Max-Discrepant Distributed Learning: Fast Risk Bound and Algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

1 Introduction

In the era of big data, the rapid expansion of computing capacities in automatic data generation and acquisition brings data of unprecedented size and complexity, and raises a series of scientific challenges such as storage bottleneck and algorithmic scalability [**?** 3**?**]. Distributed learning a feasible method to overcome the difficulty. The average mixture algorithm perhaps the simplest algorithm for distributed statistical inference. The algorithm is appealing in its simplicity: partition the dataset $\mathcal{S}$ of size N randomly into m equal sized subsets $\mathcal{S}_i$, and we compute the estimate for each of the $i = 1, \ldots, m$ subsets independently, and finally compute the average of partition-based estimate. Theoretical attempts have been recently made in [4, 3**?**] to derive learning rates for distributed learning.

This paper aims at error analysis of the distributed learning for (regularization) empirical risk minimization. Given $\mathcal{S} = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^N \in (\mathcal{Z} = \mathcal{X} \times \mathcal{Y})^N$, which is drawn identically and independently from a fixed, but unknown probability distribution $\mathbb{P}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, the (regularization) empirical risk minimization can be stated as

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}} \hat{R}(f) = \frac{1}{N} \sum_{j=1}^{N} \ell(f, z_j) + r(f) \tag{1}$$

where $\ell(f, z)$ is the loss function, and $r(f)$ is a regularizer. This learning algorithm has been well studied in learning theory, see e.g. [**? ? ? ? ?**]. The distributed learning algorithm studied in this paper starts with partitioning the data set $\mathcal{S}$ into m disjoint subsets $\{\mathcal{S}_i\}_{i=1}^m$, $|S_i| = \frac{N}{m} =: n$. Then it assigns each data subset $\mathcal{S}_i$ to one machine or processor to produce a local estimator $\hat{f}_i$:

$$\hat{f}_i = \operatorname*{arg\,min}_{f \in \mathcal{H}} \hat{R}_i(f) = \frac{1}{|\mathcal{S}_i|} \sum_{z_j \in \mathcal{S}_i} \ell(f, z_j) + r(f).$$

The finally global estimator $\bar{f}$ is synthesized by $\bar{f} = \frac{1}{m} \sum_{i=1}^m \hat{f}_i$. This algorithm has been studied with a matrix analysis approach in [4, 3**?**]. Under local strong convexity, smoothness and a reasonable set of other conditions, [4] show that the combined parameter achieves mean-squared error decays as

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}\left(\frac{1}{N} + \left(\frac{N}{m}\right)^2\right),$$

where $f^* = \arg\min_{f \in \mathcal{H}} R(f) = \mathbb{E}_{z \sim \mathbb{P}}[\ell(f, z)] + r(f)$. [3] consider the kernel ridge regression, under some eigenfunction assumption, they show that if m is not too large,

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|_2^2\right] = \mathcal{O}\left(\|f_*\|_{\mathcal{H}}^2 + \frac{\gamma(\lambda)}{N}\right),$$

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

where $\gamma(\lambda)=\sum_{j=1}^{\infty}\frac{\mu_j}{\lambda+\mu_j}$, μ_j is the eigenvalue of a Mercer kernel function. Without any eigenfunction assumption, [?] derive a novel bound for some $1\leq p\leq\infty$

$$\mathbb{E}\left[\left\|\bar{f}-f^*\right\|_2\right]=\mathcal{O}\left(\left(\frac{\gamma(\lambda)}{N}\right)^{\frac{1}{2}(1-\frac{1}{p})}\left(\frac{1}{N}\right)^{\frac{1}{2p}}\right).$$

There are two main contributions. First, under strongly convex and smooth, and a reasonable set of other conditions, we derive a risk bound of faster rate:

$$R(\bar{f})-R(f_*)=\mathcal{O}\left(\frac{R_*}{n}+\frac{1}{n^2}-\Delta(\bar{f})\right). \tag{2}$$

where $R(f)=\mathbb{E}_z\left[\ell(f,z)+r(f)\right]$, $\Delta(\bar{f})=\mathcal{O}\left(\frac{1}{m^2}\sum_{i,j=1,i\neq j}^{m}\|\hat{f}_i-\hat{f}_j\|^2\right)$ is the discrepant between all partition-based estimates. When the minimal risk is small, i.e., $R_*=\tilde{\mathcal{O}}\left(\frac{1}{n}\right)$, the rate is improved to

$$R(\bar{f})-R(f_*)=\mathcal{O}\left(\frac{1}{n^2}-\Delta(\bar{f})\right).$$

Thus, if $m\leq\sqrt{N}$, the order of $R(\bar{f})-R(f_*)$ is faster than $\mathcal{O}\left(\frac{1}{N}-\Delta(\bar{f})\right)$.

Note that if $\ell(f,z)+r(f)$ is L-Lipschitz continuous over f, the order of $R(\bar{f})-R(f^*)$ is

$$R(\bar{f})-R(f^*)=\mathcal{O}\left(L\mathbb{E}\left[\left\|\bar{f}-f^*\right\|_2\right]\right)=\mathcal{O}\left(L\sqrt{\mathbb{E}\left[\left\|\bar{f}-f^*\right\|_2\right]}\right).$$

Thus, the order of $R(\bar{f})-R(f^*)$ in [4, 3?] at most $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, which is much slower than that of our bound of $\mathcal{O}\left(\frac{1}{N}\right)$. Our second contribution is to develop a novel distributed learning algorithm. From Equation (2), we know that to guarantee good risk performance, the $\Delta(\bar{f})$ should be large. Therefore, we propose a novel max-discrepant distributed learning criterion:

$$\hat{f}_i=\arg\min_{f\in\mathcal{H}}\frac{1}{|\mathcal{S}_i|}\sum_{z_j\in\mathcal{S}_i}\ell(f,z_j)+r(f)-\gamma\|f-\bar{f}_{\setminus i}\|_{\mathcal{H}},$$

where $\bar{f}_{\setminus i}=\frac{1}{m-1}\sum_{j=1,j\neq i}^{m}\hat{f}_j$, the last term is to make $\Delta(\bar{f})$ large. We present a simple iterative algorithm to solve the above optimization problem. Experimental results on lots of datasets show that our proposed Max-Discrepant Distributed algorithm (MDD) is sound and efficient.

The rest of the paper is organized as follows. In Section 2, we derive risk bound of distributed learning with fast rates. In Section 3, we propose two novel algorithms based on the max-discrepant of the local estimate. In Section 4, we analyze the performance of our proposed criterion compared with other state-of-the-art model selection criteria. We end in Section 5 with conclusion.

2 Faster Rates of Distributed Learning

In this section, we will derive a sharper risk bound under some common assumptions.

2.1 Assumptions

In the following, we use $\|\cdot\|_{\mathcal{H}}$ to denote the norm induced by inner product of the Hilbert space $\mathcal{H}$.

Assumption 1. *The function $\nu(f,z)=\ell(f,z)+r(f)$ is η-strongly convex with respect to the first variable f, that is $\forall f,f'\in\mathcal{H}$, $z\in\mathcal{Z}$,*

$$\langle\nabla\nu(f,z),f-f'\rangle_{\mathcal{H}}+\frac{\eta}{2}\|f-f'\|_{\mathcal{H}}\leq\nu(f,z)-\nu(f',z) \tag{3}$$

or (another equivalent definition) $\forall f,f'\in\mathcal{H},z\in\mathcal{Z},t\in[0,1]$,

$$\nu(tf+(1-t)f')\leq t\nu(f,z)+(1-t)\nu(f',z)-\frac{1}{2}\eta t(t-1)\|f-f'\|_{\mathcal{H}}^2. \tag{4}$$

Assumption 2. *The function $\nu(f, z) = \ell(f, z) + r(f)$ is β-smooth with respect to the first variable f, that is $\forall f, f' \in \mathcal{H}$, $z \in \mathcal{Z}$,*

$$\|\nabla \nu(f, z) - \nabla \nu(f', z)\|_{\mathcal{H}} \leq \beta \|f - f'\|_{\mathcal{H}}. \tag{5}$$

Assumption 3. *The function $\nu(f, z) = \ell(f, z) + r(f)$ is L-Lipschitz continuous with respect to the first variable f, that is $\forall f, f' \in \mathcal{H}$,*

$$\|\nu(f, \cdot) - \nu(f', \cdot)\|_{\mathcal{H}} \leq L \|f - f'\|_{\mathcal{H}}. \tag{6}$$

Assumptions 1, 2 and 3 allow us to model some popular losses, such as square loss and logistic loss, and some regularizer, such as $r(f) = \lambda \|f\|_{\mathcal{H}}^2$.

Assumption 4. *Let $f_* = \arg\min_{f \in \mathcal{H}} R(f)$. We assume that the gradient at f_* is upper bounded by M, that is*

$$\|\nabla \ell(f^*, z)\|_{\mathcal{H}} \leq M, \forall z \in \mathcal{Z}.$$

Assumption 4 is also a common assumption, which is used in [2, 4].

2.2 Faster Rates of Distributed Learning

Let $\mathcal{N}(\mathcal{H}, \epsilon)$ be the ϵ-net of $\mathcal{H}$ with minimal cardinality, and $C(\mathcal{H}, \epsilon)$ the covering number of $|\mathcal{N}(\mathcal{H}, \epsilon)|$

Theorem 1. *For any $0 < \delta < 1$, $\epsilon \geq 0$, under **Assumptions** 1, 2, 3 and 4, and when*

$$m \leq \frac{N\eta}{4\beta \log C(\mathcal{H}, \epsilon)}, \tag{7}$$

with probability at least $1 - \delta$, we have

$$\begin{aligned} R(\bar{f}) - R(f_*) \leq & \frac{16\beta \log(4m/\delta)}{n^2 \eta} + \frac{128\beta R_* \log(4m/\delta)}{n\eta} + \frac{32\beta^2 \epsilon^2}{\eta} + \frac{64\beta L \log C(\mathcal{H}, \epsilon)\epsilon}{n\eta} \\ & \frac{64\beta \log^2 C(\mathcal{H}, \epsilon)\epsilon^2}{n^2 \eta} - \Delta(\bar{f}), \end{aligned} \tag{8}$$

where $R_* = R(f^*)$, $\Delta_{\bar{f}} = \frac{\eta}{4m^2} \sum_{i,j=1, i \neq j}^m \|\hat{f}_i - \hat{f}_j\|_{\mathcal{H}}^2$.

From the above theorem, an interesting finding is that, when the larger discrepant of each local estimate is, the tighter the risk bound is.

One can also see that when ϵ small enough, $\frac{32\beta^2\epsilon^2}{\eta} + \frac{64\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n\eta} + \frac{64\beta \log^2 C(\mathcal{H},\epsilon)\epsilon^2}{n^2\eta}$ will becomes non-dominating. To be specific, we have the following corollary:

Corollary 1. *By setting $\epsilon = \frac{1}{n}$ in Theorem 1, when $m \leq \frac{N\eta}{4\beta \log C(\mathcal{H}, 1/n)}$, with high probability, we have*

$$R(\bar{f}) - R(f_*) = \mathcal{O}\left(\frac{R_* \log(m)}{n} + \frac{\log(\mathcal{N}(\mathcal{H}, \frac{1}{n}))}{n^2} - \Delta(\bar{f})\right).$$

If the the minimal risk $R(f_*)$ is small, i.e., $R(f_*) = \mathcal{O}(\frac{1}{n})$, the rate can even reach

$$\mathcal{O}\left(\frac{\log(m)}{n^2} + \frac{\log(\mathcal{N}(\mathcal{H}, \frac{1}{n}))}{n^2} - \Delta(\bar{f})\right).$$

To the best of our knowledge, this is the first $\tilde{\mathcal{O}}\left(\frac{1}{n^2}\right)$-type of distributed risk bound of (regularization) empirical risk minimization.

In the next, we will consider two popular Hilbert spaces, linear space and reproducing kernel Hilbert space for deriving specific risk bounds.

2.2.1 Linear Space

The linear hypothesis space is defined as

$$\mathcal{H} = \left\{ f = \mathbf{w}^{\mathrm{T}}\mathbf{x} | \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq B \right\}.$$

According to the [1], the cover number of linear hypothesis space can be bounded by

$$\log\left(C(\mathcal{H}, \epsilon)\right) \leq d \log\left(6B/\epsilon\right).$$

Thus, if we set $\epsilon = \frac{1}{n}$, from Corollary 1, we have

$$R(\bar{f}) - R(f_*) = \mathcal{O}\left(\frac{R_* \log m}{n} + \frac{d \log n}{n^2} - \Delta(\bar{f})\right)$$

When the minimal risk is small, i.e., $R_* = \mathcal{O}\left(\frac{d}{n}\right)$, the rate is improved to

$$\mathcal{O}\left(\frac{d \log(mn)}{n^2} - \Delta(\bar{f})\right) = \mathcal{O}\left(\frac{d \log N}{n^2} - \Delta(\bar{f})\right).$$

Therefore, if $m \leq \sqrt{\frac{N}{d \log N}}$, the order of risk bound can even tighter than $\mathcal{O}\left(\frac{1}{N}\right)$.

2.2.2 Reproducing Kernel Hilbert Space

The reproducing kernel Hilbert space $\mathcal{H}_K$ associated with the kernel K is defined to be the closure of the linear span of the set of functions $\{K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{X}\}$ with the inner product satisfying

$$\langle K(\mathbf{x}, \cdot), f \rangle_{\mathcal{H}_K} = f(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}, f \in \mathcal{H}_K.$$

The bounded hypothesis space based on the reproducing kernel Hilbert space is defined as

$$\mathcal{H} := \{ f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq B \}.$$

From [5], if the kernel function K is the popular Gaussian kernel over $[0,1]^d$: $K(\mathbf{x}, \mathbf{x}') = \exp\left\{ -\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{\sigma^2}, \mathbf{x}, \mathbf{x}' \in [0,1]^d \right\}$, then for $0 \leq \epsilon \leq B/2$, there holds: $\log\left(C(\mathcal{H}, 1/n)\right) = \mathcal{O}\left(\log^d(nB)\right)$. From Corollary 1, if we set $\epsilon = \frac{1}{n}$, and assume $R_* = \mathcal{O}\left(\frac{1}{n}\right)$, we have

$$R(\bar{f}) - R(f_*) = \mathcal{O}\left(\frac{\log m}{n^2} + \frac{\log^d n}{n^2} - \Delta(\bar{f})\right)$$

Therefore, if $m \leq \min\left\{ \sqrt{\frac{N}{d \log N}}, \sqrt{\frac{N}{\log^d n}} \right\}$, the order can tighter than $\mathcal{O}\left(\frac{1}{N}\right)$.

2.3 Comparison with Related Work

Under the smooth, strongly convex and other some assumptions, [4] derive a distributed risk bound:

$$\mathbb{E}\left[\|\bar{f} - f_*\|^2\right] = \mathcal{O}\left(\frac{1}{N} + \frac{\log d}{n^2}\right). \tag{9}$$

[3] consider the kernel ridge regression, under some eigenfunction assumption, they show that if m is not too large,

$$\mathbb{E}\left[\left\|\bar{f} - f^*\right\|^2\right] = \mathcal{O}\left(\frac{r}{N}\right),$$

where r is the rank of the kernel function. Without any eigenfunction assumption, [**?**] derive a new bound of $\mathbb{E}\left[\left\|\bar{f} - f^*\right\|^2\right]$ of order at most $\mathcal{O}\left(\frac{1}{N}\right)$. If $\nu(f, z)$ is L-Lipschitz continuous over f, that is

$$\forall f, f \in \mathcal{H}, z \in \mathcal{Z}, |\nu(f, z) - \nu(f', z)| \leq L\|f - f'\|,$$

it is easy to verity that

$$R(f) - R(f_*) \leq L\mathbb{E}\left[\|\bar{f} - f_*\|\right] \leq L\sqrt{\mathbb{E}\left[\|\bar{f} - f_*\|^2\right]}$$

Thus, the order of [4, 3**?**] is at most $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

According to the subsections and , we know that if m is not very large, the order of this paper can faster than $\mathcal{O}\left(\frac{1}{N} - \Delta(\bar{f})\right)$, which is much sharper than the order of the related work [4, 3**?**].

Algorithm 1 Max-Discrepant Distributed Learning (MDD)

Input: $\lambda, \gamma, \mathbf{X}, m, \zeta > 0$.
For each branch node i: $\hat{\mathbf{w}}_i^0 = \mathbf{A}_i^{-1}\mathbf{b}_i$ and push $\hat{\mathbf{w}}_i^t$ to center node;
Center node: $\bar{\mathbf{w}}^0 = \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{w}}_0^t$ and push $\bar{\mathbf{w}}_{\setminus i}^0 = \frac{m\bar{\mathbf{w}}^t - \hat{\mathbf{w}}_i^0}{m-1}$ to each branch node i
for $t = 1, 2, \ldots$ **do**
    For each branch node i:
        $\mathbf{d}_i^t = \frac{\left(\bar{\mathbf{w}}_{\setminus i}^0\right)^{\mathrm{T}}\hat{\mathbf{w}}_i^0}{\mathbf{b}_i}$, $\hat{\mathbf{w}}_i^t = \hat{\mathbf{w}}_i^0 - \gamma \mathbf{d}_i^t$;
        push $\hat{\mathbf{w}}_i^t$ to center node;
    Center node:
        $\bar{\mathbf{w}}^t = \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{w}}_i^t$
        if $\|\bar{\mathbf{w}}^t - \bar{\mathbf{w}}^{t-1}\| \leq \zeta$ **end for**
    else
        push $\bar{\mathbf{w}}_{\setminus i}^t = \frac{m\bar{\mathbf{w}}^t - \hat{\mathbf{w}}_i^t}{m-1}$ to each branch node i
end for
Output: $\bar{\mathbf{w}} = \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{w}}_i^t$

3 Max-Discrepant Distributed Learning (MDD)

In this section, we will propose two novel algorithms based on the finding of the above section. From corollary 1, under some assumptions, we know that

$$R(f) - R(f_*) = \mathcal{O}\left(\frac{1}{n^2} - \frac{1}{m^2}\sum_{i,j=1, i\neq j}^m \|\hat{f}_i - \hat{f}_j\|_{\mathcal{H}}^2\right).$$

Thus, to obtain tight bound, the discrepancy of each local estimate $\hat{f}_i, i = 1, \ldots, m$ should be large. In the next, we will propose two algorithms for linear space and RKHS.

3.1 Linear Hypothesis Space

When $\mathcal{H}$ is a linear Hypothesis space, we consider the following optimization problem:

$$\hat{\mathbf{w}}_i = \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{z_i\in\mathcal{S}_i}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|_2^2 - \gamma\|\mathbf{w} - \bar{\mathbf{w}}_{\setminus i}\|_2^2, \tag{10}$$

where $\bar{\mathbf{w}}_{\setminus i} = \frac{1}{m-1}\sum_{j=1, j\neq i}\hat{\mathbf{w}}_j$ is used to make the discrepancy of each local estimate large. Note that, if given $\bar{\mathbf{w}}_{\setminus i}$, $\hat{\mathbf{w}}_i$ can be written as

$$\hat{\mathbf{w}}_i = \left(\frac{1}{n}\mathbf{X}_{\mathcal{S}_i}\mathbf{X}_{\mathcal{S}_i}^{\mathrm{T}} + \lambda\mathbf{I}_d - \gamma\mathbf{I}_d\right)^{-1}\left(\frac{1}{n}\mathbf{X}_{\mathcal{S}_i}\mathbf{y}_{\mathcal{S}_i} - \gamma\bar{\mathbf{w}}_{\setminus i}\right),$$

where $\mathbf{X}_{\mathcal{S}_i} = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \ldots, \mathbf{x}_{t_n})$, $\mathbf{y}_{\mathcal{S}_i} = (y_{t_1}, y_{t_2}, \ldots, y_{t_n})^{\mathrm{T}}$, $z_{t_j} \in \mathcal{S}_i$, $j = 1, \ldots, n$. In the next, we will give a iterative algorithm to solve the optimization problem 10, but in each iterative, we should compute $\mathbf{A}_i^{-1}\bar{\mathbf{w}}_{\setminus i}$, which is computationally intensive, where $\mathbf{A}_i = \frac{1}{n}\mathbf{X}_{\mathcal{S}_i}\mathbf{X}_{\mathcal{S}_i}^{\mathrm{T}} + \lambda\mathbf{I}_d - \gamma\mathbf{I}_d$.

Lemma 1. *If $\mathbf{A} \in \mathbb{R}^{l\times l}$ is a symmetric matrix and $\mathbf{c} = \mathbf{A}^{-1}\mathbf{b} \in \mathbb{R}^l$, then we have*

$$\mathbf{A}^{-1}\mathbf{d} = (\mathbf{d}^{\mathrm{T}}\mathbf{c})./\mathbf{c},$$

where $a./\mathbf{c} = (a/c_1, \ldots a/c_l)^{\mathrm{T}}$.

Proof. Since $\mathbf{A}$ a symmetric matrix, we have

$$\left(\mathbf{A}^{-1}\mathbf{d}\right)^{\mathrm{T}}\mathbf{b} = \mathbf{d}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b} = \mathbf{d}^{\mathrm{T}}\mathbf{c}.$$

Therefore, we can obtain that $\mathbf{A}^{-1}\mathbf{d} = (\mathbf{d}^{\mathrm{T}}\mathbf{c})./\mathbf{c}$. □

From Lemma 1, let $\mathbf{b}_i = \frac{1}{n}\mathbf{X}_{\mathcal{S}_i}\mathbf{y}_{\mathcal{S}_i}$, $\mathbf{c}_i = \mathbf{A}_i^{-1}\mathbf{b}_i$ we know that

$$\mathbf{A}_i^{-1}\bar{\mathbf{w}}_{\setminus i} = \left(\bar{\mathbf{w}}_{\setminus i}^{\mathrm{T}}\mathbf{c}_i\right)./\mathbf{c}_i,$$

which only need $\mathcal{O}(d)$.

Algorithm 2 Max-Discrepant Distributed Learning for RKHS (MDD-RKHS)

Input: $\lambda, \gamma, \mathbf{X}, m, \zeta > 0$.
For each branch node i: $\hat{\mathbf{c}}_i^0 = \mathbf{A}_i^{-1}\mathbf{b}_i$ and push $\hat{\mathbf{c}}_i^0$ to center node;
Center node: $\bar{\mathbf{c}}^0 = \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{c}}_0^t$ and push $\bar{\mathbf{c}}_{\setminus i}^0 = \frac{m\bar{\mathbf{c}}^t - \hat{\mathbf{c}}_i^0}{m-1}$ to each branch node i
for $t = 1, 2, \ldots$ **do**
  For each branch node i:
    $\mathbf{d}_i^t = \frac{\left(\bar{\mathbf{c}}_{\setminus i}^0\right)^{\mathrm{T}} \hat{\mathbf{c}}_i^0}{\mathbf{b}_i}$, $\hat{\mathbf{w}}_i^t = \hat{\mathbf{c}}_i^0 - \gamma \mathbf{d}_i^t$;
    push $\hat{\mathbf{c}}_i^t$ to center node;
  Center node:
    $\bar{\mathbf{c}}^t = \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{c}}_i^t$
    if $\|\bar{\mathbf{c}}^t - \bar{\mathbf{c}}^{t-1}\| \leq \zeta$ **end for**
  else
    push $\bar{\mathbf{c}}_{\setminus i}^t = \frac{m\bar{\mathbf{w}}^t - \hat{\mathbf{c}}_i^t}{m-1}$ to each branch node i
end for
Output: $\bar{\mathbf{c}} = \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{c}}_i^t$

3.2 Reproducing Kernel Hilbert Space

When $\mathcal{H}$ is a reproducing kernel Hilbert space, that is $f(\mathbf{x}) = \sum_{j=1}^n c_j K(\mathbf{x}_j, \mathbf{x})$, we consider the following optimization problem:

$$\hat{\mathbf{c}}_i = \underset{\mathbf{c}\in\mathbb{R}^n}{\arg\min} \frac{1}{n}\|\mathbf{K}_{\mathcal{S}_i}\mathbf{c} - \mathbf{y}_{\mathcal{S}_i}\|_2^2 + \lambda \mathbf{c}^{\mathrm{T}}\mathbf{K}_{\mathcal{S}_i}\mathbf{c} - \gamma\left(\mathbf{c} - \bar{\mathbf{c}}_{\setminus i}\right)^{\mathrm{T}} \mathbf{K}_{\mathcal{S}_i}\left(\mathbf{c} - \bar{\mathbf{c}}_{\setminus i}\right), \tag{11}$$

where $\bar{\mathbf{c}}_{\setminus i} = \frac{1}{m-1}\sum_{j=1, j\neq i} \hat{\mathbf{c}}_j$. If given $\bar{\mathbf{c}}_{\setminus i}$, $\hat{\mathbf{c}}_i$ can be written as

$$\hat{\mathbf{c}}_i = \left(\mathbf{K}_{\mathcal{S}_i} + \lambda\mathbf{I}_n - \gamma\mathbf{I}_n\right)^{-1}\left(\mathbf{y}_{\mathcal{S}_i} - \gamma\bar{\mathbf{c}}_{\setminus i}\right).$$

Let $\mathbf{A}_i = \mathbf{K}_{\mathcal{S}_i} + \lambda\mathbf{I}_n - \gamma\mathbf{I}_n$, $\mathbf{b}_i = \mathbf{y}_{\mathcal{S}_i}$.

3.3 Complexity

Linear space: for each node, we need $\min\mathcal{O}\left(\{nd^2, n^2d\}\right)$ to compute the $\mathbf{A}_i$, and $\mathcal{O}(d^3)$ to compute $\mathbf{A}_i^{-1}$, and need $\mathcal{O}(d)$ for each iterative, the communication complexity is $O(d)$ for each iterative. So, the total complexity is $\mathcal{O}\left(\{mnd^2, mn^2d\} + md^3 + Tmd\right)$, where T is the number of iterative.

RKHS: we need $\min\mathcal{O}\left(n^2d\right)$ to compute the $\mathbf{A}_i$, and $\mathcal{O}(n^3)$ to compute $\mathbf{A}_i^{-1}$, and need $\mathcal{O}(n)$ for each iterative, the communication complexity is $O(n)$ for each iterative. So, the total complexity is $\mathcal{O}\left(mn^2d + mn^3 + Tmn\right)$.

For the average mixture algorithm, for linear space, the complexity is $\mathcal{O}\left(\{mnd^2, mn^2d\} + md^3\right)$, for RKHS, the complexity is $\mathcal{O}\left(mn^2d + mn^3\right)$.

4 Analysis

4.1 The Key Idea

Since $\nu(f, z)$ is a η-strongly convex function, so both the risk $R(f) = \mathbb{E}_{z\sim\mathbb{P}}\nu(f, z)$ and empirical risk $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^n \nu(f, z_i)$ are η-strongly convex functions. By (4), we can obtain that

$$R(\bar{f}) = R\left(\frac{1}{m}\sum_{i=1}^m \hat{f}_i\right) \leq \frac{1}{m}\sum_{i=1}^m R(\hat{f}_i) - \frac{\eta}{4m^2}\sum_{i,j=1, i\neq j}^m \|\hat{f}_i - \hat{f}_j\|_{\mathcal{H}}^2.$$

Therefore, we have

$$R(\bar{f}) - R(f_*) \leq \frac{1}{m}\sum_{i=1}^m \left[R(\hat{f}_i) - R(f_*)\right] - \frac{\eta}{4m^2}\sum_{i,j=1, i\neq j}^m \|\hat{f}_i - \hat{f}_j\|_{\mathcal{H}}^2. \tag{12}$$

In the next, we will estimate $R(\hat{f}_i) - R(f_*)$, which is built upon the following inequality from (3):

$$\begin{aligned} & R(\hat{f}_i) - R(f_*) + \frac{\eta}{2}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2 \leq \langle \nabla R(\hat{f}_i), \hat{f}_i - f_* \rangle_{\mathcal{H}} \\ = & \langle \nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla \hat{R}_i(\hat{f}_i) - \nabla \hat{R}_i(f_*)], \hat{f}_i - f_* \rangle_{\mathcal{H}} \\ & + \langle \nabla R(f_*) - \nabla \hat{R}_i(f_*), \hat{f}_i - f_* \rangle_{\mathcal{H}} + \langle \nabla \hat{R}_i(\hat{f}_i), \hat{f}_i - f_* \rangle_{\mathcal{H}}. \end{aligned} \tag{13}$$

By the convexity of $\hat{R}_i(\cdot)$ and the optimality condition of $\hat{f}_i$ [**?**], we have

$$\langle \nabla \hat{R}_i(\hat{f}_i), f - \hat{f}_i \rangle_{\mathcal{H}} \geq 0, \forall f \in \mathcal{H}. \tag{14}$$

Substituting (14) into (13), we have

$$\begin{aligned} & R(\hat{f}_i) - R(f_*) + \frac{\eta}{2}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2 \\ \leq & \langle \nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla \hat{R}_i(\hat{f}_i) - \nabla \hat{R}_i(f_*)], \hat{f}_i - f_* \rangle_{\mathcal{H}} + \langle \nabla R(f_*) - \nabla \hat{R}_i(f_*), \hat{f}_i - f_* \rangle_{\mathcal{H}} \\ \leq & \left(\underbrace{\left\| \nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla \hat{R}_i(\hat{f}_i) - \nabla \hat{R}_i(f_*)] \right\|}_{:=A_1} + \underbrace{\left\| \nabla R(f_*) - \nabla \hat{R}_i(f_*) \right\|}_{=:A_2} \right) \left\| \hat{f}_i - f_* \right\| \end{aligned} \tag{15}$$

Lemma 2 (Seen in Appendix)**.** *Under **Assumptions** 2, with probability at least $1-\delta$, for any $f \in \mathcal{N}(\mathcal{H}, \epsilon)$, we have*

$$\begin{aligned} & \left\| \nabla R(f) - \nabla R(f_*) - [\nabla \hat{R}_i(f) - \nabla \hat{R}_i(f_*)] \right\| \\ \leq & \frac{\beta \log C(\mathcal{H}, \epsilon) \|f - f_*\|}{n} + \sqrt{\frac{\beta \log C(\mathcal{H}, \epsilon)(R(f) - R(f_*))}{n}}. \end{aligned} \tag{16}$$

Lemma 3 (Seen in Appendix)**.** *Under **Assumptions** 2 and 4, with probability at least $1-\delta$, we have*

$$\left\| \nabla R(f_*) - \nabla \hat{R}_i(f_*) \right\| \leq \frac{2M \log(2/\delta)}{n} + \sqrt{\frac{8\beta R_* \log(2/\delta)}{n}}. \tag{17}$$

Proof of Theorem 1. From the property of ϵ-net, we know that there exists a point $\tilde{f} \in \mathcal{N}(\mathcal{H}, \epsilon)$ such that

$$\|\hat{f}_i - \tilde{f}\| \leq \epsilon.$$

According to **Assumption** 2, we have

$$\begin{aligned} & \left\| \nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla \hat{R}_i(\hat{f}_i) - \nabla \hat{R}_i(f_*)] \right\| \\ \leq & \left\| \nabla R(\tilde{f}) - \nabla R(f_*) - [\nabla \hat{R}_i(\tilde{f}) - \nabla \hat{R}_i(f_*)] \right\| + 2\beta\epsilon \\ \overset{(16)}{\leq} & \frac{\beta \log C(\mathcal{H}, \epsilon) \|\tilde{f} - f_*\|}{n} + \sqrt{\frac{\beta \log C(\mathcal{H}, \epsilon)(R(\tilde{f}) - R(f_*))}{n}} + 2\beta\epsilon \\ \leq & \frac{\beta \log C(\mathcal{H}, \epsilon) \|\hat{f}_i - f_*\|_{\mathcal{H}}}{n} + \frac{\beta \log C(\mathcal{H}, \epsilon)\epsilon}{n} + 2\beta\epsilon \\ & + \sqrt{\frac{\beta \log C(\mathcal{H}, \epsilon)(R(\hat{f}_i) - R(f_*))}{n}} + \sqrt{\frac{\beta \log C(\mathcal{H}, \epsilon) \left(\left| R(\hat{f}_i) - R(\tilde{f}) \right| \right)}{n}} \\ \overset{(6)}{\leq} & \frac{\beta \log C(\mathcal{H}, \epsilon) \|\hat{f}_i - f_*\|_{\mathcal{H}}}{n} + \frac{\beta \log C(\mathcal{H}, \epsilon)\epsilon}{n} + 2\beta\epsilon \\ & + \sqrt{\frac{\beta \log C(\mathcal{H}, \epsilon)(R(\hat{f}_i) - R(f_*))}{n}} + \sqrt{\frac{\beta L \log C(\mathcal{H}, \epsilon)\epsilon}{n}} \end{aligned} \tag{18}$$

Substituting (18) and (17) into (15), with probability at least $1-2\delta$, we have

$$\begin{aligned}
&R(\hat{f}_i) - R(f_*) + \frac{\eta}{2}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2 \\
&\leq \frac{\beta \log C(\mathcal{H},\epsilon)\|\hat{f}_i - f_*\|_{\mathcal{H}}^2}{n} + \frac{\beta \log C(\mathcal{H},\epsilon)\epsilon\|\hat{f}_i - f_*\|_{\mathcal{H}}}{n} + 2\beta\epsilon\|\hat{f}_i - f_*\|_{\mathcal{H}} \\
&\quad + \|\hat{f}_i - f_*\|_{\mathcal{H}}\sqrt{\frac{\beta \log C(\mathcal{H},\epsilon)(R(\hat{f}_i) - R(f_*))}{n}} + \|\hat{f}_i - f_*\|_{\mathcal{H}}\sqrt{\frac{\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n}} \\
&\quad + \frac{2M\log(2/\delta)\|\hat{f}_i - f_*\|_{\mathcal{H}}}{n} + \|\hat{f}_i - f_*\|_{\mathcal{H}}\sqrt{\frac{8\beta R_*\log(2/\delta)}{n}}.
\end{aligned} \tag{19}$$

Note that

$$\sqrt{ab} \leq \frac{a}{2c} + \frac{bc}{2}, \forall a,b,c \geq 0.$$

Therefore, we can obtain that

$$\begin{aligned}
\|\hat{f}_i - f_*\|_{\mathcal{H}}\sqrt{\frac{\beta \log C(\mathcal{H},\epsilon)(R(\hat{f}_i) - R(f_*))}{n}} &\leq \frac{2\beta \log C(\mathcal{H},\epsilon)(R(\hat{f}_i) - R(f_*))}{n\eta} + \frac{\eta}{8}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2, \\
\frac{2M\log(2/\delta)\|\hat{f}_i - f_*\|_{\mathcal{H}}}{n} &\leq \frac{8M\log(2/\delta)}{n^2\eta} + \frac{\eta}{16}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2, \\
\|\hat{f}_i - f_*\|_{\mathcal{H}}\sqrt{\frac{8\beta R_*\log(2/\delta)}{n}} &\leq \frac{64\beta R_*\log(2/\delta)}{n\eta} + \frac{\eta}{32}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2, \\
2\beta\epsilon\|\hat{f}_i - f_*\|_{\mathcal{H}} &\leq \frac{32\beta^2\epsilon^2}{\eta} + \frac{\eta}{64}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2, \\
\|\hat{f}_i - f_*\|_{\mathcal{H}}\sqrt{\frac{\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n}} &\leq \frac{32\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n\eta} + \frac{\eta}{128}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2 \\
\frac{\beta \log C(\mathcal{H},\epsilon)\epsilon\|\hat{f}_i - f_*\|_{\mathcal{H}}}{n} &\leq \frac{32\beta \log^2 C(\mathcal{H},\epsilon)\epsilon^2}{n^2\eta} + \frac{\eta}{128}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2.
\end{aligned}$$

Substituting the above inequation into (19), we can obtain that

$$\begin{aligned}
&R(\hat{f}_i) - R(f_*) + \frac{\eta}{4}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2 \\
&\leq \frac{\beta \log C(\mathcal{H},\epsilon)\|\hat{f}_i - f_*\|_{\mathcal{H}}^2}{n} + \frac{2\beta \log C(\mathcal{H},\epsilon)(R(\hat{f}_i) - R(f_*))}{n\eta} + \frac{8M\log(2/\delta)}{n^2\eta} \\
&\quad + \frac{64\beta R_*\log(2/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} + \frac{32\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n\eta} + \frac{32\beta \log^2 C(\mathcal{H},\epsilon)\epsilon^2}{n^2\eta} \\
&\overset{(7)}{\leq} \frac{\eta}{4}\|\hat{f}_i - f_*\|_{\mathcal{H}}^2 + \frac{1}{2}(R(\hat{f}_i) - R(f_*)) + \frac{8\beta\log(2/\delta)}{n^2\eta} \\
&\quad + \frac{64\beta R_*\log(2/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} + \frac{32\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n\eta} + \frac{32\beta \log^2 C(\mathcal{H},\epsilon)\epsilon^2}{n^2\eta}.
\end{aligned}$$

Thus, with $1-2\delta$, we have

$$\begin{aligned}
R(\hat{f}_i) - R(f_*) \leq &\frac{16M\log(2/\delta)}{n^2\eta} + \frac{128\beta R_*\log(2/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} \\
&+ \frac{64\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n\eta} + \frac{64\beta \log^2 C(\mathcal{H},\epsilon)\epsilon^2}{n^2\eta}.
\end{aligned} \tag{20}$$

Combining (12) and (20), with $1-\delta$, we have

$$\begin{aligned}
R(\bar{f}) - R(f_*) \leq &\frac{16M\log(4m/\delta)}{n^2\eta} + \frac{128\beta R_*\log(4m/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} + \frac{64\beta L \log C(\mathcal{H},\epsilon)\epsilon}{n\eta} + \\
&+ \frac{64\beta \log^2 C(\mathcal{H},\epsilon)\epsilon^2}{n^2\eta} - \frac{\eta}{4m^2}\sum_{i,j=1,i\neq j}^{m}\|\hat{f}_i - \hat{f}_j\|_{\mathcal{H}}^2.
\end{aligned}$$

$\square$

Appendix: Proof of Lemma 2

Lemma 4 ([**?**]). *Let $\mathcal{H}$ be a Hilbert space and let ξ be a random variable with values in $\mathcal{H}$. Assume $\|\xi\| \leq M \leq \infty$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}[\|\xi\|^2]$. Let $\{\xi_i\}_{i=1}^n$ be m independent drawers of ξ. For any $0 \leq \delta \leq 1$, with confidence $1-\delta$,*

$$\left\| \frac{1}{n} \sum_{j=1}^{n} [\xi_j - \mathbb{E}[\xi_j]] \right\| \leq \frac{2M \log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2(\xi) \log(2/\delta)}{n}}.$$

Proof. Note that $\nu(f, \cdot)$ is β-smooth, so we have

$$\|\nabla \nu(f, \cdot) - \nabla \nu(f_*, \cdot)\|_{\mathcal{H}} \leq \beta \|f - f_*\|_{\mathcal{H}} \tag{21}$$

Because $\nu(f, \cdot)$ is β-smooth and convex, by (2.1.7) of [**?**], $\forall z \in \mathcal{Z}$, we have

$$\|\nabla \nu(f, z) - \nabla \nu(f_*, z)\|^2 \leq \beta \left(\nu(f, z) - \nu(f_*, z) - \langle \nabla \nu(f_*, z), f - f_* \rangle_{\mathcal{H}} \right).$$

Taking expectation over both sides, we have

$$\begin{aligned} &\mathbb{E}_{z \sim \mathbb{P}}[\|\nabla \nu(f, \cdot) - \nabla \nu(f_*, \cdot)\|^2] \\ &\leq \beta \left(R(\hat{f}_i) - R(f_*) - \langle \nabla R(f_*), f - f_* \rangle_{\mathcal{H}} \right) \\ &\leq \beta \left(R(\hat{f}_i) - R(f_*) \right) \end{aligned}$$

where the last inequality follows from the optimality condition of f_*, i.e.,

$$\langle \nabla R(f_*), f - f_* \rangle_{\mathcal{H}} \geq 0, \forall f \in \mathcal{H}.$$

Following Lemma 4, with probability at least $1 - \delta$, we have

$$\begin{aligned} &\left\| \nabla R(f) - \nabla R(f_*) - [\nabla \hat{R}_i(f) - \nabla \hat{R}_i(f_*)] \right\|_{\mathcal{H}} \\ = &\left\| \nabla R(f) - \nabla R(f_*) - \frac{1}{n} \sum_{z_i \in \mathcal{S}_i} [\nabla \nu(f, z_i) - \nabla \nu(f_*, z_i)] \right\|_{\mathcal{H}} \\ \leq &\frac{2\beta \|f - f_*\|_{\mathcal{H}} \log(2/\delta)}{n} + \sqrt{\frac{2\beta (R(f) - R(f_*)) \log(2/\delta)}{n}} \end{aligned}$$

We obtain Lemma 2 by taking the union bound over all $f \in \mathcal{N}(\mathcal{H}, \epsilon)$. □

4.2 Appendix: Proof of Lemma 3

Proof. Since $\nu(f, z_i)$ is β-smooth and nonegative, from Lemma 4 of [**?**], we have

$$\|\nabla \nu(f_*, z_i)\|^2 \leq 4\beta \nu(f_*, z_i)$$

and thus

$$\mathbb{E}_{z \sim \mathbb{P}} \left[\|\nabla \nu(f_*, z)\|^2 \right] \leq 4\beta \mathbb{E}_{z \sim \mathbb{P}}[\nu(f_*, z)] = 4\beta R(f_*).$$

From the **Assumption**, we have $\nabla \|\nu(f_*, z)\| \leq M$, $\forall z \in \mathcal{Z}$. Then, according to Lemma 4, with probability at least $1 - \delta$, we have

$$\begin{aligned} &\left\| \nabla R(f_*) - \nabla \hat{R}_i(f_*) \right\| = \left\| \nabla R(f_*) - \frac{1}{n} \sum_{z_j \in \mathcal{S}_i} \nabla \nu(f_*, z_j) \right\| \\ &\leq \frac{2\beta \log(2/\delta)}{n} + \sqrt{\frac{8\beta R_* \log(2/\delta)}{n}}. \end{aligned}$$

□

References

[1] G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

[2] L. Zhang, T. Yang, and R. Jin. Empirical risk minimization for stochastic convex optimization: $O(1/n)$-and $O(1/n^2)$-type of risk bounds. In *Proceedings of the Conference on Learning Theory (COLT 2017)*, pages 1954–1979, 2017.

[3] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. In *Proceedings of Conference on Learning Theory (COLT 2013)*, pages 592–617, 2013.

[4] Y. Zhang, M. J. Wainwright, and J. C. Duchi. Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*, pages 1502–1510, 2012.

[5] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.