



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 박 사 학 위 논 문

Use of Integral Probability Metrics in
Machine Learning

기계학습에서의 적분 확률 측도 사용

2020년 2월

서울대학교 대학원

통계학과

권 용 찬

Use of Integral Probability Metrics in Machine
Learning

기계학습에서의 적분 확률 측도 사용

지도교수 Myunghee Cho Paik

이 논문을 이학박사 학위논문으로 제출함

2019년 10월

서울대학교 대학원

통계학과

권 용 찬

권용찬의 이학박사 학위논문을 인준함

2019년 12월

위 원 장 김 용 대 (인)

부위원장 Myunghee Cho Paik (인)

위 원 장 원 철 (인)

위 원 원 중 호 (인)

위 원 Masashi Sugiyama (인)

Use of Integral Probability Metrics in Machine Learning

By

Yongchan Kwon

A Thesis

Submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
February, 2020

ABSTRACT

Use of Integral Probability Metrics in Machine Learning

Yongchan Kwon

The Department of Statistics

The Graduate School

Seoul National University

This doctoral thesis deals with two machine learning problems using integral probability metrics (IPMs). The first research problem is about learning binary classifiers using only positive and unlabeled observations, called PU learning. Recent studies in PU learning have shown promising empirical performance. However, most existing algorithms may not be suitable for large-scale datasets because they require repeated computations of a large Gram matrix. In this work, we define weighted IPMs and we propose a family of classifiers based on the metrics. We show a special case of the proposed family provides a computationally efficient PU learning algorithm. The proposed algorithm produces a closed-form classifier when the hypothesis space is a closed ball in reproducing

kernel Hilbert space. Furthermore, we present a new excess risk bound for the proposed family of classifiers. To the best of our knowledge, this is the first result to explicitly derive the excess risk bound in PU learning.

The second part is to build grounds for regularized risk minimization with augmented data in the context of Wasserstein distributionally robust optimization (WDRO). Data augmentation has recently emerged as a key technology in the field of machine learning to improve empirical performance. However, most augmentation techniques are based on heuristics, and their theoretical bases are limited to account for current successes. In this thesis, we formalize learning models with augmented data in the context of WDRO. When a loss function has Hölder continuous gradient, we show that regularized empirical risk evaluated at augmented data approximates the worst-case risk. We propose to minimize the regularized empirical risk, and we show the minimizer attains risk consistency. Based on the theoretical results, we propose a gradient-based algorithm producing a robust prediction model. Multiple real data experiments demonstrate robustness of the proposed model on noisy datasets. This is the first rigorous method to use augmented data and deep neural networks in WDRO.

Keywords: Integral probability metric, Positive-unlabeled learning, Reproducing kernel Hilbert space, Data augmentation, Distributionally robust optimization.

Student Number: 2013 – 22897

Contents

| | |
|--|----------|
| Abstract | i |
| 1 Introduction | 1 |
| 1.1 Learning binary classifiers using only positive and unlabeled observations | 1 |
| 1.2 Learning models with augmented data: Wasserstein distributionally robust optimization perspective | 5 |
| 2 Principled analytic classifier for positive-unlabeled learning via weighted integral probability metric | 7 |
| 2.1 Preliminaries | 7 |
| 2.1.1 Problem settings of PU learning | 7 |
| 2.1.2 L -risk minimization in PU learning | 9 |
| 2.2 Weighted integral probability metric and L -risk | 10 |
| 2.2.1 Relation between IPM and L -risk in supervised binary classification | 11 |
| 2.2.2 Extension to WIPM and L -risk in PU learning | 12 |
| 2.2.3 Theoretical properties of empirical WIPM optimizer | 14 |

| | | |
|-------|--|----|
| 2.3 | WIPM optimizer with reproducing kernel Hilbert space | 17 |
| 2.3.1 | An analytic classifier via WMMD optimizer | 17 |
| 2.3.2 | Explicit excess risk bound of WMMD classifier | 19 |
| 2.4 | Related work | 21 |
| 2.5 | Numerical experiments | 22 |
| 2.6 | Concluding remarks | 34 |
| 2.7 | Appendix | 34 |
| 2.7.1 | Proof of Theorem 2.2.1 | 34 |
| 2.7.2 | Proofs for Section 2.2.3: Theoretical properties of empirical WIPM optimizer | 35 |
| 2.7.3 | Proofs for Section 2.3: The empirical WMMD optimizer and the WMMD classifier | 42 |
| 2.7.4 | Implementation details | 50 |
| 2.7.5 | Comparison between Gaussian and inverse kernels | 54 |

3 Principled learning with augmented data: Wasserstein distributionally robust optimization perspective **57**

| | | |
|-------|--|----|
| 3.1 | Backgrounds | 57 |
| 3.1.1 | Statistical learning theory and distributionally robust optimization | 58 |
| 3.1.2 | Data augmentation by linear interpolation . | 59 |
| 3.2 | Wasserstein distributionally robust optimization . | 61 |
| 3.3 | Principled learning with augmented data in the context of WDRO | 62 |
| 3.4 | Numerical experiments | 64 |

| | | |
|-------|----------------------------------|-----------|
| 3.5 | Concluding remarks | 69 |
| 3.6 | Appendix | 70 |
| 3.6.1 | Proof of Theorem 3.3.1 | 70 |
| 3.6.2 | Proof of Theorem 3.3.2 | 75 |
| 3.6.3 | Implementation details | 78 |
| | Bibliography | 81 |
| | Abstract in Korean | 91 |

List of Tables

| | | |
|-----|--|----|
| 2.1 | A summary of elapsed training time and its ratio for the five PU learning algorithms | 29 |
| 2.2 | A summary of the eleven binary classification datasets | 30 |
| 2.3 | Accuracy and AUC comparison using the real datasets in cases the class-prior π_+ is known. | 32 |
| 2.4 | Accuracy and AUC comparison using the real datasets in cases the class-prior π_+ is unknown. | 33 |
| 3.1 | Accuracy comparison using the clean and noisy test datasets | 67 |
| 3.2 | The comparison of the accuracy of the three methods with the different noisy intensities | 68 |
| 3.3 | The comparison of the accuracy of MIXUP and DROID | 69 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | The illustration of the decision boundaries of WMMD classifier using <code>two_moons</code> dataset | 24 |
| 2.2 | The comparison of accuracy and AUC of the five PU learning algorithms when each n_u and π_+ changes | 25 |
| 2.3 | The comparison of accuracy and AUC of the five PU learning algorithms when each n_u and π_+ changes in case where the class-prior is unknown | 26 |
| 2.4 | The comparison of the accuracy and AUC of the LOG, DH, and WMMD algorithms with the Gaussian and the inverse kernels | 56 |
| 3.1 | Example of clean and noisy images. | 65 |

Chapter 1

Introduction

This doctoral thesis deals with two machine learning problems using integral probability metrics (IPMs) [Müller, 1997]. The first research problem is about learning binary classifiers using only positive and unlabeled observations, called PU learning. The second part is to formalize a group of data augmentation methods, including Mixup, in the context of distributionally robust optimization (DRO).

1.1 Learning binary classifiers using only positive and unlabeled observations

Supervised binary classification has shown a remarkable success in many real-world applications based on a large amount of labeled data. However, collecting such samples from the two categories is often costly, difficult, or not even possible. In contrast, unlabeled data are relatively cheap and abundant. As a consequence, semi-supervised learning is used for partially labeled data [Chapelle

et al., 2006]. In this thesis, as a special case of semi-supervised learning, we consider Positive-Unlabeled (PU) learning, the problem of building a binary classifier from only positive and unlabeled samples [Denis et al., 2005, Li and Liu, 2005]. PU learning provides a powerful framework when negative labels are impossible or very expensive to obtain, and thus has frequently appeared in many real-world applications. Examples include document classification [Elkan and Noto, 2008, Xiao et al., 2011], image classification [Zuluaga et al., 2011, Gong et al., 2018], gene identification [Yang et al., 2012, 2014], and novelty detection [Blanchard et al., 2010, Zhang et al., 2017a].

Several PU learning algorithms have been developed over the last two decades. Liu et al. [2002] and Li and Liu [2003] considered a two-step learning scheme: in Step 1, assigning negative labels to some unlabeled observations believed to be negative, and in Step 2, learning a binary classifier with existing positive samples and the negatively labeled samples from Step 1. Liu et al. [2003] pointed out that the two-step learning scheme is based on heuristics, and suggested fitting a biased support vector machine by regarding all the unlabeled observations as being negative.

Scott and Blanchard [2009] and Blanchard et al. [2010] suggested a modification of supervised Neyman-Pearson classification, whose goal is to find a classifier minimizing the false positive rate keeping the false negative rate low. To circumvent the problem of lack of negative samples, they tried to build a classifier minimizing the marginal probability of being classified as positive while keeping the false negative rate low. Solving the empirical version

of this constrained optimization problem is challenging, but the authors did not present an explicit algorithm.

Recently, many PU learning algorithms based on the empirical risk minimization principle have been studied. du Plessis et al. [2014] proposed the use of the ramp loss and provided an algorithm that requires solving a non-convex optimization problem. du Plessis et al. [2015] formulated a convex optimization problem by using the logistic loss or double hinge loss. However, all the aforementioned approaches involve solving a non-linear programming problem. This causes massive computational burdens for calculating the large Gram matrix when the sample size is large. Kiryo et al. [2017] suggested a stochastic algorithm for large-scale datasets with a non-negative risk estimator. However, to execute the algorithm, several hyperparameters are required, and choosing the optimal hyperparameter may demand substantial trials of running the algorithm, causing heavy computation costs.

In supervised binary classification, Sriperumbudur et al. [2012] proposed a computationally efficient algorithm building a closed-form binary discriminant function. The authors showed that their function estimator obtained by evaluating the negative of the empirical integral probability metric (IPM) is the minimizer of the empirical risk using the specific loss defined in Section 2.2.1. They further showed that a closed form can be derived as the result of restricting a hypothesis space to a closed unit ball in reproducing kernel Hilbert space (RKHS).

In this thesis, capitalizing on the properties shown in the supervised learning method by Sriperumbudur et al. [2012], we extend

it to PU learning settings. In addition, we derive new theoretical results on excess risk bounds. We first define a weighted version of IPM between two probability measures and call it the weighted integral probability metric (WIPM). We show that computing the negative of WIPM between the unlabeled data distribution and the positive data distribution is equivalent to minimizing the hinge risk. Based on this finding, we propose a binary discriminant function estimator that computes the negative of the empirical WIPM, and then derive associated upper bounds of the estimation error and the excess risk. Under a mild condition, our obtained upper bounds are shown to be sharper than the existing ones because of using Talagrand’s inequality over McDiarmid’s inequality [Kiryo et al., 2017]. Moreover, we pay special attention to the case where the hypothesis space is a closed ball in RKHS and propose a closed-form classifier. We show that the associated excess risk bound has an explicit form that converges to zero as the sample sizes increase. To the best of our knowledge, this is the first result to explicitly show the excess risk bound in PU learning. As a summary:

- We formally define WIPM and establish a link with the infimum of the hinge risk (Theorem 2.2.1) and derive an estimation error bound (Theorem 2.2.2).
- The proposed algorithm produces a closed-form classifier when the underlying hypothesis space is a closed ball in RKHS (Proposition 2.3.1). Furthermore, we obtain a novel excess risk bound that converges to zero as sample sizes increase (Theorem 2.3.4).

- Numerical experiments using both synthetic and real datasets show that our method is comparable to or better than existing PU learning algorithms in terms of accuracy and scalability.

Most of the PU learning work in this thesis was previously presented in Kwon et al. [2019]. In this thesis, we present a new excess risk bound in Theorem 2.3.4.

1.2 Learning models with augmented data: Wasserstein distributionally robust optimization perspective

Data augmentation is a key technique to improve empirical performance in the field of machine learning and computer vision [Cubuk et al., 2018, Lim et al., 2019]. For example, Mixup and its variants have led remarkable generalization ability in supervised and semi-supervised learning tasks [Zhang et al., 2017b, Verma et al., 2019, Berthelot et al., 2019]. However, most data augmentation techniques are based on heuristics, and their theoretical bases are limited to account for current successes.

In this thesis, we build grounds for regularized risk minimization with augmented data in the context of WDRO. When a loss function has a Hölder continuous gradient, we show that regularized empirical risk evaluated at augmented data approximates the worst-case risk (Theorem 3.3.1). While many existing results assume convexity of a loss function and limit a hypothesis space as a set of linear functions, we relax the assumption and allow deep

neural network model as hypothesis. We propose to minimize the approximation to solve WDRO, and we show that a minimizer of the approximation has risk consistency (Theorem 3.3.2). Based on the theoretical results, we propose a gradient-based algorithm producing a robust prediction model. Multiple real data experiments demonstrate robustness of the proposed model on noisy image classification datasets. This is the first rigorous method to use augmented data and deep neural networks in WDRO. As a summary:

- We build grounds for regularized risk minimization with augmented data in the context of WDRO.
- When a loss function has a Hölder continuous gradient, regularized empirical risk evaluated at augmented data approximates the worst-case risk. We further show that a risk consistency theorem.
- Our results relax convexity assumption on loss and allow deep neural network models that are not rigorously considered in WDRO literature.
- Real data experiments demonstrate robustness of the proposed model on noisy image classification datasets.

Chapter 2

Principled analytic classifier for PU learning via WIPM

In this chapter, we consider a problem of learning binary classifiers using only positive and unlabeled observations, called PU learning.

2.1 Preliminaries

In this section, we describe the L -risk for binary classification and present its PU representation. We briefly review several PU learning algorithms based on the L -risk minimization principle. We first introduce problem settings and notations.

2.1.1 Problem settings of PU learning

Let X and Y be random variables for input data and class labels, respectively, whose range is the product space $\mathcal{X} \times \{\pm 1\} \subseteq \mathbb{R}^d \times \{\pm 1\}$. The d is a positive integer. We denote the joint distribution of (X, Y) by $P_{X,Y}$ and the marginal distribution of X by

P_X . The distributions of positive and negative data are defined by conditional distributions, $P_{X|Y=1}$ and $P_{X|Y=-1}$, respectively. Let $\pi_+ := P_{X,Y}(Y = 1)$ be the marginal probability of being positive and set $\pi_- = 1 - \pi_+$. We follow the *two samples of data* scheme [Ward et al., 2009, Niu et al., 2016]. That is, let $\mathcal{X}_p = \{x_i^p\}_{i=1}^{n_p}$ and $\mathcal{X}_u = \{x_i^u\}_{i=1}^{n_u}$ be observed sets of independently identically distributed samples from the positive data distribution $P_{X|Y=1}$ and the marginal distribution P_X , respectively. Here, the n_p and n_u are the number of positive and unlabeled data points, respectively. Note that the unlabeled data distribution is the marginal distribution.

Let \mathcal{U} be a class of real-valued measurable functions defined on \mathcal{X} . A function $f \in \mathcal{U}$, often called a hypothesis, can be understood as a binary discriminant function and we classify an input x with the sign of a discriminant function, $\text{sign}(f(x))$. Define $\mathcal{M} = \{f : \mathcal{X} \rightarrow \mathbb{R} \mid \|f\|_\infty \leq 1\} \subseteq \mathcal{U}$, where $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ is the supremum norm. We restrict our attention to a class $\mathcal{F} \subseteq \mathcal{M}$ and call \mathcal{F} a hypothesis space. Throughout this paper, we assume that the hypothesis space is symmetric, *i.e.*, $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$. In PU learning, the main goal is to construct a classifier $\text{sign}(f(x))$ only from the positive dataset \mathcal{X}_p and the unlabeled dataset \mathcal{X}_u with $f \in \mathcal{F}$.

In this work, the quantity π_+ , often called the class-prior, is assumed to be known as in the literature [Kiryo et al., 2017, Kato et al., 2019] to focus on theoretical and practical benefits of our proposed algorithm.

2.1.2 L -risk minimization in PU learning

In supervised binary classification, the L -risk is defined by

$$\begin{aligned} R_L(f) &:= \int_{\mathcal{X} \times \{\pm 1\}} L(y, f(x)) dP_{X,Y}(x, y) \\ &= \pi_+ \int_{\mathcal{X}} L(1, f(x)) dP_{X|Y=1}(x) \\ &\quad + \pi_- \int_{\mathcal{X}} L(-1, f(x)) dP_{X|Y=-1}(x), \end{aligned} \quad (2.1)$$

for a loss function $L : \{\pm 1\} \times \mathbb{R} \rightarrow \mathbb{R}$ [Steinwart and Christmann, 2008, Section 2.1]. We denote the margin-based loss function by $\ell(yt) := L(y, t)$ if a loss function $L(y, t)$ can be represented as a function of margin yt , the product of a label y and a score t for all possible $y \in \{\pm 1\}$ and $t \in \mathbb{R}$.

Under the PU learning framework, however, the right-hand side of Equation (2.1) cannot be directly estimated due to lack of negatively labeled observations. To circumvent this problem, many studies in the field of PU learning exploited the relationship $P_X = \pi_+ P_{X|Y=1} + \pi_- P_{X|Y=-1}$ and replaced $P_{X|Y=-1}$ in Equation (2.1) with $(P_X - \pi_+ P_{X|Y=1})/\pi_-$ [du Plessis et al., 2014, Sakai et al., 2017]. That is, the L -risk can be alternatively expressed as:

$$\begin{aligned} R_L(f) &= \int_{\mathcal{X}} L(-1, f(x)) dP_X(x) \\ &\quad + \pi_+ \int_{\mathcal{X}} L(1, f(x)) - L(-1, f(x)) dP_{X|Y=1}(x). \end{aligned} \quad (2.2)$$

Now the right-hand side of Equation (2.2) can be empirically estimated by the positive dataset \mathcal{X}_p and the unlabeled dataset \mathcal{X}_u . However, the L -risk $R_L(f)$ is not convex with respect to f in general, and minimizing an empirical estimator for $R_L(f)$ is often formulated as a complicated non-convex optimization problem.

There have been several approaches to resolving the computational difficulty by modifying loss functions. du Plessis et al. [2014] proposed to use non-convex loss functions satisfying the symmetric condition, $L(1, f(x)) + L(-1, f(x)) = 1$. They proposed to optimize the empirical risk based on the ramp loss $\ell_{\text{ramp}}(yt) = 0.5 \times \max(0, \min(2, 1 - yt))$ via the concave-convex procedure [Collobert et al., 2006]. du Plessis et al. [2015] converted the problem to convex optimization through the linear-odd condition, $L(1, f(x)) - L(-1, f(x)) = -f(x)$. They showed that the logistic loss $\ell_{\text{log}}(yt) = \log(1 + \exp(-yt))$ and the double hinge loss $\ell_{\text{dh}}(yt) = \max(0, \max(-yt, (1 - yt)/2))$ satisfy the linear-odd condition. However, all the aforementioned methods utilized a weighted sum of $n_p + n_u$ pre-defined basis functions as a binary discriminant function, which triggered calculating the $(n_p + n_u) \times (n_p + n_u)$ Gram matrix. Hence, executing algorithms is not scalable and can be intractable when n_p and n_u are large [Sansone et al., 2019]. Our first goal is to overcome this computational problem by providing a computationally efficient method.

2.2 Weighted integral probability metric and L -risk

In this section, we formally define WIPM, a key tool for constructing the proposed algorithm, and build a link with the L -risk in Theorem 2.2.1 below. Based on the link, we propose a new binary discriminant function estimator and present its theoretical properties in Theorem 2.2.2. We first introduce the earlier work by

Sriperumbudur et al. [2012] that provided a closed-form classifier in supervised binary classification.

2.2.1 Relation between IPM and L -risk in supervised binary classification

Müller [1997] introduced an IPM for any two probability measures P and Q defined on \mathcal{X} and a class \mathcal{F} of bounded measurable functions, given by

$$\text{IPM}(P, Q; \mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x) dP(x) - \int_{\mathcal{X}} f(x) dQ(x) \right|.$$

IPM has been studied as either a metric between two probability measures [Sriperumbudur et al., 2010a, Arjovsky et al., 2017, Tolstikhin et al., 2018] or a hypothesis testing tool [Gretton et al., 2012].

Under the supervised binary classification setting, Sriperumbudur et al. [2012] showed that calculating IPM between $P_{X|Y=1}$ and $P_{X|Y=-1}$ is negatively related to minimizing the risk with a loss function, *i.e.*, $\text{IPM}(P_{X|Y=1}, P_{X|Y=-1}; \mathcal{F}) = -\inf_{f \in \mathcal{F}} R_{L_c}(f)$, where $L_c(1, t) = -t/\pi_+$ and $L_c(-1, t) = t/\pi_-$ for all $t \in \mathbb{R}$. They further showed that a discriminant function minimizing the L_c -risk can be obtained analytically when \mathcal{F} is a closed unit ball in RKHS. This result cannot be directly extended to PU learning due to absence of negatively labeled observations. In the next subsection, we define a generalized version of IPM and extend the previous results for supervised binary classification to PU learning.

2.2.2 Extension to WIPM and L -risk in PU learning

Let \mathcal{F} be a given class of bounded measurable functions and let $\tilde{w} : \mathcal{X} \rightarrow \mathbb{R}$ be a weight function such that $\|\tilde{w}\|_\infty < \infty$. We define WIPM¹ between two probability measures P and Q with a function class \mathcal{F} and a weight function \tilde{w} by

$$\text{WIPM}(P, Q; \tilde{w}, \mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x) dP(x) - \int_{\mathcal{X}} \tilde{w}(x) f(x) dQ(x) \right|. \quad (2.3)$$

Note that WIPM reduces to IPM if $\tilde{w}(x) = 1$ for all $x \in \mathcal{X}$. Other special cases of Equation (2.3) have been discussed in many applications. In the covariate shift problem, Huang et al. [2007] and Gretton et al. [2009] proposed to minimize WIPM with respect to \tilde{w} when \mathcal{F} is the unit ball in RKHS and P, Q are empirical distributions of test and training data, respectively. In unsupervised domain adaptation, Yan et al. [2017] regarded P, Q as empirical distributions of target and source data, respectively, where in this case, \tilde{w} is a ratio of two class-prior distributions.

We pay special attention to the case where $\tilde{w}(x)$ is constant, $w \in \mathbb{R}$, for every input value and denote WIPM by $\text{WIPM}(P, Q; w, \mathcal{F})$,

$$\text{WIPM}(P, Q; w, \mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{X}} f(x) dP(x) - w \int_{\mathcal{X}} f(x) dQ(x) \right|.$$

In the following theorem, we establish a link between $\text{WIPM}(P_X, P_{X|Y=1}; 2\pi_+, \mathcal{F})$ and the infimum of the ℓ_h -risk over \mathcal{F} for the hinge loss $\ell_h(yt) = \max(0, 1 - yt)$.

¹Although WIPM is not a metric in general, we keep saying the name WIPM to emphasize that it is a weighted version of IPM.

Theorem 2.2.1 (Relationship between ℓ_h -risk and WIPM). *Let \mathcal{F} be a symmetric hypothesis space in \mathcal{M} and $\ell_h(yt) = \max(0, 1 - yt)$ be the hinge loss. Then, we have*

$$\inf_{f \in \mathcal{F}} R_{\ell_h}(f) = 1 - \text{WIPM}(P_X, P_{X|Y=1}; 2\pi_+, \mathcal{F}).$$

Moreover, if $g_{\mathcal{F}}$ satisfies

$$\begin{aligned} & \text{WIPM}(P_X, P_{X|Y=1}; 2\pi_+, \mathcal{F}) \\ &= \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_X(x) - 2\pi_+ \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_{X|Y=1}(x), \end{aligned}$$

then $\inf_{f \in \mathcal{F}} R_{\ell_h}(f) = R_{\ell_h}(-g_{\mathcal{F}})$.

A proof is available in Section 2.7.1. Theorem 2.2.1 shows that the infimum of the ℓ_h -risk over a hypothesis space \mathcal{F} equals the negative WIPM between the unlabeled data distribution P_X and the positive data distribution $P_{X|Y=1}$ with the same hypothesis space \mathcal{F} and the weight $2\pi_+$ up to addition by constant. Furthermore, by negating the WIPM optimizer $g_{\mathcal{F}}$, we obtain the minimizer of the ℓ_h -risk over the hypothesis space \mathcal{F} . Here, we define a WIPM optimizer $g_{\mathcal{F}}$ as a function that attains the supremum, *i.e.*, $\text{WIPM}(P_X, P_{X|Y=1}; 2\pi_+, \mathcal{F}) = \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_X(x) - 2\pi_+ \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_{X|Y=1}(x)$ and we set $f_{\mathcal{F}} = -g_{\mathcal{F}}$ for later notational convenience. Sriperumbudur et al. [2012] derived a similar result to Theorem 2.2.1 by showing $\text{IPM}(P_{X|Y=1}, P_{X|Y=-1}; \mathcal{F}) = -\inf_{f \in \mathcal{F}} R_{L_c}(f)$ in supervised binary classification. However, as we mentioned in Section 2.2.1, their method is only applicable to supervised binary classification settings.

2.2.3 Theoretical properties of empirical WIPM optimizer

We denote the empirical distributions of $P_{X|Y=1}$ and P_X by $P_{X|Y=1, n_p}$ and P_{X, n_u} , respectively. Let $P_{X|Y=1, n_p} = n_p^{-1} \sum_{i=1}^{n_p} \delta_{x_i^p}$ and $P_{X, n_u} = n_u^{-1} \sum_{i=1}^{n_u} \delta_{x_i^u}$, where $\delta(\cdot)$ defined on \mathcal{X} is the Dirac delta function and $\delta_x(\cdot) := \delta(\cdot - x)$ for $x \in \mathcal{X}$. The empirical Rademacher complexity of \mathcal{F} given a set $S = \{z_1, \dots, z_m\}$ is defined by $\mathfrak{R}_S(\mathcal{F}) := \mathbb{E}_\sigma \left(\frac{1}{m} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^m \sigma_i f(z_i) \right| \right)$. Here, $\{\sigma_i\}_{i=1}^m$ is a set of independent Rademacher random variables taking 1 or -1 with probability 0.5 each and $\mathbb{E}_\sigma(\cdot)$ is the expectation operator over the Rademacher random variables [Bartlett and Mendelson, 2002]. Denote a maximum by $a \vee b := \max(a, b)$, a minimum by $a \wedge b := \min(a, b)$. For a probability measure Q defined on \mathcal{X} , denote the expectation of a discriminant function f by $\mathbb{E}_Q(f) := \int_{\mathcal{X}} f(x) dQ(x)$ and the variance by $\text{Var}_Q(f) := \mathbb{E}_Q(f^2) - (\mathbb{E}_Q(f))^2$.

The empirical estimator for $\text{WIPM}(P_X, P_{X|Y=1}; w, \mathcal{F})$ is given by plugging in the empirical distributions,

$$\text{WIPM}(P_{X, n_u}, P_{X|Y=1, n_p}; w, \mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n_u} \sum_{i=1}^{n_u} f(x_i^u) - \frac{w}{n_p} \sum_{i=1}^{n_p} f(x_i^p) \right|,$$

and we define an empirical WIPM optimizer $\hat{g}_{\mathcal{F}} \in \mathcal{F}$ that satisfies the following equation,

$$\text{WIPM}(P_{X, n_u}, P_{X|Y=1, n_p}; w, \mathcal{F}) = \frac{1}{n_u} \sum_{i=1}^{n_u} \hat{g}_{\mathcal{F}}(x_i^u) - \frac{w}{n_p} \sum_{i=1}^{n_p} \hat{g}_{\mathcal{F}}(x_i^p). \quad (2.4)$$

We set $\hat{f}_{\mathcal{F}} = -\hat{g}_{\mathcal{F}}$ for notational convenience as in Section 2.2.2.

We analyze the estimation error $R_{\ell_h}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_{\ell_h}(f)$ in

the following theorem. To begin, let $\chi_{n_p, n_u}^{(1)}(w) = w/\sqrt{n_p} + 1/\sqrt{n_u}$ and $\chi_{n_p, n_u}^{(2)}(w) = 2(w/n_p + 1/n_u)$.

Theorem 2.2.2 (Estimation error bound for general function space). *Let $\hat{g}_{\mathcal{F}}$ be an empirical WIPM optimizer defined in Equation (2.4) and set $\hat{f}_{\mathcal{F}} = -\hat{g}_{\mathcal{F}}$. Let \mathcal{F} be a symmetric hypothesis space such that $\|f\|_{\infty} \leq \nu \leq 1$, $\text{Var}_{P_{X|Y=1}}(f) \leq \sigma_{X|Y=1}^2$, and $\text{Var}_{P_X}(f) \leq \sigma_X^2$. Denote $\rho^2 = \sigma_{X|Y=1}^2 \vee \sigma_X^2$. Then, for all $\alpha, \tau > 0$, the following holds with probability at least $1 - e^{-\tau}$,*

$$\begin{aligned} & R_{\ell_h}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_{\ell_h}(f) \\ & \leq C_{\alpha}(\mathbb{E}_{P_X^{n_u}}(\mathfrak{R}_{\mathcal{X}_u}(\mathcal{F})) + 2\pi_+ \mathbb{E}_{P_{X|Y=1}^{n_p}}(\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F}))) \quad (2.5) \\ & + C_{\tau, \rho^2}^{(1)} \chi_{n_p, n_u}^{(1)}(2\pi_+) + C_{\tau, \nu, \alpha}^{(2)} \chi_{n_p, n_u}^{(2)}(2\pi_+), \end{aligned}$$

where $C_{\alpha} = 4(1 + \alpha)$, $C_{\tau, \rho^2}^{(1)} = 2\sqrt{2\tau\rho^2}$, $C_{\tau, \nu, \alpha}^{(2)} = 2\tau\nu \left(\frac{2}{3} + \frac{1}{\alpha}\right)$.

A proof is provided in Section 2.7.2. Due to Talagrand's inequality, Theorem 2.2.2 provides a sharper bound than the existing result based on McDiarmid's inequality. Specifically, Kiryo et al. [2017, Theorem 4] utilized McDiarmid's inequality and showed that for $\tau > 0$ and some $\Delta > 0$ the following holds with probability at least $1 - e^{-\tau}$,

$$\begin{aligned} R_{\ell_h}(\hat{f}) - \inf_{f \in \mathcal{F}} R_{\ell_h}(f) & \leq 8(\mathbb{E}_{P_X^{n_u}}(\mathfrak{R}_{\mathcal{X}_u}(\mathcal{F})) + 2\pi_+ \mathbb{E}_{P_{X|Y=1}^{n_p}}(\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F}))) \\ & + \chi_{n_p, n_u}^{(1)}(2\pi_+)(1 + \nu)\sqrt{2\tau} + \Delta. \quad (2.6) \end{aligned}$$

The following proposition shows that the proposed upper bound (2.5) is sharper than the upper bound (2.6) under a certain condition. A proof is provided in Section 2.7.2.

Proposition 2.2.3. *With the notations defined in Theorem 2.2.2, suppose that the following holds,*

$$\frac{1 + \nu}{2} - \frac{5\sqrt{2\tau}\chi_{n_p, n_u}^{(2)}(2\pi_+)\nu}{6\chi_{n_p, n_u}^{(1)}(2\pi_+)} \geq \rho. \quad (2.7)$$

Then, the proposed upper bound (2.5) is sharper than the previous result (2.6) proposed by Kiryo et al. [2017].

It is noteworthy that the second term in the left-hand side of (2.7) converges to zero as n_p and n_u increase because $\chi_{n_p, n_u}^{(1)}(2\pi_+) = O_{P_{X|Y=1}, P_X}((n_p \wedge n_u)^{-1/2})$ and $\chi_{n_p, n_u}^{(2)}(2\pi_+) = O_{P_{X|Y=1}, P_X}((n_p \wedge n_u)^{-1})$. Due to $(1 + \nu)/2 \geq \nu \geq \rho$, the condition (2.7) is quite reasonable if the upper bounds of the variances, σ_X^2 and $\sigma_{X|Y=1}^2$, are sufficiently small.

In binary classification, one ultimate goal is to find a classifier minimizing the misclassification error, or equivalently, minimizing the excess risk. Bartlett et al. [2006] showed that there is an invertible function $\psi : [-1, 1] \rightarrow [0, \infty)$ such that the excess risk $R_{\ell_{01}}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{U}} R_{\ell_{01}}(f)$ is bounded above by $\psi^{-1}(R_{\ell}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{U}} R_{\ell}(f))$ if the margin-based loss ℓ is classification-calibrated. In particular, Zhang [2004] showed that the excess risk is bounded above by the excess ℓ_h -risk, *i.e.*, $R_{\ell_{01}}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{U}} R_{\ell_{01}}(f) \leq R_{\ell_h}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{U}} R_{\ell_h}(f)$. This implies that an excess risk bound can be obtained by analyzing the excess ℓ_h -risk bound with Theorem 2.2.2. The following corollary provides the excess risk bound.

Corollary 2.2.4 (Excess risk bound for general function space).

With the notations defined in Theorem 2.2.2, for all $\alpha, \tau > 0$, the following holds with probability at least $1 - e^{-\tau}$,

$$R_{\ell_{01}}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{U}} R_{\ell_{01}}(f)$$

$$\begin{aligned}
&\leq \inf_{f \in \mathcal{F}} R_{\ell_h}(f) - \inf_{f \in \mathcal{U}} R_{\ell_h}(f) \\
&+ C_\alpha(\mathbb{E}_{P_X^{n_u}}(\mathfrak{R}_{\mathcal{X}_u}(\mathcal{F})) + 2\pi_+ \mathbb{E}_{P_{X|Y=1}^{n_p}}(\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F}))) \\
&+ C_{\tau, \rho^2}^{(1)} \chi_{n_p, n_u}^{(1)}(2\pi_+) + C_{\tau, \nu, \alpha}^{(2)} \chi_{n_p, n_u}^{(2)}(2\pi_+).
\end{aligned}$$

2.3 WIPM optimizer with reproducing kernel Hilbert space

In this section, we provide a computationally efficient PU learning algorithm which builds an analytic classifier when a hypothesis space is a closed ball in RKHS. In addition, unlike the excess risk bound in Corollary 2.2.4, we explicitly derive the bound that converges to zero when the sample sizes n_p and n_u increase.

2.3.1 An analytic classifier via WMMD optimizer

To this end, we assume that $\mathcal{X} \subseteq [0, 1]^d$ is compact. Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a reproducing kernel defined on \mathcal{X} and \mathcal{H}_k be the associated RKHS with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k} : \mathcal{H}_k \times \mathcal{H}_k \rightarrow \mathbb{R}$. We denote the induced norm by $\|\cdot\|_{\mathcal{H}_k}$. Denote a closed ball in RKHS \mathcal{H}_k with a radius $r > 0$, by $\mathcal{H}_{k,r} = \{f : \|f\|_{\mathcal{H}_k} \leq r\}$. We define the weighted maximum mean discrepancy (WMMD) between two probability measures P and Q with a weight w and a closed ball $\mathcal{H}_{k,r}$ by $\text{WMMD}_k(P, Q; w, r) := \text{WIPM}(P, Q; w, \mathcal{H}_{k,r})$. The name of WMMD comes from the maximum mean discrepancy (MMD), a popular example of the IPM whose function space is the unit ball $\mathcal{H}_{k,1}$, *i.e.*, $\text{MMD}_k(P, Q) := \text{IPM}(P, Q; \mathcal{H}_{k,1})$ [Sriperumbudur

et al., 2010a,b]. As defined in Equation (2.4), let $\hat{g}_{\mathcal{H}_{k,r}} \in \mathcal{H}_{k,r}$ be the empirical WMMD optimizer such that

$$\text{WMMD}_k(P_{X,n_u}, P_{X|Y=1,n_p}; w, r) = \frac{1}{n_u} \sum_{i=1}^{n_u} \hat{g}_{\mathcal{H}_{k,r}}(x_i^u) - \frac{w}{n_p} \sum_{i=1}^{n_p} \hat{g}_{\mathcal{H}_{k,r}}(x_i^p).$$

In addition, we set $\hat{f}_{\mathcal{H}_{k,r}} = -\hat{g}_{\mathcal{H}_{k,r}}$, which leads the corresponding classification rule to $\text{sign}(\hat{f}_{\mathcal{H}_{k,r}}(z))$. In the following proposition, we show that this classification rule has an analytic expression by exploiting the reproducing property $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$ and the Cauchy-Schwarz inequality.

Proposition 2.3.1. *Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a bounded reproducing kernel. Then, the classification rule has a closed-form expression given by*

$$\text{sign}(\hat{f}_{\mathcal{H}_{k,r}}(z)) = \begin{cases} +1 & \text{if } (2\pi_+)^{-1} < \hat{\lambda}_{n_p, n_u}(z), \\ -1 & \text{otherwise,} \end{cases} \quad (2.8)$$

where

$$\hat{\lambda}_{n_p, n_u}(z) = \frac{n_p^{-1} \sum_{i=1}^{n_p} k(z, x_i^p)}{n_u^{-1} \sum_{i=1}^{n_u} k(z, x_i^u)}.$$

We call the classifier defined in Equation (2.8) the *WMMD classifier* and the score $\hat{\lambda}_{n_p, n_u}(z)$ the *WMMD score* for z . One strength of the WMMD classifier is that the classification rule has a closed-form expression, resulting in computational efficiency. Furthermore, the WMMD score $\hat{\lambda}_{n_p, n_u}$ is independent of the class-prior π_+ , and thus we can obtain the score function without prior knowledge of the class-prior.

2.3.2 Explicit excess risk bound of WMMD classifier

Since the empirical WMMD optimizer $\hat{g}_{\mathcal{H}_{k,r}}$ is a special case of the empirical WIPM optimizer, we have an excess risk bound from the result of Corollary 2.2.4. However, without knowing convergence rates of the Rademacher complexities, $\mathbb{E}_{P_X^{n_u}}(\mathfrak{R}_{\mathcal{X}_u}(\mathcal{F}))$ and $\mathbb{E}_{P_{X|Y=1}^{n_p}}(\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F}))$, and the approximation error, the consistency of the classifier remains unclear. In this subsection, we establish an explicit excess risk bound that vanishes. We first derive an explicit estimation error bound in the following proposition.

Proposition 2.3.2 (Explicit estimation error bound). *With the notations defined in Theorem 2.2.2, assume that a reproducing kernel k defined on a compact space \mathcal{X} is bounded. Let $r_1^{-1} = \sup_{x \in \mathcal{X}} \sqrt{k(x,x)}$. Then, we have $\mathcal{H}_{k,r_1} \subseteq \mathcal{M}$. Moreover, for all $\alpha, \tau > 0$, the following holds with probability at least $1 - e^{-\tau}$,*

$$\begin{aligned} R_{\ell_h}(\hat{f}_{\mathcal{H}_{k,r_1}}) - \inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_h}(f) \\ \leq (C_\alpha + C_{\tau,\rho^2}^{(1)}) \chi_{n_p, n_u}^{(1)}(2\pi_+) + C_{\tau,\nu,\alpha}^{(2)} \chi_{n_p, n_u}^{(2)}(2\pi_+). \end{aligned}$$

While the bound in Theorem 2.2.2 is expressed in terms $\mathbb{E}_{P_X^{n_u}}(\mathfrak{R}_{\mathcal{X}_u}(\mathcal{F}))$ and $\mathbb{E}_{P_{X|Y=1}^{n_p}}(\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F}))$, these are evaluated in terms of n_p and n_u in the upper bound in Proposition 2.3.2, giving an explicit estimation error bound with $O((n_p \wedge n_u)^{-1/2})$ convergence rate. The key idea is to use reproducing property $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$ and the Cauchy-Schwarz inequality to obtain an upper bound for the Rademacher complexity. Detailed proofs are given in Section 2.7.3.

In the following proposition, we elaborate on the approximation error bound. To begin, for any $0 < \beta \leq 1$, let $\beta\mathcal{M} := \{\beta f : f \in \mathcal{M}\}$. Set $f_1^*(x) = \text{sign}(P(Y = 1 | X = x) - \frac{1}{2})$.

Proposition 2.3.3 (Approximation error bound over uniformly bounded hypothesis space). *With the notations defined in Proposition 2.3.2, we have*

$$\inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_h}(f) - \inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f) \leq \beta \inf_{g \in \mathcal{H}_{k,r_1/\beta}} \|g - f_1^*\|_{L_2(P_X)},$$

for any $0 < \beta \leq 1$.

When $\beta = 1$, Proposition 2.3.3 implies that the approximation error $\inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_h}(f) - \inf_{f \in \mathcal{U}} R_{\ell_h}(f)$ is bounded above by $\inf_{g \in \mathcal{H}_{k,r_1}} \|g - f_1^*\|_{L_2(P_X)}$ due to $\inf_{f \in \mathcal{U}} R_{\ell_h}(f) = \inf_{f \in \mathcal{M}} R_{\ell_h}(f)$ [Lin, 2002]. Hence, a naive substitution to Corollary 2.2.4 will give a sub-optimal bound because $\inf_{g \in \mathcal{H}_{k,r_1}} \|g - f_1^*\|_{L_2(P_X)}$ is non-zero in general.

In the following theorem, we rigorously establish the explicit excess risk bound which vanishes as n_p and n_u increase. To begin, we state the following assumptions.

- (A1) The distribution functions P_X and $P_{X|Y=1}$ have probability density functions $p_X(x)$ and $p_{X|Y=1}(x)$, respectively.
- (A2) The density functions $p_X(x)$ and $p_{X|Y=1}(x)$ are α_H -Hölder continuous.
- (A3) The marginal density function is bounded away from zero: $p_X(x) \geq p_{\min} > 0$ for all x on its support.

(A4) The marginal distribution P_X has Tsybakov's noise exponent $q \in [0, \infty)$, *i.e.*, there exists a constant $C_{\text{noise}} > 0$ such that for all sufficiently small $t > 0$, we have

$$P_X(\{x \in \mathcal{X} \mid |2\eta(x) - 1| \leq t\}) \leq C_{\text{noise}} t^q,$$

where $\eta(x) = P(Y = 1 \mid X = x)$.

Theorem 2.3.4. *Assume that the Gaussian kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{2h^2})$ is used. Under the assumptions (A1)-(A4), we have the following holds with probability at least $1 - 1/n_p - 1/n_u$.*

$$R_{\ell_{01}}(\hat{f}_{\mathcal{H}_{k,1}}) - \inf_{f \in \mathcal{U}} R_{\ell_{01}}(f) \leq \tilde{C}(n_p \wedge n_u)^{-\frac{\alpha_H(1+q)}{2\alpha_H+d}},$$

for some constant $\tilde{C} > 0$ and some bandwidth $h = (n_p \wedge n_u)^{-\frac{1}{2\alpha_H+d}}$.

A proof is given in Section 2.7.3. In supervised binary classification settings, a similar result is obtained by Audibert et al. [2007, Theorem 3.3] and the convergence rate is called *a super-fast rate* when $\alpha_H q > d$. However, α_H and q cannot be simultaneously very large. For more information, see a detailed discussion [Audibert et al., 2007].

2.4 Related work

Excess risk bound in noisy label literature: PU learning can be considered as a special case of classification with asymmetric label noise, and many studies in this literature have shown consistency results [Natarajan et al., 2013]. Patrini et al. [2016] derived

an explicit estimation error when \mathcal{F} is a set of linear hypotheses and Blanchard et al. [2016] showed a consistency result of the excess risk bound when the hypothesis space is RKHS with universal kernels. While the two studies assumed the *one sample of data* scheme, the proposed bound is based on the *two samples of data* scheme. Therefore, our proposed excess risk bound is expressed in n_p and n_u , giving a new consistency theory.

Closed-form classifier: Blanchard et al. [2010] suggested a score function similar to the WMMD score by using different bandwidth hyperparameters for the denominator and the numerator. However, with these differences, our method gains theoretical justification while their score function does not. du Plessis et al. [2015] derived a closed-form classifier based on the squared loss. They estimated $P(Y = 1 | X) - P(Y = -1 | X)$ and showed the consistency of the estimation error bound in the *two samples of data* scheme. However, the classifier is not scalable because it requires to compute the inverse of a $(n_p + n_u) \times (n_p + n_u)$ matrix.

2.5 Numerical experiments

In this section, we empirically analyze the proposed algorithm to demonstrate its practical efficacy using synthetic and real datasets. implementation details are available in Section 2.7.4. Pytorch implementation for the experiments is available at https://github.com/eraser347/WMMD_PU.

Synthetic data analysis We first visualize the effect of increasing the sample sizes n_p and n_u on the discriminant ability of the proposed algorithm (Experiment 1). Then we compare per-

formance with (i) the logistic loss ℓ_{\log} , denoted by LOG, (ii) the double hinge loss ℓ_{dh} , denoted by DH, both proposed by du Plessis et al. [2015], (iii) the non-negative risk estimator method, denoted by NNPU, proposed by Kiryo et al. [2017], (iv) the threshold adjustment method, denoted by tADJ, proposed by Elkan and Noto [2008], and (v) the proposed algorithm, denoted by WMMD (Experiments 2, 3, and 4).

Experiment 1 In this case, we used the `two_moons` dataset, the underlying distributions of which are

$$X|Y = y, U \sim N \left(\begin{bmatrix} 2(1+y) - 4y \cos(\pi U) \\ (1+y) - 4y \sin(\pi U) \end{bmatrix}, \begin{bmatrix} 0.4^2 & 0 \\ 0 & 0.4^2 \end{bmatrix} \right),$$

where U refers to the uniform random variable ranges from 0 to 1 and $N(\mu, \Sigma)$ is the normal distribution with mean μ and covariance Σ . We used the ‘`make_moons`’ function in the Python module ‘`sklearn.datasets`’ [Pedregosa et al., 2011] to generate the datasets.

Fig. 2.1 illustrates the decision boundaries of WMMD using the `two_moons` dataset. The first row displays the case where the unlabeled sample size is small, $n_u = 50$ and the second row displays the case where the unlabeled sample size is large, $n_u = 400$. The first and second columns display the case where the positive sample sizes are $n_p = 5$ and $n_p = 10$, respectively. The class-prior is fixed to $\pi_+ = 0.5$, and we assumed that the class-prior is known. We visualize the true mean function of the positive and negative data distributions with blue and red lines, respectively. The positive data are represented by blue diamond points, and the unlabeled data are represented by gray points. The decision boundaries of WMMD classifier tend to correctly separate the two clusters as n_p

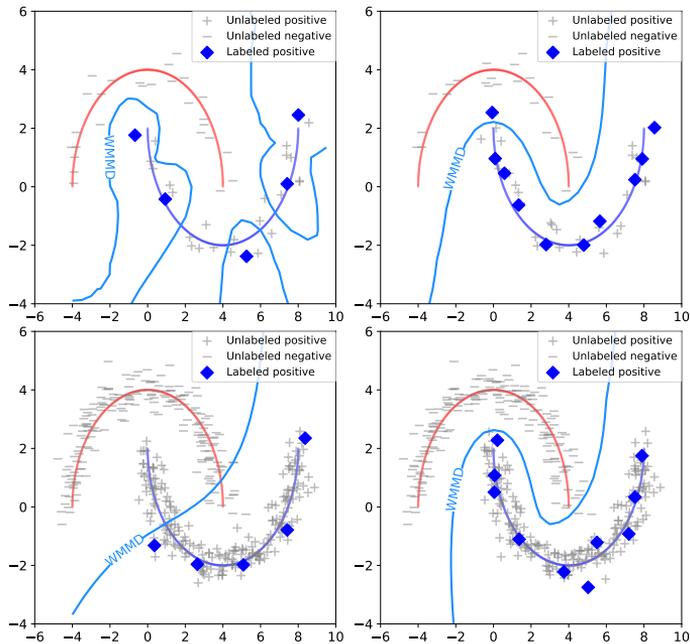


Figure 2.1: The illustration of the decision boundaries of WMMD classifier using `two_moons` dataset with the increases in the size of the positive and unlabeled samples. The true mean of positive and negative data distribution is plotted by blue and red lines respectively. The gray ‘+’ points and the gray ‘-’ points refer the unlabeled positive and unlabeled negative training data, respectively.

and n_u increase.

In Experiments 2, 3, and 4, we evaluate: (i) accuracy and area under the receiver operating characteristic curve (AUC) as n_u and π_+ change when the class-prior is known (Experiment 2) and unknown (Experiment 3); (ii) the elapsed training time (Experiment 4). In these experiments, we set up the underlying joint distribu-

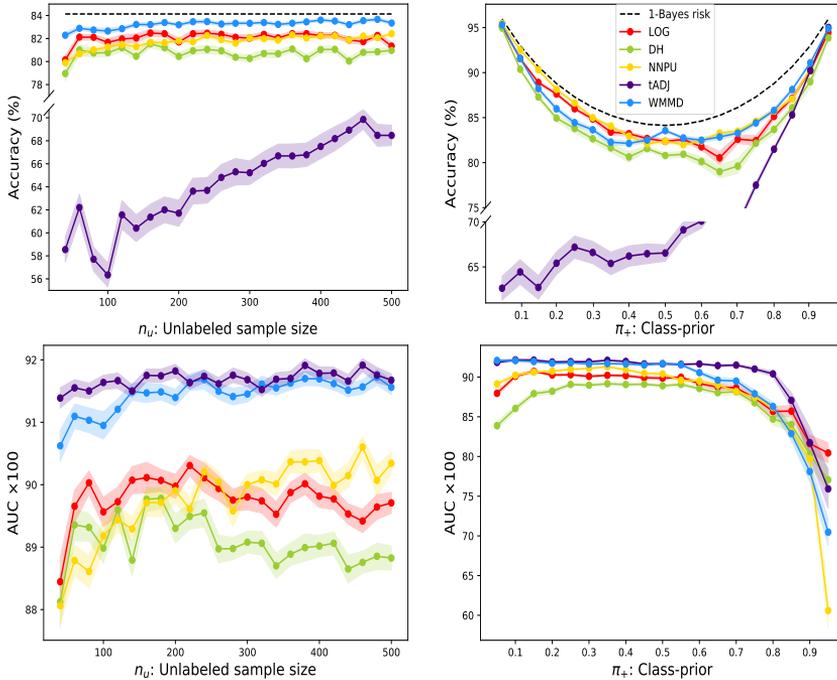


Figure 2.2: The comparison of accuracy and AUC of the five PU learning algorithms when each n_u and π_+ changes. The dashed curve represents the 1–Bayes risk. The curve and the shaded region represent the average and the standard error, respectively, based on the 100 replications.

tion as follows:

$$X \mid Y = y \sim N\left(y \frac{\mathbf{1}_2}{\sqrt{2}}, I_2\right), Y \sim 2 \times \text{Bern}(\pi_+) - 1, \quad (2.9)$$

where $\text{Bern}(p)$ is the Bernoulli distribution with mean p , $\mathbf{1}_2 = (1, 1)^T$ is the 2 dimensional vector of all ones and I_2 is the identity matrix of size 2.

Experiment 2 In this experiment, we compare the accuracy and AUC of the five PU learning algorithms when the true class-

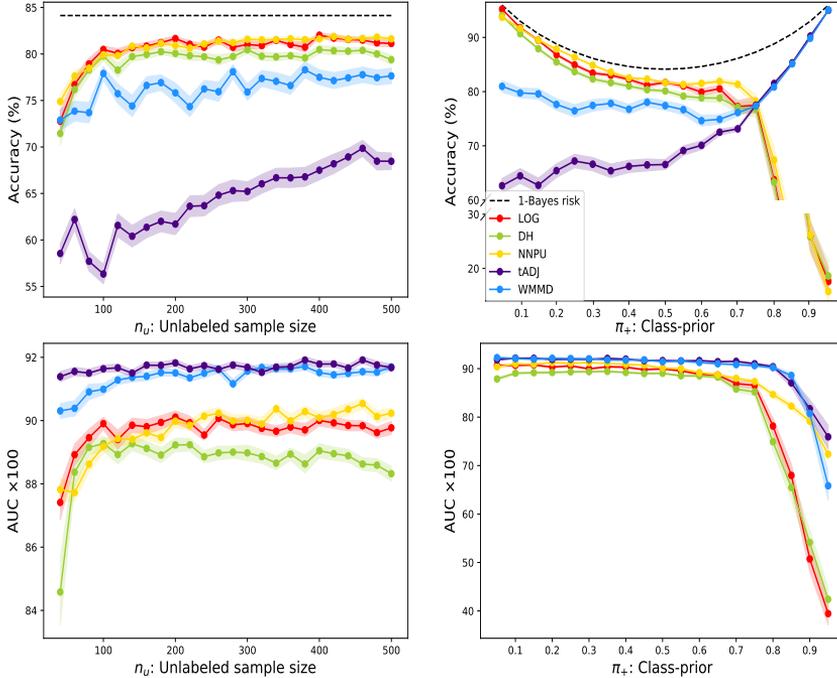


Figure 2.3: The comparison of accuracy and AUC of the five PU learning algorithms when each n_u and π_+ changes in case where the class-prior is unknown. The dashed curve represents the 1–Bayes risk. The curve and the shaded region represent the average and the standard error, respectively, based on the 100 replications. LOG, DH, and NNPU use the estimate of the class-prior from the ‘KM1’ method. Other details are given in Fig. 2.2.

prior π_+ is known. Fig. 2.2 shows the accuracy and AUC on various n_u . The training sample size for the positive data is $n_p = 100$ and the class prior is $\pi_+ = 0.5$. The unlabeled sample size changes from 40 to 500 by 20. We repeat a random generation of training and test data 100 times. For comparison purposes, we add 1–Bayes risk for each unlabeled sample size. In terms of accu-

racy, the proposed WMMD tends to be closer to 1–Bayes risk as the n_u increases. Compared with other PU learning algorithms, WMMD achieves higher accuracy in every n_u and achieves comparable to or better AUC. Also, Fig. 2.2 shows a comparison of accuracy and AUC as π_+ changes. The training sample size for the positive and unlabeled data are $n_p = 100$ and $n_u = 400$, respectively. The class-prior π_+ changes from 0.05 to 0.95 by 0.05. The test sample size is 10^3 . Training and test data are repeatedly generated 100 times. In terms of accuracy, the proposed WMMD performs comparably with LOG and NNPU, showing advantages over DH and tADJ. When the true class-prior is less than equal to 0.8, WMMD performs better in terms of AUC, except for tADJ. The tADJ achieves the highest AUC because $\eta(x)$ is proportional to $P(\{x \text{ is from the positive dataset}\} \mid X = x)$. This empirically shows that WMMD has a comparable discriminant ability to the other algorithms for a wide range of class-prior.

Experiment 3 The main goal of this subsection is to show the robustness of the proposed classifier in the case of unknown class-prior π_+ . In PU learning literature, the π_+ has been frequently assumed to be known [du Plessis et al., 2015, Niu et al., 2016, Kiryo et al., 2017, Kato et al., 2019]. However, this assumption can be considered to be strong in real-world applications, and to correctly execute existing PU learning algorithms, an accurate estimation of the π_+ is necessary. In this experiment, we compare the accuracy and AUC in cases where the class-prior π_+ is unknown. For the WMMD classifier, we used a density-based method for the class-prior estimation described in Section 2.7.4, which can be ob-

tained as a byproduct of the proposed algorithm. The results of LOG, DH, and NNPU are given for completeness sake using the ‘KM1’ method² by Ramaswamy et al. [2016]. We take these estimates as true values and repeat the same comparative numerical experiments in Experiment 2.

Since the objective functions of LOG, DH, and NNPU algorithms depend on the estimate $\hat{\pi}_+$, we anticipate that both the accuracy and AUC rely on the quality of the estimation. On the other hand, the tADJ algorithm does not depend on the class-prior, so the performance is not affected. Also, as the proposed score function does not depend on the class-prior π_+ and π_+ is used to determine a cutoff, the AUC of the proposed algorithm is less affected by the estimation of π_+ .

Fig. 2.3 compares accuracy and AUC as a function of n_u . WMMD performs worse than LOG, DH, and NNPU, while AUC is higher. Though tADJ shows poor accuracy in a wide range, it achieves high AUC comparable to WMMD. As we anticipated, WMMD is more robust than LOG, DH, and NNPU in AUC. This is possibly because our score function $\hat{\lambda}_{n_p, n_u}$ does not depend on π_+ . A similar trend can be found when π_+ changes. We note that the ‘KM1’ method is not scalable and thus may not be used for large-scale datasets.

Experiment 4 In this experiment, we compare the elapsed training time, including hyperparameter optimization, of the five

²While the ‘KM2’ method by Ramaswamy et al. [2016] is often considered to be a state-of-the-art method for estimating π_+ , in our experiments, estimates based on the ‘KM2’ have a larger estimation error than that of the ‘KM1’ method.

Table 2.1: A summary of elapsed training time and its ratio for the five PU learning algorithms based on 100 replications. We set $n_p = 100, n_u = 400$, and $\pi_+ = 0.5$. Average and standard error are denoted by ‘average \pm standard error’.

| | LOG | DH | NNPU | tADJ | WMMD |
|------------------------|------------------|------------------|----------------|---------------|---------------|
| in seconds $\times 10$ | 90.0 ± 4.7 | 96.1 ± 6.1 | 6.0 ± 0.1 | 0.4 ± 0.0 | 0.2 ± 0.0 |
| in ratio | 347.9 ± 23.4 | 371.0 ± 28.5 | 23.2 ± 1.1 | 1.8 ± 0.0 | 1.0 ± 0.0 |

PU learning algorithms. The data are generated from the distributions described in Equation (2.9), and we set $n_p = 100, n_u = 400$, and $\pi_+ = 0.5$. The elapsed time is measured with 20 Intel[®] Xeon[®] E5-2630 v4@2.20GHz CPU processors.

Table 2.1 compares the elapsed training time and its ratio relative to that of WMMD. WMMD takes the shortest time among the five baseline methods. In particular, the training time for WMMD is at least about 300 times shorter than that of LOG and DH methods. This is because the WMMD classifier has an analytic form while the LOG and DH methods require solving a quadratic programming problem.

Real data analysis We demonstrate the practical utility of the proposed algorithm using the eight real binary classification datasets from the LIBSVM³ [Chang and Lin, 2011]. Since some observations from the raw datasets are not completely recorded, we remove such observations and construct the dataset with fully recorded data. Next, to investigate the effect of varying π_+ , we artificially reconstruct \mathcal{X}_p and \mathcal{X}_u through a random sampling from the

³<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

Table 2.2: A summary of the eleven binary classification datasets. ‘# of samples’ denotes the number of total samples after removing incomplete observations. We denote the number of positive, unlabeled, and test samples, by n_p , n_u , and n_{te} after the random sampling, respectively. We categorize the eleven datasets into two groups: the first seven datasets as small-scale and the last four datasets as large-scale.

| Dataset | d | # of samples | n_p | n_u | n_{te} | π_+ | Scale |
|----------------------|-------|--------------|--------|-----------------|----------|---------|-------|
| heart_scale | 12 | 122 | 10 | 60 | 60 | 0.62 | Small |
| sonar_scale | 60 | 207 | 10 | 100 | 100 | 0.47 | Small |
| australian_scale | 12 | 449 | 20 | 220 | 220 | 0.51 | Small |
| australian_scale2 | 12 | 449 | 10 | 130 | 130 | 0.15 | Small |
| breast-cancer_scale | 10 | 683 | 20 | 340 | 340 | 0.35 | Small |
| breast-cancer_scale2 | 10 | 683 | 40 | 340 | 340 | 0.65 | Small |
| diabetes_scale | 8 | 759 | 50 | 380 | 370 | 0.65 | Small |
| skin_nonskin | 3 | 245,057 | 10^3 | 10^5 | 10^5 | 0.79 | Large |
| skin_nonskin2 | 3 | 245,057 | 10^3 | 10^5 | 10^5 | 0.21 | Large |
| epsilon_normalized | 2,000 | 500,000 | 10^3 | 4×10^5 | 10^5 | 0.50 | Large |
| HIGGS | 26 | 8,786,441 | 10^3 | 10^6 | 10^5 | 0.50 | Large |

fully recorded datasets. For the three datasets `australian_scale`, `breast-cancer_scale` and `skin_nonskin`, we reconstruct the data so that the resulting class-prior π_+ ranges from 0.15 to 0.79. We add the suffix 2 for those datasets. We sample 100 times for the seven small datasets and 10 times for the four big datasets: `skin_nonskin`, `skin_nonskin2`, `epsilon_normalized`, and `HIGGS`. Table 2.2 summarizes statistics for the eleven real datasets. We conduct two comparative numerical experiments when π_+ is known and unknown.

Table 2.3 shows the average and the standard error of accuracy and AUC when the class-prior π_+ is known. LOG and DH fail to compute the $(n_p + n_u) \times (n_p + n_u)$ Gram matrix due to out of memory in the 12 GB GPU memory limit. WMMD achieves comparable to or better accuracy and AUC than LOG, DH, and tADJ on most datasets. Compared to NNPU, WMMD performs comparably on the small datasets. However, NNPU achieves higher accuracy on `skin_nonskin`, `epsilon_normalized`, and `HIGGS`. The neural network used in NNPU fits well to the complicated and high-dimensional structure of data and shows high accuracy.

Table 2.4 compares the average and the standard error of accuracy and AUC in cases where the class-prior π_+ is unknown. As in Experiment 3 in the previous section, we estimate π_+ using the ‘KM1’ for LOG, DH, and NNPU, and using the density-based method for WMMD. LOG, DH, and NNPU algorithms are implemented on the seven small-scale datasets alone because the method by Ramaswamy et al. [2016] is not feasible with the large-scale datasets [Bekker and Davis, 2018]. Overall, WMMD shows comparable to or better performances than other PU learning algorithms on most datasets. Compared to Table 2.3, WMMD and tADJ show robustness to unknown π_+ in terms of AUC. This is because WMMD and tADJ do not require estimation of π_+ to construct score functions. In contrast, the other methods require an estimate $\hat{\pi}_+$, and we observe a substantial drop in accuracy and AUC when the ‘KM1’ estimate is used.

Table 2.3: Accuracy and AUC comparison using the real datasets in cases the class-prior π_+ is known. We denote the memory error results for LOG and DH by the hyphen. Average and standard error are denoted by ‘average \pm standard error’. Boldface numbers denote the best and equivalent algorithms with respect to a t-test with a significance level of 5%.

| Dataset | LOG | DH | NNPU | tADJ | WMMD |
|----------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| Accuracy (in %) | | | | | |
| heart_scale | 70.5 \pm 0.8 | 68.4 \pm 0.9 | 71.0 \pm 0.8 | 65.1 \pm 0.8 | 71.6 \pm 0.8 |
| sonar_scale | 55.8 \pm 0.6 | 52.9 \pm 0.6 | 63.2 \pm 0.6 | 60.7 \pm 0.6 | 62.4 \pm 0.6 |
| australian_scale | 85.4 \pm 0.4 | 84.9 \pm 0.6 | 79.2 \pm 0.5 | 80.0 \pm 0.7 | 84.2 \pm 0.6 |
| australian_scale2 | 85.7 \pm 0.2 | 85.7 \pm 0.2 | 86.7 \pm 0.3 | 75.4 \pm 2.0 | 86.2 \pm 0.2 |
| breast-cancer_scale | 95.8 \pm 0.1 | 96.0 \pm 0.3 | 90.1 \pm 0.3 | 91.0 \pm 0.4 | 89.3 \pm 0.5 |
| breast-cancer_scale2 | 95.6 \pm 0.3 | 94.4 \pm 0.8 | 95.9 \pm 0.1 | 92.5 \pm 0.2 | 94.2 \pm 0.3 |
| diabetes_scale | 66.7 \pm 0.7 | 65.5 \pm 0.9 | 69.4 \pm 0.4 | 67.9 \pm 0.3 | 66.4 \pm 0.2 |
| skin_nonskin | - | - | 98.2 \pm 0.1 | 78.0 \pm 0.4 | 85.3 \pm 0.7 |
| skin_nonskin2 | - | - | 98.6 \pm 0.0 | 93.9 \pm 0.1 | 98.1 \pm 0.2 |
| epsilon_normalized | - | - | 64.5 \pm 0.3 | 63.1 \pm 0.1 | 56.3 \pm 1.3 |
| HIGGS | - | - | 56.3 \pm 0.2 | 52.6 \pm 0.1 | 54.0 \pm 0.2 |
| AUC $\times 100$ | | | | | |
| heart_scale | 78.4 \pm 1.0 | 78.3 \pm 1.1 | 73.8 \pm 0.9 | 72.5 \pm 1.1 | 79.0 \pm 0.9 |
| sonar_scale | 61.2 \pm 0.8 | 60.6 \pm 0.9 | 67.4 \pm 0.7 | 66.2 \pm 0.7 | 68.9 \pm 0.8 |
| australian_scale | 91.1 \pm 0.2 | 91.3 \pm 0.3 | 87.8 \pm 0.4 | 87.8 \pm 0.5 | 90.4 \pm 0.4 |
| australian_scale2 | 89.2 \pm 0.4 | 87.3 \pm 0.4 | 84.3 \pm 0.6 | 85.9 \pm 0.7 | 88.6 \pm 0.6 |
| breast-cancer_scale | 99.4 \pm 0.0 | 99.3 \pm 0.0 | 97.8 \pm 0.1 | 95.6 \pm 0.4 | 99.5 \pm 0.0 |
| breast-cancer_scale2 | 99.3 \pm 0.0 | 99.2 \pm 0.1 | 99.3 \pm 0.0 | 97.2 \pm 0.2 | 98.7 \pm 0.2 |
| diabetes_scale | 74.0 \pm 0.6 | 71.5 \pm 1.1 | 73.5 \pm 0.6 | 74.7 \pm 0.5 | 74.5 \pm 0.7 |
| skin_nonskin | - | - | 99.5 \pm 0.1 | 94.8 \pm 0.1 | 99.4 \pm 0.1 |
| skin_nonskin2 | - | - | 99.7 \pm 0.0 | 94.6 \pm 0.0 | 99.8 \pm 0.0 |
| epsilon_normalized | - | - | 70.0 \pm 0.4 | 69.3 \pm 0.1 | 62.2 \pm 2.3 |
| HIGGS | - | - | 59.6 \pm 0.2 | 65.3 \pm 0.1 | 55.7 \pm 0.3 |

Table 2.4: Accuracy and AUC comparison using the real datasets in cases the class-prior π_+ is unknown. The ‘KM1’ method by Ramaswamy et al. [2016] is used for LOG, DH, and NNPU, and the density-based method is used for WMMD. We denote the infeasible cases due to ‘KM1’ method by the hyphen. Other details are given in Table 2.3.

| Dataset | LOG | DH | NNPU | tADJ | WMMD |
|----------------------|-------------------|------------|-------------------|-------------------|-------------------|
| Accuracy (in %) | | | | | |
| heart_scale | 39.5 ± 0.8 | 39.1 ± 0.7 | 42.4 ± 0.9 | 65.1 ± 0.8 | 70.5 ± 0.7 |
| sonar_scale | 53.5 ± 0.6 | 52.1 ± 0.5 | 59.9 ± 0.7 | 60.7 ± 0.6 | 54.3 ± 0.8 |
| australian_scale | 50.0 ± 0.2 | 50.0 ± 0.2 | 50.0 ± 0.2 | 80.0 ± 0.7 | 79.4 ± 1.0 |
| australian_scale2 | 84.9 ± 0.2 | 84.9 ± 0.2 | 85.6 ± 0.3 | 85.5 ± 1.0 | 80.2 ± 1.0 |
| breast-cancer_scale | 65.0 ± 0.2 | 65.2 ± 0.2 | 65.0 ± 0.2 | 91.0 ± 0.4 | 93.0 ± 1.0 |
| breast-cancer_scale2 | 35.0 ± 0.2 | 35.0 ± 0.2 | 35.0 ± 0.2 | 92.5 ± 0.2 | 92.6 ± 0.4 |
| diabetes_scale | 36.0 ± 0.3 | 41.1 ± 1.1 | 37.9 ± 0.5 | 67.9 ± 0.3 | 65.1 ± 0.2 |
| skin_nonskin | - | - | - | 78.0 ± 0.4 | 82.2 ± 0.9 |
| skin_nonskin2 | - | - | - | 93.9 ± 0.1 | 95.7 ± 0.4 |
| epsilon_normalized | - | - | - | 63.1 ± 0.1 | 49.9 ± 0.1 |
| HIGGS | - | - | - | 52.6 ± 0.1 | 50.9 ± 0.0 |
| AUC $\times 100$ | | | | | |
| heart_scale | 67.2 ± 1.8 | 67.2 ± 1.4 | 71.0 ± 0.9 | 72.5 ± 1.1 | 77.1 ± 0.9 |
| sonar_scale | 60.5 ± 0.9 | 62.2 ± 0.9 | 66.6 ± 0.8 | 66.2 ± 0.7 | 69.7 ± 0.8 |
| australian_scale | 78.4 ± 1.1 | 72.5 ± 1.4 | 80.3 ± 0.6 | 87.8 ± 0.5 | 90.3 ± 0.3 |
| australian_scale2 | 92.9 ± 0.2 | 89.6 ± 0.6 | 85.9 ± 0.7 | 92.4 ± 0.3 | 93.3 ± 0.2 |
| breast-cancer_scale | 98.9 ± 0.1 | 93.5 ± 1.8 | 54.8 ± 1.1 | 95.6 ± 0.4 | 99.5 ± 0.0 |
| breast-cancer_scale2 | 14.4 ± 2.0 | 19.4 ± 3.8 | 91.5 ± 0.3 | 97.2 ± 0.2 | 99.0 ± 0.1 |
| diabetes_scale | 64.0 ± 1.3 | 63.8 ± 1.4 | 72.6 ± 0.5 | 74.7 ± 0.5 | 75.9 ± 0.5 |
| skin_nonskin | - | - | - | 94.8 ± 0.1 | 99.5 ± 0.1 |
| skin_nonskin2 | - | - | - | 94.6 ± 0.0 | 99.8 ± 0.0 |
| epsilon_normalized | - | - | - | 69.3 ± 0.1 | 59.7 ± 1.8 |
| HIGGS | - | - | - | 65.3 ± 0.1 | 55.4 ± 0.2 |

2.6 Concluding remarks

Existing methods use different objective functions and hypothesis spaces, and as a consequence, different optimization algorithms. Hence, there is no reason that one method outperforms uniformly for all scenarios. It is possible that one particular method may outperform in one scenario, for example, NNPU proposed by Kiryo et al. [2017] would perform better in complicated data settings because of the expressive power of neural networks. However, the proposed method has a clear computational advantage due to the closed-form as well as theoretical strength in terms of the explicit excess risk bound. In this regard, we believe the proposed method can be used as a principled and easy-to-compute baseline algorithm in PU learning.

2.7 Appendix

In this section, we provide all the proofs, implementation details, and additional experiments.

2.7.1 Proof of Theorem 2.2.1

Proof of Theorem 2.2.1. Since a function $f \in \mathcal{F}$ is bounded by 1, we have $\ell_h(yf(x)) = \max(0, 1 - yf(x)) = 1 - yf(x)$. Then, from Equation (2.2),

$$\begin{aligned} R_{\ell_h}(f) &= \pi_+ \int_{\mathcal{X}} \ell_h(f(x)) - \ell_h(-f(x)) dP_{X|Y=1}(x) + \int_{\mathcal{X}} \ell_h(-f(x)) dP_X(x) \\ &= 1 + \int_{\mathcal{X}} f(x) dP_X(x) - 2\pi_+ \int_{\mathcal{X}} f(x) dP_{X|Y=1}(x). \end{aligned}$$

Thus, we have

$$\begin{aligned}
\inf_{f \in \mathcal{F}} R_{\ell_h}(f) &= 1 + \inf_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) dP_X(x) - 2\pi_+ \int_{\mathcal{X}} f(x) dP_{X|Y=1}(x) \right\} \\
&= 1 - \sup_{f \in \mathcal{F}} \left\{ - \int_{\mathcal{X}} f(x) dP_X(x) + 2\pi_+ \int_{\mathcal{X}} f(x) dP_{X|Y=1}(x) \right\} \\
&\stackrel{(*)}{=} 1 - \sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) dP_X(x) - 2\pi_+ \int_{\mathcal{X}} f(x) dP_{X|Y=1}(x) \right\} \\
&= 1 - \text{WIPM}(P_X, P_{X|Y=1}; 2\pi_+, \mathcal{F}).
\end{aligned}$$

Equation (*) holds because \mathcal{F} is symmetric.

For the second result, note that a WIPM optimizer $g_{\mathcal{F}}$ satisfies $\text{WIPM}(P_X, P_{X|Y=1}; 2\pi_+, \mathcal{F}) = \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_X(x) - 2\pi_+ \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_{X|Y=1}(x)$.

Thus, we have

$$\begin{aligned}
\inf_{f \in \mathcal{F}} R_{\ell_h}(f) &= 1 - \text{WIPM}(P_X, P_{X|Y=1}; 2\pi_+, \mathcal{F}) \\
&= 1 - \left\{ \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_X(x) - 2\pi_+ \int_{\mathcal{X}} g_{\mathcal{F}}(x) dP_{X|Y=1}(x) \right\} \\
&= 1 - \{R_{\ell_h}(g_{\mathcal{F}}) - 1\} \\
&= 2 - R_{\ell_h}(g_{\mathcal{F}}) = R_{\ell_h}(-g_{\mathcal{F}}).
\end{aligned}$$

The last equality is from $R_{\ell_h}(g_{\mathcal{F}}) + R_{\ell_h}(-g_{\mathcal{F}}) = 2$ due to $g_{\mathcal{F}} \in \mathcal{F} \subseteq \mathcal{M}$. \square

2.7.2 Proofs for Section 2.2.3: Theoretical properties of empirical WIPM optimizer

In this section, we present a proof of Theorem 2.2.2. We also provide a proof of Proposition 2.2.3. Before presenting a proof of Theorem 2.2.2, we begin with necessary technical preliminaries.

Preliminaries for Theorem 2.2.2

In supervised binary classification settings, Sriperumbudur et al. [2012] introduced an empirical estimator for IPM and developed its consistency result. In Proposition 2.7.1, we recreate theoretical results for PU learning settings, giving a consistency result of empirical WIPM estimator.

Proposition 2.7.1 (Consistency result of WIPM estimator). *Let \mathcal{F} be the symmetric function space such that $\|f\|_\infty \leq \nu$, $\text{Var}_{P_{X|Y=1}}(f) \leq \sigma_{X|Y=1}^2$, and $\text{Var}_{P_X}(f) \leq \sigma_X^2$. Denote $\rho^2 = \sigma_{X|Y=1}^2 \vee \sigma_X^2$. Then for all $w, \alpha, \tau > 0$, the following holds with probability at least $1 - e^{-\tau}$ over the choice of $\mathcal{X}_{\text{pu}} := \{x_1^{\text{p}}, \dots, x_{n_{\text{p}}}^{\text{p}}, x_1^{\text{u}}, \dots, x_{n_{\text{u}}}^{\text{u}}\} \sim P_{\text{pu}} := P_{X|Y=1}^{n_{\text{p}}} \times P_X^{n_{\text{u}}}$,*

$$\begin{aligned} & |\text{WIPM}(P_{X, n_{\text{u}}}, P_{X|Y=1, n_{\text{p}}}; w, \mathcal{F}) - \text{WIPM}(P_X, P_{X|Y=1}; w, \mathcal{F})| \\ & \leq 2(1 + \alpha) [\mathbb{E}_{P_X^{n_{\text{u}}}} \{\mathfrak{R}_{\mathcal{X}_{\text{u}}}(\mathcal{F})\} + w \mathbb{E}_{P_{X|Y=1}^{n_{\text{p}}}} \{\mathfrak{R}_{\mathcal{X}_{\text{p}}}(\mathcal{F})\}] \\ & + \chi_{n_{\text{p}}, n_{\text{u}}}^{(1)}(w) \sqrt{2\tau\rho^2} + \tau \chi_{n_{\text{p}}, n_{\text{u}}}^{(2)}(w) \nu \left(\frac{2}{3} + \frac{1}{\alpha} \right). \end{aligned} \quad (2.10)$$

Proof of Proposition 2.7.1. The following proof is a slight modification of the proof of Theorem 3.3 in Sriperumbudur et al. [2012]. Without loss of generality, by changing an order, we define a set of observations and a set of weights as follows,

$$(x_1, \dots, x_{n_{\text{p}}}, x_{n_{\text{p}}+1}, \dots, x_{n_{\text{p}}+n_{\text{u}}}) := \mathcal{X}_{\text{pu}},$$

and

$$(\tilde{y}_1, \dots, \tilde{y}_{n_{\text{p}}}, \tilde{y}_{n_{\text{p}}+1}, \dots, \tilde{y}_{n_{\text{p}}+n_{\text{u}}})$$

$$= (w/n_p, \dots, w/n_p, -1/n_u, \dots, -1/n_u),$$

respectively. For independent Rademacher random variables $\{\sigma_i\}_{i=1}^{n_p+n_u}$, we define the empirical Rademacher complexity-like term given by

$$\tilde{\mathfrak{R}}_{\mathcal{X}_{\text{pu}}}(\mathcal{F}) := \mathbb{E}_\sigma \left\{ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n_p+n_u} \sigma_i \tilde{y}_i f(X_i) \right| : (X_1, \dots, X_{n_p+n_u}) = \mathcal{X}_{\text{pu}} \right\}.$$

Note that $\tilde{\mathfrak{R}}_{\mathcal{X}_{\text{pu}}}(\mathcal{F}) \leq \mathfrak{R}_{\mathcal{X}_u}(\mathcal{F}) + w\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F})$. Define $\mu_i = P_{X|Y=1}$ for $i \in \{1, \dots, n_p\}$ and $\mu_i = P_X$ for $i \in \{n_p+1, \dots, n_p+n_u\}$, respectively. That is, $P_{\text{pu}} = \times_{i=1}^{n_p+n_u} \mu_i$. Let $(X_1, \dots, X_{n_p+n_u}) \sim P_{\text{pu}}$ and define random variables $\theta_i(f, X_i) = w\{f(X_i) - P_{X|Y=1}(f)\}/n_p$ for $i \in \{1, \dots, n_p\}$ and $\theta_i(f, X_i) = \{f(X_i) - P_X(f)\}/n_u$ for $i \in \{n_p+1, \dots, n_p+n_u\}$, respectively.

Then, using the fact that $|\sup |C| - \sup |D|| \leq \sup |C - D|$, we have

$$\begin{aligned} & |\text{WIPM}(P_{X, n_u}, P_{X|Y=1, n_p}; w, \mathcal{F}) - \text{WIPM}(P_X, P_{X|Y=1}; w, \mathcal{F})| \\ & \leq \sup_{f \in \mathcal{F}} \left| \left\{ \int f d(P_{X, n_u} - wP_{X|Y=1, n_p}) \right\} - \left\{ \int f d(P_X - wP_{X|Y=1}) \right\} \right| \\ & \hspace{20em} (2.11) \\ & = \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n_p+n_u} \theta_i(f, X_i) \right| =: h(X_1, \dots, X_{n_p+n_u}). \end{aligned}$$

Further, it is easy to verify that (i) $\int_{\mathcal{X}} \theta_i(f, z) d\mu_i(z) = 0$ for all i and $f \in \mathcal{F}$ and (ii) $\int_{\mathcal{X}} \theta_i^2(f, z) d\mu_i(z)$ is bounded by $w^2 \sigma_{X|Y=1}^2 / n_p^2$ for $i \in \{1, \dots, n_p\}$ and σ_X^2 / n_u^2 for $i \in \{n_p+1, \dots, n_p+n_u\}$, respectively. Finally, (iii) $\|\theta_i(f, \cdot)\|_\infty \leq 2(w/n_p + 1/n_u)\nu = \chi_{n_p, n_u}^{(2)}(w)\nu$ for all $i \in \{1, \dots, n_p+n_u\}$.

Then, for all $\alpha > 0$, the following holds with probability at

least $1 - e^{-\tau}$,

$$\begin{aligned}
& h(X_1, \dots, X_{n_p+n_u}) \\
& \leq (1 + \alpha) \mathbb{E}_{P_{\text{pu}}}(h) + \sqrt{2\tau \frac{(w^2 n_u + n_p)}{n_p n_u} (\sigma_{X|Y=1}^2 \vee \sigma_X^2)} + \tau \chi_{n_p, n_u}^{(2)}(w) \nu \left(\frac{2}{3} + \frac{1}{\alpha} \right) \\
& \leq 2(1 + \alpha) \mathbb{E}_{P_{\text{pu}}} \{ \tilde{\mathfrak{R}}_{\mathcal{X}_{\text{pu}}}(\mathcal{F}) \} + \chi_{n_p, n_u}^{(1)}(w) \sqrt{2\tau \rho^2} + \tau \chi_{n_p, n_u}^{(2)}(w) \nu \left(\frac{2}{3} + \frac{1}{\alpha} \right).
\end{aligned}$$

The first inequality is derived by the second inequality of Lemma 2.7.2, a variant of the Talagrand's inequality. The second inequality is from using a symmetrization lemma: with corresponding independent ghost empirical distributions \tilde{P}_{X, n_u} and $\tilde{P}_{X|Y=1, n_p}$,

$$\begin{aligned}
\mathbb{E}_{P_{\text{pu}}}(h) &= \mathbb{E}_{P_{\text{pu}}} \sup_{f \in \mathcal{F}} \left| \left\{ \int f d(P_{X, n_u} - w P_{X|Y=1, n_p}) \right\} \right. \\
&\quad \left. - \left\{ \int f d(P_X - w P_{X|Y=1}) \right\} \right| \\
&= \mathbb{E}_{P_{\text{pu}}} \sup_{f \in \mathcal{F}} \left| \left\{ \int f d(P_{X, n_u} - w P_{X|Y=1, n_p}) \right\} \right. \\
&\quad \left. - \mathbb{E}_{P_{\text{pu}}} \left\{ \int f d(\tilde{P}_{X, n_u} - w \tilde{P}_{X|Y=1, n_p}) \right\} \right| \\
&\leq \mathbb{E}_{P_{\text{pu}}} \sup_{f \in \mathcal{F}} \left| \left\{ \int f d(P_{X, n_u} - w P_{X|Y=1, n_p}) \right\} \right. \\
&\quad \left. - \left\{ \int f d(\tilde{P}_{X, n_u} - w \tilde{P}_{X|Y=1, n_p}) \right\} \right| \\
&\leq 2 \mathbb{E}_{P_{\text{pu}}} \sup_{f \in \mathcal{F}} \left| \int f d(P_{X, n_u} - w P_{X|Y=1, n_p}) \right| \\
&\leq 2 \mathbb{E}_{P_{\text{pu}}} \mathbb{E}_{\sigma} \left\{ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n_p+n_u} \sigma_i \tilde{y}_i f(X_i) \right| : (X_1, \dots, X_{n_p+n_u}) = \mathcal{X}_{\text{pu}} \right\} \\
&\leq 2 \mathbb{E}_{P_{\text{pu}}} \{ \tilde{\mathfrak{R}}_{\mathcal{X}_{\text{pu}}}(\mathcal{F}) \}
\end{aligned}$$

Next, simply using the fact $\tilde{\mathfrak{R}}_{\mathcal{X}_{\text{pu}}}(\mathcal{F}) \leq \mathfrak{R}_{\mathcal{X}_u}(\mathcal{F}) + w \mathfrak{R}_{\mathcal{X}_p}(\mathcal{F})$, we have for all $w, \alpha, \tau > 0$, the following holds with probability at

least $1 - e^{-\tau}$,

$$\begin{aligned} & |\text{WIPM}(P_{X, n_u}, P_{X|Y=1, n_p}; w, \mathcal{F}) - \text{WIPM}(P_X, P_{X|Y=1}; w, \mathcal{F})| \\ & \leq 2(1 + \alpha)[\mathbb{E}_{P_X^{n_u}}\{\mathfrak{R}_{\mathcal{X}_u}(\mathcal{F})\} + w\mathbb{E}_{P_{X|Y=1}^{n_p}}\{\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F})\}] \\ & \quad + \chi_{n_p, n_u}^{(1)}(w)\sqrt{2\tau\rho^2} + \tau\chi_{n_p, n_u}^{(2)}(w)\nu\left(\frac{2}{3} + \frac{1}{\alpha}\right). \end{aligned}$$

It concludes the proof. \square

For Lemmas 2.7.2, we quote the Proposition B.1 of Sriperumbudur et al. [2012] without proofs.

Lemma 2.7.2 (Proposition B.1 of Sriperumbudur et al. [2012]: A variant of Talagrand's inequality). *Let $B \geq 0, n \geq 1, (\Omega_i, \mathcal{A}_i, \mu_i), i = 1, \dots, n$ be a probability space and $\theta_i : \mathcal{F} \times \Omega_i \rightarrow \mathbb{R}$ be bounded measurable functions, where \mathcal{F} is the space of real-valued \mathcal{A}_i -measurable functions for all i . Suppose*

- (a) $\int_{\Omega_i} \theta_i(f, \omega) d\mu_i(\omega) = 0$ for all i and $f \in \mathcal{F}$
- (b) $\int_{\Omega_i} \theta_i^2(f, \omega) d\mu_i(\omega) \leq \rho_i^2$ for all i and $f \in \mathcal{F}$
- (c) $\|\theta_i(f, \cdot)\|_\infty \leq B$ for all i and $f \in \mathcal{F}$.

Define $Z := \times_{i=1}^n \Omega_i$ and $P := \times_{i=1}^n \mu_i$. Furthermore, define $g : Z \rightarrow \mathbb{R}$ by

$$g(z) := \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \theta_i(f, \omega_i) \right|, z = (\omega_1, \dots, \omega_n) \in Z.$$

Then, for all $\tau > 0$, we have

$$P \left(\left\{ z \in Z : g(z) \geq \mathbb{E}_P g + \sqrt{2\tau \left(\sum_{i=1}^n \rho_i^2 + 2B\mathbb{E}_P g \right)} + \frac{2\tau B}{3} \right\} \right) \leq e^{-\tau}.$$

In addition, for all $\tau > 0$ and $\alpha > 0$,

$$P \left(\left\{ z \in Z : g(z) \geq (1 + \alpha) \mathbb{E} P g + \sqrt{2\tau \sum_{i=1}^n \rho_i^2} + \tau B \left(\frac{2}{3} + \frac{1}{\alpha} \right) \right\} \right) \leq e^{-\tau}.$$

Proof of Theorem 2.2.2: estimation error bound of WIPM optimizer

Proof of Theorem 2.2.2. We first define the empirical risk estimator $\hat{R}_{\ell_h}(f)$ by replacing data distributions in Equation (2.2) with the empirical distributions. To be more specific, we define

$$\begin{aligned} \hat{R}_{\ell_h}(f) &:= \pi_+ \int_{\mathcal{X}} [\ell_h\{f(x)\} - \ell_h\{-f(x)\}] dP_{X|Y=1, n_p}(x) + \int_{\mathcal{X}} \ell_h\{-f(x)\} dP_{X, n_u}(x) \\ &= \frac{\pi_+}{n_p} \sum_{i=1}^{n_p} [\ell_h\{f(x_i^p)\} - \ell_h\{-f(x_i^p)\}] + \frac{1}{n_u} \sum_{i=1}^{n_u} \ell_h\{-f(x_i^u)\}. \end{aligned}$$

Since $\|f\|_{\infty} \leq 1$, using the similar derivations in Section 2.7.1, we have

$$\hat{R}_{\ell_h}(f) = 1 + \frac{1}{n_u} \sum_{i=1}^{n_u} f(x_i^u) - \frac{2\pi_+}{n_p} \sum_{i=1}^{n_p} f(x_i^p).$$

By the result of Theorem 2.2.1, the negative of an WIPM optimizer $f_{\mathcal{F}}$ is minimizer of $R_{\ell_h}(f)$, *i.e.*, $R_{\ell_h}(f_{\mathcal{F}}) = \inf_{f \in \mathcal{F}} R_{\ell_h}(f)$. Thus, we have

$$\begin{aligned} &R_{\ell_h}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_{\ell_h}(f) \\ &= R_{\ell_h}(\hat{f}_{\mathcal{F}}) - R_{\ell_h}(f_{\mathcal{F}}) \\ &= \left\{ R_{\ell_h}(\hat{f}_{\mathcal{F}}) - \hat{R}_{\ell_h}(\hat{f}_{\mathcal{F}}) \right\} + \left\{ \hat{R}_{\ell_h}(\hat{f}_{\mathcal{F}}) - \hat{R}_{\ell_h}(f_{\mathcal{F}}) \right\} + \left\{ \hat{R}_{\ell_h}(f_{\mathcal{F}}) - R_{\ell_h}(f_{\mathcal{F}}) \right\} \\ &\leq \sup_{f \in \mathcal{F}} |\hat{R}_{\ell_h}(f) - R_{\ell_h}(f)| + 0 + \sup_{f \in \mathcal{F}} |\hat{R}_{\ell_h}(f) - R_{\ell_h}(f)| \end{aligned}$$

$$= 2 \sup_{f \in \mathcal{F}} |\hat{R}_{\ell_h}(f) - R_{\ell_h}(f)|.$$

The first inequality holds since $\hat{R}_{\ell_h}(\hat{f}_{\mathcal{F}}) = 1 - \text{WIPM}(P_{X, n_u}, P_{X|Y=1, n_p}; 2\pi_+, \mathcal{F}) \leq \hat{R}_{\ell_h}(f)$ for any $f \in \mathcal{F}$. Thus it is enough to bound $\sup_{f \in \mathcal{F}} |\hat{R}_{\ell_h}(f) - R_{\ell_h}(f)|$, and

$$\begin{aligned} & \sup_{f \in \mathcal{F}} |\hat{R}_{\ell_h}(f) - R_{\ell_h}(f)| \\ & \leq \sup_{f \in \mathcal{F}} \left| \left\{ \frac{1}{n_u} \sum_{i=1}^{n_u} f(x_i^u) - \frac{2\pi_+}{n_p} \sum_{i=1}^{n_p} f(x_i^p) \right\} - \left\{ \int f d(P_X - 2\pi_+ P_{X|Y=1}) \right\} \right| \\ & = \sup_{f \in \mathcal{F}} \left| \left\{ \int f d(P_{X, n_u} - 2\pi_+ P_{X|Y=1, n_p}) \right\} - \left\{ \int f d(P_X - 2\pi_+ P_{X|Y=1}) \right\} \right|. \end{aligned}$$

Note that this is a special case of Equation (2.11). Therefore, applying Equation (2.10) in Proposition 2.7.1 with $w = 2\pi_+$, we have for all $\alpha, \tau > 0$, the following holds with probability at least $1 - e^{-\tau}$,

$$\begin{aligned} & R_{\ell_h}(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_{\ell_h}(f) \\ & \leq 4(1 + \alpha) [\mathbb{E}_{P_X^{n_u}} \{\mathfrak{R}_{\mathcal{X}_u}(\mathcal{F})\} + 2\pi_+ \mathbb{E}_{P_{X|Y=1}^{n_p}} \{\mathfrak{R}_{\mathcal{X}_p}(\mathcal{F})\}] \\ & \quad + 2\chi_{n_p, n_u}^{(1)}(2\pi_+) \sqrt{2\tau\rho^2} + 2\tau\chi_{n_p, n_u}^{(2)}(2\pi_+) \nu \left(\frac{2}{3} + \frac{1}{\alpha} \right). \end{aligned}$$

This concludes a proof. \square

Proof of Proposition 2.2.3

Proof of Proposition 2.2.3. After omitting the positive term Δ from the upper bound (2.6) and plugging $\alpha = 1$ in the upper bound (2.5), it is enough to show that, under the condition (2.7), the following is satisfied.

$$8R + \chi_{n_p, n_u}^{(1)}(2\pi_+)(1 + \nu)\sqrt{2\tau} \geq 8R + 2\chi_{n_p, n_u}^{(1)}(2\pi_+)\sqrt{2\tau\rho^2} + \frac{10}{3}\tau\chi_{n_p, n_u}^{(2)}(2\pi_+)\nu,$$

where $R = \mathbb{E}_{P_X^{n_u}} \{\mathfrak{R}_{\chi_u}(\mathcal{F})\} + 2\pi_+ \mathbb{E}_{P_{X|Y=1}^{n_p}} \{\mathfrak{R}_{\chi_p}(\mathcal{F})\}$.

Using simple algebras, we have

$$\begin{aligned}
& \frac{1 + \nu}{2} - \frac{5\sqrt{2\tau}\chi_{n_p, n_u}^{(2)}(2\pi_+)\nu}{6\chi_{n_p, n_u}^{(1)}(2\pi_+)} \geq \rho \\
& \implies \chi_{n_p, n_u}^{(1)}(2\pi_+)(1 + \nu)\sqrt{2\tau} - 2\tau\chi_{n_p, n_u}^{(2)}(2\pi_+)\nu \frac{5}{3} \\
& \geq 2\chi_{n_p, n_u}^{(1)}(2\pi_+)\sqrt{2\tau\rho^2} \\
& \implies 8R + \chi_{n_p, n_u}^{(1)}(2\pi_+)(1 + \nu)\sqrt{2\tau} \\
& \geq 8R + 2\chi_{n_p, n_u}^{(1)}(2\pi_+)\sqrt{2\tau\rho^2} + \frac{10}{3}\tau\chi_{n_p, n_u}^{(2)}(2\pi_+)\nu.
\end{aligned}$$

Thus, the proposed upper bound is sharper than that of Kiryo et al. [2017] if the condition (2.7) holds. \square

2.7.3 Proofs for Section 2.3: The empirical WMMD optimizer and the WMMD classifier

In this section, we first show that the empirical WMMD optimizer has a closed-form expression. We provide proofs of the two propositions: (i) explicitly showing the estimation error bound in Proposition 2.3.2 and (ii) deriving an approximation error bound in Proposition 2.3.3. Lastly, we provide a proof of Theorem 2.3.4.

Proof of Proposition 2.7.3: WMMD optimizer has a closed-form expression

We first state and prove the following Proposition 2.7.3, which is an extended version of Proposition 2.3.1. Please note that Proposition 2.3.1 can be directly obtained by plugging the two empirical distributions $P_{X|Y=1, n_p}$ and P_{X, n_u} and $w = 2\pi_+$.

Proposition 2.7.3 (Weighted maximum mean discrepancy). *Let P and Q be two probability measures defined on \mathcal{X} and let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a bounded reproducing kernel.*

(a) *WMMD between two probability measures P and Q with a weight w and a closed ball $\mathcal{H}_{k,r}$ with the radius r can be represented in a closed form,*

$$\begin{aligned} & \text{WMMD}_k(P, Q; w, r) \\ &= r \left\{ \mathbb{E}_{P^2}[k(x, x')] + w^2 \mathbb{E}_{Q^2}[k(y, y')] - 2w \mathbb{E}_{P \times Q}[k(x, y)] \right\}^{1/2}, \end{aligned}$$

where x and x' independently follow P and y , and y' independently follow Q .

(b) *we also have a closed-form expression for the unique optimizer $g_{\mathcal{H}_{k,r}} \in \mathcal{H}_{k,r}$ given by*

$$g_{\mathcal{H}_{k,r}}(z) = r \times T \left(\int k(z, x) dP(x) - w \int k(z, x) dQ(x) \right),$$

where T is the normalizing operator defined by $T(g) = g / \|g\|_{\mathcal{H}_k}$.

(c) *The associated classifier is given by*

$$\text{sign}\{f_{\mathcal{H}_{k,r}}(z)\} = \begin{cases} +1 & \text{if } w^{-1} < \lambda_{Q,P}(z) \\ -1 & \text{otherwise} \end{cases},$$

where

$$\lambda_{Q,P}(z) = \frac{\int k(z, x) dQ(x)}{\int k(z, x) dP(x)}.$$

Proof of Proposition 2.7.3. The main idea of this proof is to use $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k}$ as Gretton et al. [2012] showed. From the definition of the WMMD, we have

$$\text{WMMD}_k(P, Q; w, r)$$

$$\begin{aligned}
& := \sup_{f \in \mathcal{H}_{k,r}} \left| \int_{\mathcal{X}} f(x) dP(x) - w \int_{\mathcal{X}} f(x) dQ(x) \right| \\
& = \sup_{f \in \mathcal{H}_{k,r}} \left| \int_{\mathcal{X}} \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} dP(x) - w \int_{\mathcal{X}} \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} dQ(x) \right| \\
& = \sup_{f \in \mathcal{H}_{k,r}} \left| \langle f, \int_{\mathcal{X}} k(\cdot, x) \{dP(x) - wdQ(x)\} \rangle_{\mathcal{H}_k} \right| \\
& = r \left\| \int_{\mathcal{X}} k(\cdot, x) dP(x) - w \int_{\mathcal{X}} k(\cdot, x) dQ(x) \right\|_{\mathcal{H}_k}.
\end{aligned}$$

The last equation is obtained by Cauchy-Schwarz inequality with the WMMD optimizer $g_{\mathcal{H}_{k,r}} \in \mathcal{H}_{k,r}$ given by

$$g_{\mathcal{H}_{k,r}}(z) = r \times T \left(\int_{\mathcal{X}} k(z, x) dP(x) - w \int_{\mathcal{X}} k(z, x) dQ(x) \right).$$

It concludes a proof of (b). Furthermore,

$$\begin{aligned}
& r^{-1} \times \text{WMMD}_k(P, Q; w, r) \\
& = \left\| \int_{\mathcal{X}} k(\cdot, x) dP(x) - w \int_{\mathcal{X}} k(\cdot, x) dQ(x) \right\|_{\mathcal{H}_k} \\
& = \left\langle \int_{\mathcal{X}} k(\cdot, x) dP(x) - w \int_{\mathcal{X}} k(\cdot, x) dQ(x), \right. \\
& \quad \left. \int_{\mathcal{X}} k(\cdot, x) dP(x) - w \int_{\mathcal{X}} k(\cdot, x) dQ(x) \right\rangle_{\mathcal{H}_k}^{1/2} \\
& = \left\{ \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') dP(x) dP(x') + w^2 \int_{\mathcal{X}} \int_{\mathcal{X}} k(y, y') dQ(y) dQ(y') \right. \\
& \quad \left. - 2w \int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y) dP(x) dQ(y) \right\}^{1/2} \\
& = \left\{ \mathbb{E}_{P^2}[k(x, x')] + w^2 \mathbb{E}_{Q^2}[k(y, y')] - 2w \mathbb{E}_{P \times Q}[k(x, y)] \right\}^{1/2}.
\end{aligned}$$

It concludes a proof of (a).

Lastly, we prove the statement (c). Note that the associated classifier is determined by the sign of $f_{\mathcal{H}_{k,r}} = -g_{\mathcal{H}_{k,r}}$. From the

statement (b), we have

$$\text{sign}(g_{\mathcal{H}_{k,r}}(z)) = \begin{cases} +1 & \text{if } \frac{\int k(z,x)dP(x)}{\int k(z,x)dQ(x)} > w \\ -1 & \text{otherwise} \end{cases}.$$

Hence, we have

$$\text{sign}(f_{\mathcal{H}_{k,r}}(z)) = -\text{sign}(g_{\mathcal{H}_{k,r}}(z)) = \begin{cases} +1 & \text{if } w^{-1} < \lambda_{Q,P}(z) \\ -1 & \text{otherwise} \end{cases}.$$

□

Proof of Proposition 2.3.2

Proof of Proposition 2.3.2. We first prove $\mathcal{H}_{k,r_1} \subseteq \mathcal{M}$. By the reproducing property of \mathcal{H}_k and Cauchy-Schwarz inequality, for any $f \in \mathcal{H}_{k,r_1}, x \in \mathcal{X}$

$$|f(x)| = |\langle k(\cdot, x), f \rangle_{\mathcal{H}_k}| \leq \|k(\cdot, x)\|_{\mathcal{H}_k} \|f\|_{\mathcal{H}_k} = \sqrt{k(x, x)} \|f\|_{\mathcal{H}_k} \leq r_1^{-1} r_1 = 1.$$

Thus, $\|f\|_{\infty} \leq 1$ for all $f \in \mathcal{H}_{k,r_1}$. This proves $\mathcal{H}_{k,r_1} \subseteq \mathcal{M}$.

Now, we prove the inequality. First, we apply Theorem 2.2.2 with \mathcal{H}_{k,r_1} .

[Step 1] From the result of Theorem 2.2.2, for all $\alpha, \tau > 0$, the estimation error term is bounded above with probability at least $1 - e^{-\tau}$,

$$\begin{aligned} & R_{\ell_n}(\hat{f}_{\mathcal{H}_{k,r_1}}) - \inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_n}(f) \\ & \leq C_{\alpha}(\mathbb{E}_{P_X^{n_u}}(\mathfrak{R}_{\mathcal{X}_u}(\mathcal{H}_{k,r_1}))) + 2\pi_+ \mathbb{E}_{P_{X|Y=1}^{n_p}}(\mathfrak{R}_{\mathcal{X}_p}(\mathcal{H}_{k,r_1})) \\ & \quad + C_{\tau, \rho}^{(1)} \chi_{n_p, n_u}^{(1)}(2\pi_+) + C_{\tau, \nu, \alpha}^{(2)} \chi_{n_p, n_u}^{(2)}(2\pi_+). \end{aligned} \quad (2.12)$$

Using the notations $(\tilde{y}_1, \dots, \tilde{y}_{n_p}, \tilde{y}_{n_p+1}, \dots, \tilde{y}_{n_p+n_u}) = (2\pi_+/n_p, \dots, 2\pi_+/n_p, -1/n_u, \dots, -1/n_u)$, we obtain upper bound of the empirical Rademacher complexity of $\mathcal{H}_{k,r}$ given the positive samples $\mathfrak{R}_{\mathcal{X}_p}(\mathcal{H}_{k,r})$ as follows.

$$\begin{aligned}
& 2\pi_+ \mathfrak{R}_{\mathcal{X}_p}(\mathcal{H}_{k,r_1}) \\
&= \mathbb{E}_\sigma \left\{ \sup_{f \in \mathcal{H}_{k,r_1}} \left| \sum_{i=1}^{n_p} \sigma_i \tilde{y}_i f(X_i) \right| \right\} = \mathbb{E}_\sigma \left\{ \sup_{f \in \mathcal{H}_{k,r_1}} \left| \sum_{i=1}^{n_p} \sigma_i \tilde{y}_i \langle k(\cdot, X_i), f \rangle_{\mathcal{H}_k} \right| \right\} \\
&= \mathbb{E}_\sigma \left\{ \sup_{f \in \mathcal{H}_{k,r_1}} \left| \left\langle \sum_{i=1}^{n_p} \sigma_i \tilde{y}_i k(\cdot, X_i), f \right\rangle_{\mathcal{H}_k} \right| \right\} \leq r_1 \mathbb{E}_\sigma \left\{ \left\| \sum_{i=1}^{n_p} \sigma_i \tilde{y}_i k(\cdot, X_i) \right\|_{\mathcal{H}_k} \right\} \\
&\leq r_1 \sqrt{\mathbb{E}_\sigma \left\{ \sum_{i=1}^{n_p} \tilde{y}_i^2 k(X_i, X_i) \right\}} + r_1 \sqrt{\mathbb{E}_\sigma \left\{ \sum_{i \neq j} \sigma_i \sigma_j \tilde{y}_i \tilde{y}_j k(X_i, X_j) \right\}} \\
&= r_1 \sqrt{\mathbb{E}_\sigma \left\{ \sum_{i=1}^{n_p} \tilde{y}_i^2 k(X_i, X_i) \right\}} \leq \frac{2\pi_+}{\sqrt{n_p}}.
\end{aligned}$$

We continue the similar method to the unlabeled dataset, and applying expectation operator gives

$$\begin{aligned}
& \mathbb{E}_{P_X^{n_u}}(\mathfrak{R}_{\mathcal{X}_u}(\mathcal{H}_{k,r_1})) + 2\pi_+ \mathbb{E}_{P_{X|Y=1}^{n_p}}(\mathfrak{R}_{\mathcal{X}_p}(\mathcal{H}_{k,r_1})) \\
&\leq \frac{1}{\sqrt{n_u}} + \frac{2\pi_+}{\sqrt{n_p}} = \chi_{n_p, n_u}^{(1)}(2\pi_+). \tag{2.13}
\end{aligned}$$

[Step 2] Using Equations (2.12) and (2.13), we then conclude that for all $\alpha, \tau > 0$, the following holds with probability at least $1 - e^{-\tau}$,

$$\begin{aligned}
& R_{\ell_h}(\hat{f}_{\mathcal{H}_{k,r_1}}) - \inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_h}(f) \\
&\leq (C_\alpha + C_{\tau, \rho^2}^{(1)}) \chi_{n_p, n_u}^{(1)}(2\pi_+) + C_{\tau, \nu, \alpha}^{(2)} \chi_{n_p, n_u}^{(2)}(2\pi_+).
\end{aligned}$$

These equations conclude the proof. \square

Proof of Proposition 2.3.3

Proof of Proposition 2.3.3. [Step 1] In this step, we first claim that $\inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f) = R_{\ell_h}(\beta f_1^*)$. By Lin [2002, Lemma 3.1], the f_1^* satisfies that $\inf_{f \in \mathcal{U}} R_{\ell_h}(f) = \inf_{f \in \mathcal{M}} R_{\ell_h}(f) = R_{\ell_h}(f_1^*)$. Note that $\ell_h(yf(x)) = \max(0, 1 - yf(x)) = 1 - yf(x)$ for all $\|f\|_\infty \leq 1$. It is obvious that $\beta f_1^* \in \beta\mathcal{M}$. By definition of the infimum, we have $R_{\ell_h}(\beta f_1^*) \geq \inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f)$. Suppose $R_{\ell_h}(\beta f_1^*) > \inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f)$. Let f_β^* be a function in $\beta\mathcal{M}$ such that $R_{\ell_h}(f_\beta^*) = \inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f)$

Then,

$$\begin{aligned} R_{\ell_h}(\beta f_1^*) &> R_{\ell_h}(f_\beta^*) \\ \iff \int 1 - y\beta f_1^*(x) dP_{X,Y}(x, y) &> \int 1 - yf_\beta^*(x) dP_{X,Y}(x, y) \\ \iff \beta \int yf_1^*(x) dP_{X,Y}(x, y) &< \int yf_\beta^*(x) dP_{X,Y}(x, y). \end{aligned}$$

Since $\beta^{-1}f_\beta^* \in \mathcal{M}$,

$$\begin{aligned} R_{\ell_h}(\beta^{-1}f_\beta^*) &\geq \inf_{f \in \mathcal{M}} R_{\ell_h}(f) \\ \iff \int 1 - y\beta^{-1}f_\beta^*(x) dP_{X,Y}(x, y) &\geq \int 1 - yf_1^*(x) dP_{X,Y}(x, y) \\ \iff \int yf_\beta^*(x) dP_{X,Y}(x, y) &\leq \beta \int yf_1^*(x) dP_{X,Y}(x, y). \end{aligned}$$

Note that $\inf_{f \in \mathcal{M}} R_{\ell_h}(f) = R_{\ell_h}(f_1^*)$. This contradicts with the assumption $R_{\ell_h}(\beta f_1^*) > \inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f)$, and we have $\inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f) = R_{\ell_h}(\beta f_1^*)$.

[Step 2] By Proposition 2.3.2, we have $\mathcal{H}_{k,r_1} \subseteq \mathcal{M}$. Thus, for all $g \in \mathcal{H}_{k,r_1} \subseteq \mathcal{M}$ and $0 < \beta \leq 1$, we have

$$\begin{aligned} &\inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_h}(f) - \inf_{f \in \beta\mathcal{M}} R_{\ell_h}(f) \\ &\leq R_{\ell_h}(g) - R_{\ell_h}(\beta f_1^*) \end{aligned}$$

$$\begin{aligned}
&= \int \{(1 - yg(x)) - (1 - y\beta f_1^*(x))\} dP_{X,Y}(x, y) \\
&\leq \int |yg(x) - y\beta f_1^*(x)| dP_{X,Y}(x, y) \\
&\leq \sqrt{\int |y|^2 dP_{X,Y}(x, y)} \sqrt{\int |g(x) - \beta f_1^*(x)|^2 dP_{X,Y}(x, y)} \\
&= \|g - \beta f_1^*\|_{L_2(P_X)}.
\end{aligned}$$

The first equality holds because $\ell_h(yg(x)) = \max(0, 1 - yg(x)) = 1 - yg(x)$ for all $g \in \mathcal{H}_{k,r_1} \subseteq \mathcal{M}$. Hence,

$$\begin{aligned}
\inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_h}(f) - \inf_{f \in \beta \mathcal{M}} R_{\ell_h}(f) &\leq \inf_{g \in \mathcal{H}_{k,r_1}} \|g - \beta f_1^*\|_{L_2(P_X)} \\
&= \beta \inf_{g \in \mathcal{H}_{k,r_1}} \|g/\beta - f_1^*\|_{L_2(P_X)} \\
&= \beta \inf_{g \in \mathcal{H}_{k,r_1/\beta}} \|g - f_1^*\|_{L_2(P_X)}.
\end{aligned}$$

Therefore, by [Step 1] and [Step 2],

$$\inf_{f \in \mathcal{H}_{k,r_1}} R_{\ell_h}(f) - \inf_{f \in \beta \mathcal{M}} R_{\ell_h}(f) \leq \beta \inf_{g \in \mathcal{H}_{k,r_1/\beta}} \|g - f_1^*\|_{L_2(P_X)},$$

for any $0 < \beta \leq 1$. □

Proof of Theorem 2.3.4

Proof of Theorem 2.3.4. Let $\tilde{f}_{\mathcal{H}_k}(z) = \frac{2\pi_+}{\sqrt{2\pi n_p} h^d} \sum_{i=1}^{n_p} k(z, x_i^p) - \frac{1}{\sqrt{2\pi n_u} h^d} \sum_{i=1}^{n_u} k(z, x_i^u)$ and $\tilde{f}_{\text{Bayes}}(z) = 2\pi_+ p_{X|Y=1}(z) - p_X(z)$. Then, due to the $\ell_{01}(z) = \ell_{01}(cz)$ for any $c > 0$ and $z \in \mathbb{R}$, we have

$$\begin{aligned}
R_{\ell_{01}}(\hat{f}_{\mathcal{H}_{k,r_1}}) - \inf_{f \in \mathcal{U}} R_{\ell_{01}}(f) &= R_{\ell_{01}}(\tilde{f}_{\mathcal{H}_k}) - R_{\ell_{01}}(\tilde{f}_{\text{Bayes}}) \\
&= \mathbb{E}_X \left(|2\eta(X) - 1| \mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X) \tilde{f}_{\text{Bayes}}(X) < 0\}} \right),
\end{aligned}$$

where $\eta(x) = P(Y = 1 \mid X = x)$ and $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. The second equality is due to

$$\begin{aligned}
& R_{\ell_{01}}(\tilde{f}_{\mathcal{H}_k}) - R_{\ell_{01}}(\tilde{f}_{\text{Bayes}}) \\
&= \int \mathbb{1}_{\{Y\tilde{f}_{\mathcal{H}_k}(X) < 0\}} - \mathbb{1}_{\{Y\tilde{f}_{\text{Bayes}}(X) < 0\}} dP_{X,Y}(x, y) \\
&= \int \eta(x)(\mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X) < 0\}} - \mathbb{1}_{\{\tilde{f}_{\text{Bayes}}(X) < 0\}}) \\
&+ (1 - \eta(x))(\mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X) > 0\}} - \mathbb{1}_{\{\tilde{f}_{\text{Bayes}}(X) > 0\}}) dP_X(x) \\
&= \int (2\eta(x) - 1) \mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X)\tilde{f}_{\text{Bayes}}(X) < 0\}} \mathbb{1}_{\{\tilde{f}_{\text{Bayes}}(X) < 0\}} dP_X(x) \\
&+ \int (1 - 2\eta(x)) \mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X)\tilde{f}_{\text{Bayes}}(X) < 0\}} \mathbb{1}_{\{\tilde{f}_{\text{Bayes}}(X) > 0\}} dP_X(x) \\
&= \mathbb{E}_X \left(|2\eta(X) - 1| \mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X)\tilde{f}_{\text{Bayes}}(X) < 0\}} \right).
\end{aligned}$$

Since \tilde{f}_{Bayes} is a linear combination of two density functions, we have the following uniform convergence result due to Theorem 2 of Jiang [2017].

Lemma 2.7.4 (Theorem 2 of Jiang [2017]). *Suppose $p_X(x)$ and $p_{X|Y=1}(x)$ are α_H -Hölder continuous. Then there exist a constant $C_{(p,u)}$ such that the following holds with probability at least $1 - 1/n_p - 1/n_u$.*

$$\left\| \tilde{f}_{\mathcal{H}_k} - \tilde{f}_{\text{Bayes}} \right\|_{\infty} \leq 4\pi_+ C_{(p,u)} (n_p \wedge n_u)^{-\frac{\alpha_H}{2\alpha_H+d}}, \quad (2.14)$$

where the bandwidth $h = (n_p \wedge n_u)^{-\frac{1}{2\alpha_H+d}}$.

Denote $\varepsilon(n_p, n_u) := 4\pi_+ C_{(p,u)} (n_p \wedge n_u)^{-\frac{\alpha_H}{2\alpha_H+d}}$. Let E be the event that Equation (2.14) holds. Under the event E , $\mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X)\tilde{f}_{\text{Bayes}}(X) < 0\}} \leq \mathbb{1}_{\{|\tilde{f}_{\text{Bayes}}(X)| \leq \varepsilon(n_p, n_u)\}}$. Thus, we have

$$\mathbb{E}_X \left(|2\eta(X) - 1| \mathbb{1}_{\{\tilde{f}_{\mathcal{H}_k}(X)\tilde{f}_{\text{Bayes}}(X) < 0\}} \right)$$

$$\begin{aligned}
&\leq \mathbb{E}_X \left(|2\eta(X) - 1| \mathbb{1}_{\{|\hat{f}_{\text{Bayes}}(X)| \leq \varepsilon(n_p, n_u)\}} \right) \\
&\leq \mathbb{E}_X \left(|2\eta(X) - 1| \mathbb{1}_{\{|2\eta(X) - 1| \leq \frac{\varepsilon(n_p, n_u)}{p_{\min}}\}} \right) \\
&\leq \frac{\varepsilon(n_p, n_u)}{p_{\min}} P_X \left(|2\eta(X) - 1| \leq \frac{\varepsilon(n_p, n_u)}{p_{\min}} \right) \\
&\leq \tilde{C} (n_p \wedge n_u)^{-\frac{\alpha_H(1+q)}{2\alpha_H+d}},
\end{aligned}$$

for some constant $\tilde{C} > 0$. Thus, under the assumptions (A1)-(A4), we have the following with probability at least $1 - 1/n_p - 1/n_u$.

$$R_{\ell_{01}}(\hat{f}_{\mathcal{H}_{k,r_1}}) - \inf_{f \in \mathcal{U}} R_{\ell_{01}}(f) \leq \tilde{C} (n_p \wedge n_u)^{-\frac{\alpha_H(1+q)}{2\alpha_H+d}}.$$

□

2.7.4 Implementation details

In this section, we provide implementation details for WMMD and other baseline PU learning algorithms.

Proposed PU learning algorithm: WMMD algorithm

We divided the original training data into training and validation sets, with an 80-20 random split. Let \tilde{x}_j^{p} and \tilde{x}_j^{u} be the positive and the unlabeled samples in validation set, respectively. Similarly, \tilde{n}_p and \tilde{n}_u be the number of samples in the positive and the unlabeled validation set, respectively. With the validation set, we conducted a grid search method for the hyperparameter selection with a grid $\gamma \in \{1, 0.4, 0.2, 0.1, 0.05\}$ for all the numerical experiments. With the grids, we selected the optimal hyperparameters

γ^* which minimized

$$\begin{aligned} \hat{L}(\gamma) &:= -\pi_+ + \frac{2\pi_+}{\tilde{n}_p} \sum_{j=1}^{\tilde{n}_p} \mathbb{1} \left[\text{sign}(\hat{f}_{\mathcal{H}_{k,r}}^\gamma(\tilde{x}_j^p)) \neq 1 \right] \\ &\quad + \frac{1}{\tilde{n}_u} \sum_{j=1}^{\tilde{n}_u} \mathbb{1} \left[\text{sign}(\hat{f}_{\mathcal{H}_{k,r}}^\gamma(\tilde{x}_j^u)) \neq -1 \right], \end{aligned} \quad (2.15)$$

where

$$\text{sign}\{\hat{f}_{\mathcal{H}_{k,r}}^\gamma(z)\} = \begin{cases} +1 & \text{if } (2\pi_+)^{-1} < \hat{\lambda}_{n_p, n_u}^\gamma(z) \\ -1 & \text{otherwise} \end{cases},$$

$$\hat{\lambda}_{n_p, n_u}^\gamma(z) = \frac{n_p^{-1} \sum_{i=1}^{n_p} k_\gamma(z, x_i^p)}{n_u^{-1} \sum_{i=1}^{n_u} k_\gamma(z, x_i^u)},$$

and $k_\gamma(z_1, z_2) = \exp(-\gamma \|z_1 - z_2\|_2^2)$. Note that Equation (2.15) is an empirical estimation of the misclassification error since

$$\begin{aligned} &P_{X,Y}(f(X)Y < 0) \\ &= \pi_+ P_{X|Y=1}(\text{sign}(f(X)) = -1) + (1 - \pi_+) P_{X|Y=-1}(\text{sign}(f(X)) = 1) \\ &= \pi_+ P_{X|Y=1}(\text{sign}(f(X)) = -1) \\ &\quad + (P_X(\text{sign}(f(X)) = 1) - \pi_+ P_{X|Y=1}(\text{sign}(f(X)) = 1)) \\ &= -\pi_+ + 2\pi_+ P_{X|Y=1}(\text{sign}(f(X)) = -1) + P_X(\text{sign}(f(X)) = 1). \end{aligned}$$

Final classification for a test datum z was determined by $\text{sign}(\hat{f}_{\mathcal{H}_{k,r}}^{\gamma^*}(z))$ and AUC was computed by using $\hat{\lambda}_{n_p, n_u}^{\gamma^*}(z)$.

When the class-prior π_+ is unknown, we suggest a simple π_+ estimation method, called the density-based method, to find $\hat{\pi}_+^{\text{WMMD}}$ such that for some $\eta \in (0, 1)$,

$$\hat{\pi}_+^{\text{WMMD}}$$

$$:= \sup \left\{ t \in (0, 1) \mid \frac{\#\{j \in \{1, \dots, \tilde{n}_p\} \mid \{\hat{\lambda}_{n_p, n_u}^\gamma(\tilde{x}_j^p)\}^{-1} \leq t\}}{\tilde{n}_p} \leq \eta \right\}.$$

This estimator is sensible because $\pi_+ \leq p_X(x)/p_{X|Y=1}(x)$ and $\{\hat{\lambda}_{n_p, n_u}^\gamma(x)\}^{-1}$ can be considered as a kernel density estimation of $p_X(x)/p_{X|Y=1}(x)$ for $x \in \text{supp}(P_{X|Y=1})$. Here, we denote the density functions by p_X and $p_{X|Y=1}$. In our experiments, we fix $\eta = 0.1$ and using $\hat{\pi}_+^{\text{WMMD}}$ led to better performance than using ‘KM1’ method in terms of accuracy.

The baseline PU learning algorithms

We compared the following 4 PU learning algorithms: (i) the logistic loss ℓ_{\log} , denoted by LOG, (ii) the double hinge loss ℓ_{dh} , denoted by DH, both proposed by du Plessis et al. [2015], (iii) the non-negative risk estimator method, denoted by NNPU, proposed by Kiryo et al. [2017], and (iv) the threshold adjustment method, denoted by tADJ, proposed by Elkan and Noto [2008].

General: Similar to the WMMD procedure, we set training and validation sets with the 80-20 random split of the original training dataset and we conducted a grid search method for hyperparameter selection.

LOG and DH: As du Plessis et al. [2015] proposed, we followed a binary discriminant function as

$$g_{\alpha, b}(x) = \sum_{i=1}^N \alpha_i \varphi_i(x) + b = \alpha^T \varphi(x) + b,$$

where $N = n_p + n_u$ and $\varphi_i(x) = \exp(-\gamma \|x - c_i\|_2^2)$ for $\{c_1, \dots, c_N\} = \{x_1^p, \dots, x_{n_p}^p, x_1^u, \dots, x_{n_u}^u\}$. For a loss function $\ell \in \{\ell_{\log}, \ell_{\text{dh}}\}$, the

empirical risk function is given by

$$\begin{aligned} & \hat{J}_{\lambda, \gamma}(\alpha, b) \\ &= -\frac{\pi_+}{n_p} \sum_{i=1}^{n_p} \alpha^T \varphi(x_i^p) - \pi_+ b + \frac{1}{n_u} \sum_{i=1}^{n_u} \ell(-\alpha^T \varphi(x_i^u) - b) + \frac{\lambda}{2} \alpha^T \alpha. \end{aligned}$$

Here, the hyperparameter grids are $\lambda \in \{1, 0.4, 0.2, 0.1, 0.05\}$ and $\gamma \in \{1, 0.4, 0.2, 0.1, 0.05\}$ for all the numerical experiments. With the grids, we selected the optimal hyperparameter (λ^*, γ^*) which minimized the empirical risk on the validation set $\tilde{J}_{\lambda, \gamma}(\alpha, b)$ defined by

$$\begin{aligned} & \tilde{J}_{\lambda, \gamma}(\alpha, b) \\ &= -\frac{\pi_+}{\tilde{n}_p} \sum_{i=1}^{\tilde{n}_p} \alpha^T \varphi(\tilde{x}_i^p) - \pi_+ b + \frac{1}{\tilde{n}_u} \sum_{i=1}^{\tilde{n}_u} \ell(-\alpha^T \varphi(\tilde{x}_i^u) - b) + \frac{\lambda}{2} \alpha^T \alpha. \end{aligned}$$

After selecting the optimal hyperparameter (λ^*, γ^*) , we minimized $\hat{J}_{(\lambda^*, \gamma^*)}(\alpha, b)$ with the gradient descent algorithm. Learning rate was fixed by 0.1 and the number of epochs was 100. During the training, we applied the early stopping rule: we stopped training if the validation error is not minimized in 10 successive epochs. After the training phase, with the trained $\hat{\alpha}$ and \hat{b} , we classified a test datum z as $\text{sign}(g_{\hat{\alpha}, \hat{b}}(z))$. AUC was computed by using $g_{\hat{\alpha}, \hat{b}}(z)$.

NNPU: We followed the method by Kiryo et al. [2017]. The model for NNPU was a 5-layer multilayer perceptron with ReLU nonlinearity (d -300-300-300-1). We applied the batch normalization before each ReLU nonlinearity. Please note that this network architecture is quite similar to the model in Kiryo et al. [2017]. We used a stochastic gradient descent algorithm with a learning rate 0.01. Loss function was the sigmoid function. The number of

epochs was 100 and the optimal weights were selected at the best validation error during the training.

tADJ: We followed the method by Elkan and Noto [2008]. We used ‘LogisticRegressionCV’ function in the Python module ‘sklearn.linear_model’ [Pedregosa et al., 2011] to estimate $P(\{x \text{ is from the positive dataset} \mid X = x\})$. The hyperparameter grid for L_2 -regularizer was $\{0.01, 0.1, 1, 10, 100\}$ and the optimal hyperparameter was chosen based on 5-fold cross validation on the split training dataset, *i.e.*, 80% of the original training dataset. Then, $P(\{x \text{ is from the positive dataset} \mid Y = 1\})$ was estimated by the split validation set, *i.e.*, 20% of the original training dataset.

2.7.5 Comparison between Gaussian and inverse kernels

We compared LOG, DH, and WMMD using two kernels: (i) the Gaussian kernel $k(x, y) = \exp(-\gamma\|x - y\|_2^2)$ and (ii) the inverse kernel $k(x, y) = \frac{\gamma}{\gamma + \|x - y\|_2^2}$ for $\gamma > 0$. Figure 2.4 shows the accuracy and AUC of LOG, DH, and WMMD on various n_u and π_+ . For the two top figures, the training sample size for the positive data is $n_p = 100$ and the class prior is $\pi_+ = 0.5$. The unlabeled sample size changes from 40 to 500 by 20. We repeat a random generation of training and test data 100 times. For comparison purposes, we add the 1–Bayes risk for each unlabeled sample size. In every algorithm, using the Gaussian kernel achieves higher accuracy and AUC than using the inverse kernel in every n_u . For the two bottom figures, the training sample size for the positive and the unlabeled data are $n_p = 100$ and $n_u = 400$, respectively.

The class-prior π_+ changes from 0.05 to 0.95 by 0.05. The test sample size is 10^3 . We repeat a random generation of training and test data 100 times. Both kernels perform comparably for LOG and WMMD algorithm. The DH algorithm with the inverse kernel achieves higher accuracy when the class-prior is close to 0.5. But in terms of AUC, both kernels perform comparably in every π_+ .

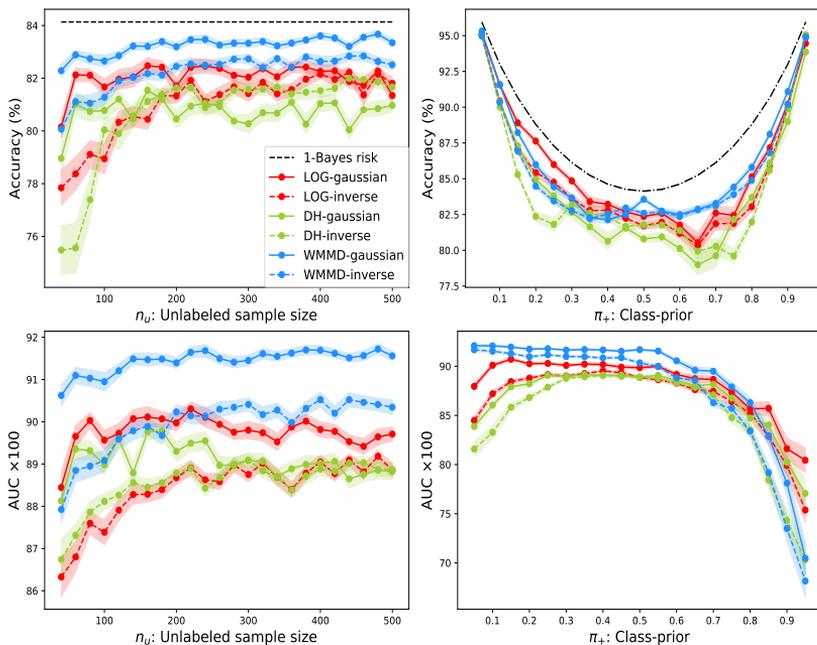


Figure 2.4: The comparison of the accuracy and AUC of the LOG, DH, and WMMD algorithms with the Gaussian and the inverse kernels when each of n_u and π_+ changes. The black dashed-dotted curve represents the 1–Bayes risk. The algorithms with the Gaussian and inverse kernel are displayed with dashed and solid lines, respectively. The curve and the shaded region represent the average and the standard error, respectively, based on 100 replications.

Chapter 3

Principled learning with augmented data: WDRO perspective

In this chapter, we build grounds for regularized risk minimization with augmented data in the context of Wasserstein distributionally robust optimization (WDRO).

3.1 Backgrounds

We first provide backgrounds on statistical learning theory, distributionally robust optimization, and data augmentation.

3.1.1 Statistical learning theory and distributionally robust optimization

Statistical learning theory provides a framework for learning models based on finite observations. A general formulation of learning is to specify our learning goal as a loss, and then to minimize the expected value of the loss [Vapnik, 1999]. Concretely, let \mathcal{Z} be a subset of \mathbb{R}^p for an integer p . Let $\mathcal{P}(\mathcal{Z})$ be a set of Borel probability measures defined on \mathcal{Z} and $\mathbb{P}_{\text{data}} \in \mathcal{P}(\mathcal{Z})$ be the underlying data distribution. Let $\mathcal{H} \subseteq \{h \mid h : \mathcal{Z} \rightarrow \mathbb{R}\}$ be a set of loss functions and denote the expected loss, called the risk, by $R(\mathbb{Q}, h) := \int_{\mathcal{Z}} h(\zeta) d\mathbb{Q}(\zeta)$ for $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$, $h \in \mathcal{H}$. In such settings, many statistical learning problems can be expressed by an optimization problem as follows:

$$\inf_{h \in \mathcal{H}} R(\mathbb{P}_{\text{data}}, h). \quad (3.1)$$

In applications, the underlying distribution \mathbb{P}_{data} is usually unknown and the exact computation of the risk in (3.1) is intractable. Instead we observe a set $\mathcal{Z}_n = \{z_1, \dots, z_n\}$ of independently identically distributed samples from \mathbb{P}_{data} . Given the dataset \mathcal{Z}_n , the empirical risk minimization (ERM) principle replaces \mathbb{P}_{data} in (3.1) with the empirical data distribution \mathbb{P}_n , where $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{z_i}$ and δ_z is the Dirac delta distribution concentrating unit mass at $z \in \mathcal{Z}$. Then, ERM can be represented as

$$\inf_{h \in \mathcal{H}} R(\mathbb{P}_n, h) = \inf_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n h(z_i). \quad (3.2)$$

The replacement \mathbb{P}_{data} with \mathbb{P}_n often yields a poor risk estimation and thus a solution of (3.2) can have a small training error but a

big test error. This phenomenon is known as the overfitting and it is frequently observed when data are high-dimensional [Bellman, 1961]. To reduce the generalization error, a great number of regularization methods have been proposed such as penalty-based methods [Tibshirani, 1996, Fan and Li, 2001], dropout [Wager et al., 2013], and early stopping [Yao et al., 2007].

Distributionally robust optimization (DRO) is an alternative approach preventing the overfitting, where the goal is to learn a model that minimizes the worst-case risk. To be more specific, let $\mathcal{A}(\mathbb{P}_n) \subseteq \mathcal{P}(\mathcal{Z})$ be an ambiguity set constructed from \mathbb{P}_n . The worst-case risk $\sup_{\mathbb{Q} \in \mathcal{A}(\mathbb{P}_n)} R(\mathbb{Q}, h)$ is defined by the supremum of risks over the $\mathcal{A}(\mathbb{P}_n)$. Then, DRO can be expressed as follows.

$$\inf_{h \in \mathcal{H}} \sup_{\mathbb{Q} \in \mathcal{A}(\mathbb{P}_n)} R(\mathbb{Q}, h). \quad (3.3)$$

If $\mathcal{A}(\mathbb{P}_n)$ is large enough to satisfy $\mathbb{P}_{\text{data}} \in \mathcal{A}(\mathbb{P}_n)$, the solution of (3.3), say h_{DRO} , is guaranteed to have the small risk, *i.e.*, $R(\mathbb{P}_{\text{data}}, h_{\text{DRO}}) \leq \sup_{\mathbb{Q} \in \mathcal{A}(\mathbb{P}_n)} R(\mathbb{Q}, h_{\text{DRO}})$. This encourages a careful choice of $\mathcal{A}(\mathbb{P}_n)$ such that $\mathbb{P}_{\text{data}} \in \mathcal{A}(\mathbb{P}_n)$. A common practice is to define $\mathcal{A}(\mathbb{P}_n)$ by a neighborhood of \mathbb{P}_n with distances on probability measures. For examples, the f -divergence [Ben-Tal et al., 2013, Hu et al., 2016, Namkoong and Duchi, 2017], the maximum mean discrepancy [Staib and Jegelka, 2019], and the Wasserstein distance [Sinha et al., 2017, Esfahani and Kuhn, 2018] are studied in the literature.

3.1.2 Data augmentation by linear interpolation

In this subsection, we introduce data augmentation by linear interpolation.

Definition 1. Given the dataset \mathcal{Z}_n , $0 \leq \gamma \leq 1$, and $z \in \mathcal{Z}_n$, we denote the index set by $I_z := \{j \in [n] \mid z_j \in \mathcal{Z}_n \setminus \{z\}\}$. We say z' is a (γ, z) -interpolated datum if z' can be expressed as

$$z' = \gamma z + \sum_{j \in I_z} \gamma_j z_j,$$

for some $\gamma_j \geq 0$ such that $\gamma + \sum_{j \in I_z} \gamma_j = 1$.

Throughout this paper, we use the prime (\prime) notation for augmented data.

Example 1 (Mixup). Suppose $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{Z}_n = \{(x_i, y_i)\}_{i=1}^n$. Let (x', y') be a Mixup datum given by

$$\begin{aligned} x' &= \gamma x + (1 - \gamma)\tilde{x} \\ y' &= \gamma y + (1 - \gamma)\tilde{y}, \end{aligned}$$

for some $(x, y), (\tilde{x}, \tilde{y}) \in \mathcal{Z}_n$ and mixing rate $0 \leq \gamma \leq 1$. Then, the Mixup datum (x', y') is a $(\gamma, (x, y))$ -interpolated datum. We denote $(x', y') = \text{Mixup}((x, y); \gamma)$.

Notations Let $\|\cdot\|$ be a norm on \mathcal{Z} . We denote a dual space by \mathcal{Z}^* and a dual norm by $\|u\|_* := \sup_{\|v\| \leq 1} u^T v$ for $v \in \mathcal{Z}, u \in \mathcal{Z}^*$. Given measurable spaces \mathcal{Z} and $\tilde{\mathcal{Z}}$ and a measurable function $T : \mathcal{Z} \rightarrow \tilde{\mathcal{Z}}$, we denote the push-forward measure of $\mu \in \mathcal{P}(\mathcal{Z})$ through T by $T\#\mu \in \mathcal{P}(\tilde{\mathcal{Z}})$. We denote a set of loss functions by \mathcal{H} . If a loss $h \in \mathcal{H}$ is a Lipschitz continuous function, we denote its Lipschitz constant by $\text{Lip}(h)$. For $n \in \mathbb{N}$, we set $[n] := \{1, \dots, n\}$. Lastly, for nonzero sequences (a_n) and (b_n) , $a_n = o(b_n)$ indicates $\lim_{n \rightarrow \infty} a_n/b_n = 0$.

3.2 Wasserstein distributionally robust optimization

In this section, we introduce Wasserstein distributionally robust optimization (WDRO) and briefly review some existing results regarding WDRO. We first define Wasserstein distance and Wasserstein ball.

Definition 2 (Wasserstein distance and Wasserstein ball). *For ν and $\mu \in \mathcal{P}(\mathcal{Z})$, the Wasserstein distance between ν and μ is defined by*

$$\mathcal{W}(\nu, \mu) := \inf_{\rho \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z})} \left\{ \int_{\mathcal{Z} \times \mathcal{Z}} \|\zeta - \tilde{\zeta}\| d\rho(\zeta, \tilde{\zeta}) \mid \pi_1 \# \rho = \nu, \pi_2 \# \rho = \mu \right\},$$

where $\pi_i : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{Z}$ is the canonical projection defined by $\pi_i(\zeta_1, \zeta_2) = \zeta_i$ for $i = 1, 2$. For $\alpha > 0$, the Wasserstein ball centered at \mathbb{P}_n with radius α is defined by

$$\mathfrak{M}_n(\alpha) := \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : \mathcal{W}(\mathbb{Q}, \mathbb{P}_n) \leq \alpha\}.$$

In this thesis, we consider the WDRO problem defined by the DRO problem with the Wasserstein ball. More precisely, by plugging the Wasserstein ball $\mathfrak{M}_n(\alpha)$ into an ambiguity set $\mathcal{A}(\mathbb{P}_n)$ in Equation (3.3) we consider the following problem

$$\inf_{h \in \mathcal{H}} \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha)} R(\mathbb{Q}, h). \quad (3.4)$$

Although the problem (3.4) involves a supremum over infinitely many distributions, which makes it difficult to solve in general,

Gao and Kleywegt [2016] derived the following strong duality reformulation.

$$\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha)} R(\mathbb{Q}, h) = \min_{\lambda \geq 0} \left\{ \lambda \alpha + \frac{1}{n} \sum_{i=1}^n \sup_{z \in \mathcal{Z}} \{h(z) - \lambda \|z - z_i\|\} \right\}. \quad (3.5)$$

Equivalent results are shown in literature [Blanchet et al., 2016, Esfahani and Kuhn, 2018]. Furthermore, based on Equation (3.5), Shafieezadeh-Abadeh et al. [2017] and Gao et al. [2017] established the relationships between (3.3) and penalty-based methods.

3.3 Principled learning with augmented data in the context of WDRO

In this section, we first show that regularized empirical risk with augmented data approximates the worst-case risk when a loss function has a Hölder continuous gradient. Throughout this section, we assume that \mathcal{Z} is compact.

Theorem 3.3.1. *Assume that $h : \mathcal{Z} \rightarrow \mathbb{R}$ is differentiable and there exists a constant $k \in (0, 1]$ and a constant $0 < C_1 < \infty$ such that*

$$\|\nabla_z h(z) - \nabla_z h(\tilde{z})\|_* \leq C_1 \|z - \tilde{z}\|^k, \quad \forall z, \tilde{z} \in \mathcal{Z}.$$

Let (α_n) be a decreasing sequence such that $\alpha_n = o(n^{-1})$. Then, the following holds almost surely.

$$\left| \text{Lip}(h)\alpha_n + \frac{1}{n} \sum_{i=1}^n \{h(z'_i) + \text{Lip}(h)\|z_i - z'_i\|\} - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) \right|$$

$$= o(\alpha_n),$$

where z'_i is any $((1 - C_2\alpha_n^2), z_i)$ -interpolated datum for all $i \in [n]$ and for some constant $C_2 > 0$.

A proof is given in Section 3.6.1. In Theorem 3.3.1, we show that the Lipschitz regularized empirical risk approximates to the worst-case risk when the loss function has Hölder continuous gradient. Esfahani and Kuhn [2018] and Gao et al. [2017] obtained the similar results when data are not augmented, *i.e.*, $\tilde{z}_i = z_i$ for all $i \in [n]$. In addition, they consider a linear hypothesis and the loss h is a composition of a univariate convex function and the hypothesis.¹ In contrast, Theorem 3.3.1 relaxes the convexity of h , which is more reasonable assumption in the machine learning literature.

We propose to minimize the approximation given by

$$\begin{aligned} R_{\alpha_n, \text{prop}}(\mathbb{P}_n, h) \\ := \text{Lip}(h)\alpha_n + \frac{1}{n} \sum_{i=1}^n \{h(z'_i) + \text{Lip}(h)\|z_i - z'_i\|\}. \end{aligned} \quad (3.6)$$

We analyze a minimizer of the objective function (3.6) in the following theorem. To begin, we denote the Rademacher complexity of \mathcal{F} given a set $\{z_1, \dots, z_m\}$ by

$$\mathfrak{R}_n(\mathcal{F}) := \mathbb{E}_{\text{data}} \mathbb{E}_{\sigma} \left(\frac{1}{m} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^m \sigma_i f(z_i) \right| \right),$$

¹Gao et al. [2017, Theorem 2] suggested the similar result for the smooth loss h but the proof is not necessarily true when the 1-Wasserstein distance is considered.

where $\{\sigma_i\}_{i=1}^m$ is a set of independent Rademacher random variables taking 1 or -1 with probability 0.5 each and $\mathbb{E}_\sigma(\cdot)$ is the expectation operator over the Rademacher random variables [Bartlett and Mendelson, 2002].

Theorem 3.3.2. *Suppose that \mathcal{H} be a set of differentiable functions $h : \mathcal{Z} \rightarrow \mathbb{R}$ such that its gradient $\nabla_z h$ is Hölder continuous as in Theorem 3.3.1 and $\text{Lip}(h) \leq L_{\mathcal{H}} < \infty$. Let $\hat{h}_{n,\text{prop}} \in \mathcal{H}$ be a minimizer of $R_{\alpha_n, \text{prop}}(\mathbb{P}_n, h)$ and (α_n) be a decreasing sequence as in Theorem 3.3.1. Then, the following holds with probability at least $1 - \delta$:*

$$\begin{aligned} & R(\mathbb{P}_{\text{data}}, \hat{h}_{n,\text{prop}}) - \inf_{h \in \mathcal{H}} R(\mathbb{P}_{\text{data}}, h) \\ & \leq 4\mathfrak{R}_n(\mathcal{H}) + 2C_{\mathcal{H}} \sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)} + o(n^{-1}), \end{aligned}$$

where $C_{\mathcal{H}}$ is a constant such that $\sup_{h \in \mathcal{H}} \|h\|_{\infty} \leq C_{\mathcal{H}}$.

A proof is given in Section 3.6.2. Theorem 3.3.2 shows that the proposed model has a risk consistency when elements in \mathcal{H} have a Lipschitz constant smaller than $L_{\mathcal{H}}$.

3.4 Numerical experiments

In this section, we empirically analyze the proposed algorithm to demonstrate its practical efficacy using the two image classification benchmark datasets. We conduct three numerical experiments: (1) comparing the accuracy on noisy datasets with two baseline methods, (2) analyzing performance changes when noise intensity increases, and (3) analyzing effects of regularization parameter. Implementation details are available in Section 3.6.3



Figure 3.1: Example of clean and noisy images. Proportions of noisy pixels are (top left) 0%, (top right) 0.5%, (bottom left) 1%, and (bottom right) 2%.

Datasets: We use CIFAR-10, and CIFAR-100, and randomly sample for the 5 independent training sets keeping the number of images in each label equally. We use the original test set consisting of 10000 images. For noisy images, we apply the salt and pepper noise on 0.5%, 1% and 2% of pixels. Figure 3.1 displays an example of the clean and the noisy test images.

Experiment settings: To evaluate the distributional robustness, we train the model with clean images and then evaluate accu-

racy on both clean and noisy test images. If the model is distributionally robust, it should perform well on other data distributions apart from that of training data. Therefore, we report the accuracy on the noisy test images to evaluate the distributional robustness of the model.

For comparison study, we consider 2 training methods as baseline methods: (i) the empirical risk minimization, denoted by ERM, (ii) the empirical risk minimization with Mixup data, denoted by MIXUP. We denote the proposed method by DROID. Comparing ERM and MIXUP with DROID, we demonstrate efficacy of the proposed method regularizing the gradient.

Experiment 1: In this experiment, we demonstrate the distributional robustness of the proposed method by evaluating accuracy on clean and noisy images. The number of training sample size varies as 2500, 5000, 25000 and 50000, and the regularization parameter λ_{grad} used in DROID is 0.004. We repeat the evaluation with five independent training sets and report the average and standard deviation of the accuracies.

Table 3.1 shows the accuracy of the three methods on clean and noisy data with various training sample sizes. For clean data, DROID performs better than ERM while showing a slight disadvantage over MIXUP. However, for noisy data, DROID achieves higher accuracy than both ERM and MIXUP in any cases, and MIXUP shows lower accuracy than ERM when the sample size is 25000 and 50000. This shows that the distributional robustness appears more clearly only when training with Mixup data while regularizing the gradient of loss function. Without the gradient

Table 3.1: Accuracy comparison using the clean and noisy test datasets. The salt and pepper noise is applied for 1% of pixels. Average and standard deviation are denoted by ‘average \pm standard deviation’. All the results are based on five independent trials. Boldface numbers denote the best and equivalent methods with respect to a t-test with a significance level of 5%.

| Sample size | Clean data | | | Noisy data | | |
|-------------|----------------|----------------------------------|----------------------------------|----------------|----------------------------------|----------------------------------|
| | ERM | MIXUP | DROID | ERM | MIXUP | DROID |
| CIFAR-10 | | | | | | |
| 2500 | 77.3 \pm 0.8 | 81.4 \pm 0.5 | 80.8 \pm 0.7 | 69.5 \pm 1.8 | 72.7 \pm 1.9 | 74.8 \pm 1.0 |
| 5000 | 83.3 \pm 0.4 | 86.7 \pm 0.2 | 85.6 \pm 0.3 | 75.2 \pm 1.2 | 76.4 \pm 1.3 | 79.5 \pm 0.7 |
| 25000 | 92.2 \pm 0.2 | 93.3 \pm 0.1 | 92.4 \pm 0.1 | 83.2 \pm 1.1 | 82.2 \pm 1.6 | 86.3 \pm 0.3 |
| 50000 | 94.1 \pm 0.1 | 94.8 \pm 0.2 | 93.5 \pm 0.2 | 83.9 \pm 1.1 | 82.6 \pm 1.1 | 87.4 \pm 0.5 |
| CIFAR-100 | | | | | | |
| 2500* | 33.8 \pm 1.0 | 38.9 \pm 0.6 | 39.4 \pm 0.2 | 29.2 \pm 0.3 | 33.4 \pm 0.8 | 34.6 \pm 0.3 |
| 5000* | 45.2 \pm 0.9 | 49.9 \pm 0.2 | 49.5 \pm 0.4 | 37.0 \pm 1.0 | 39.5 \pm 1.2 | 42.5 \pm 0.7 |
| 25000 | 67.8 \pm 0.2 | 69.3 \pm 0.3 | 68.2 \pm 0.3 | 51.1 \pm 2.0 | 49.5 \pm 1.2 | 56.0 \pm 0.4 |
| 50000 | 74.4 \pm 0.2 | 75.2 \pm 0.2 | 73.8 \pm 0.3 | 51.8 \pm 1.7 | 49.8 \pm 2.9 | 60.7 \pm 0.8 |

regularization, training with large Mixup data might not give a distributionally robust model. We can see that the distributional robustness can be demonstrated with the proper size of Mixup data and proper regularization on the gradient of loss function.

Experiment 2: In this experiment, we compare the distributional robustness of the three methods as noise level changes. In DROID, we choose the regularization parameter as 0.004. We train each model with the original clean training images and evaluate accuracy using 10000 noisy test images. We apply the salt and pepper noise and the applied noise intensities are 0.5%, 1%, and 2%.

Table 3.2: The comparison of the accuracy of the three methods with the three different noisy intensities. Other details are given in Table 3.1.

| Dataset | Probability of noisy pixels | ERM | MIXUP | DROID |
|-----------|--------------------------------|----------------|----------------|----------------------------------|
| CIFAR-10 | 0.5% | 89.7 ± 0.3 | 89.5 ± 0.8 | 90.6 ± 0.4 |
| | 1% | 83.9 ± 1.1 | 82.6 ± 1.1 | 87.4 ± 0.5 |
| | 2% | 72.9 ± 2.3 | 70.3 ± 1.6 | 80.8 ± 1.0 |
| CIFAR-100 | 0.5% | 64.4 ± 0.9 | 63.7 ± 1.3 | 68.3 ± 0.5 |
| | 1% | 51.8 ± 1.7 | 49.8 ± 2.9 | 60.7 ± 0.8 |
| | 2% | 31.5 ± 2.1 | 29.4 ± 4.1 | 44.1 ± 0.9 |

Table 3.1 shows the accuracy of the three methods on noisy images as the noise intensity changes. Compared with ERM and MIXUP, DROID achieves the highest accuracy at all noise levels, and accuracy difference becomes larger as the noise level increases. In addition, MIXUP performs slightly worse than ERM in all comparisons. This shows that DROID is more distributionally robust than ERM and MIXUP, while MIXUP is vulnerable to noise.

Experiment 3: In this experiment, we analyze the effect of regularization parameter λ_{grad} on distributional robustness. To see the effect of regularizing the gradient of the loss function, we considered MIXUP and DROID with $\lambda_{\text{grad}} = 0.001, 0.002, 0.004, 0.008$. Note that MIXUP can be considered as an extreme case of DROID with $\lambda_{\text{grad}} = 0$. We train each model with the original clean training images and evaluate accuracy using 10000 noisy test images. We apply the salt and pepper noise on 1% of pixels.

Table 3.3: The comparison of the accuracy of MIXUP and DROID with the four different regularization parameter values λ_{grad} . Other details are given in Table 3.1.

| Dataset | MIXUP | DROID with λ_{grad} | | | |
|-----------|----------------|------------------------------------|----------------|----------------------------------|----------------------------------|
| | | 0.001 | 0.002 | 0.004 | 0.008 |
| CIFAR-10 | 82.6 ± 1.1 | 86.6 ± 0.7 | 86.7 ± 0.5 | 87.4 ± 0.5 | 87.8 ± 0.2 |
| CIFAR-100 | 49.8 ± 2.9 | 57.8 ± 1.0 | 59.4 ± 0.9 | 60.7 ± 0.8 | 62.0 ± 0.5 |

Table 3.2 shows the accuracy of MIXUP and DROID with the four different regularization parameters. As the regularization parameter increases, the accuracy tends to increase. This shows that regularizing gradient of the loss function leads the model to be robust.

3.5 Concluding remarks

Existing state-of-the-art methods heavily depend on data augmentation techniques [Cubuk et al., 2018, Lim et al., 2019], and it demands for a deeper understanding of learning with augmented data. In this work, we build grounds for learning models with augmented data. We show that minimizing regularized empirical risk evaluated with augmented data can be interpreted as solving WDRO. This is the first rigorous method to use augmented data and deep neural networks in WDRO.

3.6 Appendix

In this section, we provide proofs of Theorems 3.3.1 and 3.3.2, and implementation details.

3.6.1 Proof of Theorem 3.3.1

First, we provide an upper bound of the worst-case risk in the following lemma.

Lemma 3.6.1. *Suppose $h : \mathcal{Z} \rightarrow \mathbb{R}$ is Lipschitz continuous and $\alpha \leq n^{-1} \sum_{i=1}^n \sup_{z \in \mathcal{Z}} \|z - z_i\|$. Then, we have*

$$\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha)} R(\mathbb{Q}, h) \leq \text{Lip}(h)\alpha + \frac{1}{n} \sum_{i=1}^n \{h(\tilde{z}_i) + \text{Lip}(h)\|z_i - \tilde{z}_i\|\},$$

for any $\tilde{z}_1, \dots, \tilde{z}_n \in \mathcal{Z}$.

Proof of Lemma 3.6.1. Note that for any $z_i \in \mathcal{Z}_n$ and $z, \tilde{z} \in \mathcal{Z}$, we obtain

$$\begin{aligned} h(z) - \lambda\|z - z_i\| &\leq h(\tilde{z}_i) + \text{Lip}(h)\|z - \tilde{z}_i\| - \lambda\|z - z_i\| \\ &\leq h(\tilde{z}_i) + \text{Lip}(h)(\|z - z_i\| + \|z_i - \tilde{z}_i\|) - \lambda\|z - z_i\| \\ &= h(\tilde{z}_i) + \text{Lip}(h)\|z_i - \tilde{z}_i\| + (\text{Lip}(h) - \lambda)\|z - z_i\|. \end{aligned}$$

The first inequality is from Lipschitz continuity of h and the second inequality is from triangle inequality. Therefore,

$$\begin{aligned} &\sup_{z \in \mathcal{Z}} (h(z) - \lambda\|z - z_i\|) \\ &\leq h(\tilde{z}_i) + \text{Lip}(h)\|z_i - \tilde{z}_i\| + \sup_{z \in \mathcal{Z}} \{(\text{Lip}(h) - \lambda)\|z - z_i\|\} \\ &= \begin{cases} h(\tilde{z}_i) + \text{Lip}(h)\|z_i - \tilde{z}_i\| & \text{for } \lambda \geq \text{Lip}(h), \\ h(\tilde{z}_i) + \text{Lip}(h)\|z_i - \tilde{z}_i\| + (\text{Lip}(h) - \lambda)\|z_i^* - z_i\| & \text{for } \lambda < \text{Lip}(h), \end{cases} \end{aligned}$$

where $z_i^* = \sup_{z \in \mathcal{Z}} \|z - z_i\|$. Set $\bar{D} = n^{-1} \sum_{i=1}^n \|z_i^* - z_i\|$. Then,

$$\begin{aligned}
& \min_{\lambda \geq 0} \left\{ \lambda \alpha + \frac{1}{n} \sum_{i=1}^n \sup_{z \in \mathcal{Z}} (h(z) - \lambda \|z - z_i\|) \right\} \\
& \leq \min_{\lambda \geq 0} \left\{ \lambda \alpha + \frac{1}{n} \sum_{i=1}^n \left\{ h(\tilde{z}_i) + \text{Lip}(h) \|z_i - \tilde{z}_i\| \right. \right. \\
& \quad \left. \left. + \sup_{z \in \mathcal{Z}} \{ (\text{Lip}(h) - \lambda) \|z - z_i\| \} \right\} \right\} \\
& = \frac{1}{n} \sum_{i=1}^n \{ h(\tilde{z}_i) + \text{Lip}(h) \|z_i - \tilde{z}_i\| \} \\
& \quad + \min \left\{ \text{Lip}(h) \alpha, \min_{\lambda < \text{Lip}(h)} \{ \lambda \alpha + (\text{Lip}(h) - \lambda) \bar{D} \} \right\} \\
& = \text{Lip}(h) \alpha + \frac{1}{n} \sum_{i=1}^n \{ h(\tilde{z}_i) + \text{Lip}(h) \|z_i - \tilde{z}_i\| \}.
\end{aligned}$$

Lastly, the result of Equation (3.5) concludes the proof. \square

Proof of Theorem 3.3.1. [Step 1] In this step, we first establish a lower bound for the worst-case risk $\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h)$. By the definition of the Wasserstein ball $\mathfrak{M}_n(\alpha_n)$, we have

$$\begin{aligned}
& \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) - R(\mathbb{P}_n, h) \\
& \geq \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n \{ h(\tilde{z}_i) - h(z_i) \} \mid \frac{1}{n} \sum_{i=1}^n \|z_i - \tilde{z}_i\| \leq \alpha_n \right\}.
\end{aligned}$$

The mean value theorem and the Hölder continuity assumption on $\nabla_z h$ give

$$\begin{aligned}
h(\tilde{z}_i) &= h(z_i) + \nabla_z h(c_i)^T (\tilde{z}_i - z_i) \\
&= h(z_i) + \nabla_z h(z_i)^T (\tilde{z}_i - z_i) + \{ \nabla_z h(c_i) - \nabla_z h(z_i) \}^T (\tilde{z}_i - z_i) \\
&\geq h(z_i) + \nabla_z h(z_i)^T (\tilde{z}_i - z_i) - C_1 \|\tilde{z}_i - z_i\|^{k+1},
\end{aligned}$$

where $c_i = tz_i + (1-t)\tilde{z}_i$ for some $t \in [0, 1]$. Thus, we have

$$\begin{aligned}
& \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) - R(\mathbb{P}_n, h) \\
& \geq \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n \{ \nabla_z h(z_i)^T (\tilde{z}_i - z_i) - C_1 \|\tilde{z}_i - z_i\|^{k+1} \} \right. \\
& \quad \left. \mid \frac{1}{n} \sum_{i=1}^n \|z_i - \tilde{z}_i\| \leq \alpha_n \right\} \\
& \geq \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n \{ \nabla_z h(z_i)^T (\tilde{z}_i - z_i) \} \mid \frac{1}{n} \sum_{i=1}^n \|z_i - \tilde{z}_i\| \leq \alpha_n \right\} \\
& \quad - \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n C_1 \|\tilde{z}_i - z_i\|^{k+1} \mid \frac{1}{n} \sum_{i=1}^n \|z_i - \tilde{z}_i\| \leq \alpha_n \right\} \\
& =: S_1 - S_2.
\end{aligned}$$

By definition of dual norm, we have

$$\begin{aligned}
S_1 & \leq \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n \|\nabla_z h(z_i)\|_* \|\tilde{z}_i - z_i\| \mid \frac{1}{n} \sum_{i=1}^n \|z_i - \tilde{z}_i\| \leq \alpha_n \right\} \\
& \leq \alpha_n \max_{i \in [n]} \|\nabla_z h(z_i)\|_*.
\end{aligned}$$

Let $j := \operatorname{argmax}_{i \in [n]} \|\nabla_z h(z_i)\|_*$ and set $\tilde{z}_j = z_j + n\alpha_n v^*$ where $v^* := \operatorname{argmax}_{\|v\| \leq 1} \nabla_z h(z_j)^T v$. Set $\tilde{z}_i = z_i$ for $i \in [n] \setminus j$. Then, by plugging $\tilde{z}_i \in \mathcal{Z}$, we have $S_1 = \alpha_n \max_{i \in [n]} \|\nabla_z h(z_i)\|_*$.

For S_2 , using the fact $(\sum_{i=1}^n \|\tilde{z}_i - z_i\|^{k+1})^{\frac{1}{k+1}} \leq \sum_{i=1}^n \|\tilde{z}_i - z_i\|$ and $n\alpha_n = o(1)$, we have

$$\begin{aligned}
& \sup_{\tilde{z}_i \in \mathcal{Z}} \left\{ \frac{1}{n} \sum_{i=1}^n C_1 \|\tilde{z}_i - z_i\|^{k+1} \mid \frac{1}{n} \sum_{i=1}^n \|z_i - \tilde{z}_i\| \leq \alpha_n \right\} \\
& \leq C_1 \frac{1}{n} (n\alpha_n)^{k+1} = C_1 \alpha_n (n\alpha_n)^k = o(\alpha_n).
\end{aligned}$$

Thus,

$$\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) - R(\mathbb{P}_n, h) \geq \alpha_n \max_{i \in [n]} \|\nabla_z h(z_i)\|_* + o(\alpha_n).$$

For large enough n , $\alpha_n \leq n^{-1} \sum_{i=1}^n \sup_{z \in \mathcal{Z}} \|z - z_i\|$ and with the result of Lemma 3.6.1, we have

$$\begin{aligned} & \text{Lip}(h)\alpha_n + \frac{1}{n} \sum_{i=1}^n \{h(z'_i) + \text{Lip}(h)\|z_i - z'_i\|\} \\ & \geq \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) \\ & \geq \alpha_n \max_{i \in [n]} \|\nabla_z h(z_i)\|_* + R(\mathbb{P}_n, h) + o(\alpha_n). \end{aligned}$$

[Step 2] In this step, we show the difference between the upper and lower bound is negligible. To this end, we fix $\epsilon > 0$ and let $A = \{z \in \mathcal{Z} \mid \|\nabla_z h(z)\|_* > \text{Lip}(h) - \epsilon\}$. By the assumption on $\nabla_z h$ and compactness of \mathcal{Z} , $\delta := \mathbb{P}_{\text{data}}(A) > 0$. In addition, since \mathcal{Z}_n is a set of independently identically distributed samples from \mathbb{P}_{data} , the probability of the event

$$\left\{ \max_{i \in [n]} \|\nabla_z h(z_i)\|_* \leq \text{Lip}(h) - \epsilon \right\}$$

is $(1 - \delta)^n$. Thus, the fact $\sum_{i=1}^{\infty} (1 - \delta)^n = (1 - \delta)/\delta < \infty$ implies that there exists $N_{1,\epsilon} \in \mathbb{N}$ such that for all $n \geq N_{1,\epsilon}$, we have

$$\text{Lip}(h) - \max_{i \in [n]} \|\nabla_z h(z_i)\|_* \leq \epsilon,$$

almost surely. Furthermore, the fact $h(z'_i) \leq h(z_i) + \text{Lip}(h)\|z_i - z'_i\|$ and the definition of z'_i imply that for some $\tilde{z}_i \in \mathcal{Z}$ the difference between the upper bound and the lower bound is bounded by

$$(\text{Lip}(h) - \max_{i \in [n]} \|\nabla_z h(z_i)\|_*)\alpha_n + \frac{2}{n} \sum_{i=1}^n \text{Lip}(h)\|z_i - z'_i\|$$

$$\leq \epsilon \alpha_n + \frac{2C_2 \alpha_n^2}{n} \sum_{i=1}^n \text{Lip}(h) \|z_i - \tilde{z}_i\|,$$

almost surely. Since \mathcal{Z} is bounded, $n^{-1}C_2 \sum_{i=1}^n \text{Lip}(h) \|z_i - \tilde{z}_i\| = O_p(1)$. Therefore, the difference converges to zero as $\alpha_n \rightarrow 0$.

[Step 3] By the results of [Step 1] and [Step 2], we have for large enough n

$$\begin{aligned} & \frac{1}{\alpha_n} \left| \text{Lip}(h) \alpha_n + \frac{1}{n} \sum_{i=1}^n \{h(z'_i) + \text{Lip}(h) \|z_i - z'_i\|\} - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) \right| \\ & \leq \epsilon + 2\text{Lip}(h)C_2D\alpha_n, \end{aligned}$$

almost surely. Note that D is a diameter defined by $D := \sup_{z, \tilde{z}} \|z - \tilde{z}\|$.

Since (α_n) is a decreasing sequence, there exist $N_{2,\epsilon}$ such that $2\text{Lip}(h)C_2D\alpha_n \leq \epsilon$ for all $n \geq N_{2,\epsilon}$. Hence, the following holds for all $n \geq N_\epsilon := \max\{N_{1,\epsilon}, N_{2,\epsilon}\}$,

$$\begin{aligned} & \frac{1}{\alpha_n} \left| \text{Lip}(h) \alpha_n + \frac{1}{n} \sum_{i=1}^n \{h(z'_i) + \text{Lip}(h) \|z_i - z'_i\|\} - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) \right| \\ & \leq 2\epsilon, \end{aligned}$$

almost surely. Further, ϵ can be arbitrary small with $o(1)$ order, so it concludes that the following holds almost surely.

$$\begin{aligned} & \left| \text{Lip}(h) \alpha_n + \frac{1}{n} \sum_{i=1}^n \{h(z'_i) + \text{Lip}(h) \|z_i - z'_i\|\} - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) \right| \\ & = o(\alpha_n). \end{aligned}$$

□

3.6.2 Proof of Theorem 3.3.2

Proof of Theorem 3.3.2. Before beginning, we provide notations and an outline of the proof. To this ends, we define the some terminologies. Let $\hat{h}_{n,\alpha_n} = \operatorname{argmin}_{h \in \mathcal{H}} \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h)$, $\hat{h}_n = \operatorname{argmin}_{h \in \mathcal{H}} R(\mathbb{P}_n, h)$, and $h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(\mathbb{P}_{\text{data}}, h)$. Note that $\hat{h}_{n,\text{prop}} = \operatorname{argmin}_{h \in \mathcal{H}} R_{\alpha_n,\text{prop}}(\mathbb{P}_n, h)$. As for the outline, we decompose an excess risk as follows.

$$\begin{aligned}
& R(\mathbb{P}_{\text{data}}, \hat{h}_{n,\text{prop}}) - R(\mathbb{P}_{\text{data}}, h^*) \\
&= \underbrace{R(\mathbb{P}_{\text{data}}, \hat{h}_{n,\text{prop}}) - R(\mathbb{P}_n, \hat{h}_{n,\text{prop}})}_{\text{(T1)}} \\
&+ \underbrace{R(\mathbb{P}_n, \hat{h}_{n,\text{prop}}) - R_{\alpha_n,\text{prop}}(\mathbb{P}_n, \hat{h}_{n,\text{prop}})}_{\text{(T2)}} \\
&+ \underbrace{R_{\alpha_n,\text{prop}}(\mathbb{P}_n, \hat{h}_{n,\text{prop}}) - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n,\text{prop}})}_{\text{(T3)}} \\
&+ \underbrace{\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n,\text{prop}}) - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n,\alpha_n})}_{\text{[Step 1]}} \\
&+ \underbrace{\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n,\alpha_n}) - R(\mathbb{P}_n, \hat{h}_n)}_{\text{[Step 2]}} \\
&+ \underbrace{R(\mathbb{P}_n, \hat{h}_n) - R(\mathbb{P}_{\text{data}}, h^*)}_{\text{(T4)}}.
\end{aligned}$$

[Step 1] For $\epsilon > 0$, assume that $\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n,\text{prop}}) > \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n,\alpha_n}) + \epsilon$. By Theorem 3.3.1, there exists $N_\epsilon \in \mathbb{N}$ such that for all $n \geq N_\epsilon$, the following holds almost surely.

$$\left| \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n,\text{prop}}) - R_{\alpha_n,\text{prop}}(\mathbb{P}_n, \hat{h}_{n,\text{prop}}) \right| \leq \epsilon/4 \quad (3.7)$$

$$\left| \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \alpha_n}) - R_{\alpha_n, \text{prop}}(\mathbb{P}_n, \hat{h}_{n, \alpha_n}) \right| \leq \epsilon/4.$$

Furthermore,

$$\begin{aligned} R_{\alpha_n, \text{prop}}(\mathbb{P}_n, \hat{h}_{n, \text{prop}}) &\geq \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \text{prop}}) - \epsilon/4 \\ &> \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \alpha_n}) + 3\epsilon/4 \\ &\geq R_{\alpha_n, \text{prop}}(\mathbb{P}_n, \hat{h}_{n, \alpha_n}) + \epsilon/2. \end{aligned}$$

This contradicts to the definition of $\hat{h}_{n, \text{prop}}$ and thus for large enough n we have

$$\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \text{prop}}) \leq \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \alpha_n}) + \epsilon,$$

almost surely. In the selection (3.7) of ϵ , we can choose arbitrary small ϵ with $o(\alpha_n)$ order by Theorem 3.3.1. Further, $\alpha_n = o(n^{-1})$ implies that the following holds almost surely.

$$\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \text{prop}}) - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \alpha_n}) \leq o(n^{-1}), \quad (3.8)$$

[Step 2] For all $h \in \mathcal{H}$ and small enough α_n , we have

$$\begin{aligned} &\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) \\ &\leq R_{\alpha_n, \text{prop}}(\mathbb{P}_n, h) \\ &\leq R(\mathbb{P}_n, h) + \text{Lip}(h) \left(\alpha_n + \frac{2}{n} \sum_{i=1}^n \|z_i - z'_i\| \right) \\ &\leq R(\mathbb{P}_n, h) + L_{\mathcal{H}} \left(\alpha_n + \frac{2}{n} \sum_{i=1}^n \|z_i - z'_i\| \right). \quad (3.9) \end{aligned}$$

The first inequality is due to Lemma 3.6.1, the second inequality is due to the Lipschitzness of h , and the third inequality is due to

the assumption. Applying the infimum operator to the inequality (3.9) gives

$$\begin{aligned}
& \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \alpha_n}) \\
& \leq R(\mathbb{P}_n, \hat{h}_n) + L_{\mathcal{H}} \left(\alpha_n + \frac{2}{n} \sum_{i=1}^n \|z_i - z'_i\| \right) \\
& \leq R(\mathbb{P}_n, \hat{h}_n) + o(n^{-1}).
\end{aligned}$$

The last equality is because the construction of augmented data z'_i in Theorem 3.3.1 implies the term $L_{\mathcal{H}} \left(\alpha_n + \frac{2}{n} \sum_{i=1}^n \|z_i - z'_i\| \right)$ is $O(\alpha_n)$ and thus $o(n^{-1})$. Therefore,

$$\sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \alpha_n}) - R(\mathbb{P}_n, \hat{h}_n) \leq o(n^{-1}). \quad (3.10)$$

[Step 3] In this step, we obtain an upper bound for the terms (T2) and (T3). Note that

$$R(\mathbb{P}_n, h) \leq \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, h) \leq R_{\alpha_n, \text{prop}}(\mathbb{P}_n, h).$$

The first inequality is due to $\mathbb{P}_n \in \mathfrak{M}_n(\alpha_n)$ and the second inequality is due to Lemma 3.6.1. Thus,

$$(\text{T2}) = R(\mathbb{P}_n, \hat{h}_{n, \text{prop}}) - R_{\alpha_n, \text{prop}}(\mathbb{P}_n, \hat{h}_{n, \text{prop}}) \leq 0.$$

As for the term (T3), by Theorem 3.3.1, we have

$$\begin{aligned}
(\text{T3}) &= R_{\alpha_n, \text{prop}}(\mathbb{P}_n, \hat{h}_{n, \text{prop}}) - \sup_{\mathbb{Q} \in \mathfrak{M}_n(\alpha_n)} R(\mathbb{Q}, \hat{h}_{n, \text{prop}}) \\
&= o(n^{-1}).
\end{aligned}$$

Therefore,

$$(\text{T2}) + (\text{T3}) = o(n^{-1}). \quad (3.11)$$

[Step 4] In this step, we obtain an upper bound for the terms (T1) and (T4). Note that the term (T1) is bounded by $\sup_{h \in \mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)|$. As for the term (T4), we have

$$\begin{aligned}
& R(\mathbb{P}_n, \hat{h}_n) - R(\mathbb{P}_{\text{data}}, h^*) \\
&= R(\mathbb{P}_n, \hat{h}_n) - R(\mathbb{P}_n, h^*) + R(\mathbb{P}_n, h^*) - R(\mathbb{P}_{\text{data}}, h^*) \\
&\leq 0 + R(\mathbb{P}_n, h^*) - R(\mathbb{P}_{\text{data}}, h^*) \\
&\leq \sup_{h \in \mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)|.
\end{aligned}$$

The first inequality is due to the definition of \hat{h}_n . Thus, the sum of the terms (T1) and (T4) is bounded by $2 \sup_{h \in \mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)|$. Standard concentration inequalities [Devroye et al., 2013, pages 135-136] and symmetrization arguments [Van Der Vaart and Wellner, 1996, Lemma 2.3.1] provide

$$\sup_{h \in \mathcal{H}} |R(\mathbb{P}_n, h) - R(\mathbb{P}_{\text{data}}, h)| \leq 2\mathfrak{R}_n(\mathcal{H}) + C_{\mathcal{H}} \sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)}, \quad (3.12)$$

with probability at least $1 - \delta$.

Lastly, by aggregating the inequalities (3.8), (3.10), (3.11) and (3.12), the following holds with probability at least $1 - \delta$:

$$\begin{aligned}
& R(\mathbb{P}_{\text{data}}, \hat{h}_{n,\text{prop}}) - \inf_{h \in \mathcal{H}} R(\mathbb{P}_{\text{data}}, h) \\
&\leq 4\mathfrak{R}_n(\mathcal{H}) + 2C_{\mathcal{H}} \sqrt{\frac{2}{n} \log\left(\frac{2}{\delta}\right)} + o(n^{-1}).
\end{aligned}$$

□

3.6.3 Implementation details

In all experiments, we use ‘‘Wide ResNet’’ model with depth 28 and width 2 including batch normalization and leaky ReLU activation as Oliver et al. [2018] and Berthelot et al. [2019] did.

Our implementation of the model and training hyperparameters closely matches that of Berthelot et al. [2019]. For training, we choose Adam optimizer with the learning rate fixed as 0.002 and the batch size is 64. Instead of decaying the learning rate, we use an exponential moving average of the parameters with a decay of 0.999, and apply a weight decay of 0.02 at each update for the model. In MIXUP and DROID, the images are interpolated with the coefficient sampled from Beta(0.5, 0.5). We evaluate train and test accuracy on every 2^{16} training samples, and report the accuracy of the model trained with 100×2^{16} training samples.

For the proposed method, to minimize the objective function (3.6), computation of Lipschitz constant is required. However, as Virmaux and Scaman [2018] pointed out, exact computation of Lipschitz constant is NP-hard. Miyato et al. [2018] and Tsuzuku et al. [2018] suggested to calculate an upper bound of Lipschitz constant based on power method. Though their power method-based algorithms explicitly bound Lipschitz constant, they did not perform well in our experiments. We use gradient penalty which implicitly leads Lipschitzness as described in Algorithm 1.

Algorithm 1 Distributionally robust optimization with augmented data

- 1: **Input:** training dataset $\mathcal{Z}_n = \{z_1, \dots, z_n\}$, deep neural network h_θ parametrized by θ , hyperparameters $\gamma_\alpha, \gamma_\beta, \lambda_{\text{grad}} > 0$, optimization algorithm \mathfrak{A} .
 - 2: Initialize parameters θ in h_θ
 - 3: **while** until a convergent condition is met **do**
 - 4: Sample $\{z_{(1)}, \dots, z_{(B)}\}$ from \mathcal{Z}_n
 - 5: **for** $b = 1$ **to** B **do**
 - 6: Sample γ from $Beta(\gamma_\alpha, \gamma_\beta)$
 - 7: $z'_{(b)} = \text{Mixup}(z_{(b)}; \gamma)$
 - 8: **end for**
 - 9: $\mathcal{L} = B^{-1} \sum_{b=1}^B h_\theta(z'_{(b)}) + \lambda_{\text{grad}} \left\| \nabla_z h_\theta(z'_{(b)}) \right\|_2^2$ \triangleright calculate loss
 - 10: $\theta \leftarrow \mathfrak{A}(\mathcal{L}, \theta)$ \triangleright update parameters
 - 11: **end while**
-

Bibliography

- Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.
- O Chapelle, B Schölkopf, and A Zien. *Semi-Supervised Learning*. MIT Press, 2006.
- François Denis, Rémi Gilleron, and Fabien Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83, 2005.
- Xiao-Li Li and Bing Liu. Learning from positive and unlabeled examples with different data distributions. In *European Conference on Machine Learning*, pages 218–229. Springer, 2005.
- Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220. ACM, 2008.
- Yanshan Xiao, Bo Liu, Jie Yin, Longbing Cao, Chengqi Zhang, and Zhifeng Hao. Similarity-based approach for positive and unlabelled learning. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- Maria A Zuluaga, Don Hush, Edgar JF Delgado Leyton, Marcela Hernández Hoyos, and Maciej Orkisz. Learning from

- only positive and unlabeled data to detect lesions in vascular ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 9–16. Springer, 2011.
- Tieliang Gong, Guangtao Wang, Jieping Ye, Zongben Xu, and Ming Lin. Margin based pu learning. In *AAAI Conference on Artificial Intelligence*, 2018.
- Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 2012.
- Peng Yang, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh, and See-Kiong Ng. Ensemble positive unlabeled learning for disease gene identification. *PloS One*, 9(5):e97079, 2014.
- Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009, 2010.
- Jiaqi Zhang, Zhenzhen Wang, Junsong Yuan, and Yap-Peng Tan. Positive and unlabeled learning for anomaly detection with multi-features. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 854–862. ACM, 2017a.
- Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. Partially supervised classification of text documents. In *International Conference on Machine Learning*, volume 2, pages 387–394. Citeseer, 2002.
- Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 587–592. Morgan Kaufmann Publishers Inc., 2003.
- Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Building text classifiers using positive and unlabeled examples. In *Data Mining*,

2003. *ICDM 2003. Third IEEE International Conference on*, pages 179–186. IEEE, 2003.
- Clayton Scott and Gilles Blanchard. Novelty detection: Unlabeled data definitely help. In *Artificial Intelligence and Statistics*, pages 464–471, 2009.
- Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, pages 703–711, 2014.
- Marthinus C du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pages 1386–1394, 2015.
- Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, pages 1675–1685, 2017.
- Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.
- Yongchan Kwon, Wonyoung Kim, Masashi Sugiyama, and Myunghee Cho Paik. Principled analytic classifier for positive-unlabeled learning via weighted integral probability metric. *Machine Learning*, Nov 2019.
- Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.
- Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *arXiv preprint arXiv:1905.00397*, 2019.

- Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017b.
- Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447, 2019.
- David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.
- Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, and John R Leathwick. Presence-only data and the em algorithm. *Biometrics*, 65(2): 554–563, 2009.
- Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems*, pages 1199–1207, 2016.
- Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rJzLciCqKm>.
- Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- Tomoya Sakai, Marthinus Christoffel Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *International Conference on Machine Learning*, pages 2998–3006, 2017.

- Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 201–208. ACM, 2006.
- Emanuele Sansone, Francesco GB De Natale, and Zhi-Hua Zhou. Efficient training for positive unlabeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2584–2598, 2019.
- Bharath K Sriperumbudur, Kenji Fukumizu, and Gert Lanckriet. On the relation between universality, characteristic kernels and rkhs embedding of measures. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 773–780, 2010a.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.
- Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.
- Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.
- Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.
- A Gretton, AJ Smola, J Huang, M Schmittfull, KM Borgwardt, B Schölkopf, Quiñonero Candela, M Sugiyama, A Schwaighofer, ND Lawrence, et al. Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning*, pages 131–160. MIT Press, 2009.

- Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 945–954. IEEE, 2017.
- Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.
- Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.
- Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.
- Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010b.
- Yi Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.
- Jean-Yves Audibert, Alexandre B Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.
- Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.
- Giorgio Patrini, Frank Nielsen, Richard Nock, and Marcello Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *International conference on machine learning*, pages 708–717, 2016.

- Gilles Blanchard, Marek Flaska, Gregory Handy, Sara Pozzi, and Clayton Scott. Classification with asymmetric label noise: Consistency and maximal denoising. *Electronic Journal of Statistics*, 10(2):2780–2824, 2016.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning*, pages 2052–2060, 2016.
- Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.
- Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*, 2018.
- Heinrich Jiang. Uniform convergence rates for kernel density estimation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1694–1703. JMLR. org, 2017.
- Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- Richard E Bellman. *Adaptive control processes: a guided tour*. Princeton university press, 1961.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

- Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.
- Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pages 351–359, 2013.
- Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.
- Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Meulenbergh, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? *arXiv preprint arXiv:1611.02041*, 2016.
- Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980, 2017.
- Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *arXiv preprint arXiv:1905.10943*, 2019.
- Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.
- Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance

- guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.
- Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.
- Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.
- Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *arXiv preprint arXiv:1710.10016*, 2017.
- Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.
- László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.
- Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. *arXiv preprint arXiv:1705.07815*, 2017.
- Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.
- Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.
- Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning

algorithms. In *Advances in Neural Information Processing Systems*, pages 3235–3246, 2018.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pages 3835–3844, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6541–6550, 2018.

국문초록 (Abstract in Korean)

본 학위 논문은 두 가지 종류의 기계학습 문제를 다룬다. 첫 번째 연구문제는 양성 자료집합과 미분류 자료집합만을 이용하여 이진 분류기를 학습하는 문제이다 (양성-미분류 문제). 해당 분야 종래 연구는 실제 자료에서 실험적으로 우수한 성능을 보였으나 전체 자료 수의 제공에 달하는 연산량이 필요하다. 이 연구는 재생 커널 힐버트 공간의 닫힌 구를 가설공간으로 설정하여 저연산량 알고리즘을 제안한다. 이에 더하여 제안하는 분류기의 초과 위험 상계를 유도함으로써 제안하는 알고리즘의 이론적 타당성을 보인다. 이 연구는 양성-미분류 분야에서 처음으로 위험 일치성을 유도한 연구이다.

두 번째 연구는 증대자료를 사용한 경험위험 최소화를 분포적 강건 최적화(distributionally robust optimization) 관점에서 해석한 이론 연구이다. 자료 증대법은 최근 기계학습 분야에서 성능 향상을 위한 핵심적인 기술로 부상 했으나 이에 대한 이론적 근거는 거의 전무한 상태이다. 본 연구는 자료 증대법을 미세변동으로 고려하여, 증대자료를 사용한 모형 학습을 분포적 강건 최적화 관점으로 해석한다. 구체적으로 손실 함수의 도함수가 홀더 연속 함수인 경우 증대자료를 사용한 별점경험위험(regularized empirical risk)이 최

악 위험으로 근사 됨을 보인다. 이에 더하여, 제안하는 목적함수의 최적해가 위험 일치성을 가짐을 이론적으로 증명하였다. 실제 잡음 자료를 이용한 실험에서는, 제안된 알고리즘이 종래 방법론에 비해 우수한 정분류율을 가짐을 보였다. 본 연구는 분포적 강건 최적화 문헌에서 증대 자료와 심층신경망 모형의 사용의 정당성을 엄밀하게 보인 첫 연구이다.

본 학위 논문의 두 연구 모두 적분 확률 측도를 활용한 연구이다. 본 학위 논문은 기계학습 분야의 많은 문제가 분포 간 측도를 이용하여 공식화 될 수 있으며 기계학습 문제를 새로운 관점에서 해석 및 해결될 수 있음을 보인다.

주요어 : 적분 확률 측도, 양성-미분류 학습, 재생 커널 힐버트 공간, 자료 증대법, 분포적 강건 최적화

학 번 : 2013 - 22897