RETHINKING CLASS-PRIOR ESTIMATION FOR POSITIVE-UNLABELED LEARNING

Yu Yao[1] **Tongliang Liu**[1 †] **Bo Han**[2] **Mingming Gong**[3]
Gang Niu[4] **Masashi Sugiyama**[4,5] **Dacheng Tao**[6,1]

[1]The University of Sydney [2]Hong Kong Baptist University [3]The University of Melbourne
[4]RIKEN AIP [5]The University of Tokyo [6]JD Explore Academy, China

ABSTRACT

Given only *positive* (P) and *unlabeled* (U) data, PU learning can train a binary classifier without any *negative* data. It has two building blocks: PU *class-prior estimation* (CPE) and PU classification; the latter has been well studied while the former has received less attention. Hitherto, the distributional-assumption-free CPE methods rely on a critical assumption that *the support of the positive data distribution cannot be contained in the support of the negative data distribution*. If this is violated, those CPE methods will systematically *overestimate* the class prior; it is even worse that we *cannot verify* the assumption based on the data. In this paper, we rethink CPE for PU learning—can we remove the assumption to make CPE always valid? We show an affirmative answer by proposing Regrouping CPE (ReCPE) that builds an *auxiliary* probability distribution such that the support of the positive data distribution is never contained in the support of the negative data distribution. ReCPE can work with any CPE method by treating it as the base method. Theoretically, ReCPE *does not affect* its base if the assumption already holds for the original probability distribution; otherwise, it *reduces the positive bias* of its base. Empirically, ReCPE improves all state-of-the-art CPE methods on various datasets, implying that the assumption has indeed been violated here.

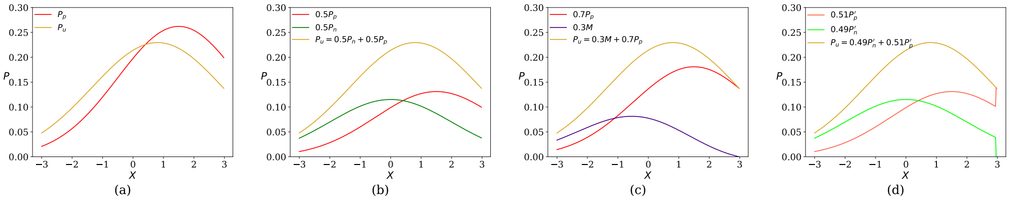
1 INTRODUCTION

Positive-unlabeled (PU) learning can date back to 1990s (Denis, 1998; De Comité et al., 1999; Letouzey et al., 2000), and there has been a surge of interest in this learning scenario in recent years because of the difficulty to annotate large-scale datasets (Ren et al., 2014; du Plessis et al., 2014; 2015; Christoffel et al., 2016; Jain et al., 2016; Ramaswamy et al., 2016; Sakai et al., 2018; Kato et al., 2018; Bekker & Davis, 2018; Gong et al., 2019; Bai et al., 2021; Xia et al., 2021; Yao et al., 2021). It is also fallen into different applications, such as knowledge-base completion (Galárraga et al., 2015; Neelakantan et al., 2015), text classification (Lee & Liu, 2003; Li & Liu, 2003), and medical diagnosis (Claesen et al., 2015; Zuluaga et al., 2011).

PU learning can be divided into two different settings based on different data generation processes. The first setting is called *censoring* PU learning (Elkan & Noto, 2008), which follows a one-sample configuration. Specifically, a sample S is randomly drawn from the unlabeled data distribution P_{u}, and a positive sample S_{p} is then distilled from it, i.e., randomly selecting some positive instances contained in the unlabeled data to be the positive sample. The second setting is called *case-control* PU learning (Kiryo et al., 2017). In this setting, a positive sample $S_{\mathrm{p}} = \{x_i\}_{i=1}^{k}$ is randomly drawn from the positive class-conditional distribution $P_{\mathrm{p}} = P(X|Y = 1)$, and an unlabeled sample $S_{\mathrm{u}} = \{x_i\}_{i=k+1}^{n}$ is randomly drawn from the unlabeled data distribution P_{u}. Because case-control PU learning is more general than censoring PU learning (Niu et al., 2016), therefore, we will focus on the setting of case-control PU learning.

Under the setting of case-control PU learning, a lot of classification methods have been proposed (Ren et al., 2014; du Plessis et al., 2014; 2015; Christoffel et al., 2016; Sakai et al., 2018; Kato et al., 2018; Bekker & Davis, 2018; Kwon et al., 2019; Tanielian & Vasile, 2019; Gong et al., 2019).

[†] Correspondence to Tongliang Liu (tongliang.liu@sydney.edu.au).



(a) (b) (c) (d)

Figure 1: (a) The unlabeled data distribution P_u and the positive class-conditional distribution P_p are given. (b) Assume that the latent negative class-conditional distribution P_n is fixed, i.e., $0.5P_\mathrm{n}$ is shown by the green curve, and that the class-prior π is 0.5, i.e., $P_\mathrm{u} = 0.5P_\mathrm{n} + 0.5P_\mathrm{p}$. (c) The existing distributional-assumption-free CPE methods will output 0.7 instead of 0.5 because they always output the maximum proportion κ^* of P_u in P_p. (d) Applying the proposed ReCPE method, a auxiliary distribution $P_{\mathrm{p}'}$ will be created and the existing CPE methods will output $\pi' = 0.49$ instead of 0.7 with input $P_{\mathrm{p}'}$ and P_u instead of P_p and P_u.

However, the *class-prior estimation* (CPE) (Elkan & Noto, 2008; Jain et al., 2016; Ramaswamy et al., 2016; Christoffel et al., 2016; Kato et al., 2018) has received less attention. Formally, CPE is defined as a problem of estimating $\pi = P(y = 1) \in (0, 1)$ given a sample from the marginal distribution P_u and a sample from positive class-conditional distribution P_p. The marginal distribution P_u is mixed with both positive and negative class-conditional distributions, i.e., $P_\mathrm{u} = \pi P_\mathrm{p} + (1 - \pi)P_\mathrm{n}$. CPE acts as a crucial building block for state-of-the-art PU classification methods, and it is essential to build statistically-consistent PU classifiers (du Plessis et al., 2014; Scott, 2015; Jain et al., 2016; Kiryo et al., 2017). The formulation of these classification methods involves the class-prior π, but π is usually unknown in practice. If π is poorly estimated, the classification accuracy of the state-of-the-art PU classification methods (du Plessis et al., 2014; 2015; Kiryo et al., 2017) could be degraded.

The mixture proportion estimation (MPE) is closely related to CPE (Blanchard et al., 2010; Scott, 2015). In the setting of MPE, there is a mixture distribution

$$F = (1 - \kappa^*)G + \kappa^* H, \tag{1}$$

where H and G are called component distributions. Given the samples randomly drawn from F and H, respectively, MPE aims to estimate the maximum proportion $\kappa^* \in (0, 1)$ of H in F. Thereby, if the maximum proportion κ^* is identical to the class-prior π, the MPE methods can be employed to obtain π by letting P_u and P_p be the mixture distribution F and the component distribution H, respectively; otherwise, the MPE methods cannot be employed. To the best of our knowledge, most of state-of-the-art CPE methods (Blanchard et al., 2010; Liu & Tao, 2015; Scott, 2015; Ramaswamy et al., 2016; Jain et al., 2016) are based on MPE, which do not rely on assumptions that the data are drawn from a given parametric family of probability distributions (i.e., they are distributional-assumption-free methods).

To let these distributional-assumption-free methods can be used to identify class-prior π, κ^* must be identical to the class-prior π. The *irreducibility* assumption (Blanchard et al., 2010) has been proposed to make them identical, which is employed by all these CPE methods implicitly or explicitly, to the best of our knowledge. It assumes that the support of the positive class-conditional distribution P_p is not contained in the support of the negative class-conditional distribution P_n. However, it is strong and hard to be verified in PU learning, since P_n is a latent distribution, such that we do not have any prior knowledge about it. Additionally, since the applications of PU learning are diverse (Hsieh et al., 2019; Bekker & Davis, 2020), it is hard to guarantee that the support of P_p is not in the support of P_n.

If the irreducibility assumption cannot be satisfied, the existing distributional-assumption-free CPE methods will suffer from an overestimation of π. For example, in Figure 1(a), we show both the unlabeled data distribution P_u and the component distribution P_p. In Figure 1(b), we assume the latent negative class-conditional distribution P_n is fixed as shown in the green color, and the positive class-prior $\pi = 0.5$. In Figure 1(c), we show the existing distributional-assumption-free CPE methods will output the biased class-prior 0.7. It is different from the ground truth 0.5, since the support of P_p is contained in the support of P_n. When the irreducibility assumption is not held, how to improve the estimations of distributional-assumption-free PU learning methods is challenging but useful.

Because the irreducibility assumption is impossible to check without making any assumption on P_{n}. Thereby, in this paper, we rethink those CPE methods and propose a novel method called ***Regrouping CPE*** (ReCPE) which improves the estimations of the current PU learning methods without irreducibility assumption. The main idea of our method is that, instead of estimating the maximum proportion of P_{p} in P_{u}, we build a new CPE problem by creating a new auxiliary distribution $P_{\mathrm{p}'}$ always guaranteeing the irreducibility assumption. Then we use the existing CPE method to obtain the maximum proportion of $P_{\mathrm{p}'}$ in P_{u}, which is denoted by π'. We show that, with both theoretical analyses and experimental validations, when the irreducibility assumption holds, our ReCPE method does not affect the prediction of the existing estimators; when the irreducibility assumption does not hold, our method will help the current estimators have less estimation bias, which could improve the performances of PU classification tasks. For example, in Figure 1(d), we create a new class-conditional (auxiliary) distribution $P_{\mathrm{p}'}$. By solving it, $\pi' = 0.51$. The estimation bias of the existing estimators will reduce to $\pi' - \pi = 0.01$ instead of $\kappa^* - \pi = 0.2$.

The rest of the paper is organized as follows. In Section 2, we review the irreducibility assumption and its variants. We discuss the difficulty of checking the assumptions. In Section 3, we provide the estimation biases of the existing consistent distributional-assumption-free CPE methods. Then we propose our method ReCPE, followed by theoretically analysis of its estimation bias and the implementation details. All the proofs are listed in Appendix A. The experimental validations are given in Section 4. Section 5 concludes the paper.

2 Irreducibility of CPE

In this section, we briefly review the assumptions used for existing distributional-assumption-free CPE estimators. Then we provide the estimation bias introduced by consistent distributional-assumption-free CPE methods when the assumptions do not hold.

The irreducibility assumption. Let P_{p} and P_{u} be probability measures (distributions) on a measurable space $(\mathcal{X}, \mathfrak{S})$, where $\mathcal{X}$ is the sample space, and $\mathfrak{S}$ is the σ-algebra. Let κ^* be the maximum proportion of P_{p} in P_{u}. To let κ^* be identical to π, the irreducibility assumption was proposed by Blanchard et al. (2010).

Definition 1 (Irreducibility). *P_{n} and P_{p} are said to satisfy the irreducibility assumption if P_{n} is not a mixture containing P_{p}. That is, there does not exist a decomposition $P_{\mathrm{n}} = (1-\beta)Q + \beta P_{\mathrm{p}}$, where Q is a probability distribution on the measurable space $(\mathcal{X}, \mathfrak{S})$, and $0 < \beta \leq 1$.*

Equivalently, the assumption assumes the support of P_{p} is hardly contained in the support of P_{n}. It means that with the selection of different sets S, the probability $P_{\mathrm{n}}(S)$ can be arbitrarily close to 0, and $P_{\mathrm{p}}(S) > 0$. Suppose we can access the distributions P_{u}, P_{p} and the set $\mathcal{C}$ containing all possible latent distributions, then the class-prior π can be found as follows:

$$\pi = \kappa^* \triangleq \sup\{\alpha | P_{\mathrm{u}} = (1-\alpha)K + \alpha P_{\mathrm{p}}, K \in \mathcal{C}\} = \inf_{S \in \mathfrak{S}, P_{\mathrm{p}}(S) > 0} \frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)}. \tag{2}$$

To the best of our knowledge, all existing distributional-assumption-free CPE methods (Blanchard et al., 2010; Scott et al., 2013; Liu & Tao, 2015; Scott, 2015; Ramaswamy et al., 2016; Ivanov, 2019) are variants of estimating the maximum proportion κ^* of P_{p} in P_{u}. Many of them are statistically consistent estimators (Blanchard et al., 2010; Scott et al., 2013; Liu & Tao, 2015; Scott, 2015).

The variants of the irreducibility. Based on the irreducibility assumption, estimators can be designed with theoretical guarantees that they will converge to the class-prior π (Blanchard et al., 2010). However, the convergence rate can be arbitrarily slow (Scott, 2015). The reason is that the irreducibility assumption implies the following fact (Blanchard et al., 2010; Scott et al., 2013)

$$\inf_{S \in \mathfrak{S}, P_{\mathrm{p}}(S) > 0} \frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)} = 0, \tag{3}$$

i.e., the maximum proportion of P_{p} in P_{n} approaches to 0. To obtain the class-prior π, it requires finding a sequence of the sets S converging to the infimum, which empirically can be hard to find. Therefore, the convergence rate of the designed estimators based on Eq. (3) will be arbitrarily slow. To ensure a fixed rate of convergence, the *anchor set* assumption, a stronger variant of the irreducibility assumption, has been proposed (Scott, 2015; Liu & Tao, 2015; Xia et al., 2019; 2020; Yao et al., 2020). It assumes that

$$\min_{S\in\mathfrak{S},P_{\mathrm{p}}(S)>0}\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}=0, \tag{4}$$

i.e., there exists a set can achieve the minimum 0, which is called an anchor set. Another stronger variant is the *separability* assumption (Ramaswamy et al., 2016) which extends the anchor set assumption to a function space. It is proposed to bound the convergence rate of the method based on kernel-mean-matching (KMM) technique (Gretton et al., 2012).

3 Regrouping for CPE (ReCPE)

In this section, we propose a general method named regrouping for CPE (ReCPE). We discuss how to theoretically and empirically mitigate the overestimation problem of the class-prior π.

3.1 Motivation

In general, it is impossible to verify the irreducibility assumption for CPE. To check the assumption, we need to make P_{n} itself to be observable and verify that whether the distribution P_{n} is a mixture containing the distribution P_{p}, which obviously contradicts the setting of PU learning. However, in practice, the irreducibility assumption may not hold for many real-world problems, because the negative class is diverse (Hsieh et al., 2019; Bekker & Davis, 2020) in PU learning. If the assumption does not hold, P_{n} is said to be reducible to P_{p}, and distributional-assumption-free CPE methods will introduce an *estimation bias*.

Proposition 1. *Let $\beta^*=\inf_{S\in\mathfrak{S},P_{\mathrm{p}}(S)>0}\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}$ be the maximum proportion of P_{p} in P_{n}, given $P_{\mathrm{u}}=(1-\pi)P_{\mathrm{n}}+\pi P_{\mathrm{p}}$, for $0<\pi\leq 1$, we have*

$$\kappa^*=\pi+(1-\pi)\inf_{S\in\mathfrak{S},P_{\mathrm{p}}(S)>0}\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}=\pi+(1-\pi)\beta^*. \tag{5}$$

According to Proposition 1, if the irreducibility assumption does not hold, then there exists $\beta>0$. In this case, maximum proportion κ^* can still be obtained, but it is different from π but equal to $\pi+(1-\pi)\beta^*$. In this case, if we directly employ existing distributional-assumption-free CPE methods, they could introduce an arbitrary estimation bias $(1-\pi)\beta^*$ which depends on P_{n}.

To reduce the estimation bias, we propose ReCPE. The process of regrouping is to change the original class-conditional distributions P_{n} and P_{p} into new class-conditional distributions $P_{\mathrm{n}'}$ and $P_{\mathrm{p}'}$ by transporting the probability mass of the set A from the negative class to the positive class. After regrouping, new class-conditional distributions are guaranteed to satisfy the irreducibility assumption, and therefore, the new positive class-prior π' can be identified by current CPE methods. To get the intuition, we provide a concrete example as follows.

Suppose that P_{p} is the uniform on $[\frac{1}{2},1]$, P_{n} is uniform on $[0,1]$, and $\pi=\frac{1}{2}$. Then we have P_{u} such that it is uniform on $[0,\frac{1}{2})$ and $[\frac{1}{2},1]$, respectively. Specifically, the probabilities are

$$P_{\mathrm{u}}\left([0,\frac{1}{2})\right)=\frac{1}{4},\;\; P_{\mathrm{u}}\left([\frac{1}{2},1]\right)=\frac{3}{4}.$$

In this case, by Eq. 5, the maximum proportion of P_{p} in P_{u} is $\kappa^*=\frac{3}{4}$. Let $\rho>0$ be a small constant, and let $A=(1-\rho,1]$. In this case, the mass of A in P_{u} from P_{n} is $\pi P_{\mathrm{n}}(A)=\frac{\rho}{2}$. After transporting the mass $\frac{\rho}{2}$ from P_{n} to P_{p}, we have a new positive class-prior π' and a new class-conditional $P_{\mathrm{p}'}$ which is uniform on $[\frac{1}{2},1-\rho)$ and $[1-\rho,1]$, respectively. Specifically,

$$\pi'=\frac{1+\rho}{2},\;\; P_{\mathrm{p}'}\left([\frac{1}{2},1-\rho)\right)=\frac{1-2\rho}{1+\rho},\;\; P_{\mathrm{p}'}\left([1-\rho,1]\right)=\frac{3\rho}{1+\rho}.$$

The left equation above shows that the new class-prior π' is dependent on ρ or the size of A. By controlling set A or ρ to be small, π' can be as close to π as possible. This is the intuition of how regrouping works.

Algorithm 1 ReCPE

```
Input: An unlabeled sample $S_\mathrm{u}$ i.i.d. drawn from $P_\mathrm{u}$, a positive sample $S_\mathrm{p}$ i.i.d. drawn from $P_\mathrm{p}$,
and the percentage $p$ of the sample needed to copy from $S_\mathrm{u}$ to $S_\mathrm{p}$.
Train a binary classifier $h$ with the unlabeled sample $S_\mathrm{u}$ and positive sample $S_\mathrm{p}$ by treating $S_\mathrm{u}$
  as a negative sample;
Assign each example $x \in S_\mathrm{u}$ with the negative class-posterior probability $P(Y=-1|X=x)$
  predicted by the trained classifier $h$;
Obtain $S_{\mathrm{p}'}$ by copying $p \times |S_\mathrm{u}|$ examples with the smallest negative class-posterior probability
  $P(Y=-1|X=x)$ from $S_\mathrm{u}$ to $S_\mathrm{p}$;
Estimate the class-prior $\pi'$ by employing an algorithm based on Eq. (5) with inputs $S_\mathrm{u}$ and $S_{\mathrm{p}'}$.
Output: The estimated new class-prior $\hat{\pi}'$.
```

3.2 Practical Implementation

In practice, we have to implement the aforementioned idea of regrouping based on positive sample S_p and unlabeled sample S_u. Since the negative sample is unavailable, we cannot "cut and paste" any example from negative class to positive sample S_p; instead, we can "copy and past" some unlabeled examples to S_p. When doing so, we should select a small set of samples $\hat{A}^*$ which look the most similar to the positive class and dissimilar to the negative class, which could encourage the difference between the original P_n, P_p, and ρ and π, $P_{\mathrm{p}'}$, and π' to be small. This is why $\hat{A}^* = (1-\rho, 1]$ was selected in the above intuitive example, i.e., $\hat{A}^*$ belongs geometrically and visually to the positive class with the highest confidence among all subsets of $[0,1]$ of size ρ.

A hyper-parameter $p \in (0,1)$ is introduced to control the size of set $\hat{A}^*$, theoretically, we prefer the set $\hat{A}^*$ to have a small size. Empirically, p cannot be so small: the existing estimators are insensitive to tiny modifications (they are designed to be robust in such a way, in order to be good estimators). For example, the difference between the estimated class-priors by employing samples S_u and S_p and the one by employing samples S_u and $S_{\mathrm{p}'}$ can be hardly observed if S_p and $S_{\mathrm{p}'}$ only differ from in one or two points. Specifically, $p = 10\%$ is selected for the experiments on all datasets, which leads to a significant improvement of the estimation accuracy. The details on the selection of the hyper-parameter value will be explained in Section 4.1. The algorithm is summarized in Algorithm 1.

There are two fundamental concerns for copying $\hat{A}^*$ to S_p. 1). When we have irreducibility, might regrouping make $\hat{\pi}'$ be a worse approximation? 2). When we lack irreducibility, must regrouping make $\hat{\pi}'$ be a better approximation? While these concerns will be formally clarified later, we give here intuitive implications of regrouping.

1). If we have irreducibility, the $P_\mathrm{n}(\hat{A}^*)$ should be rather small (if not zero), and $\hat{A}^*$ should be drawn from the positive component P_p of the mixture P_u. In this case, regrouping will generally have small influence to P_p. Hence, it will not make $\hat{\pi}'$ worse.

2). If we lack irreducibility, $\hat{A}^*$ may be drawn from either P_p or P_n. By regrouping, $\hat{A}^*$ becomes present in $S_{\mathrm{p}'}$, which encourages the probability of the set $\hat{A}^*$ in $P_{\mathrm{p}'}$ to be large. This will modify P_p as we expected towards irreducibility. As a consequence, regrouping will make $\hat{\pi}'$ better.

3.3 Theoretical Justification

In the regrouping approach described above, the auxiliary class-conditional distribution $P_{\mathrm{p}'}$ and $P_{\mathrm{n}'}$ are created by regrouping a small set A from P_p and P_n. Here, we analyze the properties of regrouping and theoretically justify it.

A formal definition of regrouping In order to analyze the properties, we need to formally define how to split, transport, and regroup a set A (or the mass of A).

Definition 2. *Let M be a probability measure on a measurable space $(\mathcal{X}, \mathfrak{S})$. Given a set $A \in \mathfrak{S}$, we define a measure M^A on the σ-algebra $\mathfrak{S}$ as follows:*

$$\forall S \in \mathfrak{S}, M^A(S) = M(S \cap A). \tag{6}$$

It is easy to see that given two measures M^A and M^{A^c} obtained according to Definition 2, where $A^c = \mathcal{X} \setminus A$, then M^A and M^{A^c} have the following property.

Lemma 1. *Let M be a probability measure over a measurable space $(\mathcal{X}, \mathfrak{S})$. For any set $A \in \mathfrak{S}$, we have $M^A + M^{A^c} = M$.*

Now, we introduce the theory of regrouping. Fixing a set $A \in \mathfrak{S}$, we split P_n as $P_\mathrm{n}^{A^c}$ and P_n^A, transport P_n^A to P_p to regroup them together, i.e.,

$$P_\mathrm{u} = (1-\pi)P_\mathrm{n} + \pi P_\mathrm{p} = (1-\pi)\underbrace{(P_\mathrm{n}^A + P_\mathrm{n}^{A^c})}_{\text{split into two}} + \pi P_\mathrm{p} = (1-\pi)P_\mathrm{n}^{A^c} + \underbrace{((1-\pi)P_\mathrm{n}^A + \pi P_\mathrm{p})}_{\text{regroup as one}}.$$

Finally, we can rewrite the unlabeled data distribution P_u as a mixture of two new class-conditional distributions $P_{\mathrm{n}'}$ and $P_{\mathrm{p}'}$ defined in Theorem 1 by normalization.

Theorem 1. *Let $P_\mathrm{u} = (1-\pi)P_\mathrm{n} + \pi P_\mathrm{p}$. Let $A \subset \mathrm{support}(P_\mathrm{u})$. By regrouping P_n^A to P_p, P_u can be written as a mixture, i.e., $P_\mathrm{u} = (1-\pi')P_{\mathrm{n}'} + \pi' P_{\mathrm{p}'}$, where*

$$\pi' = \pi + (1-\pi)P_\mathrm{n}(A), \tag{7}$$

$$P_{\mathrm{n}'} = \frac{P_\mathrm{n}^{A^c}}{P_\mathrm{n}(A^c)}, \;\; P_{\mathrm{p}'} = \frac{(1-\pi)P_\mathrm{n}^A + \pi P_\mathrm{p}}{(1-\pi)P_\mathrm{n}(A) + \pi}, \tag{8}$$

and $P_{\mathrm{n}'}$ and $P_{\mathrm{p}'}$ satisfy the anchor set assumption.

When class-conditional distributions P_n and P_p do not satisfy the irreducibility assumption, π cannot be obtained by using CPE methods based on MPE, which will lead to an estimation bias discussed before. However, Theorem 1 shows that the new proportion π' is always identifiable as $P_{\mathrm{n}'}$ and $P_{\mathrm{p}'}$ always satisfy the anchor set assumption. Thus, after regrouping, π' is identifiable and can be estimated by the existing CPE methods.

Bias reduction According to Theorem 1, to make π' closer to π, we expect to find the set A looks most dissimilar to the negative class, i.e., $P_\mathrm{n}(A)$ is small.

Theorem 2. *Let $P_{\mathrm{p}'}$ and $P_{\mathrm{n}'}$ be obtained by regrouping a set $A^* := \arg\min_{A \in \mathfrak{S}} \frac{P_\mathrm{n}(A)}{P_\mathrm{p}(A)}$*[1] *from P_p and P_n. 1). If P_n and P_p satisfy the irreducibility assumption, then $\pi' = \pi$; 2). if P_n and P_p dissatisfy the irreducibility assumption, then $\pi < \pi' < \pi + (1-\pi)\beta^* = \kappa^*$.*

Theorem 2 shows how to properly select a set used for regrouping to make π' a good approximation of π. Specifically, once A^* is selected for regrouping, if P_n and P_p satisfy the irreducibility assumption, the new estimation π' will be identical to π; if P_n and P_p dissatisfy the irreducibility assumption, π' obtained by employing the distributions P_u and $P_{\mathrm{p}'}$ will contain a smaller estimation bias compared to κ^* obtained by employing the distributions P_u and P_p.

Convergence analysis For completeness, we illustrate the convergence property of ReCPE, which is presented by employing the estimator proposed by Blanchard et al. (2010). Let S_u, S_p and $S_{\mathrm{p}'}$ be the samples i.i.d. drawn from P_u, P_p and $P_{\mathrm{p}'}$, respectively. Let A be the set used for regrouping. Let $h : \mathcal{X} \to \mathbb{R}, h \in \mathcal{H}$, be a function that predicts 1 for all elements in the set A and 0 otherwise, where $\mathcal{H}$ denotes a *hypothesis space*. Let $|S|$ denote the cardinality of a set S. Let $\mathbb{1}_{\{h(x)=1\}}$ be an indicator function which returns 1 if $h(x)$ predicts 1 and 0 otherwise. Then $P_\mathrm{u}(A)$ can be expressed as $\int_{x \in \mathcal{X}} p_\mathrm{p}(x)\mathbb{1}_{\{h(x)=1\}} \mathrm{d}x$, where p_p is the density function of the distribution P_u. Let $\hat{P}_\mathrm{u}(A)$ be the empirical version of $P_\mathrm{u}(A)$, i.e., $\hat{P}_\mathrm{u}(A) = \frac{1}{|S_\mathrm{u}|}\sum_{x \in \mathcal{X}} \mathbb{1}_{\{h(x)=1\}}$. Similarly, let $\hat{P}_{\mathrm{p}'}(A)$ be the empirical version of $P_{\mathrm{p}'}(A)$. Let the error $\epsilon_{\delta,\mathcal{H}}(S_\mathrm{u})$ denote the difference between $P_\mathrm{u}(A)$ and $\hat{P}_\mathrm{u}(A)$ obtained by exploiting the *empirical Rademacher complexity* (Mohri et al., 2018). Similarly, let $\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})$ denote the difference between $P_{\mathrm{p}'}(A)$ and $\hat{P}_{\mathrm{p}'}(A)$. We have the following theorem.

Theorem 3. *Let $P_\mathrm{u} = (1-\pi)P_\mathrm{n} + \pi P_\mathrm{p}$. By selecting a set A and regrouping P_n^A to P_p. Then, with probability $1 - 2\delta$, the estimated class-prior $\hat{\pi}'$ based on solving $\inf_{S \in \mathfrak{S}, \hat{P}_{\mathrm{p}'}(S) > 0} \frac{\hat{P}_\mathrm{u}(S)}{\hat{P}_{\mathrm{p}'}(S)}$ satisfies*

$$|\hat{\pi}' - \pi| \leq \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} + \frac{\epsilon_{\delta,\mathcal{H}}(S_\mathrm{u})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} + (1-\pi)P_\mathrm{n}(A), \tag{9}$$

[1] We have defined that the fraction tends to infinite if its numerator is larger than 0 and its denominator is 0. Additionally, the infimum may not always exist, if it does not exist, we could use a sequence of sets that converges to the infimum value, but the convergence rate can be arbitrarily slow (Scott, 2015).

where $\epsilon_{\delta,\mathcal{H}}(S) \triangleq 2\hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{4}{\delta}}{2|S|}}$, *and* $\hat{\mathfrak{R}}_S(\mathcal{H})$ *is the empirical Rademacher complexity of* $\mathcal{H}$*.*

To make $\epsilon_{\delta,\mathcal{H}}(S)$ converge to 0 with the increasing of the sample size of S, a *universal approximation* assumption has been proposed by Scott (2015) to ensure that the hypothesis space is large enough to represent a wide variety of interesting functions. Under the assumption, Scott (2015) proved that, with increasing of the size of samples S_{u} and $S_{\mathrm{p}'}$, the error between $P_{\mathrm{u}}(A)$ and $\hat{P}_{\mathrm{u}}(A)$ will converge to 0 at a rate $\mathcal{O}\left(\sqrt{\frac{\log|S_{\mathrm{u}}|}{|S_{\mathrm{u}}|}}\right)$, similarly to $P_{\mathrm{p}'}(A)$ and $\hat{P}_{\mathrm{p}'}(A)$. Since the empirical Rademacher complexity $\hat{\mathfrak{R}}_X(\mathcal{H})$ of a hypothesis space $\mathcal{H}$ can be upper-bounded by its VC-dimension (Mohri et al., 2018), the both errors based on the empirical Rademacher complexity will also converge to zero with increasing of the sample size. Consequently, the estimation $\hat{\pi}' = \frac{\hat{P}_{\mathrm{u}}(A)}{\hat{P}_{\mathrm{p}'}(A)}$ will converge to $\pi' = \frac{P_{\mathrm{u}}(A)}{P_{\mathrm{p}'}(A)} = \pi + (1-\pi)P_{\mathrm{n}}(A)$ at a rate $\mathcal{O}\left(\sqrt{\frac{\log(\min(|S_{\mathrm{u}}|,|S_{\mathrm{p}'}|))}{\min(|S_{\mathrm{u}}|,|S_{\mathrm{p}'}|)}}\right)$.

Computationally efficient identification of A^* The following theorem presents how to identify A^* with P_{u} and P_{p}. Let us define another auxiliary distribution $q(X, C)$, where $C \in \{0, 1\}$ is the positive-vs-unlabeled label i.e., a class label distinguishing between the positive component and the whole mixture. Specifically, priors are $q(C = 1) := \frac{\pi}{1-\pi}$ and $q(C = 0) := \frac{1}{1-\pi}$; conditional densities are $q(X|C = 1) := P_{\mathrm{p}}$ and $q(X|C = 0) := P_{\mathrm{u}}$; class-posterior probabilities are $q(C = 0|X)$ and $q(C = 1|X)$. We have the following theorem.

Theorem 4. *Let* p_{u} *and* p_{p} *be density functions of* P_{u} *and* P_{p}*, respectively. Let* $q = P(C = 0)p_{\mathrm{u}} + P(C = 1)p_{\mathrm{p}}$*. Let* $\mathbb{1}_A : \mathcal{X} \to \{0, 1\}$ *be the identity function which outputs* 1 *if* $x \in \mathcal{X}$ *is in the set* A*, and* 0 *otherwise. Then the set* $A^* = \arg\min_{A\in\mathfrak{S}} \frac{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=0|X=x)]}{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=1|X=x)]}$.

For the above optimization, its objective function has two expectations over q, which can have the "exact" empirical solution obtained by replacing expectations with empirical averages: $A^* = \arg\min_{A\subset S} \frac{\sum_{x\in A} q(C=0|X=x)}{\sum_{x\in A} q(C=1|X=x)}$.

Approximation of $P_{\mathrm{p}'}$ **with a surrogate** As we do not have examples drawn from P_{n}, it is hard to create $P_{\mathrm{p}'}$, let alone sample from it. We approximate $P_{\mathrm{p}'}$ by using $P_{\tilde{\mathrm{p}}'} = \frac{P_{\mathrm{u}}^A + P_{\mathrm{p}}}{P_{\mathrm{u}}(A)+1}$. The following proposition shows that when $P_{\mathrm{u}}(A)$ is small, $P_{\tilde{\mathrm{p}}'}$ is almost identical to $P_{\mathrm{p}'}$.

Proposition 2. *Let* $P_{\tilde{\mathrm{p}}'} = \frac{P_{\mathrm{u}}^A + P_{\mathrm{p}}}{P_{\mathrm{u}}(A)+1}$ *and* $P_{\mathrm{u}}(A) < \epsilon$*.* $\forall\epsilon > 0$ *and* $S \in \mathfrak{S}$*,* $|P_{\mathrm{p}'}(S) - P_{\tilde{\mathrm{p}}'}(S)| \leq \mathcal{O}(\epsilon)$*.*

Since the gap has the same order as $P_{\mathrm{u}}(A)$ uniformly over S, it is guaranteed that whenever $P_{\mathrm{u}}(A)$ is small, the gap is also small. Practically, we can control the parameter ϵ in the above proposition to be small. Specifically, using a small value of the hyper-parameter p in Algorithm 1 will lead to the set A in Theorem 1 to be small, as well as $P_{\mathrm{u}}(A)$. As a consequence, the practical implementation of regrouping is a good approximation of the theory of regrouping as we expected. By now, we have analyzed all of the properties of regrouping and theoretically justified all of the points in its design.

4 EXPERIMENTS

We run experiments on 2 synthetic datasets and 9 real word datasets[2]. The objectives of employing synthetic datasets are to validate whether the proposed regrouping CPE method reduces the estimation error of the consistent distributional-assumption-free CPE method on the dataset satisfying the irreducibility assumption and does not influence the prediction of the CPE method on the dataset dissatisfying the irreducibility assumption. The hyper-parameter p is also selected from the synthetic datasets. The real-world datasets are used to illustrate the effectiveness of our methods. Although we have introduced a hyper-parameter p and used approximations in the implementation, empirical results on all synthetic and real-world datasets consistently show the superiority of ReCPE.

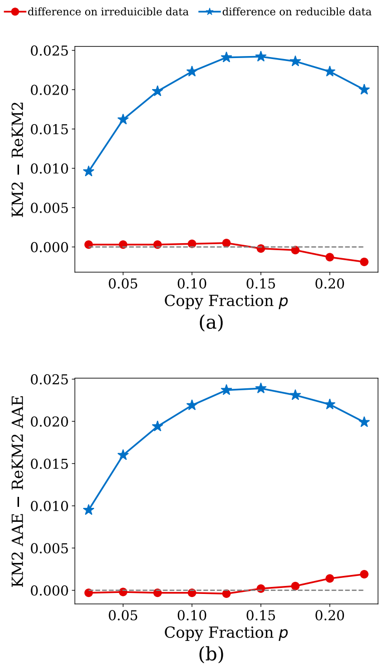
To have a rigorous performance evaluation, for each dataset, 6 × 3 × 10 experiments are conducted via random sampling. Specifically, we select $\{0.25, 0.5, 0.75\}$ fraction of positive examples to be the

[2] The real word datasets are downloaded from the UCL machine learning database. Multi-class datasets are used as binary datasets by either grouping or ignoring classes.

sample of the positive distribution P_p. We let the rest of the examples be the sample of the unlabeled distribution P_u. In such a way, 3 pairs of empirical positive and unlabeled distributions are generated. Then, we create other 3 pairs of distributions by flipping the labels of all instances in the original datasets. For each pair of distributions, we randomly draw positive and unlabeled samples with sizes of 800, 1600, and 3200, respectively, which are used as input data. Note that, the positive and unlabeled samples have the same size as did in Ramaswamy et al. (2016). For each sample size, 10 repeated experiments are carried out with random sampling. For all experiments, we employ a neural network [3] with 2 hidden layers. Each hidden layer contains 50 hidden units. The batch normalization (Ioffe & Szegedy, 2015) is also employed. The stochastic gradient descent optimizer is used with the batch size 50. The network is trained for 350 epochs with a learning rate 0.01 and momentum 0. The weight decay is set to $1e-5$. The model with the best validation accuracy is used to estimate the positive class-posterior probability $P(Y=1|X=x)$. We sample the validation set with 20% of the training data size.

4.1 Experiments on Synthetic Datasets

We create two datasets with one satisfying the irreducibility assumption while the other not. The dataset satisfying the irreducibility assumption is created by sampling from 2 different 10-dimensional Gaussian distributions as the component distributions. One of the distributions has zero means and a unit covariance matrix. Another one has unit means and unit covariance matrix. The dataset dissatisfying the irreducibility assumption is also created by drawing examples from 2 different 10-dimensional Gaussian distributions. One of the distributions has zero means and unit covariance matrix. Another one has unit means and covariance matrix. Then we remove all the data points with $P(Y=1|X) \geq 0.98$ or $P(Y=1|X) \leq 0.02$. For simplicity, in Figure 2, we name two datasets irreducible data and reducible data, respectively.



(a)

(b)

Figure 2: Experiments of the hyper-parameter selection on synthetic datasets. With increasing of the copy fraction p, (a) average estimation differences between KM2 and Regrouping KM2 (ReKM2) and (b) average differences of the absolute error between KM2 and Regrouping KM2 (ReKM2).

To validate the correctness of our method and to select a suitable value of the hyper-parameter p, we carry out two experiments. The consistent CPE method KM2 is used as the baseline, which is compared to our method ReKM2, i.e., regrouping version of the KM2. Firstly, we compare the magnitude differences between $\hat{\pi}$ and $\hat{\pi}'$ (i.e., $\hat{\pi} - \hat{\pi}'$) with the different fractions of points to be copied from the mixture sample to the component sample, which is illustrated in Figure 2(a). Then we compare differences of the absolute error (i.e., $|\hat{\pi} - \pi| - |\hat{\pi}' - \pi|$) between the baseline and our method with the increasing of the copy fractions. Note that each point in Figure 2 is obtained by averaging over $6 \times 3 \times 10$ experiments.

Figure 2(a) validates the correctness of our Theorem 2 and Eq. (7). Theorem 2 states that, by properly selecting the set A, on the dataset dissatisfying the irreducibility assumption (reducible data), π' should be smaller than the maximum proportion κ^*; on the dataset satisfying the irreducibility assumption (irreducible data), π' should be close to π. Figure 2(a) perfectly matches this statement. It shows that, on the reducible data, the values of $\hat{\pi}'$ are continuously smaller than $\hat{\pi}$ with the copy fraction $\leq 22.5\%$; on the irreducible data, $\hat{\pi}'$ and $\hat{\pi}$ have the similar values until the copy fraction $\geq 17.5\%$. According to Eq. (7), the positive bias of our estimator should become larger with the increase of $P_\mathrm{n}(A)$. This fact is reflected by the differences of $\hat{\pi} - \hat{\pi}'$ become smaller on both datasets when the copy fraction $> 15\%$.

Figure 2(b) illustrates the average differences of absolute error between the baseline and the proposed method. On the reducible data, our method continuously outperforms the baseline with the copy

[3] We employ the neural network because it has a high approximation capability (Csáji et al., 2001).

	AM	ReAM	DPL	ReDPL	EN	ReEN	KM1	ReKM1	KM2	ReKM2	ROC	ReROC	RPG	ReRPG
adult (800)	**0.127**	0.13	0.122	**0.108**∗	0.316	**0.295**	0.255	**0.132**	0.164	**0.153**	0.176	**0.153**	0.135	**0.134**
adult (1600)	**0.122**	0.124	0.089*	0.089*	0.31	**0.29**	0.131	**0.091**	**0.12**	0.13	0.121	**0.095**	**0.123**	0.137
adult (3200)	0.105	**0.086**	**0.054**	0.057	0.297	**0.279**	0.054	**0.04**∗	**0.082**	0.089	0.089	**0.067**	**0.114**	0.128
avila (800)	0.168	**0.152**	**0.129**	0.147	0.447	**0.422**	0.105	**0.075**∗	0.104	**0.081**	0.263	**0.228**	0.119	**0.111**
avila (1600)	0.165	**0.132**	0.104	**0.084**	0.439	**0.418**	0.086	**0.076**∗	0.108	**0.092**	0.191	**0.16**	0.123	**0.121**
avila (3200)	0.156	**0.133**	**0.05**∗	0.061	0.436	**0.42**	0.092	**0.078**	0.112	**0.092**	0.131	**0.095**	**0.121**	0.122
bank (800)	**0.135**	0.158	**0.116**∗	0.132	0.282	**0.264**	0.356	**0.216**	0.266	**0.238**	0.163	**0.15**	**0.163**	0.185
bank (1600)	**0.117**	0.167	**0.087**∗	0.105	0.262	**0.244**	0.178	**0.128**	0.203	**0.198**	0.129	**0.118**	**0.157**	0.167
bank (3200)	**0.104**	0.127	**0.073**∗	0.091	0.248	**0.237**	0.124	**0.09**	**0.15**	0.16	**0.093**	0.106	**0.159**	0.18
card (800)	0.131	**0.127**∗	0.174	**0.161**	0.465	**0.444**	0.293	**0.176**	0.203	**0.158**	0.247	**0.233**	0.177	**0.155**
card (1600)	0.173	**0.14**	0.14	0.14	0.459	**0.437**	0.19	**0.135**	0.159	**0.129**	0.194	**0.163**	0.126	**0.115**∗
card (3200)	0.164	**0.134**	0.127	**0.12**	0.455	**0.435**	0.161	**0.113**	0.142	**0.122**	0.159	**0.152**	0.11	**0.108**∗
covtype (800)	0.16	**0.123**	0.155	**0.151**	0.367	**0.343**	0.157	**0.142**	**0.122**	0.13	0.291	**0.258**	0.116	**0.105**∗
covtype (1600)	0.12	**0.1**∗	0.132	**0.109**	0.364	**0.339**	0.116	**0.113**	**0.121**	0.123	0.199	**0.161**	0.109	**0.108**
covtype (3200)	0.128	**0.09**	0.093	**0.083**∗	0.354	**0.334**	**0.097**	0.109	**0.124**	0.128	0.157	**0.113**	0.109	**0.107**
egg (800)	0.153	**0.106**∗	**0.218**	0.225	0.505	0.505	**0.173**	0.264	**0.119**	0.131	0.476	**0.396**	0.171	**0.124**
egg (1600)	0.137	**0.12**	**0.121**	0.142	**0.486**	0.489	0.234	**0.214**	0.116	**0.108**∗	0.315	**0.238**	0.151	**0.114**
egg (3200)	0.126	**0.113**	**0.057**∗	0.073	**0.485**	0.489	0.26	**0.193**	0.134	**0.113**	0.163	**0.139**	0.142	**0.102**
magic04 (800)	0.099	**0.077**	0.072	**0.071**	0.312	**0.296**	0.111	**0.1**	0.071	**0.064**	0.141	**0.124**	0.055	**0.054**∗
magic04 (1600)	0.071	**0.056**	0.044	**0.043**∗	0.292	**0.274**	0.084	**0.072**	0.079	**0.065**	0.1	**0.073**	0.058	**0.052**
magic04 (3200)	0.069	**0.054**	**0.035**∗	0.036	0.274	**0.258**	0.07	**0.047**	0.085	**0.063**	0.065	**0.047**	0.054	**0.052**
robot (800)	**0.053**	0.062	0.049	**0.047**∗	0.19	**0.187**	0.232	**0.215**	**0.111**	0.114	**0.119**	0.144	**0.077**	0.084
robot (1600)	0.053	**0.038**∗	0.087	**0.054**	0.139	**0.132**	0.15	**0.141**	**0.098**	0.099	0.08	**0.075**	**0.076**	0.079
robot (3200)	0.052	**0.039**∗	0.156	**0.119**	0.091	**0.085**	0.079	**0.077**	0.084	0.084	0.063	**0.043**	**0.06**	0.066
shuttle (800)	0.083	**0.031**	**0.016**∗	0.02	0.041	**0.035**	**0.058**	0.083	**0.035**	0.065	**0.042**	0.047	**0.035**	0.051
shuttle (1600)	0.09	**0.045**	**0.011**∗	0.018	0.04	**0.034**	**0.048**	0.079	**0.024**	0.05	**0.029**	0.043	**0.026**	0.039
shuttle (3200)	0.076	**0.028**	**0.012**∗	0.021	0.043	**0.038**	**0.046**	0.07	**0.018**	0.03	**0.038**	0.045	**0.028**	0.042
average	0.116	**0.1**	0.094	**0.092**∗	0.311	**0.297**	0.146	**0.121**	0.117	**0.111**	0.157	**0.136**	0.106	**0.105**

Table 1: Absolute estimation errors on real-world datasets. The first column provides the names of the datasets and sample size. We bold the smaller average estimation errors by comparing each baseline method with its regrouped version. The smallest average estimation error among all methods for each row is highlighted with $*$. The last row is obtained by averaging the results of all experiments. Variances and the results of Wilcoxon signed-rank test are reported in Appendix B. The proposed Regrouping methods are significantly better than most of the baselines.

fraction $\leq 22.5\%$. However, the differences of average absolute error start to decrease with the copy fraction $> 15\%$. On the irreducible data, the differences of average absolute error are close to zero until the copy fraction $> 15\%$.

By observing Figure 2, we found the prediction of the KM2 estimator will not change much if the copy fraction p is too small. For example, the difference between the estimated mixture proportion by employing samples S_{u} and X_P and the ones by employing samples S_{u} and X'_P can be hardly observed if X_P and X'_P only differ from in one or two points. For simplicity and consistency, we select hyper-parameter p to be 10% for all the following experiments.

4.2 Experiments on Real-world Datasets

We illustrate the absolute estimation errors of different estimators on the real-world datasets. Totally, 7 baseline methods are used in the experiments, which are AlphaMax (AM) (Jain et al., 2016), DEDPUL (DPL) (Ivanov, 2019), Elkan-Noto (EN) (Elkan & Noto, 2008), KM1, KM2 (Ramaswamy et al., 2016), ROC (Scott, 2015), and Rankpruning (RPG) (Northcutt et al., 2017). By using our method, the regrouped version of them are implemented, which are called ReAM, ReDPL, ReEN, ReKM1, ReKM2, ReROC, and ReRPG. In Table 1, we compare the absolute estimation errors of each baseline with those of its regrouped version on different datasets with different sample lengths. Each number in Table 1 is the average over 6×10 experiments.

Table 1 reflects the effectiveness of our regrouping CPE method. Overall, by using our method, the estimation accuracy is increased for most of the popular CPE methods among most of the datasets with different sample lengths. By observing the last row, the regrouped version of the estimators has much smaller average estimation errors except DPL, KM2, and RPG. On the real-world datasets, Regrouping AlphaMax (ReAM) has the smallest average estimation error among all methods.

5 Conclusion

In this paper, we investigate how to reduce the estimation bias of the distributional-assumption-free CPE method without irreducibility assumption for PU learning. We have proposed regrouping CPE which can be employed on top of most existing CPE methods. We have also theoretically analyzed the estimation bias of ReCPE. Empirically, it improves all popular CPE methods on various datasets. One future work will focus on how to generate a sample from $P_{\mathrm{p}'}$ instead of using an approximation.

Reproducibility Statement

For theoretical results, we have clearly explained any assumptions. A complete proof of the claims can be founded in the appendix. We have also included an anonymous source code in our supplementary material. For any datasets used in the experiments, a complete description of the data processing steps is provided In Section 4.

Acknowledgments

TL was partially supported by Australian Research Council Projects DP180103424, DE-190101473, IC-190100031, and DP-220102121. BH was supported by the RGC Early Career Scheme No. 22200720 and NSFC Young Scientists Fund No. 62006202. MG is supported by ARC DE210101624. GN and MS were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. MS was also supported by the Institute for AI and Beyond, UTokyo.

References

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34, 2021.

Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Mach. Learn.*, 109(4):719–760, 2020.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.

Marthinus Christoffel, Gang Niu, and Masashi Sugiyama. Class-prior estimation for learning from positive and unlabeled data. In *Asian Conference on Machine Learning*, pp. 221–236, 2016.

Marc Claesen, Frank De Smet, Pieter Gillard, Chantal Mathieu, and Bart De Moor. Building classifiers to predict the start of glucose-lowering pharmacotherapy using belgian health expenditure data. *arXiv preprint arXiv:1504.07389*, 2015.

Balázs Csanád Csáji et al. Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24(48):7, 2001.

Francesco De Comité, François Denis, Rémi Gilleron, and Fabien Letouzey. Positive and unlabeled examples help learning. In *International Conference on Algorithmic Learning Theory*, pp. 219–230. Springer, 1999.

François Denis. Pac learning from positive statistical queries. In *International Conference on Algorithmic Learning Theory*, pp. 112–126. Springer, 1998.

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

Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in neural information processing systems*, pp. 703–711, 2014.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220. ACM, 2008.

Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M Suchanek. Fast rule mining in ontological knowledge bases with amie. *The VLDB Journal*, 24(6):707–730, 2015.

Chen Gong, Hong Shi, Tongliang Liu, Chuang Zhang, Jian Yang, and Dacheng Tao. Loss decomposition and centroid estimation for positive and unlabeled learning. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. In *International Conference on Machine Learning*, pp. 2820–2829, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Dmitry Ivanov. Dedpul: Method for mixture proportion estimation and positive-unlabeled classification based on density estimation. *arXiv preprint arXiv:1902.06965*, 2019.

Shantanu Jain, Martha White, Michael W Trosset, and Predrag Radivojac. Nonparametric semi-supervised learning of class proportions. *arXiv preprint arXiv:1601.01944*, 2016.

Masahiro Kato, Liyuan Xu, Gang Niu, and Masashi Sugiyama. Alternate estimation of a classifier and the class-prior from positive and unlabeled data. *arXiv preprint arXiv:1809.05710*, 2018.

Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems*, pp. 1675–1685, 2017.

Yongchan Kwon, Wonyoung Kim, Masashi Sugiyama, and Myunghee Cho Paik. Principled analytic classifier for positive-unlabeled learning via weighted integral probability metric. *Machine Learning*, pp. 1–20, 2019.

Wee Sun Lee and Bing Liu. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*, volume 3, pp. 448–455, 2003.

Fabien Letouzey, François Denis, and Rémi Gilleron. Learning from positive and unlabeled examples. In *International Conference on Algorithmic Learning Theory*, pp. 71–85. Springer, 2000.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, volume 3, pp. 587–592, 2003.

Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

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

Arvind Neelakantan, Benjamin Roth, and Andrew McCallum. Compositional vector space models for knowledge base completion. In *ACL*, 2015.

Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in neural information processing systems*, pp. 1199–1207, 2016.

Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. *stat*, 1050:9, 2017.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning*, pp. 2052–2060, 2016.

Yafeng Ren, Donghong Ji, and Hongbin Zhang. Positive unlabeled learning for deceptive reviews detection. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 488–498, 2014.

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

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pp. 838–846, 2015.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference On Learning Theory*, pp. 489–511, 2013.

Ugo Tanielian and Flavian Vasile. Relaxed softmax for pu learning. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 119–127, 2019.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32, 2019.

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

Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. *arXiv preprint arXiv:2106.00445*, 2021.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271, 2020.

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

Maria A Zuluaga, Don Hush, Edgar JF Delgado Leyton, Marcela Hernández Hoyos, and Maciej Orkisz. Learning from only positive and unlabeled data to detect lesions in vascular ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 9–16. Springer, 2011.

APPENDIX TO "RETHINKING CLASS-PRIOR ESTIMATION FOR POSITIVE-UNLABELED LEARNING"

A PROOFS

In this section, we show all the proofs.

A.1 PROOF OF PROPOSITION 1

Proposition 1. *Let* $\beta^* = \inf_{S\in\mathfrak{S},P_\mathrm{p}(S)>0} \frac{P_\mathrm{n}(S)}{P_\mathrm{p}(S)}$ *be the maximum proportion of* P_p *in* P_n*, given* $P_\mathrm{u} = (1-\pi)P_\mathrm{n} + \pi P_\mathrm{p}$*, for* $0 < \pi \leq 1$*, we have*

$$\kappa^* = \pi + (1-\pi) \inf_{S\in\mathfrak{S},P_\mathrm{p}(S)>0} \frac{P_\mathrm{n}(S)}{P_\mathrm{p}(S)} = \pi + (1-\pi)\beta^*. \tag{1}$$

Proof. Let κ^* be the maximum proportion of P_p in P_u, which can be formulated as $\kappa^* = \inf_{S\in\mathfrak{S},P_\mathrm{p}(S)>0} \frac{P_\mathrm{u}(S)}{P_\mathrm{p}(S)}$. Then,

$$\begin{aligned}\kappa^* &= \inf_{S\in\mathfrak{S},P_\mathrm{p}(S)>0} \frac{(1-\pi)P_\mathrm{n}(S) + \pi P_\mathrm{p}(S)}{P_\mathrm{p}(S)} \\ &= \inf_{S\in\mathfrak{S},P_\mathrm{p}(S)>0} \frac{(1-\pi)P_\mathrm{n}(S)}{P_\mathrm{p}(S)} + \pi \\ &= \pi + (1-\pi) \inf_{S\in\mathfrak{S},P_\mathrm{p}(S)>0} \frac{P_\mathrm{n}(S)}{P_\mathrm{p}(S)}. \end{aligned} \tag{2}$$

By letting $\beta^* = \inf_{S\in\mathfrak{S},P_\mathrm{p}(S)>0} \frac{P_\mathrm{n}(S)}{P_\mathrm{p}(S)}$, $\kappa^* = \pi + (1-\pi)\beta^*$ which completes the proof. □

A.2 PROOF OF LEMMA 1

Lemma 1. *Let* M *be a probability measure over a measurable space* $(\mathcal{X}, \mathfrak{S})$*. For any set* $A \in \mathfrak{S}$*, we have* $M^A + M^{A^c} = M$*.*

Proof. Let $A^c = \mathcal{X} \setminus A$. Let 2^A and 2^{A^c} be the power sets on A and A^c, respectively. According to Definition 2 in the main paper, M^A and M^{A^c} are defined as follows,

$$\begin{aligned}&\forall S \in \mathfrak{S}, M^A(S) = M(S \cap A); \\ &\forall S \in \mathfrak{S}, M^{A^c}(S) = M(S \cap A^c).\end{aligned}$$

To prove $M = M^A + M^{A^c}$, we need to prove $\forall S \in \mathfrak{S}, M^A(S) + M^{A^c}(S) = M(S)$.

$\forall S \in \mathfrak{S}$,

$$\begin{aligned}M^A(S) + M^{A^c}(S) &= M(S\cap A) + M(S \cap A^c) = M((S\cap A) \cup (S \cap A^c)) \\ &= M(S \cap (A \cup A^c)) = M(S \cap \mathcal{X}) = M(S),\end{aligned}$$

which completes the proof. □

A.3 Proof of Theorem 1

Theorem 1. *Let* $P_{\mathrm{u}} = (1-\pi)P_{\mathrm{n}} + \pi P_{\mathrm{p}}$. *Let* $A \subset \mathrm{support}(P_{\mathrm{u}})$. *By regrouping* P_{n}^{A} *to* P_{p}, P_{u} *can be written as a mixture, i.e.,* $P_{\mathrm{u}} = (1-\pi')P_{\mathrm{n}'} + \pi' P_{\mathrm{p}'}$, *where*

$$\pi' = \pi + (1-\pi)P_{\mathrm{n}}(A), \tag{3}$$

$$P_{\mathrm{n}'} = \frac{P_{\mathrm{n}}^{A^c}}{P_{\mathrm{n}}(A^c)}, \;\; P_{\mathrm{p}'} = \frac{(1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}}}{(1-\pi)P_{\mathrm{n}}(A) + \pi}, \tag{4}$$

and $P_{\mathrm{n}'}$ *and* $P_{\mathrm{p}'}$ *satisfy the anchor set assumption.*

Proof. Firstly, we prove that by regrouping P_{n}^{A} to P_{p}, P_{u} is a convex combination of two new class-conditional distributions, i.e., $P_{\mathrm{u}} = (1-\pi')P_{\mathrm{n}'} + \pi' P_{\mathrm{p}'}$.

Let $A \in \mathfrak{S}$, we split P_{n} as $P_{\mathrm{n}}^{A^c}$ and P_{n}^{A}, transport P_{n}^{A} to P_{p} to regroup them together, i.e.,

$$P_{\mathrm{u}} = (1-\pi)P_{\mathrm{n}} + \pi P_{\mathrm{p}} = (1-\pi)(P_{\mathrm{n}}^{A} + P_{\mathrm{n}}^{A^c}) + \pi P_{\mathrm{p}} = (1-\pi)P_{\mathrm{n}}^{A^c} + ((1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}}). \tag{5}$$

Normalizing $P_{\mathrm{n}}^{A^c}$ and $((1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}})$ in Eq. (5) to probability measures, we have

$$\begin{aligned} P_{\mathrm{u}} &= (1-\pi)P_{\mathrm{n}}^{A^c} + ((1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}}) \\ &= ((1-\pi)P_{\mathrm{n}}^{A^c}(\mathcal{X}))\frac{P_{\mathrm{n}}^{A^c}}{P_{\mathrm{n}}^{A^c}(\mathcal{X})} + ((1-\pi)P_{\mathrm{n}}^{A}(\mathcal{X}) + \pi P_{\mathrm{p}}(\mathcal{X}))\frac{(1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}}}{(1-\pi)P_{\mathrm{n}}^{A}(\mathcal{X}) + \pi P_{\mathrm{p}}(\mathcal{X})} \\ &= ((1-\pi)P_{\mathrm{n}}^{A^c}(A^c))\frac{P_{\mathrm{n}}^{A^c}}{P_{\mathrm{n}}^{A^c}(A^c)} + (\pi + (1-\pi)P_{\mathrm{n}}^{A}(A))\frac{(1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}}}{(1-\pi)P_{\mathrm{n}}^{A}(A) + \pi} \\ &= ((1-\pi)P_{\mathrm{n}}(A^c))\frac{P_{\mathrm{n}}^{A^c}}{P_{\mathrm{n}}(A^c)} + (\pi + (1-\pi)P_{\mathrm{n}}(A))\frac{(1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}}}{(1-\pi)P_{\mathrm{n}}(A) + \pi}, \end{aligned} \tag{6}$$

where the last two qualities are obtained by the definition of P_{n}^{A} and $P_{\mathrm{n}}^{A^c}$. Let $P_{\mathrm{n}'} = \frac{P_{\mathrm{n}}^{A^c}}{P_{\mathrm{n}}(A^c)}$, $P_{\mathrm{p}'} = \frac{(1-\pi)P_{\mathrm{n}}^{A} + \pi P_{\mathrm{p}}}{(1-\pi)P_{\mathrm{n}}(A) + \pi}$ and $\pi' = \pi + (1-\pi)P_{\mathrm{n}}(A)$, then Eq. (6) becomes

$$P_{\mathrm{u}} = (1-\pi')P_{\mathrm{n}'} + \pi' P_{\mathrm{p}'},$$

which shows that P_{u} can be made to a convex combination of new class-conditional distributions $P_{\mathrm{n}'}$ and $P_{\mathrm{p}'}$ by regrouping P_{n}^{A} with P_{p}.

Now we prove that $P_{\mathrm{n}'}$ and $P_{\mathrm{p}'}$ satisfy the anchor set assumption by checking whether $P_{\mathrm{n}'}(A) = 0$ and $P_{\mathrm{p}'}(A) > 0$.

By the definition of P_{n} and $P_{\mathrm{n}}^{A^c}$, we have

$$P_{\mathrm{n}'}(A) = \frac{P_{\mathrm{n}}^{A^c}(A)}{P_{\mathrm{n}}(A^c)} = 0. \tag{7}$$

By the definition of P_{p} and P_{n}^{A}, we have

$$P_{\mathrm{p}'}(A) = \frac{(1-\pi)P_{\mathrm{n}}^{A}(A) + \pi P_{\mathrm{p}}(A)}{(1-\pi)P_{\mathrm{n}}(A) + \pi} = \frac{(1-\pi)P_{\mathrm{n}}(A) + \pi P_{\mathrm{p}}(A)}{(1-\pi)P_{\mathrm{n}}(A) + \pi} = \frac{P_{\mathrm{u}}(A)}{(1-\pi)P_{\mathrm{n}}(A) + \pi} > 0. \tag{8}$$

The last inequality holds because $A \subset \mathrm{support}(P_{\mathrm{u}})$. By combining Eq. (7) and Ineq. (8), we can conclude that $P_{\mathrm{n}'}$ and $P_{\mathrm{p}'}$ satisfy the anchor set assumption. □

A.4 Proof of Theorem 2

Theorem 2. *Let* $P_{\mathrm{p}'}$ *and* $P_{\mathrm{n}'}$ *be obtained by regrouping a set* $A^* := \arg\min_{A \in \mathfrak{S}} \frac{P_{\mathrm{n}}(A)}{P_{\mathrm{p}}(A)}$[1] *from* P_{p} *and* P_{n}. *1). If* P_{n} *and* P_{p} *satisfy the irreduciblility assumption, then* $\pi' = \pi$*; 2). if* P_{n} *and* P_{p} *dissatisfy the irreduciblility assumption, then* $\pi < \pi' < \pi + (1-\pi)\beta^* = \kappa^*$.

[1] We have defined that the fraction tends to infinite if its numerator is larger than 0 and its denominator is 0. Additionally, the infimum may not be always exist, if it does not exist, we could use a sequence of sets that converges to the infimum value, but the convergence rate can be arbitrarily slow Scott (2015).

Proof. We define that a fraction tends to infinite if its numerator is larger than 0 and its denominator is 0. In this case, we could remove the constraint $P_{\mathrm{p}}(S) > 0$ in Eq. (2) and rewrite it to $\kappa^* = \pi + (1-\pi)\inf_{S\subseteq \mathrm{support}(P_{\mathrm{u}})} \frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}$. We subtract it with the new class prior after regrouping (Eq. (3)), i.e.,

$$\begin{aligned} \kappa^* - \pi' &= \pi + (1-\pi) \inf_{S\subseteq \mathrm{support}(P_{\mathrm{u}})} \frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)} - \pi - (1-\pi)P_{\mathrm{n}}(A^*) \\ &= (1-\pi)\left(\inf_{S\subseteq \mathrm{support}(P_{\mathrm{u}})} \frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)} - P_{\mathrm{n}}(A^*) \right). \end{aligned} \tag{9}$$

Not that $A^* := \arg\min_{A\in\mathfrak{S}} \frac{P_{\mathrm{n}}(A)}{P_{\mathrm{p}}(A)}$, if P_{n} is irreducible to P_{p}, $\inf_{S\subseteq \mathrm{support}(P_{\mathrm{u}})} \frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)} = 0$, so as $P_{\mathrm{n}}(A^*)$. Therefore $\kappa^* - \pi' = 0$ and $\pi' = \pi$.
If P_{n} is reducible to P_{p}, $P_{\mathrm{p}}(A^*) < 0$, then $\inf_{S\subseteq \mathrm{support}(P_{\mathrm{u}})} \frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)} > P_{\mathrm{n}}(A^*)$ and $\kappa^* - \pi' > 0$. Therefore $\pi < \pi'$ by Eq. (3), and $\pi < \pi' < \pi + (1-\pi)\beta = \kappa^*$. □

A.5 Proof of Theorem 3

For completeness, we illustrate the convergence property of ReCPE, which is presented by employing the estimator proposed by Blanchard et al. (2010).

Theorem 3. *Let $P_{\mathrm{u}} = (1-\pi)P_{\mathrm{n}} + \pi P_{\mathrm{p}}$. By selecting a set A and regrouping P_{n}^A to P_{p}. Then, with probability $1-2\delta$, the estimated class-prior $\hat{\pi}'$ based on solving $\inf_{S\in\mathfrak{S}, \hat{P}_{\mathrm{p}'}(S)>0} \frac{\hat{P}_{\mathrm{x}}(S)}{\hat{P}_{\mathrm{p}'}(S)}$ satisfies*

$$|\hat{\pi}' - \pi| \leq \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} + \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} + (1-\pi)P_{\mathrm{n}}(A), \tag{10}$$

where $\epsilon_{\delta,\mathcal{H}}(S) \triangleq 2\hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{4}{\delta}}{2|S|}}$ for a set S, and $\hat{\mathfrak{R}}_S(\mathcal{H})$ is the empirical Rademacher complexity of $\mathcal{H}$.

Proof. Firstly, we illustrate Rademacher complexity bounds. Let $\mathcal{H}$ be a family of functions taking values in $\{-1,+1\}$, and let $\mathcal{D}$ be the distribution over the input space $\mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1-\delta/2$ over a sample $S = (x_1,\ldots,x_m)$ of size m drawn according to $\mathcal{D}$, for any function $h \in \mathcal{H}$,

$$R(h) - \hat{R}_S(h) \leq 2\hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{4}{\delta}}{2m}}, \tag{11}$$

where $R(h)$ is the expected risk of the function h, and $\hat{R}_S(h)$ is the empirical risk of the function h on the sample S (Mohri et al., 2018). Specifically, let c be a target concept, then,

$$R(h) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[\mathbb{1}_{\{h(x_i)\neq c(x_i)\}}],\ \hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h(x_i)\neq c(x_i)\}}.$$

After regrouping P_{n}^A to P_{p} and creating $P_{\mathrm{p}'} = \frac{(1-\pi)P_{\mathrm{n}'}A + \pi P_{\mathrm{p}}}{(1-\pi)P_{\mathrm{n}'}(A)+\pi}$, P_{u} can be written as a mixture, i.e., $P_{\mathrm{u}} = (1-\pi')P_{\mathrm{n}'} + \pi' P_{\mathrm{p}'}$. Additionally, $P_{\mathrm{n}'}(A) = 0$ and $P_{\mathrm{p}'}(A) > 0$. Then,

$$P_{\mathrm{u}}(A) = (1-\pi')P_{\mathrm{n}'}(A) + \pi' P_{\mathrm{p}'}(A) = \pi' P_{\mathrm{p}'}(A). \tag{12}$$

In order to bring in the Rademacher complexity bounds to the above equation, we have to connect both $P_{\mathrm{u}}(A)$ and $P_{\mathrm{p}'}(A)$ with the expected risk. Let's define a function $h \in \mathcal{H}$ which is an indicator of the anchor set A. That is, $\forall x \in \mathcal{X}$,

$$h(x) = \begin{cases} 1, & x \in A \\ -1, & x \notin A, \end{cases} \tag{13}$$

By treating the sample i.i.d. drawn from the distribution P_{u} as positive, we can rewrite the $P_{\mathrm{u}}(A)$ as follows,

$$\begin{aligned} P_{\mathrm{u}}(A) &= \int_{x\in A} p_{\mathrm{u}}(x)dx = \int_{x\in\mathcal{X}} p_{\mathrm{u}}(x)\mathbb{1}_{\{h(x)=1\}}dx \\ &= 1-\int_{x\in\mathcal{X}} p_{\mathrm{u}}(x)\mathbb{1}_{\{h(x)\neq 1\}}dx = 1-\mathop{\mathbb{E}}_{x\sim P_{\mathrm{u}}}[\mathbb{1}_{\{h(x_i)\neq 1\}}] = 1-R_1(h), \end{aligned}$$

where $R_1(h)$ represents the false negative risk of the function h.

Similarly, by treating the sample i.i.d. drawn from the distribution $P_{\mathrm{p}'}$ as negative, , we can rewrite the $P_{\mathrm{p}'}(A)$ as follows,

$$P_{\mathrm{p}'}(A) = \int_{x\in A} f_{P_{\mathrm{p}'}}(x)dx = \int_{x\in\mathcal{X}} f_{P_{\mathrm{p}'}}(x)\mathbb{1}_{\{h(x)\neq 0\}}dx = \mathop{\mathbb{E}}_{x\sim P_{\mathrm{p}'}}[\mathbb{1}_{\{h(x_i)\neq 0\}}] = R_0(h),$$

where $R_0(h)$ represents the false positive risk of the function h.

Suppose we have samples S_{u} and $S_{\mathrm{p}'}$ with sample sizes $|S_{\mathrm{u}}|$ and $|S_{\mathrm{p}'}|$ i.i.d. drawn from P_{u} and $P_{\mathrm{p}'}$, respectively. Let $\hat{P}_{\mathrm{x}}(A)$ and $\hat{P}_{\mathrm{p}'}(A)$ be the empirical version of $P_{\mathrm{u}}(A)$ and $P_{\mathrm{p}'}(A)$, which are defined uniformly over the training samples, that is,

$$\begin{aligned} \hat{P}_{\mathrm{x}}(A) &= \frac{1}{|S_{\mathrm{u}}|}\sum_{x\in S_{\mathrm{u}}} \mathbb{1}_{\{h(x_i)=1\}} = 1-\frac{1}{|S_{\mathrm{u}}|}\sum_{x\in S_{\mathrm{u}}} \mathbb{1}_{\{h(x_i)\neq 1\}} = 1-\hat{R}_{1,S_{\mathrm{u}}}(h), &(14)\\ \hat{P}_{\mathrm{p}'}(A) &= \frac{1}{|S_{\mathrm{p}'}|}\sum_{x\in S_{\mathrm{p}'}} \mathbb{1}_{h(x_i)\neq 0} = \hat{R}_{0,S_{\mathrm{p}'}}(h). &(15) \end{aligned}$$

By Eq. (12), the estimated $\hat{\pi}'$ is

$$\hat{\pi}' = \frac{\hat{P}_{\mathrm{x}}(A)}{\hat{P}_{\mathrm{p}'}(A)}. \tag{16}$$

By using the Rademacher complexity bounds and union bound, with probability $1-\delta$, we have both

$$\begin{aligned} P_{\mathrm{u}}(A) &= 1-R_1(h) \geq 1-\hat{R}_{1,S_{\mathrm{u}}}(h) - \left(2\hat{\mathfrak{R}}_{S_{\mathrm{u}}}(\mathcal{H}) + 3\sqrt{\frac{\log\frac{4}{\delta}}{|S_{\mathrm{u}}|}}\right) \\ &\triangleq 1-\hat{R}_{1,S_{\mathrm{u}}}(h) - \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}}), \end{aligned} \tag{17}$$

and

$$\begin{aligned} P_{\mathrm{p}'}(A) &= R_0(h) \leq \hat{R}_{0,S_{\mathrm{p}'}}(h) + 2\hat{\mathfrak{R}}_{S_{\mathrm{p}'}}(\mathcal{H}) + 3\sqrt{\frac{\log\frac{4}{\delta}}{|S_{\mathrm{p}'}|}} \\ &\triangleq \hat{R}_{0,S_{\mathrm{p}'}}(h) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'}). \end{aligned} \tag{18}$$

Substituting $P_{\mathrm{u}}(A)$ and $P_{\mathrm{p}'}(A)$ in Eq. (12) with Eq. (17) and Eq. (18), we have

$$1-\hat{R}_{1,S_{\mathrm{u}}}(h) - \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}}) \leq \pi' P_{\mathrm{p}'}(A) \leq \pi'\left(\hat{R}_{0,S_{\mathrm{p}'}}(h) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})\right), \tag{19}$$

By Eq. (14) and Eq. (15), the above inequality can be rewritten as,

$$\hat{P}_{\mathrm{x}}(A) - \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}}) \leq \pi'\left(\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})\right).$$

Then we have that

$$
\begin{aligned}
\pi' &\geq \frac{\hat{P}_{\mathrm{x}}(A) - \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} \\
&= \frac{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'}) - \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} \frac{\hat{P}_{\mathrm{x}}(A) - \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A)} \\
&= \left(1 - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}\right)\left(\hat{\pi}' - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A)}\right) \\
&= \left(1 - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}\right)\hat{\pi}' - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} \\
&= \hat{\pi}' - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}\hat{\pi}' - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} \\
&\geq \hat{\pi}' - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} - \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}. \qquad (20)
\end{aligned}
$$

By the symmetric property of Eq. (10), with probability $1 - 2\delta$,

$$
|\hat{\pi}' - \pi'| \leq \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} + \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}. \qquad (21)
$$

According Eq. (3), $\pi' = \pi + (1-\pi)P_{\mathrm{n}}(A)$, then, with probability $1 - 2\delta$,

$$
|\hat{\pi}' - \pi| \leq \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} + \frac{\epsilon_{\delta,\mathcal{H}}(S_{\mathrm{u}})}{\hat{P}_{\mathrm{p}'}(A) + \epsilon_{\delta,\mathcal{H}}(S_{\mathrm{p}'})} + (1-\pi)P_{\mathrm{n}}(A).
$$

□

A.6 Proof of Theorem 4

Theorem 4. *Let p_{u} and p_{p} be density functions of P_{u} and P_{p}, respectively. Let $q = P(C = 0)p_{\mathrm{u}} + P(C = 1)p_{\mathrm{p}}$. Let $\mathbb{1}_A : \mathcal{X} \to \{0,1\}$ be the identity function which outputs 1 if $x \in \mathcal{X}$ is in the set A, and 0 otherwise. Then the set $A^* = \arg\min_{A \in \mathfrak{S}} \frac{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=0|X=x)]}{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=1|X=x)]}$.*

Recall that, in the main paper, we have defined another auxiliary distribution $q(X, C)$, where $C \in \{0, 1\}$ is the positive-vs-unlabeled label i.e., a class label distinguishing between the positive component and the whole mixture. Specifically, priors are $q(C = 1) := \frac{\pi}{1-\pi}$ and $q(C = 0) := \frac{1}{1-\pi}$; conditional densities are $q(X|C = 1) := P_{\mathrm{p}}$ and $q(X|C = 0) := P_{\mathrm{u}}$; class-posterior probabilities are $q(C = 0|X)$ and $q(C = 1|X)$.

Proof. Firstly, we prove that $\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}$ is proportional to $\frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)}$.

$$
\begin{aligned}
\frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)} &= \frac{(1-\pi)P_{\mathrm{n}}(S) + \pi P_{\mathrm{p}}(S)}{P_{\mathrm{p}}(S)} \\
&= (1-\pi)\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)} + \pi.
\end{aligned}
$$

Since $\frac{1}{1-\pi}$ and $\frac{\pi}{1-\pi}$ are constants, then $\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}$ is proportional to $\frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)}$, which completes the first part of the proof.

Recall that, in the main paper, we have defined another auxiliary distribution $q(X, C)$, where $C \in \{0, 1\}$ is the positive-vs-unlabeled label i.e., a class label distinguishing between the positive component and the whole mixture. Specifically, priors are $P(C = 1) := q(C = 1) := \frac{\pi}{1-\pi}$ and $P(C = 0) := q(C = 0) := \frac{1}{1-\pi}$; conditional densities are $q(X|C = 1) := p_{\mathrm{p}}$ and $q(X|C = 0) := p_{\mathrm{u}}$; class-posterior probabilities are $q(C = 0|X)$ and $q(C = 1|X)$. We have

$$
\frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)} = \frac{\int_{x\in S} q(X = x|C = 0)dx}{\int_{x\in S} q(X = x|C = 1)dx} = \frac{\int_{x\in \mathcal{X}} \mathbb{1}_A(X = x)q(X = x|C = 0)dx}{\int_{x\in \mathcal{X}} \mathbb{1}_A(X = x)q(X = x|C = 1)dx}. \qquad (22)
$$

By using Bayesian rules, the above equation can be written as,

$$\begin{aligned}
\frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)} &= \frac{\int_{x\in\mathcal{X}} \mathbb{1}_A(X=x)q(X=x|C=0)dx}{\int_{x\in\mathcal{X}} \mathbb{1}_A(X=x)q(X=x|C=1)dx} \\
&= \frac{P(C=1)}{P(C=0)}\frac{\int_{x\in\mathcal{X}} \mathbb{1}_A(X=x)q(C=0|X=x)q(x)dx}{\int_{x\in\mathcal{X}} \mathbb{1}_A(X=x)q(C=1|X=x)q(x)dx}. \\
&= \frac{P(C=1)}{P(C=0)}\frac{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=0|X=x)]}{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=1|X=x)]}.
\end{aligned}$$

Since $\frac{P(C=1)}{P(C=0)}$ is a constant, then $\frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)}$ is proportional to $\frac{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=0|X=x)]}{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=1|X=x)]}$. Combining with the first part of the proof, i.e., $\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}$ is proportional to $\frac{P_{\mathrm{u}}(S)}{P_{\mathrm{p}}(S)}$, we can conclude that $\frac{P_{\mathrm{n}}(S)}{P_{\mathrm{p}}(S)}$ is proportional to $\frac{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=0|X=x)]}{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=1|X=x)]}$. By definition of $A^* := \arg\min_{A\in\mathfrak{S}} \frac{P_{\mathrm{n}}(A)}{P_{\mathrm{p}}(A)}$, then $A^* = \arg\min_{A\in\mathfrak{S}} \frac{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=0|X=x)]}{\mathbb{E}_{x\sim q(X)}[\mathbb{1}_A(X=x)q(C=1|X=x)]}$, which completes the proof. □

A.7 PROOF OF PROPOSITION 2

Proposition 2. *Let* $P_{\tilde{\mathrm{p}}'} = \frac{P_{\mathrm{u}}^A+P_{\mathrm{p}}}{P_{\mathrm{u}}(A)+1}$ *and* $P_{\mathrm{u}}(A) < \epsilon$*. Then,* $\forall\epsilon > 0$ *and* $\forall S \in \mathfrak{S}$, $|P_{\mathrm{p}'}(S) - P_{\tilde{\mathrm{p}}'}(S)| \leq \mathcal{O}(\epsilon)$.

Proof. To prove $P_{\mathrm{p}'}$ is a good surrogate of $P_{\mathrm{p}'}$, we show that with the decreasing of $P_{\mathrm{u}}(A)$, the difference between $P_{\mathrm{p}'}$ and $P_{\tilde{\mathrm{p}}'}$ becomes smaller. Formally, let $P_{\mathrm{u}}(A) < \epsilon$. For all $\epsilon > 0$ and for all $S \in \mathfrak{S}$, $|P_{\mathrm{p}'}(S) - P_{\tilde{\mathrm{p}}'}(S)| \leq \mathcal{O}(\epsilon)$.

Note that the definitions of $P_{\mathrm{p}'}$ and $P_{\tilde{\mathrm{p}}'}$ are

$$P_{\mathrm{p}'} = \frac{(1-\pi)P_{\mathrm{n}}^A + \pi P_{\mathrm{p}}}{(1-\pi)P_{\mathrm{n}}(A) + \pi}; P_{\tilde{\mathrm{p}}'} = \frac{P_{\mathrm{u}}^A + P_{\mathrm{p}}}{P_{\mathrm{u}}(A)+1}.$$

We firstly start to prove that for all $\epsilon > 0$ and for all $S \in \mathfrak{S}$, $P_{\mathrm{p}'}(S) - P_{\tilde{\mathrm{p}}'}(S) \leq \mathcal{O}(\epsilon)$.

$$\begin{aligned}
P_{\mathrm{p}'}(S) - P_{\tilde{\mathrm{p}}'}(S) &= \frac{(1-\pi)P_{\mathrm{n}}^A(S) + \pi P_{\mathrm{p}}(S)}{(1-\pi)P_{\mathrm{n}}(A)+\pi} - \frac{P_{\mathrm{u}}^A(S) + P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A)+1} \\
&= \frac{(P_{\mathrm{u}}^A(S) - \pi P_{\mathrm{p}}^A(S)) + \pi P_{\mathrm{p}}(S)}{(1-\pi)P_{\mathrm{n}}(A)+\pi} - \frac{P_{\mathrm{u}}^A(S) + P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A)+1} \\
&\leq \frac{P_{\mathrm{u}}^A(S) + \pi P_{\mathrm{p}}(S)}{\pi} - \frac{P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A)+1} \\
&\leq \frac{P_{\mathrm{u}}^A(A) + \pi P_{\mathrm{p}}(S)}{\pi} - \frac{P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A)+1} \\
&= \frac{P_{\mathrm{u}}(A) + \pi P_{\mathrm{p}}(S)}{\pi} - \frac{P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A)+1} \\
&= \frac{P_{\mathrm{u}}(A)^2 + P_{\mathrm{u}}(A) + \pi P_{\mathrm{p}}(S)P_{\mathrm{u}}(A) + \pi P_{\mathrm{p}}(S) - \pi P_{\mathrm{p}}(S)}{\pi P_{\mathrm{u}}(A)+\pi} \\
&= \frac{P_{\mathrm{u}}(A)(P_{\mathrm{u}}(A) + \pi P_{\mathrm{p}}(S))}{\pi P_{\mathrm{u}}(A)+\pi} \\
&\leq \frac{P_{\mathrm{u}}(A)(P_{\mathrm{u}}(A) + \pi P_{\mathrm{p}}(S))}{\pi} \\
&= \mathcal{O}(\epsilon). \qquad (23)
\end{aligned}$$

We then prove that for all $\epsilon > 0$ and for all $S \in \mathfrak{S}$, $P_{\tilde{\mathrm{p}}'}(S) - P_{\mathrm{p}'}(S) \leq \mathcal{O}(\epsilon)$.

$$\begin{aligned}
P_{\tilde{\mathrm{p}}'}(S) - P_{\mathrm{p}'}(S) &= \frac{P_{\mathrm{u}}^A(S) + P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A) + 1} - \frac{(1-\pi)P_{\mathrm{n}}^A(S) + \pi P_{\mathrm{p}}(S)}{(1-\pi)P_{\mathrm{n}}(A) + \pi} \\
&\leq \frac{P_{\mathrm{u}}^A(S) + P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A) + 1} - \frac{(1-\pi)P_{\mathrm{n}}^A(S) + \pi P_{\mathrm{p}}(S)}{(1-\pi)P_{\mathrm{n}}(A) + \pi P_{\mathrm{p}}(A) + \pi} \\
&\leq \frac{P_{\mathrm{u}}^A(S) + P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A) + 1} - \frac{\pi P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A) + \pi} \\
&\leq \frac{P_{\mathrm{u}}^A(A) + P_{\mathrm{p}}(S)}{1} - \frac{\pi P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A) + \pi} \\
&= \frac{P_{\mathrm{u}}(A) + P_{\mathrm{p}}(S)}{1} - \frac{\pi P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A) + \pi} \\
&= \frac{P_{\mathrm{u}}(A)^2 + \pi P_{\mathrm{u}}(A) + P_{\mathrm{p}}(S)P_{\mathrm{u}}(A) + \pi P_{\mathrm{p}}(S) - \pi P_{\mathrm{p}}(S)}{P_{\mathrm{u}}(A) + \pi} \\
&= \frac{P_{\mathrm{u}}(A)(P_{\mathrm{u}}(A) + \pi + P_{\mathrm{p}}(S))}{P_{\mathrm{u}}(A) + \pi} \\
&\leq \frac{P_{\mathrm{u}}(A)(P_{\mathrm{u}}(A) + \pi + P_{\mathrm{p}}(S))}{\pi} \\
&= \mathcal{O}(\epsilon). \qquad (24)
\end{aligned}$$

By combining (23) and (24), for all $\epsilon > 0$ and for all $S \subseteq \mathcal{X}$, $|P_{\mathrm{p}'}(S) - P_{\tilde{\mathrm{p}}'}(S)| \leq \mathcal{O}(\epsilon)$, which completes the proof. □

B More Experimental Results

In this section, we provide more experimental results.

B.1 Estimation Errors on UCL Datasets

In Table 1, for each baseline method and its regrouped version, we report the average and variance of the absolute estimation errors and the p-values obtained by using Wilcoxon signed rank test. Note the, a small p-value reflects the error of the Regrouped-MPE is significantly smaller than the error of its baseline. The real-word datasets are downloaded from the UCL machine learning database[2].

	AM	ReAM	DPL	ReDPL	EN	ReEN	KM1	ReKM1	KM2	ReKM2	ROC	ReROC	RPG	ReRPG
adult	**0.127**	0.13	0.122	**0.108**∗	0.316	**0.295**	0.255	**0.132**	0.164	**0.153**	0.176	**0.153**	0.135	**0.134**
(800)	±0.005	±0.005	±0.006	±0.005	±0.005	±0.005	±0.051	±0.01	±0.009	±0.007	±0.01	±0.007	±0.004	±0.004
	$p = 0.413$		$p = \underline{0.036}$		$p = \underline{0.0}$		$p = \underline{0.0}$		$p = 0.182$		$p = 0.111$		$p = 0.16$	
adult	**0.122**	0.124	0.089*	0.089*	0.31	**0.29**	0.131	**0.091**	**0.12**	0.13	0.121	**0.095**	**0.123**	0.137
(1600)	±0.005	±0.004	±0.003	±0.003	±0.004	±0.005	±0.015	±0.008	±0.007	±0.006	±0.006	±0.005	±0.002	±0.004
	$p = 0.775$		$p = 0.485$		$p = \underline{0.0}$		$p = \underline{0.0}$		$p = 0.985$		$p = \underline{0.025}$		$p = 0.934$	
adult	0.105	**0.086**	**0.054**	0.057	0.297	**0.279**	0.054	**0.04**∗	**0.082**	0.089	0.089	**0.067**	**0.114**	0.128
(3200)	±0.003	±0.004	±0.001	±0.002	±0.003	±0.004	±0.001	±0.001	±0.003	±0.003	±0.005	±0.003	±0.002	±0.004
	$p = \underline{0.001}$		$p = 0.519$		$p = \underline{0.0}$		$p = \underline{0.0}$		$p = 0.879$		$p = \underline{0.009}$		$p = 0.98$	
avila	0.168	**0.152**	**0.129**	0.147	0.447	**0.422**	0.105	**0.075**∗	0.104	**0.081**	0.263	**0.228**	0.119	**0.111**
(800)	±0.011	±0.009	±0.005	±0.004	±0.004	±0.004	±0.007	±0.003	±0.004	±0.003	±0.011	±0.012	±0.007	±0.005
	$p = \underline{0.015}$		$p = 0.978$		$p = \underline{0.0}$		$p = \underline{0.0}$		$p = \underline{0.0}$		$p = \underline{0.024}$		$p = \underline{0.047}$	
avila	0.165	**0.132**	0.104	**0.084**	0.439	**0.418**	0.086	**0.076**∗	0.108	**0.092**	0.191	**0.16**	0.123	**0.121**
(1600)	±0.011	±0.01	±0.003	±0.003	±0.003	±0.003	±0.005	±0.004	±0.004	±0.003	±0.007	±0.01	±0.005	±0.005
	$p = \underline{0.0}$		$p = \underline{0.002}$		$p = \underline{0.0}$		$p = 0.133$		$p = \underline{0.005}$		$p = \underline{0.002}$		$p = 0.369$	
avila	0.156	**0.133**	**0.05**∗	0.061	0.436	**0.42**	0.092	**0.078**	0.112	**0.092**	0.131	**0.095**	**0.121**	0.122
(3200)	±0.012	±0.01	±0.001	±0.001	±0.002	±0.002	±0.005	±0.003	±0.007	±0.003	±0.005	±0.004	±0.005	±0.005
	$p = \underline{0.001}$		$p = 0.998$		$p = \underline{0.0}$		$p = 0.658$		$p = \underline{0.008}$		$p = \underline{0.0}$		$p = 0.601$	
bank	**0.135**	0.158	**0.116**∗	0.132	0.282	**0.264**	0.356	**0.216**	0.266	**0.238**	0.163	**0.15**	**0.163**	0.185
(800)	±0.011	±0.009	±0.004	±0.004	±0.013	±0.015	±0.086	±0.029	±0.036	±0.019	±0.004	±0.006	±0.01	±0.022
	$p = 0.992$		$p = 1.0$		$p = \underline{0.0}$		$p = \underline{0.0}$		$p = 0.088$		$p = 0.103$		$p = 0.995$	
bank	**0.117**	0.167	**0.087**∗	0.105	0.262	**0.244**	0.178	**0.128**	0.203	**0.198**	0.129	**0.118**	**0.157**	0.167
(1600)	±0.007	±0.015	±0.001	±0.002	±0.009	±0.01	±0.02	±0.013	±0.021	±0.015	±0.004	±0.005	±0.01	±0.011
	$p = 1.0$		$p = 1.0$		$p = \underline{0.0}$		$p = \underline{0.0}$		$p = 0.812$		$p = 0.119$		$p = 0.453$	

[2] UCL machine learning database.

`bank` (3200)	**0.104** 0.127 ±0.009 ±0.008 $p = 0.962$	**0.073**$*$ 0.091 ±0.002 ±0.002 $p = 1.0$	0.248 **0.237** ±0.007 ±0.008 $p = \underline{0.0}$	0.124 **0.09** ±0.008 ±0.004 $p = \underline{0.008}$	**0.15** 0.16 ±0.014 ±0.005 $p = 0.986$	**0.093** 0.106 ±0.003 ±0.003 $p = 0.947$	**0.159** 0.18 ±0.005 ±0.012 $p = 0.967$
`card` (800)	0.131 **0.127**$*$ ±0.007 ±0.007 $p = 0.71$	0.174 **0.161** ±0.007 ±0.009 $p = \underline{0.018}$	0.465 **0.444** ±0.029 ±0.03 $p = \underline{0.0}$	0.293 **0.176** ±0.041 ±0.013 $p = \underline{0.0}$	0.203 **0.158** ±0.025 ±0.015 $p = \underline{0.0}$	0.247 **0.233** ±0.019 ±0.021 $p = 0.207$	0.177 **0.155** ±0.013 ±0.011 $p = \underline{0.0}$
`card` (1600)	0.173 **0.14** ±0.009 ±0.009 $p = \underline{0.027}$	0.14 0.14 ±0.004 ±0.003 $p = 0.478$	0.459 **0.437** ±0.028 ±0.028 $p = \underline{0.0}$	0.19 **0.135** ±0.009 ±0.003 $p = \underline{0.0}$	0.159 **0.129** ±0.011 ±0.004 $p = \underline{0.003}$	0.194 **0.163** ±0.01 ±0.008 $p = 0.111$	0.126 **0.115**$*$ ±0.005 ±0.008 $p = \underline{0.0}$
`card` (3200)	0.164 **0.134** ±0.006 ±0.003 $p = \underline{0.009}$	0.127 **0.12** ±0.004 ±0.002 $p = 0.204$	0.455 **0.435** ±0.025 ±0.025 $p = \underline{0.0}$	0.161 **0.113** ±0.002 ±0.002 $p = \underline{0.0}$	0.142 **0.122** ±0.004 ±0.002 $p = \underline{0.0}$	0.159 **0.152** ±0.005 ±0.004 $p = 0.268$	0.11 **0.108**$*$ ±0.004 ±0.009 $p = 0.095$
`covtype` (800)	0.16 **0.123** ±0.01 ±0.006 $p = \underline{0.0}$	0.155 **0.151** ±0.006 ±0.005 $p = 0.255$	0.367 **0.343** ±0.003 ±0.004 $p = \underline{0.0}$	0.157 **0.142** ±0.011 ±0.009 $p = \underline{0.012}$	**0.122** 0.13 ±0.008 ±0.009 $p = 0.973$	0.291 **0.258** ±0.019 ±0.016 $p = \underline{0.027}$	0.116 **0.105**$*$ ±0.003 ±0.003 $p = \underline{0.003}$
`covtype` (1600)	0.12 **0.1**$*$ ±0.006 ±0.004 $p = \underline{0.004}$	0.132 **0.109** ±0.003 ±0.004 $p = \underline{0.002}$	0.364 **0.339** ±0.002 ±0.003 $p = \underline{0.0}$	0.116 **0.113** ±0.004 ±0.003 $p = 0.359$	**0.121** 0.123 ±0.005 ±0.005 $p = 0.768$	0.199 **0.161** ±0.014 ±0.01 $p = \underline{0.011}$	0.109 **0.108** ±0.003 ±0.003 $p = 0.257$
`covtype` (3200)	0.128 **0.09** ±0.003 ±0.003 $p = \underline{0.0}$	0.093 **0.083**$*$ ±0.003 ±0.002 $p = \underline{0.032}$	0.354 **0.334** ±0.001 ±0.002 $p = \underline{0.0}$	**0.097** 0.109 ±0.004 ±0.003 $p = 0.876$	**0.124** 0.128 ±0.003 ±0.004 $p = 0.825$	0.157 **0.113** ±0.009 ±0.004 $p = \underline{0.0}$	0.109 **0.107** ±0.003 ±0.003 $p = 0.154$
`egg` (800)	0.153 **0.106**$*$ ±0.011 ±0.007 $p = \underline{0.002}$	**0.218** 0.225 ±0.013 ±0.008 $p = 0.662$	0.505 0.505 ±0.005 ±0.006 $p = 0.433$	**0.173** 0.264 ±0.032 ±0.027 $p = 0.991$	**0.119** 0.131 ±0.007 ±0.008 $p = 0.789$	0.476 **0.396** ±0.022 ±0.03 $p = \underline{0.005}$	0.171 **0.124** ±0.02 ±0.009 $p = \underline{0.009}$
`egg` (1600)	0.137 **0.12** ±0.007 ±0.008 $p = 0.076$	**0.121** 0.142 ±0.006 ±0.005 $p = 0.992$	**0.486** 0.489 ±0.006 ±0.006 $p = 0.805$	0.234 **0.214** ±0.033 ±0.02 $p = \underline{0.018}$	0.116 **0.108**$*$ ±0.007 ±0.006 $p = \underline{0.047}$	0.315 **0.238** ±0.022 ±0.019 $p = \underline{0.002}$	0.151 **0.114** ±0.011 ±0.006 $p = \underline{0.0}$
`egg` (3200)	0.126 **0.113** ±0.006 ±0.006 $p = 0.117$	**0.057**$*$ 0.073 ±0.003 ±0.004 $p = 0.938$	**0.485** 0.489 ±0.012 ±0.011 $p = 0.958$	0.26 **0.193** ±0.02 ±0.017 $p = \underline{0.0}$	0.134 **0.113** ±0.007 ±0.006 $p = \underline{0.0}$	0.163 **0.139** ±0.009 ±0.008 $p = \underline{0.015}$	0.142 **0.102** ±0.008 ±0.005 $p = \underline{0.0}$
`magic04` (800)	0.099 **0.077** ±0.006 ±0.004 $p = \underline{0.012}$	0.072 **0.071** ±0.003 ±0.002 $p = 0.357$	0.312 **0.296** ±0.003 ±0.004 $p = \underline{0.0}$	0.111 **0.1** ±0.005 ±0.006 $p = \underline{0.0}$	0.071 **0.064** ±0.002 ±0.001 $p = 0.056$	0.141 **0.124** ±0.01 ±0.007 $p = 0.181$	0.055 **0.054**$*$ ±0.001 ±0.001 $p = 0.203$
`magic04` (1600)	0.071 **0.056** ±0.002 ±0.002 $p = \underline{0.001}$	0.044 **0.043**$*$ ±0.002 ±0.001 $p = 0.497$	0.292 **0.274** ±0.002 ±0.002 $p = \underline{0.0}$	0.084 **0.072** ±0.003 ±0.004 $p = \underline{0.0}$	0.079 **0.065** ±0.003 ±0.002 $p = \underline{0.0}$	0.1 **0.073** ±0.004 ±0.003 $p = \underline{0.002}$	0.058 **0.052** ±0.001 ±0.001 $p = \underline{0.003}$
`magic04` (3200)	0.069 **0.054** ±0.002 ±0.001 $p = \underline{0.0}$	**0.035**$*$ 0.036 ±0.001 ±0.002 $p = 0.562$	0.274 **0.258** ±0.001 ±0.001 $p = \underline{0.0}$	0.07 **0.047** ±0.003 ±0.002 $p = \underline{0.0}$	0.085 **0.063** ±0.002 ±0.002 $p = \underline{0.0}$	0.065 **0.047** ±0.003 ±0.002 $p = \underline{0.007}$	0.054 **0.052** ±0.001 ±0.001 $p = 0.176$
`robot` (800)	**0.053** 0.062 ±0.004 ±0.002 $p = 0.961$	0.049 **0.047**$*$ ±0.002 ±0.001 $p = 0.681$	0.19 **0.187** ±0.001 ±0.001 $p = 0.101$	0.232 **0.215** ±0.023 ±0.02 $p = 0.108$	**0.111** 0.114 ±0.007 ±0.007 $p = 0.975$	**0.119** 0.144 ±0.006 ±0.004 $p = 0.986$	**0.077** 0.084 ±0.003 ±0.003 $p = 0.838$
`robot` (1600)	0.053 **0.038**$*$ ±0.005 ±0.001 $p = 0.129$	0.087 **0.054** ±0.007 ±0.002 $p = \underline{0.0}$	0.139 **0.132** ±0.001 ±0.001 $p = \underline{0.0}$	0.15 **0.141** ±0.018 ±0.015 $p = \underline{0.003}$	**0.098** 0.099 ±0.005 ±0.005 $p = 0.849$	0.08 **0.075** ±0.004 ±0.002 $p = 0.477$	**0.076** 0.079 ±0.002 ±0.003 $p = 0.762$
`robot` (3200)	0.052 **0.039**$*$ ±0.003 ±0.002 $p = \underline{0.001}$	0.156 **0.119** ±0.01 ±0.007 $p = \underline{0.0}$	0.091 **0.085** ±0.0 ±0.0 $p = \underline{0.0}$	0.079 **0.077** ±0.007 ±0.006 $p = 0.161$	0.084 0.084 ±0.004 ±0.004 $p = 0.395$	0.063 **0.043** ±0.004 ±0.001 $p = 0.057$	**0.06** 0.066 ±0.002 ±0.003 $p = 0.988$
`shuttle` (800)	0.083 **0.031** ±0.039 ±0.001 $p = 0.271$	**0.016**$*$ 0.02 ±0.0 ±0.001 $p = 0.898$	0.041 **0.035** ±0.001 ±0.0 $p = \underline{0.0}$	**0.058** 0.083 ±0.002 ±0.003 $p = 1.0$	**0.035** 0.065 ±0.001 ±0.005 $p = 1.0$	**0.042** 0.047 ±0.001 ±0.002 $p = 0.699$	**0.035** 0.051 ±0.001 ±0.003 $p = 1.0$
`shuttle` (1600)	0.09 **0.045** ±0.048 ±0.011 $p = 0.958$	**0.011**$*$ 0.018 ±0.0 ±0.001 $p = 0.927$	0.04 **0.034** ±0.0 ±0.0 $p = \underline{0.0}$	**0.048** 0.079 ±0.001 ±0.003 $p = 1.0$	**0.024** 0.05 ±0.0 ±0.003 $p = 1.0$	**0.029** 0.043 ±0.001 ±0.003 $p = 0.913$	**0.026** 0.039 ±0.001 ±0.002 $p = 1.0$
`shuttle` (3200)	0.076 **0.028** ±0.039 ±0.0 $p = 0.949$	**0.012**$*$ 0.021 ±0.0 ±0.001 $p = 1.0$	0.043 **0.038** ±0.0 ±0.001 $p = \underline{0.004}$	**0.046** 0.07 ±0.001 ±0.002 $p = 1.0$	**0.018** 0.03 ±0.0 ±0.001 $p = 0.999$	**0.038** 0.045 ±0.005 ±0.004 $p = 0.811$	**0.028** 0.042 ±0.001 ±0.002 $p = 1.0$
`average`	0.116 **0.1** ±0.012 ±0.007 $p = \underline{0.0}$	0.094 **0.092**$*$ ±0.006 ±0.006 $p = 0.279$	0.311 **0.297** ±0.026 ±0.026 $p = \underline{0.0}$	0.146 **0.121** ±0.022 ±0.012 $p = \underline{0.0}$	0.117 **0.111** ±0.01 ±0.008 $p = \underline{0.0}$	0.157 **0.136** ±0.018 ±0.014 $p = \underline{0.0}$	0.106 **0.105** ±0.007 ±0.008 $p = \underline{0.002}$

Table 1: The first column provides the names of the datasets and the sample lengths. We bold the smaller average estimation errors by comparing each baseline method with its regrouped version. The smallest average estimation error among all methods in each row is highlighted with $*$. p-values are obtained by using the one-sided Wilcoxon signed rank test. We underline the p-values which are smaller than the 0.05 significant level. The last column is calculated by averaging trials on all the different datasets. The proposed Regrouping method provides significantly more accurate estimations than all the baseline.

REFERENCES

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.

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

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Artificial Intelligence and Statistics*, pp. 838–846, 2015.