What Does It Take to Build a Performant Selective Classifier?

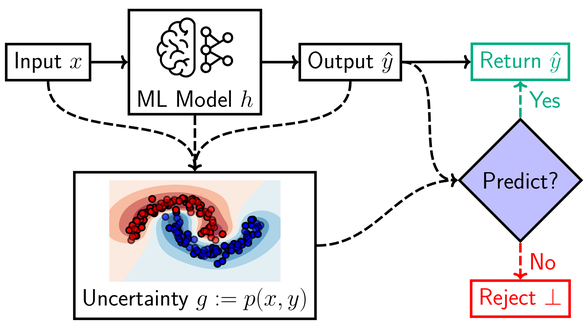
Stephan Rabanser, *Princeton University* — **Nicolas Papernot**, *University of Toronto & Vector Institute*

Main Contribution

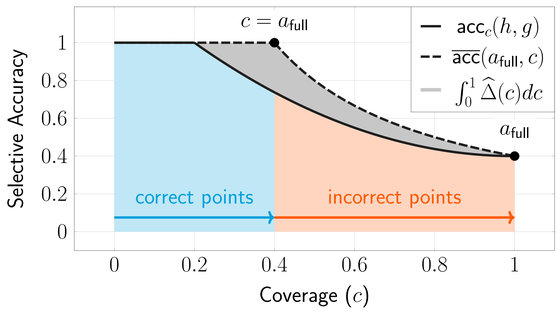
We decompose the gap between an empirically observed selective classifier and its ideal oracle into five distinct measurable sources of error. Our performance bound allows practitioners to build better selective classifiers.

Learn more: Paper + Code

Selective Classification (SC) Overview



Measuring Selective Classification Performance



Question

For my finite model h on finite data D, what aspects of the learning setup will actually move my trade-off curve $\text{acc}_c(h, g)$ closer to the perfect-ordering upper bound $\overline{\text{acc}}(a_{\text{full}}, c)$?

Insight

$$\widehat{\Delta}(c) := \overline{\text{acc}}(a_{\text{full}}, c) - \text{acc}_c(h, g) \qquad \forall c \in (0, 1]$$
$$\leq \underbrace{\varepsilon_{\text{Bayes}}(c)}_{\text{irreducible}} + \underbrace{\varepsilon_{\text{approx}}(c)}_{\text{capacity}} + \underbrace{\varepsilon_{\text{rank}}(c)}_{\text{ranking}} + \underbrace{\varepsilon_{\text{stat}}(c)}_{\text{data}} + \underbrace{\varepsilon_{\text{misc}}(c)}_{\text{optimization \& shift}}$$

Error Sources Contributing to the SC Gap

Notation: Let $\eta(x) := \Pr(Y = 1 \mid X = x)$ be the *Bayes posterior*. For a classifier $h : \mathcal{X} \to \mathcal{Y}$ define its (induced) *correctness posterior*

$$\eta_h(x) := \Pr(h(x) = Y \mid X = x) = \eta(x)\,\mathbb{I}_{\{h(x)=1\}} + (1 - \eta(x))\,\mathbb{I}_{\{h(x)=0\}}.$$

Let $g(x, h)$ be the confidence score. For a target coverage $c \in (0, 1]$

$$t_c \quad \text{s.t.} \quad \Pr(g(X, h) \geq t_c) = c$$

denotes the *population threshold*. We write the *accepted region*

$$A_c := \{x : g(x, h) \geq t_c\}.$$

The oracle that attains the perfect-ordering bound accepts

$$A_c^\star := \{x : \eta_h(x) \text{ is among the largest } c\text{-fraction}\}.$$

---

❶ **Bayes Error**: Even a Bayes-optimal rule errs on intrinsically ambiguous points (where $\max_y \Pr(Y = y \mid X) < 1$).

$$\varepsilon_{\text{Bayes}}(c) := \mathbb{E}\left[1 - \max\{\eta(X), 1 - \eta(X)\} \;\middle|\; X \in A_c\right]$$

❷ **Approximation Error**: A model h drawn from a restricted hypothesis class may misclassify inputs with high posterior confidence.

$$\varepsilon_{\text{approx}}(c) := \mathbb{E}\left[|\eta_h(X) - \eta(X)| \;\middle|\; X \in A_c\right]$$

❸ **Ranking Error**: When $g(X, h)$ approximates $\eta_h(X)$ badly, poor ranking interleaves high-confidence errors and low-confidence corrects.

$$\varepsilon_{\text{rank}}(c) := \mathbb{E}\left[\eta_h(X) \mid X \in A_c^\star\right] - \mathbb{E}\left[\eta_h(X) \mid X \in A_c\right]$$

❹ **Statistical Error**: Estimating selective accuracy from a finite validation set introduces randomness (concentration bounds).

$$\varepsilon_{\text{stat}}(c) := C\sqrt{\frac{\log(1/\delta)}{n}}$$

❺ **Shift & Optimization Error**: Additional looseness in practice.

The Role of Calibration in Ranking

Insight

The impact of calibration on the gap depends critically on the *type* of calibration: **non-monotone scoring is required**.

What is calibration? We say the model is *perfectly calibrated* if $\Pr(Y = \hat{y}(X) \mid s(X) = t) = t$ for all confidence levels $t \in [0, 1]$.

TEMP: Divide every logit vector $z(x) \in \mathbb{R}^K$ by a scalar $T > 0$,

$$p_j^{(T)}(x) = \frac{\exp(z_j(x)/T)}{\sum_k \exp(z_k(x)/T)}.$$

Can lead to **limited** re-ranking!

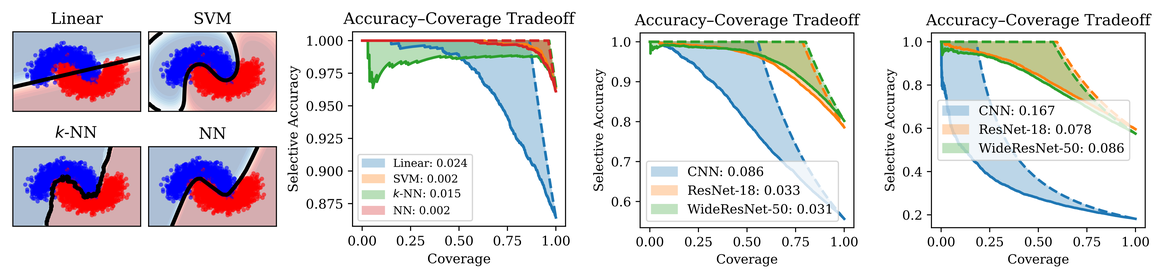
SAT: Modify loss with moving average to prevent over-confidence.

DE: Train multiple independently initialized models and aggregate.

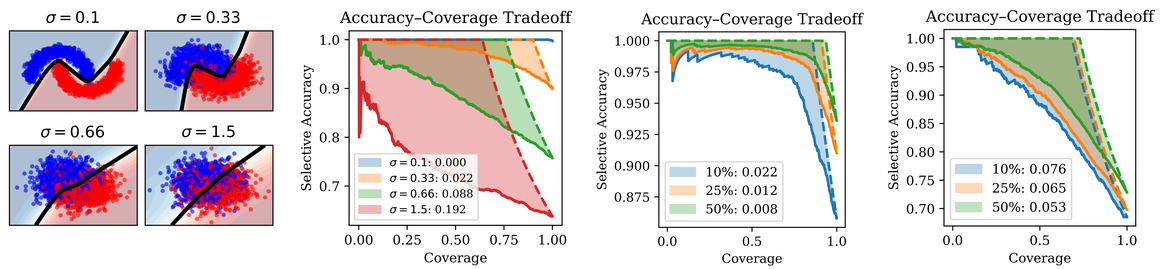
Leads to **improved** re-ranking!

Multi-Calibration Influence: The degree to which a model can predict its own loss corresponds directly to the magnitude of $\varepsilon_{\text{rank}}(c)$.

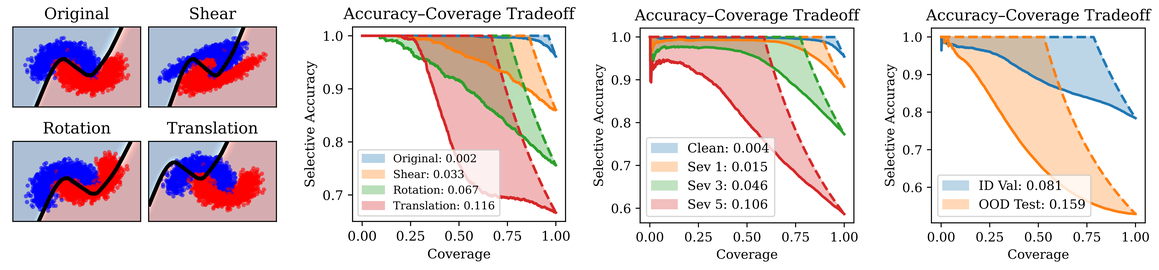
Results Across Synthetic and Real-World Selective Classification Benchmarks



(a) Approximation error with two moons dataset. (b) CIFAR-100 (c) StanfordCars



(a) Bayes error with two moons dataset. (b) CIFAR-10N (c) CIFAR-100N



(a) Distribution shifts with two moons dataset. (b) CIFAR-10C (c) Camelyon17-WILDS