Journal of Machine Learning Research 26 (2025) 1-30 Submitted 2/24; Published 3/25

On the Approximation of Kernel functions

Paul Dommel PAUL.DOMMEL@MATH.TU-CHEMNITZ.DE
Faculty of Mathematics
University of Technology, Chemnitz
09126, Chemnitz, Germany

Alois Pichler ALOIS.PICHLER@MATH.TU-CHEMNITZ.DE
Faculty of Mathematics
University of Technology, Chemnitz
09126, Chemnitz, Germany

Editor: Gabor Lugosi

Abstract

Various methods in statistical learning build on kernels considered in reproducing kernel Hilbert spaces. In applications, the kernel is often selected based on characteristics of the problem and the data. This kernel is then employed to infer response variables at points, where no explanatory data were observed.

The data considered here are located in compact sets in higher dimensions and the paper addresses approximations of the kernel itself. The new approach considers Taylor series approximations of radial kernel functions. For the Gauss kernel on the unit cube, the paper establishes an upper bound of the associated eigenfunctions, which grows only polynomially with respect to the index. The novel approach substantiates smaller regularization parameters than considered in the literature, overall leading to better approximations. This improvement confirms low rank approximation methods such as the Nyström method.

Keywords: statistical learning, kernel methods, reproducing kernel Hilbert spaces, Nyström method

1. Introduction

This paper contributes to statistical methods building on reproducing kernel Hilbert spaces. These methods have become popular in statistical learning, in inference and in support vector machines due to the kernel trick. They constitute powerful tools in different scientific areas such as geostatistics (cf. Honarkhah and Caers 2010), stochastic optimization (cf. Dommel and Pichler 2023, Park et al. 2022), digit recognition (cf. Schölkopf 1997), computer vision (cf. Zhang et al. 2007) and bio informatics (cf. Schölkopf et al. 2004).

The approach presented here approximates the kernel function by elements in the range of the associated Hilbert–Schmidt integral operator. We choose these elements so that its Taylor series expansion matches the initial coefficients. The method applies for general radial kernels with variable bandwidth. Special emphasize is given to the Gaussian kernel, which is of major importance in practical applications.

Fundamental in statistical approximation is the regularization parameter. Standard results suggest regularization parameters decreasing as $\mathcal{O}(1/n)$, where n is the sample size.

This paper, in contrast, justifies significantly smaller regularization parameters, often close to machine precision. This leads to enhanced approximation quality.

An additional consequence of our approach, not sufficiently addressed in the literature yet, are the magnitudes of the eigenfunctions of the related Hilbert–Schmidt operator. We demonstrate that this magnitude grows only polynomially in the index. Exploiting this crucial property, we present an interpolation inequality, which allows bounding the uniform error by the much weaker L^2-error.

The approximation of the kernel function builds on the function $w_m^x(\cdot)$, smallest in $L^2([0,1])$-norm, satisfying the moment constraints

$$\int_0^1 z^\ell w_m^x(z)\,dz = x^\ell, \qquad \ell = 0, \dots, m-1. \tag{1.1}$$

We provide the explicit solution, which is a polynomial with coefficients involving the Hilbert matrix.

The results have consequences for low rank kernel methods. For these methods, they ensure a stable approximation quality by building only on few supporting points. Prominent examples of these methods include the Nyström algorithms and kernel principal component analysis.

Related literature. Our results address the approximation of randomly located kernel functions. This is of particular importance for low rank kernel methods, which build on an approximation of the kernel matrix itself. The Nyström method, introduced in Williams and Seeger (2000), is a prominent example for this technique. Drineas and Mahoney (2005) analyze the error of the matrix approximation in the Nyström method and relate it to the best approximating matrix, while Bach (2013) studies the precision of predictions directly. The excellent work of Rudi et al. (2015) relates the rank of the approximating matrix to the approximation of kernel functions. The result is then employed to develop a low rank regression approach, which is significantly cheaper in computations than kernel ridge regression while maintaining stable prediction accuracy, cf. Rudi et al. (2017). Kernel principal component analysis builds on these results as well, cf. Sterge and Sriperumbudur (2021).

The approximation of kernel functions is the core research question of this paper, for which we present new bounds. The second main result of this work addresses the eigensystem of Mercer's decomposition associated with the Gaussian kernel. Shi et al. (2009) address this issue for an unbounded domain, building on the normal distribution as underlying design measure. The authors provide an explicit expression of the eigenvalues and eigenfunctions by involving the Hermite polynomials in an unbounded domain. In compact domains, Diaconis et al. (2008) consider the eigensystem for the Laplacian kernel.

The analysis of the divide and conquer approach relies on properties of the eigenfunctions as well, cf. Zhang et al. (2013), but the paper builds on unverified assumptions. The analysis of the regression error in different norms (cf. Fischer and Steinwart 2020 and Steinwart et al. 2009) can be related to bounded eigenfunctions as well. This paper presents explicit bounds of the maximum value the eigenfunction may attain.

Outline of the paper. Section 2 addresses polynomial approximations of kernel functions. Section 3 addresses the Gaussian kernel specifically and presents bounds of the eigenfunctions on bounded domains. This section contains the main results, which are considered in Section 4 in applications. Section 5 concludes the paper.

2. The minimal moment function

Reproducing kernel Hilbert spaces (RKHS) build on a kernel function, denoted k. The approximations considered here build on the point evaluation function $k_x(\cdot) \coloneqq k(x,\cdot)$. This section resumes the minimal moment function w_m^x from (1.1), for which the image $L_k w_m^x$ is a suitable estimate of k_x. We first derive the function w_m^x for the simple design space $\mathcal{X} = [0,1]$, and then extend it to the d-dimensional case with some non-uniform design measure P. Using the moment property (1.1), we derive an error bound for the approximation quality of $k_x(\cdot)$ by $L_k w_m^x$, which is based on the Taylor coefficients of the kernel.

2.1 The minimal moment function

The central element of this paper is the function with smallest L^2-norm, satisfying the moment properties (1.1). There are infinitely many functions fulfilling the condition (1.1). We refer to the function with smallest L^2-norm as the *minimal moment function*, where the inner product is $\langle f, g\rangle_{L^2} \coloneqq \int_{\mathcal{X}} f(z)g(z)p(z)\,dz$ with the density p of the underlying measure P. Throughout this section, we consider the design space $\mathcal{X} = [0,1]$ equipped with the uniform measure $P \sim \mathcal{U}[0,1]$.

In what follows we provide an explicit representation of the minimal moment function. We demonstrate that it is a polynomial of degree $m-1$, with coefficients originating from a Hilbert matrix.

Theorem 1 (Explicit minimal moment function)**.** *For $x \in [0,1]$ fixed, the optimization problem*

$$\vartheta^* \coloneqq \min\left\{ \|w\|_{L^2}^2 : \int_0^1 z^\ell w(z)dz = x^\ell,\ \ell = 0,1,\dots,m-1 \right\} \tag{2.1}$$

has the unique solution

$$w_m^x(z) = \sum_{i=1}^m \alpha_{x,i} z^{i-1}, \qquad z \in (0,1), \tag{2.2}$$

where α_x satisfies the equations $H_m \alpha_x = \bar{x}$ for the Hilbert matrix $H_m \coloneqq \left(\frac{1}{i+j-1}\right)_{i=1,j=1}^{n,n}$ and the vector $\bar{x} \coloneqq (1, x, \dots, x^{m-1}) \in \mathbb{R}^m$.

Proof The Lagrangian $\mathcal{L}$ of (2.1) is

$$\mathcal{L}(w,\mu) = \langle w(z), w(z)\rangle_{L^2} + \sum_{i=1}^m \mu_i \left(\langle z^{i-1}, w(z)\rangle_{L^2} - x^{i-1} \right),$$

where $w \in L^2$ and $\mu = (\mu_1, \dots, \mu_m) \in \mathbb{R}^m$ are Lagrange multipliers. The first order condition reads

$$(\nabla_w \mathcal{L})(w_\mu^*, \mu) = 2w_\mu^* + \sum_{i=1}^m \mu_i z^{i-1} = 0,$$

which is equivalent to $w_\mu^*(z) = -\frac{1}{2}\sum_{i=1}^m \mu_i z^{i-1}$. For $\mu = (\mu_1, \dots, \mu_m)$ fixed, the Lagrangian function $\mathcal{L}$ is convex and w_μ^* thus a minimizer. Hence, the Lagrangian dual function is

$$g(\mu) \coloneqq \min_{w \in L^2} \mathcal{L}(w,\mu) = \mathcal{L}(w_\mu^*, \mu) = \langle w_\mu^*, w_\mu^* \rangle_{L^2} + \sum_{i=1}^m \mu_i \left(\langle z^{i-1}, w_\mu^* \rangle_{L^2} - x^{i-1} \right),$$

depending only on the multipliers μ. As $\langle z^{i-1}, z^{j-1} \rangle_{L^2} = \frac{1}{i+j-1}$, we further have that

$$\langle z^{i-1}, w_\mu^* \rangle_{L^2} - x^{i-1} = -\frac{1}{2}\sum_{i=1}^m \mu_i^* \frac{1}{j+i-1} - x^{j-1} = 0$$

by setting $\mu^* = -2H_m^{-1}\bar{x}$. It follows that

$$g(\mu^*) = \langle w_{\mu^*}^*, w_{\mu^*}^* \rangle_{L^2} \geq \vartheta^*,$$

as $w_{\mu^*}^*$ is feasible in (2.1). This implies strong duality as well as the optimality of $w_{\mu^*}^* = w_m^x$, which is the assertion. ■

A minimal moment function of particular interest is the optimizer of (2.1) associated with the point $x = 1$. In contrast to the general case $x \in [0,1]$, the L^2-norm of w_m^1 can be computed explicitly. Indeed, it holds that

$$\begin{aligned}\int_0^1 w_m^1(z)^2 dz &= \int_0^1 \sum_{i=1}^m \sum_{j=1}^m \alpha_{1,i}\alpha_{1,j} z^{i+j-2} dz \\ &= \sum_{i=1}^m \sum_{j=1}^m \alpha_{1,i}\alpha_{1,j} \frac{1}{i+j-1} \\ &= \sum_{i=1}^m \alpha_{1,i} 1 = \alpha_1^\top e = e^\top H_m^{-1} e = m^2, \end{aligned} \tag{2.3}$$

as $\sum_{i,j=1}^n (H_m^{-1})_{i,j} = m^2$.

In what follows we bound the norm of the remaining moment functions. For that we relate w_m^x with a specific linear transformation of w_m^1.

Theorem 2 (Upper bound of the weight function)**.** *The weight function* w_m^x *satisfies the bound*

$$\|w_m^x\|_{L^2}^2 \leq m^2 \tag{2.4}$$

for any $x \in [0,1]$.

Proof Note first that, for $x = 0$,

$$\int_0^1 w_m^0(z)^2 dz = e_1^\top H_m^{-1} e_1 = \left(H_m^{-1}\right)_{1,1} = m^2,$$

where $e_1 = (1, 0, \dots .0)$ is the first vector in the canonical basis. To verify the bound for the remaining points $x \in (0,1)$ we relate w_m^x with an auxiliary function $\tilde{w}_m^x$ for which we are able to calculate the norm more easily. Setting $g^x(z) := w_m^1(\frac{z}{x})\mathbb{1}_{[0,x]}(z)$ we define the auxiliary function

$$\tilde{w}_m^x(z) := \begin{cases} g^x(z) & \text{if } z \leq x, \\ g^{1-x}(1-z) & \text{if } z > x, \end{cases}$$

with $x \in (0,1)$. To relate w_m^x with $\tilde{w}_m^x$, we demonstrate first that $\tilde{w}_m^x$ satisfies the moment constraints (1.1) and bound its norm afterwards. It holds that

$$\int_0^1 z^h g^x(z)dz = \int_0^x z^h\, w_m^1\Big(\frac{z}{x}\Big)\, dz = x\int_0^1 (yx)^h w_m^1(y)dy = xx^h \tag{2.5}$$

for the first part of the integral. We further have that

$$\int_0^1 z^h g^{1-x}(1-z)dz = -\int_{1-x}^0 (1-y)^h g^{1-x}(y)dz = \int_0^{1-x} (1-y)^h g^{1-x}(y)dy$$

after changing the variables. By the binomial theorem,

$$\begin{aligned}\int_0^{1-x} (1-y)^h g^{1-x}(y)dy &= \sum_{p=0}^h \binom{h}{p}(-1)^{h-p}\int_0^{1-x} y^{h-p} g^{1-x}(y)dy \\ &= \sum_{p=0}^h \binom{h}{p}(-1)^{h-p}(1-x)(1-x)^{h-p} \\ &= (1-x)(1-(1-x))^h = (1-x)x^h,\end{aligned}$$

as $\int_0^{1-x} y^{h-p}g^{1-x}(y)dy = (1-x)(1-x)^{h-p}$ by (2.5). Connecting both identities, we have that

$$\int_0^1 z^h\tilde{w}_m^x(z)dz = \int_0^x z^h g^x(z)dz + \int_0^{1-x}(1-y)^h g^{1-x}(y)dy = x^h,$$

and thus the moment property (1.1) of $\tilde{w}_m^x$. It is now evident by (2.1) that $\|\tilde{w}_m^x\|_{L^2}$ is an upper bound of $\|w_m^x\|_{L^2}$. Employing the same substitutions as above, we get that

$$\begin{aligned}\int_0^1 \tilde{w}_m^x(z)^2dz &= \int_0^x g^x(z)^2dz + \int_x^1 g^{1-x}(1-z)^2dz \\ &= x\int_0^1 w_m^1(y)^2dy + (1-x)\int_0^1 w_m^1(y)^2dz \\ &= \int_0^1 w_m^1(y)^2dy = m^2\end{aligned}$$

by (2.3), concluding the proof. ■

The squared norm of the moments functions at the boundary points is m^2. However, the norm of the minimal moment functions associated with the interior points $x \in (0,1)$ might be significantly smaller.

2.2 Extensions and error estimates

The results of the preceding Section 2.1 crucially rely on the proposed setting, i.e., the design space $[0,1]$ equipped with the uniform distribution. These assumptions are quite restrictive and need to be relaxed for situations of practical application. To this end we investigate a more general setting throughout this section.

We consider the multivariate case where $\mathcal{X} = [0,1]^d$, with some underlying design measure P. This measure has a strictly positive density with

$$\infty > C > \sup_{z \in [0,1]^d} p(z) \geq \inf_{z \in [0,1]^d} p(z) > c > 0,$$

giving rise to the inner product

$$\langle f, g \rangle_{L^2} = \int_{[0,1]^d} f(z) g(z) p(z) dz.$$

In what follows we specify the minimal moment functions for this more general setting. Building on the univariate moment property of the functions (2.2), we demonstrate that their product satisfies a multivariate version of (1.1). The next proposition reveals the precise statement.

Proposition 1 (Upper bound of the weight function in higher dimensions)**.** Let $x = (x_1, \dots, x_d) \in [0,1]^d$ and consider the function

$$W_m^x(z_1, \dots, z_d) := \left(\prod_{i=1}^d w_m^{x_i}(z_i) \right) p(z_1, \dots, z_n)^{-1}, \tag{2.6}$$

where $w_m^{x_i}$ are the functions defined in (2.2). The function W_m^x satisfies the general moment property

$$\int_{[0,1]^d} \left(\|y - z\|_2^2 \right)^\ell W_m^x(z) p(z) dz = \left(\|y - x\|_2^2 \right)^\ell \tag{2.7}$$

for all integers $\ell \leq \frac{m}{2}$. Its norm is bounded by $\|W_m^x\|_{L^2}^2 \leq c_p m^{2d}$ with $c_p = \sup_{z \in [0,1]^d} p(z)^{-1}$.

Proof The moment property (2.7) follows from

$$\begin{aligned}
\int_{[0,1]^d} \left(\|y - z\|_2^2 \right)^\ell W_m^x(z) p(z) dz &= \int_0^1 \cdots \int_0^1 \left(\sum_{i=1}^d (y_i - z_i)^2 \right)^\ell \prod_{i=1}^d w_m^{x_i}(z_i) dz_1 \dots dz_d \\
&= \sum_{h_1 + \dots + h_d = \ell} \binom{\ell}{h_1, \dots, h_d} \prod_{i=1}^d \int_0^1 (y_i - z_i)^{2h_i} w_m^{x_i}(z_i) dz_i \\
&= \sum_{h_1 + \dots + h_d = \ell} \binom{\ell}{h_1, \dots, h_d} \prod_{i=1}^d (y_i - x_i)^{2h_i} \\
&= \left(\|y - x\|_2^2 \right)^\ell,
\end{aligned}$$

as $\int_0^1 z_i^{\ell_i} w_m^{x_i}(z_i) = x_i^{\ell_i}$ holds for all integers $\ell_i \leq m - 1$. This is the first assertion.

For the second assertion note that $\|w_m^{x_i}\|_{L^2}^2 \le m^2$ holds by (2.4). Hence, we get that

$$\begin{aligned}\int_{[0,1]^d} (W_m^x(z))^2 p(z)dz &= \int_0^1 \cdots \int_0^1 \left(\prod_{i=1}^n w_m^{x_i}(z_i)\right)^2 p(z_1,\dots,z_n)^{-1}dz_1\dots dz_d \\ &\le \sup_{z\in[0,1]^d} \left|p(z)^{-1}\right| \prod_{i=1}^d \int_0^1 (w_m^{x_i}(z_i))^2 dz_i = c_p m^{2d},\end{aligned}$$

which concludes the proof. ■

Remark 1. The function $W_m^x(\cdot)$ might not be the norm minimal function satisfying (2.7). Indeed, the product structure of (2.6) leads to an exponentially (with respect to the dimension) increasing norm of W_m^x. However, the construction of the norm minimal moment function satisfying (2.7) might require a significantly more involved representation, which is out of the scope of this paper.

Continuing the general setting considered above, we now provide the first error estimate. To this end we consider a radial kernel

$$k(x,y) = \phi(\|x-y\|^2) \tag{2.8}$$

as well as the corresponding integral operator $L_k\colon L^2(\mathcal{X},P) \to L^2(\mathcal{X},P)$ defined as

$$(L_k f)(y) = \int_{\mathcal{X}} k(z,y)f(z)p(z)dz. \tag{2.9}$$

Here, ϕ is a smooth function with Taylor series expansion

$$\phi(x) = \sum_{\ell=0}^{\infty} \frac{\alpha_\ell}{\ell!} x^\ell. \tag{2.10}$$

Building on the moment property (2.7), we utilize that the ℓth moment of $k_x - L_k W_m^x$ vanishes. The subsequent theorem reveals the precise bound.

Theorem 3 (Uniform bound in d-dimensions)**.** *With the function W_m^x defined in* (2.6)*, the error estimate*

$$\sup_{x\in[0,1]^d} \|(L_k W_m^x)(y) - k_x\|_\infty \le (1 + c_p^{1/2} m^d) \sum_{\ell=\lfloor\frac{m-1}{2}\rfloor+1}^{\infty} \frac{|a_\ell|}{\ell!} d^\ell, \tag{2.11}$$

holds true. Here, $\lfloor\cdot\rfloor$ denotes the floor function.

Proof Employing the series representation (2.10) of ϕ, we have the decomposition

$$\begin{aligned}(L_k W_m^x)(y) &= \int_{\mathcal{X}} \phi\left(\|y-z\|_2^2\right) W_m^x(z)p(z)dz \\ &= \int_{\mathcal{X}} \sum_{\ell=0}^{\lfloor\frac{m-1}{2}\rfloor} \frac{a_\ell}{\ell!} (\|y-z\|_2^2)^\ell W_m^x(z)p(z)dz \\ &+ \int_{\mathcal{X}} \sum_{\ell=\lfloor\frac{m-1}{2}\rfloor+1}^{\infty} \frac{a_\ell}{\ell!} (\|y-z\|_2^2)^\ell W_m^x(z)p(z)dz.\end{aligned}$$

For the first part, it follows from the moment property (2.7) of W_m^x that

$$\int_{\mathcal{X}} \sum_{\ell=0}^{\lfloor \frac{m-1}{2} \rfloor} \frac{a_\ell}{\ell!} (\|y-z\|_2^2)^\ell W_m^x(z) p(z) dz = \sum_{\ell=0}^{\lfloor \frac{m-1}{2} \rfloor} \frac{a_\ell}{\ell!} \left(\|y-x\|_2^2\right)^\ell,$$

as the $2 \left\lfloor \frac{m-1}{2} \right\rfloor \leq m-1$. Hence, we have that

$$\begin{aligned}
&|k(y,x) - (L_k W_m^x)(y)| \\
&\quad = \left| k(y,x) - \sum_{\ell=0}^{\lfloor \frac{m-1}{2} \rfloor} \frac{a_\ell}{\ell!} (\|y-x\|_2^2)^\ell + \int_{\mathcal{X}} \sum_{\ell=\lfloor \frac{m-1}{2} \rfloor+1}^{\infty} \frac{a_\ell}{\ell!} (\|y-z\|_2^2)^\ell W_m^x(z) p(z) dz \right| \\
&\quad = \left| \sum_{\ell=\lfloor \frac{m-1}{2} \rfloor+1}^{\infty} \frac{a_\ell}{\ell!} (\|y-x\|_2^2)^\ell + \int_{\mathcal{X}} \sum_{\ell=\lfloor \frac{m-1}{2} \rfloor+1}^{\infty} \frac{a_\ell}{\ell!} (\|y-z\|_2^2)^{2\ell} W_m^x(z) p(z) dz \right| \\
&\quad \leq \sum_{\ell=\lfloor \frac{m-1}{2} \rfloor+1}^{\infty} \frac{|a_\ell|}{\ell!} d^\ell + \sum_{\ell=\lfloor \frac{m-1}{2} \rfloor+1}^{\infty} \frac{|a_\ell|}{\ell!} d^\ell \, \|W_m^x\|_{L^2} \\
&\quad \leq \sum_{\ell=\lfloor \frac{m-1}{2} \rfloor+1}^{\infty} \frac{|a_\ell|}{\ell!} d^\ell + c_p^{1/2} m^d \sum_{\ell=\lfloor \frac{m-1}{2} \rfloor+1}^{\infty} \frac{|a_\ell|}{\ell!} d^\ell.
\end{aligned}$$

The assertion follows, as the result holds for each $x \in [0,1]^d$. ■

The statement of Theorem 3 above applies for general translation invariant kernels of the shape $k(x,y) = \phi(\|x-y\|^2)$. To further specify the associated error bound, one needs to include knowledge about the decay of the Taylor coefficients of ϕ. To this end we now consider a fixed kernel for which these coefficients and their behavior are known.

3. The Gaussian kernel

The following results build on reproducing kernel Hilbert spaces (RKHS). For these spaces, point evaluations are continuous linear functionals, and this property is the decisive characteristic of RKHS. Each kernel function addressed above is associated with the corresponding space $\big(\mathcal{H}_k,\, \langle\cdot|\cdot\rangle_k\big)$, for which $\langle f|\, k(x,\cdot)\rangle_k = f(x)$ whenever $f \in \mathcal{H}_k$. For a detailed review of these spaces we refer to Berlinet and Thomas-Agnan (2004) or Steinwart and Christmann (2008).

This section addresses the most popular kernel in machine learning, the Gaussian kernel

$$k(x,y) \coloneqq e^{-\sigma\|x-y\|_2^2} = \phi\big(\sigma \cdot \|x-y\|_2^2\big), \tag{3.1}$$

where $\sigma > 0$ is a width parameter. We approximate the point evaluation function in the range of the Hilbert–Schmidt operator and analyze its error with respect to different norms. Building on these estimates, we derive essential properties of the problem

$$\inf_{w \in \mathcal{H}_k} \lambda \|w\|_{L^2}^2 + \|L_k w - k_x\|_k^2, \tag{3.2}$$

which will be used in what follows. We employ these results to establish polynomial bounds on the magnitude of the eigenfunctions of the Gaussian kernel.

3.1 Approximation of the point evaluation function

In this section we investigate the approximation of the point evaluation function k_x for the Gaussian kernel. We relate the function k_x with an approximation from the image of the corresponding integral operator L_k and provide bounds with respect to the infinity norm as well as the norm of the RKHS. The following theorem provides the result for the uniform norm first.

Theorem 4 (Uniform approximation of k_x in $\|\cdot\|_\infty$)**.** *Let k be the d-dimensional Gaussian kernel with width parameter σ. Setting $c_\sigma := \max\{1, 2e\sigma d\}$, $c_p = \sup_{z\in[0,1]^d} p(z)^{-1}$ and*

$$C(\sigma, m) := \frac{1}{1 - \frac{\sigma e d}{\lfloor \frac{m}{2} \rfloor}},$$

the uniform bound

$$\sup_{x\in[0,1]^d} \|L_k W_m^x - k_x\|_\infty \le (1 + c_p^{1/2} m^d)\, C(\sigma, m) \left(\left\lfloor \frac{m}{2} \right\rfloor \frac{1}{\sigma e d} \right)^{\lfloor \frac{m}{2} \rfloor}$$

holds for W_m^x defined in (2.6) *for $m > c_\sigma + 1$.*

Specifically, for $m(s) := 3c_\sigma s + 2$, we have that

$$\sup_{x\in[0,1]^d} \left\| L_k W_{m(td)}^x - k_x \right\|_\infty \le 3(t\, d)^{-3t\, d} \tag{3.3}$$

whenever $t \ge \max\left\{ \frac{\ln(3) + (d-1)\ln(2) + \frac{1}{2}\ln(c_p) + d\ln(3c_\sigma d)}{2d\ln(3)},\ 1 \right\}$.

Proof We defer the proof to Appendix A. ■

The uniform bound (3.3) relates $L_k W_m^x$ and k_x for all points $x \in [0,1]^d$. However, the uniform norm is slightly too weak when studying the objective (3.2), which relies on the approximation in RKHS norm. To overcome this issue we extend the result obtained and establish a dedicated bound, similar to (3.3), but with respect to the stronger norm $\|\cdot\|_k$. The next proposition reveals the desired bound.

Proposition 2 (Uniform approximation of k_x in the norm $\|\cdot\|_k$)**.** Let $m = m(s) = 3c_\sigma s + 2$ and $t \ge c := \max\{c_0, c_1, c_2, 1\}$, where

$$c_0 = \frac{\ln(3) + (d-1)\ln(2) + \frac{1}{2}\ln(c_p) + d\ln(3c_\sigma d)}{2d\ln(3)}.$$
$$c_1 = \frac{(2d-1)\ln(2) + d\ln(3c_\sigma) + \frac{1}{2}\ln(c_p)}{d} + 1,$$
$$c_2 = \frac{(2d-1)\ln(2) + \frac{1}{2}\ln(c_p)}{d}.$$

In the setting of Theorem (4), the uniform bound in $\|\cdot\|_k$-norm,

$$\sup_{x\in[0,1]^d} \left\| L_k W_{m(td)}^x - k_x \right\|_k^2 \le 9(td)^{-2td}, \tag{3.4}$$

holds true.

Proof Let $x \in [0,1]^d$ and observe from (3.3) and the reverse triangle inequality that

$$(L_k W_m^x)(x) \geq k(x,x) - \|L_k W_m^x - k_x\|_\infty \geq k(x,x) - 3(td)^{-3td}, \tag{3.5}$$

whenever m and t are chosen appropriately (see Theorem 4). Hence, setting $m = \lceil 3c_\sigma td \rceil + 2$, it follows from (3.5) that

$$\begin{aligned}\|L_k W_m^x - k_x\|_k^2 &= \langle L_k W_m^x, L_k W_m^x \rangle_k - 2\, \langle L_k W_m^x, k_x \rangle_k + k(x,x) \\ &= \langle L_k W_m^x, L_k W_m^x \rangle_k - 2\, (L_k W_m^x)\,(x) + k(x,x) \\ &\leq \langle L_k W_m^x, W_m^x \rangle_{L^2} - (L_k W_m^x)\,(x) + 3(td)^{-3td},\end{aligned}$$

as $\langle L_k W_m^x, k_x \rangle_k = (L_k W_m^x)\,(x)$ holds by the reproducing property. Furthermore, we have that

$$\langle L_k W_m^x, W_m^x \rangle_{L^2} = \langle L_k W_m^x - k_x, W_m^x \rangle_{L^2} + \langle k_x, W_m^x \rangle_{L^2} = \langle L_k W_m^x - k_x, W_m^x \rangle_{L^2} + (L_k W_m^x)\,(x).$$

Employing the bound

$$\langle L_k W_m^x - k_x, W_m^x \rangle_{L^2} \leq 6(dt)^{-2dt}$$

from Lemma A.7 in the appendix and combining the estimates above, we get

$$\|L_k w_m^x - k_x\|_k^2 \leq 6(td)^{-2td} + 3(td)^{-3td} \leq 9(td)^{-2td},$$

and thus the assertion. ■

Remark 2. The bound (3.4) depends only on the magnitude of the product td. Substituting $s = td$ gives the more convenient bound

$$\sup_{x \in [0,1]^d} \|L_k W_m^x - k_x\|_k^2 \leq 9s^{-2s}, \tag{3.6}$$

where $m = m(s) = 3c_\sigma s + 2$ and $s \geq d \cdot c$.

3.2 Implications for the weight function

Building on the bound (3.4) of Section 3.1, we now examine the optimal weight function w_λ^x solving the optimization problem

$$\inf_{w \in L^2} \lambda\, \|w\|_{L^2}^2 + \|L_k w - k_x\|_k^2\,,$$

cf. (3.2). This optimal solution w_λ^x , and its norm, determine the approximation quality of the point evaluation k_x in the range of L_k. Moreover, they relate the continuous operator L_k with discrete versions, which we address and discuss in the following sections.

By Mercer's theorem, the kernel k enjoys the representation $k(x,y) = \sum_{\ell=1}^\infty \mu_\ell\, \varphi_\ell(x)\varphi_\ell(y)$, where μ_ℓ are the eigenvalues (in non-increasing order) and φ_ℓ the eigenfunctions, $\ell = 1, 2, \dots,$ of the operator L_k introduced in (2.9). The explicit representation

$$w_\lambda^x(y) = \big((\lambda + L_k)^{-1} L_k k_x\big)(y) = \sum_{\ell=1}^\infty \frac{\mu_\ell}{\lambda + \mu_\ell} \varphi_\ell(x)\varphi_\ell(y)$$

of the optimal solution involves the regularization parameter λ and the components of Mercer's decomposition of the kernel k (cf. Cucker and Zhou 2007, Proposition 8.6).

The next theorem provides a bound on the magnitude of the worst case norm, more specifically, a bound for $\sup_{x\in[0,1]^d} \|w_\lambda^x\|_{L^2}$.

Theorem 5 (Uniform bound of the weight function)**.** *Let k be the Gaussian kernel. The weight function w_λ^x satisfies the bound*

$$\sup_{x\in[0,1]^d} \|w_\lambda^x\|_{L^2}^2 \le 9c_p\left(3c_\sigma s(\lambda)+2\right)^{2d}+1 \tag{3.7}$$

with $s(\lambda) \coloneqq \max\left\{-\frac{1}{2}\ln\left(\frac{\lambda}{9}\right),\, d\cdot c,\, e\right\}$ and c_σ, c from Proposition 2.

Proof The function w_λ^x minimizes (3.2) and thus

$$\sup_{x\in[0,1]^d} \lambda\,\|w_\lambda^x\|_{L^2}^2 \le \sup_{x\in[0,1]^d} \lambda\,\|W_m^x\|_{L^2}^2 + \|L_k W_m^x - k_x\|_k^2 \le \sup_{x\in[0,1]^d} \lambda c_p m^{2d} + \|L_k W_m^x - k_x\|_k^2\,, \tag{3.8}$$

as $\|W_m^x\|_{L^2}^2 \le c_p m^{2d}$ (see Theorem 1).

Choosing $s(\lambda) \coloneqq \max\left\{-\frac{1}{2}\ln\left(\frac{\lambda}{9}\right),\, d\cdot c,\, e\right\}$ and $m = 3c_\sigma\, s(\lambda)+2$ we derive from (3.6) that

$$\|L_k W_m^x - k_x\|_k^2 \le 9s^{-2s} \le \lambda$$

(as $s\ge e$ and $s \ge -\frac{1}{2}\ln\left(\frac{\lambda}{9}\right)$), and thus

$$\sup_{x\in[0,1]^d} \lambda\,\|w_\lambda^x\|_{L^2}^2 \le \sup_{x\in[0,1]^d} \lambda c_p m^{2d} + \|L_k W_m^x - k_x\|_k^2 \le \lambda c_p(3c_\sigma s(\lambda)+2)^{2d} + \lambda$$

by (3.8). Multiplying with λ^{-1} gives the assertion. ■

The bound (3.7) characterizes the asymptotic growth of the norm $\|w_\lambda^x\|_{L^2}$. Indeed, letting $\lambda \downarrow 0$, the bound (3.7) implies that $\|w_\lambda^x\|_{L^2}$ grows at most as $\left(\ln(\lambda^{-1})\right)^d$. The subsequent sections heavily exploit this asymptotic behavior.

3.3 Eigenvalues and eigenfunctions

This section studies the elements in Mercer's decomposition of the kernel, $k(x,y) = \sum_{\ell=1}^\infty \mu_\ell \varphi_\ell(x)\varphi_\ell(y)$, specifically for the Gaussian kernel. We first relate the maximal value of any eigenfunction with its associated eigenvalue, demonstrating that $\|\varphi_\ell\|_\infty$ is bounded by $\left(\ln(\mu_\ell^{-1})\right)^d$. This bound is significantly sharper than the standard bound $\max_{x\in\mathcal{X}} \varphi_\ell(x) \le k(x,x)^{1/2}\mu_\ell^{-1/2}$ derived from the Cauchy–Schwartz inequality, $\langle \varphi,\, k_x\rangle_k \le \|\varphi\|_k\, k(x,x)$.

We next describe the decay of the eigenvalues μ_ℓ and derive a bound on $\|\varphi_\ell\|_\infty$, which turns out to be quadratic in ℓ. The results of this section enable us to infer convergence in uniform norm from convergence in L^2. Generally, this approach leads to faster convergence rates compared to results, which are derived from convergence in the norm $\|\cdot\|_k$ (see Section 4.2).

Our approach builds on the following observation. For the regularization $\lambda = \mu_\ell$, the series

$$\|w_\lambda^x\|_{L^2}^2 = \sum_{h=1}^\infty \left(\frac{\mu_h}{\lambda+\mu_h}\right)^2 \varphi_h^2(x)$$

involves the term $\frac{1}{4}\varphi_\ell^2(x)$. To bound the maximum of $|\varphi_\ell|$, it is sufficient to assess $\sup_{x\in[0,1]^d}\|w_\lambda^x\|_{L^2}^2$, which is bounded by the inequality (3.7). The following result summarizes these relations.

Theorem 6. *For the eigenfunctions of the Gaussian kernel, the inequality*

$$\sup_{x\in[0,1]^d}|\varphi_\ell(x)| \le 6\sqrt{c_p}\left(3c_\sigma s(\mu_\ell)+2\right)^d+2 \tag{3.9}$$

holds for every $\ell\in\mathbb{N}$, where $s(\mu_\ell)=\max\left\{-\frac{1}{2}\ln\left(\frac{\mu_\ell}{9}\right),\ e,\ d\cdot c\right\}$ is as in Theorem 5.

Proof For $\ell\in\mathbb{N}$, it holds that

$$\frac{1}{4}\left|\varphi_\ell(x)\right|^2=\left(\frac{\mu_\ell}{\mu_\ell+\mu_\ell}\right)^2\varphi_\ell(x)^2\le\sum_{h=1}^\infty\left(\frac{\mu_h}{\mu_\ell+\mu_h}\right)^2\varphi_h(x)^2=\|w_{\lambda^*}^x\|_{L^2}^2\,,$$

with $\lambda^*=\mu_\ell$. Taking the maximum on both sides, we deduce from the latter inequality and (3.7) that

$$\begin{aligned}\frac{1}{4}\|\varphi_\ell\|_\infty^2&\le\sup_{x\in[0,1]^d}\|w_{\lambda^*}^x\|_{L^2}^2\\&\le 9c_p\left(3c_\sigma s(\lambda^*)+2\right)^{2d}+1=9c_p\left(3c_\sigma s(\mu_\ell)+2\right)^{2d}+1.\end{aligned}$$

Reformulating this inequality as well as using the subadditivity of the square root gives the assertion. ■

The bound (3.9) relates the eigenfunctions and the eigenvalues of the operator L_k. In what follows, we analyze the decay of $(\mu_\ell)_{\ell=1}^\infty$ to get a more concrete characterization of the bound in (3.9).

The next lemma provides a lower bound of the eigenvalues $(\mu_\ell)_{\ell=1}^\infty$.

Lemma 1 (Maximal decay of eigenvalues)**.** *Set $p_{\min}\coloneqq\inf_{x\in\mathcal{X}}p(x)$ and $p_{\max}\coloneqq\sup_{x\in\mathcal{X}}p(x)$. For the eigenvalues of the Gaussian kernel it holds that*

$$\mu_\ell\ge\frac{p_{\min}^2}{p_{\max}^2}C(d,\sigma)e^{-c_{\sigma,d}(\ell+d)^{\frac{2}{d}}},\qquad \ell=1,2,\dots, \tag{3.10}$$

where the constants $c_{d,\sigma}$ and $C(d,\sigma)$ depend on the dimension d and the bandwidth σ.

Proof See Appendix B. ■

The subsequent theorem combines the inequalities (3.9) and (3.10) and provides an explicit bound on the maximal absolute value of the eigenfunctions $(\varphi_\ell)_{\ell=1}^\infty$.

Theorem 7 (Eigenfunctions)**.** *The eigenfunctions of the d-dimensional Gaussian kernel satisfy the bound*

$$\max_{x\in[0,1]^d}|\varphi_\ell(x)|\le 6\sqrt{c_p}\max\left\{h(\ell),\ \left(3c_\sigma e+2\right)^d,\left(3c_\sigma(d\cdot c)+2\right)^d\right\}+2 \tag{3.11}$$

with

$$h(\ell)\coloneqq\left(3c_\sigma\left(\left(\frac{1}{2}\ln(9)-\frac{1}{2}\ln\left(\frac{p_{\min}^2}{p_{\max}^2}C(d,\sigma)\right)+\frac{1}{2}c_{d,\sigma}(\ell+d)^{\frac{2}{d}}\right)\right)_++2\right)^d$$

and the constants from the preceding Lemma 1.

Proof For the assertion note first that

$$s(\lambda) = \max\left\{-\frac{1}{2}\ln\left(\frac{\lambda}{9}\right),\ e,\ d\cdot c\right\} = \max\left\{\left(-\frac{1}{2}\ln\left(\frac{\lambda}{9}\right)\right)_+,\ e,\ d\cdot c\right\},$$

where $x_+ \coloneqq \max\{x,\ 0\}$ is the positive part of x. Hence, employing (3.10) in (3.9), we have the bound

$$\begin{aligned}
&|\varphi_\ell(x)| \qquad (3.12)\\
&\le 6\sqrt{c_p}\left(3c_\sigma s(\mu_\ell)+2\right)^d + 2\\
&= 6\sqrt{c_p}\left(3c_\sigma \max\left\{-\frac{1}{2}\ln\left(\frac{p_{\min}^2 C(d,\sigma)e^{-c_{\sigma,d}(\ell+d)^{\frac{2}{d}}}}{9p_{\max}^2}\right),\ e,\ d\cdot c\right\}+2\right)^d + 2\\
&= 6\sqrt{c_p}\left(3c_\sigma \max\left\{\left(\frac{1}{2}\ln(9) - \frac{1}{2}\ln\left(\frac{p_{\min}^2}{p_{\max}^2}C(d,\sigma)\right) + \frac{1}{2}c_{d,\sigma}(\ell+d)^{\frac{2}{d}}\right)_+,\ e,\ d\cdot c\right\}+2\right)^d \qquad (3.13)\\
&\qquad + 2.
\end{aligned}$$

Note now that $x \mapsto x^d$ is an increasing function on $[0,\infty)$ and thus $\max\{a,b\}^d = \max\{a^d, b^d\}$, provided that $a, b \ge 0$. Using this property in (3.13) provides the assertion (3.11). ■

Remark 3. The right-hand side of (3.11) deserves additional attention. While $h(\ell)$ grows with ℓ increasing, the other terms do not depend on ℓ. For sufficiently large ℓ, the inequality (3.11) thus reads

$$\max_{x\in[0,1]^d} |\varphi_\ell(x)| \le 6c_p^{1/2}h(\ell) + 2.$$

The function h itself grows at most as a polynomial with degree 2. That is, there exists a constant $b > 0$ with quadratic bound

$$\max_{x\in[0,1]^d} |\varphi_\ell(x)| \le b\,\ell^2, \qquad \ell = 1, 2, \dots. \tag{3.14}$$

To the best of our knowledge, this is the first non-exponential bound of $\varphi_\ell's$ absolute maximum. The following section addresses various consequences of this main result.

Remark 4. The approach chosen in this Section 3 is not limited to the Gaussian kernel. All steps extend to general kernels of the form $\phi(\|x-y\|^2)$ (cf. (2.8)), provided that a lower bound on the decay of its eigenvalues is available.

4. Main results

This section connects the results of the preceding sections for general statistical learning settings. We present improved concentration bound inequalities, an interpolation inequality relating to uniform convergence, and the Nyström method.

We adopt the well-established notation (cf. Rudi et al. 2015) and introduce

$$\mathcal{N}_\infty(\lambda) \coloneqq \sup_{x\in[0,1]^d} \|w_\lambda^x\|_{L^2}^2 + \lambda^{-1} \|L_k w_\lambda^x - k_x\|_k^2 . \tag{4.1}$$

As established above in (3.8), it holds that

$$\mathcal{N}_\infty(\lambda) = \mathcal{O}\big(\ln(\lambda^{-1})^{2d}\big)$$

for $\lambda \to 0$.

Building on the results of Section 3, we provide the results for the Gaussian kernel.

4.1 Concentration bounds

Standard kernel ridge regression minimizes the regularized squared error at the sample points. To infer the related error at other locations, one needs to connect the discrete setting with the continuous setting from Section 3. This is commonly done by relating the discrete operator

$$L_k^D \colon C(\mathcal{X}) \to \mathcal{H}_k \quad \text{with} \quad (L_k^D f)(y) \coloneqq \frac{1}{n}\sum_{i=1}^n f(x_i)k(y,x_i)$$

with its continuous version L_k, as demonstrated in Caponnetto and De Vito (2006) as well as in Fischer and Steinwart (2020). However, the underlying concentration results in these references generally require regularization sequences not faster than $\lambda_n = \mathcal{O}(1/n)$, which is too restrictive for many tasks including the analysis of low rank kernel methods (cf. Zhang et al. 2013, Bach 2013).

This section addresses this issue. We provide the following result, which allows significantly higher flexibility in choosing the regularization sequence λ_n. Specifically, the regularizing sequence may be chosen to decay faster than any polynomial.

The following proposition collects the detailed result for the concentration inequality.

Proposition 3. Assume that λ satisfies

$$\frac{4}{3}\tau\, g(\lambda)\frac{\mathcal{N}_\infty(\lambda)}{n} + \sqrt{2\tau\, g(\lambda)\frac{\mathcal{N}_\infty(\lambda)}{n}} \le \frac{1}{2} \tag{4.2}$$

with

$$g(\lambda) = \ln\left(2e\frac{(\lambda+\mu_1)\mathcal{N}(\lambda)}{\mu_1}\right), \quad \mathcal{N}(\lambda) = \sum_{\ell=1}^\infty \frac{\mu_\ell}{\lambda+\mu_\ell}$$

and $\tau > 0$. Then the inequality

$$\left\|(L_k+\lambda)^{1/2}(\lambda+L_k^D)^{-1}(L_k+\lambda)^{1/2}\right\|_{\mathcal{H}_k\to\mathcal{H}_k} \le 2 \tag{4.3}$$

holds with probability at least $1-2e^{-\tau}$.

Proof For the proof we refer to Appendix C. ■

The condition (4.2) deserves some additional commentary. The term $\mathcal{N}_\infty(\lambda)$ is asymptotically bounded by $\mathcal{O}\big(\ln(\lambda^{-1})^{2d}\big)$, and the function $g(\lambda)$ does not grow faster than $\mathcal{O}\big(\ln(\lambda^{-1})\big)$. The condition (4.2) is asymptotically satisfied, provided that the regularization parameter λ does not decay faster than $c\exp\bigl(-n^{\frac{1}{2d+1}}\bigr)$. This is a significant improvement compared to the common regularization choice n^{-1}.

The conclusion of Proposition 3 can be given more general, not necessarily involving the Gaussian kernel. Moreover, the bound (3.7) of $\mathcal{N}_\infty(\lambda)$ only relies on the decay of the Taylor coefficients associated with the kernel. The same method thus may be applied to derive concentration inequalities for any other radial kernel functions.

4.2 Interpolation inequality

This section establishes results on the uniform approximation quality for smooth functions. The following interpolation inequality ensures that the uniform norm of smooth functions is comparable its norm in L^2, although this norm is much weaker in general. We measure smoothness with reference to the norm $\|\cdot\|_s$ introduced below.

Theorem 8 (Interpolation of norm)**.** *For $f=\sum_{\ell=1}^\infty c_\ell\varphi_\ell\in L^2$ with $\|f\|_s^2 \coloneqq \sum_{\ell=1}^\infty \frac{c_\ell^2}{\ell^{-2s}}<\infty$ it holds that*

$$\|f\|_\infty \le \frac{\pi}{\sqrt{6}} b\,\|f\|_s^{\frac{3}{s}}\cdot\|f\|_2^{1-\frac{3}{s}}\,, \tag{4.4}$$

where $s>3$ and $\max_{x\in\mathcal{X}}|\varphi_\ell(x)|\le b\,\ell^2$ for all $\ell=1,2,\dots$ (cf. (3.14)).

Proof Building on the bound $\max_{x\in\mathcal{X}}|\varphi_\ell(x)|\le b\ell^2$ we have with the Cauchy–Schwarz inequality that

$$\begin{aligned} f(x)=\sum_{\ell=1}^\infty c_\ell\varphi_\ell(x) &\le b\sum_{\ell=1}^\infty \frac{c_\ell}{\ell^{-\frac{s}{p}}}\cdot \ell^{2-\frac{s}{p}} \\ &\le b\sqrt{\sum_{\ell=1}^\infty \frac{c_\ell^2}{\ell^{-\frac{2s}{p}}}}\sqrt{\sum_{\ell=1}^\infty \ell^{(2-\frac{s}{p})2}} = b\sqrt{\sum_{\ell=1}^\infty \frac{c_\ell^{\frac{2}{p}}}{\ell^{-\frac{2s}{p}}}c_\ell^{2-\frac{2}{p}}}\sqrt{\zeta\Bigl(\frac{2s}{p}-4\Bigr)}, \end{aligned}$$

where $\zeta(\cdot)$ is the Riemann zeta function and $p\in[1,2s]$. Employing Hölder's inequality, it follows for the first term that

$$\begin{aligned} \sqrt{\sum_{\ell=1}^\infty \frac{c_\ell^{\frac{2}{p}}}{\ell^{-\frac{2s}{p}}}c^{2-\frac{2}{p}}} &\le \left(\sum_{\ell=1}^\infty \frac{c_\ell^2}{\ell^{-2s}}\right)^{\frac{1}{2p}}\left(\sum_{\ell=1}^\infty c_\ell^{(2-\frac{2}{p})(\frac{p}{p-1})}\right)^{\frac{1}{2}(1-\frac{1}{p})} \\ &= \left(\sum_{\ell=1}^\infty \frac{c_\ell^2}{\ell^{-2s}}\right)^{\frac{1}{2p}}\left(\sum_{\ell=1}^\infty c_\ell^2\right)^{\frac{1}{2}(1-\frac{1}{p})}. \end{aligned}$$

Choosing $p=\frac{s}{3}$, we finally have that

$$\begin{aligned} b\left(\sum_{\ell=1}^\infty \frac{c_\ell^2}{\ell^{-2s}}\right)^{\frac{1}{2p}}\left(\sum_{\ell=1}^\infty c_\ell^2\right)^{\frac{1}{2}(1-\frac{1}{p})}\sqrt{\zeta\Bigl(\frac{2s}{p}-4\Bigr)} &= b\,\|f\|_s^{\frac{1}{p}}\,\|f\|_2^{1-\frac{1}{p}}\sqrt{\zeta\Bigl(\frac{2s}{p}-4\Bigr)} \\ &= \frac{\pi}{\sqrt{6}}b\,\|f\|_s^{\frac{3}{s}}\,\|f\|_2^{\frac{s-3}{s}} \end{aligned}$$

by involving Euler's famous formula $\zeta(2) = \frac{\pi^2}{6}$. This is the assertion. ■

Convergence in L^2 implies convergence in L^∞: Specifically, consider a sequence of functions f_n with slowly increasing norm $\|f_n\|_s$ and limit $\|f\|_s < \infty$. Then, for $\|f_n - f\|_{L^2} \to 0$ sufficiently fast, it follows from Theorem 8 that $\|f_n - f\|_\infty \to 0$.

Remark 5. Fischer and Steinwart (2020) study convergence in an interpolation space between $\mathcal{H}_k$ and L^2, which the authors call *power spaces*. The norm considered in Fischer and Steinwart (2020), however, is stronger than the norm $\|\cdot\|_s$ considered above.

4.3 Nyström method

The main result in Rudi et al. (2015, Theorem 1) relates the function $\mathcal{N}_\infty(\lambda)$ to the Nyström method. Their paper ensures that the Nyström method does not require more than $\mathcal{O}\left(\mathcal{N}_\infty(\lambda)\ln\lambda^{-1}\right)$ supporting points.

The results established in Section 3 above provides explicit access to the function $\mathcal{N}_\infty(\lambda)$, so that their main result can be refined and enhanced to the following form.

Theorem 9 (Cf. Rudi et al. 2015, Theorem 1). *Let $\mathcal{E}(f) := \iint_{\mathcal{X}\times\mathbb{R}}\big(f(x)-y\big)^2\rho(dx,dy)$ be the common error function. Given the assumptions of Theorem 1 in (Rudi et al., 2015), it holds that the Nyström-approximation $\hat{f}_{\lambda,m}$ with regression parameter λ satisfies*

$$\mathcal{E}(\hat{f}_{\lambda,m}) - \mathcal{E}(f_{\mathcal{H}}) \le q^2 n^{\frac{2\nu+1}{2nu+\gamma+1}}$$

for at least

$$m \ge (67 \vee 5\left(9c_p\,(3c_\sigma s(\lambda)+2)^{2d}+1\right)\ln\lambda^{-1} \tag{4.5}$$

supporting points. The reference provides the constant q and ν explicitly.

Proof Invoking the bound (3.7) for $\mathcal{N}_\infty(\lambda)$, the result is immediate from Rudi et al. (2015, Theorem 1). ■

The adapter result (4.5) gives an explicit selection criterion for the critical number of support points in the Nyström method.

5. Summary

The novel approach of this paper considers an explicit approximation of the kernel function in the range of the associated integral operator. To this end we provide an explicit weight function by matching the initial Taylor coefficients of the kernel.

The approach has numerous consequences in theory and in applications. We provide bounds for the eigenfunctions, which grow only quadratic in the enumeration index. An interpolation inequality provided relates convergence in uniform norm and the weaker convergence in L^2 for smooth functions. The methods established justify smaller regression parameters for regression problems, which is of particular importance for low-rank approximation techniques.

References

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 185–209, 2013. doi:10.48550/arXiv.1208.2015.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004. doi:10.1007/978-1-4419-9096-9.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006. doi:10.1007/s10208-006-0196-8.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007. doi:10.1017/CBO9780511618796.

Persi Diaconis, Sharad Goel, and Susan Holmes. Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, 2(3):777 – 807, 2008. doi:10.1214/08-AOAS165.

Benedikt Diederichs and Armin Iske. Improved estimates for condition numbers of radial basis function interpolation matrices. *Journal of Approximation Theory*, 238:38–51, 2019. doi:10.1016/j.jat.2017.10.004.

Paul Dommel and Alois Pichler. Dynamic programming for data independent decision sets. *Journal of Convex Analysis*, 30(3):897–916, 2023. doi:10.48550/arXiv.2102.07464.

Petros Drineas and Michael Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6: 2153–2175, 12 2005.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968. doi:10.1017/S0020268100037586.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(1), 2020. doi:10.48550/arXiv.1702.07254.

Mehrdad Honarkhah and Jef Caers. Stochastic simulation of patterns using distance-based pattern modeling. *Mathematical Geosciences*, 42(5):487–517, 2010. doi:10.1007/s11004-010-9276-7.

Hyuk Park, Zhuangzhuang Jia, and Grani A. Hanasusanto. Data-driven stochastic dual dynamic programming: Performance guarantees and regularization schemes. 2022. URL `https://optimization-online.org/?p=21376`.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. doi:10.48550/arXiv.1507.04717.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, 2017. URL `https://api.semanticscholar.org/CorpusID:25900554`.

Bernhard Schölkopf. *Support Vector Learning*. 1997. URL `https://pure.mpg.de/rest/items/item_1794215/component/file_3214422/content`.

Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel Methods in Computational Biology*. The MIT Press, 2004. doi:10.7551/mitpress/4057.001.0001.

John Shawe-Taylor, Chris Williams, Nello Cristianini, and Jaz Kandola. *On the Eigenspectrum of the Gram Matrix and Its Relationship to the Operator Eigenspectrum*, pages 12–12. Springer Berlin Heidelberg, 2002. doi:10.1007/3-540-36182-0_3.

Tao Shi, Mikhail Belkin, and Bin Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, 37(6B):3960 – 3984, 2009. doi:10.1214/09-AOS700.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008. doi:10.1007/978-0-387-77242-4.

Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

Nicholas Sterge and Bharath K. Sriperumbudur. Statistical optimality and computational efficiency of Nyström kernel pca. *Journal of Machine Learning Research*, 23:337:1–337:32, 2021. doi:10.48550/arXiv.2105.08875.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL `https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf`.

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007. doi:10.1007/s11263-006-9794-4.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 592–617, Princeton, NJ, USA, 2013. URL `https://proceedings.mlr.press/v30/Zhang13.html`.

Appendix A. Gaussian approximation

This section provides the proof of the formula (3.3), which is of fundamental importance for every other bound provided in Section 3. The proof of the bound (3.3) builds on the error estimate (2.11), which involves the Taylor remainder of the exponential series. To this end we utilize the formula

$$\sum_{\ell=n}^{\infty} \alpha^{\ell} = \frac{\alpha^n}{1-\alpha} \tag{A.1}$$

of the truncated geometric series as well as Stirling's approximation

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n!, \tag{A.2}$$

which relates the factorial with the exponential function. For future reference convenience of the reader, we restate Theorem 4.

Theorem 10. *Let k be the d-dimensional Gaussian kernel with width parameter σ. Setting $c_\sigma := \max\{1, 2e\sigma d\}$, $c_p = \sup_{z\in[0,1]^d} p(z)^{-1}$ and*

$$C(\sigma, m) := \frac{1}{1-\frac{\sigma e d}{\lfloor \frac{m}{2}\rfloor}},$$

the uniform bound

$$\sup_{x\in[0,1]^d} \|L_k W_m^x - k_x\|_\infty \le (1 + c_p^{1/2} m^d)\, C(\sigma, m) \left(\left\lfloor \frac{m}{2}\right\rfloor \frac{1}{\sigma e d}\right)^{\lfloor \frac{m}{2}\rfloor}$$

holds for W_m^x defined in (2.6) for $m > c_\sigma + 1$.

Specifically, for $m(s) := 3c_\sigma s + 2$, we have that

$$\sup_{x\in[0,1]^d} \left\| L_k W_{m(td)}^x - k_x \right\|_\infty \le 3(t\,d)^{-3t\,d} \tag{A.3}$$

whenever $t \ge \max\left\{\frac{\ln(3)+(d-1)\ln(2)+\frac{1}{2}\ln(c_p)+d\ln(3c_\sigma d)}{2d\ln(3)},\ 1\right\}$.

Proof Employing the inequality (2.11) involving the Taylor coefficients, we have that

$$\sup_{x\in[0,1]^d} \|L_k W_m^x - k_x\|_\infty \le (1 + c_p^{1/2} m^d) \sum_{\ell=\lfloor \frac{m-1}{2}\rfloor+1}^{\infty} \frac{\sigma^\ell}{\ell!} d^\ell \le (1 + c_p^{1/2} m^d) \sum_{\ell=\lfloor \frac{m}{2}\rfloor}^{\infty} \frac{\sigma^\ell}{\ell!} d^\ell, \tag{A.4}$$

as $\lfloor \frac{m}{2}\rfloor \le \lfloor \frac{m-1}{2}\rfloor + 1$. Further, invoking Stirling's approximation (A.2) for the right-hand side of (A.4), it follows that

$$(1 + c_p^{1/2} m^d) \sum_{\ell=\lfloor \frac{m}{2}\rfloor}^{\infty} \sigma^\ell\,(\ell!)^{-1}\, d^\ell \le (1 + c_p^{1/2} m^d) \sum_{\ell=\lfloor \frac{m}{2}\rfloor}^{\infty} \left(\sqrt{2\pi\ell}\left(\frac{\ell}{\sigma e d}\right)^\ell\right)^{-1}.$$

Now assume that m satisfies the inequality $1 < \left\lfloor \frac{m}{2} \right\rfloor (\sigma e d)^{-1}$, then we have from identity (A.1) of the truncated geometric series that

$$\begin{aligned}
\sup_{x\in[0,1]^d} \|L_k W_m^x - k_x\|_\infty &\leq (1 + c_p^{1/2} m^d) \sum_{\ell=\lfloor \frac{m}{2} \rfloor}^{\infty} \left(\frac{\ell}{\sigma e d}\right)^{-\ell} \\
&\leq (1 + c_p^{1/2} m^d) \frac{1}{1 - \frac{\sigma e d}{\lfloor \frac{m}{2} \rfloor}} \left(\left\lfloor \frac{m}{2} \right\rfloor \frac{1}{\sigma e d}\right)^{-\lfloor \frac{m}{2} \rfloor} && \text{(A.5)} \\
&= (1 + c_p^{1/2} m^d) C(\sigma, m) \left(\left\lfloor \frac{m}{2} \right\rfloor \frac{1}{\sigma e d}\right)^{-\lfloor \frac{m}{2} \rfloor}, && \text{(A.6)}
\end{aligned}$$

which is the first assertion.

For the second assertion let $m = m(td) = 6c_\sigma td + 2$ and observe for the constant $C(\sigma, m)$ that

$$C(\sigma, m) = \frac{1}{1 - \frac{\sigma e d}{\lfloor \frac{m}{2} \rfloor}} \leq \frac{1}{1 - \frac{\sigma e d}{3(\sigma e d) t d}} \leq \frac{1}{1 - \frac{1}{3td}} \leq \frac{3}{2}$$

holds for every $t \geq 1$. Furthermore, employing $m(td)$, as well as $C(\sigma, m) \leq \frac{3}{2}$, and arguing as above, we get for (A.6) that

$$\begin{aligned}
&(1 + c_p^{1/2} m^d) C(\sigma, m) \left(\left\lfloor \frac{m}{2} \right\rfloor \frac{1}{\sigma e d}\right)^{-\lfloor \frac{m}{2} \rfloor} \\
&\quad \leq \frac{3}{2}(1 + c_p^{1/2} m(t)^d) \left(\frac{3\sigma e d^2 t}{\sigma e d}\right)^{-\lfloor \frac{m}{2} \rfloor} \\
&\quad \leq \frac{3}{2}(1 + c_p^{1/2} (3c_\sigma td + 2)^d)\, (3td)^{-3dt} \\
&\quad \leq \frac{3}{2}\, (3td)^{-3dt} + \frac{3}{2} 2^{d-1} c_p^{1/2} (3c_\sigma td)^d\, (3td)^{-3dt} + \frac{3}{2} 2^{2d-1}\, (3td)^{-3dt},
\end{aligned}$$

where we utilize $(a + b)^d \leq 2^{d-1}(a^d + b^d)$ for the last inequality. The latter is bounded by

$$\frac{3}{2} 2^{2d-1}\, (3td)^{-3dt} \leq (td)^{-3dt}$$

as well as the middle term by

$$\begin{aligned}
&2^{d-1} c_p^{1/2} (3c_\sigma td)^d 3^{-3td} (td)^{-3td} \\
&\qquad = \frac{1}{2}(e^{\ln(3) + (d-1)\ln(2) + \ln(c_p^{1/2}) + d\ln(3c_\sigma td) - \ln(3)3td})(td)^{-3td} \\
&\qquad \leq \frac{1}{2}(e^{\ln(3) + (d-1)\ln(2) + \ln(c_p^{1/2}) + d\ln(3c_\sigma d) - \ln(3)2td})(dt)^{-3td} \\
&\qquad \leq \frac{1}{2}(td)^{-3td}
\end{aligned}$$

whenever $t \geq \frac{\ln(3)+(d-1)\ln(2)+\ln(c_p^{1/2})+d\ln(3\max\{c_\sigma,1\}d)}{2d\ln(3)}$. Hence, choosing

$$t \geq \max\left\{\frac{\ln(3)+(d-1)\ln(2)+\ln(c_p^{1/2})+d\ln(3\max\{c_\sigma,1\}\,d)}{2d\ln(3)}, 1\right\},$$

we have that

$$(1+c_p^{1/2}m^d)C(\sigma,m)\left(\left\lfloor\frac{m}{2}\right\rfloor\frac{1}{\sigma e d^2}\right)^{\left\lfloor\frac{m}{2}\right\rfloor} \leq \frac{3}{2}\,(3dt)^{-3dt}+\frac{3}{2}(dt)^{-3dt} \leq 3(dt)^{-3dt},$$

which is the assertion. ■

Building on the bound (3.3) in the uniform norm, we establish a bound in the RKHS norm in Proposition 2. To this end the following technical Lemma is of crucial importance.

Lemma 2. *Given the assumptions of Theorem 4, the bound*

$$\sup_{x\in[0,1]^d}\langle L_k W_m^x - k_x, W_m^x\rangle_{L^2} \leq 6(td)^{-2td} \tag{A.7}$$

holds whenever $t \geq \max\{c_0, c_1, c_2\}$ *with constants*

$$c_1 = \frac{(2d-1)\ln(2)+d\ln(3c_\sigma)+\frac{1}{2}\ln(c_p)}{d}+1 \quad \textit{and}$$
$$c_2 = \frac{(2d-1)\ln(2)+\frac{1}{2}\ln(c_p)}{d}.$$

Proof Employing the Cauchy–Schwarz inequality, we have from (3.3) that

$$\langle L_k W_m^x - k_x, W_m^x\rangle_2 \leq \|L_k W_m^x - k_x\|_{L^2}\,\|W_m^x\|_{L^2} \leq 3(td)^{-3td}c_p^{1/2}m^d$$

whenever m (and t) are chosen with respect to the constraints Theorem 4. Involving $m = m(td)$ we further observe that

$$3(td)^{-3td}c_p^{1/2}(3c_\sigma td+2)^d \leq 2^{d-1}3(td)^{-3td}c_p^{1/2}(3c_\sigma td)^d + 2^{2d-1}3(td)^{-3td}c_p^{1/2}. \tag{A.8}$$

For the second term in (A.8) it follows that

$$2^{2d-1}3(td)^{-3td}c_p^{1/2} = 2^{2d-1}(td)^{-td}c_p^{1/2}3(td)^{-2td} \leq 3(td)^{-2td}$$

whenever $t \geq \frac{(2d-1)\ln(2)+\frac{1}{2}\ln(c_p)}{d} := c_2$. Reformulating the first term in (A.8) to

$$2^{d-1}3(td)^{-3td}c_p^{1/2}(3c_\sigma td)^d = 3e^{-td\ln(td)+(2d-1)\ln(2)+\frac{1}{2}\ln(c_p)+d\ln(3c_\sigma td)}(td)^{-2td}$$

we get for the exponent that

$$\begin{aligned}
&-td\ln(td)+(d-1)\ln(2)+\frac{1}{2}\ln(c_p)+d\ln(3c_\sigma td)\\
=&-d(t-1)\ln(td)+(d-1)\ln(2)+\frac{1}{2}\ln(c_p)+d\ln(3c_\sigma)\\
\leq&-d(t-1)\ln(2)+(d-1)\ln(2)+\frac{1}{2}\ln(c_p)+d\ln(3c_\sigma).
\end{aligned}$$

Thus, for

$$t \geq \frac{(2d-1)\ln(2) + d\ln(3c_\sigma) + \frac{1}{2}\ln(c_p)}{d} + 1 =: c_1$$

it follows that

$$2^{d-1} 3(td)^{-3td} c_p^{1/2} (3c_\sigma td)^d \leq 3(td)^{-2td}.$$

Combining the estimates of the terms in (A.8), we have for $t \geq \max\{c_1, c_2\}$ that

$$\langle L_k W_m^x - k_x, W_m^x \rangle_2 \leq 6(td)^{-2td}$$

and thus the assertion. ■

Appendix B. Decay of eigenvalues

This section provides a lower bound on the eigenvalues $(\mu_\ell)_{\ell=1}^\infty$ of the operator L_k associated with the Gaussian kernel. We address the univariate case first, which is then extended to the multivariate case. As a starting point we consider the most elementary setting, i.e., $\mathcal{X} = [0,1]$, equipped with the uniform measure $P = \mathcal{U}[0,1]$. The following lemma provides the precise eigenvalue bound.

Lemma 3 (Maximal decay of eigenvalues)**.** *Let k be the Gaussian kernel, $\mathcal{X} = [0,1]$ as well as $P = \mathcal{U}[0,1]$. For every $\ell \in \mathbb{N}$ it holds that*

$$\mu_\ell \geq \frac{1}{\ell} C(\sigma) e^{-a_\sigma(\ell-1)^2}, \tag{B.1}$$

where $a_\sigma = 8 \cdot \frac{4\pi^2}{16\sigma}$ and $C(\sigma)$ is a constant depending on the width parameter σ.

Proof Let $x_1, \dots, x_\ell$ be independent random variables following the uniform measure $\mathcal{U}[0,1]$. Let $K = (k(x_i, x_j))_{i,j=1}^\ell$ be the Gramian matrix and invoking Shawe-Taylor et al. (2002, Proposition A), it follows that

$$\mu_\ell \geq \frac{1}{\ell}\, \mathbb{E}\, \lambda_{\min}(K),$$

where $\lambda_{\min}(K)$ is the smallest eigenvalue of the matrix K and the expectation is with respect to the samples. Further, employing the result of Diederichs and Iske (2019, Example 2.6), we get with $M \coloneqq \min_{i,j\leq \ell, i\neq j} |x_i - x_j|$ and (D.4) in the auxiliary Lemma 9 below (Appendix D) that

$$\mathbb{E}\, \lambda_{\min}(K) \geq \mathbb{E}\, M^{-1} \tilde{C}(\sigma) e^{-\frac{4\pi^2}{16M^2\sigma}} \geq \tilde{C}(\sigma)\, \mathbb{E}\, e^{-\frac{4\pi^2}{16M^2\sigma}} \geq C(\sigma) e^{-a_\sigma(\ell-1)^2},$$

as $M < 1$, and where $C(\sigma) = C \cdot \tilde{C}(\sigma)$ with C from as in (D.4). This is the assertion. ■

Extending the univariate case to the multivariate setting builds on the product structure of the Gaussian kernel, that is, on $\prod_{i=1}^d e^{-\sigma(x_i-y_i)^2} = e^{-\sigma\sum_{i=1}^n (x_i-y_i)^2}$. Indeed, provided that the underlying measure is $\mathcal{U}[0,1]^d$, the spectrum of the corresponding operator L_k is

$$\left\{ \prod_{i=1}^d \mu_{\ell_i} \colon \ell_1, \dots, \ell_d \in \mathbb{N} \right\},$$

where μ_{ℓ_i} is the ℓ_ith eigenvalue in the univariate setting. This is immediate, as every eigenfunction in the multivariate case is a product of elementary eigenfunctions, $\prod_{i=1}^d \varphi_{\ell_i}(x_i)$. We may assume the multivariate eigenvalues $\mu_\ell^{(d)}$ arranged in non-increasing order such that

$$\left(\mu_\ell^{(d)}\right)_{\ell=1}^\infty = \left(\prod_{i=1}^d \mu_{\ell_i} : \ell_i \in \mathbb{N}\right) \tag{B.2}$$

and $\mu_1^{(d)} \geq \mu_2^{(d)} \geq \dots$.

The subsequent auxiliary and combinatorial lemmata utilize the structure of the spectrum to infer the maximal decay rate of the sequence $\mu_\ell^{(d)}$. The first is a general combinatorial result, with which we assess the eigenvalue decay in the second.

Lemma 4. *It holds that*

$$\sum_{\substack{i_1+\dots+i_d \leq n,\\ i_j \in \mathbb{N}}} 1 \geq \frac{(n-1)^d}{(d-1)^{d-1}} - d \tag{B.3}$$

for $n \geq d \geq 2$.

Proof We proof the assertion by employing on the *stars and bars formula*

$$\sum_{i_1+\dots+i_d=i} 1 = \binom{i-1}{d-1},$$

where $i_j \geq 1$ are positive integers (cf. Feller 1968, p. 38) and $i \in \mathbb{N}$. We have that

$$\sum_{i_1+\dots+i_d \leq n} 1 = \sum_{i=d}^n \sum_{i_1+\dots+i_d=i} 1 = \sum_{i=d}^n \binom{i-1}{d-1} \geq \sum_{i=d}^n \left(\frac{i-1}{d-1}\right)^{d-1} \geq \sum_{i=1}^n \left(\frac{i-1}{d-1}\right)^{d-1} - d, \tag{B.4}$$

where we utilize that

$$\binom{i-1}{d-1} = \frac{(i-1)(i-2)\cdots(i-d+1)}{(d-1)(d-2)\cdots 1} = \prod_{j=1}^{d-1} \frac{i-j}{d-j} \geq \prod_{j=1}^{d-1} \frac{i-1}{d-1} = \left(\frac{i-1}{d-1}\right)^{d-1}.$$

Furthermore, employing

$$\frac{n^d}{d} = \int_0^n x^{d-1} dx \leq \sum_{i=1}^n k^{d-1}$$

in (B.4), it follows that

$$\begin{aligned}\sum_{i=1}^n \left(\frac{i-1}{d-1}\right)^{d-1} - d &= \frac{1}{(d-1)^{d-1}} \sum_{i=1}^n (i-1)^{d-1} - d \\ &= \frac{1}{(d-1)^{d-1}} \sum_{k=0}^{n-1} k^{d-1} - d \geq \frac{(n-1)^d}{d(d-1)^{d-1}} - d\end{aligned}$$

and thus the assertion. ■

Lemma 5. *Let $\mu_\ell \geq e^{-\rho\ell^2}$ for every $\ell \in \mathbb{N}$. Then the sequence $\mu_\ell^{(d)}$ from* (B.2) *satisfies*

$$\mu_\ell^{(d)} \geq C \exp(-c\rho(\ell+d)^{\frac{2}{d}}) \tag{B.5}$$

for all $\ell \in \mathbb{N}$ and $d \geq 2$ for some $c > 0$ and $C > 0$.

Proof We first determine the number of combinations of $\mu_{i_1}, \ldots, \mu_{i_d}$ for which the product $\prod_{k=1}^d \mu_{i_k}$ is larger than $\exp(-\rho(\lambda+2)^2)$ for some threshold parameter $\lambda > 0$. It holds that $\prod_{i=1}^d \mu_{\ell_i} \geq e^{-\rho\, i_1^2 - \ldots - \rho\, i_d^2} \geq e^{-\rho(i_1+\cdots+i_d)^2} \geq e^{-\rho(\lambda+2)^2}$, provided that $i_1 + \cdots + i_d \leq \lambda + 2$. From (B.3) we deduce that

$$\sum_{i_1+\cdots+i_d \leq \lambda+2} 1 \geq \frac{\lfloor \lambda+1 \rfloor^d}{d(d-1)^{(d-1)}} - d$$

and thus

$$\mu_\ell^{(d)} \geq e^{-\rho(\lambda+2)^2}$$

for all $\ell \leq \frac{\lfloor \lambda+1 \rfloor^d}{d(d-1)^{(d-1)}} - d$. Choosing $\lambda \coloneqq \big(d(d-1)^{d-1}(\ell+d)\big)^{1/d}$ and employing $(a+b)^2 \leq 2a^2 + 2b^2$, we get that

$$\mu_\ell^{(d)} \geq e^{-\rho\left(\left(d(d-1)^{d-1}(\ell+d)\right)^{1/d}+2\right)^2} \geq e^{-\rho 2^3} e^{-\rho 2 d^{2/d}(d-1)^{\frac{2(d-1)}{d}}(\ell+d)^{2/d}}.$$

The assertion follows by setting $C \coloneqq \exp(-8\rho)$ and $c \coloneqq 2d^{2/d}(d-1)^{\frac{2(d-1)}{d}}$.

The bound (B.5) already implies (3.10), provided that the underlying measure is uniform, i.e, the Lebesgue measure. The following lemma extends the assertion for more general probability measures. ■

Lemma 6. *Let $\mathcal{X} = [0,1]^d$ and consider the operators $L_k \colon L^2(\mathcal{X},\lambda) \to L^2(\mathcal{X},\lambda)$ and $L_k^p \colon L^2(\mathcal{X},P) \to L^2(\mathcal{X},P)$ with*

$$(L_k f)(y) = \int_{\mathcal{X}} k(x,y) f(x) dx \quad (L_k^P f)(y) = \int_{\mathcal{X}} k(x,y) f(x) p(x) dx.$$

Here, λ is the Lebesgue (uniform) measure and P is a probability measure, satisfying the condition $0 < p_{\min} \coloneqq \inf_{x\in\mathcal{X}} p(x) \leq p_{\max} \coloneqq \sup_{x\in\mathcal{X}} p(x) < \infty$. The eigenvalues $(\mu_\ell^{(d)})_\ell$ of L_k ($(\mu_{\ell,p}^{(d)})_\ell$ of L_k^p, resp.), satisfy the relation

$$\mu_{\ell,p}^{(d)} \geq \frac{p_{\min}^2}{p_{\max}^2} \mu_\ell^{(d)} \tag{B.6}$$

for all $\ell \in \mathbb{N}$.

Proof By the Courant–Fischer–Weyl min-max principle it holds that

$$\begin{aligned}
\mu_{\ell,p}^{(d)} &= \max_{\dim(S_k)=k} \min_{\substack{x\in S_k \\ \|x\|_{L^2(P)}=1}} \left\langle L_k^p x, x\right\rangle_{L^2(P)} \\
&= \max_{\dim(S_k)=k} \min_{x\in S_k} \left(\frac{\|x\|_{L^2(\lambda)}}{\|x\|_{L^2(P)}}\right)^2 \left\langle L_k^p \frac{x}{\|x\|_{L^2(\lambda)}}, \frac{x}{\|x\|_{L^2(\lambda)}}\right\rangle_{L^2(P)} \\
&\geq \max_{\dim(S_k)=k} \min_{\substack{x\in S_k \\ \|x\|_{L^2(\lambda)}=1}} p_{\max}^{-2} \left\langle L_k^p x, x\right\rangle_{L^2(P)}.
\end{aligned}$$

It follows further that

$$\begin{aligned}
\mu_{\ell,p}^{(d)} &\geq \max_{\dim(S_k)=k} \min_{\substack{x\in S_k \\ \|x\|_{L^2(\lambda)}=1}} p_{\max}^{-2} \left\langle L_k^p x, x\right\rangle_{L^2(P)} \\
&\geq \max_{\dim(S_k)=k} \min_{\substack{x\in S_k \\ \|x\|_{L^2(\lambda)}=1}} p_{\max}^{-2} p_{\min}^2 \left\langle L_k x, x\right\rangle_{L^2(\lambda)} = \frac{p_{\min}^2}{p_{\max}^2}\mu_\ell^{(d)},
\end{aligned}$$

as $0 \leq p(x)p(y) - p_{\min}^2$. Hence, the assertion. ■

We now combine the results of the preceding lemmata to bound the eigenvalues of the d-dimensional Gaussian kernel for general measures.

Lemma 7 (Maximal decay of eigenvalues)**.** *For every $\ell \in \mathbb{N}$ it holds that*

$$\mu_\ell^{(d)} \geq \frac{p_{\min}^2}{p_{\max}^2} C(d,\sigma) e^{-c_{\sigma,d}(\ell+d)^{\frac{2}{d}}}, \tag{B.7}$$

where $c_{d,\sigma}$ and $C(d,\sigma)$ are constants depending on the dimension d and the bandwidth σ.

Proof We show the assertion only for the uniform measure $\mathcal{U}[0,1]^d$, as the result for more general design measures follows immediately by (B.6). For the eigenvalues μ_ℓ in the univariate setting, we have from (B.1) that

$$\mu_\ell \geq \frac{1}{\ell} C(\sigma) e^{-a_\sigma(\ell-1)^2} \geq C(\sigma) e^{-\tilde{a}_\sigma \ell^2} \tag{B.8}$$

where $\tilde{a}_\sigma$ is chosen such that

$$\tilde{a}_\sigma \ell^2 \geq \ln(\ell) + a_\sigma(\ell-1)^2, \qquad \ell = 1, 2, \ldots.$$

Combining (B.8) with (B.5) (cf. Lemma 5), it follows that

$$\mu_\ell^{(d)} \geq C(\sigma)^d C \exp(-\tilde{a}_\sigma c(\ell+d)^{\frac{2}{d}})$$

as $C(\sigma)$ appears in every factor of the product $\prod_{i=1}^d \mu_{\ell_i}$. Setting $c_{\sigma,d} \coloneqq \tilde{a}c$ reveals the assertion. ■

Appendix C. Concentration

This section provides a proof of the operator bound (4.3). To this end, we restate the following concentration bound for random operators on Hilbert spaces, which we then subsequently.

Proposition 4 (See Fischer and Steinwart 2020, Theorem A.3)**.** Let $(\Omega, \mathcal{B}, P)$ be a probability space, H a separable Hilbert space, and $\xi\colon \Omega \to \mathcal{L}_2(H)$ be a random variable with values in the set of self-adjoint Hilbert-Schmidt operators. Furthermore, let the operator norm be uniformly, i.e., $\|\xi\|_{H\to H} \le B$ P-a.s. and V be a self-adjoint positive semi-definite trace class operator with $\mathbb{E}_P\, \xi^2 \preccurlyeq V$, i.e. $V - \mathbb{E}_P\, \xi^2$ is positive semi-definite. Then, for $g(V) \coloneqq \ln(2e \operatorname{tr}(V)\, \|V\|_{H\to H}^{-1})$, $\tau \ge 1$ and $n \ge 1$, the following concentration inequality is satisfied:

$$P^n\Biggl(\biggl\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}_P\, \xi\biggr\|_{H\to H} \ge \frac{4\tau\, B\cdot g(V)}{3n} + \sqrt{\frac{2\tau\, \|V\|_{H\to H}\, g(V)}{n}}\Biggr) \le 2e^{-\tau}. \tag{C.1}$$

To demonstrate the desired bound (4.3), we show first that $(L_k+\lambda)^{-1/2}\, L_k\, (L_k+\lambda)^{-1/2}$ and $(L_k+\lambda)^{-1/2}\, L_k^D\, (L_k+\lambda)^{-1/2}$ are close in operator norm. To this end we rephrase the operator L_k^D in terms of simple operators. Letting $x_1, \dots, x_n \sim P$, independently distributed with respect to the design measure, and defining the operators

$$T_z\colon \mathcal{H}_k \to \mathcal{H}_k \quad \text{with} \quad (T_z f)(y) = f(z)k(y,z) \tag{C.2}$$

gives the representation

$$L_k^D = \frac{1}{n}\sum_{i=1}^n T_{x_i},$$

which fits the setting of Proposition 4. With this we have the subsequent concentration bound.

Proposition 5. For $\mathcal{N}_\infty(\lambda)$ as in (4.1) it holds

$$\Bigl\|(L_k+\lambda)^{-1/2}\, (L_k - L_k^D)\, (L_k+\lambda)^{-1/2}\Bigr\|_{\mathcal{H}_k\to\mathcal{H}_k} \le \frac{4\tau\mathcal{N}_\infty(\lambda)g(\lambda)}{3n} + \sqrt{\frac{2\tau\mathcal{N}_\infty(\lambda)g(\lambda)}{n}} \tag{C.3}$$

with probability at least $1-2e^{-\tau}$, where

$$g(\lambda) \coloneqq \ln\left(2e\frac{\lambda+\mu_1}{\mu_1}\mathcal{N}(\lambda)\right) \tag{C.4}$$

and $\tau \ge 1$.

Proof For $x \sim P$ and T_x as in (C.2) we consider the operator-valued random variable $\xi(\omega) := (L_k+\lambda)^{-1/2}T_{x(\omega)}(L_k+\lambda)^{-1/2}$. We have from

$$\mathbb{E}(T_x f)(y) = \mathbb{E}_x\, f(x)k(y,x) = \int_{\mathcal{X}} f(x)k(y,x)p(x)dx = (L_k f)(y)$$

that

$$\frac{1}{n}\sum_{i=1}^n \xi_i = (L_k+\lambda)^{-1/2}L_k^D(L_k+\lambda)^{-1/2} \quad \text{and} \quad \mathbb{E}\,\xi = (L_k+\lambda)^{-1/2}L_k(L_k+\lambda)^{-1/2},$$

and thus the setting of Proposition 4. It is thus sufficient to show that ξ satisfies the requirements of Proposition 4, provided that there is an appropriate constant B as well as dominating operator V.

We bound the norm $\|\xi\|_{\mathcal{H}_k\to\mathcal{H}_k}$ first. It holds that

$$\begin{aligned}\|\xi f\|_k^2 &= \left\|(L_k+\lambda)^{-1/2}k_{x(\omega)}\big((L_k+\lambda)^{-1/2}f\big)(x(\omega))\right\|_k^2\\ &= \left\|(L_k+\lambda)^{-1/2}k_{x(\omega)}\right\|_k^2 \langle (L_k+\lambda)^{-1/2}f, k_{x(\omega)}\rangle_k^2 \le \|f\|_k^2\mathcal{N}_\infty^2(\lambda)\end{aligned}$$

and thus $\|\xi\|_{\mathcal{H}_k\to\mathcal{H}_k}\le \mathcal{N}_\infty(\lambda)$ P-almost surely. This discloses the constant $B=\mathcal{N}_\infty(\lambda)$.

To bound the second moment, note first that ξ is a positive definite operator. Hence, we have that

$$\mathbb{E}\,\xi^2 \preccurlyeq \|\xi\|_{\mathcal{H}_k\to\mathcal{H}_k}\,\mathbb{E}\,\xi = \|\xi\|_{\mathcal{H}_k\to\mathcal{H}_k}\,(L_k+\lambda)^{-1}L_k \le \mathcal{N}_\infty(\lambda)(L_k+\lambda)^{-1}L_k =: V$$

by employing the bound for $\|\xi\|_{\mathcal{H}_k\to\mathcal{H}_k}$. The operator norm of V is bounded by

$$\|V\|_{\mathcal{H}_k\to\mathcal{H}_k} = \left\|\mathcal{N}_\infty(\lambda)(L_k+\lambda)^{-1}L_k\right\|_{\mathcal{H}_k\to\mathcal{H}_k} = \mathcal{N}_\infty(\lambda)\frac{\mu_1}{\lambda+\mu_1}\le \mathcal{N}_\infty(\lambda)$$

as well as

$$\begin{aligned}g(V) &= \ln\big(2e\,\mathrm{tr}(V)\cdot\|V\|^{-1}_{\mathcal{H}_k\to\mathcal{H}_k}\big)\\ &= \ln\left(2e\mathcal{N}_\infty(\lambda)\mathcal{N}(\lambda)\cdot\frac{1}{\mathcal{N}_\infty(\lambda)\frac{\mu_1}{\lambda+\mu_1}}\right) = \ln\left(2e\frac{\lambda+\mu_1}{\mu_1}\mathcal{N}(\lambda)\right)\end{aligned}$$

corresponding to $g(\lambda)$ in (C.4). The desired inequality (C.3) follows from Proposition 4. ■

Building on the bound (C.3) above, the assertion of Proposition 3 follows from the subsequent considerations. Note first the operator identity

$$\big(I-(L_k+\lambda)^{-1/2}(L_k-L_k^D)(L_k+\lambda)^{-1/2}\big)^{-1} = (L_k+\lambda)^{1/2}(L_k^D+\lambda)^{-1}(L_k+\lambda)^{1/2},$$

from which we conclude that

$$\begin{aligned}&\left\|(L_k+\lambda)^{1/2}(L_k^D+\lambda)^{-1}(L_k+\lambda)^{1/2}\right\|_{\mathcal{H}_k\to\mathcal{H}_k}\\ &\qquad= \left\|\big(I-(L_k+\lambda)^{-1/2}(L_k-L_k^D)(L_k+\lambda)^{-1/2}\big)^{-1}\right\|_{\mathcal{H}_k\to\mathcal{H}_k}\\ &\qquad\le \frac{1}{1-\left\|(L_k+\lambda)^{-1/2}(L_k-L_k^D)(L_k+\lambda)^{-1/2}\right\|_{\mathcal{H}_k\to\mathcal{H}_k}}\end{aligned}$$

by involving the Neumann series. Assuming the condition (4.2), we have by (C.3) that

$$\left\|(L_k+\lambda)^{-1/2}(L_k-L_k^D)(L_k+\lambda)^{-1/2}\right\|_{\mathcal{H}_k\to\mathcal{H}_k}\le\frac{4}{3}\tau\,g(\lambda)\frac{\mathcal{N}_\infty(\lambda)}{n}+\sqrt{2\tau\,g(\lambda)\frac{\mathcal{N}_\infty(\lambda)}{n}}\le\frac{1}{2}$$

with probability at least $1-2e^{-\tau}$. Therefore, the bound

$$\begin{aligned}&\left\|(L_k+\lambda)^{1/2}(L_k^D+\lambda)^{-1}(L_k+\lambda)^{1/2}\right\|_{\mathcal{H}_k\to\mathcal{H}_k}\\ &\qquad\le\frac{1}{1-\left\|(L_k+\lambda)^{-1/2}(L_k-L_k^D)(L_k+\lambda)^{-1/2}\right\|_{\mathcal{H}_k\to\mathcal{H}_k}}\le\frac{1}{1-\frac{1}{2}}=2\end{aligned}$$

holds also with probability at least $1-2e^{-\tau}$. This is the desired inequality (4.3).

Appendix D. Auxiliary lemmata

The following two lemmata provide a crucial element for the proof of Lemma 3. That is, a bound on the expectation

$$\mathbb{E}\, e^{-\frac{1}{M^2}},$$

where

$$M \coloneqq \min_{\substack{i,j=1,\dots,n\\ i\neq j}} |U_i - U_j|$$

is the minimal gap between n independently chosen uniforms.

The first lemma provides the density of M explicitly, the other bounds the associated expectation.

Lemma 8. *Let $U_1,\dots,U_n \sim \mathcal{U}[0,1]$ be independent uniforms. The random variable M has the density*

$$p_M(m) = \begin{cases} n(n-1)(1-(n-1)m)^{n-1} & \mathit{for}\, m \in \left[0, \frac{1}{n-1}\right] \\ 0 & \mathit{else} \end{cases}. \tag{D.1}$$

Proof Let $U_1,\dots,U_n$ be independent uniforms on $[0,1]$ and denote the corresponding minimal absolute difference by $M \coloneqq \min_{i,j=1,\dots,n, i\neq j} |U_i - U_j|$. Note here that $M \leq \frac{1}{n-1}$, and $M = \frac{1}{n-1}$ if all $U_1,\dots,U_n$ are equidistant. Therefore, let $m \in [0, \frac{1}{n-1}]$ and observe that

$$P(M > m) = n!P(M > m, U_1 \leq U_2 \cdots \leq U_n) \tag{D.2}$$

as there are $n!$ possible rearrangements of the random variables $U_1,\dots,U_n$. For the latter probability we have that

$$\begin{aligned} &P(M > m,\ U_1 \leq U_2 \leq \ldots \leq U_n) \\ &\qquad = P(U_1 \leq U_2 - m,\ U_2 \leq U_3 - m, \dots, U_{n-1} \leq U_n - m,\ U_1 \leq \ldots \leq U_n) \\ &\qquad = \lambda\left(\mathcal{U}_m\right), \end{aligned}$$

where $\lambda\left(\mathcal{U}_m\right)$ is the Lebesgue measure of the set

$$\mathcal{U}_m = \left\{(u_1,\dots,u_n) \in [0,1]^n : u_1 \leq u_2 - m, \dots, u_{n-1} \leq u_n - m,\ u_1 \leq u_2 \cdots \leq u_n\right\}.$$

We next present a measure persevering bijection between $\mathcal{U}_m$ and

$$Y_m \coloneqq \left\{(y_1,\dots,y_n) \in [0, 1-(n-1)m]^n : y_1 \leq y_2 \leq \cdots \leq y_n\right\}.$$

To this end, define $T\colon U_m \to Y_m$ with $Tu = u - (0, m, 2m, \dots, (n-1)m)$. For $u = (u_1,\dots,u_n) \in U_m$ and $y = Tu = u - (0, m, 2m, \dots, (n-1)m)$ it is evident that $y_i \geq 0$ as well as $y_1 \leq \cdots \leq y_n$. Furthermore, it holds that $u_i \leq 1 - m\,(n-i)$ as the distance between the u_i and u_{i+1} is at least m, for every $i = 1,\dots,n$. With this we have the inequality

$$y_i = u_i - (i-1)m \leq 1 - m\,(n-i) - (i-1)m = 1 - (n-1)m$$

and therefore $y \in Y_m$. Conversely, let $y \in Y_m$ and set $u = y + (0, m, 2m, \dots, (n-1)m) = T^{-1}y$. It is again immediate that $u_i \geq 0$ as well as

$$u_i = y_i + (i-1)m \leq 1 - (n-1)m + (i-1)m \leq 1.$$

From $y_1 \leq \cdots \leq y_n$ we further get that

$$u_{i+1} - m = y_{i+1} + i\,m - m = y_{i+1} + (i-1)m \geq y_i + (i-1)m = u_i$$

for all $i = 1, \ldots, n-1$ and thus $u \in \mathcal{U}_m$. Hence, T is a bijection, from which we conclude that $\lambda(U_m) = \lambda(Y_m)$. The latter measure is

$$\begin{aligned}\lambda(Y_m) &= \lambda\left(\{(y_1, \ldots, y_n) \in [0, 1-(n-1)m]^n : y_1 \leq y_2 \leq \cdots \leq y_n\}\right)\\ &= \frac{1}{n!}\lambda\left(\{(y_1, \ldots, y_n) \in [0, 1-(n-1)m]^n\}\right) = \frac{1}{n!}(1-(n-1)m)^n. \end{aligned} \tag{D.3}$$

Combining (D.2) with (D.3), we get that

$$P(M > m) = n!\frac{1}{n!}(1-(n-1)m)^n = (1-(n-1)m)^n$$

and hence the density

$$p(m) = \frac{d}{dm}(1 - P(M > m)) = n(n-1)(1-(n-1)m)^{n-1}.$$

This is the assertion. ∎

Lemma 9. *Let $U_1, \ldots, U_n \sim \mathcal{U}[0,1]$ be independent uniforms and M the minimum gap as in Lemma 8. For any $c > 0$ it holds that*

$$\mathbb{E}\, e^{-cM^{-2}} \geq 4e^{-c\frac{2}{a^2}} \geq Ce^{-8c(n-1)^2}, \tag{D.4}$$

with $a = \min\{\frac{1}{3}c^{-\frac{2}{3}},\ \frac{1}{2(n-1)}\}$.

Proof We employ the density (D.1) of M. As $\frac{1}{n-1} \leq \frac{1}{2(n-1)}$ and $e^{-cM^{-2}} \geq 0$, we have that

$$\begin{aligned}\mathbb{E}\, e^{-cM^{-2}} &= \int_0^{\frac{1}{n-1}} e^{-c\frac{1}{m^2}}\, n(n-1)\big(1-(n-1)m\big)^{n-1} dm\\ &> \int_0^{\frac{1}{2(n-1)}} e^{-c\frac{1}{m^2}} n(n-1)(1-(n-1)m)^{n-1} dm\\ &\geq e^{-(n-1)} \int_0^{\frac{1}{2(n-1)}} e^{-c\frac{1}{m^2}}\, dm.\end{aligned}$$

To bound the latter integral term, note that

$$\frac{x^3}{c}e^{-\frac{x^2}{c}} \leq 1$$

whenever $x \geq 3c^{\frac{2}{3}}$, as

$$\ln\frac{x^3}{c}e^{-\frac{x^2}{c}} = 3\ln\frac{x}{c^{\frac{1}{3}}} - \frac{x^2}{c} \leq 3\frac{x}{c^{\frac{1}{3}}} - \frac{x^2}{c} \leq 0$$

is satisfied for all $x \geq 3c^{\frac{2}{3}}$. Thus, choosing $a = \min\{\frac{1}{3}c^{-\frac{2}{3}}, \frac{1}{2(n-1)}\}$ we get that

$$\int_0^{\frac{1}{2(n-1)}} e^{-c\frac{1}{x^2}}\,dx \geq \int_0^a \frac{c}{x^3} e^{-c\frac{2}{x^2}}\,dx = \left[4e^{-c\frac{2}{x^2}}\right]_0^a = 4e^{-c\frac{2}{a^2}},$$

which is the assertion. ■