RATE OF CONVERGENCE AND TRACTABILITY OF THE RADIAL FUNCTION APPROXIMATION PROBLEM

GREGORY E. FASSHAUER, FRED J. HICKERNELL, AND HENRYK WOŹNIAKOWSKI

ABSTRACT. This article studies the problem of approximating functions belonging to a Hilbert space H_d with an isotropic or anisotropic Gaussian reproducing kernel,

$$K_d(\boldsymbol{x},\boldsymbol{t}) = \exp\left(-\sum_{\ell=1}^{d}\gamma_\ell^2(x_\ell - t_\ell)^2\right) \quad \text{for all } \boldsymbol{x},\boldsymbol{t}\in\mathbb{R}^d.$$

The isotropic case corresponds to using the same shape parameters for all coordinates, namely $\gamma_\ell = \gamma > 0$ for all ℓ, whereas the anisotropic case corresponds to varying shape parameters γ_ℓ. We are especially interested in moderate to large d. We consider two classes of algorithms:

(1) using finitely many arbitrary linear functionals,

(2) using only finitely many function values.

The pertinent error criterion is the worst case of such an algorithm over the unit ball in H_d, with the error for a single function given by the $\mathcal{L}_2$ norm also with a Gaussian weight.

Since the Gaussian kernel is analytic, the minimal worst case errors of algorithms that use at most n linear functionals or n function values vanish like $\mathcal{O}(n^{-p})$ as n goes to infinity. Here, p can be arbitrarily large, but the leading coefficient may depend on d (Theorem 1). On the other hand, if d dependence is taken into account, the convergence rate may be quite slow. If the goal is to make the error smaller than Cn^{-p} for some *C independent of d* or *polynomially dependent on d*, then this is possible for any choice of shape parameters with the largest p equal to $1/2$, provided that arbitrary linear functional data is available (Theorem 2). If the sequence of shape parameters γ_ℓ decays to zero like $\ell^{-\omega}$ as ℓ (and therefore also d) tends to ∞, then the largest p is roughly $\max(1/2,\omega)$ (Theorem 3). If only function values are available, dimension-independent convergence rates are somewhat worse (Theorems 4 and 5).

If the goal is to make the error smaller than *Cn^{-p} times the initial ($n = 0$) error*, then the corresponding p is roughly ω. Therefore it is the same as before iff $\omega \geq 1/2$ (Theorem 7 and Corollary 2). In particular, for the isotropic case, when $\omega = 0$, the error does not even decay polynomially with n^{-1} (Theorem 6). In summary, excellent dimension independent error decay rates are only possible when the sequence of shape parameters decays rapidly.

Date: May, 2010.

2010 *Mathematics Subject Classification.* Primary 65D15; Secondary 68Q17, 41A25, 41A63.

Key words and phrases. Gaussian kernel, reproducing kernel Hilbert spaces, shape parameter, tractability.

This research was supported in part for the first two authors by NSF-DMS grant #0713848, for the second author by Department of Energy grant #SC0002100, and for the third author by NSF-DMS grant #0914345.

1. Introduction

Algorithms for function approximation based on symmetric, positive definite kernels are important and fundamental tools for numerical computation [2, 5, 13, 23], statistical learning [1, 4, 8, 12, 14, 16, 17, 20], and are often used in engineering applications [6]. These algorithms go by a variety of names, including radial basis function methods [2], scattered data approximation [23], meshfree methods [5], (smoothing) splines [20], kriging [16], Gaussian process models [12] and support vector machines [17].

In a typical application we are given noisy or noiseless scalar or vector data. For simplicity, this article treats only the noiseless scalar case in which the data is 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_i on f. Here we assume that the domain of f is a subset of $\mathbb{R}^d$. One then chooses a symmetric, positive definite kernel K_d (see (3) below for the specific requirements), ideally such that $f \in H(K_d)$, where $H(K_d)$ is a reproducing kernel Hilbert space with the reproducing kernel K_d. Then it is a good idea to construct an approximation $S_n(f)$ to f which has the minimal norm among all elements in $H(K_d)$ that interpolate the data. This corresponds to the *spline* algorithm and requires the solution of an $n \times n$ system of linear equations. While the spline algorithm is optimal in the sense explained in Section 2 below, there still remains the important questions of how fast $S_n(f)$ converges to f as the number of data n tends to infinity, and how to choose the data sites or linear functionals to maximize the rate of convergence to f. Another question is to study how the error bounds depend on d. The latter question is especially important when d is large.

The typical convergence rates (see, e.g., [5, 23]) are of the form $\mathcal{O}(n^{-p/d})$, where p denotes the smoothness of the kernel K_d, and the design is chosen optimally. Unfortunately, for a finite p, this means that as the dimension increases, these known convergence rates deteriorate dramatically. Furthermore, the dimension dependence of the leading constant in the big $\mathcal{O}$-term is usually not known in these estimates.

This article studies Hilbert spaces with reproducing kernels $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The kernel is called *translation invariant* or *stationary* if $K(\boldsymbol{x}, \boldsymbol{t}) = \widetilde{K}(\boldsymbol{x} - \boldsymbol{t})$. In particular, the kernel is *radially symmetric* or *isotropic* if $K(\boldsymbol{x}, \boldsymbol{t}) = \kappa(\|\boldsymbol{x} - \boldsymbol{t}\|^2)$, in which case the kernel is called a *radial (basic) function*.

A kernel commonly used in practice, and one which is studied here, is the isotropic Gaussian kernel:

$$K_d(\boldsymbol{x}, \boldsymbol{t}) = e^{-\gamma^2 \|\boldsymbol{x}-\boldsymbol{t}\|^2} \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{t} \in \mathbb{R}^d, \tag{1a}$$

where a positive γ is called the *shape parameter*. This parameter functions as an inverse length scale. Choosing γ very small has a beneficial effect on the rate of decay of the eigenvalues of the Gaussian kernel, as is shown below. An anisotropic, but stationary generalization of the Gaussian kernel is obtained by introducing a different positive shape parameter γ_ℓ for each variable,

$$K_d(\boldsymbol{x}, \boldsymbol{t}) = e^{-\gamma_1^2 (x_1-t_1)^2 - \cdots - \gamma_d^2 (x_d-t_d)^2} \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{t} \in \mathbb{R}^d. \tag{1b}$$

TABLE 1. Error decay rates as a function of sample size n

	Error Criterion	
Data Available	Absolute	Normalized
Arbitrary Linear functionals	$\asymp n^{-\max(r(\boldsymbol{\gamma}),1/2)}$ Theorem 3	$\asymp n^{-r(\boldsymbol{\gamma})}$ if $r(\boldsymbol{\gamma}) > 0$, Theorem 7
Function values	$\preceq n^{-\max(r(\boldsymbol{\gamma})/[1+1/(2r(\boldsymbol{\gamma}))],1/4)}$ Theorem 4 and 5	$\preceq n^{-r(\boldsymbol{\gamma})/[1+1/(2r(\boldsymbol{\gamma}))]}$ if $r(\boldsymbol{\gamma}) > 1/2$, Corollary 2

As evidence of its popularity, we note that this latter kernel is used in the Gaussian process modeling module of the JMP commercial statistical software [9]. In JMP, the values of the γ_ℓ are determined in a data-driven way[1].

We stress that the Gaussian kernels are analytic, and the smoothness parameter $p = \infty$. Therefore one can hope to obtain convergence rates of the form $\mathcal{O}(n^{-\tau})$ for an arbitrarily large τ. As we shall see, this is indeed the case. This is shown in Theorem 1 and explained in Section 4. However, the dependence on d is a function of τ and only for a relatively small τ is the dependence on d acceptable.

Given the growing number of applications with moderate to large dimension, d, it is desirable to have dimension-independent polynomial convergence rates of the form Cn^{-p} for positive C and p, which corresponds to *strong polynomial tractability*, or at worst, convergence rates that are polynomially dependent on dimension d and are of the form $Cd^q\, n^{-p}$ for positive C, q and p, which corresponds to *polynomial tractability*.

This paper establishes convergence rates with polynomial or no dimension dependence for the Gaussian kernel introduced in (1). The rates are summarized in Table 1. As explained in Section 2, the absolute error is the $\mathcal{L}_2$ worst case approximation error based on a Gaussian weight with mean zero and variance $1/2$. The normalized error is the absolute error divided by $\|I_d\|$, were I_d denotes the embedding between the radial function space H_d and the $\mathcal{L}_2$ space. Note that the norm $\|I_d\|$ is the initial error that can be achieved by the zero algorithm without sampling the functions. The dimension independent convergence rates depend to some extent on which error criterion is used. They also depend on whether the data available consists only of function values or, more generally, of arbitrary linear functionals. This latter, more generous setting may allow for faster convergence.

The notation $\preceq n^{-p}$ in Table 1 means that for all $\delta > 0$ the error is bounded *above* by $Cn^{-p+\delta}$ for some constant C that is independent of the sample size, n, and the dimension, d, but it may depend on δ. The notation $\succeq n^{-p}$ is defined analogously, and means that the error is bounded *below* by $Cn^{-p-\delta}$ for all $\delta > 0$. The notation $\asymp n^{-p}$ means that the error is both $\preceq n^{-p}$ and $\succeq n^{-p}$.

As can be seen in Table 1, the convergence rates depend strongly on how fast the sequence of shape parameters $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell\in\mathbb{N}}$ goes to zero. The term $r(\boldsymbol{\gamma})$ appearing in

[1]In the tractability literature, the shape parameters γ_ℓ are called *product weights*.

TABLE 2. Number of data, $n(\varepsilon, H_d)$, needed to obtain an error tolerance ε

	Error Criterion	
Data Available	Absolute	Normalized
Arbitrary Linear functionals	$\asymp \varepsilon^{-\min(1/r(\boldsymbol{\gamma}),2)}$ Theorem 3	$\asymp \varepsilon^{-1/r(\boldsymbol{\gamma})}$ if $r(\boldsymbol{\gamma})>0$, Theorem 7
Function values	$\preceq \varepsilon^{-\min(1/r(\boldsymbol{\gamma})+1/[2r^2(\boldsymbol{\gamma})],4)}$ Theorem 4 and 5	$\preceq \varepsilon^{-1/r(\boldsymbol{\gamma})-1/[2r^2(\boldsymbol{\gamma})]}$ if $r(\boldsymbol{\gamma})>1/2$, Corollary 2

Table 1, is defined by

$$r(\boldsymbol{\gamma}) = \sup\left\{\beta>0 \,\middle|\, \sum_{\ell=1}^{\infty} \gamma_\ell^{1/\beta} < \infty\right\} \tag{2}$$

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

For instance, for the isotropic case with $\gamma_\ell = \gamma > 0$ we have $r(\boldsymbol{\gamma}) = 0$, whereas for $\gamma_\ell = \ell^{-\alpha}$ for a nonnegative α we have $r(\boldsymbol{\gamma}) = \alpha$. If the γ_ℓ are ordered, that is, $\gamma_1 \ge \gamma_2 \ge \cdots$, then this definition is equivalent to

$$r(\boldsymbol{\gamma}) = \sup\left\{\beta \ge 0 \,\middle|\, \lim_{\ell\to\infty} \gamma_\ell\, \ell^\beta = 0\right\}.$$

For excellent *dimension independent convergence* one needs the sequence of shape parameters to decay to zero quickly, as can be seen in Table 1. These results are derived in Sections 5 and 6.

While writing the error as a function of the sample size is common in the numerical analysis literature, the computational complexity literature looks at the number of data required to obtain a given error tolerance. Let $n(\varepsilon, H_d)$ denote the minimal number of function values or linear functionals that are needed to compute an $\varepsilon\cdot\mathrm{CRI}_d$ approximation. Here, $\mathrm{CRI}_d = 1$ for the absolute error criterion, and $\mathrm{CRI}_d = \|I_d\|$ for the normalized error criterion. Again, $\|I_d\|$ is the initial error that can be achieved by the zero algorithm without sampling the functions. The tractability results presented in this paper are summarized in Table 2.

For the absolute error and algorithms that use arbitrary linear functionals, we prove *strong polynomial tractability* for *all* choices of shape parameters γ_ℓ. Furthermore, the exponent 2 of ε^{-1} is best possible for all γ_ℓ's that go to zero no faster that ℓ^{-2}. For the absolute error and algorithms that use function values, we still have strong polynomial tractability with exponent at most 4.

For the normalized error, the situation is much worse. If the sequence of shape parameters tends to zero fast enough, we still have strong polynomial tractability. However, for the isotropic case it follows that $n(\varepsilon, H_d)$ does *not* depend polynomially on ε^{-1} and d. For algorithms using arbitrary linear functionals, we have *quasi-polynomial tractability*, i.e., there are positive C and t such that

$$n(\varepsilon, H_d) \le C\,\exp\bigl(t\,\bigl(1+\ln\,d\bigr)\,\bigl(1+\ln\,\varepsilon^{-1}\bigr)\bigr) \quad \text{for all} \quad \varepsilon\in(0,1) \ \text{ and } \ d\in\mathbb{N}.$$

Furthermore, the smallest t is roughly[2]

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

As a prelude to deriving these convergence and tractability results, the next section reviews some principles of function approximation on Hilbert spaces. Section 3 applies these principles to the Gaussian kernel.

2. Function Approximation

Let $H_d = H(K_d)$ denote a reproducing kernel Hilbert space of real functions defined on a Lebesgue measurable set $D_d \subseteq \mathbb{R}^d$. The goal is to accurately approximate any function in H_d given a finite number of data about it. The reproducing kernel

$$K_d : D_d \times D_d \to \mathbb{R}$$

is symmetric, positive definite and reproduces function values. This means that for all $n \in \mathbb{N}$, $\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in D_d$, $\mathbf{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$ and $f \in H_d$, the following properties hold:

$$K_d(\cdot, \boldsymbol{x}) \in H_d, \tag{3a}$$

$$K_d(\boldsymbol{x}, \boldsymbol{t}) = K_d(\boldsymbol{t}, \boldsymbol{x}), \tag{3b}$$

$$\sum_{i=1}^{n} K_d(\boldsymbol{x}_i, \boldsymbol{x}_j) c_i c_j \geq 0, \tag{3c}$$

$$f(\boldsymbol{x}) = \langle f, K_d(\cdot, \boldsymbol{x}) \rangle_{H_d}. \tag{3d}$$

For an arbitrary $\boldsymbol{x} \in D_d$ consider the linear functional $L_{\boldsymbol{x}}(f) = f(\boldsymbol{x})$ for all $f \in H_d$. Then $L_{\boldsymbol{x}}$ is continuous and $\|L_{\boldsymbol{x}}\|_{H_d^*} = K_d^{1/2}(\boldsymbol{x}, \boldsymbol{x})$. The reader may find these and other properties in e.g., [1, 20]. Many reproducing kernels are used in practice. A popular choice is the Gaussian kernel defined in (1) for which $D_d = \mathbb{R}^d$.

It is assumed that H_d is continuously embedded in the space $\mathcal{L}_2 = \mathcal{L}_2(D_d, \varrho_d)$ of square Lebesgue integrable functions. Here, ϱ_d is a probability density function, i.e., $\varrho_d \geq 0$ and $\int_{D_d} \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t} = 1$. The norm in the space $\mathcal{L}_2$ is given by

$$\|f\|_{\mathcal{L}_2} = \left(\int_{D_d} f^2(\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}\right)^{1/2}.$$

Continuous embedding means that the linear embedding operator $I_d : H_d \to \mathcal{L}_2$ given by $I_d f = f$ is continuous,

$$\|I_d f\|_{\mathcal{L}_2} \leq \|I_d\|\ \|f\|_{H_d} \quad \text{for all} \quad f \in H_d.$$

[2]In this paper, by ln we mean the natural logarithm of base e.

Observe that

$$\begin{aligned} \|I_d f\|_{\mathcal{L}_2}^2 &= \int_{D_d} f^2(\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t} = \int_{D_d} \langle f, K_d(\cdot,\boldsymbol{t})\rangle_{H_d}^2\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t} \\ &\le \|f\|_{H_d}^2 \int_{D_d} K_d(\boldsymbol{t},\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}. \end{aligned}$$

Hence, it is enough to assume that

$$\int_{D_d} K_d(\boldsymbol{t},\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t} < \infty \tag{4}$$

to guarantee that I_d is continuous, and obviously

$$\|I_d\| \le \left[\int_{D_d} K_d(\boldsymbol{t},\boldsymbol{t})\, \varrho(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}\right]^{1/2}.$$

Functions in H_d are approximated by linear algorithms[3]

$$A_n(f) = \sum_{j=1}^{n} L_j(f) a_j \quad \text{for all} \quad f \in H_d \tag{5}$$

for some continuous linear functionals $L_j \in H_d^*$, and functions $a_j \in \mathcal{L}_2$. The *worst case error* of the algorithm A_n is then defined as

$$e^{\mathrm{wor}}(A_n) = \sup_{\|f\|_{H_d}\le 1} \|f - A_n(f)\|_{\mathcal{L}_2}.$$

The linear algorithms A_n considered here are based on function data $L_j(f)$, where the continuous linear functionals L_j may belong to one of two classes. The first class, denoted Λ^{std}, is comprised only of function values and is called *standard*. That is, $L_j \in \Lambda^{\mathrm{std}}$ iff $L_j(f) = f(\boldsymbol{t}_j)$ for all $f \in H_d$ for some $\boldsymbol{t}_j \in D_d$. The second class, denoted Λ^{all}, is comprised of arbitrary continuous functionals and is called *linear*. That is, $L_j \in \Lambda^{\mathrm{all}}$ iff $L_j \in H_d^*$. Obviously, $\Lambda^{\mathrm{std}} \subseteq \Lambda^{\mathrm{all}}$.

The aim is to determine how small the worst case error can be by choosing linear algorithms with only n linear functionals either from Λ^{std} or Λ^{all}. The *nth minimal worst case error* is defined as

$$e^{\mathrm{wor}-\vartheta}(n, H_d) = \inf_{A_n \text{ with } L_j \in \Lambda^{\vartheta}} e^{\mathrm{wor}}(A_n), \quad \vartheta \in \{\mathrm{std}, \mathrm{all}\}.$$

Here and below, for notational simplicity ϑ denotes either the standard or linear setting. Clearly, $e^{\mathrm{wor}-\mathrm{all}}(n, H_d) \le e^{\mathrm{wor}-\mathrm{std}}(n, H_d)$ since the former uses a larger class of function data.

[3]It is well known that adaption and nonlinear algorithms do not help for approximation of linear problems. A linear problem is defined as a linear operator and we approximate its values over a set that is convex and balanced. The typical example of such a set is the unit ball as taken in this paper. Then among all algorithms that use linear adaptive functionals, the worst case error is minimized by a linear algorithm that uses nonadaptive linear functionals. Adaptive choice of a linear functional means that the choice of L_j in (5) may depend on the already computed values $L_i(f)$ for $i = 1, 2, \ldots, j-1$. That is why in our case, the restriction to linear algorithms of the form (5) can be done without loss of generality, for more detail see, e.g., [19].

The case $n = 0$ means that no linear functionals of f are used to construct the algorithm. It is easy to see that the best algorithm possible is $A_0 = 0$, and then

$$e^{\mathrm{wor}-\vartheta}(0, H_d) = \|I_d\|.$$

The minimal error for $n = 0$ is called the initial error and it only depends on the formulation of the problem.

This article addresses two problems: convergence and tractability. The former considers how fast the error vanishes as n increases, and the latter considers how the error depends on the dimension, d, as well as the number of data, n.

Problem 1: Rate of Convergence

We would like to know how fast $e^{\mathrm{wor}-\vartheta}(n, H_d)$ goes to zero as n goes to infinity. In particular, we study the rate of convergence (defined by (2)) of the sequence $\{e^{\mathrm{wor}-\vartheta}(n, H_d)\}_{n\in\mathbb{N}}$. Since the numbers $e^{\mathrm{wor}-\vartheta}(n, H_d)$ are ordered, we have

$$r^{\mathrm{wor}-\vartheta}(H_d) := r\left(\{e^{\mathrm{wor}-\vartheta}(n, H_d)\}\right) = \sup\left\{\beta \geq 0 \,|\, \lim_{n\to\infty} e^{\mathrm{wor}-\vartheta}(n, H_d)\, n^{\beta} = 0\right\}. \tag{6}$$

Roughly speaking, the rate of convergence is the largest β for which the nth minimal errors behave no worse than $n^{-\beta}$. For example, if $e^{\mathrm{wor}-\vartheta}(n, H_d) = n^{-\alpha}$ for a positive α then $r^{\mathrm{wor}-\vartheta}(H_d) = \alpha$. Under this definition, even sequences of the form $e^{\mathrm{wor}-\vartheta}(n, H_d) = n^{-\alpha} \ln^p n$ for an arbitrary p still have $r^{\mathrm{wor}-\vartheta}(H_d) = \alpha$. On the other hand, if $e^{\mathrm{wor}-\vartheta}(n, H_d) = q^n$ for a number $q \in (0, 1)$ then $r^{\mathrm{wor}-\vartheta}(H_d) = \infty$.

Obviously, $r^{\mathrm{wor}-\mathrm{all}}(H_d) \geq r^{\mathrm{wor}-\mathrm{std}}(H_d)$. We would like to know both rates and whether

$$r^{\mathrm{wor}-\mathrm{all}}(H_d) > r^{\mathrm{wor}-\mathrm{std}}(H_d),$$

i.e., whether Λ^{all} admits a better rate of convergence than Λ^{std}.

Problem 2: Tractability

Assume that there is a sequence of spaces $\{H_d\}_{d\in\mathbb{N}}$ and embedding operators $\{I_d\}_{d\in\mathbb{N}}$. In this case, we would like to know how the minimal errors $e^{\mathrm{wor}-\vartheta}(n, H_d)$ depend not only on n but also on d.

More precisely, we consider the *absolute* and *normalized* error criteria. For a given (small) 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 for the absolute error criterion, and does not exceed $\varepsilon\,\|I_d\|$ for the normalized error criterion. That is,

$$n^{\mathrm{wor}-\psi-\vartheta}(\varepsilon, H_d) = \min\left\{n \,|\, e^{\mathrm{wor}-\vartheta}(n, H_d) \leq \varepsilon\, \mathrm{CRI}_d^{\psi}\right\}, \quad \psi \in \{\mathrm{abs}, \mathrm{norm}\},$$

where $\mathrm{CRI}_d^{\mathrm{abs}} = 1$ for the absolute error criterion and $\mathrm{CRI}_d^{\mathrm{nor}} = \|I_d\|$ for 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* iff there exist numbers C, p and q such that

$$n^{\mathrm{wor}-\psi-\vartheta}(\varepsilon, H_d) \leq C\, d^{\,q}\, \varepsilon^{-p} \quad \text{for all} \quad d \in \mathbb{N} \;\text{ and }\; \varepsilon \in (0, 1).$$

If $q = 0$ above then we say that $\mathcal{I}$ is *strongly polynomially tractable* and the infimum of p satisfying the bound above is called the *exponent* of strong polynomial tractability.

The essence of polynomial tractability is to guarantee that a polynomial number of linear functionals is enough to satisfy the function approximation problem to within ε. Obviously, polynomial tractability depends on which class, $\Lambda^{\rm all}$ or $\Lambda^{\rm std}$, is considered and whether the absolute or normalized error is used. As shall be shown, the results on polynomial tractability depend on the cases considered.

The property of strong polynomial tractability is especially challenging since then the number of linear functionals needed for an ε-approximation is independent of d. The reader may suspect that this property is too strong and cannot happen for function approximation. Nevertheless, there are positive results to report on strong polynomial tractability.

Besides polynomial tractability, there are the somewhat less demanding concepts such as quasi-polynomial tractability and weak tractability. The problem $\mathcal{I}$ is *quasi-polynomially tractable* iff there exist numbers C and t for which

$$n^{\rm wor-\psi-\vartheta}(\varepsilon, H_d) \le C \, \exp\left(t \, \ln(1+d) \, \ln(1+\varepsilon^{-1})\right)$$

for all $d \in \mathbb{N}$ and $\varepsilon > 0$. The exponent of quasi-polynomial tractability is defined as the infimum of t satisfying the bound above. Finally, $\mathcal{I}$ is *weakly tractable* iff

$$\lim_{\varepsilon^{-1}+d \to \infty} \frac{\ln \, n^{\rm wor-\psi-\vartheta}(\varepsilon, H_d)}{\varepsilon^{-1}+d} = 0.$$

Note that for a fixed d, quasi-polynomial tractability means that

$$n^{\rm wor-\psi-\vartheta}(\varepsilon, H_d) = \mathcal{O}\left(\varepsilon^{-t(1+\ln \, d)}\right) \quad \text{as} \quad \varepsilon \to 0.$$

Hence, the exponent of ε^{-1} may now weakly depend on d through $\ln \, d$. On the other hand, weak tractability only means that we do not have exponential dependence on ε^{-1} and d.

We will report about quasi-polynomial and weak tractability in the case when polynomial tractability does not hold. As before, quasi-polynomial and weak tractability depend on which class $\Lambda^{\rm all}$ or $\Lambda^{\rm std}$ is considered and on the error criterion.

Motivation of tractability study and more on tractability concepts can be found in [11]. Quasi-polynomial tractability has been recently studied in [7].

We end this section by briefly reviewing some general results related to the problems of convergence and tractability mentioned above. For a given design, i.e., given continuous linear functionals $L_1, \ldots, L_n$, it is known how to find functions $a_1, \ldots, a_n$ for which the worst case error of A_n is minimized. The optimal algorithm, S_n, should be taken as the *spline* or the *minimal norm interpolant*, see e.g. Section 5.7 of [19]. The spline algorithm was briefly mentioned in the introduction. It is described in more generality here.

For given $y_j = L_j(f)$ for $j = 1, 2, \ldots, n$, we take $S_n(f)$ as an element of H_d that satisfies the conditions

$$\begin{aligned} L_j(S_n(f)) &= y_j \qquad\qquad\qquad\qquad \text{for} \quad j = 1, 2, \ldots, n, \\ \|S_n(f)\|_{H_d} &= \inf_{g \in H_d,\ L_j(g) = y_j,\ j=1,2,\ldots,n} \|g\|_{H_d}. \end{aligned}$$

The construction of $S_n(f)$ may be done by solving a linear equation $\mathsf{K}\boldsymbol{c} = \boldsymbol{y}$, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ and the $n \times n$ matrix is given by

$$\mathsf{K} = (k_{i,j})_{i,j=1}^n \quad \text{with} \quad k_{i,j} = L_i(g_j) \ \text{ and } \ g_j(\boldsymbol{x}) = L_j K_d(\cdot, \boldsymbol{x}).$$

Then

$$S_n(f)(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x}) \mathsf{K}^{-1} \boldsymbol{y} \quad \text{with} \quad \boldsymbol{k}(\boldsymbol{x}) = (L_i K_d(\cdot, \boldsymbol{x}))_{i=1}^n$$

and

$$e^{\mathrm{wor}}(S_n) = \sup_{\|f\|_{H_d} \le 1,\ L_j(f) = 0,\ j=1,2,\ldots,n} \|f\|_{\mathcal{L}_2}.$$

Note that depending on the choice of linear functionals $L_1, \ldots, L_n$ the matrix K may not necessarily be invertible, however, the solution $\boldsymbol{c} = \mathsf{K}^{-1}\boldsymbol{y}$ is always well defined as the vector of minimal Euclidean norm which satisfies $\mathsf{K}\boldsymbol{c} = \boldsymbol{y}$.

The spline enjoys more optimality properties. For instance, it minimizes the *local* worst case error. Roughly speaking this means that for each $\boldsymbol{x} \in D_d$, the worst possible pointwise error $|f(\boldsymbol{x}) - A_n(f)(\boldsymbol{x})|$ over the unit ball of functions f is minimized over all possible A_n by choosing $A_n = S_n$. We do not elaborate more on this point.

It is non-trivial to find the linear functionals L_j from the class Λ^{std} that minimize the error of the spline algorithm S_n. For the class Λ^{all}, the optimal design is known, at least theoretically, see again e.g., [19]. Namely, let $W_d = I_d^* I_d : H_d \to H_d$, where $I_d^* : \mathcal{L}_2 \to H_d$ denotes the adjoint of the imbedding operator, i.e., the operator satisfying $\langle f, I_d^* h \rangle_{H_d} = \langle I_d f, h \rangle_{\mathcal{L}_2}$ for all $f \in H_d$ and $h \in \mathcal{L}_2$. As a consequence, W_d is a self adjoint and positive definite linear operator given by

$$W_d(f) = \int_{D_d} f(\boldsymbol{t})\, K_d(\cdot, \boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t} \quad \text{for all} \quad f \in H_d.$$

Clearly,

$$\langle f, g \rangle_{\mathcal{L}_2} = \langle I_d f, I_d g \rangle_{\mathcal{L}_2} = \langle W_d f, g \rangle_{H_d} = \langle f, W_d g \rangle_{H_d} \quad \text{for all} \quad f, g \in H_d.$$

It is known that $\lim_{n\to\infty} e^{\mathrm{wor-all}}(n, H_d) = 0$ iff W_d is compact. In particular, (4) implies that W_d is compact.

Assuming that W_d is compact, let us define its eigenpairs by $(\lambda_{d,j}, \eta_{d,j})$, where the eigenvalues are ordered, $\lambda_{d,1} \ge \lambda_{d,2} \ge \cdots$, and

$$W_d\, \eta_{d,j} = \lambda_{d,j}\, \eta_{d,j} \quad \text{with} \quad \langle \eta_{d,j}, \eta_{d,i} \rangle_{H_d} = \delta_{i,j} \ \text{ for all } \ i, j \in \mathbb{N}.$$

Note also that for any $f \in H_d$ we have

$$\langle f, \eta_{d,j} \rangle_{\mathcal{L}_2} = \langle I_d f, I_d \eta_{d,j} \rangle_{\mathcal{L}_2} = \langle f, W_d \eta_{d,j} \rangle_{H_d} = \lambda_{d,j} \langle f, \eta_{d,j} \rangle_{H_d}.$$

Taking $f = \eta_{d,i}$ we see that $\{\eta_{d,j}\}$ us a set of orthogonal functions in $\mathcal{L}_2$. For simplicity and without loss of generality we assume that all $\lambda_{d,j}$ are positive[4]. Letting

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

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

$$K_d(\boldsymbol{x}, \boldsymbol{t}) = \sum_{j=1}^{\infty} \eta_{d,j}(\boldsymbol{x})\, \eta_{d,j}(\boldsymbol{t}) = \sum_{j=1}^{\infty} \lambda_{d,j}\, \varphi_{d,j}(\boldsymbol{x})\, \varphi_{d,j}(\boldsymbol{t}) \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{t} \in D_d. \tag{7}$$

If (4) holds then

$$\sum_{j=1}^{\infty} \lambda_{d,j} = \int_{D_d} K_d(\boldsymbol{t}, \boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t} < \infty. \tag{8}$$

This means that (4) implies that W_d is also a finite trace operator.

It is known that the best choice of L_j for the class Λ^{all} is $L_j = \langle \cdot, \eta_{d,j} \rangle_{H_d}$. Then the spline algorithm S_n with the minimal worst case error is defined using the eigenfunctions corresponding to the n largest eigenvalues:

$$S_n(f) = \sum_{j=1}^{n} \langle f, \eta_{d,j} \rangle_{H_d}\, \eta_{d,j} \qquad \text{for all} \quad f \in H_d,$$

and

$$e^{\mathrm{wor}}(S_n) = e^{\mathrm{wor-all}}(n, H_d) = \sqrt{\lambda_{d,n+1}} \qquad \text{for all} \quad n \in \mathbb{N}.$$

The last formula for $n = 0$ yields that the initial error is $\|I_d\| = \sqrt{\lambda_{d,1}}$.

The results for the class Λ^{all} are useful for finding rates of convergence as well as necessary and sufficient conditions on polynomial, quasi-polynomial and weak tractability in terms of the behavior of the eigenvalues $\lambda_{d,j}$. This has already been done in a number of papers or books, and we will report these results later for spaces studied in this paper. For the class Λ^{std}, the situation is much harder although there are papers that relate rates of convergence and tractability conditions between classes Λ^{all} and Λ^{std}. Again we report these results later.

3. Radial Function Spaces

The focus of this article is on reproducing kernels that are *translation invariant* or *stationary*, i.e.,

$$K_d(\boldsymbol{x}, \boldsymbol{t}) = \widetilde{K}_d(\boldsymbol{x} - \boldsymbol{t}) \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{t} \in D_d = \mathbb{R}^d.$$

An even more special case is for *radially symmetric* or *isotropic* kernels, i.e.,

$$K_d(\boldsymbol{x}, \boldsymbol{t}) = \kappa(\|\boldsymbol{x} - \boldsymbol{t}\|_2^2) \quad \text{with} \quad \|\boldsymbol{x} - \boldsymbol{t}\|_2^2 = \sum_{\ell=1}^{d} (x_\ell - t_\ell)^2.$$

Here, $\widetilde{K}_d$ and κ are chosen such that K_d is a reproducing kernel.

[4]Otherwise, we should switch to a subspace of H_d spanned by eigenfunctions corresponding to k positive eigenvalues, and replace $\mathbb{N}$ by $\{1, 2, \ldots, k\}$.

Isotropic kernels also go by the name *radial basis functions*, and the spaces $H(K_d)$ are referred to as *radial function spaces*. Stationary or isotropic kernels are common in the literature on computational mathematics [2, 5, 23], statistics [1, 16, 20], statistical learning [12, 17], and engineering applications [6].

A popular isotropic kernel is the Gaussian kernel, defined in (1), which has both an isotropic version,

$$K_d(\boldsymbol{x},\boldsymbol{t}) = e^{-\gamma^2\|\boldsymbol{x}-\boldsymbol{t}\|^2} \quad \text{for all} \quad \boldsymbol{x},\boldsymbol{t}\in\mathbb{R}^d,$$

and a more general anisotropic version,

$$K_d(\boldsymbol{x},\boldsymbol{t}) = e^{-\gamma_1^2(x_1-t_1)^2-\cdots-\gamma_d^2(x_d-t_d)^2} \quad \text{for all} \quad \boldsymbol{x},\boldsymbol{t}\in\mathbb{R}^d.$$

As alluded to in the introduction, the shape parameter, γ or $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell\in\mathbb{N}}$, which functions as an inverse length scale, plays an important role in the tractability of function approximation. Choosing the γ_ℓ to decay quickly has a beneficial effect on the rate of decay of the eigenvalues of the Gaussian kernel, as we shall see below. On the other hand, a small value of γ leads to a huge condition number of the matrix K and may result in severe numerical instabilities. While this is an issue that is very important for practical implementations, and has received some attention, we will not discuss it any further here.

We now analyze the function approximation problem for the Hilbert space $H_d = H(K_d)$ with the isotropic Gaussian kernel K_d given by (1a) or, more generally, with the anisotropic Gaussian kernel given by (1b). For the space $\mathcal{L}_2(\mathbb{R}^d,\varrho_d)$ we take the Gaussian weight with zero mean and variance $1/2$, i.e.,

$$\varrho_d(\boldsymbol{t}) = \frac{1}{\pi^{d/2}}\exp\left(-(t_1^2+t_2^2+\cdots+t_d^2)\right) \qquad \text{for all} \quad \boldsymbol{t}\in\mathbb{R}^d.$$

Note that $K_d(\boldsymbol{t},\boldsymbol{t}) = 1$ for all $\boldsymbol{t}\in\mathbb{R}^d$, and therefore

$$\int_{\mathbb{R}^d} K_d(\boldsymbol{t},\boldsymbol{t})\,\varrho_d(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} = 1,$$

so that (4) holds. This means that the embedding I_d is continuous, the operator W_d is compact and a finite trace operator with

$$\sum_{j=1}^{\infty}\lambda_{d,j} = 1, \tag{9}$$

by (8).

Observe that (4) holds for all translation invariant kernels since

$$\int_{\mathbb{R}^d} K_d(\boldsymbol{t},\boldsymbol{t})\,\varrho_d(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} = \widetilde{K}_d(\boldsymbol{0}),$$

but it can now depend on d. For radially symmetric kernels we have

$$\int_{\mathbb{R}^d} K_d(\boldsymbol{t},\boldsymbol{t})\,\varrho_d(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} = \kappa(0),$$

and it is independent of d.

Since a Gaussian kernel K_d is of the product form, the space H_d is the tensor product of the Hilbert spaces of univariate spaces with the kernels $e^{-\gamma_\ell^2(x-t)^2}$ for $x, t \in \mathbb{R}$. This also implies that the operator W_d is of the product form and its eigenpairs are products of the corresponding eigenpairs for the univariate cases.

Consider now $d = 1$, and the space $H(K_1)$ with $K_1(x,t) = e^{-\gamma^2(x-t)^2}$. Then the eigenpairs $\left(\tilde{\lambda}_{\gamma,j}, \eta_{\gamma,j}\right)$ of W_1 are known, see [12]. Note that we have introduced the notation $\tilde{\lambda}_{\gamma,j}$ to emphasize the dependence of the eigenvalues on γ in the following discussion (while the dependence on d has temporarily dropped from the notation). We have

$$\tilde{\lambda}_{\gamma,j} = \frac{1}{\sqrt{\frac{1}{2}(1+\sqrt{1+4\gamma^2})+\gamma^2}} \left(\frac{\gamma^2}{\frac{1}{2}(1+\sqrt{1+4\gamma^2})+\gamma^2}\right)^{j-1} = (1-\omega_\gamma)\,\omega_\gamma^{j-1},$$

where

$$\omega_\gamma = \frac{\gamma^2}{\frac{1}{2}(1+\sqrt{1+4\gamma^2})+\gamma^2}, \tag{10}$$

and $\eta_{\gamma,j} = \sqrt{\tilde{\lambda}_{\gamma,j}}\,\varphi_{\gamma,j}$ with

$$\varphi_{\gamma,j}(x) = \sqrt{\frac{(1+4\gamma^2)^{1/4}}{2^{j-1}(j-1)!}} \exp\left(-\frac{\gamma^2 x^2}{\frac{1}{2}(1+\sqrt{1+4\gamma^2})}\right) H_{j-1}\left((1+4\gamma^2)^{1/4}x\right),$$

where H_{j-1} is the Hermite polynomial of degree $j-1$, given by

$$H_{j-1}(x) = (-1)^{j-1} e^{x^2} \frac{\mathrm{d}^{j-1}}{\mathrm{d}x^{j-1}} e^{-x^2} \quad \text{for all} \quad x \in \mathbb{R},$$

so that

$$\int_{\mathbb{R}} H_{j-1}^2(x)\, e^{-x^2}\, \mathrm{d}x = \sqrt{\pi}\, 2^{j-1}(j-1)! \qquad \text{for} \;\; j = 1, 2, \ldots .$$

Obviously, we have

$$\langle \eta_{\gamma,i}, \eta_{\gamma,j} \rangle_{H(K_1)} = \langle \varphi_{\gamma,i}, \varphi_{\gamma,j} \rangle_{\mathcal{L}_2} = \delta_{ij},$$

and applying (7) we obtain

$$K_1(x,t) = e^{-\gamma^2(x-t)^2} = \sum_{j=1}^{\infty} \tilde{\lambda}_{\gamma,j} \varphi_{\gamma,j}(x) \varphi_{\gamma,j}(y) \qquad \text{for all} \quad x, t \in \mathbb{R}.$$

Note that the eigenvalues $\tilde{\lambda}_{\gamma,j}$ are ordered. The largest eigenvalue is

$$\tilde{\lambda}_{\gamma,1} = 1 - \omega_\gamma = \sqrt{\frac{2}{1+\sqrt{1+4\gamma^2}+2\gamma^2}} = 1 - \gamma^2 + \mathcal{O}(\gamma^4) \quad \text{as} \quad \gamma \to 0.$$

Furthermore,

$$\tilde{\lambda}_{\gamma,j} = \left(1 - \gamma^2 + \mathcal{O}(\gamma^4)\right) \left(\frac{\gamma^2}{1-\gamma^2+\mathcal{O}(\gamma^4)}\right)^{j-1} \qquad \text{for} \;\; j = 1, 2, \ldots . \tag{11}$$

The space $H(K_1)$ consists of analytic functions for which

$$\|f\|^2_{H(K_1)} = \sum_{j=1}^{\infty} \langle f, \eta_{\gamma,j}\rangle^2_{H(K_1)} = \sum_{j=1}^{\infty} \frac{1}{\tilde{\lambda}_{\gamma,j}} \langle f, \varphi_{\gamma,j}\rangle^2_{\mathcal{L}_2} < \infty.$$

This means that the coefficients of f in the space $\mathcal{L}_2$ decay exponentially fast. The inner product is obviously given as

$$\langle f, g\rangle_{H(K_1)} = \sum_{j=1}^{\infty} \frac{1}{\tilde{\lambda}_{\gamma,j}} \int_{\mathbb{R}} f(t)\, \frac{\varphi_{\gamma,j}(t)}{\sqrt{\pi}}\, e^{-t^2}\, \mathrm{d}t \int_{\mathbb{R}} g(t)\, \frac{\varphi_{\gamma,j}(t)}{\sqrt{\pi}}\, e^{-t^2}\, \mathrm{d}t$$

$$\text{for all} \quad f, g \in H(K_1).$$

The reader may find more about the characterization of the space $H(K_1)$ in [18].

For $d > 1$, let $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell\in\mathbb{N}}$ and $\boldsymbol{j} = (j_1, j_2, \ldots, j_d) \in \mathbb{N}^d$. As already mentioned, the eigenpairs $\left(\tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}}, \eta_{d,\boldsymbol{\gamma},\boldsymbol{j}}\right)$ of W_d are given by the products

$$\begin{aligned}(12)\quad \tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}} &= \prod_{\ell=1}^{d} \tilde{\lambda}_{\gamma_\ell,j_\ell} = \prod_{\ell=1}^{d} \frac{1}{\sqrt{\frac{1}{2}(1+\sqrt{1+4\gamma_\ell^2})+\gamma_\ell^2}} \left(\frac{\gamma_\ell^2}{\frac{1}{2}(1+\sqrt{1+4\gamma_\ell^2})+\gamma_\ell^2}\right)^{j_\ell-1} \\ &= \prod_{\ell=1}^{d} (1-\omega_{\gamma_\ell})\, \omega_{\gamma_\ell}^{j_\ell-1},\end{aligned}$$

where ω_γ is defined above in (10), and

$$\eta_{d,\boldsymbol{\gamma},\boldsymbol{j}} = \prod_{\ell=1}^{d} \sqrt{\tilde{\lambda}_{\gamma_\ell,j_\ell}}\, \varphi_{\gamma_\ell,j_\ell}$$

with

$$\langle \eta_{d,\boldsymbol{\gamma},\boldsymbol{i}}, \eta_{d,\boldsymbol{\gamma},\boldsymbol{j}}\rangle_{H_d} = \langle \varphi_{\boldsymbol{\gamma},\boldsymbol{i}}, \varphi_{\boldsymbol{\gamma},\boldsymbol{j}}\rangle_{\mathcal{L}_2} = \delta_{\boldsymbol{ij}}.$$

This section ends with a lemma describing the convergence of the sums of powers of the eigenvalues for the multivariate problem, and how these sums depend on the dimension, d. This lemma is used in several of the theorems on convergence and tractability in the following sections.

In the next sections, it will be convenient to reorder the sequence of eigenvalues $\{\tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}}\}_{\boldsymbol{j}\in\mathbb{N}^d}$ as the sequence $\{\lambda_{d,j}\}_{j\in\mathbb{N}}$ with $\lambda_{d,1} \geq \lambda_{d,2} \geq \cdots$. Obviously, for the univariate case, $d = 1$, we have $\lambda_{1,j} = \tilde{\lambda}_{1,\gamma_1,j}$ for all $j \in \mathbb{N}$, but for the multivariate case, $d > 1$, the correspondence between $\lambda_{d,j}$ and $\tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}}$ is more complex. Obviously,

$$\lambda_{d,1} = \prod_{\ell=1}^{d} (1-\omega_{\gamma_\ell})\,.$$

We now present a simple estimate of $\lambda_{d,n+1}$ that will be needed for our analysis.

Lemma 1. *Let $\tau > 0$. Consider the Gaussian kernel with the sequence of shape parameters $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell\in\mathbb{N}}$. The sum of the τ^{th} power of the eigenvalues for the d-variate case, $d \geq 1$, is*

$$\sum_{j=1}^{\infty} \lambda_{d,j}^{\tau} = \sum_{\boldsymbol{j}\in\mathbb{N}^d} \tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}}^{\tau} = \prod_{\ell=1}^{d} \left(\sum_{j=1}^{\infty} \tilde{\lambda}_{\gamma_\ell,j}^{\tau} \right) = \prod_{\ell=1}^{d} \frac{(1-\omega_{\gamma_\ell})^{\tau}}{1-\omega_{\gamma_\ell}^{\tau}} \begin{cases} > 1, & 0 < \tau < 1, \\ = 1, & \tau = 1. \end{cases} \tag{13}$$

The $(n+1)^{st}$ largest eigenvalue satisfies

$$\lambda_{d,n+1} \leq \frac{1}{(n+1)^{1/\tau}} \prod_{\ell=1}^{d} \frac{1-\omega_{\gamma_\ell}}{(1-\omega_{\gamma_\ell}^{\tau})^{1/\tau}}. \tag{14}$$

Proof. Equation (13) follows directly from the formula for $\tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}}$ in (12). From the definition of ω_γ in (10) it follows that $0 < \omega_\gamma < 1$ for all $\gamma > 0$. For $\tau \in (0,1)$, consider the function

$$f(\omega) = (1-\omega)^{\tau} - 1 + \omega^{\tau} \quad \text{for all} \quad \omega \in [0,1].$$

Clearly, f is concave and vanishes at 0 and 1, and therefore $f(\omega) > 0$ for all $\omega \in (0,1)$. This yields the lower bound on the sum of the power of the univariate eigenvalues.

The ordering of the eigenvalues $\lambda_{d,j}$ implies that

$$\lambda_{d,n+1} \leq \left(\frac{1}{n+1} \sum_{j=1}^{n+1} \lambda_{d,j}^{\tau} \right)^{1/\tau} \leq \left(\frac{1}{n+1} \sum_{j=1}^{\infty} \lambda_{d,j}^{\tau} \right)^{1/\tau} = \frac{1}{(n+1)^{1/\tau}} \left(\sum_{j=1}^{\infty} \lambda_{d,j}^{\tau} \right)^{1/\tau}.$$

This yields the upper bound on the $n+1^{\text{st}}$ largest eigenvalue in (14), and completes the proof. □

The main point of (14) is that this estimate holds for all positive τ. This means that $\lambda_{d,n+1}$ goes to zero faster than any polynomial in $(n+1)^{-1}$.

4. Rates of Convergence for Gaussian Kernels

In this section we consider the function approximation problem for the Hilbert space $H_d = H(K_d)$ with the anisotropic Gaussian kernel given by (1b). We stress that the sequence $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell=1}^{\infty}$ of shape parameters can be arbitrary. In particular, we may consider the isotropic case for which all $\gamma_\ell = \gamma > 0$.

We want to verify how fast the minimal errors $e^{\text{wor-all}}(n, H_d)$ and $e^{\text{wor-std}}(n, H_d)$ go to zero, and what the rate of convergence of these sequences is, see (6).

Theorem 1.

$$r^{\text{wor-all}}(H_d) = r^{\text{wor-std}}(H_d) = \infty.$$

Proof. For the class Λ^{all} we know that $e^{\text{wor-all}}(n, H_d) = \sqrt{\lambda_{d,n+1}}$, where $\lambda_{d,n+1}$ is the $(n+1)^{\text{st}}$ largest eigenvalue of W_d. Lemma 1 demonstrates that $\lambda_{d,n+1}$ is proportional to $(n+1)^{-1/\tau}$ times a dimension dependent constant. This implies that $r^{\text{wor-all}}(H_d) \geq 1/(2\tau)$ and since τ can be arbitrarily small, we conclude that

$$r^{\text{wor-all}}(H_d) = \infty,$$

as claimed.

Consider now the class $\Lambda^{\rm std}$. We use Theorem 5 from [10], which states that if there exist numbers $p > 1$ and B such that

$$\lambda_{d,n} \le B\, n^{-p} \quad \text{for all} \quad n \in \mathbb{N} \tag{15}$$

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

$$e^{\rm wor}(A_n) \le B\, C_{\delta,p}\, (n+1)^{-(1-\delta)\, p^2/(2p+2)}.$$

Here, $C_{\delta,p}$ is independent of n and d and depends only on δ and p.

Note that assumption (15) holds in our case for an arbitrarily large p with B that can depend on d. Hence, $r^{\rm wor-std}(H_d) \ge (1-\delta)\, p^2/(2p+2)$, and since δ can be arbitrarily small and p can be arbitrarily large we conclude

$$r^{\rm wor-std}(H_d) = \infty,$$

as claimed. This completes the proof. □

We stress that the algorithm A_n that was used in the proof is non-constructive. However, there are known algorithms that use only function values and whose worst case error goes to zero like n^{-p} for an arbitrary large p. In fact, given a design, it is known that the spline algorithm is the best way to use the function data given via that design. Thus, the search for an algorithm with optimal convergence rates focuses on the choice of a good design. One such design was proposed by Smolyak already in 1963, see [15], and today it is usually referred to as a sparse grid, see [3] for a survey. An associated algorithm from which this design naturally arises is Smolyak's algorithm. The essence of this algorithm is to use a certain tensor product of univariate algorithms. Then, if the univariate algorithm has the worst case error of order n^{-p}, the worst case error for the d-variate case is also of order n^{-p} modulo some powers of $\ln n$, see e.g., [21].

Theorem 1 states that as long as only the rate of convergence is considered, the function approximation problem for Hilbert spaces with Gaussian kernels is easy. In fact, it is not surprising since functions of this class are very smooth. However, the rate of convergence tells us nothing about the dependence on d. As long as d is small the dependence on d is irrelevant. But if d is large we want to check the dependence on d. We are especially afraid of an exponential dependence on d which is called after Bellman the *curse of dimensionality*. It also may happen that we have a tradeoff between the rate of convergence and dependence on d. Furthermore, the results may now depend on the weights γ_ℓ. This is the subject of our next sections.

5. Tractability for the Absolute Error Criterion

As in the previous section we consider the function approximation problem for Hilbert spaces $H_d = H(K_d)$ with a Gaussian kernel. We now consider the absolute error criterion and we want to verify whether polynomial tractability holds. Let us recall that we study the minimal number of functionals from the class $\Lambda^{\rm all}$ or $\Lambda^{\rm std}$ needed to guarantee a worst case error of at most ε,

$$n^{\rm wor-abs-\vartheta}(\varepsilon, H_d) = \min\left\{ n \mid \ e^{\rm wor-\vartheta}(n, H_d) \le \varepsilon \right\}, \qquad \vartheta \in \{\rm std, all\}.$$

5.1. **Arbitrary Linear Functionals.**
We first analyze the class $\Lambda^{\rm all}$ and polynomial tractability.

Theorem 2. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with isotropic or anisotropic Gaussian kernels with arbitrary positive γ_ℓ for the class $\Lambda^{\rm all}$ and the absolute error criterion. Then*

- *$\mathcal{I}$ is strongly polynomially tractable with exponent of strong polynomial tractability at most 2. For all $d \in \mathbb{N}$ and $\varepsilon \in (0,1)$ we have*

$$\begin{aligned} e^{\rm wor-all}(n, H_d) &\le (n+1)^{-1/2}, \\ n^{\rm wor-abs-all}(\varepsilon, H_d) &\le \varepsilon^{-2}. \end{aligned}$$

- *For the isotropic Gaussian kernel the exponent of strong tractability is 2, so that the bound above is best possible in terms of the exponent of ε^{-1}. Furthermore strong polynomial tractability is equivalent to polynomial tractability.*

Proof. We use Theorem 5.1 from [11]. This theorem says that $\mathcal{I}$ is strongly polynomially tractable iff there exist two positive numbers C_1 and τ such that

$$C_2 := \sup_{d\in\mathbb{N}} \left(\sum_{j=\lceil C_1 \rceil}^{\infty} \lambda_{d,j}^{\tau} \right)^{1/\tau} < \infty.$$

If so, then

$$n^{\rm wor-abs-all}(\varepsilon, H_d) \le (C_1 + C_2^{\tau})\, \varepsilon^{-2\tau} \quad \text{for all} \quad d \in \mathbb{N} \ \text{ and } \ \varepsilon \in (0,1).$$

Furthermore, the exponent of strong polynomial tractability is

$$p^{\rm all} = \inf\{2\tau \mid \ \tau \text{ for which } C_2 < \infty\}.$$

Let $\tau = 1$. Then, by (9) it follows that no matter what the weights γ_ℓ are, we can take an arbitrarily small C_1 so that $\lceil C_1 \rceil = 1$ and $C_2 = 1$ as well as $n^{\rm wor-abs-all}(\varepsilon, H_d) \le (C_1 + 1)\, \varepsilon^{-2}$. For C_1 tending to zero, we conclude the bound

$$n^{\rm wor-abs-all}(\varepsilon, H_d) \le \varepsilon^{-2}.$$

Furthermore, by (14) in Lemma 1 it follows that

$$e^{\rm wor-all}(n, H_d) = \sqrt{\lambda_{d,n+1}} \le (n+1)^{-1/2},$$

as claimed.

Assume now the isotropic case, i.e., $\gamma_\ell = \gamma$ for all $j \in \mathbb{N}$. Then for any positive C_1 and τ we use Lemma 1 and obtain

$$\begin{aligned}
\sum_{j=\lceil C_1 \rceil}^{\infty} \lambda_{d,j}^\tau &= \sum_{j=1}^{\infty} \lambda_{d,j}^\tau - \sum_{j=1}^{\lceil C_1 \rceil - 1} \lambda_{d,j}^\tau \\
&= \left(\frac{(1-\omega_\gamma)^\tau}{1-\omega_\gamma^\tau} \right)^d - \sum_{j=1}^{\lceil C_1 \rceil - 1} \lambda_{d,j}^\tau \\
&\geq \left(\frac{(1-\omega_\gamma)^\tau}{1-\omega_\gamma^\tau} \right)^d - (\lceil C_1 \rceil - 1)\, \lambda_{d,1}^\tau \\
&= \left(\frac{(1-\omega_\gamma)^\tau}{1-\omega_\gamma^\tau} \right)^d - (\lceil C_1 \rceil - 1)\, (1-\omega_\gamma)^{\tau\, d}.
\end{aligned}$$

For $\tau \in (0,1)$, we know from Lemma 1 that $(1-\omega_\gamma)^\tau/(1-\omega_\gamma^\tau) > 1$, and therefore the last expression goes exponentially fast to infinity with d. This proves that $C_2 = \infty$ for all $\tau \in (0,1)$. Hence, the exponent of strong tractability is two.

Finally, to prove that strong polynomial tractability is equivalent to polynomial tractability, it is enough to show that polynomial tractability implies strong polynomial tractability. From Theorem 5.1 of [11] we know that polynomial tractability holds iff 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

$$n^{\mathrm{wor-abs-all}}(\varepsilon, H_d) \leq (C_1 + C_2^\tau)\, d^{\max(q_1, q_2\tau)}\, \varepsilon^{-2\tau}$$

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

$$d^{-q_2\tau} \left(\frac{(1-\omega_\gamma)^\tau}{1-\omega_\gamma^\tau} \right)^d - d^{-q_2\tau}\, (\lceil C_1 \rceil - 1)\, (1-\omega_\gamma)^{\tau\, d} \leq C_2^\tau < \infty.$$

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

We now compare Theorems 1 and 2. Theorem 1 says that for any p we have

$$e^{\mathrm{wor-all}}(n, H_d) = \mathcal{O}(n^{-p})$$

but the factor in the big $\mathcal{O}$ notation may depend on d. In fact, from Theorem 2 we conclude that, indeed, for the isotropic case it depends more than polynomially on d for all $p > 1/2$. Hence, the good rate of convergence does not necessarily mean much for large d.

The exponent of strong polynomial tractability is 2 for the isotropic case. We now check how the exponent of strong polynomial tractability depends on the sequence $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell \in \mathbb{N}}$ of shape parameters. The determining factor is the quantity $r(\boldsymbol{\gamma})$ introduced in (2), which measures the rate of decay of the shape parameter sequence.

Theorem 3. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with isotropic or anisotropic Gaussian kernels for the class $\Lambda^{\rm all}$ and the absolute error criterion. Let $r(\boldsymbol{\gamma})$ be the rate of decay of shape parameters. Then*

- *$\mathcal{I}$ is strongly polynomially tractable with exponent*
$$p^{\rm all} = \min\left(2, \frac{1}{r(\boldsymbol{\gamma})}\right) \le 2.$$
- *For all $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ we have*
$$\begin{aligned} e^{\rm wor-all}(n, H_d) &= \mathcal{O}\left(n^{-1/p^{\rm all}+\delta}\right) = \mathcal{O}\left(n^{-\max(r(\boldsymbol{\gamma}),1/2)+\delta}\right), \\ n^{\rm wor-abs-all}(\varepsilon, H_d) &= \mathcal{O}\left(\varepsilon^{-(p^{\rm all}+\delta)}\right), \end{aligned}$$
where the factors in the big $\mathcal{O}$ notation do not depend on d and ε^{-1} but may depend on δ.
- *Furthermore, in the case of ordered shape parameters, i.e., $\gamma_1 \ge \gamma_2 \ge \cdots$ if*
$$n^{\rm wor-abs-all}(\varepsilon, H_d) = \mathcal{O}\left(\varepsilon^{-p}\, d^{\,q}\right) \quad \text{for all } \varepsilon \in (0,1) \text{ and } d \in \mathbb{N},$$
then $p \ge p^{\rm all}$, which means that strong polynomial tractability is equivalent to polynomial tractability.

Proof. As in the proof of Theorem 2, $\mathcal{I}$ is strongly polynomially tractable iff there exist two positive numbers C_1 and τ such that
$$C_2 := \sup_{d\in\mathbb{N}} \left(\sum_{j=\lceil C_1\rceil}^{\infty} \lambda_{d,j}^{\tau}\right)^{1/\tau} < \infty.$$
Furthermore, the exponent $p^{\rm all}$ of strong polynomial tractability is the infimum of 2τ for which this condition holds. Proceeding similarly as before, we have
$$\sum_{j=\lceil C_1\rceil}^{\infty} \lambda_{d,j}^{\tau} \le \sum_{j=1}^{\infty} \lambda_{d,j}^{\tau} = \prod_{\ell=1}^{\infty} \frac{(1-\omega_{\gamma_\ell})^{\tau}}{1-\omega_{\gamma_\ell}^{\tau}}$$
and since $\lambda_{d,j} < 1$
$$\sum_{j=\lceil C_1\rceil}^{\infty} \lambda_{d,j}^{\tau} \ge \sum_{j=1}^{\infty} \lambda_{d,j}^{\tau} - C_1 = \prod_{\ell=1}^{\infty} \frac{(1-\omega_{\gamma_\ell})^{\tau}}{1-\omega_{\gamma_\ell}^{\tau}} - C_1.$$
Therefore, $\mathcal{I}$ is strongly polynomially tractable iff there exists a positive τ such that
$$C_3 := \prod_{\ell=1}^{\infty} \frac{1-\omega_{\gamma_\ell}}{(1-\omega_{\gamma_\ell}^{\tau})^{1/\tau}} < \infty,$$
and the exponent $p^{\rm all}$ is the infimum of 2τ for which the last condition holds.

As we already know, this holds for $\tau = 1$. Take now $\tau \in (0,1)$. Since $(1-\omega_{\gamma_\ell})/(1-\omega_{\gamma_\ell}^{\tau})^{1/\tau} > 1$ then $C_3 < \infty$ implies that
$$\lim_{\ell\to\infty} \frac{1-\omega_{\gamma_\ell}}{(1-\omega_{\gamma_\ell}^{\tau})^{1/\tau}} = 1.$$

Taking into account (10), it is easy to check that the last condition is equivalent to

$$\lim_{\ell\to\infty} \omega_{\gamma_\ell} = \lim_{\ell\to\infty} \gamma_\ell^2 = 0.$$

Furthermore, $C_3 < \infty$ implies that

$$\sum_{\ell=1}^{\infty} \gamma_\ell^{2\tau} < \infty,$$

and $r(\boldsymbol{\gamma}) \geq 1/(2\tau) > 1/2$. Hence, $p^{\rm all} < 2$ only if $r(\gamma) > 1/2$. On the other hand, $2\tau \geq 1/r(\boldsymbol{\gamma})$ and therefore $p^{\rm all} \geq 1/r(\boldsymbol{\gamma})$. This establishes the formula for $p^{\rm all}$. The estimates on $e^{\rm wor-all}(n, H_d)$ and $n^{\rm wor-abs-all}(\varepsilon, H_d)$ follow from the definition of strong tractability.

Assume now polynomial tractability with $p < 2$ and an arbitrary q. Then $\lambda_{d,n+1} \leq \varepsilon^2$ for $n = \mathcal{O}(\varepsilon^{-p} d^q)$. Hence,

$$\lambda_{d,n+1} = \mathcal{O}(d^{2q/p}(n+1)^{-2/p}).$$

This implies

$$\prod_{j=1}^{d} \frac{(1-\omega_{\gamma_\ell})^\tau}{1-\omega_{\gamma_\ell}^\tau} = \sum_{\ell=1}^{\infty} \lambda_{d,\ell}^\tau = \mathcal{O}(d^{2q\tau/p}) \quad \text{for all} \quad 2\tau > p.$$

For $\tau < 1$, this yields

$$\exp\left(\sum_{\ell=1}^{d} \gamma_\ell^{2\tau}\right) = \mathcal{O}(d^{2q\tau/p}).$$

Therefore

$$\limsup_{\ell\to\infty} \frac{\sum_{\ell=1}^{d} \gamma_\ell^{2\tau}}{\ln\, d} < \infty.$$

Since the γ_ℓ's are ordered, we have

$$\frac{d\gamma_d^{2\tau}}{\ln\, d} \leq \frac{\sum_{\ell=1}^{d} \gamma_\ell^{2\tau}}{\ln\, d} = \mathcal{O}(1),$$

and $\gamma_d = \mathcal{O}((\ln(d)/d)^{1/(2\tau)})$. Hence, $r(\boldsymbol{\gamma}) \geq 1/(2\tau)$ and $r(\boldsymbol{\gamma}) \geq 1/p$. This means that $2 > p \geq 1/r(\boldsymbol{\gamma}) = p^{\rm all}$, as claimed. □

It is interesting to notice that the last part of Theorem 3 does not hold, in general, for unordered shape parameters. Indeed, for $s > 1/2$, take

$$\begin{aligned} \gamma_{a_k} &= 1 \quad \text{for all natural } k \text{ with } \ a_k = 2^{2^k}, \\ \gamma_\ell &= \frac{1}{\ell^s} \quad \text{for all natural } \ell \text{ not equal to } a_k. \end{aligned}$$

Then strong polynomial tractability holds with the exponent 2 since $C_3 = \infty$ in the proof of Theorem 3 for all $\tau < 1$. On the other hand, we have polynomial tractability

with $p = 1/s < 2$ and q arbitrarily close to $1/(2s)$. Indeed, for $\tau = 1/(2s)$ and $q_1 = 0$ and $q_2 > 1$ we have

$$\begin{aligned} d^{-q_2} \sum_{\ell=1}^{\infty} \lambda_{d,\ell}^{\tau} &= d^{-q_2} \prod_{\ell}^{d} \frac{(1-\omega_{\gamma_\ell})^\tau}{1-\omega_{\gamma_\ell}} \\ &= d^{-q_2} \left(\frac{1-\omega_1)^\tau}{1-\omega_1} \right)^{\mathcal{O}(1)+\ln \ln d} \qquad \mathcal{O}(d) < \infty. \end{aligned}$$

This implies that

$$n^{\text{wor}-\text{abs}-\text{all}}(\varepsilon, H_d) = \mathcal{O}\left(d^{\,q_2/(2s)}\, \varepsilon^{-1/s}\right).$$

Theorem 3 states that the exponent of strong polynomial tractability is 2 for all shape parameters for which $r(\boldsymbol{\gamma}) \le 1/2$. Only if $r(\boldsymbol{\gamma}) > 1/2$ is the exponent smaller than 2. Again, although the rate of convergence of $e^{\text{wor}-\text{all}}(n, H_d)$ is always excellent, the dependence on d is eliminated only at the expense of the exponent which must be roughly $1/p^{\text{all}}$. Of course, if we take an exponentially decaying sequence of shape parameters, say, $\gamma_\ell = q^\ell$ for some $q \in (0,1)$, then $r(\boldsymbol{\gamma}) = \infty$ and $p^{\text{all}} = 0$. In this case, we have an excellent rate of convergence without any dependence on d.

Although Theorem 2 is for Gaussian kernels, it is easy to extend this theorem for other positive definite translation invariant or radially symmetric kernels. Indeed, for translation invariant kernels the only difference is that for $\tau = 1$ the sum of the eigenvalues is not necessarily one but

$$\sum_{j=1}^{\infty} \lambda_{d,j} = \widetilde{K}_d(\mathbf{0}).$$

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

$$e^{\text{wor}-\text{all}}(n, H_d) \le \left[\frac{\widetilde{K}_d(\mathbf{0})}{n+1} \right]^{1/2} \quad \text{and} \quad n^{\text{wor}-\text{abs}-\text{all}}(n, H_d) \le \widetilde{K}_d(\mathbf{0})\, \varepsilon^{-2}.$$

Tractability then depends on how $\widetilde{K}_d(\mathbf{0})$ depends on d. In particular, it is easy to check the following facts.

- If
$$\sup_{d \in \mathbb{N}} \widetilde{K}_d(\mathbf{0}) < \infty$$
then we have strong polynomial tractability with exponent at most 2, i.e.,
$$n^{\text{wor}-\text{all}}(n, H_d) = \mathcal{O}\left(\varepsilon^{-2}\right).$$
- If there exists a nonnegative q such that
$$\sup_{d \in \mathbb{N}} \widetilde{K}_d(\mathbf{0})\, d^{-q} < \infty$$
then we have polynomial tractability and
$$n^{\text{wor}-\text{all}}(n, H_d) = \mathcal{O}\left(d^{\,q}\, \varepsilon^{-2}\right).$$

- If
$$\lim_{d\to\infty} \frac{\ln \max(\widetilde{K}_d(\mathbf{0}), 1)}{d} = 0$$
then we have weak tractability.

For radially symmetric kernels, the situation is even simpler since

$$\sum_{j=1}^{\infty} \lambda_{d,j} = \kappa(0),$$

and it does not depend on d. Hence,

$$e^{\mathrm{wor-all}}(n, H_d) \le \left[\frac{\kappa(0)}{n+1}\right]^{1/2} \quad \text{and} \quad n^{\mathrm{wor-abs-all}}(n, H_d) \le \kappa(0)\,\varepsilon^{-2},$$

and we have strong polynomial tractability with exponent at most 2.

Extending Theorem 3 to arbitrary stationary or isotropic kernels is not so straightforward. To achieve smaller strong tractability exponents than 2 one needs to know the sum of the powers of eigenvalues, and their dependence on d. One would suspect, as is the case for Gaussian kernels, that some sort of anisotropy is needed to obtain better strong tractability exponents than 2.

5.2. **Only Function Values.**

We now turn to the class Λ^{std} and prove the following theorem.

Theorem 4. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with isotropic or anisotropic Gaussian kernels for the class Λ^{std} and the absolute error criterion. 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*
$$\begin{aligned} e^{\mathrm{wor-std}}(n, H_d) &\le \frac{\sqrt{2}}{n^{1/4}} \left(1 + \frac{1}{2\sqrt{n}}\right)^{1/2}, \\ n^{\mathrm{wor-abs-std}}(\varepsilon, H_d) &\le \left\lceil \frac{(1+\sqrt{1+\varepsilon^2})^2}{\varepsilon^4} \right\rceil. \end{aligned}$$
- *For the isotropic Gaussian kernel the exponent of strong tractability is at least 2. Furthermore strong polynomial tractability is equivalent to polynomial tractability.*

Proof. We now use Theorem 1 from [22]. This theorem says that

$$e^{\mathrm{wor-std}}(n, H_d) \le \min_{k=0,1,\ldots} \left([e^{\mathrm{wor-all}}(k, H_d)]^2 + \frac{k}{n} \right)^{1/2}. \tag{16}$$

Taking $k = \lceil n^{-1/2} \rceil$ and remembering that $e^{\mathrm{wor-all}}(k, H_d) \le k^{-1/2}$ we obtain

$$e^{\mathrm{wor-std}}(n, H_d) \le \left(\frac{1}{\sqrt{n}} + \frac{1+\sqrt{n}}{n} \right)^{1/2} = \frac{\sqrt{2}}{n^{1/4}} \left(1 + \frac{1}{2\sqrt{n}}\right)^{1/2},$$

as claimed. Solving $e^{\mathrm{wor-std}}(n, H_d) \le \varepsilon$, we obtain the bound on $n^{\mathrm{wor-abs-std}}(\varepsilon, H_d)$.

For the isotropic case, we know from Theorem 2 that the exponent of strong tractability for the class $\Lambda^{\rm all}$ is 2. For the class $\Lambda^{\rm std}$, the exponent cannot be smaller.

Finally, assume that we have polynomial tractability for the class $\Lambda^{\rm std}$. Then we also have polynomial tractability for the class $\Lambda^{\rm all}$. From Theorem 2 we know that then strong tractability for the class $\Lambda^{\rm all}$ holds. Furthermore we know that the exponent of strong tractability is 2 and $n^{\rm wor-abs-all}(\varepsilon, H_d) \le \varepsilon^{-2}$. As above, we then get strong tractability also for $\Lambda^{\rm std}$ with the exponent at most 4. This completes the proof. □

We do not know if the error bound of order $n^{-1/4}$ is sharp for the class $\Lambda^{\rm std}$. We suspect that it is *not* sharp and that maybe even an error bound of order $n^{-1/2}$ holds for the class $\Lambda^{\rm std}$ exactly as for the class $\Lambda^{\rm all}$.

For fast decaying shape parameters it is possible to improve Theorem 4. This is the subject of our next theorem.

Theorem 5. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with isotropic or anisotropic Gaussian kernels for the class $\Lambda^{\rm std}$ and the absolute error criterion. Let $r(\boldsymbol{\gamma}) > 1/2$. Then*

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

$$p^{\rm std} = \frac{1}{r(\boldsymbol{\gamma})} + \frac{1}{2\,r^2(\boldsymbol{\gamma})} = p^{\rm all} + \tfrac{1}{2}\left[p^{\rm all}\right]^2 < 4.$$

- *For all $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ we have*

$$\begin{aligned} e^{\rm wor-std}(n, H_d) &= \mathcal{O}\left(n^{-1/p^{\rm std}+\delta}\right) = \mathcal{O}\left(n^{-r(\boldsymbol{\gamma})/[1+1/(2r(\boldsymbol{\gamma}))]+\delta}\right), \\ n^{\rm wor-abs-std}(\varepsilon, H_d) &= \mathcal{O}\left(\varepsilon^{-(p^{\rm std}+\delta)}\right), \end{aligned}$$

  where the factors in the big $\mathcal{O}$ notation do not depend on d and ε^{-1} but may depend on δ.

Proof. For $r(\boldsymbol{\gamma}) > 1/2$, Theorem 3 for the class $\Lambda^{\rm all}$ states that the exponent of strong polynomial tractability is $p^{\rm all} = 1/r(\boldsymbol{\gamma})$. This means that for all $\eta \in (0,1)$ we have

$$\lambda_{d,n} = \mathcal{O}(n^{-2r(\boldsymbol{\gamma})+\eta}),$$

with the factor in the big $\mathcal{O}$ notation independent of n and d but dependent on δ. Since $2r(\boldsymbol{\gamma}) > 1$, it follows that for all positive η small enough, $p = 2r(\boldsymbol{\gamma}) - \eta > 1$. Applying Theorem 5 from [10] as in the proof of Theorem 1, it follows that for any $\delta_1 \in (0,1)$ we have

$$\begin{aligned} e^{\rm wor-std}(n, H_d) &= \mathcal{O}\left(n^{-(1-\delta_1)p^2/(2p+2)}\right) = \mathcal{O}\left(n^{-(1-\delta_1)(1+\mathcal{O}(\eta))2w^2(\boldsymbol{\gamma})/(2r(\boldsymbol{\gamma})+1)}\right) \\ &= \mathcal{O}\left(n^{-1/(p^{\rm std}+\delta)}\right), \end{aligned}$$

again with the factor in the big $\mathcal{O}$ notation independent of n and d but dependent on δ. This leads to the estimates of the theorem. □

Note that for large $r(\boldsymbol{\gamma})$, 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, $\gamma_\ell = q^\ell$ for some $q \in (0,1)$, we have $p^{\rm all} = p^{\rm std} = 0$, and the rates of convergence are excellent and independent of d.

6. Tractability for the Normalized Error Criterion

We now consider the function approximation problem for Hilbert spaces $H_d(K_d)$ with a Gaussian kernel for the normalized error criterion. That is, we want to find the smallest n for which

$$e^{\rm wor-\vartheta}(n, H_d) \le \varepsilon \, \|I_d\|, \qquad \vartheta \in \{\rm std, all\}.$$

Note that $\|I_d\| = \sqrt{\lambda_{d,1}} \le 1$ and it can be exponentially small in d. Therefore the normalized error criterion may be much harder than the absolute error criterion and this is the reason for a number of negative results for this error criterion. It turns out that the isotropic and anisotropic cases are quite different and we will study them in separate subsections. We begin with the case where the data are generated by arbitrary linear functionals. The class $\Lambda^{\rm std}$ is partially covered at the end.

6.1. **Isotropic Case with Arbitrary Linear Functionals.**

For the isotropic case, $\gamma_\ell = \gamma > 0$, we have

$$\|I_d\| = \tilde{\lambda}_{\gamma,1}^{d/2} = (1-\omega_\gamma)^{d/2},$$

and since $\tilde{\lambda}_{\gamma,1} = 1 - \omega_\gamma < 1$, the norm of I_d is exponentially small. We are ready to present the following theorem.

Theorem 6. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with isotropic Gaussian kernels for the class $\Lambda^{\rm all}$ and for the normalized error criterion. Then*

- *$\mathcal{I}$ is not polynomially tractable,*
- *$\mathcal{I}$ is quasi-polynomially tractable with exponent*

$$t^{\rm all} = t^{\rm all}(\gamma) = \frac{2}{\ln \frac{1+2\gamma^2+\sqrt{1+4\gamma^2}}{2\gamma^2}}.$$

That is, for all $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ we have

$$\begin{aligned} e^{\rm wor-all}(n, H_d) &= \mathcal{O}\left(\|I_d\| \left(\frac{1}{n}\right)^{\frac{1}{(t^{\rm all}+\delta)\,(1+\ln\, d)}} \left(\frac{1}{\frac{1}{2}(1+\sqrt{1+4\gamma^2})+\gamma^2} \right)^{d/4} \right), \\ n^{\rm wor-nor-all}(\varepsilon, H_d) &= \mathcal{O}\left(\exp\left((t^{\rm all}+\delta)(1+\ln\, d)(1+\ln\, \varepsilon^{-1}) \right) \right), \end{aligned}$$

where the factors in the big $\mathcal{O}$ notations are independent of n, ε^{-1} and d but may depend on δ.

Proof. The lack of polynomial tractability follows, in particular, from Theorem 5.6 of [11]. In fact, the lack of polynomial tractability for the class $\Lambda^{\rm all}$ holds for all tensor product problems with two positive eigenvalues for the univariate case.

For quasi-polynomial tractability we use Theorem 3.3 of [7], which states that quasi-polynomial tractability for the class $\Lambda^{\rm all}$ holds for tensor product problems iff the rate

$$r = \sup\left\{ \beta \ge 0 \,|\, \lim_{n\to\infty} \tilde{\lambda}_{\gamma,n}\, n^{\beta} = 0 \right\}$$

of the univariate eigenvalues is positive and the second largest univariate eigenvalue $\tilde{\lambda}_{\gamma,2}$ is strictly less than the largest univariate eigenvalue $\tilde{\lambda}_{\gamma,1}$. If so, then the exponent of quasi-polynomial tractability is

$$t^{\,\rm all} = \max\left(\frac{2}{r}, \frac{2}{\ln\ \tilde{\lambda}_{\gamma,1}/\tilde{\lambda}_{\gamma,2}}\right).$$

In our case, $r = \infty$ and

$$t^{\rm all} = \frac{2}{\ln\ \tilde{\lambda}_{\gamma,1}/\tilde{\lambda}_{\gamma,2}} = \frac{2}{-\ln\ \omega_\gamma} = \frac{2}{\ln\ \frac{1+2\gamma^2+\sqrt{1+4\gamma^2}}{2\gamma^2}}.$$

The estimates of $e^{\rm wor-all}(n, H_d)$ and $n^{\rm wor-nor-all}(\varepsilon, H_d)$ follow from the definition of quasi-polynomial tractability. This completes the proof. □

For the isotropic case we lose polynomial tractability for the normalized error criterion although even strong polynomial tractability is present for the absolute error criterion. This shows qualitatively that the normalized error criterion is much harder. In this case we only have quasi-polynomial tractability. Observe that the exponent of quasi-polynomial tractability depends on γ and we have

$$\lim_{\gamma\to 0} t^{\rm all}(\gamma) = 0 \quad \text{and} \lim_{\gamma\to\infty} t^{\rm all}(\gamma) = \infty.$$

For some specific values of γ we have

$$\begin{aligned} t^{\rm all}(2^{-1/2}) &= 1.5186\ldots, \\ t^{\rm all}(1) &= 2.0780\ldots, \\ t^{\rm all}(2^{1/2}) &= 2.8853\ldots. \end{aligned}$$

6.2. **Anisotropic Case with Arbitrary Linear Functionals.**

We now consider the sequence $\{\gamma_\ell\}_{\ell\in\mathbb{N}}$ of shape parameters and ask when we can guarantee strong polynomial tractability. As we shall see, this holds for the class $\Lambda^{\rm all}$ if $r(\boldsymbol{\gamma}) > 0$ although the exponent of strong polynomial tractability is large for small $r(\boldsymbol{\gamma})$. More precisely, we have the following theorem, which is similar to Theorem 3.

Theorem 7. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with anisotropic Gaussian kernels for the class $\Lambda^{\rm all}$ and for the normalized error criterion. Then*

- *$\mathcal{I}$ is strongly polynomially tractable if $r(\boldsymbol{\gamma}) > 0$. If so, then the exponent is*

$$p^{\rm all} = \frac{1}{r(\boldsymbol{\gamma})}.$$

- *Let $r(\boldsymbol{\gamma}) > 0$. Then for all $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ we have*

$$\begin{aligned} e^{\mathrm{wor-all}}(n, H_d) &= \mathcal{O}\left(\|I_d\| n^{-1/p^{\mathrm{all}}+\delta}\right) = \mathcal{O}\left(n^{-r(\boldsymbol{\gamma})+\delta}\right), \\ n^{\mathrm{wor-nor-all}}(\varepsilon, H_d) &= \mathcal{O}\left(\varepsilon^{-(p^{\mathrm{all}}+\delta)}\right), \end{aligned}$$

  where the factors in the big $\mathcal{O}$ notations are independent of n, ε^{-1} and d but may depend on δ.
- *Furthermore, in the case of ordered shape parameters, i.e., $\gamma_1 \geq \gamma_2 \geq \cdots$ if*

$$n^{\mathrm{wor-nor-all}}(\varepsilon, H_d) = \mathcal{O}\left(\varepsilon^{-p}\, d^{\,q}\right) \quad \text{for all } \varepsilon \in (0,1) \text{ and } d \in \mathbb{N},$$

  then $p \geq p^{\mathrm{all}} = \frac{1}{r(\boldsymbol{\gamma})}$, which means that strong polynomial tractability is equivalent to polynomial tractability.

Proof. Theorem 5.2 of [11] states that strong polynomial tractability holds iff there exits a positive number τ such that

$$\tilde{C}_2 := \sup_d \sum_{j=1}^{\infty} \left(\frac{\lambda_{d,j}}{\lambda_{d,1}}\right)^{\tau} = \prod_{\ell=1}^{\infty} \frac{1}{1-\omega_{\gamma_\ell}^{\tau}} < \infty.$$

If so, then $n^{\mathrm{wor-nor-all}}(\varepsilon, H_d) \leq \tilde{C}_2\, \varepsilon^{-2\tau}$ 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 $\tilde{C}_2 < \infty$.

Clearly, $\tilde{C}_2 < \infty$ iff

$$\sum_{\ell=1}^{\infty} \omega_{\gamma_\ell}^{\tau} < \infty \quad \text{iff} \quad \sum_{\ell=1}^{\infty} \gamma_\ell^{2\tau} < \infty.$$

This holds iff $r(\boldsymbol{\gamma}) \geq 1/(2\tau) > 0$. This also proves that $p^{\mathrm{all}} = 1/r(\boldsymbol{\gamma})$. The estimates on $e^{\mathrm{wor-all}}(n, H_d)$ and $n^{\mathrm{wor-nor-all}}(\varepsilon, H_d)$ follow from the definition of strong tractability.

The case of polynomial tractability for ordered shape parameters follows analogously to the proof in Theorem 3. From Theorem 5.2 of [11], we know that the problem is polynomially tractable with $n^{\mathrm{wor-nor-all}}(\varepsilon, H_d) = \mathcal{O}\left(\varepsilon^{-2\tau}\, d^{\,q_2\tau}\right)$ iff

$$\tilde{C}_2 := \sup_{d \in \mathbb{N}} d^{-q_2} \left[\sum_{j=1}^{\infty} \left(\frac{\lambda_{d,j}}{\lambda_{d,1}}\right)^{\tau}\right]^{1/\tau} = d^{-q_2} \prod_{\ell=1}^{d} \frac{1}{(1-\omega_\ell^{\tau})^{1/\tau}} < \infty.$$

Proceeding as in the proof of Theorem 3, this can happen for ordered shape parameters only if $\tau \geq 1/(2r(\boldsymbol{\gamma}))$. Therefore, $p \geq p^{\mathrm{all}} = 1/r(\boldsymbol{\gamma})$, as claimed. □

The essence of Theorem 7 is that under the normalized error criterion strong polynomial and polynomial tractability for the class Λ^{all} requires that the shape parameters tend to zero polynomially fast so that $r(\boldsymbol{\gamma}) > 0$. This condition is stronger than what is required for the absolute error criterion.

It is interesting to compare strong polynomial tractability for the absolute and normalized error criteria for the class Λ^{all}, see Theorems 3 and 7. This is the subject of the next corollary.

Corollary 1. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with isotropic or anisotropic Gaussian kernels for the class $\Lambda^{\rm all}$. Let $r(\boldsymbol{\gamma})$ be the rate of convergence of shape parameters.*

- *Absolute error criterion:*
  $\mathcal{I}$ is always strongly polynomially tractable with exponent
$$p^{\rm all} = \min\left(2, \frac{1}{r(\boldsymbol{\gamma})}\right) \le 2.$$
- *Normalized error criterion:*
  $\mathcal{I}$ is strongly polynomially tractable iff $r(\boldsymbol{\gamma}) > 0$. If so, the exponent is
$$p^{\rm all} = \frac{1}{r(\boldsymbol{\gamma})}.$$

The strong tractability exponents under the two error criteria are the same provided that $r(\boldsymbol{\gamma}) \ge 1/2$.

6.3. **Only Function Values.**

We now turn to the class $\Lambda^{\rm std}$. We do not know if quasi-polynomial tractability holds for the class $\Lambda^{\rm std}$ in the isotropic case. The theorems that we used for the absolute error criterion are not enough for the normalized error criterion. Indeed, no matter how a positive k is defined in (16) we must take n exponentially large in d if we want to guarantee that the error is less than $\varepsilon\|I_d\|$. Similarly, if we use (15) then we must guarantee that $p > 1$ and this makes the number B exponentially large in d. We leave as an open problem whether quasi-polynomial tractability holds for the class $\Lambda^{\rm std}$.

We now discuss the initial error for $\lim_{\ell\to\infty}\gamma_\ell = 0$. We have
$$\|I_d\| = \prod_{\ell=1}^{d}(1-\omega_{\gamma_\ell})^{1/2} = \exp\left(\mathcal{O}(1) - \tfrac{1}{2}\sum_{\ell=1}^{d}\gamma_\ell^2\right).$$
For $r(\boldsymbol{\gamma}) \in [0, 1/2)$, the initial error still goes exponentially fast to zero, whereas for $r(\boldsymbol{\gamma}) = 1/2$ it may go to zero or be uniformly bounded from below by a positive number, and finally for $r(\boldsymbol{\gamma}) > 1/2$ it is always uniformly bounded from below by a positive number. For example, take $\gamma_\ell = \ell^{-\alpha}\ln^\beta(1+\ell)$ for a positive α and real β. Then $r(\boldsymbol{\gamma}) = \alpha$. For $\alpha = \frac{1}{2}$, the initial error goes to zero for $\beta > -\frac{1}{2}$, and is of order 1 if $\beta \le -\frac{1}{2}$.

This discussion shows that for $r(\boldsymbol{\gamma}) > 1/2$ there is really no difference between the absolute and normalized error criteria. This means that for $r(\boldsymbol{\gamma}) > 1/2$ we can apply Theorem 5 for the class $\Lambda^{\rm std}$ with ε replaced by $\varepsilon\|I_d\| = \Theta(\varepsilon)$. For $r(\boldsymbol{\gamma}) = 1/2$, Theorem 4 can be applied if we assume additionally that $\sum_{\ell=1}^{\infty}\gamma_\ell^2 < \infty$. The last assumption implies that $\|I_d\| = \Theta(1)$. We summarize this discussion in the following corollary.

Corollary 2. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with anisotropic Gaussian kernels for the class $\Lambda^{\rm std}$ and for the normalized*

error criterion. Assume that

$$r(\boldsymbol{\gamma}) > \tfrac{1}{2} \quad or \quad \left(\, r(\boldsymbol{\gamma}) = \tfrac{1}{2} \ and \ \sum_{\ell=1}^{\infty} \gamma_\ell^2 < \infty \, \right).$$

Then

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

$$p^{\rm std} = \frac{1}{r(\boldsymbol{\gamma})} + \frac{1}{2\, r^2(\boldsymbol{\gamma})} = p^{\rm all} + \tfrac{1}{2} \, \left[p^{\rm all}\right]^2 \le 4.$$

- *For all* $d \in \mathbb{N}$, $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$ *we have*

$$\begin{aligned} e^{\rm wor-all}(n, H_d) &= \mathcal{O}\left(n^{-1/(p^{\rm all}+\delta)}\right), \\ n^{\rm wor-nor-all}(\varepsilon, H_d) &= \mathcal{O}\left(\varepsilon^{-(p^{\rm all}+\delta)}\right), \end{aligned}$$

  where the factors in the big $\mathcal{O}$ *notations are independent of* n, ε^{-1} *and* d *but may depend on* δ.

The case $r(\boldsymbol{\gamma}) < 1/2$ is open. We do not know if polynomial tractability holds for the class $\Lambda^{\rm std}$ in this case.

References

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

[2] M. D. Buhmann, *Radial Basis Functions: Theory and Implementations*, Cambridge University Press, Cambridge, 2003.

[3] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numer., pp. 1–123, 2004,

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

[5] G. E. Fasshauer, *Meshfree Approximation Methods with* MATLAB, World Scientific Publishers, Singapore, 2007.

[6] A. I. J. Forrester, A. Sóbester and A. J. Keane, *Engineering Design via Surrogate Modelling: A Practical Guide*, Wiley, Chichester, 2008.

[7] M. Gnewuch and H. Woźniakowski, Quasi-polynomial tractability, submitted for publication.

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

[9] SAS Institute, *JMP 8.0*, 2009.

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

[11] E. Novak and H. Woźniakowski, *Tractability of Multivariate Problems*, Volume I: Linear Information, European Math. Soc., Zürich, 2008.

[12] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006 (online version at `http://www.gaussianprocess.org/gpml/`).

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

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

[15] S. A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, *Dokl. Akad. Nauk SSSR* **4**, 240–243, 1963.

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

[17] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer Verlag, Berlin, 2008.

[18] I. Steinwart, D. Hush and C. Scovel, An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels, IEEE Transactions on Information Theory, vol. 52, no.10, 4635–4663, 2006.
[19] J. F. Traub, G. W. Wasilkowski and H. Woźniakowski, *Information-Based Complexity*, Academic Press, 1988.
[20] G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.
[21] G. W. Wasilkowski and H. Woźniakowski, Explicit error bounds of algorithms for multivariate tensor product problems, *J. Complexity* , 11, 1–56, 1995.
[22] G. W. Wasilkowski and H. Woźniakowski, On the power of standard information for weighted approximation, *Found. Comput. Math.* 1, 417–434, 2001,
[23] H. Wendland, *Scattered data approximation*, Cambridge Monographs on Applied and Computational Mathematics, no.17, Cambridge University Press, Cambridge, 2005.

Department of Applied Mathematics, Illinois Institute of Technology, Room E1-208, 10 W. 32ND Street, Chicago, IL 60616
E-mail address: fasshauer@iit.edu

Department of Applied Mathematics, Illinois Institute of Technology, Room E1-208, 10 W. 32ND Street, Chicago, IL 60616
E-mail address: hickernell@iit.edu

Department of Computer Science, Columbia University, New York, NY 10027, and Institute of Applied Mathematics, University of Warsaw, ul. Banacha 2, 02-097 Warsaw, Poland
E-mail address: henryk@cs.columbia.edu