Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels

Yu-Ting Chou [1 *] **Gang Niu** [2] **Hsuan-Tien Lin** [1] **Masashi Sugiyama** [2 3]

Abstract

In weakly supervised learning, *unbiased risk estimator* (URE) is a powerful tool for training classifiers when training and test data are drawn from different distributions. Nevertheless, UREs lead to overfitting in many problem settings when the models are complex like deep networks. In this paper, we investigate reasons for such overfitting by studying a weakly supervised problem called *learning with complementary labels*. We argue the quality of *gradient estimation* matters more in risk minimization. Theoretically, we show that a URE gives an *unbiased gradient estimator* (UGE). Practically, however, UGEs may suffer from huge variance, which causes empirical gradients to be usually far away from true gradients during minimization. To this end, we propose a novel *surrogate complementary loss* (SCL) framework that trades zero bias with reduced variance and makes empirical gradients more aligned with true gradients in the direction. Thanks to this characteristic, SCL successfully mitigates the overfitting issue and improves URE-based methods.

1. Introduction

In *weakly supervised learning* (WSL), learning algorithms have to train classifiers under incomplete, inexact or inaccurate supervision (Zhou, 2017), including but not limited to semi-supervised learning (Chapelle et al., 2009), partial labels (Jin & Ghahramani, 2002), noisy labels (Natarajan et al., 2013; Patrini et al., 2017; Han et al., 2018a;b; Yu et al., 2019; Xia et al., 2019), complementary labels (Ishida et al., 2017; Yu et al., 2018; Ishida et al., 2019; Xu et al., 2020; Feng et al., 2020), where the label distribution changes, and positive-unlabeled data (Elkan & Noto, 2008; du Plessis et al., 2014; 2015; Niu et al., 2016; Sakai et al., 2017; 2018), unlabeled-unlabeled data (Lu et al., 2019; 2020), and other similar settings (Bao et al., 2018; Ishida et al., 2018; Hsieh et al., 2019), where the data distribution changes. Among WSL methods, *unbiased risk estimator* (URE) is a powerful tool: it evaluates the *classification risk* from training data drawn from a distribution different from the test one, and thus *empirical risk minimization* (Vapnik, 1992) is possible. The success of URE is due to two orthogonal demands in WSL for handling big data and complex data: URE poses *unconstrained optimizations* so that it can handle very big data by *stochastic optimizers*; URE is *model-independent* so that it can handle complex data where the model is chosen according to the data (e.g., image, text, or speech).

An important motivation of employing URE in WSL is that URE enables *estimation error bounds* to guarantee *statistical consistency*. However, the consistency in the *asymptotic cases* is not very meaningful in the *finite-sample cases* especially in deep learning (Zhang et al., 2017; Nagarajan & Kolter, 2019). Despite its popularity and nice properties, URE in du Plessis et al. (2015), Ishida et al. (2017) or Lu et al. (2019) has inferior test performance to recent biased methods in Kiryo et al. (2017), Ishida et al. (2019) and Lu et al. (2020). When complex models like deep networks are chosen as the classifiers, UREs suffer from severe *negative empirical risks* during training, which is a sign of overfitting. Even though the overfitting issue can be relatively mitigated by keeping UREs non-negative, the mechanism behind how UREs cause overfitting is still unknown. Thus, instead of a theoretical motivation, this paper has a practical motivation and focuses on understanding how UREs cause overfitting and how to avoid such overfitting in algorithm design.

Learning with complementary labels (Ishida et al., 2017) is a WSL problem of multi-class classification where classifiers are trained from data with complementary labels (CL). A CL specifies a class that an instance *does not belong to*, but the trained classifier should still predict the correct labels. Although CLs are less informative than ordinary labels, they provide an alternative when ordinary labels are inaccessible or costly to acquire. In this paper, we choose learning with CLs to study the overfitting issue of UREs, as it combines several practical advantages: first, CLs are easy to *generate*

*Work done during an internship at RIKEN. [1]National Taiwan University [2]RIKEN [3]The University of Tokyo. Correspondence to: Yu-Ting Chou <r07922042@csie.ntu.edu.tw>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

compared with partial labels and noisy labels; second, negative empirical risks are easy to *occur*; and third, it is easy to experimentally *analyze* the *bias* and *variance* of empirical gradients. With the help of such a case study, we can gain a deep insight of UREs and lay the foundation for further studies of UREs in other WSL problem settings.

Our contributions can be summarized in two folds. First of all, we conduct a series of analyses to investigate reasons for the overfitting issue. We show that due to the linearity of the differential operator, any URE must give an *unbiased gradient estimator* (UGE); however, UGE is not necessarily good at gradient estimation though it is unbiased. During training, only a single fixed CL could be acquired for each instance, which causes empirical gradients given by a UGE to be usually far away from true gradients. This illustrates the difference between *validation* and *training*:

- In validation, the classifier is fixed and the data is repeatedly sampled, and then UGE is good at gradient estimation (which can be theoretically guaranteed by *concentration inequalities*).
- In training, the data are fixed and based on these data the classifier is iteratively updated, and then UGE might be really bad at gradient estimation.
- Theoretically speaking, good validation can imply good training if the model is simple, while good validation may still result in poor training if the model is complex (Zhang et al., 2017; Nagarajan & Kolter, 2019).

Unfortunately, UGEs in training suffer from huge variance in learning with CLs. Here, the *root cause* of overfitting is that only one fixed CL is available for each instance, and the *direct cause* is the huge variance of UGEs and the distance from empirical to true gradients. The root cause also exists in other WSL problem settings, e.g., partial or noisy labels. Notice that the quality of gradient estimation matters more than risk estimation in risk minimization, since stochastic optimizers mainly rely on empirical gradients.

Next, we propose a novel framework named *surrogate complementary loss* (SCL) to improve gradient estimation. Recall that the *classification error* is defined as the expected zero-one loss over the test distribution. Existing URE-based methods first replace the zero-one loss with a surrogate loss to obtain the risk, and then rewrite the risk into an expectation over the training distribution. We call it *complementary surrogate loss* since replacing is before rewriting. On the other hand, our framework first rewrites the error into an expectation over the training distribution and then replaces the zero-one loss with a surrogate loss, namely, rewriting before replacing. Rewriting the error is nicer since the zero-one loss has many nice properties while the surrogate loss is just arbitrary. In our experiments, SCL-based methods outperform URE-based methods, where SCL successfully reduces the variance of empirical gradients and makes them more aligned with true gradients in the direction.

The rest of the paper is organized as follows. We introduce WSL problem settings and the overfitting issue in Section 2. In Section 3, we propose the SCL framework. In Section 4, we analyze empirical gradients to justify our claims.

2. The Use of Unbiased Risk Estimators

In this section we introduce the usage of unbiased risk estimators in several weakly supervised learning settings. Then we zoom into the problem of learning with complementary labels, and show the relationship between negative risk problem and overfitting.

2.1. Related WSL Settings

The following problems are typical examples where UREs fail under weak supervision. The negative empirical risk can happen when the loss functions are not specifically restricted, causing overfitting. Biased loss functions or non-negative correction methods are introduced to mitigate such issues in related literature.

Noisy Label Learning: Noisy label learning studies about learning when training labels flip according to some underlying distribution. A common assumption is the class conditional noise setting where the noisy label depends on its ordinary label. Natarajan et al. (2013) first provided a URE for arbitrary loss in the binary case, and provided performance guarantee. To ensure the convexity of the rewritten loss function, they require the original surrogate loss to satisfy a symmetric property. Patrini et al. (2017) extends to multiclass classification and proposed two loss correction methods: backward correction and forward correction. Backward correction involves a matrix inversion and gives an unbiased estimator of the original loss. Forward correction corrects the prediction with a matrix multiplication and can be added as an additional layer to neural networks. The authors showed that forward correction performs better than backward correction, and hinted the reason to be optimization related.

Positive-Unlabeled (PU) Learning: In binary classification, the labeled data consists of two sets, the positive (P) class and the negative (N) class. PU learning studies when labeled data only consists of positive examples, while we have unlabeled (U) data consisting of both positive and negative examples. Elkan & Noto (2008) proposed to learn from assigning weights to unlabeled examples. du Plessis et al. (2014) proposed a URE of non-convex losses, and du Plessis et al. (2015) extends it further to a more general framework with convex formulation. Kiryo et al. (2017) observed the overfitting issue of unbiased PU learning and proposed a non-negative risk estimator to fix the problem.

Unlabeled-Unlabeled (UU) learning: In binary classification, UU learning considers the setting when all labels are unknown. Lu et al. (2019) discovers that if the two sets of data have different class priors, a URE can be derived to learn from such data. However, the unbiased UU learning also encounters severe overfitting due to negative empirical risk. Lu et al. (2020) proposed a non-negative corrected risk estimator to fix the problem.

2.2. Learning with Complementary Labels

In the following part, we first introduce related work of learning with complementary labels, then formally define the URE formulation and the negative risk effect.

In Ishida et al. (2017), the first work to introduce the setting of complementary labels, a URE can be obtained when a loss function satisfies the symmetric property, under uniform assumptions. Yu et al. (2018) provides a loss correction method for softmax cross entropy loss, and shows that non-uniform complementary labels can also be learned if the complementary transition matrix is known. Continuing in the uniform complementary assumption of Ishida et al. (2017), Ishida et al. (2019) generalizes the URE for arbitrary loss functions and models, and proposes a non-negative correction and a gradient ascent method to account for overfitting. Several studies have also extended to learning with multiple complementary labels (Feng et al., 2020), and its combination with unlabeled data (Cao & Xu, 2020). The flexibility of CLs makes it easy to use in settings such as online learning (Kaneko et al., 2019), generative-discriminative learning (Xu et al., 2020), and noisy label learning (Kim et al., 2019).

Ordinary Learning: We start by reviewing the setting and introduce notations in ordinary learning. Consider the problem of K class classification ($K > 2$), where $[K] = \{1, 2, ..., K\}$ is the label set. Let D be a joint distribution over the feature set X and label set Y, where we sample input feature $x \in \mathbb{R}^d$ and label $y \in [K]$. Given training samples $\{(x_i, y_i)\}_{i=1}^n$, the goal of the learning algorithm is to learn a classifier $f(x) : \mathbb{R}^d \to [K]$ which predicts the correct label from a given input x. The classifier f is implemented with a decision function $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^K$ by taking the argmax function $f(x) = \arg\max_i \boldsymbol{g}(x)_i$. For a label y and a decision function output $\boldsymbol{g}(x)$, the *loss function* is defined as a nonnegative function $\ell : [K] \times \mathbb{R}^K \to \mathbb{R}^+$. Finally, we define the *risk* as the expected loss of $\boldsymbol{g}$ over distribution D:

$$R(\boldsymbol{g}; \ell) = \mathbb{E}_{(X,Y)\sim D}[\ell(Y, \boldsymbol{g}(X))]. \tag{1}$$

Complementary Learning: In complementary learning, the data distribution is switched to $\overline{D} = X \times \overline{Y}$ where the training samples given to the learner become $\{(x_i, \overline{y}_i)\}_{i=1}^n$. For instance x_i, the complementary label (CL) $\overline{y}_i$ is a class in $[K]$ that x_i does not belong to, satisfying $\overline{y}_i \neq y_i$. In this case, the loss function ℓ cannot be used directly since the ordinary target y_i is not given. In the following part, we review the derivation of URE using backward loss rewriting process (Patrini et al., 2017; Ishida et al., 2019).

Unbiased Risk Estimator: In this part, we follow the assumption of class conditional complementary transition as in related work, assuming the transition matrix T invertible, where $T_{ij} = \mathbb{P}(\overline{Y} = j \mid Y = i)$ and $T_{ii} = 0$ for all i. We borrow the following notation from Ishida et al. (2019). The loss vector is $\ell(\boldsymbol{g}(x)) = [\ell(1, \boldsymbol{g}(x)), \ell(2, \boldsymbol{g}(x))...\ell(K, \boldsymbol{g}(x))]$, and let $e_i \in \{0, 1\}^K$ denote the one-hot vector in which the i-th entry is one.

Proposition 1. *The ordinary risk can be transformed as*

$$R(\boldsymbol{g}; \ell) = \mathbb{E}_{(X,\overline{Y})\sim\overline{D}}[e_{\overline{y}}^\top (T^{-1})\ell(\boldsymbol{g}(x))]. \tag{2}$$

That is, we obtain an unbiased risk estimator (URE):

$$\overline{R}(\boldsymbol{g}; \overline{\ell}) = \mathbb{E}_{(x,\overline{y})\sim\overline{D}}[\overline{\ell}(\overline{y}, \boldsymbol{g}(x))] = R(\boldsymbol{g}; \ell) \tag{3}$$

where $\overline{\ell}$ is the following rewritten loss:

$$\overline{\ell}(\overline{y}, \boldsymbol{g}(x)) = e_{\overline{y}}^\top (T^{-1})\ell(\boldsymbol{g}(x)). \tag{4}$$

This proposition implies the expectation of $\overline{\ell}(\overline{y}, \boldsymbol{g}(x))$ under distribution $\overline{D}$ is equivalent to the ordinary risk $R(\boldsymbol{g}; \ell)$.

Uniform Assumption: In the rest of this paper, we assume CLs are sampled uniformly from $[K] \setminus \{y\}$, for a better comparison with Ishida et al. (2019). By plugging in the uniform assumption $T = \frac{1}{K-1}(\mathbf{1}_k - \boldsymbol{I}_k)$, we have the following formulation of $\overline{\ell}$,

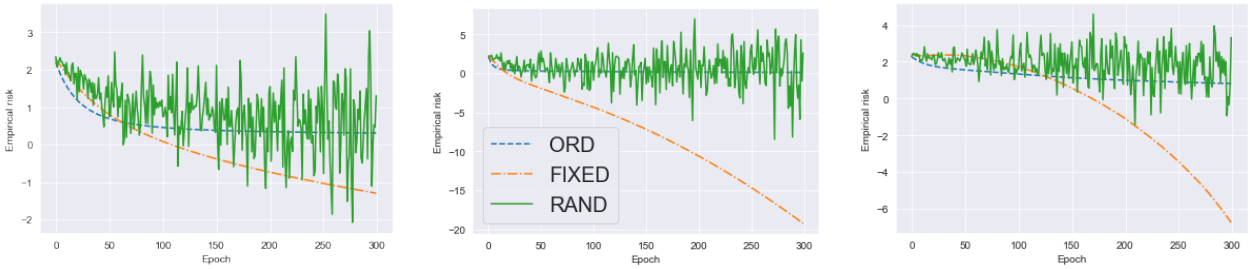
$$\overline{\ell}(\overline{y}, \boldsymbol{g}(x)) = -(K-1)\ell(\overline{y}, \boldsymbol{g}(x)) + \sum_{j=1}^{K} \ell(j, \boldsymbol{g}(x)). \tag{5}$$

This URE approach minimizes $\overline{\ell}$ over the training distribution, and theoretical results from Ishida et al. (2017) proved the consistency of the risk estimator under specific losses.

2.3. Negative Risk and Overfitting

However, URE tends to have poor empirical performance. Ishida et al. (2019) reported that minimizing URE causes the empirical risk to go negative, which is a sign of overfitting. It is clear that the negative loss term $-(K-1)\ell(\overline{y}, \boldsymbol{g}(x))$ in $\overline{\ell}$ (Equation 5) is the source of negativity. Such negative term occurs in common class conditional complementary transition as long as all diagonal elements of T are zero.

Recall the URE in Equation 3, in expectation has minimum value 0 when the classifier has no error. However, when



(a) MNIST, Linear (b) MNIST, MLP (c) CIFAR-10, DenseNet

Figure 1. Empirical risk minimization comparison

minimizing URE empirically, the non-negative lower bound does not remain. We claim that the main difference between the expectation and its empirical realization is the label distribution: only single $\overline{y}$ is given for each instance in practice, while the expectation is calculated over all possible $\overline{y}$. The URE only stays non-negative when taken expectation, which is not realistic.

Negative Risk Experiment: To show the difference between theory (expectation) and practice, we use an experiment to demonstrate how the empirical distribution of CLs leads to negative empirical risk during training. Three different label distributions are given:

1. Ordinary Learning (ORD): The supervised learning baseline, which the ordinary label y is given. This is also the case where the complementary label is marginalized out by taking expectation.
2. Fixed Complementary Learning (FIXED): The realistic complementary learning scenario, for each instance x only a fixed CL $\overline{y}$ is given.
3. Random Complementary Learning (RAND): The $\overline{y}$ of each instance is randomly sampled from $[K] \setminus \{y\}$ in each epoch. This setting acts as a stochastic version of ORD on $\overline{y}$.

In this experiment, we used the cross-entropy loss as ℓ for ordinary learning (ORD) and $\overline{\ell}$ for complementary learning (FIXED, RAND). For MNIST, we use linear model and single hidden layer MLP $(d-500-10)$ as learning models; for CIFAR-10, we used ResNet-34 (He et al., 2016) and DenseNet (Huang et al., 2017). The models are trained with Adam (Kingma & Ba, 2015) optimizer at a fixed learning rate of 10^{-5} for 300 epochs.

Results are shown in Figure 1. FIXED suffers from severe negative risk in comparison to ORD and RAND, which is a clear sign of overfitting to the given CL. The problem worsen as flexible models are used, matching results from Ishida et al. (2019). However, note that RAND yields a significantly different result from FIXED even though they are trained on the same objective. Though the risk of RAND fluctuates considerably due to the changes in each epoch, it does not stay negative, as we can view RAND as an randomized approximation of ORD. The results also show that the estimated risk diverges far from the ordinary risk as the training goes on, and the gap increases with the training epochs. In this case, consistency guarantees become ineffective since the risk estimation error keeps increasing as training goes on. That is, the behavior of URE and the ordinary risk is extremely different in the empirical setting, even if statistical properties such as unbiasedness and consistency can be proven.

Risk Correction Methods: Ishida et al. (2019) proposed two correction methods to mitigate the problem. First, the non-negative loss correction (NN), which enforces non-negativity to the decomposed risk of each class. Second, namely the gradient ascent correction (GA) which enforces a reverse gradient update to the model parameters when the decomposed risk goes negative or under a certain threshold. GA can be viewed as a more aggressive correction than NN. The correction methods show improvements in various experiments, and similar techniques have also been applied in other WSL problems (Kiryo et al., 2017; Lu et al., 2020). However, such correction methods are still based on URE and lack theoretical motivation, the fundamental difference between risk and URE are not solved. We will include experiment results of these methods in the following sections.

3. Proposed Framework

In this section, we propose a complementary learning framework that avoids the negative risk problem of URE. To clearly distinguish between complementary learning and ordinary learning, we rethink the relationship between input features and labels: An ordinary label provides a positive feedback to the given class, while a CL provides a negative feedback to the given class. The maximum likelihood approach is commonly used in ordinary learning when we have probability estimation from the model, by maximiz-

ing the conditional likelihood given the training data. The commonly used softmax cross-entropy loss function in deep learning is a typical example by combining softmax activation function and the maximum likelihood approach. In complementary learning, given only CLs as training data, we propose to apply the minimum complementary likelihood approach, through a proxy loss. In the following of this section, we propose a new framework that consists complementary 0-1 loss and its corresponding surrogate complementary loss (SCL).

3.1. Complementary 0-1 Loss

From the classification error perspective: In ordinary learning, zero error is obtained when the classifier predicts the correct class as the label, and has error otherwise. In complementary learning, given only limited information, we can only be sure that prediction error occurs when the CL is predicted by the classifier. With the rules above, we formally define the ordinary classification error and a novel complementary classification error:

Definition 1. *(Multiclass) classification error, or 0-1 loss:*

$$\ell_{01}(y, f(x)) = [\![y \neq f(x)]\!]. \tag{6}$$

Definition 2. *Complementary classification error, or complementary 0-1 loss:*

$$\overline{\ell}_{01}(\overline{y}, f(x)) = [\![\overline{y} = f(x)]\!]. \tag{7}$$

$\overline{\ell}_{01}$ is 1 when the predicted class matches the CL, which indicates classification error. By minimizing $\overline{\ell}_{01}$, we can minimize the conditional probability output of CLs.

Proposition 2. *The complementary 0-1 loss is a constant multiple of the URE of the classification error.*

$$R(\boldsymbol{g}; \ell_{01}) = (K-1)\overline{R}(\boldsymbol{g}; \overline{\ell}_{01}) \tag{8}$$

In other words, the URE of the classification error has the same minimizer with the complementary 0-1 loss:

$$\mathbb{E}_{(x,\overline{y})\sim\overline{D}}[\overline{\ell}_{01}(\overline{y}, \boldsymbol{g}(x))] \tag{9}$$

Thus, existing guarantees show that we can learn with CLs via empirical risk minimization from $\overline{R}(\boldsymbol{g}; \overline{\ell}_{01})$.

3.2. Surrogate Complementary Loss

To minimize the non-convex $\overline{\ell}_{01}$, a common approach in statistical learning is to select a convex surrogate loss to approximate the target loss. In order to minimize the output of the label prediction, which is the opposite of most common surrogate functions, we require a new type of surrogate complementary loss (SCL) for this problem setting. Different from ordinary surrogate losses which are non-increasing functions of the label class output, SCLs are non-decreasing functions of the CL class output.

Baseline Methods: To better distinguish from URE-based methods, we use ϕ to denote the SCL loss functions. Here we denote the probability output $\boldsymbol{p} \in \Delta^{K-1}$ if $\boldsymbol{g}$ passes through a softmax layer, where Δ^{K-1} is the K-dimensional simplex. Existing work on complementary learning has resulted in similar patterns that minimize label class prediction output. We include these methods as baselines in our experiments.

1. Forward correction (SCL-FWD) in Yu et al. (2018): a forward loss correction method given transition matrix T:
$$\phi_{\text{FWD}}(\overline{y}, \boldsymbol{g}(x)) = \ell(\overline{y}, T^\top \boldsymbol{p}). \tag{10}$$
2. Negative learning loss (SCL-NL) in Kim et al. (2019): a modified log loss for negative learning with CLs:
$$\phi_{\text{NL}}(\overline{y}, \boldsymbol{g}(x)) = -\log(1 - \boldsymbol{p}_{\overline{y}}). \tag{11}$$
3. Exponential loss (SCL-EXP):
$$\phi_{\text{EXP}}(\overline{y}, \boldsymbol{g}(x)) = \exp(\boldsymbol{p}_{\overline{y}}). \tag{12}$$

As we unify the above-mentioned losses into the surrogate complementary loss ϕ framework. These loss functions actually all accomplish the same purpose: minimizing the complementary 0-1 loss by using its loss as surrogate:

$$\min \overline{\ell}_{01}(\overline{y}, f(x)) \to \min \phi(\overline{y}, \boldsymbol{g}(x)). \tag{13}$$

Here we compare the proposed SCL learning process with the URE learning process, as shown in Figure 2. We use *approximation* step to denote the process of replacing 0-1 loss with its surrogate loss, and the *estimation* step represents rewriting the risk from ordinary distribution to complementary distribution. Given the same goal of minimizing the true classification risk R_{01}, the two frameworks follow a different order in the learning steps. The URE framework follows the traditional statistical learning framework by approximating R_{01} with R_ℓ, and then performs the estimation step by rewriting the risk into $\overline{R}_{\overline{\ell}}$ for the complementary distribution. The SCL framework, on the other hand, performs the approximation step after the estimation step by first rewriting the classification risk R_{01} to complementary classification risk $\overline{R}_{\overline{01}}$, then perform the approximation step by using the SCL loss ϕ, resulting in the objective $\overline{R}_\phi$.

The ordinary surrogate loss ℓ in URE is used for ordinary labels, which serves as an upper bound proxy in order to minimize the 0-1 classification error. However, when the training data distribution is changed into CLs, the loss is rewritten and the non-negativity of ℓ no longer remain, causing the negative risk term. That is, a ripple effect of error happens when the approximation error of the surrogate loss is amplified by the estimation step. In the proposed SCL framework, we sidestep this question by placing the surrogate process after the risk rewriting. In this way, the

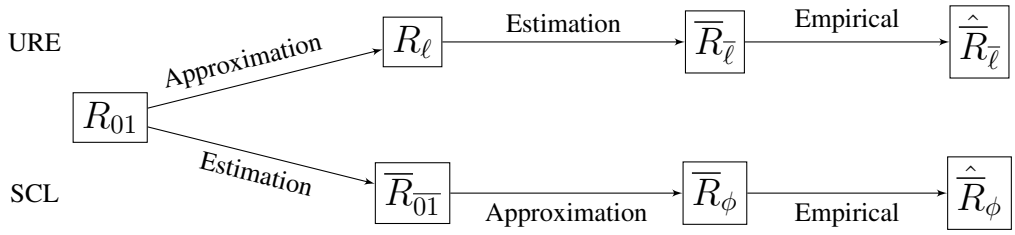


Figure 2. Comparison of URE learning process with the SCL framework.

Table 1. Classification accuracies

DATA SET + MODEL	URE	NN	GA	SCL-FWD	SCL-NL	SCL-EXP
MNIST + LINEAR	0.8503	0.8182	0.8193	0.9	0.9	**0.9019**
MNIST + MLP	0.8012	0.8665	0.9088	0.8965	**0.9469**	0.9251
KUZUSHI-MNIST + LINEAR	0.5613	0.5331	0.4992	0.6056	0.6056	**0.6132**
KUZUSHI-MNIST + MLP	0.5433	0.5683	0.6567	0.6445	**0.7644**	0.7184
FASHION-MNIST + LINEAR	0.7675	0.7755	0.7672	0.8274	0.8274	**0.8282**
FASHION-MNIST + MLP	0.7401	0.7829	0.8019	0.8372	**0.8456**	0.835
CIFAR-10 + RESNET	0.1091	0.3078	0.3738	**0.5058**	0.4713	0.492
CIFAR-10 + DENSENET	0.2909	0.3379	0.4108	**0.5457**	0.5394	0.5435

surrogate loss ϕ is directly applied on its target $\overline{\ell}_{01}$, and the negative loss problem is avoided. Furthermore, it is not only the statistical properties that matters to surrogate loss, optimization properties such as smoothness and curvature are also important to consider. As the estimation process of URE damages the original properties of ℓ, the optimization properties of ϕ are preserved.

3.3. Classification Accuracy

In this section, we use an experiment to compare the performance of each method. Specifically, the methods can be classified into two categories: URE-based methods, and SCL-based methods. In URE-based methods, we have URE, URE with negative risk correction (NN), and URE with gradient ascent (GA). In SCL-based methods, we have SCL-FWD, SCL-NL, and SCL-EXP. We used the Adam optimizer with learning rate selected from $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and trained the models for 300 epochs.

The testing accuracy is shown in Table 1. The URE performs poorly compared to other methods, especially in more flexible models. Even though NN and GA improve on URE in most tasks, the SCL methods still outperform them by a significant gap. These results justify our claims. Although URE is an estimation of the risk R_ℓ with statistical guarantees, in practice, it does not perform well as a classifier. On the other hand, although the proposed SCL framework is biased to the risk R_ℓ, introducing such bias towards minimizing the CL output yields superior results compared to URE, avoiding the negative risk issue. In the next section, we discuss why the difference between the two frameworks result in such a performance gap by analyzing the loss gradient during training.

4. Gradient Analysis

In this section, we discuss how the proposed SCL framework outperforms URE through two gradient analysis experiments. As mentioned in Section 2, the URE diverges widely from the risk itself when only a single CL is used to estimate the risk. Here we further discuss how the SCL framework gives such improvement by rearranging the learning process. The discussion will focus on the loss gradient: in the experiments, they are the stochastic gradient (SGD) in mini-batch optimization specifically. The analysis can be viewed as two parts: gradient directional estimation, and the bias-variance tradeoff of the gradient estimation error.

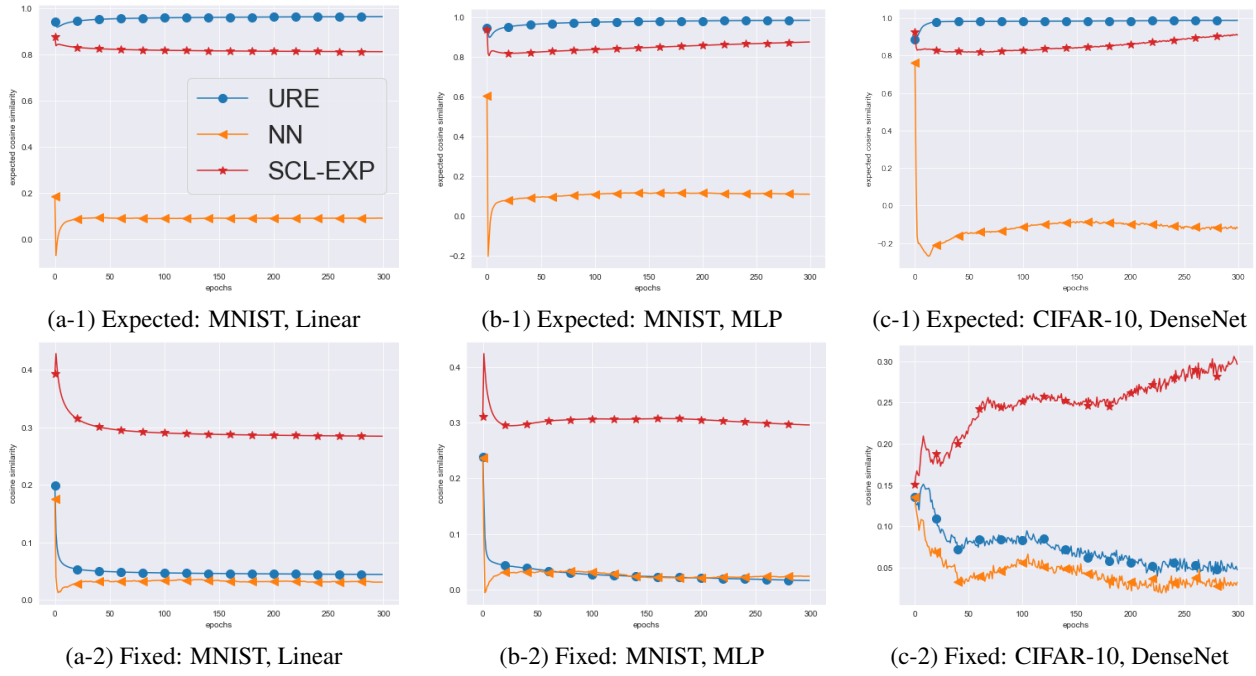
4.1. Directional Similarity

Since the URE is an estimator of the risk function, we expect its optimization to be similar to the risk function. Here we prove that the gradient of URE is also an unbiased gradient estimator (UGE) of the ordinary gradient.

Proposition 3. *The gradient of an unbiased risk estimator is unbiased to the ordinary risk gradient. That is, for an instance* (x, y) *we have,*

$$\mathbb{E}_{\overline{y}|y}\left[\nabla_\theta \overline{\ell}(\overline{y}, \boldsymbol{g}(x))\right] = \nabla_\theta \ell(y, \boldsymbol{g}(x)) \tag{14}$$

Thus, the gradient of the complementary loss $\overline{\ell}$ is unbiased with respect to the gradient of the ordinary loss, in our



(a-1) Expected: MNIST, Linear (b-1) Expected: MNIST, MLP (c-1) Expected: CIFAR-10, DenseNet

(a-2) Fixed: MNIST, Linear (b-2) Fixed: MNIST, MLP (c-2) Fixed: CIFAR-10, DenseNet

Figure 3. Cosine similarity comparison.

case the gradient of cross entropy loss. However, does that lead to similar performance with ordinary learning? Our experimental results show that URE methods learn poorly through unbiased gradient estimation.

In this section, we use an experiment to compare the gradient direction of ordinary learning and its complementary learning counterparts. We compare the complementary loss gradient directions with the ordinary gradient direction of the cross entropy loss $\nabla_\theta \ell(y, \boldsymbol{g}(x)) = -\nabla_\theta \log(\boldsymbol{p}_y)$.

The quality of the complementary gradient depends on its similarity with the ordinary gradient direction, where the similarity of two gradient directions is measured by the cosine similarity $\mathbb{S}$ of two gradient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, where $\mathbb{S}(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a} \cdot \boldsymbol{b})/|\boldsymbol{a}||\boldsymbol{b}|$. For the gradient directions, a reasonable assumption is $\mathbb{S}$ should be as large as possible, indicating a direction more similar to the ordinary gradient. In this experiment, we compare two gradient settings:

1. Expected: The averaged gradient computed over all possible CLs of an instance x.
2. Fixed: The gradient computed on a single CL of an instance x.

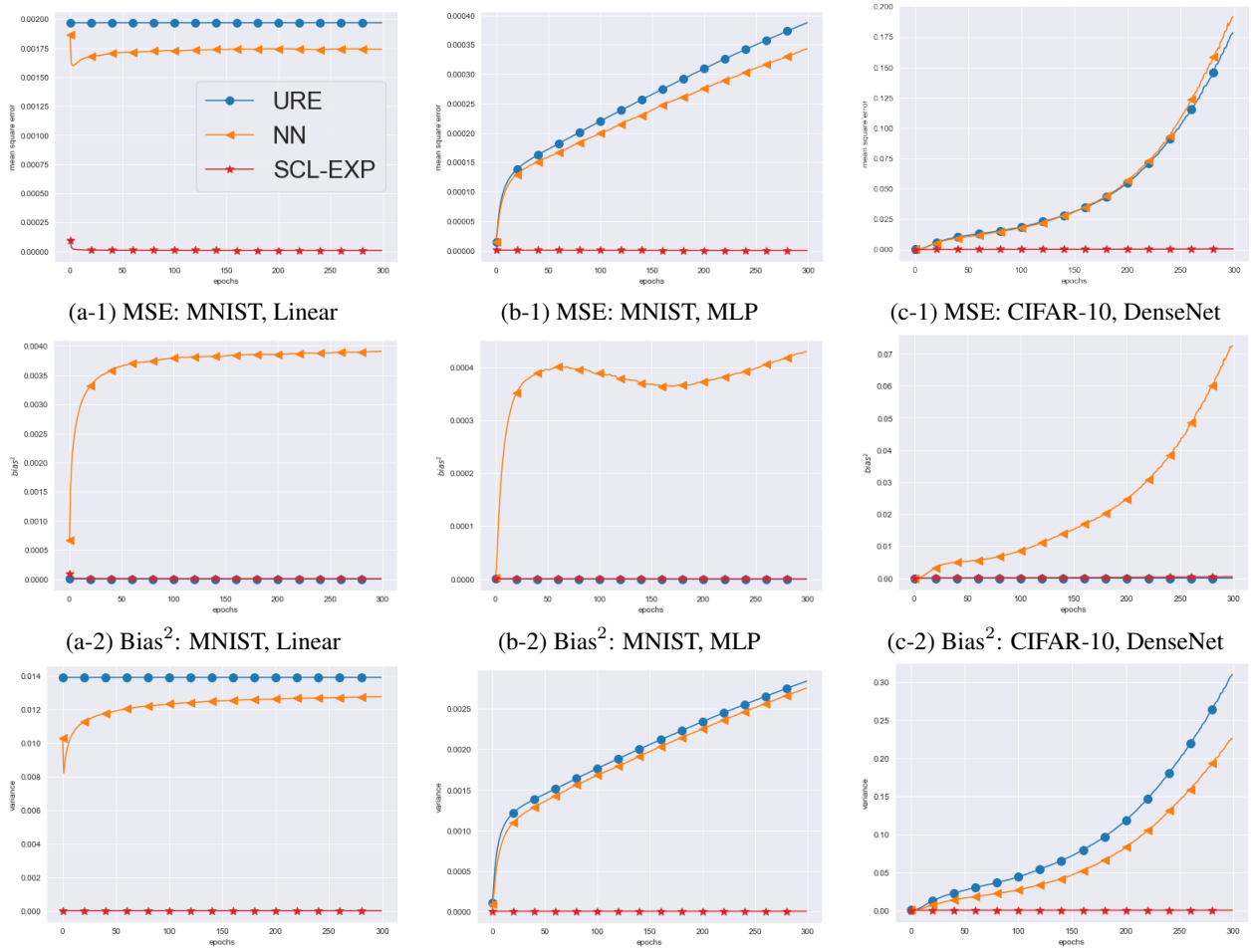
We compared three complementary learning methods on their approximation of the direction of the ordinary gradient: URE, NN, SCL-EXP. To ensure fair comparison, the model is updated only with ordinary labels in each epoch to avoid gradient error accumulation, the complementary gradients were computed only for comparison and were not updated to the model. The SGD optimizer was used with a learning rate fixed at 10^{-2}, trained for 300 epochs.

As the results show in Figure 3, URE achieves an ideal gradient direction only in the case of expected CLs. In the fixed case, UGE results in very different gradient directions with respect to the ordinary gradient direction. This shows that in the case when each x is fixed to a $\overline{y}$, UGE does not estimate a reliable direction. The UGEs of each $\overline{y}$ have diverged directions in order to maintain the unbiasedness. Unsurprisingly, the SCL methods provide better approximations of the ordinary gradient, since it does not diverge by focusing on the CL direction $\overline{\ell}_{01}$.

4.2. Bias-Variance Tradeoff

In this part, we further analyze the estimation error of the complementary gradient verses the ordinary gradient, using the bias-variance decomposition technique. Bias-variance decomposition is a common approach in statistical learning used to evaluate the complexity of a learning algorithm; instead of analyzing the error of a prediction problem, we extend this framework to evaluate the estimation error of the gradient, setting the ordinary gradient as the target. We will show that URE has much larger L_2 loss than SCL caused by its large variance, despite having no bias.

We denote $\boldsymbol{f}$ as the gradient step determined by ordinary labeled data (x, y) and ordinary loss ℓ. $\boldsymbol{c}$ denotes the complementary gradient step by complementary labeled data



(a-1) MSE: MNIST, Linear (b-1) MSE: MNIST, MLP (c-1) MSE: CIFAR-10, DenseNet

(a-2) Bias^2: MNIST, Linear (b-2) Bias^2: MNIST, MLP (c-2) Bias^2: CIFAR-10, DenseNet

(a-3) Variance: MNIST, Linear (b-3) Variance: MNIST, MLP (c-3) Variance: CIFAR-10, DenseNet

Figure 4. Error decomposition of gradient estimators.

$(x, \overline{y})$ and complementary loss $\overline{\ell}$ (or ϕ). $\boldsymbol{h}$ denotes the expected gradient step of $[K] \setminus \{y\}$, which is the average of $\boldsymbol{c}$ on every possible CL. We formalize as:

$$\boldsymbol{f} = \nabla \ell(y, \boldsymbol{g}(x)) \tag{15}$$

$$\boldsymbol{c} = \nabla \overline{\ell}(\overline{y}, \boldsymbol{g}(x)) \tag{16}$$

$$\boldsymbol{h} = \frac{1}{K-1} \sum_{y' \neq y} \nabla \overline{\ell}(y', \boldsymbol{g}(x)) \tag{17}$$

In this setting, we set $\boldsymbol{f}$ as the ground truth, which is the target for the complementary estimator $\boldsymbol{c}$. We hope the mean squared error(MSE) of gradient estimation to be small.

$$\text{MSE} = \mathbb{E}_{x,y,\overline{y}}\left[(\boldsymbol{f} - \boldsymbol{c})^2\right] \tag{18}$$

Here we can derive the bias-variance decomposition by introducing $\boldsymbol{h}$ and eliminating remaining terms:

$$\mathbb{E}\left[(\boldsymbol{f} - \boldsymbol{c})^2\right] = \mathbb{E}\left[(\boldsymbol{f} - \boldsymbol{h} + \boldsymbol{h} - \boldsymbol{c})^2\right] \tag{19}$$

$$= \underbrace{\mathbb{E}\left[(\boldsymbol{f} - \boldsymbol{h})^2\right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[(\boldsymbol{h} - \boldsymbol{c})^2\right]}_{\text{Variance}} \tag{20}$$

Since the UGE has no bias, it implies that all the estimation error of UGE comes from the variance term.

We run experiments to check how the complementary gradient $\boldsymbol{c}$ approximate the ordinary gradient $\boldsymbol{f}$, and compare with baseline methods. The training works as follows. In each epoch, we compute three gradients: the ordinary gradient $\boldsymbol{f}$, the current method $\boldsymbol{c}$, and $\boldsymbol{h}$. We measure the mean square error (MSE), the squared bias term and the variance term according to Equation 18 and Equation 20. In each epoch, we only update the model with $\boldsymbol{f}$ to maintain a fair comparison of the gradients. The optimizer is SGD with a learning rate fixed at 10^{-2}, trained for 300 epochs.

Results are showed in Figure 4 (mean statistics are shown

in Table 2 and Table 3), GA is omitted for visualization reasons. It is clear that although URE has no bias, it has very large MSE due to the large variance. On the other hand, the SCL methods though have little bias, have much smaller variance compared to URE. This justifies our claims in Section 4.1, the URE creates highly diverged gradients in order to maintain the overall unbiasedness, resulting high gradient variance. On the other hand, SCL introduces *inductive bias* towards minimizing the CL likelihood, trading zero bias with reduced variance.

5. Conclusion

In this paper, we show that unbiased risk estimator (URE) does not serve as a desirable optimization objective in weakly supervised learning problems such as learning with complementary labels. From the empirical risk aspect, the URE encounters the negative risk issue which leads to severe overfitting under weakly supervision. From the gradient aspect, the effort to maintain the unbiased gradient estimator (UGE) causes misleading direction and large variance to the loss gradient. We propose a new SCL learning framework based on the minimum likelihood principle and surrogate complementary loss functions. Though having a bias towards the CL, the SCL framework avoids the extremely noisy gradient problem encountered in URE. Empirical results show that SCL outperforms URE in classification accuracy and other gradient quality metrics.

Table 2. Gradient error decomposition of MNIST on linear model (Averaged over 300 epochs)

METHOD	MSE	BIAS^2	VARIANCE
URE	1.9692E-03	**8.0643E-07**	1.3907E-02
NN	1.7268E-03	3.7248E-03	1.2272E-02
GA	1.0436E+00	2.5829E+00	8.3596E+00
SCL-FWD	**7.7511E-06**	7.5037E-06	6.9942E-07
SCL-NL	**7.7511E-06**	7.5038E-06	6.9931E-07
SCL-EXP	7.9152E-06	7.7895E-06	**4.3945E-07**

Table 3. Gradient error decomposition of CIFAR-10 on DenseNet (Averaged over 300 epochs)

METHOD	MSE	BIAS^2	VARIANCE
URE	5.0196E-02	**6.6855E-06**	1.0101E-01
NN	5.2152E-02	2.1846E-02	7.0500E-02
GA	3.1350E+01	1.2985E+01	3.8302E+01
SCL-FWD	**2.0237E-04**	1.9225E-04	1.1051E-05
SCL-NL	**2.0237E-04**	1.9225E-04	1.1050E-05
SCL-EXP	2.0455E-04	1.9810E-04	**7.0735E-06**

Acknowledgements

GN and MS were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. YC and HL were partially supported by MOST 107-2628-E-002-008-MY3 and 108-2119-M-007-010.

References

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

Cao, Y. and Xu, Y. Multi-complementary and unlabeled learning for arbitrary losses and models. *CoRR*, abs/2001.04243, 2020. URL https://arxiv.org/abs/2001.04243.

Chapelle, O., Scholkopf, B., and Zien, A. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

du Plessis, M., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *ICML*, pp. 1386–1394, 2015.

du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *NeurIPS*, 2014.

Elkan, C. and Noto, K. Learning classifiers from only positive and unlabeled data. In *KDD*, 2008.

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

Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., and Sugiyama, M. Masking: A new perspective of noisy supervision. In *NeurIPS*, 2018a.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018b.

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

Hsieh, Y.-G., Niu, G., and Sugiyama, M. Classification from positive, unlabeled and biased negative data. In *ICML*, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017.

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

Ishida, T., Niu, G., and Sugiyama, M. Binary classification from positive-confidence data. In *NeurIPS*, 2018.

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

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *NeurIPS*, 2002.

Kaneko, T., Sato, I., and Sugiyama, M. Online multiclass classification based on prediction margin for partial feedback. *arXiv preprint arXiv:1902.01056*, 2019.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. 2015.

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

Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier from only unlabeled data. In *ICLR*, 2019.

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

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In *NeurIPS*, 2019.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *NeurIPS*, 2013.

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., and Sugiyama, M. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *NeurIPS*, 2016.

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

Sakai, T., du Plessis, M. C., Niu, G., and Sugiyama, M. Semi-supervised classification based on classification from positive and unlabeled data. In *ICML*, 2017.

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

Vapnik, V. Principles of risk minimization for learning theory. In *NeurIPS*, 1992.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, 2019.

Xu, Y., Gong, M., Chen, J., Liu, T., Zhang, K., and Batmanghelich, K. Generative-discriminative complementary learning. 2020.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *ECCV*, 2018.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., and Sugiyama, M. How does disagreement help generalization against label corruption? In *ICML*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.

Supplementary Material for Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels

A. Proofs

A.1. Proof of Proposition 1

Proof. Let $\boldsymbol{\eta}$ and $\overline{\boldsymbol{\eta}}$ denote the conditional distribution $\mathbb{P}(Y \mid X)$ and $\mathbb{P}(\overline{Y} \mid X)$ respectively, where $\boldsymbol{\eta}_k(x) = \mathbb{P}(Y = k \mid x)$ and $\overline{\boldsymbol{\eta}}_k(x) = \mathbb{P}(\overline{Y} = k \mid x)$. Since $\bar{y}$ only depends on y, we have $\bar{\boldsymbol{\eta}}(x) = T^\top \boldsymbol{\eta}(x)$. The unbiased risk estimator can be derived as follows:

$$\begin{aligned} R(\boldsymbol{g}; \ell) &= \mathbb{E}_{(x,y)\sim D}[\ell(y, \boldsymbol{g}(x))] = \mathbb{E}_X \mathbb{E}_{Y\sim\boldsymbol{\eta}(X)}[\ell(Y, \boldsymbol{g}(X))] \\ &= \mathbb{E}_X[\boldsymbol{\eta}(X)^\top \ell(\boldsymbol{g}(X))] = \mathbb{E}_X[\overline{\boldsymbol{\eta}}(X)^\top (T^{-1}) \ell(\boldsymbol{g}(X))] \\ &= \mathbb{E}_{(x,\overline{y})\sim\overline{D}}[e_{\overline{y}}^\top (T^{-1}) \ell(\boldsymbol{g}(x))] \end{aligned}$$

□

A.2. Proof of Proposition 2

Proof. Given the following two properties of ℓ_{01}:

$$\begin{aligned} &\sum_{i=1}^{K} \ell_{01}(i, \boldsymbol{g}(x)) = K - 1 \quad \text{and} \\ &\ell_{01}(\overline{y}, \boldsymbol{g}(x)) + \overline{\ell}_{01}(\overline{y}, \boldsymbol{g}(x)) = 1 \end{aligned}$$

An unbiased risk estimator of classification error can be obtained by:

$$\begin{aligned} R(\boldsymbol{g}; \ell_{01}) &= \mathbb{E}_{(x,\overline{y})\sim\overline{D}}\Big[-(K-1)\ell_{01}(\overline{y}, \boldsymbol{g}(x)) + \sum_{j=1}^{K} \ell_{01}(j, \boldsymbol{g}(x))\Big] \\ &= \mathbb{E}_{(x,\overline{y})\sim\overline{D}}\Big[(K-1)(1 - \ell_{01}(\overline{y}, \boldsymbol{g}(x)))\rrbracket\Big] \\ &= (K-1)\mathbb{E}_{(x,\overline{y})\sim\overline{D}}\Big[\overline{\ell}_{01}(\overline{y}, \boldsymbol{g}(x))\Big] = (K-1)\overline{R}(\boldsymbol{g}; \overline{\ell}_{01}) \end{aligned}$$

□

A.3. Proof of Proposition 3

Proof. The proposition can be derived by using the linearity of the gradient operator:

$$\begin{aligned} \mathbb{E}_{\overline{y}|y}\big[\nabla_\theta \overline{\ell}(\overline{y}, \boldsymbol{g}(x))\big] &= \nabla_\theta \mathbb{E}_{\overline{y}|y}\big[\overline{\ell}(\overline{y}, \boldsymbol{g}(x))\big] \\ &= \nabla_\theta \Big[\frac{1}{K-1} \sum_{y'\neq y} \big[-(K-1)\ell(y', \boldsymbol{g}(x)) + \sum_{j=1}^{K} \ell(j, \boldsymbol{g}(x))\big]\Big] \\ &= \nabla_\theta \Big[-\sum_{y'\neq y} \ell(y', \boldsymbol{g}(x)) + \sum_{j=1}^{K} \ell(j, \boldsymbol{g}(x))\Big] = \nabla_\theta \ell(y, \boldsymbol{g}(x)) \end{aligned}$$

□