Pure and Applied Functional Analysis
Volume 8, Number 5, 2023, 1377–1399

Yokohama Publishers
ISSN 2189-3764 ONLINE JOURNAL

STOCHASTIC OPTIMIZATION WITH ESTIMATED OBJECTIVES

PAUL DOMMEL AND ALOIS PICHLER

in honor of Roger J.-B. Wets on his 85th birthday

Abstract. We consider a generalization of stochastic optimization problems based on randomized first stage decisions. To estimate the objective uniformly, we address the problem in reproducing kernel Hilbert spaces. It is demonstrated that the estimator, which is often derived by employing Gaussian random fields, converges in mean norm of the reproducing kernel Hilbert space to the conditional expectation and this implies local and uniform convergence of this function estimator. By preselecting the kernel, the problem does not suffer from the curse of dimensionality.

The paper analyzes the statistical properties of the estimator. We derive convergence properties and provide a conservative rate of convergence for increasing sample sizes.

1. Introduction

Programming under uncertainty, since its early occurrence in [27], has seen an enormous development with applications in economics and various industries. This paper extends the stochastic optimization problem for nonlinear objectives, which are not known explicitly, but need to be estimated themselves.

The classical optimization problem under uncertainty (cf. [21]) is

$$\begin{aligned} &\text{minimize } f(x) \coloneqq \mathbb{E}_\xi\, f(x,\xi) \\ &\text{subject to } x \in \mathcal{X}, \end{aligned} \tag{1.1}$$

where the expectation $\mathbb{E}$ is with respect to the random variable ξ (that is, $\mathbb{E}_\xi\, f(x,\xi) = \int f(x,\xi) P(d\xi)$) and the set $\mathcal{X}$ collects possible first stage decisions.

The genuine problem exposition (1.1) requires to evaluate the function $f(x) = \mathbb{E}_\xi\, f(x,\xi)$ for varying $x \in \mathcal{X}$. However, most often, the measure P is not available explicitly but instead, only random iid observations ξ_i, $i = 1, \dots, n$, are accessible. Sample average approximation investigates the statistical properties of the solutions of (1.1) with expectation replaced by the sample mean.

This paper builds on random observations

$$(X_i, f_i) \in \mathcal{X} \times \mathbb{R}, \quad i = 1, \dots, n, \tag{1.2}$$

2020 *Mathematics Subject Classification.* 62G05, 62G08, 62G20.

Key words and phrases. Reproducing kernel Hilbert spaces, positive definite functions, Gramian matrix, Mercer kernels, Statistical learning theory.

Technische Universität Chemnitz, Faculty of mathematics. 90126 Chemnitz, Germany

DFG, German Research Foundation – Project-ID 416228727 – SFB 1410.

and does not assume any explicit functional relation of the form $f_i = f(X_i, \xi_i)$.[1] With that, the objective in (1.1) reduces to the conditional expectation

$$f(x) = \mathbb{E}\left(f \mid X = x\right), \tag{1.3}$$

where the random vector/ random pair (X, f) has the same distribution as (X_i, f_i) for all $i = 1, \dots, n$, and the problem formulation considered here thus goes beyond the classical stochastic optimization problem (1.1). Note, in addition, that estimating the conditional expectation (1.3) based on samples is challenging in general, and particularly crucial in the context of optimization considered here.

To solve the optimization problem (1.1) it is crucial to estimate the conditional expectation $\mathbb{E}\left(f \mid X\right)$, i.e., (1.3), uniformly on its entire support. The estimator we consider here derives from Gaussian random fields and is central in support vector machines as well. Here, the estimator is often inferred with least squares errors and by involving a regularization term based on a reproducing kernel Hilbert space. The literature frequently employs loss and risk functionals, and involves an L^2-error to investigate this estimator. However, this error is not adequate to investigate (1.1), where uniform convergence is crucial, and our results thus consider the estimator in its natural, genuine norm. They enable us to establish uniform convergence of the estimator by moderately regularizing the objectives.

Explicit convergence rates are presented for increasing sample sizes. The results and convergence rates correspond to other rates known from non-parametric statistics, particularly to density estimation when employing the mean (integrated) squared error.

We derive error bounds with respect to the mean of the underlying norm. This is the usual error measure for many statistical techniques, including, for example, kernel density estimation. Our methods build on conditional expectation and thus complement the predominant literature which is mainly based on concentration type results. Using this setting allows us to prove error bounds directly, without involving auxiliary quantities such as covering numbers.

[6] provide an introduction to approximation theory in a random framework. The excellent book [4, Section 2.3] gives very concrete applications in statistical learning theory, while [26] provide the mathematical foundations for approximations in reproducing kernel Hilbert spaces. The monograph [24] introduces to support vector machines, which employ kernel functions similarly to our approach presented below, see also [8]. A study, comparably to ours but employing a simpler norm, is [28]. [5] provide the state of the art for an analysis in L^2 involving the kernel operator, see also [9].

Outline of the paper. The following Section 2 repeats elements from reproducing kernel Hilbert spaces, which are of importance throughout this paper. Section 3 introduces the elementary estimator, which is employed in statistical learning. Sample average approximation (Section 3.2) address this estimator with random samples from both dimensions and Section 4 reveals related statistical results. The Sections 5 and 6 derive our main results, which is, for short, convergence of the sample average

[1] That is, the observations are (X_i, f_i) instead of $(X_i, f(X_i, \xi_i))$; the latter would require involving a function f.

optimizer in mean norm and weak consistency (Section 6.2) of this estimator. Section 7 concludes with a summary.

2. Regularization with reference to reproducing kernel Hilbert spaces

Throughout we shall expose the problem on the design space $\mathcal{X}$, an arbitrary set for which we impose more structure later; most typically, $\mathcal{X}$ is a subset of $\mathbb{R}^d$. Let (X_i, f_i), $i = 1, \dots, n$, be independent and identically distributed random vectors in $\mathcal{X} \times \mathbb{R}$ with joint probability measure ρ. For a kernel function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ we consider the estimator

$$\hat{f}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i)\, \hat{w}_i, \tag{2.1}$$

where the weights $\hat{w}_i$ satisfy the system of linear equations

$$\lambda_n\, \hat{w}_i + \frac{1}{n} \sum_{j=1}^{n} k(X_i, X_j)\, \hat{w}_j = f_i, \qquad i = 1, \dots, n, \tag{2.2}$$

for some parameter λ_n.[2] In what follows we derive the estimator (2.1) first by employing Gaussian random fields and kernel ridge regression from support vector machines and then investigate and expose its convergence properties. Specifically, we identify and characterize the function f so that

$$\mathbb{E}\, \|\hat{f}_n(\cdot) - f(\cdot)\|^2 \to 0 \tag{2.3}$$

as $n \to \infty$, where $\|\cdot\|$ is an appropriate norm and λ_n is chosen adequately; above all, we derive results for the norm of the reproducing kernel Hilbert space associated with the kernel function. We will also infer convergence results for L^2 and—most importantly—for uniform function approximations to handle stochastic optimization problems as exposed in (1.1).

2.1. Gaussian random fields.

As an initial motivation for the estimator (2.1) consider a zero mean Gaussian random field f on $\mathcal{X}$ with covariance function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, that is, $k(x, y) = \operatorname{cov}\big(f(x), f(y)\big)$. For a signal plus noise model with observations

$$f_i = f(x_i) + \epsilon_i,$$

the joint distribution, including x to the observation points $X = (x_1, \dots, x_n)$, is

$$\begin{pmatrix} f(x) \\ f \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(x, x) & k(x, X) \\ k(X, x) & k(X, X) + \lambda \end{pmatrix} \right),$$

where $\epsilon \sim \mathcal{N}(0, \lambda)$ is the independent error and where we use the compact vector notation $f := (f_1, \dots, f_n)^\top$ and $k(x, X) := \big(k(x, x_1), \dots, k(x, x_n)\big)$ for the entry of the covariance matrix; the other entries are defined analogously. With this, the conditional distribution is Gaussian (cf. [23, Theorem 13.1] or [4, Section 2.3]),

$$f(x) \;\big|\; \big(f(X) = f\big) \sim \mathcal{N}\big(\hat{\mu}(x),\, \hat{K}(x)\big),$$

[2] Note that $\hat{f}_n(\cdot)$ interpolates the data, $\hat{f}_n(X_i) = f_i$, $i = 1, \dots, n$, for the particular choice $\lambda_n = 0$ provided that all X_i are distinct and k is regular enough.

where

$$\hat{\mu}(x) := k(x, X)\big(k(X, X) + \lambda\big)^{-1} f(X) \tag{2.4}$$

is the mean and the variance is

$$\hat{K}(x) := k(x, x) - k(x, X)\big(k(X, X) + \lambda\big)^{-1} k(X, x). \tag{2.5}$$

Expanding (2.4) and setting $\hat{f}_n(x) := \hat{\mu}(x)$ reveals the initial estimator (2.1) for variance λ rescaled.

Figure 1 displays an example of the estimator $\hat{f}_n(\cdot)$ together with the range $\pm\sqrt{\hat{K}(\cdot)}$ (cf. (2.5)) coming along with the mean (2.4). The figure reveals that the estimator $\hat{f}_n(x)$ is more precise (i.e., the variance is smaller), if more observations are available locally at x.

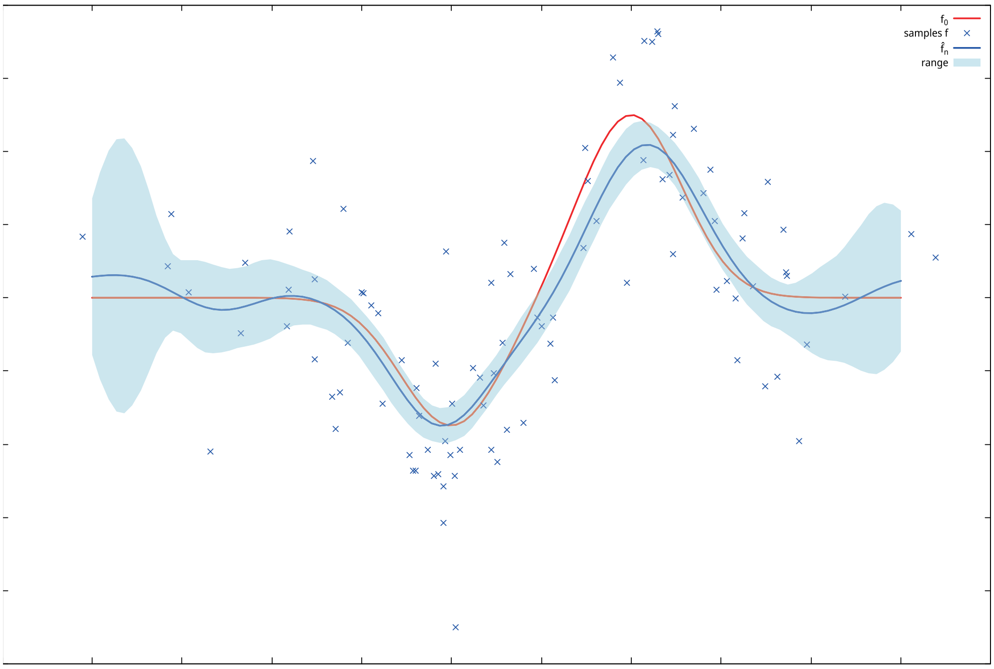

FIGURE 1. Gaussian field regression $\hat{f}_n(\cdot)$ of an exemplary function $f_0 \in \mathcal{H}_k$ using a Gaussian kernel (sample size $n = 100$, the regression parameter is $\lambda = 0.03$). The width of the blue strip indicates the local precision of the estimator.

2.2. **Reproducing kernel Hilbert space.** Every estimator $\hat{f}_n(\cdot)$ in (2.1) is an element in the reproducing kernel Hilbert space spanned by the functions $k(\cdot, y)$, $y \in \mathcal{X}$. While introducing the notation for reproducing kernel Hilbert spaces here, we briefly recall major properties, which are essential in the following exposition. For a general discussion on reproducing kernel Hilbert spaces we may refer to [14, Chapter 1].

Definition 2.1. The kernel is a symmetric and positive definite function $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$. On the linear span $\{k(\cdot,x)\colon \mathcal{X}\to\mathbb{R} \mid x\in\mathcal{X}\}$ of functions on $\mathcal{X}$, the inner product is defined by

$$\big\langle k(\cdot,x) \mid k(\cdot,y)\big\rangle_k := k(x,y). \tag{2.6}$$

The reproducing kernel Hilbert space, denoted $\big(\mathcal{H}_k, \|\cdot\|_k\big)$, is the completion with respect to the norm $\|f\|_k^2 := \langle f \mid f\rangle_k$ induced by the inner product (2.6).

Most importantly, point evaluations are continuous linear functions in reproducing kernel Hilbert spaces. Indeed, finite linear combinations $f(\cdot)=\sum_{i=1}^n k(\cdot,x_i)\,w_i$ are dense in $\mathcal{H}_k$, and it follows with (2.6) that

$$\big\langle k(\cdot,x)\big|\, f(\cdot)\big\rangle_k = \sum_{i=1}^n w_i \,\big\langle k(\cdot,x)\big|\, k(\cdot,x_i)\big\rangle_k = \sum_{i=1}^n w_i\, k(x,x_i) = f(x). \tag{2.7}$$

Although more general settings are easily possible, in what follows we convene to address only continuous and uniformly bounded kernel functions k. The space $\mathcal{H}_k$ thus is naturally embedded in $L^2(\mathcal{X},P)$, where P is a probability measure on the Borel sets of $\mathcal{X}$.

We associate the following Hilbert–Schmidt integral operator L_k with a kernel k.

Definition 2.2. Let k be a kernel. The operator $L_k\colon L^2(\mathcal{X},P)\to L^2(\mathcal{X},P)$ is

$$L_k\, w(x) := \int_{\mathcal{X}} k(x,y)\, w(y)\, P(dy). \tag{2.8}$$

Proposition 2.3. *The operator L_k is self-adjoint and positive definite with respect to the standard inner product*

$$\langle f \mid g\rangle := \int_{\mathcal{X}} f(z)\cdot g(z)\, P(dz)$$

on $\big(L^2,\|\cdot\|_2\big)$. The operator is positive definite and bounded with norm

$$\|L_k\colon L^2\to L^2\|^2 \le \iint_{\mathcal{X}^2} k(x,y)^2\, P(dx)P(dy);$$

the norm of the operator L_k is $\|L_k\colon L^2\to L^2\| := \sup\{\|L_k\, w\|_2\colon \|w\|_2\le 1\}$.

Proof. The assertion is a consequence of the Cauchy–Schwarz inequality. □

Proposition 2.4. *It holds that $\|k(\cdot,x)\|_k^2 = k(x,x)$,*

$$\langle L_k w \mid f\rangle_k = \langle w \mid f\rangle \quad \textit{and} \quad \|L_k w\|_k^2 = \langle w \mid L_k w\rangle\,. \tag{2.9}$$

Proof. The functions $f(\cdot)=\sum_{i=1}^n w_i' \, k(\cdot,x_i)$ are dense in $\mathcal{H}_k$. By linearity,

$$\begin{aligned}\langle L_k w \mid f\rangle_k &= \sum_{i=1}^n w_i' \int_{\mathcal{X}} \big\langle k(\cdot,y) \mid k(\cdot,x_i)\big\rangle_k w(y)\, P(dy)\\ &= \int_{\mathcal{X}} \sum_{i=1}^n w_i' \, k(y,x_i)\, w(y)\, P(dy)\\ &= \int_{\mathcal{X}} f(y)\, w(y)\, P(dy)\\ &= \langle w \mid f\rangle .\end{aligned}$$

The other assertions are immediate. □

Remark 2.5 (Mercer[3] and the kernel trick)**.** The operator L_k is compact and has countably many eigenfunctions. In machine learning, the decomposition is known as the *kernel trick*. It holds that $k(x,y)=\sum_{\ell=1}^\infty \sigma_\ell\, \phi_\ell(x)\, \phi_\ell(y)$, where σ_ℓ is the eigenvalue corresponding to the eigenfunction $\phi_\ell(\cdot)$. In this setting, the operator $L_k^{1/2}$ is $L_k^{1/2} f=\sum_{\ell=1}^\infty \sigma_\ell^{1/2}\, \phi_\ell \, \langle \phi_\ell \mid f\rangle$ (with $\sigma_\ell \geq 0$), cf. [17, Theorem VI.23].

Proposition 2.6 ($L_k^{1/2}\colon L^2 \to \mathcal{H}_k$ is an isometry)**.** *It holds that* $\|L_k^{1/2} f\|_k=\|f\|_2$ *and* $\|f\|_2 \leq \|L_k\|^{1/2}\cdot \|f\|_k$.

Proof. The assertion is a consequence of Mercer's theorem, cf. [12] or [10, Corollary 4]. However, for $f=L_k^{1/2}\, w$, it follows from the preceding proposition that

$$\|L_k^{1/2} f\|_k^2=\|L_k w\|_k^2=\langle w \mid L_k w\rangle=\left\langle L_k^{1/2} w \mid L_k^{1/2} w\right\rangle=\|f\|_2^2.$$

With (2.9) we have further that

$$\|f\|_2^2=\big\langle L_k^{1/2} w \mid L_k^{1/2} w\big\rangle=\big\langle w \mid L_k w\big\rangle \leq \|L_k\|\, \|w\|_2^2=\|L_k\|\, \|L_k^{1/2} w\|_k^2=\|L_k\|\, \|f\|_k^2,$$

as L_k is self-adjoint. Hence, the assertion. □

Theorem 2.7 (Continuity of the operator L_k)**.** *It holds that* $\|L_k\colon \mathcal{H}_k \to \mathcal{H}_k\| \leq \|L_k\colon L_2 \to L_2\|$, *where the norm is* $\|L_k\colon \mathcal{H}_k\to\mathcal{H}_k\| \coloneqq \sup\{\|L_k\, w\|_k \colon \|w\|_k \leq 1\}$, *cf. also Proposition 2.3.*

Proof. With (2.9) and Proposition 2.6, $\|L_k f\|_k^2=\langle f \mid L_k f\rangle \leq \|L_k\|\, \|f\|_2^2 \leq \|L_k\|^2 \|f\|_k^2$ and hence the assertion. □

We have seen in (2.7) that point evaluations are linear functionals. We shall conclude here by relating these norms to uniform convergence.

Proposition 2.8. *The point evaluation is continuous; indeed,* $|f(x)| \leq \sqrt{k(x,x)}\, \|f\|_k$ *for all* $x \in \mathcal{X}$ *and* $f \in \mathcal{H}_k$. *Further,*[4]

$$\|f\|_\infty \leq \|f\|_k \cdot \sup_{x\in \operatorname{supp} P} \sqrt{k(x,x)}, \tag{2.10}$$

where $\|f\|_\infty \coloneqq \sup_{x\in\operatorname{supp} P} |f(x)|$.

[3] The initial publication is notably due to Schmidt, see [19], and not Mercer.

[4] The support of the measure P is $\operatorname{supp} P \coloneqq \bigcap \big\{A \colon A \text{ is closed and } P(A)=1\big\} \subset \mathcal{X}$, cf. [18].

Proof. The statement is immediate from (2.7), as

$$|f(x)| = \left|\langle k(\cdot,x)|\, f\rangle_k\right| \le \|k(\cdot,x)\|_k\, \|f\|_k = \sqrt{k(x,x)}\, \|f\|_k$$

by the Cauchy–Schwartz inequality and Proposition 2.4. □

Remark 2.9. In the Gaussian process setting, the variance of the estimator $\hat{f}_n(x)$ does not exceed $k(x,x)$ (cf. (2.5)). The upper bound (2.10) is the accordant uniform estimator for the variance in the entire support. The example in Figure 1 visualizes this area.

3. The genuine approximation problem

In what follows we characterize the estimator (2.1) by involving a stochastic optimization problem. We consider the problem first in its continuous form and relate it to the data subsequently.

Let $(X_i, f_i) \in \mathcal{X} \times \mathbb{R}$, $i = 1, \dots, n$, be independent and identically distributed random vectors (cf. (1.3)) with common law ρ. By the disintegration theorem (see [7], [1] or [11, Chapter 5]), there is a family of measures $\rho(\cdot \mid x)\colon \mathcal{B}(\mathcal{X}) \to [0,1]$, $x \in \mathcal{X}$, on the Borel sets $\mathcal{B}(\mathcal{X})$ so that

$$\rho(A \times B) = \int_A \rho(B|\, x)\, P(dx), \quad A \subset \mathcal{X}, B \subset \mathbb{R} \text{ measurable},$$

where the marginal measure $P(\cdot) := \rho(\cdot \times \mathbb{R})$ on the design space $\mathcal{X}$ (cf. [24]) is called *design measure.*

For a random variable (X, f) with law ρ we recall the notational variants

$$\mathbb{E}\, g(X,f) = \iint_{\mathcal{X}\times\mathbb{R}} g(x,f)\, \rho(dx,df) = \int_{\mathcal{X}} g(x,f)\, \rho(df|\, x)\, P(dx) = \mathbb{E}\,\mathbb{E}\left(g(X,f)|\, X\right),$$

where g is measurable and

$$\mathbb{E}\left(g(x,f)|\, x\right) = \int_{\mathcal{X}} g(x,f)\, \rho(df|\, x), \qquad x \in \mathcal{X},$$

is the conditional expectation.

3.1. The continuous problem.

For the random vector (X, f) with values in $\mathcal{X}\times\mathbb{R}$, law ρ and $f \in L^2(\mathcal{X})$, consider the (stochastic) optimization problem

$$\min_{f_\lambda(\cdot)\in\mathcal{H}_k} \mathbb{E}\left(f - f_\lambda(X)\right)^2 + \lambda\, \|f_\lambda\|_k^2\,, \tag{3.1}$$

where $\lambda > 0$ is a fixed regression parameter and the expectation is with respect to the full measure ρ. The objective (3.1) is strictly convex, as the norm $\|\cdot\|_k$ is strictly convex for $\lambda > 0$ fixed.

The random variable $f_\lambda(X)$ is measurable with respect to $\sigma(X)$, the σ-algebra generated by X, and the random variable $\mathbb{E}(f \mid X)$ is the projection of f onto the closed subspace $L^2\big(\sigma(X)\big)$, see [11]. By the Pythagorean theorem, the objective in the preceding problem thus is equivalently

$$\min_{f_\lambda(\cdot)} \mathbb{E}\left(f - \mathbb{E}(f \mid X)\right)^2 + \mathbb{E}\left(\mathbb{E}(f \mid X) - f_\lambda(X)\right)^2 + \lambda\, \|f_\lambda\|_k^2\,.$$

It follows from the Doob–Dynkin lemma that there is a Borel function $f_0\colon \mathcal{X} \to \mathbb{R}$ so that $\mathbb{E}(f \mid X) = f_0(X)$. We follow the convention and denote this function also as

$$f_0(x) = \mathbb{E}(f \mid X = x). \tag{3.2}$$

The orthogonality relation characterizing f_0 is

$$\mathbb{E}\left(f - f_0(X)\right)g(X) = 0, \tag{3.3}$$

where $g\colon \mathcal{X} \to \mathbb{R}$ is any measurable test function. The objective of the optimization problem (3.1) thus is

$$\vartheta^* := \mathbb{E}\left(f - f_0(X)\right)^2 + \min_{f_\lambda(\cdot)} \mathbb{E}\left(f_0(X) - f_\lambda(X)\right)^2 + \lambda \left\|f_\lambda\right\|_k^2, \tag{3.4}$$

where the quantity $\mathbb{E}\left(f - f_0(X)\right)^2$ is the *irreducible error.*

Remark 3.1. We note that $f_0 \in L^2(\mathcal{X}, P)$, but f_0 is *not necessarily* in $\mathcal{H}_k$.

Theorem 3.2. *The solution of the optimization problem* (3.1) *is*

$$f_\lambda = L_k w_\lambda, \tag{3.5}$$

where $(\lambda + L_k)w_\lambda = f_0$*; the objective is*

$$\begin{aligned} \vartheta^* &= \|f - f_0\|_2^2 + \|f_0 - f_\lambda\|_2^2 + \lambda \|f_\lambda\|_k^2 \\ &= \|f - f_0\|_2^2 + \lambda^2 \|w_\lambda\|_2^2 + \lambda \langle w_\lambda \mid L_k w_\lambda\rangle . \end{aligned} \tag{3.6}$$

Proof. With (2.9) and Proposition 2.6 we may rewrite the objective in (3.4) by $g(w'_\lambda) := \left\|f_0 - L_k^{1/2} w'_\lambda\right\|_2^2 + \lambda \langle w'_\lambda \mid w'_\lambda\rangle$. Now note that

$$\begin{aligned} &g(w'_\lambda + h) - g(w'_\lambda) \\ &= \left\langle f_0 - L_k^{1/2} w'_\lambda - L_k^{1/2} h) \mid f_0 - L_k^{1/2} w'_\lambda - L_k^{1/2} h\right\rangle + \lambda \langle w'_\lambda + h \mid w'_\lambda + h\rangle \\ &\quad - \left\langle f_0 - L_k^{1/2} w'_\lambda \mid f_0 - L_k^{1/2} w'_\lambda\right\rangle - \lambda \langle w'_\lambda \mid w'_\lambda\rangle \\ &= -\left\langle L_k^{1/2} h \mid f_0 - L_k^{1/2} w'_\lambda\right\rangle - \left\langle f_0 - L_k^{1/2} w'_\lambda \mid L_k^{1/2} h\right\rangle + \left\langle L_k^{1/2} h \mid L_k^{1/2} h\right\rangle \\ &\quad + \lambda \langle h \mid w'_\lambda\rangle + \lambda \langle h \mid w'_\lambda\rangle + \lambda \langle h \mid h\rangle \\ &= -2\left\langle h \mid L_k^{1/2} f_0 - L_k w'_\lambda - \lambda w'_\lambda\right\rangle + \left\langle L_k^{1/2} h \mid L_k^{1/2} h\right\rangle + \lambda \langle h \mid h\rangle \end{aligned}$$

as $L_k^{1/2}$ is self-adjoint. The first, linear term vanishes if $(\lambda + L_k)w'_\lambda = L_k^{1/2} f_0$, and the second is quadratic in h – hence the infimum at $L_k^{1/2} w'_\lambda = L_k(\lambda + L_k)^{-1} f_0 = f_\lambda$, the first assertion. For the objective (3.6) note that $f_0 - f_\lambda = \lambda\, w_\lambda$, see also (3.9) below. □

Corollary 3.3 (Characterization of the coefficient function)**.** *Suppose that*

$$(\lambda + L_k)w_\lambda = f_0, \tag{3.7}$$

then

$$f_\lambda := L_k w_\lambda = (\lambda + L_k)^{-1} L_k f_0 \tag{3.8}$$

solves the Fredholm equation of the second kind $(\lambda + L_k)f_\lambda = L_k f_0$, *and it holds that*

$$f_0 - f_\lambda = \lambda\, w_\lambda. \tag{3.9}$$

Proof. Apply L_k to (3.7) to get $\lambda L_k w_\lambda + L_k L_k w_\lambda = L_k f_0$, that is, $(\lambda + L_k)f_\lambda = L_k f_0$. □

Remark 3.4. It follows from (3.8) that $f_\lambda \in \mathcal{H}_k$, even more, f_λ is in the image of L_k, although f_0 is *not necessarily* in $\mathcal{H}_k$ (cf. Remark 3.1).

Remark 3.5. The functions f_0, f_λ and w_λ are in L^2 and hence exhibit a representation in terms of the orthonormal basis $(\phi_\ell)_{\ell=1}^\infty$ of Remark 2.5. More precisely, for $f_0 = \sum_{\ell=1}^\infty c_\ell\, \phi_\ell$ we have that

$$w_\lambda(\cdot) = \sum_{\ell=1}^\infty \frac{c_\ell}{\lambda + \sigma_\ell}\, \phi_\ell(\cdot) \quad \text{and} \quad f_\lambda(\cdot) = \sum_{\ell=1}^\infty \frac{\sigma_\ell\, c_\ell}{\lambda + \sigma_\ell}\, \phi_\ell(\cdot) \tag{3.10}$$

with some coefficients $(c_\ell)_{\ell=1}^\infty \in \ell^2$. This is a consequence of the characterizing equations in Theorem 3.3 as well as the Mercer representation $L_k f = \sum_{\ell=1}^\infty \sigma_\ell\, \phi_\ell\, \langle \phi_\ell \mid f\rangle$ of the operator.

The distance of the solution f_λ to the function f_0 will be of importance in what follows. We have the following general result.

Proposition 3.6. *Suppose that* f_0 *is in the range of* L_k. *Then there is a constant* $C_0 > 0$ *so that*

$$\|f_0 - f_\lambda\|_k^2 \le C_0\, \lambda. \tag{3.11}$$

Proof. As f_0 is in the range of L_k there is some $w \in L^2$ so that $f_0 = L_k\, w$. We hence have the series representation $f_0(\cdot) = \sum_{\ell=1}^\infty \sigma_\ell\, w_\ell\, \phi_\ell(\cdot)$ with some sequence $(w_\ell)_{\ell=1}^\infty$ such that $\sum_{\ell=1}^\infty w_\ell^2 < \infty$. Thus, by (3.9) and (3.10), we observe that

$$\|f_0 - f_\lambda\|_k^2 = \|\lambda\, w_\lambda\|_k^2 = \lambda^2 \sum_{\ell=1}^\infty \frac{1}{\sigma_\ell} \left(\frac{\sigma_\ell\, w_\ell}{\lambda + \sigma_\ell}\right)^2 \le \lambda^2 \sum_{\ell=1}^\infty \frac{\sigma_\ell^2\, w_\ell^2}{2\lambda \sigma_\ell^2} = \frac{\lambda}{2}\, \|w\|_2^2$$

and thus the assertion with the constant $C_0 \coloneqq \frac{1}{2}\|w\|_2^2$. □

The following corollary to Corollary 3.3 provides the weight functions with respect to the usual Lebesgue measure. We provide this statement as it particularly useful to solving the Fredholm integral equation (3.7) numerically (by employing the Nyström method, for example, cf. [2]) to make the function f_λ available for computational purposes.

Corollary 3.7 (Coefficient function for measures with a density)**.** *Suppose that* P *has a density* $p(\cdot)$ *with respect to the Lebesgue measure,* $P(dx) = p(x)dx$, *and the coefficient function* $\tilde{w}_\lambda(\cdot)$ *satisfies*

$$\lambda\, \tilde{w}_\lambda(x) + p(x) \cdot \int_{\mathcal{X}} k(x,y)\, \tilde{w}_\lambda(y)\, dy = p(x) \cdot g_0(x). \tag{3.12}$$

Then the function $g_\lambda(\cdot) \coloneqq \int_{\mathcal{X}} k(\cdot, x)\, \tilde{w}_\lambda(x)\, dx$ *solves the integral equation*

$$(\lambda + L_k)g_\lambda = L_k g_0.$$

Proof. Multiply equation (3.12) by $k(y,x)$ and integrate with respect to dx to get

$$\lambda \int_{\mathcal{X}} k(y,x)\, \tilde{w}_\lambda(x)\, dx + \int_{\mathcal{X}} k(y,x) \cdot \int_{\mathcal{X}} k(x,z)\, \tilde{w}_\lambda(z)\, dz\, p(x)dx \\ = \int_{\mathcal{X}} k(y,x)\, g_0(x)\, p(x)dx.$$

This is

$$\lambda\, g_\lambda(y) + \int_{\mathcal{X}} k(y,x)\, g_\lambda(x)\, P(dx) = \int_{\mathcal{X}} k(y,x)\, g_0(x)\, P(dx),$$

or $(\lambda + L_k)g_\lambda = L_k g_0$, the assertion.

□

3.2. **The discrete problem and kernel ridge regression.** We now switch from the continuous problem (3.1) to learning from data. This alternative viewpoint highlights and justifies the genuine estimator (2.1) from an additional perspective.

Substituting the average for the expectation in (3.1) we consider the slightly more general objective

$$\frac{1}{n} \sum_{i,j=1}^{n} \big(f_i - f(x_i)\big)\Lambda_{ij}^{-1}\big(f_j - f(x_j)\big) + \|f\|_k^2\,, \tag{3.13}$$

where Λ is a symmetric and positive definite regularization matrix with entries Λ_{ij}. We use lowercase letters $x_i \in \mathcal{X}$ and $f_i \in \mathbb{R}$ to emphasize that these quantities are deterministic.

Proposition 3.8. *The function $f \in \mathcal{H}_k$ minimizing* (3.13) *is*

$$f(\cdot) = \frac{1}{n} \sum_{s=1}^{n} w_i \cdot k(\cdot, x_i), \tag{3.14}$$

where the weights are

$$w = n\big(K^\top \Lambda^{-1} K + n\, K\big)^{-1} K^\top \Lambda^{-1} f \tag{3.15}$$

and K is the Gram matrix with entries $K_{ij} = k(x_i, x_j)$.

Proof. Assuming that the optimal function is of the form (3.14), the objective (3.13) is

$$\frac{1}{n}(f - \frac{1}{n}Kw)^\top \Lambda^{-1}\big(f - \frac{1}{n}Kw\big) + \frac{1}{n^2} w^\top K w.$$

Differentiating with respect to w gives the first order conditions

$$0 = -\frac{1}{n^2}\left(K^\top \Lambda^{-1}\big(f - \frac{1}{n}Kw\big)\right)^\top - \frac{1}{n^2}\big(f - \frac{1}{n}Kw\big)^\top \Lambda^{-1} K + \frac{1}{n^2}(Kw)^\top + \frac{1}{n^2} w^\top K,$$

i.e.,

$$\frac{1}{n^2}\left(\frac{1}{n} K^\top \left(\Lambda^{-1} + \Lambda^{-\top}\right) K + K + K^\top\right) w = \frac{1}{n^2} K^\top \left(\Lambda^{-1} + \Lambda^{-\top}\right) f.$$

The assertion follows, as Λ^{-1} and K are both symmetric.

It remains to demonstrate that the optimal function is indeed of the form (3.14), i.e., the optimal function $f \in \mathcal{H}_k$ is located exactly on the supporting points

$x_1, \dots, x_n$. This, however, follows from the *representer theorem*, which [20] prove in the most general form. □

Corollary 3.9. *The function $f \in \mathcal{H}_k$ minimizing the objective*

$$\frac{1}{n}\sum_{i=1}^{n}\big(f_i - f(x_i)\big)^2 + \lambda\,\|f\|_k^2 \tag{3.16}$$

is $\hat{f}_n(\cdot) := \frac{1}{n}\sum_{j=1}^{n}\hat{w}_j\,k(\cdot, x_j)$ with weights $\hat{w} = \big(\lambda + \frac{1}{n}K\big)^{-1}f$. The associated optimal value is

$$\hat{\vartheta}_n = \frac{\lambda}{n} f^\top \left(\lambda + \frac{1}{n}K\right)^{-1} f. \tag{3.17}$$

Proof. The first assertion is immediate with $\Lambda = \lambda \cdot I$, the diagonal matrix with entries λ on its diagonal. Further, employing the minimizer $\hat{f}_n$ into the objective (3.16) we get that

$$\begin{aligned}\hat{\vartheta}_n &= \frac{\lambda}{n^2}\hat{w}^\top K\hat{w} + \frac{1}{n}\sum_{i=1}^{n}\big(\hat{f}_n(X_i) - f_i\big)^2 = \frac{\lambda}{n^2}\hat{w}^\top K\hat{w} + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{n}(K\hat{w})_i - f_i\right)^2\\ &= \frac{\lambda}{n^2}\hat{w}^\top K\hat{w} + \frac{1}{n}\sum_{i=1}^{n}(\lambda\hat{w}_i)^2 = \frac{\lambda}{n}\hat{w}^\top\left(\lambda + \frac{1}{n}K\right)\hat{w} = \frac{\lambda}{n}f^\top\left(\lambda + \frac{1}{n}K\right)^{-1} f\end{aligned}$$

and thus the second assertion. □

4. Elementary statistical properties

As above, let (X_i, f_i), $i = 1, \dots, n$, be independent samples from a joint measure ρ. We note that $X_i \sim P$ and the integral operator L_k in (2.8) can be restated as

$$L_k w(x) = \mathbb{E}\,k(x, X_i)\,w(X_i) = \mathbb{E}\,\big(k(X_i, X_j)\,w(X_j) \mid X_i = x\big);$$

we shall make frequent use of the latter relation.

Definition 4.1. For (X_i, f_i), $i = 1, \dots, n$, independent samples from a joint distribution ρ define the estimator

$$\hat{\vartheta}_n := \min_{\hat{f}_n(\cdot)} \frac{1}{n}\sum_{i=1}^{n}\big(f_i - \hat{f}_n(X_i)\big)^2 + \lambda\left\|\hat{f}_n\right\|_k^2. \tag{4.1}$$

It is evident that $\hat{\vartheta}_n$ is an $\mathbb{R}$-valued random variable, dependent on the samples (X_i, f_i). Further, the optimizer

$$\hat{f}_n(\cdot) := \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)\,\hat{w}_i \tag{4.2}$$

of (4.1) (cf. Corollary 3.9) is a random function, as it is supported by the samples X_i, $i = 1, \dots, n$, and the weights

$$\hat{w} = \left(\lambda + \frac{1}{n}K\right)^{-1} f \tag{4.3}$$

depend on all (X_i, f_i), $i = 1, \dots, n$. Relating to the term sample average approximation (SAA) in stochastic optimization we shall refer to the estimators $\hat{\vartheta}_n$ and $\hat{f}_n(\cdot)$ as the *SAA estimators*.

Example 4.2. A simple example is given by employing the trivial design measure $P = \delta_{x_0}$, where $x_0 \in \mathcal{X}$ is a fixed point and $\delta_{x_0}(A) := \begin{cases} 1 & \text{if } x_0 \in A, \\ 0 & \text{else} \end{cases}$ is the Dirac–measure. It is easily seen that the estimator (4.2) is the function $\hat{f}_n(\cdot) = \frac{k(\cdot,x_0)}{\lambda + k(x_0,x_0)} \cdot \frac{1}{n} \sum_{i=1}^n f_i$. It is thus clear that the estimator $f_n(\cdot)$ is biased, and all results necessarily depend on λ.

The following consistency result is related to [16, Lemma 4.1], where it is used in a different context.

Theorem 4.3 (Cf. [16, Lemma 4.1] and [21, Proposition 5.6])**.** *The estimator $\hat{\vartheta}_n$ is downwards biased and monotone in expectation for increasing sample sizes; more precisely, it holds that*

$$0 \le \mathbb{E}\,\hat{\vartheta}_n \le \mathbb{E}\,\hat{\vartheta}_{n+1} \le \vartheta^*,$$

where $\vartheta^* = \mathbb{E}\left(f - f_\lambda(X)\right)^2 + \lambda \left\|f_\lambda\right\|_k^2$ with $f_\lambda(\cdot)$ given in (3.5) *is the objective of the continuous problem* (3.1) *(see also* (3.4)*).*

Proof. It holds that

$$\begin{aligned}
\mathbb{E}\,\hat{\vartheta}_{n+1} &= \mathbb{E} \min_{\hat{f}_{n+1}(\cdot)} \frac{1}{n+1} \sum_{i=1}^{n+1} \left(f_i - \hat{f}_{n+1}(X_i)\right)^2 + \lambda \left\|\hat{f}_{n+1}\right\|_k^2 \\
&= \mathbb{E} \min_{\hat{f}_{n+1}(\cdot)} \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{1}{n} \sum_{j \ne i} \left(f_j - \hat{f}_{n+1}(X_j)\right)^2 + \lambda \left\|\hat{f}_{n+1}\right\|_k^2 \\
&\ge \mathbb{E} \frac{1}{n+1} \sum_{i=1}^{n+1} \min_{\hat{f}_i(\cdot)} \frac{1}{n} \sum_{j \ne i} \left(f_j - \hat{f}_i(X_j)\right)^2 + \lambda \left\|\hat{f}_i\right\|_k^2 \\
&= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\,\hat{\vartheta}_n = \mathbb{E}\,\hat{\vartheta}_n.
\end{aligned}$$

Further, the optimal value of (3.1) is given by f_λ (cf. (3.5) in Theorem 3.2).

Finally, we have that

$$\min_{\hat{f}_n(\cdot)} \frac{1}{n} \sum_{i=1}^n \left(f_i - \hat{f}_n(X_i)\right)^2 + \lambda \left\|\hat{f}_n\right\|_k^2 \le \frac{1}{n} \sum_{i=1}^n \left(f_i - f_\lambda(X_i)\right)^2 + \lambda \left\|f_\lambda\right\|_k^2.$$

By taking expectations and the infimum afterwards we conclude that $\mathbb{E}\,\hat{\vartheta}_n \le \vartheta^*$, the remaining inequality. □

5. Approximation in norm

Recall that the optimal solution of the continuous problem (3.1) is the function $f_\lambda(\cdot) \in \mathcal{H}_k$, while the optimal solution of the discrete analogue (4.1) is the random variable (4.2). In what follows we shall establish convergence of $\hat{f}_n(\cdot)$ towards $f_\lambda(\cdot)$ for increasing sample size n.

To establish convergence with respect to the norm we separate (3.4) into two subproblems. The first problem addresses deterministic function approximation whereas the second determines the effect of the model noise. We then relate the approximation problem with an auxiliary problem involving an auxiliary estimator $\tilde{f}_n$. Its residual then reconnects the estimator with the initial estimator $\hat{f}_n$. Finally, in the last subsection, we demonstrate that the noise included in the estimator vanishes if the regularization series is chosen properly.

5.1. **Problem decomposition.** The kernel estimator $\hat{f}_n$ descends from observations f_i which are generally contaminated by noise. This entails difficulties in analyzing its approximation behavior, as the regression function f_0 cannot be accessed directly at the sampling points $X_1, \dots, X_n$. We resolve this issue by splitting this initial problem (3.4) into the subproblems

$$\min_{f\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^n \big(f(X_i) - f_0(X_i)\big)^2 + \lambda\, \|f\|_k^2 \quad \text{and} \quad \min_{f\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^n \big(f(X_i) - \epsilon_i\big)^2 + \lambda\, \|f\|_k^2, \tag{5.1}$$

where

$$\epsilon_i := f_i - f_0(X_i), \qquad i = 1, \dots, n,$$

resembles the noise. Their solutions f_n^*, f_n^ϵ are again of the shape $\frac{1}{n}\sum_{i=1}^n w_i k(\cdot, X_i)$ with the weights

$$w^* = \left(\lambda + \frac{1}{n}K\right)^{-1} f_0 \text{ and } w^\epsilon = \left(\lambda + \frac{1}{n}K\right)^{-1}\epsilon, \tag{5.2}$$

respectively.

The next lemma justifies this separation from a perspective of approximation. It relates the expected approximation error of the initial estimator $\hat{f}_n$ to the approximation error of f_n^* and the general error due to the noise.

Lemma 5.1. *The expected approximation error for $f_0 \in \mathcal{H}_k$ is*

$$\mathbb{E}\left\|\hat{f}_n - f_0\right\|_k^2 = \mathbb{E}\,\|f_n^* - f_0\|_k^2 + \|f_n^\epsilon\|_k^2 \tag{5.3}$$

for the estimators $\hat{f}_n$ and f_n^ϵ with weights as in (5.2).

Proof. From $\hat{w} = w^* + w^\epsilon$ we have the norm decomposition

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \hat{w}_i k(\cdot, X_i) - f_0\right\|_k^2 = \mathbb{E}\left\|\hat{f}_n - f_0\right\|_k^2 + \|f_n^\epsilon\|_k^2 + 2\big\langle \hat{f}_n \,|\, f_n^\epsilon\big\rangle_k + 2\big\langle f_n^\epsilon \,|\, f_0\big\rangle_k. \tag{5.4}$$

Applying the tower property of the conditional expectation we get for the inner products that

$$\begin{aligned}\mathbb{E}\,\langle \hat{f}_n | f_n^\epsilon\rangle_k &= \mathbb{E}\Big\langle \frac{1}{n}\sum_{i=1}^n w_i^* k(\cdot, X_i) \,\Big|\, \frac{1}{n}\sum_{i=1}^n w_i^\epsilon k(\cdot, X_i)\Big\rangle_k = \frac{1}{n^2}\,\mathbb{E}\, w^{*\top} K w^\epsilon \\ &= \frac{1}{n^2}\,\mathbb{E}\, w^{*\top} K\Big(\lambda + \frac{1}{n}K\Big)^{-1}\mathbb{E}\,\big(f - f_0\,|X_1, \dots, X_n\big) = 0\end{aligned}$$

and

$$\mathbb{E}\left\langle f_n^\epsilon \middle| f_0\right\rangle_k = \mathbb{E}\left\langle \frac{1}{n}\sum_{i=1}^n w_i^\epsilon k(\cdot, X_i) \,\middle|\, f_0\right\rangle_k = \mathbb{E}\,\frac{1}{n}\sum_{i=1}^n w_i^\epsilon f_0(X_i)$$
$$= \mathbb{E}\,\frac{1}{n}\sum_{i=1}^n f_0(X_i)\,\mathbb{E}\left(w_i^\epsilon \middle| X_1,\dots,X_n\right) = 0$$

from the reproducing property. This is the assertion. □

Remark 5.2. The function f_n^* equals the ordinary kernel estimator $\hat{f}_n$ if $f_i = f_0(X_i)$ for all $i = 1,\dots,n$, or, put in different words, if the model is free of noise. If $f_0 = 0$ we have $f_n^\epsilon = \hat{f}_n$, independently of the noise involved in the model.

5.2. **Uniform approximation properties of kernel estimators towards f_λ.** In this section we study the approximation quality of different kernel estimators with respect to f_λ. In particular, we investigate the behavior of the norm

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n w_i\, k(\cdot, X_i) - f_\lambda(\cdot)\right\|_k^2$$

for differently chosen weights w_i. First we consider the weights $\tilde{w}_i = w_\lambda(X_i)$ with the weight function $w_\lambda \in L^2$ as in (3.7). The corresponding estimator $\tilde{f}_n(\cdot) \coloneqq \frac{1}{n}\sum_{i=1}^n \tilde{w}_i k(\cdot, X_i)$ is unbiased, i.e.,

$$\mathbb{E}\,\tilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\,\tilde{w}_i\, k(x, X_i) = \mathbb{E}\, w_\lambda(X_i)\, k(x, X_i) = \big(L_k w_\lambda\big)(x) = f_\lambda(x) \tag{5.5}$$

for every $x \in \mathcal{X}$. The next theorem reveals the *precise* approximation quality of this estimator.

Theorem 5.3 (Approximation in norm)**.** *It holds that*

$$\mathbb{E}\,\|f_\lambda - \tilde{f}_n\|_k^2 = \frac{1}{n}\,C_\lambda - \frac{1}{n}\,\|f_\lambda\|_k^2\,, \tag{5.6}$$

with

$$C_\lambda = \int_{\mathcal{X}} w_\lambda(x)^2\, k(x,x) P(dx) \tag{5.7}$$

for every $\lambda > 0$.

Proof. With $f_\lambda = L_k w_\lambda$ (cf. (3.8)) we have that

$$\mathbb{E}\left\|f_\lambda(\cdot) - \frac{1}{n}\sum_{j=1}^n k(\cdot, X_j)\,\tilde{w}_j\right\|_k^2 = \mathbb{E}\,\|f_\lambda\|_k^2 - 2\Big\langle f_\lambda \Big| \frac{1}{n}\sum_{j=1}^n k(\cdot, X_j)\,\tilde{w}_j\Big\rangle_k$$
$$+ \left\|\frac{1}{n}\sum_{j=1}^n k(\cdot, X_j)\,\tilde{w}_j\right\|_k^2.$$

With (5.5) and (2.9), the second term is

$$\mathbb{E}\,\frac{2}{n}\sum_{i=1}^n \int_{\mathcal{X}} w_\lambda(y)\, k(y, X_j)\, \tilde{w}_j P(dy) = 2\int_{\mathcal{X}} w_\lambda(y)\, f_\lambda(y)\, P(dy) = 2\,\|f_\lambda\|_k^2\,.$$

For the remaining term involving all combinations and by separating all combinations with $j = i$ from those with $j \neq i$ we find

$$\begin{aligned}&\frac{1}{n^2}\sum_{i,j=1}^{n} \mathbb{E}\,\tilde{w}_i\, k(X_i,X_j)\,\tilde{w}_j\\ =&\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\,w_\lambda^2(X_i)\,k(X_i,X_i)+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\,\mathbb{E}\,\mathbb{E}\left(\tilde{w}_i\sum_{j\neq i}k(X_i,X_j)\,\tilde{w}_j\middle| X_i\right)\\ =&\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\,w_\lambda^2(X_i)\,k(X_i,X_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\,\tilde{w}_i f_\lambda(X_i)\\ =&\frac{1}{n}\int_X w_\lambda^2(x)k(x,x)P(dx)+\frac{n-1}{n}\left\|f_\lambda\right\|_k^2\end{aligned}$$

by (2.9).

Collecting terms we find that

$$\begin{aligned}\mathbb{E}\left\|f_\lambda(\cdot)-\frac{1}{n}\sum_{j=1}^{n}k(\cdot,X_j)\,\tilde{w}_j\right\|_k^2 &= \left\|f_\lambda\right\|_k^2-2\left\|f_\lambda\right\|_k^2+\frac{n-1}{n}\left\|f_\lambda\right\|_k^2\\ &\quad+\frac{1}{n}\int_X w_\lambda^2(x)k(x,x)P(dx)\end{aligned}$$

and thus the assertion. □

Remark 5.4. The quality of the approximation in (5.6) depends on C_λ and therefore implicitly on the regularization parameter λ. To elaborate this dependence more clearly note that

$$C_\lambda=\lambda^{-2}\int_{\mathcal{X}}\big(f_\lambda(x)-f_0(x)\big)^2 k(x,x)P(dx)$$

by (3.9). The quantity C_λ grows, in the worst case, with rate λ^{-2}. This growth is, however, usually dampened by the latter integral term as f_λ gets a more accurate estimate of f_0 for decreasing λ.

A special situation occurs for $f_0 = L_k\, w$. Then C_λ is uniformly bounded, more precisely, from (3.10) we have the estimate

$$C_\lambda\leq\left\|k\right\|_2^2\left\|w_\lambda\right\|_2^2=\left\|k\right\|_2^2\sum_{\ell=1}^{\infty}\frac{\sigma_\ell^2 w_\ell^2}{(\lambda+\sigma_\ell)^2}\leq\left\|k\right\|_2^2\sum_{\ell=1}^{\infty}w_\ell^2=\left\|k\right\|_2^2\left\|w\right\|_2^2,$$

where the right-hand side is independent of λ.

Now we set our focus on the regression weights in (5.2) and the associate estimator f_n^*. Unlike in the considerations above, we do *not* prove the approximation properties f_n^* directly. We make use of its relationship with $\tilde{f}_n$ as well as the accompanying convergence properties. They are connected explicitly in the following way.

Lemma 5.5. *It holds that*[5]

$$f_n^*(\cdot) - \tilde f_n(\cdot) = \frac{1}{n}\sum_{j=1}^n w_j\, k(\cdot, X_j)$$

with the weights

$$w = \Big(\lambda + \frac{1}{n}K\Big)^{-1}\tilde r \tag{5.8}$$

where $\tilde r$ is the residual vector $\tilde r_i = f_\lambda(X_i) - \frac{1}{n}\sum_{j=1}^n \tilde w_i k(X_i, X_j)$.

Proof. By (3.9) and (4.3) we observe

$$\begin{aligned}\Big(\lambda + \frac{1}{n}K\Big)(w^* - \tilde w) &= f_0 - \lambda\tilde w - \frac{1}{n}\sum_{j=1}^n k(X_i, X_j)\,\tilde w_j\\ &= f_0 - (f_0 - f_\lambda) - \frac{1}{n}\sum_{j=1}^n k(X_i, X_j)\,\tilde w_j = \tilde r\end{aligned}$$

and thus

$$w^* - \tilde w = \Big(\lambda + \frac{1}{n}K\Big)^{-1}\tilde r.$$

Now recall that $f_n^*(\cdot) - \tilde f_n(\cdot) = \frac{1}{n}\sum_{j=1}^n (w_j^* - \tilde w_j)\, k(\cdot, X_j)$ to accept the assertion. □

Theorem 5.6. *It holds that*

$$\mathbb{E}\,\|f_n^* - \tilde f_n\|_k^2 + \frac{\lambda}{n}\,\mathbb{E}\sum_{i=1}^n (\tilde w_i - w_i^*)^2 \le \frac{1}{n}C_\lambda - \frac{1}{n}\,\|f_\lambda\|_k^2 \tag{5.9}$$

for C_λ as in (5.7)*.*

Proof. From (5.8) we have that

$$\tilde w - w^* = \left(\lambda + \frac{1}{n}K\right)^{-1}\tilde r$$

with the residual vector $\tilde r$ such that $\tilde r_i = f_\lambda(X_i) - \frac{1}{n}\sum_{j=1}^n \tilde w_i k(X_i, X_j)$. Defining the residual function $\tilde r(\cdot) := f_\lambda(\cdot) - \frac{1}{n}\sum_{j=1}^n \tilde w_i k(\cdot, X_j) \in \mathcal{H}_k$ we see that the weight vector $\tilde w - w^*$ is the solution of the related regression problem

$$\hat\vartheta_n = \min_{f\in\mathcal{H}_k}\frac{1}{n}\sum_{i=1}^n \big(f(X_i) - \tilde r(X_i)\big)^2 + \lambda\,\|f\|_k^2\,.$$

Its expected optimal value is

$$\begin{aligned}\mathbb{E}\,\hat\vartheta_n &= \frac{\lambda}{n}\,\mathbb{E}\,\tilde r^\top\Big(\lambda + \frac{1}{n}K\Big)^{-1}\tilde r = \frac{\lambda}{n}\,\mathbb{E}\,\tilde r^\top\Big(\lambda + \frac{1}{n}K\Big)^{-1}\Big(\lambda + \frac{1}{n}K\Big)\Big(\lambda + \frac{1}{n}K\Big)^{-1}\tilde r\\ &= \lambda\,\mathbb{E}\,\left\|\frac{1}{n}\sum_{i=1}^n(\tilde w_i - w_i^*)k(\cdot, X_i)\right\|_k^2 + \frac{\lambda^2}{n}\sum_{i=1}^n (\tilde w_i - w_i^*)^2\end{aligned}$$

[5] $\big(\lambda + \frac{1}{n}K\big)_j^{-1}$ is the j-row (or column, as K is symmetric) of the matrix $\big(\lambda + \frac{1}{n}K\big)^{-1}$.

which follows from the characterization in (3.17). Invoking this identity we get that

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(\tilde{w}_i - w_i^*)k(\cdot, X_i)\right\|_k^2 + \frac{\lambda}{n}\sum_{i=1}^{n}(\tilde{w} - w_i^*)^2 = \frac{1}{\lambda}\,\mathbb{E}\,\hat{\vartheta}_n$$

$$= \mathbb{E}\min_{f\in\mathcal{H}_k}\|f\|_k^2 + \frac{1}{n\lambda}\sum_{i=1}^{n}\left(f(X_i) - \tilde{r}(X_i)\right)^2 \leq \mathbb{E}\,\|\tilde{r}\|_k^2$$

by inserting the residual function $\tilde{r}$ into the objective function. Employing (5.6) we get that

$$\mathbb{E}\,\|\tilde{r}\|_k^2 = \mathbb{E}\,\|f_\lambda - \tilde{f}_n\|_k^2 = \frac{1}{n}C_\lambda - \frac{1}{n}\,\|f_\lambda\|_k^2\,,$$

which is the assertion. □

The following theorem connects all partial results of this section. It provides error estimates of the estimator of interest f_n^* with respect to f_λ as well as f_0.

Theorem 5.7. *For the estimator $f_n^*(\cdot)$ it holds that*

$$\mathbb{E}\,\|f_n^* - f_\lambda\|_k^2 \leq \frac{4C_\lambda}{n}$$

with C_λ as in (5.7). *Moreover, for $f_0 = L_k\, w_0$, it holds that*

$$\mathbb{E}\,\|f_n^* - f_0\|_k^2 \leq \frac{C_1}{n} + C_2\lambda \tag{5.10}$$

with C_1, C_2 independent of λ and n.

Proof. With (5.6) and (5.9) we get

$$\mathbb{E}\,\|f_n^* - f_\lambda\|_k^2 \leq 2\,\mathbb{E}\,\|f_n^* - \tilde{f}_n\|_k^2 + 2\,\mathbb{E}\,\|\tilde{f}_n - f_\lambda\|_k^2 \leq \frac{4C_\lambda}{n}$$

and hence the first assertion.

The second follows as

$$\begin{aligned}\mathbb{E}\,\|f_n^* - f_0\|_k^2 &\leq 2\,\mathbb{E}\,\|f_n^* - f_\lambda\|_k^2 + 2\,\mathbb{E}\,\|f_\lambda - f_0\|_k^2\\ &\leq 8\frac{C_\lambda}{n} + 2C_2\lambda\end{aligned}$$

by (3.11). □

5.3. **Denoising.** So far we have established convergence properties of the estimator $\hat{f}_n$ assuming a deterministic relationship between the data points X_i and observed values f_i. To finalize the considerations on asymptotic approximation it remains to demonstrate that the approach is robust with respect to noisy data. In other words, we need to show that the second term in (5.3), i.e.,

$$\left\|f_n^\epsilon\right\|_k^2 = \frac{1}{n}\epsilon^\top\left(\lambda + \frac{1}{n}K\right)^{-1}\frac{1}{n}K\left(\lambda + \frac{1}{n}K\right)^{-1}\epsilon, \tag{5.11}$$

vanishes if n tends to infinity.

In what follows we provide a first estimate on the norm in terms of the eigenvalues of $\frac{1}{n}K$ and the noise contained in the model.

Lemma 5.8. *Assume that* $\sup_{x\in\mathcal{X}} \operatorname{var}(f \mid X = x) =: \sigma^2_{\max} < \infty$. *It holds that*

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} w_i^\epsilon k(\cdot, X_i)\right\|_k^2 \leq \sigma^2_{\max}\ \mathbb{E}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\mu_i}{(\lambda+\mu_i)^2}, \tag{5.12}$$

where μ_i *are the eigenvalues of the matrix* $\frac{1}{n}K$.

Proof. The matrix $\left(\lambda + \frac{1}{n}K\right)^{-1}\frac{1}{n}K\left(\lambda + \frac{1}{n}K\right)^{-1}$ is symmetric hence and has a spectral decomposition

$$\frac{1}{n}\left(\lambda + \frac{1}{n}K\right)^{-1}\frac{1}{n}K\left(\lambda + \frac{1}{n}K\right)^{-1} = \frac{1}{n}V\Lambda V^\top$$

with matrices $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ and $V = [v_1, \dots, v_n]$ containing the eigenvalues and corresponding eigenvectors, respectively. Thus, we have that

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} w_i^\epsilon k(\cdot, X_i)\right\|_k^2 &= \mathbb{E}\,\frac{1}{n}\epsilon^\top\left(\lambda + \frac{1}{n}K\right)^{-1}\frac{1}{n}K\left(\lambda + \frac{1}{n}K\right)^{-1}\epsilon\\ &= \frac{1}{n}\,\mathbb{E}\sum_{i=1}^{n}\lambda_i\,\langle v_i, \epsilon\rangle^2\\ &= \frac{1}{n}\,\mathbb{E}\sum_{i=1}^{n}\lambda_i\,\mathbb{E}\left(\langle v_i, \epsilon\rangle^2\middle|\, X_1, \dots, X_n\right)\\ &= \frac{1}{n}\,\mathbb{E}\sum_{i=1}^{n}\lambda_i v_i^\top\,\mathbb{E}\left(\epsilon\epsilon^\top\middle|\, X_1, \dots, X_n\right)v_i\end{aligned}$$

as λ_i and v_i are continuous functions of the entries K and hence measurable with respect to $\sigma(X_1, \dots, X_n)$. Further, by the structure of the inner matrix we have $\lambda_i = \frac{\mu_i}{(\lambda+\mu_i)^2}$ with the eigenvalues μ_i of $\frac{1}{n}K$.

It is therefore sufficient to show that the spectral norm of the conditional covariance matrix

$$\mathbb{E}\left(\epsilon\epsilon^\top\middle|\, X_1, \dots, X_n\right)$$

is uniformly bounded by $\sigma^2_{\max}$. For that, recall that the samples $(X_i, f_i)_{i=1}^n$ are pairwise independent and therefore

$$\mathbb{E}\left(\epsilon_i\epsilon_j\middle|\, X_1, \dots, X_n\right) = \mathbb{E}\left(f_i - f_0(X_i)\right)\left(f_j - f_0(X_j)\right) = \begin{cases}\operatorname{var}\left(f\middle|\, X = X_i\right) & \text{if } i = j\\ 0 & \text{else}\end{cases} \tag{5.13}$$

as $\mathbb{E}\left(f_i - f_0(X_i)\middle|\, X_i\right) = 0$ for all $i = 1, \dots, n$. Thus, the conditional covariance matrix is a diagonal matrix with entries bounded by $\sigma^2_{\max}$. This proves the assertion. □

Lemma 5.8 above relates the norm $\|f_n^\epsilon\|_k^2$ with the trace of $\frac{1}{n}\left(\lambda + \frac{1}{n}K\right)^{-1}\frac{1}{n}K\left(\lambda + \frac{1}{n}K\right)^{-1}$ as well as the precision of the model which is expressed by $\sigma^2_{\max}$. To estimate the trace of the matrix we relate the spectrum of $\frac{1}{n}K$ with the spectrum of the integral operator L_k. More precisely, we make use of the crucial inequality

$$\mathbb{E}\sum_{i=\ell}^{n}\mu_i \leq \sum_{i=\ell}^{\infty}\sigma_i \qquad \text{for every } \ell\in\{1,\dots,n\}\,, \tag{5.14}$$

where $(\sigma_i)_{i=1}^{\infty}$ and $(\mu_i)_{i=1}^{n}$ are the eigenvalues of the operators L_k and the matrix $\frac{1}{n}K$, respectively (see [22]).

Based on this estimate we now state the main theorem of this section. We bound the norm (5.11) with respect to the regularization λ as well as the sample size n.

Theorem 5.9. *Assume the spectrum of the operator L_k decays exponentially, i.e., there are positive constants α and β such that*

$$\sigma_i \leq \alpha\, e^{-\beta\, i} \tag{5.15}$$

for all $i\in\mathbb{N}$. Then

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} w_i^{\epsilon}k(\cdot,X_i)\right\|_k^2 \leq \sigma_{\max}^2 c_1\frac{\log n}{p\, n\,\lambda} + c_2\frac{\sigma_{\max}^2}{\lambda^2\, n^{\frac{1}{p}+1}} \tag{5.16}$$

holds for all $p\geq 1$. Moreover, for $\lambda_n = {}^{c}\!/\!\sqrt{n}$ it holds that

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} w_i^{\epsilon}k(\cdot,X_i)\right\|_k^2 \leq \sigma_{\max}^2 c_1\frac{\log n}{\sqrt{n}} + c_2\frac{\sigma_{\max}^2}{\sqrt{n}} \tag{5.17}$$

with the constants $c_1=\frac{1}{4\beta c}$ and $c_2=\frac{\alpha}{2c^2(1-e^{-\beta})}$.

Remark 5.10 (Analytic kernels)**.** The conditions of the preceding theorem are satisfied in very general situations, cf. [13] and Remark 5.11 below.

Proof. Invoking (5.12) we have that

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n} w_i^{\epsilon}k(\cdot,X_i)\right\|_k^2 &\leq \sigma_{\max}^2\,\mathbb{E}\,\frac{1}{n}\sum_{i=1}^{n}\frac{\mu_i}{(\lambda+\mu_i)^2}\\
&\leq \sigma_{\max}^2\frac{1}{2n}\sum_{i=1}^{\ell}\frac{1}{\lambda} + \frac{\sigma_{\max}^2}{2n}\sum_{i=\ell+1}^{n}\frac{1}{\lambda^2}\,\mathbb{E}\,\mu_i\\
&= \sigma_{\max}^2\frac{\ell}{2n\lambda} + \frac{\sigma_{\max}^2}{2n}\sum_{i=\ell+1}^{n}\frac{1}{\lambda^2}\mu_i
\end{aligned}$$

for any fixed integer $\ell\in\{1,\dots,n\}$. Employing (5.14) and (5.15) we find for the latter term that

$$\sum_{i=\ell+1}^{n}\mu_i \leq \sum_{i=\ell+1}^{\infty}\sigma_i \leq \alpha\sum_{i=\ell+1}^{\infty}e^{-\beta i} = \alpha\, e^{-\beta(\ell+1)}\sum_{i=0}^{\infty}e^{-\beta i} = \frac{\alpha\, e^{-\beta(\ell+1)}}{1-e^{-\beta}} \tag{5.18}$$

using the sum formula of the geometric series. Setting $\ell = \left\lceil \frac{\log n}{p\beta} - 1 \right\rceil$ we get that

$$\begin{aligned}\sigma_{\max}^2 \frac{\ell}{2n\lambda} + \frac{\sigma_{\max}^2}{2\lambda^2 n}\, \mathbb{E} \sum_{i=\ell+1}^{n} \mu_i &\le \sigma_{\max}^2 \frac{\ell}{2n\lambda} + \frac{\sigma_{\max}^2}{2} \frac{\alpha e^{-\beta(\ell+1)}}{\lambda^2 n\big(1 - e^{-\beta}\big)} \\ &\le \sigma_{\max}^2 \frac{\log n}{2np\lambda} + \frac{\sigma_{\max}^2}{2} \frac{\alpha}{\lambda^2\big(1 - e^{-\beta}\big)} n^{-\frac{1}{p}-1}\end{aligned}$$

and thus the first assertion. Setting $\lambda_n = \frac{c}{\sqrt{n}}$ and $p = 2$ reveals the second assertion. □

Remark 5.11 (Analytic and universal kernels)**.** The condition (5.15) involves the operator L_k and hence depends on the kernel k as well as the underlying design measure P. Thus, verifying (5.15), requires prior knowledge on P, which is usually not available. However, [3] provides necessary and sufficient condition on k for which the spectrum of L_k decays exponentially *regardless* of the design measure. The class of kernels satisfying this condition includes, among others, the *Gaussian* kernel $k(x, y) := \exp(-\frac{1}{\sigma^2} \|x - y\|_2^2)$, which is the most popular kernel in machine learning.

For further discussions on this particular kernel, which is also a *universal* kernel, we refer to [15].

6. Convergence in norm and consistency

We can now connect the auxiliary and partial results of the preceding sections to present our main results. They identify the limit in the initial problem (2.3) and describe convergence of the estimator $\hat{f}_n$ towards f_0, as well as consistency of the estimators. We state our results for kernels with exponentially decaying spectrum.

6.1. Asymptotically optimal convergence rates and uniform approximation.

The results in the preceding section exhibit the typical bias variance problem: the parameter λ in (5.10), for example, should be small to increase the approximation quality of f_n^* for f_0; on the other side, λ should be large to reduce the noise in (5.16). The following statements reveal the best approximation rates asymptotically.

Theorem 6.1. *Assume the spectrum of L_k decays exponentially. For f_0 in the range of L_k and $\lambda_n = C \cdot n^{-1/2}$ it holds that*

$$\mathbb{E}\, \|f_0 - \hat{f}_n\|_k^2 \le \frac{C_1 + C_2 \log n}{n^{1/2}} \tag{6.1}$$

with constants C_1, C_2 independent of n (although dependent on f_0 and k).

Proof. The assertion derives from (5.3), (5.10) and (5.17) as

$$\mathbb{E}\, \|f_0 - \hat{f}_n\|_k^2 = \mathbb{E}\, \|f_0 - f_n^*\|_k^2 + \mathbb{E}\, \|f_n^N\|_k^2 \le C_1 \frac{1}{n} + C_2 \frac{1}{\sqrt{n}} + C_3 \frac{\log n}{\sqrt{n}}.$$

□

Corollary 6.2 (Uniform convergence)**.** *Given the conditions from Theorem 6.1 its holds that*

$$\mathbb{E}\, \|f_0 - \hat{f}_n\|_2^2 \le \mathcal{O}\big(n^{-1/2} \cdot \log n\big) \tag{6.2}$$

and

$$\mathbb{E}\,\|f_0-\hat f_n\|_\infty \le \mathcal{O}\big(n^{-1/4}\cdot(\log n)^{1/2}\big). \tag{6.3}$$

Proof. The first assertion is immediate with Proposition 2.6 and (6.1).

For the second observe from the reproducing property as well as the Cauchy Schwartz inequality that

$$|f_0(x)-\hat f_n(x)| = \big|\langle f_0-\hat f_n \mid k(\cdot,x)\rangle_k\big| \le C_k\big\|f_0-\hat f_n\big\|_k, \tag{6.4}$$

where $C_k=\sup_{x\in\mathcal{X}} k(x,x)$. This inequality is uniform in x and thus

$$\mathbb{E}\,\|f_0-\hat f_n\|_\infty \le C_k\,\mathbb{E}\,\|f_0-\hat f_n\|_k \le C_k\left(\frac{C_1+C_2\log n}{n^{1/2}}\right)^{1/2} \tag{6.5}$$

which is the assertion. □

6.2. **Weak consistency.** We have seen in Theorem 4.3 that the estimator $\hat\vartheta_n$ of the objective is downwards biased. However, weak consistency of the estimator $\hat\vartheta_n$ is immediate as the optimizers converge.

Theorem 6.3. *Given the conditions of Theorem 6.1 it holds that $\hat f_n$ converges to f_0 in probability. Further, for every $x\in\mathcal{X}$, $\hat f_n(x)\to f_0(x)$, as $n\to\infty$, in probability.*

Proof. Indeed, by Markov's inequality,

$$P(\|f_0-\hat f_n\|\ge\varepsilon)\le\frac{1}{\varepsilon^2}\,\mathbb{E}\,\|f_0-\hat f_n\|_k^2\to 0,$$

as $n\to\infty$ and thus the assertion is immediate. □

Theorem 6.4. *The estimators $\hat\vartheta_n$ are L^2-consistent.*

Proof. The assertion is immediate by Theorem 2.6 and the fact that $\hat f_n$ is optimal for $\hat\vartheta_n$ in (4.1). □

7. Discussion and summary

The motivational point of this paper is optimization under uncertainty, where the objective is not known precisely but has to be estimated instead. To ensure uniform convergence, we consider the norm of the associated reproducing kernel Hilbert space. The method investigates an unbiased functional estimator, which reconstructs the desired function under general preconditions. This estimator is closely related to a popular technique employed in machine learning. We provide results for convergence in the norm of the genuine space, the norm associated with the reproducing kernel Hilbert space.

The norm of the reproducing kernel Hilbert space bounds the residuals uniformly. For this reason, the results allow estimating functions and establish uniform convergence of the functional estimator. With that, the results are just appropriate for applications in stochastic optimization, a subject with many intersections with neural networks and deep learning.

The convergence rates presented here are in line with other results in nonparametric statistics. However, we do not have evidence from numerical computations that convergence rates can be improved.

Some results can be compared with the Nadaraya–Watson estimator (see [25] on kernel density estimation), which builds on kernels as well to estimator the conditional expectation. This method from nonparametric statistics has similar convergence properties and requires an oracle on the density function to find optimal convergence rates.

Finally, we want to mention that we have an implementation available at GitHub,

`https://github.com/aloispichler/reproducing-kernel-Hilbert-space`,

which allows assessing the theoretical results of the paper numerically.

Acknowledgment

We wish to thank Prof. Alexander Shapiro, Georgia Tech, and Prof. Tino Ullrich, TU Chemnitz, for discussions on draft versions of the manuscript.

References

[1] L. Ambrosio, N. Gigli and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, edition 2, Birkhäuser Verlag, Basel, 2005.

[2] F. Bach, *Sharp analysis of low-rank kernel matrix approximations*, in: Proceedings of the 26th Annual Conference on Learning Theory, S. Shalev-Shwartz, I. Steinwart (eds), vol. 30, PMLR, Princeton, NJ, 2013, pp. 185–209.

[3] M. Belkin, *Approximation beats concentration? An approximation view on inference with smooth radial kernels*, PMLR **75** (2018), 1348–1361.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, New York, 2006.

[5] A. Caponnetto and E. De Vito, *Optimal rates for the regularized least-squares algorithm*, **7** (2006), 331–368.

[6] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge University Press, Cambridge, New York, 2015.

[7] C. Dellacherie and P.-A. Meyer, *Probabilities and Potential*, North-Holland Publishing Co., Amsterdam, 1988.

[8] S. Fischer and I. Steinwart, *Sobolev norm learning rates for regularized least-squares algorithm*, **21** (2020), 1–38.

[9] L. Györfi, M. Kohler, A. Krzyżak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer New York, 2002.

[10] M. Hein and O. Bousquet, *Kernels, associated structures and generalizations*, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, vol. 127, Max-Planck-Gesellschaft, Biologische Kybernetik, 2004.

[11] O. Kallenberg, *Foundations of Modern Probability*, pringer, New York, 2002.

[12] H. König, *Eigenvalue Distribution of Compact Operators*, Birkhäuser Basel, 1986.

[13] G. Little and J. B. Reade, *Eigenvalues of analytic kernels* **15** (1984), 133–136.

[14] V. S. Mandrekar and L. Gawarecki, *Stochastic Analysis for Gaussian Random Processes and Fields*, CRC Press, 2015.

[15] C. A. Micchelli, Y. Xu and H. Zhang, *Universal kernels*, Journal of Machine Learning Research **7** (2006), 2651–2667.

[16] V. I. Norkin, G. Ch. Pflug and A. Ruszczyński, *A branch and bound method for stochastic global optimization*, Mathematical Programming **83** (1998), 425–450.

[17] M. Reed and B. Simon, *Methods of modern mathematical physics*, Academic Press, New York, 1980.

[18] L. Rüschendorf, *Mathematische Statistik*, Springer, Berlin Heidelberg, 2014.

[19] E. Schmidt, *Zur Theorie der linearen und nichtlinearen Integralgleichungen*, Mathematische Annalen **63** (1907), 433–476 (German).

[20] B. Schölkopf, R. Herbrich and A. J. Smola, *A generalized representer theorem*, Lecture Notes in Computer Science, Springer, Berlin Heidelberg, 2001, pp. 416–426.

[21] A. Shapiro, D. Dentcheva and A. Ruszczyński, *Lectures on Stochastic Programming*, MOS-SIAM Series on Optimization, edition 3, SIAM, 2014.

[22] J. Shawe-Taylor, C. K. I. Williams, N. Cristianini and J. Kandola, *On the Eigenspectrum of the Gram Matrix and the Generalization Error of Kernel-PCA*, IEEE Transactions on Information Theory **51** (2005), 2510–2522.

[23] A. N. Shiryaev, *Probability*, Springer, New York, 1996.

[24] I. Steinwart and A. Christmann, *Support Vector Machines*, Springer, New York, 2008.

[25] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer, New York, 2008.

[26] H. Wendland, *Scattered Data Approximation*, Cambridge University Press, 2004.

[27] R. Wets, *Programming under uncertainty: The complete problem*, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **4** (1966), 316–339.

[28] Y. Zhang, J. Duchi and M. Wainwright, *Divide and conquer kernel ridge regression*, PMLR **30** (2013), 592–617.

Manuscript received August 20 2022
revised September 20 2022

P. Dommel
Technische Universität Chemnitz, Faculty of mathematics. 90126 Chemnitz, Germany
E-mail address: `paul.dommel@math.tu-chemnitz.de`

A. Pichler
Technische Universität Chemnitz, Faculty of mathematics. 90126 Chemnitz, Germany
E-mail address: `alois.pichler@math.tu-chemnitz.de`