Theoretical Comparisons of Learning from Positive-Negative, Positive-Unlabeled, and Negative-Unlabeled Data

Gang Niu GANG@MS.K.U-TOKYO.AC.JP
Marthinus Christoffel du Plessis CHRISTO@MS.K.U-TOKYO.AC.JP
Tomoya Sakai SAKAI@MS.K.U-TOKYO.AC.JP
Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

Abstract

In PU learning, a binary classifier is trained only from *positive* (P) and *unlabeled* (U) data without *negative* (N) data. Although N data is missing, it sometimes outperforms PN learning (i.e., supervised learning) in experiments. In this paper, we theoretically compare PU (and the opposite NU) learning against PN learning, and prove that, one of PU and NU learning given infinite U data will *almost always* improve on PN learning. Our theoretical finding is also validated experimentally.

1. Introduction

PU learning, where a binary classifier is trained from *positive* (P) and *unlabeled* (U) data without *negative* (N) data, has drawn considerable attention recently (Elkan & Noto, 2008; Scott & Blanchard, 2009; Blanchard et al., 2010; du Plessis et al., 2014; 2015b). Although it uses no N data for training, it is shown useful in real-world applications (e.g., Li et al., 2011), and sometimes can be even better than PN learning (i.e., supervised learning, perhaps with class-prior change). However, it is an open question under what conditions PU learning can outperform PN learning.

In this paper, we try to answer this question by taking into account both PU and NU learning. Specifically, du Plessis et al. (2014) proved in the *cost-sensitive formula* in Elkan & Noto (2008) for PU learning, any convex surrogate loss results in a biased estimator to the risk; but there is no bias if a non-convex surrogate loss satisfies a certain *symmetric condition*. Under this symmetric condition, risk estimators in NU learning are also unbiased. Hence, we can naturally compare risk-minimizing learners in PU and NU learning against those in PN learning, theoretically.

Under review by the International Conference on Machine Learning (ICML).

We compare six learners that minimize six estimators. For every problem setting, there are a *double-expectation estimator* (DEE) and a *single-expectation estimator* (SEE). A DEE regards data as being drawn from two genuine densities, and an SEE regards data as being drawn from a single artificial density. We name a learner minimizing a DEE or SEE as *DE learner* (DEL) or *SE learner* (SEL).[1]

We establish risk bounds of all learners for comparisons in a flavor of *statistical learning theory* (Vapnik, 1998; Bousquet et al., 2004). For each learner, we firstly derive a *uniform deviation bound* using the *Rademacher complexities*[2] and secondly obtain an *estimation error bound*. Thirdly, if the loss function is *classification-calibrated* (Bartlett et al., 2006), an *excess risk bound* is an immediate corollary.

Our main results, which are totally in terms of the estimation error bounds, can be summarized as follows:

- PU-DEL and NU-DEL become better as n_{u} increases; one of them will *almost always* improve on PN-DEL, as long as $n_{\mathrm{u}} \to \infty$ faster than n_+ and n_- (in Theorem 1 and Comparisons 3 and 6).
- There are thresholds of n_{u} for PU-SEL and NU-SEL, such that they become worse as n_{u} increases further; for some n_+ and n_-, neither of them would improve on PN-SEL (in Theorem 2 and Comparisons 4 and 7).
- PN-SEL may very likely win against PN-DEL, when it beats PU-SEL and NU-SEL for optimal n_{u}; in fact, PU-DEL must be inferior to PU-SEL with its optimal n_{u}, and so must be NU-DEL to NU-SEL (in Comparisons 1, 2 and 5).

To the best of our knowledge, there is no theoretical comparison like ours. Our technique for comparisons is novel,

[1] SELs, but not DELs, can be written as special cases of learning with noisy labels (Natarajan et al., 2013).

[2] See Koltchinskii (2001); Bartlett & Mendelson (2002); Meir & Zhang (2003) for the initial and typical uses in machine learning and Mohri et al. (2012) for a comprehensive reference.

though the bounds are proved using existing techniques.

In our analyses, we assume that the true class-prior probability is known, while it can be effectively estimated from PNU (Saerens et al., 2002; du Plessis & Sugiyama, 2012) or PU (Scott & Blanchard, 2009; du Plessis et al., 2015a).

The rest of this paper is organized as follows. Three problem settings are described in Section 2, and six estimators are discussed in Section 3. Then, in Section 4 we give two main results, and in Section 5 we present all technical details. Proofs are included in Section 6. Numerical illustrations can be found in the supplementary material.

2. Problem Settings

Let random variables $X \in \mathbb{R}^d$ where $d \in \mathbb{N}$ and $Y = \pm 1$, equipped with *underlying joint probability density* $p(x, y)$. Denote by

$$\pi = p(y = +1),$$
$$p_+(x) = p(x \mid y = +1), p_-(x) = p(x \mid y = -1).$$

Suppose we have no data from $p(x, y)$, but instead $\mathcal{X}_+$ of size n_+ from $p_+(x)$ as well as $\mathcal{X}_-$ of size n_- from $p_-(x)$ are observed. Let $n_{\rm pn} = n_+ + n_-$, $\tau_{\rm pn} = n_+/n_{\rm pn}$ be the *pseudo PN class-prior*, $Z = \pm 1$ be an indicator variable of PN data, and

$$q_{\rm pn}(x, z) = \begin{cases} \tau_{\rm pn} p_+(x), & z = +1, \\ (1 - \tau_{\rm pn}) p_-(x), & z = -1, \end{cases}$$

be the *artificial joint density* of PN data. As a result,

$$\mathcal{S}_{\rm pn} = \{(x, +1) \mid x \in \mathcal{X}_+\} \cup \{(x, -1) \mid x \in \mathcal{X}_-\}$$

can be regarded as a sampling of size $n_{\rm pn}$ from $q_{\rm pn}(x, z)$. We refer to this problem setting as *PN learning*.

In addition, suppose we have $\mathcal{X}_{\rm u}$ of size $n_{\rm u}$ from $p(x)$. Let $n_{\rm pu} = n_+ + n_{\rm u}$, $\tau_{\rm pu} = n_+/n_{\rm pu}$ and

$$q_{\rm pu}(x, z) = \begin{cases} \tau_{\rm pu} p_+(x), & z = +1, \\ (1 - \tau_{\rm pu}) p(x), & z = -1, \end{cases}$$

for PU data. Now

$$\mathcal{S}_{\rm pu} = \{(x, +1) \mid x \in \mathcal{X}_+\} \cup \{(x, -1) \mid x \in \mathcal{X}_{\rm u}\}$$

can be regarded as a sampling of size $n_{\rm pu}$ from $q_{\rm pu}(x, z)$. Similarly, let $n_{\rm nu} = n_- + n_{\rm u}$, $\tau_{\rm nu} = n_{\rm u}/n_{\rm nu}$ and

$$q_{\rm nu}(x, z) = \begin{cases} \tau_{\rm nu} p(x), & z = +1, \\ (1 - \tau_{\rm nu}) p_-(x), & z = -1, \end{cases}$$

for NU data, and

$$\mathcal{S}_{\rm nu} = \{(x, +1) \mid x \in \mathcal{X}_{\rm u}\} \cup \{(x, -1) \mid x \in \mathcal{X}_-\}$$

can be regarded as a sampling of size $n_{\rm nu}$ from $q_{\rm nu}(x, z)$. We refer to these problem settings as *PU learning* and *NU learning* respectively.

3. Unbiased Estimators to the Risk

In this section, we explain six unbiased estimators.

3.1. PN learning

Let $g : \mathbb{R}^d \mapsto \mathbb{R}$ be an arbitrary real-valued *decision function* for binary classification and $\ell : \mathbb{R} \times \{\pm 1\} \mapsto \mathbb{R}$ be a *bounded Lipschitz-continuous loss function*. Denote by

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

where $\mathbb{E}_\pm[\cdot] = \mathbb{E}_{X \sim p_\pm}[\cdot]$. Then the *risk* of g w.r.t. ℓ under $p(x, y)$ is given by

$$\begin{aligned} R(g) &= \mathbb{E}_{(X,Y)}[\ell(g(X), Y)] \\ &= \pi R_+(g) + (1 - \pi) R_-(g) \qquad (1) \\ &= \tau_{\rm pn} \mathbb{E}_+ \left[\tfrac{\pi}{\tau_{\rm pn}} \ell(g(X), +1)\right] \\ &\quad + (1 - \tau_{\rm pn}) \mathbb{E}_- \left[\tfrac{1-\pi}{1-\tau_{\rm pn}} \ell(g(X), -1)\right] \\ &= \mathbb{E}_{(X,Z) \sim q_{\rm pn}}[\ell_{\rm pn}(g(X), Z)], \qquad (2) \end{aligned}$$

where the loss function $\ell_{\rm pn}$ is defined by

$$\ell_{\rm pn}(t, z) = \begin{cases} (\pi/\tau_{\rm pn}) \ell(t, +1), & z = +1, \\ ((1 - \pi)/(1 - \tau_{\rm pn})) \ell(t, -1), & z = -1. \end{cases}$$

In other words, the risk of g w.r.t. ℓ under $p(x, y)$ is equivalent to that of g w.r.t. $\ell_{\rm pn}$ under $q_{\rm pn}(x, z)$, which is known in machine learning (Quiñonero-Candela et al., 2009).

Consider the empirical versions of $R(g)$. If approximating $R(g)$ based on Eq. (1), we can obtain a *double-expectation estimator* (DEE)

$$\begin{aligned} \widehat{R}_1(g) = &\tfrac{\pi}{n_+} \textstyle\sum_{x \in \mathcal{X}_+} \ell(g(x), +1) \\ &+ \tfrac{1-\pi}{n_-} \textstyle\sum_{x \in \mathcal{X}_-} \ell(g(x), -1), \end{aligned}$$

whose convergence rate is $\mathcal{O}_p(1/\sqrt{n_+} + 1/\sqrt{n_-})$. On the other hand if approximating $R(g)$ based on Eq. (2), we can obtain a *single-expectation estimator* (SEE)

$$\widehat{R}_2(g) = \tfrac{1}{n_{\rm pn}} \textstyle\sum_{(x,z) \in \mathcal{S}_{\rm pn}} \ell_{\rm pn}(g(x), z),$$

whose convergence rate is $\mathcal{O}_p(1/\sqrt{n_{\rm pn}})$. At a glance, the empirical approximation based on (2) will converge to the risk faster in order than that based on (1). However, notice that $\lim_{n_+, n_- \to \infty} \widehat{R}_2(g)$ exists only if

$$0 < \lim_{n_+ \to \infty, n_- \to \infty} n_+/n_- < \infty, \qquad (3)$$

otherwise $\ell_{\rm pn}$ will be unbounded in the limit. Thus, $\widehat{R}_2(g)$ is not superior to but less general than $\widehat{R}_1(g)$, and (3) implies the same order of their convergence rates.[3]

[3]Note that $\widehat{R}_1(g)$ and $\widehat{R}_2(g)$ coincide, if $\mathcal{X}_+$ and $\mathcal{X}_-$ are re-

3.2. PU learning

Let $R_{\mathrm{u},-}(g) = \mathbb{E}_X[\ell(g(X),-1)]$. du Plessis et al. (2014) has shown that with the following *symmetric condition*

$$\ell(t,+1) + \ell(t,-1) = 1, \tag{4}$$

we have $R_{\mathrm{u},-}(g) = \pi(1 - R_+(g)) + (1-\pi)R_-(g)$, and hence

$$R(g) = 2\pi R_+(g) + R_{\mathrm{u},-}(g) - \pi \tag{5}$$
$$= \mathbb{E}_{(X,Z)\sim q_{\mathrm{pu}}}[\ell_{\mathrm{pu}}(g(X),Z)] - \pi, \tag{6}$$

where the loss function ℓ_{pu} is defined by

$$\ell_{\mathrm{pu}}(t,z) = \begin{cases} (2\pi/\tau_{\mathrm{pu}})\ell(t,+1), & z=+1, \\ (1/(1-\tau_{\mathrm{pu}}))\ell(t,-1), & z=-1. \end{cases}$$

Hence, the risk of g w.r.t. ℓ under $p(x,y)$ is also equivalent to that w.r.t. ℓ_{pu} under $q_{\mathrm{pu}}(x,z)$ minus π.

Similarly, if approximating $R(g)$ based on (5), we can obtain a DEE

$$\widehat{R}_3(g) = -\pi + \tfrac{2\pi}{n_+}\textstyle\sum_{x\in\mathcal{X}_+}\ell(g(x),+1) + \tfrac{1}{n_{\mathrm{u}}}\sum_{x\in\mathcal{X}_{\mathrm{u}}}\ell(g(x),-1),$$

whose convergence rate is $\mathcal{O}_p(1/\sqrt{n_+} + 1/\sqrt{n_{\mathrm{u}}})$, and if approximating $R(g)$ based on (6), we can obtain an SEE

$$\widehat{R}_4(g) = -\pi + \tfrac{1}{n_{\mathrm{pu}}}\textstyle\sum_{(x,z)\in\mathcal{S}_{\mathrm{pu}}}\ell_{\mathrm{pu}}(g(x),z),$$

which converges in $\mathcal{O}_p(1/\sqrt{n_{\mathrm{pu}}})$. $\lim_{n_+,n_{\mathrm{u}}\to\infty}\widehat{R}_4(g)$ exists only if

$$0 < \lim_{n_+\to\infty,n_{\mathrm{u}}\to\infty} n_+/n_{\mathrm{u}} < \infty, \tag{7}$$

otherwise ℓ_{pu} will be unbounded in the limit, which means $\widehat{R}_4(g)$ is also not superior to but less general than $\widehat{R}_3(g)$.

3.3. NU learning

Likewise, $R(g)$ could be estimated using NU data. Denote by $R_{\mathrm{u},+}(g) = \mathbb{E}_X[\ell(g(X),+1)]$. Under (4), we also have $R_{\mathrm{u},+}(g) = \pi R_+(g) + (1-\pi)(1-R_-(g))$, and thus

$$R(g) = R_{\mathrm{u},+}(g) + 2(1-\pi)R_-(g) - (1-\pi) \tag{8}$$
$$= \mathbb{E}_{(X,Z)\sim q_{\mathrm{nu}}}[\ell_{\mathrm{nu}}(g(X),Z)] - (1-\pi), \tag{9}$$

where the loss function ℓ_{nu} is defined by

$$\ell_{\mathrm{nu}}(t,z) = \begin{cases} (1/\tau_{\mathrm{nu}})\ell(t,+1), & z=+1, \\ (2(1-\pi)/(1-\tau_{\mathrm{nu}}))\ell(t,-1), & z=-1. \end{cases}$$

Eqs. (8) and (9) result in DEE and SEE which converge in $\mathcal{O}_p(1/\sqrt{n_-} + 1/\sqrt{n_{\mathrm{u}}})$ and $\mathcal{O}_p(1/\sqrt{n_{\mathrm{nu}}})$, respectively:

$$\widehat{R}_5(g) = -(1-\pi) + \tfrac{1}{n_{\mathrm{u}}}\textstyle\sum_{x\in\mathcal{X}_{\mathrm{u}}}\ell(g(x),+1) + \tfrac{2(1-\pi)}{n_-}\sum_{x\in\mathcal{X}_-}\ell(g(x),-1),$$
$$\widehat{R}_6(g) = -(1-\pi) + \tfrac{1}{n_{\mathrm{nu}}}\textstyle\sum_{(x,z)\in\mathcal{S}_{\mathrm{nu}}}\ell_{\mathrm{nu}}(g(x),z).$$

Again, $\lim_{n_-,n_{\mathrm{u}}\to\infty}\widehat{R}_6(g)$ exists only if

$$0 < \lim_{n_-\to\infty,n_{\mathrm{u}}\to\infty} n_-/n_{\mathrm{u}} < \infty, \tag{10}$$

and $\widehat{R}_6(g)$ is not superior to but less general than $\widehat{R}_5(g)$.

3.4. On the loss function

Let $\ell_{01}(t,y) = (1-\mathrm{sign}(ty))/2$ be the *zero-one loss* and

$$I(g) = \mathbb{E}_{(X,Y)}[\ell_{01}(g(X),Y)]$$

the risk of g w.r.t. ℓ_{01} under $p(x,y)$. Let $R^* = \inf_g R(g)$ and $I^* = \inf_g I(g)$ denote the Bayes risks w.r.t ℓ and ℓ_{01} where the infimum is over all measurable functions. If ℓ is *classification-calibrated*[4], we are able to control the excess risk w.r.t. ℓ_{01} by that w.r.t. ℓ (Bartlett et al., 2006).

Instead of the popular *hinge loss* and *ramp loss*, du Plessis et al. (2014) proposed to use a *scaled ramp loss* satisfying (4) as the surrogate loss function for ℓ_{01} in PU learning:

$$\ell_{\mathrm{sr}}(t,y) = \max\{0,\min\{1,(1-ty)/2\}\},$$

but it is still an open question whether ℓ_{sr} is classification-calibrated. Here we show indeed it is and therefore a safe surrogate loss for ℓ_{01}. Denote by $\pi_+(x) = p(y=+1 \mid x)$ and $\pi_-(x) = p(y=-1 \mid x)$, then the *conditional risk* is

$$\begin{aligned}&\mathbb{E}_Y[\ell_{\mathrm{sr}}(g(X),Y) \mid X = x] \\ &= \ell_{\mathrm{sr}}(g(x),+1)\pi_+(x) + \ell_{\mathrm{sr}}(g(x),-1)\pi_-(x) \\ &= \begin{cases} \pi_+(x), & g(x) \le -1, \\ \frac{1-g(x)}{2}\pi_+(x) + \frac{1+g(x)}{2}\pi_-(x), & -1 < g(x) < +1, \\ \pi_-(x), & g(x) \ge +1. \end{cases}\end{aligned}$$

The minimum is achieved by $g(x) = \mathrm{sign}(\pi_+(x) - \pi_-(x))$ and it is consistent with the Bayes rule. As a consequence, ℓ_{sr} is classification-calibrated according to *Theorem 1.3c* in Bartlett et al. (2006).

peatedly sampled and $\mathcal{S}_{\mathrm{pn}}$ is always constructed from them. But if τ_{pn} is fixed and $\mathcal{S}_{\mathrm{pn}}$ itself is repeatedly sampled, there will not be necessarily n_+ positive and n_- negative data every time. Moreover, n_+ and n_- can *increase independently* when sampling $\mathcal{X}_+$ and $\mathcal{X}_-$, while they should *increase together* when sampling $\mathcal{S}_{\mathrm{pn}}$ under the constraint $n_+ : n_- = \tau_{\mathrm{pn}} : (1-\tau_{\mathrm{pn}})$. This essential difference makes us to distinguish $\widehat{R}_1(g)$ and $\widehat{R}_2(g)$ for the sake of theoretical comparisons, but there will be only one implementation of the two estimators in practice.

[4] ℓ is classification-calibrated if and only if there is a convex, invertible and nondecreasing transformation ψ_ℓ with $\psi_\ell(0) = 0$, such that $\psi_\ell(I(g) - I^*) \le R(g) - R^*$ (Bartlett et al., 2006).

4. Main Results

Suppose we are given a *function class* $\mathcal{G}$. When learning is involved, let $g^* = \arg\min_{g\in\mathcal{G}} R(g)$ be the optimal decision function in $\mathcal{G}$, and $\hat{g}_i = \arg\min_{g\in\mathcal{G}} \widehat{R}_i(g)$ be arbitrary *globally optimal solution*[5] in $\mathcal{G}$ by following *empirical risk minimization* of $\widehat{R}_i(g)$ for $i = 1, \ldots, 6$.

In this paper, we derive and compare *risk bounds* for learners $\hat{g}_1, \ldots, \hat{g}_6$ by assuming $\mathcal{G}$ satisfies the following condition: There is a positive constant $C_{\mathcal{G}}$ such that

$$\mathfrak{R}_{n,q}(\mathcal{G}) = C_{\mathcal{G}}/\sqrt{n} \tag{11}$$

for any marginal density $q(x)$, where

$$\mathfrak{R}_{n,q}(\mathcal{G}) = \mathbb{E}_{\mathcal{X}\sim q^n}\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{x_i\in\mathcal{X}} \sigma_i g(x_i)\right]$$

is the *Rademacher complexity* of $\mathcal{G}$ for samplings of size n from $q(x)$ (i.e., $\mathcal{X} = \{x_1, \ldots, x_n\}$ with each x_i following $q(x)$ and $\sigma = \{\sigma_1, \ldots, \sigma_n\}$ with each σ_i as a Rademacher variable). Our comparisons also cover a special case, sets of hyperplanes with *bounded normals and feature maps*:

$$\begin{aligned}\mathcal{G} = \{g(x) = \langle w, \phi(x)\rangle_{\mathcal{H}} \mid \\ \|w\|_{\mathcal{H}} \le C_w, \|\phi(x)\|_{\mathcal{H}} \le C_\phi\},\end{aligned} \tag{12}$$

where $\mathcal{H}$ is a *Hilbert space* with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, $w \in \mathcal{H}$ is the normal vector, $\phi : \mathbb{R}^d \mapsto \mathcal{H}$ is a feature map, and C_w and C_ϕ are positive constants.

Based on our theoretical analyses (which will be presented in Section 5), two main results can be stated as follows.

Theorem 1 (Main result 1)**.** *Let $\mathcal{G}$ be defined in* (12)*, otherwise assume* (11)*. As n_{u} increases, the estimation error bounds of $\hat{g}_3$ and $\hat{g}_5$ both get improved. Moreover, assume also n_{u} increases faster in order than n_+ and n_-, i.e.,*

$$\lim_{n_+\to\infty, n_{\mathrm{u}}\to\infty} n_+/n_{\mathrm{u}} = 0, \quad \lim_{n_-\to\infty, n_{\mathrm{u}}\to\infty} n_-/n_{\mathrm{u}} = 0.$$

Then, depending on π and τ_{pn}, the estimation error bound of either $\hat{g}_3$ or $\hat{g}_5$ will ultimately become superior to that of $\hat{g}_1$ as $n_{\mathrm{u}} \to \infty$, provided that $n_+/n_- \neq \pi^2/(1-\pi)^2$.[6]

Theorem 2 (Main result 2)**.** *Let $\mathcal{G}$ be defined in* (12)*, otherwise assume* (11)*. Assume also* (3)*,* (7)*,* (10)*, as well as $\tau_{\mathrm{pu}} \le 2\pi/(1+2\pi)$ for $\hat{g}_4$ and $\tau_{\mathrm{nu}} \ge 1/(3-2\pi)$ for $\hat{g}_6$. Then the estimation error bounds of $\hat{g}_4$ and $\hat{g}_6$ are optimal at $\tau^*_{\mathrm{pu}} = 2\pi/(1+2\pi)$ and $\tau^*_{\mathrm{nu}} = 1/(3-2\pi)$ respectively.*[7] *Furthermore, if $\pi/(2-\pi) \le \tau_{\mathrm{pn}} \le 2\pi/(1+\pi)$, the estimation error bound of neither $\hat{g}_4$ nor $\hat{g}_6$ could improve on that of $\hat{g}_2$.*

[5] Loss functions satisfying (4) must be non-convex (du Plessis et al., 2014).

[6] It is included that $n_+ \to \infty$ and $n_- \to \infty$ independently of $n_{\mathrm{u}} \to \infty$ under the assumption that n_{u} increases faster in order.

[7] They cannot hold at the same time unless $\tau_{\mathrm{pn}} = \pi$, since n_{u} is the only free variable for fixed π, n_+ and n_-.

Theorem 1 guarantees the limiting behavior of $\{\hat{g}_3, \hat{g}_5\}$ as a whole. The implication is we can *almost always* improve the double-expectation learner $\hat{g}_1$ in PN learning by that in PU or NU learning, if n_{u} increases faster in order than n_+ and n_-. In PU and NU learning, it should not be difficult to collect significantly more unlabeled data so that $\mathcal{O}(n_{\mathrm{u}})$ differs from $\mathcal{O}(n_+)$ and $\mathcal{O}(n_-)$ slightly, such as $\mathcal{O}(n_{\mathrm{u}}) = \mathcal{O}(n_+ \ln n_+) + \mathcal{O}(n_- \ln n_-)$.

Although Theorem 2 does not cover all situations in which $\hat{g}_2$ outperforms $\{\hat{g}_4, \hat{g}_6\}$ as a whole, it is non-trivial to analytically yet significantly broaden or narrow the interval of τ_{pn}. In Theorem 2, the upper bound of τ_{pu} and the lower bound of τ_{nu} guarantee that n_{u} is larger than some thresholds, and the interval of τ_{pn} ensures that it is not far away from π. For example, $\pi = 1/2$ leads to the widest interval $1/3 \le \tau_{\mathrm{pn}} \le 2/3$, $\pi = 1/4$ gives $1/7 \le \tau_{\mathrm{pn}} \le 2/5$, and it has two limit cases that $|\tau_{\mathrm{pn}} - \pi| \to 0$ as $\pi \to 0$ or $\pi \to 1$. The implication is as long as τ_{pn} is close enough to π, we can *never* improve the single-expectation learner $\hat{g}_2$ in PN learning by that in PU or NU learning.

At a glance, $\hat{g}_3$ and $\hat{g}_5$ are more recommended than $\hat{g}_4$ and $\hat{g}_6$, because not only $\{\hat{g}_3, \hat{g}_5\}$ improves $\hat{g}_1$ for almost all π and τ_{pn} if given infinite unlabeled data, but also $\hat{g}_3$ and $\hat{g}_5$ improve themselves given more unlabeled data. Note that, however, this is biased: If τ_{pn} is fairly close to π, $\hat{g}_2$ is also superior to $\hat{g}_1$; actually $\hat{g}_3$ and $\hat{g}_5$ are always inferior to $\hat{g}_4$ and $\hat{g}_6$ with τ^*_{pu} and τ^*_{nu} (cf. Comparisons 1, 2 and 5).

Recall that the only difference between double-expectation and single-expectation learners is the data-generating processes, and in practice they share the implementation when the training data for $\hat{g}_2$, $\hat{g}_4$ and $\hat{g}_6$ are constructed from the training data for $\hat{g}_1$, $\hat{g}_3$ and $\hat{g}_5$ (cf. Footnote 3). Therefore, we advocate learning from positive (P), negative (N), and unlabeled (U) data in the following way:

- If U is much cheaper than P and N, collect U as many as possible and run a test (cf. the proof of Theorem 1)
  - if $\pi^2/(1-\pi)^2 < \tau_{\mathrm{pn}}/(1-\tau_{\mathrm{pn}})$, try PU learning;
  - else try NU learning;
- Else use PN learning.

5. Technical Details: Theoretical Comparisons based on Risk Bounds

In this section, we establish and analyze risk bounds of $\hat{g}_i$.

5.1. PN learning

Let L_ℓ be the *Lipschitz constant* of ℓ in its first parameter. To begin with, we derive the learning guarantee of $\hat{g}_2$ as it is technically the most basic bound.

Theorem 3. *Define*

$$\eta_{\mathrm{pn}} = \max\{\pi/\tau_{\mathrm{pn}}, (1-\pi)/(1-\tau_{\mathrm{pn}})\}.$$

Assume (3) *so that* $|\eta_{\mathrm{pn}}| < \infty$. *For any* $\delta > 0$, *with probability at least* $1-\delta$,[8]

$$R(\hat{g}_2) - R(g^*) \leq 4\eta_{\mathrm{pn}} L_\ell \mathfrak{R}_{n_{\mathrm{pn}},q_{\mathrm{pn}}}(\mathcal{G}) + \eta_{\mathrm{pn}}\sqrt{\frac{2\ln(2/\delta)}{n_{\mathrm{pn}}}}, \tag{13}$$

where $\mathfrak{R}_{n_{\mathrm{pn}},q_{\mathrm{pn}}}(\mathcal{G})$ *is the Rademacher complexity of* $\mathcal{G}$ *for samplings of size* n_{pn} *from* $q_{\mathrm{pn}}(x)$.

Moreover if ℓ *is classification-calibrated, there exists non-decreasing* φ *with* $\varphi(0) = 0$, *such that with probability at least* $1-\delta$,

$$I(\hat{g}_2) - I^* \leq \varphi\Big(R(g^*) - R^* + 4\eta_{\mathrm{pn}} L_\ell \mathfrak{R}_{n_{\mathrm{pn}},q_{\mathrm{pn}}}(\mathcal{G}) + \eta_{\mathrm{pn}}\sqrt{\frac{2\ln(2/\delta)}{n_{\mathrm{pn}}}}\Big). \tag{14}$$

The proof of Theorem 3 relies on a fundamental lemma of the *uniform deviation* of $\widehat{R}_2(g)$:

Lemma 4. *For any* $\delta > 0$, *with probability at least* $1-\delta$,

$$\sup_{g\in\mathcal{G}} |\widehat{R}_2(g) - R(g)| \leq 2\eta_{\mathrm{pn}} L_\ell \mathfrak{R}_{n_{\mathrm{pn}},q_{\mathrm{pn}}}(\mathcal{G}) + \eta_{\mathrm{pn}}\sqrt{\frac{\ln(2/\delta)}{2n_{\mathrm{pn}}}}.$$

In Lemma 4, $R(g)$ is w.r.t. ℓ and $p(x,y)$, $\widehat{R}_2(g)$ is w.r.t. ℓ_{pn} on samplings from $q_{\mathrm{pn}}(x,z)$, and $\mathfrak{R}_{n_{\mathrm{pn}},q_{\mathrm{pn}}}(\mathcal{G})$ is also w.r.t. $q_{\mathrm{pn}}(x)$. Hence, $\mathfrak{R}_{n_{\mathrm{pn}},q_{\mathrm{pn}}}(\mathcal{G})$ can easily be estimated based on $\mathcal{X}_+ \cup \mathcal{X}_-$ for some well-known $\mathcal{G}$, e.g., hyperplanes in the original data space or in kernel-induced feature spaces (Schölkopf & Smola, 2001; Mohri et al., 2012).

In Theorem 3, $R(\hat{g}_2)$ is w.r.t. ℓ and $p(x,y)$, even though $\hat{g}_2$ is learned based on ℓ_{pn} and $q_{\mathrm{pn}}(x,z)$. However, we cannot get rid of $q_{\mathrm{pn}}(x,z)$ due to the complexity term. The right-hand side (RHS) of (13) is a bound of the *estimation error* of $\hat{g}_2$, and it is small if $\mathcal{G}$ is small. RHS of (14) is a bound of the *excess risk* of $\hat{g}_2$ w.r.t. the zero-one loss, and it further involves the *approximation error* of $\mathcal{G}$ which is small if $\mathcal{G}$ is large. When $\mathcal{G}$ is fixed and it is either defined in (12) or satisfies (11), we have $\mathfrak{R}_{n_{\mathrm{pn}},q_{\mathrm{pn}}}(\mathcal{G}) = \mathcal{O}(1/\sqrt{n_{\mathrm{pn}}})$ and

$$R(\hat{g}_2) - R(g^*) \to 0, \quad I(\hat{g}_2) - I^* \to \varphi(R(g^*) - R^*)$$

in $\mathcal{O}_p(1/\sqrt{n_{\mathrm{pn}}})$. When $\mathcal{G}$ grows with n_{pn} properly,

$$R(\hat{g}_2) - R(g^*) \to 0, \quad I(\hat{g}_2) - I^* \to 0$$

in an order slower than $\mathcal{O}_p(1/\sqrt{n_{\mathrm{pn}}})$, which means $\hat{g}_2$ approaches the Bayes classifier as long as ℓ is classification-calibrated.

Next we derive the learning guarantee of $\hat{g}_1$.[9]

Theorem 5. *For any* $\delta > 0$, *with probability at least* $1-\delta$,

$$R(\hat{g}_1) - R(g^*) \leq 4\pi L_\ell \mathfrak{R}_{n_+,p_+}(\mathcal{G}) + 4(1-\pi) L_\ell \mathfrak{R}_{n_-,p_-}(\mathcal{G}) + \pi\sqrt{\frac{2\ln(4/\delta)}{n_+}} + (1-\pi)\sqrt{\frac{2\ln(4/\delta)}{n_-}}, \tag{15}$$

where $\mathfrak{R}_{n_+,p_+}(\mathcal{G})$ *and* $\mathfrak{R}_{n_-,p_-}(\mathcal{G})$ *are Rademacher complexities of* $\mathcal{G}$ *for samplings of size* n_+ *from* $p_+(x)$ *and of size* n_- *from* $p_-(x)$.

In Theorem 5, no artificial density is involved, as $R(\hat{g}_1)$ is w.r.t. $p(x,y)$ even though $\hat{g}_1$ is learned based on $p_+(x)$ and $p_-(x)$, and the complexity terms are also w.r.t. $p_+(x)$ and $p_-(x)$. Under (11) or (12), the estimation error of $\hat{g}_1$ goes to zero in $\mathcal{O}_p(1/\sqrt{n_+} + 1/\sqrt{n_-})$.[10]

It seems that the estimation error of $\hat{g}_2$ converges faster in order than that of $\hat{g}_1$. Nevertheless, this is incorrect: Theorem 3 needs (3) while Theorem 5 is *assumption-free*. As a consequence, not only $\widehat{R}_1(g)$ is more general than $\widehat{R}_2(g)$, but also the learning guarantee of $\hat{g}_1$ is *more general than* that of $\hat{g}_2$. If (3) is true, their bounds share the same order and a careful comparison of Eqs. (15) and (13) is helpful.

Comparison 1. Let $\mathcal{G}$ be defined in (12) or assume that it satisfies (11). Assume also that δ is small enough such that $\sqrt{\ln(4/\delta)} \approx \sqrt{\ln(2/\delta)}$. Then, the ratio of RHS of (15) to RHS of (13) can be approximately written as

$$\alpha_{1,2} = (\pi/\sqrt{\tau_{\mathrm{pn}}} + (1-\pi)/\sqrt{1-\tau_{\mathrm{pn}}})/\eta_{\mathrm{pn}},$$

and approximately $\hat{g}_1$ is better if $\alpha_{1,2} < 1$ while $\hat{g}_2$ is better if $\alpha_{1,2} > 1$ in terms of their estimation error bounds.[11]

For instance, in *ordinary supervised learning* where $\tau_{\mathrm{pn}} = \pi$ is claimed, we have

$$\alpha_{1,2} = \sqrt{\pi} + \sqrt{1-\pi} \geq \sqrt{2},$$

and $\hat{g}_2$ is clearly better. If $\pi = 1/\sqrt{2}$ and $\tau_{\mathrm{pn}} = 0.5$, then $\alpha_{1,2} = 1$, and thus $\hat{g}_1$ and $\hat{g}_2$ tie. Furthermore, if $\pi = 0.5$

[8] Here, the probability is over repeated sampling of $\mathcal{S}_{\mathrm{pn}}$ from $q_{\mathrm{pn}}(x,z)$ for learning $\hat{g}_2$. Explanations are similar in other theorems, but slightly different in two lemmas where repeated sampling of $\mathcal{S}_{\mathrm{pn}}$ is for evaluating $\widehat{R}_2(g)$ rather than learning $\hat{g}_2$.

[9] From now on, we will only present estimation error bounds, because excess risk bounds are immediate corollaries of the corresponding estimation error bounds. Also, necessary uniform deviation bounds will be incorporated into proofs of theorems.

[10] Note that π is always viewed as a constant, but τ_{pn} can be a constant with (3) or a function without (3). This means when we claim $\tau_{\mathrm{pn}} = \pi$, we add a constraint on τ_{pn} rather than make π a function. Hence, the order of $\pi/\sqrt{n_+}$ will always be regarded as $\mathcal{O}(1/\sqrt{n_+})$, even when $\tau_{\mathrm{pn}} = \pi$ is specified.

[11] In all of our comparisons, the quality of any learner is characterized fully by its estimation error bound. With some abuse of terminology, we will omit the dependence on this characteristic.

and $\tau_{\mathrm{pn}} = 0.1$, then $\alpha_{1,2} = 2\sqrt{2}/3\sqrt{5} \approx 0.42$, and $\hat{g}_1$ is absolutely better. To sum up, π and τ_{pn} jointly determine which of $\hat{g}_1$ and $\hat{g}_2$ is the winner. □

5.2. PU learning

The learning guarantee of $\hat{g}_3$ is similar to that of $\hat{g}_1$, and so is $\hat{g}_4$ to $\hat{g}_2$. Thus proofs of Theorems 6 and 7 are omitted.

Theorem 6. *For any $\delta > 0$, with probability at least $1-\delta$,*

$$R(\hat{g}_3) - R(g^*) \leq 8\pi L_\ell \mathfrak{R}_{n_+,p_+}(\mathcal{G}) + 4L_\ell \mathfrak{R}_{n_\mathrm{u},p}(\mathcal{G}) + 2\pi\sqrt{\frac{2\ln(4/\delta)}{n_+}} + \sqrt{\frac{2\ln(4/\delta)}{n_\mathrm{u}}}, \tag{16}$$

where $\mathfrak{R}_{n_\mathrm{u},p}(\mathcal{G})$ is the standard Rademacher complexity of $\mathcal{G}$ (for samplings of size n_u from $p(x)$).

Theorem 7. *Define*

$$\eta_{\mathrm{pu}} = \max\{2\pi/\tau_{\mathrm{pu}}, 1/(1-\tau_{\mathrm{pu}})\}.$$

Assume (7) so that $|\eta_{\mathrm{pu}}| < \infty$. Then, for any $\delta > 0$, with probability at least $1-\delta$,

$$R(\hat{g}_4) - R(g^*) \leq 4\eta_{\mathrm{pu}} L_\ell \mathfrak{R}_{n_{\mathrm{pu}},q_{\mathrm{pu}}}(\mathcal{G}) + \eta_{\mathrm{pu}}\sqrt{\frac{2\ln(2/\delta)}{n_{\mathrm{pu}}}}, \tag{17}$$

where $\mathfrak{R}_{n_{\mathrm{pu}},q_{\mathrm{pu}}}(\mathcal{G})$ is the Rademacher complexity of $\mathcal{G}$ for samplings of size n_{pu} from $q_{\mathrm{pu}}(x)$.

Compared with Theorems 5 and 3, Theorems 6 and 7 have one more assumption, namely, the symmetric condition (4). Actually Theorems 3 and 5 can be applied to any bounded Lipschitz-continuous ℓ, but (4) is indispensable for Theorems 6 and 7 in order to make sure of the unbiasedness of $\widehat{R}_3(g)$ and $\widehat{R}_4(g)$. For appropriate $\mathcal{G}$ (under (11) or (12)), the estimation error of $\hat{g}_3$ is $\mathcal{O}_p(1/\sqrt{n_+} + 1/\sqrt{n_\mathrm{u}})$ and the estimation error of $\hat{g}_4$ is $\mathcal{O}_p(1/\sqrt{n_{\mathrm{pu}}})$. The risk bound of $\hat{g}_3$ is indeed *more general than* that of $\hat{g}_4$, and if (7) is true they share the same order. Notice that this is PU learning, and these risk bounds are achieved by empirically learning $\hat{g}_3$ and $\hat{g}_4$ without any labeled negative data.

Comparison 2. To compare the estimation error bounds of $\hat{g}_3$ and $\hat{g}_4$, we assume $\sqrt{\ln(4/\delta)} \approx \sqrt{\ln(2/\delta)}$ as before. Moreover, it is natural in PU learning to assume $n_\mathrm{u} \gg n_+$ so that $\eta_{\mathrm{pu}} = 2\pi/\tau_{\mathrm{pu}}$. Given (11) the RHS-ratio of (16) to (17) can be approximately written as

$$\alpha_{3,4} = \sqrt{\tau_{\mathrm{pu}}} + \tau_{\mathrm{pu}}/(2\pi\sqrt{1-\tau_{\mathrm{pu}}}).$$

For example, $\alpha_{3,4} = (1 + 1/(6\pi))/\sqrt{10}$ when $\tau_{\mathrm{pu}} = 0.1$. Thus, $\hat{g}_3$ should be preferred as long as $\pi > 1/(6(\sqrt{10}-1)) \approx 0.077$. In fact, the smaller τ_{pu} is, the wider range of π makes $\alpha_{3,4} < 1$ and the more favorable $\hat{g}_3$ is (compared with $\hat{g}_4$). By contrast, if we take $\tau^*_{\mathrm{pu}} = 2\pi/(1+2\pi)$, then $\alpha_{3,4} = (1+\sqrt{2\pi})/\sqrt{1+2\pi} > 1$ for any $\pi > 0$. Thus, $\hat{g}_3$ is always worse than $\hat{g}_4$ with τ^*_{pu} that is optimal for $\hat{g}_4$. □

Comparison 3. PU and PN learning can also be theoretically compared in this way. Consider $\hat{g}_3$ and $\hat{g}_1$ under (11). Instead of investigating the RHS-ratio of (16) to (15), we subtract $(4\pi L_\ell \mathfrak{R}_{n_+,p_+}(\mathcal{G}) + \pi\sqrt{2\ln(4/\delta)/n_+})$ from their RHSs and examine the ratio

$$\alpha_{3,1} = (\pi + \sqrt{\tau_{\mathrm{pu}}/(1-\tau_{\mathrm{pu}})})/((1-\pi)\sqrt{\tau_{\mathrm{pn}}/(1-\tau_{\mathrm{pn}})})$$

to see whether $\alpha_{3,1} < 1$ or not.

In *ordinary semi-supervised learning* where $\tau_{\mathrm{pn}} = \pi$,[12] if we denote by $a = \sqrt{\tau_{\mathrm{pu}}/(1-\tau_{\mathrm{pu}})}$, then

$$\alpha_{3,1} = (\pi + a)/\sqrt{\pi(1-\pi)} \geq 2\sqrt{a(a+1)}$$

where the equality holds at $\pi = a/(2a+1)$. Therefore, if $\tau_{\mathrm{pu}} = 0.01$ and $\pi = 0.083$, $\alpha_{3,1} \approx 0.665$ and $\hat{g}_3$ is clearly better. If $\tau_{\mathrm{pu}} = 0.04$ and $\pi = 0.115$ or 0.18, $\alpha_{3,1} \approx 1$ and $\hat{g}_3$ and $\hat{g}_1$ tie. Whenever $\tau_{\mathrm{pu}} > 0.05$, by no means could $\hat{g}_3$ outperform $\hat{g}_1$. In fact, we can easily make $\hat{g}_3$ worse even for tiny τ_{pu} by pushing π far away from $a/(2a+1)$.

When τ_{pn} is independent of π, it is easier to analyze $\alpha_{3,1}$ since it is a monotonic function, increasing with π and τ_{pu} and decreasing with τ_{pn}. Consequently, the smaller π and τ_{pu} are or the larger τ_{pn} is, the preferable $\hat{g}_3$ is (to $\hat{g}_1$). □

Comparison 4. For a comparison of $\hat{g}_4$ and $\hat{g}_2$ we assume (3), (7), (11) and $\eta_{\mathrm{pu}} = 2\pi/\tau_{\mathrm{pu}}$. The RHS-ratio of (17) to (13) is given by

$$\alpha_{4,2} = \begin{cases} 2\sqrt{\tau_{\mathrm{pn}}/\tau_{\mathrm{pu}}}, & \tau_{\mathrm{pn}} \leq \pi, \\ 2\pi(1-\tau_{\mathrm{pn}})/((1-\pi)\sqrt{\tau_{\mathrm{pn}}\tau_{\mathrm{pu}}}), & \tau_{\mathrm{pn}} > \pi, \end{cases}$$

where the cases come from $\eta_{\mathrm{pn}} = \pi/\tau_{\mathrm{pn}}$ or $(1-\pi)/(1-\tau_{\mathrm{pn}})$.

When $\tau_{\mathrm{pn}} = \pi$ is specified, $\hat{g}_4$ would be better if and only if $4 < \tau_{\mathrm{pu}}/\pi \leq 2/(1+2\pi)$, where the bounds come from $\alpha_{4,2} < 1$ and $\eta_{\mathrm{pu}} = 2\pi/\tau_{\mathrm{pu}}$. Nonetheless, this is impossible since $1/(1+2\pi) < 1$ for any $\pi > 0$, which means $\hat{g}_4$ could never outperform $\hat{g}_2$ when $\tau_{\mathrm{pn}} = \pi$.

When τ_{pn} is independent of π, we discuss two cases separately. If $\tau_{\mathrm{pn}} \leq \pi$, $\alpha_{4,2}$ is monotonically increasing with τ_{pn} and decreasing with τ_{pu} regardless of π. The condition for $\hat{g}_4$ to win is $4\tau_{\mathrm{pn}} < \tau_{\mathrm{pu}} \leq 2\pi/(1+2\pi)$. For example $\hat{g}_4$ is better if $\pi = 0.5$, $\tau_{\mathrm{pn}} = 0.1$ and $0.4 < \tau_{\mathrm{pu}} \leq 0.5$. If $\tau_{\mathrm{pn}} > \pi$, $\alpha_{4,2}$ is monotonically increasing with π and decreasing with τ_{pn} and τ_{pu}, but this time the condition cannot be simplified. For example if $\pi = 0.5$ and $\tau_{\mathrm{pn}} = 0.9$, then $0.045 < \tau_{\mathrm{pu}} \leq 0.5$ ensures $\hat{g}_4$ being the winner. □

[12] It refers to ordinary supervised learning with unlabeled data, such that $\tau_{\mathrm{pn}} = \pi$ is inherited. Semi-supervised learning usually, but not always, assumes $\mathcal{S}_{\mathrm{pn}}$ is repeatedly sampled from $p(x,y)$ without any dataset shift (Chapelle et al., 2006).

5.3. NU learning

As the opposite of PU learning, $\hat{g}_6$ and $\hat{g}_5$ again have similar learning guarantees.

Theorem 8. *For any $\delta > 0$, with probability at least $1-\delta$,*

$$\begin{aligned} R(\hat{g}_5) - R(g^*) \leq \\ 8(1-\pi)L_\ell \mathfrak{R}_{n_-,p_-}(\mathcal{G}) + 4L_\ell \mathfrak{R}_{n_\mathrm{u},p}(\mathcal{G}) \\ + 2(1-\pi)\sqrt{\tfrac{2\ln(4/\delta)}{n_-}} + \sqrt{\tfrac{2\ln(4/\delta)}{n_\mathrm{u}}}. \end{aligned} \tag{18}$$

Theorem 9. *Define*

$$\eta_\mathrm{nu} = \max\{1/\tau_\mathrm{nu}, 2(1-\pi)/(1-\tau_\mathrm{nu})\}.$$

Assume (10) so that $|\eta_\mathrm{nu}| < \infty$. Then, for any $\delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned} R(\hat{g}_6) - R(g^*) \leq \\ 4\eta_\mathrm{nu} L_\ell \mathfrak{R}_{n_\mathrm{nu},q_\mathrm{nu}}(\mathcal{G}) + \eta_\mathrm{nu}\sqrt{\tfrac{2\ln(2/\delta)}{n_\mathrm{nu}}}, \end{aligned} \tag{19}$$

where $\mathfrak{R}_{n_\mathrm{nu},q_\mathrm{nu}}(\mathcal{G})$ is the Rademacher complexity of $\mathcal{G}$ for samplings of size n_nu from $q_\mathrm{nu}(x)$.

The symmetric condition (4) is still indispensable for Theorems 8 and 9 to make sure of the unbiasedness of $\widehat{R}_5(g)$ and $\widehat{R}_6(g)$. For appropriate $\mathcal{G}$, the estimation errors of $\hat{g}_5$ and $\hat{g}_6$ are of orders $\mathcal{O}_p(1/\sqrt{n_-} + 1/\sqrt{n_\mathrm{u}})$ and $\mathcal{O}_p(1/\sqrt{n_\mathrm{nu}})$. The risk bound of $\hat{g}_5$ is *more general than* that of $\hat{g}_6$, and if (10) is true they share the same order. These risk bounds are achieved by learning $\hat{g}_5$ and $\hat{g}_6$ empirically without any labeled positive data.

Comparison 5. The relationship of $\hat{g}_5$ and $\hat{g}_6$ is analogous to that of $\hat{g}_3$ and $\hat{g}_4$. In NU learning, assume $n_\mathrm{u} \gg n_-$ so that $\eta_\mathrm{nu} = 2(1-\pi)/(1-\tau_\mathrm{nu})$. Given (11) the RHS-ratio of (18) to (19) can be approximately written as

$$\alpha_{5,6} = \sqrt{1-\tau_\mathrm{nu}} + (1-\tau_\mathrm{nu})/(2(1-\pi)\sqrt{\tau_\mathrm{nu}}).$$

For example, when $\tau_\mathrm{nu} = 0.9$, $\hat{g}_5$ should be preferred as long as $\pi < 0.923$. The larger τ_nu is, the wider range of π makes $\alpha_{5,6} < 1$. By contrast, if we take $\tau^*_\mathrm{nu} = 1/(3-2\pi)$, then $\alpha_{5,6} = (1+\sqrt{2-2\pi})/\sqrt{3-2\pi} > 1$ for any $\pi < 1$, and $\hat{g}_5$ is always worse than $\hat{g}_6$ with τ^*_nu. □

Comparison 6. Next consider $\hat{g}_5$ and $\hat{g}_1$ under (11). Subtracting $(4(1-\pi)L_\ell \mathfrak{R}_{n_-,p_-}(\mathcal{G}) + (1-\pi)\sqrt{2\ln(4/\delta)/n_-})$ from RHSs of (18) and (15), we examine the ratio

$$\alpha_{5,1} = (1-\pi+\sqrt{(1-\tau_\mathrm{nu})/\tau_\mathrm{nu}})/(\pi\sqrt{(1-\tau_\mathrm{pn})/\tau_\mathrm{pn}})$$

to see whether $\alpha_{5,1} < 1$ or not.

When $\tau_\mathrm{pn} = \pi$ is specified, let $b = \sqrt{(1-\tau_\mathrm{nu})/\tau_\mathrm{nu}}$ then

$$\alpha_{5,1} = (1-\pi+b)/\sqrt{\pi(1-\pi)} \geq 2\sqrt{b(b+1)}$$

where the equality holds at $\pi = (b+1)/(2b+1)$, and $\hat{g}_5$ wins if $\tau_\mathrm{nu} > 0.96$ and π is very close to $(b+1)/(2b+1)$.

When τ_pn is independent of π, $\alpha_{5,1}$ becomes a monotonic function, increasing with τ_pn while decreasing with π and τ_nu. As a result, the larger π and τ_nu are or the smaller τ_pn is, the preferable $\hat{g}_5$ is. □

Comparison 7. For a comparison of $\hat{g}_6$ and $\hat{g}_2$ we assume (3), (10), (11) and $\eta_\mathrm{nu} = 2(1-\pi)/(1-\tau_\mathrm{nu})$. The RHS-ratio of (19) to (13) is given by

$$\alpha_{6,2} = \begin{cases} 2\sqrt{(1-\tau_\mathrm{pn})/(1-\tau_\mathrm{nu})}, & \tau_\mathrm{pn} \geq \pi, \\ 2(1-\pi)\tau_\mathrm{pn}/(\pi\sqrt{(1-\tau_\mathrm{pn})(1-\tau_\mathrm{nu})}), & \tau_\mathrm{pn} < \pi. \end{cases}$$

Similarly, $\hat{g}_6$ could never outperform $\hat{g}_2$ when $\tau_\mathrm{pn} = \pi$ is claimed. When τ_pn is independent of π, if $\tau_\mathrm{pn} \geq \pi$, $\alpha_{6,2}$ is monotonically increasing with τ_nu and decreasing with τ_pn regardless of π; for instance $\hat{g}_6$ wins if $\pi = 0.5$, $\tau_\mathrm{pn} = 0.9$ and $0.5 \leq \tau_\mathrm{nu} < 0.6$. If $\tau_\mathrm{pn} < \pi$, $\alpha_{6,2}$ becomes increasing with τ_pn and τ_nu and decreasing with π; for instance $\hat{g}_6$ is better if $\pi = 0.5$, $\tau_\mathrm{pn} = 0.1$ and $0.5 \leq \tau_\mathrm{nu} < 0.955$. □

5.4. On the function class

While our theoretical comparisons heavily rely on Eq. (11), we can get rid of this assumption by realizing $\mathcal{G}$ as the special case in Eq. (12). Given any sample $\mathcal{X} = \{x_1, \ldots, x_n\}$ from arbitrary marginal density $q(x)$, we have

$$\begin{aligned} \widehat{\mathfrak{R}}_\mathcal{X}(\mathcal{G}) &= \mathbb{E}_{\sigma_i}\left[\sup_{g\in\mathcal{G}} \tfrac{1}{n}\textstyle\sum_{i=1}^n \sigma_i g(x_i)\right] \\ &\leq C_w C_\phi/\sqrt{n}, \end{aligned} \tag{20}$$

where $\widehat{\mathfrak{R}}_\mathcal{X}(\mathcal{G})$ is the *empirical Rademacher complexity* of $\mathcal{G}$ conditioned on $\mathcal{X}$ (Mohri et al., 2012). Then, all uniform deviation bounds using the Rademacher complexity in our theorems can be replaced with those using the corresponding empirical Rademacher complexity, and our theoretical comparisons are all valid for $\mathcal{G}$ defined in (12) without the assumption (11). For example, the following uniform deviation bound is the alternative to Lemma 4.

Lemma 10. *Let $\mathcal{G}$ be defined in (12). For any $\delta > 0$, with probability at least $1-\delta$,*

$$\begin{aligned} \sup_{g\in\mathcal{G}} |\widehat{R}_2(g) - R(g)| \leq 2\eta_\mathrm{pn} L_\ell C_w C_\phi/\sqrt{n_\mathrm{pn}} \\ + \eta_\mathrm{pn}(\sqrt{\ln(4/\delta)} + 2\sqrt{\ln(2/\delta)})/\sqrt{2n_\mathrm{pn}}. \end{aligned}$$

6. Proofs

Proof of Theorem 3 (with Lemma 4). For simplicity let $n = n_\mathrm{pn}$. ℓ is bounded by 0 and 1 due to (4), and then ℓ_pn is bounded by 0 and η_pn. The change of $\widehat{R}_2(g)$ will be no more than η_pn/n, if a single (x, z) in $\mathcal{S}_\mathrm{pn}$ is replaced with (x', z'). Thus *McDiarmid's inequality* (McDiarmid, 1989)

implies for any fixed g, with probability at least $1-\delta$,

$$|\widehat{R}_2(g) - R(g)| \leq \eta_{\mathrm{pn}} \sqrt{\tfrac{\ln(2/\delta)}{2n}}.$$

According to the basic *uniform deviation bound* using the Rademacher complexity (Mohri et al., 2012), with probability at least $1-\delta$,

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\widehat{R}_2(g) - R(g)| \leq \\ &\quad 2\mathfrak{R}_{n,q_{\mathrm{pn}}}(\ell_{\mathrm{pn}} \circ \mathcal{G}) + \eta_{\mathrm{pn}} \sqrt{\tfrac{\ln(2/\delta)}{2n}},\end{aligned} \tag{21}$$

where

$$\begin{aligned}&\mathfrak{R}_{n,q_{\mathrm{pn}}}(\ell_{\mathrm{pn}} \circ \mathcal{G}) = \\ &\quad \mathbb{E}_{\mathcal{S}_{\mathrm{pn}}} \mathbb{E}_{\sigma} \left[\sup_{g\in\mathcal{G}} \tfrac{1}{n} \textstyle\sum_{(x_i,z_i)\in\mathcal{S}_{\mathrm{pn}}} \sigma_i \ell_{\mathrm{pn}}(g(x_i), z_i)\right]\end{aligned}$$

is the Rademacher complexity of $\ell_{\mathrm{pn}} \circ \mathcal{G}$ for samplings of size n from $q_{\mathrm{pn}}(x,z)$. Since $\ell_{\mathrm{pn}}(t,z)$ is $(\eta_{\mathrm{pn}} L_\ell)$-Lipschitz in t for every z when $\ell(t,y)$ is L_ℓ-Lipschitz in t for every y, we have $\mathfrak{R}_{n,q_{\mathrm{pn}}}(\ell_{\mathrm{pn}} \circ \mathcal{G}) \leq \eta_{\mathrm{pn}} L_\ell \mathfrak{R}_{n,q_{\mathrm{pn}}}(\mathcal{G})$ by *Talagrand's contraction lemma* (Ledoux & Talagrand, 1991), which proves Lemma 4. □

Subsequently,

$$\begin{aligned}R(\hat{g}_2) - R(g^*) &= \left(\widehat{R}_2(\hat{g}_2) - \widehat{R}_2(g^*)\right) \\ &\quad + \left(R(\hat{g}_2) - \widehat{R}_2(\hat{g}_2)\right) + \left(\widehat{R}_2(g^*) - R(g^*)\right) \\ &\leq 0 + 2\sup_{g\in\mathcal{G}} |\widehat{R}_2(g) - R(g)|\end{aligned}$$

since $\hat{g}_2 = \arg\min_{g\in\mathcal{G}} \widehat{R}_2(g)$. Applying Lemma 4 proves (13). For classification-calibrated ℓ, *Theorem 1* in Bartlett et al. (2006) implies a convex, invertible, and non-decreasing ψ_ℓ with $\psi_\ell(0) = 0$ which satisfies $\psi_\ell(I(\hat{g}_2) - I^*) \leq R(\hat{g}_2) - R^*$. Hence, let $\varphi = \psi_\ell^{-1}$, we have

$$\begin{aligned}I(\hat{g}_2) - I^* &\leq \varphi(R(\hat{g}_2) - R^*) \\ &= \varphi(R(g^*) - R^* + R(\hat{g}_2) - R(g^*)),\end{aligned}$$

and then (14) is a corollary of (13). □

Proof of Theorem 5. The proof is along the line of that of Theorem 3. Similarly to $R(g)$, $\widehat{R}_1(g)$ could also be decomposed into $\widehat{R}_1(g) = \pi\widehat{R}_+(g) + (1-\pi)\widehat{R}_-(g)$ where

$$\begin{aligned}\widehat{R}_+(g) &= \tfrac{1}{n_+} \textstyle\sum_{x\in\mathcal{X}_+} \ell(g(x), +1), \\ \widehat{R}_-(g) &= \tfrac{1}{n_-} \textstyle\sum_{x\in\mathcal{X}_-} \ell(g(x), -1).\end{aligned}$$

Due to the sub-additivity of the supremum, it holds that

$$\begin{aligned}&\sup_{g\in\mathcal{G}} |\widehat{R}_1(g) - R(g)| \leq \pi \sup_{g\in\mathcal{G}} |\widehat{R}_+(g) - R_+(g)| \\ &\quad + (1-\pi) \sup_{g\in\mathcal{G}} |\widehat{R}_-(g) - R_-(g)|.\end{aligned}$$

As a result, in order to prove Theorem 5 it suffices to show that with probability at least $1-\delta/2$, the uniform deviation bounds below hold separately:

$$\begin{aligned}\sup_{g\in\mathcal{G}} |\widehat{R}_+(g) - R_+(g)| &\leq 2L_\ell \mathfrak{R}_{n_+,p_+}(\mathcal{G}) + \sqrt{\tfrac{\ln(4/\delta)}{2n_+}}, \\ \sup_{g\in\mathcal{G}} |\widehat{R}_-(g) - R_-(g)| &\leq 2L_\ell \mathfrak{R}_{n_-,p_-}(\mathcal{G}) + \sqrt{\tfrac{\ln(4/\delta)}{2n_-}}.\end{aligned}$$

The change of $\widehat{R}_+(g)$ will be no more than $1/n_+$ if replacing a single x in $\mathcal{X}_+$ with x'. Therefore, with probability at least $1-\delta/2$,

$$\sup_{g\in\mathcal{G}} |\widehat{R}_+(g) - R(g)| \leq 2\mathfrak{R}_{n_+,p_+}(\ell \circ \mathcal{G}) + \sqrt{\tfrac{\ln(4/\delta)}{2n_+}},$$

and $\mathfrak{R}_{n_+,p_+}(\ell \circ \mathcal{G}) \leq L_\ell \mathfrak{R}_{n_+,p_+}(\mathcal{G})$ from Talagrand's contraction lemma. In the same way the part of $\widehat{R}_-(g)$ can be proved. □

Proof of Lemma 10. We continue from (21) in the proof of Lemma 4 but now adjust its probability to $1-\delta/2$. Applying McDiarmid's inequality to $\widehat{\mathfrak{R}}_{\mathcal{S}}(\ell_{\mathrm{pn}} \circ \mathcal{G})$ implies that with probability at least $1-\delta/2$,

$$\begin{aligned}\mathfrak{R}_{n,q_{\mathrm{pn}}}(\ell_{\mathrm{pn}} \circ \mathcal{G}) &\leq \widehat{\mathfrak{R}}_{\mathcal{S}}(\ell_{\mathrm{pn}} \circ \mathcal{G}) + \eta_{\mathrm{pn}} \sqrt{\tfrac{\ln(2/\delta)}{2n}} \\ &\leq \eta_{\mathrm{pn}} L_\ell \widehat{\mathfrak{R}}_{\mathcal{X}_+\cup\mathcal{X}_-}(\mathcal{G}) + \eta_{\mathrm{pn}} \sqrt{\tfrac{\ln(2/\delta)}{2n}} \\ &\leq \eta_{\mathrm{pn}} L_\ell C_w C_\phi / \sqrt{n_{\mathrm{pn}}} + \eta_{\mathrm{pn}} \sqrt{\tfrac{\ln(2/\delta)}{2n}}\end{aligned}$$

by Talagrand's contraction lemma and (20). □

Proof of Theorem 1. The first half is obvious due to the monotonicity of $\alpha_{3,1}$ and $\alpha_{5,1}$: $\alpha_{3,1}$ increases with τ_{pu} and $\alpha_{5,1}$ decreases with τ_{nu}, while τ_{pu} decreases with n_{u} and τ_{nu} increases with n_{u}, monotonically. This means both of $\alpha_{3,1}$ and $\alpha_{5,1}$ are monotonically decreasing with n_{u}. Refer to Comparisons 3 and 6 for details.

For the second half, let $\alpha^*_{3,1}$ and $\alpha^*_{5,1}$ be the limits of $\alpha_{3,1}$ and $\alpha_{5,1}$. Under our assumptions,

$$\begin{aligned}\alpha^*_{3,1} &= \pi\sqrt{1-\tau_{\mathrm{pn}}}/((1-\pi)\sqrt{\tau_{\mathrm{pn}}}), \\ \alpha^*_{5,1} &= (1-\pi)\sqrt{\tau_{\mathrm{pn}}}/(\pi\sqrt{1-\tau_{\mathrm{pn}}}).\end{aligned}$$

As a consequence, $\alpha^*_{3,1}\alpha^*_{5,1} = 1$ and either $\alpha^*_{3,1}$ or $\alpha^*_{5,1}$ is smaller than 1 depending on π and τ_{pn}. The only exception is $\alpha^*_{3,1} = \alpha^*_{5,1} = 1$, that is, $n_+/n_- = \pi^2/(1-\pi)^2$. □

Proof of Theorem 2. It is easy to show the first half due to the monotonicity of $\alpha_{4,2}$ and $\alpha_{6,2}$, where the bounds of τ_{pu} and τ_{nu} secure the correct values of η_{pu} and η_{nu}. See Comparisons 4 and 7 for details.

For the second half, there are four cases. By the definition of $\alpha_{4,2}$, if $\pi/(2+4\pi) < \tau_{\mathrm{pn}} \leq \pi$, $\alpha_{4,2} > 1$; if $\pi < \tau_{\mathrm{pn}} \leq$

$2\pi/(1+\pi)$, $\alpha_{4,2} \geq \sqrt{(1+2\pi)/(1+\pi)} > 1$. In addition, by definition, if $\pi \leq \tau_{\mathrm{pn}} < (5-3\pi)/(6-4\pi)$, $\alpha_{6,2} > 1$; if $\pi/(2-\pi) \leq \tau_{\mathrm{pn}} < \pi$, $\alpha_{6,2} \geq \sqrt{(3-2\pi)/(2-\pi)} > 1$. Finally, the proof is finished by noting that $\pi/(2+4\pi) \leq \pi/(2-\pi)$ and $2\pi/(1+\pi) \leq (5-3\pi)/(6-4\pi)$. □

7. Conclusions

We have derived and compared risk bounds of six learners in three problem settings, where each learner employs two datasets among P, N and U. In practice if they are sampled from genuine marginal densities separately, there will only be three implementations, and PU or NU learning will improve on PN learning almost certainly given infinite U.

All six learners in this paper are empirical risk minimizers to certain cost-sensitive formulas, which made our theoretical comparisons not too complicated to handle. There are novel methods that linearly combine PN learning with PU or NU learning, so that learners use all of P, N, and U data (Anonymous Authors). We are going to study these learners as our future research.

References

Anonymous Authors. Beyond the low-density separation principle: A novel approach to semi-supervised learning. *Submitted to ICML 2016*.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Blanchard, G., Lee, G., and Scott, C. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. In Bousquet, O., von Luxburg, U., and Rätsch, G. (eds.), *Advanced Lectures on Machine Learning*, pp. 169–207. Springer, 2004.

Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. MIT Press, 2006.

du Plessis, M. C. and Sugiyama, M. Semi-supervised learning of class balance under class-prior change by distribution matching. In *ICML*, 2012.

du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *NIPS*, 2014.

du Plessis, M. C., Niu, G., and Sugiyama, M. Class-prior estimation for learning from positive and unlabeled data. In *ACML*, 2015a.

du Plessis, M. C., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *ICML*, 2015b.

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

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

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

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

McDiarmid, C. On the method of bounded differences. In Siemons, J. (ed.), *Surveys in Combinatorics*, pp. 148–188. Cambridge University Press, 1989.

Meir, R. and Zhang, T. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. Learning with noisy labels. In *NIPS*, 2013.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Schölkopf, B. and Smola, A. *Learning with Kernels*. MIT Press, 2001.

Scott, C. and Blanchard, G. Novelty detection: Unlabeled data definitely help. In *AISTATS*, 2009.

Vapnik, V. N. *Statistical Learning Theory*. John Wiley & Sons, 1998.