Spectrally-truncated kernel ridge regression and its free lunch

Arash A. Amini

Department of Statistics,
University of California, Los Angeles
e-mail: aaamini@ucla.edu

Abstract: Kernel ridge regression (KRR) is a well-known and popular nonparametric regression approach with many desirable properties, including minimax rate-optimality in estimating functions that belong to common reproducing kernel Hilbert spaces (RKHS). The approach, however, is computationally intensive for large data sets, due to the need to operate on a dense $n \times n$ kernel matrix, where n is the sample size. Recently, various approximation schemes for solving KRR have been considered, and some analyzed. Some approaches such as Nyström approximation and sketching have been shown to preserve the rate optimality of KRR. In this paper, we consider the simplest approximation, namely, spectrally truncating the kernel matrix to its largest $r < n$ eigenvalues. We derive an exact expression for the maximum risk of this truncated KRR, over the unit ball of the RKHS. This result can be used to study the exact trade-off between the level of spectral truncation and the regularization parameter. We show that, as long as the RKHS is infinite-dimensional, there is a threshold on r, above which, the spectrally-truncated KRR surprisingly outperforms the full KRR in terms of the minimax risk, where the minimum is taken over the regularization parameter. This strengthens the existing results on approximation schemes, by showing that not only one does not lose in terms of the rates, truncation can in fact improve the performance, for all finite samples (above the threshold). Moreover, we show that the implicit regularization achieved by spectral truncation is not a substitute for Hilbert norm regularization. Both are needed to achieve the best performance.

MSC2020 subject classifications: Primary62G08, 62G20; secondary 62C20.
Keywords and phrases: Kernel methods, ridge regression, spectral truncation, nonparametric regression, minimax estimation.

Received July 2020.

Contents

1. Introduction

The general nonparametric regression problem can be stated as

$$y_i = f^*(x_i) + w_i, \ i = 1, \ldots, n, \quad \mathbb{E}w = 0, \ \mathrm{cov}(w) = \sigma^2 I_n \tag{1.1}$$

where $w = (w_i) \in \mathbb{R}^n$ is a noise vector and $f^* : \mathcal{X} \to \mathbb{R}$ is the function of interest to be approximated from the noisy observations $\{y_i\}$. Here, $\mathcal{X}$ is the space to which the *covariates* $\{x_i\}$ belong. We consider the fixed design regression where the covariates are assumed to be deterministic. The problem has a long history in statistics and machine learning [1, 2]. In this paper, we assume that f^* belongs to a reproducing kernel Hilbert space (RKHS), denoted as $\mathcal{H}$ [3]. Such spaces are characterized by the existence of a reproducing kernel, that is, a positive semidefinite function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that uniquely determines the underlying function space $\mathcal{H}$. RKHSs are very versatile modeling tools and include, for example, Sobolev spaces of smooth functions whose norms are measures of function roughness [4]. Throughout, we think of these Sobolev spaces as the concrete examples of $\mathcal{H}$. By assuming an upper bound on the Hilbert norm of f^*, we can encode a prior belief that the true data generating function f^* has a certain degree of smoothness. Without loss of generality, we assume that f^* belongs to the unit ball of the RKHS, that is,

$$f^* \in \mathbb{B}_{\mathcal{H}} := \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1\}. \tag{1.2}$$

A natural estimator is then, the kernel ridge regression (KRR), defined as the solution of the following optimization problem:

$$\widetilde{f}_{n,\lambda} := \min_{f \in \mathcal{H}} \ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{1.3}$$

where $\lambda > 0$ is a regularization parameter. It is well-known that this problem can be reduced to a finite-dimensional problem, by an application of the so-called representer theorem [5]:

$$\min_{\omega \in \mathbb{R}^n} \ \frac{1}{n} \|y - \sqrt{n}K\omega\|^2 + \lambda \omega^T K \omega, \quad \text{where} \quad K = \frac{1}{n}\big(\mathbb{K}(x_i, x_j)\big) \in \mathbb{R}^{n \times n} \tag{1.4}$$

is the (normalized empirical) kernel matrix. Although (1.4) has a closed form solution, it involves inverting an $n \times n$ dense matrix, with time complexity $O(n^3)$, which is prohibitive in practice.

Various approximation schemes have been proposed to mitigate the computational costs, including (i) approximating the kernel matrix or (ii) directly approximating the optimization problem (1.4). Examples of the former are the Nyström approximation, column sampling and their variants [6, 7, 8, 9, 10]. An example of the latter is sketching [11, 12] where one restricts ω to the subspace $\mathrm{ran}(S) := \{S\alpha \mid \alpha \in \mathbb{R}^r\}$, for some random matrix $S \in \mathbb{R}^{n \times r}$. It is in fact known that Nyström can be considered a special case of sketching with random standard basis vectors [12]. Sketching, with sufficiently large r, has been shown in [12] to achieve minimax optimal rates over Sobolev spaces, under mild conditions on the sketching matrix S. Similarly, the Nyström approximation has been analyzed in [13, 14, 15, 11, 16] and [17], the latter showing minimax rate optimality. In addition to the above, (iii) divide and conquer approaches have been proposed [18], where one solves the problem over subsamples and then aggregates by averaging, with some rate optimality guarantees. Other notable approaches to scaling include (iv) approximating translation-invariant kernel functions via Monte Carlo averages of tensor products of randomized feature maps [19, 20] and (v) applying stochastic gradient in the function space [21]. Memory efficiency in kernel approximation is considered in [22].

In this paper, we consider the most direct kernel approximation, namely, replacing K by its best rank r approximation (in Frobenius norm). This amounts to truncating the eigenvalue decomposition of K to its top r eigenvalues. We refer to the resulting KRR approximation as the *spectrally-truncated* KRR (ST-KRR). Although somewhat slower than the Nyström approximation and fast forms of sketching, ST-KRR can be considered an ideal rank-r spectral approximation. By analyzing it, one can also gain insights about approximate SVD truncation approaches such as Nyström or sketching. Practically, ST-KRR is a very viable solution for moderate-size problems. See Appendix A for a discussion of the time complexity of various schemes.

We derive an exact expression for the maximum (empirical) mean-squared error (MSE) of ST-KRR, uniformly over the unit ball of the RKHS. This expression is solely in terms of the eigenvalues $\{\mu_i\}$ of the kernel matrix K, the regularization parameter λ, the truncation level r, and the noise level σ^2. Thus if one has access to $\{\mu_i\}$ and the noise level (or estimates of them), one can plot the exact regularization curve (maximum MSE versus λ) for a given truncation level r and sample size n, and determine the optimal value of λ. We also note that since the empirical eigenvalues $\{\mu_i\}$ quickly approach those of the integral operator associated with $\mathbb{K}$, as $n \to \infty$ [23], one can use these idealized eigenvalues instead of $\{\mu_i\}$ to get an excellent approximation of these regularization curves.

We then show that there is an optimal threshold on r, the truncation level, which we denote as r_n, such that for all $r \geq r_n$, the minimax risk of the r-truncated KRR, with the minimum taken over the regularization parameter, is strictly smaller than that of the full KRR whenever $\mu_{r+1} > 0$. For infinite-dimensional RKHSs, we always have $\mu_{r_n+1} > 0$, hence truncating at level r_n is guaranteed to strictly improve performance. The slower the decay of the eigenvalues, the larger this gap in performance.

This result shows that although the spectral truncation is mainly used as a computational device, it also has a statistical regularization effect. The next question is whether the regularization provided by the spectral truncation renders Hilbert norm regularization (via λ) unnecessary. We answer this question in the negative by showing that for any truncation level r, the optimal maximum risk is achieved for a positive λ. Together, these results show that the "r-truncated λ-regularized KRR" defines a new class of estimators whose performance cannot be achieved (in finite sample) with either regularization alone.

We also show how the exact expression for the maximum MSE can be used to easily establish a slightly weaker bound for ST-KRR, similar to those derived in [12] for sketching. We discuss the link between the statistical dimension considered in [12] and the optimal truncation level r_n, and show how the same rate-optimality guarantees hold for ST-KRR. Rate-optimality also follows form the fact that ST-KRR, with proper r, strictly dominates full KRR and the latter is rate-optional. However, we do these calculations to make the comparison easier.

Finally, we illustrate the results with some numerical simulations showing some further surprises. For example, the Gaussian kernel has a much faster eigendecay rate than a Sobolev-1 kernel (exponential versus polynomial decay). Hence, the optimal truncation level r_n asymptotically grows much slower for the Gaussian kernel. However, for finite samples, depending on the choice of the Gaussian bandwidth, the exact optimal truncation level, computed numerically, can be larger than that of Sobolev-1.

2. Preliminaries

Let us start with some observations regarding the original KRR problem in (1.3). For $\omega \in \mathbb{R}^n$, consider the *kernel mapping*

$$f_\omega := \frac{1}{\sqrt{n}} \sum_j \omega_j \mathbb{K}(\cdot, x_j). \tag{2.1}$$

Note that $\omega \mapsto f_\omega$ is a linear map from $\mathbb{R}^n \to \mathcal{H}$. This map is the link between the solutions of the two optimization problems (1.3) and (1.4): For any optimal solution ω of (1.4), f_ω will be an optimal solution of (1.3). The link is easy to establish by observing the following two identities:

$$\|f_\omega\|_{\mathcal{H}}^2 = \omega^T K \omega, \quad f_\omega(x_i) = \sqrt{n}(K\omega)_i, \tag{2.2}$$

the first of which uses the *reproducing property* of the kernel: $\langle f, \mathbb{K}(\,\cdot\,, x) \rangle_{\mathcal{H}} = f(x)$. We will frequently use this property in the sequel. The proof of the equivalence follows from an argument similar to our discussion of the identifiability below.

2.1. Identifiability

Let us first observe that f^* in (1.1) is not (statistically) identifiable. That is, there are multiple functions f^* (in fact, infinitely many if $\mathcal{H}$ is infinite-

dimensional) for which the vector (y_i) has the exact same distribution. To see this, let

$$\mathcal{L}_X := \text{span}\{\mathbb{K}(\cdot, x_i) : i \in [n]\} = \{f_\omega : \omega \in \mathbb{R}^n\}, \tag{2.3}$$

and let f_{ω^*} be the projection of f^* onto $\mathcal{L}_X$. (It is always possible to choose at least one such ω^* by the definition of projection and since $\mathcal{L}_X$ is a closed subspace of $\mathcal{H}$.) Given observations (y_i), we can only hope to recover the following equivalence class:

$$\begin{aligned}
\{f \in \mathcal{H} : f(x_i) = f^*(x_i), \ \forall i\} &= \{f \in \mathcal{H} : \langle f - f^*, \mathbb{K}(\cdot, x_i)\rangle_\mathcal{H} = 0, \ \forall i\} \\
&= \{f^* + g : g \perp_\mathcal{H} \mathcal{L}_X\} \\
&= \{f_{\omega^*} + g : g \in \mathcal{L}_X^\perp\} = f_{\omega^*} + \mathcal{L}_X^\perp
\end{aligned}$$

where the last line follows since $f^* - f_{\omega^*} \in \mathcal{L}_X^\perp$ by the property of orthogonal projection (and can be absorbed into g).

We will use f_{ω^*} as the representative of the (identifiable) equivalence class of f^*. We are interested in measuring functional deviations (e.g., the error in our estimate relative to the true function) in the *empirical ℓ_2 norm*:

$$\|f - g\|_n = \left[\frac{1}{n}\sum_{i=1}^n \big(f(x_i) - g(x_i)\big)^2\right]^{1/2}.$$

The use of this norm is common in the literature of nonparametric regression [24, 25]. It is interesting to note that $\|f^* - f_{\omega^*}\|_n = 0$,

$$\|f - f^*\|_n = \|f - f_{\omega^*}\|_n, \quad \forall f \in \mathcal{H}, \tag{2.4}$$

and $\|f_{\omega^*}\|_\mathcal{H} \leq \|f^*\|_\mathcal{H}$, since projections are contractive. Thus, recalling (1.2), f_{ω^*} also belongs to the Hilbert unit ball: $f_{\omega^*} \in \mathbb{B}_\mathcal{H}$. It is in fact easy to see that f_{ω^*} has the least Hilbert norm among the members in the equivalence class (i.e., the smoothest version). Thus, without loss of generality, we can identify f^* with f_{ω^*}. Equivalently, we can assume from the start that f^* is of the form f_{ω^*} for some $\omega^* \in \mathbb{R}^n$. Note that the "no loss of generality" statement holds as long as we are working with the empirical ℓ_2 norm, due to (2.4).

3. Main results

Let $K = UDU^T$ be the eigenvalue decomposition (EVD) of the empirical kernel matrix defined in (1.4). Here, $U \in \mathbb{R}^{n \times n}$ is an orthogonal matrix and $D = \text{diag}(\mu_i)_{i=1}^n$ where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ are the eigenvalues of K. We assume for simplicity that $\mu_n > 0$, that is, the exact kernel matrix is invertible. Consider the rank r approximation of K, obtained by keeping the top r eigenvalues and truncating the rest to zero, that is,

$$\widetilde{K} = \widetilde{K}_r := U \begin{pmatrix} D_r & 0 \\ 0 & 0 \end{pmatrix} U^T = U_r D_r U_r^T.$$

Here, $D_r = \text{diag}(\mu_1, \ldots, \mu_r)$ and $U_r \in \mathbb{R}^{n \times r}$ collects the first r columns of U. The idea is to solve (1.4) with K replaced with $\widetilde{K}$, to obtain $\widetilde{\omega}$. We then form our functional estimate $\widetilde{f}$ by using the (exact) kernel mapping (2.1).

Definition 1. *An r-truncated λ-regularized KRR estimator with input $y \in \mathbb{R}^n$, is a function $\widetilde{f} := f_{\widetilde{\omega}} = \frac{1}{\sqrt{n}} \sum_j \widetilde{\omega}_j \mathbb{K}(\cdot, x_j)$ where*

$$\widetilde{\omega} \in \operatorname*{argmin}_{\omega \in \mathbb{R}^n} \frac{1}{n} \|y - \sqrt{n}\widetilde{K}\omega\|^2 + \lambda \omega^T \widetilde{K}\omega, \tag{3.1}$$

$$\text{such that } \widetilde{K}\widetilde{\omega} = K\widetilde{\omega}. \tag{3.2}$$

A minimizer in (3.1), without the additional condition $\widetilde{K}\widetilde{\omega} = K\widetilde{\omega}$, is not unique due to the rank deficiency of $\widetilde{K}$. Thus, we can ask for it to satisfy additional constraints. The equality condition in (3.2), which can be stated as $\widetilde{\omega} \in \ker(\widetilde{K} - K)$ can always be satisfied. It is enough to choose $\widetilde{\omega}$ to be the unique minimizer in $\operatorname{ran}(\widetilde{K}) = \operatorname{ran}(U_r)$, that is, $\widetilde{\omega} = U_r \alpha$ for some $\alpha \in \mathbb{R}^r$. This is how the estimator is often implemented in practice.

We are interested in the deviation of $\widetilde{f}$ from the true function f^* in the empirical ℓ_2 norm. More precisely, we are interested in the mean-squared error as the statistical risk:

$$\text{MSE}(\widetilde{f}, f^*) = \mathbb{E}\|\widetilde{f} - f^*\|_n^2.$$

Our main result is an expression for the worst-case risk of $\widetilde{f}$ over the unit ball of the RKHS:

Theorem 1. *Let $\widetilde{f} = \widetilde{f}_{r,\lambda}$ be an r-truncated λ-regularized KRR estimator (Definition 1) applied to input y generated from model (1.1). Let*

$$H_r(\lambda) := \max_{1 \leq i \leq r} h(\lambda; \mu_i)$$

where $h(\lambda; x) = \lambda^2 x/(x + \lambda)^2$. Then, for all $r = 1, 2, \ldots, n$ and $\lambda > 0$,

$$\sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \text{MSE}(\widetilde{f}_{r,\lambda}, f^*) = \max\{H_r(\lambda), \, \mu_{r+1}\} + \frac{\sigma^2}{n} \sum_{i=1}^r \left(\frac{\mu_i}{\mu_i + \lambda}\right)^2, \tag{3.3}$$

with $\mu_{n+1} := 0$.

The first term in (3.3) is the worst-case approximation error (WAE) and the second term the estimation error (EE). The approximation error (AE) is the risk (relative to f^*) of $\bar{f}$ which is obtained by passing the noiseless observations $(f^*(x_i))$, instead of y, through the estimation procedure. The AE is the deterministic part of the risk and is given by $\|\bar{f} - f^*\|_n^2$. The estimation error is the stochastic part of the risk and is given by $\mathbb{E}\|\widetilde{f} - \bar{f}\|_n^2$.

The function $x \mapsto h(\lambda; x)$ attains its maximum of $\lambda/4$, over $[0, \infty)$, at $x = \lambda$. Thus, as long as $\lambda \in [\mu_r, \mu_1]$, the bound $H_r(\lambda) \leq \lambda/4$ is good. In general,

$$\text{WAE}_{r,\lambda} \leq \max\left\{\frac{\lambda}{4}, \, \mu_{r+1}\right\}. \tag{3.4}$$

We note that since the KRR estimates are linear in y, Theorem 1 easily gives the maximum MSE expression over the Hilbert ball of arbitrary radius R, by replacing σ^2 in (3.3) with σ^2/R^2 and multiplying the entire right-hand side by R^2.

We also have a precise result on the regularized risk of the approximating function:

Proposition 1. *Let $\bar{f} = \bar{f}_{r,\lambda}$ be obtained by passing the noiseless observations $(f^*(x_i))$, instead of y, through the estimation procedure in Definition 1. Then,*

$$\sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \|f^* - \bar{f}\|_n^2 + \lambda \|\bar{f}\|_{\mathcal{H}}^2 = \max\left\{ \max_{1 \le i \le n} \frac{\lambda \mu_i}{\mu_i + \lambda}, \, \mu_{r+1} \right\}. \tag{3.5}$$

3.1. Maximum-risk inadmissibility

Let us now consider how the maximum risk of the truncated KKR compares with the full version. For every, $\lambda > 0$, define

$$r(\lambda) := \min\{r \in [n] : \, \mu_{r+1} \le H_n(\lambda)\}.$$

In addition, recalling that $\widetilde{f}_{n,\lambda}$ is the full KRR estimator, let

$$\lambda_n := \operatorname*{argmin}_{\lambda > 0} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \mathrm{MSE}(\widetilde{f}_{n,\lambda}, f^*), \quad \text{and} \quad r_n := r(\lambda_n). \tag{3.6}$$

That is, λ_n is the regularization parameter that achieves the minimal maximum-risk for the full KRR. We have the following corollary of Theorem 1:

Corollary 1. *For every $\lambda > 0$, and every $r \in [n]$ with $r \ge r(\lambda)$,*

$$\sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \mathrm{MSE}(\widetilde{f}_{r,\lambda}, f^*) \le \sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \mathrm{MSE}(\widetilde{f}_{n,\lambda}, f^*). \tag{3.7}$$

In particular, for every $r \ge r_n$,

$$\min_{\lambda > 0} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \mathrm{MSE}(\widetilde{f}_{r,\lambda}, f^*) \le \min_{\lambda > 0} \sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \mathrm{MSE}(\widetilde{f}_{n,\lambda}, f^*). \tag{3.8}$$

Both inequalities are strict whenever $\mu_{r+1} > 0$.

Corollary 1 shows that λ-optimized $\widetilde{f}_{r_n,\lambda}$ strictly improves on optimized full KRR whenever $\mu_{r_n+1} > 0$, in a sense rendering the full KRR inadmissible, as far as the maximum risk over $\mathbb{B}_{\mathcal{H}}$ is concerned. For infinite-dimensional RKHSs, we always have $\mu_{r_n+1} > 0$, hence truncating at level r_n is guaranteed to strictly improve performance. Note that we are not claiming inadmissibility in the classical sense which requires one estimator to improve on another for all $f^* \in \mathbb{B}_{\mathcal{H}}$. In general, the slower the decay of $\{\mu_i\}$, the more significant the improvement gained by truncation. Note that (3.6) allows one to set the precise truncation level including the exact constants if one has access to the eigenvalues of the kernel matrix. In practice, for large n, the eigenvalues of the associated kernel integral operator (if available) can act as excellent surrogates for $\{\mu_i\}$ [23].

3.2. Do we need both regularizations?

Although the spectral truncation is used as a computational device, intuitively, it also has an implicit regularization effect. This is confirmed more rigorously by Corollary 1 where truncation is shown to lead to a smaller optimal worst-case MSE. The intuition is also supported by the link between the (full) KRR and Tikhonov regularization. In both cases, one forms $(K + \lambda I_n)^{-1}$ which can be considered a form of "spectral filtering". Eigenvalue truncation followed by taking the pseudo-inverse can be considered another such filtering. A common conception is that these two approaches are performing essentially the same task, hence one of them is enough to achieve the desired regularization effect. More specifically, one can ask the following: Is Hilbert norm regularization, or λ-regularization, really needed in the presence of spectral truncation? Theorem 1 allows us to settle this question. For a given truncation level r, let

$$\lambda_r := \underset{\lambda > 0}{\operatorname{argmin}} \; \sup_{f^* \in \mathbb{B}_\mathcal{H}} \; \mathrm{MSE}(\widetilde{f}_{r,\lambda}, f^*) \tag{3.9}$$

be the optimal threshold for the r-truncated λ-regularized KRR estimator.

Corollary 2. *For every $r < n$, we have*

$$\lambda_r \geq \max \Big\{ \frac{\mu_r}{\sqrt{\mu_r/\mu_{r+1}} - 1}, \frac{\sigma^2}{n} \big(1 + B_r \big) \Big\}$$

where $B_r := \min_j \sum_{i:\, i>j} (\mu_j/\mu_i) + \sum_{i:\, i<j} (\mu_i^2/\mu_j^2)$ with i and j running in $\{1, \ldots, r\}$.

Corollary 2 shows that for any truncation level r, the optimal choice of λ is always positive, hence λ-regularization further improves the performance. The effect is more pronounced when μ_r is close to μ_{r+1} or, in general, when the spectrum decays slowly (hence $\mu_i \approx \mu_j$ for most $i, j \in [r]$). The effect is also more significant for higher effective noise levels σ^2/n.

3.3. Gaussian complexity and rates

Less precise bounds, albeit good enough to capture the correct asymptotic rate as $n \to \infty$, can be obtained in terms of the Gaussian complexity of the unit ball of the RKHS. These types of results have been obtained for the Sketched-KRR. To make a comparison easier, let us show how such bounds can be obtained from Theorem 1.

Let us define the r-truncated complexity (of the empirical Hilbert ball) as

$$R_r(\delta) = \Big(\frac{\sigma^2}{n} \sum_{i=1}^{r} \min\{\mu_i, \delta^2\} \Big)^{1/2}. \tag{3.10}$$

For the case $r = n$, this matches the definition of the kernel complexity in [12], which we refer to for the related background. In particular, (3.10) is a tight upper bound on the Gaussian complexity of the intersection of $\mathbb{B}_\mathcal{H}$ and $\{f : \|f\|_n \leq \delta\}$ [25, Chapter 13]. We have:

Corollary 3 (Looser bound). *Under the setup of Theorem 1, for $\lambda \geq \max\{\delta^2, 4\mu_{r+1}\}$,*

$$\sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \text{MSE}(\widetilde{f}_{r,\lambda}, f^*) \ \leq \ \frac{1}{4}\lambda + \Big(\frac{R_r(\delta)}{\delta}\Big)^2. \tag{3.11}$$

If $\lambda \geq \mu_1$, one can replace the first term with $\mu_1\lambda^2/(\lambda+\mu_1)^2$ for a better bound.

Choosing $\lambda = \delta^2 \geq 4\mu_{r+1}$, we obtain

$$\text{RMSE} = \sqrt{\text{MSE}} \leq \frac{\delta}{2} + \frac{R_r(\delta)}{\delta} \leq \frac{\delta}{2} + \frac{R_n(\delta)}{\delta}.$$

The latter upper bound is what one would get for the full KRR. Matching the two terms in that bound, we chooses δ_n such that $\delta_n^2 = 2R_n(\delta_n)$ which gives the well-known critical radius for the KRR problem [25]. It is known that δ_n gives the optimal rate of convergence for estimating functions in $\mathbb{B}_{\mathcal{H}}$, i.e., its rate of decay matches that of the minimax risk [12]. The above argument shows that as long as r is taken large enough so that $4\mu_{r+1} \leq \delta_n^2$, the r-truncated KRR achieves (at least) the same rate as the full KRR. For the sketching, the same conclusion is established in [12], where the smallest r satisfying $\mu_r \leq \delta_n^2$ is referred to as the *statistical dimension* of the kernel.

For Sobolev-α kernels, with eigendecay $\mu_i \asymp i^{-2\alpha}$, we obtain MSE $\lesssim \delta_n^2 \asymp (\sigma^2/n)^{-\frac{2\alpha}{2\alpha+1}}$. Interestingly, in this case, the estimate based on the weaker bound (3.11) and the exact bound (3.3) give the same rate (cf. Appendix C). This is expected since the given rate is known to be minimax optimal for Sobolev spaces. The same goes for the Gaussian kernel for which $\mu_j \asymp e^{-cj \log j}$ and the rate is $\gamma \log(1/\gamma)$ for $\gamma = \sigma^2/n$.

Order-wise, δ_n^2 will be the same as λ_n defined in (3.6), that is $\lambda_n \asymp \delta_n^2$, whenever δ_n^2 matches the optimal rate. Hence, often $\mu_1 > \lambda_n \gg \mu_n$ for large n and the argument leading to (3.4) suggests that in this case $H_n(\lambda_n) \approx \lambda_n/4$. Then, $r_n \approx \min\{r \in [n] : \mu_{r+1} \leq \frac{\lambda_n}{4}\}$.

For Sobolev-α kernels, this suggests truncation level $r_n \gtrsim (\sigma^2/n)^{\frac{1}{2\alpha+1}}$ which gives moderate savings for high smoothness levels α. Similarly, for the Gaussian kernel, it is not hard to see that truncating to $r_n \gtrsim \log(n/\sigma^2)$ is enough to get the same rate as the full KRR, which is a substantial saving.

4. Simulations

We now present some numerical experiments to corroborate the theory. We consider a Gaussian kernel $K(s,t) = e^{-(u-v)^2/2b^2}$ of bandwidth $b = 0.1$ on $[-1, 1]$, as well as the Sobolev-1 kernel $K(s,t) = \min(s,t)$ on $[0, 1]$. We take the covariates $\{x_i\}$ to be $n = 200$ equi-spaced points in each interval. The top row of Fig. 1 shows the plot of the theoretical maximum MSE as given by Theorem 1 for the two kernels, for both the full KRR $(r = n)$, and the optimally truncated version $(r = r_n)$. We have used $\sigma = 2$ in (3.3). As predicted by Theorem 1, the minimum achievable maximum MSE is smaller for the truncated KRR.
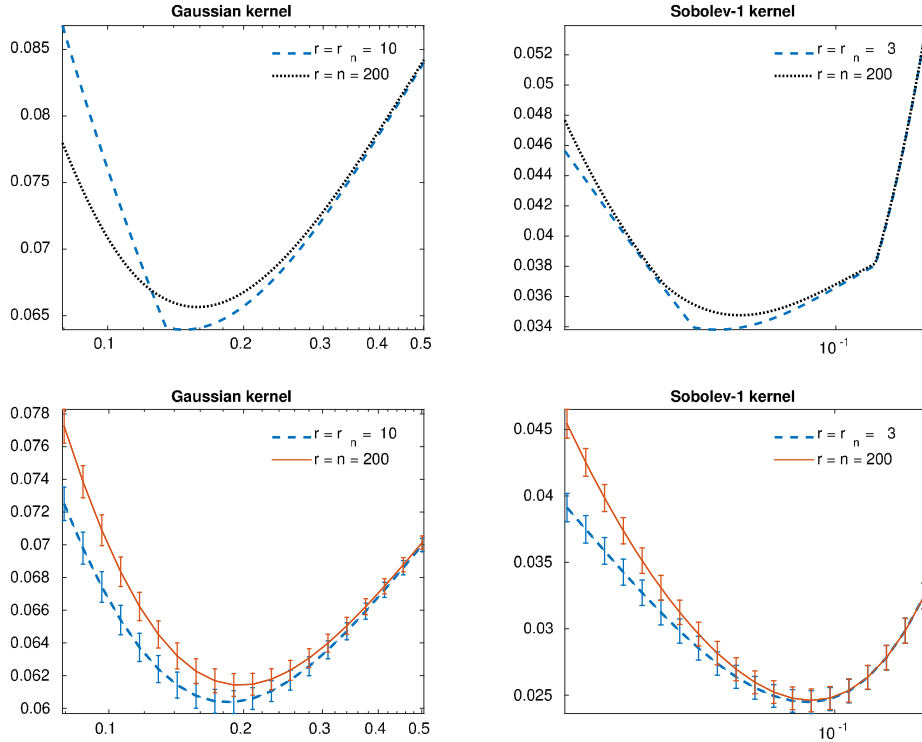
FIG 1. *Plots of (top) the maximum theoretical MSE (bottom) typical empirical MSE, versus the regularization parameter (λ) for the Gaussian (with bandwidth = 10) and Sobolev-1 kernels, on $[-1, 1]$ and $[0, 1]$ respectively, with $n = 200$ equispaced samples. The optimally-truncated KRR is shown ($r = r_n$) together with the full KRR ($r = n$).*

To compute the optimal truncation, we have evaluated the regularization curve of the full KRR first, obtained the minimizer λ_n and then used (3.6) to compute the optimal truncation level r_n. For the setup of the simulation, we get $r_n = 10$ for the Gaussian and $r_n = 3$ for the Sobolev-1. It is interesting to note that although in terms of rates, r_n for the Gaussian should be asymptotically much smaller than that of Sobolev-1, in finite samples, the truncation level for the Gaussian could be bigger as can be seen here. This is due to the unspecified, potentially large, constants in the rates (that depend on the bandwidth b as well). Also, notice how surprisingly small r_n is relative to n in both cases.

The bottom row of Fig. 1 shows the empirical MSE obtained for a typical random $f^* \in \mathbb{B}_{\mathcal{H}}$, by computing the KRR estimates for observation y and comparing with f^*. The random true function is generated as $f^* = f_{\omega^*}$ where $\omega^* \sim N(0, I_n)$ and further normalized so that $(\omega^*)^T K \omega^* = 1$. We have generated $n = 200$ observations from (1.1) with $\sigma = 2$. The plots were obtained using 1000 replications. The truncation levels are those calculated based on the maximum MSE formula (3.3). The plots show that for a typical application, the truncated
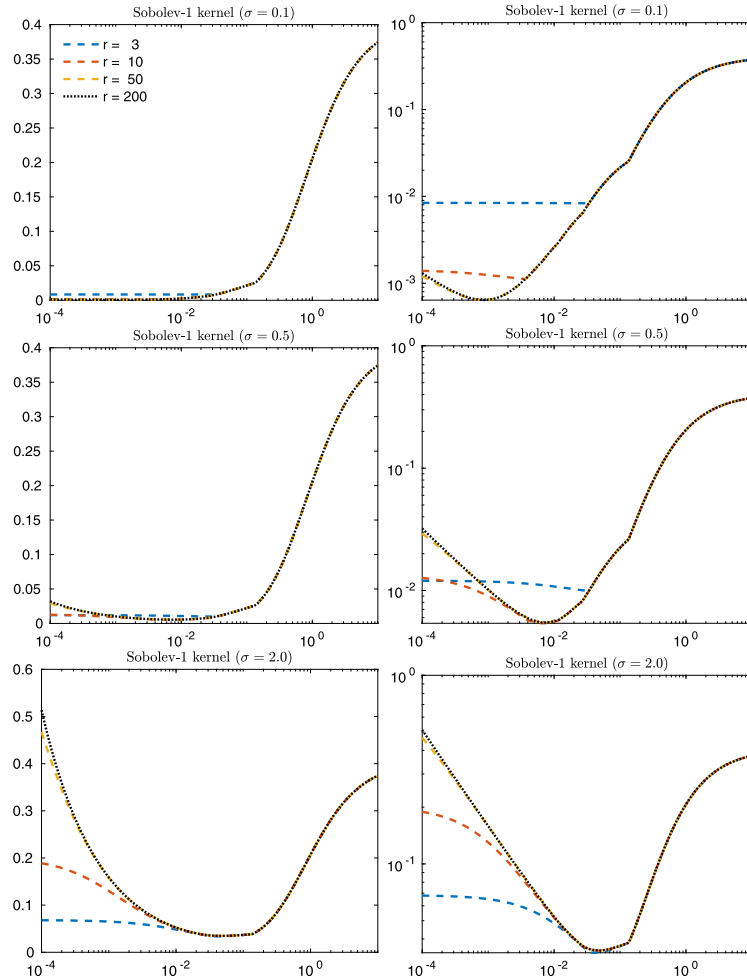
FIG 2. *The effect of the noise level on the regularization curve of the truncated KRR ($n =$ 200). The plots show the maximum theoretical MSE versus the regularization parameter (λ) for Sobolev-1 kernel on $[0, 1]$. Left panel plots have linear scale for the y-axis. The right panel shows the same plots on the* log-*scale for y-axis.*

KRR also dominates the full KRR.

Fig. 2 shows the effect of the noise level on the performance of the truncated KRR for the Sobolev-1 kernel. Each plot shows the regularization curve (maximum MSE versus λ) for various truncation levels $r = 3, 10, 50$ and the full KRR corresponding to $r = n = 200$. Generally, the higher the noise level, the smaller the truncation level needed to capture the minimum of the full KRR curve. Note that when the noise level is very small, e.g. $\sigma = 0.1$, the regularization curve on the linear scale looks as if it has its minimum achieved at $\lambda = 0$. This is what practitioners also report in applications. However, looking at the same plot on

the log-scale, one observes that the minimum is achieved at some positive value of λ, corroborating Corollary 2. The difference in MSE between the optimal λ and $\lambda = 0$ is, however, small in these low noise settings and might not be noticeable in practice. That both rank and norm regularizations are needed is quite easily observed at higher noise levels.

5. Proof of the main result

Here we give the proof of Theorem 1 and Corollaries 1 and 2. The remaining proofs can be found in Appendix B.

From the discussion in Section 2.1, both the KRR estimate and the true function belong to $\mathcal{L}_X$ given in (2.3). It is then useful to have an expression for the empirical ℓ_2 error of functions belonging to this space. First, we observe that $\|f_\omega\|_n^2 = \frac{1}{n}\sum_{i=1}^n [f_\omega(x_i)]^2 = \|K\omega\|^2$. Now, take any $\omega, \omega^* \in \mathbb{R}^d$, and let $u = K\omega$ and $u^* = K\omega^*$. Then, we have

$$\|f_\omega - f_{\omega^*}\|_n = \|f_{\omega-\omega^*}\|_n = \|K(\omega - \omega^*)\| = \|u - u^*\| \tag{5.1}$$

where the fist equality is by the linearity of $\omega \mapsto f_\omega$. For any function $f_\omega \in \mathcal{L}_X$, we call $u = K\omega$ the *u-space* representation of f_ω. Identity (5.1) shows that it is often easier to work in the u-space since the *u-transform* turns empirical ℓ_2 norms on functions into the usual ℓ_2 norms on vectors. In other words, the map $f_\omega \mapsto u$, is a Hilbert space isometry from $(\mathcal{L}_X, \|\cdot\|_n)$ to $(\mathbb{R}^n, \|\cdot\|)$. In the u-space, the KRR optimization problem can be equivalently stated as:

$$\min_{u \in \operatorname{ran}(K)} \frac{1}{n}\|y - \sqrt{n}u\|^2 + \lambda u^T K^+ u \tag{5.2}$$

where K^+ is the pseudo inverse of K, and $\operatorname{ran}(K)$ its range. More precisely:

Lemma 1. *For any $K \in \mathbb{R}^{n \times n}$, problems (1.4) and (5.2) are equivalent in the following sense:*

- *For any minimizer $\bar\omega$ of (1.4), $K\bar\omega$ is a minimizer of (5.2), and*
- *for any minimizer $\bar{u}$ of (5.2), any $\bar\omega \in \{\omega : K\omega = \bar{u}\}$ is a minimizer of (1.4).*

It is often the case that the kernel matrix itself is invertible, in which case $K^+ = K^{-1}$, $\operatorname{ran}(K) = \mathbb{R}^n$ and problem (5.2) simplifies. However, the equivalence in Lemma 1 holds even if we replace K with an approximation which is rank deficient. This observation will be useful in the sequel.

Proof of Theorem 1. Take $\widetilde\omega$ to be as in Definition 1 and let $\widetilde{y} = y/\sqrt{n}$. Since $\widetilde\omega$ is the minimizer of $F(\omega; y) = \|\widetilde{y} - \widetilde{K}\omega\|^2 + \lambda\omega^T\widetilde{K}\omega$, we have $\nabla F(\widetilde\omega; y) = 0$ or $\widetilde{K}(\widetilde{K}\widetilde\omega - \widetilde{y}) + \lambda\widetilde{K}\widetilde\omega = 0$. Hence, $(\widetilde{K} + \lambda I)\widetilde{K}\widetilde\omega = \widetilde{K}\widetilde{y}$ or

$$\widetilde{K}\widetilde\omega = \Psi_\lambda\widetilde{y}, \quad \text{where } \Psi_\lambda = (\widetilde{K} + \lambda I)^{-1}\widetilde{K}. \tag{5.3}$$

Let $w = (w_i) \in \mathbb{R}^n$ be the noise vector in (1.1) and $\widetilde{w} = w/\sqrt{n}$. We also let

$$u^* = (f_{\omega^*}(x_i))/\sqrt{n} = K\omega^*. \tag{5.4}$$

Then, we can write model (1.1) as $\widetilde{y} = u^* + \widetilde{w}$, where $\widetilde{w}$ is zero mean with $\mathrm{cov}(\widetilde{w}) = \sigma^2 I_n/n$. From (5.1), we have $\|\widetilde{f} - f^*\|_n^2 = \|K(\widetilde{\omega} - \omega^*)\|^2$, and

$$K(\widetilde{\omega} - \omega^*) = \widetilde{K}\widetilde{\omega} - u^* = \Psi_\lambda \widetilde{y} - u^* = (\Psi_\lambda - I)u^* + \Psi_\lambda \widetilde{w},$$

where the first equality uses assumption (3.2). It follows that

$$\mathbb{E}\|\widetilde{f} - f^*\|_n^2 = \|(I - \Psi_\lambda)u^*\|^2 + \mathbb{E}\|\Psi_\lambda \widetilde{w}\|^2.$$

where the first term is the approximation error (AE) and the second term, the estimation error (EE). Let us write $\widetilde{D} = \mathrm{diag}(\mu_1, \ldots, \mu_r, 0, \ldots, 0) \in \mathbb{R}^{n \times n}$ so that $\widetilde{K} = U\widetilde{D}U^T$. We define

$$\Gamma_\lambda = (\widetilde{D} + \lambda I)^{-1}\widetilde{D}, \quad \text{so that,} \quad \Psi_\lambda = U\Gamma_\lambda U^T \tag{5.5}$$

and note that Γ_λ is diagonal. Let $v^* = U^T u^*$ and $\widehat{w} = U^T \widetilde{w}$. Then, since ℓ_2 norm is unitarily invariant, we have

$$\mathbb{E}\|\widetilde{f} - f^*\|_n^2 = \|(I - \Gamma_\lambda)v^*\|^2 + \mathbb{E}\|\Gamma_\lambda \widehat{w}\|^2.$$

Controlling the estimation error: We have

$$\mathrm{EE}_{r,\lambda} := \mathbb{E}\|\Gamma_\lambda \widehat{w}\|^2 = \mathbb{E}\big[\widehat{w}^T \Gamma_\lambda^2 \widehat{w}\big] = \mathrm{tr}\left(\Gamma_\lambda^2 \mathrm{cov}(\widehat{w})\right) = \frac{\sigma^2}{n}\mathrm{tr}(\Gamma_\lambda^2),$$

using $\mathrm{cov}(\widehat{w}) = U^T \mathrm{cov}(\widetilde{w})U = (\sigma^2/n)U^T U = \sigma^2 I_n/n$ since U is an orthogonal matrix. Then,

$$(\Gamma_\lambda)_{ii} = \left(\frac{\widetilde{D}_{ii}}{\widetilde{D}_{ii} + \lambda}\right) = \begin{cases} \frac{\mu_i}{\mu_i + \lambda}, & i = 1, \ldots, r \\ 0 & i = r+1, \ldots, n, \end{cases} \tag{5.6}$$

establishing the EE part of the result.

Controlling the approximation error: Recall that we are interested in the worst-case approximation error (WAE) over the unit ball of the Hilbert space, i.e., over $f^* \in \mathbb{B}_\mathcal{H}$. Also, recall that without loss of generality, we can take $f^* = f_{\omega^*}$. Hence,

$$1 \geq \|f\|_\mathcal{H}^2 = \|f_{\omega^*}\|_\mathcal{H}^2 = (\omega^*)^T K\omega^* = (u^*)^T K^{-1} u^* = (v^*)^T D^{-1} v^* \tag{5.7}$$

where the second equality is from (2.2), and the latter two are by definitions of u^* and $v^* = U^T u^*$. We obtain

$$\mathrm{WAE}_{r,\lambda} = \sup_{(v^*)^T D^{-1} v^* \leq 1} \|(I - \Gamma_\lambda)v^*\|^2.$$

A further change of variable $v^* = D^{1/2}v$ gives

$$\text{WAE}_{r,\lambda} = \sup_{v^T v \leq 1} \|(I - \Gamma_\lambda)D^{1/2}v\|^2 = \|(I - \Gamma_\lambda)D^{1/2}\|^2,$$

where $\|\cdot\|$, applied to matrices, is the $\ell_2 \to \ell_2$ operator norm. Note that Γ_λ is a diagonal matrix with diagonal elements, $\mu_i/(\mu_i + \lambda)$ for $i = 1, \ldots, r$ followed by $n - r$ zeros. It follows that $(I - \Gamma_\lambda)D^{1/2}$ is diagonal with diagonal elements:

$$[(I - \Gamma_\lambda)D^{1/2}]_{ii} = \begin{cases} \frac{\lambda\sqrt{\mu_i}}{\lambda + \mu_i}, & i = 1, \ldots, r, \\ \sqrt{\mu_i}, & i = r + 1, \ldots, n. \end{cases} \tag{5.8}$$

Since $\{\mu_i\}$ is a non-increasing sequence, we obtain

$$\text{WAE}_{r,\lambda} = \max\left\{ \max_{1 \leq i \leq r} \frac{\lambda^2\mu_i}{(\lambda + \mu_i)^2}, \ \mu_{r+1} \right\},$$

which is the desired result.   □

Proof of Corollary 1. Let $\text{EE}_{r,\lambda} := \frac{\sigma^2}{n} \sum_{i=1}^r [\mu_i/(\mu_i + \lambda)]^2$ be the estimation error of $\widetilde{f}_{r,\lambda}$ as in (3.3). Note that as long as $\mu_{r+1} > 0$, we have $\text{EE}_{r,\lambda} < \text{EE}_{r+1,\lambda} \leq \text{EE}_{n,\lambda}$. It remains to show that the WAE of the truncated KRR is less than that of full KRR. We have for $r \geq r(\lambda)$,

$$\text{WAE}_{r,\lambda} = \max\{H_r(\lambda), \mu_{r+1}\} \leq \max\{H_n(\lambda), \mu_{r+1}\} = H_n(\lambda) = \text{WAE}_{n,\lambda}.$$

This proves (3.7). For the second assertion, it is enough to apply (3.7) with $\lambda = \lambda_n$, noting that in this case, the RHS will be the minimax risk of the full KRR and the LHS is further lower bounded by the minimax risk of the truncated KRR.   □

Proof of Corollary 2. Let us write $\text{WAE}_r(\lambda)$ and $E_r(\lambda)$ for the worst-case approximation and estimation errors, respectively, as a function of λ. Let $M_r(\lambda)$ be the worst-case MSE, so that $M_r(\lambda) = \text{WAE}_r(\lambda) + E_r(\lambda)$. The $\text{WAE}_r(\cdot)$ starts off with the constant branch $\text{WAE}_r(\lambda) = \mu_{r+1}$ for small values of λ. Let $h_i(\lambda) := h(\lambda; \mu_i)$. The constant branch starts at $\lambda = 0$ and extends to $\lambda = \lambda^{(1)}$ where $h_r(\lambda^{(1)}) = \mu_{r+1}$. Some algebra gives $\lambda^{(1)} = \mu_r/(\sqrt{\mu_r/\mu_{r+1}} - 1)$. For $\lambda \in [0, \lambda^{(1)}]$, we have $M_r'(\lambda) = E_r'(\lambda) < 0$ showing that the minimizer of M_r is $\geq \lambda^{(1)}$.

The next branch of WAE starts at $\lambda^{(1)}$ and ends at $\lambda^{(2)}$ which solves $h_r(\lambda^{(2)}) = h_{r-1}(\lambda^{(2)})$. The knots $\lambda^{(i)}$ determining subsequent branches are determined similarly: $h_{r-i+2}(\lambda^{(i)}) = h_{r-i+1}(\lambda^{(i)})$ for $i = 2, 3, \ldots, r$ and $\lambda^{(r+1)} = \infty$. We have $\text{WAE}_r(\lambda) = h_{r-i+1}(\lambda)$ for $\lambda \in I_i := [\lambda^{(i)}, \lambda^{(i+1)})$ for $i = 1, \ldots, r$. See Fig. 3.

Fix $i_* \in [r]$ and let $j_* = r - i_* + 1$. Then for $\lambda \in \text{int}(I_{i_*})$

$$M_r'(\lambda) = h_{j_*}'(\lambda) + E_r'(\lambda) = \frac{2\mu_{j_*}^2}{(\mu_{j_*} + \lambda)^3}\left[\left(\lambda - \frac{\sigma^2}{n}\right) - \frac{\sigma^2}{n}\sum_{i \neq j_*} \frac{\mu_i^2}{(\mu_i + \lambda)^3}\frac{(\mu_{j_*} + \lambda)^3}{\mu_{j_*}^2}\right]$$
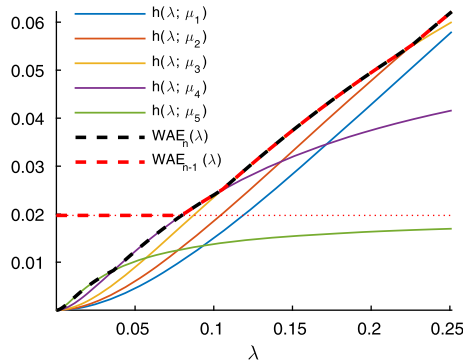
FIG 3. *An illustration of $\lambda \mapsto WAE_r(\lambda)$ for a 5×5 Gaussian kernel matrix. The case $r = n = 5$ corresponds to the full KRR and $r = n - 1 = 4$ to a truncated version.*

where i ranges over $[r] \setminus \{j^*\}$. Note that

$$\sum_{i \neq j_*} \frac{\mu_i^2}{(\mu_i + \lambda)^3} \frac{(\mu_{j^*} + \lambda)^3}{\mu_{j^*}^2} \geq \Big(\sum_{i > j_*} \frac{\mu_{j^*}}{\mu_i} + \sum_{i < j_*} \frac{\mu_i^2}{\mu_{j^*}^2} \Big) \geq B_r.$$

The first inequality is since $\lambda \mapsto (\mu_{j^*} + \lambda)/(\mu_i + \lambda)$ is increasing in $[0, \infty)$ if $\mu_{j^*} < \mu_i$, hence lower-bounded by its value at $\lambda = 0$, and is decreasing on $[0, \infty)$ if $\mu_{j^*} > \mu_i$, hence lower-bounded by its value as $\lambda \to \infty$. Then,

$$M_r'(\lambda) \leq \frac{2\mu_{j^*}^2}{(\mu_{j^*} + \lambda)^3} \Big[\Big(\lambda - \frac{\sigma^2}{n} \Big) - \frac{\sigma^2}{n} B_r \Big) \Big].$$

It follows that $M_r'(\lambda) < 0$ as long as $\lambda < \sigma^2(1 + B_r)/n$ no matter which interval I_i contains λ. This shows that the minimizer of M_r has to be $\geq \sigma^2(1 + B_r)/n$ completing the proof. □

Acknowledgments

Appendix A: Time complexity comparison

The ST-KRR and approximate versions, such as Nyström and sketching, all have time complexity of $O(nr^2 + r^3) = O(nr^2)$ for computing the r-truncated KRR estimate, once the pieces required for approximating the kernel matrix (e.g., KS and $S^T KS$ in the case of sketching, U_r and D_r in the case of ST-KRR and so on) are computed. Computing these pieces is where these methods differ. For sketching, this step could have complexity as large as $O(n^2r^2)$ for

dense sketches, $O(n^2 \log r)$ for randomized Fourier and Hadamard sketches, to as low as $O(nr)$ for the Nyström.

For the ST-KRR, this step involves computing the top-r eigenpairs of the symmetric matrix K, for which the Lanczos algorithm is the standard and for which a complexity analysis is hard to find in the literature. However, results of [26] suggest that it has average-case complexity $O(n^2(r + \log n))$. More precisely, [26] show that on average $k = O(\log n/\sqrt{\varepsilon})$ Lanczos iterations are enough to compute the top eigenvalue to within relative error ε, hence an overall average-case complexity $O(kN)$ where N is the number of nonzero entries of matrix K.

Appendix B: Remaining proofs

Proof of Proposition 1. We will use the same notation as in the proof of Theorem 1. By the same argument as in that proof, we have $\|\bar{f} - f^*\|_n^2 = \|(I - \Gamma_\lambda)v^*\|^2$ where Γ_λ is defined in (5.5) and $v^* = U^T u^*$ for u^* given in (5.4). Let $\bar{\omega}$ be the solution of (3.1) for the input $(f^*(x_i))$ (instead of y) so that $\bar{f} = f_{\bar{\omega}}$. Using the optimality condition in the proof of Theorem 1,

$$\bar{u} := K\bar{\omega} = \widetilde{K}\bar{\omega} = \Psi_\lambda u^*$$

where we have used (3.2) and (5.3), with $\widetilde{y} = u^*$ (i.e., $\widetilde{w} = 0$). We can write

$$\|\bar{f}\|_{\mathcal{H}}^2 = \bar{\omega}^T K\bar{\omega} = (\bar{u})^T K^{-1}\bar{u} = (u^*)^T \Psi_\lambda K^{-1}\Psi_\lambda u^* = (v^*)^T \Gamma_\lambda D^{-1}\Gamma_\lambda v^*$$

using $\Psi_\lambda = U\Gamma_\lambda U^T$, $K = UDU^T$ and $v^* = U^T u^*$. Recall from (5.7) that $f^* \in \mathbb{B}_{\mathcal{H}}$ is equivalent to $(v^*)^T D^{-1} v^* \leq 1$. It follows that

$$
\begin{aligned}
&\sup_{f^* \in \mathbb{B}_{\mathcal{H}}} \left(\|f^* - \bar{f}\|_n^2 + \lambda\|\bar{f}\|_{\mathcal{H}}^2 \right) \\
&= \sup_{(v^*)^T D^{-1} v^* \leq 1} \left[\|(I - \Gamma_\lambda)v^*\|^2 + \lambda(v^*)^T \Gamma_\lambda D^{-1}\Gamma_\lambda v^* \right] \\
&= \sup_{(v^*)^T D^{-1} v^* \leq 1} (v^*)^T \left[(I - \Gamma_\lambda)^2 + \lambda\Gamma_\lambda D^{-1}\Gamma_\lambda \right] v^* \\
&= \|D^{1/2}[(I - \Gamma_\lambda)^2 + \lambda\Gamma_\lambda D^{-1}\Gamma_\lambda]D^{1/2}\| \\
&= \|(I - \Gamma_\lambda)^2 D + \lambda\Gamma_\lambda^2\|
\end{aligned}
$$

where the third equality is using the change of variable $v^* = D^{1/2}v$ as in the proof of Theorem 1, and the last line follows since all the matrices are diagonal and hence commute. The result now follows by combining (5.6) and (5.8), after some algebra. □

Proof of Corollary 3. For any $a, b > 0$, we have $\frac{1}{2}(a \wedge b) \leq (a^{-1} + b^{-1})^{-1} \leq a \wedge b$, where $a \wedge b := \min\{a, b\}$. Hence, the estimation error in (3.3) is bounded as

$$\mathrm{EE}_{r,\lambda} = \frac{\sigma^2}{n\lambda^2} \sum_{i=1}^{r} \frac{\mu_i^2\lambda^2}{(\mu_i + \lambda)^2} = \frac{\sigma^2}{n\lambda^2} \sum_{i=1}^{r} (\mu_i^{-1} + \lambda^{-1})^{-2} \leq \frac{\sigma^2}{n\lambda^2} \sum_{i=1}^{r} (\mu_i \wedge \lambda)^2.$$

This upper bound is within a factor of 4 of the estimation error. Using $\mu_i/(\lambda + \mu_i) \leq 1$ to shave off the power by one, we obtain the weaker bound:

$$\mathrm{EE}_{r,\lambda} = \frac{\sigma^2}{n\lambda} \sum_{i=1}^{r} \frac{\mu_i^2 \lambda}{(\mu_i + \lambda)^2} \leq \frac{\sigma^2}{n\lambda} \sum_{i=1}^{r} \frac{\mu_i \lambda}{\mu_i + \lambda} \leq \frac{\sigma^2}{n\lambda} \sum_{i=1}^{r} \mu_i \wedge \lambda. \tag{B.1}$$

Recalling definition (3.10), we conclude that if $\lambda \geq \delta^2$,

$$\mathrm{EE}_{r,\lambda} \leq \frac{\sigma^2}{n\delta^2} \sum_{i=1}^{r} \min\{\mu_i, \delta^2\} = \Big(\frac{R_r(\delta)}{\delta}\Big)^2.$$

Combining with the WAE bound (3.4), we obtain the desired result. $\qquad\square$

Proof of Lemma 1. Let $F(\omega)$ and $G(u)$ be the objective functions in (1.4) and (5.2), respectively. We have $F(\omega) = G(K\omega)$ for any $\omega \in \mathbb{R}^n$, which follows from the identity $KK^+K = K$. Now, assume that $\bar{\omega}$ is a minimizer of F, and let $\bar{u} := K\bar{\omega}$. Pick any $u \in \mathrm{ran}(K)$; there exists ω such that $u = K\omega$, and we have $G(\bar{u}) = F(\bar{\omega}) \leq F(\omega) = G(u)$. The other direction follows similarly. $\quad\square$

Appendix C: Rate calculations

Here we compute the error rate predicted by the strong and weak bounds and show that they are the same. Let $\gamma := \sigma^2/n$. Assume the polynomial eigendecay of the Sobolev-α kernel, i.e., $\mu_i \asymp i^{-2\alpha}$. Taking k to be the smallest integer satisfying $k^{-2\alpha} \lesssim \delta^2$, we have

$$R_n^2(\delta) = \gamma\Big(k\delta^2 + \sum_{i=k+1}^{n} i^{-2\alpha}\Big) \leq \gamma(k\delta^2 + k^{-2\alpha+1}) \lesssim \gamma k\delta^2$$

where the first inequality uses an integral approximation to the sum and the second uses the definition of k. Setting $\delta^2 \asymp R_n(\delta)$ we have $\delta^2 \asymp \gamma k \asymp \gamma(\delta^2)^{-\frac{1}{2\alpha}}$, hence the critical radius $\delta_n^2 \asymp \gamma^{\frac{2\alpha}{\alpha+1}}$.

Now consider the strong bound. As discussed in the text, $\mathrm{WAE}_{n,\lambda} \asymp \lambda$. Also, as the proof of Corollary 3 shows, we have

$$\mathrm{EE}_{n,\lambda} \asymp \frac{\gamma}{\lambda^2} \sum_{i=1}^{n} \min(\mu_i^2, \lambda^2).$$

Letting k be defined as the smallest integer such that $\mu_k \lesssim \lambda$, we get $k^{-2\alpha} \lesssim \lambda$ as before. Then, the maximum MSE is bounded as

$$\mathrm{MSE} \lesssim \lambda + \frac{\gamma}{\lambda^2}\Big(k\lambda^2 + \sum_{i=k+1}^{n} i^{-4\alpha}\Big) \lesssim \lambda + \frac{\gamma}{\lambda^2}\Big(k\lambda^2 + k^{-4\alpha+1}\Big).$$

Since $k^{-4\alpha+1} \lesssim k\lambda^2$, by the definition of k, we obtain $\mathrm{MSE} \lesssim \lambda + \gamma k \lesssim \lambda + \gamma\lambda^{-\frac{1}{2\alpha}}$. Equating the two terms we obtain $\mathrm{MSE} \asymp \lambda_n \asymp \gamma^{\frac{2\alpha}{2\alpha+1}}$ as before.

For the Gaussian kernel, with $\mu_j \asymp e^{-cj \log j}$, it is not hard to verify that with $e^{-ck} \lesssim \lambda$, we get $\mathrm{MSE} \lesssim \lambda + \gamma k \lesssim \lambda + \gamma \log(1/\lambda)$. Minimizing the bound over λ, we obtain $\lambda \asymp \gamma$, hence $\mathrm{MSE} \lesssim \gamma \log(1/\gamma)$.

References

[1] Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006. MR2172729

[2] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. MR2724359

[3] Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016. MR3526117

[4] Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990. MR1045442

[5] George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971. MR0290013

[6] Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

[7] Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1232–1239. ACM, 2008.

[8] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nyström method. In *Advances in Neural Information Processing Systems*, pages 1060–1068, 2009.

[9] Mu Li, James T. Kwok, and Bao-Liang Lu. Making large-scale Nyström approximation possible. In *Proceedings of the 27th International Conference on Machine Learning*, pages 631–638, 2010.

[10] Ameet Talwalkar and Afshin Rostamizadeh. Matrix coherence and the Nyström method. *arXiv preprint arXiv:1408.2044*, 2014.

[11] Ahmed Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

[12] Yun Yang, Mert Pilanci, and Martin J. Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017. MR3662446

[13] Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 113–120, 2010.

[14] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems*, pages 476–484, 2012.

[15] Rong Jin, Tianbao Yang, Mehrdad Mahdavi, Yu-Feng Li, and Zhi-Hua Zhou. Improved bounds for the Nyström method with application to kernel classification. *IEEE Transactions on Information Theory*, 59(10):6939–6949, 2013. MR3106876

[16] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

[17] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

[18] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013. MR3450540

[19] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[20] Quoc Le, Tamás Sarlós, and Alexander Smola. Fastfood – Computing Hilbert space expansions in loglinear time. In *International Conference on Machine Learning*, pages 244–252, 2013.

[21] Bo Dai, Bo Xie, Niao He, Yingyu Liang, Anant Raj, Maria-Florina Balcan, and Le Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems*, pages 3041–3049, 2014.

[22] Si Si, Cho-Jui Hsieh, and Inderjit S. Dhillon. Memory efficient kernel approximation. *The Journal of Machine Learning Research*, 18(1):682–713, 2017. MR3634887

[23] Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000. MR1781185

[24] Sara A. van de Geer. *Empirical Processes in M-estimation.* Cambridge University Press, 2000.

[25] Martin J. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint.* Cambridge University Press, 2019. MR3967104

[26] Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and Lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, 1992. MR1182715