Learning with Multiple Complementary Labels: Supplementary Material

A. Proofs about the Problem Setting

A.1. Proofs of Theorem 1

Firstly, we define the set of all the possible label sets whose size is j as

$$\bar{\mathcal{Y}}_j := \{Y \mid Y \in \bar{\mathcal{Y}}, |Y| = j\}.$$

Then, by the definition of $\bar{p}(\boldsymbol{x}, \bar{Y})$, we can obtain

$$\begin{aligned}
\int_{\bar{\mathcal{Y}}}\int_{\mathcal{X}} \bar{p}(\boldsymbol{x}, \bar{Y})\mathrm{d}\boldsymbol{x}\ \mathrm{d}\bar{Y} &= \int \sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}} \bar{p}(\boldsymbol{x}, \bar{Y})\mathrm{d}\boldsymbol{x} \\
&= \int \sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}} \sum\nolimits_{j=1}^{k-1} \Big(\bar{p}(\boldsymbol{x}, \bar{Y} \mid s = j)p(s = j)\Big)\mathrm{d}\boldsymbol{x} \\
&= \int \sum\nolimits_{j=1}^{k-1} \sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j} \Big(\bar{p}(\boldsymbol{x}, \bar{Y} \mid s = j)p(s = j)\Big)\mathrm{d}\boldsymbol{x} \qquad (\because \bar{\mathcal{Y}}_j := \{\bar{Y} \mid \bar{Y} \in \bar{\mathcal{Y}}, |\bar{Y}| = j\}) \\
&= \int \sum\nolimits_{j=1}^{k-1} \sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j} \left(\frac{1}{\binom{k-1}{j}} \sum\nolimits_{y\notin\bar{Y}} p(\boldsymbol{x}, y)p(s = j)\right)\mathrm{d}\boldsymbol{x} \qquad (\because \text{the definition of } \bar{p}(\boldsymbol{x}, \bar{Y} \mid s = j)) \\
&= \int \sum\nolimits_{j=1}^{k-1} \left(\frac{1}{\binom{k-1}{j}} \frac{\binom{k}{j}(k-j)}{k} \sum\nolimits_{y=1}^{k} p(\boldsymbol{x}, y)p(s = j)\right)\mathrm{d}\boldsymbol{x} \qquad (\because |\bar{\mathcal{Y}}_j| = \tbinom{k}{j}) \\
&= \int \sum\nolimits_{j=1}^{k-1} p(\boldsymbol{x})p(s = j)\mathrm{d}\boldsymbol{x} \\
&= 1,
\end{aligned}$$

which concludes the proof of Theorem 1. □

A.2. Proof of Lemma 1

Let us consider the case where the correct label y is a specific label i $(i \in \{1, 2, \cdots, k\})$, then we have

$$\begin{aligned}
p(y \in \bar{Y}, y = i \mid \boldsymbol{x}, s) =& p(y \in \bar{Y} \mid y = i, \boldsymbol{x}, s)p(y = i \mid \boldsymbol{x}, s) \\
&= \sum\nolimits_{C\in\bar{\mathcal{Y}}} p(y \in \bar{Y}, \bar{Y} = C \mid y = i, \boldsymbol{x}, s)p(y = i \mid \boldsymbol{x}, s).
\end{aligned}$$

Here, $p(y = i \mid \boldsymbol{x}, s) = p(y = i \mid \boldsymbol{x})$ since the labeling rule is independent of s. In addition, $\sum_{C\in\bar{\mathcal{Y}}} p(y \in \bar{Y}, \bar{Y} = C \mid y = i, \boldsymbol{x}, s) = \sum_{C\in\bar{\mathcal{Y}}_s} p(y \in \bar{Y}, \bar{Y} = C \mid y = i, \boldsymbol{x})$ since given the size s of the label set, the whole set of all the possible label sets becomes $\bar{\mathcal{Y}}_s$. Then, we can obtain

$$\begin{aligned}
p(y \in \bar{Y}, y = i \mid \boldsymbol{x}, s) &= \sum\nolimits_{C\in\bar{\mathcal{Y}}} p(y \in \bar{Y}, \bar{Y} = C \mid y = i, \boldsymbol{x}, s)p(y = i \mid \boldsymbol{x}, s) \\
&= \sum\nolimits_{C\in\bar{\mathcal{Y}}_s} p(y \in \bar{Y}, \bar{Y} = C \mid y = i, \boldsymbol{x})p(y = i \mid \boldsymbol{x}) \\
&= \sum\nolimits_{C\in\bar{\mathcal{Y}}_s} p(y \in \bar{Y} \mid \bar{Y} = C, y = i, \boldsymbol{x})p(y = i \mid \boldsymbol{x})p(\bar{Y} = C \mid \boldsymbol{x}) \\
&= \sum\nolimits_{C\in\bar{\mathcal{Y}}_s} p(y \in \bar{Y}|\bar{Y} = C, y = i, \boldsymbol{x})p(y = i|\boldsymbol{x})p(\bar{Y} = C),
\end{aligned}$$

where the last equality holds due to the fact that for each instance $\boldsymbol{x}$, $\bar{Y}$ is uniformly and randomly chosen. Since $p(\bar{Y} = C) = \frac{1}{|\bar{\mathcal{Y}}_s|}$ if $C \in \bar{\mathcal{Y}}_s$ where $|\bar{\mathcal{Y}}_s| = \binom{k}{s}$, we have

$$\begin{aligned}
p(y \in \bar{Y}, y = i \mid \boldsymbol{x}, s) =& \textstyle\sum_{C \in \bar{\mathcal{Y}}_s} p(y \in \bar{Y} | \bar{Y} = C, y = i, \boldsymbol{x}) p(y = i | \boldsymbol{x}) p(\bar{Y} = C) \\
=& \frac{1}{\binom{k}{s}} \textstyle\sum_{C \in \bar{\mathcal{Y}}_s} p(y \in \bar{Y} | \bar{Y} = C, y = i, \boldsymbol{x}) p(y = i | \boldsymbol{x}) \\
=& \frac{1}{\binom{k}{s}} |\bar{\mathcal{Y}}_s^i| p(y = i | \boldsymbol{x}) \qquad \big(\because \bar{\mathcal{Y}}_s^i := \{\bar{Y} \in \bar{\mathcal{Y}}_s \mid i \in \bar{Y}\} \big) \\
=& \frac{\binom{k-1}{s-1}}{\binom{k}{s}} p(y = i | \boldsymbol{x}) \qquad \Big(\because |\bar{\mathcal{Y}}_s^i| = \tbinom{k-1}{s-1} \Big) \\
=& \frac{s}{k} p(y = i | \boldsymbol{x}).
\end{aligned}$$

By further summing up the both side over all the possible i, we can obtain

$$p(y \in \bar{Y} | \boldsymbol{x}, s) = \frac{s}{k},$$

which concludes the proof of Lemma 1. □

A.3. Proof of Theorem 2

Let us express $p(\bar{Y} | y \notin \bar{Y}, \boldsymbol{x}, s)$ as

$$\begin{aligned}
p(\bar{Y} | y \notin \bar{Y}, \boldsymbol{x}, s) =& \frac{p(y \notin \bar{Y}, \bar{Y} | \boldsymbol{x}, s)}{p(y \notin \bar{Y} | \boldsymbol{x}, s)} \\
=& \frac{p(y \notin \bar{Y} | \bar{Y}, \boldsymbol{x}, s) p(\bar{Y} | \boldsymbol{x}, s)}{p(y \notin \bar{Y} | \boldsymbol{x}, s)} \\
=& \frac{p(y \notin \bar{Y} | \bar{Y}, \boldsymbol{x}, s) p(\bar{Y} | s)}{p(y \notin \bar{Y} | \boldsymbol{x}, s)},
\end{aligned}$$

where the last equality holds because $\bar{Y}$ is influenced by the size s, and for each instance $\boldsymbol{x}$, $\bar{Y}$ is uniformly and randomly chosen. Note that given s, there are $|\bar{\mathcal{Y}}_s|$ possible label sets, thus $p(\bar{Y} | s) = \frac{1}{|\bar{\mathcal{Y}}_s|}$ where $|\bar{\mathcal{Y}}_s| = \binom{k}{s}$. In this way, we have

$$\begin{aligned}
p(\bar{Y} | y \notin \bar{Y}, \boldsymbol{x}, s) =& \frac{p(y \notin \bar{Y} | \bar{Y}, \boldsymbol{x}, s) p(\bar{Y} | s)}{p(y \notin \bar{Y} | \boldsymbol{x}, s)} \\
=& \frac{1}{\binom{k}{s}} \frac{p(y \notin \bar{Y} | \bar{Y}, \boldsymbol{x}, s)}{1 - p(y \in \bar{Y} | \boldsymbol{x}, s)} \\
=& \frac{1}{\binom{k}{s}} \frac{1}{1 - \frac{s}{k}} p(y \notin \bar{Y} | \bar{Y}, \boldsymbol{x}, s) \qquad \Big(\because \text{by Lemma 1, } p(y \in \bar{Y} | \boldsymbol{x}, s) = \frac{s}{k} \Big) \\
=& \frac{1}{\binom{k}{s}} \frac{k}{k - s} \textstyle\sum_{y \notin \bar{Y}} p(y | \boldsymbol{x}, s) \\
=& \frac{1}{\binom{k-1}{s}} \textstyle\sum_{y \notin \bar{Y}} p(y | \boldsymbol{x}).
\end{aligned}$$

By multiplying $p(\boldsymbol{x})$ on both side, we have

$$p(\boldsymbol{x}, \bar{Y} | y \notin \bar{Y}, s) = \frac{1}{\binom{k-1}{s}} \textstyle\sum_{y \notin \bar{Y}} p(\boldsymbol{x}, y).$$

Then taking into account the variable s, we have

$$\begin{aligned} p(\boldsymbol{x},\bar{Y}|y\notin\bar{Y}) =& \sum\nolimits_{j=1}^{k-1} p(s=j)p(\boldsymbol{x},\bar{Y}|y\notin\bar{Y},s=j) \\ =& \sum\nolimits_{j=1}^{k-1} p(s=j)\frac{1}{\binom{k-1}{j}}\sum\nolimits_{y\notin\bar{Y}} p(\boldsymbol{x},y) \\ =& \bar{p}(\boldsymbol{x},\bar{Y}), \end{aligned}$$

which concludes the proof. □

B. Proofs of the Unbiased Risk Estimator

B.1. Proof of Lemma 2

According to our defined distribution, we have already obtained

$$\bar{p}(\boldsymbol{x},\bar{Y}\mid s=j)=\frac{1}{\binom{k-1}{j}}\sum\nolimits_{y'\notin\bar{Y}}p(\boldsymbol{x},y').$$

Then, we can obtain the following equality by operating $\sum_{\bar{Y}\in\bar{\mathcal{Y}}_j^y}$ on both the left and the right hand side:

$$\sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j^y}\bar{p}(\boldsymbol{x},\bar{Y}\mid s=j)=\frac{1}{\binom{k-1}{j}}\sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j^y}\sum\nolimits_{y'\notin\bar{Y}}p(\boldsymbol{x},y'), \tag{13}$$

where $\bar{\mathcal{Y}}_j^y := \{\bar{Y}\in\bar{\mathcal{Y}}\mid y\in\bar{Y},|\bar{Y}|=j\}$. In this way, the right hand side of the above equality can be transformed by the following derivations:

$$\begin{aligned} \frac{1}{\binom{k-1}{j}}\sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j^y}\sum\nolimits_{y'\notin\bar{Y}}p(\boldsymbol{x},y') =& \frac{1}{\binom{k-1}{j}}\sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j^y}\Big(1-\sum\nolimits_{y'\in\bar{Y}}p(\boldsymbol{x},y')\Big) \\ =& \frac{|\bar{\mathcal{Y}}_j^y|}{\binom{k-1}{j}}-\frac{1}{\binom{k-1}{j}}\sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j^y}\sum\nolimits_{y'\in\bar{Y}}p(\boldsymbol{x},y') \\ =& \frac{\binom{k-1}{j-1}}{\binom{k-1}{j}}-\frac{1}{\binom{k-1}{j}}\sum\nolimits_{y'}\sum\nolimits_{\bar{Y}'\in\{\bar{Y}'\in\bar{\mathcal{Y}}_j^y|y'\in\bar{Y}'\}}p(\boldsymbol{x},y') \qquad \Big(\because |\bar{\mathcal{Y}}_j^y|=\tbinom{k-1}{j-1}\Big) \\ =& \frac{j}{k-j}-\frac{1}{\binom{k-1}{j}}\Big\{\tbinom{k-1}{j-1}p(\boldsymbol{x},y)+\tbinom{k-2}{j-2}\sum\nolimits_{y'\neq y}p(\boldsymbol{x},y')\Big\} \\ =& \frac{j}{k-j}-\frac{1}{\binom{k-1}{j}}\Big\{\tbinom{k-1}{j-1}p(\boldsymbol{x},y)+\tbinom{k-2}{j-2}\big(1-p(\boldsymbol{x},y)\big)\Big\} \\ =& \frac{j}{k-j}-\frac{1}{\binom{k-1}{j}}\Big\{\tbinom{k-2}{j-2}+\tbinom{k-2}{j-1}p(\boldsymbol{x},y)\Big\} \qquad \Big(\because \tbinom{k-2}{j-1}=\tbinom{k-1}{j-1}-\tbinom{k-2}{j-2}\Big) \\ =& \frac{j}{k-j}-\frac{j(j-1)}{(k-j)(k-1)}-\frac{j}{k-1}p(\boldsymbol{x},y) \\ =& \frac{j}{k-1}-\frac{j}{k-1}p(\boldsymbol{x},y). \end{aligned} \tag{14}$$

Combing Eq. (13) and Eq. (14), we obtain

$$p(\boldsymbol{x},y\mid s=j)=p(\boldsymbol{x},y)=1-\frac{k-1}{j}\sum\nolimits_{\bar{Y}\in\bar{\mathcal{Y}}_j^y}\bar{p}(\boldsymbol{x},\bar{Y}\mid s=j). \tag{15}$$

In the end, by taking into account the variable s, we have

$$
\begin{aligned}
p(\boldsymbol{x}, y) &= \sum\nolimits_{j=1}^{k-1} p(s=j) p(\boldsymbol{x}, y \mid s=j) \\
&= \sum\nolimits_{j=1}^{k-1} p(s=j) \Big(1 - \frac{k-1}{j} \sum\nolimits_{\bar{Y} \in \bar{\mathcal{Y}}_j^y} \bar{p}(\boldsymbol{x}, \bar{Y} \mid s=j)\Big) \\
&= 1 - \sum\nolimits_{j=1}^{k-1} \Big(\frac{k-1}{j} \sum\nolimits_{\bar{Y} \in \bar{\mathcal{Y}}_j^y} \bar{p}(\boldsymbol{x}, \bar{Y}, s=j)\Big),
\end{aligned}
$$

which concludes the proof of Lemma 2.

B.2. Proof of Theorem 3

It is intuitive to obtain

$$
R(f) = \mathbb{E}_{p(\boldsymbol{x}, y)}\big[\mathcal{L}\big(f(\boldsymbol{x}), y\big)\big] = \sum\nolimits_{j=1}^{k-1} p(s=j) \mathbb{E}_{p(\boldsymbol{x}, y \mid s=j)}\big[\mathcal{L}\big(f(\boldsymbol{x}), y\big)\big].
$$

Then, we express the right hand side for each $j \in \{1, \ldots, k-1\}$ as

$$
\begin{aligned}
\mathbb{E}_{p(\boldsymbol{x}, y \mid s=j)}\big[\mathcal{L}\big(f(\boldsymbol{x}), y\big)\big] &= \mathbb{E}_{p(\boldsymbol{x} \mid s=j)} \mathbb{E}_{p(y \mid \boldsymbol{x}, s=j)}\big[\mathcal{L}\big(f(\boldsymbol{x}), y\big)\big] \\
&= \mathbb{E}_{p(\boldsymbol{x} \mid s=j)}\Big[\sum\nolimits_{y=1}^{k} p(y \mid \boldsymbol{x}, s=j) \mathcal{L}\big(f(\boldsymbol{x}), y\big)\Big] \\
&= \mathbb{E}_{p(\boldsymbol{x} \mid s=j)}\Big[\sum\nolimits_{y=1}^{k} \Big(1 - \frac{k-1}{j} \sum\nolimits_{\bar{Y} \in \bar{\mathcal{Y}}_j^y} \bar{p}(\bar{Y} \mid \boldsymbol{x}, s=j)\Big) \mathcal{L}\big(f(\boldsymbol{x}), y\big)\Big] \qquad (\because \text{Eq. } (15)) \\
&= \mathbb{E}_{p(\boldsymbol{x} \mid s=j)}\Big[\sum\nolimits_{y=1}^{k} \mathcal{L}\big(f(\boldsymbol{x}), y\big) - \frac{k-1}{j} \sum\nolimits_{y=1}^{k} \sum\nolimits_{\bar{Y} \in \bar{\mathcal{Y}}_j^y} \bar{p}(\bar{Y} \mid \boldsymbol{x}, s=j) \mathcal{L}\big(f(\boldsymbol{x}), y\big)\Big] \\
&= \mathbb{E}_{p(\boldsymbol{x} \mid s=j)}\Big[\sum\nolimits_{y=1}^{k} \mathcal{L}\big(f(\boldsymbol{x}), y\big) - \frac{k-1}{j} \sum\nolimits_{\bar{Y} \in \bar{\mathcal{Y}}_j} \sum\nolimits_{y' \in \bar{Y}} \bar{p}(\bar{Y} \mid \boldsymbol{x}, s=j) \mathcal{L}\big(f(\boldsymbol{x}), y\big)\Big] \\
&= \mathbb{E}_{p(\boldsymbol{x} \mid s=j)}\Big[\sum\nolimits_{y=1}^{k} \mathcal{L}\big(f(\boldsymbol{x}), y\big) - \frac{k-1}{j} \sum\nolimits_{\bar{Y} \in \bar{\mathcal{Y}}_j} \bar{p}(\bar{Y} \mid \boldsymbol{x}, s=j) \Big(\sum\nolimits_{y' \in \bar{Y}} \mathcal{L}\big(f(\boldsymbol{x}), y'\big)\Big)\Big] \\
&= \mathbb{E}_{p(\boldsymbol{x} \mid s=j)} \mathbb{E}_{\bar{p}(\bar{Y} \mid \boldsymbol{x}, s=j)}\Big[\sum\nolimits_{y=1}^{k} \mathcal{L}\big(f(\boldsymbol{x}), y\big) - \frac{k-1}{j} \sum\nolimits_{y' \in \bar{Y}} \mathcal{L}\big(f(\boldsymbol{x}), y'\big)\Big] \\
&= \mathbb{E}_{\bar{p}(\boldsymbol{x}, \bar{Y} \mid s=j)}\Big[\sum\nolimits_{y \notin \bar{Y}} \mathcal{L}\big(f(\boldsymbol{x}), y'\big) - \frac{k-1-j}{j} \sum\nolimits_{y' \in \bar{Y}} \mathcal{L}\big(f(\boldsymbol{x}), y'\big)\Big] \\
&= \mathbb{E}_{\bar{p}(\boldsymbol{x}, \bar{Y} \mid s=j)}[\bar{\mathcal{L}}_j\big(f(\boldsymbol{x}), \bar{Y}\big)] \\
&= \bar{R}_j(f).
\end{aligned}
$$

In this way, we can obtain $R(f) = \sum_{j=1}^{k-1} p(s=j) \bar{R}_j(f)$, which concludes the proof of Theorem 3. □

C. Proof of Theorem 4

Recall that the expected risk and empirical risk are represented as

$$
\begin{aligned}
R(f) &= \sum\nolimits_{j=1}^{k-1} p(s=j) \bar{R}_j(f) = \sum\nolimits_{j=1}^{k-1} p(s=j) \mathbb{E}_{p(\boldsymbol{x}, \bar{Y} \mid s=j)}[\bar{\mathcal{L}}_j\big(f(\boldsymbol{x}), \bar{Y}\big)], \\
\widehat{R}(f) &= \sum\nolimits_{j=1}^{k-1} \frac{p(s=j)}{n_j} \sum\nolimits_{i=1}^{n_j} \bar{\mathcal{L}}_j\big(f(\boldsymbol{x}_i), \bar{Y}_i\big).
\end{aligned}
$$

Here, with a slight abuse of notation, we simply write $\bar{R}_j(f)$ as $R_j(f)$, and define $\widehat{R}_j(f) = 1/n_j \sum_{i=1}^{n_j} \bar{\mathcal{L}}_j\big(f(\boldsymbol{x}_i), \bar{Y}_i\big)$. Thus we have $R(f) = \sum_{j=1}^{k-1} p(s=j) R_j(f)$ and $\widehat{R}(f) = \sum_{j=1}^{k-1} p(s=j) \widehat{R}_j(f)$. Since $f^\star = \arg\min_{f \in \mathcal{F}} R(f)$ and $\widehat{f} = \arg\min_{f \in \mathcal{F}} \widehat{R}(f)$, we can obtain the following lemma.

Lemma 3. *The following inequality holds:*

$$R(\widehat{f}) - R(f^\star) \leqslant 2\sum\nolimits_{j=1}^{k-1} p(s=j) \sup_{f\in\mathcal{F}} \left|\widehat{R}_j(f) - R_j(f)\right|.$$

Proof. It would be intuitive to obtain

$$\begin{aligned} R(\widehat{f}) - R(f^\star) =& R(\widehat{f}) - \widehat{R}(\widehat{f}) + \widehat{R}(\widehat{f}) - R(f^\star) \\ \leqslant& R(\widehat{f}) - \widehat{R}(\widehat{f}) + R(\widehat{f}) - R(f^\star) \\ \leqslant& 2\sup_{f\in\mathcal{F}} \left|\widehat{R}(f) - R(f)\right| \\ =& 2\sup_{f\in\mathcal{F}} \left|\sum\nolimits_{j=1}^{k-1} p(s=j)\widehat{R}_j(f) - \sum\nolimits_{j=1}^{k-1} p(s=j)R_j(f)\right| \\ \leqslant& 2\sum\nolimits_{j=1}^{k-1} p(s=j)\sup_{f\in\mathcal{F}} \left|\widehat{R}_j(f) - R_j(f)\right|, \end{aligned}$$

which concludes the proof of Lemma 3. □

In this way, we will bound $\sup_{f\in\mathcal{F}} \left|\widehat{R}_j(f) - R_j(f)\right|$ for $j = \{1, \ldots, k-1\}$. Before that, we define a function space as

$$\mathcal{H}_j := \{(\boldsymbol{x}, \bar{Y}) \in \mathcal{X} \times \bar{\mathcal{Y}}_j \mapsto \bar{\mathcal{L}}_j\big(f(\boldsymbol{x}), \bar{Y}\big) \mid f \in \mathcal{F}\},$$

where

$$\bar{\mathcal{L}}_j\big(f(\boldsymbol{x}), \bar{Y}\big) := \sum\nolimits_{y\notin\bar{Y}} \mathcal{L}\big(f(\boldsymbol{x}), y\big) - \frac{k-1-j}{j}\sum\nolimits_{y'\in\bar{Y}} \mathcal{L}\big(f(\boldsymbol{x}), y'\big).$$

Besides, we introduce the definition of *Rademacher complexity* (Bartlett & Mendelson, 2002).

Definition 1 (Rademacher complexity (Bartlett & Mendelson, 2002))**.** *Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables drawn from a probability distribution $\mathcal{D}$, $\mathcal{H} = \{h : \mathcal{Z} \to \mathbb{R}\}$ be a class of measurable functions. Then the expected Rademacher complexity of $\mathcal{H}$ is defined as*

$$\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{Z_1,\ldots,Z_n\sim\mathcal{D}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup\nolimits_{h\in\mathcal{H}} \frac{1}{n}\sum\nolimits_{i=1}^{n} \sigma_i h(Z_i)\right],$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ are Rademacher variables taking the value from $\{-1, +1\}$ with even probabilities.

Then, we have the following lemma.

Lemma 4. *Let $C_{\mathcal{L}} = \sup_{\boldsymbol{x}\in\mathcal{X}, f\in\mathcal{F}, y\in\mathcal{Y}} \mathcal{L}\big(f(\boldsymbol{x}), y\big)$. Then, for all $j = \{1, \ldots, k-1\}$, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}} \left|\widehat{R}_j(f) - R_j(f)\right| \leqslant 2\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j) + (2k-2j-1)C_{\mathcal{L}}\sqrt{\frac{\log\frac{2}{\delta}}{2n_j}},$$

where

$$\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j) = \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_j} \frac{1}{n_j}\sum\nolimits_{i=1}^{n_j} \sigma_i h(\boldsymbol{x}_i, \bar{Y}_i)\right].$$

Proof. To prove this lemma, we first show that the single direction $\sup_{f\in\mathcal{F}}\big(\widehat{R}_j(f) - R_j(f)\big)$ is bounded with probability at least $1-\frac{\delta}{2}$, and the other direction can be similarly proved. By the definition of $\bar{\mathcal{L}}_j$, we can easily know the possible maximum of $\bar{\mathcal{L}}_j$ is $(k-j)C_{\mathcal{L}}$, and the possible minimum is $-(k-1-j)C_{\mathcal{L}}$. Suppose an example $(\boldsymbol{x}_i, \bar{Y}_i)$ is replaced by another arbitrary example $(\boldsymbol{x}_i', \bar{Y}_i')$, then the change of $\sup_{f\in\mathcal{F}}\big(\widehat{R}_j(f)-R_j(f)\big)$ is no greater than $((2k-2j-1)C_{\mathcal{L}})/n_j$. Then, by applying *McDiarmid's inequality* (McDiarmid, 1989), for any $\delta>0$, with probability at least $1-\frac{\delta}{2}$,

$$\sup_{f\in\mathcal{F}}\big(\widehat{R}_j(f)-R_j(f)\big)\leqslant\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\big(\widehat{R}_j(f)-R_j(f)\big)\Big]+(2k-2j-1)C_{\mathcal{L}}\sqrt{\frac{\log\frac{2}{\delta}}{2n_j}}. \tag{16}$$

In addition, it is routine (Mohri et al., 2012) to show

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\big(\widehat{R}_j(f)-R_j(f)\big)\Big]\leqslant 2\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j), \tag{17}$$

Combing Eq. (16) and Eq. (17), we have for any $\delta>0$, with probability at least $1-\frac{\delta}{2}$,

$$\sup_{f\in\mathcal{F}}\big(\widehat{R}_j(f)-R_j(f)\big)\leqslant 2\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j)+(2k-2j-1)C_{\mathcal{L}}\sqrt{\frac{\log\frac{2}{\delta}}{2n_j}}.$$

By further taking into account the other side $\sup_{f\in\mathcal{F}}\big(R_j(f)-\widehat{R}_j(f)\big)$, we have for any $\delta>0$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\Big|\widehat{R}_j(f)-R_j(f)\Big|\leqslant 2\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j)+(2k-2j-1)C_{\mathcal{L}}\sqrt{\frac{\log\frac{2}{\delta}}{2n_j}}.$$

which concludes the proof of Lemma 4. □

Next, we will bound the expected Rademacher complexity of the function space $\mathcal{H}_j$, i.e., $\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j)$.

Lemma 5. *Assume the loss function $\mathcal{L}\big(f(\boldsymbol{x}),y\big)$ is ρ-Lipschitz with respect to $f(\boldsymbol{x})$ $(0<\rho<\infty)$ for all $y\in\mathcal{Y}$. Then, for all $j=\{1,\ldots,k-1\}$, the following inequality holds:*

$$\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j)\leqslant\frac{\rho(k-1)}{j}\sum\nolimits_{y=1}^{k}\mathfrak{R}_{n_j}(\mathcal{G}_y),$$

where

$$\begin{aligned}\mathcal{G}_y=&\{g:\boldsymbol{x}\mapsto f_y(\boldsymbol{x})\mid f\in\mathcal{F}\},\\ \mathfrak{R}_{n_j}(\mathcal{G}_y)=&\mathbb{E}_{\boldsymbol{x}_i\sim p(\boldsymbol{x})}\mathbb{E}_{\boldsymbol{\sigma}}\Bigg[\sup_{g\in\mathcal{G}_y}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}g(\boldsymbol{x}_i)\Bigg].\end{aligned}$$

Proof. The expected Rademacher complexity of $\mathcal{H}_j$ can be expressed as

$$\begin{aligned}\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j)=&\mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\Bigg[\sup_{h\in\mathcal{H}_j}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i h(\boldsymbol{x}_i,\bar{Y}_i)\Bigg]\\ =&\mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\Bigg[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\bigg(\sum\nolimits_{y\notin\bar{Y}_i}\mathcal{L}\big(f(\boldsymbol{x}),y\big)-\frac{k-j-1}{j}\sum\nolimits_{y'\in\bar{Y}_i}\mathcal{L}\big(f(\boldsymbol{x}),y'\big)\bigg)\Bigg]\\ \leqslant&\mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\Bigg[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\bigg(\sum\nolimits_{y\notin\bar{Y}_i}\mathcal{L}\big(f(\boldsymbol{x}),y\big)\bigg)\Bigg]\\ &\qquad+\mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\Bigg[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\bigg(\frac{k-j-1}{j}\sum\nolimits_{y'\in\bar{Y}_i}\mathcal{L}\big(f(\boldsymbol{x}),y'\big)\bigg)\Bigg].\end{aligned}$$

Here, we introduce random variables $\alpha_{i,y}=\mathbb{I}[y\in\bar{Y}_i],\ \forall i\in\{1,\cdots,n\},y\in\mathcal{Y}$, where $\mathbb{I}[\cdot]$ denotes the indicator function. In other words, given a complementary label set $\bar{Y}_i$, if a specific label y satisfies the condition $y\in\bar{Y}_i$, then $\mathbb{I}[y\in\bar{Y}_i]=1$, otherwise $\mathbb{I}[y \in \bar{Y}_i] = 0$. Then, we can obtain

$$\begin{aligned}
\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j) \leqslant & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\Big(\sum\nolimits_{y\notin\bar{Y}_i}\mathcal{L}\big(f(\boldsymbol{x}),y\big)\Big)\right] \\
& + \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\Big(\frac{k-j-1}{j}\sum\nolimits_{y'\in\bar{Y}_i}\mathcal{L}\big(f(\boldsymbol{x}),y'\big)\Big)\right] \\
= & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\Big(\sum\nolimits_{y=1}^{k}(1-\alpha_{i,y})\mathcal{L}\big(f(\boldsymbol{x}),y\big)\Big)\right] \\
& + \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\Big(\frac{k-j-1}{j}\sum\nolimits_{y=1}^{k}\alpha_{i,y}\mathcal{L}\big(f(\boldsymbol{x}),y\big)\Big)\right] \\
= & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\Big(\sum\nolimits_{y=1}^{k}\frac{1}{2}(1-2\alpha_{i,y}+1)\mathcal{L}\big(f(\boldsymbol{x}),y\big)\Big)\right] \\
& + \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\Big(\frac{k-j-1}{j}\sum\nolimits_{y=1}^{k}\frac{1}{2}(2\alpha_{i,y}-1+1)\mathcal{L}\big(f(\boldsymbol{x}),y\big)\Big)\right] \\
= & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{2n_j}\sum\nolimits_{i=1}^{n_j}\Big(\sum\nolimits_{y=1}^{k}(1-2\alpha_{i,y})\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y\big)+\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y\big)\Big)\right] \\
+ & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{2n_j}\sum\nolimits_{i=1}^{n_j}\Big(\frac{k-j-1}{j}\sum\nolimits_{y=1}^{k}(2\alpha_{i,y}-1)\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y'\big)+\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y'\big)\Big)\right].
\end{aligned}$$

Here, because $(1-2\alpha_{i,y})\sigma_i$ and $(2\alpha_{i,y}-1)\sigma_i$, and σ_i follow the same distribution, we have

$$\begin{aligned}
\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j) \leqslant & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{2n_j}\sum\nolimits_{i=1}^{n_j}\Big(\sum\nolimits_{y=1}^{k}(1-2\alpha_{i,y})\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y\big)+\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y\big)\Big)\right] \\
+ & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{2n_j}\sum\nolimits_{i=1}^{n_j}\Big(\frac{k-j-1}{j}\sum\nolimits_{y=1}^{k}(2\alpha_{i,y}-1)\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y'\big)+\sigma_i\mathcal{L}\big(f(\boldsymbol{x}),y'\big)\Big)\right] \\
\leqslant & \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sum\nolimits_{y=1}^{k}\sigma_i\mathcal{L}\big(f(\boldsymbol{x}_i),y\big)\right] \\
& + \mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\frac{k-j-1}{j}\sum\nolimits_{y=1}^{k}\sigma_i\mathcal{L}\big(f(\boldsymbol{x}_i),y\big)\right] \\
= & \frac{k-1}{j}\mathbb{E}_{(\boldsymbol{x}_i,\bar{Y}_i)\sim\bar{p}(\boldsymbol{x},\bar{Y}|s=j)}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sum\nolimits_{y=1}^{k}\sigma_i\mathcal{L}\big(f(\boldsymbol{x}_i),y\big)\right] \\
\leqslant & \frac{k-1}{j}\sum\nolimits_{y=1}^{k}\mathbb{E}_{\boldsymbol{x}_i\sim p(\boldsymbol{x})}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\mathcal{L}\big(f(\boldsymbol{x}_i),y\big)\right] \qquad (\because p(\boldsymbol{x}) = \bar{p}(\boldsymbol{x} \mid s=j)),
\end{aligned}$$

Then, we have

$$\begin{aligned}
\bar{\mathfrak{R}}_{n_j}(\mathcal{H}_j) \leqslant & \frac{k-1}{j}\sum\nolimits_{y=1}^{k}\mathbb{E}_{\boldsymbol{x}_i\sim p(\boldsymbol{x})}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n_j}\sum\nolimits_{i=1}^{n_j}\sigma_i\mathcal{L}\big(f(\boldsymbol{x}_i),y\big)\right] \\
\leqslant & \frac{k-1}{j}\sum\nolimits_{y=1}^{k}\mathfrak{R}_{n_j}(\mathcal{L}\circ\mathcal{F}) \\
\leqslant & \frac{\sqrt{2}\rho k(k-1)}{j}\sum\nolimits_{y=1}^{k}\mathfrak{R}_{n_j}(\mathcal{G}_y),
\end{aligned}$$

where we applied the Rademacher vector contraction inequality (Maurer, 2016) in the last inequality. □

Table 5. Statistics of the used benchmark datasets.

Dataset	#Train	#Test	#Features	#Classes	Model
MNIST	60,000	10,000	784	10	Linear Model, MLP (d-500-10)
Fashion-MNIST	60,000	10,000	784	10	Linear Model, MLP (d-500-10)
Kuzushiji-MNIST	60,000	10,000	784	10	Linear Model, MLP (d-500-10)
20Newsgroups	16,961	1,885	1,000	20	Linear Model, MLP (d-500-20)
CIFAR-10	50,000	10,000	3,072	10	ResNet, DenseNet
Yeast	1,335	149	8	10	Linear Model
Texture	4,950	550	40	11	Linear Model
Dermatology	329	37	34	6	Linear Model
Synthetic Control	540	60	60	6	Linear Model

Under the assumptions described in the above three lemmas (Lemma 3, Lemma 4, and Lemma 5), for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(\widehat{f}) - R(f^\star) \leqslant \sum\nolimits_{j=1}^{k-1} p(s=j)\left(\frac{4\sqrt{2}\rho k(k-1)}{j}\sum\nolimits_{y=1}^{k}\mathfrak{R}_{n_j}(\mathcal{G}_y) + (4k-4j-2)C_{\mathcal{L}}\sqrt{\frac{\log\frac{2(k-1)}{\delta}}{2n_j}}\right).$$

It is clear that by combining the above three lemmas, Theorem 4 is proved. □

D. Derivations and Boundness of the Used Loss Functions

D.1. Derivations of the Used Loss Functions

Conventionally, the label for each instance $\boldsymbol{x}$ is in one-hot encoding. Concretely, if the label of $\boldsymbol{x}$ is y, then we represent the label vector as $\boldsymbol{e}_y$ where $e_{yj} = 1$ if $j = y$, otherwise 0. In this way, we provide the detailed derivations of CCE, MAE, and MSE as follows.

- Categorical Cross Entropy (CCE):

$$\mathcal{L}_{\mathrm{CCE}}(f(\boldsymbol{x}), y) = -\sum\nolimits_{j=1}^{k} e_{yj}\log p_{\boldsymbol{\theta}}(j|\boldsymbol{x}) = -\log p_{\boldsymbol{\theta}}(y|\boldsymbol{x}).$$

- Mean Absolute Error (MAE):

$$\mathcal{L}_{\mathrm{MAE}}(f(\boldsymbol{x}), y) = \sum\nolimits_{j=1}^{k} |p_{\boldsymbol{\theta}}(j|\boldsymbol{x}) - e_{yj}| = 2 - 2p_{\boldsymbol{\theta}}(y|\boldsymbol{x}).$$

- Mean Square Error (MSE):

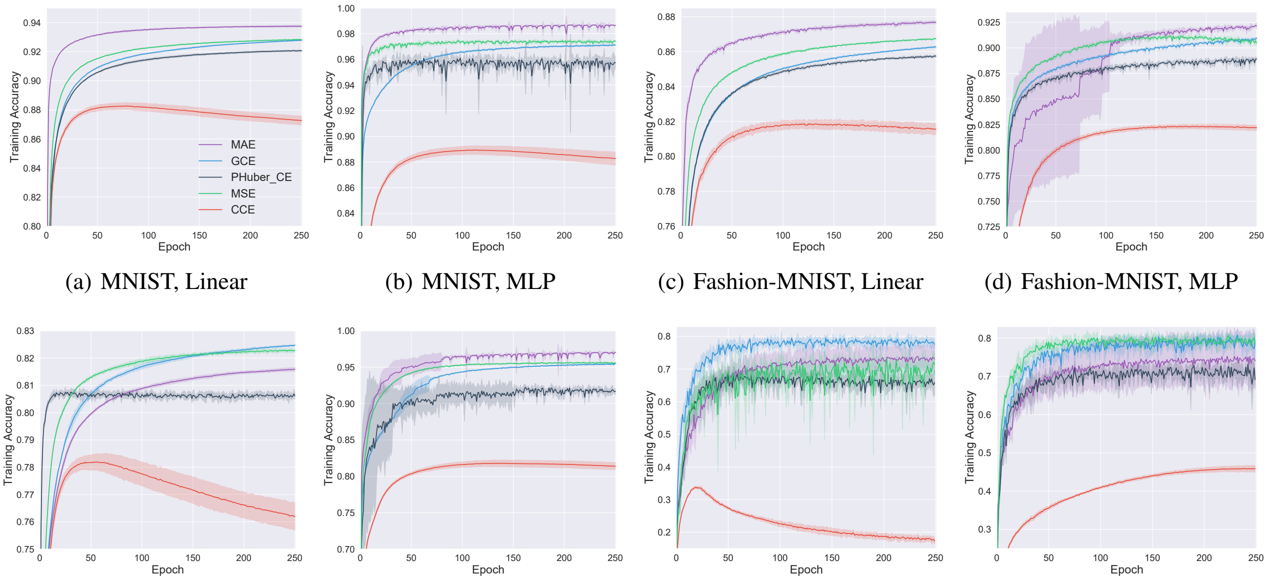
$$\mathcal{L}_{\mathrm{MSE}}(f(\boldsymbol{x}), y) = \sum\nolimits_{j=1}^{k} \big(p_{\boldsymbol{\theta}}(j|\boldsymbol{x}) - e_{yj}\big)^2 = 1 - 2p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) + \sum\nolimits_{j=1}^{k} p_{\boldsymbol{\theta}}(j|\boldsymbol{x})^2.$$

D.2. Boundness of the Used Loss Functions

Firstly, it is clear that each loss function is non-negative. Besides, for each loss function, the loss becomes larger if $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ gets smaller given the correct label y. Note that $0 < p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) < 1$, hence the upper bound of each loss function is stated as follows.

- MAE: $\mathcal{L}_{\mathrm{MAE}}(f(\boldsymbol{x}), y) < 2$.
- MSE: $\mathcal{L}_{\mathrm{MSE}}(f(\boldsymbol{x}), y) < 1 - 0 + \sum_{j=1}^{k} p_{\boldsymbol{\theta}}(j|\boldsymbol{x})^2 < 2$.
- GCE: $\mathcal{L}_{\mathrm{GCE}}(f(\boldsymbol{x}), y) < 1/q$ where $q = 0.7$.
- PHuber-CE: $\mathcal{L}_{\text{PHuber-CE}}(f(\boldsymbol{x}), y) < \log\tau + 1$ where $\tau = 10$.

Note that for CCE, $\mathcal{L}_{\mathrm{CCE}}(f(\boldsymbol{x}), y) < -\log 0 = \infty$. Therefore, we can know that MAE, MSE, GCE, and PHuber-CE are upper-bounded, while CCE is not upper-bounded.



(a) MNIST, Linear

(b) MNIST, MLP

(c) Fashion-MNIST, Linear

(d) Fashion-MNIST, MLP

(e) Kuzushiji-MNIST, Linear

(f) Kuzushiji-MNIST, MLP

(g) CIFAR-10, Resnet

(h) CIFAR-10, Densenet

Figure 2. Experimental results of different loss functions for different datasets and models. Dark colors show the mean accuracy of 5 trials and light colors show the standard deviation.

E. Additional Information of Experiments

E.1. Datasets and Models

In the experiments of Section 5, we use 5 widely-used large-scale benchmark datasets and 4 regular-scale datasets from the UCI Machine Learning Repository. The statistics of these datasets with the corresponding base models are reported in Table 5. Hyper-parameters for all the approaches are selected so as to maximize the accuracy on a validation set, which is constructed by randomly sampling 10% of the training set. We report the characteristics, the parameter settings (to reproduce the experimental results), and the sources of these datasets as follows.

- MNIST (LeCun et al., 1998): It is a 10-class dataset of handwritten digits (0 to 9). Each instance is a 28×28 grayscale image. Source: `http://yann.lecun.com/exdb/mnist/`

- Kuzushiji-MNIST (Clanuwat et al., 2018): It is a 10-class dataset of cursive Japanese ("Kuzushiji") characters. Each instance is a 28×28 grayscale image. Source: `https://github.com/zalandoresearch/fashion-mnist`

- Fashion-MNIST (Xiao et al., 2017): It is a 10-class dataset of fashion items (T-shirt/top, trouser, pullover, dress, sandal, coat, shirt, sneaker, bag, and ankle boot). Each instance is a 28×28 grayscale image. Source: `https://github.com/rois-codh/kmnist`

- CIFAR-10 (Krizhevsky et al., 2009): It is a 10-class dataset of 10 different objects (airplane, bird, automobile, cat, deer, dog, frog, horse, ship, and truck). Each instance is a 32×32×3 colored image in RGB format. This dataset is normalized with mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.247, 0.243, 0.261)$. Source: `https://www.cs.toronto.edu/~kriz/cifar.html`

- 20Newsgroups: It is a 20-class dataset of 20 different newsgroups (comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, misc.forsale, talk.politics.misc, talk.politics.guns, talk.politics.mideast, talk.religion.misc, alt.atheism, soc.religion.christian). We obtained the tf-idf features, and applied TruncatedSVD (Halko et al., 2011) to reduce the dimension to 1000. We randomly sample 90% of the examples from the whole dataset to construct the training set, and the rest 10% forms the test set. Source: `http://qwone.com/~jason/20Newsgroups/`

Table 6. Classification accuracy (%) of each approach on Kuzushiji-MNIST using linear model. The best performance is highlighted in boldface.

Approach		$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$
Upper-bound Losses	EXP	**60.87**	**62.73**	**63.53**	**64.03**	**64.55**	65.06	65.23	65.65
		(±0.38)	(±0.58)	(±0.30)	(±0.38)	(±0.41)	(±0.15)	(±0.10)	(±0.08)
	LOG	60.11	61.57	62.71	63.36	64.01	65.68	**69.35**	**70.10**
		(±0.49)	(±0.15)	(±0.32)	(±0.09)	(±0.13)	(±0.27)	(±0.22)	(±0.18)
Bounded Losses	MAE	60.43	62.71	63.51	63.75	63.94	64.61	64.82	65.10
		(±0.43)	(±0.45)	(±0.10)	(±0.31)	(±0.38)	(±0.19)	(±0.16)	(±0.16)
	MSE	58.97	62.07	63.05	63.85	64.47	64.80	65.17	65.43
		(±0.47)	(±0.54)	(±0.38)	(±0.57)	(±0.43)	(±0.34)	(±0.25)	(±0.10)
	GCE	60.48	62.71	63.13	63.87	63.91	64.28	64.38	64.33%
		(±0.55)	(±0.65)	(±0.30)	(±0.33)	(±0.30)	(±0.07)	(±0.12)	(±0.06)
	Phuber-CE	52.69	56.58	61.10	62.32	64.51	64.93	65.96	65.81
		(±4.22)	(±3.94)	(±2.58)	(±1.50)	(±0.68)	(±0.52)	(±0.37)	(±0.62)
Unbounded Loss	CCE	51.59	55.98	59.15	61.08	63.19	65.05	66.82	68.23
		(±0.64)	(±1.26)	(±1.18)	(±0.78)	(±0.54)	(±0.51)	(±0.41)	(±0.21)
Decomposition before Shuffle	GA	51.72	53.78	54.58	54.78	55.33	55.67	55.91	56.15
		(±1.04)	(±1.07)	(±0.87)	(±0.58)	(±0.29)	(±0.31)	(±0.42)	(±0.23)
	NN	55.03	57.68	58.87	59.52	60.41	60.89	61.41	61.62
		(±1.35)	(±1.29)	(±1.19)	(±0.87)	(±0.59)	(±0.53)	(±0.36)	(±0.09)
	FREE	57.26	60.69	62.77	63.91	64.54	**66.21**	67.00	67.71
		(±0.83)	(±0.96)	(±0.79)	(±0.65)	(±0.55)	(±0.56)	(±0.28)	(±0.20)
	PC	54.31	58.11	60.15	61.32	62.56	63.55	64.27	65.16
		(±1.04)	(±0.87)	(±0.79)	(±0.68)	(±0.59)	(±0.43)	(±0.20)	(±0.18)
	Forward	60.05	61.53	62.43	62.98	63.48	63.95	64.14	64.27
		(±0.43)	(±0.31)	(±0.26)	(±0.40)	(±0.34)	(±0.29)	(±0.09)	(±0.16)
Decomposition after Shuffle	GA	51.72	53.79	54.59	54.83	55.33	55.67	55.90	56.18
		(±1.05)	(±1.07)	(±0.85)	(±0.58)	(±0.35)	(±0.31)	(±0.41)	(±0.22)
	NN	55.03	58.58	60.43	61.58	62.99	64.00	65.07	66.08
		(±1.35)	(±1.11)	(±1.00)	(±0.72)	(±0.49)	(±0.48)	(±0.36)	(±0.10)
	FREE	57.26	60.32	62.11	62.98	64.30	65.18	66.02	67.02
		(±0.84)	(±0.94)	(±0.64)	(±0.67)	(±0.47)	(±0.45)	(±0.28)	(±0.18)
	PC	54.31	57.32	58.95	60.17	61.47	62.54	63.53	64.74
		(±1.04)	(±0.76)	(±0.77)	(±0.83)	(±0.45)	(±0.40)	(±0.22)	(±0.22)
	Forward	60.02	61.75	62.68	63.19	63.59	63.94	64.18	64.32
		(±0.44)	(±0.25)	(±0.23)	(±0.28)	(±0.19)	(±0.09)	(±0.14)	(±0.15)

- Yeast, Texture, Dermatology, Synthetic Control: They are all the datasets from the UCI Machine Learning Repository. Since they are all regular-scale datasets, we only apply linear model on them. For each dataset, we randomly sample 90% of the examples from the whole dataset to construct the training set, and the rest 10% forms the test set. The detailed parameter settings can be found in our provided code package. Source: `https://archive.ics.uci.edu/ml/datasets.php`

For the used models, the detailed information of the used 34-layer ResNet (He et al., 2016) and 22-layer DenseNet (Huang et al., 2017) can be found in the corresponding papers.

E.2. Experimental Results on Training Accuracy

Here, we report the mean and standard deviation of training accuracy (the training set is evaluated with ordinary labels) of 5 trials in Figure 2, to compare the bounded loss functions MAE, MSE, GCE, PHuber-CE, and the unbounded loss function CCE. The training accuracy can reflect the ability of the loss function in identifying the correct label from the non-complementary labels.

From Figure 2, we can find that CCE always achieves the worst performance among all the loss functions, which implies that unbounded loss function is worse than bounded loss function, using our provided empirical risk estimator. This observation clearly supports our conjecture that the negative term in our empirical risk estimator could cause the over-fitting issue. In addition, we can also find that compared with other bounded loss functions, MAE achieves comparable performance in most

Table 7. Classification accuracy (%) of each approach on Kuzushiji-MNIST using MLP. The best performance is highlighted in boldface.

Approach		$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$
Upper-bound Losses	EXP	71.66 (±3.48)	**82.51** (±3.08)	84.45 (±0.24)	87.10 (±0.37)	88.35 (±0.18)	**89.61** (±0.33)	90.18 (±0.37)	90.92 (±0.15)
	LOG	**77.07** (±3.00)	82.39 (±0.73)	**85.54** (±0.35)	**87.60** (±0.40)	**88.87** (±0.34)	89.25 (±0.37)	**90.22** (±0.31)	**91.19** (±0.11)
Bounded Losses	MAE	69.87 (±1.04)	73.60 (±5.77)	79.97 (±3.71)	85.34 (±2.78)	86.91 (±3.06)	89.10 (±0.46)	90.32 (±0.31)	91.06 (±0.34)
	MSE	57.56 (±0.92)	71.37 (±0.89)	78.26 (±0.49)	82.97 (±0.41)	85.37 (±0.45)	86.82 (±0.13)	88.03 (±0.11)	88.69 (±0.05)
	GCE	63.85 (±1.27)	74.11 (±2.38)	79.18 (±2.31)	83.65 (±0.15)	85.23 (±0.25)	86.32 (±0.27)	87.12 (±0.20)	87.64 (±0.09)
	Phuber-CE	10.24 (±4.09)	14.76 (±2.11)	26.60 (±1.58)	73.43 (±1.50)	81.41 (±0.58)	83.00 (±0.42)	84.69 (±0.47)	85.59 (±0.52)
Unbounded Loss	CCE	56.17 (±0.64)	60.89 (±0.61)	64.18 (±0.77)	66.57 (±0.41)	69.14 (±0.49)	71.63 (±0.31)	74.55 (±0.31)	78.22 (±0.22)
Decomposition before Shuffle	GA	70.25 (±0.24)	76.50 (±0.47)	79.77 (±0.32)	82.03 (±0.22)	84.05 (±0.64)	85.58 (±0.32)	86.40 (±0.24)	87.49 (±0.15)
	NN	65.33 (±0.51)	71.34 (±0.53)	75.46 (±0.31)	78.67 (±0.58)	81.40 (±0.28)	84.08 (±0.16)	86.56 (±0.39)	88.61 (±0.12)
	FREE	53.90 (±1.05)	60.32 (±1.14)	63.98 (±0.85)	66.79 (±0.64)	69.31 (±0.73)	71.65 (±0.73)	74.43 (±0.28)	76.61 (±0.33)
	PC	56.36 (±0.56)	62.37 (±0.50)	66.09 (±0.44)	69.51 (±0.47)	72.46 (±0.35)	75.18 (±0.33)	78.50 (±0.52)	82.40 (±0.38)
	Forward	75.40 (±2.02)	83.19 (±0.61)	85.18 (±0.48)	86.63 (±0.38)	87.51 (±0.29)	88.29 (±0.29)	88.96 (±0.26)	89.41 (±0.25)
Decomposition after Shuffle	GA	70.25 (±0.24)	75.91 (±1.37)	78.46 (±2.84)	80.60 (±3.35)	82.14 (±4.51)	83.48 (±4.92)	84.01 (±5.35)	84.65 (±6.28)
	NN	63.73 (±0.97)	67.26 (±0.82)	69.46 (±0.74)	71.25 (±0.62)	73.15 (±0.45)	74.82 (±0.35)	77.09 (±0.17)	79.39 (±0.21)
	FREE	55.33 (±0.89)	60.81 (±0.97)	64.65 (±0.89)	67.01 (±0.70)	69.60 (±0.78)	71.63 (±0.46)	74.22 (±0.40)	77.16 (±0.50)
	PC	56.68 (±1.28)	61.07 (±0.99)	63.86 (±0.67)	65.61 (±0.44)	68.03 (±0.64)	69.74 (±0.65)	72.49 (±0.37)	75.17 (±0.46)
	Forward	66.09 (±0.49)	73.20 (±3.05)	75.76 (±2.61)	82.53 (±2.60)	86.27 (±0.65)	88.05 (±0.27)	89.24 (±0.22)	90.22 (±0.20)

cases, while it is sometimes inferior to other bounded losses due to its optimization issue (Zhang & Sabuncu, 2018). All the above observations on the training accuracy (Figure 2) are very similar to those observations on the test accuracy (Figure 1 in our paper).

E.3. Experimental Results on Fixed Complementary Label Set

We also conduct additional experiments to investigate the influence of the variable s on Kuzushiji-MNIST using both linear model and MLP. Specifically, we study the case where the size of each complementary label set s is fixed at j (i.e., $p(s = j) = 1$) while increasing j from 1 to $k - 2$. The detailed experimental results are shown in Table 6 and Table 7. From the two tables, we can find that the (test) classification accuracy of our approaches increases as j increases. This observation is clearly in accordance with our derived estimation error bound (Theorem 4), as the estimation error would decrease if j increases. In addition, as shown in the two tables, our proposed upper-bound losses outperform other approaches in most cases. This observation also demonstrates the effectiveness of our proposed upper-bound losses.

References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(11): 463–482, 2002.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, pp. 4700–4708, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *ALT*, pp. 3–17, 2016.

McDiarmid, C. On the method of bounded differences. In *Surveys in Combinatorics*, 1989.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pp. 8778–8788, 2018.