# Optimal Kernel Quantile Learning with Random Features

Caixing Wang, Xingdong Feng

School of Statistics and Management & Institute of Data Science and Statistics
Shanghai University of Finance and Economics, China

July 9, 2024

# Outline



# Introduction I

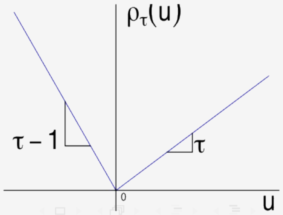
## Why Quantile Regression?

- Reason 1: Quantile regression allows us to study the impact of predictors on different quantiles of the response distribution, and thus provides a complete picture of the relationship between responses and covariates.
- Reason 2: Robust to outliers in response observations.
- Reason 3: Estimation and inference are distribution-free, and heterogeneity is usually allowed in quantile regression models.

## Loss Function

$$\rho_\tau(u) = u\{\tau - \mathcal{I}(u < 0)\},$$

where $\mathcal{I}$ is the indicator function, and $\tau$ is the quantile level.



# Introduction II


Figure 1: Financial Crisis

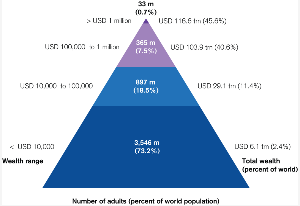



Figure 2: Income Pyramid

Figure 3: Powerful Typhoon


Figure 4: Air Pollution

# Introduction III

Suppose a random pair $(x, y)$ is drawn from an unknown joint distribution $\rho(x, y)$, consider the following quantile regression model

$$y = f^*_\tau(x) + \varepsilon, \tag{1}$$

where

- $y \in \mathbb{R}$ is the scala response;
- $x \in \mathcal{X} \subset \mathbb{R}^p$ is the p-dimensional vector of the covariate;
- $\varepsilon$ is the model error which satisfies $\mathbb{P}(\varepsilon_i < 0|x) = \tau$ for $\tau \in (0, 1)$.

Model (1) implies the following model.

**Nonparametric quantile regression**

$$Q_\tau(y_i \mid x) = f^*_\tau(x), \quad \tau \in (0, 1),$$

where $Q_\tau(\cdot|x)$ refers to the $\tau$-th conditional quantile of the response y given the covariate x.

# Introduction IV

The method of kernel quantile regression (KQR) is based on the idea of a reproducing kernel Hilbert space.

## Reproducing Kernels

Any symmetric, bounded and positive semi-definite kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines a reproducing kernel Hilbert space (RKHS), denoted by $\mathcal{H}_K$. An important property of $\mathcal{H}_K$ is the reproducing property that for any $f \in \mathcal{H}_K$, there holds

$$\langle f, K(x, \cdot) \rangle_K = f(x),$$

where $\langle \cdot, \cdot \rangle_K$ denotes the inner product in $\mathcal{H}_K$. Its equipped norm is defined as $\| \cdot \|_K^2 = \langle \cdot, \cdot \rangle_K$.

# Introduction V

Consider a standard supervised learning problem that we have a sample $D = \{(x_i, y_i)\}_{i=1}^{|D|}$, 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

$$f_{D,\lambda} = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{|D|} \sum_{i=1}^{|D|} \rho_\tau \big(y_i - f(x_i)\big) + \lambda \|f\|_K^2, \tag{2}$$

where $|D|$ is the cardinality of D and $\lambda$ is the regularization parameter controlling the model smoothness.

# Reviews and Motivation I

## Computation

According to the representer theorem (Wahba, 1990), the solution of this optimization task (2) is of finite form as given by $f_{D,\lambda}(x) = \sum_{i=1}^{|D|} \alpha_i K(x, x_i) = \boldsymbol{\alpha}^T K_N(x)$. With this solution plugged into (2), the optimization problem can be reformulated as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^{|D|}}{\operatorname{argmin}} \frac{1}{|D|} \sum_{i=1}^{|D|} \rho_\tau\left(y_i - \boldsymbol{\alpha}^T K_N(x_i)\right) + \lambda \boldsymbol{\alpha}^T K \boldsymbol{\alpha}, \tag{3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{|D|})^T \in \mathbb{R}^{|D|}$ are the representer coefficients and $K_N(x) = (K(x_1, x), \ldots, K(x_{|D|}, x))^T \in \mathbb{R}^{|D|}$, and $K = \{K(x_i, x_j)\}_{i,j=1}^{|D|}$ is the Gram matrix.

- Dual optimization (Takeuchi et al., 2006; Feng et al., 2023);
- Path-following algorithm (Li et al., 2007);
- ADMM algorithm (Boyd et al., 2011; Wang et al., 2024).

# Reviews and Motivation II

## Existing issues

- The scalability of KQR for large datasets is limited due to the expensive computational complexity ($\mathcal{O}(|\mathrm{D}|^3)$) and storage requirements ($\mathcal{O}(|\mathrm{D}|^2)$) when $|\mathrm{D}|$ is large.
- The theoretical investigation of KQR is not clear and deep enough (Suboptimal or capacity-independent).
- Most work assume the realizable setting, i.e., $f_\tau^* \in \mathcal{H}_K$, does KQR work in the agnostic setting, i.e., $f_\rho \notin \mathcal{H}_K$?

Question: Can we find some accelerated methods that can achieve a optimal trade-off between the computation and theory, especially in the agnostic settings?

# Random Fourier Features I

The following classical theorem from harmonic analysis provides the key insight behind random feature mapping:

## Bochner's theorem

A continuous kernel $K(x, x') = K(x - x')$ on $\mathbb{R}^p$ is positive definite if and only if $K(x - x')$ is the Fourier transform of a non-negative measure.

## Unbiased feature mapping

If a shift-invariant kernel $K(\cdot, \cdot)$ is properly scaled, Bochner's theorem guarantees that its Fourier transform $\pi(\boldsymbol{\omega})$ is a proper probability distribution. Define $\phi(x, \boldsymbol{\omega}) = e^{j\boldsymbol{\omega}^T x}$ we have

$$K(x, x') = K(x - x') = \int_{\mathbb{R}^p} \pi(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}^T (x - x')} d\boldsymbol{\omega} = \mathbb{E}_{\boldsymbol{\omega}} \left[ \phi(x, \boldsymbol{\omega}) \phi^*(x', \boldsymbol{\omega}) \right], \quad (4)$$

where $*$ denoting the Hermitian transpose. So $\phi(x, \boldsymbol{\omega}) \phi^*(x', \boldsymbol{\omega})$ is an unbiased estimator of $K(x, x')$ when $\boldsymbol{\omega}$ is drawn from $\pi(\boldsymbol{\omega})$.

# Random Fourier Features II

Since both $\pi(\boldsymbol{\omega})$ and $K(\cdot,\cdot)$ are real, the integral (4) converges when the complex exponentials are replaced with cosines.

**Real-valued feature mapping**

A real-valued mapping that satisfies the condition $K(x, x') = \mathbb{E}_{\boldsymbol{\omega}}\left[\phi(x, \boldsymbol{\omega})\phi^*(x', \boldsymbol{\omega})\right]$ can be obtained by setting

$$\phi(x, \boldsymbol{\omega}) = \sqrt{2}\cos(\boldsymbol{\omega}^T x + b),$$

where $\boldsymbol{\omega}$ is drawn from $\pi(\boldsymbol{\omega})$ and b is drawn uniformly from $[0, 2\pi]$.

# Random Fourier Features III

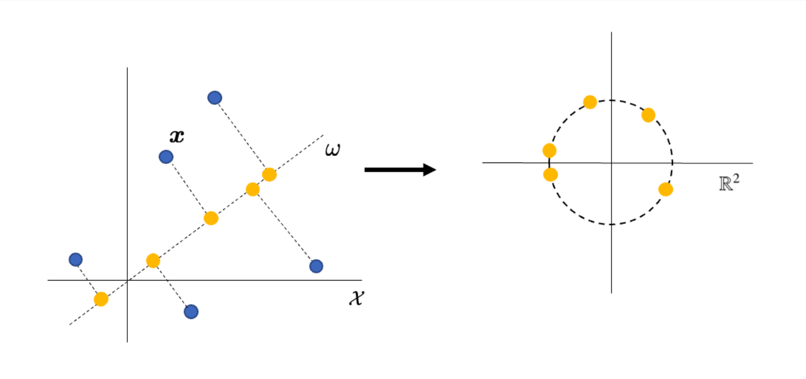



Figure 5: Random Fourier Features. Each component of the feature map $\phi(\mathrm{x}, \boldsymbol{\omega}) = \sqrt{2}\cos(\boldsymbol{\omega}^{\mathrm{T}}\mathrm{x} + \mathrm{b})$ projects x onto a random direction $\boldsymbol{\omega}$ drawn from the Fourier transform $\pi(\boldsymbol{\omega})$, and wraps this line onto the unit circle in $\mathbb{R}^2$. After transforming two points x and $\mathrm{x}'$ in this way, their inner product is an unbiased estimator of $\mathrm{K}(\mathrm{x}, \mathrm{x}')$. The mapping additionally rotates this circle by a random amount b and projects the points onto the interval $[0, 1]$ (Rahimi and Recht, 2007).

Table 1: Some examples of shift invariant kernels and their Fourier transforms (Rahimi and Recht, 2007).

| Kernel Name | $K(\Delta)$ | $\pi(\boldsymbol{\omega})$ |
|---|---|---|
| Gaussian Kernel | $e^{-\frac{\|\Delta\|_2^2}{2}}$ | $(2\pi)^{\frac{D}{2}} e^{-\frac{\|\boldsymbol{\omega}\|_2^2}{2}}$ |
| Laplacian Kernel | $e^{-\|\Delta\|_1}$ | $\prod_d \frac{1}{\pi(1+\boldsymbol{\omega}_d^2)}$ |
| Cauthy Kernel | $\prod_d \frac{1}{(1+\Delta_d^2)}$ | $e^{-\|\Delta\|_1}$ |

# Kernel Approximation

## Kernel approximation with random features

It is thus clear that we can adopt the standard Monte Carlo sampling method to estimate $\mathrm{K}(\mathrm{x}, \mathrm{x}')$ by

$$\mathrm{K}_{\mathrm{M}}(\mathrm{x}, \mathrm{x}') = \langle \boldsymbol{\phi}_{\mathrm{M}}(\mathrm{x}, \boldsymbol{\omega}), \boldsymbol{\phi}_{\mathrm{M}}(\mathrm{x}', \boldsymbol{\omega}) \rangle,$$

where $\boldsymbol{\phi}_{\mathrm{M}}(\mathrm{x}, \boldsymbol{\omega}) = \frac{1}{\sqrt{\mathrm{M}}} \left( \phi(\mathrm{x}, \boldsymbol{\omega}_1), \ldots, \phi(\mathrm{x}, \boldsymbol{\omega}_{\mathrm{M}}) \right)^{\mathrm{T}}$ is the feature map and $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{\mathrm{M}}$ are independently sampled with respect to $\pi$.

Remark: In addition to the shift invariant kernel, any kernel has the following integral representation can use the above approximation,

$$\mathrm{K}(\mathrm{x}, \mathrm{x}') = \int_{\Omega} \phi(\mathrm{x}, \boldsymbol{\omega}) \phi(\mathrm{x}', \boldsymbol{\omega}) \mathrm{d}\pi(\boldsymbol{\omega}), \tag{5}$$

# RKHS Approximation

## RKHS approximation with random features

Define a M-dimensional function space $\mathcal{H}_M$ related to $\boldsymbol{\phi}_M(x)$ as

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

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

# Illustration

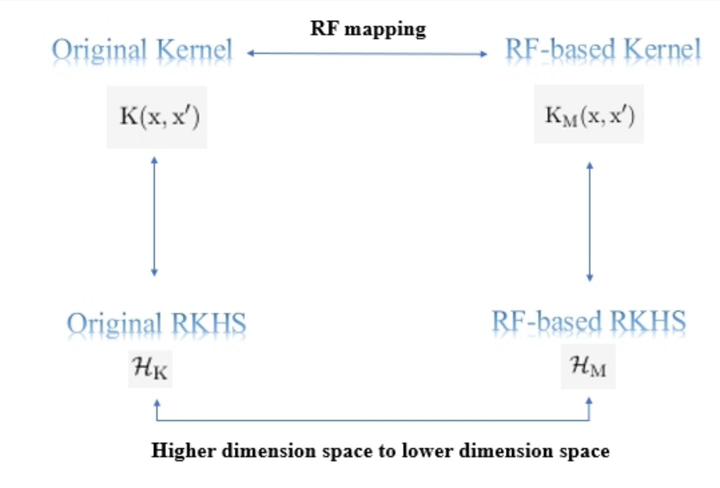



Figure 6: A simple illustration of kernel and RKHS approximation using RF.

# KQR with Random Features I

## KQR-RF

Different from KQR, KQR with random features (KQR-RF) estimates a function in the approximation RKHS $\mathcal{H}_M$

$$f_{M,D,\lambda} = \operatorname*{argmin}_{f \in \mathcal{H}_M} \frac{1}{|D|} \sum_{(x,y) \in D} \rho_\tau\left(y - f(x)\right) + \lambda \|f\|_{\mathcal{H}_M}^2, \tag{6}$$

# KQR with Random Features II

## Computation

According to the representer theorem, the solution of (3) with random features can be written as

$$f_{M,D,\lambda}(x) = \hat{u}^T \phi_M(x), \tag{7}$$

and the optimization problem becomes

$$\hat{u} = \operatorname*{argmin}_{u \in \mathbb{R}^M} \frac{1}{|D|} \sum_{i=1}^{|D|} \rho_\tau \left(y_i - u^T \phi_M(x_i)\right) + \lambda u^T u. \tag{8}$$

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 $M \ll |D|$.

# Theoretical Goal

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

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

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 $\|f - f_\tau^*\|_\rho^2$.

# Definitions and Assumptions I

## Definition 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

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

## Definition 2 (Effective dimension)

For $\lambda > 0$, we define the effective dimension of kernel $K$ and $K_M$ as

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

# Definitions and Assumptions II

## Assumption 1 (Bounded and continuous random features)

Assume kernel $K$ has the integral representation defined in (5) with $\phi$ bounded and continuous in both variables, that is, there exists some constant $\kappa \geq 1$ such that $|\phi(x, \boldsymbol{\omega})| \leq \kappa$ for any $x \in \mathcal{X}$ and $\boldsymbol{\omega} \in \Omega$. The associated RKHS $\mathcal{H}_K$ is separable.

## Assumption 2 (Source condition)

Suppose there exists $R > 0$, $r > 0$ and $h_\tau \in L^2_{\rho_\mathcal{X}}$ such that

$$f^*_\tau = L^r_K h_\tau, \tag{9}$$

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

# Definitions and Assumptions III

## Remark

- The parameter $r$ controls the size of the functional class of $f^*_\tau$. When $r \in [1/2, 1]$, the functional class $\mathcal{C}$ 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 $\mathcal{C}$ 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 $f^*_\tau \notin \mathcal{H}_K$.

## Assumption 3 (Capacity condition)

For $\lambda > 0$, there exists $Q > 0$ and $\gamma \in [0, 1]$ such that

$$\mathcal{N}(\lambda) \leq Q^2 \lambda^{-\gamma}. \tag{10}$$

# Definitions and Assumptions IV

- For kernel ridge reression (KRR) and Kernel ridge regression with random features (KRR-RF), the minimax optimal capacity-dependent rate has been shown to be $\mathcal{O}(|D|^{\frac{2r}{2r+\gamma}})$ (Caponnetto and De Vito, 2007; Rudi and Rosasco, 2017).
- Whether KQR-RF can achieve the above optimal learning rate even under the agnostic settings?

**Assumption 4 (Adaptive self-calibration condition)**

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

$$|F_{y|x}(y+\delta) - F_{y|x}(y)| \ge c_2|\delta|, \tag{11}$$

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

- For example, if y has a density that is bounded away from zero on some compact interval around $f^*_\tau(x)$, then Assumption 4 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 and 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 and Chernozhukov (2011) requiring the conditional density of Y given x to be continuously differentiable and bounded away from zero uniformly for all $\tau \in (0,1)$ and all 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 and Chatterjee (2022).

# Existing Theorem

**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[a], 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.

[a] Assumption 1, Assumption 2 with $r = 1/2$, eigenvalue decaying assumption (stronger than Assumption 3), and the local strongly convex assumption which can be derived from Assumption 4.

# Sharper Learning Rates for KQR-RF I

**Theorem 1**

Under Assumptions 1-4, if $r \in (0,1]$, $\gamma \in [0,1]$, and set $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies

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

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.

# Sharper Learning Rates for KQR-RF II

- The capacity-dependent learning rates obtained in Theorem 1 align with those of KRR (Caponnetto and De Vito, 2007) and KRR-RF (Rudi and Rosasco, 2017), which is minimax optimal and thus can not be improved any further.
- Compared to Lian (2022), we relax the regularity condition from $r \in [1/2, 1]$ to $r \in (0, 1]$, covering a wider range of scenarios.

**Remark**

Theorem 1 uses the naive uniform sampling strategy for the random features (generate $\phi(x, \boldsymbol{\omega})$ 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 and 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.

# Refined Analysis: Beyond Uniform Sampling I

**Assumption 5 (Compatibility condition)**

Define the maximum dimension of random features as

$$\mathcal{N}_\infty(\lambda) = \sup_{\boldsymbol{\omega}\in\Omega} \left\| (\mathrm{L_K} + \lambda \mathrm{I})^{-1/2} \phi(\cdot, \boldsymbol{\omega}) \right\|_{\rho_{\mathcal{X}}}^2, \tag{12}$$

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

Recall the definition of $\mathcal{N}(\lambda)$ in Definition 2. $\mathcal{N}(\lambda)$ and $\mathcal{N}_\infty(\lambda)$ measure the average and supreme capacities of $\mathcal{H}_\mathrm{K}$, respectively, so we have

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

where $\mathrm{E}_{\boldsymbol{\omega}}$ denotes the expectation taking over $\boldsymbol{\omega}$.

# Refined Analysis: Beyond Uniform Sampling II

## Theorem 2

Under Assumptions 1-5, if $r \in (0,1]$, $\gamma \in [0,1]$, and set $\lambda = |D|^{-\frac{1}{2r+\gamma}}$, when the number of random features satisfies

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

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.

# Refined Analysis: Beyond Uniform Sampling III

- The above capacity-dependent learning rate is the same as that of Theorem 1, 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 Assumption 5.
- By adopting a favorable sampling strategy called leverage scores sampling strategy, we can further reduce the required number of random features and achieve the optimal learning rates across the entire range of $r \in (0, 1]$.

**Leverage scores sampling**

Given the integral representation of kernel K as stated in (5), 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 $\phi_l(x, \boldsymbol{\omega}) = [q(\boldsymbol{\omega})]^{-1/2}\phi(x, \boldsymbol{\omega})$ and exhibit a distribution $\pi_l(\boldsymbol{\omega}) = q(\boldsymbol{\omega})\pi(\boldsymbol{\omega})$.

# Refined Analysis: Beyond Uniform Sampling IV

- As pointed out in Rudi and Rosasco (2017), the random features provide the integral representation of K and satisfy Assumption 5 with $\alpha = \gamma$ indicating that $\mathcal{N}(\lambda) = \mathcal{N}_\infty(\lambda)$.

## Corollary 1

Under Assumptions 1-5, if random features are sampled according to the leverage scores sampling strategy, $\mathrm{r} \in (0,1]$, $\gamma \in [0,1]$, and set $\lambda = |\mathrm{D}|^{-\frac{1}{2\mathrm{r}+\gamma}}$, when the number of random features satisfies

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

and $|\mathrm{D}|$ is sufficiently large, there holds

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

with probability near to 1.

# Refined Analysis: Beyond Uniform Sampling V

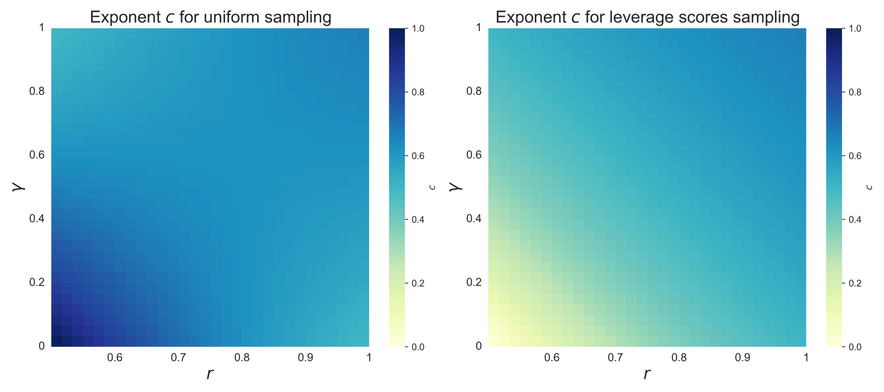



Figure 7: Comparison between the number of random features $M = \mathcal{O}(|D|^c)$ required for uniform sampling ($\alpha = 1$, left) and leverage scores sampling ($\alpha = \gamma$, right) in the realizable case.

# Refined Analysis: Beyond Uniform Sampling VI

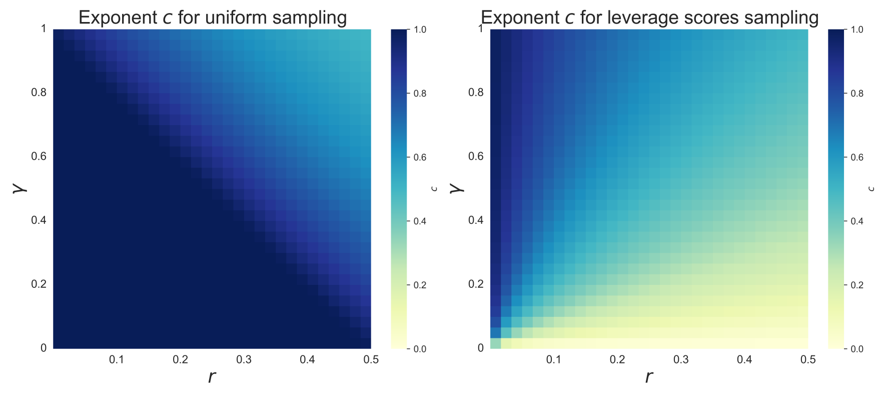



Figure 8: Comparison between the number of random features $M = \mathcal{O}(|D|^c)$ required for uniform sampling ($\alpha = 1$, left) and leverage scores sampling ($\alpha = \gamma$, right) in the agnostic case.

# Comparisons to the Related Work

Table 2: Summary of conditions for derived learning rates in different methods.

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

# Simulation Study I

## Spline kernel

We consider the spline kernel of order q, defined as

$$\Lambda_q\left(x, x'\right) = \sum_{k=-\infty}^{\infty} e^{2\pi i k x} e^{-2\pi i k x'} |k|^{-q},$$

where $x, x' \in [0,1]$, and $q \in \mathbb{R}$. According to the property of spline kernel, we have

$$\int_0^1 \Lambda_q(x, z)\Lambda_{q'}\left(x', z\right) dz = \Lambda_{q+q'}\left(x, x'\right),$$

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 $w \sim U(0,1)$.

# Simulation Study II

## Simulation setting

Data are generated from the following model

$$y = \Lambda_{\frac{r}{\gamma}+\frac{1}{2}}(x, 0) + \varepsilon,$$

where $\varepsilon \sim N(0, 0.01)$ and $x \sim U(0, 1)$.

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 $(r = 1, \gamma = 0)$.

# Results I

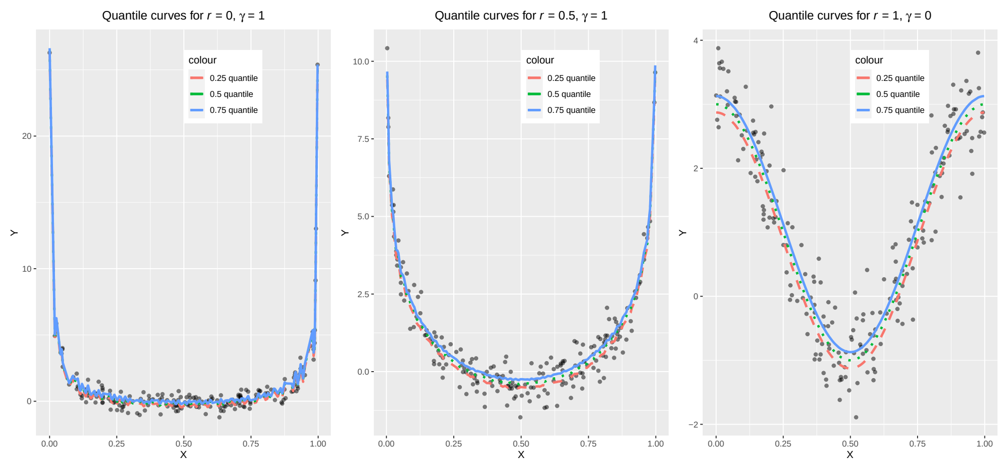



Figure 9: True quantile curves for $r = 0, \gamma = 1$ (left), $r = 1/2, \gamma = 1$ (middle), and $r = 1, \gamma = 0$ (right).

# Results II

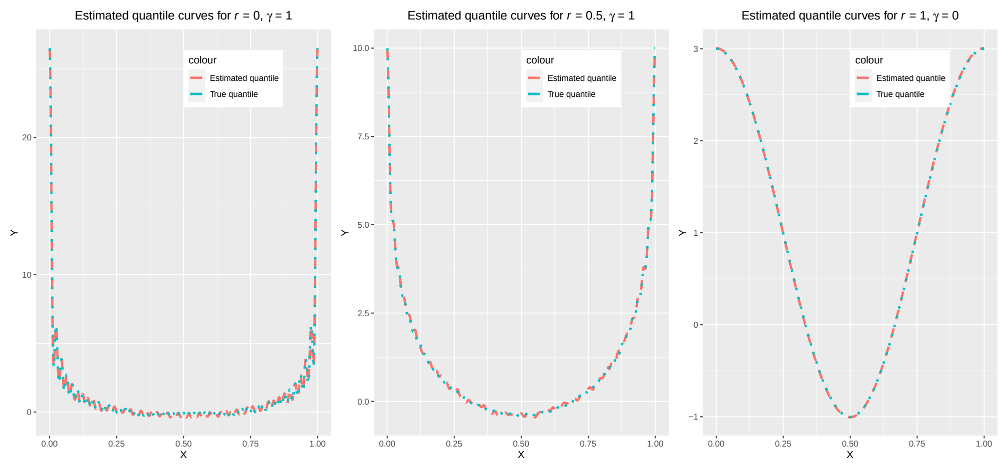



Figure 10: Estimated and true quantile curves for $r = 0, \gamma = 1$ (left), $r = 1/2, \gamma = 1$ (middle), and $r = 1, \gamma = 0$ (right) when $\tau = 0.5$.

# Learning Rates Validation

- 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 log-transformed 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.

# Results III

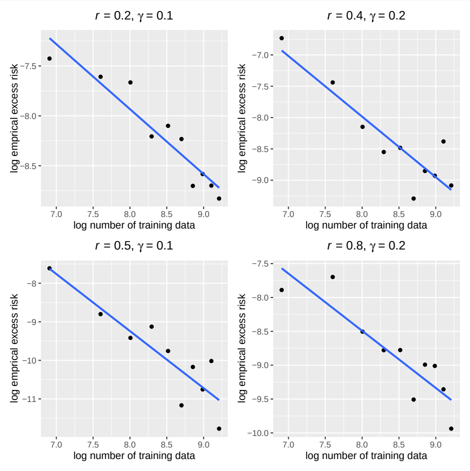



Figure 11: Log empirical excess risk for $r = 0.2, \gamma = 0.1$ (left top), $r = 0.4, \gamma = 0.2$ (right top), $r = 0.5, \gamma = 0.1$ (left bottom) and $r = 0.8, \gamma = 0.2$ (right bottom) when $\tau = 0.5$.

# Discussion

- First two figures shows that KQR-RF can estimate the quantile functions very well both in realizable and agnostic settings.
- Last figure 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.

# 谢 谢!

# Thank You!

# References


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

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

Belloni, A. and Chernozhukov, V. (2011). $\ell_1$-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics, 39(1):82.

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

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

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

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

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

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

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

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

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

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

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

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

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., and Huang, J. (2021). Deep quantile regression: Mitigating the curse of dimensionality