Some Topics on Regression Problem

Zhan Yu

City University of Hong Kong

CAST, China 9th Sep. 2021

Outline

- Background of the problem
- Distribution regression
- Multi-penalty distribution regression
- Distributed learning
- Notations and assumptions
- Main results on error bounds and learning rates
- Sketch of the way for deriving learning rates

Outline

Background of the problem: kernel-based regression

- **Working space**: In a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ induced by a Mercer kernel K on an input compact metric space X.

  • $\mathcal{H}_K$: closure of the linear span of the sets of the functions $\{K_x = K(x,\cdot) : x \in X\}$ with the (unique) inner product denoted as $\langle\cdot,\cdot\rangle_K$ satisfying $\langle K_x, K_y\rangle_K = K(x,y)$.

  • $D = \{(x_i, y_i)\}_{i=1}^{|D|} \subset X \times Y$ is the sample set with samples drawn from a probability measure ρ on $X \times Y$ where $Y = \mathbb{R}$ is the output space.

- **Regression scheme**: The classical regularized least squares scheme in machine learning can be stated as

$$f_{D,\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{i=1}^{|D|} \big(f(x_i) - y_i\big)^2 + \lambda \big\| f \big\|_K^2 \right\}, \qquad (1)$$

  Here $\lambda > 0$ is a regularization parameter. This learning algorithm is also called *kernel-based regression* in statistics and has been well studied in learning theory.

Objective of learning theory

The objective of learning theory is to find an unknown function $f : X \to Y$ from random samples $D = (x_i, y_i)_{i=1}^{|D|}$.

- Suppose that a probability measure ρ on $X \times Y$ governs random sampling. Let X be a compact metric space and $Y = \mathbb{R}$. If we define the (least square) error of f as

$$\mathcal{E}(f) = \int_{X \times Y} (f(x) - y)^2 d\rho. \tag{2}$$

  Then the function that minimizes the error is the regression function f_ρ:

$$f_\rho(x) = \int_{\mathbb{R}} y d\rho(y|x), x \in X. \tag{3}$$

- measure ρ is **unknown**. Hence neither f_ρ nor $\mathcal{E}(f)$ is computable. In learning theory, one approximates f_ρ by the function minimizing the empirical error $\mathcal{E}_D$ w.r.p.t. the sample D:

$$\mathcal{E}_D = \frac{1}{|D|} \sum_{i=1}^{|D|} (f(x_i) - y_i)^2. \tag{4}$$

Outline

From classical regression to distribution regression

Some motivations

- In the era of big data, **functional data**, **matrix-valued data** or **distribution data** has appeared in an increasingly number of practical circumstances in machine learning and statistics. Instead of scalar data setting, they cause more challenges and difficulties in handling complicated information in these settings.
- Classical regression method may not have a nice performance when treating with **non-scalar data**. Developing suitable regression scheme for these problems becomes desirable.

Brief introduction to distribution regression

Two stage scheme: With the purpose of learning the regressor from the distribution to the real valued outputs, the method contains two stages of sampling. The first stage sample is made up of **distributions** and the second stage sample is **drawn according to these distributions**.

- First stage: we define the data set as $\widetilde{D} = \{(x_i, y_i)\}_{i=1}^{|D|} \subset X \times Y$, in which $|D|$ is cardinality of $\widetilde{D}$ and each pair (x_i, y_i) is i.i.d. sampled from a meta distribution. X is the **input space** of probability measures on a compact metric space $\widetilde{X}$ and $Y = \mathbb{R}$ is the **output space** equipped with the standard Euclidean metric.
- Second stage: the samples in sample set $\hat{D} = \left\{\left(\{x_{i,s}\}_{s=1}^{d_i}, y_i\right)\right\}_{i=1}^{|D|}$ are obtained from distributions $\{x_i\}_{i=1}^{|D|}$ accordingly, where $x_{i,j} \in \widetilde{X}$.

Brief introduction to distribution regression

Mean embedding based kernel ridge regression:
In this work, we consider a **mean embedding based ridge regression** method for distribution regression.

- Let $H = H(\widetilde{K})$ be a reproducing kernel Hilbert space (RKHS) with $\widetilde{K} : \widetilde{X} \times \widetilde{X} \to \mathbb{R}$ as the reproducing kernel. Let $(\widetilde{X}, \mathcal{F})$ be a measurable space with $\mathcal{F}$ being a Borel σ-algebra on $\widetilde{X}$. Denote the set of Borel probability measures on $(\widetilde{X}, \mathcal{F})$ by $\mathcal{M}_1(\mathcal{F})$. Then the *mean embedding* of a distribution $x \in \mathcal{M}_1(\mathcal{F})$ to an element μ_x of RKHS H is given by

$$\mu_x = \int_{\widetilde{X}} \widetilde{K}(\cdot, \xi) dx(\xi).$$

- Denote the set of the mean embeddings by $X_\mu = \{\mu_x : x \in \mathcal{M}_1(\mathcal{F})\} \subseteq H$. Then we can denote the **mean embeddings** of $\widetilde{D}$ to X_μ by $D = \{(\mu_{x_i}, y_i)\}_{i=1}^{|D|}$.

Brief introduction to distribution regression

Let ρ be the μ-induced probability measure on the product space $Z = X_\mu \times Y$. The regression function of ρ w.r.t. the (μ_x, y)-pair is defined by

$$f_\rho(\mu_x) = \int_Y y d\rho(y|\mu_x), \ \mu_x \in X_\mu, \tag{5}$$

in which $\rho(\cdot|\mu_x)$ is the conditional probability measure of ρ induced at $\mu_x \in X_\mu$. f_ρ is just the **optimizer** of the least square problem

$$\min \mathcal{E}(f) = \int_{X_\mu \times Y} (f(\mu_x) - y)^2 d\rho.$$

Generally, measure ρ is **unknown**. In distribution regression circumstance, the distributions $\{x_i\}_{i=1}^{|D|}$ are **still unknown**. We are only able to approach the information of them by the random sample $\hat{D} = \left\{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\right\}_{i=1}^{|D|}$ with size $d_i \in \mathbb{N}$, $i = 1, 2, ..., |D|$ respectively.

Two-stage distribution regression scheme

In a reproducing kernel Hilbert space $(\mathcal{H}_K, \|\cdot\|_K)$ associated with a Mercer kernel $K : X_\mu \times X_\mu \to \mathbb{R}$, the classical regularization approach in distribution regression setting has the following functional optimization scheme,

$$f_{\hat{D},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{i=1}^{|D|} \left(f(\mu_{\hat{x}_i}) - y_i\right)^2 + \lambda \left\|f\right\|_K^2 \right\}, \qquad (6)$$

in which $\hat{x}_i = \frac{1}{d_i}\sum_{s=1}^{d_i} \delta_{x_{i,s}}$ serves as the empirical distribution determined by the observable quantity $\hat{D} = \left\{\left(\{x_{i,s}\}_{s=1}^{d_i}, y_i\right)\right\}_{i=1}^{|D|}$, $\mu_{\hat{x}_i} = \frac{1}{d_i}\sum_{s=1}^{d_i} \widetilde{K}(\cdot, x_{i,s})$ is the mean embedding, $\lambda > 0$ is the regularization parameter. (6) is essentially a **Tikhonov regularized scheme** in RKHS. It is an extension of one-stage kernel ridge regression scheme.

Outline

A glimpse of merits of multi-penalty regularization scheme (additional regularization term added)

- It can incorporate any prior information into additional penalties, thus can simultaneously include various features in regularized solution such as boundedness, monotonicity, smoothness.
- With these merits, multi-penalty regularization scheme has been widely used in a variety of inspiring applications like image reconstruction, earth gravity potential reconstruction, option pricing model, data detection, sparsity analysis.

Distribution regression with multi-penalty regularization

We investigate a more general framework in two-stages distribution regression by considering the novel multi-penalty regularization scheme:

$$f_{\hat{D},\lambda_1,\lambda_2} = \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{|D|}\sum_{i=1}^{|D|}\left(f(\mu_{\hat{x}_i})-y_i\right)^2+\lambda_1\left\|f\right\|_K^2+\lambda_2\left\|V_D f\right\|_K^2\right\}. \tag{7}$$

In (7), V_D is a bounded operator associated with data set $\widetilde{D}$ and its associated mean embedding set D.

- **The goal of distribution regression**: to learn the regression function f_ρ with algorithm (7) from the given training samples $\hat{D} = \left\{\left(\{x_{i,j}\}_{j=1}^{d_i}, y_i\right)\right\}_{i=1}^{|D|}$ with $x_{i,1}, x_{i,2}, ..., x_{i,d_i} \sim x_i$ (i.i.d.). We aim to investigate the learning rates of multi-penalty distribution regression (7).

Outline

Distributed learning

Some motivations for development of distributed learning algorithm:

- in real world, with the development of data mining, large-scale data are always collected in variety of application domains including financial engineering, medicine, business analysis, personal social network, sensor network and monitoring. Sensitive data such as personal data are always trained in ways of machine learning for different requirements. Hence, it is extremely important to **protect data privacy**. Recently, distributed learning has been shown to be a powerful strategy to tackle **privacy-preserving** problems.
- unprecedented large data size and complexity of distribution samples would always raise the difficulties of computing in distribution regression approach. Large-scale data would add unpredictable **storage burden and memory capacity** for a single machine.
- It would take a **huge amount of time** for a single machine to process scientific computing on distribution regression.

Distributed learning with multi-penalty distribution regression

- Based on a **divide-and-conquer** approach, we partition the data set $\widetilde{D}$ into m disjoint subsets $\{\widetilde{D}_j\}_{j=1}^m$ with corresponding decomposition of sets of mean embeddings $\{D_j\}_{j=1}^m$ of D. Then, assign each subset $\widetilde{D}_j$ to a local machine to produce a local estimator $f_{\hat{D}_j,\lambda_1,\lambda_2}$ in RKHS by the multi-penalty distribution regression scheme (7).
- After a communication of these local estimators, the global estimator $\bar{f}_{\hat{D},\lambda_1,\lambda_2}$ is obtained by taking following average

$$\bar{f}_{\hat{D},\lambda_1,\lambda_2} = \sum_{j=1}^{m} \frac{|D_j|}{|D|} f_{\hat{D}_j,\lambda_1,\lambda_2} \tag{8}$$

of the local estimators $\{f_{\hat{D}_j,\lambda_1,\lambda_2}\}_{j=1}^m$. For algorithm (8), the learning theory analysis for $\bar{f}_{\hat{D},\lambda_1,\lambda_2}$ is carried out via integral operator approach.

Notations and assumptions

- there exist a constant $M > 0$ such that $|y| \leq M$ almost surely.
- $\widetilde{K}$ and K are **bounded** Mercer kernel (symmetric, continuous, positive semidefinite) with bound $B_{\widetilde{K}}$ and κ:

$$B_{\widetilde{K}} = \sup_{v \in \widetilde{X}} \widetilde{K}(v, v) < \infty, \ \kappa = \sup_{\mu_u \in X_\mu} \sqrt{K(\mu_u, \mu_u)} < \infty.$$

- Suppose that $\alpha \in (0, 1]$ and $L > 0$. Denote $\mathcal{L}(Y, \mathcal{H}_K)$ as the Banach space of the bounded linear operators from space $Y = \mathbb{R}$ to $\mathcal{H}_K$. Denote $K_{\mu_x} = K(\mu_x, \cdot)$, $\mu_x \in X_\mu$. We treat K_{μ_x} as an element of $\mathcal{L}(Y, \mathcal{H}_K)$ by defining the linear mapping

$$K_{\mu_x}(y) = yK_{\mu_x}, \ y \in Y.$$

We assume that the mapping $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is (α, L)-**Holder continuous** in following sense

$$\left\| K_{\mu_x} - K_{\mu_y} \right\|_{\mathcal{L}(Y, \mathcal{H}_K)} \leq L \left\| \mu_x - \mu_y \right\|_H^\alpha, \ \forall (\mu_x, \mu_y) \in X_\mu \times X_\mu. \tag{9}$$

Outline

Notations and assumptions

- Denote ρ_{X_μ} to be the marginal distribution of ρ on X_μ. Let $L^2_{\rho_{X_\mu}}$ be the Hilbert space of square-integrable functions defined on X_μ.
- For $f \in L^2_{\rho_{X_\mu}}$, denote the norm of f by

$$\|f\|_\rho = \|f\|_{L^2_{\rho_{X_\mu}}} = \langle f, f \rangle^{1/2}_{\rho_{X_\mu}} = \left(\int_{X_\mu} |f(\mu_x)|^2 d\rho_{X_\mu}(\mu_x) \right)^{1/2}.$$

  Define the integral operator L_K on $L^2_{\rho_{X_\mu}}$ associated with the Mercer kernel $K : X_\mu \times X_\mu \to \mathbb{R}$ as

$$L_K(f) = \int_{X_\mu} K_{\mu_x} f(\mu_x) d\rho_{X_\mu}, \ f \in L^2_{\rho_{X_\mu}}. \tag{10}$$

- **Regularity assumption:** we assume the following *regularity condition* for the regression function f_ρ:

$$f_\rho = L^r_K(g_\rho) \text{ for some } g_\rho \in L^2_{\rho_{X_\mu}}, \ r > 0. \tag{11}$$

  The special case $r = 1/2$ corresponds to $f_\rho \in \mathcal{H}_K$.

Notations and assumptions

- We use the *effective dimension* $\mathcal{N}(\lambda_1)$ to measure the capacity of $\mathcal{H}_K$ with respect to measure ρ_{X_μ}, which is defined to be the trace of the operator $(\lambda_1 I + L_K)^{-1} L_K$, that is

$$\mathcal{N}(\lambda_1) = \mathrm{Tr}((\lambda_1 I + L_K)^{-1} L_K), \ \lambda_1 > 0. \tag{12}$$

- **Capacity assumption:** there exists a constant $\mathcal{C}_0$ such that for any $\lambda_1 > 0$,

$$\mathcal{N}(\lambda_1) \leq \mathcal{C}_0 \lambda_1^{-\beta}, \ \text{for some } 0 < \beta \leq 1. \tag{13}$$

- **Boundedness of operator** V_D: we assume that there is a constant $c_V > 0$ independent of the data set such that

$$\left\| V_D^T V_D \right\| \leq c_V \qquad \text{almost surely}, \tag{14}$$

in which V_D^T denotes the adjoint operator of V_D. $\left\| \cdot \right\|$ is the operator norm.

Notations and assumptions

- In the following, we denote the quantity $\mathcal{B}_{|D|,\lambda_1}$ as

$$\mathcal{B}_{|D|,\lambda_1} = \frac{2\kappa}{\sqrt{|D|}}\left(\frac{\kappa}{\sqrt{|D|\lambda_1}} + \sqrt{\mathcal{N}(\lambda_1)}\right). \tag{15}$$

$$\mathcal{B}'_{|D|,\lambda_1} = \frac{1}{|D|\sqrt{\lambda_1}} + \frac{\sqrt{\mathcal{N}(\lambda_1)}}{\sqrt{|D|}}. \tag{16}$$

- It can be observed that $\mathcal{B}_{|D|,\lambda_1}$ and $\mathcal{B}'_{|D|,\lambda_1}$ only differs in κ-scaling sense. $\mathcal{B}_{|D|,\lambda_1}$ and $\mathcal{B}'_{|D|,\lambda_1}$ will be used in error upper bound representation and operator estimates in subsequent analysis.

Outline

Estimates on Error bounds and Learning rates

Theorem 1

Suppose that the regularity condition (11) holds with $1/2 \leq r \leq 1$ and $|y| \leq M$ almost surely. Suppose the operator boundedness condition (14) holds. The regularization parameters $\lambda_1, \lambda_2 \in (0,1)$ and satisfy the relation $2c_V\lambda_2 = \lambda_1^{2r}$. The mapping $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is (α, L)-Holder continuous with $\alpha \in (0,1]$ and $L > 0$. If $\widetilde{d}$ satisfies $\frac{1}{\widetilde{d}^{\frac{\alpha}{2}}} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{d_i^{\alpha/2}}$. then we have

$$
\begin{aligned}
&E\left[\left\|f_{\hat{D},\lambda_1,\lambda_2} - f_\rho\right\|_\rho\right] \\
&\leq 2\left(2\sqrt{2}(2+\sqrt{\pi})^{\frac{1}{2}} L\frac{2^{\frac{\alpha+2}{2}} B_{\widetilde{K}}^{\frac{\alpha}{2}}}{\lambda_1 \widetilde{d}^{\frac{\alpha}{2}}} + 2\right)\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \frac{2^{\frac{\alpha}{2}} B_{\widetilde{K}}^{\frac{\alpha}{2}}}{\lambda_1^{\frac{1}{2}} \widetilde{d}^{\frac{\alpha}{2}}}\Bigg[(2+\sqrt{\pi})^{\frac{1}{2}} LM(2\Gamma(3) + \log^2 2) \\
&+2(2\Gamma(5) + \log^4 2)^{\frac{1}{2}}\left(2\sqrt{6}(2\Gamma(5) + \log^4 2)^{\frac{1}{2}}\frac{M}{\kappa}\frac{\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right) + \sqrt{3}(\lambda_1 + \lambda_2 c_V) \times\right. \\
&\left.\lambda_1^{r-\frac{3}{2}} 2^{r-\frac{1}{2}} \|g_\rho\|_\rho (2\Gamma(4r-1) + \log^{4r-2} 2)^{\frac{1}{2}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^{2r-1} + \sqrt{3}\kappa^{r-\frac{1}{2}}\|g_\rho\|_\rho\right)\Bigg] \\
&+4(2\Gamma(3) + \log^2 2)^{\frac{1}{2}}(2\Gamma(5) + \log^4 2)^{\frac{1}{2}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \frac{M}{\kappa}\mathcal{B}_{|D|,\lambda_1} \\
&+(2\Gamma(2r+1) + \log^{2r} 2) 2^r(\lambda_1^r + \lambda_1^{r-1}\lambda_2 c_V)\|g_\rho\|_\rho\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^{2r}.
\end{aligned}
$$

Estimates on Error bounds and Learning rates

Corollary 1

Suppose that the regularity condition (11) holds with $1/2 \leq r \leq 1$ and $|y| \leq M$ almost surely. The capacity condition (12) holds with index $\beta \in (0, 1]$. The mapping $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is (α, L)-Holder continuous with $\alpha \in (0, 1]$ and $L > 0$. If we choose the regularization parameters $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$, $\lambda_2 = \frac{1}{2c_V}|D|^{-\frac{2r}{2r+\beta}}$ and choose $d_1 = d_2 = \cdots = d_{|D|} = |D|^{\frac{1+2r}{\alpha(2r+\beta)}}$. Then we have

$$E\left[\left\|f_{\hat{D},\lambda_1,\lambda_2} - f_\rho\right\|_\rho\right] = \mathcal{O}\left(|D|^{-\frac{r}{2r+\beta}}\right). \tag{17}$$

Estimates on Error bounds and Learning rates

Above results handle **standard setting** $r \in [1/2, 1]$, which corresponds to $f_\rho \in \mathcal{H}_K$. Next two results are concerned with error bounds and learning rates in the case $f_\rho \notin \mathcal{H}_K$ that does not appear in literature on distribution regression. The following upper bound estimate for learning error holds without capacity assumption on effective dimension $\mathcal{N}(\lambda_1)$.

Estimates on Error bounds and Learning rates

Theorem 2

Suppose that the regularity condition (11) holds with $0 < r < 1/2$ and $|y| \leq M$ almost surely. Suppose the operator boundedness condition (14) holds. The regularization parameters $\lambda_1, \lambda_2 \in (0, 1)$ and satisfy the relation $2c_V\lambda_2 = \lambda_1$. The mapping $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is (α, L)-Holder continuous with $\alpha \in (0, 1]$ and $L > 0$. If $\widetilde{d}$ satisfies $\frac{1}{\widetilde{d}^{\frac{\alpha}{2}}} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{1}{d_i^{\alpha/2}}$. Then we have

Estimates on Error bounds and Learning rates

$$
\begin{aligned}
&E\left[\left\|f_{\hat{D},\lambda_1,\lambda_2}-f_\rho\right\|_\rho\right]\\
&\leq 2\left(2\sqrt{2}(2+\sqrt{\pi})^{\frac{1}{2}}L\frac{2^{\frac{\alpha+2}{2}}B_{\widetilde{K}}^{\frac{\alpha}{2}}}{\lambda_1\widetilde{d}^{\frac{\alpha}{2}}}+2\right)(2+\sqrt{\pi})^{\frac{1}{2}}LM\frac{2^{\frac{\alpha}{2}}B_{\widetilde{K}}^{\frac{\alpha}{2}}}{\lambda_1^{\frac{1}{2}}\widetilde{d}^{\frac{\alpha}{2}}}(2\Gamma(3)+\log^2 2)\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2\\
&+\left(2\sqrt{2}(2+\sqrt{\pi})^{\frac{1}{2}}L\frac{2^{\frac{\alpha+2}{2}}B_{\widetilde{K}}^{\frac{\alpha}{2}}}{\lambda_1\widetilde{d}^{\frac{\alpha}{2}}}+2\right)\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2(2\Gamma(5)+\log^4 2)^{\frac{1}{2}}\kappa L(2+\sqrt{\pi})^{\frac{1}{2}}2^{\frac{\alpha+2}{2}}B_{\widetilde{K}}^{\frac{\alpha}{2}}\times\\
&\left(2\sqrt{3}(2\Gamma(9)+\log^8 2)^{\frac{1}{4}}(2\Gamma(5)+\log^4 2)^{\frac{1}{4}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2\left[2M(\kappa+1)\frac{1}{\lambda_1^{\frac{1}{2}}\widetilde{d}^{\frac{\alpha}{2}}}\frac{1}{\sqrt{\lambda_1}}\mathcal{B}'_{|D|,\lambda_1}\right.\right.\\
&\left.+2(\kappa^2+\kappa)\|g_\rho\|_\rho\frac{\lambda_1^{r-1}}{\lambda_1^{\frac{1}{2}}\widetilde{d}^{\frac{\alpha}{2}}}\mathcal{B}'_{|D|,\lambda_1}\right]+2\sqrt{3}\|g_\rho\|_\rho\frac{\lambda_1^{r-\frac{1}{2}}}{\lambda_1^{\frac{1}{2}}\widetilde{d}^{\frac{\alpha}{2}}}\Bigg)\\
&+4(2\Gamma(5)+\log^4 2)^{\frac{1}{2}}(2\Gamma(3)+\log^2 2)^{\frac{1}{2}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2\left[M(\kappa+1)\mathcal{B}'_{|D|,\lambda_1}\right.\\
&\left.+\|g_\rho\|_\rho(\kappa^2+\kappa)\lambda_1^{r-\frac{1}{2}}\mathcal{B}'_{|D|,\lambda_1}\right]+2\|g_\rho\|_\rho c_V\lambda_2\lambda_1^{r-1}+\lambda_1^r\|g_\rho\|_\rho. \qquad (18)
\end{aligned}
$$

Estimates on Error bounds and Learning rates

When the capacity condition for $\mathcal{N}(\lambda_1)$ holds, for $r \in (0, 1/2)$, by assigning new λ_1, λ_2, $d_1, d_2, ..., d_{|D|}$, we derive following learning rates for the case $f_\rho \notin \mathcal{H}_K$.

Corollary

Suppose that the regularity condition (11) holds with $0 < r < 1/2$ and $|y| \leq M$ almost surely. The capacity condition (12) holds with index $\beta \in (0, 1]$. The mapping $K_{(.)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is (α, L)-Holder continuous with $\alpha \in (0, 1]$ and $L > 0$. If we choose the regularization parameters $\lambda_1 = |D|^{-\frac{1}{1+\beta}}$, $\lambda_2 = \frac{1}{2c_V}|D|^{-\frac{1}{1+\beta}}$ and choose $d_1 = d_2 = \cdots = d_{|D|} = |D|^{\frac{2}{\alpha(1+\beta)}}$. Then we have

$$E\left[\left\|f_{\hat{D},\lambda_1,\lambda_2} - f_\rho\right\|_\rho\right] = \mathcal{O}\left(|D|^{-\frac{r}{1+\beta}}\right). \tag{19}$$

Estimates on Error bounds and Learning rates

For the algorithm (8) of distributed learning with distribution regression, under a standard restriction on the machine number m, we derive the following optimal learning rate.

Theorem 3

Suppose that the regularity condition (11) holds with $1/2 \leq r \leq 1$ and $|y| \leq M$ almost surely. If the capacity condition (12) holds with index $\beta \in (0, 1]$. The mapping $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is (α, L)-Holder continuous with $\alpha \in (0, 1]$ and $L > 0$. Suppose that $\|V_{D_j}^T V_{D_j}\| \leq c_V$, $j = 1, 2, ..., m$ holds for a constant $c_V > 0$ independent of data sets. If $|D_j| = |D|/m$ for $j = 1, 2, ..., m$, $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$, $\lambda_2 = \frac{1}{2c_V}|D|^{-\frac{2r}{2r+\beta}}$ and the sample size for second stage are $d = |D|^{\frac{1+2r}{\alpha(2r+\beta)}}$ and m satisfies

$$m \leq |D|^{\frac{2r-1}{2r+\beta}},$$

then

$$E\left[\left\|\bar{f}_{\hat{D},\lambda_1,\lambda_2} - f_\rho\right\|_\rho\right] = \mathcal{O}(|D|^{-\frac{r}{2r+\beta}}). \tag{20}$$

Related work and discussion

To the best of our knowledge, the only existing works on learning theory analysis of learning rates of distribution regression scheme (6) are contained in

1. Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, Arthur Gretton (2016). Learning theory for distribution regression. The Journal of Machine Learning Research, 17(1), 5272-5311.
2. Zhiying Fang, Zheng-Chu Guo, Ding-Xuan Zhou (2020). Optimal learning rates for distribution regression. Journal of Complexity, 56, 101426.

Comparison:

- The optimal rates cover the case of [1] with $r = 1/2$, in contrast to the suboptimal rate of $E[\|f_{\hat{D},\lambda} - f_\rho\|^2] = \mathcal{O}(|D|^{-\frac{2}{5}})$. Also, the rates coincide with the optimal rate in [2] and this work improves the result in [2] to a more general setting by considering the additional penalty with parameter λ_2 and a bounded operator V_D in its multi-penalty regularization framework.

Related work and discussion

- Existing results in distribution regression literature consider the case when the regression function $f_\rho \in \mathcal{H}_K$. For f_ρ satisfying the regularity condition $r \in (0, 1/2)$, namely f_ρ **dose not belongs to** $\mathcal{H}_K$, the mini-max analysis on learning rate has not been carried out for distribution regression. In Theorem 2 and Corollary 2 of this work, we handle this setting and improve the analyzable regularity range to $(0, 1]$ from existing works [1] and [2].
- Based on a **divide-and-conquer** approach, we present a distributed learning algorithm with multi-penalty distribution regression scheme. The **optimal rate** is obtained for this method in Theorem 3. It can be observed that the distributed method handling **distribution or functional data** is still unexplored. By presenting the new distributed leaning algorithm with multi-penalty distribution regression and proving its optimal rates, this paper provides an effective distributed method for handling distribution data.

Outline

Starting point: error decomposition

In order to estimate expected learning rates of the algorithm, the subsequent estimates are based on the following basic **error decomposition**:

$$\|f_{\hat{D},\lambda_1,\lambda_2} - f_\rho\|_\rho \leq \|f_{\hat{D},\lambda_1,\lambda_2} - f_{D,\lambda_1,\lambda_2}\|_\rho + \|f_{D,\lambda_1,\lambda_2} - f_\rho\|_\rho. \quad (21)$$

In above decomposition, $f_{D,\lambda_1,\lambda_2}$ is the minimizer of the following classical multi-penalty regression scheme with the first stage sample $D = \{(\mu_{x_i}, y_i)\}_{i=1}^{|D|}$,

$$f_{D,\lambda_1,\lambda_2} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D|}\sum_{i=1}^{|D|} \big(f(\mu_{x_i}) - y_i\big)^2 + \lambda_1 \big\|f\big\|_K^2 + \lambda_2 \big\|V_D f\big\|_K^2 \right\}. \quad (22)$$

This quantity serves as an important **bridge** in subsequent analysis.

Representation for $f_{\hat{D},\lambda_1,\lambda_2}$ and $f_{D,\lambda_1,\lambda_2}$

We define the **sampling operator** $S_D : \mathcal{H}_K \to \mathbb{R}^{|D|}$ associated with the first stage sample as

$$S_D f = (f(\mu_{x_1}), f(\mu_{x_2}), ..., f(\mu_{x_{|D|}}))^T, \ f \in \mathcal{H}_K,$$

and the adjoint operator is given by

$$S_D^T \mathbf{c}_D = \frac{1}{|D|} \sum_{i=1}^{|D|} c_i K_{\mu_{x_i}}, \ \mathbf{c}_D = (c_1, c_2, ..., c_{|D|}) \in \mathbb{R}^{|D|}.$$

Then we can define $L_{K,D}$ as the first stage empirical operator of L_K as follow

$$L_{K,D}(f) = S_D^T S_D(f) = \frac{1}{|D|} \sum_{i=1}^{|D|} f(\mu_{x_i}) K_{\mu_{x_i}} = \frac{1}{|D|} \sum_{i=1}^{|D|} \langle K_{\mu_{x_i}}, f \rangle_K K_{\mu_{x_i}}, \ f \in \mathcal{H}$$

Representation for $f_{\hat{D},\lambda_1,\lambda_2}$ and $f_{D,\lambda_1,\lambda_2}$

We also define the **sample operator** $\hat{S}_D$ associated with the second stage sample as follows.

$$\hat{S}_D f = (f(\mu_{\hat{x}_1}), f(\mu_{\hat{x}_2}), ..., f(\mu_{\hat{x}_{|D|}}))^T, \ f \in \mathcal{H}_K,$$

Its adjoint operator $\hat{S}_D^T$ is given by

$$\hat{S}_D^T \mathbf{c}_D = \frac{1}{|D|} \sum_{i=1}^{|D|} c_i K_{\mu_{\hat{x}_i}}, \ \mathbf{c}_D = (c_1, c_2, ..., c_{|D|}) \in \mathbb{R}^{|D|}.$$

Then the empirical version operator of $L_{K,D}$ can be defined accordingly by using the second stage sample $\hat{D}$ as follow

$$L_{K,\hat{D}}(f) = \hat{S}_D^T \hat{S}_D(f) = \frac{1}{|D|} \sum_{i=1}^{|D|} f(\mu_{\hat{x}_i}) K_{\mu_{\hat{x}_i}} = \frac{1}{|D|} \sum_{i=1}^{|D|} \langle K_{\mu_{\hat{x}_i}}, f \rangle_K K_{\mu_{\hat{x}_i}}, \ f \in \mathcal{H}$$

(23)

Representation for $f_{\hat{D},\lambda_1,\lambda_2}$ and $f_{D,\lambda_1,\lambda_2}$

Basic representation:

$$f_{D,\lambda_1,\lambda_2} = (\lambda_1 I + L_{K,D} + \lambda_2 V_D^T V_D)^{-1} S_D^T y_D; \tag{24}$$

$$f_{\hat{D},\lambda_1,\lambda_2} = (\lambda_1 I + L_{K,\hat{D}} + \lambda_2 V_D^T V_D)^{-1} \hat{S}_D^T y_D \tag{25}$$

Basic fact from existing results

we use $E_{\mathbf{z}^{|D|}}[\cdot]$ to denote the expectation w.r.t. $\mathbf{z}^{|D|} = \{z_i = (\mu_{x_i}, y_i)\}_{i=1}^{|D|}$. Use $E_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}$ to denote the conditional expectation w.r.t. sample $\left\{\{x_{i,s}\}_{s=1}^{d_i}\right\}_{i=1}^{|D|}$ conditioned on $\{z_1, z_2, ..., z_{|D|}\}$. Namely

$$E_{\mathbf{z}^{|D|}}[\cdot] = E_{\{(\mu_{x_i}, y_i)\}_{i=1}^{|D|}}[\cdot], \quad E_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}[\cdot] = E_{\left\{\{x_{i,s}\}_{s=1}^{d_i}\right\}_{i=1}^{|D|}\left|\{z_i\}_{i=1}^{|D|}\right.}[\cdot].$$

From second stage to first stage:

- $E_{\{x_{i,s}\}_{s=1}^{d_i}|x_i}\left[\left\|\mu_{\hat{x}_i} - \mu_{x_i}\right\|_H^{2\alpha}\right] \leq (2+\sqrt{\pi})\frac{2^\alpha B_{\widetilde{K}}^\alpha}{d_i^\alpha}$.
- $\left\{E_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}\left[\left\|\hat{S}_D^T y_D - S_D y_D\right\|_K^2\right]\right\}^{\frac{1}{2}} \leq$

  $(2+\sqrt{\pi})^{\frac{1}{2}} ML \frac{1}{|D|}\sum_{i=1}^{|D|} \frac{2^{\frac{\alpha}{2}} B_{\widetilde{K}}^{\frac{\alpha}{2}}}{d_i^{\frac{\alpha}{2}}}$
- $\left\{E_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}\left[\left\|L_{K,\hat{D}} - L_{K,D}\right\|^2\right]\right\}^{\frac{1}{2}} \leq$

  $\kappa L(2+\sqrt{\pi})^{\frac{1}{2}} \frac{1}{|D|}\sum_{i=1}^{|D|} \frac{2^{\frac{\alpha+2}{2}} B_{\widetilde{K}}^{\frac{\alpha}{2}}}{d_i^{\frac{\alpha}{2}}}$

Basic fact from existing results

Basic lemme for handing multi-penalty regularization:

Lemme (Z. Guo et.al.)

If the regularization parameters $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2c_V\lambda_2 = \lambda_1^{\max\{2r,1\}}$ for $r \in (0, 1]$, then the following norm bound holds almost surely:

$$\left\|(\lambda_1 I + L_K)(\lambda_1 I + L_K + \lambda_2 V_D^T V_D)^{-1}\right\| \leq 2. \tag{26}$$

Basic fact from existing results

To derive the main results of the paper, a **second order decomposition** for invertible operators on Banach space is needed.

Lemma [S.Lin, X.Guo, D.X.Zhou]

For any invertible operators A and B,

$$\begin{aligned} A^{-1} - B^{-1} &= B^{-1}(B-A)A^{-1}(B-A)B^{-1} + B^{-1}(B-A)B^{-1} \\ &= B^{-1}(B-A)B^{-1}(B-A)A^{-1} + B^{-1}(B-A)B^{-1}. \end{aligned}$$

Main estimates for $r \in [1/2, 1]$

Proposition

Assume that $|y| \leq M$ and (14) hold almost surely. Let the regularity condition (11) holds with some index $1/2 \leq r \leq 1$. If $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2c_V\lambda_2 = \lambda_1^{2r}$. Then we have,

$$E_{\mathbf{z}^{|D|}}\left[\left\|f_{D,\lambda_1,\lambda_2} - f_\rho\right\|_\rho\right] \leq 4(2\Gamma(3) + \log^2 2)^{\frac{1}{2}}(2\Gamma(5) + \log^4 2)^{\frac{1}{2}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \frac{M}{\kappa}\mathcal{B}_{|D|,}$$

$$+(2\Gamma(2r+1) + \log^{2r} 2)2^r(\lambda_1^r + \lambda_1^{r-1}\lambda_2 c_V)\|g_\rho\|_\rho\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^{2r}$$

Main estimates for $r \in [1/2, 1]$

Things left is to estimate $\left\| f_{\hat{D},\lambda_1,\lambda_2} - f_{D,\lambda_1,\lambda_2} \right\|_\rho$ part in error decomposition of $\| f_{\hat{D},\lambda_1,\lambda_2} - f_\rho \|_\rho$. Denote

$$\Omega_{D,\lambda_1,\lambda_2,V_D} = \left\| L_K^{1/2} (L_{K,\hat{D}} + \lambda_1 I + \lambda_2 V_D^T V_D)^{-1} \right\|. \tag{27}$$

Proposition

There holds almost surely

$$\left\| f_{\hat{D},\lambda_1,\lambda_2} - f_{D,\lambda_1,\lambda_2} \right\|_\rho \leq \Omega_{D,\lambda_1,\lambda_2,V_D} \left(\left\| \hat{S}_D^T y_D - S_D^T y_D \right\|_K + \left\| L_{K,D} - L_{K,\hat{D}} \right\| \left\| f_{D,\lambda_1,\lambda_2} \right\|_K \right). \tag{28}$$

Main estimates for $r \in [1/2, 1]$

Only need to estimate $\Omega_{D,\lambda_1,\lambda_2,V_D}$ and $\left\|f_{D,\lambda_1,\lambda_2}\right\|_K$ in above inequality. We denote

$$\mathcal{A}_{D,\lambda_1,\lambda_2,V_D} = \left\|(\lambda_1 I + L_K + \lambda_2 V_D^T V_D)(\lambda_1 I + L_{K,D} + \lambda_2 V_D^T V_D)^{-1}\right\|.$$

Proposition

Let D be a sample drawn independently according to measure ρ and $\{x_{i,s}\}_{s=1}^{d_i}$ be a sample independently drawn according to distribution x_i, $i = 1, 2, ..., |D|$. Let the regularity condition (11) holds with some index $r \in (0, 1]$. If $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2c_V\lambda_2 = \lambda_1^{\max\{2r,1\}}$. Then we have

$$\left\{E_{\mathbf{x}^{\mathbf{d}},|D|\,|\mathbf{z}|^{D|}}\left[\Omega^2_{D,\lambda_1,\lambda_2,V_D}\right]\right\}^{\frac{1}{2}} \leq \left(\frac{2\sqrt{2}}{\lambda_1^{3/2}}(2+\sqrt{\pi})^{\frac{1}{2}} L \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{2^{\frac{\alpha+2}{2}} B_{\widetilde{K}}^{\frac{\alpha}{2}}}{d_i^{\frac{\alpha}{2}}}\right)\mathcal{A}_{D,\lambda_1,\lambda_2,V_D} + \frac{2}{\lambda_1^{\frac{1}{2}}}\mathcal{A}^{1/2}_{D,\lambda_1,\lambda_2,V_D}.$$

Main estimates for $r \in [1/2, 1]$

Proposition

Suppose $|y| \leq M$ almost surely. Let the regularity condition (11) holds with some index $1/2 \leq r \leq 1$. If $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2c_V\lambda_2 = \lambda_1^{2r}$, then we have

$$\left\{ E_{\mathbf{z}|D|} \left[\left\| f_{D,\lambda_1,\lambda_2} \right\|_K^2 \right] \right\}^{1/2} \leq 2\sqrt{6}(2\Gamma(5) + \log^4 2)^{1/2} \frac{M}{\kappa} \frac{\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} \left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1 \right)$$

$$+\sqrt{3}(\lambda_1 + \lambda_2 c_V)\lambda_1^{r-\frac{3}{2}} 2^{r-\frac{1}{2}} \left\| g_\rho \right\|_\rho (2\Gamma(4r-1) + \log^{4r-2} 2)^{1/2} \left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1 \right)^{2r-1}$$

$$+\sqrt{3}\kappa^{r-\frac{1}{2}} \left\| g_\rho \right\|_\rho$$

where $\mathcal{B}_{|D|,\lambda_1}$ is defined as in (15).

Main estimates for $r \in [1/2, 1]$

Learning rates: Learning rates can be obtained by following the selection of $\lambda_1 = |D|^{-\frac{1}{2r+\beta}}$, $\lambda_2 = \frac{1}{2c_V}|D|^{-\frac{2r}{2r+\beta}}$, $d_1 = d_2 = \cdots = d_{|D|} = |D|^{\frac{1+2r}{\alpha(2r+\beta)}}$, use the capacity assumption $\mathcal{N}(\lambda_1) \leq \mathcal{C}_0 \lambda_1^{-\beta}$, $\beta \in (0, 1]$,

Main estimates for $r \in (0, 1/2)$

we turn to prove the result in the **nonstandard setting** $r \in (0, 1/2)$ that corresponds to $f_\rho \notin \mathcal{H}_K$. In this section, we need the following norm for subsequent estimates

$$\Pi_{D,\lambda_1} = \left\|(\lambda_1 I + L_K)^{-\frac{1}{2}}(S_D^T y_D - L_K f_\rho)\right\|_K$$
$$\Xi_{D,\lambda_1} = \left\|(\lambda_1 I + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\right\|$$

Lemma

Let the sample set D be drawn independently according to probability measure ρ. If $|y| \leq M$ almost surely. Then we have,

$$E_{\mathbf{z}^{|D|}}\left[\Pi_{D,\lambda_1}^s\right] \leq (2\Gamma(s+1) + \log^s 2)\left(2M(\kappa+1)\mathcal{B}'_{|D|,\lambda_1}\right)^s, s \geq 1;$$
$$E_{\mathbf{z}^{|D|}}\left[\Xi_{D,\lambda_1}^s\right] \leq (2\Gamma(s+1) + \log^s 2)\left(2(\kappa^2+\kappa)\mathcal{B}'_{|D|,\lambda_1}\right)^s, s \geq 1.$$

in which $\mathcal{B}'_{|D|,\lambda_1}$ is defined as in (16).

Main estimates for $r \in (0, 1/2)$

to prove the result when f_ρ dose not lie in $\mathcal{H}_K$. We use the following bridge

$$f_{\lambda_1} = (L_K + \lambda_1 I)^{-1} L_K f_\rho, \tag{29}$$

which lies in $\mathcal{H}_K$. We derive the following estimate.

Proposition

Assume that $|y| \leq M$ and (14) hold almost surely. Suppose the parameters $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2c_V\lambda_2 = \lambda_1$. Then there holds almost surely

$$\max\left\{\left\|f_{D,\lambda_1,\lambda_2} - f_{\lambda_1}\right\|_\rho, \sqrt{\lambda_1}\left\|f_{D,\lambda_1,\lambda_2} - f_{\lambda_1}\right\|_K\right\} \leq 2\mathcal{A}_{D,\lambda_1,\lambda_2,v_D}\Pi_{D,\lambda_1}$$
$$+4\Xi_{D,\lambda_1}\mathcal{A}_{D,\lambda_1,\lambda_2,v_D}\left\|f_{\lambda_1}\right\|_K + \frac{2\lambda_2 c_V}{\sqrt{\lambda_1}}\left\|f_{\lambda_1}\right\|_K.$$

- Under the regularity condition $f_\rho = L_K^r(g_\rho)$, $r \in (0, 1/2)$ for some $g_\rho \in L^2_{\rho_{X_\mu}}$, the main difference with case $r \in [1/2, 1]$ is that the regression function f_ρ does not lie in $\mathcal{H}_K$ any more. To overcome this difficulty, above Proposition is derived to play a key bridge in following proofs.

Main estimates for $r \in (0, 1/2)$

Use decomposition

$$E_{\mathbf{z}|D|}\left[\left\|f_{D,\lambda_1,\lambda_2} - f_\rho\right\|_\rho\right] \le E_{\mathbf{z}|D|}\left[\left\|f_{D,\lambda_1,\lambda_2} - f_{\lambda_1}\right\|_\rho\right] + E_{\mathbf{z}|D|}\left[\left\|f_{\lambda_1} - f_\rho\right\|_\rho\right]. \qquad (30)$$

Proposition

Assume that $|y| \le M$ and (14) hold almost surely. Let the regularity condition (11) hold with some index $0 < r < 1/2$. If $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2c_V\lambda_2 = \lambda_1$. Then we have,

$$E_{\mathbf{z}|D|}\left[\left\|f_{D,\lambda_1,\lambda_2} - f_\rho\right\|_\rho\right] \le 4(2\Gamma(5) + \log^4 2)^{\frac{1}{2}}(2\Gamma(3) + \log^2 2)^{\frac{1}{2}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2$$
$$\left[M(\kappa+1)\mathcal{B}'_{|D|,\lambda_1} + \|g_\rho\|_\rho(\kappa^2+\kappa)\lambda_1^{r-\frac{1}{2}}\mathcal{B}'_{|D|,\lambda_1}\right] + 2\|g_\rho\|_\rho c_V\lambda_2$$

Proposition

Suppose $|y| \le M$ almost surely. Let the regularity condition (11) holds with some index $r \in (0, 1/2)$. If $\lambda_1, \lambda_2 \in (0, 1)$ satisfy $2\lambda_2 c_V = \lambda_1$, then we have

$$\left\{E_{\mathbf{z}|D|}\left[\left\|f_{D,\lambda_1,\lambda_2}\right\|_K^2\right]\right\}^{\frac{1}{2}} \le 2\sqrt{3}(2\Gamma(9) + \log^8 2)^{\frac{1}{4}}(2\Gamma(5) + \log^4 2)^{\frac{1}{4}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}} + 1\right)^2 \times$$
$$\times\left[2M(\kappa+1)\frac{1}{\sqrt{\lambda_1}}\mathcal{B}'_{|D|,\lambda_1} + 2(\kappa^2+\kappa)\left\|g_\rho\right\|_\rho\lambda_1^{r-1}\mathcal{B}'_{|D|,\lambda_1}\right] + 2\sqrt{3}\left\|g_\rho\right\|_\rho\lambda_1^{r-\frac{1}{2}}.$$

Main estimates for $r \in (0, 1/2)$

Learning rates: Learning rates result Corollary 2 is obtained after substituting the main parameters $\lambda_1 = |D|^{-\frac{1}{1+\beta}}$, $\lambda_2 = \frac{1}{2c_V}|D|^{-\frac{1}{1+\beta}}$, $\widetilde{d} = d_1 = d_2 = \cdots = d_{|D|} = |D|^{\frac{2}{\alpha(1+\beta)}}$ into the bound in Theorem 2.

Deriving learning rates for distributed learning

Denote $\Delta_D = \frac{1}{|D|}\sum_{z\in D}(y - f_{\lambda_1}(\mu_x))K_{\mu_x} - L_K(f_\rho - f_{\lambda_1})$, we will use the notation $\mathcal{Q}_{D_j}$, L_{K,D_j}, Δ_{D_j}, V_{D_j} involving the sample subset $\widetilde{D}_j$ and its associated mean embedding set D_j. Then the corresponding representations for $f_{\hat{D}_j,\lambda_1,\lambda_2} - f_{\lambda_1}$ are well defined for **local data sets**. The goal is the following decomposition:

$$
\begin{aligned}
\bar{f}_{\hat{D},\lambda_1,\lambda_2} - f_{\hat{D},\lambda_1,\lambda_2} &= \sum_{j=1}^{m}\frac{|D_j|}{|D|}\Big[[f_{\hat{D}_j,\lambda_1,\lambda_2} - f_{D_j,\lambda_1,\lambda_2}] - [f_{\hat{D},\lambda_1,\lambda_2} - f_{D,\lambda_1,\lambda_2}]\Big] \\
&\quad + \sum_{j=1}^{m}\frac{|D_j|}{|D|}\left[(L_{K,D_j} + \lambda_1 I + \lambda_2 V_{D_j}^T V_{D_j})^{-1} - (L_{K,D} + \lambda_1 I + \lambda_2 V_D^T V_D)^{-1}\right]\Delta_{D_j} \\
&\quad + \sum_{j=1}^{m}\frac{|D_j|}{|D|}\left[(L_{K,D_j} + \lambda_1 I + \lambda_2 V_{D_j}^T V_{D_j})^{-1}\lambda_2 V_{D_j}^T V_{D_j} f_{\lambda_1} - (L_{K,D} + \lambda_1 I\right. \\
&\quad \left. + \lambda_2 V_D^T V_D)^{-1}\lambda_2 V_D^T V_D f_{\lambda_1}\right] \\
&:= \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3. \qquad (32)
\end{aligned}
$$

Things left is to bound $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$.

Bound $\mathcal{S}_1$

$$E\left[\left\|\mathcal{S}_1\right\|_\rho\right] \lesssim \sum_{j=1}^{m} \frac{|D_j|}{|D|}\left(\frac{1}{\lambda_1 d^{\frac{\alpha}{2}}}+1\right)\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2 \frac{1}{\lambda_1^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\left[2+\right.$$
$$\left.\frac{\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)+(\lambda_1+\lambda_2 c_V)\lambda_1^{r-\frac{3}{2}}\left(\frac{2\mathcal{B}_{|D_j|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^{2r-1}\right]$$
$$+\left(\frac{1}{\lambda_1 d^{\frac{\alpha}{2}}}+1\right)\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^2 \frac{1}{\lambda_1^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\left[2+\frac{\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)\right.$$
$$\left.+(\lambda_1+\lambda_2 c_V)\lambda_1^{r-\frac{3}{2}}\left(\frac{2\mathcal{B}_{|D|,\lambda_1}}{\sqrt{\lambda_1}}+1\right)^{2r-1}\right] \lesssim |D|^{-\frac{r}{2r+\beta}}.$$

Bound $\mathcal{S}_2$, $\mathcal{S}_3$

Bound $\mathcal{S}_2$:

$$
\begin{aligned}
E_{\mathbf{z}|D|}\left[\left\|\mathcal{S}_2\right\|_\rho\right] &\lesssim \sqrt{\frac{\mathcal{N}(\lambda_1)}{\lambda_1|D|}}\lambda_1^r\left(\lambda_1^{\frac{1}{2}-r}+\frac{m}{\sqrt{|D|\lambda_1}}\right)\\
&\lesssim |D|^{-\frac{r}{2r+\beta}}\left(|D|^{-\frac{\frac{1}{2}-r}{2r+\beta}}|D|^{\frac{\frac{1}{2}-r}{2r+\beta}}+|D|^{\frac{2r-1}{2r+\beta}}|D|^{-\frac{1}{2}}|D|^{\frac{\frac{1}{2}}{2r+\beta}}|D|^{\frac{\frac{1}{2}-r}{2r+\beta}}\right)\\
&\lesssim |D|^{-\frac{r}{2r+\beta}}.
\end{aligned}
$$

Bound $\mathcal{S}_3$:

$$
E_{\mathbf{z}|D|}\left[\left\|\mathcal{S}_3\right\|_\rho\right] \lesssim |D|^{-\frac{r}{2r+\beta}}.
$$

Now combine above three estimates for $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, we arrive at

$$
E\left[\left\|\bar{f}_{\hat{D},\lambda_1,\lambda_2}-f_{\hat{D},\lambda_1,\lambda_2}\right\|_\rho\right]=\mathcal{O}(|D|^{-\frac{r}{2r+\beta}}).
$$

On the other hand, Theorem 1 has already shown

$$
E\left[\left\|f_{\hat{D},\lambda_1,\lambda_2}-f_\rho\right\|_\rho\right]=\mathcal{O}(|D|^{-\frac{r}{2r+\beta}}).
$$

Minkowski inequality implies the desired result.