SIAM J. Numer. Anal.
Vol. 50, No. 1, pp. 247–271

ON DIMENSION-INDEPENDENT RATES OF CONVERGENCE FOR FUNCTION APPROXIMATION WITH GAUSSIAN KERNELS[*]

GREGORY E. FASSHAUER[†], FRED J. HICKERNELL[†], AND HENRYK WOŹNIAKOWSKI[‡]

Abstract. This article studies the problem of approximating functions belonging to a Hilbert space $\mathcal{H}_d$ with an isotropic or anisotropic translation invariant (or stationary) reproducing kernel with special attention given to the Gaussian kernel $K_d(\boldsymbol{x},\boldsymbol{t}) = \exp\big(-\sum_{\ell=1}^d \gamma_\ell^2 (x_\ell - t_\ell)^2\big)$ for all $\boldsymbol{x},\boldsymbol{t} \in \mathbb{R}^d$. The isotropic (or radial) case corresponds to using the same shape parameters for all coordinates, i.e., $\gamma_\ell = \gamma > 0$ for all ℓ, whereas the anisotropic case corresponds to varying γ_ℓ. The approximation error of the optimal approximation algorithm, called a meshfree or kriging method, is known to decay faster than any polynomial in n^{-1}, *for fixed d*, where n is the number of data points. We are especially interested in moderate to large d, which in particular arise in the construction of surrogates for computer experiments. This article presents dimension-independent error bounds, i.e., the error is bounded by Cn^{-p}, where C and p are *independent of both d and n*. This is equivalent to strong polynomial tractability. The pertinent error criterion is the worst case of such an algorithm over the unit ball in $\mathcal{H}_d$, with the error for a single function given by the $\mathcal{L}_2$ norm whose weight is also a Gaussian which is used to "localize" $\mathbb{R}^d$. We consider two classes of algorithms: (i) using data generated by finitely many arbitrary linear functionals, and (ii) using only finitely many function values. Provided that arbitrary linear functional data is available, we show $p = 1/2$ is possible for any translation invariant positive definite kernel. We also consider the sequence of shape parameters γ_d decaying to zero like $d^{-\omega}$ as d tends to ∞. Note that for large ω this means that the function to be approximated is "essentially low-dimensional." Then the largest p is roughly $\max(1/2,\omega)$. If only function values are available, dimension-independent convergence rates are somewhat worse. If the goal is to make the error smaller than Cn^{-p} *times the initial* $(n = 0)$ *error*, then the corresponding dimension-independent exponent p is roughly ω. In particular, for the isotropic case, when $\omega = 0$, the error does not even decay polynomially with n^{-1}. In summary, excellent dimension-independent error decay rates are possible only when the sequence of shape parameters decays rapidly.

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

AMS subject classifications. 65D15, 68Q17, 41A25, 41A63

DOI. 10.1137/10080138X

1. Introduction.

This article addresses the problem of function approximation. In a typical application we are given data of the form $y_i = f(\boldsymbol{x}_i)$ or $y_i = L_i(f)$ for $i = 1,\dots,n$. That is, a function f is sampled at the locations $\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}$, 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 $\mathbb{R}^d$. The goal is to construct $A_n(f)$, a good approximation to f that is inexpensive to evaluate.

An important example is the field of computer experiments, where each datum, y_i, may be the output of some computer code implementing a realistic model of a complex system, which takes hours or days to run. The approximation, $A_n(f)$, based on a modest number of runs, n, is used as a surrogate to explore the function at values

[*] Received by the editors July 7, 2010; accepted for publication (in revised form) November 3, 2011; published electronically February 9, 2012.
http://www.siam.org/journals/sinum/50-1/80138.html

[†] Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616 (fasshauer@iit.edu, hickernell@iit.edu). The work of these authors was supported in part by NSF-DMS grants 0713848 and 1115392. The second author's work was supported by Department of Energy grant SC0002100.

[‡] 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 (henryk@cs.columbia.edu). This author's work was supported by NSF-DMS grant 0914345.

of $\boldsymbol{x}$ other than the data sites. The number of different inputs into the computer code, d, may be a dozen or more, so it is important to understand the error of $A_n(f)$ for moderate or large values of d. This is the aim of this article.

Algorithms for function approximation based on symmetric, positive definite kernels have arisen in both the numerical computation literature [2, 5, 19, 31] and the statistical learning literature [1, 4, 10, 17, 20, 22, 23, 27]. They are often used in engineering applications [7] like the one just described. These algorithms go by a variety of names, including radial basis function methods [2], scattered data approximation [31], meshfree methods [5], (smoothing) splines [27], kriging [22], Gaussian process models [17], and support vector machines [23]. As evidence of the popularity of these methods, we note that the commercial statistical software JMP [12] has a Gaussian process modeling module implementing the algorithm that uses function values.

Given the choice of a symmetric, positive definite kernel $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (see (2.1) below for the specific requirements), there is an associated Hilbert space, $\mathcal{H}_d = \mathcal{H}(K_d)$, of functions defined on $\mathbb{R}^d$ for which K_d is the reproducing kernel. The spline algorithm, $S_n(f)$, described below in (2.6), chooses the element in $\mathcal{H}(K_d)$ that interpolates the data and has minimum $\mathcal{H}(K_d)$ norm. The spline algorithm is linear in the data and can be computed by solving an $n \times n$ system of linear equations. If the data are chosen as n optimal linear functionals, then the cost of computing $S_n(f)(\boldsymbol{x})$ for one $\boldsymbol{x}$ is equal to $2n-1$ arithmetic operations plus the cost of computing these n optimal linear functionals. A wider discussion of the cost of the algorithm is given at the end of section 2.1. It is well known that the spline algorithm is the optimal approximation to functions in $\mathcal{H}(K_d)$. We explain in section 2 below how the notion of optimality is understood. Probably the first use of optimal properties of splines can be traced back to the seminal work of Golomb and Weinberger [9].

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

$$K_d(\boldsymbol{x},\boldsymbol{t}) = \mathrm{e}^{-\gamma^2\|\boldsymbol{x}-\boldsymbol{t}\|^2} \quad \text{for all} \quad \boldsymbol{x},\boldsymbol{t} \in \mathbb{R}^d, \tag{1.1a}$$

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,

$$K_d(\boldsymbol{x},\boldsymbol{t}) = \mathrm{e}^{-\gamma_1^2(x_1-t_1)^2-\cdots-\gamma_d^2(x_d-t_d)^2} = \prod_{\ell=1}^{d} \mathrm{e}^{-\gamma_\ell^2(x_\ell-t_\ell)^2} \quad \text{for all} \quad \boldsymbol{x},\boldsymbol{t} \in \mathbb{R}^d. \tag{1.1b}$$

In the tractability literature, the shape parameters γ_ℓ are called *product weights*. As evidence of its popularity, we note that the anisotropic Gaussian kernel is used in JMP [12], where the values of the γ_ℓ are determined in a data-driven way.

The error of this spline algorithm has been usually analyzed for fixed, and tacitly assumed small, d. The typical convergence rates (see, e.g., [5, 31] and—for Gaussian kernels in particular—[14, 18, 30]) 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. Even if p can be chosen to be arbitrarily large, as is the case for the Gaussian kernel, the dimension dependence of the leading factor in the big $\mathcal{O}$-term is usually not known and might prove to be disastrous.

Since a growing number of applications, such as constructing surrogates for computer experiments, deal 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 independent of d and n, which corresponds to *strong polynomial tractability.* It would also be reasonable to have convergence rates that are polynomially dependent on dimension d and are of the form $Cd^q\, n^{-p}$ for positive C, q and p independent of d and n, which corresponds to *polynomial tractability.* The substantial body of literature on tractability is summarized by Novak and Woźniakowski [15, 16]. For Hilbert spaces, tractability results utilize the eigenvalues of the Hilbert–Schmidt operator for function approximation associated with K_d (see section 2).

The functions to be approximated here lie in the Hilbert space $\mathcal{H}_d = \mathcal{H}(K_d)$, where for our most general results K_d is an arbitrary translation invariant positive definite kernel and for our more specialized results K_d is the Gaussian kernel defined in (1.1). The worst-case error of an algorithm A_n is based on the following $\mathcal{L}_2$ criterion:

$$\text{(1.2a)} \quad e^{\text{wor}}(A_n) = \sup_{\|f\|_{\mathcal{H}_d}\le 1} \|f - A_n(f)\|_{\mathcal{L}_2}, \qquad \|f\|_{\mathcal{L}_2} = \left(\int_{\mathbb{R}^d} f^2(\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\, \mathrm{d}\boldsymbol{t}\right)^{1/2}.$$

Here, ϱ_d is the probability density function defined by

$$\text{(1.2b)} \qquad \varrho_d(\boldsymbol{t}) = \frac{1}{\pi^{d/2}} \exp\left(-(t_1^2 + t_2^2 + \cdots + t_d^2)\right) \qquad \text{for all} \quad \boldsymbol{t} \in \mathbb{R}^d.$$

This specific choice of weight ϱ "localizes" the unbounded domain $\mathbb{R}^d$ by defining a natural length scale of the problem. It also provides a setting for which we can compute eigenvalues and eigenfunctions of the Hilbert–Schmidt operator associated with the Gaussian kernel K_d.

The linear functionals, L_i, used by an algorithm A_n may come either from the class of arbitrary bounded linear functionals, $\Lambda^{\text{all}} = \mathcal{H}_d^*$, or from the class of function evaluations, Λ^{std}. The *nth minimal worst-case error* over all possible algorithms is defined as

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

Since the optimal algorithm is the spline algorithm, S_n, provided the L_j are specified, the problem of computing $e^{\text{wor-}\vartheta}(n, \mathcal{H}_d)$ becomes one of finding the best sampling scheme. For notational simplicity ϑ denotes either the standard or linear class. Clearly, $e^{\text{wor-all}}(n, \mathcal{H}_d) \le e^{\text{wor-std}}(n, \mathcal{H}_d)$ since the former uses a larger class of function data. The case $n = 0$ means that no information about f is used to construct the algorithm. For $n = 0$ we approximate f by constant algorithms, i.e., $A_0(f) = c \in \mathcal{L}_2$. It is easy to see that the zero algorithm, $A_n(f) = 0$, minimizes the error and $e^{\text{wor-}\vartheta}(0, \mathcal{H}_d) = \|I_d\|$, where $I_d : \mathcal{H}_d \to \mathcal{L}_2$ is the linear embedding operator defined by $I_d(f) = f$.

This article establishes upper and lower bounds for the convergence rates for the nth minimal worst-case error with no dimension dependence using an isotropic or anisotropic Gaussian kernel. These rates are summarized in Table 1.1. The notation $\preceq n^{-p}$ means that for all $\delta > 0$ the error is bounded *above* by $C_\delta n^{-p+\delta}$ for some positive C_δ that is independent of the sample size, n, and the dimension, d, but may depend on δ. The notation $\succeq n^{-p}$ is defined analogously and means that the error is bounded from *below* by $C_\delta n^{-p-\delta}$ for all $\delta > 0$. The notation $\asymp n^{-p}$ means that the error is both $\preceq n^{-p}$ and $\succeq n^{-p}$.

Table 1.1

Error decay rates for the Gaussian kernel as a function of sample size n.

	Error criterion	
Data available	Absolute : $e^{\text{wor-}\vartheta}(n,\mathcal{H}_d)$	Normalized : $\frac{e^{\text{wor-}\vartheta}(n,\mathcal{H}_d)}{e^{\text{wor-}\vartheta}(0,\mathcal{H}_d)}$
Arbitrary linear functionals	$\asymp n^{-\max(r(\boldsymbol{\gamma}),1/2)}$ Theorem 5.2	$\asymp n^{-r(\boldsymbol{\gamma})}$ if $r(\boldsymbol{\gamma})>0$, Theorem 6.2
Function values	$\preceq n^{-\max(r(\boldsymbol{\gamma})/[1+1/(2r(\boldsymbol{\gamma}))],1/4)}$ Theorem 5.3 and 5.4	$\preceq n^{-r(\boldsymbol{\gamma})/[1+1/(2r(\boldsymbol{\gamma}))]}$ if $r(\boldsymbol{\gamma})>1/2$, Corollary 6.4

The term $r(\boldsymbol{\gamma})$ appearing in Table 1.1 denotes the *rate of convergence* of the shape parameter sequence $\boldsymbol{\gamma}$ and is defined by

$$r(\boldsymbol{\gamma}) = \sup\left\{\beta > 0 \;\middle|\; \sum_{\ell=1}^{\infty} \gamma_\ell^{1/\beta} < \infty\right\} \tag{1.3}$$

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, $\gamma_1 \geq \gamma_2 \geq \cdots$, then this definition is equivalent to

$$r(\boldsymbol{\gamma}) = \sup\left\{\beta \geq 0 \mid \lim_{\ell\to\infty} \gamma_\ell\, \ell^{\beta} = 0\right\}. \tag{1.4}$$

As can be seen in Table 1.1, *any* isotropic Gaussian kernel gives rise to dimension-independent convergence rates (when measured by the absolute error criterion) of order $n^{-1/2}$ provided that optimal linear functional data is available. Our remarks following Theorem 5.1 show that dimension-independent convergence rates of the same order can be achieved with *any* positive definite radial (isotropic) kernel as well as for certain classes of translation invariant kernels.

For arbitrary linear functionals the optimal data correspond to the first n "Fourier coefficients" of f (see section 2). For function value data one may obtain $\mathcal{O}(n^{-1/4})$ convergence, although unfortunately the current state of theory does not allow us to construct the optimal data sites. Again, our result about dimension-independent convergence rates (Theorem 5.3) generalizes to arbitrary positive definite radial (isotropic) kernels. The convergence rates for arbitrary linear functionals provide a *lower bound* on what is possible using function values. Thus, we know that for isotropic Gaussian kernels one can never obtain dimension-independent convergence rates better than of order $n^{-1/2}$, no matter how cleverly the data sites are chosen.

For *high dimension-independent convergence rates* one needs the sequence of shape parameters to decay to zero quickly, i.e., the decay rate of $\boldsymbol{\gamma} = \{\gamma_\ell\}_{\ell\in\mathbb{N}}$ must be large, as can be seen in Table 1.1. These results are derived in sections 5 and 6. The table also highlights that for the normalized error criterion, dimension-independent convergence rates with isotropic kernels are *not possible.*

Our analysis relates the decay of the shape parameters to the decay of the eigenvalues of the Hilbert–Schmidt operator associated with K_d (see section 2). Such an analysis is possible for other kernels K_d as long as the eigenvalues of the corresponding Hilbert–Schmidt operator are known. Because the Gaussian kernel is of product form, the eigenvalues for the dimension d case are products of the eigenvalues for $d = 1$, which facilitates the analysis. This fact, along with the popularity of the Gaussian kernel, is why this article focuses on this kernel.

As the vector of shape parameters $\boldsymbol{\gamma}$ changes, the Hilbert space of functions to be approximated as well as its norm change, as illustrated in the next section. Thus, the results derived here and summarized in Table 1.1 are for a *whole family* of spaces of functions indexed by $\boldsymbol{\gamma}$. If a function depends on a moderate or large number of variables and lies in a Hilbert space whose reproducing kernel is isotropic, then we should not be surprised if all algorithms give poor rates of convergence. On the other hand, if the function lies in a Hilbert space whose shape parameters decay with dimension, then there exist algorithms with good rates of convergence.

As a prelude to deriving new convergence and tractability results, the next section reviews some principles of function approximation on Hilbert spaces. Section 4 applies these principles to Hilbert spaces with translation invariant reproducing kernels.

2. Background.

2.1. Reproducing kernel Hilbert spaces. Let $\mathcal{H}_d = \mathcal{H}(K_d)$ denote a reproducing kernel Hilbert space of real functions defined on $\mathbb{R}^d$. The goal is to accurately approximate any function in $\mathcal{H}_d$ given a finite number of data about it. The reproducing kernel $K_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is symmetric and positive definite and reproduces function values. This means that for all $n \in \mathbb{N}$, $\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$, $\mathbf{c} = (c_1, c_2, \dots, c_n) \in \mathbb{R}^n$, and $f \in \mathcal{H}_d$, the following properties hold:

$$K_d(\cdot, \boldsymbol{x}) \in \mathcal{H}_d, \qquad K_d(\boldsymbol{x}, \boldsymbol{t}) = K_d(\boldsymbol{t}, \boldsymbol{x}), \qquad \sum_{i=1}^{n} \sum_{j=1}^{n} K_d(\boldsymbol{x}_i, \boldsymbol{x}_j) c_i c_j \geq 0, \tag{2.1a}$$

$$f(\boldsymbol{x}) = \langle f, K_d(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}_d}. \tag{2.1b}$$

For an arbitrary $\boldsymbol{x} \in \mathbb{R}^d$ consider the evaluation functional $L_{\boldsymbol{x}}(f) = f(\boldsymbol{x})$ for all $f \in \mathcal{H}_d$. Then $L_{\boldsymbol{x}}$ is continuous and $\|L_{\boldsymbol{x}}\|_{\mathcal{H}_d^*} = K_d^{1/2}(\boldsymbol{x}, \boldsymbol{x})$ (see [1, 27]).

It is assumed that $\mathcal{H}_d$ is continuously embedded in the space $\mathcal{L}_2 = \mathcal{L}_2(\mathbb{R}^d, \varrho_d)$ of square Lebesgue integrable functions, where the $\mathcal{L}_2$ norm was defined in (1.2). Continuous embedding means that $\|I_d f\|_{\mathcal{L}_2} = \|f\|_{\mathcal{L}_2} \leq \|I_d\| \, \|f\|_{\mathcal{H}_d}$ for all $f \in \mathcal{H}_d$. The kernels considered here are assumed to satisfy

$$\int_{\mathbb{R}^d} K_d(\boldsymbol{t}, \boldsymbol{t}) \, \varrho_d(\boldsymbol{t}) \, \mathrm{d}\boldsymbol{t} < \infty. \tag{2.2}$$

This is sufficient to imply continuous embedding since

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

Many reproducing kernels are used in practice. A kernel is called *translation invariant* or *stationary* if $K(\boldsymbol{x}, \boldsymbol{t}) = \widetilde{K}_d(\boldsymbol{x} - \boldsymbol{t})$. 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 popular choice is the Gaussian kernel defined in (1.1). The anisotropic Gaussian kernel is translation invariant, and the isotropic Gaussian kernel is radially symmetric. Stationary or isotropic kernels are common in the literature on computational mathematics [2, 5, 31], statistics [1, 22, 27], statistical learning [17, 23], and engineering applications [7]. Observe that (2.2) holds for all translation invariant kernels since

$$\int_{\mathbb{R}^d} K_d(\boldsymbol{t},\boldsymbol{t})\,\varrho_d(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} = \begin{cases} \widetilde{K}_d(\boldsymbol{0}), & \text{translation invariant,} \\ \kappa(0), & \text{radially symmetric,} \\ 1, & \text{(anisotropic) Gaussian.} \end{cases}$$

Functions in $\mathcal{H}_d$ are approximated by linear algorithms[1]

$$A_n(f)\,(\boldsymbol{x}) = \sum_{j=1}^{n} L_j(f) a_j(\boldsymbol{x}) \quad \text{for all} \quad f \in \mathcal{H}_d, \ \ \boldsymbol{x} \in \mathbb{R}^d \tag{2.3}$$

for some continuous linear functionals $L_j \in \mathcal{H}_d^*$ and functions $a_j \in \mathcal{L}_2$. In the case of minimum norm interpolation (cf. (2.5)) these functions are known as Lagrange or cardinal functions and are specified in (2.6). Note that for known functions a_j, the cost of computing $A_n(f)\,(x)$ is equal to n multiplications and $n-1$ additions of real numbers plus the cost of computing $a_j(\boldsymbol{x})$ for $j = 1, 2, \ldots, n$. That is why it is important to minimize n for which the error of the algorithm A_n meets the required error threshold. We do not consider the cost of generating the data samples, i.e., the computation of $L_j(f)$ for $j = 1, 2, \ldots, n$, even though, depending on the nature of the linear functionals, this may be nontrivial.

2.2. Convergence and tractability. This article addresses two problems: convergence and tractability. The former considers how fast the error vanishes as n increases. This is the typical point of view taken in numerical analysis and for which one can find many results in the (radial) kernel literature as summarized in, e.g., [5, 31]. However, this problem 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.

Problem 1: Rate of convergence (fixed $\boldsymbol{d}$). We would like to know how fast $e^{\text{wor-}\vartheta}(n, \mathcal{H}_d)$ goes to zero as n goes to infinity. In particular, we study the rate of convergence (defined by the notation in (1.3) and (1.4)) of the sequence $\{e^{\text{wor-}\vartheta}(n, \mathcal{H}_d)\}_{n\in\mathbb{N}}$. Since the numbers $e^{\text{wor-}\vartheta}(n, \mathcal{H}_d)$ are ordered, we have

$$r^{\text{wor-}\vartheta}(\mathcal{H}_d) := r\left(\{e^{\text{wor-}\vartheta}(n, \mathcal{H}_d)\}\right) = \sup\left\{\beta \ge 0 \mid \lim_{n\to\infty} e^{\text{wor-}\vartheta}(n, \mathcal{H}_d)\, n^{\beta} = 0\right\}. \tag{2.4}$$

Roughly speaking, the rate of convergence, $r^{\text{wor-}\vartheta}(\mathcal{H}_d)$, is the largest β for which the nth minimal errors behave like $n^{-\beta}$. For example, if $e^{\text{wor-}\vartheta}(n, \mathcal{H}_d) = n^{-\alpha}$ for a positive α, then $r^{\text{wor-}\vartheta}(\mathcal{H}_d) = \alpha$. Under this definition, even sequences of the form $e^{\text{wor-}\vartheta}(n, \mathcal{H}_d) = n^{-\alpha} \ln^p n$ for an arbitrary p still have $r^{\text{wor-}\vartheta}(\mathcal{H}_d) = \alpha$. On the other hand, if $e^{\text{wor-}\vartheta}(n, \mathcal{H}_d) = q^n$ for a number $q \in (0, 1)$, then $r^{\text{wor-}\vartheta}(\mathcal{H}_d) = \infty$.

Obviously, $r^{\text{wor-all}}(\mathcal{H}_d) \ge r^{\text{wor-std}}(\mathcal{H}_d)$. We would like to know both rates and verify if $r^{\text{wor-all}}(\mathcal{H}_d) > r^{\text{wor-std}}(\mathcal{H}_d)$, i.e., whether Λ^{all} admits a better rate of convergence than Λ^{std}.

[1] 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 (2.3) 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 (2.3) can be done without loss of generality. For more detail see, e.g., [26].

Problem 2: Tractability (unbounded d). In this case, we would like to know how $e^{\text{wor-}\vartheta}(n,\mathcal{H}_d)$ depends not only on n but also on d. Because of the focus on d-dependence, the *absolute* and *normalized* error criteria described in the previous section may lead to different answers. 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 for the absolute error criterion and does not exceed $\varepsilon\, e^{\text{wor-}\vartheta}(0,\mathcal{H}_d) = \varepsilon\, \|I_d\|$ for the normalized error criterion. That is,

$$n^{\text{wor-}\psi\text{-}\vartheta}(\varepsilon,\mathcal{H}_d) = \min\left\{ n \mid e^{\text{wor-}\vartheta}(n,\mathcal{H}_d) \le \begin{cases} \varepsilon, & \psi = \text{abs}, \\ \varepsilon\,\|I_d\|, & \psi = \text{norm}, \end{cases} \right\}.$$

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 numbers C, p, and q such that

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

If $q = 0$ above, then we say that $\mathcal{I}$ is *strongly polynomially tractable* and the infimum of p satisfying the bound above is 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, Λ^{all} or Λ^{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* if and only if there exist numbers C and t for which

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

for all $d \in \mathbb{N}$ and $\varepsilon \in (0,1)$. The exponent of quasi-polynomial tractability is defined as the infimum of t satisfying the bound above. Finally, $\mathcal{I}$ is *weakly tractable* if and only if

$$\lim_{\varepsilon^{-1}+d\to\infty} \frac{\ln\, n^{\text{wor-}\psi\text{-}\vartheta}(\varepsilon,\mathcal{H}_d)}{\varepsilon^{-1}+d} = 0,$$

which only means that we do not have exponential dependence on ε^{-1} and d. Note that for a fixed d, quasi-polynomial tractability means that

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

Hence, the exponent of ε^{-1} may now weakly depend on d through $\ln\, d$.

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 Λ^{all} or Λ^{std} is considered and on the error criterion. Motivation of tractability study and more on tractability concepts can be found in [15]. Quasi-polynomial tractability has been recently studied in [8].

2.3. The spline algorithm. As alluded to in the introduction, the optimal approximation algorithm for a function in $\mathcal{H}_d$ is known once the data functionals $L_1, \dots, L_n$ are specified. That is, the optimal $a_1, \dots, a_n$ in (2.3) for which the worst-case error of A_n is minimized can be determined explicitly. This optimal algorithm, S_n, is the *spline* or the *minimal norm interpolant*; see, e.g., section 5.7 of [26].

For given $y_j = L_j(f)$ for $j = 1, 2, \dots, n$, we take $S_n(f)$ as an element of $\mathcal{H}_d$ that satisfies the conditions

$$\begin{aligned} &L_j(S_n(f)) = y_j \qquad\qquad \text{for} \quad j = 1, 2, \dots, n, \\ &\|S_n(f)\|_{\mathcal{H}_d} = \inf_{g \in \mathcal{H}_d,\ L_j(g) = y_j,\ j = 1,2,\dots,n} \|g\|_{\mathcal{H}_d}. \end{aligned} \tag{2.5}$$

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

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

Then

$$S_n(f)(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x}) \mathsf{K}^{-1} \boldsymbol{y} \quad \text{with} \quad \boldsymbol{k}(\boldsymbol{x}) = (k_i(\boldsymbol{x}))_{i=1}^n, \tag{2.6}$$

i.e., the optimal functions a_j of (2.3) are given by $\boldsymbol{a}^T(\boldsymbol{x}) = \boldsymbol{k}^T(\boldsymbol{x}) \mathsf{K}^{-1}$ and

$$e^{\text{wor}}(S_n) = \sup_{\|f\|_{\mathcal{H}_d} \le 1,\ L_j(f) = 0,\ j = 1,2,\dots,n} \|f\|_{\mathcal{L}_2}.$$

Note that depending on the choice of linear functionals $L_1, \dots, L_n$ the matrix K may not necessarily be invertible; however, in that case $\boldsymbol{c} = \mathsf{K}^\dagger \boldsymbol{y}$ is well defined via the pseudoinverse $\mathsf{K}^\dagger$ as the vector of minimal Euclidean norm which satisfies $\mathsf{K}\boldsymbol{c} = \boldsymbol{y}$. Alternatively, one can require the linear functionals to be linearly independent.

We briefly comment on the cost of computing $S_n(f)(\boldsymbol{x})$. Assume that the matrix K is given and is nonsingular as well as that the function values $k_j(\boldsymbol{x})$ can be computed. Then $S_n(f)(\boldsymbol{x})$ can be computed by solving an $n \times n$ system of linear equations. This requires $\mathcal{O}(n^3)$ arithmetic operations if, for example, Gaussian elimination is used. But we usually can do better. For a general nonsingular matrix K, suppose we need to compute the spline $S_n(f)(\boldsymbol{x})$ for many $\boldsymbol{x}$. Then we can factorize the matrix K once at cost proportional to $\mathcal{O}(n^3)$ and then compute the solution c at cost $\mathcal{O}(n^2)$. More important, as we shall see later, for the optimally chosen linear functionals L_j the matrix K is an identity and the cost of computing $S_n(f)(\boldsymbol{x})$ equal to n multiplications, $n - 1$ additions, and the n function evaluations of $k_j(\boldsymbol{x})$. In any case, independent of the matrix K, it is clear that we should aim to work with the smallest possible n.

The spline enjoys more optimality properties. For instance, it minimizes the *local* worst-case error (see, e.g., [26, Theorem 5.7.2]). Roughly speaking this means that for each $\boldsymbol{x} \in \mathbb{R}^d$, 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 $A_n = S_n$.

2.4. The eigendecomposition of the reproducing kernel. It is nontrivial 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., [26]. Namely, let $W_d = I_d^* I_d : \mathcal{H}_d \to \mathcal{H}_d$, where $I_d^* : \mathcal{L}_2 \to \mathcal{H}_d$ denotes the adjoint of the embedding operator, i.e., the operator satisfying $\langle f, I_d^* h \rangle_{\mathcal{H}_d} = \langle I_d f, h \rangle_{\mathcal{L}_2}$ for all $f \in \mathcal{H}_d$ and $h \in \mathcal{L}_2$. As a consequence, W_d is a self-adjoint and positive definite linear operator given by

$$W_d\, f = \int_{\mathbb{R}^d} f(\boldsymbol{t})\, K_d(\cdot,\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} \quad \text{for all} \quad f \in \mathcal{H}_d.$$

In fact, W_d is a *Hilbert–Schmidt operator* (see, e.g., [11]), that is, it has a finite trace.

Clearly,

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

It is known that $\lim_{n\to\infty} e^{\text{wor-all}}(n,\mathcal{H}_d) = 0$ if and only if W_d is compact (see, e.g., [15, section 4.2]). In particular, (2.2) implies that W_d is compact.

Let us define the eigenpairs of W_d by $(\lambda_{d,j}, \eta_{d,j})$, where the eigenvalues are ordered, $\lambda_{d,1} \ge \lambda_{d,2} \ge \cdots$, and

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

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

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

Taking $f = \eta_{d,i}$ we see that $\{\eta_{d,j}\}$ is 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.[2] 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 $\mathcal{H}_d$ we have

$$K_d(\boldsymbol{x},\boldsymbol{t}) = \sum_{j=1}^{\infty} \eta_{d,j}(\boldsymbol{x})\, \eta_{d,j}(\boldsymbol{t}) = \sum_{j=1}^{\infty} \lambda_{d,j}\, \varphi_{d,j}(\boldsymbol{x})\, \varphi_{d,j}(\boldsymbol{t}) \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{t} \in \mathbb{R}^d. \tag{2.7}$$

The assumption (2.2) implies that W_d is a Hilbert–Schmidt (or a finite trace) operator:

$$\sum_{j=1}^{\infty} \lambda_{d,j} = \int_{\mathbb{R}^d} K_d(\boldsymbol{t},\boldsymbol{t})\, \varrho_d(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} < \infty. \tag{2.8}$$

It is known that the best choice of L_j for the class Λ^{all} is $L_j = \langle \cdot, \eta_{d,j}\rangle_{\mathcal{H}_d}$ (see, e.g., [15, section 4.2]). Then the spline algorithm S_n with the minimal worst-case error is defined using the eigenfunctions corresponding to the n largest eigenvalues, i.e., $a_j = \eta_{d,j}$ in (2.3):

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

and

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

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

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

The results for the class $\Lambda^{\rm all}$ 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 $\Lambda^{\rm std}$, the situation is much harder, although there are papers that relate rates of convergence and tractability conditions between classes $\Lambda^{\rm all}$ and $\Lambda^{\rm std}$. Again we report these results later.

In summary, knowing the eigenpairs of the Hilbert–Schmidt operator W_d associated with K_d provides us both with the optimal linear functionals as well as the minimal worst-case error for the minimum norm interpolant. Other power series expansions of K_d, while potentially easier to find, likely will not have these nice properties.

3. Eigenvalues of Gaussian kernels. From the previous section, it is clear that the key to dimension-independent convergence rates and tractability is to show that the eigenvalues of the reproducing kernel ordered by size decay quickly enough. While the general framework from the previous section applies to any symmetric positive definite kernel whose eigenpairs are known, we now analyze the function approximation problem for the Hilbert space with the Gaussian kernel given by (1.1b) since—as we will now show—the eigenpairs in this case are readily available.

What makes the analysis of the Gaussian kernel K_d especially attractive is its product form. This implies that the space $\mathcal{H}_d$ is the tensor product of the Hilbert spaces of univariate spaces with the kernels $\mathrm{e}^{-\gamma_\ell^2(x-t)^2}$ for $x, t \in \mathbb{R}$. As a further consequence 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 $\mathcal{H}(K_1)$ with $K_1(x,t) = \mathrm{e}^{-\gamma^2(x-t)^2}$. Then the eigenpairs $(\tilde\lambda_{\gamma,j}, \eta_{\gamma,j})$ of W_1 are known; see [17]. (This is related to Mehler's formula and appropriately rescaled Hermite functions [25, Problems and Exercises, Item 23].) 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 been 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

$$\omega_\gamma = \frac{\gamma^2}{\frac{1}{2}(1+\sqrt{1+4\gamma^2})+\gamma^2}, \tag{3.1}$$

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

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

so that

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

Note that both $\eta_{\gamma,j}(x)$ and $\varphi_{\gamma.j}(x)$ can be computed at cost proportional to j by using the three-term recurrence relation for Hermite polynomials.

Obviously, we have $\langle \eta_{\gamma,i}, \eta_{\gamma,j}\rangle_{\mathcal{H}(K_1)} = \langle \varphi_{\gamma,i}, \varphi_{\gamma,j}\rangle_{\mathcal{L}_2} = \delta_{ij}$, and applying (2.7) we obtain

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

Note that the eigenvalues $\tilde{\lambda}_{\gamma,j}$ are ordered and have the following asymptotic properties:

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

$$\tilde{\lambda}_{\gamma,j} = \left(1-\gamma^2+\mathcal{O}(\gamma^4)\right)\left(\frac{\gamma^2}{1-\gamma^2+\mathcal{O}(\gamma^4)}\right)^{j-1} \qquad \text{for } \; j=1,2,\dots . \tag{3.2}$$

The space $\mathcal{H}(K_1)$ consists of analytic functions for which

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

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

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

For more about the characterization of the space $\mathcal{H}(K_1)$ see [24].

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

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

$$= \prod_{\ell=1}^{d} (1-\omega_{\gamma_\ell})\,\omega_{\gamma_\ell}^{j_\ell-1}, \tag{3.3}$$

where ω_γ is defined above in (3.1), and

$$\eta_{d,\boldsymbol{\gamma},\boldsymbol{j}} = \prod_{\ell=1}^{d} \sqrt{\tilde{\lambda}_{\gamma_\ell,j_\ell}}\,\varphi_{\gamma_\ell,j_\ell}, \qquad \langle \eta_{d,\boldsymbol{\gamma},\boldsymbol{i}}, \eta_{d,\boldsymbol{\gamma},\boldsymbol{j}}\rangle_{\mathcal{H}_d} = \langle \varphi_{\boldsymbol{\gamma},\boldsymbol{i}}, \varphi_{\boldsymbol{\gamma},\boldsymbol{j}}\rangle_{\mathcal{L}_2} = \delta_{\boldsymbol{ij}}.$$

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 $\lambda_{d,1}\ge\lambda_{d,2}\ge\cdots$. 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, $\lambda_{d,1} = \prod_{\ell=1}^{d} (1 - \omega_{\gamma_\ell})$. 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. Also included is a simple estimate of $\lambda_{d,n+1}$. This lemma is used repeatedly in the following sections.

Lemma 3.1. *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 τth power of the eigenvalues for the d-variate case, $d \geq 1$, is*

$$\sum_{j=1}^{\infty} \lambda_{d,j}^{\tau} = \sum_{\boldsymbol{j}\in\mathbb{N}^d} \tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}}^{\tau} = \prod_{\ell=1}^{d} \left(\sum_{j=1}^{\infty} \tilde{\lambda}_{\gamma_\ell,j}^{\tau} \right) = \prod_{\ell=1}^{d} \frac{(1-\omega_{\gamma_\ell})^{\tau}}{1-\omega_{\gamma_\ell}^{\tau}} \begin{cases} > 1, & 0 < \tau < 1, \\ = 1, & \tau = 1. \end{cases} \tag{3.4}$$

The $(n+1)$st largest eigenvalue satisfies

$$\lambda_{d,n+1} \leq \frac{1}{(n+1)^{1/\tau}} \prod_{\ell=1}^{d} \frac{1-\omega_{\gamma_\ell}}{(1-\omega_{\gamma_\ell}^{\tau})^{1/\tau}}. \tag{3.5}$$

Proof. Equation (3.4) follows directly from the formula for $\tilde{\lambda}_{d,\boldsymbol{\gamma},\boldsymbol{j}}$ in (3.3). From the definition of ω_γ in (3.1) it follows that $0 < \omega_\gamma < 1$ for all $\gamma > 0$. For $\tau \in (0,1)$, consider the function $f : \omega \mapsto (1-\omega)^\tau - 1 + \omega^\tau$ defined on $[0,1]$. 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 $(n+1)$st largest eigenvalue in (3.5) and completes the proof. □

The main point of (3.5) 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 translation invariant kernels.

In this section we consider the function approximation problem for the Hilbert space $\mathcal{H}_d = \mathcal{H}(K_d)$ with translation invariant kernels and in particular the anisotropic Gaussian kernel given by (1.1b). 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 (or radial) case for which all $\gamma_\ell = \gamma > 0$.

We want to verify how fast the minimal errors $e^{\text{wor-all}}(n, \mathcal{H}_d)$ and $e^{\text{wor-std}}(n, \mathcal{H}_d)$ go to zero and what the rate of convergence of these sequences is; see (2.4). Note that the dimension d is arbitrary, but fixed, throughout this section.

Theorem 4.1. *For the anisotropic as well as the isotropic Gaussian kernel*

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

Proof. For the class Λ^{all} we know that $e^{\text{wor-all}}(n, \mathcal{H}_d) = \sqrt{\lambda_{d,n+1}}$, where $\lambda_{d,n+1}$ is the $(n+1)$st largest eigenvalue of the Hilbert–Schmidt operator W_d associated with K_d. Lemma 3.1 demonstrates that $\lambda_{d,n+1}$ is proportional to $(n+1)^{-1/\tau}$ times a dimension-dependent constant. (Note that the dependence on τ is irrelevant since (3.5) holds for all τ.) This implies that $r^{\text{wor-all}}(\mathcal{H}_d) \geq 1/(2\tau)$ and since τ can be arbitrarily small, we conclude that $r^{\text{wor-all}}(\mathcal{H}_d) = \infty$, as claimed.

Consider now the class Λ^{std}. We use [13, Theorem 5], which states that if there exist numbers $p > 1$ and B such that

$$\lambda_{d,n} \leq B\, n^{-p} \quad \text{for all} \quad n \in \mathbb{N}, \tag{4.1}$$

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

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

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

Note that assumption (4.1) holds in our case for an arbitrarily large p with B that can depend on d. Hence, $r^{\text{wor-std}}(\mathcal{H}_d) \geq (1-\delta)\, p^2/(2p+2)$, and since δ can be arbitrarily small and p can be arbitrarily large we conclude $r^{\text{wor-std}}(\mathcal{H}_d) = \infty$, as claimed. This completes the proof. □

We stress that the algorithm A_n that was used in the proof is nonconstructive. 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 in 1963 [21], 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., [28].

Theorem 4.1 states that as long as one is interested only in the rate of convergence, the function approximation problem for Hilbert spaces with infinitely smooth kernels such as the Gaussian is easy. As mentioned earlier, convergence rates for wide classes of infinitely smooth ($p = \infty$) radial kernels such as, e.g., (inverse) multiquadrics and Gaussians can be found in the literature [14, 18, 31]. However, the rate of convergence tells us nothing about the dependence on the dimension d. As long as d is small the dependence on d is irrelevant. But if d is large we want to check how the decay rate of the minimal worst-case error depends not only on the number of samples, but also on the dimension. We are especially concerned about a possible exponential dependence on d which following Bellman is called the *curse of dimensionality*. It also may happen that we have a trade-off 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 $\mathcal{H}_d = \mathcal{H}(K_d)$ 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 Λ^{all} or Λ^{std} needed to guarantee a worst-case error of at most ε,

$$n^{\text{wor-abs-}\vartheta}(\varepsilon, \mathcal{H}_d) = \min\left\{ n \mid\ e^{\text{wor-}\vartheta}(n, \mathcal{H}_d) \leq \varepsilon \right\}, \qquad \vartheta \in \{\text{std}, \text{all}\}.$$

5.1. Arbitrary linear functionals. We first analyze the class $\Lambda^{\rm all}$ and polynomial tractability. We are able to establish dimension-independent convergence rates for any translation invariant positive definite kernel. First we discuss the Gaussian kernel and then explain how to generalize our result to other radial and general translation invariant kernels.

THEOREM 5.1. *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 we have the following:*

- *$\mathcal{I}$ 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*

$$e^{\text{wor-all}}(n,\mathcal{H}_d) \le (n+1)^{-1/2}, \qquad n^{\text{wor-abs-all}}(\varepsilon,\mathcal{H}_d) \le \varepsilon^{-2}.$$

- *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 [15, Theorem 5.1]. This theorem says that $\mathcal{I}$ is strongly polynomially tractable if and only if there exist two positive numbers C_1 and τ such that

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

If so, then

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

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 (3.4) 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 $n^{\text{wor-abs-all}}(\varepsilon,\mathcal{H}_d) \le (C_1+1)\,\varepsilon^{-2}$. For C_1 tending to zero, we conclude the bound

$$n^{\text{wor-abs-all}}(\varepsilon,\mathcal{H}_d) \le \varepsilon^{-2}.$$

Furthermore, by (3.5) in Lemma 3.1 it follows that

$$e^{\text{wor-all}}(n,\mathcal{H}_d) = \sqrt{\lambda_{d,n+1}} \le (n+1)^{-1/2},$$

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 3.1 and obtain

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

For $\tau \in (0,1)$, we know from Lemma 3.1 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 [15, 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

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

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

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

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. □

Although Theorem 5.1 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

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

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.,
$$n^{\text{wor-all}}(n, \mathcal{H}_d) = \mathcal{O}\left(\varepsilon^{-2}\right).$$
- 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
$$n^{\text{wor-all}}(n, \mathcal{H}_d) = \mathcal{O}\left(d^{\,q}\, \varepsilon^{-2}\right).$$

- 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,

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

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

We now compare Theorems 4.1 and 5.1. Theorem 4.1 says that for any p we have

$$e^{\text{wor-all}}(n,\mathcal{H}_d) = \mathcal{O}(n^{-p}),$$

but the factor in the big $\mathcal{O}$ notation may depend on d. In fact, from Theorem 5.1 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—for Gaussian kernels—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 (1.3), which measures the rate of decay of the shape parameter sequence.

Theorem 5.2. *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^{all} and the absolute error criterion. Let $r(\boldsymbol{\gamma})$ be the rate of decay of shape parameters. Then we have the following:*

- *$\mathcal{I}$ is strongly polynomially tractable with exponent*
$$p^{\text{all}} = \min\left(2, \frac{1}{r(\boldsymbol{\gamma})}\right) \le 2.$$
- *For all $d \in \mathbb{N}$, $\varepsilon \in (0,1)$, and $\delta \in (0,1)$ we have*
$$e^{\text{wor-all}}(n,\mathcal{H}_d) = \mathcal{O}\left(n^{-1/p^{\text{all}}+\delta}\right) = \mathcal{O}\left(n^{-\max(r(\boldsymbol{\gamma}),1/2)+\delta}\right),$$
$$n^{\text{wor-abs-all}}(\varepsilon,\mathcal{H}_d) = \mathcal{O}\left(\varepsilon^{-(p^{\text{all}}+\delta)}\right),$$
where the factors in the big $\mathcal{O}$ notation are independent of 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*
$$n^{\text{wor-abs-all}}(\varepsilon,\mathcal{H}_d) = \mathcal{O}\left(\varepsilon^{-p}\, d^{\,q}\right) \quad \text{for all } \ \varepsilon \in (0,1) \text{ and } d \in \mathbb{N},$$
then $p \ge p^{\text{all}}$, which means that strong polynomial tractability is equivalent to polynomial tractability.

Proof. As in the proof of Theorem 5.1, $\mathcal{I}$ is strongly polynomially tractable if and only if there exist two positive numbers C_1 and τ such that

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

Furthermore, the exponent p^{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 $\gamma_\ell > 0$ ensures $\lambda_{d,j} < 1$

$$\sum_{j=\lceil C_1 \rceil}^{\infty} \lambda_{d,j}^{\tau} \ge \sum_{j=1}^{\infty} \lambda_{d,j}^{\tau} - C_1 = \prod_{\ell=1}^{\infty} \frac{(1-\omega_{\gamma_\ell})^{\tau}}{1-\omega_{\gamma_\ell}^{\tau}} - C_1.$$

Therefore, $\mathcal{I}$ is strongly polynomially tractable if and only if 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^{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 (3.1), 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 $r(\boldsymbol{\gamma}) \ge 1/(2\tau) > 1/2$. Hence, $p^{\mathrm{all}} < 2$ only if $r(\gamma) > 1/2$. On the other hand, $2\tau \ge 1/r(\boldsymbol{\gamma})$ and therefore $p^{\mathrm{all}} \ge 1/r(\boldsymbol{\gamma})$. This establishes the formula for p^{all}. The estimates on $e^{\text{wor-all}}(n,\mathcal{H}_d)$ and $n^{\text{wor-abs-all}}(\varepsilon,\mathcal{H}_d)$ follow from the definition of strong tractability.

Assume now polynomial tractability with $p < 2$ and an arbitrary q. Then $\lambda_{d,n+1} \le \varepsilon^2$ 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

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

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

It is interesting to notice that the last part of Theorem 5.2 does not hold, in general, for unordered shape parameters. Indeed, for $s > 1/2$, take

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

Then strong polynomial tractability holds with the exponent 2 since $C_3 = \infty$ in the proof of Theorem 5.2 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

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

This implies that

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

Theorem 5.2 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 $e^{\text{wor-all}}(n, \mathcal{H}_d)$ 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.

Extending Theorem 5.2 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 for Gaussian kernels. As for Theorem 5.1, it is straightforward to extend this theorem to other radially symmetric and even translation invariant positive definite kernels.

THEOREM 5.3. *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 Λ^{std} and the absolute error criterion. Then we have the following:*

- *$\mathcal{I}$ 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*

$$e^{\text{wor-std}}(n,\mathcal{H}_d) \le \frac{\sqrt{2}}{n^{1/4}}\left(1+\frac{1}{2\sqrt{n}}\right)^{1/2},$$
$$n^{\text{wor}-\text{abs}-\text{std}}(\varepsilon,\mathcal{H}_d) \le \left\lceil \frac{(1+\sqrt{1+\varepsilon^2})^2}{\varepsilon^4} \right\rceil.$$

- *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 [29, Theorem 1]. This theorem says that

$$e^{\text{wor-std}}(n,\mathcal{H}_d) \le \min_{k=0,1,\dots}\left([e^{\text{wor}-\text{all}}(k,\mathcal{H}_d)]^2 + \frac{k}{n}\right)^{1/2}. \tag{5.1}$$

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

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

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

For the isotropic case, we know from Theorem 5.1 that the exponent of strong tractability for the class Λ^{all} is 2. For the class Λ^{std}, the exponent cannot be smaller.

Finally, assume that we have polynomial tractability for the class Λ^{std}. Then we also have polynomial tractability for the class Λ^{all}. From Theorem 5.1 we know that then strong tractability for the class Λ^{all} holds. Furthermore, we know that the exponent of strong tractability is 2 and $n^{\text{wor-abs-all}}(\varepsilon,\mathcal{H}_d) \le \varepsilon^{-2}$. As above, we then get strong tractability also for Λ^{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 Λ^{std}. We suspect that it is *not* sharp and that maybe even an error bound of order $n^{-1/2}$ holds for the class Λ^{std} exactly as for the class Λ^{all}.

For fast decaying shape parameters it is possible to improve the rate obtained in Theorem 5.3. This is the subject of our next theorem.

THEOREM 5.4. *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. Let $r(\boldsymbol{\gamma}) > 1/2$. Then we have the following:*

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

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

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

$$e^{\text{wor-std}}(n, \mathcal{H}_d) = \mathcal{O}\left(n^{-1/p^{\text{std}}+\delta}\right) = \mathcal{O}\left(n^{-r(\boldsymbol{\gamma})/[1+1/(2r(\boldsymbol{\gamma}))]+\delta}\right),$$
$$n^{\text{wor-abs-std}}(\varepsilon, \mathcal{H}_d) = \mathcal{O}\left(\varepsilon^{-(p^{\text{std}}+\delta)}\right),$$

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

Proof. For $r(\boldsymbol{\gamma}) > 1/2$, Theorem 5.2 for the class Λ^{all} states that the exponent of strong polynomial tractability is $p^{\text{all}} = 1/r(\boldsymbol{\gamma})$. 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 [13, Theorem 5] as in the proof of Theorem 4.1, it follows that for any $\delta_1 \in (0,1)$ we have

$$e^{\text{wor-std}}(n, \mathcal{H}_d) = \mathcal{O}\left(n^{-(1-\delta_1)p^2/(2p+2)}\right) = \mathcal{O}\left(n^{-(1-\delta_1)(1+\mathcal{O}(\eta))2r^2(\boldsymbol{\gamma})/(2r(\boldsymbol{\gamma})+1)}\right)$$
$$= \mathcal{O}\left(n^{-1/p^{\text{std}}+\delta}\right),$$

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 Λ^{all} and Λ^{std}. For an exponentially decaying sequence of shape parameters, say, $\gamma_\ell = q^{\,\ell}$ for some $q \in (0,1)$, we have $p^{\text{all}} = p^{\text{std}} = 0$, 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 $\mathcal{H}_d(K_d)$ with a Gaussian kernel for the normalized error criterion. That is, we want to find the smallest n for which

$$e^{\text{wor-}\vartheta}(n, \mathcal{H}_d) \le \varepsilon \, \|I_d\|, \qquad \vartheta \in \{\text{std}, \text{all}\}.$$

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 Λ^{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 6.1. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d \in \mathbb{N}}$ for Hilbert spaces with isotropic Gaussian kernels for the class Λ^{all} and for the normalized error criterion. Then we have the following:*

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

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

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

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

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 [15, Theorem 5.6]. In fact, the lack of polynomial tractability for the class Λ^{all} holds for all tensor product problems with two positive eigenvalues for the univariate case.

For quasi-polynomial tractability we use [8, Theorem 3.3], which states that quasi-polynomial tractability for the class Λ^{all} holds for tensor product problems if and only if the rate

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

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

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

In our case, $r = \infty$ and

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

The estimates of $e^{\text{wor-all}}(n, \mathcal{H}_d)$ and $n^{\text{wor-nor-all}}(\varepsilon, \mathcal{H}_d)$ 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

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

For some specific values of γ we have

γ	$2^{-1/2}$	1	$2^{1/2}$
$t^{\mathrm{all}}(\gamma)$	$1.5186\ldots$	$2.0780\ldots$	$2.8853\ldots$

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 Λ^{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 5.2.

THEOREM 6.2. *Consider the function approximation problem $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ for Hilbert spaces with anisotropic Gaussian kernels for the class Λ^{all} and for the normalized error criterion. Then we have the following:*

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

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

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

$$e^{\text{wor-all}}(n,\mathcal{H}_d) = \mathcal{O}\left(\|I_d\| n^{-1/p^{\text{all}}+\delta}\right) = \mathcal{O}\left(n^{-r(\boldsymbol{\gamma})+\delta}\right),$$
$$n^{\text{wor-nor-all}}(\varepsilon,\mathcal{H}_d) = \mathcal{O}\left(\varepsilon^{-(p^{\text{all}}+\delta)}\right),$$

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*

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

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

Proof. [15, Theorem 5.2] states that strong polynomial tractability holds if and only if 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 $n^{\text{wor-nor-all}}(\varepsilon,\mathcal{H}_d) \leq \tilde{C}_2\,\varepsilon^{-2\tau}$ 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, $\tilde{C}_2 < \infty$ if and only if

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

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

The case of polynomial tractability for ordered shape parameters follows analogously from the proof in Theorem 5.2. From [15, Theorem 5.2], we know that the problem is polynomially tractable with $n^{\text{wor-nor-all}}(\varepsilon,\mathcal{H}_d) = \mathcal{O}\left(\varepsilon^{-2\tau}\, d^{\,q_2\tau}\right)$ if and only if

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

Proceeding as in the proof of Theorem 5.2, 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 6.2 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 5.2 and 6.2. This is the subject of the next corollary.

COROLLARY 6.3. *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 Λ^{all}. Let $r(\boldsymbol{\gamma})$ be the rate of convergence of shape parameters. Then we have the following:*

- *Absolute error criterion: $\mathcal{I}$ is always strongly polynomially tractable with exponent*

$$p^{\mathrm{all}} = \min\left(2, \frac{1}{r(\boldsymbol{\gamma})}\right) \leq 2.$$

- *Normalized error criterion: $\mathcal{I}$ is strongly polynomially tractable if and only if $r(\boldsymbol{\gamma}) > 0$. If so, the exponent is*

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

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

6.3. Only function values. We now turn to the class Λ^{std}. We do not know if quasi-polynomial tractability holds for the class Λ^{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 (5.1) 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 (4.1), 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 Λ^{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 $\beta \leq -\frac{1}{2}$.

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 5.4 for the class Λ^{std} with ε replaced by $\varepsilon\|I_d\| = \Theta(\varepsilon)$. For $r(\boldsymbol{\gamma}) = 1/2$, Theorem 5.3 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 6.4. *Consider the function approximation problem* $\mathcal{I} = \{I_d\}_{d\in\mathbb{N}}$ *for Hilbert spaces with anisotropic Gaussian kernels for the class* Λ^{std} *and for the normalized error criterion. Assume that*

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

Then

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

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

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

$$e^{\text{wor-all}}(n, \mathcal{H}_d) = \mathcal{O}\left(n^{-1/(p^{\mathrm{all}}+\delta)}\right),$$
$$n^{\text{wor-nor-all}}(\varepsilon, \mathcal{H}_d) = \mathcal{O}\left(\varepsilon^{-(p^{\mathrm{all}}+\delta)}\right),$$

where the factors in the big $\mathcal{O}$ *notation 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 Λ^{std} in this case.

REFERENCES

[1] A. BERLINET AND C. THOMAS-AGNAN, *Reproducing Kernel Hilbert Spaces in Probability and Statistic*, Kluwer Academic, Boston, 2004.

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

[3] H.-J. BUNGARTZ AND M. GRIEBEL, *Sparse grids*, Acta Numer., 13 (2004), pp. 1–123.

[4] F. CUCKER AND D. X. ZHOU, *Learning Theory: An Approximation Theory Viewpoint*, Camb. Monogr. Appl. Comput. Math., Cambridge University Press, Cambridge, UK, 2007.

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

[6] G. E. FASSHAUER AND M. J. MCCOURT, *Stable evaluation of Gaussian RBF interpolants*, SIAM J. Sci. Comput., to appear.

[7] A. I. J. FORRESTER, A. SÓBESTER, AND A. J. KEANE, *Engineering Design via Surrogate Modelling: A Practical Guide*, Wiley, Chichester, 2008.

[8] M. GNEWUCH AND H. WOŹNIAKOWSKI, *Quasi-polynomial tractability*, J. Complexity, 27 (2011), pp. 312–330.

[9] M. GOLOMB AND H. F. WEINBERGER, *Optimal approximation and error bounds*, in On Numerical Approximation, R. E. Langer, ed., University of Wisconsin Press, 1959, pp. 117–190.

[10] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Ser. Statist., Springer, New York, 2009.

[11] J. K. HUNTER AND B. NACHTERGAELE, *Applied Analysis*, World Scientific, Singapore, 2001.

[12] *JMP* 9.0, SAS Institute, Cary, NC, 2010.

[13] F. KUO, G. W. WASILKOWSKI, AND H. WOŹNIAKOWSKI, *On the power of standard information for multivariate approximation in the worst case setting*, J. Approx. Theory, 158 (2009), pp. 97–125.

[14] W. R. MADYCH AND S. A. NELSON, *Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation*, J. Approx. Theory, 70 (1992), pp. 94–114.

[15] E. NOVAK AND H. WOŹNIAKOWSKI, *Tractability of Multivariate Problems, Volume* I: *Linear Information*, European Mathematical Society, Zürich, 2008.

[16] E. NOVAK AND H. WOŹNIAKOWSKI, *Tractability of Multivariate Problems, Volume* II: *Standard Information for Functionals*, European Mathematical Society, Zürich, 2010.

[17] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006; also available online from http://www.gaussianprocess.org/gpml.

[18] C. Rieger and B. Zwicknagl, *Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning*, Adv. Comput. Math., 32 (2008), pp. 103–129.

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

[20] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2002.

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

[22] M. L. Stein, *Interpolation of Spatial Data. Some Theory for Kriging*, Springer-Verlag, New York, 1999.

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

[24] I. Steinwart, D. Hush, and C. Scovel, *An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4635–4663.

[25] G. Szegő, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1959.

[26] J. F. Traub, G. W. Wasilkowski, and H. Woźniakowski, *Information-Based Complexity*, Academic Press, New York, 1988.

[27] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. Appl. Math. 59, SIAM, Philadelphia, 1990.

[28] G. W. Wasilkowski and H. Woźniakowski, *Explicit error bounds of algorithms for multivariate tensor product problems*, J. Complexity, 11 (1995), pp. 1–56.

[29] G. W. Wasilkowski and H. Woźniakowski, *On the power of standard information for weighted approximation*, Found. Comput. Math., 1 (2001), pp. 417–434,

[30] H. Wendland, *Gaussian interpolation revisited*, in Trends in Approximation Theory, K. Kopotun, T. Lyche, and M. Neamtu, eds., Vanderbilt University Press, Nashville, TN, 2001, pp. 417–426.

[31] H. Wendland, *Scattered Data Approximation*, Cambridge Monogr. Appl. Comput. Math. 17, Cambridge University Press, Cambridge, UK, 2005.