\mathcal{H}},$$

which implies (6.11) provided that $\tilde{h}(n) < 1/\tau$. □

We stress that $\|A_n f\|_{\widetilde{\mathcal{H}}}$ can be computed directly from the function data. The following guaranteed automatic algorithm is straightforward.

Algorithm 1. *Given* $\tau > 0$*,* $\varepsilon > 0$*, and* $f \in \mathcal{C}_\tau = \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} \leq \tau \|g\|_{\widetilde{\mathcal{H}}}\}$*, choose* n_1 *such that* $\tilde{h}(n_1) < 1/\tau$*.*

Stage 1: *Check if* $\tau h(n_i)\|A_{n_i}f\|_{\widetilde{\mathcal{H}}} \leq \varepsilon(1-\tau\tilde{h}(n_i))$. *If it is true, return the answer* $A_{n_i}f$.

Stage 2: *Set* $n_{i+1} = \lceil an_i \rceil$, *where* $a > 1$, *increment* i *by one, and return to Stage 1.*

Note that $\tilde{h}(n)$ is non-increasing and converges to zero as $n \to \infty$. Therefore we can always find n_1 such that $\tilde{h}(n_1) < 1/\tau$. The following theorem formulates the correctness of Algorithm 1.

Theorem 9. *Algorithm 1 will terminate successfully. Suppose it terminates after* k *steps, the following error bound holds:*

$$\|f - A_{n_k}f\|_{\mathcal{L}_2} \leq \varepsilon.$$

6.3.4 An Example of Error Estimation for Minimum Kernels via Cones.

Consider the univariate case and let

$$\mathcal{H} = \{f : f(0) = 0,\ f \text{ is absolutely continuous and } f' \in \mathcal{L}_2((0,1))\}.$$

Define the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ as

$$\langle f, g\rangle_{\mathcal{H}} = \frac{1}{\gamma}\int_0^1 f'(x)g'(x)\mathrm{d}x.$$

It can be easily verified that $\mathcal{H}$ is a reproducing kernel Hilbert space with the reproducing kernel

$$K(x,y) = \gamma\min(x,y).$$

Suppose we take T to be the embedding operator, i.e., $T : \mathcal{H} \to \mathcal{L}_2((0,1))$ is defined by $T(f) = f$. Then $\tilde{h} = h$, and hence the error bound in the previous section becomes

$$\left(\int_0^1 |f(x) - (A_nf)(x)|^2\mathrm{d}x\right)^{1/2} \leq \frac{\tau h(n)\|A_nf\|_{\widetilde{\mathcal{H}}}}{1-\tau h(n)}. \tag{6.12}$$

Let the test function $f \in \mathcal{H}$ be given by $f(x) = \sin(ax)$, $a \in \mathbb{R}$. We can compute that

$$\|f\|_{\mathcal{H}} = \left(\frac{1}{\gamma}\int_0^1 a^2\cos^2(ax)\mathrm{d}x\right)^{1/2} = a\sqrt{\frac{1}{\gamma}\left(\frac{1}{2}+\frac{\sin(2a)}{4a}\right)},$$

and that

$$\|f\|_{\widetilde{\mathcal{H}}} = \|f\|_{\mathcal{L}_2} = \left(\int_0^1 \sin^2(ax)\mathrm{d}x\right)^{1/2} = \sqrt{\frac{1}{2}-\frac{\sin(2a)}{4a}}.$$

We want $f \in \mathcal{C}_\tau$, or equivalently, $\|f\|_{\mathcal{H}} \le \tau\|f\|_{\widetilde{\mathcal{H}}}$, which implies that

$$a\sqrt{\frac{1}{\gamma}\left(\frac{1}{2}+\frac{\sin(2a)}{4a}\right)} \le \tau\sqrt{\frac{1}{2}-\frac{\sin(2a)}{4a}}.$$

Solving this inequality for τ, we have that

$$\tau \ge a\sqrt{\frac{2a+\sin(2a)}{\gamma(2a-\sin(2a))}}.$$

For example, if we take $\gamma = 1$ and $a = \pi$, then $\tau = \pi$ such that $f \in \mathcal{C}_\tau$. Once τ is chosen, we can proceed to compute the error bound in (6.12).

6.3.5 Error Estimation for Gaussian Kernels. In this subsection we consider the scattered data interpolation problem in a cube

$$\mathcal{X} = \mathcal{X}(\boldsymbol{x}_0, R) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}-\boldsymbol{x}_0\|_\infty \le R\} \tag{6.13}$$

with the isotropic Gaussian kernel

$$K(\boldsymbol{x},\boldsymbol{t}) = \exp\left(-\gamma^2\|\boldsymbol{x}-\boldsymbol{t}\|_2^2\right). \tag{6.14}$$

The fill distance h is defined by

$$h = h_{X,\mathcal{X}} := \sup_{\boldsymbol{x}\in\mathcal{X}} \min_{1\le i\le n} \|\boldsymbol{x}-\boldsymbol{x}_i\|_2, \tag{6.15}$$

where the design $X := \{\boldsymbol{x}_i\}_{i=1}^n$. It turns out that we can obtain a tighter error bound and a much faster convergence rate if h is bounded.

Let D^α be the αth order differential operator. The following theorem gives an error bound in $\mathcal{L}_p = \mathcal{L}_p(\mathcal{X}, \varrho)$ for the scattered data interpolation problem with Gaussian kernels.

Theorem 10. *Consider the scattered data interpolation problem in the cube $\mathcal{X}$ given by* (6.13)*. Let the spline algorithm A_n be given by* (6.2) *with $\boldsymbol{c}$ chosen as* (6.3) *and K being the Gaussian kernel* (6.14)*. The following error bound holds for all $f \in \mathcal{C}_\tau = \{g \in \mathcal{H} : D^\alpha g \in \mathcal{L}_p(\mathcal{X}), \|g\|_{\mathcal{H}} \le \tau \|D^\alpha g\|_{\mathcal{L}_p}\}$:*

$$\|f - A_n f\|_{\mathcal{L}_p} \le \frac{\tau \mathrm{e}^{c_1 \ln(h)/h}}{1 - \tau c_2 h} \|D(A_n f)\|_{\mathcal{L}_p},$$

where the constants c_1 and c_2 are known explicitly, provided that h is sufficiently small.

Proof. We use [39, Theorem 11.22] to obtain

$$\|f - A_n f\|_{\mathcal{L}_\infty} \le \mathrm{e}^{c_1 \ln(h)/h} \|f\|_{\mathcal{H}},$$

for $h \le c_0$, where the constant $c_0 := R/(3\lambda_d)$ and λ_d is defined by

$$\lambda_1 = 2, \quad \text{and} \quad \lambda_j = 2j(1 + \lambda_{j-1}), \quad \text{for} \quad j = 2, 3, \ldots.$$

The constant c_1 is given by

$$c_1 := \min(c_0/2, 1/(\mathrm{e}C_3))/2, \tag{6.16}$$

where $C_3 := C_1 \mathrm{e}^{C_2}$, $C_1 := 576\gamma^2 d\lambda_d^2$, and $C_2 := \ln(4) + 8d\lambda_d$. In fact we can obtain the same bound for the $\mathcal{L}_p$ norm for any $1 \le p < \infty$:

$$\|f - A_n f\|_{\mathcal{L}_p} = \left(\int_{\mathcal{X}} |f(\boldsymbol{x}) - (A_n f)(\boldsymbol{x})|^p \varrho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right)^{1/p} \le \mathrm{e}^{c_1 \ln(h)/h} \|f\|_{\mathcal{H}}. \tag{6.17}$$

Now let $1 \le p \le \infty$. Note that $\mathcal{X}$ is bounded and satisfies an interior cone condition with the angle $\theta = \arcsin(1/\sqrt{d})$ and the radius R. Therefore we can apply [39, Theorem 11.14] to obtain

$$\|D^\alpha f - D^\alpha (A_n f)\|_{\mathcal{L}_\infty} \le c_2 h \|f\|_{\mathcal{H}}, \tag{6.18}$$

when h is sufficiently small. Similarly we can obtain the same bound for the $\mathcal{L}_p$ norm:

$$\|D^\alpha f - D^\alpha(A_n f)\|_{\mathcal{L}_p} \le c_2 h \|f\|_{\mathcal{H}}.$$

It follows that

$$\begin{aligned}\big|\|D^\alpha f\|_{\mathcal{L}_p} - \|D^\alpha(A_n f)\|_{\mathcal{L}_p}\big| \le \|D^\alpha f - D^\alpha(A_n f)\|_{\mathcal{L}_p} &\le c_2 h \|f\|_{\mathcal{H}} \\ &\le c_2 h \tau \|D^\alpha f\|_{\mathcal{L}_p}, \end{aligned} \tag{6.19}$$

which implies that

$$\|f\|_{\mathcal{H}} \le \tau \|D^\alpha f\|_{\mathcal{L}_p} \le \frac{\tau}{1 - \tau c_2 h} \|D^\alpha(A_n f)\|_{\mathcal{L}_p}, \tag{6.20}$$

provided that $h < 1/(\tau c_2)$. Combining (6.17) and (6.19) completes the proof. □

In order to construct a guaranteed algorithm based on the error bound given in Theorem 10, we need to know explicitly the constants c_1 and c_2, as well as how small h needs to be for the error bound (6.20) to hold. The constant c_1 is given by (6.16). It is very tedious but possible to compute the constant c_2 when $d \ge 2$ following the proof of [39, Theorem 11.14]. In the case $\alpha = d = 1$, the constant c_2 is given by

$$c_2 = 2\gamma^2 C_5 C_4^3 (6 + C_5 C_4),$$

where $C_4 := 64(\sqrt{d} + 1)^2/3$ and $C_5 := \sqrt{d}(\sqrt{d} + 1)$, and (6.20) holds for $h \le \min(R, \sqrt{3/(2\gamma)})/C_4$.

The algorithm based on this error bound is given below.

Algorithm 2. *Given the Gaussian kernel* (6.14)*, the cube* (6.13) *with* $d = 1$*,* $\tau > 0$*,* $\varepsilon > 0$*, and* $f \in \mathcal{C}_\tau = \big\{g \in \mathcal{H} : Dg \in \mathcal{L}_p(\mathcal{X}), \|g\|_{\mathcal{H}} \le \tau \|Dg\|_{\mathcal{L}_p}\big\}$*, choose a design* X *with* n_1 *data sites such that*

$$h < \min\left(\frac{1}{\tau c_2}, c_0, \frac{R}{C_4}, \frac{\sqrt{3}}{C_4\sqrt{2\gamma}}\right).$$

Stage 1: *Check if* $\tau e^{c_1 \ln(h)/h} \|D(A_n f)\|_{\mathcal{L}_p} \leq \varepsilon(1 - \tau c_2 h)$. *If it is true, return the answer* $A_{n_i} f$.

Stage 2: *Set* $n_{i+1} = \lceil a n_i \rceil$, *where* $a > 1$, *to make the design* X *with* n_{i+1} *data sites have a smaller* h, *increment* i *by one, and return to Stage 1.*

We conclude the subsection with the following theorem.

Theorem 11. *Algorithm 2 will terminate successfully. Suppose it terminates after* k *steps, the following error bound holds:*

$$\|f - A_{n_k} f\|_{\mathcal{L}_p} \leq \varepsilon.$$

We remark that for many functions the required number of data points n is quite large to guarantee that the error tolerance is met. This could result in numerical instability. See Section 7.3 for a detailed discussion.

CHAPTER 7

FUTURE WORK

7.1 Banach Spaces in the Worst Case Setting

This dissertation works with reproducing kernel Hilbert spaces in the worst case setting. The existence of complete orthonormal bases in Hilbert spaces simplifies a lot of computation. A natural generalization to reproducing kernel Hilbert spaces is reproducing kernel Banach spaces introduced in [40]. One reason of this generalization mentioned by the authors is that many training data come with intrinsic structures that make them impossible to be embedded into a Hilbert space such as $\mathcal{L}_2$. The precise definition of reproducing kernel Banach spaces is given in [40, Definition 1] and [40, Theorem 2] establishes the uniqueness and existence of reproducing kernels.

It would be interesting to see whether dimension-independent convergence rates can be achieved in this more general setting. In the case that the reproducing kernel Banach space is not a Hilbert space, it is natural to assume that it is embedded in $\mathcal{L}_p = \mathcal{L}_p(\mathbb{R}^d, \varrho_d)$, where $p \neq 2$. If we study the $\mathcal{L}_p$ approximation error, we cannot utilize the kernel decomposition (1.4) given by Mercer's theorem the same way as in the Hilbert space case. The lack of orthogonality in Banach spaces makes the search of the best algorithm a challenging task, even for the class Λ^{all}.

7.2 Designs and Algorithms for the Class Λ^{std}

For the class Λ^{all}, it is known that the optimal design is given by (2.8). For this design it is known that the best approximation algorithm is given by (2.9). However the information class Λ^{all} is not as common as the class Λ^{std}, for which we do not know the optimal design. To see why the optimal design is a difficult problem for the

class Λ^{std}, consider the squared $\mathcal{L}_2$ approximation error in the average case

$$\epsilon^2(A;\Xi^{\mathrm{avg}}) = \int_{\mathcal{F}} \|f - Af\|_{\mathcal{L}_2}\, \mu(\mathrm{d}f),$$

where we have dropped the dependence on d for simplicity. We take A to be the spline algorithm (6.2) with $\boldsymbol{c}$ given by (6.3), and a straightforward computation shows that

$$\epsilon^2(A;\Xi^{\mathrm{avg}}) = \int_{\mathbb{R}^d} \left(K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}^T(\boldsymbol{x})\mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x})\right) \varrho_d(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}.$$

We can see that the $\mathcal{L}_2$ approximation error depends on the design through the term $\boldsymbol{k}^T(\boldsymbol{x})\mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x})$ and the maximization of this term is not an easy problem. This means that even though theoretical dimension-independent convergence rates can be achieved, we do not know which algorithms yield such convergence rates. The numerical examples in Section 6.2 indicate that a simple spline algorithm using a good design appears to produce dimension-independent convergence rates, but it is still desirable to build some related theories.

Although we know that the spline algorithm is optimal when the design is given, it is worth investigating other algorithms that make full use of certain designs such as low-discrepancy sequences and random sampling. After all it would be a significant advance if we can find a good algorithm with a good design, not necessarily the optimal one, that gives the theoretical dimension-independent convergence rate.

7.3 Numerical Stability of Guaranteed Algorithms with Kernel Methods

The guaranteed error bounds in Section 6.3 are based on simple spline algorithms which require us to solve

$$\mathsf{K}\boldsymbol{c} = \boldsymbol{y}, \tag{7.1}$$

where the kernel matrix $\mathsf{K} = (K(\boldsymbol{x}_i,\boldsymbol{x}_j))_{i,j=1}^n$ and the function value vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_n))^T$. If the kernel K is positive definite,

then K is a positive definite matrix and there exists a unique solution to (7.1). If K is the Gaussian kernel, we have a tight error bound. However, in order for the error bounds to hold, we need a design that has a sufficiently small fill distance, which can only be achieved by increasing n. However as n increases, K often becomes close to singular and the condition number of K becomes very large.

It would be beneficial to find a symmetric positive definite kernel which can mitigate this conditioning problem, but still yields a tight error bound. The class of kernels which have a compact support is of particular interest because K becomes sparse in this case. Error bounds for some popular kernels are given in [39, Section 11.3]. It would be interesting to see if one can design a cone space and derive the corresponding guaranteed error bound.

Another approach is to use more stable algorithms such as the Hilbert-Schmidt SVD introduced in [4] and [11] to approximate f by a linear combination of an alternate basis. However, there are currently no known error bounds of $\|f - A_n f\|_{\mathcal{L}_2}$ in terms of $\|f\|_{\mathcal{H}}$. When we derive the first error bound in Lemma 3, we make heavy use of the reproducing property $f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}}$, where we take $\boldsymbol{x} = \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ to obtain the basis functions on the right hand side. After the change of basis using the the Hilbert-Schmidt SVD, this method no longer works and alternate approaches are needed to derive an error bound.

7.4 Guaranteed Error Bound for the class Λ^{all}

In Section 6.3 the cone space is introduced to derive guaranteed error bounds for the class Λ^{std}. It is interesting to see what assumptions need to be made on the function so that we can develop guaranteed algorithms for the class Λ^{all}. We know that the $\mathcal{L}_2$ approximation error is minimized by the algorithm (2.9). Let us denote $\langle f, \varphi_{d,i}\rangle_{\mathcal{L}_2}$ by $\hat{f}_i$. Then f has a Fourier expansion of $f = \sum_{i=1}^{\infty} \hat{f}_i \varphi_{d,i}$, where $\hat{f}_i$ are

Fourier coefficients. The algorithm (2.9) becomes $A_n f = \sum_{i=1}^{n} \hat{f}_i \varphi_{d,i}$. Since $\{\varphi_{d,i}\}_{i=1}^{\infty}$ is an orthonormal basis in $\mathcal{L}_2$, the squared $\mathcal{L}_2$ approximation error for the function f is

$$\|f - A_n f\|_{\mathcal{L}_2}^2 = \sum_{i=n+1}^{\infty} |\hat{f}_i|^2.$$

The Fourier coefficients $\hat{f}_i$ for $i > n$ are not known when we have n pieces of data information, but we may assume that they are bounded above in terms of $\hat{f}_i$ for $i \leq n$. Suppose we want to develop a guaranteed algorithm that increases the number of functional samples until the error tolerance is met, i.e., it evaluates n_j functionals at the jth iteration and the sequence $\{n_j\}_{j \in \mathbb{N}}$ is increasing. If we assume that

$$\sum_{i=n_k+1}^{n_k} |\hat{f}_i|^2 \leq C(k,l) \sum_{i=n_l+1}^{n_l} |\hat{f}_i|^2, \quad \text{for all} \quad k \geq l,$$

where $C(k,l)$ is a positive number that depends on k and l, then the squared $\mathcal{L}_2$ approximation error can be bounded above by

$$\|f - A_{n_j} f\|_{\mathcal{L}_2}^2 = \sum_{i=n_j+1}^{\infty} |\hat{f}_i|^2 = \sum_{k=0}^{\infty} \sum_{i=n_{j+k}+1}^{n_{j+k+1}} |\hat{f}_i|^2 \leq \sum_{k=0}^{\infty} C(j+k, j-1) \sum_{i=n_{j-1}+1}^{n_j} |\hat{f}_i|^2.$$

We may obtain an upper bound for $\sum_{k=0}^{\infty} C(j+k, j-1)$ for all $j > 1$ by assuming that, for example, $C(k,l) = a^{k-l}$, where $0 < a < 1$. Then we have an error bound that can be computed from the linear functional information. For the class Λ^{std}, this approach may also yield an error bound if we can estimate the Fourier coefficients $\hat{f}_i$ via a fast Fourier transform, see [20].

BIBLIOGRAPHY

[1] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*, Kluwer Academic Publishers, Boston, 2004.

[2] D. S. Bernstein, *Matrix mathematics*, Princeton University Press, 41 William Street, Princeton, New Jersey 08540, 2008.

[3] M. D. Buhmann, *Radial basis functions*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2003.

[4] R. Cavoretto, G. E. Fasshauer, and M. McCourt, *An introduction to the Hilbert-Schmidt SVD using iterated brownian bridge kernels*, Numerical Algorithms, vol. 68-2, 393–422, 2014.

[5] S.-C. T. Choi, Y. Ding, F. J. Hickernell, L. Jiang, Ll. A. Jiménez Rugama, X. Tong, Y. Zhang, and X. Zhou, *GAIL: Guaranteed Automatic Integration Library (versions 1.0–2.0)*, MATLAB software, 2013–2014, `http://code.google.com/p/gail`.

[6] N. Clancy, Y. Ding, C. Hamilton, F. J. Hickernell, and Y. Zhang, *The cost of deterministic, adaptive, automatic algorithms: Cones, not balls*, J. Complexity, vol. 30, 21–45, 2014, `doi:10.1016/j.jco.2013.09.002`.

[7] F. Cucker and D. X. Zhou, *Learning theory: An approximation theory viewpoint*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, 2007.

[8] G. E. Fasshauer, *Meshfree approximation methods with* MATLAB, Interdisciplinary Mathematical Sciences, vol. 6, World Scientific Publishing Co., Singapore, 2007.

[9] G. E. Fasshauer, F. J. Hickernell, and H. Woźniakowski, *Average case approximation: Convergence and tractability of Gaussian kernels*, Monte Carlo and Quasi-Monte Carlo Methods 2010 (L. Plaskota and H. Woźniakowski, eds.), Springer-Verlag, Berlin, 2012, pp. 329–344.

[10] G. E. Fasshauer, F. J. Hickernell, and H. Woźniakowski, *On dimension-independent rates of convergence for function approximation with Gaussian kernels*, SIAM J. Numer. Anal., vol. 50, 247–271, 2012, `doi:10.1137/10080138X`.

[11] G. E. Fasshauer and M. McCourt, *Kernel-based approximation methods using MATLAB*, World Scientific Publishing Co. Pte. Ltd., 27 Warren Street, Suite 401-402, Hackensack, NJ 07601, 2015.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of statistical learning: Data mining, inference, and prediction*, second ed., Springer Series in Statistics, Springer Science+Business Media, Inc., New York, 2009.

[13] F. J. Hickernell, L. Jiang, Y. Liu, and A. B. Owen, *Guaranteed conservative fixed width confidence intervals via Monte Carlo sampling*, Monte Carlo and Quasi-Monte Carlo Methods 2012 (J. Dick, F. Y. Kuo, G. W. Peters, and I. H. Sloan, eds.), Springer-Verlag, Berlin, 2014, pp. 105–128.

[14] F. J. Hickernell and Ll. A. Jiménez Rugama, *Reliable adaptive cubature using digital sequences*, 2014, submitted for publication, arXiv:1410.8615 [math.NA].

[15] F. J. Hickernell, G. W. Wasilkowski, and H. Woźniakowski, *Tractability of linear multivariate problems in the average case setting*, Monte Carlo and Quasi-Monte Carlo Methods 2006 (A. Keller, S. Heinrich, and H. Niederreiter, eds.), Springer-Verlag, Berlin, 2008, pp. 423–452.

[16] F. J. Hickernell and H. Woźniakowski, *The price of pessimism for multidimensional quadrature*, J. Complexity, vol. 17, 625–659, 2001.

[17] J. K. Hunter and B. Nachtergaele, *Applied analysis*, World Scientific Publishing Company, Singapore, 2001.

[18] T. Jebara, R. Kondor, and A. Howard, *Probability product kernels*, Journal of Machine Learning Research, vol. 5, 819–844, 2004.

[19] L. Jiang and F. J. Hickernell, *Guaranteed conservative confidence intervals for means of Bernoulli random variables*, 2014, submitted for publication, arXiv:1411.1151.

[20] Ll. A. Jiménez Rugama and F. J. Hickernell, *Adaptive multidimensional integration based on rank-1 lattices*, 2014, submitted for publication, arXiv:1411.1966.

[21] A. Knutson and T. Tao, *Honeycombs and sums of hermitian matrices*, Notices of the AMS, vol. 48-2, 175–186, 2001.

[22] F. Y. Kuo, G. W. Wasilkowski, and H. Woźniakowski, *On the power of standard information for multivariate approximation in the worst case setting*, J. Approx. Theory, vol. 158, 97–125, 2009.

[23] W. Liu, J. C. Principle, and S. Haykin, *Kernel adaptive filtering: a comprehensive introduction*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2010.

[24] T. M. Morton and M. Neamtu, *Error bounds for solving pseudodifferential equations on spheres by collocation with zonal kernels*, Journal of Approximation Theory, vol. 114-2, 242–268, 2002.

[25] E. Novak and H. Woźniakowski, *Tractability of multivariate problems Volume I: Linear information*, EMS Tracts in Mathematics, no. 6, European Mathematical Society, Zürich, 2008.

[26] A. Papageorgiou and G. W. Wasilkowski, *On the average complexity of multivariate problems*, J. Complexity, vol. 6, 1–23, 1990.

[27] A. Pietsch, *Operator ideals*, North-Holland Publishing Co., Amsterdam, 1980.

[28] C. E. Rasmussen and C. Williams, *Gaussian processes for machine learning*, MIT Press, Cambridge, Massachusetts, 2006, (online version at `http://www.gaussianprocess.org/gpml/`).

[29] SAS Institute, *JMP 11*, 2013.

[30] R. Schaback and H. Wendland, *Kernel techniques: From machine learning to meshless methods*, Acta Numer., vol. 15, 543–639, 2006.

[31] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT Press, Cambridge, Massachusetts, 2002.

[32] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*, Cambridge University Press, 40 West 20th Street, New York, NY 10011-4211, USA, 2004.

[33] M. L. Stein, *Interpolation of spatial data: Some theory for kriging*, Springer-Verlag, New York, 1999.

[34] I. Steinwart and A. Christmann, *Support vector machines*, Springer Science+Business Media, Inc., 2008.

[35] X. Tong, *A guaranteed, adaptive, automatic algorithm for univariate function minimization*, Master's thesis, Illinois Institute of Technology, 2014.

[36] J. F. Traub, G. W. Wasilkowski, and H. Woźniakowski, *Information-based complexity*, Academic Press, Boston, 1988.

[37] G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.

[38] G. W. Wasilkowski, *Information of varying cardinality*, J. Complexity, vol. 2, 204–228, 1986.

[39] H. Wendland, *Scattered data approximation*, Cambridge Monographs on Applied and Computational Mathematics, no. 17, Cambridge University Press, Cambridge, 2005.

[40] H. Zhang, Y. Xu, and J. Zhang, *Reproducing kernel Banach spaces for machine learning*, Journal of Machine Learning Research, vol. 10, 2741–2775, 2009.

[41] X. Zhou and F. J. Hickernell, *Tractability of function approximation with product kernels*, Monte Carlo and Quasi-Monte Carlo Methods 2014 (R. Cools and D. Nuyens, eds.), Springer-Verlag, 2015, In Press, arXiv:1411.0790 [math.NA].