Counterfactually Comparing Abstaining Classifiers

Yo Joong Choe, Aditya Gangrade, and Aaditya Ramdas

37th Conference on Neural Information Processing Systems (NeurIPS 2023)

Yo Joong "YJ" Choe
University of Chicago
(work done at CMU)

Aditya Gangrade
University of Michigan
(work done at CMU)

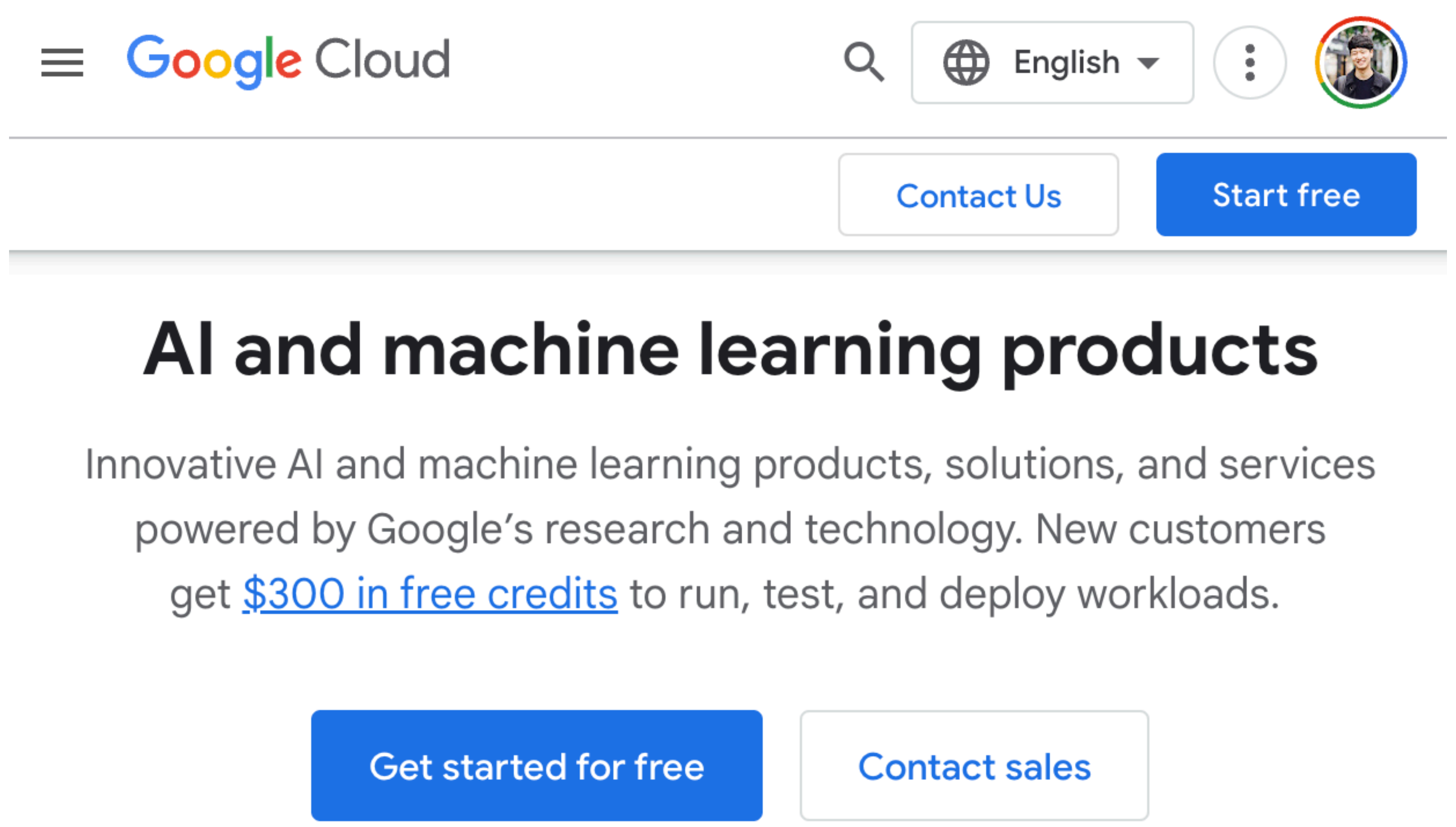
Aaditya Ramdas
Carnegie Mellon University

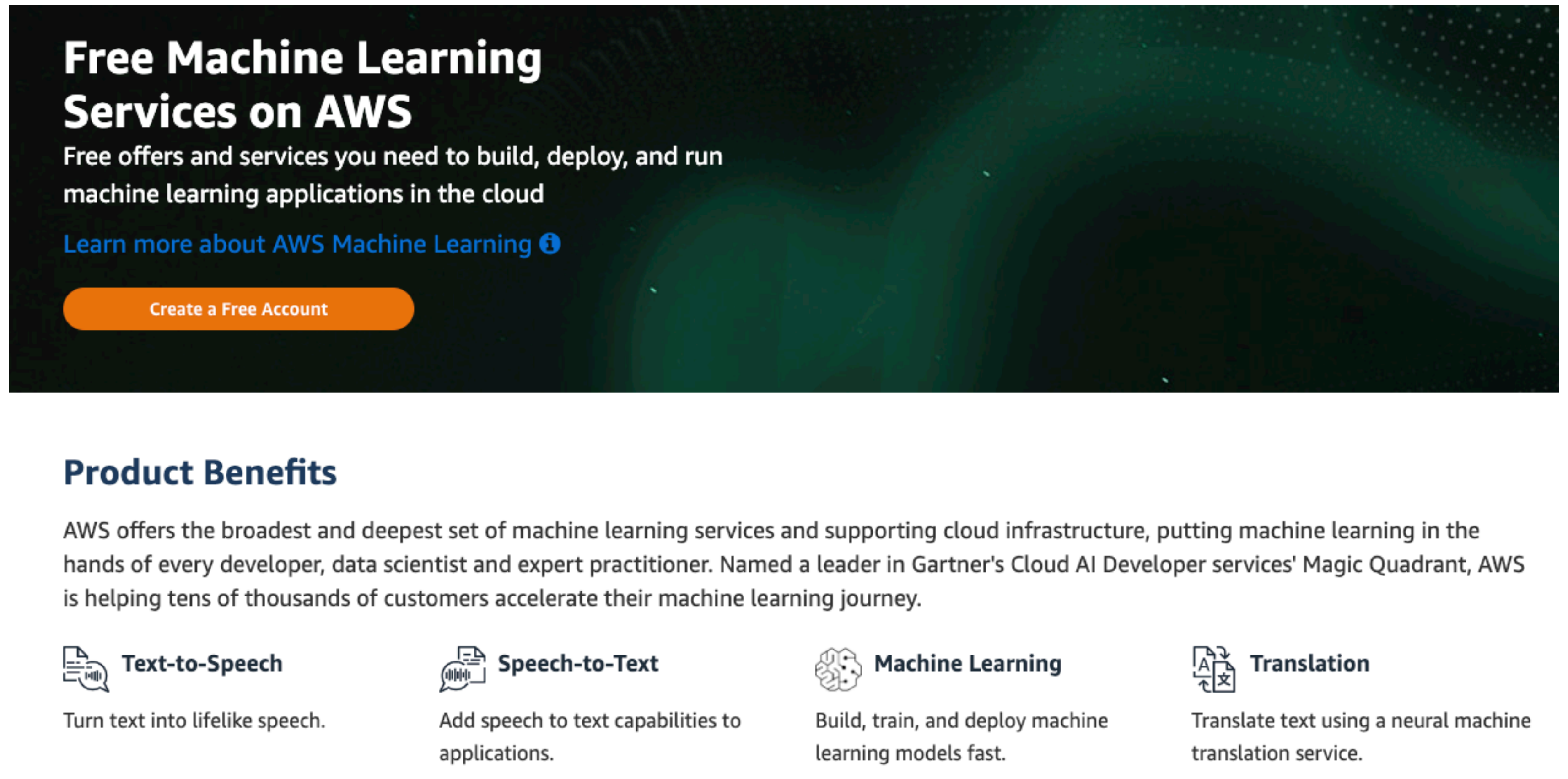
Motivation: Evaluating Free-Trial ML Services

- Suppose that we want to evaluate **black-box** ML prediction services for image classification.

- During the **free trial**, each service deploys an abstaining classifier, such that it only gives predictions on certain inputs and abstain on others.

- **The full (paid) versions do not abstain.** We want to compare the performance of the full versions.





Key Takeaway & Main Question

To the evaluator, abstentions are just **missing** predictions!

How do we compare black-box abstaining classifiers

while accounting for their missing predictions?

Problem Setup

Chow (1957); El-Yaniv & Wiener (2010)

Definition. An **abstaining classifier** is a pair of functions (f, π), where

- $f : \mathscr{X} \to \mathscr{P}(\mathscr{Y})$ is the **base classifier**, which outputs a (probabilistic) prediction; and
- $\pi : \mathscr{X} \to [0, 1]$ is the **abstention mechanism**, which outputs the probability of abstention.

Evaluating a black-box abstaining classifier (f, π).

1. Classifier receives an input X.
2. Classifier decides whether or not it will abstain: $R \mid X \sim \text{Ber}(\pi(X))$.
   - If $R = 0$, then Evaluator observes the prediction & score: $S = s(f(X), Y)$.
   - If $R = 1$ ("rejection"), then Evaluator does NOT see its prediction or score **(S is missing)**.

The 3-Step Approach To Nonparametric Causal Inference

Target Definition → **Identification** → **Estimation**

Target Definition

$$\psi \stackrel{\text{def}}{=} \mathbb{E}[S]$$

("Counterfactual Score")

Nuisance Functions (Learnable):

Abstention Mechanism

$$\pi(X) \stackrel{\text{def}}{=} \mathbb{P}(R = 1 \mid X)$$

Selective Score Predictor

$$\mu_0(X) \stackrel{\text{def}}{=} \mathbb{E}[S \mid R = 0, X]$$

Identification

$$\psi = \mathbb{E}[\mu_0(X)]$$

Conditions

1. **Missing At Random:** $S \perp\!\!\!\perp R \mid X$ — Independent Evaluation Set
2. **Positivity:** $\pi(X) \leq 1 - \epsilon$ (for some $\epsilon > 0$) — Stochastic Abstentions

Estimation

$$\sqrt{n}\left(\hat{\psi}_{\text{dr}} - \psi\right) \rightsquigarrow \mathcal{N}\left(0, \text{Var}_{\mathbb{P}}[\text{IF}]\right)$$

Conditions

1. **Double Robustness:** $\|\hat{\pi} - \pi\|_{L^2} \|\hat{\mu}_0 - \mu_0\|_{L^2} = o_{\mathbb{P}}(1/\sqrt{n})$ — Flexible Nuisance Learners (NN, RF, ...)
2. **IF Consistency:** $\|\hat{\text{IF}} - \text{IF}\| = o_{\mathbb{P}}(1)$

cf. Rubin (1974); Robins et al. (1994); many others.

Other names: augmented IPW (Robins et al., 1994);
targeted MLE (van der Laan & Rubin, 2006);
double ML (Chernozhukov et al., 2018)

The Doubly Robust Estimator $\hat{\psi}_{\mathrm{dr}}$

Given an *i.i.d.* data of potentially missing predictions, $\{(X_i, R_i, (1 - R_i)S_i)\}_{i=1}^{n} \sim \mathbb{P}$, the **doubly robust (DR) estimator for ψ** is defined as:

$$\hat{\psi}_{\mathrm{dr}} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}_0(X_i) + \frac{1 - R_i}{1 - \hat{\pi}(X_i)} \left(S_i - \hat{\mu}_0(X_i) \right) \right].$$

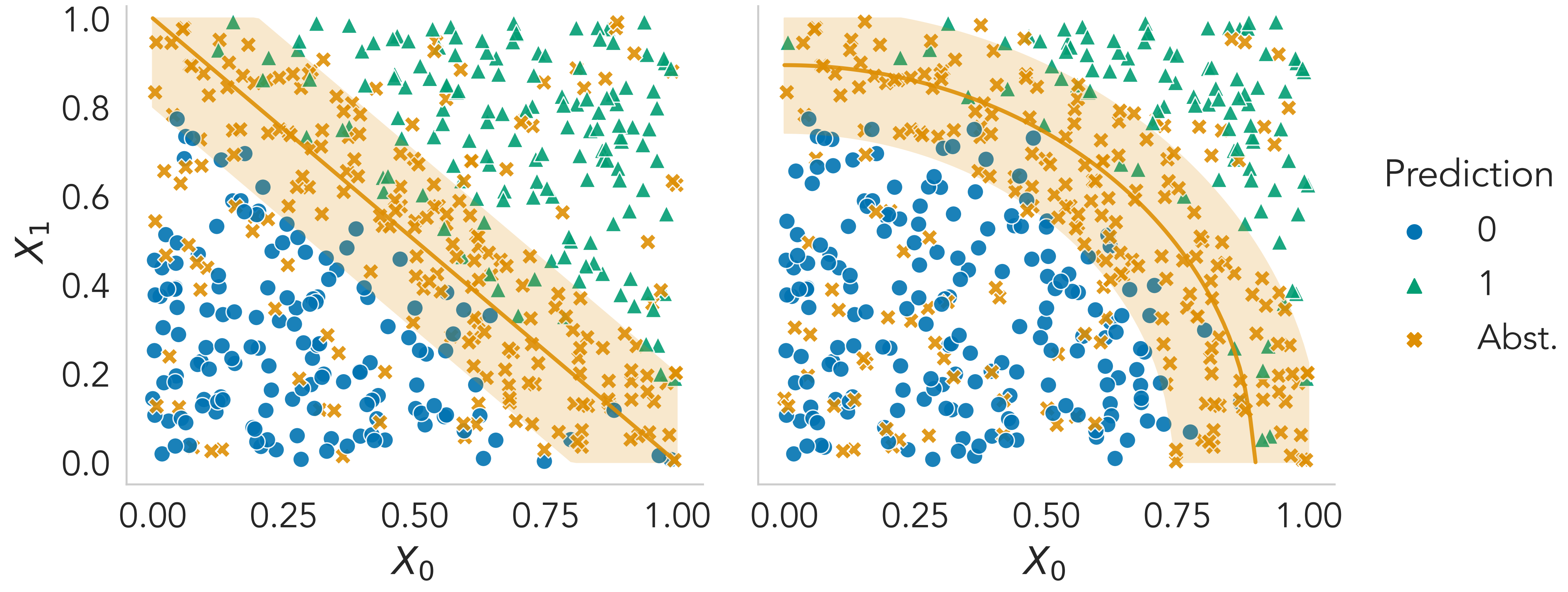
The summand is the **influence function** for $\mathbb{E}[\mu_0(X)]$ (a first-order bias correction).

For comparison, we can simply take the difference between the two classifiers ($\hat{\psi}^{A}_{\mathrm{dr}} - \hat{\psi}^{B}_{\mathrm{dr}}$).

*The nuisance functions, $\hat{\mu}_0$ and $\hat{\pi}$, are estimated via **cross-fitting** (K-fold sample splitting).

Simulated Experiment: CI Miscoverage & Width

A: linear classifier with the <u>*optimal*</u> decision boundary. **B:** <u>*biased*</u> classifier with a curved boundary.



Two abstaining classifiers, depicted using their decision boundary (orange), predictions (●/▲), and abstentions (x).

$\hat{\pi}$ / $\hat{\mu}_0$	*95% CI's*	Plug-in	IPW	DR
Random Forest	*Miscoverage*	0.64	0.14	**0.05**
	Width	*0.02*	*0.13*	*0.07*
Super Learner	*Miscoverage*	0.91	**0.03**	**0.05**
	Width	*0.01*	*0.12*	*0.06*

CI Miscoverage: rate of the 95% CI not covering the true Δ^{AB}, based on *accuracy*.
(**Blue**: valid miscoverage.)
Width: upper minus lower confidence bound.
Both averaged over 1,000 repeated simulations.

DR CI achieves the correct miscoverage rate **(small bias)**, and its width is half the width of the IPW CI **(small variance)**.

Real Data Experiment: Comparing VGG-16 Classifiers on CIFAR-100

- **Setup:** We compare abstaining classifiers based off of a pre-trained VGG-16 deep convolutional neural network* for the CIFAR-100 dataset. Evaluation set size is 5,000.
- Nuisance functions $(\hat{\pi}^{\mathrm{A}}, \hat{\mu}_0^{\mathrm{A}}, \hat{\pi}^{\mathrm{B}}, \hat{\mu}_0^{\mathrm{B}})$ are learned on top of the pre-trained VGG-16 network, but they each use a different output layer (learned via cross-fitting).

Scenarios	Base Classifier	Abstention Rule	$\bar{\Delta}^{\mathrm{AB}}$	95% DR CI	Reject H_0?
I	Same	Different	0.000	(-0.005, 0.018)	No
II	Same	Different	0.000	(-0.014, 0.008)	No
III	Different	Same	−0.029	(-0.051, -0.028)	Yes

Comparing VGG-16-Based Abstaining Classifiers on CIFAR-100 (n=5,000) using the Brier score.
Estimation target: $\Delta^{\mathrm{AB}} := \psi^{\mathrm{A}} - \psi^{\mathrm{B}}$; null hypothesis $\mathrm{H}_0 : \Delta^{\mathrm{AB}} = 0$.

*reproduced version from https://github.com/chenyaofo/pytorch-cifar-models.

Summary of Contributions

- We propose the ***counterfactual score***, a novel evaluation metric for black-box abstaining classifiers that assess the expected score had the classifier not been allowed to abstain.
- The score and its framework reveals an **underexplored connection** between abstaining classifiers, black-box evaluation, and missing data / causal inference.
- We formalize the **identifying assumptions (MAR and positivity)** for the score and give examples of settings in which they can be justified.
- We develop **nonparametrically efficient** estimators for the counterfactual score (difference), and empirically show their validity & efficiency on simulated/real datasets.

Thank You

Paper: https://arxiv.org/abs/2305.10564
Code: https://github.com/yjchoe/ComparingAbstainingClassifiers
NeurIPS Link: https://neurips.cc/virtual/2023/poster/72515
YJ's Webpage (for links to slides & poster): https://yjchoe.github.io/

Poster Session: **Tuesday Evening** (5:15-7:15pm CT on December 12th) / Poster #: **1618**