Journal of Machine Learning Research 23 (2022) 1-34 Submitted 8/21; Revised 11/22; Published 11/22

# Learning from Noisy Pairwise Similarity and Unlabeled Data

**Songhua Wu** SONGHUA.WU@SYDNEY.EDU.AU
*Sydney AI Centre*
*The University of Sydney*
*Sydney, Australia*

**Tongliang Liu**[*] TONGLIANG.LIU@SYDNEY.EDU.AU
*Sydney AI Centre*
*The University of Sydney*
*Sydney, Australia*

**Bo Han** BHANML@COMP.HKBU.EDU.HK
*Department of Computer Science*
*Hong Kong Baptist University*
*Hong Kong, China*

**Jun Yu** HARRYJUN@USTC.EDU.CN
*Department of Automation*
*University of Science and Technology of China*
*Hefei, China*

**Gang Niu** GANG.NIU@RIKEN.JP
*Center for Advanced Intelligence Project*
*RIKEN*
*Tokyo, Japan*

**Masashi Sugiyama** SUGI@K.U-TOKYO.AC.JP
*Center for Advanced Intelligence Project*
*RIKEN*
*Tokyo, Japan*
*Graduate School of Frontier Sciences*
*The University of Tokyo*
*Chiba, Japan*

**Editor:** Pradeep Ravikumar

## Abstract

SU classification employs similar (S) data pairs (two examples belong to the same class) and unlabeled (U) data points to build a classifier, which can serve as an alternative to the standard supervised trained classifiers requiring data points with class labels. SU classification is advantageous because in the era of big data, more attention has been paid to data privacy. Datasets with specific class labels are often difficult to obtain in real-world classification applications regarding privacy-sensitive matters, such as politics and religion, which can be a bottleneck in supervised classification. Fortunately, similarity labels do not reveal the explicit information and inherently protect the privacy, e.g., collecting answers to "With whom do you share the same opinion on issue $\mathcal{I}$?" instead of "What is your opinion on issue $\mathcal{I}$?". Nevertheless, SU classification still has an

[*] The corresponding author.

obvious limitation: respondents might answer these questions in a manner that is viewed favorably by others instead of answering truthfully. Therefore, there exist some dissimilar data pairs labeled as similar, which significantly degenerates the performance of SU classification. In this paper, we study how to learn from noisy similar (nS) data pairs and unlabeled (U) data, which is called *nSU classification*. Specifically, we carefully model the similarity noise and estimate the noise rate by using the *mixture proportion estimation* technique. Then, a clean classifier can be learned by minimizing a denoised and unbiased classification risk estimator, which only involves the noisy data. Moreover, we further derive a theoretical generalization error bound for the proposed method. Experimental results demonstrate the effectiveness of the proposed algorithm on several benchmark datasets.

**Keywords:** privacy concern, similarity learning, unbiased classifier

## 1. Introduction

In standard supervised classification scenarios, the performances of classifiers crucially rely on the amount of accurately labeled data. However, in many real-world classification problems, collecting large-scale fully labeled data is expensive and time-consuming, and sometimes even infeasible. Therefore, weakly supervised learning (WSL) (Zhou, 2017; Han et al., 2019; Wang et al., 2019; Li et al., 2017, 2018; Krause et al., 2016; Khetan et al., 2018; Hu et al., 2019; Liu and Tao, 2016; Han et al., 2018; Wu et al., 2021; Sugiyama et al., 2022; Bai et al., 2022; Li et al., 2022b; Cheng et al., 2022b; Yang et al., 2022; Wu et al., 2022; Li et al., 2022a; Cheng et al., 2022a; Yao et al., 2022; Xia et al., 2022), which utilizes weaker supervisions, is becoming more and more prominent.

Within WSL, similarity learning (Gionis et al., 1999; Kulis et al., 2012) is one of the most notable emerging problems, where recent approaches primarily explore two directions. One direction focuses on semi-supervised clustering (Wagstaff et al., 2001). For example, similar and dissimilar data pairs are treated as must-link pairs and cannot-link pairs, which are used as constraints on clustering (Wagstaff et al., 2001; Basu et al., 2002; Li and Liu, 2009; Hu et al., 2008; Calandriello et al., 2014); similarity supervision is used for metric learning, which learns a distance function over instances and can easily convert to clustering tasks (Xing et al., 2003; Davis et al., 2007; Niu et al., 2014). Another direction aims at classification. For example, the similarity (dissimilarity) and unlabeled learning was proposed by designing unbiased risk estimators for binary classification (Bao et al., 2018; Shimada et al., 2021; Bao et al., 2022); a multi-class classifier can be learned from similarity supervision (Hsu et al., 2019).

However, the above methods are based on the strong assumption that similarity labels are entirely accurate, which is hard to meet for many applications. For example, when it comes to privacy-sensitive matters, such as politics and religion, people often hesitate to directly answer questions like “What is your opinion on issue $\mathcal{I}$?” and prefer to answer questions like “With whom do you share the same opinion on issue $\mathcal{I}$?”. Questions in this form can be regarded as one type of randomized response technique, which is a commonly used indirect questioning survey method that reduces the social desirability bias and increases data reliability (Warner, 1965; Fisher, 1993; Bao et al., 2018). To some degree, similarity information avoids embarrassment and protects personal privacy. Nevertheless, in practice, respondents might answer these questions in a manner that is viewed favorably by others instead of answering truthfully (Oral, 2019). As a result, the collected similarity data not only contain similar but also dissimilar data pairs. This kind of noise is random noise rather than adversarial noise because adversarial noise is often designed by considering the properties of algorithms so that it can confuse the algorithm (Szegedy et al., 2014; Madry et al., 2018); however, the noise induced by humans is independent of machine learning algorithms. In this case, if we directly employ the existing algorithms for clean similarity learning to deal with noisy similarity supervision, the classification performance will inevitably degenerate because the model will overfit the noisy data (Zhang et al., 2017). For example, if we directly employ the estimator designed for SU classification (Bao et al., 2018), an estimation bias would be introduced, and the learned classifier would thereby no longer be optimal.

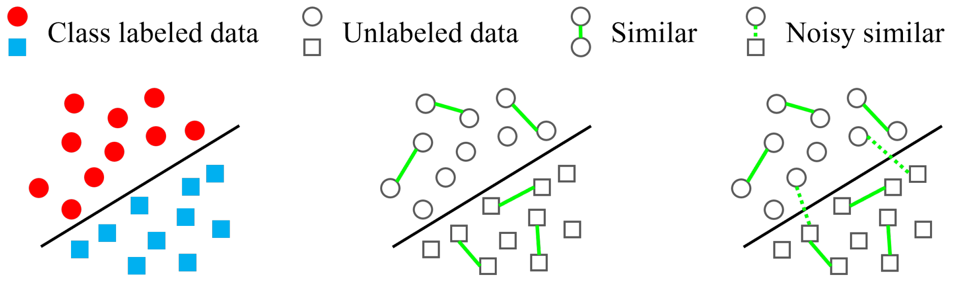
In this paper, we study the problem of how to learn a robust consistent classifier from noisy similar data pairs and unlabeled (nSU) data, which is called *nSU classification*. As shown in Figure 1, there is no class supervision in SU or nSU classification. In addition, there are some wrong links of dissimilar data pairs in nSU classification, which makes the problem more difficult. To this end, we propose an empirical risk minimization (ERM) (Vapnik, 1991) framework to learn classifiers from only nSU data. Although the unbiased classification risk estimator for SU data has been studied, how to properly model the noise is a significant bottleneck for solving the nSU problem. There are two widely-used noise models, i.e., the class-conditional label noise (CCN) model (Angluin and Laird, 1988) and mutually contaminated distributions (MCD) model (Scott et al., 2013). We employ MCD to model the noise: the distribution of noisy similar data pairs is a mixture proportion of the similar and dissimilar data pairs, where the noise rate is defined as the proportion of dissimilar data pairs in noisy similar data pairs. We choose MCD because CCN is a noise model for labeling noise and MCD is a noise model for sampling noise, which is why MCD is more suitable for the scenario of surveying sensitive topics with indirect questioning (more discussion can be found in Appendix G). To estimate the noise rate, we decompose the nSU data distributions and convert them to a standard mixture proportion estimation (MPE)[1] format. We prove that our MPE problem is well defined such that the noise rate is identifiable and can be consistently estimated using MPE methods. Then we theoretically build an unbiased estimator for the classification risk with respect to the fully and accurately labeled data.

The contributions of this paper are summarized as follows:

- Under the ERM framework, we propose a denoised and unbiased estimator for the classification risk with respect to the accurately labeled data by employing the nSU data.
- We prove that the MPE problem raised in this work is well-defined and thus the solutions are identifiable, i.e., the noise rate and class-prior can be consistently estimated from the nSU data.
- We theoretically establish a generalization error bound for the proposed nSU classification method, showing that the learned classifier will converge to the optimal classifier with respect to accurately labeled data.
- We empirically demonstrate that the proposed method can effectively reduce the side effect of noisy similarity data.

The rest of this paper is organized as follows. In Section 2, we formalize the fundamental definitions, goals, and assumptions of the nSU classification problem. In Section 3, we propose the nSU classification method and practical implementation. In Section 4, we discuss how to estimate the parameters. In Section 5, we analyze the generalization error. In Section 6, we discuss the experimental results. In Section 7, we conclude our paper.

[1] Let $F, G$, and $H$ be distributions on $(\mathcal{X}, \mathfrak{S})$ such that $F = (1-\kappa)G + \kappa H$, where $0 \leq \kappa \leq 1$. MPE is to estimate $\kappa$, given i.i.d. samples from both $F$ and $H$ (Blanchard et al., 2010).



(a) Supervised Classification (b) SU Classification (c) nSU Classification

Figure 1: The illustration of supervised classification, SU classification, and nSU classification. SU classification learns from similar data pairs and unlabeled data while nSU classification learns from noisy similar data pairs and unlabeled data.

## 2. Problem Setup and Related Work

In this section, we formalize the nSU classification problem and introduce related works.

### 2.1 Preliminaries

We consider the binary classification problem. Let $\mathcal{D}$ be the distribution of a pair of random variables $(x, y) \in \mathcal{X} \times \{-1, 1\}$, where $\mathcal{X} \subset \mathbb{R}^d$ and $d$ represents the dimension. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a *hypothesized classifier* in the predefined *hypothesis class* $\mathcal{F}$, and $\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$ be the *loss function* measuring how well the true class label $Y$ is estimated by the prediction of a hypothesis. The *optimal classifier* $f^*$ is therefore defined by the hypothesis that minimizes the *expected classification risk*:

$$R(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)], \tag{1}$$

i.e.,

$$f^* = \mathop{\mathrm{argmin}}_{f \in \mathcal{F}} R(f). \tag{2}$$

Often, the distribution of data is unknown. The standard supervised binary classification method utilizes positive and negative training data drawn i.i.d. from $\mathcal{D}$ to learn the optimal classifier by minimizing the empirical classification risk, which is an approximation of the expected risk in Eq. (1). While in our setting, we only have noisy similar (nS) data pairs, i.e., $\{(\boldsymbol{x}_{\mathrm{S},1}, \boldsymbol{x'}_{\mathrm{S},1}), \ldots, (\boldsymbol{x}_{\mathrm{S},n_{\mathrm{nS}}}, \boldsymbol{x'}_{\mathrm{S},n_{\mathrm{nS}}})\}$, and some unlabeled (U) data, i.e., $\{\boldsymbol{x}_{\mathrm{U},1}, \ldots, \boldsymbol{x}_{\mathrm{U},n_{\mathrm{U}}}\}$, which are called the nSU data. A pair of instances is said to be similar if they are from the same class. Noisy similar data pairs mean that the data may come from different classes but are treated as similar data pairs. Therefore, our research problem in this paper is to build an empirical classification risk by only employing the nSU data that will approximate the expected risk in Eq. (1).

### 2.2 Noisy Pairwise Similarity and Unlabeled Data

Below, we provide detailed definitions and assumptions in the nSU classification.

**Assumption 1** *Data independence (Bao et al., 2018). Without any assumptions, the pairwise data cannot effectively be used to approximate the risk. We assume that each instance is independently drawn from joint distribution $\mathcal{D}$. For example, for every data pair $(\boldsymbol{x}, \boldsymbol{x}')$, if two instances are similar, they follow the probability density as*

$$\begin{aligned} p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') &= p(\boldsymbol{x}, \boldsymbol{x}' | y = y' = +1 \vee y = y' = -1) \\ &= \frac{\pi_+^2 p_+(\boldsymbol{x}) p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}) p_-(\boldsymbol{x}')}{\pi_+^2 + \pi_-^2}, \end{aligned} \tag{3}$$

*where $p(A \vee B)$ represents the probability density that either $A$ or $B$ occurs, and*

- *$\pi_+ = p(y = +1)$ and $\pi_- = p(y = -1)$ are the class-prior probabilities, which satisfy $\pi_+ + \pi_- = 1$,*
- *$p_+(\boldsymbol{x}) = p(\boldsymbol{x}|y = +1)$ and $p_-(\boldsymbol{x}) = p(\boldsymbol{x}|y = -1)$ are the class-condition probability density.*

More discussion about this assumption can be found in Appendix C.

Eq. (3) shows that two instances are drawn independently following $p(\boldsymbol{x}, y)$, which corresponds to the density of $\mathcal{D}$, and they belong to the similar data pairs if they have the same label. In contrast, if the two instances of a data pair belong to the different classes, they follow the probability density as

$$\begin{aligned} p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}') &= p(\boldsymbol{x}, \boldsymbol{x}' | (y = +1 \wedge y' = -1) \vee (y = -1 \wedge y' = +1)) \\ &= \frac{\pi_+ \pi_- p_+(\boldsymbol{x}) p_-(\boldsymbol{x}') + \pi_+ \pi_- p_-(\boldsymbol{x}) p_+(\boldsymbol{x}')}{2\pi_+ \pi_-} \\ &= \frac{p_+(\boldsymbol{x}) p_-(\boldsymbol{x}') + p_-(\boldsymbol{x}) p_+(\boldsymbol{x}')}{2}, \end{aligned} \tag{4}$$

where $p(A \wedge B)$ represents the probability density that both $A$ and $B$ occur.

For the unlabeled data, we assume that they are independently drawn from the marginal density $p(\boldsymbol{x})$, which can be decomposed as

$$p(\boldsymbol{x}) = \pi_+ p_+(\boldsymbol{x}) + \pi_- p_-(\boldsymbol{x}). \tag{5}$$

**Lemma 1** *Assume that Assumption I holds. If two instances $\boldsymbol{x}$ and $\boldsymbol{x}'$ are drawn independently from the unlabeled data density, Eq. (5) can easily convert to a pairwise marginal version such that*

$$p(\boldsymbol{x}, \boldsymbol{x}') = \pi_{\mathrm{s}} p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') + \pi_{\mathrm{d}} p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{6}$$

*where $\pi_{\mathrm{s}} = \pi_+^2 + \pi_-^2$ and $\pi_{\mathrm{d}} = 2\pi_+ \pi_-$.*

A detailed proof is provided in Appendix A.

**Assumption 2** *Contamination model. To handle the noise, we consider the contamination model proposed by Huber et al. (1964), which has been widely used in the label noise learning community (Menon et al., 2015; Scott et al., 2013). Specifically, the noisy similarity data consist of both similar data pairs and dissimilar data pairs:*

$$\tilde{p}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') = (1 - \rho_{\mathrm{d}}) p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') + \rho_{\mathrm{d}} p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{7}$$

*where $\rho_{\mathrm{d}} \in [0, 1)$ is regarded as the noise rate and $\tilde{p}_{\mathrm{s}}$ denotes the density of noisy similar data pairs. More specifically, in $\tilde{p}_{\mathrm{s}}$, a proportion $\rho_{\mathrm{d}}$ of data pairs are contaminated by dissimilar data pairs, while the $(1 - \rho_{\mathrm{d}})$ proportion remains similar.*

We employ the contamination model rather than the CCN model (see Section 2.3 for details) because the former is more suitable to describe the noise pattern of nSU data. We take the "opinion-on-issue-$\mathcal{I}$" case as an example: In practice, the data generation procedure is to collect answers to the question "With whom do you share the same opinion on issue $\mathcal{I}$?". In statistics, it is to sample examples of similar data pairs from $p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$. However, some people give wrong answers, which makes the selected examples contain dissimilar data pairs from $p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}')$. Overall, the data generation procedure is an imbalanced sample from both $p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$ and $p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}')$, which can be exactly formulated by a contamination model in Eq. (7).

The above two assumptions overlook the data-dependence in pairwise data (i.e., pairwise data is independently sampled from $p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$) and the instance-dependence of noise (i.e., $(1 - \rho_{\mathrm{d}})$ is independent of $\boldsymbol{x}$ and $\boldsymbol{x}'$). However, this simplification has been widely accepted in statistical learning theory and the label-noise learning communities, and the empirical results on benchmark datasets verify the efficiency of the assumptions (Patrini et al., 2017; Xia et al., 2019). We could then build a denoised and unbiased estimator for the classification risk with respect to the latent clean data with the nSU data and provide a theoretical error bound for the proposed method.

We denote the noisy similar data pairs and the unlabeled data by $\tilde{D}_{\mathrm{s}}$ and $D_{\mathrm{u}}$, respectively.

$$\tilde{D}_{\mathrm{s}} \triangleq \{(\boldsymbol{x}_{\mathrm{S},1}, \boldsymbol{x}'_{\mathrm{S},1}), \ldots, (\boldsymbol{x}_{\mathrm{S},n_{\mathrm{nS}}}, \boldsymbol{x}'_{\mathrm{S},n_{\mathrm{nS}}})\} \overset{\text{i.i.d.}}{\sim} \tilde{p}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{8}$$

$$D_{\mathrm{u}} \triangleq \{\boldsymbol{x}_{\mathrm{U},1}, \ldots, \boldsymbol{x}_{\mathrm{U},n_{\mathrm{U}}}\} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}), \tag{9}$$

where $n_{\mathrm{nS}}$ is the size of $\tilde{D}_{\mathrm{s}}$ and $n_{\mathrm{U}}$ is the size of $D_{\mathrm{u}}$. We show that a consistent classifier could be learned with theoretical guarantee.

### 2.3 Related Work

There are a few works regarding the label noise issue on similarity learning. However, the employed noise model, the essential parameter estimation method, and the classifier models of our work are all different from the related works'.

First, Dan et al. (2021) studied a similar but different problem of how to learn a binary classifier from noisy similar data pairs and dissimilar data pairs. The noise model Dan et al. (2021) employed is CCN, where the clean labels are assumed to flip into other classes with a certain probability. Specifically, in Dan et al. (2021), similar data pairs are corrupted into dissimilar data pairs with probability $\alpha$, and dissimilar data pairs are corrupted into similar data pairs with probability $\beta$:

$$\begin{bmatrix} p(\bar{S} = 0|\boldsymbol{x}, \boldsymbol{x}') \\ p(\bar{S} = 1|\boldsymbol{x}, \boldsymbol{x}') \end{bmatrix} = \begin{bmatrix} 1 - \beta & \alpha \\ \beta & 1 - \alpha \end{bmatrix} \begin{bmatrix} p(S = 0|\boldsymbol{x}, \boldsymbol{x}') \\ p(S = 1|\boldsymbol{x}, \boldsymbol{x}') \end{bmatrix}, \tag{10}$$

where $S$ and $\bar{S}$ denote the similarity label and noisy similarity label[2].

This model is different from the MCD model, where the noise distribution is a mixture proportion of the clean distributions. Specifically, for the MCD model, in $\tilde{p}_{\mathrm{s}}$ ($\tilde{p}_{\mathrm{d}}$), a proportion $\rho_{\mathrm{d}}$ ($\rho_{\mathrm{s}}$) of data

[2] Dan et al. (2021) called this noise model *pairing corruption*. They also discussed another noise model called *labeling corruption*, where labels $Y$ and $Y'$ are corrupted in an instance-independent manner.

pairs are contaminated by dissimilar (similar) data pairs, while the remaining $1 - \rho_{\mathrm{d}}$ $(1 - \rho_{\mathrm{s}})$ proportion remains similar (dissimilar):

$$\begin{bmatrix} p(\boldsymbol{x}, \boldsymbol{x}' | \bar{S} = 0) \\ p(\boldsymbol{x}, \boldsymbol{x}' | \bar{S} = 1) \end{bmatrix} = \begin{bmatrix} 1 - \rho_{\mathrm{s}} & \rho_{\mathrm{s}} \\ \rho_{\mathrm{d}} & 1 - \rho_{\mathrm{d}} \end{bmatrix} \begin{bmatrix} p(\boldsymbol{x}, \boldsymbol{x}' | S = 0) \\ p(\boldsymbol{x}, \boldsymbol{x}' | S = 1) \end{bmatrix}. \tag{11}$$

It is notable that the middle matrix in Eq. (10) is column normalized while the middle matrix in Eq. (11) is row normalized. Moreover, the CCN model is a strict special case of the MCD model (Menon et al., 2015). It has been studied in Lu et al. (2019) that $p(\bar{S})$ is fixed in the CCN model once $p(\bar{S}|\boldsymbol{x}, \boldsymbol{x}')$ is specified while $p(\bar{S})$ is free in the MCD model after $p(\boldsymbol{x}, \boldsymbol{x}'|\bar{S})$ is specified. Furthermore, for $\tilde{p}(\boldsymbol{x})$ being the distribution of noisy $\boldsymbol{x}$, $\tilde{p}(\boldsymbol{x}) = p(\boldsymbol{x})$ holds in the CCN model but $\tilde{p}(\boldsymbol{x}) \neq p(\boldsymbol{x})$ holds in the MCD model. Due to this covariate shift (Sugiyama and Kawanabe, 2012), CCN methods do not fit the MCD problem setting, while the MCD methods fit the CCN problem setting conversely.

For our nSU contamination model, we only have the noisy similar data pairs, and Eq. (7) is consistent with part of Eq. (11). Besides, there is no restriction on the noisy dissimilar data pairs. Namely, we can set the imaginary noisy dissimilar data pairs to obey the distribution in Eq. (11). Moreover, the distribution of the unlabeled data is free to label noise. Therefore, the MCD model, as well as the CCN model, is a special case of the nSU contamination model.

Second, to build an unbiased classification risk framework from the observation, there are some essential parameters to estimate. Dan et al. (2021) roughly tuned the noise rate parameter by cross-validation on the noisy data. Then, based on the tuned parameter, the prior was empirically estimated. Bao et al. (2018) used an MPE method to solve the estimation problem. However, the solution of that MPE problem is not guaranteed to be identifiable. By contrast, we prove that the new MPE problem in this work is *irreducible* (Scott, 2015), and thereby can be solved with a theoretical guarantee (see Section 4 for details).

Third, Dan et al. (2021) used a neural network to approximately learn the classifier. However, we use not only the neural network but also the linear model, of which there exist analytical solutions to the latter model.

## 3. Learning from nSU Data

In this section, we rewrite the classification risk in Eq. (1) by only employing nSU data. Then, we discuss its practical implementation under different surrogate loss functions. Lastly, we show an effective method for estimating class-prior $\pi_+$ and noise rate $\rho_{\mathrm{d}}$.

### 3.1 Risk Expression with nSU Data

**Theorem 1** *Assume that* $\pi_+ \neq 0.5$. *The classification risk in Eq.* (1) *can be equivalently expressed only in terms of nSU data as follows:*

$$R_{\mathrm{nSU},\ell}(f) = \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{x}')\sim\tilde{p}_{\mathrm{s}}} [\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x}, \boldsymbol{x}')] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p} [\mathcal{L}_{\mathrm{U}}(\boldsymbol{x})], \tag{12}$$

*where*

$$\begin{aligned}\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x}, \boldsymbol{x}') &= \frac{\pi_{\mathrm{s}}(1-\pi_{\mathrm{s}})}{2(1-\rho_{\mathrm{d}}-\pi_{\mathrm{s}})(2\pi_{+}-1)}[\tilde{l}(\boldsymbol{x}) + \tilde{l}(\boldsymbol{x}')],\\ \tilde{l}(\boldsymbol{x}) &= \ell(f(\boldsymbol{x}), +1) - \ell(f(\boldsymbol{x}), -1),\\ \mathcal{L}_{\mathrm{U}}(\boldsymbol{x}) &= \frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_{-}(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_{+}-1)}\ell(f(\boldsymbol{x}), +1) - \frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_{+}(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_{+}-1)}\ell(f(\boldsymbol{x}), -1).\end{aligned}$$

A detailed proof and discussion are provided in Appendix B.

Theorem 1 immediately leads to an unbiased risk estimator:

$$\hat{R}_{\mathrm{nSU},\ell}(f) = \frac{1}{n_{\mathrm{nS}}}\sum_{i=1}^{n_{\mathrm{nS}}}\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x}_{\mathrm{S},i}, \boldsymbol{x}'_{\mathrm{S},i}) + \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\mathcal{L}_{\mathrm{U}}(\boldsymbol{x}_{\mathrm{U},i}). \tag{13}$$

### 3.2 Practical Implementation

Here, we employ the linear-in-parameter-model $f(x) = \boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x})$, where $\boldsymbol{w}$ and $\boldsymbol{\phi}$ are vectors of parameters and basis functions with the same dimension. Then employing Eq. (13) with the $\ell_2$ regularization, the nSU classification can be formulated as the following regularized empirical risk minimization problem:

$$\hat{\boldsymbol{w}} = \min_{\boldsymbol{w}} \hat{J}_{\ell}(\boldsymbol{w}), \tag{14}$$

where

$$\begin{aligned}\hat{J}_{\ell}(\boldsymbol{w}) &= \frac{1}{n_{\mathrm{nS}}}\sum_{i=1}^{n_{\mathrm{nS}}}\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x}_{\mathrm{S},i}, \boldsymbol{x}'_{\mathrm{S},i}) + \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\mathcal{L}_{\mathrm{U}}(\boldsymbol{x}_{\mathrm{U},i}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\\ &= \frac{A}{2n_{\mathrm{nS}}}\sum_{i=1}^{2n_{\mathrm{nS}}}[\ell(\boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}), +1) - \ell(\boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}), -1)]\\ &\quad + \frac{B}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}[\ell(\boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), +1)] - \frac{C}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}[\ell(\boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), -1)] + \frac{\lambda}{2}\|\boldsymbol{w}\|^2,\end{aligned} \tag{15}$$

and

$$A = \frac{\pi_{\mathrm{s}}(1-\pi_{\mathrm{s}})}{(1-\rho_{\mathrm{d}}-\pi_{\mathrm{s}})(2\pi_{+}-1)}, \tag{16}$$

$$B = \frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_{-}(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_{+}-1)}, \tag{17}$$

$$C = \frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_{+}(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_{+}-1)}, \tag{18}$$

and $\lambda$ $(\geq 0)$ in Eq. (15) is the regularization parameter. Note that since the loss form is symmetric to $\boldsymbol{x}_{\mathrm{S},i}$ and $\boldsymbol{x}'_{\mathrm{S},i}$, we use $\boldsymbol{x}_{\mathrm{S},i}$ uniformly in Eq. (15). To solve this optimization problem, we need the knowledge of class-prior $\pi_{+}$ ($\pi_{\mathrm{s}}$ can be calculated from $\pi_{+}$) and the noise rate $\rho_{\mathrm{d}}$. In Section 4, we discuss how to estimate them from nSU data.

Inspired by Bao et al. (2018), Natarajan et al. (2013), and du Plessis et al. (2015), we surprisingly find that adopting certain loss functions, i.e., the margin loss function[3] (Mohri et al., 2018), will result in a convex objective function.

**Theorem 2** *Assume that the loss function $\ell(z,t)$ is a convex margin loss function, and for every fixed $t \in \{\pm 1\}$, $\ell(z,t)$ is twice differentiable with respect to $z$. If $\ell(z,t)$ satisfies the condition*

$$\ell(z,+1) - \ell(z,-1) = -z,$$

*then $\hat{J}_\ell(\boldsymbol{w})$ is convex.*

A detailed proof and more discussion are provided in Appendix D.

Below, we consider the squared loss function, which satisfies the conditions in Theorem 2.

The squared loss function is defined as $\ell_{\mathrm{SQ}}(z,t) = \frac{1}{4}(tz-1)^2$. Substituting $\ell_{\mathrm{SQ}}$ into Eq. (15), we have

$$\hat{J}_{\mathrm{SQ}}(\boldsymbol{w}) = \boldsymbol{w}^\top \left( \frac{1}{4n_{\mathrm{U}}} X_{\mathrm{U}}^\top X_{\mathrm{U}} + \frac{\lambda}{2} I \right) \boldsymbol{w} - \left( \frac{A}{2n_{\mathrm{nS}}} \mathbf{1}^\top X_{\mathrm{S}} + \frac{B+C}{2n_{\mathrm{U}}} \mathbf{1}^\top X_{\mathrm{U}} \right) \boldsymbol{w},$$

where $I$ is the identity matrix, $\mathbf{1}$ represents the vector whose elements are all ones, $X_{\mathrm{S}} = [\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{s},1}) \ \cdots \ \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{s},2n_{\mathrm{nS}}})]^\top$, and $X_{\mathrm{U}} = [\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{u},1}) \ \cdots \ \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{u},n_{\mathrm{U}}})]^\top$. Then we have the analytical solution of this minimization problem as

$$\boldsymbol{w} = n_{\mathrm{U}} \cdot \left( X_{\mathrm{U}}^\top X_{\mathrm{U}} + 2\lambda n_{\mathrm{U}} I \right)^{-1} \left( \frac{A}{n_{\mathrm{nS}}} X_{\mathrm{S}}^\top \mathbf{1} + \frac{B+C}{n_{\mathrm{U}}} X_{\mathrm{U}}^\top \mathbf{1} \right). \tag{19}$$

Besides, we provide a deep learning method where we can obtain an approximation to the optimal solution. Specifically, we employ a deep model $f$ with logistic loss: $\ell_{\mathrm{LG}}(z,t) = \log(1 + \exp(-tz))$. Then we directly optimize the objective function Eq. (13), into which $\ell_{\mathrm{LG}}$ substituted:

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_{\mathrm{nSU},\ell_{\mathrm{LG}}}(f). \tag{20}$$

## 4. Estimating $\pi_+$, $\pi_{\mathrm{s}}$, and $\rho_{\mathrm{d}}$ with the MPE method

In the aforementioned method, similar rate $\pi_{\mathrm{s}}$, class-prior $\pi_+$, and noise rate $\rho_{\mathrm{d}}$ are assumed to be given in advance, which is not always true. Here, we thus provide a practical method to estimate them. Mixture proportion estimation (MPE) (Blanchard et al., 2010) is the following problem: Let $F, G$, and $H$ be probability distributions on $(\mathcal{X}, \mathfrak{S})$ such that

$$F = (1-\kappa)G + \kappa H, \tag{21}$$

where $0 \leq \kappa \leq 1$. Given random samples from $F$ and $H$, estimate $\kappa$.

Similarly, the distributions of both $D_{\mathrm{u}}$ and $\tilde{D}_{\mathrm{s}}$ (see Eqs. (8) and (9)) have mixture representations according to Eq. (6) and Eq. (7). By substituting $(1-\pi_{\mathrm{s}})$ for $\pi_{\mathrm{d}}$, we obtain

$$P(\boldsymbol{x},\boldsymbol{x}') = (1-\pi_{\mathrm{s}}) P_{\mathrm{d}}(\boldsymbol{x},\boldsymbol{x}') + \pi_{\mathrm{s}} P_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}'), \tag{22}$$

$$\tilde{P}_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') = (1-\rho_{\mathrm{d}}) P_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') + \rho_{\mathrm{d}} P_{\mathrm{d}}(\boldsymbol{x},\boldsymbol{x}'). \tag{23}$$

[3] $\ell$ is said to be a margin loss function if there exists $\psi : \mathbb{R} \to \mathbb{R}_+$ such that $\ell(z,t) = \psi(tz)$.

The above equations are similar to the standard MPE format but the accessible samples are not from the corresponding $F$ and $H$. Therefore, by further calculating [(22)×$(1-\rho_{\mathrm{d}})$ − (23)×$\pi_{\mathrm{s}}$], [(22)×$\rho_{\mathrm{d}}$ − (23)×$(1-\pi_{\mathrm{s}})$] respectively and organizing, we obtain

$$P(\boldsymbol{x},\boldsymbol{x}') = (1-\frac{\pi_{\mathrm{s}}}{1-\rho_{\mathrm{d}}})P_{\mathrm{d}}(\boldsymbol{x},\boldsymbol{x}') + \frac{\pi_{\mathrm{s}}}{1-\rho_{\mathrm{d}}}\tilde{P}_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}'), \tag{24}$$

$$\tilde{P}_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') = (1-\frac{\rho_{\mathrm{d}}}{1-\pi_{\mathrm{s}}})P_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') + \frac{\rho_{\mathrm{d}}}{1-\pi_{\mathrm{s}}}P(\boldsymbol{x},\boldsymbol{x}'). \tag{25}$$

According to these mixture representations, i.e., Eq. (22) and Eq. (23), we assume that $1-\rho_{\mathrm{d}} > \pi_{\mathrm{s}}$, which can easily be held because the proportion of similar data pairs in $\tilde{D}_{\mathrm{s}}$ (denoted by $1-\rho_{\mathrm{d}}$) which it collects the similar data pairs purposely, is apparently bigger than that in unlabeled data (denoted by $\pi_{\mathrm{s}}$). More discussion can be found in Appendix H.

Based on this reasonable assumption, we have the following lemma:

**Lemma 2** *Assume* $1-\rho_{\mathrm{d}} > \pi_{\mathrm{s}}$. *Eq.* (24) *and Eq.* (25) *can be equivalently rewritten as a standard MPE format such that*

$$\begin{aligned} P(\boldsymbol{x},\boldsymbol{x}') &= (1-\gamma)P_{\mathrm{d}}(\boldsymbol{x},\boldsymbol{x}') + \gamma\tilde{P}_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}'), & (26)\\ \tilde{P}_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') &= (1-\kappa)P_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') + \kappa P(\boldsymbol{x},\boldsymbol{x}'). & (27)\end{aligned}$$

*where* $\gamma = \frac{\pi_{\mathrm{s}}}{1-\rho_{\mathrm{d}}} \in [0,1)$, $\kappa = \frac{\rho_{\mathrm{d}}}{1-\pi_{\mathrm{s}}} \in [0,1)$.

**Proof** Lemma 2 directly follows from Eq. (24) and Eq. (25) under the assumption $1-\rho_{\mathrm{d}} > \pi_{\mathrm{s}}$. ■

Note that since we have no information about $G$ in the original MPE problem, without additional assumptions, MPE is ill-defined and the mixture proportion $\kappa$ is not identifiable.

The weakest and most common assumption to yield the identifiability of the mixture proportion $\kappa$ is the *irreducibility assumption* (Blanchard et al., 2010):

**Definition 1** *(Scott, 2015) Let* $G$*, and* $H$ *be probability distributions. We say that* $G$ *is irreducible with respect to* $H$ *if there exists no decomposition of the form* $G = \gamma H + (1-\gamma)F'$*, where* $F'$ *is some probability distribution and* $0 < \gamma \leq 1$. *We say that* $G$ *and* $H$ *are mutually irreducible if* $G$ *is irreducible with respect to* $H$ *and vice versa.*

The irreducibility assumption states that the maximum proportion of $H$ in $G$ approaches to 0, otherwise there would exist such an $F'$. Consider $\mathfrak{S}$ is the set of measurable sets in $\mathcal{X}$ and above discussion straightforwardly implies the following fact (Scott et al., 2013; Blanchard et al., 2010): $G$ being irreducible with respect to $H$ is equivalent to $\operatorname{supp}(H) \not\subset \operatorname{supp}(G)$, i.e.,

$$\inf_{S\in\mathfrak{S},H(S)>0} \frac{G(S)}{H(S)} = 0.$$

In general, the distributions of positive data and negative data are assumed to be mutually irreducible (Scott et al., 2013), which leads to the following theorem.

**Theorem 3** *Assume that the positive data distribution* $P_{+}$ *and the negative data distribution* $P_{-}$ *are mutually irreducible, then* $P_{\mathrm{d}}$ *is irreducible with respect to* $\tilde{P}_{\mathrm{s}}$*, and* $P_{\mathrm{s}}$ *is irreducible with respect to* $P$. *Thus, the mixture proportions* $\gamma$ *and* $\kappa$ *in Lemma 2 is identifiable.*

**Algorithm 1** nSU classification.

```
**Input:** Noisy similar data pairs $\tilde{D}_{\mathrm{s}}$ and unlabeled data $D_{\mathrm{u}}$;
**Output:** The classifier $\hat{f}$;
**Stage 1. Estimate the similar rate** $\pi_{\mathrm{s}}$ **and noise rate** $\rho_{\mathrm{d}}$
Intermediate parameters $(\gamma, \kappa)$ = *MPE*$(\tilde{D}_{\mathrm{s}}, D_{\mathrm{u}})$;
Compute $(\pi_{\mathrm{s}}, \rho_{\mathrm{d}}, \pi_+, \pi_-)$ from $(\gamma, \kappa)$;
**Stage 2. Obtain classifier** $\hat{f}$
**if** Squared loss **then**
    Compute the analytical solution $\hat{\boldsymbol{w}}$ by Eq. (19);
**end if**
**if** Logistic loss **then**
    Approximate the optimal classifier $f^*$ by the SGD;
**end if**
**return** $\hat{f}$;
```

A detailed proof is provided in Appendix E.

Based on Lemma 2 and Theorem 3, we can effectively estimate $\gamma$ and $\kappa$ by the MPE method (Ramaswamy et al., 2016). After estimating $\gamma$ and $\kappa$, we can reversely calculate $\pi_{\mathrm{s}}$, $\rho_{\mathrm{d}}$, and $\pi_+$ according to the definitions of $\gamma$ and $\kappa$ in Lemma 2 as follows:

- Similar rate: $\pi_{\mathrm{s}} = \frac{\gamma(1-\kappa)}{1-\gamma\kappa}$,
- Noise rate: $\rho_{\mathrm{d}} = \frac{\kappa(1-\gamma)}{1-\gamma\kappa}$,
- Class-prior[4]: $\pi_+ = \frac{\sqrt{2\frac{\gamma(1-\kappa)}{1-\gamma\kappa}-1}+1}{2}$..

Overall, our method for nSU classification is summarized in Algorithm (1).

## 5. Error Bound Analysis

In this section, we derive a generalization error bound for the nSU classification.

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a function class of the linear-in-parameter model, and $f^* = \operatorname*{argmin}_{f \in \mathcal{F}} R(f)$ be the true risk minimizer, and $\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}_{\mathrm{nSU},\ell}(f)$ be the empirical risk minimizer. By introducing a Rademacher Complexity bound assumption (Bao et al., 2018), i.e., for any probability density $\mu$, $\mathfrak{R}(\mathcal{F}; n, \mu) \le \frac{C_{\mathcal{F}}}{\sqrt{n}}$ for some constant $C_{\mathcal{F}} > 0$, then we can obtain the following theorem.

**Theorem 4** *Assume the loss function $\ell$ is $\rho$-Lipschitz with respect to the input instance $\boldsymbol{x}$ ($\rho \in (0, \infty)$), and all functions in the hypothesis class $\mathcal{F}$ are bounded by $C_{\mathrm{b}}$, i.e., $\|f\|_\infty \le C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_\ell = \sup_{t \in \{\pm 1\}} \ell(C_{\mathrm{b}}, t)$. For any $\delta > 0$, with probability at least $1 - \delta$,*

$$R(\hat{f}) - R(f^*) \le \frac{4A\rho C_{\mathcal{F}} + A\sqrt{2C_\ell^2 \log\frac{4}{\delta}}}{\sqrt{2n_{\mathrm{nS}}}} + \frac{2(-B-C)\rho C_{\mathcal{F}} + (-B-C)\sqrt{\frac{1}{2}C_\ell^2 \log\frac{4}{\delta}}}{\sqrt{n_{\mathrm{U}}}}, \quad (28)$$

[4] $2\pi_{\mathrm{s}} - 1 = \pi_{\mathrm{s}} - \pi_{\mathrm{d}} = (\pi_+ - \pi_-)^2 = (2\pi_+ - 1)^2$, such that $\pi_+ = \frac{\sqrt{2\frac{\gamma(1-\kappa)}{1-\gamma\kappa}-1}+1}{2}$

*where $A$, $B$, and $C$ are defined in Eqs.* (16), (17), *and* (18).

A detailed proof is provided in Appendix F.

Theorem 4 implies that the expected risk of the classifier learned from nSU data is consistent with that of the classifier learned from standard positive and negative data if we have $\pi_+$ and $\rho_{\mathrm{d}}$ in advance. The convergence rate is $\mathcal{O}_p(1/\sqrt{n_{\mathrm{nS}}} + 1/\sqrt{n_{\mathrm{U}}})$, which achieves the optimal parametric rate for the empirical risk minimization without additional assumptions (Mendelson, 2008).

## 6. Experiments

In this section, we experimentally investigate the behavior of the proposed method and the baselines for nSU classification on both synthetic and benchmark datasets. All experiments were conducted with 3.10GHz Intel(R) Core(TM) i9-9900 CPU and NVIDIA 2080Ti.

### 6.1 Data Generation and Common Setup

To obtain nSU data, first, we collected raw binary classification datasets which consist of positive and negative data, while leaving 10% of the data as test data. Then we converted the labeled data to noisy similar data pairs according to class-prior $\pi_+$ and noise rate $\rho_{\mathrm{d}}$. Specifically, we randomly subsampled similar data and dissimilar data pairs following the ratio of $1 - \rho_{\mathrm{d}}$ and $\rho_{\mathrm{d}}$. The similar data pairs consisted of positive and negative pairs with a ratio of $\pi_+^2$ and $\pi_-^2$. After that, we randomly selected unlabeled data samples from positive data and negative data with a ratio of $\pi_+$ and $\pi_-$.

For all the experiments, the sample size of the noisy similar data pairs was fixed to 4000, while the sample size of the unlabeled data was fixed to 2000. The class-prior $\pi_+$ and noise rate $\rho_{\mathrm{d}}$ were estimated by the MPE method (Ramaswamy et al., 2016). For the first MPE for $\gamma$, we used the default parameters. For the second MPE for $\kappa$, to ensure the term in the square root in the formula $\pi_+ = \frac{\sqrt{2\frac{\gamma(1-\kappa)}{1-\gamma\kappa}-1}+1}{2}$ to be greater than zero, we set $\lambda_{\mathrm{right}} = 2 - 1/\hat{\gamma}$ while kept all other parameters as default, where $\lambda_{\mathrm{right}}$ is a parameter in the MPE method (Ramaswamy et al., 2016) and $\hat{\gamma}$ is the estimated value.

We used the linear basis functions and the regularization parameter $\lambda$ was fixed to $10^{-4}$. For the deep model, we employed a 3-layer MLP (multilayer perceptron) with the softsign active function ($\mathrm{Saf}(x) = x/(1 + |x|)$). We used the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.002, which decays every 40 epochs by a factor of 0.1 with 200 epochs in total.

### 6.2 Baselines

We implemented our method with two models. In Stage 1, we employed the KM2 algorithm (Ramaswamy et al., 2016) to estimate the class-prior and noise rate. In Stage 2, for the linear model, we obtained the analytical solution by Eq. (19) (denoted by -LC (Linear Classifier)); for the deep neural network, we used the SGD to optimize the model (denoted by -DC (Deep Classifier)). We compared our proposed method with state-of-the-art methods:

- SU classification (Bao et al., 2018), which is the state-of-the-art method for learning from similarity and unlabeled data. This method was also implemented with two models, i.e., linear and deep models.

- nSD classification (Dan et al., 2021), which learns a classifier from noisy similarity and dissimilarity data pairs. To make it compatible with the nSU data, we treated the unlabeled data as the noisy dissimilarity data and fed the ground-truth class-prior and noises rates to the nSD algorithm.
- Information-theoretic metric learning (ITML) (Davis et al., 2007), which utilizes pairwise similarity and dissimilarity as constraints to learn a metric. Then, $k$-means clustering is applied on test data with the learned metric.

Moreover, we also implemented some unsupervised learning methods, i.e., $k$-means (KM) (MacQueen, 1967) and hierarchical clustering schemes (HC) (Johnson, 1967). For unsupervised learning methods, we directly used the implementations on scikit-learn (Pedregosa et al., 2011). We also employed a support vector classifier (SVC) (Cortes and Vapnik, 1995) with a linear kernel learned from fully-supervised data as a benchmark. Note that learning a classifier without class information will lose the mapping between the cluster nodes and the semantic classes. The semantic classes can be identified with some prior knowledge, e.g., the exact $\pi_+$ or the sign of $(\pi_+ - \pi_-)$. Here, we employed the Hungarian algorithm (Kuhn, 1955), which is a commonly used method for evaluating the clustering accuracy, to assign the output nodes to the dominant semantic classes by using some training examples with class labels.

### 6.3 Experiments on Synthetic Datasets

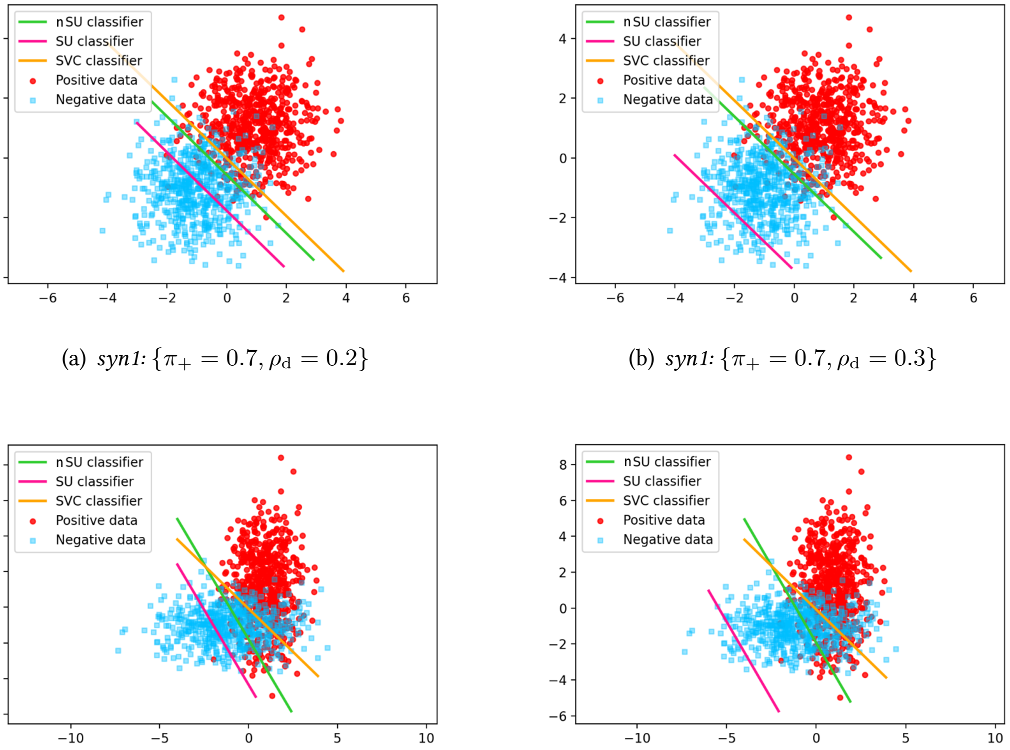
We generated synthetic data drawn from two-dimension normal distributions. Settings with various combinations of parameters were tested. The results are shown in Figures 2 and 3.

For one case, the positive data follows the Gaussian distribution $\mathcal{N}\left(\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$, and the negative data follows $\mathcal{N}\left(\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)$, which we called the syn1 dataset. For another case, the positive data follows $\mathcal{N}\left(\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}1&0\\0&2\end{pmatrix}\right)$, and the negative data follows $\mathcal{N}\left(\begin{pmatrix}-1\\-1\end{pmatrix},\begin{pmatrix}2&0\\0&1\end{pmatrix}\right)$, having more overlap, which we called the syn2 dataset. From such two binary datasets, we generated the nSU data with $\{\pi_+ = 0.7, \rho_\mathrm{d} = 0.2\}$ (Figures 2.a & 2.c) and $\{\pi_+ = 0.7, \rho_\mathrm{d} = 0.3\}$ (Figures 2.b & 2.d). Here we employed an SVC with a linear kernel learned from fully-supervised data as a benchmark. The class-prior and noise rate were assumed to be known.

In Figure 2, we can see that the nSU classifier is closer to the SVC classifier than the SU classifier in all four setups. As the noise rate increases from 0.2 to 0.3, the SU classifier moves further away from the SVC classifier, while the nSU classifier is hardly affected. From Figure 3, we can see that the accuracy of the SU classifier drops dramatically with the increased noise rate on both datasets. Meanwhile, there is only a slight fluctuation of the nSU classifier. Note that the gap between the accuracy of the nSU classifier and the SVC classifier trained on clean data is small, i.e., within two percentage points.

### 6.4 Experiments on UCI and LIBSVM Datasets

Here datasets were obtained from the LIBSVM data (Chang and Lin, 2011) and UCI Machine Learning Repository (Dua and Graff, 2017). Then we generated the corresponding nSU data following the data generation method. Tables 1 and 2 demonstrate the remarkable superiority of the nSU classifier



(a) *syn1:* $\{\pi_+ = 0.7, \rho_\text{d} = 0.2\}$

(b) *syn1:* $\{\pi_+ = 0.7, \rho_\text{d} = 0.3\}$

(c) *syn2:* $\{\pi_+ = 0.7, \rho_\text{d} = 0.2\}$

(d) *syn2:* $\{\pi_+ = 0.7, \rho_\text{d} = 0.3\}$

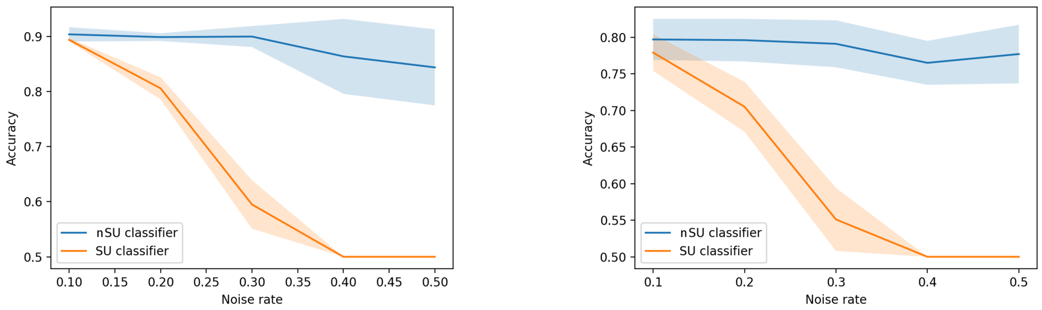
Figure 2: Illustrations based on a single trail of the four setups. The green and pink lines are decision boundaries learned by nSU and SU classification from nSU data respectively. The orange boundary is obtained by SVC with a linear kernel using the fully-supervised data.

over the SU classifier and nSD classifier on all the benchmark-simulated datasets. For some datasets, ITML also achieved comparable accuracy. It is because ITML is a cluster-based method, whose performance is closely associated with the natural structure of the dataset, i.e., whether the *low density separation*[5] condition holds.

### 6.5 Experiments on Text Datasets

SMS Spam (Almeida et al., 2011) is a public set of short message service (SMS) labeled messages that have been collected for mobile phone spam research, which is composed of 5,574 English, real, and non-encoded messages, tagged according to being legitimate or spam. News20 is a collection of approximately 20,000 newsgroup documents, partitioned nearly evenly across 20 different newsgroups. We selected ten newsgroups and paired them into five datasets, i.e., News_05, . . ., News_49. The specific class information is provided in appendix I. For these two datasets, we

[5] The decision boundary should lie in a low-density region.



(a) *syn1 with stepwise noise*

(b) *syn2 with stepwise noise*

Figure 3: Means and Standard Deviations (Percentage) of Classification Accuracy over 5 trials on syn1 and syn2 with stepwise noise. The class-prior $\pi_+$ is fixed at 0.7. When the noise rate is higher than 0.3, the SU classifier even loses the identification ability, and assigns all the test instances the same label, which leads to a steady 0.5 accuracy.

Table 1: Means and Standard Deviations (Percentage) of Classification Accuracy over 5 trials on the UCI and LIBSVM datasets with $\{\pi_+ = 0.7, \rho_\mathrm{d} = 0.2\}$. The best (except SVC) and comparable methods (paired t-test at significance level 5%) are highlighted in bold.

| Dataset | nSU-LC | nSU-DC | SU-LC | SU-DC | nSD | KM | ITML | HC | SVC |
|---|---|---|---|---|---|---|---|---|---|
| australian | **83.1±4.4** | 68.0±8.8 | 63.7±5.4 | 54.3±2.9 | 60.0±2.9 | 69.1±4.2 | **82.2±4.4** | 75.4±11.3 | 84.9±4.1 |
| breast-cancer | **93.6±1.7** | 85.8±7.2 | 84.1±3.8 | 79.4±4.0 | 93.0±1.2 | **95.9±2.4** | 64.1±1.3 | **95.7±2.0** | 96.2±1.3 |
| fourclass | **71.5±6.7** | 64.8±6.2 | 63.7±1.5 | 63.4±2.1 | 64.7±4.8 | 64.8±5.0 | **80.6±5.8** | 63.4±15.8 | 98.2±1.0 |
| magic | **76.2±1.5** | 66.2±2.6 | 69.5±3.5 | 64.5±4.1 | 70.5±3.2 | 61.8±2.3 | 68.8±5.9 | 64.7±1.1 | 80.3±1.0 |
| cod-rna | **89.8±2.7** | **85.6±3.6** | 62.1±12.0 | 63.3±3.7 | 72.3±7.9 | 76.8±0.5 | 83.2±0.3 | 76.5±2.8 | 77.2±0.6 |
| adult | 67.9±1.0 | 59.1±6.8 | 63.6±4.2 | 58.6±3.0 | 75.7±0.3 | 71.1±0.6 | **84.4±4.2** | 72.6±2.3 | 70.5±1.0 |
| banknote | **98.0±1.1** | 62.2±8.2 | **94.9±5.6** | 60.6±8.9 | 62.6±3.9 | 62.9±2.9 | 59.1±7.0 | 66.1±2.0 | 100.0±0.0 |
| heart | **83.3±3.7** | 61.5±12.7 | 58.3±7.6 | 56.3±4.8 | 70.4±6.4 | 63.7±6.6 | **80.0±10.0** | 62.2±8.4 | 60.0±5.5 |
| svmguide1 | **76.1±7.0** | 56.1±4.3 | 52.8±3.4 | 56.6±6.6 | 71.1±0.6 | 72.6±1.0 | **80.3±1.3** | 75.5±12.0 | 93.5±1.2 |
| htru_2 | **96.4±1.3** | 81.1±8.4 | **96.1±1.2** | 86.3±5.7 | 92.0±1.6 | 73.4±2.3 | 86.0±5.0 | 71.0±8.5 | 97.1±0.2 |

used GloVe (Pennington et al., 2014) to extract vector representations from raw text data. Tables 3 and 4 show the consistent superiority of the nSU classifier over the SU classifier and nSD classifier on all the benchmark-simulated datasets. Due to the natural structure of the dataset, the clustering methods occasionally achieved the best performance. However, the clustering methods were not stable, with larger standard deviations.

Table 2: Means and Standard Deviations (Percentage) of Classification Accuracy over 5 trials on the UCI and LIBSVM datasets with $\{\pi_+ = 0.8, \rho_\mathrm{d} = 0.1\}$. The best (except SVC) and comparable methods (paired t-test at significance level 5%) are highlighted in bold.

| Dataset | nSU-LC | nSU-DC | SU-LC | SU-DC | nSD | KM | ITML | HC | SVC |
|---|---|---|---|---|---|---|---|---|---|
| australian | **85.4±3.1** | 55.1±2.4 | 82.6±3.6 | 54.6±2.9 | 64.6±6.7 | 69.1±4.2 | 80.8±5.2 | 78.9±4.1 | 86.7±4.2 |
| breast-cancer | **94.8±1.7** | 66.7±3.7 | 93.0±2.4 | 64.9±2.6 | 89.0±3.3 | **95.9±2.4** | 67.7±8.0 | 94.8±1.9 | 95.4±2.4 |
| fourclass | **70.8±5.9** | 61.8±5.2 | 57.5±5.3 | 62.1±5.1 | 64.3±4.9 | 62.3±6.4 | **80.9±5.7** | 55.2±5.2 | 97.5±1.0 |
| magic | **75.6±2.2** | 63.2±7.4 | 71.4±1.6 | 65.9±1.4 | 68.5±3.0 | 58.3±0.6 | **71.7±19.2** | 63.4±3.6 | 75.9±0.7 |
| cod-rna | **90.1±0.9** | 58.7±10.9 | 71.9±7.0 | 66.7±0.1 | 66.7±7.8 | 77.2±0.5 | 83.2±0.3 | 77.8±0.3 | 59.2±0.3 |
| adult | 74.0±2.2 | 60.9±5.7 | 73.2±2.2 | 57.7±3.4 | 71.1±0.6 | 75.8±0.1 | **84.4±4.2** | 65.3±8.2 | 65.5±0.8 |
| banknote | **97.9±0.7** | 66.7±6.4 | **96.7±2.3** | 63.2±9.5 | 94.1±7.6 | 63.8±3.4 | 83.4±13.5 | 64.1±4.3 | 100.0±0.0 |
| heart | **77.0±8.4** | 63.0±9.4 | 61.5±10.0 | 55.6±2.6 | 63.7±4.1 | 63.7±6.6 | **80.7±8.8** | 63.0±10.1 | 56.6±0.0 |
| svmguide1 | 73.5±2.5 | 65.4±1.8 | 67.3±4.1 | 57.0±4.6 | 67.8±3.5 | 73.3±1.0 | **81.5±2.3** | 70.6±13.1 | 91.1±1.4 |
| htru_2 | **97.8±0.2** | 76.4±9.5 | 97.1±0.5 | 84.0±7.2 | 87.5±3.8 | 77.2±1.5 | 81.3±13.2 | 82.2±7.3 | 97.2±0.3 |

Table 3: Means and Standard Deviations (Percentage) of Classification Accuracy over 5 trials on text datasets with $\{\pi_+ = 0.7, \rho_\mathrm{d} = 0.2\}$. The best (except SVC) and comparable methods (paired t-test at significance level 5%) are highlighted in bold.

| Dataset | nSU-LC | nSU-DC | SU-LC | SU-DC | nSD | KM | ITML | HC | SVC |
|---|---|---|---|---|---|---|---|---|---|
| SMS Spam | 72.4±2.3 | 71.1±5.3 | 56.5±4.4 | 70.4±2.9 | 86.5±0.1 | 69.3±0.7 | **93.1±4.4** | 58.6±9.4 | 84.2±1.5 |
| News_05 | **82.9±2.0** | 59.4±4.9 | 72.4±5.3 | 54.3±1.9 | 56.6±2.3 | 62.7±3.7 | 72.9±8.0 | 61.0±3.6 | 92.4±1.9 |
| News_16 | **77.7±1.9** | 56.8±4.6 | 64.7±2.0 | 52.2±1.4 | 60.0±9.4 | **75.6±3.1** | 71.3±9.2 | 69.3±4.8 | 90.2±2.3 |
| News_27 | **82.9±2.8** | 52.1±1.1 | 78.1±6.3 | 54.3±1.8 | 57.6±8.0 | 62.8±2.9 | **80.7±14.5** | 61.3±3.7 | 96.5±1.7 |
| News_38 | **75.7±3.3** | 55.3±5.0 | 65.4±5.1 | 51.9±1.5 | 55.5±7.0 | 59.2±1.7 | **76.6±4.3** | 59.3±5.5 | 85.4±3.3 |
| News_49 | **84.0±2.2** | 53.4±4.4 | 74.0±5.7 | 54.3±3.9 | 54.1±3.8 | 67.6±5.1 | 75.8±9.4 | 66.3±9.9 | 92.4±1.0 |

Table 4: Means and Standard Deviations (Percentage) of Classification Accuracy over 5 trials on text datasets with $\{\pi_+ = 0.8, \rho_\mathrm{d} = 0.1\}$. The best (except SVC) and comparable methods (paired t-test at significance level 5%) are highlighted in bold.

| Dataset | nSU-LC | nSU-DC | SU-LC | SU-DC | nSD | KM | ITML | HC | SVC |
|---|---|---|---|---|---|---|---|---|---|
| SMS Spam | 75.9±2.2 | 63.0±6.4 | 58.1±3.1 | 70.3±1.5 | 86.5±0.1 | 70.0±1.0 | **93.3±1.4** | 74.3±3.7 | 79.0±1.4 |
| News_05 | **88.5±1.7** | 54.9±3.6 | 83.4±3.5 | 56.6±3.8 | 59.2±5.0 | 63.4±3.3 | 76.0±3.6 | 63.1±2.9 | 89.2±2.4 |
| News_16 | **83.2±2.4** | 53.6±3.6 | 74.7±3.2 | 55.9±7.0 | 53.5±5.9 | 74.5±3.1 | 71.4±5.4 | 67.0±5.4 | 85.3±1.7 |
| News_27 | **92.8±3.0** | 64.6±13.6 | 88.3±1.7 | 60.6±8.8 | 64.4±9.5 | 62.6±2.9 | 74.0±9.4 | 61.5±9.2 | 94.7±1.0 |
| News_38 | **80.6±4.0** | 55.3±2.4 | 75.9±3.5 | 55.4±4.1 | 55.8±9.2 | 58.2±1.8 | 71.7±5.9 | 53.4±2.3 | 81.8±2.0 |
| News_49 | **87.1±3.2** | 58.1±8.6 | 83.8±4.1 | 51.7±1.1 | 59.8±8.7 | 67.7±4.4 | 74.1±4.5 | 65.2±5.8 | 87.8±2.6 |

Table 5: Means and Standard Deviations (Percentage) of Classification Accuracy over 5 trials on image datasets with $\{\pi_+ = 0.7, \rho_\mathrm{d} = 0.2\}$. The best (except SVC) and comparable methods (paired t-test at significance level 5%) are highlighted in bold.

| Dataset | nSU-LC | nSU-DC | SU-LC | SU-DC | nSD | KM | ITML | HC | SVC |
|---|---|---|---|---|---|---|---|---|---|
| Cifar_03 | **70.4±2.9** | 57.8±6.5 | 64.8±4.1 | 51.9±1.5 | 58.6±8.2 | 67.1±1.1 | 64.9±2.6 | 68.7±2.1 | 84.7±1.0 |
| Cifar_14 | **75.0±3.0** | 65.0±3.1 | 73.2±3.3 | 54.0±2.2 | 56.3±7.6 | 63.1±1.9 | 67.3±9.8 | 69.3±4.8 | 88.7±1.0 |
| Cifar_25 | **72.6±1.6** | 56.6±3.9 | 68.0±3.4 | 54.0±2.6 | 62.6±5.4 | 62.8±2.9 | **70.1±9.1** | 61.3±3.7 | 87.3±0.8 |

Table 6: Means and Standard Deviations (Percentage) of Classification Accuracy over 5 trials on image datasets with $\{\pi_+ = 0.8, \rho_\mathrm{d} = 0.1\}$. The best (except SVC) and comparable methods (paired t-test at significance level 5%) are highlighted in bold.

| Dataset | nSU-LC | nSU-DC | SU-LC | SU-DC | nSD | KM | ITML | HC | SVC |
|---|---|---|---|---|---|---|---|---|---|
| Cifar_03 | **72.7±2.2** | 59.0±6.6 | 69.6±2.5 | 51.2±0.7 | 57.0±4.4 | 66.2±0.9 | 65.6±1.4 | 67.2±3.5 | 77.2±1.4 |
| Cifar_14 | **76.7±1.4** | 60.2±7.2 | **75.6±2.6** | 54.7±2.1 | 59.5±4.7 | 62.0±1.4 | **71.3±9.2** | 57.4±4.3 | 85.2±1.8 |
| Cifar_25 | **76.9±1.9** | 64.8±4.0 | 75.1±2.1 | 53.6±2.1 | 65.1±6.3 | 66.4±2.0 | 66.8±3.2 | 66.1±3.3 | 84.1±0.6 |

### 6.6 Experiments on Image Datasets

*CIFAR-10* (Krizhevsky et al., 2009) has $32 \times 32 \times 3$ color images including 50,000 training images and 10,000 test images of 10 classes. Similarly, we selected six classes and paired them into three datasets, i.e., Cifar_03, Cifar_14, and Cifar_25. The specific class information is provided in appendix I. Since the raw feature of *CIFAR-10* is far from a good representation, we extracted the 32-dimensional features of images by a deep variational autoencoder (Kingma and Welling, 2014). Namely, both linear methods (-LC) and deep methods (-DC) in our experiment are fitted on top of the same pre-trained embedding, and take the same advantage of *deep models* extracting good representations. Besides, the linear methods (-LC) have the analytical solution for the objective Eq (14) while the deep methods (-DC) use the SGD method to obtain an approximation to the optimal solution, which introduces additional optimization errors. Therefore, the linear methods (-LC) could outperform the deep methods (-DC) in our experiment, which is not conflict with the fact that generally *deep models* do better than purely *linear models* on image datasets like *CIFAR-10*. Tables 5 and 6 show the consistent superiority of the nSU classifier over the SU classifier, the nSD classifier, and the clutering methods on all the benchmark-simulated datasets.

## 7. Conclusion

In this paper, we proposed a novel weakly supervised learning (WSL) problem named nSU classification, which considers the case where similar data pairs are corrupted with the mutually contaminated distributions model. To tackle this problem, nSU classification provided a robust risk-consistent estimator for learning from nSU data. The mixture proportion estimation (MPE) method was employed to estimate the noise rate and the class-prior probabilities. When utilizing proper models and loss functions, we showed that our optimization problem becomes convex.

Specifically, there exists a closed-form solution to the objective function with a linear-in-parameter model combined with the squared loss. We also established a generalization error bound for the proposed method. Experiments conducted on benchmark datasets demonstrated that our method can excellently solve the aforementioned WSL problem and showed the superiority over the baseline methods. One of the limitations of our work is that for some extreme cases, i.e., heavily unbalanced data with a large noise rate, the parameters cannot be estimated in an MPE manner. Another limitation is that our method cannot deal with the multi-class setting, investigations on which might prove important in future work.

## Acknowledgments

SHW and TLL were supported by Australian Research Council Project No. DE-190101473 and RIKEN Collaborative Research Fund. TLL was also supported by Australian Research Council Projects No. IC-190100031, DP-220102121, and FT-220100318. BH was supported by the RGC Early Career Scheme No. 22200720, NSFC Young Scientists Fund No. 62006202, and Guangdong Basic and Applied Basic Research Foundation No. 2022A1515011652. JY was supported by Natural Science Foundation of China No. 62276242, CAAI-Huawei MindSpore Open Fund No. CAAIXSJLJJ-2021-016B and CAAIXSJLJJ-2022-001A, Anhui Province Key Research and Development Program No. 202104a05020007. GN and MS were supported by JST AIP Acceleration Research Grant No. JPMJCR20U3, Japan. MS was also supported by the Institute for AI and Beyond, UTokyo.

## Appendix A. Proof of Lemma 1

**Proof** Since $\boldsymbol{x}$ and $\boldsymbol{x}'$ are drawn independently, we have

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

■

## Appendix B. Proof and discussion about Theorem 1

**Proof** To begin with, we introduce the following lemma:

**Lemma 3** *(Bao et al., 2018) The classification risk* (1) *can be equivalently expressed as*

$$R_{\mathrm{SU}}(f) = \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{x}')\sim p_{\mathrm{s}}}[\mathcal{L}_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p}[\mathcal{L}'_{\mathrm{U}}(\boldsymbol{x})], \tag{A.1}$$

*where*

$$\begin{aligned}
\mathcal{L}_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}') &= \frac{\pi_{\mathrm{s}}}{2(2\pi_+ - 1)}[\tilde{l}(\boldsymbol{x}) + \tilde{l}(\boldsymbol{x}')], \\
\tilde{l}(\boldsymbol{x}) &= \ell(f(\boldsymbol{x}),+1) - \ell(f(\boldsymbol{x}),-1), \\
\mathcal{L}'_{\mathrm{U}}(\boldsymbol{x}) &= \frac{-\pi_-}{2\pi_+ - 1}\ell(f(\boldsymbol{x}),+1) + \frac{\pi_+}{2\pi_+ - 1}\ell(f(\boldsymbol{x}),-1).
\end{aligned}$$

Combining Eq. (6) and Eq. (7) and organizing, we have

$$p_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') = \frac{1-\pi_{\mathrm{s}}}{1-\rho_{\mathrm{d}}-\pi_{\mathrm{s}}}\tilde{p}_{\mathrm{s}}(\boldsymbol{x},\boldsymbol{x}') - \frac{\rho_{\mathrm{d}}}{1-\rho_{\mathrm{d}}-\pi_{\mathrm{s}}}p(\boldsymbol{x},\boldsymbol{x}'). \tag{A.2}$$

Substituting Eq.(A.2) into Eq.(A.1) we have

$$\begin{aligned}
R(f) &= R_{\mathrm{SU}}(f) \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{x}')\sim\tilde{p}_{\mathrm{s}}}\left[\frac{\pi_{\mathrm{s}}(1-\pi_{\mathrm{s}})}{2(1-\rho_{\mathrm{d}}-\pi_{\mathrm{s}})(2\pi_+-1)}[\ell(f(\boldsymbol{x}),+1) - \ell(f(\boldsymbol{x}),-1) + \ell(f(\boldsymbol{x}'),+1) - \ell(f(\boldsymbol{x}'),-1)]\right] \\
&\quad + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p}\left[\frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_-(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_+-1)}\ell(f(\boldsymbol{x}),+1) - \frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_+(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_+-1)}\ell(f(\boldsymbol{x}),-1)\right] \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{x}')\sim\tilde{p}_{\mathrm{s}}}\left[\frac{\pi_{\mathrm{s}}(1-\pi_{\mathrm{s}})}{2(1-\rho_{\mathrm{d}}-\pi_{\mathrm{s}})(2\pi_+-1)}[\tilde{l}(\boldsymbol{x}) + \tilde{l}(\boldsymbol{x}')]\right] \\
&\quad + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p}\left[\frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_-(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_+-1)}\ell(f(\boldsymbol{x}),+1) - \frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}} - \pi_+(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_+-1)}\ell(f(\boldsymbol{x}),-1)\right] \\
&= \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{x}')\sim\tilde{p}_{\mathrm{s}}}[\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x},\boldsymbol{x}')] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p}[\mathcal{L}_{\mathrm{U}}(\boldsymbol{x})] \\
&= R_{\mathrm{nSU}}(f),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x},\boldsymbol{x}') &= \frac{\pi_{\mathrm{s}}(1-\pi_{\mathrm{s}})}{2(1-\rho_{\mathrm{d}}-\pi_{\mathrm{s}})(2\pi_+-1)}[\tilde{l}(\boldsymbol{x})+\tilde{l}(\boldsymbol{x}')],\\
\tilde{l}(\boldsymbol{x}) &= \ell(f(\boldsymbol{x}),+1)-\ell(f(\boldsymbol{x}),-1),\\
\mathcal{L}_{\mathrm{U}}(\boldsymbol{x}) &= \frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}}-\pi_-(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_+-1)}\ell(f(\boldsymbol{x}),+1)-\frac{\pi_{\mathrm{s}}\rho_{\mathrm{d}}-\pi_+(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)}{(\rho_{\mathrm{d}}+\pi_{\mathrm{s}}-1)(2\pi_+-1)}\ell(f(\boldsymbol{x}),-1).
\end{aligned}$$

■

### B.1 Discussion about $\pi_+ \neq 0.5$ and how a value of $\pi_+$ close to 0.5 affects the results

The direct cause of this requirement $\pi_+ \neq 0.5$ is that the denominator of the risk contains the term $(2\pi_+ - 1)$, which cannot be zero. The intuitive and ultimate cause is that if $\pi_+ = 0.5$, the marginal distributions of similarity data and unlabeled data are the same, i.e., $p_{\mathrm{s}}(\boldsymbol{x}) = p(\boldsymbol{x}) = 0.5p_+(\boldsymbol{x}) + 0.5p_-(\boldsymbol{x})$, but at least 2 marginals with different class priors are required to make comparison and extract contrastive information to make the prediction (Lu et al., 2019). Technically, if $\pi_+ = 0.5$, those terms in the original risk expression, e.g., $\underset{\boldsymbol{x},\boldsymbol{x}'\sim p_+}{\mathbb{E}}\left[\frac{\ell(f(\boldsymbol{x}),+1)+\ell(f(\boldsymbol{x}'),+1)}{2}\right]$ cannot be rewritten as a linear combination of $\underset{(\boldsymbol{x},\boldsymbol{x}')\sim p_{\mathrm{s}}}{\mathbb{E}}\left[\frac{\ell(f(\boldsymbol{x}),+1)+\ell(f(\boldsymbol{x}'),+1)}{2}\right]$ and $\underset{\boldsymbol{x}\sim p}{\mathbb{E}}[\ell(f(\boldsymbol{x}),+1)]$.

A $\pi_+$ value close to 0.5 makes the marginal distributions of $p_s$ and $p$ more similar and makes $p_s$ and $p$ more entangled. Thus, it becomes more difficult to solve the MPE problem and thereby leads to a poor estimation of the parameters. We investigate the effect of $\pi_+$ by conducting experiments on UCI and LIBSVM datasets with $\pi_+ = [0.55, 0.6, 0.7]$ and a constant noise rate $\rho_{\mathrm{d}} = 0.2$. From Table 7, we can see that the classification accuracy of nSU-LC decreases as the class prior approaches 0.5, indicating the parameters are poorly estimated. Most of SU-LC's results decrease while some of them get increased. The reason could be that the poorly estimated parameters accidentally are close to the true values for SU-LC. Overall, our method nSU-LC performs better than the baseline with a $\pi_+$ value close to 0.5.

Table 7: Means and Standard Deviations (Percentage) of Classification Accuracy on the UCI and LIBSVM datasets with $\pi_+ = [0.55, 0.6, 0.7]$ and $\rho_{\mathrm{d}} = 0.2$.

| Dataset | australian | breast-cancer | fourclass | magic | cod-rna | adult | banknote | heart | svmguide1 | htru_2 |
|---|---|---|---|---|---|---|---|---|---|---|
| {0.7, 0.2} | | | | | | | | | | |
| nSU-LC | 83.1±4.2 | 93.6±1.7 | 71.5±6.7 | 76.2±1.5 | 89.8±2.7 | 67.9±1.0 | 98.0±1.1 | 83.3±3.7 | 76.1±7.0 | 96.4±1.3 |
| SU-LC | 63.7±5.4 | 84.1±3.8 | 63.7±1.5 | 69.5±3.5 | 62.1±12.0 | 63.6±4.2 | 94.9±5.6 | 58.3±7.6 | 52.8±3.4 | 96.1±1.2 |
| {0.6, 0.2} | | | | | | | | | | |
| nSU-LC | 75.1±6.1 | 91.3±3.4 | 69.9±7.5 | 72.4±0.8 | 87.7±2.4 | 62.3±3.1 | 97.3±1.4 | 81.5±6.8 | 75.1±6.3 | 92.5±4.6 |
| SU-LC | 59.7±5.5 | 85.2±3.3 | 62.4±0.5 | 69.6±3.2 | 61.2±3.5 | 59.8±3.3 | 93.7±7.5 | 53.1±2.1 | 54.6±6.8 | 95.6±0.5 |
| {0.55, 0.2} | | | | | | | | | | |
| nSU-LC | 66.9±9.2 | 86.4±4.9 | 69.4±9.4 | 67.1±4.3 | 78.4±5.3 | 54.4±4.5 | 90.9±3.6 | 74.1±6.8 | 69.7±8.8 | 86.4±6.3 |
| SU-LC | 55.7±0.4 | 79.4±3.3 | 64.2±0.1 | 66.3±3.6 | 65.4±2.7 | 52.4±3.0 | 90.3±5.6 | 54.6±1.9 | 58.8±7.7 | 93.1±1.9 |

## Appendix C. Discussion about the Data independence assumption

Following Bao et al. (2018), we assume that each instance is independently drawn from the joint distribution, i.e.,

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

However, in real-world surveys, a given positive example might occur very frequently with a few positive examples and very in-frequently with other positive examples. Then the distribution of similar pairs is shifted from the original one. We denote the shifted distribution by $p'_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$ and let $p'_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') = w(\boldsymbol{x}, \boldsymbol{x}')\, p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$. Then we have

$$
\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{x}') &= \pi_{\mathrm{s}} p'_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') + \pi_{\mathrm{d}} p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}'), && \text{(A.3)} \\
\tilde{p}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') &= (1 - \rho_{\mathrm{d}}) p'_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') + \rho_{\mathrm{d}} p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}'), && \text{(A.4)} \\
p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') &= \frac{1 - \pi_{\mathrm{s}}}{(1 - \rho_{\mathrm{d}} - \pi_{\mathrm{s}}) w(\boldsymbol{x}, \boldsymbol{x}')} \tilde{p}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') - \frac{\rho_{\mathrm{d}}}{(1 - \rho_{\mathrm{d}} - \pi_{\mathrm{s}}) w(\boldsymbol{x}, \boldsymbol{x}')} p(\boldsymbol{x}, \boldsymbol{x}'). && \text{(A.5)}
\end{aligned}
$$

Then, by substituting Eq.(A.5) into Eq.(A.1), the original risk can be equivalently expressed in terms of data sampled from the shifted distribution $p'_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$ and the unlabeled data as

$$
\begin{aligned}
\mathop{\mathbb{E}}_{(\boldsymbol{x}, y) \sim p} [\ell(f(\boldsymbol{x}), y)] = {} & \frac{\pi_{\mathrm{s}}(1 - \pi_{\mathrm{s}})}{(1 - \rho_{\mathrm{d}} - \pi_{\mathrm{s}})(2\pi_+ - 1)} \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{x}') \sim \tilde{p}_{\mathrm{s}}} \left[ \frac{\tilde{l}(\boldsymbol{x}) + \tilde{l}(\boldsymbol{x}')}{2w(\boldsymbol{x}, \boldsymbol{x}')} \right] \\
& - \frac{\pi_{\mathrm{s}} \rho_{\mathrm{d}}}{(1 - \rho_{\mathrm{d}} - \pi_{\mathrm{s}})(2\pi_+ - 1)} \mathop{\mathbb{E}}_{(\boldsymbol{x}, \boldsymbol{x}') \sim p} \left[ \frac{\tilde{l}(\boldsymbol{x}) + \tilde{l}(\boldsymbol{x}')}{2w(\boldsymbol{x}, \boldsymbol{x}')} \right] \\
& - \frac{\pi_-}{(2\pi_+ - 1)} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p} [\ell(f(\boldsymbol{x}), +1)] + \frac{\pi_+}{(2\pi_+ - 1)} \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p} [\ell(f(\boldsymbol{x}), -1)] .
\end{aligned}
$$

Therefore, given the weight $w(\boldsymbol{x}, \boldsymbol{x}')$, we can design an unbiased risk estimator according to the above equation. To validate this setting experimentally, we assign $w(\boldsymbol{x}, \boldsymbol{x}')$ to every pair, and the new data is sampled from $p'_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$. We compare the weighted-nSU (wnSU-DC) with the original nSU-DC and SU-DC on UCI and LIBSVM datasets. From Table 8, we can see that in this shifted setting, wnSU-DC outperforms the original nSU-DC and SU-DC. The weight function $w(\boldsymbol{x}, \boldsymbol{x}')$ must be away from 0, and this gap can affect both the optimization stability and the generalization bound where 1/min($w$) will also be in the convergence rate. Another implication of having a weight function is that unbiased risk estimator should be no longer very good to use in practice, because we need to round up too small w which leads to a benign bias (benign in both optimization stability and generalization bound).

Table 8: Means and Standard Deviations (Percentage) of Classification Accuracy on the UCI and LIBSVM datasets with $\{\pi_+ = 0.7, \rho_\mathrm{d} = 0.2\}$ and $\{\pi_+ = 0.8, \rho_\mathrm{d} = 0.1\}$.

| Dataset | australian | breast-cancer | fourclass | magic | cod-rna | adult | banknote | heart | svmguide1 | htru_2 |
|---|---|---|---|---|---|---|---|---|---|---|
| {0.7, 0.2} | | | | | | | | | | |
| wnSU-DC | 63.4±9.2 | 93.6±4.7 | 73.6±8.4 | 66.2±1.1 | 78.6±8.8 | 58.6±6.4 | 81.2±5.8 | 72.2±2.6 | 64.4±10.4 | 90.8±4.0 |
| nSU-DC | 62.9±8.7 | 82.9±4.3 | 72.6±8.3 | 64.2±5.2 | 74.0±7.0 | 58.7±7.1 | 67.6±4.6 | 70.4±10.5 | 64.1±8.0 | 86.8±7.0 |
| SU-DC | 55.7±4.8 | 88.7±7.6 | 71.0±5.7 | 64.1±4.4 | 72.9±6.6 | 63.8±4.5 | 68.1±3.1 | 63.0±5.2 | 54.9±3.9 | 90.0±0.7 |
| {0.8, 0.1} | | | | | | | | | | |
| wnSU-DC | 61.4±8.6 | 90.1±9.6 | 73.2±6.2 | 68.1±4.2 | 78.2±4.0 | 73.8±2.3 | 75.8±9.9 | 71.1±8.0 | 65.0±12.3 | 93.7±2.3 |
| nSU-DC | 60.9±5.2 | 71.0±6.2 | 72.6±6.6 | 64.7±4.8 | 77.4±10.2 | 72.9±3.3 | 75.1±9.1 | 65.9±4.1 | 60.6±4.4 | 85.9±7.7 |
| SU-DC | 58.0±4.5 | 81.4±12.8 | 72.6±9.6 | 66.6±1.8 | 73.9±8.1 | 75.0±0.9 | 75.1±6.4 | 61.5±4.2 | 61.3±3.5 | 85.5±5.1 |

## Appendix D. Proof and discussion about Theorem 2

**Proof** There exists a twice differentiable function $\psi : \mathbb{R} \to \mathbb{R}_+$ such that $\ell(z,t) = \psi(tz)$, because $\ell$ is a twice differentiable margin loss function. Taking the derivative of

$$\begin{aligned}\hat{J}_\ell(\boldsymbol{w}) &= \frac{1}{n_\mathrm{nS}} \sum_{i=1}^{n_\mathrm{nS}} \mathcal{L}_\mathrm{nS}(\boldsymbol{x}_{\mathrm{S},i}, \boldsymbol{x}'_{\mathrm{S},i}) + \frac{1}{n_\mathrm{U}} \sum_{i=1}^{n_\mathrm{U}} \mathcal{L}_\mathrm{U}(\boldsymbol{x}_{\mathrm{U},i}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &= \frac{A}{2n_\mathrm{nS}} \sum_{i=1}^{2n_\mathrm{nS}} [\ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}), +1) - \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}), -1)] \\ &\quad + \frac{B}{n_\mathrm{U}} \sum_{i=1}^{n_\mathrm{U}} [\ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), +1)] - \frac{C}{n_\mathrm{U}} \sum_{i=1}^{n_\mathrm{U}} [\ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), -1)] + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \\ &= \frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} - \frac{A}{2n_\mathrm{nS}} \sum_{i=1}^{2n_\mathrm{nS}} \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}) \\ &\quad + \frac{B}{n_\mathrm{U}} \sum_{i=1}^{n_\mathrm{U}} [\ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), +1)] - \frac{C}{n_\mathrm{U}} \sum_{i=1}^{n_\mathrm{U}} [\ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), -1)],\end{aligned}$$

with respect to $\boldsymbol{w}$,

$$\frac{\partial}{\partial \boldsymbol{w}} \hat{J}_\ell(\boldsymbol{w}) = \lambda \boldsymbol{w} - \frac{A}{2n_\mathrm{nS}} \sum_{i=1}^{2n_\mathrm{nS}} \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}) + \frac{1}{n_\mathrm{U}} \sum_{i=1}^{n_\mathrm{U}} \left\{ B \frac{\partial \ell(\xi_i, +1)}{\partial \xi_i} - C \frac{\partial \ell(\xi_i, -1)}{\partial \xi_i} \right\} \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}),$$

where $\xi_i = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})$.

Note that the second-order derivative of $\ell(z,t)$ with respect to $z$ is

$$\frac{\partial^2 \ell(z,t)}{\partial z^2} = \frac{\partial^2 \psi(tz)}{\partial z^2} = \frac{\partial}{\partial z}\left( t \frac{\partial \psi(\xi)}{\partial \xi} \right) = t^2 \frac{\partial^2 \psi(\xi)}{\partial \xi^2} = \frac{\partial^2 \psi(\xi)}{\partial \xi^2},$$

where $\xi = tz$ is employed in the second equality and the last equality holds because $t \in \{+1, -1\}$.

The Hessian matrix is a square matrix of second-order partial derivatives of a scalar-valued function. A twice continuously differentiable function of several variables is convex on a convex set if and only if its Hessian matrix of second partial derivatives is positive semidefinite on the interior of the convex set. For $\hat{J}_\ell$, its Hessian matrix is

$$\begin{aligned}
\boldsymbol{H}\hat{J}_\ell(\boldsymbol{w}) &= \lambda I + \frac{1}{n_\mathrm{U}}\sum_{i=1}^{n_\mathrm{U}}\left\{B\frac{\partial}{\partial w}\frac{\partial \ell(\xi_i,+1)}{\partial \xi_i} - C\frac{\partial}{\partial w}\frac{\partial \ell(\xi_i,-1)}{\partial \xi_i}\right\}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})^\top \\
&= \lambda I + \frac{1}{n_\mathrm{U}}\sum_{i=1}^{n_\mathrm{U}}\left\{B\frac{\partial^2 \ell(\xi_i,+1)}{\partial \xi_i^2}\frac{\partial \xi_i}{\partial w} - C\frac{\partial^2 \ell(\xi_i,-1)}{\partial \xi_i^2}\frac{\partial \xi_i}{\partial w}\right\}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})^\top \\
&= \lambda I + \frac{1}{n_\mathrm{U}}\sum_{i=1}^{n_\mathrm{U}}\left\{B\frac{\partial^2 \ell(\xi_i,+1)}{\partial \xi_i^2} - C\frac{\partial^2 \ell(\xi_i,-1)}{\partial \xi_i^2}\right\}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})^\top \\
&= \lambda I + \frac{1}{n_\mathrm{U}}\sum_{i=1}^{n_\mathrm{U}}(B-C)\frac{\partial^2 \psi(\xi)}{\partial \xi^2}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})^\top \\
&= \lambda I + \frac{1}{n_\mathrm{U}}\frac{\partial^2 \psi(\xi)}{\partial \xi^2}\sum_{i=1}^{n_\mathrm{U}}\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})^\top.
\end{aligned}$$

Since $\ell$ is convex, $\frac{\partial^2 \psi(\xi)}{\partial \xi^2} \geq 0$. Besides, $\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})^\top \succeq 0$. Therefore $\boldsymbol{H}\hat{J}_\ell(\boldsymbol{w}) \succeq 0$, and $\hat{J}_\ell(\boldsymbol{w})$ is convex. ■

Examples of marginal loss functions other than the squared loss function that satisfy the condition in Theorem 2 are logistic loss and double hinge loss.

If we focus on the margin loss function $\psi$, the condition in Theorem 2 can be relaxed to that the corresponding loss $\ell$ is $\alpha$-linear-odd: $\ell(x,1) - \ell(x,-1) = \alpha x$, for any $\alpha \in \mathbb{R}$ (Patrini et al., 2016). This condition is both sufficient and necessary for the composite loss to be convex. The sufficiency can be proved in the same way above. Below we prove this condition is necessary.

Without additional assumptions, the objective function can be decomposed as follows

$$\begin{aligned}
\widehat{J}_\ell(\boldsymbol{w}) &\triangleq \frac{\pi_\mathrm{S}}{2n_\mathrm{S}}\sum_{i=1}^{2n_\mathrm{S}}\mathcal{L}_{\mathrm{S},\ell}\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i})\right)+\frac{1}{n_\mathrm{U}}\sum_{i=1}^{n_\mathrm{U}}\mathcal{L}_{\mathrm{U},\ell}\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})\right)+\frac{\lambda}{2}\|\boldsymbol{w}\|^2\\
&=\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}+\frac{\pi_\mathrm{S}}{2n_\mathrm{S}\left(2\pi_+-1\right)}\sum_{i=1}^{2n_\mathrm{S}}\left\{\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}),+1\right)-\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}),-1\right)\right\}\\
&\quad+\frac{1}{n_\mathrm{U}\left(2\pi_+-1\right)}\sum_{i=1}^{n_\mathrm{U}}\left\{-\pi_-\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}),+1\right)+\pi_+\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}),-1\right)\right\}\\
&=\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}+\frac{\pi_\mathrm{S}}{2n_\mathrm{S}\left(2\pi_+-1\right)}\sum_{i=1}^{2n_\mathrm{S}}\left\{\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}),+1\right)-\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{S},i}),-1\right)\right\}\\
&\quad-\frac{\pi_-}{n_\mathrm{U}\left(2\pi_+-1\right)}\sum_{i=1}^{n_\mathrm{U}}\left\{\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}),+1\right)-\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}),-1\right)\right\}\\
&\quad+\frac{1}{n_\mathrm{U}\left(2\pi_+-1\right)}\sum_{i=1}^{n_\mathrm{U}}\left\{(\pi_+-\pi_-)\ell\left(\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}),-1\right)\right\}.
\end{aligned}$$

Notice that the first and the last terms are both convex with respect to $w$. Since the second and third terms cannot be convex simultaneous unless they are both linear in $w$. Therefore, to make this objective function convex for arbitrary data and hyper-parameters, $(\ell(x,1)-\ell(x,-1))$ must be linear in $x$.

## Appendix E. Proof of Theorem 3

**Theorem 5** *Assume the positive data distribution $P_+$ and the negative data distribution $P_-$ are mutually irreducible, then $P_\mathrm{d}$ is irreducible with respect to $\tilde{P}_\mathrm{s}$, and $P_\mathrm{s}$ is irreducible with respect to $P$. Thus the mixture proportion $\gamma$ and $\kappa$ in Lemma 2 is identifiable.*

**Proof** Since the positive data distribution $P_+$ and the negative data distribution $P_-$ are mutually irreducible, the following two equations hold:

$$\inf_{S\in\mathfrak{S},P_-(S)>0}\frac{P_+(S)}{P_-(S)}=0, \tag{A.6}$$

$$\inf_{S\in\mathfrak{S},P_+(S)>0}\frac{P_-(S)}{P_+(S)}=0. \tag{A.7}$$

For $P_\mathrm{s}$ and $P_\mathrm{d}$,

$$\begin{aligned}
\frac{P_\mathrm{d}(S,S')}{P_\mathrm{s}(S,S')}&=\frac{(\pi_+^2+\pi_-^2)(P_+(S)P_-(S')+P_-(S)P_+(S'))}{2(\pi_+^2P_+(S)P_+(S')+\pi_-^2P_-(S)P_-(S'))}\\
&=\frac{(\pi_+^2+\pi_-^2)(P_-(S')/P_+(S')+P_-(S)/P_+(S))}{2(\pi_+^2+\pi_-^2(P_-(S)/P_+(S))(P_-(S')/P_+(S')))}.
\end{aligned}$$

According to Eq. (A.7),

$$\inf_{S,S'\in\mathfrak{S},P_\mathrm{s}(S,S')>0}\frac{P_\mathrm{d}(S,S')}{P_\mathrm{s}(S,S')}=0. \tag{A.8}$$

Similarly we have,

$$\inf_{S,S' \in \mathfrak{S}, P_{\mathrm{d}}(S,S')>0} \frac{P_{\mathrm{s}}(S,S')}{P_{\mathrm{d}}(S,S')} = 0. \tag{A.9}$$

Thus, $P_{\mathrm{s}}$ and $P_{\mathrm{d}}$ are mutually irreducible.
For $P_{\mathrm{d}}$ and $\tilde{P}_{\mathrm{s}}$,

$$\begin{aligned} \frac{P_{\mathrm{d}}(S,S')}{\tilde{P}_{\mathrm{s}}(S,S')} &= \frac{P_{\mathrm{d}}(S,S')}{(1-\rho_{\mathrm{d}})P_{\mathrm{s}}(S,S') + \rho_{\mathrm{d}}P_{\mathrm{d}}(S,S')} \\ &= \frac{P_{\mathrm{d}}(S,S')/P_{\mathrm{s}}(S,S')}{(1-\rho_{\mathrm{d}}) + \rho_{\mathrm{d}}P_{\mathrm{d}}(S,S')/P_{\mathrm{s}}(S,S')}. \end{aligned}$$

According to Eq. (A.8),

$$\inf_{S,S' \in \mathfrak{S}, \tilde{P}_{\mathrm{s}}(S,S')>0} \frac{P_{\mathrm{d}}(S,S')}{\tilde{P}_{\mathrm{s}}(S,S')} = 0. \tag{A.10}$$

Similarly we have,

$$\inf_{S,S' \in \mathfrak{S}, P(S,S')>0} \frac{P_{\mathrm{s}}(S,S')}{P(S,S')} = 0. \tag{A.11}$$

Therefore, $P_{\mathrm{d}}$ is irreducible with respect to $\tilde{P}_{\mathrm{s}}$, and $P_{\mathrm{s}}$ is irreducible with respect to $P$. ■

## Appendix F. Proof of Theorem 4

Note that the loss function is symmetric to $\boldsymbol{x}_{\mathrm{S},i}$ and $\boldsymbol{x}'_{\mathrm{S},i}$, such that the expected and empirical risks can be expressed as

$$\begin{aligned} R_{\mathrm{nSU}}(f) &= \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\tilde{p}_{\mathrm{s}}}[\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x})] + \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p}[\mathcal{L}_{\mathrm{U}}(\boldsymbol{x})], \\ \hat{R}_{\mathrm{nSU}}(f) &= \frac{1}{2n_{\mathrm{nS}}}\sum_{i=1}^{2n_{\mathrm{nS}}} \mathcal{L}_{\mathrm{nS}}(\boldsymbol{x}_{\mathrm{S},i}) + \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}} \mathcal{L}_{\mathrm{U}}(\boldsymbol{x}_{\mathrm{U},i}). \end{aligned}$$

Let $R_{\mathrm{nS}}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\tilde{p}_{\mathrm{s}}}[\mathcal{L}_{\mathrm{nS}}(\boldsymbol{x})]$, $R_{\mathrm{U}}(f) = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim p}[\mathcal{L}_{\mathrm{U}}(\boldsymbol{x})]$,$\hat{R}_{\mathrm{nS}}(f) = \frac{1}{2n_{\mathrm{nS}}}\sum_{i=1}^{2n_{\mathrm{nS}}} \mathcal{L}_{\mathrm{nS}}(\boldsymbol{x}_{\mathrm{S},i})$ and $\hat{R}_{\mathrm{U}}(f) = \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}} \mathcal{L}_{\mathrm{U}}(\boldsymbol{x}_{\mathrm{U},i})$, we have

$$\begin{aligned} R(\hat{f}) - R(f^*) &= R_{\mathrm{nSU}}(\hat{f}) - R_{\mathrm{nSU}}(f^*) \\ &= (R_{\mathrm{nSU}}(\hat{f}) - \hat{R}_{\mathrm{nSU}}(\hat{f})) + (\hat{R}_{\mathrm{nSU}}(\hat{f}) - \hat{R}_{\mathrm{nSU}}(f^*)) \\ &\quad + (\hat{R}_{\mathrm{nSU}}(f^*) - R_{\mathrm{nSU}}(f^*)) \\ &\le (R_{\mathrm{nSU}}(\hat{f}) - \hat{R}_{\mathrm{nSU}}(\hat{f})) + 0 + (\hat{R}_{\mathrm{nSU}}(f^*) - R_{\mathrm{nSU}}(f^*)) \\ &\le 2\sup_{f\in\mathcal{F}} \left| R_{\mathrm{nSU}}(f) - \hat{R}_{\mathrm{nSU}}(f) \right| \\ &\le 2\sup_{f\in\mathcal{F}} \left| R_{\mathrm{nS}}(f) - \hat{R}_{\mathrm{nS}}(f) \right| + 2\sup_{f\in\mathcal{F}} \left| R_{\mathrm{U}}(f) - \hat{R}_{\mathrm{U}}(f) \right|. \end{aligned} \tag{A.12}$$

The third inequality holds because of the definition of $f^*$.

Here, we introduce the generalization error with Rademacher complexity.

**Lemma 4** *(Bartlett and Mendelson, 2002) Let the loss function be upper bounded by $M$. Then, for any $\delta > 0$, with the probability $1 - \delta$, we have*

$$\sup_{f\in\mathcal{F}} |\,\mathbb{E}[f(x)] - \frac{1}{n}\sum_{i=1}^{n} f(x_i)| \leq 2\mathfrak{R}_n(\ell \circ \mathcal{F}) + M\sqrt{\frac{\log 1/\delta}{2n}}, \tag{A.13}$$

*where $\mathfrak{R}_n(\ell \circ \mathcal{F})$ is the Rademacher complexity defined by*

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(\boldsymbol{x}_i), f(\boldsymbol{x}_{i'}), \bar{S}_{ii'})\right], \tag{A.14}$$

*and $\{\sigma_1, \cdots, \sigma_n\}$ are Rademacher variables uniformly distributed from $\{-1, 1\}$.*

Now we can bound two terms in Eq.(A.12) with the next two lemmas.

**Lemma 5** *Assume the loss function $\ell$ is $\rho$-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant $C_\mathrm{b}$ such that $\|f\|_\infty \leq C_\mathrm{b}$ for any $f \in \mathcal{F}$. Let $C_\ell \triangleq \sup_{t\in\{\pm 1\}} \ell(C_\mathrm{b}, t)$. For any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$,*

$$\sup_{f\in\mathcal{F}} \left| R_{N\tilde{S}}(f) - \hat{R}_{N\tilde{S}}(f) \right| \leq \frac{4A\rho C_\mathcal{F} + A\sqrt{2C_\ell^2 \log\frac{4}{\delta}}}{\sqrt{2n_\mathrm{nS}}}.$$

**Proof** By Lemma 4,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left| R_{N\tilde{S}}(f) - \hat{R}_{N\tilde{S}}(f) \right| \\
&\quad = A \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\tilde{p}_\mathrm{s}} [\ell(f(\boldsymbol{x}), +1) - \ell(f(\boldsymbol{x}), -1)] - \frac{1}{2n_\mathrm{nS}} \sum_{i=1}^{2n_\mathrm{nS}} [\ell(f(\boldsymbol{x}_{\mathrm{S},i}), +1) - \ell(f(\boldsymbol{x}_{\mathrm{S},i}), -1)] \right| \\
&\quad \leq A \left\{ \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\tilde{p}_\mathrm{s}} [\ell(f(\boldsymbol{x}), +1)] - \frac{1}{2n_\mathrm{nS}} \sum_{i=1}^{2n_\mathrm{nS}} \ell(f(\boldsymbol{x}_{\mathrm{S},i}), +1) \right| \right. \\
&\qquad \left. + \sup_{f\in\mathcal{F}} \left| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\tilde{p}_\mathrm{s}} [\ell(f(\boldsymbol{x}), -1)] - \frac{1}{2n_\mathrm{nS}} \sum_{i=1}^{2n_\mathrm{nS}} \ell(f(\boldsymbol{x}_{\mathrm{S},i}), -1) \right| \right\} \\
&\quad \leq A \left\{ 4\mathfrak{R}(\ell \circ \mathcal{F}; 2n_\mathrm{nS}, \tilde{p}_\mathrm{s}) + \sqrt{\frac{2C_\ell^2 \log\frac{4}{\delta}}{2n_\mathrm{nS}}} \right\},
\end{aligned}$$

where $\ell \circ \mathcal{F}$ in the last line means $\{\ell \circ f | f \in \mathcal{F}\}$. The last inequality holds from Lemma 4. By Talagrand's lemma (Lemma 4.2 in (Mohri et al., 2018)),

$$\mathfrak{R}(\ell \circ \mathcal{F}; 2n_\mathrm{nS}, \tilde{p}_\mathrm{s}) \leq \rho\mathfrak{R}(\mathcal{F}; 2n_\mathrm{nS}, \tilde{p}_\mathrm{s}).$$

Together with $\mathfrak{R}(\mathcal{F}; n, \mu) \leq \frac{C_{\mathcal{F}}}{\sqrt{n}}$, we obtain

$$\sup_{f \in \mathcal{F}} \left| R_{N\tilde{S}}(f) - \hat{R}_{N\tilde{S}}(f) \right| \leq A \left\{ 4\rho \frac{C_{\mathcal{F}}}{\sqrt{2n_{\mathrm{nS}}}} + \sqrt{\frac{2C_{\ell}^2 \log \frac{4}{\delta}}{2n_{\mathrm{nS}}}} \right\}$$

$$= \frac{4A\rho C_{\mathcal{F}} + A\sqrt{2C_{\ell}^2 \log \frac{4}{\delta}}}{\sqrt{2n_{\mathrm{nS}}}}.$$

**Lemma 6** *Assume the loss function $\ell$ is $\rho$-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant $C_{\mathrm{b}}$ such that $\|f\|_{\infty} \leq C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_{\ell} \triangleq \sup_{t \in \{\pm 1\}} \ell(C_{\mathrm{b}}, t)$. For any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$,*

$$\sup_{f \in \mathcal{F}} \left| R_{\mathrm{U}}(f) - \hat{R}_{\mathrm{U}}(f) \right| \leq \frac{2(-B - C)\rho C_{\mathcal{F}} + (-B - C)\sqrt{\frac{1}{2}C_{\ell}^2 \log \frac{4}{\delta}}}{\sqrt{n_{\mathrm{U}}}}.$$

This lemma can be proven similarly to Lemma 5.

Combining Lemma 5, Lemma 6 and Eq. (A.12), Theorem 4 is proven. ■

If we further take the MPE error into consideration, there is a gap between the ground-truth empirical risk $\hat{R}_{\mathrm{nSU}}(f)$ and the approximated empirical risk $\hat{\hat{R}}_{\mathrm{nSU}}(f)$, which uses the estimated class prior and noise rate. We have

$$\begin{aligned} |\hat{R}_{\mathrm{nSU}}(f) - \hat{\hat{R}}_{\mathrm{nSU}}(f)| = &|\frac{A - \hat{A}}{2N_{\mathrm{nS}}} \sum_{i=1}^{2N_{\mathrm{nS}}} [\ell(f(\boldsymbol{x}_{\mathrm{S},i}), +1) - \ell(f(\boldsymbol{x}_{\mathrm{S},i}), -1)] \\ &+ \frac{B - \hat{B}}{N_{\mathrm{U}}} \sum_{i=1}^{N_{\mathrm{U}}} [\ell(f(\boldsymbol{x}_{\mathrm{U},i}, +1)] - \frac{C - \hat{C}}{N_{\mathrm{U}}} \sum_{i=1}^{N_{\mathrm{U}}} [\ell(f(\boldsymbol{x}_{\mathrm{U},i}), -1)]| \\ \leq &\left| \frac{A - \hat{A}}{2N_{\mathrm{nS}}} \sum_{i=1}^{2N_{\mathrm{nS}}} [\ell(f(\boldsymbol{x}_{\mathrm{S},i}), +1) - \ell(f(\boldsymbol{x}_{\mathrm{S},i}), -1)] \right| \\ &+ \left| \frac{B - \hat{B}}{N_{\mathrm{U}}} \sum_{i=1}^{N_{\mathrm{U}}} [\ell(f(\boldsymbol{x}_{\mathrm{U},i}), +1)] \right| + \left| \frac{C - \hat{C}}{N_{\mathrm{U}}} \sum_{i=1}^{N_{\mathrm{U}}} [\ell(f(\boldsymbol{x}_{\mathrm{U},i}), -1)] \right|, \end{aligned}$$

where $\hat{A}, \hat{B}, \hat{C}$ are the corresponding estimated ones.

Ideally, if we have the knowledge of the exact class priors and noise rate parameters, those risks can be directly calculated since they are both empirical risks. If not, according to the Theorem 12 in (Ramaswamy et al., 2016), we know that the estimated value $\hat{\lambda}$ converges to the true value $\lambda$ with a rate $\mathcal{O}(m^{-\frac{1}{2}})$:

$$|\lambda - \hat{\lambda}| \leq |\alpha(\lambda) m^{-\frac{1}{2}}|,$$

where $\alpha(\lambda)$ is the coefficient and $m$ is the smaller number of data from two proportions, i.e., $F$ and $H$ in the MPE. Let $\pi_{\mathrm{s}} = g(\kappa, \gamma) = \frac{\gamma(1-\kappa)}{1-\gamma\kappa}$. The Taylor series of $\pi_{\mathrm{s}}$ at the true value $\pi_{\mathrm{s}}^*$ $(\kappa^*, \gamma^*)$ is:

$$\pi_{\mathrm{s}} = g(\kappa, \gamma) = g(\kappa^*, \gamma^*) + (\kappa - \kappa^*) g_{\kappa}^{'}(\kappa^*, \gamma^*) + (\gamma - \gamma^*) g_{\kappa}^{'}(\kappa^*, \gamma^*) + o^n.$$

By omitting the high-order terms, we have:

$$\begin{aligned}|\pi_{\mathrm{s}} - \pi_{\mathrm{s}}^*| &\leq |(\kappa - \kappa^*)g_{\kappa}^{'}(\kappa^*, \gamma^*) + (\gamma - \gamma^*)g_{\kappa}^{'}(\kappa^*, \gamma^*)| \\ &\leq \left(|\alpha(\kappa^*)g_{\kappa}^{'}(\kappa^*, \gamma^*)| + |\alpha(\gamma^*)g_{\kappa}^{'}(\kappa^*, \gamma^*)|\right)\left|m^{-\frac{1}{2}}\right|,\end{aligned}$$

which indicates that the convergence rate for $\pi_{\mathrm{s}}$ is $\mathcal{O}(m^{-\frac{1}{2}})$. Likewise, $\rho_{\mathrm{d}}$ and $\pi_+$ both have the convergence rate of $\mathcal{O}(m^{-\frac{1}{2}})$. Further, given that $\pi_{\mathrm{s}}$, $\rho_{\mathrm{d}}$, and $\pi_+$ all converge with a rate of $\mathcal{O}(m^{-\frac{1}{2}})$, similarly, we have that $A$, $B$, and $C$ all have the same convergence rate of $\mathcal{O}(m^{-\frac{1}{2}})$.

## Appendix G. Motivation for the noise model

The class-conditional noise model is compelling for the setting that the label of a pair of examples $(\boldsymbol{x}, \boldsymbol{x}')$ is annotated by human who can inevitably make mistakes, i.e., the corruption is in $P\left(S|\boldsymbol{x}, \boldsymbol{x}'\right)$. However, this setting is not very suitable for our problem because we only collect similar data pairs. Therefore, we do not have seeming dissimilar data pairs and only modelling the corruption of $P\left(S|\boldsymbol{x}, \boldsymbol{x}'\right)$ cannot solve our problem. Besides, as we discuss in the related work, the CCN model is a special case of the MCD model (Menon et al., 2015). CCN model does not fit the MCD setting problem, though the MCD model fits the CCN setting problem conversely. Namely, our method, which is developed under the MCD model, can also solve the CCN problem.

In addition to its good generality, we employ the MCD model because CCN is a noise model for labeling noise and MCD is a noise model for sampling noise, which is why MCD is more suitable for the scenario of surveying sensitive topics with indirect questioning. Data reliability is a common concern especially when asking about sensitive topics such as sexual misconduct, or drug and alcohol abuse. Sensitive topics might cause refusals in surveys due to privacy concerns of the subjects (Oral, 2019). This nonresponse reduces sample size and study power and increases bias. Various indirect questioning methods have been developed to reduce the social desirability bias and increase data reliability. Questions in the form of 'With whom do you share the same opinion on issue $\mathcal{I}$?' can be regarded as one type of randomized response technique, which is a commonly used indirect questioning survey method (Warner, 1965; Fisher, 1993; Bao et al., 2018). Such questioning is to sample examples of similar data pairs from $p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}')$. Besides, due to the sensitivity of the questions, respondents might answer them in a manner that will be viewed favorably by others instead of answering truthfully (Oral, 2019), which makes the selected examples contain dissimilar data pairs from $p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}')$. These phenomena motivate us to employ the contamination model to describe the noisy similar data pairs, i.e., $\tilde{p}_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') = (1 - \rho_{\mathrm{d}})p_{\mathrm{s}}(\boldsymbol{x}, \boldsymbol{x}') + \rho_{\mathrm{d}}p_{\mathrm{d}}(\boldsymbol{x}, \boldsymbol{x}')$.

## Appendix H. Why assumption $1 - \rho_{\mathrm{d}} > \pi_{\mathrm{s}}$ generally holds true

Note that the physical meanings of $\pi_{\mathrm{s}}$ and $1 - \rho_{\mathrm{d}}$ are the pro portion of similar data pairs in unlabeled data $D_{\mathrm{u}}$ and noisy similarity data $\tilde{D}_{\mathrm{s}}$. Generally, even though $\tilde{D}_{\mathrm{s}}$ contains some noise, it still collects the similar data pairs purposely, and thereby the noise is not too large. Thus, the proportion of similar data pairs in purposely collected noisy similarity data $\tilde{D}_{\mathrm{s}}$ (i.e., $1 - \rho_{\mathrm{d}}$) is generally bigger than than in unlabeled data $D_{\mathrm{u}}$ (i.e., $\pi_{\mathrm{s}}$). Besides, if this condition is not satisfied, it becomes a different problem and thus we need a new solution rather than still solving this problem in the proposed way.

## Appendix I. Specific class information regarding *News20* and *CIFAR-10* datasets

Table 9: The relationships between the semantic classes in the original News20 and the classes selected in the News_05, $\cdots$, News_49 datasets.

| Dataset | Positive | Negative |
|---|---|---|
| News_05 | alt.atheism | comp.graphics |
| News_16 | misc.forsale | rec.autos |
| News_27 | talk.politics.mideast | comp.sys.ibm.pc.hardware |
| News_38 | comp.os.ms-windows.misc | sci.crypt |
| News_49 | sci.space | sci.med |

Table 10: The relationships between the semantic classes in the original CIFAR-10 and the classes selected in the Cifar_03, Cifar_14, and Cifar_25 datasets.

| Dataset | Positive | Negative |
|---|---|---|
| Cifar_03 | airplane | dog |
| Cifar_14 | cat | ship |
| Cifar_25 | deer | truck |

## References

Tiago A Almeida, José María G Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: new collection and results. In *Proceedings of the 11th ACM Symposium on Document Engineering*, 2011.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Yingbin Bai, Erkun Yang, Zhaoqing Wang, Yuxuan Du, Bo Han, Cheng Deng, Dadong Wang, and Tongliang Liu. Msr: Making self-supervised learning robust to aggressive augmentations. In *NeurIPS*, 2022.

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *ICML*, 2018.

Han Bao, Takuya Shimada, Liyuan Xu, Issei Sato, and Masashi Sugiyama. Pairwise supervision can provably elicit a decision boundary. In *AISTATS*, 2022.

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

Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In *ICML*, 2002.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.

Daniele Calandriello, Gang Niu, and Masashi Sugiyama. Semi-supervised information-maximization clustering. *Neural Networks*, 57:103–111, 2014.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *CVPR*, 2022a.

De Cheng, Yixiong Ning, Nannan Wang, Xinbo Gao, Heng Yang, Yuxuan Du, Bo Han, and Tongliang Liu. Class-dependent label-noise learning with cycle-consistency regularization. In *NeurIPS*, 2022b.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Soham Dan, Han Bao, and Masashi Sugiyama. Learning from noisy similar and dissimilar data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2021.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

Marthinus du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *ICML*, 2015.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Robert J Fisher. Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2):303–315, 1993.

Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, 1999.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *ICCV*, 2019.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *ICLR*, 2019.

Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. Weakly supervised image classification through noise regularization. In *CVPR*, 2019.

Yang Hu, Jingdong Wang, Nenghai Yu, and Xian-Sheng Hua. Maximum margin clustering with pairwise constraints. In *ICDM*, 2008.

Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *ICLR*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *ECCV*, 2016.

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

Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Brian Kulis et al. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4): 287–364, 2012.

Chenglong Li, Chengli Zhu, Yan Huang, Jin Tang, and Liang Wang. Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking. In *ECCV*, 2018.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *CVPR*, 2022a.

Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label cor- relations for noisy multi-label learning. In *NeurIPS*, 2022b.

Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, 2017.

Zhenguo Li and Jianzhuang Liu. Constrained clustering by spectral kernel learning. In *ICCV*, 2009.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2016.

Nan Lu, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In *ICLR*, 2019.

James MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, 2015.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NeurIPS*, 2013.

Gang Niu, Bo Dai, Makoto Yamada, and Masashi Sugiyama. Information-theoretic semi-supervised metric learning via entropy regularization. *Neural Computation*, 26(8):1717–1762, 2014.

Evrim Oral. Surveying sensitive topics with indirect questioning. In *Statistical Methodologies*. IntechOpen, 2019.

Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *ICML*, 2016.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *ICML*, 2016.

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

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT*, 2013.

Takuya Shimada, Han Bao, Issei Sato, and Masashi Sugiyama. Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization. *Neural Computation*, 33 (5):1234–1268, 2021.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation.* MIT Press, 2012.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, and Tomoya Sakai. *Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach.* MIT Press, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *NeurIPS*, 1991.

Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl. Constrained k-means clustering with background knowledge. In *ICML*, 2001.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Nannan Wang, Haifeng Liu, and Gang Niu. Class2simi: A noise reduction perspective on learning with noisy labels. In *ICML*, 2021.

Songhua Wu, Mingming Gong, Bo Han, Yang Liu, and Tongliang Liu. Fair classification with instance-dependent label noise. In *First Conference on Causal Learning and Reasoning*, 2022.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, 2019.

Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *ICLR*, 2022.

Eric P Xing, Michael I Jordan, Stuart J Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *NeurIPS*, 2003.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *ICML*, 2022.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Gang Niu, Masashi Sugiyama, and Dacheng Tao. Rethinking class-prior estimation for positive-unlabeled learning. In *ICLR*, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53, 2017.