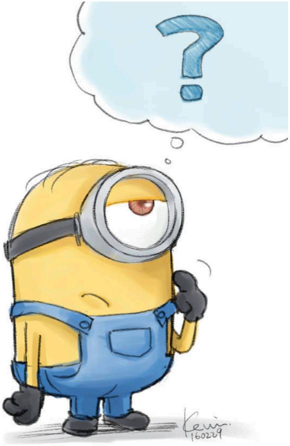
Towards Failure Detection With Statistical Guarantees

Bellairs Worshop

Matteo Sammut – Dec. 17 2025

Supervised by: Pablo Piantanida (ILLS), Yann Chevaleyre (Dauphine PSL), Rafael Pinot (LPSM - Sorbonne)

Outline of the Talk



I. Background on Failure Detection

II. Impossibility of Estimating the Point-wise Error Probability

III. Estimation at a Lower Resolution

IV. Error-Detector with Statistical Guarantees

IV. Experiments

Background on Failure Detection

Formalization

- Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = [K]$ be the input and label spaces, respectively.
- Let $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a data distribution.
- Let $f : \mathcal{X} \to \mathcal{Y}$ be a pretrained classifier.

Current Goal: Construct an **uncertainty score** $u : \mathcal{X} \to [0, 1]$ for predicting the occurrence of error:

$$E := Y \neq f(\mathbf{x})$$

for any input $\mathbf{x} \in \mathcal{X}$.

Implicit Goal: Approximate the error probability function:

$$\eta_{f,P}(\mathbf{x}) := \mathbb{P}\{Y \neq f(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}\},$$

i.e., the **regression function** of this **binary classification** problem.

Limitations of Current Work

- ✓ Current uncertainty score u performs well **on average** (e.g., AUROC, FPR@95).
- ✗ But they provide **no statistical guarantees** on their **approximation error**.

Our goal: Estimate the error probability function $\eta_{f,P}$ with valid confidence bounds.

$\alpha-$Coverage

An algorithm $\widehat{C}_n$ provides an $\alpha-$confidence interval if for any data distribution $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ it holds that for any $\mathsf{x} \in \mathcal{X}$,

$$\mathbb{P}_{\mathcal{D}_n}\left\{\eta_{f,P}(\mathsf{x}) \in \widehat{C}_n(\mathsf{x}; \mathcal{D}_n, f)\right\} \geq 1 - \alpha,$$

where $\mathcal{D}_n \overset{\text{i.i.d.}}{\sim} P$.

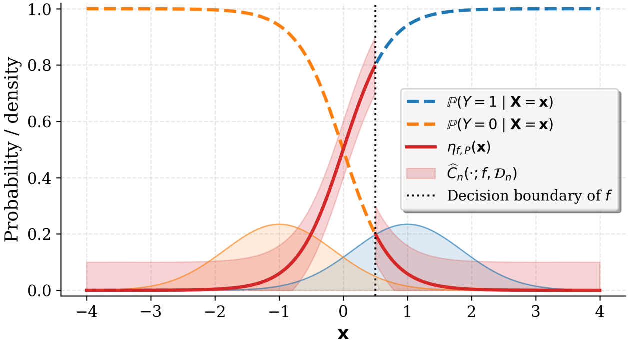
Precise Inference: We say that an $\alpha-$confidence interval $\widehat{C}_n$ is *precise* with respect to $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ if, for any ,

$$\forall \mathbf{x} \in \mathcal{X}, \quad \lim_{n \to \infty} \mathbb{E}\left[\mathrm{leb}(\widehat{C}_n(\mathsf{x}))\right] = 0$$

.

Illustrative Example

- Let $P_{\mathcal{X}|\mathcal{Y}}(\cdot \mid 0) = \mathcal{N}(-1, \sigma)$, $P_{\mathcal{X}|\mathcal{Y}}(\cdot \mid 1) = \mathcal{N}(1, \sigma)$, $f(\mathbf{x}) := \mathbb{1}\{\mathbf{x} \geq 0.5\}$.
- $$\eta_{f,P}(\mathbf{x}) = \begin{cases} \mathbb{P}(Y = 1 \mid X = \mathbf{x}) & \text{if } \mathbf{x} < 0.5, \\ \mathbb{P}(Y = 0 \mid X = \mathbf{x}) & \text{if } \mathbf{x} \geq 0.5. \end{cases}$$



Impossibility of Estimating the Point-wise Error Probability

Impossibility in the Distribution-Free setting

Informal Theorem restated from Barber (2020)

Let $\widehat{C}_n$ that provides an $\alpha-$ confidence interval. For any $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that $P_{\mathcal{X}}$ is nonatomic[1], then there **exists a constant** $C_\alpha(f, P)$ **independant of** n such that:

$$\mathbb{E}_{(D_n, X)}\left[\text{leb}\left(\widehat{C}_n(\mathsf{X}; \mathcal{D}_n, f)\right)\right] \geq C_\alpha(f, P) > 0.$$

Intuitions:

- In the **distribution-free** setting, to infer $\eta_{f,P}(\mathbf{x})$ you can **only use** calibration data $(X_i, Y_i) \in \mathcal{D}_n$ for which $X_i = \mathbf{x}$.
- ✗ If $P_{\mathcal{X}}(\mathbf{x}) = 0$, you will never get in off such calibration point.

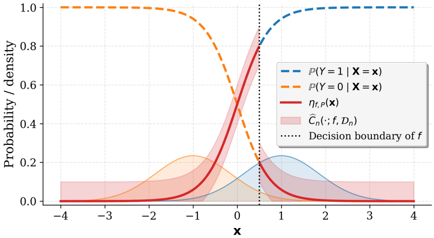
[1] We say that the marginal $P_{\mathcal{X}}$ is *nonatomic* if for any $\mathbf{x} \in \mathcal{X}, \quad P_{\mathcal{X}}\{\mathsf{x}\} = 0$.

- If you assume that your **regression function** lies in a **smooth class** of functions, precise inference should be possible.

On **which assumptions** does $\eta_{f,P}$ become **smooth**?

Theorem for $\mathcal{Y} = \{0, 1\}$:

- The regularity of $\eta_{f,P}$ in the **interior of the level sets** of f is inherited from the one of $\mathbf{x} \mapsto \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$.
- $\eta_{f,P}$ is continuous at the **decision boundary** of f iff f has the **same decision boundary** than the **Bayes classifier**.



Overview of Impossibility Results

Constructing **distribution**-free α-confidence intervals is fundamentally hard:

.

- The **distribution-free requirement** renders **precise inference infeasible** for many distributions.
- Restricting coverage to **"smooth" distributions** is not applicable to our problem, since the **error-probability** function $\eta_{f,P}$ typically **exhibits irregularities** (except in degenerate cases)

Estimation of the Error Probability at a Lower Resolution

Coverage at resolution r

Let $r : \mathcal{X} \to \mathcal{Z}$ be a **resolution function** and define the **levels set** of r as:

$$\mathcal{X}_{\mathbf{z}} := \{\mathbf{x} \in \mathcal{X} : r(\mathbf{x}) = \mathbf{z}\}.$$

Define the **error-probability function at resolution** r by

$$\forall \mathsf{z} \in \mathcal{Z}, \quad \eta_{f,P,r}(\mathsf{z}) := \mathbb{P}\{E = 1 \mid \mathsf{X} \in \mathcal{X}_{\mathbf{z}}\} = \mathbb{E}\left[\eta_{f,P}(\mathsf{X}) \mid \mathsf{X} \in \mathcal{X}_{\mathbf{z}}\right]$$

Examples :

- If $r = \mathrm{Id} \implies \mathcal{X}_{\mathbf{z}} = \{\mathbf{z}\} \implies \eta_{f,P,r} = \eta_{f,P}$ - **high resolution.**
- If r constant $\implies \mathcal{X}_{\mathbf{z}} = \mathcal{X} \implies \eta_{f,P,r} = \mathbb{P}(E = 1)$ - **low resolution.**

Feasibility of coverage at resolution r : The partition $\{\mathcal{X}_{\mathbf{z}} : \mathbf{z} \in \mathcal{Z}\}$ being **countable** is a **necessary condition** for **precise inference** at resolution r for any $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

- Let $r : \mathcal{X} \to \mathcal{Z}$ be any resolution function with $|\mathcal{Z}| < \infty$.
- Let $\mathcal{D}_n = \{(\mathsf{X}_1, Y_1), \ldots, (\mathsf{X}_n, Y_n)\} \overset{\text{i.i.d.}}{\sim} P$ be a calibration set.

For any $\mathbf{z} \in \mathcal{Z}$, denote by

$$N_{\mathbf{z}} := \Big|\{\, i \in [n] : \mathsf{X}_i \in \mathcal{X}_{\mathbf{z}} \,\}\Big|. \tag{1}$$

the (random) number of calibration points that fall in the cell $\mathcal{X}_{\mathbf{z}}$. When $N_{\mathbf{z}} > 0$, define the empirical estimator $\widehat{\eta}_n(\mathsf{z}) = \widehat{\eta}_n(\mathsf{z}; \mathcal{D}_n, f)$ of $\eta_r(z)$ by

$$\widehat{\eta}_n(\mathsf{z}) := \frac{1}{N_{\mathbf{z}}} \sum_{i=1}^{n} E_i \cdot \mathbb{1}\{\mathsf{X}_i \in \mathcal{X}_{\mathbf{z}}\}, \tag{2}$$

where $E_i := \mathbb{1}\{Y_i \neq f(\mathsf{X}_i)\}$. Finally, define the intervals

$$\widehat{C}_n(\mathbf{z}, \mathcal{D}_n, f) := \left[\widehat{\eta}_n(\mathbf{z}) \pm \sqrt{\frac{\ln(2/\alpha)}{2N_{\mathbf{z}}}}\right] \cap [0, 1], \tag{3}$$

if $N_{\mathbf{z}} > 0$ and set $\widehat{C}_n(\mathbf{z}, \mathcal{D}_n, f) = [0, 1]$ if $N_{\mathbf{z}} = 0$.

Partition Algorithm Guarantees

Theorem

The confidence interval $\widehat{C}_n$ defined in (3) provides an $\alpha-$coverage at resolution r. Moreover,

$$\forall \mathbf{z} \in \mathcal{Z}, \quad \mathbb{E}_{\mathcal{D}_n}\left[\mathrm{leb}(\widehat{C}_n(\mathbf{z}; D_n, f))\right] \leq \min\left\{1, \frac{c(\alpha)}{\sqrt{n P_{\mathcal{X}}\{\mathcal{X}_{\mathbf{z}}\}}}\right\}, \qquad (4)$$

where $c(\alpha)$ is a universal constant depending only on α.

Precise Inference: The confidence interval is **precise** for any $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$

Trade Off: **Larger** level sets $\mathcal{X}_{\mathbf{z}}$ **improve** the **convergence rate** in (4) but may lead to **less qualitative resolution**.

Error-Detector with Statistical Guarantee

Misclassifications Detection at resolution r: Given a loss function l_τ consider the learning problem

$$\inf_{d:\mathcal{Z}\to\{0,1\}} \mathbb{E}\big[\ell_\tau(E,\ d(r(\mathbf{X})))\big]. \tag{5}$$

Bayes detector: $d^*_{f,P,r}(\mathbf{z}) := \mathbb{1}\{\eta_{f,P,r}(\mathbf{z}) \geq \tau\}.$

Conservative Detector: $\widehat{d}_n(\mathbf{z};\mathcal{D}_n,f) := \mathbb{1}\{\sup \widehat{C}_n(\mathbf{z};\mathcal{D}_n,f) \ \geq\ \tau\}.$

Agreement with Bayes Decision

For any $\mathbf{z} \in \mathcal{Z}$ such that $d^*_{f,P,r}(\mathbf{z}) = 1$ then,

$$\mathbb{P}_{\mathcal{D}_n}\{\widehat{d}_n(\mathbf{z};\mathcal{D}_n,f) = 1\} \geq 1-\alpha,$$

and

$$\mathbb{P}_{\mathcal{D}_n}\{\widehat{d}_n(\mathbf{z};\mathcal{D}_n,f) = 0\} \leq \alpha.$$

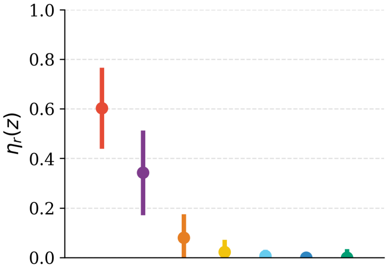
Experiments

Quantization Algorithm: **Gaussian Mixture Model** on the softmax output of the model f.

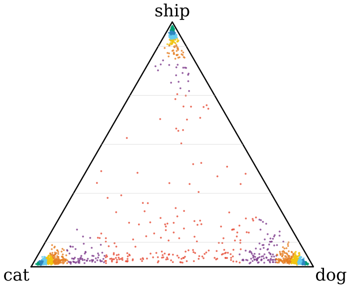
Preprocessing: **Reordering** the softmax output $\implies$ **invariant** to the predicted class.

Data Splitting: In our results, the resolution function r was **fixed** ! We use distinct data sets $\mathcal{D}_{\mathrm{res}}$ and $\mathcal{D}_{\mathrm{cal}}$ to **learn** the **resolution function** and **construct** $\widehat{C}_n$ respectively.

Setting: CIFAR10, ResNet34 with $\mathbb{P}(E = 1) \approx 5\%$.



(a) Confidence Intervals $\widehat{C}_n$ per level sets $\mathcal{X}_z$



(b) Level sets $\mathcal{X}_z$ visualization

Setting: ImageNet, ViT-Base-16 with $\mathbb{P}(E = 1) \approx 20\%$.

(a) $\eta_r(z) \in [38.8,\ 84.6]\%$

(b) $\eta_r(z) \in [0,\ 1.3]\%$

Results

- Achieves competitive performance compared to SOTA heuristic methods.

Dataset	Method Model	Ours	Doctor	ODIN	Rel-U
CIFAR-10	DenseNet-121	29.2/91.0/**15.1**	**24.1**/91.6/**15.1**	31.4/<u>91.7</u>/16.1	<u>27.1</u>/**92.2**/16.0
	ResNet-34	<u>23.9</u>/<u>93.2</u>/14.1	**22.8**/**93.6**/14.1	27.0/92.6/<u>14.0</u>	26.8/90.2/**12.0**
CIFAR-100	DenseNet-121	48.9/84.8/ <u>47.8</u>	48.4/**86.0**/48.7	<u>48.3</u>/<u>85.5</u>/48.6	**46.5**/82.3/**44.7**
	ResNet-34	44.3/85.6/**41.4**	<u>42.1</u>/<u>86.8</u>/43.1	42.5/**87.4**/44.0	**41.2**/86.7/<u>41.6</u>
ImageNet-1k	ViT-Tiny-16	46.6/84.6/<u>44.6</u>	**46.0**/<u>86.5</u>/47.6	**46.0**/**86.7**/47.8	51.2/80.3/**40.6**
	ViT-Base-16	**42.3**/86.4/<u>37.2</u>	**42.3**/**87.7**/39.0	**42.3**/**87.7**/39.1	49.0/82.9/**33.9**

Table 1: MisD results in terms of FPR@95/AUROC/AURC.

References

Barber, R. F. (2020). Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487 – 3524.