Learning with Complementary Labels Revisited: The Selected-Completely-at-Random Setting Is More Practical

Wei Wang [1 2] **Takashi Ishida** [2 1] **Yu-Jie Zhang** [1] **Gang Niu** [2] **Masashi Sugiyama** [2 1]

Abstract

Complementary-label learning is a weakly supervised learning problem in which each training example is associated with one or multiple complementary labels indicating the classes to which it does not belong. Existing consistent approaches have relied on the uniform distribution assumption to model the generation of complementary labels, or on an ordinary-label training set to estimate the transition matrix in non-uniform cases. However, either condition may not be satisfied in real-world scenarios. In this paper, we propose a novel consistent approach that does not rely on these conditions. Inspired by the positive-unlabeled (PU) learning literature, we propose an *unbiased risk estimator* based on the *Selected-Completely-at-Random assumption* for complementary-label learning. We then introduce a *risk-correction approach* to address overfitting problems. Furthermore, we find that complementary-label learning can be expressed as a set of *negative-unlabeled binary classification* problems when using the one-versus-rest strategy. Extensive experimental results on both synthetic and real-world benchmark datasets validate the superiority of our proposed approach over state-of-the-art methods.

1. Introduction

Complementary-label learning is a weakly supervised learning problem that has received a lot of attention recently (Ishida et al., 2017; Feng et al., 2020a; Gao & Zhang, 2021; Liu et al., 2023). In complementary-label learning, we are given training data associated with complementary labels that specify the classes to which the examples do not belong. The task is to learn a multi-class classifier that assigns correct labels to test data as in the standard supervised learning. Collecting training data with complementary labels is much easier and cheaper than collecting ordinary-label data. For example, when asking workers on crowdsourcing platforms to annotate training data, we only need to randomly select a candidate label and then ask them whether the example belongs to that class or not. Such "yes" or "no" questions are much easier to answer than asking workers to determine the ground-truth label from a large set of candidate labels. The benefits and effectiveness of complementary-label learning have also been demonstrated in several machine learning problems and applications, such as domain adaptation (Zhang et al., 2021; Han et al., 2023), semi-supervised learning (Chen et al., 2020b; Ma et al., 2023; Deng et al., 2024), noisy-label learning (Kim et al., 2019), adversarial robustness (Zhou et al., 2022), few-shot learning (Wei et al., 2022), and medical image analysis (Rezaei et al., 2020).

Existing research works with *consistency guarantees* have attempted to solve complementary-label learning problems by making assumptions about the distribution of complementary labels. The remedy started with Ishida et al. (2017), which proposed the *uniform distribution assumption* that a label other than the ground-truth label is sampled from the uniform distribution to be the complementary label. A subsequent work extended it to arbitrary loss functions and models (Ishida et al., 2019) based on the same distribution assumption. Then, Feng et al. (2020a) extended the problem setting to the existence of multiple complementary labels. Recent works have proposed discriminative methods that work by modeling the posterior probabilities of complementary labels instead of the generation process (Chou et al., 2020; Gao & Zhang, 2021; Liu et al., 2023; Lin & Lin, 2023). However, the uniform distribution assumption is still necessary to ensure the classifier consistency property (Liu et al., 2023). Yu et al. (2018) proposed the *biased distribution assumption*, elaborating that the generation of complementary labels follows a *transition matrix*, i.e., the complementary-label distribution is determined by the true label.

In summary, previous complementary-label learning approaches all require either the uniform distribution assump-

[1]The University of Tokyo [2]RIKEN. Correspondence to: Masashi Sugiyama <sugi@k.u-tokyo.ac.jp>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Table 1. Comparison between SCARCE and previous risk-consistent or classifier-consistent complementary-label learning methods.

Method	Uniform distribution assumption-free	Ordinary-label training set-free	Classifier-consistent	Risk-consistent
PC (Ishida et al., 2017)	✗	✓	✓	✓
Forward (Yu et al., 2018)	✓	✗	✓	✗
NN (Ishida et al., 2019)	✗	✓	✓	✓
LMCL (Feng et al., 2020a)	✗	✓	✓	✓
OP (Liu et al., 2023)	✗	✓	✓	✗
SCARCE (Ours)	✓	✓	✓	✓

tion or the biased distribution assumption to guarantee the consistency property, to the best of our knowledge. However, such assumptions may not be satisfied in real-world scenarios. On the one hand, the uniform distribution assumption is too strong, since the transition probability for different complementary labels is undifferentiated, i.e., the transition probability from the true label to a complementary label is constant for all labels. Such an assumption is not realistic since the annotations may be imbalanced and biased (Wei et al., 2023b; Wang et al., 2023a). On the other hand, although the biased distribution assumption is more practical, an ordinary-label training set with *deterministic labels*, also known as *anchor points* (Liu & Tao, 2015), is essential for estimating transition probabilities during the training phase (Yu et al., 2018). However, the collection of ordinary-label data with deterministic labels is often unrealistic in complementary-label learning problems (Feng et al., 2020a; Gao & Zhang, 2021).

To this end, we propose a novel risk-consistent approach named SCARCE, i.e., *Selected-Completely-At-Random ComplEmentary-label learning*, without relying on the uniform distribution assumption or an additional ordinary-label training set. Inspired by the PU learning literature, we propose the Selected-Completely-at-Random (SCAR) assumption for complementary-label learning and propose an unbiased risk estimator accordingly. We then introduce a risk-correction approach to mitigate overfitting issues with risk consistency maintained. Furthermore, we show that complementary-label learning can be expressed as a set of negative-unlabeled (NU) binary classification problems when using the one-versus-rest (OVR) strategy. Table 1 shows the comparison between SCARCE and previous methods. The main contributions of this work are summarized as follows:

- Methodologically, we propose the first consistent complementary-label learning approach without relying on the uniform distribution assumption or an additional ordinary-label dataset in non-uniform cases.
- Theoretically, we uncover the relation between complementary-label learning and NU learning, which provides a new perspective for understanding complementary-label learning. We also prove the convergence rate of the proposed risk estimator by providing an estimation error bound.
- Empirically, the proposed approach is shown to achieve superior performance over state-of-the-art methods on both synthetic and real-world benchmark datasets.

2. Preliminaries

In this section, we review the background of learning with ordinary labels, complementary labels, and PU learning. Then, we introduce a new data distribution assumption for generating complementary labels.

2.1. Learning with Ordinary Labels

Let $\mathcal{X} = \mathbb{R}^d$ denote the d-dimensional feature space and $\mathcal{Y} = \{1, 2, \ldots, q\}$ denote the label space with q class labels. Let $p(\boldsymbol{x}, y)$ be the joint probability density over the random variables $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, then the classification risk is

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)} \left[\mathcal{L}(f(\boldsymbol{x}), y)\right], \tag{1}$$

where $f(\boldsymbol{x})$ is the model prediction and $\mathcal{L}$ can be any *classification-calibrated* loss function, such as the cross-entropy loss (Bartlett et al., 2006). Let $p(\boldsymbol{x})$ denote the marginal density of unlabeled data. Besides, let $\pi_k = p(y = k)$ be the class-prior probability of the k-th class and $p(\boldsymbol{x}|y = k)$ denote the class-conditional density. Then, the classification risk in Eq. (1) can be written as

$$R(f) = \sum_{k=1}^{q} \left(\pi_k \mathbb{E}_{p(\boldsymbol{x}|y=k)} \left[\mathcal{L}(f(\boldsymbol{x}), k)\right]\right). \tag{2}$$

2.2. Learning with Complementary Labels

In complementary-label learning, each training example is associated with one or multiple complementary labels specifying the classes to which the example does not belong. Let $\mathcal{D} = \left\{\left(\boldsymbol{x}_i, \bar{Y}_i\right)\right\}_{i=1}^{n}$ denote the complementary-label training set sampled i.i.d. from an unknown density

$p(\boldsymbol{x}, \bar{Y})$. Here, $\boldsymbol{x} \in \mathcal{X}$ is a feature vector, and $\bar{Y} \subseteq \mathcal{Y}$ is a complementary-label set associated with $\boldsymbol{x}$. In the literature, complementary-label learning can be categorized into single complementary-label learning when we have $|\bar{Y}| = 1$ (Ishida et al., 2017; Gao & Zhang, 2021; Liu et al., 2023), and multiple complementary-label learning when we have $1 \leq |\bar{Y}| \leq q-1$ (Feng et al., 2020a).

In this paper, we consider a more general case where $\bar{Y}$ can contain *any number* of complementary labels, ranging from zero to $q-1$. This means that our method can cover a wider range of applications and can handle additional unlabeled data without complementary labels. For ease of notation, we use a q-dimensional label vector $\bar{\boldsymbol{y}} = [\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_q] \in \{0,1\}^q$ to denote the vector version of $\bar{Y}$, where $\bar{y}_k = 1$ when $k \in \bar{Y}$ and $\bar{y}_k = 0$ otherwise. Let $\bar{\pi}_k = p(\bar{y}_k = 1)$ denote the fraction of training data where the k-th class is considered as a complementary label. Let $p(\boldsymbol{x}|\bar{y}_k = 1)$ and $p(\boldsymbol{x}|\bar{y}_k = 0)$ denote the marginal densities where the k-th class is considered as a complementary label or not. The task of complementary-label learning is to learn a multi-class classifier $f : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{D}$.

2.3. Learning from Positive and Unlabeled Data

In PU learning (Elkan & Noto, 2008; du Plessis et al., 2014; Kiryo et al., 2017), the goal is to learn a binary classifier only from a positive dataset $\mathcal{D}^{\mathrm{P}} = \{(\boldsymbol{x}_i, +1)\}_{i=1}^{n^{\mathrm{P}}}$ and an unlabeled dataset $\mathcal{D}^{\mathrm{U}} = \{\boldsymbol{x}_i\}_{i=1}^{n^{\mathrm{U}}}$. There are mainly two problem settings for PU learning, i.e., the two-sample setting (du Plessis et al., 2014; Niu et al., 2016; Chen et al., 2020a) and the one-sample setting (Elkan & Noto, 2008; Coudray et al., 2023). In the two-sample setting, we assume that $\mathcal{D}^{\mathrm{P}}$ is sampled from the positive-class density $p(\boldsymbol{x}|y = +1)$ and $\mathcal{D}^{\mathrm{U}}$ is sampled from the marginal density $p(\boldsymbol{x})$. In contrast, in the one-sample setting, we assume that an unlabeled dataset is first sampled from the marginal density $p(\boldsymbol{x})$. Then, if a training example is positive, its label is observed with a *constant probability* c, and the example remains unlabeled with probability $1-c$. If a training example is negative, its label is never observed and the example remains unlabeled with probability 1. In this paper, we make use of the one-sample setting for complementary-label learning.

2.4. Generation Process of Complementary Labels

Inspired by the SCAR assumption in PU learning (Elkan & Noto, 2008; Coudray et al., 2023), we introduce the SCAR assumption for generating complementary labels, which can be summarized as follows.

Assumption 1 (Selected-Completely-at-Random (SCAR) Assumption)**.** The complementary-label data with the k-th class as a complementary label are sampled completely at random from the marginal density of the data not belonging to the k-th class, i.e.,

$$p\left(k \in \bar{Y} | \boldsymbol{x}, k \in \mathcal{Y} \backslash \{y\}\right) = p\left(k \in \bar{Y} | k \in \mathcal{Y} \backslash \{y\}\right) = c_k, \tag{3}$$

where $c_k = \bar{\pi}_k / (1 - \pi_k)$ is a constant specifying the fraction of data with the k-th class as a complementary label and $(\boldsymbol{x}, y)$ is sampled from the density $p(\boldsymbol{x}, y)$.

Our motivation is that complementary labels are often generated in a *class-wise* manner. They can be collected by answering "yes" or "no" questions given a pair of an example and a candidate label (Hu et al., 2019; Wang et al., 2021). During an annotation round, we randomly select a candidate label and ask the annotators whether the example belongs to that class or not. The process is repeated iteratively, so that each example may be annotated with *multiple* complementary labels. The SCAR assumption differs from the biased distribution assumption, where only one *single* complementary label is generated by sampling only once from a multinomial distribution. Moreover, the SCAR assumption can be generalized to non-uniform cases by setting c_k to different values for different labels. Therefore, our assumption is more practical in real-world scenarios.

We generate the complementary-label training set $\mathcal{D}$ as follows. First, an unlabeled dataset is sampled from $p(\boldsymbol{x})$. Then, if the latent ground-truth label of an example is not the k-th class, we assign it a complementary label k with probability c_k and still consider it to be an unlabeled example with probability $1 - c_k$. We generate complementary labels for all the examples by following the procedure w.r.t. each of the q labels. The data generation process is summarized in Appendix B.2.

3. Methodology

In this section, we first introduce a risk rewrite formulation for complementary-label learning. Then, we propose an unbiased risk estimator, followed by its theoretical analysis. Finally, we present a risk-correction approach to improve the generalization performance.

3.1. Risk Rewrite

Under the SCAR assumption, the ordinary multi-class classification risk in Eq. (1) can be rewritten as follows (the proof is given in Appendix C).

Theorem 2. *Under Assumption 1, the classification risk in Eq. (1) can be equivalently expressed as*

$$R(f) = \sum_{k=1}^{q} \left(\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)} \left[(\bar{\pi}_k + \pi_k - 1) \mathcal{L}(f(\boldsymbol{x}), k) \right] + \mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=0)} \left[(1 - \bar{\pi}_k) \mathcal{L}(f(\boldsymbol{x}), k) \right] \right). \tag{4}$$

Theorem 2 shows that the ordinary classification risk in Eq. (1) can be equivalently expressed using densities $p(\boldsymbol{x}|\bar{y}_k = 1)$ and $p(\boldsymbol{x}|\bar{y}_k = 0)$. Therefore, we can perform *empirical risk minimization* by minimizing an unbiased estimation of Eq. (4) with training data sampled from $p(\boldsymbol{x}|\bar{y}_k = 1)$ and $p(\boldsymbol{x}|\bar{y}_k = 0)$.

Ideally, *any multi-class loss function* can be used to instantiate $\mathcal{L}$, such as the cross-entropy loss. In addition, *any model and optimizer* can be used, which reveals the universality of our proposed approach. According to our experimental results, we find that the cross-entropy loss is not robust and often leads to inferior performance, possibly due to its unboundedness (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Feng et al., 2020a; Wei et al., 2023a), which will be discussed in Section 4.3. In Section 3.2, we provide an instantiation based on the OVR strategy.

3.2. OVR Strategy

The OVR strategy decomposes multi-class classification into a series of binary classification problems, which is a common strategy with extensive theoretical guarantees and sound performance (Rifkin & Klautau, 2004; Zhang, 2004). It instantiates the loss function $\mathcal{L}$ in Eq. (1) with the OVR loss, i.e.

$$R(f_1, f_2, \ldots, f_q) = \mathbb{E}_{p(\boldsymbol{x},y)} \left[\ell\left(f_y(\boldsymbol{x})\right) + \sum_{k\in\mathcal{Y}\setminus\{y\}} \ell\left(-f_k(\boldsymbol{x})\right)\right]. \quad (5)$$

Here, f_k is a binary classifier w.r.t. the k-th class, $\mathbb{E}$ denotes the expectation, and $\ell : \mathbb{R} \to \mathbb{R}_+$ is a non-negative binary-class loss function. Then, the predicted label for a test instance $\boldsymbol{x}$ is determined as

$$f(\boldsymbol{x}) = \arg\max_{k\in\mathcal{Y}} \; f_k(\boldsymbol{x}). \quad (6)$$

The goal is to find optimal classifiers $f_1^*, f_2^*, \ldots, f_q^*$ in a function class $\mathcal{F}$ which achieve the minimum classification risk in Eq. (5), i.e.,

$$\left(f_1^*, f_2^*, \ldots, f_q^*\right) = \arg\min_{f_1, f_2, \ldots, f_q \in \mathcal{F}} R(f_1, f_2, \ldots, f_q). \quad (7)$$

We show that the OVR risk can be rewritten using densities $p(\boldsymbol{x}|\bar{y}_k = 1)$ and $p(\boldsymbol{x}|\bar{y}_k = 0)$ as well.

Theorem 3. *When the OVR loss is used, the classification risk in Eq. (5) can be equivalently expressed as* $R(f_1, f_2, \ldots, f_q) = \sum_{k=1}^{q} R_k(f_k)$, *where*

$$R_k(f_k) = \mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)} \left[(1-\pi_k)\ell\left(-f_k(\boldsymbol{x})\right) + (\bar{\pi}_k + \pi_k - 1)\ell\left(f_k(\boldsymbol{x})\right)\right] + \mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=0)} \left[(1-\bar{\pi}_k)\,\ell\left(f_k(\boldsymbol{x})\right)\right]. \quad (8)$$

The proof is given in Appendix D. Since the true densities $p(\boldsymbol{x}|\bar{y}_k = 1)$ and $p(\boldsymbol{x}|\bar{y}_k = 0)$ are not directly accessible, we approximate the risk *empirically*. Suppose we have binary-class datasets $\mathcal{D}_k^{\mathrm{N}}$ and $\mathcal{D}_k^{\mathrm{U}}$ sampled i.i.d. from $p(\boldsymbol{x}|\bar{y}_k = 1)$ and $p(\boldsymbol{x}|\bar{y}_k = 0)$, respectively. Then, an unbiased risk estimator can be derived from these binary-class datasets to approximate the classification risk in Theorem 3 as $\widehat{R}(f_1, f_2, \ldots, f_q) = \sum_{k=1}^{q} \widehat{R}_k(f_k)$, where

$$\widehat{R}_k(f_k) = \frac{1}{n_k^{\mathrm{N}}} \sum_{i=1}^{n_k^{\mathrm{N}}} \left((1-\pi_k)\,\ell\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right) + (\bar{\pi}_k + \pi_k - 1)\,\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)\right) + \frac{(1-\bar{\pi}_k)}{n_k^{\mathrm{U}}} \sum_{i=1}^{n_k^{\mathrm{U}}} \ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}})\right). \quad (9)$$

We may add regularization terms to $\widehat{R}(f_1, f_2, \ldots, f_q)$ when necessary (Loshchilov & Hutter, 2019). This paper considers generating the binary-class datasets $\mathcal{D}_k^{\mathrm{N}}$ and $\mathcal{D}_k^{\mathrm{U}}$ by *duplicating* instances of $\mathcal{D}$. Specifically, if the k-th class is a complementary label of a training example, we regard its duplicated instance as a *negative example* sampled from $p(\boldsymbol{x}|\bar{y}_k = 1)$ and put the duplicated instance in $\mathcal{D}_k^{\mathrm{N}}$. If the k-th class is not a complementary label of a training example, we regard its duplicated instance as an *unlabeled example* sampled from $p(\boldsymbol{x}|\bar{y}_k = 0)$ and put the duplicated instance in $\mathcal{D}_k^{\mathrm{U}}$. In this way, we can obtain q negative binary-class datasets and q unlabeled binary-class datasets ($k \in \mathcal{Y}$):

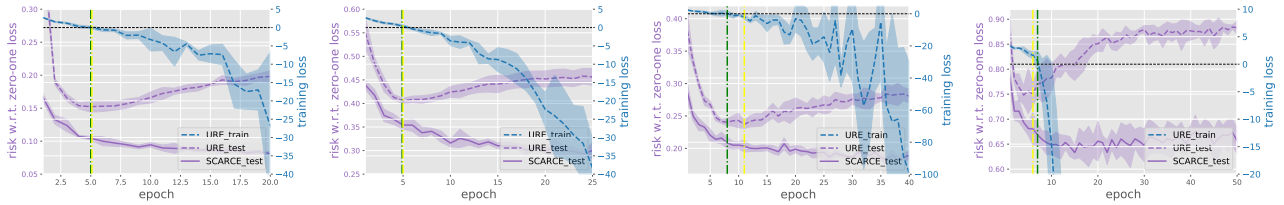
$$\mathcal{D}_k^{\mathrm{N}} = \left\{(\boldsymbol{x}_{k,i}^{\mathrm{N}}, -1)\right\}_{i=1}^{n_k^{\mathrm{N}}} = \left\{(\boldsymbol{x}_j, -1) | (\boldsymbol{x}_j, \bar{Y}_j) \in \mathcal{D}, k \in \bar{Y}_j\right\}; \quad (10)$$

$$\mathcal{D}_k^{\mathrm{U}} = \left\{\boldsymbol{x}_{k,i}^{\mathrm{U}}\right\}_{i=1}^{n_k^{\mathrm{U}}} = \left\{\boldsymbol{x}_j | (\boldsymbol{x}_j, \bar{Y}_j) \in \mathcal{D}, k \notin \bar{Y}_j\right\}. \quad (11)$$

The details of our algorithm are summarized in Algorithm 1. When the class priors π_k are not accessible to the learning algorithm, they can be estimated by off-the-shelf mixture proportion estimation approaches (Scott, 2015; Ramaswamy et al., 2016; Zhang et al., 2020; Garg et al., 2021; Yao et al., 2022) with $\mathcal{D}_k^{\mathrm{N}}$ and $\mathcal{D}_k^{\mathrm{U}}$. Notably, the *irreducibility* (Blanchard et al., 2010; Scott et al., 2013) assumption is necessary for class-prior estimation. However, it is still less demanding than the biased distribution assumption, which requires additional ordinary-label training data with deterministic labels, a.k.a. anchor points, to estimate the transition matrix (Yu et al., 2018). We present the details of a class-prior estimation algorithm in Appendix A.

3.3. Relation to Negative-Unlabeled Learning

First, we provide the risk rewrite results of NU learning (Niu et al., 2016). Consider the binary classification problem with a class prior $\pi_+ = p(y = +1)$. Suppose an NU dataset is generated by following the SCAR assumption in PU learning with the positive and negative



(a) MNIST (b) Kuzushiji-MNIST (c) Fashion-MNIST (d) CIFAR-10

Figure 1. Training curves and test curves of the method that minimizes the URE and test curves of our proposed risk-correction approach SCARCE. The green dashed lines indicate when the URE becomes negative while the yellow dashed lines indicate when the overfitting phenomena occur. The complementary labels are generated by following the uniform distribution assumption. ResNet is used as the model architecture for CIFAR-10 and MLP is used for other datasets.

Algorithm 1 SCARCE

Input: Complementary-label training set $\mathcal{D}$, class priors π_k ($k \in \mathcal{Y}$), unseen instance $\boldsymbol{x}_*$, epoch $T_{\max}$, iteration $I_{\max}$.

for $t = 1, 2, \ldots, T_{\max}$ **do**
  Shuffle the complementary-label training set $\mathcal{D}$;
  for $j = 1, \ldots, I_{\max}$ **do**
    Fetch mini-batch $\mathcal{D}_j$ from $\mathcal{D}$;
    Update the shared representation layers and specific classification layers $f_1, f_2, \ldots, f_q$ by minimizing the corrected risk estimator $\widetilde{R}(f_1, f_2, \ldots, f_q)$ in Eq. (17);
  end for
end for
Return $y_* = \arg\max_{k \in \mathcal{Y}} \; f_k(\boldsymbol{x}_*)$;

Output: Predicted label y_*.

Algorithm 2 SCARCE-Meta

Input: Complementary-label training set $\mathcal{D}$, PU learning algorithm $\mathcal{A}$, unseen instance $\boldsymbol{x}_*$, epoch $T_{\max}$, iteration $I_{\max}$, number of labels q.

for $t = 1, 2, \ldots, T_{\max}$ **do**
  Shuffle the complementary-label training set $\mathcal{D}$;
  for $j = 1, \ldots, I_{\max}$ **do**
    Fetch a mini-batch $\mathcal{D}_j$ from $\mathcal{D}$;
    for $k = 1, \ldots, q$ **do**
      Construct a negative dataset $\mathcal{D}_k^{\mathrm{N}}$ and an unlabeled dataset $\mathcal{D}_k^{\mathrm{U}}$ according to Eq. (10) and (11);
      Train a binary classifier $f_k \leftarrow \mathcal{A}(\mathcal{D}_k^{\mathrm{N}}, \mathcal{D}_k^{\mathrm{U}})$ by regarding negative data as positive.
    end for
  end for
end for
Return $y_* = \arg\min_{k \in \mathcal{Y}} \; f_k(\boldsymbol{x}_*)$;

Output: Predicted label y_*.

classes swapped. Our goal is to learn a binary classifier f^{NU} from the NU dataset. Let $\bar{\pi}_- = p(\bar{y} = -1)$ denote the proportion of negative data in the entire NU dataset. Let $p(\boldsymbol{x}|\bar{y} = -1)$ and $p(\boldsymbol{x}|\bar{y} = 0)$ denote the density of observed negative data and unlabeled data, respectively.

Lemma 4. *Based on the assumptions above, the classification risk of binary classification can be equivalently expressed as*

$$\mathbb{E}_{p(\boldsymbol{x}|\bar{y}=-1)}\left[(1-\pi_+)\ell\left(-f^{\mathrm{NU}}(\boldsymbol{x})\right) + (\bar{\pi}_- + \pi_+ - 1)\right.$$
$$\left.\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] + \mathbb{E}_{p(\boldsymbol{x}|\bar{y}=0)}\left[(1-\bar{\pi}_-)\,\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right]. \quad (12)$$

We observe that the multi-class classification risk in Theorem 3 is the sum of Eq. (12) by considering each class as a positive class in turn. Furthermore, besides minimizing the NU classification risk $R_k(f_k)$, we can adopt any other PU learning approach (Chen et al., 2020a; Garg et al., 2021; Li et al., 2022; Wang et al., 2023; Jiang et al., 2023; Dai et al., 2023) to derive the binary classifier f_k by swapping the positive class and the negative class. Finally, we can predict the label for a test instance as the class of the minimum model output, since the positive and negative classes are swapped. Therefore, the proposal can be considered as a *general framework* for solving complementary-label learning problems. Based on this finding, we propose a *meta* complementary-label learning algorithm in Algorithm 2, and the proposed method SCARCE can be considered as an instantiation. In particular, when employing deep neural networks as the model architecture for PU learning algorithms, we can share the representation learning layers and use specific classification layers for different labels, which may allow training different classifiers simultaneously.

3.4. Theoretical Analysis

Calibration. We show that the proposed risk can be calibrated to the 0-1 loss (Zhang, 2004). Let $R_{0-1}(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\mathbb{I}(f(\boldsymbol{x}) \neq y)$ denote the expected 0-1 loss where $f(\boldsymbol{x}) = \arg\max_{k \in \mathcal{Y}} \; f_k(\boldsymbol{x})$ and $R_{0-1}^* = \min_f \; R_{0-1}(f)$ denote the Bayes error. Besides, let $R^* = \min_{f_1, f_2, \ldots, f_q} \; R(f_1, f_2, \ldots, f_q)$ denote the minimum risk

of the proposed risk. Then we have the following theorem (its proof is given in Appendix F).

Theorem 5. *Suppose the binary-class loss function ℓ is convex, bounded below, differential, and satisfies $\ell(z) \leq \ell(-z)$ when $z > 0$. Then we have that for any $\epsilon_1 > 0$, there exists an $\epsilon_2 > 0$ such that*

$$R\left(f_1, f_2, \ldots, f_q\right) \leq R^* + \epsilon_2 \Rightarrow R_{0-1}(f) \leq R_{0-1}^* + \epsilon_1. \tag{13}$$

Remark 6. The infinite-sample consistency elucidates that the proposed risk can be calibrated to the 0-1 loss. Therefore, if we minimize the proposed risk and obtain the optimal classifier, the classifier also achieves the Bayes error.

Estimation error bound. We further elaborate the convergence property of the empirical risk estimator $\widehat{R}(f_1, f_2, \ldots, f_q)$ by providing its estimation error bound. The optimal classifiers w.r.t. $\widehat{R}(f_1, f_2, \ldots, f_q)$ are

$$\left(\widehat{f}_1, \widehat{f}_2, \ldots, \widehat{f}_q\right) = \underset{f_1, f_2, \ldots, f_q \in \mathcal{F}}{\arg\min} \widehat{R}\left(f_1, f_2, \ldots, f_q\right). \tag{14}$$

Theorem 7. *Based on the above assumptions, for any $\delta > 0$, the following inequality holds with probability at least $1-\delta$:*

$$\begin{aligned} &R\left(\widehat{f}_1, \widehat{f}_2, \ldots, \widehat{f}_q\right) - R\left(f_1^*, f_2^*, \ldots, f_q^*\right) \leq \\ &\sum_{k=1}^{q}\Bigg((4-4\bar{\pi}_k)L_\ell \mathfrak{R}_{n_k^{\mathrm{U}}, p_k^{\mathrm{U}}}(\mathcal{F}) + (1-\bar{\pi}_k)C_\ell\sqrt{\frac{2\ln(2/\delta)}{n_k^{\mathrm{U}}}} \\ &+ (8-8\pi_k-4\bar{\pi}_k)L_\ell \mathfrak{R}_{n_k^{\mathrm{N}}, p_k^{\mathrm{N}}}(\mathcal{F}) \\ &+(2-2\pi_k-\bar{\pi}_k)C_\ell\sqrt{\frac{2\ln(2/\delta)}{n_k^{\mathrm{N}}}}\Bigg), \end{aligned} \tag{15}$$

where $\mathfrak{R}_{n_k^{\mathrm{U}}, p_k^{\mathrm{U}}}(\mathcal{F})$ and $\mathfrak{R}_{n_k^{\mathrm{N}}, p_k^{\mathrm{N}}}(\mathcal{F})$ denote the Rademacher complexity of $\mathcal{F}$ given n_k^{U} unlabeled data sampled from $p(\boldsymbol{x}|\bar{y}_k = 0)$ and n_k^{N} negative data sampled from $p(\boldsymbol{x}|\bar{y}_k = 1)$ respectively.

Remark 8. Theorem 7 elucidates an estimation error bound of our proposed risk estimator. When n_k^{U} and $n_k^{\mathrm{N}} \to \infty$, $R\left(\widehat{f}_1, \widehat{f}_2, \ldots, \widehat{f}_q\right) \to R\left(f_1^*, f_2^*, \ldots, f_q^*\right)$ because $\mathfrak{R}_{n_k^{\mathrm{U}}, p_k^{\mathrm{U}}}(\mathcal{F}) \to 0$ and $\mathfrak{R}_{n_k^{\mathrm{N}}, p_k^{\mathrm{N}}}(\mathcal{F}) \to 0$ for all parametric models with a bounded norm such as deep neural networks with weight decay (Golowich et al., 2018). Furthermore, the estimation error bound converges in $\mathcal{O}_p\left(\sum_{k=1}^{q}\left(1/\sqrt{n_k^{\mathrm{N}}} + 1/\sqrt{n_k^{\mathrm{U}}}\right)\right)$, where $\mathcal{O}_p$ denotes the order in probability.

3.5. Risk-Correction Approach

Although the unbiased risk estimator (URE) has sound theoretical properties, we have found that it can encounter several overfitting problems when using complex models such as deep neural networks. The training curves and test curves of the method that works by minimizing the URE in Eq. (9) are shown in Figure 1. We can observe that the overfitting phenomena often occur almost simultaneously when the training loss becomes negative. We conjecture the overfitting problems are related with the negative terms in Eq. (9) (Kiryo et al., 2017; Lu et al., 2020; Cao et al., 2021). Therefore, following Lu et al. (2020); Wang et al. (2023b), we wrap each potentially negative term with a *non-negative risk-correction function* $g(z)$, such as the absolute value function $g(z) = |z|$. For ease of notation, we introduce

$$\begin{aligned} \widehat{R}_k^{\mathrm{P}}(f_k) =& \frac{\bar{\pi}_k + \pi_k - 1}{n_k^{\mathrm{N}}} \sum_{i=1}^{n_k^{\mathrm{N}}} \ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right) \\ &+ \frac{1-\bar{\pi}_k}{n_k^{\mathrm{U}}} \sum_{i=1}^{n_k^{\mathrm{U}}} \ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}})\right). \end{aligned} \tag{16}$$

Then, the corrected risk estimator can be written as $\widetilde{R}\left(f_1, f_2, \ldots, f_q\right) = \sum_{k=1}^{q} \widetilde{R}_k(f_k)$, where

$$\widetilde{R}_k(f_k) = g\left(\widehat{R}_k^{\mathrm{P}}(f_k)\right) + \frac{1-\pi_k}{n_k^{\mathrm{N}}} \sum_{i=1}^{n_k^{\mathrm{N}}} \ell\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right). \tag{17}$$

It is obvious that Eq. (17) is an upper bound of Eq. (9), so the bias is always present. Next, we perform a theoretical analysis to clarify that the corrected risk estimator is *biased but consistent*. Since $\mathbb{E}\left[\widehat{R}_k^{\mathrm{P}}(f_k)\right] = \pi_k \mathbb{E}_{p(\boldsymbol{x}|y=k)} \ell\left(f_k(\boldsymbol{x})\right)$ is non-negative, we assume that there exists a *non-negative constant* β such that for $\forall k \in \mathcal{Y}, \mathbb{E}\left[\widehat{R}_k^{\mathrm{P}}(f_k)\right] \geq \beta$. We also assume that the risk-correction function $g(z)$ is Lipschitz continuous with a Lipschitz constant L_g. Besides, we assume that there exists some constant $C_{\mathfrak{R}}$ such that the Rademacher complexity $\mathfrak{R}_{n,p}(\mathcal{F})$ for unlabeled (with $n = n_k^{\mathrm{U}}, p = p_k^{\mathrm{U}}$) and negative data (with $n = n_k^{\mathrm{N}}, p = p_k^{\mathrm{N}}$) satisfies $\mathfrak{R}_{n,p}(\mathcal{F}) \leq C_{\mathfrak{R}}/\sqrt{n}$. This assumption holds for many models, such as fully connected neural networks and linear-in-parameter models with a bounded norm (Golowich et al., 2018; Lu et al., 2020). We introduce $\left(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q\right) = \arg\min_{f_1, f_2, \ldots, f_q \in \mathcal{F}} \widetilde{R}\left(f_1, f_2, \ldots, f_q\right)$ and $\Delta_k = \exp\left(-2\beta^2 / \left((1-\pi_k-\bar{\pi}_k)^2 C_\ell^2 / n_k^{\mathrm{N}} + (1-\bar{\pi}_k)^2 C_\ell^2 / n_k^{\mathrm{U}}\right)\right)$. Then we have the following theorems (the proofs are given in Appendix H and I respectively).

Theorem 9. *Based on the above assumptions, the bias of the expectation of the corrected risk estimator has the following lower and upper bounds:*

$$\begin{aligned} 0 \leq& \mathbb{E}[\widetilde{R}(f_1, f_2, \ldots, f_q)] - R(f_1, f_2, \ldots, f_q) \\ \leq& \sum_{k=1}^{q} \left(2 - 2\bar{\pi}_k - \pi_k\right)\left(L_g + 1\right) C_\ell \Delta_k. \end{aligned} \tag{18}$$

Table 2. Classification accuracy (mean±std) of each method on MNIST. The best performance is shown in bold (pairwise *t*-test at the 0.05 significance level).

Setting	Uniform		Biased-a		Biased-b		SCAR-a		SCAR-b	
Model	MLP	LeNet	MLP	LeNet	MLP	LeNet	MLP	LeNet	MLP	LeNet
PC	71.11 ±0.83	82.69 ±1.15	69.29 ±0.97	87.82 ±0.69	71.59 ±0.85	87.66 ±0.66	66.97 ±1.03	11.00 ±0.79	57.67 ±0.98	49.17 ±35.9
NN	67.75 ±0.96	86.16 ±0.69	30.59 ±2.31	46.27 ±2.61	38.50 ±3.93	63.67 ±3.75	67.39 ±0.68	86.58 ±0.95	63.95 ±0.56	79.94 ±0.48
GA	88.00 ±0.85	96.02 ±0.15	65.97 ±7.87	94.55 ±0.43	75.77 ±1.48	94.87 ±0.28	62.62 ±2.29	90.23 ±0.92	56.91 ±2.08	78.66 ±0.61
L-UW	73.49 ±0.88	77.74 ±0.97	39.63 ±0.57	32.21 ±1.20	42.77 ±1.42	34.57 ±1.90	35.08 ±1.59	33.82 ±2.44	30.24 ±1.81	24.28 ±2.74
L-W	62.24 ±0.50	63.04 ±1.58	36.90 ±0.34	29.25 ±0.94	41.55 ±0.63	32.98 ±2.25	33.53 ±2.08	26.02 ±1.31	28.99 ±2.38	23.69 ±2.94
OP	78.87 ±0.46	88.76 ±1.68	73.46 ±0.71	85.96 ±1.02	74.16 ±0.52	87.23 ±1.31	76.29 ±0.23	86.94 ±1.94	68.12 ±0.51	71.67 ±2.30
SCARCE	**91.27 ±0.20**	**97.00 ±0.30**	**88.14 ±0.70**	**96.14 ±0.32**	**89.51 ±0.44**	**96.62 ±0.10**	**90.98 ±0.27**	**96.72 ±0.16**	**81.85 ±0.25**	**87.05 ±0.28**

Furthermore, for any $\delta > 0$, the following inequality holds with probability at least $1-\delta$:

$$|\widetilde{R}(f_1, f_2, \ldots, f_q) - R(f_1, f_2, \ldots, f_q)| \leq \mathcal{O}_p\left(\sum_{k=1}^{q}\left(1/\sqrt{n_k^{\mathrm{N}}} + 1/\sqrt{n_k^{\mathrm{U}}}\right)\right). \tag{19}$$

Theorem 10. *Based on the above assumptions, for any $\delta > 0$, the following inequality holds with probability at least $1-\delta$:*

$$R(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q) - R(f_1^*, f_2^*, \ldots, f_q^*) \leq \mathcal{O}_p\left(\sum_{k=1}^{q}\left(1/\sqrt{n_k^{\mathrm{N}}} + 1/\sqrt{n_k^{\mathrm{U}}}\right)\right). \tag{20}$$

Remark 11. Theorem 9 shows that $\widetilde{R}(f_1, f_2, \ldots, f_q) \to R(f_1, f_2, \ldots, f_q)$ as n_k^{U} and $n_k^{\mathrm{N}} \to \infty$, indicating that the corrected risk estimator is biased but consistent. An estimation error bound is also shown in Theorem 10. The convergence rate of the estimation error bound is still the same after employing the risk-correction function.

4. Experiments

In this section, we validate the effectiveness of SCARCE through extensive experiments.[1]

4.1. Experiments on Synthetic Benchmark Datasets

We conducted experiments on synthetic benchmark datasets, including MNIST (LeCun et al., 1998), Kuzushiji-MNIST (Clanuwat et al., 2018), Fashion-MNIST (Xiao et al., 2017), and CIFAR-10 (Krizhevsky & Hinton, 2009). We considered various generation processes of complementary labels by following the uniform, biased, and SCAR assumptions. We evaluated the classification performance of SCARCE against six single complementary-label learning methods, including PC (Ishida et al., 2017), NN (Ishida et al., 2019), GA (Ishida et al., 2019), L-UW (Gao & Zhang, 2021), L-W (Gao & Zhang, 2021), and OP (Liu et al., 2023). The logistic loss was adopted to instantiate the binary-class loss function l, and the absolute value function was used as the risk-correction function g for SCARCE. All the methods were implemented in PyTorch (Paszke et al., 2019). We used the Adam optimizer (Kingma & Ba, 2015). For a fair comparison, we use the same hyperparameters for all methods. The learning rate and batch size were fixed to 1e-3 and 256 for all the datasets, respectively. The weight decay was 1e-3 for CIFAR-10 and 1e-5 for the other three datasets. The number of epochs was set to 200, and we recorded the mean accuracy in the last ten epochs. Details of the datasets, models, and method descriptions can be found in Appendix J. We assumed that the class priors were accessible to the learning algorithm. We randomly generated complementary labels five times with different seeds and recorded the mean accuracy and standard deviations. In addition, a pairwise *t*-test at the 0.05 significance level is performed to show whether the performance advantages are significant.

Tables 2, 3, and 4 show the classification performance of each method with different models and generation settings of complementary labels on MNIST, Kuzushiji-MNIST, and Fashion-MNIST respectively. The experimental re-

[1]Our implementation of SCARCE is available at https://github.com/wwangwitsel/SCARCE.

Table 3. Classification accuracy (mean±std) of each method on Kuzushiji-MNIST. The best performance is shown in bold (pairwise *t*-test at the 0.05 significance level).

Setting	Uniform		Biased-a		Biased-b		SCAR-a		SCAR-b	
Model	MLP	LeNet	MLP	LeNet	MLP	LeNet	MLP	LeNet	MLP	LeNet
PC	42.93 ±0.33	56.79 ±1.54	41.60 ±0.97	67.39 ±1.04	42.53 ±0.80	66.81 ±1.33	39.58 ±1.35	42.59 ±29.8	33.95 ±1.14	37.67 ±25.3
NN	39.42 ±0.68	58.57 ±1.15	23.97 ±2.53	31.10 ±2.95	29.93 ±1.80	48.72 ±2.89	39.31 ±1.18	56.84 ±2.10	38.68 ±0.58	56.70 ±1.08
GA	60.83 ±1.37	76.17 ±0.44	43.22 ±3.03	**75.04 ±0.92**	48.03 ±2.93	**77.05 ±1.67**	36.56 ±2.96	59.16 ±3.30	33.02 ±2.31	52.92 ±2.39
L-UW	43.00 ±1.20	49.31 ±1.95	27.89 ±0.51	25.82 ±0.78	31.53 ±0.42	30.05 ±1.63	21.49 ±0.57	19.71 ±1.44	18.36 ±1.23	16.67 ±1.86
L-W	37.21 ±0.59	42.69 ±2.54	26.75 ±0.61	25.86 ±0.64	30.10 ±0.57	27.94 ±1.68	21.22 ±0.77	18.28 ±2.11	18.41 ±1.66	16.25 ±1.51
OP	51.78 ±0.41	65.94 ±1.38	45.66 ±0.90	65.59 ±1.71	47.47 ±1.26	64.65 ±1.68	49.95 ±0.79	59.93 ±1.38	42.72 ±0.95	56.36 ±2.15
SCARCE	**67.95 ±1.29**	**79.81 ±1.19**	**62.43 ±1.02**	**75.99 ±0.91**	**64.98 ±0.72**	**78.53 ±0.57**	**66.72 ±0.69**	**78.27 ±1.09**	**61.78 ±0.36**	**72.03 ±0.45**

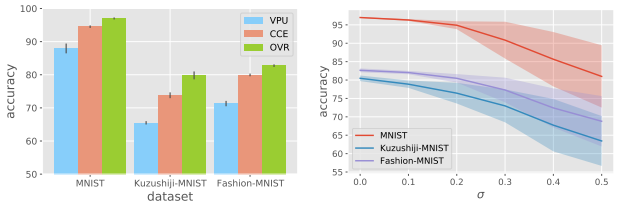
sult on CIFAR-10 is shown in Appendix K. We can observe that: a) Out of 40 cases of different distributions and datasets, SCARCE achieves the best performance in 39 cases, which clearly validates its effectiveness. b) Some consistent approaches based on the uniform distribution assumption can achieve comparable or better performance for the "uniform" setting. For example, GA outperforms SCARCE on CIFAR-10. However, its performance drops significantly on other distribution settings.

4.2. Experiments on Real-World Benchmark Datasets

We also verified the effectiveness of SCARCE on two real-world complementary-label datasets CLCIFAR-10 and CLCIFAR-20 (Wang et al., 2023a). The datasets were annotated by human annotators from Amazon Mechanical Turk (MTurk). The distribution of complementary labels is too complex to be captured by any of the above assumptions. Moreover, the complementary labels may be *noisy*, which means that the complementary labels may be annotated as ground-truth labels by mistake. There are three human-annotated complementary labels for each example, so they can be considered as multiple complementary-label datasets. We evaluated the classification performance of SCARCE against nine multiple complementary-label learning or partial-label learning methods, including CC (Feng et al., 2020b), PRODEN (Lv et al., 2020), EXP (Feng et al., 2020a), MAE (Feng et al., 2020a), Phuber-CE (Feng et al., 2020a), LWS (Wen et al., 2021), CAVL (Zhang et al., 2022), IDGP (Qiao et al., 2023), and POP (Xu et al., 2023). For a fair comparison, we set the same hyperparameters for all the methods as those of CIFAR-10 in Section 4.1. Details of the models and method descriptions can be found in Appendix J. We found that the performance of some approaches was unstable with different network initialization, so we randomly initialized the network five times with different seeds and recorded the mean accuracy and standard deviations.

Table 5 shows the experimental results on CLCIFAR-10 and CLCIFAR-20 with different models. We can observe that: a) SCARCE achieves the best performance in all cases, further confirming its effectiveness. b) The superiority is even more evident on CLCIFAR-20, a more complex dataset with extremely limited supervision. It demonstrates the advantages of SCARCE in dealing with real-world datasets.

4.3. Further Analysis



(a) Comparison between Different Instantiations of SCARCE

(b) Sensitivity Analysis

Figure 2. (a) Classification accuracy of different instantiations of SCARCE on different datasets. (b) Classification accuracy given inaccurate class priors.

Comparison between different instantiations of SCARCE. In Theorem 2, any multi-class loss function can be used to instantiate $\mathcal{L}$. Therefore, we also investigated the classification performance of the cross-entropy

Table 4. Classification accuracy (mean±std) of each method on Fashion-MNIST. The best performance is shown in bold (pairwise *t*-test at the 0.05 significance level).

Setting	Uniform		Biased-a		Biased-b		SCAR-a		SCAR-b	
Model	MLP	LeNet	MLP	LeNet	MLP	LeNet	MLP	LeNet	MLP	LeNet
PC	64.82 ±1.27	69.56 ±1.82	61.14 ±1.09	72.89 ±1.26	61.20 ±0.79	73.04 ±1.38	63.08 ±0.88	23.28 ±29.7	47.23 ±2.38	37.53 ±25.2
NN	63.89 ±0.92	70.34 ±1.09	25.66 ±2.12	36.93 ±3.86	30.75 ±0.96	40.88 ±3.71	63.47 ±0.70	70.83 ±0.87	55.96 ±1.55	63.06 ±1.38
GA	77.04 ±0.95	81.91 ±0.43	50.04 ±4.30	74.73 ±0.96	49.02 ±5.76	75.66 ±1.10	54.74 ±3.04	74.75 ±1.17	44.75 ±3.04	60.01 ±1.47
L-UW	**80.29 ±0.44**	72.43 ±2.07	40.26 ±2.49	29.46 ±1.70	43.55 ±1.61	33.53 ±1.35	35.71 ±1.50	30.73 ±1.64	31.43 ±2.98	22.03 ±3.62
L-W	75.14 ±0.40	61.89 ±0.88	39.87 ±0.95	27.57 ±1.70	42.02 ±1.41	32.69 ±0.68	31.86 ±2.16	27.37 ±2.30	30.26 ±1.68	21.61 ±2.12
OP	69.03 ±0.71	71.28 ±0.94	62.93 ±1.25	70.82 ±1.15	62.25 ±0.36	68.94 ±2.78	66.29 ±0.60	69.52 ±1.18	56.55 ±1.39	56.39 ±3.03
SCARCE	**80.44 ±0.19**	**82.74 ±0.39**	**70.08 ±2.53**	**79.74 ±1.10**	**71.97 ±1.09**	**80.43 ±0.69**	**79.75 ±0.60**	**82.55 ±0.30**	**71.16 ±0.66**	**72.79 ±0.62**

Table 5. Classification accuracy (mean±std) of each method on CLCIFAR-10 and CLCIFAR-20. The best performance is shown in bold (pairwise *t*-test at the 0.05 significance level).

Dataset	Model	CC	PRODEN	EXP	MAE	Phuber-CE	LWS	CAVL	IDGP	POP	SCARCE
CLCIFAR-10	ResNet	31.56 ±2.17	26.37 ±0.98	34.84 ±4.19	19.48 ±2.88	**41.13 ±0.74**	13.05 ±4.18	24.12 ±3.32	10.00 ±0.00	26.75 ±1.28	**42.04 ±0.96**
	DenseNet	37.03 ±1.77	31.31 ±1.06	**43.27 ±1.33**	22.77 ±0.22	39.92 ±0.91	10.00 ±0.00	25.31 ±4.06	10.00 ±0.00	31.45 ±1.16	**44.41 ±0.43**
CLCIFAR-20	ResNet	5.00 ±0.00	6.69 ±0.31	7.21 ±0.17	5.00 ±0.00	8.10 ±0.18	5.20 ±0.45	5.00 ±0.00	4.96 ±0.09	6.40 ±0.33	**20.08 ±0.62**
	DenseNet	5.00 ±0.00	5.00 ±0.00	7.51 ±0.91	5.67 ±1.49	7.22 ±0.39	5.00 ±0.00	5.09 ±0.13	5.00 ±0.00	5.00 ±0.00	**19.91 ±0.68**

loss (CCE). Furthermore, we adopted the risk-correction approach to mitigate overfitting problems. We also included another instantiation of the meta-algorithm in Section 3.3. We used VPU (Chen et al., 2020a) as the PU learning approach. For a fair comparison, we did not use the mixup loss. We compared them with the default instantiation of SCARCE, i.e. the OVR loss, and Figure 2 (a) shows the experimental results. We generated complementary labels with the uniform distribution assumption and used LeNet as the model architecture. We can observe that the OVR loss outperforms CCE and VPU. We conjecture that the inferior performance of CCE may be related to its unboundedness (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Wei et al., 2023a).

Sensitivity analysis. We investigated the influence of inaccurate class priors on the classification performance of SCARCE. Specifically, let $\bar{\pi}_k = \xi_k \pi_k$ denote the corrupted class prior probability for the k-th class where ξ_k is sampled from a normal distribution $\mathcal{N}(1, \sigma^2)$. We further normalized the obtained corrupted class priors to ensure that they sum up to one. Figure 2 (b) shows the classification performance given inaccurate class priors using the uniform generation process and LeNet as the model architecture. From Figure 2 (b), we can see that the performance is still satisfactory with small perturbations of the class priors. However, the performance will degenerate if the class priors deviate too much from the ground-truth values.

5. Conclusion

In this paper, we proposed the first attempt towards consistent complementary-label learning without relying on the uniform distribution assumption or an ordinary-label training set to estimate the transition matrix in non-uniform cases. Based on a more practical distribution assumption, a consistent approach was proposed with theoretical guarantees. We also observed that complementary-label learning could be expressed as a set of NU classification problems when using the OVR strategy. Extensive experimental results on benchmark datasets validated the effectiveness of our proposed approach.

Acknowledgements

The authors thank Hsuan-Tien Lin, Deng-Bao Wang, and Yi Gao for helpful comments and discussions. The authors thank the anonymous reviewers for their helpful comments and suggestions. WW was supported by the SGU MEXT Scholarship, by the Junior Research Associate (JRA) program of RIKEN, and by Microsoft Research Asia. TI was supported by KAKENHI Grant Number 22K17946. MS was supported by Institute for AI and Beyond, UTokyo, and by a grant from Apple, Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple Inc.

Impact Statement

This paper presents work that aims to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel need to be highlighted here.

References

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Blanchard, G., Lee, G., and Scott, C. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

Cao, Y., Feng, L., Xu, Y., An, B., Niu, G., and Sugiyama, M. Learning from similarity-confidence data. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1272–1282, 2021.

Chen, H., Liu, F., Wang, Y., Zhao, L., and Wu, H. A variational approach for learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 33*, pp. 14844–14854, 2020a.

Chen, J., Shah, V., and Kyrillidis, A. Negative sampling in semi-supervised learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1704–1714, 2020b.

Chou, Y.-T., Niu, G., Lin, H.-T., and Sugiyama, M. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1929–1938, 2020.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical Japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Coudray, O., Keribin, C., Massart, P., and Pamphile, P. Risk bounds for positive-unlabeled learning under the selected at random assumption. *Journal of Machine Learning Research*, 24(107):1–31, 2023.

Dai, S., Li, X., Zhou, Y., Ye, X., and Liu, T. GradPU: Positive-unlabeled learning via gradient penalty and positive upweighting. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pp. 7296–7303, 2023.

Deng, Q., Guo, Y., Yang, Z., Pan, H., and Chen, J. Boosting semi-supervised learning with contrastive complementary labeling. *Neural Networks*, 170:417–426, 2024.

du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 27*, pp. 703–711, 2014.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008.

Feng, L., Kaneko, T., Han, B., Niu, G., An, B., and Sugiyama, M. Learning with multiple complementary labels. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3072–3081, 2020a.

Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B., and Sugiyama, M. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems 33*, pp. 10948–10960, 2020b.

Gao, Y. and Zhang, M.-L. Discriminative complementary-label learning with weighted loss. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3587–3597, 2021.

Garg, S., Wu, Y., Smola, A. J., Balakrishnan, S., and Lipton, Z. C. Mixture proportion estimation and PU learning: A modern approach. In *Advances in Neural Information Processing Systems 34*, pp. 8532–8544, 2021.

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *Proceedings of the 31st AAAI conference on artificial intelligence*, pp. 1919–1925, 2017.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *Proceedings of the 31st Conference On Learning Theory*, pp. 297–299, 2018.

Han, J., Zeng, L., Du, L., Ding, W., and Feng, J. Rethinking precision of pseudo label: Test-time adaptation via complementary learning. *arXiv preprint arXiv:2301.06013*, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hu, P., Lipton, Z. C., Anandkumar, A., and Ramanan, D. Active learning with partial feedback. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from complementary labels. In *Advances in Neural Information Processing Systems 30*, pp. 5644–5654, 2017.

Ishida, T., Niu, G., Menon, A. K., and Sugiyama, M. Complementary-label learning for arbitrary losses and models. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2971–2980, 2019.

Jiang, Y., Xu, Q., Zhao, Y., Yang, Z., Wen, P., Cao, X., and Huang, Q. Positive-unlabeled learning with label distribution alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15345–15363, 2023.

Kim, Y., Yim, J., Yun, J., and Kim, J. NLNL: Negative learning for noisy labels. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*, pp. 101–110, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems 30*, pp. 1674–1684, 2017.

Krizhevsky, A. and Hinton, G. E. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

Li, C., Li, X., Feng, L., and Ouyang, J. Who is your right mixup partner in positive and unlabeled learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Lin, W.-I. and Lin, H.-T. Reduction from complementary-label learning to probability estimates. In *Proceedings of the 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 469–481, 2023.

Liu, S., Cao, Y., Zhang, Q., Feng, L., and An, B. Consistent complementary-label learning via order-preserving losses. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 8734–8748, 2023.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2015.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Lu, N., Zhang, T., Niu, G., and Sugiyama, M. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 1115–1125, 2020.

Lv, J., Xu, M., Feng, L., Niu, G., Geng, X., and Sugiyama, M. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6500–6510, 2020.

Ma, Q., Gao, J., Zhan, B., Guo, Y., Zhou, J., and Wang, Y. Rethinking safe semi-supervised learning: Transferring the open-set problem to a close-set one. In *Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*, pp. 16370–16379, 2023.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, 2012.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., and Sugiyama, M. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems 29*, pp. 1199–1207, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8026–8037, 2019.

Qiao, C., Xu, N., and Geng, X. Decompositional generation process for instance-dependent partial label learning. In *Proceedings of the 11th International Conference on Learning Representations*, 2023.

Ramaswamy, H. G., Scott, C., and Tewari, A. Mixture proportion estimation via kernel embeddings of distributions. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 2052–2060, 2016.

Rezaei, M., Yang, H., and Meinel, C. Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation. *Multimedia Tools and Applications*, 79(21-22):15329–15348, 2020.

Rifkin, R. and Klautau, A. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5: 101–141, 2004.

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 838–846, 2015.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proceedings of the 26th Annual Conference on Learning Theory*, pp. 489–511, 2013.

Wang, D., Feng, L., and Zhang, M. Learning from complementary labels via partial-output consistency regularization. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pp. 3075–3081, 2021.

Wang, H.-H., Lin, W.-I., and Lin, H.-T. CLCIFAR: CIFAR-derived benchmark datasets with human annotated complementary labels. *arXiv preprint arXiv:2305.08295*, 2023a.

Wang, W., Feng, L., Jiang, Y., Niu, G., Zhang, M.-L., and Sugiyama, M. Binary classification with confidence difference. In *Advances in Neural Information Processing Systems 36*, 2023b.

Wang, X., Wan, W., Geng, C., Li, S.-Y., and Chen, S. Beyond myopia: Learning from positive and unlabeled data through holistic predictive trends. In *Advances in Neural Information Processing Systems 36*, 2023.

Wei, H., Zhuang, H., Xie, R., Feng, L., Niu, G., An, B., and Li, Y. Mitigating memorization of noisy labels by clipping the model prediction. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 36868–36886, 2023a.

Wei, M., Zhou, Y., Li, Z., and Xu, X. Class-imbalanced complementary-label learning via weighted loss. *Neural Networks*, 166:555–565, 2023b.

Wei, X.-S., Xu, H.-Y., Zhang, F., Peng, Y., and Zhou, W. An embarrassingly simple approach to semi-supervised few-shot learning. In *Advances in Neural Information Processing Systems 35*, pp. 14489–14500, 2022.

Wen, H., Cui, J., Hang, H., Liu, J., Wang, Y., and Lin, Z. Leveraged weighted loss for partial label learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11091–11100, 2021.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, N., Liu, B., Lv, J., Qiao, C., and Geng, X. Progressive purification for instance-dependent partial label learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 38551–38565, 2023.

Yao, Y., Liu, T., Han, B., Gong, M., Niu, G., Sugiyama, M., and Tao, D. Rethinking class-prior estimation for positive-unlabeled learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *Proceedings of the 15th European Conference on Computer Vision*, pp. 68–83, 2018.

Zhang, F., Feng, L., Han, B., Liu, T., Niu, G., Qin, T., and Sugiyama, M. Exploiting class activation value for partial-label learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.

Zhang, Y., Liu, F., Fang, Z., Yuan, B., Zhang, G., and Lu, J. Learning from a complementary-label source domain: Theory and algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7667–7681, 2021.

Zhang, Y.-J., Zhao, P., Ma, L., and Zhou, Z.-H. An unbiased risk estimator for learning with augmented classes. In *Advances in Neural Information Processing Systems 33*, pp. 10247–10258, 2020.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31*, pp. 8792–8802, 2018.

Zhao, Y., Xu, Q., Jiang, Y., Wen, P., and Huang, Q. Dist-PU: Positive-unlabeled learning from a label distribution perspective. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14441–14450, 2022.

Zhou, J., Zhu, J., Zhang, J., Liu, T., Niu, G., Han, B., and Sugiyama, M. Adversarial training with complementary labels: On the benefit of gradually informative attacks. In *Advances in Neural Information Processing Systems 35*, pp. 23621–23633, 2022.

A. Class-Prior Estimation

When the class priors π_k are not accessible to the learning algorithm, they can be estimated by off-the-shelf mixture proportion estimation approaches (Ramaswamy et al., 2016; Scott, 2015; Garg et al., 2021; Yao et al., 2022). In this section, we discuss the problem formulation and how to adapt a state-of-the-art class-prior estimation method to our problem as an example.

Mixture proportion estimation. Let F be a mixture distribution of two component distributions G and H with a proportion θ^*, i.e.,

$$F = (1-\theta^*)G + \theta^* H.$$

The task of mixture proportion estimation problems is to estimate θ^* given training examples sampled from F and H. For PU learning, we consider $F = p(\boldsymbol{x})$, $G = p(\boldsymbol{x}|y=-1)$, and $H = p(\boldsymbol{x}|y=+1)$. Then, the estimation of θ^* corresponds to the estimation of the class prior $p(y=+1)$. It is shown that θ^* cannot be identified without any additional assumptions (Scott et al., 2013; Scott, 2015). Hence, various assumptions have been proposed to ensure the identifiability, including the irreducibility assumption (Scott et al., 2013), the anchor point assumption (Scott, 2015; Liu & Tao, 2015), the separability assumption (Ramaswamy et al., 2016), etc.

Best Bin Estimation. We use Best Bin Estimation (BBE) (Garg et al., 2021) as the base algorithm for class-prior estimation because it can achieve nice performance with simple implementations. First, they split the PU data into PU training data $\mathcal{D}^{\mathrm{PTr}} = \left\{\left(\boldsymbol{x}_i^{\mathrm{PTr}}, +1\right)\right\}_{i=1}^{n^{\mathrm{PTr}}}$ and $\mathcal{D}^{\mathrm{UTr}} = \left\{\boldsymbol{x}_i^{\mathrm{UTr}}\right\}_{i=1}^{n^{\mathrm{UTr}}}$, and PU validation data $\mathcal{D}^{\mathrm{PVal}} = \left\{\left(\boldsymbol{x}_i^{\mathrm{PVal}}, +1\right)\right\}_{i=1}^{n^{\mathrm{PVal}}}$ and $\mathcal{D}^{\mathrm{UVal}} = \left\{\boldsymbol{x}_i^{\mathrm{UVal}}\right\}_{i=1}^{n^{\mathrm{UVal}}}$. Then, they train a positive-versus-unlabeled (PvU) classifier f^{PvU} with $\mathcal{D}^{\mathrm{PTr}}$ and $\mathcal{D}^{\mathrm{UTr}}$. They collect the model outputs of the PU validation data $\mathcal{Z}^{\mathrm{P}} = \left\{z_i^{\mathrm{P}}\right\}_{i=1}^{n^{\mathrm{PVal}}}$ and $\mathcal{Z}^{\mathrm{U}} = \left\{z_i^{\mathrm{U}}\right\}_{i=1}^{n^{\mathrm{UVal}}}$ where $z_i^{\mathrm{P}} = f^{\mathrm{PvU}}\left(\boldsymbol{x}_i^{\mathrm{PVal}}\right)$ and $z_i^{\mathrm{U}} = f^{\mathrm{PvU}}\left(\boldsymbol{x}_i^{\mathrm{UVal}}\right)$. They also introduce $q(z) = \int_{A_z} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ where $A_z = \left\{\boldsymbol{x} \in \mathcal{X} | f^{\mathrm{PvU}}(\boldsymbol{x}) \geq z\right\}$. Then, $q(z)$ can be regarded as the proportion of data with the model output not less than z. For $p\left(\boldsymbol{x}|y=+1\right)$ and $p\left(\boldsymbol{x}\right)$, they define $q^{\mathrm{P}}(z)$ and $q^{\mathrm{U}}(z)$ respectively. They empirically estimate them as

$$\widehat{q}^{\mathrm{P}}(z) = \frac{\sum_{i=1}^{n^{\mathrm{PVal}}} \mathbb{I}\left(f^{\mathrm{PvU}}\left(\boldsymbol{x}_i^{\mathrm{PVal}}\right) \geq z\right)}{n^{\mathrm{PVal}}} \quad \text{and} \quad \widehat{q}^{\mathrm{U}}(z) = \frac{\sum_{i=1}^{n^{\mathrm{UVal}}} \mathbb{I}\left(f^{\mathrm{PvU}}\left(\boldsymbol{x}_i^{\mathrm{UVal}}\right) \geq z\right)}{n^{\mathrm{UVal}}}. \tag{21}$$

Then, they obtain $\widehat{z}$ as

$$\widehat{z} = \underset{z\in[0,1]}{\arg\max}\left(\frac{\widehat{q}^{\mathrm{U}}(z)}{\widehat{q}^{\mathrm{P}}(z)} + \frac{1+\gamma}{\widehat{q}^{\mathrm{P}}(z)}\left(\sqrt{\frac{\ln\left(4/\delta\right)}{2n^{\mathrm{PVal}}}} + \sqrt{\frac{\ln\left(4/\delta\right)}{2n^{\mathrm{UVal}}}}\right)\right) \tag{22}$$

where γ and δ are hyperparameters respectively. Finally, they calculate the estimation value of the mixture proportion as

$$\widehat{\theta} = \frac{\widehat{q}^{\mathrm{U}}\left(\widehat{z}\right)}{\widehat{q}^{\mathrm{P}}\left(\widehat{z}\right)} \tag{23}$$

and they prove that $\widehat{\theta}$ is an unbiased estimator of θ^* when satisfying the *pure positive bin assumption*, a variant of the irreducibility assumption. More detailed descriptions of the approach can be found in Garg et al. (2021).

Class-prior estimation for SCARCE. Our class-prior estimation approach is based on BBE. First, we split the complementary-label data into training and validation data. Then, we generate q negative binary-class datasets $\mathcal{D}_k^{\mathrm{NTr}}$ and q unlabeled binary-class datasets $\mathcal{D}_k^{\mathrm{UTr}}$ by Eq. (10) and Eq. (11) with training data ($k \in \mathcal{Y}$). We also generate q negative binary-class datasets $\mathcal{D}_k^{\mathrm{NVal}}$ and q unlabeled binary-class datasets $\mathcal{D}_k^{\mathrm{UVal}}$ by Eq. (10) and Eq. (11) with validation data ($k \in \mathcal{Y}$). Then, we estimate the class priors $(1-\pi_k)$ for each label $k \in \mathcal{Y}$ using BBE adapted by swapping the positive and negative classes. Finally, we normalize π_k to ensure that they sum up to one. The details of the algorithm are summarized in Algorithm 3.

Experimental results. We assumed that the ground-truth class priors for all labels and datasets were 0.1, meaning that the test set was balanced. We generated complementary labels using the SCAR assumption with $\bar{\pi}_k = 0.5$. We repeated the generation process 5 times with different random seeds. Table 6 shows the experimental results of the proposed class-prior estimation approach. We can observe that the class priors are generally accurately estimated with the proposed class-prior estimation method.

Algorithm 3 Class-prior Estimation

Input: Complementary-label training set $\mathcal{D}$.

for $k \in \mathcal{Y}$ **do**

Generate training datasets $\mathcal{D}_k^{\mathrm{NTr}}$, $\mathcal{D}_k^{\mathrm{UTr}}$, validation data $\mathcal{D}_k^{\mathrm{NVal}}$, and $\mathcal{D}_k^{\mathrm{UVal}}$ by Eq. (10) and Eq. (11);

Estimate the value of $(1-\pi_k)$ by employing the BBE algorithm and interchanging the positive and negative classes;

end for

Normalize π_k to ensure they sum up to one;

Output: Class priors π_k ($k \in \mathcal{Y}$).

Table 6. Estimated values (mean±std) of class priors.

Label Index	1	2	3	4	5
MNIST	0.104±0.011	0.119±0.012	0.110±0.009	0.099±0.008	0.101±0.010
Kuzushiji-MNIST	0.108±0.026	0.098±0.011	0.087±0.012	0.104±0.004	0.101±0.021
Fashion-MNIST	0.091±0.016	0.118±0.005	0.090±0.024	0.090±0.009	0.077±0.020
CIFAR-10	0.085±0.016	0.102±0.039	0.073±0.019	0.109±0.047	0.100±0.031
Label Index	6	7	8	9	10
MNIST	0.087±0.007	0.089±0.005	0.106±0.019	0.091±0.008	0.096±0.016
Kuzushiji-MNIST	0.095±0.010	0.105±0.025	0.095±0.007	0.094±0.016	0.113±0.035
Fashion-MNIST	0.117±0.007	0.070±0.010	0.114±0.023	0.117±0.016	0.117±0.016
CIFAR-10	0.098±0.013	0.115±0.023	0.120±0.033	0.097±0.041	0.100±0.013

B. Discussions

B.1. Related Work on PU Learning

PU learning methods can be roughly categorized into two groups: sample selection methods and cost-sensitive methods. Sample selection methods try to identify reliable negative examples from the unlabeled dataset and then use supervised learning methods to learn the classifier (Wang et al., 2023; Dai et al., 2023; Garg et al., 2021). Cost-sensitive methods are based on an unbiased risk estimator, which rewrites the classification risk as that only on positive and unlabeled data (Kiryo et al., 2017; Jiang et al., 2023; Zhao et al., 2022).

B.2. Data Generation Process

The generation process of complementary labels is summarized in Algorithm 4.

B.3. Use of Ordinary-Label Data

If an additional ordinary-label training set $\mathcal{D}^{\mathrm{O}} = \{(\boldsymbol{x}_i^{\mathrm{O}}, y_i^{\mathrm{O}})\}_{i=1}^{n^{\mathrm{O}}}$ consisting of n^{O} training examples is available, we can include it in the training as well. Let $\widehat{R}^{\mathrm{O}}(f_1, f_2, \ldots, f_q) = \sum_{i=1}^{n^{\mathrm{O}}} \left(\ell\left(f_{y_i^{\mathrm{O}}}(\boldsymbol{x}_i^{\mathrm{O}})\right) + \sum_{k \in \mathcal{Y} \setminus \{y_i^{\mathrm{O}}\}} \ell\left(-f_k(\boldsymbol{x}_i^{\mathrm{O}})\right) \right) / n^{\mathrm{O}}$ denote the empirical risk estimator calculated on $\mathcal{D}^{\mathrm{O}}$, it is easy to see that $\alpha \widehat{R}(f_1, f_2, \ldots, f_q) + (1-\alpha) \widehat{R}^{\mathrm{O}}(f_1, f_2, \ldots, f_q)$ is still an *unbiased risk estimator* w.r.t. Eq. (5) where $\alpha \in [0, 1]$ is a weight.

B.4. Summary of Assumptions

In this section, we briefly summarize the assumptions used in this paper.

- The SCAR assumption, which is used to describe the data generation process.
- The irreducibility assumption, which is used to allow the estimation of class priors.
- The assumptions used to prove the classification-calibrated property in Theorem 5, i.e. ℓ is convex, bounded below, differential, and satisfies $\ell(z) \leq \ell(-z)$ when $z > 0$.
- The assumptions used to prove the estimation error bounds in Theorems 7, 9, and 10, e.g., $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq C_f$,

Algorithm 4 Data Generation Process

Input: Marginal density $p(\boldsymbol{x})$, Posterior probability distribution $p(y|\boldsymbol{x})$, number of training data n, probabilities $c_k (k \in \mathcal{Y})$.

```
Initialize a complementary-label dataset D = ∅;
for i = 1, 2, ..., n do
    Sample an unlabeled example x_i from p(x);
    Sample the label y_i from p(y|x_i);
    Initialize Ȳ_i = ∅;
    for k ∈ Y do
        if y_i ≠ k then
            Assign Ȳ_i = Ȳ_i ∪ {k} with c_k;
        end if
    end for
    Assign D = D ∪ {(x_i, Ȳ_i)};
end for
```

Output: Complementary-label dataset $\mathcal{D}$.

$\sup_{|z| \leq C_f} \ell(z) \leq C_\ell$, $\ell(z)$ is Lipschitz continuous, $\forall k \in \mathcal{Y}, \mathbb{E}\left[\widehat{R}_k^{\mathrm{P}}(f_k)\right] \geq \beta$, the risk-correction function $g(z)$ is Lipschitz continuous, $\mathfrak{R}_{n,p}(\mathcal{F}) \leq C_{\mathfrak{R}}/\sqrt{n}$, etc.

B.5. Limitations

Our method is designed for instance-independent settings, where the generation of complementary labels is determined only by the ground-truth label and is not related to the feature. Moreover, if we adopt some PU learning methods that cannot share the representation layers for different labels, we may need additional storage and computational cost.

C. Proof of Theorem 2

First, we introduce the following lemma.

Lemma 12. *Under Assumption 1, we have* $p\left(\boldsymbol{x}|\bar{y}_k = 1\right) = p(\boldsymbol{x}|y \neq k)$.

Proof. On one hand, we have

$$p(\boldsymbol{x}|\bar{y}_k = 1, y \neq k) = \frac{p\left(\boldsymbol{x}|\bar{y}_k = 1\right) p(y \neq k|\boldsymbol{x}, \bar{y}_k = 1)}{p(y \neq k|\bar{y}_k = 1)}.$$

According to the definition of complementary labels, we have $p(y \neq k|\boldsymbol{x}, \bar{y}_k = 1) = p(y \neq k|\bar{y}_k = 1) = 1$. Therefore, we have $p(\boldsymbol{x}|\bar{y}_k = 1, y \neq k) = p\left(\boldsymbol{x}|\bar{y}_k = 1\right)$. On the other hand, we have

$$p(\boldsymbol{x}|\bar{y}_k = 1, y \neq k) = \frac{p(\boldsymbol{x}|y \neq k) p(\bar{y}_k = 1|\boldsymbol{x}, y \neq k)}{p(\bar{y}_k = 1|y \neq k)} = p(\boldsymbol{x}|y \neq k),$$

where the first equation is derived from Assumption 1. The proof is completed. □

Then, the proof of Theorem 2 is given.

Proof of Theorem 2.

$$
\begin{aligned}
R(f) =& \mathbb{E}_{p(\boldsymbol{x},y)}\left[\mathcal{L}(f(\boldsymbol{x}),y)\right] \\
=& \sum_{k=1}^{q}\left(\pi_k \mathbb{E}_{p(\boldsymbol{x}|y=k)}\left[\mathcal{L}(f(\boldsymbol{x}),k)\right]\right) \\
=& \sum_{k=1}^{q}\left(\mathbb{E}_{p(\boldsymbol{x})}\left[\mathcal{L}(f(\boldsymbol{x}),k)\right] - (1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|y\neq k)}\left[\mathcal{L}(f(\boldsymbol{x}),k)\right]\right) \\
=& \sum_{k=1}^{q}\left(\mathbb{E}_{p(\boldsymbol{x})}\left[\mathcal{L}(f(\boldsymbol{x}),k)\right] - (1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[\mathcal{L}(f(\boldsymbol{x}),k)\right]\right) \\
=& \sum_{k=1}^{q}\left(\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[\left(\bar{\pi}_k+\pi_k-1\right)\mathcal{L}(f(\boldsymbol{x}),k)\right] + \mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=0)}\left[\left(1-\bar{\pi}_k\right)\mathcal{L}(f(\boldsymbol{x}),k)\right]\right),
\end{aligned}
$$

which concludes the proof. □

D. Proof of Theorem 3

Proof.

$$
\begin{aligned}
R(f_1,f_2,\ldots,f_q) =& \mathbb{E}_{p(\boldsymbol{x},y)}\left[\ell\left(f_y\left(\boldsymbol{x}\right)\right) + \sum_{k=1,k\neq y}^{q}\ell\left(-f_k\left(\boldsymbol{x}\right)\right)\right] \\
=& \mathbb{E}_{p(\boldsymbol{x},y)}\left[\sum_{k=1}^{q}\left(\mathbb{I}(k=y)\ell(f_k(\boldsymbol{x})) + \mathbb{I}(k\neq y)\ell(-f_k(\boldsymbol{x}))\right)\right] \\
=& \sum_{k=1}^{q}\mathbb{E}_{p(\boldsymbol{x},y)}\left[\mathbb{I}(k=y)\ell\left(f_k(\boldsymbol{x})\right) + \mathbb{I}(k\neq y)\ell\left(-f_k(\boldsymbol{x})\right)\right] \\
=& \sum_{k=1}^{q}\left(\pi_k\mathbb{E}_{p(\boldsymbol{x}|y=k)}\left[\ell\left(f_k(\boldsymbol{x})\right)\right] + \left(1-\pi_k\right)\mathbb{E}_{p(\boldsymbol{x}|y\neq k)}\left[\ell\left(-f_k(\boldsymbol{x})\right)\right]\right) \\
=& \sum_{k=1}^{q}\left(\mathbb{E}_{p(\boldsymbol{x})}\left[\ell(f_k(\boldsymbol{x}))\right] - (1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|y\neq k)}\left[\ell(f_k(\boldsymbol{x}))\right]\right. \\
& \left.+(1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|y\neq k)}\left[\ell(-f_k(\boldsymbol{x}))\right]\right) \\
=& \sum_{k=1}^{q}\left(\mathbb{E}_{p(\boldsymbol{x})}\left[\ell(f_k(\boldsymbol{x}))\right] - (1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[\ell(f_k(\boldsymbol{x}))\right]\right. \\
& \left.+(1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[\ell(-f_k(\boldsymbol{x}))\right]\right) \\
=& \sum_{k=1}^{q}\left(\bar{\pi}_k\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[\ell(f_k(\boldsymbol{x}))\right] + (1-\bar{\pi}_k)\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=0)}\left[\ell(f_k(\boldsymbol{x}))\right]\right. \\
& \left.-(1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[\ell(f_k(\boldsymbol{x}))\right] + (1-\pi_k)\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[\ell(-f_k(\boldsymbol{x}))\right]\right) \\
=& \sum_{k=1}^{q}\left(\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=1)}\left[(1-\pi_k)\ell\left(-f_k(\boldsymbol{x})\right) + \left(\bar{\pi}_k+\pi_k-1\right)\ell\left(f_k(\boldsymbol{x})\right)\right]\right. \\
& \left.+\mathbb{E}_{p(\boldsymbol{x}|\bar{y}_k=0)}\left[\left(1-\bar{\pi}_k\right)\ell\left(f_k(\boldsymbol{x})\right)\right]\right).
\end{aligned}
$$

Here, $\mathbb{I}(\cdot)$ returns 1 if predicate holds. Otherwise, $\mathbb{I}(\cdot)$ returns 0. The proof is completed. □

E. Proof of Lemma 4

Proof. Under the binary classification setting, we have the marginal density $p(\boldsymbol{x}) = \pi_+ p(\boldsymbol{x}|y=+1) + (1-\pi_+)p(\boldsymbol{x}|y=-1)$, where $p(\boldsymbol{x}|y=+1)$ and $p(\boldsymbol{x}|y=-1)$ denote the densities of positive and negative data respectively. Then, we have

the classification risk of binary classification

$$\begin{aligned}
R(f) =& \mathbb{E}_{p(\boldsymbol{x},y)}\left[\ell\left(yf^{\mathrm{NU}}(\boldsymbol{x})\right)\right] \\
=& \pi_+\mathbb{E}_{p(\boldsymbol{x}|y=+1)}\left[\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] + (1-\pi_+)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\left[\ell\left(-f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] \\
=& \mathbb{E}_{p(\boldsymbol{x})}\left[\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] - (1-\pi_+)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\left[\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] + (1-\pi_+)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\left[\ell\left(-f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] \\
=& \bar{\pi}_-\mathbb{E}_{p(\boldsymbol{x}|\bar{y}=-1)}\left[\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] + (1-\bar{\pi}_-)\mathbb{E}_{p(\boldsymbol{x}|\bar{y}=0)}\left[\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] \\
&+ (1-\pi_+)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\left[\ell\left(-f^{\mathrm{NU}}(\boldsymbol{x})\right) - \ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] \\
=& \mathbb{E}_{p(\boldsymbol{x}|\bar{y}=-1)}\left[(1-\pi_+)\ell\left(-f^{\mathrm{NU}}(\boldsymbol{x})\right) + (\bar{\pi}_- + \pi_+ - 1)\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right] + \mathbb{E}_{p(\boldsymbol{x}|\bar{y}=0)}\left[(1-\bar{\pi}_-)\ell\left(f^{\mathrm{NU}}(\boldsymbol{x})\right)\right],
\end{aligned}$$

where the last equation results from Lemma 12 under the binary-class setting. The proof is completed. □

F. Proof of Theorem 5

To begin with, we show the following theoretical results about infinite-sample consistency from Zhang (2004). For ease of notations, let $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_q(\boldsymbol{x})]$ denote the vector form of modeling outputs of all the binary classifiers. First, we elaborate the infinite-sample consistency property of the OVR strategy.

Theorem 13 (Theorem 10 of Zhang (2004))**.** *Consider the OVR strategy, whose surrogate loss function is defined as* $\Psi_y\left(\boldsymbol{f}(\boldsymbol{x})\right) = \ell\left(f_y(\boldsymbol{x})\right) + \sum_{k\in\mathcal{Y}\setminus\{y\}}\ell\left(-f_k(\boldsymbol{x})\right)$*. Assume* ℓ *is convex, bounded below, differentiable, and* $\ell(z) < \ell(-z)$ *when* $z > 0$*. Then, the OVR strategy is infinite-sample consistent on* $\Omega = \mathbb{R}^K$ *with respect to the 0-1 classification risk.*

Then, we elaborate the relationship between the minimum classification risk of an infinite-sample consistent method and the Bayes error.

Theorem 14 (Theorem 3 of Zhang (2004))**.** *Let* $\mathcal{B}$ *be the set of all vector Borel measurable functions, which take values in* $\mathbb{R}^q$*. For* $\Omega \subset \mathbb{R}^q$*, let* $\mathcal{B}_\Omega = \{\boldsymbol{f} \in \mathcal{B} : \forall \boldsymbol{x}, \boldsymbol{f}(x) \in \Omega\}$*. If* $\Psi_y(\cdot)$ *is infinite-sample consistent on* Ω *with respect to the 0-1 classification risk, then for any* $\epsilon_1 > 0$*, there exists an* $\epsilon_2 > 0$ *such that for all underlying Borel probability measurable* p*, and* $\boldsymbol{f}(\cdot) \in \mathcal{B}_\Omega$*,*

$$\mathbb{E}_{(\boldsymbol{x},y)\sim p}\left[\Psi_y(\boldsymbol{f}(\boldsymbol{x}))\right] \leq \inf_{\boldsymbol{f}'\in\mathcal{B}_\Omega}\mathbb{E}_{(\boldsymbol{x},y)\sim p}\left[\Psi_y(\boldsymbol{f}'(\boldsymbol{x}))\right] + \epsilon_2 \tag{24}$$

implies

$$R_{0-1}\left(T(\boldsymbol{f}(\cdot))\right) \leq R^*_{0-1} + \epsilon_1, \tag{25}$$

where $T(\cdot)$ *is defined as* $T(\boldsymbol{f}(\boldsymbol{x})) := \arg\max_{k=1,\ldots,q} f_k(\boldsymbol{x})$*.*

Then, we give the proof of Theorem 5.

Proof of Theorem 5. According to Theorem 3, the proposed classification risk $R(f_1, f_2, \ldots, f_q)$ is equivalent to the OVR risk. Therefore, it is sufficient to elaborate the theoretical properties of the OVR risk to prove Theorem 5. □

G. Proof of Theorem 7

First, we give the definition of Rademacher complexity.

Definition 15 (Rademacher complexity)**.** Let $\mathcal{X}_n = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_n\}$ denote n i.i.d. random variables drawn from a probability distribution with density $p(\boldsymbol{x})$, $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$ denote a class of measurable functions, and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ denote Rademacher variables taking values from $\{+1, -1\}$ uniformly. Then, the (expected) Rademacher complexity of $\mathcal{F}$ is defined as

$$\mathfrak{R}_{n,p}(\mathcal{F}) = \mathbb{E}_{\mathcal{X}_n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right]. \tag{26}$$

For ease of notation, we define $\bar{\mathcal{D}} = \mathcal{D}^{\mathrm{U}}_1 \bigcup \mathcal{D}^{\mathrm{U}}_2 \bigcup \ldots \bigcup \mathcal{D}^{\mathrm{U}}_q \bigcup \mathcal{D}^{\mathrm{N}}_1 \bigcup \mathcal{D}^{\mathrm{N}}_2 \bigcup \ldots \bigcup \mathcal{D}^{\mathrm{N}}_q$ denote the set of all the binary-class training data. Then, we have the following lemma.

Lemma 16. *For any $\delta > 0$, the inequalities below hold with probability at least $1-\delta$:*

$$\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right| \leq \sum_{k=1}^{q}\left((1-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}}\right.$$
$$\left.+(2-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+(4-4\pi_k-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})+(2-2\pi_k-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}\right). \tag{27}$$

Proof. In the following proofs, we consider a general case where all the datasets $\mathcal{D}_k^{\mathrm{N}}$ and $\mathcal{D}_k^{\mathrm{U}}$ are *mutually independent*. We can observe that when an unlabeled example $\boldsymbol{x}_{k,i}^{\mathrm{U}} \in \mathcal{D}_k^{\mathrm{U}}$ is substituted by another unlabeled example $\boldsymbol{x}_{k,j}^{\mathrm{U}}$, the value of $\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right|$ changes at most $(1-\bar{\pi}_k)C_\ell/n_k^{\mathrm{U}}$. Besides, when a negative example $\boldsymbol{x}_{k,i}^{\mathrm{N}} \in \mathcal{D}_k^{\mathrm{N}}$ is substituted by another negative example $\boldsymbol{x}_{k,j}^{\mathrm{N}}$, the value of $\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right|$ changes at most $(2-2\pi_k-\bar{\pi}_k)C_\ell/n_k^{\mathrm{N}}$. According to the McDiarmid's inequality, for any $\delta > 0$, the following inequality holds with probability at least $1-\delta/2$:

$$\begin{aligned}&\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left(R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right)\\ \leq&\mathbb{E}_{\bar{\mathcal{D}}}\left[\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left(R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right)\right]\\ &+\sum_{k=1}^{q}\left((1-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}}+(2-2\pi_k-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}\right),\end{aligned} \tag{28}$$

where the inequality is deduced since $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$. It is a routine work to show by symmetrization (Mohri et al., 2012) that

$$\begin{aligned}&\mathbb{E}_{\bar{\mathcal{D}}}\left[\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left(R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right)\right]\\ \leq&\sum_{k=1}^{q}\left((2-2\bar{\pi}_k)\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\ell\circ\mathcal{F})+(4-4\pi_k-2\bar{\pi}_k)\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\ell\circ\mathcal{F})\right),\end{aligned} \tag{29}$$

where $\mathfrak{R}_{n,p}(\ell\circ\mathcal{F})$ is the Rademacher complexity of the composite function class $(\ell\circ\mathcal{F})$. According to Talagrand's contraction lemma (Ledoux & Talagrand, 1991), we have

$$\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\ell\circ\mathcal{F}) \leq L_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F}), \tag{30}$$
$$\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\ell\circ\mathcal{F}) \leq L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F}). \tag{31}$$

By combining Inequality (28), Inequality (29), Inequality (30), and Inequality (31), the following inequality holds with probability at least $1-\delta/2$:

$$\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left(R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right) \leq \sum_{k=1}^{q}\left((2-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})\right.$$
$$\left.+(1-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}}+(4-4\pi_k-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})+(2-2\pi_k-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}\right). \tag{32}$$

In the same way, we have the following inequality with probability at least $1-\delta/2$:

$$\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left(\widehat{R}(f_1,f_2,\ldots,f_q)-R(f_1,f_2,\ldots,f_q)\right) \leq \sum_{k=1}^{q}\left((1-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}}\right.$$
$$\left.+(2-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+(4-4\pi_k-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})+(2-2\pi_k-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}\right). \tag{33}$$

By combining Inequality (32) and Inequality (33), we have the following inequality with probability at least $1-\delta$:

$$\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right|\leq\sum_{k=1}^{q}\left((1-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}}\right.$$
$$\left.+(2-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+(4-4\pi_k-2\bar{\pi}_k)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})+(2-2\pi_k-\bar{\pi}_k)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}\right),\tag{34}$$

which concludes the proof. □

Proof of Theorem 7.

$$\begin{aligned}&R(\widehat{f}_1,\widehat{f}_2,\ldots,\widehat{f}_q)-R(f_1^*,f_2^*,\ldots,f_q^*)\\=&R(\widehat{f}_1,\widehat{f}_2,\ldots,\widehat{f}_q)-\widehat{R}(\widehat{f}_1,\widehat{f}_2,\ldots,\widehat{f}_q)+\widehat{R}(\widehat{f}_1,\widehat{f}_2,\ldots,\widehat{f}_q)-\widehat{R}(f_1^*,f_2^*,\ldots,f_q^*)\\&+\widehat{R}(f_1^*,f_2^*,\ldots,f_q^*)-R(f_1^*,f_2^*,\ldots,f_q^*)\\\leq&R(\widehat{f}_1,\widehat{f}_2,\ldots,\widehat{f}_q)-\widehat{R}(\widehat{f}_1,\widehat{f}_2,\ldots,\widehat{f}_q)+\widehat{R}(f_1^*,f_2^*,\ldots,f_q^*)-R(f_1^*,f_2^*,\ldots,f_q^*)\\\leq&2\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right|\end{aligned}\tag{35}$$

The first inequality is deduced because $(\widehat{f}_1,\widehat{f}_2,\ldots,\widehat{f}_q)$ is the minimizer of $\widehat{R}(f_1,f_2,\ldots,f_q)$. Combining Inequality (35) and Lemma 16, the proof is completed. □

H. Proof of Theorem 9

Let $\mathfrak{D}_k^+(f_k)=\left\{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)|\widehat{R}_k^{\mathrm{P}}(f_k)\geq 0\right\}$ and $\mathfrak{D}_k^-(f_k)=\left\{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)|\widehat{R}_k^{\mathrm{P}}(f_k)<0\right\}$ denote the sets of NU data pairs having positive and negative empirical risk respectively. Then we have the following lemma.

Lemma 17. *The probability measure of $\mathfrak{D}_k^-(f_k)$ can be bounded as follows:*

$$\mathbb{P}\left(\mathfrak{D}_k^-(f_k)\right)\leq\exp\left(\frac{-2\beta^2}{(1-\pi_k-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{N}}+(1-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{U}}}\right).\tag{36}$$

Proof. Let

$$p\left(\mathcal{D}_k^{\mathrm{N}}\right)=p\left(\boldsymbol{x}_{k,1}^{\mathrm{N}}|\bar{y}_k=1\right)p\left(\boldsymbol{x}_{k,2}^{\mathrm{N}}|\bar{y}_k=1\right)\ldots p\left(\boldsymbol{x}_{k,n_k^{\mathrm{N}}}^{\mathrm{N}}|\bar{y}_k=1\right)$$

and

$$p\left(\mathcal{D}_k^{\mathrm{U}}\right)=p\left(\boldsymbol{x}_{k,1}^{\mathrm{U}}|\bar{y}_k=0\right)p\left(\boldsymbol{x}_{k,2}^{\mathrm{U}}|\bar{y}_k=0\right)\ldots p\left(\boldsymbol{x}_{k,n_k^{\mathrm{U}}}^{\mathrm{U}}|\bar{y}_k=0\right)$$

denote the probability density of $\mathcal{D}_k^{\mathrm{N}}$ and $\mathcal{D}_k^{\mathrm{U}}$ respectively. The joint probability density of $\mathcal{D}_k^{\mathrm{N}}$ and $\mathcal{D}_k^{\mathrm{U}}$ is

$$p\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)=p\left(\mathcal{D}_k^{\mathrm{N}}\right)p\left(\mathcal{D}_k^{\mathrm{U}}\right).$$

Then, the probability measure $\mathbb{P}\left(\mathfrak{D}_k^-(f_k)\right)$ is defined as

$$\begin{aligned}\mathbb{P}\left(\mathfrak{D}_k^-(f_k)\right)&=\int_{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\in\mathfrak{D}_k^-(f_k)}p\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\mathrm{d}\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\\&=\int_{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\in\mathfrak{D}_k^-(f_k)}p\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\mathrm{d}\boldsymbol{x}_{k,1}^{\mathrm{N}}\ldots\mathrm{d}\boldsymbol{x}_{k,n_k^{\mathrm{N}}}^{\mathrm{N}}\,\mathrm{d}\boldsymbol{x}_{k,1}^{\mathrm{U}}\ldots\mathrm{d}\boldsymbol{x}_{k,n_k^{\mathrm{U}}}^{\mathrm{U}}.\end{aligned}$$

When a negative example in $\mathcal{D}_k^{\mathrm{N}}$ is substituted by another different negative example, the change of the value of $\widehat{R}_k^{\mathrm{P}}(f_k)$ is no more than $(1-\pi_k-\bar{\pi}_k)C_\ell/n_k^{\mathrm{N}}$; when an unlabeled example in $\mathcal{D}_k^{\mathrm{U}}$ is substituted by another different unlabeled

example, the change of the value of $\widehat{R}_k^{\mathrm{P}}(f_k)$ is no more than $(1-\bar{\pi}_k)C_\ell/n_k^{\mathrm{U}}$. Therefore, by applying the McDiarmid's inequality, we can obtain the following inequality:

$$\mathbb{P}\left(\mathbb{E}\left[\widehat{R}_k^{\mathrm{P}}(f_k)\right]-\widehat{R}_k^{\mathrm{P}}(f_k)\geq\beta\right)\leq\exp\left(\frac{-2\beta^2}{(1-\pi_k-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{N}}+(1-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{U}}}\right). \tag{37}$$

Therefore, we have

$$\begin{aligned}\mathbb{P}\left(\mathfrak{D}_k^-(f_k)\right)=&\mathbb{P}\left(\widehat{R}_k^{\mathrm{P}}(f_k)\leq 0\right)\\ \leq&\mathbb{P}\left(\widehat{R}_k^{\mathrm{P}}(f_k)\leq\mathbb{E}\left[\widehat{R}_k^{\mathrm{P}}(f_k)\right]-\beta\right)\\ =&\mathbb{P}\left(\mathbb{E}\left[\widehat{R}_k^{\mathrm{P}}(f_k)\right]-\widehat{R}_k^{\mathrm{P}}(f_k)\geq\beta\right)\\ \leq&\exp\left(\frac{-2\beta^2}{(1-\pi_k-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{N}}+(1-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{U}}}\right),\end{aligned} \tag{38}$$

which concludes the proof. □

We present a more complete version of Theorem 9 and its proof.

Theorem 18. *Based on the above assumptions, the bias of the expectation of the corrected risk estimator has the following lower and upper bounds:*

$$0\leq\mathbb{E}[\widetilde{R}(f_1,f_2,\ldots,f_q)]-R(f_1,f_2,\ldots,f_q)\leq\sum_{k=1}^{q}\left(2-2\bar{\pi}_k-\pi_k\right)\left(L_g+1\right)C_\ell\Delta_k, \tag{39}$$

where $\Delta_k=\exp\left(-2\beta^2/\left((1-\pi_k-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{N}}+(1-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{U}}\right)\right)$. *Furthermore, for any* $\delta>0$*, the following inequality holds with probability at least* $1-\delta$*:*

$$\begin{aligned}|\widetilde{R}(f_1,f_2,\ldots,f_q)-R(f_1,f_2,\ldots,f_q)|\leq\sum_{k=1}^{q}\Bigg(&\left(1-\bar{\pi}_k\right)C_\ell L_g\sqrt{\frac{\ln\left(2/\delta\right)}{2n_k^{\mathrm{U}}}}\\ &+\left(2-2\bar{\pi}_k-\pi_k\right)\left(L_g+1\right)C_\ell\Delta_k+\left(\left(1-\pi_k-\bar{\pi}_k\right)L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln\left(2/\delta\right)}{2n_k^{\mathrm{N}}}}\Bigg).\end{aligned}$$

Proof. First, we have

$$\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-R(f_1,f_2,\ldots,f_q)=\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)-\widehat{R}(f_1,f_2,\ldots,f_q)\right].$$

Since $\widetilde{R}(f_1,f_2,\ldots,f_q)$ is an upper bound of $\widehat{R}(f_1,f_2,\ldots,f_q)$, we have

$$\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-R(f_1,f_2,\ldots,f_q)\geq 0.$$

Besides, we have

$$\begin{aligned}&\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-R(f_1,f_2,\ldots,f_q)\\ =&\sum_{k=1}^{q}\int_{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\in\mathfrak{D}_k^-(f_k)}\left(g\left(\widehat{R}_k^{\mathrm{P}}(f_k)\right)-\widehat{R}_k^{\mathrm{P}}(f_k)\right)p\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\mathrm{d}\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\\ \leq&\sum_{k=1}^{q}\sup_{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\in\mathfrak{D}_k^-(f_k)}\left(g\left(\widehat{R}_k^{\mathrm{P}}(f_k)\right)-\widehat{R}_k^{\mathrm{P}}(f_k)\right)\int_{(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}^{\mathrm{U}})\in\mathfrak{D}_k^-(f_k)}p\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\mathrm{d}\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\\ =&\sum_{k=1}^{q}\sup_{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\in\mathfrak{D}_k^-(f_k)}\left(g\left(\widehat{R}_k^{\mathrm{P}}(f_k)\right)-\widehat{R}_k^{\mathrm{P}}(f_k)\right)\mathbb{P}\left(\mathfrak{D}_k^-(f_k)\right)\\ \leq&\sum_{k=1}^{q}\sup_{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\in\mathfrak{D}_k^-(f_k)}(L_g\left|\widehat{R}_k^{\mathrm{P}}(f_k)\right|+\left|\widehat{R}_k^{\mathrm{P}}(f_k)\right|)\mathbb{P}\left(\mathfrak{D}_k^-(f_k)\right).\end{aligned}$$

Besides,

$$\begin{aligned}\left|\widehat{R}_k^{\mathrm{P}}(f_k)\right| &= \left|\frac{\bar{\pi}_k+\pi_k-1}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)+\frac{1-\bar{\pi}_k}{n_k^{\mathrm{U}}}\sum_{i=1}^{n_k^{\mathrm{U}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}})\right)\right| \\ &\leq \left|\frac{\bar{\pi}_k+\pi_k-1}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)\right|+\left|\frac{1-\bar{\pi}_k}{n_k^{\mathrm{U}}}\sum_{i=1}^{n_k^{\mathrm{U}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}})\right)\right| \\ &\leq (1-\pi_k-\bar{\pi}_k)C_\ell+(1-\bar{\pi}_k)C_\ell=\left(2-2\bar{\pi}_k-\pi_k\right)C_\ell.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-R(f_1,f_2,\ldots,f_q) \\ \leq&\sum_{k=1}^{q}\sup_{\left(\mathcal{D}_k^{\mathrm{N}},\mathcal{D}_k^{\mathrm{U}}\right)\in\mathfrak{D}_k^-(f_k)}\left(L_g\left|\widehat{R}_k^{\mathrm{P}}(f_k)\right|+\left|\widehat{R}_k^{\mathrm{P}}(f_k)\right|\right)\mathbb{P}(\mathfrak{D}_k^-(f_k)). \\ \leq&\sum_{k=1}^{q}\left(2-2\bar{\pi}_k-\pi_k\right)\left(L_g+1\right)C_\ell\exp\left(\frac{-2\beta^2}{(1-\pi_k-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{N}}+(1-\bar{\pi}_k)^2C_\ell^2/n_k^{\mathrm{U}}}\right) \\ =&\sum_{k=1}^{q}\left(2-2\bar{\pi}_k-\pi_k\right)\left(L_g+1\right)C_\ell\Delta_k,\end{aligned}$$

which concludes the first part of the proof of Theorem 3. Before giving the proof for the second part, we give the upper bound of $\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|$. When an unlabeled example from $\mathcal{D}_k^{\mathrm{U}}$ is substituted by another unlabeled example, the value of $\widetilde{R}(f_1,f_2,\ldots,f_q)$ changes at most $(1-\bar{\pi}_k)\,C_\ell L_g/n_k^{\mathrm{U}}$. When a negative example from $\mathcal{D}_k^{\mathrm{N}}$ is substituted by a different example, the value of $\widetilde{R}(f_1,f_2,\ldots,f_q)$ changes at most $\left((1-\pi_k-\bar{\pi}_k)\,L_g+1-\pi_k\right)C_\ell/n_k^{\mathrm{N}}$. By applying McDiarmid's inequality, we have the following inequalities with probability at least $1-\delta/2$:

$$\begin{aligned}\widetilde{R}(f_1,f_2,\ldots,f_q)-\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]\leq&\sum_{k=1}^{q}(1-\bar{\pi}_k)\,C_\ell L_g\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}} \\ &+\sum_{k=1}^{q}\left((1-\pi_k-\bar{\pi}_k)\,L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}, \\ \mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\leq&\sum_{k=1}^{q}(1-\bar{\pi}_k)\,C_\ell L_g\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}} \\ &+\sum_{k=1}^{q}\left((1-\pi_k-\bar{\pi}_k)\,L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}.\end{aligned}$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\leq&\sum_{k=1}^{q}(1-\bar{\pi}_k)\,C_\ell L_g\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{U}}}} \\ &+\sum_{k=1}^{q}\left((1-\pi_k-\bar{\pi}_k)\,L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln(2/\delta)}{2n_k^{\mathrm{N}}}}.\end{aligned}$$

Therefore, with probability at least $1-\delta$ we have

$$\begin{aligned}
&\left|\widetilde{R}(f_1,f_2,\ldots,f_q)-R(f_1,f_2,\ldots,f_q)\right| \\
=&\left|\widetilde{R}(f_1,f_2,\ldots,f_q)-\mathbb{E}[\widetilde{R}(f_1,f_2,\ldots,f_q)]+\mathbb{E}[\widetilde{R}(f_1,f_2,\ldots,f_q)]-R(f_1,f_2,\ldots,f_q)\right| \\
\leq&\left|\widetilde{R}(f_1,f_2,\ldots,f_q)-\mathbb{E}[\widetilde{R}(f_1,f_2,\ldots,f_q)]\right|+\left|\mathbb{E}[\widetilde{R}(f_1,f_2,\ldots,f_q)]-R(f_1,f_2,\ldots,f_q)\right| \\
=&\left|\widetilde{R}(f_1,f_2,\ldots,f_q)-\mathbb{E}[\widetilde{R}(f_1,f_2,\ldots,f_q)]\right|+\mathbb{E}[\widetilde{R}(f_1,f_2,\ldots,f_q)]-R(f_1,f_2,\ldots,f_q) \\
\leq&\sum_{k=1}^{q}\left(1-\bar{\pi}_k\right)C_\ell L_g\sqrt{\frac{\ln\left(2/\delta\right)}{2n_k^{\mathrm{U}}}}+\sum_{k=1}^{q}\left(\left(1-\pi_k-\bar{\pi}_k\right)L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln\left(2/\delta\right)}{2n_k^{\mathrm{N}}}} \\
&+\sum_{k=1}^{q}\left(2-2\bar{\pi}_k-\pi_k\right)\left(L_g+1\right)C_\ell\Delta_k,
\end{aligned}$$

which concludes the proof. □

I. Proof of Theorem 10

In this section, we adopt an alternative definition of Rademacher complexity:

$$\mathfrak{R}'_{n,p}(\mathcal{F})=\mathbb{E}_{\mathcal{X}_n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)\right|\right]. \tag{40}$$

Then, we introduce the following lemmas.

Lemma 19. *Without any composition, for any $\mathcal{F}$, we have $\mathfrak{R}'_{n,p}(\mathcal{F})\geq\mathfrak{R}_{n,p}(\mathcal{F})$. If $\mathcal{F}$ is closed under negation, we have $\mathfrak{R}'_{n,p}(\mathcal{F})=\mathfrak{R}_{n,p}(\mathcal{F})$.*

Lemma 20 (Theorem 4.12 in (Ledoux & Talagrand, 1991))**.** *If $\psi:\mathbb{R}\to\mathbb{R}$ is a Lipschitz continuous function with a Lipschitz constant L_ψ and satisfies $\psi(0)=0$, we have*

$$\mathfrak{R}'_{n,p}(\psi\circ\mathcal{F})\leq 2L_\psi\mathfrak{R}'_{n,p}(\mathcal{F}),$$

where $\psi\circ\mathcal{F}=\{\psi\circ f|f\in\mathcal{F}\}$.

Before giving the proof of Theorem 10, we give the following lemma.

Lemma 21. *For any $\delta>0$, the inequalities below hold with probability at least $1-\delta$:*

$$\begin{aligned}
&\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\widetilde{R}(f_1,f_2,\ldots,f_q)\right| \\
\leq&\sum_{k=1}^{q}\left(1-\bar{\pi}_k\right)C_\ell L_g\sqrt{\frac{\ln\left(1/\delta\right)}{2n_k^{\mathrm{U}}}}+\sum_{k=1}^{q}\left(\left(1-\pi_k-\bar{\pi}_k\right)L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln\left(1/\delta\right)}{2n_k^{\mathrm{N}}}} \\
&+\sum_{k=1}^{q}\left(\left(4-4\bar{\pi}_k\right)L_gL_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+\left(\left(4-4\pi_k-4\bar{\pi}_k\right)L_g+4-4\pi_k\right)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})\right) \\
&+\sum_{k=1}^{q}\left(2-2\bar{\pi}_k-\pi_k\right)\left(L_g+1\right)C_\ell\Delta_k.
\end{aligned}$$

Proof. Similar to previous proofs, we can observe that when an unlabeled example from $\mathcal{D}_k^{\mathrm{U}}$ is substituted by another unlabeled example, the value of $\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|$ changes at most $\left(1-\bar{\pi}_k\right)C_\ell L_g/n_k^{\mathrm{U}}$. When a negative example from $\mathcal{D}_k^{\mathrm{N}}$ is substituted by a different example, the value of

$\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|$ changes at most $\left(\left(1-\pi_k-\bar{\pi}_k\right)L_g+1-\pi_k\right)C_\ell/n_k^{\mathrm{N}}$. By applying McDiarmid's inequality, we have the following inequality with probability at least $1-\delta$:

$$\begin{aligned}&\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\\&-\mathbb{E}\left[\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\right]\\&\leq\sum_{k=1}^{q}\left(1-\bar{\pi}_k\right)C_\ell L_g\sqrt{\frac{\ln\left(1/\delta\right)}{2n_k^{\mathrm{U}}}}+\sum_{k=1}^{q}\left(\left(1-\pi_k-\bar{\pi}_k\right)L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln\left(1/\delta\right)}{2n_k^{\mathrm{N}}}}.\end{aligned}\tag{41}$$

Besides, we have

$$\begin{aligned}&\mathbb{E}\left[\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\right]\\&=\mathbb{E}_{\bar{\mathcal{D}}}\left[\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}_{\bar{\mathcal{D}}'}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\right]\\&\leq\mathbb{E}_{\bar{\mathcal{D}},\bar{\mathcal{D}}'}\left[\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\widetilde{R}(f_1,f_2,\ldots,f_q;\bar{\mathcal{D}})-\widetilde{R}(f_1,f_2,\ldots,f_q;\bar{\mathcal{D}}')\right|\right],\end{aligned}\tag{42}$$

where the last inequality is deduced by applying Jensen's inequality twice since the absolute value function and the supremum function are both convex. Here, $\widetilde{R}(f_1,f_2,\ldots,f_q;\bar{\mathcal{D}})$ denotes the value of $\widetilde{R}(f_1,f_2,\ldots,f_q)$ calculated on $\bar{\mathcal{D}}$. To ensure that the conditions in Lemma 20 hold, we introduce $\tilde{\ell}(z)=\ell(z)-\ell(0)$. It is obvious that $\tilde{\ell}(0)=0$ and $\tilde{\ell}(z)$ is also a Lipschitz continuous function with a Lipschitz constant L_ℓ. Then, we have

$$\begin{aligned}&\left|\widetilde{R}(f_1,f_2,\ldots,f_q;\bar{\mathcal{D}})-\widetilde{R}(f_1,f_2,\ldots,f_q;\bar{\mathcal{D}}')\right|\\&\leq\sum_{k=1}^{q}\left|g\left(\frac{\bar{\pi}_k+\pi_k-1}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)+\frac{1-\bar{\pi}_k}{n_k^{\mathrm{U}}}\sum_{i=1}^{n_k^{\mathrm{U}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}})\right)\right)\right.\\&\quad\left.-g\left(\frac{\bar{\pi}_k+\pi_k-1}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}'})\right)+\frac{1-\bar{\pi}_k}{n_k^{\mathrm{U}}}\sum_{i=1}^{n_k^{\mathrm{U}}}\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}'})\right)\right)\right|\\&\quad+\sum_{k=1}^{q}\left|\frac{1-\pi_k}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\ell\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)-\frac{1-\pi_k}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\ell\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}'})\right)\right|\\&\leq\sum_{k=1}^{q}L_g\left|\frac{\bar{\pi}_k+\pi_k-1}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\left(\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)-\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}'})\right)\right)+\frac{1-\bar{\pi}_k}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{U}}}\left(\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}})\right)-\ell\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}'})\right)\right)\right|\\&\quad+\sum_{k=1}^{q}\left|\frac{1-\pi_k}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\left(\ell\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)-\ell\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}'})\right)\right)\right|.\end{aligned}\tag{43}$$

Besides, we can observe $\ell(z_1)-\ell(z_2)=\tilde{\ell}(z_1)-\tilde{\ell}(z_2)$. Therefore, the RHS of Inequality (43) can be expressed as

$$\begin{aligned}&\sum_{k=1}^{q}L_g\left|\frac{\bar{\pi}_k+\pi_k-1}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\left(\tilde{\ell}\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)-\tilde{\ell}\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}'})\right)\right)+\frac{1-\bar{\pi}_k}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{U}}}\left(\tilde{\ell}\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}})\right)-\tilde{\ell}\left(f_k(\boldsymbol{x}_{k,i}^{\mathrm{U}'})\right)\right)\right|\\&+\sum_{k=1}^{q}\left|\frac{1-\pi_k}{n_k^{\mathrm{N}}}\sum_{i=1}^{n_k^{\mathrm{N}}}\left(\tilde{\ell}\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}})\right)-\tilde{\ell}\left(-f_k(\boldsymbol{x}_{k,i}^{\mathrm{N}'})\right)\right)\right|.\end{aligned}$$

Then, it is a routine work to show by symmetrization (Mohri et al., 2012) that

$$
\begin{aligned}
&\mathbb{E}_{\bar{\mathcal{D}},\bar{\mathcal{D}}'}\left[\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\widetilde{R}(f_1,f_2,\ldots,f_q;\bar{\mathcal{D}})-\widetilde{R}(f_1,f_2,\ldots,f_q;\bar{\mathcal{D}}')\right|\right]\\
\leq&\sum_{k=1}^{q}\left(\left(2-2\bar{\pi}_k\right)L_g\mathfrak{R}'_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\tilde{\ell}\circ\mathcal{F})+\left(\left(2-2\pi_k-2\bar{\pi}_k\right)L_g+2-2\pi_k\right)\mathfrak{R}'_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\tilde{\ell}\circ\mathcal{F})\right)\\
\leq&\sum_{k=1}^{q}\left(\left(4-4\bar{\pi}_k\right)L_gL_\ell\mathfrak{R}'_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+\left(\left(4-4\pi_k-4\bar{\pi}_k\right)L_g+4-4\pi_k\right)L_\ell\mathfrak{R}'_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})\right)\\
=&\sum_{k=1}^{q}\left(\left(4-4\bar{\pi}_k\right)L_gL_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+\left(\left(4-4\pi_k-4\bar{\pi}_k\right)L_g+4-4\pi_k\right)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})\right),
\end{aligned}
\tag{44}
$$

where the second inequality is deduced according to Lemma 20 and the last equality is based on the assumption that $\mathcal{F}$ is closed under negation. By combing Inequality (41), Inequality (42), and Inequality (44), we have the following inequality with probability at least $1-\delta$:

$$
\begin{aligned}
&\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\\
\leq&\sum_{k=1}^{q}\left(1-\bar{\pi}_k\right)C_\ell L_g\sqrt{\frac{\ln\left(1/\delta\right)}{2n_k^{\mathrm{U}}}}+\sum_{k=1}^{q}\left(\left(1-\pi_k-\bar{\pi}_k\right)L_g+1-\pi_k\right)C_\ell\sqrt{\frac{\ln\left(1/\delta\right)}{2n_k^{\mathrm{N}}}}\\
&+\sum_{k=1}^{q}\left(\left(4-4\bar{\pi}_k\right)L_gL_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+\left(\left(4-4\pi_k-4\bar{\pi}_k\right)L_g+4-4\pi_k\right)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})\right).
\end{aligned}
\tag{45}
$$

Then, we have

$$
\begin{aligned}
&\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\\
=&\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]\right.\\
&\left.+\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|\\
\leq&\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|R(f_1,f_2,\ldots,f_q)-\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]\right|\\
&+\sup_{f_1,f_2,\ldots,f_q\in\mathcal{F}}\left|\mathbb{E}\left[\widetilde{R}(f_1,f_2,\ldots,f_q)\right]-\widetilde{R}(f_1,f_2,\ldots,f_q)\right|.
\end{aligned}
\tag{46}
$$

Combining Inequality (46) with Inequality (45) and Inequality (18), the proof is completed. □

We present a more complete version of Theorem 10 and its proof.

Theorem 22. *Based on the above assumptions, for any $\delta>0$, the following inequality holds with probability at least $1-\delta$:*

$$
\begin{aligned}
&R(\widetilde{f}_1,\widetilde{f}_2,\ldots,\widetilde{f}_q)-R(f_1^*,f_2^*,\ldots,f_q^*)\leq\sum_{k=1}^{q}\left(4-4\bar{\pi}_k-2\pi_k\right)\left(L_g+1\right)C_\ell\Delta_k\\
&+\sum_{k=1}^{q}\left(\left(1-\bar{\pi}_k\right)C_\ell L_g\sqrt{\frac{2\ln\left(1/\delta\right)}{n_k^{\mathrm{U}}}}+\left(\left(1-\pi_k-\bar{\pi}_k\right)L_g+1-\pi_k\right)C_\ell\sqrt{\frac{2\ln\left(1/\delta\right)}{n_k^{\mathrm{N}}}}\right)\\
&+\sum_{k=1}^{q}\left(\left(8-8\bar{\pi}_k\right)L_gL_\ell\mathfrak{R}_{n_k^{\mathrm{U}},p_k^{\mathrm{U}}}(\mathcal{F})+\left(\left(8-8\pi_k-8\bar{\pi}_k\right)L_g+8-8\pi_k\right)L_\ell\mathfrak{R}_{n_k^{\mathrm{N}},p_k^{\mathrm{N}}}(\mathcal{F})\right).
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&R(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q) - R(f_1^*, f_2^*, \ldots, f_q^*) \\
=&R(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q) - \widetilde{R}(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q) + \widetilde{R}(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q) - \widetilde{R}(f_1^*, f_2^*, \ldots, f_q^*) \\
&+ \widetilde{R}(f_1^*, f_2^*, \ldots, f_q^*) - R(f_1^*, f_2^*, \ldots, f_q^*) \\
\leq&R(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q) - \widetilde{R}(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q) + \widetilde{R}(f_1^*, f_2^*, \ldots, f_q^*) - R(f_1^*, f_2^*, \ldots, f_q^*) \\
\leq&2 \sup_{f_1, f_2, \ldots, f_q \in \mathcal{F}} \left| R(f_1, f_2, \ldots, f_q) - \widetilde{R}(f_1, f_2, \ldots, f_q) \right| . \qquad (47)
\end{aligned}
$$

The first inequality is deduced because $(\widetilde{f}_1, \widetilde{f}_2, \ldots, \widetilde{f}_q)$ is the minimizer of $\widetilde{R}(f_1, f_2, \ldots, f_q)$. Combining Inequality (47) and Lemma 21, the proof is completed. □

J. Details of Experimental Setup

J.1. Details of Synthetic Benchmark Datasets

We considered the single complementary-label setting and similar results could be observed with multiple complementary labels.

For the "uniform" setting, a label other than the ground-truth label was sampled randomly following the uniform distribution to be the complementary label.

For the "biased-a" and "biased-b" settings, we adopted the following row-normalized transition matrices of $p(\bar{y}|y)$ to generate complementary labels:

$$
\text{biased-a:} \begin{bmatrix}
0 & 0.250 & 0.043 & 0.040 & 0.043 & 0.040 & 0.250 & 0.040 & 0.250 & 0.043 \\
0.043 & 0 & 0.250 & 0.043 & 0.040 & 0.043 & 0.040 & 0.250 & 0.040 & 0.250 \\
0.250 & 0.043 & 0 & 0.250 & 0.043 & 0.040 & 0.043 & 0.040 & 0.250 & 0.040 \\
0.040 & 0.250 & 0.043 & 0 & 0.250 & 0.043 & 0.040 & 0.043 & 0.040 & 0.250 \\
0.250 & 0.040 & 0.250 & 0.043 & 0 & 0.250 & 0.043 & 0.040 & 0.043 & 0.040 \\
0.040 & 0.250 & 0.040 & 0.250 & 0.043 & 0 & 0.250 & 0.043 & 0.040 & 0.043 \\
0.043 & 0.040 & 0.250 & 0.040 & 0.250 & 0.043 & 0 & 0.250 & 0.043 & 0.040 \\
0.040 & 0.043 & 0.040 & 0.250 & 0.040 & 0.250 & 0.043 & 0 & 0.250 & 0.043 \\
0.043 & 0.040 & 0.043 & 0.040 & 0.250 & 0.040 & 0.250 & 0.043 & 0 & 0.250 \\
0.250 & 0.043 & 0.040 & 0.043 & 0.040 & 0.250 & 0.040 & 0.250 & 0.043 & 0
\end{bmatrix},
$$

$$
\text{biased-b:} \begin{bmatrix}
0 & 0.220 & 0.080 & 0.033 & 0.080 & 0.033 & 0.220 & 0.033 & 0.220 & 0.080 \\
0.080 & 0 & 0.220 & 0.080 & 0.033 & 0.080 & 0.033 & 0.220 & 0.033 & 0.220 \\
0.220 & 0.080 & 0 & 0.220 & 0.080 & 0.033 & 0.080 & 0.033 & 0.220 & 0.033 \\
0.033 & 0.220 & 0.080 & 0 & 0.220 & 0.080 & 0.033 & 0.080 & 0.033 & 0.220 \\
0.220 & 0.033 & 0.220 & 0.080 & 0 & 0.220 & 0.080 & 0.033 & 0.080 & 0.033 \\
0.033 & 0.220 & 0.033 & 0.220 & 0.080 & 0 & 0.220 & 0.080 & 0.033 & 0.080 \\
0.080 & 0.033 & 0.220 & 0.033 & 0.220 & 0.080 & 0 & 0.220 & 0.080 & 0.033 \\
0.033 & 0.080 & 0.033 & 0.220 & 0.033 & 0.220 & 0.080 & 0 & 0.220 & 0.080 \\
0.080 & 0.033 & 0.080 & 0.033 & 0.220 & 0.033 & 0.220 & 0.080 & 0 & 0.220 \\
0.220 & 0.080 & 0.033 & 0.080 & 0.033 & 0.220 & 0.033 & 0.220 & 0.080 & 0
\end{bmatrix}.
$$

For each example, we sample a complementary label from a multinomial distribution parameterized by the row vector of the transition matrix indexed by the ground-truth label.

For the "SCAR-a" and "SCAR-b" settings, we followed the generation process in Section 3.1 with the following class priors of complementary labels:

$$
\begin{aligned}
&\text{SCAR-a: } [0.05, 0.05, 0.2, 0.2, 0.1, 0.1, 0.05, 0.05, 0.1, 0.1] , \\
&\text{SCAR-b: } [0.1, 0.1, 0.2, 0.05, 0.05, 0.1, 0.1, 0.2, 0.05, 0.05] .
\end{aligned}
$$

We repeated the sampling procedure to ensure that each example had a single complementary label.

J.2. Descriptions of Compared Approaches

The compared methods in the experiments on synthetic benchmark datasets:

- PC (Ishida et al., 2017): A risk-consistent complementary-label learning approach using the pairwise comparison loss.
- NN (Ishida et al., 2019): A risk-consistent complementary-label learning approach using the non-negative risk estimator.
- GA (Ishida et al., 2019): A variant of the non-negative risk estimator of complementary-label learning by using the gradient ascent technique.
- L-UW (Gao & Zhang, 2021): A discriminative approach by minimizing the outputs corresponding to complementary labels.
- L-W (Gao & Zhang, 2021): A weighted loss based on L-UW by considering the prediction uncertainty.
- OP (Liu et al., 2023): A classifier-consistent complementary-label learning approach by minimizing the outputs of complementary labels.

The compared methods in the experiments on real-world benchmark datasets:

- CC (Feng et al., 2020b): A classifier-consistent partial-label learning approach based on the assumption of uniform distribution of partial labels.
- PRODEN (Lv et al., 2020): A risk-consistent partial-label learning approach using the self-training strategy to identify the ground-truth labels.
- EXP (Feng et al., 2020a): A classifier-consistent multiple complementary-label learning approach by using the exponential loss function.
- MAE (Feng et al., 2020a): A classifier-consistent multiple complementary-label learning approach by using the Mean Absolute Error loss function.
- Phuber-CE (Feng et al., 2020a): A classifier-consistent multiple complementary-label learning approach by using the Partially Huberised Cross Entropy loss function.
- LWS (Wen et al., 2021): A partial-label learning approach by leveraging a weight to account for the tradeoff between losses on partial and non-partial labels.
- CAVL (Zhang et al., 2022): A partial-label learning approach by using the class activation value to identify the true labels.
- IDGP (Qiao et al., 2023): An instance-dependent partial-label learning approach by modeling the generation process of partial labels.
- POP (Xu et al., 2023): A partial-label learning approach by progressively purifying candidate label sets.

J.3. Details of Models

For CIFAR-10, we used 34-layer ResNet (He et al., 2016) and 22-layer DenseNet (Huang et al., 2017) as the model architectures. For the other three datasets, we used a multilayer perceptron (MLP) with a hidden layer of width 500 equipped with the ReLU (Nair & Hinton, 2010) activation function and 5-layer LeNet (LeCun et al., 1998) as the model architectures.

For CLCIFAR-10 and CLCIFAR-20, we adopted the same data augmentation techniques for all the methods, including random horizontal flipping and random cropping. We used 34-layer ResNet (He et al., 2016) and 22-layer DenseNet (Huang et al., 2017) as the model architectures.

K. More Experimental Results

K.1. Experimental Results on CIFAR-10

Table 7 shows the experimental results on CIFAR-10 with synthetic complementary labels.

Table 7. Classification accuracy (mean±std) of each method on CIFAR-10 with a single complementary label. The best performance is shown in bold (pairwise *t*-test at the 0.05 significance level).

Setting	Uniform		Biased-a		Biased-b		SCAR-a		SCAR-b	
Model	ResNet	DenseNet	ResNet	DenseNet	ResNet	DenseNet	ResNet	DenseNet	ResNet	DenseNet
PC	14.33 ±0.73	17.44 ±0.52	25.46 ±0.69	34.01 ±1.47	23.04 ±0.33	29.27 ±1.05	14.94 ±0.88	17.11 ±0.87	17.16 ±0.86	21.14 ±1.34
NN	19.90 ±0.73	30.55 ±1.01	24.88 ±1.01	24.48 ±1.50	26.59 ±1.33	24.51 ±1.24	21.11 ±0.94	29.48 ±1.05	23.56 ±1.25	30.67 ±0.73
GA	**37.59 ±1.76**	**46.86 ±0.84**	20.01 ±1.96	22.41 ±1.33	16.74 ±2.64	21.48 ±1.46	**24.17 ±1.32**	29.04 ±1.84	23.47 ±1.30	30.72 ±1.44
L-UW	19.58 ±1.77	17.25 ±3.03	24.83 ±2.67	29.46 ±1.03	20.73 ±2.41	25.41 ±2.61	14.56 ±2.71	10.69 ±0.94	10.39 ±0.50	10.04 ±0.09
L-W	18.05 ±3.02	13.97 ±2.55	22.65 ±2.70	27.64 ±0.80	22.70 ±2.33	24.86 ±1.34	13.72 ±2.60	10.00 ±0.00	10.25 ±0.49	10.00 ±0.00
OP	23.78 ±2.80	39.32 ±2.46	29.47 ±2.71	41.99 ±1.54	25.60 ±4.18	39.61 ±2.26	17.55 ±1.38	27.12 ±1.17	20.08 ±2.96	27.24 ±2.62
SCARCE	**35.63 ±3.23**	42.65 ±2.00	**39.70 ±3.79**	**51.42 ±1.81**	**37.82 ±2.72**	**50.52 ±2.18**	**29.04 ±3.70**	**36.38 ±2.56**	**35.71 ±1.16**	**38.43 ±0.85**