Towards Theoretical Understanding of Learning Large-scale Dependent Data via Random Features

Chao Wang [1] Xin Bing [2] Xin He [1] Caixing Wang [1]

Abstract

Random feature (RF) mapping is an attractive and powerful technique for solving large-scale nonparametric regression. Yet, the existing theoretical analysis crucially relies on the i.i.d. assumption that individuals in the data are independent and identically distributed. It is still unclear whether learning accuracy would be compromised when the i.i.d. assumption is violated. This paper aims to provide theoretical understanding of the kernel ridge regression (KRR) with RFs for large-scale dependent data. Specifically, we consider two types of data dependence structure, namely, the τ-mixing process with exponential decay coefficient, and that with polynomial decay coefficient. Theoretically, we prove that the kernel ridge estimator with RFs achieves the minimax optimality under the exponential decay scenario, but yields a sub-optimal result under the polynomial decay case. Our analysis further reveals how the decay rate of the τ-mixing coefficient impacts the learning accuracy of the kernel ridge estimator with RFs. Extensive numerical experiments on both synthetic and real examples further validate our theoretical findings and support the effectiveness of the KRR with RFs in dealing with dependent data.

1. Introduction

Kernel-based methods (Kimeldorf & Wahba, 1971; Wahba, 1990; Vapnik, 1999; Schölkopf & Smola, 2002) stand as a cornerstone in machine learning community due to their computational convenience, flexible framework, and rich capacity of offering efficient and powerful tools for statistical analysis. Theoretical guarantees for the kernel-based methods have been widely studied in literature (Smale & Zhou, 2007; Caponnetto & De Vito, 2007), just to name a few. Despite their attractive theoretical properties, these methods could suffer severe computational burden in dealing with large-scale data, mainly due to the problem of solving and storing the inverse of the kernel matrix. To alleviate such issue, a variety of methods have been proposed, and one major class is known as kernel approximation, including Nyström subsampling (Rudi et al., 2015), random sketching (Yang et al., 2017) and random features (Rahimi & Recht, 2007). Interested readers are referred to Section 3.3 of Rudi & Rosasco (2017) for detailed comparison and discussion on these methods.

Among them, random feature technique involves constructing an explicit feature mapping with a dimension much lower than the number of observations. It is proven to be state-of-the-art from both computational and theoretical aspects (Rudi & Rosasco, 2017; Bach, 2017; Avron et al., 2017; Sun et al., 2018; Liu & Lian, 2023). Specifically, Rudi & Rosasco (2017) shows the kernel ridge regression estimator with random features (KRR-RF) not only enjoys computational efficiency due to the reduced dimension of RFs, but also preserves the minimax optimal learning rate, compared to the standard KRR, provided that the number of RFs is not chosen too small. However, it is worth pointing out that the results in Rudi & Rosasco (2017) are established under the i.i.d. assumption which could be restrictive in many real applications, including clinical medicine, speech recognition, and traffic data (Ralanamahatana et al., 2005; Fu, 2011). It is unclear whether the aforementioned properties of KRR-RF still hold when the i.i.d. assumption is violated. This leaves an open question in theoretical understanding of KRR-RF in dealing with the dependent data.

This paper attempts to fulfill this gap and provide an answer to the question: *whether, and if so, to what extent does the dependency structure within the data affect the performance of learning with random features?* In the existing literature, data dependence is often characterized by the mixing structure of stochastic processes, such as α-mixing (Modha & Masry, 1996), β-mixing (Yu, 1994) and ϕ-mixing (Birman & Solomyak, 1967). It is worthy pointing out that although

[1] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China [2] Department of Statistical Sciences, University of Toronto. Correspondence to: Xin He <he.xin17@mail.shufe.edu.cn>, Caixing Wang <wang.caixing@stu.sufe.edu.cn>.

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

α-mixing is considered much weaker than the β-mixing and ϕ-mixing conditions, it still excludes some popular types of real applications, including causal linear processes and certain Markov chains. In this paper we focus on a less stringent mixing condition known as τ-mixing (Dedecker & Prieur, 2004; 2005). Indeed Dedecker & Prieur (2004) shows that a α-mixing process must also be τ-mixing, but the reverse may not hold. This paper aims to provide a comprehensive theoretical understanding of KRR-RF for large-scale τ-mixing data. The theoretical investigation faces two major challenges: (1) the resulting estimator can no longer be analyzed by either the classical empirical process theory or the concentration of measures tailored to the i.i.d. case; (2) the approximated kernel function arising from random features is random, thus needs to be quantified carefully. By borrowing a recently developed Bernstein inequality tailored for the τ-mixing process (Blanchard & Zadorozhnyi, 2019) together with advanced operator theory, we are able to establish theoretical guarantees of KRR-RF for τ-mixing dependent data.

1.1. Contributions

Our primary contribution is to provide a comprehensive theoretical understanding of both computational and theoretical aspects of the KRR estimator with RFs for large-scale dependent data. In particular, we begin by introducing two standard capacity and regularity conditions in learning theory. To maintain the mixing property in our technical analysis, we need an additional mild condition that assumes the tail behavior of the derivative of the random feature mapping to be sub-exponential. By utilizing the integral operator (Smale & Zhou, 2007) and a Bernstein-type concentration inequality tailored for the τ-mixing process (Blanchard & Zadorozhnyi, 2019), we derive two key results for the τ-mixing processes with exponential decay coefficient (Theorem 4.2) and polynomial decay coefficient (Theorem 4.3), respectively. Our results further characterize how the degree of dependency structure among the data affects the convergence rate of KRR-RF. Specifically, for the exponential decay τ-mixing process, KRR-RF can achieve the same minimax optimal rate as established under the i.i.d. setting, provided that the number of RFs is not chosen too small. However, for the polynomial decay τ-mixing process, the convergence rate of KRR-RF becomes slower. We also verify the tighter lower bounds on the required number of RFs in Section 4.2 for both decay rates as first considered in Rudi & Rosasco (2017) under i.i.d. case. Extensive simulation studies and a real data analysis corroborate our theoretical findings and demonstrate the effectiveness of KRR-RF for dealing with large-scale dependent data.

1.2. Related Works

We summarize below some of the most related works on random features and learning with dependent data.

Random Features. Random feature (RF) approximation is a popular and powerful tool to approximate kernel matrix using some explicit feature mapping (Rahimi & Recht, 2007; 2008; Li et al., 2019). The theoretical properties of learning with random features have been extensively studied under the regression and classification settings (Rudi & Rosasco, 2017; Avron et al., 2017; Bach, 2017; Sun et al., 2018; Li et al., 2019). The most related work is Rudi & Rosasco (2017) in which the authors establish the capacity-dependent optimal rate that $\mathcal{O}_P(n^{-\frac{2r}{2r+\alpha}})$ for the KRR estimator with RFs. Yet, theoretical investigation in the aforementioned works require i.i.d. assumption, hence are not applicable to dependent data.

Learning with Dependent Data. Many existing works focus on the theoretical behaviors of kernel-based methods for dealing with the dependent data under several mixing conditions, including α-mixing (Modha & Masry, 1996), β-mixing (Yu, 1994), and τ-mixing (Blanchard & Zadorozhnyi, 2019). Yet, most established theoretical results of the estimators based on these mixing data are sub-optimal, largely due to the loss of efficient samples in dealing with dependent data. For example, Modha & Masry (1996); Yu (1994) obtain slower convergence rates of their estimators for the α-mixing data compared to the i.i.d. case under minimum complexity regression estimation framework. A recent study (Blanchard & Zadorozhnyi, 2019) introduces a Bernstein-type inequality for sums of Banach-valued random variables satisfying τ-mixing condition. Based on this technical tool, Blanchard & Zadorozhnyi (2019) and Sun et al. (2022) establish upper bounds for the standard KRR estimator and the KRR estimator with Nyström subsampling under the dependent case. Moreover, Sun & Lin (2022) studies the distributed KRR focusing on the α-mixing dependent data (Modha & Masry, 1996). To the best of our knowledge, theoretical investigation of KRR-RF for large-scale dependent data is still lacking.

2. Preliminaries

In this paper, we consider a stochastic process of random pairs $\{\boldsymbol{Z}_i\}_{i\geq 1}$ that is defined over some probability measure space $(\mathcal{B}, \mathcal{F}, \mathbb{P})$. Suppose $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ is supported on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a compact and convex subset of $\mathcal{R}^d$ and $\mathcal{Y} \subset \mathcal{R}$. We assume that $\{\boldsymbol{Z}_i\}_{i\geq 1}$ is stationary with marginal distribution $\rho_{\boldsymbol{X}}$ on $\mathcal{X}$ and conditional distribution $\rho(y|\,\mathbf{x})$ on $\mathcal{Y}$ given $\boldsymbol{X} = \mathbf{x}$. Moreover, Y_i is assumed to follow

$$Y_i = f_\rho(\boldsymbol{X}_i) + \varepsilon_i, \tag{1}$$

with ε_i being the additive noise and $f_\rho(\mathbf{x}) = E[Y_i|\boldsymbol{X}_i = \mathbf{x}]$ denoting the target function of interest. Note that the stationary assumption is rather mild and it generalizes the i.i.d. assumption. In Section 3, we state the detailed dependence structure of $\{\boldsymbol{Z}_i\}_{i\geq 1}$. For technical reason, we assume that $\mathcal{Y}$ is contained in a finite set $[-U, U]$, which further implies $\|f_\rho\|_\infty \leq U$ with $\|\cdot\|_\infty$ denoting the sup-norm. The same assumption is commonly used in the literature of kernel-based method (Smale & Zhou, 2005; 2007; Blanchard & Zadorozhnyi, 2019) for analytical simplicity. It can be relaxed to assume some moment conditions on ε_i. We defer detailed discussion to Appendix D.

2.1. Kernel Ridge Regression

Under the nonparametric setting, the true target function f_ρ in (1) is often assumed to belong to some function class. In this paper, we assume that f_ρ belongs to a separable reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ induced by some positive symmetric kernel function $K(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \to \mathcal{R}^+$, and we further denote the inner product equipped with $\mathcal{H}_K$ as $\langle\cdot,\cdot\rangle_K$ and the endowed norm as $\|\cdot\|_K^2 = \langle\cdot,\cdot\rangle_K$. Note that the RKHS $\mathcal{H}_K$ is a particular type of Hilbert space of real-value functions f with domain $\mathcal{X}$. It enjoys several nice properties that make it particularly attractive and useful in nonparametric modeling. In particular, it is known that the RKHS $\mathcal{H}_K$ induced by the universal kernels, such as Gaussian and Laplace kernel, is dense in the continuous function space under the infinity norm, and thus leads to small approximation error in estimating any continuous target function. In this paper, we assume the commonly used boundedness condition on the kernel function, that is, $\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} K(\mathbf{x},\mathbf{x}') \leq \kappa^2$ with κ being some positive constant.

Suppose that the data $\{\mathbf{z}_i\}_{i=1}^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are sequentially collected observations of the first n random pairs of $\{\boldsymbol{Z}_i\}_{i\geq1}$. Then, the kernel ridge regression (KRR) estimator is defined as

$$\widehat{f} = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 + \lambda\|f\|_K^2 \right\}, \tag{2}$$

where $\lambda > 0$ is some tuning parameter. By the representer theorem (Kimeldorf & Wahba, 1971), the minimizer of the optimization task (2) must have a closed form that

$$\widehat{f}(\mathbf{x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{x}_i, \mathbf{x}) \quad \forall\ \mathbf{x}\in\mathcal{X}, \tag{3}$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_1,\ldots,\widehat{\alpha}_n)^\top = (\mathbf{K} + \lambda\mathbf{I}_n)^{-1}\mathbf{y} \in \mathcal{R}^n$ with $\mathbf{I}_n$ being the $n\times n$ identity matrix, $\mathbf{K} = \{\frac{1}{n}K(\mathbf{x}_i,\mathbf{x}_j)\}_{i,j=1}^n$ denotes the $n\times n$ kernel matrix and $\mathbf{y} = \frac{1}{\sqrt{n}}(y_1,\ldots,y_n)^\top$. It is worth pointing out that although it has the explicit solution, computing (3) may suffer severe computational and storage burdens when n is large. Precisely, it requires $\mathcal{O}(n^3)$ time complexity to invert $(\mathbf{K} + \lambda\,\mathbf{I}_n)$ and $\mathcal{O}(n^2)$ memory to store $\mathbf{K}$.

2.2. Random Feature Mapping

Recently, the random feature technique has attracted tremendous attentions in the literature (Rahimi & Recht, 2007; Rudi & Rosasco, 2017; Li et al., 2019) because it is powerful and computationally efficient for kernel approximation. Specifically, let (Ω, π) be a probability space. In the rest of this paper, we focus on the RKHS that is induced by the kernel with an integral representation,

$$K(\mathbf{x},\mathbf{x}') = \int_\Omega \psi(\mathbf{x},\boldsymbol{\omega})\psi(\mathbf{x}',\boldsymbol{\omega})d\pi(\boldsymbol{\omega}), \tag{4}$$

where $\psi : \mathcal{X}\times\Omega \to \mathcal{R}$ is a continuous function. It is worth pointing out that various widely used kernels admit this integral representation. For instance, the Bochner's theorem (Rahimi & Recht, 2007) ensures that for any continuous, positive definite and scaled shift-invariant kernel $K(\mathbf{x},\mathbf{x}') = v(\mathbf{x}-\mathbf{x}')$ with some function $v : \mathcal{R}^d \to \mathcal{R}^+$ satisfying $v(\mathbf{0}) = 1$, the Fourier transform $\widehat{v}$ of v can be regarded as a probability density such that for any $\mathbf{x},\mathbf{x}'\in\mathcal{X}$,

$$\begin{aligned} K(\mathbf{x},\mathbf{x}') &= \int_{\mathcal{R}^d} \widehat{v}(\mathbf{w})e^{j\,\mathbf{w}^\top(\mathbf{x}-\mathbf{x}')}\mathrm{d}\,\mathbf{w} \\ &= E_{\mathbf{w}\sim\widehat{v}}[\varphi(\mathbf{x},\mathbf{w})\varphi(\mathbf{x}',\mathbf{w})^*], \end{aligned}$$

where $\varphi(\mathbf{x},\mathbf{w}) = e^{-j\,\mathbf{w}^\top\mathbf{x}}$ with $\mathbf{w}\in\mathcal{R}^d$ and $*$ denotes the complex conjugate transpose. Note that since $\widehat{v}$ and K are both real-valued, $e^{j\,\mathbf{w}^\top(\mathbf{x}-\mathbf{x}')}$ can be replaced with its real-valued part. Indeed, simple algebra demonstrates that the random feature mapping $\psi(\mathbf{x},\boldsymbol{\omega}) = \sqrt{2}\cos(\mathbf{w}^\top\mathbf{x} + b)$ with $\boldsymbol{\omega} = (\mathbf{w}, b)$, $\mathbf{w} \sim \widehat{v}(\mathbf{w})$ and $b \sim \text{Uniform}(0, 2\pi)$ satisfies (4).[1] Note that many popular kernels, including the Gaussian kernel, are shift-invariant.

Based on the integral representation (4), we adopt the Monte Carlo sampling strategy to approximate $K(\mathbf{x},\mathbf{x}')$ as

$$\begin{aligned} K_M(\mathbf{x},\mathbf{x}') &= \frac{1}{M}\sum_{i=1}^M \psi(\mathbf{x},\boldsymbol{\omega}_i)\psi(\mathbf{x}',\boldsymbol{\omega}_i) \\ &= \boldsymbol{\phi}_M(\mathbf{x})^\top\boldsymbol{\phi}_M(\mathbf{x}'), \end{aligned} \tag{5}$$

where $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_M$ are independently drawn from π with $M \ll n$ and $\boldsymbol{\phi}_M(\mathbf{x}) = \frac{1}{\sqrt{M}}(\psi(\mathbf{x},\boldsymbol{\omega}_1), \ldots, \psi(\mathbf{x},\boldsymbol{\omega}_M))^\top \in$

[1] Note that $E_{\boldsymbol{\omega}}[\psi(\mathbf{x},\boldsymbol{\omega})\psi(\mathbf{x}',\boldsymbol{\omega})]$ equals to

$$\begin{aligned} &E_{\boldsymbol{\omega}}[\cos(\mathbf{w}^\top(\mathbf{x}-\mathbf{x}')) + \cos(\mathbf{w}^\top(\mathbf{x}+\mathbf{x}') + 2b)] \\ &= E_{\boldsymbol{\omega}}[\cos(\mathbf{w}^\top(\mathbf{x}-\mathbf{x}'))] = E_{\boldsymbol{\omega}}[e^{j\mathbf{w}^\top(\mathbf{x}-\mathbf{x}')}]. \end{aligned}$$

$\mathcal{R}^{\mathcal{M}}$ denotes the pre-specified M-dimensional random feature mapping. By writing

$$\widehat{\mathbf{S}}_M = \frac{1}{\sqrt{n}}(\boldsymbol{\phi}_M(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_M(\mathbf{x}_n))^\top \in \mathcal{R}^{n \times M},$$

replacing the kernel K in (3) by K_M in (5) and using the Woodbury matrix identity, we obtain the following KRR estimator with random features (KRR-RF):

$$\widehat{f}_{M,\lambda}(\mathbf{x}) = \boldsymbol{\phi}_M(\mathbf{x})^\top (\widehat{\mathbf{S}}_M^\top \widehat{\mathbf{S}}_M + \lambda \mathbf{I}_M)^{-1} \widehat{\mathbf{S}}_M^\top \mathbf{y}. \quad (6)$$

Computationally, the matrix multiplication $\widehat{\mathbf{S}}_M^\top \widehat{\mathbf{S}}_M$ requires time complexity $\mathcal{O}(nM^2)$. Since inverting $\widehat{\mathbf{S}}_M^\top \widehat{\mathbf{S}}_M + \lambda\, \mathbf{I}_M$ requires time complexity $\mathcal{O}(M^3)$, the overall time complexity of computing (6) is $\mathcal{O}(nM^2 + M^3)$. Compared to the standard KRR, the computational complexity is largely reduced as long as $M \ll n$. Moreover, we only need to store the nM entries in $\widehat{\mathbf{S}}_M$, as opposed to the original n^2 entries.

Let $\mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$ denote the square integrable function space with respect to $\rho_{\boldsymbol{X}}$. We denote the inner product endowed with $\mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$ as $\langle \cdot, \cdot \rangle_\rho$ and use $\| \cdot \|_\rho$ to represent the corresponding norm. For any estimator $\widehat{f}$, its estimation accuracy is evaluated in terms of the squared norm

$$\|\widehat{f} - f_\rho\|_\rho^2 = E_{\boldsymbol{X} \sim \rho_{\boldsymbol{X}}}[(\widehat{f}(\boldsymbol{X}) - f_\rho(\boldsymbol{X}))^2].$$

In literature, some regularity conditions on the capacity of $\mathcal{H}_K$ and the smoothness of the target function f_ρ are needed to derive theoretical guarantees. We start with the definition of the integral operator that is widely used in literature (Smale & Zhou, 2007; Caponnetto & De Vito, 2007).

Definition 2.1. The integral operator $L_K : \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}}) \to \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$ is defined as

$$L_K f = \int_{\mathcal{X}} K(\mathbf{x}, \cdot) f(\mathbf{x}) d\rho_{\boldsymbol{X}}(\mathbf{x}).$$

It is known that L_K is positive, trace-class and self-adjoint, hence compact. By the Mercer's theorem, L_K admits a spectral decomposition of the form

$$L_K = \sum_{i=1}^{\infty} \mu_i \langle \cdot, \psi_i \rangle_\rho \psi_i, \quad (7)$$

where $\{\mu_i\}_{i \geq 1}$ are the non-negative eigenvalues in descending order, and $\{\psi_i\}_{i \geq 1}$ are the corresponding eigenfunctions in $\mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$.

Definition 2.2. The effective dimension of the RKHS $\mathcal{H}_K$ is defined as

$$\mathcal{N}(\lambda) = \mathrm{Tr}((L_K + \lambda I)^{-1} L_K),$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a trace-class operator.

Note that $\mathcal{N}(\lambda)$ measures the capacity of $\mathcal{H}_K$ with respect to $\rho_{\boldsymbol{X}}$ and it is closely related to the covering number (Steinwart & Christmann, 2008) and the kernel complexity function (Guo et al., 2017; Ma et al., 2023). The following assumption further characterizes the richness of $\mathcal{H}_K$.

Assumption 2.3 (Capacity condition)**.** There exist two constants $\alpha \in [0, 1]$ and $c_0 \geq 0$ such that $\mathcal{N}(\lambda) \leq c_0 \lambda^{-\alpha}$.

Note that we assume $\lambda \in (0, 1]$ and the theoretical choice of λ typically depends on n and tends to zero as $n \to \infty$. Assumption 2.3 is commonly imposed in the literature of learning theory (Guo et al., 2017; Rudi & Rosasco, 2017). Precisely, α controls the richness of $\mathcal{H}_K$ in the sense that a larger α implies a larger capacity of $\mathcal{H}_K$. It can be easily verified that Assumption 2.3 always holds when $\alpha = 1$ by taking $c_0 = \mathrm{Tr}(L_K) \leq \kappa^2$. Assumption 2.3 is more general than the polynomial decay condition in Caponnetto & De Vito (2007): $\mu_j \leq Cj^{-1/\alpha}$ for $\alpha \in (0, 1)$ with μ_j's being the eigenvalues of L_K given by (7). Moreover, Assumption 2.3 holds with $\alpha = 0$ if the kernel function has finite rank, such as the linear kernel $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$.

Assumption 2.4 (Regularity condition)**.** There exist some $r \in [1/2, 1]$ and $g \in \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$ satisfying $f_\rho = L_K^r g$ with $L_K^r = \sum_{i=1}^\infty \mu_i^r \langle \cdot, \psi_i \rangle_\rho \psi_i$.

Assumption 2.4 is also commonly assumed in literature (Smale & Zhou, 2007; Caponnetto & De Vito, 2007; Guo et al., 2017). It requires that f_ρ belongs to the range of L_K^r and when $r = 1/2$, it only requires that f_ρ belongs to $\mathcal{H}_K$. Note that r controls the smoothness of f_ρ and a larger r corresponds to a smoother f_ρ.

To be self-contained, we restate Theorem 2 in Rudi & Rosasco (2017) that provides theoretical guarantees for the KRR-RF estimator under the i.i.d data case.

Theorem 2.5. *Under Assumptions 2.3 and 2.4, assume that $\{\boldsymbol{Z}_i\}_{i \geq 1}$ are i.i.d. For any $\delta \in (0, 1)$, if the number of random features satisfies $M \geq Cn^{\frac{1+\alpha(2r-1)}{2r+\alpha}} \log \frac{108\kappa^2 n}{\delta}$ and $\lambda \asymp n^{-\frac{1}{2r+\alpha}}$, then with probability at least $1 - \delta$, the following holds for sufficiently large n,*

$$\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho^2 \leq Cn^{-\frac{2r}{2r+\alpha}} \log^2(1/\delta),$$

where $C > 0$ is some constant independent of n, α, r.

Theorem 2.5 shows that under the i.i.d. case, the KRR-RF estimator can achieve the minimax rate $\mathcal{O}_P(n^{-\frac{2r}{2r+\alpha}})$ (Caponnetto & De Vito, 2007) with the number of random features chosen properly. However, it is unclear yet important whether the above optimal learning rate can be achieved in the presence of data dependence. In the sequel, we answer this question by analyzing the theoretical behavior of the KRR-RF estimator for dependent data.

3. Learning with Dependent Data

For measuring dependency among individual data points, we consider the τ-mixing condition (Dedecker & Prieur, 2004; 2005). Recall that the pairs $\{\boldsymbol{Z}_i\}_{i\geq 1}$ are defined over some probability measure space $(\mathcal{B}, \mathcal{F}, \mathbb{P})$. We start by defining an increasing sequence of σ-fields $\{\mathcal{F}_k\}_{k\geq 0}$ as

$$\mathcal{F}_k = \sigma\big(\boldsymbol{Z}_i, 1 \leq i \leq k\big) \subset \mathcal{F},$$

which is induced by the first k random pairs. We also define a real-valued, bounded Lipschitz function class $\mathcal{C}_{\text{Lip}}$ on $\mathcal{X} \times \mathcal{Y}$ with a finite Lipschitz semi-norm

$$\|g\|_{\text{Lip}(\mathcal{X})} = \sup\left\{\frac{|g(\mathbf{z}) - g(\mathbf{z}')|}{\|\,\mathbf{z} - \mathbf{z}'\|_2} : \mathbf{z}, \mathbf{z}' \in \mathcal{X} \times \mathcal{Y}, \mathbf{z} \neq \mathbf{z}'\right\}.$$

Based on the function class $\mathcal{C}_{\text{Lip}}$, the dependence structure can be characterized by the following mixing condition.

Definition 3.1. For $k \geq 1$, we define $\tau(k)$ as

$$\tau(k) = \sup\Big\{\big\|E\big[f(\boldsymbol{Z}_{i+k}) \mid \mathcal{F}_i\big] - E\big[f(\boldsymbol{Z}_{i+k})\big]\big\|_\infty : \\ f \in \mathcal{C}_{\text{Lip}}, i \geq 1\Big\},$$

where $\|\cdot\|_\infty$ denotes the sup-norm in $\mathcal{L}_\infty(\mathcal{B}, \mathcal{F}, \mathbb{P})$. A stochastic process $\{\boldsymbol{Z}_i\}_{i\geq 1}$ is a τ-mixing process with rate $\tau(k)$ if $\lim_{k\to\infty} \tau(k) = 0$.

This mixing condition commonly appears in literature (Dedecker & Prieur, 2004; 2005; Maume-Deschamps, 2006; Blanchard & Zadorozhnyi, 2019) with $\tau(k)$ often referred to as the τ-mixing coefficient. Note that based on the function class $\mathcal{C}_{\text{Lip}}$, $\tau(k)$ can effectively quantify the distance between the expectation of $\boldsymbol{Z}_{i+k}$ conditioning on the information from the past k periods and the unconditional expectation of $\boldsymbol{Z}_{i+k}$ for each i. Therefore, $\tau(k)$ measures the dependence of lag k periods. Precisely, a faster convergence rate of $\tau(k) \to 0$ implies weaker dependence among individual data points, and vice versa. In particular, when the data are i.i.d. we have $\tau(k) = 0$ for all $k \geq 1$. It is worth pointing out that the τ-mixing condition is weaker than the α-mixing condition (Modha & Masry, 1996) as well as the well-known ϕ-mixing (Birman & Solomyak, 1967) and β-mixing (Yu, 1994) conditions. The following proposition plays a crucial role in the subsequent technical analysis.

Proposition 3.2. *If $\{\boldsymbol{Z}_i\}_{i\geq 1}$ is a τ-mixing process with rate $\tau(k)$ and $g : \mathcal{X} \times \mathcal{R} \to \mathcal{B}$ is c-Lipschitz continuous, where $\mathcal{B}$ is some Banach space. Then $\{g(\boldsymbol{Z}_i)\}_{i\geq 1}$ is also τ-mixing process with τ-mixing coefficient $c\,\tau(k)$.*

We consider two types of decay rates of $\tau(k)$.

(i) **Exponential decay:** $\tau(k) \leq b_0 \exp\big(-(b_1 k)^{\gamma_0}\big)$, for some constants $\gamma_0, b_1 > 0, b_0 \geq 0$;

(ii) **Polynomial decay:** $\tau(k) \leq b_2 k^{-\gamma_1}$, for some constants $\gamma_1 > 0, b_2 \geq 0$.

Cases (i) and (ii) cover a wide range of practical applications, including: (i) causal Bernoulli shifts such as causal linear processes; (ii) iterative random functions such as autoregressive processes; and (iii) other Markov chains (Dedecker & Prieur, 2004; 2005). We refer the readers to our numerical experiments in Section 5 for concrete examples. We also remark that the i.i.d. assumption can be regarded as an exponential τ-mixing process with $\gamma_0 = \infty$, or a polynomial τ-mixing process with $\gamma_1 = \infty$.

4. Theoretical Analysis

In the literature of KRR with dependent data, a key technical step is to establish the Lipschitz continuity of the kernel function $K(\cdot,\cdot)$ so that Proposition 3.2 can be invoked. To this end, it is commonly assumed (Blanchard & Zadorozhnyi, 2019; Sun et al., 2022) that

$$\max_{1\leq l,k\leq d} \sup_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \left|\frac{\partial^2 K(\mathbf{x}, \mathbf{x}')}{\partial x_l \partial x_k'}\right| \leq B, \tag{8}$$

for some constant $B > 0$. Here $\mathbf{x} = (x_1, ..., x_d)^\top$.

Differently, our theoretical framework is built upon the Lipschitz continuity of the M-dimensional random feature mapping $\boldsymbol{\phi}_M(\cdot)$, more specifically, $\boldsymbol{\phi}_M(\cdot)$ in (5) which is inherently non-deterministic due to its dependence on the RFs $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M$. Let $\partial_{\mathbf{x}}\psi(\mathbf{x}, \boldsymbol{\omega}) \in \mathcal{R}^d$ be the partial derivative of $\psi(\mathbf{x}, \boldsymbol{\omega})$ with respect to $\mathbf{x}$. Our analysis relies on the following assumption on the tail probability of $T(\boldsymbol{\omega}) := \sup_{\mathbf{x}\in\mathcal{X}} \|\partial_{\mathbf{x}}\psi(\mathbf{x}, \boldsymbol{\omega})\|_2^2/d$.

Assumption 4.1. There exist some positive constants a_0 and a_1 such that for all $t > 0$,

$$P_{\boldsymbol{\omega}}\big(T(\boldsymbol{\omega}) > t\big) \leq a_0 \exp(-a_1 t).$$

Assumption 4.1 requires the sub-exponential tail of $T(\boldsymbol{\omega})$, and is used to establish the Lipschitz property of $\boldsymbol{\phi}_M(\cdot)$. Indeed, we prove Lemma C.1 of Appendix C that for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

$$\|\boldsymbol{\phi}_M(\mathbf{x}) - \boldsymbol{\phi}_M(\mathbf{x}')\|_2^2 \leq d\|\,\mathbf{x} - \mathbf{x}'\|_2^2\, \frac{1}{M}\sum_{i=1}^M T(\boldsymbol{\omega}_i).$$

Since the rightmost term is the empirical average of $T(\boldsymbol{\omega})$, an application of Bernstein's inequality leads to the Lipschitz continuity of $\boldsymbol{\phi}_M(\cdot)$. On the other hand, it is worth pointing out that Assumption 4.1 is mild, and can be verified for many standard kernels. For the linear kernel $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top\mathbf{x}'$, we have $\psi(\mathbf{x}, \boldsymbol{\omega}) = \mathbf{x}^\top\boldsymbol{\omega}$ with $\boldsymbol{\omega} \sim N(0, \mathbf{I}_d)$ whence $T(\boldsymbol{\omega}) = \|\boldsymbol{\omega}\|_2^2/d$ and Assumption 4.1 holds. For the shift-invariant kernel class, since

$\psi(\mathbf{x},\boldsymbol{\omega}) = \sqrt{2}\cos(\mathbf{w}^\top \mathbf{x} + b)$ with $\boldsymbol{\omega} = (\mathbf{w}, b)$, we have $T(\boldsymbol{\omega}) \le 2\|\boldsymbol{\omega}\|_2^2/d$ so that Assumption 4.1 holds if the random vector $\mathbf{w}$ is sub-Gaussian, such as $\mathbf{w} \sim N(0, 2\mathbf{I}_d)$ for the Gaussian kernel (Rahimi & Recht, 2007).

4.1. Main Results

In this section, we derive non-asymptotic upper bounds of the KRR-RF estimator $\widehat{f}_{M,\lambda}$ in (6) under two types of the τ-mixing decay rate. Specifically, the learning rates for the τ-mixing process with exponential and polynomial decay coefficients are provided in Theorem 4.2 and 4.3, respectively. Recall that we assume $|Y| \le U$, $K(\mathbf{x},\mathbf{x}') \le \kappa^2$ for any $\mathbf{x},\mathbf{x}' \in \mathcal{X}$. We also assume in the rest of this paper that $|\psi(\mathbf{x},\boldsymbol{\omega})| \le \kappa$ for any $\mathbf{x},\mathbf{x}' \in \mathcal{X}$ and $\boldsymbol{\omega} \in \Omega$. Both U and κ will be treated as absolute constants throughout our analysis.

Theorem 4.2. *Suppose that Assumptions 2.3, 2.4 and 4.1 are satisfied and $\{\boldsymbol{Z}_i\}_{i\ge 1}$ is a τ-mixing process with exponential decay coefficient. For any $\delta \in (0,1)$, if the number of RFs satisfies*

$$M \ge C n^{\frac{1+\alpha(2r-1)}{2r+\alpha}} (\log n)^{-\frac{1}{4\gamma_0}} \log(n/\delta)$$

and $\lambda \asymp (n(\log n)^{-\frac{1}{2\gamma_0}})^{-\frac{1}{2r+\alpha}}$, then when n is sufficiently large, with probability at least that $1-\delta$, one has

$$\left\|\widehat{f}_{M,\lambda} - f_\rho\right\|_\rho^2 \le C n^{-\frac{2r}{2r+\alpha}} (\log n)^{\frac{r}{\gamma_0(2r+\alpha)}} \log(1/\delta).$$

The constant $C > 0$ is independent of n, α, r, γ_0.

Theorem 4.2 states that the learning rate of the KRR-RF estimator is $\mathcal{O}_P(n^{-\frac{2r}{2r+\alpha}})$ (up to some logarithmic factor) for the τ-mixing data with exponential decay coefficient. Since this rate matches the minimax lower bound in the i.i.d. case (Caponnetto & De Vito, 2007), Theorem 4.2 extends the minimax optimality of the KRR-RF estimator from the i.i.d. case to exponential decay τ-mixing data.

In another word, the dependence of a τ-mixing process does not compromise the effectiveness of learning with KRR-RF as long as the τ-mixing coefficient decays exponentially. Theorem 4.2 also states that the optimal learning rate requires to choose the number of random features $M \ge C n^{\frac{1+\alpha(2r-1)}{2r+\alpha}} (\log n)^{-\frac{1}{4\gamma_0}}$ which matches that required in Theorem 2.5 (up to some logarithmic factor). It suggests that data dependence does not necessitate a larger number of random features, thus with no additional computation burden.

Theorem 4.3. *Suppose that Assumptions 2.3, 2.4 and 4.1 are satisfied and $\{\boldsymbol{Z}_i\}_{i\ge 1}$ is a τ-mixing process with polynomial decay coefficient. For any $\delta \in (0,1)$, if the number of RFs satisfies*

$$M \ge C n^{\frac{2\gamma_1(1+\alpha(2r-1))}{4\gamma_1 r + 2r + 2\alpha\gamma_1 + 1}} \log(n/\delta)$$

and $\lambda \asymp n^{-\frac{2\gamma_1}{4\gamma_1 r + 2r + 2\gamma_1\alpha + 1}}$, then when n is sufficiently large, with probability at least $1-\delta$, one has

$$\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho^2 \le C n^{-\frac{4\gamma_1 r}{4\gamma_1 r + 2r + 2\gamma_1\alpha + 1}} \log(1/\delta).$$

The constant $C > 0$ is independent of n, α, r, γ_1.

The learning rate established in Theorem 4.3 aligns with that of the full KRR for the polynomial τ-mixing process (Blanchard & Zadorozhnyi, 2019). On the other hand, Theorem 4.3 shows that the KRR-RF estimator can not achieve the optimal rate established under the i.i.d. setting when dealing with the τ-mixing data with polynomial decay coefficient. This indicates potential loss of efficiency of the KRR-RF estimator in the presence of strong data dependence. The parameter γ_1 in Theorem 4.3 controls the degree of dependency among the data in the sense that the larger value of γ_1, the weaker dependence in the data. As $\gamma_1 \to \infty$, the data becomes less dependent and the learning rate in Theorem 4.3 gets closer to the optimal rate $n^{-\frac{2r}{2r+\alpha}}$. We remark that the sub-optimality comes from the fact that the effective sample size is degraded to $n^{\frac{2\gamma_1(2r+\alpha)}{2\gamma_1(2r+\alpha)+2r+1}}$, a quantity that increases as γ_1 gets larger. Simply increasing the number of RFs will not lead to the improvement of the rate in Theorem 4.3. Finally, the required lower bound of RFs in Theorem 4.3 is $n^{\frac{2\gamma_1(1+\alpha(2r-1))}{4\gamma_1 r + 2r + 2\alpha\gamma_1 + 1}}$ which gets smaller as γ_1 decreases.

4.2. Tighter Bounds on the number of random features

In this section, we further improve the required lower bounds of M in Theorems 4.2 and 4.3. This is particularly useful as a smaller value of M reduces more computational cost. We first introduce the random feature maximum effective dimension, as introduced in the literature (Rudi et al., 2015; Rudi & Rosasco, 2017).

Definition 4.4. The maximum random feature effective dimension related to $\mathcal{H}_K$ is defined as

$$\mathcal{N}_\infty(\lambda) = \sup_{\boldsymbol{\omega}\in\Omega} \|(L_K + \lambda I)^{-1/2}\psi(\cdot,\boldsymbol{\omega})\|_\rho^2.$$

Note that $\mathcal{N}_\infty(\lambda)$ also measures the capacity of $\mathcal{H}_K$ with respect to $\rho_{\boldsymbol{X}}$ in a RF-dependent manner. It is related with $\mathcal{N}(\lambda)$ in Definition 2.2 in that $\mathcal{N}(\lambda) \le \mathcal{N}_\infty(\lambda)$. We assume the following compatibility condition on $\mathcal{N}_\infty(\lambda)$.

Assumption 4.5 (Compatibility condition)**.** There exist two constants $\beta \in [0,1]$ and $c_1 \ge 1$ such that $\mathcal{N}_\infty(\lambda) \le c_1\lambda^{-\beta}$.

Since $|\psi(\mathbf{x},\boldsymbol{\omega})| \le \kappa$ for any $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\omega} \in \Omega$, we always have $\mathcal{N}_\infty(\lambda) \le \kappa^2\lambda^{-1}$. Assumption 4.5 considers a tighter bound on $\mathcal{N}_\infty(\lambda)$, which is also assumed in Rudi & Rosasco (2017), and interested readers are referred therein for detailed discussions. By the fact that $\mathcal{N}(\lambda) \le \mathcal{N}_\infty(\lambda)$, we have $0 < \alpha \le \beta \le 1$. Armed with Assumption 4.5, a much tighter lower bound of M can be derived without sacrificing the learning accuracy. The results are stated in the following two corollaries for both exponential and polynomial decay coefficients.

Corollary 4.6. *Suppose that Assumptions 2.3, 2.4, 4.1 and 4.5 are satisfied and $\{\boldsymbol{Z}_i\}_{i\geq 1}$ is a τ-mixing process with exponential decay coefficient. For any $\delta \in (0,1)$, if the number of RFs satisfies*

$$M \geq Cn^{\frac{\beta+(1+\alpha-\beta)(2r-1)}{2r+\alpha}}(\log n)^{-\frac{1}{4\gamma_0}}\log(n/\delta)$$

and $\lambda \asymp (n(\log n)^{-\frac{1}{2\gamma_0}})^{-\frac{1}{2r+\alpha}}$, then when n is sufficiently large, with probability at least $1-\delta$, one has

$$\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho^2 \leq Cn^{-\frac{2r}{2r+\alpha}}(\log n)^{\frac{r}{\gamma_0(2r+\alpha)}}\log(1/\delta).$$

The constant $C > 0$ is independent of $n, \alpha, \beta, r, \gamma_0$.

For the exponential τ-mixing process, the learning rate in Corollary 4.6 is the same as that in Theorem 4.2 whereas the required number of RFs is substantially reduced due to the usage of maximum effective dimension. Specifically, we only need $M \geq Cn^{\frac{\beta+(1+\alpha-\beta)(2r-1)}{2r+\alpha}}(\log n)^{-\frac{1}{4\gamma_0}}$ which could be much smaller than that in Theorem 4.2 for $\beta < 1$. This also aligns with that required in Theorem 3 in Rudi & Rosasco (2017) under the i.i.d. setting.

Corollary 4.7. *Suppose that Assumptions 2.3, 2.4, 4.1 and 4.5 are satisfied and $\{\boldsymbol{Z}_i\}_{i\geq 1}$ is a τ-mixing process with polynomial coefficient decay. For any $\delta \in (0,1)$, if the number of RFs satisfies*

$$M \geq Cn^{\frac{2\gamma_1\beta+2\gamma_1(1+\alpha-\beta)(2r-1)}{4\gamma_1 r+2r+2\alpha\gamma_1+1}}\log(n/\delta)$$

and $\lambda \asymp n^{-\frac{2\gamma_1}{4\gamma_1 r+2r+2\gamma_1\alpha+1}}$, then when n is sufficiently large, with probability at least $1-\delta$, one has

$$\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho^2 \leq Cn^{-\frac{4\gamma_1 r}{4\gamma_1 r+2r+2\gamma_1\alpha+1}}\log(1/\delta).$$

The constant $C > 0$ is independent of $n, \alpha, \beta, r, \gamma_1$.

For the polynomial τ-mixing process, Corollary 4.7 presents a similar improvement over Theorem 4.3. Furthermore, we notice that the required number of RFs in Corollary 4.7 decreases as β becomes smaller. Since $\beta \geq \alpha$, we need the fewest number of RFs when $\beta = \alpha$. It is worthy pointing out that β depends on the choice of the distribution of π from which the random features are sampled. As discussed in Rudi et al. (2015); Rudi & Rosasco (2017), the favorable situation $\beta = \alpha$ could be achieved through a data-dependent sampling strategy. We refer to Appendix F for details of such sampling strategy.

It is also interesting to point out that the significant reduction in the required number of RFs for the polynomial τ-mixing process is largely due to the fact that the stronger dependency structure among the data may diminish the useful information contained in the kernel matrix $\mathbf{K}$. As a consequence, a smaller number of RFs are needed for retaining the information in $\mathbf{K}$.

5. Numerical Experiments

In this section, we validate our theoretical findings through extensive numerical experiments on both synthetic and real-life examples. In all the experiments, the RKHS $\mathcal{H}_K$ is induced by the Gaussian kernel $K(\mathbf{x},\mathbf{x}') = \exp(-\|\mathbf{x}-\mathbf{x}'\|^2)$. And its corresponding random feature is $\psi(\mathbf{x},\boldsymbol{\omega}) = \sqrt{2}\cos(\mathbf{x}^\top\mathbf{w}+b)$ with $\mathbf{w} \sim N(0, 2\mathbf{I}_d)$ and $b \sim \text{Uniform}(0, 2\pi)$. The parameter λ is chosen from $\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ via cross-validation, and the performance of the estimator $\widehat{f}$ is evaluated by the prediction error that $\|\widehat{f} - f_\rho\|_m = \sqrt{\frac{1}{m}\sum_{i=1}^m(\widehat{f}(\mathbf{x}_i) - f_\rho(\mathbf{x}_i))^2}$ using a new test data of size m drawn from the specified model. The Python code for reproducing the numerical experiments is available in `https://github.com/wangchao-afk/KRR-RF-DP`.

5.1. Synthetic Data Analysis

The following two data generating schemes are considered, involving a nonparametric time series model and a Markov chain.

Example 5.1 (**Nonparametric time series**).

$$X_i = f_\rho(X_{i-1}) + \varepsilon_i,$$

where $f_\rho(\mathbf{x}) = 0.8\sin(\pi\mathbf{x})$ and $\{\varepsilon_i\}_{i\geq 1}$ is an i.i.d. noise sequence with $\varepsilon_i \sim \text{Uniform}(-0.6, 0.6)$. Note that the above stochastic process $\{\boldsymbol{Z}_i\}_{i\geq 1}$ with $\boldsymbol{Z}_i = (X_{i-1}, X_i)$ is a α-mixing process hence a τ-mixing process (Dedecker & Prieur, 2004; 2005).

Example 5.2 (**Markov chain**).

$$X_i = 0.5(X_{i-1} + \varepsilon_i),$$

where $P(\varepsilon_i = -1) = P(\varepsilon_i = 1) = 0.5$. Note that this case is shown to be τ-mixing but not α-mixing (Dedecker & Prieur, 2004; 2005).

For both examples, we sequentially generate $(n+m)$ samples and use the first n samples as the training data and the remaining m samples as the test data. Here we fix $n = 2m$ and the generating scheme for each example is replicated 100 times in all the settings.

We first investigate how the training sample size n influences the numerical performance of KRR-RF in both examples. Specifically, we vary n within $\{1000, 2000, 3000, 5000, 7500, 10000\}$ and also vary M

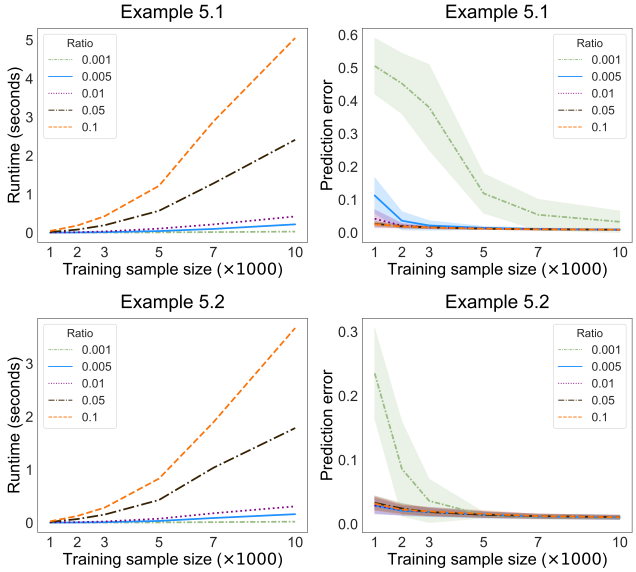

Figure 1. The left panel reports the averaged runtimes vs the training sample size n; the right panel reports the averaged prediction error vs the training sample size n.

such that the ratio of the training sample size to the number of RFs, M/n, varies within $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. The obtained numerical results are shown in Figure 1.

As shown in the right panels of Figure 1, as n increases, the prediction errors decrease at first and then become stable, which is aligned with our theoretical findings. We also observe that for the small ratios 0.001 and 0.005, the running time is reduced, but the performance is significantly undermined when n is not relatively large, especially in Example 5.1. Moreover, the error curves of the ratio 0.01 are comparable to those of the ratios 0.05 and 0.1, while its running time is much shorter. This observation is especially noticeable when the training sample size n is relatively large, suggesting to choose ratio as 0.01.

We also investigate the effect of M on the numerical performance of KRR-RF by varying the ratios of the training sample size to the number of RFs and fixing the training sample size $n = 4000$. The results are shown in Figure 2. It is clear that the consumed time increases with the growth of the ratio, aligning with the theoretical prediction that the computational complexity monotonically grows as M increases. Conversely, the prediction performance improves as the ratio increases and eventually becomes stable. The results also suggest an optimal number of RFs could be around $\exp(-5) \times 4000 \approx 27$ for Example 5.1, and $\exp(-6.5) \times 4000 \approx 6$ for Example 5.2, which leads to low-level computational cost while preserving comparable prediction accuracy to that achieved by using more RFs.

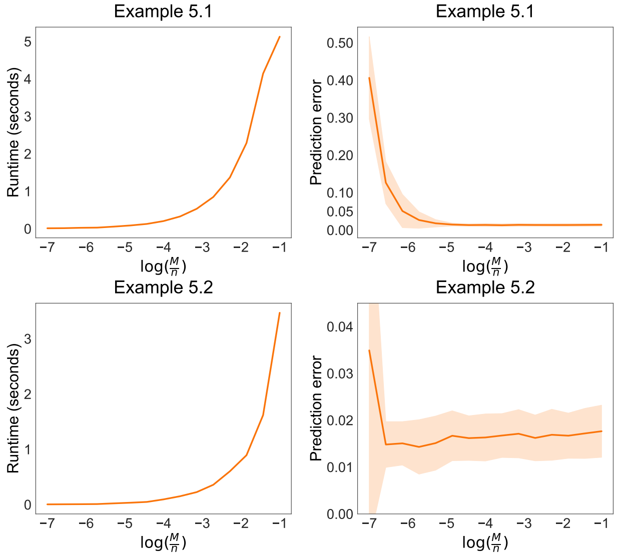

Figure 2. The left panel reports the averaged runtime vs the log ratio $\log(M/n)$; the right panel reports the averaged prediction error vs the log ratio $\log(M/n)$.

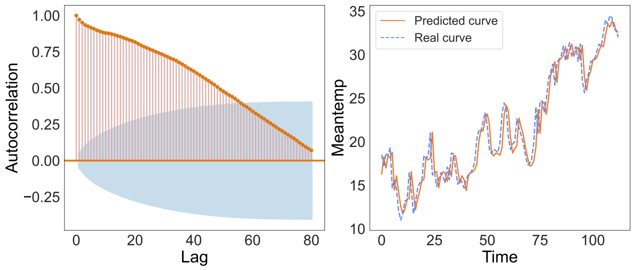

Figure 3. The left panel reports the autocorrelation vs the lag time, the shadow zone represents the confidence interval with level 0.05; the right panel exhibits the real curve and the predicted curve, respectively, where we use $M = 30$.

5.2. Real Data Analysis

In this part, we apply KRR-RF to a daily climate time series data.[2] This dataset contains climate data from 1st January 2013 to 24th April 2017 for Delhi, India, and includes four features: mean temperature (Meantemp), humidity value, wind speed and mean pressure. It can be regarded as a baseline for understanding long-term climate dynamics in urban settings by offering a detailed and extended record of climate variables. We apply KRR-RF to this data for predicting mean temperature. Specifically, we chose the first 1462 samples from 2013 to 2016 as the training data and the remaining 114 samples in 2017 as the test data. The obtained numerical results are shown in Figure 3.

[2] This dataset is available in https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data/data.

From the left panel of Figure 3, we can observe the mean temperature is strongly correlated at nearby moments and this dependency becomes weaker as the lag time increases. The right panel in Figure 3 exhibits a close alignment between the predicted curve and the real curve, suggesting that KRR-RF can effectively model the dependent data in practical scenarios and achieve accurate predictions, meanwhile enjoying computational efficiency.

6. Discussions and Conclusion

6.1. Comparison to Rudi & Rosasco (2017)

Rudi & Rosasco (2017) establishes a theoretical foundation for the KRR-RF method under the i.i.d. case. In contrast, our work aims to understand the behavior of KRR-RF under a more challenging scenario when the data are dependent. Clearly, there exist many significant technical differences between our work and the work by Rudi & Rosasco (2017), and some of them are summarized as follows.

(i) Due to the dependent structure of the τ-mixing process, the theoretical tools used in Rudi & Rosasco (2017), such as the classical empirical process theory or the concentration of measures, can not be used in our technical treatment. Instead, we use different treatment by leveraging recently developed concentration inequality tailored to the τ-mixing (Blanchard & Zadorozhnyi, 2019). We want to emphasize that this difference necessitates a more fine-grained and different theoretical treatment, rather than simply substituting theoretical tools. For instance, in the proof of Lemmas C.2 and D.1, to apply the concentration inequality for the τ-mixing process, the key step is to verify the τ-mixing condition for the random processes $\{\zeta(\boldsymbol{X}_i)\}_{i\geq 1}$ and $\{\xi(\boldsymbol{X}_i, Y_i)\}_{i\geq 1}$.

(ii) As discussed earlier, maintaining the mixing property in our technical analysis requires $\boldsymbol{\phi}_M(\cdot)$ to be Lipschitz continuous. However, a theoretical challenge arises from the fact that the RFs are randomly generated. To overcome this challenge, we derive a sufficient condition in Assumption 4.1, which requires the tail behavior of the partial derivative of the random feature mapping to be sub-exponential. In Lemma C.1, we proved that the random mapping $\psi(\mathbf{x}, \boldsymbol{\omega})$ has uniformly bounded partial derivative with probability at least $1-\delta$ as long as the number of RFs is large enough that $M \geq C\log(2/\delta)$. Note that Assumption 4.1 is satisfied by many widely used kernels, including the linear kernel and the Gaussian kernel. Moreover, this assumption may be further relaxed by directly assuming the Lipschitz continuity of the M-dimensional random feature mapping $\boldsymbol{\phi}_M(\cdot)$. More detailed discussions are presented below in Assumption 4.1. Despite such complication of dealing with dependent data, the required number of RFs for achieving the minimax optimality is the same as that in Rudi & Rosasco (2017).

6.2. Conclusion and future work

This paper focuses on the theoretical understanding of nonparametric learning with random features under the large-scale dependent data setting. Our analysis shed light on the understanding of how the learning rates with random features depend on the degree of dependency among data. Extensive experiments on both synthetic and real-world data further support the applicability of the nonparametric method with random features in complex practical scenarios. The main contribution of this work is to bridge the gap between theoretical understanding and empirical advantages of learning large-scale dependent data via random features.

Our studies also raise an intriguing question: *is the slower rate of KRR-RF under the polynomial τ-mixing process case due to the technical limitation or an inherent limitation arising from the stronger dependency within data?* To answer this question, some advanced tools are required to establish the minimax lower bound under this case and we leave this for future research.

Acknowledgements

Wang's research is supported by the Fundamental Research Funds for the Central Universities. Bing's research is supported in part by a discovery grant from the Natural Sciences and Engineering Research Council of Canada. Xin He's research is supported by NSFC-11901375, Shanghai Science and Technology Development Funds (23JC1402100), Program for Innovative Research Team of Shanghai University of Finance and Economics and Shanghai Research Center for Data Science and Decision Technology.

Impact Statement

This paper presents work whose goal is to fill the gaps in the literature on kernel-based methods, which is an important area of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Avron, H., Clarkson, K. L., and Woodruff, D. P. Faster kernel ridge regression using sketching and preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 38 (4):1116–1138, 2017.

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

Bhatia, R. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

Birman, M. S. and Solomyak, M. Z. Piecewise-polynomial approximations of functions of the classes w_p^α. *Matematicheskii Sbornik*, 115(3):331–355, 1967.

Blanchard, G. and Zadorozhnyi, O. Concentration of weakly dependent Banach-valued sums and applications to statistical learning methods. *Bernoulli*, 25(4B):3421–3458, 2019.

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

Dedecker, J. and Prieur, C. Coupling for τ-dependent sequences and applications. *Journal of Theoretical Probability*, 17(4):861–885, 2004.

Dedecker, J. and Prieur, C. New dependence coefficients. Examples and applications to statistics. *Probability Theory and Related Fields*, 132:203–236, 2005.

Fu, T. C. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

Guo, Z.-C., Lin, S.-B., and Zhou, D.-X. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7): 074009, 2017.

Kimeldorf, G. and Wahba, G. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

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.

Liu, J. and Lian, H. On optimal learning with random features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):9536–9541, 2023.

Ma, C., Pathak, R., and Wainwright, M. J. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

Maume-Deschamps, V. Exponential inequalities and functional estimations for weak dependent data: applications to dynamical systems. *Stochastics and Dynamics*, 6(04): 535–560, 2006.

Modha, D. S. and Masry, E. Minimum complexity regression estimation with weakly dependent observations. *IEEE Transactions on Information Theory*, 42(6):2133–2145, 1996.

Papaspiliopoulos, O. High-dimensional probability: An introduction with applications in data science. *Quantitative Finance*, 20(10):1591–1594, 2020.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20:1177–1184, 2007.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in Neural Information Processing Systems*, 21:1313–1320, 2008.

Ralanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., and Das, G. Mining time series data. *Data Mining and Knowledge Discovery Handbook*, pp. 1069–1103, 2005.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. *Advances in Neural Information Processing Systems*, 30:3218–3228, 2017.

Rudi, A., Camoriano, R., and Rosasco, L. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28:1657–1665, 2015.

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

Smale, S. and Zhou, D.-X. Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302, 2005.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Science & Business Media, 2008.

Sun, Y., Gilbert, A., and Tewari, A. But how does it work in theory? Linear SVM with random features. *Advances in Neural Information Processing Systems*, 31:3383–3392, 2018.

Sun, Z. and Lin, S.-B. Distributed learning with dependent samples. *IEEE Transactions on Information Theory*, 68 (9):6003–6020, 2022.

Sun, Z., Dai, M., Wang, Y., and Lin, S.-B. Nyström regularization for time series forecasting. *The Journal of Machine Learning Research*, 23(1):14082–14123, 2022.

Vapnik, V. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 1999.

Wahba, G. Spline models for observational data. In *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, 1990.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Yang, Y., Pilanci, M., and Wainwright, M. J. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991 – 1023, 2017.

Yu, B. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

Appendix

This appendix is organized as follows. In Section A, we provide more numerical studies. In Section B, we introduce the definitions for the operators, along with an error decomposition. Moving to Section C, we prove some technical lemmas needed for the proofs of the main theorems. In Section D, we derive upper bounds separately for three components of error and then present the proof of the main theorems. Section E is devoted to listing auxiliary lemmas that involve two concentration inequalities utilized in our proofs. Section F provides more details of the data-dependent sampling strategy.

A. Additional numerical experiments

We consider a nonparametric time series model as follows. The setup of the numerical experiments is the same as those in Section 5 of the main text, including kernel selection, random feature mapping, evaluation standard, and tuning procedure for λ.

Example A.1.

$$X_i = 0.8\sin(\pi(0.7X_{i-1} + 0.2X_{i-2} + 0.1X_{i-3})) + \varepsilon_i,$$

where $\{\varepsilon_i\}_{i\geq 1}$ is an i.i.d. noise sequence with $\varepsilon_i \sim \text{Uniform}(-0.6, 0.6)$.

We first investigate the numerical performance of KRR-RF by varying n within $\{1000, 2000, 3000, 5000, 7500, 10000\}$ and varying the ratio of the training sample size to the number of RFs, M/n, within $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. The averaged numerical performance is illustrated in Figure 4.

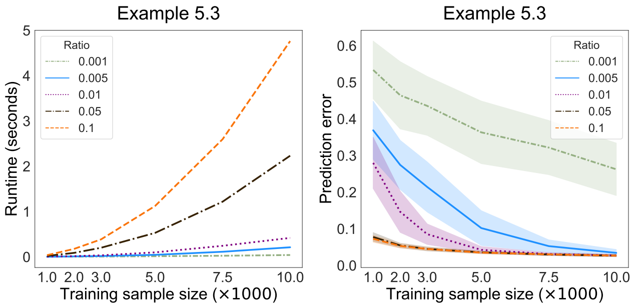

Figure 4. The left panel reports the averaged runtimes vs the training sample size n; the right panel reports the averaged prediction error vs the training sample size n.

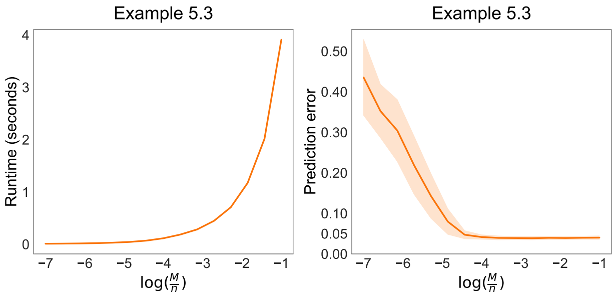

Figure 5. The left panel reports the averaged runtimes vs the log ratio $\log(M/n)$; the right panel reports the averaged prediction error vs the log ratio $\log(M/n)$.

It is clear from Figure 4 that the obtained results are comparable to those reported in the main text. It can be seen that the prediction error curves exhibit roughly the same decline trends as n grows for the ratios in $\{0.05, 0.1\}$ and has significant improvement over that for other ratios. Whereas, the consumed time for the ratio 0.1 is significantly higher than for the ratio 0.05. This observation suggests the best choice of ratio is 0.05. The second experiment is designed to investigate the effect of M on the numerical performance of KRR-RF by varying the ratios and fixing the training sample size $n = 4000$. The experiment result is reported in Figure 5, which shows that the choice of $M = 4000 \exp(-4.5) \approx 44$ leads to a low-level computational cost while preserving an almost optimal learning efficiency.

B. Operators and Error Decomposition

Notation. For an operator A, we use $A^\top$ to represent its adjoint operator, use I to denote the identity operator, and use $\boldsymbol{\nu}^\top$ to represent the transport of a vector $\boldsymbol{\nu}$. We denote the inner product endowed with Euclidean space as $\langle \cdot, \cdot \rangle_2$ and the ℓ_2-vector norm as $\| \cdot \|_2$. Note that the operators involved in the technical proof may be defined on different domains and ranges. Specifically, we use $\| \cdot \|_{\mathcal{L},\mathcal{L}}$, $\| \cdot \|_{\ell_2,\ell_2}$ and $\| \cdot \|_{\ell_2,\mathcal{L}}$ to denote the operator norm for the operator from $\mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$ to $\mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$, from $\mathcal{R}^M$ to $\mathcal{R}^M$, from $\mathcal{R}^M$ to $\mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$, respectively. Moreover, we use $\| \cdot \|$ to denote the operator norm when not specifying any particular domain and range. Throughout the rest of this paper, we use $E_{\boldsymbol{\omega}}[\cdot]$ (and $P_{\boldsymbol{\omega}}[\cdot]$) to denote expectation (and probability) taken over the random feature $\boldsymbol{\omega}$ or its i.i.d. copies $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M$. Otherwise, we use $E[\cdot]$ to denote expectation taken over all random quantities while conditioning on $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M$.

Operator representations. Given the random features $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M$ sampled from the distribution π, recall that

$$\boldsymbol{\phi}_M(\mathbf{x}) = \frac{1}{\sqrt{M}} (\psi(\mathbf{x}, \boldsymbol{\omega}_1), ..., \psi(\mathbf{x}, \boldsymbol{\omega}_M))^\top \in \mathcal{R}^M,$$

$$\widehat{\mathbf{S}}_M = \frac{1}{\sqrt{n}} (\boldsymbol{\phi}_M(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_M(\mathbf{x}_n))^\top \in \mathcal{R}^{n \times M}.$$

Then, the KRR-RF estimator has a closed form in (6) that

$$\widehat{f}_{M,\lambda}(\mathbf{x}) = \boldsymbol{\phi}_M(\mathbf{x})^\top (\widehat{\mathbf{S}}_M^\top \widehat{\mathbf{S}}_M + \lambda \mathbf{I}_M)^{-1} \widehat{\mathbf{S}}_M^\top \mathbf{y}, \quad \forall\ \mathbf{x} \in \mathcal{X},$$

with $\mathbf{y} = \frac{1}{\sqrt{n}}(y_1, \ldots, y_n)^\top \in \mathcal{R}^n$. Moreover, we define the M-dimensional function space induced by $\boldsymbol{\phi}_M(\mathbf{x})$ as

$$\mathcal{H}_M := \left\{ f : f(\mathbf{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}_M(\mathbf{x}), \boldsymbol{\alpha} \in \mathcal{R}^M \right\}.$$

Note that if we equip the inner product in $\mathcal{H}_M$ as $\langle f, g \rangle_{K_M} = \boldsymbol{\alpha}^\top \boldsymbol{\beta}$ for $f(\cdot) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}_M(\cdot), g(\cdot) = \boldsymbol{\beta}^\top \boldsymbol{\phi}_M(\cdot)$, then $\mathcal{H}_M$ is a RKHS associated with kernel $K_M(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}_M(\mathbf{x}), \boldsymbol{\phi}_M(\mathbf{x}') \rangle_2$. Denote $K_M(\cdot, \mathbf{x}) = \boldsymbol{\phi}_M(\mathbf{x})^\top \boldsymbol{\phi}_M(\cdot) \in \mathcal{H}_M$ which can also be considered as an operator from $\mathcal{X}$ to $\mathcal{H}_M$, and the above statement can be verified by the fact that for any $f(\cdot) = \boldsymbol{\alpha}^\top \boldsymbol{\phi}_M(\cdot)$, we have

$$\langle f, K_M(\cdot, \mathbf{x}) \rangle_{K_M} = \boldsymbol{\alpha}^\top \boldsymbol{\phi}_M(\mathbf{x}) = f(\mathbf{x}),$$

and then, the reproducing property holds. Note that for the kernel function K, we have $\sup_{\mathbf{x},\mathbf{x}'} |K(\mathbf{x}, \mathbf{x}')| \leq \kappa^2$, and this property also holds for K_M, since $\sup_{\mathbf{x},\mathbf{x}'} |K_M(\mathbf{x}, \mathbf{x}')| \leq \kappa^2$ due to the assumption that $|\psi(\mathbf{x}, \boldsymbol{\omega})| \leq \kappa$. It is worthy pointing out that $\mathcal{H}_M$ can be viewed as the finite-dimensional approximation of $\mathcal{H}_K$.

We introduce some useful operators that are commonly used in literature (Smale & Zhou, 2007; Rudi et al., 2015; Rudi & Rosasco, 2017). For any $g \in \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}})$, $\boldsymbol{\alpha} \in \mathcal{R}^M$ and $\boldsymbol{\beta} \in \mathcal{R}^n$, we define the data-free operators as

- $S_M : \mathcal{R}^M \to \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}}), \quad (S_M \boldsymbol{\beta})(\cdot) = \boldsymbol{\phi}_M(\cdot)^\top \boldsymbol{\beta};$
- $S_M^\top : \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}}) \to \mathcal{R}^M, \quad \left(S_M^\top g\right)_i = \frac{1}{\sqrt{M}} \int_{\mathcal{X}} \psi(\mathbf{x}, \boldsymbol{\omega}_i) g(\mathbf{x}) d\rho_{\boldsymbol{X}}(\mathbf{x}), \quad \text{for } \ i = 1, \ldots, M;$
- $L_M : \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}}) \to \mathcal{L}(\mathcal{X}, \rho_{\boldsymbol{X}}), \quad L_M f = \int_{\mathcal{X}} K_M(\cdot, \mathbf{x}) f(\mathbf{x}) d\rho_{\boldsymbol{X}}(\mathbf{x});$
- $C_M : \mathcal{R}^M \to \mathcal{R}^M, \quad C_M = \int_{\mathcal{X}} \boldsymbol{\phi}_M(\mathbf{x}) \boldsymbol{\phi}_M(\mathbf{x})^\top d\rho_{\boldsymbol{X}}(\mathbf{x}).$

We also define the data-dependent operator as

- $\widehat{S}_M : \mathcal{R}^M \to \mathcal{R}^n, \quad \widehat{S}_M \boldsymbol{\alpha} = \widehat{\mathbf{S}}_M \boldsymbol{\alpha};$

- $\widehat{S}_M^\top : \mathcal{R}^n \to \mathcal{R}^M, \quad \widehat{S}_M^\top \boldsymbol{\beta} = \widehat{\mathbf{S}}_M^\top \boldsymbol{\beta}$;
- $\widehat{C}_M : \mathcal{R}^M \to \mathcal{R}^M, \quad \widehat{C}_M = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}_M(\mathbf{x}_i)\boldsymbol{\phi}_M(\mathbf{x}_i)^\top$.

Note that for any $\mathbf{x} \in \mathcal{X}$, the matrix $\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top$ can be viewed as an operator $\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top : \mathcal{R}^M \to \mathcal{R}^M$ that maps $\boldsymbol{\alpha} \in \mathcal{R}^M$ to $\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top \boldsymbol{\alpha} \in \mathcal{R}^M$. It can be verified that

$$C_M = S_M^\top S_M, \quad \widehat{C}_M = \widehat{S}_M^\top \widehat{S}_M, \quad L_M = S_M S_M^\top,$$

and all these operators are all self-adjoint and positive operators (Rudi & Rosasco, 2017). For ease of presentation and henceforth, the statements that condition on the random features $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M$, such as $C_M = S_M^\top S_M$ above, should be understood to hold almost surely.

Clearly, using the introduced operators, $\widehat{f}_{M,\lambda}$ can be rewritten as

$$\widehat{f}_{M,\lambda} = S_M(\widehat{C}_M + \lambda I)^{-1}\widehat{S}_M^\top \mathbf{y}\,.$$

We also define two intermediate functions as

$$\widetilde{f}_{M,\lambda} = S_M(\widehat{C}_M + \lambda I)^{-1} S_M^\top f_\rho$$

and

$$\bar{f}_{M,\lambda} = L_M(L_M + \lambda)^{-1} f_\rho.$$

For any $\lambda > 0$, define $\mathcal{N}_M(\lambda)$ as

$$\mathcal{N}_M(\lambda) = \mathrm{Tr}((L_M + \lambda I)^{-1} L_M).$$

It is worthy pointing out that in the following proofs, we primarily rely on $\mathcal{N}_M(\lambda)$ rather than $\mathcal{N}(\lambda)$. However, if $\lambda \le \|L_K\|_{\mathcal{L},\mathcal{L}}$, it can be proved that these two quantities are equivalent under the requirement on the number of random features that $M \ge (4 + 18\mathcal{N}_\infty(\lambda)) \log \frac{12\kappa}{\lambda\delta}$ (Proposition 10 in Rudi & Rosasco (2017)), i.e.,

$$\frac{1}{2}\mathcal{N}(\lambda) \le \mathcal{N}_M(\lambda) \le 3\mathcal{N}(\lambda)$$

with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Error decomposition. By using triangle inequality, we can decompose the total error $\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho$ into three terms that

$$\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho \le \underbrace{\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|_\rho}_{\text{Sample variance}} + \underbrace{\|\widetilde{f}_{M,\lambda} - \bar{f}_{M,\lambda}\|_\rho}_{\text{Empirical error}} + \underbrace{\|\bar{f}_{M,\lambda} - f_\rho\|_\rho}_{\text{Approximation error}}. \tag{9}$$

The above decomposition typically identifies three components of error: Sample variance, empirical error, and approximation error. Sample variance is attributed to the noise in responses, while empirical error controls the discrepancy between the empirical covariance operator $\widehat{C}$ and the population integral operator L_M. Additionally, the approximation error arises from the penalty. All these three components depend on the regularization parameter λ. To be more specific, a larger λ value may lead to reduced sample variance and empirical error but simultaneously result in an increased approximation error. Consequently, an optimal selection of λ should be chosen to balance the trade-off among these three components.

C. Technical Lemmas

In this section, we provide some technical lemmas that are used to complete the proof of the main results in Section D. Lemma C.1 is devoted to verifying the uniform boundedness of the partial derivative of the random feature mapping $\psi(\cdot, \boldsymbol{\omega})$ with a high probability if the number of RFs is sufficiently large. Under some event, we can verify the Lipschitz continuity of the M-dimensional random feature mapping $\boldsymbol{\phi}_M(\cdot)$. Lemmas C.2 and C.3 aim to bound the similarity between $C_M + \lambda I$ and $\widehat{C}_M + \lambda I$ in different manners.

For notation simplicity, we write

$$L_{M,\lambda} = L_M + \lambda I, \qquad L_{K,\lambda} = L_K + \lambda I, \qquad C_{M,\lambda} = C_M + \lambda I, \qquad \widehat{C}_{M,\lambda} = \widehat{C}_M + \lambda I.$$

Recall from Assumption 4.1 that

$$T(\boldsymbol{\omega}) = \sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{d}\left\|\partial_{\mathbf{x}}\psi(\mathbf{x},\boldsymbol{\omega})\right\|_2^2, \quad \forall\, \boldsymbol{\omega}\in\Omega. \tag{10}$$

Define the event $\mathcal{E}_{\boldsymbol{\omega}}$ that

$$\mathcal{E}_{\boldsymbol{\omega}} = \left\{\frac{1}{M}\sum_{i=1}^{M} T(\boldsymbol{\omega}_i) \le B\right\},$$

for some positive constant B to be clarified in the proof. The following lemma shows that by choosing a sufficiently large number of RFs, the event $\mathcal{E}_{\boldsymbol{\omega}}$ holds with high probability.

Lemma C.1. *For any $\delta\in(0,1)$, under Assumption 4.1, if $M\ge C\log(2/\delta)$, we have*

$$P(\mathcal{E}_{\boldsymbol{\omega}}) \ge 1-\delta.$$

Moreover, on the event $\mathcal{E}(\boldsymbol{\omega})$,

$$\|\boldsymbol{\phi}_M(\mathbf{x}) - \boldsymbol{\phi}_M(\mathbf{x}')\|_2 \le \sqrt{dB}\|\mathbf{x}-\mathbf{x}'\|_2, \quad \forall\ \mathbf{x},\mathbf{x}'\in\mathcal{X}. \tag{11}$$

Proof. Recall that $T(\boldsymbol{\omega})$ from (10). Write $\bar{T} = \frac{1}{M}\sum_{i=1}^{M} T(\boldsymbol{\omega}_i)$. We first establish the proof for the first part of Lemma C.1 in two steps.

Step 1: By Lemma E.3, Assumption 4.1 is equivalent to

$$E_{\boldsymbol{\omega}}[\exp(\sigma(T(\boldsymbol{\omega}) - E_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})]))] \le \exp\Big(\frac{\sigma^2 C_0^2}{2}\Big), \quad \text{for all } |\sigma| \le \frac{1}{C_1}, \tag{12}$$

where C_0, C_1 are some positive constants. Then, for $|\sigma|\le\frac{1}{C_1}$, the moment-generating function of $M\bar{T} - E_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})]$ can be bounded by

$$\begin{aligned} E_{\boldsymbol{\omega}}[\exp(\sigma(M\bar{T} - ME_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})]))] &= \prod_{i=1}^{M} E_{\boldsymbol{\omega}}[\exp(\sigma(T(\boldsymbol{\omega}_i) - E_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})]))] \\ &\le \prod_{i=1}^{M} \exp\Big(\frac{\sigma^2 C_0^2}{2}\Big) = \exp\Big(\frac{M\sigma^2 C_0^2}{2}\Big), \end{aligned}$$

where the first equality follows from the fact that $\boldsymbol{\omega}_1, ..., \boldsymbol{\omega}_M$ are independent and the inequality follows from (12). This proves that $M\bar{T} - E_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})]$ is sub-exponential with parameter $(\sqrt{M}C_0, C_1)$.

Step 2: By the Bernstein inequality stated in Lemma E.4, we have

$$P_{\boldsymbol{\omega}}\left[|M\bar{T} - ME_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})]| \ge t\right] \le \begin{cases} 2\exp(-\frac{t^2}{2MC_0^2}) & \text{if } 0\le t\le \frac{MC_0^2}{C_1}, \\ 2\exp(-\frac{t}{2C_1}) & \text{for } t > \frac{MC_0^2}{C_1}. \end{cases}$$

By plugging $t = MC_0^2/C_1$ into the above inequality, we have

$$P_{\boldsymbol{\omega}}\Big(\big|\bar{T} - E_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})]\big| \ge \frac{C_0^2}{C_1}\Big) \le 2\exp\Big(-\frac{MC_0^2}{2C_1^2}\Big).$$

Therefore, if we take $M\ge C\log(2/\delta)$, with probability at least $1-\delta$, there holds

$$\bar{T} \le E_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})] + \frac{C_0^2}{C_1} \le \frac{a_0}{a_1} + \frac{C_0^2}{C_1} = B,$$

as desired. The second inequality follows from

$$E_{\boldsymbol{\omega}}[T(\boldsymbol{\omega})] = \int_0^\infty P(T(\boldsymbol{\omega}) > t)dt \le \int_0^\infty a_0\exp(-a_1 t)dt = \frac{a_0}{a_1}.$$

We proceed to prove the second part of Lemma C.1. By definition,

$$\begin{aligned}
\|\boldsymbol{\phi}_M(\mathbf{x}) - \boldsymbol{\phi}_M(\mathbf{x}')\|_2^2 &= \frac{1}{M}\sum_{i=1}^{M}\left(\psi(\mathbf{x},\boldsymbol{\omega}_i) - \psi(\mathbf{x}',\boldsymbol{\omega}_i)\right)^2 \\
&= \frac{1}{M}\sum_{i=1}^{M}\left(\int_0^1 (\mathbf{x} - \mathbf{x}')^\top \partial_{\mathbf{x}}\psi(\mathbf{x}_u,\boldsymbol{\omega}_i)du\right)^2 && \text{by } \mathbf{x}_u = u\,\mathbf{x} + (1-u)\mathbf{x}' \\
&\leq \frac{1}{M}\sum_{i=1}^{M}\sup_{u\in[0,1]}\|\mathbf{x} - \mathbf{x}'\|_2^2\,\|\partial_{\mathbf{x}}\psi(\mathbf{x}_u,\boldsymbol{\omega}_i)\|_2^2 && \text{by Cauchy Schwarz inequality} \\
&\leq d\,\|\mathbf{x} - \mathbf{x}'\|_2^2\,\frac{1}{M}\sum_{i=1}^{M}T(\boldsymbol{\omega}_i) && \text{by (10)}
\end{aligned}$$

implying the claim by invoking $\mathcal{E}_{\boldsymbol{\omega}}$, where $\mathbf{x}_u \in \mathcal{X}$ due to convexity of $\mathcal{X}$. This completes the proof of Lemma C.1. □

Recall $\tau(k)$ from Definition 3.1, and we define

$$\begin{aligned}
\ell_1 &= \left(\max\left\{1, \log(\kappa^{-1}b_0b_1n\sqrt{dB})\right\}\right)^{-1/\gamma_0}\frac{b_1n}{2}, && \text{if} \quad \tau(k)\leq b_0\exp(-(b_1k)^{\gamma_0}); \\
\ell_1 &= \left(\frac{\sqrt{\lambda\mathcal{N}_M(\lambda)}}{2b_2\sqrt{dB}}\right)^{\frac{2}{2\gamma_1+1}}(\frac{n}{2})^{\frac{2\gamma_1}{2\gamma_1+1}}, && \text{if} \quad \tau(k)\leq b_2k^{-\gamma_1}.
\end{aligned} \tag{13}$$

The following lemma controls the operator norm of $C_{M,\lambda}^{-1/2}(\widehat{C}_M - C_M)$.

Lemma C.2. *For any $\delta\in(0,1)$ and on the event $\mathcal{E}_{\boldsymbol{\omega}}$, with probability at least $1-\delta/5$, there holds*

$$\left\|C_{M,\lambda}^{-1/2}(\widehat{C}_M - C_M)\right\|_{\ell_2,\ell_2} \leq 21\Big(\frac{2\kappa^2}{\ell_1\sqrt{\lambda}} + \sqrt{\frac{\kappa^2\mathcal{N}_M(\lambda)}{\ell_1}}\Big)\log\frac{10}{\delta}. \tag{14}$$

Proof. To start with, we first define the operator $\zeta(\mathbf{x})$ from $\mathcal{R}^M$ to $\mathcal{R}^M$ as

$$\zeta(\mathbf{x}) = C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M).$$

It can be verified that $E[\zeta(\boldsymbol{X})] = 0$ by the definition of C_M and

$$C_{M,\lambda}^{-1/2}(\widehat{C}_M - C_M) = \frac{1}{n}\sum_{i=1}^{n}\zeta(\mathbf{x}_i),$$

by the definition of $\widehat{C}_M$. Now, we aim to bound its operator norm by utilizing the concentration inequality in Lemma E.1. Specifically, we have

$$\begin{aligned}
\|\zeta(\mathbf{x})\|_{\ell_2,\ell_2} = \left\|C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M)\right\|_{\ell_2,\ell_2} &\leq \left\|C_{M,\lambda}^{-1/2}\right\|_{\ell_2,\ell_2}\left\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M\right\|_{\ell_2,\ell_2} \\
&\leq \lambda^{-1/2}\left\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M\right\|_{\ell_2,\ell_2} \\
&\leq \lambda^{-1/2}\left(\left\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\right\|_{\ell_2,\ell_2} + \left\|C_M\right\|_{\ell_2,\ell_2}\right),
\end{aligned}$$

where the second inequality follows from the fact that $\|C_{M,\lambda}^{-1/2}\|_{\ell_2,\ell_2}\leq\lambda^{-1/2}$. To bound $\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\|_{\ell_2,\ell_2}$, we have

$$\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\|_{\ell_2,\ell_2} = \|\boldsymbol{\phi}_M(\mathbf{x})\|_2^2 = \frac{1}{M}\sum_{i=1}^{M}(\psi(\mathbf{x},\boldsymbol{\omega}_i))^2 \leq \sup_{1\leq i\leq M}|\psi(\mathbf{x},\boldsymbol{\omega}_i)|^2 \leq \kappa^2. \tag{15}$$

Moreover, by using Jensen's inequality, we have

$$\|C_M\|_{\ell_2,\ell_2} = \Big\| \int_{\mathcal{X}} \boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top d\rho_{\boldsymbol{X}}(\mathbf{x})\Big\|_{\ell_2,\ell_2} \le \int_{\mathcal{X}} \big\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big\|_{\ell_2,\ell_2} d\rho_{\boldsymbol{X}}(\mathbf{x}) \le \kappa^2.$$

Combining the above results, we have $\|\zeta(\mathbf{x})\|_{\ell_2,\ell_2} \le 2\kappa^2\lambda^{-1/2}$.

To verify the second bounded condition required in Lemma E.1, we have

$$\begin{aligned}
E\big[\|\zeta(\boldsymbol{X})\|^2_{\ell_2,\ell_2}\big] &= \int_{\mathcal{X}} \big\|C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M)\big\|^2_{\ell_2,\ell_2} d\rho_{\boldsymbol{X}}(\mathbf{x}) \\
&\overset{(i)}{\le} \int_{\mathcal{X}} \mathrm{Tr}\big((\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M)C_{M,\lambda}^{-1}(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M)\big) d\rho_{\boldsymbol{X}}(\mathbf{x}) \\
&\overset{(ii)}{\le} \int_{\mathcal{X}} \mathrm{Tr}\big(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top C_{M,\lambda}^{-1}\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big) d\rho_{\boldsymbol{X}}(\mathbf{x}) \\
&\overset{(iii)}{\le} \int_{\mathcal{X}} \big\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big\|_{\ell_2,\ell_2} \mathrm{Tr}\big(C_{M,\lambda}^{-1}\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big) d\rho_{\boldsymbol{X}}(\mathbf{x}) \\
&\overset{(iv)}{\le} \kappa^2 \int_{\mathcal{X}} \mathrm{Tr}\big(C_{M,\lambda}^{-1}\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big) d\rho_{\boldsymbol{X}}(\mathbf{x}) \\
&\overset{(v)}{\le} \kappa^2 \mathcal{N}_M(\lambda),
\end{aligned}$$

where (i) follows from $\|A\|^2 \le \mathrm{Tr}(A^\top A)$ for any trace class operator A, (ii) follows from the fact that

$$\begin{aligned}
&\int_{\mathcal{X}} \mathrm{Tr}\big((\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M)C_{M,\lambda}^{-1}(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - C_M)\big) d\rho_{\boldsymbol{X}}(\mathbf{x}) \\
&= \int_{\mathcal{X}} \mathrm{Tr}\big(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top C_{M,\lambda}^{-1}\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big) d\rho_{\boldsymbol{X}}(\mathbf{x}) - \mathrm{Tr}\big(C_M C_{M,\lambda}^{-1} C_M\big),
\end{aligned}$$

(iii) follows from Von-Neumann's trace inequality that $\mathrm{Tr}(AB) \le \|A\|\mathrm{Tr}(B)$ valid for the trace class operators A and B, (iv) follows from (15), and (v) follows from

$$\begin{aligned}
\int_{\mathcal{X}} \mathrm{Tr}\big(C_{M,\lambda}^{-1}\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big) d\rho_{\boldsymbol{X}}(\mathbf{x}) &= \mathrm{Tr}\Big(C_{M,\lambda}^{-1}\int_{\mathcal{X}} \boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top d\rho_{\boldsymbol{X}}(\mathbf{x})\Big) \\
&= \mathrm{Tr}\big(C_{M,\lambda}^{-1}C_M\big) =: J_M(\lambda)
\end{aligned} \tag{16}$$

and the fact that

$$J_M(\lambda) = \mathrm{Tr}(S_M^\top L_{M,\lambda}^{-1} S_M) = \mathrm{Tr}(L_{M,\lambda}^{-1} L_M) = \mathcal{N}_M(\lambda).$$

See Rudi & Rosasco (2017).

To apply Lemma E.1, we additionally need to verify that the stochastic process $\{\zeta(\boldsymbol{X}_i)\}_{i\ge1}$ remains a τ-mixing process. Note that $\{\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)\}_{i\ge1}$ is assumed to be τ-mixing process with coefficient $\tau(k)$, it immediately follows from Proposition 3.2 that $\{\boldsymbol{X}_i\}_{i\ge1}$ is also τ-mixing process with coefficient $\tau(k)$.

To verify the τ-mixing condition for $\{\zeta(\boldsymbol{X}_i)\}_{i\ge1}$, first note that

$$\begin{aligned}
\big\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - \boldsymbol{\phi}_M(\mathbf{x}')\boldsymbol{\phi}_M(\mathbf{x}')^\top\big\|_{\ell_2,\ell_2} &\le (\|\boldsymbol{\phi}_M(\mathbf{x})\|_2 + \|\boldsymbol{\phi}_M(\mathbf{x}')\|_2)\,\|\boldsymbol{\phi}_M(\mathbf{x}) - \boldsymbol{\phi}_M(\mathbf{x}')\|_2 \\
&\le 2\kappa\|\boldsymbol{\phi}_M(\mathbf{x}) - \boldsymbol{\phi}_M(\mathbf{x}')\|_2 && \text{by (15)} \\
&\le 2\kappa\sqrt{dB}\|\,\mathbf{x} - \mathbf{x}'\|_2. && \text{by (11)}
\end{aligned}$$

Together with the fact $\|C_{M,\lambda}^{-1/2}\|_{\ell_2,\ell_2} \le \lambda^{-1/2}$, we thus have

$$\begin{aligned}
\|\zeta(\mathbf{x}) - \zeta(\mathbf{x}')\|_{\ell_2,\ell_2} &= \big\|C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - \boldsymbol{\phi}_M(\mathbf{x}')\boldsymbol{\phi}_M(\mathbf{x}')^\top)\big\|_{\ell_2,\ell_2} \\
&\le \big\|C_{M,\lambda}^{-1/2}\big\|_{\ell_2,\ell_2}\big\|\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top - \boldsymbol{\phi}_M(\mathbf{x}')\boldsymbol{\phi}_M(\mathbf{x}')^\top\big\|_{\ell_2,\ell_2} \\
&\le 2\kappa\sqrt{dB\lambda}\|\,\mathbf{x} - \mathbf{x}'\|_2.
\end{aligned}$$

This verifies the Lipschitz continuity of $\{\zeta(\boldsymbol{X}_i)\}_{i\geq 1}$, thus proving that the process $\{\zeta(\boldsymbol{X}_i)\}_{i\geq 1}$ is τ-mixing with coefficient $2\kappa\sqrt{dB\lambda}\tau(k)$ by using Proposition 3.2.

By applying Lemma E.1 with $L = 2\lambda^{-1/2}\kappa^2$ and $\sigma = \kappa\sqrt{\mathcal{N}_M(\lambda)}$, and the corresponding Hilbert space consists of the bounded linear operators on $\mathcal{R}^M$, with probability at least $1-\delta/5$, the following result holds

$$\left\|C_{M,\lambda}^{-1/2}(\widehat{C}_M - C_M)\right\|_{\ell_2,\ell_2} \leq 21\left(\frac{2\kappa^2}{\ell_1\sqrt{\lambda}} + \sqrt{\frac{\kappa^2\mathcal{N}_M(\lambda)}{\ell_1}}\right)\log\frac{10}{\delta}.$$

This completes the proof of Lemma C.2. □

Lemma C.3. *For any $\delta \in (0,1)$ and if*

$$\ell_1\lambda \geq C\kappa^2 \max\{\mathcal{N}_M(\lambda), 1\}\log^2\frac{10}{\delta},$$

and on the event $\mathcal{E}_{\boldsymbol{\omega}}$, with probability at least $1-\delta/5$, there holds

$$\left\|\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}\right\|_{\ell_2,\ell_2} \leq 2.$$

Proof. Note that

$$\left\|C_{M,\lambda}^{-1}(\widehat{C}_M - C_M)\right\|_{\ell_2,\ell_2} \leq \left\|C_{M,\lambda}^{-1/2}\right\|_{\ell_2,\ell_2}\left\|C_{M,\lambda}^{-1/2}(\widehat{C}_M - C_M)\right\|_{\ell_2,\ell_2} \leq \frac{1}{\sqrt{\lambda}}\left\|C_{M,\lambda}^{-1/2}(\widehat{C}_M - C_M)\right\|_{\ell_2,\ell_2}.$$

Then, by (14), with probability at least $1-\delta/5$, there holds

$$\left\|C_{M,\lambda}^{-1}(\widehat{C}_M - C_M)\right\|_{\ell_2,\ell_2} \leq 21\Big(\frac{2\kappa^2}{\ell_1\lambda} + \sqrt{\frac{\kappa^2\mathcal{N}_M(\lambda)}{\ell_1\lambda}}\Big)\log\frac{10}{\delta}.$$

Moreover, if $\ell_1\lambda \geq C\kappa^2\max\{\mathcal{N}_M(\lambda), 1\}\log^2\frac{10}{\delta}$, we have

$$\left\|C_{M,\lambda}^{-1}(\widehat{C}_M - C_M)\right\|_{\ell_2,\ell_2} \leq \frac{1}{2}.$$

Then, we have

$$\begin{aligned}
\left\|\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}\right\|_{\ell_2,\ell_2} &\overset{(i)}{\leq} 1 + \left\|(\widehat{C}_{M,\lambda}^{-1} - C_{M,\lambda}^{-1})C_{M,\lambda}\right\|_{\ell_2,\ell_2}\\
&\overset{(ii)}{=} 1 + \left\|\widehat{C}_{M,\lambda}^{-1}(C_M - \widehat{C}_M)\right\|_{\ell_2,\ell_2}\\
&= 1 + \left\|\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}C_{M,\lambda}^{-1}(C_M - \widehat{C}_M)\right\|_{\ell_2,\ell_2}\\
&\leq 1 + \left\|\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}\right\|_{\ell_2,\ell_2}\left\|C_{M,\lambda}^{-1}(C_M - \widehat{C}_M)\right\|_{\ell_2,\ell_2}\\
&\leq 1 + \frac{1}{2}\left\|\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}\right\|_{\ell_2,\ell_2},
\end{aligned}$$

where (i) follows from the triangle inequality and (ii) follows from the fact that that $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$ for any invertible bounded operators A and B. Thus, we have

$$\left\|\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}\right\|_{\ell_2,\ell_2} \leq 2.$$

This completes the proof of Lemma C.3. □

D. Proofs of the Main Results

Based on the error decomposition (9), we separately derive the upper bounds for the sample variance in Section D.1, the approximation error in D.2, and the empirical error D.3. Then we provide the proof for the main results in Section D.4. Recall that $|Y| \leq U$ and $\|f_\rho\|_\infty \leq U$.

D.1. To bound the sample variance

Lemma D.1. *Fix any $\delta \in (0,1)$. With ℓ_1 defined in Lemma C.2, suppose that*

$$\ell_1 \lambda \geq C\kappa^2 \max\{\mathcal{N}_M(\lambda), 1\} \log^2 \frac{10}{\delta}.$$

Then, under Assumption 4.1 and on the event $\mathcal{E}_{\boldsymbol{\omega}}$, the following inequality holds

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|_\rho \leq \big\|\widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}\big\|_{\ell_2,\ell_2} \big\|C_{M,\lambda}^{-1/2}(\widehat{S}_M^\top \mathbf{y} - S_M^\top f_\rho)\big\|_2 \leq 42\Big(\frac{2\kappa U}{\ell_2\sqrt{\lambda}} + \sqrt{\frac{U^2\mathcal{N}_M(\lambda)}{\ell_2}}\Big) \log \frac{10}{\delta}, \tag{17}$$

with probability at least $1 - \frac{2\delta}{5}$, where

$$\ell_2 = \begin{cases} \frac{b_1 n}{2(\max\{1, \log(2^{-1}\kappa - U^{-1}(\kappa + U\sqrt{dB})b_0 b_1 n)\})^{1/\gamma_0}}, & \textit{if} \quad \tau(k) \leq b_0 \exp(-(b_1 k)^{\gamma_0}); \\ \big(\frac{U\sqrt{\mathcal{N}_M(\lambda)\lambda}}{(\kappa + U\sqrt{dB})b_2}\big)^{\frac{2}{2\gamma_1+1}} \big(\frac{n}{2}\big)^{\frac{2\gamma_1}{2\gamma_1+1}}, & \qquad\qquad \textit{if} \quad \tau(k) \leq b_2 k^{-\gamma_1}. \end{cases} \tag{18}$$

Proof. Note that

$$\begin{aligned} \|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|_\rho =& \big\|S_M \widehat{C}_{M,\lambda}^{-1} \widehat{S}_M^\top \mathbf{y} - S_M \widehat{C}_{M,\lambda}^{-1} S_M^\top f_\rho\big\|_\rho \\ =& \big\|S_M \widehat{C}_{M,\lambda}^{-1} (\widehat{S}_M^\top \mathbf{y} - S_M^\top f_\rho)\big\|_\rho \\ =& \big\|S_M \widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2} C_{M,\lambda}^{-1/2} (\widehat{S}_M^\top \mathbf{y} - S_M^\top f_\rho)\big\|_\rho \\ =& \big\|S_M \widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\mathcal{L}} \big\|C_{M,\lambda}^{-1/2} (\widehat{S}_M^\top \mathbf{y} - S_M^\top f_\rho)\big\|_2. \end{aligned}$$

We first derive the upper bound for $\|S_M \widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2}\|_{\ell_2,\mathcal{L}}$. Note that

$$\begin{aligned} \big\|S_M \widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\mathcal{L}} &\leq \big\|S_M \widehat{C}_{M,\lambda}^{-1/2}\big\|_{\ell_2,\mathcal{L}} \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\ell_2} \\ &\overset{(i)}{=} \big\|\widehat{C}_{M,\lambda}^{-1/2} S_M^\top S_M \widehat{C}_{M,\lambda}^{-1/2}\big\|_{\ell_2,\ell_2}^{1/2} \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\ell_2} \\ &= \big\|\widehat{C}_{M,\lambda}^{-1/2} C_M \widehat{C}_{M,\lambda}^{-1/2}\big\|_{\ell_2,\ell_2}^{1/2} \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\ell_2} \\ &\overset{(ii)}{\leq} \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\ell_2}^2, \end{aligned}$$

where (i) follows from that $\|A\| = \|A^\top A\|^{1/2}$ for any bounded operator A and the fact that $\widehat{C}_{M,\lambda}^{-1/2}$ is self-adjoint, and (ii) follows from the fact that $\|AB\| = \|BA\|$ for two self-adjoint operators A and B (Caponnetto & De Vito, 2007) and

$$\begin{aligned} \big\|\widehat{C}_{M,\lambda}^{-1/2} C_M \widehat{C}_{M,\lambda}^{-1/2}\big\|_{\ell_2,\ell_2} &= \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2} C_{M,\lambda}^{-1/2} C_M C_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2} \widehat{C}_{M,\lambda}^{-1/2}\big\|_{\ell_2,\ell_2} \\ &\leq \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\ell_2} \big\|C_{M,\lambda}^{-1/2} C_M C_{M,\lambda}^{-1/2}\big\|_{\ell_2,\ell_2} \big\|C_{M,\lambda}^{1/2} \widehat{C}_{M,\lambda}^{-1/2}\big\|_{\ell_2,\ell_2} \\ &\leq \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\ell_2}^2. \end{aligned}$$

By applying Lemmas C.3 and the Cordes inequality presented in Lemma E.2, with probability at least $1 - \delta/5$, we have

$$\big\|S_M \widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\mathcal{L}} \leq \big\|\widehat{C}_{M,\lambda}^{-1/2} C_{M,\lambda}^{1/2}\big\|_{\ell_2,\ell_2}^2 \leq \big\|\widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}\big\|_{\ell_2,\ell_2} \leq 2. \tag{19}$$

We proceed to bound from above $\big\|C_{M,\lambda}^{-1/2}(\widehat{S}_M^\top \mathbf{y} - S_M^\top f_\rho)\big\|_2$. Recall $\boldsymbol{\phi}_M(\mathbf{x}) = \frac{1}{\sqrt{M}}(\psi(\mathbf{x}, \boldsymbol{\omega}_1), \ldots, \psi(\mathbf{x}, \boldsymbol{\omega}_M))^\top$. We define

$$\xi(\mathbf{x}, y) = C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\mathbf{x}) y - S_M^\top f_\rho).$$

Then, we have

$$\begin{aligned}
E[\xi(\boldsymbol{X},Y)] =& \int_{\mathcal{X}\times\mathcal{Y}} C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\mathbf{x})y - S_M^\top f_\rho) d\rho_{\boldsymbol{X}}(\mathbf{x}) \\
=& \int_{\mathcal{X}} C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}) \int_{\mathcal{Y}} y d\rho(y|\,\mathbf{x}) d\rho_{\boldsymbol{X}}(\mathbf{x}) - C_{M,\lambda}^{-1/2} S_M^\top f_\rho \\
=& \int_{\mathcal{X}} C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}) f_\rho(\mathbf{x}) d\rho_{\boldsymbol{X}}(\mathbf{x}) - C_{M,\lambda}^{-1/2} S_M^\top f_\rho \\
=& C_{M,\lambda}^{-1/2} S_M^\top f_\rho - C_{M,\lambda}^{-1/2} S_M^\top f_\rho = 0.
\end{aligned}$$

It is also clear by the definition of $\widehat{S}_M^\top$ that

$$C_{M,\lambda}^{-1/2}(\widehat{S}_M^\top\,\mathbf{y} - S_M^\top f_\rho) = \frac{1}{n}\sum_{i=1}^n \xi(\mathbf{x}_i, y_i).$$

Then, there holds

$$\begin{aligned}
\|\xi(\mathbf{x},y)\|_2 =& \big\|C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\mathbf{x})y - S_M^\top f_\rho)\big\|_2 \\
\le& \big\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x})y\big\|_2 + \big\|C_{M,\lambda}^{-1/2} S_M^\top f_\rho(\mathbf{x})\big\|_2 \\
\overset{(i)}{\le}& \frac{U\kappa}{\sqrt{\lambda}} + \big\|C_{M,\lambda}^{-1/2} S_M^\top f_\rho\big\|_2 \\
\overset{(ii)}{\le}& \frac{2U\kappa}{\sqrt{\lambda}},
\end{aligned}$$

where (i) follows from (15) and $\|C_{M,\lambda}^{-1/2}\|_{\ell_2,\ell_2} \le \lambda^{-1/2}$, $|y| \le U$, $\|f_\rho\|_\infty \le U$, (ii) follows from Jensen's inequality that

$$\big\|C_{M,\lambda}^{-1/2} S_M^\top f_\rho\big\|_2 = \Big\| \int_{\mathcal{X}\times\mathcal{Y}} C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x}) y d\rho_{\boldsymbol{X}}(\mathbf{x})\Big\|_2 \le \int_{\mathcal{X}\times\mathcal{Y}} \big\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x})y\big\|_2 d\rho_{\boldsymbol{X}}(\mathbf{x}) \le \frac{U\kappa}{\sqrt{\lambda}}.$$

Now, we turn to verify the second-moment condition. Note that

$$\begin{aligned}
E[\|\xi(\boldsymbol{X},Y)\|_2^2] &= E\big[\big\|C_{M,\lambda}^{-1/2}(\boldsymbol{\phi}_M(\boldsymbol{X})Y - S_M^\top f_\rho)\big\|_2^2\big] \\
&\overset{(i)}{\le} E\big[\big\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\boldsymbol{X})Y\big\|_2^2\big] \\
&\le U^2 \int \big\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x})\big\|_2^2 d\rho_{\boldsymbol{X}}(\mathbf{x}) && \text{by } |Y| \le U \\
&\overset{(ii)}{=} U^2 J_M(\lambda) \\
&= U^2 \mathcal{N}_M(\lambda) && \text{by } J_M(\lambda) = \mathcal{N}_M(\lambda),
\end{aligned}$$

where (i) follows from $E[\|\,\boldsymbol{Z} - E\,\boldsymbol{Z}\,\|_2^2] \le E\|\,\boldsymbol{Z}\,\|_2^2$ for any random vector $\boldsymbol{Z}$ and (ii) follows from

$$\int_{\mathcal{X}} \big\|C_{M,\lambda}^{-1/2}\boldsymbol{\phi}_M(\mathbf{x})\big\|_2^2 d\rho_{\boldsymbol{X}}(\mathbf{x}) = \int_{\mathcal{X}} \mathrm{Tr}\big(C_{M,\lambda}^{-1}\boldsymbol{\phi}_M(\mathbf{x})\boldsymbol{\phi}_M(\mathbf{x})^\top\big) d\rho_{\boldsymbol{X}}(\mathbf{x})$$

and (16) .

Now we turn to verify the function $\xi(\mathbf{x},y)$ is Lipschitz continuous. Specifically, for $\mathbf{z} = (\mathbf{x},y), \mathbf{z}' = (\mathbf{x}',y') \in \mathcal{X}\times\mathcal{Y}$, we have

$$\begin{aligned}\left\|\xi(\mathbf{x},y)-\xi(\mathbf{x}',y')\right\|_2 &\leq \left\|C_{M,\lambda}^{-1/2}\right\|_{\ell_2,\ell_2}\left\|\boldsymbol{\phi}_M(\mathbf{x})y-\boldsymbol{\phi}_M(\mathbf{x}')y'\right\|_2\\ &\leq \lambda^{-1/2}\big(\left\|\boldsymbol{\phi}_M(\mathbf{x})y-\boldsymbol{\phi}_M(\mathbf{x})y'\right\|_2+\left\|\boldsymbol{\phi}_M(\mathbf{x})y'-\boldsymbol{\phi}_M(\mathbf{x}')y'\right\|_2\big)\\ &\leq \lambda^{-1/2}\big(|y-y'|\,\|\boldsymbol{\phi}_M(\mathbf{x})\|_2+U\left\|\boldsymbol{\phi}_M(\mathbf{x})-\boldsymbol{\phi}_M(\mathbf{x}')\right\|_2\big)\\ &\overset{(i)}{\leq} \lambda^{-1/2}\big(\kappa|y-y'|+U\sqrt{dB}\|\mathbf{x}-\mathbf{x}'\|_2\big)\\ &\leq \lambda^{-1/2}\left(\kappa+U\sqrt{dB}\right)\|\mathbf{z}-\mathbf{z}'\|_2,\end{aligned}$$

where (i) follows from (11).

Then, the process $\{\xi(\boldsymbol{X}_i,Y_i)\}_{i\geq 1}$ is τ-mixing with coefficient $\lambda^{-1/2}(\kappa+U\sqrt{dB})\tau(k)$. Clearly, by applying Lemma E.1 with $L=2\kappa\lambda^{-1/2}U$ and $\sigma^2=U^2\mathcal{N}_M(\lambda)$, and the corresponding Hilbert space is $\mathcal{R}^M$, we have

$$\left\|C_{M,\lambda}^{-1/2}(\widehat{S}_M^\top\mathbf{y}-S_M^\top f_\rho)\right\|_2\leq 21\Big(\frac{2\kappa U}{\ell_2\sqrt{\lambda}}+\sqrt{\frac{U^2\mathcal{N}_M(\lambda)}{\ell_2}}\Big)\log\frac{10}{\delta},$$

with probability at least $1-\delta/5$, where ℓ_2 is defined in (18). Together with (19), we complete the proof. □

Remark D.2. In this proof, we use the assumption $|Y|\leq U$, which further implies $|f_\rho(\mathbf{x})|=\big|E[Y\,|\,\boldsymbol{X}=\mathbf{x}]\big|\leq E[|Y|\,|\,\boldsymbol{X}=\mathbf{x}]\leq U$. We remark that this assumption essentially requires the random noise to be uniformly bounded, which can be extended to sub-exponential noise condition with a slight order sacrifice of an additional $\log n$ term in the upper bound on $\|\widehat{f}_{M,\lambda}-\widetilde{f}_{M,\lambda}\|_\rho$. To be clear, suppose that $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. sub-exponential variables: that is, there exist positive constants K_1,K_2 such that $P\big(|\varepsilon_i|>t\big)\leq K_1\exp(-K_2t)$ for all $t\geq 0$. Then, by the union bound, we have

$$P\Big(\max_{i=1,\ldots,n}|\varepsilon_i|>t\Big)\leq P\Big(\bigcup_{i=1}^n\big\{|\varepsilon_i|>t\big\}\Big)\leq\sum_{i=1}^n K_1\exp(-K_2t)=nK_1\exp(-K_2t).$$

Consequently, for any $\delta\in(0,1)$, by taking $t=\frac{1}{K_2}\log\big(\frac{nK_1}{\delta}\big)$, it holds with probability at least $1-\delta$ that

$$\max_{i=1,\ldots,n}|\varepsilon_i|\leq\frac{1}{K_2}\log\big(\frac{nK_1}{\delta}\big)\lesssim\log\frac{1}{\delta}+\log n.$$

D.2. To bound the approximation error

Note that the approximation error does not involve dependent data specifically considered in this paper. Therefore, we can directly apply the existing upper bound for the approximation error established in Rudi & Rosasco (2017). We summarize some important results in Rudi & Rosasco (2017, Theorems 4 & 6) in the following lemma which will be used in our proof.

Lemma D.3. *Suppose $\lambda\leq\frac{3}{4}\|L_K\|_{\mathcal{L},\mathcal{L}}$ and the number of RFs satisfies that*

$$M\geq\max\left\{4\kappa^2\Big(\frac{\mathcal{N}(\lambda)}{\lambda}\Big)^{2r-1}\Big(\mathcal{N}_\infty(\lambda)\log\frac{55\kappa^2}{\lambda}\Big)^{2-2r},\ 18\,(q_0+\mathcal{N}_\infty(\lambda))\log\frac{540\kappa^2}{\lambda\delta}\right\}\tag{20}$$

with $q_0=2(2+\kappa/\|L_K\|_{\mathcal{L},\mathcal{L}}+\kappa^2)$. Then, for any $\delta\in(0,1)$, the following inequality holds with probability at least $1-\delta/5$

$$\left\|L_{M,\lambda}^{-1/2}L_{K,\lambda}^{1/2}\right\|_{\mathcal{L},\mathcal{L}}\leq 2.\tag{21}$$

Moreover, we have the following upper bound on the approximation error

$$\|\bar{f}_{M,\lambda}-f_\rho\|_\rho\leq 3R\lambda^r,$$

where $R=\|g_\rho\|_\rho$.

D.3. To bound the empirical error

To establish the upper bound on the empirical error, we first borrow a lemma from Rudi & Rosasco (2017).

Lemma D.4 (Lemma 3 in Rudi & Rosasco (2017))**.** *Under Assumption 2.4, the following inequality holds for any* $\lambda > 0, M, n$

$$\|\widetilde{f}_{M,\lambda} - \bar{f}_{M,\lambda}\|_\rho \leq R\max\{1,\kappa\}\big\|L_{M,\lambda}^{-1/2}L_K^{1/2}\big\|_{\mathcal{L},\mathcal{L}}\|S_M\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{-1/2}\|_{\ell_2,\mathcal{L}}\big\|C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)\big\|_{\ell_2,\ell_2},$$

where $R = \|g_\rho\|_\rho$*.*

Lemma D.5. *Suppose* $\lambda \leq \frac{3}{4}\|L_K\|_{\mathcal{L},\mathcal{L}}$ *and* $\ell_1\lambda \geq C\kappa^2\max\{\mathcal{N}_M(\lambda),1\}\log^2\frac{10}{\delta}$*, where* ℓ_1 *is defined in Lemma 13. Furthermore, we assume that the number of RFs satisfies (20). Then, on the event* $\mathcal{E}_{\boldsymbol{\omega}}$*, for any* $\delta \in (0,1)$*, the following inequality holds*

$$\|\widetilde{f}_{M,\lambda} - \bar{f}_{M,\lambda}\|_\rho \leq 84R\max\{1,\kappa\}\Big(\frac{2\kappa^2}{\ell_1\sqrt{\lambda}} + \sqrt{\frac{\kappa^2\mathcal{N}_M(\lambda)}{\ell_1}}\Big)\log\frac{10}{\delta},$$

with probability at least $1 - \frac{2\delta}{5}$*.*

Proof. By applying Lemmas C.2 and D.4, and together with the upper bound (21) on $\big\|L_{M,\lambda}^{-1/2}L_K^{1/2}\big\|_{\mathcal{L},\mathcal{L}}$ and the upper bound (19) on $\|S_M\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{-1/2}\|_{\ell_2,\mathcal{L}}$, we immediately obtain the desired result. □

D.4. The proof for the main theorems

In this section, we complete the proof for Theorems 4.2 and 4.3 and the corresponding corollaries. Note that the results in Theorems 4.2 and 4.3 hold without requiring Assumption 4.5. Since Corollaries 4.6 and 4.7 are general versions of the two theorem above, we can specialize the results in Theorems 4.2 and 4.3 with $\beta = 1$, we only provide the proof for the Corollaries 4.6 and 4.7.

Before providing the formal proof, we recall that the involved random events consist of:

- If $M \geq C\log\frac{10d}{\delta}$, we have $P(\mathcal{E}_{\boldsymbol{\omega}}) \geq 1 - \delta/5$ (Lemma C.1);
- On the event $\mathcal{E}_{\boldsymbol{\omega}}$, it holds with probability at least $1 - \delta/5$ that (Lemma C.2)
$$\big\|C_{M,\lambda}^{-1/2}(\widehat{C}_M - C_M)\big\|_{\ell_2,\ell_2} \leq 21\Big(\frac{2\kappa^2}{\ell_1\sqrt{\lambda}} + \sqrt{\frac{\kappa^2\mathcal{N}_M(\lambda)}{\ell_1}}\Big)\log\frac{10}{\delta}.$$
- If $\lambda \leq \|L_K\|_{\mathcal{L},\mathcal{L}}$ and $M \geq (4 + 18\mathcal{N}_\infty(\lambda))\log\frac{60\kappa}{\lambda\delta}$, with probability at least $1 - \delta/5$, one has $\frac{1}{2}\mathcal{N}(\lambda) \leq \mathcal{N}_M(\lambda) \leq 3\mathcal{N}(\lambda)$ (Proposition 10 Rudi & Rosasco (2017));
- On the event $\mathcal{E}_{\boldsymbol{\omega}}$, it holds with probability at least $1 - \delta/5$ that (Lemma D.1)
$$\big\|C_{M,\lambda}^{-1/2}(\widehat{S}_M^\top\mathbf{y} - S_M^\top f_\rho)\big\|_2 \leq 21\Big(\frac{2\kappa U}{\ell_2\sqrt{\lambda}} + \sqrt{\frac{U^2\mathcal{N}_M(\lambda)}{\ell_2}}\Big)\log\frac{10}{\delta}.$$
- If $\lambda \leq \frac{3}{4}\|L_K\|_{\mathcal{L},\mathcal{L}}$ and M satisfies the lower bound (20), with probability at least $1-\delta/5$, one has $\|L_{M,\lambda}^{-1/2}L_{K,\lambda}^{1/2}\|_{\mathcal{L},\mathcal{L}} \leq 2$ and $\|\bar{f}_{M,\lambda} - f_\rho\|_\rho \leq 3R\lambda^r$ (Lemma D.3).

Proof. Note that if $\ell_1\lambda \geq C\kappa^2\max\{\mathcal{N}_M(\lambda),1\}\log^2\frac{10}{\delta}$, by applying Lemma D.1 , it holds with probability $1 - \frac{2\delta}{5}$ that

$$\|\widehat{f}_{M,\lambda} - \widetilde{f}_{M,\lambda}\|_\rho \leq 42\Big(\frac{2\kappa U}{\ell_2\sqrt{\lambda}} + \sqrt{\frac{U^2\mathcal{N}_M(\lambda)}{\ell_2}}\Big)\log\frac{10}{\delta}, \tag{22}$$

and if the number of RFs satisfies (20), by applying Lemma D.3, it holds with probability $1-\delta/5$ that

$$\|\bar{f}_{M,\lambda}-f_\rho\|_\rho \leq 3R\lambda^r. \tag{23}$$

In addition, by applying Lemma D.5 , it holds with probability $1-\frac{2\delta}{5}$ that

$$\|\widetilde{f}_{M,\lambda}-\bar{f}_{M,\lambda}\|_\rho \leq 84R\max\{1,\kappa\}\Big(\frac{2\kappa^2}{\ell_1\sqrt{\lambda}}+\sqrt{\frac{\kappa^2\mathcal{N}_M(\lambda)}{\ell_1}}\Big)\log\frac{10}{\delta}. \tag{24}$$

Recall the error decomposition in (9), by combining (22), (23) and (24), with probability at least $1-\delta$, we have

$$\|\widehat{f}_{M,\lambda}-f_\rho\|_\rho \leq 3R\lambda^r+168d_1d_2\Big(\frac{2\kappa d_3}{\ell_3\sqrt{\lambda}}+\sqrt{\frac{d_3^2\mathcal{N}_M(\lambda)}{\ell_3}}\Big)\log\frac{10}{\delta},$$

where $d_1=\max\{1,R\}$, $d_2=\max\{1,\kappa\}$, $d_3=\max\{\kappa,U\}$ and $\ell_3=\min\{\ell_1,\ell_2\}$.

Note that the requirement in Lemma C.1 that $M\geq C\log\frac{10d}{\delta}$ is also satisfied since the requirement in (20) is a more strict constraint on the number of RFs.

Then, it remains to verify these three conditions:

i) $\ell_1\lambda\geq C\kappa^2\max\{\mathcal{N}_M(\lambda),1\}\log^2\frac{10}{\delta}$;

ii) $M\geq(4+18\mathcal{N}_\infty(\lambda))\log\frac{60\kappa}{\lambda\delta}$;

iii)

$$M\geq 4\kappa^2\Big(\frac{\mathcal{N}(\lambda)}{\lambda}\Big)^{2r-1}\Big(\mathcal{N}_\infty(\lambda)\log\frac{55\kappa^2}{\lambda}\Big)^{2-2r}\vee 18\,(q_0+\mathcal{N}_\infty(\lambda))\log\frac{540\kappa^2}{\lambda\delta}$$

with $q_0=2(2+\frac{\kappa}{\|L_K\|_{\mathcal{L},\mathcal{L}}}+\kappa^2)$.

Indeed, by applying Proposition 10 in Rudi & Rosasco (2017) and Assumption 2.3, we have

$$\mathcal{N}_M(\lambda)\leq 3\mathcal{N}(\lambda)\leq 3c_0\lambda^{-\alpha}$$

with probability at least $1-\delta/5$. Note that whether the τ-mixing coefficient is exponential decay or polynomial decay, we always have $\ell_1\asymp\ell_2\asymp\ell_3$. We always use ℓ_3 in the remaining proof.

Then, with the choice of $\lambda\asymp\ell_3^{-\frac{1}{2r+\alpha}}$, by using Assumption 2.3 that $\mathcal{N}(\lambda)\leq c_0\lambda^{-\alpha}$ and Assumption 4.5 that $\mathcal{N}_\infty(\lambda)\leq c_1\lambda^{-\beta}$, the required number of RFs turns to

$$M\geq C\ell_3^{\frac{(\alpha+1)(2r-1)}{2r+\alpha}}\ell_3^{\frac{(2-2r)\beta}{2r+\alpha}}\log\frac{2\ell_3}{\delta}=C\ell_3^{\frac{\beta+(1+\alpha-\beta)(2r-1)}{2r+\alpha}}\log\frac{2\ell_3}{\delta},$$

where C hides several constants, including $c_0,c_1,\kappa,\|L_K\|_{\mathcal{L}\to\mathcal{L}}$.

In addition, $\ell_1\lambda\geq C\kappa^2\max\{\mathcal{N}_M(\lambda),1\}\log^2\frac{10}{\delta}$ reduces to

$$\ell_3\ell_3^{-\frac{1}{2r+\alpha}}\geq C\kappa^2\max\{\ell_3^{\frac{\alpha}{2r+\alpha}},1\}\log^2\frac{10}{\delta},$$

which holds as long as n is sufficiently large.

Next, we separately consider two cases: (1) $\tau(k)\leq b_0\exp(-(b_1k)^{\gamma_0})$, and (2) $\tau(k)\leq b_2k^{-\gamma_1}$.

Case 1: For $\tau(k)\leq b_0\exp(-(b_1k)^{\gamma_0})$, we have

$$\ell_1\asymp\ell_2\asymp\ell_3\asymp n(\log n)^{-\frac{1}{2\gamma_0}}.$$

Then, λ turns to $n^{-\frac{1}{2r+\alpha}}(\log n)^{\frac{1}{2\gamma_0(2r+\alpha)}}$ and the number of RFs must satisfy

$$M\geq Cn^{\frac{\beta+(1+\alpha-\beta)(2r-1)}{2r+\alpha}}(\log n)^{-\frac{\beta+(1+\alpha-\beta)(2r-1)}{2\gamma_0(2r+\alpha)}}\log\frac{2n}{\delta}.$$

Due to $\frac{\beta+(1+\alpha-\beta)(2r-1)}{2r+\alpha} \geq \frac{1}{2}$, $M \geq Cn^{\frac{\beta+(1+\alpha-\beta)(2r-1)}{2r+\alpha}}(\log n)^{-\frac{1}{4\gamma_0}} \log \frac{2n}{\delta}$ will be enough.

Therefore, by applying the union bound, it holds with probability at least $1-\delta$ that

$$\begin{aligned}
\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho &\lesssim \lambda^r + \Big(\frac{1}{\ell_3\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}_M(\lambda)}{\ell_3}}\Big) \log \frac{10}{\delta} \\
&\lesssim \ell_3^{-\frac{r}{2r+\alpha}} + \big(\ell_3^{-\frac{4r+2\alpha-1}{4r+2\alpha}} + \ell_3^{-\frac{r}{2r+\alpha}}\big) \log \frac{10}{\delta} \\
&\overset{(i)}{\lesssim} \ell_3^{-\frac{r}{2r+\alpha}} \log \frac{10}{\delta} \\
&\lesssim n^{-\frac{r}{2r+\alpha}} (\log n)^{\frac{r}{2\gamma_0(2r+\alpha)}} \log \frac{10}{\delta}
\end{aligned}$$

where (i) holds if $r \geq \frac{1}{2}$.

Case 2: For $\tau(k) \leq b_2 k^{-\gamma_1}$, we have

$$\ell_1 \asymp \ell_2 \asymp \ell_3 \asymp (\lambda^{\frac{1-\alpha}{2}})^{\frac{2}{2\gamma_1+1}} n^{\frac{2\gamma_1}{2\gamma_1+1}} = \lambda^{\frac{1-\alpha}{2\gamma_1+1}} n^{\frac{2\gamma_1}{2\gamma_1+1}}.$$

Then, λ turns to $n^{-\frac{2\gamma_1}{4\gamma_1 r+2r+2\alpha\gamma_1+1}}$ and the number of RFs must satisfy

$$M \geq Cn^{\frac{2\gamma_1\beta+2\gamma_1(1+\alpha-\beta)(2r-1)}{4\gamma_1 r+2r+2\alpha\gamma_1+1}} \log \frac{2n}{\delta}.$$

Similarly, by applying the union bound, it holds with probability at least $1-\delta$ that

$$\begin{aligned}
\|\widehat{f}_{M,\lambda} - f_\rho\|_\rho &\lesssim \lambda^r + \Big(\frac{1}{\ell_3\sqrt{\lambda}} + \sqrt{\frac{\mathcal{N}_M(\lambda)}{\ell_3}}\Big) \log \frac{10}{\delta} \\
&\lesssim \ell_3^{-\frac{r}{2r+\alpha}} + \big(\ell_3^{-\frac{4r+2\alpha-1}{4r+2\alpha}} + \ell_3^{-\frac{r}{2r+\alpha}}\big) \log \frac{10}{\delta} \\
&\lesssim n^{-\frac{2\gamma_1 r}{4\gamma_1 r+2r+2\alpha\gamma_1+1}} \log \frac{10}{\delta}.
\end{aligned}$$

This completes the proof. □

E. Auxiliary Lemmas

The following lemma provides Bernstein's inequality for the sum of zero-mean random elements (Blanchard & Zadorozhnyi, 2019).

Lemma E.1. *If the zero-mean stochastic process $\{(\xi_i)_{i\geq 1}\} \subset \mathcal{H}$ is τ-mixing process with rate $\tau(k)$ in exponentially or polynominally decay, where $\mathcal{H}$ be a real separable Hilbert space equipped with norm $\|\cdot\|_{\mathcal{H}}$. Assume that there exist some constants $L, \sigma > 0$ such that*

$$\|\xi\|_{\mathcal{H}} \leq L \quad \textit{a.s.} \quad \textit{and} \quad E\|\xi\|_{\mathcal{H}}^2 \leq \sigma^2.$$

Then, let $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$\Big\|\frac{1}{n}\sum_{i=1}^n \xi_i\Big\|_{\mathcal{H}} \leq 21 \left(\frac{L}{\ell^*} + \frac{\sigma}{\sqrt{\ell^*}}\right) \log \frac{2}{\delta},$$

where

$$\begin{aligned}
\ell^* &= (\max\{1, \log(b_0 b_1 n/L)\})^{-1/\gamma_0} \frac{b_1 n}{2}, && \textit{if} \quad \tau(k) \leq b_0 \exp(-(b_1 k)^{\gamma_0}); \\
\ell^* &= \left(\frac{\sigma}{b_2}\right)^{\frac{2}{2\gamma_1+1}} \left(\frac{n}{2}\right)^{\frac{2\gamma_1}{2\gamma_1+1}}, && \textit{if} \quad \tau(k) \leq b_2 k^{-\gamma_1}.
\end{aligned}$$

The following lemma is known as the Cordes inequality (Bhatia, 2013).

Lemma E.2. *Let A and B be positive operators on a Hilbert space. Then, for any $0 < r \leq 1$, we have $\|A^r B^r\| \leq \|AB\|^r$.*

Lemma E.3. *For a random variable X, the following statements are equivalent:*

(1) There exist constants K_1, K_2 such that $P\big(|X| > t\big) \leq K_1 \exp(-K_2 t)$ for all $t \geq 0$.

(2) There exist constants K_3, K_4 such that $E[\exp(\sigma(X - EX))] \leq \exp(K_3^2\sigma^2/2)$ for all σ satisfying $|\sigma| < 1/K_4$.

The above lemma characterizes the sub-exponential variable in two different manners: tail probability and moment generating function, one can refer to Wainwright (2019); Papaspiliopoulos (2020) for detailed proof and discussion. Formally, a random variable X is called sub-exponential if there are non-negative parameters (ν, α) such that

$$E[\exp(\sigma(X - EX))] \leq \exp\big(\frac{\sigma^2\nu^2}{2}\big) \quad \text{for all } |\sigma| \leq \frac{1}{\alpha}.$$

The following concentration inequality is known as Bernstein inequality, whose detailed proof can be found in Section 2 of Wainwright (2019).

Lemma E.4 (Bernstein inequality for sub-exponential variable)**.** *Suppose a random variable is sub-exponential with parameters (ν, α), then we have*

$$P[|X - EX| \geq t] \leq \exp\Big(-\frac{1}{2}\min\big\{\frac{t^2}{\nu^2}, \frac{t}{\alpha}\big\}\Big) = \begin{cases} 2\exp(-\frac{t^2}{2\nu^2}), & \textit{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ 2\exp(-\frac{t}{2\alpha}), & \textit{if } t > \frac{\nu^2}{\alpha}. \end{cases}$$

F. Data-dependent sampling strategy

In this section, we introduce a data-dependent sampling strategy proposed in Rudi & Rosasco (2017).

Example F.1 (Data dependent RF, Example 2 of Rudi & Rosasco (2017))**.** Suppose that kernel K has the integral representation (4). Let

$$s(\boldsymbol{\omega}) = \|(L_K + \lambda I)^{-1/2}\psi(\cdot, \boldsymbol{\omega})\|_\rho^{-2}, \quad \text{and} \quad L_s := \int \frac{1}{s(\boldsymbol{\omega})} d\pi(\boldsymbol{\omega}).$$

We consider random features

$$\psi_s(\mathbf{x}, \boldsymbol{\omega}) = \psi(\mathbf{x}, \boldsymbol{\omega})\sqrt{L_s s(\boldsymbol{\omega})}$$

with distribution $\pi_s(\boldsymbol{\omega}) := \pi(\boldsymbol{\omega})/(s(\boldsymbol{\omega})L_s)$. One can prove that these random features ensure the integral representation of K and satisfy Assumption 4.5 with $\beta = \alpha$. The proof of Example F.1 is given in Rudi & Rosasco (2017). We state it below for completeness.

Proof. We first prove random features $\psi_s(\cdot, \boldsymbol{\omega})$ ensure the integral representation of K, note that

$$\int_\Omega \psi_s(\mathbf{x}, \boldsymbol{\omega})\psi_s(\mathbf{x}', \boldsymbol{\omega}) d\pi_s(\boldsymbol{\omega}) = \int_\Omega \psi(\mathbf{x}, \boldsymbol{\omega})\psi(\mathbf{x}', \boldsymbol{\omega}) L_s s(\boldsymbol{\omega}) \frac{1}{L_s s(\boldsymbol{\omega})} d\pi_(\boldsymbol{\omega}) = \int_\Omega \psi(\mathbf{x}, \boldsymbol{\omega})\psi(\mathbf{x}', \boldsymbol{\omega}) d\pi(\boldsymbol{\omega}) = K(\mathbf{x}, \overset{\prime}{\mathbf{x}}).$$

Then we prove random features $\psi_s(\mathbf{x}, \boldsymbol{\omega})$ satisfy Assumption 4.5 with $\beta = \alpha$,

$$\begin{aligned} \mathcal{N}_\infty(\lambda) &= \sup_{\boldsymbol{\omega}\in\Omega} \|(L_K + \lambda I)^{-1/2}\psi_s(\cdot, \boldsymbol{\omega})\|_\rho^2 \\ &= \sup_{\boldsymbol{\omega}\in\Omega} \Big\|(L_K + \lambda I)^{-1/2}\psi(\cdot, \boldsymbol{\omega})\sqrt{L_s s(\boldsymbol{\omega})}\Big\|_\rho^2 \\ &= \sup_{\boldsymbol{\omega}\in\Omega} \Big\|(L_K + \lambda I)^{-1/2}\psi(\mathbf{x}, \boldsymbol{\omega})\}\Big\|_\rho^2 \Big\|(L_K + \lambda I)^{-1/2}\psi(\mathbf{x}, \boldsymbol{\omega})\}\Big\|_\rho^{-2} \int_\Omega \|(L_K + \lambda I)^{-1/2}\psi(\cdot, \boldsymbol{\omega})\|_\rho^2 d\pi(\boldsymbol{\omega}) \\ &= \int_\Omega \|(L_K + \lambda I)^{-1/2}\psi(\cdot, \boldsymbol{\omega})\|_\rho^2 d\pi(\boldsymbol{\omega}) \\ &= \mathcal{N}(\lambda) \leq c_0\lambda^{-\alpha}. \end{aligned}$$

This completes the proof. □