Positive-Unlabeled Learning with Non-Negative Risk Estimator

Xinsong Ma, Chenglin Yu[*]

June 20, 2020

Contents

[*] V587 Group

1 Background

1.1 What is PU learning?

We have only Positive and Unlabeled examples. We need to learning something. For example,

- train a binary classifier (binary classification)
- matrix completion

In this paper, we focus on binary classification.

1.2 Why PU learning?

Classification when engative data is unavailable. Example: click advertisement

- clicked: positive (interesting)
- non-clicked: unlabeled (not interesting or unseen)

1.3 Notation

1. $X \in \mathbb{R}^d$, $d \in \mathbb{N}$ is the input random variable
2. $Y \in \{\pm 1\}$ output random variable
3. $p(x, y)$ be the underlying joint density of (X,Y)
4. $p_{\mathrm{p}}(x) = p(x|Y = +1)$ the conditional pdf of X given $Y = +1$
5. $p_{\mathrm{n}}(x) = p(x|Y = -1)$ the conditional pdf of X given $Y = -1$
6. $\pi_p = p(Y = +1)$ the class-prior probability
7. $\pi_n = p(Y = -1) = 1 - p(Y = +1) = 1 - \pi_p$
8. $g : \mathbb{R}^d \to \mathbb{R}$ decision function
9. $\ell : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$ loss function, the value of $\ell(t, y)$ means the loss incurred by predicting an ouput t when the ground truth is y.
10. $R(g) = \mathbb{E}_{(X,Y)\sim p(x,y)}[\ell(g(X), Y)]$
$= p(Y = +1)\mathbb{E}_{X\sim p_{\mathrm{p}}}[\ell(g(X), +1)] + p(Y = -1)\mathbb{E}_{X\sim p_{\mathrm{n}}}[\ell(g(X), -1)]$
$= \pi_{\mathrm{p}}\mathbb{E}_{\mathrm{p}}[\ell(g(X), +1)] + \pi_{\mathrm{n}}\mathbb{E}_{\mathrm{n}}[\ell(g(X), -1)]$
$= \pi_{\mathrm{p}} R_{\mathrm{p}}^{+}(g) + \pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g)$
11. $R_{\mathrm{p}}^{-}(g) = \mathbb{E}_{\mathrm{p}}[\ell(g(X), -1)]$
12. $R_{\mathrm{u}}^{-}(g) = \mathbb{E}_{\mathrm{u}}[\ell(g(X), -1)]$
13. $\widehat{R}_{\mathrm{p}}^{+}(g) = (1/n_{\mathrm{p}}) \sum_{i=1}^{n_{\mathrm{p}}} \ell\left(g\left(x_i^{\mathrm{p}}\right), +1\right)$

14. $\widehat{R}_{\mathrm{n}}^{-}(g)=(1/n_{\mathrm{n}})\sum_{i=1}^{n_{\mathrm{n}}}\ell\left(g\left(x_i^{\mathrm{n}}\right),-1\right)$

15. $\widehat{R}_{\mathrm{u}}^{-}(g)=(1/n_{\mathrm{u}})\sum_{i=1}^{n_{\mathrm{u}}}\ell\left(g\left(x_i^{\mathrm{u}}\right),-1\right)$

16. $\widehat{R}_{\mathrm{pn}}(g)=\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{+}(g)+\pi_{\mathrm{n}}\widehat{R}_{\mathrm{n}}^{-}(g)$

17. $\mathcal{X}_{\mathrm{p}}=\{x_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}}$

18. $\mathcal{X}_{\mathrm{u}}=\{x_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}}$

19. $\mathcal{G}$ function class (hypothesis class)

20. $\mathfrak{R}_{n,q}(\mathcal{G})=\mathbb{E}_{\mathcal{X}\sim q^n}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{x_i\in\mathcal{X}}\sigma_i g\left(x_i\right)\right]$ Rademacher complexity of $\mathcal{G}$ for the sampling of size n from $q(x)$.

21. $\widehat{g}_{\mathrm{pn}}$ the empirical minimizers of $\widehat{R}_{\mathrm{pn}}(g)$

22. $\widehat{R}_{\mathrm{pu}}(g)=\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{+}(g)-\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)+\widehat{R}_{\mathrm{u}}^{-}(g)$

23. $\widehat{g}_{\mathrm{pu}}$ the empirical minimizers of $\widehat{R}_{\mathrm{pu}}(g)$

24. $\widetilde{R}_{\mathrm{pu}}(g)=\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{+}(g)+\max\left\{0,\widehat{R}_{\mathrm{u}}^{-}(g)-\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)\right\}$

25. $\widetilde{g}_{\mathrm{pu}}$ the empirical minimizers of $\widetilde{R}_{\mathrm{pu}}(g)$

2 Preliminaries

2.1 The bias of a point estimator

According to Definition 7.3.2 of [Casella and Berger, 2002]

Definition 2.1. *The bias of a point estimator W of a parameter θ is the difference between the expected value of W and θ; that is,* $\operatorname{Bias}_\theta W=\mathrm{E}_\theta W-\theta$.

An estimator whose bias is identically equal to 0 called unbiased and satisifes $\mathrm{E}_\theta W=\theta$ *for all* θ

2.2 MSE of an estimator

Mean Squared Error is a method of evaluating esimators. We refer to the Definition 7.3.1 of [Casella and Berger, 2002].

Definition 2.2. *The mean squared error (MSE) of an estimator W of a parameter θ is the function of θ defined by* $\mathrm{E}_\theta(W-\theta)^2$

2.3 Consistency of an estimator

Consistency is an asymptotic property[1] of an estimator. We refer to section 10.1.1 of [Casella and Berger, 2002].

The property of consistency seems to be quite a fundamental one, requiring that the estimator converges to the ”correct” value as the sample size becomes infinite. It is such a fundamental property that the worth of an inconsistent estimator should be questioned (or at least vigorously investigated).

Consistency (as well as all asymptotic properties) concerns a sequence of estimators rather than a single estimator, although it is common to speak of a ”consistent estimator.” If we observe $X_l, X_2,$ according to a distribution $f(x|\theta)$, we can construct a sequence of estimators $W_n = W_n(X_1, ..., X_n)$ merely by performing the same estimation procedure for each sample size n. For example, $\bar{X}_1 = X_1$, $\bar{X}_2 = (X_1 + X_2)/2, \bar{X}_3 = (X_1 + X_2 + X_3)/3$, etc. We can now define a consistent sequence.

Definition 2.3. *A sequence of estimators $W_n = W_n(X_1, \ldots, X_n)$ is a consistent sequence of estimators of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$*

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| < \epsilon) = 1$$

2.4 Almost surely

We refer to wiki

In probability theory, an event is said to happen almost surely if it happens with probability 1.

Why ”almost” surely? math.stackexchange

Consider choosing a real number uniformly at random from the interval [0,1]. The event ”1/2 will not be chosen” has probability 1, but it is still possible for 1/2 to be chosen. Thus, even if the event happens with probability 1, we say the event ”1/2 will not be chosen” happen almost surely, rather than happen surely!

To sum up, almost surely means ”happen with probability 1”, rather than surely happen.

The concept is essentially analogous to the concept of ”almost everywhere” in measure theory.

2.5 Estimation error bound of a hypothesis

According to section 2.4.3 Estimation and approximation errors of ”Foundation of Machine Learning”, The difference between the error of a hypothesis $h \in H$ and the Bayes error can be decomposed as:The difference between the error of a hypothesis $h \in H$ and the Bayes error can be decomposed as:

[1] properties describing the behavior of a procedure as the sample size becomes infinite

$$R(h)-R^{*}=\underbrace{\left(R(h)-R\left(h^{*}\right)\right)}_{\text{estimation}}+\underbrace{\left(R\left(h^{*}\right)-R^{*}\right)}_{\text{approximation}}$$

where h^{*} is a hypothesis in H with minimal error, or a *best-in-class hypothesis.*

2.6 The infinity norm of a function

We refer to the definition of L^{∞} -Norm in this lecture[2].

Definition 2.4. $\boldsymbol{L}^{\infty}$ *-norm Let* (X,μ) *be a measure space, and let* f *be a measurable function on* X*. The* $\boldsymbol{L}^{\infty}$ *-norm of* f *is defined as follows:*

$$\|f\|_{\infty}=\min\{M\in[0,\infty]\,|\,|f|\leq M\ \textit{almost everywhere}\ \}$$

We say that f *is an* $\boldsymbol{L}^{\infty}$ *function if* $\|f\|_{\infty}<\infty$

” Almost everywhere” from wiki

If (X,Σ,μ) is a measure space, a property P is said to hold almost everywhere in X if there exists a set $N\in\Sigma$ with $\mu(N)=0$, and all $x\in X\backslash N$ have the property P.

Another common way of expressing the same thing is to say that ”almost every point satisfies P ”, or that ”for almost every $x, P(x)$ holds”. It is not required that the set $\{x\in X:\neg P(x)\}$ has measure 0 ; it may not belong to Σ. By the above definition, it is sufficient that $\{x\in X:\neg P(x)\}$ be contained in some set N that is measurable and has measure 0

2.7 Mcdiarmid's inequality

Refer to McDiarmid's inequality in wiki.

Consider independent random variables $X_1,X_2,\ldots X_n$ on probability space $(\Omega,\mathcal{F},\mathrm{P})$ where $X_i\in\mathcal{X}_i$ for all i and a mapping $f:\mathcal{X}_1\times\mathcal{X}_2\times\cdots\times\mathcal{X}_n\rightarrow\mathbb{R}$. Assume there exist constant $c_1,c_2,\ldots,c_n$ such that for all i

$$\sup_{x_1,\cdots,x_{i-1},x_i,x_i',x_{i+1},\cdots,x_n}\left|f\left(x_1,\ldots,x_{i-1},x_i,x_{i+1},\cdots,x_n\right)-f\left(x_1,\ldots,x_{i-1},x_i',x_{i+1},\cdots,x_n\right)\right|\leq c_i$$

(In other words, changing the value of the i th coordinate x_i changes the value of f by at most c_i.) Then, for any $\epsilon>0$

$$\begin{aligned}&\mathrm{P}\left(f\left(X_1,X_2,\cdots,X_n\right)-\mathbb{E}\left[f\left(X_1,X_2,\cdots,X_n\right)\right]\geq\epsilon\right)\leq\exp\left(-\tfrac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)\\&\mathrm{P}\left(f\left(X_1,X_2,\cdots,X_n\right)-\mathbb{E}\left[f\left(X_1,X_2,\cdots,X_n\right)\right]\leq-\epsilon\right)\leq\exp\left(-\tfrac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)\end{aligned}$$

and

$$\mathrm{P}\left(\left|f\left(X_1,X_2,\cdots,X_n\right)-\mathbb{E}\left[f\left(X_1,X_2,\cdots,X_n\right)\right]\right|\geq\epsilon\right)\leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

[2] http://faculty.bard.edu/belk/math461/LpFunctions.pdf

2.8 Contraction Lemma

Lemma 26.9 of [Shalev-Shwartz and Ben-David, 2014]

Lemma 2.5. *(Contraction Lemma). For each $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function; namely, for all $\alpha, \beta \in \mathbb{R}$ we have $|\phi_i(\alpha) - \phi_i(\beta)| \leq \rho|\alpha - \beta|$. For $\mathbf{a} \in \mathbb{R}^m$ let $\boldsymbol{\phi}(\mathbf{a})$ denote the vector $(\phi_1(a_1), \ldots, \phi_m(a_m))$. Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ be a finite set of vectors in $\mathbb{R}^m$. Let $\boldsymbol{\phi} \circ A = \{\boldsymbol{\phi}(\mathbf{a}) : a \in A\}$. Then*

$$R(\boldsymbol{\phi} \circ A) \leq \rho R(A)$$

Theorem 4.12 of [Ledoux and Talagrand, 2013]

Theorem 2.6. Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ *be convex and increasing. Let further $\varphi_i : \mathbb{R} \to \mathbb{R}, i \leq N$, be contractions such that $\varphi_i(0) = 0$. Then, for any bounded subset T in $\mathbb{R}^N$*

$$\mathbb{E}F\left(\frac{1}{2}\left\|\sum_{i=1}^{N} \varepsilon_i \varphi_i(t_i)\right\|_T\right) \leq \mathbb{E}F\left(\left\|\sum_{i=1}^{N} \varepsilon_i t_i\right\|_T\right)$$

3 Progress and Motivation

3.1 Earlier

3.2 unbiased PU learning

Since our goal is minimize the same expected riks as PN learning:

$$R(g) = \mathbb{E}_{(X,Y)\sim p(x,y)}[\ell(g(X), Y)] = \pi_{\mathrm{p}} R_{\mathrm{p}}^{+}(g) + \pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g)$$

Since we don't know the true distribution, we often minimize the empirical risk (which is an estimator of the true risk),

In PN learning, thanks to the availability of $\mathcal{X}_p$ and $\mathcal{X}_n$, $R(g)$ can be approximated directly by:

$$\widehat{R}_{\mathrm{pn}}(g) = \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) + \pi_{\mathrm{n}} \widehat{R}_{\mathrm{n}}^{-}(g) \tag{1}$$

Since $\mathbb{E}[\widehat{R}_{\mathrm{pn}}(g)] = R(g)$, $\widehat{R}_{\mathrm{pn}}(g)$ is unbiased.

In PU learning, $\mathcal{X}_n$ is unavailable. We can not get a unbiased estimator of $R_{\mathrm{n}}^{-}(g)$ directly. However, we often assume the *Two-sample problem setting of PU learning:*

- $\mathcal{X}_{\mathrm{p}} = \{x_i^{\mathrm{p}}\}_{i=1}^{n_{\mathrm{p}}} \sim p_{\mathrm{p}}(x)$
- $\mathcal{X}_{\mathrm{u}} = \{x_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{y}}} \sim p(x)$
- $p(x) = \pi_{\mathrm{p}} p_{\mathrm{p}}(x) + \pi_{\mathrm{n}} p_{\mathrm{n}}(x)$

We want to use the assumption to approximate $R_{\mathrm{n}}^{-}(g)$ indirectly.

Since

$$R_{\mathrm{n}}^{-}(g) = \mathbb{E}_{\mathrm{n}}[\ell(g(X), -1)]$$

Motivated by this, we consider this random variable, $\ell(g(X), -1)$,
We know

$$p(x) = \pi_{\mathrm{p}} p_{\mathrm{p}}(x) + \pi_{\mathrm{n}} p_{\mathrm{n}}(x)$$

Taking expectation of $\ell(g(X), -1)$ according to $p(x)$, $p_{\mathrm{p}}(x)$, $p_{\mathrm{n}}(x)$, we have

$$\pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g) = R_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} R_{\mathrm{p}}^{-}(g)$$

Taking it into

$$R(g) = \pi_{\mathrm{p}} R_{\mathrm{p}}^{+}(g) + \pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g)$$

We have

$$R(g) = \pi_{\mathrm{p}} R_{\mathrm{p}}^{+}(g) + R_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} R_{\mathrm{p}}^{-}(g)$$

Then $R(g)$ can be approximated indirectly by

$$\widehat{R}_{\mathrm{pu}}(g) = \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g) + \widehat{R}_{\mathrm{u}}^{-}(g) \tag{2}$$

This estimator is also unbiased.

[Niu et al., 2016] analyzed the estimation error bound of $\widehat{g}_{\mathrm{pn}}$ and $\widehat{g}_{\mathrm{pu}}$, respectively. And [Niu et al., 2016] proved that the estimation error bound (EEB) of $\widehat{g}_{\mathrm{pu}}$ is tighter than EEB of $\widehat{g}_{\mathrm{pn}}$ if the followingt 2 conditions are satisifed:

1. ℓ satisfies $\ell(t, +1) + \ell(t, -1) = 1$ and is Lipschitz continuous
2. The rademacher complexity of $\mathcal{G}$ satisifes: There is a constant $C_{\mathcal{G}} > 0$ such that $\mathfrak{R}_{n,q}(\mathcal{G}) \leq C_{\mathcal{G}}/\sqrt{n}$ for any marginal density $q(x) \in \{p(x), p_{+}(x), p_{-}(x)\}$

The estimation error bounds contains the Rademacher complexity. Therefore, the critical assumption on the Rademacher complexity is indisensable, otherwise it will be difficult for EEB of $\widehat{g}_{\mathrm{pu}}$ to be tighter than $\widehat{g}_{\mathrm{pn}}$

For example, If $\mathcal{G} = \{g | \|g\|_{\infty} \leq C_g\}$ where $C_g > 0$ is a constant, i.e., it has all measurable functions with some bounded norm, then $\mathfrak{R}_{n,q}(\mathcal{G}) = \mathcal{O}(1)$ for any n and $q(x)$ and all bounds become trivial[3].

since

$$\begin{aligned}
&\widehat{R}_{\mathrm{pu}}(g) = \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g) + \widehat{R}_{\mathrm{u}}^{-}(g) \\
&= \pi_{\mathrm{p}} \left(1/n_{\mathrm{p}}\right) \sum_{i=1}^{n_{\mathrm{p}}} \ell\left(g\left(x_i^{\mathrm{p}}\right), +1\right) \\
&- \pi_{\mathrm{p}} \left(1/n_{\mathrm{p}}\right) \sum_{i=1}^{n_{\mathrm{p}}} \ell\left(g\left(x_i^{\mathrm{p}}\right), -1\right) + \left(1/n_{\mathrm{u}}\right) \sum_{i=1}^{n_{\mathrm{u}}} \ell\left(g\left(x_i^{\mathrm{u}}\right), -1\right)
\end{aligned}$$

If there is a classifier that can perfectly separate P and U, Then training error w.r.t. 0-1 loss is:

$$\pi_{\mathrm{p}} \cdot 0 - \pi_{\mathrm{p}} \cdot 1 + 0 < 0$$

[3] Q: How to obtain the $\mathcal{O}(1)$ result?

Moreover, if ℓ is not bounded from above, $\widehat{R}_{\mathrm{pu}}(g)$ becomes not bounded from below, i.e., it may diverge to $-\infty$. The problem of overfitting may occur.

Thus, in oder to obtain high quality $\widehat{g}_{\mathrm{pu}}$, $\mathcal{G}$ cannot be too complex, or equivalently, the model of g cannot be too flexible, such as being deep neural network.

Nevertheless, we have no choice sometimes: we are interested in using flexible models, while labeling more data is out of our control. Can we alleviate the overfitting problem with neither changing the model nor labeling more data?

In the expermiment, we note that $\widehat{R}_{\mathrm{pu}}\left(\widehat{g}_{\mathrm{pu}}\right)$ keeps decreasing and goes negative. This should be fixed since $R(g) \geq 0$ for any g. Specifically, it holds that $R_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} R_{\mathrm{p}}^{-}(g) = \pi_{\mathrm{n}} R_{\mathrm{n}}^{-}(g) \geq 0$, but $\widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g) \geq 0$ is not always true, which is a potential reason for uPU to overfit. Based on this key observation, we propose a *non-negative risk estimator* for PU learning:

$$\widetilde{R}_{\mathrm{pu}}(g) = \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) + \max\left\{0, \widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g)\right\} \tag{3}$$

We refer to the process of obtaining $\widehat{R}_{\mathrm{pu}}$ as non-negative PU(nnPU) learning.

We will analyze $\widetilde{R}_{\mathrm{pu}}(g)$ and $\widetilde{g}_{\mathrm{pu}}$ respectively.

4 $\widetilde{R}_{\mathrm{pu}}(g)$ as an estimator

4.1 Bias

Bias: $\mathbb{E}_{\mathcal{X}_{\mathrm{p}}, \mathcal{X}_{\mathrm{u}}}\left[\widetilde{R}_{\mathrm{pu}}(g)\right] - R(g)$

$$\widetilde{R}_{\mathrm{pu}}(g) = \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) + \max\left\{0, \widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g)\right\}$$

Let $\mathfrak{D}^{-}(g) = \left\{(\mathcal{X}_{\mathrm{p}}, \mathcal{X}_{\mathrm{u}}) \,|\, \widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \hat{R}_{\mathrm{p}}^{-}(g) < 0\right\}$,

$\mathfrak{D}^{+}(g) = \left\{(\mathcal{X}_{\mathrm{p}}, \mathcal{X}_{\mathrm{u}}) \,|\, \widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \hat{R}_{\mathrm{p}}^{-}(g) \geq 0\right\}$

We have

$$\begin{aligned} \widetilde{R}_{\mathrm{pu}}(g) = \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) + \max\left\{0, \widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g)\right\} &= \begin{cases} \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) + 0 & \mathfrak{D}^{-}(g) \\ \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) + \widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{-}(g) & \mathfrak{D}^{+}(g) \end{cases} \\ &= \begin{cases} \pi_{\mathrm{p}} \widehat{R}_{\mathrm{p}}^{+}(g) & \mathfrak{D}^{-}(g) \\ \widehat{R}_{\mathrm{pu}}(g) & \mathfrak{D}^{+}(g) \end{cases} \end{aligned}$$

Let

$$p_{\mathrm{p}}\left(\mathcal{X}_{\mathrm{p}}\right) = p_{\mathrm{p}}\left(x_1^{\mathrm{p}}\right) \cdots p_{\mathrm{p}}\left(x_{n_{\mathrm{p}}}^{\mathrm{p}}\right), \quad p\left(\mathcal{X}_{\mathrm{u}}\right) = p\left(x_1^{\mathrm{u}}\right) \cdots p\left(x_{n_{\mathrm{u}}}^{\mathrm{u}}\right)$$

be the probability density functions of $\mathcal{X}_{\mathrm{p}}$ and $\mathcal{X}_{\mathrm{u}}$. Then let $F_{\mathrm{p}}\left(\mathcal{X}_{\mathrm{p}}\right)$ be the cumulative distribution function of $\mathcal{X}_{\mathrm{p}}$, $F_{\mathrm{u}}\left(\mathcal{X}_{\mathrm{u}}\right)$ be that of $\mathcal{X}_{\mathrm{u}}$, and $F\left(\mathcal{X}_{\mathrm{p}}, \mathcal{X}_{\mathrm{u}}\right)$ be the joint cumulative distribution function of $\left(\mathcal{X}_{\mathrm{p}}, \mathcal{X}_{\mathrm{u}}\right)$

Then we have [4]

$$F\left(\mathcal{X}_{\mathrm{p}}, \mathcal{X}_{\mathrm{u}}\right) = F_{\mathrm{p}}\left(\mathcal{X}_{\mathrm{p}}\right) \cdot F_{\mathrm{u}}\left(\mathcal{X}_{\mathrm{u}}\right)$$

[4] Q: $\mathcal{X}_{\mathrm{p}}$ and $\mathcal{X}_{\mathrm{u}}$ is not necessarily independent

Given the above definitions, the measure of $\mathfrak{D}^-(g)$ is defined by

$$\Pr\left(\mathfrak{D}^-(g)\right) = \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \mathrm{d}F\left(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}\right)$$

where Pr denotes the probability.

Since $\widetilde{R}_\mathrm{pu}(g)$ is identical to $\widehat{R}_\mathrm{pu}(g)$ on $\mathfrak{D}^+(g)$ and different from $\widehat{R}_\mathrm{pu}(g)$ on $\mathfrak{D}^-(g)$, we have $\Pr\left(\mathfrak{D}^-(g)\right) = \Pr\left\{\widetilde{R}_\mathrm{pu}(g) \neq \widehat{R}_\mathrm{pu}(g)\right\}$. That is, the measure of $\mathfrak{D}^-(g)$ is non-zero if and only if $\widetilde{R}_\mathrm{pu}(g)$ differs from $\widehat{R}_\mathrm{pu}(g)$ with a non-zero probability. Based on the facts that $\widehat{R}_\mathrm{pu}(g)$ is unbiased and $\widetilde{R}_\mathrm{pu}(g) - \widehat{R}_\mathrm{pu}(g) = 0$ on $\mathfrak{D}^+(g)$, we have

$$\begin{aligned}
\mathbb{E}\left[\widetilde{R}_\mathrm{pu}(g)\right] - R(g) =& \mathbb{E}\left[\widetilde{R}_\mathrm{pu}(g) - \widehat{R}_\mathrm{pu}(g)\right] \\
=& \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^+(g)} \widetilde{R}_\mathrm{pu}(g) - \widehat{R}_\mathrm{pu}(g)\mathrm{d}F\left(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}\right) \\
&+ \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \widetilde{R}_\mathrm{pu}(g) - \widehat{R}_\mathrm{pu}(g)\mathrm{d}F\left(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}\right) \\
=& \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \widetilde{R}_\mathrm{pu}(g) - \widehat{R}_\mathrm{pu}(g)\mathrm{d}F\left(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}\right) \\
\leq& \sup_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \left(\widetilde{R}_\mathrm{pu}(g) - \widehat{R}_\mathrm{pu}(g)\right) \cdot \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \mathrm{d}F\left(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}\right) \\
=& \sup_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \left(\pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g) - \widehat{R}_\mathrm{u}^-(g)\right) \cdot \Pr\left(\mathfrak{D}^-(g)\right)
\end{aligned}$$

That is,

$$\begin{aligned}
\mathbb{E}\left[\widetilde{R}_\mathrm{pu}(g)\right] - R(g) \leq& \sup_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \left(\widetilde{R}_\mathrm{pu}(g) - \widehat{R}_\mathrm{pu}(g)\right) \cdot \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \mathrm{d}F\left(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}\right) \\
=& \sup_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} \left(\pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g) - \widehat{R}_\mathrm{u}^-(g)\right) \cdot \Pr\left(\mathfrak{D}^-(g)\right)
\end{aligned}$$

Taking into account that

$$\Pr\left(\mathfrak{D}^-(g)\right) = \Pr\left\{\widehat{R}_\mathrm{u}^-(g) - \pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g) < 0\right\}$$

We made an assumption:

Assumption There is $\alpha > 0$, such that $R_\mathrm{n}^-(g) \geq \alpha$

Then we have

$$\begin{aligned}
\Pr\left(\mathfrak{D}^-(g)\right) &= \Pr\left\{\widehat{R}_\mathrm{u}^-(g) - \pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g) < 0\right\} \\
&\leq \Pr\left\{\widehat{R}_\mathrm{u}^-(g) - \pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g) \leq R_\mathrm{n}^-(g) - \alpha\right\} \\
&= \Pr\left\{R_\mathrm{n}^-(g) - \left(\widehat{R}_\mathrm{u}^-(g) - \pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g)\right) \geq \alpha\right\}
\end{aligned}$$

Since $\mathbb{E}\left[\widehat{R}_{\mathrm{u}}^{-}(g)-\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)\right]=R_{\mathrm{n}}^{-}(g)$

We want to use McDiarmid's inequality to upper bound this probability,

We made another assumption:

Assumption $0 \leq \ell(t, \pm 1) \leq C_{\ell}$

We have assumed that $0 \leq \ell(t, \pm 1) \leq C_{\ell}$, and thus the change of $\widehat{R}_{\mathrm{p}}^{-}(g)$ will be no more than C_{ℓ}/n_{p} if some $x_i^{\mathrm{p}} \in \mathcal{X}_{\mathrm{p}}$ is replaced, or the change of $\widehat{R}_{\mathrm{u}}^{-}(g)$ will be no more than C_{ℓ}/n_{u} if some $x_i^{\mathrm{u}} \in \mathcal{X}_{\mathrm{u}}$ is replaced. Subsequently, Mc Diarmid's inequality implies

$$\begin{aligned}\Pr\left\{R_{\mathrm{n}}^{-}(g)-\left(\widehat{R}_{\mathrm{u}}^{-}(g)-\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)\right)\geq\alpha\right\} &\leq \exp\left(-\frac{2\alpha^2}{n_{\mathrm{p}}\left(C_{\ell}\pi_{\mathrm{p}}/n_{\mathrm{p}}\right)^2+n_{\mathrm{u}}\left(C_{\ell}/n_{\mathrm{u}}\right)^2}\right)\\ &= \exp\left(-\frac{2\alpha^2/C_{\ell}^2}{\pi_{\mathrm{p}}^2/n_{\mathrm{p}}+1/n_{\mathrm{u}}}\right)\end{aligned} \tag{4}$$

Denote the RHS of this equation by Δ_g

We have

$$\mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)\right]-R(g)\leq\sup_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})\in\mathfrak{D}^{-}(g)}\left(\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)-\widehat{R}_{\mathrm{u}}^{-}(g)\right)\cdot\Delta(g)$$

$$\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)-\widehat{R}_{\mathrm{u}}^{-}(g)=\pi_{\mathrm{p}}\left(1/n_{\mathrm{p}}\right)\sum_{i=1}^{n_{\mathrm{p}}}\ell\left(g\left(x_i^{\mathrm{p}}\right),-1\right)-\left(1/n_{\mathrm{u}}\right)\sum_{i=1}^{n_{\mathrm{u}}}\ell\left(g\left(x_i^{\mathrm{u}}\right),-1\right)$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)\right]-R(g) &\leq \sup_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})\in\mathfrak{D}^{-}(g)}\left(\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)-\widehat{R}_{\mathrm{u}}^{-}(g)\right)\cdot\Delta(g)\\ &\leq C_{\ell}\pi_{\mathrm{p}}\Delta_g\end{aligned}$$

We also have

$$\mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)\right]-R(g)=\mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)-\widehat{R}_{\mathrm{pu}}(g)\right]\geq 0$$

Therefore, As n_{p},$n_{\mathrm{u}}\to\infty$, the bias of $\widetilde{R}_{\mathrm{pu}}(g)$ decays exponentially:

$$0\leq\mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)\right]-R(g)=\mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)-\widehat{R}_{\mathrm{pu}}(g)\right]\leq C_{\ell}\pi_{\mathrm{p}}\Delta_g \tag{5}$$

4.2 consistency

$$\begin{aligned}\left|\widetilde{R}_{\mathrm{pu}}(g)-R(g)\right| &= \left|\widetilde{R}_{\mathrm{pu}}(g)-\mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]+\mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]-R(g)\right|\\ &\leq |\widetilde{R}_{\mathrm{pu}}(g)-\mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]|+|\mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]-R(g)|\\ &\leq |\widetilde{R}_{\mathrm{pu}}(g)-\mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]|+C_{\ell}\pi_{\mathrm{p}}\Delta_g\end{aligned} \tag{6}$$

We can Apply Mcdiarmid's inequality to $\widetilde{R}_{\mathrm{pu}}(g)$,

The change of $\widetilde{R}_{\mathrm{pu}}(g)$ will be no more than $2C_\ell/n_\mathrm{p}$ if some $x_i^\mathrm{p} \in \mathcal{X}_\mathrm{p}$ is replaced, or it will be no more than C_ℓ/n_u if some $x_i^\mathrm{u} \in \mathcal{X}_\mathrm{u}$ is replaced, and McDiarmid's inequality gives us

$$\Pr\left\{\left|\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)\right]\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2\epsilon^2}{n_\mathrm{p}\left(2C_\ell\pi_\mathrm{p}/n_\mathrm{p}\right)^2 + n_\mathrm{u}\left(C_\ell/n_\mathrm{u}\right)^2}\right)$$

Let δ the RHS of the equation. We obtain the equivalent statement: or equivalently, with probability at least $1-\delta$

$$\begin{aligned}\left|\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}\left[\widetilde{R}_{\mathrm{pu}}(g)\right]\right| &\leq \sqrt{\frac{\ln(2/\delta)C_\ell^2}{2}\left(\frac{4\pi_\mathrm{p}^2}{n_\mathrm{p}} + \frac{1}{n_\mathrm{u}}\right)} \\ &= C_\delta\sqrt{\frac{4\pi_\mathrm{p}^2}{n_\mathrm{p}} + \frac{1}{n_\mathrm{u}})} \\ &\leq C_\delta\left(\frac{2\pi_\mathrm{p}}{\sqrt{n_\mathrm{p}}} + \frac{1}{\sqrt{n_\mathrm{u}}}\right) \\ &= C_\delta \cdot \chi_{n_\mathrm{p},n_\mathrm{u}}\end{aligned}$$

where $C_\delta = C_\ell\sqrt{\ln(2/\delta)/2}$, $\chi_{n_\mathrm{p},n_\mathrm{u}} = 2\pi_\mathrm{p}/\sqrt{n_\mathrm{p}} + 1/\sqrt{n_\mathrm{u}}$

This inequality indicates for fixed $g, \widetilde{R}_{\mathrm{pu}}(g) \to R(g)$ in $\mathcal{O}_p\left(\pi_\mathrm{p}/\sqrt{n_\mathrm{p}} + 1/\sqrt{n_\mathrm{u}}\right)$ This convergence rate is optimal according to the central limit theorem, which means the proposed estimator is a biased yet optimal estimator to the risk.

4.3 Mean Squared error

After introducing the bias, $\widetilde{R}_{\mathrm{pu}}(g)$ tends to overestimate $R(g)$.(Since the bias is nonnegative). It is not a shrinkage estimator, so that its mean squared error (MSE) is not necessarily smaller than that of $\widehat{R}_{\mathrm{pu}}(g)$. However, we can still characterize this reduction in MSE.

For convenience, let $A = \pi_\mathrm{p}\widehat{R}_\mathrm{p}^+(g)$ and $B = \widehat{R}_\mathrm{u}^-(g) - \pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g)$, so that

$$R(g) = \mathbb{E}[A+B], \quad \widehat{R}_{\mathrm{pu}}(g) = A+B, \quad \widetilde{R}_{\mathrm{pu}}(g) = A+B_+,$$

where $B_+ = \max\{0, B\}$. Subsequently, let $R = R(g)$ for short, and then by definition,

$$\begin{aligned}\mathrm{MSE}(\widehat{R}_{\mathrm{pu}}(g)) &= \mathbb{E}[(A+B-R)^2] \\ &= \mathbb{E}[(A+B)^2] - 2R\cdot\mathbb{E}[A+B] + R^2, \\ \mathrm{MSE}(\widetilde{R}_{\mathrm{pu}}(g)) &= \mathbb{E}[(A+B_+-R)^2] \\ &= \mathbb{E}[(A+B_+)^2] - 2R\cdot\mathbb{E}[A+B_+] + R^2.\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{MSE}(\widehat{R}_{\mathrm{pu}}(g)) - \mathrm{MSE}(\widetilde{R}_{\mathrm{pu}}(g)) &= \mathbb{E}[(A+B)^2] - \mathbb{E}[(A+B_+)^2] \\ &\quad - 2R\cdot(\mathbb{E}[A+B] - \mathbb{E}[A+B_+]).\end{aligned}$$

The first part $\mathbb{E}[(A+B)^2] - \mathbb{E}[(A+B_+)^2]$ can be rewritten as

$$\begin{aligned}\mathbb{E}[(A+B)^2] - \mathbb{E}[(A+B_+)^2] &= \mathbb{E}[2A(B-B_+) + B^2 - B_+^2] \\ &= \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^+(g)} 2A(B-B) + B^2 - B^2 \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}) \\ &\quad + \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} 2A(B-0) + B^2 - 0^2 \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}) \\ &= \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} 2AB + B^2 \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}).\end{aligned}$$

The second part $2R \cdot (\mathbb{E}[A+B] - \mathbb{E}[A+B_+])$ can be rewritten as

$$\begin{aligned}2R \cdot (\mathbb{E}[A+B] - \mathbb{E}[A+B_+]) &= 2R \cdot \mathbb{E}[B - B_+] \\ &= 2R \cdot \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^+(g)} B - B \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}) \\ &\quad + 2R \cdot \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} B - 0 \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}) \\ &= \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} 2RB \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}).\end{aligned}$$

As a consequence,

$$\mathrm{MSE}(\widehat{R}_\mathrm{pu}(g)) - \mathrm{MSE}(\widetilde{R}_\mathrm{pu}(g)) = \int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} (2A + B - 2R)B \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}),$$

That is,

Theorem 4.1 (MSE reduction). *It holds that* $\mathrm{MSE}(\widetilde{R}_\mathrm{pu}(g)) < \mathrm{MSE}(\widehat{R}_\mathrm{pu}(g))$,[5] *if and only if*

$$\int_{(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})\in\mathfrak{D}^-(g)} (\widehat{R}_\mathrm{pu}(g) + \widetilde{R}_\mathrm{pu}(g) - 2R(g))(\widehat{R}_\mathrm{u}^-(g) - \pi_\mathrm{p}\widehat{R}_\mathrm{p}^-(g)) \,\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}) > 0, \quad (7)$$

where $\mathrm{d}F(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u}) = \prod_{i=1}^{n_\mathrm{p}} p_\mathrm{p}(x_i^\mathrm{p})\mathrm{d}x_i^\mathrm{p} \cdot \prod_{i=1}^{n_\mathrm{u}} p(x_i^\mathrm{u})\mathrm{d}x_i^\mathrm{u}$.

Now we analyze the sufficient condition of Eq. (7) being valid.

We want to get a lower bound of the LHS of Eq. (7). Then it suffices to let the lower bound be positive.

$$\ell(t,+1) + \ell(t,-1) = 1, \quad (8)$$

Since $B \leq 0$ on $\mathfrak{D}^-(g)$, we need to get an upper bound of $2A + B - 2R$.

[5] Here, $\mathrm{MSE}(\cdot)$ is over repeated sampling of $(\mathcal{X}_\mathrm{p},\mathcal{X}_\mathrm{u})$.

$$\begin{aligned}A - R &= A - \mathbb{E}[A] - \mathbb{E}[B] \\ &= \pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{+}(g) - \pi_{\mathrm{p}}R_{\mathrm{p}}^{+}(g) - \mathbb{E}[B] \\ &= \pi_{\mathrm{p}}R_{\mathrm{p}}^{-}(g) - \pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g) - \mathbb{E}[B].\end{aligned}$$

where the third equation is by the assumption that ℓ satisfies (8).

Thus, with probability one,

$$\begin{aligned}A - R &= \pi_{\mathrm{p}}R_{\mathrm{p}}^{-}(g) - \pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g) - \mathbb{E}[B] + (\widehat{R}_{\mathrm{u}}^{-}(g) - \widehat{R}_{\mathrm{u}}^{-}(g)) + (R_{\mathrm{u}}^{-}(g) - R_{\mathrm{u}}^{-}(g)) \\ &= (\widehat{R}_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g)) - (R_{\mathrm{u}}^{-}(g) - \pi_{\mathrm{p}}R_{\mathrm{p}}^{-}(g)) - \mathbb{E}[B] + (R_{\mathrm{u}}^{-}(g) - \widehat{R}_{\mathrm{u}}^{-}(g)) \\ &= B - 2\mathbb{E}[B] + (R_{\mathrm{u}}^{-}(g) - \widehat{R}_{\mathrm{u}}^{-}(g)) \\ &\le B,\end{aligned}$$

where we used the assumptions that $\mathbb{E}[B] \ge \alpha$ and $R_{\mathrm{u}}^{-}(g) - \widehat{R}_{\mathrm{u}}^{-}(g) \le 2\alpha$ almost surely on $\mathfrak{D}^{-}(g)$.

To sum up, we have established that

$$\int_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})\in\mathfrak{D}^{-}(g)} (2A + B - 2R)B\,\mathrm{d}F(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) \ge 3\int_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})\in\mathfrak{D}^{-}(g)} B^2\,\mathrm{d}F(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}).$$

Due to the fact that $B^2 > 0$ on $\mathfrak{D}^{-}(g)$ and the assumption that $\Pr(\mathfrak{D}^{-}(g)) > 0$, we know Eq. (7) is valid.

Finally, for any $0 \le \beta \le C_\ell\pi_{\mathrm{p}}$, it is clear that

$$\{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) \mid B < -\beta\} \subseteq \{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) \mid B < 0\} = \mathfrak{D}^{-}(g),$$

and $B < -\beta$ if and only if $\widetilde{R}_{\mathrm{pu}}(g) - \widehat{R}_{\mathrm{pu}}(g) > \beta$. These two facts imply that

$$\begin{aligned}\int_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})\in\mathfrak{D}^{-}(g)} B^2\,\mathrm{d}F(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) &\ge \int_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})|B<-\beta} B^2\,\mathrm{d}F(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) \\ &\ge \beta^2\int_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})|B<-\beta} \mathrm{d}F(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) \\ &= \beta^2\Pr\{B < -\beta\} \\ &= \beta^2\Pr\{\widetilde{R}_{\mathrm{pu}}(g) - \widehat{R}_{\mathrm{pu}}(g) > \beta\},\end{aligned}$$

Then we have we have for any $0 \le \beta \le C_\ell\pi_{\mathrm{p}}$,

$$\mathrm{MSE}(\widehat{R}_{\mathrm{pu}}(g)) - \mathrm{MSE}(\widetilde{R}_{\mathrm{pu}}(g)) \ge 3\beta^2\Pr\{\widetilde{R}_{\mathrm{pu}}(g) - \widehat{R}_{\mathrm{pu}}(g) > \beta\}. \tag{9}$$

To sum up, we use the following 4 assumptions:

1. $\Pr(\mathfrak{D}^{-}(g)) > 0$;
2. ℓ satisfies Eq. (8);
3. $R_{\mathrm{n}}^{-}(g) \ge \alpha > 0$;

4. $n_\mathrm{u} \gg n_\mathrm{p}$, such that we have $R_\mathrm{u}^-(g) - \widehat{R}_\mathrm{u}^-(g) \le 2\alpha$ almost surely on $\mathfrak{D}^-(g)$.

The 4th assumption is explained as follows. Since *U data can be much cheaper than P data* in practice, it would be natural to assume n_u is much larger and grows much faster than n_p, hence $\Pr\{R_\mathrm{u}^-(g) - \widehat{R}_\mathrm{u}^-(g) \ge \alpha\}/\Pr\{\widehat{R}_\mathrm{p}^-(g) - R_\mathrm{p}^-(g) \ge \alpha/\pi_\mathrm{p}\} \propto \exp(n_\mathrm{p} - n_\mathrm{u})$ asymptotically.[6] This means the contribution of $\mathcal{X}_\mathrm{u}$ is negligible for making $(\mathcal{X}_\mathrm{p}, \mathcal{X}_\mathrm{u}) \in \mathfrak{D}^-(g)$ so that $\Pr(\mathfrak{D}^-(g))$ exhibits exponential decay mainly in n_p. As $\Pr\{R_\mathrm{u}^-(g) - \widehat{R}_\mathrm{u}^-(g) \ge 2\alpha\}$ has stronger exponential decay in n_u than $\Pr\{R_\mathrm{u}^-(g) - \widehat{R}_\mathrm{u}^-(g) \ge \alpha\}$ as well as $n_\mathrm{u} \gg n_\mathrm{p}$, we made the 4th assumption.

5 The estimation error bound of $\widetilde{g}_\mathrm{pu}$

We are likewise interested in its use for training classifiers. In what follows, we analyze the estimation error $R(\widetilde{g}_\mathrm{pu}) - R(g^*)$, where g^* is the true risk minimizer in $\mathcal{G}$, i.e., $g^* = \arg\min_{g\in\mathcal{G}} R(g)$.

As a common practice, we assume that $\ell(t, y)$ is Lipschitz continuous in t for all $|t| \le C_g$ with a Lipschitz constant L_ℓ.

Theorem 5.1 (Estimation error bound)**.** *Assume that*

1. *(a)* $\inf_{g\in\mathcal{G}} R_\mathrm{n}^-(g) \ge \alpha > 0$
2. *(b) $\mathcal{G}$ is closed under negation, i.e., $g \in \mathcal{G}$ if and only if $-g \in \mathcal{G}$.*

denote by Δ the right-hand side of Eq. (4);
Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$R(\widetilde{g}_\mathrm{pu}) - R(g^*) \le 16 L_\ell \pi_\mathrm{p} \mathfrak{R}_{n_\mathrm{p},p_\mathrm{p}}(\mathcal{G}) + 8 L_\ell \mathfrak{R}_{n_\mathrm{u},p}(\mathcal{G}) + 2C'_\delta \cdot \chi_{n_\mathrm{p},n_\mathrm{u}} + 2C_\ell \pi_\mathrm{p} \Delta, \quad (10)$$

where $C'_\delta = C_\ell\sqrt{\ln(1/\delta)/2}$, and $\mathfrak{R}_{n_\mathrm{p},p_\mathrm{p}}(\mathcal{G})$ and $\mathfrak{R}_{n_\mathrm{u},p}(\mathcal{G})$ are the Rademacher complexities of $\mathcal{G}$ for the sampling of size n_p from $p_\mathrm{p}(x)$ and of size n_u from $p(x)$, respectively.

Proof.

$$\begin{aligned}
R(\widetilde{g}_\mathrm{pu}) - R(g^*) &= \Big(R(\widetilde{g}_\mathrm{pu}) - \widetilde{R}_\mathrm{pu}(\widetilde{g}_\mathrm{pu})\Big) + \Big(\widetilde{R}_\mathrm{pu}(\widetilde{g}_\mathrm{pu}) - \widetilde{R}_\mathrm{pu}(g^*)\Big) + \Big(\widetilde{R}_\mathrm{pu}(g^*) - R(g^*)\Big) \\
&= \Big(\widetilde{R}_\mathrm{pu}(\widetilde{g}_\mathrm{pu}) - \widetilde{R}_\mathrm{pu}(g^*)\Big) + \Big(R(\widetilde{g}_\mathrm{pu}) - \widetilde{R}_\mathrm{pu}(\widetilde{g}_\mathrm{pu})\Big) + \Big(\widetilde{R}_\mathrm{pu}(g^*) - R(g^*)\Big) \\
&\le 0 + 2\sup\nolimits_{g\in\mathcal{G}} |\widetilde{R}_\mathrm{pu}(g) - R(g)|
\end{aligned}$$

where $\widetilde{R}_\mathrm{pu}(\widetilde{g}_\mathrm{pu}) \le \widetilde{R}_\mathrm{pu}(g^*)$ by the definition of $\widetilde{g}_\mathrm{pu}$. □ □

Next, we will focus on $\sup_{g\in\mathcal{G}} |\widetilde{R}_\mathrm{pu}(g) - R(g)|$

[6] This can be derived as $n_\mathrm{p}, n_\mathrm{u} \to \infty$ by applying the *central limit theorem* to the two differences and then *L'Hôpital's rule* to the ratio of *complementary error functions* .

Lemma 5.2. *Under the assumptions of Theorem 5.1, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - R(g)| \le 8L_\ell \pi_{\mathrm{p}} \mathfrak{R}_{n_{\mathrm{p}},p_{\mathrm{p}}}(\mathcal{G}) + 4L_\ell \mathfrak{R}_{n_{\mathrm{u}},p}(\mathcal{G}) + C'_\delta \cdot \chi_{n_{\mathrm{p}},n_{\mathrm{u}}} + C_\ell \pi_{\mathrm{p}} \Delta. \tag{11}$$

In the next, we will prove Lemma 5.2

Preliminary An alternative definition of the Rademacher complexity will be used in the proof:

$$\mathfrak{R}'_{n,q}(\mathcal{G}) = \mathbb{E}_{\mathcal{X}} \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum\nolimits_{x_i \in \mathcal{X}} \sigma_i g(x_i) \right| \right].$$

For the sake of comparison, the one we have used in the statements of theoretical results is

$$\mathfrak{R}_{n,q}(\mathcal{G}) = \mathbb{E}_{\mathcal{X}} \mathbb{E}_{\sigma_1,\dots,\sigma_n} \left[\sup_{g\in\mathcal{G}} \frac{1}{n} \sum\nolimits_{x_i \in \mathcal{X}} \sigma_i g(x_i) \right].$$

This alternative version comes from [Koltchinskii, 2001] of which authors are the pioneers of error bounds based on the Rademacher complexity. Without any composition, $\mathfrak{R}'_{n,q}(\mathcal{G}) \ge \mathfrak{R}_{n,q}(\mathcal{G})$ for arbitrary $\mathcal{G}$ and $\mathfrak{R}'_{n,q}(\mathcal{G}) = \mathfrak{R}_{n,q}(\mathcal{G})$ if $\mathcal{G}$ is closed under negation. However, with a composition

$$\ell \circ \mathcal{G} = \{\ell \circ g \mid g \in \mathcal{G}\}$$

where the loss ℓ is non-negative, the Rademacher complexity of the *composite function class* would generally not satisfy $\mathfrak{R}'_{n,q}(\ell \circ \mathcal{G}) = \mathfrak{R}_{n,q}(\ell \circ \mathcal{G})$ since $\ell \circ \mathcal{G}$ is generally not closed under negation. Furthermore, a vital disagreement arises when considering the contraction principle or property: if $\psi : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous function with a Lipschitz constant L_ψ and satisfies $\psi(0) = 0$, we have

$$\mathfrak{R}_{n,q}(\psi \circ \mathcal{G}) \le L_\psi \mathfrak{R}_{n,q}(\mathcal{G}),$$
$$\mathfrak{R}'_{n,q}(\psi \circ \mathcal{G}) \le 2L_\psi \mathfrak{R}'_{n,q}(\mathcal{G}),$$

according to *Talagrand's contraction lemma* [Ledoux and Talagrand, 2013] and its extension [Mohri et al., 2012, Shalev-Shwartz and Ben-David, 2014]. Here, for $\mathfrak{R}_{n,q}(\psi \circ \mathcal{G})$ we can use Lemma 4.2 in [Mohri et al., 2012] or Lemma 26.9 in [Shalev-Shwartz and Ben-David, 2014] where $\psi(0) = 0$ is safely dropped, while for $\mathfrak{R}'_{n,q}(\psi \circ \mathcal{G})$ we have to use the original Theorem 4.12 in [Ledoux and Talagrand, 2013] where $\psi(0) = 0$ is required. In fact, the name of the lemma is after that ψ is a contraction if $\psi(0) = 0$ and $L_\psi = 1$.

Proof Firstly, we deal with the bias of $\widetilde{R}_{\mathrm{pu}}(g)$:

$$\begin{aligned}
\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - R(g)| &= \sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)] + \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)] - R(g)| \qquad (12)\\
&\le \sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]| + \sup_{g\in\mathcal{G}} |\mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)] - R(g)|\\
&\le \sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]| + C_\ell \pi_{\mathrm{p}} \Delta, \qquad (13)
\end{aligned}$$

where we followed the assumption that $\inf_{g\in\mathcal{G}} R_{\mathrm{n}}^{-}(g) \geq \alpha > 0$ and Theorem (5).

Secondly, we apply McDiarmid's inequality to the uniform deviation $\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]|$ to get that with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]| - \mathbb{E}[\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]|] \leq C'_{\delta} \cdot \chi_{n_{\mathrm{p}},n_{\mathrm{u}}}. \quad (14)$$

Notice that this concentration inequality is single-sided even though the uniform deviation itself is double-sided, which is different from the non-uniform deviation in Theorem 5.

Thirdly, we make *symmetrization.* Suppose that $(\mathcal{X}'_{\mathrm{p}}, \mathcal{X}'_{\mathrm{u}})$ is a *ghost sample*, then

$$\begin{aligned}
\mathbb{E}[\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]|] &= \mathbb{E}_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})}[\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) - \mathbb{E}_{(\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})}[\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})]|] \\
&= \mathbb{E}_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})}[\sup_{g\in\mathcal{G}} |\mathbb{E}_{(\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})}[\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})] - \mathbb{E}_{(\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})}[\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})]|] \\
&= \mathbb{E}_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})}[\sup_{g\in\mathcal{G}} |\mathbb{E}_{(\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})}[\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) - \widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})]|] \\
&\leq \mathbb{E}_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}})}[\sup_{g\in\mathcal{G}} \mathbb{E}_{(\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})}[|\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) - \widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})|]] \\
&\leq \mathbb{E}_{(\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}),(\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})}[\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) - \widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})|],
\end{aligned}$$

where we applied *Jensen's inequality* twice since the absolute value and the supremum are convex.

By decomposing the difference $|\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) - \widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})|$, we can know that

$$\begin{aligned}
&|\widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}_{\mathrm{p}},\mathcal{X}_{\mathrm{u}}) - \widetilde{R}_{\mathrm{pu}}(g;\mathcal{X}'_{\mathrm{p}},\mathcal{X}'_{\mathrm{u}})| \\
&\quad = |\pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{+}(g;\mathcal{X}_{\mathrm{p}}) - \pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{+}(g;\mathcal{X}'_{\mathrm{p}}) \\
&\qquad + \max\{0, \widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{u}}) - \pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g;\mathcal{X}_{\mathrm{p}})\} - \max\{0, \widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}'_{\mathrm{u}}) - \pi_{\mathrm{p}}\widehat{R}_{\mathrm{p}}^{-}(g;\mathcal{X}'_{\mathrm{p}})\}| \\
&\quad \leq \pi_{\mathrm{p}}|\widehat{R}_{\mathrm{p}}^{+}(g;\mathcal{X}_{\mathrm{p}}) - \widehat{R}_{\mathrm{p}}^{+}(g;\mathcal{X}'_{\mathrm{p}})| + \pi_{\mathrm{p}}|\widehat{R}_{\mathrm{p}}^{-}(g;\mathcal{X}_{\mathrm{p}}) - \widehat{R}_{\mathrm{p}}^{-}(g;\mathcal{X}'_{\mathrm{p}})| + |\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{u}}) - \widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}'_{\mathrm{u}})|
\end{aligned}$$

where we employed $|\max\{0,z\} - \max\{0,z'\}| \leq |z-z'|$. This decomposition results in

$$\begin{aligned}
\mathbb{E}[\sup_{g\in\mathcal{G}} |\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]|] \leq\ & \pi_{\mathrm{p}}\mathbb{E}_{\mathcal{X}_{\mathrm{p}},\mathcal{X}'_{\mathrm{p}}}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{p}}^{+}(g;\mathcal{X}_{\mathrm{p}}) - \widehat{R}_{\mathrm{p}}^{+}(g;\mathcal{X}'_{\mathrm{p}})|] \\
&+ \pi_{\mathrm{p}}\mathbb{E}_{\mathcal{X}_{\mathrm{p}},\mathcal{X}'_{\mathrm{p}}}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{p}}^{-}(g;\mathcal{X}_{\mathrm{p}}) - \widehat{R}_{\mathrm{p}}^{-}(g;\mathcal{X}'_{\mathrm{p}})|] \\
&+ \mathbb{E}_{\mathcal{X}_{\mathrm{u}},\mathcal{X}'_{\mathrm{u}}}[\sup_{g\in\mathcal{G}} |\widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}_{\mathrm{u}}) - \widehat{R}_{\mathrm{u}}^{-}(g;\mathcal{X}'_{\mathrm{u}})|]. \quad (15)
\end{aligned}$$

Fourthly, we relax those expectations in (15) to Rademacher complexities. The original ℓ may miss the origin, i.e., $\ell(0,y) \neq 0$, with which we need to cope. Let

$$\tilde{\ell}(t,y) = \ell(t,y) - \ell(0,y)$$

be a *shifted loss* so that $\tilde{\ell}(0,y) = 0$. Note that for all $t, t' \in \mathbb{R}$ and $y = \pm 1$,

$$\ell(t,y) - \ell(t',y) = \tilde{\ell}(t,y) - \tilde{\ell}(t',y).$$

Hence,

$$\begin{aligned}\widehat{R}^+_{\mathrm{p}}(g;\mathcal{X}_{\mathrm{p}}) - \widehat{R}^+_{\mathrm{p}}(g;\mathcal{X}'_{\mathrm{p}}) &= (1/n_{\mathrm{p}})\textstyle\sum_{x_i\in\mathcal{X}_{\mathrm{p}}}\ell(g(x_i),+1) - (1/n_{\mathrm{p}})\sum_{x'_i\in\mathcal{X}'_{\mathrm{p}}}\ell(g(x'_i),+1)\\ &= (1/n_{\mathrm{p}})\textstyle\sum_{i=1}^{n_{\mathrm{p}}}(\ell(g(x_i),+1) - \ell(g(x'_i),+1))\\ &= (1/n_{\mathrm{p}})\textstyle\sum_{i=1}^{n_{\mathrm{p}}}(\tilde{\ell}(g(x_i),+1) - \tilde{\ell}(g(x'_i),+1)).\end{aligned}$$

This is already a standard form where we can attach Rademacher variables to every $\tilde{\ell}(g(x_i),+1) - \tilde{\ell}(g(x'_i),+1)$, and it is a routine work to show that

$$\mathbb{E}_{\mathcal{X}_{\mathrm{p}},\mathcal{X}'_{\mathrm{p}}}[\sup_{g\in\mathcal{G}}|\widehat{R}^+_{\mathrm{p}}(g;\mathcal{X}_{\mathrm{p}}) - \widehat{R}^+_{\mathrm{p}}(g;\mathcal{X}'_{\mathrm{p}})|] \le 2\mathfrak{R}_{n_{\mathrm{p}},p_{\mathrm{p}}}(\tilde{\ell}(\cdot,+1)\circ\mathcal{G}).$$

The other two expectations can be handled analogously. As a result, (15) can be reduced to

$$\begin{aligned}\mathbb{E}[\sup_{g\in\mathcal{G}}|\widetilde{R}_{\mathrm{pu}}(g) - \mathbb{E}[\widetilde{R}_{\mathrm{pu}}(g)]|] \le{}& 2\pi_{\mathrm{p}}\mathfrak{R}'_{n_{\mathrm{p}},p_{\mathrm{p}}}(\tilde{\ell}(\cdot,+1)\circ\mathcal{G})\\ &+ 2\pi_{\mathrm{p}}\mathfrak{R}'_{n_{\mathrm{p}},p_{\mathrm{p}}}(\tilde{\ell}(\cdot,-1)\circ\mathcal{G}) + 2\mathfrak{R}'_{n_{\mathrm{u}},p}(\tilde{\ell}(\cdot,-1)\circ\mathcal{G}).\end{aligned} \tag{16}$$

Finally, we transform the Rademacher complexities of composite function classes in (16) to those of the original function class. It is obvious that $\tilde{\ell}$ shares the same Lipschitz constant L_ℓ with ℓ, and consequently

$$\begin{aligned}\mathfrak{R}'_{n_{\mathrm{p}},p_{\mathrm{p}}}(\tilde{\ell}(\cdot,+1)\circ\mathcal{G}) &\le 2L_\ell\mathfrak{R}'_{n_{\mathrm{p}},p_{\mathrm{p}}}(\mathcal{G}) = 2L_\ell\mathfrak{R}_{n_{\mathrm{p}},p_{\mathrm{p}}}(\mathcal{G})\\ \mathfrak{R}'_{n_{\mathrm{p}},p_{\mathrm{p}}}(\tilde{\ell}(\cdot,-1)\circ\mathcal{G}) &\le 2L_\ell\mathfrak{R}'_{n_{\mathrm{p}},p_{\mathrm{p}}}(\mathcal{G}) = 2L_\ell\mathfrak{R}_{n_{\mathrm{p}},p_{\mathrm{p}}}(\mathcal{G})\\ \mathfrak{R}'_{n_{\mathrm{u}},p}(\tilde{\ell}(\cdot,-1)\circ\mathcal{G}) &\le 2L_\ell\mathfrak{R}'_{n_{\mathrm{u}},p}(\mathcal{G}) = 2L_\ell\mathfrak{R}_{n_{\mathrm{u}},p}(\mathcal{G}),\end{aligned} \tag{17}$$

where we used Talagrand's contraction lemma and the assumption that $\mathcal{G}$ is closed under negation. Combining (13), (14), (16) and (17) finishes the proof of the uniform deviation bound (11). □

Notice that $\widehat{R}_{\mathrm{pu}}(g)$ is point-wise while $\widetilde{R}_{\mathrm{pu}}(g)$ is not due to the maximum, which makes Lemma 5.2 much more difficult to prove than Lemma 8 of [Niu et al., 2016]. The key trick is that after *symmetrization*, we employ $|\max\{0,z\} - \max\{0,z'\}| \le |z - z'|$, making three differences of partial risks point-wise (see (15) in the proof). As a consequence, we have to use a different Rademacher complexity *with the absolute value inside the supremum* whose *contraction* makes the coefficients of (11) doubled compared with Lemma 8 of [Niu et al., 2016] moreover, we have to assume $\mathcal{G}$ is closed under negation to change back to the standard Rademacher complexity *without the absolute value*

6 Assumption

- π_{p} is known

References

[Casella and Berger, 2002] Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

[Koltchinskii, 2001] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

[Ledoux and Talagrand, 2013] Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

[Mohri et al., 2012] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). Foundations of machine learning (2012). *Google Scholar*, pages 198–199.

[Niu et al., 2016] Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., and Sugiyama, M. (2016). Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in neural information processing systems*, pages 1199–1207.

[Shalev-Shwartz and Ben-David, 2014] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.