Distributed Learning: Risk Bound and Algorithm

Anonymous Author(s)
Affiliation
Address
email

Abstract

1 Preliminaries

We consider the supervised learning where a learning algorithm receives a sample of N labeled points

$$\mathcal{S} = \{z_i = (\mathbf{x}_i, y_i)\}_{i=1}^N \in (\mathcal{Z} = \mathcal{X} \times \mathcal{Y})^N,$$

where $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ denotes the output space. We assume $\mathcal{S}$ is drawn identically and independently from a fixed, but unknown probability distribution $\mathbb{P}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The goal is to learn a good prediction model $f \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$, whose prediction accuracy at instance $z = (\mathbf{x}, y)$ is measured by a loss function $\ell(f, z)$.

In this paper, we focus on the supervised learning over the some Hilbert space $\mathcal{H}$:

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

where $\ell(f, z)$ is the loss function, and $r(f)$ is a regularizer.

The expect of $\hat{R}$ is defined as

$$R(f) = \mathbb{E}_{z \sim \mathbb{P}}[\ell(f, z)] + r(f). \tag{2}$$

Let $f^* = \arg\min_{f \in \mathcal{H}} R(f)$, and $R_* = R(f^*)$.

In the distributed setting, we divide evenly amongst m processors or inference procedures. Let $\mathcal{S}_i, i \in (1, 2, \ldots, m)$, denote a subsampled dataset of size $n = \frac{N}{m}$. For each $i = 1, 2, \ldots, m$, the local estimate

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

The average local estimates is denote as

$$\bar{f} = \frac{1}{m} \sum_{i=1}^m \hat{f}_i.$$

In the next, we will estimate the discrepancy of $R(\bar{f})$ and $R(f^*)$.

2 Faster Rates of Distributed Learning

2.1 Assumptions

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

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

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

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

$$\|\nabla \nu(f, z) - \nabla \nu(f', z)\|_{\mathcal{H}} \leq \beta \|f - f'\|_{\mathcal{H}}. \tag{4}$$

The above assumptions allow us to model many popular losses, such as square loss and logistic loss, and the regularizer, $r(f) = \lambda \|f\|_{\mathcal{H}}^2$.

Assumption 2. *Let $f_* = \arg\min_{f \in \mathcal{H}} R(f)$. We assume that the gradient at f_* is upper bounded by M, that is*

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

The above assumption is a common assumption can be seen in [4, 2].

3 Faster Rates of Distributed Learning

Theorem 1. *For any $0 < \delta < 1$, ϵ, the cover number of the Hilbert space $\mathcal{H}$ is defined as $\mathcal{N}(\mathcal{H}, \epsilon)$. Under **Assumptions** 1 and 2, and if*

$$m \leq \frac{N\eta}{4\beta \log \mathcal{N}(\mathcal{H}, \epsilon)}, \tag{5}$$

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

$$\begin{aligned} R(\bar{f}) - R(f_*) \leq & \frac{16\beta \log(4m/\delta)}{n^2\eta} + \frac{128\beta R_* \log(4m/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} + \frac{64\beta L \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon}{n\eta} \\ & \frac{64\beta \log^2\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon^2}{n^2\eta} - \Delta(\bar{f}), \end{aligned} \tag{6}$$

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

By choosing ϵ small enough,

$$\frac{32\beta^2\epsilon^2}{\eta} + \frac{64\beta L \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon}{n\eta} + \frac{64\beta \log^2\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon^2}{n^2\eta}$$

will becomes non-dominating. To be specific, we have the following corollary:

Corollary 1. *By setting $\epsilon = \frac{1}{n}$ in Theorem 1, with high probability, we have*

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

3.1 Linear Hypothesis Space

If we consider the linear hypothesis space, that is

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

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

$$\log\left(\mathcal{N}(\mathcal{H}, 1/n)\right) \leq d \log\left(6Bn\right).$$

Thus, 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 $m \leq \sqrt{\frac{N}{d \log N}}$, we have

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

3.2 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

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

In this subsection, we consider hypothesis space as the reproducing kernel Hilbert space,

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

From [5], if the Mercer kernel

$$K(\mathbf{x},\mathbf{x}')=k(\mathbf{x}-\mathbf{x}'),k(\mathbf{x})=\exp\left\{-\frac{\|\mathbf{x}\|^2}{\sigma^2},\mathbf{x},\mathbf{x}'\in[0,1]^d,\right\},$$

then for $0\leq\epsilon\leq B/2$, there holds:

$$\log\left(\mathcal{N}(\mathcal{H},1/n)\right)=\mathcal{O}\left(\log^d(nB)\right)$$

According to Corollary 1, we can obtain that

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

When the minimal risk R_* is small, $R_*=\mathcal{O}\left(\frac{\log^{(d-1)}n}{n}\right)$, if $m\leq n$, we have

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

Therefore, if $m\leq\sqrt{\frac{N}{\log^2 n}}$, we have

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

3.3 Comparison with Related Work

Under the smooth, strongly convex and other some common assumption, [4] shows that

$$\mathbb{E}\left[\|\bar{f}-f_*\|^2\right]=\mathcal{O}\left(\frac{1}{N}+\frac{\log d}{n^2}\right). \tag{7}$$

If $\nu(f,z)$ is L-Lipschitz continuous over f, that is

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

it is easy to verity that

$$\begin{aligned}R(f)-R(f_*)&\leq L\mathbb{E}\left[\|\bar{f}-f_*\|_{\mathcal{H}}\right]\leq L\sqrt{\mathbb{E}\left[\|\bar{f}-f_*\|_{\mathcal{H}}^2\right]}\\&=\mathcal{O}\left(\frac{1}{\sqrt{N}}+\frac{\sqrt{\log d}}{n}\right).\end{aligned}\tag{8}$$

According to the subsections and , we know that if m is not very large, the order of this paper can reach $\mathcal{O}\left(\frac{1}{N}-\Delta(\bar{f})\right)$, which is much sharper than the order of (8).

[3] consider the kernel ridge regression, under some assumptions over the feature map induced by the kernel function, and if m is not very large, they show that

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

If $\nu(f,z)$ is L-Lipschitz continuous over f, same as the above analysis, it is easy to verity that

$$R(f)-R(f_*)=\mathcal{O}\left(\frac{1}{\sqrt{N}}\right),$$

which is much looser than of propose bound.

4 Discrepant Distributed Algorithms (DDA)

According to the above results, under some assumptions, we know that

$$R(f) - R(f_*) = \mathcal{O}\left(\frac{1}{N} - \Delta(\bar{f})\right),$$

where $\Delta_{\bar{f}} = \frac{\eta}{4m^2} \sum_{i,j=1,i\neq j}^{m} \|\hat{f}_i - \hat{f}_j\|_{\mathcal{H}}^2$. Thus, to obtain tight bound, the discrepancy of each local estimate $\hat{f}_i, i = 1, \ldots, m$ should be large. Therefore, it is reasonable to derive the following optimization problem:

$$\hat{f}_i = \arg\min_{f\in\mathcal{H}} \frac{1}{n} \sum_{z_j\in\mathcal{S}_i} \ell(f, z_i) + r(f) - \gamma\|f - \bar{f}_{\setminus i}\|_{\mathcal{H}}, \tag{9}$$

where $\bar{f}_{\setminus i} = \frac{1}{m-1} \sum_{j=1,j\neq i}^{m} \hat{f}_j$.

4.1 Linear Hypothesis Space

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

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

where $\bar{\mathbf{w}}_{\setminus i} = \frac{1}{m-1} \sum_{j=1,j\neq i} \hat{\mathbf{w}}_j$. If given $\bar{\mathbf{w}}_{\setminus i} = \frac{1}{m-1} \sum_{j=1,j\neq i} \hat{\mathbf{w}}_j$, it is easy to verity that $\hat{\mathbf{w}}_i$ can be written as

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

where $\mathbf{X}_{\mathcal{S}_i} = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \ldots, \mathbf{x}_{t_n})$, $\mathbf{y}_{\mathcal{S}_i} = (y_{t_1}, y_{t_2}, \ldots, y_{t_n})^{\mathrm{T}}$, $z_{t_i} \in \mathcal{S}_i$, $i = 1, \ldots, n$.
Let $\mathbf{A}_i = \frac{1}{n}\mathbf{X}_{\mathcal{S}_i}\mathbf{X}_{\mathcal{S}_i}^{\mathrm{T}} + \lambda\mathbf{I}_d - \gamma\mathbf{I}_d$, $\mathbf{b}_i = \frac{1}{n}\mathbf{X}_{\mathcal{S}_i}^{\mathrm{T}}\mathbf{y}_{\mathcal{S}_i}$.
Let $\mathbf{d}_i = \mathbf{A}_i^{-1}\bar{\mathbf{w}}_{\setminus i}$, $\hat{\mathbf{w}}_i = \mathbf{A}_i^{-1}\mathbf{b}_i$, we have

$$\bar{\mathbf{w}}_{\setminus i}^{\mathrm{T}}\hat{\mathbf{w}}_i = \bar{\mathbf{w}}_{\setminus i}^{\mathrm{T}}\mathbf{A}_i^{-1}\mathbf{b}_i = \left(\mathbf{A}_i^{-1}\bar{\mathbf{w}}_{\setminus i}\right)^{\mathrm{T}}\mathbf{b}_i = \mathbf{d}_i^{\mathrm{T}}\mathbf{b}_i,$$

thus $\mathbf{d}_i = \frac{\bar{\mathbf{w}}_{\setminus i}^{\mathrm{T}}\hat{\mathbf{w}}_i}{\mathbf{b}_i}$

The DDA Algorithm is given as follows:

Input : $\lambda, \gamma, \mathbf{X}, m, \zeta > 0$.
  For $t = 0, 1, \ldots, T$
    Each branch node i:
      If $t = 0$
        $\hat{\mathbf{w}}_i^0 = \mathbf{A}_i^{-1}\mathbf{b}_i$;
    else
        $\mathbf{d}_i^t = \frac{\left(\bar{\mathbf{w}}_{\setminus i}^0\right)^{\mathrm{T}}\hat{\mathbf{w}}_i^0}{\mathbf{b}_i}$
        $\hat{\mathbf{w}}_i^t = \hat{\mathbf{w}}_i^0 - \gamma\mathbf{d}_i^t$;
        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**
        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
Output : $\bar{\mathbf{w}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\mathbf{w}}_i^T$

4.2 Reproducing Kernel Hilbert Space

When $\mathcal{H}$ is a reproducing kernel Hilbert space, we consider the following problem:

$$\hat{f}_i = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{z_i\in\mathcal{S}_i} (f(\mathbf{x}_i) - y_i)^2 + \lambda\|f\|^2_{\mathcal{H}} + \gamma\|f - \bar{f}_{\setminus i}\|^2_{\mathcal{H}}, \tag{10}$$

where $f(\mathbf{x}) = \sum_{j=1}^n c_j K(\mathbf{x}_j, \mathbf{x})$, which can be written as

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

If given $\bar{\mathbf{c}}_{\setminus i} = \frac{1}{m-1}\sum_{j=1, j\neq i}\hat{\mathbf{c}}_j$, it is easy to verity that $\hat{\mathbf{c}}_i$ 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 $\mathbf{A}_i = \mathbf{K}_{\mathcal{S}_i} + \lambda\mathbf{I}_n - \gamma\mathbf{I}_n$, $\mathbf{b}_i = \mathbf{y}_{\mathcal{S}_i}$.

Input : $\lambda, \gamma, \mathbf{X}, m, \zeta > 0$.
For $t = 0, 1, \ldots, T$
Each branch node i:
If $t = 0$
$\hat{\mathbf{c}}_i^0 = \mathbf{A}_i^{-1}\mathbf{b}_i$;
else
$\mathbf{d}_i^t = \frac{(\bar{\mathbf{c}}_{\setminus i}^0)^{\mathrm{T}}\hat{\mathbf{c}}_i^0}{\mathbf{b}_i}$
$\hat{\mathbf{c}}_i^t = \hat{\mathbf{c}}_i^0 - \gamma\mathbf{d}_i^t$;
push $\hat{\mathbf{w}}_i^t$ 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**
else push $\bar{\mathbf{c}}_{\setminus i}^t = \frac{m\bar{\mathbf{c}}^t - \hat{\mathbf{c}}_i^t}{m-1}$ to each branch node i
End
Output : $\bar{\mathbf{c}} = \frac{1}{m}\sum_{i=1}^m \hat{\mathbf{c}}_i^T$

5 Analysis

5.1 The Key Idea

Since $R(f)$ is η-strongly convex function, we have

$$R(\bar{f}) \leq \frac{1}{m}\sum_{i=1}^m R(\hat{f}_i) - \frac{\eta}{2m^2}\sum_{i,j=1}^m \|\hat{f}_i - \hat{f}_j\|^2.$$

Therefore, we have

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

In the next, we will estimate $R(\hat{f}_i) - R(f_*)$.

Our theoretical analysis is built upon the following inequality:

$$\begin{aligned}
&R(\hat{f}_i) - R(f_*) + \frac{\eta}{2}\|\hat{f}_i - f_*\|^2 \leq \langle\nabla R(\hat{f}_i), \hat{f}_i - f_*\rangle \\
&= \langle\nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla\hat{R}(\hat{f}_i) - \nabla\hat{R}(f_*)], \hat{f}_i - f_*\rangle + \langle\nabla\hat{R}(\hat{f}_i) - \nabla\hat{R}(f_*) + \nabla R(f_*), \hat{f}_i - f_*\rangle \\
&= \langle\nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla\hat{R}(\hat{f}_i) - \nabla\hat{R}(f_*)], \hat{f}_i - f_*\rangle + \langle\nabla R(f_*) - \nabla\hat{R}(f_*), \hat{f}_i - f_*\rangle \\
&\leq \left(\underbrace{\left\|\nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla\hat{R}(\hat{f}_i) - \nabla\hat{R}(f_*)]\right\|}_{:=A_1} + \underbrace{\left\|\nabla R(f_*) - \nabla\hat{R}(f_*)\right\|}_{:A_2}\right)\left\|\hat{f}_i - f_*\right\|
\end{aligned} \tag{13}$$

where η is the strong convexity modulus of $R(\cdot)$ if exists otherwise it is zero.

Lemma 1. *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 $1-\delta$,*

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

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

$$\begin{aligned} &\left\|\nabla R(f) - \nabla R(f_*) - [\nabla \hat{R}(f) - \nabla \hat{R}(f_*)]\right\| \\ &\leq \frac{\beta \log\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\|f - f_*\|}{n} + \sqrt{\frac{\beta\log\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\left(R(f) - R(f_*)\right)}{n}}. \end{aligned} \tag{14}$$

Proof. Let $\nu(f) = \ell(f, \cdot) + r(f)$. Note that $\nu(f)$ is β-smooth, so we have

$$\|\nabla\nu(f) - \nabla\nu(f_*)\| \leq \beta\|f - f_*\| \tag{15}$$

Because ν is β-smooth and convex, by (2.1.7) of?, we have

$$\|\nabla\nu(f) - \nabla\nu(f_*)\|^2 \leq \beta\left(\nu(f) - \nu(f_*) - \langle\nabla\nu(f_*), f - f_*\rangle\right).$$

Taking expectation over both sides, we have

$$\begin{aligned} &\mathbb{E}[\|\nabla\nu(f) - \nabla\nu(f_*)\|^2] \\ &\leq \beta\left(R(\hat{f}_i) - R(f_*) - \langle\nabla R(f_*), f - f_*\rangle\right) \\ &\leq \beta\left(R(\hat{f}_i) - R(f_*)\right) \end{aligned}$$

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

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

$\square$

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

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

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

Lemma 3. *with probability at least $1-\delta$, we have*

$$\left\|\nabla R(f_*) - \nabla\hat{R}(f_*)\right\| \leq \frac{2M\log(2/\delta)}{n} + \sqrt{\frac{8\beta R_*\log(2/\delta)}{n}}. \tag{16}$$

Proof. Let $\nu(f, z_i) = \ell(f, z_i) + r(f)$ Since $\nu(\cdot, z_i)$ is β-smooth and nonegative, from Lemma 4 of Srebro et al. (2010), 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 1, with probability at least $1-\delta$, we have

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

□

Theorem 2. *At least* $1-2\delta$*, if*

$$
n \geq \frac{4\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)}{\eta}, \tag{17}
$$

we have

$$
\begin{aligned}
R(\bar{f}) - R(f_*) \leq & \frac{16\beta \log(2m/\delta)}{n^2\eta} + \frac{128\beta R_* \log(2m/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} \\
& + \frac{64\beta L \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon}{n\eta} + \frac{64\beta \log^2\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon^2}{n^2\eta} \\
& - \frac{\eta}{4m^2}\sum_{i,j=1, i\neq j}^{m} \|\hat{f}_i - \hat{f}_j\|^2.
\end{aligned} \tag{18}
$$

Proof. 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 ??**, we have

$$
\begin{aligned}
& \left\|\nabla R(\hat{f}_i) - \nabla R(f_*) - [\nabla \hat{R}(\hat{f}_i) - \nabla \hat{R}(f_*)]\right\| \\
\leq & \left\|\nabla R(\tilde{f}) - \nabla R(f_*) - [\nabla \hat{R}(\tilde{f}) - \nabla \hat{R}(f_*)]\right\| + 2\beta\epsilon \\
\overset{(14)}{\leq} & \frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\|\tilde{f} - f_*\|}{n} + \sqrt{\frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\left(R(\tilde{f}) - R(f_*)\right)}{n}} + 2\beta\epsilon \\
\leq & \frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\|\hat{f}_i - f_*\|}{n} + \frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon}{n} + 2\beta\epsilon \\
& + \sqrt{\frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\left(R(\hat{f}_i) - R(f_*)\right)}{n}} + \sqrt{\frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\left(\left|R(\hat{f}_i) - R(\tilde{f})\right|\right)}{n}} \\
\overset{(??)}{\leq} & \frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\|\hat{f}_i - f_*\|}{n} + \frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon}{n} + 2\beta\epsilon \\
& + \sqrt{\frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\left(R(\hat{f}_i) - R(f_*)\right)}{n}} + \sqrt{\frac{\beta L \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon}{n}}
\end{aligned} \tag{19}
$$

Substituting (19) and (16) into (13), with probability at least $1-2\delta$, we have

$$
\begin{aligned}
& R(\hat{f}_i) - R(f_*) + \frac{\eta}{2}\|\hat{f}_i - f_*\|^2 \\
\leq & \frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\|\hat{f}_i - f_*\|^2}{n} + \frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon\|\hat{f}_i - f_*\|}{n} + 2\beta\epsilon\|\hat{f}_i - f_*\| \\
& + \|\hat{f}_i - f_*\|\sqrt{\frac{\beta \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\left(R(\hat{f}_i) - R(f_*)\right)}{n}} + \|\hat{f}_i - f_*\|\sqrt{\frac{\beta L \log\left(\mathcal{N}(\mathcal{H}, \epsilon)\right)\epsilon}{n}} \\
& + \frac{2\beta \log(2/\delta)\|\hat{f}_i - f_*\|}{n} + \|\hat{f}_i - f_*\|\sqrt{\frac{8\beta R_* \log(2/\delta)}{n}}.
\end{aligned} \tag{20}
$$

Note that

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

Therefore, we can obtain that

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

Substituting the above inequation into (20), we can obtain that

$$\begin{aligned}
& R(\hat{f}_i) - R(f_*) + \frac{\eta}{4}\|\hat{f}_i - f_*\|^2 \\
\leq & \frac{\beta \log\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\|\hat{f}_i - f_*\|^2}{n} + \frac{2\beta \log\left(\mathcal{N}(\mathcal{H},\epsilon)\right)(R(\hat{f}_i) - R(f_*))}{n\eta} + \frac{8\beta\log(2/\delta)}{n^2\eta} \\
& + \frac{64\beta R_*\log(2/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} + \frac{32\beta L \log\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\epsilon}{n\eta} + \frac{32\beta \log^2\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\epsilon^2}{n^2\eta} \\
\overset{(17)}{\leq} & \frac{\eta}{4}\|\hat{f}_i - f_*\|^2 + \frac{1}{2}(R(\hat{f}_i) - R(f_*)) + \frac{8\beta\log(2/\delta)}{n^2\eta} \\
& + \frac{64\beta R_*\log(2/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} + \frac{32\beta L \log\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\epsilon}{n\eta} + \frac{32\beta \log^2\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\epsilon^2}{n^2\eta}.
\end{aligned}$$

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

$$\begin{aligned}
R(\hat{f}_i) - R(f_*) \leq & \frac{16\beta\log(2/\delta)}{n^2\eta} + \frac{128\beta R_*\log(2/\delta)}{n\eta} + \frac{32\beta^2\epsilon^2}{\eta} \\
& + \frac{64\beta L \log\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\epsilon}{n\eta} + \frac{64\beta \log^2\left(\mathcal{N}(\mathcal{H},\epsilon)\right)\epsilon^2}{n^2\eta}.
\end{aligned} \tag{21}$$

Combining (12) and (21), with $1 - 2\delta$, we have

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

□

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.