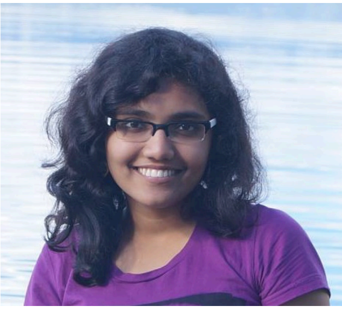
Carnegie Mellon University

IBM Research

Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing



Sanghamitra Dutta
sanghamd@andrew.cmu.edu

Dennis Wei
dwei@us.ibm.com

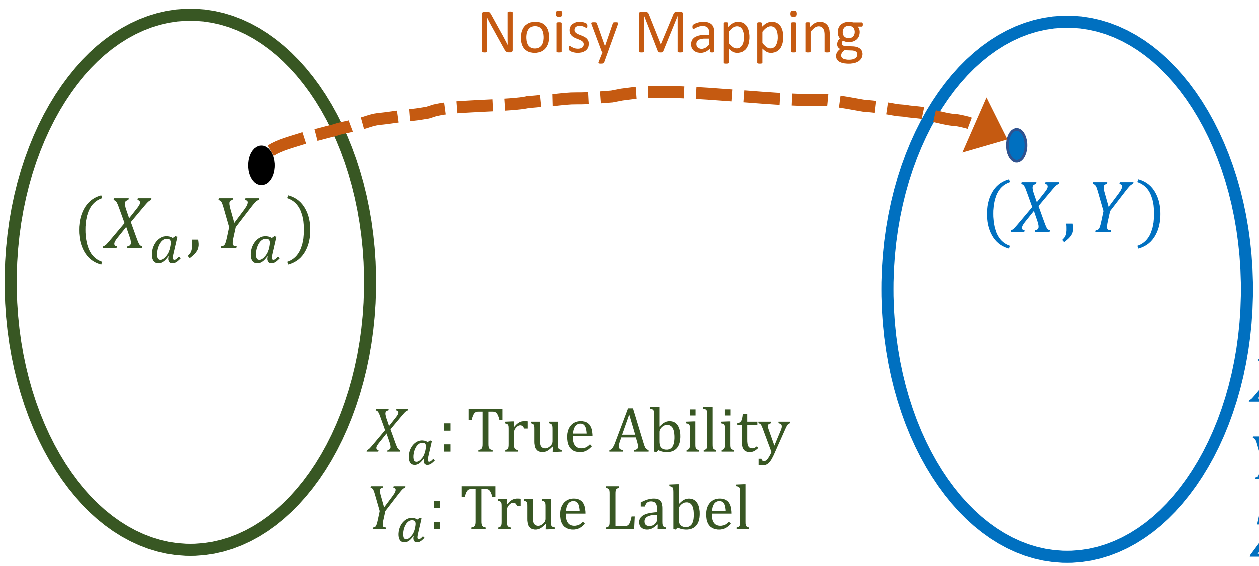
Hazar Yueksel
hazar.yueksel@ibm.com

Pin-Yu Chen
pin-yu.chen@ibm.com

Sijia Liu
sijia.liu@ibm.com

Kush Varshney
krvarshn@us.ibm.com

Motivational Example



X_a: True Ability
Y_a: True Label

Construct Space

X: Exam Score
Y: Data Label (0) or (1)
Z: Protected Attribute (Gender, Race, etc.)

Observed Space

No trade-off between accuracy and fairness

Bayes optimal classifier achieves fairness (Equal Opportunity)

Accuracy-fairness trade-off in observed space is due to noisier mappings for one group making the 0 and 1 labels “less separable”

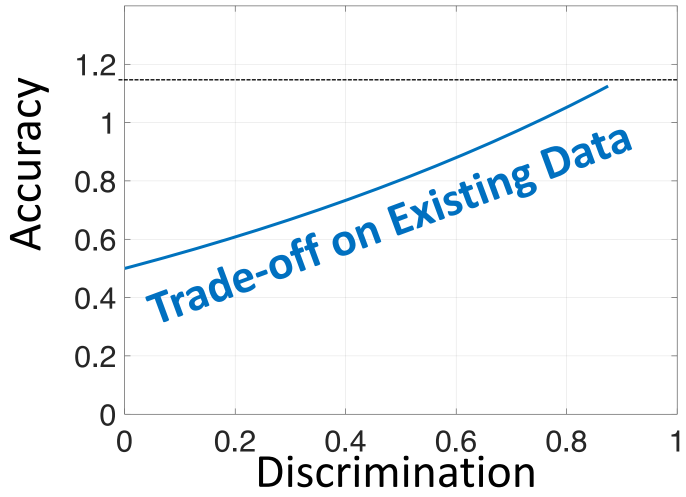
Setup inspired from [Friedler et al. ’16] [Yeom et al. ’18]; Definition of Equal Opportunity [Hardt et al. ‘16]

Main Contributions

Concept of Separability

Chernoff Information: approximation to best error exponent in binary classification

- **Explain the trade-off (Theorem 1)**
- **Compute fundamental limits**

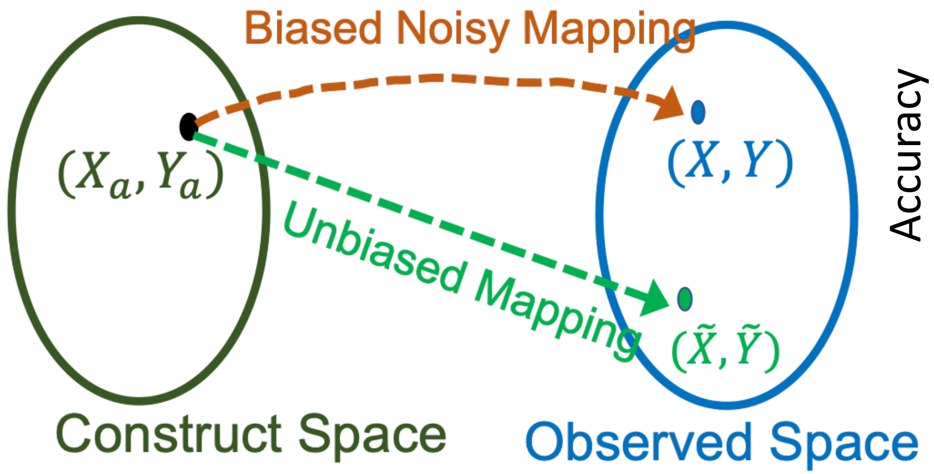


Accuracy with respect to observed dataset is a problematic measure of performance

Ideal Distributions

where accuracy and fairness are in accord

- **Proof of existence (Theorem 2)**
  With analytical forms
- **Interpretation**

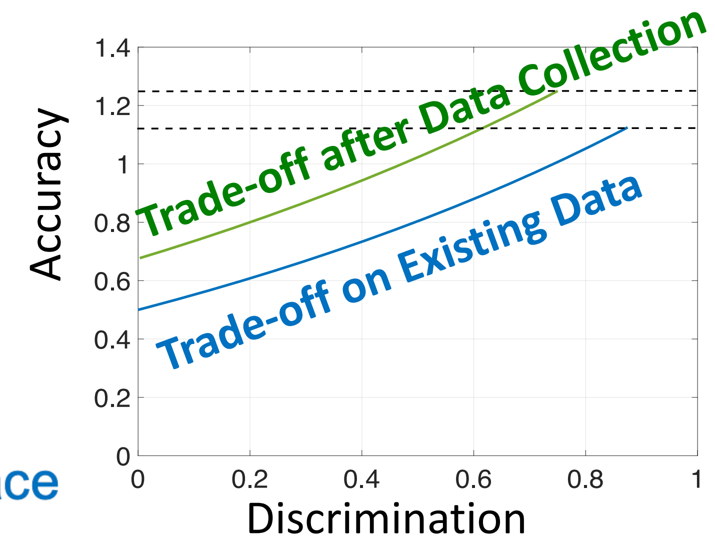


Plausible distributions in observed space, or distributions in the construct space

Alleviate Trade-off in Real World

Gather knowledge from active data collection, often improving separability

- **Criterion to alleviate (Theorem 3)**
- **Compute alleviated trade-off**



These results also explain why active fairness works

Related Works

- Characterizing Accuracy-Fairness Trade-Off

[Menon & Williamson ‘18] [Garg et al. ‘19]
[Chen et al. ‘18] [Zhao & Gordon ‘19]

Exponent Analysis with
Geometric Interpretability

- Empirical Datasets for Accuracy Evaluation

[Wick et al. ’19] [Sharma et al. ‘19]

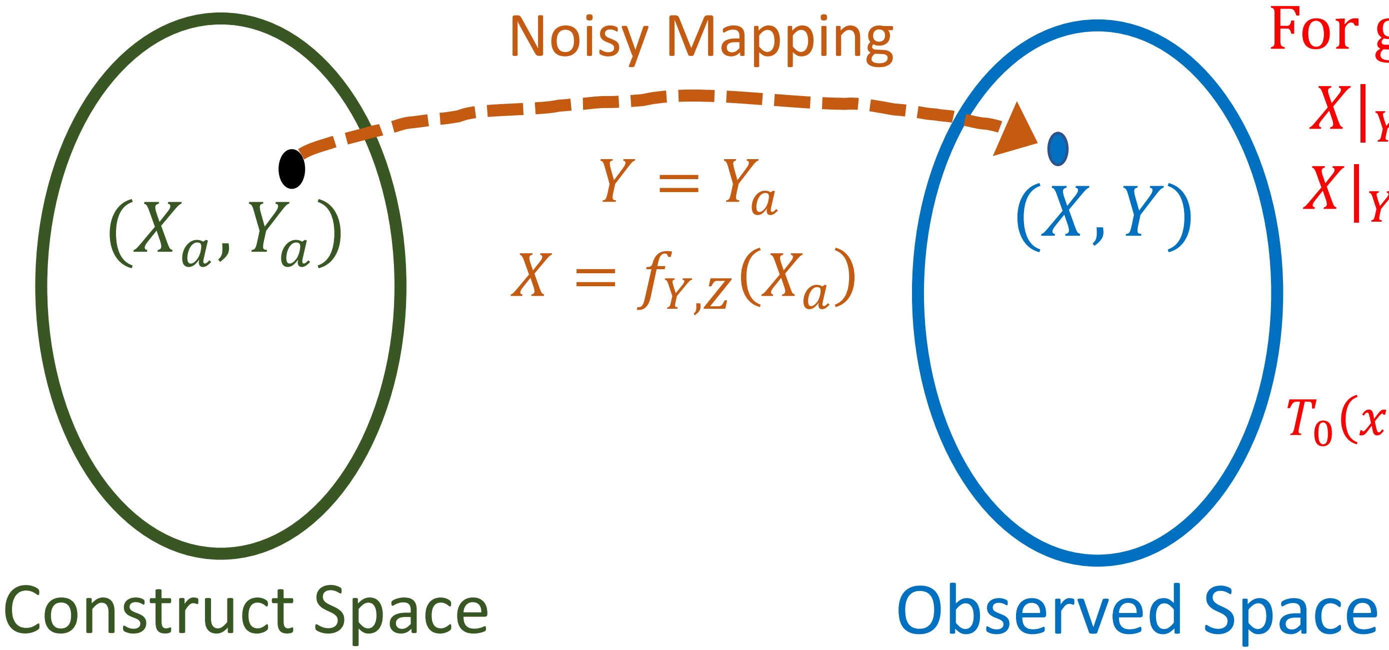
- Pre-processing Datasets for Fairness

[Calmon et al. ‘18] [Feldman et al. ‘15] [Zemel et al. ‘13]

- Explainability/ Active Fairness

[Varshney et al. ‘18] [Noriega-Campero et al. ‘19]

Preliminaries



For group Z=0,

$X|_{Y=0,Z=0} \sim P_0(x)$

$X|_{Y=1,Z=0} \sim P_1(x)$

For group Z=1,

$X|_{Y=0,Z=1} \sim Q_0(x)$

$X|_{Y=1,Z=1} \sim Q_1(x)$

$$T_0(x) = \log \frac{P_1(x)}{P_0(x)} \geq \tau_0 \qquad T_1(x) = \log \frac{Q_1(x)}{Q_0(x)} \geq \tau_1$$

EQUAL OPPORTUNITY→ EQUAL Prob. of FN

- Probability of **F**alse **N**egative(FN): $P_{FN,T_z}(\tau_z) = \Pr(T_z(x) < \tau_z | Y = 1, Z = z)$

  Wrongful Reject of True (+), i.e., True Y=1
- Probability of **F**alse **P**ositive(FP): $P_{FP,T_z}(\tau_z) = \Pr(T_z(x) \geq \tau_z | Y = 0, Z = z)$

  Wrongful Accept of True (−), i.e., True Y=0
- Probability of error: $P_{e,T}(\tau) = \pi_0 P_{FP,T}(\tau) + \pi_1 P_{FN,T}(\tau)$

Prior probabilities (assume $\pi_0 = \pi_1 = 1/2$)

Quick Background on Chernoff Error Exponents

$$P_{FN,T_z}(\tau_z) \lesssim e^{-E_{FN,T_z}(\tau_z)}$$

$$P_{FP,T_z}(\tau_z) \lesssim e^{-E_{FP,T_z}(\tau_z)}$$

Chernoff exponents of probabilities of FN and FP (Larger exponent → lower error)

Since $P_{e,T}(\tau) = \frac{1}{2} P_{FP,T}(\tau) + \frac{1}{2} P_{FN,T}(\tau)$, we define the Chernoff exponent of overall error probability as

$$E_{e,T_z}(\tau_z) = \min\{E_{FN,T_z}(\tau_z), E_{FP,T_z}(\tau_z)\}$$

(Larger exponent
→ lower error
→ higher accuracy)

Lemma: Chernoff exponent of error probability for Bayes optimal classifier between distributions $P_0(x)$ under $Y = 0$ and $P_1(x)$ under $Y = 1$:

Chernoff information $C(P_0, P_1) = -\log \min_{\alpha \in [0,1]} \sum P_0(x)^{\alpha} P_1(x)^{1-\alpha}$

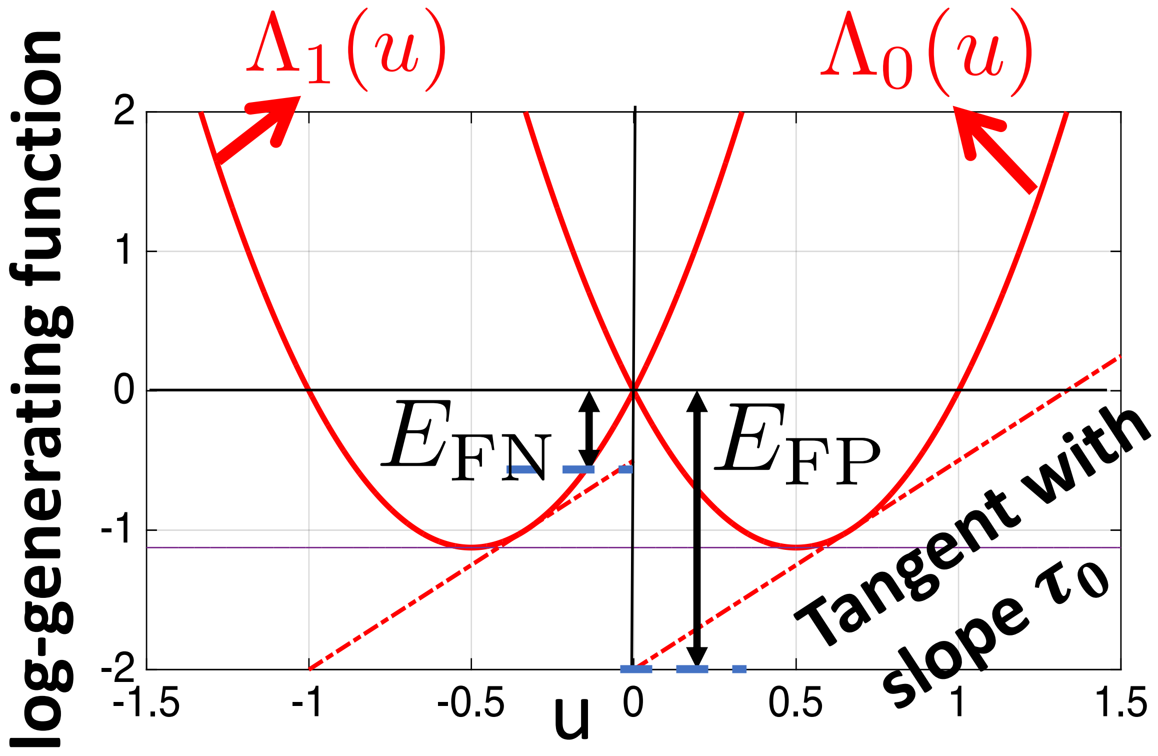
[Cover & Thomas]

Our Proposition: Concept of Separability

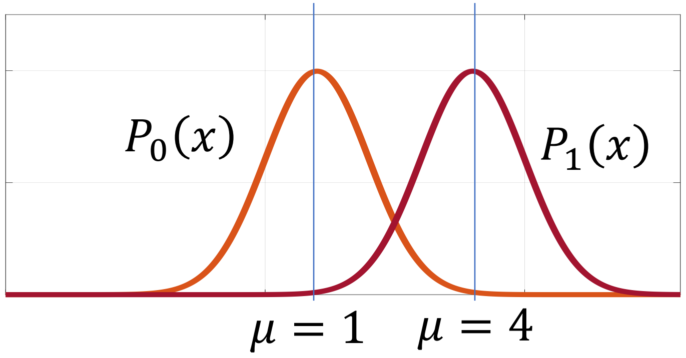
- **Definition of Separability:** For a group of people with data distributions $P_0(x)$ and $P_1(x)$ under hypotheses $Y = 0$ and $Y = 1$, we define the separability as their Chernoff information $C(P_0, P_1)$.

Geometric interpretability makes them tractable

Geometric understanding of the results



For group Z=0,
$P_0(x) \sim N\ (1,1)$
$P_1(x) \sim N(4,1)$



$T_0(x) \geq \tau_0$

$$\Lambda_0(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 0, Z = 0\right) = \frac{9}{2}u(u-1)$$

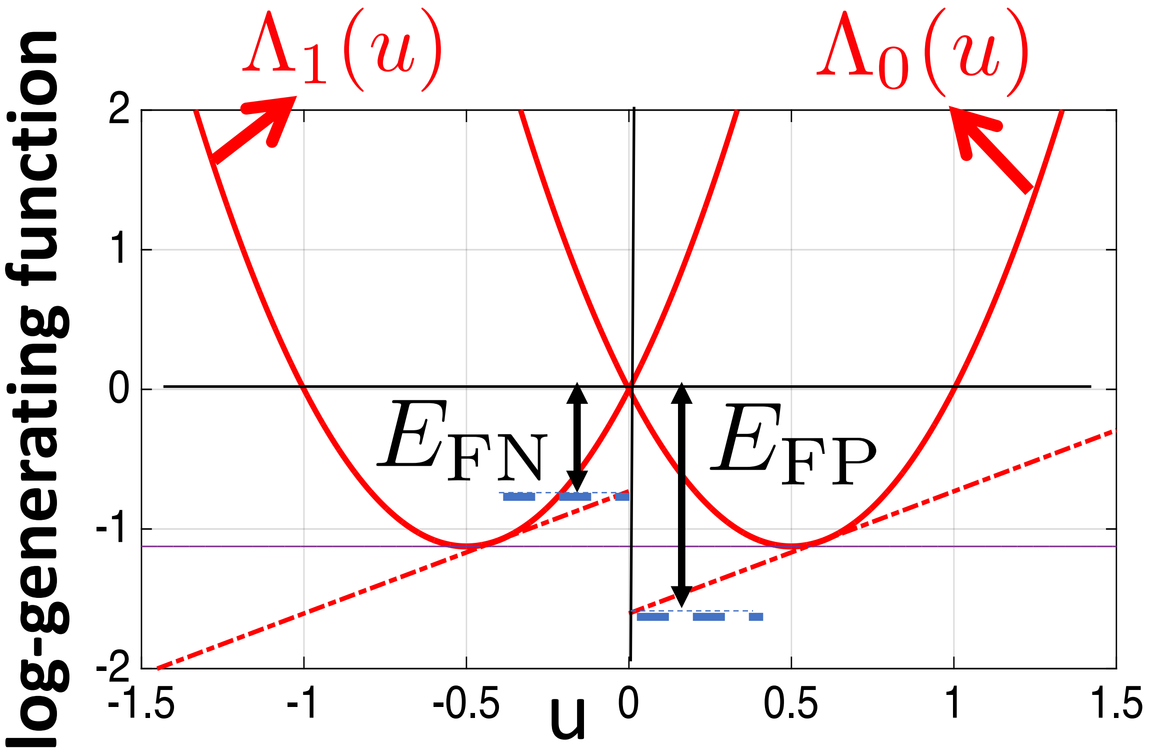
$$\Lambda_1(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 1, Z = 0\right) = \frac{9}{2}u(u+1)$$

$$E_{FP,T_0}(\tau_0) = \sup_{u>0}(u\tau_0 - \Lambda_0(u))$$

$$E_{FN,T_0}(\tau_0) = \sup_{u<0}(u\tau_0 - \Lambda_1(u))$$

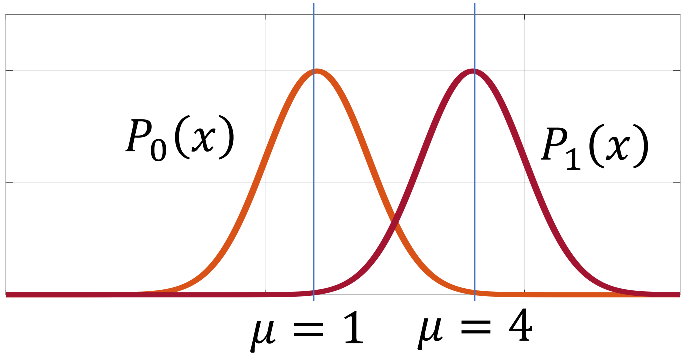
$$E_{e,T_0}(\tau_0) = \min\{E_{FN,T_0}(\tau_0), E_{FP,T_0}(\tau_0)\}$$

Geometric understanding of the results



For group Z=0,
$P_0(x) \sim N\ (1,1)$
$P_1(x) \sim N(4,1)$

$T_0(x) \geq \tau_0$



$$\Lambda_0(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 0, Z = 0\right) = \frac{9}{2}u(u-1)$$

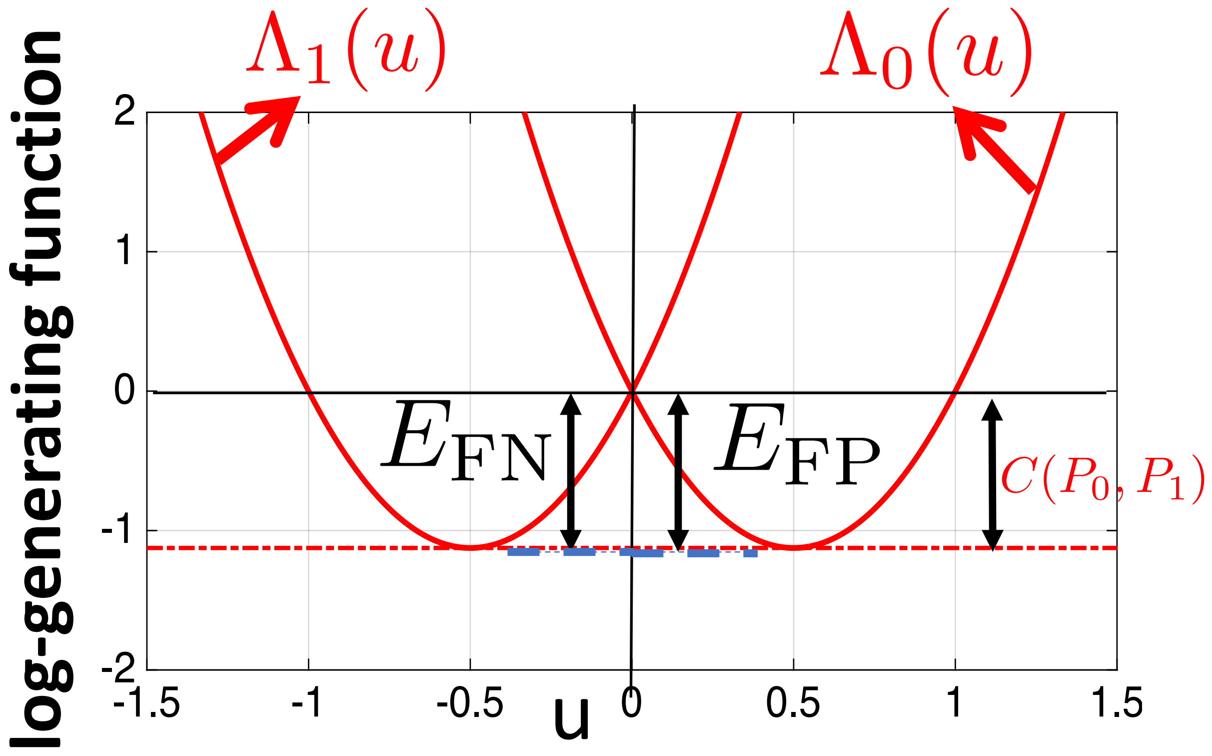
$$\Lambda_1(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 1, Z = 0\right) = \frac{9}{2}u(u+1)$$

$$E_{FP,T_0}(\tau_0) = \sup_{u>0}(u\tau_0 - \Lambda_0(u))$$

$$E_{FN,T_0}(\tau_0) = \sup_{u<0}(u\tau_0 - \Lambda_1(u))$$

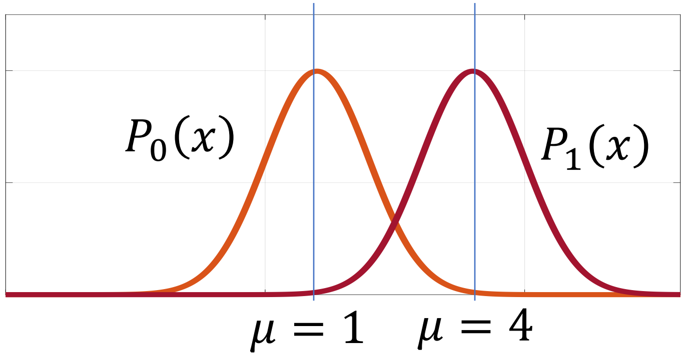
$$E_{e,T_0}(\tau_0) = \min\{E_{FN,T_0}(\tau_0), E_{FP,T_0}(\tau_0)\}$$

Geometric understanding of the results



$$E_{FN} = E_{FP} = C(P_0, P_1)$$

For group Z=0,
$P_0(x) \sim N\ (1,1)$
$P_1(x) \sim N(4,1)$



$$T_0(x) \geq \tau_0$$

$$\Lambda_0(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 0, Z = 0\right) = \frac{9}{2}u(u-1)$$

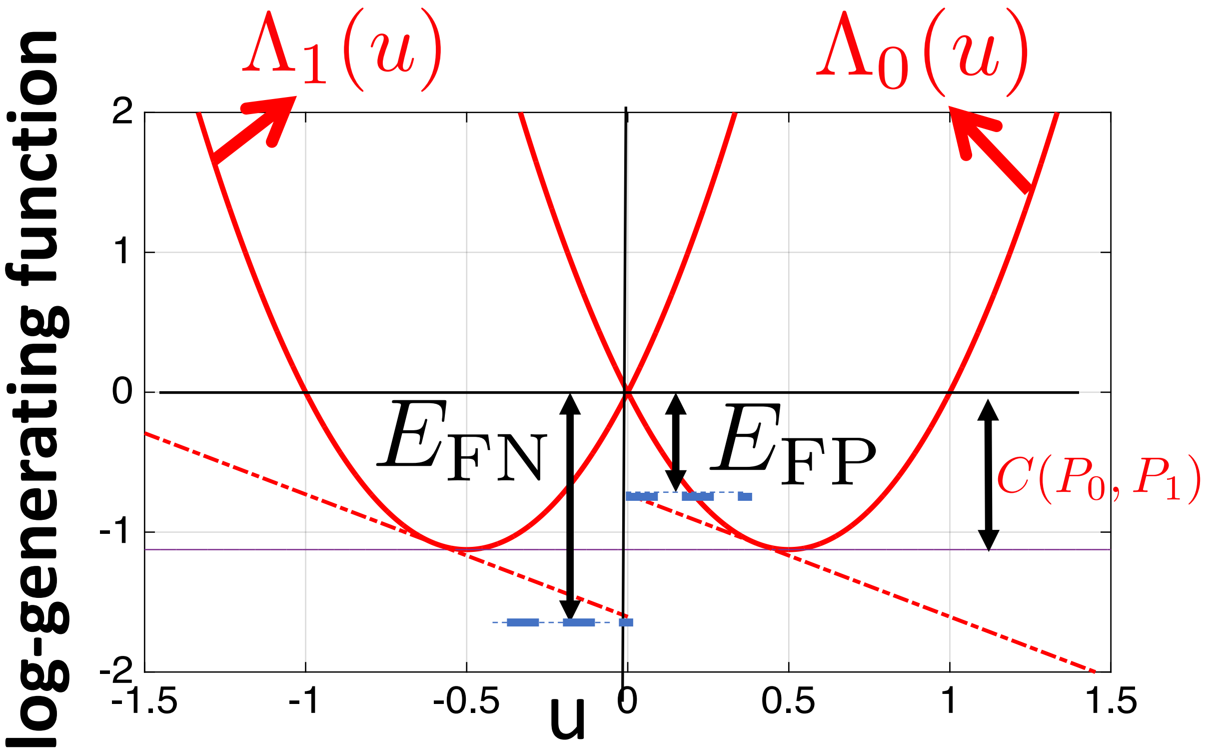
$$\Lambda_1(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 1, Z = 0\right) = \frac{9}{2}u(u+1)$$

$$E_{FP,T_0}(\tau_0) = \sup_{u>0}(u\tau_0 - \Lambda_0(u))$$

$$E_{FN,T_0}(\tau_0) = \sup_{u<0}(u\tau_0 - \Lambda_1(u))$$

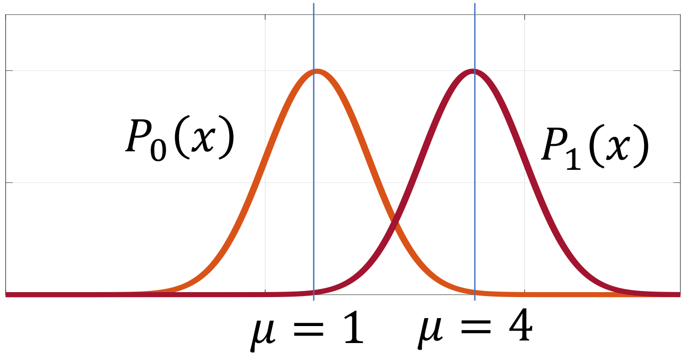
$$E_{e,T_0}(\tau_0) = \min\{E_{FN,T_0}(\tau_0), E_{FP,T_0}(\tau_0)\}$$

Geometric understanding of the results



For group Z=0,
$P_0(x) \sim N(1,1)$
$P_1(x) \sim N(4,1)$

$T_0(x) \geq \tau_0$



$$\Lambda_0(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 0, Z = 0\right) = \frac{9}{2}u(u-1)$$

$$\Lambda_1(u) = \mathbf{log\ E}\left(e^{uT_0(x)} \middle| Y = 1, Z = 0\right) = \frac{9}{2}u(u+1)$$

$$E_{FP,T_0}(\tau_0) = \sup_{u>0}(u\tau_0 - \Lambda_0(u))$$

$$E_{FN,T_0}(\tau_0) = \sup_{u<0}(u\tau_0 - \Lambda_1(u))$$

$$E_{e,T_0}(\tau_0) = \min\{E_{FN,T_0}(\tau_0), E_{FP,T_0}(\tau_0)\}$$

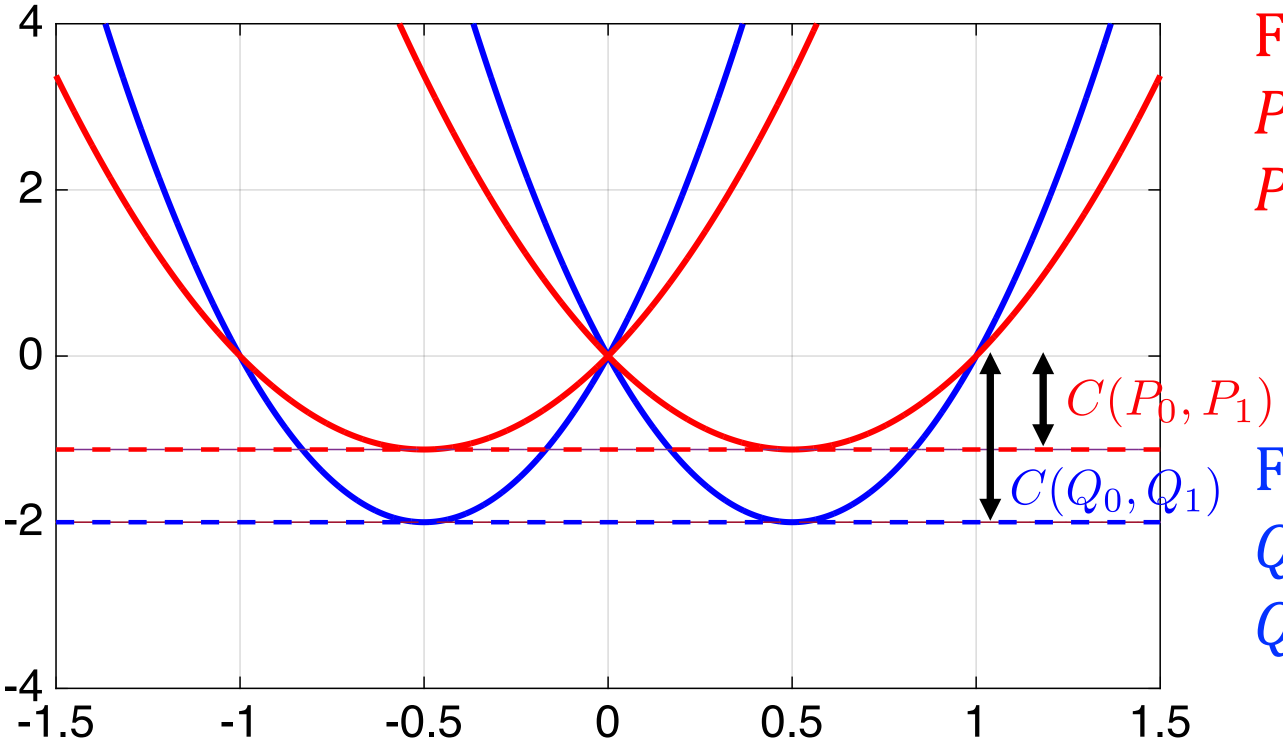
Accuracy-fairness trade-off is due to difference in separability of one group of people over another

Theorem 1 (informal): One of the following is true in observed space:

- Unbiased Mappings $C(P_0, P_1) = C(Q_0, Q_1)$: Bayes optimal classifiers for both groups also satisfy equal opportunity, i.e., $E_{FN,T_0}(\tau_0) = E_{FN,T_1}(\tau_1)$.
- Biased Mappings $C(P_0, P_1) < C(Q_0, Q_1)$: Given two classifiers (one for each group) that satisfy equal opportunity, for at least one of the groups it is not the Bayes optimal classifier, i.e.,

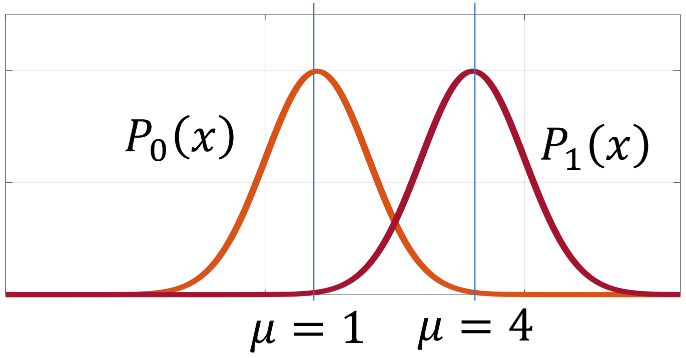
Either $E_{e,T_0}(\tau_0) < C(P_0, P_1)$ or $E_{e,T_1}(\tau_1) < C(Q_0, Q_1)$ or both

Geometric understanding of the results



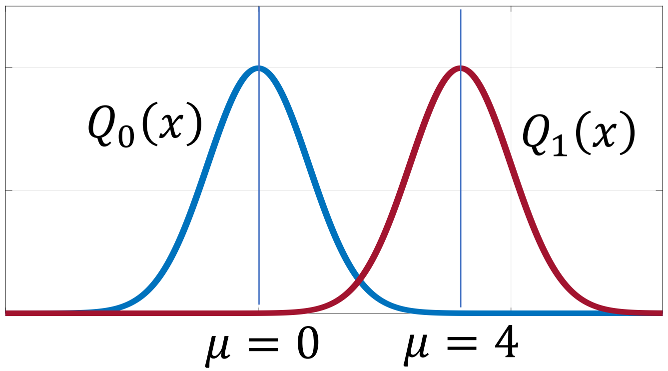
For group Z=0,
$P_0(x) \sim N\ (1{,}1)$
$P_1(x) \sim N(4{,}1)$

$T_0(x) \geq \tau_0$



For group Z=1,
$Q_0(x) \sim N\ (0{,}1)$
$Q(x) \sim N(4{,}1)$

$T_1(x) \geq \tau_1$

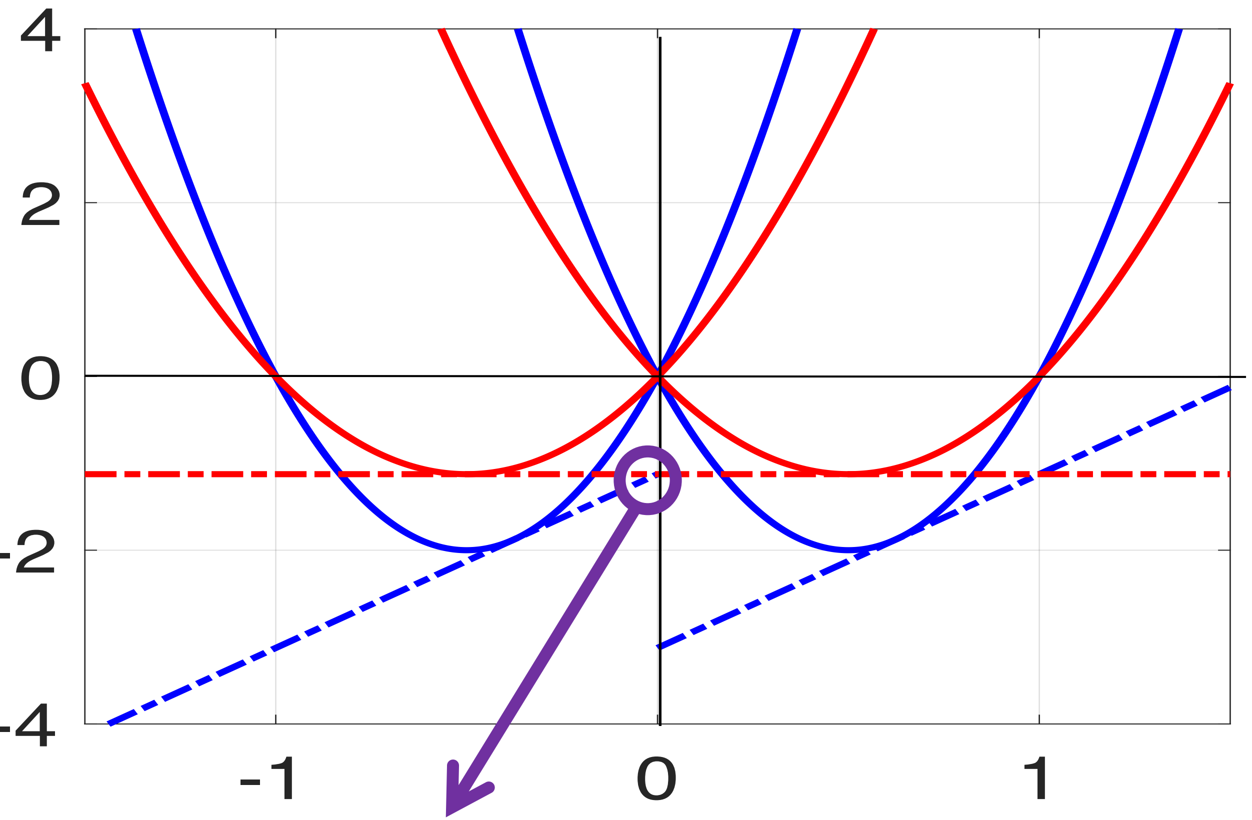


For group Z=0, we have $E_{FN} = E_{FP} = C(P_0, P_1)$

For group Z=1, we have $E_{FN} = E_{FP} = C(Q_0, Q_1)$

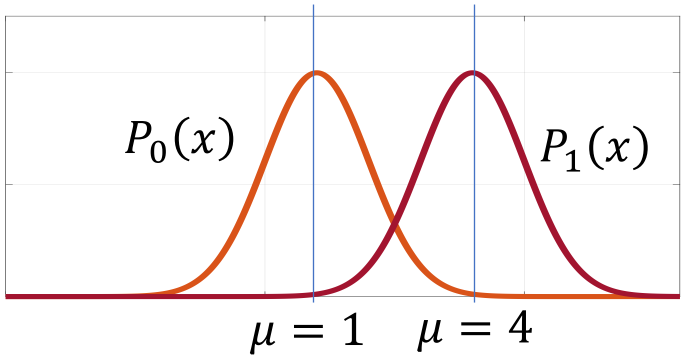
Bayes optimal classifiers do not satisfy Equal Opportunity (unequal E_{FN})

Geometric understanding of the results



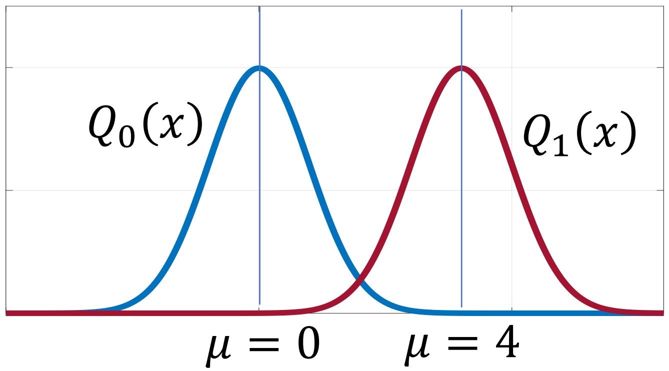
For group Z=0,
$P_0(x) \sim N\ (1,1)$
$P_1(x) \sim N(4,1)$

$T_0(x) \geq \tau_0$



For group Z=1,
$Q_0(x) \sim N\ (0,1)$
$Q(x) \sim N(4,1)$

$T_1(x) \geq \tau_1$

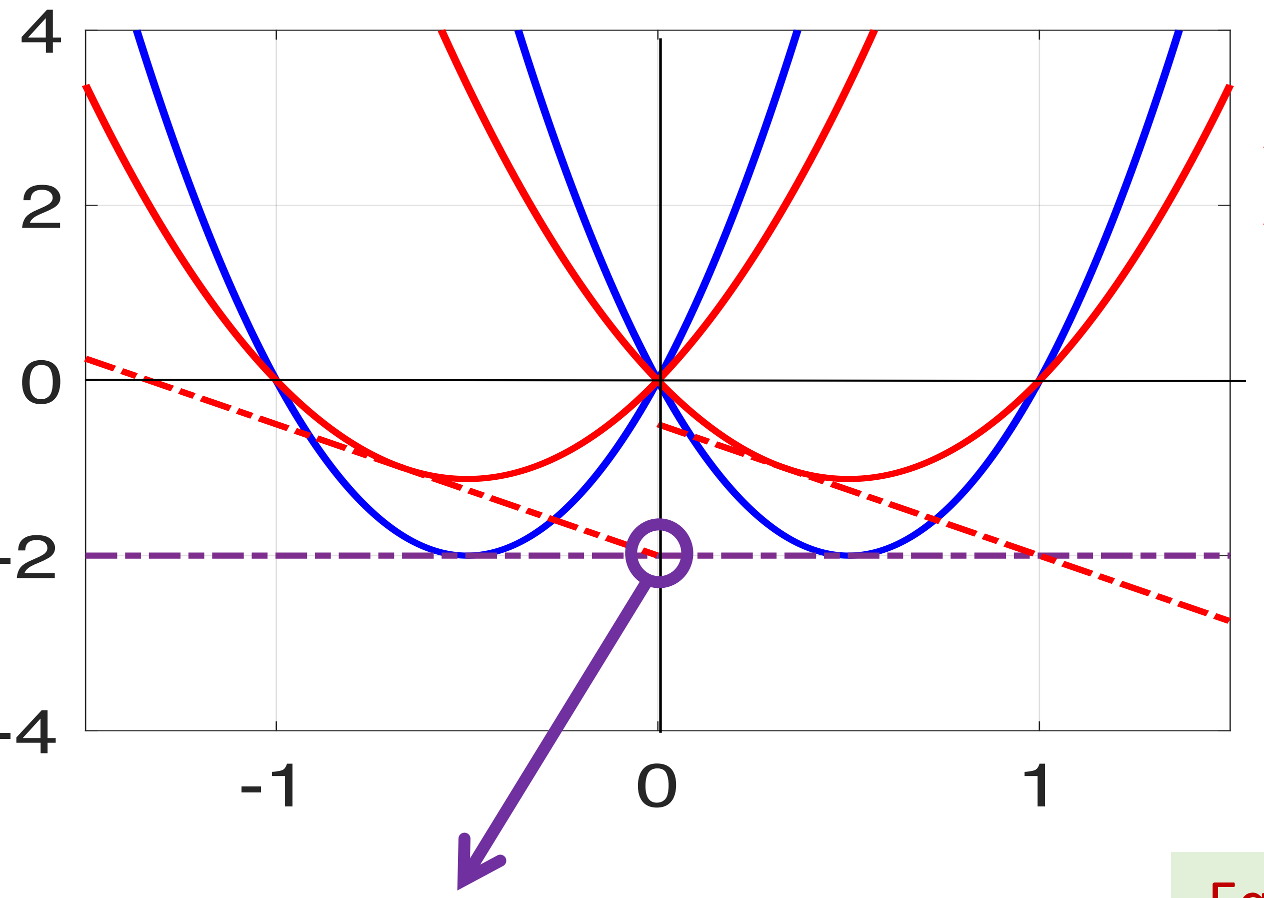


$E_{\mathrm{FN},T_0}(\tau_0) = E_{\mathrm{FN},T_1}(\tau_1)$

Equal Opportunity (equal E_{FN}) satisfied but sub-optimal for privileged group Z=1

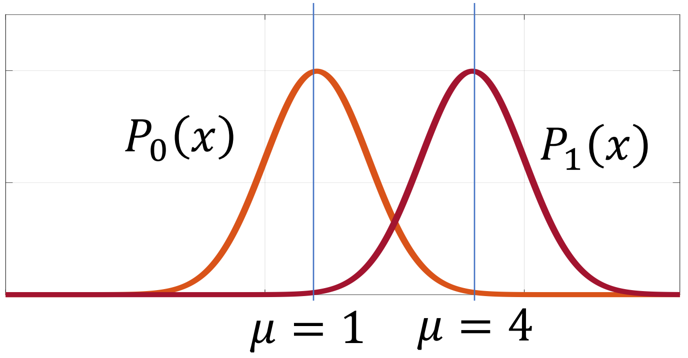
Avoid active harm to privileged group?

Geometric understanding of the results



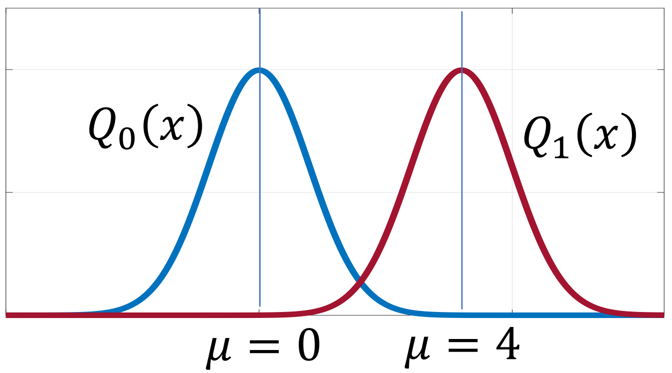
For group Z=0,
$P_0(x) \sim N\ (1,1)$
$P_1(x) \sim N(4,1)$

$T_0(x) \geq \tau_0$



For group Z=1,
$Q_0(x) \sim N\ (0,1)$
$Q(x) \sim N(4,1)$

$T_1(x) \geq \tau_1$



$E_{\mathrm{FN},T_0}(\tau_0) = E_{\mathrm{FN},T_1}(\tau_1)$

Equal Opportunity (equal E_{FN}) satisfied but sub-optimal for unprivileged group Z=0

For at least one of the groups, accuracy on given data is compromised for fairness.

Ideal distributions where accuracy and fairness are in accord

Theorem 2 (informal): Fix Bayes optimal classifier for privileged group *Z=1*. Then, for group *Z=0*, there exists ideal distributions of the forms

$$\widetilde{P}_0(x) = \frac{P_0(x)^{(1-w)} P_1(x)^w}{\sum_x P_0(x)^{(1-w)} P_1(x)^w}$$ and $$\widetilde{P}_1(x) = \frac{P_0(x)^{(1-v)} P_1(x)^v}{\sum_x P_0(x)^{(1-v)} P_1(x)^v}$$

such that:

- Fairness on given data: The Bayes optimal classifier for the new distributions is fair on given data (in fact it is the same classifier $T_0^*(x) \geq \tau_0^*$ that was sub-optimal but fair on the given data).
- Fairness and Optimal Accuracy on ideal data: On the ideal data, this Bayes optimal classifier also has $E_{\mathrm{FN}} = C\left(\tilde{P}_0, \tilde{P}_1\right) = C(Q_0, Q_1)$.

Proof of existence of ideal distributions (with analytical forms)

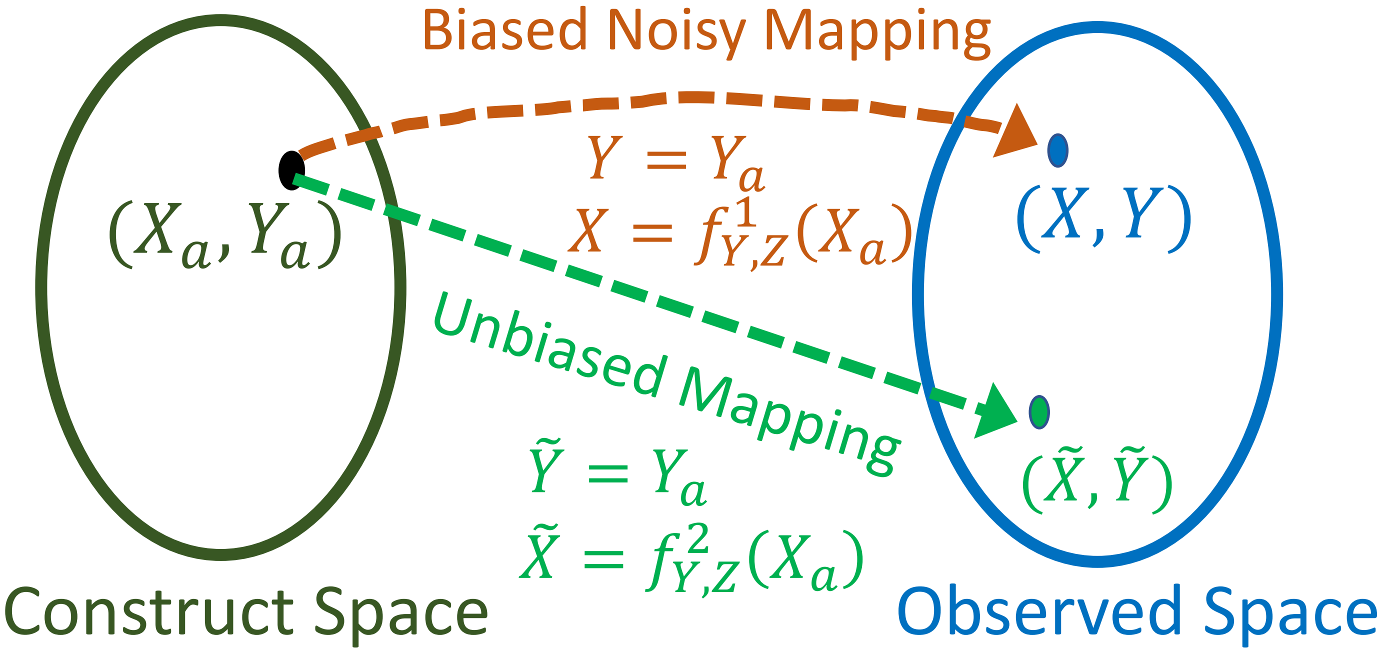
How to go about finding such ideal distributions?

$$\min_{\widetilde{P}_0, \widetilde{P}_1} \pi_0 \mathrm{D}(\widetilde{P}_0 || P_0) + \pi_1 \mathrm{D}(\widetilde{P}_1 || P_1)$$

$$\text{such that, } E_{\mathrm{FN}, \widetilde{T_0}}(0) = \mathrm{C}(Q_0, Q_1)$$

where $\widetilde{T_0}(x) = \log \frac{\widetilde{P_1}(x)}{\widetilde{P_0}(x)} \geq 0$ is the Bayes optimal classifier for the ideal distributions.

How to interpret these ideal distributions?



For group Z=1,

$\tilde{X}|_{Y=0,Z=1} \sim Q_0(x)$

$\tilde{X}|_{Y=1,Z=1} \sim Q_1(x)$

For group Z=0,

$\tilde{X}|_{Y=0,Z=0} \sim \widetilde{P_0}(x)$

$\tilde{X}|_{Y=1,Z=0} \sim \widetilde{P_1}(x)$

Plausible distributions in observed space under unbiased mappings, or candidate distributions in the construct space under identity mappings

When does active data collection alleviate the accuracy-fairness trade-off in the real world?

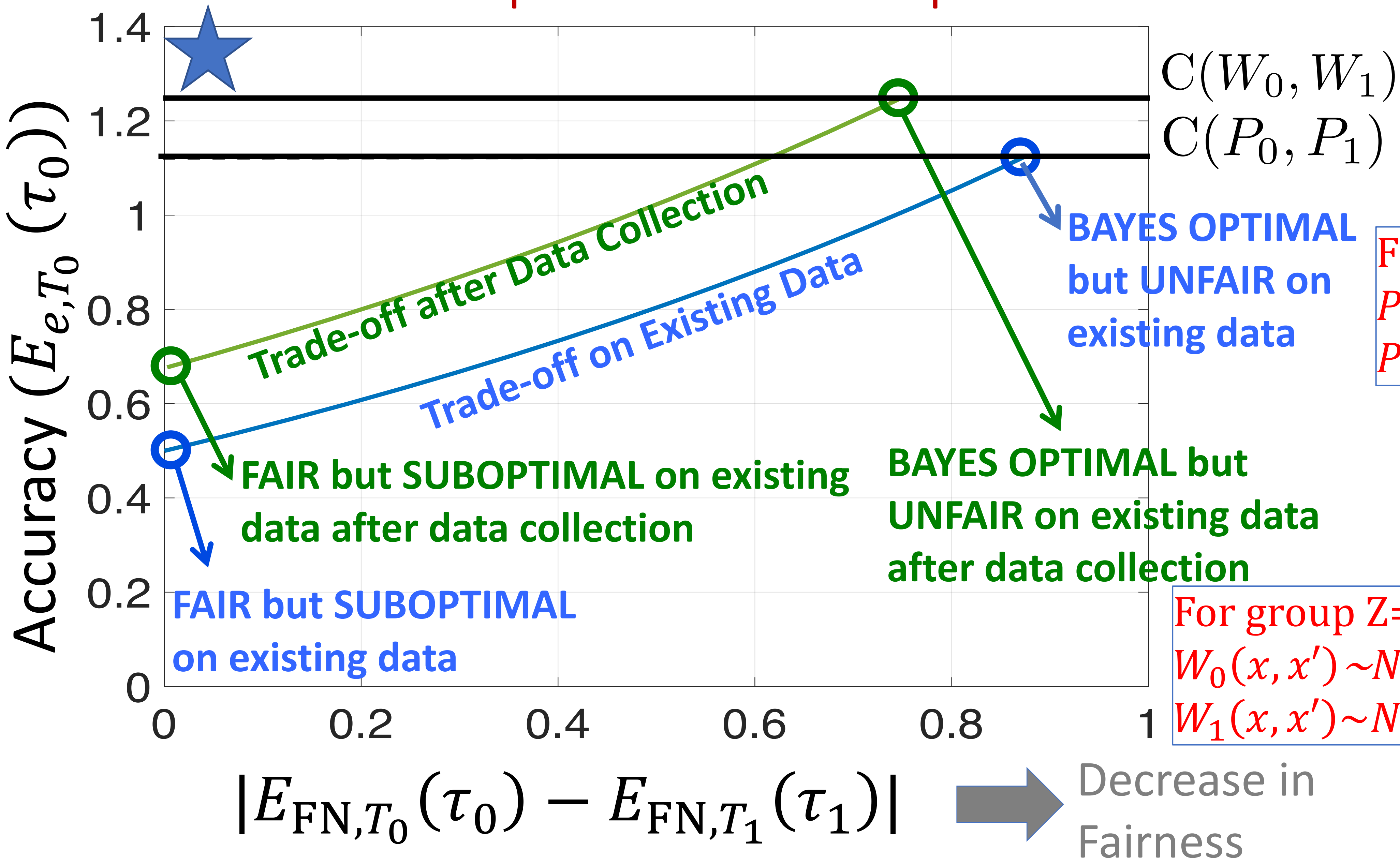
X' : New feature collected for *Z=0*

$$X, X'|_{Y=0,Z=0} \sim W_0(x, x') \qquad X, X'|_{Y=1,Z=0} \sim W_1(x, x')$$

Theorem 3: The separability $C(W_0, W_1)$ is strictly greater than $C(P_0, P_1)$ if and only if the conditional mutual information $I(X'; Y|X, Z = 0) > 0$.

Improving separability alleviates the accuracy-fairness trade-off in the real world

Numerical example: Exact computation of the trade-off



For group Z=0,
$P_0(x) \sim N(1,1)$
$P_1(x) \sim N(4,1)$

For group Z=0,
$W_0(x, x') \sim N((1,1), \mathbf{I}_{2x2})$
$W_1(x, x') \sim N((4,2), \mathbf{I}_{2x2})$

Summary

- Provides new tools that go beyond explaining accuracy-fairness trade-off
- Geometric interpretability helps exact quantification of this trade-off
- Separability, ideal distributions and their connection to construct space
- Criterion to alleviate the trade-off explains success of active fairness

Thank You!