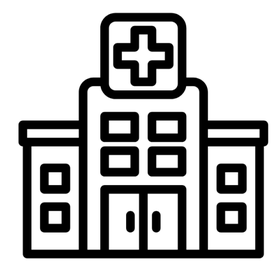
A Causal Framework for Evaluating Deferring Systems

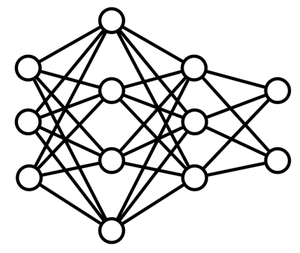
F. Palomba† A. Pugnana* J. Alvarez* S. Ruggieri*
† Princeton University, Princeton, USA * University of Pisa, Pisa, Italy

Why do we care?



Created by Muh Syafri from Noun Project

Created by Wilson Joseph from Noun Project



Created by Lucas Rathgeb from Noun Project

Theoretical Framework

Causal inference

- $D_i \in 0, 1$ prescribes treatment
- $O_i(0), O_i(1)$ potential outcomes;
- $O_i = (1 - D_i)O_i(0) + O_i(1)D_i$
- $\tau_i = O_i(1) - O_i(0)$
- Regression Discontinuity Design (RDD)
- $D_i = \mathbb{1}\{V_i \geq v\}$;
- You can compare instances close to the cutoff v!

Deferring systems

- ML model $f : \mathcal{X} \to \mathcal{Y}$
- Human expert $h : \mathcal{X} \to \mathcal{Y}$
- Deferring system:

$$\vartheta(\mathbf{x}) = (f, g, h)(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if} \quad g(\mathbf{x}) = 0 \\ h(\mathbf{x}) & \text{if} \quad g(\mathbf{x}) = 1 \end{cases}$$

- $k : \mathcal{X} \to \mathbb{R}$
- $g(\mathbf{x}) = \mathbb{1}\{k(\mathbf{x}) \geq \overline{\kappa}\}$

Bridging the two Worlds

- $G_i = g(\mathbf{X}_i)$
- $T_i(0) = \mathbb{1}\{f(\mathbf{X}_i) = Y_i\}$
- $T_i(1) = \mathbb{1}\{h(\mathbf{X}_i) = Y_i\}$
- $T_i = (1 - G_i)T_i(0) + T_i(1)G_i$

Scenario 1

- $\tau_i = T_i(1) - T_i(0), \quad \tau_{\texttt{ATD}} = \mathbb{E}[T(1) - T(0)|G = 1]$

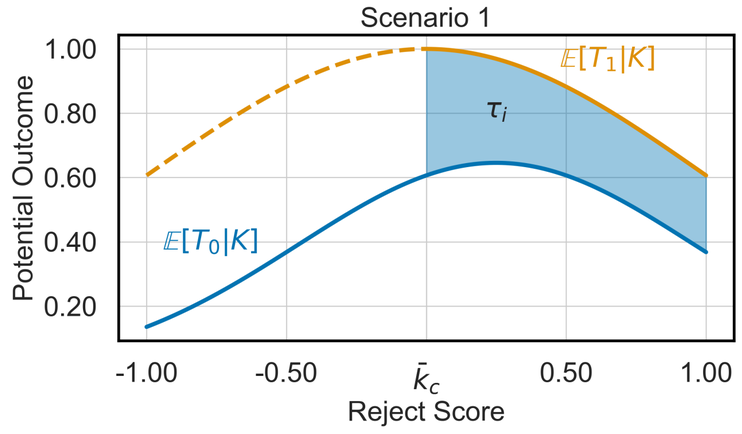


Figure 1: Scenario 1: dashed lines are unobserved values and thick lines observed ones. The coloured area represents where the effects can be estimated (i.e., $k(\mathbf{x}) \geq \overline{\kappa}_c$).

Scenario 2

- $\tau_{RD} = \mathbb{E}[T(1) - T(0) \mid K = \overline{\kappa}_c]$

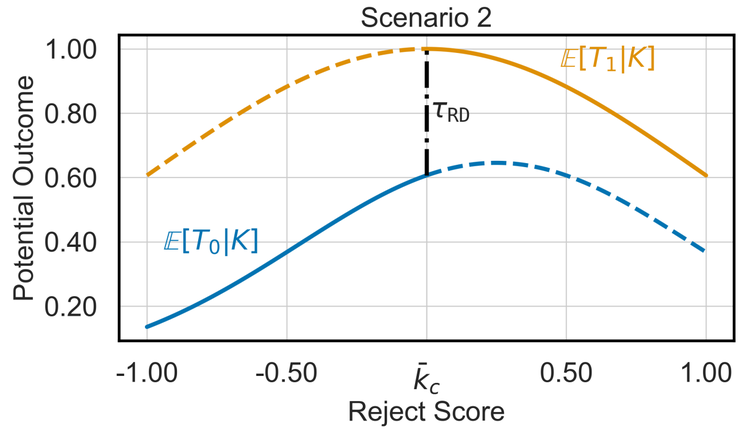


Figure 2: Scenario 2: dashed lines are unobserved values and thick lines observed ones. We can estimate τ_{RD} at the cutoff value (i.e., $k(\mathbf{x}) = \overline{\kappa}_c$).

Experimental Evaluation

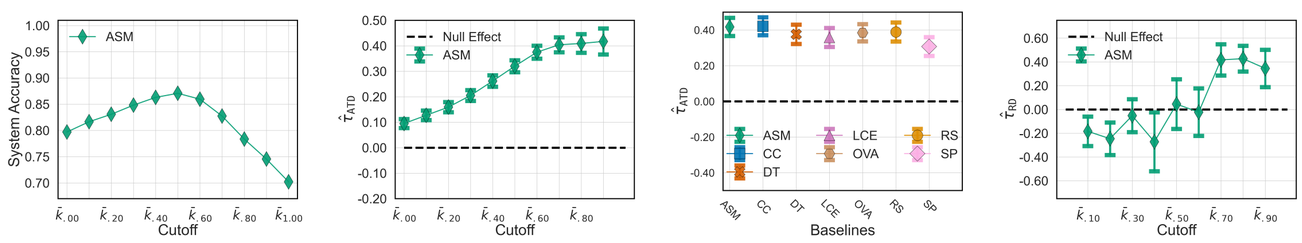


Figure 3: Performance on synthetic data. (a) reports the deferring system accuracy when varying cutoff $\overline{\kappa}_c$ for the best baseline *Asymmetric SoftMax* (ASM) w.r.t. accuracy. (b) reports estimated $\hat{\tau}_{\texttt{ATD}}$ when varying cutoff $\overline{\kappa}_c$ on synthetic data for the best baseline. (c) compares the $\hat{\tau}_{\texttt{ATD}}$ of multiple baselines at a fixed coverage $c = .90$. (d) reports estimated $\hat{\tau}_{\texttt{RD}}$ when varying cutoff $\overline{\kappa}_c$ for the best baseline.

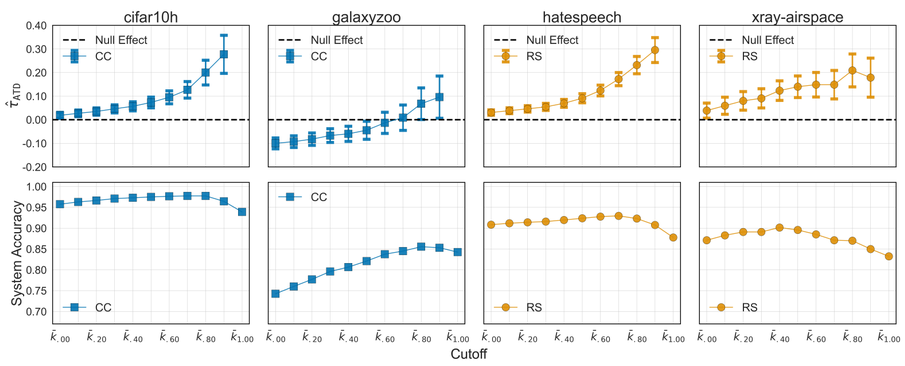


Figure 4: Best deferring system performances on real data when varying the cutoff $\overline{\kappa}_c$. Top: estimated τ_{ATD}. Bottom: accuracy. CC is *Compare Confidence*, RS is *Realizable Surrogate*.

Contacts

mail: andrea.pugnana@di.unipi.it, X: @andrepugni