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

Gang Niu
Marthins Christoffel du Plessis
Tomoya Sakai
Masashi Sugiyama
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.

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).

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 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 $n_u \to \infty$ faster than $n_+ + n_-$ (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,

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

2See 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 $X_+$ of size $n_+$ from $p_+(x)$ as well as $X_-$ of size $n_-$ from $p_-(x)$ are observed. Let $n_{pn} = n_+ + n_-$, $\tau_{pn} = n_+/n_{pn}$ be the pseudo PN class-prior, $Z = \pm 1$ be an indicator variable of PN data, and

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

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

$$
S_{pn} = \{(x, +1) \mid x \in X_+\} \cup \{(x, -1) \mid x \in X_\}\}
$$

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

In addition, suppose we have $X_u$ of size $n_u$ from $p(x)$. Let $n_{pu} = n_+ + n_u$, $\tau_{pu} = n_+/n_{pu}$ and

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

for PU data. Now

$$
S_{pu} = \{(x, +1) \mid x \in X_+\} \cup \{(x, -1) \mid x \in X_u\}
$$

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

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

for NU data, and

$$
S_{nu} = \{(x, +1) \mid x \in X_u\} \cup \{(x, -1) \mid x \in X_-\}
$$

can be regarded as a sampling of size $n_{nu}$ from $q_{nu}(x, z)$. 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 $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be an arbitrary real-valued decision function for binary classification and $\ell : \mathbb{R} \times \{\pm 1\} \rightarrow \mathbb{R}$ be a bounded Lipschitz-continuous loss function. Denote by

$$
R(g) = E_{X,Y}[\ell(g(X), Y)] = \pi R_+(g) + (1 - \pi) R_-(g)
$$

where $E_{X,Y}[\cdot] = E_{X \sim p_+}[\cdot]$. Then the risk of $g$ w.r.t. $\ell$ under $p(x, y)$ is given by

$$
R(g) = E_{X,Y}[\ell(g(X), Y)] = \pi R_+(g) + (1 - \pi) R_-(g)
$$

where

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

In other words, the risk of $g$ w.r.t. $\ell$ under $p(x, y)$ is equivalent to that of $g$ w.r.t. $\ell_{pn}$ under $q_{pn}(x, z)$, 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)

$$
\hat{R}_1(g) = \frac{1}{n_{pu}} \sum_{x \in X_+} \ell(g(x), +1)
$$

$$
+ \frac{1 - \pi}{n_{pu}} \sum_{x \in X_-} \ell(g(x), -1),
$$

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

$$
\hat{R}_2(g) = \frac{1}{n_{pn}} \sum_{(x, z) \in S_{pn}} \ell_{pn}(g(x, z), z),
$$

whose convergence rate is $O_p(1/\sqrt{n_{pn}})$. 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 $\lim_{n^+ \rightarrow \infty} \hat{R}_2(g)$ exists only if

$$
0 < \lim_{n^+ \rightarrow \infty} \lim_{n_- \rightarrow \infty} n^+/n_- < \infty,
$$

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

\(^3\)Note that $\hat{R}_1(g)$ and $\hat{R}_2(g)$ coincide, if $X_+$ and $X_-$ are re-
3.2. PU learning

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

\[
\ell(t, +1) + \ell(t, -1) = 1,
\]

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

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

where the loss function \( \ell_{pu} \) is defined by

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

Hence, the risk of \( g \) w.r.t. \( \ell \) under \( p(x, y) \) is also equivalent to that w.r.t. \( \ell_{pu} \) under \( q_{pu}(x, z) \) minus \( \pi \).

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

\[
\hat{R}_3(g) = -\pi + \frac{2\pi}{n_+} \sum_{x \in X_+} \ell(g(x), +1)
+ \frac{1}{n_u} \sum_{x \in X_u} \ell(g(x), -1),
\]

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

\[
\hat{R}_4(g) = -\pi + \frac{1}{n_+} \sum_{(x, z) \in S_{pu}} \ell_{pu}(g(x), z),
\]

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

\[
0 < \lim_{n_+ \to \infty, n_u \to \infty} n_+/n_u < \infty,
\]

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

3.3. NU learning

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

\[
R_{u, +}(g) = \pi R_{+}(g) + (1 - \pi)(1 - R_{-}(g)),
\]

and thus

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

where the loss function \( \ell_{nu} \) is defined by

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

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

\[
\hat{R}_5(g) = -(1 - \pi) + \frac{1}{n_u} \sum_{x \in X_u} \ell(g(x), +1)
+ \frac{2(1 - \pi)}{n_{nu}} \sum_{x \in X_u} \ell(g(x), -1),
\]

\[
\hat{R}_6(g) = -(1 - \pi) + \frac{1}{n_{nu}} \sum_{(x, z) \in S_{nu}} \ell_{nu}(g(x), z).
\]

Again, \( \lim_{n_+, n_u \to \infty} \hat{R}_6(g) \) exists only if

\[
0 < \lim_{n_+, n_u \to \infty} n_+/n_u < \infty,
\]

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

3.4. On the loss function

Let \( \ell_{o1}(t, y) = (1 - \text{sign}(ty))/2 \) be the zero-one loss and

\[
I(g) = \mathbb{E}_{(X, Y)}[\ell_{o1}(g(X, Y))]
\]

the risk of \( g \) w.r.t. \( \ell_{o1} \) under \( p(x, y) \). Let \( R^* = \inf_g R(g) \) and \( I^* = \inf_g I(g) \) denote the Bayes risks w.r.t \( \ell \) and \( \ell_{o1} \) where the infimum is over all measurable functions. If \( \ell \) is classification-calibrated\(^4\), we are able to control the excess risk w.r.t. \( \ell \) by that w.r.t. \( \ell_{o1} \) (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 \( \ell_{o1} \) in PU learning:

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

but it is still an open question whether \( \ell_{sr} \) is classification-calibrated. Here we show indeed it is and therefore a safe surrogate loss for \( \ell_{o1} \). Denote by \( \pi_+(x) = p(y = +1 \mid x) \) and \( \pi_-(x) = p(y = -1 \mid x) \), then the conditional risk is

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

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

---

\(^4\) is classification-calibrated if and only if there is a convex, invertible and nondecreasing transformation \( \psi_\ell \) with \( \psi_\ell(0) = 0 \), such that \( \psi_\ell(I(g) - I^*) \leq R(g) - R^* \) (Bartlett et al., 2006).
4. Main Results

Suppose we are given a function class $G$. When learning is involved, let $g^* = \arg \min_{g \in G} R(g)$ be the optimal decision function in $G$, and $\hat{g}_i = \arg \min_{g \in G} \hat{R}_i(g)$ be arbitrary globally optimal solution in $G$ by following empirical risk minimization of $\hat{R}_i(g)$ 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 $G$ satisfies the following condition: There is a positive constant $C_G$ such that

$$\mathcal{R}_{n,q}(G) = C_G / \sqrt{n}$$

(11)

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

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

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

$$G = \{g(x) = \langle w, \phi(x) \rangle_H \mid \|w\|_H \leq C_w, \|\phi(x)\|_H \leq C_\phi \},$$

(12)

where $H$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$, $w \in H$ is the normal vector, $\phi : \mathbb{R}^d \rightarrow H$ is a feature map, and $C_w$ and $C_\phi$ 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 $G$ 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.,

$$\lim_{n_+ \rightarrow \infty, n_u \rightarrow \infty} n_+ / n_u = 0, \lim_{n_- \rightarrow \infty, n_u \rightarrow \infty} n_- / n_u = 0.$$

Then, depending on $\pi$ and $\tau_{pu}$, the estimation error bound of either $\hat{g}_3$ or $\hat{g}_5$ will ultimately become superior to that of $\hat{g}_1$ as $n_u \rightarrow \infty$, provided that $n_+ / n_- \neq \pi^2 / (1 - \pi)^2$.6

Theorem 2 (Main result 2). Let $G$ be defined in (12), otherwise assume (11). Assume also (3), (7), (10), as well as $\tau_{pu} \leq 2\pi / (1 + 2\pi)$ for $\hat{g}_3$ and $\tau_{pu} \geq 1/(3 - 2\pi)$ for $\hat{g}_6$. Then the estimation error bounds of $\hat{g}_3$ and $\hat{g}_6$ are optimal at $\tau_{pu}^* = 2\pi / (1 + 2\pi)$ and $\tau_{pu}^* = 1/(3 - 2\pi)$ respectively.7 Furthermore, if $\pi / (2 - \pi) \leq \tau_{pu} \leq 2\pi / (1 + \pi)$, the estimation error bound of neither $\hat{g}_3$ nor $\hat{g}_6$ could improve on that of $\hat{g}_2$.

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 $O(n_u)$ differs from $O(n_+)$ and $O(n_-)$ slightly, such as $O(n_u) = O(n_+ \ln n_+) + O(n_- \ln n_-)$.

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 $\tau_{pu}$. In Theorem 2, the upper bound of $\tau_{pu}$ and the lower bound of $\tau_{nu}$ guarantee that $n_u$ is larger than some thresholds, and the interval of $\tau_{pu}$ ensures that it is not far away from $\pi$. For example, $\pi = 1/2$ leads to the widest interval $1/3 \leq \tau_{pu} \leq 2/3, \pi = 1/4$ gives $1/7 \leq \tau_{pu} \leq 2/5$, and it has two limit cases that $|\tau_{pu} - \pi| \rightarrow 0$ as $\pi \rightarrow 0$ or $\pi \rightarrow 1$. The implication is as long as $\tau_{pu}$ is close enough to $\pi$, 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 $\pi$ and $\tau_{pu}$ 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 $\tau_{pu}$ is fairly close to $\pi$, $\hat{g}_2$ 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 $\tau_{pu}$ and $\tau_{pu}^*$ (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 $\pi^2/(1 - \pi)^2 < \tau_{pu}/(1 - \tau_{pu})$, 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 $\hat{g}_1$.

5.1. PN learning

Let $L_\ell$ be the Lipschitz constant of $\ell$ 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
\[ \eta_{pn} = \max\{\pi/\tau_{pn}, (1 - \pi)/(1 - \tau_{pn})\}. \]
Assume (3) so that \( \eta_{pn} < \infty \). For any \( \delta > 0 \), with probability at least \( 1 - \delta \),
\[ R(\hat{g}_2) - R(\hat{g}^*) \leq 4\eta_{pn}L_\ell \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} + \eta_{pn} \sqrt{\frac{2\ln(2/\delta)}{n_{pn}}}, \] (13)
where \( \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} \) is the Rademacher complexity of \( \mathcal{G} \) for samplings of size \( n_{pn} \) from \( p_+(x) \).

Moreover if \( \ell \) is classification-calibrated, there exists non-decreasing \( \varphi \) with \( \varphi(0) = 0 \), such that with probability at least \( 1 - \delta \),
\[ I(\hat{g}_2) - I^* \leq \varphi(R(\hat{g}^*) - R^*) + 4\eta_{pn}L_\ell \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} + \eta_{pn} \sqrt{\frac{2\ln(2/\delta)}{n_{pn}}}. \] (14)

The proof of Theorem 3 relies on a fundamental lemma of the uniform deviation of \( \hat{R}_2(\mathcal{G}) \):

Lemma 4. For any \( \delta > 0 \), with probability at least \( 1 - \delta \),
\[ \sup_{g \in \mathcal{G}} |\hat{R}_2(g) - R(g)| \leq 2\eta_{pn}L_\ell \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} + \eta_{pn} \sqrt{\frac{\ln(2/\delta)}{2\eta_{pn}}}. \]

In Lemma 4, \( R(g) \) is w.r.t. \( \ell \) and \( p(x, y) \), \( \hat{R}_2(g) \) is w.r.t. \( \ell_{pn} \) on samplings from \( q_{pn}(x, z) \), and \( \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} \) is also w.r.t. \( q_{pn}(x) \). Hence, \( \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} \) 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. \( \ell \) and \( p(x, y) \), even though \( \hat{g}_2 \) is learned based on \( \ell_{pn} \) and \( q_{pn}(x, z) \). However, we cannot get rid of \( \eta_{pn} \) 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 \( \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} = O(1/\sqrt{n_{pn}}) \) and
\[ R(\hat{g}_2) - R(\hat{g}^*) \to 0, \quad I(\hat{g}_2) - I^* \to \varphi(R(\hat{g}^*) - R^*) \]
in \( O_p(1/\sqrt{n_{pn}}) \). When \( \mathcal{G} \) grows with \( n_{pn} \) properly,
\[ R(\hat{g}_2) - R(\hat{g}^*) \to 0, \quad I(\hat{g}_2) - I^* \to 0 \]
in an order slower than \( O_p(1/\sqrt{n_{pn}}) \), which means \( \hat{g}_2 \) approaches the Bayes classifier as long as \( \ell \) is classification-calibrated.

Next we derive the learning guarantee of \( \hat{g}_1 \).

Theorem 5. For any \( \delta > 0 \), with probability at least \( 1 - \delta \),
\[ R(\hat{g}_1) - R(\hat{g}^*) \leq 4\pi L_\ell \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} + 4(1 - \pi) L_\ell \mathcal{R}_{\eta_{pn}, p_-(\mathcal{G})} + \pi \sqrt{\frac{2\ln(4/\delta)}{n_{pn}}} + (1 - \pi) \sqrt{\frac{2\ln(4/\delta)}{n_{pn}}}, \] (15)
where \( \mathcal{R}_{\eta_{pn}, p_+(\mathcal{G})} \) and \( \mathcal{R}_{\eta_{pn}, p_-(\mathcal{G})} \) 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 \( O_p(1/\sqrt{n_+} + 1/\sqrt{n_-}) \).

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 \( \hat{R}_1(\mathcal{g}) \) is more general than \( \hat{R}_2(\mathcal{g}) \), 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 \( \delta \) 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
\[ \alpha_{1,2} = (\pi/\sqrt{\tau_{pn} + (1 - \pi)/\sqrt{1 - \tau_{pn}}})/\eta_{pn}, \]
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.

For instance, in ordinary supervised learning where \( \tau_{pn} = \pi \) 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 \( \tau_{pn} = 0.5 \), then \( \alpha_{1,2} = 1 \), and thus \( \hat{g}_1 \) and \( \hat{g}_2 \) tie. Furthermore, if \( \pi = 0.5 \)

---

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.

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

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 $\tau_{pn} = 0.1$, then $\alpha_{1,2} = 2\sqrt{2/3}\sqrt{5} \approx 0.42$, and $\hat{g}_1$ is absolutely better. To sum up, $\pi$ and $\tau_{pn}$ 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 $1 - \delta$,

$$ R(\hat{g}_3) - R(g^*) \leq 8\pi L_\ell R_{\mathcal{R}_{n\alpha,p}}(G) + 4L_\ell R_{\mathcal{R}_{n\alpha,p}}(G) + 2\pi \sqrt{\frac{2\ln(4/\delta)}{n_+}} + \sqrt{\frac{2\ln(4/\delta)}{n_+}}, $$

(16)

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

**Theorem 7.** Define

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

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

$$ R(\hat{g}_3) - R(g^*) \leq 4\eta_{pu} L_\ell R_{\mathcal{R}_{n\eta_{pu},p_{\mathcal{G}}}}(G) + \eta_{pu} \sqrt{\frac{2\ln(2/\delta)}{n_+}}, $$

(17)

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

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 $\ell$, but (4) is indispensable for Theorems 6 and 7 in order to make sure of the unbiasedness of $\mathbb{R}_\alpha(g)$ and $\mathbb{R}_\alpha(g)$. For appropriate $G$ (under (11) or (12)), the estimation error of $\hat{g}_3$ is $O_{\mathbb{P}}(1/\sqrt{n_+} + 1/\sqrt{n_\alpha})$ and the estimation error of $\hat{g}_3$ is $O_{\mathbb{P}}(1/\sqrt{n_{\eta_{pu}}})$. The risk bound of $\hat{g}_3$ is indeed more general than that of $\hat{g}_1$, 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 $n_\alpha \gg n_+$ so that $\eta_{pu} = 2\pi/\tau_{pu}$. Given (11) the RHS-ratio of (16) to (17) can be approximately written as

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

For example, $\alpha_{3,4} = (1 + 1/(6\pi))/\sqrt{10}$ when $\tau_{pu} = 0.1$. Thus, $\hat{g}_4$ should be preferred as long as $\pi > 1/(6(\sqrt{10} - 1)) \approx 0.077$. In fact, the smaller $\tau_{pu}$ is, the wider range of $\pi$ makes $\alpha_{3,4} < 1$ and the more favorable $\hat{g}_3$ is (compared with $\hat{g}_4$). By contrast, if we take $\tau_{pu} = 2\pi/(1 + 2\pi)$, 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 $\tau_{pu}$ 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 $\hat{g}_4$ under (11). Instead of investigating the RHS-ratio of (16) to (15), we subtract $(4\pi L_\ell \mathcal{R}_{n\alpha,p}(G) + \sqrt{2\ln(4/\delta)/n_+})$ from their RHSs and examine the ratio

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

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

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

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

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

When $\tau_{pn}$ is independent of $\pi$, it is easier to analyze $\alpha_{3,1}$ since it is a monotonic function, increasing with $\pi$ and $\tau_{pu}$ and decreasing with $\tau_{pn}$. Consequently, the smaller $\pi$ and $\tau_{pn}$ are or the larger $\tau_{pu}$ 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 $\eta_{pu} = 2\pi/\tau_{pu}$. The RHS-ratio of (17) to (13) is given by

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

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

When $\tau_{pn} = \pi$ is specified, $\hat{g}_4$ would be better if and only if $4 < \tau_{pu}/\pi \leq 2/(1 + 2\pi)$, where the bounds come from $\alpha_{4,2} < 1$ and $\eta_{pu} = 2\pi/\tau_{pu}$. 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 $\tau_{pn} = \pi$.

When $\tau_{pn}$ is independent of $\pi$, we discuss two cases separately. If $\tau_{pn} \leq \pi$, $\alpha_{4,2}$ is monotonically increasing with $\tau_{pu}$ and decreasing with $\tau_{pu}$ regardless of $\pi$. The condition for $\hat{g}_2$ to win is $4\tau_{pu} < \tau_{pu} \leq 2\pi/(1 + 2\pi)$. For example, $\hat{g}_4$ is better if $\pi = 0.5$, $\tau_{pu} = 0.1$ and $0.4 < \tau_{pu} \leq 0.5$. If $\tau_{pu} > \pi$, $\alpha_{4,2}$ is monotonically increasing with $\pi$ and decreasing with $\tau_{pu}$ and $\tau_{pu}$, but this time the condition cannot be simplified. For example if $\pi = 0.5$ and $\tau_{pu} = 0.9$, then $0.045 < \tau_{pu} < 0.5$ ensures $\hat{g}_4$ being the winner.

\(^{21}\)It refers to ordinary supervised learning with unlabeled data, such that $\tau_{pu} = \pi$ is inherited. Semi-supervised learning usually, but not always, assumes $S_{pu}$ 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 \( 1 - \delta \),
\[
R(\hat{g}_5) - R(g^*) \leq 8(1 - \pi)L_{\delta}R_{\nu, p}(\mathcal{G}) + 4L_{\delta}R_{\mu, p}(\mathcal{G}) + 2(1 - \pi)\sqrt{\frac{2\ln(4/\delta)}{\nu_m}} + \sqrt{\frac{2\ln(4/\delta)}{\nu_m}}.
\]

**Theorem 9.** Define
\[
\eta_{nu} = \max\{1/\tau_{nu}, 2(1 - \pi)/(1 - \tau_{nu})\}.
\]

Assume (10) so that \( |\eta_{nu}| < \infty \). Then, for any \( \delta > 0 \), with probability at least \( 1 - \delta \),
\[
R(\hat{g}_6) - R(g^*) \leq 4\eta_{nu}L_{\delta}R_{\nu, q_m}(\mathcal{G}) + \eta_{nu}\sqrt{\frac{2\ln(2/\delta)}{\nu_m}},
\]
where \( R_{\nu, q_m}(\mathcal{G}) \) is the Rademacher complexity of \( \mathcal{G} \) for samplings of size \( \nu_m \) from \( q_m(x) \).

The symmetric condition (4) is still indispensable for Theorems 8 and 9 to make sure of the unbiasedness of \( \hat{R}_6(g) \) and \( \hat{R}_6(g) \). For appropriate \( \mathcal{G} \), the estimation errors of \( \hat{g}_5 \) and \( \hat{g}_6 \) are of orders \( O_p(1/\sqrt{\nu_m}) \) and \( O_p(1/\sqrt{\nu_m}) \). 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 \( \nu_m \gg \nu_\ldots \) so that \( \eta_{nu} = 2(1 - \pi)/(1 - \tau_{nu}) \). Given (11) the RHS-ratio of (18) to (19) can be approximately written as
\[
\alpha_{5,6} = \sqrt{1 - \tau_{nu}} + (1 - \tau_{nu})/(2(1 - \pi)\sqrt{\nu_m}).
\]

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

**Comparison 6.** Next consider \( \hat{g}_5 \) and \( \hat{g}_1 \) under (11). Subtracting \( 4(1 - \pi)L_{\delta}R_{\nu, p}(\mathcal{G}) + (1 - \pi)\sqrt{2\ln(4/\delta)}/\nu_\ldots \) from RHSs of (18) and (15), we examine the ratio
\[
\alpha_{5,1} = (1 - \pi + \sqrt{(1 - \tau_{nu})/\nu_m})/(\pi \sqrt{(1 - \tau_{pn})/\nu_m})
\]
to see whether \( \alpha_{5,1} < 1 \) or not.

When \( \tau_{pn} = \pi \) is specified, let \( b = \sqrt{(1 - \tau_{nu})/\nu_m} \) 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 \( \tau_{nu} > 0.96 \) and \( \pi \) is very close to \( (b + 1)/(2b + 1) \).

When \( \tau_{pn} \) is independent of \( \pi \), \( \alpha_{5,1} \) becomes a monotonic function, increasing with \( \tau_{pn} \) while decreasing with \( \pi \) and \( \tau_{nu} \). As a result, the larger \( \pi \) and \( \tau_{nu} \) are or the smaller \( \tau_{pn} \) 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 \( \eta_{nu} = 2(1 - \pi)/(1 - \tau_{nu}) \). The RHS-ratio of (19) to (13) is given by
\[
\alpha_{0,2} = \begin{cases} 
2\sqrt{(1 - \tau_{pn})/(1 - \tau_{nu})}, & \tau_{pn} \geq \pi, \\
2(1 - \pi)\tau_{pn}/(\pi\sqrt{(1 - \tau_{pn})(1 - \tau_{nu})}), & \tau_{pn} < \pi.
\end{cases}
\]

Similarly, \( \hat{g}_6 \) could never outperform \( \hat{g}_2 \) when \( \tau_{pn} = \pi \) is claimed. When \( \tau_{pn} \) is independent of \( \pi \), if \( \tau_{pn} \geq \pi, \alpha_{0,2} \) is monotonically increasing with \( \tau_{nu} \) and decreasing with \( \tau_{pn} \) regardless of \( \pi \); for instance \( \hat{g}_6 \) wins if \( \pi = 0.5, \tau_{pn} = 0.9 \) and \( 0.5 \leq \tau_{nu} < 0.6 \). If \( \tau_{pn} < \pi, \alpha_{0,2} \) becomes increasing with \( \tau_{pn} \) and \( \tau_{nu} \) and decreasing with \( \pi \); for instance \( \hat{g}_6 \) is better if \( \pi = 0.5, \tau_{pn} = 0.1 \) and \( 0.5 \leq \tau_{nu} < 0.955 \).

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
\[
\hat{\mathcal{R}}_{\mathcal{X}}(\mathcal{G}) = \mathbb{E}_{\sigma_i} \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i) \right] \leq C_wC\phi/\sqrt{\nu_m},
\]
where \( \hat{\mathcal{R}}_{\mathcal{X}}(\mathcal{G}) \) 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 \( 1 - \delta \),
\[
\sup_{g \in \mathcal{G}} |\hat{R}_2(g) - R(g)| \leq 2\eta_{pn}L_{\delta}C_wC\phi/\sqrt{\nu_m} + \eta_{pn}(\sqrt{\ln(4/\delta)} + 2\ln(2/\delta))/\sqrt{2\nu_m}.
\]

6. Proofs

**Proof of Theorem 3 (with Lemma 4).** For simplicity let \( n = \nu_{pn} \), \( \ell \) is bounded by 0 and 1 due to (4), and then \( \ell_{pn} \) is bounded by 0 and \( \eta_{pn} \). The change of \( \hat{R}_2(g) \) will be no more than \( \eta_{pn} \), if a single \( (x, z) \) in \( \mathcal{S}_{pn} \) is replaced with \( (x', z') \). Thus McDiarmid’s inequality (McDiarmid, 1989)
implies for any fixed \( g \), with probability at least \( 1 - \delta \),

\[
|\hat{R}_2(g) - R(g)| \leq \eta_{pn} \sqrt{\frac{\ln(2/\delta)}{2n}}.
\]

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

\[
\sup_{g \in \mathcal{F}} |\hat{R}_2(g) - R(g)| \leq 2\mathfrak{R}_{n,\eta_{pn}}(\ell_{pn} \circ \mathcal{G}) + \eta_{pn} \sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{21}
\]

where

\[
\mathfrak{R}_{n,\eta_{pn}}(\ell_{pn} \circ \mathcal{G}) = \mathbb{E}_{\mathcal{S}_{pn}} \mathbb{E}_\sigma \left[ \sup_{g \in \mathcal{F}} \frac{1}{n} \sum_{(x_i, z_i) \in \mathcal{S}_{pn}} \sigma_i \ell_{pn}(g(x_i), z_i) \right]
\]

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

Subsequently,

\[
R(\hat{g}_2) - R(g^*) = (\hat{R}_2(\hat{g}_2) - \hat{R}_2(g^*)) + (\hat{R}_2(g^*) - R(g^*)) \\
\leq 0 + 2 \sup_{g \in \mathcal{F}} |\hat{R}_2(g) - R(g)|
\]

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

\[
I(\hat{g}_2) - I^* \leq \varphi(R(\hat{g}_2) - R^*) \\
= \varphi(R(g^*) - R^* + R(\hat{g}_2) - R(g^*)),
\]

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 \( \hat{R}(g) \), \( \hat{R}_1(g) \) could also be decomposed into \( \hat{R}_1(g) = \pi \hat{R}_+(g) + (1 - \pi) \hat{R}_-(g) \) where

\[
\hat{R}_+(g) = \frac{1}{n_+} \sum_{x \in \mathcal{X}_+} \ell(g(x), +1), \\
\hat{R}_-(g) = \frac{1}{n_-} \sum_{x \in \mathcal{X}_-} \ell(g(x), -1).
\]

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

\[
\sup_{g \in \mathcal{F}} |\hat{R}_1(g) - R(g)| \leq \pi \sup_{g \in \mathcal{F}} |\hat{R}_+(g) - R_+(g)| + (1 - \pi) \sup_{g \in \mathcal{F}} |\hat{R}_-(g) - R_-(g)|.
\]

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:

\[
\sup_{g \in \mathcal{F}} |\hat{R}_+(g) - R_+(g)| \leq 2L_\pi \mathfrak{R}_{n_+,\pi,\eta_{pn}}(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2n}}, \\
\sup_{g \in \mathcal{F}} |\hat{R}_-(g) - R_-(g)| \leq 2L_\pi \mathfrak{R}_{n_-,\pi,\eta_{pn}}(\mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2n}}.
\]

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

\[
\sup_{g \in \mathcal{F}} |\hat{R}_+(g) - R(g)| \leq 2\mathfrak{R}_{n_+,\pi,\eta_{pn}}(\ell \circ \mathcal{G}) + \sqrt{\frac{\ln(4/\delta)}{2n}},
\]

and \( \mathfrak{R}_{n_+,\pi,\eta_{pn}}(\ell \circ \mathcal{G}) \) from Talagrand’s contraction lemma. In the same way the part of \( \hat{R}_-(g) \) 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 \( \mathfrak{R}_{\mathcal{S}}(\ell_{pn} \circ \mathcal{G}) \) implies that with probability at least \( 1 - \delta/2 \),

\[
\mathfrak{R}_{n,\eta_{pn}}(\ell_{pn} \circ \mathcal{G}) \leq \mathfrak{R}_{\mathcal{S}}(\ell_{pn} \circ \mathcal{G}) + \eta_{pn} \sqrt{\frac{\ln(2/\delta)}{2n}} \\
\leq \eta_{pn} \frac{L_\pi}{\delta} \mathfrak{R}_{\mathcal{X}_+,\mathcal{X}_-}(\mathcal{G}) + \eta_{pn} \sqrt{\frac{\ln(2/\delta)}{2n}} \\
\leq \eta_{pn} L_\pi C_{u} C_{\phi} / \sqrt{\mathcal{S}_{pn}} + \eta_{pn} \sqrt{\frac{\ln(2/\delta)}{2n}}
\]

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 \( \tau_{pn} \) and \( \alpha_{3,1} \) decreases with \( \tau_{nu} \) while \( \tau_{pn} \) decreases with \( n_+ \) and \( \tau_{nu} \) 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,

\[
\alpha_{3,1}^* = \pi \sqrt{1 - \tau_{pn}/(1 - \pi)\sqrt{\tau_{pn}}}, \\
\alpha_{5,1}^* = (1 - \pi)\sqrt{\tau_{pn}/(1 - \tau_{pn})}.
\]

As a consequence, \( \alpha_{3,1}^* \cdot \alpha_{5,1}^* = 1 \) and either \( \alpha_{3,1}^* \) or \( \alpha_{5,1}^* \) is smaller than 1 depending on \( \pi \) and \( \tau_{pn} \). 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 \( \tau_{pn} \) and \( \tau_{nu} \) secure the correct values of \( \eta_{pn} \) and \( \eta_{nu} \). See Comparisons 4 and 7 for details.

For the second half, there are four cases. By the definition of \( \alpha_{4,2} \), if \( \pi/(2 + 4\pi) < \tau_{pn} \leq \pi, \alpha_{4,2} > 1 \); if \( \pi < \tau_{pn} \leq \pi/(2 + 4\pi) \),
2π/(1 + π), α_{4,2} ≥ √[(1 + 2π)/(1 + π)] > 1. In addition, by definition, if π ≤ τm < (5 − 3π)/(6 − 4π), α_{6,2} > 1; if π/(2 − π) ≤ τm < π, α_{6,2} ≥ √[(3 − 2π)/(2 − π)] > 1. Finally, the proof is finished by noting that π/(2 + 4π) ≤ π/(2 − π) and 2π/(1 + π) ≤ (5 − 3π)/(6 − 4π).

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


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


