Machine Learning (2023) 112:889–919
https://doi.org/10.1007/s10994-022-06190-z

Positive-unlabeled classification under class-prior shift: a prior-invariant approach based on density ratio estimation

Shota Nakajima[1] · **Masashi Sugiyama**[1,2]

Received: 29 October 2021 / Revised: 16 February 2022 / Accepted: 15 May 2022 /
Published online: 27 June 2022

Abstract
Learning from positive and unlabeled (PU) data is an important problem in various applications. Most of the recent approaches for PU classification assume that the class-prior (the ratio of positive samples) in the training unlabeled dataset is identical to that of the test data, which does not hold in many practical cases. In addition, we usually do not know the class-priors of the training and test data, thus we have no clue on how to train a classifier without them. To address these problems, we propose a novel PU classification method based on density ratio estimation. A notable advantage of our proposed method is that it does not require the class-priors in the training phase; class-prior shift is incorporated only in the test phase. We theoretically justify our proposed method and experimentally demonstrate its effectiveness.

Keywords Positive-unlabeled classification · Class-prior shift · Density ratio estimation

1 Introduction

Positive-unlabeled (PU) classification is a problem of training a binary classifier from only positive and unlabeled data (Letouzey et al., 2000; Elkan & Noto, 2008). This problem is important when it is difficult to gather negative data, and appears in many applications, such as inlier-based outlier detection (Blanchard et al., 2010), land-cover classification (Li et al., 2011), matrix completion (Hsieh et al., 2015), and sequential data classification (Li et al., 2009; Nguyen et al., 2011). Several heuristic approaches for PU classification have been proposed in the past (Liu et al., 2003; Li & Liu, 2003),

Editors: Dana Drachsler Cohen, Javier Garcia, Mohammad Ghavamzadeh, Marek Petrik, Philip S. Thomas.

✉ Shota Nakajima
nakajima@alumni.u-tokyo.ac.jp

Masashi Sugiyama
sugi@k.u-tokyo.ac.jp

[1] The University of Tokyo, Tokyo, Japan

[2] RIKEN, Tokyo, Japan

which aim to identify negative samples in the unlabeled dataset, yet they heavily rely on the heuristic strategy and data separability assumption. One of the most theoretically and practically effective methods for PU classification was established by Plessis et al. (2014, 2015), called *unbiased PU classification*. It rewrites the classification risk in terms of the distributions over positive and unlabeled samples, and obtains an unbiased estimator of the risk without negative samples. Although unbiased PU classification works well with simple models such as linear-in-parameter models, it easily suffers from overfitting with flexible models such as deep neural networks. To overcome this problem, a *non-negative risk estimator* (Kiryo et al., 2017) for PU classification was proposed.

Besides unbiased PU classification, various approaches for PU classification have also been proposed recently. For example, generative adversarial networks (GAN) have been applied to PU classification by Hou et al. (2018), which allows one to learn from a small number of positive samples. Zhang et al. (2019) introduced *ambiguity* to unlabeled samples and performed PU label disambiguation (PULD) based on margin maximization to determine the true labels of all unlabeled examples. A variational approach and a data augmentation method based on *Mixup* (Zhang et al., 2018) were proposed by Chen et al. (2020) for PU classification without explicit estimation of the class-prior of the training data.

One of the drawbacks of these approaches is that the distribution of the test data must be identical to that of the training data, which may be violated in practice (Quionero-Candela et al., 2009). For example, the class-prior (the ratio of positive data) in the training unlabeled dataset might be different from that of the test data, known as the *class-prior shift* problem. To cope with this problem, Charoenphakdee and Sugiyama (2019) showed that classification under class-prior shift can be written as cost-sensitive classification, and proposed a risk-minimization approach and a density ratio estimation (Sugiyama et al., 2012) approach. In their study, both the class-priors of the training and test data are assumed to be given in advance, but this assumption does not hold in many practical cases. Therefore, we need to estimate them with the training and test data.

However, it is usually hard to obtain samples from the test distribution at the training time, and this is not natural because we do not know whether the prior shift would occur or not in advance. Furthermore, the training data would be inaccessible once the training has been completed, especially in privacy-concerned situations such as click analysis (McMahan et al., 2013), purchase prediction (Martínez et al., 2018), and voting prediction (Coletto et al., 2015). In that kind of problem, the model is trained with data including personal information, and only the trained model is kept while the data must be discarded. This implies that we are not allowed to use training data when a classifier is adapted to an unknown test distribution.

To overcome these problems, we propose an approach based on density ratio estimation (Sugiyama et al., 2012). Density ratio estimation for PU classification has appeared in several existing works (Charoenphakdee & Sugiyama, 2019; Kato et al., 2019), yet their studies have no guarantees on the theoretical relationship between binary classification and density ratio estimation. Our proposed method can train a classifier without given knowledge of the class-priors, and adapt to the test-time class-prior shift without the training data. Table 1 summarizes comparisons of representative existing methods and our proposed method. Our main contributions are: (i) We propose a method for PU classification under test-time class-prior shift with unknown class-priors, (ii) We theoretically justify the proposed method, and (iii) Experimental results show the effectiveness of the proposed method.

Table 1 Comparisons of representative existing PU classification methods. uPU was proposed by Plessis et al. (2014, 2015), nnPU was proposed by Kiryo et al. (2017), GenPU was proposed by Hou et al. (2018), PULD was proposed by Zhang et al. (2019), VPU was proposed by Chen et al. (2020), and PUa was proposed by Charoenphakdee and Sugiyama (2019)

	uPU	nnPU	GenPU	PULD	VPU	PUa	Ours
Excess risk bound and its convergence rate analysis	✓	✓	×	✓	×	×	✓
Learning a classifier without given class-prior(s)	×	×	×	×	✓	×	✓
Adaptable to test-time class-prior shift	×	×	×	×	×	✓	✓

2 Preliminaries

In this section, we introduce the notations, and review the concepts of unbiased/non-negative PU classification, cost-sensitive classification, and density ratio estimation.

2.1 Problem formulation

Let $X \in \mathbb{R}^d$ and $Y \in \{\pm 1\}$ be the input and output random variables, where d denotes the dimensionality of the input variable. Let $p(x, y)$ be the *underlying joint density* of (X, Y) and $p(x)$ be the input marginal density. We denote the positive and negative class-conditional densities as

$$\begin{aligned} p_+(x) &= p(x \mid Y = +1) \\ p_-(x) &= p(x \mid Y = -1). \end{aligned} \tag{1}$$

Let $\pi = p(Y = +1)$ be the positive *class-prior probability*. Assume we have i.i.d. sample sets $\mathcal{X}_\mathrm{P}$ and $\mathcal{X}_\mathrm{U}$ from $p_+(x)$ and $p(x)$ respectively, and let $n_\mathrm{P} = |\mathcal{X}_\mathrm{P}|$ and $n_\mathrm{U} = |\mathcal{X}_\mathrm{U}|$, where $|\cdot|$ denotes the cardinality of a set. We denote the expectations over each class-conditional density as

$$\begin{aligned} \mathbb{E}_\mathrm{P}[\cdot] &= \mathbb{E}_{X \sim p_+}[\cdot] \\ \mathbb{E}_\mathrm{N}[\cdot] &= \mathbb{E}_{X \sim p_-}[\cdot] \\ \mathbb{E}_\mathrm{U}[\cdot] &= \mathbb{E}_{X \sim p}[\cdot] = \mathbb{E}_X[\cdot], \end{aligned} \tag{2}$$

and their empirical counterparts as

$$\begin{aligned} \widehat{\mathbb{E}}_\mathrm{P}\big[f(X)\big] &= \frac{1}{n_\mathrm{P}} \sum_{x \in \mathcal{X}_\mathrm{P}} f(x) \\ \widehat{\mathbb{E}}_\mathrm{U}\big[f(X)\big] &= \frac{1}{n_\mathrm{U}} \sum_{x \in \mathcal{X}_\mathrm{U}} f(x), \end{aligned} \tag{3}$$

where f is an arbitrary function of $x \in \mathbb{R}^d$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a real-valued decision function. The purpose of binary classification is to minimize the expected classification risk

$$R(g) = \mathbb{E}_{X,Y}\big[1\{\mathrm{sign}(g(X)) \neq Y\}\big], \tag{4}$$

where $1\{\cdot\}$ denotes the indicator function. Since the optimization problem based on the zero-one loss is computationally infeasible (Arora et al., 1997; Bartlett et al., 2006), a surrogate loss function $\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$ is used in practice. Classification risk with respect to surrogate loss is defined as

$$R_\ell(g) = \mathbb{E}_{X,Y}\big[\ell(g(X), Y)\big]. \tag{5}$$

2.2 Unbiased/non-negative PU classification

The surrogate classification risk can be written as

$$R_\ell(g) = \pi\mathbb{E}_{\mathrm{P}}\big[\ell(g(X), +1)\big] + (1-\pi)\mathbb{E}_{\mathrm{N}}\big[\ell(g(X), -1)\big]. \tag{6}$$

Since negative samples are unavailable in PU classification, we rewrite the expectation over the negative class-conditional distribution as

$$(1-\pi)\mathbb{E}_{\mathrm{N}}\big[\ell(g(X), -1)\big] = \mathbb{E}_{\mathrm{U}}\big[\ell(g(X), -1)\big] - \pi\mathbb{E}_{\mathrm{P}}\big[\ell(g(X), -1)\big], \tag{7}$$

where $p(x) = \pi p_+(x) + (1-\pi)p_-(x)$ is used (Plessis et al., 2014). Then, the risk can be approximated directly with $\mathcal{X}_{\mathrm{P}}$ and $\mathcal{X}_{\mathrm{U}}$ as

$$\widehat{R}_\ell(g) = \pi\widehat{\mathbb{E}}_{\mathrm{P}}\big[\ell(g(X), +1)\big] - \pi\widehat{\mathbb{E}}_{\mathrm{P}}\big[\ell(g(X), -1)\big] + \widehat{\mathbb{E}}_{\mathrm{U}}\big[\ell(g(X), -1)\big]. \tag{8}$$

The *empirical risk estimator* $\widehat{R}_\ell(g)$ is unbiased and consistent (Niu et al., 2016), i.e., $\mathbb{E}\big[\widehat{R}_\ell(g)\big] = R_\ell(g)$ where the expectation $\mathbb{E}$ is taken over all of the samples, and $\widehat{R}_\ell(g) \to R_\ell(g)$ as $n_{\mathrm{P}}, n_{\mathrm{U}} \to \infty$.

Unbiased PU classification easily suffers from overfitting when we use a flexible model such as neural networks, because the model can be so powerful that it fits all of the given samples, and then the empirical risk goes negative (Kiryo et al., 2017). To mitigate this problem, a non-negative risk correction approach was proposed (Kiryo et al., 2017). Since

$$\mathbb{E}_{\mathrm{U}}\big[\ell(g(X), -1)\big] - \pi\mathbb{E}_{\mathrm{P}}\big[\ell(g(X), -1)\big] = (1-\pi)\mathbb{E}_{\mathrm{N}}\big[\ell(g(X), -1)\big] \geq 0 \tag{9}$$

holds for any non-negative loss function, we correct the corresponding part of the expected risk to be non-negative. Approximating the expectations by sample averages gives the non-negative risk estimator:

$$\widetilde{R}_\ell(g) = \pi\widehat{\mathbb{E}}_{\mathrm{P}}\big[\ell(g(X), +1)\big] + \Big(\widehat{\mathbb{E}}_{\mathrm{U}}\big[\ell(g(X), -1)\big] - \pi\widehat{\mathbb{E}}_{\mathrm{P}}\big[\ell(g(X), -1)\big]\Big)_+, \tag{10}$$

where $(\cdot)_+ = \max(0, \cdot)$. The non-negative risk estimator is biased yet consistent, and its bias decreases exponentially with respect to $n_{\mathrm{P}} + n_{\mathrm{U}}$ (Kiryo et al., 2017).

2.3 Cost-sensitive classification

For arbitrary *false-positive cost* parameter $c \in (0, 1)$, cost-sensitive classification is defined as a problem of minimize the following risk (Elkan, 2001; Scott, 2012) :

$$R_{\pi,c}(g) = (1-c)\pi\mathbb{E}_{\mathrm{P}}\big[1\{\mathrm{sign}(g(X)) \neq +1\}\big] + c(1-\pi)\mathbb{E}_{\mathrm{N}}\big[1\{\mathrm{sign}(g(X)) \neq -1\}\big]. \tag{11}$$

When $c = 1/2$, cost-sensitive classification reduces to ordinary binary classification, up to unessential scaling factor 1/2. (Charoenphakdee & Sugiyama, 2019) showed that classification under class-prior shift can be formulated as cost-sensitive classification. For example, let $\pi' \in (0, 1)$ be the class-prior of the test distribution, then $R_{\pi',1/2} \propto R_{\pi,c}$ with $c = \frac{\pi(1-\pi')}{\pi(1-\pi')+(1-\pi)\pi'}$.

2.4 Class-prior estimation

In unbiased/non-negative PU classification, the class-prior is assumed to be given, which does not hold in many practical cases. Unfortunately, we cannot treat π as a hyperparameter to be tuned, because there exists a trivial solution such as $\pi = 0$ and $g(x) \equiv \mathrm{argmin}_v \ell(-1, v)$. One of the solutions to this problem is to estimate both the training and test class-priors by existing methods respectively with positive, training-unlabeled, and test-unlabeled datasets. In fact, it is known that class-prior estimation is an ill-posed problem, without any additional assumptions (Blanchard et al., 2010; Scott et al., 2013). For example, if

$$p(x) = \kappa p_+(x) + (1 - \kappa)p_-(x) \tag{12}$$

holds, then there exists a density $p'_-(x)$ such that

$$p(x) = (\kappa - \delta)p_+(x) + (1 - \kappa + \delta)p'_-(x) \tag{13}$$

for $0 \leq \delta \leq \kappa$. In practice, an alternative goal of estimating the maximum mixture proportion

$$\kappa^* = \max\{\kappa \in [0, 1] : \exists p_- \text{ s.t. } p(x) = \kappa p_+(x) + (1 - \kappa)p_-(x)\} \tag{14}$$

is pursued (Blanchard et al., 2010; Scott et al., 2013). The *irreducibility assumption* (Blanchard et al., 2010; Scott et al., 2013) gives a constraint on the true underlying densities which ensures that κ^* is the unique solution of prior estimation.

Definition 1 (Irreducibility (Blanchard et al., 2010; Scott et al., 2013)) Let G and H be probability distributions on $(\mathbb{R}^d, \mathfrak{S})$ where $\mathfrak{S}$ is a Borel algebra on $\mathbb{R}^d$. We say that G is irreducible with respect to H, if there is no decomposition of the form $G = \kappa H + (1 - \kappa)H'$ where H' is some probability distribution and $0 < \kappa \leq 1$.

Let P, P_+, and P_- be the cumulative distribution functions of p, p_+, and p_- respectively. Under the irreducibility assumption, the class-prior is identical to the maximum mixture proportion.

Proposition 1 [(*Blanchard et al.*, 2010; *Scott et al.*, 2013)] *Let P, P_+, and P_- be probability distributions on $(\mathbb{R}^d, \mathfrak{S})$. If $P = \pi P_+ + (1 - \pi)P_-$ and P_- is irreducible with respect to P_+, then*

$$\begin{aligned} \pi &= \max\{\kappa \in [0, 1] : \exists Q \text{ s.t. } P = \kappa P_+ + (1 - \kappa)Q\} \\ &= \inf_{S \in \mathfrak{S}, P_+(S) > 0} \frac{P(S)}{P_+(S)}. \end{aligned} \tag{15}$$

Note that the set $S \in \mathfrak{S}$ corresponds to a measurable hypothesis $h : \mathbb{R}^d \to \{-1, +1\}$ bijectively. Based on these facts, several works for class-prior estimation have been proposed (Blanchard et al., 2010; Scott et al., 2013; Scott, 2015; Ramaswamy et al., 2016; Plessis et al., 2016). However, they usually work with kernel methods which are computationally hard to apply to large-scale and high-dimensional data. Furthermore, since the unbiased/non-negative risk estimators depend on the class-prior, an estimation error of the class-prior directly affects the optimization. In addition, we usually do not have a sample set from the test distribution at the training time, and thus cannot even estimate the class-prior of the test data by such existing methods.

2.5 Density ratio estimation

The ratio of two probability densities has attracted attention in various problems (Sugiyama et al., 2009, 2012). Density ratio estimation (DRE) (Sugiyama et al., 2012) aims to directly estimate the ratio of two probabilities, instead of estimating the two densities separately. Sugiyama et al. (2011) showed that various existing DRE methods (Sugiyama et al., 2008; Kanamori et al., 2009; Kato et al., 2019) can be unified from the viewpoint of Bregman divergence minimization, so we consider the DRE problem as a Bregman divergence minimization problem.

Here we consider estimating the ratio of the positive class-conditional density to the input marginal density. Let $r^*(x) = p_+(x)/p(x)$ be the true density ratio and $r : \mathbb{R}^d \to [0, \infty)$ be a density ratio model. For a convex and differentiable function $f : [0, \infty) \to \mathbb{R}$, the expected Bregman divergence, which measures the discrepancy from r^* to r, is defined as

$$\begin{aligned} \mathrm{BR}_f(r^* \parallel r) &= \int \Big(f(r^*(x)) - f(r(x)) - f'(r(x))(r^*(x) - r(x))\Big)p(x)\mathrm{d}x \\ &= \mathbb{E}_{\mathrm{P}}\big[-f'(r(X))\big] + \mathbb{E}_{\mathrm{U}}\big[f'(r(X))r(X) - f(r(X))\big] + \text{const.}, \end{aligned} \tag{16}$$

where the constant term does not include r. The function f is called the generator function of the Bregman divergence (Menon & Ong, 2016). We can see that the Bregman divergence of DRE does not contain the class-prior π, and can be approximated by taking empirical averages over the positive and unlabeled datasets, except the constant term.

Similarly to the case of unbiased PU classification, it was revealed that empirical Bregman divergence minimization often suffers from severe overfitting when we use a highly flexible model (Kato & Teshima, 2021). To mitigate this problem, non-negative risk correction for the Bregman divergence minimization was proposed, based on the idea of non-negative PU classification (Kiryo et al., 2017).

The objective function for Bregman divergence minimization is defined by Eq. (16) without the constant term

$$\mathcal{L}_f(r) = \mathbb{E}_{\mathrm{P}}\big[-f'(r(X))\big] + \mathbb{E}_{\mathrm{U}}\big[f'(r(X))r(X) - f(r(X))\big]. \tag{17}$$

We also consider its empirical counterpart $\widehat{\mathcal{L}}_f(r)$. Let us denote

$$\begin{aligned} f^*(t) &= tf'(t) - f(t) \\ \mathfrak{F}(t) &= f^*(t) - f^*(0), \end{aligned} \tag{18}$$

then $\mathfrak{F}$ is non-negative on $[0, \infty)$ because $(f^*)'(t) = tf''(t) \geq 0$ (i.e., f is convex.). We pick a lower bound of π as α and then we have

$$\mathbb{E}_{\mathrm{U}}\left[\mathfrak{F}(r(X))\right] - \alpha\mathbb{E}_{\mathrm{P}}\left[\mathfrak{F}(r(X))\right] \geq (1-\pi)\mathbb{E}_{\mathrm{N}}\left[\mathfrak{F}(r(X))\right] \geq 0. \tag{19}$$

Thus, we define the corrected empirical estimator of $\mathcal{L}$ as

$$\begin{aligned}\widetilde{\mathcal{L}}_f(r) = \widehat{\mathbb{E}}_{\mathrm{P}}\left[-f'(r(X)) + \alpha\mathfrak{F}(r(X))\right] \\ + \left(\widehat{\mathbb{E}}_{\mathrm{U}}\left[\mathfrak{F}(r(X))\right] - \alpha\widehat{\mathbb{E}}_{\mathrm{P}}\left[\mathfrak{F}(r(X))\right]\right)_+ + f^*(0),\end{aligned} \tag{20}$$

where $(\cdot)_+ = \max(0, \cdot)$. $\widetilde{\mathcal{L}}_f$ is consistent as long as $0 \leq \alpha \leq \pi$, and its bias decreases exponentially with respect to $n_{\mathrm{P}} + n_{\mathrm{U}}$. Even though we do not have any knowledge of π, we can tune α as a hyperparameter to minimize the empirical estimator without non-negative correction $\widehat{\mathcal{L}}_f(r)$, which contains neither π nor α, with the positive and unlabeled validation datasets.

3 Density ratio estimation for PU learning

In this section, we formulate a cost-sensitive binary classification problem as a density ratio estimation problem, and propose a method of Density Ratio estimation for PU learning (DRPU). All proofs are given in Appendix A.

3.1 Excess risk bounds

From Bayes' rule, we have $p(Y = +1 \mid X) = \pi p_+(x)/p(x) = \pi r^*(x)$. Therefore, the optimal solution of the Bregman divergence minimization, $r = r^*$ gives a Bayes optimal classifier by thresholding $p(Y = +1 \mid X) = 1/2$, and it motivates us to use r for binary classification. However, this statement only covers the optimal solution, and it is unclear how the classification risk grows along with the Bregman divergence. To cope with this problem, we interpret the DRE as minimization of an upper bound of the excess classification risk. Although the relationship between DRE and class probability estimation has been studied by Menon and Ong (2016), differently from that, our work focuses on the ratio of the densities of positive and unlabeled data, and the value of the Bregman divergence does not depend on the class-prior π.

We denote the *Bayes optimal risk* as $R^*_{\pi,c} = \inf_{g\in\mathcal{F}} R_{\pi,c}(g)$, where $\mathcal{F}$ is the set of all measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$, and the difference $R_{\pi,c(g)} - R^*_{\pi,c}$ is called the *excess risk* for $R_{\pi,c}$. The following theorem associates DRE with cost-sensitive classification under a strong convexity assumption on f, justifying solving binary classification by DRE.

Theorem 1 *Let f be a μ-strongly convex function, i.e., $\mu = \inf_{t\in[0,\infty)} f''(t) > 0$. Then, for any $\pi \in (0,1)$, $c \in (0,1)$, $r : \mathbb{R}^d \to [0,\infty)$, and $h_\theta = \mathrm{sign}(r - \theta)$, we have*

$$R_{\pi,c}(h_\theta) - R^*_{\pi,c} \leq \pi\sqrt{\frac{2}{\mu}\mathrm{BR}_f(r^* \parallel r)}, \tag{21}$$

where $\theta = c/\pi$.

As we have already seen in Sect. 2.3, the class-prior shift problem can be transformed into a cost-sensitive classification problem. Next, we extend the excess risk bound in Theorem 1 to the case of prior shift. Let π' be the test-time class-prior and c' be the test-time

false positive cost. Note that $c' = 1/2$ corresponds to standard binary classification. The classification risk with respect to the test distribution is defined as

$$\begin{aligned} R_{\pi',c'}(g) = (1-c')\pi' \mathbb{E}_{\mathrm{P}}\left[1\{\mathrm{sign}(g(X)) \neq +1\}\right] \\ + c'(1-\pi')\mathbb{E}_{\mathrm{N}}\left[1\{\mathrm{sign}(g(X)) \neq -1\}\right]. \end{aligned} \tag{22}$$

The following theorem gives an excess risk bound for $R_{\pi',c'}$.

Theorem 2 *Let f be a μ-strongly convex function. Then, for any $\pi, \pi' \in (0,1)$, $c' \in (0,1)$, $r : \mathbb{R}^d \to [0,\infty)$, and $h_\theta = \mathrm{sign}(r - \theta)$, we have*

$$R_{\pi',c'}(h_\theta) - R^*_{\pi',c'} \leq C\sqrt{\frac{2}{\mu}\mathrm{BR}_f(r^* \parallel r)}, \tag{23}$$

where $c_0 = \frac{c'\pi(1-\pi')}{(1-c')(1-\pi)\pi' + c'\pi(1-\pi')}$, $\theta = c_0/\pi$, and $C = \pi\frac{c'+\pi'-2c'\pi'}{c_0+\pi-2c_0\pi}$.

Note that this is a generalized version of Theorem 1, which is the case of $\pi' = \pi$ and $c' = c$. This result shows that even when the class-prior and cost are shifted, by changing the classification threshold to c_0, the classification risk can still be bounded by the Bregman divergence of DRE.

3.2 Estimating the class-priors

Although we can train a model r and deal with a prior shift problem without the knowledge of the class-priors, we still need to estimate them to determine the classification threshold c_0/π. Based on Proposition 1, we propose the following estimator of π:

$$\hat{\pi}(r) = \inf_{h \in \mathcal{H}_r} \frac{\widehat{P}(h)}{\widehat{P}_+(h)}, \tag{24}$$

where

$$\begin{aligned} \widehat{P}(h) &= \widehat{\mathbb{E}}_{\mathrm{U}}[1\{h(X) = +1\}] \\ \widehat{P}_+(h) &= \widehat{\mathbb{E}}_{\mathrm{P}}[1\{h(X) = +1\}] \end{aligned} \tag{25}$$

and

$$\mathcal{H}_r = \left\{h : \mathbb{R}^d \to \{\pm 1\} \mid \exists \theta \in \mathbb{R}, h(x) = \mathrm{sign}(r(x) - \theta) \wedge \widehat{P}_+(h) > \bar{\gamma}\right\}. \tag{26}$$

Here,

$$\bar{\gamma} = \frac{1}{\gamma}\max(\varepsilon(n_{\mathrm{P}}, 1/n_{\mathrm{P}}), \varepsilon(n_{\mathrm{U}}, 1/n_{\mathrm{U}})) \tag{27}$$

and,

$$\varepsilon(n,\delta) = \sqrt{\frac{4\log(en/2)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{28}$$

for some $n > 0$ and $0 < \delta < 1$, and $0 < \gamma < 1$ is an arbitrarily fixed constant. The main difference from the estimator proposed by Blanchard et al. (2010) and Scott et al. (2013) is that the hypothesis h is determined by thresholding the trained density ratio model r, thus we need no additional training of the model.

To consider convergence of the proposed estimator, we introduce the concept of the *Area Under the receiver operating characteristic Curve (AUC)*, which is a criterion to measure the performance of a score function for bipartite ranking (Menon & Williamson, 2016). For any real-valued score function $s : \mathbb{R}^d \to \mathbb{R}$, AUC is defined as

$$\mathrm{AUC}(s) = \mathbb{E}\big[1\{(Y - Y')(s(X) - s(X')) > 0\} \mid Y \neq Y'\big], \tag{29}$$

where the expectation is taken over X, X', Y, and Y'. In addition, we define the AUC risk and the optimal AUC risk as $R_{\mathrm{AUC}}(r) = 1 - \mathrm{AUC}(r)$ and $R^*_{\mathrm{AUC}} = \inf_{s \in \mathcal{F}} R_{\mathrm{AUC}}(s)$, where $\mathcal{F}$ is a set of all measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$. Then, the following theorem gives a convergence guarantee of the estimator.

Theorem 3 *For $\hat{\pi}(r)$ defined by Eq. (24), with probability at least $(1 - 1/n_{\mathrm{P}})(1 - 1/n_{\mathrm{U}})$, we have*

$$|\hat{\pi}(r) - \pi| \leq \xi\big(R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}}\big) + \mathcal{O}\left(\sqrt{\frac{\log n_{\mathrm{P}}}{n_{\mathrm{P}}}} + \sqrt{\frac{\log n_{\mathrm{U}}}{n_{\mathrm{U}}}}\right). \tag{30}$$

Here, ξ is an increasing function such that

$$\xi\big(R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}}\big) \leq \frac{2(1-\pi)}{1-\bar{\gamma}^2} R_{\mathrm{AUC}}(r), \tag{31}$$

and $\xi(0) \to 0$ as $\bar{\gamma} \to 0$.

This result shows that a *better score function* in the sense of AUC tends to result in a *better estimation* of π. Furthermore, we can see that the scale of r is not important for the estimator $\hat{\pi}(r)$; it just needs to be a good score function, therefore r can be used not only to estimate π but also to estimate π'. Given a sample set $\mathcal{X}'_{\mathrm{U}}$ from the test density $p'(x) = \pi' p_+(x) + (1 - \pi') p_-(x)$, we propose the following estimator of the test prior π':

$$\hat{\pi}'(r) = \inf_{h \in \mathcal{H}_r} \frac{\widehat{P}'(h)}{\widehat{P}_+(h)}, \tag{32}$$

where $\widehat{P}'(h) = \widehat{\mathbb{E}}_{\mathrm{U}'}[1\{h(X) = +1\}]$. Replacing π by π', $\hat{\pi}$ by $\hat{\pi}'$, and n_{U} by $n'_{\mathrm{U}} = |\mathcal{X}'_{\mathrm{U}}|$ in Theorem 3 , we can obtain an error bound of $\hat{\pi}'$.

In Eq. (32), we require the dataset from the positive class-conditional distribution and the test-time input marginal distribution. As described in Sects. 1 and 2, we sometimes do not have access to the training data at the test-time. Fortunately in Eq. (32), we need only the value of $\widehat{P}_+(h)$ for each h, and we do not care about the samples themselves. Also, $\widehat{P}_+(h)$ takes ascending piece-wise constant values from 0 to 1 with interval $1/n_{\mathrm{P}}$. Hence, preserving the list of intervals $\{\Theta_i\}_{i=0}^{n_{\mathrm{P}}}$ such that $\widehat{P}_+(\mathrm{sign}(r(X) - \theta)) = i/n_{\mathrm{P}}$ for all $\theta \in \Theta_i$ at the training time, we can use it to estimate the test-time class-prior, without preserving the training data themselves.

3.3 Practical implementation

The entire flow of our proposed method is described in Algorithm 1. Since the strong convexity of the generator f of the Bregman divergence is desired, we may employ the quadratic function $f(t) = t^2/2$. In this case, the DRE method is called *Least-Square Importance Fitting (LSIF)* (Kanamori et al., 2009). As a parametric model r, a deep neural network may be used and optimized by stochastic gradient descent. Details of the stochastic optimization method for $\widetilde{\mathcal{L}}_f$ are described in Algorithm 2. In the prior estimation step, it is recommended to use data that is not used in the training step to avoid overfitting, especially when we are using flexible models. So we split the given data into the training and validation sets, then use the validation set to tune hyperparameters and estimate the class-priors.

Algorithm 1 DRPU

Require: Training datasets $(\mathcal{X}_{\mathrm{P}}, \mathcal{X}_{\mathrm{U}})$, test dataset $\mathcal{X}'_{\mathrm{U}}$,
Ensure: Classifier $h : \mathbb{R}^d \to \{-1, +1\}$
Split $(\mathcal{X}_{\mathrm{P}}, \mathcal{X}_{\mathrm{U}})$ into training set $(\mathcal{X}_{\mathrm{P}}^{\mathrm{tr}}\mathcal{X}_{\mathrm{U}}^{\mathrm{tr}})$ and validation set $(\mathcal{X}_{\mathrm{P}}^{\mathrm{val}}, \mathcal{X}_{\mathrm{U}}^{\mathrm{val}})$.
while no stopping criterion has been met **do**
    Optimize r with $(\mathcal{X}_{\mathrm{P}}^{\mathrm{tr}}, \mathcal{X}_{\mathrm{U}}^{\mathrm{tr}})$ by minimizing $\widetilde{\mathcal{L}}_f$.
end while
Estimate $\hat{\pi}$ from r and $(\mathcal{X}_{\mathrm{P}}^{\mathrm{val}}, \mathcal{X}_{\mathrm{U}}^{\mathrm{val}})$.
Preserve the list of intervals $\{\Theta_i\}_{i=0}^{n_{\mathrm{P}}}$.
// Obtain $\mathcal{X}'_{\mathrm{U}}$ at the test-time.
Estimate $\hat{\pi}'$ from r, $\mathcal{X}'_{\mathrm{U}}$, and $\{\Theta_i\}_{i=0}^{n_{\mathrm{P}}}$.
Determine the classification threshold $\hat{\theta}$ from $\hat{\pi}$ and $\hat{\pi}'$.
return $h = \mathrm{sign}(r - \hat{\theta})$

Algorithm 2 Stochastic optimization for non-negative Bregman divergence (Kato and Teshima, 2021)

Require: Positive and unlabeled dataset $(\mathcal{X}_{\mathrm{P}}, \mathcal{X}_{\mathrm{U}})$,
Ensure: A trained model $\hat{r} : \mathbb{R}^d \to [0, \infty)$
while no stopping criterion has been met **do**
    Create N mini-batches $B_1, \ldots, B_N$
    for $i = 1$ to N **do**
        if $\widehat{\mathbb{E}}_{\mathrm{U}}\left[\mathfrak{F}(r(X))\right] - \alpha\widehat{\mathbb{E}}_{\mathrm{P}}\left[\mathfrak{F}(r(X))\right] \geq 0$ **then**
            Set gradient: $\nabla_r \left(\widehat{\mathbb{E}}_{\mathrm{P}}\left[-f'(r(X))\right] + \widehat{\mathbb{E}}_{\mathrm{U}}\left[f^*(r(X))\right]\right)$
        else
            Set gradient: $\nabla_r \left(-\widehat{\mathbb{E}}_{\mathrm{U}}\left[\mathfrak{F}(r(X))\right] + \alpha\widehat{\mathbb{E}}_{\mathrm{P}}\left[\mathfrak{F}(r(X))\right]\right)$
        end if
        Update r
    end for
end while

4 Discussions

In this section, we provide further theoretical analysis and compare the convergence rate of the proposed method to that of unbiased/non-negative PU classification.

4.1 Selection of the Bregman generator function

Theorem 2 needs the assumption of strong convexity on the Bregman generator function f. We used a quadratic function in the proposed method; nevertheless there could be other choices of f. The following proposition shows that the tightest excess risk bound is achieved by a quadratic function.

Proposition 2 *Let f be a strongly convex function and $\mu = \inf_{t\in[0,\infty)} f''(t)$. Then the quadratic function $f_{\mathrm{S}}(t) = \mu t^2/2$ satisfies*

$$\mathrm{BR}_{f_{\mathrm{S}}}(r^* \parallel r) \leq \mathrm{BR}_f(r^* \parallel r). \tag{33}$$

Furthermore, Bregman divergence with respect to the quadratic function can be related to the excess classification risk with respect to the squared loss function. Let us denote the classification risk w.r.t. the squared loss as

$$R_{\mathrm{sq}}(g) = \mathbb{E}_{X,Y}\left[\frac{1}{4}(Yg(X) - 1)^2\right], \tag{34}$$

where $g : \mathbb{R}^d \to [-1, 1]$, and the optimal risk as

$$R^*_{\mathrm{sq}} = \inf_g R_{\mathrm{sq}}(g) = R_{\mathrm{sq}}(2\eta - 1), \tag{35}$$

where $\eta(x) = P(Y = +1 \mid X = x)$. Then, the Bregman divergence is decomposed into the excess risk w.r.t. the squared loss and a superfluous term.

Proposition 3 *Let $g_r = 2\min(\pi r, 1) - 1$ for any $r : \mathbb{R}^d \to [0, \infty)$. Then,*

$$\frac{2\pi^2}{\mu}\mathrm{BR}_{f_{\mathrm{S}}}(r^* \parallel r) = R_{\mathrm{sq}}(g_r) - R^*_{\mathrm{sq}} + \chi_r \mathfrak{D}_r, \tag{36}$$

where $\chi_r = (1/\pi^2)\mathbb{E}_{X|\pi r(X)>1}[(\pi r(X) - 1)(\pi r(X) - 2\eta(X) + 2)]$ and $\mathfrak{D}_r = P(\pi r(X) > 1)$.

If r is bounded above by $1/\pi$, the superfluous term is canceled and the Bregman divergence corresponds to the excess risk w.r.t. the squared loss, up to the scaling factor.

4.2 Excess risk bound for AUC

Here we consider the relationship between AUC optimization and DRE. It is clear that the optimal density ratio r^* is the optimal score function (Menon & Williamson, 2016), and as Theorem 1, we can obtain an excess AUC risk bound by the Bregman divergence of DRE as follows:

Theorem 4 *Let f be a μ-strongly convex function. Then, for any $r : \mathbb{R} \to [0, \infty)$, we have*

$$R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}} \leq \frac{1}{1-\pi}\sqrt{\frac{2}{\mu}\mathrm{BR}_f(r^* \parallel r)}. \tag{37}$$

Theorem 4 implies that a better estimation of the density ratio in the sense of the Bregman divergence tends to result in a better score function in the sense of AUC.

4.3 Excess risk bound with the estimated threshold

Theorem 2 gives an excess risk bound with the optimal threshold. However, in practice, we need to use an estimated threshold. Here we also consider an excess risk bound for that case. Let θ be the true classification threshold for $h = \mathrm{sign}(r - \theta)$, defined as $\theta = c_0/\pi$, and $\hat{\theta}$ be the empirical version of θ, obtained from $\hat{\pi}(r)$ and $\hat{\pi}'(r)$. Then, we have the following excess risk bound.

Theorem 5 *Let f be a μ-strongly convex function and $\theta = c_0/\pi$ where c_0 is defined in Theorem 2. Then for $\hat{\theta} \in (0, 1)$ and $h_{\hat{\theta}} = \mathrm{sign}(r - \hat{\theta})$, we have*

$$R_{\pi',c'}(h_{\hat{\theta}}) - R^*_{\pi',c'} \leq \left(C + \pi'\omega_{\hat{\theta}}\right)\sqrt{\frac{2}{\mu}\mathrm{BR}_f(r^* \parallel r)} + \pi'\left|\hat{\theta} - \theta\right|, \tag{38}$$

where C is defined in Theorem 2 and $\omega_{\hat{\theta}}$ is a constant such that $0 \leq \omega_{\hat{\theta}} \leq 1$ and $\omega_\theta = 0$.

Theorem 5 reduces to Theorem 2 when $\hat{\theta} = \theta$. We can also prove that the estimation error of the threshold decays at the linear order of the estimation error of the class-priors as follows:

Proposition 4 *Let $\hat{\pi}, \hat{\pi}'$ be estimated class-priors and $\hat{\theta}$ be an estimated threshold by $\hat{\pi}, \hat{\pi}'$. Then,*

$$\left|\hat{\theta} - \theta\right| \leq \mathcal{O}\left(|\hat{\pi} - \pi| + \left|\hat{\pi}' - \pi'\right|\right) \quad \text{as } |\hat{\pi} - \pi|, \left|\hat{\pi}' - \pi'\right| \to 0. \tag{39}$$

Combining Corollary 5 and Proposition 4, we can see that the excess risk decays at the linear order of the estimation error of the class-priors.

4.4 Convergence rate comparison to unbiased/non-negative PU classification

From the above results and theoretical analysis for non-negative Bregman divergence minimization provided by Kato and Teshima (2021), we can derive the convergence rate for our proposed method. Let $\mathcal{H}$ be a hypothesis space of density ratio model $r : \mathbb{R}^d \to [0, \infty)$ and let us denote the minimizer of the empirical risk as $\hat{r} = \mathrm{argmin}_{r \in \mathcal{H}} \widetilde{\mathcal{L}}_f(r)$ where $\mathcal{L}_f$ and $\widetilde{\mathcal{L}}_f$ are defined in Sect. 3.3. Theorem 1 in Kato and Teshima (2021) states that if f satisfies some appropriate conditions and the Rademacher complexity of $\mathcal{H}$ decays at $\mathcal{O}(1/\sqrt{n})$ w.r.t. sample size n, for example, linear-in-parameter models with a bounded norm or neural networks with a bounded Frobenius norm (Golowich et al., 2018; Lu et al., 2020), the estimation error $\mathcal{L}(\hat{r}) - \inf_{r \in \mathcal{H}} \mathcal{L}(r)$ decays at $\mathcal{O}(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{U}}})$ with high probability. Applying this to Corollary 5 and Proposition 4, the following theorem is induced.

Corollary 1 *Let f be a μ-strongly convex function and satisfy Assumption* 3 *in Kato and Teshima* (2021). *Then, for $\hat{h}_{\hat{\theta}} = \mathrm{sign}(\hat{r} - \hat{\theta})$, with probability at least $1 - \delta$, we have*

$$R_{\pi',c'}(\hat{h}_{\hat{\theta}}) - R^*_{\pi',c'} \leq A_{\mathcal{H}} + D_\delta \cdot \mathcal{O}\left(\frac{1}{n_{\mathrm{P}}^{1/4}} + \frac{1}{n_{\mathrm{U}}^{1/4}}\right) + \mathcal{O}\left(|\hat{\pi} - \pi| + |\hat{\pi}' - \pi'|\right), \quad (40)$$

where $A_{\mathcal{H}} = (C + \pi')\sqrt{\frac{2}{\mu}\left(\inf_{r \in \mathcal{H}} \mathcal{L}_f(r) - \mathcal{L}_f(r^*)\right)}$ with the constant C defined in Theorem 2 *and $D_\delta = \sqrt{\log(1/\delta)}$.*

For comparison, we consider the convergence of the excess risk based on the theoretical analysis of unbiased/non-negative PU classification provided by Kiryo et al. (2017) and the properties of *classification calibrated* loss functions provided by Bartlett et al. (2006) and Scott (2012). Let $\mathcal{G}$ be a hypothesis space of decision function $g : \mathbb{R}^d \to \mathbb{R}$ and let us denote the minimizer of the empirical risk as $\hat{g} = \inf_{g \in \mathcal{G}} \widetilde{R}_\ell(g)$ where R_ℓ and $\widetilde{R}_\ell$ are defined in Sect. 2. Assume that the loss function ℓ satisfies some appropriate conditions and if the Rademacher complexity of $\mathcal{G}$ decays at $\mathcal{O}(1/\sqrt{n})$, the estimation error $R_\ell(\hat{g}) - \inf_{g \in \mathcal{G}} R_\ell(g)$ decays at $\mathcal{O}(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{U}}})$ with high probability. In addition, if ℓ is classification calibrated (Bartlett et al., 2006; Scott, 2012), there exists a strictly increasing function ψ and the excess risk w.r.t. the zero-one loss is bounded above by the surrogate excess risk. That is, with probability at least $1 - \delta$, we have

$$R_{\pi,c}(\hat{g}) - R^*_{\pi,c} \leq \psi^{-1}\left(A_{\mathcal{G}} + D_\delta \cdot \mathcal{O}\left(\frac{1}{\sqrt{n_{\mathrm{P}}}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right)\right), \quad (41)$$

where $A_{\mathcal{G}} = \inf_{g \in \mathcal{G}} R_\ell(g) - R^*_\ell$ and $D_\delta = \sqrt{\log(1/\delta)}$.

For specific loss functions such as the hinge loss or the sigmoid loss, ψ is the identity function (Bartlett et al., 2006; Steinwart, 2007), hence the convergence rate of unbiased/non-negative PU classification would be faster than that of the density ratio estimation approach for PU classification (DRPU). This result is intuitively reasonable, because a method solving a specific problem tends to have better performance than a method solving more general problems (Vapnik, 1995). That is, the hinge loss and sigmoid loss are not *proper losses* in the context of class-posterior probability estimation (Buja et al., 2005; Reid & Williamson, 2009), and risk minimization with respect to these losses allows one to bypass the estimation of the posterior probability and obtain a classifier directly, while DRE does not.

Based on the above discussions, we should choose nnPU when we know the class-prior of the training data and it is assured that there is no class-prior shift in the test phase. In other cases, DRPU could be a better choice to solve PU classification more stably. Also, we should notice that nnPU with the sigmoid loss or the hinge loss does not provide the class-posterior probability.

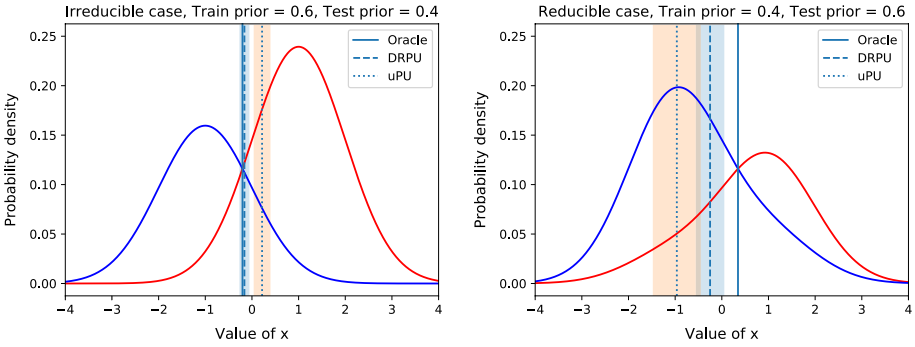

Fig. 1 Visualized classification boundaries of uPU and DRPU, averaged over 10 trials. Each of the vertical lines are the boundaries and the colored areas are the standard deviations. "Oracle" is the optimal classification boundary. The red curve means the probability density p_+ scaled by π', and the blue curve means p_- scaled by $1-\pi'$. The upper graph corresponds to the case where irreduciblity assumption holds, while the lower one does not (Color figure online)

5 Experiments

In this section, we report our experimental results. All the experiments were done with *PyTorch* (Paszke et al., 2019).[1]

5.1 Test with synthetic data

We conducted experiments with synthetic data to confirm the effectiveness of the proposed method via numerical visualization. Firstly, we define $p_+(x) = \mathcal{N}(+1, 1)$ and $p_-(x) = \mathcal{N}(-1, 1)$ where $\mathcal{N}(\mu, \sigma^2)$ denotes the univariate Gaussian distribution with mean μ and variance σ^2, and $p(x) = \pi p_+(x) + (1-\pi)p_-(x)$. We generated samples from $p_+(x)$ and $p(x)$ with $\pi = 0.4$ for training data, and from $p(x)$ with $\pi' = 0.6$ for test data.

The training dataset contained 200 positively labeled samples and 1000 unlabeled samples, and the validation dataset contained 100 positively labeled samples and 500 unlabeled samples. The test dataset consisted of 1000 samples. As a parametric model, linear-in-parameter model with Gaussian basis functions $\varphi(x) = \exp(-(x-x_i)^2/2)$, where $\{x_1, \ldots, x_{n_U}\} = \mathcal{X}_U$, was used. Adam with default momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$ and ℓ_2 regularization parameter 0.1 was used as an optimizer. Training was performed for 200 epochs with the batch size 200 and the learning rate 2×10^{-5}.

We did experiments with unbiased PU learning (uPU) (Plessis et al., 2014, 2015) with the logistic loss and our proposed method (DRPU) with LSIF. In uPU, the class-prior of the training data was estimated by KM2 (Ramaswamy et al., 2016), and for DRPU, the test unlabeled dataset was used as an unlabeled dataset to estimate the test prior π'. The left-hand side of Fig. 1 shows the obtained classification boundaries, and the boundary of DRPU was closer to the optimal one than that of uPU.

[1] We downloaded the source-codes of nnPU from https://github.com/kiryor/nnPUlearning, VPU from https://github.com/HC-Feynman/vpu, and KM2 from http://web.eecs.umich.edu/~cscott/code.html. Our implementation is available at https://github.com/csnakajima/pu-learning.

Secondly, we tested the case where the irreducibility assumption does not hold. Let $p_+(x) = 0.8\mathcal{N}(+1, 1) + 0.2\mathcal{N}(-1, 1)$, $p_-(x) = 0.2\mathcal{N}(+1, 1) + 0.8\mathcal{N}(-1, 1)$, and set the training prior $\pi = 0.6$, the test prior $\pi' = 0.4$. The result is illustrated in the right-hand side of Fig. 1. Class-prior estimation by KM2 was inaccurate since the irreducibility assumption did not hold, and then uPU led to a large error. DRPU also gave inaccurate estimations of the class-priors, but they did not affect the training step, so the influence of the estimation error was relatively mitigated.

5.2 Test with benchmark data

We also measured the performances of nnPU (Kiryo et al., 2017), PUa (Charoenphakdee & Sugiyama, 2019), VPU (Chen et al., 2020), and DRPU (the proposed method) on MNIST (Lecun et al., 1998), Fashion-MNIST (Xiao et al., 2017), Kuzushiji-MNIST (Lamb et al., 2018), and CIFAR-10 (Krizhevsky, 2012). Here we summarize the descriptions of the datasets and the training settings.

- MNIST (Lecun et al., 1998) is a gray-scale 28×28 image dataset of handwritten digits from 0 to 9, which contains 60000 training samples and 10000 test samples. Since it has 10 classes, we treated the even digits as the positive class and the odd digits as the negative class respectively. We prepared 2500 positively labeled (P) samples and 50000 unlabeled (U) samples as the training data, and 500 P samples and 10000 U samples as the validation data. The test dataset was made up of 5000 samples for each of the test distributions with different class-priors respectively. As a parametric model, 5-layer MLP : 784-300-300-300-1 with ReLU activation was used, and trained by Adam with default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and ℓ_2 regularization parameter 5×10^{-3}. Training was performed for 50 epochs with the batch size 500. The learning rate was set to 10^{-4} for nnPU/PUa and 2×10^{-5} for VPU/DRPU, which is halved for every 20 epochs. In VPU, we set hyperparameters for Mixup as $\alpha = 0.3$ and $\lambda = 2.0$. In DRPU, we set a hyperparameter for non-negative correction as $\alpha = 0.475$.
- Fashion-MNIST (Xiao et al., 2017) is a gray-scale 28×28 image dataset of 10 kinds of fashion items, which contains 60000 training samples and 10000 test samples. We treated 'Pullover', 'Dress', 'Coat', 'Sandal', 'Bag', and 'Ankle boot' as the positive class, and 'T-shirt', 'Trouser', 'Shirt', and 'Sneaker' as the negative class respectively. We prepared 2500 P samples and 50000 U samples as training data, and 500 P samples and 10000 U samples for validation data. The test dataset was made up of 5000 samples for each of the test distributions with different class-priors respectively. As a parametric model, we used LeNet (Lecun et al., 1998) -based CNN : $(1 \times 32 \times 32)$ - C(6, 5×5, pad=2) - MP(2) - C(16, 5×5, pad=2) - MP(2) - C(120, 5×5) - 120 - 84 - 1, where C(c, $h \times w$, pad=p) means c channels of $h \times w$ convolutions with zero-padding p (abbreviated if $p = 0$) followed by activation function (ReLU), and MP(k) means $k \times k$ max pooling. Batch normalization was applied after the first fully-connected layer. The model was trained by Adam, with the same settings as the case of MNIST. Training was performed for 100 epochs with the batch size 500 and the learning rate 2×10^{-5}, which is halved for every 20 epochs. In VPU, we set hyperparameters for Mixup as $\alpha = 0.3$ and $\lambda = 0.5$. In DRPU, we set a hyperparameter for non-negative correction as $\alpha = 0.6$.
- Kuzushiji-MNIST (Lamb et al., 2018) is a gray-scale 28×28 image dataset of 10 kinds of cursive Japanese characters, which contains 60000 training samples and

10000 test samples. We treated 'o', 'ki', 're', 'wo' as the positive class, and 'su', 'tsu', 'na', 'ha', 'ma', 'ya' as the negative class respectively. We prepared 2500 P samples and 50000 U samples as the training data, and 500 P samples and 10000 U samples as the validation data. The test dataset was made up of 5000 samples for each of the test distributions with different class-priors respectively. The model and optimization settings were the same as the cases of Fashion-MNIST. In VPU, we set hyperparameters for Mixup as $\alpha = 0.3$ and $\lambda = 0.5$. In DRPU, we set a hyperparameter for non-negative correction as $\alpha = 0.375$.

- CIFAR-10 (Krizhevsky, 2012) is a colored 32×32 image dataset, which contains 50000 training samples and 10000 test samples. We treated 'airplane', 'automobile', 'ship', and 'truck' as the positive class, and 'bird', 'cat', 'deer', 'dog', 'frog', and 'horse' as the negative class respectively. We prepared 2500 P samples and 45000 U samples as the training data, and 500 P samples and 5000 U samples as the validation data. The test dataset was made up of 5000 samples for each the test distributions with different class-priors respectively. As a parametric model, we used the CNN introduced in Springenberg et al. (2015) : $(3 \times 32 \times 32)$ - C(96, 5×5, pad=2) - MP(2×2) - C(96, 5×5, pad=2) - MP(2×2) - C(192, 5×5, pad=2) - C(192, 5×3, pad=1) - C(192, 1×1) - C(10, 1×1) with ReLU activation. Batch normalization was applied after the max pooling layers and the third, fourth, fifth convolution layers. The model was trained by Adam, with the same settings as the case of MNIST. Training was performed for 100 epochs with the batch size 500 and the learning rate 10^{-5}, which is halved for every 20 epochs. In VPU, we set hyperparameters for Mixup as $\alpha = 0.3$ and $\lambda = 4.0$. In DRPU, we set a hyperparameter for non-negative correction as $\alpha = 0.425$.

In nnPU, the class-prior of the training data was estimated by KM2 (Ramaswamy et al., 2016). In PUa, we estimated both the training and test priors by KM2, then performed cost-sensitive non-negative PU classification (Charoenphakdee & Sugiyama, 2019). Note that in this setting, PUa needs the unlabeled test dataset at the training-time to estimate the test prior by KM2, while DRPU needs it at only the test-time. Moreover, PUa needs to train a model for each time the test prior changes. In nnPU and PUa, the sigmoid loss was used as a loss function.

Table 2 shows the results of the experiments. nnPU and PUa unintentionally achieved high accuracy in some cases because of estimation errors of the class-priors, while they had poor results in the other cases. VPU achieved good results in several cases where the scale of class-prior shift was small since it does not need the class-prior in the training phase, but it was not adapted to large class-prior shift. DRPU outperformed the other methods in almost all cases, and was the most robust to the test-time class-prior shift. Figure 2 gives the classification errors in the experiments. For example, in the case of Fashion-MNIST with $\pi' = 0.8$, nnPU and PUa suffered from overfitting due to the estimation error of the training class-prior. Also, as seen in the case of Kuzushiji-MNIST with $\pi' = 0.2$, DRPU gave a better result than the other methods, and was the most stable, i.e., it had the smallest variance.

In addition, Table 3 reports the computed AUC values on the experiments for each of the methods. The results were picked from $\pi' = 0.6$ case. DRPU had a bit worse results than VPU on MNIST and Fashion-MNIST, while it performed well on Kuzushiji-MNIST and CIFAR-10. Table 4 summarizes the absolute error of the class-prior estimation by KM2 and our method described in Sect. 3.2. For KM2, we used 2000 positive samples

Table 2 The means and standard deviations of the classification accuracy in percent on benchmark datasets over 10 trials. "Train" and "Test" denote the class-priors of the training and test data respectively. "Avg" is the averaged accuracy of the four results with different priors. The best results with respect to the one-sided t-test at the significance level 0.05 are highlighted in boldface (for the "Avg" case, just picking the highest one)

Dataset	Train	Test	nnPU	PUa	VPU	DRPU
MNIST	0.5	0.2	92.98 ± 1.72	93.11 ± 1.51	**96.21 ± 0.20**	95.78 ± 0.54
		0.4	93.76 ± 1.02	93.70 ± 0.94	**94.93 ± 0.51**	**94.85 ± 0.58**
		0.6	**94.52 ± 0.31**	**94.79 ± 0.48**	93.71 ± 0.89	**94.67 ± 0.46**
		0.8	95.23 ± 0.76	95.28 ± 0.90	92.43 ± 1.39	**95.91 ± 0.48**
		Avg.	94.12	94.22	94.32	**95.30**
Fashion-MNIST	0.6	0.2	**91.67 ± 0.97**	**91.05 ± 0.61**	89.22 ± 1.13	**91.59 ± 0.57**
		0.4	88.95 ± 1.22	87.89 ± 1.10	**90.11 ± 0.51**	**90.42 ± 0.92**
		0.6	86.26 ± 1.50	86.07 ± 1.35	**90.80 ± 0.48**	**90.46 ± 0.72**
		0.8	83.26 ± 2.29	84.39 ± 2.27	91.70 ± 0.98	**92.97 ± 0.58**
		Avg.	87.54	87.35	90.46	**91.36**
Kuzushiji-MNIST	0.4	0.2	81.88 ± 2.52	82.81 ± 2.99	85.78 ± 2.81	**91.41 ± 0.56**
		0.4	85.18 ± 1.64	85.11 ± 2.12	85.93 ± 1.77	**88.62 ± 0.52**
		0.6	**88.35 ± 0.95**	**87.87 ± 0.94**	86.11 ± 1.36	**88.22 ± 0.67**
		0.8	**91.28 ± 0.53**	**90.67 ± 1.73**	86.10 ± 2.09	**91.32 ± 0.62**
		Avg.	86.67	86.62	85.98	**89.89**
CIFAR-10	0.4	0.2	80.32 ± 1.55	80.33 ± 2.34	**91.94 ± 0.67**	**91.59 ± 0.37**
		0.4	84.38 ± 1.15	84.03 ± 1.59	**89.00 ± 1.21**	**88.67 ± 0.56**
		0.6	88.16 ± 0.54	88.05 ± 0.95	85.99 ± 2.45	**89.22 ± 0.56**
		0.8	**92.28 ± 0.51**	**92.21 ± 0.41**	82.92 ± 3.92	**92.45 ± 0.35**
		Avg.	86.28	86.16	87.46	**90.48**

from the training dataset and 2000 unlabeled samples from the test dataset. The inputs were transformed into 50 dimensions by PCA (Jolliffe & Cadima, 2016). For our method, we used 500 positive samples from the validation dataset and 5000 unlabeled samples from the test dataset. It is observed that our class-prior estimation method outperformed KM2 in almost all cases.

5.3 Comparisons under different numbers of labeled samples

To numerically verify the theoretical insight provided in Sect. 4.4, we compared nnPU and DRPU with different sizes of the positively labeled dataset. In this experiment, we assumed that the true class-prior π was known and no class-prior shift would occur. We performed nnPU and DRPU on MNIST and CIFAR-10, with $n_{\mathrm{P}} \in \{500, 1000, 2000, 4000\}$. Note that we skipped the class-prior estimation step of DRPU because the class-priors were given. Figure 3 shows the results of the experiments. On MNIST, the performance of DRPU was comparable to that of nnPU when n_{P} was small, yet it got outperformed under larger n_{P}. On CIFAR-10, unlike the MNIST

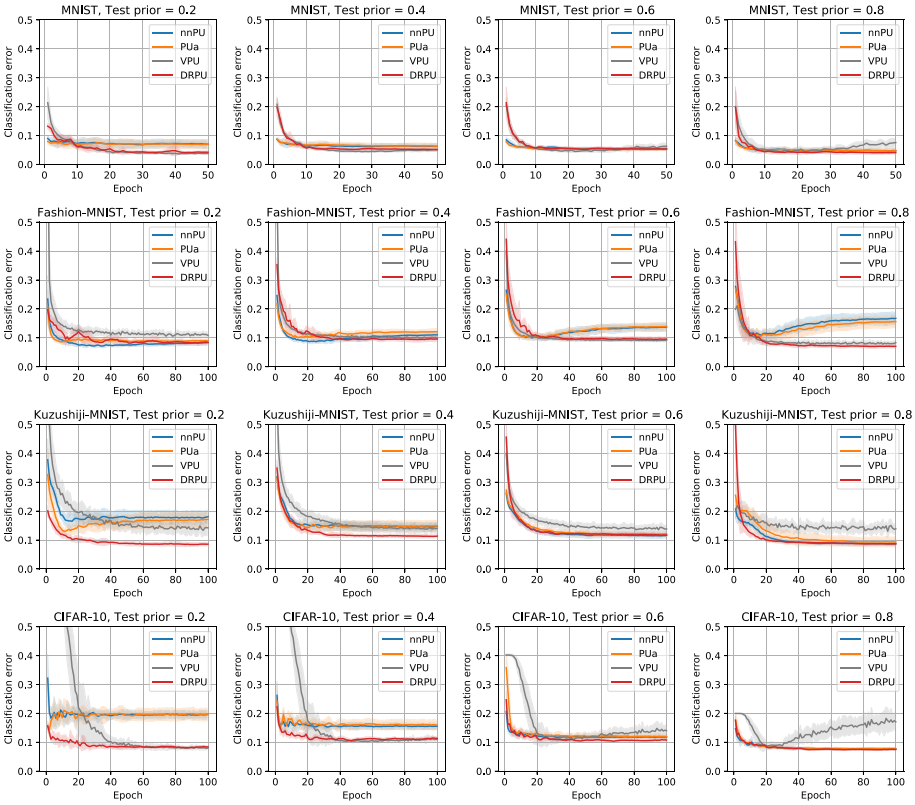

Fig. 2 The means and standard deviations of the classification error as functions of the training epoch

Table 3 The means and standard deviations of the AUC on benchmark datasets over 10 trials. The best results with respect to the one-sided t-test at the significance level 0.05 are highlighted in boldface

Dataset	nnPU	PUa	VPU	DRPU
MNIST	0.9855 ± 0.0022	0.9861 ± 0.0024	**0.9902 ±0.0010**	0.9815 ±0.0023
F-MNIST	0.9371 ±0.0129	0.9359 ±0.0065	**0.9656 ±0.0032**	0.9607 ±0.0055
K-MNIST	**0.9505 ±0.0050**	**0.9485 ±0.0044**	0.9347 ±0.0129	**0.9507 ±0.0033**
CIFAR-10	0.9509 ±0.0062	0.9512 ±0.0071	**0.9594 ±0.0024**	**0.9577 ±0.0028**

case, the difference in the classification error was larger when n_{P} was smaller. As a whole, nnPU stably outperformed DRPU, and this experimental result supports the theoretical discussion in Sect. 4.4.

Table 4 The means and standard deviations of the absolute error of the class-prior estimation on benchmark datasets over 10 trials. The best results with respect to one-sided *t*-test at the significance level 0.05 are highlighted in boldface

Dataset	Prior	KM2	Ours
MNIST	0.2	0.0679 ± 0.0147	$\mathbf{0.0101 \pm 0.0128}$
	0.4	0.0410 ± 0.0169	$\mathbf{0.0279 \pm 0.0265}$
	0.6	$\mathbf{0.0209 \pm 0.0110}$	$\mathbf{0.0361 \pm 0.0372}$
	0.8	0.2885 ± 0.0560	$\mathbf{0.0485 \pm 0.0429}$
F-MNIST	0.2	$\mathbf{0.0133 \pm 0.0105}$	0.0230 ± 0.0120
	0.4	$\mathbf{0.0170 \pm 0.0151}$	$\mathbf{0.0181 \pm 0.0115}$
	0.6	0.0843 ± 0.0297	$\mathbf{0.0183 \pm 0.0148}$
	0.8	0.2534 ± 0.0340	$\mathbf{0.0249 \pm 0.0231}$
K-MNIST	0.2	0.0326 ± 0.0155	$\mathbf{0.0220 \pm 0.0130}$
	0.4	0.1029 ± 0.0201	$\mathbf{0.0704 \pm 0.0211}$
	0.6	0.2685 ± 0.0246	$\mathbf{0.1088 \pm 0.0310}$
	0.8	0.4869 ± 0.0361	$\mathbf{0.1496 \pm 0.0394}$
CIFAR-10	0.2	0.1880 ± 0.0097	$\mathbf{0.0151 \pm 0.0112}$
	0.4	0.1399 ± 0.0163	$\mathbf{0.0189 \pm 0.0142}$
	0.6	0.0738 ± 0.0167	$\mathbf{0.0305 \pm 0.0184}$
	0.8	0.1242 ± 0.0784	$\mathbf{0.0397 \pm 0.0322}$

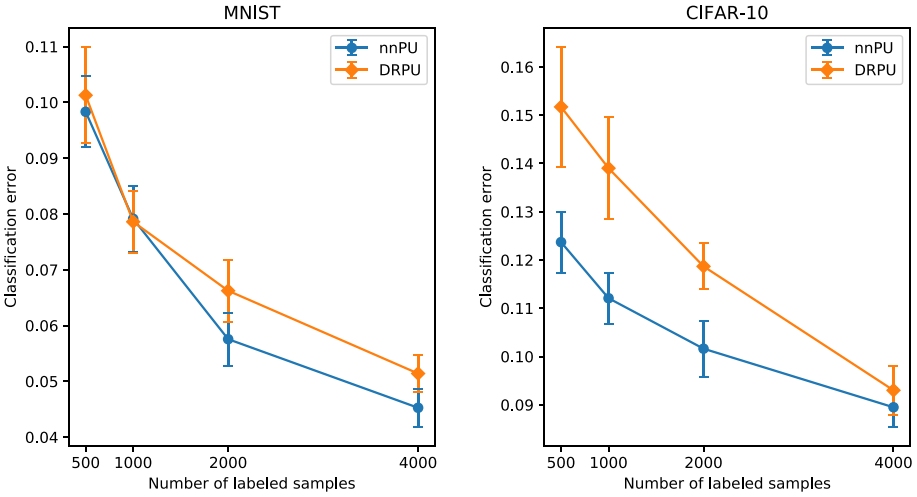

Fig. 3 Classification errors of nnPU and DRPU on MNIST and CIFAR-10, averaged over 10 trials for each settings of the number of labeled samples. The vertical bars at each of the points refer the standard deviations

6 Conclusions

In this paper, we investigated positive-unlabeled (PU) classification from a perspective of density ratio estimation, and proposed a novel PU classification method based on density ratio estimation. The proposed method does not require the class-priors in the training phase, and it can cope with class-prior shift in the test phase. We provided theoretical analysis for the proposed method, and demonstrated its effectiveness in the experiments. Extending our work to other weakly-supervised learning problems (Lu et al., 2019; Bao et al., 2018) or multi-class classification settings (Xu et al., 2017) is a promising future work.

Appendix A proofs

A.1 Proof of Theorem 1

From Lemma 1 of Scott (2012), we have

$$R_{\pi,c}(g) - R^*_{\pi,c} = \mathbb{E}_X\big[1\{\mathrm{sign}(g(X)) \neq \mathrm{sign}(\eta(X) - c)\}|\eta(X) - c|\big],$$

where $\eta(x) = p(Y = +1 \mid X = x)$. Then, for $h_\theta = \mathrm{sign}(r - \theta) = \mathrm{sign}(\pi r - c)$,

$$\begin{aligned}
R_{\pi,c}(h_\theta) - R^*_{\pi,c} &= \mathbb{E}_X[1\{(\pi r(X) - c)(\eta(X) - c) < 0\}|\eta(X) - c|] \\
&= \mathbb{E}_X[1\{\pi r(X) < c < \eta(X)\}|\eta(X) - c|] \\
&\quad + \mathbb{E}_X[1\{\eta(X) < c < \pi r(X)\}|\eta(X) - c|] \\
&\leq \mathbb{E}_X[1\{\pi r(X) < \eta(X)\}|\eta(X) - \pi r(X)|] \\
&\quad + \mathbb{E}_X[1\{\eta(X) < \pi r(X)\}|\eta(X) - \pi r(X)|] \\
&= \mathbb{E}_X[|\eta(X) - \pi r(X)|] \\
&= \pi\mathbb{E}_X\big[|r^*(X) - r(X)|\big] \\
&\leq \pi\sqrt{\mathbb{E}_X\big[(r^*(X) - r(X))^2\big]} \\
&= \pi\sqrt{\frac{2}{\mu}\mathbb{E}_X\Big[\frac{\mu}{2}(r^*(X) - r(X))^2\Big]} \\
&\leq \pi\sqrt{\frac{2}{\mu}\mathbb{E}_X\big[f(r^*(X)) - f(r(X)) - f'(r(X))(r^*(X) - r(X))\big]} \\
&= \pi\sqrt{\frac{2}{\mu}\mathrm{BR}_f(r^* \parallel r)},
\end{aligned}$$

where the second inequality is Jensen's, and the third inequality comes from the definition of strong convexity. □

A.2 Proof of Theorem 2

Same as Theorem 1 of Charoenphakdee and Sugiyama (2019), we normalize coefficients of $R_{\pi,c}$ and $R_{\pi',c'}$ and determine c to satisfy

$$\frac{R_{\pi,c}(g)}{(1-c)\pi + c(1-\pi)} = \frac{R_{\pi',c'}(g)}{(1-c')\pi' + c'(1-\pi')}.$$

Compare the coefficient of the $\mathbb{E}_{\mathrm{P}}[\cdot]$,

$$\frac{(1-c)\pi}{(1-c)\pi + c(1-\pi)} = \frac{(1-c')\pi'}{(1-c')\pi' + c'(1-\pi')}.$$

Solve this equation with respect to c and denote it as c_0,

$$c_0 = \frac{c'\pi(1-\pi')}{(1-c')(1-\pi)\pi' + c'\pi(1-\pi')}.$$

Therefore, we obtain

$$\begin{aligned} R_{\pi',c'}(h_{c_0/\pi}) - R^*_{\pi',c'} &= \frac{(1-c')\pi' + c'(1-\pi')}{(1-c_0)\pi + c_0(1-\pi)}\left(R_{\pi,c_0}(h_{c_0/\pi}) - R^*_{\pi,c_0}\right) \\ &\leq \pi\frac{c' + \pi' - 2c'\pi'}{c_0 + \pi - 2c_0\pi}\sqrt{\frac{2}{\mu}\mathrm{BR}_f(r^* \parallel r)}, \end{aligned}$$

where $h_{c_0/\pi} = \mathrm{sign}(r - c_0/\pi)$ and the inequality is from Theorem 1. □

A.3 Proof of Theorem 3

We separate $|\hat{\pi} - \pi|$ as follows

$$|\hat{\pi} - \pi| \leq \left|\hat{\pi} - \inf_{h\in\mathcal{H}_r}\frac{P(h)}{P_+(h)}\right| + \left|\inf_{h\in\mathcal{H}_r}\frac{P(h)}{P_+(h)} - \pi\right| \tag{42}$$

The first term of Eq. (42) is upper bounded by the uniform bound

$$\begin{aligned} \left|\hat{\pi} - \inf_{h\in\mathcal{H}_r}\frac{P(h)}{P_+(h)}\right| &= \left|\inf_{h\in\mathcal{H}_r}\frac{\widehat{P}(h)}{\widehat{P}_+(h)} - \inf_{h\in\mathcal{H}_r}\frac{P(h)}{P_+(h)}\right| \\ &\leq \max\left(\frac{\widehat{P}(h_1)}{\widehat{P}_+(h_1)} - \frac{P(h_1)}{P_+(h_1)}, \frac{P(h_2)}{P_+(h_2)} - \frac{\widehat{P}(h_2)}{\widehat{P}_+(h_2)}\right) \\ &\leq \sup_{h\in\mathcal{H}_r}\left|\frac{\widehat{P}(h)}{\widehat{P}_+(h)} - \frac{P(h)}{P_+(h)}\right| \end{aligned}$$

where $h_1 = \mathrm{argmin}_{h\in\mathcal{H}_r}\frac{P(h)}{P_+(h)}$, $h_2 = \mathrm{argmin}_{h\in\mathcal{H}_r}\frac{\widehat{P}(h)}{\widehat{P}_+(h)}$.

From McDiarmid's inequality, with probability at least $1-\delta$, we have

$$\begin{aligned}\sup_{h\in\mathcal{H}_r}\left|\widehat{P}(h)-P(h)\right| &\le \mathfrak{R}^p_{n_{\mathrm{U}}}(\mathcal{H}_r)+\sqrt{\frac{\log\frac{2}{\delta}}{2n_{\mathrm{U}}}}\\ &\le \sqrt{\frac{4\log(en_{\mathrm{U}}/2)}{n_{\mathrm{U}}}}+\sqrt{\frac{\log\frac{2}{\delta}}{2n_{\mathrm{U}}}}\\ &= \varepsilon(n_{\mathrm{U}},\delta)\end{aligned}$$

for any $0<\delta<1$. We used $\mathfrak{R}^p_n(\mathcal{H})\le\sqrt{\frac{2d\log(en/d)}{n}}$ where d is VC-dimension of $\mathcal{H}$, and $\mathrm{VCdim}(\mathcal{H}_r)\le 2$ for a fixed r. The same discussion holds for $\widehat{P}_+(h)$, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{H}_r}\left|\widehat{P}_+(h)-P_+(h)\right|\le\varepsilon(n_{\mathrm{P}},\delta)$$

Thus, with probability at least $(1-\delta_{\mathrm{P}})(1-\delta_{\mathrm{U}})$, we have

$$\frac{\widehat{P}(h)-\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)+\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}\le\frac{P(h)}{P_+(h)}\le\frac{\widehat{P}(h)+\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)-\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}$$

For the left hand side,

$$\begin{aligned}\frac{\widehat{P}(h)-\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)+\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})} &= \left(\frac{\widehat{P}(h)}{\widehat{P}_+(h)}-\frac{\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)}\right)\sum_{i=0}^{\infty}\left(-\frac{\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}{\widehat{P}_+(h)}\right)^i\\ &= \frac{\widehat{P}(h)}{\widehat{P}_+(h)}-\frac{\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)}\\ &\quad-\left(\frac{\widehat{P}(h)}{\widehat{P}_+(h)}-\frac{\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)}\right)\frac{\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}{\widehat{P}_+(h)}\sum_{i=1}^{\infty}\left(-\frac{\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}{\widehat{P}_+(h)}\right)^{i-1}\\ &\ge \frac{\widehat{P}(h)}{\widehat{P}_+(h)}-\mathcal{O}(\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}}))-\mathcal{O}\big(\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})\big).\end{aligned}$$

Note that $\frac{\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}{\widehat{P}_+(h)}<\gamma<1$ by the assumption on $\mathcal{H}_r$. And for the right hand side,

$$\begin{aligned}\frac{\widehat{P}(h)+\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)-\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})} &= \left(\frac{\widehat{P}(h)}{\widehat{P}_+(h)}+\frac{\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)}\right)\sum_{i=0}^{\infty}\left(\frac{\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}{\widehat{P}_+(h)}\right)^i\\ &= \frac{\widehat{P}(h)}{\widehat{P}_+(h)}+\frac{\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)}\\ &\quad+\left(\frac{\widehat{P}(h)}{\widehat{P}_+(h)}+\frac{\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}})}{\widehat{P}_+(h)}\right)\frac{\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}{\widehat{P}_+(h)}\sum_{i=1}^{\infty}\left(\frac{\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})}{\widehat{P}_+(h)}\right)^{i-1}\\ &\le \frac{\widehat{P}(h)}{\widehat{P}_+(h)}+\mathcal{O}(\varepsilon(n_{\mathrm{U}},\delta_{\mathrm{U}}))+\mathcal{O}\big(\varepsilon(n_{\mathrm{P}},\delta_{\mathrm{P}})\big).\end{aligned}$$

Therefore, assign $\delta_{\mathrm{P}}=1/n_{\mathrm{P}}$, $\delta_{\mathrm{U}}=1/n_{\mathrm{U}}$ to obtain

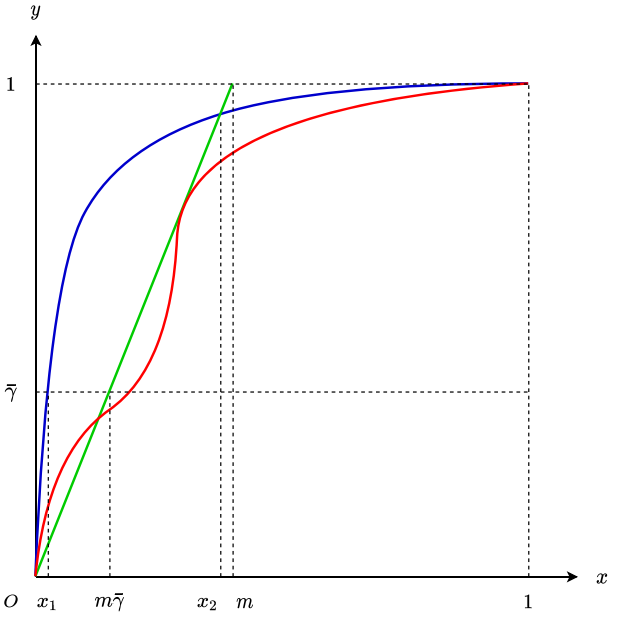

Fig. 4 Green: $y = \frac{1}{m}x$ / Red: ROC curve of r / Blue: ROC curve of r^*

$$\sup_{h\in\mathcal{H}_r}\left|\frac{\widehat{P}(h)}{\widehat{P}_+(h)}-\frac{P(h)}{P_+(h)}\right|\leq\mathcal{O}\left(\sqrt{\frac{\log n_{\mathrm{P}}}{n_{\mathrm{P}}}}+\sqrt{\frac{\log n_{\mathrm{U}}}{n_{\mathrm{U}}}}\right).$$

Next, we consider the second term of Eq. (42). At first we show that the term is bounded above by $\frac{2(1-\pi)}{1-\bar{\gamma}^2}R_{\mathrm{AUC}}(r)$. Utilize $P(h)=\pi P_+(h)+(1-\pi)P_-(h)$ to obtain

$$\left|\inf_{h\in\mathcal{H}_r}\frac{P(h)}{P_+(h)}-\pi\right|=(1-\pi)\left|\inf_{h\in\mathcal{H}_r}\frac{P_-(h)}{P_+(h)}\right|.$$

Let $m=\inf_{h\in\mathcal{H}_r}\frac{P_-(h)}{P_+(h)}$, and consider a ROC curve where P_- (False Positive Rate) plotted as x-axis and P_+ (True Positive Rate) plotted as y-axis. As seen in Fig. 4, the trapezoid surrounded by $x=0$, $y=1$, $y=\bar{\gamma}$, and $y=x/m$ is in the area over the ROC curve of r, since $\sup_{h\in\mathcal{H}_r}\frac{P_+(h)}{P_-(h)}=\frac{1}{m}$. Thus

$$\frac{(m\bar{\gamma}+m)(1-\bar{\gamma})}{2}\leq 1-\mathrm{AUC}(r)=R_{\mathrm{AUC}}(r)$$
$$\frac{m(1-\bar{\gamma}^2)}{2}\leq R_{\mathrm{AUC}}(r).$$

Then,

$$\inf_{h\in\mathcal{H}_r}\frac{P_-(h)}{P_+(h)}=(1-\pi)m\leq\frac{2(1-\pi)}{1-\bar{\gamma}^2}R_{\mathrm{AUC}}(r).$$

Secondly, we prove that there exists an increasing function ξ such that $m \le \xi\left(R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}}\right)$. For a fixed m, the ROC curve of r is always under $y = x/m$ in $m\bar{\gamma} \le x \le 1$, and the optimal ROC curve always dominates all other ROC curves (Menon & Williamson, 2016). Therefore, at least, the area surrounded by $y = \bar{\gamma}$, $y = x/m$, and the optimal ROC curve is assured. That is,

$$\begin{aligned} R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}} &= \mathrm{AUC}(r^*) - \mathrm{AUC}(r) \\ &\ge \int_{x_1}^{m\bar{\gamma}} (\rho(x) - \bar{\gamma})\mathrm{d}x + \int_{m\bar{\gamma}}^{x_2} \left(\rho(x) - \frac{1}{m}x\right)\mathrm{d}x, \end{aligned}$$

where $\rho(x)$ means the optimal ROC curve and $\rho(x_1) = \bar{\gamma}$, $\rho(x_2) = x_2/m$. Denote the right-hand side of the above inequality as $U(m)$, then

$$\begin{aligned} \frac{\partial U}{\partial m}(m) &= \rho(m\bar{\gamma})\bar{\gamma} - \bar{\gamma}^2 + \rho(x_2)x_2' - \rho(m\bar{\gamma})\bar{\gamma} - \frac{1}{m}x_2x_2' + \frac{x_2^2}{2m^2} + \frac{\bar{\gamma}^2}{2} \\ &= \frac{x_2^2}{2m^2} + \frac{\bar{\gamma}^2}{2} > 0. \end{aligned}$$

Note that x_1 is independent of m. This shows that $U(m)$ is a strictly increasing function of $m \in [x_1/\bar{\gamma}, 1]$, therefore, there exists an increasing function $U^{-1} : [0, 1] \to [x_1/\bar{\gamma}, 1]$ and

$$\begin{aligned} U(m) &\le R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}} \\ m &\le U^{-1}\left(R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}}\right), \end{aligned}$$

then we can define an increasing function

$$\begin{aligned} (1-\pi)m &\le \xi(R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}}) \\ &= \min\left(\frac{2(1-\pi)}{1-\bar{\gamma}^2}R_{\mathrm{AUC}}(r), (1-\pi)U^{-1}\left(R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}}\right)\right). \end{aligned}$$

Thirdly, we check that $\xi(0) \to 0$ when $\bar{\gamma} \to 0$. Since the optimal ROC curve is concave (Menon & Williamson, 2016), we have $U(x_1/\bar{\gamma}) = 0$. And from the irreducibility assumption, we have $\frac{x_1}{\bar{\gamma}} \to 0$ when $\bar{\gamma} \to 0$. Therefore, $U^{-1}(0) = \frac{x_1}{\bar{\gamma}} \to 0$. and we concludes the proof. □

A.4 Proof of Proposition 2

From Reid and Williamson (2009), we have the integral representation of the Bregman divergence:

$$\mathrm{BR}_f(r^* \parallel r) = \mathbb{E}_X\left[\int_{r(X)}^{r^*(X)} (r^*(X) - t)f''(t)\mathrm{d}t\right].$$

Then,

$$\begin{aligned}\mathrm{BR}_f(r^* \parallel r) &= \mathbb{E}_X\left[\int_r^{r^*} (r^* - t) f''(t)\mathrm{d}t\right] \\ &\geq \mathbb{E}_X\left[\int_r^{r^*} (r^* - t)\mu\mathrm{d}t\right] \\ &= \mathrm{BR}_{f_\mathrm{S}}(r^* \parallel r).\end{aligned}$$

We used $f''(t) \geq \inf_{t\in[0,\infty)} f''(t) = \mu$ at the second line and $f_\mathrm{S}''(t) = \mu$ at the third line.

□

A.5 Proof of Proposition 3

As mentioned in the Proof of Proposition 2, we have the integral representation of the Bregman divergence. Then,

$$\begin{aligned}\mathrm{BR}_{f_\mathrm{S}}(r^* \parallel r) &= \mathbb{E}_X\left[\int_{r(X)}^{r^*(X)} (r^*(X) - t) f_\mathrm{S}''(t)\mathrm{d}t\right] \\ &= \mu\mathbb{E}_X\left[\int_{r^*}^{r} (t - r^*)\mathrm{d}t\right] \\ &= \mu\mathbb{E}_X\left[\int_{r^*}^{\min(r,1/\pi)} (t - r^*)\mathrm{d}t + 1\{\pi r > 1\}\int_{1/\pi}^{r} (t - r^*)\mathrm{d}t\right].\end{aligned}$$

The first term in the expectation can be written as:

$$\int_{r^*}^{\min(r,1/\pi)} (t - r^*)\mathrm{d}t = \frac{1}{\pi^2}\int_{\eta}^{\min(\pi r,1)} (t - \eta)\mathrm{d}t,$$

where $\eta(x) = \pi r^*(x) = P(Y = +1 \mid X = x)$. On the other hand, the classification risk w.r.t the squared loss can be written as

$$\begin{aligned}R_{\mathrm{sq}}(g) &= \mathbb{E}_{X,Y}\left[\frac{1}{4}(Yg(X) - 1)^2\right] \\ &= \mathbb{E}_X\left[C_\eta(g)\right]\end{aligned}$$

where $g \in [-1, 1]$ and $C_\eta(g)$ is the conditional risk

$$C_\eta(g) = \eta\frac{(g-1)^2}{4} + (1-\eta)\frac{(g+1)^2}{4},$$

and the optimal classification risk w.r.t the squared loss as

$$C_\eta^* = \inf_g C_\eta(g) = C_\eta(2\eta - 1).$$

Then, we have

$$\begin{aligned} C_\eta(g) - C_\eta^* &= \eta(\min(\pi r, 1) - 1)^2 + (1-\eta)(\min(\pi r, 1))^2 - \eta(1-\eta) \\ &= (\min(\pi r, 1) - \eta)^2 \\ &= 2\int_\eta^{\min(\pi r, 1)} (t - \eta)\mathrm{d}t. \end{aligned}$$

Thus,

$$\begin{aligned} R_{\mathrm{sq}}(g) - R_{\mathrm{sq}}^* &= \mathbb{E}_X\left[C_\eta(g) - C_\eta^*\right] \\ &= 2\mathbb{E}_X\left[\int_\eta^{\min(\pi r, 1)} (t - \eta)\mathrm{d}t.\right] \end{aligned}$$

Next, the second term in the expectation is

$$1\{\pi r > 1\}\int_{1/\pi}^{r} (t - r^*)\mathrm{d}t = 1\{\pi r > 1\} \cdot \frac{1}{2}(\pi r - 1)(\pi r - 2\eta + 2),$$

then,

$$\begin{aligned} &\mathbb{E}_X\left[1\{\pi r > 1\}\int_{1/\pi}^{r} (t - r^*)\mathrm{d}t\right] \\ &= \frac{1}{2}\mathbb{E}_{X|\pi r(X)>1}[(\pi r - 1)(\pi r - 2\eta + 2)]P(\pi r(X) > 1). \end{aligned}$$

Finally, we have

$$\frac{2\pi^2}{\mu}\mathrm{BR}_{f_{\mathrm{s}}}(r^* \parallel r) = R_{\mathrm{sq}}(g) - R_{\mathrm{sq}}^* + \chi_r\mathfrak{D}_r.$$

□

A.6 Proof of Theorem 4

From the result of Clémençon et al. (2006), we have

$$\begin{aligned} &R_{\mathrm{AUC}}(s) - R_{\mathrm{AUC}}^* \\ &= \frac{1}{2\pi(1-\pi)}\mathbb{E}_{X,X'}\left[\left|\eta(X) - \eta(X')\right|1\{(s(X) - s(X')(\eta(X) - \eta(X')) \le 0\}\right], \end{aligned}$$

where $\eta(x) = p(Y = +1 \mid X = x)$. Then for any r,

$$
\begin{aligned}
&R_{\mathrm{AUC}}(r) - R^*_{\mathrm{AUC}} \\
&= \frac{1}{2\pi(1-\pi)} \mathbb{E}_{X,X'} \left[\left| \eta(X) - \eta(X') \right| 1\{(r(X) - r(X'))(\eta(X) - \eta(X')) \le 0\} \right] \\
&= \frac{1}{2\pi(1-\pi)} \mathbb{E}_{X,X'} \left[\left| \eta(X) - \eta(X') \right| 1\{(r(X) - r(X'))(\eta(X) - \eta(X')) \le 0\} \right] \\
&= \frac{1}{2(1-\pi)} \mathbb{E}_{X,X'} \Big[\left| r^*(X) - r^*(X') \right| \Big(1\{r(X) < r(X')\} 1\{r^*(X) \ge r^*(X')\} \\
&\quad + 1\{r(X) \ge r(X')\} 1\{r^*(X) < r^*(X')\} \Big) \Big] \\
&= \frac{1}{2(1-\pi)} \mathbb{E}_{X,X'} \Big[(r^*(X) - r^*(X')) 1\{r(X) < r(X')\} \\
&\quad + (r^*(X') - r^*(X)) 1\{r(X) \ge r(X')\} \Big] \\
&\le \frac{1}{2(1-\pi)} \mathbb{E}_{X,X'} \left[\left(r^*(X) - r(X) + r(X') - r^*(X') \right) 1\{r(X) < r(X')\} \right. \\
&\quad \left. + \left(r^*(X') - r(X') + r(X) - r^*(X) \right) 1\{r(X) \ge r(X')\} \right] \\
&\le \frac{1}{2(1-\pi)} \mathbb{E}_{X,X'} \left[|r^*(X) - r(X)| + \left| r^*(X') - r(X') \right| \right] \\
&= \frac{1}{1-\pi} \mathbb{E}_X \left[|r^*(X) - r(X)| \right] \\
&\le \frac{1}{1-\pi} \sqrt{\frac{2}{\mu} \mathrm{BR}_f(r^* \parallel r)}.
\end{aligned}
$$

Used the same technique as the proof of Theorem 1 for the last inequality, and this concludes the proof. □

A.7 Proof of Theorem 5

We utilize the following equality (see the proof of Theorem 2.)

$$
R_{\pi',c'}(h_{\hat{\theta}}) - R^*_{\pi',c'} = \left(R_{\pi',c'}(h_{\hat{\theta}}) - R_{\pi',c'}(h_\theta) \right) + \left(R_{\pi',c'}(h_\theta) - R^*_{\pi',c'} \right).
$$

where $h_\theta = \mathrm{sign}(r - \theta)$. The second term $R_{\pi',c'}(h_\theta) - R^*_{\pi',c'}$ is processed by Theorem 2, so we consider the first term. By the definition of cost-sensitive classification risk,

$$
\begin{aligned}
&R_{\pi',c'}(h_{\hat{\theta}}) - R_{\pi',c'}(h_\theta) \\
&\quad = \mathbb{E}\Big[(1-c')1\{Y=+1\}\big(1\{\mathrm{sign}(r(X)-\hat{\theta})=-1\} - 1\{\mathrm{sign}(r(X)-\theta)=-1\}\big) \\
&\qquad + c'1\{Y=-1\}\big(1\{\mathrm{sign}(r(X)-\hat{\theta})=+1\} - 1\{\mathrm{sign}(r(X)-\theta)=+1\}\big)\Big] \\
&\quad = \mathbb{E}_X\Big[(1-c')\eta(X)\big(1\{\theta < r(X) < \hat{\theta}\} - 1\{\hat{\theta} < r(X) < \theta\}\big) \\
&\qquad + c'(1-\eta(X))\big(1\{\hat{\theta} < r(X) < \theta\} - 1\{\theta < r(X) < \hat{\theta}\}\big)\Big] \\
&\quad = \mathbb{E}_X\big[(\eta(X)-c')\big(1\{\theta < r(X) < \hat{\theta}\} - 1\{\hat{\theta} < r(X) < \theta\}\big)\big] \\
&\quad = \pi'\mathbb{E}_X\big[(r^*(X)-\theta)\big(1\{\theta < r(X) < \hat{\theta}\} - 1\{\hat{\theta} < r(X) < \theta\}\big)\big] \\
&\quad = \pi'\mathbb{E}_X\big[(r^*(X)-r(X)+r(X)-\theta)\big(1\{\theta < r(X) < \hat{\theta}\} - 1\{\hat{\theta} < r(X) < \theta\}\big)\big] \\
&\quad \le \pi'\omega_{\hat{\theta}}\mathbb{E}_X\big[|r^*(X)-r(X)|\big] + \pi'\Big|\hat{\theta}-\theta\Big| \\
&\quad \le \pi'\omega_{\hat{\theta}}\sqrt{\frac{2}{\mu}\mathrm{BR}_f(r^* \parallel r)} + \pi'\Big|\hat{\theta}-\theta\Big|.
\end{aligned}
$$

The first inequality holds because $\mathbb{E}\Big[\Big|1\{\theta < r(X) < \hat{\theta}\} - 1\{\hat{\theta} < r(X) < \theta\}\Big|\Big] \le 1$ and $\mathbb{E}[1\{\theta < r(X) < \theta\}] = 0$. We used the same technique as the proof of Theorem 1 for the second inequality, and this concludes the proof. □

A.8 Proof of Proposition 4

We first consider the bound for $\Big|\hat{\theta}-\theta\Big|$. Let $\hat{\pi} = \pi + d$, $\hat{\pi}' = \pi' + d'$ where $-\pi < d < 1-\pi$ and $-\pi' < d' < 1-\pi'$. Then,

$$
\begin{aligned}
\hat{\theta} &= \frac{c'(1-\hat{\pi}')}{(1-c')(1-\hat{\pi})\hat{\pi}' + c'\hat{\pi}(1-\hat{\pi}')} \\
&= \frac{c'(1-\pi') - c'd'}{(1-c')(1-\pi)\pi' + c'\pi(1-\pi') + (c'-\pi')d + (1-c'-\pi)d' - dd'}.
\end{aligned}
$$

We denote $A = c'(1-\pi')$, $B(d') = -c'd'$, $C = (1-c')(1-\pi)\pi' + c'\pi(1-\pi')$, $D(d,d') = (c'-\pi')d + (1-c'-\pi)d' - dd'$. Using these notations, we have

$$
\Big|\hat{\theta}-\theta\Big| = \Big|\frac{A+B}{C+D} - \frac{A}{C}\Big|.
$$

Thus, if $D \ge 0$,

$$\begin{aligned}\left|\hat{\theta}-\theta\right| &= \left|\frac{A+B}{C+D}-\frac{A}{C}\right| \\ &= \left|\frac{A+B}{C+D}-\frac{A}{C+D}\frac{C+D}{C}\right| \\ &= \left|\frac{A}{C+D}+\frac{B}{C+D}-\frac{A}{C+D}\left(1+\frac{D}{C}\right)\right| \\ &= \left|\frac{1}{C+D}\left(B-\frac{AD}{C}\right)\right| \\ &\leq \frac{1}{C}\left|B-\frac{AD}{C}\right| \\ &\leq \mathcal{O}\left(|d|+\left|d'\right|\right) \quad (\text{as } d, d' \to 0),\end{aligned}$$

where the first inequality holds because $C > 0$ and $D > 0$. And if $D < 0$, we have $C > D$ from $\hat{\theta} \geq 0$. Then,

$$\begin{aligned}\left|\hat{\theta}-\theta\right| &= \left|\frac{A+B}{C+D}-\frac{A}{C}\right| \\ &= \left|\frac{A+B}{C}\sum_{i=0}^{\infty}\left(-\frac{D}{C}\right)^{i}-\frac{A}{C}\right| \\ &= \left|\frac{A+B}{C}+\frac{A+B}{C}\sum_{i=1}^{\infty}\left(-\frac{D}{C}\right)^{i}-\frac{A}{C}\right| \\ &= \frac{1}{C}\left|B+(A+B)\sum_{i=1}^{\infty}\left(-\frac{D}{C}\right)^{i}\right| \\ &\leq \mathcal{O}\left(|d|+\left|d'\right|\right) \quad (\text{as } d, d' \to 0).\end{aligned}$$

and we complete the proof.

A.9 Proof of Corollary 1

It immediately holds from Theorem 1 of Kato and Teshima (2021) and Theorem 5. □

Acknowledgements MS was supported by KAKENHI 20H04206.

Author Contributions All authors contributed to the study conception and design. Theoretical analysis and experimental setup were performed by Shota Nakajima. The first draft of the manuscript was written by Shota Nakajima, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Masashi Sugiyama was supported by KAKENHI 20H04206.

Availability of data and material Only public datasets and frameworks were used.

Declarations

Conflict of interest Not Applicable.

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication Not Applicable.

Code availability Our code is available at https://github.com/csnakajima/pu-learning.

References

Arora, S., Babai, L., Stern, J., et al. (1997). The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences, 54*(2), 317–331.

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

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

Blanchard, G., Lee, G., & Scott, C. (2010). Semi-supervised novelty detection. *Journal of Machine Learning Research, 11*(99), 2973–3009.

Buja, A., Stuetzle, W., & Shen, Y. (2005). *Loss functions for binary class probability estimation and classification: Structure and applications*. Philadelphia: University of Pennsylvania.

Charoenphakdee, N., Sugiyama, M. (2019). Positive-unlabeled classification under class prior shift and asymmetric error. In: SDM.

Chen, H., Liu, F., Wang, Y., et al (2020). A variational approach for learning from positive and unlabeled data. In: NeurIPS.

Clémençon, S., Lugosi, G., & Vayatis, N. (2006). Ranking and empirical minimization of u-statistics. *The Annals of Statistics, 36*, 844–874.

Coletto, M., Lucchese, C., Orlando, S., et al (2015). Electoral predictions with twitter: A machine-learning approach. In: CEUR Workshop Proceedings 1404.

Elkan, C. (2001). The foundations of cost-sensitive learning. In: IJCAI.

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

Golowich, N., Rakhlin, A., Shamir, O. (2018). Size-independent sample complexity of neural networks. In: COLT.

Hou, M., Chaib-Draa, B., Li, C., et al. (2018). Generative adversarial positive-unlabeled learning. In: IJCAI.

Hsieh, C.J., Natarajan, N., Dhillon, I.S. (2015). Pu learning for matrix completion. In: ICML.

Jolliffe, I., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(20150), 202.

Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research, 10*(48), 1391–1445.

Kato, M., Teshima, T. (2021). Non-negative bregman divergence minimization for deep direct density ratio estimation. In: ICML.

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

Kiryo, R., Niu, G., du Plessis, M.C., et al. (2017). Positive-unlabeled learning with non-negative risk estimator. In: NeurIPS.

Krizhevsky, A. (2012). *Learning multiple layers of features from tiny images*. Tech. Rep.: University of Toronto.

Lamb, A., Kitamoto, A., Ha, D., et al. (2018). Deep learning for classical japanese literature. arXiv:1812.01718.

Lecun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*, 2278–2324.

Letouzey, F., Denis, F., Gilleron, R. (2000). Learning from positive and unlabeled examples. In: ALT.

Li, W., Guo, Q., & Elkan, C. (2011). A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geoscience and Remote Sensing, 49*(2), 717–725.

Li, X., Liu, B. (2003). Learning to classify texts using positive and unlabeled data. In: IJCAI.

Li, X., Yu, P., Liu, B., et al. (2009). Positive unlabeled learning for data stream classiflcation. In: SDM.

Liu, B., Yu, P., Li, X. (2003). Partially supervised classification of text documents. In: ICML.

Lu, N., Niu, G., Menon, AK., et al. (2019). On the minimal supervision for training any binary classifier from only unlabeled data. In: ICLR.

Lu, N., Zhang, T., Niu, G., et al. (2020). Mitigating overfitting in supervised classification from two unlabeled datasets: A consistent risk correction approach. In: ICAIS.
Martínez, A., Schmuck, C., Pereverzyev, S., et al. (2018). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research, 281*(3), 588–96.
McMahan, H.B., Holt, G., Sculley, D., et al. (2013). Ad click prediction: a view from the trenches. In: KDD.
Menon, A.K., Ong, C.S. (2016). Linking losses for density ratio and class-probability estimation. In: ICML.
Menon, A. K., & Williamson, R. C. (2016). Bipartite ranking: A risk-theoretic perspective. *Journal of Machine Learning Research, 17*(195), 1–102.
Nguyen, M.N., Li, X.L., Ng, S.K. (2011). Positive unlabeled learning for time series classification. In: IJCAI.
Niu, G., du Plessis, M.C., Sakai, T., et al. (2016). Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In: NeurIPS.
Paszke, A., Gross, S., Massa, F., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS.
Plessis, M.C.d., Niu, G., Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. In: NeurIPS.
Plessis, M.C.d., Niu, G., Sugiyama, M. (2015). Convex formulation for learning from positive and unlabeled data. In: ICML.
Plessis, M.C.d., Niu, G., Sugiyama, M. (2016). Class-prior estimation for learning from positive and unlabeled data. In: ACML.
Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., et al. (2009). *Dataset Shift in Machine Learning*. United States: The MIT Press.
Ramaswamy, H., Scott, C., Tewari, A. (2016). Mixture proportion estimation via kernel embeddings of distributions. In: ICML.
Reid, M.D., Williamson, R.C. (2009) Surrogate regret bounds for proper losses. In: ICML.
Scott, C. (2012). Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics, 6*, 958–992.
Scott, C. (2015). A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In: AISTATS.
Scott, C., Blanchard, G., Handy, G. (2013). Classification with asymmetric label noise: Consistency and maximal denoising. In: COLT.
Springenberg, J., Dosovitskiy, A., Brox, T., et al. (2015). Striving for simplicity: The all convolutional net. In: ICLR.
Steinwart, I. (2007). How to compare different loss functions and their risks. *Constructive Approximation, 26*, 225–287.
Sugiyama, M., Suzuki, T., Nakajima, S., et al. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics, 60*, 699–746.
Sugiyama, M., Kanamori, T., Suzuki, T., et al. (2009). A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications, 1*, 183–208.
Sugiyama, M., Suzuki, T., & Kanamori, T. (2011). Density ratio matching under the bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics, 64*(5), 1009–44.
Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge: Cambridge University Press.
Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer.
Xiao, H., Rasul, K., Vollgraf, R. (2017), Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747.
Xu, Y., Xu, C., Xu, C., et al. (2017). Multi-positive and unlabeled learning. In: IJCAI.
Zhang, C., Ren, D., Liu, T., et al. (2019). Positive and unlabeled learning with label disambiguation. In: IJCAI.
Zhang, H., Cisse, M., Dauphin, Y.N., et al. (2018). mixup: Beyond empirical risk minimization. In: ICLR.