Mitigating Overfitting in Supervised Classification from Two Unlabeled Datasets: A Consistent Risk Correction Approach

Nan Lu[1,2] **Tianyi Zhang**[1] **Gang Niu**[2] **Masashi Sugiyama**[2,1]
[1]The University of Tokyo, Japan [2]RIKEN, Japan

Abstract

The recently proposed unlabeled-unlabeled (UU) classification method allows us to train a binary classifier only from two *unlabeled* datasets with different class priors. Since this method is based on the empirical risk minimization, it works as if it is a *supervised* classification method, compatible with any model and optimizer. However, this method sometimes suffers from severe overfitting, which we would like to prevent in this paper. Our empirical finding in applying the original UU method is that overfitting often co-occurs with the empirical risk *going negative*, which is not legitimate. Therefore, we propose to *wrap* the terms that cause a negative empirical risk by certain *correction functions*. Then, we prove the consistency of the corrected risk estimator and derive an estimation error bound for the corrected risk minimizer. Experiments show that our proposal can successfully mitigate overfitting of the UU method and significantly improve the classification accuracy.

1 Introduction

In traditional supervised classification, we always assume a vast amount of labeled data in the training phase. However, labeling industrial-level data can be expensive and time-consuming due to laborious manual annotations. Furthermore, in some real-world problems such as medical diagnosis (Li and Zhou, 2007; Fakoor et al., 2013; Sun et al., 2017), massive labeled data may not even be possible to collect. This has led to the development of machine learning algorithms to leverage large-scale unlabeled (U) data, including but not limited to semi-supervised learning (Grandvalet and Bengio, 2004; Mann and McCallum, 2007; Niu et al., 2013; Miyato et al., 2016; Laine and Aila, 2017; Luo et al., 2018; Oliver et al., 2018) and positive-unlabeled learning (Elkan and Noto, 2008; du Plessis et al., 2014, 2015; Niu et al., 2016; Kiryo et al., 2017; Kato et al., 2019).

In this paper, we consider a more challenging setting of learning from only U data. A naïve approach to this problem is to use *discriminative clustering* (Xu et al., 2004; Valizadegan and Jin, 2006; Gomes et al., 2010; Sugiyama et al., 2014; Hu et al., 2017), which is also known as *unsupervised classification*. But this solution is usually suboptimal due to the tacit *clustering assumption* that *one cluster corresponds to one class* (Chapelle et al., 2002), which is often violated in practice. For example, when one cluster is formed by a few geometrically close classes, or one class is formed by several geometrically separated clusters, even perfect clustering may still result in poor classification.

In order to avoid the unrealistic clustering assumption, we prefer to utilize U data for *risk evaluation* and then optimize the obtained risk estimator by *empirical risk minimization* (ERM), as what has been carried out in standard supervised classification methods. This line of research was pioneered by du Plessis et al. (2013) and Menon et al. (2015), where a binary classifier is trained from *two sets of U data with different class priors*. However, a critical limitation is that their performance measure has to be the *balanced error* (Brodersen et al., 2010), which is the classification accuracy when the class prior is 1/2. Recently, Lu et al. (2019) extended these works and developed the first ERM method based on unbiased risk estimators for learning from two sets of U data. Their method, called *UU classification* achieved the state-of-the-art performance in experiments.

However, we found that, depending on the situation, the state-of-the-art unbiased UU method still suffers from severe overfitting as demonstrated in Figure 1. Based on our empirical explorations to this problem,

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

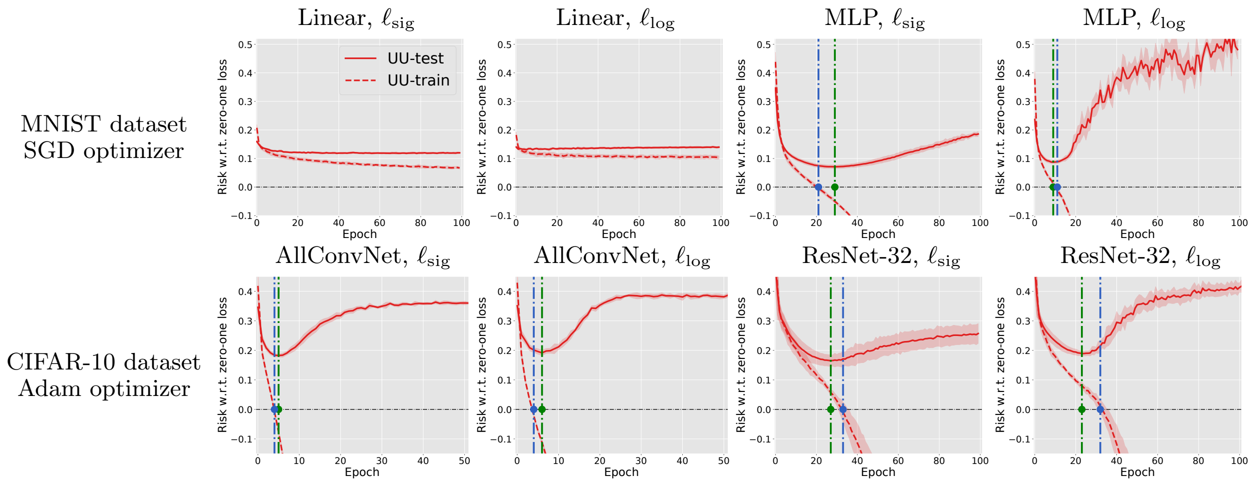

Figure 1: Co-occurrence between severe overfitting and a negative empirical risk in UU classification. By co-occurrence, we mean both of them can be observed, or neither of them can be observed. In the upper row, on MNIST (even vs. odd), a linear-in-input model (Linear) and a 5-layer *multi-layer perceptron* (MLP) were trained by stochastic gradient descent (SGD) using the sigmoid (ℓ_{sig}) and logistic (ℓ_{log}) losses. In the bottom row, on CIFAR-10 (transportation vs. animal), the *all convolutional net* (AllConvNet) (Springenberg et al., 2015) and the 32-layer *residual network* (ResNet-32) (He et al., 2016) were trained by Adam (Kingma and Ba, 2015) using the same losses. The class priors θ and θ' were set to be 0.6 and 0.4 (see Sec. 2 for details). The blue dashed lines indicate when the empirical risk computed from UU training data goes negative; the green dashed lines indicate when the test error turns around and severe overfitting begins. We can clearly see a high co-occurrence in the figure regardless of datasets, optimizers, models, and losses. Details of how to reproduce the figure can be found in Appendix B.

we conjecture that the overfitting issue of the unbiased UU method is strongly connected to the empirical risk on training data going *negative* due to the co-occurrence of them regardless of datasets, models, optimizers and loss functions. This negative empirical training risk should be fixed since the empirical training risk in standard supervised classification is always non-negative as long as the loss function is non-negative, which might be a potential reason for the unbiased risk estimator based method to overfit. [1]

In this paper, we focus on mitigating this overfitting, where our goal is to learn a robust binary classifier from two U sets with different class priors following the ERM principle. To this end, we propose a novel *consistent risk correction* technique that follows and improves the state-of-the-art unbiased UU method. The proposed method has the following advantages:

- Empirically, the proposed corrected risk estimators are robust against overfitting. Theoretically, they are asymptotically unbiased and thus may be properly used for hyparameter tuning with only UU data, which is a clear advantage in deep learning since no labeled validation data are needed. Furthermore, the corrected minimizers possess an *estimation error bound* which guarantees the *consistency of learning* (Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014b);
- We do not have implicit assumptions on the loss function, model architecture, and optimization, thus allowing the use of any loss (convex and non-convex), any model (e.g., the linear-in-parameter model and deep neural network) and any off-the-shelf *stochastic optimization algorithms* (e.g., Duchi et al., 2011; Kingma and Ba, 2015).

Organization The rest of the paper is organized as follows. We formalize our research problem in Sec. 2. In Sec. 3, we propose the consistent risk correction with theoretical analysis. Experimental results are discussed in Sec. 4, and conclusions are given in Sec. 5. Proofs are presented in the supplementary material.

2 Preliminaries

In this section, we introduce some notations and review the formulations of standard supervised classification and learning from two sets of U data with different class priors.

2.1 Learning from fully labeled data

We begin with the standard supervised classification setup. Let $\mathcal{X}$ be the example space and $\mathcal{Y} = \{+1, -1\}$

[1] Note that the general-purpose regularization techniques, such as weight decay and dropout, fail to mitigate this overfitting as illustrated and analyzed in Appendix D.

be a binary label space. Denote by $\mathcal{D}$ the *underlying joint distribution* over $\mathcal{X} \times \mathcal{Y}$. Any $\mathcal{D}$ may be decomposed into *class-conditional distributions* $(P_\mathrm{p}, P_\mathrm{n}) = (p(x \mid y = +1), p(x \mid y = -1))$ and *class-prior probability* $\pi_\mathrm{p} = p(y = +1)$.

Let $g : \mathcal{X} \to \mathbb{R}$ be an arbitrary binary classifier and $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ be a *loss function*, such that the value $\ell(t, y)$ means the loss for predicting the ground truth label y by t. We assume that the loss function is non-negative. Denote by $R_\mathrm{p}^+(g) = \mathbb{E}_{x \sim P_\mathrm{p}}[\ell(g(x), +1)]$ and $R_\mathrm{n}^-(g) = \mathbb{E}_{x \sim P_\mathrm{n}}[\ell(g(x), -1)]$. The goal of binary classification is to obtain a classifier g which minimizes the risk defined as

$$R(g) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(g(x), y)] = \pi_\mathrm{p} R_\mathrm{p}^+(g) + \pi_\mathrm{n} R_\mathrm{n}^-(g), \tag{1}$$

where $\mathbb{E}_{(x,y)\sim\mathcal{D}}$ denotes the expectation over $\mathcal{D}$, and $\pi_\mathrm{n} = p(y = -1) = 1 - \pi_\mathrm{p}$. If ℓ is the *zero-one loss* defined by $\ell_{01}(t, y) = (1 - \mathrm{sign}(ty))/2$, the risk is named the *classification error* (or the *misclassification rate*), which is the standard performance measure in classification (Mohri et al., 2012).

Since the joint distribution $\mathcal{D}$ is unknown, the ordinary ERM approach approximates the expectation by the average over training samples drawn i.i.d. from $\mathcal{D}$ (Vapnik, 1998). More specifically, given $\mathcal{X}_\mathrm{p} = \{x_1^+, \ldots, x_{n_\mathrm{p}}^+\} \overset{\text{i.i.d.}}{\sim} P_\mathrm{p}$ and $\mathcal{X}_\mathrm{n} = \{x_1^-, \ldots, x_{n_\mathrm{n}}^-\} \overset{\text{i.i.d.}}{\sim} P_\mathrm{n}$, $R(g)$ can be approximated by

$$\widehat{R}_\mathrm{pn}(g) = \pi_\mathrm{p} \widehat{R}_\mathrm{p}^+(g) + \pi_\mathrm{n} \widehat{R}_\mathrm{n}^-(g), \tag{2}$$

where $\widehat{R}_\mathrm{p}^+(g) = (1/n_\mathrm{p}) \sum_{i=1}^{n_\mathrm{p}} \ell(g(x_i^+), +1)$ and $\widehat{R}_\mathrm{n}^-(g) = (1/n_\mathrm{n}) \sum_{i=1}^{n_\mathrm{n}} \ell(g(x_i^-), -1)$.

2.2 Learning from two sets of U data with different class priors

Next we consider the problem of learning from two sets of U data with different class priors, which is called *unlabeled-unlabeled* (UU) *classification* in Lu et al. (2019). We are given only unlabeled samples drawn from the following marginal distributions:

$$\begin{aligned} p_\mathrm{tr}(x) &= \theta P_\mathrm{p} + (1 - \theta) P_\mathrm{n}, \\ p'_\mathrm{tr}(x) &= \theta' P_\mathrm{p} + (1 - \theta') P_\mathrm{n}, \end{aligned} \tag{3}$$

where θ and θ' are two class priors such that $\theta \neq \theta'$. This implies there are $p_\mathrm{tr}(x, y)$ and $p'_\mathrm{tr}(x, y)$, whose class-conditional densities are same and equal to those of $\mathcal{D}$, but whose class priors are different, i.e.,

$$\begin{gathered} p_\mathrm{tr}(x \mid y) = p'_\mathrm{tr}(x \mid y) = p(x \mid y), \\ p_\mathrm{tr}(y = +1) = \theta \neq \theta' = p'_\mathrm{tr}(y = +1). \end{gathered}$$

More specifically, we have $\mathcal{X}_\mathrm{tr} = \{x_1, \ldots, x_n\} \overset{\text{i.i.d.}}{\sim} p_\mathrm{tr}(x)$ and $\mathcal{X}'_\mathrm{tr} = \{x'_1, \ldots, x'_{n'}\} \overset{\text{i.i.d.}}{\sim} p'_\mathrm{tr}(x)$, and our goal is to train a binary classifier that can generalize well with respect to the original $\mathcal{D}$.

In the standard supervised classification setting where training data are directly drawn from $\mathcal{D}$, the expectation in (1) can be estimated by the corresponding sample average. However, in the UU classification setting, no labeled samples are available and therefore the risk may not be estimated directly.

This problem can be avoided by the *risk rewriting* approach (Lu et al., 2019; van Rooyen and Williamson, 2018): the risk (1) is firstly rewritten into an equivalent expression such that it only involves the same distributions from which two sets of U data are sampled, and then it is estimated by plugging in the given U data. Let $R_\mathrm{u}^+(g) = \mathbb{E}_{x \sim p_\mathrm{tr}}[\ell(g(x), +1)]$, $R_\mathrm{u}^-(g) = \mathbb{E}_{x \sim p_\mathrm{tr}}[\ell(g(x), -1)]$, $R_{\mathrm{u}'}^+(g) = \mathbb{E}_{x \sim p'_\mathrm{tr}}[\ell(g(x'), +1)]$ and $R_{\mathrm{u}'}^-(g) = \mathbb{E}_{x \sim p'_\mathrm{tr}} \ell(g(x'), -1)]$. $R(g)$ can be expressed by

$$R(g) = a R_\mathrm{u}^+(g) - b R_\mathrm{u}^-(g) - c R_{\mathrm{u}'}^+(g) + d R_{\mathrm{u}'}^-(g),$$

where $a = \frac{(1-\theta')\pi_\mathrm{p}}{\theta - \theta'}$, $b = \frac{\theta'(1-\pi_\mathrm{p})}{\theta - \theta'}$, $c = \frac{(1-\theta)\pi_\mathrm{p}}{\theta - \theta'}$, and $d = \frac{\theta(1-\pi_\mathrm{p})}{\theta - \theta'}$. Then with empirical estimates $\widehat{R}_\mathrm{u}^+(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), +1)$, $\widehat{R}_\mathrm{u}^-(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), -1)$, $\widehat{R}_{\mathrm{u}'}^+(g) = \frac{1}{n'} \sum_{j=1}^{n'} \ell(g(x'_j), +1)$, and $\widehat{R}_{\mathrm{u}'}^-(g) = \frac{1}{n'} \sum_{j=1}^{n'} \ell(g(x'_j), -1)$, $R(g)$ can be approximated as

$$\widehat{R}_\mathrm{uu}(g) = a \widehat{R}_\mathrm{u}^+(g) - b \widehat{R}_\mathrm{u}^-(g) - c \widehat{R}_{\mathrm{u}'}^+(g) + d \widehat{R}_{\mathrm{u}'}^-(g). \tag{4}$$

The *empirical risk estimators* in Eqs. (2) and (4) are *unbiased* and *consistent*[2] w.r.t. all loss functions. When they are used for evaluating the classification accuracy, ℓ is by default ℓ_{01}; when they are used for training, it is replaced with a *surrogate loss* since ℓ_{01} is discontinuous and therefore difficult to optimize (Ben-David et al., 2003; Bartlett et al., 2006).

The unbiased risk estimator methods (Lu et al., 2019; van Rooyen and Williamson, 2018) use classification error (1) as the performance measure and assume the knowledge of class priors. Note that given only U data, by no means could we learn the class priors without any assumptions (Menon et al., 2015). But, by introducing the *mutually irreducible condition* (Scott et al., 2013), the class priors become identifiable and can be estimated in some cases (Menon et al., 2015; Liu and Tao, 2016; Jain et al., 2016; Blanchard et al., 2016).

[2] The consistency here means for fixed g, $\widehat{R}_\mathrm{pn}(g) \to R(g)$ and $\widehat{R}_\mathrm{uu}(g) \to R(g)$ as $n_\mathrm{p}, n_\mathrm{n}, n, n' \to \infty$.

To simplify analysis, we assume the class priors to be known in this paper.

Another line of research on UU classification focuses on the *balanced error* (BER), which is a special case of the classification error (1), defined by $B(g) = \frac{1}{2}\mathbb{E}_{x\sim P_\mathrm{p}}[\ell_{01}(g(x), +1)] + \frac{1}{2}\mathbb{E}_{x\sim P_\mathrm{n}}[\ell_{01}(g(x), -1)]$. Though BER minimization methods do not need the knowledge of class priors, they assume that the class prior is balanced (i.e., $\pi_\mathrm{p} = \frac{1}{2}$) (du Plessis et al., 2013; Menon et al., 2015; Charoenphakdee et al., 2019). Note that $B(g) = R(g)$ for any g if and only if $\pi_\mathrm{p} = \frac{1}{2}$, which indicates that BER is a meaningful performance measure for classification when $\pi_\mathrm{p} \approx \frac{1}{2}$ while it definitely biases learning when $\pi_\mathrm{p} \approx \frac{1}{2}$ is not the case. Therefore, through out this paper, we consider the more natural classification error metric (1).

3 Consistent Risk Correction

In this section, we first study the overfitting issue of the unbiased risk estimator of UU classification, and then propose our consistent risk correction method with theoretical guarantees.

3.1 Is an unbiased risk estimator really good?

As discussed in Sec. 2.2, the state-of-the-art method of UU classification uses the risk rewriting technique to obtain an unbiased risk estimator. However, the derived unbiased UU risk estimator (4) contains two negative partial risks $-b\widehat{R}^-_\mathrm{u}(g)$ and $-c\widehat{R}^+_{\mathrm{u}'}(g)$. This may be problematic since the original expression of the classification risk (1) only includes expectations over non-negative loss $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ and is by definition non-negative. In practice, we find that the unbiased UU method may suffer severe overfitting and observe a high co-occurrence between overfitting and their empirical risk going negative. Thus, we conjecture that the negative empirical risk might be a potential reason that results in overfitting.

We elaborate on the issue in Figure 1, where we trained various models on MNIST and CIFAR-10 using different optimizers and loss functions. From the experimental results, we can see a strong co-occurrence of severe overfitting and a negative empirical risk regardless of datasets, models, optimizers, and loss functions: in the experiments of MNIST and CIFAR-10 with different deep neural network models, optimizers, and loss functions, overfitting is observable when the empirical risk on the training data goes negative; in the experiments on MNIST with the linear model and SGD optimizer, the test performance is reasonably good while the empirical risk on the training data is kept non-negative. The overfitting is more severe when flexible models such as deep neural networks are used, since they have larger capacity to fit data and thus they make the negative partial risks $-b\widehat{R}^-_\mathrm{u}(g)$ and $-c\widehat{R}^+_{\mathrm{u}'}(g)$ more negative.

3.2 Corrected risk estimator

Now we face a dilemma: in many real-world problems, we may only collect large unlabeled datasets and still wish our classifier trained from them generalize well. So the question arises: can we alleviate the aforementioned overfitting problem with neither labeling more training data nor turning to a suboptimal solution (e.g., clustering)?

The answer is affirmative. In Figure 1, we observed that the resulting empirical risk $\widehat{R}_\mathrm{uu}(g)$ keeps decreasing and goes negative. This issue should be fixed since the empirical training risk in standard supervised classification is always non-negative for non-negative loss functions. Note that the two terms (i.e., $R^+_\mathrm{p}(g)$ and $R^-_\mathrm{n}(g)$) in the original classification risk (1), which correspond to the risks of the P and N classes, are both non-negative. Thus our basic idea is reformulating the rewritten risk (4) to find the counterparts for the risks of the P and N classes in (1):

$$\begin{aligned}\pi_\mathrm{p} R^+_\mathrm{p}(g) &= aR^+_\mathrm{u}(g) - cR^+_{\mathrm{u}'}(g),\\ \pi_\mathrm{n} R^-_\mathrm{n}(g) &= dR^-_{\mathrm{u}'}(g) - bR^-_\mathrm{u}(g).\end{aligned}$$

We then enforce non-negativity to these counterparts. More specifically, we have

$$\begin{aligned}\widehat{R}_\text{uu-max}(g) = \max\left\{0, a\widehat{R}^+_\mathrm{u}(g) - c\widehat{R}^+_{\mathrm{u}'}(g)\right\}&\\ + \max\left\{0, d\widehat{R}^-_{\mathrm{u}'}(g) - b\widehat{R}^-_\mathrm{u}(g)\right\}.&\end{aligned} \tag{5}$$

This is motivated by Kiryo et al. (2017), which considered the problem of the rewritten risk going negative in the context of positive-unlabeled learning. In their setting, the reformulated P risk is exactly the same as its counterpart in the original classification risk (1) (i.e., $R^+_\mathrm{p}(g)$), since they are given positive data with true labels. So there was only one max operator in the reformulated N risk. Our setting differs from them since we are given only unlabeled data and therefore needs the "max" correction for both the reformulated P and N risks.

However, the max operator completely ignores the training data that yield a negative risk. We argue that the information in those data is also useful for training and should not be dropped. Following this idea, we propose a generalized *consistent correction function* as follows:

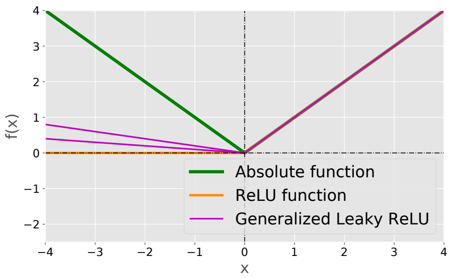

Figure 2: Examples of consistent correction functions.

Definition 1 (Consistent correction function)**.** *A function $f : \mathbb{R} \to \mathbb{R}$ is called a consistent correction function if it is Lipschitz continuous, non-negative and $f(x) = x$ for all $x \geq 0$. Let $\mathcal{F}$ be a class of all consistent correction functions.*

For example, the rectified linear unit (ReLU) function and absolute value function belong to $\mathcal{F}$. Based on this definition, we propose a family of consistently corrected risk estimators $\widehat{R}_{\mathrm{cc}}$ by

$$\widehat{R}_{\mathrm{cc}}(g) = f_1\left(a\widehat{R}^+_{\mathrm{u}}(g) - c\widehat{R}^+_{\mathrm{u}'}(g)\right) + f_2\left(d\widehat{R}^-_{\mathrm{u}'}(g) - b\widehat{R}^-_{\mathrm{u}}(g)\right), \tag{6}$$

where f_1 and f_2 can be any consistent correction funtions. The proposed corrected risk estimator is by nature ERM-based, and consequently the empirical risk minimizer of (6), i.e., $\widehat{g}_{\mathrm{cc}} = \arg\min_{g\in\mathcal{G}} \widehat{R}_{\mathrm{cc}}(g)$ can be obtained by flexible models and powerful stochastic optimization algorithms.

The corrected UU classification algorithm is described in Algorithm 1. In the implementation, we propose to use the generalized leaky ReLU function, i.e., $f(x) = f_1(x) = f_2(x) = \mathbb{I}_{\{x\geq 0\}}x + \mathbb{I}_{\{x<0\}}\lambda x$, where $\lambda \leq 0$. The intuition behind is that instead of completely ignoring the training data that yield a negative risk by the "max" correction, we propose to actively control learning on those sensitive data by adding weights on the negative partial risks. Note that the ReLU function and the absolute value function are special cases of the generalized leaky ReLU function as illustrated in Figure 2.[3]

3.3 Theoretical analysis

In this section, we analyze the consistently corrected risk estimator (6) and its minimizer.

Algorithm 1 Corrected UU classification

```
Input: two sets of U training data $(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}})$
Output: learned model parameter $\theta$
Initialize $\theta$
Let $\mathcal{A}$ be an SGD-like optimizer working on $\theta$
for $t = 1$ to number_of_epochs:
    Shuffle $(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}})$
    for $i = 1$ to number_of_mini-batches:
        Let $(\overline{\mathcal{X}}_{\mathrm{tr}}, \overline{\mathcal{X}}'_{\mathrm{tr}})$ be the current mini-batch
        Forward $\overline{\mathcal{X}}_{\mathrm{tr}}$ and $\overline{\mathcal{X}}'_{\mathrm{tr}}$
        Compute
        $L^+ = a\ell(g(\overline{\mathcal{X}}_{\mathrm{tr}}), +1)/|\overline{\mathcal{X}}_{\mathrm{tr}}| - c\ell(g(\overline{\mathcal{X}}'_{\mathrm{tr}}), +1)/|\overline{\mathcal{X}}'_{\mathrm{tr}}|$
        $L^- = d\ell(g(\overline{\mathcal{X}}'_{\mathrm{tr}}), -1)/|\overline{\mathcal{X}}'_{\mathrm{tr}}| - b\ell(g(\overline{\mathcal{X}}_{\mathrm{tr}}), -1)/|\overline{\mathcal{X}}_{\mathrm{tr}}|$
        Correct them by $L^+_{\mathrm{cc}} = f(L^+), L^-_{\mathrm{cc}} = f(L^-)$
        Backward $L_{\mathrm{cc}} = L^+_{\mathrm{cc}} + L^-_{\mathrm{cc}}$
        Update $\theta$ by $\mathcal{A}$
```

Bias and consistency The proposed corrected risk estimator $\widehat{R}_{\mathrm{cc}}(g)$ is no longer unbiased due to the fact that $\widehat{R}_{\mathrm{uu}}(g)$ is unbiased and $\widehat{R}_{\mathrm{cc}}(g) \geq \widehat{R}_{\mathrm{uu}}(g)$ for any $(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}})$ if we fix g. The question then arises: is $\widehat{R}_{\mathrm{cc}}(g)$ consistent? Next we prove the consistency.

First, let $A = a\widehat{R}^+_{\mathrm{u}}(g)$, $B = b\widehat{R}^-_{\mathrm{u}}(g)$, $C = c\widehat{R}^+_{\mathrm{u}'}(g)$, $D = d\widehat{R}^-_{\mathrm{u}'}(g)$ and L_f be the Lipschitz constant of f_1 and f_2. Then partition all possible $(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}})$ into: $\mathfrak{D}^+(g) = \{(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}}) \mid A - C \geq 0, D - B \geq 0\}$, $\mathfrak{D}^-(g) = \{(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}}) \mid A - C < 0\} \cup \{(\mathcal{X}_{\mathrm{tr}}, \mathcal{X}'_{\mathrm{tr}}) \mid D - B < 0\}$. Assume there are $C_g > 0$ and $C_\ell > 0$ such that $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq C_g$ and $\sup_{|z|\leq C_g} \ell(z) \leq C_\ell$. By *McDiarmid's inequality* (McDiarmid, 1989), we can prove the following lemma.

Lemma 2. *The bias of $\widehat{R}_{\mathrm{cc}}(g)$ is positive if and only if the probability measure of $\mathfrak{D}^-(g)$[4] is non-zero. Further, by assuming that there is $\alpha_g > 0$ and $\beta_g > 0$ such that $R^+_{\mathrm{p}}(g) \geq \alpha_g/\pi_{\mathrm{p}}$ and $R^-_{\mathrm{n}}(g) \geq \beta_g/\pi_{\mathrm{n}}$, the probability measure of $\mathfrak{D}^-(g)$ can be bounded by*

$$\Pr(\mathfrak{D}^-(g)) \leq \exp\left(-\frac{2\alpha_g^2/C_\ell^2}{a^2/n + c^2/n'}\right) + \exp\left(-\frac{2\beta_g^2/C_\ell^2}{b^2/n' + d^2/n}\right). \tag{7}$$

Based on Lemma 2, we can show the exponential decay of the bias and also the consistency.

Theorem 3 (Bias and consistency)**.** *Let $\Delta_g = \exp\left(-\frac{2\alpha_g^2/C_\ell^2}{a^2/n+c^2/n'}\right) + \exp\left(-\frac{2\beta_g^2/C_\ell^2}{b^2/n'+d^2/n}\right)$. By assumption in Lemma 2, the bias of $\widehat{R}_{\mathrm{cc}}(g)$ decays exponen-*

[3] Using the ReLU and absolute value function to prevent the negative risk problem has been studied in the context of positive-unlabeled learning (Kiryo et al., 2017) and complementary-label learning (Ishida et al., 2019). Our proposal can be regarded as their extension to a family of correction functions applied in the UU classification setting.

[4] The probability measure is induced by the randomness of the two unlabeled datasets, see Appendix A.1 for the formal definition.

Table 1: Specification of benchmark datasets and models.

Dataset	# Train	# Test	# Feature	π_{p}	Simple $g(x)$	Deep $g(x)$
MNIST	60,000	10,000	784	0.50	Linear model	5-layer MLP
Fashion-MNIST	60,000	10,000	784	0.40	Linear model	5-layer MLP
Kuzushiji-MNIST	60,000	10,000	784	0.30	Linear model	5-layer MLP
CIFAR-10	50,000	10,000	3,072	0.60	Linear model	ResNet-32

tially as $n, n' \to \infty$:

$$\begin{aligned} 0 &\le \mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}}}[\widehat{R}_{\mathrm{cc}}(g)] - R(g) \\ &\le (L_f+1)(a+b+c+d)C_\ell\Delta_g. \end{aligned} \tag{8}$$

Moreover, for any $\delta > 0$, let $C_\delta = C_\ell L_f\sqrt{\ln(2/\delta)/2}$, $\chi_{n,n'} = (a+b)/\sqrt{n} + (c+d)/\sqrt{n'}$, then we have with probability at least $1-\delta$,

$$\begin{aligned} |\widehat{R}_{\mathrm{cc}}(g) - R(g)| \le (L_f+1)(a+b+c+d)C_\ell\Delta_g \\ + C_\delta \cdot \chi_{n,n'}, \end{aligned} \tag{9}$$

and with probability at least $1-\delta-\Delta_g$,

$$|\widehat{R}_{\mathrm{cc}}(g) - R(g)| \le C_\delta \cdot \chi_{n,n'}. \tag{10}$$

Either (9) or (10) in Theorem 3 indicates for fixed g, $\widehat{R}_{\mathrm{cc}}(g) \to R(g)$ in $\mathcal{O}_p(1/\sqrt{n} + 1/\sqrt{n'})$. This convergence rate is optimal according to the *central limit theorem* (Chung, 1968), which means the proposed estimator is a biased yet optimal estimator to the risk.

Estimation error bound While Theorem 3 addressed the use of (6) when the risk is evaluated, in what follows we study the estimation error $R(\widehat{g}_{\mathrm{cc}}) - R(g^*)$ when classifiers are trained, where g^* is the true risk minimizer in the model class $\mathcal{G}$, i.e., $g^* = \arg\min_{g\in\mathcal{G}} R(g)$. As a common practice (Mohri et al., 2012; Boucheron et al., 2005), assume that the instances are upper bounded, i.e., $\|x\| \le C_x$, and that the loss function $\ell(t, y)$ is Lipschitz continuous in t for all $|t| \le C_g$ with a Lipschitz constant L_ℓ.

Theorem 4 (Estimation error bound)**.** *Assume that (a) $\inf_{g\in\mathcal{G}} R_{\mathrm{p}}^+(g) \ge \alpha/\pi_{\mathrm{p}} > 0$, $\inf_{g\in\mathcal{G}} R_{\mathrm{n}}^-(g) \ge \beta/\pi_{\mathrm{n}} > 0$; (b) $\mathcal{G}$ is closed under negation, i.e., $g \in \mathcal{G}$ if and only if $-g \in \mathcal{G}$. Let $\Delta = \exp\left(-\frac{2\alpha^2/C_\ell^2}{a^2/n + c^2/n'}\right) + \exp\left(-\frac{2\beta^2/C_\ell^2}{b^2/n' + d^2/n}\right)$. Then, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\begin{aligned} &R(\widehat{g}_{\mathrm{cc}}) - R(g^*) \\ &\le 8(a+b)L_f L_\ell \mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G}) + 8(c+d)L_f L_\ell \mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}) \\ &\quad + 2(L_f+1)(a+b+c+d)C_\ell\Delta + 2C'_\delta \cdot \chi_{n,n'}, \end{aligned} \tag{11}$$

where $C'_\delta = C_\ell L_f\sqrt{\ln(1/\delta)/2}$, and $\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})$ and $\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G})$ are the Rademacher complexities of $\mathcal{G}$ for the sampling of size n from $p_{\mathrm{tr}}(x)$ and of size n' from $p'_{\mathrm{tr}}(x)$, respectively.

Theorem 4 ensures that learning with (6) is also consistent: as $n, n' \to \infty$, $R(\widehat{g}_{\mathrm{cc}}) \to R(g^*)$, since $\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})$, $\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}) \to 0$ for all parametric models with a bounded norm and $\Delta \to 0$. Specifically, for linear-in-parameter models with a bounded norm, $\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G}) = \mathcal{O}(1/\sqrt{n})$ and $\mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}) = \mathcal{O}(1/\sqrt{n'})$, and thus $R(\widehat{g}_{\mathrm{cc}}) \to R(g^*)$ in $\mathcal{O}_p(1/\sqrt{n} + 1/\sqrt{n'})$. Furthermore, for deep neural networks, we can obtain the following corollary based on the results in Golowich et al. (2017).

Consider neural networks of the form $g : x \mapsto W_m\sigma_{m-1}(W_{m-1}\sigma_{m-2}(\dots\sigma_1(W_1x)))$, where m is the depth of the neural network, $W_1, \dots, W_m$ are weight matrices, and $\sigma_1, \dots, \sigma_{m-1}$ are activation functions for each layer.

Corollary 5. *Assume the Frobenius norm of the weight matrices W_j are at most $M_F(j)$. Let σ be a positive-homogeneous (i.e., it is element-wise and satisfies $\sigma(\alpha z) = \alpha\sigma(z)$ for all $\alpha \ge 0$ and $z \in \mathbb{R}$), 1-Lipschitz activation function which is applied element-wise (such as the ReLU). Then, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\begin{aligned} &R(\widehat{g}_{\mathrm{cc}}) - R(g^*) \le \\ &\left(8L_f L_\ell C_x\left(\sqrt{2m\log 2} + 1\right)\prod_{j=1}^{m} M_F(j) + 2C'_\delta\right) \cdot \chi_{n,n'} \\ &+ 2(L_f+1)(a+b+c+d)C_\ell\Delta. \end{aligned}$$

The factor $(\sqrt{2m\log 2} + 1)\prod_{j=1}^m M_F(j)$ is induced by the hypothesis complexity of the deep neural network and could be improved (Golowich et al., 2017). From Corollary 5, for fully connected neural networks, we obtain the same convergence rate as the linear-in-parameter models.

4 Experiments

In this section, we verify the effectiveness of the proposed consistent risk correction methods on various models and datasets, and test under different class prior settings for an extensive investigation.

Datasets We train on widely adopted benchmarks MNIST, Fashion-MNIST, Kuzushiji-MNIST and CIFAR-10. Table 1 summarizes the benchmark

Table 2: Means (standard deviations) of the classification accuracy (Acc) and the drop (Δ_A) over five trials in percentage with simple models. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Dataset	θ, θ'	UU-Biased		UU-Unbiased		UU-ABS		UU-ReLU		UU-LReLU	
		Acc	Δ_A	Acc	Δ_A	Acc	Δ_A	Acc	Δ_A	Acc	Δ_A
MNIST	0.8, 0.2	89.30 (0.09)	0.28 (0.12)	**89.76** **(0.13)**	0.10 (0.04)	**89.80** **(0.20)**	0.10 (0.03)	**89.68** **(0.14)**	0.14 (0.07)	**89.70** **(0.14)**	0.12 (0.07)
	0.7, 0.3	88.59 (0.22)	0.42 (0.11)	**89.26** **(0.09)**	0.11 (0.07)	**89.19** **(0.20)**	0.21 (0.06)	**89.24** **(0.11)**	0.14 (0.07)	**89.15** **(0.27)**	0.20 (0.06)
	0.6, 0.4	84.64 (0.42)	2.03 (0.15)	**87.15** **(0.34)**	0.54 (0.22)	**87.28** **(0.38)**	0.49 (0.23)	**87.13** **(0.33)**	0.60 (0.25)	**87.26** **(0.37)**	0.40 (0.08)
Fashion-MNIST	0.8, 0.2	**87.27** **(0.83)**	0.82 (0.73)	**87.73** **(0.11)**	0.43 (0.06)	**87.72** **(0.11)**	0.44 (0.06)	**87.78** **(0.20)**	0.35 (0.14)	**87.78** **(0.20)**	0.39 (0.13)
	0.7, 0.3	85.53 (0.93)	2.02 (0.77)	**86.99** **(0.17)**	0.73 (0.14)	**87.02** **(0.35)**	0.71 (0.26)	**87.07** **(0.28)**	0.72 (0.20)	**86.84** **(0.56)**	0.97 (0.52)
	0.6, 0.4	80.66 (2.22)	4.59 (1.91)	**83.69** **(0.53)**	2.70 (0.46)	**84.18** **(0.57)**	2.41 (0.57)	**84.20** **(0.44)**	2.08 (0.63)	**83.92** **(1.07)**	2.59 (1.09)
Kuzushiji-MNIST	0.8, 0.2	72.73 (0.39)	1.59 (0.45)	**79.19** **(0.29)**	0.39 (0.18)	**79.28** **(0.29)**	0.39 (0.18)	**79.28** **(0.29)**	0.39 (0.18)	**79.32** **(0.19)**	0.35 (0.20)
	0.7, 0.3	72.21 (0.74)	1.91 (0.52)	**78.67** **(0.34)**	0.75 (0.25)	**78.89** **(0.40)**	0.75 (0.23)	**78.79** **(0.21)**	0.63 (0.25)	**78.90** **(0.40)**	0.63 (0.28)
	0.6, 0.4	69.76 (0.46)	3.14 (0.70)	**77.73** **(0.37)**	1.24 (0.24)	**77.95** **(0.71)**	1.20 (0.43)	**77.84** **(0.65)**	1.26 (0.36)	**77.86** **(0.72)**	1.19 (0.29)
CIFAR-10	0.8, 0.2	76.94 (5.49)	4.62 (5.35)	**80.50** **(1.20)**	1.49 (1.22)	**80.48** **(1.19)**	1.50 (1.21)	**80.82** **(0.69)**	1.07 (0.58)	**81.13** **(0.51)**	0.76 (0.41)
	0.7, 0.3	**78.04** **(2.02)**	2.22 (2.18)	**79.68** **(0.66)**	1.54 (0.56)	**80.12** **(0.42)**	1.20 (0.35)	**80.28** **(0.14)**	1.03 (0.21)	**79.95** **(0.67)**	1.32 (0.59)
	0.6, 0.4	67.23 (6.68)	9.05 (6.77)	**76.34** **(1.41)**	3.74 (1.51)	**75.21** **(1.95)**	4.81 (1.93)	**76.24** **(0.96)**	3.85 (0.99)	**76.28** **(0.92)**	3.72 (1.06)

datasets. Following Lu et al. (2019), we manually corrupted the 10-class datasets into binary classification datasets (see Appendix C for details). Two unlabeled training datasets $\mathcal{X}_{\text{tr}}$ and $\mathcal{X}'_{\text{tr}}$ of the same sample size are drawn according to Eq. (3). And the risk is evaluated on them during training. Test data are just drawn from $p(x, y)$ for evaluations.

Baselines In order to analyze the proposed methods, we compare them with two baselines:

- *UU-Biased* means supervised classification taking the U set with larger class prior as P data and the other U set with smaller class prior as N data, which is a straightforward method to handle UU classification problem. In our setup, two U sets are of the same sample size, thus UU-biased method reduces to the BER minimization method (Menon et al., 2015);
- *UU-Unbiased* means the state-of-the-art UU method proposed in Lu et al. (2019).

For our proposed methods, *UU-ABS*, *UU-ReLU*, *UU-LReLU* are short for the unbiased UU method using ABSolute function, ReLU function and generalized Leaky ReLU function as consistent correction function in Eq. (6) respectively.

Experimental setup We first demonstrate that the aforementioned overfitting problem cannot be solved by simply applying the regularization techniques in deep learning, such as dropout and weight decay. For space reasons, we defer the experimental results on general-purpose regularization and discussions to Appendix D. We then test our proposed methods under different training class prior settings: (θ, θ') are chosen as $(0.9, 0.1)$, $(0.8, 0.2)$, and $(0.7, 0.3)$; and using different models which are summarized in Table 1: MLP refers to *multi-layer perceptron*, ResNet refers to *residual networks* (He et al., 2016) and their detailed architectures are in Appendix C.

We implemented all the methods by Keras[5], and conducted the experiments on a NVIDIA Tesla P100 GPU. As a common practice, Adam (Kingma and Ba, 2015) with logistic loss $\ell_{\log}(z) = \ln(1 + \exp(-z))$ was used for optimization. We trained 200 epochs and besides the final classification accuracy (Acc) we also report the classification accuracy drop (Δ_A), which is the difference between the best accuracy of all training epochs and the accuracy at the end of training, to demonstrate overfitting. Note that for fair comparison, we use the same models and hyperparameters (see Appendix C) for the implementation of all methods.

[5] https://keras.io

Table 3: Means (standard deviations) of the classification accuracy (Acc) and the drop (Δ_A) over five trials in percentage with deep models. The best and comparable methods based on the paired t-test at the significance level 5% are highlighted in boldface.

Dataset	θ, θ'	UU-Biased		UU-Unbiased		UU-ABS		UU-ReLU		UU-LReLU	
		Acc	Δ_A	Acc	Δ_A	Acc	Δ_A	Acc	Δ_A	Acc	Δ_A
MNIST	0.8, 0.2	80.56 (0.62)	14.99 (0.74)	78.01 (0.45)	18.07 (0.55)	**95.19** **(0.12)**	0.86 (0.13)	**95.15** **(0.43)**	1.11 (0.36)	**95.21** **(0.42)**	1.06 (0.42)
	0.7, 0.3	70.55 (0.66)	20.80 (0.75)	64.74 (0.78)	29.78 (0.84)	91.69 (1.13)	2.77 (1.08)	**93.01** **(0.39)**	1.88 (0.53)	**93.29** **(0.81)**	1.60 (0.76)
	0.6, 0.4	59.85 (0.59)	19.37 (1.11)	53.34 (0.88)	37.74 (1.31)	78.54 (1.19)	11.73 (1.27)	**88.11** **(1.48)**	3.37 (1.38)	**90.34** **(0.84)**	1.13 (0.66)
Fashion-MNIST	0.8, 0.2	81.51 (0.77)	9.56 (0.65)	80.19 (0.81)	11.40 (0.81)	**90.41** **(0.56)**	1.13 (0.43)	90.20 (0.53)	1.35 (0.37)	**90.90** **(0.26)**	0.64 (0.25)
	0.7, 0.3	72.07 (0.94)	16.23 (0.63)	71.93 (1.42)	18.48 (1.29)	87.84 (0.80)	2.43 (0.70)	**88.14** **(0.90)**	2.22 (1.00)	**89.39** **(0.18)**	0.97 (0.15)
	0.6, 0.4	61.58 (1.30)	17.89 (0.75)	63.01 (1.07)	25.10 (1.17)	80.86 (1.38)	6.83 (1.59)	83.78 (1.00)	4.11 (0.84)	**86.25** **(0.32)**	1.63 (0.24)
Kuzushiji-MNIST	0.8, 0.2	78.10 (0.69)	8.22 (0.83)	74.60 (0.71)	14.76 (0.77)	**86.62** **(1.11)**	2.85 (1.19)	**87.13** **(0.99)**	2.28 (0.87)	**87.56** **(0.62)**	1.85 (0.45)
	0.7, 0.3	70.77 (0.58)	10.95 (0.63)	66.40 (0.49)	21.12 (0.48)	83.79 (0.66)	3.81 (0.55)	**85.35** **(0.60)**	2.20 (0.41)	**85.65** **(0.29)**	1.60 (0.23)
	0.6, 0.4	61.70 (0.76)	11.44 (1.41)	60.12 (0.90)	23.59 (1.12)	77.82 (1.12)	5.79 (1.19)	80.52 (1.35)	3.32 (1.10)	**82.22** **(0.52)**	1.61 (0.27)
CIFAR-10	0.8, 0.2	74.28 (0.94)	10.76 (1.37)	76.12 (3.51)	11.48 (3.21)	**84.39** **(1.34)**	3.22 (1.04)	**84.47** **(1.68)**	3.18 (1.32)	**84.51** **(1.33)**	3.11 (0.95)
	0.7, 0.3	65.06 (0.46)	14.09 (1.59)	67.52 (3.07)	17.84 (2.85)	**80.53** **(1.52)**	4.84 (0.72)	**81.64** **(1.46)**	3.73 (1.19)	**81.26** **(2.51)**	4.08 (2.44)
	0.6, 0.4	57.12 (0.46)	12.52 (2.25)	57.26 (1.33)	23.95 (1.35)	71.53 (1.40)	9.30 (0.66)	76.83 (1.26)	4.10 (0.91)	**78.34** **(1.00)**	2.62 (0.69)

Experimental results with simple models We firstly test on simple models and report our results in Table 2. We can see that the UU-Unbiased method and three consistent risk correction methods outperform the UU-biased method. The advantage increases as the classification task becomes harder, that is, the class priors move closer[6]. Moreover, the overfitting issue in UU-Unbiased method is not severe for linear models, but we can see the tendency that overfitting gets slightly worse when class priors are closer.

Experimental results with deep models We now test on deep neural networks which are more flexible than the aforementioned simple models. Our observations of the experimental results in Table 3 are as follows. First, compared to simple model experiments, the overfitting of the UU-biased and UU-Unbiased methods become catastrophic: the performance drops behind their linear counterparts. This may be explained by that flexible models have larger capacity to fit patterns (making negative partial risks $-b\widehat{R}_{\mathrm{u}}^{-}(g)$ and $-c\widehat{R}_{\mathrm{u}'}^{+}(g)$ in (4) as negative as possible) and thus the empirical risk tends to be negative. And we observe that the closer the class priors are, the more severe the overfitting is. Second, the proposed consistent risk correction methods significantly alleviate the overfitting even in the hardest learning scenario, and their classification accuracy improves compared to the simple model experiments. Among all methods, the UU-LReLU method achieves the best performance for all the datasets and class prior settings, and has the smallest performance drop when the class priors get closer, which implies that it is relatively robust against the closeness of class priors.

5 Conclusions

We focused on mitigating the overfitting problem of the state-of-the-art unbiased UU method. Based on our empirical observations, we conjecture the negative empirical training risk as a potential reason for the overfitting and proposed a correction method that wraps the false positive and false negative parts of the empirical risk in a family of consistent correction functions. Furthermore, we proved the consistency of the proposed risk estimators and their minimizers. Experiments demonstrated the superiority of our proposed methods, especially for using flexible neural network models.

[6] Intuitively, as the class priors move closer, two U sets would be more similar and thus less informative, which is significantly harder than assuming that they are sufficiently far away.

Acknowledgments

We thank Aditya Krishna Menon, Takashi Ishida, Ikko Yamane and Wenkai Xu for helpful discussions. NL was supported by the MEXT scholarship No. 171536. MS was supported by JST CREST Grant Number JPMJCR18A2.

References

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3: 463–482, 2002.

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

S. Ben-David, N. Eiron, and P.M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3): 496–514, 2003.

G. Blanchard, M. Flaska, G. Handy, S. Pozzi, and C. Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10(2):2780–2824, 2016.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In *ICPR*, 2010.

O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *NeurIPS*, 2002.

N. Charoenphakdee, J. Lee, and M. Sugiyama. On symmetric losses for learning from corrupted labels. In *ICML*, 2019.

K.-L. Chung. *A Course in Probability Theory*. Academic Press, 1968.

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

M. C. du Plessis, G. Niu, and M. Sugiyama. Clustering unclustered data: Unsupervised binary labeling of two datasets having different class balances. In *TAAI*, 2013.

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

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

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

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

R. Fakoor, F. Ladhak, A. Nazi, and M. Huber. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning*, volume 28. ACM New York, USA, 2013.

N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *NeurIPS*, 2010.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, 2004.

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

W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama. Learning discrete representations via information maximizing self augmented training. In *ICML*, 2017.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

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

S. Jain, M. White, and P. Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. In *NeurIPS*, pages 2693–2701, 2016.

M. Kato, T. Teshima, and J. Honda. Learning from positive and unlabeled data with a selection bias. In *ICLR*, 2019.

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

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

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

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

S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

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

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

M. Li and Z.H. Zhou. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):1088–1098, 2007.

T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3): 447–461, 2016.

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

Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *CVPR*, 2018.

G. S. Mann and A. McCallum. Simple, robust, scalable semi-supervised learning via expectation regularization. In *ICML*, 2007.

C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

A. K. Menon, B. van Rooyen, C. S. Ong, and R. C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, 2015.

T. Miyato, S. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing with virtual adversarial training. In *ICLR*, 2016.

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

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

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

A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018.

C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT*, 2013.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014a.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014b.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR*, 2015.

M. Sugiyama, G. Niu, M. Yamada, M. Kimura, and H. Hachiya. Information-maximization clustering based on squared-loss mutual information. *Neural Computation*, 26(1):84–131, 2014.

W. Sun, T. B. Tseng, J. Zhang, and W. Qian. Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics*, 57:4–9, 2017.

H. Valizadegan and R. Jin. Generalized maximum margin clustering and unsupervised kernel learning. In *NeurIPS*, 2006.

B. van Rooyen and R. C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

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

L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *NeurIPS*, 2004.

Supplementary Material for Mitigating Overfitting in Supervised Classification from Two Unlabeled Datasets: A Consistent Risk Correction Approach

A Proofs

In this appendix, we prove all theorems.

A.1 Proof of Lemma 2

Let

$$p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}}) = p_{\mathrm{tr}}(x_1)\cdots p_{\mathrm{tr}}(x_n), \quad p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}}) = p'_{\mathrm{tr}}(x'_1)\cdots p'_{\mathrm{tr}}(x'_{n'})$$

be the probability density functions of $\mathcal{X}_{\mathrm{tr}}$ and $\mathcal{X}'_{\mathrm{tr}}$ (due to the i.i.d. sample assumption). Then, the measure of $\mathfrak{D}^-(g)$ is defined by

$$\Pr(\mathfrak{D}^-(g)) = \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}},$$

where Pr denotes the probability, $\mathrm{d}\mathcal{X}_{\mathrm{tr}} = \mathrm{d}x_1\cdots\mathrm{d}x_n$ and $\mathrm{d}\mathcal{X}'_{\mathrm{tr}} = \mathrm{d}x'_1\cdots\mathrm{d}x'_{n'}$. Since $\widehat{R}_{\mathrm{uu}}(g)$ is unbiased and $\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g) = 0$ on $\mathfrak{D}^+(g)$, the bias of $\widehat{R}_{\mathrm{cc}}(g)$ can be formulated as:

$$\begin{aligned}
\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)] - R(g) &= \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)] \\
&= \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^+(g)} \left(\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)\right) p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}} \\
&\quad + \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} \left(\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)\right) p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}} \\
&= \int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} \left(\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)\right) p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}}
\end{aligned}$$

Thus we have $\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)] - R(g) > 0$ if and only if $\int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^-(g)} p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}} > 0$ due to the fact that $\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g) > 0$ on $\mathfrak{D}^-(g)$. That is, the bias of $\widehat{R}_{\mathrm{cc}}(g)$ is positive if and only if the measure of $\mathfrak{D}^-(g)$ is non-zero.

Next we study the probability measure of $\mathfrak{D}^-(g)$ by *the method of bounded differences*. Since $R_{\mathrm{p}}^+(g) \geq \alpha_g/\pi_{\mathrm{p}}$ and $R_{\mathrm{n}}^-(g) \geq \beta_g/\pi_{\mathrm{n}}$, then

$$\mathbb{E}[A - C] = \pi_{\mathrm{p}}R_{\mathrm{p}}^+(g) \geq \alpha_g, \quad \mathbb{E}[D - B] = \pi_{\mathrm{n}}R_{\mathrm{n}}^-(g) \geq \beta_g.$$

We have assumed that $0 \leq \ell(z) \leq C_\ell$, and thus the change of $a\widehat{R}_{\mathrm{u}}^+(g)$ and $b\widehat{R}_{\mathrm{u}}^-(g)$ will be no more than aC_ℓ/n and bC_ℓ/n if some $x_i \in \mathcal{X}_{\mathrm{tr}}$ is replaced, or the change of $c\widehat{R}_{\mathrm{u}'}^+(g)$ and $d\widehat{R}_{\mathrm{u}'}^-(g)$ will be no more than cC_ℓ/n' and dC_ℓ/n' if some $x'_j \in \mathcal{X}'_{\mathrm{tr}}$ is replaced. Subsequently, *McDiarmid's inequality* (McDiarmid, 1989) implies

$$\begin{aligned}
\Pr\{\pi_{\mathrm{p}}R_{\mathrm{p}}^+(g) - (A - C) \geq \alpha_g\} &\leq \exp\left(-\frac{2\alpha_g^2}{n(aC_\ell/n)^2 + n'(cC_\ell/n')^2}\right) \\
&= \exp\left(-\frac{2\alpha_g^2/C_\ell^2}{a^2/n + c^2/n'}\right),
\end{aligned}$$

and

$$\begin{aligned}\Pr\{\pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g)-(D-B)\geq \beta_g\} &\leq \exp\left(-\frac{2\beta_g^2}{n'(dC_\ell/n')^2+n(bC_\ell/n)^2}\right)\\ &= \exp\left(-\frac{2\beta_g^2/C_\ell^2}{b^2/n'+d^2/n}\right).\end{aligned}$$

Then the probability measure of $\mathfrak{D}^{-}(g)$ can be bounded by

$$\begin{aligned}\Pr(\mathfrak{D}^{-}(g)) &\leq \Pr\{A-C\leq 0\}+\Pr\{D-B<0\}\\ &\leq \Pr\{A-C\leq \pi_{\mathrm{p}} R_{\mathrm{p}}^{+}(g)-\alpha_g\}+\Pr\{D-B\leq \pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g)-\beta_g\}\\ &= \Pr\{\pi_{\mathrm{p}} R_{\mathrm{p}}^{+}(g)-(A-C)\geq \alpha_g\}+\Pr\{\pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g)-(D-B)\geq \beta_g\}\\ &\leq \exp\left(-\frac{2\alpha_g^2/C_\ell^2}{a^2/n+c^2/n'}\right)+\exp\left(-\frac{2\beta_g^2/C_\ell^2}{b^2/n'+d^2/n}\right),\end{aligned}$$

we complete the proof. □

A.2 Proof of Theorem 3

Based on Lemma 2, we can show the exponential decay of the bias and also the consistency of the proposed non-negative risk estimator $\widehat{R}_{\mathrm{cc}}(g)$. It has been proved in Lemma 2 that

$$\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]-R(g)=\int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}\left(\widehat{R}_{\mathrm{cc}}(g)-\widehat{R}_{\mathrm{uu}}(g)\right)p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}}.$$

Therefore the exponential decay of the bias can be obtained via

$$\begin{aligned}\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]-R(g) &\leq \sup\nolimits_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}\left(\widehat{R}_{\mathrm{cc}}(g)-\widehat{R}_{\mathrm{uu}}(g)\right)\cdot\int_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}p_{\mathrm{tr}}(\mathcal{X}_{\mathrm{tr}})p'_{\mathrm{tr}}(\mathcal{X}'_{\mathrm{tr}})\mathrm{d}\mathcal{X}_{\mathrm{tr}}\mathrm{d}\mathcal{X}'_{\mathrm{tr}}\\ &= \sup\nolimits_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}(f_1(A-C)+f_2(D-B)-(A-C)-(D-B))\cdot\Pr(\mathfrak{D}^{-}(g))\\ &\leq \sup\nolimits_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}(|f_1(A-C)|+|f_2(D-B)|+|A-C|+|D-B|)\cdot\Pr(\mathfrak{D}^{-}(g))\\ &\leq \sup\nolimits_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}(L_f|A-C|+L_f|D-B|+|A-C|+|D-B|)\cdot\Pr(\mathfrak{D}^{-}(g))\\ &= \sup\nolimits_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}((L_f+1)|A-C|+(L_f+1)|D-B|)\cdot\Pr(\mathfrak{D}^{-}(g))\\ &\leq \sup\nolimits_{(\mathcal{X}_{\mathrm{tr}},\mathcal{X}'_{\mathrm{tr}})\in\mathfrak{D}^{-}(g)}((L_f+1)(a+c)C_\ell+(L_f+1)(d+b)C_\ell)\cdot\Pr(\mathfrak{D}^{-}(g))\\ &= (L_f+1)(a+b+c+d)C_\ell\Delta_g,\end{aligned}$$

where we employed the Lipschitz condition, i.e., $|f_1(x)-f_1(y)|\leq L_f|x-y|$ (also holds for f_2), and the assumption $f(0)=0$ in Definition 1. Then the deviation bound (9) is due to

$$\begin{aligned}|\widehat{R}_{\mathrm{cc}}(g)-R(g)| &\leq |\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|+|\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]-R(g)|\\ &\leq |\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|+(L_f+1)(a+b+c+d)C_\ell\Delta_g.\end{aligned}$$

Denote by A', B', C' and D' that differs from A, B ,C and D on a single example. Then

$$\begin{aligned}&|f_1(A-C)+f_2(D-B)-f_1(A'-C)-f_2(D-B')|\\ &\quad\leq |f_1(A-C)-f_1(A'-C)|+|f_2(D-B)-f_2(D-B')|\\ &\quad\leq L_f|A-C-A'+C|+L_f|D-B-D+B'|\\ &\quad= L_f|A-A'|+L_f|B'-B|\\ &\quad\leq (a+b)L_fC_\ell/n. \end{aligned}\tag{12}$$

Similarly, we can obtain

$$|f_1(A-C)+f_2(D-B)-f_1(A-C')-f_2(D'-B)|\leq (c+d)L_fC_\ell/n'. \tag{13}$$

Therefore the change of $\widehat{R}_{\mathrm{cc}}(g)$ will be no more than $(a+b)L_f C_\ell/n$ if some $x_i \in \mathcal{X}_{\mathrm{tr}}$ is replaced, or it will be no more than $(c+d)L_f C_\ell/n'$ if some $x'_j \in \mathcal{X}'_{\mathrm{tr}}$ is replaced, and McDiarmid's inequality gives us

$$\Pr\{|\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]| \geq \epsilon\} \leq 2\exp\left(-\frac{2\epsilon^2}{n((a+b)L_f C_\ell/n)^2 + n'((c+d)L_f C_\ell/n')^2}\right).$$

Setting the above right-hand side to be equal to δ and solving for ϵ yields immediately the following bound. For any $\delta > 0$, the following inequality holds with probability at least $1-\delta$,

$$\begin{aligned}|\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]| &\leq \sqrt{\frac{\ln(2/\delta)C_\ell^2 L_f^2}{2}\left(\frac{(a+b)^2}{n} + \frac{(c+d)^2}{n'}\right)}\\ &\leq C_\delta\left(\frac{(a+b)}{\sqrt{n}} + \frac{(c+d)}{\sqrt{n'}}\right)\\ &= C_\delta \cdot \chi_{n,n'},\end{aligned}$$

where $C_\delta = C_\ell L_f\sqrt{\ln(2/\delta)/2}$ and $\chi_{n,n'} = (a+b)/\sqrt{n} + (c+d)/\sqrt{n'}$. Thus we obtain

$$|\widehat{R}_{\mathrm{cc}}(g) - R(g)| \leq C_\delta \cdot \chi_{n,n'} + (L_f+1)(a+b+c+d)C_\ell \Delta_g.$$

On the other hand, the deviation bound (10) is due to

$$|\widehat{R}_{\mathrm{cc}}(g) - R(g)| \leq |\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)| + |\widehat{R}_{\mathrm{uu}}(g) - R(g)|,$$

where $|\widehat{R}_{\mathrm{cc}}(g) - \widehat{R}_{\mathrm{uu}}(g)| > 0$ with probability at most Δ_g, and $|\widehat{R}_{\mathrm{uu}}(g) - R(g)|$ shares the same concentration inequality with $|\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|$. □

A.3 Proof of Theorem 4

First, we introduce the definitions of Rademacher complexity.

Definition 6 (Rademacher complexity)**.** *Let $\mathcal{G} = \{g : \mathcal{Z} \to \mathbb{R}\}$ be a class of measurable functions, $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a fixed sample of size n i.i.d. drawn from a probability distribution p, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ be Rademacher variables, i.e., independent uniform random variables taking values in $\{-1, +1\}$. For any integer $n \geq 1$, the Rademacher complexity of $\mathcal{G}$ (Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014a) is defined as*

$$\mathfrak{R}_{n,p}(\mathcal{G}) = \mathbb{E}_{\mathcal{X}}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum\nolimits_{x_i\in\mathcal{X}} \varepsilon_i g(x_i)\right].$$

An alternative definition of the Rademacher complexity (Koltchinskii, 2001; Bartlett and Mendelson, 2002) will be used in the proof is:

$$\mathfrak{R}'_{n,p}(\mathcal{G}) = \mathbb{E}_{\mathcal{X}}\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum\nolimits_{x_i\in\mathcal{X}} \varepsilon_i g(x_i)\right|\right].$$

Then, we list all the lemmas that will be used to derive the estimation error bound in Theorem 4.

Lemma 7. *For arbitrary $\mathcal{G}$, $\mathfrak{R}'_{n,p}(\mathcal{G}) \geq \mathfrak{R}_{n,p}(\mathcal{G})$; if $\mathcal{G}$ is closed under negation, $\mathfrak{R}'_{n,p}(\mathcal{G}) = \mathfrak{R}_{n,p}(\mathcal{G})$.*

Lemma 8 (Theorem 4.12 in Ledoux and Talagrand (1991))**.** *If $\psi : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous function with a Lipschitz constant L_ψ and satisfies $\psi(0) = 0$, we have*

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

where $\psi \circ \mathcal{G} = \{\psi \circ g | g \in \mathcal{G}\}$ and $\circ$ is a composition operator.

Lemma 9. *Under the assumptions of Theorem 4, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\begin{aligned}\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - R(g)| \leq{}& 4(a+b)L_f L_\ell \mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G}) + 4(c+d)L_f L_\ell \mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G})\\ &+ (L_f+1)(a+b+c+d)C_\ell\Delta + C'_\delta \cdot \chi_{n,n'}.\end{aligned} \tag{14}$$

Proof. Firstly, we deal with the bias of $\widehat{R}_{\rm cc}(g)$. Noticing that the assumptions $\inf_{g\in\mathcal{G}} R_{\rm p}^{+}(g) \geq \alpha/\pi_{\rm p} > 0$ and $\inf_{g\in\mathcal{G}} R_{\rm n}^{-}(g) \geq \beta/\pi_{\rm n} > 0$ imply $\Delta = \sup_{g\in\mathcal{G}} \Delta_g$. By (8) we have:

$$\begin{aligned}\sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - R(g)| &\leq \sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - \mathbb{E}[\widehat{R}_{\rm cc}(g)]| + \sup\nolimits_{g\in\mathcal{G}} |\mathbb{E}[\widehat{R}_{\rm cc}(g)] - R(g)| \\ &\leq \sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - \mathbb{E}[\widehat{R}_{\rm cc}(g)]| + (L_f+1)(a+b+c+d)C_\ell\Delta. \qquad (15)\end{aligned}$$

Secondly, we consider the double-sided uniform deviation $\sup_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - \mathbb{E}[\widehat{R}_{\rm cc}(g)]|$. Denote by $\mathcal{X}_{\rm s} = \{(\mathcal{X}_{\rm tr}, \mathcal{X}'_{\rm tr})\}$, and $\mathcal{X}'_{\rm s}$ that differs from $\mathcal{X}_{\rm s}$ on a single example. Then we have

$$\begin{aligned}&\big|\sup\nolimits_{g\in\mathcal{G}} \big|\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm s}) - \mathbb{E}_{\mathcal{X}_{\rm s}}[\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm s})]\big| - \sup\nolimits_{g\in\mathcal{G}} \big|\widehat{R}_{\rm cc}(g;\mathcal{X}'_{\rm s}) - \mathbb{E}_{\mathcal{X}'_{\rm s}}[\widehat{R}_{\rm cc}(g;\mathcal{X}'_{\rm s})]\big|\big| \\ &\leq \sup\nolimits_{g\in\mathcal{G}} \big||\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm s}) - \mathbb{E}_{\mathcal{X}_{\rm s}}[\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm s})]| - |\widehat{R}_{\rm cc}(g;\mathcal{X}'_{\rm s}) - \mathbb{E}_{\mathcal{X}'_{\rm s}}[\widehat{R}_{\rm cc}(g;\mathcal{X}'_{\rm s})]|\big| \\ &\leq \sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm s}) - \widehat{R}_{\rm cc}(g;\mathcal{X}'_{\rm s})|,\end{aligned}$$

where we applied the *triangle inequality*. According to (12) and (13), we see that the change of $\sup_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - \mathbb{E}[\widehat{R}_{\rm cc}(g)]|$ will be no more than $(a+b)L_f C_\ell/n$ if some $x_i \in \mathcal{X}_{\rm tr}$ is replaced, or it will be no more than $(c+d)L_f C_\ell/n'$ if some $x'_j \in \mathcal{X}'_{\rm tr}$ is replaced. Similar to the proof technique of Theorem 3, by applying McDiarmid's inequality to the uniform deviation we have with probability at least $1-\delta$,

$$\begin{aligned}&\sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - \mathbb{E}[\widehat{R}_{\rm cc}(g)]| - \mathbb{E}[\sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - \mathbb{E}[\widehat{R}_{\rm cc}(g)]|] \\ &\leq \sqrt{\frac{\ln(1/\delta)C_\ell^2 L_f^2}{2}\left(\frac{(a+b)^2}{n} + \frac{(c+d)^2}{n'}\right)} \\ &= C'_\delta \cdot \chi_{n,n'}, \qquad (16)\end{aligned}$$

where $C'_\delta = C_\ell L_f \sqrt{\ln(1/\delta)/2}$. Thirdly, we make *symmetrization* (Vapnik, 1998). Suppose that $(\mathcal{X}_{\rm tr}^{gh}, \mathcal{X}_{\rm tr}^{'gh})$ is a *ghost sample*, then

$$\begin{aligned}&\mathbb{E}[\sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g) - \mathbb{E}[\widehat{R}_{\rm cc}(g)]|] \\ &= \mathbb{E}_{(\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr})}[\sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr}) - \mathbb{E}_{(\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})}\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})|] \\ &\leq \mathbb{E}_{(\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr})}[\sup\nolimits_{g\in\mathcal{G}} \mathbb{E}_{(\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})}|\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr}) - \widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})|] \\ &\leq \mathbb{E}_{(\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr}),(\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})}[\sup\nolimits_{g\in\mathcal{G}} |\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr}) - \widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})|],\end{aligned}$$

where we applied *Jensen's inequality* twice since the absolute value and the supremum are convex. By decomposing the difference $|\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr}) - \widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})|$, we can know that

$$\begin{aligned}&|\widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr},\mathcal{X}'_{\rm tr}) - \widehat{R}_{\rm cc}(g;\mathcal{X}_{\rm tr}^{gh},\mathcal{X}_{\rm tr}^{'gh})| \\ &= \Big|f_1\left(a\widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}) - c\widehat{R}_{\rm u'}^{+}(g;\mathcal{X}'_{\rm tr})\right) - f_1\left(a\widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}^{gh}) - c\widehat{R}_{\rm u'}^{+}(g;\mathcal{X}_{\rm tr}^{'gh})\right) \\ &+ f_2\left(-b\widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}) + d\widehat{R}_{\rm u'}^{-}(g;\mathcal{X}'_{\rm tr})\right) - f_2\left(-b\widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}^{gh}) + d\widehat{R}_{\rm u'}^{-}(g;\mathcal{X}_{\rm tr}^{'gh})\right)\Big| \\ &\leq \Big|f_1\left(a\widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}) - c\widehat{R}_{\rm u'}^{+}(g;\mathcal{X}'_{\rm tr})\right) - f_1\left(a\widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}^{gh}) - c\widehat{R}_{\rm u'}^{+}(g;\mathcal{X}_{\rm tr}^{'gh})\right)\Big| \\ &+ \Big|f_2\left(-b\widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}) + d\widehat{R}_{\rm u'}^{-}(g;\mathcal{X}'_{\rm tr})\right) - f_2\left(-b\widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}^{gh}) + d\widehat{R}_{\rm u'}^{-}(g;\mathcal{X}_{\rm tr}^{'gh})\right)\Big| \\ &\leq \Big|L_f\left(a\widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}) - c\widehat{R}_{\rm u'}^{+}(g;\mathcal{X}'_{\rm tr}) - a\widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}^{gh}) + c\widehat{R}_{\rm u'}^{+}(g;\mathcal{X}_{\rm tr}^{'gh})\right)\Big| \\ &+ \Big|L_f\left(-b\widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}) + d\widehat{R}_{\rm u'}^{-}(g;\mathcal{X}'_{\rm tr}) + b\widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}^{gh}) - d\widehat{R}_{\rm u'}^{-}(g;\mathcal{X}_{\rm tr}^{'gh})\right)\Big| \\ &\leq \Big|aL_f\left(\widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}) - \widehat{R}_{\rm u}^{+}(g;\mathcal{X}_{\rm tr}^{gh})\right)\Big| + \Big|cL_f\left(\widehat{R}_{\rm u'}^{+}(g;\mathcal{X}'_{\rm tr}) - \widehat{R}_{\rm u'}^{+}(g;\mathcal{X}_{\rm tr}^{'gh})\right)\Big| \\ &+ \Big|bL_f\left(\widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}) - \widehat{R}_{\rm u}^{-}(g;\mathcal{X}_{\rm tr}^{gh})\right)\Big| + \Big|dL_f\left(\widehat{R}_{\rm u'}^{-}(g;\mathcal{X}'_{\rm tr}) - \widehat{R}_{\rm u'}^{-}(g;\mathcal{X}_{\rm tr}^{'gh})\right)\Big|,\end{aligned}$$

where we employed the Lipschitz condition. This decomposition results in

$$\begin{aligned}\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|] \leq{}& aL_f\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh})\Big)\Big|\Big]\\ &+cL_f\mathbb{E}_{\mathcal{X}_{\mathrm{tr}}',\mathcal{X}_{\mathrm{tr}}'^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}_{\mathrm{tr}}')-\widehat{R}_{\mathrm{u}'}^{+}(g;\mathcal{X}_{\mathrm{tr}}'^{gh})\Big)\Big|\Big]\\ &+bL_f\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{tr}}^{gh})\Big)\Big|\Big]\\ &+dL_f\mathbb{E}_{\mathcal{X}_{\mathrm{tr}}',\mathcal{X}_{\mathrm{tr}}'^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|\Big(\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}_{\mathrm{tr}}')-\widehat{R}_{\mathrm{u}'}^{-}(g;\mathcal{X}_{\mathrm{tr}}'^{gh})\Big)\Big|\Big].\end{aligned}$$

Fourthly, we relax those expectations to Rademacher complexities. The original ℓ may miss the origin, i.e., $\ell(0,y)\neq 0$, with which we need to cope. Let

$$\bar{\ell}(t,y)=\ell(t,y)-\ell(0,y)$$

be a *shifted loss* so that $\bar{\ell}(0,y)=0$. Hence,

$$\begin{aligned}\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh}) &= (1/n)\textstyle\sum_{x_i\in\mathcal{X}_{\mathrm{tr}}}\ell(g(x_i),+1)-(1/n)\sum_{x_i^{gh}\in\mathcal{X}_{\mathrm{tr}}^{gh}}\ell(g(x_i^{gh}),+1)\\ &=(1/n)\textstyle\sum_{i=1}^{n}(\ell(g(x_i),+1)-\ell(g(x_i^{gh}),+1))\\ &=(1/n)\textstyle\sum_{i=1}^{n}(\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1)).\end{aligned}$$

This is already a standard form where we can attach Rademacher variables to every $\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1)$, so we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}[\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}})-\widehat{R}_{\mathrm{u}}^{+}(g;\mathcal{X}_{\mathrm{tr}}^{gh})|]\\ &=\mathbb{E}_{\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}(\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1))\Big|\Big]\\ &=\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i),+1)-\bar{\ell}(g(x_i^{gh}),+1))\Big|\Big]\\ &\leq\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i),+1)\Big|\Big]\\ &+\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}}^{gh}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i^{gh}),+1)\Big|\Big]\\ &=2\mathbb{E}_{\boldsymbol{\varepsilon},\mathcal{X}_{\mathrm{tr}}}\Big[\sup_{g\in\mathcal{G}}\Big|(1/n)\sum_{i=1}^{n}\varepsilon_i(\bar{\ell}(g(x_i),+1)\Big|\Big]\\ &=2\mathfrak{R}_{n,p_{\mathrm{tr}}}'(\bar{\ell}(\cdot,+1)\circ\mathcal{G})\end{aligned}$$

The other three expectations can be handled analogously. As a result,

$$\begin{aligned}\mathbb{E}[\sup_{g\in\mathcal{G}}|\widehat{R}_{\mathrm{cc}}(g)-\mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|] \leq{}& 2aL_f\mathfrak{R}_{n,p_{\mathrm{tr}}}'(\bar{\ell}(\cdot,+1)\circ\mathcal{G})+2cL_f\mathfrak{R}_{n',p_{\mathrm{tr}}'}'(\bar{\ell}(\cdot,+1)\circ\mathcal{G})\\ &+2bL_f\mathfrak{R}_{n,p_{\mathrm{tr}}}'(\bar{\ell}(\cdot,-1)\circ\mathcal{G})+2dL_f\mathfrak{R}_{n',p_{\mathrm{tr}}'}'(\bar{\ell}(\cdot,-1)\circ\mathcal{G}).\end{aligned}$$

Finally, we transform the Rademacher complexities of composite function classes to the original function class. It is obvious that $\bar{\ell}$ shares the same Lipschitz constant L_ℓ with ℓ, and consequently

$$\begin{aligned}\mathfrak{R}_{n,p_{\mathrm{tr}}}'(\bar{\ell}(\cdot,+1)\circ\mathcal{G})&\leq 2L_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}'(\mathcal{G})=2L_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})\\ \mathfrak{R}_{n',p_{\mathrm{tr}}'}'(\bar{\ell}(\cdot,+1)\circ\mathcal{G})&\leq 2L_\ell\mathfrak{R}_{n',p_{\mathrm{tr}}'}'(\mathcal{G})=2L_\ell\mathfrak{R}_{n',p_{\mathrm{tr}}'}(\mathcal{G})\\ \mathfrak{R}_{n,p_{\mathrm{tr}}}'(\bar{\ell}(\cdot,-1)\circ\mathcal{G})&\leq 2L_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}'(\mathcal{G})=2L_\ell\mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G})\\ \mathfrak{R}_{n',p_{\mathrm{tr}}'}'(\bar{\ell}(\cdot,-1)\circ\mathcal{G})&\leq 2L_\ell\mathfrak{R}_{n',p_{\mathrm{tr}}'}'(\mathcal{G})=2L_\ell\mathfrak{R}_{n',p_{\mathrm{tr}}'}(\mathcal{G})\end{aligned}$$

where we used the assumption that $\mathcal{G}$ is closed under negation, Lemma 7 and Lemma 8. So we have

$$\mathbb{E}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - \mathbb{E}[\widehat{R}_{\mathrm{cc}}(g)]|] \leq 4(a+b)L_f L_\ell \mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G}) + 4(c+d)L_f L_\ell \mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}). \tag{17}$$

Combining (15), (16) and (17) finishes the proof of the uniform deviation bound (14). □

We are now ready to prove our estimation error bound based on the uniform deviation bound in Lemma 9.

$$\begin{aligned}
R(\widehat{g}_{\mathrm{cc}}) - R(g^*) &= \Big(\widehat{R}_{\mathrm{cc}}(\widehat{g}_{\mathrm{cc}}) - \widehat{R}_{\mathrm{cc}}(g^*)\Big) + \Big(R(\widehat{g}_{\mathrm{cc}}) - \widehat{R}_{\mathrm{cc}}(\widehat{g}_{\mathrm{cc}})\Big) + \Big(\widehat{R}_{\mathrm{cc}}(g^*) - R(g^*)\Big) \\
&\leq 0 + 2\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{cc}}(g) - R(g)| \\
&\leq 8(a+b)L_f L_\ell \mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G}) + 8(c+d)L_f L_\ell \mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}) \\
&+ 2(L_f+1)(a+b+c+d)C_\ell\Delta + 2C'_\delta \cdot \chi_{n,n'},
\end{aligned}$$

where $\widehat{R}_{\mathrm{cc}}(\widehat{g}_{\mathrm{cc}}) \leq \widehat{R}_{\mathrm{cc}}(g^*)$ by the definition of g^* and $\widehat{g}_{\mathrm{cc}}$. □

A.4 Proof of Corollary 5

We further get bounds on the Rademacher complexity of deep neural networks by the following Theorem.

Theorem 10 (Theorem 1 in Golowich et al. (2017))**.** *Assume the Frobenius norm of the weight matrices W_j are at most $M_F(j)$, and the activation function σ satisfying the assumption that it is 1-Lipschitz, positive-homogeneous which is applied element-wise (such as the ReLU). Let x is upper bounded by C_x. Then,*

$$\mathfrak{R}_{n,p}(\mathcal{G}) \leq \frac{1}{n}\prod_{j=1}^{m} M_F(j)\cdot(\sqrt{2m\log 2}+1)\sqrt{\sum_{i=1}^{n}\|\mathbf{x}_i\|^2} \leq \frac{C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m} M_F(j)}{\sqrt{n}}. \tag{18}$$

Based on Theorem 10, we proved

$$\begin{aligned}
R(\widehat{g}_{\mathrm{cc}}) - R(g^*) &\leq 8(a+b)L_f L_\ell \mathfrak{R}_{n,p_{\mathrm{tr}}}(\mathcal{G}) + 8(c+d)L_f L_\ell \mathfrak{R}_{n',p'_{\mathrm{tr}}}(\mathcal{G}) \\
&+ 2(L_f+1)(a+b+c+d)C_\ell\Delta + 2C'_\delta \cdot \chi_{n,n'} \\
&\leq 8(a+b)L_f L_\ell \frac{C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m} M_F(j)}{\sqrt{n}} \\
&+ 8(c+d)L_f L_\ell \frac{C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m} M_F(j)}{\sqrt{n'}} \\
&+ 2(L_f+1)(a+b+c+d)C_\ell\Delta + 2C'_\delta \cdot \chi_{n,n'} \\
&= \left(8L_f L_\ell C_x(\sqrt{2m\log 2}+1)\prod_{j=1}^{m} M_F(j) + 2C'_\delta\right)\cdot\chi_{n,n'} \\
&+ 2(L_f+1)(a+b+c+d)C_\ell\Delta.
\end{aligned}$$

B Supplementary information on Figure 1

In Sec. 3.1, we illustrated the overfitting issue of state-of-the-art unbiased UU method using different datasets, different models, different optimizers and different loss functions. The details of these demonstration results are presented here.

In the upper row, the dataset used was MNIST and we artifically corrupt it into a binary classification dataset: even digits form the P class and odd digits form the N class. The models used were a linear-in-input model (Linear) $g(x) = \boldsymbol{\omega}^T x + b$ where $\boldsymbol{\omega} \in \mathbb{R}^{784}$ and $b \in \mathbb{R}$, and a 5-layer *multi-layer perceptron* (MLP): d-300-300-300-300-1. And the optimizer was SGD with momentum (momentum=0.9) with logistic loss $\ell_{\mathrm{log}}(z) = \ln(1+\exp(-z))$ or sigmoid loss $\ell_{\mathrm{sig}}(z) = 1/(1+\exp(z))$. For linear model experiments, the batch size was fixed to be 1000 and the initial learning rate was $5e-2$. For MLP model experiments, the batch size was fixed to be 3000 and the initial learning rate was $1e-3$.

In the bottom row, the dataset used was CIFAR-10 and we artifically corrupt it into a binary classification dataset: the P class is composed of 'bird', 'cat', 'deer', 'dog', 'frog', and 'horse', and the N class is composed of 'airplane', 'automobile', 'ship' and 'truck'. The models used were *all convolutional net* (AllConvNet) (Springenberg et al., 2015) as follows:

0th (input) layer:	(32*32*3)-
1st to 3rd layers:	[C(3*3, 96)]*2-C(3*3, 96, 2)-
4th to 6th layers:	[C(3*3, 192)]*2-C(3*3, 192, 2)-
7th to 9th layers:	C(3*3, 192)-C(1*1, 192)-C(1*1, 10)-
10th to 12th layers:	1000-1000-1

where C(3*3, 96) means 96 channels of 3*3 convolutions followed by ReLU, [$\cdot$]*2 means 2 such layers, C(3*3, 96, 2) means a similar layer but with stride 2, etc; and a 32-layer *residual networks* (ResNet32) (He et al., 2016) as follows:

0th (input) layer:	(32*32*3)-
1st to 11th layers:	C(3*3, 16)-[C(3*3, 16), C(3*3, 16)]*5-
12th to 21st layers:	[C(3*3, 32), C(3*3, 32)]*5-
22nd to 31st layers:	[C(3*3, 64), C(3*3, 64)]*5-
32nd layer:	Global Average Pooling-1

where [$\cdot, \cdot$] means a building block (He et al., 2016). Batch normalization (Ioffe and Szegedy, 2015) was applied before hidden layers. An ℓ_2-regularization was added, where the regularization parameter was fixed to 5e-3. The models were trained by Adam (Kingma and Ba, 2015) with the default momentum parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and the loss function was $\ell_{\mathrm{sig}}(z)$ or $\ell_{\log}(z)$. For AllConvNet experiments, the batch size and the learning rate were fixed to be 500 and $1e-5$ respectively. For ResNet32 experiments, the batch size and the learning rate were fixed to be 3000 and $3e-5$ respectively.

The two training distributions were created following (3) with class priors $\theta = 0.6$ and $\theta' = 0.4$. Subsequently, the two sets of U training data were sampled from those distributions with sample sizes $n = 30000$ and $n' = 30000$.

The results demonstrate the concurrence of empirical training risk going negative (blue dashed line) and the test accuracy overfitting (green dashed line) regardless of datasets, models, optimizers and loss functions.

C Supplementary information on the experiments

MNIST (LeCun et al., 1998) This is a grayscale image dataset of handwritten digits from 0 to 9 where the size of the images is 28*28. It contains 60,000 training images and 10,000 test images. See http://yann.lecun.com/exdb/mnist/ for details. Since it has 10 classes originally, we used the even digits as the P class and the odd digits as the N class, respectively.

The simple model used for training MNIST was a linear-in-input model $g(x) = \boldsymbol{\omega}^T x + b$ where $\boldsymbol{\omega} \in \mathbb{R}^{784}$ and $b \in \mathbb{R}$ with ℓ_2-regularization (the regularization parameter was fixed to be $1e-4$). The batch size and learning rate were set to be 3000 and $1e-3$ respectively. The deep model uased was a 5-layer FC with ReLU as the activation function: d-300-300-300-300-1 with ℓ_2-regularization (the regularization parameter was fixed to be $5e-3$). The batch size and learning rate were set to be 3000 and $5e-5$ respectively. For both models, batch normalization (Ioffe and Szegedy, 2015) with the default *momentum* $= 0.99$ and $\epsilon = 1e-3$ was applied before hidden layers, and the model was trained by Adam with the default momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For all the experiments, the generalized leaky ReLU hyperparameter λ was selected from -1 to 0.01.

Fashion-MNIST (Xiao et al., 2017) This is also a grayscale fashion image dataset similarly to MNIST, but here each data is associated with a label from 10 fashion item classes. See https://github.com/zalandoresearch/fashion-mnist for details. It was converted into a binary classification dataset as follows:

- the P class is formed by 'T-shirt', 'Trouser', 'Shirt', and 'Sneaker';
- the N class is formed by 'Pullover', 'Dress', 'Coat', 'Sandal', 'Bag', and 'Ankle boot'.

The models and optimizers were the same as MNIST, where the learning rate for the simple and deep models were set to be $5e-3$ and $3e-5$ and the other hyperparameters remain the same.

Kuzushiji-MNIST (Clanuwat et al., 2018) This is another variant of MNIST dataset consisting of 60,000 training images and 10,000 test images of cursive Japanese (Kuzushiji) characters. See https://github.com/rois-codh/kmnist for details. For Kuzushi-MNIST dataset,

- 'ki', 're', 'wo' made up the P class;
- 'o', 'su', 'tsu', 'na', 'ha', 'ma', 'ya' made up the N class.

The models and optimizers were the same as MNIST, where the learning rate for the deep models was set to be $3e-5$ and the other hyperparameters remain the same.

CIFAR-10 (Krizhevsky, 2009) This dataset consists of 60,000 32∗32 color images in 10 classes, and there are 5,000 training images and 1,000 test images per class. See https://www.cs.toronto.edu/~kriz/cifar.html for details. For CIFAR-10 dataset,

- the P class is composed of 'bird', 'cat', 'deer', 'dog', 'frog' and 'horse';
- the N class is composed of 'airplane', 'automobile', 'ship' and 'truck'.

The simple model used for training CIFAR-10 was also a linear-in-input model $g(x) = \boldsymbol{\omega}^T x + b$ where $\boldsymbol{\omega} \in \mathbb{R}^{3072}$ and $b \in \mathbb{R}$ with ℓ_2-regularization (the regularization parameter was fixed to be $5e-3$). The batch size and learning rate were set to be 3000 and $5e-3$ respectively. The deep model was again ResNet-32 (He et al., 2016) that can be find in Appendix B. The batch size and learning rate were set to be 3000 and $3e-5$ respectively. For both models, batch normalization (Ioffe and Szegedy, 2015) with the default $momentum = 0.99$ and $\epsilon = 1e-3$ was applied before hidden layers, and the model was trained by Adam with the default momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

D Supplementary experiments on general-purpose regularization

Regularization is the most common technique that lower the complexity of a neural network model during training, and thus prevent the overfitting (Goodfellow et al., 2016). In this section, we demonstrate that the general-purpose regularization methods fail to mitigate the overfitting in UU classification scenario.

We tested the unbiased UU method using two most popular regularization techniques, i.e., dropout and weight decay. The dataset and model used were again MNIST (the class priors θ and θ' were set to be 0.6 and 0.4) and the 5-layer MLP d-300-300-300-300-1 with ℓ_2-regularization, where dropout layers were added between the existing layers. The optimizer was SGD with momentum ($momentum = 0.9$) and logistic loss. Batch size was fixed to be 3000 and the initial learning rate was $1e-3$. For dropout experiments, we fixed the weight decay parameter to be $1e-4$ and change the dropout parameter from 0 to 0.8. For weight decay experiments, we fix the dropout parameter to be 0.2 and change the weight decay parameter from 0.0005 to 5.

Empirical results in Figure 3 show that the unbiased UU method with slightly strong regularizations outperforms the one with weaker regularizations, but still suffers from overfitting. It is because adding strong regularization may prohibit the high representation power of deep models, which in turn may cause underfitting.

Discussion Instead of the general-purpose regularization, our proposed method explicitly utilizes the additional knowledge that the empirical risk goes negative. By that, we can more "effectively" constrain the model without too much sacrificing the representation power of deep models. Note that our correction can also be regarded as a regularization in its general sense for fighting against overfitting, but differently from weight decay or dropout, it is exclusively designed for UU classification and hence it is no surprising that our regularization fits UU classification better than other regularizations as discussed in Sec. 3 theoretically and demonstrated in Sec. 4 empirically.

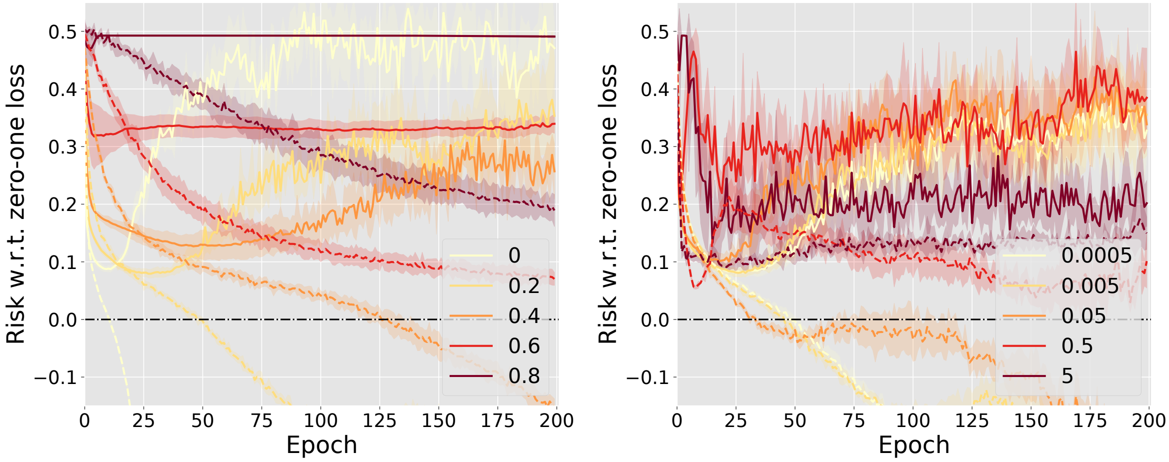

Figure 3: Supplementary experimental results on general regularization. Left: dropout. Right: weight decay. Solid curves are $\widehat{R}_{\mathrm{uu}}(g)$ on test data and dashed curves are $\widehat{R}_{\mathrm{uu}}(g)$ on training data.