Consistent Complementary-Label Learning via Order-Preserving Losses

Shuqi Liu[1] **Yuzhou Cao**[2] **Qiaozhen Zhang**[1] **Lei Feng**[2] **Bo An**[2]
[1]School of Statistics and Data Science, Nankai University, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
(correspondence to zhangqz@nankai.edu.cn)

Abstract

In contrast to ordinary supervised classification tasks that require massive data with high-quality labels, complementary-label learning (CLL) deals with the weakly-supervised learning scenario where each instance is equipped with a complementary label, which specifies a class the instance does not belong to. However, most of the existing statistically consistent CLL methods suffer from overfitting intrinsically due to the negative empirical risk issue. In this paper, we aim to propose overfitting-resistant and theoretically grounded methods for CLL. Considering the unique property of the distribution of complementarily labeled samples, we provide a risk estimator via order-preserving losses, which is naturally non-negative and thus can avoid overfitting. Moreover, we provide *classifier-consistency* analysis and statistical guarantee for this estimator. Furthermore, we provide a weighted version of the proposed risk estimator to further enhance its generalization ability and prove its statistical consistency. Experiments on benchmark datasets demonstrate the effectiveness of our proposed methods.

1 Introduction

Ordinary supervised classification tasks require each instance to be equipped with a ground-truth label, while vast data with high-quality labels is costly to acquire or even inaccessible. Supervised classification tasks also ignore the data with inexact, incomplete, or inaccurate supervision, e.g., partially-labeled, unlabeled, and noisy-labeled data, which are ubiquitous in reality. In order to efficiently utilize various types of weak supervision, weakly supervised learning (WSL) (Zhou, 2018; Sugiyama, 2015) has been widely studied in recent years, such as noisy-label learning (Ghosh et al., 2017; Zhang and Sabuncu, 2018; Ma et al., 2018; Kim et al., 2019; Liu and Guo, 2020; Han et al., 2020), positive and unlabeled learning (Du Plessis et al., 2014; Kiryo et al., 2017; Sakai et al., 2018), partial-label learning (Cour et al., 2011; Feng and An, 2018; Lv et al., 2020), semi-supervised learning (Chapelle et al., 2009; Miyato et al., 2018; Niu et al., 2013), similar-unlabeled learning (Bao et al., 2018), and unlabeled and uareabeled learning (Lu et al., 2018; Golovnev et al., 2019). In this paper, we consider complementary-label learning (CLL) (Ishida et al., 2017, 2019; Yu et al., 2018; Feng et al., 2020; Cao et al., 2022), which is a weakly supervised learning problem where the classifier is trained only from examples equipped with labels that denote a class they do not belong to.

In Ishida et al. (2017), the risk rewriting technique was applied to construct the unbiased risk estimators (UREs) from only data with complementary labels, which enables consistent learning results via empirical risk minimization (ERM) in this task. However, unlike classification risks that are always non-negative, the obtained UREs of them contain some negative terms and thus may not be lower-bounded, which can lead to serious overfitting according to previous studies (Ishida et al., 2017, 2019; Chou et al., 2020). To mitigate this problem, various corrections (Ishida et al., 2019; Chou et al., 2020; Gao and Zhang, 2021) on the UREs are conducted to enforce their non-negativity.

Though non-negative correction methods have been widely applied in the field of WSL (Kiryo et al., 2017; Lu et al., 2020), it has been shown in recent works that the risk estimators of CLL can be further improved by utilizing the properties of complementary labels. Chou et al. (2020) provided a surrogate complementary loss framework and proposed several risk estimators based on this framework. Another non-negative risk estimator for CLL (Gao and Zhang, 2021) was proposed by modeling the posterior probability of complementary labels from the output of trained classifiers. The risk estimators above are non-negative, so they can naturally prevent overfitting caused by negative terms in the risk estimator. However, Chou et al. (2020) did not provide statistical consistency guarantees for the ERM with its

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

proposed estimators. Though Gao and Zhang (2021) gave theoretical analysis with a statistical consistency guarantee, strong restriction on loss function was required, i.e., only softmax cross-entropy loss is allowed. Meanwhile, the improved version of their method via weighting also lacked statistical consistency guarantees.

To tackle the problems above, we propose non-negative complementary learning via order-preservation losses. Considering the unique property of the distribution of complementarily labeled examples, we derive a risk estimator from order-preserving losses which is non-negative, thereby avoiding overfitting caused by negative terms in the risk estimator. Then we provide statistical guarantee for the estimator we proposed. Moreover, our method is compatible with arbitrary models and various losses as long as they are order-preserving. Furthermore, we introduce weighted loss based on the estimator we provided and investigate the statistical consistency of the risk correction method via weighted loss. The experiment results show that our method achieves the best performance among all the methods on various benchmark datasets.

The rest of this paper is organized as follows. Section 2 gives formal definitions and briefly reviews existing approaches to CLL. Section 3 presents the proposed non-negative complementary label learning via order-preserving. Section 4 shows the risk correction method via weighted loss. Section 5 states the theoretical analysis of our method. Section 6 reports the results of comparative experimental studies. Finally, Section 7 concludes this paper.

2 Preliminaries

In this section, we first introduce the problem formulations of ordinary multi-class classification and complementary-label learning, and then review the (improved) unbiased estimators of the classification risk.

2.1 Problem Formulations

Here, we introduce notations used in this paper and briefly review the existing methods for CLL.

Ordinary multi-class classification. In ordinary multi-class classification, suppose the feature space is $\mathcal{X} \in \mathbb{R}^d$ and let $\mathcal{Y} = \{1, 2, \ldots, K\}$ be the label set with K classes. The training examples at hand $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are drawn from an unknown distribution $p(\boldsymbol{x}, y)$, where each instance $\boldsymbol{x}_i$ is associated with the maximum possible label y_i. Ordinary multi-class classification aims at learning a scoring function $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^K$ that minimizes the classification risk:

$$R(\boldsymbol{g}; \ell_{01}) = \mathbb{E}_{p(\boldsymbol{x},y)} \left[\ell_{01} \left(\boldsymbol{g}(\boldsymbol{x}), y\right)\right], \tag{2.1}$$

where $\mathbb{E}$ and ℓ_{01} denote the expectation and the 0-1 loss: $\ell_{01}(\boldsymbol{g}(\boldsymbol{x}), y) = [\![y \neq \operatorname{argmax}_i \boldsymbol{g}_i(\boldsymbol{x})]\!]$, respectively. The prediction is usually generated by taking $\operatorname{argmax}_y \boldsymbol{g}_y$. Since 0-1 loss is not continuous and its optimization is NP-hard, ℓ_{01} is often substituted by some other surrogate loss functions $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$. The classifier f is implemented with a scoring function $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^K$ by taking the argmax function $f(\boldsymbol{x}) = \operatorname{argmax}_i \boldsymbol{g}_i(\boldsymbol{x})$, where $\boldsymbol{g}_i(\boldsymbol{x})$ denotes the i-th coordinate of $\boldsymbol{g}(\boldsymbol{x})$. The corresponding risk related to scorig function $\boldsymbol{g}$ and loss ℓ can be defined as:

$$R(\boldsymbol{g}; \ell) = \mathbb{E}_{p(\boldsymbol{x},y)} \left[\ell \left(\boldsymbol{g}(\boldsymbol{x}), y\right)\right]. \tag{2.2}$$

In practice, the classification risk can be approximated by the empirical risk, which requires to be minimized in the training stage:

$$\hat{R}(\boldsymbol{g}; \ell) = \frac{1}{n} \sum_{i=1}^{n} \ell \left(\boldsymbol{g}\left(\boldsymbol{x}_i\right), y_i\right). \tag{2.3}$$

Complementary-label learning. In this paper, we consider the single complementary-label learning task where each instance $\boldsymbol{x}_i$ has only one complementary label $\bar{y}_i$ which specifies one of the classes that the example does not belong to. Let $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$ denote the training examples that are drawn from an unknown distribution $\bar{p}(\boldsymbol{x}, \bar{y})$. Following the work of Ishida et al. (2017), the complementarily labeled examples are drawn form the following data distribution:

$$\bar{p}(\boldsymbol{x}, \bar{y}) = \frac{1}{K-1} \sum_{y \neq \bar{y}} p(\boldsymbol{x}, y). \tag{2.4}$$

The assumption (2.4) implies that except the correct label, all other labels are chosen with uniform probability. The first attempt towards learning from complementary labels was Ishida et al. (2017), which can recover the unbiased risk estimator when the used loss functions are limited to the one-versus-all (OvA) loss and the pairwise-comparison loss (Zhang, 2004).

To remove the restrictions on the available loss functions, Ishida et al. (2019) proposed a more general URE that is suitable for arbitrary losses and models:

$$\bar{R}(\boldsymbol{g}; \ell) = \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})} \left[\sum_{y=1}^{K} \ell(\boldsymbol{g}(\boldsymbol{x}), y) - (K-1)\ell(\boldsymbol{g}(\boldsymbol{x}), \bar{y})\right]. \tag{2.5}$$

Though this URE has no constraints on the loss and model, the empirical risk can go negative or even not lower-bounded when common losses are used (e.g., the cross-entropy loss) due to the negative terms in (2.5), which contradicts with the fact that its expectation is always non-negative. Many previous works (Ishida et al., 2019; Chou et al., 2020; Feng et al., 2020) have shown that, in practical implementation, the empirical risk would continue decreasing and go below zero, leading to the test accuracy dropping significantly. This indicates that UREs tend to have poor empirical performance and can be further improved.

Table 1: An overview of the properties of commonly used multi-class loss functions.

Loss	$\ell(\boldsymbol{u}, y)$	Order-Preserving	Convexity
Cross-Entropy	$-\log \boldsymbol{s}_y(\boldsymbol{u})$	✓	✓
Focal	$-(1-\boldsymbol{s}_y(\boldsymbol{u}))^\gamma \log \boldsymbol{s}_y(\boldsymbol{u})$	✓	✓
OvA Logistic	$\log(1+\exp(-\boldsymbol{u}_y)) + \sum_{y'\neq y} \log(1+\exp(\boldsymbol{u}_{y'}))$	✓	✓
Pairwise Logistic	$\sum_{y'\neq y} \log(1+\exp(\boldsymbol{u}_{y'}-\boldsymbol{u}_y))$	✓	✓
Pairwise Sigmoid	$\sum_{y'\neq y} \frac{1}{1+\exp(\boldsymbol{u}_y-\boldsymbol{u}_{y'})}$	×	×
MAE	$1-\boldsymbol{s}_y(\boldsymbol{u})$	×	×

2.2 Improved Risk Estimators

To alleviate the overfitting issue caused by the negative terms in the URE (2.5), Ishida et al. (2019) further proposed two non-negative corrections to the original URE (Ishida et al., 2019): the non-negative (NN) strategy and the gradient ascent (GA) strategy by employing the ReLU function and the absolute value function respectively on the risk estimator (2.5) to enforce the non-negativity to the risk estimator. The above correction methods are post-hoc corrections based on the URE (2.5)fromich can only mitigate overfitting caused by the negative risk issue instead of avoiding it. Later, Chou et al. (2020) proposed a naturally non-negative risk estimator by defining a complemfor entary 0-1 loss as

$$\bar{\ell}_{01}(\boldsymbol{g}(\boldsymbol{x}), \bar{y}) = [\![\bar{y} = \operatorname{argmax}_i \boldsymbol{g}_i(\boldsymbol{x})]\!]. \tag{2.6}$$

Based on this complementary 0-1 loss, a URE can be obtained from only complementary labeled data and this URE can be used in the validation process. They further selected convex surrogate losses to approximate the complementary 0-1 loss in empirical risk minimization, but their statistical consistency is not proved.

Gao and Zhang (2021) designed another complementary loss function that is also naturally non-negative using the predictive probability of the complementary label and proved its statistical consistency:

$$\bar{R}(\hat{\boldsymbol{p}}; \ell) = \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(1-\hat{\boldsymbol{p}}(y|\boldsymbol{x}), \bar{y})], \tag{2.7}$$

where $\hat{\boldsymbol{p}}(y|\boldsymbol{x})$ is the *softmax* output of the trained classifier $\boldsymbol{g}(\boldsymbol{x})$, i.e., $\hat{\boldsymbol{p}}(y=i|\boldsymbol{x}) = e^{\boldsymbol{g}_i(\boldsymbol{x})}/\sum_{j=1}^K e^{\boldsymbol{g}_j(\boldsymbol{x})}$, which can be interpreted as an approximation of the posterior probability $\boldsymbol{p}(y|\boldsymbol{x})$, and only the cross-entropy loss is allowed in Gao and Zhang (2021). Meanwhile, they introduced a weighted complementary loss by associating their proposed complementary loss function with a weight vector $\boldsymbol{\omega}$. Though the weighted loss strategy performed better than the unweighted version (2.7), it is worth noting that their weighting strategy does not have theoretical supports.

In this paper, considering the unique property of the distribution of complementarily labeled data, we propose a risk estimator via order-preserving losses. The proposed risk estimator is non-negative which naturally prevents overfitting caused by negative terms in the risk estimator, and can work with various losses as long as they satisfy the order-preserving property. Then, we provide statistical guarantee for the estimator we proposed. Furthermore, we improve the risk estimator by weighting, i.e., making more use of complementary labels when tackling a ranking problem via ERM, and we analyse the consistency of the weighted strategy for various losses.

3 Non-Negative Complementary Label Learning via Order-Preservation

In this section, we introduce the framework of non-negative complementary-label learning via order-preserving losses and provide statistical consistency guarantees for the estimator we proposed.

3.1 Risk Formulation and Consistency Analysis

Based on the relationship between the ordinary-label and complementary-label distribution (2.4), we can deduce that for any instance $\boldsymbol{x}$ and any label $i \in \{1, 2, \ldots, K\}$, $\bar{p}(\bar{y}=i|\boldsymbol{x})$ is inversely proportional to $p(y=i|\boldsymbol{x})$, which motivates us to propose the following risk estimator.

Definition 3.1. (Complementary Risk Estimator via Order-Preserving Classifier) With the scoring function $\boldsymbol{g}: \mathcal{X} \to \mathbb{R}^K$ and the order-preserving loss $\ell: \mathbb{R}^K \times \mathcal{Y} \to R_+$, we define a complementary risk estimator as

$$\bar{R}(\boldsymbol{g}; \ell) = \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y})\right], \tag{3.1}$$

where the order-preserving loss ℓ (Zhang, 2004) satisfies that for any $\boldsymbol{\eta} \in \Delta^K$:

$$\boldsymbol{\eta}_y > \boldsymbol{\eta}_{y'} \to \boldsymbol{u}^*_y > \boldsymbol{u}^*_{y'},$$

where $\boldsymbol{u}^* \in \mathbb{R}^K$ is the minimizer of the conditional risk $\boldsymbol{\eta}^T \boldsymbol{\ell}(\boldsymbol{u})$ and $\boldsymbol{\ell}(\boldsymbol{u}) = [\ell(\boldsymbol{u}, 1), \cdots, \ell(\boldsymbol{u}, K)]^T$.

We list the commonly used order-preserving and non-order-preserving losses and their properties in Table 1. Given

training examples $\{(\boldsymbol{x}_i, \bar{y})\}_{i=1}^n$, we can get the following unbiased empirical approximation of the risk estimator above.

$$\hat{\bar{R}}(\boldsymbol{g}, \ell) = \frac{1}{n}\sum_{i=1}^{n} \ell(-\boldsymbol{g}(\boldsymbol{x}_i), \bar{y}_i). \tag{3.2}$$

The derived risk estimator (3.1) includes an expectation over a non-negative loss $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_+$, so the risk and its empirical form (3.2) are both non-negative, which naturally avoid the overfitting issue caused by the negative terms in risk estimators. Meanwhile, our method can work with various losses as long as they satisfy the order-preserving property (e.g., cross-entropy, Focal, OvA, and pairwise logistic losses). Moreover, there are no implicit assumptions on the used classifiers, hence both linear and non-linear classifiers are allowed. Based on Definition 3.1, we give the statistical consistency analysis of the risk estimator we proposed.

Theorem 3.1. The risk formulation (3.1) is *classifier-consistent*, i.e., for any $\boldsymbol{x} \in \mathcal{X}$:

$$\operatorname{argmax}_y \bar{\boldsymbol{g}}_y^*(\boldsymbol{x}) = \operatorname{argmax}_y \boldsymbol{g}_y^*(\boldsymbol{x}),$$

where $\bar{\boldsymbol{g}}^*(\boldsymbol{x})$ and $\boldsymbol{g}^*(\boldsymbol{x})$ are the minimizer of (3.1) and (2.1), respectively.

Proof. First, we denote the minimizer of $\bar{R}(\boldsymbol{g}; \ell)$ (3.1) and $R(\boldsymbol{g}; \ell_{01})$ (2.1) as $\bar{\boldsymbol{g}}^*$ and $\boldsymbol{g}^*$ respectively.

For any $\boldsymbol{x}$, we can learn that:

$$\operatorname{argmax}_y \boldsymbol{g}_y^*(\boldsymbol{x}) = \operatorname{argmax}_y p(y|\boldsymbol{x}).$$

Based on the definition of order-preserving losses, we can learn that:

$$\operatorname{argmin}_y -\bar{\boldsymbol{g}}^*(\boldsymbol{x}) = \operatorname{argmin}_y \bar{p}(y|\boldsymbol{x}),$$

which indicates $\operatorname{argmax}_y \bar{\boldsymbol{g}}^*(\boldsymbol{x}) = \operatorname{argmin}_y \bar{p}(y|\boldsymbol{x})$,

From the definition of the generation process of complementary labels, we can learn that:

$$\bar{p}(y|\boldsymbol{x}) = \frac{\sum_{y' \neq y} p(y'|\boldsymbol{x})}{K-1} = \frac{1 - p(y|\boldsymbol{x})}{K-1}.$$

According to the equation above, we can immediately learn that:

$$\operatorname{argmin}_y \bar{p}(y|\boldsymbol{x}) = \operatorname{argmax}_y p(y|\boldsymbol{x}),$$

which indicates that $\operatorname{argmax}_y \bar{\boldsymbol{g}}^*(\boldsymbol{x}) = \operatorname{argmax}_y \boldsymbol{g}^*(\boldsymbol{x})$. Combining the conclusions above and we can conclude the proof. □

With this theorem, we can learn that our method is statistically consistent given infinite *i.i.d.* examples. In the next section and Section 5, we will further provide conclusions about the non-asymptotic properties of the proposed method.

3.2 Regret Transfer Bound

Based on Theorem 3.1, we can learn that the minimization of our risk formulation (3.1) yields a consistent classifier that can minimize the classification risk (2.1). However, this guarantee is still limited in two perspectives.

Firstly, since we only have access to the *i.i.d.* complementarily labeled examples drawn from the probability density $\bar{p}(\boldsymbol{x}, \bar{y})$, we can only approximate the risk formulation (3.1) using its unbiased estimation (3.2) and conduct ERM to obtain an empirically optimal classifier. Generally speaking, there usually exists a gap between the performance of the empirically optimal and globally optimal classifiers, and Theorem 3.1 fails to give a guarantee for such an empirically optimal classifier. Secondly, even if a large number of *i.i.d.* examples are provided to make the generated classifier's risk (3.1) close enough to its infimum, its classification accuracy still remains unclear since the decrease of (3.1) does not immediately yield that of (2.1).

An issue naturally arises from these gaps: to what degree the decrease of (3.1) leads to that of (2.1)? To answer this question, we provide the following regret transfer bound that quantifies the relation between $\bar{R}(\boldsymbol{g}, \ell)$ and $R(\boldsymbol{g}, \ell_{01})$ when using cross-entropy loss as the surrogate loss ℓ in (3.1):

Theorem 3.2. (Regret transfer bound of the proposed risk)

$$\Delta_{01}(\boldsymbol{g}) \leq \sqrt{2}(K-1)\sqrt{\bar{\Delta}_{\mathrm{CE}}(\boldsymbol{g})}, \tag{3.3}$$

where $\bar{\Delta}_{\mathrm{CE}}(\boldsymbol{g}) = \bar{R}(\boldsymbol{g}, \ell_{\mathrm{CE}}) - \min_{\boldsymbol{g}'} \bar{R}(\boldsymbol{g}', \ell_{\mathrm{CE}})$ and $\Delta_{01}(\boldsymbol{g}) = R(\boldsymbol{g}, \ell_{01}) - \min_{\boldsymbol{g}'} R(\boldsymbol{g}', \ell_{01})$.

Proof. We begin with the point-wise regret bound and then generalize it to the conclusion via Jensen's inequality.

Suppose $\boldsymbol{\eta}(\boldsymbol{x})$ is the class-posterior probability of the ordinary labels and $\bar{\boldsymbol{\eta}}(\boldsymbol{x})$ is the class-posterior probability of the complementary labels. According to Pinsker's inequality:

$$\bar{\Delta}_{\mathrm{CE}}(\boldsymbol{g}, \boldsymbol{x}) \geq \frac{1}{2}\|\bar{\boldsymbol{\eta}}(\boldsymbol{x}) - \boldsymbol{s}(-\boldsymbol{g}(\boldsymbol{x}))\|_1^2,$$

where $\boldsymbol{s}(-\boldsymbol{g}(\boldsymbol{x}))$ is the softmax output. When $\operatorname{argmax}_y g_y(\boldsymbol{x}) = \operatorname{argmax}_y p(y|\boldsymbol{x})$, we can learn that $\bar{\Delta}_{CE}(\boldsymbol{g}, \boldsymbol{x}) \geq \Delta_{01}(\boldsymbol{g}, \boldsymbol{x}) = 0$. When the equality does not hold, suppose $\operatorname{argmax}_y g_y(\boldsymbol{x}) = y_1$ and $\operatorname{argmax}_y p(y|\boldsymbol{x}) = \operatorname{argmin}_{\bar{y}} \bar{p}(\bar{y}|\boldsymbol{x}) = y_2$, we can further learn that:

$$\begin{aligned}
&\frac{1}{2}\|\bar{\boldsymbol{\eta}}(\boldsymbol{x}) - \boldsymbol{s}(-\boldsymbol{g}(\boldsymbol{x}))\|_1^2 \\
&\geq \frac{1}{2}\left(|\bar{p}(y_1|\boldsymbol{x}) - \boldsymbol{s}_{y_1}(-\boldsymbol{g}(\boldsymbol{x}))| + |\bar{p}(y_2|\boldsymbol{x}) - \boldsymbol{s}_{y_2}(-\boldsymbol{g}(\boldsymbol{x}))|\right)^2 \\
&\geq \frac{1}{2}\left(|\bar{p}(y_1|\boldsymbol{x}) - \boldsymbol{s}_{y_1}(-\boldsymbol{g}(\boldsymbol{x})) - \bar{p}(y_2|\boldsymbol{x}) + \boldsymbol{s}_{y_2}(-\boldsymbol{g}(\boldsymbol{x}))|\right)^2
\end{aligned}$$

Since $\boldsymbol{g}_{y_2}(\boldsymbol{x}) < \boldsymbol{g}_{y_1}(\boldsymbol{x})$, we can learn that $\boldsymbol{s}_{y_1}(-\boldsymbol{g}(\boldsymbol{x})) < \boldsymbol{s}_{y_2}(-\boldsymbol{g}(\boldsymbol{x}))$, then we can continue the proof:

$$
\begin{aligned}
&\frac{1}{2}\left(|\bar{p}(y_1|\boldsymbol{x}) - \boldsymbol{s}_{y_1}(-\boldsymbol{g}(\boldsymbol{x})) - \bar{p}(y_2|\boldsymbol{x}) + \boldsymbol{s}_{y_2}(-\boldsymbol{g}(\boldsymbol{x}))|\right)^2 \\
&= \frac{1}{2}\left(\bar{p}(y_1|\boldsymbol{x}) - \bar{p}(y_2|\boldsymbol{x}) + \boldsymbol{s}_{y_2}(-\boldsymbol{g}(\boldsymbol{x})) - \boldsymbol{s}_{y_1}(-\boldsymbol{g}(\boldsymbol{x}))\right)^2 \\
&\geq \frac{1}{2}\left(\bar{p}(y_1|\boldsymbol{x}) - \bar{p}(y_2|\boldsymbol{x})\right)^2 \\
&= \frac{1}{2}\left(\frac{1-p(y_1|\boldsymbol{x})}{K-1} - \frac{1-p(y_2|\boldsymbol{x})}{K-1}\right)^2 \\
&= \frac{(p(y_1|\boldsymbol{x}) - p(y_2|\boldsymbol{x}))^2}{2(K-1)^2} \\
&= \frac{\Delta_{01}(\boldsymbol{g}, \boldsymbol{x})^2}{2(K-1)^2}
\end{aligned}
$$

Then we can learn that $\bar{\Delta}_{\mathrm{CE}}(\boldsymbol{g}, \boldsymbol{x}) \geq \frac{\Delta_{01}(\boldsymbol{g}, \boldsymbol{x})^2}{2(K-1)^2}$, which concludes the proof. □

The proof of this theorem is conducted using the Pinsker's inequality and the definition of $\bar{p}(\boldsymbol{x}, \bar{y})$. From this theorem, we can learn that the upper bound of the gap between the misclassification rate of $\boldsymbol{g}$ and the Bayes optimal classifier $\boldsymbol{g}^*$ is guaranteed to be smaller and vanishes as $\bar{R}(\boldsymbol{g}, \ell_{\mathrm{CE}})$ moves closer to its minimum. Compared with the bound in the scenario of ordinary classification (Ni et al., 2019)[1], the regret transfer bound above is $K-1$ times larger, which implicates the inherent difficulty of complementary-label learning.

4 Risk Correction via Weighting

In Section 3, we show that based on the unique property of the distribution of complementary labels, a naturally non-negative risk estimator can be constructed by considering a label ranking problem (Fotakis et al., 2022; Vogel and Clémençon, 2020; Brinker and Hüllermeier, 2019). In this section, we first discuss some potential shortcomings of the classifier-consistent risk estimator proposed in the previous section, and then give a novel weighting strategy to further improve its performance. The statistical consistency of the improved risk estimator is also provided for various loss functions.

4.1 Adaptive Risk Weighting Strategy

It is worth noting that the performance of our obtained classifier $\boldsymbol{g}$ is determined by whether $\bar{\boldsymbol{g}} = -\boldsymbol{g}$ can give a correct prediction on the complementary label with minimal likelihood for each instance $\boldsymbol{x}$, i.e., $\min_{\bar{y}} \bar{p}(\bar{y}|\boldsymbol{x})$. However, such dependence can be unreliable due to the intrinsic difficulties of the ranking problem.

[1]In Ni et al. (2019), the bound for ordinary classification can be deduced by setting $c = 1$ in Appendix A.4.

To be detailed, when tackling a ranking problem via ERM, the algorithm can often achieve better results on predicting the top items compared with the bottom ones since the top items are given larger weights (posterior probabilities in the classification scenario) and incur more losses when sorted incorrectly. When solving a standard classification problem via such a ranking approach, the property mentioned above is not disadvantageous because people are more concerned about the most likely class label (ranked on the top) and other labels are of less importance or even dropped simply. However, the influence of such a property must not be overlooked in our risk minimization setting (3.1) since we aim at predicting the **least likely** complementary label, which is at the bottom of the ranking list.

To mitigate the dichotomy mentioned above, we improve the proposed risk estimator (3.1) through a weighting strategy, which can be combined with any order-preserving loss functions mentioned in Table 1. The improved risk estimator is defined as follows:

Definition 4.1. (Weighted Order-Preserving Complementary Risk Formulation) With the classifier $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^K$ and the order-preserving loss $\ell : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}^+$, the weighted order-preserving complementary risk is defined as

$$
\bar{R}^w(\boldsymbol{g}; \ell) = \mathbb{E}_{\bar{p}(\boldsymbol{x}, \bar{y})}\left[w(\boldsymbol{g}(\boldsymbol{x}), \bar{y}) \cdot \ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y})\right], \quad (4.1)
$$

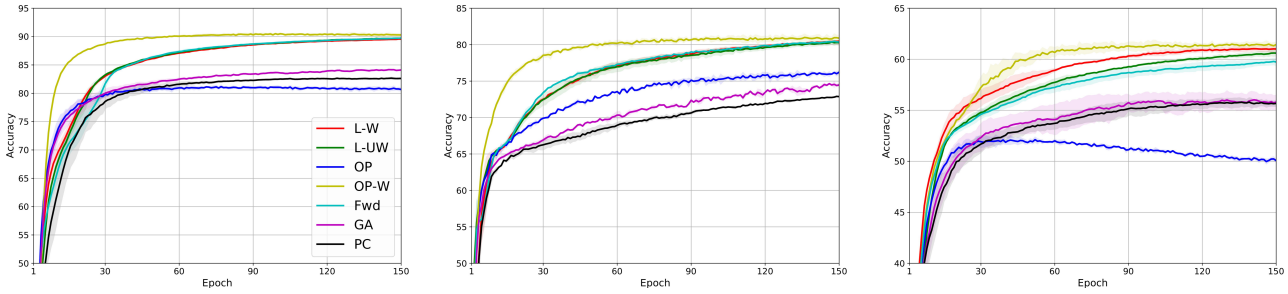
where $w(\cdot, \cdot)$ is the weight function that for any $y \in \mathcal{Y}$, it is strictly decreasing $w.r.t.$ u_y and strictly increasing $w.r.t.$ $u_{y'}$ where $y' \neq y$, i.e., for any $\delta > 0$:

$$
w(\boldsymbol{u} + \delta \boldsymbol{e}^y, y) > w(\boldsymbol{u}, y),
$$

where $\boldsymbol{e}^y$ is the canonical basis for $\mathbb{R}^K$ with a unique non-zero entry $e^y_y = 1$.

This correction can be simply implemented by optimizing the product of the loss function and the weight function instead. Let us illustrate how this weighting strategy helps mitigate the problem mentioned in this section. Intuitively, this weighting strategy can enhance the importance of examples that receive too much attention to ensure that the most likely complementary label $\bar{y}^*$ is ranked at the top. For any $\boldsymbol{x} \in \mathcal{X}$, when the model increases the score of its most likely complementary label $\bar{y}^*$, $-\boldsymbol{g}(\boldsymbol{x})_{\bar{y}^*}$, to decrease its corresponding loss $\ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y}^*)$, the weight function $w(\boldsymbol{g}(\boldsymbol{x}), \bar{y}^*)$ will increase to stop the weighted loss $w(\boldsymbol{g}(\boldsymbol{x}), \bar{y}^*)\ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y}^*)$ from further decreasing, which can prevent the algorithm from paying too much attention to make $-\boldsymbol{g}(\boldsymbol{x})_{\bar{y}^*}$ larger.

Though intuitively plausible, the statistical consistency of the proposed weighting strategy is not straightforward. For example, if we set $w(\boldsymbol{u}, y) = \exp(u_y)$, we can simply let $u_y \to -\infty$ for any $y \in \mathcal{Y}$ to make the weighted risk converge to be 0, which can lead to a degenerated solution. In the following section, we will show the condition under which the proposed $\bar{R}^w(\boldsymbol{g}, \ell)$ can yield consistent results.



(a) MNIST, linear (b) Fashion, linear (c) Kuzushiji, linear

(d) MNIST, MLP (e) Fashion, MLP (f) Kuzushiji, MLP

Figure 1: The experimental results on various test datasets with different methods and models for 150 epochs for 5 trails. The dark color is the mean accuracy and the light corresponds to the std.

4.2 Consistency of the Weighting Strategy

Though the intuitive motivation of our proposed weighting strategy is given in the previous section, its classifier-consistency (stated in Theorem 3.1) is still unclear, i.e., if we can obtain a classifier which can generate the most likely ordinary label for each instance $\boldsymbol{x}$ by minimizing (4.1)? In the following theorem, we show that with mild conditions, we can safely claim the consistency of our weighting strategy:

Theorem 4.1. The weighted order-preserving risk (4.1) is classifier-consistent when combined with the order-preserving losses listed in Table 1 if the weight function $w(\cdot,\cdot) \in \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}^+$ and $\boldsymbol{w}(\boldsymbol{u}) = [w(\boldsymbol{u},1),\cdots,w(\boldsymbol{u},K)]^T$ meet the following three conditions:

- w is differentiable and lower bounded by $\epsilon > 0$,
- w is symmetric, i.e., $P\boldsymbol{w}(\boldsymbol{u}) = \boldsymbol{w}(P\boldsymbol{u})$ for all the permutation matrix P.
- Function $w(\boldsymbol{u},y)$ is strictly decreasing $w.r.t.$ u_y and strictly increasing $w.r.t.u_{y'}$, $y' \neq y$, i.e., for $\delta > 0$:

$$w(\boldsymbol{u}+\delta\boldsymbol{e}^y,y) > w(\boldsymbol{u},y),$$
$$w(\boldsymbol{u}+\delta\boldsymbol{e}^{y'},y) < w(\boldsymbol{u},y),$$

  where $\boldsymbol{e}^y$ is the canonical basis for $\mathbb{R}^K$ with a unique non-zero entry $e^y_y = 1$.

The proof is shown in the appendix. After showing that our weighting strategy is statistically valid, we will further demonstrate its performance in Section 6.

5 Estimation Error Analysis

In this section, we establish the estimation error bounds of our proposed methods. The proof of the conclusions in this section are all provided in the appendix. Let $\mathcal{G} \subset \mathcal{X} \to \mathbb{R}^K$ be the model class and each of its dimension is constructed by $\mathcal{G}_y \subset \mathcal{X} \to \mathbb{R}$. Assume there exists $C_{\boldsymbol{g}} > 0$ that $\sup_{\boldsymbol{g}\in\mathcal{G}} \|\boldsymbol{g}\|_\infty \leq C_{\boldsymbol{g}}$ and $C_\ell > 0$ such that $\sup_{\|\boldsymbol{z}\|_\infty \leq C_{\boldsymbol{g}}} \ell(\boldsymbol{z},y) \leq C_\ell$ for any $y \in \mathcal{Y}$. We also assume that $\ell(\boldsymbol{z},y)$ is L_ℓ-Lipschitz continuous $w.r.t.$ $\boldsymbol{z}$ following the common practice (Mohri et al., 2018). Suppose $\mathfrak{R}_n(\mathcal{G}_y)$ is the Rademacher complexity (Bartlett and Mendelson, 2002; Bartlett et al., 2002) of $\mathcal{G}_y$ given n $i.i.d.$ samples drawn from distribution with density $\bar{p}(\boldsymbol{x},\bar{y})$. We show the definition of Rademacher complexity in the appendix.

Suppose $\bar{\boldsymbol{g}}^*$ and $\hat{\bar{\boldsymbol{g}}}$ are the minimizers of (3.1) and its empirical version, respectively. Then we can get the following estimation error bound.

Theorem 5.1. For any $\delta > 0$, the following inequality holds with probability at least $1-\delta$:

$$\bar{R}(\hat{\bar{\boldsymbol{g}}},\ell) - \bar{R}(\bar{\boldsymbol{g}}^*,\ell) \leq 4\sqrt{2}L_\ell \sum_{y=1}^{K} \mathfrak{R}_n(\mathcal{G}_y) + 4C_\ell\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

Here we also provide the estimation error bound for our weighted risk formulation. Suppose that the weight function is upper-bounded by C_w and is L_w-Lipschitz continuous for any $y \in \mathcal{Y}$. Suppose $\bar{\boldsymbol{g}}^{w*}$ and $\hat{\bar{\boldsymbol{g}}}^w$ are the minimizers of (4.1) and its empirical version, respectively. Then we can get the following conclusion.

Table 2: Specification of benchmark datasets and models.

Statistics	MNIST	Fashion-MNIST	Kuzushiji-MNIST	SVHN
#Train	60,000	60,000	60,000	73257
#Test	10,000	10,000	10,000	26032
#Feature	784	784	784	3072
Simple Model	Linear Model	Linear Model	Linear Model	—-
Deep Model	784-500-10 MLP	784-500-10 MLP	784-500-10 MLP	ResNet-18

Table 3: Classification accuracy (mean±std) of each algorithm on three datasets using linear models for 5 trails. The best performance among all the approaches is highlighted in boldface.

Approach	MNIST	Fashion-MNIST	Kuzushiji-MNIST
PC	82.66±0.81	72.91±0.05	54.96±1.28
Forward	89.73±0.34	80.47±0.10	59.89±0.31
GA	84.18±0.35	74.80±0.17	56.25±0.70
L-UW	89.72±0.24	80.35±0.33	60.65±0.26
L-W	89.56±0.25	80.49±0.03	60.88±0.32
OP (Naive)	81.16±0.46	76.25±0.23	51.96±0.50
OP-W (Ours)	**90.51±0.23**	**81.13±0.38**	**61.67±0.26**

Theorem 5.2. For any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$\bar{R}^w(\hat{\bar{\boldsymbol{g}}}, \ell) - \bar{R}^w(\bar{\boldsymbol{g}}^*, \ell) \leq 4\sqrt{2}L' \sum_{y=1}^{K} \mathfrak{R}_n(\mathcal{G}_y) + 4C' \sqrt{\frac{\ln \frac{2}{\delta}}{2n}},$$

where $L' = L_\ell C_w + L_w C_\ell$ and $C' = C_w C_\ell$.

Notice that the Rademacher complexity $\mathfrak{R}_n(\mathcal{G})$ is usually assumed to be smaller than $\mathcal{R}/\sqrt{n}$, which holds for various models like linear-in-input model and fully-connected neural network . We make this assumption in the rest analysis of this paper. Then we can learn that risk minimization with the empirical version of proposed risks (3.1) and (4.1) converges in the rate of $\mathcal{O}_p(1/\sqrt{n})$, which is the optimal parametric convergence rate without additional assumptions (Mendelson, 2008).

Furthermore, with cross-entropy loss and the identifiable condition, which is a common assumption when the model class is complex (Bao et al., 2020; Lu et al., 2021; Hsu et al., 2019), we can combine the regret transfer bound in Theorem 3.2 and the estimation error bound in Theorem 5.1 to further get a high-probability bound for the expected 0-1 risk, i.e., misclassification rate, for learning with our method (3.1).

Corollary 5.1. Assume that the identifiable condition holds, i.e., $\min_{\boldsymbol{g} \in \mathcal{X} \to \mathbb{R}^K} \bar{R}(\boldsymbol{g}, \ell_{\mathrm{CE}}) = \min_{\boldsymbol{g} \in \mathcal{G}} \bar{R}(\boldsymbol{g}, \ell_{\mathrm{CE}})$. For any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$:

$$R(\hat{\bar{\boldsymbol{g}}}, \ell_{01}) - \min_{\boldsymbol{g}'} R(\boldsymbol{g}', \ell_{01}) \leq \sqrt{2}(K-1) \sqrt{\sqrt{2}L_\ell \sum_{y=1}^{K} \mathfrak{R}_n(\mathcal{G}_y) + \frac{C_\ell}{2} \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}}.$$

From the corollary above, we can learn that with identifiable condition and cross-entropy loss, the classification accuracy of the empirical risk minimizer of (3.1) converges in the rate of $\mathcal{O}_p(1/n^{1/4})$. However, the term $K - 1$ still indicates that the naive order-preserving method (3.1) can be quite inefficient under the multi-class classification scenario. In the following experiment, we show that this shortcoming can be effectively mitigated by our consistent weighting strategy proposed in Section 4.

6 Experiments

In this section, we provide the experimental results, i.e., the classification accuracy of our proposed methods and baseline methods with different models on several commonly used benchmark datasets based on the original data generation process defined in Ishida et al. (2017, 2019) that is mentioned in Section 2.1. The specific parameter setting and description of different datasets and models are provided in the appendix. We provide an overview of our used models and datasets in Table 2.

Table 4: Classification accuracy (mean±std) of each algorithm on three datasets using complex models for 3 trails. The best performance among all the approaches is highlighted in boldface.

Approach	MNIST	Fashion-MNIST	Kuzushiji-MNIST	SVHN
PC	85.03±0.21	78.22±0.44	58.42±0.83	59.33±2.74
Forward	92.66±0.10	83.87±0.19	65.13±0.58	81.54±0.53
GA	88.62±0.26	80.16± 0.17	64.91±0.57	77.12±0.09
L-UW	92.12± 0.08	83.22±0.2	64.62±0.58	73.13±0.49
L-W	91.82±0.39	83.39± 0.15	65.91±2.25	79.54±0.21
OP (Naive)	91.26±0.13	81.90± 0.09	68.47±0.46	63.75±0.94
OP-W (Ours)	**94.13±0.29**	**84.58±0.39**	**73.13±0.49**	**83.34±0.63**

6.1 Experimental Setup

Baselines. We compare our methods with the state-of-the-art methods in learning with single complementary labels, including pairwise-comparison (PC) & gradient ascent (GA) (Ishida et al., 2019), forward correction (Fwd) (Patrini et al., 2017), unweighted loss (L-UW) & weighted loss (L-W) (Gao and Zhang, 2021). We denote our classifier-consistent order-preserving method (3.1) and weighted order-preserving method (4.1) with OR and OR-W, respectively. Cross-entropy loss is used in all the methods. We implemented all the methods by Pytorch (Paszke et al., 2019) and conducted all the experiments on NVIDIA GeForce 3090 GPUs.

Datasets and models. We conduct experiments with both a linear-in-parameter model and an MLP with one hidden layer on three commonly used benchmark datasets: MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), and Kuzushiji-MNIST (Clanuwat et al., 2018). To further validate the performance of our proposed methods with complex models, we add an extra experiment on SVHN (Netzer et al., 2011) with ResNet-18 (He et al., 2016). We use Adam (Kingma and Ba, 2015) as the optimizer for all the experiments.

6.2 Experimental Results

We show the experimental results of linear models in Table 3 and those of MLP and the result of ResNet-18 on SVHN in Table 5. The epoch-wise testing accuracy of each method on MNIST, Fashion-MNIST, and Kuzushiji-MNIST are provided in Figure 1.

Specifically, we can observe that:

- From Table 3 and 5, it can be seen that our proposed weighted order-preserving method (4.1), i.e., OP-W, outperforms other methods on all the datasets with different models.
- It can also be learned from Figure 1 that OP-W not only outperforms other methods but also converges faster, which means that it needs less training epochs to achieve a satisfying performance than other methods.
- Comparing the experimental results of OP and OP-W, we can conclude that the weighting strategy boosts the performance of our naive OP method, which often suffers from suboptimal performance.
- Though the performance of L-W is close to that of OP-W when linear model is used, it is far inferior to OP-W when more complex models, i.e., MLP and ResNet-18 are used, which can be observed in the experimental results of Kuzushiji-MNIST and SVHN in Table 5.

In conclusion, the experimental results show that our weighting strategy can effectively enhance the performance of the naive OP method and t weighted order-preserving method (OP-W) (4.1) can benefit from both outstanding performance and fast convergence on both simple linear model and complex deep models.

6.3 Additional Experiments

We conduct experiments based on the setup in the previous sections with more weighting strategies and loss functions. Due to the page limitation, we focus on SVHN and CIFAR-10. We use ResNet-34 on CIFAR-10 for all the methods and other settings are the same as those in the previous sections. Aside from CE loss, we also use Focal and Generalized-CE (GCE) losses in the experiments, where $\ell_{foc}(\boldsymbol{u}, y) = -(1 - s_y(\boldsymbol{u}))^\gamma \log s_y(\boldsymbol{u})$, $\ell_{GCE}(\boldsymbol{u}, y) = (1 - s_y(\boldsymbol{u})^\alpha)/\alpha$, and $s(\cdot)$ is the softmax function. We set $\gamma = 2$ as the default value and $\alpha = 0.3$. We also conduct experiments with a new simple family of weighting strategies: $\tilde{w}_y(\boldsymbol{u}) = s_y(u)^q$, where $q \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. We set proposed methods that perform comparable to or better than compared methods in gray and show the best methods in boldface.

We first analyze the result of CIFAR-10. We can find that all three losses outperform the compared methods when combined with the new weighting strategy $\tilde{w}$ under all the

Table 5: Classification accuracy (mean±std) offewerpared and proposed methods on CIFAR-10 and SVHN.

CIFAR-10-Compared					
PC	Forward	GA	L-UW	L-W	
31.47±0.76	41.78±0.14	37.44± 0.71	38.18±0.81	38.65±0.09	

CIFAR-10-Proposed								
Weight	OP	OP-W	$q = 0.5$	$q = 0.6$	$q = 0.7$	$q = 0.8$	$q = 0.9$	$q = 1.0$
CE	26.57±1.05	41.75± 0.94	42.96±0.26	43.12±0.17	44.31± 0.47	42.75± 0.27	43.92± 0.23	42.32± 0.36
Focal	27.06±0.66	28.96±0.36	44.50± 0.31	43.31±0.31	44.30±0.13	44.11±0.21	44.37±0.16	44.08±0.40
GCE	23.76±0.34	28.88± 0.29	44.02±0.35	44.35±0.35	**45.52±0.27**	43.37±0.48	44.56±0.33	44.75±0.39

SVHN-Proposed								
Weight	OP	OP-W	$q = 0.5$	$q = 0.6$	$q = 0.7$	$q = 0.8$	$q = 0.9$	$q = 1.0$
CE	63.71±0.94	**83.34±0.63**	81.20±0.26	81.38±0.31	81.94±0.48	82.30±0.34	82.96±0.17	82.96±0.17
Focal	65.67±0.47	81.52±0.29	81.49± 0.46	81.78±0.66	82.47±0.19	**83.21±0.23**	**83.50±0.33**	**83.04±0.53**
GCE	66.70±0.38	69.72± 0.13	79.83±0.56	80.16±0.21	80.29±0.38	82.37±0.25	82.42±0.41	82.50±0.16

selections of parameter q. Though Forward has the best performance among all the compared methods, it is still not comparable to the proposed methods with weighting strategy $\tilde{w}$. Regarding SVHN, we can find that the new weighting strategy is still comparable to the compared methods and outperforms those when Focal loss is used.

According to the experimental results in this section and Section 6.2, it can also be observed that our proposed methods outperform compared methods on all the datasets. Furthermore, on the complex dataset CIFAR-10, Focal loss and GCE loss with weighting strategy $\tilde{w}$ have better performance than the CE loss, which indicates that the access to various kinds of order-preserving losses and weighting strategies is a reason for the superiority of our method. With complex models, the improvement of our methods is 8% on Kuzushiji-MNIST and $1-4\%$ on other datasets. Though not as remarkable as that on Kuzushiji-MNIST, the $1-4\%$ improvement over compared methods is also common and significant in CL as shown in the previous studies Ishida et al. (2019); Feng et al. (2020); Gao and Zhang (2021).

7 Conclusion

In this paper, we focus on the problem of learning from single complementary labels. We first propose a naturally non-negative classifier-consistent risk formulation based on the connection between complementary labels and ordinary labels with order-preserving loss functions. To enhance the efficiency of the proposed naive order-preserving (OR) method, we further propose a weighting strategy (OR-W) for it and prove its statistical consistency. We also provide estimation error analyses for both methods we proposed and discuss the regret transfer bound for the naive OR method when using cross-entropy loss as the loss function. Experimental results show that our proposed weighting strategy can boost the performance of the naive OR method and outperform the state-of-the-art methods on different benchmark datasets with both linear and deep models.

Acknowledgement

This work was supported by the State Key Laboratory of CEMEE (CEMEE2020K0301A). Yuzhou Cao and Bo An are supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Lei Feng was supported by the National Natural Science Foundation of China (Grant No. 62106028), Chongqing Overseas Chinese Entrepreneurship and Innovation Support Program, and CAAI-Huawei MindSpore Open Fund.

References

Bao, H., Niu, G., and Sugiyama, M. (2018). Classification from pairwise similarity and unlabeled data. In *ICML*.

Bao, H., Scott, C., and Sugiyama, M. (2020). Calibrated surrogate losses for adversarially robust classification. In *COLT*.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2002). Localized rademacher complexities. In *COLT*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*

Brinker, K. and Hüllermeier, E. (2019). A reduction of label ranking to multiclass classification. In *ECML/PKDD*.

Cao, Y., Liu, S., and Xu, Y. (2022). Multi-complementary and unlabeled learning for arbitrary losses and models. *Pattern Recognition*.

Chapelle, O., Scholkopf, B., and Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*.

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

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

Cour, T., Sapp, B., and Taskar, B. (2011). Learning from partial labels. *J. Mach. Learn. Res.*

Du Plessis, M. C., Niu, G., and Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. *NeurIPS*.

Feng, L. and An, B. (2018). Leveraging latent label distributions for partial label learning. In *IJCAI*.

Feng, L., Kaneko, T., Han, B., Niu, G., An, B., and Sugiyama, M. (2020). Learning with multiple complementary labels. In *ICML*.

Fotakis, D., Kalavasis, A., and Psaroudaki, E. (2022). Label ranking through nonparametric regression. In *ICML*.

Gao, Y. and Zhang, M. (2021). Discriminative complementary-label learning with weighted loss. In *ICML*.

Ghosh, A., Kumar, H., and Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. In *AAAI*.

Golovnev, A., Pál, D., and Szorenyi, B. (2019). The information-theoretic value of unlabeled data in semi-supervised learning. In *ICML*.

Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I., and Sugiyama, M. (2020). Sigua: Forgetting may make learning with noisy labels more robust. In *ICML*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

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

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. (2017). Learning from complementary labels. In *NeurIPS*.

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

Kim, Y., Yim, J., Yun, J., and Kim, J. (2019). Nlnl: Negative learning for noisy labels. In *ICCV*.

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

Kiryo, R., Niu, G., Du Plessis, M. C., and Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. *NeurIPS*.

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

Liu, Y. and Guo, H. (2020). Peer loss functions: Learning from noisy labels without knowing noise rates. In *ICML*.

Lu, N., Lei, S., Niu, G., Sato, I., and Sugiyama, M. (2021). Binary classification from multiple unlabeled datasets via surrogate set classification. In *ICML*.

Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. (2018). On the minimal supervision for training any binary classifier from only unlabeled data. *arXiv preprint arXiv:1808.10585*.

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

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

Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S., Xia, S., Wijewickrema, S., and Bailey, J. (2018). Dimensionality-driven learning with noisy labels. In *ICML*.

Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In *ALT*.

Mendelson, S. (2008). Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inf. Theory*.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on Pattern Analysis and Machine Intelligence*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. MIT press.

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

Ni, C., Charoenphakdee, N., Honda, J., and Sugiyama, M. (2019). On the calibration of multiclass classification with rejection. In *NeurIPS*.

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

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

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*.

Sakai, T., Niu, G., and Sugiyama, M. (2018). Semi-supervised auc optimization based on positive-unlabeled learning. *Machine Learning*.

Sugiyama, M. (2015). *Introduction to Statistical Machine learning*. Morgan Kaufmann.

Vogel, R. and Clémençon, S. (2020). A multiclass classification approach to label ranking. In *AISTATS*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747.

Yu, X., Liu, T., Gong, M., and Tao, D. (2018). Learning with biased complementary labels. In *ECCV*, pages 68–83.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*

Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *NeurIPS*.

Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*.

A Proof of Theorem 4.1

Proof. In this section, we show the classifier-consistency by showing that $\operatorname{argmax}_y u_y = \operatorname{argmax}_y p(y|\boldsymbol{x})$, where $\boldsymbol{u} = \operatorname{argmin}_{\boldsymbol{u}'} r(\boldsymbol{u}') = \operatorname{argmin}_{\boldsymbol{u}'} \sum_{y=1}^{K} \bar{p}(y|\boldsymbol{x}) w(\boldsymbol{u}', y)) \ell(-\boldsymbol{u}', y)$.

First of all, we show that for any $y, y' \in \mathcal{Y}$, if $p(y|\boldsymbol{x}) > p(y'|\boldsymbol{x})$, then $u_y \geq u_{y'}$. According to the theorem, it is easy to find that if $u_y < u_{y'}$, then $w(\boldsymbol{u}, y))\ell(-\boldsymbol{u}, y) < w(\boldsymbol{u}, y'))\ell(-\boldsymbol{u}, y')$. According to the definition of complementary label, we can learn that $\bar{p}(y|\boldsymbol{x}) < \bar{p}(y'|\boldsymbol{x})$. Then we can get a new $\boldsymbol{u}'$ by swapping u_y and $u_{y'}$ to get a $r(\boldsymbol{u}') < r(\boldsymbol{u})$.

Then we show that it is necessary for $\boldsymbol{u}$ to fulfill the condition below if it is also a stationary point of $r(\boldsymbol{u})$:

$$p(y|\boldsymbol{x}) > p(y'|\boldsymbol{x}) \rightarrow u_y > u_{y'}.$$

Since we have proved that $p(y|\boldsymbol{x}) > p(y'|\boldsymbol{x}) \rightarrow u_y \geq u_{y'}$, we only have to show that the equality does not hold. If $u_y = u_{y'}$ but $p(y|\boldsymbol{x}) > p(y'|\boldsymbol{x})$, we can learn from the monotonicity and symmetricity that $\frac{\partial r(\boldsymbol{u})}{\partial u_y} < \frac{\partial r(\boldsymbol{u})}{\partial u_{y'}}$, which indicates that only one of the partial derivatives can be zero, i.e., $\boldsymbol{u}$ is not the stationary point.

Then we show that the points out of the interior of $\mathbb{R}^K$ cannot be the minima of $r(\boldsymbol{u})$, based on the fact that the weighted function is lower bounded by a positive number.

For OvA Logistic loss, we can easily find that if there exists any $u_y = +\infty$ or $-\infty$, $r(\boldsymbol{u})$ is also $+\infty$.

For Pairwise Logistic loss, we can easily learn that $r(\boldsymbol{u}) = +\infty$ if there exists $u_y = \infty$ and $u_{y'} < \infty$. When $u_y = \infty$ for all the $y \in \mathcal{Y}$, we can learn that $\ell(-\boldsymbol{u}, y) = \log 2$ for any y. Notice that we can set all the $u_y = a \neq \infty$ to get the same value. From the derivative analysis, we can learn that the gradient of $u_y = a$ is non-zero from the previous analysis, which indicates that $u_y = \infty$ for all the $y \in \mathcal{Y}$ cannot be a solution.

For CE loss and Focal loss, we assume that $p(y|\boldsymbol{x}) \neq 0$ without loss of generality. If not all the $u_y = +\infty$ or $-\infty$, we can learn that there exists y that $s_y(-\boldsymbol{u}) = 0$ and hence $r(\boldsymbol{u}) = +\infty$. When all the $u_y = +\infty$ or $-\infty$, we can learn that $s_y(-\boldsymbol{u}) = \frac{1}{K}$ for any y. In the same way as in the previous paragraph, we can also learn the suboptimality of this $\boldsymbol{u}$.

Combining the conclusions above and we can conclude the proof. □

B Proofs of Theorem 5.1

Our proof of the estimation error bound is based on Rademacher complexity:

Definition B.1. (Rademacher complexity) Let $Z_1, \ldots, Z_n$ be n i.i.d. random variables drawn from a probability distribution $\mu, \mathcal{H} = \{h : \mathcal{Z} \rightarrow \mathbb{R}\}$ be a class of measurable functions. Then the expected Rademacher complexity of $\mathcal{H}$ is defined as

$$\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{Z_1,\ldots,Z_n \sim \mu} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h\left(Z_i\right) \right]$$

where $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$ are Rademacher variables taking the value from $\{-1, +1\}$ with even probabilities.

First, let $\mathcal{G} \subset \mathcal{X} \rightarrow \mathbb{R}^K$ be the model class and each of its dimension is constructed by $\mathcal{G}_y \subset \mathcal{X} \rightarrow \mathbb{R}$. Then, we define the following function space for our order-preserving method

$$\mathcal{L} \circ \mathcal{G} = \{h : (\boldsymbol{x}, \bar{y}) \mapsto \ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y}) \mid \boldsymbol{g} \in \mathcal{G}\}$$

So the Rademacher complexity of $\{\mathcal{L} \circ \mathcal{G}\}$ given n $i.i.d.$ samples drawn from distribution with density $\bar{p}(\boldsymbol{x}, \bar{y})$ can be defined as

$$\mathfrak{R}_n(\mathcal{L} \circ \mathcal{G}) = \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(\boldsymbol{x}_i, \bar{y}_i) \right].$$

Before proving Theorem 5.1, we introduce the following lemmas.

Lemma B.1. Suppose $\hat{\bar{\boldsymbol{g}}}$ is the empirical risk minimizer (i.e., $\hat{\bar{\boldsymbol{g}}} = \operatorname{argmin}_{\boldsymbol{g} \in \mathcal{G}} \hat{\bar{R}}(\boldsymbol{g}, \ell)$) and $\bar{\boldsymbol{g}}*$ is the true minimizer (i.e., $\bar{\boldsymbol{g}}^* = \operatorname{argmin}_{\boldsymbol{g} \in \mathcal{G}} \bar{R}(\boldsymbol{g}, \ell)$), then the following inequality holds:

$$\bar{R}(\hat{\bar{\boldsymbol{g}}}) - \bar{R}(\bar{\boldsymbol{g}}^*) \leq 2 \sup_{\boldsymbol{g} \in \mathcal{G}} |\hat{\bar{R}}(\boldsymbol{g}) - \bar{R}(\boldsymbol{g})|$$

Proof. It is intuitive to obtain

$$\begin{aligned}\bar{R}(\hat{\boldsymbol{g}}) - \bar{R}(\bar{\boldsymbol{g}}^*) &= \bar{R}(\hat{\boldsymbol{g}}) - \hat{\bar{R}}(\hat{\boldsymbol{g}}) + \hat{\bar{R}}(\hat{\boldsymbol{g}}) - \bar{R}(\bar{\boldsymbol{g}}^*) \\ &\leq \bar{R}(\hat{\boldsymbol{g}}) - \hat{\bar{R}}(\hat{\boldsymbol{g}}) + \hat{\bar{R}}(\bar{\boldsymbol{g}}^*) - \bar{R}(\bar{\boldsymbol{g}}^*) \\ &\leq 2\sup_{\boldsymbol{g}\in\mathcal{G}} \left|\hat{\bar{R}}(\boldsymbol{g}) - \bar{R}(\boldsymbol{g})\right|\end{aligned}$$

which completes the proof. □

Then we have the following lemma.

Lemma B.2. Assume there exists $C_{\boldsymbol{g}} > 0$ that $\sup_{\boldsymbol{g}\in\mathcal{G}} \|\boldsymbol{g}\|_\infty \leq C_{\boldsymbol{g}}$ and $C_\ell > 0$ such that $\sup_{\|\boldsymbol{z}\|_\infty \leq C_{\boldsymbol{g}}} \ell(\boldsymbol{z}, y) \leq C_\ell$ for any $y \in \mathcal{Y}$. We also assume the loss function $\ell(\boldsymbol{z}, y)$ is L_ℓ-Lipschitz continuous $w.r.t.$ $\boldsymbol{z}$ for all $y \in \mathcal{Y}$. Then for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{\boldsymbol{g}\in\mathcal{G}} \left|\hat{\bar{R}}(\boldsymbol{g}) - \bar{R}(\boldsymbol{g})\right| \leq 2\sqrt{2}L_\ell \sum_{y=1}^{K} \mathfrak{R}_n(\mathcal{G}_y) + 2C_\ell\sqrt{\frac{ln(2/\delta)}{2n}}$$

Proof. We will only discuss a one-sided bound on $\sup_{\boldsymbol{g}\in\mathcal{G}} \left(\hat{\bar{R}}(\boldsymbol{g}) - \bar{R}(\boldsymbol{g})\right)$ that holds with probability at least $1 - \frac{\delta}{2}$. The other side can be derived in a similar way. To begin with, we bound the change of $\sup_{\boldsymbol{g}\in\mathcal{G}} \left(\hat{\bar{R}}(\boldsymbol{g}) - \bar{R}(\boldsymbol{g})\right)$ when a single entry $z_i = (\boldsymbol{x}_i, y_i)$ of $(z_1, \ldots, z_n)$ is replaced with $z_i' = (\boldsymbol{x}_i', y_i')$. Define $A(z_1, \ldots, z_n) = \sup_{\boldsymbol{g}\in\mathcal{G}} \left(\hat{\bar{R}}(\boldsymbol{g}) - \bar{R}(\boldsymbol{g})\right)$. Then it holds that

$$\begin{aligned}&A(z_1, \ldots, z_i, \ldots, z_n) - A(z_1, \ldots, z_i', \ldots, z_n) \\ &\quad= \sup_{\boldsymbol{g}\in\mathcal{G}} \left[\frac{1}{n}\sum_{j=1}^{n} \ell(-\boldsymbol{g}(\boldsymbol{x}_j), \bar{y}_j) - \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y})]\right] \\ &\qquad- \sup_{\boldsymbol{g}'\in\mathcal{G}} \left[\frac{1}{n}\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} \ell(-\boldsymbol{g}'(\boldsymbol{x}_j), \bar{y}_j) + \frac{1}{n}\ell(-\boldsymbol{g}'(\boldsymbol{x}_i'), \bar{y}_i') - \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(-\boldsymbol{g}'(\boldsymbol{x}), \bar{y})]\right] \\ &\quad\leq \sup_{\boldsymbol{g}\in\mathcal{G}} \left[\frac{1}{n}\sum_{j=1}^{n} \ell(-\boldsymbol{g}'(\boldsymbol{x}_j), \bar{y}_j) - \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y})]\right. \\ &\qquad\left.-\frac{1}{n}\sum_{j\in\{1,\ldots,n\}\setminus\{i\}} \ell(-\boldsymbol{g}(\boldsymbol{x}_j), \bar{y}_j) - \frac{1}{n}\ell(-\boldsymbol{g}(\boldsymbol{x}_i'), \bar{y}_i') + \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\ell(-\boldsymbol{g}(\boldsymbol{x}), \bar{y})]\right] \\ &\quad= \sup_{\boldsymbol{g}\in\mathcal{G}} \left[\frac{1}{n}\ell(-\boldsymbol{g}(\boldsymbol{x}_i), \bar{y}_i) - \frac{1}{n}\ell(-\boldsymbol{g}(\boldsymbol{x}_i'), \bar{y}_i')\right] \\ &\quad= \frac{1}{n}\sup_{\boldsymbol{g}\in\mathcal{G}} [\ell(-\boldsymbol{g}(\boldsymbol{x}_i), \bar{y}_i) - \ell(-\boldsymbol{g}(\boldsymbol{x}_i'), \bar{y}_i')] \leq \frac{2C_\ell}{n}\end{aligned}$$

By applying McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta/2$,

$$\sup_{\boldsymbol{g}\in\mathcal{G}} \bar{R}(\hat{\boldsymbol{g}}) - \bar{R}(\bar{\boldsymbol{g}}) \leq \mathbb{E}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \bar{R}(\hat{\boldsymbol{g}}) - \bar{R}(\bar{\boldsymbol{g}})\right] + 2C_\ell\sqrt{\frac{ln(2/\delta)}{2n}}$$

Since $\bar{R}(\boldsymbol{g}) = \mathbb{E}\left[\hat{\bar{R}}(\boldsymbol{g})\right]$, by applying the symmetrization (Mohri et al., 2018), we get

$$\mathbb{E}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \left(\hat{\bar{R}}(\boldsymbol{g}) - \bar{R}(\boldsymbol{g})\right)\right] \leq 2\mathfrak{R}_n(\mathcal{L}\circ\mathcal{G}).$$

By further taking into account the other side $\sup_{\boldsymbol{g}\in\mathcal{G}}\left(\bar{R}(\boldsymbol{g})-\hat{\bar{R}}(\boldsymbol{g})\right)$, we have for any $\delta>0$, with probability at least $1-\delta$,

$$\sup_{\boldsymbol{g}\in\mathcal{G}}\left|\hat{\bar{R}}(\boldsymbol{g})-\bar{R}(\boldsymbol{g})\right|\leq 2\mathfrak{R}_n(\mathcal{L}\circ\mathcal{G})+2C_\ell\sqrt{\frac{ln(2/\delta)}{2n}},$$

Since the loss function $\ell(\boldsymbol{z},y)$ is L_ℓ-Lipschitz continuous $w.r.t.$ $\boldsymbol{z}$ for all $y\in\mathcal{Y}$, by the Rademacher vector contraction inequality (), we have $\mathfrak{R}_n(\mathcal{L}\circ\mathcal{G})\leq\sqrt{2}L_\ell\sum_{y=1}^{K}\mathfrak{R}_n(\mathcal{G}_y)$, which concludes the proof. □

Combing Lemma B.1 and Lemma B.2, Theorem 5.1 is proved.

C Proofs of Theorem 5.2

Since this proof is somewhat similar to the proof of Theorem 5.1, we briefly sketch the key points.

We define a function space for our weighting method as

$$\mathcal{L}_w\circ\mathcal{G}=\{h:(\boldsymbol{x},\bar{y})\mapsto w(\boldsymbol{g}(\boldsymbol{x}),\bar{y})\cdot\ell(-\boldsymbol{g}(\boldsymbol{x}),\bar{y})\mid\boldsymbol{g}\in\mathcal{G}\}$$

So the Rademacher complexity of $\{\mathcal{L}_w\circ\mathcal{G}\}$ given n $i.i.d.$ samples drawn from distribution with density $\bar{p}(\boldsymbol{x},\bar{y})$ can be defined as

$$\mathfrak{R}_n(\mathcal{L}_w\circ\mathcal{G})=\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(\boldsymbol{x}_i,\bar{y}_i)\right].$$

Then we have the following lemma.

Lemma C.1. Suppose the weight function is upper bounded by C_w and is L_w-Lipschitz continuous for any $y\in\mathcal{Y}$. And we assume that $\hat{\bar{\boldsymbol{g}}}^w$ is the empirical risk minimizer (i.e., $\hat{\bar{\boldsymbol{g}}}^w=\operatorname{argmin}_{\boldsymbol{g}\in\mathcal{G}}\hat{\bar{R}}^w(\boldsymbol{g},\ell)$) and $\bar{\boldsymbol{g}}^{*w}$ is the true minimizer (i.e., $\bar{\boldsymbol{g}}^*=\operatorname{argmin}_{\boldsymbol{g}\in\mathcal{G}}\bar{R}^w(\boldsymbol{g},\ell)$), then for any $\delta>0$, with probability at least $1-\delta$,

$$\sup_{\boldsymbol{g}\in\mathcal{G}}\left|\bar{R}^w(\hat{\bar{\boldsymbol{g}}})-\bar{R}^w(\bar{\boldsymbol{g}})\right|\leq 2\sqrt{2}L'\sum_{y=1}^{K}\mathfrak{R}_n(\mathcal{G}_y)+2C'\sqrt{\frac{ln(2/\delta)}{2n}},$$

where $L'=L_\ell C_w+L_w C_\ell$ and $C'=C_w C_\ell$.

Proof. In order to prove this lemma, we first show that the one direction $\sup_{\boldsymbol{g}\in\mathcal{G}}\bar{R}^w(\hat{\bar{\boldsymbol{g}}})-\bar{R}^w(\bar{\boldsymbol{g}})$ is bounded with probability at least $1-\delta/2$, and the other direction can be similarly shown. Suppose an example $(\boldsymbol{x}_i,\bar{y}_i)$ is replaced by another arbitrary example $(\boldsymbol{x}_i',\bar{y}_i')$, then the change of $\sup_{\boldsymbol{g}\in\mathcal{G}}\bar{R}^w(\hat{\bar{\boldsymbol{g}}})-\bar{R}^w(\bar{\boldsymbol{g}})$ is no greater than $2C_wC_\ell/n$. By applying McDiarmid's inequality, for any $\delta>0$, with probability at least $1-\delta/2$,

$$\sup_{\boldsymbol{g}\in\mathcal{G}}\bar{R}^w(\hat{\bar{\boldsymbol{g}}})-\bar{R}^w(\bar{\boldsymbol{g}})\leq\mathbb{E}\left[\sup_{\boldsymbol{g}\in\mathcal{G}}\bar{R}^w(\hat{\bar{\boldsymbol{g}}})-\bar{R}^w(\bar{\boldsymbol{g}})\right]+2C_wC_\ell\sqrt{\frac{ln(2/\delta)}{2n}}$$

Since $\bar{R}^w(\boldsymbol{g})=\mathbb{E}\left[\hat{\bar{R}}^w(\boldsymbol{g})\right]$, by applying the symmetrization (Mohri et al., 2018), we get

$$\mathbb{E}\left[\sup_{\boldsymbol{g}\in\mathcal{G}}\left(\hat{\bar{R}}^w(\boldsymbol{g})-\bar{R}(\boldsymbol{g})\right)\right]\leq 2\mathfrak{R}_n(\mathcal{L}_w\circ\mathcal{G}).$$

By further taking into account the other side $\sup_{\boldsymbol{g}\in\mathcal{G}}\left(\bar{R}^w(\boldsymbol{g})-\hat{\bar{R}}^w(\boldsymbol{g})\right)$, we have for any $\delta>0$, with probability at least $1-\delta$,

$$\sup_{\boldsymbol{g}\in\mathcal{G}}\left|\hat{\bar{R}}(\boldsymbol{g})-\bar{R}(\boldsymbol{g})\right|\leq 2\mathfrak{R}_n(\mathcal{L}_w\circ\mathcal{G})+2C_wC_\ell\sqrt{\frac{ln(2/\delta)}{2n}},$$

Since the loss function $\ell(\boldsymbol{z},y)$ is L_ℓ-Lipschitz continuous $w.r.t.$ $\boldsymbol{z}$ for all $y\in\mathcal{Y}$, and the weight function is L_w-Lipschitz continuous for any $y\in\mathcal{Y}$, by the Rademacher vector contraction inequality (Maurer, 2016), we have $\mathfrak{R}_n(\mathcal{L}_w\circ\mathcal{G})\leq\sqrt{2}(L_\ell C_w+L_wC_\ell)\sum_{y=1}^{K}\mathfrak{R}_n(\mathcal{G}_y)$, which concludes the proof. □

By taking into account Lemma B.1 and Lemma C.1, Theorem 5.2 is proved.

D Details of Experimental Setup

We generate the complementary labels following the setting in Ishida et al. (2017). In the experiments of MNIST, Fashion-MNIST and Kuzushiji-MNIST, we use Adam with default momentum and learning rate, batch size and weight decay were set to 5e-5, 256, and 1e-4, respectively. For SVHN, the learning rate was set to 5e-4 and other parameters remain the same. The weight function for our OP-W is set to: $w(\boldsymbol{g}(\boldsymbol{x}), y) = s_y(\boldsymbol{u}(\boldsymbol{x}) + 1) * s_y(\boldsymbol{g}(\boldsymbol{x})) + \epsilon$, where $\epsilon = 1e - 6$ and $u_y(\boldsymbol{x}) = 1/s_y(-\boldsymbol{g}(\boldsymbol{x}))$. For all the datasets, we split 10% of the training set as the validation set.

E Limitations and Societal Impact

This framework is used for single complementary label learning, while it is also available to learn from multiple complementary labels (Feng et al., 2020). We believe that extensions to MCL is a promising future direction. The use of complementary-label learning can improve the ability of privacy protection as stated in Ishida et al. (2019); Feng et al. (2020), and thus there may not be severe negative societal impact.