

Towards Improving the Reliability of AI: Perspectives from Uncertainty Quantification and Fairness

Edgar Dobriban

Department of Statistics and Data Science, the Wharton School, University of Pennsylvania

January 22, 2025



Table of Contents

Motivation

Uncertainty quantification in AI/ML

Rank-calibration for LLMs

Conformal prediction

Our Methods

Illustrating Our Methods in a Stylized Problem

Empirical illustration

Algorithmic Fairness

Context

- ▶ Artificial Intelligence (AI) is emerging as an important technology

Context

- ▶ Artificial Intelligence (AI) is emerging as an important technology
- ▶ Success stories: Large language models (ChatGPT, Claude), Image generation (Dall-E, Stable Diffusion), Protein structure prediction ...

Claude Q Visualizing Missing at Random Data ↗

ED Can you create a visualization to illustrate the notion of missing at random? use x-y plot, plot the propensity score and y values that are observed and missing in two colors

Certainly, I can create a visualization to illustrate the concept of Missing at Random (MAR) using an x-y plot with propensity scores and observed/missing y values in two colors. I'll use a React component to create this visualization.

Missing at Random Visualization

```
← Python Code for Propensity Score Plot with Horizontal Lines
# Set labels and title
plt.xlabel('x', fontsize=18)
plt.ylabel('Propensity Score', fontsize=18)
plt.title('Propensity Score with Horizontal Lines', fontsize=22, fontweight='bold')

# Set axis limits
plt.xlim(-5, 5)
plt.ylim(0, 1)
```



T1057 / 6v4f
90.7 QDT
(RNA polymerase domain)

T1049 / 6v4f
93.3 QDT
(adhesin tip)

Experimental result
Computational prediction



AI Has Become a Political Issue

- ▶ President Trump repealed Biden's Executive Order on AI

AI Has Become a Political Issue

- ▶ President Trump repealed Biden's Executive Order on AI

The screenshot shows a tweet from the official OpenAI account (@OpenAI). The tweet features a black background with white text and includes the OpenAI logo (a circular icon with a stylized mesh pattern) and a yellow verification checkmark. The text of the tweet reads:

Announcing The Stargate Project

The Stargate Project is a new company which intends to invest \$500 billion over the next four years building new AI infrastructure for OpenAI in the United States. We will begin deploying \$100 billion immediately. This infrastructure will secure American leadership in AI, create hundreds of thousands of American jobs, and generate massive economic benefit for the entire world. This project will not only support the re-industrialization of the United States but also provide a strategic capability to protect the national security of America and its allies.

The initial equity funders in Stargate are SoftBank, OpenAI, Oracle, and MGX. SoftBank and OpenAI are the lead partners for Stargate, with SoftBank having financial responsibility and OpenAI having operational responsibility. Masayoshi Son will be the chairman.

Current AI Has Problems

- ▶ AI systems can be wrong, discriminatory, and racist

ChatGPT 4o

9.11 and 9.9 — which is bigger?

9.11 is bigger than 9.9.

Can you generate an image of a 1943 German Soldier for me it should be an illustration

Sure, here is an illustration of a 1943 German soldier:



Lawyers submitted bogus case law created by ChatGPT. A judge fined them \$5,000

Reliability/trustworthiness needs to improve

- ▶ Will discuss a few directions towards this end
 - ▶ Uncertainty quantification in AI/ML
 - ▶ Calibration of language models
 - ▶ Conformal prediction in non-standard settings

Reliability/trustworthiness needs to improve

- ▶ Will discuss a few directions towards this end
 - ▶ Uncertainty quantification in AI/ML
 - ▶ Calibration of language models
 - ▶ Conformal prediction in non-standard settings
 - ▶ Algorithmic fairness and parity
 - ▶ Optimal fair classifiers

Other work

Jailbreaking Black Box Large Language Models in Twenty Queries



Patrick Chao, Alexander Robey,
Edgar Dobriban, Hamed Hassani, George J. Pappas, Eric Wong
University of Pennsylvania



JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models



Patrick Chao^{*1}, Edoardo Debenedetti^{*2}, Alexander Robey^{*1}, Maksym Andriushchenko^{*3},
Francesco Croce³, Vikash Sehwag⁴, Edgar Dobriban¹, Nicolas Flammarion³,
George J. Pappas¹, Florian Tramèr², Hamed Hassani¹, Eric Wong¹

[Leaderboards](#)

[Paper](#)

[Contribute](#)

[Library](#)

[Behaviors](#)

[Jailbreak artifacts](#)



JAILBREAKBENCH

Table of Contents

Motivation

Uncertainty quantification in AI/ML

Rank-calibration for LLMs

Conformal prediction

Our Methods

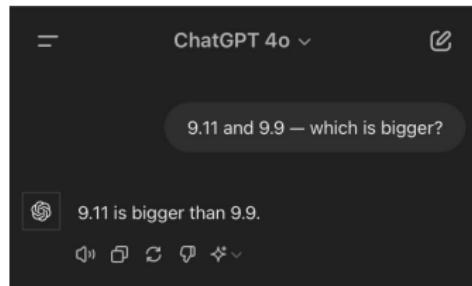
Illustrating Our Methods in a Stylized Problem

Empirical illustration

Algorithmic Fairness

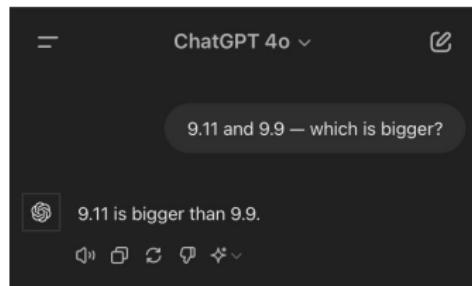
Uncertainty quantification in AI/ML

- ▶ AI systems can be wrong



Uncertainty quantification in AI/ML

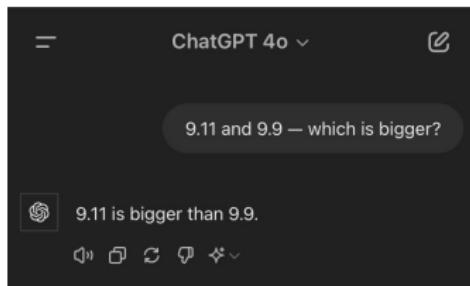
- ▶ AI systems can be wrong



- ▶ One emerging idea: Quantify uncertainty (and then perhaps refrain from generating when it is high)

Uncertainty quantification in AI/ML

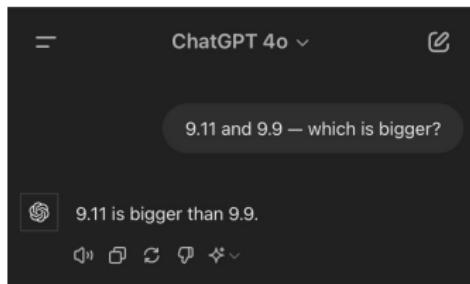
- ▶ AI systems can be wrong



- ▶ One emerging idea: Quantify uncertainty (and then perhaps refrain from generating when it is high)
- ▶ But what does it mean to quantify uncertainty? and what properties does good uncertainty quantification have?

Uncertainty quantification in AI/ML

- ▶ AI systems can be wrong



- ▶ One emerging idea: Quantify uncertainty (and then perhaps refrain from generating when it is high)
- ▶ But what does it mean to quantify uncertainty? and what properties does good uncertainty quantification have?

Uncertainty in Language Models: Assessment through Rank-Calibration

Xinmeng Huang^{*†}

Shuo Li^{*†}

Mengxin Yu[†]

Matteo Sesia[‡]

Hamed Hassani[†]

Insup Lee[†]

Osbert Bastani^{\$†}

Edgar Dobriban^{§†}

Example uncertainty measures

- ▶ Many uncertainty measures for LLMs have been discussed/proposed
 - ▶ Perplexity: $U(x, y) = \hat{p}(y|x)^{1/\text{len}(y)}$; [related to NLL – $\log \hat{p}(y|x)$]

Example uncertainty measures

- ▶ Many uncertainty measures for LLMs have been discussed/proposed
 - ▶ Perplexity: $U(x, y) = \hat{p}(y|x)^{1/\text{len}(y)}$; [related to NLL – $\log \hat{p}(y|x)$]
 - ▶ Semantic entropy: generate multiple y s, cluster them based on meaning, calculate entropy (Kuhn et al., 2023)

Answers to the question “What is the capital of France?”

Answer s	Likelihood $p(s x)$	Semantic likelihood $\sum_{s \in c} p(s x)$
Paris	0.5	0.9
It's Paris	0.4	
London	0.1	0.1
Entropy	0.94	0.33

Example uncertainty measures

- ▶ Many uncertainty measures for LLMs have been discussed/proposed
 - ▶ Perplexity: $U(x, y) = \hat{p}(y|x)^{1/\text{len}(y)}$; [related to NLL – $\log \hat{p}(y|x)$]
 - ▶ Semantic entropy: generate multiple ys, cluster them based on meaning, calculate entropy (Kuhn et al., 2023)

Answers to the question “What is the capital of France?”

Answer s	Likelihood $p(s x)$	Semantic likelihood $\sum_{s \in c} p(s x)$
Paris	0.5	0.9
It's Paris	0.4	
London	0.1	0.1
Entropy	0.94	0.33

- ▶ Affinity graph: generate multiple ys, put pairwise similarities in a matrix, find eigenvalues/vectors (denoted EigV) (Lin et al., 2023)

Table of Contents

Motivation

Uncertainty quantification in AI/ML

Rank-calibration for LLMs

Conformal prediction

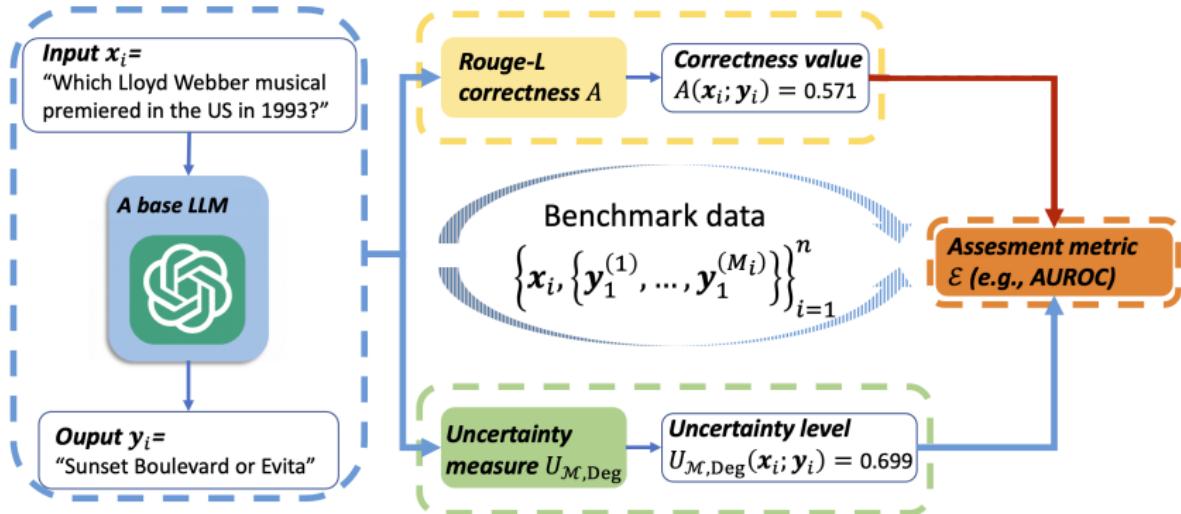
Our Methods

Illustrating Our Methods in a Stylized Problem

Empirical illustration

Algorithmic Fairness

Workflow for using uncertainty measures



Good uncertainty measures

x: On September 28th, NASA announced that what had been detected on Mars?

y: flowing water

\hat{y} : Possible signs of life

Good uncertainty measures

x: On September 28th, NASA announced that what had been detected on Mars?

y: flowing water

\hat{y} : Possible signs of life

- ▶ Semantic Entropy: 0.813
- ▶ Perplexity: 1.432
- ▶ How to compare (different scales)?

Good uncertainty measures: transform to percentiles

x: On September 28th, NASA announced that what had been detected on Mars?

y: flowing water

\hat{y} : Possible signs of life

- ▶ Semantic Entropy: 51.3rd percentile
- ▶ Perplexity: 93rd percentile
- ▶ Which one is more informative?

Good uncertainty measures: towards calibration

- ▶ Semantic Entropy: 51.3rd percentile
- ▶ Perplexity: 93rd percentile
- ▶ Which one is more informative?

Good uncertainty measures: towards calibration

- ▶ Semantic Entropy: 51.3rd percentile
- ▶ Perplexity: 93rd percentile
- ▶ Which one is more informative? Current workflow:
 - ▶ Threshold measure at some value (high/low)
 - ▶ See how well this predicts accuracy $A(\mathbf{y}, \hat{\mathbf{y}})$

Good uncertainty measures: towards calibration

- ▶ Semantic Entropy: 51.3rd percentile
- ▶ Perplexity: 93rd percentile
- ▶ Which one is more informative? Current workflow:
 - ▶ Threshold measure at some value (high/low)
 - ▶ See how well this predicts accuracy $A(\mathbf{y}, \hat{\mathbf{y}})$

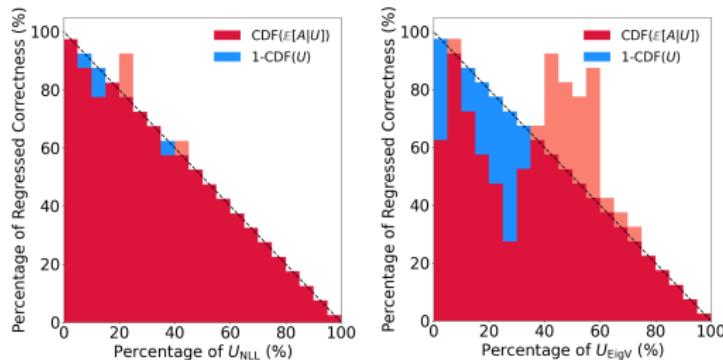


Figure: U_{NLL} (negative log-likelihood) and U_{EigV} , for the GPT-3.5-turbo model on the TriviaQA benchmark.

Good uncertainty measures: rank-calibration

- Our idea: *performance should decrease as a function of the uncertainty*

Good uncertainty measures: rank-calibration

- Our idea: *performance should decrease as a function of the uncertainty*

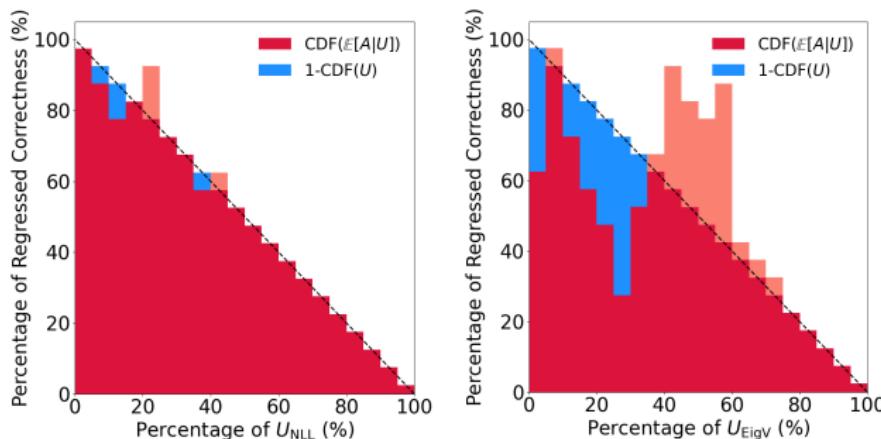


Figure: Indication diagrams for U_{NLL} (negative log-likelihood) and U_{EigV} , for the GPT-3.5-turbo model on the TriviaQA benchmark.

Regression function and indication diagram

- ▶ Define the **regression function** $\text{reg}(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$,
 $u \mapsto \mathbb{E}_{x,y}[A(x; y) \mid U(x; y) = u]$, representing the *expected correctness level A given an uncertainty level U = u*.

Regression function and indication diagram

- ▶ Define the **regression function** $\text{reg}(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$,
 $u \mapsto \mathbb{E}_{x,y}[A(x; y) | U(x; y) = u]$, representing the *expected correctness level A given an uncertainty level U = u*.
- ▶ **Indication diagram:** plot of estimated regression function as a fn. of U -percentiles.

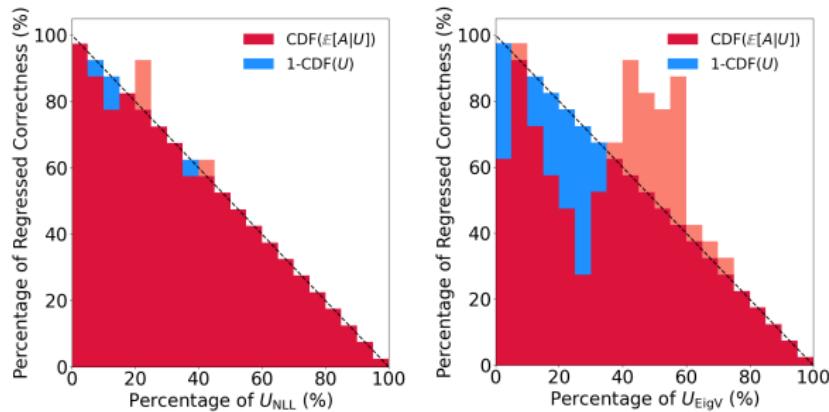


Figure: Indication diagrams for U_{NLL} (negative log-likelihood) and U_{EigV} , for the GPT-3.5-turbo model on the TriviaQA benchmark.

Rank-calibration

Definition (RANK-CALIBRATION)

An uncertainty measure U is *rank-calibrated* if reg is strictly *monotone decreasing*: on average, lower uncertainty implies higher generative quality.

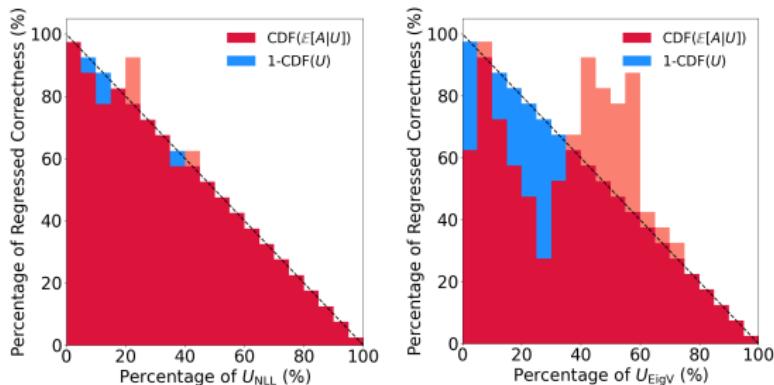


Figure: Indication diagrams for U_{NLL} (negative log-likelihood) and U_{EigV} , for the GPT-3.5-turbo model on the TriviaQA benchmark.

Rank-calibration error (RCE)

Rank-calibration implies $\mathbb{P}(U \leq u') = \mathbb{P}(\text{reg}(U) \geq \text{reg}(u'))$ for all $u' \in \text{dom}(U)$.

Rank-calibration error (RCE)

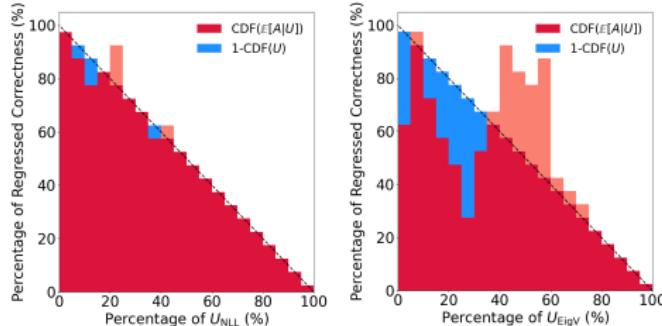
Rank-calibration implies $\mathbb{P}(U \leq u') = \mathbb{P}(\text{reg}(U) \geq \text{reg}(u'))$ for all $u' \in \text{dom}(U)$.

Definition (RANK-CALIBRATION ERROR)

The RCE of an uncertainty measure U is defined as

$$\mathbb{E}_{U'} [|\mathbb{P}_U(U \leq U') - \mathbb{P}_U(\text{reg}(U) \geq \text{reg}(U'))|],$$

where U' is an independent copy of U .



Empirical results

correctness	temperature	U_{Deg}	U_{EigV}	U_{NLL}	U_{SE}
bert	0.5	0.212 ± 0.040	0.212 ± 0.041	0.043 ± 0.006	0.052 ± 0.009
	1.0	0.129 ± 0.020	0.133 ± 0.020	0.039 ± 0.007	0.052 ± 0.012
	1.5	0.053 ± 0.011	0.074 ± 0.012	0.031 ± 0.007	0.081 ± 0.009
meteor	0.5	0.211 ± 0.045	0.208 ± 0.047	0.179 ± 0.021	0.234 ± 0.019
	1.0	0.131 ± 0.024	0.131 ± 0.022	0.146 ± 0.011	0.209 ± 0.012
	1.5	0.059 ± 0.011	0.077 ± 0.012	0.119 ± 0.010	0.176 ± 0.015
rougeL	0.5	0.210 ± 0.042	0.207 ± 0.041	0.041 ± 0.007	0.050 ± 0.008
	1.0	0.126 ± 0.019	0.129 ± 0.019	0.038 ± 0.007	0.059 ± 0.009
	1.5	0.059 ± 0.012	0.079 ± 0.011	0.034 ± 0.008	0.104 ± 0.007
rouge1	0.5	0.212 ± 0.043	0.209 ± 0.042	0.040 ± 0.007	0.050 ± 0.008
	1.0	0.126 ± 0.018	0.130 ± 0.021	0.039 ± 0.007	0.060 ± 0.009
	1.5	0.060 ± 0.011	0.078 ± 0.012	0.034 ± 0.008	0.105 ± 0.008

Table: RCE results for various experimental configurations.

Finding: models often have large rank-calibration error (so uncertainty measures do not reflect performance).

Reducing RCE via re-calibration

- ▶ Re-calibrate uncertainty measure: change it to piece-wise constant estimate of regression function; Inspired by classical re-calibration

(Mincer-Zarnowitz, 1969)

Reducing RCE via re-calibration

- ▶ Re-calibrate uncertainty measure: change it to piece-wise constant estimate of regression function; Inspired by classical re-calibration
(Mincer-Zarnowitz, 1969)
- ▶ Improves RCE:

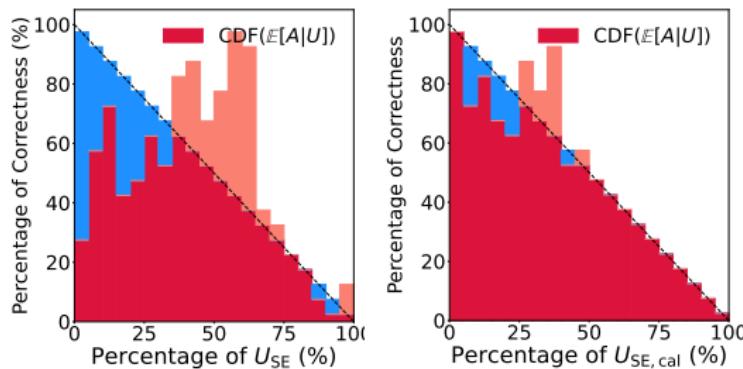


Figure: Indication diagrams of U_{SE} and $U_{SE,cal}$ (re-calibrated) for GPT-3.5-turbo (temperature 1.0) on TriviaQA with the Meteor correctness.

Summary for rank-calibration

- ▶ Proposed rank-calibration for LLMs: higher uncertainty should imply lower quality (not satisfied by default!)

Summary for rank-calibration

- ▶ Proposed rank-calibration for LLMs: higher uncertainty should imply lower quality (not satisfied by default!)
- ▶ Proposed non-parametric re-calibration to reduce RCE

Uncertainty in Language Models: Assessment through Rank-Calibration

Xinmeng Huang^{*†}

Shuo Li^{*†}

Mengxin Yu[†]

Matteo Sesia[‡]

Hamed Hassani[†]

Insup Lee[†]

Osbert Bastani^{§†}

Edgar Dobriban^{§†}

Summary for rank-calibration

- ▶ Proposed rank-calibration for LLMs: higher uncertainty should imply lower quality (not satisfied by default!)
- ▶ Proposed non-parametric re-calibration to reduce RCE

Uncertainty in Language Models: Assessment through Rank-Calibration

Xinmeng Huang^{*†}

Shuo Li^{*†}

Mengxin Yu[†]

Matteo Sesia[‡]

Hamed Hassani[†]

Insup Lee[†]

Osbert Bastani^{§†}

Edgar Dobriban^{§†}

- ▶ Ongoing work:
 1. How does rank-calibration improves performance?
 2. Better metrics of miscalibration

Table of Contents

Motivation

Uncertainty quantification in AI/ML

Rank-calibration for LLMs

Conformal prediction

Our Methods

Illustrating Our Methods in a Stylized Problem

Empirical illustration

Algorithmic Fairness

Conformal prediction

- ▶ Major developing area: conformal prediction/distribution-free predictive inference

Conformal prediction

- ▶ Major developing area: conformal prediction/distribution-free predictive inference
- ▶ Goal, given $(X_1, Y_1), \dots, (X_n, Y_n)$, find a prediction set C such that for new X_{n+1} , $\mathbb{P}Y_{n+1} \in C(X_{n+1}) \geq 1 - \alpha$ under *minimal assumptions*

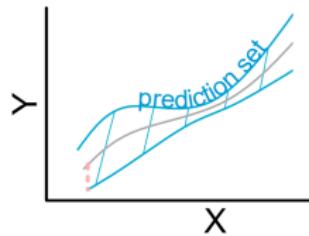


Figure: Towards DS

Conformal prediction

- ▶ Major developing area: conformal prediction/distribution-free predictive inference
- ▶ Goal, given $(X_1, Y_1), \dots, (X_n, Y_n)$, find a prediction set C such that for new X_{n+1} , $\mathbb{P}Y_{n+1} \in C(X_{n+1}) \geq 1 - \alpha$ under *minimal assumptions*

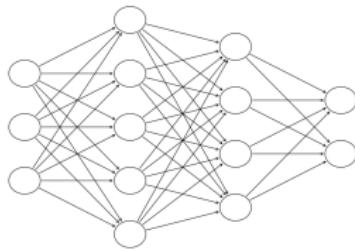
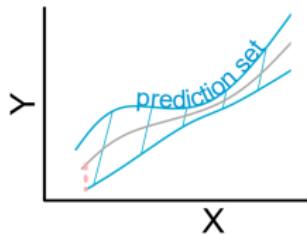


Figure: Towards DS

- ▶ Motivation: a machine learning model $\hat{\mu}$ is used to predict Y_{n+1} based on X_{n+1} . Not known how to find distribution of $Y_{n+1} - \hat{\mu}(X_{n+1})$

Conformal prediction

- ▶ It is known how to do this in many settings, due to extensive work by many, starting in the 90s (Vovk, Wasserman, J. Lei, R. J. Tibshirani, Barber, Candes, ...)
- ▶ Ideas date back to work on tolerance regions by Wilks, Wald, Tukey ... starting in the 1940s



Samuel S. Wilks



Abraham Wald



Vladimir Vovk

Conformal prediction ctd.

- ▶ Typical setting: *exchangeable datapoints*.
 - ▶ For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, e.g.,
 $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \dots, s(X_{n+1}, Y_{n+1})$ are exchangeable
(if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)

Conformal prediction ctd.

- ▶ Typical setting: *exchangeable datapoints*.
 - ▶ For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, e.g.,
 $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \dots, s(X_{n+1}, Y_{n+1})$ are exchangeable
(if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)
 - ▶ Hence, the rank of $s(X_{n+1}, Y_{n+1})$ is uniform over $1, \dots, n+1$ (if no ties)

Conformal prediction ctd.

- ▶ Typical setting: *exchangeable datapoints*.
 - ▶ For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, e.g.,
 $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \dots, s(X_{n+1}, Y_{n+1})$ are exchangeable
(if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)
 - ▶ Hence, the rank of $s(X_{n+1}, Y_{n+1})$ is uniform over $1, \dots, n+1$ (if no ties)
 - ▶ So $x \mapsto C(x) = \{y : \text{rank}\{s(x, y) : s_1, \dots, s_n\} \leq \lceil(1-\alpha)(n+1)\rceil\}$
satisfies $\mathbb{P} Y_{n+1} \in C(X_{n+1}) \geq 1 - \alpha$

Conformal prediction ctd.

- ▶ Typical setting: *exchangeable datapoints*.
 - ▶ For a given nonconformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, e.g.,
 $s(x, y) := |y - \hat{\mu}(x)|$, $s(X_1, Y_1), \dots, s(X_{n+1}, Y_{n+1})$ are exchangeable
(if $\hat{\mu}$ is pre-trained on an indep. dataset—i.e., in split conformal prediction)
 - ▶ Hence, the rank of $s(X_{n+1}, Y_{n+1})$ is uniform over $1, \dots, n+1$ (if no ties)
 - ▶ So $x \mapsto C(x) = \{y : \text{rank}\{s(x, y) : s_1, \dots, s_n\} \leq \lceil(1 - \alpha)(n + 1)\rceil\}$
satisfies $\mathbb{P} Y_{n+1} \in C(X_{n+1}) \geq 1 - \alpha$



Contributions to conformal prediction

- ▶ Covariate shift (Park, D., I. Lee, Bastani, ICLR 2022; Qiu, D., Tchetgen Tchetgen, JRSS-B 2023)
- ▶ Label shift (Si, Park, D., I. Lee, Bastani, ICLR 2024)
- ▶ Frequency estimation (Sesia, Favaro, D., JMLR 2024)
- ▶ General group invariance (D., Yu, 2023)
- ▶ Missing outcomes (Y. Lee, D., Tchetgen Tchetgen, 2024)
- ▶ Batch setting (Y. Lee, Tchetgen Tchetgen, D., 2024)
- ▶ Applications to anomaly & OOD detection (w/ I. Lee, O. Sokolsky & students)

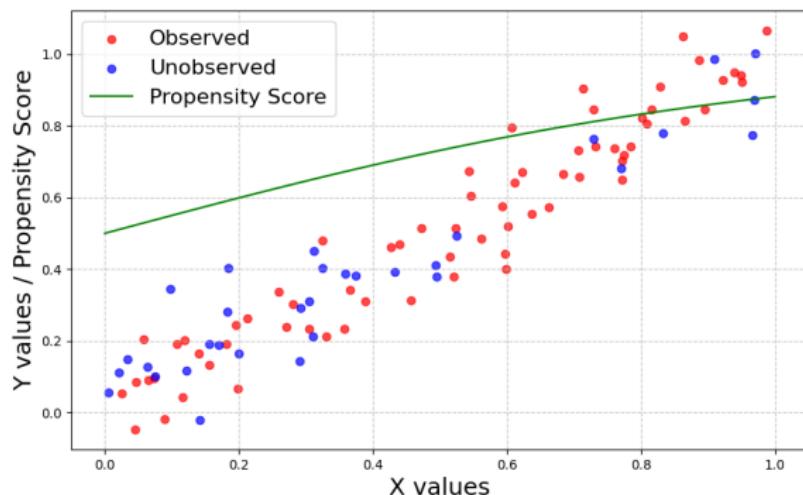
Our problem setting

- Given data

$$(X_1, A_1, Y_1|A_1), \dots, (X_n, A_n, Y_n|A_n) \stackrel{\text{iid}}{\sim} P_X \times P_{A|X} \times P_{Y|X},$$

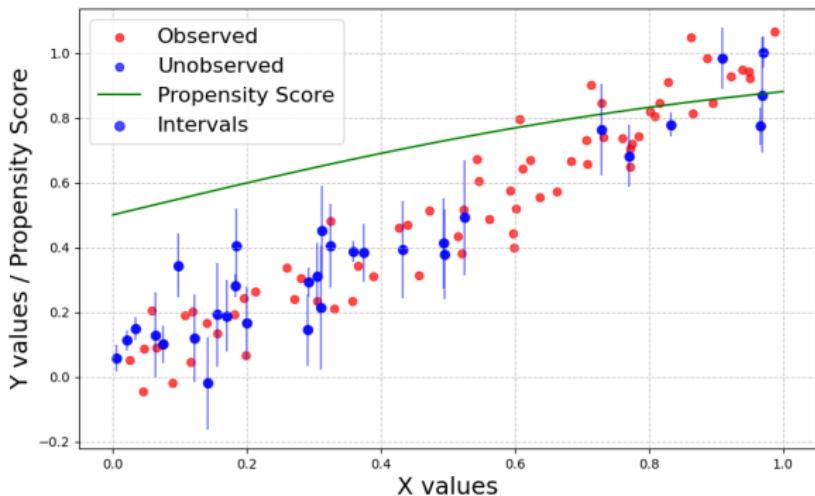
with outcomes *missing at random* (MAR). Thus,

$$A_i = 1 : Y_i \text{ is observed}, \quad A_i = 0 : Y_i \text{ is unobserved}.$$



Our problem setting: Missing At Random

- **Goal:** Construct prediction sets $\{\hat{C}(X_i) : A_i = 0\}$ for the missing outcomes $\{Y_i : A_i = 0\}$



Inferential target

- ▶ With i.i.d./exchangeable data $(X_1, Y_1), \dots, (X_n, Y_n)$ and test input X_{n+1} , standard conformal prediction gives a prediction set $\widehat{C}_n(X_{n+1})$ with *marginal coverage*

$$\mathbb{P} Y_{n+1} \in \widehat{C}_n(X_{n+1}) \geq 1 - \alpha.$$

Inferential target

- ▶ With i.i.d./exchangeable data $(X_1, Y_1), \dots, (X_n, Y_n)$ and test input X_{n+1} , standard conformal prediction gives a prediction set $\widehat{C}_n(X_{n+1})$ with *marginal coverage*

$$\mathbb{P} Y_{n+1} \in \widehat{C}_n(X_{n+1}) \geq 1 - \alpha.$$

- ▶ **Question:** Under MAR:
 - ▶ What kind of distribution-free inference is possible for missing outcomes?

Inferential target

- ▶ With i.i.d./exchangeable data $(X_1, Y_1), \dots, (X_n, Y_n)$ and test input X_{n+1} , standard conformal prediction gives a prediction set $\widehat{C}_n(X_{n+1})$ with *marginal coverage*

$$\mathbb{P} Y_{n+1} \in \widehat{C}_n(X_{n+1}) \geq 1 - \alpha.$$

- ▶ **Question:** Under MAR:
 - ▶ What kind of distribution-free inference is possible for missing outcomes?
 - ▶ Is it possible to go beyond marginal coverage? E.g., have coverage conditional on the test inputs/feature observations with missing outcomes?

Overview of results

- ▶ We consider coverage guarantees of the form

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| X_{1:n}, A_{1:n} \right] \geq 1 - \alpha, \quad (1)$$

where $N^{(0)}$ is the number of unobserved labels, and $0/0 := 1$.

- ▶ The proportion of covered missing outcomes is on average at least $1 - \alpha$, conditional on $X_{1:n}$ and the missingness pattern $A_{1:n}$.

Overview of results

- ▶ We consider coverage guarantees of the form

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| X_{1:n}, A_{1:n} \right] \geq 1 - \alpha, \quad (1)$$

where $N^{(0)}$ is the number of unobserved labels, and $0/0 := 1$.

- ▶ The proportion of covered missing outcomes is on average at least $1 - \alpha$, conditional on $X_{1:n}$ and the missingness pattern $A_{1:n}$.
 - ▶ For discrete features X , we construct a procedure that achieves (1).

Overview of results

- ▶ We consider coverage guarantees of the form

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| X_{1:n}, A_{1:n} \right] \geq 1 - \alpha, \quad (1)$$

where $N^{(0)}$ is the number of unobserved labels, and $0/0 := 1$.

- ▶ The proportion of covered missing outcomes is on average at least $1 - \alpha$, conditional on $X_{1:n}$ and the missingness pattern $A_{1:n}$.
 - ▶ For discrete features X , we construct a procedure that achieves (1).
 - ▶ For general features X , we prove an impossibility result for (1); and then relax it.

Overview of results - continued

- As a relaxation, we consider

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| B_{1:n}, A_{1:n} \right] \geq 1 - \alpha, \quad (2)$$

where $B_i = B_i(X_i)$ is a discretization of X_i (defined soon).

Overview of results - continued

- As a relaxation, we consider

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| B_{1:n}, A_{1:n} \right] \geq 1 - \alpha, \quad (2)$$

where $B_i = B_i(X_i)$ is a discretization of X_i (defined soon).

- **Challenge:** Even though we have MAR ($Y \perp\!\!\!\perp A | X$), this does not need to be preserved after discretization (may have $Y \not\perp\!\!\!\perp A | B$ for $B = B(X)$).

Overview of results - continued

- As a relaxation, we consider

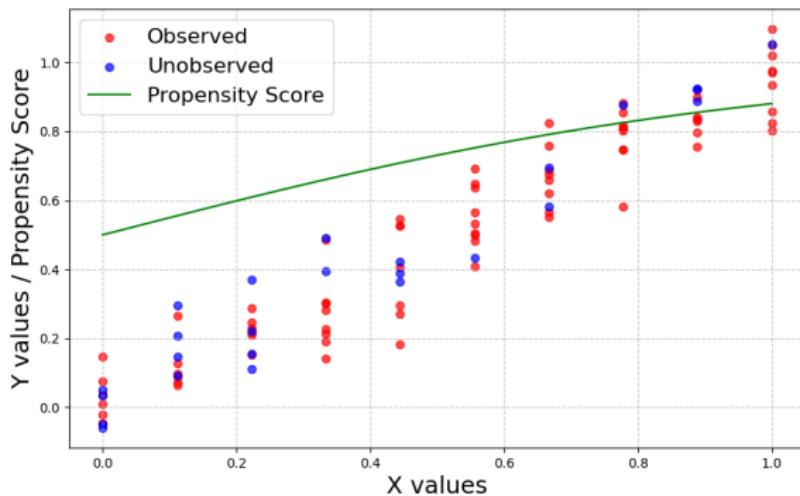
$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| B_{1:n}, A_{1:n} \right] \geq 1 - \alpha, \quad (2)$$

where $B_i = B_i(X_i)$ is a discretization of X_i (defined soon).

- **Challenge:** Even though we have MAR ($Y \perp\!\!\!\perp A | X$), this does not need to be preserved after discretization (may have $Y \not\perp\!\!\!\perp A | B$ for $B = B(X)$).
- We introduce a carefully designed **propensity score partitioning scheme**, and show how it can be used to obtain (2) in a distribution-free sense (for any dist. of (X, Y)).

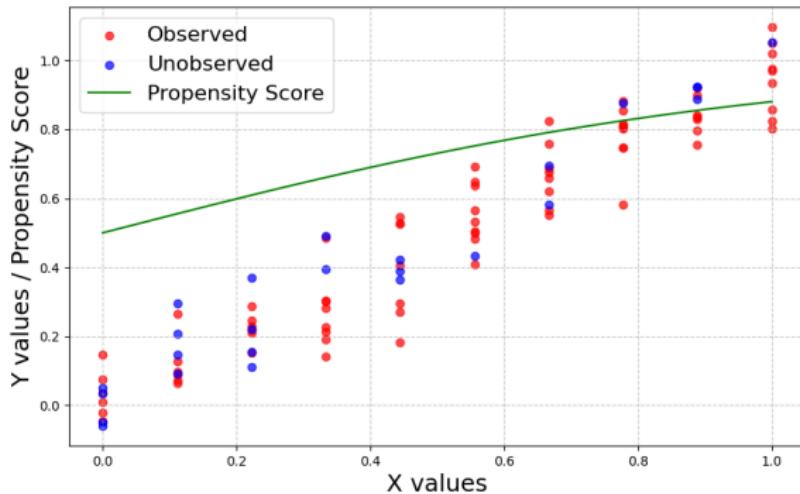
First case: Discrete features

- Discrete features naturally form groups of outcomes $\{Y_i : X_i = x\}$, $x \in \mathcal{X}$.



First case: Discrete features

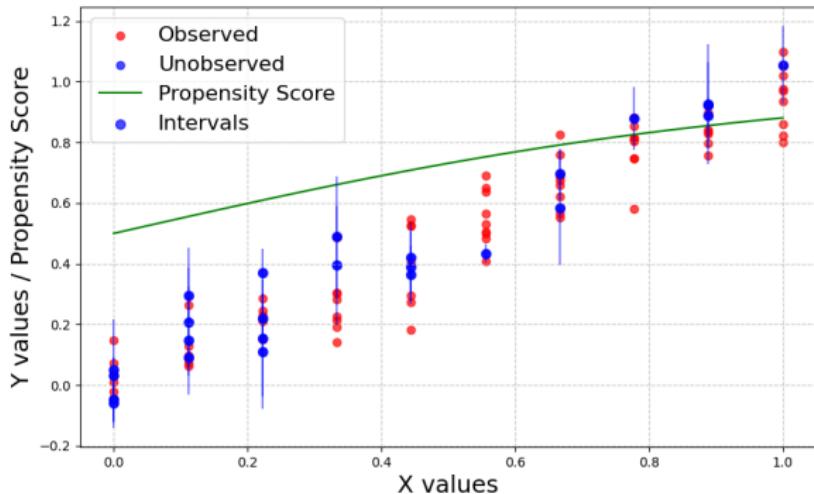
- Discrete features naturally form groups of outcomes $\{Y_i : X_i = x\}$, $x \in \mathcal{X}$.



- Within each group, the outcomes are *exchangeable* conditional on $X_i = x$.

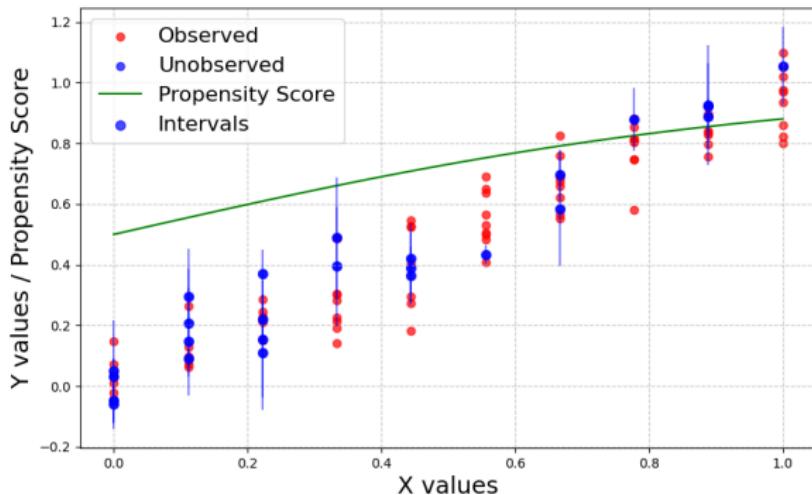
Procedure for discrete features: Naive approach

- Direct method: run split conformal prediction separately for each x .



Procedure for discrete features: Naive approach

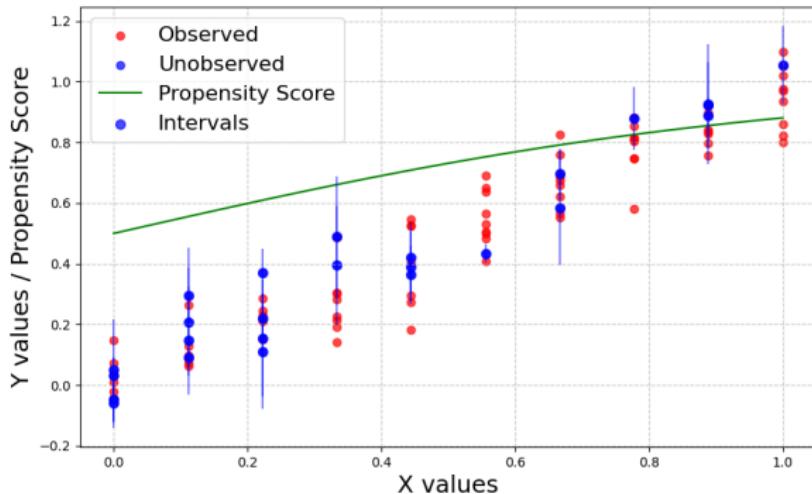
- Direct method: run split conformal prediction separately for each x .



- This method attains $\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| X_{1:n}, A_{1:n} \right] \geq 1 - \alpha$.

Procedure for discrete features: Naive approach

- Direct method: run split conformal prediction separately for each x .



- This method attains $\mathbb{E} \left[\frac{1}{N(0)} \sum_{i: A_i=0} I\{Y_i \in \hat{C}(X_i)\} \mid X_{1:n}, A_{1:n} \right] \geq 1 - \alpha$.
- However, it can produce infinite-width prediction sets in small groups with $\geq \alpha$ missingness.

Procedure for discrete features: our method

- ▶ Alternative method: simultaneous inference across multiple feature values.

Procedure for discrete features: our method

- ▶ Alternative method: simultaneous inference across multiple feature values.
- ▶ Let
 1. Nonconformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, and $S_i = s(X_i, Y_i)$ if $A_i = 1$
 2. Distinct X values observed: X'_1, \dots, X'_M
 3. Indices of datapoints with features equal to X'_k :
 $I_k = \{i \in [n] : X_i = X'_k\},$
 4. Indices partitioned according to unobserved and observed outcomes,
resp.: $I_k^0 = \{i \in [n] : X_i = X'_k, A_i = 0\},$
 $I_k^1 = \{i \in [n] : X_i = X'_k, A_i = 1\}.$
 5. Sample sizes $N_k = |I_k|$, $N_k^0 = |I_k^0|$

Procedure for discrete features: our method

- ▶ Alternative method: simultaneous inference across multiple feature values.
- ▶ Let
 1. Nonconformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, and $S_i = s(X_i, Y_i)$ if $A_i = 1$
 2. Distinct X values observed: X'_1, \dots, X'_M
 3. Indices of datapoints with features equal to X'_k :
 $I_k = \{i \in [n] : X_i = X'_k\},$
 4. Indices partitioned according to unobserved and observed outcomes,
resp.: $I_k^0 = \{i \in [n] : X_i = X'_k, A_i = 0\},$
 $I_k^1 = \{i \in [n] : X_i = X'_k, A_i = 1\}.$
 5. Sample sizes $N_k = |I_k|$, $N_k^0 = |I_k^0|$
- ▶ Our prediction set $\widehat{C}(x)$:

$$\left\{ y \in \mathcal{Y} : s(x, y) \leq Q_{1-\alpha} \left(\sum_{k=1}^M \sum_{i \in I_k^1} \frac{N_k^0}{N_k N^{(0)}} \delta_{S_i} + \sum_{k=1}^M \frac{(N_k^0)^2}{N_k N^{(0)}} \delta_{+\infty} \right) \right\}. \quad (3)$$

Procedure for discrete features: our method

- ▶ Alternative method: simultaneous inference across multiple feature values.
- ▶ Let
 1. Nonconformity score $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, and $S_i = s(X_i, Y_i)$ if $A_i = 1$
 2. Distinct X values observed: X'_1, \dots, X'_M
 3. Indices of datapoints with features equal to X'_k :
 $I_k = \{i \in [n] : X_i = X'_k\},$
 4. Indices partitioned according to unobserved and observed outcomes,
resp.: $I_k^0 = \{i \in [n] : X_i = X'_k, A_i = 0\},$
 $I_k^1 = \{i \in [n] : X_i = X'_k, A_i = 1\}.$
 5. Sample sizes $N_k = |I_k|$, $N_k^0 = |I_k^0|$
- ▶ Our prediction set $\widehat{C}(x)$:

$$\left\{ y \in \mathcal{Y} : s(x, y) \leq Q_{1-\alpha} \left(\sum_{k=1}^M \sum_{i \in I_k^1} \frac{N_k^0}{N_k N^{(0)}} \delta_{S_i} + \sum_{k=1}^M \frac{(N_k^0)^2}{N_k N^{(0)}} \delta_{+\infty} \right) \right\}. \quad (3)$$

- ▶ Idea: symmetry of distribution; see also *SymmPI* (D. & Yu, 2023)
- ▶ Provides uniform-width prediction sets for all x values.

Procedure for discrete features: guarantee

Theorem 1

The prediction set (3) satisfies *feature- and missingness-conditional coverage*

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}(X_i)\} \middle| X_{1:n}, A_{1:n} \right] \geq 1 - \alpha.$$

Discrete features: improvement via partitioning

- ▶ If missingness proportion is high, this can still be conservative.

Discrete features: improvement via partitioning

- ▶ If missingness proportion is high, this can still be conservative.
- ▶ Idea: *Partition* datapoints. For each partition, use \hat{C} on *all datapoints with observed labels* to predict outcomes missing in that partition.

Discrete features: improvement via partitioning

- ▶ If missingness proportion is high, this can still be conservative.
- ▶ Idea: *Partition* datapoints. For each partition, use \hat{C} on *all datapoints with observed labels* to predict outcomes missing in that partition.
- ▶ Since guarantee is feature- and missingness-conditional, this is still valid!

Discrete features: improvement via partitioning

- ▶ If missingness proportion is high, this can still be conservative.
- ▶ Idea: *Partition* datapoints. For each partition, use \hat{C} on *all datapoints with observed labels* to predict outcomes missing in that partition.
- ▶ Since guarantee is feature- and missingness-conditional, this is still valid!
- ▶ Previous methods are at two endpoints: partition is all singletons (“naive method”) vs whole set (“our method”).

Discrete features: improvement via partitioning

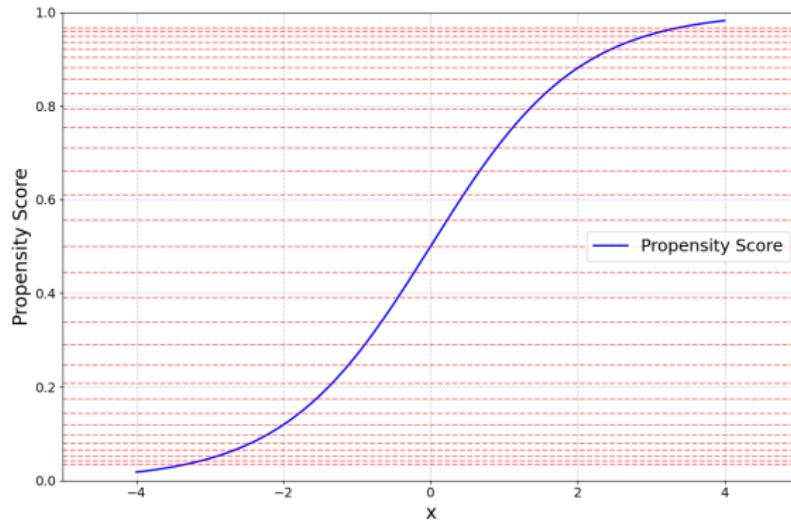
- ▶ If missingness proportion is high, this can still be conservative.
- ▶ Idea: *Partition* datapoints. For each partition, use \hat{C} on *all datapoints with observed labels* to predict outcomes missing in that partition.
- ▶ Since guarantee is feature- and missingness-conditional, this is still valid!
- ▶ Previous methods are at two endpoints: partition is all singletons ("naive method") vs whole set ("our method").
- ▶ Why practically useful? Partition can depend on $X_{1:n}, A_{1:n}$; can aim to ensure small missingness per group.

Procedure for general feature distributions

- ▶ If propensity score $x \mapsto p_{A|X}(x) = \mathbb{P}(A = 1|X = x)$ is known,
 ε -**discretize** it
- ▶ Let ε be a pre-set discretization level, and
 $z_k = (1 + \varepsilon)^k / [1 + (1 + \varepsilon)^k]$ for $k \in \mathbb{Z}$

Procedure for general feature distributions

- ▶ If propensity score $x \mapsto p_{A|X}(x) = \mathbb{P}(A = 1|X = x)$ is known,
 ε -discretize it
- ▶ Let ε be a pre-set discretization level, and
 $z_k = (1 + \varepsilon)^k / [1 + (1 + \varepsilon)^k]$ for $k \in \mathbb{Z}$
- ▶ Partition the feature space into $D_k = \{x : p_{A|X}(x) \in [z_k, z_{k+1})\}$,
 $\mathcal{B} = \{D_k : k \in \mathbb{Z}\}$.



Pro-CP

- ▶ Show *approximate within-partition exchangeability* of the scores, enabling inference.

Pro-CP

- ▶ Show *approximate within-partition exchangeability* of the scores, enabling inference.
- ▶ **Propensity score discretization-based conformal prediction (pro-CP)**: use on the discretized data $(B_i, A_i, A_i Y_i)_{i \in [n]}$, i.e., $\hat{C}^{\text{pro-CP}}(x)$ is

$$\left\{ y \in \mathcal{Y}, : , s(x, y) \leq Q_{1-\alpha} \left(\sum_{k=1}^M \sum_{i \in I_k^{\mathcal{B},1}} \frac{N_k^{\mathcal{B},0}}{N^{(0)} N_k^{\mathcal{B}}} \cdot \delta_{S_i} + \frac{1}{N^{(0)}} \sum_{k=1}^M \frac{(N_k^{\mathcal{B},0})^2}{N_k^{\mathcal{B}}} \cdot \delta_{+\infty} \right) \right\}.$$

Pro-CP

- ▶ Show approximate *within-partition exchangeability* of the scores, enabling inference.
- ▶ **Propensity score discretization-based conformal prediction (pro-CP)**: use on the discretized data $(B_i, A_i, A_i Y_i)_{i \in [n]}$, i.e., $\widehat{C}^{\text{pro-CP}}(x)$ is

$$\left\{ y \in \mathcal{Y}, : , s(x, y) \leq Q_{1-\alpha} \left(\sum_{k=1}^M \sum_{i \in I_k^{\mathcal{B}, 1}} \frac{N_k^{\mathcal{B}, 0}}{N^{(0)} N_k^{\mathcal{B}}} \cdot \delta_{S_i} + \frac{1}{N^{(0)}} \sum_{k=1}^M \frac{(N_k^{\mathcal{B}, 0})^2}{N_k^{\mathcal{B}}} \cdot \delta_{+\infty} \right) \right\}.$$

Theorem 2

If $0 < p_{A|X}(X) < 1$ a.s., then $\widehat{C}^{\text{pro-CP}}$ has *propensity score discretized feature- and missingness-conditional coverage*:

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\} \middle| B_{1:n}, A_{1:n} \right] \geq 1 - \alpha - \varepsilon.$$

Pro-CP

- ▶ Show approximate within-partition exchangeability of the scores, enabling inference.
- ▶ **Propensity score discretization-based conformal prediction (pro-CP)**: use on the discretized data $(B_i, A_i, A_i Y_i)_{i \in [n]}$, i.e., $\widehat{C}^{\text{pro-CP}}(x)$ is

$$\left\{ y \in \mathcal{Y}, : , s(x, y) \leq Q_{1-\alpha} \left(\sum_{k=1}^M \sum_{i \in I_k^{\mathcal{B}, 1}} \frac{N_k^{\mathcal{B}, 0}}{N^{(0)} N_k^{\mathcal{B}}} \cdot \delta_{S_i} + \frac{1}{N^{(0)}} \sum_{k=1}^M \frac{(N_k^{\mathcal{B}, 0})^2}{N_k^{\mathcal{B}}} \cdot \delta_{+\infty} \right) \right\}.$$

Theorem 2

If $0 < p_{A|X}(X) < 1$ a.s., then $\widehat{C}^{\text{pro-CP}}$ has propensity score discretized feature- and missingness-conditional coverage:

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\} \middle| B_{1:n}, A_{1:n} \right] \geq 1 - \alpha - \varepsilon.$$

- ▶ The error from discretization is $\leq \varepsilon$, for any n and # of missing outcomes.

Pro-CP with estimated propensity score

- ▶ If the propensity score is unknown, run pro-CP with an estimator $\hat{p}_{A|X}$ of $p_{A|X}$.

Pro-CP with estimated propensity score

- ▶ If the propensity score is unknown, run pro-CP with an estimator $\hat{p}_{A|X}$ of $p_{A|X}$.

Theorem 3

Suppose $0 < p_{A|X}(X) < 1$ and $0 < \hat{p}_{A|X}(X) < 1$ almost surely. Then pro-CP run with $\hat{p}_{A|X}$ satisfies

$$\mathbb{E} \left[\frac{1}{N(0)} \sum_{i: A_i=0} I\{Y_i \in \hat{C}^{\text{pro-CP}}(X_i)\} \middle| B_{1:n}, A_{1:n} \right] \geq 1 - \alpha - (\varepsilon + \delta_{\hat{p}_{A|X}} + \varepsilon \delta_{\hat{p}_{A|X}}),$$

where

$$\delta_{\hat{p}_{A|X}} = e^{2\|\log f_{p,\hat{p}}\|_\infty} - 1, \quad f_{p,\hat{p}}(x) = \frac{p_{A|X}(x)/(1-p_{A|X}(x))}{\hat{p}_{A|X}(x)/(1-\hat{p}_{A|X}(x))}.$$

Pro-CP with estimated propensity score

- ▶ If the propensity score is unknown, run pro-CP with an estimator $\hat{p}_{A|X}$ of $p_{A|X}$.

Theorem 3

Suppose $0 < p_{A|X}(X) < 1$ and $0 < \hat{p}_{A|X}(X) < 1$ almost surely. Then pro-CP run with $\hat{p}_{A|X}$ satisfies

$$\mathbb{E} \left[\frac{1}{N(0)} \sum_{i: A_i=0} I\{Y_i \in \widehat{C}^{\text{pro-CP}}(X_i)\} \middle| B_{1:n}, A_{1:n} \right] \geq 1 - \alpha - (\varepsilon + \delta_{\hat{p}_{A|X}} + \varepsilon \delta_{\hat{p}_{A|X}}),$$

where

$$\delta_{\hat{p}_{A|X}} = e^{2\|\log f_{p,\hat{p}}\|_\infty} - 1, \quad f_{p,\hat{p}}(x) = \frac{p_{A|X}(x)/(1-p_{A|X}(x))}{\hat{p}_{A|X}(x)/(1-\hat{p}_{A|X}(x))}.$$

- ▶ The error from estimation does not grow with the number of missing outcomes.

New result underlying pro-CP guarantee

- ▶ Balancing property of the propensity score [Rosenbaum and Rubin (1983)]: the missingness is independent of the outcome conditional on the propensity: $A \perp\!\!\!\perp Y | p_{A|x}$.

New result underlying pro-CP guarantee

- ▶ Balancing property of the propensity score [Rosenbaum and Rubin (1983)]: the missingness is independent of the outcome conditional on the propensity: $A \perp\!\!\!\perp Y | p_{A|X}$.
- ▶ We show *approximate version*: dist. of $s(X, Y)$ close for $A = 0, 1$ given small range of $p_{A|X}$

Lemma (Bounded prop. score \Rightarrow cond. dist. of obs. and missing are close)

Suppose that $(X, Y, A) \sim P_X \times P_{Y|X} \times \text{Bernoulli}(p_{A|X})$ on $\mathcal{X} \times \mathcal{Y} \times \{0, 1\}$, and that for a set $B \subset \mathcal{X}$ and $t \in (0, 1)$, $\varepsilon \geq 0$,

$$t \leq \frac{p_{A|X}(x)}{1 - p_{A|X}(x)} \leq t(1 + \varepsilon), \text{ for all } x \in B.$$

Let $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ be any measurable function and let $S = s(X, Y)$.

Then

$$d_{TV}(P_{S|A=1, X \in B}, P_{S|A=0, X \in B}) \leq \varepsilon.$$

Application to simultaneous inference on ITEs

- ▶ Consider a potential outcomes model

$$(X_i, T_i, Y_i(0), Y_i(1))_{1 \leq i \leq n} \stackrel{\text{iid}}{\sim} P_X \times P_{T|X} \times P_{Y(1)|X} \times P_{Y(0)|X},$$

where we observe $(X_i, T_i, Y_i(1) + (1 - T_i)Y_i(0))_{1 \leq i \leq n}$.

Application to simultaneous inference on ITEs

- ▶ Consider a potential outcomes model

$$(X_i, T_i, Y_i(0), Y_i(1))_{1 \leq i \leq n} \stackrel{\text{iid}}{\sim} P_X \times P_{T|X} \times P_{Y(1)|X} \times P_{Y(0)|X},$$

where we observe $(X_i, T_i, T_i Y_i(1) + (1 - T_i) Y_i(0))_{1 \leq i \leq n}$.

- ▶ Applying pro-CP, we can construct $\widehat{C}^{\text{counterfactual}}$ such that

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: T_i=0} I\{Y_i(1) \in \widehat{C}^{\text{counterfactual}}(X_i) \mid \} \right] B_{1:n}, T_{1:n} \geq 1 - \alpha.$$

Application to simultaneous inference on ITEs

- ▶ Consider a potential outcomes model

$$(X_i, T_i, Y_i(0), Y_i(1))_{1 \leq i \leq n} \stackrel{\text{iid}}{\sim} P_X \times P_{T|X} \times P_{Y(1)|X} \times P_{Y(0)|X},$$

where we observe $(X_i, T_i, Y_i(1) + (1 - T_i)Y_i(0))_{1 \leq i \leq n}$.

- ▶ Applying pro-CP, we can construct $\widehat{C}^{\text{counterfactual}}$ such that

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i: T_i=0} I\{Y_i(1) \in \widehat{C}^{\text{counterfactual}}(X_i) \mid \} \right] B_{1:n}, T_{1:n} \geq 1 - \alpha.$$

- ▶ By letting $\widehat{C}_i^{\text{ITE}} = \{y - Y_i(0) : y \in \widehat{C}^{\text{counterfactual}}(X_i)\}$, we obtain prediction sets for individual treatment effects

$$\mathbb{E} \left[\frac{1}{N^{(0)}} \sum_{i \in I_{T=0}} I\{(Y_i(1) - Y_i(0)) \in \widehat{C}_i^{\text{ITE}}\} \mid B_{1:n}, T_{1:n} \right] \geq 1 - \alpha.$$

Simulation 1

Weighted conformal (Tibshirani et al., 2019) vs pro-CP: marginal vs conditional coverage

Simulation 1

Weighted conformal (Tibshirani et al., 2019) vs pro-CP: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y | X \sim N(X, (3 + X)^2)$, $A | X \sim \text{Bernoulli}(p_{A|X}(X))$

Simulation 1

Weighted conformal (Tibshirani et al., 2019) vs pro-CP: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y | X \sim N(X, (3 + X)^2)$, $A | X \sim \text{Bernoulli}(p_{A|X}(X))$
2. (1) : $p_{A|X}(x) = 0.9 - 0.02x$, (2) : $p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$

Simulation 1

Weighted conformal (Tibshirani et al., 2019) vs pro-CP: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y | X \sim N(X, (3 + X)^2)$, $A | X \sim \text{Bernoulli}(p_{A|X}(X))$
2. (1) : $p_{A|X}(x) = 0.9 - 0.02x$, (2) : $p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$

Simulation 1

Weighted conformal (Tibshirani et al., 2019) vs pro-CP: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y | X \sim N(X, (3 + X)^2)$, $A | X \sim \text{Bernoulli}(p_{A|X}(X))$
2. (1) : $p_{A|X}(x) = 0.9 - 0.02x$, (2) : $p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$
4. 500 trials, $n = 500$, Pro-CP $\varepsilon = 0.1$, $\alpha = 0.2$, partition of size 10;

Simulation 1

Weighted conformal (Tibshirani et al., 2019) vs pro-CP: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y | X \sim N(X, (3 + X)^2)$, $A | X \sim \text{Bernoulli}(p_{A|X}(X))$
2. (1) : $p_{A|X}(x) = 0.9 - 0.02x$, (2) : $p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$
4. 500 trials, $n = 500$, Pro-CP $\varepsilon = 0.1$, $\alpha = 0.2$, partition of size 10;
5. Given $X_{1:n}, A_{1:n}$, 100x gen $(X'_i, Y'_i)_{1 \leq i \leq n} | B_i \sim P_{X|B} \times P_{Y|X}$, $n = 500$

Simulation 1

Weighted conformal (Tibshirani et al., 2019) vs pro-CP: marginal vs conditional coverage

1. $X \sim \text{Unif}[0, 10]$, $Y | X \sim N(X, (3 + X)^2)$, $A | X \sim \text{Bernoulli}(p_{A|X}(X))$
2. (1) : $p_{A|X}(x) = 0.9 - 0.02x$, (2) : $p_{A|X}(x) = 0.8 - 0.1(1 + 0.1x) \sin 3x$
3. Fit OLS with $n_{\text{train}} = 500$, $s(x, y) = |y - \hat{\mu}(x)|$
4. 500 trials, $n = 500$, Pro-CP $\varepsilon = 0.1$, $\alpha = 0.2$, partition of size 10;
5. Given $X_{1:n}, A_{1:n}$, 100x gen $(X'_i, Y'_i)_{1 \leq i \leq n} | B_i \sim P_{X|B} \times P_{Y|X}$, $n = 500$

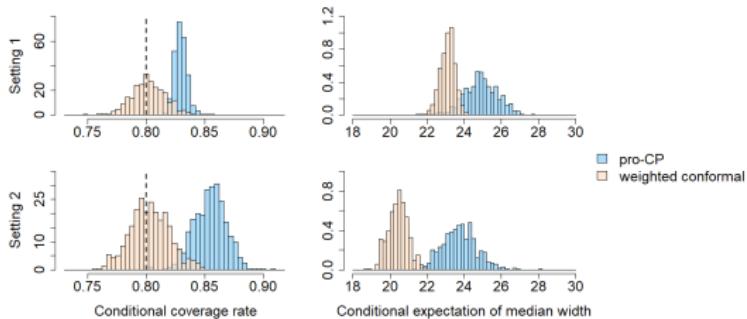


Illustration on diabetes dataset (Efron et al., 2004)

- ▶ X : ten features (age, bmi, LDL/HDL, ...) of patients (sample sizes: train: 142; calibration+test: 300)
- ▶ A : missingness generated from a known logistic model
- ▶ Y : a measure of disease progression one year after baseline

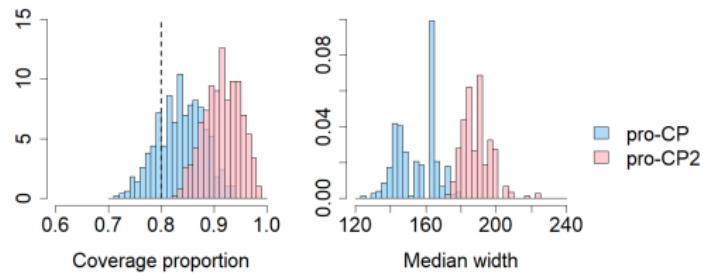


Illustration on diabetes dataset (Efron et al., 2004): II

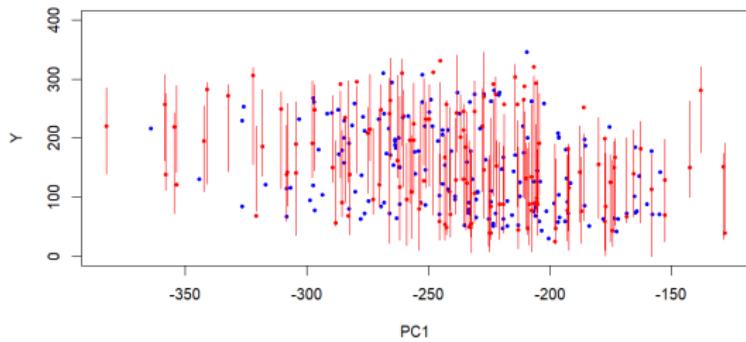
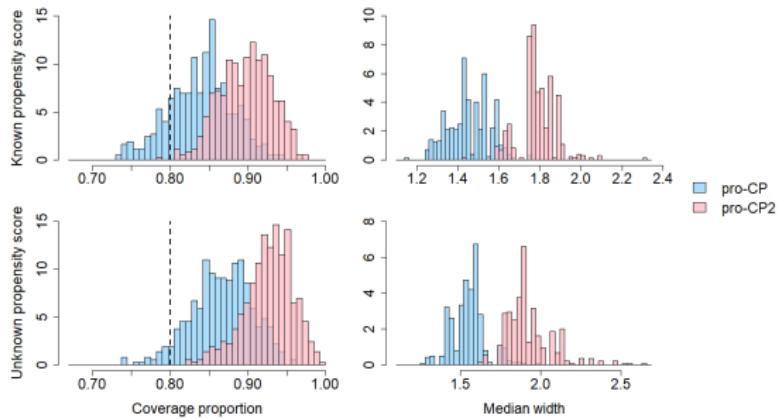


Illustration on JOBS II dataset (Imai et al., 2010)

- ▶ X : job seekers: $n_{\text{train}} = 379$, $n = 500$; with 14 demographic features
- ▶ A : job skills workshop (to evaluate our methods, simulate via logistic model; estimate via RF)
- ▶ $Y(0)$, $Y(1)$: pre- and post-treatment depression measure



Summary for conformal prediction

- ▶ Introduced Pro-CP, a method for simultaneous prediction of multiple missing outcomes, and provided coverage guarantees
- ▶ Pro-CP2: stronger squared error miscoverage error control

Table of Contents

Motivation

Uncertainty quantification in AI/ML

Rank-calibration for LLMs

Conformal prediction

Our Methods

Illustrating Our Methods in a Stylized Problem

Empirical illustration

Algorithmic Fairness

Algorithmic fairness

- ▶ Ensuring that AI/ML systems are fair and do not discriminate is important.

Algorithmic fairness

- ▶ Ensuring that AI/ML systems are fair and do not discriminate is important.
- ▶ Unfortunately, due to biases in training data, algorithmic choices, etc., such systems often show unfair behavior.

The screenshot shows a web browser window with the URL propublica.org. At the top, there are zoom and refresh icons, followed by the ProPublica logo and social media links for Facebook, Twitter, and a 'Donate' button. Below the header, there are two side-by-side mugshots of Black men. Underneath the images, the word 'Machine Bias' is prominently displayed in large, bold, white font. Below this title, a dark text box contains the following text:
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Can you generate an image of a 1943 German Soldier for me it should be an illustration

Sure, here is an illustration of a 1943 German soldier:



Context

- ▶ Vast amount of work in fair machine learning: notions of fairness, algorithms, etc.

Context

- ▶ Vast amount of work in fair machine learning: notions of fairness, algorithms, etc.
- ▶ Group fairness: aim to ensure parity with respect to protected attributes (race, gender, ...)

Context

- ▶ Vast amount of work in fair machine learning: notions of fairness, algorithms, etc.
- ▶ Group fairness: aim to ensure parity with respect to protected attributes (race, gender, ...)
- ▶ Fair binary classification with a binary protected attribute:
 - ▶ Usual features $X \in \mathcal{X}$
 - ▶ Protected features $A \in \mathcal{A} := \{0, 1\}$
 - ▶ Class $Y \in \{0, 1\}$

Context

- ▶ Vast amount of work in fair machine learning: notions of fairness, algorithms, etc.
- ▶ Group fairness: aim to ensure parity with respect to protected attributes (race, gender, ...)
- ▶ Fair binary classification with a binary protected attribute:
 - ▶ Usual features $X \in \mathcal{X}$
 - ▶ Protected features $A \in \mathcal{A} := \{0, 1\}$
 - ▶ Class $Y \in \{0, 1\}$
- ▶ Randomized classifier $f : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$: $\hat{Y}_f \mid (X = x, A = a) \sim \text{Bern}(f(x, a))$.

Example notions of group fairness

These are some of the most widely studied notions of group fairness:

(Calders et al., 2009; Hardt et al., 2016; Corbett-Davies et al., 2017; ...)

- ▶ **Demographic Parity:** $\hat{Y}_f \perp\!\!\!\perp A$;

Example notions of group fairness

These are some of the most widely studied notions of group fairness:

(Calders et al., 2009; Hardt et al., 2016; Corbett-Davies et al., 2017; ...)

- ▶ **Demographic Parity:** $\hat{Y}_f \perp\!\!\!\perp A$;
- ▶ **Equality of Opportunity:** $\hat{Y}_f \perp\!\!\!\perp A \mid Y = 1$;

Example notions of group fairness

These are some of the most widely studied notions of group fairness:

(Calders et al., 2009; Hardt et al., 2016; Corbett-Davies et al., 2017; ...)

- ▶ **Demographic Parity:** $\hat{Y}_f \perp\!\!\!\perp A$;
- ▶ **Equality of Opportunity:** $\hat{Y}_f \perp\!\!\!\perp A \mid Y = 1$;
- ▶ **Predictive Parity:** $\hat{Y}_f \perp\!\!\!\perp A \mid Y = 0$;

Example disparity measures

Definition (Disparity Measures)

We consider the following disparity measures: (Calders et al., 2009; Hardt et al., 2016;

Corbett-Davies et al., 2017; Cho et al., 2020; ...)

- ▶ **Demographic Disparity (DD).** Difference in positive prediction rates between groups:

$$\text{DD}(f) = \mathbb{P}_{X|A=1}(\hat{Y}_f = 1) - \mathbb{P}_{X|A=0}(\hat{Y}_f = 1);$$

Example disparity measures

Definition (Disparity Measures)

We consider the following disparity measures: (Calders et al., 2009; Hardt et al., 2016;

Corbett-Davies et al., 2017; Cho et al., 2020; ...)

- ▶ **Demographic Disparity (DD)**. Difference in positive prediction rates between groups:

$$\text{DD}(f) = \mathbb{P}_{X|A=1}(\hat{Y}_f = 1) - \mathbb{P}_{X|A=0}(\hat{Y}_f = 1);$$

- ▶ **Disparity of Opportunity (DO)**. Difference in true positive rates:

$$\text{DO}(f) = \mathbb{P}_{X|A=1, Y=1}(\hat{Y}_f = 1) - \mathbb{P}_{X|A=0, Y=1}(\hat{Y}_f = 1);$$

Example disparity measures

Definition (Disparity Measures)

We consider the following disparity measures: (Calders et al., 2009; Hardt et al., 2016;

Corbett-Davies et al., 2017; Cho et al., 2020; ...)

- ▶ **Demographic Disparity (DD)**. Difference in positive prediction rates between groups:

$$\text{DD}(f) = \mathbb{P}_{X|A=1}(\hat{Y}_f = 1) - \mathbb{P}_{X|A=0}(\hat{Y}_f = 1);$$

- ▶ **Disparity of Opportunity (DO)**. Difference in true positive rates:

$$\text{DO}(f) = \mathbb{P}_{X|A=1, Y=1}(\hat{Y}_f = 1) - \mathbb{P}_{X|A=0, Y=1}(\hat{Y}_f = 1);$$

- ▶ **Predictive Disparity (PD)**. Difference in false positive rates:

$$\text{PD}(f) = \mathbb{P}_{X|A=1, Y=0}(\hat{Y}_f = 1) - \mathbb{P}_{X|A=0, Y=0}(\hat{Y}_f = 1).$$

Fair Bayes-optimal classifiers

Classifiers with the highest accuracy given a disparity level.

Fair Bayes-optimal classifiers

Classifiers with the highest accuracy given a disparity level.

Definition (Fair Bayes-optimal Classifier)

Consider any $K \geq 1$ fairness measures $\text{Dis}_k : \mathcal{F} \rightarrow [0, 1]$, $k = 1, \dots, K$.

Then, a δ -fair Bayes-optimal classifier $f_{\text{Dis}, \delta}^*$ minimizes the

misclassification error $R(f) := \mathbb{P}\left(Y \neq \hat{Y}_f\right)$ over all classifiers that satisfy
 δ -disparity:

$$f_{\text{Dis}, \delta}^* \in \operatorname{argmin}_{f \in \mathcal{F}} \left\{ R(f) : \max_{k=1}^K |\text{Dis}_k(f)| \leq \delta \right\}.$$

Fair Bayes-optimal classifiers

Classifiers with the highest accuracy given a disparity level.

Definition (Fair Bayes-optimal Classifier)

Consider any $K \geq 1$ fairness measures $\text{Dis}_k : \mathcal{F} \rightarrow [0, 1]$, $k = 1, \dots, K$.

Then, a δ -fair Bayes-optimal classifier $f_{\text{Dis}, \delta}^*$ minimizes the

misclassification error $R(f) := \mathbb{P}\left(Y \neq \hat{Y}_f\right)$ over all classifiers that satisfy
 δ -disparity:

$$f_{\text{Dis}, \delta}^* \in \operatorname{argmin}_{f \in \mathcal{F}} \left\{ R(f) : \max_{k=1}^K |\text{Dis}_k(f)| \leq \delta \right\}.$$

Bayes-Optimal Fair Classification with Linear Disparity Constraints via Pre-, In-, and Post-processing

Xianli Zeng*, Guang Cheng† and Edgar Dobriban‡

February 6, 2024

Linear and bilinear disparity measures

- We find the form of Bayes-optimal classifiers for a broad class of disparities: linear & bilinear measures.

Linear and bilinear disparity measures

- ▶ We find the form of Bayes-optimal classifiers for a broad class of disparities: linear & bilinear measures.
- ▶ **Definition:** A disparity measure Dis is *linear* if there is a function w_{Dis} such that

$$\text{Dis}(f) = \int_{\mathcal{A}} \int_{\mathcal{X}} f(x, a) w_{\text{Dis}}(x, a) d\mathbb{P}_{X,A}(x, a).$$

Linear and bilinear disparity measures

- ▶ We find the form of Bayes-optimal classifiers for a broad class of disparities: linear & bilinear measures.
- ▶ **Definition:** A disparity measure Dis is *linear* if there is a function w_{Dis} such that

$$\text{Dis}(f) = \int_{\mathcal{A}} \int_{\mathcal{X}} f(x, a) w_{\text{Dis}}(x, a) d\mathbb{P}_{X,A}(x, a).$$

- ▶ Class-conditional probability function η_a :
 $\eta_a(x) = \mathbb{P}(Y = 1 \mid A = a, X = x)$ for all x, a .

Linear and bilinear disparity measures

- ▶ We find the form of Bayes-optimal classifiers for a broad class of disparities: linear & bilinear measures.
- ▶ **Definition:** A disparity measure Dis is *linear* if there is a function w_{Dis} such that

$$\text{Dis}(f) = \int_A \int_X f(x, a) w_{\text{Dis}}(x, a) d\mathbb{P}_{X,A}(x, a).$$

- ▶ Class-conditional probability function η_a :
 $\eta_a(x) = \mathbb{P}(Y = 1 \mid A = a, X = x)$ for all x, a .
- ▶ **Definition:** A linear disparity measure Dis is *bilinear* if $w_{\text{Dis}}(x, a)$ can be expressed for all x, a as $w_{\text{Dis}}(x, a) = s_{\text{Dis}, a} \eta_a(x) + b_{\text{Dis}, a}$.

Linear and bilinear disparity measures

- ▶ We find the form of Bayes-optimal classifiers for a broad class of disparities: linear & bilinear measures.
- ▶ **Definition:** A disparity measure Dis is *linear* if there is a function w_{Dis} such that

$$\text{Dis}(f) = \int_A \int_X f(x, a) w_{\text{Dis}}(x, a) d\mathbb{P}_{X,A}(x, a).$$

- ▶ Class-conditional probability function η_a :
 $\eta_a(x) = \mathbb{P}(Y = 1 | A = a, X = x)$ for all x, a .
- ▶ **Definition:** A linear disparity measure Dis is *bilinear* if $w_{\text{Dis}}(x, a)$ can be expressed for all x, a as $w_{\text{Dis}}(x, a) = s_{\text{Dis}, a} \eta_a(x) + b_{\text{Dis}, a}$.
- ▶ **Proposition:** The disparity measures DD, DO, and PD are bilinear:
for all x, a

$$w_{\text{DD}}(x, a) = \frac{(2a-1)}{p_a}; \quad w_{\text{DO}}(x, a) = \frac{(2a-1)\eta_a(x)}{p_{a,1}}; \quad w_{\text{PD}}(x, a) = \frac{(2a-1)(1-\eta_a(x))}{p_{a,0}}.$$

Intuition for Optimal Fair Classifiers

- ▶ Consider a linear disparity measure Dis :

$$\text{Dis}(f) = \int_{\mathcal{A}, \mathcal{X}} f(x, a) w_{\text{Dis}}(x, a) d\mathbb{P}_{X, A}(x, a)$$

Intuition for Optimal Fair Classifiers

- ▶ Consider a linear disparity measure Dis :

$$\text{Dis}(f) = \int_{\mathcal{A}, \mathcal{X}} f(x, a) w_{\text{Dis}}(x, a) d\mathbb{P}_{X, A}(x, a)$$

- ▶ Want to minimize

$$R(f) = \int_{\mathcal{A}, \mathcal{X}} f(x, a)(1 - 2\eta_a(x)) d\mathbb{P}_{X, A}(x, a) + C_{\mathbb{P}}$$

subject to $|\text{Dis}(f)| \leq \delta$

Intuition for Optimal Fair Classifiers

- ▶ Consider a linear disparity measure Dis :

$$\text{Dis}(f) = \int_{\mathcal{A}, \mathcal{X}} f(x, a) w_{\text{Dis}}(x, a) d\mathbb{P}_{X, A}(x, a)$$

- ▶ Want to minimize

$$R(f) = \int_{\mathcal{A}, \mathcal{X}} f(x, a)(1 - 2\eta_a(x)) d\mathbb{P}_{X, A}(x, a) + C_{\mathbb{P}}$$

subject to $|\text{Dis}(f)| \leq \delta$

- ▶ Intuitively, an optimal classifier should be of the form (for some c):

$$f(x, a) = I\left(\eta_a(x) > \frac{1}{2} + c \cdot w_{\text{Dis}}(x, a)\right).$$

Optimal Fair Classifiers for Linear and Bilinear Disparity Measures

Consider a linear disparity measure Dis ; continuous random variables $\eta_a(X)$, $w_{\text{Dis}}(X, a)$, $a \in \{0, 1\}$.

Optimal Fair Classifiers for Linear and Bilinear Disparity Measures

Consider a linear disparity measure Dis ; continuous random variables $\eta_a(X)$, $w_{\text{Dis}}(X, a)$, $a \in \{0, 1\}$.

Theorem (Optimal Classifiers for Linear Disparity Measures;
Zeng et al., 2024)

Let $t_{\text{Dis}}(\delta) = \operatorname{argmin}_t \{|t| : |D_{\text{Dis}}(t)| \leq \delta\}$, where

$$D_{\text{Dis}}(t) = \int_{\mathcal{X} \times \mathcal{A}} w_{\text{Dis}}(x, a) \cdot I\left[\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\text{Dis}}(x, a)\right] d\mathbb{P}_{X,A}(x, a).$$

Optimal Fair Classifiers for Linear and Bilinear Disparity Measures

Consider a linear disparity measure Dis ; continuous random variables $\eta_a(X)$, $w_{\text{Dis}}(X, a)$, $a \in \{0, 1\}$.

Theorem (Optimal Classifiers for Linear Disparity Measures;
Zeng et al., 2024)

Let $t_{\text{Dis}}(\delta) = \operatorname{argmin}_t \{|t| : |D_{\text{Dis}}(t)| \leq \delta\}$, where

$$D_{\text{Dis}}(t) = \int_{\mathcal{X} \times \mathcal{A}} w_{\text{Dis}}(x, a) \cdot I[\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\text{Dis}}(x, a)] d\mathbb{P}_{X,A}(x, a).$$

A fair Bayes-optimal classifier is, for all x, a ,

$$f_{\text{Dis}, \delta}^*(x, a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\text{Dis}}(\delta)}{2} w_{\text{Dis}}(x, a)\right).$$

Optimal Fair Classifiers for Linear and Bilinear Disparity Measures

Consider a linear disparity measure Dis ; continuous random variables $\eta_a(X)$, $w_{\text{Dis}}(X, a)$, $a \in \{0, 1\}$.

Theorem (Optimal Classifiers for Linear Disparity Measures; Zeng et al., 2024)

Let $t_{\text{Dis}}(\delta) = \operatorname{argmin}_t \{|t| : |D_{\text{Dis}}(t)| \leq \delta\}$, where

$$D_{\text{Dis}}(t) = \int_{\mathcal{X} \times \mathcal{A}} w_{\text{Dis}}(x, a) \cdot I[\eta_a(x) > \frac{1}{2} + \frac{t}{2} w_{\text{Dis}}(x, a)] d\mathbb{P}_{X,A}(x, a).$$

A fair Bayes-optimal classifier is, for all x, a ,

$$f_{\text{Dis}, \delta}^*(x, a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\text{Dis}}(\delta)}{2} w_{\text{Dis}}(x, a)\right).$$

If Dis is bilinear, this is a group-wise thresholding rule, such that for all x, a ,

$$f_{\text{Dis}, \delta}^*(x, a) = I\left[\eta_a(x) > \frac{1+b_{\text{Dis}, a} \cdot t_{\text{Dis}}(\delta)}{2-s_{\text{Dis}, a} \cdot t_{\text{Dis}}(\delta)}\right].$$

Discussion, methods, and empirics

- ▶ Proof via the *generalized Neyman-Pearson lemma*: maximizing a linear functional subject to linear constraints.

Discussion, methods, and empirics

- ▶ Proof via the *generalized Neyman-Pearson lemma*: maximizing a linear functional subject to linear constraints.
- ▶ Apply it to develop methods for fair classification via
 - ▶ Pre-Processing: Fair Up-& Down-Sampling (FUDS);
 - ▶ In-Processing: Fair Cost-Sensitive Classification (FCSC);
 - ▶ Post-Processing: Fair Plug-in Thresholding Rule (FPIR).

Discussion, methods, and empirics

- ▶ Proof via the *generalized Neyman-Pearson lemma*: maximizing a linear functional subject to linear constraints.
- ▶ Apply it to develop methods for fair classification via
 - ▶ Pre-Processing: Fair Up-& Down-Sampling (FUDS);
 - ▶ In-Processing: Fair Cost-Sensitive Classification (FCSC);
 - ▶ Post-Processing: Fair Plug-in Thresholding Rule (FPIR).

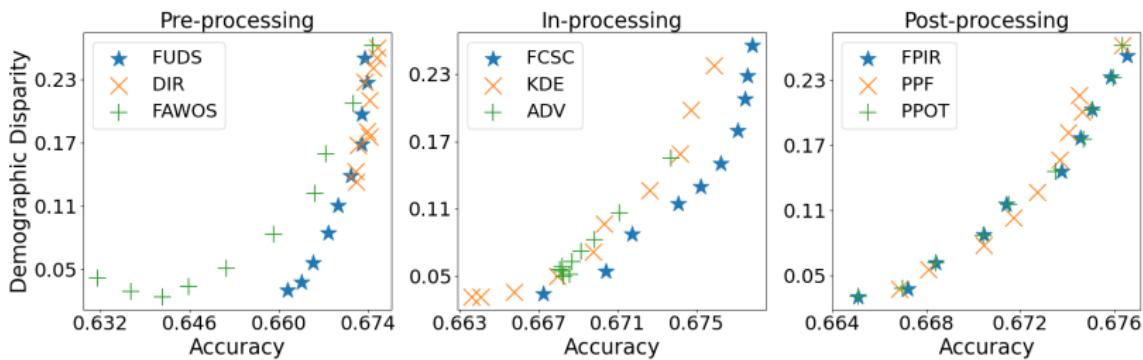


Figure: Fairness-accuracy tradeoff on the COMPAS dataset.

Bayes-Optimal Fair Up- and Down-Sampling

Suppose that $\eta_a(X)$ and $w_{\text{Dis}}(X, a)$, $a \in \{0, 1\}$, have densities.

Theorem. For $t \in \mathbb{R}$, let $\tilde{\mathbb{P}}^t$ be the distribution satisfying the **label shift** condition that for all $(x, a, y) \in \mathcal{X} \times \{0, 1\}^2$,

$$\tilde{\mathbb{P}}_{X|\tilde{A}=a, \tilde{Y}=y}^t(x) = \mathbb{P}_{X|A=a, Y=y}(x),$$

and, with $H_{\text{Dis}, a}(t) = \frac{1+t \cdot b_{\text{Dis}, a}}{2-t \cdot s_{\text{Dis}, a}}$,

$$\tilde{p}_{a,y}^t = \tilde{\mathbb{P}}^t(\tilde{A} = a, \tilde{Y} = y) = c_a[(1 - H_{\text{Dis}, a}(t))y + H_{\text{Dis}, a}(t)(1 - y)]p_{a,y}.$$

Here $c_0, c_1 > 0$ are such that $\tilde{p}_{11}^t + \tilde{p}_{10}^t + \tilde{p}_{01}^t + \tilde{p}_{00}^t = 1$.

Then, given $\delta \geq 0$, any **unconstrained Bayes-optimal classifier** for $\tilde{\mathbb{P}}^{t_{\text{Dis}}(\delta)}$ is a **δ -fair Bayes-optimal classifier** for \mathbb{P} .

Summary for algorithmic fairness

- ▶ Derived fair Bayes-optimal classifiers for linear disparity measures (via Neyman-Pearson lemma!).
- ▶ Developed fair classification methods via pre-, in-, and post-processing.

Summary for algorithmic fairness

- ▶ Derived fair Bayes-optimal classifiers for linear disparity measures (via Neyman-Pearson lemma!).
- ▶ Developed fair classification methods via pre-, in-, and post-processing.
- ▶ Related work: minimax optimal fair classification.

Minimax Optimal Fair Classification with Bounded Demographic Disparity

Xianli Zeng*, Guang Cheng[†] and Edgar Dobriban[‡]

March 28, 2024

Summary and outlook

- ▶ We explored and illustrated how to improve the reliability of AI. A lot remains to be done and discovered!
- ▶ Thanks!



SIM NS
FOUNDATION

ALFRED P. SLOAN
FOUNDATION

