



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

From standard classification to Positive Unlabeled (PU) classification setting

General setting: $(X_i, Y_i, S_i)_{1 \leq i \leq n}$ 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 :

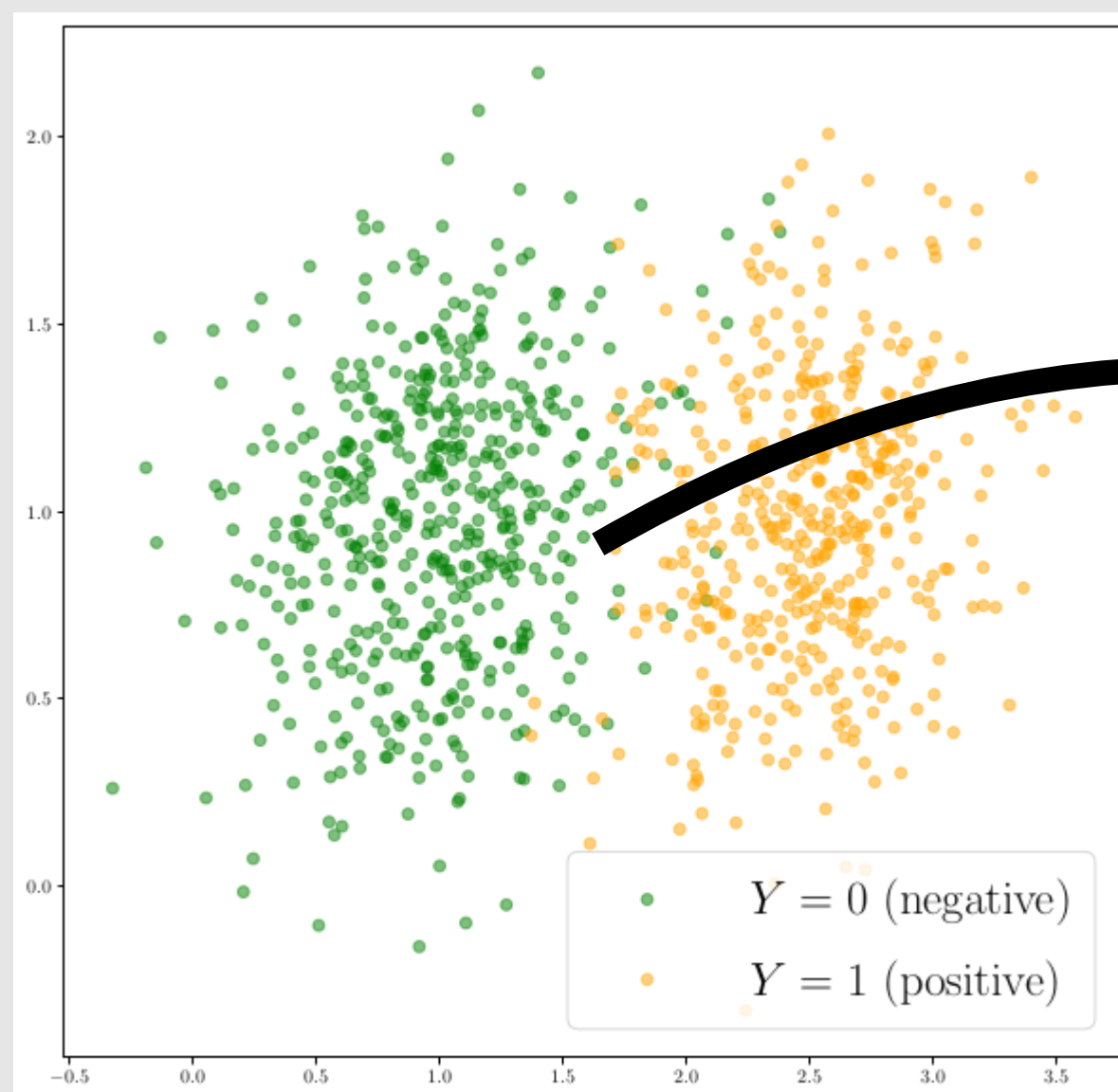
$$\mathbb{P}(Y = 1 | X = x) = \eta(x).$$

Propensity or selection bias:

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

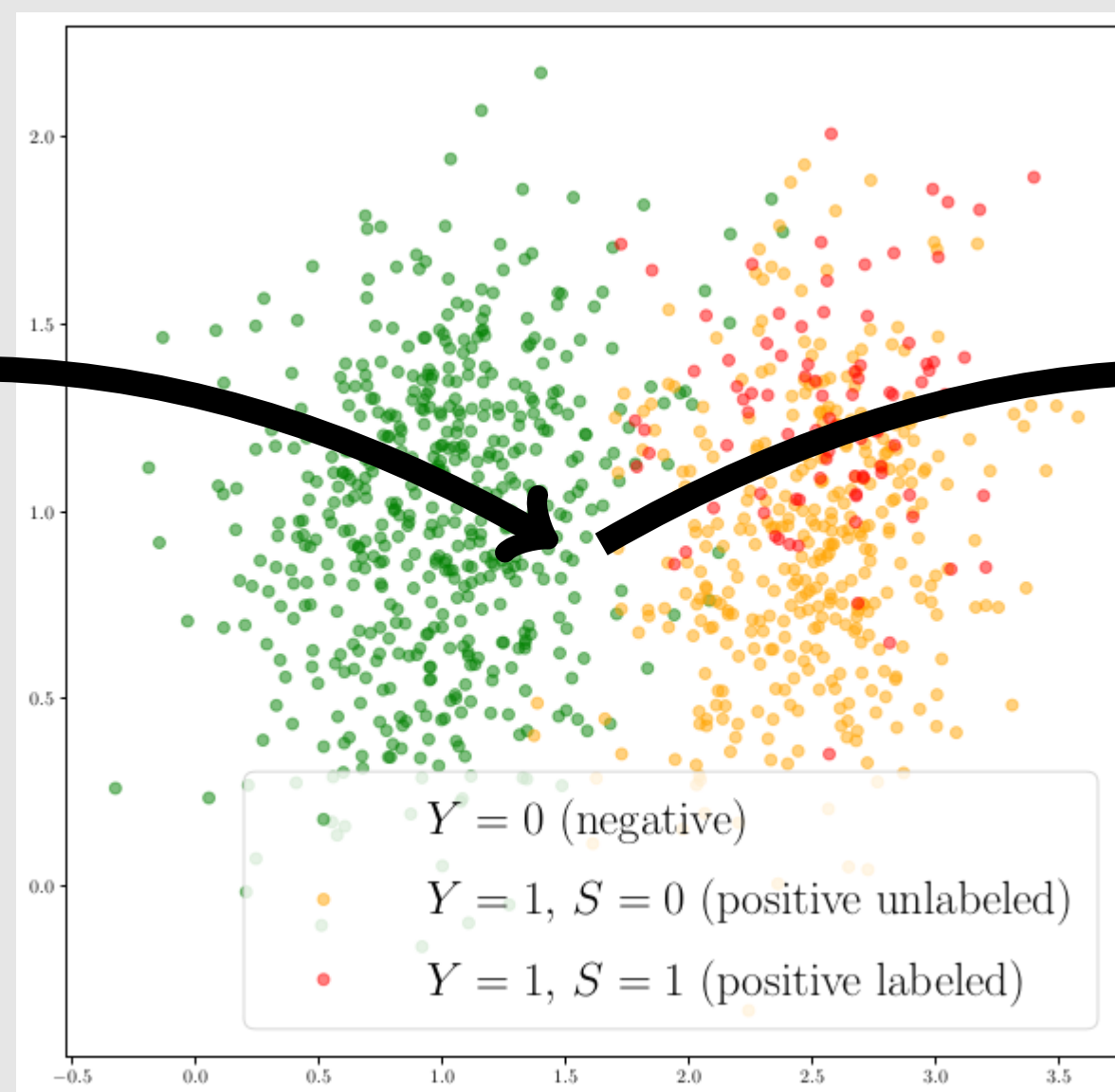
Biased classification of S given X :

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

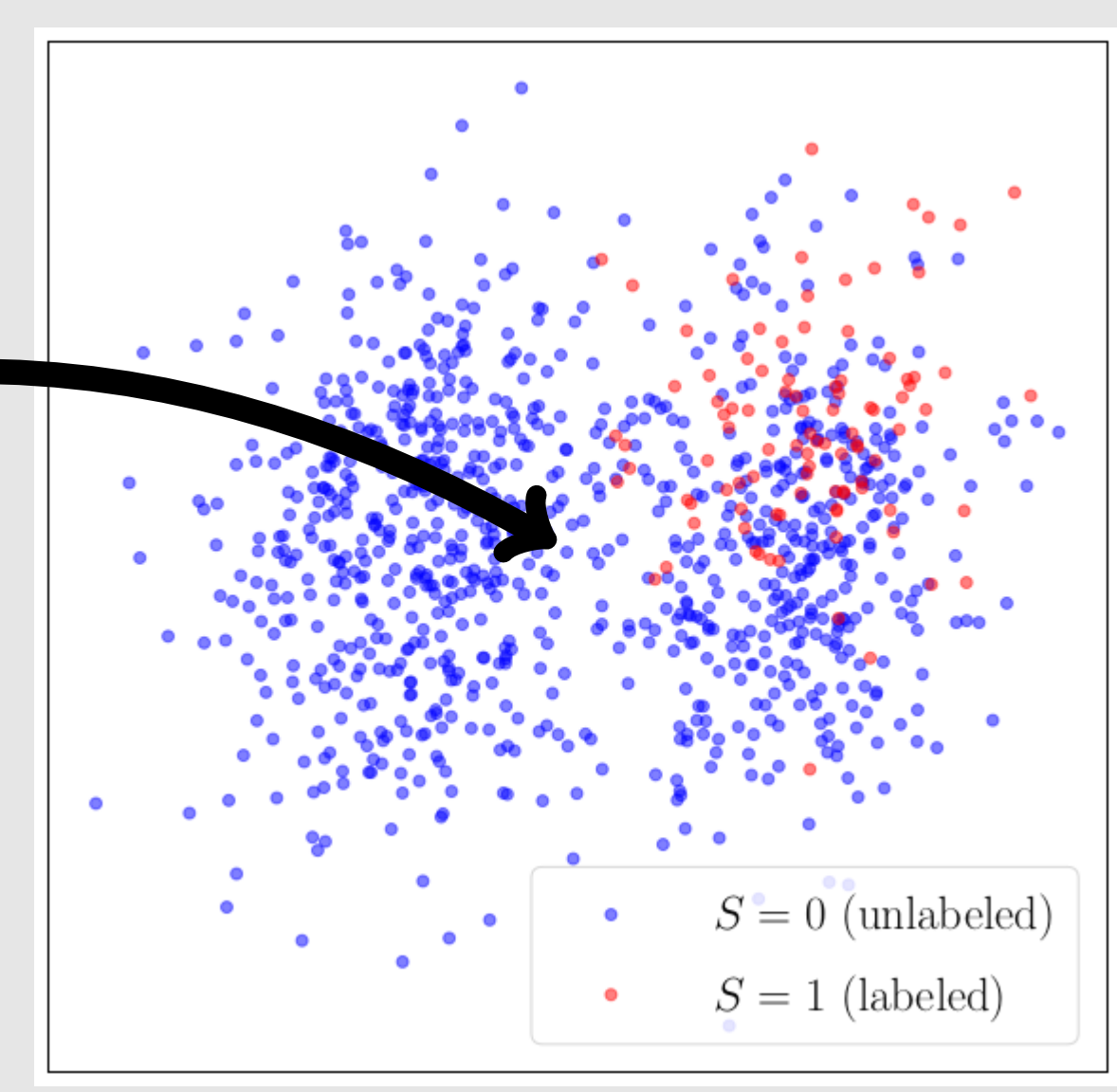


Standard classification setting

$(X_i, Y_i) \rightarrow$ Estimate η



Selection of positive instances through e



PU classification setting

$(X_i, S_i) \rightarrow$ Estimate η

Estimating the classifier in PU learning setting

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

$$f^* \in \text{Argmin}_f R(f).$$

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

Empirical risks:

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

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

- Non traditional empirical risk: ignoring the propensity e .

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

\Leftrightarrow **biased** estimate of the true risk.

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

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

Theoretical risk bounds for PU learning^[3]

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

$$\hat{f}_{PU} = \text{Argmin}_{f \in \mathcal{F}} \hat{R}_{PU}(f).$$

Excess risk:

$$\ell(\hat{f}_{PU}, f^*) = R(\hat{f}_{PU}) - R(f^*).$$

Assumptions:

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

Result: two convergence rates

$$\mathbb{E}[\ell(\hat{f}, f^*)] \leq \kappa \sqrt{\frac{V}{n e_m}} \quad \text{if } h \leq \sqrt{\frac{V}{n e_m}};$$

$$\leq \kappa \left[\frac{V(2 - e_m)}{n e_m h} \left(1 + \log \left(\frac{n h^2}{V} \vee 1 \right) \right) \right] \quad \text{if } h \geq \sqrt{\frac{V}{n e_m}}.$$

Simulation setting

- One-dimensional setting ($d = 1$)
- $\mathcal{F} = \{x \mapsto \mathbb{1}_{x \geq m}, m \in \mathbb{R}\}$
- $X_i \sim \mathcal{N}(0, 1)$
- $Y_i \sim \mathcal{B}\left(\frac{1+h}{2} \mathbb{1}_{X_i \geq 0} + \frac{1-h}{2} \mathbb{1}_{X_i < 0}\right)$, $h = 0.25$.
- $S_i \sim \mathcal{B}(e(x))$ if $Y_i = 1$ (else $S_i = 0$)

- Propensity functions:
 - SCAR constant propensity, $e(x) = e_m$, with $e_m > 0$;
 - SAR logistic propensity (cf. Fig. 1), $e(x) = \max\left(e_m, \frac{1}{1 + e^{-x}}\right)$, with $e_m > 0$.

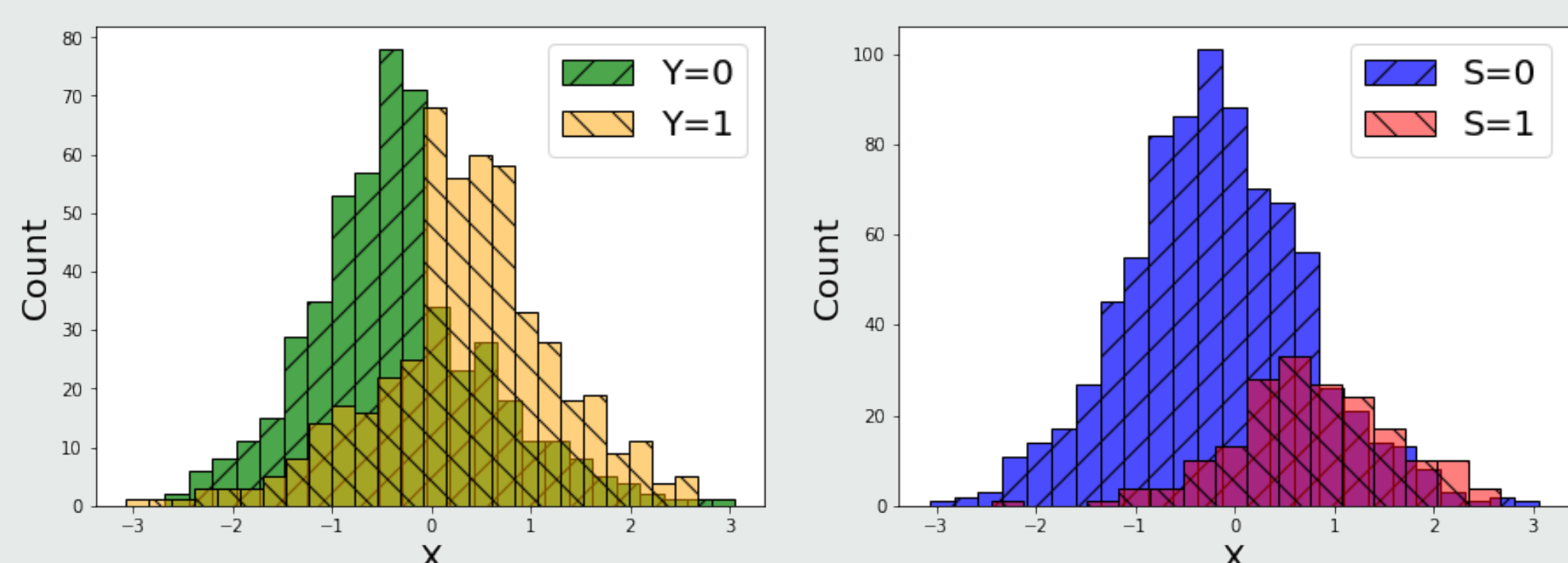
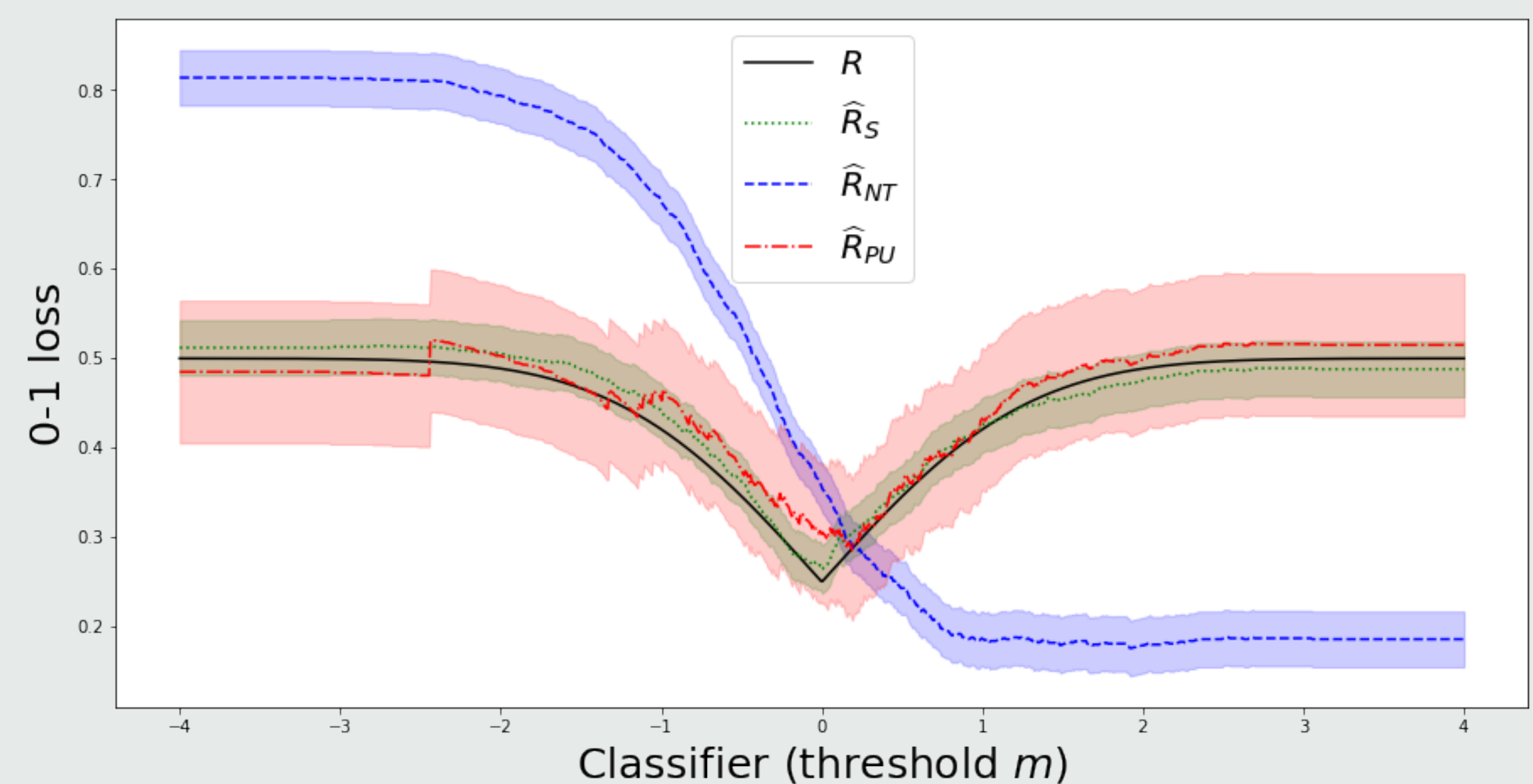


Figure 1. Simulation example (SAR)

Comparison of the empirical risk functions



Conclusion

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

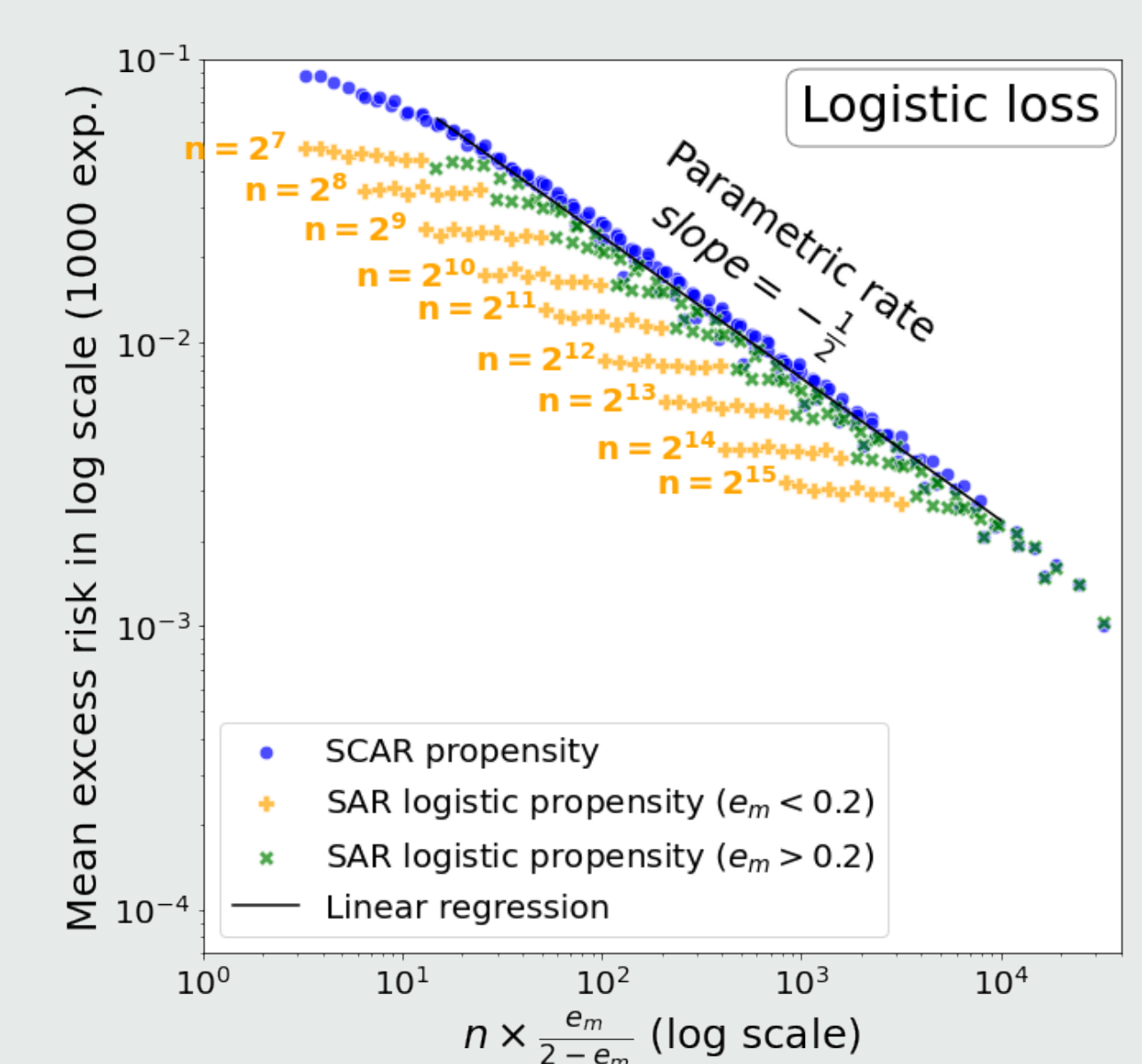
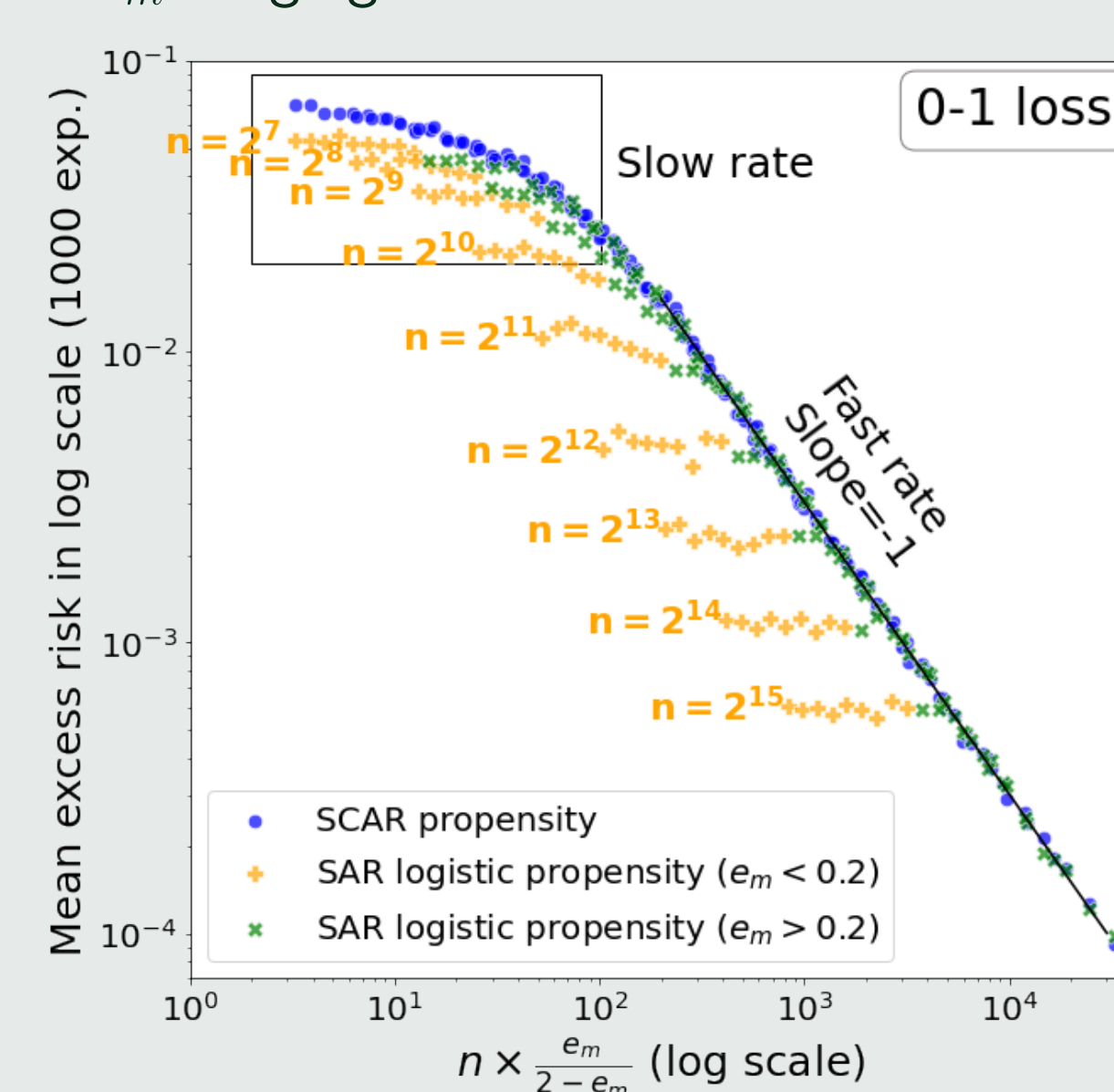
Convergence rates: simulation study

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:

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



Conclusion

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

Perspectives

Ongoing / future work

- Dealing with an unknown propensity: joint estimation of the classifier η 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  . N d lec. "Risk bounds for statistical learning". In: *Annals of Statistics* 34.5 (Oct. 2006).