Classification with Noisy Labels by Importance Reweighting

Tongliang Liu and Dacheng Tao, *Fellow, IEEE*

Abstract—In this paper, we study a classification problem in which sample labels are randomly corrupted. In this scenario, there is an unobservable sample with noise-free labels. However, before being observed, the true labels are independently flipped with a probability $\rho \in [0, 0.5)$, and the random label noise can be class-conditional. Here, we address two fundamental problems raised by this scenario. The first is how to best use the abundant surrogate loss functions designed for the traditional classification problem when there is label noise. We prove that any surrogate loss function can be used for classification with noisy labels by using importance reweighting, with consistency assurance that the label noise does not ultimately hinder the search for the optimal classifier of the noise-free sample. The other is the open problem of how to obtain the noise rate ρ. We show that the rate is upper bounded by the conditional probability $P(\hat{Y}|X)$ of the noisy sample. Consequently, the rate can be estimated, because the upper bound can be easily reached in classification problems. Experimental results on synthetic and real datasets confirm the efficiency of our methods.

Index Terms—Classification, label noise, noise rate estimation, consistency, importance reweighting.

1 Introduction

Classification crucially relies on the accuracy of the dataset labels. In some situations, observation labels are easily corrupted and, therefore, inaccurate. Designing learning algorithms that account for noisy labeled data is therefore of great practical importance and has attracted a significant amount of interest in the machine learning community.

The random classification noise (RCN), in which each label is flipped independently with a probability $\rho \in [0, 0.5)$, has been proposed; it was proven to be PAC-learnable by Angluin and Laird [1] soon after the noise-free PAC learning model was introduced by Valiant [2]. Many related works then followed: Kearns [3] proposed the statistical query model to learn with RCN. The restriction he enforced is that learning is based not on the particular properties of individual random examples, but instead on the global statistical properties of large samples. Such an approach to learning seems intuitively more robust. Lawrence and Scholköpf [4] proposed a Bayesian model for this noise and applied it to sky moving in images. Biggio et al. [5] enabled support vector machine learning with RCN via a kernel matrix correction. And Yang et al. [6] developed multiple kernel learning for classification with noisy labels using stochastic programming. The interested reader is referred to further examples in the survey [7]. However, most of these algorithms are designed for specific surrogate loss functions, and the use and benefit of the large number of surrogate loss functions designed for the traditional (noise-free) classification problem is important to investigate in order to solve classification problems in the presence of label noise.

Aslam and Decatur [8] proved that the RCN exploited using a 0-1 loss function is PAC-learnable if the function class is of finite VC-dimension. Manwani and Sastry [9] analyzed the tolerance properties of RCN for risk minimization under several frequently used surrogate loss functions and showed that many of them do not tolerate RCN. Natarajan et al. [10] reported two methods for learning asymmetric RCN models, in which the random label noise is class-conditional. Their methods exploit many different surrogate loss functions [11]: the first model uses unbiased estimators of surrogate loss functions for empirical risk minimization, but the unbiased estimator may be non-convex, even if the original surrogate loss function is convex; their second method uses label-dependent costs. The latter approach is based on the idea that there exists an $\alpha \in (0, 1)$ such that the minimizer of the expected risk as assessed using the α-weighted 0-1 loss function $\ell_\alpha(t, y) = (1 - \alpha)1_{y=1}1_{t\leq 0} + \alpha 1_{y=-1}1_{t>0}$ over the noisy sample distribution, where t is the predicted value and y is the label of the example, has the same sign as that of the Bayes classifier which minimizes the expected risk as assessed using the 0-1 loss function over the clean sample distribution; see, for example, Theorem 9 in [10]. The method is notable because it can be applied to all the convex and classification-calibrated surrogate loss functions (If a surrogate loss function is classification-calibrated and the sample size is sufficiently large, the surrogate loss function will help learn the same optimal classifier as the 0-1 loss function does, see Theorem 1 in Bartlett et al. [11]). This modification is based on the asymmetric classification-calibrated results [12] and cannot be used to improve the performance of symmetric RCN problems or the algorithms that employ the non-classification-calibrated surrogate loss functions.

To best use and benefit from the abundant surrogate loss functions designed for the traditional classification prob-

• *T. Liu and D. Tao are with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 81 Broadway Street, Broadway, NSW 2007, Australia.*
E-mail: tliang.liu@gmail.com, dacheng.tao@uts.edu.au.

lems, here we propose an importance reweighting method in which any surrogate loss function designed for a traditional classification problem can be used for classification with noisy labels. In our method, the weights are non-negative, so the convexity of objective functions does not change. In addition, our method inherits most batch learning optimization procedures designed for traditional classification problems with different regularizations; see, for examples, [13]–[17].

Although many works have focused on the RCN model, how to best estimate the noise rate ρ remains an open problem [6] and severely limits the practical application of the existing algorithms. Most previous works make the assumption that the noise rate is known or learn it using cross-validation, which is time-consuming and lacks a guarantee of generalization accuracy. In this paper, we set the noise rate to be asymmetric and unknown and denote the flip probability of positive labels $P(\hat{Y} = -1|Y = +1)$ and the flip probability of negative labels $P(\hat{Y} = +1|Y = -1)$ by ρ_{+1} and ρ_{-1}, respectively. We show that the noise rate ρ_{+1} (or ρ_{-1}) is upper bounded by the conditional probability $P(-1|X)$ (or $P(+1|X)$) of the noisy data. Moreover, the upper bound can be reached if there exists an $x \in \mathcal{X}$ such that the probability $P(+1|x)$ (or $P(-1|x)$) of the "clean" sample is zero, which is very likely to hold for classification problems. The noise rates $\rho_{\pm 1}$ are therefore estimated by finding the minimal $P(\mp 1|X)$ of the noisy training sample.

1.1 Related Works

Kearns and Li [18] introduced the malicious noise (MN) model, in which an adversary can access the sample and randomly replace a fraction of them with adversarial ones. It has been proven that any nontrivial target function class cannot be PAC learned with accuracy ϵ and malicious noise rate $\eta \geq \epsilon(1 + \epsilon)$; see, for examples, [18]–[20]. Long and Servedio [21] proved that an algorithm for learning γ-margin half-spaces that minimizes a convex surrogate loss function for misclassification risk cannot tolerate malicious noise at a rate greater than $\mathcal{O}(\epsilon\gamma)$. They therefore proposed an algorithm, that does not optimize a convex loss function and that can tolerate a higher rate of malicious noise than order $\mathcal{O}(\epsilon\gamma)$. Further details about the MN model can be found in [22].

Cesa-Bianchi et al. [23] considered a more complicated model in which the features and labels are both added with zero-mean and variance-bounded noise. They used unbiased estimates of the gradient of the surrogate loss function to learn from the noisy sample in an online learning setting. Perceptron algorithms that tolerate RCN have also been widely studied; see, for examples, [24]–[27]. See Khardon and Wachman [28] for a survey of noise-tolerant variants of perceptron algorithms.

As well as these model-motivated algorithms, many algorithms that exploit robust surrogate loss functions have been designed for learning with any kind of feature and label noise. Robust surrogate loss functions, such as the Cauchy loss function [29] and correntropy (also known as the Welsch loss function), [30], [31], have been empirically proven to be robust to noise. Some other algorithms, such as confidence weighted learning [32], have also been proposed for noise-tolerant learning.

To the best of our knowledge, the only work that related to learn the unknown noise rate was proposed by Scott et al. [33]. Inspired by the theory of mixture proposition estimation [34], they provided estimators for the inversed noise rates $\pi_{+1} = P(Y = -1|\hat{Y} = +1)$ and $\pi_{-1} = P(Y = +1|\hat{Y} = -1)$. However, there were no efficient algorithms that can be used to calculate the estimators until Scott [35] proposed an efficient algorithm for optimizing them during the preparation of this manuscript. By using Bayes' rule, we have $P(\hat{Y}|Y) = P(Y|\hat{Y})P(\hat{Y})/P(Y)$. However, our method for estimation the noise rates is essentially different from that of Scott et al. [33] because $P(Y)$ is unknown. The inversed noise rates can be used to design algorithms for classification with label noise; see, for example, [35]. In this paper, we also design importance reweighting algorithms for classification with label noise by employing the inversed noise rates.

The rest of this paper is organized as follows. The problem is set up in Section 2. Section 3 presents some useful results applied to the traditional classification problem. In Section 4, we discuss how to perform classification in the presence of RCN and benefit from the abundant surrogate loss functions and algorithms designed for the traditional classification problem. In Section 5, we discuss how to reduce the uncertainty introduced by RCN by estimating the conditional probability $P(\hat{Y}|X)$ of the noisy sample; theoretical guarantees for the consistency of the learned classifiers are provided; certain convergence rates are also characterized in this section. In Section 6, an approach for estimating the noise rates is proposed. We also provide a detailed comparison between the theory of noise rate estimation and that of the inversed noise rate estimation in this section. We present the proofs of our assertions in Section 7. In Section 8, we present experimental results on synthetic and benchmark datasets, before concluding in Section 9.

2 Problem Setup

Let D be the distribution of a pair of random variables $(X, Y) \in \mathcal{X} \times \{\pm 1\}$, where $\mathcal{X} \subseteq \mathbb{R}^m$. Our goal is to predict a label for any given observation $X \in \mathcal{X}$ using a sample drawn i.i.d. from the distribution D. However, in many real-world classification problems, sample labels are randomly corrupted. We therefore consider the asymmetric RCN model (see [10]). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an i.i.d. sample drawn from the distribution D and $(X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n)$ the corresponding corrupted ones. The asymmetric RCN model is given by:

$$P(\hat{Y} = +1|Y = -1) = \rho_{-1}, P(\hat{Y} = -1|Y = +1) = \rho_{+1},$$

where $\rho_{+1}, \rho_{-1} \in [0, 1)$ and $\rho_{+1} + \rho_{-1} < 1$.

We denote by D_ρ the distribution of the corrupted variables $(X, \hat{Y})$. In our setting, the "clean" sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the noise rates ρ_{+1} and ρ_{-1} are not available for learning algorithms. The classifier and noise rates are learned only by using the knowledge from the corrupted sample $(X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n)$.

3 The Traditional Classification Pr-oblem

Classification is a fundamental machine learning problem. One intuitive way to learn the classifier is to find a decision function $f \in F$, such that the expected risk $R_{1,D}(f) = E_{(X,Y)\sim D}[1_{\text{sign}(f(X))\neq Y}]$ is minimized, where F is the function class for searching. However, two problems remain when minimizing the expected risk: first, that the 0-1 loss function is neither convex nor smooth, and second that the distribution D is unknown. The solutions to these two problems are summarized below.

For the problem that the 0-1 loss function is neither convex nor smooth, abundant convex surrogate loss functions (most are smooth) with the classification-calibrated property [11], [12] have been proposed. These surrogate loss functions, such as square loss, logistic loss, and hinge loss, have been proven useful in many real-world applications. Apart from the convex classification-calibrated surrogate loss functions, many other non-convex surrogate loss functions empirically proven to be robust to noise, such as Cauchy loss and Welsch loss, are also frequently used. In this paper, we show that all these surrogate loss functions, as well as the non-classification-calibrated surrogate loss functions, such as the asymmetric exponential loss function (see Example 8 in [11])

$$\ell(t,y) = \begin{cases} \exp(-2ty) & t \leq 0 \\ \exp(-ty) & t > 0, \end{cases}$$

can be used directly for classification in the presence of RCN by employing the importance reweighting method.

For the problem that distribution D is unknown, empirical risk is proposed to approximate the expected risk. The empirical risk is defined as

$$\hat{R}_{\ell,D}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i),$$

where the corresponding expected risk is

$$R_{\ell,D}(f) = R[D,f,\ell] = E_{(X,Y)\sim D}[\hat{R}_{\ell,D}(f)]$$

and ℓ denotes any surrogate loss function. The classifier is then learned by empirical risk minimization (ERM) [36]:

$$f_n = \arg\min_{f\in F} \hat{R}_{\ell,D}(f).$$

The consistency of $R_{\ell,D}(f_n)$ to $\min_{f\in F} R_{\ell,D}(f)$ is therefore essential for designing surrogate loss functions and learning algorithms. Let

$$f^* = \arg\min_{f\in F} R_{\ell,D}(f).$$

It [37] is easily proven that

$$R_{\ell,D}(f_n) - R_{\ell,D}(f^*) \leq 2\sup_{f\in F} |R_{\ell,D}(f) - \hat{R}_{\ell,D}(f)|.$$

The right hand side term is known as the generalization error, and the consistency is guaranteed by convergence of the generalization error. We note that learning algorithms which are based on ERM, such as those using Tikhonov or manifold regularization, will not have a slower convergence rate of consistency than that of ERM. In this paper, we therefore provide consistency guarantees for learning algorithms dealing with RCN by deriving the generalization error bounds of the corresponding ERM algorithms.

4 Learning with Importance Reweigh-ting

Importance reweighting is widely used for domain adaptation [38], but here we introduce it to classification in the presence of label noise. One observation [39] from the field of importance reweighting is as follows:

$$\begin{aligned} R_{\ell,D}(f) &= R[D,f,\ell] = E_{(X,Y)\sim D}[\ell(f(X),Y)] \\ &= E_{(X,Y)\sim D_\rho}\left[\frac{P_D(X,Y)}{P_{D_\rho}(X,Y)}\ell(f(X),Y)\right] \\ &= R\left[D_\rho, f, \frac{P_D(X,Y)}{P_{D_\rho}(X,Y)}\ell(f(X),Y)\right] \\ &= R\left[D_\rho, f, \beta(X,Y)\ell(f(X),Y)\right] \\ &= R_{\beta\ell,D_\rho}(f), \end{aligned}$$

where $\beta(X,Y) = \frac{P_D(X,Y)}{P_{D_\rho}(X,Y)}$.

For the problem of classification in the presence of label noise, note that $P_D(X) = P_{D_\rho}(X)$. We therefore have

$$\begin{aligned} \beta(X,Y) &= \frac{P_D(X,Y)}{P_{D_\rho}(X,Y)} = \frac{P_D(Y|X)P_D(X)}{P_{D_\rho}(Y|X)P_{D_\rho}(X)} \\ &= \frac{P_D(Y|X)}{P_{D_\rho}(Y|X)}. \end{aligned}$$

Thus, even though the labels are corrupted, classification can still be implemented if only the weight $\beta(X,\hat{Y}) = P_D(\hat{Y}|X)/P_{D_\rho}(\hat{Y}|X)$ could be accessed to the loss $\ell(f(X),\hat{Y})$.

Lemma 1. The asymmetric RCN problem can be addressed by reweighting the surrogate loss functions of the traditional classification problem via importance reweighting. The weight given to a noisy example $(X,Y) \sim D_\rho$ is

$$\begin{aligned} \beta(X,Y) &= \frac{P_D(Y|X)}{P_{D_\rho}(Y|X)} \\ &= \frac{P_{D_\rho}(Y|X) - \rho_{-Y}}{(1-\rho_{+1}-\rho_{-1})P_{D_\rho}(Y|X)}. \end{aligned}$$

The weight $\beta(X,Y)$ is non-negative if $P_{D_\rho}(Y|X) \neq 0$. If $P_{D_\rho}(Y|X) = 0$, we intuitively let $\beta(X,Y) = 0$.

A classifier can therefore be learned for the "clean" data in the presence of asymmetric RCN by minimizing the following reweighted empirical risk:

$$\begin{aligned} \hat{f}_n &= \arg\min_{f\in F} \hat{R}_{\beta\ell,D_\rho} \\ &= \arg\min_{f\in F} \frac{1}{n}\sum_{i=1}^{n} \beta(X_i,\hat{Y}_i)\ell(f(X_i),\hat{Y}_i), \end{aligned}$$

where

$$\beta(X_i,\hat{Y}_i) = \frac{P_{D_\rho}(\hat{Y}_i|X_i) - \rho_{-\hat{Y}_i}}{(1-\rho_{+1}-\rho_{-1})P_{D_\rho}(\hat{Y}_i|X_i)}.$$

By the following proposition, based on Talagrand's Lemma (see, e.g., Lemma 4.2 in [40]), we show that, given $P_{D_\rho}(\hat{Y}|X)$, the above weighted empirical risk will converge to the unweighted expected risk of the "clean" data for any $f \in F$. So, $R_{\ell,D}$ can be approximated by $\hat{R}_{\beta\ell,D_\rho}$.

Proposition 1. Given the conditional probability $P_{D_\rho}(\hat{Y}|X)$ and the noise rates ρ_{+1} and ρ_{-1}. Let $\beta(X,\hat{Y})\ell(f(X),\hat{Y})$ be upper bounded by b. Then, for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\begin{aligned}&\sup_{f\in F}|R_{\ell,D}(f) - \hat{R}_{\beta\ell,D_\rho}(f)|\\ &= \sup_{f\in F}\left|E_{(X,\hat{Y})\sim D_\rho}\left[\hat{R}_{\beta\ell,D_\rho}(f)\right] - \hat{R}_{\beta\ell,D_\rho}(f)\right|\\ &\le \frac{1-U}{1-\rho_{-1}-\rho_{+1}}\mathfrak{R}(\ell\circ F) + b\sqrt{\frac{\log(1/\delta)}{2n}},\end{aligned}$$

where $U = \min_{(X,\hat{Y})}\frac{\rho_{-\hat{Y}}}{P_{D_\rho}(\hat{Y}|X)}$, and the Rademacher complexity $\mathfrak{R}(\ell\circ F)$ [41] is defined by

$$\mathfrak{R}(\ell\circ F) = E_{(X,\hat{Y})\sim D_\rho,\sigma}\left[\sup_{f\in F}\frac{2}{n}\sum_{i=1}^{n}\sigma_i\ell(f(X_i),\hat{Y}_i)\right]$$

and $\sigma_1,\ldots,\sigma_n$ are i.i.d. Rademacher variables.

The Rademacher complexity has a convergence rate of order $\mathcal{O}(\sqrt{1/n})$ [41]. If the function class has proper conditions on its variance, the Rademacher complexity will quickly converge and is of order $\mathcal{O}(1/n)$; see, for example, [42]. The generalization bound in Proposition 1 is derived using the Rademacher complexity method. Many other hypothesis complexities and methods can also be employed to derive the generalization bound.

Since

$$\begin{aligned}&R_{\ell,D}(f_n) - R_{\ell,D}(f^*)\\ &= R_{\beta\ell,D_\rho}(f_n) - R_{\beta\ell,D_\rho}(f^*)\\ &\le 2\sup_{f\in F}|R_{\beta\ell,D_\rho}(f) - \hat{R}_{\beta\ell,D_\rho}(f)|,\end{aligned}$$

the consistency rate will therefore be inherited for learning with label noise, provided that the conditional probability $P_{D_\rho}(\hat{Y}|X)$ and noise rates $\rho_{\pm1}$ are accurately estimated.

Based on Proposition 1, we can now state our first main result for classification in the presence of label noise using our framework of importance reweighting.

Theorem 1. Any surrogate loss functions designed for the traditional classification problem can be used for classification in the presence of asymmetric RCN by employing the importance reweighting method. The consistency rate for classification with asymmetric RCN will be the same as that of the corresponding traditional classification algorithm, provided that the conditional probability $P_{D_\rho}(\hat{Y}|X)$ and noise rates $\rho_{\pm1}$ are accurately estimated.

The trade-off for using and benefitting from the abundant surrogate loss functions designed for traditional classification problems is the need to estimate the distribution $P_{D_\rho}(\hat{Y}|X)$ and noise rates $\rho_{\pm1}$. Next, we address how to estimate the distribution and the noise rates separately.

5 Estimating $P_{D_\rho}(\hat{Y}|X)$

We have shown that the uncertainty introduced by classification label noise can be reduced by the knowledge of weight

$$\beta(X,\hat{Y}) = \frac{P_D(X,\hat{Y})}{P_{D_\rho}(X,\hat{Y})} = \frac{P_D(\hat{Y}|X)}{P_{D_\rho}(\hat{Y}|X)}.$$

In the asymmetric RCN problem,

$$\beta(X,\hat{Y}) = \frac{P_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1-\rho_{+1}-\rho_{-1})P_{D_\rho}(\hat{Y}|X)},$$

and therefore the weight can be learned by using the noisy sample and the noise rates. In this section, we present three methods to estimate the conditional probability $P_{D_\rho}(\hat{Y}|X)$ with consistency analyses; how to estimate the noise rates is discussed in the next section.

5.1 The Probabilistic Classification Method

The conditional probability $P_{D_\rho}(\hat{Y}|X)$ can be estimated by a simple probabilistic classification method, where the corresponding link function maps the outputs of the learned predictor to the interval $[0,1]$ and thus can be interpreted as probabilities. However, such a method is parametric, which has a strong assumption that the target conditional distribution is of the form of the link function used. For example, if the logistic loss function is employed, the learned distribution will be the form of

$$P_{D_\rho}(\hat{Y}|X,f) = \frac{1}{1+\exp\left(-\hat{Y}f(X)\right)}.$$

When the logistic regression is correctly specified, i.e., there exists $f^* \in F$ such that $P_{D_\rho}(\hat{Y}|X,f^*)$ is equal to the target conditional distribution $P^*_{D_\rho}(\hat{Y}|X)$, the logistic regression is optimal in the sense that the approximation error is minimized (being zero). However, when the model is misspecified, which would be the case in practice, a large approximation error may be introduced even if the hypothesis class F is chosen to be relatively large, which will hinder the statistical consistency for learning the target weight function $\beta^*(X,\hat{Y})$.

Remark 1. We found that employing the probabilistic classification method to estimate the conditional probability $P_{D_\rho}(\hat{Y}|X)$ did not perform well. Its empirical validation is therefore omitted in this paper.

5.2 The Kernel Density Estimation Method

In this subsection, we introduce the kernel density estimation method to estimate the conditional probability $P_{D_\rho}(\hat{Y}|X)$, which has the consistency property for learning the target weight function $\beta^*(X,\hat{Y})$.

Using Bayes' rule, we have

$$P_{D_\rho}(\hat{Y}|X) = \frac{P_{D_\rho}(X|\hat{Y})P_{D_\rho}(\hat{Y})}{P_{D_\rho}(X)}. \tag{1}$$

When the dimensionality of $\mathcal{X}$ is low and the sample size is sufficiently large, the probabilities $P_{D_\rho}(x|y), P_{D_\rho}(y)$ and $P_{D_\rho}(x)$ can be easily and efficiently estimated using the noisy sample.

If we use

$$\hat{P}_{D_\rho}(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}1_{\hat{Y}_i=\hat{Y}} \tag{2}$$

and the kernel density estimation method

$$\hat{P}_{D_\rho}(X) = \frac{1}{n}\sum_{i=1}^{n}K(X,X_i) \tag{3}$$

to estimate $P_{D_\rho}(\hat{Y})$ and $P_{D_\rho}(X)$, respectively (where $K(X, X_i) = k(X)k(X_i)$ is a universal kernel, see [43]), the consistency of classification with label noise (or learning the target weight function $\beta^*(X, \hat{Y})$) is guaranteed by the following theorem.

Theorem 2. Let $\hat{P}_{D_\rho}(\hat{Y}|X)$ be an estimator for $P_{D_\rho}(\hat{Y}|X)$ using equations (1), (2) and (3), and

$$\hat{\beta}(X, \hat{Y}) = \frac{\hat{P}_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1 - \rho_{-1} - \rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)}.$$

Let

$$\hat{f}_{n,\hat{\beta}} = \min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}(X_i, \hat{Y}_i)\ell(f(k(X_i)), \hat{Y}_i)$$

and

$$f^* = \min_{f \in F} R[D, f, \ell(f(k(X)), Y)].$$

For any $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(R[D, \hat{f}_{n,\hat{\beta}}, \ell(\hat{f}_{n,\hat{\beta}}(k(X)), Y)] - R[D, f^*, \ell(f^*(k(X)), Y)] > \epsilon) = 0.$$

When $P_{D_\rho}(X|\hat{Y})$ and $P_{D_\rho}(X)$ are estimated separately, although the consistency property is guaranteed by mapping features into a universal kernel induced reproducing kernel Hilbert space (RKHS), the convergence rate may be slow. Note that the kernel density estimation method is non-parametric and thus it often requires a large sample size. Since density estimation is known to be a hard problem for high-dimensional variables, in practice, it is preferable to directly estimate the density ratio [44] and avoid estimating the densities separately.

5.3 The Density Ratio Estimation Method

Density ratio estimation [45] provides a way to significantly reduce the curse of dimensionality for kernel density estimation and can be estimated accurately for high-dimensional variables. Therefore, in this subsection, we introduce density ratio estimation to estimate the conditional probability distribution $P_{D_\rho}(\hat{Y}|X)$ for classification in the presence of RCN.

Three methods are frequently used for density ratio estimation, including the moment matching approach, the probabilistic classification approach and the ratio matching approach; see [46]. Since the probabilistic classification approach may introduce a large approximation error, in practice, the moment matching and ratio matching methods are more preferable [47], where the density ratio $r(X) = P_1(X)/P_2(X)$ can be modelled by employing linear or non-linear functions. If proper reproducing kernel Hilbert spaces are chosen to be the hypothesis classes, the approximation errors of the moment matching and ratio matching methods could be small. Although these methods introduce approximation errors for learning the weight $\beta^*(X, \hat{Y})$, their efficiency has been widely and empirically proven [48]–[50].

In this paper, we exploit the ratio matching approach that employs the Bregman divergence [51] (KLIEP [52]) to estimate the conditional probability distribution $P_{D_\rho}(\hat{Y}|X)$. It is proven that the ratio matching approach exploiting the Bregman divergence [51] is consistent with the optimal approximation in the hypothesis class[1].

The following theorem provides an assurance that our importance reweighting method that exploits density ratio estimation is consistent.

Theorem 3. When employing the density ratio estimation method to estimate the conditional probability distribution $P_{D_\rho}(\hat{Y}|X)$ (and $\beta(X, \hat{Y})$), if the hypothesis class for estimating the density ratio is chosen properly so that the approximation error is zero, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(R[D, \hat{f}_{n,\hat{\beta}}, \ell(\hat{f}_{n,\hat{\beta}}(X), Y)] - R[D, f^*, \ell(f^*(X), Y)] > \epsilon) = 0,$$

where $\hat{\beta}(X, \hat{Y})$ is the same as that defined in Theorem 2, $\hat{f}_{n,\hat{\beta}} = \min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}(X_i, \hat{Y}_i)\ell(f(X_i), \hat{Y}_i)$ and $f^* = \min_{f \in F} R[D, f, \ell(f(X), Y)]$.

The convergence rate is characterized in the following proposition.

Proposition 2. Under the settings of Theorem 3, if the Bregman divergence degenerates to square distance, for any $\delta > 0$, with probability at least $1 - 3\delta$, the following holds:

$$R[D, \hat{f}_{n,\hat{\beta}}, \ell(\hat{f}_{n,\hat{\beta}}(X), Y)] - R[D, f^*, \ell(f^*(X), Y)] \leq \mathcal{O}\left(\mathfrak{R}(\ell \circ F) + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\text{BDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right),$$

where $\mathfrak{R}_{\text{BDR}}$ is the Rademacher complexity induced by estimating the density ratio using square distance and is defined in Section 7.6.

The proofs of Theorem 3 and Proposition 2 are provided in the supplementary material.

The convergence rate in Proposition 2 could be a certain rate of order $\mathcal{O}(1/\min(n_+, n_-)^{1/4})$ because $\mathfrak{R}_{\text{SDR}} \leq \mathcal{O}(\sqrt{1/\min(n_+, n_-)})$, where n_+ and n_- denote the number of positive labels and negative labels of the noisy sample, respectively.

6 Estimating the Noise Rates

Most existing algorithms designed for RCN problems need the knowledge of the noise rates. Scott et al. [33], [34] developed lower bounds for the inversed noise rates $\pi_{+1} = P(Y = -1|\hat{Y} = +1)$ and $\pi_{-1} = P(Y = +1|\hat{Y} = -1)$, under the irreducibility assumption, which are consistent with the target inversed noise rates and can therefore be used as estimators for the inversed noise rates. However, the convergence rate could be slow. Then, during the preparation of this manuscript, Scott [35] released an efficient implementation to estimate the inversed noise rates and introduced the distributional assumption $\{(X, Y)|Y = 1\} \not\subset \{(X, Y)|Y = -1\}$ and $\{(X, Y)|Y = -1\} \not\subset \{(X, Y)|Y = 1\}$ to the label noise classification problem. The distributional assumption is sufficient for the irreducibility assumption and thus is slightly stronger. Scott then proved that the distributional assumption ensures an asymptotic convergence rate of order $\mathcal{O}(\sqrt{1/n})$ for estimating the inversed noise rates.

1. Parametric modeling is used for estimating density ratio. We provide the proof of consistency in Section 7.6.

To the best of our knowledge, no efficient method has been proposed to estimate the noise rates and how to estimate them remains an open problem [6]. We first provide upper bounds for the noise rates and show that with a mild assumption on the "clean" data, they can be used to efficiently estimate the noise rates.

Theorem 4. We have that

$$\rho_{\hat{Y}} \leq P_{D_\rho}(-\hat{Y}|X).$$

Moreover, if the assumption holds that there exists $x_{-1}, x_{+1} \in \mathcal{X}$, such that $P_D(Y = +1|x_{-1}) = P_D(Y = -1|x_{+1}) = 0$, we have

$$\rho_{-1} = P_{D_\rho}(\hat{Y} = +1|x_{-1})$$

and

$$\rho_{+1} = P_{D_\rho}(\hat{Y} = -1|x_{+1}),$$

which means

$$\rho_{-\hat{Y}} = \min_{X \in \mathcal{X}} P_{D_\rho}(\hat{Y}|X).$$

Theorem 4 shows that under the assumption that there exists $x_{-1}, x_{+1} \in \mathcal{X}$, such that $P_D(Y = +1|x_{-1}) = P_D(Y = -1|x_{+1}) = 0$, $\min_{X \in \mathcal{X}} P_{D_\rho}(\hat{Y}|X)$ is a consistent estimator for the noise rates. The convergence rate for estimating the noise rates is the same as that of estimating the conditional distribution $P_{D_\rho}(\hat{Y}|X)$. We therefore could obtain fast convergence rates for estimating the noise rates via finite sample analysis. For example, if the hypothesis class has proper conditions on its variance, the Rademacher complexity will quickly converge and is of order $\mathcal{O}(1/n)$ [42].

Remark 2. We have proven the consistency property of the joint estimation of the weight and classifier in Theorems 2 and 3, and characterized the convergence rates of the joint estimation in Proposition 2. According to Theorem 4, the results can be easily extended to the joint estimation of the weight, noise rate and classifier of our importance reweighting method. We provide detailed proofs in the supplementary material.

For classification problems, the assumption in Theorem 4 can be easily held. If an observation $x \in \mathcal{X}$ is far from the target classifier, it is likely that the conditional probability $P_D(y = +1|x)$ (or $P_D(y = -1|x)$) is equal to zero or very small. With the assumption that there exist $x_{-1}, x_{+1} \in \mathcal{X}$ such that $P_D(y = +1|x_{-1})$ and $P_D(y = -1|x_{+1})$ are very small, we can efficiently estimate ρ_y by

$$\min_{X \in \mathcal{X}} P_{D_\rho}(\hat{Y}|X).$$

In our experiments, we estimate ρ_y by

$$\hat{\rho}_{-\hat{Y}} = \min_{X \in \{X_1, \ldots, X_n\}} \hat{P}_{D_\rho}(\hat{Y}|X).$$

It is not hard to see that Scott's distributional assumption is equal to ours in Theorem 4. Interestingly, this is not a coincidence. In Proposition 3 of [35], Scott derived that

$$P_{D_\rho}(X|\hat{Y}) = \left(1 - \frac{\pi_{\hat{Y}}}{1 - \pi_{-\hat{Y}}}\right) P_D(X|\hat{Y}) + \frac{\pi_{\hat{Y}}}{1 - \pi_{-\hat{Y}}} P_{D_\rho}(X|-\hat{Y}).$$

Note that $\left(1 - \frac{\pi_{\hat{Y}}}{1-\pi_{-\hat{Y}}}\right) P_D(X|\hat{Y}) \geq 0$. Thus, $\frac{\pi_{\hat{Y}}}{1-\pi_{-\hat{Y}}}$ can be consistently estimated by $\min_{X \in \mathcal{X}} \frac{P_{D_\rho}(X|\hat{Y})}{P_{D_\rho}(X|-\hat{Y})}$ if there exists an $x \in \mathcal{X}$ such that $P_D(x|Y) = 0$, where $\frac{P_{D_\rho}(X|\hat{Y})}{P_{D_\rho}(X|-\hat{Y})}$ is the slope to the point $(1, 1)$ in the receiver operating characteristic (ROC) space defined in [33], [35]. In the proof of Theorem 4, we also derived that

$$P_{D_\rho}(\hat{Y}|X) = (1 - \rho_{-1} - \rho_{+1}) P_D(\hat{Y}|X) + \rho_{-\hat{Y}}.$$

Since $(1-\rho_{-1}-\rho_{+1}) P_D(\hat{Y}|X)$ is non-negative, our estimator $\rho_{-\hat{Y}} = \min_{X \in \mathcal{X}} P_{D_\rho}(\hat{Y}|X)$ is consistent based on the assumption that there exists an $x \in X$ such that $P_D(\hat{Y}|x) = 0$. Having the above knowledge in mind, we can improve the theoretical analysis for estimation the inversed noise rates in [33], [35] (and the mixture proportion estimation) by employing finite sample analysis.

We can design importance reweighting algorithms for classification with label noise by employing the inversed noise rates.

Lemma 2. When using the importance reweighting method to address the asymmetric RCN problem, the weight given to a noisy example $(X, \hat{Y}) \sim D_\rho$ can be derived by exploiting the inversed noise rates:

$$\begin{aligned} \beta(X, \hat{Y}) &= \frac{P_D(\hat{Y}|X)}{P_{D_\rho}(\hat{Y}|X)} \\ &= \frac{(1 - \pi_{-1} - \pi_{+1}) P_{D_\rho}(\hat{Y}|X) + \pi_{-\hat{Y}}}{P_{D_\rho}(\hat{Y}|X)}. \end{aligned}$$

The weight $\beta(X, \hat{Y})$ is non-negative[2] if $P_{D_\rho}(\hat{Y}|X) \neq 0$. If $P_{D_\rho}(\hat{Y}|X) = 0$, we intuitively let $\beta(X, \hat{Y}) = 0$.

Remark 3. We employed Scott's method [35] to estimate the inversed noise rates and found that the importance reweighting method exploiting the estimated inversed noise rates did not perform well, so the results are omitted. There might be two reasons which could possibly explain the poor performance: (1) Scott's estimator $\frac{\pi_{\hat{Y}}}{1-\pi_{-\hat{Y}}} = \min_{X \in \mathcal{X}} \frac{P_{D_\rho}(X|\hat{Y})}{P_{D_\rho}(X|-\hat{Y})}$ has the form of density ratio estimation, and is more complex than our estimator $\rho_{-\hat{Y}} = \min_{X \in \mathcal{X}} P_{D_\rho}(\hat{Y}|X)$, which has the form of the conditional distribution. (2) How to choose the kernel width to obtain the ROC in Scott's method has remained elusive.

7 PROOF

In this section, we provide detailed proofs of the assertions made in previous sections.

7.1 Proof of Lemma 1

For label noise problem, we have shown that

$$\beta(X, \hat{Y}) = \frac{P_D(\hat{Y}|X)}{P_{D_\rho}(\hat{Y}|X)}.$$

2. The inversed noise rates are defined so that $\pi_{-1} + \pi_{+1} \leq 1$, see, [35].

When the label noise is of asymmetric RCN, we have

$$\begin{aligned}
&P_{D_\rho}(+1|X)\\
&= P(\hat{Y}=+1, Y=+1|X) + P(\hat{Y}=+1, Y=-1|X)\\
&= P(\hat{Y}=+1|Y=+1, X)P_D(Y=+1|X)\\
&\quad + P(\hat{Y}=+1|Y=-1, X)P_D(Y=-1|X)\\
&= P(\hat{Y}=+1|Y=+1)P_D(Y=+1|X)\\
&\quad + P(\hat{Y}=+1|Y=-1)P_D(Y=-1|X)\\
&= (1-\rho_{+1})P_D(Y=+1|X)\\
&\quad + \rho_{-1}(1-P_D(Y=+1|X))\\
&= (1-\rho_{-1}-\rho_{+1})P_D(Y=+1|X) + \rho_{-1}\\
&\geq \rho_{-1}.
\end{aligned}$$

Similarly, it gives

$$\begin{aligned}
P_{D_\rho}(-1|X) &= (1-\rho_{-1}-\rho_{+1})P_D(Y=-1|X) + \rho_{+1}\\
&\geq \rho_{+1}.
\end{aligned}$$

We therefore have

$$P_D(\hat{Y}|X) = \frac{P_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})}.$$

Thus,

$$\begin{aligned}
\beta(X,\hat{Y}) &= \frac{P_D(\hat{Y}|X)}{P_{D_\rho}(\hat{Y}|X)}\\
&= \frac{P_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1-\rho_{+1}-\rho_{-1})P_{D_\rho}(\hat{Y}|X)}.
\end{aligned}$$

We intuitively let $\beta(X,\hat{Y}) = 0$, if $P_{D_\rho}(\hat{Y}|X) = 0$. Since $P_{D_\rho}(\hat{Y}|X) \geq \rho_{-\hat{Y}}$, we can conclude that $\beta(X,\hat{Y}) \geq 0$. ■

7.2 Proofs of Proposition 1 and Theorem 1

We start by introducing the Rademacher complexity method [41] for deriving generalization bounds.

Let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher variables, $X_1, \ldots, X_n$ be i.i.d. variables and F be a real-valued function class. The Rademacher complexity of the function class over the variable is defined as

$$\mathfrak{R}(F) = E_{X,\sigma}\left[\sup_{f\in F}\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(X_i)\right].$$

Theorem 5 ([41]). Let F be a real-valued function class on $\mathcal{X}$, $S = \{X_1, \ldots, X_n\} \in \mathcal{X}^n$ and

$$\Phi(S) = \sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n}E[f(X)] - f(X_i)\right|.$$

Then, $E_S[\Phi(S)] \leq \mathfrak{R}(F)$.

The following theorem, proven utilizing Theorem 5 and Hoeffding's inequality, plays an important role in deriving the generalization bounds.

Theorem 6 ([41]). Let F be an $[a,b]$-valued function class on $\mathcal{X}$, and $S = \{X_1, \ldots, X_n\} \in \mathcal{X}^n$. Then, for any $f \in F$ and any $\delta > 0$, with probability at least $1-\delta$, we have

$$E_X[f(X)] - \frac{1}{n}\sum_{i=1}^{n}f(X_i) \leq \mathfrak{R}(F) + (b-a)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

According to Theorem 6, we can easily prove that for any $[0,b]$-valued function class and $\delta > 0$, with probability at least $1-\delta$, the following holds

$$\begin{aligned}
&\sup_{f\in F}|E_{(X,\hat{Y})\sim D_\rho}\hat{R}_{\beta\ell,D_\rho} - \hat{R}_{\beta\ell,D_\rho}|\\
&\leq \mathfrak{R}(\beta\circ\ell\circ F) + b\sqrt{\frac{\log(1/\delta)}{2n}}.
\end{aligned}$$

Since β is upper bounded by

$$\frac{1-U}{1-\rho_{-1}-\rho_{+1}},$$

where

$$U = \min_{(X,\hat{Y})}\frac{\rho_{-\hat{Y}}}{P_{D_\rho}(\hat{Y}|X)},$$

using the Lipschitz composition property of Rademacher complexity, which is also known as the Talagrand's Lemma (see, e.g., Lemma 4.2 in [40]), we have

$$\mathfrak{R}(\beta\circ\ell\circ F) \leq \frac{1-U}{1-\rho_{-1}-\rho_{+1}}\mathfrak{R}(\ell\circ F).$$

Propostion 1 can be proven together with the fact that $E_{(X,\hat{Y})\sim D_\rho}[\hat{R}_{\beta\ell,D_\rho}] = R_{\beta\ell,D_\rho} = R_{\ell,D}$. ■

Theorem 1 follows from Proposition 1.

7.3 Proof of Theorem 2

We begin with the following lemma.

Lemma 3. Let $K(X_1,X_2) = k(X_1)k(X_2)$ be a universal kernel, where $k: \mathcal{X} \to \mathcal{H}$ is a feature map into a feature space. Let

$$\hat{P}_{D_\rho}(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}1_{\hat{Y}_i=\hat{Y}}$$

and

$$\hat{P}_{D_\rho}(X) = \frac{1}{n}\sum_{i=1}^{n}K(X,X_i).$$

Then, $\hat{P}_{D_\rho}(\hat{Y})$ and $\hat{P}_{D_\rho}(X)$ will converge to their target distributions $P_{D_\rho}(\hat{Y})$ and $P_{D_\rho}(X)$ in the induced RKHS $\mathcal{H}$, respectively.

The proof relies on the following theorem proven by Gretton et al. [39].

Theorem 7. Let $\mathcal{P}$ be the space of all probability distributions on an RKHS $\mathcal{H}$ induced by a universal kernel $K(X_1,X_2) = k(X_1)k(X_2)$. Define $\mu: \mathcal{P} \to \mathcal{H}$ as the expectation operator that $\mu(P) = E_{X\sim P(X)}[k(X)]$. The operator μ is a bijection between $\mathcal{P}$ and $\{\mu(P)|P\in\mathcal{P}\}$.

Proof of Lemma 3. Since

$$E[\hat{P}_{D_\rho}(\hat{Y})] = \frac{1}{n}\sum_{i=1}^{n}E1_{\hat{Y}_i=\hat{Y}} = P_{D_\rho}(\hat{Y}),$$

using the weak law of large numbers, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty}P\left(|\hat{P}_{D_\rho}(\hat{Y}) - P_{D_\rho}(\hat{Y})| \geq \epsilon\right) = 0.$$

So, $\hat{P}_{D_\rho}(\hat{Y})$ will converge to its target distribution $P_{D_\rho}(\hat{Y})$.

We then prove that $\hat{P}_{D_\rho}(X)$ converges to $P_{D_\rho}(X)$ in the RKHS by using Theorem 7 and showing that

$$\int \hat{P}_{D_\rho}(X)k(X)dX = E_{X\sim P_{D_\rho}(X)}[k(X)], \text{ when } n \to \infty.$$

We have that

$$\begin{aligned}\hat{P}_{D_\rho}(X) &= \frac{1}{n}\sum_{i=1}^{n} K(X, X_i) = \frac{1}{n}\sum_{i=1}^{n} k(X)k(X_i) \\ &= \frac{k(X)}{n}\sum_{i=1}^{n} k(X_i).\end{aligned}$$

By properly modifying the kernel map k by a constant so that $\int k^2(X)dX = 1$, we have

$$\begin{aligned}\int \hat{P}_{D_\rho}(X)k(X)dX &= \frac{1}{n}\sum_{i=1}^{n} k(X_i)\int k^2(X)dX \\ &= \frac{1}{n}\sum_{i=1}^{n} k(X_i).\end{aligned}$$

Moreover, for any $\epsilon > 0$, using Hoeffding's inequality, the following holds

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} k(X_i) - E_{X\sim P_{D_\rho}(X)}[k(X)]\right| \geq \epsilon\right) = 0.$$

By combing the above two equations, we can conclude that

$$\begin{aligned}\lim_{n\to\infty} P\Big(\Big|\int \hat{P}_{D_\rho}(X)k(X)dX \\ -E_{X\sim P_{D_\rho}(X)}[k(X)]\Big| \geq \epsilon\Big) = 0.\end{aligned}$$

According to Theorem 7, we have that the estimator $\hat{P}_{D_\rho}(x)$ will converge to $P_{D_\rho}(X)$ in $\mathcal{H}$. ■

Proof of Theorem 2. In the universal kernel induced RKHS, we have proven that

$$\hat{P}_{D_\rho}(\hat{Y}) = P_{D_\rho}(\hat{Y}), \text{ when } n \to \infty$$

and

$$\hat{P}_{D_\rho}(X) = P_{D_\rho}(X), \text{ when } n \to \infty.$$

Thus, we have

$$\hat{\beta}(X, \hat{Y}) = \beta(X, \hat{Y}), \text{ when } n \to \infty. \tag{4}$$

In Proposition 1, we have proven that

$$\begin{aligned}\sup_{f\in F} |R[D_\rho, f, \beta(X, \hat{Y})\ell(f(X), \hat{Y})] \\ -\hat{R}[D_\rho, f, \beta(X, \hat{Y})\ell(f(X), \hat{Y})]| = 0, \\ \text{when } n \to \infty.\end{aligned} \tag{5}$$

By substitution from equation (4) into equation (5), we have

$$\begin{aligned}\sup_{f\in F} |R[D_\rho, f, \hat{\beta}(X, \hat{Y})\ell(f(X), \hat{Y})] \\ -\hat{R}[D_\rho, f, \hat{\beta}(X, \hat{Y})\ell(f(X), \hat{Y})]| = 0, \\ \text{when } n \to \infty.\end{aligned} \tag{6}$$

Let

$$\hat{f}_{n,\hat{\beta}} = \min_{f\in F} \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}(X_i, \hat{Y}_i)\ell(f(k(X_i)), \hat{Y}_i)$$

and

$$f^* = \min_{f\in F} R[D, f, \ell(f(k(X)), Y)].$$

We have inequalities (7). The first inequality in inequalities (7) holds because of the definition of $\hat{f}_{n,\hat{\beta}}$.

For sufficiently large n, using equations (4), (6) and (7), we have

$$\begin{aligned}&R[D, \hat{f}_{n,\hat{\beta}}, \ell(\hat{f}_{n,\hat{\beta}}(k(X)), Y)] \\ &\quad -R[D, f^*, \ell(f^*k(X)), Y)] \\ &= R[D_\rho, \hat{f}_{n,\hat{\beta}}, \beta(X, \hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)), \hat{Y})] \\ &\quad -R[D_\rho, f^*, \beta(X, \hat{Y})\ell(f^*(k(X)), \hat{Y})] \\ &= R[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X, \hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)), \hat{Y})] \\ &\quad -R[D_\rho, f^*, \hat{\beta}(X, \hat{Y})\ell(f^*(k(X)), \hat{Y})] \\ &\leq 2\sup_{f\in F}\Big|\hat{R}[D_\rho, f, \hat{\beta}(X, \hat{Y})\ell(f(k(X)), \hat{Y})] \\ &\quad -R[D_\rho, f, \hat{\beta}(X, \hat{Y})\ell(f(k(X)), \hat{Y})]\Big| \\ &= 0.\end{aligned}$$

This concludes the proof of Theorem 2. ■

7.4 Proof of Theorem 4

In the proof of Lemma 1, we have proven that

$$P_{D_\rho}(+1|X) = (1 - \rho_{-1} - \rho_{+1})P_D(Y = +1|X) + \rho_{-1},$$

If there exists $x_{-1} \in \mathcal{X}$ such that

$$P_D(Y = +1|x_{-1}) = 0,$$

then

$$P_{D_\rho}(\hat{Y} = +1|x_{-1}) = \rho_{-1}.$$

Similarly,

$$P_{D_\rho}(-1|X) = (1 - \rho_{-1} - \rho_{+1})P_D(Y = -1|X) + \rho_{+1},$$

and if there exists $x_{+1} \in \mathcal{X}$ such that

$$P_D(Y = -1|x_{+1}) = 0,$$

which means

$$P_{D_\rho}(\hat{Y} = -1|x_{+1}) = \rho_{+1}.$$

We therefore have

$$\rho_{-\hat{Y}} = \min_{X\in\mathcal{X}} P_{D_\rho}(\hat{Y}|X),$$

which concludes the proof. ■

$$
\begin{aligned}
&R[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)),\hat{Y})] - R[D_\rho, f^*, \hat{\beta}(X,\hat{Y})\ell(f^*(k(X)),\hat{Y})] \\
&= R[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)),\hat{Y})] - \hat{R}[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)),\hat{Y})] \\
&\quad + \hat{R}[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)),\hat{Y})] - \hat{R}[D_\rho, f^*, \hat{\beta}(X,\hat{Y})\ell(f^*(k(X)),\hat{Y})] \\
&\quad + \hat{R}[D_\rho, f^*, \hat{\beta}(X,\hat{Y})\ell(f^*(k(X)),\hat{Y})] - R[D_\rho, f^*, \hat{\beta}(X,\hat{Y})\ell(f^*(k(X)),\hat{Y})] \\
&\leq R[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)),\hat{Y})] - \hat{R}[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(k(X)),\hat{Y})] \\
&\quad + \hat{R}[D_\rho, f^*, \hat{\beta}(X,\hat{Y})\ell(f^*(k(X)),\hat{Y})] - R[D_\rho, f^*, \hat{\beta}(X,\hat{Y})\ell(f^*(k(X)),\hat{Y})] \\
&\leq 2 \sup_{f \in F} |\hat{R}[D_\rho, f, \hat{\beta}(X,\hat{Y})\ell(f(k(X)),\hat{Y})] - R[D_\rho, f, \hat{\beta}(X,\hat{Y})\ell(f(k(X)),\hat{Y})]|. \qquad (7)
\end{aligned}
$$

$$
\mathfrak{R}_{\mathrm{BDR}} = E_{X \sim D_\rho, \sigma}\left[\frac{2}{n_2} \sum_{i=1}^{n_2} \sigma_i \nabla f(r(X_i^{\mathrm{de}})) r(X_i^{\mathrm{de}}) - \frac{2}{n_2} \sum_{i=1}^{n_2} \sigma_i f(r(X_i^{\mathrm{de}})) - \frac{2}{n_1} \sum_{i=1}^{n_1} \sigma_i \nabla f(r(X_i^{\mathrm{nu}})) \right]. \qquad (8)
$$

7.5 Proof of Lemma 2

Similar to the proof of Lemma 1, we have

$$
\begin{aligned}
&P_D(+1|X) \\
&= P(Y=+1, \hat{Y}=+1|X) + P(Y=+1, \hat{Y}=-1|X) \\
&= P(Y=+1|\hat{Y}=+1, X) P_{D_\rho}(\hat{Y}=+1|X) \\
&\quad + P(Y=+1|\hat{Y}=-1, X) P_{D_\rho}(\hat{Y}=-1|X) \\
&= P(Y=+1|\hat{Y}=+1) P_{D_\rho}(\hat{Y}=+1|X) \\
&\quad + P(Y=+1|\hat{Y}=-1) P_{D_\rho}(\hat{Y}=-1|X) \\
&= (1-\pi_{+1}) P_{D_\rho}(\hat{Y}=+1|X) \\
&\quad + \pi_{-1}(1 - P_{D_\rho}(\hat{Y}=+1|X)) \\
&= (1-\pi_{-1}-\pi_{+1}) P_{D_\rho}(\hat{Y}=+1|X) + \pi_{-1} \\
&\geq \pi_{-1}.
\end{aligned}
$$

We also have

$$
\begin{aligned}
&P_D(-1|X) \\
&= (1-\pi_{-1}-\pi_{+1}) P_{D_\rho}(\hat{Y}=-1|X) + \pi_{+1} \\
&\geq \pi_{+1}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\beta(X,\hat{Y}) &= \frac{P_D(Y|X)}{P_{D_\rho}(\hat{Y}|X)} \\
&= \frac{(1-\pi_{-1}-\pi_{+1}) P_{D_\rho}(\hat{Y}|X) + \pi_{-\hat{Y}}}{P_{D_\rho}(\hat{Y}|X)}.
\end{aligned}
$$

We intuitively let $\beta(X,\hat{Y}) = 0$, if $P_{D_\rho}(\hat{Y}|X) = 0$. Then, we can conclude that $\beta(X,\hat{Y}) \geq 0$. ■

7.6 Consistency of Density Ratio Estimation

We first introduce how to use the ratio matching method under the Bregman divergence to estimate

$$
r^*(X) = \frac{P_{D_\rho}(X|\hat{Y})}{P_{D_\rho}(X)}.
$$

The discrepancy from the true density ratio r^* to a density ratio model r measured by the Bregman divergence (BD) is as follows:

$$
\begin{aligned}
\mathrm{BD}_f(r^* \| r) = \int P_{D_\rho}(X) \{ &f(r^*(X)) - f(r(X)) \\
&- \nabla f(r(X))(r^*(X) - r(X)) \} \, dX,
\end{aligned}
$$

where f is a convex function and $\nabla f(X)$ denotes the subgradient of $f(X)$.

Let $X_1^{\mathrm{nu}}, \ldots, X_{n_1}^{\mathrm{nu}}$ be the i.i.d. sample of the numerator distribution and $X_1^{\mathrm{de}}, \ldots, X_{n_2}^{\mathrm{de}}$ the i.i.d. sample of the denominator distribution. An empirical approximation of $\mathrm{BD}_f(r^* \| r)$ is given by

$$
\begin{aligned}
\hat{\mathrm{BD}}_f(r^* \| r) \quad &= \frac{1}{n_2} \sum_{i=1}^{n_2} \nabla f(r(X_i^{\mathrm{de}})) r(X_i^{\mathrm{de}}) \\
&\quad - \frac{1}{n_2} \sum_{i=1}^{n_2} f(r(X_i^{\mathrm{de}})) - \frac{1}{n_1} \sum_{i=1}^{n_1} \nabla f(r(X_i^{\mathrm{nu}})).
\end{aligned}
$$

Let

$$
\hat{r}(X) = \arg\min_r \hat{\mathrm{BD}}_f(r^* \| r)
$$

and

$$
r' = \arg\min_r \mathrm{BD}_f(r^* \| r).
$$

If the hypothesis class includes r^*, we have

$$
\begin{aligned}
&\mathrm{BD}_f(r^* \| \hat{r}) \\
&= \mathrm{BD}_f(r^* \| \hat{r}) - \mathrm{BD}_f(r^* \| r') \\
&= \mathrm{BD}_f(r^* \| \hat{r}) - \hat{\mathrm{BD}}_f(r^* \| \hat{r}) + \hat{\mathrm{BD}}_f(r^* \| \hat{r}) \\
&\quad - \hat{\mathrm{BD}}_f(r^* \| r') + \hat{\mathrm{BD}}_f(r^* \| r') - \mathrm{BD}_f(r^* \| r') \\
&\leq \mathrm{BD}_f(r^* \| \hat{r}) - \hat{\mathrm{BD}}_f(r^* \| \hat{r}) \\
&\quad + \hat{\mathrm{BD}}_f(r^* \| r') - \mathrm{BD}_f(r^* \| r') \\
&\leq 2 \sup_r |\hat{\mathrm{BD}}_f(r^* \| r) - \mathrm{BD}_f(r^* \| r)|,
\end{aligned}
$$

where the first inequality holds because of the definition of $\hat{r}$.

Since $r(x)$ is usually modeled by linear or non-linear functions, we can assume that $r(x)$ has the range $[a, b]$ for all observations. Using the Rademacher method again, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
\mathrm{BD}_f(r^* \| \hat{r}) \quad &\leq 2 \sup_r |\hat{\mathrm{BD}}_f(r^* \| r) - \mathrm{BD}_f(r^* \| r)| \\
&\leq 2\mathfrak{R}_{\mathrm{BDR}} + C \sqrt{\frac{\log(1/\delta)}{2n}}, \qquad (9)
\end{aligned}
$$

where $\mathfrak{R}_{\mathrm{BDR}}$ defined in (8)) is the Rademacher complexity induced by estimating the density ratio exploiting Bregman divergence and C is a constant. The convergence

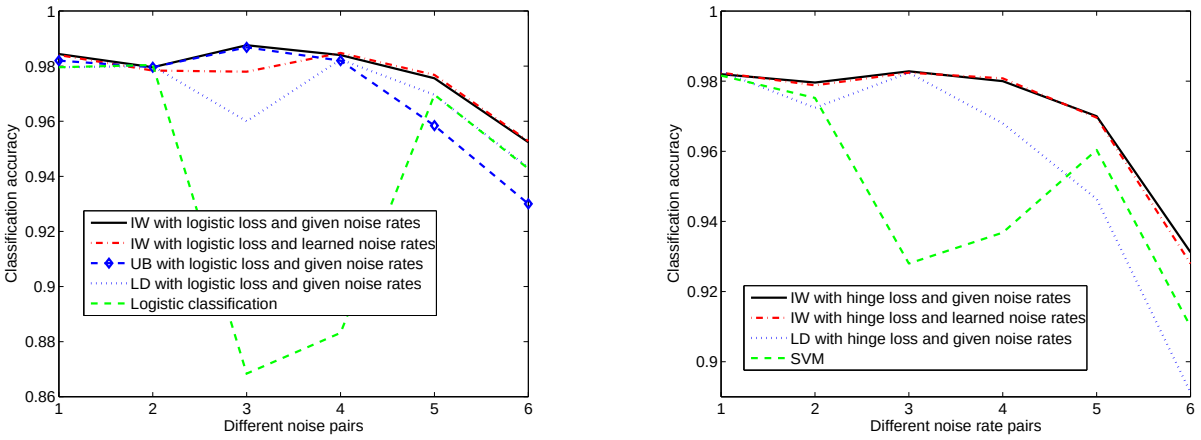

Fig. 1: Accuracy comparison of classification algorithms on synthetic data (m=2, n=1000). The six different noise rate pairs (ρ_{+1}, ρ_{-1}) are: (0.1, 0.1), (0.2, 0.2), (0.3, 0.1), (0.1, 0.3), (0.3, 0.3), and (0.4, 0.4). UB method employing hinge loss was not implemented due to non-convexity.

TABLE 1: Estimating the Noise Rates (Means and Standard Deviations) on Synthetic Data

Dataset (m, n)	True (ρ_{+1}, ρ_{-1})	Cross-validation	Scott's Method	Our Method
Synthetic dataset (2, 1000)	(0, 0.4)	**(0.036±0.032, 0.402±0.153)**	**(0.000±0.000, 0.382±0.032)**	**(0.019±0.001, 0.422±0.019)**
	(0.1, 0.3)	(0.322±0.163, 0.352±0.158)	**(0.089±0.028, 0.297±0.032)**	**(0.125±0.023, 0.325±0.023)**
	(0.2, 0.2)	(0.424±0.098, 0.352±0.171)	**(0.176±0.035, 0.203±0.036)**	**(0.213±0.031, 0.230±0.023)**
Synthetic dataset (20, 1000)	(0, 0.4)	(0.445±0.016, 0.235±0.024)	**(0.000±0.000, 0.320±0.049)**	**(0.074±0.021, 0.458±0.112)**
	(0.1, 0.3)	(0.440±0.021, 0.310±0.091)	**(0.078±0.022, 0.255±0.047)**	**(0.140±0.022, 0.328±0.090)**
	(0.2, 0.2)	(0.425±0.043, 0.445±0.016)	**(0.159±0.022, 0.168±0.037)**	**(0.214±0.064, 0.226±0.023)**

rate of $\mathfrak{R}_{\text{BDR}}$ can be proven to be as fast as the order $\mathcal{O}(\sqrt{1/\min(n_+, n_-)})$, where n_+ and n_- denote the number of positive labels and negative labels of the noisy sample, respectively. So, the ratio matching approach exploiting Bregman divergence is consistent to the optimal approximation in the hypothesis class.

8 Experiments

We next conducted experiments on synthetic and real data to illustrate the performance of the proposed approaches. Each dataset was randomly split 10 times, 75% for training and 25% for testing, and then the labels of the training sample flipped according to given noise rates ρ_{+1} and ρ_{-1}. The mean accuracies of the 10 datasets are presented.

To show the efficiency of our method for estimating the noise rates, we employed two baselines for comparison: the simple cross-validation method used in [10] and Scott's method [35] for estimating the inversed noise rates. Note that Scott's method can not be exploited to estimate the noise rates unless the knowledge $P_D(\pm 1)$ is given, which is often unknown in practice. To make the comparison, we assumed that the knowledge is known.

For the task of classification with noisy labels, the unbiased estimator (UBℓ) and label-dependent costs (LDℓ) models, developed by Natarajan et al. [10], and empirically shown to be competitive with, and perform better more often than, three of state-of-the-art robust methods (random projection classifier [27], NHERD [32], and the perceptron algorithm with margin [28]) for dealing with asymmetric RCN, were chosen as baselines for comparison[3]. We denote our importance reweighting method by IWℓ that estimates the conditional distribution $P_{D_\rho}(\hat{Y}|X)$ by employing the KLIEP method and the noise rates by using the cross-validation method, and denote our importance reweighting method by eIWℓ that exploits the KLIEP method to estimate the conditional distribution $P_{D_\rho}(\hat{Y}|X)$ and the noise rates jointly. Three-fold cross-validation[4] was used to tune the noise rates on the training sets when needed.

8.1 Synthetic Data

We first tested the performance of noise rate estimation on the synthetic dataset, where the data were uniformly distributed from the interval $[0, 1]$ and then linearly separated into two classes such that $P(Y = +1) = P(Y = -1)$. We used kernel density and density ratio estimation methods to estimate the noise rates on 2-dimensional and 20-dimensional synthetic data, respectively. The kernel width for kernel density estimation method was chosen as the standard deviation, and the density ratio was estimated using the KLIEP method [46]. The performances of the different methods are shown in Table 1, with entries having errors less than 0.1 shown in bold. Table 1 shows that our and Scott's methods for estimating the noise rates is far more

3. Note that comparisons in our paper and those in [10] are implemented on the same standard UCI classification datasets provided by Gunnar Rätsch: http://theoval.cmp.uea.ac.uk/matlab.

4. To achieve high performance for UBℓ, LDℓ and IWℓ, the optimal noise rates were chosen by the criterion that the classification accuracy rate, instead of the weighted objective function, is minimized on the validation set.

TABLE 2: Estimating the Noise Rates (Means and Standard Deviations) on UCI Benchmarks

Dataset (m, n)	True (ρ_{+1}, ρ_{-1})	Cross-validation	Scott's Method	Our Method
Heart (13, 270)	(0, 0.4)	(0.186±0.150, 0.050±0.047)	(0.073±0.043, 0.225±0.080)	**(0.027±0.013, 0.364±0.078)**
	(0.1, 0.3)	**(0.134±0.102, 0.220±0.168)**	**(0.089±0.031, 0.231±0.116)**	**(0.070±0.020, 0.272±0.070)**
	(0.2, 0.2)	**(0.110±0.109, 0.260±0.172)**	**(0.105±0.055, 0.179±0.047)**	**(0.121±0.063, 0.131±0.065)**
Diabetes (8, 768)	(0, 0.4)	(0.288±0.123, 0.206±0.118)	(0.122±0.036, 0.310±0.050)	**(0.026±0.012, 0.402±0.062)**
	(0.1, 0.3)	(0.154±0.099, 0.162±0.087)	(0.254±0.045, 0.249±0.057)	**(0.096±0.056, 0.304±0.060)**
	(0.2, 0.2)	**(0.138±0.103, 0.168±0.131)**	(0.361±0.127, 0.185±0.053)	**(0.135±0.036, 0.215±0.074)**

TABLE 3: Means and Standard Deviations (Percentage) of Classification Accuracies of all Kernel Hinge-loss-based Methods on UCI benchmarks

Bechmark dataset (m, n_+, n_-)	Noise rate (ρ_{+1}, ρ_{-1})	ℓ_{hinge}	LDℓ_{hinge}	StPMKL	eIWℓ_{hinge}	IWℓ_{hinge}
Breast cancer (9, 77, 186)	(0.2, 0.2)	64.24±6.59	65.76±9.29	**71.36±5.95**	69.39±5.91	**71.06±4.13**
	(0.3, 0.1)	67.12±8.69	70.61±5.16	**71.82±5.21**	68.79±7.57	**72.73±5.15**
	(0.4, 0.4)	57.88±5.52	54.18±11.84	**67.12±8.72**	65.30±7.64	**68.79±8.09**
Diabetes (8, 268, 500)	(0.2, 0.2)	71.56±4.20	71.77±4.51	65.00±2.50	**73.02±3.09**	**72.92±3.53**
	(0.3, 0.1)	**73.59±2.63**	73.23±2.37	66.46±2.75	**74.27±2.77**	71.46±3.04
	(0.4, 0.4)	66.77±2.37	66.25±4.31	**73.18±4.01**	**71.98±2.50**	71.77±3.38
German (20, 300, 700)	(0.2, 0.2)	67.20±3.55	**68.68±2.84**	**69.80±2.23**	67.20±3.45	68.08±2.85
	(0.3, 0.1)	68.56±2.62	**70.84±2.87**	67.24±1.78	68.76±2.29	**69.56±2.37**
	(0.4, 0.4)	62.32±2.81	62.04±5.90	**71.96±3.41**	**63.36±2.83**	63.04±2.89
Heart (13, 120, 150)	(0.2, 0.2)	67.21±5.33	**70.15±6.20**	**77.21±9.88**	68.82±5.62	68.82±5.08
	(0.3, 0.1)	70.59±8.05	**72.21±8.16**	54.71±7.96	70.74±8.42	**72.06±6.69**
	(0.4, 0.4)	68.68±6.12	67.94±13.28	59.12±13.30	**70.29±5.62**	**69.71±6.09**
Image (18, 1188, 898)	(0.2, 0.2)	**92.80±1.19**	92.16±0.95	73.35±1.80	**92.82±1.14**	92.49±0.93
	(0.3, 0.1)	91.30±1.99	**91.74±2.29**	58.89±6.72	91.02±1.70	**92.07±2.27**
	(0.4, 0.4)	**91.97±3.18**	90.98±1.49	57.24±2.83	**92.13±1.33**	89.37±3.45
Thyroid (5, 65, 150)	(0.2, 0.2)	**89.81±3.18**	**90.74±3.38**	70.93±3.50	88.52±3.58	87.41±4.43
	(0.3, 0.1)	**87.22±6.95**	**90.93±5.89**	69.81±6.93	85.19±7.81	81.85±6.58
	(0.4, 0.4)	**91.85±4.02**	88.52±13.26	82.22±12.60	91.76±2.66	**93.15±3.50**
Average		75.04	75.44	68.19	**76.29**	**76.48**

TABLE 4: Means and Standard Deviations (Percentage) of Classification Accuracies of all Kernel Logistic-loss-based Methods on UCI benchmarks

Bechmark dataset (m, n_+, n_-)	Noise rate (ρ_{+1}, ρ_{-1})	ℓ_{log}	LDℓ_{log}	UBℓ_{log}	eIWℓ_{log}	IWℓ_{log}
Breast cancer (9, 77, 186)	(0.2, 0.2)	73.48±5.16	**73.48±4.47**	72.73±4.46	72.88±6.04	**73.94±4.33**
	(0.3, 0.1)	**73.33±3.86**	70.91±5.09	71.67±5.49	71.36±5.41	**71.97±5.99**
	(0.4, 0.4)	66.36±8.41	**67.73±11.50**	67.09±8.24	**71.76±6.89**	65.61±7.69
Diabetes (8, 268, 500)	(0.2, 0.2)	**74.43±2.67**	72.24±2.78	72.92±2.95	**73.70±2.49**	72.45±2.74
	(0.3, 0.1)	**73.54±2.98**	73.12±4.29	72.34±4.71	**73.70±2.47**	73.33±3.62
	(0.4, 0.4)	70.21±4.56	71.04±5.10	**71.30±4.56**	**73.85±3.50**	70.83±3.67
German (20, 300, 700)	(0.2, 0.2)	69.28±2.20	68.80±2.66	**69.52±2.11**	**69.72±2.02**	69.00±2.76
	(0.3, 0.1)	**67.36±1.91**	67.20±2.17	67.28±1.89	**67.36±1.92**	67.32±2.04
	(0.4, 0.4)	60.60±7.13	60.36±9.15	**65.16±6.66**	**64.96±6.19**	64.56±6.95
Heart (13, 120, 150)	(0.2, 0.2)	**82.21±5.61**	80.29±7.44	**82.94±5.01**	81.32±10.36	81.91±4.44
	(0.3, 0.1)	69.41±9.37	**77.06±8.80**	75.88±9.02	75.44±9.33	**76.91±7.84**
	(0.4, 0.4)	69.56±10.17	**78.24±5.62**	76.18±7.03	**78.38±9.40**	77.50±7.04
Image (18, 1188, 898)	(0.2, 0.2)	**62.84±3.02**	59.85±7.36	**65.84±3.70**	62.16±4.68	61.72±4.88
	(0.3, 0.1)	**58.56±2.72**	57.47±1.82	56.15±1.90	**58.91±3.02**	58.26±2.69
	(0.4, 0.4)	60.48±7.60	63.72±4.36	**65.27±3.95**	**64.69±5.60**	62.26±4.07
Thyroid (5, 65, 150)	(0.2, 0.2)	89.07±4.40	**90.93±3.08**	90.37±3.47	86.11±6.37	**92.41±2.82**
	(0.3, 0.1)	84.26±4.12	**88.89±4.78**	85.04±6.36	82.59±4.38	**87.96±5.26**
	(0.4, 0.4)	86.48±6.88	87.04±9.28	86.30±9.16	**88.33±6.11**	**88.70±4.23**
Average		71.75	72.69	73.00	**73.13**	**73.23**

accurate than the simple cross-validation method and that our method is comparable with that of Scott on the synthetic datasets.

We next tested the performance of IWℓ, UBℓ, and LDℓ on the synthetic data. For fair comparison, we used the true noise rates for each model so that there was no tuning parameter for all the methods. Even though our method still needs to estimate the conditional probability $P_{D_\rho}(y|x)$, Fig. 1 shows that our method is more effective than the baselines when tested on these synthetic data. The empirical results also show that the logistic classification and SVM perform very bad when the noise rates are asymmetric.

8.2 Comparison on UCI Benchmarks

The comparison of noise rate estimation on two UCI datasets are shown in Table 2, with entries having error less than 0.1 shown in bold. The results show that the cross-validation method can estimate some noise rates with errors less than 0.1. However, this method produced large standard deviations. We employed the KLIEP method to estimate the noise rates. Table 2 illustrates that our method is more accurate than the baselines and has errors and standard deviations less than 0.1. These errors in our method occurred for two reasons: the first is that no example has very small $P_D(Y|X)$ in the Euclidian space (according to the

TABLE 5: Means and Standard Deviations (Percentage) of Classification Accuracies of all Linear Hinge-loss-based Methods on UCI benchmarks

Bechmark dataset (m, n_+, n_-)	Noise rate (ρ_{+1}, ρ_{-1})	ℓ_{hinge}	LDℓ_{hinge}	eIWℓ_{hinge}	IWℓ_{hinge}
Breast cancer (9, 77, 186)	(0.2, 0.2)	**71.82±3.93**	69.24±5.94	**71.36±4.65**	70.30±4.58
	(0.3, 0.1)	71.52±3.56	71.36±3.67	**72.42±3.83**	**72.42±3.90**
	(0.4, 0.4)	66.36±7.55	63.64±13.59	**71.67±4.58**	**67.88±4.83**
Diabetes (8, 268, 500)	(0.2, 0.2)	**76.88±1.89**	75.68±2.37	**75.99±6.55**	75.63±2.86
	(0.3, 0.1)	68.39±5.31	**73.59±5.19**	70.52±4.90	**74.22±4.99**
	(0.4, 0.4)	73.54±4.65	**75.05±5.43**	**76.41±2.38**	74.53±4.96
German (20, 300, 700)	(0.2, 0.2)	71.20±2.80	71.08±2.87	**71.56±2.62**	**71.76±1.90**
	(0.3, 0.1)	67.16±1.91	**67.16±1.78**	**67.24±1.78**	67.16±1.91
	(0.4, 0.4)	70.24±3.99	70.96±3.21	**71.08±2.41**	**72.56±3.44**
Heart (13, 120, 150)	(0.2, 0.2)	78.82±5.20	77.35±5.29	**81.18±5.49**	**78.97±6.51**
	(0.3, 0.1)	**75.88±8.06**	74.12±9.49	**79.26±4.46**	75.59±7.08
	(0.4, 0.4)	75.88±4.96	74.71±10.18	**78.97±5.05**	**78.38±6.76**
Image (18, 1188, 898)	(0.2, 0.2)	79.62±2.55	**82.30±2.03**	79.69±2.66	**82.55±2.21**
	(0.3, 0.1)	76.13±4.38	**82.09±1.74**	75.70±4.21	**82.78±1.30**
	(0.4, 0.4)	73.74±2.00	**81.84±2.53**	73.83±1.92	**80.21±4.20**
Thyroid (5, 65, 150)	(0.2, 0.2)	85.00±6.32	**87.78±4.11**	82.59±7.67	**88.33±2.90**
	(0.3, 0.1)	82.22±5.03	**85.19±6.23**	77.59±6.95	**84.26±4.96**
	(0.4, 0.4)	**86.11±5.18**	**85.93±4.55**	85.56±7.40	85.00±7.06
Average		75.03	**76.06**	75.72	**76.81**

TABLE 6: Means and Standard Deviations (Percentage) of Classification Accuracies of all Linear Logistic-loss-based Methods on UCI benchmarks

Bechmark dataset (m, n_+, n_-)	Noise rate (ρ_{+1}, ρ_{-1})	ℓ_{log}	LDℓ_{log}	UBℓ_{log}	eIWℓ_{log}	IWℓ_{log}
Breast cancer (9, 77, 186)	(0.2, 0.2)	70.00±5.88	71.52±4.83	**71.82±5.31**	**73.48±4.91**	71.52±4.51
	(0.3, 0.1)	**72.42±3.96**	70.76±4.63	**72.73±4.23**	71.06±4.49	71.06±4.76
	(0.4, 0.4)	63.33±6.42	60.30±16.24	61.52±16.12	**66.97±6.95**	**65.61±12.00**
Diabetes (8, 268, 500)	(0.2, 0.2)	**77.14±1.84**	**77.29±2.00**	76.51±2.22	74.79±2.52	76.72±2.00
	(0.3, 0.1)	74.48±2.33	**74.79±2.96**	**74.64±3.30**	74.48±3.44	74.37±3.49
	(0.4, 0.4)	71.20±3.17	72.86±5.29	71.93±4.89	**77.03±3.39**	**74.06±5.01**
German (20, 300, 700)	(0.2, 0.2)	**72.80±1.73**	72.68±1.93	**72.92±2.14**	71.56±2.95	72.64±1.28
	(0.3, 0.1)	**70.88±2.40**	68.76±1.96	70.04±2.03	**71.76±3.13**	69.08±2.52
	(0.4, 0.4)	70.60±3.23	70.88±5.23	**71.56±3.96**	**71.20±3.70**	70.84±5.20
Heart (13, 120, 150)	(0.2, 0.2)	77.50±3.80	76.91±6.01	77.79±5.74	**78.97±6.61**	**79.85±5.33**
	(0.3, 0.1)	74.85±7.02	74.26±6.36	74.41±9.56	**78.68±5.85**	**75.00±7.17**
	(0.4, 0.4)	74.71±5.27	73.82±7.49	**77.21±5.60**	**77.79±6.81**	72.50±6.09
Image (18, 1188, 898)	(0.2, 0.2)	**82.22±2.31**	82.05±2.21	81.99±2.76	82.18±2.47	**82.43±2.23**
	(0.3, 0.1)	73.24±4.28	80.29±2.67	**80.63±2.88**	72.76±3.92	**81.19±2.35**
	(0.4, 0.4)	77.59±1.79	81.32±2.03	**82.09±2.27**	77.93±1.85	**81.59±2.52**
Thyroid (5, 65, 150)	(0.2, 0.2)	**85.37±5.20**	85.00±4.14	85.00±5.12	84.07±6.66	**86.11±3.52**
	(0.3, 0.1)	82.22±4.88	**84.07±5.74**	82.22±5.60	81.30±4.14	**84.63±4.46**
	(0.4, 0.4)	80.37±6.99	85.37±4.66	84.26±5.67	**86.30±6.43**	**86.48±5.24**
Average		75.05	75.72	76.07	**76.24**	**76.43**

theory part, in this case, both Scott's and our methods cannot perform well), and the second is that the KLIEP method for estimating the conditional distribution is inaccurate due to the difficulty in choosing kernel width. However, when using the estimated noise rates, the performance of our method (eIWℓ) for classification with noisy labels performs better more often than the baselines (see Tables 3, 4, 5 and 6). Our noise rate estimation method is also valuable in the sense that some methods can benefit when the noise rates are approximately known [10].

To further demonstrate the efficiency of our importance reweighting method, we also tested multiple kernel learning from noisy labels by stochastic programming (StPMKL) [6] as a baseline on six UCI datasets, where all single kernel learning methods used Gaussian kernel with width 1 and StPMKL used Gaussian kernels with 10 different widths $\{2^{-3}, 2^{-2}, \ldots, 2^{6}\}$. The performances of the methods for different noise rates are shown in Tables 3, 4, 5 and 6, separately, with the two highest values in each row shown in bold. The results validate that our methods eIWℓ, which employ the estimated noise rates and run fast, perform better more often than the baselines in all datasets and that our methods IWℓ have average performances better than those of all the baselines. While our methods IW have average performances better than those of all the baselines, the methods as well as the baselines LD and UB need to learn the noise rates via cross-validation, which is time-consuming. Note that the kernel logistic-loss-based average performances in Tables 4 are lower than those in Tables 3, 5 and 6 because the kernel logistic-loss-based methods have low accuracies on the Image dataset.

According to Tables 3, 4, 5 and 6, the method IWℓ_{hinge}, which has the highest average performances, is preferred in practice. However, the method eIWℓ_{hinge}, which is time efficient and is competitive with all the baselines on all the datasets, should be sometimes preferred because of its low training time complexity.

9 Conclusions and Future Work

In this paper, we presented an importance reweighting framework for classification in the presence of label noise.

Theoretical analyses were provided to assure that the learned classifier will converge to the optimal one for the noise-free sample. Empirical studies on synthetic and real-world datasets verified the effectiveness and robustness of our proposed learning framework. We also provided a method to estimate the noise rates.

All our proposed methods crucially depend on the accuracy of the estimation of the conditional distribution $P_{D_\rho}(\hat{Y}|X)$. In future work, we need to consider how to accurately learn the conditional probability distribution for the noisy sample.

ACKNOWLEDGMENT

We greatly thank the handling Associate Editor and all anonymous reviewers for their valuable comments. The work was supported in part by Australian Research Council Projects DP-140102164 and FT-130101457.

REFERENCES

[1] D. Angluin and P. Laird, "Learning from noisy examples," *Machine Learning*, vol. 2, no. 4, pp. 343–370, 1988.
[2] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.
[3] M. Kearns, "Efficient noise-tolerant learning from statistical queries," *Journal of the ACM*, vol. 45, no. 6, pp. 983–1006, 1998.
[4] N. D. Lawrence and B. Schölkopf, "Estimating a kernel Fisher discriminant in the presence of label noise," in *ICML*, pp. 306–313, 2001.
[5] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *ACML*, pp. 97–112, 2011.
[6] T. Yang, M. Mahdavi, R. Jin, L. Zhang, and Y. Zhou, "Multiple kernel learning from noisy labels by stochastic programming," in *ICML*, 2012.
[7] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, 2014.
[8] J. A. Aslam and S. E. Decatur, "On the sample complexity of noise-tolerant learning," *Information Processing Letters*, vol. 57, no. 4, pp. 189–195, 1996.
[9] N. Manwani and P. Sastry, "Noise tolerance under risk minimization," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 1146–1151, 2013.
[10] N. Natarajan, I. Dhillon, P. Ravikumar, and A. Tewari, "Learning with noisy labels," in *NIPS*, pp. 1196–1204, 2013.
[11] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.
[12] C. Scott, "Calibrated asymmetric surrogate losses," *Electronic Journal of Statistics*, vol. 6, pp. 958–992, 2012.
[13] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.
[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.
[15] T. K. Pong, P. Tseng, S. Ji, and J. Ye, "Trace norm regularization: reformulations, algorithms, and multi-task learning," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3465–3489, 2010.
[16] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.
[17] P. Gong, C. Zhang, Z. Lu, J. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *ICML*, pp. 37–45, 2013.
[18] M. Kearns and M. Li, "Learning in the presence of malicious errors," *SIAM Journal on Computing*, vol. 22, no. 4, pp. 807–837, 1993.
[19] N. Cesa-Bianchi, E. Dichterman, P. Fischer, E. Shamir, and H. U. Simon, "Sample-efficient strategies for learning in the presence of noise," *Journal of the ACM*, vol. 46, no. 5, pp. 684–719, 1999.
[20] N. H. Bshouty, N. Eiron, and E. Kushilevitz, "PAC learning with nasty noise," *Theoretical Computer Science*, vol. 288, no. 2, pp. 255–275, 2002.
[21] P. M. Long and R. A. Servedio, "Learning large-margin halfspaces with more malicious noise," in *NIPS*, pp. 91–99, 2011.
[22] A. R. Klivans, P. M. Long, and R. A. Servedio, "Learning halfspaces with malicious noise," *Journal of Machine Learning Research*, vol. 10, pp. 2715–2740, 2009.
[23] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir, "Online learning of noisy data," *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp. 7907–7931, 2011.
[24] T. Bylander, "Learning linear threshold functions in the presence of classification noise," in *COLT*, pp. 340–347, ACM, 1994.
[25] E. Cohen, "Learning noisy perceptrons by a perceptron in polynomial time," in *FOCS*, pp. 514–523, IEEE, 1997.
[26] A. Blum, A. Frieze, R. Kannan, and S. Vempala, "A polynomial-time algorithm for learning noisy linear threshold functions," *Algorithmica*, vol. 22, no. 1-2, pp. 35–52, 1998.
[27] G. Stempfel and L. Ralaivola, "Learning kernel perceptrons on noisy data using random projections," in *ALT*, pp. 328–342, Springer, 2007.
[28] R. Khardon and G. Wachman, "Noise tolerant variants of the perceptron algorithm," *Journal of Machine Learning Research*, vol. 8, pp. 227–248, 2007.
[29] H. Moore and N. P. S. M. CALIF., *Robust Regression Using Maximum-Likelihood Weighting and Assuming Cauchy-Distributed Random Error*. Defense Technical Information Center, 1977.
[30] W. Liu, P. P. Pokharel, and J. C. Príncipe, "Correntropy: properties and applications in non-gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.
[31] R. He, W. Zheng, and B. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1561–1576, 2011.
[32] K. Crammer and D. D. Lee, "Learning via Gaussian herding," in *NIPS*, pp. 451–459, 2010.
[33] C. Scott, G. Blanchard, and G. Handy, "Classification with asymmetric label noise: Consistency and maximal denoising," in *COLT*, pp. 489–511, 2013.
[34] G. Blanchard, G. Lee, and C. Scott, "Semi-supervised novelty detection," *Journal of Machine Learning Research*, vol. 11, pp. 2973–3009, 2010.
[35] C. Scott, "A rate of convergence for mixture proportion estimation, with application to learning from noisy labels," in *AISTATS*, pp. 838–846, 2015.
[36] V. Vapnik, *The Nature of Statistical Learning Theory*. springer, 2000.
[37] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. cambridge university press, 2009.
[38] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 770–787, 2010.
[39] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," in *Dataset shift in machine learning* (J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, eds.), ch. 8, pp. 131–160, MIT press, 2009.
[40] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, 2012.
[41] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2003.
[42] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.
[43] I. Steinwart, "Support vector machines are universally consistent," *Journal of Complexity*, vol. 18, no. 3, pp. 768–791, 2002.
[44] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, "Relative density-ratio estimation for robust distribution comparison," in *NIPS*, pp. 594–602, 2011.
[45] V. Vapnik, I. Braga, and R. Izmailov, "Constructive setting of the density ratio estimation problem and its rigorous solution," *arXiv preprint arXiv:1306.0407*, 2013.
[46] M. Sugiyama, T. Suzuki, and T. Kanamori, "Density ratio estimation: A comprehensive review," *RIMS Kokyuroku*, pp. 10–31, 2010.
[47] T. Kanamori, T. Suzuki, and M. Sugiyama, "Theoretical analysis of density ratio estimation," *IEICE transactions on fundamentals*

of electronics, communications and computer sciences, vol. 93, no. 4, pp. 787–798, 2010.
[48] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *NIPS*, pp. 601–608, 2006.
[49] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation," *Journal of Machine Learning Research*, vol. 10, pp. 1391–1445, 2009.
[50] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye, "A two-stage weighting framework for multi-source domain adaptation," in *NIPS*, pp. 505–513, 2011.
[51] R. Nock and F. Nielsen, "Bregman divergences and surrogates for learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2048–2059, 2009.
[52] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *NIPS*, pp. 1433–1440, 2008.

Tongliang Liu received the B.E. degree in electronics engineering and information science from the University of Science and Technology of China, in 2012. He is currently pursuing the Ph.D. degree in computer science from the University of Technology, Sydney. He won the best paper award in the IEEE International Conference on Information Science and Technology 2014.

His research interests include statistical learning theory, computer vision, and optimization.

Dacheng Tao (F'15) is Professor of Computer Science with the Centre for Quantum Computation & Intelligent Systems, and the Faculty of Engineering and Information Technology in the University of Technology, Sydney. He mainly applies statistics and mathematics to data analytics problems and his research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 100+ publications at prestigious journals and prominent conferences, such as IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM07, the best student paper award in IEEE ICDM13, and the 2014 ICDM 10 Year Highest-Impact Paper Award.

The following is the supplementary material to the main manuscript providing proofs of Theorem 3, Proposition 2 and the assertions in Remark 2.

APPENDIX A
PROOFS OF THEOREM 3 AND PROPOSITION 2

Proof of Theorem 3. If the hypothesis class for estimating the density ratio is set properly so that the approximation error is zero, the target density ratio $r^*(X) = P_{D_\rho}(X|Y)/P_{D_\rho}(X)$ will be included in the hypothesis class. The consistency of the ratio matching approach exploiting Bregman divergence (proven in Section 7.6 of the submission) guarantees that the target density ratio $r^*(X)$ can be learned when n is sufficiently large. Using the proof method of Theorem 2, for any $\epsilon > 0$, we can prove that

$$\lim_{n\to\infty} P(R[D, \hat{f}_{n,\hat{\beta}}, \ell(\hat{f}_{n,\hat{\beta}}(X), Y)] - R[D, f^*, \ell(f^*(X), Y)] > \epsilon) = 0.$$

■

Proof of Proposition 2. If we let $f(r) = (t-1)^2/2$, the Bregman divergence degenerates to the square distance as follows

$$\mathrm{BD}_f(r^*\|r) = \mathrm{SD}(r^*\|r) = \frac{1}{2}(r^* - r)^2.$$

According to (9), for any $\delta > 0$, with probability at least $1-\delta$, we have

$$\begin{aligned}\mathrm{BD}(r^*\|\hat{r}) &= \frac{1}{2}(r^* - \hat{r})^2 \\ &\le 2\sup_r |\hat{\mathrm{BD}}(r^*\|r) - \mathrm{BD}(r^*\|r)| \\ &\le 2\mathfrak{R}_{\mathrm{BDR}} + C\sqrt{\frac{\log(1/\delta)}{2n}},\end{aligned}$$

where $\mathfrak{R}_{\mathrm{BDR}}$ is defined in (8). Thus, with probability at least $1-\delta$, the following holds

$$|r^* - \hat{r}| \le \mathcal{O}\left(\sqrt{\mathfrak{R}_{\mathrm{BDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right). \quad \text{(A.1)}$$

The convergence rate of $\mathfrak{R}_{\mathrm{BDR}}$ can be proven to be of order $\mathcal{O}(\sqrt{1/\min(n_+, n_-)})$, where n_+ and n_- denote the number of positive labels and negative labels of the noisy sample, respectively.

Note that $P_{D_\rho}(\hat{Y}|X) = \frac{P_{D_\rho}(X|\hat{Y})P_{D_\rho}(\hat{Y})}{P_{D_\rho}(X)}$, and that we estimate $\frac{P_{D_\rho}(X|\hat{Y})}{P_{D_\rho}(X)}$ by employing the Bregman divergence based ratio matching method and $P_{D_\rho}(\hat{Y} = \pm 1)$ by $\frac{1}{n}\sum_{i=1}^n 1_{\{\hat{Y}=\pm 1\}}$.

Using Hoeffding's inequality, with probability at least $1-\delta$, we have

$$\left|P_{D_\rho}(\hat{Y} = \pm 1) - \hat{P}_{D_\rho}(\hat{Y} = \pm 1)\right| \le \sqrt{\frac{\log(1/\delta)}{2n}}. \quad \text{(A.2)}$$

Combining Equations (A.1) and (A.2), with probability at least $1-2\delta$, we have

$$\begin{aligned}&\left|P_{D_\rho}(\hat{Y}|X) - \hat{P}_{D_\rho}(\hat{Y}|X)\right| \\ &\le \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right).\end{aligned}$$

Then, with probability at least $1-2\delta$, we have

$$\begin{aligned}&\beta(X, \hat{Y}) \\ &= \frac{P_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})P_{D_\rho}(\hat{Y}|X)} \\ &= \frac{1 - \frac{\rho_{-\hat{Y}}}{P_{D_\rho}(\hat{Y}|X)}}{1-\rho_{-1}-\rho_{+1}} \\ &\le \frac{1 - \frac{\rho_{-\hat{Y}}}{\hat{P}_{D_\rho}(\hat{Y}|X) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right)}}{1-\rho_{-1}-\rho_{+1}}\end{aligned}$$

Let $\Delta(n) \triangleq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right)$, we have

$$\begin{aligned}&\beta(X, \hat{Y}) \\ &\le \frac{\hat{P}_{D_\rho}(\hat{Y}|X) + \Delta(n) - \rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})(\hat{P}_{D_\rho}(\hat{Y}|X) + \Delta(n))} \\ &\le \frac{\hat{P}_{D_\rho}(\hat{Y}|X) + \Delta(n) - \rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)} \\ &\le \frac{\hat{P}_{D_\rho}(\hat{Y}|X) - \rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)} \\ &\quad + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right) \\ &= \hat{\beta}(X, \hat{Y}) \\ &\quad + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right)\end{aligned}$$

Hence, with probability at least $1-2\delta$, we have that

$$\begin{aligned}&R_{\beta\ell, D_\rho}(\hat{f}_{n,\hat{\beta}}) \\ &= R[D_\rho, \hat{f}_{n,\hat{\beta}}, \beta(X, \hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(X), \hat{Y})] \\ &= E_{(X,\hat{Y})\sim D_\rho}\left[\beta(X, \hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(X), \hat{Y})\right] \\ &\le E_{(X,\hat{Y})\sim D_\rho}\Big[\Big(\hat{\beta}(X, \hat{Y}) \\ &\quad + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right)\Big) \\ &\quad \ell(\hat{f}_{n,\hat{\beta}}(X), \hat{Y})\Big] \\ &= R\left[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X, \hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(X), \hat{Y})\right] \\ &\quad + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right) \\ &= R_{\hat{\beta}\ell, D_\rho}(\hat{f}_{n,\hat{\beta}}) \quad \text{(A.3)} \\ &\quad + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\mathrm{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right).\end{aligned}$$

Using the proof method of Proposition 1, with probability at least $1-\delta$, we have

$$\begin{aligned}
&R_{\hat{\beta}\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) - R_{\hat{\beta}\ell,D_\rho}(f^*)\\
&\leq 2\sup_{f\in F}\left|E_{(X,\hat{Y})\sim D_\rho}\left[\hat{R}_{\hat{\beta}\ell,D_\rho}\right] - \hat{R}_{\hat{\beta}\ell,D_\rho}\right|\\
&\leq 2\frac{1-\min_{(X,\hat{Y})}\frac{\rho_{-\hat{Y}}}{\hat{P}_{D_\rho}(\hat{Y}|X)}}{1-\rho_{-1}-\rho_{+1}}\mathfrak{R}(\ell\circ F) + 2b\sqrt{\frac{\log(1/\delta)}{2n}}\\
&\leq 2\frac{1-\min(\rho_{-1},\rho_{+1})}{1-\rho_{-1}-\rho_{+1}}\mathfrak{R}(\ell\circ F) + 2b\sqrt{\frac{\log(1/\delta)}{2n}},
\end{aligned}$$

or

$$\begin{aligned}
&R_{\hat{\beta}\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) - R_{\hat{\beta}\ell,D_\rho}(f^*)\\
&\leq \mathcal{O}\left(\mathfrak{R}(\ell\circ F) + \sqrt{\frac{\log(1/\delta)}{n}}\right). \qquad \text{(A.4)}
\end{aligned}$$

Combining Equations (A.3) and (A.4), with probability at least $1-3\delta$, we have

$$\begin{aligned}
&R_{\beta\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) - R_{\hat{\beta}\ell,D_\rho}(f^*)\\
&\leq \mathcal{O}\left(\mathfrak{R}(\ell\circ F) + \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\text{SDR}} + \sqrt{\frac{\log(1/\delta)}{n}}}\right),
\end{aligned}$$

which completes the proof. ■

APPENDIX B
PROOFS OF THE ASSERTIONS IN REMARK 2

B.1 Consistency of the Joint Estimation of the Noise Rate, Weight and Classifier

We have considered the consistency of the joint estimation of the weight and classifier in Theorems 2 and 3. Those consistency results can be easily extended to the joint estimation of the noise rate, weight and classifier. In Lemma 3, we have proven that the kernel density estimation method can learn the target weight $\beta^*(X,\hat{Y})$ by employing a universal kernel. Thus, given sufficiently large data and the assumption in Theorem 4 that $\exists x_{-1}, x_{+1} \in \mathcal{X}, P_D(Y=+1|x_{-1}) = P_D(Y=-1|x_{+1}) = 0$, Theorem 4 guarantees that we will learn the target noise rates. Theorem 2 therefore can be extended to provide a theoretical justification for the consistency of learning the optimal classifier in the hypothesis class with the estimated noise rates and weights. In Theorem 3, we have assumed that when employing the density ratio estimation method to learn the conditional distribution $P_{D_\rho}(\hat{Y}|X)$ and the hypothesis class is properly chosen, the corresponding approximation error is zero. The estimated noise rates therefore will converge to the target noise rates and the estimated weights will approach to the target weights under the assumption in Theorem 4. Thus, Theorem 3 can be extended to the consistency of the joint estimation of the noise rate, weight and classifier as well.

B.2 Convergence rate of the Joint Estimation of the Noise Rate, Weight and Classifier

We have discussed the convergence rate of the joint estimation of the weight and classifier in Proposition 2. In this subsection, we characterize a convergence rate for the joint estimation of the noise rate, weight and classifier.

Let $\hat{P}_{D_\rho}(\hat{Y}|X)$ be an estimator for $P_{D_\rho}(\hat{Y}|X)$ using equations (1), (2) and (3), and

$$\hat{\rho}_{-\hat{Y}} = \min_{X\in\{X_1,\ldots,X_n\}} \hat{P}_{D_\rho}(\hat{Y}|X). \qquad \text{(B.1)}$$

be the estimators for $\rho_{\pm 1}$.

Let defined the (learned) weight function comprised of the learned conditional distribution and learned noise rates as follows:

$$\hat{\beta}(X,\hat{Y}) = \frac{\hat{P}_{D_\rho}(\hat{Y}|X) - \hat{\rho}_{-\hat{Y}}}{(1-\hat{\rho}_{-1}-\hat{\rho}_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)}.$$

Note that the weight function is different from that defined in Theorem 2, where the noise rates are known and not estimated.

We also let

$$\hat{f}_{n,\hat{\beta}} = \min_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}(X_i,\hat{Y}_i)\ell(f(X_i),\hat{Y}_i)$$

and

$$f^* = \min_{f\in F} R[D,f,\ell(f(X),Y)].$$

Then, the convergence rate for the joint estimation of the learned noise rate, weight and classifier is characterized as follows:

Proposition 3. Under the settings of Theorem 3, if the Bregman divergence degenerates to square distance, for any $\delta>0$, with probability at least $1-9\delta$, the following holds:

$$\begin{aligned}
&R[D,\hat{f}_{n,\hat{\beta}},\ell(\hat{f}_{n,\hat{\beta}}(X),Y)] - R[D,f^*,\ell(f^*(X),Y)]\\
&\leq \frac{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)}{\left(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right)^2}\\
&\quad+\mathcal{O}\left(\mathfrak{R}(\ell\circ F)+\sqrt{\frac{\log(1/\delta)}{n}}\right),
\end{aligned}$$

Proof of Proposition 3. In the proof of Proposition 2, we have proven that with probability at least $1-2\delta$, we have

$$\begin{aligned}
&\left|P_{D_\rho}(\hat{Y}|X) - \hat{P}_{D_\rho}(\hat{Y}|X)\right| \qquad \text{(B.2)}\\
&\leq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).
\end{aligned}$$

We have also proven that with probability at least $1-2\delta$, we have

$$\begin{aligned}
&\frac{P_{D_\rho}(\hat{Y}|X)-\rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})P_{D_\rho}(\hat{Y}|X)}\\
&\leq \frac{\hat{P}_{D_\rho}(\hat{Y}|X)-\rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)}\\
&\quad+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).
\end{aligned}$$

According to (B.1) and (B.2), with probability at least $1-2\delta$, we have

$$|\hat{\rho}_{\pm 1}-\rho_{\pm 1}| \leq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).$$

This is because, with probability at least $1-2\delta$, we have

$$\begin{aligned}\hat{\rho}_{\pm 1}-\rho_{\pm 1} &= \min_{X\in\{X_1,\ldots,X_n\}} \hat{P}_{D_\rho}(\hat{Y}|X) - \min_{X\in\mathcal{X}} P_{D_\rho}(\hat{Y}|X)\\ &\leq \max_{X\in\mathcal{X}} \hat{P}_{D_\rho}(\hat{Y}|X) - P_{D_\rho}(\hat{Y}|X)\\ &\leq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\end{aligned}$$

and

$$\begin{aligned}\rho_{\pm 1}-\hat{\rho}_{\pm 1} &= \min_{X\in\mathcal{X}} P_{D_\rho}(\hat{Y}|X) - \min_{X\in\{X_1,\ldots,X_n\}} \hat{P}_{D_\rho}(\hat{Y}|X)\\ &\leq \max_{X\in\mathcal{X}} P_{D_\rho}(\hat{Y}|X) - \hat{P}_{D_\rho}(\hat{Y}|X)\\ &\leq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).\end{aligned}$$

Thus, with probability at least $1-4\delta$, it holds that

$$\begin{aligned}&\frac{\hat{P}_{D_\rho}(\hat{Y}|X)-\rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)}\\ &\quad+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\\ &\leq \frac{\hat{P}_{D_\rho}(\hat{Y}|X)-\hat{\rho}_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)} \qquad \text{(B.3)}\\ &\quad+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).\end{aligned}$$

Since with probability at least $1-2\delta$, it holds that $\hat{\rho}_{\pm 1} \geq \rho_{\pm 1} - \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)$, with probability at least $1-4\delta$, we have

$$\frac{1}{1-\rho_{-1}-\rho_{+1}} \leq \frac{1}{1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\Delta(n)}, \qquad \text{(B.4)}$$

where $\Delta(n) \triangleq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)$.

We now prove that for any $a>0, b>0, a+b<1$, it holds that

$$\frac{1}{1-a-b} \leq \frac{1}{1-a}+\frac{b}{(1-a-b)^2}. \qquad \text{(B.5)}$$

This is because

$$\begin{aligned}&\frac{1}{1-a-b} \leq \frac{1}{1-a}+\frac{b}{(1-a-b)^2}\\ \Leftrightarrow\quad & 1-a \leq 1-a-b+\frac{b(1-a)}{1-a-b}\\ \Leftrightarrow\quad & 1 \leq \frac{1-a}{1-a-b}.\end{aligned}$$

Combining inequalities (B.4) and (B.5), with probability at least $1-4\delta$, we have

$$\begin{aligned}&\frac{1}{1-\rho_{-1}-\rho_{+1}}\\ &\leq \frac{1}{1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\Delta(n)} \qquad \text{(B.6)}\\ &\leq \frac{1}{1-\hat{\rho}_{-1}-\hat{\rho}_{+1}}+\frac{\Delta(n)}{(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\Delta(n))^2}\end{aligned}$$

Combining inequalities (B.3) and (B.6), we have that with probability at least $1-8\delta$, the following holds

$$\begin{aligned}&\beta(X,\hat{Y})\\ &\leq \frac{\hat{P}_{D_\rho}(\hat{Y}|X)-\rho_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)}\\ &\quad+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\\ &\leq \frac{\hat{P}_{D_\rho}(\hat{Y}|X)-\hat{\rho}_{-\hat{Y}}}{(1-\rho_{-1}-\rho_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)}\\ &\quad+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).\\ &\quad\text{(According to } \hat{P}_{D_\rho}(\hat{Y}|X)-\hat{\rho}_{-\hat{Y}} \geq 0 \text{ and (B.6))}\\ &\leq \frac{\hat{P}_{D_\rho}(\hat{Y}|X)-\hat{\rho}_{-\hat{Y}}}{(1-\hat{\rho}_{-1}-\hat{\rho}_{+1})\hat{P}_{D_\rho}(\hat{Y}|X)}\\ &\quad+\frac{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)}{\left(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right)^2}\\ &\quad+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).\\ &= \hat{\beta}(X,\hat{Y}) \qquad \text{(B.7)}\\ &\quad+\frac{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)}{\left(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right)^2}\\ &\quad+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right).\end{aligned}$$

Hence, with probability at least $1-8\delta$, we have that

$$
\begin{aligned}
&R_{\beta\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) \\
&= R[D_\rho, \hat{f}_{n,\hat{\beta}}, \beta(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(X),\hat{Y})] \\
&= E_{(X,\hat{Y})\sim D_\rho}\left[\beta(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(X),\hat{Y})\right] \\
&\le E_{(X,\hat{Y})\sim D_\rho}\left[\left(\hat{\beta}(X,\hat{Y})\right.\right. \\
&\quad + \frac{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)}{\left(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right)^2} \\
&\quad \left.+\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right) \\
&\quad \left.\ell(\hat{f}_{n,\hat{\beta}}(X),\hat{Y})\right] \\
&= R\left[D_\rho, \hat{f}_{n,\hat{\beta}}, \hat{\beta}(X,\hat{Y})\ell(\hat{f}_{n,\hat{\beta}}(X),\hat{Y})\right] \\
&\quad + \frac{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)}{\left(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right)^2} \\
&\quad +\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right) \\
&= R_{\hat{\beta}\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) \\
&\quad + \frac{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)}{\left(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right)^2} \\
&\quad +\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right). \qquad \text{(B.8)}
\end{aligned}
$$

Using the proof method of Proposition 1, with probability at least $1-\delta$, we have

$$
\begin{aligned}
&R_{\hat{\beta}\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) - R_{\hat{\beta}\ell,D_\rho}(f^*) \\
&\le 2\sup_{f\in F}\left|E_{(X,\hat{Y})\sim D_\rho}\left[\hat{R}_{\hat{\beta}\ell,D_\rho}\right] - \hat{R}_{\hat{\beta}\ell,D_\rho}\right| \\
&\le 2\frac{1-\min_{(X,\hat{Y})}\frac{\hat{\rho}_{-\hat{Y}}}{\hat{P}_{D_\rho}(\hat{Y}|X)}}{1-\hat{\rho}_{-1}-\hat{\rho}_{+1}}\mathfrak{R}(\ell\circ F) + 2b\sqrt{\frac{\log(1/\delta)}{2n}} \\
&\le 2\frac{1-\min(\hat{\rho}_{-1},\hat{\rho}_{+1})}{1-\hat{\rho}_{-1}-\hat{\rho}_{+1}}\mathfrak{R}(\ell\circ F) + 2b\sqrt{\frac{\log(1/\delta)}{2n}},
\end{aligned}
$$

or

$$
\begin{aligned}
&R_{\hat{\beta}\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) - R_{\hat{\beta}\ell,D_\rho}(f^*) \\
&\le \mathcal{O}\left(\mathfrak{R}(\ell\circ F) + \sqrt{\frac{\log(1/\delta)}{n}}\right). \qquad \text{(B.9)}
\end{aligned}
$$

Combining Equations (B.8) and (B.9), with probability at least $1-9\delta$, we have

$$
\begin{aligned}
&R_{\beta\ell,D_\rho}(\hat{f}_{n,\hat{\beta}}) - R_{\hat{\beta}\ell,D_\rho}(f^*) \\
&\le \frac{\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)}{\left(1-\hat{\rho}_{-1}-\hat{\rho}_{+1}-\mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}+\sqrt{\mathfrak{R}_{\text{SDR}}+\sqrt{\frac{\log(1/\delta)}{n}}}\right)\right)^2} \\
&\quad +\mathcal{O}\left(\mathfrak{R}(\ell\circ F) + \sqrt{\frac{\log(1/\delta)}{n}}\right),
\end{aligned}
$$

which completes the proof. ■