Convergence rates for Positive-Unlabeled learning under Selected At Random assumption: sensitivity analysis with respect to propensity

Olivier Coudray, Christine Keribin, Patrick Pamphile

Convergence rates for Positive-Unlabeled learning under Selected At Random assumption: sensitivity analysis with respect to propensity

Olivier COUDRAY[1,2], Christine KERIBIN[2], et Patrick PAMPHILE[2]

[1]Stellantis, Centre d'Expertise Métier et Région, Poissy, 78300, France
[2]Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, Orsay, 91405, France

March 18, 2022

Abstract

Positive-Unlabeled learning (PU learning) is a binary classification task where only a subset of positive instances are labeled. The objective is then to find the correct classifier using positive labeled instances and unlabeled instances that contain a mixture of positive and negative data. In this paper, we illustrate some recent results about the convergence rates of PU learning under the general Selected At Random assumption, meaning that the labeled instances are not assumed to be a representative sample of the positive instances. We show that the simulations support the theoretical results highlighting the two regimes of convergence. We finally extend the simulations by relaxing some assumptions.

Keywords: Semi-supervised classification, Label noise, PU learning.

Introduction

Classic binary classification is a supervised machine learning task in which, from i.i.d. training observations $(X_i, Y_i)_{1\leq i\leq n}$ in $\mathbb{R}^d \times \{0,1\}$ with given classes (positive, $Y_i = 1$ or negative, $Y_i = 0$), one seeks to predict the class of new data. However, in many realistic situations, the observed classes can be noisy. In this paper, we are interested in a completely asymmetric 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 semi-supervised classification setting is called *Positive-Unlabeled Learning* (PU learning). PU learning is used in many practical situations: for instance in automatic diagnosis [CCC+20], spam review detection [LCL+14], gene disease identification [YLM+12] and anomaly detection [FEAS14]. This work was motivated by the fatigue design of structures in mechanics where testing can assert the presence of design flaws on a mechanical part, but cannot prove its absence [CBD+21]. In this paper, we only focus on methodological aspects.

In PU learning, the classes $(Y_i)_{1\leq i\leq n}$ are not observed. Instead we have access to $(S_i)_{1\leq i\leq n}$ where S_i indicates whether or not instance i is labeled. Of course, labeled instances ($S = 1$) are always positive instances and negative instances are never labeled:

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

The probability for a positive instance to be labeled is called the propensity (cf. [BRD20]) and is denoted e:

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

The SCAR assumption assumes a propensity independent of x, namely $e(x) = e$ (Selected Completely At Random assumption, SCAR). Here we are interested in the general case where the propensity may depend on x (Selected At Random assumption, SAR). Therefore, the labeled instances are a biased sub-sample of positive instances.

In Section 1, we recall a recent result on general risk bounds for PU learning under Selected At Random assumption. This result highlights two convergence rates that we empirically illustrate in Section 2. Finally, in Section 3, we relax some assumptions and study empirically there impact on PU learning convergence rates.

1 Risk bounds for PU learning under SAR assumption

Let $\mathcal{F}$ be a class of classifiers, *i.e.* binary functions on $\mathbb{R}^d$. And for every $f \in \mathcal{F}$, let $R(f)$ be the missclassification risk of f:

$$R(f) = \mathbb{P}\left(f(X) \neq Y\right) = \mathbb{E}\left[\mathbb{1}_{f(X)\neq Y}\right] \ .$$

We assume that Bayes classifier f^*, the minimizer of the missclassification risk R, belongs to $\mathcal{F}$. The objective is to use the training dataset $(X_i, S_i)_{1\leq i\leq n}$ to estimate f^*. Note that we cannot use the classes $(Y_i)_{1\leq i\leq n}$ since they are not observed.

A natural idea to address PU learning would be to minimize an empirical risk as if there were no label noise:

$$\widehat{f}_{NT} \in \operatorname*{Argmin}_{f\in\mathcal{F}} \widehat{R}_{NT}(f)$$

where $\widehat{R}_{NT}(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f(X_i)\neq S_i}$.

The difficulty of PU learning is that this classic empirical risk minimization procedure fails to estimate the right classifier because the so-called non-traditional empirical risk $\widehat{R}_{NT}(f)$ is a biased estimate of $R(f)$. When the propensity is partially known, an unbiased empirical risk is available and thus can be minimized to retrieve the right classifier (cf. [DPNS14, BRD20]).

$$\widehat{f} \in \operatorname*{Argmin}_{f\in\mathcal{F}} \widehat{R}(f)$$

$$\widehat{R}(f) = \frac{1}{n}\sum_{i=1}^n \left[\frac{\mathbb{1}_{S_i=1}}{e(X_i)}\left(2\,\mathbb{1}_{f(X_i)\neq 1} - 1\right) + \mathbb{1}_{f(X_i)\neq 0}\right].$$

In addition, [CKMP22] recently provided a general upper bound on the generalization risk under the following conditions:

(*A*1) $\exists\, h > 0$ such that $\forall\, x \in \mathbb{R}^d$,

$$\left|2\,\mathbb{P}\left(Y = 1 \,|\, X = x\right) - 1\right| \geq h \ .$$

(*A*2) The propensity $e(\cdot)$ is lower bounded by $e_m > 0$.

(*A*3) $\mathcal{F}$ has Vapnik dimension $V < +\infty$.

Assumptions (*A*1) and (*A*2) are assumptions on the label noise controlling the difficulty of the PU learning task, (*A*3) controls the complexity of class $\mathcal{F}$.

Let $\ell\left(\widehat{f}, f^*\right)$ be the excess risk, by definition:

$$\ell\left(\widehat{f}, f^*\right) = R(\widehat{f}) - R(f^*)$$

Assuming (*A*1), (*A*2) and (*A*3) are satisfied and under measurability conditions that are not detailed here, there exists an absolute constant κ such that:

$$\mathbb{E}\left[\ell\left(\widehat{f}, f^*\right)\right] \leq \kappa\sqrt{\frac{V}{n\,e_m}} \ . \tag{1}$$

Besides, if h given by assumption (*A*1) is greater than $1/\sqrt{n\,e_m}$, then the upper bound can be improved:

$$\mathbb{E}\left[\ell\left(\widehat{f}, f^*\right)\right] \leq \kappa\left[\frac{V}{n\,e_m\,h}\left(1 + \log\left(\frac{n\,h^2}{V}\vee 1\right)\right)\right] \ . \tag{2}$$

This result shows that under some conditions on the label noise involving h and e_m, fast convergence rates up to $\mathcal{O}\left(1/(n\,e_m\,h)\right)$ can be achieved. Thus, it extends an already well known result showing that the convergence rates for empirical risk minimization in classification are at least in $\mathcal{O}\left(1/\sqrt{n}\right)$ but can reach $\mathcal{O}\left(1/n\right)$ if h given by Massart's noise assumption (*A*1) is high enough (cf. [MN06]). The bounds explicitly show how a small propensity can hamper the generalization performances of PU learning. Finally, note that under the SCAR assumption ($e(x) = e_m$), $n\,e_m$ is proportional to the expected number of labeled instances.

The upper bounds 1 and 2 are almost optimal in the minimax sense as [CKMP22] provides lower bounds on the minimax risk exactly matching 1 or matching 2 up to the logarithmic term when h is high enough.

2 Sensitivity analysis with respect to propensity

We now want to illustrate the bounds on PU learning empirical risk minimizers through numerical simulations. We first describe the simulation setting. Then, we show that using the PU empirical learning risk enables to estimate the right classifier when the naive non-traditional approach fails to do so. Finally, we empirically study the convergence rates, emphasizing how they are affected by the propensity.

2.1 Simulation setting

We consider examples of PU learning tasks in one dimension ($d = 1$). The covariates $(X_i)_{1\leq i\leq n}$ are drawn i.i.d. according to a centered normal distribution with unit variance. Then, the class Y_i corresponds to the sign of X_i with probability $\frac{1+h}{2}$ where h is a fixed real in $(0, 1)$. More formally, let $(U_i)_{1\leq i\leq n}$ be i.i.d. uniform random variables on $[0, 1]$ (independent from

$(X_i)_{1\leq i\leq n}$), for every i we define:

$$Y_i = \mathbb{1}_{X_i\geq 0,\ U_i\geq \frac{1-h}{2}} + \mathbb{1}_{X_i<0,\ U_i<\frac{1-h}{2}}$$

Under this setting, the Bayes classifier f^* is known explicitly:

$$f^*(x) = \mathbb{1}_{x\geq 0}\ .$$

Besides, the distribution on (X, Y) satisfies Massart's noise condition $(A1)$ with constant $h > 0$.

In order to generate the labels $(S_i)_{1\leq i\leq n}$, we consider two models of propensity:

1. constant propensity (SCAR assumption):

$$e(x) = e_m,\ \text{with } e_m > 0 \tag{3}$$

2. logistic propensity (SAR assumption):

$$e(x) = \max\left(e_m, \frac{1}{1+e^{x-1}}\right),\ \text{with } e_m > 0\ . \tag{4}$$

   Note that this propensity is lower bounded by $e_m > 0$ and thus respects assumption $(A2)$.

An example of simulation is shown in Figure 1. In these simulations, the objective is to use only the observations $(X_i, S_i)_{1\leq i\leq n}$ (cf. Fig. 1, right) to estimate the classifier.

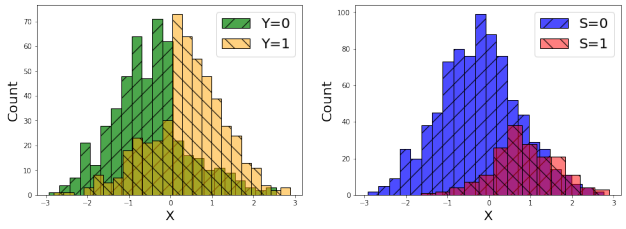

Figure 1: Example of simulation with $n = 1000$, $h = 0.5$ and logistic propensity ($e_m = 0.05$). On the left, the histograms of the positive and negative instances (true classes); on the right, the histograms of labeled and unlabeled instances (noisy labels). Note that there is a significant overlap between the distributions in both figures.

2.2 PU learning empirical risks

In this simulation setting, estimating the classifier is equivalent to identifying a threshold $m \in \mathbb{R}$ for the classification. Hence, we consider the following hypothesis space $\mathcal{F} = \{x \mapsto \mathbb{1}_{x\geq m},\ m \in \mathbb{R}\}$.

We recall that different empirical risks exist to approximate the true risk $R(f)$:

1. the traditional approach in standard binary classification is to compute the proportion of missclassified training instances:

$$\widehat{R}_T(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(X_i)\neq Y_i} \tag{5}$$

   which is inapplicable in PU learning context since the true classes are unobserved;

2. the non-traditional approach uses an analogous empirical risk by ignoring the label noise due to PU learning:

$$\widehat{R}_{NT}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{f(X_i)\neq S_i}\ ; \tag{6}$$

3. the unbiased empirical risk that accounts for the propensity:

$$\widehat{R}(f) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}_{S_i=1}}{e(X_i)}\left(2\mathbb{1}_{f(X_i)\neq 1} - 1\right) + \mathbb{1}_{f(X_i)\neq 0}\right]. \tag{7}$$

These three empirical risks are represented on Fig. 2 along with the true risk. Despite a higher variance, the PU learning unbiased empirical risk correctly estimates the true risk and can at least identify its minimum. Instead, the non-traditional empirical risk is clearly a biased estimate of the true risk and fails to identify the right classifier.

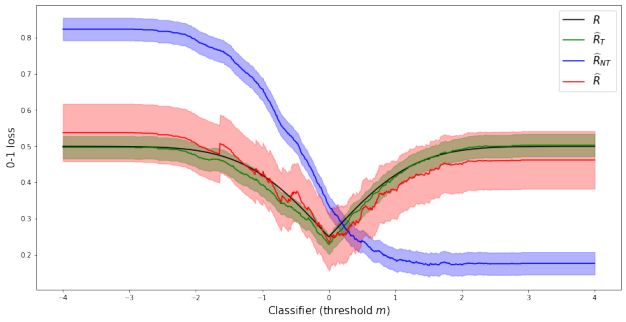

Figure 2: Comparison between the different empirical risk functions (cf. Equations 1, 6 and 7) evaluated on the simulation from Fig. 1. Abscissa represent the threshold m corresponding to the classifier $x \mapsto \mathbb{1}_{x\geq m}$. Around the curves are represented the 95% confidence intervals.

2.3 Convergence rates

We now want to illustrate numerically the rates of convergence of PU learning empirical risk minimizers

when the number of observations n and the minimum propensity e_m change. To do so, we repeat N times the following steps:

1. simulate a training set of size n with propensity $e(\cdot)$ (chosen following one of the models described in Equations 3 and 4)

2. estimate a classifier $\widehat{f}$ as a minimizer of PU learning empirical risk. Normally, minimizing such a risk is NP-hard. This is why restricting ourselves to $d = 1$ allows us to minimize this PU learning risk by performing a simple grid search over the set of possible thresholds $m \in \mathbb{R}$ for classification. Else, it is essential to resort to continuous and convex loss functions, which will be discussed in the next section.

3. evaluate the excess risk $\ell(\widehat{f}, f^*) = R\left(\widehat{f}\right) - R\left(f^*\right)$

We then estimate the mean excess risk by the empirical average over the N runs. Multiple experiments were realized with n ranging from 2^7 to 2^{16} and e_m ranging from 0.05 to 1. Massart's noise parameter is fixed for these experiments: $h = 0.25$. This value for parameter h was chosen low enough to allow both convergence regimes (Eq. 1 and 2) to be observable.

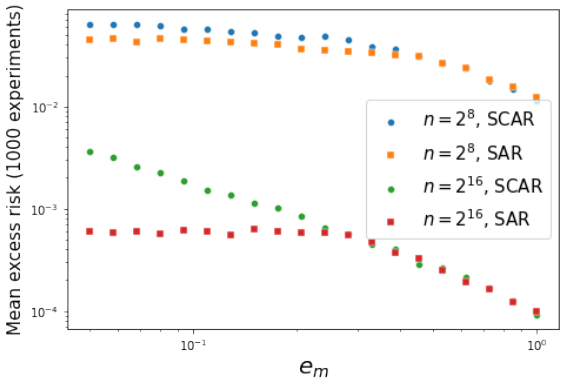

Figure 3: Mean excess risk as a function of e_m for n fixed ($n = 2^8$ and $n = 2^{16}$)

The results for both propensity models are presented in Fig. 3, 4 and 5, each on logarithmic scale. In Fig. 3, we clearly see that the mean excess risk decreases when e_m increases but the decrease happens faster when n is high. The performances under SAR assumption are always better than under SCAR assumption. This is due to the fact that, for a same value of e_m, there are more labeled instances in the SAR situation (because the propensity is generally greater than e_m) than

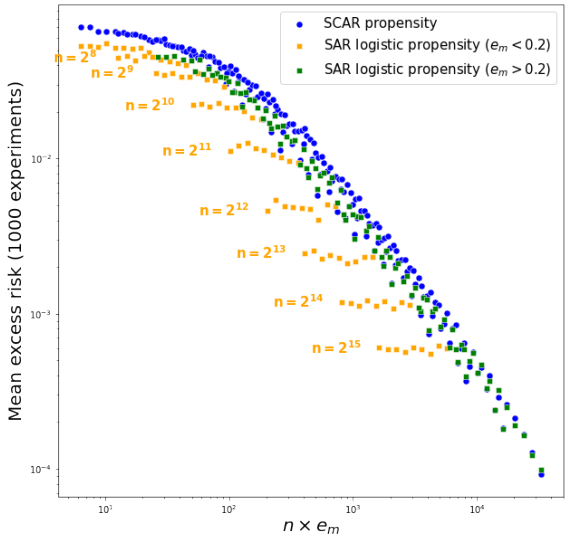

Figure 4: Mean excess risk for both propensity models (SCAR and logistic SAR) for different values of n and e_m. Experiments under logistic SAR model are split in two: in orange those for which $e_m \leq 0.2$, in green the rest. Aligned orange points correspond to equal values of n (cf. annotations)

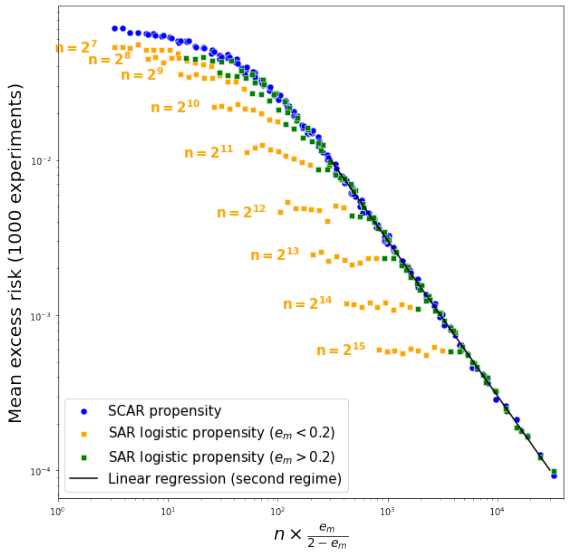

Figure 5: Estimated mean excess risk for both propensity models (SCAR and logistic SAR). Contrary to Fig. 4, abscissa corresponds to $n \times \frac{e_m}{2-e_m}$

in the SCAR situation where the propensity is always equal to its minimum e_m. A small value of e_m in the SAR propensity model does not alter much the propen-

sity function, we could even choose $e_m = 0$. This means that, in practice, we can allow the propensity to take arbitrary small values (hence violating assumption $(A2)$) as far as this occurs with small probability. When e_m is greater, the SAR propensity model behaves almost like the SCAR model because the effect of the maximum (cf. Eq. 4) is prominent.

Fig. 4 shows that the mean excess risk effectively depends on the term $n \times e_m$ which is closely related to the expected number of labeled instances. We clearly identify the different rates of convergence: fast when $n \times e_m$ is high, slower when $n \times e_m$ is low. The behaviour of the mean excess risk under SAR assumption confirms the observations of Fig. 3: when n is fixed, the performances remain almost identical for low values of e_m and follow SCAR propensity for higher values.

The results of mean excess risk as a function of $n \times e_m$ under SCAR assumption remain a bit scattered even if the general trend is well captured. Representing $n \times \frac{e_m}{2-e_m}$ in abscissa seems to better explain the observed results (cf. Fig. 5). In fact, looking closer at the theoretical result, [CKMP22] show in the proof that the risk upper bound depends on e_m through the term $\frac{2-e_m}{e_m}$ which was then upper bounded by $\frac{2}{e_m}$ in the final result. A linear regression is performed at the logarithmic scale on the results under SCAR assumption for $n\,e_m$ high enough. The estimated slope asserts the linear decrease in $\mathcal{O}\left(\frac{2-e_m}{n\,e_m}\right)$ of the excess risk.

3 Using tractable loss functions

In this section, we investigate the performances of PU learning using a continuous and convex loss function which is of course more suitable for applications.

The theoretical results of Section 1 are based on a procedure that consists in minimizing an empirical risk based on $0-1$ loss. If this framework is convenient to study theoretical properties of PU learning, it is not directly useful for applications because the minimization of $0-1$ loss requires solving difficult combinatorial optimization problems. It is thus natural to resort to convex loss functions instead. The use of convex loss functions adapted to PU learning was discussed in [DPNS14] under SCAR assumption. [BRD20] present a natural extension to SAR assumption.

Coming back to our simulation example ($d = 1$), we change the estimation of the classifier. Replacing the $0-1$ loss function by a logistic loss yields the following optimization problem:

$$\widehat{g} \in \underset{g \in \mathcal{G}}{\operatorname{Argmin}}\, \widehat{R}_C(g),$$

where $\mathcal{G} = \{x \mapsto x - m,\ m \in \mathbb{R}\}$ and where

$$\widehat{R}_C(g) = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\mathbb{1}_{S_i=1}}{e(X_i)}g(X_i) + \log\left(1 + e^{g(X_i)}\right)\right]. \quad (8)$$

The corresponding classifier is $\widehat{f}(x) = \mathbb{1}_{\widehat{g}(x)\geq 0}$. This time the loss function (as a function of $m \in \mathbb{R}$) is continuous, convex and one can check that it remains an unbiased estimate of the logistic risk:

$$\mathbb{E}\left[\widehat{R}_C(g)\right] = \mathbb{E}\left[-Y g(X) + \log\left(1 + e^{g(X)}\right)\right]. \quad (9)$$

As in Subsection 2.2, we can check that the PU learning empirical risk provides a good estimate of the true one contrary to the non-traditional risk, *i.e.* ignoring the label noise (cf. Fig. 6).

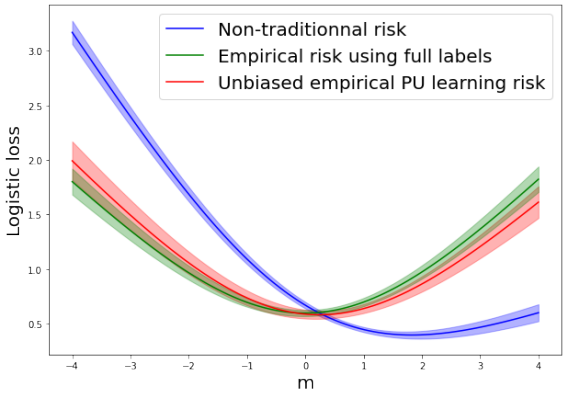

Figure 6: Comparison between the different loss functions: in green the traditional logistic loss function using the true classes $(Y_i)_{1\leq i\leq n}$, in blue the non-traditional logistic loss function ignoring the label noise, in red the logistic loss adapted to PU learning (cf. Eq. 8). Around the curves are represented the 95% confidence intervals.

We perform similar experiments as in subsection 2.3, this time using the logistic loss function to estimate the classifier. We study the mean excess risk under both propensity models (cf. Fig. 7). The numerical results confirm, at least for the SCAR propensity model, that the mean excess risk depends on $\frac{n\,e_m}{2-e_m}$. Unless, this time, the rate of convergence is parametric. With a linear regression, we find a slope close to $-\frac{1}{2}$ which suggests a decrease of the mean excess risk at the parametric rate. This is not surprising as the logistic regression is a parametric model optimized through maximum of likelihood, it is then normal that the convergence of the excess risk is bounded by the parametric rate.

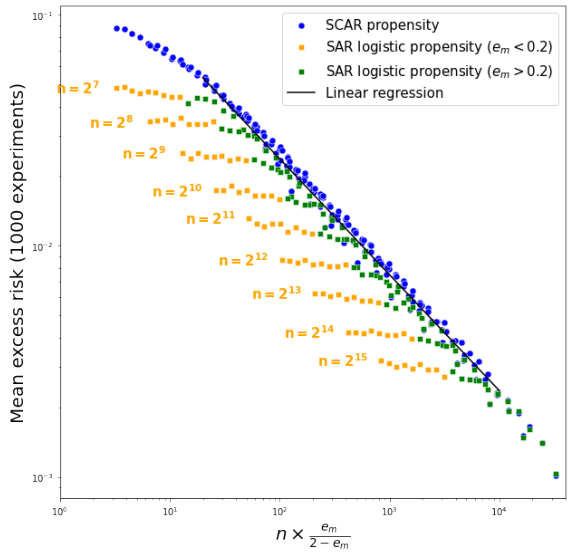

Figure 7: Mean excess risk as a function of $\frac{2-e_m}{n\,e_m}$. A linear regression on the experiments under the SCAR assumption allows to estimate a slope close to $\frac{1}{2}$ which asserts a convergence rate in $\mathcal{O}\left(\sqrt{\frac{n\,e_m}{2-e_m}}\right)$.

Conclusion

In this paper, we provided an empirical study of convergence rates for PU learning under SAR assumption. The simulations illustrate the theoretical convergence rates, highlighting the two regimes. Besides, we extended our experiments to tractable loss functions that are suitable for applications. In this case, we observed a parametric convergence rate on the mean excess risk. Future work could investigate what happens when the propensity is not assumed to be partially known like here, and thus has to be estimated.

Acknowledgements

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.

References

[BRD20] 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.

[CBD+21] 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*, Online, France, April 2021.

[CCC+20] 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.

[CKMP22] Olivier Coudray, Christine Keribin, Pascal Massart, and Patrick Pamphile. Risk bounds for PU learning under Selected At Random assumption. preprint, January 2022.

[DPNS14] M. C. Du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, 1:703–711, Jan 2014.

[FEAS14] 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.

[LCL+14] 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.

[MN06] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Annals of Statistics*, 34(5), Oct 2006.

[YLM+12] 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.