LETTER Communicated by Yang Liu

Classification From Pairwise Similarities/Dissimilarities and Unlabeled Data via Empirical Risk Minimization

Takuya Shimada
shima@ms.k.u-tokyo.ac.jp
Han Bao
tsutsumi@ms.k.u-tokyo.ac.jp
Issei Sato
issei.sato@is.s.u-tokyo.ac.jp
University of Tokyo, Bunkyo-ku, Tokyo, 113-0333, Japan, and RIKEN Center for Advanced Intelligence Project, Chuo-ku, Tokyo 103-0027, Japan

Masashi Sugiyama
sugi@k.u-tokyo.ac.jp
RIKEN Center for Advanced Intelligence Project, Chuo-ku, Tokyo 103-0027, Japan, and University of Tokyo, Bunkyo-ku, Tokyo, 113-0333, Japan

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Pairwise similarities and dissimilarities between data points are often obtained more easily than full labels of data in real-world classification problems. To make use of such pairwise information, an empirical risk minimization approach has been proposed, where an unbiased estimator of the classification risk is computed from only pairwise similarities and unlabeled data. However, this approach has not yet been able to handle pairwise dissimilarities. Semisupervised clustering methods can incorporate both similarities and dissimilarities into their framework; however, they typically require strong geometrical assumptions on the data distribution such as the manifold assumption, which may cause severe performance deterioration. In this letter, we derive an unbiased estimator of the classification risk based on all of similarities and dissimilarities and unlabeled data. We theoretically establish an estimation error bound and experimentally demonstrate the practical usefulness of our empirical risk minimization method.

1 Introduction

In supervised classification, we need a vast amount of labeled data to train our classifiers. However, it is often not easy to obtain such labels due to high labeling costs (Chapelle, Schölkopf, & Zien, 2010), privacy concerns

T.S. is now with Preferred Networks, Japan.

Neural Computation **33, 1234–1268** (2021) © 2021 Massachusetts Institute of Technology
https://doi.org/10.1162/neco_a_01373

(Warner, 1965), and social bias (Nederhof, 1985). In real-world classification problems, pairwise similarities (i.e., pairs of samples in the same class) and pairwise dissimilarities (i.e., pairs of samples in different classes) are often collected more easily than full labels of data. For example, in protein function prediction, the knowledge about similarities and dissimilarities can be obtained by experimental means as additional supervision (Klein, Kamvar, & Manning, 2002). In video object classification, knowledge of temporal relations can be used to generate pairwise labels in an algorithmic way; for example, an object staying in temporally adjacent frames must be the same, and two objects in the same frame must be different (Yan, Zhang, Yang, & Hauptmann, 2006; Zhang & Yan, 2007). To make use of such pairwise information, similar-unlabeled (SU) classification (Bao, Niu, & Sugiyama, 2018) has been proposed, where the classification risk is estimated in an unbiased fashion from only similar pairs and unlabeled data. Although their method can handle only similar data and unlabeled data, we may also obtain dissimilar pairs in practice. In such a case, we can expect that the use of dissimilarities, in addition to similarities and unlabeled data, improves the classification accuracy.

Semisupervised clustering (Wagstaff, Cardie, Rogers, & Schrödl, 2001) is a method that can incorporate both similar and dissimilar pairs into their framework, where must-link pairs (i.e., similar pairs) and cannot-link pairs (i.e., dissimilar pairs) are used to obtain meaningful clusters. Existing literature provides useful semisupervised clustering methods based on the ideas that (1) must/cannot-links are treated as constraints (Basu, Banerjee, & Mooney, 2002; Wagstaff et al., 2001; Li & Liu, 2009; Hu, Wang, Yu, & Hua, 2008), (2) clustering is performed with metrics learned by semisupervised metric learning (Xing, Jordan, Russell, & Ng, 2003; Bilenko, Basu, & Mooney, 2004; Weinberger & Saul, 2009; Davis, Kulis, Jain, Sra, & Dhillon, 2007; Niu, Dai, Yamada, & Sugiyama, 2012), and (3) missing links are predicted by matrix completion (Yi, Zhang, Jin, Qian, & Jain, 2013; Chiang, Hsieh, & Dhillon, 2015). However, the motivation of clustering, finding a meaningful cluster structure, is different from that of classification, finding a classifier that allows prediction of labels for unseen data. Therefore, applying a semisupervised clustering method to classification does not necessarily give an appropriate solution. For example, most of the semisupervised clustering methods rely on geometrical or margin-based assumptions such as the cluster assumption and manifold assumption (Basu, Davidson, & Wagstaff, 2008), and without such assumptions, semisupervised clustering methods do not work well. In addition, the objective of semisupervised clustering is usually not the minimization of the classification risk, which may lead to suboptimal performance in terms of the classification accuracy.

In contrast, more and more discriminative training approaches have been studied recently. One is contrastive representation learning (Chopra, Hadsell, & LeCun, 2005; Hadsell, Chopra, & LeCun, 2006; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Kiros et al., 2015; Sohn, 2016;

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Logeswaran & Lee, 2018; Peters et al., 2018; Oord, Li, & Vinyals, 2018; Hjelm et al., 2019; Arora, Khandeparkar, Khodak, Plevrakis, & Saunshi, 2019), which tries to obtain good data representations by bringing an anchor data point close to a given similar data point (a positive sample) and far from randomly sampled data points (negative samples). Resulting representations can be used for downstream classification. The other is the meta-classification approach (Hsu, Lv, Schlosser, Odom, & Kira, 2019; Wu et al., 2020), which performs the maximum likelihood estimation of similar and dissimilar data points. The likelihood is modeled with the inner product between two logits, and the individual logit models are expected to perform well on classification of single data points. While both approaches incorporate similar and dissimilar data points into their formulations, it is not clear whether these methods perform good classification performance from the theoretical perspective; indeed, their objective functions have not been directly connected to the classification risk.

In this letter, we propose a similar-dissimilar-unlabeled (SDU) classification method, where all of pairwise similarities and dissimilarities and unlabeled data serve for unbiased estimation of the classification risk. Like the SU classification method (Bao et al., 2018), our method does not require any geometrical assumptions on the data distribution and enables us to minimize the classification risk via empirical risk minimization. As preparation for constructing our SDU classification, we first develop a dissimilar-unlabeled (DU) classification method and a similar-dissimilar (SD) classification method, where only dissimilar and unlabeled data or similar and dissimilar data are required, respectively. We also show that these methods can be regarded as a special case of a very general framework of classification from unlabeled data (Lu, Niu, Menon, & Sugiyama, 2019). Then we combine the three risks in SU, DU, and SD classification, in a similar manner to positive-negative-unlabeled classification (Sakai, du Plessis, Niu, & Sugiyama, 2017) and train a classifier based on empirical risk minimization at last. We further propose a strategy to reduce the computation cost of hyperparameter tuning by ignoring the SU risk and combine the SD and DU risks for estimation of the classification risk. This strategy comes from the analysis of estimation error bounds for each of SU, DU, and SD classification methods; the bounds for DU/SD classification methods tend to be tighter than the bound of SD classification method. Finally, we experimentally demonstrate the practical usefulness of our method.

Our contributions can be summarized as follows:

- We develop DU and SD classification methods by extending the SU classification method and propose an SDU classification method as a general form of those methods (see section 3).
- We establish estimation error bounds for each method and confirm that unlabeled data help the estimation of the classification risk (see sections 4.1 and 4.3).

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

- From theoretical analysis, we find that estimation error bounds for the DU/SD classification methods tend to be tighter than that for the SU classification method and propose a strategy to reduce the computation cost in the SDU classification method (see section 4.2).

2 Preliminary

In this section, we introduce our problem setting and a generation model of pairwise similarities and dissimilarities and unlabeled data. Thereafter, we review the existing SU classification method.

2.1 Problem Setting. Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{+1, -1\}$ be a d-dimensional example space and binary label space, respectively. Suppose that each labeled example $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is generated independently from the joint probability distribution with density $p(\boldsymbol{x}, y)$. For simplicity, let π_+ and π_- be class priors $p(y = +1)$ and $p(y = -1)$, which satisfy the condition $\pi_+ + \pi_- = 1$, and $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$ be class-conditional densities $p(\boldsymbol{x} \mid y = +1)$ and $p(\boldsymbol{x} \mid y = -1)$.

The standard goal of supervised binary classification is to obtain a classifier $f : \mathcal{X} \to \mathbb{R}$, which minimizes the classification risk defined by

$$R(f) := \mathbf{E}_{(X,Y)\sim p(\boldsymbol{x},y)} \left[\ell(f(X), Y) \right], \tag{2.1}$$

where $\mathbf{E}_{(X,Y)\sim p(\boldsymbol{x},y)}[\cdot]$ denotes the expected value with respect to (X, Y) over the joint density $p(\boldsymbol{x}, y)$ and $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a loss function.

2.2 Generation Model of Training Data. We formulate the data generation process of pairwise data and unlabeled data as follows. First, two examples $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}', y')$ are drawn from $p(\boldsymbol{x}, y)$ independently, $p(\boldsymbol{x}, \boldsymbol{x}', y, y') = p(\boldsymbol{x}, y)p(\boldsymbol{x}', y')$, which also implies that $p(y, y') = p(y)p(y')$. After that, the pairwise information $\tau \in \{+1, -1\}$ is associated with $(\boldsymbol{x}, \boldsymbol{x}')$, where $\tau = +1$ if $y = y'$ and $\tau = -1$ if $y \neq y'$. We represent a pairwise similarity and dissimilarity by the triplet $(\boldsymbol{x}, \boldsymbol{x}', \tau)$. In addition, we suppose that each pairwise example is generated independently. Under these assumptions, we can describe the generation model for n_{SD} pairwise training data as

$$\mathcal{D}_{\mathrm{SD}} := \{(\boldsymbol{x}_{\mathrm{SD},i}, \boldsymbol{x}'_{\mathrm{SD},i}, \tau_i)\}_{i=1}^{n_{\mathrm{SD}}} \sim p(\boldsymbol{x}, \boldsymbol{x}', \tau), \tag{2.2}$$

where

$$p(\boldsymbol{x}, \boldsymbol{x}', \tau = +1) = p(\tau = +1)p(\boldsymbol{x}, \boldsymbol{x}' \mid \tau = +1) = \pi_{\mathrm{S}} p_{\mathrm{S}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.3}$$

$$p(\boldsymbol{x}, \boldsymbol{x}', \tau = -1) = p(\tau = -1)p(\boldsymbol{x}, \boldsymbol{x}' \mid \tau = -1) = \pi_{\mathrm{D}} p_{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}'), \tag{2.4}$$

$$\pi_{\mathrm{S}} := p(\tau = +1) = p(y=+1)p(y'=+1) + p(y=-1)p(y'=-1) = \pi_+^2 + \pi_-^2, \tag{2.5}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

$$\pi_{\mathrm{D}} := p(\tau = -1) = p(y = +1)p(y' = -1) + p(y = -1)p(y' = +1) = 2\pi_+\pi_-, \tag{2.6}$$

$$\begin{aligned} p_{\mathrm{S}}(\boldsymbol{x}, \boldsymbol{x}') &:= p(\boldsymbol{x}, \boldsymbol{x}' \mid \tau = +1) \\ &= \frac{1}{\pi_{\mathrm{S}}} \left\{ p(\boldsymbol{x}, \boldsymbol{x}', y = +1, y' = +1) + p(\boldsymbol{x}, \boldsymbol{x}', y = -1, y' = -1)) \right\} \\ &= \frac{1}{\pi_{\mathrm{S}}} \left\{ p(\boldsymbol{x}, y = +1)p(\boldsymbol{x}', y' = +1) + p(\boldsymbol{x}, y = -1)p(\boldsymbol{x}', y' = -1) \right\} \\ &= \frac{\pi_+^2}{\pi_{\mathrm{S}}} p_+(\boldsymbol{x})p_+(\boldsymbol{x}') + \frac{\pi_-^2}{\pi_{\mathrm{S}}} p_-(\boldsymbol{x})p_-(\boldsymbol{x}'), \end{aligned} \tag{2.7}$$

$$\begin{aligned} p_{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}') &:= p(\boldsymbol{x}, \boldsymbol{x}' \mid \tau = -1) \\ &= \frac{1}{\pi_{\mathrm{D}}} \left\{ p(\boldsymbol{x}, \boldsymbol{x}', y = +1, y' = -1) + p(\boldsymbol{x}, \boldsymbol{x}', y = -1, y' = +1) \right\} \\ &= \frac{1}{\pi_{\mathrm{D}}} \left\{ p(\boldsymbol{x}, y = +1)p(\boldsymbol{x}', y' = -1) + p(\boldsymbol{x}, y = -1)p(\boldsymbol{x}', y' = +1) \right\} \\ &= \frac{1}{2} p_+(\boldsymbol{x})p_-(\boldsymbol{x}') + \frac{1}{2} p_-(\boldsymbol{x})p_+(\boldsymbol{x}'). \end{aligned} \tag{2.8}$$

Similarly, we assume that n_{U} unlabeled examples are drawn from the marginal distribution of $\boldsymbol{x}$ independently:

$$\mathfrak{D}_{\mathrm{U}} := \{\boldsymbol{x}_{\mathrm{U},i}\}_{i=1}^{n_{\mathrm{U}}} \sim p_{\mathrm{U}}(\boldsymbol{x}), \tag{2.9}$$

$$\text{where} \quad p_{\mathrm{U}}(\boldsymbol{x}) := \pi_+ p_+(\boldsymbol{x}) + \pi_- p_-(\boldsymbol{x}). \tag{2.10}$$

On the basis of pairwise information τ, we can divide n_{SD} pairs in $\mathfrak{D}_{\mathrm{SD}}$ into n_{S} similar pairs and n_{D} dissimilar pairs, where $n_{\mathrm{SD}} = n_{\mathrm{S}} + n_{\mathrm{D}}$:

$$\mathfrak{D}_{\mathrm{S}} := \{(\boldsymbol{x}_{\mathrm{S},i}, \boldsymbol{x}'_{\mathrm{S},i})\}_{i=1}^{n_{\mathrm{S}}} = \left\{ (\boldsymbol{x}, \boldsymbol{x}') \mid (\boldsymbol{x}, \boldsymbol{x}', \tau = +1) \in \mathfrak{D}_{\mathrm{SD}} \right\}, \tag{2.11}$$

$$\mathfrak{D}_{\mathrm{D}} := \{(\boldsymbol{x}_{\mathrm{D},i}, \boldsymbol{x}'_{\mathrm{D},i})\}_{i=1}^{n_{\mathrm{D}}} = \left\{ (\boldsymbol{x}, \boldsymbol{x}') \mid (\boldsymbol{x}, \boldsymbol{x}', \tau = -1) \in \mathfrak{D}_{\mathrm{SD}} \right\}. \tag{2.12}$$

With this notation, we can treat pairwise similarities and dissimilarities as if they were drawn from the above conditional distributions: $\mathfrak{D}_{\mathrm{S}} \sim p_{\mathrm{S}}(\boldsymbol{x}, \boldsymbol{x}')$ and $\mathfrak{D}_{\mathrm{D}} \sim p_{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}')$.

2.3 SU Classification. In the seminal paper by Bao et al. (2018), the first method of SU classification was proposed, where the classification risk is estimated in an unbiased fashion from only similar pairs and unlabeled data as follows:

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Proposition 1 *(Theorem 1 in Bao et al., 2018). Suppose $\pi_+ \neq \frac{1}{2}$. The classification risk in equation 2.1 can be equivalently represented as*

$$R_{\mathrm{SU}}(f) = \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X)) + \widetilde{\mathcal{L}}(f(X'))}{2}\right] + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X), -1)\right], \tag{2.13}$$

where

$$\mathcal{L}(z,t) := \frac{\pi_+}{\pi_+ - \pi_-}\ell(z,t) - \frac{\pi_-}{\pi_+ - \pi_-}\ell(z,-t), \tag{2.14}$$

$$\widetilde{\mathcal{L}}(z) := \mathcal{L}(z,+1) - \mathcal{L}(z,-1). \tag{2.15}$$

We can train a classifier by minimizing $\widehat{R}_{\mathrm{SU}}$, that is, the empirical approximation of R_{SU}, computed from $(\mathfrak{D}_{\mathrm{S}}, \mathfrak{D}_{\mathrm{U}})$:

$$\widehat{R}_{\mathrm{SU}}(f) := \frac{\pi_{\mathrm{S}}}{n_{\mathrm{S}}}\sum_{i=1}^{n_{\mathrm{S}}}\frac{\widetilde{\mathcal{L}}(f(\boldsymbol{x}_{\mathrm{S},i})) + \widetilde{\mathcal{L}}(f(\boldsymbol{x}'_{\mathrm{S},i}))}{2} + \frac{1}{n_{\mathrm{U}}}\sum\mathcal{L}(f(\boldsymbol{x}_{\mathrm{U},i}), -1). \tag{2.16}$$

Note that the positive class proportion π_+ is needed to compute $\widehat{R}_{\mathrm{SU}}$. Bao et al. (2018) proposed its estimation procedure as well. Although pairwise similarities and unlabeled data are sufficient to solve a binary classification problem, we incorporate pairwise dissimilarities into their framework to further improve the classification performance of a classifier.

3 Proposed Method

In this section, we propose an SDU classification method, where the classification risk is estimated from pairwise similarities and dissimilarities and unlabeled data. As the first step to construct our method, we extend the SU classification method to DU and SD classification methods.

3.1 DU and SD Classification.

As well as the SU classification method, the classification risk can be estimated from only dissimilar pairs and unlabeled data (DU), or similar pairs and dissimilar pairs (SD) as follows.

Theorem 1. *Suppose $\pi_+ \neq \frac{1}{2}$. The classification risk in equation 2.1 can be equivalently represented as*

$$R_{\mathrm{DU}}(f) = \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}\left[-\frac{\widetilde{\mathcal{L}}(f(X)) + \widetilde{\mathcal{L}}(f(X'))}{2}\right] + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X), +1)\right], \tag{3.1}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

$$R_{\mathrm{SD}}(f) = \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(x,x')}\left[\frac{\mathcal{L}(f(X),+1)+\mathcal{L}(f(X'),+1)}{2}\right] + \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(x,x')}\left[\frac{\mathcal{L}(f(X),-1)+\mathcal{L}(f(X'),-1)}{2}\right], \quad (3.2)$$

where $\mathcal{L}(z,t)$ and $\widetilde{\mathcal{L}}(z)$ are defined in equations 2.14 and 2.15, respectively.

These alternative forms of the classification risk give us empirical risk minimization methods with the empirical risks $\widehat{R}_{\mathrm{DU}}$ and $\widehat{R}_{\mathrm{SD}}$ defined as

$$\widehat{R}_{\mathrm{DU}}(f) = \frac{\pi_{\mathrm{D}}}{n_{\mathrm{D}}}\sum_{i=1}^{n_{\mathrm{D}}} -\frac{\widetilde{\mathcal{L}}(f(x_{\mathrm{D},i}))+\widetilde{\mathcal{L}}(f(x'_{\mathrm{D},i}))}{2} + \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\mathcal{L}(f(x_{\mathrm{U},i}),+1), \quad (3.3)$$

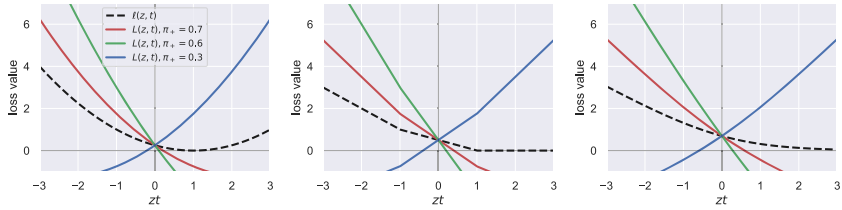
$$\widehat{R}_{\mathrm{SD}}(f) = \frac{\pi_{\mathrm{S}}}{n_{\mathrm{S}}}\sum_{i=1}^{n_{\mathrm{S}}}\frac{\mathcal{L}(f(x_{\mathrm{S},i}),+1)+\mathcal{L}(f(x'_{\mathrm{S},i}),+1)}{2} + \frac{\pi_{\mathrm{D}}}{n_{\mathrm{D}}}\sum_{i=1}^{n_{\mathrm{D}}}\frac{\mathcal{L}(f(x_{\mathrm{D},i}),-1)+\mathcal{L}(f(x'_{\mathrm{D},i}),-1)}{2}. \quad (3.4)$$

Note that $\widehat{R}_{\mathrm{DU}}$ can be computed from $(\mathfrak{D}_{\mathrm{D}}, \mathfrak{D}_{\mathrm{U}})$, and $\widehat{R}_{\mathrm{SD}}$ can be computed from $(\mathfrak{D}_{\mathrm{S}}, \mathfrak{D}_{\mathrm{D}})$, respectively. We call the training method with $\widehat{R}_{\mathrm{DU}}$ a DU classification method and that with $\widehat{R}_{\mathrm{SD}}$ a SD classification method.

3.1.1 Interpretation of SD Risk. The SD risk is not only an equivalent expression of the classification risk, but can also be interpreted as a binary classification risk that aims to predict "similar" and "dissimilar" as labels. We can rewrite the SD risk as

$$\begin{aligned}
R_{\mathrm{SD}}(f) &= \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(x,x')}\left[\frac{\mathcal{L}(f(X),+1)+\mathcal{L}(f(X'),+1)}{2}\right] \\
&\quad + \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(x,x')}\left[\frac{\mathcal{L}(f(X),-1)+\mathcal{L}(f(X'),-1)}{2}\right] \\
&= \iint p(x,x',\tau=+1)\left\{\frac{\mathcal{L}(f(x),+1)+\mathcal{L}(f(x'),+1)}{2}\right\}dxdx' \\
&\quad + \iint p(x,x',\tau=-1)\left\{\frac{\mathcal{L}(f(x),-1)+\mathcal{L}(f(x'),-1)}{2}\right\}dxdx' \\
&= \iint \sum_{\tau\in\{+1,-1\}} p(x,x',\tau)\left\{\frac{\mathcal{L}(f(x),\tau)+\mathcal{L}(f(x'),\tau)}{2}\right\}dxdx' \\
&= \mathbf{E}_{(X,X',T)\sim p(x,x',\tau)}\left[\frac{\mathcal{L}(f(X),T)+\mathcal{L}(f(X'),T)}{2}\right].
\end{aligned}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021



(a) squared loss (b) double hinge loss (c) logistic loss

Figure 1: Visualization of loss $\mathcal{L}$ defined in equation 2.14. ℓ is set to squared, double hinge, and logistic loss. The details of these loss functions are described in section 3.3 (see Table 1). As shown in the graphs, $\mathcal{L}(z, t)$ approaches $\ell(z, t)$ as π_+ gets larger, and $\mathcal{L}(z, t)$ approaches $\ell(z, -t)$ as π_+ gets smaller.

To interpret this expression of the risk from a different perspective, we visualize the landscape of loss $\mathcal{L}$ in Figure 1, with ℓ set to several standard margin-based losses. As can be seen in the figure, $\mathcal{L}(z, t)$ has a similar profile to $\ell(z, t)$ when $\pi_+ > \frac{1}{2}$. Otherwise, $\mathcal{L}(z, t)$ is similar to $\ell(z, -t)$. This enables us to give another interpretation of SD classification that it is a binary classification with loss function $\mathcal{L}$, where a classifier f takes input X and predicts its associated pairwise label T. From this point of view, the relationship among SD, SU, and DU classification corresponds to that among positive-negative (PN), positive-unlabeled (PU), and negative-unlabeled (NU) classification (du Plessis, Niu, & Sugiyama, 2014, 2015). The main idea of the PU (resp. NU) classification method is to complement missing negative (resp. positive) information with unlabeled data. For example, the classification risk can be represented by positive and unlabeled data as follows:

$$\begin{aligned} R(f) &= \mathbf{E}_{(X,Y)\sim p(\boldsymbol{x},y)}\left[\ell(f(X), Y)\right] \\ &= \pi_+ \mathbf{E}_{X\sim p_+(\boldsymbol{x})}\left[\ell(f(X), +1)\right] + \pi_- \mathbf{E}_{X\sim p_-(\boldsymbol{x})}\left[\ell(f(X), -1)\right] \\ &= \pi_+ \mathbf{E}_{X\sim p_+(\boldsymbol{x})}\left[\ell(f(X), +1)\right] \\ &\quad + \underbrace{\left\{\mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\ell(f(X), -1)\right] - \pi_+ \mathbf{E}_{X\sim p_+(\boldsymbol{x})}\left[\ell(f(X), -1)\right]\right\}}_{\text{since } \pi_- p(\boldsymbol{x}) = p_{\mathrm{U}}(\boldsymbol{x}) - \pi_+ p_+(\boldsymbol{x})} \\ &= \pi_+ \mathbf{E}_{X\sim p_+(\boldsymbol{x})}\left[\ell(f(X), +1) - \ell(f(X), -1)\right] \\ &\quad + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\ell(f(X), -1)\right]. \end{aligned} \tag{3.5}$$

This derivation is the same as that of theorem 1.

Furthermore, when loss ℓ is symmetric (Ghosh, Manwani, & Sastry, 2015; Charoenphakdee, Lee, & Sugiyama, 2019), that is, for some $K \in \mathbb{R}$, $\ell(z, +1) + \ell(z, -1) = K$, then the following relationship holds:

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Corollary 1. *Assume $\pi_+ \neq \frac{1}{2}$. We define Q_{SD} by replacing $\mathcal{L}$ in R_{SD} with ℓ as follows:*

$$Q_{\mathrm{SD}}(f) := \mathbf{E}_{(X,X',T)\sim p(\boldsymbol{x},\boldsymbol{x}',\tau)}\left[\frac{\ell(f(X),T)+\ell(f(X'),T)}{2}\right]. \tag{3.6}$$

Suppose that ℓ is a symmetric loss. Then R_{SD} and Q_{SD} share the optimal solution as

$$\operatorname*{arg\,min}_{f\in\mathcal{F}} R_{\mathrm{SD}}(f) = \begin{cases} \operatorname*{arg\,min}_{f\in\mathcal{F}} Q_{\mathrm{SD}}(f) & \left(\pi_+ > \frac{1}{2}\right), \\ \operatorname*{arg\,max}_{f\in\mathcal{F}} Q_{\mathrm{SD}}(f) & \left(\pi_+ < \frac{1}{2}\right). \end{cases} \tag{3.7}$$

Proof. We show that $\mathcal{L}(z,t)$ is a linear function of $\ell(z,t)$:

$$\begin{aligned} \mathcal{L}(z,t) &= \frac{\pi_+}{\pi_+-\pi_-}\ell(z,t) - \frac{\pi_-}{\pi_+-\pi_-}\ell(z,-t) \\ &= \frac{\pi_+}{\pi_+-\pi_-}\ell(z,t) - \frac{\pi_-}{\pi_+-\pi_-}\left\{K-\ell(z,t)\right\} \\ &= \frac{1}{\pi_+-\pi_-}\ell(z,t) - \frac{\pi_-}{\pi_+-\pi_-}K. \end{aligned}$$

By using the above relationship, we obtain

$$R_{\mathrm{SD}}(f) = \frac{1}{\pi_+-\pi_-}Q_{\mathrm{SD}}(f) - \frac{\pi_-}{\pi_+-\pi_-}K,$$

which results in equation 3.7. □

If we use a symmetric loss in the classification risk, corollary 1 gives us practical advantages. For instance, when writing program code for a training algorithm, we do not have to implement loss $\mathcal{L}$ by ourselves. Instead, we can treat "similar" and "dissimilar" labels as positive and negative labels associated with each point in a pair and use any standard binary classification algorithm with a loss function ℓ.

Actually, the relationship in corollary 1 holds for R_{SU} and R_{DU} as well. The alternative objective functions Q_{SU} and Q_{DU} are defined as

$$\begin{aligned} Q_{\mathrm{SU}}(f) := {} & \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\tilde{\ell}(f(X))+\tilde{\ell}(f(X'))}{2}\right] \\ & + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\ell(f(X),-1)\right], \end{aligned} \tag{3.8}$$

$$\begin{aligned} Q_{\mathrm{DU}}(f) := {} & \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}\left[-\frac{\tilde{\ell}(f(X))+\tilde{\ell}(f(X'))}{2}\right] \\ & + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\ell(f(X),+1)\right], \end{aligned} \tag{3.9}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

where

$$\tilde{\ell}(z) := \ell(z, +1) - \ell(z, -1). \tag{3.10}$$

Similar to equation 3.7, we observe that Q_{SU} and Q_{DU} have the same optimizers as R_{SU} and R_{DU}, respectively. Moreover, we can confirm that Q_{SU} corresponds to the PU risk in equation 3.5 (Q_{DU} corresponds to the NU risk as well), which gives us an intuitive interpretation of the SU (resp. DU) classification method as the PU (resp. NU) classification method.

3.1.2 Relation to UU Classification. Each of SU, DU, and SD classification is regarded as a special case of unlabeled-unlabeled (UU) classification (Lu et al., 2019), a very general framework in weakly supervised learning and enables us to train a classifier without any labeled data. In UU classification, we assume that two unlabeled training sets $\mathfrak{D}_{\mathrm{tr}}$ and $\mathfrak{D}'_{\mathrm{tr}}$ are available, which are drawn from two distinct marginal densities p_{tr} and p'_{tr}, respectively:

$$\begin{aligned}
\mathfrak{D}_{\mathrm{tr}} &:= \{\boldsymbol{x}_{\mathrm{tr},i}\}_{i=1}^{n_{\mathrm{tr}}} \sim p_{\mathrm{tr}}(\boldsymbol{x}),\\
\mathfrak{D}'_{\mathrm{tr}} &:= \{\boldsymbol{x}'_{\mathrm{tr},i}\}_{i=1}^{n'_{\mathrm{tr}}} \sim p'_{\mathrm{tr}}(\boldsymbol{x}),\\
p_{\mathrm{tr}}(\boldsymbol{x}) &:= \theta p_+(\boldsymbol{x}) + (1-\theta)p_-(\boldsymbol{x}),\\
p'_{\mathrm{tr}}(\boldsymbol{x}) &:= \theta' p_+(\boldsymbol{x}) + (1-\theta')p_-(\boldsymbol{x}),
\end{aligned}$$

where θ and θ' are some constants satisfying $\theta, \theta' \in [0, 1]$ and $\theta \neq \theta'$. Then we can rewrite the classification risk with these densities as follows.

Proposition 2 *(Theorem 4 in Lu et al., 2019). Assume that $\theta > \theta'$; otherwise, swap p_{tr} and p'_{tr} to make sure $\theta > \theta'$. Then, the classification risk in equation 2.1 can be equivalently represented as*

$$\begin{aligned}
&\mathbf{E}_{X\sim p_{\mathrm{tr}}(x)}\left[a\ell(f(X), +1) + b\ell(f(X), -1)\right]\\
&+ \mathbf{E}_{X\sim p'_{\mathrm{tr}}(x)}\left[c\ell(f(X), -1) + d\ell(f(X), +1)\right],\\
&\quad \textit{where} \quad a := \frac{(1-\theta')\pi_+}{\theta-\theta'}, \quad b := -\frac{\theta'\pi_-}{\theta-\theta'}, \quad c := \frac{\theta\pi_-}{\theta-\theta'},\\
&\qquad\qquad d := -\frac{(1-\theta)\pi_+}{\theta-\theta'}.
\end{aligned} \tag{3.11}$$

The risk expression in equation 3.11 enables us to train a classifier by minimizing the empirical risk computed from $(\mathfrak{D}_{\mathrm{tr}}, \mathfrak{D}'_{\mathrm{tr}})$. Now, we review the relationship between UU classification and SU, DU, SD classification. Since we assume that each example in a pair is drawn independently (see section 2.2), we can reduce the pairwise distributions into the pointwise

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

distributions as

$$\tilde{p}_{\mathrm{S}}(\boldsymbol{x}) := \int p_{\mathrm{S}}(\boldsymbol{x}, \boldsymbol{x}')d\boldsymbol{x}' = \frac{\pi_+^2}{\pi_{\mathrm{S}}} p_+(\boldsymbol{x}) + \frac{\pi_-^2}{\pi_{\mathrm{S}}} p_-(\boldsymbol{x}), \tag{3.12}$$

$$\tilde{p}_{\mathrm{D}}(\boldsymbol{x}) := \int p_{\mathrm{D}}(\boldsymbol{x}, \boldsymbol{x}')d\boldsymbol{x}' = \frac{1}{2} p_+(\boldsymbol{x}) + \frac{1}{2} p_-(\boldsymbol{x}). \tag{3.13}$$

With this notation, each single point in $\mathfrak{D}_{\mathrm{S}}$ and $\mathfrak{D}_{\mathrm{D}}$ can be treated as if it was drawn from $\tilde{p}_{\mathrm{S}}(\boldsymbol{x})$ and $\tilde{p}_{\mathrm{D}}(\boldsymbol{x})$, respectively. Therefore, SU, DU, and SD classification correspond to special cases in UU classification:

$$(\theta, \theta') = \begin{cases} \left(\frac{\pi_+^2}{\pi_{\mathrm{S}}}, \pi_+\right) & \text{(SU classification)}, \\ \left(\frac{1}{2}, \pi_+\right) & \text{(DU classification)}, \\ \left(\frac{\pi_+^2}{\pi_{\mathrm{S}}}, \frac{1}{2}\right) & \text{(SD classification)}. \end{cases} \tag{3.14}$$

Note that the condition $\theta \neq \theta'$ in UU classification corresponds to $\pi_+ \neq \frac{1}{2}$ in SU, DU, and SD classification. If such a condition is not satisfied, none of them can be solved because unbiased risk estimators degenerate.

3.2 SDU Classification. Here, we propose an SDU classification method that incorporates all of pairwise similarities, pairwise dissimilarities, and unlabeled data into the empirical risk minimization framework. Our main idea is to combine the risks computed from SU, DU, and SD data, in a similar manner to positive-negative-unlabeled classification (Sakai et al., 2017) that is an unbiased risk estimation approach to semisupervised classification. Since each of $R_{\mathrm{SU}}(f)$, $R_{\mathrm{DU}}(f)$, and $R_{\mathrm{SD}}(f)$ is an equivalent expression of the true classification risk, the following convex combination of those risks is still equivalent to $R(f)$:

$$R^{\boldsymbol{\gamma}}_{\mathrm{SDU}}(f) := \gamma_1 R_{\mathrm{SU}}(f) + \gamma_2 R_{\mathrm{DU}}(f) + \gamma_3 R_{\mathrm{SD}}(f), \tag{3.15}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ is the hyperparameter that satisfies $\gamma_1, \gamma_2, \gamma_3 \geq 0$, and $\gamma_1 + \gamma_2 + \gamma_3 = 1$. In section 4.2, from the theoretical analysis, we propose a strategy to reduce the tuning cost of $\boldsymbol{\gamma}$ by fixing $\gamma_1 = 0$.

3.3 Practical Implementation. We investigate the objective function with a linear classifier $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) + b$, where $\boldsymbol{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$ are weights and $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^k$ is a mapping function. Then the empirical risk minimization with L_2 regularization can be described by

$$\min_{\boldsymbol{w}} \widehat{R}^{\boldsymbol{\gamma}}_{\mathrm{SDU}}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2, \tag{3.16}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Table 1: Margin-Based Loss Functions That Satisfy the Conditions in Theorem 2.

Loss Name	$\psi(tz)$
Squared loss	$\frac{1}{4}(tz-1)^2$
Logistic loss	$\log(1+\exp(-tz))$
Double hinge loss	$\max(-m, \max(0, \frac{1}{2}-\frac{1}{2}tz))$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

where $\widehat{R}^{\gamma}_{\mathrm{SDU}}$ is an empirical estimator of $R^{\gamma}_{\mathrm{SDU}}$ and $\lambda > 0$ is a parameter of L_2 regularization. In the rest of this letter, we suppose that the loss function ℓ is a margin-based loss function. As defined in Mohri, Rostamizadeh, Bach, and Talwalkar (2012), we call ℓ a margin-based loss function if there exists $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that $\ell(z,t) = \psi(tz)$. In general, the optimization problem in equation 3.16 is nonconvex even if ℓ is a convex margin-based loss. However, if we choose ℓ that satisfies the following property, the optimization problem becomes convex.

Theorem 2. *Suppose that the loss function $\ell(z,t)$ is a convex margin-based loss, twice differentiable in z almost everywhere (for every fixed $t \in \{\pm 1\}$), and satisfies the following condition:*

$$\ell(z,+1) - \ell(z,-1) = -z. \tag{3.17}$$

Then the optimization problem in equation 3.16 is convex.

Several loss functions that satisfy the conditions in theorem 2 are shown in Table 1, borrowed from Patrini, Nielsen, Nock, and Carioni (2016) and Bao, Niu, & Sugiyama (2018). Next, we consider the optimization problem with the squared loss and the double hinge loss, respectively.

3.3.1 Squared Loss. We consider the optimization problem in equation 3.16 with the squared loss defined by

$$\ell_{\mathrm{SQ}}(z,t) = \frac{1}{4}(tz-1)^2. \tag{3.18}$$

For convenience, we denote pointwise samples in $\mathfrak{D}_{\mathrm{S}}$ and $\mathfrak{D}_{\mathrm{D}}$ as

$$\widetilde{\mathfrak{D}}_{\mathrm{S}} := \{\widetilde{\boldsymbol{x}}_{\mathrm{S},i}\}_{i=1}^{2n_{\mathrm{S}}} = \bigcup \{\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}'_{\mathrm{S}} \mid (\boldsymbol{x}_{\mathrm{S}}, \boldsymbol{x}'_{\mathrm{S}}) \in \mathfrak{D}_{\mathrm{S}}\}, \tag{3.19}$$

$$\widetilde{\mathfrak{D}}_{\mathrm{D}} := \{\widetilde{\boldsymbol{x}}_{\mathrm{D},i}\}_{i=1}^{2n_{\mathrm{D}}} = \bigcup \{\boldsymbol{x}_{\mathrm{D}}, \boldsymbol{x}'_{\mathrm{D}} \mid (\boldsymbol{x}_{\mathrm{D}}, \boldsymbol{x}'_{\mathrm{D}}) \in \mathfrak{D}_{\mathrm{D}}\}. \tag{3.20}$$

Then the objective function, $\widehat{R}^{\boldsymbol{\gamma}}_{\mathrm{SDU}}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$, can be written as

$$\begin{aligned}
&\frac{1}{4}\boldsymbol{w}^\top \left\{ \gamma_3 \left(\frac{\pi_{\mathrm{S}}}{2n_{\mathrm{S}}} X_{\mathrm{S}}^\top X_{\mathrm{S}} + \frac{\pi_{\mathrm{D}}}{2n_{\mathrm{D}}} X_{\mathrm{D}}^\top X_{\mathrm{D}} \right) + \frac{\gamma_1+\gamma_2}{n_{\mathrm{U}}} X_{\mathrm{U}}^\top X_{\mathrm{U}} + 2\lambda I \right\} \boldsymbol{w} \\
&+ \frac{1}{\pi_+ - \pi_-} \left\{ -\frac{\pi_{\mathrm{S}}}{2n_{\mathrm{S}}} \left(\gamma_1 + \frac{\gamma_3}{2} \right) X_{\mathrm{S}}^\top \mathbf{1} + \frac{\pi_{\mathrm{D}}}{2n_{\mathrm{D}}} \left(\gamma_2 + \frac{\gamma_3}{2} \right) X_{\mathrm{D}}^\top \mathbf{1} \right. \\
&\qquad \left. + \frac{1}{2n_{\mathrm{U}}} (\gamma_1 - \gamma_2) X_{\mathrm{U}}^\top \mathbf{1} \right\} \boldsymbol{w} + \text{const.}, \qquad (3.21)
\end{aligned}$$

where

$$\begin{aligned}
X_{\mathrm{S}} &:= [\boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{S},1}), \ldots, \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{S},2n_{\mathrm{S}}})]^\top, \\
X_{\mathrm{D}} &:= [\boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{D},1}), \ldots, \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{D},2n_{\mathrm{D}}})]^\top, \\
X_{\mathrm{U}} &:= [\boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},1}), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},n_{\mathrm{U}}})]^\top.
\end{aligned}$$

We denote $\mathbf{1}$ as the vector whose elements are all ones and I as the identity matrix. Since this function has a nondegenerate quadratic form with respect to $\boldsymbol{w}$, the solution of this minimization problem can be obtained analytically as

$$\begin{aligned}
\widehat{\boldsymbol{w}} = &\frac{1}{\pi_+ - \pi_-} \left\{ \gamma_3 \left(\frac{\pi_{\mathrm{S}}}{2n_{\mathrm{S}}} X_{\mathrm{S}}^\top X_{\mathrm{S}} + \frac{\pi_{\mathrm{D}}}{2n_{\mathrm{D}}} X_{\mathrm{D}}^\top X_{\mathrm{D}} \right) + \frac{\gamma_1+\gamma_2}{n_{\mathrm{U}}} X_{\mathrm{U}}^\top X_{\mathrm{U}} + 2\lambda I \right\}^{-1} \\
&\times \left\{ \frac{\pi_{\mathrm{S}}}{n_{\mathrm{S}}} \left(\gamma_1 + \frac{\gamma_3}{2} \right) X_{\mathrm{S}}^\top \mathbf{1} + \frac{\pi_{\mathrm{D}}}{n_{\mathrm{D}}} \left(\gamma_2 + \frac{\gamma_3}{2} \right) X_{\mathrm{D}}^\top \mathbf{1} + \frac{1}{n_{\mathrm{U}}} (\gamma_1 - \gamma_2) X_{\mathrm{U}}^\top \mathbf{1} \right\}. \qquad (3.22)
\end{aligned}$$

3.3.2 Double Hinge Loss. Standard hinge loss $\ell_{\mathrm{H}}(z,t) = \max(0, 1-tz)$ does not satisfy the condition in equation 3.17. As an alternative, the double hinge loss $\ell_{\mathrm{DH}}(z,t) = \max(-tz, \max(0, \frac{1}{2} - \frac{1}{2}tz)$ was proposed by du Plessis et al. (2015). The optimization problem in equation 3.16 with the double hinge loss can be solved by quadratic programming. The objective function, $\widehat{R}^{\boldsymbol{\gamma}}_{\mathrm{SDU}}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$, can be represented by

$$\begin{aligned}
&-\frac{\gamma_1 \pi_{\mathrm{S}}}{2n_{\mathrm{S}}(\pi_+ - \pi_-)} \sum_{i=1}^{2n_{\mathrm{S}}} \boldsymbol{w}^\top \phi(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}) + \frac{\gamma_2 \pi_{\mathrm{D}}}{2n_{\mathrm{D}}(\pi_+ - \pi_-)} \sum_{i=1}^{2n_{\mathrm{D}}} \boldsymbol{w}^\top \phi(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}) \\
&+\frac{\gamma_3 \pi_{\mathrm{S}}}{2n_{\mathrm{S}}(\pi_+ - \pi_-)} \sum_{i=1}^{2n_{\mathrm{S}}} \left(\pi_+ \ell_{\mathrm{DH}}(\boldsymbol{w}^\top \phi(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}), +1) - \pi_- \ell_{\mathrm{DH}}(\boldsymbol{w}^\top \phi(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}), -1) \right)
\end{aligned}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

$$
\begin{aligned}
&-\frac{\gamma_3\pi_{\mathrm{D}}}{2n_{\mathrm{D}}(\pi_+-\pi_-)}\sum_{i=1}^{2n_{\mathrm{D}}}\left(\pi_-\ell_{\mathrm{DH}}(\boldsymbol{w}^\top\phi(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}),+1)-\pi_+\ell_{\mathrm{DH}}(\boldsymbol{w}^\top\phi(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}),-1)\right)\\
&+\frac{1}{n_{\mathrm{U}}(\pi_+-\pi_-)}\sum_{i=1}^{n_{\mathrm{U}}}\left\{(\gamma_2\pi_+-\gamma_1\pi_-)\ell_{\mathrm{DH}}(\boldsymbol{w}^\top\phi(\boldsymbol{x}_{\mathrm{U},i}),+1)\right.\\
&\qquad\qquad\left.+(\gamma_1\pi_+-\gamma_2\pi_-)\ell_{\mathrm{DH}}(\boldsymbol{w}^\top\phi(\boldsymbol{x}_{\mathrm{U},i}),-1)\right\}\\
&+\frac{\lambda}{2}\boldsymbol{w}^\top\boldsymbol{w}. \qquad (3.23)
\end{aligned}
$$

Using slack variables $\boldsymbol{\xi}=\{\boldsymbol{\xi}_{\mathrm{S}},\boldsymbol{\xi}_{\mathrm{D}},\boldsymbol{\xi}_{\mathrm{U}}\}$ and $\boldsymbol{\eta}=\{\boldsymbol{\eta}_{\mathrm{S}},\boldsymbol{\eta}_{\mathrm{D}},\boldsymbol{\eta}_{\mathrm{U}}\}$, we can rewrite the optimization problem in equation 3.16 as

$$
\begin{aligned}
\min_{\boldsymbol{w},\boldsymbol{\xi},\boldsymbol{\eta}}\ &-\frac{\gamma_1\pi_{\mathrm{S}}}{2n_{\mathrm{S}}(\pi_+-\pi_-)}\mathbf{1}^\top X_{\mathrm{S}}\boldsymbol{w}+\frac{\gamma_2\pi_{\mathrm{D}}}{2n_{\mathrm{D}}(\pi_+-\pi_-)}\mathbf{1}^\top X_{\mathrm{D}}\boldsymbol{w}\\
&+\frac{\gamma_3\pi_+\pi_{\mathrm{S}}}{2n_{\mathrm{S}}(\pi_+-\pi_-)}\mathbf{1}^\top\boldsymbol{\xi}_{\mathrm{S}}-\frac{\gamma_3\pi_-\pi_{\mathrm{S}}}{2n_{\mathrm{S}}(\pi_+-\pi_-)}\mathbf{1}^\top\boldsymbol{\eta}_{\mathrm{S}}\\
&-\frac{\gamma_3\pi_-\pi_{\mathrm{D}}}{2n_{\mathrm{D}}(\pi_+-\pi_-)}\mathbf{1}^\top\boldsymbol{\xi}_{\mathrm{D}}+\frac{\gamma_3\pi_+\pi_{\mathrm{D}}}{2n_{\mathrm{D}}(\pi_+-\pi_-)}\mathbf{1}^\top\boldsymbol{\eta}_{\mathrm{D}}\\
&+\frac{-\gamma_1\pi_-+\gamma_2\pi_+}{n_{\mathrm{U}}(\pi_+-\pi_-)}\mathbf{1}^\top\boldsymbol{\xi}_{\mathrm{U}}+\frac{\gamma_1\pi_+-\gamma_2\pi_-}{n_{\mathrm{U}}(\pi_+-\pi_-)}\mathbf{1}^\top\boldsymbol{\eta}_{\mathrm{U}}+\frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}\\
\text{s.t.}\quad &\boldsymbol{\xi}_{\mathrm{S}}\geq 0,\boldsymbol{\xi}_{\mathrm{S}}\geq\frac{1}{2}-\frac{1}{2}X_{\mathrm{S}}\boldsymbol{w},\boldsymbol{\xi}_{\mathrm{S}}\geq -X_{\mathrm{S}}\boldsymbol{w},\\
&\boldsymbol{\eta}_{\mathrm{S}}\geq 0,\boldsymbol{\eta}_{\mathrm{S}}\geq\frac{1}{2}+\frac{1}{2}X_{\mathrm{S}}\boldsymbol{w},\boldsymbol{\eta}_{\mathrm{S}}\geq X_{\mathrm{S}}\boldsymbol{w},\\
&\boldsymbol{\xi}_{\mathrm{D}}\geq 0,\boldsymbol{\xi}_{\mathrm{D}}\geq\frac{1}{2}-\frac{1}{2}X_{\mathrm{D}}\boldsymbol{w},\boldsymbol{\xi}_{\mathrm{D}}\geq -X_{\mathrm{D}}\boldsymbol{w},\\
&\boldsymbol{\eta}_{\mathrm{D}}\geq 0,\boldsymbol{\eta}_{\mathrm{D}}\geq\frac{1}{2}+\frac{1}{2}X_{\mathrm{D}}\boldsymbol{w},\boldsymbol{\eta}_{\mathrm{D}}\geq X_{\mathrm{D}}\boldsymbol{w},\\
&\boldsymbol{\xi}_{\mathrm{U}}\geq 0,\boldsymbol{\xi}_{\mathrm{U}}\geq\frac{1}{2}-\frac{1}{2}X_{\mathrm{U}}\boldsymbol{w},\boldsymbol{\xi}_{\mathrm{U}}\geq -X_{\mathrm{U}}\boldsymbol{w},\\
&\boldsymbol{\eta}_{\mathrm{U}}\geq 0,\boldsymbol{\eta}_{\mathrm{U}}\geq\frac{1}{2}+\frac{1}{2}X_{\mathrm{U}}\boldsymbol{w},\boldsymbol{\eta}_{\mathrm{U}}\geq X_{\mathrm{U}}\boldsymbol{w}, \qquad (3.24)
\end{aligned}
$$

where $\geq$ for vectors indicates elementwise inequality.

3.4 Class Prior Estimation from Pairwise Data. Although the exact positive class proportion π_+ has to be known in advance to compute the empirical risk $\widehat{R}_{\mathrm{SDU}}$, it is often unknown in practice. Here, we show that π_+ can be estimated from the number of similar pairs n_{S} and the number of

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

dissimilar pairs n_{D}. The positive ratio in pointwise data π_+ and the similar ratio in pairwise data π_{S} have the following relationship:

$$\pi_+ = \begin{cases} \frac{1+\sqrt{2\pi_{\mathrm{S}}-1}}{2} & (\pi_+ \geq \frac{1}{2}), \\ \frac{1-\sqrt{2\pi_{\mathrm{S}}-1}}{2} & (\text{otherwise}). \end{cases} \tag{3.25}$$

The above equality is derived from $\pi_{\mathrm{S}} = \pi_+^2 + (1-\pi_+)^2$. Since $\widehat{\pi}_{\mathrm{S}} = n_{\mathrm{S}}/(n_{\mathrm{S}} + n_{\mathrm{D}})$ is an unbiased estimator of π_{S}, π_+ can be estimated by plugging $\widehat{\pi}_{\mathrm{S}}$ into equation 3.25.

4 Theoretical Analysis

In this section, we analyze estimation error bounds for our methods. We first derive estimation error bounds for the SU, DU, and SD classification methods via Rademacher complexity. By comparing these bounds, we find a nontrivial relationship in the performances of these methods. It also gives a strategy to reduce the cost of hyperparameter tuning in the SDU classification method. Finally, we derive an estimation error bound for the SDU classification method.

4.1 Estimation Error Bounds for SU, DU, and SD Classification. We investigate estimation error bounds for the SU, DU, and SD classification methods. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a function class of the specified model:

Definition 1 *(Rademacher Complexity). Let n be a positive integer, $Z_1, \ldots, Z_n$ be independent and identically distributed (i.i.d.) random variables drawn from a probability distribution with density μ, $\mathcal{H} = \{h : \mathcal{Z} \to \mathbb{R}\}$be a class of measurable functions, and $\sigma = (\sigma_1, \ldots, \sigma_n)$ be Rademacher variables, that is, random variables taking $+1$ and -1 with even probabilities. Then the (expected) Rademacher complexity of $\mathcal{H}$ is defined as*

$$\mathfrak{R}(\mathcal{H}; n, \mu) := \mathbf{E}_{Z_1,\ldots,Z_n \sim \mu} \mathbf{E}_{\sigma} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(Z_i) \right]. \tag{4.1}$$

For the function class $\mathcal{F}$ and any probability density μ, we assume

$$\mathfrak{R}(\mathcal{F}; n, \mu) \leq \frac{C_{\mathcal{F}}}{\sqrt{n}}. \tag{4.2}$$

This assumption holds for many models, such as linear-in-parameter model class $\mathcal{F} = \left\{ f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) \right\}$ as shown in Mohri et al. (2012). Partially based on Bao et al. (2018), we have estimation error bounds for the SU, DU, and SD classification methods as follows.

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Theorem 3. *Let $R(f) = \mathbf{E}[\ell(f(\boldsymbol{x}), y)]$ be a classification risk for function f, $f^* \in \mathcal{F}$ be its minimizer, and $\widehat{f}_{\mathrm{SU}}$, $\widehat{f}_{\mathrm{DU}}$, $\widehat{f}_{\mathrm{SD}}$ be minimizers of the empirical SU, DU, SD risks in $\mathcal{F}$, respectively. Assume that $\pi_+ \neq \frac{1}{2}$, the loss function ℓ is ρ-Lipschitz function with respect to the first argument ($0 < \rho < \infty$), and all functions in the model class $\mathcal{F}$ are bounded, that is, there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t \in \{\pm 1\}} \ell(C_b, t)$. For any $\delta > 0$, each of the following inequalities holds independently with probability at least $1 - \delta$:*

$$R(\widehat{f}_{\mathrm{SU}}) - R(f^*) \leq C_{\mathcal{F},\ell,\delta} \left(\frac{2\pi_{\mathrm{S}}}{\sqrt{2n_{\mathrm{S}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}} \right), \tag{4.3}$$

$$R(\widehat{f}_{\mathrm{DU}}) - R(f^*) \leq C_{\mathcal{F},\ell,\delta} \left(\frac{2\pi_{\mathrm{D}}}{\sqrt{2n_{\mathrm{D}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}} \right), \tag{4.4}$$

$$R(\widehat{f}_{\mathrm{SD}}) - R(f^*) \leq C_{\mathcal{F},\ell,\delta} \left(\frac{\pi_{\mathrm{S}}}{\sqrt{2n_{\mathrm{S}}}} + \frac{\pi_{\mathrm{D}}}{\sqrt{2n_{\mathrm{D}}}} \right), \tag{4.5}$$

where

$$C_{\mathcal{F},\ell,\delta} = \frac{1}{|\pi_+ - \pi_-|} \left(4\rho C_{\mathcal{F}} + \sqrt{2C_\ell^2 \log \frac{8}{\delta}} \right). \tag{4.6}$$

4.2 Comparison of SU, DU, and SD Bounds. Here, we compare the SU, DU, and SD classification methods from the perspective of their estimation error bounds. Under the generation process of similar and dissimilar pairs in equation 2.2, we have the following claim:

Theorem 4. *Suppose similar and dissimilar pairs follow the generation process in equation 2.2. We denote each right-hand side in equations 4.3 to 4.5 by V_{SD}, V_{SU}, and V_{DU}, respectively. Then, $V_{\mathrm{DU}} \leq V_{\mathrm{SU}}$ and $V_{\mathrm{SD}} \leq V_{\mathrm{SU}}$ hold with the probability at least $1 - \exp(-cn_{\mathrm{SD}})$ for some constant $c > 0$.*

Proof. If $\pi_{\mathrm{S}}/\sqrt{2n_{\mathrm{S}}} > \pi_{\mathrm{D}}/\sqrt{2n_{\mathrm{D}}}$ holds, we have

$$\frac{V_{\mathrm{SU}} - C_{\mathcal{F},\ell,\delta}/\sqrt{n_{\mathrm{U}}}}{V_{\mathrm{DU}} - C_{\mathcal{F},\ell,\delta}/\sqrt{n_{\mathrm{U}}}} = \frac{\pi_{\mathrm{S}}/\sqrt{2n_{\mathrm{S}}}}{\pi_{\mathrm{D}}/\sqrt{2n_{\mathrm{D}}}} > 1$$

and

$$V_{\mathrm{SU}} - V_{\mathrm{SD}} = C_{\mathcal{F},\ell,\delta} \left(\frac{\pi_{\mathrm{S}}}{\sqrt{2n_{\mathrm{S}}}} - \frac{\pi_{\mathrm{D}}}{\sqrt{2n_{\mathrm{D}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}} \right) > \frac{C_{\mathcal{F},\ell,\delta}}{\sqrt{n_{\mathrm{U}}}} > 0.$$

These two inequalities imply $V_{\mathrm{DU}} < V_{\mathrm{SU}}$ and $V_{\mathrm{SD}} < V_{\mathrm{SU}}$. Since we assume the generation process in equation 2.2, the class of each pair (i.e., similar or dissimilar) follows a Bernoulli distribution. Therefore, the number of pairs in each class follows a binomial distribution, namely, $n_{\mathrm{D}} \sim$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Binomial($n_{\mathrm{SD}}, \pi_{\mathrm{D}}$) and $n_{\mathrm{S}} = n_{\mathrm{SD}} - n_{\mathrm{D}}$. By using Chernoff's inequality in Okamoto (1959), we have

$$p\left(\frac{\pi_{\mathrm{S}}}{\sqrt{2n_{\mathrm{S}}}} \leq \frac{\pi_{\mathrm{D}}}{\sqrt{2n_{\mathrm{D}}}}\right) = p\left(n_{\mathrm{D}} \leq \frac{n_{\mathrm{SD}}\pi_{\mathrm{D}}^2}{\pi_{\mathrm{S}}^2 + \pi_{\mathrm{D}}^2}\right)$$
$$\leq \exp\left(-\frac{n_{\mathrm{SD}}\pi_{\mathrm{D}}}{2(1-\pi_{\mathrm{D}})}\left(1 - \frac{\pi_{\mathrm{D}}}{\pi_{\mathrm{S}}^2 + \pi_{\mathrm{D}}^2}\right)^2\right).$$

Finally, we obtain

$$p(V_{\mathrm{DU}} \leq V_{\mathrm{SU}} \wedge V_{\mathrm{SD}} \leq V_{\mathrm{SU}}) \geq 1 - \exp\left(-\frac{n_{\mathrm{SD}}\pi_{\mathrm{D}}}{2(1-\pi_{\mathrm{D}})}\left(1 - \frac{\pi_{\mathrm{D}}}{\pi_{\mathrm{S}}^2 + \pi_{\mathrm{D}}^2}\right)^2\right).$$

□

4.2.1 SDDU Classification for Efficient Hyperparameter Search. Theorem 4 states that $\max\{V_{\mathrm{SD}}, V_{\mathrm{DU}}\} \leq V_{\mathrm{SD}}$ holds with high probability when n_{SD} is sufficiently large. It suggests that both DU and SD classification methods are likely to outperform the SU classification method when all of pairwise similarities and dissimilarities and unlabeled data are given in advance. Inspired by this result, we propose a strategy to reduce the computation cost by fixing $\gamma_1 = 0$ in equation 3.15, that is, the classification risk is always estimated with the DU and SD risks. We call this method the *SDDU classification method* to distinguish it from the general SDU classification method. In sections 5.2 and 5.3, we experimentally demonstrate that the SDDU classification method performs at the same level as or better than the SDU classification method.

4.3 Estimation Error Bound for SDU Classification. We derive an estimation error bound for the SDU classification method. With the same technique as in theorem 3, we have the following bound:

Theorem 5. *Let $R(f) = \mathbf{E}[\ell(f(\boldsymbol{x}), y)]$ be a classification risk for function f, $f^* \in \mathcal{F}$ be its minimizer, and $\widehat{f}_{\mathrm{SDU}} \in \mathcal{F}$ be a minimizer of the empirical risk $\widehat{R}^{\gamma}_{\mathrm{SDU}}$. Assume that $\pi_+ \neq \frac{1}{2}$, the loss function ℓ is ρ-Lipschitz function with respect to the first argument ($0 < \rho < \infty$), and all functions in the model class $\mathcal{F}$ are bounded, that is, there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t\in\{\pm1\}} \ell(C_b, t)$. For any $\delta > 0$, with probability at least $1 - \delta$,*

$$R(\widehat{f}_{\mathrm{SDU}}) - R(f^*) \leq C'_{\mathcal{F},\ell,\delta}\left((2\gamma_1 + \gamma_3)\frac{\pi_{\mathrm{S}}}{\sqrt{2n_{\mathrm{S}}}} + (2\gamma_2 + \gamma_3)\frac{\pi_{\mathrm{D}}}{\sqrt{2n_{\mathrm{D}}}}\right.$$
$$\left. + (|\gamma_1\pi_- - \gamma_2\pi_+| + |\gamma_1\pi_+ - \gamma_2\pi_-|)\frac{1}{\sqrt{n_{\mathrm{U}}}}\right), \quad (4.7)$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

where

$$C'_{\mathcal{F},\ell,\delta} = \frac{1}{|\pi_+ - \pi_-|}\left(4\rho C_{\mathcal{F}} + \sqrt{2C_\ell^2 \log\frac{12}{\delta}}\right). \tag{4.8}$$

Theorem 5 ensures that the estimation error bound of $\widehat{f}_{\mathrm{SDU}}$ diminishes asymptotically, that is, $R(f^*) - R(\widehat{f}_{\mathrm{SDU}}) \to 0$ as $n_{\mathrm{S}}, n_{\mathrm{D}}, n_{\mathrm{U}} \to \infty$. As a negative aspect, it should also be noted that $C'_{\mathcal{F},\ell,\delta}$ is inversely proportional to $|\pi_+ - \pi_-|$, which implies that the estimation error can increase as π_+ and π_- are approaching each other.

5 Experiments

In this section, we experimentally investigate the behavior of the proposed methods on benchmark data sets. First, we compare the performances of the SU, DU, and SD classification methods to confirm that the SD and DU classification methods are likely to perform better than the SU classification method, as discussed in section 4.2. Second, we demonstrate that unlabeled data can improve the classification accuracy of the SDU classification method. Finally, we compare the performance of the SDU classification method and those of baseline methods.

We conducted experiments on 10 benchmark data sets obtained from the UCI Machine Learning Repository (Dua & Graff, 2017) and LIBSVM (Chang & Lin, 2011). To obtain pairwise training data, we first converted pointwise labeled data into pairs by coupling. Then we randomly subsampled pairwise similar and dissimilar data following the ratio of π_{S} and π_{D}. To obtain unlabeled data, we randomly picked positive and negative data following the ratio of π_+ and π_-. The labeled data for testing are created in the same way as with unlabeled data, and the number of test data was set to 500.

In the SDU classification method, including the SU, DU, and SD classification methods, a linear-in-input model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ is used as a classifier. The weight of L_2 regularization was chosen from $\{10^{-1}, 10^{-4}, 10^{-7}\}$. Each of the coefficient parameters in $(\gamma_1, \gamma_2, \gamma_3)$ was chosen from $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ subject to $\gamma_1 + \gamma_2 + \gamma_3 = 1$. All hyperparameters were tuned with 5-fold cross validation on the empirical classification error computed from similarities and dissimilarities, that is, $\widehat{R}_{\mathrm{SD}}$ equipped with the zero-one loss. The squared loss is used for experiments in sections 5.1 and 5.2, and the double hinge loss is used in section 5.3. We assumed that the true positive class proportion π_+ is known for computing the empirical risk.

5.1 Comparison of SU, DU, and SD Performances. We compared the performances of the SU, DU, and SD classification methods. We set the number of unlabeled training data to 500 and the number of pairwise training data to each of {50, 100, 200, 300, 400, 500}. Training and test data were

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

generated with maintaining $\pi_+ = 0.7$. The misclassification rates for each method are plotted in Figure 2.

5.2 Performance Improvement with Unlabeled Data. We investigated the effect of unlabeled data in the SDU classification method. The number of pairwise data was fixed: $n_{\mathrm{SD}} = 50$. As with the previous experiment, training and test data were generated with maintaining $\pi_+ = 0.7$. Three methods, the SD, SDU, and SDDU classification methods, are evaluated in each setting. The misclassification rates for each method are plotted in Figure 3.

5.3 Benchmark Comparison of SDU and Existing Methods. We evaluated the performances of the SDU/SDDU classification methods and six baseline methods on benchmark data sets. We set $n_{\mathrm{U}} = 500$ and $n_{\mathrm{SD}} = \{50, 200\}$. In each trial, the misclassification rate was measured with 500 test examples. To see the influence of the class prior on our methods, we conducted experiments in a moderately imbalanced case ($\pi_+ = 0.7$) and a fairly imbalanced case ($\pi_+ = 0.9$), respectively. We report the results for each setup in Table 2. The details of the baseline methods are described below.

5.3.1 SU Classification (SU). The first baseline method is the SU classification method (Bao et al., 2018), where the classification risk is estimated from similar pairs and unlabeled data in an unbiased manner. This method is a special case of SDU classification where the coefficient parameters are fixed as $(\gamma_1, \gamma_2, \gamma_3) = (1, 0, 0)$.

5.3.2 KMeans Clustering (KM). K-means clustering (MacQueen, 1967) is one of the most popular unsupervised methods. It is applied to training data by ignoring all pairwise information. We predicted labels of test data with learned clusters.

5.3.3 Constrained KMeans Clustering (CKM). Constrained K-means clustering (Wagstaff et al., 2001) is a semisupervised clustering method based on K-means clustering, where pairwise similarities and dissimilarities are treated as must-links or cannot-links.

5.3.4 Semisupervised Spectral Clustering (SSP). Semisupervised spectral clustering was proposed in Chen and Feng (2012), where similar and dissimilar labels are propagated through an affinity matrix. We set $k = 5$ for constructing the affinity matrix with k-nearest-neighbors graph and $\sigma^2 = 1$ for a precision parameter used in similarity measurement.

5.3.5 Information-Theoretical Metric Learning (ITML). Information-theoretical metric learning (Davis et al., 2007) is an algorithm to learn a matrix that parameterizes the Mahalanobis distance on given data points.

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

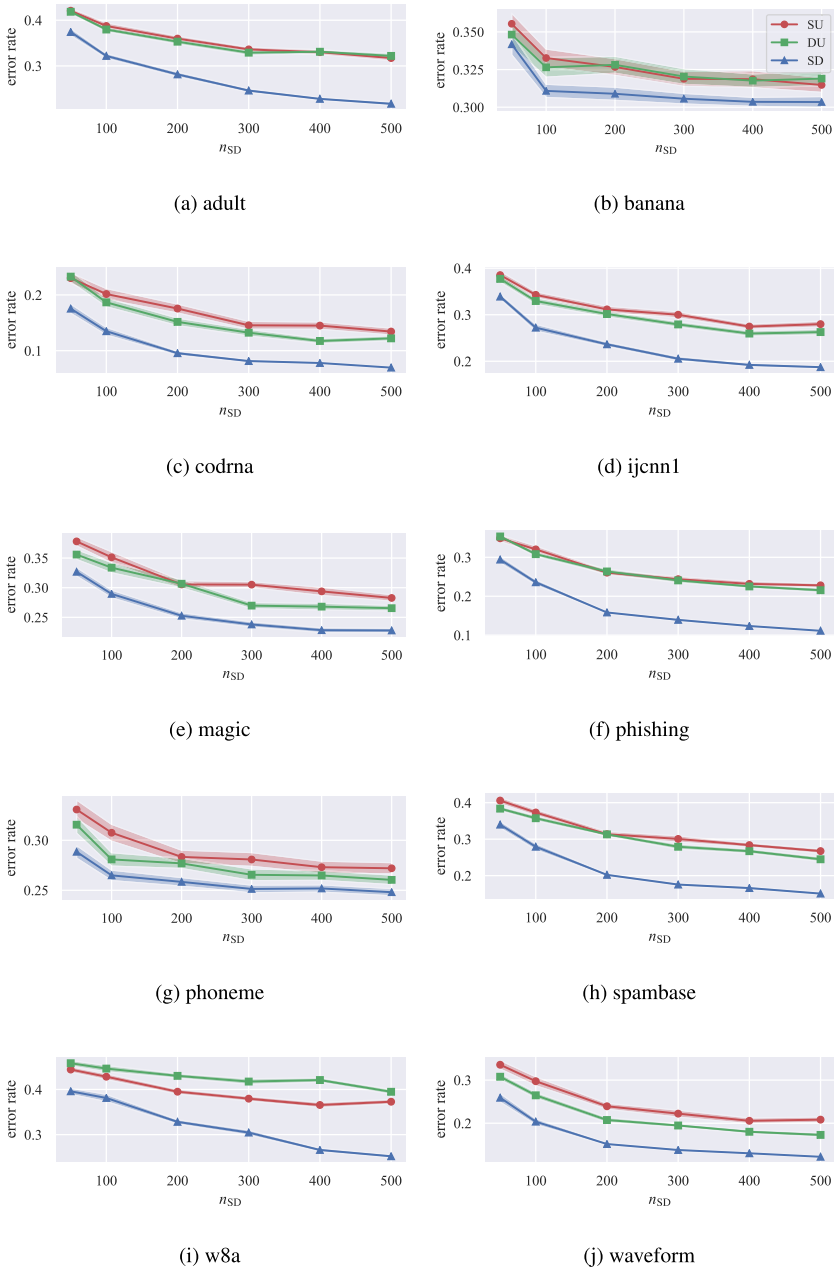


Figure 2: Average misclassification rate and standard error as a function of the number of similar and dissimilar pairs over 50 trials. Performances are shown for the SU (red), DU (green), and SD (blue) classification methods.

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

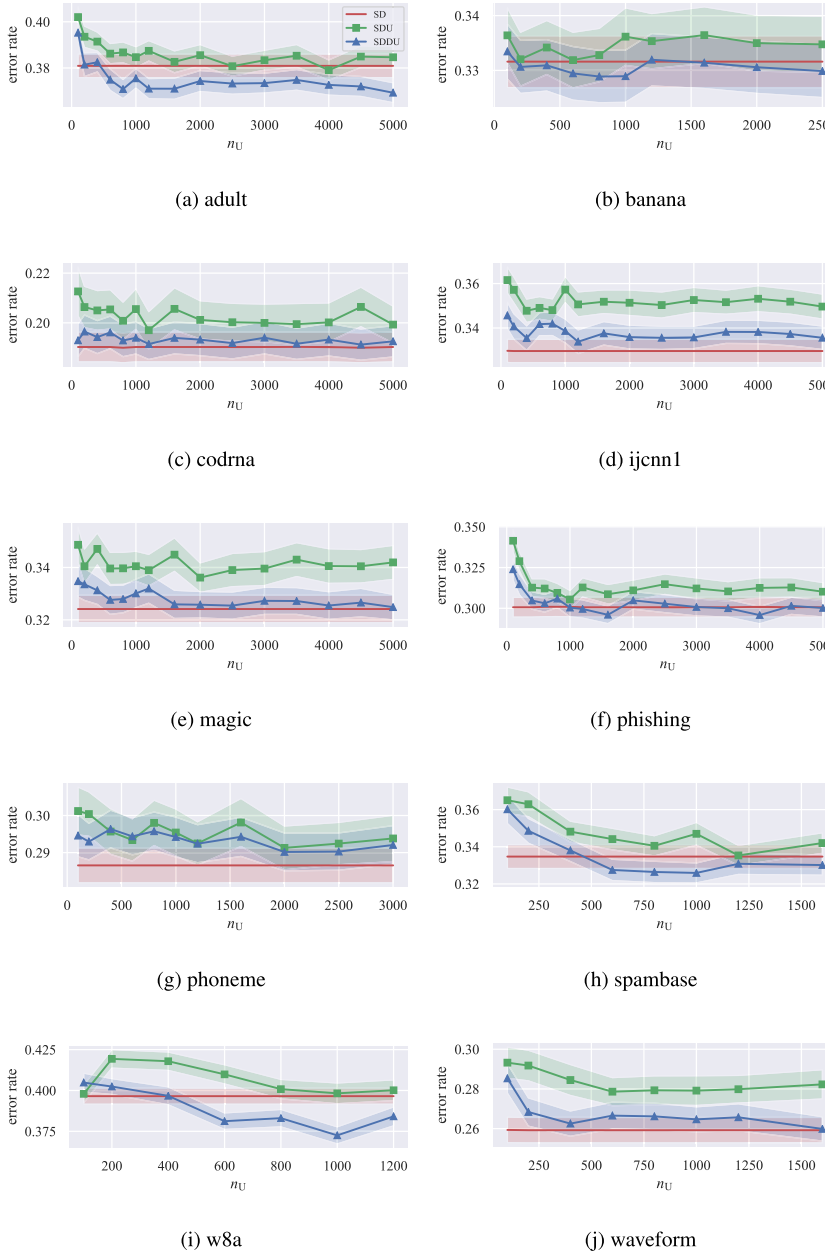


Figure 3: Average misclassification rate and standard error as a function of the number of unlabeled samples over 100 trials. Performances are shown for the SD (red), SDU (green), and SDDU (blue) classification methods.

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Table 2: Mean Misclassification Rate and Standard Error on Different Benchmark Data Sets over 50 Trials.

(a) $n_{\mathrm{SD}} = 50, \pi_+ = 0.7$

		Proposed		Baselines							
Data Set	# Dim.	SDU	SDDU	SU	KM	CKM	SSP	ITML	CRL	OVPC	MCL
adult	123	26.4 (1.12)	**23.6 (0.83)**	35.2 (1.03)	35.0 (0.81)	33.4 (1.05)	30.6 (0.29)	38.2 (0.64)	38.3 (0.99)	41.4 (0.87)	26.0 (1.10)
banana	2	**33.9 (0.73)**	**33.5 (0.68)**	35.7 (0.72)	47.1 (0.36)	47.3 (0.35)	41.3 (0.73)	47.1 (0.36)	43.7 (0.63)	38.0 (0.98)	**33.0 (0.77)**
codrna	8	**20.1 (1.14)**	**18.9 (1.17)**	24.6 (1.01)	37.4 (0.50)	38.5 (0.40)	45.4 (0.99)	37.4 (0.50)	41.0 (0.66)	37.1 (0.99)	32.1 (1.20)
ijcnn1	22	**33.0 (0.73)**	**32.2 (0.61)**	36.5 (0.96)	44.5 (0.63)	45.3 (0.50)	40.0 (0.79)	44.9 (0.71)	44.9 (0.60)	42.1 (0.89)	**32.9 (1.13)**
magic	10	34.5 (0.83)	34.2 (0.87)	40.0 (0.82)	47.6 (0.21)	48.2 (0.18)	47.3 (0.27)	47.6 (0.21)	42.2 (0.52)	43.2 (0.74)	**29.9 (1.20)**
phishing	68	22.1 (1.21)	21.4 (1.11)	27.3 (1.30)	37.4 (0.33)	37.4 (0.31)	31.9 (0.30)	37.4 (0.34)	33.8 (1.09)	39.7 (0.81)	**17.8 (1.58)**
phoneme	5	**29.6 (0.78)**	**29.2 (0.79)**	32.7 (0.98)	32.2 (0.34)	31.1 (0.45)	33.5 (1.01)	32.2 (0.34)	35.5 (1.01)	37.8 (0.92)	32.1 (0.91)
spambase	57	21.0 (1.40)	20.4 (1.27)	31.7 (1.71)	36.3 (1.08)	35.8 (1.06)	29.6 (0.31)	39.1 (1.10)	37.1 (1.14)	39.8 (0.92)	**14.3 (0.83)**
w8a	300	36.2 (1.26)	33.0 (1.19)	41.3 (0.80)	**30.7 (0.26)**	34.0 (0.67)	36.0 (0.69)	**31.1 (0.29)**	31.7 (0.36)	43.4 (0.75)	39.3 (1.07)
waveform	21	17.3 (0.98)	15.8 (0.88)	26.7 (1.48)	48.5 (0.17)	48.4 (0.18)	46.7 (0.35)	48.5 (0.17)	30.6 (1.66)	36.4 (1.43)	**12.2 (0.31)**
# Outperform		4	5	0	1	0	0	1	0	0	6

(b) $n_{\mathrm{SD}} = 50, \pi_+ = 0.9$

		Proposed		Baselines							
Data Set	# Dim.	SDU	SDDU	SU	KM	CKM	SSP	ITML	CRL	OVPC	MCL
adult	123	**9.8 (0.43)**	**9.7 (0.41)**	23.7 (0.66)	22.0 (1.55)	33.7 (1.42)	11.5 (0.26)	28.5 (1.46)	41.7 (0.79)	38.5 (0.94)	14.6 (1.25)
banana	2	**10.4 (0.23)**	**10.4 (0.22)**	12.7 (0.63)	45.5 (0.31)	46.0 (0.30)	31.7 (1.47)	45.4 (0.31)	39.3 (0.80)	23.9 (1.97)	**10.7 (0.32)**
codrna	8	**6.9 (0.39)**	**6.9 (0.45)**	15.8 (0.70)	32.9 (1.12)	36.6 (1.05)	42.3 (1.93)	32.8 (1.13)	38.0 (1.54)	33.1 (1.65)	10.0 (0.22)
ijcnn1	22	**10.0 (0.31)**	**9.9 (0.31)**	14.3 (0.62)	40.0 (1.00)	41.0 (0.81)	29.4 (1.34)	39.1 (1.12)	41.1 (0.73)	36.7 (1.55)	14.0 (1.32)
magic	10	**11.7 (0.33)**	**11.6 (0.29)**	21.8 (0.85)	36.4 (0.45)	38.8 (0.38)	45.6 (0.44)	36.4 (0.45)	29.5 (1.15)	41.1 (0.99)	14.5 (0.96)
phishing	68	**8.7 (0.38)**	**8.5 (0.32)**	18.0 (0.85)	24.7 (0.36)	25.7 (0.38)	13.6 (0.41)	24.8 (0.36)	33.1 (1.35)	40.5 (0.94)	**9.2 (0.94)**
phoneme	5	**11.1 (0.27)**	**11.3 (0.27)**	15.8 (0.67)	40.4 (0.51)	40.8 (0.35)	26.7 (1.76)	40.1 (0.56)	36.2 (1.39)	31.3 (1.62)	**11.7 (0.85)**
spambase	57	**8.8 (0.23)**	**8.5 (0.24)**	16.7 (0.53)	18.6 (1.48)	32.3 (1.12)	11.6 (0.34)	21.7 (1.46)	33.2 (1.88)	40.0 (0.98)	**8.1 (0.78)**
w8a	300	**8.3 (0.50)**	**8.3 (0.50)**	26.0 (0.56)	11.2 (0.19)	11.7 (0.22)	18.3 (1.00)	11.7 (0.26)	14.6 (0.80)	35.8 (1.23)	27.4 (1.52)
waveform	21	**5.2 (0.24)**	**5.1 (0.24)**	8.2 (0.73)	48.6 (0.17)	48.5 (0.18)	47.6 (0.23)	48.6 (0.17)	44.1 (0.71)	35.1 (1.52)	**5.4 (0.74)**
# Outperform		10	10	0	0	0	0	0	0	0	5

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Table 2: Continued.

(c) $n_{\mathrm{SD}} = 200, \pi_+ = 0.7$

		Proposed		Baselines							
Data Set	# Dim.	SDU	SDDU	SU	KM	CKM	SSP	ITML	CRL	OVPC	MCL
adult	123	**18.0 (0.46)**	**17.8 (0.31)**	21.3 (0.54)	36.7 (0.76)	28.1 (0.90)	30.7 (0.31)	39.1 (0.64)	33.0 (0.95)	42.7 (0.79)	19.7 (0.39)
banana	2	**30.5 (0.42)**	**31.0 (0.42)**	32.8 (0.52)	47.5 (0.24)	47.5 (0.24)	33.5 (1.29)	47.5 (0.24)	42.3 (0.50)	35.2 (0.77)	32.0 (0.67)
codrna	8	10.8 (0.69)	**9.4 (0.41)**	18.1 (0.88)	37.2 (0.51)	40.5 (0.50)	46.8 (0.72)	37.4 (0.54)	36.6 (1.33)	27.6 (1.28)	**10.5 (0.79)**
ijcnn1	22	23.3 (0.44)	22.6 (0.40)	28.4 (0.64)	45.8 (0.29)	46.9 (0.35)	40.5 (0.78)	46.4 (0.35)	44.3 (0.66)	43.1 (0.81)	**16.2 (0.64)**
magic	10	25.9 (0.65)	25.6 (0.55)	30.2 (0.71)	48.0 (0.18)	48.3 (0.17)	47.4 (0.29)	48.0 (0.18)	40.3 (0.59)	43.4 (0.77)	**22.1 (0.68)**
phishing	68	12.0 (0.42)	12.0 (0.41)	17.2 (0.87)	37.4 (0.31)	37.2 (0.30)	31.6 (0.29)	37.4 (0.31)	22.8 (1.34)	42.9 (0.67)	**7.0 (0.18)**
phoneme	5	**25.5 (0.49)**	**25.5 (0.49)**	27.5 (0.67)	32.2 (0.31)	29.0 (0.60)	28.0 (0.73)	32.0 (0.37)	32.9 (1.07)	37.1 (0.90)	**25.5 (0.51)**
spambase	57	12.8 (0.27)	12.3 (0.23)	16.2 (0.62)	38.2 (1.20)	29.6 (0.56)	29.4 (0.29)	40.4 (1.14)	34.8 (1.34)	39.6 (1.05)	**9.5 (0.20)**
w8a	300	**20.4 (0.80)**	**18.8 (0.69)**	35.8 (0.71)	30.7 (0.27)	43.7 (0.61)	32.6 (0.60)	31.7 (0.33)	33.2 (0.69)	45.1 (0.57)	27.6 (0.76)
waveform	21	12.8 (0.29)	12.6 (0.29)	15.6 (0.59)	48.5 (0.16)	48.5 (0.14)	46.9 (0.31)	48.4 (0.16)	17.5 (1.40)	35.3 (1.26)	**10.7 (0.19)**
# Outperform		4	5	0	0	0	0	0	0	0	7

(d) $n_{\mathrm{SD}} = 200, \pi_+ = 0.9$

		Proposed		Baselines							
Data Set	# Dim.	SDU	SDDU	SU	KM	CKM	SSP	ITML	CRL	OVPC	MCL
adult	123	**8.4 (0.24)**	**8.3 (0.24)**	11.2 (0.32)	27.4 (1.41)	43.6 (0.95)	11.1 (0.27)	27.8 (1.25)	43.3 (0.76)	39.3 (1.11)	9.0 (0.21)
banana	2	**10.2 (0.19)**	**10.2 (0.19)**	**10.5 (0.24)**	45.5 (0.30)	46.6 (0.32)	25.8 (1.77)	45.5 (0.29)	40.0 (0.81)	24.1 (1.63)	**10.2 (0.19)**
codrna	8	**4.1 (0.18)**	**4.0 (0.19)**	9.8 (0.39)	32.2 (1.10)	40.4 (0.79)	40.7 (2.11)	32.5 (1.17)	38.1 (1.16)	29.1 (1.53)	7.6 (0.20)
ijcnn1	22	8.4 (0.20)	8.3 (0.20)	9.4 (0.24)	40.4 (0.82)	43.1 (0.63)	27.7 (1.38)	41.1 (0.90)	41.2 (1.00)	38.9 (1.26)	**7.7 (0.16)**
magic	10	**10.3 (0.23)**	**10.2 (0.22)**	16.8 (0.73)	37.0 (0.33)	41.4 (0.34)	45.2 (0.43)	37.0 (0.32)	32.7 (1.50)	38.6 (1.35)	**10.0 (0.19)**
phishing	68	6.3 (0.21)	6.3 (0.22)	8.9 (0.38)	24.4 (0.26)	27.6 (0.33)	13.7 (0.38)	24.5 (0.28)	38.4 (1.22)	40.8 (0.81)	**3.7 (0.13)**
phoneme	5	**10.3 (0.21)**	**10.3 (0.19)**	12.4 (0.41)	40.2 (0.54)	40.5 (0.39)	24.9 (1.70)	40.3 (0.54)	33.2 (1.53)	34.1 (1.50)	**10.2 (0.19)**
spambase	57	7.5 (0.19)	7.5 (0.19)	8.1 (0.24)	20.2 (1.33)	40.4 (0.60)	10.9 (0.25)	22.9 (1.18)	31.8 (1.31)	40.6 (1.17)	**5.9 (0.15)**
w8a	300	**6.0 (0.18)**	**6.0 (0.18)**	17.2 (0.41)	11.2 (0.21)	18.1 (1.07)	12.6 (0.67)	11.7 (0.24)	12.8 (0.48)	38.8 (0.99)	9.1 (0.30)
waveform	21	**4.5 (0.13)**	**4.6 (0.14)**	5.2 (0.22)	48.5 (0.17)	48.5 (0.20)	47.6 (0.22)	48.5 (0.17)	39.3 (1.13)	34.7 (1.54)	**4.4 (0.14)**
# Outperform		7	7	1	0	0	0	0	0	0	7

Note: Bold numbers indicate outperforming methods, chosen by one-sided t-test with a significance level of 5% # Dim. = number of dimensions.

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Similar and dissimilar pairs are used for regularizing the covariance matrix. For test samples prediction, k-means clustering was applied with the obtained metric. We used the identity matrix as prior information, and a slack parameter γ was set to 1.

5.3.6 Contrastive Learning (CRL). Contrastive learning (Arora et al., 2019) is another framework for learning a useful representation by leveraging similarity information. We used a linear model $g(\boldsymbol{x}) = W\boldsymbol{x}$, where $W \in \mathbb{R}^{d' \times d}$ as an embedding function from input space to representation space. In this experiment, we fixed d' to 10 for all data sets. Each triplet used for training was created by concatenating a similar pair and an example randomly picked from unlabeled data. With the learned representations, K-means clustering was applied in the same manner as the KM method.

5.3.7 On the Value of Pairwise Constraints (OVPC). A classification-based approach was proposed in Zhang and Yan (2007), where an auxiliary classifier is trained on the feature vectors obtained from pairwise examples. The trained classifier is converted into a function that can be applied to pointwise prediction. The weight of L_2 regularization is chosen from $\{10^{-1}, 10^{-4}, 10^{-7}\}$ by five-fold cross-validation.

5.3.8 Meta-Classification Likelihood (MCL). A meta-learning approach was recently proposed by Hsu et al. (2019), where the objective is maximum likelihood estimation over similar and dissimilar labels. The conditional class probability was modeled by $p(y = 1 \mid \boldsymbol{x}) = \{1 + \exp(\boldsymbol{w}^\top \boldsymbol{x} + b)\}^{-1}$. A stochastic gradient descent algorithm was applied for optimization.

5.3.9 Setup for Clustering Algorithms. For clustering methods, the number of clusters was set to two. To evaluate the accuracy of k-means-based clustering methods (i.e., KM, CKM, and ITML), test samples were completely separated from training samples. The labels of test samples are predicted based on the clusters obtained from only training samples. For SSP, the clustering algorithm was applied to both train and test samples so that we could predict for test samples. Since there is no explicit positive or negative assignment in clustering methods, their performances are evaluated by $\min(r, 1 - r)$, where r is misclassification rate.

5.4 Discussion. In section 5.3, we stated that the SD and DU classification methods are likely to outperform the SU classification method, which comes from the comparison of their estimation error bounds. As shown in Figure 2, we confirmed that the misclassification rates of the SU, DU, and SD classification methods are consistent with this statement.

Figure 3 indicates that more unlabeled data lead to better classification performance for the SDU and SDDU classification methods. We also found that the SDDU classification method not only reduces the computation cost

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

for tuning the coefficient parameters but also often outperforms the SDU classification method. It might indicate the difficulty in tuning the coefficient parameters ($\gamma_1, \gamma_2, \gamma_3$) only with similarities and dissimilarities for cross validation.

Table 2 demonstrates that the SDU and SDDU classification methods perform better than or comparable to other baselines in many scenarios. Specifically, we observed that the superiority of the proposed methods becomes outstanding in the situations where the number of pairwise data is limited and the positive and negative class priors are fairly imbalanced (see Table 2b). The first property suggests that the advantage gained from unlabeled data becomes significant when the amount of pairwise supervision is relatively small. The second one is consistent with the theoretical analysis in section 4.3, which states that the estimation error of the proposed method can increase as two class priors are approaching each other. Furthermore, we confirmed that our methods always benefit from the increased number of pairwise data, while most other clustering-based methods do not.

6 Conclusion and Future Work

In this letter, we proposed a novel weakly supervised classification method, similar-dissimilar-unlabeled (SDU) classification, where the classification risk is computed from pairwise similarities and dissimilarities and unlabeled data. We derived the estimation error bound for the proposed method and confirmed convergence to the optimal solution. From the theoretical analysis, we developed a strategy to reduce the computation cost for tuning the hyperparameter. Through experiments on benchmark data sets, we demonstrated that our SDU classification method performs better than baseline methods.

We discuss three important directions for future work. First, further research in a multiclass classification scenario is required. Our formulation relies on the connection between classification of similarity and classification of binary class labels. Since both are classification with binary outcomes, the extension to the multiclass case is not straightforward unless additional information is available. Second, in the SDU classification method, the positive and negative class proportions must not be equal, that is, $\pi_+ \neq \pi_-$. Even if they are not exactly equal, the estimation error can increase when $\pi_+ \to \frac{1}{2}$, as mentioned in section 4.3. Our recent study (Bao, Shimada, Xu, Sato, & Sugiyama, 2020) partially overcomes this problem by ignoring the sign identification of a classifier, that is, a classifier is trained to minimize or maximize the classification error, but we cannot know which optimization is achieved without auxiliary information. Finally, the use of different types of pairwise supervision should be explored. Although this letter focused on binary representations of similarity and dissimilarity information, it would be more appealing if we can extend our method to handle other types of pairwise supervision, for example, confidence score (Ishida, Niu, &

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Sugiyama, 2018) and triplet comparison (Schroff, Kalenichenko, & Philbin, 2015; Cui, Charoenphakdee, Sato, & Sugiyama, 2020).

Appendix: Proofs of Theorems

In this appendix, we give complete proofs of the theorems in sections 3 and 4.

A.1 Proof of Theorem 1. We can express the joint density for pairwise unlabeled examples as

$$p(\boldsymbol{x}, \boldsymbol{x}') = p(\boldsymbol{x}, \boldsymbol{x}', \tau = +1) + p(\boldsymbol{x}, \boldsymbol{x}', \tau = -1) = \pi_{\mathrm{S}} p_{\mathrm{S}}(\boldsymbol{x}.\boldsymbol{x}') + \pi_{\mathrm{D}} p_{\mathrm{D}}(\boldsymbol{x}.\boldsymbol{x}'). \tag{A.1}$$

This provides the following relationship in conditional expectations:

$$\mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}[\cdot] = \pi_{\mathrm{S}} \mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}[\cdot] + \pi_{\mathrm{D}} \mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}[\cdot]. \tag{A.2}$$

In addition, we can rewrite the conditional expectation over pointwise unlabeled data into that over pairwise unlabeled data. For a binary variable $t \in \{+1, -1\}$, we have

$$\begin{aligned}
\mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X), t)\right] &= \frac{1}{2}\mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X), t)\right] + \frac{1}{2}\mathbf{E}_{X'\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X'), t)\right] \\
&= \frac{1}{2}\mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\mathcal{L}(f(X), t)\right] \\
&\quad + \frac{1}{2}\mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\mathcal{L}(f(X'), t)\right] \\
&= \mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X), t) + \mathcal{L}(f(X'), t)}{2}\right].
\end{aligned} \tag{A.3}$$

With equations A.2 and A.3, we can transform the SU risk defined in equation 2.13 into the DU risk and the SD risk as follows:

$$\begin{aligned}
R_{\mathrm{SU}}(f) &= \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X)) + \widetilde{\mathcal{L}}(f(X'))}{2}\right] + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X), -1)\right] \\
&= \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X)) + \widetilde{\mathcal{L}}(f(X'))}{2}\right] \\
&\quad + \mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X), -1) + \mathcal{L}(f(X'), -1)}{2}\right]
\end{aligned}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

$$
\begin{aligned}
&= \mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X))+\widetilde{\mathcal{L}}(f(X'))}{2}\right] \\
&\quad - \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X))+\widetilde{\mathcal{L}}(f(X'))}{2}\right] \\
&\quad + \mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X),-1)+\mathcal{L}(f(X'),-1)}{2}\right] \\
&= \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}\left[-\frac{\widetilde{\mathcal{L}}(f(X))+\widetilde{\mathcal{L}}(f(X'))}{2}\right] \\
&\quad + \mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X),+1)+\mathcal{L}(f(X'),+1)}{2}\right] \\
&= \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}\left[-\frac{\widetilde{\mathcal{L}}(f(X))+\widetilde{\mathcal{L}}(f(X'))}{2}\right] \\
&\quad + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X),+1)\right] \\
&= R_{\mathrm{DU}}(f),
\end{aligned}
$$

$$
\begin{aligned}
R_{\mathrm{SU}}(f) &= \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X))+\widetilde{\mathcal{L}}(f(X'))}{2}\right] + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X),-1)\right] \\
&= \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X))+\widetilde{\mathcal{L}}(f(X'))}{2}\right] \\
&\quad + \mathbf{E}_{(X,X')\sim p(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X),-1)+\mathcal{L}(f(X'),-1)}{2}\right] \\
&= \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\widetilde{\mathcal{L}}(f(X))+\widetilde{\mathcal{L}}(f(X'))}{2}\right] \\
&\quad + \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X),-1)+\mathcal{L}(f(X'),-1)}{2}\right] \\
&\quad + \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X),-1)+\mathcal{L}(f(X'),-1)}{2}\right] \\
&= \pi_{\mathrm{S}}\mathbf{E}_{(X,X')\sim p_{\mathrm{S}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X),-1)+\mathcal{L}(f(X'),+1)}{2}\right] \\
&\quad + \pi_{\mathrm{D}}\mathbf{E}_{(X,X')\sim p_{\mathrm{D}}(\boldsymbol{x},\boldsymbol{x}')}\left[\frac{\mathcal{L}(f(X),+1)+\mathcal{L}(f(X'),-1)}{2}\right] \\
&= R_{\mathrm{SD}}(f).
\end{aligned}
$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

As shown in proposition 1, R_{SU} is an equivalent expression of the classification risk. Therefore, R_{DU} and R_{SD} are also equivalent expressions of the classification risk. □

A.2 Proof of Theorem 2. We prove this theorem based on the positive semidefiniteness of the Hessian matrix similarly to SU classification in Bao et al. (2018). Since ℓ is a twice-differentiable margin-based loss, there is a twice differentiable function $\psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that $\ell(z,t) = \psi(tz)$. Here, our objective function, $J(\boldsymbol{w}) := \widehat{R}^{\boldsymbol{\gamma}}_{\mathrm{SDU}}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$, can be written as

$$
\begin{aligned}
J(\boldsymbol{w}) = {} & \frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} - \frac{\gamma_1 \pi_{\mathrm{S}}}{2n_{\mathrm{S}}(\pi_+ - \pi_-)} \sum_{i=1}^{2n_{\mathrm{S}}} \boldsymbol{w}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}) + \frac{\gamma_2 \pi_{\mathrm{D}}}{2n_{\mathrm{D}}(\pi_+ - \pi_-)} \sum_{i=1}^{2n_{\mathrm{D}}} \boldsymbol{w}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}) \\
& + \frac{\gamma_3 \pi_{\mathrm{S}}}{2n_{\mathrm{S}}(\pi_+ - \pi_-)} \sum_{i=1}^{2n_{\mathrm{S}}} \left(\pi_+ \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}), +1) - \pi_- \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}), -1)\right) \\
& - \frac{\gamma_3 \pi_{\mathrm{D}}}{2n_{\mathrm{D}}(\pi_+ - \pi_-)} \sum_{i=1}^{2n_{\mathrm{D}}} \left(\pi_- \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}), +1) - \pi_+ \ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}), -1)\right) \\
& + \frac{1}{n_{\mathrm{U}}(\pi_+ - \pi_-)} \sum_{i=1}^{n_{\mathrm{U}}} \big((\gamma_2 \pi_+ - \gamma_1 \pi_-)\ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), +1) \\
& \qquad + (\gamma_1 \pi_+ - \gamma_2 \pi_-)\ell(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}), -1)\big). \qquad \text{(A.4)}
\end{aligned}
$$

The second-order derivative of $\ell(z,t)$ with respect to z can be computed as

$$
\frac{\partial^2 \ell(z,t)}{\partial z^2} = \frac{\partial^2 \psi(tz)}{\partial z^2} = t^2 \frac{\partial^2 \psi(\xi)}{\partial \xi^2} = \frac{\partial^2 \psi(\xi)}{\partial \xi^2}, \qquad \text{(A.5)}
$$

where $\xi = tz$ is employed in the second equality and $t \in \{+1, -1\}$ is employed in the last equality. Here, the Hessian of $J(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ is

$$
\begin{aligned}
\boldsymbol{H}J(\boldsymbol{w}) = \lambda I + \frac{\partial^2 \psi(\xi)}{\partial \xi^2} & \left(\frac{\gamma_3}{2n_{\mathrm{S}}} \sum_{i=1}^{2n_{\mathrm{S}}} \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}) \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{S},i})^\top + \frac{\gamma_3}{2n_{\mathrm{D}}} \sum_{i=1}^{2n_{\mathrm{D}}} \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}) \boldsymbol{\phi}(\widetilde{\boldsymbol{x}}_{\mathrm{D},i})^\top \right. \\
& \left. + \frac{\gamma_1 + \gamma_2}{n_{\mathrm{U}}} \sum_{i=1}^{n_{\mathrm{U}}} \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i}) \boldsymbol{\phi}(\boldsymbol{x}_{\mathrm{U},i})^\top \right) \succeq 0, \qquad \text{(A.6)}
\end{aligned}
$$

where $A \succeq 0$ means that a matrix A is positive semidefinite. Positive semidefiniteness of $\boldsymbol{H}J(\boldsymbol{w})$ follows from $\frac{\partial^2 \psi(\xi)}{\partial \xi^2} \geq 0$ ($\because \ell$ is convex) and $\boldsymbol{\phi}(\widetilde{\boldsymbol{x}})\boldsymbol{\phi}(\widetilde{\boldsymbol{x}})^\top \succeq 0$. Therefore, $J(\boldsymbol{w})$ is convex with respect to $\boldsymbol{w}$. □

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

A.3 Proof of Theorem 3. We apply a similar technique with the SU classification method to the DU and SD classification methods. Using pointwise distributions defined in equations 3.12 and 3.13, we have the following lemma.

Lemma 1. *Assume that $\pi_+ \neq \frac{1}{2}$. Given any function $f : \mathcal{X} \to \mathbb{R}$, we denote $R_{\widetilde{\mathrm{SU}}}$, $R_{\widetilde{\mathrm{DU}}}$, and $R_{\widetilde{\mathrm{SD}}}$ by*

$$R_{\widetilde{\mathrm{SU}}} := \pi_{\mathrm{S}}\mathbf{E}_{X\sim\widetilde{p}_{\mathrm{S}}(\boldsymbol{x})}\left[\widetilde{\mathcal{L}}(f(X))\right] + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X),-1)\right], \tag{A.7}$$

$$R_{\widetilde{\mathrm{DU}}} := \pi_{\mathrm{D}}\mathbf{E}_{X\sim\widetilde{p}_{\mathrm{D}}(\boldsymbol{x})}\left[-\widetilde{\mathcal{L}}(f(X))\right] + \mathbf{E}_{X\sim p_{\mathrm{U}}(\boldsymbol{x})}\left[\mathcal{L}(f(X),+1)\right], \tag{A.8}$$

$$R_{\widetilde{\mathrm{SD}}} := \pi_{\mathrm{S}}\mathbf{E}_{X\sim\widetilde{p}_{\mathrm{S}}(\boldsymbol{x})}\left[\mathcal{L}(f(X),+1)\right] + \pi_{\mathrm{D}}\mathbf{E}_{X\sim\widetilde{p}_{\mathrm{D}}(\boldsymbol{x})}\left[\mathcal{L}(f(X),-1)\right]. \tag{A.9}$$

Then, $R_{\widetilde{\mathrm{SU}}}$, $R_{\widetilde{\mathrm{DU}}}$, and $R_{\widetilde{\mathrm{SD}}}$ are equivalent to R_{SU}, R_{DU}, and R_{SD}, respectively.

Here, empirical versions of the above risks are defined as

$$\widehat{R}_{\widetilde{\mathrm{SU}}} := \frac{\pi_{\mathrm{S}}}{2n_{\mathrm{S}}}\sum_{i=1}^{2n_{\mathrm{S}}}\widetilde{\mathcal{L}}(f(\widetilde{\boldsymbol{x}}_{\mathrm{S},i})) + \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\mathcal{L}(f\boldsymbol{x}_{\mathrm{U},i}),-1), \tag{A.10}$$

$$\widehat{R}_{\widetilde{\mathrm{DU}}} := -\frac{\pi_{\mathrm{D}}}{2n_{\mathrm{D}}}\sum_{i=1}^{2n_{\mathrm{D}}}\widetilde{\mathcal{L}}(f(\widetilde{\boldsymbol{x}}_{\mathrm{D},i})) + \frac{1}{n_{\mathrm{U}}}\sum_{i=1}^{n_{\mathrm{U}}}\mathcal{L}(f(\boldsymbol{x}_{\mathrm{U},i}),+1), \tag{A.11}$$

$$\widehat{R}_{\widetilde{\mathrm{SD}}} := \frac{\pi_{\mathrm{S}}}{2n_{\mathrm{S}}}\sum_{i=1}^{2n_{\mathrm{S}}}\mathcal{L}(f(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}),+1) + \frac{\pi_{\mathrm{D}}}{2n_{\mathrm{D}}}\sum_{i=1}^{2n_{\mathrm{D}}}\mathcal{L}(f(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}),-1). \tag{A.12}$$

Note that these empirical risks are also equivalent to $\widehat{R}_{\mathrm{SU}}$, $\widehat{R}_{\mathrm{DU}}$, and $\widehat{R}_{\mathrm{SD}}$. Now, we introduce the uniform deviation bound, which is useful to derive estimation error bounds. The proof can be found in the textbooks such as Mohri et al. (2012).

Lemma 2. *Let Z be a random variable drawn from a probability distribution with density μ, $\mathcal{H} = \{h : \mathcal{Z} \to [0, M]\}(M > 0)$ be a class of measurable functions, and $\{z_i\}_{i=1}^n$ be i.i.d. samples drawn from the distribution with density μ. Then, for any $\delta > 0$, with the probability at least $1-\delta$,*

$$\sup_{h\in\mathcal{H}}\left|\mathbf{E}_{Z\sim\mu}[h(Z)] - \frac{1}{n}\sum_{i=1}^{n}h(z_i)\right| \leq 2\mathfrak{R}(\mathcal{H};\mu,n) + \sqrt{\frac{M^2\log\frac{2}{\delta}}{2n}}. \tag{A.13}$$

We can derive the estimation error bound for the SU classification method as

$$\begin{aligned} R(\widehat{f}_{\mathrm{SU}}) - R(f^*) &= R_{\mathrm{SU}}(\widehat{f}_{\mathrm{SU}}) - R_{\mathrm{SU}}(f^*) \\ &\leq \left(R_{\mathrm{SU}}(\widehat{f}_{\mathrm{SU}}) - \widehat{R}_{\mathrm{SU}}(\widehat{f}_{\mathrm{SU}})\right) + \left(\widehat{R}_{\mathrm{SU}}(f^*) - R_{\mathrm{SU}}(f^*)\right) \end{aligned}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

$$
\begin{aligned}
&\leq 2 \sup_{f \in \mathcal{F}} \left| R_{\mathrm{SU}}(f) - \widehat{R}_{\mathrm{SU}}(f) \right| \\
&= 2 \sup_{f \in \mathcal{F}} \left| R_{\widetilde{\mathrm{SU}}}(f) - \widehat{R}_{\widetilde{\mathrm{SU}}}(f) \right| \\
&= 2\pi_{\mathrm{S}} \sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim \widetilde{p}_{\mathrm{S}}} \left[\widetilde{\mathcal{L}}(f(X)) \right] - \frac{1}{2n_{\mathrm{S}}} \sum_{i=1}^{2n_{\mathrm{S}}} \widetilde{\mathcal{L}}(f(\widetilde{\boldsymbol{x}}_{\mathrm{S},i})) \right| \\
&\quad + 2 \sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim p_{\mathrm{U}}} \left[\mathcal{L}(f(X), -1) \right] - \frac{1}{n_{\mathrm{U}}} \sum_{i=1}^{n_{\mathrm{U}}} \mathcal{L}(f(\boldsymbol{x}_{\mathrm{U},i}), -1) \right|.
\end{aligned} \tag{A.14}
$$

In the same way, for DU and SD, we have

$$
\begin{aligned}
R(\widehat{f}_{\mathrm{DU}}) - R(f^*) &\leq 2\pi_{\mathrm{D}} \sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim \widetilde{p}_{\mathrm{D}}} \left[\widetilde{\mathcal{L}}(f(X)) \right] - \frac{1}{2n_{\mathrm{D}}} \sum_{i=1}^{2n_{\mathrm{D}}} \widetilde{\mathcal{L}}(f(\widetilde{\boldsymbol{x}}_{\mathrm{D},i})) \right| \\
&\quad + 2 \sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim p_{\mathrm{U}}} \left[\mathcal{L}(f(X), +1) \right] - \frac{1}{n_{\mathrm{U}}} \sum_{i=1}^{n_{\mathrm{U}}} \mathcal{L}(f(\boldsymbol{x}_{\mathrm{U},i}), +1) \right|,
\end{aligned} \tag{A.15}
$$

$$
\begin{aligned}
R(\widehat{f}_{\mathrm{SD}}) - R(f^*) &\leq 2\pi_{\mathrm{S}} \sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim \widetilde{p}_{\mathrm{S}}} \left[\mathcal{L}(f(X), +1) \right] - \frac{1}{2n_{\mathrm{S}}} \sum_{i=1}^{2n_{\mathrm{S}}} \mathcal{L}(f(\widetilde{\boldsymbol{x}}_{\mathrm{S},i}), +1) \right| \\
&\quad + 2\pi_{\mathrm{D}} \sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim \widetilde{p}_{\mathrm{D}}} \left[\mathcal{L}(f(X), -1) \right] - \frac{1}{2n_{\mathrm{D}}} \sum_{i=1}^{2n_{\mathrm{D}}} \mathcal{L}(f(\widetilde{\boldsymbol{x}}_{\mathrm{D},i}), -1) \right|.
\end{aligned} \tag{A.16}
$$

To obtain the upper bound of the right-hand side for each algorithm, we derive the uniform deviation bound for $\widetilde{\mathcal{L}}(f(\cdot))$ and $\mathcal{L}(f(\cdot), \pm 1)$ as follows:

Lemma 3. *Assume that $\pi_+ \neq \frac{1}{2}$, the loss function ℓ is ρ-Lipschitz function with respect to the first argument ($0 < \rho < \infty$), and all functions in the model class $\mathcal{F}$ are bounded, that is, there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t \in \{\pm 1\}} \ell(C_b, t)$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d. samples drawn from a probability distribution with density p. For any $\delta > 0$, each of the following inequality holds with probability at least $1 - \delta$*

$$
\sup_{f \in \mathcal{F}} \left| \mathbf{E}_{X \sim p} \left[\widetilde{\mathcal{L}}(f(X)) \right] - \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathcal{L}}(f(\boldsymbol{x}_i)) \right| \leq \frac{4\rho C_{\mathcal{F}} + \sqrt{2C_\ell^2 \log \frac{4}{\delta}}}{|\pi_+ - \pi_-| \sqrt{n}}, \tag{A.17}
$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

$$\sup_{f\in\mathcal{F}}\left|\mathbf{E}_{X\sim p}\left[\mathcal{L}(f(X),+1)\right]-\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f(\boldsymbol{x}_i),+1)\right|\leq\frac{2\rho C_{\mathcal{F}}+\sqrt{\frac{1}{2}C_\ell^2\log\frac{4}{\delta}}}{|\pi_+-\pi_-|\sqrt{n}},\tag{A.18}$$

$$\sup_{f\in\mathcal{F}}\left|\mathbf{E}_{X\sim p}\left[\mathcal{L}(f(X),-1)\right]-\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f(\boldsymbol{x}_i),-1)\right|\leq\frac{2\rho C_{\mathcal{F}}+\sqrt{\frac{1}{2}C_\ell^2\log\frac{4}{\delta}}}{|\pi_+-\pi_-|\sqrt{n}}.\tag{A.19}$$

Proof.

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\mathbf{E}_{X\sim p}\left[\widetilde{\mathcal{L}}(f(X))\right]-\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{L}}(f(\boldsymbol{x}_i))\right|\\
&\leq\frac{1}{|\pi_+-\pi_-|}\underbrace{\sup_{f\in\mathcal{F}}\left|\mathbf{E}_{X\sim p}\left[\ell(f(X),+1)\right]-\frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}_i),+1)\right|}_{\text{with the probability at least }1-\delta/2}\\
&\quad+\frac{1}{|\pi_+-\pi_-|}\underbrace{\sup_{f\in\mathcal{F}}\left|\mathbf{E}_{X\sim p}\left[\ell(f(X),-1)\right]-\frac{1}{n}\sum_{i=1}^{n}\ell(f(\boldsymbol{x}_i),-1)\right|}_{\text{with the probability at least }1-\delta/2}\\
&\leq\frac{1}{|\pi_+-\pi_-|}\underbrace{\left\{4\mathfrak{R}(\ell\circ\mathcal{F};n,p)+\sqrt{\frac{2C_\ell^2\log\frac{4}{\delta}}{n}}\right\}}_{\text{with the probability at least }1-\delta},
\end{aligned}\tag{A.20}$$

where $\ell\circ\mathcal{F}$ indicates the class $\{\ell\circ f\mid f\in\mathcal{F}\}$. By applying Talagrand's lemma,

$$\mathfrak{R}(\ell\circ\mathcal{F};n,p)\leq\rho\mathfrak{R}(\mathcal{F};n,p).\tag{A.21}$$

With the assumption in equation 4.2, we obtain

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\mathbf{E}_{X\sim p}\left[\widetilde{\mathcal{L}}(f(X))\right]-\frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{L}}(f(\boldsymbol{x}_i))\right|\\
&\leq\frac{1}{|\pi_+-\pi_-|}\left\{4\rho\frac{C_{\mathcal{F}}}{\sqrt{n}}+\sqrt{\frac{2C_\ell^2\log\frac{4}{\delta}}{n}}\right\},
\end{aligned}$$

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

$$= \frac{4\rho C_{\mathcal{F}} + \sqrt{2C_\ell^2 \log \frac{4}{\delta}}}{|\pi_+ - \pi_-|\sqrt{n}}. \tag{A.22}$$

The bounds for $\mathcal{L}(f(\cdot), \pm 1)$ can be proven similarly to $\widetilde{\mathcal{L}}(f(\cdot))$. □

By combining lemma 3 and equations A.14 to A.16, we complete the proof of theorem 3. □

A.4 Proof of Theorem 5. Let $R^{\boldsymbol{\gamma}}_{\mathrm{SDU}}(f) := \gamma_1 R_{\mathrm{SU}}(f) + \gamma_2 R_{\mathrm{DU}}(f) + \gamma_3 R_{\mathrm{SD}}(f)$. We can rewrite this risk as follows:

$$\begin{aligned} R^{\boldsymbol{\gamma}}_{\mathrm{SDU}}(f) = {} & \frac{\pi_{\mathrm{S}}}{\pi_+ - \pi_-} \mathbf{E}_{X \sim \tilde{p}_{\mathrm{S}}(x)} \big[(\gamma_1 + \gamma_3 \pi_+) \ell(f(X), +1) \\ & \qquad - (\gamma_1 + \gamma_3 \pi_-) \ell(f(X), -1) \big] \\ & + \frac{\pi_{\mathrm{D}}}{\pi_+ - \pi_-} \mathbf{E}_{X \sim \tilde{p}_{\mathrm{D}}(x)} \big[-(\gamma_2 + \gamma_3 \pi_-) \ell(f(X), +1) \\ & \qquad + (\gamma_2 + \gamma_3 \pi_+) \ell(f(X), -1) \big] \\ & + \frac{1}{\pi_+ - \pi_-} \mathbf{E}_{X \sim p_{\mathrm{U}}(x)} \big[(\gamma_2 \pi_+ - \gamma_1 \pi_-) \ell(f(X), +1) \\ & \qquad + (\gamma_1 \pi_+ - \gamma_2 \pi_-) \ell(f(X), -1) \big]. \end{aligned} \tag{A.23}$$

Applying the uniform deviation bounds for each S, D, and U term as in theorem 3, theorem 5 can be proven. □

Acknowledgments

H.B. was supported by JST ACT-I grant JPMJPR18UI. I.S. was supported by JST CREST grant JPMJCR17A1, Japan. M.S. was supported by JST CREST grant JPMJCR1403.

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., & Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*.

Bao, H., Niu, G., & Sugiyama, M. (2018). Classification from pairwise similarity and unlabeled data. In *Proceedings of the 35th International Conference on Machine Learning* (p. 452).

Bao, H., Shimada, T., Xu, L., Sato, I., & Sugiyama, M. (2020). *Similarity-based classification: Connecting similarity learning to binary classification*. arXiv:2006.06207.

Basu, S., Banerjee, A., & Mooney, R. (2002). Semi-supervised clustering by seeding. In *Proceedings of 19th International Conference on Machine Learning* (p. 27).

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Basu, S., Davidson, I., & Wagstaff, K. (2008). *Constrained clustering: Advances in algorithms, theory, and applications*. Boca Raton, FL: CRC Press.

Bilenko, M., Basu, S., & Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the 21st International Conference on Machine Learning* (p. 839).

Chang, C.-C. & Lin, C.-J. (2011, May). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, art. 27.

Chapelle, O., Schölkopf, B., & Zien, A. (2010). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Charoenphakdee, N., Lee, J., & Sugiyama, M. (2019). On symmetric losses for learning from corrupted labels. In *Proceedings of the 36th International Conference on Machine Learning* (p. 961).

Chen, W., & Feng, G. (2012). Spectral clustering: A semi-supervised approach. *Neurocomputing*, *77*, 229–242.

Chiang, K.-Y., Hsieh, C.-J., & Dhillon, I. S. (2015). Matrix completion with noisy side information. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 3447–3455). Red Hook, NY: Curran.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 539–546). Piscataway, NJ: IEEE.

Cui, Z., Charoenphakdee, N., Sato, I., & Sugiyama, M. (2020). Classification from triplet comparison data. *Neural Computation*, *32*(3), 659–681.

Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 209–216).

du Plessis, M. C., Niu, G., & Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27* (pp. 703–711). Red Hook, NY: Curran.

du Plessis, M. C., Niu, G., & Sugiyama, M. (2015). Convex formulation for learning from positive and unlabeled data. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 1386–1394).

Dua, D., & Graff, C. (2017). *UCI machine learning repository*.

Ghosh, A., Manwani, N., & Sastry, P. (2015). Making risk minimization tolerant to label noise. *Neurocomputing*, *160*, 93–107.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1735–1742). Piscataway, NJ: IEEE.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2019). Learning deep representations by mutual information estimation and maximization. In *Proceedings of the International Conference on Learning Representations*. Madison, WI: Omnipress.

Hsu, Y.-C., Lv, Z., Schlosser, J., Odom, P., & Kira, Z. (2019). Multi-class classification without multi-class labels. In *Proceedings of the International Conference on Learning Representations*. Madison, WI: Omnipress.

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Hu, Y., Wang, J., Yu, N., & Hua, X.-S. (2008). Maximum margin clustering with pairwise constraints. In *Proceedings of the Eighth IEEE International Conference on Data Mining* (pp. 253–262). Piscataway, NJ: IEEE.

Ishida, T., Niu, G., & Sugiyama, M. (2018). Binary classification from positive- confidence data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems*, *31* (pp. 5917–5928). Red Hook, NY: Curran.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 3294–3302). Red Hook, NY: Curran.

Klein, D., Kamvar, S. D., & Manning, C. D. (2002). From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the 19th International Conference on Machine Learning* (pp. 307–314).

Li, Z., & Liu, J. (2009). Constrained clustering by spectral kernel learning. In *Proceedings of the IEEE 12th International Conference on Computer Vision* (pp. 421–427). Piscataway, NJ: IEEE.

Logeswaran, L., & Lee, H. (2018). An efficient framework for learning sentence representations. In *Proceedings of the International Conference on Learning Representations*. Madison, WI: Omnipress.

Lu, N., Niu, G., Menon, A. K., & Sugiyama, M. (2019). On the minimal supervision for training any binary classifier from only unlabeled data. In *International Conference on Learning Representations*. Madison, WI: Omnipress.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281–297). Berkeley: University of California Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *26* (pp. 3111–3119). Red Hook, NY: Curran.

Mohri, M., Rostamizadeh, A., Bach, F., & Talwalkar, A. (2012). *Foundations of machine learning*. Cambridge, MA: MIT Press.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280.

Niu, G., Dai, B., Yamada, M., & Sugiyama, M. (2012). Information-theoretic semisupervised metric learning via entropy regularization. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 89–96).

Okamoto, M. (1959). Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, *10*(1), 29–35.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). *Representation learning with contrastive predictive coding*. arXiv:1807.03748.

Patrini, G., Nielsen, F., Nock, R., & Carioni, M. (2016). Loss factorization, weakly supervised learning and label noise robustness. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 708–717).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018*

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2227–2237). Stroudsburg, PA: ACL.

Sakai, T., du Plessis, M. C., Niu, G., & Sugiyama, M. (2017). Semi-supervised classification based on classification from positive and unlabeled data. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 2998–3006).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823). Piscataway, NJ: IEEE.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, *29* (pp. 1857–1865). Red Hook, NY: Curran.

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained K-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 577–584).

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

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*, 207–244.

Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., . . . Niu, G. (2020). *Class2Simi: A new perspective on learning with label noise*. arXiv:2006.07831.

Xing, E. P., Jordan, M. I., Russell, S. J., & Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In S. Becker, S. Thrun, & K. Overmayer (Eds.), *Advances in neural information processing systems*, *15* (pp. 521–528). Cambridge, MA: MIT Press.

Yan, R., Zhang, J., Yang, J., & Hauptmann, A. G. (2006). A discriminative learning framework with pairwise constraints for video object classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 578–593.

Yi, J., Zhang, L., Jin, R., Qian, Q., & Jain, A. (2013). Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 1400–1408).

Zhang, J., & Yan, R. (2007). On the value of pairwise constraints in classification and consistency. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 1111–1118).

Received January 31, 2020; accepted November 18, 2020.

Downloaded from http://direct.mit.edu/neco/article-pdf/33/5/1234/1908985/neco_a_01373.pdf by guest on 23 September 2021