Multi-class Classification from Multiple Unlabeled Datasets with Partial Risk Regularization

Yuting Tang   TANG@MS.K.U-TOKYO.AC.JP
Nan Lu   LU@MS.K.U-TOKYO.AC.JP
Tianyi Zhang   ZHANG@MS.K.U-TOKYO.AC.JP
The University of Tokyo/RIKEN, Japan.

Masashi Sugiyama   SUGI@K.U-TOKYO.AC.JP
RIKEN/The University of Tokyo, Japan.

Abstract

Recent years have witnessed a great success of *supervised* deep learning, where predictive models were trained from a large amount of *fully* labeled data. However, in practice, labeling such big data can be very costly and may not even be possible for privacy reasons. Therefore, in this paper, we aim to learn an accurate classifier without any class labels. More specifically, we consider the case where multiple sets of *unlabeled* data and only their *class priors*, i.e., the proportions of each class, are available. Under this problem setup, we first derive an unbiased estimator of the classification risk that can be estimated from the given unlabeled sets and theoretically analyze the generalization error of the learned classifier. We then find that the classifier obtained as such tends to cause overfitting as its empirical risks go negative during training. To prevent overfitting, we further propose a partial risk regularization that maintains the partial risks with respect to unlabeled datasets and classes to certain levels. Experiments demonstrate that our method effectively mitigates overfitting and outperforms state-of-the-art methods for learning from multiple unlabeled sets.

Keywords: Unlabeled data; class prior; unbiased estimator; overfitting; regularization.

1. Introduction

Supervised deep learning approaches are very successful but data-hungry (Goodfellow et al., 2016). As it is very expensive and time-consuming to collect full labels for big data, it is desirable for machine learning techniques to work with weaker forms of supervision (Zhou, 2018; Sugiyama et al., 2022).

In this paper, we consider a challenging weakly supervised multi-class classification problem: instead of fully labeled data, we are given multiple sets of only unlabeled (U) data for classifier training. We assume that each available unlabeled dataset varies only in class-prior probabilities, and these class priors are known (Lu et al., 2019, 2020). Our problem setting traces back to a classical problem of *learning with label proportions* (LLP) (Quadrianto et al., 2009); however, there are two key differences between our work and previous LLP studies. First, most of the existing LLP papers consider U sets sampled from the identical distribution (Liu et al., 2019; Tsai and Lin, 2020); while in our case, the U sets are generated from distributions with different class priors. Second, for the

solution, the majority of deep LLP methods are based on the *empirical proportion risk minimization* (EPRM) (Yu et al., 2013, 2015) principle, which aims at predicting accurate label proportions for the U sets; while our method is based on the ordinary *empirical risk minimization* (ERM) (Vapnik, 1998) principle, which is superior to EPRM as the consistency of learning can be guaranteed.

Such a framework of learning from multiple unlabeled datasets can find applications in various real-world scenarios. For example, in politics, predicting the demographic information of users in social networks is critical to policy making (Culotta et al., 2015); however, labeling individual users according to their demographic information may violate data privacy, and thus the traditional supervised learning methods cannot be applied. Fortunately, multiple unlabeled datasets can be obtained from different time points or geographic locations, and their class priors can be collected from pre-existing census data, which suffices for our learning framework.

Under this problem setup, the first challenge is to estimate the classification risk from only unlabeled datasets and their class priors, since the typical ERM approach used in supervised learning cannot be directly applied to this unlabeled setting. A promising direction to solve the problem is risk rewriting (Sugiyama et al., 2022), i.e., rewrite the classification risk into an equivalent form that can be estimated from the given data. Existing risk rewriting results on learning from multiple U sets focused on binary classification (Lu et al., 2019, 2020). In this paper, we extend them and derive an unbiased risk estimator for multi-class classification from multiple U sets. In this way, ERM is enabled and statistical consistency (Mohri et al., 2018) is guaranteed, i.e., the risk of the learned classifier converges to the risk of the optimal classifier in the model, as the amount of training data approaches infinity.

However, in practice, with only finite training data, we found that the risk rewriting method suffers from severe overfitting, which is another challenge we are facing. For tackling overfitting, a common practice is to use regularization techniques to limit the flexibility of the model. Unfortunately, we also observed that these general-propose regularization techniques cannot mitigate this overfitting. Based on another observation that the empirical training risks can go negative during training, we conjecture that the overfitting is due to the negative terms included in the rewritten risk function: when unbounded loss functions, e.g., the cross-entropy loss, are used, the empirical training risk can even diverge to negative infinity. Therefore, it is necessary to fix this negative risk issue.

Recently, Ishida et al. (2020) proposed the *flooding* regularization to combat overfitting in supervised learning, which intentionally prevents further reduction of the empirical training risk when it reaches a reasonably small value called the *flood level*. We could naively apply the flooding regularization in our unlabeled setting and observed that overfitting can be mitigated to some extent. However, its effect was still on the same level as naive early stopping (Morgan and Bourlard, 1989).

To further improve the performance, we propose a fine-grained *partial risk regularization* inspired by the flooding method. Specifically, we decompose the rewritten risk into partial risks, i.e., the risks regarding all data in each U set being in same class, and then maintain each partial risk to a certain level as flooding does. We find that under some idealistic conditions, the optimal levels can be determined by the class priors, which serves as a guidance for selecting the hyper-parameters of these levels. Experimental results show that

our proposed partial risk regularization approach can successfully mitigate overfitting and outperform existing methods for learning from multiple U sets.

The rest of the paper is organized as follows. In Section 2, we formulate the problem of learning from multiple U sets, and in Section 3, we propose an ERM method and analyze its estimation error. The partial risk regularizer is proposed in Section 4, and experimental results are discussed in Section 5. Finally, conclusions are given in Section 6.

2. From Supervised Learning to Learning from Multiple Unlabeled Datasets

To begin with, let us consider a multi-class classification task with the input feature space $\mathcal{X} \subset \mathbb{R}^d$ and the output label space $\mathcal{Y} = \{1, 2, \ldots, K\} =: [K]$, where d is the input dimension and K is the number of classes. Let $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ be the input and output random variables following an underlying joint distribution with density $p(\boldsymbol{x}, y)$, which can be decomposed using the class priors $\{\pi_k = \Pr(y = k)\}_{k=1}^K$ and the class-conditional densities $\{p_k(\boldsymbol{x}) = p(\boldsymbol{x} \mid y = k)\}_{k=1}^K$ as $p(\boldsymbol{x}, y) = \sum_{k=1}^K \pi_k p_k(\boldsymbol{x})$. The goal of multi-class classification is to learn a classifier $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^K$ that minimizes the following *classification error*, also known as the *risk*:

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

where $\mathbb{E}$ denotes the expectation, and $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$ denotes a real-valued *loss function* that measures the discrepancy between the true label y and its prediction $\boldsymbol{g}(\boldsymbol{x})$. Typically, the predicted label is given by $\widehat{y} = \arg\max_{k \in [K]}(\boldsymbol{g}(\boldsymbol{x}))_k$, where $(\boldsymbol{g}(\boldsymbol{x}))_k$ is the k-th element of $\boldsymbol{g}(\boldsymbol{x})$. Note that when (1) is used for evaluation, ℓ is often chosen as the *zero-one loss* $\ell_{01}(\boldsymbol{g}(\boldsymbol{x}), y) = I(\widehat{y} \neq y)$, where I is the indicator function; when (1) is used for training, ℓ_{01} is replaced by a *surrogate loss* due to its difficulty for optimization, e.g., by the softmax cross-entropy loss.

As the density $p(\boldsymbol{x}, y)$ in (1) remains unknown, in traditional supervised learning, we normally require a vast amount of fully labeled training data $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ (where n is the training sample size and is assumed to be a sufficiently large number) drawn independently from $p(\boldsymbol{x}, y)$ to learn an accurate classifier (Wahba, 1990; Vapnik, 1998; Hastie et al., 2009; Sugiyama, 2015). With such supervised data, *empirical risk minimization* (ERM) (Vapnik, 1998) is a common practice that approximates (1) by replacing the expectation with the average over the training data: $\widehat{R}(\boldsymbol{g}) = \frac{1}{n} \sum_{i=1}^n [\ell(\boldsymbol{g}(\boldsymbol{x}_i), y_i)]$; then minimizing the empirical risk $\widehat{R}(\boldsymbol{g})$ over a class of functions, also known as the *model* $\mathcal{G}$: $\widehat{\boldsymbol{g}} = \arg\min_{\boldsymbol{g} \in \mathcal{G}} \widehat{R}(\boldsymbol{g})$.

However, in practice, collecting massive fully labeled data may be difficult due to the high labeling cost. In this paper, we consider a challenging setting of learning from multiple unlabeled datasets where *no* explicit labels are given. More specifically, assume that we have access to M sets of unlabeled samples $\mathcal{D}_U := \bigcup_{m=1}^M \mathcal{X}_m$, and each unlabeled set $\mathcal{X}_m = \{\boldsymbol{x}_{m,i}\}_{i=1}^{n_m}$ is a collection of n_m data points drawn from a mixture of class-conditional densities $\{p_k = p(\boldsymbol{x} \mid y = k)\}_{k=1}^K$:

$$\boldsymbol{x}_{m,i} \overset{\text{i.i.d.}}{\sim} q_m(\boldsymbol{x}) = \sum_{k=1}^K \theta_{m,k} p_k, \tag{2}$$

where $\theta_{m,k} \geq 0$ denotes the k-th class prior of the m-th unlabeled set. The class priors are assumed to form a full column rank matrix $\Theta := (\theta_{m,k}) \in \mathbb{R}^{M \times K}$ with the constraint that $\sum_{k=1}^{K} \theta_{m,k} = 1$. Through the paper, we assume that the class priors of each unlabeled set are given, including training class priors Θ and test class priors $\{\pi_k\}_{k=1}^{K}$.[1] Our goal is the same as standard supervised learning where we wish to learn a classifier $\boldsymbol{g}$ that generalizes well with respect to $p(\boldsymbol{x}, y)$, despite the fact that it is unobserved and we can only access unlabeled training sets $\{\mathcal{X}_m\}_{m=1}^{M}$.

3. Learning from Multiple Unlabeled Datasets

In this section, we derive an unbiased estimator of the classification risk in learning from multiple unlabeled datasets and analyze its estimation error.

3.1. Estimation of the Multi-Class Risk

Unbiased risk estimator based approaches have shown promising results in many weakly-supervised learning scenarios (du Plessis et al., 2014; Du Plessis et al., 2015; Ishida et al., 2017; Lu et al., 2019; Van Rooyen and Williamson, 2017), most of which mainly focused on binary classification problems. In what follows, we extend the results of unbiased risk estimation in their papers to multi-class classification from multiple unlabeled datasets.

The key technique for deriving unbiased risk estimators is risk rewriting, which enables the risk to be calculated only from observable distributions. In the context of our setting, risk rewriting means to evaluate the risk using $\{q_m\}_{m=1}^{M}$ instead of unknown $\{p_k\}_{k=1}^{K}$. Formally, we extend the definition of rewritability of the risk introduced in Lu et al. (2019) to our setting and show that $R(\boldsymbol{g})$ is rewritable as follows.

Definition 1 *We say that $R(\boldsymbol{g})$ is rewritable given $\{q_m\}_{m=1}^{M}$, if and only if there exist constants $\{w_{m,k}\}_{m \in [M], k \in [K]}$ such that for any model $\boldsymbol{g}$ it holds that*

$$R(\boldsymbol{g}) = \sum_{m=1}^{M} \mathbb{E}_{\boldsymbol{x} \sim q_m} \left[\bar{\ell}_m(\boldsymbol{g}(\boldsymbol{x})) \right], \tag{3}$$

where $\bar{\ell}_m(\boldsymbol{z}) = \sum_{k=1}^{K} w_{m,k} \ell(\boldsymbol{z}, k)$.

Theorem 2 *Assume that Θ has full column rank. Then $R(\boldsymbol{g})$ is rewritable by letting*

$$W = (\Pi \Theta^{\dagger})^{\top}, \tag{4}$$

where $W := (w_{m,k}) \in \mathbb{R}^{M \times K}$, $\Pi := \mathrm{diag}\{\pi_1, \pi_2, \ldots, \pi_K\}$, and $\dagger$ denotes the Moore–Penrose generalized inverse.

Based on Theorem 2, we can immediately obtain a rewritten risk function as

$$R_{\mathrm{U}}(\boldsymbol{g}) = \sum_{m=1}^{M} \sum_{k=1}^{K} w_{m,k} \mathbb{E}_{q_m} \left[\ell(\boldsymbol{g}(\boldsymbol{x}), k) \right], \tag{5}$$

---

1. Later in Section 5.3, we will experimentally show that the use of approximate class-priors is sufficient to obtain an accurate classifier.

and its unbiased risk estimator

$$\widehat{R}_{\mathrm{U}}(\boldsymbol{g}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{w_{m,k}}{n_m} \sum_{i=1}^{n_m} \ell(\boldsymbol{g}(\boldsymbol{x}_{m,i}), k), \tag{6}$$

where $w_{m,k}$'s are given by (4). Then, any off-the-shelf optimization algorithms, such as stochastic gradient descent (Robbins and Monro, 1951) and Adam (Kingma and Ba, 2015), can be applied for minimizing the empirical risk given by (6).

3.2. Theoretical Analysis

Here, we establish an estimation error bound for the unbiased risk estimator.

Theorem 3 *Let $\mathfrak{R}_{n_m}(\mathcal{H}_k)$ be the Rademacher complexity of $\mathcal{H}_k$ over $q_m(\boldsymbol{x})$ (Mohri et al., 2018), i.e., $\mathfrak{R}_{n_m}(\mathcal{H}_k) = \mathbb{E}_{q_m(\boldsymbol{x})} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}_k} \frac{1}{n_m} \sum_{i=1}^{n_m} \sigma_i h(\boldsymbol{x}_i) \right]$, where σ_i's are independent uniform random variables taking values in $\{+1, -1\}$ and $\mathcal{H}_k = \{h : \boldsymbol{x} \mapsto (\boldsymbol{g}(\boldsymbol{x}))_k | \boldsymbol{g} \in \mathcal{G}\}$. Assume the loss function $\ell(\boldsymbol{g}(\boldsymbol{x}), y)$ is L-Lipschitz with respect to $\boldsymbol{g}(\boldsymbol{x})$ for all $y \in \mathcal{Y}$ and upper-bounded by C_ℓ, i.e., $C_\ell = \sup_{\boldsymbol{g} \in \mathcal{G}, \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}} \ell(\boldsymbol{g}(\boldsymbol{x}), y)$. $\widehat{\boldsymbol{g}}_{\mathrm{U}} := \arg\min_{\boldsymbol{g} \in \mathcal{G}} \widehat{R}_U(\boldsymbol{g})$ is the risk of the learned classifier and $\boldsymbol{g}^* := \arg\min_{\boldsymbol{g} \in \mathcal{G}} R(\boldsymbol{g})$ is the risk of the optimal classifier in the given model class. Then, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$R(\widehat{\boldsymbol{g}}_{\mathrm{U}}) - R(\boldsymbol{g}^*) \le 4\sqrt{2} L K C_w \sum_{m=1}^{M} \sum_{k=1}^{K} \mathfrak{R}_{n_m}(\mathcal{H}_k) + 2 C_w K C_\ell \sqrt{\frac{M \ln \frac{2}{\delta}}{2 N_{\min}}}, \tag{7}$$

where $C_w = \max_{m \in [M], k \in [K]} |w_{m,k}|$ and $N_{\min} = \min_{m \in [M]} n_m$.

The proof of Theorem 3 is provided in Appendix B. Generally, $\mathfrak{R}_{n_m}(\mathcal{H}_k)$ can be bounded by $C_{\mathcal{G}}/\sqrt{N_{\min}}$ for a positive constant $C_{\mathcal{G}}$ (Golowich et al., 2018; Xia et al., 2019). The Theorem 3 shows that as $N_{\min} \to \infty$, $\widehat{\boldsymbol{g}}_{\mathrm{U}}$ converges to $\boldsymbol{g}^*$.

4. Partial Risk Regularization

In this section, we empirically analyze the unbiased risk estimator in a finite-sample case, and then propose partial risk regularization for further boosting its performance.

4.1. Overfitting of the Unbiased Risk Estimator

So far, we have obtained an unbiased risk estimator for learning from unlabeled datasets and derived an estimation error bound which guarantees the consistency of learning. However, we find that in practice, with only finite training data, the unbiased method tends to overfit especially when flexible models are used (see Figure 1 (a)).

For reducing overfitting, a common method is to use regularization techniques to limit the flexibility of the model. Popular choices of general-propose regularization techniques are dropout and weight decay (Goodfellow et al., 2016). In Figure 1,[2] (b) and (c) show

---

2. The class priors $\boldsymbol{\Theta}$ were set to be an asymmetric diagonal dominated square matrix, see the detailed experimental setup in Section 5.2.

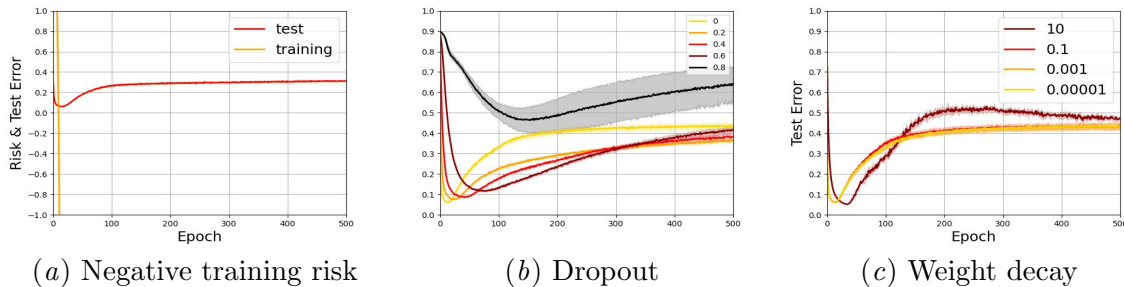(a) Negative training risk   (b) Dropout   (c) Weight decay

Figure 1: The negative training risk and overfitting problem on test data are shown in (a). A 5-layer multi-layer perception (MLP) was trained by Adam (Kingma and Ba, 2015) using the cross-entropy loss as a surrogate loss function. And the test error was calculated by zero-one loss. The results of general-propose regularization techniques dropout and weight decay are presented in (b) and (c), respectively. Each method was trained for 500 epochs.

the results of applying them on MNIST. For the dropout experiment, we fixed the weight decay parameter to be 1×10^{-5} and changed the dropout probability from 0 to 0.8. For the weight decay experiment, we fixed the dropout probability to be 0 and changed the weight decay parameter from 1×10^{-5} to 1×10^{1}. The results show that general-propose regularization techniques cannot mitigate the overfitting of the unbiased risk estimation method in learning from multiple unlabeled datasets.

To further analyze this phenomenon, [3] let us take a close look at the unbiased risk estimator (6). We find that the coefficients $w_{m,k}$'s given by Theorem 2 are not necessarily non-negative, some of which are exactly negative numbers. These negative terms may lead the empirical training risk to go negative during training (see Figure 1 (a)). When overparameterized models, e.g., deep neural networks, and unbounded loss functions, e.g., the cross-entropy loss, are used, the empirical training risk may even diverge to negative infinity; the model can memorize all training data and become overly confident. This may be the reasons why severe overfitting occurs here and why general-propose regularizations cannot solve it.

4.2. Partial Risk Regularization

Based on the analysis in Section 4.1, the issue that the empirical training risk goes negative and may even diverge should be fixed. Flooding (Ishida et al., 2020) is a regularization method dedicated to maintaining a fixed-level empirical training risk, which has been shown to mitigate overfitting effectively in the supervised setting. This method seems promising to solve our problem, however, it is based on supervised learning and needs full labels to calculate the risk. Thanks to the unbiased risk estimator derived in Section 3.1, we can straightforwardly apply the flooding method to it as follows:

$$R_{\text{U-flood}}(\boldsymbol{g}) = |R_{\text{U}}(\boldsymbol{g}) - b| + b, \tag{8}$$

---

3. This phenomenon is also observed in other weakly supervised learning settings using the unbiased risk estimator method, for example, in learning from positive and unlabeled data (Kiryo et al., 2017) and learning from two unlabeled datasets (Lu et al., 2020).

where $b \geq 0$ is a hyperparameter called the flood level. The experimental results in Section 5 show this simple method relieves overfitting to some extent, but only performs on the same level of early stopping.

To further boost performance, we decompose the total risk $R_U(\boldsymbol{g})$ into partial risks:

$$R_{q_m,k}(\boldsymbol{g}) := \mathbb{E}_{q_m} \left[\ell(\boldsymbol{g}(\boldsymbol{x}), k) \right], \tag{9}$$

that is, the risk of considering all data in one unlabeled dataset to belong to the same class. Then we propose a fine-grained *partial risk regularization* (PRR) approach that maintains each partial risk in the rewritten risk function (5) to a certain level:

$$R_{\text{U-PRR}}(\boldsymbol{g}) = \alpha R_U(\boldsymbol{g}) + (1 - \alpha) R_{\text{reg}}(\boldsymbol{g}), \tag{10}$$

where

$$R_{\text{reg}}(\boldsymbol{g}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \lambda_{m,k} \left| R_{q_m,k}(\boldsymbol{g}) - b_{m,k} \right|, \tag{11}$$

$0 \leq \alpha \leq 1$ and $\lambda_{m,k}$'s are trade-off parameters, and $b_{m,k}$'s are the flood levels for the corresponding partial risks. To select the flood levels, we provide the following lemma which studies the relationship between them and the Bayes optimal classifier $\boldsymbol{g}^\star$ where $R(\boldsymbol{g}^\star) = \inf_{\boldsymbol{g}} R(\boldsymbol{g})$.

Lemma 4 *Let $R_{p,k}(\boldsymbol{g}) = \mathbb{E}_p \left[\ell(\boldsymbol{g}(\boldsymbol{x}), k) \right]$ where p is any probability density function, $k = 1, \ldots, K$, and $m = 1, \ldots, M$. We have $R_{q_m,k}(\boldsymbol{g}^\star) = \sum_{i=1}^{K} \theta_{m,i} R_{p_i,k}(\boldsymbol{g}^\star)$. When all classes are separable, we have $R_{q_m,k}^{01}(\boldsymbol{g}^\star) = 1 - \theta_{m,k}$, where $R_{q_m,k}^{01}(\boldsymbol{g})$ denotes $R_{q_m,k}(\boldsymbol{g})$ calculated by the zero-one loss.*

Guided by Lemma 4, in order to push $\boldsymbol{g}$ towards the optimal classifier $\boldsymbol{g}^\star$, we propose to choose $b_{m,k} = 1 - \theta_{m,k}$ as the flood level of $R_{q_m,k}^{01}(\boldsymbol{g})$, and replace $R_{q_m,k}(\boldsymbol{g})$ by $R_{q_m,k}^{01}(\boldsymbol{g})$ in equation (11). Therefore, the empirical learning objective of PRR approach is

$$\widehat{R}_{\text{U-PRR}}(\boldsymbol{g}) = \alpha \widehat{R}_U(\boldsymbol{g}) + (1 - \alpha) \widehat{R}_{\text{reg}}(\boldsymbol{g}), \tag{12}$$

where

$$\widehat{R}_{\text{reg}}(\boldsymbol{g}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \lambda_{m,k} \left| \widehat{R}_{q_m,k}^{01}(\boldsymbol{g}) - (1 - \theta_{m,k}) \right|, \tag{13}$$

and

$$\widehat{R}_{q_m,k}^{01}(\boldsymbol{g}) = \frac{1}{n_m} \sum_{i=1}^{n_m} \ell_{01}(\boldsymbol{g}(\boldsymbol{x}_{m,i}), k). \tag{14}$$

In the implementation (see Algorithm 1), we use the zero-one loss for determining the sign of the term in the absolute function and use a surrogate loss in calculating the gradients due to the difficulty of optimizing the zero-one loss. For dealing with absolute functions, when the term in it is greater than 0, we perform gradient descent as usual; when the term is less than 0, we perform gradient ascent instead and add a hyperparameter s_{GA} in the gradient ascent process.

In addition, we have to determine hyperparameters $\lambda_{m,k}$ in the regularization part. Here, we set them as $\lambda_{m,k} = |w_{m,k}|$ since choosing such a large number of hyperparameters separately is not straightforward. This choice means that the term that receives more attention in the unbiased risk estimator also receives more attention in partial risk regularization.

Algorithm 1 Partial risk regularization approach for learning from multiple U datasets

---

Input: M sets of U training data $\mathcal{D}_{\mathrm{U}} = \bigcup_{m=1}^{M} \mathcal{X}_m$ with known class prior Θ
Output: model parameter γ for classifier $\boldsymbol{g}$

1: Initialize γ
2: Let $\mathcal{A}$ be an SGD-like optimizer working on γ
3: **for** $t = 0$; $t <$ number of epochs; $t++$ **do**
4:    Shuffle $\mathcal{D}_{\mathrm{U}}$
5:    **for** $i = 0$; $i <$ number of mini-batches; $i++$ **do**
6:       Let $\overline{\mathcal{D}}_{\mathrm{U}} = \bigcup_{m=1}^{M} \overline{\mathcal{X}}_m$ be the current mini-batch
7:       Compute $\widehat{R}_{\mathrm{U}}(\boldsymbol{g})$ followed Eq. (6)
8:       Compute $\widehat{R}_{\mathrm{reg}}(\boldsymbol{g})$:
9:       **if** $\mathbb{E}_{\overline{\mathcal{X}}_m} [\ell_{01}(\boldsymbol{g}(\boldsymbol{x}), k)] \geq (1 - \theta_{m,k})$ **then**
10:          $\left| \mathbb{E}_{\overline{\mathcal{X}}_m} [\ell(\boldsymbol{g}(\boldsymbol{x}), k)] - (1 - \theta_{m,k}) \right| \leftarrow \mathbb{E}_{\overline{\mathcal{X}}_m} [\ell(\boldsymbol{g}(\boldsymbol{x}), k)] - (1 - \theta_{m,k})$
11:       **else**
12:          $\left| \mathbb{E}_{\overline{\mathcal{X}}_m} [\ell(\boldsymbol{g}(\boldsymbol{x}), k)] - (1 - \theta_{m,k}) \right| \leftarrow -s_{\mathrm{GA}}(\mathbb{E}_{\overline{\mathcal{X}}_m} [\ell(\boldsymbol{g}(\boldsymbol{x}), k)] - (1 - \theta_{m,k}))$
13:       **end if**
14:       Compute $\widehat{R}_{\mathrm{U\text{-}PRR}}(\boldsymbol{g})$ followed Eq. (12)
15:       Compute gradient $\bigtriangledown_\gamma \widehat{R}_{\mathrm{U\text{-}PRR}}(\boldsymbol{g})$
16:       Update γ by $\mathcal{A}$
17:    **end for**
18: **end for**

Table 1: Specification of datasets and corresponding models.

Dataset	Feature	Class	Training Data	Test Data	Model
Pendigits	16	10	7,494	3,498	FC with ReLU (depth 3)
USPS	16*16	10	7,291	2,007	FC with ReLU (depth 3)
MNIST	28*28	10	60,000	10,000	FC with ReLU (depth 5)
Fashion-MNIST	28*28	10	60,000	10,000	FC with ReLU (depth 5)
Kuzushiji-MNIST	28*28	10	60,000	10,000	FC with ReLU (depth 5)
CIFAR-10	32*32*3	10	50,000	10,000	ResNet (depth 20)

5. Experiments

In this section, we verify the effectiveness of the proposed method[4] on various datasets. We also show the robustness against noisy class priors of the proposed method.

5.1. Experimental Setup

We describe the details of the experimental setup as follows.

Datasets   We used widely adopted benchmarks: Pendigits (Alimoglu and Alpaydin, 1996), USPS (Hull, 1994), MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), Kuzushiji-MNIST (Clanuwat et al., 2018) and, CIFAR-10 (Krizhevsky et al., 2009). Table 1 briefly summaries the benchmark datasets.

The m-th U set contains n_m samples where its proportions $\theta_{m,k}$ were randomly chosen from class k for $k = 1, \ldots, K$. The size of each unlabeled dataset n_m is set to $1/M$ of the total number of training datasets in all these settings.

---

4. Our implementation is available at https://github.com/Tang-Yuting/U-PRR.

Baselines To better analyze the performance of the proposed unbiased risk estimator (6) (*Unbiased*) and partial risk regularization approach (12) (*U-PRR*), we compared them with following baseline methods:

- Biased proportion (*Biased*): We consider the label of the largest category in each unlabeled dataset as the label of all samples in it, and perform supervised learning from such pseudo-label samples.

- Proportion loss (*Prop/Prop-CR*): *Prop* uses class priors of each unlabeled dataset as weak supervision and aims to minimize the difference between true class priors and the predicted class priors (Yu et al., 2015). *Prop-CR* builds upon the proportion loss baseline by adding a consistency regularization term (Tsai and Lin, 2020).

- Corrected unbiased risk estimator (*U-correct*): We apply a corrected risk estimation method (Lu et al., 2020), which aims to correct partial risks corresponding to each class to be non-negative, to the unbiased risk estimator. The learning objective is $\widehat{R}_{\text{U-correct}}(\boldsymbol{g}) = \sum_{k=1}^{K} \left| \sum_{m=1}^{M} w_{m,k} \frac{1}{n_m} \sum_{i=1}^{n_m} \ell(\boldsymbol{g}(\boldsymbol{x}_{m,i}), k) \right|$.

- Unbiased risk estimator with early stopping (*U-stop*): We apply the early stopping on the proposed unbiased risk estimator, and stop the training when the empirical risk goes negative since Lu et al. (2020) showed a high co-occurrence between overfitting and negative empirical risk.

- Unbiased risk estimator with flooding (*U-flood*) (Ishida et al., 2020): We directly apply the flooding method to the unbiased risk estimator. The learning objective is $\widehat{R}(\boldsymbol{g}) = |\widehat{R}_{\text{U}}(\boldsymbol{g}) - b| + b$. We choose b from $\{0, 0.05, 0.1\}$ and report the best test error.

Training Details The models are described in Table 1, where FC refers to the fully connected neural network and ResNet refers to the residual network (He et al., 2016). As a common practice, Adam (Kingma and Ba, 2015) with the cross-entropy loss was used for optimization. We ran each experiment five times, and we trained the model for 500 epochs on all datasets except that CIFAR-10 only used 200 epochs since 200 epochs were sufficient for convergence. The classification error rate was used for evaluating the test performance. The learning rates were chosen from $\{5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$ and the batch sizes were chosen from $\{\frac{n}{500}, \frac{n}{200}, \frac{n}{100}, \frac{n}{50}, \frac{n}{20}, \frac{n}{10}\}$, where n is the size of the training dataset. In the PRR method, α's were chosen from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and s_{GA}'s were chosen from $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. Note that for fair comparison, we used the same model and the same hyperparameters for the implementation of all methods in each benchmark. The implementation in our experiments was based on PyTorch[5], and the experiments were conducted on an NVIDIA Tesla P100 GPU.

5.2. Comparison with Baseline Methods

Here, we report the final test error (Err) and the test error drop (Δ_E), which is the difference between the smallest test error among all training epochs and the test error at the end of training, for the baseline methods and proposed methods. We designed different Θ for evaluating the proposed methods in various scenarios:

---

5. See https://pytorch.org.

Table 2: Experimental results with a symmetric class-prior matrix. Means (standard deviations) of the classification error (Err) and the drop (Δ_E) over five trials in percentage. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Dataset	a,b	Biased Err	Biased Δ_E	Prop Err	Prop Δ_E	Prop-CR Err	Prop-CR Δ_E	Unbiased Err	Unbiased Δ_E	U-correct Err	U-correct Δ_E	U-stop Err	U-stop Δ_E	U-flood Err	U-flood Δ_E	U-PRR Err	U-PRR Δ_E
Pendigits	0.5, 0.05	11.54 (1.19)	7.42 (1.05)	**5.80** (**1.08**)	0.23 (0.19)	5.85 (1.01)	0.31 (0.22)	17.99 (1.88)	13.09 (1.67)	**5.58** (**0.48**)	1.25 (0.38)	7.80 (1.13)	0 (0)	**5.29** (**0.99**)	1.18 (0.96)	**4.61** (**0.77**)	0.26 (0.30)
	0.1, 0.09	61.30 (1.50)	41.27 (2.16)	37.97 (1.30)	22.44 (0.54)	36.20 (1.98)	20.90 (1.62)	47.16 (2.24)	19.45 (2.77)	36.25 (2.92)	16.28 (1.68)	20.12 (1.75)	0 (0)	29.44 (3.34)	11.12 (3.96)	**12.64** (**2.48**)	0.59 (0.39)
USPS	0.5, 0.05	23.23 (1.30)	14.73 (1.41)	7.92 (0.15)	0.62 (0.27)	6.51 (0.24)	0.33 (0.20)	26.98 (2.51)	18.64 (2.35)	11.14 (0.25)	2.11 (0.58)	9.39 (0.51)	0 (0)	11.41 (0.65)	2.53 (0.68)	**6.01** (**0.25**)	0.30 (0.11)
	0.1, 0.09	70.42 (0.60)	37.39 (1.95)	66.97 (3.14)	49.25 (3.28)	56.52 (3.00)	44.26 (3.42)	75.86 (1.50)	57.28 (1.44)	69.97 (2.08)	35.17 (3.94)	22.32 (2.09)	0 (0)	56.09 (3.42)	31.83 (3.16)	**11.15** (**1.20**)	0.64 (0.12)
MNIST	0.5, 0.05	25.19 (0.70)	19.02 (0.70)	5.56 (0.11)	0.09 (0.06)	3.86 (0.11)	0.08 (0.06)	31.14 (0.49)	24.93 (0.65)	5.59 (0.29)	0.24 (0.20)	6.67 (0.25)	0 (0)	6.34 (0.33)	0.42 (0.22)	**2.99** (**0.10**)	0.19 (0.07)
	0.1, 0.09	78.19 (0.37)	35.31 (0.91)	66.21 (1.60)	50.56 (1.38)	43.35 (13.76)	40.37 (13.77)	82.51 (0.52)	63.83 (0.52)	53.16 (3.41)	24.39 (3.31)	20.95 (1.31)	0 (0)	27.72 (1.74)	6.34 (2.02)	**5.84** (**0.19**)	0.26 (0.09)
Fashion-MNIST	0.5, 0.05	32.57 (0.81)	17.79 (0.85)	14.75 (0.36)	0.29 (0.33)	15.25 (0.19)	0.26 (0.05)	36.03 (0.73)	21.36 (0.79)	15.14 (0.36)	0.75 (0.33)	14.88 (0.31)	0.03 (0.04)	15.95 (0.56)	0.92 (0.64)	**13.09** (**0.17**)	0.50 (0.12)
	0.1, 0.09	79.28 (0.54)	41.39 (0.88)	56.34 (1.15)	35.16 (1.37)	34.81 (1.60)	9.76 (1.94)	82.85 (0.64)	60.68 (0.62)	39.45 (2.95)	14.07 (2.87)	23.39 (0.72)	0 (0)	26.43 (1.67)	2.74 (1.25)	**21.91** (**0.90**)	2.78 (0.76)
Kuzushiji-MNIST	0.5, 0.05	39.35 (0.49)	15.84 (0.55)	20.07 (0.53)	0.41 (0.22)	17.30 (0.91)	0.17 (0.17)	43.39 (0.70)	21.38 (0.93)	20.05 (0.40)	0.96 (0.33)	22.94 (0.60)	0 (0)	21.04 (1.04)	0.99 (0.74)	**14.33** (**0.14**)	0.88 (0.13)
	0.1, 0.09	81.26 (0.58)	17.19 (1.15)	73.10 (2.54)	28.41 (2.15)	39.33 (3.30)	11.72 (3.28)	84.57 (0.46)	39.26 (0.51)	61.77 (2.33)	8.64 (1.48)	46.57 (1.13)	0 (0)	47.23 (2.38)	5.28 (2.92)	**31.17** (**0.70**)	3.09 (0.89)
CIFAR-10	0.5, 0.05	51.41 (3.10)	7.65 (2.95)	65.24 (1.87)	0.20 (0.19)	67.76 (1.68)	0.10 (0.21)	57.44 (3.20)	9.39 (2.18)	57.72 (5.93)	8.40 (6.77)	51.94 (2.64)	3.14 (2.39)	56.49 (4.19)	8.06 (3.16)	**44.75** (**1.85**)	1.37 (0.86)
	0.1, 0.09	84.67 (0.25)	16.66 (1.68)	72.73 (2.83)	0.41 (0.37)	74.22 (1.32)	0.17 (0.23)	86.82 (1.71)	19.41 (1.39)	76.46 (2.19)	8.96 (1.17)	**69.99** (**1.25**)	1.99 (1.76)	71.08 (2.15)	3.71 (2.15)	**69.24** (**1.55**)	3.14 (1.46)

- Symmetric square class-prior matrix. It is defined as

$$\theta_{m,k} = \begin{cases} a + b & m = k, \\ b & m \neq k, \end{cases}$$

  where $a > 0$ and $b > 0$ are constants satisfying $a + Kb = 1$. We tested our proposed methods under two different training class-prior settings: (a, b) was chosen as $(0.5, 0.05)$ and $(0.1, 0.09)$. The experimental results are reported in Table 2.

- Asymmetric diagonal-dominated square class-prior matrix. In this setting, the values on the diagonals of the matrix Θ are larger than the other values, and they can be different from each other. We randomly and uniformly generated the class priors $\theta_{m,k}$ from range $[0, 1/M]$ when $m \neq k$, and set $\theta_{k,k} = 1 - \sum_{m \neq k} \theta_{m,k}$. The experimental results are shown in Table 3.

- Non-square class-prior matrix. We also conduct experiments when the number of unlabeled datasets M is more than the number of classes K. Here, we set $M = 2K$. The class-prior matrix Θ is a concatenation of two asymmetric diagonal-dominated square matrices. Compared with the previous experiments, this setting is more in line with real-world situations since it is rare that the number of unlabeled datasets is exactly equal to the number of classes. The experimental results are reported in Table 4.

The experimental results show: the unbiased risk estimator suffers severe overfitting and this issue is significantly alleviated by correction methods (U-correct, U-stop, U-flood

Table 3: Experimental results on diagonal-dominated square class-prior matrix. Means (standard deviations) of the classification error (Err) and the drop (Δ_E) over five trials in percentage. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Dataset	Biased		Prop		Prop-CR		Unbiased		U-correct		U-stop		U-flood		U-PRR	
	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E
Pendigits	15.19 (5.35)	9.56 (3.94)	**6.75** (**0.59**)	0.35 (0.25)	**6.58** (**0.48**)	0.45 (0.26)	27.06 (7.82)	20.17 (6.06)	9.75 (3.87)	1.88 (1.51)	10.29 (2.89)	0.01 (0.01)	7.92 (1.75)	2.04 (0.79)	**5.57** (**1.78**)	0.86 (0.41)
USPS	27.85 (8.12)	17.18 (5.68)	**10.90** (**3.45**)	2.18 (2.41)	**8.99** (**2.64**)	1.13 (1.60)	46.06 (14.14)	35.98 (12.47)	26.14 (7.42)	11.73 (5.41)	11.27 (1.86)	0 (0)	14.70 (4.25)	3.71 (2.03)	**7.31** (**1.02**)	0.48 (0.29)
MNIST	28.42 (0.60)	21.31 (0.68)	5.70 (0.17)	0.11 (0.09)	4.02 (0.08)	0.05 (0.05)	43.36 (2.01)	37.12 (1.90)	5.96 (0.44)	0.37 (0.27)	6.62 (0.24)	0 (0)	6.10 (0.30)	0.24 (0.21)	**3.65** (**0.08**)	0.38 (0.08)
Fashion-MNIST	35.27 (0.23)	19.93 (0.36)	**14.75** (**0.20**)	0.21 (0.21)	14.96 (0.12)	0.09 (0.05)	45.09 (0.89)	30.08 (0.99)	15.35 (0.41)	0.76 (0.22)	15.10 (0.18)	0.09 (0.08)	15.60 (0.11)	0.72 (0.14)	**14.25** (**0.19**)	1.21 (0.29)
Kuzushiji-MNIST	38.49 (0.63)	14.88 (0.49)	20.77 (0.60)	0.68 (0.44)	**17.38** (**0.64**)	0.24 (0.15)	52.17 (0.97)	29.19 (0.86)	20.18 (0.33)	1.07 (0.40)	23.89 (1.24)	0 (0)	20.50 (0.47)	0.59 (0.52)	**17.15** (**0.71**)	1.28 (0.45)
CIFAR-10	50.05 (1.64)	6.38 (1.13)	64.96 (2.09)	0.14 (0.15)	68.67 (2.44)	0.34 (0.07)	61.45 (7.26)	10.28 (7.21)	56.04 (2.42)	6.07 (1.96)	53.32 (0.95)	3.13 (0.70)	57.49 (1.91)	7.33 (1.78)	**45.02** (**1.85**)	1.92 (0.87)

Table 4: Experimental results on non-square class-prior matrix. Means (standard deviations) of the classification error (Err) and the drop (Δ_E) over five trials in percentage. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Dataset	Biased		Prop		Prop-CR		Unbiased		U-correct		U-stop		U-flood		U-PRR	
	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E	Err	Δ_E
Pendigits	14.03 (3.69)	9.65 (2.79)	**5.04** (**0.85**)	0.46 (0.18)	**5.05** (**1.00**)	0.46 (0.25)	23.37 (6.59)	17.68 (5.58)	6.51 (1.27)	1.46 (0.63)	9.51 (2.44)	0 (0)	6.90 (1.81)	2.04 (1.49)	**4.75** (**1.22**)	0.68 (0.77)
USPS	30.05 (4.94)	19.59 (3.71)	13.38 (0.59)	4.47 (0.50)	**8.45** (**0.29**)	1.27 (0.26)	45.04 (2.00)	34.82 (2.31)	21.02 (1.15)	6.95 (1.51)	10.89 (0.65)	0.03 (0.04)	13.38 (0.74)	2.59 (0.88)	**8.00** (**0.30**)	0.82 (0.22)
MNIST	36.15 (0.43)	26.56 (0.41)	7.00 (0.40)	0.69 (0.39)	5.60 (0.14)	0.29 (0.08)	51.55 (1.45)	43.82 (1.57)	9.80 (0.68)	1.19 (0.78)	8.03 (0.30)	0 (0)	6.31 (0.56)	0.60 (0.44)	**3.32** (**0.13**)	0.22 (0.09)
Fashion-MNIST	37.93 (0.90)	20.60 (0.78)	**16.01** (**0.21**)	0.55 (0.28)	**15.27** (**0.39**)	0.23 (0.15)	54.78 (2.69)	38.31 (2.37)	19.51 (0.47)	2.44 (0.53)	16.64 (0.25)	0.02 (0.04)	17.21 (0.85)	0.95 (0.88)	**16.43** (**0.86**)	2.01 (0.97)
Kuzushiji-MNIST	46.61 (0.98)	19.04 (0.90)	25.87 (0.68)	0.49 (0.81)	**20.28** (**0.58**)	0.84 (0.53)	57.52 (1.49)	30.68 (1.78)	27.39 (0.44)	2.21 (0.49)	27.53 (0.74)	0 (0)	23.33 (0.91)	1.12 (0.70)	26.40 (0.74)	2.95 (1.15)
CIFAR-10	55.84 (2.04)	7.47 (2.04)	63.45 (2.02)	0.21 (0.17)	67.58 (1.74)	0.10 (0.17)	71.46 (11.81)	15.30 (11.97)	63.98 (2.55)	9.82 (2.60)	59.53 (6.53)	7.05 (5.84)	59.90 (1.40)	7.32 (0.45)	**47.18** (**2.20**)	1.95 (0.86)

and U-PRR) across various scenarios and datasets; the performance of directly applying flooding on the unbiased risk estimator can only reach the same level as early stopping; the partial risk regularization method outperformed all baseline methods and achieved the best performance for all the datasets and class prior settings.

5.3. Robustness against Noisy Class Priors

Hitherto, we have assumed that the values of the class priors were accurately accessible, which may not be true in practice. We designed experiments where random noise was added to each class prior in order to simulate real-world situations. In the experiment, we added different levels (1%, 3%, and 5%) of perturbation to each element of Θ to obtain Θ', and the noisy Θ' was treated as the true Θ during the whole learning process.

The results on the MNIST dataset with different class priors are shown in Fig. 2, where 0% noise rate means that we used the true class priors. This figure shows that the proposed

(a) Symmetric(a=0.5) (b) Symmetric(a=0.1) (c) Asymmetric (d) Non-square
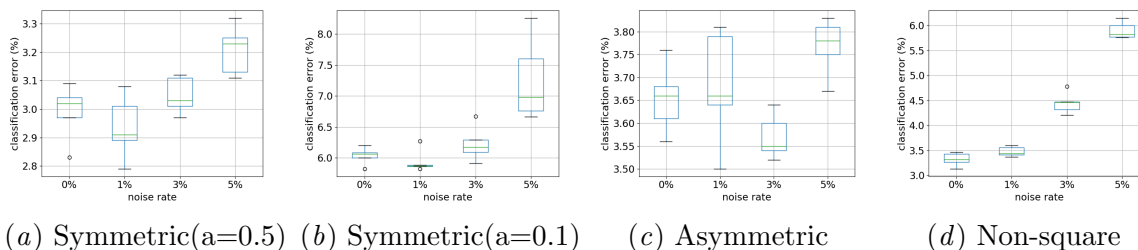
Figure 2: Experiments of the robustness of the proposed method to different (1%, 3%, and 5%) noise rate. The class-prior matrices in experiments are symmetric matrix with $a = 0.5$ and $a = 0.1$, asymmetric diagonal dominated matrix, and non-square matrix. Means (standard deviations) of the classification error over five trials in percentage.

method is not strongly affected by noise in the class priors and thus it may be safely applied in the real world.

6. Conclusion

In this work, we focused on the problem of multi-class classification from multiple unlabeled datasets. We established an unbiased risk estimator and provided a generalization error bound for it. Based on our empirical observations, the negative empirical training risk was a potential reason why the unbiased risk estimation suffers severe overfitting. To overcome overfitting, we proposed a partial risk regularization approach customized for learning from multiple unlabeled datasets. Experiments demonstrated the superiority of our proposed partial risk regularization approach.

Acknowledgments

References

Fevzi Alimoglu and Ethem Alpaydin. Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition. In *Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN) 96*. Citeseer, 1996.

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

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. Predicting the demographics of twitter users from website traffic data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394. PMLR, 2015.

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in neural information processing systems*, 27: 703–711, 2014.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

J.J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. *arXiv preprint arXiv:1710.07138*, 2017.

Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Do we need zero training loss after achieving zero training error? In *International Conference on Machine Learning*, pages 4604–4614. PMLR, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *Advances in neural information processing systems*, 2017.

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

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

Jiabin Liu, Bo Wang, Zhiquan Qi, YingJie Tian, and Yong Shi. Learning from label proportions with generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Nan Lu, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Nan Lu, Tianyi Zhang, Gang Niu, and Masashi Sugiyama. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1115–1125. PMLR, 2020.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Nelson Morgan and Hervé Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. *Advances in neural information processing systems*, 2, 1989.

Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(10), 2009.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Masashi Sugiyama. *Introduction to statistical machine learning*. Morgan Kaufmann, 2015.

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

Kuen-Han Tsai and Hsuan-Tien Lin. Learning from label proportions with consistency regularization. In Sinno Jialin Pan and Masashi Sugiyama, editors, *Proceedings of The 12th Asian Conference on Machine Learning*, volume 129 of *Proceedings of Machine Learning Research*, pages 513–528. PMLR, 18–20 Nov 2020.

Brendan Van Rooyen and Robert C Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(1):8501–8550, 2017.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Grace Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, and Shih-Fu Chang. \proptosvm for learning with label proportions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 504–512. PMLR, 2013.

Felix X. Yu, Krzysztof Choromanski, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. On learning from label proportions, 2015.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

Appendix A. Proof of Theorem 2

Notice that

$$R(\boldsymbol{g}) = \sum_{k=1}^{K} \pi_k \mathbb{E}_{p_k}[\ell(\boldsymbol{g}(\boldsymbol{x}), k)] = \sum_{i=1}^{K} \sum_{j=1}^{K} (\Pi)_{i,j} \mathbb{E}_{p_i}[\ell(\boldsymbol{g}(\boldsymbol{x}), j)]. \tag{15}$$

If the classification risk is rewritable, i.e.,

$$R(\boldsymbol{g}) = \sum_{m=1}^{M} \mathbb{E}_{q_m}\left[\sum_{k=1}^{K} w_{m,k}\ell(\boldsymbol{g}(\boldsymbol{x}), k)\right] = \sum_{m=1}^{M} \sum_{k'=1}^{K} \theta_{m,k'} \mathbb{E}_{p_{k'}}\left[\sum_{k=1}^{K} w_{m,k}\ell(\boldsymbol{g}(\boldsymbol{x}), k)\right]$$

$$= \sum_{i=1}^{K} \sum_{j=1}^{K} (W^\top \Theta)_{i,j} \mathbb{E}_{p_i}[\ell(\boldsymbol{g}(\boldsymbol{x}), j)],$$

then comparing the coefficients with (15), we have $W^\top \Theta = \Pi$. Therefore, $R(\boldsymbol{g})$ is rewritable by letting $W = (\Pi\Theta^\dagger)^\top$.

Appendix B. Proof of Theorem 3

Our proof of the estimation error bound is based on *Rademacher complexity* (Bartlett and Mendelson, 2002). First, we show the uniform deviation bound, which is useful to derive the estimation error bound.

Lemma 5 *For any $\delta > 0$, we have the probability at least $1 - \delta$,*

$$\sup_{\boldsymbol{g}\in\mathcal{G}} \left|\widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g})\right| \le 2\sqrt{2}LKC_w \sum_{m=1}^{M} \sum_{y=1}^{K} \mathfrak{R}_{n_m}(\mathcal{H}_y) + 2C_w KC_\ell \sqrt{\frac{M\ln\frac{2}{\delta}}{2N_{min}}}. \tag{16}$$

where the probability is over repeated sampling of data for evaluating $\widehat{R}_U(\boldsymbol{g})$.

Proof Consider the one-side uniform deviation $\sup_{\boldsymbol{g}\in\mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g})$. The change of it will be no more than $\frac{C_w KC_\ell}{n_m}$ if some $\boldsymbol{x} \in \mathcal{X}_m$ $(1 \le m \le M)$ is replaced by $\boldsymbol{x}'$. Subsequently, *McDiarmid's inequality* (McDiarmid et al., 1989) tells us that

$$\Pr\left\{\sup_{\boldsymbol{g}\in\mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g}) - \mathbb{E}[\sup_{\boldsymbol{g}\in\mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g})] \ge \epsilon\right\} \le \exp(-\frac{2\epsilon^2}{(\sum_m \frac{1}{n_m})(C_w KC_\ell)^2}).$$

or equivalently, with probability at least $1 - \delta/2$,

$$\sup_{\boldsymbol{g} \in \mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g}) \leq \mathbb{E}[\sup_{\boldsymbol{g} \in \mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g})] + \sqrt{\frac{(\sum_m \frac{1}{n_m})(C_w K C_\ell)^2 \ln \frac{2}{\delta}}{2}}$$

$$\leq \mathbb{E}[\sup_{\boldsymbol{g} \in \mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g})] + C_w K C_\ell \sqrt{\frac{M \ln \frac{2}{\delta}}{2 N_{\min}}}.$$

By symmetrization (Vapnik, 1998) and using the same trick in Maurer (2016), it is a routine work to show that

$$\mathbb{E}[\sup_{\boldsymbol{g} \in \mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g})] \leq 2\sqrt{2}L \sum_{m=1}^{M} \sum_{k=1}^{K} |w_{m,k}| \sum_{k'=1}^{K} \mathfrak{R}_{n_m}(\mathcal{H}_{k'})$$

$$\leq 2\sqrt{2}LKC_w \sum_{m=1}^{M} \sum_{k=1}^{K} \mathfrak{R}_{n_m}(\mathcal{H}_k).$$

The one-side uniform deviation $R(\boldsymbol{g}) - \sup_{\boldsymbol{g} \in \mathcal{G}} \widehat{R}_{\mathrm{U}}(\boldsymbol{g})$ can be bounded similarly. ∎

Based on Lemma 5, the estimation error bound (7) is proven through

$$R(\widehat{\boldsymbol{g}}_{\mathrm{U}}) - R(\boldsymbol{g}^*) = \left(\widehat{R}_{\mathrm{U}}(\widehat{\boldsymbol{g}}_{\mathrm{U}}) - \widehat{R}_{\mathrm{U}}(\boldsymbol{g}^*) \right) + \left(R(\widehat{\boldsymbol{g}}_{\mathrm{U}}) - \widehat{R}_{\mathrm{U}}(\widehat{\boldsymbol{g}}_{\mathrm{U}}) \right) + \left(\widehat{R}_{\mathrm{U}}(\boldsymbol{g}^*) - R(\boldsymbol{g}^*) \right)$$

$$\leq 0 + 2 \sup_{\boldsymbol{g} \in \mathcal{G}} \left| \widehat{R}_{\mathrm{U}}(\boldsymbol{g}) - R(\boldsymbol{g}) \right|$$

$$\leq 4\sqrt{2}LKC_w \sum_{m=1}^{M} \sum_{k=1}^{K} \mathfrak{R}_{n_m}(\mathcal{H}_k) + 2C_w K C_\ell \sqrt{\frac{M \ln \frac{2}{\delta}}{2 N_{\min}}}.$$

where $\widehat{R}_{\mathrm{U}}(\widehat{\boldsymbol{g}}_{\mathrm{U}}) \leq \widehat{R}_{\mathrm{U}}(\boldsymbol{g}^*)$ by the definition of $\widehat{\boldsymbol{g}}_{\mathrm{U}}$.

Appendix C. Proof of Lemma 4

By the relationship of distributions p_k and q_m, we have

$$b_{m,k} = \sum_{i=1}^{K} \theta_{m,i} R_{p_i,k}^{01}(\boldsymbol{g}^\star) = R_{\sum_{i=1}^{K} \theta_{m,i} p_i, k}^{01}(\boldsymbol{g}^\star) = R_{q_m,k}^{01}(\boldsymbol{g}^\star).$$

When $\boldsymbol{g}^\star$ is the Bayes optimal classifier and all classes are separable, we have

$$R_{p_i,k}^{01}(\boldsymbol{g}^\star) = \begin{cases} 0 & \text{if } i = k, \\ 1 & \text{otherwise.} \end{cases}$$

Then

$$b_{m,k} = \sum_{i=1}^{K} \theta_{m,i} R_{p_i,k}^{01}(\boldsymbol{g}^\star) = \sum_{i \neq k} \theta_{m,i} = 1 - \theta_{m,k}$$

holds.