Appendix of "Binary Classification with Confidence Difference"

Anonymous Author(s)
Affiliation
Address
`email`

A Proof of Theorem 1

Before giving the proof of Theorem 1, we begin with the following lemmas:

Lemma 2. *The confidence difference $c(\boldsymbol{x}, \boldsymbol{x}')$ can be equivalently expressed as*

$$\begin{aligned} c(\boldsymbol{x}, \boldsymbol{x}') &= \frac{\pi_+ p(\boldsymbol{x}) p_+(\boldsymbol{x}') - \pi_+ p_+(\boldsymbol{x}) p(\boldsymbol{x}')}{p(\boldsymbol{x}) p(\boldsymbol{x}')} && (1) \\ &= \frac{\pi_- p_-(\boldsymbol{x}) p(\boldsymbol{x}') - \pi_- p(\boldsymbol{x}) p_-(\boldsymbol{x}')}{p(\boldsymbol{x}) p(\boldsymbol{x}')} && (2) \end{aligned}$$

Proof. On one hand,

$$\begin{aligned} c(\boldsymbol{x}, \boldsymbol{x}') &= p(y' = 1|\boldsymbol{x}') - p(y = 1|\boldsymbol{x}) \\ &= \frac{p(\boldsymbol{x}', y' = 1)}{p(\boldsymbol{x}')} - \frac{p(\boldsymbol{x}, y = 1)}{p(\boldsymbol{x})} \\ &= \frac{\pi_+ p_+(\boldsymbol{x}')}{p(\boldsymbol{x}')} - \frac{\pi_+ p_+(\boldsymbol{x})}{p(\boldsymbol{x})} \\ &= \frac{\pi_+ p(\boldsymbol{x}) p_+(\boldsymbol{x}') - \pi_+ p_+(\boldsymbol{x}) p(\boldsymbol{x}')}{p(\boldsymbol{x}) p(\boldsymbol{x}')}. \end{aligned}$$

On the other hand,

$$\begin{aligned} c(\boldsymbol{x}, \boldsymbol{x}') &= p(y' = 1|\boldsymbol{x}') - p(y = 1|\boldsymbol{x}) \\ &= (1 - p(y' = 0|\boldsymbol{x}')) - (1 - p(y = 0|\boldsymbol{x})) \\ &= p(y = 0|\boldsymbol{x}) - p(y' = 0|\boldsymbol{x}') \\ &= \frac{p(\boldsymbol{x}, y = 0)}{p(\boldsymbol{x})} - \frac{p(\boldsymbol{x}', y = 0)}{p(\boldsymbol{x}')} \\ &= \frac{\pi_- p_-(\boldsymbol{x})}{p(\boldsymbol{x})} - \frac{\pi_- p_-(\boldsymbol{x}')}{p(\boldsymbol{x}')} \\ &= \frac{\pi_- p_-(\boldsymbol{x}) p(\boldsymbol{x}') - \pi_- p(\boldsymbol{x}) p_-(\boldsymbol{x}')}{p(\boldsymbol{x}) p(\boldsymbol{x}')}, \end{aligned}$$

which concludes the proof. □

Lemma 3. *The following equations hold:*

$$\mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{x}')}[(\pi_+ - c(\boldsymbol{x}, \boldsymbol{x}'))\ell(g(\boldsymbol{x}), +1)] = \pi_+ \mathbb{E}_{p_+(\boldsymbol{x})}[\ell(g(\boldsymbol{x}), +1)], \quad (3)$$

$$\mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{x}')}[(\pi_- + c(\boldsymbol{x}, \boldsymbol{x}'))\ell(g(\boldsymbol{x}), -1)] = \pi_- \mathbb{E}_{p_-(\boldsymbol{x})}[\ell(g(\boldsymbol{x}), -1)], \quad (4)$$

$$\mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{x}')}[(\pi_+ + c(\boldsymbol{x}, \boldsymbol{x}'))\ell(g(\boldsymbol{x}'), +1)] = \pi_+ \mathbb{E}_{p_+(\boldsymbol{x}')}[\ell(g(\boldsymbol{x}'), +1)], \quad (5)$$

$$\mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{x}')}[(\pi_- - c(\boldsymbol{x}, \boldsymbol{x}'))\ell(g(\boldsymbol{x}'), -1)] = \pi_- \mathbb{E}_{p_-(\boldsymbol{x}')}[\ell(g(\boldsymbol{x}'), -1)]. \quad (6)$$

Proof. Firstly, the proof of Eq. (3) is given:

$$
\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\pi_+ - c(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),+1)]\\
=&\int\int \frac{\pi_+ p(\boldsymbol{x})p(\boldsymbol{x}') - \pi_+ p(\boldsymbol{x})p_+(\boldsymbol{x}') + \pi_+ p_+(\boldsymbol{x})p(\boldsymbol{x}')}{p(\boldsymbol{x})p(\boldsymbol{x}')}\ell(g(\boldsymbol{x}),+1)p(\boldsymbol{x},\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}'\\
=&\int\int (\pi_+ p(\boldsymbol{x})p(\boldsymbol{x}') - \pi_+ p(\boldsymbol{x})p_+(\boldsymbol{x}') + \pi_+ p_+(\boldsymbol{x})p(\boldsymbol{x}'))\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}'\\
=&\int \pi_+ p(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x}\int p(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}' - \int \pi_+ p(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x}\int p_+(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'\\
&+\int \pi_+ p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x}\int p(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'\\
=&\int \pi_+ p(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x} - \int \pi_+ p(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x} + \int \pi_+ p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x}\\
=&\int \pi_+ p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)\,\mathrm{d}\boldsymbol{x}\\
=&\pi_+\mathbb{E}_{p_+(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),+1)].
\end{aligned}
$$

After that, the proof of Eq. (4) is given:

$$
\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\pi_- + c(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),-1)]\\
=&\int\int \frac{\pi_- p(\boldsymbol{x})p(\boldsymbol{x}') + \pi_- p_-(\boldsymbol{x})p(\boldsymbol{x}') - \pi_- p(\boldsymbol{x})p_-(\boldsymbol{x}')}{p(\boldsymbol{x})p(\boldsymbol{x}')}\ell(g(\boldsymbol{x}),-1)p(\boldsymbol{x},\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}'\\
=&\int\int (\pi_- p(\boldsymbol{x})p(\boldsymbol{x}') + \pi_- p_-(\boldsymbol{x})p(\boldsymbol{x}') - \pi_- p(\boldsymbol{x})p_-(\boldsymbol{x}'))\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}'\\
=&\int \pi_- p(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x}\int p(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}' + \int \pi_- p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x}\int p(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'\\
&-\int \pi_- p(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x}\int p_-(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'\\
=&\int \pi_- p(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x} + \int \pi_- p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x} - \int \pi_- p(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x}\\
=&\int \pi_- p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)\,\mathrm{d}\boldsymbol{x}\\
=&\pi_-\mathbb{E}_{p_-(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),-1)].
\end{aligned}
$$

It can be noticed that $c(\boldsymbol{x},\boldsymbol{x}') = -c(\boldsymbol{x}',\boldsymbol{x})$ and $p(\boldsymbol{x},\boldsymbol{x}') = p(\boldsymbol{x}',\boldsymbol{x})$. Therefore, it can be deduced naturally that $\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\pi_+ - c(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),+1)] = \mathbb{E}_{p(\boldsymbol{x}',\boldsymbol{x})}[(\pi_+ + c(\boldsymbol{x}',\boldsymbol{x}))\ell(g(\boldsymbol{x}),+1)]$. Because $\boldsymbol{x}$ and $\boldsymbol{x}'$ are symmetric, we can swap them and deduce Eq. (5). Eq. (6) can be deduced in the same manner, which concludes the proof. □

Based on Lemma 3, the proof of Theorem 1 is given.

Proof of Theorem 1. To begin with, it can be noticed that $\mathbb{E}_{p_+(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),+1)] = \mathbb{E}_{p_+(\boldsymbol{x}')}[\ell(g(\boldsymbol{x}'),+1)]$ and $\mathbb{E}_{p_-(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),-1)] = \mathbb{E}_{p_-(\boldsymbol{x}')}[\ell(g(\boldsymbol{x}'),-1)]$. Then, by summing up all the equations from Eq. (3) to Eq. (6), we can get the following equation:

$$
\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[\mathcal{L}_+(g(\boldsymbol{x}),g(\boldsymbol{x}')) + \mathcal{L}_-(g(\boldsymbol{x}),g(\boldsymbol{x}'))]\\
&= 2\pi_+\mathbb{E}_{p_+(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),+1)] + 2\pi_-\mathbb{E}_{p_-(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),-1)]
\end{aligned}
$$

After dividing each side of the equation above by 2, we can obtain Theorem 1. □

B Analysis on Variance of Risk Estimator

B.1 Proof of Lemma 1 in the Main Paper

Based on Lemma 3, it can be observed that

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[\mathcal{L}(\boldsymbol{x},\boldsymbol{x}')] =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\pi_+ - c(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),+1) + (\pi_- - c(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}'),-1)] \\ =& \pi_+\mathbb{E}_{p_+(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),+1)] + \pi_-\mathbb{E}_{p_-(\boldsymbol{x}')}[\ell(g(\boldsymbol{x}'),-1)] \\ =& \pi_+\mathbb{E}_{p_+(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),+1)] + \pi_-\mathbb{E}_{p_-(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),-1)] \\ =& R(g)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[\mathcal{L}(\boldsymbol{x}',\boldsymbol{x})] =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\pi_+ + c(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}'),+1) + (\pi_- + c(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),-1)] \\ =& \pi_-\mathbb{E}_{p_-(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),-1)] + \pi_+\mathbb{E}_{p_+(\boldsymbol{x}')}[\ell(g(\boldsymbol{x}'),+1)] \\ =& \pi_-\mathbb{E}_{p_-(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),-1)] + \pi_+\mathbb{E}_{p_+(\boldsymbol{x})}[\ell(g(\boldsymbol{x}),+1)] \\ =& R(g).\end{aligned}$$

Therefore, for an arbitrary weight $\alpha \in [0,1]$,

$$\begin{aligned}R(g) =& \alpha R(g) + (1-\alpha)R(g) \\ =& \alpha\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[\mathcal{L}(\boldsymbol{x},\boldsymbol{x}')] + (1-\alpha)\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[\mathcal{L}(\boldsymbol{x}',\boldsymbol{x})],\end{aligned}$$

which indicates that

$$\frac{1}{n}\sum_{i=1}^{n}(\alpha\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i') + (1-\alpha)\mathcal{L}(\boldsymbol{x}_i',\boldsymbol{x}_i))$$

is also an unbiased risk estimator and concludes the proof. □

B.2 Proof of Theorem 2

In this subsection, we show that Eq. (8) in the main paper achieves the minimum variance of

$$S(g;\alpha) = \frac{1}{n}\sum_{i=1}^{n}(\alpha\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i') + (1-\alpha)\mathcal{L}(\boldsymbol{x}_i',\boldsymbol{x}_i))$$

w.r.t. any $\alpha \in [0,1]$. To begin with, we introduce the following notations:

$$\begin{aligned}\mu_1 &\triangleq \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i'))^2] = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i',\boldsymbol{x}_i))^2], \\ \mu_2 &\triangleq \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[\frac{1}{n^2}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i')\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i',\boldsymbol{x}_i)].\end{aligned} \tag{7}$$

Furthermore, according to Lemma 1 in the main paper, we have

$$\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[S(g;\alpha)] = R(g).$$

Then, we provide the proof of Theorem 2 as follows.

Proof of Theorem 2.

$$\begin{aligned}\mathrm{Var}(S(g;\alpha)) =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(S(g;\alpha) - R(g))^2] \\ =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[S(g;\alpha)^2] - R(g)^2 \\ =& \alpha^2\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i'))^2] + (1-\alpha)^2\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i',\boldsymbol{x}_i))^2] \\ &+ 2\alpha(1-\alpha)\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[\frac{1}{n^2}\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i')\sum_{i=1}^{n}\mathcal{L}(\boldsymbol{x}_i',\boldsymbol{x}_i)] - R(g)^2 \\ =& \mu_1\alpha^2 + \mu_1(1-\alpha)^2 + 2\mu_2\alpha(1-\alpha) - R(g)^2 \\ =& (2\mu_1 - 2\mu_2)(\alpha - \frac{1}{2})^2 + \frac{1}{2}(\mu_1 + \mu_2) - R(g)^2.\end{aligned}$$

Besides, it can be observed that

$$2\mu_1 - 2\mu_2 = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}')}[(\frac{1}{n}\sum_{i=1}^{n}(\mathcal{L}(\boldsymbol{x}_i,\boldsymbol{x}_i') - \mathcal{L}(\boldsymbol{x}_i',\boldsymbol{x}_i))^2] \geq 0.$$

Therefore, $\mathrm{Var}(S(g;\alpha))$ achieves the minimum value when $\alpha = 1/2$, which concludes the proof. $\square$

C Proof of Theorem 3

To begin with, we give the definition of Rademacher complexity.

Definition 1 (Rademacher complexity)**.** *Let $\mathcal{X}_n = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_n\}$ denote n i.i.d. random variables drawn from a probability distribution with density $p(\boldsymbol{x})$, $\mathcal{G} = \{g : \mathcal{X} \mapsto \mathbb{R}\}$ denote a class of measurable functions, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_n)$ denote Rademacher variables taking values from $\{+1, -1\}$ uniformly. Then, the (expected) Rademacher complexity of $\mathcal{G}$ is defined as*

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{\mathcal{X}_n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(\boldsymbol{x}_i)\right]. \tag{8}$$

Let $\mathcal{D}_n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x},\boldsymbol{x}')$ denote n pairs of ConfDiff data and $\mathcal{L}_{\mathrm{CD}}(g;\boldsymbol{x}_i,\boldsymbol{x}_i') = (\mathcal{L}(\boldsymbol{x},\boldsymbol{x}') + \mathcal{L}(\boldsymbol{x}',\boldsymbol{x}))/2$, then we introduce the following lemma.

Lemma 4.

$$\bar{\mathfrak{R}}_n(\mathcal{L}_{\mathrm{CD}} \circ \mathcal{G}) \leq 2L_\ell \mathfrak{R}_n(\mathcal{G}),$$

where $\mathcal{L}_{\mathrm{CD}} \circ \mathcal{G} = \{\mathcal{L}_{\mathrm{CD}} \circ g | g \in \mathcal{G}\}$ and $\bar{\mathfrak{R}}_n(\cdot)$ is the Rademacher complexity over ConfDiff data pairs $\mathcal{D}_n$ of size n.

Proof.

$$\begin{aligned}
\bar{\mathfrak{R}}_n(\mathcal{L}_{\mathrm{CD}} \circ \mathcal{G}) =& \mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{\boldsymbol{\sigma}}[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\mathcal{L}_{\mathrm{CD}}(g;\boldsymbol{x}_i,\boldsymbol{x}_i')] \\
=& \mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{\boldsymbol{\sigma}}[\sup_{g\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i((\pi_+ - c_i)\ell(g(\boldsymbol{x}_i),+1) + (\pi_- - c_i)\ell(g(\boldsymbol{x}_i'),-1) \\
&+ (\pi_+ + c_i)\ell(g(\boldsymbol{x}_i'),+1) + (\pi_- + c_i)\ell(g(\boldsymbol{x}_i),-1))].
\end{aligned}$$

Then, we can induce that

$$\begin{aligned}
&\|\nabla\mathcal{L}_{\mathrm{CD}}(g;\boldsymbol{x}_i,\boldsymbol{x}_i')\|_2 \\
=&\|\nabla(\frac{(\pi_+ - c_i)\ell(g(\boldsymbol{x}_i),+1) + (\pi_- - c_i)\ell(g(\boldsymbol{x}_i'),-1)}{2} \\
&+ \frac{(\pi_+ + c_i)\ell(g(\boldsymbol{x}_i'),+1) + (\pi_- + c_i)\ell(g(\boldsymbol{x}_i),-1)}{2})\|_2 \\
\leq&\|\nabla(\frac{(\pi_+ - c_i)\ell(g(\boldsymbol{x}_i),+1)}{2})\|_2 + \|\nabla(\frac{(\pi_- - c_i)\ell(g(\boldsymbol{x}_i'),-1)}{2})\|_2 \\
&+ \|\nabla(\frac{(\pi_+ + c_i)\ell(g(\boldsymbol{x}_i'),+1)}{2})\|_2 + \|\nabla(\frac{(\pi_- + c_i)\ell(g(\boldsymbol{x}_i),-1)}{2})\|_2 \\
\leq&\frac{|\pi_+ - c_i|L_\ell}{2} + \frac{|\pi_- - c_i|L_\ell}{2} + \frac{|\pi_+ + c_i|L_\ell}{2} + \frac{|\pi_- + c_i|L_\ell}{2}.
\end{aligned} \tag{9}$$

Suppose $\pi_+ \geq \pi_-$, the value of RHS of Eq. (9) can be determined as follows: when $c_i \in [-1, -\pi_+)$, the value is $-2c_iL_\ell$; when $c_i \in [-\pi_+, -\pi_-)$, the value is $(\pi_+ - c_i)L_\ell$; when $c_i \in [-\pi_-, \pi_-)$, the value is L_ℓ; when $c_i \in [\pi_-, \pi_+)$, the value is $(\pi_+ + c_i)L_\ell$; when $c_i \in [\pi_+, 1]$, the value is $2c_iL_\ell$. To sum up, when $\pi_+ \geq \pi_-$, the value of RHS of Eq. (9) is less than $2L_\ell$. When $\pi_+ \leq \pi_-$, we can deduce that the value of RHS of Eq. (9) is less than $2L_\ell$ in the same way. Therefore,

$$\begin{aligned}\bar{\mathfrak{R}}_n(\mathcal{L}_{\mathrm{CD}} \circ \mathcal{G}) \leq & 2L_\ell \mathbb{E}_{\mathcal{D}_n} \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(\boldsymbol{x}_i)] \\ = & 2L_\ell \mathbb{E}_{\mathcal{X}_n} \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(\boldsymbol{x}_i)] \\ = & 2L_\ell \mathfrak{R}_n(\mathcal{G}),\end{aligned}$$

which concludes the proof. □

After that, we introduce the following lemma.

Lemma 5. *The inequality below hold with probability at least* $1-\delta$*:*

$$\sup_{g\in\mathcal{G}} |R(g) - \widehat{R}_{\mathrm{CD}}(g)| \leq 4L_\ell\mathfrak{R}_n(\mathcal{G}) + 2C_\ell\sqrt{\frac{\ln 2/\delta}{2n}}.$$

Proof. To begin with, we introduce $\Phi = \sup_{g\in\mathcal{G}}(R(g) - \widehat{R}_{\mathrm{CD}}(g))$ and $\bar{\Phi} = \sup_{g\in\mathcal{G}}(R(g) - \widehat{\bar{R}}_{\mathrm{CD}}(g))$, where $\widehat{R}_{\mathrm{CD}}(g)$ and $\widehat{\bar{R}}_{\mathrm{CD}}(g)$ denote the empirical risk over two sets of training examples with exactly one different point $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i), c_i\}$ and $\{(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}'_i), c(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}'_i)\}$ respectively. Then we have

$$\begin{aligned}\bar{\Phi} - \Phi &\leq \sup_{g\in\mathcal{G}}(\widehat{R}_{\mathrm{CD}}(g) - \widehat{\bar{R}}_{\mathrm{CD}}(g)) \\ &\leq \sup_{g\in\mathcal{G}}(\frac{\mathcal{L}_{\mathrm{CD}}(g;\boldsymbol{x}_i,\boldsymbol{x}'_i) - \mathcal{L}_{\mathrm{CD}}(g;\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{x}}'_i)}{n}) \\ &\leq \frac{2C_\ell}{n}.\end{aligned}$$

Accordingly, $\Phi - \bar{\Phi}$ can be bounded in the same way. The following inequalities holds with probability at least $1-\delta/2$ by applying McDiarmid's inequality:

$$\sup_{g\in\mathcal{G}}(R(g) - \widehat{R}_{\mathrm{CD}}(g)) \leq \mathbb{E}_{\mathcal{D}_n}[\sup_{g\in\mathcal{G}}(R(g) - \widehat{R}_{\mathrm{CD}}(g))] + 2C_\ell\sqrt{\frac{\ln 2/\delta}{2n}},$$

Furthermore, we can bound $\mathbb{E}_{\mathcal{D}_n}[\sup_{g\in\mathcal{G}}(R(g) - \widehat{R}_{\mathrm{CD}}(g))]$ with Rademacher complexity. It is a routine work to show by symmetrization [16] that

$$\mathbb{E}_{\mathcal{D}_n}[\sup_{g\in\mathcal{G}}(R(g) - \widehat{R}_{\mathrm{CD}}(g))] \leq 2\bar{\mathfrak{R}}_n(\mathcal{L}_{\mathrm{CD}} \circ \mathcal{G}) \leq 4L_\ell\mathfrak{R}_n(\mathcal{G}),$$

where the second inequality is from Lemma 4. Accordingly, $\sup_{g\in\mathcal{G}}(\widehat{R}_{\mathrm{CD}}(g) - R(g))$ has the same bound. By using the union bound, the following inequality holds with probability at least $1-\delta$:

$$\sup_{g\in\mathcal{G}} |R(g) - \widehat{R}_{\mathrm{CD}}(g)| \leq 4L_\ell\mathfrak{R}_n(\mathcal{G}) + 2C_\ell\sqrt{\frac{\ln 2/\delta}{2n}},$$

which concludes the proof. □

Finally, the proof of Theorem 3 is provided.

Proof of Theorem 3.

$$\begin{aligned}R(\widehat{g}_{\mathrm{CD}}) - R(g^*) &= (R(\widehat{g}_{\mathrm{CD}}) - \widehat{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}})) + (\widehat{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}}) - \widehat{R}_{\mathrm{CD}}(g^*)) + (\widehat{R}_{\mathrm{CD}}(g^*) - R(g^*)) \\ &\leq (R(\widehat{g}_{\mathrm{CD}}) - \widehat{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}})) + (\widehat{R}_{\mathrm{CD}}(g^*) - R(g^*)) \\ &\leq |R(\widehat{g}_{\mathrm{CD}}) - \widehat{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}})| + \left|\widehat{R}_{\mathrm{CD}}(g^*) - R(g^*)\right| \\ &\leq 2\sup_{g\in\mathcal{G}} |R(g) - \widehat{R}_{\mathrm{CD}}(g)| \\ &\leq 8L_\ell\mathfrak{R}_n(\mathcal{G}) + 4C_\ell\sqrt{\frac{\ln 2/\delta}{2n}}.\end{aligned}$$

The first inequality is derived because $\widehat{g}_{\mathrm{CD}}$ is the minimizer of $\widehat{R}_{\mathrm{CD}}(g)$. The last inequality is derived according to Lemma 5, which concludes the proof. □

D Proof of Theorem 4

To begin with, we provide the following inequality:

$$\begin{aligned}
&\sup_{g\in\mathcal{G}}|\bar{R}_{\mathrm{CD}}(g)-\widehat{R}_{\mathrm{CD}}(g)|\\
=&\frac{1}{2n}|\sum_{i=1}^{n}((\bar{\pi}_+-\pi_++c_i-\bar{c}_i)\ell(g(\boldsymbol{x}_i),+1)+(\bar{\pi}_--\pi_-+c_i-\bar{c}_i)\ell(g(\boldsymbol{x}_i'),-1)\\
&+(\bar{\pi}_+-\pi_++\bar{c}_i-c_i)\ell(g(\boldsymbol{x}_i'),+1)+(\bar{\pi}_--\pi_-+\bar{c}_i-c_i)\ell(g(\boldsymbol{x}_i),-1))|\\
\leq&\frac{1}{2n}\sum_{i=1}^{n}(|(\bar{\pi}_+-\pi_++c_i-\bar{c}_i)\ell(g(\boldsymbol{x}_i),+1)|+|(\bar{\pi}_--\pi_-+c_i-\bar{c}_i)\ell(g(\boldsymbol{x}_i'),-1)|\\
&+|(\bar{\pi}_+-\pi_++\bar{c}_i-c_i)\ell(g(\boldsymbol{x}_i'),+1)|+|(\bar{\pi}_--\pi_-+\bar{c}_i-c_i)\ell(g(\boldsymbol{x}_i),-1)|)\\
=&\frac{1}{2n}\sum_{i=1}^{n}(|\bar{\pi}_+-\pi_++c_i-\bar{c}_i|\ell(g(\boldsymbol{x}_i),+1)+|\bar{\pi}_--\pi_-+c_i-\bar{c}_i|\ell(g(\boldsymbol{x}_i'),-1)\\
&+|\bar{\pi}_+-\pi_++\bar{c}_i-c_i|\ell(g(\boldsymbol{x}_i'),+1)+|\bar{\pi}_--\pi_-+\bar{c}_i-c_i|\ell(g(\boldsymbol{x}_i),-1))\\
\leq&\frac{1}{2n}\sum_{i=1}^{n}((|\bar{\pi}_+-\pi_+|+|c_i-\bar{c}_i|)\ell(g(\boldsymbol{x}_i),+1)+(|\bar{\pi}_--\pi_-|+|c_i-\bar{c}_i|)\ell(g(\boldsymbol{x}_i'),-1)\\
&+(|\bar{\pi}_+-\pi_+|+|\bar{c}_i-c_i|)\ell(g(\boldsymbol{x}_i'),+1)+(|\bar{\pi}_--\pi_-|+|\bar{c}_i-c_i|)\ell(g(\boldsymbol{x}_i),-1))\\
=&\frac{1}{2n}\sum_{i=1}^{n}((|\bar{\pi}_+-\pi_+|+|c_i-\bar{c}_i|)\ell(g(\boldsymbol{x}_i),+1)+(|\pi_+-\bar{\pi}_+|+|c_i-\bar{c}_i|)\ell(g(\boldsymbol{x}_i'),-1)\\
&+(|\bar{\pi}_+-\pi_+|+|\bar{c}_i-c_i|)\ell(g(\boldsymbol{x}_i'),+1)+(|\pi_+-\bar{\pi}_+|+|\bar{c}_i-c_i|)\ell(g(\boldsymbol{x}_i),-1))\\
\leq&\frac{2C_\ell\sum_{i=1}^{n}|\bar{c}_i-c_i|}{n}+2C_\ell|\bar{\pi}_+-\pi_+|.
\end{aligned}$$

Then, we deduce the following inequality:

$$\begin{aligned}
R(\bar{g}_{\mathrm{CD}})-R(g^*)=&(R(\bar{g}_{\mathrm{CD}})-\widehat{R}_{\mathrm{CD}}(\bar{g}_{\mathrm{CD}}))+(\widehat{R}_{\mathrm{CD}}(\bar{g}_{\mathrm{CD}})-\bar{R}_{\mathrm{CD}}(\bar{g}_{\mathrm{CD}}))+(\bar{R}_{\mathrm{CD}}(\bar{g}_{\mathrm{CD}})-\bar{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}}))\\
&+(\bar{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}})-\widehat{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}}))+(\widehat{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}})-R(\widehat{g}_{\mathrm{CD}}))+(R(\widehat{g}_{\mathrm{CD}})-R(g^*))\\
\leq&2\sup_{g\in\mathcal{G}}|R(g)-\widehat{R}_{\mathrm{CD}}(g)|+2\sup_{g\in\mathcal{G}}|\bar{R}_{\mathrm{CD}}(g)-\widehat{R}_{\mathrm{CD}}(g)|+(R(\widehat{g}_{\mathrm{CD}})-R(g^*))\\
\leq&4\sup_{g\in\mathcal{G}}|R(g)-\widehat{R}_{\mathrm{CD}}(g)|+2\sup_{g\in\mathcal{G}}|\bar{R}_{\mathrm{CD}}(g)-\widehat{R}_{\mathrm{CD}}(g)|\\
\leq&16L_\ell\mathfrak{R}_n(\mathcal{G})+8C_\ell\sqrt{\frac{\ln 2/\delta}{2n}}+\frac{4C_\ell\sum_{i=1}^{n}|\bar{c}_i-c_i|}{n}+4C_\ell|\bar{\pi}_+-\pi_+|.
\end{aligned}$$

The first inequality is derived because $\bar{g}_{\mathrm{CD}}$ is the minimizer of $\bar{R}(g)$. The second and third inequality are derived according to the proof of Theorem 3 and Lemma 5 respectively. □

E Proof of Theorem 5

To begin with, let $\mathfrak{D}_n^+(g)=\{\mathcal{D}_n|\widehat{A}(g)\geq 0\cap\widehat{B}(g)\geq 0\cap\widehat{C}(g)\geq 0\cap\widehat{D}(g)\geq 0\}$ and $\mathfrak{D}_n^-(g)=\{\mathcal{D}_n|\widehat{A}(g)\leq 0\cup\widehat{B}(g)\leq 0\cup\widehat{C}(g)\leq 0\cup\widehat{D}(g)\leq 0\}$. Before giving the proof of Theorem 5, we give the following lemma based on the assumptions in Section 3.

Lemma 6. *The probability measure of $\mathfrak{D}_n^-(g)$ can be bounded as follows:*

$$\mathbb{P}(\mathfrak{D}_n^-(g))\leq\exp\big(\frac{-2a^2n}{C_\ell^2}\big)+\exp\big(\frac{-2b^2n}{C_\ell^2}\big)+\exp\big(\frac{-2c^2n}{C_\ell^2}\big)+\exp\big(\frac{-2d^2n}{C_\ell^2}\big).\qquad(10)$$

Proof. It can be observed that

$$\begin{aligned}p(\mathcal{D}_n) &= p(\boldsymbol{x}_1, \boldsymbol{x}_1') \cdots p(\boldsymbol{x}_n, \boldsymbol{x}_n') \\ &= p(\boldsymbol{x}_1) \cdots p(\boldsymbol{x}_n') p(\boldsymbol{x}_1) \cdots p(\boldsymbol{x}_n').\end{aligned}$$

Therefore, the probability measure $\mathbb{P}(\mathfrak{D}_n^-(g))$ can be defined as follows:

$$\begin{aligned}\mathbb{P}(\mathfrak{D}_n^-(g)) &= \int_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} p(\mathcal{D}_n)\, \mathrm{d}\mathcal{D}_n \\ &= \int_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} p(\mathcal{D}_n)\, \mathrm{d}\boldsymbol{x}_1 \cdots \mathrm{d}\boldsymbol{x}_n\, \mathrm{d}\boldsymbol{x}_1' \cdots \mathrm{d}\boldsymbol{x}_n'.\end{aligned}$$

When exactly one ConfDiff data pair in S_n is replaced, the change of $\widehat{A}(g), \widehat{B}(g), \widehat{C}(g)$ and $\widehat{D}(g)$ will be no more than C_ℓ/n. By applying McDiarmid's inequality, we can obtain the following inequalities:

$$\begin{aligned}\mathbb{P}(\mathbb{E}[\widehat{A}(g)] - \widehat{A}(g) \geq a) &\leq \exp\big(\frac{-2a^2 n}{C_\ell^2}\big), \\ \mathbb{P}(\mathbb{E}[\widehat{B}(g)] - \widehat{B}(g) \geq b) &\leq \exp\big(\frac{-2b^2 n}{C_\ell^2}\big), \\ \mathbb{P}(\mathbb{E}[\widehat{C}(g)] - \widehat{C}(g) \geq c) &\leq \exp\big(\frac{-2c^2 n}{C_\ell^2}\big), \\ \mathbb{P}(\mathbb{E}[\widehat{D}(g)] - \widehat{D}(g) \geq d) &\leq \exp\big(\frac{-2d^2 n}{C_\ell^2}\big).\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{P}(\mathfrak{D}_n^-(g)) \leq& \mathbb{P}(\widehat{A}(g) \leq 0) + \mathbb{P}(\widehat{B}(g) \leq 0) + \mathbb{P}(\widehat{C}(g) \leq 0) + \mathbb{P}(\widehat{D}(g) \leq 0) \\ \leq& \mathbb{P}(\widehat{A}(g) \leq \mathbb{E}[\widehat{A}(g)] - a) + \mathbb{P}(\widehat{B}(g) \leq \mathbb{E}[\widehat{B}(g)] - b) \\ &+ \mathbb{P}(\widehat{C}(g) \leq \mathbb{E}[\widehat{C}(g)] - c) + \mathbb{P}(\widehat{D}(g) \leq \mathbb{E}[\widehat{D}(g)] - d) \\ =& \mathbb{P}(\mathbb{E}[\widehat{A}(g)] - \widehat{A}(g) \geq a) + \mathbb{P}(\mathbb{E}[\widehat{B}(g)] - \widehat{B}(g) \geq b) \\ &+ \mathbb{P}(\mathbb{E}[\widehat{C}(g)] - \widehat{C}(g) \geq c) + \mathbb{P}(\mathbb{E}[\widehat{D}(g)] - \widehat{D}(g) \geq d) \\ \leq& \exp\big(\frac{-2a^2 n}{C_\ell^2}\big) + \exp\big(\frac{-2b^2 n}{C_\ell^2}\big) + \exp\big(\frac{-2c^2 n}{C_\ell^2}\big) + \exp\big(\frac{-2d^2 n}{C_\ell^2}\big),\end{aligned}$$

which concludes the proof. $\square$

Then, the proof of Theorem 5 is given.

Proof of Theorem 5. To begin with, we prove the first inequality in Theorem 5.

$$\begin{aligned}&\mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] - R(g) \\ =& \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g) - \widehat{R}_{\mathrm{CD}}(g)] \\ =& \int_{\mathcal{D}_n \in \mathfrak{D}_n^+(g)} (\widetilde{R}_{\mathrm{CD}}(g) - \widehat{R}_{\mathrm{CD}}(g)) p(\mathcal{D}_n)\, \mathrm{d}\mathcal{D}_n \\ &+ \int_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} (\widetilde{R}_{\mathrm{CD}}(g) - \widehat{R}_{\mathrm{CD}}(g)) p(\mathcal{D}_n)\, \mathrm{d}\mathcal{D}_n \\ =& \int_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} (\widetilde{R}_{\mathrm{CD}}(g) - \widehat{R}_{\mathrm{CD}}(g)) p(\mathcal{D}_n)\, \mathrm{d}\mathcal{D}_n \geq 0,\end{aligned}$$

where the last inequality is derived because $\widetilde{R}_{\mathrm{CD}}(g)$ is an upper bound of $\widehat{R}_{\mathrm{CD}}(g)$. Furthermore,

$$
\begin{aligned}
&\mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] - R(g) \\
&= \int_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} (\widetilde{R}_{\mathrm{CD}}(g) - \widehat{R}_{\mathrm{CD}}(g)) p(\mathcal{D}_n) \,\mathrm{d}\mathcal{D}_n \\
&\leq \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} (\widetilde{R}_{\mathrm{CD}}(g) - \widehat{R}_{\mathrm{CD}}(g)) \int_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} p(\mathcal{D}_n) \,\mathrm{d}\mathcal{D}_n \\
&= \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} (\widetilde{R}_{\mathrm{CD}}(g) - \widehat{R}_{\mathrm{CD}}(g)) \mathbb{P}(\mathfrak{D}_n^-(g)) \\
&= \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} (f(\widehat{A}(g)) + f(\widehat{B}(g)) + f(\widehat{C}(g)) + f(\widehat{D}(g)) \\
&\quad - \widehat{A}(g) - \widehat{B}(g) - \widehat{C}(g) - \widehat{D}(g)) \mathbb{P}(\mathfrak{D}_n^-(g)) \\
&\leq \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} (L_f|\widehat{A}(g)| + L_f|\widehat{B}(g)| + L_f|\widehat{C}(g)| + L_f|\widehat{D}(g)| \\
&\quad + |\widehat{A}(g)| + |\widehat{B}(g)| + |\widehat{C}(g)| + |\widehat{D}(g)|) \mathbb{P}(\mathfrak{D}_n^-(g) \\
&= \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} \frac{L_f+1}{2n} (|\sum_{i=1}^n (\pi_+ - c_i)\ell(g(\boldsymbol{x}_i), +1)| + |\sum_{i=1}^n (\pi_- - c_i)\ell(g(\boldsymbol{x}_i'), -1)| \\
&\quad + |\sum_{i=1}^n (\pi_+ + c_i)\ell(g(\boldsymbol{x}_i'), +1)| + |\sum_{i=1}^n (\pi_- + c_i)\ell(g(\boldsymbol{x}_i), -1)|) \mathbb{P}(\mathfrak{D}_n^-(g)) \\
&\leq \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} \frac{L_f+1}{2n} (\sum_{i=1}^n |(\pi_+ - c_i)\ell(g(\boldsymbol{x}_i), +1)| + \sum_{i=1}^n |(\pi_- - c_i)\ell(g(\boldsymbol{x}_i'), -1)| \\
&\quad + \sum_{i=1}^n |(\pi_+ + c_i)\ell(g(\boldsymbol{x}_i'), +1)| + \sum_{i=1}^n |(\pi_- + c_i)\ell(g(\boldsymbol{x}_i), -1)|) \mathbb{P}(\mathfrak{D}_n^-(g)) \\
&= \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} \frac{L_f+1}{2n} \sum_{i=1}^n (|(\pi_+ - c_i)\ell(g(\boldsymbol{x}_i), +1)| + |(\pi_- - c_i)\ell(g(\boldsymbol{x}_i'), -1)| \\
&\quad + |(\pi_+ + c_i)\ell(g(\boldsymbol{x}_i'), +1)| + |(\pi_- + c_i)\ell(g(\boldsymbol{x}_i), -1)|) \mathbb{P}(\mathfrak{D}_n^-(g)) \\
&\leq \sup_{\mathcal{D}_n \in \mathfrak{D}_n^-(g)} \frac{(L_f+1)C_\ell}{2n} \sum_{i=1}^n (|\pi_+ - c_i| + |\pi_- - c_i| + |\pi_+ + c_i| + |\pi_- + c_i|) \mathbb{P}(\mathfrak{D}_n^-(g)).
\end{aligned}
$$

Similar to the proof of Theorem 3, we can obtain

$$|\pi_+ - c_i| + |\pi_- - c_i| + |\pi_+ + c_i| + |\pi_- + c_i| \leq 4.$$

Therefore, we have

$$\mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] - R(g) \leq 2(L_f+1)C_\ell\Delta,$$

which concludes the proof of the first inequality in Theorem 5. Before giving the proof of the second inequality, we give the upper bound of $|\widetilde{R}_{\mathrm{CD}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)]|$. When exactly one ConfDiff data pair in $\mathcal{D}_n$ is replaced, the change of $\widetilde{R}_{\mathrm{CD}}(g)$ is no more than $2C_\ell L_f/n$. By applying McDiarmid's inequality, we have the following inequalities with probability at least $1-\delta/2$:

$$
\begin{aligned}
\widetilde{R}_{\mathrm{CD}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] &\leq 2C_\ell L_f \sqrt{\frac{\ln 2/\delta}{2n}}, \\
\mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] - \widetilde{R}_{\mathrm{CD}}(g) &\leq 2C_\ell L_f \sqrt{\frac{\ln 2/\delta}{2n}}.
\end{aligned}
$$

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

$$|\widetilde{R}_{\mathrm{CD}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)]| \leq 2C_\ell L_f \sqrt{\frac{\ln 2/\delta}{2n}}.$$

Table 1: Characteristics of experimental data sets.

Data Set	# Train	# Test	# Features	# Class Labels	Model
MNIST	60,000	10,000	784	10	MLP
Kuzushiji	60,000	10,000	784	10	MLP
Fashion	60,000	10,000	784	10	MLP
CIFAR-10	50,000	10,000	3,072	10	ResNet-34
Optdigits	4,495	1,125	62	10	MLP
USPS	7,437	1,861	256	10	MLP
Pendigits	8,793	2,199	16	10	MLP
Letter	16,000	4,000	16	26	MLP

Finally, we have

$$\begin{aligned}|\widetilde{R}_{\mathrm{CD}}(g) - R(g)| &= |\widetilde{R}_{\mathrm{CD}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] + \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] - R(g)| \\ &\leq |\widetilde{R}_{\mathrm{CD}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)]| + |\mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] - R(g)| \\ &= |\widetilde{R}_{\mathrm{CD}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)]| + \mathbb{E}[\widetilde{R}_{\mathrm{CD}}(g)] - R(g) \\ &\leq 2C_\ell L_f\sqrt{\frac{\ln 2/\delta}{2n}} + 2(L_f+1)C_\ell\Delta, \end{aligned} \tag{11}$$

with probability at least $1-\delta$, which concludes the proof. □

F Proof of Theorem 6

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

$$\begin{aligned}R(\widetilde{g}_{\mathrm{CD}}) - R(g^*) =&(R(\widetilde{g}_{\mathrm{CD}}) - \widetilde{R}_{\mathrm{CD}}(\widetilde{g}_{\mathrm{CD}})) + (\widetilde{R}_{\mathrm{CD}}(\widetilde{g}_{\mathrm{CD}}) - \widetilde{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}})) \\ &+ (\widetilde{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}}) - R(\widehat{g}_{\mathrm{CD}})) + (R(\widehat{g}_{\mathrm{CD}}) - R(g^*)) \\ \leq&|R(\widetilde{g}_{\mathrm{CD}}) - \widetilde{R}_{\mathrm{CD}}(\widetilde{g}_{\mathrm{CD}})| + |\widetilde{R}_{\mathrm{CD}}(\widehat{g}_{\mathrm{CD}}) - R(\widehat{g}_{\mathrm{CD}})| + (R(\widehat{g}_{\mathrm{CD}}) - R(g^*)) \\ \leq&4C_\ell(L_f+1)\sqrt{\frac{\ln 2/\delta}{2n}} + 4(L_f+1)C_\ell\Delta + 8L_\ell\mathfrak{R}_n(\mathcal{G}).\end{aligned}$$

The first inequality is derived because $\widetilde{g}_{\mathrm{CD}}$ is the minimizer of $\widetilde{R}_{\mathrm{CD}}(g)$. The second inequality is derived from Theorem 5 and Theorem 3. The proof is completed. □

G Related Work

Learning with pairwise comparisons has been investigated pervasively in the community [1, 2, 10, 11, 17, 19, 23], with applications in information retrieval [15], computer vision [6], regression [21, 22], crowdsourcing [3, 24], and graph learning [9]. It is noteworthy that there exist distinct differences between our work and previous works on learning with pairwise comparisons. Previous works have mainly tried to learn a ranking function that can rank candidate examples according to relevance or preference. In this paper, we try to learn a *pointwise binary classifier* by conducting empirical risk minimization under the binary classification setting.

H Limitations and Potential Negative Social Impacts

H.1 Limitations

This work focuses on binary classification problems. To generalize it to multi-class problems, we need to convert multi-class classification to a set of binary classification problems via the one-versus-rest or the one-versus-one strategies. In the future, developing methods directly handling multi-class classification problems is promising.

H.2 Potential Negative Social Impacts

This work is within the scope of weakly supervised learning, which aims to achieve comparable performance while reducing labeling costs. Therefore, when this technique is very effective and prevalent in society, the demand for data annotations may be reduced, leading to the increasing unemployment rate of data annotation workers.

I Additional Information about Experiments

In this section, the details of experimental data sets and hyperparameters are provided.

I.1 Details of Experimental Data Sets

The detailed statistics and corresponding model architectures are summarized in Table 1 while the basic information, sources and data split details are elaborated in this subsection.

For the four benchmark data sets,

- MNIST [14]: It is a grayscale handwritten digits recognition data set. It is composed of 60,000 training examples and 10,000 test examples. The original feature dimension is 28*28, and the label space is 0-9. The even digits are regarded as the positive class while the odd digits are regarded as the negative class. We sampled 15,000 unlabeled data pairs as training data. The data set can be downloaded from `http://yann.lecun.com/exdb/mnist/`.
- Kuzushiji-MNIST [4]: It is a grayscale Japanese character recognition data set. It is composed of 60,000 training examples and 10,000 test examples. The original feature dimension is 28*28, and the label space is {'o', 'su','na', 'ma', 're', 'ki','tsu','ha', 'ya','wo'}. The positive class is composed of 'o', 'su','na', 'ma', and 're' while the negative class is composed of 'ki','tsu','ha', 'ya', and 'wo'. We sampled 15,000 unlabeled data pairs as training data. The data set can be downloaded from `https://github.com/rois-codh/kmnist`.
- Fashion-MNIST [20]: It is a grayscale fashion item recognition data set. It is composed of 60,000 training examples and 10,000 test examples. The original feature dimension is 28*28, and the label space is {'T-shirt', 'trouser', 'pullover', 'dress', 'sandal', 'coat', 'shirt', 'sneaker', 'bag', 'ankle boot'}. The positive class is composed of 'T-shirt', 'pullover', 'coat', 'shirt', and 'bag' while the negative class is composed of 'trouser', 'dress', 'sandal', 'sneaker', and 'ankle boot'. We sampled 15,000 unlabeled data pairs as training data. The data set can be downloaded from `https://github.com/zalandoresearch/fashion-mnist`.
- CIFAR-10 [13]: It is a colorful object recognition data set. It is composed of 50,000 training examples and 10,000 test examples. The original feature dimension is 32*32*3, and the label space is {'airplane', 'bird', 'automobile', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', 'truck'}. The positive class is composed of 'bird', 'deer', 'dog', 'frog', 'cat', and 'horse' while the negative class is composed of 'airplane', 'automobile', 'ship', and 'truck'. We sampled 10,000 unlabeled data pairs as training data. The data set can be downloaded from `https://www.cs.toronto.edu/~kriz/cifar.html`.

For the four UCI data sets, they can be downloaded from [5].

- Optdigits, USPS, Pendigits [5]: They are handwritten digit recognition data set. The train-test split can be found in Table 1. The feature dimensions are 62, 256, and 16 respectively and the label space is 0-9. The even digits are regarded as the positive class while the odd digits are regarded as the negative class. We sampled 1,200, 2,000, and 2,500 unlabeled data pairs for training respectively.
- Letter [5]: It is a letter recognition data set. It is composed of 16,000 training examples and 4,000 test examples. The feature dimension is 16 and the label space is the 26 capital letters in the English alphabet. The positive class is composed of the top 13 letters while the negative class is composed of the latter 13 letters. We sampled 4,000 unlabeled data pairs for training.

I.2 Details of Experiments on the KuaiRec Data Set

We used the small matrix of the KuaiRec [7] data set since it has dense confidence scores. It has 1,411 users and 3,327 items. We clipped the watching ratio above 2 and regarded the examples with watching ratio greater than 2 as positive examples. Following the experimental protocol of [8], we regarded the latest positive example foe each user as the positive testing data, and sampled 49 negative testing data to form the testing set for each user. The HR and NDCG were calculated at top 10. The learning rate was set to 1e-3 and the dropout rate was set to 0.5. The number of epochs was set to 50 and the batch size was set to 256. The number of MLP layers was 2 and the embedding dimension was 128. The hyperparameters was the same for all the approaches for a fair comparison.

I.3 Details of Hyperparameters

All the experiments were conducted on NVIDIA GeForce RTX 3090. The number of training epoches was set to 200 and we obtained the testing accuracy by averaging the results in the last 10 epoches. All the methods were implemented in Pytorch [18]. We used the Adam optimizer [12]. To ensure fair comparisons, We set the same hyperparameter values for all the compared approaches.

For MNIST, Kuzushiji-MNIST and Fashion-MNIST, the learning rate was set to 1e-3 and the weight decay was set to 1e-5. The batch size was set to 256 data pairs. For training the probabilistic classifier to generate confidence, the batch size was set to 256 and the epoch number was set to 10.

For CIFAR10, the learning rate was set to 5e-4 and the weight decay was set to 1e-5. The batch size was set to 128 data pairs. For training the probabilistic classifier to generate confidence, the batch size was set to 128 and the epoch number was set to 10.

For all the UCI data sets, the learning rate was set to 1e-3 and the weight decay was set to 1e-5. The batch size was set to 128 data pairs. For training the probabilistic classifier to generate confidence, the batch size was set to 128 and the epoch number was set to 10.

The learning rate and weight decay for training the probabilistic classifier were the same as the setting for each data set correspondingly.

J More Experimental Results with Fewer Training Data

Figure 1 shows extra experimental results with fewer training data on other data sets with different class priors.

References

[1] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96, 2005.

[2] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136, 2007.

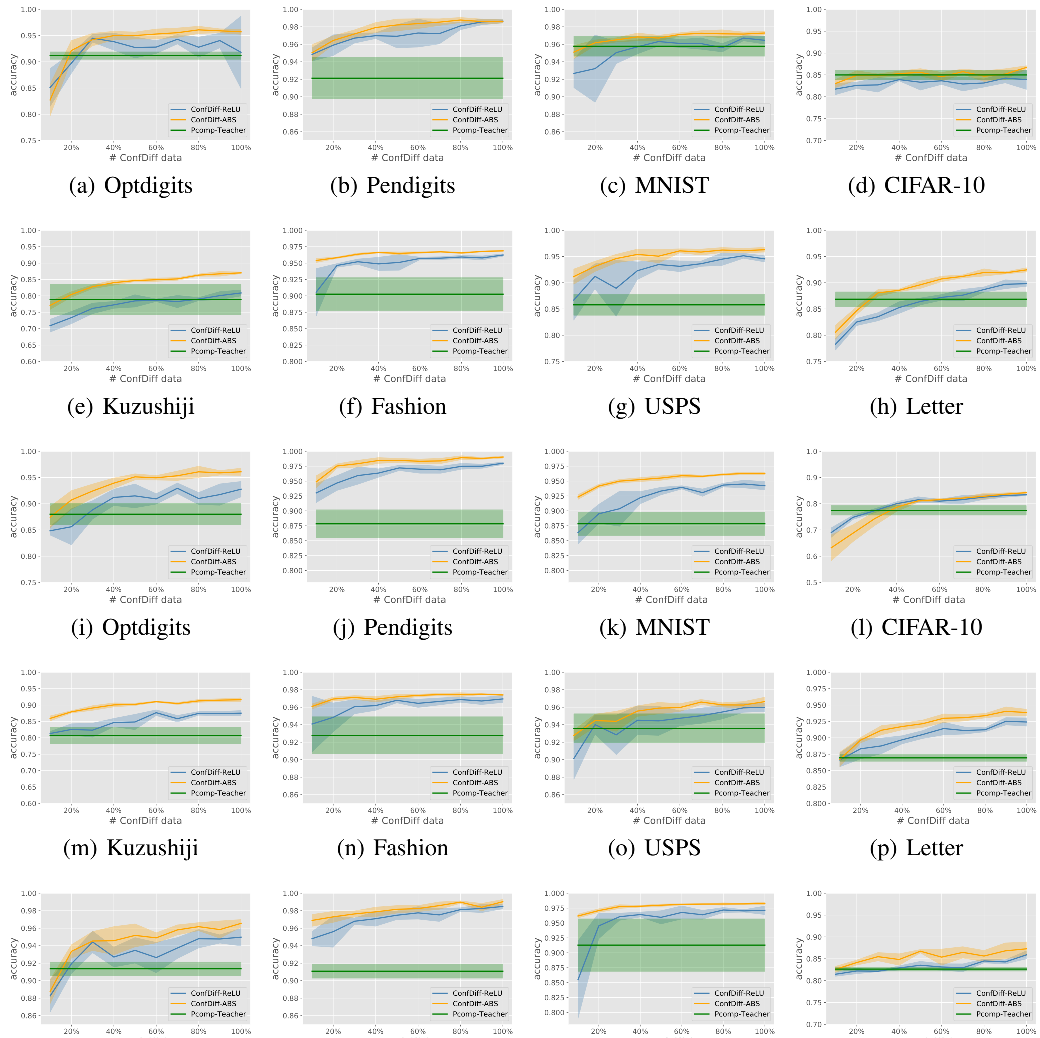
[3] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 193–202, 2013.

[4] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature. *CoRR*, abs/1812.01718, 2018.

[5] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[6] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):563–577, 2015.

[7] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 540–550, 2022.



(a) Optdigits (b) Pendigits (c) MNIST (d) CIFAR-10

(e) Kuzushiji (f) Fashion (g) USPS (h) Letter

(i) Optdigits (j) Pendigits (k) MNIST (l) CIFAR-10

(m) Kuzushiji (n) Fashion (o) USPS (p) Letter

(q) Optdigits (r) Pendigits (s) MNIST (t) CIFAR-10

Figure 1: Classification performance of ConfDiff-ReLU and ConfDiff-ABS given a fraction of training data as well as Pcomp-Teacher given 100% of training data with different prior settings ($\pi_+ = 0.2$ for the first row, $\pi_+ = 0.5$ for the second and the third row, and $\pi_+ = 0.8$ for the fourth and the fifth row).

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182, 2017.

[9] Yixuan He, Quan Gan, David Wipf, Gesine D. Reinert, Junchi Yan, and Mihai Cucuringu. GNNRank: Learning global rankings from pairwise comparisons via directed graph neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8581–8612, 2022.

[10] Kevin G. Jamieson and Robert D. Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems 24*, pages 2240–2248, 2011.

[11] Daniel M. Kane, Shachar Lovett, Shay Moran, and Jiapeng Zhang. Active classification with comparison queries. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*, pages 355–366, 2017.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[13] Alex Krizhevsky and Geoffrey E. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[16] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[17] Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit S. Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1907–1916, 2015.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037, 2019.

[19] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. *IEEE Transactions on Information Theory*, 65(8):4854–4874, 2019.

[20] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

[21] Liyuan Xu, Junya Honda, Gang Niu, and Masashi Sugiyama. Uncoupled regression from pairwise comparison data. In *Advances in Neural Information Processing Systems 32*, pages 3992–4002, 2019.

[22] Yichong Xu, Sivaraman Balakrishnan, Aarti Singh, and Artur Dubrawski. Regression with comparisons: Escaping the curse of dimensionality with ordinal information. *Journal of Machine Learning Research*, 21(1):6480–6533, 2020.

[23] Yichong Xu, Hongyang Zhang, Kyle Miller, Aarti Singh, and Artur Dubrawski. Noise-tolerant interactive learning using pairwise comparisons. In *Advances in Neural Information Processing Systems 30*, pages 2428–2437, 2017.

[24] Shiwei Zeng and Jie Shen. Efficient pac learning from the crowd with pairwise comparisons. In *Proceedings of the 39th International Conference on Machine Learning*, pages 25973–25993, 2022.