Achievable Fairness on Your Data With Utility Guarantees

Muhammad Faaiz Taufiq, Jean-François Ton, Yang Liu
NeurIPS ‘24

Motivational Example

A bank uses a predictive model ***h*** to decide whether to grant loans to applicants, based on their data.

Features for each applicant include race, gender, annual income, age, etc.

Motivational Example

A bank uses a predictive model ***h*** to decide whether to grant loans to applicants, based on their data.

Features for each applicant include race, gender, annual income, age, etc.

We would like the model ***h*** to not discriminate against applicants based on their gender.

Fairness losses (like demographic parity) measure how much the model depends on gender

Motivational Example

A bank uses a predictive model ***h*** to decide whether to grant loans to applicants, based on their data.

Features for each applicant include race, gender, annual income, age, etc.

We would like the model ***h*** to not discriminate against applicants based on their gender.

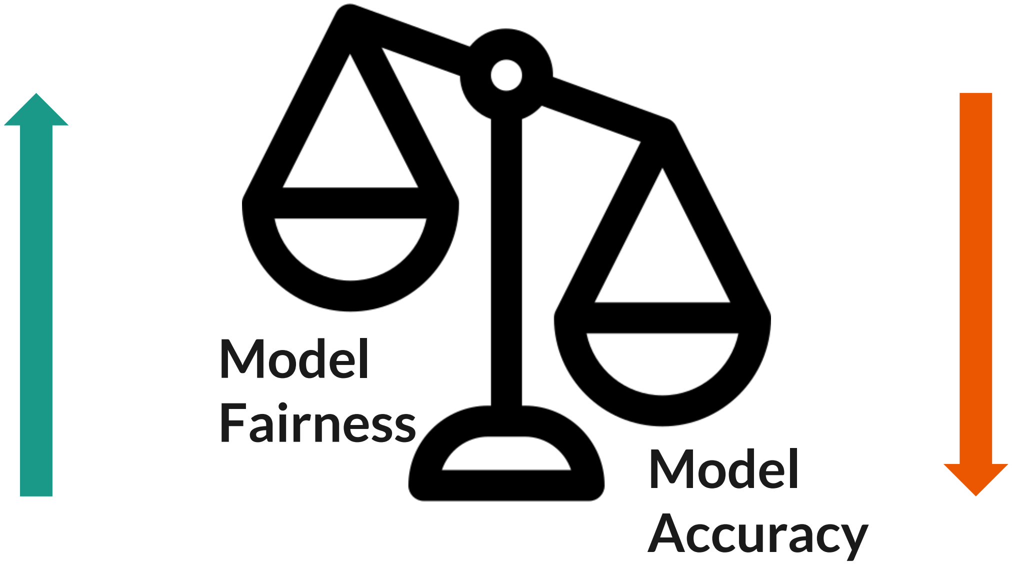
Fairness losses (like demographic parity) measure how much the model depends on gender

"Why is the fairness loss for this model not 0?"

Problem!

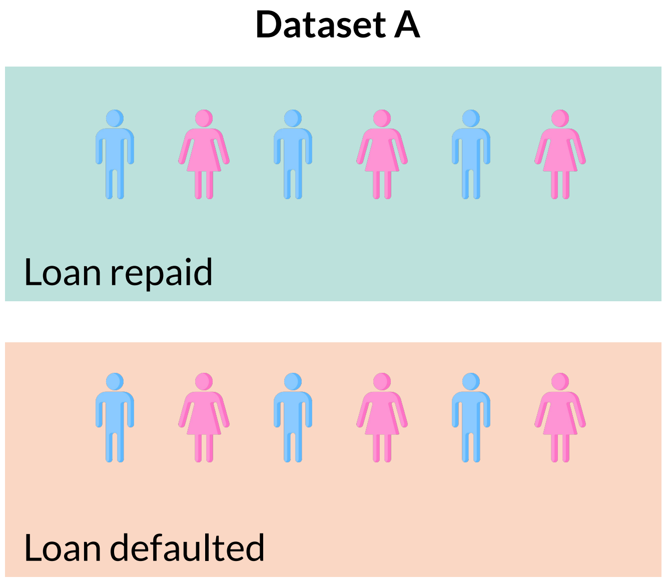
Making the model fairer can reduce model accuracy.

As the model becomes fairer

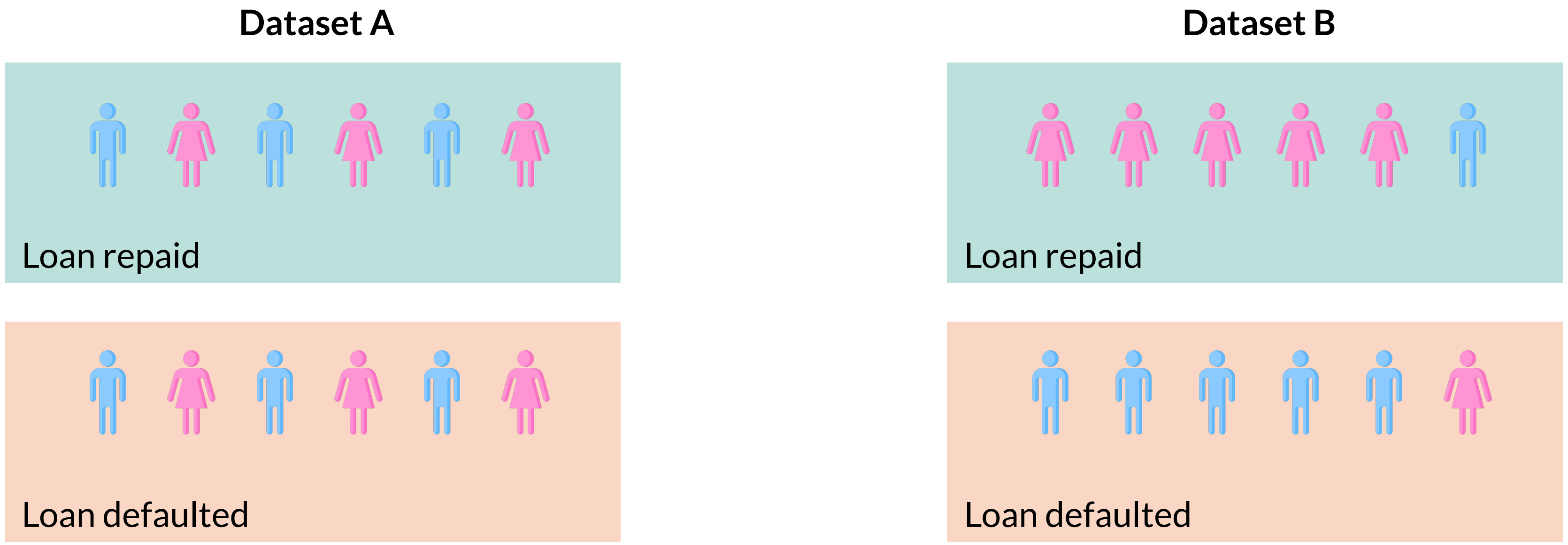


the model accuracy decreases

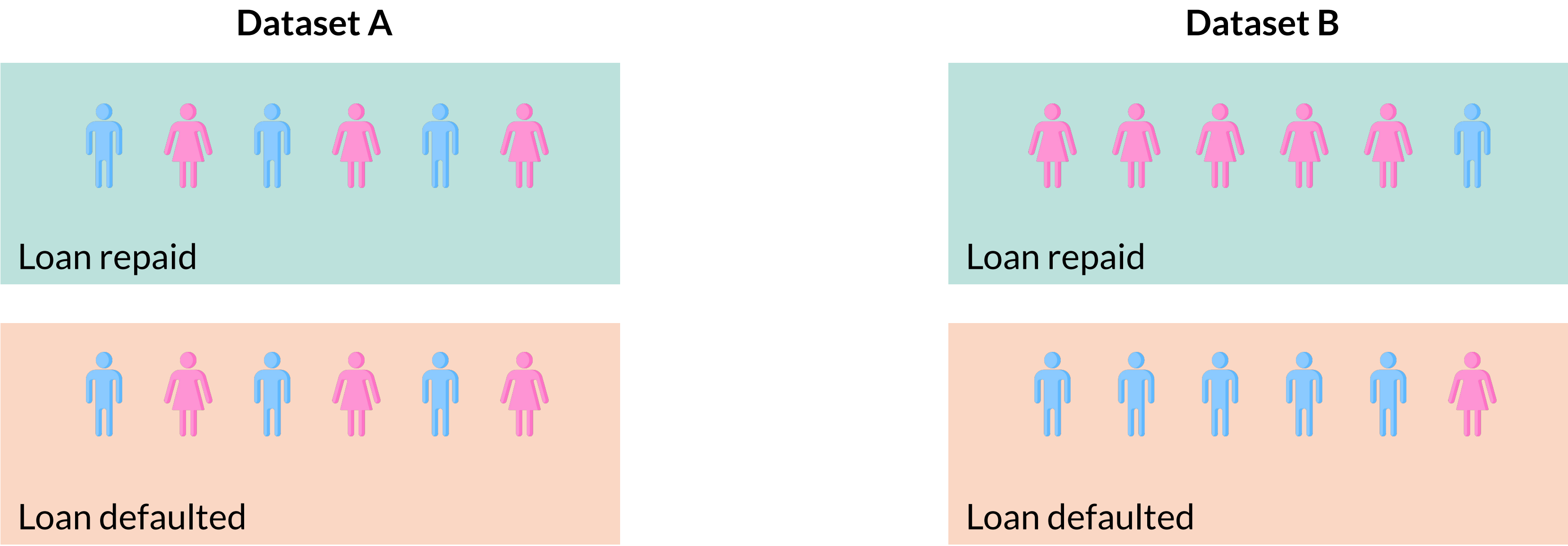
Accuracy-fairness trade-off is data dependent



Accuracy-fairness trade-off is data dependent

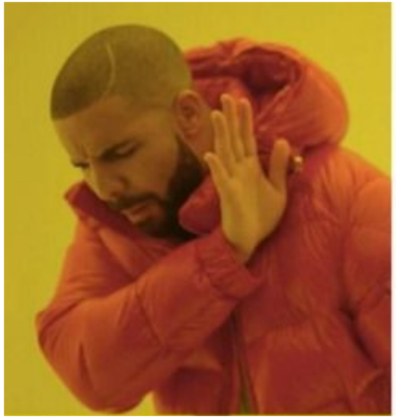


Accuracy-fairness trade-off is data dependent



Training classifiers which are gender agnostic is more challenging for Dataset B than for Dataset A

Problem statement



“Why is the fairness loss for this model not 0?”

Problem statement

“Why is the fairness loss for this model not 0?”

“For this dataset, what is the minimum attainable fairness loss corresponding to <u>each accuracy threshold</u>?”

Problem statement

Our problem can be formalised as finding $l(\delta)$ defined as:

$$l(\delta) := \min_{h \in \mathcal{H}} \mathcal{L}_{\mathrm{f}}(h) \quad \text{subject to} \quad \mathrm{acc}(h) \geq \delta$$

Problem statement

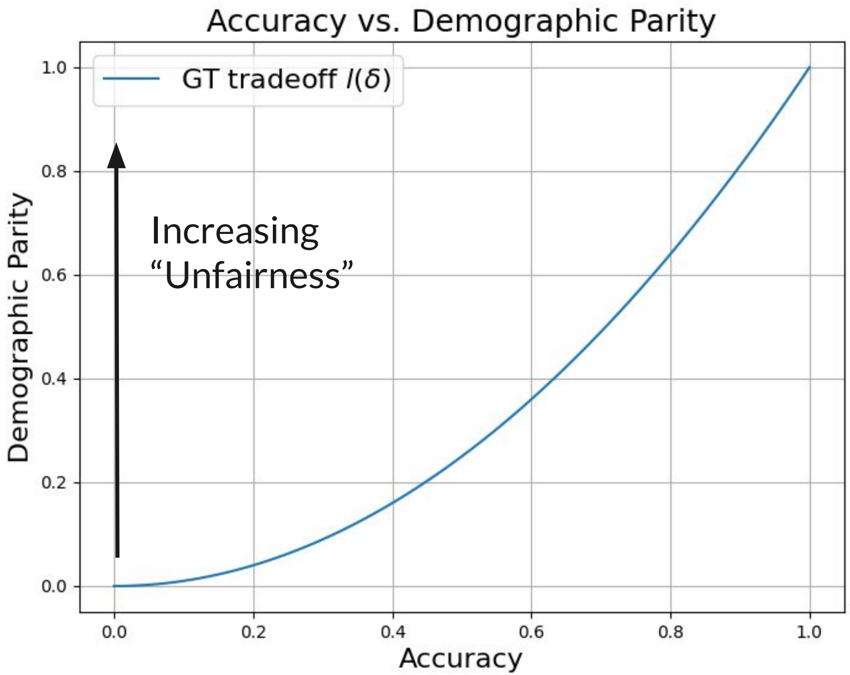
Our problem can be formalised as finding $l(\delta)$ defined as:

$$l(\delta) := \min_{h \in \mathcal{H}} \boxed{\mathcal{L}_{\mathrm{f}}(h)} \quad \text{subject to} \quad \boxed{\mathrm{acc}(h)} \geq \delta$$

Fairness loss
E.g. demographic parity

Model accuracy

Problem statement

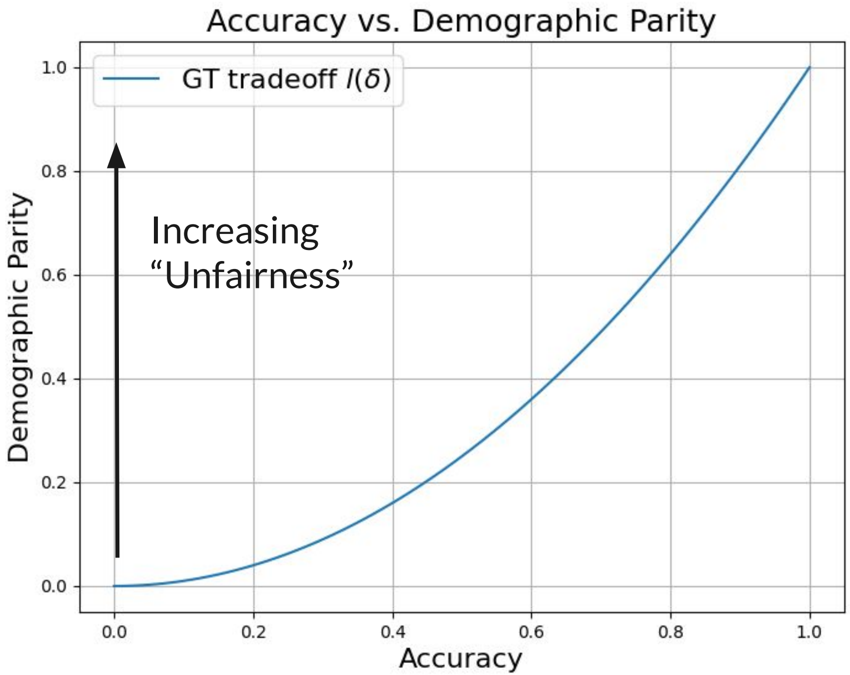


$$l(\delta) := \min_{h\in\mathcal{H}} \boxed{\mathcal{L}_{\mathrm{f}}(h)} \quad \text{subject to} \quad \boxed{\mathrm{acc}(h)} \geq \delta$$

Fairness loss
E.g. demographic parity

Model accuracy

Problem statement



$$l(\delta) := \min_{h\in\mathcal{H}} \boxed{\mathcal{L}_{\mathrm{f}}(h)} \quad \text{subject to} \quad \boxed{\mathrm{acc}(h)} \geq \delta$$

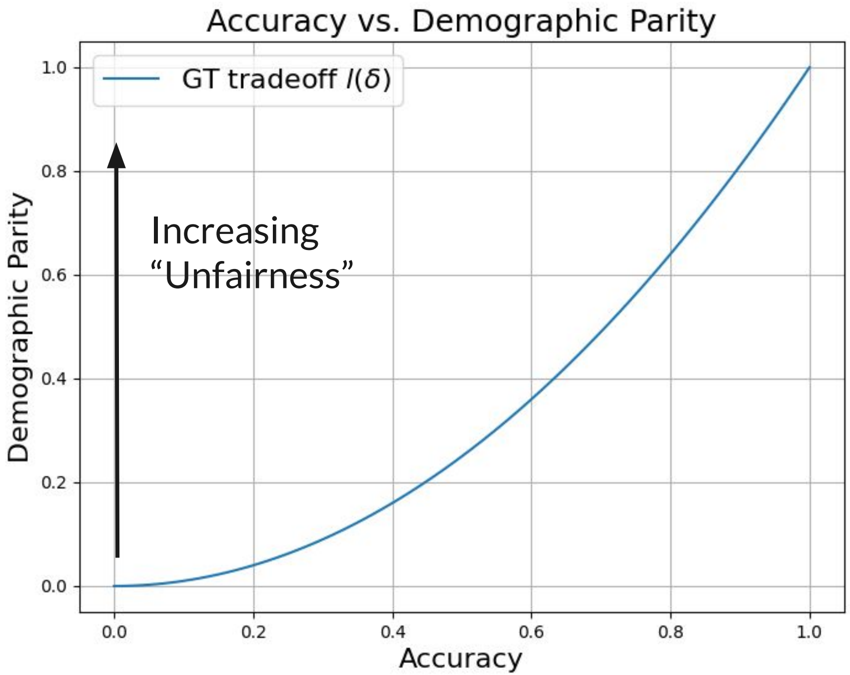
Fairness loss
E.g. demographic parity

Model accuracy

We cannot obtain the **exact** ground-truth tradeoff curve $l(\delta)$:

- We only have access to finite dataset

Problem statement



$$l(\delta) := \min_{h \in \mathcal{H}} \boxed{\mathcal{L}_{\mathrm{f}}(h)} \quad \text{subject to} \quad \boxed{\mathrm{acc}(h)} \geq \delta$$

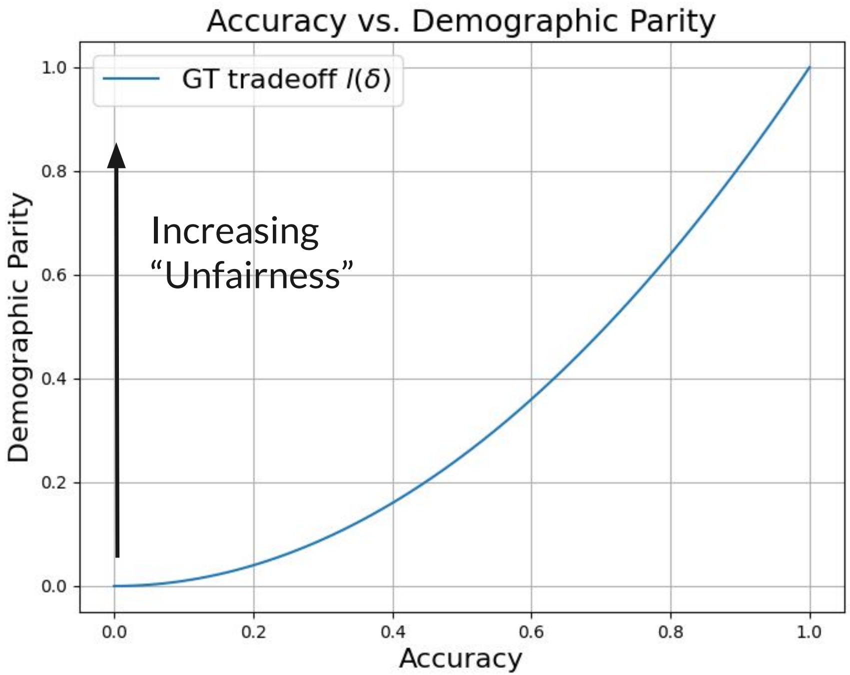
Fairness loss
E.g. demographic parity

Model accuracy

We cannot obtain the **exact** ground-truth tradeoff curve $l(\delta)$:

- We only have access to finite dataset
- The constrained optimisation problem shown above is non-trivial to solve

Problem statement



$$l(\delta) := \min_{h \in \mathcal{H}} \boxed{\mathcal{L}_{\mathrm{f}}(h)} \quad \text{subject to} \quad \boxed{\mathrm{acc}(h)} \geq \delta$$

Fairness loss
E.g. demographic parity

Model accuracy

We cannot obtain the **exact** ground-truth tradeoff curve $l(\delta)$:

- We only have access to finite dataset
- The constrained optimisation problem shown above is non-trivial to solve
- Can be computationally expensive

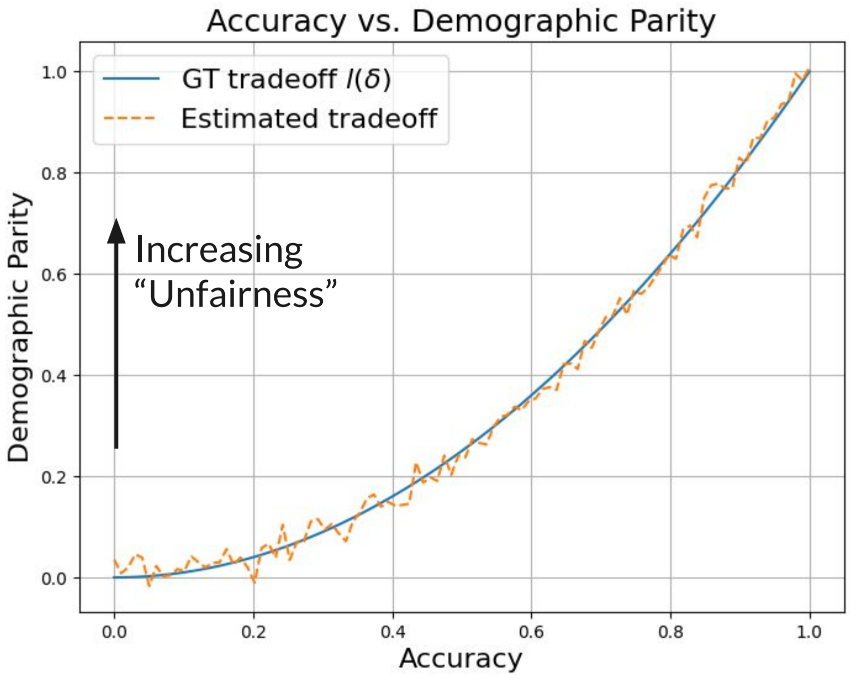
Methodology – Overview

Methodology

$$l(\delta) := \min_{h\in\mathcal{H}} \boxed{\mathcal{L}_{\mathrm{f}}(h)} \quad \text{subject to} \quad \boxed{\mathrm{acc}(h)} \geq \delta$$

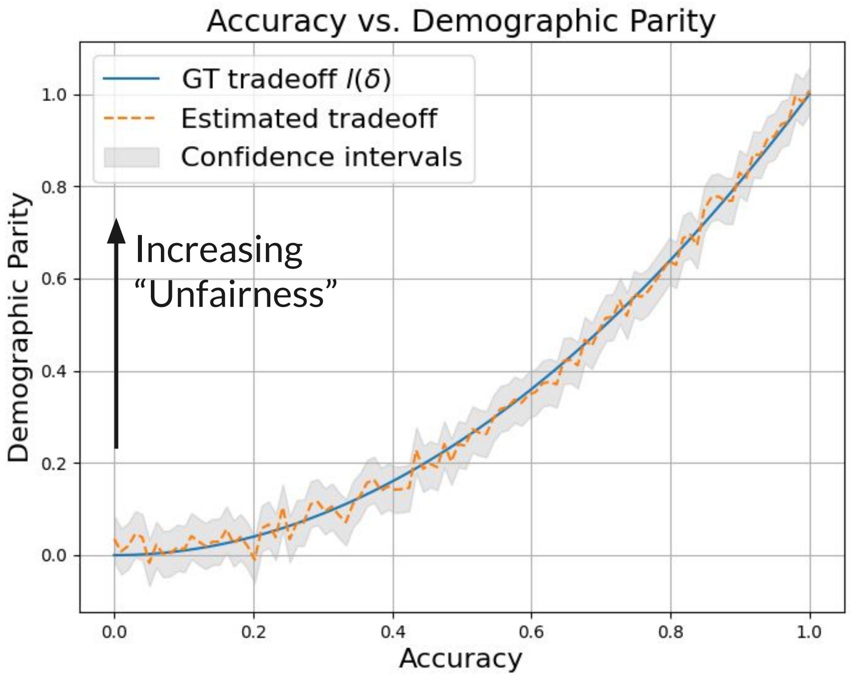
Fairness loss
E.g. demographic parity

Model accuracy



Step I – Computationally Efficient Estimation:
Estimate the trade-off curve $l(\delta)$ by training a single model

Methodology



$$l(\delta) := \min_{h \in \mathcal{H}} \boxed{\mathcal{L}_{\mathrm{f}}(h)} \quad \text{subject to} \quad \boxed{\mathrm{acc}(h)} \geq \delta$$

Fairness loss
E.g. demographic parity

Model accuracy

Step I – Computationally Efficient Estimation:
Estimate the trade-off curve $l(\delta)$ by training a single model

Step II – Calibration:
Using a held-out dataset, we construct confidence intervals which are going to contain the ground truth with probability at least $1 - \alpha$

Experimental results

Adult data experiments

X: data for some employees

A: gender

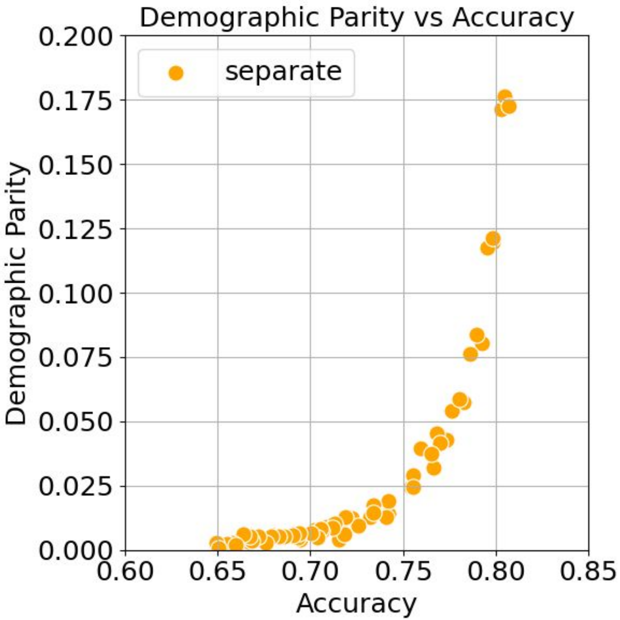
Y: whether salary is above $50k

Adult data experiments

X: data for some employees

A: gender

Y: whether salary is above $50k

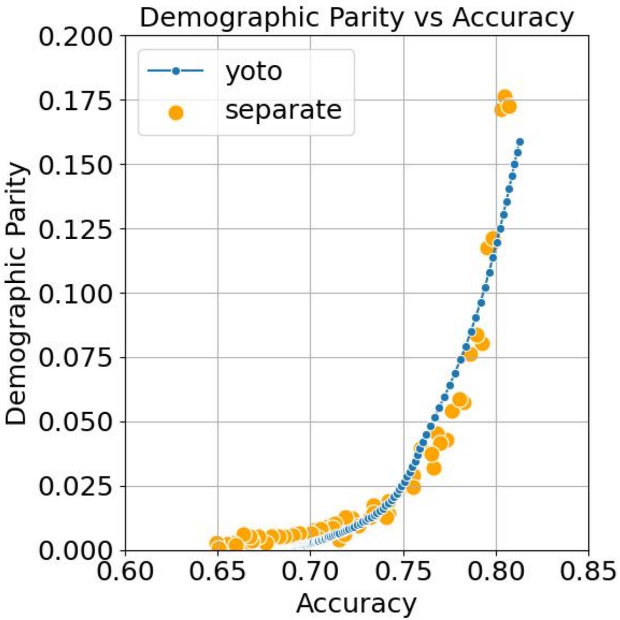


Adult data experiments

X: data for some employees

A: gender

Y: whether salary is above $50k



Trade-off Estimation

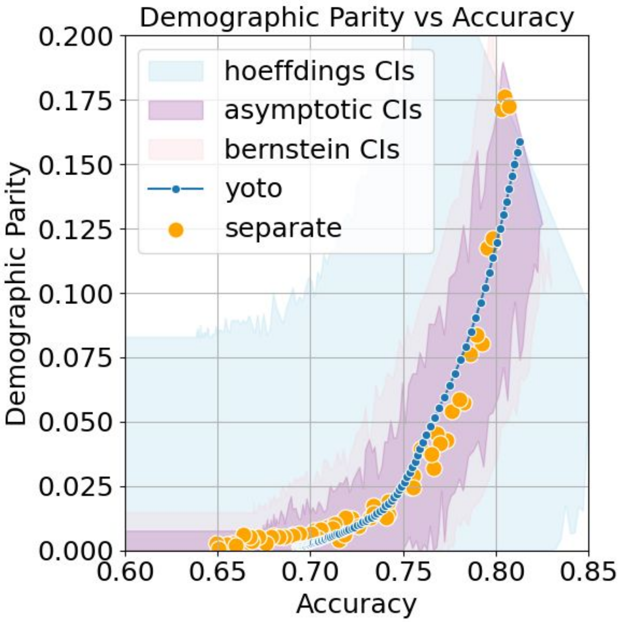
- YOTO trade-off curve is consistent with separately trained model

Adult data experiments

X: data for some employees

A: gender

Y: whether salary is above $50k



Trade-off Estimation

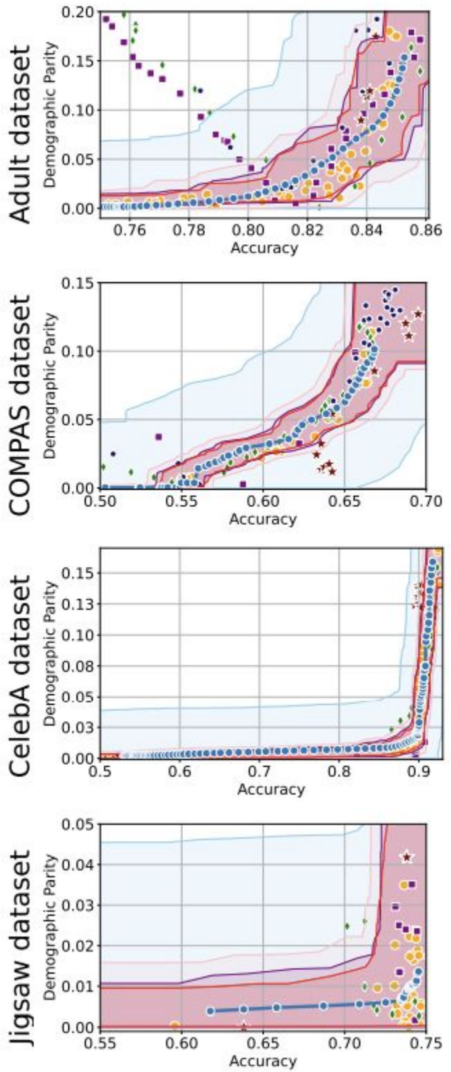
- YOTO trade-off curve is consistent with separately trained model

Confidence Intervals

- The Asymptotic Intervals are informative and cover the baselines
- Hoeffding's Intervals are conservative

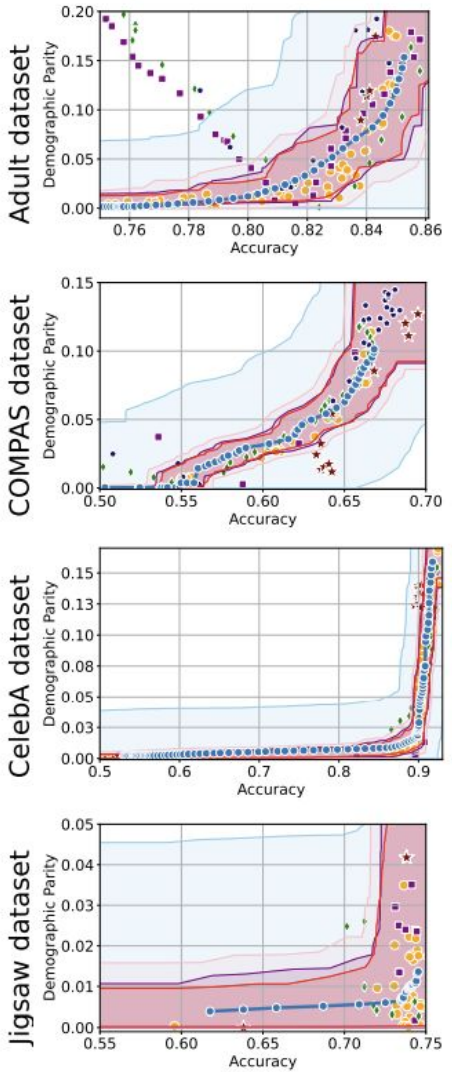
Key Takeaways

- The severity of accuracy-fairness trade-off fundamentally depends on dataset characteristics such as dataset imbalances or biases.



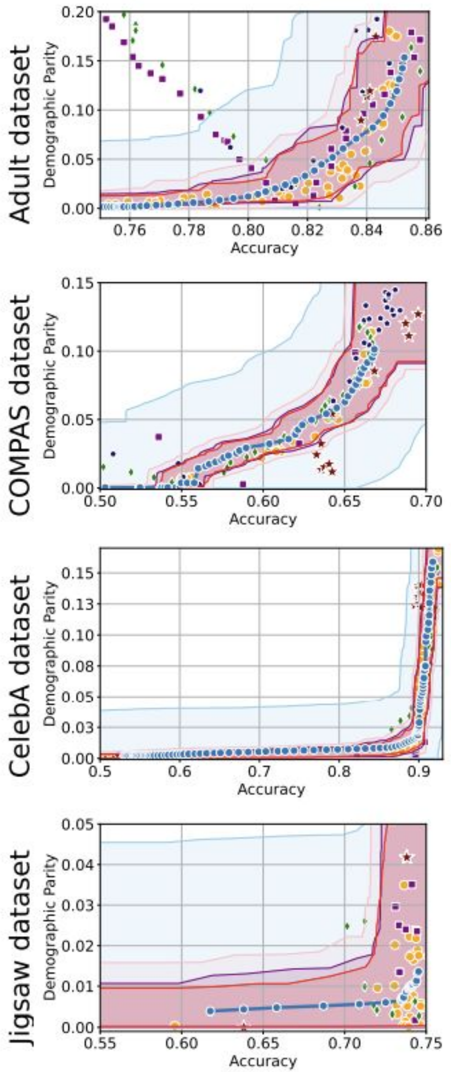
Key Takeaways

- The severity of accuracy-fairness trade-off fundamentally depends on dataset characteristics such as dataset imbalances or biases.
- We propose a computationally efficient approach to capture the fairness-accuracy trade-offs inherent to individual datasets, backed by sound statistical guarantees.



Key Takeaways

- The severity of accuracy-fairness trade-off fundamentally depends on dataset characteristics such as dataset imbalances or biases.
- We propose a computationally efficient approach to capture the fairness-accuracy trade-offs inherent to individual datasets, backed by sound statistical guarantees.
- The methodology provides the capability to specify desired accuracy levels and promptly receive corresponding admissible fairness violation ranges at inference time.



Thank you!

Check out our paper for additional details