Discriminative Complementary-Label Learning with Weighted Loss

Yi Gao [1 2] **Min-Ling Zhang** [2 3]

Abstract

Complementary-label learning (CLL) deals with the weak supervision scenario where each training instance is associated with one *complementary* label, which specifies the class label that the instance does *not* belong to. Given the training instance $\boldsymbol{x}$, existing CLL approaches aim at modeling the *generative* relationship between the complementary label $\bar{y}$, i.e. $P(\bar{y} \mid \boldsymbol{x})$, and the ground-truth label y, i.e. $P(y \mid \boldsymbol{x})$. Nonetheless, as the ground-truth label is not directly accessible for complementarily labeled training instance, strong generative assumptions may not hold for real-world CLL tasks. In this paper, we derive a simple and theoretically-sound *discriminative* model towards $P(\bar{y} \mid \boldsymbol{x})$, which naturally leads to a risk estimator with estimation error bound at $\mathcal{O}(1/\sqrt{n})$ convergence rate. Accordingly, a practical CLL approach is proposed by further introducing weighted loss to the empirical risk to maximize the predictive gap between potential ground-truth label and complementary label. Extensive experiments clearly validate the effectiveness of the proposed discriminative complementary-label learning approach.

1. Introduction

Ordinary classification tasks generally require vast data with high-quality labels, while accurately annotating large-scale datasets is costly and time-consuming. The weakly supervised learning (WSL) paradigm has brought a new inspiration to alleviate this problem, which allows learning algorithms to train classifiers with less expensive data (Zhou, 2017; Ishida et al., 2017). The researchers have studied various frameworks based on weak supervision information, including but not limited to, *semi-supervised learning* (Chapelle et al., 2006; Oliver et al., 2018; Calder et al., 2020; Izmailov et al., 2020), *noisy-label learning* (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Ma et al., 2018; Kim et al., 2019; Liu & Guo, 2020; Han et al., 2020), *positive-unlabeled learning* (du Plessis et al., 2014; Sakai et al., 2018; Chapel et al., 2020; Hammoudeh & Lowd, 2020; Su et al., 2021), *unlabeled-unlabeled learning* (Lu et al., 2019; Golovnev et al., 2019) and *partial label learning* (Wu & Zhang, 2019; Lv et al., 2020).

Here, we consider another natural scenario of WSL – *complementary-label learning (CLL)* (Ishida et al., 2017; Yu et al., 2018; Ishida et al., 2019; Xu et al., 2020; Chou et al., 2020; Feng et al., 2020) – the class label specifies one of the classes that the instance does *not* belong to, while the learned classifier is expected to predict the ground-truth label of each instance. Collection of data with complementary labels is obviously much easier and less time-consuming than that of ordinary labels. To solve the CLL problem, previous approaches mainly focus on assuming the generative relationship between the complementary label $\bar{y}$ and the ground-truth label y of each instance, which could be roughly divided into two categories: the first category assumes that the relationship between $\bar{y}$ and y is unbiased based on an uniform distribution, i.e. $P(\bar{Y} = \bar{y} \mid X = \boldsymbol{x}) = \frac{1}{c-1}\sum_{y \neq \bar{y}} P(Y = y \mid X = \boldsymbol{x})$ (c refers to the number of classes) (Ishida et al., 2017; 2019; Feng et al., 2020), while the second one assumes that the relationship is biased, i.e. $P(\bar{Y} = \bar{y} \mid X = \boldsymbol{x}) = \sum_{y \neq \bar{y}} P(\bar{Y} = \bar{y} \mid Y = y) P(Y = y \mid X = \boldsymbol{x})$ (Yu et al., 2018; Xu et al., 2020).

As a pioneering work, the approach proposed by Ishida et al. (2017) designed an unbiased risk estimator (URE) with a solid theoretical analysis according to the assumption of the first category, which enables multi-class classification with only complementary labels. However, this approach only works with a limited group of loss functions, i.e., the one-versus-all (OVA) and the pairwise comparison (PC) loss functions (Zhang, 2004). With the same unbiased generation assumption, Ishida et al. (2019) proposed a general URE framework of complementary-label learning, which is unrestricted in models and loss functions. Nonetheless, these URE-based approaches in CLL may suffer from overfitting, as the empirical gradients may deviate from true

[1]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China [2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China [3]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. Correspondence to: Min-Ling Zhang <zhangml@seu.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

gradients during the optimization procedure (Chou et al., 2020).

Unlike previous studies, Yu et al. (2018) take the assumption that a biased relationship exists between the complementary label $\bar{y}$ and the ground-truth label y which is modeled by the transition probability, i.e. $P(\bar{Y} = \bar{y} \mid Y = y), \forall \bar{y} \neq y \in \{1, \ldots, c\}$. Their approach makes the widely-used multi-class Cross-Entropy (CE) loss be amenable for solving CLL tasks. Subsequently, Xu et al. (2020) applied Conditional Generative Adversarial Net (CGAN) (Goodfellow et al., 2014; Mirza & Osindero, 2014) based on the same biased assumption to improve the classification accuracy of CLL. However, this method requires extra conditions to be satisfied such as the availability of a set of anchor instances to enable transition probability estimation, which may not be satisfied in reality.

Overall, existing approaches rely on modeling the generative relationship between the complementary label and the ground-truth label of each training instance. Nevertheless, the ground-truth label is unknown for CLL training examples such that these strong generative assumption may be not suitable for solving real-world CLL problem. To tackle this problem, we propose a *discriminative* solution to directly model $P(\bar{y} \mid \boldsymbol{x})$ from the output of trained classifiers, which naturally leads to a novel CLL risk estimator. Specifically, a weighted loss is introduced to the empirical risk yielding a practical discriminative CLL approach. Experimental results on benchmark datasets demonstrate the effectiveness of the proposed discriminative CLL approach. The main contributions are summarized as follows:

(1) We directly model $P(\bar{y} \mid \boldsymbol{x})$ from the predictive probability of learned classifiers in a simple yet effective manner. Correspondingly, we derive a risk estimator with guaranteed estimation error bound at $\mathcal{O}(1/\sqrt{n})$ convergence rate.

(2) A practical CLL approach is proposed by introducing weighted loss to enforce predictive gap between potential ground-truth label and complementary label.

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 discriminative CLL approach with theoretical analyses and algorithmic details. Section 4 reports the results of comparative experimental studies. Finally, Section 5 concludes this paper.

2. Background and Formulation

In this section, we give notations used in this paper, and briefly discuss ordinary multi-class classification and complementary-label learning.

2.1. Ordinary Multi-Class Classification

In ordinary multi-class classification, let $\mathcal{X} \subset \mathbb{R}^d$ be the feature space and $\mathcal{Y} = \{1, \ldots, c\}$ be the label space, where c is the number of classes and $c \geq 2$. Let $p(\boldsymbol{x}, y)$ be the unknown probability density function over random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, and $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a set of n training examples each associated with a ground-truth label. Ordinary multi-class classification tasks aim to learn a classifier that maps from the feature space to the label space $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^c$, which is trained by minimizing the following classification risk:

$$R(\boldsymbol{f}) = \mathbb{E}_{(X,Y)\sim p(\boldsymbol{x},y)} \left[\ell(\boldsymbol{f}(X), e^Y) \right] \tag{1}$$

where $e^Y \in \{0, 1\}^c$ is the one-hot encoded label of X, and the Y-th element of e^Y is one with all other elements being zero. $\mathbb{E}$ and ℓ denote the expectation and the loss function, respectively. Accordingly, the most possible predicted label $\widehat{y}$ of an instance $\boldsymbol{x}$ is determined as

$$\widehat{y} = \operatorname*{argmax}_{k \in \mathcal{Y}} f_k(\boldsymbol{x}) \tag{2}$$

where $f_k(\cdot)$ denotes the k-th element of $\boldsymbol{f}(\cdot)$, referring to the posterior probability of the k-th label being the ground-truth one, i.e., $f_k(X) = P(Y = k|X)$. The optimal classifier $\boldsymbol{f}^*$ in function class $\mathcal{F}$ corresponds to the minimizer of classification risk $R(\boldsymbol{f})$: $\boldsymbol{f}^* = \operatorname{argmin}_{\boldsymbol{f} \in \mathcal{F}} R(\boldsymbol{f})$. As the underlying distribution $p(\boldsymbol{x}, y)$ is unknown, the classification risk in Eq.(1) is usually approximated by the empirical risk $R_n(\boldsymbol{f})$, i.e. $R_n(\boldsymbol{f}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{f}(\boldsymbol{x}_i), e^{y_i})$. Similarly, the optimal classifier w.r.t. the empirical risk corresponds to: $\boldsymbol{f}_n = \operatorname{argmin}_{\boldsymbol{f} \in \mathcal{F}} R_n(\boldsymbol{f})$.

2.2. Complementary-Label Learning

Different from ordinary multi-class classification, each instance only has one complementary label in CLL. Let $\bar{\mathcal{D}} = \{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$ denote the set of complementarily labeled training examples, where $\bar{y}_i \in \mathcal{Y} \setminus \{y_i\}$ is the complementary label of the instance $\boldsymbol{x}_i$ and each example is sampled from $\bar{p}(\boldsymbol{x}, \bar{y})$ which denotes an unknown probability distribution. As discussed in Section 1, existing approaches generally aim at modeling generative relationship between $P(\bar{Y} = \bar{y} \mid X = \boldsymbol{x})$ and $P(Y = y \mid X = \boldsymbol{x})$ (WLOG, we rewrite these terms as $P(\bar{y} \mid \boldsymbol{x})$ and $P(y \mid \boldsymbol{x})$ in the rest of this paper), which can be categorized into the unbiased generative assumption and the biased one, respectively. The work of Ishida et al. (2017) follows the first assumption to define $P(\bar{y} \mid \boldsymbol{x})$ as

$$\begin{aligned} \bar{p}(\boldsymbol{x}, \bar{y}) &= \frac{1}{c-1} \sum_{y \neq \bar{y}} p(\boldsymbol{x}, y) \\ \Leftrightarrow P(\bar{y} \mid \boldsymbol{x}) \bar{p}(\boldsymbol{x}) &= \frac{1}{c-1} \sum_{y \neq \bar{y}} P(y \mid \boldsymbol{x}) p(\boldsymbol{x}). \end{aligned} \tag{3}$$

Since $\bar{p}(\boldsymbol{x}) = p(\boldsymbol{x})$, we have $P(\bar{y} \mid \boldsymbol{x}) = \frac{1}{c-1} \sum_{y \neq \bar{y}} p(y \mid \boldsymbol{x})$. Based on Eq.(3), the OVA loss and PC loss for CLL, which naturally lead to an URE serving as an alternative formulation to Eq.(1), are defined as

$$\bar{\mathcal{L}}_{OVA}(\boldsymbol{f}(X), \bar{Y}) = \frac{1}{c-1}\sum_{Y\neq\bar{Y}} \ell(f_Y(X)) + \ell(-f_{\bar{Y}}(X))$$
$$\bar{\mathcal{L}}_{PC}(\boldsymbol{f}(X), \bar{Y}) = \sum_{Y\neq\bar{Y}} \ell(f_Y(X) - f_{\bar{Y}}(X)) \quad (4)$$

where $\ell(z)$ is a binary loss which satisfies $\ell(z)+\ell(-z)=1$, such as the sigmoid loss $\ell_S(z)=\frac{1}{1+e^z}$.

Different from Ishida et al. (2017; 2019), Yu et al. (2018) took another assumption which considers that the generative relationship between $P(\bar{y} \mid \boldsymbol{x})$ and $P(y \mid \boldsymbol{x})$ is biased, i.e. the complementary label of an instance X is non-uniformly selected from $\mathcal{Y}\setminus\{Y\}$. Therefore, this biased assumption can be formalized as $P(\bar{Y}=j \mid X) = \sum_{k\neq j} P(\bar{Y}=j \mid Y=k)P(Y=k \mid X)$, where this biased generative relationship can be characterized by a transition probability matrix Q, i.e. $Q_{kj}=P(\bar{Y}=j \mid Y=k)$ and $Q_{kk}=0$, $\forall k,j\in\{1,\ldots,c\}$.

Although feasible CLL approaches have been developed by exploiting either the unbiased (uniform) or biased (transition-based) generative assumption, their performance may be suboptimal for real-world CLL tasks where the two assumptions do not necessarily hold. In this paper, we propose a simple yet effective discriminative solution towards CLL which directly models $P(\bar{y} \mid \boldsymbol{x})$ from the predictive probability, and the solution naturally results in a CLL risk estimator with estimation error bound. We further introduce the weighted loss to maximize the predictive gap between the potential ground-truth label and complementary label.

3. The Proposed Approach

In this section, we introduce the proposed discriminative model and weighted loss. Accordingly, we further derive the estimation error bound of our approach.

3.1. The Discriminative Model

In ordinary multi-class classification, we aim to optimize Eq.(1) where the predictive probability of the ground-truth label approaches one, i.e. $f_Y(X)\to 1$, and other other labels to zero. In contrast, due to the complementary label is obviously not the ground-truth label of an instance, CLL expects that the predictive probability of the complementary label approaches zero, i.e. $f_{\bar{Y}}(X)\to 0$ (Kim et al., 2019).

The idea of Kim et al. (2019) also brings a strong motivation for us to propose the discriminative model that directly estimates $P(\bar{y} \mid \boldsymbol{x})$ from the classifiers' output. Different from the approach proposed by Chou et al. (2020), we directly define the prediction probability of complementary label as $\bar{\boldsymbol{f}}(X) = 1 - \boldsymbol{f}(X)$. Hence, the complementary loss $\bar{\ell}$ can be expressed as

$$\bar{\ell}(\boldsymbol{f}(X), e^{\bar{Y}}) = \ell(\bar{\boldsymbol{f}}(X), e^{\bar{Y}}) = \ell(1-\boldsymbol{f}(X), e^{\bar{Y}}) \quad (5)$$

where $e^{\bar{Y}}\in\{0,1\}^c$ is a one-hot vector for label $\bar{Y}$, in which the $\bar{Y}$-th element of $e^{\bar{Y}}$ is one and all other elements being zero. As discussed above, the novel risk estimator for CLL can be described as

$$\bar{R}(\boldsymbol{f}) = \mathbb{E}_{(X,\bar{Y})\sim\bar{p}(\boldsymbol{x},\bar{y})}\left[\bar{\ell}(\boldsymbol{f}(X), e^{\bar{Y}})\right]. \quad (6)$$

Correspondingly, the empirical risk estimator corresponds to

$$\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} \ell(1-f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i}) \quad (7)$$

where $e_k^{\bar{y}_i}$ is the k-th element of $e^{\bar{y}_i}$, $\bar{f}_k$ denotes the k-th element of $\bar{\boldsymbol{f}}$.

3.2. Estimation Error Bound

Let $\mathcal{F}=\{\boldsymbol{f}(\boldsymbol{x})\}$ be a c-valued function class to minimize empirical risk, where $\boldsymbol{f}=\{f_1,\ldots,f_c\}\in\mathcal{F}$. We denote $\widehat{\mathfrak{R}}_n(\mathcal{F})$ as the Rademacher complexity of $\mathcal{F}$ for $\mathcal{X}$ with data size n (Mohri et al., 2012) and $\bar{\boldsymbol{f}}_n^* = \operatorname{argmin}_{\boldsymbol{f}\in\mathcal{F}}\bar{R}_n(\boldsymbol{f})$. Using M and L_ℓ to denote the upper bound and Lipschitz constant of ordinary loss function ℓ respectively. Here, we first establish the upper bound of $\widehat{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{F})$ in Lemma 1, which naturally leads to the uniform deviation bound that further guarantees to derive the estimation error bound. We start investigating Lemma 1 from Assumption 1.

Assumption 1. *The loss function $\ell(\cdot,\cdot)$ satisfies*

$$\ell(1-f_k(X), 1-e_k^Y) \le \ell(f_k(X), e_k^Y). \quad (8)$$

Such an assumption holds for some commonly used loss functions, such as MSE (Mean Squared Error) loss and MAE (Mean Absolute Error) loss.

Lemma 1. *Based on Eq.(5) and Assumption 1, it holds that*

$$\widehat{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{F}) \le c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k). \quad (9)$$

The proof is provided in Appendix. Given the upper bound for $\widehat{\mathfrak{R}}_n(\bar{\ell}\circ\mathcal{F})$, we can directly obtain Lemma 2 based on McDiarmid's inequality (McDiarmid, 2013) and symmetrization (Mohri et al., 2012), which defines the uniform deviation bound.

Lemma 2. *For any $\delta>0$, with probability at least $1-\delta$,*

$$\sup_{\boldsymbol{f}\in\mathcal{F}} \left|\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f})\right| \le 2c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M\sqrt{\frac{log(2/\delta)}{2n}} \quad (10)$$

where $\bar{R}(\boldsymbol{f})$ and $\bar{R}_n(\boldsymbol{f})$ is defined by Eq.(6) and Eq.(7) respectively. The proof is given in Appendix.

According to Lemma 2, we can establish the estimation error bound for the proposed CLL risk estimator. The estimation error bound is shown in Theorem 1, whose proof is presented in Appendix.

Algorithm 1 CLL with weighted loss

Input:
$\bar{\mathcal{D}}$: the complementary-label training set $\{\boldsymbol{x}_i, \bar{y}_i\}_{i=1}^n$;
T : the number of epochs;
$\mathcal{A}$: an external stochastic optimization algorithm;
Output:
θ : model parameter for $\boldsymbol{f}(\boldsymbol{x}; \theta)$;
$f_k(\cdot)$: the k-th element of $\boldsymbol{f}(\boldsymbol{x}; \theta)$ and the predictive probability of the k-th label being the ground-truth label of an instance;
for $t = 1$ to T **do do**
  Shuffle $\bar{\mathcal{D}}$ into $\mathcal{B}$ mini-batchs each with size s;
  for $b = 1$ to $\mathcal{B}$ **do do**
    Let $\boldsymbol{x}_i^b$ be the i-th instance in b-th mini-batch, and $\bar{y}_i^b$ be the corresponding complementary label;
    Set $w_i^k = \frac{1-f_k(\boldsymbol{x}_i^b)}{\sum_{j=1}^c (1-f_j(\boldsymbol{x}_i^b))}$;
    Let $\mathcal{L}^b$ be the risk of b-th mini-batches, $\mathcal{L}^b = \frac{1}{s}\sum_{i=1}^s \sum_{k=1}^c (1 + w_i^k)\bar{\ell}(f_k(\boldsymbol{x}_i^b), e_k^{\bar{y}_i^b})$;
    Set gradient $-\nabla_\theta \mathcal{L}^b$;
    Update θ by $\mathcal{A}$;
  end for
end for

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

$$\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}(\bar{\boldsymbol{f}}^*) \leq 4c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M\sqrt{\frac{2log(2/\delta)}{n}}. \tag{11}$$

For all parametric models with a bounded norm, as $n \to \infty$, $\bar{R}(\bar{\boldsymbol{f}}_n^*) \to \bar{R}(\bar{\boldsymbol{f}}^*)$. Theorem 1 shows that the proposed risk estimator exists an estimation error bound and convergence rate is $\mathcal{O}(1\sqrt{n})$. Note that in Eq.(11), c^2 shows that the number of labels have a strong impact on our empirical performance. This implication agrees well with our expectation: the fewer number of labels, the more effective our proposed CLL method.

3.3. The Weighted Loss

Commonly, the estimated posterior probability can be regarded as one metric to measure the prediction uncertainty and increasing uncertainty could lead to a deteriorated prediction performance (Yao et al., 2020). In Subsection 3.1, we propose to estimate the posterior probability of the complementary label. Although the proposed method is theoretically sound, its performance still depends heavily on the number of instances and the number of labels. In this part, we consider employing the prediction uncertainty in our proposed method. In this way, the highly confident predictions in the early stage of learning can be employed to boost the performance of succeeding updating of the model. Our solution is to introduce a weighted loss term to $\bar{\ell}$ to minimize the loss value in CLL, which is defined as

$$\bar{\ell}(\boldsymbol{f}(X), e^{\bar{Y}}) = w\ell(1 - \boldsymbol{f}(X), e^{\bar{Y}}) \tag{12}$$

where w corresponds to a *loss weight vector* in the c-dimensional simplex Δ^{c-1}.

Intuitively, w should be related to the prediction uncertainty and should be updated constantly through the whole learning process. A recent line of work proposes strategies, including maximum likelihood and maximum margin, to highlight the ground-truth label (Nguyen & Caruana, 2008; Liu & Dietterich, 2012; Yao et al., 2020; Jin & Ghahramani, 2002). In order to stand out the ground-truth label from all labels, maximum likelihood methods generally adopt EM procedure to optimize their models, which firstly use an independent E-step to learn weights, then train the models until convergence in the M-step (Jin & Ghahramani, 2002; Liu & Dietterich, 2012; Lv et al., 2020). However, E-step of these methods are separated from the M-step. In this way, these methods are easy to have a greedy solution, which will lead to the overfitting problem(Lv et al., 2020). Maximum margin methods maximize the margin between the ground-truth label and other labels to make the ground-truth label gradually prominent (Nguyen & Caruana, 2008). As thoroughly discussed in Yao et al. (2020), these methods are difficult to be calibrated in the later processing when false positive is selected in the current step.

To address aforementioned problems, we use the current prediction probability of the complementary label to make more use of highly possible complementary labels. Moreover, E-step and M-step are considered as a whole during the training procedure and weights can be updated easily as well. Specially, we set $w = [w_1, w_2, \ldots, w_c]$, where w_k is the k-th element of w and is defined as

$$w_k = \frac{1 - f_k(X)}{\sum_{j=1}^c (1 - f_j(X))}. \tag{13}$$

Note that $\sum_{k=1}^c w_k = 1$ and $w_k \geq 0$. Let us explain the setting of w with a simple CLL task of three labels ($c = 3$). Let $\bar{\boldsymbol{f}}(\boldsymbol{x}_i) = [0.1, 0.7, 0.2]$ denote the predicted posterior probability of the three labels for $\boldsymbol{x}_i$, we can infer that the first label is the potential ground-truth label because $\boldsymbol{f}(x_i) = 1 - \bar{\boldsymbol{f}}(x_i)$. By our setting of w, the weight vector in Eq.(13) become $[0.1, 0.7, 0.2]$ as well, and we apply a smaller weight to treating the ground-truth label as the complementary label, and a larger weight to treating two other labels, especially the highly confident second label as complementary labels. Therefore, the potential ground-truth label will be prominent gradually as the increasing of the predictive gap between the potential ground-truth label and the complementary label of each instance. Accordingly, we add the weighted loss and the unweighted loss together, resulting in our targeted loss

$$\bar{\ell}(\boldsymbol{f}(X), e^{\bar{Y}}) = \sum_{k=1}^c (1 + \lambda w^k)\ell(1 - f_k(X), e_k^{\bar{Y}}) \tag{14}$$

and the empirical risk estimator for CLL is described as

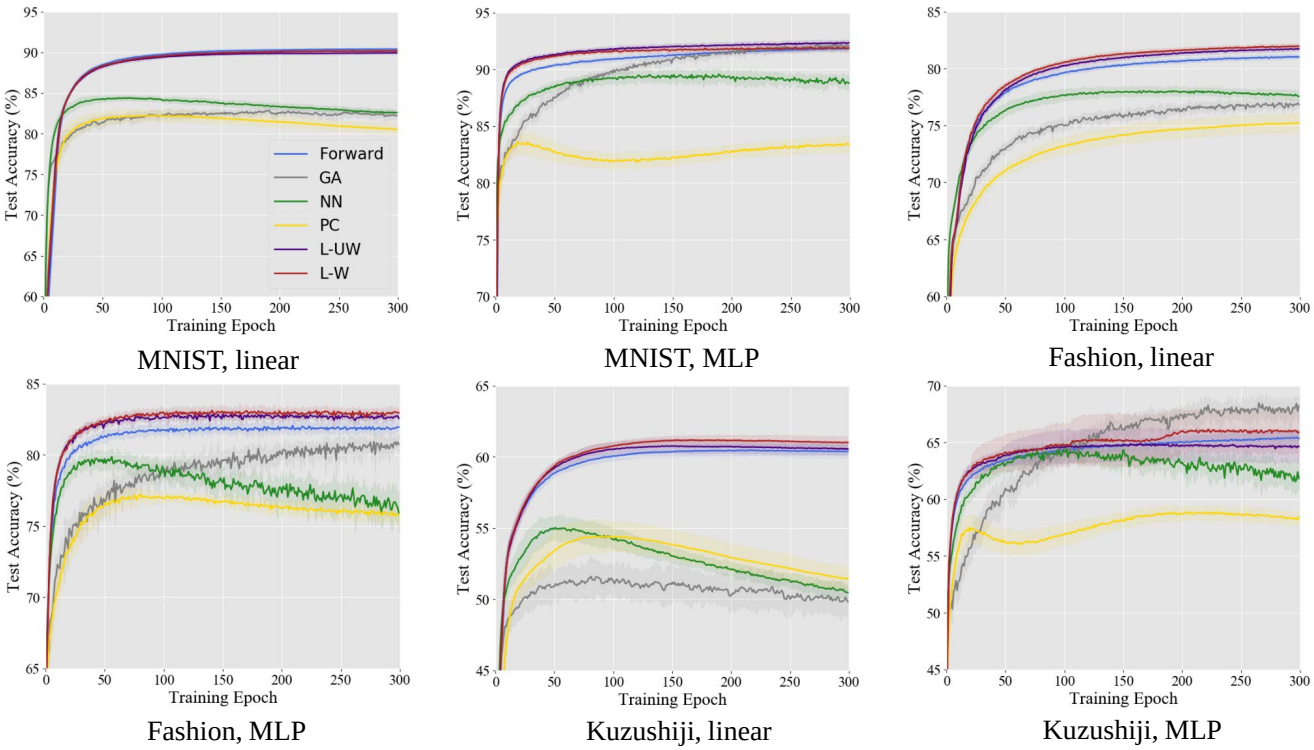

Figure 1. The experimental results on various test datasets with different loss functions and models for 300 epochs. The dark color is the mean accuracy and the light color corresponds to the std.

$$\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}(1+\lambda w_i^k)\ell(1-f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i}) \qquad (15)$$

where w_i^k corresponds to the k-th element of loss weight vector w_i for instance $\boldsymbol{x}_i$. λ is the tradeoff parameter between the original loss function and the weighted loss function, and we set it simply as 1. The overall procedure of the proposed approach is shown in Algorithm 1.

4. Experiments

In this section, we evaluate the performance of the proposed approach with comparative studies against state-of-the-art complementary-label learning approaches. We use L-UW and L-W to denote the proposed CLL approach instantiated with complementary loss function in Eq.(5) and Eq.(14) respectively. Due to $\bar{\boldsymbol{f}}$ is directly applied to a model training, which will result in the gradient diffusion problem of a model, we employ the constraint $\sum_{k=1}^{c}\bar{f}_k(X)=1$, where softmax function is used to normalize $\bar{\boldsymbol{f}}$ to make $\bar{\boldsymbol{f}}$ satisfy the constraint. Then, $\bar{\boldsymbol{f}}$ can be immediately defined as $\bar{\boldsymbol{f}} = \mathrm{softmax}(1-\boldsymbol{f})$, where $\bar{f}_k = \exp(1-f_k)/\sum_{j=1}^{c}\exp(1-f_k)$. The cross-entropy loss is commonly applied to multi-class classification tasks, which is adopted to replace ℓ in this paper. All experiments are implemented based on PyTorch (Paszke et al., 2019) and Colab [1]. The code is available at https://github.com/Yolk-justlike/complementary-label-learning.

[1]https://colab.research.google.com

4.1. Experimental Settings

Datasets Following Ishida et al. (2017; 2019); Yu et al. (2018); Feng et al. (2020), three widely-used benchmark datasets, namely MNIST (Lecun et al., 1998), Fashion-MNIST (Fashion) (Xiao et al., 2017), and Kuzushiji-MNIST (Kuzushiji) (Clanuwat et al., 2018), are used for experimental studies.

- MNIST dataset (Lecun et al., 1998) is a handwritten digits dataset that consists of 10 classes, which has 60,000 training examples and 10,000 test examples.
- Fashion dataset is collected by Xiao et al. (2017) from standardized images of fashion items, which has 60,000 training images and 10,000 test images from 10 classes.
- The size of Kuzushiji dataset (Clanuwat et al., 2018) is similar to MNIST dataset. Kuzushiji dataset derives from Kuzushiji which includes 60,000 training images and 10,000 test images from 10 classes.

Base models Two base models are utilized: linear model and MLP model ($d-500-c$).

Baselines We employ four state-of-the-art CLL approaches to be compared with, including Pairwise Comparison (PC) with sigmoid loss (Ishida et al., 2017), forward loss correction (Forward) (Yu et al., 2018), Gradient Ascent (GA)

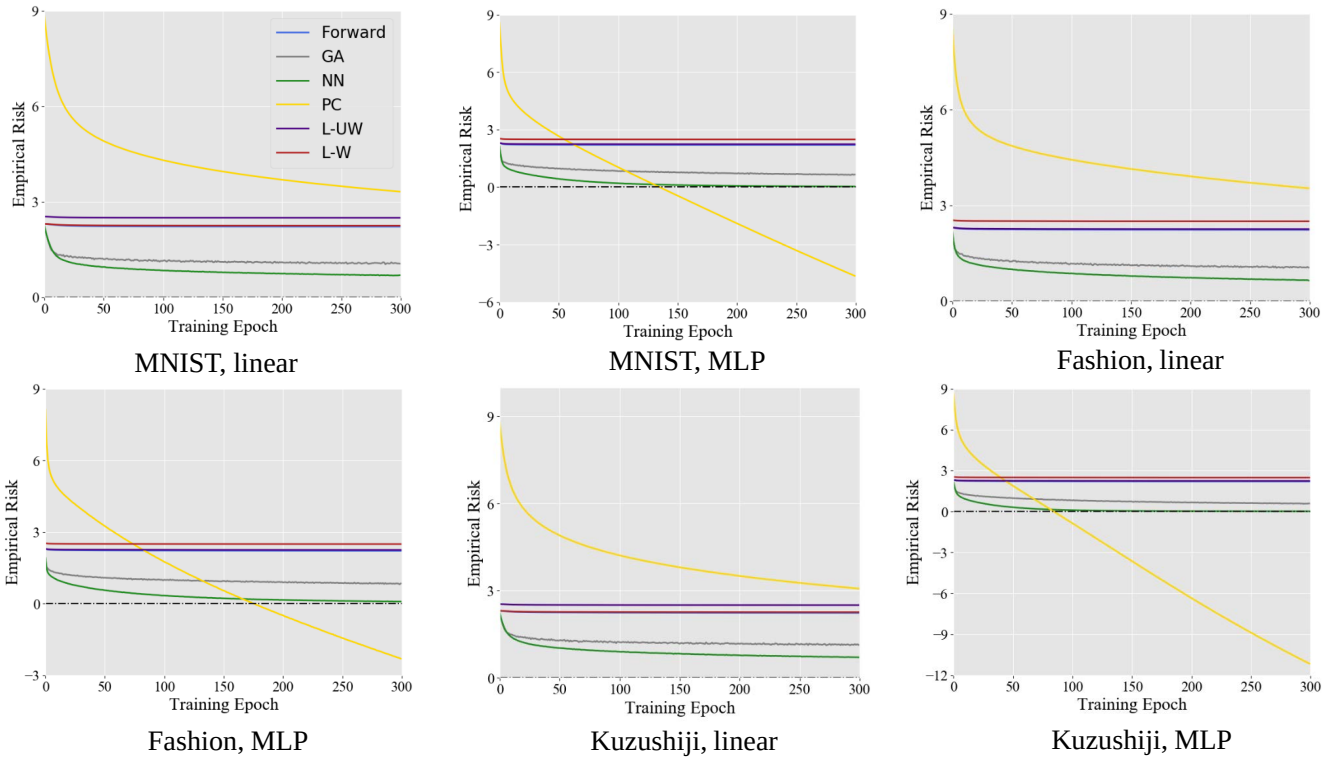

Figure 2. Empirical risk minimization procedure for various models and loss functions.

Table 1. Test accuracy (mean±std) out of 10 trials (in %), where data with unbiased complementary labels is used to train. The best performance on each data set is shown in boldface.

Dataset	Model	PC	Forward	GA	NN	L-UW	L-W
MNIST	linear	82.31±0.72	**90.42±0.17**	83.23±0.43	84.56±0.31	89.98±0.20	90.22±0.11
	MLP	84.04±0.55	91.93±0.25	**92.49±0.25**	89.99±0.42	92.45±0.24	92.08±0.28
Fashion	linear	75.29±0.83	81.14±0.20	77.41±0.30	78.32±0.31	81.79±0.22	**82.04±0.21**
	MLP	77.55±0.39	82.31±0.24	81.62±0.19	80.29±0.47	83.15±0.20	**83.40±0.32**
Kuzushiji	linear	54.57±1.13	60.57±0.42	52.52±1.12	55.27±0.85	60.87±0.48	**61.29±0.31**
	MLP	59.32±0.59	65.59±0.54	**69.56±0.53**	65.44±0.51	65.17±1.43	66.98±1.63

Table 2. The Win/Loss statistics for the proposed approach of Table 1. If our approach outperforms comparison baselines, add 1 to the count of ours; otherwise, add 1 to the comparison baselines.

Baselines	PC	Forward	GA	NN
L-UW	6/0	4/2	4/2	5/1
L-W	6/0	5/1	4/2	6/0

Table 3. The Win/Loss statistics for the proposed approach of Table 4. If our approach outperforms comparison baselines, add 1 to the count of ours; otherwise, add 1 to the comparison baselines.

Baselines	PC	Forward	GA
L-UW	14/4	13/5	15/3
L-W	15/3	16/2	15/3

(Ishida et al., 2019) and Non-Negative loss (NN) (Ishida et al., 2019).

4.2. Comparison on Unbiased Complementary Labels

Setup Weight decay is set as 1e-4 and learning rate of 5e-5 is used for MNIST, Fashion and Kuzushiji. Adam (Kingma & Ba, 2015) optimization method is applied. For all datasets, the number of epoch and mini-batch size are set as 300 and 256 respectively.

We divide the original training dataset into training and validation parts with proportion 9/1, where complementary labels are generated by randomly choosing one of the labels other than the ground-truth one (unbiased complementary-label generation). Test set with ordinary labels is used to evaluate the performance of each comparing approach. The mean and standard deviation (std) of test accuracy out of 10

Table 4. Test accuracy (mean±std) on three datasets out of 5 trials (in %), where data with biased complementary labels is used to train. The best performance on each data set is shown in boldface.

Set 1						
Baselines		PC	Forward	GA	L-UW	L-W
MNIST	linear	**19.66±0.28**	19.54±0.58	9.86±0.15	18.23±0.17	18.57±0.55
	MLP	19.34±0.69	20.44±0.15	9.80±0.00	19.46±0.34	**21.13±2.06**
Fanshion	linear	10.65±1.11	12.71±2.73	10.01±0.17	13.73±0.28	**14.40±0.55**
	MLP	15.40±3.55	16.94±0.37	10.06±0.87	20.81±1.32	**21.77±0.93**
Kuzushiji	linear	**14.10±0.80**	13.40±0.88	10.63±0.30	13.17±0.62	13.83±0.24
	MLP	14.43±0.87	12.97±0.97	10.22±0.38	13.56±0.59	**14.75±0.10**
Set 2						
Baselines		PC	Forward	GA	L-UW	L-W
MNIST	linear	19.69±0.63	20.31±0.10	10.19±0.16	23.55±2.05	**23.67±0.74**
	MLP	22.59±2.32	20.44±0.20	10.09±0.00	23.35±0.66	**126.76±2.00**
Fanshion	linear	12.43±2.51	12.71±2.73	9.75±0.32	17.30±0.35	**20.14±0.70**
	MLP	15.41±3.74	16.94±0.37	10.29±0.50	**23.54±0.93**	23.17±0.56
Kuzushiji	linear	12.73±0.14	**13.35±0.40**	9.89±0.57	12.43±0.27	12.51±0.25
	MLP	14.93±1.03	12.71±0.66	10.00±0.00	16.45±0.26	**17.28±0.45**
Set 3						
Baselines		PC	Forward	GA	L-UW	L-W
MNIST	linear	72.22±1.43	78.53±4.41	78.55±0.80	**81.16±0.12**	79.72±0.27
	MLP	84.46±0.23	80.67±5.34	85.13±0.10	84.98±0.10	**85.91±0.11**
Fanshion	linear	58.34±0.61	60.28±3.76	**65.62±0.17**	59.71±4.08	61.57±0.45
	MLP	58.00±0.91	61.92±3.56	**63.80±0.07**	62.19±0.13	63.11±0.12
Kuzushiji	linear	46.54±0.43	54.41±1.95	50.16±0.41	54.47±1.10	**57.85±2.32**
	MLP	51.85±1.58	51.05±1.52	52.24±0.72	**52.56±3.62**	52.02±3.68

trials on the model that corresponds to the best validation score on 300 epochs are shown in Table 1.

Results We show the mean and std of test accuracy for 300 epochs on MNIST, Fanshion, and Kuzushiji in Figure 1. Figure 2 illustrates corresponding empirical risk of PC (Ishida et al., 2017), Forward (Yu et al., 2018), GA (Ishida et al., 2019), NN (Ishida et al., 2019), L-UW and L-W on three benchmark datasets for all epochs during the process of training.

Based on the reported results in Figure 1, we can observe that the proposed discriminative CLL approach achieves better or at least comparable performance against the comparing approaches on different datasets. As shown in Figure 1, the std of GA is greater than that of L-UW and L-W, which demonstrates that the performance of our approaches is more stable than GA on different training data partitioning. Furthermore, the test accuracy of approaches in later training epochs gradually decrease when the more complex models are applied, which is especially prominent under the case of using MLP model. This is because overparameterized deep neural networks are available to make the training loss go zero via memorizing training data, while the model becomes overconfident with a weak generalization performance that result in the degraded test performance (Ishida et al., 2020).

The viewpoint as mentioned earlier is reflected in Figure 2 as well, all approaches work normally with linear base model on MNIST, Fashion and Kuzushiji, while empirical risk of URE-based methods, such as PC and NN, goes zero or even negative when MLP model is applied (Ishida et al., 2019). We observe that the test accuracy of PC starts decreasing when its empirical risk trends to negative. In comparison, the gradient ascent trick is used in GA when the empirical risk approaches negative, which saves GA from the deteriorated performance.

In Table 1, we report the mean and std of the classification accuracy on test data out of 10 trials, where Table 2 shows the Win/Loss statistics of the proposed approach outperforming other baselines. The following observations can be made based on Table 1 and Table 2: 1) L-UW (without weighted loss term) achieves comparable test accuracy to PC, Forward, GA and NN on different datasets, which indicates that the proposed simple discriminative model (Eq.(5)) serves as a feasible solution to CLL problem; 2) L-W (with weighted loss term) works well under all cases; it shows that the introduction of weighted loss to the discriminative model does help improve the generalization performance by maximizing the predictive gap between potential ground-truth label and complementary label.

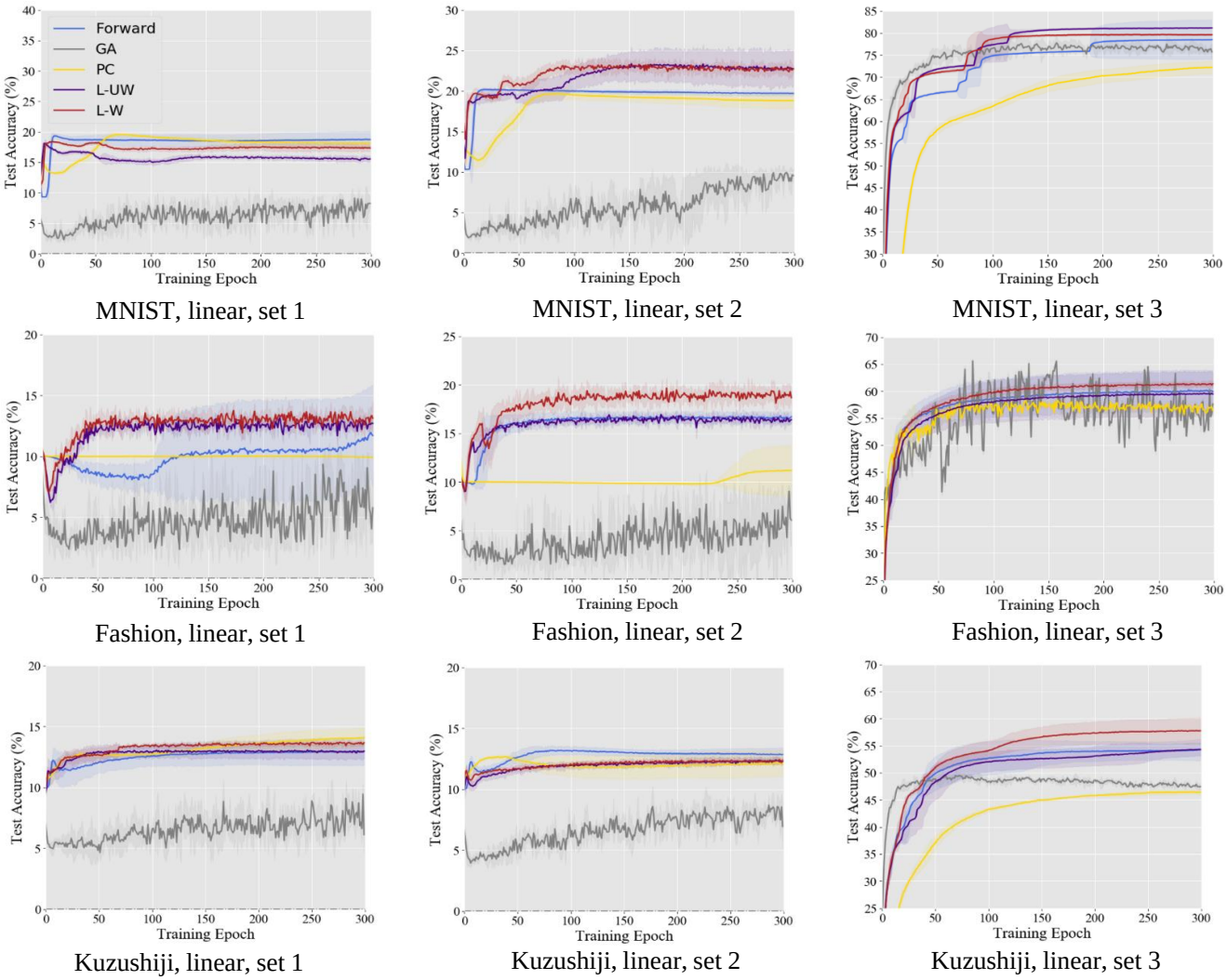

Figure 3. The experimental results on various biased settings on the linear model for 300 epochs. The dark color is the mean accuracy and the light color corresponds to the std.

4.3. Comparison on Biased Complementary Labels

Setup We use training data which is associated with biased complementary labels to evaluate the effectiveness of our approach. The biased complementary-label generation is similar to Yu et al. (2018). More specifically, we adopt three settings to generate biased complementary labels. For all settings, the complementary label is selected from $\mathcal{Y} \setminus \{y\}$, which is divided into three subsets randomly, each including three class labels. For set 1: the selected probabilities of each complementary label in three subsets are $0.75/3, 0.24/3$ and $0.01/3$ respectively; the selected probabilities of that are $0.66/3, 0.24/3$ and $0.1/3$ respectively for set 2; for set 3, the probability of each label is selected as the complementary label at probabilities $0.45/3, 0.3/3$ and $0.25/3$.

We utilize the training dataset associated with biased complementary labels to train the model, while test set with ordinary labels is applied to evaluate the performance of approaches. The other experimental settings are same with Subsection 4.2. The mean and std of test accuracy out of 5 trials on the model that corresponds to the best validation score on 300 epochs are shown in Table 4. Table 3 is used to count the Win/Loss results of the proposed approach that is superior to other baselines.

Results From the results shown in Table 4, we can find that the test accuracy of all approaches has improved as the biased degree of complementary labels decreasing, which also demonstrates that the performance of approaches will suffer from the non-uniform selection of complementary labels. Furthermore, experimental results in Table 4 and Table 3 show that our proposal is better to other baselines in most cases.

Figure 3 and Figure 4 illustrate the mean and std of test accuracy for all epochs on the linear model and MLP model

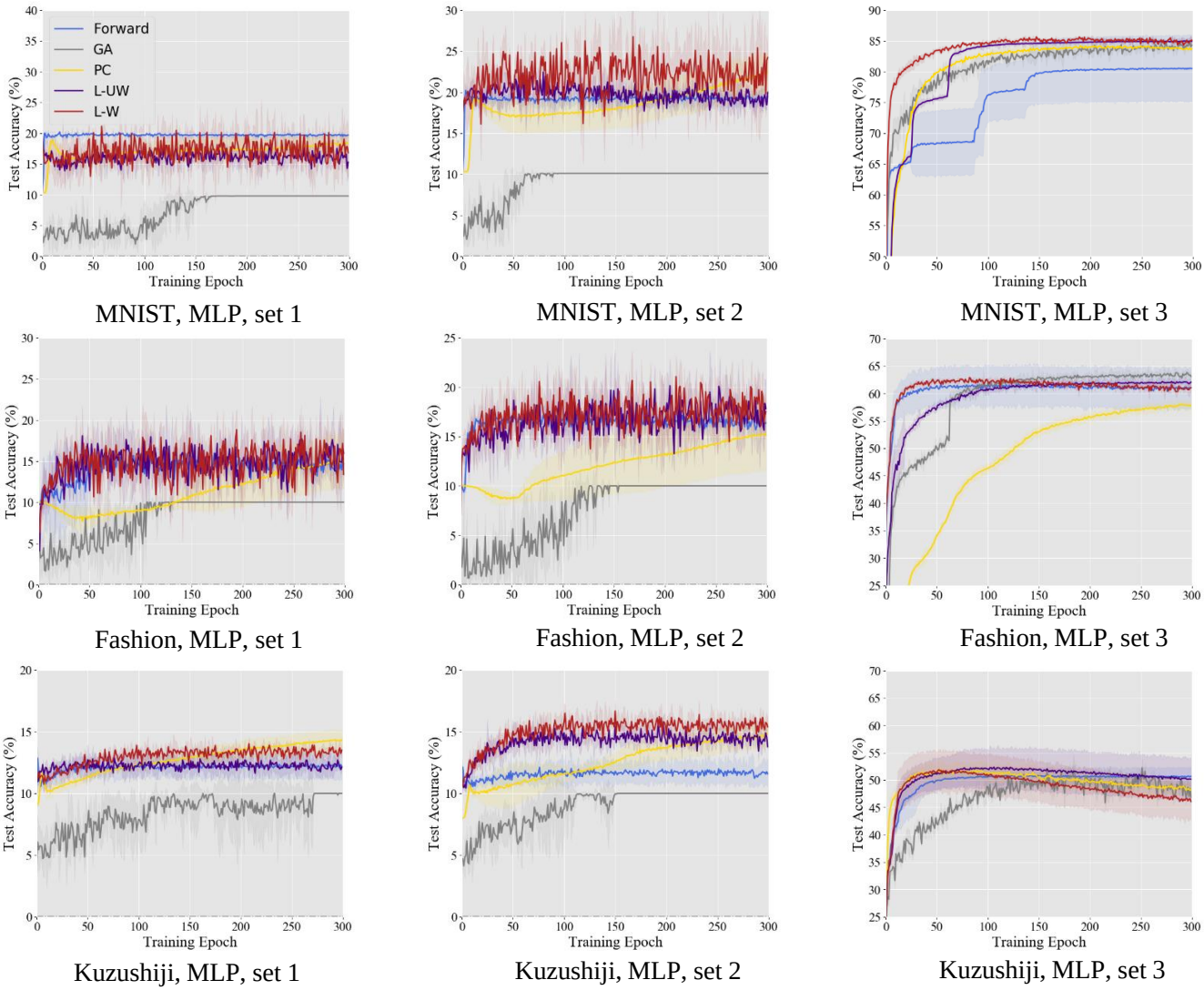

Figure 4. The experimental results on various biased settings on the MLP model for 300 epochs. The dark color is the mean accuracy and the light color corresponds to the std.

respectively with different biased settings. As shown in Figure 3 and Figure 4, L-W gets comparable test accuracy on various biased setting to Forward when the biased transition matrix with no additional information is available for Forward. Moreover, the fluctuation frequency of L-UW and L-W is less than that of GA in Figure 3, which indicates that L-UW and L-W have a stable performance in the biased complementary-label case. Due to the corresponding empirical risk of biased setting follow the same trend as the unbiased one, it is put in the Appendix.

5. Conclusion

In this paper, a risk estimator with guaranteed estimation error bound based on discriminative model is proposed for CLL. It estimates the complementary label predictions $P(\bar{y} \mid \boldsymbol{x})$ by the output of discriminative classifiers with sound theoretical properties. Accordingly, the weighted loss which makes use of the output of current classification model during the training procedure is further introduced to the classification risk to yield the empirical risk for CLL model training. The effectiveness of the proposed discriminative CLL model is clearly validated with extensive comparative studies over benchmark datasets.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Key R&D Program of China (2018YFB1004300), and the China University S&T Innovation Plan Guided by the Ministry of Education. We thank the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Calder, J., Cook, B., Thorpe, M., and Slepcev, D. Poisson learning: Graph based semi-supervised learning at very low label rates. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1306–1316, Virtual Event, 2020.

Chapel, L., Alaya, M. Z., and Gasso, G. Partial optimal tranport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems 33*, Virtual Event, 2020.

Chapelle, O., Schölkopf, B., and Zien, A. *Introduction to Semi-Supervised Learning*. The MIT Press, 2006.

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

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

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

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

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pp. 1919–1925, San Francisco, CA, 2017.

Golovnev, A., Pál, D., and Szörényi, B. The information-theoretic value of unlabeled data in semi-supervised learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2328–2336, Long Beach, CA, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, Montreal, Canada, 2014.

Hammoudeh, Z. and Lowd, D. Learning from positive and unlabeled data with arbitrary positive shift. In *Advances in Neural Information Processing Systems 33*, Virtual Event, 2020.

Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I. W., and Sugiyama, M. SIGUA: forgetting may make learning with noisy labels more robust. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4006–4016, Virtual Event, 2020.

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from complementary labels. In *Advances in Neural Information Processing Systems 30*, pp. 5639–5649, Long Beach, CA, 2017.

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

Ishida, T., Yamane, I., Sakai, T., Niu, G., and Sugiyama, M. Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4604–4614, Virtual Event, 2020.

Izmailov, P., Kirichenko, P., Finzi, M., and Wilson, A. G. Semi-supervised learning with normalizing flows. In *Proceedings of the 37th International Conference on Machine Learningt*, pp. 4615–4630, Virtual Event, 2020.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Advances in Neural Information Processing Systems 15*, pp. 897–904, Vancouver, Canada, 2002.

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

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

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

Liu, L. and Dietterich, T. G. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25*, pp. 557–565, Lake Tahoe, NV, 2012.

Liu, Y. and Guo, H. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6226–6236, Virtual Event, 2020.

Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier

from only unlabeled data. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, 2019.

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

Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S. M., Xia, S., Wijewickrema, S. N. R., and Bailey, J. Dimensionality-driven learning with noisy labels. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3361–3370, Stockholm, Sweden, 2018.

McDiarmid, C. On the method of bounded differences. *Surveys in Combinatorics*, pp. 148–188, 2013.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

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

Nguyen, N. and Caruana, R. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2008*, pp. 551–559, Las Vegas, NV, 2008.

Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., and Goodfellow, I. J. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems 31*, pp. 3239–3250, Montréal, Canada, 2018.

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. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Vancouver, Canada, 2019.

Sakai, T., Niu, G., and Sugiyama, M. Semi-supervised AUC optimization based on positive-unlabeled learning. *Machine Learning*, 107(4):767–794, 2018.

Su, G., Chen, W., and Xu, M. Positive-unlabeled learning from imbalanced data. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, Virtual Event, 2021.

Wu, J. and Zhang, M. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 416–424, Anchorage, AK, 2019.

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

Xu, Y., Gong, M., Chen, J., Liu, T., Zhang, K., and Batmanghelich, K. Generative-discriminative complementary learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pp. 6526–6533, New York, NY, 2020.

Yao, Y., Deng, J., Chen, X., Gong, C., Wu, J., and Yang, J. Deep discriminative CNN with temporal ensembling for ambiguously-labeled image classification. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pp. 12669–12676, New York, NY, 2020.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *Proceedings of the 15th European Conference on Computer Vision*, pp. 69–85, Munich, Germany, 2018.

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

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

Zhou, Z. H. A brief introduction to weakly supervised learning. *National Science Review*, (1):1, 2017.

Supplementary Material for Discriminative Complementary-Label Learning with Weighted Loss

A. The Proof of Lemma 1

Lemma 1. *Based on Eq.(5) and Assumption 1, it holds that*

$$\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) \leq c^2 L_\ell \widehat{\mathfrak{R}}_n\left(\mathcal{F}_k\right)$$

Proof. Given

$$\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(\boldsymbol{f}(\boldsymbol{x}_i), e^{\bar{y}_i}) \right]$$

where $\sigma = [\sigma_1, \ldots, \sigma_n]$, which denotes n Rademacher varibles. Let us first assume $c = 2$ and use the max operator $\max(a, b) = \frac{1}{2}(a + b + |a - b|)$. Thus, we have

$$\begin{aligned}
&\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(\boldsymbol{f}(\boldsymbol{x}_i), e^{\bar{y}_i}) \right] \\
&= \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \sum_{k=1}^{c} \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i}) \right] \\
&\leq \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} c\sigma_i \max_{k \in \{1, \ldots, c\}} \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i}) \right] \\
&\leq \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{n} c\sigma_i \bar{\ell}(f_1(\boldsymbol{x}_i), e_1^{\bar{y}_i}) \right] \\
&+ \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{n} c\sigma_i \bar{\ell}(f_2(\boldsymbol{x}_i), e_2^{\bar{y}_i}) \right] \\
&+ \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{n} c\sigma_i |\bar{\ell}(f_1(\boldsymbol{x}_i), e_1^{\bar{y}_i}) - \bar{\ell}(f_2(\boldsymbol{x}_i), e_2^{\bar{y}_i})| \right] \\
&\leq 2\mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} c\sigma_i \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i}) \right]
\end{aligned}$$

When there are c classes, the general case can be derived from $\max\{z_1, \ldots, z_c\} = \max\{z_1, \max\{z_2, \ldots, z_c\}\}$, by recurrence, we will have

$$\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) \leq c^2 \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i}) \right]$$

By our Assumption 1 and the definition of $\bar{\ell}(\cdot)$, we further have

$$\begin{aligned}
&\mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i}) \right] \\
&\leq \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(1 - f_k(\boldsymbol{x}_i), 1 - e_k^{y_i}) \right]
\end{aligned}$$

$$\leq \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f_k(\boldsymbol{x}_i), e_k^{y_i}) \right] = \widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}_k)$$

According to Talagrand's contraction lemma (Ledoux & Talagrand, 1991), we have $\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) \leq c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k)$. □

B. The Proof of Lemma 2

Given the upper bound for $\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F})$, we can prove Lemma 2 that defines the uniform deviation bound.

Lemma 2. *For any $\delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{\boldsymbol{f} \in \mathcal{F}} \left| \bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}) \right| \leq 2c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M \sqrt{\frac{log(2/\delta)}{2n}}$$

where $\bar{R}(\boldsymbol{f})$ and $\bar{R}_n(\boldsymbol{f})$ is defined by Eq.(6) and Eq.(7) respectively.

Proof. Consider the single direction $sup_{\boldsymbol{f} \in \mathcal{F}}(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}))$ with probability at least $1 - \delta/2$. Because M is the upper bound of ℓ, the change of $sup_{\boldsymbol{f} \in \mathcal{F}} \left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}) \right)$ is no greater than M/n after some x are replaced. So using McDiarmid's inequality (McDiarmid, 2013) to $sup_{\boldsymbol{f} \in \mathcal{F}} \left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}) \right)$, we have

$$\sup_{\boldsymbol{f} \in \mathcal{F}} \left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}) \right) \leq \mathbb{E} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}) \right) \right] + M \sqrt{\frac{log\,(2/\delta)}{2n}}$$

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

$$\mathbb{E} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}) \right) \right] \leq 2 \widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) = 2c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k)$$

□

C. The Proof of Theorem 1

According to Lemma 2, we can establish the estimation error bound for the proposed CLL risk estimator. The estimation error bound is shown in Theorem 1.

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

$$\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}(\bar{\boldsymbol{f}}^*) \leq 4c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M \sqrt{\frac{2log(2/\delta)}{n}}$$

Proof.

$$\begin{aligned}
\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}(\bar{\boldsymbol{f}}^*) &= (\bar{R}_n(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}^*)) + (\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}_n^*)) + (\bar{R}_n(\bar{\boldsymbol{f}}^*) - \bar{R}(\bar{\boldsymbol{f}}^*)) \\
&\leq \bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}_n^*) + \bar{R}_n(\bar{\boldsymbol{f}}^*) - \bar{R}(\bar{\boldsymbol{f}}^*) \\
&\leq 2 \sup_{\boldsymbol{f} \in \mathcal{F}} \left| \bar{R}_n(\bar{\boldsymbol{f}}) - \bar{R}(\bar{\boldsymbol{f}}) \right| \\
&\leq 4c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M \sqrt{\frac{2 \log (2/\delta)}{n}}
\end{aligned}$$

Since $\bar{R}_n(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}^*) \leq 0$, the second step in the above equation naturally follows from the first step. The proof is complete. □

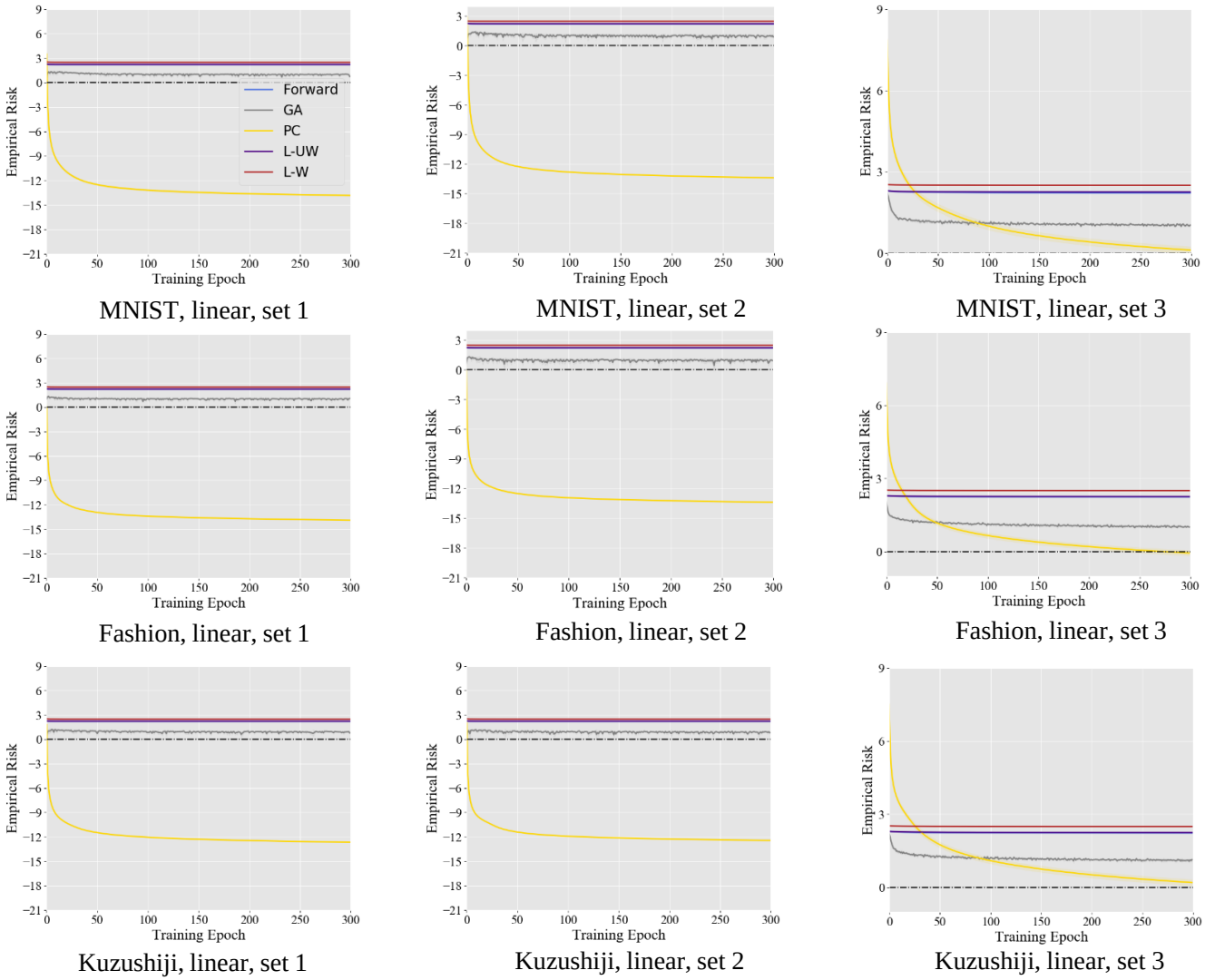

Figure 1. The experimental results on various biased settings on the linear model for 300 epochs. The dark color is the mean accuracy and the light color corresponds to the std.

D. Empirical Risk for Biased Settings

Figure 1 and Figure 2 are corresponding empirical risks for the linear model and MLP model on various datasets and biased settings.

Results From Figure 1, the empirical risk of PC on three datasets goes non-negative when the generation setting of complementary labels gradually becomes uniform. Furthermore, as shown in Figure 1, all approaches work normally with linear base model on MNIST, Fashion-MNIST and Kuzushiji-MNIST, while empirical risk of URE-based methods, such as PC and GA, goes zero or even negative when the more complex models are applied (shown in 2). Specifically, under the case of using MLP model, the performance of PC becomes the worst. This is due to the property that URE-based methods are easy to suffer from over-fitting problem when using complex models (Chou et al., 2020).

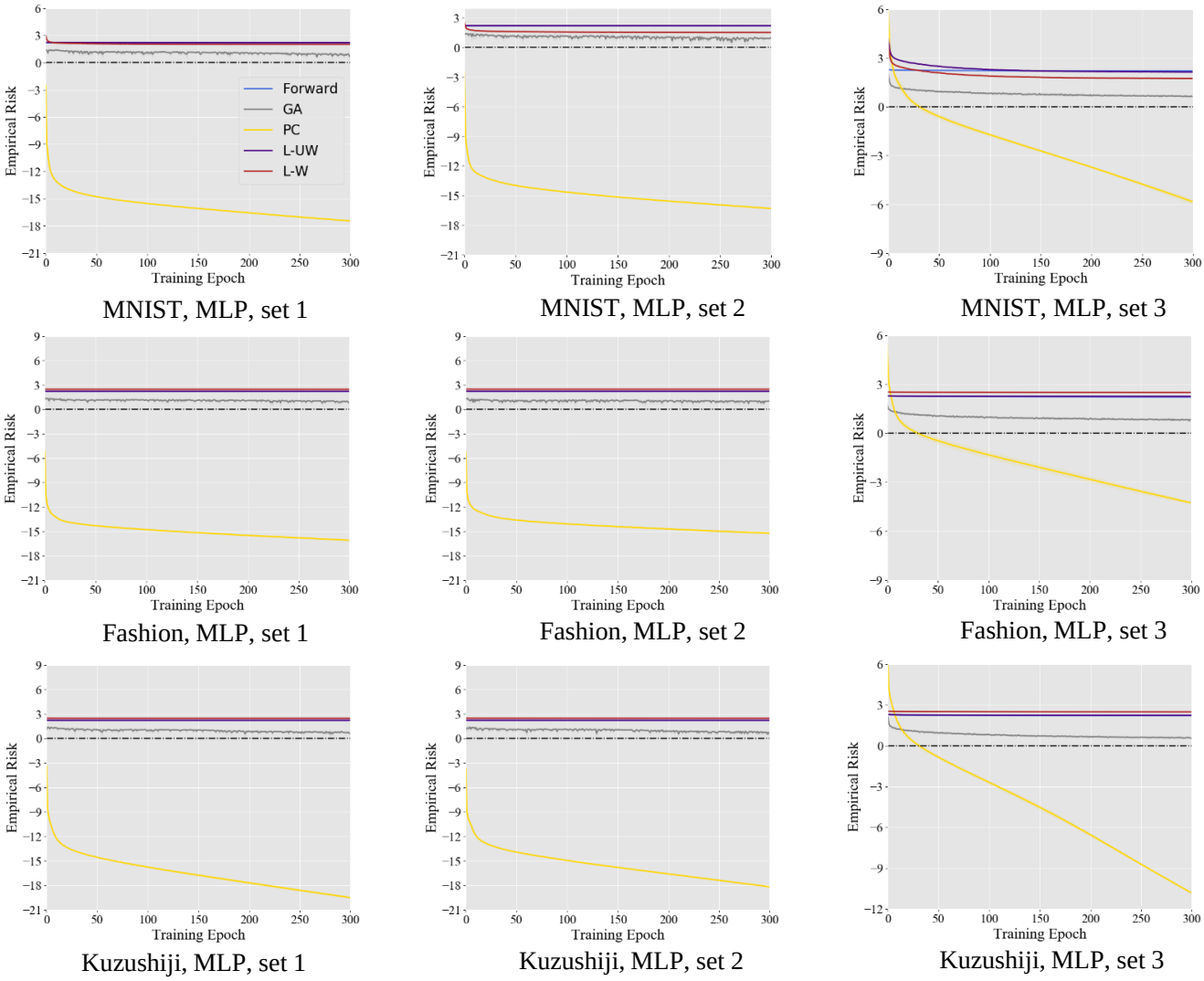

Figure 2. The experimental results on various biased settings on the MLP model for 300 epochs. The dark color is the mean accuracy and the light color corresponds to the std.