A. Proof of Lemma 1

Proof.

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

□

B. Proof of Theorem 1

We give a technical lemma before proving Theorem 1:

Lemma 3.

$$p_S(\boldsymbol{x},\boldsymbol{x}') = \frac{\pi_+^2 p_+(\boldsymbol{x})p_+(\boldsymbol{x}')+\pi_-^2 p_-(\boldsymbol{x})p_-(\boldsymbol{x}')}{\pi_S}.$$

Proof. According to the independence assumption $(\boldsymbol{x},y) \perp (\boldsymbol{x}',y')$, we can immediately get the independence between $\boldsymbol{x},\ \boldsymbol{x}'$ and $y,\ y'$. Then the following equations hold:

$$\begin{aligned}
p_S(\boldsymbol{x},\boldsymbol{x}') &= p(\boldsymbol{x},\boldsymbol{x}'|y=y') = \frac{p(\boldsymbol{x},\boldsymbol{x}',y=y')}{p(y=y')} \\
&= \frac{p(\boldsymbol{x},y=+1,\boldsymbol{x}',y'=+1)+p(\boldsymbol{x},y=-1,\boldsymbol{x}',y'=-1)}{p(y=+1)p(y'=+1)+p(y=-1)p(y'=-1)} \\
&= \frac{p(\boldsymbol{x},y=+1)p(\boldsymbol{x}',y'=+1)+p(\boldsymbol{x},y=-1)p(\boldsymbol{x}',y'=-1)}{\pi_+^2+\pi_-^2} \\
&= \frac{\pi_+^2 p_+(\boldsymbol{x})p_+(\boldsymbol{x}')+\pi_-^2 p_-(\boldsymbol{x})p_-(\boldsymbol{x}')}{\pi_+^2+\pi_-^2} \\
&= \frac{\pi_+^2 p_+(\boldsymbol{x})p_+(\boldsymbol{x}')+\pi_-^2 p_-(\boldsymbol{x})p_-(\boldsymbol{x}')}{\pi_S}
\end{aligned}$$

□

Then we can prove the Theorem 1

Proof.

$$
\begin{aligned}
&\mathbb{E}_{p_S(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_S(s(\boldsymbol{x},\boldsymbol{x}')-\pi_-)(\ell(g(\boldsymbol{x}),+1)+\ell(g(\boldsymbol{x}'),+1))}{2(\pi_+-\pi_-)s(\boldsymbol{x},\boldsymbol{x}')}\right]\\
&=\int\frac{\pi_S(s(\boldsymbol{x},\boldsymbol{x}')-\pi_-)(\ell(g(\boldsymbol{x}),+1)+\ell(g(\boldsymbol{x}'),+1))}{2(\pi_+-\pi_-)s(\boldsymbol{x},\boldsymbol{x}')}*\frac{\pi_+^2p_+(\boldsymbol{x})p_+(\boldsymbol{x}')+\pi_-^2p_-(\boldsymbol{x})p_-(\boldsymbol{x}')}{\pi_S}d\boldsymbol{x}d\boldsymbol{x}'\\
&=\int\frac{(\pi_+^2p_+(\boldsymbol{x})p_+(\boldsymbol{x}')+\pi_-^2p_-(\boldsymbol{x})p_-(\boldsymbol{x}')-\pi_-p(\boldsymbol{x})p(\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),+1)+\ell(g(\boldsymbol{x}'),+1))}{2(\pi_+-\pi_-)}d\boldsymbol{x}d\boldsymbol{x}'\\
&=\int\frac{(\pi_+^2p_+(\boldsymbol{x})+\pi_-^2p_-(\boldsymbol{x})-\pi_-p(\boldsymbol{x}))\ell(g(\boldsymbol{x}),+1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}+\int\frac{(\pi_+^2p_+(\boldsymbol{x}')+\pi_-^2p_-(\boldsymbol{x}')-\pi_-p(\boldsymbol{x}'))\ell(g(\boldsymbol{x}'),+1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}'\\
&=\int\frac{\pi_+(\pi_+-\pi_-)p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}+\int\frac{\pi_+(\pi_+-\pi_-)p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}'),+1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}'\\
&=\int\frac{\pi_+p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}),+1)}{2}d\boldsymbol{x}+\int\frac{\pi_+p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}'),+1)}{2}d\boldsymbol{x}'\\
&=\frac{\pi_+\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)]}{2}+\frac{\pi_+\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)]}{2}\\
&=\pi_+\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)]
\end{aligned}
$$

Symmetrically, we have:

$$
\begin{aligned}
&\mathbb{E}_{p_S(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_S(\pi_+-s(\boldsymbol{x},\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)s(\boldsymbol{x},\boldsymbol{x}')}\right]\\
&=\int\frac{\pi_S(\pi_+-s(\boldsymbol{x},\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)s(\boldsymbol{x},\boldsymbol{x}')}*\frac{\pi_+^2p_+(\boldsymbol{x})p_+(\boldsymbol{x}')+\pi_-^2p_-(\boldsymbol{x})p_-(\boldsymbol{x}')}{\pi_S}d\boldsymbol{x}d\boldsymbol{x}'\\
&=\int\frac{(\pi_+p(\boldsymbol{x})p(\boldsymbol{x}')-\pi_+^2p_+(\boldsymbol{x})p_+(\boldsymbol{x}')-\pi_-^2p_-(\boldsymbol{x})p_-(\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)}d\boldsymbol{x}d\boldsymbol{x}'\\
&=\int\frac{(\pi_+p(\boldsymbol{x})-\pi_+^2p_+(\boldsymbol{x})-\pi_-^2p_-(\boldsymbol{x}))\ell(g(\boldsymbol{x}),-1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}+\int\frac{(\pi_+p(\boldsymbol{x}')-\pi_+^2p_+(\boldsymbol{x}')-\pi_-^2p_-(\boldsymbol{x}'))\ell(g(\boldsymbol{x}'),-1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}'\\
&=\int\frac{\pi_-(\pi_+-\pi_-)p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}+\int\frac{\pi_-(\pi_+-\pi_-)p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}'),-1)}{2(\pi_+-\pi_-)}d\boldsymbol{x}'\\
&=\int\frac{\pi_-p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)}{2}d\boldsymbol{x}+\int\frac{\pi_-p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}'),-1)}{2}d\boldsymbol{x}'\\
&=\frac{\pi_-\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]}{2}+\frac{\pi_-\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]}{2}\\
&=\pi_-\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]
\end{aligned}
$$

Then we have:

$$
\begin{aligned}
R_{\mathrm{S}}(g)=&\mathbb{E}_{p_S(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_S(s(\boldsymbol{x},\boldsymbol{x}')-\pi_-)(\ell(g(\boldsymbol{x}),+1)+\ell(g(\boldsymbol{x}'),+1))}{2(\pi_+-\pi_-)s(\boldsymbol{x},\boldsymbol{x}')}\right]\\
&+\mathbb{E}_{p_S(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_S(\pi_+-s(\boldsymbol{x},\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)s(\boldsymbol{x},\boldsymbol{x}')}\right].
\end{aligned}\tag{10}
$$

and we can give the unbiased estimator of classification risk according to the risk expression above:

$$
\hat{R}_{\mathrm{S}}(g)=\pi_S\sum\nolimits_{i=1}^{n}\frac{(s_i-\pi_-)(\ell(g(\boldsymbol{x}_i),+1)+\ell(g(\boldsymbol{x}_i'),+1))}{2n(\pi_+-\pi_-)s_i}+\pi_S\sum\nolimits_{i=1}^{n}\frac{(\pi_+-s_i)(\ell(g(\boldsymbol{x}_i),-1)+\ell(g(\boldsymbol{x}_i'),-1))}{2n(\pi_+-\pi_-)s_i}.
$$

which concludes the proof. □

C. Proof of Theorem 2

Proof. We aim to solve the following optimization problem when conducting ERM algorithm according to Theorem 1:

$$\min_{g\in\mathcal{G}} \pi_S \sum_{i=1}^{n} \left(\frac{(s_i-\pi_-)(\ell(g(\boldsymbol{x}_i),+1)+\ell(g(\boldsymbol{x}_i'),+1))}{2n(\pi_+-\pi_-)s_i} + \frac{(\pi_+-s_i)(\ell(g(\boldsymbol{x}_i),-1)+\ell(g(\boldsymbol{x}_i'),-1))}{2n(\pi_+-\pi_-)s_i} \right). \tag{11}$$

Notice that since $\pi_+ > \pi_-$ and $s_i \geq \pi_+$ for all $i \in [n]$, we have the following

$$\begin{cases} \frac{s_i-\pi_-}{2n(\pi_+-\pi_-)s_i} \geq 0, \ i = 1\cdots, n \\ \frac{\pi_+-s_i}{2n(\pi_+-\pi_-)s_i} \leq 0, \ i = 1\cdots, n \end{cases}$$

Since 0-1 loss is used, we have the conclusion that $\ell(g(\boldsymbol{x}), y) \in [0, 1]$ for any g, $\boldsymbol{x}$, and y. According to the discussion above, by setting all the $\ell(\cdot, +1)$ to 0 and $\ell(\cdot, -1)$ to 1, we can get the lower bound of (11):

$$(11) \geq \sum_{i=1}^{n} \frac{\pi_S(\pi_+ - s_i)}{n(\pi_+ - \pi_-)s_i}.$$

It is obvious that such setting can be realized if we let $g(\boldsymbol{x}) > 0$ for all the $\boldsymbol{x}$, which means that g classifies all the examples as positive. □

D. Proof of Lemma 2

Before proving the Lemma 2, we begin with the proof of two important technical Lemmas:

Lemma 4. *For any binary loss function $\ell(\cdot,\cdot) : \mathbb{R} \times \{+1, -1\} \to \mathbb{R}^+$:*

$$\mathbb{E}_{U^2}[s(\boldsymbol{x},\boldsymbol{x}')\ell(g(\boldsymbol{x}),+1)] = \pi_+^2\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)] + \pi_-^2\mathbb{E}_-[\ell(g(\boldsymbol{x}),+1)] \tag{12}$$
$$\mathbb{E}_{U^2}[s(\boldsymbol{x},\boldsymbol{x}')\ell(g(\boldsymbol{x}),-1)] = \pi_+^2\mathbb{E}_+[\ell(g(\boldsymbol{x}),-1)] + \pi_-^2\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)] \tag{13}$$

Proof. We only prove the first equation since the second one can be deduced in the same manner.

$$\begin{aligned}
\mathbb{E}_{U^2}[s(\boldsymbol{x},\boldsymbol{x}')\ell(g(\boldsymbol{x}),+1)] &= \int\int \frac{\pi_+^2 p_+(\boldsymbol{x})p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x})p_-(\boldsymbol{x}')}{p(\boldsymbol{x})p(\boldsymbol{x}')} p(\boldsymbol{x})p(\boldsymbol{x}')\ell(g(\boldsymbol{x}),+1)d\boldsymbol{x}d\boldsymbol{x}' \\
&= \int \pi_+^2\ell(g(\boldsymbol{x}),+1)p_+(\boldsymbol{x})d\boldsymbol{x} \int p_+(\boldsymbol{x}')d\boldsymbol{x}' \\
&\quad + \int \pi_-^2\ell(g(\boldsymbol{x}),+1)p_-(\boldsymbol{x})d\boldsymbol{x} \int p_-(\boldsymbol{x}')d\boldsymbol{x}' \\
&= \pi_+^2 \int \ell(g(\boldsymbol{x}),+1)p_+(\boldsymbol{x})d\boldsymbol{x} + \pi_-^2 \int \ell(g(\boldsymbol{x}),+1)p_-(\boldsymbol{x})d\boldsymbol{x} \\
&= \pi_+^2\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)] + \pi_-^2\mathbb{E}_-[\ell(g(\boldsymbol{x}),+1)]
\end{aligned}$$

□

Lemma 5.

$$\mathbb{E}_{U^2}[(1-s(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),+1)] = \pi_+\pi_-(\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)] + \mathbb{E}_-[\ell(g(\boldsymbol{x}),+1)]) \tag{14}$$
$$\mathbb{E}_{U^2}[(1-s(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),-1)] = \pi_+\pi_-(\mathbb{E}_+[\ell(g(\boldsymbol{x}),-1)] + \mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]) \tag{15}$$

Proof. First, we note that

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

Then we can prove the first equation:

$$\begin{aligned}\mathbb{E}_{U^2}[(1-s(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),+1)] &= \int\int \frac{\pi_+\pi_-(p_+(\boldsymbol{x})p_-(\boldsymbol{x}')+p_-(\boldsymbol{x})p_+(\boldsymbol{x}'))}{p(\boldsymbol{x})p(\boldsymbol{x}')}p(\boldsymbol{x})p(\boldsymbol{x}')\ell(g(\boldsymbol{x}),+1)d\boldsymbol{x}d\boldsymbol{x}'\\ &= \int \pi_+\pi_-\ell(g(\boldsymbol{x}),+1)p_+(\boldsymbol{x})d\boldsymbol{x}\int p_-(\boldsymbol{x}')d\boldsymbol{x}'\\ &\quad+\int \pi_+\pi_-\ell(g(\boldsymbol{x}),+1)p_-(\boldsymbol{x})d\boldsymbol{x}\int p_+(\boldsymbol{x}')d\boldsymbol{x}'\\ &= \pi_+\pi_-\left(\int \ell(g(\boldsymbol{x}),+1)p_+(\boldsymbol{x})d\boldsymbol{x}+\int \ell(g(\boldsymbol{x}),+1)p_-(\boldsymbol{x})d\boldsymbol{x}\right)\\ &= \pi_+\pi_-(\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)]+\mathbb{E}_-[\ell(g(\boldsymbol{x}),+1)])\end{aligned}$$

□

Note that similar conclusions for $\boldsymbol{x}'$ can be derived by switching $\boldsymbol{x}$ and $\boldsymbol{x}'$ in the lemmas above since they are completely symmetric.

Based on the lemmas above, we give the proof of Lemma 2.

Proof. We first prove the first equation. It can be deduced from Lemma 4 and 5 that:

$$\begin{aligned}&\mathbb{E}_{U^2}[s(\boldsymbol{x},\boldsymbol{x}')\ell(g(\boldsymbol{x}),+1)]-\frac{\pi_-}{\pi_+}\mathbb{E}_{U^2}[(1-s(\boldsymbol{x},\boldsymbol{x}'))\ell(g(\boldsymbol{x}),+1)]\\ &=\pi_+^2\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)]+\pi_-^2\mathbb{E}_-[\ell(g(\boldsymbol{x}),+1)]-\pi_-^2(\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)]+\mathbb{E}_-[\ell(g(\boldsymbol{x}),+1)])\\ &=(\pi_+^2-\pi_-^2)\mathbb{E}_+[\ell(g(\boldsymbol{x},+1))]\\ &=(\pi_+-\pi_-)\mathbb{E}_+[\ell(g(\boldsymbol{x},+1))]\end{aligned}$$

Dividing each side by $\pi_+-\pi_-$, we can get an equivalent expression of $R_+(g)$ and $\hat{R}_+(g)$ is its unbiased estimator, which we can conclude the proof of the first equation.

The proof of the second equation is omitted since it can be proved in a completely symmetric way. As shown in Bao et al. (2018), though any convex combination of the loss terms of $\boldsymbol{x}$ and $\boldsymbol{x}'$ can be the unbiased estimator, the formulation above can achieve minimal variance among all the potential candidates, which can be helpful for better generalization. □

E. Proof of Theorem 4

For convenience, we make the following notations:

$$\begin{aligned}\mathcal{L}_{Sconf}(g,(\boldsymbol{x},\boldsymbol{x}')) \triangleq& \frac{(s(\boldsymbol{x},\boldsymbol{x}')-\pi_-)(\ell(g(\boldsymbol{x}),+1)+\ell(g(\boldsymbol{x}'),+1))}{2(\pi_+-\pi_-)}\\ &-\frac{(s(\boldsymbol{x},\boldsymbol{x}')-\pi_+)(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)}\end{aligned}$$

Denote the Sconf data pairs of size n with $S_n \overset{i.i.d.}{\sim} p(\boldsymbol{x}, \boldsymbol{x}')$. We first introduce the Rademacher complexity and give the following technical lemma:

Definition 2. (Rademacher complexity (Bartlett & Mendelson, 2001))). *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ be i.i.d. random variables drawn from a probability distribution $\mathcal{D}$, $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R}\}$ be a class of measurable functions. Then the Rademacher complexity of $\mathcal{G}$ is defined as:*

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i) \right]. \tag{16}$$

where $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_n)$ are Rademacher variables taking from $\{-1, +1\}$ uniformly.

Lemma 6.

$$\bar{\mathfrak{R}}_n(\mathcal{L}_{Sconf}) \leq \frac{L_\ell}{|\pi_+ - \pi_-|} \mathfrak{R}_n(\mathcal{G})$$

where $\bar{\mathfrak{R}}_n(\mathcal{L}_{Sconf})$ is the Rademacher complexity of $\mathcal{L}_{Sconf}$ over Sconf data pairs of size n drawn from U^2.

Proof. Due to the sub-additivity of supremum, symmetry between $\boldsymbol{x}$ and $\boldsymbol{x}'$ and the property of Rademacher variable:

$$\begin{aligned}
\bar{\mathfrak{R}}_n(\mathcal{L}_{Sconf}) &= \mathbb{E}_{\mathcal{S}_n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i \frac{\mathcal{L}_{Sconf}(g, \boldsymbol{x}_i, \boldsymbol{x}_i')}{n} \right] \\
&\leq \mathbb{E}_{\mathcal{S}_n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i \frac{(s(\boldsymbol{x}_i, \boldsymbol{x}_i') - \pi_-)\ell(g(\boldsymbol{x}_i), +1) + (\pi_+ - (s(\boldsymbol{x}_i, \boldsymbol{x}_i'))\ell(g(\boldsymbol{x}_i), -1))}{n(\pi_+ - \pi_-)} \right]
\end{aligned}$$

Suppose $\pi_+ > \pi_-$. We also have the following results:

$$\begin{aligned}
&\left\| \nabla \left(\frac{(s(\boldsymbol{x}, \boldsymbol{x}') - \pi_-)\ell(g(\boldsymbol{x}), +1) + (\pi_+ - s(\boldsymbol{x}, \boldsymbol{x}'))\ell(g(\boldsymbol{x}), -1))}{(\pi_+ - \pi_-)} \right) \right\|_2 \\
&\leq \left\| \nabla \left(\frac{(s(\boldsymbol{x}, \boldsymbol{x}') - \pi_-)\ell(g(\boldsymbol{x}), +1)}{(\pi_+ - \pi_-)} \right) \right\|_2 + \left\| \nabla \left(\frac{(\pi_+ - (s(\boldsymbol{x}, \boldsymbol{x}'))\ell(g(\boldsymbol{x}), -1))}{(\pi_+ - \pi_-)} \right) \right\|_2 \\
&\leq \frac{|s(\boldsymbol{x}, \boldsymbol{x}') - \pi_-| L_\ell}{(\pi_+ - \pi_-)} + \frac{|\pi_+ - (s(\boldsymbol{x}, \boldsymbol{x}')| L_\ell}{(\pi_+ - \pi_-)}
\end{aligned} \tag{17}$$

We can further bound (17) under different conditions:

$$\frac{|s(\boldsymbol{x}, \boldsymbol{x}') - \pi_-| L_\ell}{(\pi_+ - \pi_-)} + \frac{|\pi_+ - (s(\boldsymbol{x}, \boldsymbol{x}')| L_\ell}{(\pi_+ - \pi_-)} \leq \begin{cases} L_\ell, & s(\boldsymbol{x}, \boldsymbol{x}') \in [\pi_-, \ \pi_+], \\ \frac{L_\ell}{|\pi_+ - \pi_-|}, & s(\boldsymbol{x}, \boldsymbol{x}') \notin [\pi_-, \ \pi_+]. \end{cases}$$

which shows that (17) is upper bounded by $\frac{L_\ell}{|\pi_+ - \pi_-|}$ According to Talagrand's lemma(Ledoux & Talagrand, 2013) and the result above, we can further get the following inequality:

$$\begin{aligned}
\bar{\mathfrak{R}}_n(\mathcal{L}_{Sconf}) &\leq \frac{L_\ell}{|\pi_+ - \pi_-|} \mathbb{E}_{\mathcal{S}_n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i) \right] \\
&= \frac{L_\ell}{|\pi_+ - \pi_-|} \mathbb{E}_{\mathcal{X}_n} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i) \right] \\
&= \frac{L_\ell}{|\pi_+ - \pi_-|} \mathfrak{R}_n(\mathcal{G})
\end{aligned}$$

□

Then we can bound $\sup_{g \in \mathcal{G}} \left| \hat{R}(g) - R(g) \right|$ using McDiarmid's inequality:

Lemma 7. *The inequalities below hold with probability at least* $1-\delta$*:*

$$\sup_{g\in\mathcal{G}}\left|R(g)-\hat{R}(g)\right| \leq \frac{L_\ell}{|\pi_+-\pi_-|}\mathfrak{R}_n(\mathcal{G}) + \frac{C_\ell}{|\pi_+-\pi_-|}\sqrt{\frac{\ln 2/\delta}{2n}}. \tag{18}$$

Proof. To begin with, we first bound the one-side supremum $\sup_{g\in\mathcal{G}}\left(R(g)-\hat{R}(g)\right)$. Denote $\Phi = \sup_{g\in\mathcal{G}}\left(R(g)-\hat{R}(g)\right)$ and $\bar{\Phi} = \sup_{g\in\mathcal{G}}\left(R(g)-\hat{\bar{R}}(g)\right)$, where $\hat{R}(g)$ and $\hat{\bar{R}}(g)$ are empirical risk over two samples differing by exactly one point: $\{(\boldsymbol{x}_n,\boldsymbol{x}'_n),s_n\}$ and $\{(\bar{\boldsymbol{x}}_n,\bar{\boldsymbol{x}}'_n),\bar{s}_n\}$. Then we have:

$$\begin{aligned}
\bar{\Phi}-\Phi &\leq \sup_{g\in\mathcal{G}}\left(\hat{R}(g)-\hat{\bar{R}}(g)\right)\\
&\leq \sup_{g\in\mathcal{G}}\left(\frac{\mathcal{L}_{Sconf}(g,\boldsymbol{x}_n,\boldsymbol{x}'_n)-\mathcal{L}_{Sconf}(g,\bar{\boldsymbol{x}}_n,\bar{\boldsymbol{x}}'_n)}{n}\right)\\
&\leq \frac{C_\ell}{n|\pi_+-\pi_-|}
\end{aligned}$$

and $\Phi-\bar{\Phi}$ has the same upper bound symmetrically. By applying McDiarmid's inequality, the inequality below holds with probability at least $1-\frac{\delta}{2}$:

$$\sup_{g\in\mathcal{G}}\left(R(g)-\hat{R}(g)\right) \leq \mathbb{E}_{S_n}\left[\sup_{g\in\mathcal{G}}\left(R(g)-\hat{R}(g)\right)\right] + \frac{C_\ell}{|\pi_+-\pi_-|}\sqrt{\frac{\ln 2/\delta}{2n}}. \tag{19}$$

The following step is to bound $\mathbb{E}_{S_n}\left[\sup_{g\in\mathcal{G}}\left(R(g)-\hat{R}(g)\right)\right]$ with Rademacher complexity. It is a routine work to show by symmetrization(Mohri et al., 2012) and Lemma 6 that

$$\begin{aligned}
\mathbb{E}_{S_n}\left[\sup_{g\in\mathcal{G}}\left(R(g)-\hat{R}(g)\right)\right] &\leq \bar{\mathfrak{R}}_n(\mathcal{L}_{Sconf})\\
&\leq \frac{L_\ell}{|\pi_+-\pi_-|}\bar{\mathfrak{R}}_n(\mathcal{G})
\end{aligned}$$

The other direction $\sup_{g\in\mathcal{G}}\left(\hat{R}(g)-R(g)\right)$ is similar. Using the union bound, the following inequality holds with probability at least $1-\delta$:

$$\sup_{g\in\mathcal{G}}\left|R(g)-\hat{R}(g)\right| \leq \frac{L_\ell}{|\pi_+-\pi_-|}\mathfrak{R}_n(\mathcal{G}) + \frac{C_\ell}{|\pi_+-\pi_-|}\sqrt{\frac{\ln 2/\delta}{2n}}. \tag{20}$$

□

Then we can prove Theorem 4:

Proof.

$$\begin{aligned}
R(\hat{g})-R(g^*) &= (R(\hat{g})-\hat{R}(\hat{g})) + (\hat{R}(\hat{g})-\hat{R}(g^*)) + (\hat{R}(g^*)-R(g^*))\\
&\leq (R(\hat{g})-\hat{R}(\hat{g})) + (\hat{R}(g^*)-R(g^*))\\
&\leq |R(\hat{g})-\hat{R}(\hat{g})| + |\hat{R}(g^*)-R(g^*)|\\
&\leq 2\sup_{g\in\mathcal{G}}|R(g)-\hat{R}(g)|
\end{aligned}$$

The first inequality holds due to the definition of ERM. We can conclude the proof by applying Lemma 7. □

F. Proof of Theorem 5

Proof. We first prove the unbiasedness of proposed class-prior estimator:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_n}\left[\frac{\sum_{i=1}^n s(\boldsymbol{x}_i,\boldsymbol{x}_i')}{n}\right] &= \sum_{i=1}^n \mathbb{E}_{U^2}\left[\frac{s(\boldsymbol{x},\boldsymbol{x}')}{n}\right] = \mathbb{E}_{U^2}[s(\boldsymbol{x},\boldsymbol{x}')] \\
&= \int\int p(y=+1|\boldsymbol{x})p(y'=+1|\boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}' \\
&\quad + \int\int p(y=-1|\boldsymbol{x})p(y'=-1|\boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}' \\
&= \int\int p(\boldsymbol{x},y=+1)p(\boldsymbol{x}',y'=+1)d\boldsymbol{x}d\boldsymbol{x}' \\
&\quad + \int\int p(\boldsymbol{x},y=-1)p(\boldsymbol{x}',y'=-1)d\boldsymbol{x}d\boldsymbol{x}' \\
&= p(y=+1)p(y'=+1) + p(y=-1)p(y'=-1) \\
&= \pi_+^2 + \pi_-^2
\end{aligned}$$

Note that for any different Sconf data pairs $(\boldsymbol{x},\boldsymbol{x}')$ and $(\bar{\boldsymbol{x}},\bar{\boldsymbol{x}}')$: $\frac{|s(\boldsymbol{x},\boldsymbol{x}')-s(\bar{\boldsymbol{x}},\bar{\boldsymbol{x}}')|}{n} \leq \frac{1}{n}$. Then we can simply prove the consistency of proposed estimator using McDiarmid's inequality, which can be formulated as the following theorem:

Theorem 9. *For any $\delta > 0$ and $\mathcal{S}_n \overset{i.i.d.}{\sim} U^{2n}$, the following inequality holds with probability at least $1-\delta$:*

$$\left|\sum_{i=1}^n \frac{s(\boldsymbol{x}_i,\boldsymbol{x}_i')}{n} - (\pi_+^2+\pi_-^2)\right| \leq \sqrt{\frac{\ln 2/\delta}{2n}} \tag{21}$$

□

G. Proof of Theorem 6

Proof. According to the definition of empirical minimizers $\bar{g}$, $\hat{g}$ and the proof of Theorem 4:

$$\begin{aligned}
R(\bar{g}) - R(g^*) &= \left(R(\bar{g}) - \hat{R}(\bar{g})\right) + \left(\hat{R}(\bar{g}) - \bar{R}(\bar{g})\right) + \left(\bar{R}(\bar{g}) - \bar{R}(\hat{g})\right) + \left(\bar{R}(\hat{g}) - \hat{R}(\hat{g})\right) \\
&\quad + \left(\hat{R}(\hat{g}) - R(\hat{g})\right) + \left(R(\hat{g}) - R(g^*)\right) \\
&\leq 2\sup_{g\in\mathcal{G}}|R(g) - \hat{R}(g)| + 2\sup_{g\in\mathcal{G}}|\bar{R}(g) - \hat{R}(g)| + \left(\bar{R}(\hat{g}) - \hat{R}(\hat{g})\right) \\
&\leq 4\sup_{g\in\mathcal{G}}|R(g) - \hat{R}(g)| + 2\sup_{g\in\mathcal{G}}|\left(\bar{R}(g) - \hat{R}(g)\right)| \\
&\leq 4\sup_{g\in\mathcal{G}}|R(g) - \hat{R}(g)| + \frac{2\sum_{i=1}^n C_\ell\sigma_n}{n(\pi_+ - \pi_-)}
\end{aligned}$$

According to Lemma 7, the following inequality holds with probability at least $1-\delta$:

$$R(\bar{g}) - R(g^*) \leq \frac{4L_\ell\mathfrak{R}_n(\mathcal{G})}{|\pi_+ - \pi_-|} + \frac{4C_\ell}{|\pi_+ - \pi_-|}\sqrt{\frac{\ln 2/\delta}{2n}} + \frac{2C_\ell\sigma_n}{n|\pi_+ - \pi_-|},$$

which concludes the proof. □

H. Proof of Theorem 7

Denote the Sconf data pairs of size n with $\mathcal{S}_n = \{(\boldsymbol{x}_{,i}, \boldsymbol{x}_i')\}_{i=1}^n$. We first make the following notation: $\mathfrak{D}_-^n(g) = \{\mathcal{S}_n|\hat{R}_+(g) < 0\} \cup \{\mathcal{S}_n|\hat{R}_-(g) < 0\}$, $\mathfrak{D}_+^n(g) = \{\mathcal{S}_n|\hat{R}_+(g) \geq 0\} \cap \{\mathcal{S}_n|\hat{R}_-(g) \geq 0\}$, $R_+(g) = \pi_+\mathbb{E}_+[\ell(g(\boldsymbol{x}),+1)]$, $R_-(g) = \pi_-\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]$. Before proving Theorem 7, we begin with the proof of a technical lemma.

Lemma 8. *Assume that there is $\alpha > 0$ and $\beta > 0$ such that $R_+(g) \geq \alpha$ and $R_-(g) \geq \beta$. By assumptions in Theorem 4, the probability measure of $\mathfrak{D}_-(g)$ can be upper bounded by:*

$$\mathbb{P}(\mathfrak{D}_-(g)) \leq \exp\left(-\frac{(\pi_+ - \pi_-)^2 n}{2C_\ell^2}\right)\Delta$$

where $\Delta = \exp(\alpha^2) + \exp(\beta^2)$.

Proof. According to the data generation process:

$$p(S_n) = p(\boldsymbol{x}_1)\cdots p(\boldsymbol{x}_n)p(\boldsymbol{x}'_1)\cdots p(\boldsymbol{x}'_n),$$

and the probability measure of $\mathfrak{D}_-(g)$ is defined as below:

$$\mathbb{P}(\mathfrak{D}_-(g)) = \int_{\mathcal{S}_n \in \mathfrak{D}_-(g)} p(S_n)dS_n = \int_{\mathcal{S}_n \in \mathfrak{D}_-(g)} p(S_n)d\boldsymbol{x}_1\cdots d\boldsymbol{x}_n d\boldsymbol{x}'_1\cdots d\boldsymbol{x}'_n,$$

where $\mathbb{P}$ is the probability.

By assumptions in Theorem 4, the change of $\hat{R}_+(g)$ and $\hat{R}_-(g)$ will be no more that $2C_\ell/n|\pi_+ - \pi_-|$ if exactly one pair of Sconf data $(\boldsymbol{x}_i, \boldsymbol{x}'_i) \in \mathcal{S}_n$ is replaced. According to McDiarmid's inequality:

$$\mathbb{P}(R_+(g) - \hat{R}_+(g) \geq \alpha) \leq \exp\left(-\frac{\alpha^2(\pi_+ - \pi_-)^2 n}{2C_\ell^2}\right)$$

and

$$\mathbb{P}(R_-(g) - \hat{R}_-(g) \geq \beta) \leq \exp\left(-\frac{\beta^2(\pi_+ - \pi_-)^2 n}{2C_\ell^2}\right)$$

Then we can bound $\mathbb{P}(\mathfrak{D}_-(g))$ in this manner:

$$\begin{aligned}
\mathbb{P}(\mathfrak{D}_-(g)) &\leq \mathbb{P}(\hat{R}_+(g) \leq 0) + \mathbb{P}(\hat{R}_-(g) \leq 0) \\
&\leq \mathbb{P}(\hat{R}_+(g) \leq R_+(g) - \alpha) + \mathbb{P}(\hat{R}_-(g) \leq R_-(g) - \beta) \\
&= \mathbb{P}(R_+(g) - \hat{R}_+(g) \geq \alpha) + \mathbb{P}(R_-(g) - \hat{R}_-(g) \geq \beta) \\
&\leq \exp\left(-\frac{\alpha^2(\pi_+ - \pi_-)^2 n}{2C_\ell^2}\right) + \exp\left(-\frac{\beta^2(\pi_+ - \pi_-)^2 n}{2C_\ell^2}\right) \\
&= \exp\left(-\frac{(\pi_+ - \pi_-)^2 n}{2C_\ell^2}\right)\Delta
\end{aligned}$$

The first inequality holds due to union bound and the second one is deduced according to the assumptions. □

Then we prove the Theorem 7:

Proof. According to the definition of consistent correction function:

$$\begin{aligned}
&\mathbb{E}[\widetilde{R}(g)] - R(g) = \mathbb{E}[\widetilde{R}(g) - \hat{R}(g)] \\
&= \int_{S_n \in \mathfrak{D}_+(g)} (\widetilde{R}(g) - \hat{R}(g))p(S_n)dS_n + \int_{S_n \in \mathfrak{D}_-(g)} (\widetilde{R}(g) - \hat{R}(g))p(S_n)dS_n \\
&= \int_{S_n \in \mathfrak{D}_-(g)} (\widetilde{R}(g) - \hat{R}(g))p(S_n)dS_n
\end{aligned}$$

According to the definition of $\widetilde{R}(g)$, we know that it can upper bound $\hat{R}(g)$: $\widetilde{R}(g) \geq \hat{R}(h)$. Then we can get the *l.h.s.* inequality:

$$\mathbb{E}[\widetilde{R}(g) - \hat{R}(g)] \geq 0$$

Note that the consistent correction function is Lipschitz continuous with Lipschitz constant $L_f = \max\{1, k\}$ and $f(0) = 0$. Then we upper bound $\mathbb{E}[\widetilde{R}(g)] - R(g)$ based on the assumptions in Theorem 4 and the fact that $\left|\hat{R}_+(g)\right|$ and $\left|\hat{R}_-(g)\right|$ can be bounded by $2C_\ell/|\pi_+ - \pi_-|$:

$$\begin{aligned}
\mathbb{E}[\widetilde{R}(g)] - R(g) &= \int_{S_n \in \mathfrak{D}_-(g)} (\widetilde{R}(g) - \hat{R}(g)) p(S_n) dS_n \\
&\leq \sup_{S_n \in \mathfrak{D}_-(g)} \left((\widetilde{R}(g) - \hat{R}(g) \right) \int_{S_n \in \mathfrak{D}_-(g)} p(S_n) dS_n \\
&= \sup_{S_n \in \mathfrak{D}_-(g)} \left((\widetilde{R}(g) - \hat{R}(g) \right) \mathbb{P}(\mathfrak{D}_-(g)) \\
&= \sup_{S_n \in \mathfrak{D}_-(g)} \left(f\left(\hat{R}_+(g)\right) + f\left(\hat{R}_-(g)\right) - \hat{R}_+(g) - \hat{R}_-(g) \right) \mathbb{P}(\mathfrak{D}_-(g)) \\
&\leq \sup_{S_n \in \mathfrak{D}_-(g)} \left(L_f \left|\hat{R}_+(g)\right| + L_f \left|\hat{R}_-(g)\right| + \left|\hat{R}_+(g)\right| + \left|\hat{R}_-(g)\right| \right) \mathbb{P}(\mathfrak{D}_-(g)) \\
&\leq \sup_{S_n \in \mathfrak{D}_-(g)} \left(\frac{(L_f + 1)C_\ell}{|\pi_+ - \pi_-|} \right) \mathbb{P}(\mathfrak{D}_-(g)) \\
&= \frac{(L_f + 1)C_\ell}{|\pi_+ - \pi_-|} \exp\left(-\frac{(\pi_+ - \pi_-)^2 n}{2C_\ell^2} \right) \Delta
\end{aligned}$$

Then we give the high-probability bound of consistent risk estimator $\widetilde{R}(g)$ by bounding $\left|\widetilde{R}(g) - R(g)\right|$. We first give the following inequality according to the discussions above:

$$\begin{aligned}
\left|\widetilde{R}(g) - R(g)\right| &\leq \left|\widetilde{R}(g) - \mathbb{E}[\widetilde{R}(g)]\right| + \left|\mathbb{E}[\widetilde{R}(g)] - R(g)\right| \\
&\leq \left|\widetilde{R}(g) - \mathbb{E}[\widetilde{R}(g)]\right| + \frac{(L_f + 1)C_\ell}{|\pi_+ - \pi_-|} \exp\left(-\frac{(\pi_+ - \pi_-)^2 n}{2C_\ell^2} \right) \Delta
\end{aligned} \tag{22}$$

Then we can focus on bounding $\left|\widetilde{R}(g) - \mathbb{E}[\widetilde{R}(g)]\right|$. According to the definition of $\widetilde{R}(g)$ and the Lipschitzness of $f(\cdot)$, the change of $\widetilde{R}(g)$ will be no more than $L_\ell C_\ell / n |\pi_+ - \pi_-|$. Then we can simply bound $\left|\widetilde{R}(g) - \mathbb{E}[\widetilde{R}(g)]\right|$ using McDiarmid's inequality. With probability at least $1 - \delta$, the following inequality holds:

$$\left|\widetilde{R}(g) - \mathbb{E}[\widetilde{R}(g)]\right| \leq \frac{L_\ell C_\ell}{|\pi_+ - \pi_-|} \sqrt{\frac{\ln 2/\delta}{2n}}$$

We can conclude the proof by combining the inequality above and (22). □

I. Proof of Theorem 8

Based on Theorem 7 and the proof of Theorem 4, we prove Theorem 8:

Proof. We first give the following inequalities:

$$\begin{aligned}
R(\tilde{g}) - R(g^*) &= \left(R(\tilde{g}) - \widetilde{R}(\tilde{g})\right) + \left(\widetilde{R}(\tilde{g}) - \widetilde{R}(\hat{g})\right) + \left(\widetilde{R}(\hat{g}) - R(\hat{g})\right) + (R(\hat{g}) - R(g^*)) \\
&\leq \left|R(\tilde{g}) - \widetilde{R}(\tilde{g})\right| + \left|\widetilde{R}(\hat{g}) - R(\hat{g})\right| + (R(\hat{g}) - R(g^*))
\end{aligned}$$

Then we can conclude the proof by combining the high-probability bound in Theorem 7, Theorem 4 and union bound. With probability at least $1 - \delta$, the following inequality holds:

$$\begin{aligned}
R(\tilde{g}) - R(g^*) &\leq \left|R(\tilde{g}) - \widetilde{R}(\tilde{g})\right| + \widetilde{R}(\tilde{g}) - \widetilde{R}(\hat{g}) + \left|\widetilde{R}(\hat{g}) - R(\hat{g})\right| + (R(\hat{g}) - R(g^*)) \\
&\leq \frac{2L_\ell}{|\pi_+ - \pi_-|} \mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{\ln 6/\delta}{2n}} \left(\frac{2L_\ell C_\ell + 2C_\ell}{|\pi_+ - \pi_-|} \right) + \frac{2(L_f + 1)C_\ell}{|\pi_+ - \pi_-|} \exp\left(-\frac{(\pi_+ - \pi_-)^2 n}{2C_\ell^2} \right) \Delta
\end{aligned}$$

□

J. Symmetric Conclusions of Theorem 1 and Theorem 2 for Dissimilar Data Pairs

Suppose the dissimilar data pairs $\{(\boldsymbol{x}_i, \boldsymbol{x}_i')\}_{i=1}^n$ are drawn from the distribution with density $p_D(\boldsymbol{x}, \boldsymbol{x}') = p(\boldsymbol{x}, \boldsymbol{x}'|y \neq y')$. We give an unbiased risk estimator of classification risk with only dissimilar data pairs and their similarity confidence:

Theorem 10. *With dissimilar data pairs and their similarity confidence, assuming that $s(\boldsymbol{x}, \boldsymbol{x}') < 1$ for all the pair $(\boldsymbol{x}, \boldsymbol{x}')$, we can get the unbiased estimator of classification risk (1), i.e., $\mathbb{E}_{p(\boldsymbol{x},\boldsymbol{x}'|y\neq y')}[\hat{R}_{\mathrm{D}}(g)] = R(g)$, where*

$$\hat{R}_{\mathrm{D}}(g) = 2\pi_+\pi_- \sum_{i=1}^n \frac{(s_i-\pi_-)(\ell(g(\boldsymbol{x}_i),+1)+\ell(g(\boldsymbol{x}_i'),+1))}{2n(\pi_+-\pi_-)(1-s_i)} + 2\pi_+\pi_- \sum_{i=1}^n \frac{(\pi_+-s_i)(\ell(g(\boldsymbol{x}_i),-1)+\ell(g(\boldsymbol{x}_i'),-1))}{2n(\pi_+-\pi_-)(1-s_i}). \tag{23}$$

Proof. First we show the equivalent expression of $p(\boldsymbol{x}, \boldsymbol{x}'|y \neq y')$. According to the independence assumption $(\boldsymbol{x}, y) \perp (\boldsymbol{x}', y')$, we can immediately get the independence between $\boldsymbol{x},\ \boldsymbol{x}'$ and $y,\ y'$. Then the following equations hold:

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

Denote $2\pi_+\pi_-$ with π_{D}. Then we can prove the theorem above.

$$\begin{aligned}
&\mathbb{E}_{p_D(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_D(s(\boldsymbol{x}, \boldsymbol{x}') - \pi_-)(\ell(g(\boldsymbol{x}), +1) + \ell(g(\boldsymbol{x}'), +1))}{2(\pi_+ - \pi_-)(1 - s(\boldsymbol{x}, \boldsymbol{x}'))}\right] \\
&= \int \frac{\pi_D(s(\boldsymbol{x}, \boldsymbol{x}') - \pi_-)(\ell(g(\boldsymbol{x}), +1) + \ell(g(\boldsymbol{x}'), +1))}{2(\pi_+ - \pi_-)(1 - s(\boldsymbol{x}, \boldsymbol{x}'))} * \frac{p_+(\boldsymbol{x})p_-(\boldsymbol{x}') + p_-(\boldsymbol{x})p_+(\boldsymbol{x}')}{2} d\boldsymbol{x}d\boldsymbol{x}' \\
&= \int \frac{(\pi_+^2 p_+(\boldsymbol{x})p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x})p_-(\boldsymbol{x}') - \pi_- p(\boldsymbol{x})p(\boldsymbol{x}'))(\ell(g(\boldsymbol{x}), +1) + \ell(g(\boldsymbol{x}'), +1))}{4(\pi_+ - \pi_-)} d\boldsymbol{x}d\boldsymbol{x}' \\
&= \int \frac{(\pi_+^2 p_+(\boldsymbol{x}) + \pi_-^2 p_-(\boldsymbol{x}) - \pi_- p(\boldsymbol{x}))\ell(g(\boldsymbol{x}), +1)}{2(\pi_+ - \pi_-)} d\boldsymbol{x} + \int \frac{(\pi_+^2 p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}') - \pi_- p(\boldsymbol{x}'))\ell(g(\boldsymbol{x}'), +1)}{2(\pi_+ - \pi_-)} d\boldsymbol{x}' \\
&= \int \frac{\pi_+(\pi_+ - \pi_-)p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}), +1)}{2(\pi_+ - \pi_-)} d\boldsymbol{x} + \int \frac{\pi_+(\pi_+ - \pi_-)p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}'), +1)}{2(\pi_+ - \pi_-)} d\boldsymbol{x}' \\
&= \int \frac{\pi_+ p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}), +1)}{2} d\boldsymbol{x} + \int \frac{\pi_+ p_+(\boldsymbol{x})\ell(g(\boldsymbol{x}'), +1)}{2} d\boldsymbol{x}' \\
&= \frac{\pi_+\mathbb{E}_+[\ell(g(\boldsymbol{x}), +1)]}{2} + \frac{\pi_+\mathbb{E}_+[\ell(g(\boldsymbol{x}), +1)]}{2} \\
&= \pi_+\mathbb{E}_+[\ell(g(\boldsymbol{x}), +1)]
\end{aligned}$$

Symmetrically, we have:

$$
\begin{aligned}
&\mathbb{E}_{p_D(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_D(\pi_+ - s(\boldsymbol{x},\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)(1-s(\boldsymbol{x},\boldsymbol{x}'))}\right]\\
&=\int \frac{\pi_S(\pi_+ - s(\boldsymbol{x},\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)(1-s(\boldsymbol{x},\boldsymbol{x}'))} * \frac{p_+(\boldsymbol{x})p_-(\boldsymbol{x}')+p_-(\boldsymbol{x})p_+(\boldsymbol{x}')}{2} d\boldsymbol{x}d\boldsymbol{x}'\\
&=\int \frac{(\pi_+ p(\boldsymbol{x})p(\boldsymbol{x}') - \pi_+^2 p_+(\boldsymbol{x})p_+(\boldsymbol{x}') - \pi_-^2 p_-(\boldsymbol{x})p_-(\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)} d\boldsymbol{x}d\boldsymbol{x}'\\
&=\int \frac{(\pi_+ p(\boldsymbol{x}) - \pi_+^2 p_+(\boldsymbol{x}) - \pi_-^2 p_-(\boldsymbol{x}))\ell(g(\boldsymbol{x}),-1)}{2(\pi_+-\pi_-)} d\boldsymbol{x} + \int \frac{(\pi_+ p(\boldsymbol{x}') - \pi_+^2 p_+(\boldsymbol{x}') - \pi_-^2 p_-(\boldsymbol{x}'))\ell(g(\boldsymbol{x}'),-1)}{2(\pi_+-\pi_-)} d\boldsymbol{x}'\\
&=\int \frac{\pi_-(\pi_+-\pi_-)p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)}{2(\pi_+-\pi_-)} d\boldsymbol{x} + \int \frac{\pi_-(\pi_+-\pi_-)p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}'),-1)}{2(\pi_+-\pi_-)} d\boldsymbol{x}'\\
&=\int \frac{\pi_- p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}),-1)}{2} d\boldsymbol{x} + \int \frac{\pi_- p_-(\boldsymbol{x})\ell(g(\boldsymbol{x}'),-1)}{2} d\boldsymbol{x}'\\
&=\frac{\pi_-\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]}{2} + \frac{\pi_-\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]}{2}\\
&=\pi_-\mathbb{E}_-[\ell(g(\boldsymbol{x}),-1)]
\end{aligned}
$$

Then we have:

$$
\begin{aligned}
R_{\mathrm{D}}(g) =& \mathbb{E}_{p_D(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_D(s(\boldsymbol{x},\boldsymbol{x}')-\pi_-)(\ell(g(\boldsymbol{x}),+1)+\ell(g(\boldsymbol{x}'),+1))}{2(\pi_+-\pi_-)(1-s(\boldsymbol{x},\boldsymbol{x}'))}\right]\\
&+\mathbb{E}_{p_D(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\pi_D(\pi_+ - s(\boldsymbol{x},\boldsymbol{x}'))(\ell(g(\boldsymbol{x}),-1)+\ell(g(\boldsymbol{x}'),-1))}{2(\pi_+-\pi_-)(1-s(\boldsymbol{x},\boldsymbol{x}'))}\right].
\end{aligned} \tag{24}
$$

and we can give the unbiased estimator of classification risk according to the risk expression above:

$$
\hat{R}_{\mathrm{D}}(g) = 2\pi_+\pi_- \textstyle\sum_{i=1}^n \frac{(s_i-\pi_-)(\ell(g(\boldsymbol{x}_i),+1)+\ell(g(\boldsymbol{x}_i'),+1))}{2n(\pi_+-\pi_-)(1-s_i)} + 2\pi_+\pi_- \sum_{i=1}^n \frac{(\pi_+-s_i)(\ell(g(\boldsymbol{x}_i),-1)+\ell(g(\boldsymbol{x}_i'),-1))}{2n(\pi_+-\pi_-)(1-s_i)}. \tag{25}
$$

which concludes the proof.

□

Denote the empirical risk minimizer of $\hat{R}_{\mathrm{D}}(g)$ with $\hat{g}_{\mathrm{D}}$. We theoretically show that learning with only dissimilar data pairs can result in collapsed solution:

Theorem 11. *Suppose $\pi_+ > \pi_-$ and 0-1 loss is used. For similar data pairs, we assume that $s_i \leq \pi_-$ for $i = 1, \cdots, n$. Then $\hat{g}_{\mathrm{D}}$ is a collapsed solution that classifies all the examples as negative.*

Proof. We aim to solve the following optimization problem when conducting ERM algorithm according to Theorem 1:

$$
\min_{g\in\mathcal{G}} \pi_D \sum\nolimits_{i=1}^n \left(\frac{(s_i-\pi_-)(\ell(g(\boldsymbol{x}_i),+1)+\ell(g(\boldsymbol{x}_i'),+1))}{2n(\pi_+-\pi_-)(1-s_i)} + \frac{(\pi_+-s_i)(\ell(g(\boldsymbol{x}_i),-1)+\ell(g(\boldsymbol{x}_i'),-1))}{2n(\pi_+-\pi_-)(1-s_i)}\right). \tag{26}
$$

Notice that since $\pi_+ > \pi_-$ and $s_i \leq \pi_-$ for all $i \in [n]$, we have the following

$$
\begin{cases}
\frac{s_i-\pi_-}{2n(\pi_+-\pi_-)(1-s_i)} \leq 0, \ i = 1\cdots, n\\
\frac{\pi_+-s_i}{2n(\pi_+-\pi_-)(1-s_i)} \geq 0, \ i = 1\cdots, n
\end{cases}
$$

Since 0-1 loss is used, we have the conclusion that $\ell(g(\boldsymbol{x}), y) \in [0, 1]$ for any g, $\boldsymbol{x}$, and y. According to the discussion above, by setting all the $\ell(\cdot, -1)$ to 0 and $\ell(\cdot, +1)$ to 1, we can get the lower bound of (26):

$$
(26) \geq \sum_{i=1}^n \frac{\pi_D(s_i-\pi_-)}{n(\pi_+-\pi_-)(1-s_i)}.
$$

It is obvious that such setting can be realized if we let $g(\boldsymbol{x}) < 0$ for all the $\boldsymbol{x}$, which means that g classifies all the examples as negative. □

Table 3. Detailed Statistics of benchmark datasets and models

Datasets	# Train	# Validation	# Test	π_+	Dim	Model $g(\boldsymbol{x})$
MNIST	54000	6000	10000	0.3	784	3-layer MLP with ReLU (d-500-500-1)
Kuzushiji-MNIST	54000	6000	10000	0.7	784	3-layer MLP with ReLU (d-500-500-1)
Fashion-MNIST	54000	6000	10000	0.4	784	3-layer MLP with ReLU (d-500-500-1)
EMNIST-Digits	216000	24000	40000	0.6	784	3-layer MLP with ReLU (d-500-500-1)
EMNIST-Letters	112320	12480	20800	0.6153	784	3-layer MLP with ReLU (d-500-500-1)
EMNIST-Balanced	101520	11280	18800	0.5744	784	3-layer MLP with ReLU (d-500-500-1)
CIFAR-10	54000	6000	10000	0.6	3072	ResNet-34
SVHN	65931	7326	26032	0.7085	3072	ResNet-18

K. Additional Information of Experiments

K.1. Detailed Setup of Figure 2

We generated 500 positive data and 300 negative data according to the 2-dimensional Gaussian distributions with different means and covariance for $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$. The parameters are listed below:

$$\boldsymbol{\mu}_+ = [-4,\ 0]^\top,\ \boldsymbol{\mu}_- = [2, 2]^\top,\ \boldsymbol{\Sigma}_+ = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},\ \boldsymbol{\Sigma}_- = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}.$$

Adam was chosen as the optimizer with default momentum parameters ($\beta_1 = 0.9,\ \beta_2 = 0.999$) and the learning rate, epoch, weight decay, and batch size were fixed to be 1e-1, 30, 1e-3, and 128, respectively.

K.2. Detailed Setup of Synthetic Experiments

In the synthetic experiments in Section 7.1, we generate 4 synthetic datasets to show the validity of our methods. The detailed parameters for generating different synthetic datasets are listed below. $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ are the means for two Gaussian distributions and $\boldsymbol{\Sigma}_+$ and $\boldsymbol{\Sigma}_-$ are the covariance for two Gaussian distributions:

- Setup A: $\boldsymbol{\mu}_+ = [0,\ 0]^\top, \boldsymbol{\mu}_- = [-2,\ 5]^\top, \boldsymbol{\Sigma}_+ = \begin{bmatrix} 7 & -6 \\ -6 & 7 \end{bmatrix}, \boldsymbol{\Sigma}_- = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$
- Setup B: $\boldsymbol{\mu}_+ = [0,\ 0]^\top, \boldsymbol{\mu}_- = [4,\ 0]^\top, \quad \boldsymbol{\Sigma}_+ = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \boldsymbol{\Sigma}_- = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$
- Setup C: $\boldsymbol{\mu}_+ = [0,\ 0]^T, \boldsymbol{\mu}_- = [3,\ -3]^T, \boldsymbol{\Sigma}_+ = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \boldsymbol{\Sigma}_- = \begin{bmatrix} 4 & -3 \\ -3 & 4 \end{bmatrix}.$
- Setup D: $\boldsymbol{\mu}_+ = [0,\ 0]^T, \boldsymbol{\mu}_- = [4,\ 4]^T, \quad \boldsymbol{\Sigma}_+ = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \boldsymbol{\Sigma}_- = \begin{bmatrix} 6 & -5 \\ -5 & 6 \end{bmatrix}.$

K.3. Detailed Setup of Benchmark Experiments

In Section 7.2, we use 8 widely-used large-scale benchmark datasets. The detailed statistics of the datasets and the corresponding models are listed in Table 3: We report the sources of these datasets and the way we corrupt these datasets into binary datasets.

- MNIST (LeCun et al., 1998). It is a grayscale dataset of handwritten digits from 0 to 9, where the size of the images is 28*28. Source: `http://yann.lecun.com/exdb/mnist/`.

  The digits $0 \sim 2$ are used as the positive class and the rest digits are used as the negative class.

- Kuzushiji-MNIST (Clanuwat et al., 2018). It is a 10-class dataset of cursive Japanese characters ('Kuzushiji'). Source: `https://github.com/rois-codh/kmnist`.

  The positive class includes 'O', 'Ki', 'Su', 'Tsu', 'Na', 'Ha', and 'Ma'. The negative class includes 'Ya', 'Re', and 'Wo'.

- Fashion-MNIST (Xiao et al.). It is a 10-class dataset of fashion items. Each instance is a 28*28 grayscale image. Source: https://github.com/zalandoresearch/fashion-mnist.

  'T-short', 'Pullover', 'Dress', and 'Shirt' make up the positive class and the negative class is made up of 'Trouser', 'Coat', 'Sandal', 'Sneaker', 'Bag', and 'Ankle boot'.

- EMNIST (Cohen et al., 2017). A dataset that contain both letters and digits. Source: https://www.westernsydney.edu.au/icns/reproducible_research/publication_support_materials/emnist.

  The splits 'Digits', 'Letters', and 'Balanced' are used and the details of each split are listed below:

  - For 'Digits', $0 \sim 5$ are used as the positive class and $6 \sim 9$ are used as the negative class;
  - For 'Letters', 'a'$\sim$'p' is used as the positive class and 'q'$\sim$'z' are used as the negative class;
  - For 'Balanced', instances with class labels in $[0,\ 26]$ are used as the rest of the instances are used as the negative class.

- CIFAR-10 (Krizhevsky, 2012). It is a 10-class dataset for 10 different objects and each instance is a 32*32*3 colored image in RGB format. Source: https://www.cs.toronto.edu/~kriz/cifar.html.

  'Bird', 'Cat', 'Dog', 'Deer', 'Frog', and 'Horse' form the positive class. The negative class is formed by 'Airplane', 'Automobile', 'Ship', and 'Truck'.

- SVHN (Netzer et al., 2011), a real-world image dataset of digits from 0 to 9. Each instance is a 32*32*3 colored image in RGB format. Source: http://ufldl.stanford.edu/housenumbers/.

  The positive class is composed of digits $0 \sim 5$ and the negative class is composed of $6 \sim 9$.

The hyper-parameters for optimization algorithms are shown below:

Adam with default momentum was used for optimization in this paper. For generating similarity confidence, the epoch number, batch size, and learning rate are 10, 3000, and 1e-2, respectively.

For Sconf-Unbiased, Sconf-ABS, Sconf-NN, and SD, the epoch number, batch size, and weight decay are 60, 3000, and 1e-3, respectively. The initial learning rate was set to 1e-3 and divided by 10 every 20 epochs.

For Siamese and Contrastive, the epoch number, batch size, weight decay, and learning rate are 10, 3000, 1e-3, and 1e-3.