A Novel Characterization of the Population Area Under the Risk Coverage Curve (AURC) and Rates of Finite Sample Estimators

Han Zhou[1] Jordy Van Landeghem[2] Teodora Popordanoska[1] Matthew B. Blaschko[1]
[1] Processing Speech and Images, Department of Electrical Engineering, KU Leuven, Belgium
[2] Instabase, San Francisco, USA

Motivation

Selective classifier
Given a classifier $f : \mathcal{X} \to \Delta^k$, the selective classifier (f, g) at an input x is given by

$$(f,g)(x) := \begin{cases} f(x) & \text{if } g(x) \geq \tau, \\ \text{“abstain”} & \text{otherwise.} \end{cases} \quad (1)$$

where “abstain” is triggered when the *confidence scoring function* $g(x) < \tau \in \mathbb{R}$.

- selective risk w.r.t. $\mathrm{P}(x, y)$ is

$$R(f, \tilde{g}) := \frac{\mathbb{E}_{\mathrm{P}(x,y)}\left[\ell(f(x), y)\mathbb{I}[g(x) \geq \tau]\right]}{\phi(f, g)}. \quad (2)$$

- *coverage*: $\phi(f, g) = \mathbb{E}_{\mathrm{P}(x)}\left[\mathbb{I}[g(x) \geq \tau]\right]$ represents the probability mass over the accepted samples.

Area Under the Risk Coverage curve (AURC)
The AURC [1] is typically specified as an empirical quantity from a finite sample, from which we derive the population AURC as

$$\mathrm{AURC}_p(f) = \mathbb{E}_{\tilde{x}\sim\mathrm{P}(x)} \frac{\mathbb{E}_{(x,y)\sim\mathrm{P}(x,y)}\ell(f(x), y))\mathbb{I}\left[g(x) \geq g(\tilde{x})\right]}{\mathbb{E}_{x'\sim\mathrm{P}(x)}\mathbb{I}\left[g(x') \geq g(\tilde{x})\right]}. \quad (3)$$

Problem

- Finite sample limitation: Prior works compute AURC empirically rather than at the population level, with little analysis of its statistical properties—estimator like SELE [2] can remain biased even with large sample sizes.
- Optimization gap: Few methods optimize AURC directly, and existing estimator do not guarantee convergence to the population AURC.

Our interpretation

Define function $G(x)$ as the cumulative distribution function(CDF) of the CSF $g(x)$ such that

$$G(x) = \Pr\left(g(x') \leq g\left(x\right)\right) = \int \mathbb{I}\left[g(x') \leq g\left(x\right)\right] d\mathrm{P}(x').$$

Under this definition, the population AURC in Eq. (3) is equivalent to:

$$\mathrm{AURC}_a(f) = \int \alpha(x)\ell(f(x), y))d\mathrm{P}(x, y) \quad (4)$$

where $\alpha(x) = -\ln\left(1 - G(x)\right)$.

- redistribution of the risk
- $G(x)$: the population rank percentile based on the CSF sorted in ascending order.

Proposed estimators

Estimators for $\alpha(x)$ can be achieved via Monte Carlo:

$$\hat{\alpha}_i = H_n - H_{n-r_i} \text{ and } \hat{\alpha}'_i = -\ln(1 - \frac{r_i}{n+1}). \quad (5)$$

- Both are consistent.
- $\hat{\alpha}_i$ upper bounds the $\hat{\alpha}'_i$, leading to

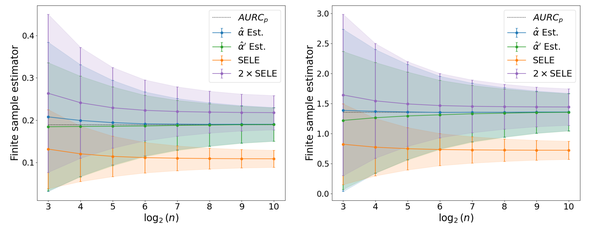
$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{\alpha}'_i\ell(f(x_i), y_i)}_{\widehat{\mathrm{AURC}}'_p(f)} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{\alpha}_i\ell(f(x_i), y_i)}_{\widehat{\mathrm{AURC}}_p(f)}.$$

Proposition 1 (MSE of $\hat{\alpha}_i$ or $\hat{\alpha}'_i$) *Both* $\mathrm{MSE}(\hat{\alpha}_i)$ *and* $\mathrm{MSE}(\hat{\alpha}'_i)$ *are asymptotically bounded by* $\mathcal{O}(\frac{\beta_i}{n(1-\beta_i)+1})$.

Proposition 2 (Convergence Rate of the estimators) *Assume that the loss function* ℓ *is square-integrable, the plug-in estimators with* $\hat{\alpha}_i$ *or* $\hat{\alpha}'_i$ *as the weight estimator, converges at a rate of* $\mathcal{O}(\sqrt{\ln(n)/n})$.

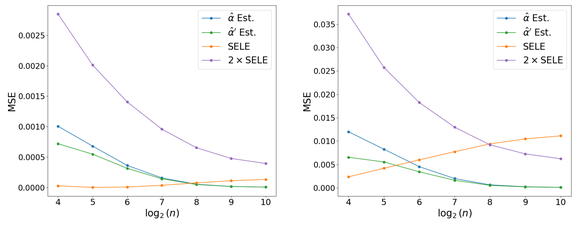
Experiments & Results

Datasets. CIFAR10/100, ImageNet and a text dataset i.e Amazon Reviews.
Models. pre-trained from those datasets.
Metrics. plug-in estimators with $\hat{\alpha}$ or $\hat{\alpha}'$, the SELE score [2].



(a) BERT (**0/1**) (b) BERT (**CE**)

Figure 1: (**Amazon**) Finite-sample estimators with **0/1** or **CE** loss. We utilize a pre-trained model and randomly divide the test set into batch samples of size n. Subsequently, we compute the mean and standard deviation of various estimators applied to these batch samples.



(a) Swin-Base (**0/1**) (b) Swin-Base (**CE**)

Figure 2: (**ImageNet**) MSE of finite sample estimators with **0/1** or **CE** loss. For each model architecture, we calculate the MSE of the estimators using a pre-trained model on batch samples derived from the test set.

	CIFAR10				CIFAR100			
Model	CE	SELE	$\hat{\alpha}$ Est.	$\hat{\alpha}'$ Est.	CE	SELE	$\hat{\alpha}$ Est.	$\hat{\alpha}'$ Est.
ResNet18	$4.967_{\pm 0.038}$	$\mathbf{4.470}_{\pm 0.030}$	$\mathbf{4.473}_{\pm 0.030}$	$\mathbf{4.471}_{\pm 0.030}$	$6.648_{\pm 0.021}$	$\mathbf{6.577}_{\pm 0.011}$	$\mathbf{6.532}_{\pm 0.012}$	$\mathbf{6.533}_{\pm 0.014}$
ResNet34	$6.464_{\pm 0.036}$	$\mathbf{5.661}_{\pm 0.039}$	$\mathbf{5.652}_{\pm 0.036}$	$\mathbf{5.651}_{\pm 0.036}$	$6.023_{\pm 0.016}$	$\mathbf{5.862}_{\pm 0.012}$	$\mathbf{5.825}_{\pm 0.011}$	$\mathbf{5.826}_{\pm 0.011}$
ResNet50	$8.318_{\pm 0.002}$	$\mathbf{7.892}_{\pm 0.046}$	$\mathbf{7.921}_{\pm 0.047}$	$\mathbf{7.918}_{\pm 0.049}$	$6.225_{\pm 0.009}$	$\mathbf{6.043}_{\pm 0.015}$	$\mathbf{6.007}_{\pm 0.008}$	$\mathbf{6.007}_{\pm 0.009}$
VGG16BN	$7.922_{\pm 0.002}$	$\mathbf{7.010}_{\pm 0.018}$	$\mathbf{7.064}_{\pm 0.014}$	$\mathbf{7.060}_{\pm 0.015}$	$10.790_{\pm 0.001}$	$\mathbf{10.586}_{\pm 0.029}$	$\mathbf{10.559}_{\pm 0.029}$	$\mathbf{10.560}_{\pm 0.030}$
VGG19BN	$9.813_{\pm 0.192}$	$\mathbf{8.475}_{\pm 0.061}$	$\mathbf{8.528}_{\pm 0.059}$	$\mathbf{8.524}_{\pm 0.059}$	$10.633_{\pm 0.001}$	$\mathbf{10.421}_{\pm 0.026}$	$\mathbf{10.393}_{\pm 0.025}$	$\mathbf{10.391}_{\pm 0.024}$
WideResNet28x10	$4.137_{\pm 0.046}$	$\mathbf{3.867}_{\pm 0.049}$	$\mathbf{3.864}_{\pm 0.049}$	$\mathbf{3.863}_{\pm 0.049}$	$5.912_{\pm 0.652}$	$\mathbf{5.607}_{\pm 0.707}$	$\mathbf{5.836}_{\pm 0.652}$	$\mathbf{5.607}_{\pm 0.707}$

Table 1: Summary of population AURC_p (mean $\pm$ std, scaled by 10^{-2}) on the test set for models fine-tuned with various loss functions. Each entry aggregates results over five seeds using the same pre-trained model.

Conclusion

- We extend empirical AURC to a true population quantity and show it admits a reweighted-risk interpretation.
- We propose two plug-in estimators via Monte Carlo method and show their bias, MSE, and an $\mathcal{O}(\sqrt{\ln n/n})$ convergence rate.
- Experiments demonstrate these consistent estimators not only outperform SELE in terms of estimation but also serve as effective objectives for directly fine-tuning networks to minimize AURC.

References

[1] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019.
[2] Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023.

Contact Information

Paper https://arxiv.org/pdf/2410.15361
Email han.zhou@esat.kuleuven.be