HAL open science

# Convergence rates for PU learning under the SAR assumption: influence of propensity

Olivier Coudray, Christine Keribin, Patrick Pamphile

▶ To cite this version:

Olivier Coudray, Christine Keribin, Patrick Pamphile. Convergence rates for PU learning under the SAR assumption: influence of propensity. CAp&RFIAP 2022 - Conférence sur l'Apprentissage automatique, Jul 2022, Vannes, France. hal-03738282

HAL Id: hal-03738282
https://inria.hal.science/hal-03738282

Submitted on 25 Jul 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Convergence rates for PU learning under the SAR assumption: influence of propensity


Olivier COUDRAY [1,2] Christine KERIBIN [2] 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
**OpenLab AI between Stellantis and Inria**


## From standard classification to Positive Unlabeled (PU) classification setting

**General setting:** $(X_i, Y_i, S_i)_{1\le i\le n}$ **i.i.d.**

- $X_i \in \mathbb{R}^d$ *covariate vector*
- $Y_i \in \{0, 1\}$ *class* of interest
- $S_i \in \{0, 1\}$ observed *label*

**Classification of $Y$ given $X$:**

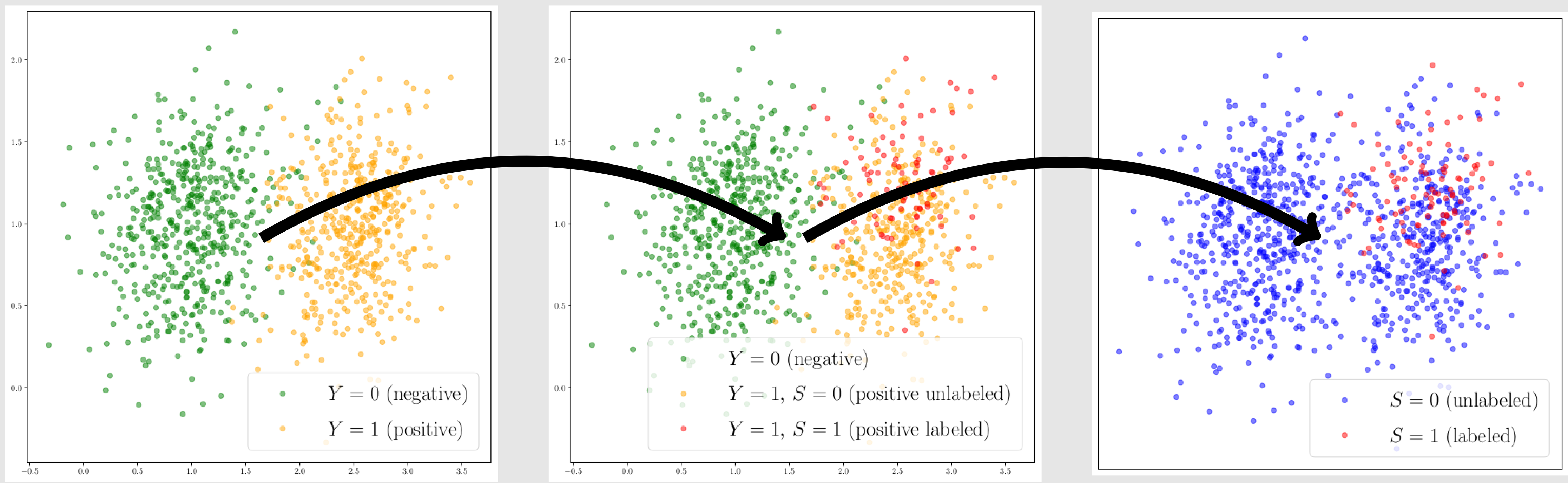
$$\mathbb{P}\left(Y = 1 \,|\, X = x\right) = \eta(x) \;.$$

**Propensity or selection bias:**

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

**Biased classification of $S$ given $X$:**

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




**Standard classification setting** $(X_i, Y_i) \to$ Estimate $\eta$

*Selection of positive instances through $e$*

**PU classification setting** $(X_i, S_i) \to$ Estimate $\eta$

## Estimating the classifier in PU learning setting

**Objective:** given a loss function $\boldsymbol{L}$, find the classification rule $f : \mathbb{R}^d \to \{0, 1\}$ minimizing the associated risk $R(f) = \mathbb{E}\left[L\left(f(X), Y\right)\right]$.

$$f^* \in \underset{f}{\operatorname{Argmin}}\, R(f) \;.$$

**Estimation of $f^*$ using the training sample** $(X_i, S_i)_{1\le i\le n}$: find $f : \mathbb{R}^d \to \{0, 1\}$ in a predefined class $\mathcal{F}$ minimizing an empirical risk.

**Empirical risks:**

1. Standard empirical risk: unavailable because the classes $(Y_i)_{1\le i\le n}$ are unobserved...

$$\widehat{R}_S(f) = \frac{1}{n}\sum_{i=1}^{n} L\left(f(X_i), Y_i\right)$$

2. Non traditional empirical risk: ignoring the propensity $e$.

$$\widehat{R}_{NT}(f) = \frac{1}{n}\sum_{i=1}^{n} L\left(f(X_i), S_i\right)$$

$\hookrightarrow$ biased estimate of the true risk.

3. Unbiased PU empirical risk[1],[2]: assuming the propensity scores $e(X_i)$ of labeled observations to be known.

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

## Theoretical risk bounds for PU learning[3]

**PU learning classifier using $0-1$ loss:** $L\left(f(x), y\right) = \mathbb{1}_{f(x)\neq y}$.

$$\widehat{f}_{PU} = \underset{f\in\mathcal{F}}{\operatorname{Argmin}}\, \widehat{R}_{PU}(f) \;.$$

**Excess risk:**

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

**Assumptions:**

- $\mathcal{F}$ has finite VC dimension $V$ (complexity)
- $f^* \in \mathcal{F}$
- $\exists\, h > 0$ such that $\forall x \in \mathbb{R}^d$, $|2\,\mathbb{P}\left(Y = 1 \,|\, X = x\right) - 1| \ge h$ (Massart noise assumption[4]).
- $\exists\, e_m > 0$, $\forall x \in \mathbb{R}^d\; e(x) \ge e_m$.

**Result: two convergence rates**

$$\mathbb{E}\left[\ell\left(\widehat{f}, f^*\right)\right] \le \kappa\sqrt{\frac{V}{n\, e_m}} \qquad \text{if } h \le \sqrt{\frac{V}{n\, e_m}} \;;$$

$$\le \kappa\left[\frac{V\,(2-e_m)}{n\, e_m\, h}\left(1 + \log\left(\frac{n\, h^2}{V} \vee 1\right)\right)\right] \qquad \text{if } h \ge \sqrt{\frac{V}{n\, e_m}} \;.$$

## Simulation setting

- One-dimensional setting ($d = 1$)
- $\mathcal{F} = \{x \mapsto \mathbb{1}_{x\ge m},\; m \in \mathbb{R}\}$
- $X_i \sim \mathcal{N}(0, 1)$
- $Y_i \sim \mathcal{B}\left(\frac{1+h}{2}\mathbb{1}_{X_i\ge 0} + \frac{1-h}{2}\mathbb{1}_{X_i<0}\right)$, $h = 0.25$.
- $S_i \sim \mathcal{B}\left(e(x)\right)$ if $Y_i = 0$ (else $S_i = 0$)
- Propensity functions:
  1. SCAR constant propensity,
  $$e(x) = e_m, \text{ with } e_m > 0 \;;$$
  2. SAR logistic propensity (cf. Fig. 1),
  $$e(x) = \max\left(e_m, \frac{1}{1 + e^{x-1}}\right), \text{ with } e_m > 0 \;.$$

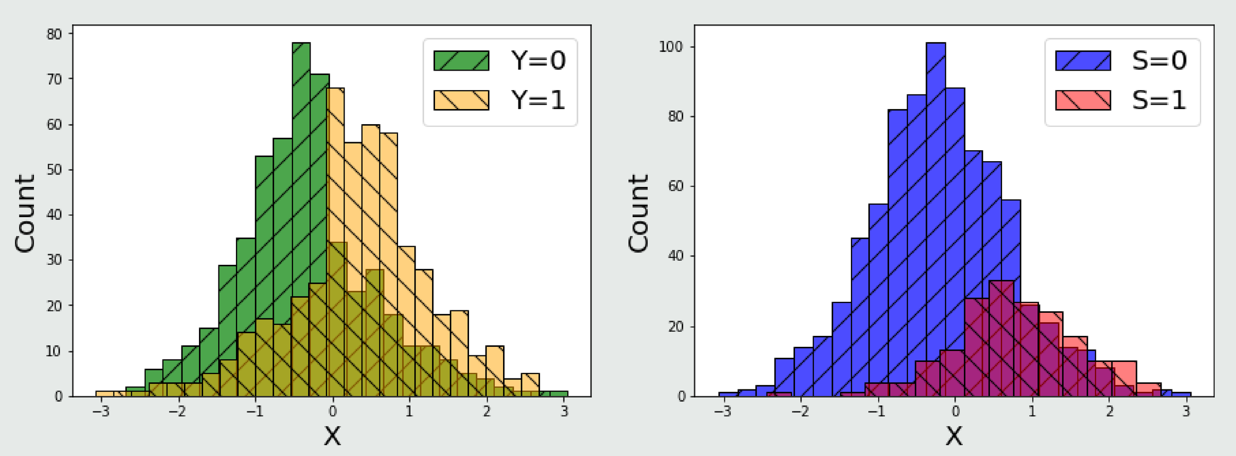
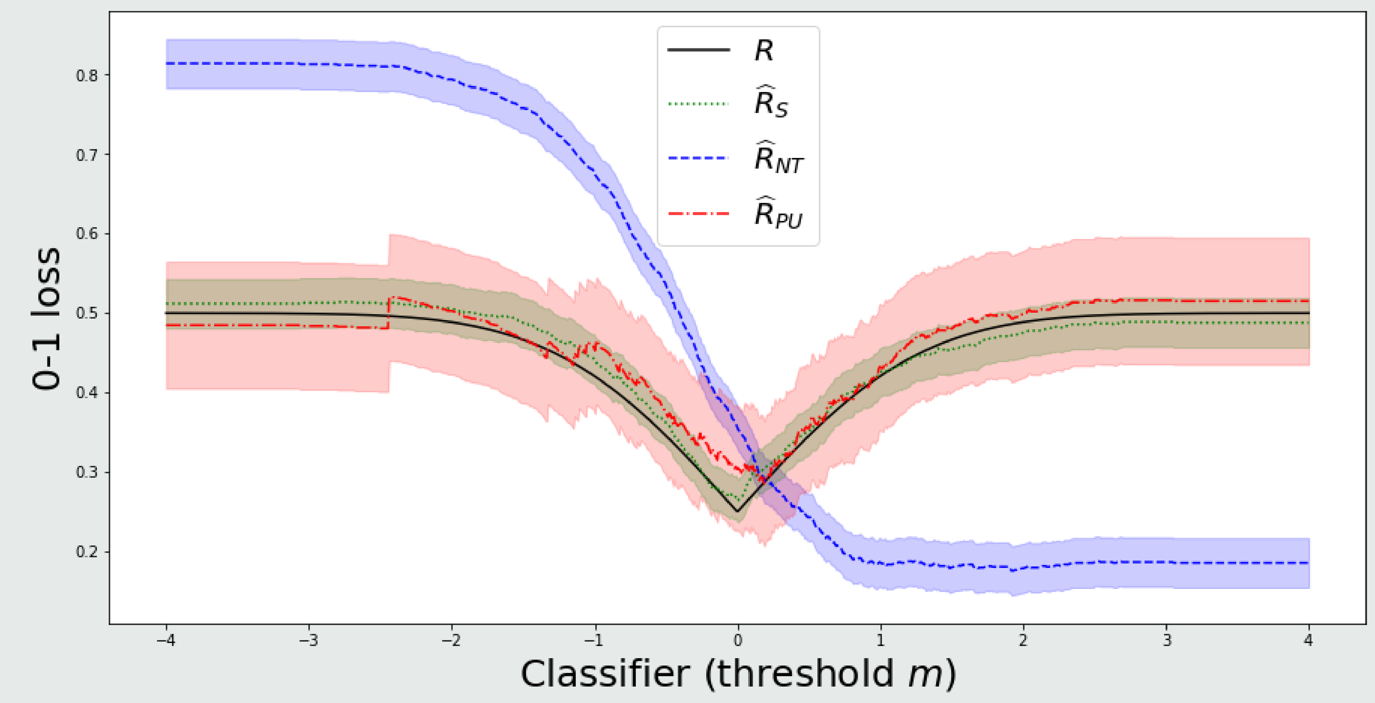



Figure 1. Simulation example (SAR)

## Comparison of the empirical risk functions




**Conclusion**

- The non-traditional loss $\widehat{R}_{NT}$ is a biased estimate of the true risk $R$
  $\hookrightarrow$ fails to provide a good estimate of the optimal classifier $f^*$.
- Despite a higher variance, the PU loss function $\widehat{R}_{PU}$ provides a better estimate of the true risk.

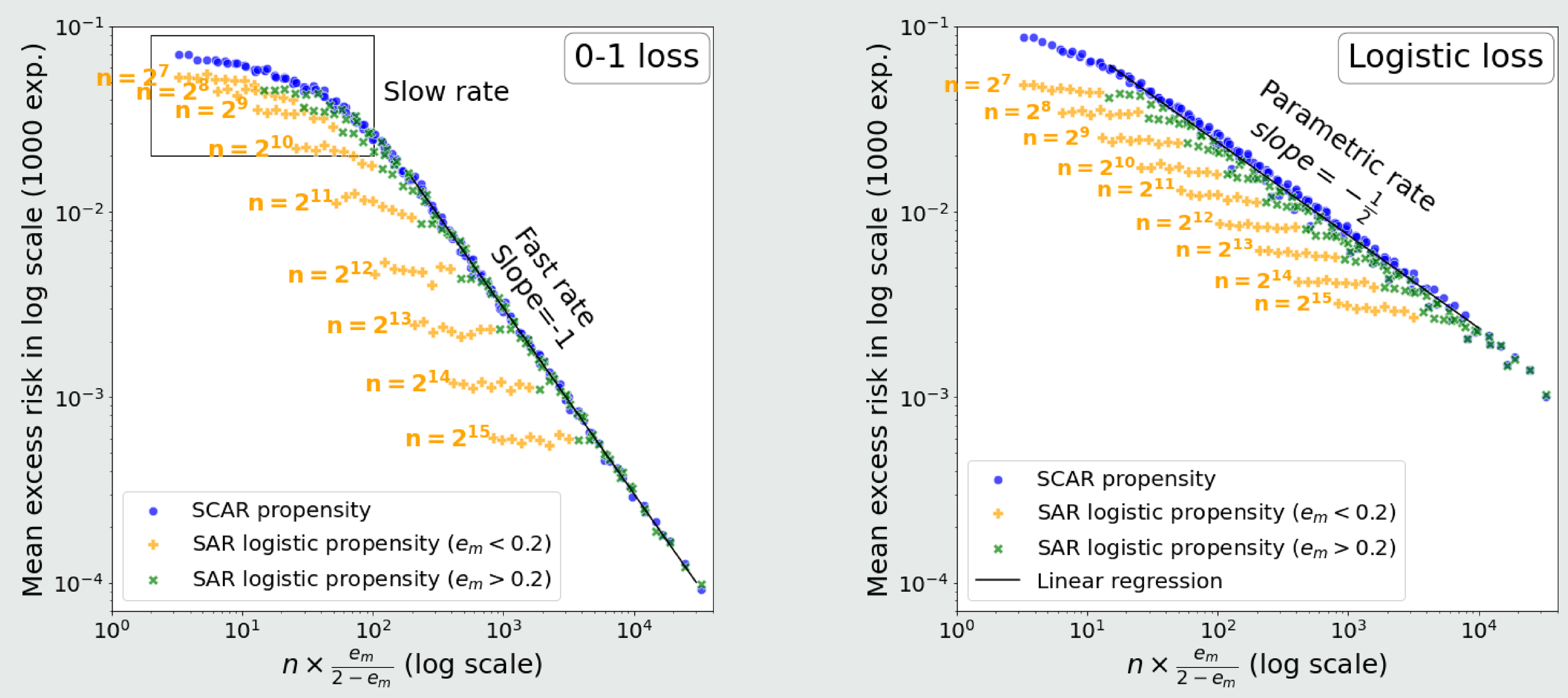
## Convergence rates: simulation study

**Estimation of the mean excess risk for:**

- $n$ ranging from $100$ to $30{,}000$
- $e$ constant propensity or logistic propensity
- $e_m$ ranging from $0.05$ to $1$.

**Two series of experiments:**

1. using the PU unbiased $0-1$ loss function
2. using a PU unbiased logistic loss function.




**Conclusion**

- Illustration of theoretical risk bounds for PU learning under $0-1$ loss:
  1. slow rate for $n\,e_m/\left(2 - e_m\right) < 100$
  2. fast rate for $n\,e_m/\left(2 - e_m\right) \ge 100$: empirical rate in $\mathcal{O}\left(1/\left(n\,e_m\right)\right)$ matching the theory
- Extension to a tractable logistic loss highlighting a unique parametric convergence rate $\mathcal{O}\left(1/\sqrt{n\,e_m}\right)$

## Perspectives

**Ongoing / future work**

- Dealing with an unknown propensity: joint estimation of the classifier $\eta$ and the propensity $e$.
- Practical applications: fatigue design of mechanical parts.
- Extension of the theoretical study to convex loss functions.

## Main references


[1] J. Bekker, P. Robberechts, and J. Davis. "Beyond the Selected Completely at Random Assumption for Learning from Positive and Unlabeled Data". In: *Machine Learning and Knowledge Discovery in Databases*. Vol. 11907. 2020, pp. 71–85.

[2] M. C. Du Plessis, G. Niu, and M. Sugiyama. "Analysis of learning from positive and unlabeled data". In: *NIPS* 1 (Jan. 2014), pp. 703–711.

[3] O. Coudray, C. Keribin, P. Massart, and P. Pamphile. "Risk bounds for PU learning under Selected At Random assumption". Preprint. Jan. 2022. URL: https://hal.inria.fr/hal-03526889.

[4] P. Massart and É. Nédélec. "Risk bounds for statistical learning". In: *Annals of Statistics* 34.5 (Oct. 2006).




olivier.coudray@universite-paris-saclay.fr