RETHINKING ONE-VS-THE-REST LOSS FOR INSTANCE-DEPENDENT COMPLEMENTARY LABEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Complementary Label Learning (CLL) is a typical weakly supervised learning protocol, where each instance is associated with one complementary label, which specifies the class that the instance does not belong to. Existing CLL methods assume that the complementary label is sampled uniformly from all non-ground-truth labels, or from a biased probability depending on the ground-truth label. However, these assumptions are normally unrealistic, for example, an annotator tends to choose a label that is largely irrelevant to the instance to avoid mistaking the ground-truth label as the complementary one. Therefore, in this paper, we introduce instance-dependent CLL (IDCLL), where non-ground-truth labels that are less relevant to the instances are more likely to be selected as the complementary ones. Accordingly, we present our generation process for instance-dependent complementary label and observe that directly applying existing CLL methods to IDCLL results in poor performance. We further empirically analyze this phenomenon and identify: Existing methods exhibit a decline in their capacity to share complementary labels under the instance-dependent setting, resulting in small logit margins, thus difficult to identify ground-truth labels. To address this problem, we introduce *complementary logit margin loss* (CLML) and demonstrate CLML can enhance the capacity to share complementary labels. Additionally, we propose a novel form of the complementary one-vs-the-rest loss (COVR) as the surrogate loss for CLML, and provide theoretical proof that COVR can decrease CLML to a greater extent compared to existing CLL methods. The estimation error bound of the proposed COVR is also theoretically characterized. Extensive experiments conducted on benchmark datasets demonstrate the superiority of the proposed method compared to the existing CLL methods under our instance-dependent setting.

1 INTRODUCTION

In ordinary supervised classification, each instance is specified to the class it belongs to (Xue & Hauskrecht, 2019). However, accurately annotating large-scale datasets in a fully supervised manner can be both time-consuming and costly. To overcome this problem, weakly supervised learning protocols have gained increasing attention in recent years, including partial label learning (Tian et al., 2023; Wang et al., 2022), semi-supervised learning (Kou et al., 2023; Guo et al., 2022), noisy label learning (Natarajan et al., 2013; Li et al., 2022), and positive-unlabeled learning (Elkan & Noto, 2008; Wilton et al., 2022).

Particularly, in this paper, we consider another weakly supervised learning protocol called complementary label learning (CLL) (Ishida et al., 2017; 2019), where each instance is only associated with a complementary label to specify a class that the instance *does not* belong to. Then, CLL aims to learn a classifier to predict the ground-truth label for each instance by leveraging the complementary labels. Existing CLL methods primarily focus on two assumptions for the generation of the complementary label: the uniform assumption and the biased one. The uniform assumption assumes that each complementary label is sampled *uniformly* from all labels except the ground-truth label (Ishida et al., 2017; 2019; Chou et al., 2020), while the biased assumption considers that the complementary label is selected from a biased probabilities only depending on the ground-truth label (Yu et al., 2018). In other words, traditional CLL methods consider that the selection of complementary labels is *independent* of the instance.

Nevertheless, in real-world scenarios, the selection of complementary labels is mostly instance-dependent: An annotator tends to choose a complementary label that is largely unrelated to the instance, in order to ensure the correctness of the selection. Consequently, in this paper, we propose a realistic instance-dependent assumption:

The non-ground-truth labels with lower relevance to the instance are more likely to be chosen as the complementary labels than the ones with higher relevance.

Based on this assumption, we formally introduce the protocol of instance-dependent complementary label learning (IDCLL), which assumes that the selection of each instance-dependent complementary label (IDCL) explicitly/implicitly depends on the instance. In particular, to facilitate the generation of IDCL, we propose a systematic mechanism named "Min$k$" that utilizes a selection probability $\bar{p}(\bar{Y}|Y, X)$, which indicates the likelihood of selecting a non-ground-truth label as the complementary one based on its relevance to the instance. Specifically, a pre-trained model is used to obtain the selection probability and k non-ground-truth labels with the highest selection probabilities are chosen. Subsequently, we randomly designate one label as the IDCL from the k selected labels. Moreover, we present the definition of *complementary label distribution* and compare its difference between uniform and instance-dependent settings.

A straightforward attempt to solve IDCLL is to use existing CLL methods (Ishida et al., 2017; Yu et al., 2018; Chou et al., 2020) employed for instance-independent settings. However, our empirical study suggests that directly applying existing CLL methods to IDCLL leads to poor performance. In order to further investigate the underlying reason, we conduct in-depth investigation on existing CLL losses and explain why these methods work well under uniform setting. We first introduce the share complementary label hypothesis and logit margin loss (LML), and provide insight into the root cause of the poor performance: The sparser complementary label distribution under instance-dependent settings diminishes the capacity of existing CLL methods to effectively share complementary labels, resulting in small logit margin losses and making it challenging to disambiguate for the potential ground-truth labels. To tackle this challenge, we introduce complementary logit margin loss (CLML) and demonstrate that minimizing CLML can enhance model's capacity to share complementary labels. We also tailor a novel complementary one-versus-rest loss (COVR) as a surrogate loss to CLML. Our empirical and theoretical analysis has verified that the proposed COVR significantly enhances the capacity to share complementary labels and decreases CLML, which ultimately leads to an improved performance in identifying the ground-truth labels.

Our contributions are summarized in the following four levels:

- Problem setting level: We for the first time introduce the concept of IDCLL and provide a systematic mechanism for generating IDCLs according to our assumption.
- Empirical study level: We demonstrate the poor performance of existing CLL methods under the instance-dependent setting and conduct thorough empirical analysis to identify the underlying cause, i.e., the weaken capacity to share complementary labels.
- Methodology level: To enhance the capacity to share complementary label, we accordingly introduce CLML and propose COVR as the surrogate loss to CLML.
- Experimental level: We conduct extensive experiments on benchmark datasets to verify the effectiveness and superiority of the proposed method under different instance-dependent settings.

2 RELATED WORK

Complementary label learning (CLL) To solve CLL under uniform assumption, an unbiased risk estimator (URE) was derived using a specific complementary loss function (e.g., one-versus-all or pairwise comparison) that satisfies a symmetric condition (Ishida et al., 2017). Additionally, a more general URE for arbitrary loss functions and models was derived: To alleviate the overfitting issue of this URE in practice, non-negative correction and gradient ascent methods were further proposed (Ishida et al., 2019). However, URE suffers from huge gradient variance and results in unsatisfactory performance in practice. In order to mitigate this issue, the surrogate complementary loss framework (SCL) was proposed to reduce the gradient variance (Chou et al., 2020). Moreover, several promising approaches have been proposed to solve CLL, such as using weighted complementary loss (Gao & Zhang, 2021) and introducing partial-output regularization (Liu et al., 2023). Different from the uniform assumption, biased CLL was considered where the selection of complementary labels depends on a biased probabilities (Yu et al., 2018). We should emphasize that IDCLL is different from biased CLL and a comprehensive understanding of the distinctions can be found in Appendix G.

Instance-dependent weakly supervised learning Instance-dependent assumption has been introduced into two weakly supervised learning protocols: Instance-dependent partial label learning (IDPLL) and instance-dependent label noise learning (IDN). IDPLL assumes that the labels with high relevance to the instance are more likely to be selected in the candidate set (Xu et al., 2021). Recovering the latent label distribution (Xu et al., 2021), performing Maximum A Posterior (MAP) based on an explicitly model generation process of candidate labels (Qiao et al., 2023), and purifying the candidate set with dynamic thresholds (Xu et al., 2022) has been proposed to solve IDPLL. Similarly, IDN assumes that poor quality or ambiguous instances in real-world datasets are more likely to be mislabeled (Garg et al., 2023), including the instance-dependent dichotomous label noise (Menon et al., 2018), bounded instance-dependent label noise (Cheng et al., 2020), and instance-dependent noise with confidence scores (Berthon et al., 2021). Moreover, several work directly estimated the transition matrix to solve IDN (Yang et al., 2022; Cheng et al., 2022; Yao et al., 2021; Xia et al., 2020). Our work is the first to introduce the instance-dependent assumption to CLL.

3 PRELIMINARIES

Different from supervised classification, each instance is only provided with *single* complementary label in CLL. Let $\bar{\mathcal{D}} = \{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$ denotes the complementary dataset, sampled from an unknown probability distribution $\bar{p}(\boldsymbol{x}, \bar{y})$, where $\bar{y}_i \in \mathcal{Y}\backslash\{y_i\}$ is the complementary label of the instance $\boldsymbol{x}_i$. The goal of CLL is to learn a DNN classifier $h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) : \mathcal{X} \rightarrow \mathcal{Y}$, which is expected to predict the ground-truth label of an input $\boldsymbol{x}_i$:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \underset{k\in\{1,2,\ldots,K\}}{\operatorname{argmax}} f_k(\boldsymbol{x}_i, \boldsymbol{\theta}), \quad f_k(\boldsymbol{x}_i, \boldsymbol{\theta}) = e^{g_k(\boldsymbol{x}_i,\boldsymbol{\theta})} / \sum_{j=1}^{K} e^{g_j(\boldsymbol{x}_i,\boldsymbol{\theta})}, \tag{1}$$

where $\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) = [g_1(\boldsymbol{x}_i, \boldsymbol{\theta}), g_2(\boldsymbol{x}_i, \boldsymbol{\theta}), .., g_K(\boldsymbol{x}_i, \boldsymbol{\theta})]^T$ and $g_k(\boldsymbol{x}_i, \boldsymbol{\theta})$ is the k-th logit of the model with respect to the instance $\boldsymbol{x_i}$. Meanwhile, $\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\theta}) = [f_1(\boldsymbol{x}_i, \boldsymbol{\theta}), f_2(\boldsymbol{x}_i, \boldsymbol{\theta}), .., f_k(\boldsymbol{x}_i, \boldsymbol{\theta})]^T$ and $f_k(\boldsymbol{x}_i, \boldsymbol{\theta})$ is the predicted probability of $\boldsymbol{x}_i$ belonging to class k, i.e., $p_{\boldsymbol{\theta}}(Y = k|X = \boldsymbol{x}_i)$.

A commonly used CLL method based on DNNs is the Surrogate Complementary Loss (SCL) framework(Chou et al., 2020), which defines a novel Complementary 0-1 Loss:

$$\bar{\ell}_{01}(\bar{y}, h_{\boldsymbol{\theta}}(\boldsymbol{x})) = \mathbb{1}(h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \bar{y}), \tag{2}$$

where $\mathbb{1}$ is the indicator function. Based on the Complementary 0-1 Loss, the expected complementary classification risk of classifier $h_{\boldsymbol{\theta}}$ on $\bar{p}(\boldsymbol{x}, \bar{y})$ can be defined as:

$$\bar{R}(h_{\boldsymbol{\theta}}; \bar{\ell}_{01}) = \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\mathbb{1}(h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \bar{y})]. \tag{3}$$

Consequently, the expected complementary classification risk $\bar{R}(h_{\boldsymbol{\theta}}; \bar{\ell}_{01})$ is an Unbiased Risk Estimator (URE) of the expected risk of supervised classification $R(h_{\boldsymbol{\theta}}; \ell_{01})$ (Chou et al., 2020):

$$R(h_{\boldsymbol{\theta}}; \ell_{01}) = (K - 1)\bar{R}(h_{\boldsymbol{\theta}}; \bar{\ell}_{01}). \tag{4}$$

In other words, minimizing the empirical complementary 0-1 risk is equivalent to the supervised learning. In order to mitigate the overfitting issue and improve traditional URE-based methods, Chou et al. (2020) proposed to use negative learning loss (SCL_NL) as the Surrogate Complementary Loss:

$$\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{f}(\boldsymbol{x})) = -\log(1 - f_{\bar{y}}(\boldsymbol{x})) = -\log(1 - e^{g_{\bar{y}}(\boldsymbol{x})} / \sum_{j=1}^{K} e^{g_j(\boldsymbol{x})}). \tag{5}$$

4 INSTANCE-DEPENDENT COMPLEMENTARY LABEL LEARNING

4.1 GENERATION PROCESS FOR IDCLS

In this section, we introduce the generation process for IDCLs based on the instance-dependent assumption. In general, we select the IDCL for each instance $\boldsymbol{x}$ according to the probability $\bar{p}(\bar{Y} =$

Table 1: Classification accuracy (mean±std) of each baseline approach on benchmark datasets under uniform and "Min3" settings, respectively. The symbol ↓ indicates a decrease in classification accuracy for these approaches under the "Min3" setting compared to the uniform setting.

Dataset	MNIST		Kuzushiji-MNIST		Fashion-MNIST		CIFAR-10		SVHN	
Setting	Uniform	Min3	Uniform	Min3	Uniform	Min3	Uniform	Min3	Uniform	Min3
W_Loss	97.14 ± 0.28	46.19 ± 4.56 ↓	77.01 ± 1.16	46.73 ± 3.94 ↓	83.42 ± 0.66	42.61 ± 6.14 ↓	72.11 ± 1.14	45.57 ± 0.63 ↓	79.54 ± 0.21	38.46 ± 3.13 ↓
Forward	98.00 ± 0.07	51.82 ± 6.81 ↓	78.05 ± 0.74	53.69 ± 4.51 ↓	85.16 ± 0.32	39.54 ± 0.13 ↓	76.56 ± 0.97	45.12 ± 2.03 ↓	87.22 ± 2.43	31.43 ± 2.86 ↓
SCL_NL	98.06 ± 0.11	55.56 ± 6.91 ↓	78.21 ± 0.35	51.33 ± 1.76 ↓	85.18 ± 0.31	39.53 ± 0.08 ↓	74.94 ± 4.06	38.40 ± 0.34 ↓	82.32 ± 4.72	32.97 ± 2.18 ↓
SCL_EXP	97.73 ± 0.15	33.78 ± 4.03 ↓	76.86 ± 0.36	42.78 ± 4.20 ↓	84.70 ± 0.13	37.06 ± 1.60 ↓	74.02 ± 0.26	30.78 ± 0.94 ↓	79.58 ± 5.06	20.59 ± 0.06 ↓

$\bar{y}|Y=y, X=\boldsymbol{x})$, which reflects the likelihood of $\bar{y}$ being selected as the IDCL for $\boldsymbol{x}$. Recent work suggests that the prediction probability of a DNN classifier, i.e., $p(Y|X)$, indicates the label's relevance to the corresponding instance (Wu et al., 2018). Specifically, lower prediction probability means less association between the label and instance. Hence, we consider to utilize the prediction probability of a pre-trained model to generate the non-noisy IDCL by introducing the following assumption (Gao & Zhang, 2021):

Assumption 1. *For an instance $\boldsymbol{x}$ with ground-truth y, the selection probability $\bar{p}(\bar{Y}=j|Y=y, X=\boldsymbol{x})=\frac{\exp(1-p(Y=j|X=\boldsymbol{x}))}{\sum_{k\neq y}\exp(1-p(Y=k|X=\boldsymbol{x}))}, j\in\mathcal{Y}\backslash\{y\}$, and $\bar{p}(\bar{Y}=y|Y=y, X=\boldsymbol{x})=0$.*

Based on Assumption 1, the non-ground-truth labels has lower relevance to the instance are prone to be chosen as IDCL due to the higher selection probabilities. To further study IDCLL, we present our generation process of IDCLs named "Min$k$" through the selection probability for each instance.

Definition 1 (Min$k$)**.** *We utilize a pre-trained model to obtain prediction probability $p(Y|X=\boldsymbol{x})$ for each instance $\boldsymbol{x}$, and calculate the selection probability $\bar{p}(\bar{Y}|Y=y, X=\boldsymbol{x})$ based on Assumption 1. Subsequently, we randomly select one complementary label for each instance from the labels with the k highest selection probabilities.*

The rationale behind the above "Min$k$" stems from the belief that, although individuals may adhere to our assumption when selecting IDCLs, they actually do not consider the labels with high correlation to the ground-truth one. More importantly, this design allows us to generate various settings by varying the value of k and recovery the uniform setting with $k=K-1$, thus simulating the diverse scenarios encountered in practice. Further analysis on the rationality can be found in Appendix E.

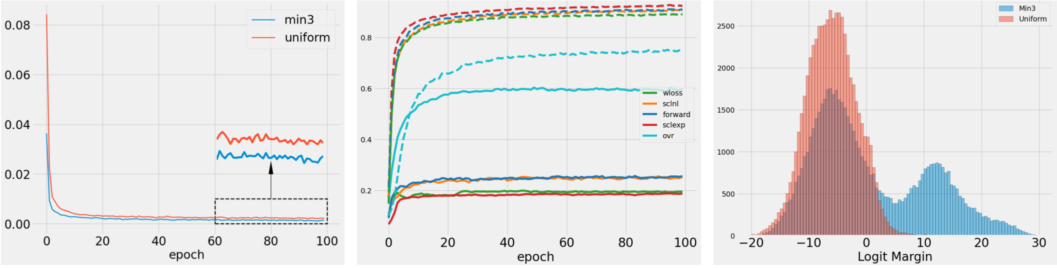
4.2 EMPIRICAL ANALYSIS OF EXISTING CLL METHODS

In this section, we first evaluate the performance of existing CLL methods under "Min3" setting. We select four baseline methods originally employed for solving uniform CLL: W_loss (Gao & Zhang, 2021), Forward (Yu et al., 2018), SCL_NL (Chou et al., 2020), and SCL_EXP (Chou et al., 2020). Table 1 presents the classification accuracy of these baseline methods on different benchmark datasets under the uniform and "Min3" settings, respectively. The results demonstrate a consistent decrease in accuracy for all the methods when switching from the uniform to "Min3" setting, confirming that most existing CLL methods are not well-suited for our instance-dependent setting.

We conducted experiments on the Kuzushiji-MNIST (KMNIST) dataset to further investigate the cause of the accuracy drop. When considering the existing CLL methods, most of them mathematically try to minimize the predictions of complementary label $p_{\boldsymbol{\theta}}(Y=\bar{y}|X=\boldsymbol{x})$ in one way or another. We calculate the average prediction of complementary label $p_{\boldsymbol{\theta}}(Y=\bar{y}|X=\boldsymbol{x})$ for SCL_NL over all instances in every epoch and Figure 1a presents SCL_NL effectively minimizes $p_{\boldsymbol{\theta}}(Y=\bar{y}|X=\boldsymbol{x})$ under "Min3" setting, which even lower than the value under the uniform setting. This phenomenon reveals the ineffectiveness of existing complementary losses: Although these complementary losses optimize their objectives, i.e., minimizing $p_{\boldsymbol{\theta}}(Y=\bar{y}|X=\boldsymbol{x})$, they still fail to make the ground-truth label prominent from the non-complementary labels.

To further explain this phenomenon, we first present the definition of *complementary label distribution* and introduce *share complementary label* hypothesis (Lin et al., 2023). Subsequently, we take an insight in why existing CLL methods demonstrate effectiveness under uniform setting.

Definition 2 (Complementary label distribution)**.** *The complementary label distribution refers to the distribution of complementary labels given the same ground-truth label, denoted as $p(\bar{Y}|Y=y)$. In addition, we use the entropy $H(\bar{Y}|Y)$ to characterize the dispersion of this distribution.*



(a) $p_\theta(Y = \bar{y}|X = \boldsymbol{x})$ for SCL_NL (b) Implicit Sharing Efficiency (c) Logit Margin Loss for SCL_NL

Figure 1: (a) The prediction of complementary label $p_\theta(Y = \bar{y}|X = \boldsymbol{x})$; (b) Implicit sharing efficiency for various methods. Dotted line represents the uniform setting and the solid line represents the "Min3" setting; (c) Logit margin loss for SCL_NL under both uniform and "Min3" settings.

Remark 1. *Under the uniform setting, the complementary label distribution is assumed to be a uniform distribution over $K-1$ non-ground-truth labels. However, under the "Mink" setting, the complementary label distribution becomes relatively sparser, i.e., $H(\bar{Y}|Y)$ becomes smaller. As k increases, the complementary label distribution approaches a uniform distribution, thus the "Min9" setting is equivalent to the uniform setting. More details can be found in Appendix F.*

Hypothesis 1 (Share complementary label)**.** *Instances that exhibit proximity in the feature space tend to share their complementary labels with each other. Under the uniform assumption, the complementary label distribution for each class is close to a uniform one, and each instance has access to all non-ground-truth labels as complementary labels shared by the neighboring instances.*

Next, we elucidate the significance of the ability to share complementary labels for CLL. We consider the logit margin loss (LML) ℓ_{LM} in supervised classification as (Kanai et al., 2023):

$$\ell_{\text{LM}} = \max_{j \neq y} g_j(\boldsymbol{x}) - g_y(\boldsymbol{x}). \tag{6}$$

When $\ell_{\text{LM}} < 0$, the model correctly classifies the instance, whereas when $\ell_{\text{LM}} > 0$, the model misclassifies the instance. Under the uniform setting, due to the share complementary label mechanism, existing CLL methods can potentially decrease the logit of all non-ground-truth label, i.e., $g_j(\boldsymbol{x}), \forall j \neq y$, thereby making ℓ_{LM} small enough to correctly identify ground-truth label. However, the sparsity of the complementary label distribution under the instance-dependent setting diminishes the ability of existing CLL methods to share complementary labels, which makes the classifier hardly distinguishes the ground-truth label from the non-complementary labels. In order to validate our conjecture, we borrow the implicit sharing efficiency (ISE) from Lin et al. (2023):

$$\text{ISE} = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{K-1}{K-2}\sum_{j \notin \{\bar{y}_i, y_i\}} p_{\boldsymbol{\theta}}(Y = j|X = \boldsymbol{x}_i). \tag{7}$$

This metric quantifies the decrease in the model's confidence regarding the unseen CLs. If implicit sharing helps identify all complementary labels, then $p_{\boldsymbol{\theta}}(Y = j|X = \boldsymbol{x}_i)$ will become zero, resulting in an ISE of one. In the absence of implicit sharing among instances, $p_{\boldsymbol{\theta}}(Y = j|X = \boldsymbol{x}_i)$ will be average $\frac{1}{K-1}$ and the ISE becomes zero. We calculate ISE during the training process for various CLL methods under both the "Min3" setting and the uniform setting. As shown in Figure 1b, different colors denote different CLL methods, with the dotted line representing ISE under the uniform setting and the solid line representing the "Min3" setting, respectively. Figure 1b reveals a substantial decrease in ISE for the existing CLL methods under the "Min3" setting compared to the uniform one. Furthermore, we present histograms of ℓ_{LM} for SCL_NL under both uniform and "Min3" settings in Figure 1c, which clearly illustrates that SCL_NL exhibits smaller ℓ_{LM} values under the uniform setting compared to the "Min3" setting. This observation solidifies the notion that greater ability of sharing complementary labels leads to smaller ℓ_{LM}, and consequently, improves the identification of potential ground-truth labels. From the empirical results, it is evident that existing CLL methods experience a diminished capacity to share complementary labels under the instance-dependent settings, which results in smaller ℓ_{LM} and hinders their ability to effectively identify the ground-truth label.

5 METHODOLOGY

5.1 COMPLEMENTARY LOGIT MARGIN LOSS

Our empirical analysis highlights the connection between large ℓ_{LM} and label disambiguation. However, in CLL problem, we are given only *single* complementary label for each instance and ℓ_{LM} is underdetermined. In order to enhance ISE and identify the ground-truth label, we instead consider a novel complementary logit margin loss (CLML) $\bar{\ell}_{\mathrm{LM}}$ as:

$$\bar{\ell}_{\mathrm{LM}} = g_{\bar{y}}(\boldsymbol{x}) - \max_{j \neq \bar{y}} g_j(\boldsymbol{x}), \tag{8}$$

and define complementary logit margin (CLM) as $\max_{j\neq\bar{y}} g_j(\boldsymbol{x}) - g_{\bar{y}}(\boldsymbol{x})$. If $\bar{\ell}_{\mathrm{LM}} > 0$, the model misclassifies instance, while $\bar{\ell}_{\mathrm{LM}} < 0$ does not mean correct classification. However, $\bar{\ell}_{\mathrm{LM}}$ of correct classification should be negative and small, which suggests to penalize small CLM.

By decreasing CLML, we should minimizes $g_{\bar{y}}(\boldsymbol{x})$, which aligns with the existing CLL methods. However, CLML also takes into account $\max_{j\neq\bar{y}} g_j(\boldsymbol{x})$ and aims to maximize it. According to Assumption 1, IDCL tends to be the least relevant label to the instance, while the ground-truth label is typically the most relevant one and has the largest logit. This allows instances that exhibit proximity in the feature space to obtain not only the complementary label from neighbors, but also their ground-truth labels, which enhance the ISE and the ability to identify the ground-truth label.

5.2 COVR LOSS FOR IDCLL

For the purpose to increase CLML, it is essential to design a loss function that penalizes the small complementary logit margins. The CLML is a straightforward choice as the loss function. However, it exclusively considers the pair of the largest logit $g_j(\boldsymbol{x})(j \neq \bar{y})$ and the logit for the complementary label $g_{\bar{y}}(\boldsymbol{x})$, leading to a significant loss of information. In order to take all the logits into consideration, we utilize the Complementary One-Versus-Rest Loss (COVR) (Ishida et al., 2017):

$$\bar{\ell}_{\mathrm{COVR}}(y, \boldsymbol{g}(\boldsymbol{x})) = \ell(-g_{\bar{y}}(\boldsymbol{x})) + \frac{1}{K-1}\sum_{j\neq\bar{y}} \ell(g_j(\boldsymbol{x})). \tag{9}$$

In particular, we set $\ell(z) = \log(1+e^{-z})$ and the COVR takes the form of:

$$\bar{\ell}_{\mathrm{COVR}}(y, \boldsymbol{g}(\boldsymbol{x})) = \log(1+e^{g_{\bar{y}}(\boldsymbol{x})}) + \frac{1}{K-1}\sum_{j\neq\bar{y}} \log(1+e^{-g_j(\boldsymbol{x})}). \tag{10}$$

We should point out that this novel form of COVR has not been proposed in previous work. As illustrated in Figure 1b, the implicit sharing efficiency of COVR demonstrates a substantial enhancement in its capacity to share complementary labels, allowing to better identify the ground-truth label. However, it is worth noting that ISE of COVR diminishes under the uniform setting. This decline is attributed to the fact that complementary labels obtained in the uniform setting are not necessarily the least relevant labels. Consequently, using COVR to decrease CLML may lead to overfitting. We also derived an estimation error bound for COVR to justify its convergence: To what extent, the empirical complementary risk minimization leads to the expected complementary risk minimization. The details of the estimation error bound can be found in Appendix C.

5.3 THEORETICAL ANALYSIS FOR COVR

To illustrate the effectiveness of COVR in decreasing CLML, we conduct a theoretical comparison between COVR and SCL_NL. First, we observe that COVR is the upper bound of SCL_NL:

Theorem 1. *If we consider COVR and SCL_NL, we can have*

$$0 \leq \bar{\ell}_{\mathrm{SCL\_NL}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x})) \leq \bar{\ell}_{\mathrm{COVR}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x})), \forall (\boldsymbol{x}, \bar{y}) \in \{(\boldsymbol{x}_i, \bar{y})\}_{i=1}^{n}. \tag{11}$$

When $g_{\bar{y}}(\boldsymbol{x}) \to -\infty$ *and* $g_j(\boldsymbol{x}) \to +\infty, \forall j \neq \bar{y}$, *we have* $\bar{\ell}_{\mathrm{COVR}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x})) \to 0$, *and then* $\bar{\ell}_{\mathrm{SCL\_NL}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x})) \to 0$.

For IDCLL, $\bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$ is consistently larger or at least equal to the baseline $\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$. Furthermore, as $|\bar{\ell}_{LM}|$ increases towards infinity, both $\bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$ and $\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$ converge towards zero asymptotically. Consequently, we anticipate that COVR penalize the large CLML more strongly than SCL_NL.

In addition, we further explore the impact of COVR on CLML to behavior of CLM by the problem:

$$\min_{g} \bar{\ell}(\bar{y}, \boldsymbol{g}), \tag{12}$$

where $\bar{\ell}$ is set to $\bar{\ell}_{\text{COVR}}$ or $\bar{\ell}_{\text{SCL\_NL}}$, and $\boldsymbol{g} \in \mathbb{R}^K$ is the logit vector for a instance $\boldsymbol{x}$. To dissect the training dynamics of Eq. (12), we use the following assumption (Kanai et al., 2023):

Assumption 2. *The logit vector* $\mathbf{g}$ *follows the following gradient flow to solve Eq.(12):*

$$\frac{d\boldsymbol{g}}{dt} = -\nabla_{\boldsymbol{g}} \bar{\ell}(\bar{y}, \boldsymbol{g}), \tag{13}$$

where t *is the time step of training. We assume that* $\boldsymbol{g}$ *is initialized to zeros* $\boldsymbol{g} = \mathbf{0}$ *at* $t = 0$ *.*

Eq. (13) serves as a continuous approximation of the gradient descent equation $\boldsymbol{g}^{\tau+1} = \boldsymbol{g}^{\tau} - \eta \nabla_{\boldsymbol{g}} \bar{\ell}$ and aligns with it as the learning rate η approaches zero. It is a commonly used method to analyze the training dynamics (Kunin et al., 2021; Elkabetz & Cohen, 2021).

Under Assumption 2, we obtain the following two lemmas for the logits in the training of Eq.(12):

Lemma 1. *If we use* $\bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g})$ *in Eq.(12), the* j*-th logit* g_j *at time* t *is*

$$g_j(t) = \begin{cases} -t - 1 + W(e^{t+1}) & j \neq \bar{y}, \\ \frac{1}{K-1}t + 1 - W(e^{\frac{1}{K-1}t+1}) & j = \bar{y}, \end{cases} \tag{14}$$

where W *is Lambert* W *function, which is a function satisfying* $x = W(xe^x)$ *(Corless et al., 1996).*

Lemma 2. *If we use* $\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g})$ *in Eq. (12), the* j*-th logit* g_j *at time* t *is*

$$g_j(t) = \begin{cases} \frac{1}{K-1}t + \frac{K-1}{K} - \frac{1}{K}W[(K-1)e^{\frac{K}{K-1}t+K-1}] & j \neq \bar{y}, \\ -t - \frac{(K-1)^2}{K} + \frac{K-1}{K}W[(K-1)e^{\frac{K}{K-1}t+K-1}] & j = \bar{y}. \end{cases} \tag{15}$$

These two lemmas provide insights into the trajectories of the logit vectors during the minimization of $\bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$ and $\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$, respectively. Both methods decrease the complementary label logit $g_{\bar{y}}$ and increase the logits for non-complementary labels, but they exhibit different rates of update. From the above lemmas, we derive the trajectory of the complementary logit margins:

Theorem 2. *Complementary logit margin loss for the logit vector* $\boldsymbol{g}^{\text{COVR}}$ *in the minimization of* $\bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$ *and logit vector* $\boldsymbol{g}^{\text{SCL\_NL}}$ *in the minimization of* $\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x}))$ *at time* t *are*

$$\bar{\ell}_{\text{LM}}(\boldsymbol{g}^{\text{COVR}}(t)) = -\frac{K}{K-1}t - 2 + W(e^{t+1}) + W(e^{\frac{1}{K-1}t+1}), \tag{16}$$

$$\bar{\ell}_{\text{LM}}(\boldsymbol{g}^{\text{SCL\_NL}}(t)) = -\frac{K}{K-1}t - K + 1 + W[(K-1)e^{\frac{K}{K-1}t+K-1}]. \tag{17}$$

For large t*, they can be approximated by*

$$\bar{\ell}_{\text{LM}}(\boldsymbol{g}^{\text{COVR}}(t)) \approx -\log[(\frac{1}{K-1}t + 1)(t+1)], \tag{18}$$

$$\bar{\ell}_{\text{LM}}(\boldsymbol{g}^{\text{SCL\_NL}}(t)) \approx -\log[\frac{K}{(K-1)^2}t + 1 + \log(K-1)^{\frac{1}{K-1}}]. \tag{19}$$

Then, we have $\lim_{t\to\infty} \frac{\bar{\ell}_{\text{LM}}(\boldsymbol{g}^{\text{COVR}}(t))}{\bar{\ell}_{\text{LM}}(\boldsymbol{g}^{\text{SCL\_NL}}(t))} = 2$ *for any fixed* K*.*

Theorem 2 elucidates the disparity in the trajectories of CLML between COVR and SCL_NL under Assumption 2 and shows that SCL_NL does not decrease the CLML as small as COVR for sufficiently large t. As a result, COVR is proved to be more effective and efficient at decreasing the large CLML.

Table 2: Classification accuracy (mean±std) of various methods on benchmark datasets under "Min3" setting. The symbol • indicates the statistically significant superiority (at a significance level of 0.05).

	MNIST	Kuzushiji-MNIST	Fashion-MNIST	CIFAR-10	SVHN
PRODEN(PLL)	63.84 ± 1.12 •	46.37 ± 0.74 •	40.74 ± 3.77 •	40.72 ± 1.69 •	33.05 ± 1.06 •
W_Loss	46.19 ± 4.56 •	46.73 ± 3.94 •	42.61 ± 6.14 •	45.57 ± 0.63 •	38.46 ± 3.13 •
Forward	51.82 ± 6.81 •	53.69 ± 4.50 •	39.54 ± 0.13 •	45.12 ± 2.03 •	31.43 ± 2.86 •
SCL_NL	55.56 ± 6.91 •	51.33 ± 1.76 •	39.53 ± 0.08 •	38.40 ± 0.34 •	32.97 ± 2.18 •
SCL_EXP	33.78 ± 4.03 •	42.78 ± 4.20 •	37.06 ± 1.60 •	30.78 ± 0.94 •	20.59 ± 0.06 •
NN	60.55 ± 4.88 •	37.67 ± 2.24 •	39.15 ± 2.22 •	34.03 ± 1.51 •	33.67 ± 2.84 •
PC	83.94 ± 2.20 •	64.99 ± 1.44 •	45.10 ± 1.15	37.51 ± 1.21 •	72.53 ± 1.67
COVR (ours)	**87.52 ± 0.02**	**67.31 ± 1.53**	**47.41 ± 0.60**	**63.21 ± 0.18**	**74.81 ± 1.19**

Table 3: Classification accuracy (mean±std) of different CLL methods on CIFAR10 under different instance-dependent settings. The symbol • indicates the statistically significant superiority (at a significance level of 0.05).

Method	CIFAR10				CLCIFAR10
	CIFAR10_Min1	CIFAR10_Min3	CIFAR10_Min5	CIFAR10_Min7	
Forward	30.48 ± 3.47 •	45.12 ± 2.03 •	49.18 ± 0.27 •	65.54 ± 1.97	36.83 ± 1.17 •
SCL_NL	30.32 ± 3.34 •	38.40 ± 0.34 •	48.71 ± 0.45 •	**66.01 ± 1.92**	37.81 ± 2.21
SCL_EXP	25.97 ± 0.90 •	30.78 ± 0.94 •	42.74 ± 3.31 •	55.05 ± 1.32 •	36.96 ± 0.18 •
PC	63.88 ± 0.54	37.51 ± 1.21 •	49.76 ± 1.02 •	44.20 ± 1.62 •	35.88 ± 0.98 •
COVR (ours)	**66.22 ± 1.34**	**63.21 ± 0.18**	**62.23 ± 0.59**	64.34 ± 1.08	**38.29 ± 0.21**

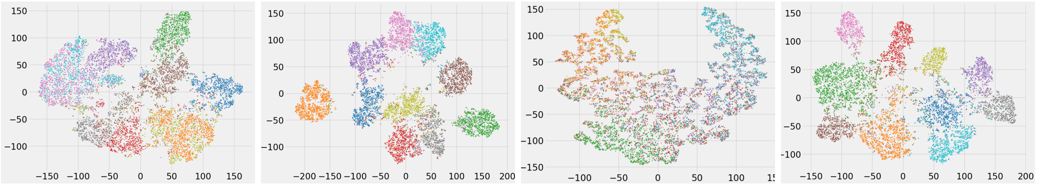
6 EXPERIMENTS

6.1 SETUPS

Datasets and Baselines In our experiments, we adopt five widely used benchmark datasets including MNIST , Fashion-MNIST (FMNIST) , Kuzushiji-MNIST (KMNIST) , CIFAR-10, and SVHN . For MNIST, FMNIST and KMNIST, we train a convolutional neural network with two convolutional layers and two fully-connected layers for 100 epochs with batch size of 256. Adam optimizer (Kingma & Ba, 2014) is used with the learning rate$=1\times10^{-3}$, weight decay$=1\times10^{-4}$. For CIFAR-10 and SVHN, we train ResNet-18 (He et al., 2016) for 200 epochs with batch size of 128. SGD optimizer is used with the learning rate$=1\times10^{-1}$, weight decay$=5\times10^{-4}$. We choose six existing CLL methods to date, including W_Loss (Gao & Zhang, 2021), Forward (Yu et al., 2018), SCL_NL (Chou et al., 2020), SCL_EXP (Chou et al., 2020), NN (Ishida et al., 2019), PC (Ishida et al., 2017), and a partial label learning method PRODEN (Lv et al., 2020). To generate IDCLs for each dataset, we use MLP and ResNet-18 trained with ground-truth labels as the pre-trained models and follow the generation processes of IDCLs defined in Section 4. More experimental details can be found in Appendix D.

6.2 EXPERIMENTAL RESULTS

Classification accuracy on benchmark datasets Table 2 reports the classification accuracy of each CLL method on benchmark datasets under "Min3" setting. The best results are highlighted in bold and the symbol • indicates that our method is statistically superior to the comparing methods on each dataset (pairwise t-test at a significance level of 0.05). From Table 2, COVR demonstrates the best performance and significantly outperforms other existing CLL methods under "Min3" setting. Furthermore, we find that COVR exhibits a lower standard deviation compared to other baseline methods, indicating its improved consistency and stability under the instance-dependent setting.



(a) KMNIST SCL_NL (b) KMNIST COVR (c) SVHN SCL_NL (d) SVHN COVR

Figure 2: The t-SNE visualization of the image representation on KMNIST (left panel) and SVHN (right panel) under “Min3” setting. Different colors represent the corresponding classes.

The effectiveness of COVR under different instance-dependent settings To validate the effectiveness of COVR, we conducted experiments using the CIFAR10 with various instance-dependent settings, including “Min1”, “Min5”, and “Min7”. A human-annotated CLL dataset CLCIFAR10 (Wang et al., 2023) (more details in Appendix H) is also considered. Table 3 presents the accuracy of different CLL methods across these instance-dependent settings on the CIFAR10 dataset. We observe that COVR consistently performs well under various instance-dependent settings, and as k increases (becoming more uniform), COVR does not exhibit significant performance degradation. As Figure 1b illustrates, COVR exhibits a more stable performance across both uniform and instance-dependent settings. Moreover, our proposed approach demonstrates notable improvement compared to other methods when evaluated on the human-annotated CLCIFAR10 dataset. These results emphasize the effectiveness of our approach on the CLCIFAR10 dataset, further validating its superiority in practice.

COVR learns more distinguishable representations In Figure 2, we present a visualization of the image representations generated by the feature encoder using t-SNE (Torralba et al., 2008). Different colors represent distinct ground-truth labels. This analysis was conducted using the KMNIST and SVHN datasets under the “Min3” setting, respectively. We compare the t-SNE embeddings of two approaches: (a) The best-performing baseline method, i.e., SCL_NL, and (b) Our method, i.e., COVR. It is shown that on both datasets, representation learned by SCL_NL is indistinguishable, especially for the more complex dataset SVHN. Conversely, COVR consistently generates well-separated clusters and learns distinct representations on both datasets. This observation confirms that COVR allows the model to learn the feature that is more relevant to the instance (Lv et al., 2020; Wang et al., 2023).

COVR effectively increases the logit margin In Figure 3, we display histograms of ℓ_{LM} for both SCL_NL and COVR under the “Min3” setting on SVHN. Instances with $\ell_{\mathrm{LM}} > 0$ are misclassified. Figure 3 reveals that by reducing $\bar{\ell}_{\mathrm{LM}}$, COVR enhances the capacity to share complementary labels and subsequently decreases ℓ_{LM}. This improvement allows COVR to better identify potential ground-truth labels compared to SCL_NL, which validates our explanation.

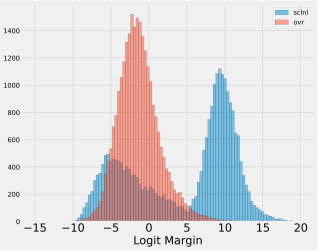


Figure 3: ℓ_{LM} for SCL_NL and COVR under “Min3”.

7 CONCLUSION

In this paper, we for the first time introduce the instance-dependent assumption to complementary label learning and present the generation process for IDCL according to our assumption. Subsequently, we discover that directly applying existing CLL methods to IDCLL results in a poor performance and further empirically identify the underlying reason as the diminishing capacity to share complementary labels. To address this problem, we introduce CLML to enhance the share of complementary labels. We additionally propose COVR as the surrogate loss, and thoroughly investigate its advantages in theoretical and empirical under our instance-dependent setting.

Nevertheless, there are three limitations in our study. First, although we explore the share complementary label mechanism, COVR cannot explicitly utilize this mechanism, especially under uniform setting. Second, the generation process of IDCL presented in our work may not be the most realistic one. Third, the estimation of instance-dependent transition matrix may be a better solution for IDCLL. To overcome these limitations will shed light on the promising improvement of our current work.

REFERENCES

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. International Conference on Machine Learning, pp. 825–836, 2021.

De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16630–16639, 2022.

Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. International Conference on Machine Learning, pp. 1789–1799, 2020.

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels. International Conference on Machine Learning, pp. 1929–1938, 2020.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambert w function. Advances in Computational Mathematics, 5:329–359, 1996.

Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks. Advances in Neural Information Processing Systems, 34:4947–4960, 2021.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 213–220, 2008.

Yi Gao and Min-Ling Zhang. Discriminative complementary-label learning with weighted loss. International Conference on Machine Learning, pp. 3587–3597, 2021.

Arpit Garg, Cuong Nguyen, Rafael Felix, Thanh-Toan Do, and Gustavo Carneiro. Instance-dependent noisy label learning via graphical modelling. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2288–2298, 2023.

Lan-Zhe Guo, Yi-Ge Zhang, Zhi-Fan Wu, Jie-Jing Shao, and Yu-Feng Li. Robust semi-supervised learning when not all classes have labels. Advances in Neural Information Processing Systems, 35: 3305–3317, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

Abdolhossein Hoorfar and Mehdi Hassani. Approximation of the lambert w function and hyperpower function. Research Report Collection, 10(2), 2007.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. Advances in Neural Information Processing Systems, 30, 2017.

Takashi Ishida, Gang Niu, Aditya Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. International Conference on Machine Learning, pp. 2971–2980, 2019.

Sekitoshi Kanai, Shin'ya Yamaguchi, Masanori Yamada, Hiroshi Takahashi, Kentaro Ohno, and Yasutoshi Ida. One-vs-the-rest loss to focus on important samples in adversarial training. pp. 15669–15695, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Yiwen Kou, Zixiang Chen, Yuan Cao, and Quanquan Gu. How does semi-supervised learning with pseudo-labelers work? A case study. International Conference on Learning Representations, 2023.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. Advances in Neural Information Processing Systems, 2022.

Wei-I Lin, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Enhancing label sharing efficiency in complementary-label learning with label augmentation. arXiv preprint arXiv:2305.08344, 2023.

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Consistent complementary-label learning via order-preserving losses. International Conference on Artificial Intelligence and Statistics, pp. 8734–8748, 2023.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. International Conference on Machine Learning, pp. 6500–6510, 2020.

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

Aditya Krishna Menon, Brendan Van Rooyen, and Nagarajan Natarajan. Learning from binary labels with instance-dependent noise. Machine Learning, 107:1561–1595, 2018.

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

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. Advances in Neural Information Processing Systems, 26, 2013.

Congyu Qiao, Ning Xu, and Xin Geng. Decompositional generation process for instance-dependent partial label learning. International Conference on Learning Representations, 2023.

Yingjie Tian, Xiaotong Yu, and Saiji Fu. Partial label learning: Taxonomy, analysis and outlook. Neural Networks, 2023.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1958–1970, 2008.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. PiCO: Contrastive label disambiguation for partial label learning. International Conference on Learning Representations, 2022.

Hsiu-Hsuan Wang, Wei-I Lin, and Hsuan-Tien Lin. Clcifar: Cifar-derived benchmark datasets with human annotated complementary labels. arXiv preprint arXiv:2305.08295, 2023.

Jonathan Wilton, Abigail Koay, Ryan Ko, Miao Xu, and Nan Ye. Positive-unlabeled learning using random forests via recursive greedy risk minimization. Advances in Neural Information Processing Systems, 35:24060–24071, 2022.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742, 2018.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. Advances in Neural Information Processing Systems, 33:7597–7610, 2020.

Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. Instance-dependent partial label learning. Advances in Neural Information Processing Systems, 34:27119–27130, 2021.

Ning Xu, Jiaqi Lv, Biao Liu, Congyu Qiao, and Xin Geng. Progressive purification for instance-dependent partial label learning. arXiv preprint arXiv:2206.00830, 2022.

Yanbing Xue and Milos Hauskrecht. Active learning of multi-class classification models from ordered class sets. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):5589–5596, 2019.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent bayes-label transition matrix using a deep neural network. International Conference on Machine Learning, pp. 25302–25312, 2022.

Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. Advances in Neural Information Processing Systems, 34:4409–4420, 2021.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. Proceedings of the European Conference on Computer Vision, pp. 68–83, 2018.

APPENDIX

We first show the detailed complementary loss functions in Appendix A. Then, the proofs of Lemma 1, Lemma 2, Theorem 1 and Theorem 2 are given in Appendix B. Later, we derive the estimation error bound for COVR to verify its convergence and consistency in Appendix C. In Appendix D, there are more details of our experiments. Moreover, in order to confirm the consistency of our generation process of IDCLs under different pre-trained models, more experiments are shown in Appendix E. To further validate the effectiveness of our proposed method, in Appendix F, we conduct experiments under different "Min$k$" settings, respectively. In Appendix G, we compare the difference between "biased" and instance-dependent settings. Finally, we conduct experiments on a human-annotated CLL dataset CLCIFAR10 in Appendix H.

A THE DETAILED COMPLEMENTARY LOSS FUNCTIONS

In this section, we introduce the popular complementary loss functions in CLL, which are used as the baselines in our paper. Here, we mainly consider the single complementary label learning.

PC loss PC (Ishida et al., 2017) is an unbiased risk estimator (URE) for CLL under uniform setting, which takes the form of:

$$\bar{\ell}_{\mathrm{PC}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x})) = (K-1)\sum_{j\neq\bar{y}} \sigma(g_j(\boldsymbol{x}) - g_{\bar{y}}(\boldsymbol{x})) - M_1 + M_2, \tag{20}$$

where the constants $M_1 = K(K-1)/2n$ and $M_2 = (K-1)/n$ are used to derive the URE for CLL.

SCL_NL and SCL_EXP To mitigate the huge empirical gradient variance caused by URE, the surrogate complementary loss (SCL) framework was proposed (Chou et al., 2020), where SCL_NL takes the form of:

$$\bar{\ell}_{\mathrm{SCL\_NL}}(\bar{y}, \boldsymbol{f}(\boldsymbol{x})) = -\log(1 - p_{\boldsymbol{\theta}}(Y = \bar{y}|X = \boldsymbol{x})) = \log(1 - f_{\bar{y}}(\boldsymbol{x})), \tag{21}$$

and SCL_EXP takes the form of:

$$\bar{\ell}_{\mathrm{SCL\_EXP}}(\bar{y}, \boldsymbol{f}(\boldsymbol{x})) = \exp(p_{\boldsymbol{\theta}}(Y = \bar{y}|X = \boldsymbol{x})) = \exp(f_{\bar{y}}(\boldsymbol{x})). \tag{22}$$

Forward loss Forward is the complementary loss derived by forward correction technique (Yu et al., 2018). In detail, Q is the transition matrix with each element $Q_{ij} = p(\bar{Y} = j|Y = i)$. By using the transition matrix, Forward loss takes the form of:

$$\bar{\ell}_{\mathrm{Forward}}(\bar{y}, \boldsymbol{f}(\boldsymbol{x})) = -\log(\sum_{j\neq\bar{y}} Q_{j\bar{y}} \cdot p_{\boldsymbol{\theta}}(Y = j|X = \boldsymbol{x})) = -\log(\sum_{j\neq\bar{y}} Q_{j\bar{y}} \cdot f_j(\boldsymbol{x})). \tag{23}$$

In particular, under the uniform setting, the transition matrix Q takes 0 on diagonals and $\frac{1}{K-1}$ on non-diagonals.

W_Loss W_Loss is proposed by introducing weighted loss to the SCL_NL to maximize the predictive gap between potential ground-truth label and complementary label, which takes the form of:

$$\bar{\ell}_{\mathrm{W\_Loss}}(\bar{y}, \boldsymbol{f}(\boldsymbol{x})) = -(1 + \lambda w_{\bar{y}}) \log(1 - f_{\bar{y}}(\boldsymbol{x})), \tag{24}$$

where $w_j,\ j = 1, 2, \cdots, K$ are defined as:

$$w_j = \frac{1 - f_j(\boldsymbol{x})}{\sum_{k=1}^{K}(1 - f_k(\boldsymbol{x}))},\ j = 1, 2, \cdots, K. \tag{25}$$

B PROOFS

B.1 PROOF OF THEOREM 1

Proof. By the definition, we have

$$\begin{aligned}
\bar{\ell}_{\text{COVR}} &= \log(1+e^{g_{\bar{y}}}) + \frac{1}{K-1}\sum_{j\neq\bar{y}}\log(1+e^{-g_j}) && (26)\\
&= \frac{1}{K-1}\sum_{j\neq\bar{y}}[\log(1+e^{g_{\bar{y}}}) + \log(1+e^{-g_j})] && (27)\\
&= \frac{1}{K-1}\sum_{j\neq\bar{y}}\log[(1+e^{g_{\bar{y}}})(1+e^{-g_j})] && (28)\\
\bar{\ell}_{\text{SCL\_NL}} &= -\log(1-\frac{e^{g_{\bar{y}}}}{\sum_{k=1}^K e^{g_k}}) = -\log(\sum_{m\neq\bar{y}}\frac{e^{g_m}}{\sum_{k=1}^K e^{g_k}}). && (29)
\end{aligned}$$

We want to prove $\bar{\ell}_{\text{SCL\_NL}}(\bar{y},\boldsymbol{g}(\boldsymbol{x})) \leq \bar{\ell}_{\text{COVR}}(\bar{y},\boldsymbol{g}(\boldsymbol{x}))$, i.e., $-\log(\sum_{m\neq\bar{y}}\frac{e^{g_m}}{\sum_{k=1}^K e^{g_k}})^{K-1} \leq \sum_{j\neq\bar{y}}\log[(1+e^{g_{\bar{y}}})(1+e^{-g_j})]$. We consider one term of both the right hand and left hand of the equation, and we should prove

$$-\log(\sum_{m\neq\bar{y}}\frac{e^{g_m}}{\sum_{k=1}^K e^{g_k}}) \leq \log[(1+e^{g_{\bar{y}}})(1+e^{-g_j})]. \tag{30}$$

Since the $\log(x)$ is a monotonically increasing function, and we should prove

$$\begin{aligned}
\frac{\sum_{k=1}^K e^{g_k}}{\sum_{m\neq\bar{y}} e^{g_m}} &\leq (1+e^{g_{\bar{y}}})(1+e^{-g_j}) && (31)\\
1+\frac{e^{g_{\bar{y}}}}{\sum_{m\neq\bar{y}} e^{g_m}} &\leq 1+e^{g_{\bar{y}}}+e^{-g_j}+e^{g_{\bar{y}}-g_j} && (32)\\
\frac{e^{g_{\bar{y}}}}{\sum_{m\neq\bar{y}} e^{g_m}} &\leq e^{g_{\bar{y}}}+e^{-g_j}+e^{g_{\bar{y}}-g_j}. && (33)
\end{aligned}$$

We have

$$\begin{aligned}
&\frac{e^{g_{\bar{y}}}+e^{-g_j}+e^{g_{\bar{y}}-g_j}}{\frac{e^{g_{\bar{y}}}}{\sum_{m\neq\bar{y}} e^{g_m}}} && (34)\\
=&\sum_{m\neq\bar{y}} e^{g_m}(1+e^{-g_j-g_{\bar{y}}}+e^{-g_j}) && (35)\\
=&\sum_{m\neq\bar{y}} e^{g_m}+\sum_{m\neq\bar{y}} e^{g_m-g_j-g_{\bar{y}}}+\sum_{m\neq\bar{y}} e^{g_m-g_j} && (36)\\
\geq&\sum_{m\neq\bar{y}} e^{g_m-g_j} && (37)\\
=&\sum_{m\neq\bar{y},j} e^{g_m-g_j}+1 && (38)\\
\geq&1. && (39)
\end{aligned}$$

Then we complete the proof of $\bar{\ell}_{\text{SCL\_NL}}(\bar{y},\boldsymbol{g}(\boldsymbol{x})) \leq \bar{\ell}_{\text{COVR}}(\bar{y},\boldsymbol{g}(\boldsymbol{x}))$. When $g_{\bar{y}}(x) \to -\infty$ and $g_k(\boldsymbol{x}) \to +\infty$ for $k \neq \bar{y}$, we have

$$\begin{aligned}
&\lim_{g_{\bar{y}}\to+\infty, g_j\to-\infty} \bar{\ell}_{\text{COVR}}(\bar{y},\boldsymbol{g}(\boldsymbol{x})) && (40)\\
=&\lim_{g_{\bar{y}}\to+\infty, g_j\to-\infty} \log(1+e^{g_{\bar{y}}}) + \frac{1}{K-1}\sum_{j\neq\bar{y}}\log(1+e^{-g_j}) && (41)\\
=&0. && (42)
\end{aligned}$$

And $g_j(\boldsymbol{x}) - g_{\bar{y}}(\boldsymbol{x}) \to +\infty, \forall j \neq \bar{y}$, we have

$$\lim_{g_{\bar{y}} \to +\infty, g_j \to -\infty} \bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g}(\boldsymbol{x})) \tag{43}$$

$$= \lim_{g_{\bar{y}} \to +\infty, g_j \to -\infty} -\log(1 - \frac{e^{g_{\bar{y}}}}{\sum_{j=1}^{K} e^{g_j}}) \tag{44}$$

$$= \lim_{g_{\bar{y}} \to +\infty, g_j \to -\infty} -\log(1 - \frac{1}{1 + \sum_{j \neq \bar{y}} e^{g_j - g_{\bar{y}}}}) \tag{45}$$

$$=0. \tag{46}$$

Then we complete the proof of Theorem 1. □

B.2 PROOF OF LEMMA 1

Proof. From the Assumption 2 , we consider the following ordinary differential equation (ODE):

$$\frac{dg_j}{dt} = -\frac{\partial \bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g})}{\partial g_j}. \tag{47}$$

The initial condition is $\boldsymbol{g}(0) = \mathbf{0}$. For the complementary label $\bar{y}$, the gradient of COVR is given by

$$\frac{\partial \bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g})}{\partial g_{\bar{y}}} = \frac{e^{g_{\bar{y}}}}{1 + e^{g_{\bar{y}}}}. \tag{48}$$

Thus, the ODE becomes

$$\frac{dg_{\bar{y}}}{dt} = -\frac{e^{g_{\bar{y}}}}{1 + e^{g_{\bar{y}}}} \tag{49}$$

$$-\frac{1 + e^{g_{\bar{y}}}}{e^{g_{\bar{y}}}} dg_{\bar{y}} = dt \tag{50}$$

$$-g_{\bar{y}} + e^{-g_{\bar{y}}} = t + c \tag{51}$$

$$e^{-g_{\bar{y}} + e^{-g_{\bar{y}}}} = e^{t+c}, \tag{52}$$

where c is a constant, which is determined by the initial condition. From the Assumption 2, we have $\boldsymbol{g}(0) = \mathbf{0}$, and thus, $c = 1$. We apply the Lambert W function (Corless et al., 1996) for both sides and use $W(xe^x) = x$ and $\log W(x) = \log x - W(x)$ for $x > 0$ as

$$W(e^{-g_{\bar{y}} + e^{-g_{\bar{y}}}}) = W(e^{t+1}) \tag{53}$$

$$e^{-g_{\bar{y}}} = W(e^{t+1}) \tag{54}$$

$$-g_{\bar{y}} = \log W(e^{t+1}) = t + 1 - W(e^{t+1}) \tag{55}$$

$$g_{\bar{y}} = -t - 1 + W(e^{t+1}). \tag{56}$$

Next, we consider the logit of non-complementary label g_j for $j \neq \bar{y}$. Since the gradient of g_j is

$$\frac{\partial \bar{\ell}_{\text{COVR}}(\bar{y}, \boldsymbol{g})}{\partial g_j} = -\frac{1}{(K-1)(1 + e^{g_j})}, \tag{57}$$

we have

$$\frac{dg_j}{dt} = \frac{1}{(K-1)(1 + e^{g_j})} \tag{58}$$

$$(1 + e^{g_j}) dg_j = \frac{1}{K-1} dt. \tag{59}$$

We can use the same way to solve this ODE, and get

$$g_j = \frac{1}{K-1} t + 1 - W(e^{\frac{1}{K-1} t + 1}), \tag{60}$$

for $j \neq \bar{y}$. Then we complete the proof. □

B.3 PROOF OF LEMMA 2

Proof. By definition, we have $\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g}) = \log(\sum_{j=1}^{K} e^{g_j}) - \log(\sum_{k\neq\bar{y}} e^{g_k})$. Same as the proof of Lemma 1, we first solve the ODE for the logit of the complementary label $g_{\bar{y}}$. The gradient of $\bar{\ell}_{\text{SCL\_NL}}$ is

$$\frac{\partial\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g})}{\partial g_{\bar{y}}} = -\frac{e^{g_{\bar{y}}}}{\sum_{k=1}^{K} e^{g_k}} \tag{61}$$

$$\frac{\partial\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g})}{\partial g_j} = \frac{e^{g_j}}{\sum_{k=1}^{K} e^{g_k}} - \frac{e^{g_j}}{\sum_{k=1}^{K} e^{g_k}}. \tag{62}$$

Since $\boldsymbol{g} = \mathbf{0}$ at $t = 0$, we have $g_i = g_j$ for $\forall i, j \neq \bar{y}$. We can also find that $\sum_k \frac{\partial\bar{\ell}_{\text{SCL\_NL}}(\bar{y}, \boldsymbol{g})}{\partial g_k} = 0$ for $\forall t$. Thus, the logits satisfy the following equality:

$$g_{\bar{y}} = -(K-1)g_j, \tag{63}$$

for $j \neq \bar{y}$. Then we have

$$\frac{dg_{\bar{y}}}{dt} = -\frac{e^{g_{\bar{y}}}}{\sum_{k=1}^{K} e^{g_k}} \tag{64}$$

$$-\frac{\sum_{k=1}^{K} e^{g_k}}{e^{g_{\bar{y}}}} dg_{\bar{y}} = dt \tag{65}$$

$$-(1 + \sum_{j\neq\bar{y}} e^{g_j - g_{\bar{y}}}) dg_{\bar{y}} = dt \tag{66}$$

$$-1 - (K-1)e^{-\frac{K}{K-1}g_{\bar{y}}} dg_{\bar{y}} = dt \tag{67}$$

$$(\frac{1}{K-1} + e^{-\frac{K}{K-1}g_{\bar{y}}}) dg_{\bar{y}} = -\frac{1}{K-1}dt \tag{68}$$

$$\frac{1}{K-1}g_{\bar{y}} - \frac{K-1}{K}e^{-\frac{K}{K-1}g_{\bar{y}}} = -\frac{1}{K-1}t + c \tag{69}$$

$$\frac{1}{K-1}g_{\bar{y}} - \frac{K-1}{K}e^{-\frac{K}{K-1}g_{\bar{y}}} = -\frac{1}{K-1}t + c \tag{70}$$

$$-\frac{K}{K-1}g_{\bar{y}} + (K-1)e^{-\frac{K}{K-1}g_{\bar{y}}} = \frac{K}{K-1}t + c \tag{71}$$

$$(K-1)e^{-\frac{K}{K-1}g_{\bar{y}}} e^{(K-1)e^{-\frac{K}{K-1}g_{\bar{y}}}} = (K-1)e^{\frac{K}{K-1}t+c} \tag{72}$$

$$(K-1)e^{-\frac{K}{K-1}g_{\bar{y}}} = W[(K-1)e^{\frac{K}{K-1}t+c}] \tag{73}$$

$$-\frac{K}{K-1}g_{\bar{y}} = \log W[(K-1)e^{\frac{K}{K-1}t+c}] - \log(K-1) \tag{74}$$

$$-\frac{K}{K-1}g_{\bar{y}} = \frac{K}{K-1}t + c - W[(K-1)e^{\frac{K}{K-1}t+c}] \tag{75}$$

$$g_{\bar{y}} = -t - \frac{K-1}{K}c + \frac{K-1}{K}W[(K-1)e^{\frac{K}{K-1}t+c}], \tag{76}$$

and we have $\boldsymbol{g}(0) = \mathbf{0}$, and thus, $c = K-1$, and then

$$g_{\bar{y}} = -t - \frac{(K-1)^2}{K} + \frac{K-1}{K}W[(K-1)e^{\frac{K}{K-1}t+K-1}], \tag{77}$$

Then we use Eq.(63) and have

$$g_j = \frac{1}{K-1}t + \frac{K-1}{K} - \frac{1}{K}W[(K-1)e^{\frac{K}{K-1}t+K-1}], \tag{78}$$

for $j \neq \bar{y}$. Then we complete the proof. □

B.4 PROOF OF THEOREM 2

Proof. From Lemma 1 and Lemma 2, we can calculate the complementary logit margin loss for $\bar{\ell}_{\mathrm{COVR}}$ and $\bar{\ell}_{\mathrm{SCL\_NL}}$, respectively. We have

$$\bar{\ell}_{\mathrm{LM}}(\boldsymbol{g}^{\mathrm{COVR}}(t)) = -\frac{K}{K-1}t - 2 + W(e^{t+1}) + W(e^{\frac{1}{K-1}t+1}), \tag{79}$$

$$\bar{\ell}_{\mathrm{LM}}(\boldsymbol{g}^{\mathrm{SCL\_NL}}(t)) = -\frac{K}{K-1}t - K + 1 + W[(K-1)e^{\frac{K}{K-1}t+K-1}]. \tag{80}$$

Since $W(x) = \log(x) - \log(\log(x)) + O(1)$ for large x (Hoorfar & Hassani, 2007), we have

$$\bar{\ell}_{\mathrm{LM}}(\boldsymbol{g}^{\mathrm{COVR}}(t)) \approx -\log(t+1) - \log(\frac{1}{K-1}t+1) + O(1) \tag{81}$$

$$= -\log[(t+1)(\frac{1}{K-1}t+1)] + O(1), \tag{82}$$

$$\bar{\ell}_{\mathrm{LM}}(\boldsymbol{g}^{\mathrm{SCL\_NL}}(t)) \approx -\log[\log(K-1) + \frac{K}{K-1}t + K - 1] + \log(K-1) + O(1). \tag{83}$$

Then, we can have

$$\lim_{t\to\infty}\frac{\bar{\ell}_{\mathrm{LM}}(\boldsymbol{g}^{\mathrm{COVR}}(t))}{\bar{\ell}_{\mathrm{LM}}(\boldsymbol{g}^{\mathrm{SCL\_NL}}(t))} = \lim_{t\to\infty}\frac{-\log[(t+1)(\frac{1}{K-1}t+1)] + O(1)}{-\log[\frac{\log(K-1)}{K-1} + \frac{K}{(K-1)^2}t + 1] + O(1)} \tag{84}$$

$$= \lim_{t\to\infty}\frac{2\log t + \log(\frac{1}{K-1} + \frac{K}{(K-1)t} + \frac{1}{t^2}) + O(1)}{\log t + \log(\frac{\log(K-1)}{K-1}\frac{1}{t} + \frac{K}{(K-1)^2} + \frac{1}{t}) + O(1)} \tag{85}$$

$$= \lim_{t\to\infty}\frac{2 + \frac{1}{\log t}\log(\frac{1}{K-1} + \frac{K}{(K-1)t} + \frac{1}{t^2}) + \frac{O(1)}{\log t}}{1 + \frac{1}{\log t}\log(\frac{\log(K-1)}{K-1}\frac{1}{t} + \frac{K}{(K-1)^2} + \frac{1}{t}) + \frac{O(1)}{\log t}} \tag{86}$$

$$= 2. \tag{87}$$

Then we complete the proof. □

C ESTIMATION ERROR BOUND

In this section, we derive an upper bound for the estimation error of $\bar{\ell}_{\mathrm{COVR}}$. To this end, let $\mathcal{G} = \{g(\boldsymbol{x})\}$ be a function class for empirical risk minimization. For the simplicity, we denote $\bar{\ell} = \bar{\ell}_{\mathrm{COVR}}$ in this section. We define the complementary expected risk as $\bar{R}(h;\bar{\ell})$, and the complementary empirical risk as $\bar{R}_n(h;\bar{\ell})$. Let g^* be the expected risk minimizer and $\hat{g}^*$ be the empirical risk minimizer, i.e.,

$$g^* = \operatorname*{argmin}_{g\in\mathcal{G}} \bar{R}(h;\bar{\ell}) \quad \text{and} \quad \hat{g}^* = \operatorname*{argmin}_{g\in\mathcal{G}} \bar{R}_n(h;\bar{\ell}), \tag{88}$$

and

$$h^*(\boldsymbol{x}) = \operatorname*{argmax}_{y\in\mathcal{Y}} g_y^*(\boldsymbol{x}) \quad \text{and} \quad \hat{h}^*(\boldsymbol{x}) = \operatorname*{argmax}_{y\in\mathcal{Y}} \hat{g}_y^*(\boldsymbol{x}). \tag{89}$$

Let L_ℓ be any Lipschitz constant of ℓ and $M = \sup_{\boldsymbol{x}\in\bar{\mathcal{X}}, g\in\mathcal{G}} \ell(g(\boldsymbol{x}))$.

Lemma 3. *Let $\bar{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{G})$ be the Rademacher complexity of $\mathcal{G}$ for $\bar{\mathcal{X}}$ with data size n drawn from $\bar{\mathcal{D}}$. Then,*

$$\bar{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{G}) \leq KL_\ell\bar{\mathfrak{R}}_n(\mathcal{G}). \tag{90}$$

The proof of Lemma 3 can be found in Appendix C.1.Given the upper bound for $\bar{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{G})$, we can obtain Lemma 4 based on McDiarmid's inequality McDiarmid et al. (1989) and symmetrization Mohri et al. (2018), which defines the uniform deviation bound.

Lemma 4. *For any $\delta > 0$, with the probability at least $1-\delta$, we have*

$$\sup_{g\in\mathcal{G}} |\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell})| \leq 2\bar{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{G}) + M\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{91}$$

The proof is given in Appendix C.2.

Finally, based on Lemma 4, we can establish the estimation error bound for COVR as follows:

Theorem 3. *For any $\delta > 0$, with the probability at least $1-\delta$, we have*

$$\bar{R}(\hat{h}^*;\bar{\ell}) - \bar{R}(h^*;\bar{\ell}) \leq 4KL_{\ell}\bar{\mathfrak{R}}_n(\mathcal{G}) + 2M\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{92}$$

The proof can be found in Appendix C.3. For all parametric models with a bounded norm, $\bar{R}(\hat{h}^*;\bar{\ell}) \rightarrow \bar{R}(h^*;\bar{\ell})$ as $n \rightarrow \infty$. Theorem 3 shows that the proposed risk estimator exists an estimation error bound and the convergence rate is $\mathcal{O}(1/\sqrt{n})$.

C.1 PROOF OF LEMMA 3

Proof. By the definition of the Rademacher complexity, we have

$$\begin{aligned}
\bar{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{G}) &= \mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\{\ell(-g_{\bar{y}_i}(\boldsymbol{x}_i)) + \frac{1}{K-1}\sum_{j\neq\bar{y}_i}\ell(g_j(\boldsymbol{x}_i))\}\right] \qquad (93)\\
&= \mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\{\frac{1}{K-1}\sum_{j=1}^{K}\ell(g_j(\boldsymbol{x}_i)) + \frac{K-2}{K-1}\ell(-g_{\bar{y}_i}(\boldsymbol{x}_i))\}\right] \qquad (94)\\
&\leq \frac{1}{K-1}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{K}\ell(g_j(\boldsymbol{x}_i))\right] + \frac{K-2}{K-1}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(-g_{\bar{y}_i}(\boldsymbol{x}_i))\right]. \qquad (95)
\end{aligned}$$

The first term of Eq. (95) is

$$\begin{aligned}
&\frac{1}{K-1}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{K}\ell(g_j(\boldsymbol{x}_i))\right] \qquad (96)\\
\leq&\frac{1}{K-1}\sum_{j=1}^{K}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(g_j(\boldsymbol{x}_i))\right] \qquad (97)\\
=&\frac{K}{K-1}\bar{\mathfrak{R}}_n(\ell\circ\mathcal{G}). \qquad (98)
\end{aligned}$$

Let $\mathbb{1}$ be the indicator function and $\alpha_i = 2\mathbb{1}(j=\bar{y}_i)-1$. Then, the second term of Eq. (95) is

$$\begin{aligned}
&\frac{K-2}{K-1}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(-g_{\bar{y}_i}(\boldsymbol{x}_i))\right] \qquad (99)\\
=&\frac{K-2}{K-1}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{K}\ell(-g_j(\boldsymbol{x}_i))\mathbb{1}(j=\bar{y}_i)\right] \qquad (100)\\
=&\frac{K-2}{K-1}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{2n}\sum_{i=1}^{n}\sigma_i(\alpha_i+1)\sum_{j=1}^{K}\ell(-g_j(\boldsymbol{x}_i))\right] \qquad (101)\\
=&\frac{K-2}{K-1}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{K}\ell(-g_j(\boldsymbol{x}_i))\right] \qquad (102)\\
\leq&\frac{K-2}{K-1}\sum_{j=1}^{K}\mathbb{E}_{\bar{\mathcal{X}}}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(-g_j(\boldsymbol{x}_i))\right] \qquad (103)\\
=&\frac{K(K-2)}{K-1}\bar{\mathfrak{R}}_n(\ell\circ\mathcal{G}). \qquad (104)
\end{aligned}$$

Then we have

$$\bar{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{G}) = \frac{K}{K-1}\bar{\mathfrak{R}}_n(\ell \circ \mathcal{G}) + \frac{K(K-2)}{K-1}\bar{\mathfrak{R}}_n(\ell \circ \mathcal{G}) \tag{105}$$

$$= K\bar{\mathfrak{R}}_n(\ell \circ \mathcal{G}) \tag{106}$$

$$= KL_\ell \bar{\mathfrak{R}}_n(\mathcal{G}), \tag{107}$$

which completes the proof. □

C.2 PROOF OF LEMMA 4

Proof. We only consider the single direction $\sup_{g\in\mathcal{G}}(\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell}))$ with probability at least $1-\delta/2$, and the other direction is similar. Let M be the upper bound of $\bar{\ell}$ and a single $(x_i, \bar{y}_i)$ be replaced with $(x_i', \bar{y}_i')$, then the change of $\sup_{g\in\mathcal{G}}(\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell}))$ is no greater than $2M/n$. Applying McDiarmid's inequality McDiarmid et al. (1989) to the single-direction uniform deviation $\sup_{g\in\mathcal{G}}(\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell}))$, we have

$$\sup_{g\in\mathcal{G}}(\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell})) \leq \mathbb{E}\left[\sup_{g\in\mathcal{G}}(\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell}))\right] + M\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{108}$$

By symmetrization Mohri et al. (2018), we have

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}(\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell}))\right] \leq 2\bar{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{G}) = 2KL_\ell\mathfrak{R}_n(\mathcal{G}), \tag{109}$$

which completes the proof. □

C.3 PROOF OF THEOREM 3

Proof. Based on Lemma 4, the estimation error bound can be proven through

$$\bar{R}(\hat{h}^*;\bar{\ell}) - \bar{R}(h^*;\bar{\ell}) = (\bar{R}_n(\hat{h}^*;\bar{\ell}) - \bar{R}_n(h^*;\bar{\ell})) + (\bar{R}(\hat{h}^*;\bar{\ell}) - \bar{R}_n(\hat{h}^*;\bar{\ell})) + (\bar{R}_n(h^*;\bar{\ell}) - \bar{R}(h^*\bar{\ell})) \tag{110}$$

$$\leq 0 + 2\sup_{g\in\mathcal{G}}|\bar{R}(h;\bar{\ell}) - \bar{R}_n(h;\bar{\ell})|, \tag{111}$$

then we complete the proof. □

D MORE DETAILS OF THE EXPERIMENTS

D.1 EXPERIMENTAL ENVIRONMENT

The experiments are conducted on an NVIDIA GeForce RTX 3090 GPU with CUDA version 11.4. Then, we use Python 3.9.15 as the programming language and PyTorch 1.13.0 as the deep learning framework.

D.2 THE DATASETS

We adopt four widely-used benchmark datasets. MNIST is a benchmark dataset for hand-written digit recognition [1]. Fashion-MNIST (FMNIST) is a clothing image dataset [2]. Besides, Kuzushiji-MNIST (KMNIST) is a Japanese language dataset [3]. These datasets are consisting of 10 classes with 60,000 training samples and 10,000 testing samples. Each sample is a 28 × 28 gray-scale image. Their sizes are similar to MNIST. In addition, CIFAR-10 is an image dataset containing 10 different classes of objects, including 50,000 training samples and 10,000 testing samples. Each sample is a 32 × 32 × 3 RGB color image [4]. SVHN is the street view house numbers dataset, including 73257 samples for training, 26032 samples for testing and each sample is a 32 × 32 × 3 RGB color image [5].

---

[1] http://yann.lecun.com/exdb/mnist/

[2] https://www.worldlink.com.cn/en/osdir/fashion-mnist.html

[3] http://codh.rois.ac.jp/char-shape/book/

[4] http://www.cs.toronto.edu/ kriz/cifar.html

[5] http://ufldl.stanford.edu/housenumbers/

Table 4: Classification accuracy on CIFAR-10 under different pre-trained models.

Pre-trained model	**Accuracy(%)**
ResNet18	95.44
ResNet34	95.01
VGG16	93.97
GoogleNet	95.60

Table 5: The KL divergences between complementary label distributions generated by three pre-trained models and ResNet18.

KL	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
ResNet34	0.0007	0.0018	0.0002	0.0006	0.0007	0.0006	0.0006	0.0004	0.0005	0.0008
VGG16	0.0005	0.0008	0.0002	0.0014	0.0008	0.0006	0.0006	0.0015	0.0011	0.0008
GoogleNet	0.0007	0.0011	0.0009	0.0005	0.0006	0.0006	0.0003	0.0015	0.0009	0.0008

D.3 THE GENERATION OF IDCLL DATASETS

To generate IDCLL datasets, we use MLP as pre-trained model for MNIST, KMNIST, and FMNIST. For CIFAR-10 and SVHN, RseNet-18 is used. Then, we follow the definitions "Min$k$" in the main text, and conduct different CLs selections. Moreover, we conducted multiple experiments by setting different random seeds and compared the performance of each method on the same generated IDCLL datasets.

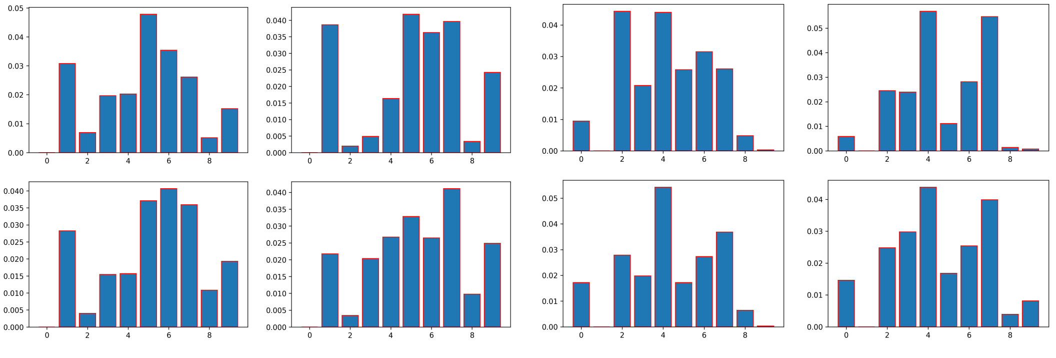
E IDCLS' CONSISTENCY ACROSS DIFFERENT PRE-TRAINED MODELS

Specifically, our generation process of IDCLs depends on the prediction probabilities of the pre-trained model, while different pre-trained models tend to provide different prediction probabilities. Therefore, in order to confirm the consistency of our generation process of IDCLs under different pre-trained models, we investigate the IDCLs produced by four different pre-trained models on CIFAR-10, including ResNet18 (He et al., 2016), ResNet34 (He et al., 2016), VGG16 (Simonyan & Zisserman, 2014), and GoogleNet (Szegedy et al., 2015).

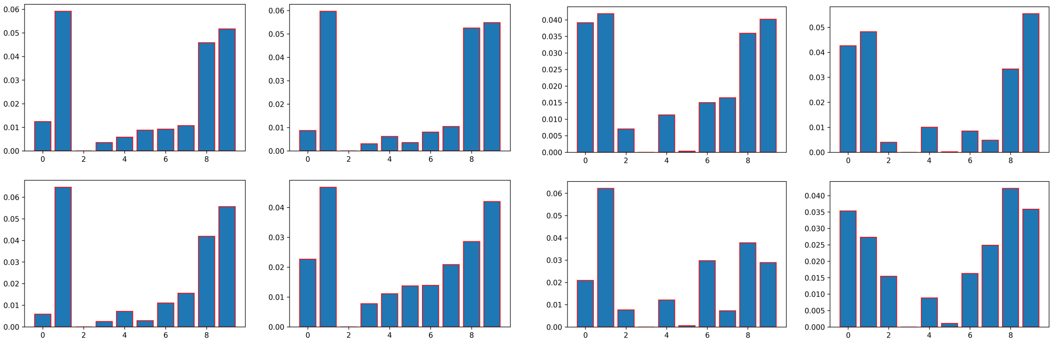
Table 4 shows the classification accuracy of the pre-trained models we used to generate IDCLs for CIFAR-10. All four pre-trained models achieve high classification accuracy. Figure 4 displays the complementary label distribution with ground-truth labels of "0","1","2", and "3" under "Min3" setting. In detail, each subfigure represents the result under a pre-trained model, i.e., ResNet18 (upper left) , ResNet34 (upper right) , VGG16 (lower left) , and GoogleNet (bottom right). We observe that IDCLs generated by different pre-trained models exhibit a similar complementary label distribution under "Min3" setting. This observation suggests the consistency in the IDCLs across different pre-trained models, despite the inherent randomness in "Min3". To further quantify this consistency, we calculate the KL divergence of the complementary label distribution generated by ResNet34, VGG16, and GoogleNet to the distribution generated by ResNet18 in Table 5. Remarkably, we find that all the KL divergences are extremely small, indicating that the complementary label distributions generated by the four pre-trained models closely resemble to each other.

F EXPERIMENTS UNDER "MINk" SETTINGS

To validate the performance of COVR loss, we conduct experiments under different "Min$k$" settings. Table 6 presents the classification accuracy of each approach on benchmark datasets under "Min$k$" settings. Several key observations can be made from the table. First, as k decreases, the accuracy of



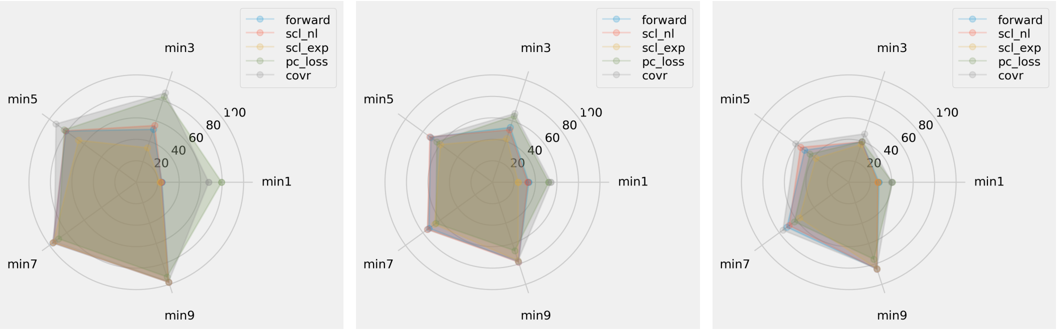
(a) ground truth: “0” (b) ground truth: “1”



(c) ground truth: “2” (d) ground truth: “3”

Figure 4: Complementary label distributions when the true labels are “0”, “1”, “2”, and “3” under different pre-trained models. In each subfigure, the four distributions are similar to each other.

all methods gradually decreases. However, the PC loss and our COVR loss still maintain a relatively higher accuracy. Second, COVR loss continues to achieve high accuracy even under the uniform setting, surpassing the performance of PC loss, which suggests that our approach enjoys a better balance between the uniform and instance-dependent situations. Third, Figure 5a, 5b, and 5c illustrate the classification accuracy under different “Min$k$” settings for three benchmark dataset. It is evident that COVR loss outperforms other methods overall, which further reinforces the superior performance of COVR loss compared to other approaches.



(a) MNIST (b) KMNIST (c) FMNIST

Figure 5: (a) (b) (c) Radar charts of the classification accuracy of different methods on three benchmark datasets under various “Min$k$” settings.

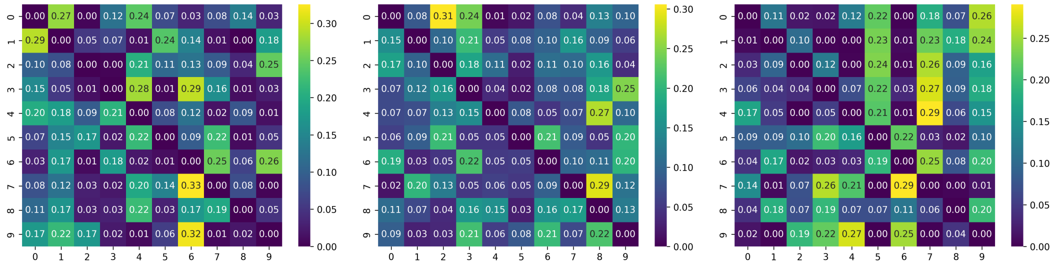
Table 6: Classification accuracy (mean±std) of each approach on benchmark datasets under "Min$k$" settings, respectively.

Dataset	Method	Min1	Min3	Min5	Min7	Min9
MNIST	**Forward**	23.96 ± 0.27	51.82 ± 6.81	81.07 ± 0.97	**96.19 ± 0.34**	98.00 ± 0.07
	SCL_NL	23.55 ± 1.18	55.56 ± 6.91	81.45 ± 0.88	96.06 ± 0.37	**98.06 ± 0.11**
	SCL_EXP	22.07 ± 2.63	33.78 ± 4.03	66.59 ± 3.76	95.56 ± 0.13	97.73 ± 0.15
	PC	**79.51 ± 1.96**	83.94 ± 2.20	83.02 ± 1.73	89.60 ± 0.78	92.72 ± 0.05
	COVR (ours)	67.51 ± 2.53	**87.52 ± 0.02**	**92.47 ± 0.11**	94.29 ± 0.47	97.49 ± 0.21
Kuzushiji-MNIST	**Forward**	33.83 ± 2.30	53.69 ± 4.51	71.29 ± 0.96	73.03 ± 2.36	78.05 ± 0.74
	SCL_NL	32.81 ± 3.24	51.33 ± 1.76	**72.06 ± 0.31**	**74.70 ± 3.30**	**78.21 ± 0.35**
	SCL_EXP	24.06 ± 0.52	42.78 ± 4.20	59.76 ± 4.07	66.70 ± 3.04	76.86 ± 0.36
	PC	52.37 ± 0.94	64.99 ± 1.44	63.97 ± 0.22	65.10 ± 1.01	67.07 ± 0.97
	COVR (ours)	**54.40 ± 0.32**	**67.31 ± 1.53**	70.22 ± 1.12	74.91 ± 1.01	76.79 ± 0.79
Fashion-MNIST	**Forward**	28.17 ± 0.80	39.54 ± 0.16	50.80 ± 1.17	71.22 ± 3.66	85.16 ± 0.32
	SCL_NL	27.07 ± 0.33	39.53 ± 0.10	55.53 ± 0.30	68.70 ± 0.46	**85.18 ± 0.31**
	SCL_EXP	27.17 ± 0.32	37.07 ± 1.96	38.00 ± 1.85	56.73 ± 3.82	84.70 ± 0.13
	PC	**40.61 ± 0.46**	39.96 ± 0.29	44.72 ± 0.64	62.63 ± 0.16	75.38 ± 0.60
	COVR (ours)	39.94 ± 2.09	**47.41 ± 0.60**	**61.29 ± 1.61**	**75.86 ± 1.20**	84.34 ± 0.18

G THE DIFFERENCE BETWEEN BIASED CL AND IDCL

In this section, we discuss the difference between biased CL and our IDCL. To this end, firstly, we applied the "Min3" method to generate complementary labels for each instance, thus forming the IDCL dataset. Subsequently, we calculated the true transition matrix based on the generated IDCL dataset, as illustrated in Figure 6, which presented the transition matrix for three different benchmark datasets. Furthermore, we generated the biased CL dataset using this transition matrix. Third, we compared various CLL methods under both "Biased" and our "Min3" settings. From Table 7, compared to the "Min3" setting, the CLL methods perform significantly better under the biased setting. This observation suggests that our instance-dependent setting poses a much more challenging CLL task than the biased setting, despite both settings having the same class-level transition matrix.

Moreover, we calculate the ISE during the training process for various CLL methods under both the "Min3" setting and the biased setting. Figure 7 reveals a substantial decrease in ISE for the existing CLL methods under the "Min3" setting, when compared to that of the biased setting. This result further illustrates that our IDCLL is a more challenging CLL task than the biased setting.



	0	1	2	3	4	5	6	7	8	9
0	0.00	0.27	0.00	0.12	0.24	0.07	0.03	0.08	0.14	0.03
1	0.29	0.00	0.05	0.07	0.01	0.24	0.14	0.01	0.00	0.18
2	0.10	0.08	0.00	0.00	0.21	0.11	0.13	0.09	0.04	0.25
3	0.15	0.05	0.01	0.00	0.28	0.01	0.29	0.16	0.01	0.03
4	0.20	0.18	0.09	0.21	0.00	0.08	0.12	0.02	0.09	0.01
5	0.07	0.15	0.17	0.02	0.22	0.00	0.09	0.22	0.01	0.05
6	0.03	0.17	0.01	0.18	0.02	0.01	0.00	0.25	0.06	0.26
7	0.08	0.12	0.03	0.02	0.20	0.14	0.33	0.00	0.08	0.00
8	0.11	0.17	0.03	0.03	0.22	0.03	0.17	0.19	0.00	0.05
9	0.17	0.22	0.17	0.02	0.01	0.06	0.32	0.01	0.02	0.00

(a) MNIST

	0	1	2	3	4	5	6	7	8	9
0	0.00	0.08	0.31	0.24	0.01	0.02	0.08	0.04	0.13	0.10
1	0.15	0.00	0.10	0.21	0.05	0.08	0.10	0.16	0.09	0.06
2	0.17	0.10	0.00	0.18	0.11	0.02	0.11	0.10	0.16	0.04
3	0.07	0.12	0.16	0.00	0.04	0.02	0.08	0.08	0.18	0.25
4	0.07	0.07	0.13	0.15	0.00	0.08	0.05	0.07	0.27	0.10
5	0.06	0.09	0.21	0.05	0.05	0.00	0.21	0.09	0.05	0.20
6	0.19	0.03	0.05	0.22	0.05	0.05	0.00	0.10	0.11	0.20
7	0.02	0.20	0.13	0.05	0.06	0.05	0.09	0.00	0.29	0.12
8	0.11	0.07	0.04	0.16	0.15	0.03	0.16	0.17	0.00	0.13
9	0.09	0.03	0.03	0.21	0.06	0.08	0.21	0.07	0.22	0.00

(b) KMNIST

	0	1	2	3	4	5	6	7	8	9
0	0.00	0.11	0.02	0.02	0.12	0.22	0.00	0.18	0.07	0.26
1	0.01	0.00	0.10	0.00	0.00	0.23	0.01	0.23	0.18	0.24
2	0.03	0.09	0.00	0.12	0.00	0.24	0.01	0.26	0.09	0.16
3	0.06	0.04	0.04	0.00	0.07	0.22	0.03	0.27	0.09	0.18
4	0.17	0.05	0.00	0.05	0.00	0.21	0.01	0.29	0.06	0.15
5	0.09	0.09	0.10	0.20	0.16	0.00	0.22	0.03	0.02	0.10
6	0.04	0.17	0.02	0.03	0.03	0.19	0.00	0.25	0.08	0.20
7	0.14	0.01	0.07	0.26	0.21	0.00	0.29	0.00	0.00	0.01
8	0.04	0.18	0.07	0.19	0.07	0.07	0.11	0.06	0.00	0.20
9	0.02	0.00	0.19	0.22	0.27	0.00	0.25	0.00	0.04	0.00

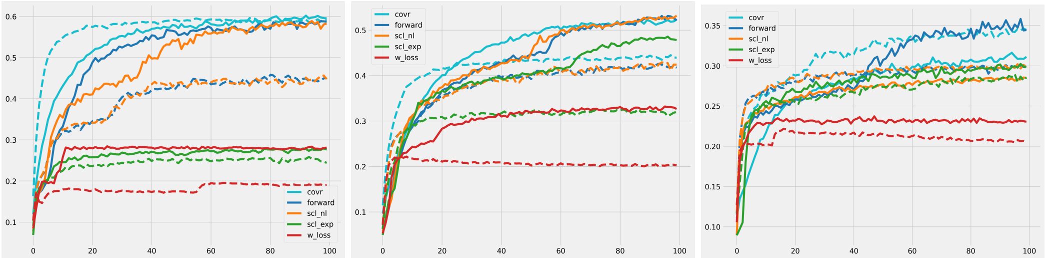
(c) FMNIST

Figure 6: The transition matrix generated by calculating the probabilities of each class label under "Min3" on three datasets, which was utilized to create the biased datasets.

We additionally illustrated this with an experiment by separately considering the confident and unconfident samples. To be specific, we selected all instances with ground-truth label "0", and

Table 7: Classification accuracy (mean±std) of each baseline approach on benchmark datasets under "Biased" and "Min3" settings, respectively. The symbol ↑ indicates an increase in classification accuracy for these approaches under the Biased setting compared to the "Min3" setting.

	MNIST		Kuzushiji-MNIST		Fashion-MNIST	
	Min3	**Biased**	**Min3**	**Biased**	**Min3**	**Biased**
Forward	51.82 ± 6.81	70.70 ± 4.55 ↑	53.69 ± 4.51	76.57 ± 0.14 ↑	39.54 ± 0.13	71.83 ± 0.96 ↑
SCL_NL	55.56 ± 6.91	71.10 ± 4.71 ↑	51.33 ± 1.76	76.89 ± 0.23 ↑	39.53 ± 0.08	72.25 ± 1.59 ↑
SCL_EXP	33.78 ± 4.03	40.27 ± 4.25 ↑	42.78 ± 4.20	51.91 ± 4.36 ↑	37.06 ± 1.60	52.22 ± 4.95 ↑
PC	83.94 ± 2.20	92.44 ± 2.71 ↑	64.99 ± 1.44	95.09 ± 0.44 ↑	45.10 ± 1.15	71.78 ± 3.44 ↑
COVR	87.52 ± 0.02	95.37 ± 0.46 ↑	67.31 ± 1.53	72.83 ± 0.47 ↑	47.41 ± 0.60	48.71 ± 2.13 ↑



(a) MNIST (b) KMNIST (c) FMNIST

Figure 7: Implicit sharing efficiency for various methods on MNIST,KMNIST and FMNIST datasets under both biased and "Min3" settings. Different colors denote different CLL methods, with the dotted line representing ISE under the "Min3" setting and the solid line representing the biased setting. It is evident that the solid line is higher than the dotted one with the same color, which means the CLL methods can share the complementary labels more efficiently under the biased setting than the "Min3" setting.

calculated the cross-entropy loss of all instances through a pre-trained model as a criterion for distinguishing whether the instances are confident: CE loss of the confident one should be small. We believe that the confident instances are easy to be classified, while the unconfident one are difficult to be classified. Then, we calculated the complementary label distribution for both confident and unconfident instances under the "Min3" and biased settings, respectively. The results are shown in Figure 8, denoted by confidence/unconfidence. It is evident that under the biased setting, both confident and unconfident instances share the same distribution. Conversely, when considering "Min3" setting, substantial disparities arise in the complementary label distributions. This inconsistency in complementary label distributions further increases the difficulty for the following CLL.

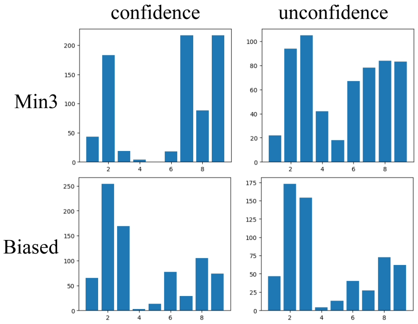


Figure 8: The complementary label distribution of confidence and unconfidence instances under "Min3" and "Biased" setting.

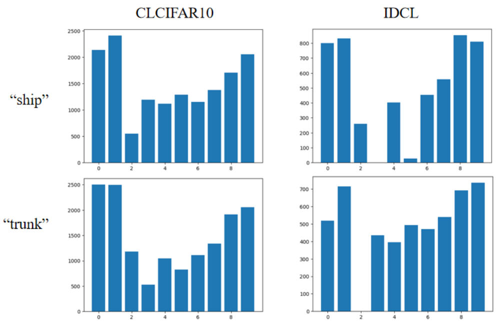


Figure 9: The distributions of CLCIFAR10 and IDCL when the ture labels are "ship" and "trunk". CLCIFAR10 contains noise because when annotating, the true label is not excluded from the randomly assigned four labels.

H EXPERIMENTS ON CLCIFAR10

CLCIFAR10 is considered as one of the earliest real-world datasets for CLL (Wang et al., 2023). To collect the complementary labels of instances in CIFAR10 annotated by human annotators, they first randomly sample four distinct labels for each instance (may contain the ground-truth label) and ask the human annotators to select any of the incorrect one from them as the complementary label of the instance. Each image in the dataset was labeled by three different annotators, and subsequently, one of the three selected complementary labels was randomly chosen as the final complementary label for each instance. This approach ensured the generation of reliable and diverse complementary labels for the instances in CLCIFAR10, and the resulting CLs can be regarded as the instance-dependent ones.

To further provide empirical support for our instance-dependent assumption, we conducted an analysis of the complementary label distribution within the real-world human-annotated CLCIFAR10 dataset. Figure 9 illustrates the complementary label distribution for two distinct classes in CLCIFAR10, namely “ship” and “trunk”, along with the proposed IDCL (“Min3”) method. From Figure 9, it becomes apparent that the complementary labels generated by “Min3” method closely resemble the complementary labels manually assigned by human annotators, regardless the noise in CLCIFAR10. This observation suggests that human annotators tend to select labels that are entirely unrelated to the actual instances as complementary labels. This alignment between the two distributions empirically validates our instance-dependent assumption and reveals the significance of the proposed approach.

REFERENCES (APPENDIX)

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels. International Conference on Machine Learning, pp. 1929–1938, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. Advances in Neural Information Processing Systems, 30, 2017.

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

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9, 2015.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. Proceedings of the European Conference on Computer Vision, pp. 68–83, 2018.