Quasi-Monte Carlo Features for Kernel Approximation

Zhen Huang[1] **Jiajin Sun**[1] **Yian Huang**[1]

Abstract

Random features (Rahimi & Recht, 2007), based on Monte Carlo (MC) method, is one of the most popular approximation techniques to accelerate kernel methods. We show for a class of kernels, including Gaussian kernels, quasi-Monte Carlo (QMC) methods can be used in place of MC to improve the approximation error from $O_P(1/\sqrt{M})$ to $O(1/M)$ (up to logarithmic factors), for estimating both the kernel function itself and the associated integral operator, where M is the number of features being used. Furthermore, we demonstrate the advantage of QMC features in the case of kernel ridge regression, where theoretically, fewer random features suffice to guarantee the same convergence rate of the excess risk. In practice, the QMC kernel approximation approach is easily implementable and shows superior performance, as supported by the empirical evidence provided in the paper.

1. Introduction

Kernel methods offer a mathematically well-founded and practically powerful nonparametric modeling framework for a wide range of problems in machine learning (Wahba, 1990; Schölkopf & Smola, 2002; Cucker & Smale, 2002). Random features (Rahimi & Recht, 2007) is one of the most popular approximation techniques to accelerate kernel methods. Its idea proceeds as follows: first suppose a kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (where $\mathcal{X}$ is a subset of $\mathbb{R}^d$) has an integral representation:

$$K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\pi(\omega), \tag{1}$$

where π is a probability measure over some space Ω and $\psi(\cdot, \cdot)$ is a function on $\mathcal{X} \times \Omega$. Note that an integral representation in the form of (1) exists under very mild conditions (see e.g., Proposition A.8 in Appendix). Explicit examples include any shift invariant kernel $K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$, for which Bochner's theorem (Bochner, 1933) implies the existence of a finite non-negative symmetric Borel measure μ on $\mathbb{R}^d$ such that

$$h(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^d} e^{-i(\mathbf{x}-\mathbf{x}')^\top \omega} \mathrm{d}\mu(\omega)$$
$$= \int_{\mathbb{R}^d} \int_0^{2\pi} \frac{1}{\pi} \cos\left(\mathbf{x}^\top \omega + b\right) \cos\left((\mathbf{x}')^\top \omega + b\right) \mathrm{d}b\, \mathrm{d}\mu(\omega). \tag{2}$$

Shift invariant kernels cover many popular kernels such as

1. Gaussian kernel $e^{-\|\sigma(\mathbf{x}-\mathbf{x}')\|_2^2/2}$: $\mu \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_d)$.
2. Laplacian kernel $e^{-\|\gamma(\mathbf{x}-\mathbf{x}')\|_1}$: μ has Lebesgue density $\prod_{i=1}^d \frac{1}{\pi\gamma(1+(\omega_i/\gamma)^2)}$ (Cauchy distribution).
3. Cauchy kernel $\prod_{i=1}^d \frac{1}{1+(x_i - x_i')^2/\lambda^2}$: μ has Lebesgue density $\frac{\lambda}{2} e^{-\lambda\|\omega\|_1}$ (Laplace distribution).

Given the kernel function has integral representation (1), $K(\mathbf{x}, \mathbf{x}')$ can be approximated by

$$K_M(\mathbf{x}, \mathbf{x}') = \frac{1}{M} \sum_{i=1}^M \psi(\mathbf{x}, \omega_i)\psi(\mathbf{x}', \omega_i), \tag{3}$$

with $\omega_1, \ldots, \omega_M$ i.i.d. from π. Note that (3) is an inner product on $\mathbb{R}^M$. This reduces the computational complexity of the kernel ridge regression ($O(n^3)$ in time; $O(n^2)$ in space) to that of the ordinary ridge regression on $\mathbb{R}^M$ ($O(nM^2 + M^3)$ in time; $O(nM)$ in space), if $M \ll n$.

For kernel ridge regression, suppose $(\mathbf{X}, Y) \in \mathcal{X} \times \mathbb{R}$ follows a distribution $P_{\mathbf{X}Y}$ with marginal distributions $P_{\mathbf{X}}$ and P_Y. Given the kernel function K, the integral operator $L : L^2(P_{\mathbf{X}}) \to L^2(P_{\mathbf{X}})$ is defined as:

$$Lf(\mathbf{x}) := \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[K(\mathbf{X}, \mathbf{x})f(\mathbf{X})\right]. \tag{4}$$

If the true regression function in the kernel ridge regression belongs to the range of L^r (here $r \in [1/2, 1]$ can be viewed as a complexity or smoothness parameter), then Rudi & Rosasco (2017) shows that $M \asymp n^{\frac{2r}{2r+1}}$ (up to logarithmic factors) random features can ensure the same convergence rate of the excess risk as the exact kernel ridge regression.

[1] Department of Statistics, Columbia University, New York, NY 10027, USA. Correspondence to: Zhen Huang <zh2395@columbia.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

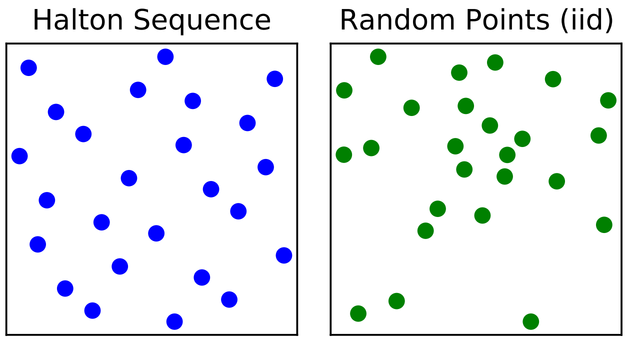

Figure 1. Left: the first 25 points of the two-dimensional Halton sequence. Right: 25 i.i.d. random points from $\text{Unif}[0,1]^2$.

Our contributions: In this paper, we show that compared with the random error bound from Monte Carlo (MC) method: $|K(x,x') - K_M(x,x')| = O_P(1/\sqrt{M})$ (Dick et al., 2013), quasi-Monte Carlo (QMC) method uses a deterministic sequence $\omega_1, \dots, \omega_M$ to achieve a deterministic error bound $|K(x,x') - K_M(x,x')| = O(\log^a M/M)$ for some integer a under some conditions (Theorems 2.2 and 2.5) — this convergence rate is much faster and non-random — and such an improvement holds true for a class of kernels including the Gaussian kernel and Cauchy kernel mentioned above. The improved approximation to the kernel also leads to improved approximations of the integral operator and the spectrum of the kernel matrix (see Propositions 2.6 and 2.7). Further, we demonstrate the usefulness of QMC features in the application of kernel ridge regression, by showing that with QMC method, $M \approx n^{\frac{1}{2r+1}}$ is enough to guarantee the same convergence rate of the excess risk as exact kernel ridge regression (Theorem 2.2). This is an enormous reduction from MC based random features (which require $M \approx n^{\frac{2r}{2r+1}}$) when $r > 1/2$ (i.e., the true regression function has a smoothness condition beyond simply lying in the *reproducing kernel Hilbert space*[1], a.k.a. RKHS, associated with the kernel K; see Section 3 for more details). In practice, the QMC kernel approximation approach is easily implementable and demonstrates superior performance, as supported by empirical evidence provided in the paper.

1.1. Quasi-Monte Carlo Method

QMC is a powerful tool in numerical integration. Its primary focus is to approximate integrals over the unit cube with respect to the uniform measure. In order to approximate $\int_{[0,1]^d} f(\mathbf{x})\mathrm{d}\mathbf{x}$ with a sum $\frac{1}{M}\sum_{i=1}^M f(\mathbf{x}_i)$, MC uses i.i.d. random $\{\mathbf{x}_i\}_{i=1}^M$, while QMC uses some well-chosen deterministic $\{\mathbf{x}_i\}_{i=1}^M$ that are spread out "more uniformly" in a certain sense. In this section, we will cover some background that is necessary for subsequent discussions. More details can be found in textbooks such as Niederreiter (1992) and Owen (2023). We first introduce an important inequality in QMC:

Theorem 1.1 (Koksma-Hlawka inequality, Hlawka, 1961)**.** *Suppose $f : [0,1]^d \to \mathbb{R}$ has bounded variation in the sense of Hardy and Krause $V_{\text{HK}}(f)$.*[2] *Then for any $\mathbf{x}_1, \dots, \mathbf{x}_M \in [0,1]^d$, we have*

$$\left| \int_{[0,1]^d} f(\mathbf{x})\mathrm{d}\mathbf{x} - \frac{1}{M}\sum_{i=1}^M f(\mathbf{x}_i) \right| \le V_{\text{HK}}(f)\mathcal{D}^*(\{\mathbf{x}_i\}_{i=1}^M),$$

where $\mathcal{D}^*(\{\mathbf{x}_i\}_{i=1}^M)$ is the star discrepancy[3] *of the point set $\{\mathbf{x}_i\}_{i=1}^M$.*

QMC is useful thanks to the existence of some low-discrepancy sequences. One notable example is the *Halton sequence* $\mathbf{h}_1, \mathbf{h}_2, \dots$ which satisfies

$$\mathcal{D}^*(\{\mathbf{h}_i\}_{i=1}^M) \le C_H(d)(\log M)^d/M \tag{5}$$

for some $C_H(d) > 0$ that depends on d, and all $M \ge 2$ (Halton, 1964; Atanassov, 2004). This is a substantial improvement from random sampling, whose star discrepancy is of order $O_P(M^{-1/2})$. It may be seen from Figure 1 that points from Halton sequence appear "more uniform" than i.i.d. points from the uniform distribution. In practice, Halton sequence can be easily generated from an elegant formula and is directly accessible in major computational softwares. Compared with other QMC sequences, Halton sequence has a distinct feature: it avoids the boundary of the unit cube (Owen, 2006). This characteristic makes it particularly useful in approximating a class of shift-invariant kernels, including the Gaussian kernel and Cauchy kernel (see Section 2.1 for details). Hence, the Halton sequence will be the primary choice of QMC sequence in this paper.

Other low-discrepancy QMC sequences include *Sobol' sequence* (Sobol', 1967) and *Faure sequence* (Faure, 1982), which were combined by Niederreiter (1987) to formulate the concepts of *digital nets* and *sequences*. Note that digital sequences also satisfy (5), but with $C_H(d)$ replaced by a different constant (Niederreiter, 1992, Theorem 4.17). The lead

[1] The RKHS (Aronszajn, 1950) $\mathcal{H}$ is a space of function over $\mathcal{X}$ consisting of $\text{span}\{K(\mathbf{x},\cdot) : \mathbf{x} \in \mathcal{X}\}$ and their limits, equipped with an inner product given by $\langle K(\mathbf{x},\cdot), K(\mathbf{x}',\cdot)\rangle_{\mathcal{H}} = K(\mathbf{x},\mathbf{x}')$.

[2] In one-dimension, Hardy-Krause variation coincides with the usual total variation. In general dimensions, $V_{\text{HK}}(f;[0,1]^d) = \sum_{I\subset\{1,\dots,d\}, I\neq\emptyset} \int_{[0,1]^{|I|}} \left| \frac{\partial f}{\partial u_I}\Big|_{u_j=1, j\notin I} \right| \mathrm{d}u_I$, provided that f has all the related derivatives. For definition in general situation, see e.g., Niederreiter (1992); Owen (2005).

[3] The star discrepancy of the point set $\{\mathbf{x}_i\}_{i=1}^M$ is defined as $\mathcal{D}^*(\{\mathbf{x}_i\}_{i=1}^M) := \sup_{\mathbf{t}\in[0,1]^d} \left| \text{Vol}(J_{\mathbf{t}}) - \frac{|\{i\in\{1,\dots,M\}:\mathbf{x}_i\in J_{\mathbf{t}}\}|}{M} \right|$, where $J_{\mathbf{t}} := [0,t_1) \times [0,t_2) \times \cdots \times [0,t_d)$ and $\text{Vol}(J_{\mathbf{t}}) := \prod_{i=1}^d t_i$ is the volume.

constant on the dominating term $(\log M)^d/M$ (for digital sequences) used to be much smaller than that for the Halton sequences for large d, but that was changed by Atanassov (2004) who considerably sharpened the bounds for Halton squence (the constant was shown to converge to 0 as $d \to \infty$). It is conjectured that the $O((\log M)^d/M)$ rate for star discrepancy decay is optimal for infinite sequences, and Schmidt (1972) proved this in the case $d = 1$. For $d > 1$, the question remains open; a lower bound $(\log M)^{\frac{d}{2}}/M$ was provided by Roth (1954), which was slightly improved by Baker (1999).

When applying QMC to kernel approximation, a negative result was found by Avron et al. (2016) that the integral representation (2) from Bochner's theorem, when written as an integral over the unit cube, has infinite variation. Consequently, the Koksma-Hlawka inequality (Theorem 1.1) cannot provide a meaningful bound.

To overcome this difficulty, we show that for a class of shift invariant kernels including the Gaussian kernel, even though the integrand has infinite variation, the singularity is mild, so the approximation error can still be well controlled. Our result relies on the geometry of Halton sequence that it avoids the boundary of the unit cube (Owen, 2006). In addition to shift invariant kernels, We also provide examples of non-shift invariant kernels which have integral representation (1) with the integrand having bounded variation. Our results continue to hold true for such non-shift invariant kernels.

1.2. Related Literature

Kernel methods provide a mathematically rigorous nonparametric modeling approach that finds applications across a broad spectrum of machine learning (Fukumizu et al., 2004; Belkin et al., 2006; Fukumizu et al., 2009; Sriperumbudur et al., 2011; Gretton et al., 2012; Fukumizu et al., 2013; Klebanov et al., 2020; Huang et al., 2022). Despite being remarkably effective in small and medium size problems with certain optimal statistical results (Kimeldorf & Wahba, 1970; Schölkopf et al., 2001; Caponnetto & De Vito, 2007), exact kernel methods become infeasible for large scale problems due to its time and memory requirements (Rudi & Rosasco, 2017). To overcome this difficulty, various approximation techniques have been proposed (Smola, 2000; Williams & Seeger, 2000; Rahimi & Recht, 2007). One notable approach is *random features* (RF) (Rahimi & Recht, 2007) which have been well-understood theoretically (Sutherland & Schneider, 2015; Sriperumbudur & Szabó, 2015; Choromanski et al., 2018; Jacot et al., 2020; Lanthaler & Nelsen, 2023). In particular, RF demonstrates nice generalization properties, achieving the same rate of prediction accuracy as the exact kernel ridge regression estimator but at a much lower computational cost (Rudi & Rosasco, 2017; Li et al., 2019; Mei et al., 2022; Liu et al., 2022). In this paper, we show that with QMC, the computational cost can be further reduced.

QMC is effective for numerical integration, which was born in the 1950s and 1960s (Korobov, 1963) from the successful attempt to achieve a faster convergence rate than MC. Contemporary reviews of QMC can be found in the books of Niederreiter (1992); Leobacher & Pillichshammer (2014) and articles of Owen (2005); Dick et al. (2013).

Efforts have been made to employ QMC to random features: Yang et al. (2014) and Avron et al. (2016) considered shift-invariant kernels, and found that Koksma-Hlawka inequality cannot be applied due to the integrand having unbounded variation; they instead proposed a framework based on a modified discrepancy measure called *box discrepancy*. QMC points on the high-dimensional sphere were adopted in Lyu (2017), which can be used to approximate shift and rotation invariant kernels and the arc-cosine kernels. Applying QMC to the graph kernels and graph random features was considered in Reid et al. (2023). Nevertheless, none of the above proposals provides a deterministic error bound of the kernel approximation in finite samples, nor do they show whether QMC could attain comparable performance as the exact kernel computation in learning problems, while reducing computational costs. We aim to answer these questions in this paper.

1.3. Organization

In Section 2, we describe how QMC can be used to approximate a kernel function, and provide a deterministic approximation error bound (of the same order as that in the QMC literature) in finite samples. In Section 3, we show that the same error rate in kernel ridge regression can be achieved with much lower computational costs (compared with the exact computation and MC based random features). Simulation evidence will be given in Section 4. Proofs of our results, further discussions, additional simulation studies and real data examples are provided in Appendices A-D.

2. Approximate Kernel Functions with QMC

In this section, we show how QMC can be used to approximate kernel functions, and provide deterministic error bounds for the approximation. In Section 2.1, we consider shift invariant kernels; in Section 2.2, we consider non-shift invariant kernels. Approximation error bounds for the associated integral operator and the spectrum of the kernel matrix will be given in Section 2.3.

2.1. Shift Invariant Kernels

Recall that we are considering kernels on a space $\mathcal{X} \subset \mathbb{R}^d$. We assume that the μ from Bochner's theorem (2) is a probability measure with independent components[4], with the i-th component having cumulative distribution function $\Phi_i(t)$ $(i = 1, 2, \ldots, d)$. Let $\mathbf{\Phi}(\mathbf{t}) := (\Phi_1(\mathbf{t}), \ldots, \Phi_d(\mathbf{t}))^\top$, and $\mathbf{\Phi}^{-1}(\mathbf{t}) := (\Phi_1^{-1}(\mathbf{t}), \ldots, \Phi_d^{-1}(\mathbf{t}))^\top$, where $\Phi_i^{-1}(\mathbf{t})$ denotes the inverse function of the monotone function $\Phi_i(\mathbf{t})$. By a change of variable, (2) reduces to

$$K(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}') = \int_{[0,1]^{d+1}} 2\cos\left(\mathbf{x}^\top \mathbf{\Phi}^{-1}(\mathbf{t}) + 2\pi b\right) \cos\left((\mathbf{x}')^\top \mathbf{\Phi}^{-1}(\mathbf{t}) + 2\pi b\right) \mathrm{d}b \mathrm{d}\mathbf{t}. \tag{6}$$

Therefore, the integral representation (1) holds with $\omega = (\mathbf{t}, b)$ following $\text{Unif}[0,1]^{d+1}$ and $\psi(\mathbf{x}, \omega) = \sqrt{2}\cos\left(\mathbf{x}^\top \mathbf{\Phi}^{-1}(\mathbf{t}) + 2\pi b\right)$. We propose to set $\omega_1, \ldots, \omega_M$ as the first M points in the Halton sequence, and define the approximate kernel function $K_M(\cdot, \cdot)$ as in (3).

As $\mathbf{t}$ approaches the boundary of $[0,1]^d$, the integrand in (6) oscillates back and forth and has unbounded variation. We therefore need a condition to characterize the situation where the singularity is mild so that K can still be well-approximated by K_M:

QMC Condition 1. $K(\cdot, \cdot)$ is shift invariant with Φ_i defined as above $(i = 1, \ldots, d)$ satisfying $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_i^{-1}(t) \leq \frac{C_i}{\min(t, 1-t)}$ for some constant $C_i > 0$ and all $t \in (0,1)$. $\mathcal{X}$ is compact.

QMC Condition 1 helps control the derivatives of the integrand in (6) as $\mathbf{t}$ approaches the boundary of $[0,1]^d$. Two important kernels that satisfy QMC Condition 1 are given by the proposition below: (see Appendix A.1 for a proof)

Proposition 2.1. *The Gaussian kernel and Cauchy kernel over a compact domain satisfy QMC Condition 1.*

Gaussian kernel and Cauchy kernel over a compact domain are examples of *universal kernels* (Micchelli et al., 2006), i.e., the function class associated with the kernel can approximate (uniformly) any continuous function arbitrarily well. This property makes them particularly useful in machine learning applications such as kernel ridge regression, where an unknown regression function needs to be estimated from data. Laplacian kernel, although being universal, unfortunately does not satisfy QMC Condition 1. The following theorem (proved in Appendix B.1) shows that if QMC Condition 1 is satisfied, then the approximation error of K_M to K is of order $1/M$, up to logarithmic factors.

Theorem 2.2. *Suppose $K(\cdot, \cdot)$ satisfies QMC Condition 1. Then there exists a constant $C > 0$ (depending on $\mathcal{X} \subset \mathbb{R}^d$ and K) such that for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $M \geq 2$,*

$$|K_M(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{x}')| \leq \frac{C(\log M)^{2d+1}}{M}.$$

This error rate is significantly better than that of the MC-based random features, which is of order $O_P(1/\sqrt{M})$.

Remark 2.3 (On the proof of Theorem 2.2). The general idea is to study the singularity of the integrand (6) near the boundary of the unit cube, which will be mild if QMC Condition 1 holds true. The classical Koksma-Hlawka inequality (Theorem 1.1) can then be applied to a large sub-cube within the unit cube. The fact that Halton sequence avoids the boundary of the unit cube (Owen, 2006) is useful, which ensures that the first M points of the Halton sequence do not lie too close to the boundary.

Remark 2.4 (On the constant C). The exact expression of the error bound can be found in our proof (in Appendix B.1). In particular, the constant multiplied to the dominating term $\frac{(\log M)^{2d+1}}{M}$ is $C_H(d+1)2^{2d+1}B$, where $C_H(d+1)$ is the constant from the Halton sequence (5), and $B = 4\pi \max_{u \subset \{1,\ldots,d\}} \big\{ \max^{|u|}_{\mathbf{x},\mathbf{y} \in \mathcal{X}} \{\|\mathbf{x} - \mathbf{y}\|_\infty, \|\mathbf{x} + \mathbf{y}\|_\infty\} \prod_{i \in u} C_i \big\}$, with the convention that $(\max\{\cdot\})^0 = 1$.

2.2. Non-Shift Invariant Kernels

For non-shift invariant kernels, Bochner's theorem is no longer applicable. Consequently, whether $K(\cdot, \cdot)$ possesses an integration representation in the form of (1) needs to be considered on a case-by-case basis. In this section, we will provide a collection of non-shift invariant kernels which have representation (1), and QMC can be applied.

Motivated by the Koksma-Hlawka inequality (Theorem 1.1), we introduce the following general condition:

QMC Condition 2. Suppose there exists a function $\psi : \mathcal{X} \times [0,1]^p \to \mathbb{R}$ such that

$$K(\mathbf{x}, \mathbf{x}') = \int_{[0,1]^p} \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\omega,$$

and for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $g(\omega) = \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)$ is of bounded Hardy-Krause variation $V_{\text{HK}}(g) \leq C_0$, for some $C_0 > 0$.

Note that if all derivatives of g are well-bounded, then $V_{\text{HK}}(g)$ can be bounded; see footnote 2. When QMC Condition 2 is satisfied, we set $\omega_1, \ldots, \omega_M$ as the first M points in the Halton sequence, and define the approximate kernel function $K_M(\cdot, \cdot)$ as in (3). A direct application of the Koksma-Hlawka inequality (Theorem 1.1) yields the following error bound:

Theorem 2.5. *Suppose $K(\cdot, \cdot)$ satisfies QMC Condition 2. Let $C_H(p)$ be the Halton sequence constant as in (5). For any $x, x' \in \mathcal{X}$ and $M \geq 2$, we have*

$$|K_M(\mathbf{x}, \mathbf{x}') - K(\mathbf{x}, \mathbf{x}')| \leq C_H(p) \cdot \frac{(\log M)^p}{M}.$$

One may notice that the Halton sequence is not essential here for Theorem 2.5; other QMC sequences can also be used, though the constant in front of $(\log M)^p/M$ may vary.

[4] This assumption has been adopted in previous works (e.g., Avron et al., 2016) and is satisfied for many common kernels.

In the following, we present some kernels for which QMC Condition 2 is satisfied, and thereby Theorem 2.5 is valid.

Example 1 (Min kernel). For $u, v \in [0,1]$,

$$K(u,v) = \min\{u,v\} = \int_0^1 1_{t<u} 1_{t<v} \mathrm{d}t.$$

$1_{t<u}1_{t<v} = 1_{t<\min\{u,v\}}$ is of bounded variation 1. *Min kernel* is the covariance kernel of the Brownian motion, and is also the *distance kernel* in one-dimension — a famous example of a characteristic, but not universal kernel (Sejdinovic et al., 2013). High-dimensional *min kernel* is also available, by taking the product of univariate *min kernels*.

Example 2 (Brownian bridge). For $u, v \in [0,1]$,

$$K(u,v) = \min\{u,v\} - uv = \int_0^1 (1_{t<u} - u)(1_{t<v} - v)\mathrm{d}t.$$

The integrand $(1_{t<u} - u)(1_{t<v} - v)$ has variation bounded by 3. The *Brownian bridge kernel* has been used to analyze average-case errors in numerical problems (Ritter, 2000).

Example 3 (Iterative kernel). Suppose $K_1(\cdot,\cdot)$ is a continuous kernel on $[0,1]^d$, and μ is a positive integrable function on $[0,1]^d$. The *iterative kernel* (Courant & Hilbert, 1953, Section III.5.3) of K_1 is defined as

$$K_2(\mathbf{x},\mathbf{z}) := \int_{[0,1]^d} K_1(\mathbf{x},\mathbf{t})K_1(\mathbf{z},\mathbf{t})\mu(\mathbf{t})\mathrm{d}\mathbf{t}.$$

If the Hardy-Krause variation of the integrand $f_{\mathbf{x},\mathbf{z}}(\mathbf{t}) := K_1(\mathbf{x},\mathbf{t})K_1(\mathbf{z},\mathbf{t})\mu(\mathbf{t})$ is bounded by some $C_0 > 0$ for all $\mathbf{x},\mathbf{z} \in [0,1]^d$, then QMC Condition 2 is satisfied. A sufficient condition for the existence of such a C_0 is that there exists a constant $\tilde{C}_0 > 0$ such that for all $u \subset \{1,\dots,d\}$ and $\mathbf{x},\mathbf{z},\mathbf{t} \in [0,1]^d$, $|(\prod_{i\in u} \frac{\partial}{\partial t_i}) f_{\mathbf{x},\mathbf{z}}(\mathbf{t})| \le \tilde{C}_0$. If K_1 has Mercer series (w.r.t. μ): $K_1(\mathbf{x},\mathbf{z}) = \sum_{i=1}^\infty \lambda_i e_i(\mathbf{x})e_i(\mathbf{z})$, then it can be shown that $K_2(\mathbf{x},\mathbf{z}) = \sum_{i=1}^\infty \lambda_i^2 e_i(\mathbf{x})e_i(\mathbf{z})$ (which is smoother than the original K_1). See Appendix A.2.1 for more detailed discussions.

Example 4 (Natural cubic spline). For $u, v \in [0,1]$,

$$\begin{aligned} K(u,v) &= \int_0^1 (u \wedge t - ut)(v \wedge t - vt)\mathrm{d}t \\ &= \begin{cases} \frac{1}{6}u(1-v)(1-u^2-(1-v)^2), 0 \le u \le v \le 1 \\ \frac{1}{6}v(1-u)(1-v^2-(1-u)^2), 0 \le v \le u \le 1 \end{cases}. \end{aligned}$$

The integrand above has variation bounded by 4. For any fixed value of v, it is a natural cubic spline that interpolates zero at $u = 0$ and $u = 1$. This kernel is also the iterative kernel of Brownian bridge with $\mu(t) \equiv 1$.

Example 5 (Product kernel). Suppose for $i = 1,\dots,d$, $K_i(u,v) = \int_0^1 \psi_i(u,t)\psi_i(v,t)\mathrm{d}t$ satisfies QMC Condition 2 with $|\psi_i(u,t)| \le \kappa_i$ for some κ_i and all u, t. Then

$$K(\mathbf{u},\mathbf{v}) = \prod_{i=1}^d K_i(u_i,v_i) = \int_{[0,1]^d} \psi(\mathbf{u},\mathbf{t})\psi(\mathbf{v},\mathbf{t})\mathrm{d}\mathbf{t},$$

where $\psi(\mathbf{x},\mathbf{t}) = \prod_{i=1}^d \psi_i(x_i,t_i)$, satisfies QMC Condition 2. See Appendix A.2.2 for the detailed proof.

Example 5 allows us to construct high-dimensional kernels that satisfy QMC Condition 2 with Examples 1-4.

2.3. Approximate Integral Operator and Kernel Matrix

When the kernel function $K(\cdot,\cdot)$ is well approximated by $K_M(\cdot,\cdot)$, the associated integral operator and the kernel matrix can also be well approximated. Recall the integral operator L defined in (4). Define its approximation $L_M : L^2(P_{\mathbf{X}}) \to L^2(P_{\mathbf{X}})$ as

$$L_M f(\mathbf{x}) := \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[K_M(\mathbf{X},\mathbf{x})f(\mathbf{X})\right].$$

The following result (proved in Appendix B.2) shows that the error in estimating the kernel function propagates to that in estimating the integral operator:

Proposition 2.6. *Suppose two kernels K and K_M satisfy*

$$\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} |K(\mathbf{x},\mathbf{x}') - K_M(\mathbf{x},\mathbf{x}')| \le C \cdot \frac{(\log M)^a}{M}$$

for some positive constants C and a. Then we have

$$\|L_M - L\| \le C \cdot \frac{(\log M)^a}{M},$$

where $\|\cdot\|$ denotes the operator norm.

Proposition 2.6 will be useful in showing the superior performance of QMC features (compared with MC based random features) in kernel ridge regression (see Section 3).

Another advantage of using QMC features concerns the spectral approximation of the kernel matrix:

Proposition 2.7 (Spectrum approximation)**.** *Suppose two kernels K and K_M satisfy*

$$\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} |K(\mathbf{x},\mathbf{x}') - K_M(\mathbf{x},\mathbf{x}')| \le C \cdot \frac{(\log M)^a}{M}$$

for some constants $C, a > 0$. Let $\mathbf{K} := [K(\mathbf{x}_i,\mathbf{x}_j)]_{(i,j)} \in \mathbb{R}^{n\times n}$, $\mathbf{K}_M := [K_M(\mathbf{x}_i,\mathbf{x}_j)]_{(i,j)} \in \mathbb{R}^{n\times n}$, and $\lambda, \Delta > 0$. When $\frac{M}{(\log M)^a} \ge \frac{Cn}{\Delta\lambda}$, we have

$$(1-\Delta)(\mathbf{K}+\lambda\mathbf{I}_n) \preceq \mathbf{K}_M + \lambda\mathbf{I}_n \preceq (1+\Delta)(\mathbf{K}+\lambda\mathbf{I}_n). \quad (7)$$

Compare with (MC based) random features that require M to be of order $\frac{n}{\Delta^2\lambda}$ to achieve a Δ-spectral approximation (7) with high probability (Avron et al., 2017, Theorem 7), QMC features only require M of order $\frac{n}{\Delta\lambda}$ (ignoring logarithmic factors). Moreover, (7) holds true with probability 1. Such a spectral bound is useful in analyzing the statistical performance of some downstream learning tasks (see e.g., Musco & Musco, 2017, Appendix E).

3. Application in Kernel Ridge Regression

In this section, we will show how QMC features introduced in Section 2 can be used to further accelerate the computation of kernel ridge regression (compared with MC random features), while still maintaining the same statistical accuracy. We will first give a brief review on the kernel ridge regression with random features, and then provide the theoretical results for our method.

3.1. Brief Review on Kernel Ridge and Random Feature

Brief Review on Kernel Ridge Regression. Consider the usual setting where we have n i.i.d. samples $(\mathbf{x}_i, y_i)_{i=1}^n$ drawn from a joint distribution $P_{\mathbf{X}Y}$ on $\mathcal{X} \times \mathbb{R}$. The goal is to learn the regression function $f_\star(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$ which minimizes the expected risk $\mathcal{E}(f) = \mathbb{E}[Y - f(\mathbf{X})]^2$. Given a kernel K, denote the reproducing kernel Hilbert space associated with K by $\mathcal{H}$. The kernel ridge regression (KRR) defines a penalized estimator for the above learning problem

$$\hat{f}_\lambda := \underset{f \in \mathcal{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\} \tag{8}$$

for some $\lambda > 0$, and has the explicit solution

$$\hat{f}_\lambda(\mathbf{x}) = \sum_{i=1}^n \hat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}), \quad \hat{\boldsymbol{\alpha}} = (\mathbf{K} + n\lambda \mathbf{I}_n)^{-1} \mathbf{y} \tag{9}$$

where $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{(i,j)} \in \mathbb{R}^{n \times n}$, $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\mathbf{y} = (y_1, \dots, y_n)^\top$, and $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n)^\top$. Statistically, the KRR estimator (8) is minimax rate optimal for the sqaure loss $\mathcal{E}(\cdot)$ over $\mathcal{H}$ (Schölkopf & Smola, 2002; Caponnetto & De Vito, 2007). Computationally, the time and space complexities of KRR are of order $O(n^3)$ and $O(n^2)$, respectively, which could be costly when n is large.

Brief Review on KRR with Random Features. Recall from Section 1 that when the kernel function K has an integral representation $K(\mathbf{x}, \mathbf{x}') = \int_\Omega \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega) \mathrm{d}\pi(\omega)$ as in (1), one could use a Monte Carlo integration $K_M(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_M(\mathbf{x})^\top \boldsymbol{\phi}_M(\mathbf{x}') = \frac{1}{M} \sum_{i=1}^M \psi(\mathbf{x}, \omega_i)\psi(\mathbf{x}', \omega_i)$ as in (3) to approximate K, where the ω_i's are i.i.d. sampled from π, and $\boldsymbol{\phi}_M(\mathbf{x}) := M^{-1/2}(\psi(\mathbf{x}, \omega_1), \dots, \psi(\mathbf{x}, \omega_M))^\top$. Substituting $\mathbf{K}_M = [K_M(\mathbf{x}_i, \mathbf{x}_j)]_{(i,j)}$ for $\mathbf{K}$ in the KRR estimator (9) gives the random feature kernel ridge regression (RF-KRR) estimator (Rudi & Rosasco, 2017; Avron et al., 2017). The time and space complexities of RF-KRR are $O(nM^2 + M^3)$ and $O(nM)$, respectively, which indicates a reduction in computational cost compared to (9) if $M \ll n$. Furthermore, this computational gain does not bring an additional cost in statistical error: Rudi & Rosasco (2017) shows that RF-KRR with $M \asymp n^{\frac{2r}{2r+1}}$ (up to logarithmic factors) guarantees the same error rate as the exact KRR, where $r \in [\frac{1}{2}, 1]$ is a measure of complexity to be rigorously defined in Section 3.3 below.

3.2. Kernel Ridge Regression with QMC Features

The nice properties of RF-KRR rely on the fact that K_M approximates K with an $O_P(M^{-1/2})$ error rate when sampling $\psi(\mathbf{x}, \omega_i)$ with i.i.d. ω_i. Enlightened by the even better error $O(M^{-1})$ QMC approximation (up to logarithmic factors) of K_M to K as shown in Theorems 2.2 and 2.5, we naturally consider the quasi-Monte Carlo feature kernel ridge regression (QMCF-KRR) estimator by employing $\mathbf{K}_M$ in lieu of $\mathbf{K}$ in the KRR estimator (9), where $\mathbf{K}_M$ is the QMC approximation to $\mathbf{K}$. Through algebraic transformations (Bach, 2017), we have the explicit formula of the QMCF-KRR estimator given by

$$\hat{f}_{\lambda,M}(\mathbf{x}) = \boldsymbol{\phi}_M(\mathbf{x})^\top \left(\hat{\boldsymbol{\Phi}}_M^\top \hat{\boldsymbol{\Phi}}_M + n\lambda \mathbf{I}_M \right)^{-1} \hat{\boldsymbol{\Phi}}_M^\top \mathbf{y} \tag{10}$$

where $\hat{\boldsymbol{\Phi}}_M := (\boldsymbol{\phi}_M(\mathbf{x}_1), \dots, \boldsymbol{\phi}_M(\mathbf{x}_n))^\top \in \mathbb{R}^{n \times M}$, with $\boldsymbol{\phi}_M(\mathbf{x})$ still defined as $M^{-1/2}(\psi(\mathbf{x}, \omega_1), \dots, \psi(\mathbf{x}, \omega_M))^\top$ but ω_i's are now generated from a QMC sequence.

From (10) it is clear that, the time and space complexities of QMCF-KRR, like those of the RF-KRR estimator, are $O(nM^2 + M^3)$ and $O(nM)$, respectively. It remains to answer the question: how large should M be to guarantee good statistical accuracy? In the next subsection, we show that to achieve the same error rate as RF-KRR and the exact KRR, QMCF-KRR requires only $M \asymp n^{\frac{1}{2r+1}}$ (up to logarithmic factors) number of random features, which further reduces the computational cost compared with RF-KRR while maintaining the same statistical accuracy.

3.3. Theoretical Results for QMCF-KRR

Under the settings of Kernel Ridge Regression in subsection 3.1, for $\hat{f}_{\lambda,M}$ as defined by (10), we will establish the statistical excess error rate in terms of $\mathcal{E}(f) = \mathbb{E}[Y - f(\mathbf{X})]^2$ over the class $\mathcal{H}$. To formally state our result, we postulate the following regularity conditions:

KRR Condition 1. (i) $K(\mathbf{x}, \mathbf{x}')$ is continuous and has the integral representation (1), in which $|\psi(\mathbf{x}, \omega)| \le \kappa$ for some constant $\kappa > 0$. Assume $\mathbf{X}$ has full support on $\mathcal{X}$, and $\omega \mapsto \psi(\cdot, \omega)$, as a map from Ω to $L^2(P_{\mathbf{X}})$, is continuous.

(ii) π in (1) is the uniform distribution over $[0, 1]^p$ for some $p \ge 1$, and quasi-Monte Carlo method is used for approximating the kernel as in (3), from which we have

$$\sup_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} |K(\mathbf{x}, \mathbf{x}') - K_M(\mathbf{x}, \mathbf{x}')| \le C \cdot \frac{\log^a M}{M}$$

for some positive constants C and a (see Theorems 2.2, 2.5).

KRR Condition 2. The distribution of Y satisfies a Bernstein condition: there exist positive constants σ and D such that $\mathbb{E}[|Y|^k \mid \mathbf{X}] \le \frac{1}{2} k! \sigma^2 D^{k-2}$ for all $k \ge 2$.

KRR Condition 3. There exists $r \in [1/2, 1]$ such that $f_{\mathcal{H}} = L^r g$ for some $g \in L^2(P_{\mathbf{X}})$, where $f_{\mathcal{H}}$ solves

$\min_{f\in\mathcal{H}}\mathcal{E}(f)$, and L is the integral operator defined in (4). Let $R := \max\{\|g\|_{L^2(P_{\mathbf{X}})}, 1\}$ be a positive constant.

Remark 3.1 (On the conditions). In KRR Condition 1(i), the continuity of K and the full support of $\mathbf{X}$ are standard assumptions for a Mercer kernel, which implies that $\mathcal{H}$ is essentially $\operatorname{ran} L^{1/2}$ (see Theorem A.7 in the Appendix). The continuity of $\omega \mapsto \psi(\cdot,\omega) \in L^2(P_{\mathbf{X}})$ is weaker than the continuity of $\psi(\mathbf{x},\omega)$ (which was assumed in Rudi & Rosasco 2017); it includes *min kernel* where $\psi(x,\omega) = 1_{\omega<x}$ is not continuous in the usual sense. For the boundedness assumption of ψ, we note that all examples in Section 2 have bounded feature functions. KRR Condition 1(ii), as shown in Theorems 2.2 and 2.5, is satisfied under the case of either QMC Condition 1 or 2. KRR Condition 2 is a moderate and usual condition on the tail of the response distribution, which holds when the conditional distribution of Y is sub-exponential. KRR Condition 3 is widely-used in the kernel machine literature (Smale & Zhou, 2003; Caponnetto & De Vito, 2007), whose implications manifest in two folds. First, it assumes the existence of a minimizer $f_{\mathcal{H}}$ of the loss $\mathcal{E}(\cdot)$ in the class $\mathcal{H}$. Second, it further assumes that $f_{\mathcal{H}}$ lies in the range of L^r. Here $r \in [1/2, 1]$ is interpreted as a measure of complexity of $f_{\mathcal{H}}$, and can be intuitively understood as a smoothness parameter: the larger r is, the smoother $f_{\mathcal{H}}$ is, as L is a convolution. In particular, $r = 1/2$ represents the basic case, equivalent to assuming only that $f_{\mathcal{H}}$ exists in $\mathcal{H}$ (see Theorem A.7 in the Appendix).

Thoerem 3.2 below (proved in Appendix C) establishes the statistical error rate of the proposed QMCF-KRR estimator (10).

Theorem 3.2. *Assume KRR Conditions 1, 2, 3. Let* $\lambda = \tilde{C}n^{-\frac{1}{2r+1}} \in (0, e^{-1}]$*, and* $\hat{f}_{\lambda,M}$ *be defined as in* (10)*. Then* $M = \frac{\log^a(1/\lambda)}{\lambda} = n^{\frac{1}{2r+1}}\log^a(n^{\frac{1}{2r+1}}/\tilde{C})/\tilde{C}$ *is enough to guarantee that, for any* $\delta \in (0,1]$*, there exists* n_0 *(of order* $(\log\frac{1}{\delta})^{1+\frac{1}{2r}}$*), such that when* $n \geq n_0$*, with probability at least* $1-\delta$*, the excess risk*

$$\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f\in\mathcal{H}}\mathcal{E}(f) \leq C_1 n^{-\frac{2r}{2r+1}}\log^2\frac{6}{\delta}, \qquad (11)$$

where C_1 *is a constant depending only on* $\kappa, \sigma, D, R, r, \tilde{C}$*,* C *and* a *(see Appendix C.2 for the exact expression of* C_1*).*

Remark 3.3 (Implications of Theorem 3.2). Our error bound (11) achieves the same error rate as in exact KRR (Caponnetto & De Vito, 2007, Theorem 1) and RF-KRR (Rudi & Rosasco, 2017, Theorem 2). However, our QMC approach is computationally more efficient in smoother or less complexity cases: RF-KRR (Rudi & Rosasco, 2017, Theorem 2) requires $M \asymp n^{\frac{2r}{2r+1}}\log\frac{108\kappa^2 n}{\delta}$ many random features to achieve an excess risk of $\tilde{C}_1 n^{-\frac{2r}{2r+1}}\log^2\frac{18}{\delta}$, while our QMCF-KRR method needs only $M = n^{\frac{1}{2r+1}}\log^a(n^{\frac{1}{2r+1}}/\tilde{C})/\tilde{C}$ many features. Recall that $r \in [1/2, 1]$. The improvement in the required number of features is reflected in two aspects. Firstly, when r is greater than $1/2$, QMCF-KRR allows for a substantial reduction in the rate of M, which indicates a sharp diminution in computational costs. For the ease of understanding, when ignoring the constants and logarithmic terms, QMCF-KRR requires $M \asymp n^{\frac{1}{2r+1}}$, which is of smaller order than the $n^{\frac{2r}{2r+1}}$ number of features needed by RF-KRR to achieve the same statistical error rate. Secondly, as QMC generates a non-random sequence of ω_i's, the choice of M in QMCF-KRR does not depend on the "small probability" δ (as opposed to RF-KRR), which facilitates a more straightforward selection of M in practical applications with theoretical guarantees.

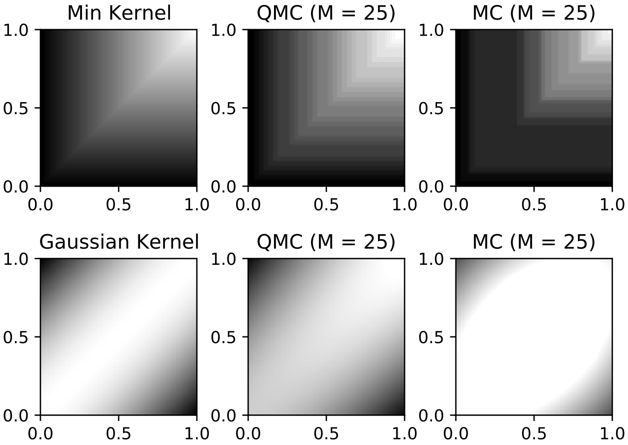

Figure 2. Min kernel and Gaussian kernel over $[0,1]^2$, and their approximated versions using QMC and MC based random features.

4. Simulations

In this section, we demonstrate the usefulness of QMC features in kernel approximation and kernel ridge regression through simulations. Additional simulation studies and real data examples will be referred to Appendix D. Halton sequence implemented in the SciPy package in Python (Virtanen et al., 2020) is used.

4.1. Visualization of Kernel Approximation

We consider the (non-shift invariant) *min kernel* $\min\{x, x'\}$ and the (shift invariant) Gaussian kernel $\exp(-|x-x'|^2)$. For the *min kernel*, we use the integral representation as in Example 1. For Gaussian kernel, we use the integral representation (6) from Bochner's theorem. The values of the kernel functions over $[0,1]^2$ are shown in the left column of Figure 2, with the approximated kernels from QMC and MC shown in the middle column and the right column, respectively. Here the MC plot is produced using one realization of M i.i.d. random features. Figure 2 is in grayscale: brighter pixels correspond to larger values. It can be seen that with $M = 25$ features, QMC method has already provided a reasonably good approximation to the true kernel function. Whereas for MC, the same number of random features are clearly not enough to well approximate the kernel.

4.2. Simulations on Kernel Ridge Regression

In this subsection, we compare the performances of RF-KRR and QMCF-KRR. It has been shown in Theorem 3.2 that the QMC approach is guaranteed to be more efficient than the MC approach in smoother cases, where $r \in [1/2, 1]$ is large. Therefore, we will present simulations for cases where $r = 1$ to support our findings here. In fact, we observe that even in cases where $r = 1/2$, the QMC approach can still have superior performance compared with MC. These additional simulations are given in Appendix D.1.

The training and test data are generated from $Y = f(\mathbf{X}) + \varepsilon$, where f is the regression function, $\mathbf{X} \sim \text{Unif}[0, 1]^d$, and $\varepsilon \sim N(0, 1)$. We consider two choices of kernels: (i) the *min kernel* $K(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d} \min(x_i, x_i')$, and (ii) the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{x}'\|_2^2)$, with the bandwidth σ set as the median of $\|\mathbf{X} - \mathbf{X}'\|$ (computed numerically), where $\mathbf{X}, \mathbf{X}'$ i.i.d. $\sim \text{Unif}[0, 1]^d$.

By definition, a function in $\operatorname{ran} L^r$ for $r = 1$ has the form $\tilde{f}(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{z}) g(\mathbf{z}) \mathrm{d}P_{\mathbf{X}}(\mathbf{z})$ for some $g \in L^2(P_{\mathbf{X}})$. For the *min kernel*, we set $g(\mathbf{z}) = (\prod_{i=1}^{d} z_i)^{\frac{1}{d}}$ as the geometric mean function, which leads to $\tilde{f}(\mathbf{x}) = (\frac{d}{d+1})^d \prod_{j=1}^{d} \left(x_j - \frac{d}{2d+1} x_j^{2+\frac{1}{d}} \right)$, and for the Gaussian kernel, we set $g(\mathbf{z}) = \exp(\frac{1}{2\sigma^2} \|\mathbf{z}\|_2^2)$, which yields $\tilde{f}(\mathbf{x}) = \sigma^{2d} \exp(-\frac{1}{2\sigma^2} \|\mathbf{x}\|_2^2) \prod_{j=1}^{d} \frac{\exp(x_j/\sigma^2) - 1}{x_j}$. To approximately control the signal-noise-ratio, we set the regression function $f(\mathbf{x}) = C_{\tilde{f}} \cdot \tilde{f}(\mathbf{x})$ for some constant $C_{\tilde{f}}$ such that $\mathbb{E} f(\mathbf{X}) = 5$. The kernel ridge regularization parameter is set as $\lambda = 0.25 n^{-\frac{1}{2r+1}}$.

In Figures 3 and 4, we plot the test mean square error (MSE) against the number of random features, for exact KRR, RF-KRR and QMCF-KRR. For each combination of kernel and d, 10^6 test data points are first generated and held fixed. We consider 1000 realizations of training samples of size 10^4. For each of the realization, we fit a kernel ridge regression and compute its test error (i.e., MSE on the test set). The solid lines in Figures 3, 4 are obtained by averaging over the 1000 realizations. We also provide confidence bands using the 25% and 75% error quantiles from the 1000 realizations. For the MC method, the randomness comes from re-generating the training set and the MC random features. Whereas for the QMC method, it only comes from the training set re-generation as the QMC features are deterministic.

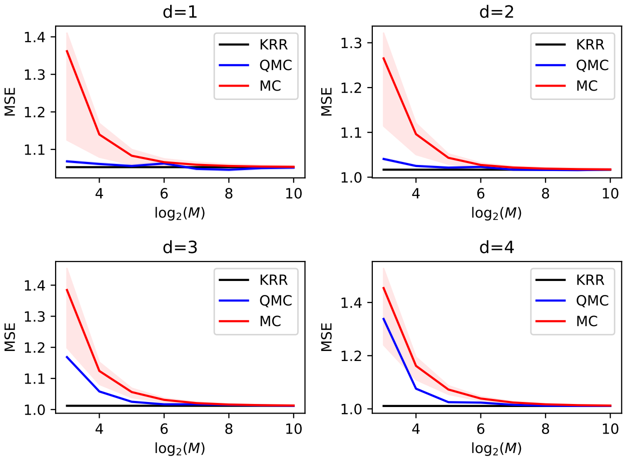

Figure 3. Gaussian Kernel ($r = 1$): the test MSE against the number of random features for exact KRR, RF-KRR and QMCF-KRR.

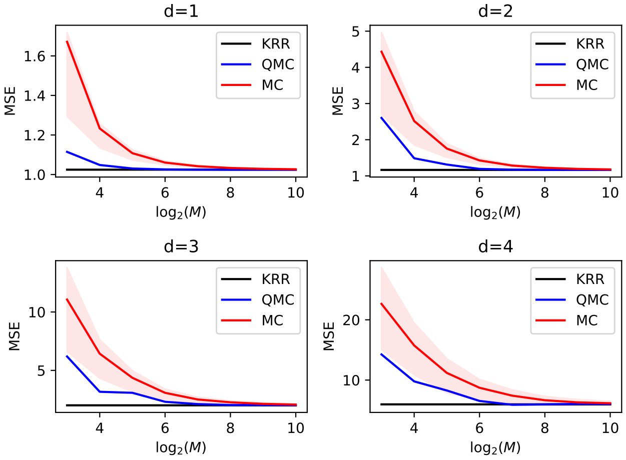

Figure 4. Min Kernel ($r = 1$): the test MSE against the number of random features, for exact KRR, RF-KRR and QMCF-KRR.

It can be seen from Figures 3 and 4 that for all combinations of kernel and d, QMC features exhibit superior performance in kernel ridge regression: it achieves the same generalization error as MC random features with much fewer random features (smaller M), and converges to the error of the exact KRR at a much faster rate. Note that the confidence bands of QMC method are almost negligible, showing very stable performance compared with the MC method, and in many cases, their confidence bands do not intersect. Similar phenomena are observed even when $r = 0.5$. We refer readers to Appendix D.1 for more details.

Superior performance of QMC is also observed in real data, where RF-KRR is often seen to have a wide confidence band. For some real-world examples, see Appendix D.2.

Note that the reported plots (Figures 3, 4) are for $d \in [1, 4]$, which are all low-dimensional, as we observe that QMC features are most effective when the dimension of the ambient space $\mathcal{X}$ is low, which aligns with the practical discovery that the best use case for QMC often arises when the integrand can be well approximated by a sum of functions involving only a small number of its input variables (Owen, 2023; Adcock & Brugiapaglia, 2022). Ongoing research has been studying the high-dimensional situations where QMC methods will be successful (Dick et al., 2013), with a famous empirical finding by Paskov & Taub (1995) that showed integrands from finance in 360 dimensions could be well integrated by QMC. However, for the examples above, we do not observe ideal performance of the QMC features in high-dimensional cases where $d > 10$ (see Appendix D.3). In practice, a dimension reduction prior to applying the methodology may be helpful. Exploring the effectiveness of QMC features in high dimensions could be an interesting research direction to pursue in the future.

Acknowledgements

The authors would like to thank Samory Kpotufe, Bodhisattva Sen, Liu Sifan, and the anonymous reviewers for their careful reading of the paper and helpful suggestions and comments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There might be potential societal consequences of our work, none of which we feel need to be specifically highlighted here.

References

Adcock, B. and Brugiapaglia, S. Is monte carlo a bad sampling strategy for learning smooth functions in high dimensions? *arXiv preprint arXiv:2208.09045*, 2022.

Aronszajn, N. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

Atanassov, E. On the discrepancy of the halton sequences. *Mathematica Balkanica*, 18:15–32, 2004.

Aubin, J.-P. *Applied Functional Analysis*, volume 47. John Wiley & Sons, 2011.

Avron, H., Sindhwani, V., Yang, J., and Mahoney, M. W. Quasi-monte carlo feature maps for shift-invariant kernels. *Journal of Machine Learning Research*, 17(1):4096–4133, 2016.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning*, pp. 253–262. PMLR, 2017.

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

Baker, R. C. On irregularities of distribution. II. *J. London Math. Soc. (2)*, 59(1):50–64, 1999.

Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(11), 2006.

Bochner, S. Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108 (1):378–410, 1933.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Chamakh, L., Gobet, E., and Szabó, Z. Orlicz random fourier features. *The Journal of Machine Learning Research*, 21(1):5739–5775, 2020.

Choromanski, K., Rowland, M., Sarlos, T., Sindhwani, V., Turner, R., and Weller, A. The geometry of random features. In *International Conference on Artificial Intelligence and Statistics*, pp. 1–9. PMLR, 2018.

Cohn, D. L. *Measure Theory*. Springer, New York, second edition, 2013.

Courant, R. and Hilbert, D. *Methods of Mathematical Physics. Vol. I.* Interscience Publishers, Inc., New York, N.Y., 1953.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

De Vito, E., Caponnetto, A., and Rosasco, L. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5: 59–85, 2005.

Dick, J., Kuo, F. Y., and Sloan, I. H. High-dimensional integration: the quasi-monte carlo way. *Acta Numerica*, 22:133–288, 2013.

Faure, H. Discrépance de suites associées à un système de numération (en dimension s). *Acta Arithmetica*, 41(4): 337–351, 1982.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5 (Jan):73–99, 2004.

Fukumizu, K., Bach, F. R., and Jordan, M. I. Kernel dimension reduction in regression. *Ann. Statist.*, 37(4): 1871–1905, 2009.

Fukumizu, K., Song, L., and Gretton, A. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Halton, J. H. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12): 701–702, 1964.

Hein, M. and Bousquet, O. Kernels, Associated Structures and Generalizations. Technical Report of the Max Planck Institute for Biological Cybernetics 127, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2004.

Hlawka, E. Funktionen von beschränkter variatiou in der theorie der gleichverteilung. *Annali di Matematica Pura ed Applicata*, 54(1):325–333, 1961.

Huang, Z., Deb, N., and Sen, B. Kernel partial correlation coefficient — a measure of conditional dependence. *Journal of Machine Learning Research*, 23(216):1–58, 2022.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pp. 4631–4640. PMLR, 2020.

Kimeldorf, G. S. and Wahba, G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502, 1970.

Klebanov, I., Schuster, I., and Sullivan, T. J. A rigorous theory of conditional mean embeddings. *SIAM J. Math. Data Sci.*, 2(3):583–606, 2020.

Korobov, N. M. *Number-Theoretic Methods in Approximate Analysis*. Fizmatgiz, Moscow, 1963.

Lanthaler, S. and Nelsen, N. H. Error bounds for learning with vector-valued random features. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71834–71861, 2023.

Leobacher, G. and Pillichshammer, F. *Introduction to Quasi-Monte Carlo Integration and Applications*. Compact Textbooks in Mathematics. Birkhäuser/Springer, Cham, 2014.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. Towards a unified analysis of random Fourier features. In *International Conference on Machine Learning*, pp. 3905–3914. PMLR, 2019.

Liao, Z., Couillet, R., and Mahoney, M. W. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13939–13950, 2020.

Liu, F., Huang, X., Chen, Y., and Suykens, J. A. K. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2022.

Lyu, Y. Spherical structured feature maps for kernel approximation. In *International Conference on Machine Learning*, pp. 2256–2264. PMLR, 2017.

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022. Special Issue on Harmonic Analysis and Machine Learning.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

Musco, C. and Musco, C. Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Niederreiter, H. Point sets and sequences with small discrepancy. *Monatsh. Math.*, 104(4):273–337, 1987.

Niederreiter, H. *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, PA, 1992.

Owen, A. B. Multidimensional variation for quasi-monte carlo. In *Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday*, pp. 49–74. World Scientific, 2005.

Owen, A. B. Halton sequences avoid the origin. *SIAM Review*, 48(3):487–503, 2006.

Owen, A. B. *Practical Quasi-Monte Carlo Integration*. https://artowen.su.domains/mc/practicalqmc.pdf, 2023.

Pace, R. K. and Barry, R. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

Paskov, S. H. and Taub, J. F. Faster valuation of financial derivatives. *Journal of Portfolio Management*, 22(1):113, 1995.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Reid, I., Choromanski, K., and Weller, A. Quasi-monte carlo graph random features. *arXiv preprint arXiv:2305.12470*, 2023.

Ritter, K. *Average-Case Analysis of Numerical Problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

Roth, K. F. On irregularities of distribution. *Mathematika*, 1:73–79, 1954.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Rynne, B. P. and Youngson, M. A. *Linear Functional Analysis*. Springer Undergraduate Mathematics Series. Springer-Verlag London, Ltd., London, second edition, 2008.

Schmidt, W. Irregularities of distribution, vii. *Acta Arithmetica*, 21(1):45–50, 1972.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pp. 416–426. Springer, 2001.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, 41(5):2263–2291, 2013.

Smale, S. and Zhou, D.-X. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01): 17–41, 2003.

Smola, A. J. Sparse greedy matrix approximation for machine learning. In *International Conference on Machine Learning*. Morgan Kaufmann, 2000.

Sobol', I. M. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 7(4): 784–802, 1967.

Sriperumbudur, B. and Sterge, N. Approximate kernel pca using random features: Computational vs. statistical tradeoff. *The Annals of Statistics*, 50(5):2713 – 2736, 2022.

Sriperumbudur, B. and Szabó, Z. Optimal rates for random fourier features. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

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

Sutherland, D. J. and Schneider, J. On the error of random fourier features. In *Conference on Uncertainty in Artificial Intelligence*, pp. 862–871, 2015.

Szabó, Z. and Sriperumbudur, B. On kernel derivative approximation with random fourier features. In *International Conference on Artificial Intelligence and Statistics*, pp. 827–836. PMLR, 2019.

Ullah, E., Mianjy, P., Marinov, T. V., and Arora, R. Streaming kernel PCA with $\tilde{O}(\sqrt{n})$ random features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Uzilov, A. V., Keegan, J. M., and Mathews, D. H. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7:1–30, 2006.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Wahba, G. *Spline Models for Observational Data*. SIAM, 1990.

Williams, C. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Yang, J., Sindhwani, V., Avron, H., and Mahoney, M. Quasi-monte carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, pp. 485–493. PMLR, 2014.

A. Some Further Discussions

In this section, we elaborate on some parts of the main text that were initially deferred so as not to impede the flow of the paper.

A.1. Kernels Satisfying QMC Condition 1

Here we provide a proof for Proposition 2.1 (i.e., Gaussian kernel and Cauchy kernel over a compact domain satisfy QMC Condition 1).

A.1.1. Gaussian Kernel

To show that Gaussian kernel over a compact domain satisfies QMC Condition 1, we first show the following lemma:

Lemma A.1. *If Φ is the distribution function of the standard normal distribution, then there exists $C > 0$ such that*

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{-1}(t) \le C \min(t, 1-t)^{-1}, \quad \text{for all } t \in (0,1). \tag{A.1}$$

Proof. Note that $\frac{\mathrm{d}}{\mathrm{d}t}\Phi^{-1}(t) = \frac{1}{\phi(\Phi^{-1}(t))} = \frac{1}{\phi(\Phi^{-1}(1-t))}$, where ϕ is the density function of the standard normal distribution. Therefore, by the symmetry about $t = 1/2$, we only need to consider $t \in (0, 1/2]$ in (A.1), which reduces to

$$\frac{1}{\phi(\Phi^{-1}(t))} \le C t^{-1}.$$

Let $x = \Phi^{-1}(t) \in (-\infty, 0]$. The above inequality is equivalent to

$$\Phi(x) \le C\phi(x).$$

If $x \le -1$, then

$$\frac{\Phi(x)}{\phi(x)} = \frac{\int_{-\infty}^{x} \phi(u)\mathrm{d}u}{\phi(x)} < \frac{\frac{1}{\sqrt{2\pi}}\frac{1}{|x|}e^{-\frac{x^2}{2}}}{\phi(x)} = \frac{1}{|x|} \le 1.$$

If $-1 < x \le 0$, then

$$\frac{\Phi(x)}{\phi(x)} \le \frac{\Phi(0)}{\phi(-1)}.$$

Hence, by taking $C = \max\left(1, \frac{\Phi(0)}{\phi(-1)}\right)$, the inequality (A.1) holds. □

For a Gaussian kernel with a general bandwidth, the cumulative distribution function $\Phi_i(t)$ (as defined in QMC Condition 1) is given by $\Phi_i(t) = \Phi(t/\sigma)$ for some $\sigma > 0$. Therefore, $\Phi_i^{-1}(t) = \sigma\Phi^{-1}(t)$. By Lemma A.1, $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_i^{-1}(t) \le \frac{C_i}{\min(t,1-t)}$ for all $t \in (0,1)$, where $C_i = \sigma C$.

A.1.2. Cauchy Kernel

For Cauchy kernel $K(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d} \frac{1}{1+\lambda^2(x_i - x_i')^2}$, $\Phi_i(t)$ (as defined in QMC Condition 1) is the cumulative distribution function of the symmetrized exponential distribution (a.k.a. Laplace distribution) with Lebesgue density $\phi_i(x) = \frac{\lambda}{2}e^{-\lambda|x|}$. Therefore, $\Phi_i(x) = \frac{1}{2}e^{\lambda x}1_{x\le 0} + \left(1 - \frac{1}{2}e^{-\lambda x}\right)1_{x>0}$, and $\Phi_i^{-1}(t) = \lambda^{-1}(\log(2t)1_{t\le 1/2} - \log(2(1-t))1_{t\ge 1/2})$. By taking the derivative w.r.t. t, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_i^{-1}(t) = \lambda^{-1}\left(\frac{1}{t}1_{t\le 1/2} + \frac{1}{1-t}1_{t\ge 1/2}\right) \le 2\lambda^{-1}\min(t, 1-t)^{-1}.$$

Therefore, Cauchy kernel over a compact domain satisfies QMC Condition 1. □

A.2. Kernels Satisfying QMC Condition 2

A.2.1. Iterative Kernel

Suppose $K_1(\cdot,\cdot)$ is a continuous kernel on $[0,1]^d$, and μ is a positive integrable function on $[0,1]^d$. The iterative kernel (Courant & Hilbert, 1953, Section III.5.3) of K_1 is defined as

$$K_2(\mathbf{x},\mathbf{z}) := \int_{[0,1]^d} K_1(\mathbf{x},\mathbf{t})K_1(\mathbf{z},\mathbf{t})\mu(\mathbf{t})\mathrm{d}\mathbf{t}.$$

QMC Condition 2 holds if the Hardy-Krause variation of $f_{\mathbf{x},\mathbf{z}}(\mathbf{t}) := K_1(\mathbf{x},\mathbf{t})K_1(\mathbf{z},\mathbf{t})\mu(\mathbf{t})$ is bounded by some $C_0 > 0$ for all $\mathbf{x},\mathbf{z} \in [0,1]^d$. A sufficient condition for the existence of such a C_0 is that there exists a constant $\tilde{C}_0 > 0$ such that for all $u \subset \{1,\ldots,d\}$ and $\mathbf{x},\mathbf{z},\mathbf{t} \in [0,1]^d$, $|(\prod_{i\in u}\frac{\partial}{\partial t_i})f_{\mathbf{x},\mathbf{z}}(\mathbf{t})| \le \tilde{C}_0$, due to the fact that (Niederreiter, 1992, Section 2.2):

$$V_{\mathrm{HK}}(f;[0,1]^d) = \sum_{I\subset\{1,\ldots,d\}, I\neq\emptyset} \int_{[0,1]^{|I|}} \left| \left.\frac{\partial f}{\partial u_I}\right|_{u_j=1, j\notin I} \right| \mathrm{d}u_I.$$

Note that μ can be viewed as a Lebesgue density function and thus induces a strictly positive and finite Borel measure on $[0,1]^d$, with the measure of a set B defined as $\int_B \mu(\mathbf{x})\mathrm{d}\mathbf{x}$. Recall the integral operator $L : L^2(\mu) \to L^2(\mu)$ defined as:

$$Lf(\mathbf{x}) := \int_{[0,1]^d} K_1(\mathbf{x},\mathbf{t})f(\mathbf{t})\mu(\mathbf{t})\mathrm{d}\mathbf{t}.$$

Mercer's theorem (Steinwart & Scovel, 2012) implies that there exists a continuous orthonormal basis $\{e_i\}$ of $L^2(\mu)$ consisting of eigenfunctions of the integral operator L, with corresponding nonnegative eigenvalues $\{\lambda_i\}$ satisfying $\sum_i \lambda_i < +\infty$. Moreover,

$$K_1(\mathbf{x},\mathbf{y}) = \sum_i \lambda_i e_i(\mathbf{x})e_i(\mathbf{y}),$$

where the convergence is absolute and uniform.

Observe that

$$K_2(\mathbf{x},\mathbf{z}) = \int_{[0,1]^d} K_1(\mathbf{x},\mathbf{t})K_1(\mathbf{z},\mathbf{t})\mu(\mathbf{t})\mathrm{d}\mathbf{t} = \int_{[0,1]^d} K_1(\mathbf{x},\mathbf{t})\sum_i \lambda_i e_i(\mathbf{t})e_i(\mathbf{z})\mu(\mathbf{t})\mathrm{d}\mathbf{t}.$$

Since $\sup_{\mathbf{x},\mathbf{t}\in[0,1]^d} |K_1(\mathbf{x},\mathbf{t})| < \infty$ by the continuity of K_1, and the convergence of Mercer's series is absolute and uniform, we can exchange the summation and integration to obtain

$$\begin{aligned}
K_2(\mathbf{x},\mathbf{z}) &= \sum_i \int_{[0,1]^d} K_1(\mathbf{x},\mathbf{t})\lambda_i e_i(\mathbf{t})e_i(\mathbf{z})\mu(\mathbf{t})\mathrm{d}\mathbf{t} \\
&= \sum_i \lambda_i e_i(\mathbf{z}) \int_{[0,1]^d} K_1(\mathbf{x},\mathbf{t})e_i(\mathbf{t})\mu(\mathbf{t})\mathrm{d}\mathbf{t} \\
&= \sum_i \lambda_i e_i(\mathbf{z}) L e_i(\mathbf{x}) \\
&= \sum_i \lambda_i^2 e_i(\mathbf{z})e_i(\mathbf{x}).
\end{aligned}$$

We can see that the integral operators of K_1 and K_2 share the same set of eigenfunctions, while the eigenvalues of K_2 (i.e., $\{\lambda_i^2\}$) decay faster than those of K_1 (i.e., $\{\lambda_i\}$). □

A.2.2. Product Kernel

Here we give a proof for Example 5 in Section 2.2: if for $i = 1,\ldots,d$, $K_i(u,v) = \int_0^1 \psi_i(u,t)\psi_i(v,t)\mathrm{d}t$ satisfies QMC Condition 2 and $|\psi_i(u,t)| \le \kappa_i$ for some κ_i and all u,t, then

$$K(\mathbf{u},\mathbf{v}) := \prod_{i=1}^d K_i(u_i,v_i) = \int_{[0,1]^d} \psi(\mathbf{u},\mathbf{t})\psi(\mathbf{v},\mathbf{t})\mathrm{d}\mathbf{t},$$

where $\psi(\mathbf{x},\mathbf{t}) = \prod_{i=1}^d \psi_i(x_i, t_i)$, satisfies QMC Condition 2.

Definition A.2 (Vitali Variation)**.** Consider a rectangle $[\mathbf{a},\mathbf{b}] := [a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d$ and a function $f : [\mathbf{a},\mathbf{b}] \to \mathbb{R}$. Define

$$\Delta_{h_k}(f,\mathbf{x}) := f(x_1,\ldots,x_k + h_k,\ldots,x_d) - f(x_1,\ldots,x_k,\ldots,x_d),$$

and, recursively,

$$\Delta_{h_1 h_2 \ldots h_k}(f,\mathbf{x}) := \Delta_{h_k}\left(\Delta_{h_1\ldots h_{k-1}},\mathbf{x}\right).$$

Consider the collection Π_k of finite ordered families π_k of points $t_k^1 < t_k^2 < \ldots < t_k^{N_k+1} \in [a_k, b_k]$. Denote $h_k^i = t_k^{i+1} - t_k^i$. The *Vitali variation* of f is defined as

$$V_{[\mathbf{a},\mathbf{b}]}(f) := \sup_{(\pi_1,\ldots,\pi_d)\in\Pi_1\times\cdots\times\Pi_d} \sum_{i_1=1}^{N_1} \cdots \sum_{i_n=1}^{N_d} \left|\Delta_{h_1^{i_1}\cdots h_n^{i_d}}\left(f,(x_1^{i_1},\ldots,x_n^{i_d})\right)\right|.$$

Note that on $\mathbb{R}^1$, Vitali variation coincides with the usual notion of total variation. If $f_i : [a_i, b_i] \to \mathbb{R}$ has total variation $V_{[a_i,b_i]}(f_i)$, then from Definition A.2 we deduce that $f(\mathbf{x}) = \prod_{i=1}^d f_i(x_i)$ has Vitali variation $V_{[\mathbf{a},\mathbf{b}]}(f) = \prod_{i=1}^d V_{[a_i,b_i]}(f_i)$. With this fact, we can prove the following proposition.

Proposition A.3. *Consider a rectangle* $[\mathbf{a},\mathbf{b}] := [a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d$*, and* $f_i : [a_i, b_i] \to \mathbb{R}$ *has total variation* $V_{[a_i,b_i]}(f_i) \le C_i$ *and* $|f_i(b_i)| \le \kappa_i$*. Then the Hardy-Krause variation of* $f(\mathbf{x}) := \prod_{i=1}^d f_i(x_i)$ *is bounded by* $C = \sum_{u\subset\{1,\ldots,d\},u\neq\emptyset} \prod_{i\in u} C_i \prod_{j\notin u} \kappa_j$.

Proof. By Owen (2005, Definition 2), the Hardy-Krause variation can be written in terms of Vitali variation:

$$V_{\text{HK}}(f;[\mathbf{a},\mathbf{b}]) = \sum_{u\subset\{1,\ldots,d\},u\neq\emptyset} V_{[\mathbf{a}_u,\mathbf{b}_u]} f(\mathbf{x}_u;\mathbf{b}_{-u}) \le \sum_{u\subset\{1,\ldots,d\},u\neq\emptyset} \prod_{i\in u} C_i \prod_{j\notin u} \kappa_j.$$

Here, $-u$ denotes the complement $\{1,\ldots,d\}\backslash u$; $\mathbf{x}_u$ denotes the sub-vector from $\mathbf{x}$ by taking all x_j with $j \in u$; $\mathbf{x}_u : \mathbf{b}_{-u}$ denotes a vector $\mathbf{y} \in \mathbb{R}^d$ with $y_j = x_j$ for $j \in u$ and $y_j = b_j$ for $j \notin u$; and when $\mathbf{b}_{-u}$ is held fixed, $f(\mathbf{x}_u : \mathbf{b}_{-u})$ is a function of $\mathbf{x}_u$, and this function is denoted by $f(\mathbf{x}_u;\mathbf{b}_{-u})$. □

By applying Proposition A.3 to $f_i(t) = \psi_i(u,t)\psi_i(v,t)$ and $[\mathbf{a},\mathbf{b}] = [0,1]^d$, we see that the product kernel in Example 5 satisfies QMC Condition 2.

A.3. A Review of Concepts from Functional Analysis

We start with a brief review of some notions from functional analysis that will be important for subsequent discussions; see e.g., Rynne & Youngson (2008); Aubin (2011) for a detailed study of these concepts. By a *bounded linear operator* A from a Hilbert space $\mathcal{F}$ to a Hilbert space $\mathcal{G}$ we mean a linear map $A : \mathcal{F} \to \mathcal{G}$ such that, for some $L \ge 0$, $\|Av\|_{\mathcal{G}} \le L\|v\|_{\mathcal{F}}$, for all $v \in \mathcal{F}$. The smallest L such that the previous inequality holds is called the *operator norm* of A, denoted by $\|A\|$. There is a unique bounded operator $A^* : \mathcal{G} \to \mathcal{F}$, called the *adjoint* of A, such that $\langle u, Av\rangle_{\mathcal{G}} = \langle A^*u, v\rangle_{\mathcal{F}}$, for all $u \in \mathcal{G}, v \in \mathcal{F}$. Let $\operatorname{ran} A := \{Av : v \in \mathcal{F}\}$ denote the *range* of the operator A and $\ker A := \{v \in \mathcal{F} : Av = 0\}$ denote the *kernel* of A. A bounded linear operator A from a Hilbert space $\mathcal{H}$ to itself is *nonnegative* if $\langle u, Au\rangle_{\mathcal{H}} \ge 0$ for all $u \in \mathcal{H}$. We say that A is *self-adjoint* if $A^* = A$. Suppose that $\mathcal{H}$ is separable with orthonormal basis $\{e_i\}_{i\ge1}$. Then the *trace* of a non-negative A is defined as $\operatorname{tr}(A) := \sum_i \langle Ae_i, e_i\rangle_{\mathcal{H}}$. For a nonnegative operator A, if $\operatorname{tr}(A) < \infty$, then A is said to be a *trace-class* operator. A *compact* operator from a Hilbert space $\mathcal{F}$ to another Hilbert space $\mathcal{G}$ is a linear operator L such that the image under L of any bounded subset of $\mathcal{F}$ is a relatively compact subset (has compact closure) of $\mathcal{G}$. Such an operator is necessarily a bounded operator, and thus is continuous. For a bounded linear operator L from a Hilbert space $\mathcal{F}$ to another Hilbert space $\mathcal{G}$, L is compact if and only if L^* is compact, and they are also equivalent to LL^* being compact or L^*L being compact. For every compact self-adjoint operator L from a Hilbert space $\mathcal{H}$ to itself, the spectral theorem states that there exists an orthonormal basis $\{\mathbf{x}_i\}$ of $\mathcal{H}$ consisting of eigenvectors of L, i.e., $L\mathbf{x}_i = \lambda_i\mathbf{x}_i$ where nonzero λ_i's are at most countable and converge to 0. If furthermore L is nonnegative, then L^r with $r > 0$ can be defined by $L^r\mathbf{x}_i = \lambda_i^r\mathbf{x}_i$ for all i.

Let $\mathcal{F}$ and $\mathcal{G}$ be separable Hilbert spaces. Let $\{f_i\}_{i\in I}$ to be an orthonormal basis for $\mathcal{F}$, and let $\{g_j\}_{j\in J}$ be an orthonormal basis for $\mathcal{G}$; here I and J are indexing sets being either finite or countably infinite.

Definition A.4 (Hilbert-Schmidt operators)**.** The Hilbert-Schmidt norm of a compact operator $L : \mathcal{G} \to \mathcal{F}$ is defined to be

$$\|L\|_{\mathrm{HS}}^2 := \sum_{j \in J} \|Lg_j\|_{\mathcal{F}}^2.$$

The operator L is Hilbert-Schmidt when this norm is finite. The Hilbert-Schmidt operators mapping from $\mathcal{G}$ to $\mathcal{F}$ form a Hilbert space, written $\mathrm{HS}(\mathcal{G}, \mathcal{F})$, with inner product

$$\langle L, M \rangle_{\mathrm{HS}} := \sum_{j \in J} \langle Lg_j, Mg_j \rangle_{\mathcal{F}}, \tag{A.2}$$

which is independent of the orthonormal basis chosen. Here $L : \mathcal{G} \to \mathcal{F}$ and $M : \mathcal{G} \to \mathcal{F}$ are two Hilbert-Schmidt operators.

A.4. A Review of Mercer Kernels

By a kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ we mean a symmetric and nonnegative definite function such that $K(\mathbf{x}, \cdot)$ is a (real-valued) measurable function on $\mathcal{X}$, for all $\mathbf{x} \in \mathcal{X}$. Denote by $\mathcal{H}$ the reproducing kernel Hilbert space associated with K, and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the associated inner product . Then $\mathcal{H}$ is a Hilbert space of real-valued functions on $\mathcal{X}$ such that, for any $f \in \mathcal{H}$, we have $f(\mathbf{x}) = \langle f, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$, for all $\mathbf{x} \in \mathcal{X}$; this is usually referred to as the *reproducing property* of the kernel K. Let μ be a measure on $\mathcal{X}$.

Definition A.5 (Mercer kernel)**.** We call a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a *Mercer kernel* (w.r.t. μ) if it is continuous and $\int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) < \infty$.

If $\mathcal{X}$ is a separable space and K is continuous, then $\mathcal{H}$ is separable; see e.g., Hein & Bousquet (2004, Theorem 7), Steinwart & Christmann (2008, Lemma 4.33). Therefore, a Mercer kernel on $\mathcal{X} \subset \mathbb{R}^d$ automatically induces a separable RKHS.

Definition A.6 (Inclusion operator)**.** Suppose K is a Mercer kernel. Define the *inclusion operator* $I : \mathcal{H} \to L^2(\mu)$ by identifying a function in $\mathcal{H}$ as a function in $L^2(\mu)$, i.e., $I(f) = f \in L^2(\mu)$. Note that $\|If\|_{L^2(\mu)}^2 = \int_{\mathcal{X}} f^2(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \leq \int_{\mathcal{X}} \|f\|_{\mathcal{H}}^2 \cdot \|K(\mathbf{x}, \cdot)\|_{\mathcal{H}}^2 \mathrm{d}\mu(\mathbf{x}) = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) \cdot \|f\|_{\mathcal{H}}^2$, which implies I is a bounded linear operator. Its adjoint operator $I^* : L^2(\mu) \to \mathcal{H}$ is given by:

$$I^* g(\mathbf{x}) = \langle I^* g, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = \langle g, IK(\mathbf{x}, \cdot) \rangle_{L^2(\mu)} = \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{t}) g(\mathbf{t}) \mathrm{d}\mu(\mathbf{t}).$$

An important property of the inclusion operator I is that it is injective as long as μ is supported on the entire $\mathcal{X}$. To see this, suppose $f \in \mathcal{H}$ with $If = 0$. The continuity of K implies that f is continuous. If a continuous function has $L^2(\mu)$ norm being 0, then it must be identically 0. Hence, the injectivity is shown.

It follows from a direct verification that the integral operator defined in (4) is $L = II^*$. Let $L^{1/2}$ be the unique self-adjoint nonnegative square root operator of L (see Appendix A.3 for the definition). Then $L^{1/2}$ induces an isometry from $L^2(\mu)$ to $I(\mathcal{H})$:

Theorem A.7 (Isometry induced by $L^{1/2}$)**.** *Suppose K is a Mercer kernel (w.r.t. μ) and μ has full support on $\mathcal{X}$. Then $L^{1/2}$ induces an isometry from $(\ker L^{1/2})^\perp$ to $I(\mathcal{H})$, where $I(\mathcal{H})$ is equipped with the inner product $\langle I(f), I(g) \rangle := \langle f, g \rangle_{\mathcal{H}}$ for $f, g \in \mathcal{H}$.*

Proof. Let I be the inclusion operator defined above, and $\{e_n\}_{n \geq 1}$ be an orthonormal basis of $\mathcal{H}$. Observe that

$$\begin{aligned} \operatorname{tr}(I^* I) &= \sum_n \langle e_n, I^* I e_n \rangle_{\mathcal{H}} = \sum_n \langle Ie_n, Ie_n \rangle_{L^2(\mu)} = \sum_n \int e_n^2(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}) = \sum_n \int \langle K(\mathbf{x}, \cdot), e_n \rangle_{\mathcal{H}}^2 \mathrm{d}\mu(\mathbf{x}) \\ &= \int \|K(\mathbf{x}, \cdot)\|_{\mathcal{H}}^2 \mathrm{d}\mu(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{x}) \mathrm{d}\mu(\mathbf{x}) < \infty. \end{aligned}$$

Hence, $I^* I$ is trace-class and therefore a compact operator. As a result, I, I^*, II^* are all compact operators. By the spectral theorem of the compact self-adjoint operator on a Hilbert space, there exists an orthonormal basis $\{f_i\}_{i \geq 1}$ of $L^2(\mu)$ which are eigenvectors of II^*. Assume that $\{\lambda_i\}_{i \geq 1}$ are the corresponding eigenvalues, i.e., $II^* f_i = \lambda_i f_i$ for all i. Since I is injective (by the fact that K is a Mercer kernel and μ has full support), we have $(\operatorname{ran} I^*)^\perp = \ker I = \{0\}$. Therefore, the closure of $\operatorname{ran} I^*$ is the entire $\mathcal{H}$. Moreover, $\langle I^* f_i, I^* f_j\rangle_{\mathcal{H}} = \langle f_i, II^* f_j\rangle_{L^2(\mu)} = \langle f_i, \lambda_j f_j\rangle_{L^2(\mu)} = 0$ for $i \neq j$, and $\langle I^* f_i, I^* f_i\rangle_{\mathcal{H}} = \langle f_i, II^* f_i\rangle_{L^2(\mu)} = \langle f_i, \lambda_i f_i\rangle_{L^2(\mu)} = \lambda_i$. Hence, $\{g_i = I^* f_i/\sqrt{\lambda_i} : \lambda_i > 0\}$ is an orthonormal basis of $\mathcal{H}$.

We first establish the bijection between $(\ker L^{1/2})^\perp$ and $I(\mathcal{H})$. Any function in $(\ker L^{1/2})^\perp$ can be uniquely written as $\sum_i a_i f_i$ with $\sum_i a_i^2 < \infty$, where f_i corresponds to a positive eigenvalue $\lambda_i > 0$. This function is mapped by $L^{1/2}$ to $\sum_i \sqrt{\lambda_i} a_i f_i$. On the other hand, any function in $\mathcal{H}$ can be uniquely written as $\sum_i a_i g_i$ with $\sum_i a_i^2 < \infty$, which gets mapped to $\sum_i \sqrt{\lambda_i} a_i f_i$ by I. Therefore, a bijection between $(\ker L^{1/2})^\perp$ and $\mathcal{H}$ is established, by mapping $\sum_i a_i f_i$ to $\sum_i a_i g_i$. Note that for any $\{a_i\}_{i\geq 1}$, $\{b_i\}_{i\geq 1}$ satisfying $\sum_i a_i^2 < \infty$ and $\sum_i b_i^2 < \infty$, we have

$$\left\langle \sum_i a_i f_i, \sum_i b_i f_i \right\rangle_{L^2(\mu)} = \sum_i a_i b_i = \left\langle \sum_i a_i g_i, \sum_i b_i g_i \right\rangle_{\mathcal{H}}.$$

This shows that the bijection preserves inner products, and is therefore an isometry. □

A direct consequence of Theorem A.7 is the existence of an integral representation in the form of (1):

Proposition A.8 (Existence of integral representation)**.** *Suppose K is a Mercer kernel and μ has full support on $\mathcal{X}$. Then there exists $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that*

$$K(\mathbf{x}, \mathbf{x}') = \int_{\mathcal{X}} \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\mu(\omega).$$

Proof. From Theorem A.7, there exists an isometry between $\mathcal{H}$ and a subspace of $L^2(\mu)$. Suppose $K(\mathbf{x}, \cdot)$ is mapped to $\psi(\mathbf{x}, \cdot) \in L^2(\mu)$ under this isometry. Then

$$K(\mathbf{x}, \mathbf{x}') = \langle K(\mathbf{x}, \cdot), K(\mathbf{x}', \cdot)\rangle_{\mathcal{H}} = \langle \psi(\mathbf{x}, \cdot), \psi(\mathbf{x}', \cdot)\rangle_{L^2(\mu)} = \int_{\mathcal{X}} \psi(\mathbf{x}, \omega)\psi(\mathbf{x}', \omega)\mathrm{d}\mu(\omega).$$

□

B. Proofs of the Results in Section 2

The main idea of the proof of Theorem 2.2 is to find a low variation function $\tilde{f}$ that coincides with the integrand f in (6) on a "large set", and apply the classical Koksma-Hlawka inequality (Theorem 1.1) to $\tilde{f}$. QMC Condition 1 is used to control both $V_{\mathrm{HK}}(\tilde{f})$ and the behavior the integrand outside the "large set". The fact that Halton sequence avoids the boundary of the unit cube (Owen, 2006) will be useful, which allows us to claim $\tilde{f} = f$ on the first n points of the Halton sequence.

B.1. Proof of Theorem 2.2

We first introduce some notations: for a non-empty set $u \subset \{1, \dots, d+1\}$ and a function f on $\mathbb{R}^{d+1}$, $\partial^u f(\mathbf{x})$ represents $(\prod_{j\in u} \partial/\partial x_j) f(\mathbf{x})$. Let $-u$ denote the complement $\{1, \dots, d+1\}\backslash u$, and $\mathbf{x}_u$ denote the sub-vector from $\mathbf{x}$ by taking all x_j with $j \in u$. For $\mathbf{x}, \mathbf{z} \in [0,1]^{d+1}$, $\mathbf{x}_u : \mathbf{z}_{-u}$ denotes a vector $\mathbf{y} \in \mathbb{R}^{d+1}$ with $y_j = x_j$ for $j \in u$ and $y_j = z_j$ for $j \notin u$. When $\mathbf{z}_{-u}$ is held fixed, then $f(\mathbf{x}_u : \mathbf{z}_{-u})$ is a function of $\mathbf{x}_u$, and this function is denoted by $f(\mathbf{x}_u; \mathbf{z}_{-u})$.

Observe that the integrand in (6) can be re-written as

$$f(\mathbf{t}, b) = \cos\left((\mathbf{x} - \mathbf{x}')^\top \mathbf{\Phi}^{-1}(\mathbf{t})\right) - \cos\left((\mathbf{x} + \mathbf{x}')^\top \mathbf{\Phi}^{-1}(\mathbf{t}) + 4\pi b\right).$$

Let $D = \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}, i\in\{1,\dots,d\}} \{|x_i - y_i|, |x_i + y_i|\}$. Then for any non-empty set $u \subset \{1, \dots, d+1\}$ and $(\mathbf{t}, b) \in (0,1)^{d+1}$,

$$|\partial^u f(\mathbf{t}, b)| \leq 4\pi D^{|u\backslash\{d+1\}|} \prod_{i\in u\backslash\{d+1\}} \frac{\mathrm{d}}{\mathrm{d}t_i} \Phi_i^{-1}(t_i).$$

Theorem 2.2 is a direct consequence of the following lemma:

Lemma B.1. *Suppose* $f : (0,1)^{d+1} \to \mathbb{R}$ *satisfies* $|\partial^u f(\mathbf{x})| \leq B \prod_{i \in u \setminus \{d+1\}} \frac{1}{\min(x_i, 1-x_i)}$ *for some* $B > 0$ *and any non-empty set* $u \subset \{1, \ldots, d+1\}$*. Let* $\mathbf{h}_1, \ldots, \mathbf{h}_n$ *be the first* n *points of the* $(d+1)$*-dimensional Halton sequence. There exists a constant* $C(d)$ *depending only on* d *such that for all* $n \geq 2$*,*

$$\left| \int_{[0,1]^{d+1}} f(\mathbf{x}) \mathrm{d}\mathbf{x} - \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{h}_i) \right| \leq BC(d) \frac{(\log n)^{2d+1}}{n}.$$

Proof of Lemma B.1. The main idea is to find a low variation function $\tilde{f}_n$ that coincides with f on a "large set" K_n, so that the classical Koksma-Hlawka inequality (Theorem 1.1) can be applied to $\tilde{f}_n$. Outside K_n, f can be controlled with the given bounds on the derivatives. Note that if $\mathbf{h}_1, \ldots, \mathbf{h}_n \in K_n$, then

$$\begin{aligned} \left| \int_{[0,1]^{d+1}} f(\mathbf{x}) \mathrm{d}\mathbf{x} - \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{h}_i) \right| &\leq \int_{[0,1]^{d+1}} |f(\mathbf{x}) - \tilde{f}_n(\mathbf{x})| \mathrm{d}\mathbf{x} + \mathcal{D}^*(\{\mathbf{h}_i\}_{i=1}^n) V_{\mathrm{HK}}(\tilde{f}_n) + \frac{1}{n} \sum_{i=1}^{n} |\tilde{f}_n(\mathbf{h}_i) - f_n(\mathbf{h}_i)| \\ &= \int_{[0,1]^{d+1}} |f(\mathbf{x}) - \tilde{f}_n(\mathbf{x})| \mathrm{d}\mathbf{x} + \mathcal{D}^*(\{\mathbf{h}_i\}_{i=1}^n) V_{\mathrm{HK}}(\tilde{f}_n). \end{aligned} \tag{B.1}$$

We will construct $\tilde{f}_n, K_n$, and bound each term on the right-hand side above.

Let $\mathbf{c} = (1/2, \ldots, 1/2)^\top \in \mathbb{R}^{d+1}$. We can write $f(\mathbf{x})$ as an integral

$$f(\mathbf{x}) = f(\mathbf{c}) + \sum_{u \neq \emptyset} \int_{[\mathbf{c}_u, \mathbf{x}_u]} \partial^u f(\mathbf{z}_u : \mathbf{c}_{-u}) \mathrm{d}\mathbf{z}_u.$$

For $\varepsilon_n \in (0, 1/2)$ and $K_n = [\varepsilon_n, 1-\varepsilon_n]^{d+1}$, we use the same low variation extension $\tilde{f}_n$ of f as in Owen (2006, Equation 2.9):

$$\tilde{f}_n(\mathbf{x}) = f(\mathbf{c}) + \sum_{u \neq \emptyset} \int_{[\mathbf{c}_u, \mathbf{x}_u]} 1_{\mathbf{z}_u : \mathbf{c}_{-u} \in K_n} \partial^u f(\mathbf{z}_u : \mathbf{c}_{-u}) \mathrm{d}\mathbf{z}_u.$$

Then $\tilde{f}_n$ coincides with f on K_n. By definition (see e.g., Owen, 2005, Definition 2), the Hardy-Krause variation of $\tilde{f}_n$ over $[0,1]^{d+1}$ is:

$$V_{\mathrm{HK}}(\tilde{f}_n; [0,1]^{d+1}) = \sum_{u \neq \emptyset} V_{[\mathbf{0}^u, \mathbf{1}^u]} \tilde{f}_n(\mathbf{x}_u; \mathbf{1}_{-u}), \tag{B.2}$$

where $V_{[\mathbf{a},\mathbf{b}]}(\cdot)$ is the Vitali variation. Note that the Vitali variation satisfies the bound (see e.g., Owen, 2005, Proposition 13)

$$V_{[\mathbf{a},\mathbf{b}]}(g) \leq \int_{[\mathbf{a},\mathbf{b}]} |\partial^{\{1,\ldots,p\}} g(\mathbf{x})| \mathrm{d}\mathbf{x}, \quad \text{for } g(\cdot) \text{ defined on a hyperrectangle } [\mathbf{a}, \mathbf{b}] \subset \mathbb{R}^p. \tag{B.3}$$

Plugging the bound (B.3) into the definition (B.2), we obtain

$$\begin{aligned} V_{\mathrm{HK}}(\tilde{f}_n; [0,1]^{d+1}) &\leq \sum_{u \neq \emptyset} \int_{[\varepsilon_n, 1-\varepsilon_n]^{|u|}} |\partial^u f(\mathbf{x}_u : \mathbf{c}_{-u})| \mathrm{d}\mathbf{x}_u \\ &\leq \sum_{u \neq \emptyset} \int_{[\varepsilon_n, 1-\varepsilon_n]^{|u|}} B \prod_{j \in u \setminus \{d+1\}} \min(x_j, 1-x_j)^{-1} \mathrm{d}\mathbf{x}_u \\ &= B \sum_{u \neq \emptyset} ((1-2\varepsilon_n) \cdot 1_{d+1 \in u} + 1_{d+1 \notin u}) \prod_{j \in u \setminus \{d+1\}} 2(\log(1/2) - \log \varepsilon_n) \\ &\leq B \sum_{u \neq \emptyset} \prod_{j \in u \setminus \{d+1\}} 2(\log(1/2) - \log \varepsilon_n). \end{aligned}$$

We may simplify the above sum by considering whether $d+1 \in u$:

$$\begin{aligned}
&\sum_{u\neq\emptyset}\prod_{j\in u\setminus\{d+1\}} 2(\log(1/2)-\log\varepsilon_n)\\
&= \sum_{u\neq\emptyset, u\subset\{1,\dots,d\}}\prod_{j\in u} 2(\log(1/2)-\log\varepsilon_n) + \sum_{u\subset\{1,\dots,d+1\}, d+1\in u}\prod_{j\in u\setminus\{d+1\}} 2(\log(1/2)-\log\varepsilon_n)\\
&= \left(-1+\prod_{j=1}^{d}(1+2\log(1/2)-2\log\varepsilon_n)\right)+\prod_{j=1}^{d}(1+2\log(1/2)-2\log\varepsilon_n)\\
&\leq 2\prod_{j=1}^{d}(1+2\log(1/2)-2\log\varepsilon_n) = 2(1+2\log(1/2)-2\log\varepsilon_n)^d.
\end{aligned}$$

Therefore, we obtain the following bound for $V_{\mathrm{HK}}(\tilde{f}_n;[0,1]^{d+1})$:

$$V_{\mathrm{HK}}(\tilde{f}_n;[0,1]^{d+1}) \leq 2B(1-2\log 2-2\log\varepsilon_n)^d. \tag{B.4}$$

The star discrepancy of the $(d+1)$-dimensional Halton sequence satisfies

$$\mathcal{D}^*(\{\mathbf{h}_i\}_{i=1}^n) \leq C_H(d+1)\cdot\frac{(\log n)^{d+1}}{n}, \quad \text{for } n\geq 2, \tag{B.5}$$

where $C_H(d+1)$ is a constant depending on the dimension $d+1$ (see e.g., Niederreiter, 1992, Theorem 3.6 and Atanassov, 2004). Hence, it remains to bound the term $\int_{[0,1]^{d+1}}|f(\mathbf{x})-\tilde{f}_n(\mathbf{x})|\mathrm{d}\mathbf{x}$ in (B.1). Observe that for $\mathbf{x}\in(0,1)^{d+1}$,

$$\begin{aligned}
|f(\mathbf{x})-\tilde{f}_n(\mathbf{x})| &\leq \sum_{u\neq\emptyset}\left|\int_{[\mathbf{c}_u,\mathbf{x}_u]} 1_{\mathbf{z}_u:\mathbf{c}_{-u}\notin K_n}|\partial^u f(\mathbf{z}_u:\mathbf{c}_{-u})|\mathrm{d}\mathbf{z}_u\right|\\
&\leq \sum_{u\neq\emptyset}\left|\int_{[\mathbf{c}_u,\mathbf{x}_u]} 1_{\mathbf{z}_u:\mathbf{c}_{-u}\notin K_n} B\prod_{j\in u\setminus\{d+1\}}\min(z_j,1-z_j)^{-1}\mathrm{d}\mathbf{z}_u\right|\\
&\leq B\sum_{u\neq\emptyset}\left|\int_{[\mathbf{c}_u,\mathbf{x}_u]}\prod_{j\in u\setminus\{d+1\}}\min(z_j,1-z_j)^{-1}\mathrm{d}\mathbf{z}_u\right|.
\end{aligned}$$

By considering whether $d+1\in u$, we have

$$\begin{aligned}
&\sum_{u\neq\emptyset}\left|\int_{[\mathbf{c}_u,\mathbf{x}_u]}\prod_{j\in u\setminus\{d+1\}}\min(z_j,1-z_j)^{-1}\mathrm{d}\mathbf{z}_u\right|\\
&= \sum_{u\neq\emptyset,u\subset\{1,\dots,d\}}\left|\int_{[\mathbf{c}_u,\mathbf{x}_u]}\prod_{j\in u\setminus\{d+1\}}\min(z_j,1-z_j)^{-1}\mathrm{d}\mathbf{z}_u\right|\\
&\quad+\sum_{u\subset\{1,\dots,d+1\},d+1\in u}\left|\int_{[\mathbf{c}_u,\mathbf{x}_u]}\prod_{j\in u\setminus\{d+1\}}\min(z_j,1-z_j)^{-1}\mathrm{d}\mathbf{z}_u\right|\\
&= \sum_{u\neq\emptyset,u\subset\{1,\dots,d\}}\prod_{j\in u}|\log(1/2)-\log(\min(x_j,1-x_j))|\\
&\quad+\sum_{u\subset\{1,\dots,d+1\},d+1\in u}|1/2-x_{d+1}|\prod_{j\in u\setminus\{d+1\}}|\log(1/2)-\log(\min(x_j,1-x_j))|\\
&= -1+\prod_{j=1}^{d}\big(1+|\log(1/2)-\log(\min(x_j,1-x_j))|\big)\\
&\quad+|1/2-x_{d+1}|\cdot\prod_{j=1}^{d}\big(1+|\log(1/2)-\log(\min(x_j,1-x_j))|\big)
\end{aligned}$$

$$\le \frac{3}{2}\prod_{j=1}^{d}\big(1+|\log(1/2)-\log(\min(x_j,1-x_j))|\big).$$

Therefore, $|f(\mathbf{x})-\tilde{f}_n(\mathbf{x})|\le\frac{3B}{2}\prod_{j=1}^{d}\big(1+|\log(1/2)-\log(\min(x_j,1-x_j))|\big)$. Recall that $\tilde{f}_n$ coincides with f on $K_n=[\varepsilon_n,1-\varepsilon_n]^{d+1}$. Hence, we have

$$\int_{(0,1/2)^{d+1}}|f(\mathbf{x})-\tilde{f}_n(\mathbf{x})|\mathrm{d}\mathbf{x}\le\frac{3B}{2}\int_{(0,1/2)^{d+1}\setminus[\varepsilon_n,1/2]^{d+1}}\prod_{j=1}^{d}(1+\log(1/2)-\log(x_j))\mathrm{d}\mathbf{x}$$
$$\le\frac{3B}{2}\sum_{j=1}^{d}\int_0^{\varepsilon_n}(1+\log(1/2)-\log(x_j))\mathrm{d}x_j\int_{(0,1/2)^d}\prod_{k\in\{1,\dots,d\}\setminus\{j\}}(1+\log(1/2)-\log(x_k))\mathrm{d}\mathbf{x}_{-j}$$
$$+\frac{3B}{2}\int_0^{\varepsilon_n}\mathrm{d}x_{d+1}\prod_{k=1}^{d}\int_0^{1/2}(1+\log(1/2)-\log(x_k))\mathrm{d}x_k$$
$$=\frac{3B}{2}\sum_{j=1}^{d}((2-\log 2)\varepsilon_n-\varepsilon_n\log\varepsilon_n)\cdot\frac{1}{2}1^{d-1}+\frac{3B}{2}\varepsilon_n\cdot 1^d$$
$$=\frac{3B\varepsilon_n}{4}\big(2+(2-\log 2)d-d\log\varepsilon_n\big).$$

The same argument can be applied to all 2^{d+1} subcubes of $(0,1)^{d+1}$, which yields

$$\int_{(0,1)^{d+1}}|f(\mathbf{x})-\tilde{f}_n(\mathbf{x})|\mathrm{d}\mathbf{x}\le 3\cdot 2^{d-1}B\varepsilon_n\big(2+(2-\log 2)d-d\log\varepsilon_n\big). \tag{B.6}$$

Finally, we use the fact that Halton sequence avoids the boundary of the unit cube: by choosing $\varepsilon_n=\frac{1}{n(n+1)}\prod_{j=1}^{d}p_j^{-1}$, $\mathbf{h}_1,\dots,\mathbf{h}_n$ belong to $K_n=[\varepsilon_n,1-\varepsilon_n]^{d+1}$ (Owen, 2006, Theorem 3.1). Now, we can apply the above bounds on $\int_{[0,1]^{d+1}}|f(\mathbf{x})-\tilde{f}_n(\mathbf{x})|\mathrm{d}\mathbf{x}$ (B.6), $\mathcal{D}^*(\{\mathbf{h}_i\}_{i=1}^n)$ (B.5), and $V_{\mathrm{HK}}(\tilde{f}_n)$ (B.4) back to (B.1) to obtain:

$$\left|\int_{[0,1]^{d+1}}f(\mathbf{x})\mathrm{d}\mathbf{x}-\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{h}_i)\right|\le\int_{[0,1]^{d+1}}|f(\mathbf{x})-\tilde{f}_n(\mathbf{x})|\mathrm{d}\mathbf{x}+\mathcal{D}^*(\{\mathbf{h}_i\}_{i=1}^n)V_{\mathrm{HK}}(\tilde{f}_n)$$
$$\le 3\cdot 2^{d-1}B\varepsilon_n\big(2+(2-\log 2)d-d\log\varepsilon_n\big)+C_H(d+1)\cdot\frac{(\log n)^{d+1}}{n}\cdot 2B(1-2\log 2-2\log\varepsilon_n)^d$$
$$\le BC(d)\frac{(\log n)^{2d+1}}{n},$$

for some $C(d)>0$ and all $n\ge 2$.

Note that the coefficient of the dominating term $\frac{(\log n)^{2d+1}}{n}$ is $C_H(d+1)2^{2d+1}B$. □

B.2. Proof of Proposition 2.6

For any $g\in L^2(P_{\mathbf{X}})$, we have

$$\|(L_M-L)g\|^2_{L^2(P_{\mathbf{X}})}=\left\|\int(K_M(\cdot,\mathbf{x})-K(\cdot,\mathbf{x}))g(\mathbf{x})\mathrm{d}P_{\mathbf{X}}(\mathbf{x})\right\|^2_{L^2(P_{\mathbf{X}})}$$
$$=\int(K_M(\mathbf{z},\mathbf{x})-K(\mathbf{z},\mathbf{x}))(K_M(\mathbf{z},\mathbf{x}')-K(\mathbf{z},\mathbf{x}'))g(\mathbf{x})g(\mathbf{x}')\mathrm{d}P_{\mathbf{X}}(x)\mathrm{d}P_{\mathbf{X}}(\mathbf{x}')\mathrm{d}P_{\mathbf{X}}(\mathbf{z})$$
$$\le\frac{C^2\log^{2a}M}{M^2}\int|g(\mathbf{x})g(\mathbf{x}')|\mathrm{d}P_{\mathbf{X}}(\mathbf{x})\mathrm{d}P_{\mathbf{X}}(\mathbf{x}')$$
$$\le\frac{C^2\log^{2a}M}{M^2}\|g\|^2_{L^2(P_{\mathbf{X}})}.$$

Therefore, by the definition of operator norm, $\|L_M-L\|=\sup_{g\in L^2(P_{\mathbf{X}})}\frac{\|(L_M-L)g\|_{L^2(P_{\mathbf{X}})}}{\|g\|_{L^2(P_{\mathbf{X}})}}\le\frac{C\log^a M}{M}$. □

B.3. Proof of Proposition 2.7

Note that the Δ-spectral approximation (7) is equivalent to

$$\mathbf{K} - \Delta(\mathbf{K} + \lambda \mathbf{I}_n) \preceq \mathbf{K}_M \preceq \mathbf{K} + \Delta(\mathbf{K} + \lambda \mathbf{I}_n).$$

Let $\mathbf{K} + \lambda \mathbf{I}_n = \mathbf{V}^\top \boldsymbol{\Sigma}^2 \mathbf{V}$ be the eigen-decomposition of $\mathbf{K} + \lambda \mathbf{I}_n$. By multiplying by $\boldsymbol{\Sigma}^{-1}\mathbf{V}$ on the left and $\mathbf{V}^\top \boldsymbol{\Sigma}^{-1}$ on the right, it suffices to show that

$$\|\boldsymbol{\Sigma}^{-1}\mathbf{V}(\mathbf{K}_M - \mathbf{K})\mathbf{V}^\top \boldsymbol{\Sigma}^{-1}\|_2 \le \Delta,$$

where $\|\cdot\|_2$ denotes the matrix L_2 norm, or the spectral norm. Since the eigenvalues of $\mathbf{K} + \lambda \mathbf{I}_n$ is lower bounded by λ, we have that $\|\boldsymbol{\Sigma}^{-1}\|_2 \le \frac{1}{\sqrt{\lambda}}$. Together with the facts that (i) $\|UAV\|_2 = \|A\|_2$ for any orthogonal matrices U, V, and (ii) the spectral norm is upper bounded by the Frobenius norm, we have that

$$\begin{aligned}\|\boldsymbol{\Sigma}^{-1}\mathbf{V}(\mathbf{K}_M - \mathbf{K})\mathbf{V}^\top \boldsymbol{\Sigma}^{-1}\|_2 &\le \frac{1}{\lambda}\|\mathbf{V}(\mathbf{K}_M - \mathbf{K})\mathbf{V}^\top\|_2 \\ &= \frac{1}{\lambda}\|\mathbf{K}_M - \mathbf{K}\|_2 \le \frac{1}{\lambda}\|\mathbf{K}_M - \mathbf{K}\|_F \\ &\le \frac{1}{\lambda}\sqrt{\frac{n^2 C^2 (\log M)^{2a}}{M^2}} \le \Delta.\end{aligned}$$

□

C. Proofs of the Results in Section 3

The goal of this section is to prove Theorem 3.2. We will first define some useful operators and clarify some notions about the excess risk with the inclusion operator notations in subsection C.1, and then provide the proof in subsection C.2.

C.1. Preliminaries

Some Useful Operators. We first introduce some useful operators between Hilbert spaces (Rudi & Rosasco, 2017).

Definition C.1. Recall $\boldsymbol{\phi}_M(\mathbf{x}) = M^{-1/2}(\psi(\mathbf{x}, \omega_1), \ldots, \psi(\mathbf{x}, \omega_M))^\top$ and the approximated kernel $K_M(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}_M(\mathbf{x})^\top \boldsymbol{\phi}_M(\mathbf{x}')$. Define:

- $S_M : \mathbb{R}^M \to L^2(P_\mathbf{X}), \quad (S_M \beta)(\cdot) = \boldsymbol{\phi}_M(\cdot)^\top \beta.$
- $S_M^* : L^2(P_\mathbf{X}) \to \mathbb{R}^M, \quad (S_M^* g)_i = \frac{1}{\sqrt{M}} \int_\mathcal{X} \psi(\mathbf{x}, \omega_i) g(\mathbf{x}) \mathrm{d}P_\mathbf{X}(\mathbf{x}).$
- $L_M : L^2(P_\mathbf{X}) \to L^2(P_\mathbf{X}), \quad (L_M g)(\cdot) = \int_\mathcal{X} K_M(\cdot, \mathbf{z}) g(\mathbf{z}) \mathrm{d}P_\mathbf{X}(\mathbf{z}).$
- $C_M : \mathbb{R}^M \to \mathbb{R}^M, \quad C_M = \int_\mathcal{X} \boldsymbol{\phi}_M(\mathbf{x}) \boldsymbol{\phi}_M(\mathbf{x})^\top \mathrm{d}P_\mathbf{X}(\mathbf{x}).$
- $\hat{C}_M : \mathbb{R}^M \to \mathbb{R}^M, \quad \hat{C}_M = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}_M(\mathbf{x}_i) \boldsymbol{\phi}_M(\mathbf{x}_i)^\top.$
- $\hat{S}_M : \mathbb{R}^M \to \mathbb{R}^n, \quad \hat{S}_M = \frac{1}{\sqrt{n}}(\boldsymbol{\phi}_M(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_M(\mathbf{x}_n))^\top.$
- $\hat{S}_M^* : \mathbb{R}^n \to \mathbb{R}^M, \quad \hat{S}_M^* = \frac{1}{\sqrt{n}}(\boldsymbol{\phi}_M(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_M(\mathbf{x}_n)).$

It follows from direct verification that S_M^* is the *adjoint* of S_M, $\hat{S}_M^*$ is the *adjoint* of $\hat{S}_M$, $L_M = S_M S_M^*$, $C_M = S_M^* S_M$, and $\hat{C}_M = \hat{S}_M^* \hat{S}_M$. Let $C_K = I^* I$ where I is the inclusion operator defined in Appendix A.4. Recall (from Appendix A.4) that $L = II^*$. From Caponnetto & De Vito (2007) and Rudi & Rosasco (2017), we have that $L, C_K, L_M, C_M, \hat{C}_M$ are trace class operators, and $I, S_M, \hat{S}_M$ are compact operators.

Excess Risk. Recall the excess risk: $\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)$ defined in (11), where $\mathcal{E}(f) = \mathbb{E}[Y - f(\mathbf{X})]^2$. In order to more clearly distinguish between the inner products $\langle \cdot, \cdot \rangle_\mathcal{H}$ and $\langle \cdot, \cdot \rangle_{L^2(P_\mathbf{X})}$, we identify $f \in \mathcal{H}$ as $If \in L^2(P_\mathbf{X})$ when the associated inner product is $\langle \cdot, \cdot \rangle_{L^2(P_\mathbf{X})}$, where I is the inclusion operator (introduced in Appendix A.4). With these notations, the excess risk is (more rigorously) written as $\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(If)$.

Recall that $f_\star(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$. A variance-bias decomposition of the risk gives:

$$\begin{aligned}\mathcal{E}(f) = \mathbb{E}[(Y - f(\mathbf{X}))^2] &= \mathbb{E}[(Y - f_\star(\mathbf{X}) + f_\star(\mathbf{X}) - f(\mathbf{X}))^2] \\ &= \mathbb{E}[(Y - f_\star(\mathbf{X}))^2] + \mathbb{E}[(f_\star(\mathbf{X}) - f(\mathbf{X}))^2].\end{aligned} \tag{C.1}$$

Therefore, the existence of an $f_\mathcal{H} \in \mathcal{H}$ that minimizes $\mathcal{E}(If)$ (KRR Condition 3) is equivalent to the existence of an $If_\mathcal{H}$ in $I(\mathcal{H})$ being closest to $f_\star$ in $L^2(P_\mathbf{X})$ distance. In other words, let $P : L^2(P_\mathbf{X}) \to L^2(P_\mathbf{X})$ be the projection operator from $L^2(P_\mathbf{X})$ to the closure of $I(\mathcal{H})$ in $L^2(P_\mathbf{X})$; if KRR Condition 3 holds, we have $If_\mathcal{H} = Pf_\star = L^r g$.

Hence, from (C.1) we can re-write the excess risk as follows (the last equality follows from Pythagorean theorem and the fact that $\hat{f}_{\lambda,M} \in I(\mathcal{H})$):

$$\begin{aligned}\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f\in\mathcal{H}} \mathcal{E}(If) &= \mathcal{E}(\hat{f}_{\lambda,M}) - \mathcal{E}(If_\mathcal{H}) \\ &= \|f_\star - \hat{f}_{\lambda,M}\|^2_{L^2(P_\mathbf{X})} - \|f_\star - Pf_\star\|^2_{L^2(P_\mathbf{X})} = \|\hat{f}_{\lambda,M} - Pf_\star\|^2_{L^2(P_\mathbf{X})}.\end{aligned} \tag{C.2}$$

C.2. Proof of Theorem 3.2

For any operator T on some space $\mathcal{S}$, define $T_\lambda := T + \lambda \mathbf{I}_\mathcal{S}$, where $\mathbf{I}_\mathcal{S}$ denotes the identity operator on $\mathcal{S}$. For $\mathcal{S} = \mathbb{R}^M$, $\mathbf{I}_\mathcal{S}$ is the $M \times M$ identity matrix $\mathbf{I}_M$. The regression function $\hat{f}_{\lambda,M}(\mathbf{x})$ defined in (10) can then be written as

$$\hat{f}_{\lambda,M} = S_M(\hat{C}_M + \lambda \mathbf{I}_M)^{-1}\hat{S}_M^* \mathbf{y}/\sqrt{n} = S_M \hat{C}_{M,\lambda}^{-1} \hat{S}_M^* \mathbf{y}/\sqrt{n}, \quad \text{where } \mathbf{y} = (y_1, \ldots, y_n)^\top. \tag{C.3}$$

Using the operators introduced in the previous subsection, we define

$$\tilde{f} = I(C_K + \lambda \mathbf{I}_\mathcal{H})^{-1} I^* P f_\star = I C_{K,\lambda}^{-1} I^* P f_\star \quad \text{and} \tag{C.4}$$

$$\tilde{f}_M = S_M (C_M + \lambda \mathbf{I}_M)^{-1} S_M^* P f_\star = S_M C_{M,\lambda}^{-1} S_M^* P f_\star. \tag{C.5}$$

Here, $\hat{f}_{\lambda,M}$ can be viewed as the empirical version of $\tilde{f}_M$, and $\tilde{f}_M$ can be viewed as the random feature approximation of $\tilde{f}$. With triangle inequality, we decompose the excess risk (C.2) into $\|\hat{f}_{\lambda,M} - Pf_\star\|_{L^2(P_\mathbf{X})} \le E_1 + E_2 + E_3$, where

$$\begin{aligned}E_1 &:= \|\tilde{f} - Pf_\star\|_{L^2(P_\mathbf{X})}, \\ E_2 &:= \|\tilde{f}_M - \tilde{f}\|_{L^2(P_\mathbf{X})}, \\ E_3 &:= \|\hat{f}_{\lambda,M} - \tilde{f}_M\|_{L^2(P_\mathbf{X})}.\end{aligned} \tag{C.6}$$

In the above decomposition (C.6), E_1, E_2, E_3 may be understood as follows:

- E_1 is the *bias* introduced by the ridge regression penalty, which also appears in the analysis of the exact kernel ridge regression (De Vito et al., 2005), and is not affected by the application of random feature approximation.
- E_2 is the random feature *computational error* arising from approximating the kernel K with K_M. It is worth noting that it is in this term that our error rate has substantial improvement compared to the bound for the Monte Carlo random features (Rudi & Rosasco, 2017).
- E_3 is the *variance* coming from n i.i.d. draws from the population. While E_3 exhibits an error rate comparable to that in the Monte Carlo random feature case, our analysis differs in several aspects. In particular, results that hold almost surely may need to be replaced by results that hold surely, and some bounds also have tighter and non-random versions under the QMC setting.

In the following, we will bound these three terms respectively. Throughout the paper, norms without subscripts are considered operator norms unless otherwise specified.

Bound for E_1: The following expressions

$$\tilde{f} = LL_\lambda^{-1} P f_\star, \quad \tilde{f}_M = L_M L_{M,\lambda}^{-1} P f_\star$$

follow from the definitions of $\tilde{f}$, $\tilde{f}_M$ in (C.4), (C.5), together with the identity that: for a bounded linear operator $A : \mathcal{H}_1 \to \mathcal{H}_2$,

$$A^*(AA^* + c\mathbf{I}_{\mathcal{H}_2})^{-1} = (A^*A + c\mathbf{I}_{\mathcal{H}_1})^{-1}A^*. \tag{C.7}$$

From KRR Condition 3, we have that $Pf_\star = L^r g$ with $\|g\|_{L^2(P_{\mathbf{X}})} \leq R$. By the identity $L(L + \lambda\mathbf{I}_{L^2(P_{\mathbf{X}})})^{-1} = \mathbf{I}_{L^2(P_{\mathbf{X}})} - \lambda(L + \lambda\mathbf{I}_{L^2(P_{\mathbf{X}})})^{-1}$, we have

$$\begin{aligned}
E_1 = \|LL_\lambda^{-1}Pf_\star - Pf_\star\|_{L^2(P_{\mathbf{X}})} &= \|(LL_\lambda^{-1} - \mathbf{I}_{L^2(P_{\mathbf{X}})})Pf_\star\|_{L^2(P_{\mathbf{X}})} = \| - \lambda L_\lambda^{-1}Pf_\star\|_{L^2(P_{\mathbf{X}})} \\
&= \| - \lambda L_\lambda^{-1}L^r g\|_{L^2(P_{\mathbf{X}})} \\
&\leq \lambda \cdot \|L_\lambda^{-1+r}\| \cdot \|L_\lambda^{-r}L^r\| \cdot \|g\|_{L^2(P_{\mathbf{X}})} \\
&\leq \lambda \cdot \lambda^{-1+r} \cdot 1 \cdot R \\
&= R\lambda^r.
\end{aligned}$$

Bound for E_2: Denote a general identity operator whose associated space is unspecified by $\mathbf{I}$. By the algebraic identities $A(A + \lambda\mathbf{I})^{-1} = \mathbf{I} - \lambda(A + \lambda\mathbf{I})^{-1}$ and $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we have that

$$(LL_\lambda^{-1} - L_M L_{M,\lambda}^{-1})Pf_\star = \lambda(L_{M,\lambda}^{-1} - L_\lambda^{-1})Pf_\star = \lambda L_{M,\lambda}^{-1}(L - L_M)L_\lambda^{-1}Pf_\star.$$

From KRR Condition 3, we have that $Pf_\star = L^r g$ with $\|g\|_{L^2(P_{\mathbf{X}})} \leq R$. From KRR Condition 1 and Proposition 2.6, we have that $\|L - L_M\| \leq C \cdot \frac{\log^a M}{M}$. Therefore,

$$\begin{aligned}
E_2 = \|(LL_\lambda^{-1} - L_M L_{M,\lambda}^{-1})Pf_\star\|_{L^2(P_{\mathbf{X}})} &= \|\lambda L_{M,\lambda}^{-1}(L - L_M)L_\lambda^{-1}L^r g\|_{L^2(P_{\mathbf{X}})} \\
&\leq \lambda\|L_{M,\lambda}^{-1}\| \cdot \|L - L_M\| \cdot \|L_\lambda^{-1+r}\| \cdot \|L_\lambda^{-r}L^r\| \cdot \|g\|_{L^2(P_{\mathbf{X}})} \\
&\leq \lambda \cdot \lambda^{-1} \cdot \frac{C\log^a M}{M} \cdot \lambda^{-1+r} \cdot 1 \cdot R \\
&= R\lambda^{-1+r}\frac{C\log^a M}{M}.
\end{aligned}$$

Bound for E_3: Recall that $\hat{f}_{\lambda,M} = S_M\hat{C}_{M,\lambda}^{-1}\hat{S}_M^*\mathbf{y}/\sqrt{n}$ and $\tilde{f}_M = S_M C_{M,\lambda}^{-1}S_M^*Pf_\star$. With triangle inequality, we decompose E_3 into $E_3 \leq E_{31} + E_{32} + E_{33}$, where

$$E_{31} = \|S_M\hat{C}_{M,\lambda}^{-1}(\hat{S}_M^*\mathbf{y}/\sqrt{n} - S_M^*f_\star)\|_{L^2(P_{\mathbf{X}})}, \tag{C.8}$$

$$E_{32} = \|S_M\hat{C}_{M,\lambda}^{-1}S_M^*(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)f_\star\|_{L^2(P_{\mathbf{X}})}, \tag{C.9}$$

$$E_{33} = \|S_M(\hat{C}_{M,\lambda}^{-1} - C_{M,\lambda}^{-1})S_M^*Pf_\star\|_{L^2(P_{\mathbf{X}})}. \tag{C.10}$$

We will show that $E_{32} = 0$, and when $\frac{19\kappa^2}{n}\log\frac{3n}{2\delta} \leq \lambda \leq \min\{\|L\| - \frac{C\log^a M}{M}, 1/e\}$ and $n \geq \max\{405\kappa^2, 67\kappa^2\log\frac{3\kappa^2}{2\delta}\}$ (the constants a, κ and C are from KRR Condition 1), with probability at least $1 - \delta$,

$$E_{31} \leq \frac{3}{2} \cdot \left(\frac{4D\kappa\log\frac{6}{\delta}}{\sqrt{\lambda}n} + \sqrt{\frac{8\sigma^2\kappa^2\log\frac{6}{\delta}}{\lambda n}}\right), \tag{C.11}$$

$$E_{33} \leq \frac{3}{2} \cdot \sqrt{C\left(1+a\right)^a + 1} \cdot \left(4R\kappa^{2r+1}\log\frac{6}{\delta} + 2R\kappa^{2r}\sqrt{\log\frac{6}{\delta}}\right)\frac{1}{\sqrt{n\lambda}}. \tag{C.12}$$

The proofs of the bounds for E_{31}, E_{32} and E_{33} are given in Appendix C.3.

Combining the bounds: Combining the bounds for E_1, E_2 and E_3 above, we have that when $\frac{19\kappa^2}{n}\log\frac{3n}{2\delta} \leq \lambda \leq$

$\min\{\|L\| - \frac{C\log^a M}{M}, 1/e\}$ and $n \ge \max\{405\kappa^2, 67\kappa^2\log\frac{3\kappa^2}{2\delta}\}$, with probability at least $1-\delta$:

$$\begin{aligned}\left|\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f\in\mathcal{H}}\mathcal{E}(f)\right|^{1/2} &= \|\hat{f}_{\lambda,M} - Pf_\star\|_{L^2(P_{\mathbf{X}})} \le E_1 + E_2 + E_3 \\ &\le R\lambda^r + R\lambda^{-1+r}\frac{C\log^a M}{M} \\ &\quad + \frac{3}{2}\cdot\left(\frac{4D\kappa\log\frac{6}{\delta}}{\sqrt{\lambda}n} + \sqrt{\frac{8\sigma^2\kappa^2\log\frac{6}{\delta}}{\lambda n}}\right) \\ &\quad + \frac{3}{2}\cdot\sqrt{C\left(1+a\right)^a + 1}\cdot\left(4R\kappa^{2r+1}\log\frac{6}{\delta} + 2R\kappa^{2r}\sqrt{\log\frac{6}{\delta}}\right)\frac{1}{\sqrt{n\lambda}}.\end{aligned}$$

As $M = \frac{\log^a(1/\lambda)}{\lambda}$ and $\lambda \in (0, 1/e]$, we can derive that $\frac{C\log^a M}{\lambda M} \le C\left(1+a\right)^a$; see (C.16). Thus we may further bound the right-hand side by:

$$\begin{aligned}\left|\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f\in\mathcal{H}}\mathcal{E}(f)\right|^{1/2} \le &\left(6D\kappa\log\frac{6}{\delta} + 3\kappa\sigma\sqrt{2\log\frac{6}{\delta}} + 3R\kappa^{2r}\sqrt{C\left(1+a\right)^a+1}\left(2\kappa\log\frac{6}{\delta} + \sqrt{\log\frac{6}{\delta}}\right)\right)\frac{1}{\sqrt{n\lambda}} \\ &+ R\left(C\left(1+a\right)^a + 1\right)\lambda^r.\end{aligned}$$

Let $\lambda = \tilde{C}n^{-\frac{1}{2r+1}}$. When $n \ge C_2\max\left\{(\log\frac{1}{\delta})^{1+\frac{1}{2r}}, 1\right\}$ for a large enough constant C_2 depending on $\kappa, C, a, \|L\|, \tilde{C}$ and r (but not on δ), all the above conditions on n and λ are satisfied. Since $1 \le \sqrt{\log\frac{6}{\delta}} \le \log\frac{6}{\delta}$ for $\delta \le 1$, the above bound can be further upper bounded: $\mathcal{E}(\hat{f}_{\lambda,M}) - \inf_{f\in\mathcal{H}}\mathcal{E}(f) \le C_1 n^{-\frac{2r}{2r+1}}\log^2\frac{6}{\delta}$, where

$$C_1 = \left[\left(6D\kappa + 3\sqrt{2}\kappa\sigma + 3R\kappa^{2r}(2\kappa+1)\sqrt{C\left(1+a\right)^a+1}\right)\frac{1}{\sqrt{\tilde{C}}} + R\left(C\left(1+a\right)^a+1\right)\tilde{C}^r\right]^2.$$

□

C.3. Proof of Theorem 3.2: Proof of Bounds for E_{31}, E_{32} and E_{33}

We prove the bound (C.11) in Appendix C.3.1, $E_{32} = 0$ in Appendix C.3.2, and the bound (C.12) in Appendix C.3.3. Some supporting technical lemmas are presented in Appendix C.3.4.

C.3.1. Proof of Bound for E_{31}

Recall that $\|\cdot\|$ is the operator norm; in particular, for a vector $v \in \mathbb{R}^M$, $\|v\|$ is its vector ℓ^2-norm. From the expression of E_{31} in (C.8), we have $E_{31} \le b_1 \cdot A$, where $b_1 = \|S_M\hat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2}\|$, and $A = \|C_{M,\lambda}^{-1/2}(\hat{S}_M^*\mathbf{y}/\sqrt{n} - S_M^* f_\star)\|$. We bound these two terms respectively.

Bounding b_1: We have

$$\begin{aligned}b_1^2 &= \|S_M\hat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2}\|^2 \\ &= \|C_{M,\lambda}^{1/2}\hat{C}_{M,\lambda}^{-1}S_M^*S_M\hat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2}\| \\ &= \|C_{M,\lambda}^{1/2}\hat{C}_{M,\lambda}^{-1}C_M\hat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2}\| \\ &\le \|C_{M,\lambda}^{1/2}\hat{C}_{M,\lambda}^{-1/2}\|\cdot\|\hat{C}_{M,\lambda}^{-1/2}C_M^{1/2}\|\cdot\|C_M^{1/2}\hat{C}_{M,\lambda}^{-1/2}\|\cdot\|\hat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\| \\ &= \|\hat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\|^2\cdot\|\hat{C}_{M,\lambda}^{-1/2}C_M^{1/2}\|^2.\end{aligned}$$

Note that $\|\hat{C}_{M,\lambda}^{-1/2}C_M^{1/2}\| \le \|\hat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\|\cdot\|C_{M,\lambda}^{-1/2}C_M^{1/2}\| \le \|\hat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\|$. Therefore, we have that $b_1^2 \le \|\hat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\|^4$ and thus $b_1 \le \|\hat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\|^2$.

A direct application of Proposition C.5 (with $A = \hat{C}_M$, $B = C_M$) gives us $b_1 \leq \frac{1}{1-\lambda_{\max}(C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)C_{M,\lambda}^{-1/2})}$.

In order to bound $\frac{1}{1-\lambda_{\max}(C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)C_{M,\lambda}^{-1/2})}$, we verify the conditions needed to apply Proposition C.4. Recall that $\boldsymbol{\phi}_M(\mathbf{x}) = M^{-1/2}(\psi(\mathbf{x},\omega_1),\ldots,\psi(\mathbf{x},\omega_M))^\top$, $C_M = \mathbb{E}[\boldsymbol{\phi}_M(\mathbf{X}) \otimes \boldsymbol{\phi}_M(\mathbf{X})]$ (from Definition C.1), and $\hat{C}_M = \frac{1}{n}\sum_{i=1}^n [\boldsymbol{\phi}_M(\mathbf{x}_i) \otimes \boldsymbol{\phi}_M(\mathbf{x}_i)]$. For $v := \boldsymbol{\phi}_M(\mathbf{X})$, we have $\langle v, C_{M,\lambda}^{-1} v\rangle \leq \|v\|^2 \|C_{M,\lambda}^{-1}\| \leq \frac{\kappa^2}{\lambda}$. Hence, $\mathcal{F}_\infty(\lambda)$ in Proposition C.4 can be taken as $\frac{\kappa^2}{\lambda}$. From Proposition C.4, for any $\tau > 0$, when $\frac{19\kappa^2}{n}\log\frac{n}{2\tau} \leq \lambda \leq \|C_M\|$ and $n \geq \max\{405\kappa^2, 67\kappa^2 \log\frac{\kappa^2}{2\tau}\}$, we have

$$\lambda_{\max}(C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)C_{M,\lambda}^{-1/2}) \leq \frac{1}{3}$$

with probability at least $1-\tau$. Note that

$$\|C_M\| = \|S_M^* S_M\| = \|S_M S_M^*\| = \|L_M\| \geq \|L\| - \|L - L_M\| \geq \|L\| - \frac{C\log^a M}{M}.$$

Summarizing the above conditions, when $\frac{19\kappa^2}{n}\log\frac{n}{2\tau} \leq \lambda \leq \|L\| - \frac{C\log^a M}{M}$ and $n \geq \max\{405\kappa^2, 67\kappa^2\log\frac{\kappa^2}{2\tau}\}$, we have $b_1 \leq \frac{1}{1-1/3} = 3/2$ with probability at least $1-\tau$.

Bounding A: By the definition of $\hat{S}_M^*$ (see Definition C.1), we have that $\hat{S}_M^*\mathbf{y}/\sqrt{n} = n^{-1}\sum_{i=1}^n \boldsymbol{\phi}_M(\mathbf{x}_i)y_i$. Therefore,

$$C_{M,\lambda}^{-1/2}(\hat{S}_M^*\mathbf{y}/\sqrt{n} - S_M^* f_\star) = \frac{1}{n}\sum_{i=1}^n (C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)y_i - C_{M,\lambda}^{-1/2}S_M^* f_\star).$$

To bound the above sum by concentration inequalities, we apply Proposition C.2 with $z_i := C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)y_i$ and $\mu := C_{M,\lambda}^{-1/2}S_M^* f_\star \in \mathbb{R}^M$, whose conditions are verified as follows. First, we verify that $\mathbb{E}z_i = \mu$:

$$\mathbb{E}z_i = C_{M,\lambda}^{-1/2}\mathbb{E}[\boldsymbol{\phi}_M(\mathbf{x}_i)y_i] = C_{M,\lambda}^{-1/2}\mathbb{E}[\boldsymbol{\phi}_M(\mathbf{x}_i)\mathbb{E}[y_i \mid \mathbf{x}_i]] = C_{M,\lambda}^{-1/2}\mathbb{E}[\boldsymbol{\phi}_M(\mathbf{x}_i)f_\star(\mathbf{x}_i)] = C_{M,\lambda}^{-1/2}S_M^* f_\star = \mu.$$

Next, we bound the k-th moment ($k \geq 2$) of z_i using KRR Condition 2:

$$\begin{aligned}\mathbb{E}\|z_i\|^k &= \mathbb{E}[\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)y_i\|^k] = \mathbb{E}[\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)\|^k \cdot |y_i|^k] = \mathbb{E}[\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)\|^k \cdot \mathbb{E}[|y_i|^k \mid \mathbf{x}_i]] \\ &\leq \mathbb{E}[\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)\|^k] \cdot \frac{1}{2}k!\sigma^2 D^{k-2}.\end{aligned}$$

Recall that $\boldsymbol{\phi}_M(\mathbf{x}) = M^{-1/2}(\psi(\mathbf{x},\omega_1),\ldots,\psi(\mathbf{x},\omega_M))^\top$ and ψ is bounded by κ (KRR Condition 1). Hence,

$$\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)\|^2 \leq \frac{1}{\lambda}\|\boldsymbol{\phi}_M(\mathbf{x}_i)\|^2 \leq \frac{\kappa^2}{\lambda}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\|z_i\|^k &\leq \mathbb{E}[\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}_i)\|^k] \cdot \frac{1}{2}k!\sigma^2 D^{k-2} \\ &\leq \left(\frac{\kappa^2}{\lambda}\right)^{\frac{k}{2}} \cdot \frac{1}{2}k!\sigma^2 D^{k-2} \\ &= \frac{1}{2}k!\left(\frac{\sigma\kappa}{\sqrt{\lambda}}\right)^2\left(\frac{D\kappa}{\sqrt{\lambda}}\right)^{k-2}.\end{aligned} \tag{C.13}$$

By Jensen's inequality, triangle inequality, and generalized Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}\|z_i - \mu\|^k &= \mathbb{E}\|z_i - \mathbb{E}z_{i+1}\|^k \leq \mathbb{E}\|z_i - z_{i+1}\|^k \\ &\leq \mathbb{E}[(\|z_i\| + \|z_{i+1}\|)^k] \leq \mathbb{E}[2^{k-1}(\|z_i\|^k + \|z_{i+1}\|^k)] = 2^k\mathbb{E}\|z_i\|^k.\end{aligned}$$

With our bound for $\mathbb{E}\|z_i\|^k$ in (C.13),

$$\mathbb{E}\|z_i - \mu\|^k \leq 2^k\mathbb{E}\|z_i\|^k \leq \frac{1}{2}k!\left(\frac{2\sigma\kappa}{\sqrt{\lambda}}\right)^2\left(\frac{2D\kappa}{\sqrt{\lambda}}\right)^{k-2}.$$

Now, we are ready to apply Proposition C.2. With probability at least $1-\tau$, we have that

$$A = \|C_{M,\lambda}^{-1/2}(\hat{S}_M^* \mathbf{y}/\sqrt{n} - S_M^* f_\star)\| = \left\| \frac{1}{n}\sum_{i=1}^n z_i - \mu \right\| \le \frac{4D\kappa \log \frac{2}{\tau}}{\sqrt{\lambda} n} + \sqrt{\frac{8\sigma^2\kappa^2 \log \frac{2}{\tau}}{\lambda n}}.$$

Combining the bound $b_1 \le 3/2$ with the above bound on A, by setting τ to be $\delta/3$, we have that (C.11) holds with probability at least $1-2\delta/3$ when $\frac{19\kappa^2}{n}\log\frac{3n}{2\delta} \le \lambda \le \|L\| - \frac{C\log^a M}{M}$ and $n \ge \max\{405\kappa^2, 67\kappa^2 \log\frac{3\kappa^2}{2\delta}\}$.

C.3.2. Proof of $E_{32} = 0$

Let ψ_ω denote the function $\psi(\cdot,\omega)$. Observe that for $f, g \in L^2(P_{\mathbf{X}})$, we have

$$\begin{aligned}
\langle f, Lg\rangle_{L^2(P_{\mathbf{X}})} &= \int f(\mathbf{x})K(\mathbf{x},\mathbf{z})g(\mathbf{z})\mathrm{d}P_{\mathbf{X}}(\mathbf{x})\mathrm{d}P_{\mathbf{X}}(\mathbf{z}) \\
&= \int f(\mathbf{x})\psi(\mathbf{x},\omega)\psi(\mathbf{z},\omega)g(\mathbf{z})\mathrm{d}P_{\mathbf{X}}(\mathbf{x})\mathrm{d}P_{\mathbf{X}}(\mathbf{z})\mathrm{d}\pi(\omega) \\
&= \int \langle f, \psi_\omega\rangle_{L^2(P_{\mathbf{X}})}\langle g, \psi_\omega\rangle_{L^2(P_{\mathbf{X}})}\mathrm{d}\pi(\omega).
\end{aligned}$$

We will use the isometric isomorphism between the Hilbert tensor product space $L^2(P_{\mathbf{X}}) \otimes L^2(P_{\mathbf{X}})$ and the space of Hilbert-Schmidt operators from $L^2(P_{\mathbf{X}})$ to itself (see e.g., Aubin 2011, Section 12). In particular, this isomorphism $\Phi : \mathcal{H}_1 \otimes \mathcal{H}_2 \to \mathrm{HS}(\mathcal{H}_2, \mathcal{H}_1)$ is given by $\langle f, \Phi(a)g\rangle_{\mathcal{H}_1} = \langle a, f\otimes g\rangle_{\mathcal{H}_1\otimes\mathcal{H}_2}$ for $a \in \mathcal{H}_1\otimes\mathcal{H}_2$, $f\in\mathcal{H}_1$, $g\in\mathcal{H}_2$; and if $a = \tilde{f}\otimes\tilde{g}$ for $\tilde{f}\in\mathcal{H}_1$ and $\tilde{g}\in\mathcal{H}_2$, then $\langle a, f\otimes g\rangle_{\mathcal{H}_1\otimes\mathcal{H}_2} = \langle\tilde{f}, f\rangle_{\mathcal{H}_1}\cdot\langle\tilde{g}, g\rangle_{\mathcal{H}_2}$.

Therefore, with this isomorphism, we can write

$$\begin{aligned}
\langle f, Lg\rangle_{L^2(P_{\mathbf{X}})} &= \int \langle f, \psi_\omega\rangle_{L^2(P_{\mathbf{X}})}\langle g, \psi_\omega\rangle_{L^2(P_{\mathbf{X}})}\mathrm{d}\pi(\omega) \\
&= \int \langle f, (\psi_\omega\otimes\psi_\omega)g\rangle_{L^2(P_{\mathbf{X}})}\mathrm{d}\pi(\omega) \\
&= \left\langle f, \int \psi_\omega\otimes\psi_\omega \mathrm{d}\pi(\omega) g\right\rangle_{L^2(P_{\mathbf{X}})}.
\end{aligned}$$

Note that the last equality follows from the fact that if X is Bochner integrable, then $\mathbb{E}\varphi(X) = \varphi(\mathbb{E}X)$ for any bounded linear functional φ (see e.g., Cohn, 2013, Appendix E). The HS norm of $\int \psi_\omega\otimes\psi_\omega\mathrm{d}\pi(\omega)$ is finite since $\psi(\mathbf{x},\omega)$ is bounded (KRR Condition 1). Hence, we have that

$$L = \int \psi_\omega\otimes\psi_\omega \mathrm{d}\pi(\omega). \tag{C.14}$$

Since P is the projection operator onto the closure of $I(\mathcal{H})$, which is $\overline{\mathrm{ran}\, L^{1/2}}$ by Theorem A.7, we have that $(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)L^{1/2} = 0$. Let $\{e_i\}_{i\ge1}$ be an orthonormal basis of $L^2(P_{\mathbf{X}})$. Note that $\mathbf{I}_{L^2(P_{\mathbf{X}})} - P$ is self-adjoint. Hence, we have

$$\begin{aligned}
0 = \mathrm{tr}\left((\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)L(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)\right) &= \sum_i \langle(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)L(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)e_i, e_i\rangle_{L^2(P_{\mathbf{X}})} \\
&= \sum_i \langle L(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)e_i, (\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)e_i\rangle_{L^2(P_{\mathbf{X}})} \\
&= \sum_i \left\langle \int \psi_\omega\otimes\psi_\omega\mathrm{d}\pi(\omega)(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)e_i, (\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)e_i\right\rangle_{L^2(P_{\mathbf{X}})} \\
&= \sum_i \int \left\langle\psi_\omega, (\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)e_i\right\rangle^2_{L^2(P_{\mathbf{X}})}\mathrm{d}\pi(\omega) = \int \|(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)\psi_\omega\|^2_{L^2(P_{\mathbf{X}})}\mathrm{d}\pi(\omega).
\end{aligned}$$

Hence, $(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)\psi_\omega = 0$ for Lebesgue almost every $\omega \in [0,1]^p$. Since $\omega \mapsto \psi_\omega \in L^2(P_{\mathbf{X}})$ is continuous (KRR

Condition 1), we have that $(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)\psi_\omega = 0$ for all $\omega \in [0,1]^p$. Consequently, for any $\beta \in \mathbb{R}^M$, we have

$$\langle \beta, S_M^*(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)f_\star\rangle_{\mathbb{R}^M} = \frac{1}{\sqrt{M}}\sum_{i=1}^M \beta_i \langle (\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)\psi_{\omega_i}, f_\star\rangle_{L^2(P_{\mathbf{X}})} = 0.$$

This shows that $E_{32} = \|S_M \hat{C}_{M,\lambda}^{-1} S_M^*(\mathbf{I}_{L^2(P_{\mathbf{X}})} - P)f_\star\|_{L^2(P_{\mathbf{X}})} = 0$.

C.3.3. Proof of Bound for E_{33}

From KRR Condition 3, we have $Pf_\star = If_{\mathcal{H}} = L^r g$, where $g \in L^2(P_{\mathbf{X}})$. From the definition of E_{33} in (C.10), together with the equality $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$, we have

$$\begin{aligned}
E_{33} &= \|S_M(\hat{C}_{M,\lambda}^{-1} - C_{M,\lambda}^{-1})S_M^* P f_\star\|_{L^2(P_{\mathbf{X}})} \\
&= \|S_M \hat{C}_{M,\lambda}^{-1}(C_M - \hat{C}_M)C_{M,\lambda}^{-1} S_M^* P f_\star\|_{L^2(P_{\mathbf{X}})} \\
&= \|S_M \hat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2} C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)C_{M,\lambda}^{-1} S_M^* L_{M,\lambda}^{1/2} L_{M,\lambda}^{-1/2} L^r g\|_{L^2(P_{\mathbf{X}})} \\
&\le \|S_M \hat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2}\| \cdot \|C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)\| \cdot \|C_{M,\lambda}^{-1} S_M^* L_{M,\lambda}^{1/2}\| \cdot \|L_{M,\lambda}^{-1/2} L^{1/2}\| \cdot \|L^{r-1/2}\| \cdot \|g\|_{L^2(P_{\mathbf{X}})}.
\end{aligned}$$

From KRR Condition 3, we have $\|g\|_{L^2(P_{\mathbf{X}})} \le R$. Note that

$$\begin{aligned}
\|L^{1/2}\|^2 &= \sup_{\|f\|_{L^2(P_{\mathbf{X}})}=1} \|L^{1/2} f\|_{L^2(P_{\mathbf{X}})}^2 = \sup_{\|f\|_{L^2(P_{\mathbf{X}})}=1} \langle Lf, f\rangle_{L^2(P_{\mathbf{X}})} \\
&= \sup_{\|f\|_{L^2(P_{\mathbf{X}})}=1} \left\langle \int \psi_\omega \otimes \psi_\omega \mathrm{d}\pi(\omega) f, f \right\rangle_{L^2(P_{\mathbf{X}})} \\
&= \sup_{\|f\|_{L^2(P_{\mathbf{X}})}=1} \int \langle \psi_\omega, f\rangle_{L^2(P_{\mathbf{X}})}^2 \mathrm{d}\pi(\omega) \\
&\le \sup_{\|f\|_{L^2(P_{\mathbf{X}})}=1} \int \|\psi_\omega\|_{L^2(P_{\mathbf{X}})}^2 \cdot \|f\|_{L^2(P_{\mathbf{X}})}^2 \mathrm{d}\pi(\omega) \le \kappa^2,
\end{aligned}$$

where the second line follows from (C.14), and the last inequality is based on the boundedness of ψ (KRR Condition 1). This implies that all eigenvalues of L are in $[0, \kappa^2]$. Therefore, we have $\|L^{r-1/2}\| \le \kappa^{2r-1}$ for $r \ge 1/2$.

From (C.7) we have $S_M(S_M^* S_M + \lambda \mathbf{I}_M)^{-2} = (S_M S_M^* + \lambda \mathbf{I}_{L^2(P_{\mathbf{X}})})^{-2} S_M$. Thus,

$$\begin{aligned}
\|C_{M,\lambda}^{-1} S_M^* L_{M,\lambda}^{1/2}\|^2 &= \|L_{M,\lambda}^{1/2} S_M C_{M,\lambda}^{-2} S_M^* L_{M,\lambda}^{1/2}\| = \|L_{M,\lambda}^{1/2} S_M (S_M^* S_M + \lambda \mathbf{I}_M)^{-2} S_M^* L_{M,\lambda}^{1/2}\| \\
&= \|L_{M,\lambda}^{1/2}(S_M S_M^* + \lambda \mathbf{I}_{L^2(P_{\mathbf{X}})})^{-2} S_M S_M^* L_{M,\lambda}^{1/2}\| \\
&= \|L_{M,\lambda}^{1/2}(L_M + \lambda \mathbf{I}_{L^2(P_{\mathbf{X}})})^{-2} L_M L_{M,\lambda}^{1/2}\| \\
&= \|L_{M,\lambda}^{-3/2} L_M L_{M,\lambda}^{1/2}\| = \|L_M L_{M,\lambda}^{-3/2} L_{M,\lambda}^{1/2}\| = \|L_M L_{M,\lambda}^{-1}\| \le 1.
\end{aligned}$$

Here, we use the fact that L_M and $L_{M,\lambda}^{-3/2}$ commute, as they share the same set of eigen-vectors.

Combining the bounds above, we have that

$$E_{33} \le \|S_M \hat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2}\| \cdot \|C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)\| \cdot 1 \cdot \|L_{M,\lambda}^{-1/2} L^{1/2}\| \cdot \kappa^{2r-1} \cdot R = R\kappa^{2r-1} b_1 b_2 B, \tag{C.15}$$

where $b_2 = \|L_{M,\lambda}^{-1/2} L^{1/2}\|$, $B = \|C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)\|$, and $\|S_M \hat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2}\|$ is the term b_1 already defined and bounded in Appendix C.3.1. In what follows we bound b_2 and B respectively.

Bounding b_2: From KRR Condition 1 and Proposition 2.6, we have that $\|L - L_M\| \le C \cdot \frac{\log^a M}{M}$. Therefore,

$$\begin{aligned}
b_2 = \|L_{M,\lambda}^{-1/2} L^{1/2}\| &= \sqrt{\|L_{M,\lambda}^{-1/2} L L_{M,\lambda}^{-1/2}\|} \\
&\le \sqrt{\|L_{M,\lambda}^{-1/2}(L - L_M) L_{M,\lambda}^{-1/2}\| + \|L_{M,\lambda}^{-1/2} L_M L_{M,\lambda}^{-1/2}\|} \\
&\le \sqrt{\|L_{M,\lambda}^{-1/2}(L - L_M) L_{M,\lambda}^{-1/2}\| + 1} \\
&\le \sqrt{\|L_{M,\lambda}^{-1/2}\| \cdot \|L - L_M\| \cdot \|L_{M,\lambda}^{-1/2}\| + 1} \\
&\le \sqrt{\frac{C \log^a M}{\lambda M} + 1}.
\end{aligned}$$

Recall that $M = \frac{\log^a(1/\lambda)}{\lambda}$. We have

$$\begin{aligned}
\frac{C \log^a M}{\lambda M} &= \frac{C \log^a \left(\frac{\log^a(1/\lambda)}{\lambda}\right)}{\log^a(1/\lambda)} = \frac{C\left(\log(1/\lambda) + \log(\log^a(1/\lambda))\right)^a}{\log^a(1/\lambda)} && \text{(C.16)} \\
&= C\left(1 + a\frac{\log\log(1/\lambda)}{\log(1/\lambda)}\right)^a \le C\left(1+a\right)^a, && \text{(C.17)}
\end{aligned}$$

where we have used a loose bound that $\frac{\log\log(1/\lambda)}{\log(1/\lambda)} \le 1$ for $\lambda \in (0, 1/e]$. In conclusion, when $\lambda \in (0, 1/e]$, we have $b_2 \le \sqrt{C(1+a)^a + 1}$. Note that this is a deterministic rather than a probabilistic bound.

Bounding B: To bound $B = \|C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)\|$, we apply Proposition C.3 with $v = z = \boldsymbol{\phi}_M(\mathbf{X})$ and $v_i = z_i = \boldsymbol{\phi}_M(\mathbf{X}_i)$. We first verify its conditions: $\|v\| = \|z\| \le \kappa$ by $|\psi(\mathbf{x}, \omega)| \le \kappa$ from KRR Condition 1; for $v = \boldsymbol{\phi}_M(\mathbf{X})$, $Q = \mathbb{E} v \otimes v = C_M$, we have $\|(C_M + \lambda I)^{-1/2} v\|^2 \le \|v\|^2 \lambda^{-1} \le \kappa^2 \lambda^{-1}$. Hence we can take $\tilde{F}_\infty(\lambda) = \kappa^2 \lambda^{-1}$. By definition,

$$\begin{aligned}
\tilde{\mathcal{N}}(\lambda) &= \operatorname{tr}\left((Q + \lambda I)^{-1} Q\right) = \operatorname{tr}\left((Q + \lambda I)^{-1} \mathbb{E} v \otimes v\right) = \mathbb{E}[v^\top (Q + \lambda I)^{-1} v] \\
&\le \operatorname{ess\,sup}_{v \in \mathcal{H}} \|(Q + \lambda I)^{-1/2} v\|^2 \le \tilde{F}_\infty(\lambda).
\end{aligned}$$

Thus, by Proposition C.3, for any $\tau > 0$, with probability at least $1 - \tau$,

$$\begin{aligned}
B = \|C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)\| \le \|C_{M,\lambda}^{-1/2}(C_M - \hat{C}_M)\|_{\mathrm{HS}} &\le \frac{4\kappa^2 \log\frac{2}{\tau}}{n\sqrt{\lambda}} + \sqrt{\frac{4\kappa^4 \log \frac{2}{\tau}}{n\lambda}} \\
&\le \left(4\kappa^2 \log \frac{2}{\tau} + \sqrt{4\kappa^2 \log \frac{2}{\tau}}\right) \frac{1}{\sqrt{n\lambda}}.
\end{aligned}$$

Combining the bound on b_1 in Appendix C.3.1 and $b_2 \le \sqrt{C(1+a)^a + 1}$ with the above bound on B and setting τ to be $\delta/3$, we have from (C.15) that

$$E_{33} \le R\kappa^{2r-1} b_1 b_2 B \le \frac{3}{2} \cdot \sqrt{C\left(1+a\right)^a + 1} \cdot \left(4R\kappa^{2r+1} \log\frac{6}{\delta} + 2R\kappa^{2r}\sqrt{\log\frac{6}{\delta}}\right) \frac{1}{\sqrt{n\lambda}}$$

with probability $1 - 2\delta/3$, which proves the statement of (C.12).

Moreover, combining the bounds on b_1, b_2, A and B, we have that (C.11) and (C.12) hold simultaneously with probability at least $1 - \delta$ when $\frac{19\kappa^2}{n} \log \frac{3n}{2\delta} \le \lambda \le \min\{\|L\| - \frac{C \log^a M}{M}, 1/e\}$ and $n \ge \max\{405\kappa^2, 67\kappa^2 \log \frac{3\kappa^2}{2\delta}\}$.

C.3.4. Supplementary Technical Lemmas

Proposition C.2 (Bernstein's inequality, Rudi & Rosasco 2017, Appendix B, Proposition 2)**.** *Let $z_1, \ldots, z_n$ be a sequence of i.i.d. random vectors on a separable Hilbert space $\mathcal{H}$. Assume $\mu = \mathbb{E} z_i$ exists and there exist $\sigma, D \ge 0$ such that*

$$\mathbb{E}\|z_i - \mu\|_{\mathcal{H}}^k \le \frac{1}{2} k! \sigma^2 D^{k-2}, \quad \textit{for all } k \ge 2,\ i \in \{1, \ldots, n\}.$$

Then for any $\delta \in (0,1]$, with probability at least $1-\delta$,

$$\left\| \frac{1}{n}\sum_{i=1}^{n} z_i - \mu \right\|_{\mathcal{H}} \le \frac{2D\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\sigma^2\log\frac{2}{\delta}}{n}}.$$

Proposition C.3 (Rudi & Rosasco 2017, Appendix D, Proposition 5)**.** *Let $\mathcal{H}$, $\mathcal{K}$ be two separable Hilbert spaces and $(v_1, z_1), \ldots, (v_n, z_n) \in \mathcal{H}\times\mathcal{K}$ be n i.i.d. random vectors such that $\|v\|_{\mathcal{H}} \le \kappa$ and $\|z\|_{\mathcal{K}} \le \kappa$ almost surely for some constant $\kappa > 0$. Let $Q = \mathbb{E}v\otimes v$, $T = \mathbb{E}v\otimes z$, and $T_n = \frac{1}{n}\sum_{i=1}^{n} v_i \otimes z_i$. Define*

$$\tilde{\mathcal{F}}_\infty(\lambda) := \operatorname{ess\,sup}_{v\in\mathcal{H}} \|(Q+\lambda I)^{-1/2}v\|^2, \quad \tilde{\mathcal{N}}(\lambda) := \operatorname{tr}\left((Q+\lambda I)^{-1}Q\right).$$

For any $0 < \lambda < \|Q\|$ and any $\tau > 0$, with probability at least $1-\tau$, the following holds

$$\|(Q+\lambda I)^{-1/2}(T - T_n)\|_{\mathrm{HS}} \le \frac{4\sqrt{\tilde{\mathcal{F}}_\infty(\lambda)}\kappa\log\frac{2}{\tau}}{n} + \sqrt{\frac{4\kappa^2\tilde{\mathcal{N}}(\lambda)\log\frac{2}{\tau}}{n}}.$$

Proposition C.4 (Rudi & Rosasco 2017, Appendix D, Proposition 6)**.** *Let $v_1, \ldots, v_n$ be n i.i.d. random vectors on a separable Hilbert space $\mathcal{H}$ such that $Q = \mathbb{E}v\otimes v$ is trace-class, and for any $\lambda > 0$, there exists a constant $\mathcal{F}_\infty(\lambda) < \infty$ such that $\langle v, (Q+\lambda I)^{-1}v\rangle \le \mathcal{F}_\infty(\lambda)$ almost surely. Let $Q_n = \frac{1}{n}\sum_{i=1}^{n} v_i\otimes v_i$ and take $0 < \lambda \le \|Q\|$. Then for any $\delta \ge 0$, the following holds with probability at least $1-2\delta$:*

$$\|(Q+\lambda I)^{-1/2}(Q - Q_n)(Q+\lambda I)^{-1/2}\| \le \frac{2\beta(1+\mathcal{F}_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}},$$

where $\beta = \log\frac{4\operatorname{tr}Q}{\lambda\delta}$. Moreover, with the same probability,

$$\lambda_{\max}\left((Q+\lambda I)^{-1/2}(Q - Q_n)(Q+\lambda I)^{-1/2}\right) \le \frac{2\beta}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}.$$

Consequently, if $\|v\| \le \kappa$ almost surely, when $\frac{19\kappa^2}{n}\log\frac{n}{4\delta} \le \lambda \le \|Q\|$ and $n \ge \max\left\{405\kappa^2, 67\kappa^2\log\frac{\kappa^2}{2\delta}\right\}$, with probability at least $1-\delta$, we have

$$\lambda_{\max}\left((Q+\lambda I)^{-1/2}(Q - Q_n)(Q+\lambda I)^{-1/2}\right) \le 1/3.$$

Proposition C.5 (Rudi & Rosasco 2017, Appendix D, Proposition 8)**.** *Let $\mathcal{H}$ be a separable Hilbert space, A, B be two bounded self-adjoint positive linear operators on $\mathcal{H}$, and $\lambda > 0$. Then*

$$\|(A+\lambda I)^{-1/2}B^{1/2}\| \le \|(A+\lambda I)^{-1/2}(B+\lambda I)^{1/2}\| \le (1-\beta)^{-1/2},$$

where $\beta = \lambda_{\max}\left[(B+\lambda I)^{-1/2}(B - A)(B+\lambda I)^{-1/2}\right]$ satisfies $\beta \le \frac{\lambda_{\max}(B)}{\lambda_{\max}(B)+\lambda} < 1$.

D. Additional Simulation Studies and Real Data Examples

D.1. Additional Simulation Studies

Here, we present results corresponding to the $r = 0.5$ case as mentioned in Section 4.2. Similar to what we did before, the training and test data are generated from $Y = f(\mathbf{X}) + \varepsilon$, where f is the regression function, $\mathbf{X} \sim \text{Unif}[0,1]^d$, and $\varepsilon \sim N(0,1)$. We consider two choices of kernels: (i) the *min kernel* $K(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d}\min(x_i, x_i')$, and (ii) the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}'\|_2^2)$, with the bandwidth σ set as the median of $\|\mathbf{X} - \mathbf{X}'\|$ (computed numerically), where $\mathbf{X}, \mathbf{X}'$ i.i.d. $\sim \text{Unif}[0,1]^d$.

From Theorem A.7, when $r = 0.5$, $\operatorname{ran} L^r$ is essentially $\mathcal{H}$. Therefore, for the *min kernel*, we set $\tilde{f}(\mathbf{x}) = 2K(\mathbf{1}_d, \mathbf{x}) - K(\frac{3}{4}\mathbf{1}_d, \mathbf{x}) + 2K(\frac{1}{2}\mathbf{1}_d, \mathbf{x}) - K(\frac{1}{4}\mathbf{1}_d, \mathbf{x})$, and for the Gaussian kernel, we set $\tilde{f}(\mathbf{x}) = K(\frac{1}{3}\mathbf{1}_d, \mathbf{x}) + K(\frac{2}{3}\mathbf{1}_d, \mathbf{x})$. In both cases, we have $\tilde{f} \in \operatorname{ran} L^r$. To approximately control the signal-noise-ratio, we set the regression function as $f(\mathbf{x}) = C_{\tilde{f}} \cdot \tilde{f}(\mathbf{x})$ for some constant $C_{\tilde{f}}$ such that $\mathbb{E}f(\mathbf{X}) = 5$. The kernel ridge regularization constant is set as $\lambda = 0.25n^{-\frac{1}{2r+1}}$.

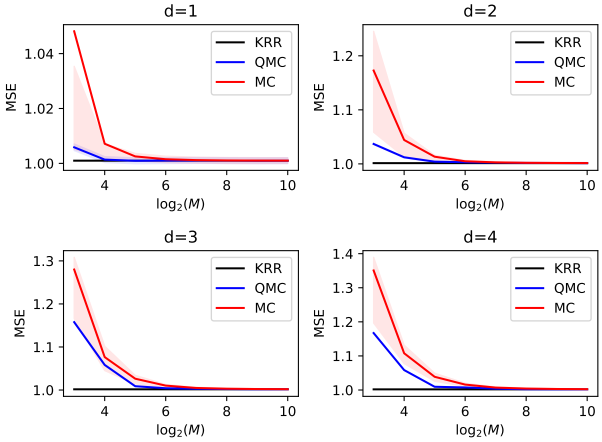

Figure 5. Gaussian Kernel ($r = 0.5$): the test MSE against the number of random features for KRR, RF-KRR and QMCF-KRR.

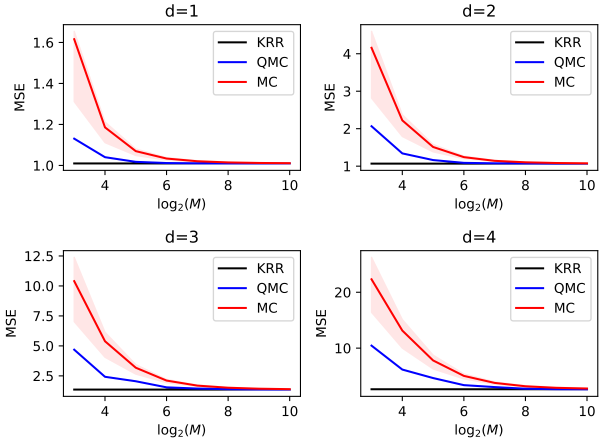

Figure 6. Min Kernel ($r = 0.5$): the test MSE against the number of random features, for KRR, RF-KRR and QMCF-KRR.

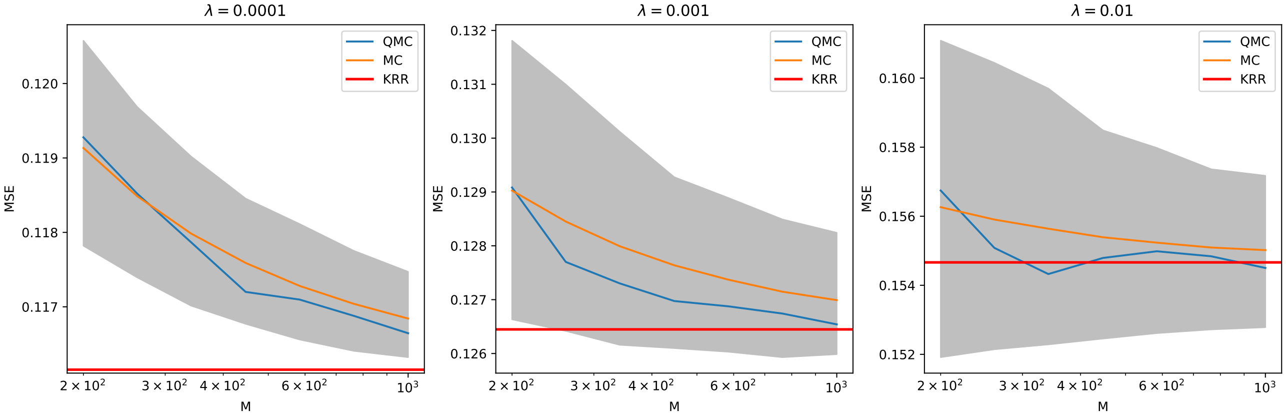

Figure 7. The test MSE against the number of random features for the 'Cadata' data set.

We plot the test MSE against the number of random features, for exact KRR, RF-KRR and QMCF-KRR in Figures 5 and 6. For each combination of kernel and d, 10^6 test data are first generated and held fixed. We consider 1000 realizations of training samples of size 10^4. For each of the realization, we fit a kernel ridge regression and compute its test error. The MSE (solid lines) in Figure 3 and 4 are obtained by averaging over the 1000 realizations. We also provide confidence bands using the 25% and 75% error quantiles from the 1000 realizations.

It is observed that, empirically, the case with $r = 0.5$ case exhibits patterns similar to those of the $r = 1$ case in Section 4.2, which demonstrates the superior performance of QMC features.

D.2. Real Data

We consider the following two real-world examples in this subsection.

Cadata (Pace & Barry, 1997): In this data set ($n = 20640$, $d = 6$), the response is the log of the median house price, and the predictors are median income, housing median age, total rooms, total bedrooms, population, and households. We first perform a random train-test split, allocating 25% of the data to the test set. The response is normalized to have mean 0, and all predictors are normalized to have mean 0 and variance 1, using statistics from the training set. Gaussian kernel is used with the bandwidth empirically set as the median of pairwise distance within the training set.

Cod-rna (Uzilov et al., 2006): This benchmark dataset ($n = 59535$ (train) + 271617 (test), $d = 8$) was developed for

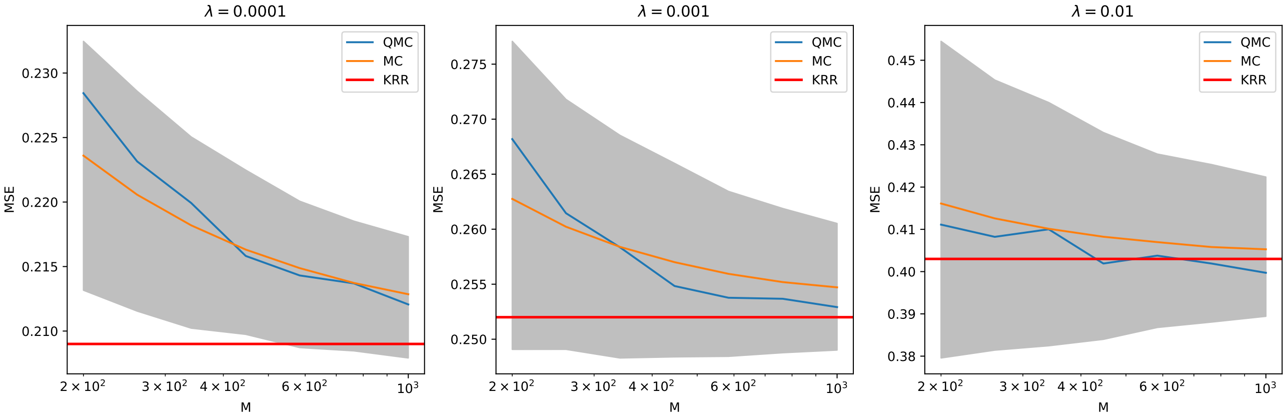

Figure 8. The test MSE against the number of random features for the ‘Cod-rna’ data set.

detecting non-coding RNAs. The response is binary with imbalanced class sizes: in the training set there are 39690 ‘-1’s and 19845 ‘1’s; in the test set there are 181078 ‘-1’s and 90539 ‘1’s. Eight predictors are available: $\Delta G^{\circ}_{\text{total}}$ value computed by the Dynalign algorithm (Uzilov et al., 2006), the length of shorter sequence, and respective ‘A’, ‘U’, ‘C’ frequencies of sequence 1 and sequence 2. We normalize all predictors to have mean 0 and variance 1, using statistics from the training set. The Gaussian kernel is used with the bandwidth empirically set as the median of pairwise distance within 10000 random samples from the training set.

Note that our result (Theorem 3.2) depends on λ of order between $n^{-1/2}$ and $n^{-1/3}$. Thus, for the above two data sets (Cadata: $n^{-1/2} \approx 0.007$, $n^{-1/3} \approx 0.037$; Cod-rna: $n^{-1/2} \approx 0.004$, $n^{-1/3} \approx 0.026$) we consider three choices of λ that are approximately within this range: $\lambda = 0.0001, 0.001, 0.01$. We let M vary from 200 to 1000. Given the randomness of the MC approach, we repeated the RF-KRR process 1000 times with 1000 different seeds; subsequently, we plotted the mean test MSE along with the 95% confidence band (for RF-KRR). The results are shown in Figures 7 and 8.

We observe that for all choices of λ: (i) As M increases, the MSE of QMCF-KRR decreases quickly and converges to that of the exact KRR. (ii) When M is not too small ($M > 200$ for Cadata; $M > 400$ for Cod-rna), QMC outperforms the average performance of MC. (iii) RF-KRR has a very wide confidence band. In terms of providing a high probability MSE upper bound, QMCF-KRR significantly outperforms RF-KRR. For example, when $\lambda = 0.01$, the high probability (97.5% quantile) MSE upper bound for RF-KRR with $M = 1000$ is comparable to that of QMCF-KRR with $M \approx 200$.

D.3. Simulation Results in High Dimensions

We have seen superior performance of QMC features compared with classical random features in low-dimensional settings (in Section 4, Appendix D.1, and Appendix D.2). In this subsection, we present simulation results of RF-KRR and QMCF-KRR in higher dimensions. These results show that the performance of QMC features may be less satisfactory as the dimension increases.

We will follow the same simulation setting as in Section 4.2 and Appendix D.1. We consider both $r = 1$ and $r = 1/2$. The training and test data are generated from $Y = f(\mathbf{X}) + \varepsilon$, where f is the regression function, $\mathbf{X} \sim \text{Unif}[0,1]^d$, and $\varepsilon \sim N(0,1)$. We consider the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}'\|_2^2)$, with the bandwidth σ set as the median of $\|\mathbf{X} - \mathbf{X}'\|$ (computed numerically), where $\mathbf{X}, \mathbf{X}'$ i.i.d. $\sim \text{Unif}[0,1]^d$.

For $r = 1$, we use the same regression function $f(\mathbf{x})$ as in Section 4.2. For $r = 0.5$, we use the same regression function $f(\mathbf{x})$ as in Appendix D.1. The kernel ridge regularization parameter is set as $\lambda = 0.25 n^{-\frac{1}{2r+1}}$.

In Figures 9 and 10, we plot the test MSE against the number of random features, for exact KRR, RF-KRR and QMCF-KRR. For each combination of r and d, 10^6 test data are first generated and held fixed. We consider 1000 realizations of training samples of size 10^4. For each of the realization, we fit a kernel ridge regression and compute its test error (i.e., MSE on the test set). The solid lines in Figures 9, 10 are obtained by averaging over the 1000 realizations. We also provide confidence bands using the 25% and 75% error quantiles from the 1000 realizations. For the MC method, the randomness comes from re-generating the training set and the MC random features. Whereas for the QMC method, it only comes from the training

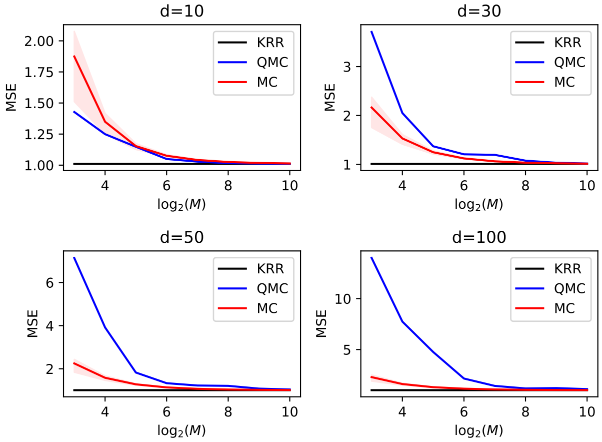

Figure 9. Negative results in higher dimensions: the test MSE against the number of random features for exact KRR, RF-KRR and QMCF-KRR. Gaussian Kernel is used with $r = 1$.

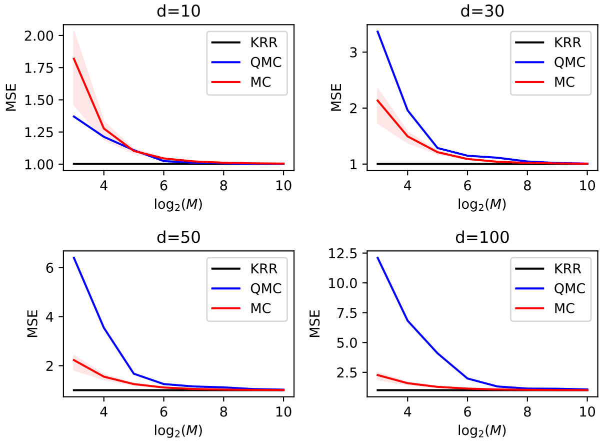

Figure 10. Negative results in higher dimensions: the test MSE against the number of random features, for exact KRR, RF-KRR and QMCF-KRR. Gaussian Kernel is used with $r = 0.5$.

set re-generation as the QMC features are deterministic.

It can be seen from Figures 9 and 10 that for both $r = 1$ and $r = 0.5$, QMC features may not outperform classical random features as the dimension increases, which aligns with the practical observation that the best use case for QMC often arises when the integrand can be well approximated by a sum of functions involving only a small number of its input variables (Owen, 2023; Adcock & Brugiapaglia, 2022). As the dimension increases, the improvement of QMC over MC may be less significant or even worse, as seen in Figures 9 and 10 for $d = 30$, 50 and 100. Nevertheless, for suitably large M, it can be seen that the MSE of QMCF-KRR still decreases quickly, approaching the performance of the exact KRR.