Sharper Generalization Bounds for Clustering: Supplementary Material

Shaojie Li [1 2] **Yong Liu** [1 2]

Abstract

In this supplementary material, we provide the proofs of the theoretical results in the main paper.

A. Proof of Theorem 1

[Sketch of proof techniques.] We first prove that the expected excess clustering risk $\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^*$ can be bounded by $2\mathbb{E}\sup_{f_{W,Z}\in\mathcal{F}}\left|L(f_{W,Z}) - \hat{L}_n(f_{W,Z})\right|$. Based on U-process (Clémençon et al., 2008), the standard symmetrization technique (Bartlett & Mendelson, 2002) and Jensen's inequality (Mohri et al., 2018), this term can be bounded by $2R(\mathcal{F})$. Furthermore, $R(\mathcal{F})$ can be bounded by $K\max_k R(\mathcal{F}_k)$, that is

$$R(\mathcal{F}) \leq K\max_k \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{f_{W,Z_k}\in\mathcal{F}_k}\left|\frac{2}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i f_{W,Z_k}\left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor\frac{n}{2}\rfloor}\right)\right|\right],$$

where $\mathcal{F}_k$ is a function space of the output coordinate k of $\mathcal{F}$, and where $K\max_k R(\mathcal{F}_k)$ means the maximum Rademacher complexity of the restrictions of the function class along each coordinate with timing a factor of $\mathcal{O}(K)$. Finally, $K\max_k R(\mathcal{F}_k)$ can be bounded by $\frac{4KM}{\sqrt{n}}$. For $L(\hat{f}^*_{W,Z}) - L^*$, we can bound it by $2\sup_{f_{W,Z}\in\mathcal{F}}\left|L(f_{W,Z}) - \hat{L}_n(f_{W,Z})\right|$ in a similar method. Since $\left|L(f_{W,Z}) - \hat{L}_n(f_{W,Z})\right|$ is a bounded difference function, so the term $L(\hat{f}^*_{W,Z}) - L^*$ can be proved by McDiarmid inequality (Mohri et al., 2018).

Proof. **(1.)** We first prove that $\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq \frac{4KM}{\sqrt{\lfloor n/2\rfloor}}$.

$$\begin{aligned}
\mathbb{E}\left[L(\hat{f}^*_{W,Z})\right] - L^* &= \mathbb{E}\left[L(\hat{f}^*_{W,Z}) - \hat{L}_n(\hat{f}^*_{W,Z}) + \hat{L}_n(\hat{f}^*_{W,Z}) - L^*\right] \\
&= \mathbb{E}\left[L(\hat{f}^*_{W,Z}) - \hat{L}_n(\hat{f}^*_{W,Z})\right] + \mathbb{E}\left[\hat{L}_n(\hat{f}^*_{W,Z}) - L^*\right] \\
&\leq \mathbb{E}\sup_{f_{W,Z}\in\mathcal{F}}\left|L(f_{W,Z}) - \hat{L}_n(f_{W,Z})\right| + \mathbb{E}\sup_{f_{W,Z}\in\mathcal{F}}\left|\hat{L}_n(f_{W,Z}) - L(f_{W,Z})\right| \\
&= 2\mathbb{E}\sup_{f_{W,Z}\in\mathcal{F}}\left|L(f_{W,Z}) - \hat{L}_n(f_{W,Z})\right| \\
&\leq 2\mathbb{E}\sup_{f_{W,Z}\in\mathcal{F}}\left|L(f_{W,Z}) - \frac{1}{\lfloor n/2\rfloor}\sum_{k=1}^{K}\sum_{i=1}^{\lfloor n/2\rfloor} f_{W,Z_k}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2\rfloor})\right|.
\end{aligned}$$

The last inequality is obtained by the Lemma A.1 in (Clémençon et al., 2008), which refers to the U-process technique. Let $\bar{S} = \bar{\mathbf{x}}_1, ..., \bar{\mathbf{x}}_n$ be an independent copy of $S = \mathbf{x}_1, ..., \mathbf{x}_n$, independent of $\sigma_1, \ldots, \sigma_{\lfloor n/2\rfloor}$, then by a standard symmetrization

technique and the Jensen's inequality (Mohri et al., 2018), the last inequality can be bounded by:

$$
\begin{aligned}
&2\mathbb{E}_{S,\bar{S}} \sup_{f_{W,Z}\in\mathcal{F}} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{k=1}^{K} f_{W,Z_k}\left(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_{i+\lfloor n/2 \rfloor}\right) - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{k=1}^{K} f_{W,Z_k}\left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}\right) \right| \\
&= 2\mathbb{E}_{S,\bar{S},\boldsymbol{\sigma}} \sup_{f_{W,Z}\in\mathcal{F}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{k=1}^{K} \sigma_i \left[f_{W,Z_k}\left(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_{i+\lfloor n/2 \rfloor}\right) - f_{W,Z_k}\left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}\right) \right] \right| \\
&= 4\mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{f_{W,Z}\in\mathcal{F}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{k=1}^{K} \sigma_i f_{W,Z_k}\left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}\right) \right| \\
&= 2R(\mathcal{F}) \\
&\leq 2K \max_k R(\mathcal{F}_k) \\
&= 4K \max_k \mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{f_{W,Z_k}\in\mathcal{F}_k} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i f_{W,Z_k}\left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor}\right) \right| \\
&\leq 4K \max_k \mathbb{E}_{S} \sup_{f_{W,Z_k}\in\mathcal{F}_k} \frac{1}{\lfloor n/2 \rfloor} \left(\sum_{i=1}^{\lfloor n/2 \rfloor} \left(f_{W,Z_k}(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor})\right)^2 \right)^{1/2} \\
&\quad \text{Use Khintchine-Kahane inequality (Latała \& Oleszkiewicz, 1994),} \\
&\leq 4KM \frac{1}{\sqrt{\lfloor n/2 \rfloor}} \quad \text{Use Assumption 1 in the main paper.}
\end{aligned}
$$

Based on the above results, we have $\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq \frac{8KM}{\sqrt{n}}$.

(2.) We then prove that $L(\hat{f}^*_{W,Z}) - L^* \leq \frac{8KM}{\sqrt{n}} + E\sqrt{\frac{8\log\frac{1}{\delta}}{n}}$ with probability $1-\delta$.

Similarly, we can derive that

$$
\begin{aligned}
L(\hat{f}^*_{W,Z}) - L^* &= L(\hat{f}^*_{W,Z}) - \hat{L}_n(\hat{f}^*_{W,Z}) + \hat{L}_n(\hat{f}^*_{W,Z}) - L^* \\
&\leq \sup_{f_{W,Z}\in\mathcal{F}} \left| L(f_{W,Z}) - \hat{L}_n(f_{W,Z}) \right| + \sup_{f_{W,Z}\in\mathcal{F}} \left| \hat{L}_n(f_{W,Z}) - L(f_{W,Z}) \right| \\
&= 2 \sup_{f_{W,Z}\in\mathcal{F}} \left| L(f_{W,Z}) - \hat{L}_n(f_{W,Z}) \right|.
\end{aligned}
$$

Let $\bar{S} = \{\mathbf{x}_1, ..., \bar{\mathbf{x}}_t, ..., \mathbf{x}_n\}$, which are different from S in $\mathbf{x}_t$, and denote $\hat{L}'_n(f_{W,Z})$ as the empirical clustering risk of hypothesis function $f_{W,Z}$ on samples $\bar{S}$, then we have:

$$
\begin{aligned}
&\left| \sup_{f_{W,Z}\in\mathcal{F}} \left| L(f_{W,Z}) - \hat{L}_n(f_{W,Z}) \right| - \sup_{f_{W,Z}\in\mathcal{F}} \left| L(f_{W,Z}) - \hat{L}'_n(f_{W,Z}) \right| \right| \\
&\leq \sup_{f_{W,Z}\in\mathcal{F}} \left| \hat{L}_n(f_{W,Z}) - \hat{L}'_n(f_{W,Z}) \right| \\
&\leq \frac{2}{n(n-1)} \sup_{f_{W,Z}\in\mathcal{F}} \sum_{j=1, j\neq t}^{n} \left(\left| \sum_{k=1}^{K} f_{W,Z_k}(\mathbf{x}_t, \mathbf{x}_j) \right| + \left| \sum_{k=1}^{K} f_{W,Z_k}(\bar{\mathbf{x}}_t, \mathbf{x}_j) \right| \right) \\
&\leq \frac{4}{n} E.
\end{aligned}
$$

The last inequality is obtained because of Assumption 1 in the main paper. So, by McDiarmid inequality (Mohri et al., 2018) with increments bounded by $\frac{4}{n}E$, the term $L(\hat{f}^*_{W,Z}) - L^*$ can be bounded by $\frac{8KM}{\sqrt{n}} + E\sqrt{\frac{8\log\frac{1}{\delta}}{n}}$ with probability $1-\delta$. □

B. Proof of Theorem 2

B.1. Preliminaries

To improve the readability of this paper, we further simplify the notations. Let $g : \mathbb{R}^K \to \mathbb{R}$ be a summation function:

$$\forall \alpha \in \mathbb{R}^K, g(\alpha) = \sum_{i=1}^{K} \alpha_i,$$

and let

$$\ell_{f_{W,Z}}(X, X') = g(f_{W,Z})(X, X') = \sum_{k=1}^{K} f_{W,Z_k}(X, X').$$

Assume that $\mathcal{L}$ is a function class defined as

$$\mathcal{L} := \left\{ \ell_{f_{W,Z}} = \sum_{k=1}^{K} f_{W,Z_k} \middle| f_{W,Z} \in \mathcal{F} \right\}, \tag{1}$$

the Eqs. (2) and (3) in the main paper can thus be written as

$$\hat{L}_n(\ell_{f_{W,Z}}) := \hat{L}_n(f_{W,Z}) = \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} \ell_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_j),$$
$$L(\ell_{f_{W,Z}}) := L(f_{W,Z}) = \mathbb{E}\ell_{f_{W,Z}}(X, X').$$

Furthermore, we define the following local clustering Rademacher complexity:

Definition 1. *For any $r > 0$, the expectation local Rademacher complexity of a function space $\mathcal{L}$ for clustering learning is defined as:*

$$R(\mathcal{L}^r) := R\left(\left\{ \alpha \ell_{f_{W,Z}} \middle| \alpha \in [0,1], \ell_{f_{W,Z}} \in \mathcal{L}, L\left[\left(\alpha \ell_{f_{W,Z}}\right)^2 \right] \leq r \right\} \right),$$

where $\mathcal{L}^r = \left\{ \alpha \ell_{f_{W,Z}} \middle| \alpha \in [0,1], \ell_{f_{W,Z}} \in \mathcal{L}, L\left[\left(\alpha \ell_{f_{W,Z}}\right)^2 \right] \leq r \right\}$ *and* $L\left[\left(\alpha \ell_{f_{W,Z}}\right)^2 \right] := \mathbb{E}\left[\left(\alpha \ell_{f_{W,Z}}\right)^2 \right]$.

From Definition 1, one can easily verify that $R(\mathcal{L})$ is equal to $R(\mathcal{F})$ defined in the main paper, and also there holds that

$$R(\mathcal{L}^r) = R(\mathcal{F}^r),$$

where $R(\mathcal{F}^r)$ is the corresponding local Rademacher complexity of function class $\mathcal{F}$. In this section, we will use the above defined concise notations to finish the proofs.

[Sketch of proof techniques.] We first prove that the generalization error can be bounded through an assumption over the uniform deviation: if uniform deviation $\hat{U}_n(\bar{\mathcal{L}}) \leq \frac{r}{Eh}$, where $\forall h > \max\left(1, \frac{\sqrt{2}}{2E}\right)$ and $\bar{\mathcal{L}}$ is the normalized loss space:

$$\bar{\mathcal{L}} = \left\{ \frac{r}{max\left(L\left(\ell^2_{f_{W,Z}}\right), r\right)} \middle| \ell_{f_{W,Z}} \in \mathcal{L} \right\},$$

for $\forall \ell_{f_{W,Z}} \in \mathcal{L}$,

$$L \leq \max\left\{ \left(\frac{h}{h-1} \hat{L}_n \right), \left(\hat{L}_n + \frac{r}{Eh} \right) \right\}.$$

Then, we propose the upper bound of $\hat{U}_n(\bar{\mathcal{L}})$ with $R(\mathcal{L}^r)$: $\hat{U}_n(\bar{\mathcal{L}}) \leq 2R(\mathcal{L}^r) + \sqrt{\frac{2r \ln \delta}{\lfloor n/2 \rfloor}} + \frac{4 \ln \delta}{3\lfloor n/2 \rfloor}$. The above results show that we can choose a suitable r to satisfy the assumption $\hat{U}_n(\bar{\mathcal{L}}) \leq \frac{r}{Eh}$ to accomplish this proof. Finally, we show that the suitable r can be chosen with the fixed point r^* of $R(\mathcal{L}^r)$. Therefore we obtain that with probability $1 - \delta$:

$$L(\ell_{f_{W,Z}}) \leq \frac{2h+1}{h-1} \hat{L}_n(\ell_{f_{W,Z}}) + c_1 r^* + \frac{c_2}{n-1},$$

where $c_1 = 8Eh$ and $c_2 = 8h \ln \delta + 6$. This proof is inspired by (Liu et al., 2017). By replacing the function class $\mathcal{L}$ with another function class, we can finish the proof.

B.2. Proof of Theorem 2

We first prove the following five lemmas.

Lemma 1. *Let $\bar{\mathcal{L}}$ be the normalized loss space*

$$\bar{\mathcal{L}} = \left\{ \frac{r}{\max(L(\ell^2_{f_{W,Z}}), r)} \ell_{f_{W,Z}} \Big| \ell_{f_{W,Z}} \in \mathcal{L} \right\}. \tag{2}$$

Suppose that, $\forall h > 1$,

$$\hat{U}_n(\bar{\mathcal{L}}) := \sup_{\bar{\ell}_{f_{W,Z}} \in \bar{\mathcal{L}}} \left\{ L(\bar{\ell}_{f_{W,Z}}) - \hat{L}_n(\bar{\ell}_{f_{W,Z}}) \right\} \leq \frac{r}{Eh}.$$

Then, $\forall \ell_{f_{W,Z}} \in \mathcal{L}$, we have

$$L(\ell_{f_{W,Z}}) \leq \max \left\{ \left(\frac{h}{h-1} \hat{L}_n(\ell_{f_{W,Z}}) \right), \left(\hat{L}_n(\ell_{f_{W,Z}}) + \frac{r}{Eh} \right) \right\}.$$

Proof. Note that, $\forall \bar{\ell}_{f_{W,Z}} \in \bar{\mathcal{L}}$:

$$L(\bar{\ell}_{f_{W,Z}}) \leq \hat{L}_n(\bar{\ell}_{f_{W,Z}}) + \hat{U}_n(\bar{\mathcal{L}}) \leq \hat{L}_n(\bar{\ell}_{f_{W,Z}}) + \frac{r}{Eh}. \tag{3}$$

Let us consider the two cases:

1) $L(\ell^2_{f_{W,Z}}) \leq r, \ell_{f_{W,Z}} \in \mathcal{L}$.

2) $L(\ell^2_{f_{W,Z}}) > r, \ell_{f_{W,Z}} \in \mathcal{L}$.

In the first case $\bar{\ell}_{f_{W,Z}} = \ell_{f_{W,Z}}$, by (3), we have

$$L(\ell_{f_{W,Z}}) = L(\bar{\ell}_{f_{W,Z}}) \leq \hat{L}_n(\bar{\ell}_{f_{W,Z}}) + \frac{r}{Eh} = \hat{L}_n(\ell_{f_{W,Z}}) + \frac{r}{Eh}. \tag{4}$$

In the second case, $\bar{\ell}_{f_{W,Z}} = \frac{r}{L(\ell^2_{f_{W,Z}})} \ell_{f_{W,Z}}$, then

$$\begin{aligned} L(\ell_{f_{W,Z}}) - \hat{L}_n(\ell_{f_{W,Z}}) \leq \hat{U}_n(\mathcal{L}) &= \frac{L(\ell^2_{f_{W,Z}})}{r} \hat{U}_n(\bar{\mathcal{L}}) \\ &\leq \frac{E \cdot L(\ell_{f_{W,Z}})}{r} \frac{r}{Eh} = \frac{L(\ell_{f_{W,Z}})}{h}, \end{aligned} \tag{5}$$

where $\hat{U}_n(\mathcal{L}) := \sup_{\ell_{f_{W,Z}} \in \mathcal{L}} \left\{ L(\ell_{f_{W,Z}}) - \hat{L}_n(\ell_{f_{W,Z}}) \right\}$. By combining the results of Eqs. (4) and (5), the proof is over. □

Lemma 2. $\bar{\mathcal{L}} \subseteq \mathcal{L}^r$.

Proof. Let us consider $\mathcal{L}^r$ in the two cases:

1) $L(\ell^2_{f_{W,Z}}) \leq r, \ell_{f_{W,Z}} \in \mathcal{L}$.

2) $L(\ell^2_{f_{W,Z}}) > r, \ell_{f_{W,Z}} \in \mathcal{L}$.

In the first case, $\bar{\ell}_{f_{W,Z}} = \ell_{f_{W,Z}}$ and then:

$$L(\ell^2_{f_{W,Z}}) = L(\bar{\ell}^2_{f_{W,Z}}) \leq r.$$

In the second case, $L(\ell^2_{f_{W,Z}}) > r$, so we have that

$$\bar{\ell}_{f_{W,Z}} = \left[\frac{r}{L(\ell^2_{f_{W,Z}})}\right] \ell_{f_{W,Z}}, \frac{r}{L(\ell^2_{f_{W,Z}})} \leq 1,$$

and the following bound holds:

$$L(\bar{\ell}^2_{f_{W,Z}}) = \left[\frac{r}{L(\ell^2_{f_{W,Z}})}\right]^2 L(\ell^2_{f_{W,Z}}) \leq \left[\frac{r}{L(\ell^2_{f_{W,Z}})}\right] L(\ell^2_{f_{W,Z}}) = r.$$

Thus, the lemma is proved. □

Lemma 3. *$\psi_n(r) = R(\mathcal{L}^r)$ is a sub-root function.*

Proof. In order to prove the lemma, the following properties mush apply:

1) $\psi_n(r)$ is positive
2) $\psi_n(r)$ is non-decreasing
3) $\psi_n(r)/\sqrt{r}$ is non-increasing

By the definition of $R(\mathcal{L}^r)$, it is easy to verity that $R(\mathcal{L}^r)$ is positive.

Concerning the second property, we have that, for $0 \leq r_1 \leq r_2$: $\mathcal{L}^{r_1} \subseteq \mathcal{L}^{r_2}$, therefore

$$\begin{aligned}
\psi_n(r_1) &= \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}} \in \mathcal{L}^{r_1}} \left| \frac{2}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \ell_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_{\lfloor \frac{n}{2} \rfloor + i}) \right| \right] \\
&\leq \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}} \in \mathcal{L}^{r_2}} \left| \frac{2}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \ell_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_{\lfloor \frac{n}{2} \rfloor + i}) \right| \right] \\
&= \psi_n(r_2).
\end{aligned}$$

Finally, concerning the third property, for $0 \leq r_1 \leq r_2$, let

$$\ell^{r_2}_{f_{W,Z}} = \arg\sup_{\ell_{f_{W,Z}} \in \mathcal{L}^{r_2}} \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}} \in \mathcal{L}^{r_2}} \left| \frac{2}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \ell_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_{\lfloor \frac{n}{2} \rfloor + i}) \right| \right].$$

Note that, since $\frac{r_1}{r_2} \leq 1$, we have that $\sqrt{\frac{r_1}{r_2}} \ell^{r_2}_{f_{W,Z}} \in \mathcal{L}^{r_2}$. Consequently:

$$L\left[\left(\sqrt{\frac{r_1}{r_2}} \ell^{r_2}_{f_{W,Z}}\right)^2\right] = \frac{r_1}{r_2} L\left[(\ell^{r_2}_{f_{W,Z}})^2\right] \leq r_1.$$

Thus, we have that:

$$\begin{aligned}
\psi_n(r_1) &= \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}} \in \mathcal{L}^{r_1}} \left| \frac{2}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \ell_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_{\lfloor \frac{n}{2} \rfloor + i}) \right| \right] \\
&\geq \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\left| \frac{2}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \sqrt{\frac{r_1}{r_2}} \ell^{r_2}_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_{\lfloor \frac{n}{2} \rfloor + i}) \right| \right] \\
&= \sqrt{\frac{r_1}{r_2}} \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}} \in \mathcal{L}^{r_2}} \left| \frac{2}{\lfloor \frac{n}{2} \rfloor} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \sigma_i \ell_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_{\lfloor \frac{n}{2} \rfloor + i}) \right| \right] \\
&= \sqrt{\frac{r_1}{r_2}} \psi_n(r_2),
\end{aligned}$$

which allows proving the claim since

$$\frac{\psi_n(r_2)}{\sqrt{r_2}} \leq \frac{\psi_n(r_1)}{\sqrt{r_1}}.$$

□

Lemma 4. *With probability at least* $1-\delta$*,*

$$\hat{U}_n(\bar{\mathcal{L}}) \leq 2R(\bar{\mathcal{L}}) + \sqrt{\frac{2r\ln\delta}{\lfloor n/2 \rfloor}} + \frac{4\ln\delta}{3\lfloor n/2 \rfloor}.$$

Proof. Note that $\hat{L}_n(\ell_{f_{W,Z}}) = \frac{1}{n(n-1)}\sum_{i\neq j}\ell_{f_{W,Z}}(\mathbf{x}_i,\mathbf{x}_j)$ is a non-sum-of-i.i.d. pairwise loss. According to (Clémençon et al., 2005; 2008), we introduce permutations to convert the non-sum-of-i.i.d pairwise loss to a sum-of-i.i.d form. Assume Γ is the symmetric group of degree n and $\pi \in \Gamma$ which permutes the n samples. Then we have,

$$\hat{L}_n(\ell_{f_{W,Z}}) \overset{P}{=} \frac{1}{n!}\sum_{\pi\in\Gamma}\frac{1}{\lfloor \frac{n}{2} \rfloor}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor}\ell_{f_{W,Z}}(\mathbf{x}_j,\mathbf{x}_{\lfloor \frac{n}{2} \rfloor+j}), \tag{6}$$

where $\overset{P}{=}$ means identity in distribution. Denote

$$G(S,\bar{\mathcal{L}}) = \sup_{\ell_{f_{W,Z}}\in\bar{\mathcal{L}}}\left|\frac{1}{\lfloor \frac{n}{2} \rfloor}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor}\ell_{f_{W,Z}}(\mathbf{x}_j,\mathbf{x}_{\lfloor \frac{n}{2} \rfloor+j}) - L(\ell_{f_{W,Z}})\right|, \tag{7}$$

then, we have

$$\begin{aligned}
&U_n(\bar{\mathcal{L}}) \\
&= \mathbb{E}_S \sup_{\ell_{f_{W,Z}}\in\bar{\mathcal{L}}}\left[L(\ell_{f_{W,Z}}) - \hat{L}_n(\ell_{f_{W,Z}})\right] \\
&\leq \frac{1}{n!}\sum_{\pi\in\Gamma}\mathbb{E}_S\left[\sup_{\ell_{f_{W,Z}}\in\bar{\mathcal{L}}}\left|\frac{1}{\lfloor \frac{n}{2} \rfloor}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor}\ell_{f_{W,Z}}(\mathbf{x}_j,\mathbf{x}_{\lfloor \frac{n}{2} \rfloor+j}) - L(\ell_{f_{W,Z}})\right|\right] \\
&= \mathbb{E}_S\left[G(S,\bar{\mathcal{L}})\right].
\end{aligned} \tag{8}$$

Next, we give a bound for $\mathbb{E}_S\left[G(S,\bar{\mathcal{L}})\right]$ by use of symmetrization. We introduce a ghost data set

$$S' = \{\mathbf{x}'_1,\ldots,\mathbf{x}'_n\}$$

that is independent of S and identically distributed. Assume $\sigma_1,\ldots,\sigma_{\lfloor n/2 \rfloor}$ are independent Rademacher random variables, independent of S and S'.

$$\begin{aligned}
&\mathbb{E}_S\left[G(S,\bar{\mathcal{L}})\right] \\
&\leq \mathbb{E}_{S,S'}\left[\sup_{\ell_{f_{W,Z}}\in\bar{\mathcal{L}}}\left|\frac{1}{\lfloor \frac{n}{2} \rfloor}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor}\left(\ell_{f_{W,Z}}(\mathbf{x}_j,\mathbf{x}_{\lfloor \frac{n}{2} \rfloor+j}) - \ell_{f_{W,Z}}(\mathbf{x}'_j,\mathbf{x}'_{\lfloor \frac{n}{2} \rfloor+j})\right)\right|\right] \\
&= \mathbb{E}_{S,S',\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}}\in\bar{\mathcal{L}}}\left|\frac{1}{\lfloor \frac{n}{2} \rfloor}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor}\sigma_j\left(\ell_{f_{W,Z}}(\mathbf{x}_j,\mathbf{x}_{\lfloor \frac{n}{2} \rfloor+j}) - \ell_{f_{W,Z}}(\mathbf{x}'_j,\mathbf{x}'_{\lfloor \frac{n}{2} \rfloor+j})\right)\right|\right] \\
&= \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}}\in\bar{\mathcal{L}}}\left|\frac{2}{\lfloor \frac{n}{2} \rfloor}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor}\sigma_j\ell_{f_{W,Z}}(\mathbf{x}_j,\mathbf{x}_{\lfloor \frac{n}{2} \rfloor+j})\right|\right] = R(\bar{\mathcal{L}}).
\end{aligned} \tag{9}$$

In the following, we will bound the $G(S, \bar{\mathcal{L}})$. Note that, for all $\ell_{f_{W,Z}} \in \bar{\mathcal{L}}$,

$$V^2(\ell_{f_{W,Z}}) = L(\ell^2_{f_{W,Z}}) - [L(\ell_{f_{W,Z}})]^2 \leq L(\ell^2_{f_{W,Z}}) = r,$$

where $V^2(\ell_{f_{W,Z}})$ is the variance of $\ell_{f_{W,Z}} \in \bar{\mathcal{L}}$. Thus, according to the Bennett concentration inequality (Bousquet, 2002), with probability at least $1-\delta$, we have

$$G(S, \bar{\mathcal{L}}) \leq \mathbb{E}_S[G(S, \bar{\mathcal{L}})] + \sqrt{\frac{2r \ln \frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor} + \frac{4\mathbb{E}_S[G(S, \bar{\mathcal{L}})] \ln \frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{\ln \frac{1}{\delta}}{3\lfloor \frac{n}{2} \rfloor}.$$

From (9), we know that $\mathbb{E}_S[G(S, \bar{\mathcal{L}})] \leq R(\bar{\mathcal{L}})$, so

$$G(S, \bar{\mathcal{L}}) \leq R(\bar{\mathcal{L}}) + \sqrt{\frac{2r \ln \frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor} + \frac{4R(\bar{\mathcal{L}}) \ln \frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{\ln \frac{1}{\delta}}{3\lfloor \frac{n}{2} \rfloor}. \tag{10}$$

Note that, for $u, v \geq 0$,

$$\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}, 2\sqrt{uv} \leq u + v.$$

So, by (10), the following inequality holds:

$$\begin{aligned} & G(S, \bar{\mathcal{L}}) \\ \leq & R(\bar{\mathcal{L}}) + \sqrt{\frac{2r \ln \frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + 2\sqrt{\frac{\ln \frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor} R(\bar{\mathcal{L}})} + \frac{\ln \frac{1}{\delta}}{3\lfloor \frac{n}{2} \rfloor} \\ \leq & 2R(\bar{\mathcal{L}}) + \sqrt{\frac{2r \ln \frac{1}{\delta}}{\lfloor \frac{n}{2} \rfloor}} + \frac{4 \ln \frac{1}{\delta}}{3\lfloor \frac{n}{2} \rfloor}. \end{aligned} \tag{11}$$

Similar with the proof (8), it is easy to verity that

$$\hat{U}_n(\bar{\mathcal{L}}) \leq \frac{1}{n!} \sum_{\pi \in \Gamma} G(S, \bar{\mathcal{L}}). \tag{12}$$

By combining the results of (11) and (12), the proof is over. □

Lemma 5. *Assume that r^* is the fixed point of $R(\mathcal{L}^r)$, that is, r^* is the solution of $R(\mathcal{L}^r) = r$ with respect to r. Then, $\forall h > \max\left(1, \frac{\sqrt{2}}{2E}\right)$, with probability $1-\delta$:*

$$L(\ell_{f_{W,Z}}) \leq \max\left\{\frac{h}{h-1}\hat{L}_n(\ell_{f_{W,Z}}), \hat{L}_n(\ell_{f_{W,Z}}) + c_1 r^* + \frac{c_2}{n-1}\right\},$$

where $c_1 = 8hE$ and $c_2 = 8h \ln \frac{1}{\delta} + 6 \ln \frac{1}{\delta}$.

Proof. According to Lemma 2, we know that $\bar{\mathcal{L}} \subseteq \mathcal{L}^r$. Therefore, from Lemma 4, with probability $1-\delta$, we have

$$\begin{aligned} \hat{U}_n(\bar{\mathcal{L}}) \leq & 2R(\bar{\mathcal{L}}) + \sqrt{\frac{2r \ln \frac{1}{\delta}}{\lfloor n/2 \rfloor} + \frac{4 \ln \frac{1}{\delta}}{3\lfloor n/2 \rfloor}} \\ \leq & 2R(\mathcal{L}^r) + \sqrt{\frac{2r \ln \frac{1}{\delta}}{\lfloor n/2 \rfloor} + \frac{4 \ln \frac{1}{\delta}}{3\lfloor n/2 \rfloor}}. \end{aligned}$$

By Lemma 3, we know that $R(\mathcal{L}^r)$ is a sub-root function. Thus, $R(\mathcal{L}^r) \leq \sqrt{rr^*}$ for all $r \geq r^*$. Then,

$$\hat{U}_n(\bar{\mathcal{L}}) \leq 2\sqrt{rr^*} + \sqrt{\frac{2r \ln \frac{1}{\delta}}{\lfloor n/2 \rfloor} + \frac{4 \ln \frac{1}{\delta}}{3\lfloor n/2 \rfloor}}.$$

The last step of the proof consists in showing that r can be chosen, such that $\hat{U}_n(\bar{\mathcal{L}}) \leq \frac{r}{Eh}$ and $r \geq r^*$, so that we can exploit Lemma 1 and conclude the proof. For this purpose, we set

$$A = 2\sqrt{r^*} + \sqrt{\frac{2\ln\frac{1}{\delta}}{\lfloor n/2 \rfloor}}, B = \frac{4\ln\frac{1}{\delta}}{3\lfloor n/2 \rfloor}.$$

Thus, we have to find the solution of

$$A\sqrt{r} + B = \frac{r}{Eh},$$

which is

$$r = \frac{\left[\left(\frac{2B}{hE} + A^2\right) + \sqrt{\left(\frac{2B}{hE} + A^2\right)^2 - \frac{4B^2}{E^2h^2}}\right]}{\frac{2}{E^2h^2}} \tag{13}$$

Since $h \geq \max(1, \frac{\sqrt{2}}{2E})$, $h^2E^2 \geq \frac{1}{2}$. Therefore, from (13), we have

$$\begin{aligned} r &\geq A^2E^2h^2 \geq \frac{A^2}{2} = r^*, \\ r &\leq A^2E^2h^2 + 2BEh. \end{aligned}$$

Thus, we have

$$\begin{aligned} \frac{r}{Eh} &\leq A^2Eh + 2B \\ &= \left(2\sqrt{r^*} + \sqrt{\frac{2\ln\frac{1}{\delta}}{\lfloor n/2 \rfloor}}\right)^2 Eh + \frac{8\ln\frac{1}{\delta}}{3\lfloor n/2 \rfloor}. \end{aligned}$$

Note that, $\forall a, b > 0$, $(a+b)^2 \leq 2a^2 + 2b^2$, so we have that

$$\begin{aligned} \frac{r}{Eh} &\leq 8Ehr^* + \frac{8h}{n-1}\ln\frac{1}{\delta} + \frac{16}{3n-3}\ln\frac{1}{\delta} \\ &\leq 8Ehr^* + \frac{8h+6}{n-1}\ln\frac{1}{\delta}. \end{aligned}$$

By substituting the above inequality into Lemma 1, we can prove that $\forall h > \max\left(1, \frac{\sqrt{2}}{2E}\right)$, with probability $1-\delta$,

$$L(\ell_{f_{W,Z}}) \leq \max\left\{\frac{h}{h-1}\hat{L}_n(\ell_{f_{W,Z}}), \hat{L}_n(\ell_{f_{W,Z}}) + c_1r^* + \frac{c_2}{n-1}\right\},$$

where $c_1 = 8hE$ and $c_2 = 8h\ln\frac{1}{\delta} + 6\ln\frac{1}{\delta}$. □

Proof of Theorem 2. By Lemma 5, $\forall h > \max\left(1, \frac{\sqrt{2}}{2E}\right)$, with probability $1-\delta$ we obtain that

$$L(\ell_{f_{W,Z}}) \leq \frac{2h-1}{h-1}\hat{L}_n(\ell_{f_{W,Z}}) + c_1r^* + \frac{c_2}{n-1},$$

where r^* is the fixed point of $R(\mathcal{L}^r)$.

Assume that $\hat{\ell}^*_{f_{W,Z}} = \arg\min_{\ell_{f_{W,Z}} \in \mathcal{L}} \hat{L}_n(\ell_{f_{W,Z}})$ and $\ell^*_{f_{W,Z}} = \inf_{\ell_{f_{W,Z}} \in \mathcal{L}} L(\ell_{f_{W,Z}})$, so there holds that $L(\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}}) \geq 0$. And since $\ell_{f_{W,Z}} \leq E$ due to Assumption 1, so there holds that $L((\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}})^2) \leq 2EL(\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}})$. If we apply Lemmas 1-5 to the class $\{\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}}\}$, we will get $\forall h > \max\left(1, \frac{\sqrt{2}}{4E}\right)$, with probability $1-\delta$

$$L(\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}}) \leq \max\left\{\frac{h}{h-1}\left[\hat{L}_n(\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}})\right], \hat{L}_n(\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}}) + c_1r^* + \frac{c_2}{n-1}\right\},$$

where $c_1 = 16hE$ and $c_2 = 8h\ln\frac{1}{\delta} + 6\ln\frac{1}{\delta}$, and where r^* is the fixed point of the local Rademacher complexity of function class $\{\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}}\}$. Note that $\hat{L}_n(\hat{\ell}^*_{f_{W,Z}} - \ell^*_{f_{W,Z}}) \leq 0$, so we have

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_1 r^* + \frac{c_2}{n-1}.$$

And, note that from the Definition 1, the local Rademacher complexity of the function class $\{\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}}\}$ is equal to the local Rademacher complexity of the excess function class $\mathcal{F}_{exc}$.

Therefore, we obtain that under Assumption 1 in the main paper, and let r^* be the fixed point of $R(\mathcal{F}^r_{exc})$, that is r^* is the solution of $R(\mathcal{F}^r_{exc}) = r$ with respect to r. Then, $\forall h > \max\left(1, \frac{\sqrt{2}}{4E}\right)$, with probability $1-\delta$:

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_{h,E} r^* + \frac{c_{h,\delta}}{n-1}, \tag{14}$$

where $c_{h,E}$ and $c_{h,\delta}$ are constants dependent on h, E and h, δ respectively. □

C. Proof of Theorem 3

Lemma 6. *Let $\mathcal{L}$ be a function class satisfying Eq. (1). The excess loss class is defined as: $\mathcal{L}_{exc} := \{\ell_{f_{W,Z}} - \ell^*_{f_{W,Z}}\}$. Since $\|\ell_{f_{W,Z}}\|_\infty \leq E$, $\forall \ell_{f_{W,Z}} \in \mathcal{L}$, there holds the following inequality:*

$$R(\mathcal{L}^r_{exc}) \leq \inf_{\epsilon>0}\left[2R\left\{\ell_{f_{W,Z}} \in \widetilde{\mathcal{L}} : \hat{L}_n(\ell^2_{f_{W,Z}}) \leq \epsilon^2\right\} + \frac{64E\log\mathcal{N}(\epsilon/2, \mathcal{L}, \|\cdot\|_2)}{n} + \sqrt{\frac{8r\log\mathcal{N}(\epsilon/2, \mathcal{L}, \|\cdot\|_2)}{n}}\right],$$

where $\widetilde{\mathcal{L}} := \left\{\ell_{f_{W,Z}} - \ell'_{f_{W,Z}} : \ell_{f_{W,Z}}, \ell'_{f_{W,Z}} \in \mathcal{L}\right\}$ and $\hat{L}_n(\ell^2_{f_{W,Z}}) = \frac{1}{n(n-1)}\sum_{i\neq j}\ell^2_{f_{W,Z}}(\mathbf{x}_i, \mathbf{x}_j)$.

Proof. Let $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{x}_{i+\lfloor\frac{n}{2}\rfloor})$, one can see that $\mathbf{z}_1, ..., \mathbf{z}_{\lfloor\frac{n}{2}\rfloor}$ are i.i.d. samples. It is easy to write the following local Rademacher complexity of class $\mathcal{L}$:

$$R\left(\left\{\ell_{f_{W,Z}} \in \mathcal{L} : L(\ell^2_{f_{W,Z}}) \leq r\right\}\right) = \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{\ell_{f_{W,Z}} \in \left\{\ell_{f_{W,Z}} \in \mathcal{L} : L(\ell^2_{f_{W,Z}}) \leq r\right\}}\left|\frac{2}{\lfloor n/2\rfloor}\sum_{j=1}^{\lfloor n/2\rfloor}\sigma_j \ell_{f_{W,Z}}(\mathbf{z}_j)\right|\right].$$

Use the proof method of Theorem 2 in paper (Lei et al., 2016), it is easy to obtain:

$$\begin{aligned} &R\left(\left\{\ell_{f_{W,Z}} \in \mathcal{L} : L(\ell^2_{f_{W,Z}}) \leq r\right\}\right) \leq \\ &\inf_{\epsilon>0}\left[2R\left\{\ell_{f_{W,Z}} \in \widetilde{\mathcal{L}} : \hat{L}_n(\ell^2_{f_{W,Z}}) \leq \epsilon^2\right\} + \frac{64E\log\mathcal{N}(\epsilon/2, \mathcal{L}, \|\cdot\|_2)}{n} + \sqrt{\frac{8r\log\mathcal{N}(\epsilon/2, \mathcal{L}, \|\cdot\|_2)}{n}}\right]. \end{aligned} \tag{15}$$

Note that there is no difference between the metric entropy of the function class $\mathcal{L}_{exc}$ and mertic entropy of the loss class $\mathcal{L}$ itself: that is, from the definition of covering number, one has

$$\log\mathcal{N}_\infty(\epsilon, \mathcal{L}, S) = \log\mathcal{N}_\infty(\epsilon, \mathcal{L}_{exc}, S). \tag{16}$$

This implies that we can bound the local Rademacher complexity of the excess loss class $\mathcal{L}_{exc}$ by:

$$\inf_{\epsilon>0}\left[2R\left\{\ell_{f_{W,Z}} \in \widetilde{\mathcal{L}} : \hat{L}_n(\ell^2_{f_{W,Z}}) \leq \epsilon^2\right\} + \frac{64E\log\mathcal{N}(\epsilon/2, \mathcal{L}, \|\cdot\|_2)}{n} + \sqrt{\frac{8r\log\mathcal{N}(\epsilon/2, \mathcal{L}, \|\cdot\|_2)}{n}}\right],$$

where $R\left\{\ell_{f_{W,Z}} \in \widetilde{\mathcal{L}} : \hat{L}_n(\ell^2_{f_{W,Z}}) \leq \epsilon^2\right\}$ is obtained by using Dudley entropy integral inequality (Lemma A.5 in (Lei et al., 2016)) and Eq. (16).

□

Proof of Theorem 3. Similar to Section A, $R(\mathcal{F}^r_{exc})$ in this proof can be bounded by:

$$K\max_k R(\mathcal{F}^r_{exc,k}) = K\max_k \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{f_{W,Z_k}\in\mathcal{F}^r_{exc,k}}\left|\frac{2}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\sigma_i f_{W,Z_k}(\mathbf{x}_i,\mathbf{x}_{i+\lfloor\frac{n}{2}\rfloor})\right|\right],$$

where $\mathcal{F}^r_{exc,k}$ is a function class of the output coordinate k of $\mathcal{F}^r_{exc}$.

By Lemma 6, it is easy to verify that:

$$\begin{aligned}&R(\mathcal{F}^r_{exc}) \le K\max_k R(\mathcal{F}^r_{exc,k})\\ &\le K\max_k\inf_{\epsilon>0}\left[2R\left\{f_{W,Z_k}\in\widetilde{\mathcal{F}_k}:\hat{L}_n(f^2_{W,Z_k})\le\epsilon^2\right\}+\frac{64M\log\mathcal{N}(\epsilon/2,\mathcal{F}_k,\|\cdot\|_2)}{n}+\sqrt{\frac{8r\log\mathcal{N}(\epsilon/2,\mathcal{F}_k,\|\cdot\|_2)}{n}}\right],\end{aligned} \tag{17}$$

where $\widetilde{\mathcal{F}_k} := \{f_{W,Z_k} - f'_{W,Z_k} : f_{W,Z_k}, f'_{W,Z_k}\in\mathcal{F}_k\}$ and $\hat{L}_n(f^2_{W,Z_k}) = \frac{1}{n(n-1)}\sum_{i\neq j} f^2_{W,Z_k}(\mathbf{x}_i,\mathbf{x}_j)$.

After obtaining the relationships between the expected clustering local Rademacher complexity and the covering number, we can use some mild assumptions of the covering number and to obtain the suitable fixed point r^*.

(1). Assume that there exist three positive constants γ, d and p satisfying $\log\mathcal{N}(\epsilon,\mathcal{F}_k,\|\cdot\|_2)\le d\log^p(\gamma/\epsilon)$ for any $0<\epsilon\le\gamma$ and $k=1,...,K$. Based on Eq. (17) and the Corollary 1 in (Lei et al., 2016), for any $0<r<\gamma^2$ and $n\ge\gamma^{-2}$ it is easy to verify that:

$$R(\mathcal{F}^r_{exc}) \le c_{M,p,\gamma}K\min\left[\left(\sqrt{\frac{dr\log^p(2\gamma r^{-1/2})}{n}}+\frac{d\log^p(2\gamma r^{-1/2})}{n}\right),\left(\frac{d\log^p(2\gamma n^{1/2})}{n}+\sqrt{\frac{rd\log^p(2\gamma n^{1/2})}{n}}\right)\right],$$

where $c_{M,p,\gamma}$ is a constant dependent on M, γ and p. Then, we can set

$$R(\mathcal{F}^r_{exc}) \le c_{M,p,\gamma}K\left(\frac{d\log^p(2\gamma n^{1/2})}{n}+\sqrt{\frac{rd\log^p(2\gamma n^{1/2})}{n}}\right).$$

The sub-root function can be set as:

$$\psi(r) := c_{M,p,\gamma}K\left[\frac{d\log^p(2\gamma n^{1/2})}{n}+\sqrt{\frac{rd\log^p(2\gamma n^{1/2})}{n}}\right].$$

Let r^* be its fixed point then we have:

$$r^* = c_{M,p,\gamma}K\left[\frac{d\log^p(2\gamma n^{1/2})}{n}+\sqrt{\frac{r^*d\log^p(2\gamma n^{1/2})}{n}}\right].$$

Denote $\frac{d\log^p(2\gamma n^{1/2})}{n}$ as x, we get an equation:

$$r^* = c_{M,p,\gamma}K\left(x+\sqrt{xr^*}\right),$$

Solving this equation, it is easy to verify that $r^*\le c_{M,p,\gamma}K^2x$. That is $r^*\le\frac{c_{M,p,\gamma,d}K^2\log^p(2\gamma n^{1/2})}{n}$, so finally we obtain that $r^*\le\frac{c_{M,p,\gamma,d}K^2\log^p(n)}{n}$. By substituting this into Eq. (14), we obtain that: with probability $1-\delta$,

$$L(\hat{f}^*_{W,Z}) - L^* \le c_{M,p,\gamma,d,h,E}K^2\frac{\log^p(n)}{n}+\frac{c_{h,\delta}}{n-1}.$$

(2). Assume that there exist two constants $\gamma>0$ and $p>0$ satisfying $\log\mathcal{N}(\epsilon,\mathcal{F}_k,\|\cdot\|_2)\le\gamma\epsilon^{-p}$ for any $k=1,...,K$. Based on Eq. (17) and the Corollary 3 in (Lei et al., 2016), it is easy to verify that

$$R(\mathcal{F}^r_{exc}) \le c_{M,p,\gamma}K\left[n^{-1/2}\epsilon^{1-p/2}+\epsilon^{-p}n^{-1}+\sqrt{r\epsilon^{-p}n^{-1}}\right],$$

where $c_{M,p,\gamma}$ is a constant dependent on M, γ and p. We can set:

$$\psi(r) := c_{M,p,\gamma} K \left[n^{-1/2} \epsilon^{1-p/2} + \epsilon^{-p} n^{-1} + \sqrt{r \epsilon^{-p} n^{-1}} \right].$$

Let r^* be its fixed point then we have:

$$r^* = c_{M,p,\gamma} K \left[n^{-1/2} \epsilon^{1-p/2} + \epsilon^{-p} n^{-1} + \sqrt{r^* \epsilon^{-p} n^{-1}} \right].$$

Denote $n^{-1/2}\epsilon^{1-p/2}$ as x and $\epsilon^{-p}n^{-1}$ as y, we get an equation:

$$r^* = c_{M,p,\gamma} K \left(x + y + \sqrt{y r^*} \right).$$

Solving this equation, it is easy to verify that $r^*(\epsilon) \leq c_{M,p,\gamma} K^2 [n^{-1/2}\epsilon^{1-p/2} + \epsilon^{-p} n^{-1}]$. Since $\epsilon > 0$, we can choose $\epsilon = n^{-\frac{1}{2+P}}$, then we obtain

$$r^* = c_{M,p,\gamma} K^2 n^{-\frac{2}{p+2}}.$$

By substituting this into Eq. (14), we obtain that: with probability $1 - \delta$,

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_{M,p,\gamma,h,E} K^2 n^{-\frac{2}{p+2}} + \frac{c_{h,\delta}}{n-1}.$$

(3). Assume that there exist two constants $\gamma > 0$ and $p > 0$ satisfying $\log \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_2) \leq \gamma \epsilon^{-p} \log^2 \frac{2}{\epsilon}$ for any $k = 1, ..., K$. Based on Eq. (17) and the Corollary 2 in (Lei et al., 2016), it is easy to verify that

$$R(\mathcal{F}^r_{exc}) \leq c_{M,p,\gamma} K \left[n^{-1/2} \epsilon^{1-p/2} \log \frac{1}{\epsilon} + \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon} + \sqrt{r \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon}} \right],$$

where $c_{M,p,\gamma}$ is a constant dependent on M, γ and p. We can set:

$$\psi(r) := c_{M,p,\gamma} K \left[n^{-1/2} \epsilon^{1-p/2} \log \frac{1}{\epsilon} + \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon} + \sqrt{r \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon}} \right].$$

Let r^* be its fixed point then we have:

$$r^* = c_{M,p,\gamma} K \left[n^{-1/2} \epsilon^{1-p/2} \log \frac{1}{\epsilon} + \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon} + \sqrt{r^* \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon}} \right].$$

Denote $n^{-1/2}\epsilon^{1-p/2} \log \frac{1}{\epsilon}$ as x and $\epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon}$ as y, we get an equation:

$$r^* = c_{M,p,\gamma} K \left(x + y + \sqrt{y r^*} \right).$$

Solving this equation, it is easy to verify that $r^*(\epsilon) \leq c_{M,p,\gamma} K^2 [n^{-1/2}\epsilon^{1-p/2} \log \frac{1}{\epsilon} + \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon}]$. Since $\epsilon > 0$, we can choose $\epsilon = (\log n)^{\frac{2}{p+2}} n^{-\frac{1}{2+P}}$, then we obtain

$$r^* = c_{M,p,\gamma} K^2 n^{-\frac{2}{p+2}} (\log n)^{\frac{2-p}{p+2}} \log \frac{n}{(\log n)^{\frac{2}{p+2}}}.$$

By substituting this into Eq. (14), we obtain that: with probability $1 - \delta$,

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_{M,p,\gamma,h,E} K^2 n^{-\frac{2}{p+2}} (\log n)^{\frac{2-p}{p+2}} \log \frac{n}{(\log n)^{\frac{2}{p+2}}} + \frac{c_{h,\delta}}{n-1}.$$

□

D. Proof of Theorem 4

Lemma 7. *$\sum_{k=1}^{K} f_{W,Z_k}$ is K-Lipschitz with respect to the L_∞ norm in the worst case. For the hard clustering scheme, $\sum_{k=1}^{K} f_{W,Z_k}$ is 1-Lipschitz with respect to the L_∞ norm.*

Proof. **(1.)** For $\forall f_{W,Z}, f'_{W,Z} \in \mathcal{F}$,

$$\left\|\sum_{k=1}^{K} f_{W,Z_k} - \sum_{k=1}^{K} f'_{W,Z_k}\right\|_\infty = \left\|f_{W,Z_1} + \cdots + f_{W,Z_K} - f'_{W,Z_1} - \cdots - f'_{W,Z_K}\right\|_\infty \leq K \left\|f_{W,Z} - f'_{W,Z}\right\|_\infty .$$

(2.) We have mentioned in the main paper that in the hard clustering scheme, a pair of observations can at most correspond to one cluster, which means that Z_k is valued either 0 or 1 for a pair of observations where $k = 1, ..., K$, and at most one valued 1. Thus, for the hard clustering scheme,

$$\left\|\sum_{k=1}^{K} f_{W,Z_k} - \sum_{k=1}^{K} f'_{W,Z_k}\right\|_\infty = \left\|f_{W,Z_1} + \cdots + f_{W,Z_K} - f'_{W,Z_1} - \cdots - f'_{W,Z_K}\right\|_\infty \leq \left\|f_{W,Z} - f'_{W,Z}\right\|_\infty .$$

Lemma 7 suggests that Assumption 5 in the main paper is a very mild assumption. □

Lemma 8. *(Foster & Rakhlin, 2019) Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^K\}$, and let $\phi : \mathbb{R}^K \to \mathbb{R}$ be L-lipschitz with respect to the L_∞ norm, that is $\|\phi(V) - \phi(V')\|_\infty \leq L\|V - V'\|_\infty$, $\forall V, V' \in \mathbb{R}^K$. For any $\delta > 0$, there exists a constant $C > 0$ such that if $|\phi(f(x))| \vee \|f(x)\|_\infty \leq \beta$, then*

$$R_n(\phi \circ \mathcal{F}) \leq C \cdot L\sqrt{K} \max_i \widetilde{R}_n(\mathcal{F}_i) \log^{\frac{3}{2}+\delta}\left(\frac{\beta n}{\max_i \widetilde{R}_n(\mathcal{F}_i)}\right),$$

where $R_n(\phi \circ \mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\phi(f(\mathbf{x}_i))\right|\right]$, $\widetilde{R}_n(\mathcal{F}_i) = \sup_{\mathcal{S}\in\mathcal{X}^n} R_n(\mathcal{F}_i)$.

Proof of Theorem 4. Assume that $\mathcal{Z} = \mathcal{X} \times \mathcal{X}$, based on Assumption 1 in the main paper, we have $\left|\sum_{k=1}^{K} f_{W,Z_k}(z)\right| \vee \|f_{W,Z}(z)\|_\infty \leq E$ for all $z \in \mathcal{Z}$. Let $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{x}_{i+\lfloor \frac{n}{2} \rfloor})$, and let $S = \{\mathbf{z}_1, ..., \mathbf{z}_{\lfloor n/2 \rfloor} \in \mathcal{Z}^{\lfloor \frac{n}{2} \rfloor}\}$. Note that $\mathbf{z}_i$ in S are i.i.d. samples, thus Lemma 8 can be applied to our defined empirical clustering Rademacher complexity $R_n(\mathcal{F})$, where $R_n(\mathcal{F})$ is defined by considering the U-process technique. Based on Assumption 5 in the main paper and Lemma 8, we then bound $R_n(\mathcal{F})$ in the following form: for any $\eta > 0$, there exists a constant $C > 0$ such that

$$R_n(\mathcal{F}) \leq CL\sqrt{K} \max_k \widetilde{R}_n(\mathcal{F}_k) \log^{\frac{3}{2}+\eta}\left(\frac{E\lfloor n/2 \rfloor}{\max_k \widetilde{R}_n(\mathcal{F}_k)}\right) \leq CL\sqrt{K} \max_k \widetilde{R}_n(\mathcal{F}_k) \log^{\frac{3}{2}+\eta}\left(\frac{En}{\max_k \widetilde{R}_n(\mathcal{F}_k)}\right).$$

Furthermore, we refine Lemma 8 and bound $\widetilde{R}_n(\mathcal{F}_k)$ by:

$$\begin{aligned}\widetilde{R}_n(\mathcal{F}_k) = \sup_{S\in\mathcal{Z}^{\lfloor \frac{n}{2} \rfloor}} R_n(\mathcal{F}_k) &= \sup_{S\in\mathcal{Z}^{\lfloor \frac{n}{2} \rfloor}} \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f_{W,Z_k}\in\mathcal{F}_k}\left|\frac{2}{\lfloor \frac{n}{2} \rfloor}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor}\sigma_j f_{W,Z_k}(\mathbf{z}_j)\right|\right] \\ &\geq 2 \sup_{S\in\mathcal{Z}^{\lfloor \frac{n}{2} \rfloor}} \frac{1}{\lfloor \frac{n}{2} \rfloor}\left(\sup_{f_{W,Z_k}\in\mathcal{F}_k}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} f^2_{W,Z_k}(\mathbf{z}_j)\right)^{\frac{1}{2}},\end{aligned}$$

where the last inequality is obtained by Khintchine inequality with $p = 1$ in (Haagerup, 1981). Since $f_{W,Z_k} \leq M$, we set $\sup_{S\in\mathcal{Z}^{\lfloor \frac{n}{2} \rfloor}} \frac{1}{\lfloor \frac{n}{2} \rfloor}\left(\sup_{f_{W,Z_k}\in\mathcal{F}_k}\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} f^2_{W,Z_k}(\mathbf{z}_j)\right)^{\frac{1}{2}} = M\frac{1}{\sqrt{\lfloor n/2 \rfloor}}$. So

$$\forall k, \quad \widetilde{R}_n(\mathcal{F}_k) \geq \frac{2M}{\sqrt{n}}.$$

Thus, we can prove that:

$$\frac{En}{\max_k \widetilde{R}_n(\mathcal{F}_k)} \leq \frac{En^{3/2}}{2M},$$

Based on the above results, we finally obtain that under Assumption 1 and 5 in the main paper, for any $\eta > 0$ and $S = \mathbf{x}_{i=1}^n \in \mathcal{X}^n$, there exists a constant $C > 0$ such that

$$R_n(\mathcal{F}) \leq CL\sqrt{K} \max_k \widetilde{R}_n(\mathcal{F}_k) \log^{\frac{3}{2}+\eta}(\sqrt{n}).$$

□

E. Proof of Theorem 5

Proof. In Section A, we have proved that:

$$\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq 4\mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{f_{W,Z} \in \mathcal{F}} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \sum_{k=1}^{K} \sigma_i f_{W,Z_k} \left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor} \right) \right| = 2R(\mathcal{F}).$$

Based on Theorem 4, for any $\eta > 0$, there exists a constant $C > 0$ that makes the term $2R(\mathcal{F})$ can be bounded by

$$\begin{aligned}
&2CL\sqrt{K} \max_k \mathbb{E}\widetilde{R}_n(\mathcal{F}_k) \log^{\frac{3}{2}+\eta}(\sqrt{n}) \\
&= 4CL\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \max_k \left(\mathbb{E}_{S,\boldsymbol{\sigma}} \sup_{S \in \mathcal{X}^n} \sup_{f_{W,Z_k} \in \mathcal{F}_k} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i f_{W,Z_k} \left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor} \right) \right| \right) \\
&= 4CL\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \max_k \mathbb{E}_S \frac{1}{\lfloor n/2 \rfloor} \left(\sup_{S \in \mathcal{X}^n} \sup_{f_{W,Z_k} \in \mathcal{F}_k} \sum_{i=1}^{\lfloor n/2 \rfloor} f_{W,Z_k} \left(\mathbf{x}_i, \mathbf{x}_{i+\lfloor n/2 \rfloor} \right)^2 \right)^{1/2} \\
&\quad \text{Use Khintchine-Kahane inequality (Latała \& Oleszkiewicz, 1994),} \\
&\leq 4MCL\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \frac{1}{\sqrt{\lfloor n/2 \rfloor}} \quad \text{Use Assumption 1 in the main paper.}
\end{aligned}$$

So based on the above results, $\mathbb{E}[L(\hat{f}^*_{W,Z})] - L^* \leq \frac{8MCL\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n})}{\sqrt{n}}$.

Based on the analysis in Section A and the McDiarmid inequality (Mohri et al., 2018), the term $L(\hat{f}^*_{W,Z}) - L^*$ can be bounded by $\frac{8MCL\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n})}{\sqrt{n}} + E\sqrt{\frac{8 \log \frac{1}{\delta}}{n}}$ with probability $1 - \delta$. □

F. Proof of Theorem 6

Proof. Based on Theorem 5, we can bound the expected local clustering Rademacher complexity in the following formula:

$$\mathbb{E}R_n(\mathcal{F}^r_{exc}) \leq CL\sqrt{K} \max_k \mathbb{E}\widetilde{R}_n(\mathcal{F}^r_{exc,k}) \log^{\frac{3}{2}+\eta}(\sqrt{n}).$$

According to Lemma 6, we have

$$\begin{aligned}
R(\mathcal{F}^r_{exc}) \leq CL\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \max_k \inf_{\epsilon} \Bigg[&2R\left\{ f_{W,Z_k} \in \widetilde{\mathcal{F}_k} : \hat{L}_n(f^2_{W,Z_k}) \leq \epsilon^2 \right\} \\
&+ \frac{64M \log \mathcal{N}(\epsilon/2, \mathcal{F}_k, \|\cdot\|_2)}{n} + \sqrt{\frac{8r \log \mathcal{N}(\epsilon/2, \mathcal{F}_k, \|\cdot\|_2)}{n}} \Bigg],
\end{aligned}$$

where $\widetilde{\mathcal{F}_k} := \left\{ f_{W,Z_k} - f'_{W,Z_k} : f_{W,Z_k}, f'_{W,Z_k} \in \mathcal{F}_k \right\}$ and $\hat{L}_n(f^2_{W,Z_k}) = \frac{1}{n(n-1)} \sum_{i \neq j} f^2_{W,Z_k}(\mathbf{x}_i, \mathbf{x}_j)$.

The following steps are also to use the covering number assumptions to obtain the suitable fixed point r^*.

(1). Assume that there exist three positive constants γ, d and p satisfying $\log \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_2) \leq d \log^p(\gamma/\epsilon)$ for any $0 < \epsilon \leq \gamma$ and $k = 1, ..., K$. Based on the analysis in Section C, for any $0 < r < \gamma^2$, $n \geq \gamma^{-2}$ and $\eta > 0$, it is easy to verify that:

$$R(\mathcal{F}^r_{exc}) \leq c_{\gamma,d,p,M,C} L\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \min \left[\left(\sqrt{\frac{dr \log^p(2\gamma r^{-1/2})}{n}} + \frac{d \log^p(2\gamma r^{-1/2})}{n} \right), \left(\frac{d \log^p(2\gamma n^{1/2})}{n} + \sqrt{\frac{rd \log^p(2\gamma n^{1/2})}{n}} \right) \right].$$

Obviously, we have

$$R(\mathcal{F}^r_{exc}) \leq c_{\gamma,d,p,M,C} L\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \left(\frac{d \log^p(2\gamma n^{1/2})}{n} + \sqrt{\frac{rd \log^p(2\gamma n^{1/2})}{n}} \right).$$

Then, the sub-root function can be set as:

$$\psi(r) := c_{\gamma,d,p,M,C} L\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \left(\frac{d \log^p(2\gamma n^{1/2})}{n} + \sqrt{\frac{rd \log^p(2\gamma n^{1/2})}{n}} \right).$$

Let r^* be its fixed point then we have:

$$r^* = c_{\gamma,d,p,M,C} L\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \left(\frac{d \log^p(2\gamma n^{1/2})}{n} + \sqrt{\frac{r^* d \log^p(2\gamma n^{1/2})}{n}} \right).$$

Solving this equation, we get:

$$r^* \leq c_{M,p,\gamma,d,C} L^2 K \frac{\log^{3+p+2\eta}(n^{1/2})}{n}.$$

By substituting this into Eq. (14), we obtain that: for any $\eta > 0$, with probability $1 - \delta$,

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_{M,p,\gamma,d,C,h,E} L^2 K \frac{\log^{3+p+2\eta}(n^{1/2})}{n} + \frac{c_{h,\delta}}{n-1}.$$

(2). Assume that there exist two constants $\gamma > 0$ and $p > 0$ satisfying $\log \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_2) \leq \gamma \epsilon^{-p}$ for any $k = 1, ..., K$. Based on the analysis in Section C, it is easy to verify that

$$R(\mathcal{F}^r_{exc}) \leq c_{M,p,\gamma,C} L\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \left[n^{-1/2} \epsilon^{1-p/2} + \epsilon^{-p} n^{-1} + \sqrt{r\epsilon^{-p} n^{-1}} \right].$$

So we can set:

$$\psi(r) := c_{M,p,\gamma,C} L\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \left[n^{-1/2} \epsilon^{1-p/2} + \epsilon^{-p} n^{-1} + \sqrt{r\epsilon^{-p} n^{-1}} \right].$$

Let r^* be its fixed point then we have:

$$r^* = c_{M,p,\gamma,C} L\sqrt{K} \log^{\frac{3}{2}+\eta}(\sqrt{n}) \left[n^{-1/2} \epsilon^{1-p/2} + \epsilon^{-p} n^{-1} + \sqrt{r^*\epsilon^{-p} n^{-1}} \right].$$

Solving this equation, it is easy to verify that

$$r^*(\epsilon) \leq c_{M,p,\gamma,C} L^2 K \log^{3+2\eta}(n^{1/2}) [n^{-1/2} \epsilon^{1-p/2} + \epsilon^{-p} n^{-1}].$$

Since $\epsilon > 0$, we can choose $\epsilon = n^{-\frac{1}{2+P}}$, then we obtain

$$r^* = c_{M,p,\gamma,C} L^2 K \log^{3+2\eta}(n^{1/2}) n^{-\frac{2}{p+2}}.$$

By substituting this into Eq. (14), we obtain that: for any $\eta > 0$, with probability $1-\delta$,

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_{M,p,\gamma,C,h,E} L^2 K \log^{3+2\eta}(n^{1/2}) n^{-\frac{2}{p+2}} + \frac{c_{h,\delta}}{n-1}.$$

(3). Assume that there exist two constants $\gamma > 0$ and $p > 0$ satisfying $\log \mathcal{N}(\epsilon, \mathcal{F}_k, \|\cdot\|_2) \leq \gamma \epsilon^{-p} \log^2 \frac{2}{\epsilon}$ for any $k = 1, ..., K$. Based on the analysis in Section C, it is easy to verify that

$$R(\mathcal{F}^r_{exc}) \leq c_{M,p,\gamma,C} L \sqrt{K} \log^{\frac{3}{2}+\eta}(n^{1/2}) \left[n^{-1/2} \epsilon^{1-p/2} \log(\frac{1}{\epsilon}) + \epsilon^{-p} n^{-1} \log^2(\frac{4}{\epsilon}) + \sqrt{r \epsilon^{-p} n^{-1} \log^2(\frac{4}{\epsilon})} \right].$$

So we can set:

$$\psi(r) := c_{M,p,\gamma,C} L \sqrt{K} \log^{\frac{3}{2}+\eta}(n^{1/2}) \left[n^{-1/2} \epsilon^{1-p/2} \log(\frac{1}{\epsilon}) + \epsilon^{-p} n^{-1} \log^2(\frac{4}{\epsilon}) + \sqrt{r \epsilon^{-p} n^{-1} \log^2(\frac{4}{\epsilon})} \right].$$

Let r^* be its fixed point then we have:

$$r^* = c_{M,p,\gamma,C} L \sqrt{K} \log^{\frac{3}{2}+\eta}(n^{1/2}) \left[n^{-1/2} \epsilon^{1-p/2} \log(\frac{1}{\epsilon}) + \epsilon^{-p} n^{-1} \log^2(\frac{4}{\epsilon}) + \sqrt{r^* \epsilon^{-p} n^{-1} \log^2(\frac{4}{\epsilon})} \right].$$

Solving this equation, it is easy to verify that

$$r^*(\epsilon) \leq c_{M,p,\gamma,C} L^2 K \log^{3+2\eta}(\sqrt{n}) \left[n^{-1/2} \epsilon^{1-p/2} \log \frac{1}{\epsilon} + \epsilon^{-p} n^{-1} \log^2 \frac{4}{\epsilon} \right].$$

Since $\epsilon > 0$, we can choose $\epsilon = (\log n)^{\frac{2}{p+2}} n^{-\frac{1}{2+P}}$, then we obtain

$$\begin{aligned} r^* &= c_{M,p,\gamma,C} L^2 K \log^{3+2\eta}(\sqrt{n}) n^{-\frac{2}{p+2}} (\log n)^{\frac{2-p}{p+2}} \log \frac{n}{(\log n)^{\frac{2}{p+2}}} \\ &= c_{M,p,\gamma,C,\eta} L^2 K n^{-\frac{2}{p+2}} (\log n)^{\frac{2-p}{p+2}+3+2\eta} \log \frac{n}{(\log n)^{\frac{2}{p+2}}}. \end{aligned}$$

By substituting this into Eq. (14), we obtain that: for any $\eta > 0$, with probability $1-\delta$,

$$L(\hat{f}^*_{W,Z}) - L^* \leq c_{M,p,\gamma,C,h,E,\eta} L^2 K n^{-\frac{2}{p+2}} (\log n)^{\frac{2-p}{p+2}+3+2\eta} \log \frac{n}{(\log n)^{\frac{2}{p+2}}} + \frac{c_{h,\delta}}{n-1}.$$

□

References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bousquet, O. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and scoring using empirical risk minimization. In *International Conference on Computational Learning Theory (COLT 2005)*, pp. 1–15. Springer, 2005.

Clémençon, S., Lugosi, G., Vayatis, N., et al. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Foster, D. J. and Rakhlin, A. l_∞ vector contraction for rademacher complexity. *arXiv preprint arXiv:1911.06468*, 2019.

Haagerup, U. The best constants in the khintchine inequality. *Studia Mathematica*, 70:231–283, 1981.

Latała, R. and Oleszkiewicz, K. On the best constant in the khinchin-kahane inequality. *Studia Mathematica*, 109(1): 101–104, 1994.

Lei, Y., Ding, L., and Bi, Y. Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218: 320–330, 2016.

Liu, Y., Liao, S., Lin, H., Yue, Y., and Wang, W. Generalization analysis for ranking using integral operator. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017)*, pp. 2273–2279, 2017.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.