Sampling-based Nyström Approximation and Kernel Quadrature

Satoshi Hayakawa [1] **Harald Oberhauser** [1] **Terry Lyons** [1]

Abstract

We analyze the Nyström approximation of a positive definite kernel associated with a probability measure. We first prove an improved error bound for the conventional Nyström approximation with i.i.d. sampling and singular-value decomposition in the continuous regime; the proof techniques are borrowed from statistical learning theory. We further introduce a refined selection of subspaces in Nyström approximation with theoretical guarantees that is applicable to non-i.i.d. landmark points. Finally, we discuss their application to convex kernel quadrature and give novel theoretical guarantees as well as numerical observations.

1. Introduction

Kernel methods form a prominent part among modern machine learning tools. However, making kernel methods scalable to large datasets is an ongoing challenge. The main bottleneck is that the kernel Gram matrix scales quadratically in the number of data points. For large scale problems the number of matrix entries can easily be of the order hundred-thousands or millions so that even storing the full Gram matrix can become too costly. Several approaches have been developed to deal with these, among the most prominent are the Random Fourier Features and the Nyström method. In this article, we revisit and generalize the Nyström method and provide new error estimates. Consequences are theoretical guarantees for kernel quadrature and improvements on the standard Nyström method that go beyond uniform subsampling of data points.

Nyström Approximation. The main idea of the Nyström method is to replace the original kernel k by another kernel k_{app} that is constructed by random projection of the elements in the (in general infinite-dimensional) RKHS associated with k into a low-dimensional RKHS. A consequence of this is that the Gram matrix of k_{app} is a low-rank approximation of the original Gram matrix. Concretely, let μ denote a probability measure on a (Hausdorff) space $\mathcal{X}$ and k a kernel on $\mathcal{X}$; then the standard Nyström approximation uses the random kernel

$$k^Z(x, y) := k(x, Z)k(Z, Z)^+k(Z, y). \tag{1}$$

where $Z = (z_i)_{i=1}^{\ell}$ is an ℓ-point subset of $\mathcal{X}$ usually taken i.i.d. from μ (Drineas et al., 2005; Kumar et al., 2012).

Further s-rank Approximation. While less common, the following rank-reduced version is of our interest:

$$k_{\mathrm{app}}(x, y) = k_s^Z(x, y) := k(x, Z)k(Z, Z)_s^+k(Z, y), \tag{2}$$

where $k(Z, Z)_s^+$ is the Moore–Penrose pseudo-inverse of the best s-rank approximation of the Gram matrix $k(Z, Z) = (k(z_i, z_j))_{i,j=1}^{\ell}$ with $s \le \ell$. Note that $k_\ell^Z = k^Z$.

Our motivation for this rank reduction comes from kernel-based numerical integration. Indeed, if we are given an s-rank kernel k_{app} and a probability measure μ, by Tchakaloff's theorem there is a discrete probability measure ν supported over at most $s + 1$ points satisfying $\int_{\mathcal{X}} f \,\mathrm{d}\mu = \int_{\mathcal{X}} f \,\mathrm{d}\nu$ for all $f \in \mathcal{H}_{k_{\mathrm{app}}}$, where $\mathcal{H}_{k_{\mathrm{app}}}$ is the finite-dimensional RKHS associated with the kernel k_{app}. Such a measure ν works as a kernel quadrature rule if the k_{app} well approximates the original kernel k, and the rank s directly affects the number of (possibly expensive) function evaluations we need to estimate each integral. The primary error criterion in this paper is

$$\int_{\mathcal{X}} \sqrt{k(x, x) - k_{\mathrm{app}}(x, x)} \,\mathrm{d}\mu(x), \tag{3}$$

which arises from the error estimate in kernel/Bayesian quadrature (Hayakawa et al., 2022; Adachi et al., 2022).

Contribution. Our first theoretical result is that the expectation of (3) is of the order $\mathcal{O}\big(\sqrt{\sum_{i>s} \sigma_i} + \mathrm{polylog}(\ell)/\ell\big)$ when the eigenvalues $(\sigma_i)_{i=1}^{\infty}$ of the kernel integral operator induced by (k, μ) enjoy exponential convergence (the expectation is taken over the empirical sample Z). Key to the proof of this bound is the use of concepts from statistical learning theory; in particular, the (local)

[1] Mathematical Institute, University of Oxford, Oxford, United Kingdom. Correspondence to: Satoshi Hayakawa <hayakawa@maths.ox.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Table 1: Main quantitative results. Individual bounds are available in Remark 1, Theorem 2, and Proposition 4. For the explanation on each kernel, see at the end of **Contribution** section. Here are remarks on the notation. (a) σ_i is the i-th eigenvalue of the integral operator $\mathcal{K} : L^2(\mu) \to L^2(\mu); g \mapsto \int_{\mathcal{X}} k(\cdot, x)g(x)\,\mathrm{d}\mu(x)$. (b) μ_X denotes the equally weighted empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ given by $X = (x_i)_{i=1}^N$. (c) $\mu(\cdot)$ and $\mu_X(\cdot)$ denote the integrals over the diagonal. See (4).

Quantity	Bound	Assumption
$\mathbb{E}[\mu(\sqrt{k - k_s^Z})]$ $\mathbb{E}[\mu(k - k_s^Z)]$	$\mathcal{O}\left(\sqrt{\sum_{i>s}\sigma_i} + \frac{(\log \ell)^{2d+1}}{\ell}\right)$	$\begin{cases} Z \sim_{\mathrm{iid}} \mu, \ k\text{: bounded} \\ \sigma_i \lesssim \exp(-\beta i^{1/d}) \end{cases}$
$\left(\mu(\sqrt{k^Z - k_{s,\mu}^Z})^2 \le\right) \mu(k^Z - k_{s,\mu}^Z)$	$\sum_{i>s}\sigma_i$	Z: fixed
$\left(\mathbb{E}[\mu_X(\sqrt{k^Z - k_{s,X}^Z})]^2 \le\right) \mathbb{E}[\mu_X(k^Z - k_{s,X}^Z)]$		Z: fixed, $X \sim_{\mathrm{iid}} \mu$

Rademacher complexity. This error estimate is far better than the bound $\mathcal{O}\big(\text{spectral term} + s^{1/2}/\ell^{1/4}\big)$ that follows from the existing high-probability estimate $\int_{\mathcal{X}}(k(x,x) - k_s^Z(x,x))\,\mathrm{d}\mu(x) = \mathcal{O}(s\sigma_s + \sum_{i>s}\sigma_i + s/\sqrt{\ell})$ (Hayakawa et al., 2022, Corollary 4). By combining our new bound with known kernel quadrature estimates this explains the strong empirical performance of the random kernel quadrature, see Hayakawa et al. (2022); previously the theoretical bounds were not even better than Monte-Carlo in terms of ℓ.

Our second contribution is the use of other k_{app} than k_s^Z with better bounds of (3), for a general class of landmark points Z rather than just an i.i.d. sample from μ. This generalization allows to use other sets Z in (2) to achieve better overall performance; e.g. sampling Z from determinantal point processes (DPPs) on $\mathcal{X}$ is known to be advantageous in applications. To construct and provide theoretical guarantees for such improved Nyström constructions we revisit and generalize a method that was proposed in Santin & Schaback (2016) and give further theoretical guarantees applicable to kernel quadrature rules.

The following is the list of low-rank approximations presented in the paper:

- k^Z and k_s^Z: Usual Nyström approximations using landmark points Z. See (1) and (2).
- $k_{s,\mu}^Z$: The s-rank truncated Mercer decomposition of the kernel k^Z with respect to the measure μ. See (11).
- $k_{s,X}^Z$: A version of $k_{s,\mu}^Z$ with μ given by the empirical measure $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ of the set $X = (x_i)_{i=1}^N$. This actually coincides with k_s^Z when $X = Z$; see (6).

See Table 1 for a summary of our quantitative results.

Outline. Section 2 discusses the existing literature and introduces some notation. Section 3 contains our first main result, namely the analysis of k_s^Z for an i.i.d. Z; Appendix A provides the necessary background from statistical learning theory. In Section 4, we then treat a general Z to give refined low-rank approximations together with theoretical guarantees, rather than the conventional k_s^Z. In Section 5, we discuss how our bounds yields new theories and methods for the recent random kernel quadrature construction, which enables us to explain the empirical performance as well as to build some strong candidates whose performance is assessed by numerical experiments. All the omitted proofs are given in Appendix B.

2. Related Literature and Notation

To simplify the notation, we denote

$$\nu(f) := \int_{\mathcal{X}} f(x)\,\mathrm{d}\nu(x), \quad \nu(h) := \int_{\mathcal{X}} h(x,x)\,\mathrm{d}\nu(x) \quad (4)$$

for any functions $f : \mathcal{X} \to \mathbb{R}$, $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a (probability) measure ν on $\mathcal{X}$, whenever the integrals are well-defined. In this notation, the aim of this paper is to bound $\mu(\sqrt{k - k_{\mathrm{app}}})$ or $\mu(k - k_{\mathrm{app}})$ for a class of low-rank approximation k_{app}. Also, A^+ denotes the Moore–Penrose pseudo-inverse of a matrix A.

Approximation of the Gram Matrix. The standard use of the Nyström method in ML is to replace the Gram matrix $k(X,X)$ for a set $X = (x_i)_{i=1}^N$ by the low-rank matrix $k^Z(X,X)$ where k^Z is defined as in (1). A well-developed literature studies the case when $Z = (z_i)_{i=1}^\ell$ is uniformly and independently sampled from X, see Drineas et al. (2005); Kumar et al. (2012); Yang et al. (2012); Jin et al. (2013); Li et al. (2015). Further, the cases of leverage-based sampling (Gittens & Mahoney, 2016), DPPs (Li et al., 2016), and kernel K-means samples (Oglic & Gärtner, 2017) have received attention. Moreover, two variants of the standard Nyström method have been studied: the first replaces the Moore-Penrose inverse of $k(Z,Z)$ in (1) with the pseudo-inverse of the best s-rank approximation of $k(Z,Z)$ as in (2) via SVD (Drineas et al., 2005; Kumar et al., 2012; Li et al., 2015); the second uses the best s-rank approximation of $k^Z(X,X)$, see (Tropp et al., 2017; Wang et al., 2019).

For a brief overview in this regard, see Wang et al. (2019, Remark 1).

Approximation of the Integral Operator. The matrix $k(X, X)$ can be regarded as a finite-dimensional representation of the linear (integral) operator

$$\mathcal{K} : L^2(\mu) \to L^2(\mu), \quad (\mathcal{K}f)(x) = \int_{\mathcal{X}} k(x, y)f(y)\,\mathrm{d}\mu(y).$$

We denote with $(\sigma_i, e_i)_{i=1}^{\infty}$ the eigenpairs of the operator $\mathcal{K}$, and assume the eigenvalues are ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. The Mercer decomposition exists under mild assumptions (for example, $\operatorname{supp}\mu = \mathcal{X}$, k is continuous and $\int_{\mathcal{X}} k(x, x)\,\mathrm{d}\mu(x) < \infty$ (Steinwart & Scovel, 2012) are sufficient) and gives the representation

$$k(x, y) = \sum_{i=1}^{\infty} \sigma_i e_i(x)e_i(y), \tag{5}$$

where $\|e_i\|_{L^2(\mu)} = 1$, and $(\sqrt{\sigma_i}e_i)_{i=1}^{\infty}$ is an orthonormal basis of the RKHS $\mathcal{H}_k$ of k. Hence, a natural approach is to just truncate this expansions after s terms, $k_{\mathrm{app}} = \sum_{i=1}^{s} \sigma_i e_i(x)e_i(y)$, to get a finite-dimensional approximation of the kernel k. This approach is natural since the approximation quality of the operator $\mathcal{K}$ determines the resulting error estimates. Unfortunately, it is often rendered useless since the Mercer decomposition depends on the tuple (k, μ) and while explicit expression are known for special choices, in general it is unlikely to have a closed-form representation of the eigenpairs $(\sigma_i, e_i)_{i=1}^{\infty}$.

Other Approximations. A compromise which is relevant to our work is proposed in Santin & Schaback (2016). Instead of using the Mercer decomposition of $\mathcal{K}$ one uses the Mercer decomposition of (1). Our main result allows to generalize this approach and to provide theoretical guarantees missing in the reference. Related is the article Gauthier (2021) that studies the interactions of several Hilbert-Schmidt spaces of (integral) operators given by a Nyström approximation/projection of a kernel-measure pair as in the present paper; further, Chatalic et al. (2022) considers a low-rank approximation of an empirical kernel mean embedding by using a Nyström-based projection. The leverage-based sampling studied in Gittens & Mahoney (2016) has continuous counterparts. One with a slight modification is in the kernel literature (Bach, 2017), while the exact counterpart can be found in a context from approximation theory (Cohen & Migliorati, 2017) under the name of *optimally-weighted sampling*, which essentially proposes sampling from $s^{-1}\sum_{i=1}^{s} e_i^2(x)\,\mathrm{d}\mu(x)$.

The Power Function. Finally, the square root of the diagonal term $\sqrt{k(x, x) - k^Z(x, x)}$ or its generalization is known as the *power function* in the literature on kernel-based interpolation (De Marchi, 2003; Santin & Haasdonk, 2017; Karvonen et al., 2021). There the primary interest is its L^∞ (uniform) norm, rather than the $L^1(\mu)$ norm, $\mu(\sqrt{k - k_{\mathrm{app}}})$, or the $L^2(\mu)$ norm, $\mu(k - k_{\mathrm{app}})$, that appear in kernel quadrature estimates and error estimates of the Nyström/Mercer type decompositions.

Kernel Quadrature. The literature on kernel quadrature includes herding (Chen et al., 2010; Bach et al., 2012; Huszár & Duvenaud, 2012; Tsuji et al., 2022), weighted/correlated sampling (Bach, 2017; Belhadji et al., 2019; 2020; Belhadji, 2021), a subsampling method called thinning (Dwivedi & Mackey, 2021; 2022; Shetty et al., 2022) and a positively weighted kernel quadrature (Hayakawa et al., 2022) that motivated our work. We refer to Hayakawa et al. (2022, Table 1) for comparison of existing algorithms in terms of their convergence guarantees and computational complexities.

3. Analyzing k_s^Z for i.i.d. Z via Statistical Learning Theory

Let $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$ and k_s^Z be the s-dimensional kernel given by $k_s^Z(x, y) = k(x, Z)k(Z, Z)_s^+k(Z, y)$ as in the usual Nyström approximation. Throughout the paper, suppose we are provided the singular value decomposition of the matrix $k(Z, Z) = U\operatorname{diag}(\lambda_1, \ldots, \lambda_\ell)U^\top$ with an orthogonal matrix $U = [u_1, \ldots, u_\ell]$ and $\lambda_1 \geq \cdots \geq \lambda_\ell \geq 0$. Note that

$$k_s^Z(x, y) = \sum_{i=1}^{s} \mathbf{1}_{\{\lambda_i > 0\}} \frac{1}{\lambda_i}(u_i^\top k(Z, x))(u_i^\top k(Z, y)) \tag{6}$$

is actually a truncated Mercer decomposition of k^Z with regard to the measure $\mu_Z = \frac{1}{\ell}\sum_{i=1}^{\ell}\delta_{z_i}$, since

$$\begin{aligned}&\left\langle u_i^\top k(Z, \cdot), u_j^\top k(Z, \cdot)\right\rangle_{L^2(\mu_Z)} \\ &= \frac{1}{\ell}u_i^\top k(Z, Z)k(Z, Z)u_j = \frac{\lambda_i\lambda_j}{\ell}\delta_{ij}.\end{aligned}$$

This fact is at the heart of our analysis: k_s^Z is 'optimal' s-rank approximation for the measure μ_Z, and the statistical learning theory connects estimates in empirical measure and the original measure.

Let us denote by $P_{Z,s} : \mathcal{H}_k \to \mathcal{H}_k$ the linear operator given by $k(\cdot, x) \mapsto k_s^Z(\cdot, x)$ for all $x \in \mathcal{X}$. We shall also simply write $P_Z = P_{Z,\ell}$.

Lemma 1. *$P_{Z,s}$ is an orthogonal projection in $\mathcal{H}$.*

This projection is related the quantity of interest, in that $k_s^Z(x, x) = \langle k(\cdot, x), P_{Z,s}k(\cdot, x)\rangle_{\mathcal{H}_k} = \|P_{Z,s}k(\cdot, x)\|_{\mathcal{H}_k}^2$. Thus, we have $k(x, x) - k_s^Z(x, x) = \|P_{Z,s}^\perp k(\cdot, x)\|_{\mathcal{H}_k}^2$ by using $P_{Z,s}^\perp$, the orthogonal complement of $P_{Z,s}$. So we are now interested in estimating the integral $\mu(\sqrt{k-k_s^Z}) = \int_{\mathcal{X}}\|P_{Z,s}^\perp k(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mu(x)$ from the viewpoint of the projection operator. We first estimate its empirical counterpart $\mu_Z(\sqrt{k-k_s^Z}) = \frac{1}{\ell}\sum_{i=1}^\ell\|P_{Z,s}^\perp k(\cdot,z_i)\|_{\mathcal{H}_k}$, where $\mu_Z = \frac{1}{\ell}\sum_{i=1}^\ell \delta_{z_i}$ is the empirical measure. Indeed, we have the following identity regarding $\mu_Z(k-k_s^Z)$:

Lemma 2. *For any ℓ-point sample $Z\subset\mathcal{X}$, we have*

$$\mu_Z(\sqrt{k-k_s^Z})^2 \le \mu_Z(k-k_s^Z) = \frac{1}{\ell}\sum_{i=s+1}^{\ell}\lambda_i$$

where $\lambda_1\ge\cdots\ge\lambda_\ell$ are eigenvalues of $k(Z,Z)$.

When Z is given by an i.i.d. sampling, the decay of eigenvalues λ_i enjoys the rapid decay given by σ_i in the following sense:

Lemma 3. *Let $Z=(z_i)_{i=1}^\ell$ be an ℓ-point independent sample from μ. Then, for the eigenvalues $\lambda_1\ge\cdots\ge\lambda_\ell$ of $k(Z,Z)$, we have*

$$\mathbb{E}\left[\frac{1}{\ell}\sum_{i=s+1}^{\ell}\lambda_i\right] \le \sum_{i>s}\sigma_i.$$

For a general random orthogonal projection operator, we can prove the following bound by using arguments in statistical learning theory (Section A):

Theorem 1. *Let $Z=(z_i)_{i=1}^\ell$ be an ℓ-point independent sample from μ and P be a random orthogonal projection in $\mathcal{H}_k$ possibly depending on Z. For any integer $m\ge 1$, we have the following bound:*

$$\mathbb{E}\left[\int_{\mathcal{X}}\|Pk(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mu(x)\right] \le \mathbb{E}\left[\frac{2}{\ell}\sum_{i=1}^{\ell}\|Pk(\cdot,z_i)\|_{\mathcal{H}_k}\right] + 4\sqrt{\sum_{i>m}\sigma_i} + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2\log(1+2\ell)}{9}+69\right),$$

where the expectation is taken regarding the draws of Z.

Recall that $\mu(\sqrt{k-k_s^Z}) = \int_{\mathcal{X}}\|P_{Z,s}^\perp k(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mu(x)$. By combining this theorem when $P=P_{Z,s}^\perp$ and Lemma 2 & 3, we can obtain the following:

Corollary 1. *Let $Z=(z_i)_{i=1}^\ell$ be an ℓ-point independent sample from μ. Then, for any integer $m\ge 1$, we have*

$$\mathbb{E}\left[\mu(\sqrt{k-k_s^Z})\right] \le 2\sqrt{\sum_{i>s}\sigma_i} + 4\sqrt{\sum_{i>m}\sigma_i} + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2\log(1+2\ell)}{9}+69\right).$$

Remark 1. When $\sigma_j \lesssim e^{-\beta i^{1/d}}$ with a constant $\beta>0$ and a positive integer d (typical for d-dimensional Gaussian kernel, see, e.g., Adachi et al., 2022, Section A.2), by taking $m\sim(\log\ell)^d$, we have a bound

$$\mathbb{E}\left[\mu(\sqrt{k-k_s^Z})\right] = \mathcal{O}\left(\sqrt{\sum_{i>s}\sigma_i} + \frac{(\log\ell)^{2d+1}}{\ell}\right)$$

for $\ell\ge 3$; see Appendix B.6 for the proof. Since $k-k_s^Z \le \sqrt{k_{\max}}\sqrt{k-k_s^Z}$, the same estimate applies to $\mathbb{E}[\mu(\sqrt{k-k_s^Z})]$. These also lead to an $(s+1)$-point randomized convex kernel quadrature Q_{s+1} with the same order of $\mathbb{E}[\mathrm{wce}(Q_{s+1})]$. See Section 5 for details.

4. A Refined Low-rank Approximation with General Z

The process of obtaining a good approximation k_{app} of k using k^Z can be decomposed into two parts:

$$k-k_{\mathrm{app}} = \underbrace{k-k^Z}_{\mathrm{A}} + \underbrace{k^Z-k_{\mathrm{app}}}_{\mathrm{B}}.$$

In the previous section, we have analyzed the case Z is i.i.d. and $k_{\mathrm{app}}=k_s^Z$. However, we can consider more general Z, and indeed we actually have a better way to select a subspace (i.e., k_{app}) from the finite-rank kernel k^Z rather than just using k_s^Z.

4.1. Part A: Estimating the Error of k^Z for General Z

This part is relatively well-studied. Indeed, $\mu(k-k^Z) = \int_{\mathcal{X}}(k(x,x)-k^Z(x,x))\,\mathrm{d}\mu(x)$ for some non-i.i.d. Z can be bounded by using the results of weighted kernel quadrature. For example, Belhadji et al. (2019) consider the worst-case error for the weighted integral

$$\mu(fg) = \int_{\mathcal{X}} f(x)g(x)\,\mathrm{d}\mu(x) \approx \sum_{i=1}^{\ell} w_i f(z_i) \tag{7}$$

for any $\|f\|_{\mathcal{H}_k}\le 1$ and a fixed $g\in L^2(\mu)$ with $Z=(z_i)_{i=1}^\ell$ following a certain DPP. Now consider the optimal worst-case error in the above approximation for the fixed point configuration Z:

$$\begin{aligned}
&\inf_{w_i}\sup_{\|f\|_{\mathcal{H}_k}\le 1}\left|\mu(fg)-\sum_{i=1}^{\ell}w_i f(z_i)\right| \\
&= \sup_{\|f\|\le 1}\left|\left\langle f, \int_{\mathcal{X}} k(\cdot,x)g(x)\,\mathrm{d}\mu(x) - \sum_{i=1}^{\ell}w_i k(\cdot,z_i)\right\rangle_{\mathcal{H}_k}\right| \\
&= \inf_{w_i}\left\|\mathcal{K}g - \sum_{i=1}^{\ell}w_i k(\cdot,z_i)\right\|_{\mathcal{H}_k} = \|P_Z^\perp\mathcal{K}g\|_{\mathcal{H}_k}.
\end{aligned} \tag{8}$$

By using this, we can prove the following estimate:

Proposition 1. *For any finite subset $Z \subset \mathcal{X}$ and any integer $m \geq 0$, we have*

$$\mu(k - k^Z) = \sum_{i=1}^{\infty} \|P_Z^\perp \mathcal{K} e_i\|_{\mathcal{H}_k}^2 \leq \sum_{i=1}^{m} \|P_Z^\perp \mathcal{K} e_i\|_{\mathcal{H}_k}^2 + \sum_{i>m} \sigma_i$$

where $(\sigma_i, e_i)_{i=1}^\infty$ are the eigenpairs of $\mathcal{K}$.

The papers Belhadji et al. (2019; 2020); Belhadji (2021) give bounds on the worst-case error of the weighted kernel quadrature (8) when Z is given by some correlated sampling, whereas Bach (2017) gives another bound when Z is given by an optimized weighted sampling rather than sampling from μ. By using (8) and Proposition 1, we can import their bounds on weighted kernel quadrature with non-i.i.d. Z to the estimate of $\mu(k - k^Z) = \int_X \|P_Z^\perp k(\cdot, x)\|_{\mathcal{H}_k}^2 \,\mathrm{d}\mu(x)$. Here, we just give one such example:

Corollary 2. *Let $Z = (z_i)_{i=1}^\ell$ be taken from a DPP given by the projection kernel $p(x, y) = \sum_{i=1}^\ell e_i(x) e_i(y)$ with a reference measure μ, i.e., $\mathbb{P}(Z \in A) = \frac{1}{\ell!} \int_A \det p(Z, Z) \,\mathrm{d}\mu^{\otimes \ell}(Z)$ for any Borel set $A \subset \mathcal{X}^d$. Then, for any integer $m \geq 0$, we have*

$$\mathbb{E}\big[\mu(k - k^Z)\big] \leq \sum_{i>m} \sigma_i + 4m \sum_{i>\ell} \sigma_i,$$

where the expectation is taken regarding the draws of Z.

In any case, by using those non-i.i.d. points, we can obtain a better Z in the sense that $\int_{\mathcal{X}} (k(x, x) - k^Z(x, x)) \,\mathrm{d}\mu(x)$ attains a sharper upper bound than the bound given in the previous section for an ℓ-point i.i.d. sample from μ. However, for a general Z, it is not necessary sensible to execute the SVD of $k(Z, Z)$ and get k_s^Z accordingly, as a SVD of $k(Z, Z)$ corresponds to approximating μ by the empirical measure $\frac{1}{\ell} \sum_{i=1}^\ell \delta_{z_i}$ (indeed, this observation is the key to the results in the previous section). Thus, for points Z not given by i.i.d. sampling, there should exist a better choice of k_{app} than k_s^Z. We discuss this in the following section.

4.2. Part B: Mercer Decomposition of k^Z

Instead of using k_s^Z, we propose to compute the Mercer decomposition of k^Z with respect to μ and truncate it to get $k_{s,\mu}^Z$, which is defined in the following. This is doable if we have knowledge of $h_\mu(x, y) := \int_{\mathcal{X}} k(x, t) k(t, y) \,\mathrm{d}\mu(t)$, since k^Z is a finite-dimensional kernel. We can prove the following:

Lemma 4. *We have $h_\mu(x, y) = \sum_{i=1}^\infty \sigma_i^2 e_i(x) e_i(y)$.*

We now discuss how h_μ can be used to derive the Mercer decomposition of k^Z. Note that this can be regarded as a generalization of Santin & Schaback (2016, Section 6). Let $\mathcal{K}^Z : L^2(\mu) \to L^2(\mu)$ be the integral operator given by $g \mapsto \int_{\mathcal{X}} k^Z(\cdot, x) g(x) \,\mathrm{d}\mu(x)$.

For functions of the form $f = a^\top k(Z, \cdot)$ and $g = b^\top k(Z, \cdot)$ with $a, b \in \mathbb{R}^\ell$, we have

$$\begin{aligned} \langle f, g \rangle_{L^2(\mu)} &= \int_{\mathcal{X}} a^\top k(Z, x) k(x, Z) b \,\mathrm{d}\mu(x) \\ &= a^\top h_\mu(Z, Z) b. \end{aligned} \tag{9}$$

So, if we write $h_\mu(Z, Z) = H^\top H$ by using an $H \in \mathbb{R}^{\ell \times \ell}$ (since $h_\mu(Z, Z)$ is positive semi-definite), an element $f = a^\top k(Z, \cdot) \in L^2(\mu)$ is non-zero if and only if $Ha \neq 0$. Furthermore, we have

$$\begin{aligned} \mathcal{K}^Z f &= \int_{\mathcal{X}} k(\cdot, Z) k(Z, Z)^+ k(Z, x) k(x, Z) a \,\mathrm{d}\mu(x) \\ &= k(\cdot, Z) k(Z, Z)^+ h_\mu(Z, Z) a \\ &= \big[k(Z, Z)^+ h_\mu(Z, Z) a\big]^\top k(Z, \cdot). \end{aligned} \tag{10}$$

Thus, f is a nontrivial eigenfunction of $\mathcal{K}^Z$, if $Ha \neq 0$ and a is an eigenvector of $k(Z, Z)^+ h_\mu(Z, Z)$. It is equivalent to $c = Ha$ being an eigenvector of $Hk(Z, Z)^+ H^\top$.

Let us decompose this matrix by SVD as $Hk(Z, Z)^+ H^\top = V \operatorname{diag}(\kappa_1, \ldots, \kappa_\ell) V^\top$, where the $V = [v_1, \ldots, v_\ell] \in \mathbb{R}^{\ell \times \ell}$ is an orthogonal matrix and $\kappa_1 \geq \cdots \geq \kappa_\ell \geq 0$. Then, we have

$$Hk(Z, Z)^+ H^\top = \sum_{i=1}^\ell \kappa_i v_i v_i^\top.$$

Let us consider $f_i = (H^+ v_i)^\top k(Z, \cdot) = v_i^\top (H^+)^\top k(Z, \cdot)$ for $i = 1, \ldots, \ell$ as candidates of eigenfunctions of $\mathcal{K}^Z$. We can actually prove the following:

Lemma 5. *The set $\{f_i \mid i \geq 1, \kappa_i > 0\}$ forms an orthonormal subset of $L^2(\mu)$ whose elements are eigenfunctions of $\mathcal{K}^Z$.*

Let us define $k_\mu^Z(x, y) := \sum_{i=1}^\ell \kappa_i f_i(x) f_i(y)$; note that this is computable. From the above lemma, this expression is a natural candidate for "Mercer decomposition" of k^Z. We can prove that it actually coincides with $k^Z(x, y)$ μ-almost everywhere, and so the decomposition is independent of the choice of H up to μ-null sets:

Proposition 2. *There exists a measurable set $A \subset \mathcal{X}$ depending on Z with $\mu(A) = 1$ such that $k^Z(x, y) = k_\mu^Z(x, y)$ holds for all $x, y \in A$. Moreover, we can take $A = \mathcal{X}$ if $\ker h_\mu(Z, Z) \subset \ker k(Z, Z)$.*

Now we just define $k_{s,\mu}^Z$ for $s \leq \ell$ as follows:

$$k_{s,\mu}^Z(x, y) := \sum_{i=1}^s \kappa_i f_i(x) f_i(y). \tag{11}$$

Theorem 2. *We have $\mu(k_\mu^Z - k_{s,\mu}^Z) \leq \sum_{i=s+1}^\ell \sigma_i$ for any $Z = (z_i)_{i=1}^\ell \subset \mathcal{X}$.*

Proof. The left-hand side is equal to $\sum_{i=s+1}^{\ell}\kappa_i$ from Lemma 5 and the definition of the kernels. Thus, it suffices to prove $\kappa_i \le \sigma_i$ for each i. It directly follows from the min-max principle (or Weyl's inequality) as $k - k_\mu^Z$ is positive definite on an $A \subset \mathcal{X}$ with $\mu(A) = 1$ from Proposition 2. □

Remark 2. The choice of the matrix H with $H^\top H = h_\mu(Z,Z)$ does not affect the theory but might affect the numerical errors. We have used the matrix square-root $h_\mu(Z,Z)^{1/2}$, i.e., the symmetric and positive semi-definite matrix H with $H^2 = h_\mu(Z,Z)$, throughout the experiments in Section 5, so that we just need to take the pseudo-inverse of positive semi-definite matrices.

Approximate Mercer Decomposition. When we have no access to the function h_μ, we can just approximate it by using an empirical measure. For a $X = (x_j)_{j=1}^M \subset \mathcal{X}$, denote by h_X the function given by replacing μ in h_μ with the empirical measure with points X:

$$h_X(x,y) = \frac{1}{M}\sum_{j=1}^M k(x,x_j)k(x_j,y) = \frac{1}{M}k(x,X)k(X,y).$$

We can actually replace every h_μ by h_X in the above construction to define k_X^Z and $k_{s,X}^Z$. This approximation is already mentioned by Santin & Schaback (2016) without theoretical guarantee. Another remark is that, when restricted on the set X, it is equivalent to the best s-rank approximation of $k^Z(X,X)$ in the Gram-matrix case (Tropp et al., 2017; Wang et al., 2019), since the L^2-norm for the uniform measure on X just corresponds to the ℓ^2-norm in $\mathbb{R}^{|X|}$.

Note that we have $k_X^Z(X,X) = k^Z(X,X)$ from Proposition 2 in the discrete case. As we have $\ker h_X(Z,Z) = \ker k(Z,X)k(X,Z) = \ker k(X,Z)$, we additionally obtain the following sufficient condition from Proposition 2.

Proposition 3. *$k_X^Z(x,y) = k^Z(x,y)$ holds for all $x,y \in X$. Moreover, if $\ker k(X,Z) \subset \ker k(Z,Z)$, then we have $k_X^Z = k^Z$ over the whole $\mathcal{X}$.*

In particular, we have $k_X^Z = k^Z$ whenever $Z \subset X$. These (at least μ-a.s.) equalities given in Proposition 2 & 3 are necessary for the applications to kernel quadrature, since we need $k - k_{\mathrm{app}}$ to be positive definite for exploiting the existing guarantees such as Theorem 3 in the next section.

Although checking $k_X^Z = k^Z$ is not an easy task, from the first part of Proposition 3, $k_{s,X}^Z$ satisfies the following estimate in terms of the empirical measure μ_X.

Proposition 4. *Let $Z \subset \mathcal{X}$ be a fixed subset and X be an M-point independent sample from μ. Then, we have*

$$\mathbb{E}\big[\mu_X(k^Z - k_{s,X}^Z)\big] = \mathbb{E}\big[\mu_X(k_X^Z - k_{s,X}^Z)\big] \le \sum_{i>s}\sigma_i,$$

where the expectation is taken regarding the draws of X.

We can also give a bound of the resulting error $\mu(k^Z - k_{s,X}^Z)$ again by using the arguments from learning theory, but under an additional assumption as stated in the following. Nevertheless, Proposition 4 is already sufficient for our application in kernel quadrature; see Theorem 4.

Proposition 5. *Under the same setting as in Proposition 4, if $\ker k(X,Z) \subset \ker k(Z,Z)$ holds almost surely for the draws of X, we have*

$$\mathbb{E}\Big[\mu\big(\sqrt{k^Z - k_{s,X}^Z}\big)\Big] \le 2\sqrt{\sum_{i>M}\sigma_i} + 4\sqrt{\sum_{i>m}\sigma_i} + \frac{\sqrt{k_{\max}}}{M}\left(\frac{80m^2\log(1+2M)}{9} + 69\right).$$

for any integer $m \ge 1$.

Remark 3. The assumption $\ker k(X,Z) \subset \ker k(Z,Z)$ seems to be very hard to check in practice. An example with this property is $(\mathcal{X}, k, \mu)$ such that $\mathcal{X} = \mathbb{R}^D$ with $D, M > \ell$, the kernel k is just the Euclidean inner product on $\mathbb{R}^D$, and μ is given by a Gaussian distribution with a nonsingular covariance matrix.

This said, we have some ways to avoid this issue in practice. One way is to use $X \cup Z$ instead of X so that the condition automatically holds. Then, the above order of estimate should still hold when $\ell \ll M$, though it complicates the analysis. Another way is effective when we use k_X^Z for constructing a kernel quadrature from an empirical measure given by X itself; see the next section for details.

5. Application to Kernel Quadrature

Let us give error bounds for kernel quadrature as a consequence of the previous sections. We are mainly concerned with the kernel quadrature of the form (7) without weight, i.e., the case when $g = 1$ for efficiently discretizing the probability measure μ.

Given an n-point quadrature rule $Q_n : f \mapsto \sum_{i=1}^n w_i f(x_i)$ with weights $w_i \in \mathbb{R}$ and points $x_i \in \mathcal{X}$, the worst-case error of Q_n with respect to the RKHS $\mathcal{H}_k$ and the target measure μ is defined as

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) := \sup_{\|f\|_{\mathcal{H}_k}\le 1} |Q_n(f) - \mu(f)|.$$

Note that it is equal to $\mathrm{MMD}_k(Q_n, \mu)$, the maximum mean discrepancy (with k) between Q_n regarded as a (signed) measure and μ (Gretton et al., 2006). We call Q_n *convex* if it is a probability measure, i.e., $w_i \ge 0$ and $\sum_{i=1}^n w_i = 1$.

Suppose we are given an s-rank kernel approximation $k_{\mathrm{app}}(x,y) = \sum_{i=1}^s c_i\varphi_i(x)\varphi_i(y)$ with $c_i \ge 0$ and $k - k_{\mathrm{app}}$ being positive definite (μ-almost surely). The following is taken from Hayakawa et al. (2022, Theorem 6 & 8).

Theorem 3. *If an n-point convex quadrature Q_n satisfies $Q_n(\varphi_i) = \mu(\varphi_i)$ for $1 \le i \le s$ and $Q_n(\sqrt{k - k_{\mathrm{app}}}) \le \mu(\sqrt{k - k_{\mathrm{app}}})$, then we have*

$$\mathrm{wce}(Q_n; \mathcal{H}_k, \mu) \le 2\mu(\sqrt{k - k_{\mathrm{app}}}).$$

Moreover, such a quadrature Q_n exists with $n = s + 1$.

Although there is a randomized algorithm for constructing the Q_n stated in the above theorem (Hayakawa et al., 2022, Algorithm 2 with modification), it has two issues; it requires exact values of $\mu(\varphi_i)$ (and $\mu(\sqrt{k - k_{\mathrm{app}}})$) and its computational complexity has no useful upper bound unless we have additional assumptions such as well-behaved moments of test functions φ_i (Hayakawa et al., 2023a) or structure like a product kernel with a product measure (Hayakawa et al., 2023b). This said, we can deduce updated convergence results for outputs of the algorithm as in Remark 1.

5.1. Kernel Recombination

Instead of considering an "exact" quadrature, what we do in practice in this low-rank approach is matching the integrals against a large empirical measure (see also Adachi et al., 2022, Section 6), say $\mu_Y = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$ with $Y = (y_i)_{i=1}^N$. If we have

$$\begin{cases} Q_n(\varphi_i) = \mu_Y(\varphi_i), & 1 \le i \le s, \\ Q_n(\sqrt{k - k_{\mathrm{app}}}) \le \mu_Y(\sqrt{k - k_{\mathrm{app}}}), \end{cases} \tag{12}$$

then, from Theorem 3 with a target measure μ_Y and the triangle inequality of MMD, we have

$$\begin{aligned} \mathrm{wce}(Q_n; \mathcal{H}_k, \mu) &\le \mathrm{MMD}_k(Q_n, \mu_Y) + \mathrm{MMD}_k(\mu_Y, \mu) \\ &\le 2\mu_Y(\sqrt{k - k_{\mathrm{app}}}) + \mathrm{MMD}_k(\mu_Y, \mu). \end{aligned} \tag{13}$$

Indeed, such a quadrature Q_n with $n = s + 1$ and points given by a subset of Y can be constructed via an algorithm called *recombination* (Litterer & Lyons, 2012; Tchernychova, 2016; Cosentino et al., 2020; Hayakawa et al., 2022).

Existing approaches of this kernel recombination have then been using an approximation k_{app} typically given by k_s^Z whose randomness is independent from the sample Y, but it is not a necessary requirement as long as we can expect an efficient bound of $\mu_Y(\sqrt{k - k_{\mathrm{app}}})$ in some sense. Another small but novel observation is that $k - k_{\mathrm{app}}$ being positive definite is only required on the sample Y in deriving the estimate (13); not over the support of μ in contrast to Theorem 3. These observations circumvent the issues mentioned in Remark 3 when using $k_{\mathrm{app}} = k_Y^Z$ ($k_{s,X}^Z$ with $X = Y$).

Let us now denote the kernel recombination in a form of function as $Q_n = \mathrm{KQuad}(k_{\mathrm{app}}, Y)$, where the output Q_n is an n-point convex quadrature satisfying $n = s + 1$ and (12); note that the constraint is slightly different from what is given in Hayakawa et al. (2022, Algorithm 1), but we can achieve (12) by replacing $k_{1,\mathrm{diag}}$ with $\sqrt{k_{1,\mathrm{diag}}}$ in the cited algorithm.

We can now prove the performance of low-rank approximations given in the previous section. Indeed, $k_{s,Y}^Z$ and $k_{s,\mu}^Z$ have the following same estimate.

Theorem 4. *Let $Z \subset \mathcal{X}$ be a fixed subset and Y be an N-point independent sample from μ. Then, a random convex quadrature $Q_n = \mathrm{KQuad}(k_{s,Y}^Z, Y)$ satisfies*

$$\begin{aligned} &\mathbb{E}[\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)] \\ &\le 2\mu(\sqrt{k - k^Z}) + 2\sqrt{\sum_{i>s}\sigma_i} + \sqrt{\frac{c_{k,\mu}}{N}}, \end{aligned} \tag{14}$$

where $c_{k,\mu} := \mu(k) - \iint_{\mathcal{X}\times\mathcal{X}} k(x, y)\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(y)$ and the expectation is taken regarding the draws of Y. The estimate (14) holds also for $Q_n = \mathrm{KQuad}(k_{s,\mu}^Z, Y)$.

5.2. Numerical Examples

In this section, we compare the numerical performance of $k_{s,Y}^Z$ and $k_{s,\mu}^Z$ for kernel quadrature with the conventional Nyström approximation for a non-i.i.d. Z in the setting that we can explicitly compute the worst-case error.

Periodic Sobolev Spaces. The class of RKHS we use is called periodic Sobolev spaces of functions on $\mathcal{X} = [0, 1]$ (a.k.a. Korobov spaces), and given by the following kernel for a positive integer r:

$$k_r(x, y) = 1 + \frac{(-1)^{r-1}(2\pi)^{2r}}{(2r)!} B_{2r}(|x - y|),$$

where B_{2r} is the $2r$-th Bernoulli polynomial (Wahba, 1990; Bach, 2017). We consider the case μ being the uniform measure, where the eigenfunctions of the integral operator $\mathcal{K}$ are known to be $1, \sqrt{2}\cos(2\pi m\,\cdot), \sqrt{2}\sin(2\pi m\,\cdot)$ with eigenvalues respectively $1, m^{-2r}, m^{-2r}$ for each positive integer m. This RKHS is commonly used for measuring the performance of kernel quadrature methods (Kanagawa et al., 2016; Bach, 2017; Belhadji et al., 2019; Hayakawa et al., 2022). We also consider its products: $k_r^{\otimes d}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^d k_r(x_i, y_i)$ and μ being the uniform measure on the hypercube $\mathcal{X} = [0, 1]^d$.

By considering the eigenvalues, we can see that $h_\mu = k_{2r}^{\otimes d}$ for each kernel $k_r^{\otimes d}$ from Remark 4.

Experiments. In the experiments for the kernel $k_r^{\otimes d}$, we compared the worst-case error of n-point kernel quadrature rules given by $Q_n = \mathrm{KQuad}(k_{\mathrm{app}}, Y)$ with $k_{\mathrm{app}} = k_s^H, k_s^Z, k_{s,Y}^Z, k_{s,\mu}^Z$ ($s = n - 1$) under the following setting:

- Y is an N-point independent sample from μ with $N = n^2$ (Figure 1) or $N = n^3$ (Figure 2).

- H is the uniform grid $\{i/n \mid i = 1, \ldots, n\}$ ($d = 1$) or the Halton sequence with Owen scrambling (Halton, 1960; Owen, 2017) ($d \geq 2$).

- Z is the union of H and another $20n$-point independent sample from $\nu^{\otimes d}$, where ν is the 1-dimensional $(2, 5)$-Beta distribution, whose density is proportional to $x(1-x)^4$ for $x \in [0, 1]$.

We additionally compared '**Monte Carlo**': uniform weights $1/n$ with i.i.d. sample $(x_i)_{i=1}^n$ from μ, '**Uniform Grid**' ($d = 1$): points in H with uniform weights $1/n$ (known to be *optimal* for each n), and '**Halton**' ($d \geq 2$): points in an independent copy of H with uniform weights $1/n$.

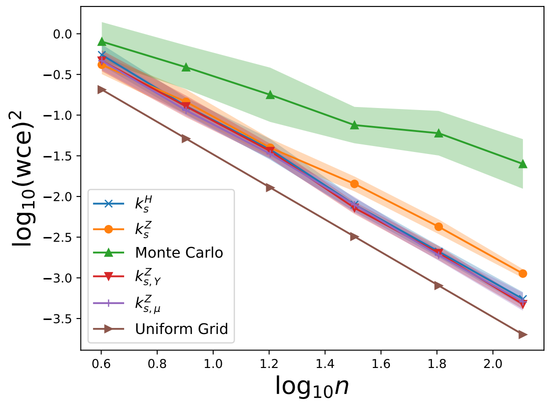
The aim of this experiment was to see if the proposed methods ($k^Z_{s,Y}$ and $k^Z_{s,\mu}$) can actually recover a 'good' subspace of the RKHS given by k^Z with Z not summarizing μ. To do so, we mixed H (a 'good' summary of μ) and an i.i.d. sample from ν to determine Z.

Figure 1 shows the results for $(d, r) = (1, 1), (2, 1), (3, 3)$ with $N = n^2$ and $n = 4, 8, 16, 32, 64, 128$. From Figure 1(a, b), we can see that our methods indeed recover (and perform slightly better than) the rate of k^H from a contaminated sample Z. In Figure 1(c), the four low-rank methods all perform equally well, and it seems that the dominating error is given by the term caused by $\mathrm{MMD}_k(\mu_Y, \mu)$.

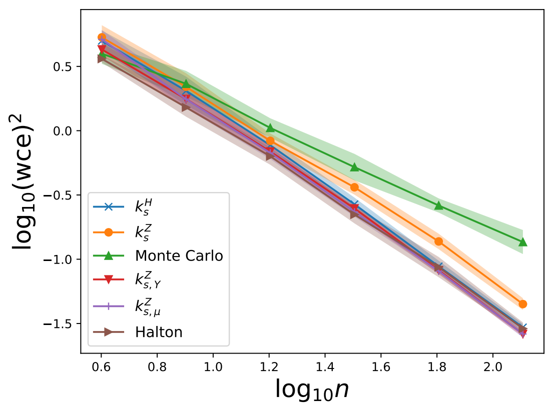
Figure 2 shows the results for $(d, r) = (1, 2)$ with $N = n^2$ or $N = n^3$ and $n = 4, 8, 16, 32, 64$. In this case, we can see that $k^Z_{s,Y}$ or $k^Z_{s,\mu}$ eventually suffers from the numerical instability, which is also reported by Santin & Schaback (2016). Since their error inflation is not completely hidden even in the case $N = n^2$ unlike the previous experiments, one possible reason for the instability is that taking the pseudo-inverse of $k(Z, Z)$ or $h_\mu(Z, Z)^{1/2}$ in the algorithm becomes highly unstable when the spectral decay is fast. Although they have preferable guarantees in theory, its numerical error seems to harm the overall efficiency, and this issue needs to be addressed e.g. by circumventing the use of pseudo-inverse in future work.

Remark 4. Unlike the kernel quadrature with $k^Z_{s,\mu}$ or $k^Z_{s,Y}$, that with k^Z_s does not suffer from a similar numerical instability despite the use of $k(Z, Z)^+_s$. This phenomenon can be explained by the nature of Hayakawa et al. (2022, Algorithm 1); it only requires (stable) *test functions* $\varphi_i = u_i^\top k(Z, \cdot)$ ($i = 1, \ldots, s$) for its equality constraints, where u_i is the i-th eigenvector of $k(Z, Z)$, while the (possibly unstable) diagonal term $k^Z_s(x, x)$ appears in the inequality constraint, which can empirically be omitted (Hayakawa et al., 2022, Section E.2).

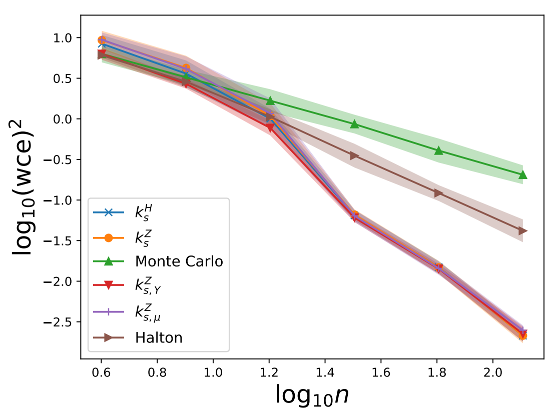
Computational Complexity. By letting ℓ, N (larger than s) respectively be the cardinality of Z and Y, we can express



(a) $d = 1, r = 1$

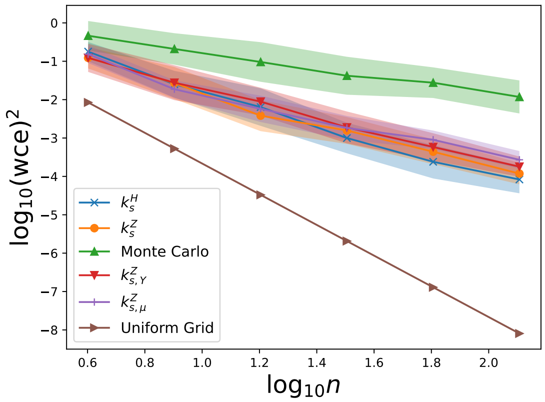


(b) $d = 2, r = 1$

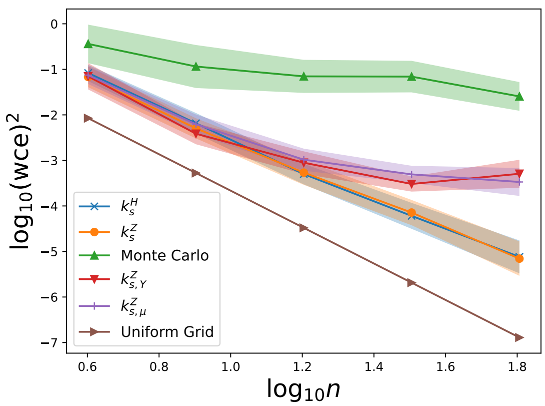


(c) $d = 3, r = 3$

Figure 1: Experiments in periodic Sobolev spaces with reproducing kernel $k_r^{\otimes d}$. Average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 samples plotted with their standard deviation.



(a) $N = n^2$



(b) $N = n^3$

Figure 2: Experiments in k_2 with $N = n^2, n^3$ for recombination algorithms. Average of $\log_{10}(\mathrm{wce}(Q_n; \mathcal{H}_k, \mu)^2)$ over 20 samples plotted with their standard deviation.

the computational steps of $\mathrm{KQuad}(k_{\mathrm{app}}, Y)$ with $k_{\mathrm{app}} = k_s^Z, k_{s,Y}^Z, k_{s,\mu}^Z$ as follows:

- Using k_s^Z takes $\mathcal{O}\big(s\ell N + s\ell^2 + s^3 \log(N/s)\big)$, but it can be reduced to $\mathcal{O}\big(\ell N + s\ell^2 + s^3 \log(N/s)\big)$ by omitting the (empirically unnecessary) inequality constraint (Hayakawa et al., 2022, Remark 2).
- Using $k_{s,Y}^Z$ takes $\mathcal{O}\big(\ell^3 + \ell^2 N + s^3 \log(N/s)\big)$, where $\mathcal{O}\big(\ell^3\big)$ and $\mathcal{O}\big(\ell^2 N\big)$ respectively come from computing $k(Z, Z)^+$ and $h_Y(Z, Z)$.
- Using $k_{s,\mu}^Z$ takes $\mathcal{O}\big(\ell^3 + s\ell N + s^3 \log(N/s)\big)$ (if h_μ available), where $\mathcal{O}\big(\ell^3\big)$ is from computing $k(Z, Z)^+$.

For example, in the case of Figure 1(c) with $n = 128$, the average time per one trial was respectively 26.5, 226, 216 seconds for $k_s^Z, k_{s,Y}^Z, k_{s,\mu}^Z$, while it was 52.6, 57.8, 41.2 seconds for the case of Figure 2(b) with $n = 64$.[1]

6. Concluding Remarks

In this paper, we have studied the performance of several Nyström-type approximations k_{app} of a positive definite kernel k associated with a probability measure μ, in terms of the error $\mu(\sqrt{k - k_{\mathrm{app}}})$. We first improved the bounds for k_s^Z, the conventional Nyström approximation based on an i.i.d. Z and the use of SVD, by leveraging results in statistical learning theory. We then went beyond the i.i.d. setting and considered general Z including DPPs; we further introduced two competitors of k_s^Z, i.e., $k_{s,\mu}^Z$ and $k_{s,X}^Z$, which are given by directly computing the Mercer decomposition of the finite-rank kernel k^Z against the measure μ and the empirical measure μ_X, respectively. Finally, we used our results to improve the theoretical guarantees for convex kernel quadrature Hayakawa et al. (2022), and provided numerical results to illustrate the difference between the conventional k_s^Z and the newly proposed $k_{s,\mu}^Z$ and $k_{s,X}^Z$.

Despite its nice theoretical properties, a limitation of our second contribution, i.e., the proposed kernel approximations, is that they involve the computation of a pseudo-inverse, which can be numerically unstable when there is a rapid spectral decay. This point should be addressed in future work, but one promising approach in the context of kernel quadrature is to conceptually learn from the stability of k_s^Z mentioned in Remark 4; if we see the construction of the low-rank kernel as optimization of the vectors u_i for which functions $u_i^\top k(Z, \cdot)$ well approximate $\mathcal{H}_{k^Z}$ in terms of $L^2(\mu)$ metric, we can possibly leverage the stability of convex optimization for instance.

Acknowledgements

The authors would like to thank Ken'ichiro Tanaka and the anonymous reviewers for helpful comments. This work was supported in part by the EPSRC [grant number EP/S026347/1], in part by The Alan Turing Institute under the EPSRC grant EP/N510129/1, the Data Centric Engineering Programme (under the Lloyd's Register Foundation grant G0095), the Defence and Security Programme (funded by the UK Government) and the Office for National Statistics & The Alan Turing Institute (strategic partnership) and in part by the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA).

[1] All the experiments were conducted on a MacBook Pro with Apple M1 Max chip and 32GB unified memory. Code is available at the `nystrom` folder in https://github.com/satoshi-hayakawa/kernel-quadrature.

References

Adachi, M., Hayakawa, S., Jørgensen, M., Oberhauser, H., and Osborne, M. A. Fast Bayesian inference with batch Bayesian quadrature via kernel recombination. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16533–16547, 2022.

Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

Bach, F., Lacoste-Julien, S., and Obozinski, G. On the equivalence between herding and conditional gradient algorithms. In *International Conference on Machine Learning*, pp. 1355–1362, 2012.

Belhadji, A. An analysis of Ermakov–Zolotukhin quadrature using kernels. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Belhadji, A., Bardenet, R., and Chainais, P. Kernel quadrature with DPPs. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12907–12917, 2019.

Belhadji, A., Bardenet, R., and Chainais, P. Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning*, pp. 725–735. PMLR, 2020.

Chatalic, A., Schreuder, N., Rosasco, L., and Rudi, A. Nyström kernel mean embeddings. In *International Conference on Machine Learning*, pp. 3006–3024. PMLR, 2022.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence*, pp. 109–116, 2010.

Cohen, A. and Migliorati, G. Optimal weighted least-squares methods. *The SMAI journal of computational mathematics*, 3:181–203, 2017.

Cosentino, F., Oberhauser, H., and Abate, A. A randomized algorithm to reduce the support of discrete measures. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15100–15110, 2020.

De Marchi, S. On optimal center locations for radial basis function interpolation: computational aspects. *Rendiconti del Seminario Matematico*, 61(3):343–358, 2003.

Drineas, P., Mahoney, M. W., and Cristianini, N. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6(12):2153–2175, 2005.

Dwivedi, R. and Mackey, L. Kernel thinning. In *Conference on Learning Theory*, pp. 1753–1753. PMLR, 2021.

Dwivedi, R. and Mackey, L. Generalized kernel thinning. In *International Conference on Learning Representations*, 2022.

Gauthier, B. Nyström approximation and reproducing kernels: embeddings, projections and squared-kernel discrepancy. *preprint*, 2021. URL https://hal.archives-ouvertes.fr/hal-03207443.

Giné, E. and Koltchinskii, V. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

Gittens, A. and Mahoney, M. W. Revisiting the Nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1):3977–4041, 2016.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer, 2006.

Halton, J. H. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik*, 2(1):84–90, 1960.

Hayakawa, S. and Suzuki, T. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks*, 123:343–361, 2020.

Hayakawa, S., Oberhauser, H., and Lyons, T. Positively weighted kernel quadrature via subsampling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 6886–6900, 2022.

Hayakawa, S., Lyons, T., and Oberhauser, H. Estimating the probability that a given vector is in the convex hull of a random sample. *Probability Theory and Related Fields*, 185:705–746, 2023a.

Hayakawa, S., Oberhauser, H., and Lyons, T. Hypercontractivity meets random convex hulls: analysis of randomized multivariate cubatures. *Proceedings of the Royal Society A*, 479(2273):20220725, 2023b.

Huszár, F. and Duvenaud, D. Optimally-weighted herding is Bayesian quadrature. In *Conference on Uncertainty in Artificial Intelligence*, pp. 377–386, 2012.

Jin, R., Yang, T., Mahdavi, M., Li, Y.-F., and Zhou, Z.-H. Improved bounds for the Nyström method with application to kernel classification. *IEEE Transactions on Information Theory*, 59(10):6939–6949, 2013.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3296–3304, 2016.

Karvonen, T., Särkkä, S., and Tanaka, K. Kernel-based interpolation at approximate Fekete points. *Numerical Algorithms*, 87(1):445–468, 2021.

Koltchinskii, V. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Kumar, S., Mohri, M., and Talwalkar, A. Sampling methods for the Nyström method. *The Journal of Machine Learning Research*, 13(1):981–1006, 2012.

Li, C., Jegelka, S., and Sra, S. Fast DPP sampling for Nyström with application to kernel methods. In *International Conference on Machine Learning*, pp. 2061–2070. PMLR, 2016.

Li, M., Bi, W., Kwok, J. T., and Lu, B.-L. Large-scale Nyström kernel matrix approximation using randomized SVD. *IEEE Transactions on Neural Networks and Learning Systems*, 1(26):152–164, 2015.

Litterer, C. and Lyons, T. High order recombination and an application to cubature on Wiener space. *The Annals of Applied Probability*, 22(4):1301–1327, 2012.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Second edition, 2018.

Oglic, D. and Gärtner, T. Nyström method with kernel k-means++ samples as landmarks. In *International Conference on Machine Learning*, pp. 2652–2660. PMLR, 2017.

Owen, A. B. A randomized Halton algorithm in R. *arXiv preprint arXiv:1706.02808*, 2017.

Penrose, R. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

Santin, G. and Haasdonk, B. Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10, 2017.

Santin, G. and Schaback, R. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, 2016.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Shetty, A., Dwivedi, R., and Mackey, L. Distribution compression in near-linear time. In *International Conference on Learning Representations*, 2022.

Shinozaki, N., Sibuya, M., and Tanabe, K. Numerical algorithms for the Moore–Penrose inverse of a matrix: iterative methods. *Annals of the Institute of Statistical Mathematics*, 24(1):621–629, 1972.

Srebro, N. and Sridharan, K. Note on refined Dudley integral covering number bound. *Unpublished results*, 2010. URL https://www.cs.cornell.edu/~sridharan/dudley.pdf.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Steinwart, I. and Scovel, C. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

Tchernychova, M. *Carathéodory cubature measures*. PhD thesis, University of Oxford, 2016.

Tropp, J. A., Yurtsever, A., Udell, M., and Cevher, V. Fixed-rank approximation of a positive-semidefinite matrix from streaming data. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tsuji, K., Tanaka, K., and Pokutta, S. Pairwise conditional gradients without swap steps and sparser kernel herding. In *International Conference on Machine Learning*, pp. 21864–21883. PMLR, 2022.

Wahba, G. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, S., Gittens, A., and Mahoney, M. W. Scalable kernel k-means clustering with Nyström approximation: relative-error bounds. *The Journal of Machine Learning Research*, 20(1):431–479, 2019.

Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., and Zhou, Z.-H. Nyström method vs random Fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

A. Tools from statistical learning theory

In this section, $\mathcal{F}$ always denotes a class of functions from $\mathcal{X}$ to $\mathbb{R}$, i.e., $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$. Let us define the Rademacher complexity of $\mathcal{F}$ with respect to the sample $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$ as follows (e.g., Mohri et al., 2018, Definition 3.1):

$$\mathcal{R}_Z(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{\ell} \sum_{j=1}^{\ell} s_j f(z_j) \,\middle|\, Z\right]$$

where the conditional expectation is taken with regard to the Rademacher variables, i.e., i.i.d. variables s_j uniform in $\{\pm 1\}$.

The following is a version of the uniform law of large numbers, though we only use the one side of the inequality.

Proposition 6 (Mohri et al., 2018, Theorem 3.3)**.** *Let Z be an ℓ-point independent sample from μ. If there is a $B > 0$ such that $\|f\|_\infty \le B$ for every $f \in \mathcal{F}$, then with probability at least $1 - \delta$, we have*

$$\sup_{f \in \mathcal{F}} (\mu(f) - \mu_Z(f)) \le 2\mathbb{E}[\mathcal{R}_Z(\mathcal{F})] + \sqrt{\frac{2B^2}{\ell} \log \frac{1}{\delta}}.$$

For a pseudo metric d on $\mathcal{F}$, we denote the ε-*convering number* of $\mathcal{F}$ by $\mathcal{N}(\mathcal{F}, d; \varepsilon)$. Namely, $\mathcal{N}(\mathcal{F}, d; \varepsilon)$ is the infimum of positive integers N such that there exist $f_1, \dots, f_N \in \mathcal{F}$ satisfying $\min_{1 \le i \le N} d(f_i, g) \le \varepsilon$ for all $g \in \mathcal{F}$.

Let us define a pseudo-metric $d_Z(f, g) := \sqrt{\frac{1}{\ell} \sum_{j=1}^{\ell} (f(z_i) - g(z_i))^2}$. The following assertion is a version of Dudley's integral entropy bound (Srebro et al., 2010, Lemma A.3; see Srebro & Sridharan (2010) for a correction of the constant).

Proposition 7 (Dudley integral)**.** *For any ℓ-point sample $Z = (z_i)_{i=1}^{\ell} \subset \mathcal{X}$, we have*

$$\mathcal{R}_Z(\mathcal{F}) \le \frac{12}{\sqrt{\ell}} \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, d_Z; \varepsilon)} \, \mathrm{d}\varepsilon.$$

The following is a straightforward modification of Schmidt-Hieber (2020, Lemma 4) tailored to our setting. It originates from an analysis of empirical risk minimizers, and this kind of technique has also been known in earlier work under the name of local Rademacher complexities (Györfi et al., 2006; Koltchinskii, 2006; Giné & Koltchinskii, 2006).

Proposition 8. *Let $\mathcal{F} \subset L^\infty(\mu)$ be a set of functions with $f \ge 0$ and $\|f\|_{L^\infty(\mu)} \le F$ for all $f \in \mathcal{F}$, where $F > 0$ is a constant. If $\hat{f}$ is a random function in $\mathcal{F}$ possibly depending on Z, then, for every $\varepsilon > 0$, we have*

$$\mathbb{E}\Big[\mu(\hat{f})\Big] \le 2\mathbb{E}\Big[\mu_Z(\hat{f})\Big] + \frac{F}{\ell}\left(\frac{80}{9}\log N + 64\right) + 5\varepsilon,$$

where $N := \max\{3, \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^1(\mu)}; \varepsilon)\}$.

Proof. The proof here essentially follows the original proof, where we re-compute the constants as the condition is slightly different; see also Hayakawa & Suzuki (2020, Theorem 2.6) and its remark.

Let $Z' = (z_1', \dots z_\ell')$ be an independent copy of Z. Let $\mathcal{F}_\varepsilon$ be an ε-covering of $\mathcal{F}$ in $L^1(\mu)$ with the cardinality N and f^* be a random element of $\mathcal{F}_\varepsilon$ such that $\mu(|\hat{f} - f^*|) \le \varepsilon$. Then, we have

$$\left|\mathbb{E}\Big[\mu_Z(\hat{f})\Big] - \mathbb{E}\Big[\mu(\hat{f})\Big]\right| = \left|\mathbb{E}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}(\hat{f}(z_i) - \hat{f}(z_i'))\right]\right| \le \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{\ell}(f^*(z_i) - f^*(z_i'))\right|\right] + 2\varepsilon \tag{15}$$

Define $T := \max_{f \in \mathcal{F}_\varepsilon} \sum_{i=1}^{\ell} (f(z_i) - f(z_i'))/r(f)$, where we let $r(f) := \max\{c\sqrt{\ell^{-1}\log N}, \sqrt{\mu(f)}\}$ for each $f \in \mathcal{F}_\varepsilon$ with a constant $c > 0$ fixed afterwards. Thus, we obtain

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{\ell}(f^*(z_i) - f^*(z_i'))\right|\right] \le \mathbb{E}\left[\frac{r(f^*)T}{\ell}\right] \le \frac{1}{2}\mathbb{E}\big[r(f^*)^2\big] + \frac{1}{2\ell^2}\mathbb{E}\big[T^2\big]\,. \tag{16}$$

The first term can evaluated as

$$\mathbb{E}\big[r(f^*)^2\big] \le c^2\frac{\log N}{\ell} + \mathbb{E}[\mu(f^*)] \le c^2\frac{\log N}{\ell} + \mathbb{E}[\mu(\hat{f})] + \varepsilon. \tag{17}$$

For the second term, we first have

$$\sum_{i=1}^{\ell}\mathbb{E}\left[\left(\frac{f(z_i)-f(z_i')}{r(f)}\right)^2\right] \le \sum_{i=1}^{\ell}\mathbb{E}\left[\frac{f(z_i)^2+f(z_i')^2}{r(f)^2}\right] \le 2F\ell, \qquad f\in\mathcal{F}_\varepsilon.$$

Since we have $|f(z_i)-f(z_i')|/r(f) \le 2F/r(f) \le 2F\frac{\sqrt{\ell}}{c\sqrt{\log N}}$ uniformly for $f\in\mathcal{F}_\varepsilon$, Bernstein's inequality combined with the union bound yields

$$\mathbb{P}\big(T^2\ge t\big) = \mathbb{P}\Big(T\ge\sqrt{t}\Big) \le 2N\exp\left(-\frac{t}{4F(\ell+\frac{\sqrt{\ell t}}{3c\sqrt{\log N}})}\right) \le 2N\exp\left(-\frac{3c\sqrt{\log N}}{8F\sqrt{\ell}}\sqrt{t}\right)$$

for $t\ge 9c^2\ell\log N$. Therefore, we have

$$\begin{aligned}\mathbb{E}\big[T^2\big] = \int_0^\infty \mathbb{P}\big(T^2\ge t\big)\,\mathrm{d}t &\le 9c^2\ell\log N + \int_{9c^2\ell\log N}^{\infty} 2N\exp\left(-\frac{3c\sqrt{\log N}}{8F\sqrt{\ell}}\sqrt{t}\right)\mathrm{d}t\\ &= 9c^2\ell\log N + 4N\left(8F\ell + \frac{64F^2\ell}{9c^2\log N}\right)\exp\left(-\frac{9c^2\log N}{8F}\right)\end{aligned}$$

Let us now set $c=\sqrt{8F/9}$ so that $9c^2=8F$. Then, we obtain $\mathbb{E}\big[T^2\big]\le 8F\ell\log N + 64F\ell$ since $N\ge 3$ by assumption. By combining it with (15)–(17), we finally obtain

$$\left|\mathbb{E}\Big[\mu_Z(\hat{f})\Big] - \mathbb{E}\Big[\mu(\hat{f})\Big]\right| \le \frac{1}{2}\mathbb{E}\Big[\mu(\hat{f})\Big] + \frac{(\frac{40}{9}F\log N + 32F)}{\ell} + \frac{5}{2}\varepsilon,$$

from which the desired inequality readily follows. □

B. Proofs

B.1. Properties of the pseudo-inverse

For a matrix $A\in\mathbb{R}^{m\times n}$, its Moore–Penrose pseudo-inverse A^+ (Penrose, 1955) is defined as the unique matrix $X\in\mathbb{R}^{n\times m}$ that satisfies

$$AXA = A,\quad XAX = X,\quad (AX)^\top = AX,\quad (XA)^\top = XA.$$

It also satisfies that A^+A is the orthogonal projection onto the orthogonal complement of $\ker A$ (the range of $A^\top$), while AA^+ is the orthogonal projection onto the range of A (Penrose, 1955; Shinozaki et al., 1972). We use these general properties of A^+ throughout Section B. See e.g. Drineas et al. (2005) for the concrete construction of such a matrix.

B.2. Proof of Lemma 1

Proof. Recall that we have the SVD $k(Z,Z) = U\operatorname{diag}(\lambda_1,\dots,\lambda_\ell)U^\top$ with an orthogonal matrix $U=[u_1,\dots,u_\ell]$. and $\lambda_1\ge\cdots\ge\lambda_\ell\ge 0$. By using this notation, we have

$$k_s^Z(x,y) = \sum_{\substack{1\le j\le s\\ \lambda_j>0}}\frac{1}{\lambda_j}(u_j^\top k(Z,x))(u_j^\top k(Z,y)). \tag{18}$$

If we denote by $Q_j:\mathcal{H}_k\to\mathcal{H}_k$ the projection onto $\operatorname{span}\{u_j^\top k(Z,\cdot)\}$, we have

$$\begin{aligned}(u_j^\top k(Z,x))(u_j^\top k(Z,y)) &= \big\langle u_j^\top k(Z,\cdot), k(\cdot,x)\big\rangle_{\mathcal{H}_k}\big\langle u_j^\top k(Z,\cdot), k(\cdot,y)\big\rangle_{\mathcal{H}_k}\\ &= \|u_j^\top k(Z,\cdot)\|_{\mathcal{H}_k}^2\left\langle Q_jk(\cdot,x), Q_jk(\cdot,y)\right\rangle_{\mathcal{H}_k}\\ &= \lambda_j\left\langle Q_jk(\cdot,x), Q_jk(\cdot,y)\right\rangle_{\mathcal{H}_k},\end{aligned} \tag{19}$$

where the last inequality follows from $\left\langle u_i^\top k(Z,\cdot), u_j^\top k(Z,\cdot)\right\rangle_{\mathcal{H}_k} = u_i^\top k(Z,Z)u_j = \delta_{ij}\lambda_j$. Now let $\widetilde{P}_{Z,s}$ be the orthogonal projection onto $\operatorname{span}\{u_j^\top k(Z,\cdot)\}_{j=1}^s$ in $\mathcal{H}_k$. We prove $\widetilde{P}_{Z,s} = P_{Z,s}$. From the orthogonality of $\{u_j^\top k(Z,\cdot)\}_{j=1}^s$ we have $\widetilde{P}_{Z,s} = \sum_{j=1}^s Q_j$ and

$$\begin{aligned}\left\langle k(\cdot,x), k_s^Z(\cdot,y)\right\rangle_{\mathcal{H}_k} = k_s^Z(x,y) &= \sum_{j=1}^s \langle Q_j k(\cdot,x), Q_j k(\cdot,y)\rangle_{\mathcal{H}_k} \\ &= \left\langle \widetilde{P}_{Z,s}k(\cdot,x), \widetilde{P}_{Z,s}k(\cdot,y)\right\rangle_{\mathcal{H}_k} = \left\langle k(\cdot,x), \widetilde{P}_{Z,s}k(\cdot,y)\right\rangle_{\mathcal{H}_k}\end{aligned}$$

for all $x, y \in \mathcal{X}$. In particular, $k_s^Z(\cdot,y) = \widetilde{P}_{Z,s}k(\cdot,y)$, so we have $\widetilde{P}_{Z,s} = P_{Z,s}$. □

B.3. Proof of Lemma 2

Proof. The inequality follows from Cauchy–Schwarz. Let us prove the equality.

We use the notation Q_j from the proof of Lemma 1. We first obtain $P_Z k(\cdot,z_i) = k(\cdot,z_i)$ for $i = 1,\dots,\ell$, since P_Z is a projection onto $\operatorname{span}\{k(\cdot,z_i)\}_{i=1}^\ell$. Thus, we have $P_{Z,s}^\perp k(\cdot,z_i) = (P_Z - P_{Z,s})k(\cdot,z_i) = (Q_{s+1}+\cdots+Q_\ell)k(\cdot,z_i)$, and so

$$\frac{1}{\ell}\sum_{i=1}^\ell \|P_{Z,s}^\perp k(\cdot,z_i)\|_{\mathcal{H}_k}^2 = \frac{1}{\ell}\sum_{i=1}^\ell \sum_{\substack{s+1\le j\le \ell\\ \lambda_j>0}} \frac{1}{\lambda_j}(u_j^\top k(Z,z_i))^2$$

by using (19). Since $k(Z,Z) = U\operatorname{diag}(\lambda_1,\dots,\lambda_\ell)U^\top = \sum_{i=1}^\ell \lambda_i u_i u_i^\top$, we can explicitly calculate

$$u_j^\top k(Z,z_i) = u_j^\top \sum_{i=1}^\ell \lambda_i u_i u_i^\top \mathbf{1}_j = \lambda_j u_i^\top \mathbf{1}_j,$$

where $\mathbf{1}_j \in \mathbb{R}^\ell$ is the vector with 1 in the j-th coordinate and 0 in the other coordinates. As U is an $\ell\times\ell$ orthogonal matrix, we actually have $\sum_{i=1}^\ell (u_i^\top \mathbf{1}_j)^2 = 1$ for each $j = 1,\dots,\ell$.

$$\frac{1}{\ell}\sum_{i=1}^\ell \sum_{\substack{s+1\le j\le \ell\\ \lambda_j>0}} \frac{1}{\lambda_j}(u_j^\top k(Z,z_i))^2 = \frac{1}{\ell}\sum_{i=1}^\ell\sum_{j=s+1}^\ell \lambda_j (u_i^\top \mathbf{1}_j)^2 = \frac{1}{\ell}\sum_{j=s+1}^\ell \lambda_j, \tag{20}$$

and the proof is complete. □

B.4. Proof of Lemma 3

Proof. From the min-max principle, we have

$$\lambda_j = \min_{\substack{V_{j-1}\subset\mathbb{R}^\ell\\ \dim V_{j-1}\le j-1}} \max_{x_j\in V_{j-1}^\perp,\ \|x_j\|_2=1} x_j^\top k(Z,Z)x_j, \tag{21}$$

where V_{j-1} is a linear subspace of $\mathbb{R}^\ell$. Recall the Mercer expansion $k(x,y) = \sum_{i=1}^\infty \sigma_i e_i(x)e_i(y)$. By letting $e_j(Z) = (e_j(z_1),\dots,e_j(z_\ell))^\top \in \mathbb{R}^\ell$, we can write $k(Z,Z) = \sum_{i=1}^\infty \sigma_i e_i(Z)e_i(Z)^\top$. We assume that this equality holds in the following. We especially write the remainder term as $k_{s+1}(Z,Z) := k(Z,Z) - \sum_{i=1}^s \sigma_i e_i(Z)e_i(Z)^\top$

Consider taking $V_s = \operatorname{span}\{e_1(Z),\dots,e_s(Z)\}$ and

$$x_j \in \operatorname*{argmax}_{x\in V_{j-1}^\perp,\ \|x\|_2=1} x^\top k(Z,Z)x, \qquad V_j = \operatorname{span}(V_{j-1}\cup\{x_j\})$$

for $j = s+1,\dots,\ell$ in (21). Then, $\lambda_j' := x_j^\top k(Z,Z)x$ satisfies $\lambda_j \le \lambda_j'$, and so we have

$$\sum_{j=s+1}^\ell \lambda_j \le \sum_{j=s+1}^\ell \lambda_k' = \sum_{j=s+1}^\ell x_j^\top k(Z,Z)x_j = \sum_{j=s+1}^\ell x_j^\top k_{s+1}(Z,Z)x_j,$$

where we have used that $x_j^\top e_i(Z) = 0$ for any $i \le s < j$ in the last inequality. By taking some $\{x_1, \dots, x_s\} \subset \mathbb{R}^\ell$, we can make $\{x_1, \dots, x_\ell\}$ a orthonormal basis of $\mathbb{R}^\ell$, so we obtain

$$\sum_{j=s+1}^{\ell} \lambda_j \le \sum_{j=s+1}^{\ell} x_j^\top k_{s+1}(Z, Z) x_j \le \sum_{j=1}^{\ell} x_j^\top k_{s+1}(Z, Z) x_j = \operatorname{tr} k_{s+1}(Z, Z).$$

Therefore, we have

$$\frac{1}{\ell} \sum_{j=s+1}^{\ell} \lambda_j \le \frac{1}{\ell} \operatorname{tr} k_{s+1}(Z, Z) = \frac{1}{\ell} \sum_{i=1}^{\ell} k_{s+1}(z_i, z_i),$$

and we obtain the desired inequality in expectation since $\mathbb{E}[k_{s+1}(z_i, z_i)] = \sum_{j=s+1}^{\infty} \sigma_j$. $\square$

B.5. Proof of Theorem 1

We first prove the following generic proposition by exploiting the ingredients given in Section A.

Proposition 9. *Let Q be an arbitrary deterministic m-dimensional orthogonal projection in $\mathcal{H}_k$. Then, for any random orthogonal projection P possibly depending on Z, we have*

$$\mu(\|PQk(\cdot, x)\|_{\mathcal{H}_k}) \le \mu_Z(\|PQk(\cdot, x)\|_{\mathcal{H}_k}) + \sqrt{\frac{k_{\max}}{\ell}} \left(36m + \sqrt{2 \log \frac{1}{\delta}} \right) \tag{22}$$

with probability at least $1 - \delta$.

Furthermore, with regard to the expectation, we also have

$$\mathbb{E}[\mu(\|PQk(\cdot, x)\|_{\mathcal{H}_k})] \le 2\mathbb{E}[\mu_Z(\|PQk(\cdot, x)\|_{\mathcal{H}_k})] + \frac{\sqrt{k_{\max}}}{\ell} \left(\frac{80m^2 \log(1 + 2\ell)}{9} + 69 \right). \tag{23}$$

Proof. Let $\{v_1, \dots, v_m\}$ be an orthonormal basis of $Q\mathcal{H}_k$. Let also $\{u_i\}_{i \in I}$ and $\{u_i\}_{i \in J}$ be respectively an orthonormal basis of $P\mathcal{H}_k$ and $(P\mathcal{H}_k)^\perp$, so $\{u_i\}_{i \in I \cup J}$ is an orthonormal basis of $\mathcal{H}_k$.

Let us compute $\|PQk(\cdot, x)\|_{\mathcal{H}_k}^2$. Since we have

$$PQk(\cdot, x) = P\left(\sum_{j=1}^{m} \langle v_j, k(\cdot, x) \rangle_{\mathcal{H}_k} v_j \right) = \sum_{j=1}^{m} v_j(x) P v_j = \sum_{i \in I} \sum_{j=1}^{m} v_j(x) \langle u_i, v_j \rangle_{\mathcal{H}_k} u_i$$

(where we can exchange the summation as they converge in $\mathcal{H}_k$), we obtain

$$\|PQk(\cdot, x)\|_{\mathcal{H}_k}^2 = \sum_{i \in I} \left(\sum_{j=1}^{m} v_j(x) \langle u_i, v_j \rangle_{\mathcal{H}_k} \right)^2 = \|A_{P,Q} \boldsymbol{v}_x\|_{\ell^2(I)}^2 = \boldsymbol{v}_x^\top A_{P,Q}^* A_{P,Q} \boldsymbol{v}_x,$$

where $\boldsymbol{v}_x = (v_1(x), \dots, v_m(x))^\top \in \mathbb{R}^m$ and $A_{P,Q}$ is a linear operator $\mathbb{R}^m \to \ell^2(I)$ given by $a = (a_1, \dots, a_m)^\top \mapsto (\sum_{j=1}^{m} \langle u_i, v_j \rangle_{\mathcal{H}_k} a_j)_{i \in I}$, and $A_{P,Q}^* : \ell^2(I) \to \mathbb{R}^m$ is its dual (defined by the property $\langle a, A_{P,Q}^* b \rangle_{\mathbb{R}^m} = \langle A_{P,Q} a, b \rangle_{\ell^2(I)}$), which can be understood as the "transpose" of $A_{P,Q}$. Note that $A_{P,Q}^* A_{P,Q}$ can be regarded as an $m \times m$ matrix and we have

$$(A_{P,Q}^* A_{P,Q})_{j,h} = \sum_{i \in I} \langle u_i, v_j \rangle_{\mathcal{H}_k} \langle u_i, v_h \rangle_{\mathcal{H}_k} = \langle P v_j, P v_h \rangle_{\mathcal{H}_k}.$$

We can also define $B_{P,Q} = A_{P^\perp, Q}$ by replacing P with $P^\perp$. Then we have

$$(A_{P,Q}^* A_{P,Q})_{j,h} + (B_{P,Q}^* B_{P,Q})_{j,h} = \langle P v_j, P v_h \rangle_{\mathcal{H}_k} + \langle P^\perp v_j, P^\perp v_h \rangle_{\mathcal{H}_k} = \langle v_j, v_h \rangle_{\mathcal{H}_k} = \delta_{jh},$$

so $A_{P,Q}^\top A_{P,Q}$ is an $m \times m$ positive semi-definite matrix with $A_{P,Q}^\top A_{P,Q} \le I_m$.

It thus suffices to consider a uniform estimate of $\mu(\sqrt{\boldsymbol{v}_x^\top S \boldsymbol{v}_x}) - \mu_Z(\sqrt{\boldsymbol{v}_x^\top S \boldsymbol{v}_x})$ with a positive semi-definite matrix $S \le I_m$. This S can be written as $S = U^\top U$ by using a $U \in \mathbb{R}^{m \times m}$ with $\|U\|_2 \le 1$, so we shall solve the following problem:

Find a uniform upper bound of $\mu(\|U\boldsymbol{v}_x\|_2) - \mu_Z(\|U\boldsymbol{v}_x\|_2)$ for any matrix $U \in \mathbb{R}^{m\times m}$ with $\|U\|_2 \le 1$.

Now we can reduce our problem to a routine work of bounding the covering number of the function class $\mathcal{F} := \{f_U := x \mapsto \|U\boldsymbol{v}_x\|_2 \mid U \in \mathcal{U}\}$, where $\mathcal{U} := \{U \in \mathbb{R}^{m\times m} \mid \|U\|_2 \le 1\}$.

For any $x \in \mathcal{X}$, we have

$$\|\boldsymbol{v}_x\|_2^2 = \sum_{j=1}^{\ell} v_j(x)^2 = \|Qk(\cdot,x)\|_{\mathcal{H}_k}^2 \le \|k(\cdot,x)\|_{\mathcal{H}_k}^2 = k(x,x).$$

If $\mathcal{U}_\delta$ is a δ-covering of $\mathcal{U}$, then $\{f_U\}_{U\in\mathcal{U}_\delta}$ gives a $\delta\sqrt{k_{\max}}$-covering. Indeed, for any $U, V \in \mathcal{U}$ with $\|U - V\|_2 \le \delta$, we have

$$d_Z(f_U, f_V)^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}(\|U\boldsymbol{v}_{z_i}\|_2 - \|V\boldsymbol{v}_{z_i}\|_2)^2 \le \frac{1}{\ell}\sum_{i=1}^{\ell}\|(U-V)\boldsymbol{v}_{z_i}\|_2^2 \le \delta^2\frac{1}{\ell}\sum_{i=1}^{\ell}\|\boldsymbol{v}_{z_i}\|_2^2 \le \delta^2 k_{\max}.$$

Here, we have the covering number bound $\log\mathcal{N}(\mathcal{U}, \|\cdot\|_2; \delta) \le m^2 \log\left(1 + \frac{2}{\delta}\right)$ for $\delta \le 1$ (and 0 for $\delta \ge 1$) as $\mathcal{U}$ can be seen as a unit ball of $\mathbb{R}^{m^2}$ in a certain norm (Wainwright, 2019, Example 5.8), so $\log\mathcal{N}(\mathcal{F}, d_Z; \varepsilon) \le m^2\log(1 + 2\sqrt{k_{\max}}/\varepsilon)$ for $\varepsilon \le \sqrt{k_{\max}}$.

Therefore, from Proposition 7, we have

$$\begin{aligned}\mathcal{R}_Z(\mathcal{F}) &\le \frac{12}{\sqrt{\ell}}\int_0^{\sqrt{k_{\max}}}\sqrt{m^2\log\left(1 + \frac{2\sqrt{k_{\max}}}{\varepsilon}\right)}\,\mathrm{d}\varepsilon \\ &= \frac{12m\sqrt{k_{\max}}}{\sqrt{\ell}}\int_0^1\sqrt{\log\left(1+\frac{2}{t}\right)}\,\mathrm{d}t \le \frac{18m\sqrt{k_{\max}}}{\sqrt{\ell}},\end{aligned}$$

where we have used the estimate

$$\int_0^1\sqrt{\log\left(1+\frac{2}{t}\right)}\,\mathrm{d}t \le \int_0^1\frac{1}{2}\left(1+\log\left(1+\frac{2}{t}\right)\right)\,\mathrm{d}t = \frac{1}{2}+\frac{1}{2}\log\frac{27}{4} \le \frac{3}{2}.$$

Since we also have a bound $\|f_U\|_\infty \le \|U\|_2\sqrt{k_{\max}}$, we can use Proposition 6 to obtain

$$\mu(\|PQk(\cdot,x)\|_{\mathcal{H}_k}) - \mu_Z(\|PQk(\cdot,x)\|_{\mathcal{H}_k}) \le \sup_{f\in\mathcal{F}}(\mu_Z(f) - \mu(f)) \le \sqrt{\frac{k_{\max}}{\ell}}\left(36m + \sqrt{2\log\frac{1}{\delta}}\right)$$

with probability at least $1-\delta$. So we have proven (22).

We next prove (23) by using Proposition 8. We have the same bound for $\log\mathcal{N}(\mathcal{F}, \|\cdot\|_{L^1}(\mu); \varepsilon)$ from the same argument as above, and so we especially get

$$\log N\left(\mathcal{F}, \|\cdot\|_{L^1(\mu)}; \frac{\sqrt{k_{\max}}}{\ell}\right) \le m^2\log(1+2\ell).$$

As $\|f\|_{L^\infty(\mu)} \le \sqrt{k_{\max}} =: F$ holds for all $f \in \mathcal{F}$, we can now apply Proposition 8 with $\varepsilon = F/\ell$ to obtain the desired conclusion. □

We next prove the following proposition that includes the desired assertion by using Proposition 9.

Proposition 10. *Let $Z = (z_i)_{i=1}^{\ell}$ be an ℓ-point independent sample from μ. Let P be a random orthogonal projection in $\mathcal{H}_k$ possibly depending on Z. For any integer $m \ge 1$, with probability at least $1-\delta$, we have*

$$\int_{\mathcal{X}}\|Pk(\cdot,x)\|_{\mathcal{H}_k}\,\mathrm{d}\mu(x) \le \frac{1}{\ell}\sum_{i=1}^{\ell}\|Pk(\cdot,z_i)\|_{\mathcal{H}_k} + \sqrt{\frac{k_{\max}}{\ell}}\left(36m + \sqrt{\frac{9}{2}\log\frac{2}{\delta}}\right) + 3\sqrt{\sum_{j>m}\sigma_j}.$$

Furthermore, in expectation, we have the following bound:

$$\mathbb{E}\left[\int_{\mathcal{X}} \|Pk(\cdot, x)\|_{\mathcal{H}_k} \,\mathrm{d}\mu(x)\right] \le \mathbb{E}\left[\frac{2}{\ell}\sum_{i=1}^{\ell} \|Pk(\cdot, z_i)\|_{\mathcal{H}_k}\right] + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2\log(1+2\ell)}{9} + 69\right) + 4\sqrt{\sum_{j>m}\sigma_j}. \tag{24}$$

Proof. Note that we use the fact that for any projection operator P $\|Pf\| \le \|f\|$ frequently within the proof. For an ℓ-point sample $Z = (z_1, \dots, z_\ell) \subset \mathcal{X}$, let us denote μ_Z be the mapping $f \mapsto \frac{1}{\ell}\sum_{i=1}^{\ell} f(z_i)$. If we have $f_-, f \in L^1(\mu)$ with $f_- \le f$, we can generally obtain

$$\begin{aligned}\mu(f) - \mu_Z(f) &= (\mu(f) - \mu(f_-)) + (\mu(f_-) - \mu_Z(f_-)) + (\mu_Z(f_-) - \mu_Z(f)) \\ &\le \mu(f - f_-) + (\mu(f_-) - \mu_Z(f_-)).\end{aligned} \tag{25}$$

We here use $f(x) = \|Pk(\cdot, x)\|_{\mathcal{H}_k}$ and $f_-(x) = \|PP_mk(\cdot, x)\|_{\mathcal{H}_k} - \|PP_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}$ for an m, where P_m is the projection operator onto $\operatorname{span}\{e_1, \dots, e_m\}$ in $\mathcal{H}_k$ and $P_m^\perp$ is its orthogonal complement. In this case, $\mu(f - f_-)$ can easily be estimated by Cauchy–Schwarz as follows:

$$\begin{aligned}\mu(f - f_-) &\le \mu(2\|PP_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}) \le 2\mu(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}) \\ &\le 2\sqrt{\mu(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}^2)} = 2\sqrt{\sum_{j>m}\sigma_j},\end{aligned} \tag{26}$$

where we have used the fact

$$\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}^2 = \|k(\cdot, x)\|_{\mathcal{H}_k}^2 - \|P_mk(\cdot, x)\|_{\mathcal{H}_k}^2 = k(x, x) - \sum_{i=1}^{m}\sigma_i e_i(x)^2 = \sum_{i=m+1}^{\infty}\sigma_i e_i(x)^2.$$

We also bound $\mu(f_-) - \mu_Z(f_-)$ by

$$\mu(f_-) - \mu_Z(f_-) \le \mu(\|PP_mk(\cdot, x)\|_{\mathcal{H}_k}) - \mu_Z(\|PP_mk(\cdot, x)\|_{\mathcal{H}_k}) + \mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}), \tag{27}$$

where we have used the second inequality in (26) for μ_Z. The last term $\mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k})$ above is estimated either in expectation or in high probability as follows:

$$\begin{cases}\mathbb{E}\big[\mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k})\big] \le \sqrt{\sum_{j>m}\sigma_j}. \\ \mu_Z(\|P_m^\perp k(\cdot, x)\|_{\mathcal{H}_k}) \le \sqrt{\sum_{j>m}\sigma_j} + \sqrt{\frac{k_{\max}}{2\ell}\log\frac{1}{\delta}} \text{ with probability at least } 1 - \delta.\end{cases} \tag{28}$$

The latter follows from a simple calculation of Hoeffing's inequality.

Thus, it suffices to derive a bound for $\mu(\|PP_mk(\cdot, x)\|_{\mathcal{H}_k}) - \mu_Z(\|PP_mk(\cdot, x)\|_{\mathcal{H}_k})$ or its expectation; we do it by letting $Q = P_m$ and $\hat{f} = f$ in Proposition 9. By combining (just summing up) the inequalities (25)–(28), and (22), we obtain the desired inequality in high probability. For the result in expectation, we first combine the inequalities (25)–(28), and (23) to get the bound

$$\mathbb{E}[\mu(f)] - \mathbb{E}[\mu_Z(f)] \le \mathbb{E}[\mu_Z(\|PP_mk(\cdot, x)\|_{\mathcal{H}_k})] + \frac{\sqrt{k_{\max}}}{\ell}\left(\frac{80m^2\log(1+2\ell)}{9} + 69\right) + 3\sqrt{\sum_{j>m}\sigma_j}$$

(recall $f(x) = \|Pk(\cdot, x)\|_{\mathcal{H}_k}$). Since we can also estimate $\mathbb{E}[\mu_Z(\|PP_mk(\cdot, x)\|_{\mathcal{H}_k})]$ as

$$\begin{aligned}\mathbb{E}[\mu_Z(\|PP_mk(\cdot, x)\|_{\mathcal{H}_k})] &\le \mathbb{E}[\mu_Z(\|Pk(\cdot, x)\|_{\mathcal{H}_k})] + \mathbb{E}\big[\mu_Z(\|PP_m^\perp k(\cdot, x)\|_{\mathcal{H}_k})\big] \\ &\le \mathbb{E}[\mu_Z(\|Pk(\cdot, x)\|_{\mathcal{H}_k})] + \sqrt{\sum_{j>m}\sigma_j},\end{aligned}$$

we obtain the desired conclusion. $\square$

B.6. Proof of Remark 1

Proof. We assume $\ell \geq 3$ here. Let $F(x) := -\beta^{-1}x^{1-1/d}\exp(-\beta x^{1/d})$. If $d \geq 2$, its derivative is

$$F'(x) = \exp(-\beta x^{1/d}) - \frac{1-1/d}{\beta}x^{-1/d}\exp(-\beta x^{1/d}) = \left(1 - \frac{1-1/d}{\beta}x^{-1/d}\right)\exp(-\beta x^{1/d}).$$

Thus, if $x \geq (\log \ell)^d/\beta^d$, we have $F'(x) \geq d\exp(-\beta x^{1/d})$. This inequality is still true if $d = 1$. By taking $m = \lfloor (2\log \ell)^d/\beta^d \rfloor$, we obtain

$$\sum_{i>m}\sigma_i \lesssim \int_{2(\log \ell)^d/\beta^d}^{\infty} \exp(-\beta x^{1/d})\,\mathrm{d}x \leq -dF(2(\log \ell)^d/\beta^d) = \frac{2^{d-1}d}{\beta^d}\cdot\frac{(\log \ell)^{d-1}}{\ell^2}.$$

Therefore, this choice of m satisfies

$$\sqrt{\sum_{i>m}\sigma_i} = \mathcal{O}\left(\frac{(\log \ell)^{(d-1)/2}}{\ell}\right), \qquad m^2 = \mathcal{O}\big((\log \ell)^{2d}\big)\,.$$

Combining these with the inequality in Corollary 1 gives the desired estimate. □

B.7. Proof of Proposition 1

Proof. We basically just compute the trace of the operator $P_Z^\perp \mathcal{K}$. Indeed, we have

$$\int_{\mathcal{X}} \|P_Z^\perp k(\cdot, x)\|_{\mathcal{H}_k}^2 = \int_{\mathcal{X}} (k(x,x) - k^Z(x,x))\,\mathrm{d}\mu(x), \tag{29}$$

and, from (5), we also have the following identity:

$$\int_{\mathcal{X}} k(x,x)\,\mathrm{d}\mu(x) = \sum_{i=1}^{\infty}\langle e_i, \mathcal{K}e_i\rangle_{L^2(\mu)}\,. \tag{30}$$

For k^Z, as we can write $k^Z(x,y) = \sum_{i=1}^{\ell} g_i(x)g_i(y)$ by using $g_i \in L^2(\mu)$ (see e.g., (18)), we can also have

$$\int_{\mathcal{X}} k^Z(x,x)\,\mathrm{d}\mu(x) = \sum_{i\in I}\left\langle e_i, \mathcal{K}^Z e_i\right\rangle_{L^2(\mu)} = \sum_{i=1}^{\infty}\left\langle e_i, \mathcal{K}^Z e_i\right\rangle_{L^2(\mu)}, \tag{31}$$

where $\mathcal{K}^Z : L^2(\mu) \to L^2(\mu)$ is the integral operator given by $g \mapsto \int_{\mathcal{X}} k^Z(\cdot, x)g(x)\,\mathrm{d}\mu(x)$, and $(e_i)_{i\in I}$ is an orthonormal basis of $L^2(\mu)$ including $(e_i)_{i=1}^{\infty}$. The second equality follows from the fact that $\mathcal{K} - \mathcal{K}^Z$ is a (semi-)positive definite operator since $k - k^Z$ is a positive definite kernel, and so we have $0 \leq \left\langle e_i, \mathcal{K}^Z e_i\right\rangle_{L^2(\mu)} \leq \langle e_i, \mathcal{K}e_i\rangle_{L^2(\mu)} = 0$ for any $i \in I \setminus \mathbb{Z}_{>0}$. For this integral operator, since we have $k^Z(\cdot, x) = P_Z k(\cdot, x)$, we can prove

$$\mathcal{K}^Z g = \int_{\mathcal{X}} P_Z k(\cdot,x)g(x)\,\mathrm{d}\mu(x) = P_Z\int_{\mathcal{X}} k(\cdot,x)g(x)\,\mathrm{d}\mu(x) = P_Z\mathcal{K}g$$

for any $g \in L^2(\mu)$ under the well-definedness of $\mathcal{K}$. Thus, from (29)–(31), we have

$$\int_{\mathcal{X}} \|P_Z^\perp k(\cdot,x)\|_{\mathcal{H}_k}^2 = \sum_{i=1}^{\infty}\left\langle e_i, (\mathcal{K} - \mathcal{K}^Z)e_i\right\rangle_{L^2(\mu)} = \sum_{i=1}^{\infty}\left\langle e_i, P_Z^\perp\mathcal{K}e_i\right\rangle_{L^2(\mu)}\,. \tag{32}$$

For general $f \in \mathcal{H}_k$ and $g \in L^2(\mu)$, we can prove

$$\langle f, \mathcal{K}g\rangle_{\mathcal{H}_k} = \left\langle f, \int_{\mathcal{X}} k(\cdot,x)g(x)\,\mathrm{d}\mu(x)\right\rangle_{\mathcal{H}_k} = \int_{\mathcal{X}}\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k} g(x)\,\mathrm{d}\mu(x) = \langle f, g\rangle_{L^2(\mu)}\,,$$

so that in particular

$$\left\langle g, P_Z^\perp\mathcal{K}g\right\rangle_{L^2(\mu)} = \left\langle \mathcal{K}g, P_Z^\perp\mathcal{K}g\right\rangle_{\mathcal{H}_k} = \|P_Z^\perp\mathcal{K}g\|_{\mathcal{H}_k}^2.$$

By letting $g = e_i$ in the above equation, we can deduce the desired equality from (32). For the inequality, use the bound

$$\|P_Z^\perp\mathcal{K}e_i\|_{\mathcal{H}_k}^2 \leq \|\mathcal{K}e_i\|_{\mathcal{H}_k}^2 = \|\sigma_i e_i\|_{\mathcal{H}_k}^2 = \sigma_i\|\sqrt{\sigma_i}e_i\|_{\mathcal{H}_k}^2 = \sigma_i$$

for each $i > m$. □

B.8. Proof of Corollary 2

Proof. From Proposition 1 and (8), it suffices to prove for an arbitrary $g \in L^2(\mu)$ that

$$\|P_Z^\perp \mathcal{K} g\|_{\mathcal{H}_k}^2 = \inf_{w_i} \sup_{\|f\|_{\mathcal{H}_k} \le 1} \left| \mu(fg) - \sum_{i=1}^{\ell} w_i f(z_i) \right|^2 \le 4 \sum_{i>\ell} \sigma_i.$$

It is indeed an immediate consequence of Belhadji (2021, Theorem 4). □

B.9. Proof of Lemma 4

Proof. Given the Mercer decomposition $k(x, y) = \sum_{i=1}^\infty \sigma_i e_i(x) e_i(y)$, we can compute

$$\begin{aligned} h_\mu(x, y) &= \int_{\mathcal{X}} k(x, t) k(t, y) \, \mathrm{d}\mu(t) \\ &= \sum_{i,j=1}^{\infty} \sigma_i \sigma_j e_i(x) e_j(y) \int_{\mathcal{X}} e_i(t) e_i(t) \, \mathrm{d}\mu(t) \\ &= \sum_{i,j=1}^{\infty} \delta_{ij} \sigma_i \sigma_j e_i(x) e_j(y) = \sum_{i=1}^{\infty} \sigma_i^2 e_i(x) e_i(y), \end{aligned}$$

where we have used the fact that $(e_i)_{i=1}^\infty$ is an orthonormal set in $L^2(\mu)$. □

B.10. Proof of Lemma 5

Proof. From (9), we have

$$\langle f_i, f_j \rangle_{L^2(\mu)} = v_i^\top (H^+)^\top H^\top H H^+ v_j = (HH^+ v_i)^\top (HH^+ v_j). \tag{33}$$

Here, note that $\{v_i,\ \kappa_i > 0\} \subset (\ker H^\top)^\perp$ as we have, for any $v \in \ker H^\top$,

$$0 = v^\top H k(Z, Z)^+ H^\top v = \sum_{i=1}^{\ell} \kappa_i v^\top v_i v_i^\top v = \sum_{i=1}^{\ell} \kappa_i (v^\top v_i)^2.$$

Therefore, $HH^+ v_i = v_i$ if $\kappa_i > 0$ since HH^+ is the projection onto $(\ker H^\top)^\perp$, and so $\{f_i,\ \kappa_i > 0\}$ is orthonormal from (33). We can also see that $f_i = (H^+ v_i)^\top k(Z, \cdot)$ is an eigenfunction of $\mathcal{K}^Z$ from the remark below (10) and $HH^+ v_i = v_i$. □

B.11. Proof of Proposition 2

Proof. We rewrite k_μ^Z in terms of another summation as follows:

$$\begin{aligned} k_\mu^Z(x, y) &:= \sum_{i=1}^{\ell} \kappa_i f_i(x) f_i(y) \\ &= k(x, Z) H^+ \left(\sum_{i=1}^{\ell} \kappa_i v_i v_i^\top \right) (H^\top)^+ k(Z, y) \\ &= k(x, Z) H^+ H k(Z, Z)^+ H^\top (H^\top)^+ k(Z, y) \\ &= \sum_{\lambda_i > 0} \frac{1}{\lambda_i} u_i^\top H^\top (H^+)^\top k(Z, x) k(y, Z) H^+ H u_i, \end{aligned} \tag{34}$$

where (λ_i, u_i) are eigenpairs of $k(Z, Z)$. Recall also that we have

$$k^Z(x, y) = k(x, Z) k(Z, Z)^+ k(Z, y) = \sum_{\lambda_i > 0} \frac{1}{\lambda_i} u_i^\top k(Z, x) k(y, Z) u_i. \tag{35}$$

From (34) and this, it suffices to prove $u^\top k(Z,\cdot) = u^\top H^\top (H^+)^\top k(Z,\cdot)$ in $L^2(\mu)$ for any $u \in \mathbb{R}^\ell$. Indeed, we have

$$\begin{aligned}
&\int_{\mathcal{X}} \left(u^\top k(Z,x) - u^\top H^\top (H^+)^\top k(Z,x)\right)^2 \,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \left(u^\top \left(I_\ell - H^\top (H^+)^\top\right) k(Z,x)\right)^2 \,\mathrm{d}\mu(x) \\
&= u^\top \left(I_\ell - H^\top (H^+)^\top\right) \left(\int_{\mathcal{X}} k(Z,x)k(x,Z)\,\mathrm{d}\mu(x)\right)(I_\ell - H^+H)u \\
&= u^\top \left(I_\ell - H^\top (H^+)^\top\right) H^\top H (I_\ell - H^+H)u = 0
\end{aligned}$$

since $H^\top (H^+)^\top H^\top = H^\top$ and $HH^+H = H$ hold (I_ℓ is the identity matrix). Thus, we obtain the desired assertion.

Finally, we prove that k^Z_μ and k^Z coincide when $\ker h_\mu(Z,Z) \subset \ker k(Z,Z)$. From (34) and (35), it suffices to prove $H^+Hu_i = u_i$ for indices i with $\lambda_i > 0$. Note that H^+H is the orthogonal projection onto the orthogonal complement of $\ker H = \ker H^\top H = h_\mu(Z,Z)$ from a general property of the pseudo-inverse. Since u_i is an eigenvector of $k(Z,Z)$ with a positive eigenvalue λ_i, it is orthogonal to any $v \in \ker k(Z,Z)$ (as $u_i^\top v = \lambda_i^{-1} u_i^\top k(Z,Z)v = 0$). Therefore, if we have $\ker h_\mu(Z,Z) \subset \ker k(Z,Z)$, u_i is also orthogonal to $\ker h_\mu(Z,Z)$ and so $H^+Hu_i = u_i$ as desired. □

B.12. Proof of Proposition 4

First, we give a proof for a folklore property of products of positive semi-definite matrices.

Lemma 6. *Let $\ell, m \ge n$ be positive integers and $A, B \in \mathbb{R}^{n\times n}$ be (symmetric) positive semi-definite matrices. Assume $B = C^\top C = D^\top D$ for a real matrix $C \in \mathbb{R}^{m\times n}$ and $D \in \mathbb{R}^{\ell\times n}$. Then, $CAC^\top$ and $DAD^\top$ have the same set of nonzero eigenvalues with the same multiplicity (in terms of real eigenvectors).*

Proof. For a real square matrix $M \in \mathbb{R}^{j\times j}$ and a real number λ, let us define $S_\lambda(M) := \{v \in \mathbb{R}^j \mid Mv = \lambda v\}$ be the real eigenspace of M corresponding to λ.

We shall prove there is a bijection between $S_\lambda(AB)$ and $S_\lambda(CAC^\top)$ for each real $\lambda \ne 0$ (and the same for $S_\lambda(DAD^\top)$ by symmetry). Once we establish this, we see that each $\lambda \ne 0$ has the same multiplicity as an eigenvalue of $CAC^\top$ and $DAD^\top$ (multiplicity can be zero; in that case λ is not an eigenvalue), and the desired assertion follows.

Let us fix $\lambda \ne 0$. If $v \in S_\lambda(CAC^\top)$, we have $CAC^\top(Cv) = CABv = \lambda(Cv)$, so $Cv \in S_\lambda(CAC^\top)$. We also have $Cv' \ne Cv$ for another element $(v \ne) v' \in S_\lambda(AB)$ since $AC^\top(Cv' - Cv) = AB(v'-v) = \lambda(v'-v) \ne 0$. Thus, matrix multiplication by C is an injective map from $S_\lambda(AB)$ to $S_\lambda(CAC^\top)$.

Let us finally prove $S_\lambda(AB) \ni v \mapsto Cv \in S_\lambda(CAC^\top)$ is surjective. Let $u \in S_\lambda(CAC^\top)$. Then, $u = \lambda^{-1}(\lambda u) = \lambda^{-1}CAC^\top u = C(\lambda^{-1}AC^\top u)$, so we can write $u = Cv$ for $v = \lambda^{-1}AC^\top u$. It remains to prove $v \in S_\lambda(AB)$, but we can see it as follows:

$$ABv = AB\left(\frac{1}{\lambda}AC^\top u\right) = \frac{1}{\lambda}(AC^\top C)AC^\top u = \frac{1}{\lambda}AC^\top(CAC^\top u) = \frac{1}{\lambda}AC^\top(\lambda u) = \lambda v.$$

Therefore, we have a bijection between $S_\lambda(AB)$ and $S_\lambda(CAC^\top)$ and we are done. □

Recall $\mu(k^Z_\mu - k^Z_{s,\mu}) \le \sum_{i=s+1}^\ell \kappa_i$ holds for eigenvalues $\kappa_1 \ge \cdots \kappa_\ell \ge 0$ of $H_\mu k(Z,Z)^+ H_\mu^\top$ with $H_\mu^\top H_\mu = h_\mu(Z,Z)$ (that immediately follows from the definitions of k^Z_μ and $k^Z_{s,\mu}$, and that f_i are $L^2(\mu)$-orthonormal). By replacing μ with μ_X, we have $\mu_X(k^Z_X - k^Z_{s,X}) \le \sum_{i=s+1}^\ell \kappa_i^X$ for eigenvalues of $\kappa_1^X \ge \cdots \ge \kappa_\ell^X \ge 0$ of $H_X k(Z,Z)^+ H_X^\top$, where $H_X^\top H_X = h_X(Z,Z) = \frac{1}{M}k(Z,X)k(X,Z)$.

By using the lemma, we can see that κ_i^X are actually the same as the eigenvalues of $\frac{1}{M}k(X,Z)k(Z,Z)^+k(Z,X) = \frac{1}{M}k^Z(X,X)$. As $k - k^Z$ is a positive definite kernel, $k(X,X) - k^Z(X,X)$ is a positive semi-definite matrix, the i-th largest eigenvalue of $k^Z(X,X)$ is bounded by the i-th largest eigenvalue of $k(X,X)$ (Weyl's inequality).

Now, let $\lambda_1^X \ge \lambda_2^X \ge \cdots \ge 0$ be the eigenvalues of $k(X,X)$. From the above argument, we have

$$\mu_X(k^Z_X - k^Z_{s,X}) \le \sum_{i=s+1}^\ell \kappa_i^X \le \frac{1}{M}\sum_{i=s+1}^\ell \lambda_i^X \le \frac{1}{M}\sum_{i=s+1}^M \lambda_i^X.$$

Notice that we can apply Lemma 3 with X instead of Z, and obtain $\mathbb{E}\bigl[\mu_X(k_X^Z - k_{s,X}^Z)\bigr] \le \sum_{i>s} \sigma_i$ as desired.

B.13. Proof of Proposition 5

Proof. Fix a sample X with $\ker k(X,Z) \subset \ker k(Z,Z)$ and let us use the same notation as in μ, i.e.,

- $H^\top H = h_X(Z,Z) = \frac{1}{M} k(Z,X)k(X,Z)$;
- $Hk(Z,Z)^+ H^\top = V \operatorname{diag}(\kappa_1, \ldots, \kappa_\ell) V^\top$ with $\kappa_1 \ge \cdots \kappa_\ell \ge 0$ and V being orthogonal;
- $f_i = (H^+ v_i)^\top k(Z,\cdot)$ and $k_X^Z(x,y) = \sum_{i=1}^\ell \kappa_i f_i(x) f_i(y)$.

In this case, from the same argument as the last paragraph in the proof of Proposition 2, we have H^+H is an identity map over $(\ker h_X(Z,Z))^\perp = (\ker k(X,Z))^\perp \supset (\ker k(Z,Z))^\perp$. By considering the SVD of $k(Z,Z)$, we see that $(\ker k(Z,Z))^\perp$ is exactly the linear subspace of $\mathbb{R}^\ell$ spanned by eigenvectors of $k(Z,Z)$ with nonzero eigenvalues, which is equal to $\{k(Z,Z)v \mid v \in \mathbb{R}^\ell\} = \{k(Z,Z)^+ v \mid v \in \mathbb{R}^\ell\}$. In particular, we have $H^+Hk(Z,Z)^+ = k(Z,Z)^+$.

We now prove that $\{\sqrt{\kappa_i} f_i \mid i \ge 1,\ \kappa_i > 0\}$ actually forms an orthonoramal set in $\mathcal{H}_k$. Indeed, if $\kappa_i, \kappa_j > 0$, we have

$$\begin{aligned}
\bigl\langle \sqrt{\kappa_i} f_i, \sqrt{\kappa_j} f_j \bigr\rangle_{\mathcal{H}_k} &= \sqrt{\kappa_i \kappa_j} v_i^\top (H^+)^\top k(Z,Z) H^+ v_j \\
&= \frac{1}{\sqrt{\kappa_i \kappa_j}} v_i^\top \left[Hk(Z,Z)^+ H^\top\right] (H^+)^\top k(Z,Z) H^+ \left[Hk(Z,Z)^+ H^\top\right] v_j \\
&= \frac{1}{\sqrt{\kappa_i \kappa_j}} v_i^\top Hk(Z,Z)^+ k(Z,Z) k(Z,Z)^+ H^\top v_j \\
&= \frac{1}{\sqrt{\kappa_i \kappa_j}} v_i^\top Hk(Z,Z)^+ H^\top v_j = \delta_{ij},
\end{aligned}$$

where we have used the fact that v_i and v_j are eigenvectors of $Hk(Z,Z)^+H^\top$ with eigenvalues κ_i and κ_j, respectively.

Let $P : \mathcal{H}_k \to \mathcal{H}_k$ be the orthogonal projection onto $\operatorname{span}\{\sqrt{\kappa_i} f_i \mid i > s,\ \kappa_i > 0\}$. Then, we have

$$P k(\cdot, x) = \sum_{i=s+1}^{\ell} \langle \sqrt{\kappa_i} f_i, k(\cdot,x) \rangle_{\mathcal{H}_k} \sqrt{\kappa_i} f_i = \sum_{i=s+1}^{\ell} \sqrt{\kappa_i} f_i(x) \sqrt{\kappa_i} f_i,$$

and so $\|Pk(\cdot,x)\|_{\mathcal{H}_k}^2 = \sum_{i=s+1}^\ell \kappa_i f_i(x)^2 = k^Z(x,x) - k_{s,X}^Z(x,x)$. Note that the projection P is a random operator depending on the sample X. Now, we can use Theorem 1 with the empirical measure given by X instead of Z to obtain

$$\mathbb{E}\Bigl[\mu\bigl(\sqrt{k^Z - k_{s,X}^Z}\bigr)\Bigr] \le 2\mathbb{E}\Bigl[\mu_X\bigl(\sqrt{k_X^Z - k_{s,X}^Z}\bigr)\Bigr] + 4\sqrt{\sum_{i>m} \sigma_i} + \frac{\sqrt{k_{\max}}}{M}\left(\frac{80m^2 \log(1+2M)}{9} + 69\right). \tag{36}$$

for any integer $m \ge 1$, where we have used $\|Pk(\cdot,x)\|_{\mathcal{H}_k} = \sqrt{k^Z(x,x) - k_{s,X}^Z(x,x)} = \sqrt{k_X^Z(x,x) - k_{s,X}^Z(x,x)}$ almost surely. From Proposition 4, we have

$$\mathbb{E}\Bigl[\mu_X\bigl(\sqrt{k_X^Z - k_{s,X}^Z}\bigr)\Bigr]^2 \le \mathbb{E}\Bigl[\mu_X\bigl(\sqrt{k_X^Z - k_{s,X}^Z}\bigr)^2\Bigr] \le \mathbb{E}\bigl[\mu_X(k_X^Z - k_{s,X}^Z)\bigr] \le \sum_{i>s} \sigma_i,$$

and combining it with (36) leads to the desired conclusion. □

B.14. Proof of Theorem 4

Proof. We first prove the result for $Q_n = \mathrm{KQuad}(\mathrm{k_{s,Y}}, \mathrm{Y})$. Since $k(x,x) \ge k^Z(x,x) = k_Y^Z(x,x) \ge k_{s,Z}^Y(x,x)$ for $x \in Y$ from Proposition 3, we have

$$\mu_Y\bigl(\sqrt{k - k_{s,Y}^Z}\bigr) \le \mu_Y\bigl(\sqrt{k - k^Z}\bigr) + \mu_Y\bigl(\sqrt{k_Y^Z - k_{s,Y}^Z}\bigr).$$

From Proposition 4, by taking the expectation with regard to Y, we have

$$\mathbb{E}\Big[\mu_Y(\sqrt{k_Y^Z - k_{s,Y}^Z})\Big] \le \sqrt{\mathbb{E}\Big[\mu_Y(k_Y^Z - k_{s,Y}^Z)\Big]} \le \sqrt{\sum_{i>s}\sigma_i},$$

and so we obtain

$$\mathbb{E}\Big[\mu_Y(\sqrt{k - k_{s,Y}^Z})\Big] \le \mu(\sqrt{k - k^Z}) + \sqrt{\sum_{i>s}\sigma_i}$$

By combining it with (13), it is now sufficient to show $\mathbb{E}[\mathrm{MMD}_k(\mu_Y, \mu)] \le \sqrt{c_{k,\mu}/N}$, but actually it follows from the identity $\mathbb{E}\big[\mathrm{MMD}_k(\mu_Y, \mu)^2\big] = c_{k,\mu}/N$, which can be shown by a straightforward calculation (see, e.g., Hayakawa et al., 2022, Proof of Theorem 7).

In the case of $Q_n = \mathrm{KQuad}(k_{s,\mu}^Z, Y)$, we instead have the decomposition

$$\mu_Y(\sqrt{k - k_{s,\mu}^Z}) \le \mu_Y(\sqrt{k - k^Z}) + \mu_Y(\sqrt{k_\mu^Z - k_{s,\mu}^Z});$$

Theorem 2 yields the desired estimate for expectation. □