Learning from Similarity-Confidence Data

Yuzhou Cao [1] **Lei Feng** [2 3] **Yitian Xu** [1] **Bo An** [4] **Gang Niu** [3] **Masashi Sugiyama** [3 5]

Abstract

Weakly supervised learning has drawn considerable attention recently to reduce the expensive time and labor consumption of labeling massive data. In this paper, we investigate a novel weakly supervised learning problem called *learning from similarity-confidence (Sconf) data*, where we aim to learn an effective binary classifier from only unlabeled data pairs equipped with confidence that illustrates their degree of similarity (two examples are similar if they belong to the same class). To solve this problem, we propose an unbiased estimator of the classification risk that can be calculated from only Sconf data and show that the estimation error bound achieves the optimal convergence rate. To alleviate potential overfitting when flexible models are used, we further employ a risk correction scheme on the proposed risk estimator. Experimental results demonstrate the effectiveness of the proposed methods.

1. Introduction

In supervised classification, a vast quantity of exactly labeled data are required for training effective classifiers. However, the collection of massive data with exact supervision is laborious and expensive in many real-world problems. To overcome this bottleneck, *weakly supervised learning* (Zhou, 2018) has been proposed and explored under various settings, including but not limited to, *semi-supervised learning* (Chapelle et al., 2006; Zhu & Goldberg, 2009; Niu et al., 2013; Li & Zhou, 2015; Sakai et al., 2017; Li & Liang, 2019; Guo et al., 2020), *positive-unlabeled learning* (Elkan & Noto, 2008; du Plessis et al., 2014; 2015; Kiryo et al., 2017; Sansone et al., 2019), *noisy-label learning* (Natarajan et al., 2013; Han et al., 2018b;a; Zhang et al., 2019; Xia et al., 2019; 2020), *partial-label learning* (Cour et al., 2011; Wang et al., 2019; Feng et al., 2020b; Lv et al., 2020), *complementary-label learning* (Ishida et al., 2017; Yu et al., 2018; Ishida et al., 2019; Feng et al., 2020a; Katsura & Uchida, 2020; Chou et al., 2020), *similarity-unlabeled learning* (Bao et al., 2018), and *similarity-dissimilarity learning* (Shimada et al., 2019).

In this paper, we consider a novel weakly supervised learning setting called *similarity-confidence (Sconf) learning*. Under this setting, we aim to train a binary classifier from only unlabeled data pairs equipped with *similarity confidence* that demonstrates the degree of their pairwise similarity, i.e., the possibility that two data points share the same label, without any ordinarily labeled data. The Sconf learning setting exists in many practical scenarios. Compared with ordinary class labels, similarity labels are more easily accessible in many applications (e.g., protein function prediction (Klein et al., 2002)) and can alleviate potentially biased decisions (Fisher, 1993). However, such similarity labels could cause severe privacy leakage: for a data pair equipped with a similarity label, the disclosure of the class label of either of the two examples can subsequently reveal the class label of another one. When the collected data are sensitive (e.g., political opinions and religious orientations), such leakage will lead to serious consequences. In this scenario, similarity confidence is more favorable in the sense of privacy-preserving: given the similarity confidence of a data pair, people are uncertain if they share the same label because the confidence only gives the *probability* that they belong to the same class, e.g., even if the similarity confidence is 70%, we still cannot assert that the data pair shares the same label since the data pair can also be dissimilar with possibility 30%. As a result, it is unable to exactly figure out the underlying similarity label of the data pair from only similarity confidence.

Another example is crowdsourcing (Howe, 2009). When the data are annotated by crowdworkers, it is difficult for us to always obtain high-quality crowdsourcing labels (Wang & Zhou, 2015) due to the crowdworkers' lack of domain knowledge. When a data pair is annotated with both pairwise similarity and dissimilarity labels by different crowdworkers, we can generate the similarity confidence by averaging instead of choosing the majority of crowdsourcing

[1]College of Science, China Agricultural University, Beijing, China [2]College of Computer Science, Chongqing University, Chongqing, China [3]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan [4]Nanyang Technological University, School of Computer Science and Engineering, Singapore [5]The University of Tokyo, Tokyo, Japan. Correspondence to: Yitian Xu <xytshuxue@126.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

labels, which can alleviate noisy supervision. In these scenarios, Sconf learning makes it possible to learn an effective binary classifier from only unlabeled data pairs equipped with similarity confidence instead of hard labels.

Our main contributions in this paper are the following:

- We propose a novel Sconf learning framework (in Section 5) that allows the use of ERM by constructing an unbiased estimator of the classification risk with only unlabeled data pairs with similarity confidence, where any loss functions, models, and optimizers are applicable in this setting.
- In Section 5, we derive an estimation error bound for Sconf learning and show that it achieves the optimal parametric convergence rate. Analysis of the influence of noisy confidence also shows the robustness of our Sconf learning framework.
- We leverage an effective empirical risk correction scheme (Kiryo et al., 2017; Lu et al., 2020) for correcting the obtained unbiased risk estimator to alleviate potential overfitting and further show the consistency of its risk minimizer (in Section 6).
- Extensive experiments on various datasets and deep neural networks clearly demonstrate the effectiveness of the proposed Sconf learning method and risk correction scheme (in Section 7).

2. Related Work

We illustrate Sconf learning and related problems in Figure 1. In what follows, we briefly review *semi-supervised clustering* and *similarity-based learning*.

The research on similarity-based learning was pioneered by the semi-supervised clustering (SSC) paradigm, where pairwise similarity/dissimilarity is utilized to enhance the clustering performance (Wagstaff et al., 2001; Basu et al., 2002; Xing et al., 2002; Niu et al., 2012; Yi et al., 2013). From the learning theory viewpoint, the SSC methods are confined in the clustering setting and have no generalization guarantee.

Recently, many studies have tried to solve the similarity-based learning problem by *empirical risk minimization* (ERM) with rigorous consistency analysis. In Bao et al. (2018), it was shown that the classification risk can be recovered from similar data pairs and unlabeled data, which enables the use of ERM and analysis on the estimation error. However, the dissimilar data pairs are ignored in this work and the collection of additional unlabeled data is inevitable.

Later, Shimada et al. (2019) made it possible to learn from both similar and dissimilar data pairs by ERM, yet it is still confined within the hard-label setting. Bao et al. (2020) introduced a new performance metric for the binary discriminative model and developed a surrogate risk minimization framework with both similar and dissimilar data pairs.

On the other hand, the likelihood-based models (Hsu et al., 2019; Wu et al., 2020) were proposed to conduct similarity-based learning for multi-class classification tasks. The loss functions in these methods are fixed and we cannot directly optimize the classification-oriented losses in Bao et al. (2020). Compared with these works, our proposed Sconf learning framework is assumption-free on models, loss functions, and optimizers, which makes it a flexible framework when we use deep learning.

3. Preliminaries

In this section, we first briefly review the ordinary classification problem and then show our problem setting where each unlabeled data pair is merely equipped with similarity confidence. Proofs are presented in supplementary materials.

3.1. Ordinary Classification Problem

Suppose that the feature space is $\mathcal{X} \subset \mathbb{R}^d$ and the label space is $\mathcal{Y}=\{-1,+1\}$, the instance and its ordinary class label $(\boldsymbol{x}, y)$ obey an unknown distribution with density $p(\boldsymbol{x}, y)$. Then the critical work is to find a decision function $g(\cdot): \mathcal{X} \to \mathbb{R}$ that minimizes the classification risk:

$$R(g) = \mathbb{E}_{p(\boldsymbol{x},y)}[\ell(g(\boldsymbol{x}), y)], \tag{1}$$

where $\ell(\cdot,\cdot): \mathbb{R} \times \mathcal{Y} \to \mathbb{R}^+$ is a binary loss function, e.g., the 0-1 loss and logistic loss. An equivalent expression of classification risk (1) used in the following sections is:

$$R(g) = \underbrace{\pi_+ \mathbb{E}_+[\ell(g(\boldsymbol{x}), +1)]}_{R_+(g)} + \underbrace{\pi_- \mathbb{E}_-[\ell(g(\boldsymbol{x}), -1)]}_{R_-(g)}$$

where $\pi_+ = p(y=+1)$, $\pi_- = p(y=-1)$ denote the class prior probabilities. $\mathbb{E}_+[\cdot]$ and $\mathbb{E}_-[\cdot]$ are expectations on class-conditional distributions with densities $p_+(\boldsymbol{x}) = p(\boldsymbol{x}|y=+1)$ and $p_-(\boldsymbol{x}) = p(\boldsymbol{x}|y=-1)$, respectively. The class posterior probabilities are denoted by $r_+(\boldsymbol{x}) = p(y=+1|\boldsymbol{x})$ and $r_-(\boldsymbol{x}) = p(y=-1|\boldsymbol{x})$.

3.2. Generation of Similarity-Confidence Data

To conduct ERM with only Sconf data, we first give the underlying distributions of Sconf data pairs and further discuss the expression and property of similarity confidence.

In this setting, we only have access to the unlabeled data pairs with **similarity confidence**: $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{x}'_i), s_i\}_{i=1}^n$, where the similarity confidence $s_i = s(\boldsymbol{x}_i, \boldsymbol{x}'_i) = p(y_i = y'_i|\boldsymbol{x}_i, \boldsymbol{x}'_i)$ denotes the probability that $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$ share the same label $y_i = y'_i$. Each unlabeled data pair in $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}_{i=1}^n$ is drawn independently from a simple distribution U^2 with density $p(\boldsymbol{x}, \boldsymbol{x}') = p(\boldsymbol{x})p(\boldsymbol{x}')$ and we

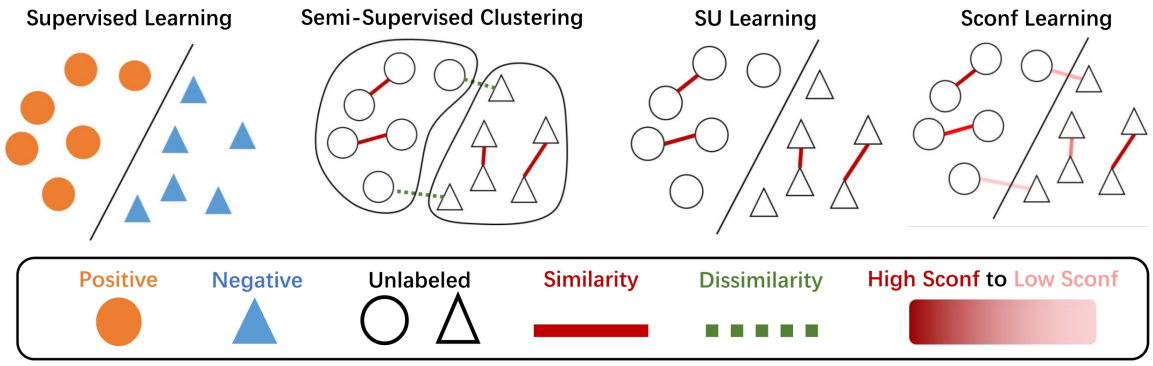


Figure 1. Illustration of Sconf learning and related problems.

further denote by $\mathcal{S}_n$ the unlabeled data pairs with similarity confidence $\{(\boldsymbol{x}_i, \boldsymbol{x}_i')\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, \boldsymbol{x}')$. This formulation implies that we can regard the decoupled unlabeled samples $\{\boldsymbol{x}_i\}_{i=1}^n \cup \{\boldsymbol{x}_i'\}_{i=1}^n$ as drawn from the marginal distribution $p(\boldsymbol{x})$ independently, which can be easily implemented in real-world data collection. We also assume the sample independence: $(\boldsymbol{x}, y) \perp (\boldsymbol{x}', y')$, which is an implicit assumption used in Bao et al. (2018). Furthermore, we show that the similarity confidence has the following property:

Lemma 1. *(Equivalent expression of similarity confidence)*

$$s(\boldsymbol{x}, \boldsymbol{x}') = \frac{\pi_+^2 p_+(\boldsymbol{x}) p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}) p_-(\boldsymbol{x}')}{p(\boldsymbol{x}) p(\boldsymbol{x}')}. \tag{2}$$

4. Failure of Learning with One-Sided Similarity Relation

As mentioned in Section 1, though we can recover classification risk from both similar and dissimilar data (Shimada et al., 2019), such a type of hard label could cause serious privacy leakage, which indicates that it is not favorable when the data are sensitive. Such leakage may be alleviated by learning from only one-sided similarity relation: we only have similar (dissimilar) data pairs and no dissimilar (similar) data pairs are provided. For example, when investigating political or religious orientations, people with dissimilar opinions may refuse to give the answer in case of potential conflicts. In these scenarios, only similar data pairs are accessible. As reported in Hadsell et al. (2006), learning with only similar data pairs can lead to degenerated solutions. A natural optional idea is to combine one-sided similarity relation with similarity confidence.

Can we learn an effective classifier from only one-sided similarity relation and similarity confidence? Unfortunately, the following experimental and theoretical results give a negative answer to this question. Due to the space limitation, we only provide the result when we only have similar data pairs, and a completely analogous result with only dissimilar data pairs is listed in the supplementary materials.

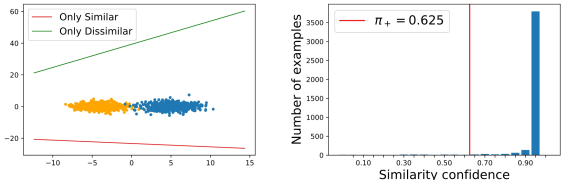


Figure 2. Decision boundaries and the distribution of similarity confidence.

Suppose we have a training set including only similar data pairs, i.e., $\mathcal{S}_{\mathrm{S}}^n = \{(\boldsymbol{x}_i, \boldsymbol{x}_i')\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, \boldsymbol{x}' | y = y')$ and their similarity confidences $\{s_i\}_{i=1}^n$. The following theorem shows that it is *theoretically* possible to conduct ERM with only data provided above:

Theorem 1. *With similar data pairs and their similarity confidence, assuming that $s(\boldsymbol{x}, \boldsymbol{x}') > 0$ for all the pair $(\boldsymbol{x}, \boldsymbol{x}')$, we can get the unbiased estimator of classification risk (1), i.e., $\mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{x}' | y=y')}[\hat{R}_{\mathrm{S}}(g)] = R(g)$, where*

$$\hat{R}_{\mathrm{S}}(g) = (\pi_+^2 + \pi_-^2) \sum_{i=1}^n \frac{(s_i - \pi_-)(\ell(g(\boldsymbol{x}_i), +1) + \ell(g(\boldsymbol{x}_i'), +1))}{2n(\pi_+ - \pi_-) s_i}$$
$$+ (\pi_+^2 + \pi_-^2) \sum_{i=1}^n \frac{(\pi_+ - s_i)(\ell(g(\boldsymbol{x}_i), -1) + \ell(g(\boldsymbol{x}_i'), -1))}{2n(\pi_+ - \pi_-) s_i}. \tag{3}$$

It seems that we can conduct ERM on the obtained unbiased risk estimator to get a binary classifier. Unfortunately, with only one-sided similarity relation, we can only get collapsed solutions *empirically*. Denote the empirical risk minimizer of Eq. (3) with $\hat{g}_{\mathrm{S}}$. Then we come to the following conclusion:

Theorem 2. *Suppose $\pi_+ > \pi_-$ and 0-1 loss is used. For similar data pairs, we assume that $s_i \geq \pi_+$ for $i = 1, \cdots, n$. Then $\hat{g}_{\mathrm{S}}$ is a collapsed solution that classifies all the examples as positive.*

A rough proof intuition for Theorem 2 is that the coefficients of positive loss terms are always positive and those of negative loss terms are always negative, then minimizing Eq. (3)

is equivalent to minimizing the positive counterpart and maximizing the negative counterpart of classification risk, which can lead to the collapsed solution. We can conclude that though Eq. (3) is unbiased, it cannot well represent the classification risk in Eq. (1).

To empirically illustrate the failure of learning with one-sided similarity relation, we conducted experiments on a synthetic dataset and show the distribution of similarity confidence. The detailed statistics of the synthetic dataset are provided in the supplementary materials. According to the experimental results in Figure 2, the empirical minimizers of learning with only similar or dissimilar data pairs yield collapsed results and their classification boundaries are severely biased, which aligns with Theorem 2. The distribution of similarity confidence also meets our assumption on $\{s_i\}_{i=1}^n$.

As shown above, the incorporation of one-sided similarity relation can lead to collapsed solution due to the highly skewed distribution of similarity confidence. A potential remedy for such failure is an underlying non-skewed distribution of similarity confidence. Fortunately, we can achieve this goal with only unlabeled data pairs. In the following section, we show that given unlabeled data pairs with similarity confidence, the hard similarity labels are all completely unnecessary, which means that we can successfully train an effective classifier from only unlabeled data pairs with similarity confidence.

5. Learning from Similarity-Confidence Data

In this section, we propose an unbiased risk estimator for learning from only unlabeled data pairs with similarity confidence and show the consistency of the proposed estimator by giving its estimation error bound. Finally, we propose an effective class-prior estimator for estimating π_+ when it is not given in advance. An analysis of the influence of inaccurate similarity confidence is also provided by giving a high-probability bound.

5.1. Unbiased Risk Estimator with Sconf Data

In this section, we derive an unbiased estimator of the classification risk in Eq. (1) with only Sconf data and establish its risk minimization framework.

Based on the settings in Section 3.2, we first derive the crucial lemma before deriving the unbiased estimator of classification risk (1) from only Sconf data:

Lemma 2. *The following equalities hold:*

$R_+(g) = \mathbb{E}_{U^2}[\hat{R}_+(g)]$, $R_-(g) = \mathbb{E}_{U^2}[\hat{R}_-(g)]$, *where*

$$\hat{R}_+(g) = \sum_{i=1}^n \frac{(s_i-\pi_-)(\ell(g(\boldsymbol{x}_i),+1)+\ell(g(\boldsymbol{x}'_i),+1))}{2n(\pi_+-\pi_-)}, \quad (4)$$

$$\hat{R}_-(g) = \sum_{i=1}^n \frac{(\pi_+-s_i)(\ell(g(\boldsymbol{x}_i),-1)+\ell(g(\boldsymbol{x}'_i),-1))}{2n(\pi_+-\pi_-)}. \quad (5)$$

According to Lemma 2 above, we get the unbiased estimator of each counterpart of the classification risk in Eq. (1). Then we can simply derive the unbiased estimator of the classification risk in Eq. (1) with only Sconf data:

Theorem 3. *We can construct the unbiased estimator $\hat{R}(g)$ of the classification risk (1), i.e., $\mathbb{E}_{U^2}[\hat{R}(g)] = R(g)$, with only Sconf data as in the formulation below:*

$$\hat{R}(g) = \hat{R}_+(g) + \hat{R}_-(g). \quad (6)$$

Since there are no implicit assumptions on models, losses, and optimizers in our analysis, any convex/non-convex loss and deep/linear model can be used for Sconf learning.

5.2. Estimation Error Bound

Here we show the consistency of proposed risk estimator $\hat{R}(g)$ in Eq. (6) by giving an estimation error bound. To begin with, let $\mathcal{G}$ be our function class for ERM. Assume there exists $C_g > 0$ that $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq C_g$ and $C_\ell > 0$ such that $\sup_{|z|\leq C_g} \ell(z, y) \leq C_\ell$ holds for all y. Following the usual practice (Mohri et al., 2012), we assume $\ell(z, y)$ is Lipschitz continuous *w.r.t.* z for all $|z| \leq C_g$ and all y with a Lipschitz constant L_ℓ.

Let $g^* = \arg\min_{g\in\mathcal{G}} R(g)$ be the minimizer of classification risk in Eq. (1), and $\hat{g} = \arg\min_{g\in\mathcal{G}} \hat{R}(g)$ be the minimizer of empirical risk in Eq. (6). Then we can derive the following estimation error bound for Sconf learning:

Theorem 4. *For any $\delta > 0$, the following inequality holds with probability at least $1-\delta$:*

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

where $\mathfrak{R}_n(\mathcal{G})$ is the Rademacher complexity of $\mathcal{G}$ for unlabeled data of size n drawn from the marginal distribution with density $p(\boldsymbol{x})$.

The definition of the Rademacher complexity (Bartlett & Mendelson, 2001) is provided in the supplementary material. Note that the estimation error bound converges in the rate of $\mathcal{O}_p(1/\sqrt{n})$ if we assume that $\mathfrak{R}_n(\mathcal{G}) \leq C_\mathcal{G}/\sqrt{n}$, where $\mathcal{O}_p$ denotes the order in probability and $C_\mathcal{G}$ is a non-negative constant determined by the model complexity. This is a natural assumption since many model classes (e.g., linear-in-parameter models and fully-connected neural network (Golowich et al., 2018)) satisfy this condition. We make this assumption in the rest of this paper.

Theorem 4 shows that the empirical risk minimizer converges to the classification risk minimizer with high-probability in the rate of $\mathcal{O}_p(1/\sqrt{n})$. As shown in Mendelson (2008), this is the optimal parametric convergence rate without additional assumptions.

5.3. Class-Prior Estimation from Similarity Confidence

In our Sconf learning, class-prior π_+ plays an important role in the construction of the unbiased risk estimator. Compared with the previous work (Bao et al., 2018; Shimada et al., 2019), we make a milder assumption on the data distribution, which aligns with the practical data collection process. However, when the class-prior π_+ is not given, we cannot estimate it by mixture proportion estimation (Scott, 2015) as in Bao et al. (2018) since we only have data drawn from a single distribution U^2. In this section, we propose a simple yet effective class-prior estimator with only Sconf data.

We have the following theorem for the sample average of similarity confidence:

Theorem 5. *The sample average of similarity confidence is an unbiased estimator of $\pi_+^2 + \pi_-^2$:*

$$\mathbb{E}_{S_n}\left[\frac{\sum_{i=1}^n s(\boldsymbol{x}_i, \boldsymbol{x}_i')}{n}\right] = \pi_+^2 + \pi_-^2.$$

Furthermore, according to McDiarmid's inequality (McDiarmid, 1989), the sample average of similarity confidence is consistent and converges to $\pi_+^2 + \pi_-^2$ in the rate of $\mathcal{O}_p(\exp(-n))$, which is the optimal rate according to the central limit theorem (Chung, 1974).

Let us denote $\pi_+^2 + \pi_-^2$ by π_{S}. Assuming $\pi_+ > \pi_-$, we can calculate the class prior by $\pi_+ = \frac{\sqrt{2\pi_{\mathrm{S}} - 1} + 1}{2}$. According to Theorem 5, we can approximate π_+ with the average of similarity confidence and the formulation above.

5.4. Analysis with Noisy Similarity Confidence

In the previous sections, we assumed that accurate confidence is accessible. However, this assumption may not be realistic in some practical tasks. We may have the question that how the noisy similarity confidence can affect the learning performance? If our Sconf learning is not robust and even a slight noise on the similarity confidence can cause catastrophic degradation of performance? In this section, we theoretically justify that the Sconf learning framework is robust to noise on similarity confidence by bounding the estimation error of learning with noisy confidence.

Suppose we are given the noisy Sconf data pairs: $\bar{\mathcal{S}}_n = \{(\boldsymbol{x}_i, \boldsymbol{x}_i'), \bar{s}_i\}_{i=1}^n$, where $\bar{s}_i$ is the noisy similarity confidence and is not necessary equal to $s(\boldsymbol{x}_i, \boldsymbol{x}_i')$ (in fact, it can take the form of any real number in $[0,\ 1]$). For simplicity, we replace the accurate confidence $\{s_i\}_{i=1}^n$ in Eq. (6) with noisy ones $\{\bar{s}_i\}_{i=1}^n$ and denote the noisy empirical risk with $\bar{R}(g)$. The minimizer of noisy risk is $\bar{g} = \arg\min_{g \in \mathcal{G}} \bar{R}(g)$. To quantify the influence of noisy similarity confidence, we deduce the following estimation error bound:

Theorem 6. *For any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:*

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

where $\sigma_n = \sum_{i=1}^n |s_i - \bar{s}_i|$ is the summation of the deviation of noisy similarity confidence.

In a straightforward way, the deduced estimation error bound demonstrates the magnitude of the influence of noisy similarity confidence: the estimation error of $\bar{g}$ is affected up to the mean absolute error of noisy confidence and the noisy confidence only has limited influence on the performance of Sconf learning. If the summation of noise σ_n has a sublinear growth rate in high probability, Sconf learning can even remain consistent, which shows that our Sconf learning framework is robust to the noisy confidence.

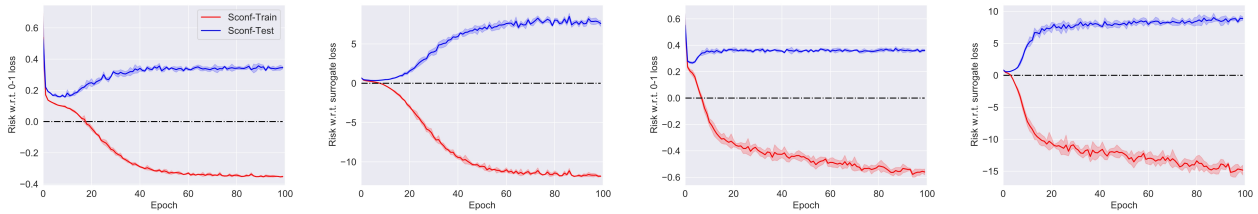
6. Consistent Risk Correction

In the previous section, we showed the unbiased risk estimator that can recover the classification risk in Eq. (1) from only Sconf data with rigorous consistency analysis. It is noticeable that the positive and negative counterparts of the empirical risk, i.e., $\hat{R}_+(g)$ and $\hat{R}_-(g)$, are not bounded below and can go negative, while their expectations are non-negative by definition. This contradiction can be problematic since as in previous works (Kiryo et al., 2017; Lu et al., 2020) that severe overfitting usually occurs when the empirical risk goes negative, especially when flexible models (e.g., deep models) are used. This phenomenon can be also observed in Sconf learning, as shown in Figure 3. The detailed setting of optimization algorithm is provided in supplementary materials.

In this section, we alleviate this problem with a simple yet effective *risk correction* on the proposed estimator (6). We further show that the proposed corrected estimator can preserve its consistency by bounding its estimation error.

6.1. General Risk Formulation

Can we alleviate the overfitting in Sconf learning without collecting more data or changing the model? Here we give a positive answer to this question by giving a slightly modified empirical risk estimator. Since the overfitting is caused by negative empirical risk, it is a natural idea to make a correction on the risk estimator when it goes negative. This idea was first proposed in Kiryo et al. (2017), where the data that yield a negative risk are ignored by applying a non-negative risk estimator. Lu et al. (2020) further showed



(a) Kuzushiji-MNIST, MLP (ReLU, d-500-500-1) (b) CIFAR-10, ResNet-34

Both Kuzushiji-MNIST (Clanuwat et al., 2018) and CIFAR-10 (Krizhevsky, 2012) are manually corrupted into binary classification datasets. In (a), we trained a 3-layer *multi-layer perceptron* (MLP) with ReLU (Nair & Hinton, 2010) on Kuzushiji-MNIST. In (b), ResNet-34 (He et al., 2016) was trained on CIFAR-10. Adam (Kingma & Ba, 2015) was chosen as the optimization algorithm. Logistic loss was used as the loss function. The generation of similarity confidence and the details of corrupted datasets are the same as those in Section 7. The occurrence of overfitting and negative empirical risk is almost simultaneous: when the empirical risk (red line) goes negative, the risk on test set (blue line) stops dropping and increases rapidly.

Figure 3. Illustration of the overfitting of unbiased Sconf learning.

that the information in those data can be helpful for generalization and should not be dropped. Based on the previous works, we propose the *consistently corrected risk estimator* for Sconf learning to enforce the non-negativity:

Definition 1. *(Lu et al., 2020) A risk estimator $\widetilde{R}$ is called the consistently corrected risk estimator if it takes the following form:*

$$\widetilde{R}(g) = f\big(\hat{R}_+(g)\big) + f\big(\hat{R}_-(g)\big), \tag{7}$$

where $f(x) = \begin{cases} x, & x \geq 0, \\ k|x|, & x < 0. \end{cases}$ *and* $k > 0$.

Denote the minimizer of consistently corrected Sconf risk estimator (7) with $\tilde{g} = \arg\min_{g\in\mathcal{G}} \widetilde{R}(g)$, which can be obtained by ERM. Two representative correction functions are **N**on-**N**egative correction (Kiryo et al., 2017) and **ABS**olute function (Lu et al., 2020), with $k = 0$ and 1 respectively. Their explicit formulations are shown below:

$$\widetilde{R}_{\mathrm{NN}}(g) = \max\left\{0, \hat{R}_+(g)\right\} + \max\left\{0, \hat{R}_-(g)\right\}, \tag{8}$$

$$\widetilde{R}_{\mathrm{ABS}}(g) = \left|\hat{R}_+(g)\right| + \left|\hat{R}_-(g)\right|. \tag{9}$$

In Section 7, we will experimentally show their efficiency in alleviating overfitting.

6.2. Consistency Guarantee

It is noticeable that $\widetilde{R}(g)$ is an upper bound of the unbiased risk estimator $\hat{R}(g)$ for any fixed classifier g, which means that $\widetilde{R}(g)$ is generally *biased* and does not align with the consistency analysis in the previous section. Here we justify the use of ERM by analyzing the consistency of $\widetilde{R}(g)$ and its minimizer $\tilde{g}$. We first show that the corrected estimator is consistent and the bias decays exponentially.

Theorem 7. *(Consistency of $\widetilde{R}(g)$) Assume that there are $\alpha > 0$ and $\beta > 0$ such that $R_+(g) \geq \alpha$ and $R_-(g) \geq \beta$. According to the assumptions in Theorem 4, the bias of $\widetilde{R}(g)$ decays exponentially as $n \to \infty$:*

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

where $\Delta = \exp(\alpha^2) + \exp(\beta^2)$ and $L_f = \max\{1, k\}$ is the Lipschitz constant of $f(\cdot)$. Furthermore, with probability at least $1-\delta$:

$$|\widetilde{R}(g) - R(g)| \leq \frac{L_\ell C_\ell}{|\pi_+ - \pi_-|}\sqrt{\frac{\ln 2/\delta}{2n}} + \frac{(L_f+1)C_\ell}{|\pi_+ - \pi_-|} \exp\left(-\frac{(\pi_+ - \pi_-)^2 n}{2C_\ell^2}\right) \Delta.$$

Based on Theorem 7, we show that the empirical risk minimizer $\tilde{g}$ obtained by ERM converges to g^* in the same rate of $\mathcal{O}_p(1/\sqrt{n})$.

Theorem 8. *(Estimation error bound of $\tilde{g}$) Based on the assumptions and notations above, with probability at least $1-\delta$:*

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

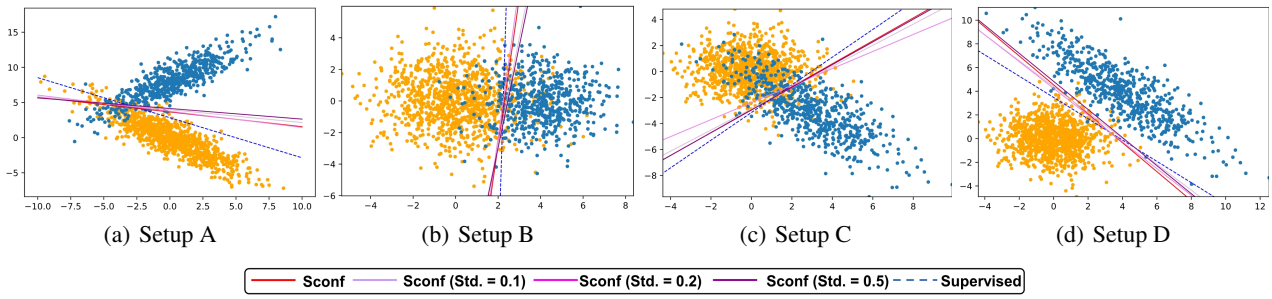
Theorem 8 shows that learning with $\widetilde{R}(g)$ is also consistent and it has the same convergence rate as learning with $R(g)$ since the additional exponential term is of lower order.

7. Experiments

In this section, we demonstrate the usefulness of proposed methods on both synthetic and benchmark datasets with data generation process in Section 5. Sconf-Unbiased, Sconf-ABS and Sconf-NN are short for ERM with risk estimators in Eqs. (6), (8), and (9), respectively.

Table 1. Mean and standard deviation of the classification accuracy over 5 trails in percentage with linear models on various synthetic datasets. Std. is the standard deviation of Gaussian noise. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Setup	Sconf	Sconf (std. = 0.1)	Sconf (std. = 0.2)	Sconf (std. = 0.3)	Supervised
A	$\mathbf{89.91 \pm 0.18}$	89.84 ± 0.03	$\mathbf{89.55 \pm 0.64}$	89.64 ± 0.33	89.66 ± 0.41
B	$\mathbf{90.62 \pm 0.28}$	$\mathbf{90.34 \pm 0.40}$	90.03 ± 0.33	89.49 ± 0.92	90.71 ± 0.17
C	$\mathbf{88.05 \pm 0.30}$	$\mathbf{88.14 \pm 0.14}$	$\mathbf{87.92 \pm 0.57}$	$\mathbf{87.91 \pm 0.38}$	88.14 ± 0.16
D	$\mathbf{90.43 \pm 0.14}$	$\mathbf{90.29 \pm 0.30}$	$\mathbf{90.40 \pm 0.15}$	90.20 ± 0.14	90.56 ± 0.16



(a) Setup A (b) Setup B (c) Setup C (d) Setup D

Figure 4. Illustration of Sconf learning on different scales of noise and Gaussian distributions on a single trail.

7.1. Synthetic Experiments

We experimentally characterize the behavior of Sconf learning and show its robustness to noisy confidence on the synthetic datasets.

Setup: We generated the positive and negative data according to the 2-dimensional Gaussian distributions with different means and covariance for $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$. The setups of data generation distributions are provided in the supplementary material.

500 positive data and 300 negative data were generated independently from each distribution for training. We dropped the class labels for Sconf learning and generated the unlabeled data pair according to the data generation process in Section 5. Then we analytically computed the class posterior probability $r_+(\boldsymbol{x})$ from the two Gaussian densities and equipped the unlabeled data pairs with true similarity confidence, which was obtained based on Lemma 1. 1000 positive data and 600 negative data were generated in the same way for testing.

The linear-in-input model $f(\boldsymbol{x}) = \boldsymbol{\omega}^\top \boldsymbol{x} + b$ and logistic loss were used. We trained the model with Adam for 100 epochs (full-batch size) and default momentum parameter. The learning rate was initially set to 0.1 and divided by 10 every 30 epochs. To generate noisy similarity confidence, we added zero-mean Gaussian noise with different scales of standard deviation chosen from $\{0.1\ , 0.2,\ 0.3\}$ on the obtained similarity confidence. When the noisy similarity confidence was over 1 or below 0, we clipped it to 1 or rounded up to 0, respectively. The results of fully supervised learning are also provided.

Experimental results: The results are shown in Table 1 and Figure 4. Compared with fully supervised learning, Sconf learning has similar accuracy on all synthetic datasets. The decline in performance under different scales of noise is not significant, which shows the robustness of Sconf learning.

7.2. Benchmark Experiments

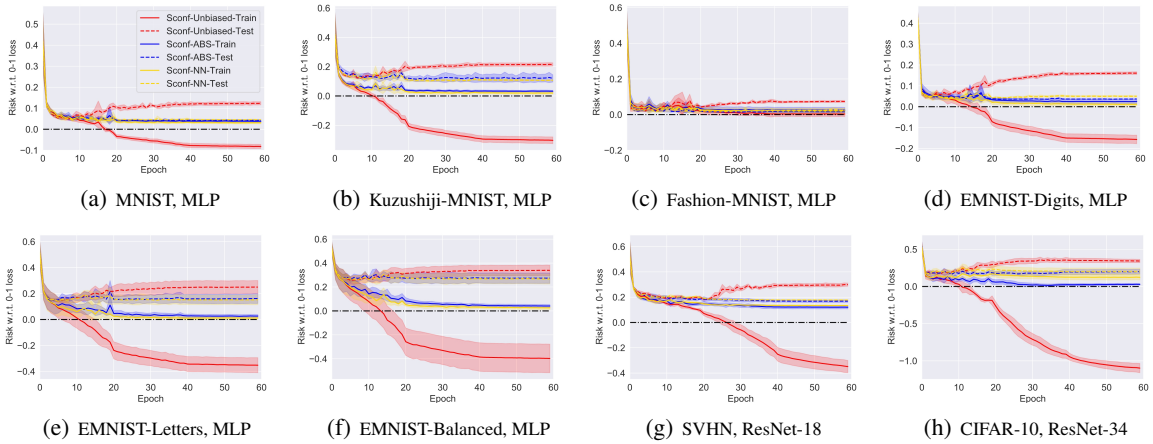
Here we conducted experiments with deep neural networks on the more realistic benchmark datasets.

Datasets: We evaluated the performance of proposed methods on six widely-used benchmarks MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al.), Kuzushiji-MNIST (Clanuwat et al., 2018), EMNIST (Cohen et al., 2017), SVHN (Netzer et al., 2011), and CIFAR-10 (Krizhevsky, 2012). Following Lu et al. (2020), we manually corrupted the multi-class datasets into binary classification datasets. The detailed statistics of datasets are in the supplementary materials.

Baselines: We compared our methods with both statistical learning-based and representation learning-based similarity learning baselines, including similarity-dissimilarity learning (SD) (Shimada et al., 2019), Siamese network (Koch et al., 2015), and contrastive loss (Hadsell et al., 2006). Since we can only get the vector representation rather than class label using Siamese network and contrastive loss, we adopted the one-shot setting in (Koch et al., 2015) and randomly chose two samples with different labels as the prototypes. Then prediction is determined according to the

Table 2. Mean and standard deviation of the classification accuracy over 5 trials in percentage with deep models. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Datasets	Proposed			Baselines		
	Sconf-Unbiased	Sconf-ABS	Sconf-NN	SD	Siamese	Contrastive
MNIST	87.22 ± 2.11	$\mathbf{96.12 \pm 2.31}$	$\mathbf{96.04 \pm 1.23}$	86.57 ± 0.78	55.08 ± 3.94	71.91 ± 2.39
Kuzushiji-MNIST	78.12 ± 3.08	$\mathbf{89.25 \pm 1.58}$	$\mathbf{90.00 \pm 0.55}$	76.42 ± 4.09	59.82 ± 6.15	67.18 ± 5.41
Fashion-MNIST	86.28 ± 7.03	$\mathbf{91.44 \pm 0.39}$	$\mathbf{91.37 \pm 0.30}$	83.61 ± 8.94	58.29 ± 4.42	64.97 ± 5.76
EMNIST-Digits	87.96 ± 1.67	$\mathbf{96.62 \pm 0.06}$	96.21 ± 0.11	76.18 ± 11.21	53.08 ± 2.55	66.37 ± 5.40
EMNIST-Letters	77.14 ± 3.71	$\mathbf{86.32 \pm 1.20}$	$\mathbf{86.72 \pm 1.39}$	76.18 ± 11.21	55.76 ± 3.95	60.29 ± 3.14
EMNIST-Balanced	68.61 ± 10.21	$\mathbf{74.94 \pm 2.92}$	$\mathbf{74.83 \pm 3.40}$	64.03 ± 14.66	52.61 ± 1.22	58.30 ± 2.19
CIFAR-10	65.68 ± 5.03	$\mathbf{84.71 \pm 1.41}$	$\mathbf{84.49 \pm 1.14}$	60.39 ± 6.56	59.83 ± 2.75	54.38 ± 1.48
SVHN	72.88 ± 3.15	$\mathbf{83.51 \pm 0.65}$	82.37 ± 0.23	71.48 ± 5.43	60.90 ± 5.01	69.26 ± 2.97



(a) MNIST, MLP (b) Kuzushiji-MNIST, MLP (c) Fashion-MNIST, MLP (d) EMNIST-Digits, MLP

(e) EMNIST-Letters, MLP (f) EMNIST-Balanced, MLP (g) SVHN, ResNet-18 (h) CIFAR-10, ResNet-34

Figure 5. Experimental Results of proposed methods. Dark colors show the mean accuracy and light colors show the standard deviation.

similarity between each test instance and the prototypes.

Experimental setup: We trained the proposed methods and baseline methods with the same model on a certain dataset with logistic loss. Different models are used on different datasets as summarized in Figure 5. Since Siamese network and contrastive loss is representation learning-based, the output dimensions for them were changed to 300 on MNIST, Kuzushiji-MNIST, Fashion-MNIST, EMNIST and further increased to 1000 on CIFAR-10 and SVHN. For all the methods, the optimization algorithm was chosen to be Adam (Kingma & Ba, 2015) and the detailed setting is listed in supplementary materials. For ERM-based methods: Sconf-Unbiased, Sconf-ABS, Sconf-NN, and SD, the validation accuracy was also calculated according to their empirical risk estimators on a validation set consisted of Sconf data, which means that we do not have to collect additional ordinarily labeled data for validation when using ERM-based methods.

To simulate real similarity confidence, We generated the class posterior probability $p(y=+1|\boldsymbol{x})$ using logistic regression with the same network for each dataset, and obtained the similarity confidence according to Lemma 1. Since the baseline methods requires data with hard similarity labels, we generated the similarity label for each data pair according to the Bernoulli distribution determined by their similarity confidence. Note that we ask the labelers for similarity confidence values in real-world Sconf learning, but we generated them through a probability classifier here. The class labels are only used for estimating the similarity confidence and the test sets are not used in any process of experiments except for testing steps.

We implemented all the methods by Pytorch (Paszke et al., 2019), and conducted the experiments on NVIDIA Tesla P4 GPUs. Experimental results are reported in Figure 5 and Table 2.

Experimental results: It can be observed from Table 2 that the proposed methods: Sconf-Unbiased, Sconf-ABS, and Sconf-NN outperformed the baseline methods on all the datasets. Among all the methods, Siamese network and contrastive loss performed poorly since they are representation learning-based rather than classification-oriented. Though the goal of SD learning is classification, it failed to compete with the proposed methods because it can only utilize the similarity labels degenerating from similarity confidence, which can cause the loss of supervision information.

The efficiency of the risk correction schemes on mitigat-

ing overfitting is illustrated in Figure 5 and Table 2. For Sconf-Unbiased, the learning curves in Figure 5 show that when the empirical risk of Sconf-Unbiased (red full line) goes negative, the test loss increases rapidly, which indicates the occurrence of overfitting. As a consequence, the performance of Sconf-Unbiased became catastrophic compared with Sconf-ABS and Sconf-NN as shown in Table 2. Thanks to the risk correction schemes that enforce the non-negativity of empirical risk, Sconf-ABS and Sconf-NN did not suffer from the overfitting caused by negative empirical risk and greatly outperformed Sconf-Unbiased and all the baseline methods. For Sconf-ABS and Sconf-NN, the learning curves of their test and training loss are consistent, i.e., the minimization of training loss corresponding to the corrected risk estimators implies the minimization of classification risk. This observation indicates that the corrected risk estimators can better represent the classification risk compared with the unbiased risk estimator.

8. Conclusion

We proposed a novel weakly supervised learning setting and effective algorithms for learning from unlabeled data pairs equipped with similarity confidence, where no class labels or similarity labels are needed. We proposed the unbiased risk estimator from unlabeled data pairs with similarity confidence and further improved its performance against overfitting via a risk correction scheme. Furthermore, we proved the consistency of the minimizers of the risk estimator and corrected risk estimators. Experimental results showed that the proposed methods outperform baseline methods and our proposed risk correction scheme can effectively mitigate overfitting caused by negative empirical risk.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 12071475, 11671010), Beijing Natural Science Foundation (No.4172035), the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2019-0013), National Satellite of Excellence in Trustworthy Software Systems (Award No: NSOE-TSS2019-01), and NTU. GN and MS were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. MS was also supported by the Institute for AI and Beyond, UTokyo.

References

Bao, H., Niu, G., and Sugiyama, M. Classification from pairwise similarity and unlabeled data. In *ICML*, pp. 461–470, 2018.

Bao, H., Shimada, T., Xu, L., Sato, I., and Sugiyama, M. Similarity-based classification: Connecting similarity learning to binary classification. *CoRR*, abs/2006.06207, 2020.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. In *COLT*, pp. 224–240, 2001.

Basu, S., Banerjee, A., and Mooney, R. J. Semi-supervised clustering by seeding. In *ICML*, pp. 27–34, 2002.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. The MIT Press, 2006.

Chou, Y., Niu, G., Lin, H., and Sugiyama, M. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *ICML*, pp. 1929–1938, 2020.

Chung, K. *A Course in Probability Theory*. Academic Press, 1974.

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

Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. EMNIST: an extension of MNIST to handwritten letters. *CoRR*, abs/1702.05373, 2017.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *J. Mach. Learn. Res.*, 12:1501–1536, 2011.

du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *NeurIPS*, pp. 703–711, 2014.

du Plessis, M. C., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *ICML*, pp. 1386–1394, 2015.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *KDD*, pp. 213–220, 2008.

Feng, L., Kaneko, T., Han, B., Niu, G., An, B., and Sugiyama, M. Learning from multiple complementary labels. In *ICML*, pp. 3072–3081, 2020a.

Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B., and Sugiyama, M. Provably consistent partial-label learning. In *ICML*, 2020b.

Fisher, R. J. Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, (2): 303–315, 1993.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *COLT*, pp. 297–299, 2018.

Guo, L., Zhou, Z., and Li, Y. Record: Resource constrained semi-supervised learning under distribution shift. In *KDD*, pp. 1636–1644, 2020.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *CVPR*, pp. 1735–1742, 2006.

Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I. W., Zhang, Y., and Sugiyama, M. Masking: A new perspective of noisy supervision. In *NeurIPS*, pp. 5841–5851, 2018a.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pp. 8536–8546, 2018b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 463–469, 2016.

Howe, J. Crowdsourcing: Why the power of the crowd is driving the future of business. *Crwon Publishing Group*, 2009.

Hsu, Y., Lv, Z., Schlosser, J., Odom, P., and Kira, Z. Multi-class classification without multi-class labels. In *ICLR*, 2019.

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from complementary labels. In *NeurIPS*, pp. 5639–5649, 2017.

Ishida, T., Niu, G., Menon, A. K., and Sugiyama, M. Complementary-label learning for arbitrary losses and models. In *ICML*, pp. 2971–2980, 2019.

Katsura, Y. and Uchida, M. Bridging ordinary-label learning and complementary-label learning. In *ACML*, pp. 161–176, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, pp. 1675–1685, 2017.

Klein, D., Kamvar, S. D., and Manning, C. D. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *ICML*, pp. 307–314, 2002.

Koch, G., Zemel, R., and Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015.

Krizhevsky, A. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

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

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Li, Y. and Liang, D. Safe semi-supervised learning: A brief introduction. *Frontiers Comput. Sci.*, 13(4):669–676, 2019.

Li, Y. and Zhou, Z. Towards making unlabeled data never hurt. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(1): 175–188, 2015.

Lu, N., Zhang, T., Niu, G., and Sugiyama, M. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In *AISTATS*, pp. 1115–1125, 2020.

Lv, J., Xu, M., Feng, L., Niu, G., Geng, X., and Sugiyama, M. Progressive identification of true labels for partial-label learning. In *ICML*, pp. 6500–6510, 2020.

McDiarmid, C. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

Mendelson, S. Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inf. Theory*, 54(8):3797–3803, 2008.

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

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, pp. 807–814, 2010.

Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. Learning with noisy labels. In *NeurIPS*, pp. 1196–1204, 2013.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Reading digits in natural images with unsupervised feature learning. *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 1 2011.

Niu, G., Dai, B., Yamada, M., and Sugiyama, M. Information-theoretic semi-supervised metric learning via entropy regularization. In *ICML*, 2012.

Niu, G., Jitkrittum, W., Dai, B., Hachiya, H., and Sugiyama, M. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *ICML*, pp. 10–18, 2013.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

Sakai, T., du Plessis, M. C., Niu, G., and Sugiyama, M. Semi-supervised classification based on classification from positive and unlabeled data. In *ICML*, pp. 2998–3006, 2017.

Sansone, E., Natale, F. G. B. D., and Zhou, Z. Efficient training for positive unlabeled learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2584–2598, 2019.

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, 2015.

Shimada, T., Bao, H., Sato, I., and Sugiyama, M. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *CoRR*, abs/1904.11717, 2019.

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. Constrained k-means clustering with background knowledge. In *ICML*, pp. 577–584, 2001.

Wang, Q., Li, Y., and Zhou, Z. Partial label learning with unlabeled data. In *IJCAI*, pp. 3755–3761, 2019.

Wang, W. and Zhou, Z. Crowdsourcing label quality: A theoretical analysis. *Sci. China Inf. Sci.*, 58(11):1–12, 2015.

Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., Liu, H., and Niu, G. Multi-class classification from noisy-similarity-labeled data. *CoRR*, abs/2002.06508, 2020.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, pp. 6835–6846, 2019.

Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Part-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*, 2020.

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

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. J. Distance metric learning with application to clustering with side-information. In *NeurIPS*, pp. 505–512, 2002.

Yi, J., Zhang, L., Jin, R., Qian, Q., and Jain, A. K. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *ICML*, pp. 1400–1408, 2013.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *ECCV*, pp. 69–85, 2018.

Zhang, Z., Zhao, P., Jiang, Y., and Zhou, Z. Learning from incomplete and inaccurate supervision. In *KDD*, pp. 1017–1025, 2019.

Zhou, Z. A brief introduction to weakly supervised learning. *National Science Review*, 005(001):44–53, 2018.

Zhu, X. and Goldberg, A. B. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

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.