Journal of Machine Learning Research 24 (2023) 1-31 Submitted 1/22; Revised 1/23; Published 1/23

Risk Bounds for Positive-Unlabeled Learning Under the Selected At Random Assumption

Olivier Coudray olivier.coudray@universite-paris-saclay.fr
Stellantis, Centre d'Expertise Métier et Région, Poissy, 78300, France
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, Orsay, 91405, France

Christine Keribin christine.keribin@universite-paris-saclay.fr
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, Orsay, 91405, France

Pascal Massart pascal.massart@fondation-hadamard.fr
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, Orsay, 91405, France

Patrick Pamphile patrick.pamphile@universite-paris-saclay.fr
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, Orsay, 91405, France

Editor: Jean-Philippe Vert

Abstract

Positive-Unlabeled learning (PU learning) is a special case of semi-supervised binary classification where only a fraction of positive examples is labeled. The challenge is then to find the correct classifier despite this lack of information. Recently, new methodologies have been introduced to address the case where the probability of being labeled may depend on the covariates. In this paper, we are interested in establishing risk bounds for PU learning under this general assumption. In addition, we quantify the impact of label noise on PU learning compared to the standard classification setting. Finally, we provide a lower bound on the minimax risk proving that the upper bound is almost optimal.

Keywords: Statistical learning theory, Classification, Label noise, PU learning, Risk bounds.

1. Introduction

Classic binary classification is a supervised machine learning task in which, from training observations with given *classes* (*positive* or *negative*), one seeks to predict the class of new data. However, in many realistic situations, the observed *classes* can be noisy. A case in point is when the class assignment is subject to errors. In this paper, we are interested in a special case of label noise, occurring when a fraction of positive instances is labeled and none of the negative instances are. The *unlabeled* instances are either positive or negative: their class is unknown.

This can be seen as a semi-supervised classification setting because only a fraction of the observations is labeled. This semi-supervised classification task is called *Positive-Unlabeled Learning* (*PU learning*). PU learning aims to build classifiers that find the right class (positive or negative) of a new data point given a training dataset of positive and unlabeled observations.

PU learning is used in situations where it is difficult or costly to obtain or identify reliable negative instances. For example, in the diagnosis of a disease, given the incubation period of the disease, a patient with a negative test may still be carrying the disease (cf. Chen et al., 2020). PU learning approach could be helpful in *fatigue design of structures* in mechanics where testing can prove the presence of design flaws on a mechanical part but cannot prove its absence. In the automotive industry, fatigue tests are performed to determine if a part is critical: if a crack is observed before the end of the test, then the part is declared critical. However, if no crack is observed, it does not mean that the part is not critical. It may be possible to observe a crack by extending the test (cf. Coudray et al., 2021). Other applications of PU learning exist in *spam review detection* (cf. Li et al., 2014; Fusilier et al., 2015; He et al., 2020), *text classification* (cf. Liu et al., 2002, 2003), *gene-disease identification* (cf. Yang et al., 2012, 2014; Nikdelfaz and Jalili, 2018), and *anomaly detection* (cf. Ferretti et al., 2014; Luo et al., 2018; Jiang et al., 2018).

PU learning is, therefore, much more complex than learning from fully labeled data. The situation is asymmetric as one usually wants to understand the positive class in contrast with an unidentified negative class. The number of positive labeled examples is critical as it is the only reliable information.

Different methodologies have been developed to address PU learning. A first class of heuristic methods proceeds in two steps. The first step consists in identifying reliable negative instances among the unlabeled observations: various methods exist like *Spy* (cf. Liu et al., 2002) or *Rocchio* (cf. Li and Liu, 2003) methods. In the second step, a standard supervised or semi-supervised classification method is used to build the PU classifier from the positive labeled instances, reliable negative instances, and the remaining unlabeled ones. A typical choice is Support Vector Machine (SVM). Some methods repeat both steps iteratively until convergence. Bekker and Davis (2020) gave an exhaustive list of existing methods for both steps. Good empirical results support these methods, but theoretical guarantees are not discussed.

Another class of methodologies resorts to a modeling of the label noise and adapts existing supervised classification methods to the PU learning setting. Most PU learning methods in this category assume that the probability for a positive instance to be labeled is constant and thus independent from the covariates. This situation is called Selected Completely At Random (SCAR). However, in certain cases, the probability for a positive instance to be labeled is influenced by its covariates. For example, in the diagnosis of a disease, a carrier of the disease with symptoms is more likely to see a doctor and be diagnosed than a carrier who is asymptomatic. This situation with a selection bias is called Selected At Random (SAR). Under the SCAR assumption, since the noise for positive instances is constant, the probability for an instance to be labeled is then proportional to the probability for it to be positive: fully labeled classification and PU learning are then connected. Hence, some algorithms use this property to derive consistent classifiers: Blanchard et al. (2010) use Neyman-Pearson classification and Du Plessis et al. (2014) rewrite PU learning as a cost-sensitive binary classification that can be solved through empirical risk minimization. These approaches are supported by theoretical guarantees: consistency and risk bounds. Mordelet and Vert (2014) suggest a *bagging SVM* method to solve PU learning tasks under the SCAR assumption, which proves efficient empirically. As mentioned above, the SCAR assumption is unlikely to hold in many practical situations. Recently, several publications have addressed PU learning when the probability of being labeled is instance-dependent (Bekker et al., 2020; Gong et al., 2021).

From a theoretical point of view, risk bounds in the standard classification setting have been extensively studied in the literature. The convergence rate of the excess risk in classification is known to be less than a quantity proportional to $\sqrt{1/n}$ where n is the size of the training set (cf. Lugosi, 2002). In addition, this rate can be refined, reaching $1/n$ under margin assumptions (Massart and Nédélec, 2006). Finally, these rates are proved to be optimal in the minimax sense (cf. Lugosi, 2002; Massart and Nédélec, 2006).

Missing labels in PU learning can arise from different settings. In the *two-sample setting*, the positive and unlabeled instances are sampled separately and are therefore not identically distributed: it is a *case-control* situation. In the *one-sample setting*, all the instances are i.i.d. and some positive instances are labeled. In the past few years, several papers have studied excess risk upper bounds for PU learning classifiers in the case-control setting (cf. Du Plessis et al., 2015; Niu et al., 2016; Vogel et al., 2020). In particular, Du Plessis et al. (2014) showed a convergence rate in $\mathcal{O}\left(\sqrt{1/n_L} + \sqrt{1/n_U}\right)$, where n_L (n_U) denotes the number of labeled (unlabeled) instances. More recently, Bekker et al. (2020) and Gong et al. (2021) studied theoretical properties of PU learning under selection bias with specific assumptions: the former establishes an upper bound on an empirical risk minimizer under partial knowledge of the labeling mechanism, the latter focus on a parametric model.

In this work, we focus on PU learning in the one-sample setting. We establish new risk bounds specifically adapted to PU learning under selection bias, meaning that the label noise specific to PU learning is instance-dependent (Selected At Random assumption). Unlike Gong et al. (2021), we do not make parametric assumptions. Contrary to Bekker et al. (2020), who focused on the deviations between PU learning empirical risk and fully supervised empirical risk, we provide an upper bound on the excess risk. The novelty of this result also lies in its ability to quantify explicitly the impact of the label noise specific to PU learning (*propensity*). We show that fast convergence rates can be achieved under margin conditions similar to Massart and Nédélec (2006). Finally, we discuss the optimality of this result by identifying a lower bound on the minimax risk.

The paper is organized as follows. In Section 2, we define the standard binary classification setting and recall some existing risk bounds. In Section 3, we move to the PU learning setting, discuss the bias issue with labeled-unlabeled classification and introduce an unbiased empirical risk. In Section 4, we present the main results of this paper: a general upper bound on the excess risk for PU learning under instance-dependent label noise and a lower bound on the minimax risk. In Section 5, we conclude and discuss some future perspectives.

2. Standard Classification Setting

In this section, we introduce the standard classification setting and recall some risk bounds results. This will be the opportunity to introduce general notations used throughout the paper.

2.1 General Setting

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be independent couples of random variables in $\mathbb{R}^d \times \{0,1\}$ identically distributed according to some unknown probability distribution denoted $\mathbb{P}$. For each i, X_i is a *covariate* vector with marginal distribution $\mathbb{P}_X$ and Y_i is the *class*, either *negative* ($Y_i = 0$) or *positive* ($Y_i = 1$). Let $\alpha = \mathbb{P}(Y = 1)$ denote the class prior. Using $\mathbb{P}_0$ ($\mathbb{P}_1$) the conditional distribution of X given that the class is negative, $Y = 0$ (positive, $Y = 1$), we write the convenient decomposition:

$$\mathbb{P}_X = (1-\alpha)\mathbb{P}_0 + \alpha\mathbb{P}_1 \ . \tag{1}$$

In classification, the goal is to find a classifier, *i.e.* a binary function $g : \mathbb{R}^d \to \{0,1\}$, minimizing some risk function R. In this paper, R will denote the misclassification risk:

$$R(g) = \mathbb{P}(g(X) \neq Y) \ .$$

Given the regression function $\eta(x) = \mathbb{P}(Y = 1|X = x)$, the minimizer of misclassification risk is Bayes classifier g^* that depends explicitly on $\mathbb{P}$:

$$g^*(x) = \mathbb{1}_{\eta(x)\geq\frac{1}{2}}.$$

In order to assess how close a given classifier g is to the optimal one g^*, we are interested in the excess risk $\ell(g, g^*)$:

$$\ell(g, g^*) = R(g) - R(g^*) .$$

Since $\mathbb{P}$ is unknown, neither g^* nor the risk function R can be computed. We rely instead on the training sample $(X_1, Y_1), ..., (X_n, Y_n)$ to build a classifier $\widehat{g}$. Let $r(g, (X,Y)) = \mathbb{1}_{g(X)\neq Y}$ the misclassification error for one observation, the true risk R can be estimated by the empirical mean:

$$\widehat{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} r(g, (X_i, Y_i)) \ .$$

An empirical classifier $\widehat{g}$ is then identified as a minimizer of the empirical risk over a predefined class of classifiers $\mathcal{G}$.

$$\widehat{g} \in \underset{g\in\mathcal{G}}{\operatorname{Argmin}} \ \widehat{R}_n(g) .$$

This procedure is known as empirical risk minimization. Let $g^{\mathcal{G}}$ be the minimizer of the true risk R over $\mathcal{G}$. The excess risk of the classifier $\widehat{g}$ can be decomposed as follows:

$$\ell(\widehat{g}, g^*) = \left(R\left(g^{\mathcal{G}}\right) - R(g^*)\right) + \left(R(\widehat{g}) - R\left(g^{\mathcal{G}}\right)\right)$$

where the first term is the approximation error depending on $\mathcal{G}$, and the second one is the statistical error. Since we are only interested in assessing the statistical error, we assume that Bayes classifier g^* belongs to $\mathcal{G}$. Hence the first term vanishes. Note that $\ell(\widehat{g}, g^*)$ depends on $\mathbb{P}$ (through the risk R) and on the training sample $(X_1, Y_1), ..., (X_n, Y_n)$.

2.2 Risk Bounds in the Standard Classification Setting

In order to assess the convergence rate of the excess risk $\ell\left(\widehat{g}, g^*\right)$ in a non-asymptotic framework, we need an upper bound on $\mathbb{E}\left[\ell\left(\widehat{g}, g^*\right)\right]$. Note that the expectation is taken with respect to the distribution of the training sample $\mathbb{P}^{\otimes n}$. Moreover, the upper bound needs to be uniform over a set of distributions $\mathbb{P}$. We introduce $\mathcal{P}\left(\mathcal{G}\right)$ a set of probability distributions on $\mathbb{R}^d \times \{0,1\}$ such that g^* belongs to $\mathcal{G}$. In this case, Lugosi (2002) proved that for some absolute constant $C_1 > 0$:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{G})} \mathbb{E}\left[\ell\left(\widehat{g}, g^*\right)\right] \leq C_1\sqrt{\frac{V}{n}}\ , \tag{2}$$

where V is the *Vapnik-Chervonenkis dimension* of $\mathcal{G}$ (VC dimension, see Vapnik, 1999, Chapter 3). We recall that the VC dimension is the maximum integer V such that there exist V points $x_1, ..., x_V$ in $\mathbb{R}^d$ *shattered* by $\mathcal{G}$, namely classified in every way possible by elements of $\mathcal{G}$. In other words:

$$V = \sup_{v\in\mathbb{N}^*}\left\{v\ s.t.\ \exists\, x_1, ..., x_v \in \mathbb{R}^d,\ \left|\left\{\left(g(x_1), ..., g(x_v)\right),\ g\in\mathcal{G}\right\}\right| = 2^v\right\}\ .$$

The VC dimension V measures the complexity of class $\mathcal{G}$ and has to be finite for Equation 2 to be meaningful, which we assume for the rest of the paper.

The upper bound in Equation 2 remains true regardless of the form of the regression function η. Actually, η is closely linked to the difficulty of the classification task: when $\eta(x)$ is close to $1/2$, the observed class can be positive or negative with probability close to $1/2$, which makes the classification of x more difficult. Massart and Nédélec (2006) showed that when $\eta(x)$ is uniformly and symmetrically bounded away from $1/2$ by a *margin* $h > \sqrt{V/n}$, the upper bound on the risk excess can be improved. Let $\mathcal{P}\left(\mathcal{G}, h\right)$ denote the subset of probability distributions in $\mathcal{P}\left(\mathcal{G}\right)$ such that for every $x\in\mathbb{R}^d$, $|2\eta(x) - 1| \geq h$. Massart and Nédélec (2006) showed that there exists an absolute constant $C_2 > 0$ such that:

$$\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{G},h)} \mathbb{E}\left[\ell\left(\widehat{g}, g^*\right)\right] \leq C_2\frac{V}{nh}\left(1 + \log\left(\frac{nh^2}{V}\right)\right)\ . \tag{3}$$

Hence, as the margin h gets higher, the classification task gets easier, and the convergence rate can be improved up to $1/n$, letting aside the logarithm. However, when h is smaller than $\sqrt{1/n}$, Equation 2 remains better. Equation 3 provides fine control on the excess risk depending on the difficulty of the classification task, accounted through h.

A lower bound was obtained by Lugosi (2002), extended by Massart and Nédélec (2006), allowing to prove the optimality of the convergence rates. Note that the optimality of the refined bound Equation 3 is up to the logarithmic term.

3. PU Learning Context

In the standard classification setting, the classes $(Y_i)_{1\leq i\leq n}$ are observed. This is no longer the case in PU learning where only an incomplete set of positive data is available, the remaining is unlabeled. For each i, the observed label S_i is 1 if the class Y_i is positive and *selected* (*i.e.* labeled). Otherwise, the label S_i is 0. The true classes are affected by a class-dependent (thus asymmetric) label noise. The probability for a positive instance to be labeled is generally called the *propensity* (Bekker and Davis, 2020), and it may depend on the covariates:

$$e(x) = \mathbb{P}\left(S = 1|Y = 1, X = x\right).$$

Negative instances are never labeled:

$$\mathbb{P}\left(S = 1|Y = 0, X = x\right) = 0\ .$$

Note that the regression function associated with S is $\tilde{\eta}(x) = \mathbb{P}\left(S = 1|X = x\right)$. It depends on this additional label noise:

$$\tilde{\eta}(x) = e(x)\,\eta(x)\ . \tag{4}$$

The objective of PU learning is to use the incomplete information $(X_1, S_1), ...(X_n, S_n)$ to build a classifier able to predict the class Y given a new instance with covariates X.

This concept of completely asymmetric label noise was first pointed out by Elkan and Noto (2008). It is now common to define two general types of assumptions: Selected Completely At Random (SCAR) and Selected At Random (SAR).

SCAR: PU learning without selection bias. The propensity $e(x) = e$ does not depend on the covariates x. This applies in situations where every positive instance has an equal probability of being selected (labeled). In this case, the conditional distributions of X given $Y = 1$ ($\mathbb{P}_1$) and given $S = 1$ ($\tilde{\mathbb{P}}_1$) are the same. In other words, labeled instances are a representative sub-sample of positive instances.

SAR: PU learning with selection bias. The probability for an instance to be selected depends on its covariates. Hence, labeled instances are a biased sample of positive instances. For example, in mechanical design, a specimen subjected to higher stress is more likely to break, which results in a higher probability of a crack being detected. This is clearly a situation where the SCAR assumption does not hold.

In this section, we focus on the definition of loss functions that enable learning in PU learning setting. After explaining why labeled-unlabeled classifiers are limited, we will introduce an unbiased empirical risk for PU learning under the SCAR assumption (cf. Du Plessis et al., 2014), which generalizes to the SAR assumption (cf. Bekker et al., 2020).

3.1 Bias Issue with Labeled-Unlabeled Classification

A natural idea to address a PU learning problem is to consider labeled instances as positive and every unlabeled instance as negative. Standard classification methods then allow to identify a classifier $\widehat{g}_{NT}$. In the literature, such a classifier is called a *non-traditional classifier* (Elkan and Noto, 2008) because it is meant to give good predictions on S instead of Y. As the number of training examples increases, we can then expect $\widehat{g}_{NT}$ to get closer to Bayes classifier $\tilde{g}^*$ for the classification of S given X, which is not what we are looking for. Indeed, $\tilde{g}^*$ is *a priori* different from g^* as the regression function $\tilde{\eta}(x) = (Y = 1|X = x)$ is different from $\eta(x)$ (cf. Equation 4).

Nevertheless, in specific situations, the non-traditional classifier is robust to PU learning label noise. Cannings et al. (2020) showed for example that $\tilde{g}^* = g^*$ if:

$$e(x) \geq \frac{1}{2\eta(x)},\ \text{for all } x \in \mathbb{R}^d \text{ such that } \eta(x) \geq \frac{1}{2}\ . \tag{5}$$

Note that this is part of a more general result from Cannings et al. (2020) that encompasses binary classification with asymmetric and instance-dependent label noise. Under the condition from Equation 5, any consistent non-traditional classifier is a consistent traditional classifier. In other words, as the training sample size increases, $\widehat{g}_{NT}$ gets closer to $\tilde{g}^*$, which is identical to g^*.

This condition requires every positive instance ($\eta(x) > \frac{1}{2}$) difficult to classify ($\eta(x)$ close to $\frac{1}{2}$) to have propensity close enough to 1. Instances easier to classify ($\eta(x)$ close to 1) can undergo label noise without harming the consistency. However, the label noise cannot exceed $\frac{1}{2}$ or, in other words, the propensity can never be smaller than $\frac{1}{2}$.

This condition is thus restrictive in the context of PU learning under the SAR assumption for two main reasons. On the one hand, in many realistic situations, the propensity (*i.e.* the probability for a positive instance to be labeled) is correlated to the difficulty of classifying the observation. A positive instance difficult to classify tends to have a low propensity, which clearly violates the condition given in Equation 5. On the other hand, we cannot expect the propensity to be greater than $\frac{1}{2}$. In *text classification* or *spam review detection*, as the process of labeling is both difficult and time-consuming, only a small fraction of positive instances gets labeled, which suggests a propensity lower than $\frac{1}{2}$.

Before dealing with convergence rates, it is crucial to have methods for building consistent classifiers under more general conditions than Equation 5.

3.2 Unbiased Empirical Risk Minimization Under the SCAR Assumption

In this subsection, we assume that the SCAR assumption is satisfied, which means that the propensity is constant:

$$e(x) = e_m > 0 \ .$$

In order to compensate for label noise due to PU Learning under the SCAR assumption, Du Plessis et al. (2014) showed in the case-control setting that a consistent classifier could be found by minimizing an unbiased version of the risk. Using the convenient decomposition of $\mathbb{P}_X$ distribution (Equation 1), the misclassification risk can be rewritten only with $\mathbb{P}_X$ and $\mathbb{P}_1$.

$$\begin{aligned} R\left(g\right) &= \alpha\, \mathbb{P}_1\left(g(X) \neq 1\right) + (1-\alpha)\, \mathbb{P}_0\left(g(X) \neq 0\right) \\ &= \alpha\left(\mathbb{P}_1\left(g(X) \neq 1\right) - \mathbb{P}_1\left(g(X) \neq 0\right)\right) + \mathbb{P}_X\left(g(X) \neq 0\right). \end{aligned} \tag{6}$$

Therefore, as labeled instances are a representative sub-sample of positive instances, a consistent classifier can be found by minimizing the following risk:

$$\widehat{R}_n^{SCAR}(g) = \frac{\alpha}{N_L} \sum_{i=1}^{n} \mathbb{1}_{S_i=1}\left[\mathbb{1}_{g(X_i)\neq 1} - \mathbb{1}_{g(X_i)\neq 0}\right] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(X_i)\neq 0} \tag{7}$$

where $N_L = \sum_{i=1}^{n} \mathbb{1}_{S_i=1}$ is the number of labeled instances. Note that Du Plessis et al. (2014) considered the case-control setting where the number of labeled instances N_L is fixed, which is slightly different from our setting. One of the main properties of $\widehat{R}_n^{SCAR}(g)$ is that it is an unbiased estimate of the true risk, as we have:

$$\mathbb{E}\left[\widehat{R}_n^{SCAR}(g)\right] = \mathbb{P}\left(g(X) \neq Y\right) \ .$$

The proof of Du Plessis et al. (2014) extends to the one-sample-setting where N_L is random:

$$\begin{aligned}
\mathbb{E}\left[\widehat{R}_n^{SCAR}(g)\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\alpha}{N_L}\mathbb{1}_{S_i=1}\mathbb{E}\left[\mathbb{1}_{g(X_i)\neq 1} - \mathbb{1}_{g(X_i)\neq 0} \,|\, S_i\right]\right] + \mathbb{P}_X\left(g(X) \neq 0\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\alpha}{N_L}\mathbb{1}_{S_i=1}\left[\mathbb{P}\left(g(X_i) \neq 1 \,|\, S_i\right) - \mathbb{P}\left(S_i = 1, g(X_i) \neq 0 \,|\, S_i\right)\right]\right] \\
&+ \mathbb{P}_X\left(g(X) \neq 0\right) \\
&= \alpha \sum_{i=1}^{n} \mathbb{E}\left[\frac{\mathbb{1}_{S_i=1}}{N_L}\left(\mathbb{P}_1\left(g(X) \neq 1\right) - \mathbb{P}_1\left(g(X) \neq 0\right)\right)\right] + \mathbb{P}_X\left(g(X) \neq 0\right) \quad (8a) \\
&= \alpha\left[\mathbb{P}_1\left(g(X) \neq 1\right) - \mathbb{P}_1\left(g(X) \neq 0\right)\right] + \mathbb{P}_X\left(g(X) \neq 0\right) \ . \quad (8b)
\end{aligned}$$

Equation 8a results from the fact that under the SCAR assumption, the conditional distribution of X given $S = 1$ is the same as the conditional distribution of X given $Y = 1$ ($\mathbb{P}_1$). Finally, Equation 8b matches the decomposition of Equation 6, ending the proof.

Computing the risk $\widehat{R}_n^{SCAR}$ requires α to be known. Alternatively, another empirical risk can be written:

$$\widehat{R'}_n^{SCAR}(g) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}_{S_i=1}}{e_m}\left(\mathbb{1}_{g(X_i)\neq 1} - \mathbb{1}_{g(X_i)\neq 0}\right) + \mathbb{1}_{g(X_i)\neq 0}\right] \ . \quad (9)$$

This risk remains unbiased and consistent but requires the knowledge of the constant propensity e_m instead of the class prior α. The unbiasedness of $\widehat{R'}_n^{SCAR}$ will be proved in Subsection 3.3 as a special case of the more general SAR setting.

3.3 Extension to PU Learning Under the SAR Assumption

For now, PU learning under the SAR assumption is a difficult problem and there are only a few results in the literature (cf. Bekker et al., 2020; He et al., 2018; Gong et al., 2021). We recall that empirical risk minimization under the SCAR assumption requires extra knowledge on the model (class prior or propensity). In practice, these parameters are usually estimated (cf. Blanchard et al., 2010; Du Plessis and Sugiyama, 2014; Jain et al., 2016; Ramaswamy et al., 2016; Bekker and Davis, 2018). In order to provide a consistent empirical risk in the SAR setting, additional assumptions are needed to avoid identifiability issues. In the literature, different settings have been studied. He et al. (2018) assume that the propensity $e(x)$ is an increasing function of $\eta(x)$. Bekker et al. (2020) and Gong et al. (2021) suggest a *parametric* model on the propensity. Bekker et al. (2020) also study the case where the propensity is known for labeled instances which enables an empirical risk minimization approach similar to Du Plessis et al. (2014).

In this paper, following Bekker et al. (2020), we will focus on PU learning under the SAR assumption where the propensity is known for labeled instances. We argue that this setting is sufficient to derive interesting risk bounds and assess the difficulty of PU learning tasks. However restrictive this assumption may seem, we insist that only the propensity for labeled instances is needed; therefore, an exhaustive knowledge of the propensity is not required. When the propensity is unknown, it could be estimated using prior knowledge on the labeling mechanism (when available) or by defining a parametric model on the propensity (Bekker et al., 2020; Gong et al., 2021). In this case though, the theoretical analysis would have to account for the difference between the estimated propensity and the true one. However, it is worth noting that the lower bound on the minimax risk established in Subsection 4.2 does not require this assumption on the knowledge of the propensity.

Assuming that the propensity is known for labeled instances, Bekker et al. (2020) generalized the empirical risk in Equation 9 to obtain an unbiased empirical risk for PU learning under the SAR assumption. More particularly, they define the following loss function:

$$\begin{aligned} r_{SAR}\left(g,(X,S)\right) &= \frac{\mathbb{1}_{S=1}}{e(X)}\left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right) + \mathbb{1}_{g(X)\neq 0} \\ &= \frac{\mathbb{1}_{S=1}}{e(X)}\left(2\,\mathbb{1}_{g(X)\neq 1} - 1\right) + \mathbb{1}_{g(X)\neq 0}\ . \end{aligned}$$

The empirical risk is then the empirical mean:

$$\widehat{R}_n^{SAR}(g) = \frac{1}{n}\sum_{i=1}^{n} r_{SAR}\left(g,(X_i,S_i)\right)\ . \tag{10}$$

This time, the labeled instances are weighted by the inverse of their propensity. Clearly, $\widehat{R'}_n^{SCAR}$ in Equation 9 is a special case of $\widehat{R}_n^{SAR}$ under the SCAR assumption ($e(x) = e_m$).

Bekker et al. (2020) studied maximum deviations between this latter empirical risk $\widehat{R}_n^{SAR}$ and the empirical risk for standard classification $\widehat{R}_n$. They then used concentration inequalities to derive an upper bound on the deviations between the two quantities with high probability. As we are interested in studying the deviations between $\widehat{R}_n^{SAR}$ and the true risk R directly, we compute the total expectation of $\mathbb{E}\left[\widehat{R}_n^{SAR}(g)\right] = \mathbb{E}\left[r_{SAR}\left(g,(X,S)\right)\right]$ shedding light on the fact that for any g, $\widehat{R}_n^{SAR}(g)$ is an unbiased estimate of the true risk $R(g)$.

$$\begin{aligned} \mathbb{E}\left[r_{SAR}\left(g,(X,S)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[r_{SAR}\left(g,(X,S)\right)|X\right]\right] \\ &= \mathbb{E}\left[\frac{1}{e(X)}\left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right)\mathbb{P}\left(S=1|X\right)\right] + \mathbb{P}_X\left(g(X)\neq 0\right) \\ &= \mathbb{E}\left[\frac{1}{e(X)}\left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right)\eta(X)e(X)\right] + \mathbb{P}_X\left(g(X)\neq 0\right) \\ &= \mathbb{E}\left[\left(\mathbb{1}_{g(X)\neq 1} - \mathbb{1}_{g(X)\neq 0}\right)\mathbb{1}_{Y=1}\right] + \mathbb{P}_X\left(g(X)\neq 0\right) \\ &= \alpha\left(\mathbb{P}_1\left(g(X)\neq 1\right) - \mathbb{P}_1\left(g(X)\neq 0\right)\right) + \mathbb{P}_X\left(g(X)\neq 0\right) \\ &= R(g)\ . \end{aligned}$$

where the last line comes from Equation 6. Then, $\widehat{R}_n^{SAR}$ is indeed unbiased:

$$\mathbb{E}\left[\widehat{R}_n^{SAR}(g)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[r_{SAR}\left(g,(X_i,S_i)\right)\right] = \mathbb{P}\left(g(X)\neq Y\right) = R(g)\ . \tag{11}$$

It is important to note that, while the empirical risks $\widehat{R}_n^{SCAR}$ and $\widehat{R}_n^{SAR}$ in Equations 7 and 10 are here studied using the $0-1$ loss, they can be defined for any arbitrary loss function (cf. Du Plessis et al., 2014; Bekker et al., 2020).

4. Main Results

We are now in a position to state our results. First, we present an upper bound on the excess risk for PU learning under the SAR assumption. We then show that the rate achieved is almost optimal by providing a lower bound on the minimax risk. Both bounds explicitly quantify the impact of label noise specific to PU learning.

4.1 An Upper Bound for PU Learning Excess Risk Under the SAR Assumption

We recall that, in PU learning, the true classes $(Y_i)_{1\leq i\leq n}$ are no longer available for training. A classifier is then built as a minimizer of the unbiased empirical risk introduced in Equation 10:

$$\widehat{g}_{PU} \in \underset{g\in\mathcal{G}}{\operatorname{Argmin}}\, \widehat{R}_n^{SAR}(g)\ .$$

We recall that the risk $\widehat{R}_n^{SAR}$ is unbiased (Equation 11). Let $\overline{R_n^{SAR}}$ denote the centered empirical risk:

$$\overline{R_n^{SAR}}(g) = \widehat{R}_n^{SAR}(g) - \mathbb{P}\left(g(X) \neq Y\right)\ .$$

Bekker et al. (2020) study the deviations between $\widehat{R}_n^{SAR}\left(\widehat{g}_{PU}\right)$ and $\widehat{R}_n\left(\widehat{g}_{PU}\right)$ and provide an upper bound in the case where $\mathcal{G}$ is a *finite* family of classifiers. Besides, the influence of $e(\cdot)$ on the upper bound is not discussed. Our objective here is to provide a uniform upper bound on $\ell(\widehat{g}_{PU}, g^*)$ and explicitly show its dependence in $e(\cdot)$. In our setting, $\mathcal{G}$ is an *infinite* set of functions. Its complexity is controlled by its VC dimension $V < +\infty$. Following Massart and Nédélec (2006), we consider the following separability assumption, which is key to work with the possibly uncountable class $\mathcal{G}$:

(A_1) There exists a countable subset $\mathcal{G}'$ dense in $\mathcal{G}$ in the sense that for each $g \in \mathcal{G}$, there exists a sequence $(g_k)_{k\geq 0}$ of classifiers of $\mathcal{G}'$ such that, for every $(x, s) \in \mathbb{R}^d \times \{0, 1\}$:

$$r_{SAR}\left(g_k, (x, s)\right) \underset{k\to+\infty}{\longrightarrow} r_{SAR}\left(g, (x, s)\right)\ .$$

In other words, any classifier of g of $\mathcal{G}$ is "arbitrarily close" to some element of $\mathcal{G}'$.

In addition, we want our upper bound on the excess risk to account for the difficulty of the classification task explicitly. Then, as $|2\eta(x) - 1|$ quantify the difficulty of classifying x, we introduce the following assumption (Massart and Nédélec, 2006):

(A_2) $\quad \exists\, h > 0,\ \forall x \in \mathbb{R}^d,\ |2\eta(x) - 1| \geq h\ .$

Assumption (A_2) will be referred to as *Massart margin* assumption in the rest of the paper.

Before stating our upper bound for PU learning under the SAR assumption, we introduce the standard notations for maximum and minimum: for any couple of real (a, b), $a \vee b$ $(a \wedge b)$ denotes the maximum (minimum) between a and b.

We are now able to state our first result.

Theorem 1 (Upper risk bound for PU learning under the SAR assumption) *Let $\widehat{g}_{PU}$ be a minimizer of the unbiased empirical risk for PU learning under the SAR assumption:*

$$\widehat{g}_{PU} \in \underset{g \in \mathcal{G}}{\operatorname{Argmin}}\, \widehat{R}_n^{SAR}(g) \ .$$

Suppose that separability (A_1) and Massart margin (A_2) assumptions hold, and that the propensity $e(\cdot)$ is greater than $e_m > 0$. Then, we have the following upper bound on the excess risk:

$$\mathbb{E}\left[\ell\left(\widehat{g}_{PU}, g^*\right)\right] \leq \kappa_1 \left[\frac{V}{n\, e_m\, h}\left(1 + \log\left(\frac{n\, h^2}{V} \vee 1\right)\right) \wedge \sqrt{\frac{V}{n\, e_m}}\right] \tag{12}$$

where $\kappa_1 > 0$ is an absolute constant.

Remarks: The upper bound in Equation 12 is uniform on the set of probability distributions for which $g^* \in \mathcal{G}$ and Massart margin condition (A_2) is satisfied with constant h $(\mathcal{P}\left(\mathcal{G}, h\right))$. This can be re-written as follows:

$$\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{G}, h)} \mathbb{E}\left[\ell\left(\widehat{g}_{PU}, g^*\right)\right] \leq \kappa_1 \left[\frac{V}{n\, e_m\, h}\left(1 + \log\left(\frac{n\, h^2}{V} \vee 1\right)\right) \wedge \sqrt{\frac{V}{n\, e_m}}\right] \tag{13}$$

Note that the assumption $e(x) \geq e_m$ is an assumption on the label noise. As the biased regression function is $\tilde{\eta}(x) = \eta(x)\, e(x)$ (cf. Equation 4), this assumption together with assumption (A_2) control the difficulty of the PU learning task.

In Equation 12, the convergence rate is of order $\mathcal{O}\left(\frac{V}{n\, h\, e_m}\right)$ (if we let aside the logarithmic term) when h is higher than $\sqrt{V/n\, e_m}$. When h becomes smaller than $\sqrt{V/n\, e_m}$, the rate is of order $\mathcal{O}\left(\sqrt{V/n\, e_m}\right)$. These two regimes are analogous to standard classification risk bounds as recalled in Subsection 2.2. In particular, when $e_m = 1$, all positive examples are labeled and we are then in a standard classification setting $(S = Y)$. In this case, the upper bound exactly matches the known upper bound rates in the standard classification setting (Equation 3 and Equation 2). Conversely, as e_m gets lower, the upper bound increases. This means without surprise that PU learning deteriorates the generalization bound: Theorem 1 quantifies this effect through the coefficients $1/e_m$ and $1/\sqrt{e_m}$.

Let N_L be the number of labeled instances in the training set. Under the SCAR assumption $(e(x) = e_m)$, $n\, e_m$ from Equation 12 is linked to the expectation of the number of labeled instances in the training set:

$$\mathbb{E}\left[N_L\right] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{S_i = 1}\right] = n\, \mathbb{P}\left(S = 1\right) = n\, \alpha\, e_m$$

where $\alpha = \mathbb{P}(Y=1)$ is the class prior. This illustrates a natural intuition on PU learning: the upper bound on the excess risk is related to the number of fully labeled examples. Hence, good prediction performances cannot be expected if the number of labeled examples among the positives is too low or equivalently if the propensity is too low.

The detailed proof of Theorem 1 can be found in Appendix A. It consists in establishing controls on the variance of increments of $r_{SAR}(\cdot)$ and uniform bounds on the empirical process $\left(\overline{R_n^{SAR}}(g)\right)_{g\in\mathcal{G}}$. A general risk bound result for empirical risk minimizers is then applied.

So far, we have provided an upper bound on generalization risk for unbiased empirical risk minimization in PU learning under the SAR assumption. There is, however, no proof that this rate is optimal. In other words, is there another procedure that can learn a classifier $\widehat{g}$ that outperforms $\widehat{g}_{PU}$? A lower bound will help to answer this question.

4.2 A Lower Bound on the Minimax Risk

In order to assess the optimality of the upper bound (Equation 12), we analyze and provide a lower bound on the minimax risk.

The minimax risk is the risk of the classification procedure that performs best in the worst case. For any given estimate $\widehat{g}$, we recall that its generalization risk is measured as $\mathbb{E}\left[\ell\left(\widehat{g}, g^*\right)\right]$. The minimax risk is denoted $\mathcal{R}\left(\mathcal{G}, h\right)$ and is defined as follows:

$$\mathcal{R}\left(\mathcal{G}, h\right) = \inf_{\widehat{g}\in\mathcal{G}}\left[\sup_{\mathbb{P}\in\mathcal{P}(\mathcal{G},h)} \mathbb{E}\left[\ell\left(\widehat{g}, g^*\right)\right]\right]$$

where the infimum is taken over the set of functions $\widehat{g}$ of $(X_i, S_i)_{1\leq i\leq n}$ such that $\widehat{g}$ belongs to $\mathcal{G}$.

The bound in Equation 13 is an obvious upper bound on the minimax risk. Theorem 2 establishes a lower bound on the minimax risk for PU learning under the SCAR assumption. Proposition 3 extends it to the SAR assumption.

Theorem 2 (Lower bound on the minimax risk under the SCAR assumption) *Suppose that $V \geq 2$ and $n\, e_m \geq V$. Let $h' = \sqrt{\frac{V}{n\, e_m}}$.*
Assuming $e(x) = e_m,\ \forall x \in \mathbb{R}^d$, there exists an absolute constant $\kappa_2 > 0$ such that:

(C_1) *if $h \geq h'$:*

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \kappa_2 \frac{V-1}{h\, n\, e_m}\ ; \tag{14}$$

(C_2) *if $h \leq h'$:*

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \kappa_2 \sqrt{\frac{V-1}{n\, e_m}}\ . \tag{15}$$

Remarks

The lower bounds in Theorem 2 explicitly depend on V, n, h and e_m. The cases (C_1) and (C_2) highlight a trade-off between the expected number of fully labeled instances (proportional to $n\,e_m$), the complexity of the model V and the margin condition (A_2) represented by h. The restriction of these results to the standard classification setting ($e_m = 1$) exactly matches existing results (see Massart and Nédélec, 2006). Theorem 2 moreover provides the influence of propensity e_m in PU learning framework under the SCAR assumption. As for the upper bound (cf. Theorem 1), the lower bound (Equation 14) is affected the same way with a degradation of order $1/e_m$ over the minimax rate when Massart margin condition (A_2) is satisfied with h high enough, in case (C_1). In this case, the lower bound rate almost matches the upper bound up to a logarithmic factor. In the second case (C_2), the lower bound (Equation 15) is of order $\sqrt{V/n\,e_m}$ which exactly matches the rate of the upper bound in this regime. In this sense, $\widehat{g}_{PU}$ obtained through unbiased empirical risk minimization is almost optimal as it almost achieves the minimax convergence rates. Finally, it is important to note that this lower bound result remains valid when the propensity is unknown, contrary to the upper bound where the knowledge of the propensity is required to obtain $\widehat{g}_{PU}$.

The detailed proof of Theorem 2 can be found in Appendix B.1. It makes use of similar arguments as for minimax lower bounds in the standard classification setting. First, the expression of the minimax risk is simplified by choosing a specific set of probabilities satisfying the margin and noise conditions. Then Assouad lemma (Yu, 1997) is applied to provide a lower bound on this simplified expression, where the singularity of PU learning mainly interferes.

To extend the result to the SAR assumption, we need an extra condition:

(A_3) $\forall \varepsilon > 0,\ \exists (x_1, ..., x_V) \in \left(\mathbb{R}^d\right)^V$ scattered by $\mathcal{G}$ and such that:

$$\sup_{i \in \{1,...,V\}} e(x_i) \le e_m + \varepsilon\ .$$

This assumption is technical. It is used in the first step of the proof of the minimax lower bound as it allows us to choose a convenient family of discrete probability distributions satisfying the noise assumptions. Assumption (A_3) is fulfilled in natural situations, for example, when $e(\cdot)$ is continuous and $\mathcal{G}$ is the set of linear classifiers in $\mathbb{R}^d$.

Proposition 3 (Lower bound on the minimax risk under the SAR assumption)

Theorem 2 extends to the SAR assumption if the propensity $e(\cdot)$ greater than $e_m > 0$ and if assumption (A_3) is satisfied.

The proof of the above proposition can be found in Appendix B.2. The same remarks as for Theorem 2 remain valid under the SAR assumption when assumption (A_3) is satisfied. In particular, in regimes (C_1) and (C_2), the minimax rate still matches the upper bound rate Equation 12 up to the logarithmic factor.

5. Conclusion

In this paper, we provided a theoretical study of PU learning under the SAR assumption, *i.e.* when the probability for an instance to be labeled depends on its covariates. Assuming partial knowledge of the propensity, a consistent classifier can be identified by minimizing a conveniently weighted empirical risk. We established a general non-asymptotic upper bound on the excess risk that naturally extends known risk bounds in the standard classification setting. By providing a minimax lower bound, we then showed that the convergence rates are optimal up to a logarithmic term. From a practical point of view, these bounds help to understand the difficulty of the PU learning task in terms of the propensity. A low propensity results in fewer positive instances labeled and, thus, a more difficult task. Conversely, as the propensity tends to 1, the performances of PU learning tend to those of fully supervised classification. Finally, both results show that fast rates can be achieved under margin conditions.

As a future perspective, it would be interesting to study how some assumptions made on the propensity could be relaxed. In particular, future work could assess whether or not the theoretical guarantees proved in this paper still hold when the propensity is estimated. Likewise, we may wonder if these results could be extended if the lower bound on the propensity only holds with high probability. To bridge the gap between these theoretical results and practical PU learning methodologies adapted to the SAR assumption, several challenges remain open. For instance, the estimation of the propensity is a difficult problem. Besides, minimizing the unbiased empirical risk in PU learning based on $0-1$ loss requires solving computationally difficult combinatorial optimization problems. The use of convex loss functions would facilitate the optimization. Then, it could be interesting to study how our theoretical results extend to such cases.

Acknowledgments

This work was carried out within the framework of the partnership between Stellantis and the OpenLab AI with the financial support of the ANRT for the CIFRE contract n°2019/1131.

Appendix A. Proof of Theorem 1

The proof is organized as follows. We first state a general upper bound result for empirical risk minimizers adapted to the case where the loss function takes values in an arbitrary interval $[a, b]$ with $a < b$ (cf. Subsection A.1). Then, we show that the PU learning loss function satisfies the assumptions of this general result (cf. Subsection A.2). Finally, we deduce the upper bound as the solution of a fixed point equation (cf. Subsection A.3).

A.1 General Risk Upper Bound on Empirical Risk Minimizers

We begin by stating a general upper bound theorem for empirical risk minimizers.

Theorem 4 (General upper bound for empirical risk minimizers)
Let r be an unbiased loss function with values in $[a, b]$, $\widehat{R}_n$ the empirical risk, and $\overline{R_n}$ the centered empirical risk. Let g^* denote the Bayes classifier and let $\widehat{g}$ be a minimizer of the empirical risk over a class $\mathcal{G}$ for which we assume separability condition (A_1). Let ℓ denote the excess risk. We assume that:

(B_1) *there exists a positive and symmetric function d such that for any couple of classifiers (g, g'):*

$$Var\left[r\left(g', (X, S)\right) - r\left(g, (X, S)\right)\right] \leq d^2\left(g', g\right);$$

(B_2) *there exists a non-decreasing function w continuous on $\mathbb{R}_+$, such that $x \mapsto \frac{w(x)}{x}$ is non-increasing on $\mathbb{R}_+^*$, with $w(\sqrt{b-a}) \geq b - a$ and ensuring for any classifier g:*

$$d(g^*, g) \leq w\left(\sqrt{\ell(g^*, g)}\right);$$

(B_3) *there exists a non-decreasing function Φ continuous on $\mathbb{R}_+$, such that $x \mapsto \frac{\Phi(x)}{x}$ is non-increasing with $\Phi(b - a) \geq b - a$ and ensuring:*

$$\forall h \in \mathcal{G}', \sqrt{n}\,\mathbb{E}\left[\sup_{g\in\mathcal{G}', d(g,h)\leq\sigma} \overline{R_n}(h) - \overline{R_n}(g)\right] \leq \Phi(\sigma).$$

for every positive σ such that $\Phi(\sigma) \leq \sqrt{n}\frac{\sigma^2}{b-a}$, where $\mathcal{G}'$ comes from separability condition (A_1).

Then there exists an absolute constant $\kappa > 0$ such that:

$$\mathbb{E}\left[\ell(g^*, \widehat{g})\right] \leq \kappa\,\varepsilon_*^2, \tag{16}$$

where ε_* is the unique positive solution of the following equation:

$$\sqrt{n}\,\varepsilon_*^2 = \Phi\left(w\left(\varepsilon_*\right)\right). \tag{17}$$

Proof The above result follows from the application of Massart and Nédélec's theorem (2006, Theorem 2) using the re-scaled risk $\tilde{r} = \frac{r-a}{b-a}$ and the functions $\tilde{d}(g, g') = \frac{d(g,g')}{b-a}$, $\tilde{w}(x) = \frac{1}{b-a}\,w\left(x\sqrt{b-a}\right)$ and $\tilde{\Phi}(x) = \frac{1}{b-a}\Phi\left((b-a)\,x\right)$. This leads to the upper bound in Equation 16 solution of Equation 17. ■

Note that now, contrary to Massart and Nédélec's original result, (B_2) and (B_3) explicitly involve the length of the interval $[a, b]$. This will be accounted for in our proof.

A.2 Verification of Assumptions of Theorem 4 in the PU Learning Setting

We first recall the definition and the main property of PU learning loss function as defined in Subsection 3.3. We then exhibit three functions d, w, Φ fulfilling conditions (B_1), (B_2) and (B_3). Hence we show that the general upper bound result (*i.e.* Theorem 4) can be applied in PU learning context.

In the context of PU learning under the SAR assumption, we recall that the loss function r_{SAR} is defined as follows:

$$r_{SAR}\left(g,(X,S)\right) = \frac{\mathbb{1}_{S=1}}{e(X)}\left(2\,\mathbb{1}_{g(X)\neq 1} - 1\right) + \mathbb{1}_{g(X)\neq 0}$$

where $e(x) = \mathbb{P}\left(S=1\,|\,Y=1, X=x\right)$ is the propensity assumed to be known for labeled observations. Knowing that the propensity greater than $e_m > 0$, the loss function is then at values in $\left[1-\frac{1}{e_m}, \frac{1}{e_m}\right]$, an interval of length:

$$C_e = \frac{2}{e_m} - 1\ . \tag{18}$$

We have seen that this empirical risk is an unbiased estimate of the true risk (cf. Equation 11):

$$\mathbb{E}\left[r_{SAR}\left(g,(X,S)\right)\right] = \mathbb{P}\left(g(X)\neq Y\right)\ .$$

In order to apply the general upper bound theorem (Theorem 4) to the PU learning risk minimizer, we need to identify three functions d, w, Φ satisfying conditions (B_1), (B_2) and (B_3). These functions are crucial since the upper bound is the solution of a fixed point equation involving them. The choice of functions d, w and Φ will be a consequence of Propositions 5, 6 and 7.

Proposition 5 *For any pair of classifiers (g,g'):*

$$Var\left[r_{SAR}\left(g',(X,S)\right) - r_{SAR}\left(g,(X,S)\right)\right] \leq 2\,C_e\,\mathbb{E}\left[\left|g(X)-g'(X)\right|^2\right]\ ,$$

where C_e is given by Equation 18.

Remark A direct consequence of the above proposition is that the function d defined as:

$$d(g,g') = \sqrt{2C_e}\sqrt{\mathbb{E}\left[\left|g(X)-g'(X)\right|^2\right]} \tag{19}$$

satisfies condition (B_1).

Proof We first provide an upper bound on the variance of increments of r_{SAR}:

$$\begin{aligned}
Var\left[r_{SAR}\left(g\right) - r_{SAR}\left(g'\right)\right] &\leq \mathbb{E}\left[\left(r_{SAR}\left(g\right) - r_{SAR}\left(g'\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(g(X) - g'(X)\right)^2\left(1 - \frac{2\mathbf{1}_{S=1}}{e(X)}\right)^2\right] \\
&= \mathbb{E}\left[\left(g(X) - g'(X)\right)^2\mathbb{E}\left[\left(1 - \frac{2\mathbf{1}_{S=1}}{e(X)}\right)^2|X\right]\right] \\
&= \mathbb{E}\left[\left(g(X) - g'(X)\right)^2\left(1 + 4\eta(X)\frac{1-e(X)}{e(X)}\right)\right] && \text{(20a)} \\
&\leq \left(1 + 4\frac{1-e_m}{e_m}\right)\mathbb{E}\left[\left(g(X) - g'(X)\right)^2\right] && \text{(20b)} \\
&\leq 2\,C_e\,\mathbb{E}\left[\left(g(X) - g'(X)\right)^2\right] \ .
\end{aligned}$$

We then use the fact that $\mathbb{E}\left[\mathbb{1}_{S=1}|X\right] = \mathbb{E}\left[\eta(X)\,e(X)\right]$ to get Equation 20a. And Equation 20b results from the fact that $\eta(X)$ is less than 1 and $e(X)$ is greater than e_m. ■

Proposition 6 *For any classifier g:*

$$d(g, g^*) \leq \sqrt{\frac{2\,C_e}{h}}\sqrt{\ell(g, g^*)} \ .$$

for d defined in Equation 19.

Remark As a consequence, the function w defined as:

$$w(x) = \sqrt{\frac{2\,C_e}{h}}\,x \ . \tag{21}$$

satisfies Assumption (B_2): w is continuous on $\mathbb{R}_+$, non-decreasing, such that $x \mapsto \frac{w(x)}{x}$ is non-increasing and $w\left(\sqrt{C_e}\right) \geq C_e$, and such that:

$$d(g^*, g) \leq w\left(\sqrt{\ell\left(g^*, g\right)}\right) \ .$$

Let $h' = \sqrt{V/n\,e_m}$. Note that the function

$$w_0(x) = \sqrt{2\,C_e} \vee x\,\sqrt{2\,C_e/h'} \tag{22}$$

also satisfies assumption (B_2).

Proof The excess risk can be expressed in terms of $\eta(X)$ as follows:

$$\begin{aligned}
\ell(g, g^*) &= \mathbb{P}\left(g(X) \neq Y\right) - \mathbb{P}\left(g^*(X) \neq Y\right) \\
&= \mathbb{E}\left[\left|g(X) - g^*(X)\right|^2\left|2\,\eta(X) - 1\right|\right] \ .
\end{aligned} \tag{23}$$

Then, using the margin assumption (A_2) and the definition of d (cf. Equation 19), we have the following lower bound on the excess risk:

$$\begin{aligned}\ell(g, g^*) &= \mathbb{E}\left[(g(X) - g^*(X))^2\, |2\,\eta(X) - 1|\right] \\ &\geq h\,\mathbb{E}\left[(g(X) - g^*(X))^2\right] \\ &= \frac{h}{2\,C_e} d^2(g, g^*)\ .\end{aligned}$$

Taking the square root on both sides finishes the proof. ■

The next proposition states the existence of Φ fulfilling (B_3). We recall that the subset $\mathcal{G}' \subset \mathcal{G}$ is given by the separability assumption (A_1) and that the constant C_e is defined in Equation 18.

Proposition 7 *Assume $\mathcal{G}$ has finite VC dimension V and $\mathcal{G}'$ is given by separability assumption (A_1). There exists an absolute constant $K \geq 1$ such that the function Φ defined as*

$$\Phi(\sigma) = K\sigma\sqrt{V\left[1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)\right]} \tag{24}$$

satisfies:

$$\sqrt{n}\,\mathbb{E}\left[\sup_{g\in\mathcal{G}', d(g,h)\leq\sigma} \overline{R_n^{SAR}}(g_0) - \overline{R_n^{SAR}}(g)\right] \leq \Phi(\sigma)$$

for all $g_0 \in \mathcal{G}'$ and for every σ such that $\Phi(\sigma) \leq \sqrt{n}\,\frac{\sigma^2}{C_e}$.

Proof We consider a fixed $g_0 \in \mathcal{G}'$ along the proof and use the notation:

$$W = \sup_{g\in\mathcal{G}', d(g,g_0)\leq\sigma} \overline{R_n^{SAR}}(g_0) - \overline{R_n^{SAR}}(g)\ .$$

The main steps of the proof are: (i) rewrite W as the supremum of an empirical process over a class of functions; (ii) split the expression of W into two terms depending on the sign of $(g_0(x) - g(x))$ (W^+ and W^-) that will be processed similarly and independently; (iii) provide an upper bound on $\mathbb{E}[W^+]$ using a symmetrization principle (cf. Bousquet et al., 2003); (iv) apply a chaining inequality and Haussler bound (Bousquet et al., 2003; Massart and Nédélec, 2006); (v) a few calculations finish the proof. This proof uses the notion of entropy metrics: the definition and some useful properties are recalled in Appendix C.

(i) We start by rewriting the expression inside the supremum in W:

$$\begin{aligned}\overline{R_n^{SAR}}(g_0) - \overline{R_n^{SAR}}(g) &= \widehat{R}_n^{SAR}(g_0) - \widehat{R}_n^{SAR}(g) - \mathbb{E}\left[\widehat{R}_n^{SAR}(g_0) - \widehat{R}_n^{SAR}(g)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(r_{SAR}\left(g_0, (X_i, S_i)\right) - r_{SAR}\left(g, (X_i, S_i)\right)\right) - \mathbb{E}\left[\widehat{R}_n^{SAR}(g_0) - \widehat{R}_n^{SAR}(g)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(g_0(X_i) - g(X_i))\left(\frac{2\,\mathbb{1}_{S_i=1}}{e(X_i)} - 1\right) - \mathbb{E}\left[(g(X) - g_0(X))\left(\frac{2\,\mathbb{1}_{S=1}}{e(X)} - 1\right)\right] \\ &= (\mathbb{P}_n - \mathbb{P})\,(f_g),\end{aligned}$$

where $\mathbb{P}_n f_g$ and $\mathbb{P} f_g$ denote the empirical mean and the expectation of the function f_g:

$$f_g : (x, s) \mapsto (g_0(x) - g(x)) \left(\frac{2\,\mathbb{1}_{s=1}}{e(x)} - 1 \right).$$

Hence, denoting $\mathcal{F}(\sigma) = \{f_g : g \in \mathcal{G}',\ d(g_0, g) \leq \sigma\}$, we can write W as the supremum of the empirical process $(\mathbb{P}_n - \mathbb{P})(\cdot)$ over the set of functions $\mathcal{F}(\sigma)$:

$$W = \sup_{f \in \mathcal{F}(\sigma)} (\mathbb{P}_n - \mathbb{P})(f).$$

(ii) For any $g \in \mathcal{G}'$, we can decompose f_g depending on the sign of $(g_0(x) - g(x))$:

$$f_g\,(x, s) = \left(\frac{2\,\mathbb{1}_{s=1}}{e(x)} - 1 \right) \mathbb{1}_{g(x) > g_0(x)} - \left(\frac{2\,\mathbb{1}_{s=1}}{e(x)} - 1 \right) \mathbb{1}_{g_0(x) > g(x)}\ .$$

Then, introducing the following classes of functions

$$\mathcal{F}^+(\sigma) = \left\{ f : \mathbb{R}^d \times \{0,1\} \to \mathbb{R},\ \exists\, g \in \mathcal{G}',\ f(x,s) = \left[\frac{2\mathbb{1}_{s=1}}{e(X)} - 1 \right] \mathbb{1}_{g(x) > g_0(x)}\ ,\ d(g_0, g) \leq \sigma \right\}$$
$$\mathcal{F}^-(\sigma) = \left\{ f : \mathbb{R}^d \times \{0,1\} \to \mathbb{R},\ \exists\, g \in \mathcal{G}',\ f(x,s) = \left[\frac{2\mathbb{1}_{s=1}}{e(X)} - 1 \right] \mathbb{1}_{g(x) < g_0(x)}\ ,\ d(g_0, g) \leq \sigma \right\}$$

and the corresponding suprema

$$W^+ = \sup_{f \in \mathcal{F}^+(\sigma)} (\mathbb{P}_n - \mathbb{P})(f)$$
$$W^- = \sup_{f \in \mathcal{F}^-(\sigma)} (\mathbb{P} - \mathbb{P}_n)(f),$$

we decompose $\mathbb{E}[W]$ as follows:

$$\mathbb{E}[W] \leq \mathbb{E}\left[W^+\right] + \mathbb{E}\left[W^-\right].$$

We now process both terms separately, focusing on W^+ (the proof for the other term is almost identical).

(iii) We first apply a symmetrization principle to provide an upper bound on $\mathbb{E}[W^+]$ depending on a Rademacher average (cf. Bousquet et al., 2003):

$$\mathbb{E}\left[W^+\right] \leq \frac{2}{n} \mathbb{E}\left[\sup_{f \in \mathcal{F}^+(\sigma)} \sum_{i=1}^n \varepsilon_i f(X_i, S_i) \right]$$

where $(\varepsilon_i)_{1 \leq i \leq n}$ are i.i.d. Rademacher variables (*i.e.* $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}$).

(iv) Let $\delta^2 = \sup_{f\in\mathcal{F}^+(\sigma)} \mathbb{P}_n\left(f^2\right) \vee \sigma^2$. We apply a chaining inequality (lemma A.2, Massart and Nédélec 2006) which gives us the following inequality:

$$\mathbb{E}\left[W^+\right] \leq \frac{6}{\sqrt{n}}\mathbb{E}\left[\delta \sum_{j=0}^{+\infty} 2^{-j}\sqrt{H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right)}\right] \tag{25}$$

where H is the universal entropy metric (cf. Appendix C).

Let $\mathcal{A}_+ = \left\{\mathbb{1}_{g(x)>g_0(x)}, g \in \mathcal{G}'\right\}$, which can be considered as a set of classifiers and has VC dimension V at most. Using the fact that $H\left(\cdot, \mathcal{F}_+(\sigma)\right)$ is non-increasing (cf. Proposition 9), we have $\forall j \geq 0$:

$$H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right) \leq H\left(2^{-j-1}\sigma, \mathcal{F}_+(\sigma)\right)\ .$$

Applying Proposition 10, we obtain the following upper bound on the entropy of $\mathcal{F}_+(\sigma)$ in terms of the entropy of $\mathcal{A}_+$:

$$H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right) \leq H\left(2^{-j-1}\frac{\sigma}{C_e},\ \mathcal{A}_+(\sigma)\right)\ .$$

We are then in a position to apply Haussler bound (Proposition 11), to get an upper bound on the entropy in terms of the VC dimension of $\mathcal{A}_+$, which is no more than V:

$$H\left(2^{-j-1}\delta, \mathcal{F}_+(\sigma)\right) \leq \kappa\, V\left(1 + \log\left(2^{j+1}\frac{C_e}{\sigma} \vee 1\right)\right) \tag{26}$$

for some absolute constant $\kappa > 1$.

(v) Injecting Equation 26 in Equation 25, we get:

$$\begin{aligned}\mathbb{E}\left[W^+\right] &\leq 6\sqrt{\frac{\kappa\, V}{n}}\left[\sum_{j=0}^{+\infty} 2^{-j}\sqrt{1 + \log\left(2^{j+1}\frac{C_e}{\sigma} \vee 1\right)}\right]\mathbb{E}\left[\delta\right] \\ &\leq C(\sigma)\sqrt{\frac{V}{n}}\mathbb{E}\left[\delta\right] &(27a) \\ &\leq C(\sigma)\sqrt{\frac{V}{n}}\sqrt{\mathbb{E}\left[\delta^2\right]} &(27b)\end{aligned}$$

where $C(\sigma) = 12\left(1 + \log(2)\right)\sqrt{\kappa}\sqrt{1 + \log\left(\frac{C_e}{\sigma} \vee 1\right)}$. Equation 27a is a consequence of technical Lemma 12 in Appendix D, Equation 27b follows from Cauchy-Schwartz inequality.

Now, we provide an upper bound on $\mathbb{E}\left[\delta^2\right]$ in terms of $\mathbb{E}\left[W^+\right]$:

$$\begin{aligned}\mathbb{E}\left[\delta^2\right] &\leq \sigma^2 + \mathbb{E}\left[\sup_{f\in\mathcal{F}_+(\sigma)} \mathbb{P}_n\left(f^2\right)\right] \\ &\leq \sigma^2 + C_e\,\mathbb{E}\left[\sup_{f\in\mathcal{F}_+(\sigma)} \mathbb{P}_n\left(f\right)\right] \\ &\leq \sigma^2 + C_e\,\mathbb{E}\left[\sup_{f\in\mathcal{F}_+(\sigma)} \left(\mathbb{P}_n - \mathbb{P}\right)(f)\right] + C_e \sup_{f\in\mathcal{F}_+(\sigma)} \mathbb{P}(f) &(28)\end{aligned}$$

Let $f \in \mathcal{F}_+(\sigma)$ and define $g \in \mathcal{G}'$ such that $f(x,s) = \left[\frac{2\mathbb{1}_{s=1}}{e(x)} - 1\right]\mathbb{1}_{g_0(x)>g(x)}$ (and $d(g_0, g) \leq \sigma$). We have:

$$\begin{aligned}
\mathbb{P}(f) &= \mathbb{E}\left[\mathbb{E}\left[\frac{2\mathbb{1}_{S=1}}{e(X)} - 1|X\right]\mathbb{1}_{g_0(X)>g(X)}\right] \\
&= \mathbb{E}\left[(2\eta(X)-1)\,\mathbb{1}_{g_0(X)>g(X)}\right] \\
&\leq \mathbb{E}\left[|g_0(X) - g(X)|^2\right] \\
&= \frac{d^2(g_0, g)}{2\,C_e} \\
&\leq \frac{\sigma^2}{2\,C_e}
\end{aligned}$$

using Equation 19 and the definition of $\mathcal{F}_+(\sigma)$. Note that the above upper bound does not depend on $f \in \mathcal{F}_+(\sigma)$. Hence, we can use it in Equation 28 to obtain:

$$\mathbb{E}\left[\delta^2\right] \leq C_e\,\mathbb{E}\left[W^+\right] + \frac{3}{2}\sigma^2$$

Hence, coming back to $\mathbb{E}\left[W^+\right]$:

$$\mathbb{E}\left[W^+\right] \leq C(\sigma)\sqrt{\frac{V}{n}}\sqrt{C_e\,\mathbb{E}\left[W^+\right] + \frac{3}{2}\sigma^2}\ .$$

Taking the square on both sides and solving the second-order inequation in $\mathbb{E}\left[W^+\right]$ yields:

$$\mathbb{E}\left[W^+\right] \leq \frac{1}{2}C(\sigma)\sqrt{\frac{V}{n}}\left(C(\sigma)\,C_e\sqrt{\frac{V}{n}} + \sqrt{\frac{C(\sigma)^2\,C_e^2\,V}{n} + 6\sigma^2}\right)\ .$$

Therefore, whenever $\sigma \geq C(\sigma)\,C_e\,\sqrt{\frac{V}{n}}$:

$$\sqrt{n}\,\mathbb{E}\left[W^+\right] \leq 2\,\sigma\,C(\sigma)\sqrt{V}\ .$$

We can prove a similar upper bound on $\mathbb{E}\left[W^-\right]$. If we define $\Phi(\sigma) = 4\,\sigma\,C(\sigma)\sqrt{V}$, for all σ such that $\Phi(\sigma) \leq \sqrt{n}\,\frac{\sigma^2}{C_e}$ (condition of Proposition 7):

$$\sigma \geq C(\sigma)\,C_e\sqrt{\frac{V}{n}}\ .$$

Hence, we have the desired upper bound on $\mathbb{E}\left[W\right]$:

$$\sqrt{n}\,\mathbb{E}\left[W\right] \leq \Phi(\sigma)\ .$$

Note that the constant $K = 4\,C(\sigma)$ is greater than 1. ■

A.3 Upper Bounds on the Risk

In the previous subsection, we checked that Theorem 4 can be applied to PU learning under the SAR assumption. Hence, the upper bound on risk excess ε_*^2 is the unique solution to the fixed point equation:

$$\sqrt{n}\,\varepsilon_*^2 = \Phi\left(w(\varepsilon_*)\right) \tag{29}$$

where w is given in Equation 21 (or w_0 in Equation 22) and Φ in Equation 24.

$$w(x) = \sqrt{\frac{2\,C_e}{h}}\,x\ ,$$

$$w_0(x) = \sqrt{2\,C_e} \vee x\sqrt{\frac{2\,C_e}{h'}}\ ,$$

$$\Phi(\sigma) = K\sigma\sqrt{V\left[1+\log\left(\frac{C_e}{\sigma}\vee 1\right)\right]}\ .$$

We cannot explicitly solve this equation, but we can provide an upper bound on the solution which is enough to complete the proof of Theorem 1. The choice of w as Equation 21 or Equation 22 leads to two different upper bounds (Subsections A.3.1 and A.3.2) that together complete the proof of Theorem 1.

A.3.1 First Case

Using the known definitions of w in Equation 21 and Φ in Equation 24, Equation 29 can be rewritten as:

$$\sqrt{n}\,\varepsilon_*^2 = K\,\varepsilon_*\sqrt{\frac{2\,C_e}{h}}\sqrt{V\,\left[1+\log\left(\frac{\sqrt{C_e\,h}}{\sqrt{2}\varepsilon_*}\vee 1\right)\right]}$$

Because the logarithmic term is always non-negative and $K \geq 1$, we get:

$$\varepsilon_* \geq \sqrt{\frac{2\,C_e\,V}{n\,h}}\ .$$

Using this on the logarithmic term, we obtain the following upper bound on ε_*:

$$\begin{aligned}\varepsilon_* &\leq K\sqrt{\frac{2\,C_e\,V}{n\,h}}\sqrt{1+\log\left(\frac{\sqrt{n}\,h}{2\sqrt{V}}\vee 1\right)}\\ &\leq K\sqrt{\frac{2\,C_e\,V}{n\,h}}\sqrt{1+\log\left(\frac{n\,h^2}{V}\vee 1\right)}\end{aligned}$$

Note that $C_e \leq \frac{2}{e_m}$. Finally, we get the desired result:

$$\varepsilon_*^2 \leq 4\,K^2\,\frac{V}{n\,h\,e_m}\left[1+\log\left(\frac{n\,h^2}{V}\vee 1\right)\right]\ .$$

■

A.3.2 Second Case

We now consider Equation 29 where w is given by Equation 22. Note that the logarithmic term is necessarily 0. If we assume that the solution ε_* of Equation 29 satisfies $\varepsilon_* \geq \sqrt{h'}$, then $w(x) = \varepsilon_* \sqrt{\frac{2\,C_e}{h'}}$. We obtain:

$$\varepsilon_*^2 \leq 4\,K^2 \sqrt{\frac{V}{n\,e_m}} \ .$$

Else, $\varepsilon_* \leq \sqrt{h'}$ which implies that

$$\varepsilon_*^2 \leq h' = \sqrt{\frac{V}{n\,e_m}} \ .$$

Both bounds provide the same convergence rate.

Paragraphs A.3.1 and A.3.2 together complete the proof of Theorem 1.

■

Appendix B. Proof of Minimax Lower Bounds

We remind the reader that the minimax risk is defined as:

$$\mathcal{R}\left(\mathcal{G}, h\right) = \inf_{\widehat{g} \in \mathcal{G}} \left[\sup_{\mathbb{P} \in \mathcal{P}(\mathcal{G}, h)} \mathbb{E}\left[\ell\left(\widehat{g}, g^*\right)\right] \right] \ .$$

The lower bound on the minimax risk is proved in Subsection B.1 for the SCAR assumption (cf. Theorem 2) and in Subsection B.2 for the SAR assumption (cf. Proposition 3).

B.1 Under the SCAR Assumption (Proof of Theorem 2)

The proof consists in exhibiting a finite subset of probability distributions on which the excess risk is worst. It is organized as follows: (i) we provide a lower bound on the minimax risk expression by restricting ourselves to this subset of distributions; (ii) we use Massart margin condition and simplify the remaining expression; (iii) the application of Assouad lemma finishes the proof.

(i) We start by introducing a family of probability distributions that plainly exploits the margin condition (A_2). Let $x_1, ..., x_V$ be V points of $\mathbb{R}^d$ shattered by $\mathcal{G}$. This is possible because the VC dimension of $\mathcal{G}$ is V. For some parameter $p < \frac{1}{V-1}$, we define a discrete probability distribution on $\{x_1, ..., x_V\} \subset \mathbb{R}^d$ verifying:

$$\mathbb{P}\left(X = x_i\right) = p \ \ \forall\, i \leq V-1 \text{ and } \mathbb{P}\left(X = x_V\right) = 1 - p\ (V-1) \ .$$

For some binary vector $b \in \{0,1\}^{V-1}$, we consider $\mathbb{P}_b$ the probability distribution such that:

$$\forall\, 1 \leq i \leq V-1,\ \mathbb{P}_b\left(Y = 1 \,|\, X = x_i\right) = \frac{1}{2}\left[1 + \left(2\,b_i - 1\right)\,h\right]$$

for $h > 0$. We can consider by default that each point in $\mathbb{R}^d \setminus \{x_1, ..., x_{V-1}\}$ has class 0 almost surely. This has no incidence on the rest of the proof. Moreover:

$$\mathbb{P}_b\left(S = 1 \mid X = x_i, Y = y\right) = y\, e(x_i)$$

following the definition of propensity.

Hence, $(\mathbb{P}_b)_{b \in \{0,1\}^{V-1}}$ defines a family of distributions on (X, S) that satisfies Massart margin condition (A_2) at its limit: the regression function $|2\eta(x_i) - 1|$ equals h for every $i \in \{1, ...V-1\}$. Furthermore, for every $b \in \{0,1\}^{V-1}$, the Bayes classifier g_b^* is known:

$$\forall\, 1 \leq i \leq V-1,\ g_b^*(x_i) = b_i\ .$$

As $(x_1, ..., x_V)$ is shattered by $\mathcal{G}$, g_b^* necessarily belongs to $\mathcal{G}$.
Hence, $(\mathbb{P}_b)_{b \in \{0,1\}^{V-1}} \subset \mathcal{P}\left(\mathcal{G}, h\right)$ and therefore:

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \inf_{\widehat{g} \in \mathcal{G}} \left[\sup_{b \in \{0,1\}^{V-1}} \mathbb{E}_b\left[\ell\left(\widehat{g}, g_b^*\right)\right] \right]$$

where $\mathbb{E}_b$ denotes the expectation according to $\mathbb{P}_b$ distribution.

(ii) Let $\widehat{g}$ be a classifier, function of the training sample $(X_i, S_i)_{1 \leq i \leq n}$. We use the following decomposition of ℓ (cf. Equation 23):

$$\ell(\widehat{g}, g_b^*) = \mathbb{E}\left[|2\eta(X) - 1|\, |\widehat{g}(X) - g_b^*(X)|\right]\ .$$

Combined with Massart margin condition (A_2), this yields:

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq h \inf_{\widehat{g} \in \mathcal{G}} \left[\sup_{b \in \{0,1\}^{V-1}} \mathbb{E}_b\left[|\widehat{g}(X) - g_b^*(X)|\right] \right]$$

For every $\widehat{g}$, we define $\widehat{b}$ such that:

$$\widehat{b} = \operatorname*{Argmin}_{b \in \{0,1\}^{V-1}} \mathbb{E}_X\left[|g_b^*(X) - \widehat{g}(X)|\right]$$

where the expectation is taken with respect to the marginal distribution of X and conditionally to the training sample. Hence, $\widehat{b}$ is a function of the training sample $(X_i, S_i)_{1 \leq i \leq n}$. By triangular inequality and then by definition of $\widehat{b}$:

$$\left|g_{\widehat{b}}^*(X) - g_b^*(X)\right| \leq \left|g_{\widehat{b}}^*(X) - \widehat{g}(X)\right| + |\widehat{g}(X) - g_b^*(X)| \leq 2\ |\widehat{g}(X) - g_b^*(X)|\ .$$

Hence:

$$\begin{aligned}
\mathcal{R}\left(\mathcal{G}, h\right) &\geq \frac{h}{2} \inf_{\widehat{g} \in \mathcal{G}} \left[\sup_{b \in \{0,1\}^{V-1}} \mathbb{E}_b\left[\left|g_{\widehat{b}}^*(X) - g_b^*(X)\right|\right] \right] \\
&= \frac{h}{2} \inf_{\widehat{b} \in \{0,1\}^{V-1}} \left[\sup_{b \in \{0,1\}^{V-1}} \mathbb{E}_b\left[\left|g_{\widehat{b}}^*(X) - g_b^*(X)\right|\right] \right] \\
&= \frac{p\, h}{2} \inf_{\widehat{b} \in \{0,1\}^{V-1}} \left[\sup_{b \in \{0,1\}^{V-1}} \mathbb{E}_b\left[\sum_{i=1}^{V-1} \mathbb{1}_{b_i \neq \widehat{b}_i} \right] \right]
\end{aligned}$$

where the last line is obtained by developing the expectation according to the marginal distribution of X, which is discrete.

(iii) With this simplified expression, we apply Assouad lemma (cf. Yu, 1997) which provides the following general lower bound:

$$\inf_{\widehat{b}\in\{0,1\}^{V-1}} \left[\sup_{b\in\{0,1\}^{V-1}} \mathbb{E}_b \left[\sum_{i=1}^{V-1} \mathbb{1}_{b_i\neq\widehat{b}_i} \right]\right] \geq \frac{V-1}{2} \left(1-\sqrt{\gamma\, n}\right)$$

where γ is an upper bound on the square Hellinger distance between probability distributions $\mathbb{P}_b$ and $\mathbb{P}'_b$ on (X, S) when b and b' only differ on one coordinate. Using technical Lemma 13 in Appendix D, we have the following upper bound on the square Hellinger distance $\mathcal{H}^2\left(\mathbb{P}_b, \mathbb{P}_{b'}\right)$:

$$\mathcal{H}^2\left(\mathbb{P}_b, \mathbb{P}_{b'}\right) \leq 2\, p\, e_m\, h^2 \ . \tag{30}$$

Applying Assouad lemma together with Equation 30, we get the following inequality:

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \frac{p\, h}{4} \left(V-1\right) \left(1-\sqrt{2\, p\, e_m\, h^2\, n}\right) \ .$$

In case (C_1), we choose $p = \frac{2}{9\, e_m\, h^2\, n}$ that is lower than $\frac{1}{V-1}$, we obtain the desired lower bound on $\mathcal{R}\left(\mathcal{G}, h\right)$:

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \frac{V-1}{54\, e_m\, h\, n} \ .$$

Else, in case (C_2), we choose $p = \frac{2}{9\, e_m\, h'^2\, n}$ where we recall that $h' = \sqrt{\frac{V}{n\, e_m}}$. As $h \leq h'$:

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \mathcal{R}\left(\mathcal{G}, h'\right) \geq \frac{V-1}{54\, e_m\, h'\, n} \geq \frac{1}{54\sqrt{2}} \sqrt{\frac{V-1}{n\, e_m}} \ .$$

∎

B.2 Proof of Proposition 3

This proof relies on the same tools as SCAR assumption case. We alter (i) by choosing $x_1, ..., x_V$ satisfying assumption (A_3) for $\varepsilon > 0$. (ii) remains unchanged. In (iii), the upper bound in Equation 30 has to be replaced by $2\, p\, h^2\, \left(e_m + \varepsilon\right)$. This yields the following lower bounds:

1. in case (C_1):

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \frac{V-1}{54\, \left(e_m+\varepsilon\right)\, h\, n} \ ;$$

2. in case (C_2):

$$\mathcal{R}\left(\mathcal{G}, h\right) \geq \frac{1}{54\sqrt{2}} \sqrt{\frac{V-1}{\left(e_m+\varepsilon\right)\, h\, n}} \ .$$

It remains to note that these lower bounds are valid for any $\varepsilon > 0$ to complete the proof.

∎

Appendix C. Universal Entropy Metric and Related Properties

In this section, we recall some definitions and properties concerning the universal entropy metric. These properties are used for the proof of Proposition 7 in Appendix A.

Let us consider $(X_i, S_i)_{1\leq i\leq n}$ i.i.d. random variables with values in $\mathbb{R}^d \times \{0,1\}$ and $\mathcal{F}$ a set of functions on $\mathbb{R}^d \times \{0,1\}$.

Definition 8 (Universal entropy metric, cf. Massart and Nédélec 2006)
Let $\varepsilon > 0$ and $\mathbb{Q}$ be a probability measure.
Define $h\left(\mathcal{F}, \varepsilon, \mathbb{Q}\right)$ as the logarithm of the largest number N of functions $f_1, ..., f_N$ separated by a distance ε, namely $\mathbb{E}_{\mathbb{Q}}\left[\left(f_i(X,S) - f_j(X,S)\right)^2\right] > \varepsilon^2$, $\forall i \neq j$.
Then the universal entropy metric $H\left(\mathcal{F}, \varepsilon\right)$ is defined as:

$$H\left(\mathcal{F}, \varepsilon\right) = \sup_{\mathbb{Q}} h\left(\mathcal{F}, \varepsilon, \mathbb{Q}\right) \ .$$

Proposition 9 *For a fixed $\mathcal{F}$, $H\left(\mathcal{F}, \cdot\right)$ is a decreasing function.*

Proposition 10 *Let ψ be a function defined on $\mathbb{R}^d \times \{0,1\}$ and $\mathcal{F}$ be a family of functions such that:*

$$\mathcal{F} = \left\{(x,s) \mapsto \psi(x,s)\, g(x,s) \ , \ g \in \mathcal{G}\right\}$$

where $\mathcal{G}$ is another family of functions on $\mathbb{R}^d \times \{0,1\}$. Then:

$$\forall \varepsilon > 0, \ H\left(\mathcal{F}, \varepsilon\right) \leq H\left(\mathcal{G}, \frac{\varepsilon}{||\psi||_\infty}\right) \ .$$

Proof Let $\mathbb{Q}$ be a probability distribution and N such that $h\left(\mathcal{G}, \frac{\varepsilon}{||\psi||_\infty}, \mathbb{Q}\right) < \log(N)$. Then, for any set of functions $g_1, ..., g_N$, there is $i \neq j$ such that $\mathbb{E}_{\mathbb{Q}}\left[\left(g_i(X,S) - g_j(X,S)\right)^2\right] \leq \left(\frac{\varepsilon}{||\psi||_\infty}\right)^2$. This implies that $\mathbb{E}_{\mathbb{Q}}\left[\left(\psi(X,S)\ \left[g_i(X,S) - g_j(X,S)\right]\right)^2\right] \leq \varepsilon^2$ and then that $h\left(\mathcal{F}, \varepsilon, \mathbb{Q}\right) < \log(N)$.
Then, we have that $h\left(\mathcal{F}, \varepsilon, \mathbb{Q}\right) \leq h\left(\mathcal{G}, \frac{\varepsilon}{||\psi||_\infty}, \mathbb{Q}\right)$. Considering the supremum over the probability distributions $\mathbb{Q}$, we obtain the desired result. ■

Finally, we recall Haussler bound, which provides an upper bound on the universal entropy metric of a set of classifiers in terms of its VC dimension.

Proposition 11 (Haussler bound, cf. Bousquet et al. 2003)
Assuming that $\mathcal{F}$ is a set of indicator functions with finite Vapnik dimension V. Then, $\forall \varepsilon > 0$:

$$H\left(\mathcal{F}, \varepsilon\right) \leq \kappa\, V\left(1 + \log\left(\varepsilon^{-1} \vee 1\right)\right)$$

where $\kappa \geq 1$ is an absolute constant.

Appendix D. Technical Lemmas

Lemma 12 *Let $C_e > 1$ and $\sigma > 0$. Then:*

$$\sum_{j=0}^{+\infty} 2^{-j}\sqrt{1+\log\left(2^{j+1}\frac{C_e}{\sigma}\vee 1\right)} \leq 2\,\left(1+\log(2)\right)\sqrt{1+\log\left(\frac{C_e}{\sigma}\vee 1\right)}$$

Proof

$$\begin{aligned}
\sum_{j=0}^{+\infty} 2^{-j}\sqrt{1+\log\left(2^{j+1}\frac{C_e}{\sigma}\vee 1\right)} &\leq \sum_{j=0}^{+\infty} 2^{-j}\sqrt{1+(j+1)\log(2)+\log\left(\frac{C_e}{\sigma}\vee 1\right)}\\
&\leq \sum_{j=0}^{+\infty} 2^{-j}\sqrt{1+(j+1)\log(2)}\sqrt{1+\log\left(\frac{C_e}{\sigma}\vee 1\right)}\\
&\leq \sum_{j=0}^{+\infty} 2^{-j}\left(1+(j+1)\frac{\log(2)}{2}\right)\sqrt{1+\log\left(\frac{C_e}{\sigma}\vee 1\right)}\\
&= 2\,\left(1+\log(2)\right)\sqrt{1+\log\left(\frac{C_e}{\sigma}\vee 1\right)}
\end{aligned}$$

■

Lemma 13 *Let $x_1, ..., x_V$ be vectors of $\mathbb{R}^d$. Let e be a function on R^d with values in $(0,1]$. Let $p \leq \frac{1}{V-1}$ and consider $(\mathbb{P}_b)_{b\in\{0,1\}^{V-1}}$ the family of probability distributions on $\{x_1, ..., x_V\} \times \{0,1\}$ defined in (i) (cf. Appendix B.1). If b and b' are binary vectors of $\{0,1\}^{V-1}$ which only differ at coordinate i, then:*

$$\mathcal{H}\left(\mathbb{P}_b, \mathbb{P}_{b'}\right) \leq 2\,p\,e(x_i)\,h^2\ .$$

Proof Recall that b and b' only differ at coordinate i, hence:

$$\begin{aligned}
\mathcal{H}^2\left(\mathbb{P}_b, \mathbb{P}_{b'}\right) &= \frac{1}{2}\sum_{j=1}^{V}\left(\sqrt{\mathbb{P}_b\left(X=x_j,\ S=1\right)}-\sqrt{\mathbb{P}_{b'}\left(X=x_j,\ S=1\right)}\right)^2\\
&+\frac{1}{2}\sum_{j=1}^{V}\left(\sqrt{\mathbb{P}_b\left(X=x_j,\ S=0\right)}-\sqrt{\mathbb{P}_{b'}\left(X=x_j,\ S=0\right)}\right)^2\\
&=\frac{1}{2}\left(\sqrt{\mathbb{P}_b\left(X=x_i,\ S=1\right)}-\sqrt{\mathbb{P}_{b'}\left(X=x_i,\ S=1\right)}\right)^2\\
&+\frac{1}{2}\left(\sqrt{\mathbb{P}_b\left(X=x_i,\ S=0\right)}-\sqrt{\mathbb{P}_{b'}\left(X=x_i,\ S=0\right)}\right)^2\ .
\end{aligned}$$

Let us now calculate the probabilities using the definition of $\mathbb{P}_b$:

$$\begin{aligned}
\mathbb{P}_b\left(X=x_i, S=1\right) &= p\,\frac{e(x_i)}{2}\left[1+(2b_i-1)\ h\right]\ ,\\
\mathbb{P}_b\left(X=x_i, S=0\right) &= p\,\left(1-\frac{e(x_i)}{2}\left[1+(2b_i-1)\ h\right]\right)\ .
\end{aligned}$$

Noting that either $(b_i, b'_i) = (0, 1)$ or $(b_i, b'_i) = (1, 0)$, we have in both cases:

$$\left(\sqrt{\mathbb{P}_b\left(X = x_i,\ S = 1\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_i,\ S = 1\right)}\right)^2 = p\, e(x_i)\, \left[1 - \sqrt{1 - h^2}\right] \ ,$$

and

$$\begin{aligned}
&\left(\sqrt{\mathbb{P}_b\left(X = x_i,\ S = 0\right)} - \sqrt{\mathbb{P}_{b'}\left(X = x_i,\ S = 0\right)}\right)^2 \\
&= p\left[2 - e(x_i) - 2\sqrt{1 - \frac{e(x_i)}{2}\left(1 + h\right)}\sqrt{1 - \frac{e(x_i)}{2}\left(1 - h\right)}\right] \ .
\end{aligned}$$

We then sum the two results together:

$$\begin{aligned}
\mathcal{H}^2\left(\mathbb{P}_b, \mathbb{P}_{b'}\right) &= \frac{p}{2}\left[2 - e(x_i)\sqrt{1 - h^2} - 2\sqrt{1 - e(x_i) + \frac{e(x_i)^2}{4}\left(1 - h^2\right)}\right] \\
&= p\left[1 - \frac{e(x_i)}{2}\sqrt{1 - h^2} - \sqrt{\left(1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}\right)^2 - e(x_i)\left(1 - \sqrt{1 - h^2}\right)}\right] \\
&= p\left[1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}\right]\left[1 - \sqrt{1 - \frac{e(x_i)\left(1 - \sqrt{1 - h^2}\right)}{\left[1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}\right]^2}}\right] \\
&\leq \frac{p\ e(x_i)\left(1 - \sqrt{1 - h^2}\right)}{1 - \frac{e(x_i)}{2}\sqrt{1 - h^2}} \\
&\leq 2\, p\, e(x_i)\, h^2
\end{aligned}$$

In the above calculation, we applied the inequality $1 - \sqrt{1 - h^2} \leq h^2$ for $h^2 \in [0, 1]$. ■

References

Jessa Bekker and Jesse Davis. Estimating the class prior in positive and unlabeled data through decision tree induction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Mach. Learn.*, 109(4):719–760, Apr 2020.

Jessa Bekker, Pieter Robberechts, and Jesse Davis. Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data. In *Machine Learning and Knowledge Discovery in Databases*, volume 11907, pages 71–85, 2020.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 11:2973–3009, 2010.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207, 2003.

Timothy I Cannings, Yingying Fan, and Richard J Samworth. Classification with imperfect training labels. *Biometrika*, 107(2):311–330, Apr 2020.

Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. Self-PU: Self boosted and calibrated positive-unlabeled training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1510–1519, Jul 2020.

Olivier Coudray, Philippe Bristiel, Miguel Dinis, Christine Keribin, and Patrick Pamphile. Fatigue Data-Based Design: statistical methods for the identification of critical zones. In *SIA Simulation Numérique*, April 2021.

Marthinus Christoffel Du Plessis and Masashi Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE Transactions on Information and Systems*, 97(5):1358–1362, 2014.

Marthinus Christoffel Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, 1:703–711, Jan 2014.

Marthinus Christoffel Du Plessis, Gang Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning*, pages 1386–1394. PMLR, 2015.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 213. ACM Press, 2008.

Edgardo Ferretti, Marcelo L. Errecalde, Maik Anderka, and Benno Stein. On the Use of Reliable-Negatives Selection Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia. *25th International Workshop on Database and Expert Systems Applications*, pages 211–215, Sep 2014.

Donato Hernández Fusilier, Manuel Montes-y-Gómez, Paolo Rosso, and Rafael Guzmán Cabrera. Detecting positive and negative deceptive opinions using pu-learning. *Information processing and management*, 51(4):433–443, 2015.

Chen Gong, Qizhou Wang, Tongliang Liu, Bo Han, Jane J You, Jian Yang, and Dacheng Tao. Instance-dependent positive and unlabeled learning with labeling bias estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Daojing He, Menghan Pan, Kai Hong, Yao Cheng, Sammy Chan, Xiaowen Liu, and Nadra Guizani. Fake Review Detection Based on PU Learning and Behavior Density. *IEEE Network*, 34(4):298–303, Feb 2020.

Fengxiang He, Tongliang Liu, Geoffrey I Webb, and Dacheng Tao. Instance-dependent pu learning by bayesian optimal relabeling. *arXiv:1808.02180*, 2018.

Shantanu Jain, Martha White, and Predrag Radivojac. Estimating the class prior and posterior from noisy positives and unlabeled data. *Advances in neural information processing systems*, 29:2693–2701, 2016.

Yufeng Jiang, E. Haihong, Meina Song, and Ken Zhang. Research and Application of Newborn Defects Prediction Based on Spark and PU-learning. *5th IEEE International Conference on Cloud Computing and Intelligence Systems*, pages 657–663, Nov 2018.

Huayi Li, Zhiyuan Chen, Bing Liu, Xiaokai Wei, and Jidong Shao. Spotting Fake Reviews via Collective Positive-Unlabeled Learning. *IEEE International Conference on Data Mining*, pages 899–904, Dec 2014.

Xiaoli Li and Bing Liu. Learning to classify texts using positive and unlabeled data. In *IJCAI*, volume 3, pages 587–592, 2003.

Bing Liu, Wee Sun Lee, Philip S. Yu, and Xiaoli Li. Partially supervised classification of text documents. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 387–394, 2002.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining*, pages 179–186. IEEE, 2003.

Gábor Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer, 2002.

Yuxuan Luo, Shaoyin Cheng, Chong Liu, and Fan Jiang. PU Learning in Payload-based Web Anomaly Detection. *Third International Conference on Security of Smart Cities, Industrial Control System and Communications*, pages 1–5, Oct 2018.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5), Oct 2006.

Fantine Mordelet and J-P Vert. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

Ozra Nikdelfaz and Saeed Jalili. Disease genes prediction by HMM based PU-learning using gene expression profiles. *J. Biomed. Inf.*, 81:102–111, May 2018.

Gang Niu, Marthinus Christoffel Du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. *Advances in neural information processing systems*, 29:1199–1207, 2016.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International conference on machine learning*, pages 2052–2060. PMLR, 2016.

Vladimir Vapnik. *The nature of statistical learning theory.* Springer science & business media, 1999.

Robin Vogel, Mastane Achab, Stéphan Clémençon, and Charles Tillier. Weighted empirical risk minimization: Sample selection bias correction based on importance sampling. In *Proceedings of ICMA 2020*, 2020.

Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, Aug 2012.

Peng Yang, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh, and See-Kiong Ng. Ensemble Positive Unlabeled Learning for Disease Gene Identification. *PLoS One*, 9(5):e97079, May 2014.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.