Optimal Kernel Quantile Learning with Random Features

Caixing Wang [1] **Xingdong Feng** [1]

Abstract

The random feature (RF) approach is a well-established and efficient tool for scalable kernel methods, but existing literature has primarily focused on kernel ridge regression with random features (KRR-RF), which has limitations in handling heterogeneous data with heavy-tailed noises. This paper presents a generalization study of kernel quantile regression with random features (KQR-RF), which accounts for the non-smoothness of the check loss in KQR-RF by introducing a refined error decomposition and establishing a novel connection between KQR-RF and KRR-RF. Our study establishes the capacity-dependent learning rates for KQR-RF under mild conditions on the number of RFs, which are minimax optimal up to some logarithmic factors. Importantly, our theoretical results, utilizing a data-dependent sampling strategy, can be extended to cover the agnostic setting where the target quantile function may not precisely align with the assumed kernel space. By slightly modifying our assumptions, the capacity-dependent error analysis can also be applied to cases with Lipschitz continuous losses, enabling broader applications in the machine learning community. To validate our theoretical findings, simulated experiments and a real data application are conducted.

1. Introduction

Kernel methods are pivotal in statistical analysis and have seen extensive applications in nonparametric regression (Wahba, 1990; Vapnik, 1999) and classification (Schölkopf & Smola, 2002; Steinwart & Christmann, 2008). Despite their empirical success, typical kernel algorithms struggle with large-scale datasets, hindered by substantial computational cost, scaling as $\mathcal{O}(|D|^3)$, and considerable storage requirements, scaling as $\mathcal{O}(|D|^2)$, where $|D|$ is the sample size of the dataset. This limitation has motivated significant research efforts towards scalable kernel methods, such as distributed learning (Zhang et al., 2015; Lin et al., 2017; 2020; Lian, 2022), Nyström subsampling (Williams & Seeger, 2001; Rudi et al., 2015; Li et al., 2023a), random features (Rahimi & Recht, 2007; 2008; Rudin, 2017; Rudi & Rosasco, 2017), stochastic gradient methods (Lin & Rosasco, 2016; Lin & Cevher, 2020), Falkon (Rudi et al., 2017) and EigenPro 3.0 (Abedsoltan et al., 2023).

Among these popular accelerated methods, random features (Rahimi & Recht, 2007) is a kernel approximation technique that maps the attribute space to a finite and low-dimensional space through Bochner's theorem (Bochner, 2005; Rudin, 2017). Recent attention in theoretical analysis has been directed toward kernel methods employing random features (Sutherland & Schneider, 2015; Rudi & Rosasco, 2017; Avron et al., 2017; Bach, 2017; Carratino et al., 2018). In the context of kernel ridge regression with random features, significant efforts have focused on achieving optimal learning rates (Rudi & Rosasco, 2017; Avron et al., 2017; Bach, 2017; Li et al., 2023b; Liu & Lian, 2023), aligning with the minimax optimal rates of exact KRR (Caponnetto & De Vito, 2007) under mild conditions on the number of random features. However, it is worth pointing out that these works mainly focus on the least square loss which exclusively estimates the conditional mean of the response given the covariate variables. Broader investigations encompass losses that are Lipschitz continuous, as in support vector machine (SVM) and logistic regression (Bach, 2017; Sun et al., 2018; Li et al., 2021). It is worth noting that their statistical guarantees are capacity-independent and rely on the stringent assumption that the true target function f_ρ lies in the assumed kernel space, i.e., $f_\rho \in \mathcal{H}_K$, known as the realizable setting. However, the agnostic setting, where the true target function f_ρ is not in the assumed kernel space, i.e., $f_\rho \notin \mathcal{H}_K$, is more common in practice. *This leads to a motivating question: can the capacity-dependent optimal rates for some general losses using random features be maintained in the agnostic settings?*

This paper primarily investigates the statistical and computational trade-offs inherent in random feature approximation for nonparametric quantile regression within a reproducing kernel Hilbert space (RKHS), also known as

[1]School of Statistics and Management, Shanghai University of Finance and Economics. Correspondence to: Xingdong Feng <feng.xingdong@mail.shufe.edu.cn>.

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

kernel quantile regression with random features (KQR-RF). In contrast to KRR-RF, KQR-RF models the entire conditional quantiles of the response, enhancing robustness against outliers and handling data heterogeneity more effectively (Koenker, 2005; Takeuchi et al., 2006; Li et al., 2007; Lian, 2022; Wang et al., 2024+). Our objective is to establish the capacity-dependent optimal rates for KQR-RF applicable to both realizable and agnostic settings. Based on this special check loss, we extend the theoretical framework to a broader family of Lipschitz continuous loss functions. This expansion encompasses various commonly employed methodologies as specialized instances, including mean regression, quantile regression, likelihood-based classification, and margin-based classification.

1.1. Our Contributions

The main contributions of this paper are multi-folds.

Comprehensive theoretical assessments. We propose a comprehensive theoretical analysis of KQR-RF, offering deep insights into the impact of random features on kernel quantile learning. To the best of our knowledge, this is the first work to provide generalization analysis for random features in kernel quantile learning. Moreover, the optimal learning rates we derived can be directly extended to the general Lipschitz loss functions. Compared to the existing results, which are either capacity-independent (Rahimi & Recht, 2008; Bach, 2017; Li et al., 2021) or suboptimal (Sun et al., 2018), we provide the capacity-dependent optimal learning rates (Theorem 3.13) for KQR-RF (and RF for Lipschitz loss) for both realizable and agnostic settings.

Efficient computational improvement. For both uniform sampling and data-dependent sampling schemes, we obtain, to the best of our knowledge, the minimum number of random features required for the optimal learning rates in the literature. Specifically, we reduce the number of random features from $\mathcal{O}(|D|), r = 1/2$ (Rahimi & Recht, 2008; Li et al., 2021) to $\mathcal{O}(|D|^{\frac{1}{2r+\gamma}}) \vee \mathcal{O}(|D|^{\frac{(2r-1)\gamma+1}{2r+\gamma}}),\ r \in (0,1]$ (Theorem 3.9) for the uniform sampling scheme; and $\mathcal{O}(|D|^{\frac{2\gamma}{2\gamma+1}}),\ r = 1/2$ (Sun et al., 2018) to $\mathcal{O}(|D|^{\frac{\gamma}{2r+\gamma}}) \vee \mathcal{O}(|D|^{\frac{2r+\gamma-1}{2r+\gamma}}),\ r \in (0,1]$ (Corollary 3.16) for the data-dependent sampling scheme. Here, $|D|$ is the sample size, r, and γ are some key parameters defined in Section 3. The improvement notably enhances computational efficiency.

Novel proof skills. In contrast to existing related work on KRR and its RF variants (Caponnetto & De Vito, 2007; Rudi & Rosasco, 2017; Li et al., 2023b), the estimator of KQR-RF (random features for Lipschitz loss) lacks an explicit solution, posing challenges in deriving the learning rates. In our proof, we first provide a novel error decomposition including a least square approximation (LS-approximation) error term (Lemma B.3). Leveraging the empirical process and a self-calibration assumption, we successfully establish a connection between the KQR-RF estimator $f_{M,D,\lambda}$ and its KRR-RF approximation $f^{\diamond}_{M,D,\lambda}$. The theoretical extension to the entire regularity setting when $r \in (0,1]$ is achieved by using the nontrivial Cauthy-Schwarz and Young's inequalities, along with sharper estimates for the differences between the operators. A more detailed proof sketch will be provided in Section 4.

Diverse numerical verification. Another contribution of this work is the exploration of KQR-RF's efficacy across diverse synthetic and real-world examples, further validating the theoretical findings in this paper.

1.2. Related work

Some most related works are presented below.

Random features approximation. Rahimi & Recht (2007); Sutherland & Schneider (2015); Sriperumbudur & Szabó (2015) have investigated the approximation error between the approximated kernel function $K_M(\boldsymbol{x}, \boldsymbol{x}')$ and the original kernel Gram-matrix $K(\boldsymbol{x}, \boldsymbol{x}')$, requiring $\mathcal{O}(|D|)$ features to guarantee the accuracy of the approximation, i.e., $\sup_{\boldsymbol{x},\boldsymbol{x}'} |K_M(\boldsymbol{x}, \boldsymbol{x}') - K(\boldsymbol{x}, \boldsymbol{x}')| = \mathcal{O}(|D|^{-1/2})$. Another line of research delves into the generalization properties of random features in various supervised learning tasks, such as kernel ridge regression (Bach, 2017; Avron et al., 2017; Rudi & Rosasco, 2017), kernel support vector machine (KSVM) (Sun et al., 2018), and kernel learning with Lipschitz loss (Rahimi & Recht, 2008; Li et al., 2021; Li, 2022). However, the success of these works depends on the realizable setting where the true function satisfies $f_\rho \in \mathcal{H}_K$.

Agnostic kernel learning. Recent studies have established the capacity-dependent optimal learning rates in the agnostic kernel learning, such as KRR (Smale & Zhou, 2007; Zhang et al., 2023), along with its variations including random features (Li et al., 2023b; Li & Liu, 2023) and Nyström subsampling (Li et al., 2023a). However, these studies primarily concentrate on the least square loss, while our focus lies on the KQR-RF with a non-smooth check loss function (and Lispschitz loss functions), posing additional challenges due to the lack of explicit solutions (refer to Section 4 for a detailed discussion).

Data-dependent sampling. Data-dependent sampling based on an importance ratio was initially introduced by Alaoui & Mahoney (2015) for Nyström subsampling and has been integrated into random features (Bach, 2017; Avron et al., 2017; Rudi & Rosasco, 2017; Li et al., 2021), facilitating faster learning rates with fewer random features. Despite its computational efficiency, there remains an open question regarding its impact on the theoretical results for KQR-RF (and RF for Lipschitz loss), particularly in the agnostic settings.

2. Methodology

2.1. Preambles

Consider a standard supervised learning problem that we have a sample $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{|D|}$, which are the independent copies of $(\boldsymbol{x}, y)$ drawn from an unknown joint distribution $\rho(\boldsymbol{x}, y)$ over $\mathcal{X} \times \mathbb{R}$. The τ-th conditional quantile of the response y is the minimizer of the expected quantile risk across all measurable functions, given by:

$$f_\tau^* = \operatorname*{argmin}_{f \in L^2_{\rho_\mathcal{X}}} \int_{\mathcal{X}\times\mathbb{R}} \rho_\tau\big(y - f(\boldsymbol{x})\big) d\rho(\boldsymbol{x}, y), \qquad (1)$$

where $\rho_\tau(u) = u(\tau - I\{u \le 0\})$ denotes the check loss function at τ-th quantile level with the indicator function $I(\cdot)$, and $L^2_{\rho_\mathcal{X}} = \{f : \mathcal{X} \to \mathbb{R} | \int_\mathcal{X} f^2(\boldsymbol{x}) d\rho_\mathcal{X} < \infty\}$ is the space of square integral function with respect to the distribution of the covariates $\rho_\mathcal{X}$. We also denote the $L^2_{\rho_\mathcal{X}}$-norm of f as $\|f\|_\rho^2 = \langle f, f\rangle_\rho = \int_\mathcal{X} f^2(\boldsymbol{x}) d\rho_\mathcal{X}$ for $f \in L^2_{\rho_\mathcal{X}}$. From the definition of quantile regression model, we have $P(\epsilon \le 0 | \boldsymbol{x}) = \tau$, where $\epsilon = y - f_\tau^*(\boldsymbol{x})$ is the noise term.

2.2. Kernel Quantile Regression

Kernel methods are one of the most powerful nonparametric tools, particularly for learning predictive functions within an RKHS (Vapnik, 1999). Let $\mathcal{H}_K$ denotes the RKHS induced by a symmetric, positive and semi-definite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and we define its equipped norm as $\|\cdot\|_K^2 = \langle \cdot, \cdot \rangle_K$ with the endowed inner product $\langle \cdot, \cdot \rangle_K$.

KQR estimates a function in the RKHS $\mathcal{H}_K$ by minimizing the check loss function combined with a penalty based on the squared Hilbert norm

$$f_{D,\lambda} = \operatorname*{argmin}_{f \in \mathcal{H}_K} \frac{1}{|D|} \sum_{i=1}^{|D|} \rho_\tau\big(y_i - f(\boldsymbol{x}_i)\big) + \lambda \|f\|_K^2, \quad (2)$$

where $|D|$ is the cardinality of D and λ is the regularization parameter. According to the representer theorem (Wahba, 1990), the solution of this optimization task (2) is of finite form as given by $f_{D,\lambda}(\boldsymbol{x}) = \sum_{i=1}^{|D|} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) = \boldsymbol{\alpha}^T \mathbf{K}_N(\boldsymbol{x})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{|D|})^T \in \mathbb{R}^{|D|}$ are the representer coefficients and $\mathbf{K}_N(\boldsymbol{x}) = (K(\boldsymbol{x}_1, \boldsymbol{x}), \ldots, K(\boldsymbol{x}_{|D|}, \boldsymbol{x}))^T \in \mathbb{R}^{|D|}$. With this solution plugged into (2), the optimization problem can be reformulated as

$$\hat{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^{|D|}} \frac{1}{|D|} \sum_{i=1}^{|D|} \rho_\tau\big(y_i - \boldsymbol{\alpha}^T \mathbf{K}_N(\boldsymbol{x}_i)\big) + \lambda \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha},$$

where $\mathbf{K} = \{K(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^{|D|}$ is the Gram matrix. In literature, this problem can be solved by dual optimization (Takeuchi et al., 2006; Feng et al., 2023), path-following algorithm (Li et al., 2007), and ADMM algorithm (Boyd et al., 2011; Wang et al., 2024+). However, its scalability for large datasets is limited due to the expensive computational complexity ($\mathcal{O}(|D|^3)$ and storage requirements ($\mathcal{O}(|D|^2)$ when $|D|$ is large. Consequently, a surge in research investigating scalable kernel methods and analyzing their performance has surfaced (Lin et al., 2017; Rudi et al., 2015; Rudi & Rosasco, 2017; Li et al., 2021).

2.3. KQR with Random Features

Random features prove advantageous in kernel approximation. Assuming the kernel K has an integral representation,

$$K(\boldsymbol{x}, \boldsymbol{x}') = \int_\Omega \phi(\boldsymbol{x}, \boldsymbol{\omega}) \phi(\boldsymbol{x}', \boldsymbol{\omega}) d\pi(\boldsymbol{\omega}), \qquad (3)$$

for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, where (Ω, π) is a probability space and $\phi : \mathcal{X} \times \Omega \to \mathbb{R}$, it is thus clear that we can adopt the standard Monte Carlo sampling method (Rahimi & Recht, 2007) to estimate $K(\boldsymbol{x}, \boldsymbol{x}')$ by

$$K_M(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}_M(\boldsymbol{x}, \boldsymbol{\omega}), \boldsymbol{\phi}_M(\boldsymbol{x}', \boldsymbol{\omega}) \rangle,$$

where $\boldsymbol{\phi}_M(\boldsymbol{x}, \boldsymbol{\omega}) = \frac{1}{\sqrt{M}} \big(\phi(\boldsymbol{x}, \boldsymbol{\omega}_1), \ldots, \phi(\boldsymbol{x}, \boldsymbol{\omega}_M)\big)^T$ is the feature map and $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_M$ are independently sampled with respect to π. Henceforth, we use $\boldsymbol{\phi}_M(\boldsymbol{x})$ to denote $\boldsymbol{\phi}_M(\boldsymbol{x}, \boldsymbol{\omega})$ for notation simplicity. Consequently, the solution of (2) with random features can be written as

$$f_{M,D,\lambda}(\boldsymbol{x}) = \hat{\boldsymbol{u}}^T \boldsymbol{\phi}_M(\boldsymbol{x}), \qquad (4)$$

and the optimization problem becomes

$$\hat{\boldsymbol{u}} = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^M} \frac{1}{|D|} \sum_{i=1}^{|D|} \rho_\tau\big(y_i - \boldsymbol{u}^T \boldsymbol{\phi}_M(\boldsymbol{x}_i)\big) + \lambda \boldsymbol{u}^T \boldsymbol{u}. \quad (5)$$

Notably, leveraging random features allows us to reformulate the initial problem into linear quantile regression augmented by a ridge penalty, reducing the number of parameters to be $M \ll |D|$. In our simulation study, we utilize the ADMM algorithm with the proximal operator (Boyd et al., 2011; Gu et al., 2018) to solve (5). Although random features can achieve significant success in both computation and storage by approximating the kernel, the detailed trade-off between the number of features required and the statistical prediction accuracy is still an open question, especially when the non-smooth check loss is considered and the true quantile function lies outside of the exact RKHS $\mathcal{H}_K$. This paper aims to answer these theoretical questions of KQR-RF in subsequent sections.

3. Theoretical Guarantee

In this section, we first present an existing bound for KQR-RF (Li et al., 2021, Theorem 3.8), where they focus on

the Lipschitz continuous loss family including the check loss. Subsequently, we provide our capacity-dependent and shaper learning rates for KQR-RF (Theorem 3.9), which can not only recover those of Li et al. (2021), but also can be applied to the case with the agnostic settings where the true quantile functions may not lie in the considered function space. Furthermore, we consider the data-dependent sampling strategy, which achieves the same rates (Corollary 3.16) with fewer random features and pertains its applicability to the entire agnostic settings. At last, we extend our theoretical results to a wider array of Lipschitz continuous losses with a modified local strong convexity assumption (Assumption 3.17).

The objective of KQR-RF is to find an estimator that minimizes the following expected risk

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathbb{R}} \rho_\tau\big(y - f(\boldsymbol{x})\big) d\rho(\boldsymbol{x}, y),$$

and we evaluate the performance of KRR-RF by the excess risk $\mathcal{E}(f) - \mathcal{E}(f_\tau^*)$, or the $L^2_{\rho_\mathcal{X}}$-norm of the difference $\|f - f_\tau^*\|_\rho^2$. The following are some standard definitions and assumptions needed to establish the theoretical results.

Definition 3.1 (Integral operators)**.** For any $f \in L^2_{\rho_\mathcal{X}}$, we define the integral operators by the kernel K and K_M as

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

Definition 3.2 (Effective dimension)**.** For $\lambda > 0$, we define the effective dimension of kernel K and K_M as

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

The effective dimension $\mathcal{N}(\lambda)$ serves as a common metric in kernel learning theory literature, measuring the complexity of the RKHS $\mathcal{H}_K$ (Caponnetto & De Vito, 2007; Smale & Zhou, 2007; Rudi et al., 2015; Rudi & Rosasco, 2017). Similarly, we introduce $\mathcal{N}_M(\lambda)$ as the effective dimension induced by the approximation kernel K_M. As indicated in Lemma E.7 (Rudi & Rosasco, 2017) in the appendix, $\mathcal{N}_M(\lambda)$ has been shown to be equivalent to $\mathcal{N}(\lambda)$ under mild conditions on the number of random features.

Assumption 3.3 (Bounded and continuous random features)**.** Assume kernel K has the integral representation defined in (3) with ϕ bounded and continuous in both variables, that is, there exists some constant $\kappa \geq 1$ such that $|\phi(\boldsymbol{x}, \boldsymbol{\omega})| \leq \kappa$ for any $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\omega} \in \Omega$. The associated RKHS $\mathcal{H}_K$ is separable.

Assumption 3.3 is a common condition in the literature of random features (Rudi & Rosasco, 2017; Liu et al., 2020; Li et al., 2021), which can be satisfied when the random features are continuous and bounded and $\mathcal{X}$ is separable. Note that this assumption implies that $\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{x}') \leq \kappa^2$ and $\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} K_M(\boldsymbol{x}, \boldsymbol{x}') \leq \kappa^2$.

Assumption 3.4 (Source condition)**.** Suppose there exists $R > 0$, $r > 0$ and $h_\tau \in L^2_{\rho_\mathcal{X}}$ such that

$$f_\tau^* = L_K^r h_\tau, \tag{6}$$

where $\|h_\tau\|_\rho \leq R$ and L_K^r is the r-th power of L_K.

The parameter r controls the size of the functional class of f_τ^*, denoted as $\mathcal{F} = L_K^r(L^2_{\rho_\mathcal{X}})$. According to Steinwart & Christmann (2008); Lin & Rosasco (2016), we have $\mathcal{H}_K = L_K^{1/2}(L^2_{\rho_\mathcal{X}})$, and $L_K^{r_1}(L^2_{\rho_\mathcal{X}}) \subseteq L_K^{r_2}(L^2_{\rho_\mathcal{X}})$ if $r_1 \geq r_2$. When $r \in [1/2, 1]$, the functional class $\mathcal{F}$ is a subset of the assumed RKHS $\mathcal{H}_K$, so we have $f_\tau^* \in \mathcal{H}_K$. When $r \in (0, 1/2)$, the functional class $\mathcal{F}$ is larger than the assumed RKHS $\mathcal{H}_K$, and there exists some cases where $f_\tau^* \notin \mathcal{H}_K$. Existing literature on KQR and kernel methods with Lipschitz continuous loss functions often assumes that $r = 1/2$ (Bach, 2017; Sun et al., 2018; Li et al., 2021) or $r \in [1/2, 1]$ (Lian, 2022), corresponding to the realizable setting $f_\tau^* \in \mathcal{H}_K$. However, our analysis further allows $r \in (0, 1/2)$, relating to the agnostic setting $f_\tau^* \notin \mathcal{H}_K$. This is a non-trivial extension since we consider a non-smooth loss with random feature approximation.

Assumption 3.5 (Capacity condition)**.** For $\lambda > 0$, there exists $Q > 0$ and $\gamma \in [0, 1]$ such that

$$\mathcal{N}(\lambda) \leq Q^2 \lambda^{-\gamma}. \tag{7}$$

Note that this assumption always holds when $\gamma = 1$ by taking $Q = \mathrm{Tr}(L_K) \leq \kappa^2$, and $\gamma = 0$ corresponds to some more benign cases. It is more general than the eigenvalue decay assumption (Li et al., 2021; Li, 2022; Lian, 2022), since it is satisfied when the eigenvalues $\{\mu'\}_{i\geq 1}$ of L_K have a polynomial decay, i.e., $i^{-1/\gamma}$. For KRR and KRR-RF, the minimax optimal capacity-dependent rate has been shown to be $\mathcal{O}(|D|^{\frac{2r}{2r+\gamma}})$ (Caponnetto & De Vito, 2007; Rudi & Rosasco, 2017). In the case of KQR, Lian (2022) also derive the same capacity-dependent rate $\mathcal{O}(|D|^{\frac{2r}{2r+\gamma}})$. We want to emphasize that these works mainly focus on the realizable setting with $r \in [1/2, 1]$, while our result first extends the capacity-dependent rate analysis of KQR-RF to the agnostic setting.

Assumption 3.6 (Adaptive self-calibration condition)**.** Let $f_{y|\boldsymbol{x}}(\cdot)$ denote the conditional density function of y given $\boldsymbol{x}$. Suppose that $\sup_{t\in\mathbb{R}} f_{y|\boldsymbol{x}}(t) \leq c_1$ for $c_1 > 0$. Furthermore, there exist some universal constants $\varepsilon, \varepsilon', c_2 > 0$ that are independent with $\boldsymbol{x}$ and y, such that for any $y \in \mathcal{B}(f_\tau^*(\boldsymbol{x}), \varepsilon)$ and $|\delta| \leq \varepsilon'$, the following inequality holds almost surely,

$$|F_{y|\boldsymbol{x}}(y + \delta) - F_{y|\boldsymbol{x}}(y)| \geq c_2|\delta|, \tag{8}$$

where $\mathcal{B}(f_\tau^*(\boldsymbol{x}), \varepsilon) = \{y \mid |y - f_\tau^*(\boldsymbol{x})| \leq \varepsilon\}$ denotes the ball centered at $f_\tau^*(\boldsymbol{x})$ with radius ε, and $F_{y|\boldsymbol{x}}(\cdot)$ is the cumulative distribution function of y given $\boldsymbol{x}$.

Assumption 3.6 serves as an adaptive self-calibration condition for the conditional distribution of y given $\boldsymbol{x}$. It is a mild condition intended to hold for most realistic sequences of distributions. For example, if y has a density that is bounded away from zero on some compact interval around $f_\tau^*(\boldsymbol{x})$, then Assumption 3.6 holds. More importantly, we do not impose any moment condition on the distribution of y. It is also worth noting that Assumption 3.6 is weaker than Condition 2 in He & Shi (1994) where the density function of y is lower bounded everywhere by some positive constant. It is also weaker than Condition D.1 in Belloni & Chernozhukov (2011) requiring the conditional density of Y given $\boldsymbol{x}$ to be continuously differentiable and bounded away from zero uniformly for all $\tau \in (0, 1)$ and all $\boldsymbol{x}$ in the support $\mathcal{X}$. The special case when $\varepsilon = 0$ aligning with the self-calibration condition also appeared in Shen et al. (2021); Madrid Padilla & Chatterjee (2022).

Remark 3.7. This adaptive self-calibration condition plays a pivotal role in our novel error decomposition as shown in Lemma B.3 of the appendix, which leads to an adaptive local strong convexity condition of the expected check loss in a small ball around f_τ^*. It is worth noting that the self-calibration condition is weaker than Assumption (A2') of Lian (2022) and Assumption (B2) of Li et al. (2021) where the conditional density of y given $\boldsymbol{x}$ is assumed to be bounded away from zero across all quantile levels and $\boldsymbol{x} \in \mathcal{X}$. Under this assumption, we derive a tight bound for a novel least square approximation (LS-approximation) error between the KQR-RF estimator $f_{M,D,\lambda}$ and its KRR-RF approximation estimator $f^\diamond_{M,D,\lambda}$, detailed in Lemma B.8 of the appendix.

3.1. Existing Learning Rates for KQR-RF

To facilitate a clear comparison between our findings and existing results, we first introduce the best learning rates so far for KQR-RF (Li et al., 2021).

Theorem 3.8 (Existing learning rates for KQR-RF (random features with Lipschitz loss), Theorem 19 of Li et al. (2021))**.** *Assume there exists a function $f_\mathcal{H}$ such that $f_\mathcal{H} = \operatorname{argmin}_{f \in \mathcal{H}_K} \mathcal{E}(f)$. Under some technical assumptions*[1]*, and $\lambda = \mathcal{O}(|D|^{-1})$, when the number of random features satisfies*

$$M \gtrsim |D|^{\frac{\gamma}{2}} \log |D|,$$

and $|D|$ is sufficiently large, there holds

$$\mathcal{E}(f_{M,D,\lambda}) - \mathcal{E}(f_\mathcal{H}) \asymp \|f_{M,D,\lambda} - f_\mathcal{H}\|_\rho^2 = \mathcal{O}(|D|^{-\frac{1}{2}}),$$

with probability near to 1.

[1]Assumption 3.3, Assumption 3.4 with $r = 1/2$, eigenvalue decaying assumption (stronger than Assumption 3.5), and the local strongly convex assumption which can be derived from Assumption 3.6.

Theorem 3.8 establishes an upper bound for KQR-RF in the worst case, requiring only the existence of $f_\mathcal{H}$. In this scenario, if the number of random features scales as $|D|^{\frac{\gamma}{2}} \log |D|$, KQR-RF can achieve the capacity-independent optimal generalization properties. This represents a significant improvement over previous work, which required a larger number of random features to guarantee similar learning rates. Note that Rahimi & Recht (2008) proved $\mathcal{O}(|D|)$ random features to guarantee the learning rates at $\mathcal{O}(|D|^{-\frac{1}{2}})$. However, these results are capacity-independent and can not apply to the agnostic setting when the size of RKHS is small. In our subsequent analysis, we will present a sharper and capacity-dependent learning rate, allowing $r \in (0, 1]$, which covers the entire source condition space. This particularly marks the primary novelty and advancement in the theoretical understanding of KQR-RF.

3.2. Sharper Learning Rates for KQR-RF

Theorem 3.9 (Worst case)**.** *Under Assumptions 3.3-3.6, if $r \in (0, 1]$, $\gamma \in [0, 1]$, and set $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies*

$$M \gtrsim |D|^{\frac{1}{2r+\gamma}}, \quad \textit{for} \quad r \in (0, 1/2);$$
$$M \gtrsim |D|^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, \quad \textit{for} \quad r \in [1/2, 1],$$

and $|D|$ is sufficiently large, there holds

$$\mathcal{E}(f_{M,D,\lambda}) - \mathcal{E}(f_\tau^*) \asymp \|f_{M,D,\lambda} - f_\tau^*\|_\rho^2 = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log^2 |D|),$$

with probability near to 1.

The capacity-dependent learning rates obtained in Theorem 3.9 align with those of KRR (Caponnetto & De Vito, 2007) and KRR-RF (Rudi & Rosasco, 2017), which is minimax optimal and thus can not be improved any further. Specifically, in scenario of highest regularity ($r = 1$) and a small RKHS ($\gamma = 0$), it approaches the standard parametric bound $\mathcal{O}(1/|D|)$. For $r = 1/2$ and $\gamma = 1$, corresponding to the worst case, our learning rates and the requirements on the number of the random features match those in Theorem 3.8. More interestingly, our results extend the optimal learning rates to the agnostic case where the true quantile function is located outside of the RKHS $\mathcal{H}_K$. Specifically, we relax the regularity condition from $r \in [1/2, 1]$ to $r \in (0, 1]$, covering a wider range of scenarios.

Remark 3.10. Recent studies have explored the generalization performance of kernel-based methods in the agnostic setting, including kernel ridge regression (Zhang et al.,

2023), kernel ridge regression with Nyström subsampling (Lu et al., 2019; Li et al., 2023a), and kernel ridge regression with random features (Li et al., 2023b; Li & Liu, 2023). However, these studies primarily focus on the least square loss, contrasting with our work that delves into more complex non-smooth check loss and a broader Lipschitz loss family. Our theory requires a distinct set of proof techniques compared to the work grounded in the least square loss paradigm which has an implicit solution, necessitating the use of the empirical process. Specifically, we introduce a novel error decomposition including an LS-approximation error term, which bridges the excess risk for the check loss with the $L^2_{\rho_\mathcal{X}}$ error of an intermediate estimator $f^\diamond_{M,D,\lambda}$ (see details in Lemmas B.3 and B.8 of the appendix). To derive the faster learning rates for both realizable and agnostic settings, we use different technical skills to take the regularity condition into the LS-approximation error term, such as the non-trivial Young's inequality and Cauthy-Schwarz inequality tailored for operators.

Remark 3.11. Theorem 3.9 broadens the regularity condition for optimal learning rates from $r \in [1/2, 1]$ to $(0, 1]$. However, it uses the naive uniform sampling strategy for the random features (generate $\phi(\boldsymbol{x}, \boldsymbol{\omega})$ with $\pi(\boldsymbol{\omega})$), which is independent of the training samples. This may lead to an unnecessary burden in computation. Inspired by the data-dependent sampling strategy (Bach, 2017; Avron et al., 2017; Rudi & Rosasco, 2017), we aim to demonstrate in the upcoming section how these strategies enable attaining optimal learning rates across all agnostic settings $r \in (0, 1]$ with a reduced number of random features in the next section.

3.3. Refined Analysis: Beyond Uniform Sampling

To obtain shaper learning rates for the entire setting $r \in (0, 1]$ with fewer random features, we first introduce a compatibility condition that is commonly used in the literature (Rudi et al., 2015; Rudi & Rosasco, 2017; Liu et al., 2020).

Assumption 3.12 (Compatibility condition)**.** Define the maximum dimension of random features as

$$\mathcal{N}_\infty(\lambda) = \sup_{\boldsymbol{\omega}\in\Omega} \left\| (L_K + \lambda I)^{-1/2} \phi(\cdot, \boldsymbol{\omega}) \right\|^2_{\rho_\mathcal{X}}, \quad (9)$$

where $\lambda > 0$. There exist constants $\alpha \in [0, 1]$ and $F > 0$, such that $\mathcal{N}_\infty(\lambda) \le F\lambda^{-\alpha}$.

The maximum dimension of random features in (9) correlates with the data-generating distribution through the integral operator L_K, which is always satisfied for $\alpha = 1$ and $F = \kappa^2$. Recall the definition of $\mathcal{N}(\lambda)$ in Definition 3.2. $\mathcal{N}(\lambda)$ and $\mathcal{N}_\infty(\lambda)$ measure the average and supreme capacities of $\mathcal{H}_K$, respectively, so we have $\mathcal{N}(\lambda) = E_{\boldsymbol{\omega}} \left\| (L_K + \lambda I)^{-1/2} \phi(\cdot, \boldsymbol{\omega}) \right\|^2_{\rho_\mathcal{X}} \le \sup_{\boldsymbol{\omega}\in\Omega} \left\| (L_K + \lambda I)^{-1/2} \phi(\cdot, \boldsymbol{\omega}) \right\|^2_{\rho_\mathcal{X}} = \mathcal{N}_\infty(\lambda)$, where $E_{\boldsymbol{\omega}}$ denotes the expectation taking over $\boldsymbol{\omega}$.

Theorem 3.13. *Under Assumptions 3.3-3.6 and 3.12, if $r \in (0, 1]$, $\gamma \in [0, 1]$, and set $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies*

$$M \gtrsim |D|^{\frac{\alpha}{2r+\gamma}}, \quad \textit{for} \quad r \in (0, 1/2);$$
$$M \gtrsim |D|^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}}, \quad \textit{for} \quad r \in [1/2, 1],$$

and $|D|$ is sufficiently large, there holds

$$\mathcal{E}(f_{M,D,\lambda}) - \mathcal{E}(f^*_\tau) \asymp \|f_{M,D,\lambda} - f^*_\tau\|^2_\rho = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log^2 |D|),$$

with probability near to 1.

The above capacity-dependent learning rate is the same as that of Theorem 3.9, while the required number of random features reduces from $\mathcal{O}(|D|^{\frac{1}{2r+\gamma}})$ to $\mathcal{O}(|D|^{\frac{\alpha}{2r+\gamma}})$ when $r \in (0, 1/2)$ and $\mathcal{O}(|D|^{\frac{(2r-1)\gamma+1}{2r+\gamma}})$ to $\mathcal{O}(|D|^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}})$ when $r \in [1/2, 1]$, owing to the additional imposition of the compatibility condition $\mathcal{N}_\infty(\lambda) \le F\lambda^{-\alpha}$. By adopting a favorable sampling strategy, as demonstrated in Example 3.14, we can further reduce the required number of random features and achieve the optimal learning rates across the entire range of $r \in (0, 1]$.

Example 3.14 (Leverage scores sampling)**.** Given the integral representation of kernel K as stated in (3), we adopt the leverage scores sampling strategy (Bach, 2017; Avron et al., 2017) by employing an importance ratio denoted as $q(\boldsymbol{\omega}) = l_\lambda(\boldsymbol{\omega}) / \int_{\boldsymbol{\omega}} l_\lambda(\boldsymbol{\omega}) d\pi(\boldsymbol{\omega})$, where $l_\lambda(\boldsymbol{\omega}) = \|(L_K + \lambda I)^{-1/2} \phi(\cdot, \boldsymbol{\omega})\|^2_{\rho_\mathcal{X}}$. Consequently, the random features are computed as $\phi_l(\boldsymbol{x}, \boldsymbol{\omega}) = [q(\boldsymbol{\omega})]^{-1/2} \phi(\boldsymbol{x}, \boldsymbol{\omega})$ and exhibit a distribution $\pi_l(\boldsymbol{\omega}) = q(\boldsymbol{\omega})\pi(\boldsymbol{\omega})$. As pointed out in Rudi & Rosasco (2017), the random features provide the integral representation of K and satisfy Assumption 3.12 with $\alpha = \gamma$ indicating that $\mathcal{N}(\lambda) = \mathcal{N}_\infty(\lambda)$.

Remark 3.15. We call $\alpha = 1$ as the worst case (Theorem 3.9) when considering the random features with uniform sampling in (3) which is independent of the training samples, and $\alpha = \gamma$ as the benign case (Corollary 3.16) when adopting the data-dependent sampling strategy in Example 3.14.

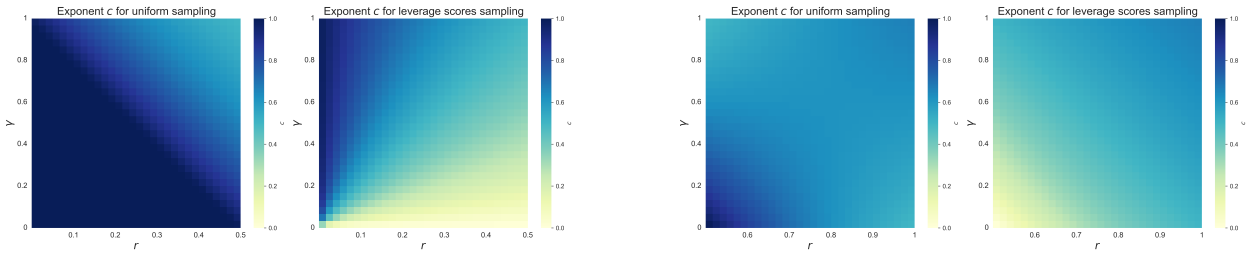
Corollary 3.16 (Benign case)**.** *Under Assumptions 3.3-3.6, if random features are sampled according to the strategy in Example 3.14, $r \in (0, 1]$, $\gamma \in [0, 1]$, and set $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies*

$$M \gtrsim |D|^{\frac{\gamma}{2r+\gamma}}, \quad \textit{for} \quad r \in (0, 1/2);$$
$$M \gtrsim |D|^{\frac{2r+\gamma-1}{2r+\gamma}}, \quad \textit{for} \quad r \in [1/2, 1],$$

and $|D|$ is sufficiently large, then there holds

$$\mathcal{E}(f_{M,D,\lambda}) - \mathcal{E}(f^*_\tau) \asymp \|f_{M,D,\lambda} - f^*_\tau\|^2_\rho = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log^2 |D|),$$

with probability near to 1.



(a) Agnostic case

(b) Realizable case

Figure 1. Comparison between the number of random features $M = \mathcal{O}(|D|^c)$ required for uniform sampling ($\alpha = 1$, left) and leverage scores sampling ($\alpha = \gamma$, right), Figure 1(a) is the agnostic case and Figure 1(b) is the realizable case, respectively.

Theorem 3.9 is the worst case of Theorem 3.13 with $\alpha = 1$, while Corollary 3.16 is the benign case of Theorem 3.13 with $\alpha = \gamma$. This distinction arises from the choice of the uniform sampling strategy $\pi(\boldsymbol{\omega})$, which typically yields an approximate estimate where α tends to 1. Conversely, employing data-dependent random features assures a favorable scenario where $\alpha = \gamma$. To better illustrate the computational improvement for different cases, we depict a comparison in Figure 1(a) and 1(b) between the number of random features required to ensure the optimal learning rates using uniform sampling (left panel) and data-dependent sampling (right panel) for the agnostic case when $r \in (0, 1/2)$ and the realizable case when $r \in [1/2, 1]$, respectively.

3.4. Extension to Lipschitz Loss

Note that the check loss belongs to the family composed of Lipschitz continuous losses. We aim to extend our theoretical results to the general Lipschitz continuous loss family, including other kernel-based methods, such as kernel support vector machines (Sun et al., 2018, KSVM) and kernel logistic regression (Keerthi et al., 2005, KLR). Similar to the quantile regression estimation in (5), we formulate the following general learning problem

$$\widetilde{\boldsymbol{u}} = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^M} \frac{1}{|D|}\sum_{i=1}^{|D|} L\big(y_i, \boldsymbol{u}^T\boldsymbol{\phi}_M(\boldsymbol{x}_i)\big) + \lambda \boldsymbol{u}^T\boldsymbol{u},$$

where $L(y, \cdot)$ is a Lipschitz continuous loss such that for some $V \geq 0$, there exists a constant $c_L > 0$ such that $|L(y, x) - L(y, x')| \leq c_L|x - x'|$ holds for all pairs $x, x' \in [-V, V]$ and $y \in \mathbb{R}$. We can refer to Feng et al. (2023) for more specific examples satisfying this property.

Our objective is to replace the check loss ρ_τ with some general Lipschitz continuous loss and construct a unified theoretical framework. In our proof of the main theorems for KQR-RF, a pivotal step involves controlling the LS-approximation error therm in Lemmas B.8 and B.9. To facilitate this, we merely need to substitute Assumption 3.6 with the following substantial assumption.

Assumption 3.17 (Local strong convexity)**.** There exist some constants $u, u', c_3, c_4 > 0$ such that for any f and f' satisfying $\|f - f'\|_\rho \leq u$ and $\|f' - f^*\|_\rho \leq u'$, there holds

$$\mathcal{E}_L(f) - \mathcal{E}_L(f') \geq c_3\|f - f'\|_\rho^2, \tag{10}$$

or

$$\mathcal{E}_L(f) - \mathcal{E}_L(f') + \|f' - f^*\|_\rho^2 \geq c_4\|f - f'\|_\rho^2, \tag{11}$$

where $\mathcal{E}_L(f) = E(L(y, f(\boldsymbol{x})))$ and $f^* = \operatorname{argmin}_f \mathcal{E}_L(f)$. Here, we refer to (10) as the local strong convexity of $L(y, f)$, and (11) as the adaptive local strong convexity of $L(y, f)$.

It is worth pointing out that we can verify conditions (10) and (11) for the check loss $\rho_\tau(\cdot)$ by using (8) in Assumption 3.6 with $\varepsilon = 0$ and $\varepsilon \neq 0$, respectively. With Assumption 3.6 replaced by Assumption 3.17 and keeping all other conditions unchanged, we can similarly establish the same learning rates for the Lipschitz loss L. Specifically,

$$\|f^L_{M,D,\lambda} - f^*\|_\rho^2 = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log^2|D|),$$

with probability near to 1, where $f^L_{M,D,\lambda} = \widetilde{\boldsymbol{u}}^T\phi_M$. The detailed results for Lipschitz continuous loss and their proofs are deferred to Appendix C due to the space limit.

4. Comparisons to Related Work

In this section, we compare the conditions and learning rates of our method with related existing approaches including KRR, KRR-RF, KQR, and Lipschitz loss with RF, which are summarized in Table 1, where "Uniform" and "Leverage" stand for the uniformly and leverage scores sampling strategies, respectively, and "Lip" is short for Lipschitz continuous loss. Note that under Assumption 3.17, the results in the last three lines of Table 1 can also be directly extended to the cases with the general Lipschitz continuous losses.

Compared to KRR and its RF variants. Previous studies have extensively pursued the optimal learning rates for

Table 1. Summary of conditions for derived learning rates in different methods.

Methods	Regularity condition	Capacity condition	Random centers M	Learning rate
KRR (Caponnetto & De Vito, 2007)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$\times$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KRR (Zhang et al., 2023)	$r \in (0, 1]$	$\gamma \in [0, 1]$	$\times$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KRR-RF-Uniform (Rudi & Rosasco, 2017)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$\lvert D\rvert^{-\frac{(2r-1)\gamma+1}{2r+\gamma}}$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KRR-RF-Leverage (Rudi & Rosasco, 2017)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$\lvert D\rvert^{-\frac{2r+\gamma-1}{2r+\gamma}}$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KRR-RF-Uniform (Li et al., 2023b)	$r \in (0, 1], 2r+\gamma \geq 1$	$\gamma \in [0, 1]$	$\lvert D\rvert^{-\frac{1}{2r+\gamma}}$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KRR-RF-Leverage (Li et al., 2023b)	$r \in (0, 1]$	$\gamma \in [0, 1]$	$\lvert D\rvert^{-\frac{\gamma}{2r+\gamma}}$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KQR (Lian, 2022)	$r \in [1/2, 1]$	$\gamma \in [0, 1]$	$\times$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
Lip-RF-Uniform (Rahimi & Recht, 2008)	$r = 1/2$	$\gamma \in [0, 1]$	$\lvert D\rvert$	$\lvert D\rvert^{-1/2}$
Lip-RF-Leverage (Bach, 2017)	$r = 1/2$	$\gamma \in [0, 1]$	$\lvert D\rvert^{\frac{\gamma}{2}}$	$\lvert D\rvert^{-1/2}$
Lip-RF-Uniform (Li et al., 2021)	$r = 1/2$	$\gamma \in [0, 1]$	$\lvert D\rvert$	$\lvert D\rvert^{-1/2}$
Lip-RF-Leverage (Li et al., 2021)	$r = 1/2$	$\gamma \in [0, 1]$	$\lvert D\rvert^{\frac{\gamma}{2}}$	$\lvert D\rvert^{-1/2}$
KSVM-RF (Sun et al., 2018)	$r = 1/2$	$\gamma \in [0, 1]$	$\lvert D\rvert^{\frac{2\gamma}{2\gamma+1}}$	$\lvert D\rvert^{-\frac{1}{2\gamma+1}}$
KQR-RF (Theorem 3.13)	$r \in (0, 1]$	$\gamma \in [0, 1]$	$\lvert D\rvert^{\frac{\alpha}{2r+\gamma}}, r \in (0, 1/2)$ $\lvert D\rvert^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}}, r \in [1/2, 1]$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KQR-RF-Uniform (Theorem 3.9)	$r \in (0, 1]$	$\gamma \in [0, 1]$	$\lvert D\rvert^{\frac{1}{2r+\gamma}}, r \in (0, 1/2)$ $\lvert D\rvert^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, r \in [1/2, 1]$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$
KQR-RF-Leverage (Corollary 3.16)	$r \in (0, 1]$	$\gamma \in [0, 1]$	$\lvert D\rvert^{\frac{\gamma}{2r+\gamma}}, r \in (0, 1/2)$ $\lvert D\rvert^{\frac{2r+\gamma-1}{2r+\gamma}}, r \in [1/2, 1]$	$\lvert D\rvert^{-\frac{2r}{2r+\gamma}}$

KRR (Caponnetto & De Vito, 2007; Smale & Zhou, 2007) and KRR-RF (Rudi & Rosasco, 2017; Avron et al., 2017). Recent extensions (Zhang et al., 2023; Li et al., 2023b; Li & Liu, 2023) have enlarged the regularity condition to the agnostic setting when the regression function lies outside of the RKHS. However, we focus on KQR-RF with the non-smooth check loss, which is more challenging since we have no explicit solutions. Notably, deriving a capacity-dependent learning rate for $r \in (0, 1/2)$ demands distinct technical skills compared to those required for KRR and its RF variations. Moreover, our results can be easily extended to the Lipschitz losses with a modified assumption, signifying the added novelty of our analysis.

Compared to kernel methods with Lipschitz loss and their RF variants. Existing literature for random features with Lipschitz loss (Rahimi & Recht, 2008; Sun et al., 2018; Li et al., 2021; Li, 2022) only consider the ideal case when $r = 1/2$, and their learning rates are either capacity-independent (Rahimi & Recht, 2008; Li et al., 2021) or sub-optimal (Sun et al., 2018). Lian (2022) studied the capacity-dependent learning rate for KQR when $r \in [1/2, 1]$. However, their work can not be directly applied to the random feature setting. In contrast, our study offers a comprehensive analysis of the capacity-dependent learning rates for KQR-RF (Lip-RF), exhibiting broader applicability across scenarios where the true regression function resides in the entire agnostic setting.

We also provide a brief proof sketch to emphasize the theoretical contributions of this paper.

A novel error decomposition and least square approximation $\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_\rho$. Unlike existing RF work (Rudi & Rosasco, 2017; Li et al., 2021; 2023b), we first introduce a novel error decomposition in Lemma B.3. Except for the standard empirical, RF approximation, and kernel approximation errors, we have an extra least square approximation (LS-approximation) error term $\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_\rho$. By the adaptive self-calibration assumption, we build an adaptive local strong convexity condition of the expected loss on a small neighborhood of f^*_τ. This promises the convergence of the LS-approximation error term. We also use the non-trivial Cauthy-Schwarz and Young's inequalities to take into account the source index when $r \in (0, 1/2)$ and $r \in [1/2, 1]$, respectively.

Sharper analysis for $\|f^{\diamond}_{M,D,\lambda} - f^*_\tau\|_\rho$. As indicated in Lemma B.3, we divide the $\|f^{\diamond}_{M,D,\lambda} - f^*_\tau\|_\rho$ into three terms. To get tighter bounds of the empirical errors $\|f^{\diamond}_{M,D,\lambda} - f_{M,\lambda}\|_\rho$ and $\|f_{M,\lambda} - f_\lambda\|_\rho$, we utilize the compatibility condition to Bernstein's inequalities among operators L_K, L_M, and $C_M, C_{M,D}$. This refined procedure helps to relax the conditions on M and $|D|$, and further enlarges the regularity condition to $r \in (0, 1]$. In fact, the convergence of term $\|f^{\diamond}_{M,D,\lambda} - f^*_\tau\|_\rho$ is also an important premise for the convergence of the LS-approximation error $\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_\rho$.

5. Numerical Experiments

Inspired by the simulation setup in Rudi & Rosasco (2017); Li et al. (2021), we also consider the spline kernel of order

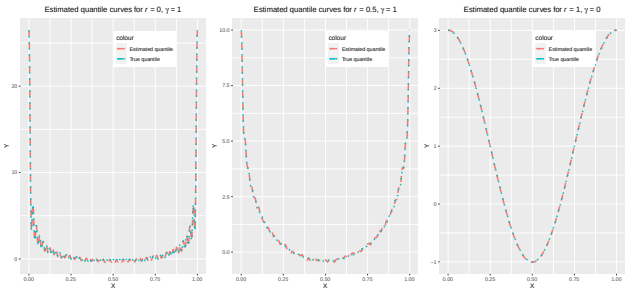


Figure 2. Estimated and true quantile curves for $r = 0, \gamma = 1$ (left), $r = 1/2, \gamma = 1$ (middle), and $r = 1, \gamma = 0$ (right) when $\tau = 0.5$.

q, defined as $\Lambda_q(x, x') = \sum_{k=-\infty}^{\infty} e^{2\pi ikx} e^{-2\pi ikx'} |k|^{-q}$, where $x, x' \in [0, 1]$, and $q \in \mathbb{R}$. According to the property of spline kernel, we have $\int_0^1 \Lambda_q(x, z)\Lambda_{q'}(x', z)\, dz = \Lambda_{q+q'}(x, x')$, for any $q, q' \in \mathbb{R}$. Consequently, for $r \in (0, 1]$ and $\gamma \in [0, 1]$, let $K(x, x') = \Lambda_{\frac{1}{\gamma}}(x, x')$, and its corresponding random feature is $\phi(x, w) = \Lambda_{\frac{1}{2\gamma}}(x, w)$ with $w \sim U(0, 1)$. We consider the model $y = \Lambda_{\frac{r}{\gamma}+\frac{1}{2}}(x, 0) + \varepsilon$, where $\varepsilon \sim N(0, 0.01)$ and $x \sim U(0, 1)$. Then Assumptions 3.3-3.5 and 3.11 are satisfied and $\alpha = \gamma$ (Rudi & Rosasco, 2017). To graphically show the true and estimated quantile function, we consider three different settings: (1) worst case ($r = 0, \gamma = 1$); (2) general case ($r = 1/2, \gamma = 1$); (3) most benign case ($r = 1, \gamma = 0$). Without loss of generality, we fix $\tau = 0.5$. We generate training data with size $N_{tr} = 1000$, and testing data with size $N_{te} = 10000$. The regularization parameter λ is selected via a grid search based on a validation set with 1000 samples, where the grid is set as $\{10^{0.5s} : s = -20, -19, \ldots, 2\}$. The number of random features is selected according to Theorem 3.13. The estimated and true quantile curves on the testing data are shown in Figure 2. From the results, we can conclude that KQR-RF can estimate the quantile functions very well both in realizable and agnostic settings.

To validate the derived learning rates, i.e., $\mathcal{E}(f_{M,D,\lambda}) - \mathcal{E}(f_\tau^*) = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}})$, we estimate the excess risk on the testing data and compared it with the theoretical one. We consider two agnostic cases ($r = 0.2, \gamma = 0.1$ and $r = 0.4, \gamma = 0.2$) and two realizable cases ($r = 0.5, \gamma = 0.1$ and $r = 0.8, \gamma = 0.2$) for better illustration. The setting is the same as the above except that the training data size varies in $\{1000, 2000, \ldots, 10000\}$. We perform a log transform on the empirical excess risk and the number of training data and plot them in Figure 3. From the results, we can see that the data points are uniformly distributed on both sides of a straight line, which verifies the derived learning rate. To further investigate the constants in the big-$\mathcal{O}$ bounds, we calculate the slope of each learning curve and compare it to $-\frac{2r}{2r+\gamma}$. The slope constants are $0.81, 1.21, 1.63, 0.95$ in four scenarios. This also highlights our contribution in deriving the sharper and capacity-dependent learning rates.

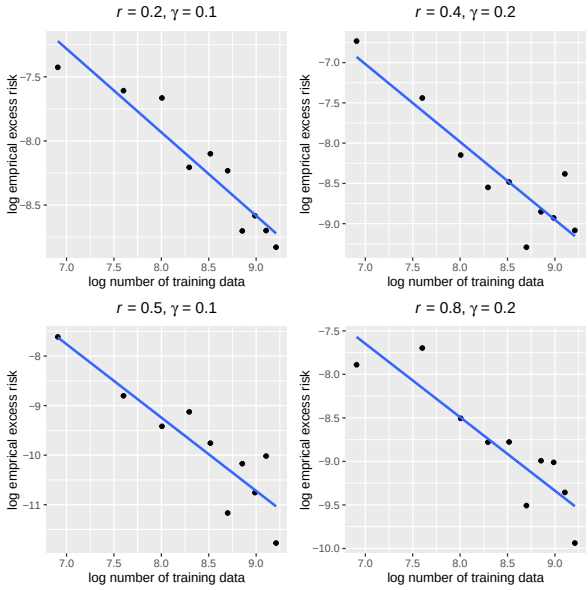


Figure 3. Log empirical excess risk for $r = 0.2, \gamma = 0.1$ (left top), $r = 0.4, \gamma = 0.2$ (right top), $r = 0.5, \gamma = 0.1$ (left bottom) and $r = 0.8, \gamma = 0.2$ (right bottom) when $\tau = 0.5$.

6. Conclusion

This paper investigates kernel quantile regression with random features and derives capacity-dependent optimal learning rates for both realizable and agnostic settings. By introducing a modified local strong convexity assumption, our theoretical analysis seamlessly extends to the entire Lipschitz continuous loss family, leading to the sharpest result so far to our best knowledge. Extensive experiments are conducted on both simulated and real case studies, providing empirical evidence that supports the theoretical findings of our paper. Furthermore, it is feasible to extend the theoretical results of random features to incorporate other accelerated approaches, such as stochastic gradient methods or distributed techniques, or consider the parallel problem in deep over-parameterized quantile regression.

Acknowledgement

The authors thank the area chair and the anonymous referees for their constructive suggestions, which significantly improved this article. This research is supported in part by NSFC-12371270 and Shanghai Science and Technology Development Funds (23JC1402100). This research is also supported by Shanghai Research Center for Data Science and Decision Technology.

Impact Statement

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

References

Abedsoltan, A., Belkin, M., and Pandit, P. Toward large kernel models. In *International Conference on Machine Learning*, pp. 61–78. PMLR, 2023.

Alaoui, A. and Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. *Advances in Neural Information Processing Systems*, 28:775–783, 2015.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

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.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Belloni, A. and Chernozhukov, V. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82, 2011.

Blanchard, G. and Krämer, N. Optimal learning rates for kernel conjugate gradient regression. *Advances in Neural Information Processing Systems*, 23:226–234, 2010.

Bochner, S. *Harmonic analysis and the theory of probability*. Courier Corporation, 2005.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

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

Carratino, L., Rudi, A., and Rosasco, L. Learning with sgd and random features. *Advances in Neural Information Processing Systems*, 31:10213–10224, 2018.

Daniely, A., Frostig, R., and Singer, Y. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in Neural Information Processing Systems*, 29:2261–2269, 2016.

Feng, X., He, X., Wang, C., Wang, C., and Zhang, J. Towards a unified analysis of kernel-based methods under covariate shift. *Advances in Neural Information Processing Systems*, 36:73839–73851, 2023.

Furuta, T. *Invitation to linear operators: From matrices to bounded linear operators on a Hilbert space*. CRC Press, 2001.

Gu, Y., Fan, J., Kong, L., Ma, S., and Zou, H. Admm for high-dimensional sparse penalized quantile regression. *Technometrics*, 60(3):319–331, 2018.

He, X. and Shi, P. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journaltitle of Nonparametric Statistics*, 3(3-4):299–308, 1994.

Keerthi, S. S., Duan, K., Shevade, S. K., and Poo, A. N. A fast dual algorithm for kernel logistic regression. *Machine Learning*, 61:151–165, 2005.

Koenker, R. *Quantile regression*, volume 38. Cambridge University Press, 2005.

Li, J. and Liu, Y. Towards sharp analysis for distributed learning with random features. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3920–3928, 2023.

Li, J., Liu, Y., and Wang, W. Optimal convergence rates for agnostic nyström kernel learning. In *International Conference on Machine Learning*, pp. 19811–19836. PMLR, 2023a.

Li, J., Liu, Y., and Wang, W. Optimal convergence for agnostic kernel learning with random features. *IEEE Transactions on Neural Networks and Learning Systems*, 28:1–11, 2023b.

Li, Y., Liu, Y., and Zhu, J. Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

Li, Z. Sharp analysis of random fourier features in classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7444–7452, 2022.

Li, Z., Ton, J.-F., Oglic, D., and Sejdinovic, D. Towards a unified analysis of random fourier features. *The Journal of Machine Learning Research*, 22(1):4887–4937, 2021.

Lian, H. Distributed learning of conditional quantiles in the reproducing kernel hilbert space. *Advances in Neural Information Processing Systems*, 35:11686–11696, 2022.

Lin, J. and Cevher, V. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *The Journal of Machine Learning Research*, 21(1):5852–5914, 2020.

Lin, J. and Rosasco, L. Optimal learning for multi-pass stochastic gradient methods. *Advances in Neural Information Processing Systems*, 29:4563–4571, 2016.

Lin, S.-B., Guo, X., and Zhou, D.-X. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Lin, S.-B., Wang, D., and Zhou, D.-X. Distributed kernel ridge regression with communications. *The Journal of Machine Learning Research*, 21:3718–3755, 2020.

Liu, F., Huang, X., Chen, Y., and Suykens, J. A. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.

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

Liu, Y., Liu, J., and Wang, S. Effective distributed learning with random features: Improved bounds and algorithms. In *International Conference on Learning Representations*, pp. 1–13, 2020.

Lu, S., Mathé, P., and Pereverzyev Jr, S. Analysis of regularized nyström subsampling for regression functions of low smoothness. *Analysis and Applications*, 17(06):931–946, 2019.

Madrid Padilla, O. H. and Chatterjee, S. Risk bounds for quantile trend filtering. *Biometrika*, 109(3):751–768, 2022.

Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N., and Kong, L. Random feature attention. In *International Conference on Learning Representations (ICLR 2021)*, 2021.

Pollard, D. *Convergence of stochastic processes*. Springer Science & Business Media, 2012.

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.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. *Advances in Neural Information Processing Systems*, 30:3215–3225, 2017.

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

Rudi, A., Carratino, L., and Rosasco, L. Falkon: An optimal large scale kernel method. *Advances in Neural Information Processing Systems*, 30:3891–3901, 2017.

Rudin, W. *Fourier analysis on groups*. Courier Dover Publications, 2017.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., and Huang, J. Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*, 2021.

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

Sriperumbudur, B. and Szabó, Z. Optimal rates for random fourier features. *Advances in Neural Information Processing Systems*, 28:1144–1152, 2015.

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*, 31:3379–3388, 2018.

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

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. Nonparametric quantile estimation. *The Journal of Machine Learning Research*, 7:1231–1264, 2006.

Vapnik, V. *The nature of statistical learning theory*. Springer Science & Business Media, 1999.

Wahba, G. *Spline models for observational data*. SIAM, 1990.

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

Wang, C., Li, T., Zhang, X., Feng, X., and He, X. Communication-efficient nonparametric quantile regression via random features. *Journal of Computational and Graphical Statistics*, to appear, 2024+.

Williams, C. and Seeger, M. Using the nystroem method to speed up kernel machines. *Advances in Neural Information Processing Systems*, pp. 682–688, 2001.

Zambon, D., Alippi, C., and Livi, L. Graph random neural features for distance-preserving graph representations. In *International Conference on Machine Learning*, pp. 10968–10977. PMLR, 2020.

Zandieh, A., Han, I., Avron, H., Shoham, N., Kim, C., and Shin, J. Scaling neural tangent kernels via sketching and random features. *Advances in Neural Information Processing Systems*, 34:1062–1073, 2021.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, H., Li, Y., Lu, W., and Lin, Q. On the optimality of misspecified kernel ridge regression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 41331–41353. PMLR, 2023.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

A. Random features in kernel methods and deep neural networks

Random features mapping is a powerful tool for scaling up kernel methods (Rudi & Rosasco, 2017), neural tangent kernel (Zandieh et al., 2021), graph neural networks (Zambon et al., 2020) and attention in Transformers (Peng et al., 2021). In fact, random features can be viewed as a class of two-layer neural networks with fixed weights in their first layer (Liu et al., 2021). For example, we consider a two-layer neural network, i.e., $f(\boldsymbol{x}, \boldsymbol{\theta}) = \sqrt{\frac{2}{M}} \sum_{j=1}^{M} \alpha_j \sigma(\boldsymbol{\omega}_j^T \boldsymbol{x})$ for some activation function σ, where $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{\omega} \sim N(0, \mathbf{I}_d)$. Its corresponding random features mapping is $k(\boldsymbol{x}, \boldsymbol{x}') = E_{\boldsymbol{\omega}} \left[\sigma\left(\boldsymbol{\omega}^T \boldsymbol{x}\right) \sigma\left(\boldsymbol{\omega}^T \boldsymbol{x}'\right)\right]$. If the commonly used ReLU activation $\sigma(x) = \max\{0, x\}$ is adopted, then the kernel is the first order arc-cosine kernel, i.e., $k\left(\boldsymbol{x}, \boldsymbol{x}'\right) \equiv \kappa_1(u) = \frac{1}{\pi}\left(u(\pi - \arccos(u)) + \sqrt{1-u^2}\right)$ with $u = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle / \left(\|\boldsymbol{x}\| \, \|\boldsymbol{x}''\|\right)$. This relationships helps to explain phenomena such as the fit the random labels (Zhang et al., 2021) and double descent (Belkin et al., 2019) in the two-layer overparameterized neural networks (Arora et al., 2019). For a deep neural network with more than two layers and fixed weights except for the output layer, we can also find a compositional kernel with its widths tending to infinity (Daniely et al., 2016). In view of the connection of random features for kernel methods and neural networks, it is meaningful to study the generalization properties of random features in over-parameterized quantile neural networks by those in KQR, especially in the agnostic setting.

B. Proofs of the Learning Rate KQR-RF

To start with, we define a M-dimensional function space $\mathcal{H}_M$ related to $\boldsymbol{\phi}_M(\boldsymbol{x})$ as

$$\mathcal{H}_M = \left\{ f \mid f(\boldsymbol{x}) = \boldsymbol{u}^T \phi_M(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}, \boldsymbol{u} \in \mathbb{R}^M \right\}.$$

It thus clear that $\mathcal{H}_M$ is a RKHS induced by kernel function $K_M(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}_M(\boldsymbol{x}, \boldsymbol{\omega}), \boldsymbol{\phi}_M(\boldsymbol{x}', \boldsymbol{\omega}) \rangle$. For $f = \boldsymbol{u}^T \phi_M(\boldsymbol{x}) \in \mathcal{H}_M, g = \boldsymbol{z}^T \phi_M(\boldsymbol{x}) \in \mathcal{H}_M$, we define their inner product in $\mathcal{H}_M$ as $\langle f, g \rangle_{\mathcal{H}_M} = \boldsymbol{u}^T \boldsymbol{z}$. And the corresponding norm of f in $\mathcal{H}_M$ is $\|f\|_{\mathcal{H}_M} = \sqrt{\boldsymbol{u}^T \boldsymbol{u}} = \|\boldsymbol{u}\|_2$.

In the rest of this paper, we denote $\|\cdot\|$ as the operatorial norm, $\|\cdot\|_{HS}$ as the Hilbert-Schmidt norm and $\|\cdot\|_2$ as the Euclidean norm of a vector in $\mathbb{R}^n$. Let $\mathcal{H}$ be a Hilbert space, we denote with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the associated inner product, with $\|\cdot\|_{\mathcal{H}}$ the norm and with $\mathrm{Tr}(\cdot)$ the trace.

B.1. Kernel and Random Feature Operators

In this section, we provide some popular kernel and random feature operators used in the proofs.

Definition B.1. For any $g \in L^2_{\rho_{\mathcal{X}}}, \boldsymbol{\beta} \in \mathbb{R}^M$, we define

- $S_M : \mathbb{R}^M \to L^2_{\rho_{\mathcal{X}}}, \ \ (S_M \boldsymbol{\beta})(\cdot) = \boldsymbol{\beta}^T \boldsymbol{\phi}_M(\cdot)$,
- $S_M^* : L^2_{\rho_{\mathcal{X}}} \to \mathbb{R}^M, \ S_M^* g = \int_{\mathcal{X}} \boldsymbol{\phi}_M(\boldsymbol{x}) g(\boldsymbol{x}) d\rho_{\mathcal{X}}(\boldsymbol{x})$,
- $S_{M,D}^* : L^2_{\rho_{\mathcal{X}}} \to \mathbb{R}^M, \ S_{M,D}^* g = \frac{1}{|D|} \sum_{i=1}^{|D|} \boldsymbol{\phi}_M(\boldsymbol{x}_i) g(\boldsymbol{x}_i)$,
- $C_M : \mathbb{R}^M \to \mathbb{R}^M, \ C_M = \int_{\mathcal{X}} \boldsymbol{\phi}_M(\boldsymbol{x}) \boldsymbol{\phi}_M(\boldsymbol{x})^T d\rho_{\mathcal{X}}(\boldsymbol{x})$,
- $C_{M,D} : \mathbb{R}^M \to \mathbb{R}^M, \ C_{M,D} = \frac{1}{|D|} \sum_{i=1}^{|D|} \boldsymbol{\phi}_M(\boldsymbol{x}_i) \boldsymbol{\phi}_M(\boldsymbol{x}_i)^T$.

According to Definition 3.1, B.1 and Assumption 3.3, we have $L_M, C_M, S_M, C_{M,D}$ are finite dimensional. Moreover, we have $L_M = S_M S_M^*, C_M = S_M^* S_M$ and $C_{M,D} = S_{M,D}^* S_M$. Finally, $L_M, C_M, C_{M,D}$ are self-adjoint and positive operator, with spectrum is $[0, \kappa^2]$.

B.2. Error Decomposition

In this section, we first introduce some intermediate estimators based on check loss and least square loss and then explain the relationship between the estimators. Finally, we give a tight error decomposition for KQR-RF.

Definition B.2. We define the following intermediate estimators:

$$f_{M,D,\lambda} = \operatorname*{argmin}_{f\in\mathcal{H}_M} \left\{ \frac{1}{|D|} \sum_{(\boldsymbol{x},y)\in D} \rho_\tau\left(y - f(\boldsymbol{x})\right) + \lambda \|f\|^2_{\mathcal{H}_M} \right\},$$

$$f^\diamond_{M,D,\lambda} = \operatorname*{argmin}_{f\in\mathcal{H}_M} \left\{ \frac{1}{|D|} \sum_{(\boldsymbol{x},y)\in D} \left(f(\boldsymbol{x}) - f^*_\tau(\boldsymbol{x})\right)^2 + \lambda \|f\|^2_{\mathcal{H}_M} \right\},$$

$$f_{M,\lambda} = \operatorname*{argmin}_{f\in\mathcal{H}_M} \left\{ \int_{\mathcal{X}} \left(f(\boldsymbol{x}) - f^*_\tau(\boldsymbol{x})\right)^2 d\rho_{\mathcal{X}}(\boldsymbol{x}) + \lambda \|f\|^2_{\mathcal{H}_M} \right\},$$

$$f_{\lambda} = \operatorname*{argmin}_{f\in\mathcal{H}_K} \left\{ \int_{\mathcal{X}} \left(f(\boldsymbol{x}) - f^*_\tau(\boldsymbol{x})\right)^2 d\rho_{\mathcal{X}}(\boldsymbol{x}) + \lambda \|f\|^2_{K} \right\}.$$

We can also write $f^\diamond_{M,D,\lambda} = \boldsymbol{\phi}_M(\cdot)^T \boldsymbol{\omega}^\diamond_{M,D,\lambda}$ and $\boldsymbol{\omega}^\diamond_{M,D,\lambda} = \operatorname{argmin}_{\boldsymbol{\omega}\in\mathbb{R}^M} \frac{1}{|D|}\sum_{(\boldsymbol{x},y)\in D}(\boldsymbol{\phi}_M(\boldsymbol{x})^T\boldsymbol{\omega} - f^*_\tau(\boldsymbol{x}))^2 + \lambda\|\boldsymbol{\omega}\|_2^2$, $f_{M,\lambda} = \boldsymbol{\phi}_M(\cdot)^T\boldsymbol{\omega}_{M,\lambda}$ and $\boldsymbol{\omega}_{M,\lambda} = \operatorname{argmin}_{\boldsymbol{\omega}\in\mathbb{R}^M} \int_{\mathcal{X}}(\boldsymbol{\phi}_M(\boldsymbol{x})^T\boldsymbol{\omega} - f^*_\tau(\boldsymbol{x}))^2 d\rho_{\mathcal{X}}(\boldsymbol{x}) + \lambda\|\boldsymbol{\omega}\|_2^2$.

Note that $f_{M,D,\lambda}$ is the global estimator of KQR-RF, and it does not have an explicit form due to the non-smoothness of the check loss function, while the other three estimators are defined by the least square loss function. Recall the operators defined in Definition 3.1 and B.1, we have

$$\begin{aligned}
&f^\diamond_{M,D,\lambda} = S_M(C_{M,D} + \lambda I)^{-1} S^*_{M,D} f^*_\tau,\\
&f_{M,\lambda} = (L_M + \lambda I)^{-1} L_M f^*_\tau = S_M(C_M + \lambda I)^{-1} S^*_M f^*_\tau,\\
&f_\lambda = (L_K + \lambda I)^{-1} L_K f^*_\tau.
\end{aligned}$$

According to the definition of these estimators, we summarized a relationship chain from the KQR-RF estimator to the true quantile function f^*_τ in $L^2_{\rho_{\mathcal{X}}}$:

$$f_{M,D,\lambda} \overset{\rho_\tau\to ls}{\longrightarrow} f^\diamond_{M,D,\lambda} \overset{\rho_{\mathcal{X}}(\boldsymbol{x})}{\longrightarrow} f_{M,\lambda} \overset{\mathcal{H}_M\to\mathcal{H}_K}{\longrightarrow} f_\lambda \overset{\mathcal{H}_K\to L_2}{\longrightarrow} f^*_\tau.$$

Consequently, we can decompose the error in terms of $L^2_{\rho_{\mathcal{X}}}$-norm into four parts: $\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_\rho$ is the LS-approximation error from the non-smooth check loss to the least square loss; $\|f^\diamond_{M,D,\lambda} - f_{M,\lambda}\|_\rho$ is the empirical error from the sample to the expectation; $\|f_{M,\lambda} - f_\lambda\|_\rho$ is the approximation error introduced by the random features; and $\|f_\lambda - f^*_\tau\|_\rho$ is the approximation error between $\mathcal{H}_K$ and $L^2_{\rho_{\mathcal{X}}}$.

Lemma B.3. *Let $f_{M,D,\lambda}$, $f^\diamond_{M,D,\lambda}$, $f_{M,\lambda}$ and f_λ be defined in Definition B.2, we have the following error decomposition for KQR-RF,*

$$\|f_{M,D,\lambda} - f^*_\tau\|_\rho \le \underbrace{\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_\rho}_{\textit{LS-approximation error}} + \underbrace{\|f^\diamond_{M,D,\lambda} - f_{M,\lambda}\|_\rho}_{\textit{Empirical error}} + \underbrace{\|f_{M,\lambda} - f_\lambda\|_\rho}_{\textit{RF error}} + \underbrace{\|f_\lambda - f^*_\tau\|_\rho}_{\textit{Approximation error}}. \tag{12}$$

Proof. According to the triangle inequality, we can obtain the result directly. □

B.3. Error Bounds

In this section, we provide the bounds for the four error terms in Lemma B.3. By utilizing the operator representation of $f^\diamond_{M,D,\lambda}$, $f_{M,\lambda}$ and f_λ, we first bound the last three error terms. Benefiting from the maximum dimension of random features and Berntein's inequalities, our refined convergence results allow the source condition index $r \in [0,1]$ which can also explain the agnostic case when $f^*_\tau \notin \mathcal{H}_K$. Thus we can show that $\|f^\diamond_{M,D,\lambda} - f^*_\tau\|_\rho$ is small enough, i.e., $\mathcal{O}_P(\lambda^r)$. Based on this result, we finally bound the LS-approximation error by using the empirical process and some properties of the check loss function.

B.3.1. Approximation Errors

Lemma B.4. *Let f_λ and f^*_τ be defined in Definition B.2 and (1), respectively. Under Assumption 3.4, for any $\lambda \in (0,1)$ and $r \in (0,1]$, there holds*

$$\|f_\lambda - f^*_\tau\|_\rho \le R\lambda^r.$$

Proof. Recall that $f_\lambda = (L_K + \lambda I)^{-1} L_K f_\tau^*$ and Assumption 3.4 that $f_\tau^* = L_K^r h_\rho$ with $\|h_\rho\|_\rho \le R$, we have

$$\begin{aligned}\|f_\lambda - f_\tau^*\|_\rho &= \|(L_K + \lambda I)^{-1} L_K f_\tau^* - f_\tau^*\|_\rho = \|\lambda (L_K + \lambda I)^{-1} f_\tau^*\|_\rho \\ &= \lambda \|(L_K + \lambda I)^{-1} L_K^r h_\rho\|_\rho \\ &= \lambda^r \|\lambda^{1-r} (L_K + \lambda I)^{r-1} (L_K + \lambda I)^{-r} L_K^r h_\rho\|_\rho \\ &\le \lambda^r \|\lambda (L_K + \lambda I)^{-1}\|^{1-r} \|(L_K + \lambda I)^{-1} L_K\|^r \|h_\rho\|_\rho \le R\lambda^r,\end{aligned}$$

where the first inequality is from Lemma E.1 and the fact that $\|\lambda(L_K + \lambda I)^{-1}\|^{1-r} \le 1$ and $\|(L_K + \lambda I)^{-1} L_K\|^r \le 1$ for $r \in (0, 1)$. Thus we complete the proof. □

B.3.2. Random Feature Error

Lemma B.5. *Let $f_{M,\lambda}$ and f_λ be defined in Definition B.2, for any $0 < \lambda \le \|L_K\|$ and $\delta \in (0, 1)$, if the number of random features satisfies the following inequalities*

$$\begin{aligned}&M \ge 16(\mathcal{N}_\infty(\lambda) + 1)\log(4/\delta), \quad \textit{for} \quad r \in (0, 1/2); \\ &M \ge 16(\mathcal{N}_\infty(\lambda) + 1)\log(4/\delta) \vee 128\kappa^2 \lambda^{1-2r} \mathcal{N}(\lambda)^{2r-1} \mathcal{N}_\infty(\lambda)^{2-2r} \log(4/\delta), \quad \textit{for} \quad r \in [1/2, 1],\end{aligned}$$

then under Assumptions 3.3, 3.4 and 3.12, there holds

$$\|f_{M,\lambda} - f_\lambda\|_\rho \le R\lambda^r,$$

with probability at least $1 - \delta$.

Proof. From the definition of $f_{M,\lambda}$ and f_λ, we have

$$\begin{aligned}f_{M,\lambda} - f_\lambda &= ((L_M + \lambda I)L_M - (L_K + \lambda I)L_K) f_\tau^* \\ &= \lambda((L_K + \lambda I)^{-1} - (L_M + \lambda I)^{-1}) f_\tau^* \\ &= \lambda (L_M + \lambda I)^{-1} (L_M - L_K)(L_K + \lambda I)^{-1} f_\tau^* \\ &= \lambda^{1/2} (\lambda^{1/2} (L_M + \lambda I)^{-1/2})((L_M + \lambda I)^{-1/2} (L_K + \lambda I)^{1/2})[(L_K + \lambda I)^{-1/2} (L_M - L_K)(L_K + \lambda I)^{r-1}] \\ &\quad ((L_K + \lambda I)^{-r} L_K^r) h_\rho,\end{aligned}$$

where the second and third equalities use $(A + \lambda I)^{-1} A = I - \lambda(A + \lambda I)^{-1}$ and $A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}$, and the last inequality we use Assumption 3.4 that $f_\tau^* = L_K^r h_\rho$ with $\|h_\rho\|_\rho \le R$. Note that $\|\lambda^{1/2}(L_M + \lambda I)^{-1/2}\| \le 1$, $\|(L_K + \lambda I)^{-r} L_K^r\| \le 1$ and $\|(L_M + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\| \le \sqrt{2}$ in (32) from Lemma D.2 when $M \ge 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$, thus for any $\delta \in (0, 1)$, there holds

$$\|f_{M,\lambda} - f_\lambda\|_\rho \le R\sqrt{2\lambda} \|(L_K + \lambda I)^{-1/2} (L_M - L_K)(L_K + \lambda I)^{r-1}\| \tag{13}$$

with probability at least $1 - \delta$. We next to bound $\|f_{M,\lambda} - f_\lambda\|_\rho$ for two cases:

For the case when $r \in (0, 1/2)$, according to (13), we have

$$\begin{aligned}\|f_{M,\lambda} - f_\lambda\|_\rho &\le R\sqrt{2\lambda} \|(L_K + \lambda I)^{-1/2} (L_M - L_K)(L_K + \lambda I)^{-1/2}\| \|(L_K + \lambda I)^{r-1/2}\| \\ &\le \sqrt{2} R\lambda^r \|(L_K + \lambda I)^{-1/2} (L_M - L_K)(L_K + \lambda I)^{-1/2}\| \|\lambda^{1/2-r} (L_K + \lambda I)^{r-1/2}\| \\ &\le \frac{\sqrt{2}}{2} R\lambda^r < R\lambda^r,\end{aligned}$$

where the third inequality is from $\|\lambda^{1/2-r}(L_K + \lambda I)^{r-1/2}\| \le 1$ for $r \in (0, 1/2)$, and $\|(L_K + \lambda I)^{-1/2}(L_M - L_K)(L_K + \lambda I)^{-1/2}\| \le 1/2$ in (31) from Lemma D.2 when $M \ge 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$.

For the case when $r \in [1/2, 1]$, according to (13), we apply Lemma E.2 by letting $s = 2 - 2r \in [0, 1]$, $X = (L_K + \lambda I)^{-1/2}(L_M - L_K)$ and $A = (L_K + \lambda I)^{-1/2}$,

$$\|f_{M,\lambda} - f_\lambda\|_\rho \le R\sqrt{2\lambda} \|(L_K + \lambda I)^{-1/2} (L_M - L_K)\|^{2r-1} \|(L_K + \lambda I)^{-1/2} (L_M - L_K)(L_K + \lambda I)^{-1/2}\|^{2-2r}.$$

Note that from Lemmas D.1 and D.3, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\|f_{M,\lambda} - f_\lambda\|_\rho \le R\sqrt{2\lambda}\|(L_K+\lambda I)^{-1/2}(L_M - L_K)\|^{2r-1}\|(L_K+\lambda I)^{-1/2}(L_M-L_K)(L_K+\lambda I)^{-1/2}\|^{2-2r} \\
\le& R\sqrt{2\lambda}\left(\frac{4\kappa\sqrt{\mathcal{N}_\infty(\lambda)}\log(4/\delta)}{M} + \sqrt{\frac{4\kappa^2\mathcal{N}(\lambda)\log(4/\delta)}{M}}\right)^{2r-1}\left(\frac{2(\mathcal{N}_\infty(\lambda)+1)\log(4/\delta)}{M} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(4/\delta)}{M}}\right)^{2-2r} \\
\le& R\sqrt{2\lambda}\left[\left(\frac{4\kappa\sqrt{\mathcal{N}_\infty(\lambda)}\log(4/\delta)}{M}\right)^{2r-1} + \left(\sqrt{\frac{4\kappa^2\mathcal{N}(\lambda)\log(4/\delta)}{M}}\right)^{2r-1}\right]\left(\sqrt{\frac{4\mathcal{N}_\infty(\lambda)\log(4/\delta)}{M}}\right)^{2-2r} \\
\le& 4\sqrt{2}R\kappa^{2r-1}\left(\frac{\sqrt{\lambda\mathcal{N}_\infty(\lambda)}(\log(4/\delta))^r}{M^r} + \sqrt{\frac{\lambda\mathcal{N}(\lambda)^{2r-1}\mathcal{N}_\infty(\lambda)^{2-2r}\log(4/\delta)}{M}}\right),
\end{aligned} \tag{14}$$

where the second inequality follows from the inequality that $(a+b)^{2r-1} \le a^{2r-1} + b^{2r-1}$ for $r \in [1/2, 1]$.

Next, we need to add a condition on M to bound (14) with $R\lambda^r$. We consider

$$M \ge 128\kappa^2\lambda^{1-2r}\mathcal{N}(\lambda)^{2r-1}\mathcal{N}_\infty(\lambda)^{2-2r}\log(4/\delta),$$

plug this condition into (14), we get

$$\begin{aligned}
\|f_{M,\lambda} - f_\lambda\|_\rho &\le 4\sqrt{2}R\kappa^{2r-1}\left(\sqrt{\frac{\lambda^{4r^2-2r+1}\mathcal{N}_\infty(\lambda)^{4r^2-4r+1}}{128^{2r}\kappa^{4r}\mathcal{N}(\lambda)^{4r^2-2r}}} + \sqrt{\frac{\lambda^{2r}}{128\kappa^2}}\right) \\
&\le 4\sqrt{2}R\left(\sqrt{\frac{\lambda^{2r}}{128^{2r}\kappa^{8r-8r^2}\mathcal{N}(\lambda)^{4r^2-2r}}} + \sqrt{\frac{\lambda^{2r}}{128\kappa^{4-4r}}}\right) \le R\lambda^r,
\end{aligned}$$

where the second inequality is obtained from $\mathcal{N}_\infty(\lambda) = \sup_{\boldsymbol{\omega}} \|(L_K+\lambda)^{-1/2}\phi_{\boldsymbol{\omega}}\|_\rho \le \kappa^2/\lambda$ due to Assumptions 3.3 and 3.12, the third inequality follows from: (1) $128^{2r}\mathcal{N}(\lambda)^{4r^2-2r} \ge 128$, this is from the fact that $2r \ge 1$ and $4r^2 - 2^r \ge 0$ with $r \in [1/2, 1)$, and $\mathcal{N}(\lambda) = \mathrm{Tr}((L_K+\lambda I)^{-1}L_K) \ge \frac{|L_K|}{|L_K|+\lambda} \ge 1/2$ with $0 \le \lambda \le |L_K|$; (2) $\kappa^{8r-8r^2} \ge 1$ and $\kappa^{4-4r} \ge 1$ with $\kappa \ge 1$ from Assumption 3.3 and $r \in [1/2, 1]$.

Combining the results of two cases, we complete the proof.

□

B.3.3. Empirical Error

The empirical error is related to the similarity between C_M and $C_{M,D}$, so we first define two important quantities measuring this similarity

$$\begin{aligned}
\mathcal{Q}_{M,D,\lambda} &= \|(C_M+\lambda I)^{1/2}(C_{M,D}+\lambda I)^{-1/2}\|, \\
\mathcal{R}_{M,D,\lambda} &= \|(C_M+\lambda I)^{-1/2}(C_M - C_{M,D})(C_M+\lambda I)^{-1/2}\|.
\end{aligned}$$

Lemma B.6. *Let $f^\diamond_{M,D,\lambda}$ and $f_{M,\lambda}$ be defined in Definition B.2, if the number of random features and the total sample size satisfy inequalities*

$$\begin{aligned}
&M \ge 16(\mathcal{N}_\infty(\lambda)+1)\log(4/\delta), \quad \textit{for} \quad r \in (0, 1/2); \\
&M \ge 16(\mathcal{N}_\infty(\lambda)+1)\log(4/\delta) \vee 128\kappa^2\lambda^{1-2r}\mathcal{N}(\lambda)^{2r-1}\mathcal{N}_\infty(\lambda)^{2-2r}\log(4/\delta), \quad \textit{for} \quad r \in [1/2, 1],
\end{aligned}$$

and $|D| \ge 16(\kappa^2\lambda^{-1}+1)\log(6/\delta)$, respectively, then under Assumption 3.3, 3.4 and 3.12, for any $\delta \in (0,1)$, there holds

$$\|f^\diamond_{M,D,\lambda} - f_{M,\lambda}\|_{\mathcal{H}_M} \le \sqrt{2}\widetilde{C}_1\lambda^{r-1/2},$$

and

$$\|f^\diamond_{M,D,\lambda} - f_{M,\lambda}\|_\rho \le \sqrt{2}\widetilde{C}_1\lambda^r,$$

with probability at least $1-\delta$, where $\widetilde{C}_1$ is a constant defined in the proof.

Proof. From the definition of $f^{\diamond}_{M,D,\lambda}$ and $f_{M,\lambda}$, we have

$$\|f^{\diamond}_{M,D,\lambda} - f_{M,\lambda}\|_{\mathcal{H}_M} = \|\boldsymbol{\omega}^{\diamond}_{M,D,\lambda} - \boldsymbol{\omega}_{M,\lambda}\|_2,$$

and by the equality that $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$ for A and B are invertible operators, we have

$$\begin{aligned}
\boldsymbol{\omega}^{\diamond}_{M,D,\lambda} - \boldsymbol{\omega}_{M,\lambda} =& (C_{M,D}+\lambda I)^{-1} S^*_{M,D} f^*_\tau - (C_M+\lambda I)^{-1} S^*_M f^*_\tau \\
=& (C_{M,D}+\lambda I)^{-1}(S^*_{M,D} - S^*_M) f^*_\tau + [(C_{M,D}+\lambda I)^{-1} - (C_M+\lambda I)^{-1}] S^*_M f^*_\tau \\
=& (C_{M,D}+\lambda I)^{-1}(S^*_{M,D} - S^*_M) f^*_\tau + (C_{M,D}+\lambda I)^{-1}(C_M - C_{M,D})\boldsymbol{\omega}_{M,\lambda} \\
=& (C_{M,D}+\lambda I)^{-1}(S^*_{M,D} - S^*_M) f^*_\tau + (C_{M,D}+\lambda I)^{-1}(S^*_M - S^*_{M,D}) S_M \boldsymbol{\omega}_{M,\lambda} \\
=& (C_{M,D}+\lambda I)^{-1} S^*_{M,D}(f^*_\tau - f_{M,\lambda}) + (C_{M,D}+\lambda I)^{-1} S^*_M (f_{M,\lambda} - f^*_\tau),
\end{aligned}$$

where the fourth equality uses $C_{M,D} = S^*_{M,D} S_M$ and $C_M = S^*_M S_M$. Thus we have

$$\begin{aligned}
\|\boldsymbol{\omega}^{\diamond}_{M,D,\lambda} - \boldsymbol{\omega}_{M,\lambda}\|_2 \le& \left(\|(C_{M,D}+\lambda I)^{-1} S^*_{M,D}\| + \|(C_{M,D}+\lambda I)^{-1} S^*_M\|\right) \|f_{M,\lambda} - f^*_\tau\|_\rho \\
\le& \|(C_{M,D}+\lambda I)^{-1/2}\| \left(\|(C_{M,D}+\lambda I)^{-1/2} S^*_{M,D}\| + \|(C_{M,D}+\lambda I)^{-1/2} S^*_M\|\right) \|f_{M,\lambda} - f^*_\tau\|_\rho.
\end{aligned}$$

Note that $C_{M,D}$ is self-adjoint and positive operator, we have $\|(C_{M,D}+\lambda I)^{-1/2}\| \le \lambda^{-1/2}$. On the other hand, it holds that

$$\|(C_{M,D}+\lambda I)^{-1/2} S^*_{M,D}\| = \|(C_{M,D}+\lambda I)^{-1/2} C_{M,D} (C_{M,D}+\lambda I)^{-1/2}\|^{1/2} \le 1,$$

and

$$\begin{aligned}
\|(C_{M,D}+\lambda I)^{-1/2} S^*_M\| &= \|(C_{M,D}+\lambda I)^{-1/2}(C_M+\lambda I)^{1/2}(C_M+\lambda I)^{-1/2} S^*_M\| \\
&\le \|(C_{M,D}+\lambda I)^{-1/2}(C_M+\lambda I)^{1/2}\| \|(C_M+\lambda I)^{-1/2} S^*_M\| \\
&= \mathcal{Q}_{M,D,\lambda} \|(C_M+\lambda I)^{-1/2} C_M (C_M+\lambda I)^{-1/2}\|^{1/2} \le \sqrt{2},
\end{aligned}$$

where the second equality uses the fact that $\|AB\| = \|BA\|$ for A and B are self-adjoint operators, and the last inequality uses Lemma D.5 that $\mathcal{Q}_{M,D,\lambda} \le \sqrt{2}$ and $\|(C_M+\lambda I)^{-1/2} C_M (C_M+\lambda I)^{-1/2}\|^{1/2} \le 1$. Combine these inequalities and Lemma B.5, we get that

$$\|\boldsymbol{\omega}^{\diamond}_{M,D,\lambda} - \boldsymbol{\omega}_{M,\lambda}\|_2 \le (1+\sqrt{2})\lambda^{-1/2} \|f_{M,\lambda} - f^*_\tau\|_\rho \le \widetilde{C}_1 \lambda^{r-1/2},$$

where $\widetilde{C}_1 = (1+\sqrt{2})R$. Similarly, we have

$$\begin{aligned}
& f^{\diamond}_{M,D,\lambda} - f_{M,\lambda} = S_M(\boldsymbol{\omega}^{\diamond}_{M,D,\lambda} - \boldsymbol{\omega}_{M,\lambda}) \\
=& S_M (C_{M,D}+\lambda I)^{-1} S^*_{M,D}(f^*_\tau - f_{M,\lambda}) + S_M (C_{M,D}+\lambda I)^{-1} S^*_M (f_{M,\lambda} - f^*_\tau) \\
=& S_M (C_M+\lambda I)^{-1/2}(C_M+\lambda I)^{1/2}(C_{M,D}+\lambda I)^{-1/2}(C_{M,D}+\lambda I)^{-1/2} S^*_{M,D}(f^*_\tau - f_{M,\lambda}) + S_M (C_M+\lambda I)^{-1/2} \\
& (C_M+\lambda I)^{1/2}(C_{M,D}+\lambda I)^{-1/2}(C_{M,D}+\lambda I)^{-1/2}(C_M+\lambda I)^{1/2}(C_M+\lambda I)^{-1/2} S^*_M (f_{M,\lambda} - f^*_\tau).
\end{aligned}$$

Note that

$$\|S_M (C_M+\lambda I)^{-1/2}\| = \|(C_M+\lambda I)^{-1/2} C_M (C_M+\lambda I)^{-1/2}\|^{1/2} \le 1,$$

and the above inequalities that $\|(C_{M,D}+\lambda I)^{-1/2} S^*_{M,D}\| \le 1$ and $\|(C_M+\lambda I)^{-1/2} S^*_M\| \le 1$, we have

$$\|f^{\diamond}_{M,D,\lambda} - f_{M,\lambda}\|_\rho \le (\mathcal{Q}_{M,D,\lambda} + \mathcal{Q}^2_{M,D,\lambda}) \|f_{M,\lambda} - f^*_\tau\|_\rho \le (2+\sqrt{2}) R \lambda^r = \sqrt{2}\widetilde{C}_1 \lambda^r.$$

Thus we complete the proof. □

The following proposition states the convergence rate of $f^{\diamond}_{M,D,\lambda}$ under some mild conditions on r, γ, λ, and M.

Proposition B.7. *Under Assumptions 3.3-3.5 and 3.12, if $r \in [0,1], \gamma \in [0,1]$, and $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies the following inequalities*

$$M \gtrsim |D|^{\frac{\alpha}{2r+\gamma}}, \quad \textit{for} \quad r \in (0, 1/2),$$
$$M \gtrsim |D|^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}}, \quad \textit{for} \quad r \in [1/2, 1],$$

and $|D|$ is sufficiently large, then with probability near to 1, there holds

$$\|f^{\diamond}_{M,D,\lambda} - f^*_\tau\|_\rho \leq \widetilde{C}_2 |D|^{-\frac{r}{2r+\gamma}}, \tag{15}$$

where $\widetilde{C}_2 = 2R + \sqrt{2}\widetilde{C}_1$.

Proof. Combining Lemmas B.4-B.6, and setting $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, we can obtain that

$$\|f^{\diamond}_{M,D,\lambda} - f^*_\tau\|_\rho \leq (2R + \sqrt{2}\widetilde{C}_1)\lambda^r = (2R + \sqrt{2}\widetilde{C}_1)|D|^{-\frac{r}{2r+\gamma}},$$

with probability near to 1. Now we check the following conditions for M and $|D|$,

$$M \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(4/\delta), \quad \text{for} \quad r \in (0, 1/2);$$
$$M \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(4/\delta) \vee 128\kappa^2\lambda^{1-2r}\mathcal{N}(\lambda)^{2r-1}\mathcal{N}_\infty(\lambda)^{2-2r}\log(4/\delta), \quad \text{for} \quad r \in [1/2, 1];$$
$$|D| \geq 32(\kappa^2\lambda^{-1} + 1)\log(6/\delta).$$

Recalling Assumptions 3.5 and 3.12 that $\mathcal{N}(\lambda) \leq Q^2\lambda^{-\gamma}$, $\mathcal{N}_\infty(\lambda) \leq F\lambda^{-\alpha}$, and $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, we have

$$M \gtrsim \lambda^{-\alpha} = |D|^{\frac{\alpha}{2r+\gamma}}, \quad \text{for} \quad r \in (0, 1/2),$$

and

$$M \gtrsim \lambda^{-\alpha} \vee \lambda^{(2r-2)\alpha+(1-2r)(\gamma+1)} = |D|^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}}, \quad \text{for} \quad r \in [1/2, 1],$$

and

$$|D| \gtrsim \lambda^{-1} = |D|^{\frac{1}{2r+\gamma}} \longrightarrow |D| \text{ is sufficiently large.}$$

Thus we complete the proof. □

B.3.4. LS-APPROXIMATION ERROR

Now we are ready to provide the bound for the LS-approximation error. We first give a lemma that establishes the connection between the $L^2_{\rho_\mathcal{X}}$ error term $\|f - f^{\diamond}_{M,D,\lambda}\|^2_\rho$ and the excess risk error term $E\big[\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big]$ for any $f \in L^2_{\rho_\mathcal{X}}$. This lemma heavily relies on the adaptive self-calibration condition governing the conditional distribution of y (see Assumption 3.6). To use this assumption, we need the conclusion on Proposition B.7 that under mild condition that $f^{\diamond}_{M,D,\lambda}$ lies in the ball center at f^*_τ with radius $\|f^{\diamond}_{M,D,\lambda} - f^*_\tau\|_\rho \leq \varepsilon$ for $\varepsilon \leq 1$ when $|D|$ is large enough.

Lemma B.8. *Suppose that Assumptions 3.1-3.6 and the conditions in Proposition B.7 are satisfied, for any $f \in L^2_{\rho_\mathcal{X}}$, if $|D| \geq \widetilde{C}_3$, then with probability near to 1, there holds*

$$\|f - f^{\diamond}_{M,D,\lambda}\|^2_\rho \leq \frac{4}{c_2}E\big[\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big] + \frac{4c_1^2\widetilde{C}_2^2}{c_2^2}\|f^{\diamond}_{M,D,\lambda} - f^*_\tau\|^2_\rho,$$

where $c_1, c_2, \widetilde{C}_2$ and $\widetilde{C}_3$ are some universal positive constants .

Proof. Using Knight's identity that $\rho_\tau(u - v) - \rho_\tau(u) = -v\big(\tau - I(u \leq 0)\big) + \int_0^v \big(I(u \leq t) - I(u \leq 0)\big)dt$, we have

$$\begin{aligned}
&\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x})) = -(f(\boldsymbol{x}) - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big(\tau - I(y \leq f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big) \\
&+ \int_0^{f(\boldsymbol{x}) - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x})} \big(I(y \leq f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}) + t) - I(y \leq f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big)dt \\
&= -(f(\boldsymbol{x}) - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big(\tau - I(y \leq f^*_\tau(\boldsymbol{x}))\big) - (f(\boldsymbol{x}) - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big(I(y \leq f^*_\tau(\boldsymbol{x})) - I(y \leq f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big) \\
&+ \int_0^{f(\boldsymbol{x}) - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x})} \big(I(y \leq f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}) + t) - I(y \leq f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}))\big)dt.
\end{aligned}$$

Here we take the expectation and using Fubini's theorem,

$$\begin{aligned}&E\big[\rho_\tau(y-f(\boldsymbol{x}))-\rho_\tau(y-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\big]=-E\big[(f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))E((\tau-I(y\le f^*_\tau(\boldsymbol{x}))|\boldsymbol{x})\big]\\&-E\big[(f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))E((I(y\le f^*_\tau(\boldsymbol{x}))-I(y\le f^\diamond_{M,D,\lambda}(\boldsymbol{x})))|\boldsymbol{x})\big]\\&+E\left[\int_0^{f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})}\big[E(I(y\le f^\diamond_{M,D,\lambda}(\boldsymbol{x})+t)|\boldsymbol{x})-E(I(y\le f^\diamond_{M,D,\lambda}(\boldsymbol{x}))|\boldsymbol{x})\big]dt\right].\end{aligned}\tag{16}$$

The first term on the right side of (16) is 0 due to the fact that $P(y\le f^*_\tau(\boldsymbol{x})|\boldsymbol{x})=\tau$. For the second term,

$$\begin{aligned}&E\big[(f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))E((I(y\le f^*_\tau(\boldsymbol{x}))-I(y\le f^\diamond_{M,D,\lambda}(\boldsymbol{x})))|\boldsymbol{x})\big]\\&\le E\big[|f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})||E((I(y\le f^*_\tau(\boldsymbol{x}))-I(y\le f^\diamond_{M,D,\lambda}(\boldsymbol{x})))|\boldsymbol{x})|\big]\\&=E\big[|f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})||F_{y|\boldsymbol{x}}(f^\diamond_{M,D,\lambda}(\boldsymbol{x}))-F_{y|\boldsymbol{x}}(f^*_\tau(\boldsymbol{x}))|\big]\\&=E\big[|f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})||f_{y|\boldsymbol{x}}(\xi)(f^\diamond_{M,D,\lambda}(\boldsymbol{x})-f^*_\tau(\boldsymbol{x}))|\big]\\&\le c_1E\big[|f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})||f^\diamond_{M,D,\lambda}(\boldsymbol{x})-f^*_\tau(\boldsymbol{x})|\big]\\&\le c_1\sqrt{E\big[(f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))^2\big]}\sqrt{E\big[(f^\diamond_{M,D,\lambda}(\boldsymbol{x})-f^*_\tau(\boldsymbol{x}))^2\big]}=c_1\|f-f^\diamond_{M,D,\lambda}\|_\rho\|f^\diamond_{M,D,\lambda}-f^*_\tau\|_\rho,\end{aligned}$$

where the first inequality is from $E(AB)\le E(|A||B|)$ for any random variable A and B, the second equality is from the mean value theorem with $\xi\in[f^\diamond_{M,D,\lambda}(\boldsymbol{x}),f^*_\tau(\boldsymbol{x})]$ or $\xi\in[f^*_\tau(\boldsymbol{x}),f^\diamond_{M,D,\lambda}(\boldsymbol{x})]$ together with Assumption 3.6 that the conditional density $f_{y|\boldsymbol{x}}(\cdot)$ is uniformly bounded, and the last inequality is from the Cauchy-Schwarz inequality. Similarly, for the third term on the right side of (16),

$$\begin{aligned}&E\left[\int_0^{f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})}\big[E(I(y\le f^\diamond_{M,D,\lambda}(\boldsymbol{x})+t)|\boldsymbol{x})-E(I(y\le f^\diamond_{M,D,\lambda}(\boldsymbol{x}))|\boldsymbol{x})\big]dt\right]\\&=E\left[\int_0^{f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})}\big[F_{y|\boldsymbol{x}}(f^\diamond_{M,D,\lambda}(\boldsymbol{x})+t)-F_{y|\boldsymbol{x}}(f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\big]dt\right]\\&\ge c_2E\left[\int_0^{f(\boldsymbol{x})-f^\diamond_{M,D,\lambda}(\boldsymbol{x})}tdt\right]=\frac{c_2}{2}\|f-f^\diamond_{M,D,\lambda}\|^2_\rho,\end{aligned}$$

where the inequality is from Assumption 3.6 and Propsition B.7 that $\|f^\diamond_{M,D,\lambda}-f^*_\tau\|_\rho\le\xi$ when $|D|\ge(\widetilde{C}_2/\xi)^{2+\gamma/r}$.

Plug these results into (16), we get

$$\begin{aligned}\frac{c_2}{2}\|f-f^\diamond_{M,D,\lambda}\|^2_\rho&\le c_1\|f-f^\diamond_{M,D,\lambda}\|_\rho\|f^\diamond_{M,D,\lambda}-f^*_\tau\|_\rho+E\big[\rho_\tau(y-f(\boldsymbol{x}))-\rho_\tau(y-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\big]\\&\le\frac{c_1}{4\beta}\|f-f^\diamond_{M,D,\lambda}\|^2_\rho+c_1\beta\|f^\diamond_{M,D,\lambda}-f^*_\tau\|^2_\rho+E\big[\rho_\tau(y-f(\boldsymbol{x}))-\rho_\tau(y-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\big],\end{aligned}$$

then we set $\beta=c_1/c_2$, it holds that

$$\begin{aligned}\|f-f^\diamond_{M,D,\lambda}\|^2_\rho&\le\frac{4}{c_2}E\big[\rho_\tau(y-f(\boldsymbol{x}))-\rho_\tau(y-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\big]+\frac{4c_1^2}{c_2^2}\|f^\diamond_{M,D,\lambda}-f^*_\tau\|^2_\rho\\&\le\frac{4}{c_2}E\big[\rho_\tau(y-f(\boldsymbol{x}))-\rho_\tau(y-f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\big]+\frac{4c_1^2\widetilde{C}_2^2}{c_2^2}|D|^{-\frac{2r}{2r+\gamma}},\end{aligned}$$

with probability near to 1. Thus we completes the proof. □

The following lemma bounds the supremum of the difference between the empirical average dependent on the data $\frac{1}{|D|}\sum_{i=1}^{|D|}[\rho_\tau(y_i-f(\boldsymbol{x}_i))-\rho_\tau(y_i-f^\diamond_{M,D,\lambda}(\boldsymbol{x}_i))]$ and its expectation $E[\rho_\tau(y-f(\boldsymbol{x}))-\rho_\tau(y-f^\diamond_{M,D,\lambda}(\boldsymbol{x})]$ within a local ball using the Rademacher complexity function on $\mathcal{H}_M$

$$\mathcal{R}_M(\delta)=\sqrt{\frac{1}{|D|}\sum_{j=1}^{\infty}\min\{\mu_j,\delta^2\}},$$

where μ_j's are the eigenvalues of the spectral decomposition of L_M. Recall the definition of the effective dimension of $\mathcal{H}_M$ that $\mathcal{N}_M(\lambda) = \mathrm{Tr}((L_M + \lambda I)^{-1} L_M) = \sum_{j=1}^{\infty} \mu_j/(\mu_j + \lambda)$. It is easy to verify that $\mathcal{N}_M(\lambda) \asymp |D|\mathcal{R}_M^2(\sqrt{\lambda})/\lambda$ (using inequality that $\min(a,b)/2 \leq \frac{ab}{a+b} \leq \min(a,b)$ for $a, b \in \mathbb{R}$). So these two quantities are equivalent to some extent, and according to Lemma E.8, under some mild conditions on the number of random features, we have $\mathcal{R}_M(\delta) \asymp \mathcal{R}(\delta)$, where $\mathcal{R}(\delta)$ is the Rademacher complexity function on $\mathcal{H}_K$ defined by

$$\mathcal{R}(\delta) = \sqrt{\frac{1}{|D|} \sum_{j=1}^{\infty} \min\{\mu_j', \delta^2\}},$$

where μ_j''s are the eigenvalues of the spectral decomposition of L_K.

Lemma B.9. *For any $\delta > 0$ and $f \in \mathcal{H}_M$, we define the event $\mathcal{M}(\delta)$ as*

$$\left\{ \sup_{f \in \Theta(\delta)} \left| \frac{1}{|D|} \sum_{i=1}^{|D|} [\rho_\tau(y_i - f(\boldsymbol{x}_i)) - \rho_\tau(y_i - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}_i))] - E[\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x})] \right| \leq C \log |D| \mathcal{R}_M(\delta) \right\},$$

where $\Theta(\delta) := \{f \in \mathcal{H}_M \mid \|f - f^{\diamond}_{M,D,\lambda}\|_\rho \leq \delta,$ and $\|f - f^{\diamond}_{M,D,\lambda}\|_{\mathcal{H}_M} \leq 1\}$, then $\mathcal{M}(\delta)$ holds with probability near to 1.

Proof. For the notation simplify, we denote

$$A = \left| \frac{1}{|D|} \sum_{i=1}^{|D|} [\rho_\tau(y_i - f(\boldsymbol{x}_i)) - \rho_\tau(y_i - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}_i))] - E[\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x})] \right|,$$

and C is a universal positive constant that may be different from line to line in this lemma.

We first use the standard symmetrization argument in the empirical process (Pollard, 2012) to bound $E[A]$ such that

$$\begin{aligned} E[A] &\leq 2E \left[\sup_{f \in \Theta(\delta)} \left| \frac{1}{|D|} \sum_{i=1}^{|D|} \sigma_i \big(\rho_\tau(y_i - f(\boldsymbol{x}_i)) - \rho_\tau(y_i - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}_i)) \big) \right| \right] \\ &\leq 4E \left[\sup_{f \in \Theta(\delta)} \left| \frac{1}{|D|} \sum_{i=1}^{|D|} \sigma_i \big(f(\boldsymbol{x}_i) - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}_i) \big) \right| \right], \end{aligned} \tag{17}$$

where $\{\sigma_i\}'s$ denote the Rademacher variables taking values in $\{-1, 1\}$ with equal probability, the second inequality follows from the fact that $\rho_\tau(\cdot)$ is 1-Lipschitz continuous and the Ledoux–Talagrand contraction inequality (Wainwright, 2019).

For any $f \in \Theta(\delta)$, we denote $g = f - f^{\diamond}_{M,D,\lambda} \in \mathcal{H}_M$, and $g = \sum_{j=1}^{\infty} g_j \psi_j$ with $g_j = \int_{\mathcal{X}} f(\boldsymbol{x}) \psi_j(\boldsymbol{x}) \rho_{\mathcal{X}}(\boldsymbol{x}) d\boldsymbol{x}$. Note that $\|g\|_\rho \leq \delta$ and $\|g\|_{\mathcal{H}_M} \leq 1$, this implies that $\sum_{j=1}^{\infty} g_j^2 \leq \delta^2$ and $\sum_{j=1}^{\infty} g_j^2/\mu_j \leq 1$. Combine these two inequalities, we have

$$\sum_{j=1}^{\infty} \frac{g_j^2}{\min\{\mu_j, \delta^2\}} \leq 2. \tag{18}$$

Then we have

$$\begin{aligned} &\left| \sum_{i=1}^{|D|} \sigma_i \big(f(\boldsymbol{x}_i) - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x}_i) \big) \right| = \left| \sum_{i=1}^{|D|} \sigma_i \sum_{j=1}^{\infty} g_j \psi_j(\boldsymbol{x}_i) \right| \\ &= \left| \sum_{j=1}^{\infty} \frac{g_j}{\sqrt{\min\{\mu_j, \delta^2\}}} \sqrt{\min\{\mu_j, \delta^2\}} \sum_{i=1}^{|D|} \sigma_i \psi_j(\boldsymbol{x}_i) \right| \leq \sqrt{2} \left\{ \sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\} \left(\sum_{i=1}^{|D|} \sigma_i \psi_j(\boldsymbol{x}_i) \right)^2 \right\}^{1/2}, \end{aligned} \tag{19}$$

where the inequality is from Cauthy-Schwarz inequality and (18). Plug (19) into (17), we have

$$\begin{aligned}
E[A] &\leq \frac{4\sqrt{2}}{|D|} E\left\{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\}\left(\sum_{i=1}^{|D|} \sigma_i \psi_j(\boldsymbol{x}_i)\right)^2\right\}^{1/2} \\
&\leq \frac{4\sqrt{2}}{|D|} \left\{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\} E_{\boldsymbol{x},\sigma}\left(\sum_{i=1}^{|D|} \sigma_i \psi_j(\boldsymbol{x}_i)\right)^2\right\}^{1/2} \\
&= \frac{4\sqrt{2}}{|D|} \left\{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\} \sum_{i=1}^{|D|} E_{\boldsymbol{x},\sigma}\left(\sigma_i^2 \psi_j^2(\boldsymbol{x}_i)\right)\right\}^{1/2} \\
&\leq \frac{4\sqrt{2}C}{|D|} \sqrt{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\}},
\end{aligned}$$

where the second inequality follows from Jensen's inequality, and the first equality is from the fact that $E_{\boldsymbol{x},\sigma}(\sigma_i \psi_j(\boldsymbol{x}_i)) = 0$ for each i. Thus we have

$$E[A] \leq C\mathcal{R}_M(\delta). \tag{20}$$

Next, we turn to bound $A - E[A]$, note that

$$|g(\boldsymbol{x})| = \left|\sum_{j=1}^{\infty} g_j \psi_j(\boldsymbol{x})\right| \leq \sqrt{\sum_{j=1}^{\infty} \frac{g_j^2}{\min\{\mu_j, \delta^2\}}} \sqrt{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\} \psi_j^2(\boldsymbol{x})} \leq C\sqrt{\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\}} = C\sqrt{|D|}\mathcal{R}_M(\delta).$$

Thus we have

$$\left|\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\right| \leq \left|f(\boldsymbol{x}) - f^\diamond_{M,D,\lambda}(\boldsymbol{x})\right| = |g(\boldsymbol{x})| \leq C\sqrt{|D|}\mathcal{R}_M(\delta),$$

and

$$E\left[\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x}))\right]^2 \leq E\left(f(\boldsymbol{x}) - f^\diamond_{M,D,\lambda}(\boldsymbol{x})\right)^2 = E(g(\boldsymbol{x}))^2 \leq C|D|\mathcal{R}_M^2(\delta).$$

With these two inequalities, we use the Bousquet bound inequality in Lemma E.7 and set $t = C\sqrt{\frac{\log|D|}{|D|}}$, then

$$A - E[A] \leq C \log|D| \mathcal{R}_M(\delta) \tag{21}$$

holds with probability at least $1 - n^{-C}$.

Combine (20) and (21), we can obtain the inequality in the lemma. Thus we complete the proof.

□

According to Lemma B.9, we can also get the following inequality by some normalized procedure,

$$\begin{aligned}
&\left|\frac{1}{|D|}\sum_{i=1}^{|D|}[\rho_\tau(y_i - f(\boldsymbol{x}_i)) - \rho_\tau(y_i - f^\diamond_{M,D,\lambda}(\boldsymbol{x}_i))] - E[\rho_\tau(y - f(\boldsymbol{x})) - \rho_\tau(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x})]\right| \\
\leq & C\log|D|\mathcal{R}_M(\delta)\left(\frac{\|f - f^\diamond_{M,D,\lambda}\|_\rho}{\delta} + \|f - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M}\right)
\end{aligned} \tag{22}$$

Lemma B.10. *Suppose that Assumptions 3.3-3.6 and 3.12 and the conditions in Proposition B.7 are satisfied, if $|D| \geq \widetilde{C}_3$, then with probability near to 1, there holds*

$$\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_\rho \leq C|D|^{-\frac{r}{2r+\gamma}} \log|D|,$$

where C is a universal positive constant.

Proof. Recall the definition of $f_{M,D,\lambda}$, we have

$$\frac{1}{|D|}\sum_{i=1}^{|D|}\rho_\tau(y_i - f_{M,D,\lambda}(\boldsymbol{x}_i)) + \lambda\|f_{M,D,\lambda}\|_{\mathcal{H}_M} \le \frac{1}{|D|}\sum_{i=1}^{|D|}\rho_\tau(y_i - f^\diamond_{M,D,\lambda}(\boldsymbol{x}_i)) + \lambda\|f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M}.$$

We can not directly obtain a upper bound of $E[\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x})]$ according to (22) from Lemma B.9, because $E[\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x})] \ge 0$ does not always hold. Thus we use Lemma B.8 and note that $E[\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x})] + C\|f^\diamond_{M,D,\lambda} - f^*_\tau\| \ge 0$ holds. Thus combine Lemma B.8 and (22) from Lemma B.9 (plus a $C|D|^{-\frac{r}{2r+\gamma}}$ term and minus the same term in the left of (22)) , with probability near to 1, we have

$$\begin{aligned}
&E[\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x})] + C|D|^{-\frac{2r}{2r+\gamma}}\\
\le&\lambda\|f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M} - \lambda\|f_{M,D,\lambda}\|_{\mathcal{H}_M} + C\log|D|\frac{\mathcal{R}_M(\delta)}{\delta}\|f - f^\diamond_{M,D,\lambda}\|_\rho + C\log|D|\mathcal{R}_M(\delta)\|f - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M} + C|D|^{-\frac{2r}{2r+\gamma}}\\
=&-2\lambda\langle f^\diamond_{M,D,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M} - \lambda\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|^2_{\mathcal{H}_M} + C\log|D|\frac{\mathcal{R}_M(\delta)}{\delta}\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_\rho\\
&+C\log|D|\mathcal{R}_M(\delta)\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M} + C|D|^{-\frac{2r}{2r+\gamma}}\\
=&-2\lambda\langle f^\diamond_{M,D,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M} - \lambda\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|^2_{\mathcal{H}_M} + C\lambda^r\log|D|\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_\rho\\
&+C\lambda^{r+1/2}\log|D|\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M} + C\lambda^{2r},
\end{aligned} \tag{23}$$

where in the last equality, we choose δ satisfying that $\mathcal{R}_M(\delta) = \delta^{1+2r}$. Note that $\mathcal{R}_M(\delta) \asymp \mathcal{R}(\delta) \asymp \delta^{1-\gamma}/\sqrt{|D|}$ (Lian, 2022), we can obtain that $\delta = |D|^{-\frac{1}{4r+2\gamma}}$, and $\lambda = \delta^2 = |D|^{-\frac{1}{2r+\gamma}}$.

We now establish the bound of the term $-2\lambda\langle f^\diamond_{M,D,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M}$. Note that by the triangle inequality, we have

$$\begin{aligned}
&|\lambda\langle f^\diamond_{M,D,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M}|\\
\le&|\lambda\langle f^\diamond_{M,D,\lambda} - f_{M,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M}| + |\lambda\langle f_{M,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M}|.
\end{aligned} \tag{24}$$

For the first term of the right side of (24), we use the Cauchy–Schwarz inequality and Lemma B.6 and obtain that

$$\begin{aligned}
|\lambda\langle f^\diamond_{M,D,\lambda} - f_{M,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M}| &\le \lambda\|f^\diamond_{M,D,\lambda} - f_{M,\lambda}\|_{\mathcal{H}_M}\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M}\\
&\le \widetilde{C}_1\lambda^{r+1/2}\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M}.
\end{aligned}$$

For the second term in the right side of (24), we consider the following two cases:

(i). For the case when $r \in (0, 1/2)$, recall the definition of $f_{M,\lambda}$ we get

$$\begin{aligned}
\|f_{M,\lambda}\|_{\mathcal{H}_M} &= \|(L_M + \lambda I)^{-1}L_M f_\rho\|_{\mathcal{H}_M} = \|(L_M + \lambda I)^{-1}L_M L_K^r h_\rho\|_{\mathcal{H}_M} \le \|(L_M + \lambda I)^{-1}L_M\|\|L_K^r h_\rho\|_\rho\\
&\le \|L_K^r h_\rho\|_\rho \le \|L_K^r\|\|h_\rho\|_\rho \le R\kappa^{2r},
\end{aligned}$$

where the first and third inequality is from the fact that $(L_M + \lambda I)^{-1}L_M$ and L_K^r are linear operators, the last inequality is from $\|L_K^r\| \le \kappa^{2r}$ for $r \in (0, 1/2)$ and $\|h_\rho\|_\rho \le R$. Then by the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
|\lambda\langle f_{M,\lambda}, f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\rangle_{\mathcal{H}_M}| &\le \lambda\|f_{M,\lambda}\|_{\mathcal{H}_M}\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M} \le R\phi^{2r}\lambda\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M}\\
&\le R\kappa^{2r}\lambda^{r+1/2}\|f_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M}.
\end{aligned}$$

(ii). For the case when $r \in [1/2, 1]$, we have

$$
\begin{aligned}
&|\lambda\langle f_{M,\lambda}, f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\rangle_{\mathcal{H}_M}| \\
=&|\lambda\langle f_{M,\lambda}, L_M^{-1}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\rho}| \\
=&\lambda|\langle (L_M + \lambda I)^{-1} L_M L_M^r h_\tau^*, L_M^{-1}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\rho}| \\
\leq& R\lambda^r \|\lambda^{1-r}(L_M + \lambda I)^{-1} L_M^r (f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\|_{\rho} \\
=& R\lambda^r \sqrt{\langle f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}, \lambda^{2-2r} L_M^{2r}(L_M + \lambda I)^{-2}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\rho}} \\
\leq& R\lambda^r \sqrt{\langle f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}, ((2-2r)\lambda + (2r-1)L_M) L_M (L_M + \lambda I)^{-2}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\rho}} \\
\leq& R\lambda^r \sqrt{\langle f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}, \lambda L_M (L_M + \lambda I)^{-2}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\rho}} + \\
&R\lambda^r \sqrt{\langle f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}, L_M^2 (L_M + \lambda I)^{-2}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\rho}} \\
=& R\lambda^{r+1/2} \sqrt{\langle f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}, L_M^2 (L_M + \lambda I)^{-2}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\mathcal{H}_M}} + \\
&R\lambda^r \sqrt{\langle f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}, L_M^2 (L_M + \lambda I)^{-2}(f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda})\rangle_{\rho}} \\
&\leq R\lambda^{r+1/2}\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\mathcal{H}_M} + R\lambda^r \|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho},
\end{aligned}
$$

where we use the fact that $\|f\|_{\rho} = \|L_M^{1/2} f\|_{\mathcal{H}_M}$ for any $f \in L^2_{\rho_{\mathcal{X}}}$, the the second inequality uses the Young's inequality that $\lambda^{2-2r} L_M^{2r} \leq (2-2r)\lambda + 2r L_M$ for the positive operator L_M, $\lambda > 0$, and $r \in [1/2, 1]$, the last inequality is from $\|(L_M + \lambda I)^{-1} L_M\| \leq 1$. This technical proof taking consideration of $r \in [1/2, 1]$ is inspired from that of KQR in Lian (2022).

Plug the aforementioned two results (i) and (ii) into (24), we have

$$
|\lambda\langle f^{\diamond}_{M,D,\lambda}, f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\rangle_{\mathcal{H}_M}| \leq C\lambda^{r+1/2}\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\mathcal{H}_M} + C\lambda^r \|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho},
$$

where $C = (\widetilde{C}_1 + R(\phi^{2r} + 1)) + R$. Plug this result into (23), we get

$$
\begin{aligned}
&E[\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x})] \\
\leq& C\lambda^{r+1/2}\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\mathcal{H}_M} + C\lambda^r \|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho} - \lambda\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|^2_{\mathcal{H}_M} \\
&+ C\lambda^r \log|D| \|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho} + C\lambda^{r+1/2}\log|D|\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\mathcal{H}_M} + C\lambda^{2r} \\
\leq& C\lambda^{2r} + \frac{\lambda}{4}\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|^2_{\mathcal{H}_M} + C\lambda^r \log|D|\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho} - \lambda\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|^2_{\mathcal{H}_M} \\
&+ C\lambda^{2r}\log^2|D| + \frac{\lambda}{4}\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|^2_{\mathcal{H}_M} \\
\leq& C\lambda^{2r}\log^2|D| + C\lambda^r \log|D|\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho} = C|D|^{-\frac{2r}{2r+\gamma}}\log^2|D| + C|D|^{-\frac{r}{2r+\gamma}}\log|D|\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho}.
\end{aligned}
$$

By Lemma B.8, we have

$$
\begin{aligned}
\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|^2_{\rho} &\leq CE[\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f^{\diamond}_{M,D,\lambda}(\boldsymbol{x})] + C|D|^{-\frac{2r}{2r+\gamma}} \\
&\leq C|D|^{-\frac{2r}{2r+\gamma}}\log^2|D| + C|D|^{-\frac{r}{2r+\gamma}}\log|D|\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho}.
\end{aligned}
$$

Solve the above inequality we can finally obtain that

$$
\|f_{M,D,\lambda} - f^{\diamond}_{M,D,\lambda}\|_{\rho} \leq C|D|^{-\frac{r}{2r+\gamma}}\log|D|.
$$

Thus we complete the proof. □

B.4. Proofs of Theorems 3.9 and 3.13 and Corollary 3.16

Now we are ready to prove Theorems 3.9 and 3.13 and Corollary 3.16.

Proof. By Proposition B.7 and Lemma B.10, if $r \in [0,1], \gamma \in [0,1], 2r+\gamma \geq \alpha$, and $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, and the number of random features satisfies the following two inequalities

$$M \gtrsim |D|^{\frac{\alpha}{2r+\gamma}}, \quad \text{for} \quad r \in (0, 1/2),$$
$$M \gtrsim |D|^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}}, \quad \text{for} \quad r \in [1/2, 1],$$

then with probability near to 1, there holds

$$\|f_{M,D,\lambda} - f_\tau^*\|_\rho \leq \|f_{M,D,\lambda} - f_{M,D,\lambda}^\diamond\|_\rho + \|f_{M,D,\lambda}^\diamond - f_\tau^*\|_\rho \leq (\widetilde{C}_2 + C)|D|^{-\frac{r}{2r+\gamma}} \log|D|,$$

so we have $\|f_{M,D,\lambda} - f_{M,D,\lambda}^\diamond\|_\rho \leq C|D|^{-\frac{r}{2r+\gamma}} \log^2|D|$. Recall the knight inequality that

$$\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f_\tau^*(\boldsymbol{x})) = -(f_{M,D,\lambda}(\boldsymbol{x}) - f_\tau^*(\boldsymbol{x}))\big(\tau - I(y \leq f_\tau^*(\boldsymbol{x}))\big)$$
$$+ \int_0^{f_{M,D,\lambda}(\boldsymbol{x}) - f_\tau^*(\boldsymbol{x})} \big(I(y \leq f_\tau^*(\boldsymbol{x}) + t) - I(y \leq f_\tau^*(\boldsymbol{x}))\big) dt.$$

Taking the expectation and using Fubini's theorem, we obtain that

$$E\big[\rho_\tau(y - f_{M,D,\lambda}(\boldsymbol{x})) - \rho_\tau(y - f_\tau^*(\boldsymbol{x}))\big] = -E\big[(f_{M,D,\lambda}(\boldsymbol{x}) - f_{M,D,\lambda}^\diamond(\boldsymbol{x}))E((\tau - I(y \leq f_\tau^*(\boldsymbol{x}))|\boldsymbol{x})\big]$$
$$+E\left[\int_0^{f_{M,D,\lambda}(\boldsymbol{x}) - f_\tau^*(\boldsymbol{x})} \big[E(I(y \leq f_\tau^*(\boldsymbol{x}) + t)|\boldsymbol{x}) - E(I(y \leq f_\tau^*(\boldsymbol{x}))|\boldsymbol{x})\big] dt\right].$$

The first term is 0 due to that fact that $E((\tau - I(y \leq f_\tau^*(\boldsymbol{x}))|\boldsymbol{x}) = 0$ and the second term can be bounded by

$$E\left[\int_0^{f_{M,D,\lambda}(\boldsymbol{x}) - f_\tau^*(\boldsymbol{x})} \big[E(I(y \leq f_\tau^*(\boldsymbol{x}) + t)|\boldsymbol{x}) - E(I(y \leq f_\tau^*(\boldsymbol{x}))|\boldsymbol{x})\big] dt\right]$$
$$= E\left[\int_0^{f_{M,D,\lambda}(\boldsymbol{x}) - f_\tau^*(\boldsymbol{x})} \big[F_{y|\boldsymbol{x}}(f_\tau^*(\boldsymbol{x}) + t) - F_{y|\boldsymbol{x}}(f_\tau^*(\boldsymbol{x}))\big] dt\right]$$
$$\leq c_1 E\left[\int_0^{f_{M,D,\lambda}(\boldsymbol{x}) - f_\tau^*(\boldsymbol{x})} t dt\right] = \frac{c_1}{2}\|f_{M,D,\lambda} - f_\tau^*\|_\rho^2,$$

where the inequality uses Assumption 3.12 that $\sup_{t\in\mathbb{R}} f_{y|\boldsymbol{x}}(t) \leq c_1$. Therefore, we have

$$\mathcal{E}(f_{M,D,\lambda}) - \mathcal{E}(f_\tau^*) \leq c_1/2\|f_{M,D,\lambda} - f_\tau^*\|_\rho^2 \leq C|D|^{-\frac{2r}{2r+\gamma}} \log^2|D|.$$

Thus we complete the proof of Theorem 3.13. By Theorem 3.13 with $\alpha = 1$ and $\alpha = \gamma$, we can estabilsh the proofs of Theorem 3.9 and Corollary 3.16.

□

C. Extension to the Lipschitz Loss

In this section, we consider random feature method with Lipschitz continuous loss function $L(\cdot,\cdot)$. Similar to the check loss case in (5), we approximate y_i with $f_{M,D,\lambda}^L = \widetilde{\boldsymbol{u}}\phi_M$ and formulate the following general learning problem

$$\widetilde{\boldsymbol{u}} = \underset{\boldsymbol{u}\in\mathbb{R}^M}{\operatorname{argmin}} \frac{1}{|D|}\sum_{i=1}^{|D|} L\big(y_i, \boldsymbol{u}^T\boldsymbol{\phi}_M(\boldsymbol{x}_i)\big) + \lambda \boldsymbol{u}^T\boldsymbol{u}.$$

The following theorem shows the capacity-dependent learning rates for the RF estimator with Lipschitz continuous loss function (Lip-RF), which is sharper than those of the existing literature (Li et al., 2021; Li, 2022) and can be applied to the agnostic setting.

Theorem C.1. *Under Assumptions 3.3-3.6 and 3.17, if $r \in (0,1]$, $\gamma \in [0,1]$, and $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies*

$$M \gtrsim |D|^{\frac{\alpha}{2r+\gamma}}, \quad \textit{for} \quad r \in (0, 1/2),$$
$$M \gtrsim |D|^{\frac{(2r-1)(1+\gamma-\alpha)+\alpha}{2r+\gamma}}, \quad \textit{for} \quad r \in [1/2, 1],$$

and $|D|$ is sufficiently large, then there holds

$$\|f^L_{M,D,\lambda} - f^*\|_\rho^2 = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log |D|),$$

with probability near to 1, where $f^* = \operatorname{argmin}_f \mathcal{E}_L(f)$.

Proof. Similar to Lemma B.3, we decompose the error for Lip-RF in the following

$$\|f^L_{M,D,\lambda} - f^*\|_\rho \leq \underbrace{\|f^L_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_\rho}_{\text{LS-approximation error}} + \underbrace{\|f^\diamond_{M,D,\lambda} - f_{M,\lambda}\|_\rho}_{\text{Empirical error}} + \underbrace{\|f_{M,\lambda} - f_\lambda\|_\rho}_{\text{RF error}} + \underbrace{\|f_\lambda - f^*_\tau\|_\rho}_{\text{Approximation error}}. \tag{25}$$

For the last three error terms, we have established their upper bounds in Lemma B.4-B.6. So we only need to bound the first LS-approximation error term.

With the similar argument in the proof of Lemma B.8, we have the similar (adaptive) local strongly convexity condition on L with Assumption 3.17,

$$\mathcal{E}_L(f) - \mathcal{E}_L(f^\diamond_{M,D,\lambda}) \geq c_3 \|f - f^\diamond_{M,D,\lambda}\|_\rho^2, \tag{26}$$

or

$$\mathcal{E}_L(f) - \mathcal{E}_L(f^\diamond_{M,D,\lambda}) + \|f^\diamond_{M,D,\lambda} - f^*\|_\rho^2 \geq c_4 \|f - f^\diamond_{M,D,\lambda}\|_\rho^2. \tag{27}$$

Note that L is Lipschitz continuous, we can also replace ρ_τ with L and obtain a similar inequality with that in (22)

$$\begin{aligned} &\left| \frac{1}{|D|} \sum_{i=1}^{|D|} [L(y_i - f(\boldsymbol{x}_i)) - L(y_i - f^\diamond_{M,D,\lambda}(\boldsymbol{x}_i))] - E[L(y - f(\boldsymbol{x})) - L(y - f^\diamond_{M,D,\lambda}(\boldsymbol{x})] \right| \\ \leq & C \log |D| \mathcal{R}_M(\delta) \left(\frac{\|f - f^\diamond_{M,D,\lambda}\|_\rho}{\delta} + \|f - f^\diamond_{M,D,\lambda}\|_{\mathcal{H}_M} \right). \end{aligned} \tag{28}$$

Using (26)-(28), we perform a similar procedure in Lemma B.10 with ρ_τ replaced by L and get the upper bound for the LS-approximation error term

$$\|f^L_{M,D,\lambda} - f^\diamond_{M,D,\lambda}\|_\rho \leq C|D|^{-\frac{r}{2r+\gamma}} \log |D|, \tag{29}$$

with probability near to 1.

Combining Lemmas B.4-B.6, (25) and (29), we have

$$\|f^L_{M,D,\lambda} - f^*\|_\rho^2 = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log |D|),$$

with probability near to 1. □

With $\alpha = 1$ and $\alpha = \gamma$, we can derive the following corollaries for Lip-RF with uniformly sampling and data-dependent sampling strategies.

Corollary C.2. *Under Assumptions 3.3-3.6 and 3.17, if random features are sampled according to the uniform strategy, $r \in (0,1]$, $\gamma \in [0,1]$, and $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies*

$$M \gtrsim |D|^{\frac{1}{2r+\gamma}}, \quad \textit{for} \quad r \in (0, 1/2),$$
$$M \gtrsim |D|^{\frac{(2r-1)\gamma+1}{2r+\gamma}}, \quad \textit{for} \quad r \in [1/2, 1],$$

and $|D|$ is sufficiently large, then there holds

$$\|f^L_{M,D,\lambda} - f^*\|_\rho^2 = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log |D|)$$

with probability near to 1, where $f^* = \operatorname{argmin}_f \mathcal{E}_L(f)$.

Corollary C.3. *Under Assumptions 3.3-3.6 and 3.17, if random features are sampled according to the strategy in Example 3.14, $r \in (0,1]$, $\gamma \in [0,1]$, and $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies*

$$M \gtrsim |D|^{\frac{\gamma}{2r+\gamma}}, \quad \textit{for} \quad r \in (0, 1/2),$$
$$M \gtrsim |D|^{\frac{2r+\gamma-1}{2r+\gamma}}, \quad \textit{for} \quad r \in [1/2, 1],$$

and $|D|$ is sufficiently large, then there holds

$$\|f^L_{M,D,\lambda} - f^*\|_\rho^2 = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log |D|),$$

with probability near to 1, where $f^* = \operatorname{argmin}_f \mathcal{E}_L(f)$.

Remark C.4. Note that if we further pose an assumption which is widely used in the literature (Feng et al., 2023)

$$\mathcal{E}_L(f) - \mathcal{E}_L(f^*) \leq C\|f - f^*\|_\rho^2,$$

we can also establish the learning rates for the excess risk of Lip-RF as given by

$$\mathcal{E}_L(f^L_{M,D,\lambda}) - \mathcal{E}_L(f^*) \asymp \|f^L_{M,D,\lambda} - f^*\|_\rho^2 = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}} \log^2 |D|).$$

D. Operator Similarities

In this section, we provide some tight bounds of operator similarities. We first analyze the similarity between L_K and L_M.

Lemma D.1. *For any $\delta \in (0,1)$, under Assumption 3.12, there holds*

$$\|(L_K + \lambda I)^{-1/2}(L_K - L_M)(L_K + \lambda I)^{-1/2}\| \leq \frac{2(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)}{M} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{M}}, \tag{30}$$

with probability at least $1 - \delta$.

Proof. We denote $\phi_{\boldsymbol{\omega}}$ as the the function $\phi(\cdot, \boldsymbol{\omega})$ for any $\boldsymbol{\omega} \in \Omega$. Note that

$$L_M = \frac{1}{M}\sum_{i=1}^M \phi_{\boldsymbol{\omega}_i} \otimes \phi_{\boldsymbol{\omega}_i} \quad \text{and} \quad L_K = \mathbb{E}_{\boldsymbol{\omega}}[\phi_{\boldsymbol{\omega}} \otimes \phi_{\boldsymbol{\omega}}].$$

and Assumption 3.12 that the inequality $\|(L_K + \lambda I)^{-1/2}\phi_{\boldsymbol{\omega}}\|_\rho^2 \leq \mathcal{N}_\infty(\lambda)$ holds almost everywhere. By Lemma E.6 with $Q = L_K$ and $v_i = \phi_{\boldsymbol{\omega}_i}$ for $i = 1, \ldots, M$ we can get (30) with probability at least $1 - \delta$. Thus we complete the proof. □

Lemma D.2. *If the number of random features $M \geq 16(\mathcal{N}_\infty(\lambda) + 1)\log(2/\delta)$, then under Assumption 3.12, for any $\delta \in (0,1)$, there holds*

$$\|(L_K + \lambda I)^{-1/2}(L_K - L_M)(L_K + \lambda I)^{-1/2}\| \leq \frac{1}{2}, \tag{31}$$

and

$$\|(L_M + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\| \leq \sqrt{2}, \tag{32}$$

with probability at least $1 - \delta$.

Proof. If $M \geq 16(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)$, from Lemma D.1, we have

$$\|(L_K+\lambda I)^{-1/2}(L_K-L_M)(L_K+\lambda I)^{-1/2}\| \leq \frac{2(\mathcal{N}_\infty(\lambda)+1)\log(2/\delta)}{M} + \sqrt{\frac{2\mathcal{N}_\infty(\lambda)\log(2/\delta)}{M}} \leq \frac{1}{2}.$$

By Lemma E.3 with $A = L_K$, $B = L_M$, and $\eta = 1/2$, we have

$$\|(L_M+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\| \leq \sqrt{2}.$$

Thus we complete the proof. □

Lemma D.3. *For any $\delta \in (0,1)$, under Assumptions 3.3 and 3.12, there holds*

$$\|(L_K+\lambda I)^{-1/2}(L_K-L_M)\| \leq \frac{4\kappa\sqrt{\mathcal{N}_\infty(\lambda)}\log(2/\delta)}{M} + \sqrt{\frac{4\kappa^2\mathcal{N}(\lambda)\log(2/\delta)}{M}}, \tag{33}$$

with probability at least $1-\delta$.

Proof. Let $v_i = z_i = \phi_{\boldsymbol{\omega}_i}$ for $i = 1,\ldots,M$, then

$$T_M = \frac{1}{M}\sum_{i=1}^M v_i \otimes z_i = \frac{1}{M}\sum_{i=1}^M \phi_{\boldsymbol{\omega}_i} \otimes \phi_{\boldsymbol{\omega}_i} = L_M,$$

and

$$Q = T = \mathbb{E}[v \otimes v] = \mathbb{E}[v \otimes z] = \mathbb{E}[\phi_{\boldsymbol{\omega}} \otimes \phi_{\boldsymbol{\omega}}] = L_K.$$

Note that $\|v\|^2 = \|\phi_{\boldsymbol{\omega}}\|^2 \leq \kappa^2$ from Assumption 3.3, and $\|(L_K+\lambda)^{-1/2}\phi_{\boldsymbol{\omega}}\|^2 \leq \mathcal{N}_\infty(\lambda)$ from Assumption 3.12, and $\mathcal{N}(\lambda) = \mathrm{Tr}((L_K+\lambda I)^{-1}L_K)$, then by Lemma E.5, we have

$$\|(L_K+\lambda I)^{-1/2}(L_K-L_M)\| \leq \frac{4\kappa\sqrt{\mathcal{N}_\infty(\lambda)}\log(2/\delta)}{M} + \sqrt{\frac{4\kappa^2\mathcal{N}(\lambda)\log(2/\delta)}{M}}.$$

Thus we complete the proof. □

Lemma D.4. *Under Assumption 3.12, for any $\delta \in (0,1)$, there holds*

$$\mathcal{R}_{M,D,\lambda} \leq \frac{2(\kappa^2\lambda^{-1}+1)\log(2/\delta)}{|D|} + \sqrt{\frac{2\kappa^2\lambda^{-1}\log(2/\delta)}{|D|}}, \tag{34}$$

with probability at least $1-\delta$.

Proof. Recall that $\mathcal{R}_{M,D,\lambda} = \|(C_M+\lambda I)^{-1/2}(C_M - C_{M,D})(C_M+\lambda I)^{-1/2}\|$, we have

$$C_{M,D} = \frac{1}{|D|}\sum_{\boldsymbol{x}\in D} \boldsymbol{\phi}_M(\boldsymbol{x}) \otimes \boldsymbol{\phi}_M(\boldsymbol{x}) \quad \text{and} \quad C_M = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{\phi}_M(\boldsymbol{x}) \otimes \boldsymbol{\phi}_M(\boldsymbol{x})].$$

Note that

$$\begin{aligned}\|(C_M+\lambda I)^{-1/2}\boldsymbol{\phi}_M(\boldsymbol{x})\|_2^2 &\leq \frac{1}{\lambda}\sup_{\boldsymbol{x}\in\mathcal{X}}\|\boldsymbol{\phi}_M(\boldsymbol{x})\|_2^2 = \frac{1}{\lambda M}\sup_{\boldsymbol{x}\in\mathcal{X}}\sum_{i=1}^M |\phi(\boldsymbol{x},\boldsymbol{\omega}_i)|^2 \\ &\leq \frac{1}{\lambda M}\sum_{i=1}^M \sup_{\boldsymbol{x}\in\mathcal{X}}|\phi(\boldsymbol{x},\boldsymbol{\omega}_i)|^2 \leq \frac{1}{\lambda M}\sum_{i=1}^M \sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\omega}\in\Omega}|\phi(\boldsymbol{x},\boldsymbol{\omega}_i)|^2 \leq \kappa^2\lambda^{-1},\end{aligned} \tag{35}$$

where the second inequality is from Lemma D.2. By Lemma E.6 with $Q = C_M$ and $v_i = \boldsymbol{\phi}_M(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in D(\boldsymbol{x})$, we can get (30) with probability at least $1-\delta$. Thus we complete the proof. □

Lemma D.5. *If the number of sample $|D| \geq 16(\kappa^2\lambda^{-1}+1)\log(2/\delta)$, then under Assumption 3.12, for any $\delta \in (0,1)$, there holds*

$$\mathcal{R}_{M,D,\lambda} \leq \frac{1}{2}, \tag{36}$$

and

$$\mathcal{Q}_{M,D,\lambda} \leq \sqrt{2}, \tag{37}$$

with probability at least $1-\delta$.

Proof. If $|D| \geq 16(\kappa^2\lambda^{-1}+1)\log(2/\delta)$, from Lemma D.4, we have

$$\|(C_M+\lambda I)^{-1/2}(C_M - C_{M,D})(C_M+\lambda I)^{-1/2}\| \leq \frac{2(\kappa^2\lambda^{-1}+1)\log(2/\delta)}{M} + \sqrt{\frac{2\kappa^2\lambda^{-1}\log(2/\delta)}{M}} \leq \frac{1}{2}.$$

By Lemma E.3 with $A = C_M$, $B = C_{M,D}$, and $\eta = 1/2$, we have

$$\|(C_{M,D}+\lambda I)^{-1/2}(C_M+\lambda I)^{1/2}\| \leq \sqrt{2}.$$

Thus we complete the proof. □

E. Technical Lemmas

Lemma E.1 (Cordes Inequality (Furuta, 2001))**.** *Let A and B be positive bounded linear operators on a separable Hilbert space. Then, for any $0 < \tau \leq 1$, we have*

$$\|A^\tau B^\tau\| \leq \|AB\|^\tau.$$

Lemma E.2 (Proposition 9 in (Rudi & Rosasco, 2017))**.** *Let $\mathcal{H}, \mathcal{K}$ be two separable Hilbert spaces and X, A be bounded linear operators, with $X : \mathcal{H} \to \mathcal{K}$ and $A : \mathcal{H} \to \mathcal{H}$ be positive semi-definite, then there holds*

$$\|XA^s\| \leq \|X\|^{1-s}\|XA\|^s, \quad \forall s \in [0,1].$$

Lemma E.3 (Lemma E.2 in (Blanchard & Krämer, 2010))**.** *For any self-adjoint and positive semi-definite operators A and B, if there exists some $\eta \in [0,1]$ such that*

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

then we have

$$\|(A+\lambda I)^{1/2}(B+\lambda I)^{-1/2}\| \leq \frac{1}{\sqrt{\eta}}.$$

Lemma E.4 (Bernstein's inequality for sum of random vectors (Proposition 2 in (Rudi & Rosasco, 2017)))**.** *Let $\xi_1, \ldots, \xi_n$ be a sequence of i.i.d random variables on a separable Hilbert space $\mathcal{H}$, if there exists $\widetilde{\sigma}, \widetilde{B} \geq 0$ such that*

$$\mathbb{E}\|\xi_i - \mathbb{E}\xi_i\|_{\mathcal{H}}^p \leq \frac{1}{2}p!\widetilde{\sigma}^2\widetilde{B}^{p-2}, \quad \forall p \geq 2, \tag{38}$$

for any $0 \leq i \leq n$, then for any $\delta \in (0,1]$, there holds

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}\xi_i\right\|_{\mathcal{H}} \leq \frac{2\widetilde{B}\log(2/\delta)}{n} + \sqrt{\frac{2\widetilde{\sigma}^2\log(2/\delta)}{n}},$$

with probability at least $1-\delta$. Particularly, (38) is satisfied if

$$\|\xi\|_{\mathcal{H}} \leq \frac{\widetilde{B}}{2}, a.s. \quad \text{and} \quad \mathbb{E}\|\xi\|_{\mathcal{H}}^2 \leq \widetilde{\sigma}^2, \quad \text{or} \quad \|\xi - \mathbb{E}\xi\|_{\mathcal{H}} \leq \widetilde{B}, a.s. \quad \text{and} \quad \mathbb{E}\|\xi - \mathbb{E}\xi\|_{\mathcal{H}}^2 \leq \widetilde{\sigma}^2.$$

Lemma E.5 (Proposition 5 in (Rudi & Rosasco, 2017)). *Let $\mathcal{H}$ and $\mathcal{K}$ be two separable Hilbert spaces and $(v_1, z_1), \ldots, (v_n, z_n) \in \mathcal{H} \times \mathcal{K}$ for $n \geq 1$ be i.i.d. random variables such that there exists some constant τ such that $\|v\|_{\mathcal{H}} \leq \tau$ and $\|z\|_{\mathcal{H}} \leq \tau$ almost everywhere. Let $Q = \mathbb{E}v \otimes v$ and $T = \mathbb{E}v \otimes z$ and $T_n = \frac{1}{n}\sum_{i=1}^n v_i \otimes z_i$, then for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$,*

$$\left\|(Q + \lambda I)^{-1/2}(T - T_n)\right\|_{HS} \leq \frac{4\tau\sqrt{\mathcal{Q}_\infty(\lambda)}\log(2/\delta)}{n} + \sqrt{\frac{4\tau^2\mathcal{Q}(\lambda)\log(2/\delta)}{n}},$$

where $\mathcal{Q}_\infty(\lambda) = \sup_{v\in\mathcal{H}} \|(Q + \lambda I)^{-1/2}v\|^2$ and $\mathcal{Q}(\lambda) = \mathrm{Tr}((Q + \lambda I)^{-1/2}Q)$.

Lemma E.6 (Proposition 6 in (Rudi & Rosasco, 2017)). *Let $v_1, \ldots, v_n$ be a sequence of i.i.d random variables on a separable Hilbert spaces $\mathcal{H}$ such that $Q = \mathbb{E}v \otimes v$ is trace class, and for any $\lambda > 0$ there exists a constant $\mathcal{Q}_\infty(\lambda) < \infty$ such that $\left\langle v, (Q + \lambda I)^{-1}v\right\rangle \leq \mathcal{Q}_\infty(\lambda)$ almost everywhere. Let $Q_n = \frac{1}{n}\sum_{i=1}^n v_i \otimes v_i$, then for any $\delta \in (0, 1]$, the following holds with probability at least $1 - \delta$,*

$$\left\|(Q + \lambda I)^{-1/2}(Q - Q_n)(Q + \lambda I)^{-1/2}\right\| \leq \frac{2(\mathcal{Q}_\infty(\lambda) + 1)\log(2/\delta)}{n} + \sqrt{\frac{2\mathcal{Q}_\infty(\lambda)\log(2/\delta)}{n}}.$$

Lemma E.7 (Bousquet Inequality). *Let $Z_1, \ldots, Z_n$ be independent random elements taking values in some space $\mathcal{Z}$ and let Ξ be a class of real-valued functions on $\mathcal{Z}$, if we have*

$$\|\xi\| \leq \eta_n \quad \textit{and} \quad \frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left(\xi\left(Z_i\right)\right) \leq \zeta_n^2, \quad \forall \xi \in \Xi.$$

Define $\boldsymbol{Z} := \sup_{\xi\in\Xi}\left|\frac{1}{n}\sum_{i=1}^n \left(\xi\left(Z_i\right) - E\xi\left(Z_i\right)\right)\right|$. Then for $t > 0$

$$P\left(\boldsymbol{Z} \geq E(\boldsymbol{Z}) + t\sqrt{2\left(\zeta_n^2 + 2\eta_n E(\boldsymbol{Z})\right)} + \frac{2\eta_n t^2}{3}\right) \leq \exp\left(-nt^2\right).$$

Lemma E.8 (Proposition 10 in (Rudi & Rosasco, 2017)). *If the number of random features $M \geq (4 + 18\mathcal{N}_\infty(\lambda))\log(12\kappa^2/\lambda\delta)$, then under Assumption 3.3, for any $\delta \in (0, 1)$, there holds*

$$|\mathcal{N}_M(\lambda) - \mathcal{N}(\lambda)| \leq 1.55\mathcal{N}(\lambda),$$

with probability at least $1 - \delta$.

F. Additional numerical Experiments

In this section, we conduct some additional numerical experiments on both simulated and real-world data to demonstrate the effectiveness of random features in large kernel quantile learning tasks.

F.1. Simulated Data

For the simulated data, we consider the following two data-generating schemes that

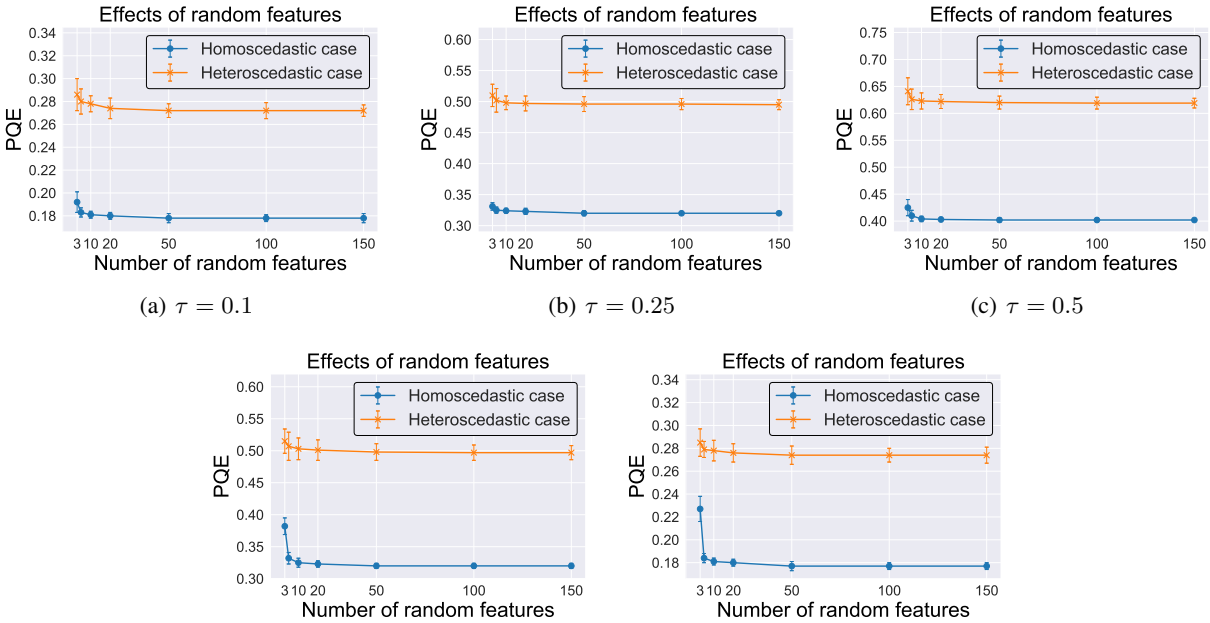
(i) Homoscedastic case:

$$y_i = \exp(-x_{i1} + x_{i2}) - x_{i2}x_{i3} + \bar{x}_i + \epsilon_i, \quad i = 1, 2, \ldots, N.$$

(ii) Heteroscedastic case:

$$y_i = \sum_{j=1}^3 \beta_i \sin(2\pi x_{ij}) + (1 + \bar{x}_i)\big(\epsilon_i - F_\epsilon^{-1}(\tau)\big), \quad i = 1, 2, \ldots, N.$$

In both cases, $x_{ij} \sim U(0, 1)$ with $\bar{x}_i = \frac{1}{p}\sum_{j=1}^p x_{ij}$, and ϵ_i follows the standard normal distribution. Moreover, in the heteroscedastic case, $\beta_i \sim U(0, 1)$ and F_ϵ^{-1} denotes the quantile function of ϵ. Clearly, the τ-th conditional quantile of y



(a) $\tau = 0.1$ (b) $\tau = 0.25$ (c) $\tau = 0.5$

(d) $\tau = 0.75$ (e) $\tau = 0.9$

Figure 4. Averaged PQE and its standard deviation against the number of random features used in KQR-RF under various scenarios.

given $\boldsymbol{x}$ is given by $f_\tau^*(\boldsymbol{x}) = f(\boldsymbol{x}) + F_\epsilon^{-1}(\tau)$ with $f(\boldsymbol{x}) = \exp(-x_{i1} + x_{i2}) - x_{i2}x_{i3} + \bar{x}_i$; and in the homoscedastic case, the τ-th conditional quantile of y given $\boldsymbol{x}$ is $f_\tau^*(\boldsymbol{x}) = \sum_{j=1}^3 \beta_i \sin(2\pi x_{ij})$.

Parameters setting. In our simulation, several scenarios are considered by varying quantile level τ from $\{0.1, 0.25, 0.5, 0.75, 0.9\}$. In all the scenarios, we employ the standard Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2)$. As suggested by Rahimi & Recht (2007) and Rudi & Rosasco (2017), the corresponding random features are taken as $\phi(\boldsymbol{x}, \boldsymbol{\omega}) = \sqrt{2}\cos(\boldsymbol{\omega}^T\boldsymbol{x} + b)$, where $\boldsymbol{\omega} \sim N(0, I)$ and $b \sim U(0, 2\pi)$. The regularization parameter λ is selected via a grid search based on a validation set with 1000 samples, where the grid is set as $\{10^{0.5s} : s = -20, -19, ..., 2\}$.

Performance evaluation. To assess the numerical performance of KQR-RF, we use the predicted quantile error (PQE) defined on a testing dataset with $n_{te} = 10000$ samples $\{\boldsymbol{x}_{te}^i, y_{te}^i\}$ as follows

$$\widehat{\mathcal{E}}_\tau(\hat{f}) = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \rho_\tau\big(y_{te}^i - \widehat{f}(\boldsymbol{x}_{te}^i)\big).$$

Our investigation delves into evaluating the influence of several critical factors including the number of random features, the sample size, and the sampling type of random features. All the reported numerical results are obtained from an average of 50 independently repeated experiments.

F.1.1. Effect of Random Feature Size

In this part, we evaluate the performance of KQR-RF by using various random features. Specifically, we consider the fixed setting that $N = 1000$ and vary the number of random features $M \in \{2, 5, 10, 50, 100, 150\}$. The results of KQR-RF are summarized in Figure 4.

As indicated in Figure 4, we can conclude that the PQE of KQR-RF tends to be smaller if a larger number of random features are used. The curves of PQEs become relatively flat as the number of random features reaches 50, which implies that the marginal gain from those increased random features is limited as enough features have been taken into account.

Table 2. Averaged PQE and its standard deviation against N of different methods in the homoscedastic case.

Quantile level	Method	Sample size				
		$N = 500$	$N = 1000$	$N = 2000$	$N = 5000$	$N = 10000$
$\tau = 0.1$	KQR-RF	0.193(0.014)	0.185(0.010)	0.178(0.007)	0.176(0.005)	0.174(0.003)
	KQR	0.184(0.005)	0.175(0.004)	0.168(0.001)	0.164(0.001)	0.162(0.001)
$\tau = 0.25$	KQR-RF	0.332(0.007)	0.325(0.005)	0.321(0.002)	0.319(0.002)	0.317(0.002)
	KQR	0.323(0.002)	0.315(0.002)	0.312(0.001)	0.308(0.001)	0.306(0.001)
$\tau = 0.5$	KQR-RF	0.419(0.008)	0.407(0.006)	0.402(0.002)	0.400(0.002)	0.398(0.001)
	KQR	0.405(0.002)	0.395(0.002)	0.391(0.002)	0.388(0.001)	0.386(0.001)
$\tau = 0.75$	KQR-RF	0.338(0.010)	0.327(0.008)	0.322(0.004)	0.319(0.002)	0.316(0.002)
	KQR	0.321(0.003)	0.313(0.003)	0.306(0.002)	0.304(0.001)	0.302(0.001)
$\tau = 0.9$	KQR-RF	0.191(0.016)	0.183(0.011)	0.177(0.005)	0.175(0.004)	0.174(0.003)
	KQR	0.183(0.005)	0.173(0.004)	0.168(0.001)	0.166(0.001)	0.165(0.001)

Table 3. Averaged PQE and its standard deviation against N and n of different methods in the heteroscedastic case.

Quantile level	Method	Sample size				
		$N = 500$	$N = 1000$	$N = 2000$	$N = 5000$	$N = 10000$
$\tau = 0.1$	KQR-RF	0.284(0.009)	0.275(0.008)	0.268(0.005)	0.266(0.004)	0.264(0.003)
	KQR	0.271(0.005)	0.266(0.004)	0.261(0.001)	0.259(0.001)	0.256(0.001)
$\tau = 0.25$	KQR-RF	0.502(0.007)	0.495(0.006)	0.487(0.003)	0.483(0.002)	0.479(0.002)
	KQR	0.492(0.002)	0.484(0.002)	0.479(0.001)	0.472(0.001)	0.470(0.001)
$\tau = 0.5$	KQR-RF	0.621(0.009)	0.614(0.006)	0.605(0.003)	0.601(0.002)	0.598(0.001)
	KQR	0.609(0.002)	0.601(0.002)	0.597(0.002)	0.594(0.001)	0.591(0.001)
$\tau = 0.75$	KQR-RF	0.501(0.009)	0.495(0.008)	0.482(0.004)	0.479(0.002)	0.476(0.002)
	KQR	0.493(0.003)	0.483(0.003)	0.474(0.002)	0.467(0.001)	0.463(0.001)
$\tau = 0.9$	KQR-RF	0.295(0.013)	0.283(0.011)	0.270(0.006)	0.266(0.004)	0.264(0.003)
	KQR	0.284(0.005)	0.275(0.004)	0.261(0.001)	0.256(0.001)	0.254(0.001)

F.1.2. EFFECT OF SAMPLE SIZE

In this part, we investigate how the performance of KQR-RF is affected by the sample size N, and we also compare it with the exact kernel quantile regression (KQR without random feature). Specifically, all the settings are exactly the same as those in Section F.1.1 except that we set $M = 50$ and vary $N \in \{500, 1000, 2000, 5000, 10000\}$, respectively.

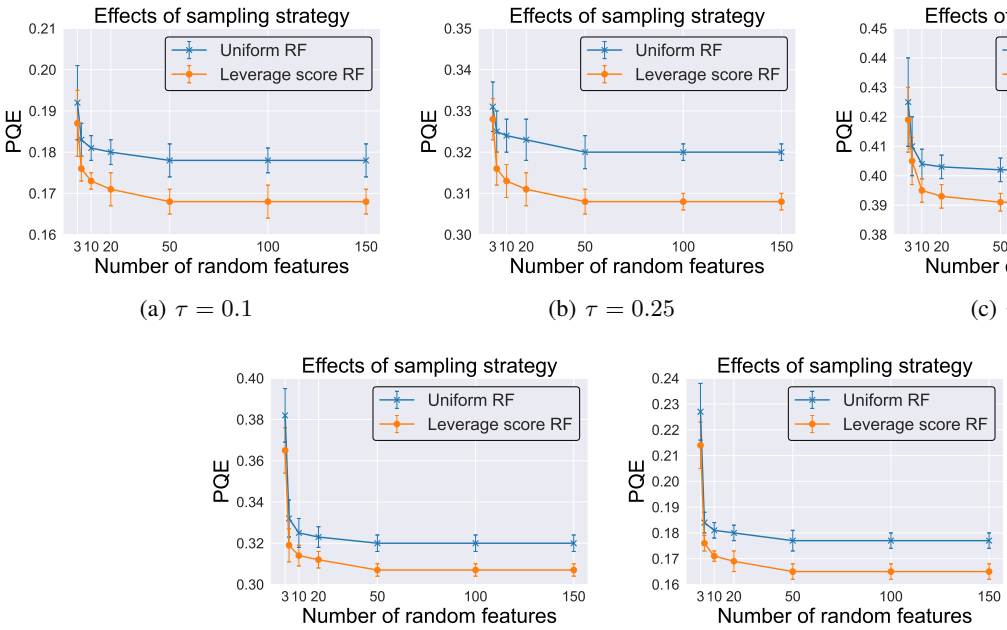
It is clear from Table F.1.2 that under the homoscedastic case, the PQE of KQR-RF decreases as N is increased at all the quantile levels, which is consistent with the theoretical result given in Theorem 3.13. Moreover, the performance of KQR-RF is near to that of KQR as the sample size increases, which shows the consistency of random feature approximation. Similar conclusions can also be drawn in the heteroscedastic case as indicated in Table F.1.2.

F.1.3. EFFECT OF SAMPLING STRATEGY

In this part, we compare the PQE of KQR-RF with different sampling strategies. Specifically, we consider the following two strategies,

1. *Uniform RF:* We generate the random features with uniform sampling strategy in (3), so the corresponding random features are $\phi(\boldsymbol{x}, \boldsymbol{\omega}_i) = \sqrt{2}\cos(\boldsymbol{\omega}^T\boldsymbol{x} + b)$, where $\boldsymbol{\omega}_i \sim N(0, I)$ and $b \sim U(0, 2\pi)$ for $i = 1, \ldots, M$.
2. *Leverage scores RF:* We generate the random features with leverage scores sampling strategy in Example 3.14. By adopting the idea of Sun et al. (2018) and Li et al. (2023b), we consider the importance ratio $q(\boldsymbol{\omega}_i) = r_i / \sum_{i=1}^M r_i$, where $\{r_i\}_{i=1}^M$ is the the diagonal of
$$\boldsymbol{\phi}_M(\boldsymbol{X})^T\boldsymbol{\phi}_M(\boldsymbol{X})\big(\boldsymbol{\phi}_M(\boldsymbol{X})^T\boldsymbol{\phi}_M(\boldsymbol{X}) + \lambda N I\big)^{-1},$$
with $\boldsymbol{\phi}_M(\boldsymbol{X}) = (\boldsymbol{\phi}_M(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}_M(\boldsymbol{x}_N))^T \in \mathbb{R}^{N\times M}$. The corresponding random features are then given as $\phi_l(\boldsymbol{x}, \boldsymbol{\omega}_i) = [q(\boldsymbol{\omega}_i)]^{-1/2}\phi(\boldsymbol{x}, \boldsymbol{\omega}_i)$ where $\phi(\boldsymbol{x}, \boldsymbol{\omega}_i)$ is the uniform RF.

Specifically, we consider the same settings as those in Section F.1.1 except that we additionally consider the data-dependent sampling strategy. The results of KQR-RF are summarized in Figures 5-6 for the homoscedastic and heteroscedastic cases, respectively.



(a) $\tau = 0.1$ (b) $\tau = 0.25$ (c) $\tau = 0.5$

(d) $\tau = 0.75$ (e) $\tau = 0.9$

Figure 5. Averaged PQE and its standard deviation against the number of random features used in KQR-RF for different sampling strategies in the homoscedastic case.

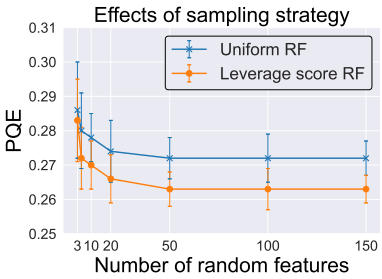
From the results in Figures 5-6, we can see that both uniform random features and leverage scores random features can achieve better performance as the number of random features increases. The data-dependent sampling strategy is more effective than the uniform sampling strategy with a fixed number of random features, which confirms our theoretical findings in Theorem 3.9 and Corollary 3.16.

Theoretical and empirical leverage scores sampling. Example 3.14 consider a leverage scores sampling strategy by using an importance ratio denoted as $q(\boldsymbol{\omega}) = l_\lambda(\boldsymbol{\omega}) / \int_{\boldsymbol{\omega}} l_\lambda(\boldsymbol{\omega}) d\pi(\boldsymbol{\omega})$, where $l_\lambda(\boldsymbol{\omega}) = \|(L_K + \lambda I)^{-1/2}\psi(\cdot, \boldsymbol{\omega})\|^2_{\rho_{\mathcal{X}}}$. The corresponding parameters $\boldsymbol{\omega}$ are sampled from distribution $\pi_l(\boldsymbol{\omega}) = q(\boldsymbol{\omega})\pi(\boldsymbol{\omega})$, and random features are $\phi_l(\boldsymbol{x}, \boldsymbol{\omega}) = [q(\boldsymbol{\omega})]^{-1/2}\phi(\boldsymbol{x}, \boldsymbol{\omega})$. This reweighted sampling ensure that $K(\boldsymbol{x}, \boldsymbol{x}') = E_{\boldsymbol{\omega}\sim\pi_l(\boldsymbol{\omega})}[\langle \phi_l(\boldsymbol{x}, \boldsymbol{\omega}), \phi_l(\boldsymbol{x}', \boldsymbol{\omega})\rangle]$. Here, we want to emphasize that Example 3.14 is a theoretical construction, and the data dependent sampling scheme is implicit due to the integral operator L_K and the expectation with respect to $\boldsymbol{x}$ in $\|(L_K + \lambda I)^{-1/2}\psi(\cdot, \boldsymbol{\omega})\|^2_{\rho_{\mathcal{X}}}$. However, in the literature, there are a lot of empirical leverage score sampling strategies that highly depend on the data. For example, the empirical random features leverages scores $\widehat{l}_\lambda(\boldsymbol{\omega}) = \widehat{\Xi}(\boldsymbol{\omega})^T(\mathbf{K} + \lambda I)^{-1}\widehat{\Xi}(\boldsymbol{\omega})^T$, with $\widehat{\Xi}(\boldsymbol{\omega}) \in \mathbb{R}^{|D|}$, $(\widehat{\Xi}(\boldsymbol{\omega}))_i = \phi_M(\boldsymbol{x}_i)$ and $\mathbf{K} = \{k(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{ij}$ is the data kernel matrix (see Remark 4 in Rudi & Rosasco (2017)). There are also some approximate leverage score sampling strategies to save the computation cost, see in Sun et al. (2018); Li et al. (2021).

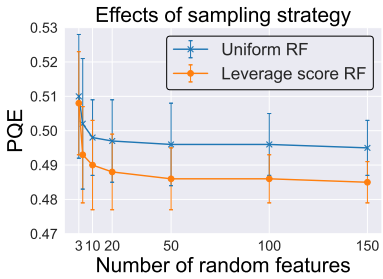
F.2. Real Case Study

In this study, we consider the UK used car prices dataset from Kaggle (https://www.kaggle.com/datasets/kukuroo3/used-car-price-dataset-competition-format). In the raw dataset, there are $N = 7632$ samples after excluding those with missing values. The response is the price of each used car, while the covariates include crucial information about the used cars, such as the registration year, mileage, road tax, miles per gallon (mpg), and engine size. We mean to predict the prices of used cars, thereby assisting car buyers in making optimal purchasing decisions.

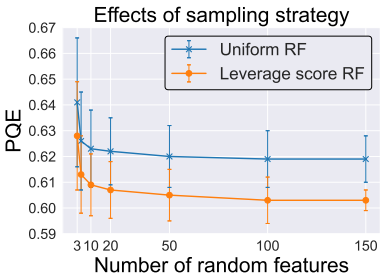
In Figure 7, we plot the histogram of the used car price, as well as the skewness and kurtosis, revealing a notable right-skew in the raw price distribution. Although this skewness is partially mitigated by applying a log transformation to the price, there is still a tail dragging on the right. To obtain more robust estimates, we consider the quantile regression with the log-transformed price as the response. Note that the sample size is large, thus it is natural to use random features to save the computing cost. In our experimental setup, we randomly choose $N_{tr} = 5000$ samples as the training data and assume



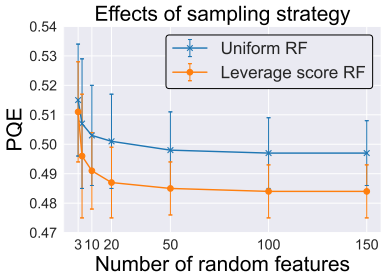
(a) $\tau = 0.1$



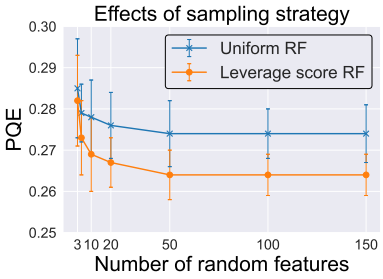
(b) $\tau = 0.25$



(c) $\tau = 0.5$

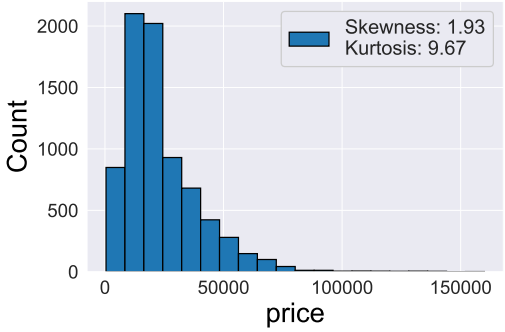


(d) $\tau = 0.75$

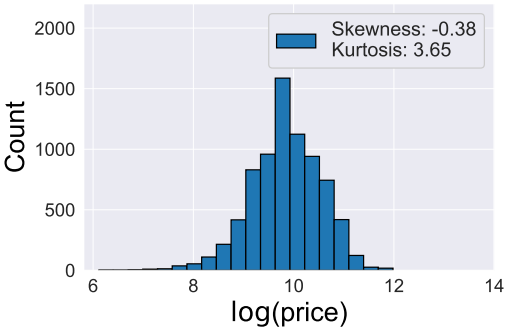


(e) $\tau = 0.9$

Figure 6. Averaged PQE and its standard deviation against the number of random features used in KQR-RF for different sampling strategies in the heteroscedastic case.



(a) Raw price



(b) Log-transformed price

Figure 7. The Histogram of the used car price and the log-transformed price.

Table 4. Averaged PQE and its standard deviation for $\tau = 0.5$ of different methods in used car price dataset.

Method	Exact KQR	KQR-RF(Uniform RF)	KQR-RF(Leverage scores RF)
PQE	0.094(0.000)	0.105(0.004)	0.098(0.003)

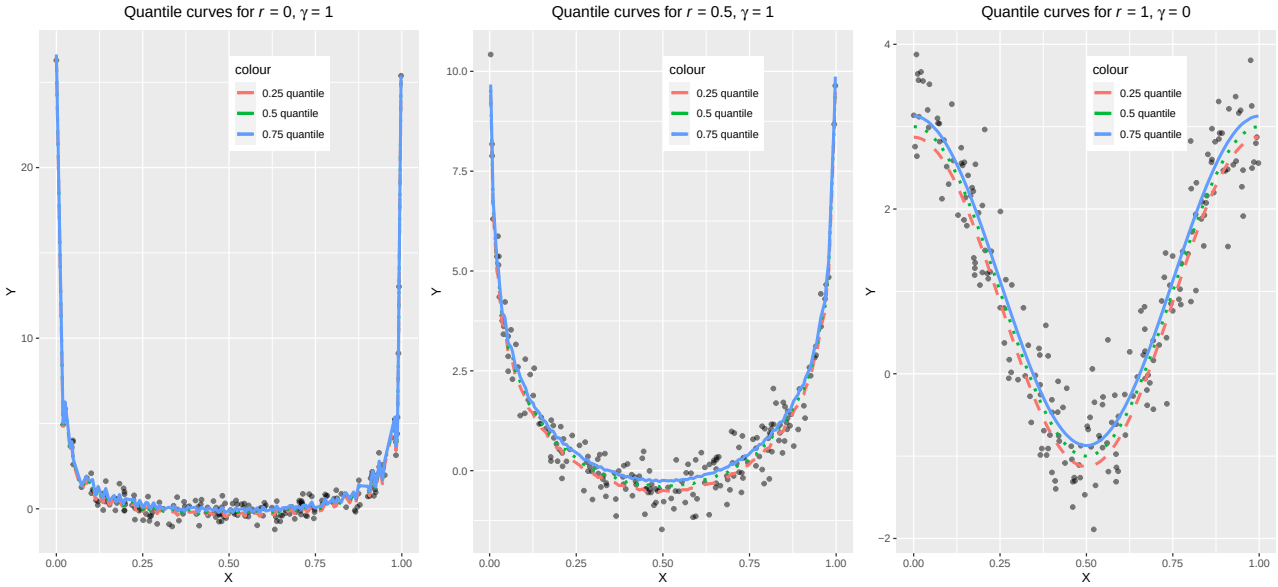


Figure 8. True quantile curves for $r = 0, \gamma = 1$ (left), $r = 1/2, \gamma = 1$ (middle), and $r = 1, \gamma = 0$ (right).

they are randomly distributed, $N_{va} = 1000$ samples as the validation data, and the rest as the testing data. For each dataset, we perform the min-max normalization for each covariate, i.e., x_{ij} is rescaled by $x_{ij} = (x_{ij} - x^i_{min})/(x^i_{max} - x^i_{min})$, where x^i_{min} and x^i_{max} is the minimum and maximum values of the i-th covariate within the entire dataset. Following the suggestion in Section F.1.1, we select the number of random features $M = 100$, and the choices of regularization parameters λ are the same as that in Section F.1.

Considering that car buyers primarily focus on the average price of used cars, we only consider the quantile level $\tau = 0.5$. Table F.2 depicts the averaged PQE and its standard deviation (50 repeats) of three methods, including the exact KQR, KQR-RF with uniform random features, and KQR-RF with leverage scores random features. Clearly, two random feature methods exhibit close performance compared to the exact KQR, especially when we use the data-dependent random features. These results further support the effectiveness of KQR-RF and substantiate the theoretical results provided in the main text.

F.3. True quantile functions in the simulation of the main text

To graphically show the quantile function at different quantile levels in the simulation of the main text , we consider three different settings: (1) worst case ($r = 0, \gamma = 1$); (2) general case ($r = 1/2, \gamma = 1$); (3) most benign case ($r = 1, \gamma = 0$). We generate data with size $N = 200$, and plot the quantile curves for $\tau \in \{0.25, 0.5, 0.75\}$ in Figure 8.