POINTWISE BINARY CLASSIFICATION WITH PAIRWISE CONFIDENCE COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Binary classification generally requires pointwise labeled data. But can we learn a binary classifier *without any pointwise data*? In this paper, we show that with *only pairwise comparison data*, i.e., pairs of unlabeled data that we know one is more likely to be positive than the other, we can still successfully learn a binary classifier. We name this problem setting *pairwise comparison (Pcomp) classification*, which is useful to handle, e.g., private classification tasks where we are not allowed to directly access labels and subjective classification tasks where labelers have different labeling standards. For the Pcomp classification problem, we present a mathematical formulation for the generation process of pairwise comparison data, based on which we explore an *unbiased risk estimator* (URE) to train a binary classifier by *empirical risk minimization* and establish an *estimation error bound*. We first prove that an URE can be obtained and improve it using *correction functions*. Then, we start from the *noisy-label learning* perspective to introduce a *progressive* URE and improve it by imposing *consistency regularization*. Finally, experiments demonstrate the effectiveness of our proposed methods.

1 INTRODUCTION

Traditional supervised learning techniques have achieved great advances, while they are demanding for precisely labeled data. In many real-world scenarios, it may be too difficult to collect such data. To alleviate this problem, a large number of weakly supervised learning problems (Zhou, 2018) have been extensively studied, including *semi-supervised learning* (Zhu & Goldberg, 2009; Niu et al., 2013; Sakai et al., 2018), *multi-instance learning* (Zhou et al., 2009; Sun et al., 2016; Zhang & Zhou, 2017), *noisy-label learning* (Han et al., 2018; Xia et al., 2019; Wei et al., 2020), *partial-label learning* (Zhang et al., 2017; Feng & An, 2019; Lv et al., 2020), *complementary-label learning* (Ishida et al., 2017; Yu et al., 2018; Ishida et al., 2019; Feng et al., 2020), *positive-unlabeled classification* (Gong et al., 2019), *positive-confidence classification* (Ishida et al., 2018), *similar-unlabeled classification* (Bao et al., 2018), and *unlabeled-unlabeled classification* (Lu et al., 2019; 2020), and *triplet classification* (Cui et al., 2020).

This paper considers another novel weakly supervised learning setting called *pairwise comparison (Pcomp) classification*, where we aim to perform binary classification from only *pairwise comparison data*, instead of pointwise labeled data. A pairwise comparison $(\boldsymbol{x}, \boldsymbol{x}')$ represents that the instance $\boldsymbol{x}$ has a larger probability of belonging to the positive class than the instance $\boldsymbol{x}'$. Such pairwise comparison data could be used to deal with the binary classification problem regarding sensitive or private matters. For example, when collecting some sensitive or private data, it would be prohibited or illegal to ask for the true label. In this case, we may collect other indirect labeling information like the pairwise comparison between two examples.

In addition, for binary classification with class overlapping, the labeling task becomes difficult, and even experienced labelers could provide noisy labels. Faced with such a difficult labeling task, different labelers may hold different labeling standards and provide different labels for the same instance. Let us denote the labeling standard of a labeler as $\tilde{p}(y|\boldsymbol{x})$. Even if an instance $\boldsymbol{x}_1$ is more positive than $\boldsymbol{x}_2$, different labelaers may give different labels depending on their labeling standards: $\tilde{p}(y=+1|\boldsymbol{x}_1) > \tilde{p}(y=+1|\boldsymbol{x}_2) > 1/2$, $\tilde{p}(y=+1|\boldsymbol{x}_1) > 1/2 > \tilde{p}(y=+1|\boldsymbol{x}_2)$, $1/2 > \tilde{p}(y=+1|\boldsymbol{x}_1) > \tilde{p}(y=+1|\boldsymbol{x}_2)$. In other words, different labelers provide the same pairwise comparison of the probabilities of two examples belonging to the positive class, even though they may provide

different labels for the same instance. This case also demonstrates a potential application of our proposed Pcomp classification.

Our contributions in this paper can be summarized as follows:

- We propose Pcomp classification, a novel weakly supervised learning setting, and present a mathematical formulation for the generation process of pairwise comparison data.
- We prove that an *unbiased risk estimator* (URE) can be obtained, propose an *empirical risk minimization* (ERM) based method, and present an improvement using correction functions (Lu et al., 2020) for alleviating overftting when complex models are used.
- We start from the noisy-label learning perspective to introduce the *RankPruning* method (Northcutt et al., 2017) that holds a *progressive* URE for solving our proposed Pcomp classification problem and improve it by imposing *consistency regularization*.
- We experimentally demonstrate the effectiveness of our proposed methods.

2 PRELIMINARIES

In this section, we briefly introduce some notations and review the related problem formulations of binary classification, positive-unlabeled learning, unlabeled-unlabeled learning.

Binary Classification. Since this paper focuses on how to train a binary classifier from pairwise comparison data, we first review the problem formulation of binary classification. Let the feature space be $\mathcal{X}$ and the label space be $\mathcal{Y} = \{+1, -1\}$. Suppose the collected dataset is denoted by $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where each example $(\boldsymbol{x}_i, y_i)$ is independently sampled from the joint distribution with density $p(\boldsymbol{x}, y)$, which includes an instance $\boldsymbol{x}_i \in \mathcal{X}$ and a label $y_i \in \mathcal{Y}$. The goal of binary classification is to train an optimal classifier $f : \mathcal{X} \mapsto \mathbb{R}$ by minimizing the following expected classification risk:

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\big[\ell(f(\boldsymbol{x}), y)\big] = \pi_+\mathbb{E}_{p_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), +1)\big] + \pi_-\mathbb{E}_{p_-(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), -1)\big], \qquad (1)$$

where $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ denotes a binary loss function, $\pi_+ := p(y = +1)$ (or $\pi_- := p(y = -1)$) denotes the *positive* (or *negative*) *class prior probability*, and $p_+(\boldsymbol{x}) := p(\boldsymbol{x}|y = +1)$ (or $p_-(\boldsymbol{x}) := p(\boldsymbol{x}|y = -1)$) denotes the *class-conditional probability density* of the positive (or negative) data. ERM approximates the expectations over $p_+(x)$ and $p_-(x)$ by the empirical averages of positive and negative data and the empirical risk is minimized with respect to the classifier f.

Positive-Unlabeled (PU) Classification. In some real-world scenarios, it may be difficult to collect negative data, and only positive (P) and unlabeled (U) data are available. PU classification aims to train an effective binary classifier in this weakly supervised setting. Previous studies (du Plessis et al., 2014; 2015; Kiryo et al., 2017) showed that the classification risk $R(f)$ in Eq. (1) can be rewritten only in terms of positive and unlabeled data as

$$R(f) = R_{\mathrm{PU}}(f) = \pi_+\mathbb{E}_{p_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), +1) - \ell(f(\boldsymbol{x}), -1)\big] + \mathbb{E}_{p(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), -1)\big], \qquad (2)$$

where $p(\boldsymbol{x}) = \pi_+ p_+(\boldsymbol{x}) + \pi_- p_-(\boldsymbol{x})$ denotes the probability density of unlabeled data. This risk expression immediately allows us to employ ERM in terms of positive and unlabeled data.

Unlabeled-Unlabeled (UU) Classification. The recent studies (Lu et al., 2019; 2020) showed that it could be possible to train a binary classifier only from two unlabeled datasets with different class priors. Lu et al. (2019) showed that the classification risk can be rewritten as

$$\begin{aligned} R(f) = R_{\mathrm{UU}}(f) = \mathbb{E}_{p_{\mathrm{tr}}(\boldsymbol{x})}\Big[\frac{(1-\theta')\pi_+}{\theta-\theta'}\ell(f(\boldsymbol{x}), +1) - \frac{\theta'(1-\pi_+)}{\theta-\theta'}\ell(f(\boldsymbol{x}), -1)\Big] \\ + \mathbb{E}_{p_{\mathrm{tr}'}(\boldsymbol{x}')}\Big[\frac{\theta(1-\pi_+)}{\theta-\theta'}\ell(f(\boldsymbol{x}'), -1) - \frac{(1-\theta)\pi_+}{\theta-\theta'}\ell(f(\boldsymbol{x}'), +1)\Big], \qquad (3) \end{aligned}$$

where θ and θ' are different class priors of two unlabeled datasets, and $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{tr}'}(\boldsymbol{x}')$ are the densities of two datasets of unlabeled data, respectively. This risk expression immediately allows us to employ ERM only from two sets of unlabeled data. For $R_{\mathrm{UU}}(f)$ in Eq. (3), if we set $\theta = 1$, $\theta' = \pi_+$, and replace $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{tr}'}(\boldsymbol{x}')$ by $p_+(\boldsymbol{x})$ and $p(\boldsymbol{x})$ respectively, then we can recover $R_{\mathrm{PU}}(f)$ in Eq. (2). Therefore, UU classification could be taken as a generalized framework of PU

classification in terms of the unbiased risk estimator. Besides, Eq. (3) also recovers a complicated unbiased risk estimator of similar-unlabeled classification (Bao et al., 2018) by setting $\theta = \pi_+$ and $\theta' = \pi_+^2/(2\pi_+^2 - 2\pi_+ + 1)$.

To solve our proposed Pcomp classification problem, we will present a mathematical formulation for the generation process of pairwise comparison data, based on which we will explore two UREs to train a binary classifier by ERM and establish the corresponding *estimation error bounds*.

3 Data Generation Process

In order to derive UREs for performing ERM, we first formulate the underlying generation process of pairwise comparison data[1], which consists of pairs of unlabeled data that we know which one is more likely to be positive. Suppose the provided dataset is denoted by $\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \boldsymbol{x}_i')\}_{i=1}^n$ where $(\boldsymbol{x}_i, \boldsymbol{x}_i')$ (with their unknown true labels (y_i, y_i')) is expected to satisfy $p(y_i = +1|\boldsymbol{x}_i) > p(y_i' = +1|\boldsymbol{x}_i')$.

It is clear that we could easily collect pairwise comparison data if the positive confidence (i.e., $p(y = +1|\boldsymbol{x})$) of each instance could be obtained. However, such information is much harder to be obtained than full labels in real-world scenarios. Therefore, unlike some studies (Ishida et al., 2018; Shinoda et al., 2020) that assume the positive confidence is known, we only assume that the labeling system has access to the full labels. Specifically, we adopt the assumption (Cui et al., 2020) that weakly supervised examples are first sampled from the true data distribution, but the labels are only accessible by the labeling system, not us. Then, the labeling system would provide us weakly supervised information (i.e., pairwise comparison information) according to the received labels. We could infer that the collected labels (y, y') of our defined pairwise comparison data $(\boldsymbol{x}, \boldsymbol{x}')$ only appear to be one of the three cases $\{(+1, -1), (+1, +1), (-1, -1)\}$, as the condition $p(y = +1|\boldsymbol{x}) \geq p(y' = +1|\boldsymbol{x}')$ is definitely violated if $(y, y') = (-1, +1)$. Therefore, the labeling system would take $(\boldsymbol{x}, \boldsymbol{x}')$ as a pairwise comparison example in the dataset $\widetilde{\mathcal{D}}$, if the collected labels (y, y') of $(\boldsymbol{x}, \boldsymbol{x}')$ belong to one of the three cases. Based on the above descriptions, we have the following theorem.

Theorem 1. *According to the generation process of pairwise comparison data described above, let*

$$\widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') = \frac{q(\boldsymbol{x}, \boldsymbol{x}')}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-}, \tag{4}$$

where $q(\boldsymbol{x}, \boldsymbol{x}') = \pi_+^2 p_+(\boldsymbol{x})p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x})p_-(\boldsymbol{x}') + \pi_+\pi_- p_+(\boldsymbol{x})p_-(\boldsymbol{x}')$. *Then we have* $\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \boldsymbol{x}_i')\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}')$.

The proof is provided in Appendix A. Theorem 1 provides an explicit expression of the probability density of pairwise comparison data.

Next, we would like to exract pointwise information from pairwise information, since our goal is to perform pointwise binary classification. Let $\widetilde{\pi} = \pi_+^2 + \pi_-^2 + \pi_+\pi_- = \pi_+ + \pi_-^2 = \pi_+^2 + \pi_-$ and we denote the pointwise data collected from $\widetilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \boldsymbol{x}_i')\}_{i=1}^n$ by breaking the pairwise comparison relation as $\widetilde{\mathcal{D}}_+ = \{\boldsymbol{x}_i\}_{i=1}^n$ and $\widetilde{\mathcal{D}}_- = \{\boldsymbol{x}_i'\}_{i=1}^n$. Then we can obtain the following theorem.

Theorem 2. *Pointwise examples in* $\widetilde{\mathcal{D}}_+$ *and* $\widetilde{\mathcal{D}}_-$ *are independently drawn from* $\widetilde{p}_+(\boldsymbol{x})$ *and* $\widetilde{p}_-(\boldsymbol{x}')$, *where*

$$\widetilde{p}_+(\boldsymbol{x}) = \frac{\pi_+}{\pi_-^2 + \pi_+} p_+(\boldsymbol{x}) + \frac{\pi_-^2}{\pi_-^2 + \pi_+} p_-(\boldsymbol{x}), \quad \widetilde{p}_-(\boldsymbol{x}') = \frac{\pi_+^2}{\pi_+^2 + \pi_-} p_+(\boldsymbol{x}') + \frac{\pi_-}{\pi_+^2 + \pi_-} p_-(\boldsymbol{x}').$$

The proof is provided in Appendix B. Theorem 2 shows the relationships between the pointwise densities and the class-conditional densities. Besides, it indicates that from pairwise comparison data, we can essentially obtain examples that can be drawn independently from $\widetilde{p}_+(\boldsymbol{x})$ and $\widetilde{p}_-(\boldsymbol{x}')$.

[1]Unlike Xu et al. (2019) who utilized pairwise comparison data to perform uncoupled regression, we focus on binary classification.

4 THE PROPOSED METHODS

In this section, we explore two UREs to train a binary classifier by ERM from only pairwise comparison data with the above generation process.

4.1 CORRECTED PCOMP CLASSIFICATION

As shown in Eq. (1), the classification risk $R(f)$ could be separately expressed as the expectations over $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$. Although we do not have access to the two class-conditional densities $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$, we can represent them by our introduced pointwise densities $\widetilde{p}_+(\boldsymbol{x})$ and $\widetilde{p}_-(\boldsymbol{x})$.

Lemma 1. *We can express $p_+(\boldsymbol{x})$ and $p_-(\boldsymbol{x})$ in terms of $\widetilde{p}_+(\boldsymbol{x})$ and $\widetilde{p}_+(\boldsymbol{x})$ as*

$$p_+(\boldsymbol{x}) = \frac{1}{\pi_+}\big(\widetilde{p}_+(\boldsymbol{x}) - \pi_-\widetilde{p}_-(\boldsymbol{x})\big), \quad p_-(\boldsymbol{x}) = \frac{1}{\pi_-}\big(\widetilde{p}_-(\boldsymbol{x}) - \pi_+\widetilde{p}_+(\boldsymbol{x})\big).$$

The proof is provided in Appendix C. As a result of Lemma 1, we can express the classification risk $R(f)$ using only pairwise comparison data sampled from $\widetilde{p}_+(\boldsymbol{x})$ and $\widetilde{p}_-(\boldsymbol{x})$.

Theorem 3. *The classification risk $R(f)$ can be equivalently expressed as*

$$R_{\mathrm{PC}}(f) = \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1) - \pi_+\ell(f(\boldsymbol{x}),-1)\big] + \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}'),-1) - \pi_-\ell(f(\boldsymbol{x}'),+1)\big]. \tag{5}$$

The proof is provided in Appendix D. In this way, we could train a binary classifier by minimizing the following empirical approximation of $R_{\mathrm{PC}}(f)$:

$$\widehat{R}_{\mathrm{PC}}(f) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\Big(\ell(f(\boldsymbol{x}_i),+1) - \pi_+\ell(f(\boldsymbol{x}_i),-1) + \ell(f(\boldsymbol{x}_i'),-1) - \pi_-\ell(f(\boldsymbol{x}_i'),+1)\Big). \tag{6}$$

Estimation Error Bound. Here, we establish an estimation error bound for the proposed unbiased risk estimator. Let $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$ be the model class, $\widehat{f}_{\mathrm{PC}} = \arg\min_{f\in\mathcal{F}} \widehat{R}_{\mathrm{PC}}(f)$ be the empirical risk minimizer, and $f^\star = \arg\min_{f\in\mathcal{F}} R(f)$ be the true risk minimizer. Let $\widetilde{\mathfrak{R}}_n^+(\mathcal{F})$ and $\widetilde{\mathfrak{R}}_n^-(\mathcal{F})$ be the *Rademacher complexities* (Bartlett & Mendelson, 2002) of $\mathcal{F}$ with sample size n over $\widetilde{p}_+(\boldsymbol{x})$ and $\widetilde{p}_-(\boldsymbol{x})$ respectively.

Theorem 4. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 \leq \rho \leq \infty$), and all functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a positive constant C_{b} such that $\|f\| \leq C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{z\leq C_{\mathrm{b}}, t=\pm1}\ell(z,t)$. Then for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$R(\widehat{f}_{\mathrm{PC}}) - R(f^\star) \leq (1+\pi_+)4\rho\widetilde{\mathfrak{R}}_n^+(\mathcal{F}) + (1+\pi_-)4\rho\widetilde{\mathfrak{R}}_n^-(\mathcal{F}) + 6C_\ell\sqrt{\frac{\log\frac{8}{\delta}}{2n}}.$$

The proof is provided in Appendix E. Theorem 4 shows that our proposed method is consistent, i.e., as $n \to \infty$, $R(\widehat{f}_{\mathrm{PC}}) \to R(f^\star)$, since $\widetilde{\mathfrak{R}}_n^+(\mathcal{F})$, $\widetilde{\mathfrak{R}}_n^-(\mathcal{F}) \to 0$ for all parametric models with a bounded norm (Golowich et al., 2017; Lu et al., 2019).

Relation to UU Classification. It is worth noting that the unbiased risk estimator of UU classification $R_{\mathrm{UU}}(f)$ is quite general for binary classification with weak supervision. Hence we also would like to show the relationships between our proposed estimator $R_{\mathrm{PC}}(f)$ and $R_{\mathrm{UU}}(f)$. We demonstrate by the following corollary that under some conditions, $R_{\mathrm{UU}}(f)$ is equivalent to $R_{\mathrm{PC}}(f)$.

Corollary 1. *By setting $p_{\mathrm{tr}} = \widetilde{p}_+(\boldsymbol{x})$, $p'_{\mathrm{tr}} = \widetilde{p}_-(\boldsymbol{x})$, $\theta = \pi_+/(1-\pi_+ + \pi_+^2)$, and $\theta' = \pi_+^2/(1-\pi_+ + \pi_+^2)$, Eq. (3) is equivalent to Eq. (5), which means that $R_{\mathrm{UU}}(f)$ is equivalent to $R_{\mathrm{PC}}(f)$.*

We omit the proof of Corollary 1 since it is straightforward to derive Eq. (5) from Eq. (3) by inserting the required notations. As shown in Lu et al. (2020), directly minimizing $\widehat{R}_{\mathrm{PC}}$ would suffer from overfitting when complex models are used due to the negative risk issue. More specifically, since negative terms are included in Eq. (6), the empirical risk can be negative even though the original true risk can never be negative. To prevent from overfitting, they wrapped the terms in $\widehat{R}_{\mathrm{UU}}(f)$ that

cause a negative empirical risk by certain *consistent correction functions* such as the rectified linear unit (ReLU) function and absolute value function. This solution could also be applied to $\widehat{R}_{\mathrm{PC}}$. In this way, we could obtain the following corrected empirical risk estimator:

$$\widehat{R}_{\mathrm{cPC}}(f) = g\Big(\frac{1}{n}\sum\nolimits_{i=1}^{n}\big(\ell(f(\boldsymbol{x}_i),+1) - \pi_-\ell(f(\boldsymbol{x}_i'),+1)\big)\Big) + g\Big(\frac{1}{n}\sum\nolimits_{i=1}^{n}\big(\ell(f(\boldsymbol{x}_i'),-1) - \pi_+\ell(f(\boldsymbol{x}_i),-1)\big)\Big), \quad (7)$$

where $g:\mathbb{R}\mapsto\mathbb{R}$ denotes a consistent correction function (Lu et al., 2020).

4.2 Progressive Pcomp Classification

Here, we start from the noisy-label learning perspective to solve the Pcomp classification problem. Intuitively, we could simply perform binary classification by regarding the data from $\widetilde{p}_+(\boldsymbol{x})$ as (noisy) positive data and the data from $\widetilde{p}_-(\boldsymbol{x})$ as (noisy) negative data. However, this naive solution could be inevitably affected by noisy labels. In this scenario, we denote the noise rates as $\rho_- = p(\widetilde{y}=+1|y=-1)$ and $\rho_+ = p(\widetilde{y}=-1|y=+1)$, where $\widetilde{y}$ is the observed (noisy) label and y is the true label, and the inverse noise rates as $\phi_+ = p(y=-1|\widetilde{y}=+1)$ and $\phi_- = p(y=+1|\widetilde{y}=-1)$. According to the defined generation process of pairwise comparison data, we have the following theorem.

Theorem 5. *The following equalities hold:*

$$\phi_+ = \frac{\pi_-^2}{\pi_+^2+\pi_-^2+\pi_+\pi_-}, \quad \phi_- = \frac{\pi_+^2}{\pi_+^2+\pi_-^2+\pi_+\pi_-},$$
$$\rho_+ = \frac{\pi_+}{2(\pi_+^2+\pi_-^2+\pi_+\pi_-)}, \quad \rho_- = \frac{\pi_-}{2(\pi_+^2+\pi_-^2+\pi_+\pi_-)}.$$

The proof is provided in Appendix F.

By regarding the Pcomp classification problem as the noisy-label learning problem with known noise rates, we could obtain UREs (Natarajan et al., 2013; Northcutt et al., 2017). We adopt the RankPruning method (Northcutt et al., 2017) because it holds a progressive URE by selecting confident examples using the learning model and achieves state-of-the-art performance. Specifically, we denote by the dataset composed of all the observed positive data $\widetilde{\mathcal{P}}$, i.e., $\widetilde{\mathcal{P}} = \{\boldsymbol{x}_i\}_{i=1}^n$, where $\boldsymbol{x}_i$ is independently sampled from $\widetilde{p}_+(\boldsymbol{x})$. Similarly, the dataset composed of all the observed negative data is denoted by $\widetilde{\mathcal{N}}$, i.e., $\widetilde{\mathcal{N}} = \{\boldsymbol{x}_i'\}_{i=1}^n$, where $\boldsymbol{x}_i'$ is independently sampled from $\widetilde{p}_-(\boldsymbol{x}')$. Then, confident examples will be selected from $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{N}}$ by ranking the outputs of the model f. We denote the selected positive data from $\widetilde{\mathcal{P}}$ as $\mathcal{P}\widetilde{\mathcal{P}}$, and the selected negative data from $\widetilde{\mathcal{N}}$ as $\mathcal{N}\widetilde{\mathcal{N}}$:

$$\mathcal{P}\widetilde{\mathcal{P}} = \underset{\mathcal{P}:|\mathcal{P}|=(1-\phi_+)|\widetilde{\mathcal{P}}|}{\arg\max}\sum\nolimits_{\boldsymbol{x}\in\{\mathcal{P}\cap\widetilde{\mathcal{P}}\}} f(\boldsymbol{x}), \quad \mathcal{N}\widetilde{\mathcal{N}} = \underset{\mathcal{N}:|\mathcal{N}|=(1-\phi_-)|\widetilde{\mathcal{N}}|}{\arg\min}\sum\nolimits_{\boldsymbol{x}\in\{\mathcal{N}\cap\widetilde{\mathcal{N}}\}} f(\boldsymbol{x}).$$

Then we show that if the model f satisfies the *separability condition*, i.e., the output of every true positive example is always larger than that of every true negative example, we could obtain a URE. We name it progressive URE, as the model f is progressively optimized.

Theorem 6 (Theorem 5 in (Northcutt et al., 2017))**.** *Assume that the model f satisfies the above separability condition, then the classification risk $R(f)$ can be equivalently expressed as*

$$R_{\mathrm{pPC}}(f) = \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\Big[\frac{\ell(f(\boldsymbol{x}),+1)}{1-\rho_+}\mathbb{I}[\boldsymbol{x}\in\mathcal{P}\widetilde{\mathcal{P}}]\Big] + \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\Big[\frac{\ell(f(\boldsymbol{x}'),-1)}{1-\rho_-}\mathbb{I}[\boldsymbol{x}'\in\mathcal{N}\widetilde{\mathcal{N}}]\Big],$$

where $\mathbb{I}[\cdot]$ is the indicator function.

In this way, we have the following empirical approximation of R_{pPC}:

$$\widehat{R}_{\mathrm{pPC}}(f) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\Big(\frac{\ell(f(\boldsymbol{x}_i),+1)}{1-\rho_+}\mathbb{I}[\boldsymbol{x}_i\in\mathcal{P}\widetilde{\mathcal{P}}] + \frac{\ell(f(\boldsymbol{x}_i'),-1)}{1-\rho_-}\mathbb{I}[\boldsymbol{x}_i'\in\mathcal{N}\widetilde{\mathcal{N}}]\Big). \quad (8)$$

Estimation Error Bound. It worth noting that Northcutt et al. (2017) did not prove the learning consistency for the RankPruning method. Here, we establish an estimation error bound for this method, which guarantees the learning consistency. Let $\widehat{f}_{\mathrm{pPC}} = \arg\min_{f\in\mathcal{F}}\widehat{R}_{\mathrm{pPC}}(f)$ be the empirical risk minimizer of the RankPruning method, then we have the following theorem.

Theorem 7. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 \leq \rho \leq \infty$), and all all functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a positive constant C_{b} such that $\|f\| \leq C_{\mathrm{b}}$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{z \leq C_{\mathrm{b}}, t=\pm 1} \ell(z,t)$. Then for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$R(\widehat{f}_{\mathrm{pPC}}) - R(f^\star) \leq \frac{2}{1-\rho_+}\Big(2\rho\widetilde{\mathfrak{R}}_n^+(\mathcal{F}) + C_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2n}}\Big) + \frac{2}{1-\rho_-}\Big(2\rho\widetilde{\mathfrak{R}}_n^-(\mathcal{F}) + C_\ell\sqrt{\frac{\log\frac{4}{\delta}}{2n}}\Big).$$

The proof is provided in Appendix G. Theorem 7 shows that the above method is consistent, i.e., as $n \to \infty$, $R(\widehat{f}_{\mathrm{pPC}}) \to R(f^\star)$, since $\widetilde{\mathfrak{R}}_n^+(\mathcal{F}), \widetilde{\mathfrak{R}}_n^-(\mathcal{F}) \to 0$ for all parametric models with a bounded norm (Golowich et al., 2017; Lu et al., 2019).

Regularization. For the above RankPruning method, its URE is based on the assumption that the learning model could satisfy the separability condition. Thus, its performance heavily depends on the accuracy of the learning model. However, as the learning model is progressively updated, some of the selected confident examples may still contain label noise during the training process. As a result, the RankPruning method would be affected by incorrectly selected data. A straightforward improvement could be to improve the output quality of the learning model. Motivated by Mean Teacher used in semi-supervised learning (Tarvainen & Valpola, 2017), we also resort to a teacher model that is an exponential moving average of model snapshots, i.e., $\boldsymbol{\Theta}'_t = \alpha\boldsymbol{\Theta}'_{t-1} + (1-\alpha)\boldsymbol{\Theta}_t$, where $\boldsymbol{\Theta}'$ denotes the parameters of the teacher model, $\boldsymbol{\Theta}$ denotes the parameters of the learning model, the subscript t denotes the training step, and α is a smoothing coefficient hyper-parameter. Such a teacher model could guide the learning model to produce high-quality outputs. To learn from the teacher model, we leverage consistency regularization (Laine & Aila, 2016; Tarvainen & Valpola, 2017) $\Omega(f) = \mathbb{E}_{\boldsymbol{x}}\big[\|f_{\boldsymbol{\Theta}}(\boldsymbol{x}) - f_{\boldsymbol{\Theta}'}(\boldsymbol{x})\|^2\big]$ to make the learning model consistent with the teacher model for improving the RankPruning method.

5 Experiments

In this section, we conduct experiments to evaluate the practical performance of our proposed methods on various benchmark datasets.

Datasets. We employ four widely used benchmark datasets, including MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), Kuzushiji-MNIST (Clanuwat et al., 2018), and CIFAR-10 (Krizhevsky et al., 2009). The detailed descriptions of these datasets are provided in Appendix. Since these datasets are specially used for multi-class classification, we manually transformed them into binary classification datasets (see Appendix for details).

Methods. We propose the following methods to solve the Pcomp classification problem:

- Pcomp-Unbiased, which denotes the proposed method that minimizes $\widehat{R}_{\mathrm{pc}}(f)$ in Eq. (6).
- Pcomp-ReLU, which denotes the proposed method that minimizes $\widehat{R}_{\mathrm{cpc}}(f)$ in Eq. (7) with the ReLU function.
- Pcomp-ABS, which denotes the proposed method that minimizes $\widehat{R}_{\mathrm{cpc}}(f)$ in Eq. (7) with the absolute value function.
- Pcomp-Teacher, which improves the RankPruning method by imposing regularization on the learning model to boost the consistency with the mean teacher model.

We compare with the following baselines:

- Pcomp-Biased, which conducts binary classification by regarding the data from $\widetilde{p}_+(\boldsymbol{x})$ as positive data and the data from $\widetilde{p}_-(\boldsymbol{x})$ as negative data. This is a straightforward method to handle the Pcomp classification problem. In this setting, Pcomp-Biased reduces to the BER minimization method (Menon et al., 2015).
- Noisy-Unbiased, which is a noisy-label learning method that minimizes the empirical approximation of the URE proposed by Natarajan et al. (2013).
- RankPruning, which is a noisy-label learning method (Northcutt et al., 2017) that minimizes $\widehat{R}_{\mathrm{ppc}}(f)$ in Eq. (8).

- Supervised, which conducts binary classification with pointwise labeled data.

Experimental setup. We could exactly estimate $\widetilde{\pi}$ by counting the fraction of collected pairwise comparison data in all the sampled paired data. Since $\widetilde{\pi} = \pi_+^2 + \pi_- = \pi_+^2 + 1 - \pi_+$, we have $\pi_+ = 1/2 - \sqrt{\widetilde{\pi} - 3/4}$ (if $\pi_+ < 1/2$) or $\pi_+ = 1/2 + \sqrt{\widetilde{\pi} - 3/4}$ (if $\pi_+ \geq 1/2$). Hence we could exactly estimate the true class prior π_+ if we know whether π_+ is larger than π_-. For simplicity, we assume that the class prior is known for all the methods. We test the performance of the above methods under different class prior settings, i.e., π_+ is set to $0.2, 0.5, 0.8$.

6 Conclusion

In this paper, we proposed a novel weakly supervised learning setting called *pairwise comparison (Pcomp) classification*, where we aim to train a binary classifier from only *pairwise comparison data*, i.e., two examples that we know one is more likely to be positive than the other instead of pointwise labeled data. Pcomp classification is useful to handle, e.g., private classification tasks where we are not allowed to directly access labels and subjective classification tasks where labelers have different labeling standards. To solve this problem, we presented a mathematical formulation for the generation process of pairwise comparison data, based on which we explored two *unbiased risk estimators* (UREs) to train a binary classifier by *empirical risk minimization* and established the corresponding *estimation error bounds*. We first proved that an URE can be obtained and improved it using correction functions. Then, we started from the *noisy-label learning* perspective to introduce a *progressive* URE and improved it by imposing *consistency regularization*. Finally, experiments demonstrated the effectiveness of our proposed methods. In future work, we will apply Pcomp classification to solve some challenging real-world problems such as binary classification with privacy matters or class overlapping.

REFERENCES

Han Bao, Gang Niu, and Masashi Sugiyama. Classification from pairwise similarity and unlabeled data. In *ICML*, pp. 452–461, 2018.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(11):463–482, 2002.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Zheng-Hang Cui, Nontawat Charoenphakdee, Issei Sato, and Masashi Sugiyama. Classification from triplet comparison data. *Neural Computation*, 32(3):659–681, 2020.

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *NeurIPS*, pp. 703–711, 2014.

Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *ICML*, pp. 1386–1394, 2015.

Lei Feng and Bo An. Partial label learning with self-guided retraining. In *AAAI*, pp. 3542–3549, 2019.

Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *ICML*, pp. in press, 2020.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

Chen Gong, Hong Shi, Tong-Liang Liu, Chuang Zhang, Jian Yang, and Da-Cheng Tao. Loss decomposition and centroid estimation for positive and unlabeled learning. *TPAMI*, 2019.

Bo Han, Quan-Ming Yao, Xing-Rui Yu, Gang Niu, Miao Xu, Wei-Hua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pp. 8527–8537, 2018.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *NeurIPS*, pp. 5644–5654, 2017.

Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification for positive-confidence data. In *NeurIPS*, pp. 5917–5928, 2018.

Takashi Ishida, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *ICML*, pp. 2971–2980, 2019.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, pp. 1674–1684, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Nan Lu, Gang Niu, Aditya K. Menon, and Masashi Sugiyama. On the minimal supervision for training any binary classifier from only unlabeled data. In *ICLR*, 2019.

Nan Lu, Tian-Yi Zhang, Gang Niu, and Masashi Sugiyama. Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In *AISTATS*, 2020.

Jia-Qi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *ICML*, 2020.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *ICML*, pp. 125–134, 2015.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NeurIPS*, pp. 1196–1204, 2013.

Gang Niu, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, and Masashi Sugiyama. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *ICML*, pp. 10–18, 2013.

Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *UAI*, 2017.

Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Semi-supervised auc optimization based on positive-unlabeled learning. *MLJ*, 107(4):767–794, 2018.

Kazuhiko Shinoda, Hirotaka Kaji, and Masashi Sugiyama. Binary classification from positive data with skewed confidence. In *IJCAI*, pp. 3328–3334, 2020.

Miao Sun, Tony X Han, Ming-Chang Liu, and Ahmad Khodayari-Rostamabad. Multiple instance learning convolutional neural networks for object recognition. In *ICPR*, pp. 3270–3275, 2016.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pp. 1195–1204, 2017.

Hong-Xin Wei, Lei Feng, Xiang-Yu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, pp. 13726–13735, 2020.

Xiao-Bo Xia, Tong-Liang Liu, Nan-Nan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, pp. 6835–6846, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Li-Yuan Xu, Junya Honda, Gang Niu, and Masashi Sugiyama. Uncoupled regression from pairwise comparison data. In *Advances in Neural Information Processing Systems*, pp. 3992–4002, 2019.

Xi-Yu Yu, Tong-Liang Liu, Ming-Ming Gong, and Da-Cheng Tao. Learning with biased complementary labels. In *ECCV*, pp. 68–83, 2018.

Min-Ling Zhang, Fei Yu, and Cai-Zhi Tang. Disambiguation-free partial label learning. *TKDE*, 29 (10):2155–2167, 2017.

Ya-Lin Zhang and Zhi-Hua Zhou. Multi-instance learning with key instance shift. In *IJCAI*, pp. 3441–3447, 2017.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53, 2018.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *ICML*, pp. 1249–1256, 2009.

Xiao-Jin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.

A PROOF OF THEOREM 1

It is clear that each pair of examples $(\boldsymbol{x}, \boldsymbol{x}')$ is independently drawn from the following data distribution:

$$\widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') = p((\boldsymbol{x}, \boldsymbol{x}') \mid (y, y') \in \widetilde{\mathcal{Y}}) = \frac{p((\boldsymbol{x}, \boldsymbol{x}'), (y, y') \in \widetilde{\mathcal{Y}})}{p((y, y') \in \widetilde{\mathcal{Y}})},$$

where $p((y, y') \in \widetilde{\mathcal{Y}}) = \pi_+^2 + \pi_-^2 + \pi_+\pi_-$ and

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{x}', (y, y') \in \widetilde{\mathcal{Y}}) &= \sum\nolimits_{(y,y') \in \widetilde{\mathcal{Y}}} p(\boldsymbol{x}, \boldsymbol{x}' \mid (y, y')) \cdot p(y, y') \\ &= \pi_+^2 p_+(\boldsymbol{x}) p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}) p_-(\boldsymbol{x}') + \pi_+\pi_- p_+(\boldsymbol{x}) p_-(\boldsymbol{x}'). \end{aligned}$$

Finally, let $\widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') = p((\boldsymbol{x}, \boldsymbol{x}') \mid (y, y') \in \widetilde{\mathcal{Y}})$, the proof is completed. □

B PROOF OF THEOREM 2

In order to decompose the pairwise comparison data distribution into pointwise distribution, we marginalize $\widetilde{p}(\boldsymbol{x}, \boldsymbol{x}')$ with respect to $\boldsymbol{x}$ or $\boldsymbol{x}'$. Then we can obtain

$$\begin{aligned} \int \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') \mathrm{d}\boldsymbol{x}' =& \frac{1}{\widetilde{\pi}} \Big(\pi_+^2 p_+(\boldsymbol{x}) + \pi_-^2 p_-(\boldsymbol{x}) + \pi_+\pi_- p_+(\boldsymbol{x}) \Big) \\ =& \frac{\pi_+}{\pi_-^2 + \pi_+} p_+(\boldsymbol{x}) + \frac{\pi_-^2}{\pi_-^2 + \pi_+} p_-(\boldsymbol{x}) \\ =& \widetilde{p}_+(\boldsymbol{x}), \end{aligned}$$

and

$$\begin{aligned} \int \widetilde{p}(\boldsymbol{x}, \boldsymbol{x}') \mathrm{d}\boldsymbol{x} =& \frac{1}{\widetilde{\pi}} \big(\pi_+^2 p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}') + \pi_+\pi_- p_-(\boldsymbol{x}') \big) \\ =& \frac{\pi_+^2}{\pi_+^2 + \pi_-} p_+(\boldsymbol{x}') + \frac{\pi_-}{\pi_+^2 + \pi_-} p_-(\boldsymbol{x}') \\ =& \widetilde{p}_-(\boldsymbol{x}'), \end{aligned}$$

which concludes the proof of Theorem 2. □

C PROOF OF LEMMA 1

Based on Theorem 2, we can obtain the following linear equation:

$$\begin{bmatrix} \widetilde{p}_+(\boldsymbol{x}) \\ \widetilde{p}_-(\boldsymbol{x}) \end{bmatrix} = \frac{1}{\widetilde{\pi}} \begin{bmatrix} \pi_+ & \pi_-^2 \\ \pi_+^2 & \pi_- \end{bmatrix} \begin{bmatrix} p_+(\boldsymbol{x}) \\ p_-(\boldsymbol{x}) \end{bmatrix}.$$

By solving the above equation, we obtain

$$\begin{aligned} p_+(\boldsymbol{x}) &= \frac{1}{\pi_+ - \pi_-\pi_+^2} \big(\widetilde{\pi} \cdot \widetilde{p}_+(\boldsymbol{x}) - \pi_- \widetilde{\pi} \cdot \widetilde{p}_-(\boldsymbol{x}) \big) = \frac{1}{\pi_+} \big(\widetilde{p}_+(\boldsymbol{x}) - \pi_- \widetilde{p}_-(\boldsymbol{x}) \big), \\ p_-(\boldsymbol{x}) &= \frac{1}{\pi_- - \pi_+\pi_-^2} \big(\widetilde{\pi} \cdot \widetilde{p}_-(\boldsymbol{x}) - \pi_+ \widetilde{\pi} \cdot \widetilde{p}_+(\boldsymbol{x}) \big) = \frac{1}{\pi_-} \big(\widetilde{p}_-(\boldsymbol{x}) - \pi_+ \widetilde{p}_+(\boldsymbol{x}) \big), \end{aligned}$$

which concludes the proof of Lemma 1. □

D PROOF OF THEOREM 3

It is quite intuitive to derive

$$
\begin{aligned}
R(f) &= \mathbb{E}_{p(\boldsymbol{x},y)}\big[\ell(f(\boldsymbol{x}),y)\big] \\
&= \pi_+\mathbb{E}_{p_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)\big] + \pi_-\mathbb{E}_{p_-(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),-1)\big] \\
&= \frac{\pi_+\widetilde{\pi}}{\pi_+-\pi_-\pi_+^2}\mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)\big] - \frac{\pi_+\pi_-\widetilde{\pi}}{\pi_+-\pi_-\pi_+^2}\mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}),+1)\big] \qquad \text{(Lemma 1)} \\
&\qquad + \frac{\pi_-\widetilde{\pi}}{\pi_--\pi_+\pi_-^2}\mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}),-1)\big] - \frac{\pi_+\pi_-\widetilde{\pi}}{\pi_--\pi_+\pi_-^2}\mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),-1)\big] \\
&= \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)-\pi_+\ell(f(\boldsymbol{x}),-1)\big] + \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}),-1)-\pi_-\ell(f(\boldsymbol{x}),+1)\big] \\
&= R_{\mathrm{pc}}(f),
\end{aligned}
$$

which concludes the proof of Theorem 3. □

E PROOF OF THEOREM 4

First of all, we introduce the following notations:

$$
\begin{aligned}
R_{\mathrm{pc}}^+(f) &= \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}),+1)-\pi_+\ell(f(\boldsymbol{x}),-1)\big], \\
\widehat{R}_{\mathrm{pc}}^+(f) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(f(\boldsymbol{x}_i),+1)-\pi_+\ell(f(\boldsymbol{x}_i),-1)\Big), \\
R_{\mathrm{pc}}^-(f) &= \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}'),-1)-\pi_-\ell(f(\boldsymbol{x}'),+1)\big], \\
\widehat{R}_{\mathrm{pc}}^-(f) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\ell(f(\boldsymbol{x}_i'),-1)-\pi_-\ell(f(\boldsymbol{x}_i'),+1)\Big).
\end{aligned}
$$

In this way, we could simply represent $R_{\mathrm{pc}}(f)$ and $\widehat{R}_{\mathrm{pc}}(f)$ as

$$
R_{\mathrm{pc}}(f) = R_{\mathrm{pc}}^+(f) + R_{\mathrm{pc}}^-(f), \quad \widehat{R}_{\mathrm{pc}}(f) = \widehat{R}_{\mathrm{pc}}^+(f) + \widehat{R}_{\mathrm{pc}}^-(f).
$$

Then we have the following lemma.

Lemma 2. *The following inequality holds:*

$$
R(\widehat{f}_{\mathrm{pc}}) - R(f^\star) \le 2\sup_{f\in\mathcal{F}}\left|R_{\mathrm{pc}}^+(f)-\widehat{R}_{\mathrm{pc}}^+(f)\right| + 2\sup_{f\in\mathcal{F}}\left|R_{\mathrm{pc}}^-(f)-\widehat{R}_{\mathrm{pc}}^-(f)\right|. \tag{9}
$$

Proof. We could intuitively express $R(\widehat{f}_{\mathrm{pc}}) - R(f^\star)$ as

$$
\begin{aligned}
R(\widehat{f}_{\mathrm{pc}}) - R(f^\star) &= R(\widehat{f}_{\mathrm{pc}}) - \widehat{R}_{\mathrm{pc}}(\widehat{f}_{\mathrm{pc}}) + \widehat{R}_{\mathrm{pc}}(\widehat{f}_{\mathrm{pc}}) - \widehat{R}_{\mathrm{pc}}(f^\star) + \widehat{R}_{\mathrm{pc}}(f^\star) - R(f^\star) \\
&= R_{\mathrm{pc}}(\widehat{f}_{\mathrm{pc}}) - \widehat{R}_{\mathrm{pc}}(\widehat{f}_{\mathrm{pc}}) + \widehat{R}_{\mathrm{pc}}(\widehat{f}_{\mathrm{pc}}) - \widehat{R}_{\mathrm{pc}}(f^\star) + \widehat{R}_{\mathrm{pc}}(f^\star) - R_{\mathrm{pc}}(f^\star) \\
&\le \sup_{f\in\mathcal{F}}\left|R_{\mathrm{pc}}(f)-\widehat{R}_{\mathrm{pc}}(f)\right| + 0 + \sup_{f\in\mathcal{F}}\left|R_{\mathrm{pc}}(f)-\widehat{R}_{\mathrm{pc}}(f)\right| \\
&= 2\sup_{f\in\mathcal{F}}\left|R_{\mathrm{pc}}(f)-\widehat{R}_{\mathrm{pc}}(f)\right| \\
&\le 2\sup_{f\in\mathcal{F}}\left|R_{\mathrm{pc}}^+(f)-\widehat{R}_{\mathrm{pc}}^+(f)\right| + 2\sup_{f\in\mathcal{F}}\left|R_{\mathrm{pc}}^-(f)-\widehat{R}_{\mathrm{pc}}^-(f)\right|,
\end{aligned}
$$

where the second inequality holds due to Theorem 3. □

As suggested by Lemma 2, we need to further upper bound the right hand size of Eq. (9). Before doing that, we introduce the *uniform deviation bound*, which is useful to derive estimation error bounds. The proof can be found in some textbooks such as Mohri et al. (2012) (Theorem 3.1).

Lemma 3. *Let Z be a random variable drawn from a probability distribution with density μ, $\mathcal{H} = \{h : \mathcal{Z} \mapsto [0, M]\}$ ($M > 0$) be a class of measurable functions, $\{z_i\}_{i=1}^n$ be i.i.d. examples drawn from the distribution with density μ. Then, for any $delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{h \in \mathcal{H}} \left| \mathbb{E}_{Z \sim \mu}\big[h(Z)\big] - \frac{1}{n} \sum_{i=1}^n h(z_i) \right| \leq 2\mathfrak{R}_n(\mathcal{H}) + M \sqrt{\frac{\log \frac{2}{\delta}}{2n}},$$

where $\mathfrak{R}_n(\mathcal{H})$ denotes the (expected) Rademacher complexity *(Bartlett & Mendelson, 2002) of $\mathcal{H}$ with sample size n over μ.*

Lemma 4. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t=\pm 1} \ell(C_b, t)$. For any $\delta > 0$, with probability $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} \left| R_{\mathrm{pc}}^+(f) - \widehat{R}_{\mathrm{pc}}^+(f) \right| \leq (1 + \pi_+) 2\rho \widetilde{\mathfrak{R}}_n^+(\mathcal{F}) + (1 + \pi_+) C_\ell \sqrt{\frac{\log \frac{4}{\delta}}{2n}}.$$

Proof. By the definition of $R_{\mathrm{pc}}^+(f)$ and $\widehat{R}_{\mathrm{pc}}^+(f)$, we can obtain

$$\begin{aligned} \sup_{f \in \mathcal{F}} \left| R_{\mathrm{pc}}^+(f) - \widehat{R}_{\mathrm{pc}}^+(f) \right| \leq & \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), +1)\big] - \frac{1}{n} \sum_{i=1}^n \ell(f(\boldsymbol{x}), +1) \right| \\ & + \pi_+ \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), -1)\big] - \frac{1}{n} \sum_{i=1}^n \ell(f(\boldsymbol{x}), -1) \right|. \end{aligned} \tag{10}$$

By applying Lemma 3, we have for any $\delta > 0$, with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), +1)\big] - \frac{1}{n} \sum_{i=1}^n \ell(f(\boldsymbol{x}), +1) \right| \leq 2\widetilde{\mathfrak{R}}_n^+(\ell \circ \mathcal{F}) + C_\ell \sqrt{\frac{\log \frac{2}{\delta}}{2n}}, \tag{11}$$

and for any for any $\delta > 0$, with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), -1)\big] - \frac{1}{n} \sum_{i=1}^n \ell(f(\boldsymbol{x}), -1) \right| \leq 2\widetilde{\mathfrak{R}}_n^+(\ell \circ \mathcal{F}) + C_\ell \sqrt{\frac{\log \frac{2}{\delta}}{2n}}, \tag{12}$$

where $\ell \circ \mathcal{F}$ means $\{\ell \circ f \mid f \in \mathcal{F}\}$. By Talagrand's lemma (Lemma 4.2 in Mohri et al. (2012)),

$$\widetilde{\mathfrak{R}}_n^+(\ell \circ \mathcal{F}) \leq \rho \widetilde{\mathfrak{R}}_n^+(\mathcal{F}). \tag{13}$$

Finally, by combing Eqs. (10), (11), (12), and (13), we have for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left| R_{\mathrm{pc}}^+(f) - \widehat{R}_{\mathrm{pc}}^+(f) \right| \leq (1 + \pi_+) 2\rho \widetilde{\mathfrak{R}}_n^+(\mathcal{F}) + (1 + \pi_+) C_\ell \sqrt{\frac{\log \frac{4}{\delta}}{2n}}, \tag{14}$$

which concludes the proof of Lemma 4. □

Lemma 5. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t=\pm 1} \ell(C_b, t)$. For any $\delta > 0$, with probability $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} \left| R_{\mathrm{pc}}^-(f) - \widehat{R}_{\mathrm{pc}}^-(f) \right| \leq (1 + \pi_-) 2\rho \widetilde{\mathfrak{R}}_n^-(\mathcal{F}) + (1 + \pi_-) C_\ell \sqrt{\frac{\log \frac{4}{\delta}}{2n}}.$$

Proof. Lemma 5 can be proved similarly to Lemma 4. □

By combining Lemma 2, Lemma 4, and Lemma 5, Theorem 4 is proved. □

F PROOF OF THEOREM 5

Suppose there are n pairs of paired data points, which means there are in total $2n$ data points. For our Pcomp classification problem, we could simply regard $\boldsymbol{x}$ sampled from $\widetilde{p}_+(\boldsymbol{x})$ as (noisy) positive data and $\boldsymbol{x}'$ sampled from $\widetilde{p}_-(\boldsymbol{x}')$ as (noisy) negative data. Given n pairs of examples $\{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}_{i=1}^n$, for the n observed positive examples, there are actually $n \cdot p(y=+1|\widetilde{y}=+1)$ true positive examples; for the n observed negative examples, there are actually $n \cdot p(y=-1|\widetilde{y}=-1)$ true negative examples. From our defined data generation process in Theorem 1, it is intuitive to obtain

$$p(y=+1 \mid \widetilde{y}=+1) = \frac{\pi_+^2 + \pi_+\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-},$$
$$p(y=-1 \mid \widetilde{y}=-1) = \frac{\pi_-^2 + \pi_+\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-}.$$

Since $\phi_+ = p(y=-1 \mid \widetilde{y}=+1) = 1 - p(y=+1 \mid \widetilde{y}=+1)$ and $\phi_- = p(y=+1 \mid \widetilde{y}=-1) = 1 - p(y=-1 \mid \widetilde{y}=-1)$, we can obtain

$$\phi_+ = p(y=-1 \mid \widetilde{y}=+1) = 1 - \frac{\pi_+^2 + \pi_+\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-} = \frac{\pi_-^2}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-},$$
$$\phi_- = p(y=+1 \mid \widetilde{y}=-1) = 1 - \frac{\pi_-^2 + \pi_+\pi_-}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-} = \frac{\pi_+^2}{\pi_+^2 + \pi_-^2 + \pi_+\pi_-}.$$

In this way, we can further obtain the following noise transition ratios:

$$\rho_+ = p(\widetilde{y}=-1 \mid y=+1) = \frac{p(y=+1 \mid \widetilde{y}=-1)p(\widetilde{y}=-1)}{p(y=+1)} = \frac{\pi_+}{2(\pi_+^2 + \pi_-^2 + \pi_+\pi_-)},$$
$$\rho_- = p(\widetilde{y}=+1 \mid y=-1) = \frac{p(y=-1 \mid \widetilde{y}=+1)p(\widetilde{y}=+1)}{p(y=-1)} = \frac{\pi_-}{2(\pi_+^2 + \pi_-^2 + \pi_+\pi_-)},$$

where $p(\widetilde{y}=1) = p(\widetilde{y}=-1) = \frac{1}{2}$, because we have the same number of observed positive examples and negative examples.

G PROOF OF THEOREM 7

First of all, we introduce the following notations:

$$R_{\mathrm{ppc}}^+(f) = \mathbb{E}_{\widetilde{p}_+(\boldsymbol{x})}\big[\ell(f(\boldsymbol{x}), +1)\mathbb{I}[\boldsymbol{x} \in \widetilde{\mathrm{P}}]\big],$$
$$\widehat{R}_{\mathrm{no}}^+(f) = \frac{1}{n}\sum_{i=1}^n \big(\ell(f(\boldsymbol{x}_i), +1)\mathbb{I}[\boldsymbol{x}_i \in \widetilde{\mathrm{P}}]\big),$$
$$R_{\mathrm{ppc}}^-(f) = \mathbb{E}_{\widetilde{p}_-(\boldsymbol{x}')}\big[\ell(f(\boldsymbol{x}'), -1)\mathbb{I}[\boldsymbol{x}' \in \widetilde{\mathrm{N}}]\big],$$
$$\widehat{R}_{\mathrm{no}}^-(f) = \frac{1}{n}\sum_{i=1}^n \big(\ell(f(\boldsymbol{x}'_i), -1)\mathbb{I}[\boldsymbol{x}'_i \in \widetilde{\mathrm{N}}]\big).$$

In this way, we could simply represent $R_{\mathrm{ppc}}(f)$ and $\widehat{R}_{\mathrm{no}}(f)$ as

$$R_{\mathrm{ppc}}(f) = \frac{1}{1-\rho_+}R_{\mathrm{no}}^+(f) + \frac{1}{1-\rho_-}R_{\mathrm{no}}^-(f), \quad \widehat{R}_{\mathrm{no}}(f) = \frac{1}{1-\rho_+}\widehat{R}_{\mathrm{no}}^+(f) + \frac{1}{1-\rho_-}\widehat{R}_{\mathrm{no}}^-(f).$$

Then we have the following lemma.

Lemma 6. *The following inequality holds:*

$$R(\widehat{f}_{\mathrm{no}}) - R(f^\star) \leq \frac{2}{1-\rho_+}\sup_{f\in\mathcal{F}}\left|R_{\mathrm{ppc}}^+(f) - \widehat{R}_{\mathrm{no}}^+(f)\right| + \frac{2}{1-\rho_-}\sup_{f\in\mathcal{F}}\left|R_{\mathrm{ppc}}^-(f) - \widehat{R}_{\mathrm{no}}^-(f)\right|. \quad (15)$$

Proof. We omit the proof of Lemma 6 since it is quite similar to that of Lemma 2. □

As suggested by Lemma 6, we need to further upper bound the right hand size of Eq. (15). According to Lemma 3, we have the following two lemmas.

Lemma 7. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t=\pm 1} \ell(C_b, t)$. For any $\delta > 0$, with probability $1-\delta$,*

$$\sup_{f\in\mathcal{F}} \left| R^+_{\mathrm{ppc}}(f) - \widehat{R}^+_{\mathrm{no}}(f) \right| \leq 2\rho \widetilde{\mathfrak{R}}^+_n(\mathcal{F}) + C_\ell \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Lemma 8. *Suppose the loss function ℓ is ρ-Lipschitz with respect to the first argument ($0 < \rho < \infty$), and all the functions in the model class $\mathcal{F}$ are bounded, i.e., there exists a constant C_b such that $\|f\|_\infty \leq C_b$ for any $f \in \mathcal{F}$. Let $C_\ell := \sup_{t=\pm 1} \ell(C_b, t)$. For any $\delta > 0$, with probability $1-\delta$,*

$$\sup_{f\in\mathcal{F}} \left| R^-_{\mathrm{ppc}}(f) - \widehat{R}^-_{\mathrm{no}}(f) \right| \leq 2\rho \widetilde{\mathfrak{R}}^-_n(\mathcal{F}) + C_\ell \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

We omit the proofs of Lemma 7 and Lemma 8 since they are similar to that of Lemma 4.

By combing Lemma 6, Lemma 7, and Lemma 8, Theorem 7 is proved.