On the Role of Nonparametric Statistics in the Age of AI

Edgar Dobriban

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

August 8, 2024

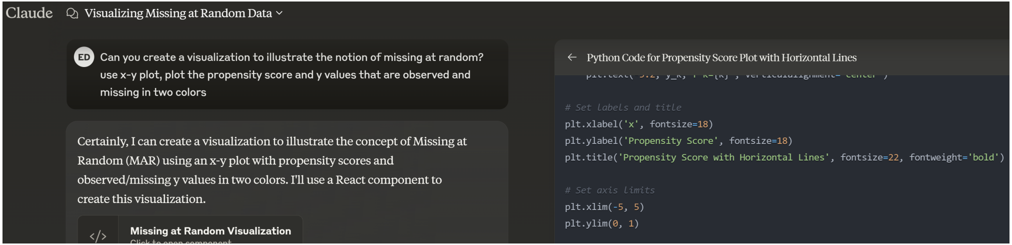
Table of Contents

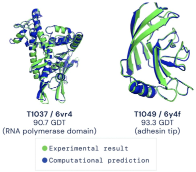
Context

- 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 ...





Is this a real trend?

- Statisticians, skeptical as we often are, may wonder if this is a real trend. Here are a few facts:

Is this a real trend?

- Statisticians, skeptical as we often are, may wonder if this is a real trend. Here are a few facts:
  - ChatGPT has 100 million weekly active users

Is this a real trend?

- Statisticians, skeptical as we often are, may wonder if this is a real trend. Here are a few facts:
  - ChatGPT has 100 million weekly active users
  - Nvidia is one of the world's most valuable publicly traded companies

Is this a real trend?

- Statisticians, skeptical as we often are, may wonder if this is a real trend. Here are a few facts:
  - ChatGPT has 100 million weekly active users
  - Nvidia is one of the world's most valuable publicly traded companies
  - AI starting to be used in products: Code assistants (Copilot), Customer service, Web search

Is this a real trend?

- Statisticians, skeptical as we often are, may wonder if this is a real trend. Here are a few facts:
  - ChatGPT has 100 million weekly active users
  - Nvidia is one of the world's most valuable publicly traded companies
  - AI starting to be used in products: Code assistants (Copilot), Customer service, Web search
  - Several papers in Nature/Science reporting discoveries using AI

Article

Scaling deep learning for materials discovery

https://doi.org/10.1038/s41586-023-06735-9
Received: 8 May 2023
Accepted: 10 October 2023
Published online: 29 November 2023

Amil Merchant[1,3], Simon Batzner[1,3], Samuel S. Schoenholz[1,3], Muratahan Aykol[1], Gowoon Cheon[2] & Ekin Dogus Cubuk[1,3]

Novel functional materials enable fundamental breakthroughs across technological applications from clean energy to information processing[1–11]. From microchips to

But what is AI?

- Relies on a predictive model $\hat{p}$, such that for all $x \in \mathcal{X}$ (images, text), $\hat{p}(\cdot|x)$ is a probability distribution over a set $\mathcal{Y}$ (images, text) from which we can sample

Predictive Model $\hat{p}$	Input Space $\mathcal{X}$	Sample Space $\mathcal{Y}$
Large Language Models	Text	Text
Diffusion Models	Text	Images
Multimodal Language Models	Images, text	Images, text, sound, video
Protein Structure Prediction	Amino acid sequence	3D structure

But what is AI?

- Relies on a predictive model $\hat{p}$, such that for all $x \in \mathcal{X}$ (images, text), $\hat{p}(\cdot|x)$ is a probability distribution over a set $\mathcal{Y}$ (images, text) from which we can sample

Predictive Model $\hat{p}$	Input Space $\mathcal{X}$	Sample Space $\mathcal{Y}$
Large Language Models	Text	Text
Diffusion Models	Text	Images
Multimodal Language Models	Images, text	Images, text, sound, video
Protein Structure Prediction	Amino acid sequence	3D structure

- Enormous engineering effort around it (input/output/UX, specialized hardware, multiple calls, ...)

What is the connection to statistics?

- The predictive model $\hat{p}$ is obtained by minimizing a simple loss function over a non-parametric function class using a massive data set

What is the connection to statistics?

- The predictive model $\hat{p}$ is obtained by minimizing a simple loss function over a non-parametric function class using a massive data set
- E.g., for LLMs:
  - Data type: strings $x = (x_1, x_2, \dots, x_{\mathrm{len}(x)})$;
  - Loss: $-\sum_{x \in \mathcal{D}} \sum_{i \leq \mathrm{len}(x)} \log p_\theta(x_i | x_1, \dots, x_{i-1}) = -\sum_{x \in \mathcal{D}} \log p_\theta(x)$
  - Function class $\theta \mapsto p_\theta$: Huge neural nets
  - Dataset: Mixture of internet data sets (books, Wikipedia, arxiv, Common Crawl, ...?)

What is the role of (nonparametric) statistics in the age of AI?

- What should we, statisticians, do to maximize our impact during this time?

What is the role of (nonparametric) statistics in the age of AI?

- What should we, statisticians, do to maximize our impact during this time?
- What new methods should we be developing?

What is the role of (nonparametric) statistics in the age of AI?

- What should we, statisticians, do to maximize our impact during this time?
- What new methods should we be developing?
- Should we change anything in our training programs?

What is the role of (nonparametric) statistics in the age of AI?

- What should we, statisticians, do to maximize our impact during this time?
- What new methods should we be developing?
- Should we change anything in our training programs?
- Will discuss a few emerging areas of research where statistics could have an impact:
  - Uncertainty quantification in AI/ML

What is the role of (nonparametric) statistics in the age of AI?

- What should we, statisticians, do to maximize our impact during this time?
- What new methods should we be developing?
- Should we change anything in our training programs?
- Will discuss a few emerging areas of research where statistics could have an impact:
  - Uncertainty quantification in AI/ML
  - Algorithmic fairness

What is the role of (nonparametric) statistics in the age of AI?

- What should we, statisticians, do to maximize our impact during this time?
- What new methods should we be developing?
- Should we change anything in our training programs?
- Will discuss a few emerging areas of research where statistics could have an impact:
  - Uncertainty quantification in AI/ML
  - Algorithmic fairness
  - AI evaluation
  - ...

Table of Contents

Motivation

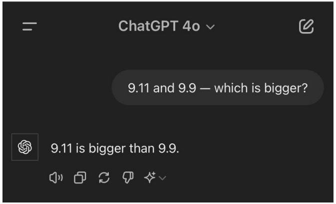
Uncertainty quantification in AI/ML

Algorithmic Fairness

AI Evaluation

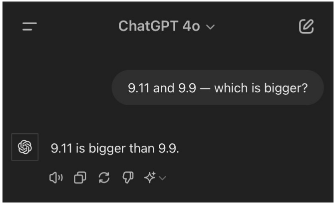
Uncertainty quantification in AI/ML

- AI systems can be wrong



Uncertainty quantification in AI/ML

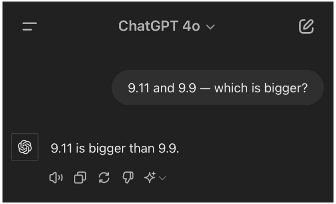
- AI systems can be wrong



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

Uncertainty quantification in AI/ML

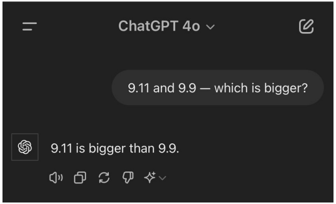
- AI systems can be wrong



- 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: $U(x, y) = \hat{p}(y|x)^{1/\mathrm{len}(y)}$; [related to NLL $-\log \hat{p}(y|x)$]

Example uncertainty measures

- Many uncertainty measures for LLMs have been discussed/proposed
  - Perplexity: $U(x, y) = \hat{p}(y|x)^{1/\mathrm{len}(y)}$; [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 $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
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: $U(x, y) = \hat{p}(y|x)^{1/\mathrm{len}(y)}$; [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 $\mathbf{s}$	Likelihood $p(\mathbf{s} \mid x)$	Semantic likelihood $\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x)$
Paris	0.5	0.9
It's Paris	0.4	
London	0.1	0.1
Entropy	0.94	0.33

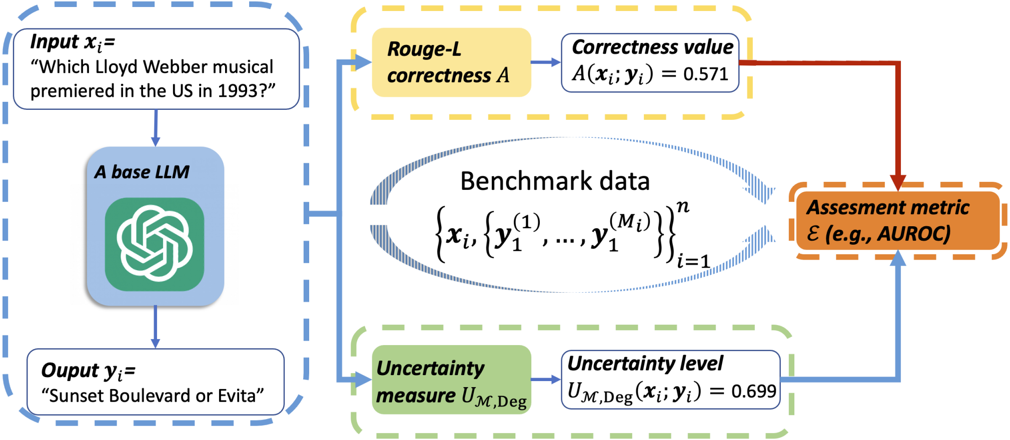
  - Affinity graph: generate multiple ys, put pairwise similarities in a matrix, find eigenvalues/vectors (denoted EigV) (Lin et al., 2023)

Good uncertainty measures

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

Good uncertainty measures

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



Regression function and indication diagram

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

Regression function and indication diagram

- Define the regression function $\mathrm{reg}(\cdot) : \mathbb{R} \to \mathbb{R}, u \mapsto \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[A(\boldsymbol{x};\boldsymbol{y}) \mid U(\boldsymbol{x};\boldsymbol{y}) = u]$, 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.

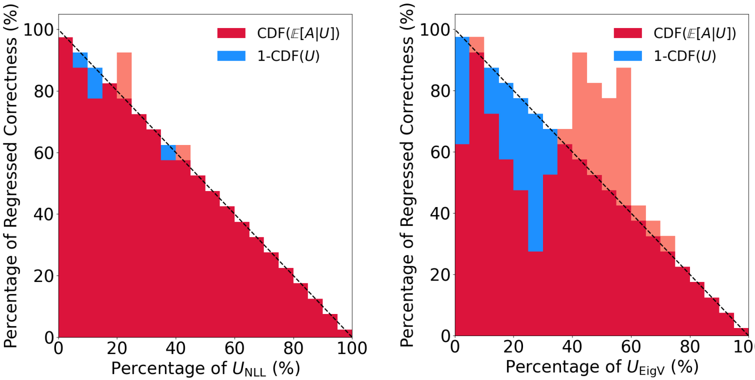


Figure: *Indication diagrams* for U_{NLL} (negative log-likelihood) and U_{EigV}, for the GPT-3.5-turbo model on the TriviaQA benchmark.

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.

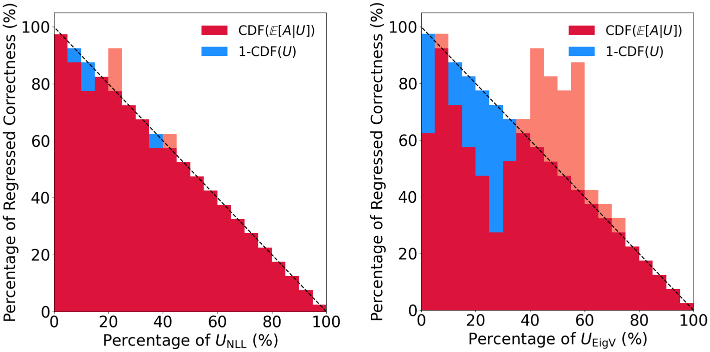


Figure: *Indication diagrams* for U_{NLL} (negative log-likelihood) and U_{EigV}, for the GPT-3.5-turbo model on the TriviaQA benchmark.

Rank-calibration error (RCE)

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

Rank-calibration error (RCE)

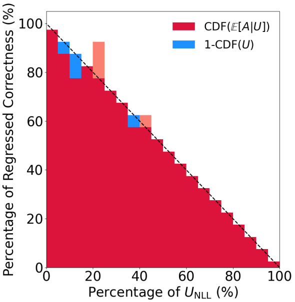
Rank-calibration implies $\mathbb{P}(U \leq u') = \mathbb{P}(\mathrm{reg}(U) \geq \mathrm{reg}(u'))$ for all $u' \in \mathrm{dom}(U)$.

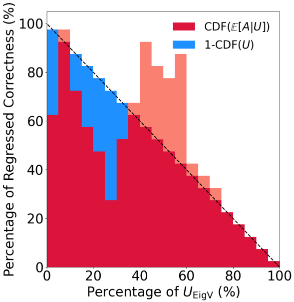
Definition (Rank-Calibration Error)

The RCE of an uncertainty measure U is defined as

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

where U' is an independent copy of U.





Empirical results

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

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-Zamowitz, 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-Zamowitz, 1969)
- Improves RCE:

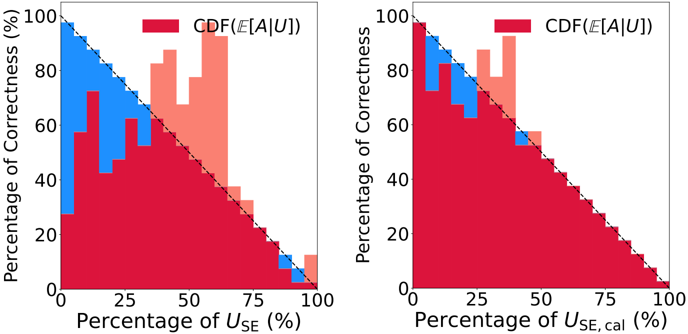


Figure: Indication diagrams of U_{SE} and $U_{\mathrm{SE,cal}}$ (re-calibrated) for GPT-3.5-turbo (temperature 1.0) on TriviaQA with the Meteor correctness.

Summary for uncertainty quantification in AI/ML

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

Summary for uncertainty quantification in AI/ML

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

Summary for uncertainty quantification in AI/ML

- Proposed rank-calibration for LLMs: higher uncertainty should imply lower quality (not satisfied by default!)
- Proposed non-parametric re-calibration to reduce RCE
- Showcased the utility of ideas from non-parametric statistics

Uncertainty in Language Models: Assessment through Rank-Calibration

Xinmeng Huang[*†] **Shuo Li**[*†] **Mengxin Yu**[†] **Matteo Sesia**[‡] **Hamed Hassani**[†]

Insup Lee[†] **Osbert Bastani**[§†] **Edgar Dobriban**[§†]

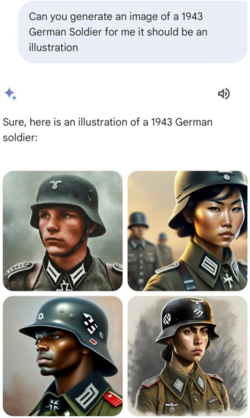
Table of Contents

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., asuch systems often show unfair behavior.



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 $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$: $\widehat{Y}_f \mid (X = x, A = a) \sim \text{Bern}(f(x, a))$.

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**: $\widehat{Y}_f \perp\!\!\!\perp A$;

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**: $\widehat{Y}_f \perp\!\!\!\perp A$;
- **Equality of Opportunity**: $\widehat{Y}_f \perp\!\!\!\perp A \mid Y = 1$;

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

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:

$$\mathrm{DD}(f) = \mathbb{P}_{X|A=1}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_f = 1\right);$$

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:

$$\mathrm{DD}(f) = \mathbb{P}_{X|A=1}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_f = 1\right);$$

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

$$\mathrm{DO}(f) = \mathbb{P}_{X|A=1,Y=1}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=0,Y=1}\left(\widehat{Y}_f = 1\right);$$

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:
$$\mathrm{DD}(f) = \mathbb{P}_{X|A=1}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=0}\left(\widehat{Y}_f = 1\right);$$
- **Disparity of Opportunity (DO)**. Difference in true positive rates:
$$\mathrm{DO}(f) = \mathbb{P}_{X|A=1,Y=1}(\widehat{Y}_f = 1) - \mathbb{P}_{X|A=0,Y=1}\left(\widehat{Y}_f = 1\right);$$
- **Predictive Disparity (PD)**. Difference in false positive rates:
$$\mathrm{PD}(f) = \mathbb{P}_{X|A=1,Y=0}\left(\widehat{Y}_f = 1\right) - \mathbb{P}_{X|A=0,Y=0}\left(\widehat{Y}_f = 1\right).$$

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 $\mathrm{Dis}_k : \mathcal{F} \to [0,1]$, $k = 1, \ldots, K$. Then, a *δ-fair Bayes-optimal classifier* $f^\star_{\mathrm{Dis},\delta}$ minimizes the misclassification error $R(f) := \mathbb{P}\left(Y \neq \widehat{Y}_f\right)$ over all classifiers that satisfy δ-disparity:

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

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 $\text{Dis}_k : \mathcal{F} \to [0,1]$, $k = 1, ..., K$. Then, a *δ-fair Bayes-optimal classifier* $f^\star_{\text{Dis},\delta}$ minimizes the misclassification error $R(f) := \mathbb{P}\left(Y \neq \widehat{Y}_f\right)$ over all classifiers that satisfy δ-disparity:

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

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

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

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

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

- 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 $w_{\mathrm{Dis}}(x, a)$ can be expressed for all x, a as $w_{\mathrm{Dis}}(x, a) = s_{\mathrm{Dis},a}\eta_a(x) + b_{\mathrm{Dis},a}$.

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

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

- 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 $w_{\mathrm{Dis}}(x, a)$ can be expressed for all x, a as $w_{\mathrm{Dis}}(x, a) = s_{\mathrm{Dis},a} \eta_a(x) + b_{\mathrm{Dis},a}$.
- **Proposition**: The disparity measures DD, DO, and PD are bilinear: for all x, a

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

Intuition for optimal fair classifiers

- Consider a linear disparity measure Dis:

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

Intuition for optimal fair classifiers

- Consider a linear disparity measure Dis:

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

- Want to minimize

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

s.t. $|\mathrm{Dis}(f)| \leq \delta$

Intuition for optimal fair classifiers

- Consider a linear disparity measure Dis:

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

- Want to minimize

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

s.t. $|\mathrm{Dis}(f)| \leq \delta$

- Intuitive that an optimal classifier should be of the form (for some c)

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

Optimal fair classifiers for linear and bilinear disparity measures

Consider linear disp. measure Dis; Continuous r.v.s $\eta_a(X)$, $w_{\mathrm{Dis}}(X, a)$, $a \in \{0, 1\}$.

Optimal fair classifiers for linear and bilinear disparity measures

Consider linear disp. measure Dis; Continuous r.v.s $\eta_a(X)$, $w_{\mathrm{Dis}}(X, a)$, $a \in \{0, 1\}$.

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

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

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

Optimal fair classifiers for linear and bilinear disparity measures

Consider linear disp. measure Dis; Continuous r.v.s $\eta_a(X)$, $w_{\mathrm{Dis}}(X, a)$, $a \in \{0, 1\}$.

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

Let $t_{\mathsf{Dis}}(\delta) = \mathrm{argmin}_t\{|t| : |D_{\mathsf{Dis}}(t)| \le \delta\}$, *where*

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

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

$$f^{\star}_{\mathsf{Dis},\delta}(x, a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\mathsf{Dis}}(\delta)}{2} w_{\mathsf{Dis}}(x, a)\right).$$

Optimal fair classifiers for linear and bilinear disparity measures

Consider linear disp. measure Dis; Continuous r.v.s $\eta_a(X)$, $w_{\mathrm{Dis}}(X, a)$, $a \in \{0, 1\}$.

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

Let $t_{\mathrm{Dis}}(\delta) = \mathrm{argmin}_t\{|t| : |D_{\mathrm{Dis}}(t)| \le \delta\}$, *where*

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

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

$$f^{\star}_{\mathrm{Dis},\delta}(x, a) = I\left(\eta_a(x) > \frac{1}{2} + \frac{t_{\mathrm{Dis}}(\delta)}{2} w_{\mathrm{Dis}}(x, a)\right).$$

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

$$f^{\star}_{\mathrm{Dis},\delta}(x, a) = I\left(\eta_a(x) > \frac{1 + b_{\mathrm{Dis},a} \cdot t_{\mathrm{Dis}}(\delta)}{2 - s_{\mathrm{Dis},a} \cdot t_{\mathrm{Dis}}(\delta)}\right).$$

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).

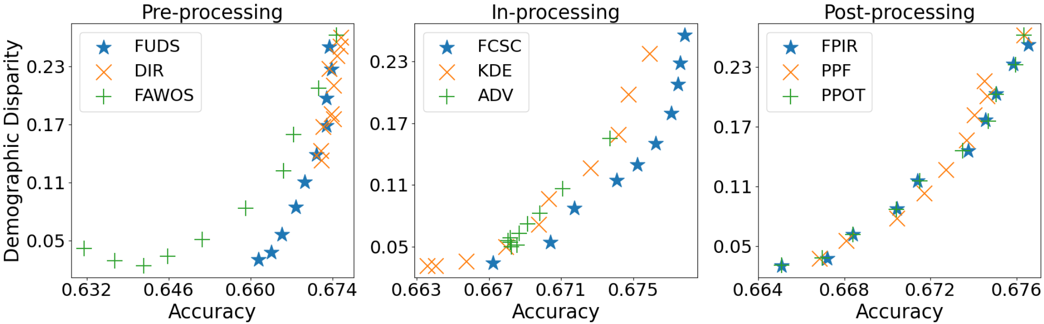


Figure: Fairness-accuracy tradeoff on the COMPAS dataset.

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

Table of Contents

AI evaluation

- Evaluating AI can be hard

AI evaluation

- Evaluating AI can be hard
- Say want to measure *reasoning ability*:
  - Collect test data on reasoning problems (e.g., high school math)
  - Evaluate error rate

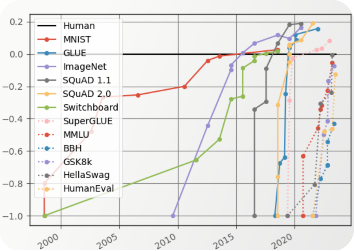


Figure: Plotting progress in AI (Kiela et al. 2023)

AI evaluation

- Evaluating AI can be hard
- Say want to measure *reasoning ability*:
  - Collect test data on reasoning problems (e.g., high school math)
  - Evaluate error rate

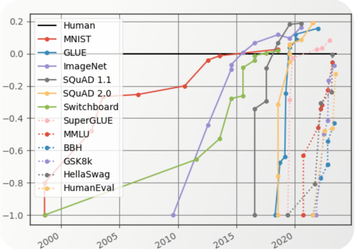


Figure: Plotting progress in AI (Kiela et al. 2023)

- Many challenges
  - Test data is necessarily arbitrary and limited
  - Correctness evaluation can be ambiguous (often use other LLMs!)
  - Evaluating the biggest models can be expensive (lots of API calls)

A statistical approach to efficient AI evaluation

tinyBenchmarks: evaluating LLMs with fewer examples

Felipe Maia Polo [1] **Lucas Weber** [2] **Leshem Choshen** [3 4] **Yuekai Sun** [1] **Gongjun Xu** [1] **Mikhail Yurochkin** [3 5]

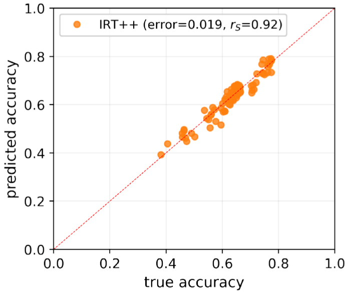


Figure 1. Estimating accuracy on MMLU (true accuracy) using 100 curated examples (predicted accuracy). IRT++, our best-performing evaluation strategy, predicts the accuracy of recent LLMs released between December 30th and January 18th within 1.9% of their true accuracy on all of MMLU (14K examples).

Item response theory (IRT) for AI evaluation

4. Better performance estimation with IRT

4.1. The IRT model

The two-parameter multidimensional IRT model assumes that the probability of the LLM j getting example i correctly is given by

$$p_{il} \triangleq \mathbb{P}(Y_{il} = 1 \mid \theta_l, \alpha_i, \beta_i) = \frac{1}{1+\exp(-\alpha_i^\top \theta_l + \beta_i)}, \quad (4.1)$$

where $\theta_l \in \mathbb{R}^d$ denotes the unobserved abilities of LLM l, while $\alpha_i \in \mathbb{R}^d$ dictates which dimensions of θ_l are required from model l to respond to example i correctly. In this formulation, $\beta_i \in \mathbb{R}$ can be viewed as a bias term that regulates the probability of correctness when $\theta_l = 0$.

Formally, we are interested in approximating

$$Z_{jl} \triangleq \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} Y_{il} \quad (4.2)$$

Now, assume that we have run model l on a subset of examples from scenario j, obtaining responses $\{Y_{i_0 l}, \cdots, Y_{i_k l}\}$ for the examples $\widehat{\mathcal{I}}_j = \{i_0, \cdots, i_k\}$. Let $\widehat{\theta}_l$ denote the estimate for θ_l after observing $\widehat{\mathcal{I}}_j$ and possibly a bigger set of examples coming from different scenarios. To obtain that estimate, we maximize the log-likelihood of the freshly observed data with respect to θ_l, fixing examples' parameters. This procedure is equivalent to fitting a logistic regression model.

Because Z_{jl} is a random variable, we approximate it by estimating the conditional expectation

$$\mathbb{E}[Z_{jl} \mid Y_{i_0 l}, \cdots, Y_{i_k l}] = \\ = \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} \mathbb{E}[Y_{il} \mid Y_{i_0 l}, \cdots, Y_{i_k l}]$$

The probability $p_{il} = \mathbb{P}(Y_{il} = 1 \mid \theta_l, \alpha_i, \beta_i)$ is given by the IRT model in Equation 4.1. The estimator for the conditional expectation is then given by

$$\hat{Z}_{jl}^{\text{p-IRT}} \triangleq \widehat{\mathbb{E}}[Z_{jl} \mid Y_{i_0 l}, \cdots, Y_{i_k l}] \quad (4.3) \\ = \frac{\hat{\lambda}}{|\widehat{\mathcal{I}}_j|} \sum_{i \in \widehat{\mathcal{I}}_j} Y_{il} + \frac{1-\hat{\lambda}}{|\mathcal{I}_j \setminus \widehat{\mathcal{I}}_j|} \sum_{i \in \mathcal{I}_j \setminus \widehat{\mathcal{I}}_j} \hat{p}_{il}$$

where $\hat{p}_{il} \triangleq \mathbb{P}(Y_{il} = 1 \mid \widehat{\theta}_l, \widehat{\alpha}_i, \widehat{\beta}_i)$.

Summary and outlook

- We explored and illustrated how (nonparametric) statistics could have an impact in the world of AI. The most exciting discoveries are yet to come!
- Thanks!