The Science of Fair ML

Things That Can Go Wrong

- Distribution of x differs between groups
- Some group underrepresented in data S
- Features in x are less predictive for some group
- Some group is just harder to predict
- The y's are biased in the first place

Fairness in ML

- Typically a property of a **model** (ML algo output)
- Exceptions: online decision-making, RL, bandit settings
- Multiple **types** of fairness definitions

Types of Model Fairness

- Group fairness (most common)
- Individual fairness
- Interpolations between the two
- Others (causal, fair representations, ...)

Group Fairness Notions

Start by identifying:

- groups or attributes we wish to "protect" (e.g. race, gender)
- what constitutes harm (e.g. error, false pos/neg)

Choices are subjective & domain-specific

Then seek to equalize rates of harm across groups.

Example:

- domain: consumer lending
- groups: male & female
- harm: false rejection (negs)

Want to find model $h(x)$ s.t.

$$FN(h, male) \approx FN(h, female)$$

↳ allows for optimization of overall error

Confusion Matrices

$$FN(h, \text{male}) \triangleq \Pr_{\langle x,y\rangle \sim P}\left[h(x) = -1 \mid x \text{ male} \;\&\; y = +1\right]$$

$$y:$$

$$\begin{array}{cc} & \begin{array}{cc} +1 & -1 \end{array} \\ h(x): \begin{array}{c} +1 \\ -1 \end{array} & \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array} \end{array} \qquad\qquad \begin{array}{|c|c|} \hline a' & b' \\ \hline c' & d' \\ \hline \end{array}$$

Group 1 $\qquad\qquad$ Group 2

$$\frac{c}{a+c} \qquad\qquad \frac{c'}{a'+c'}$$

Note: We can achieve = FN rates by randomization.

If individual x, predict $\hat{y} = +$ with prob. p

If $y = -$, can't be a FN

If $y = +$, $\hat{y} = -$ w.p. $1-p$

$$\therefore FN(p,*) = 1-p$$

If we are given a model $h(x)$ & have access to group membership, easy to audit $h(x)$ for fairness.

How can we learn a fair model $h(x)$?

Why won't standard ML algos work?

A Post-Processing Approach ("bolt on")

- start with non-fair $h(x)$, want to $\approx$ M/F error rates
- build a probabilistic classifier on top of $h(x)$:

$$h(x): \quad \begin{array}{c|c|c} & M & F \\ \hline + & p & q \\ \hline - & r & s \end{array}$$

prob. $\tilde{h}(x) = +$

$$\tilde{h}(x)$$

(closed under mixtures)

$p = q = 1, \; r = s = 0$:

$\tilde{h} \equiv h, \; \varepsilon(\tilde{h}) = \varepsilon(h)$

$p = q = r = s = 1/2$:

$\varepsilon(\tilde{h}) = 1/2$, perfectly fair

$p = r = 1/2, \; q = s = 1$:

error on men $= 1/2$

error on women $=$ same as h

etc.

Equalizing Error Rates

- Let group error rates be $\varepsilon_M, \varepsilon_F$
- Suppose $1/2 \geq \varepsilon_F \geq \varepsilon_M$
- Pick $q=1, s=0 \Rightarrow \tilde{\varepsilon}_F = \varepsilon_F$
- Pick $r = 1-p$, agree with h with prob p pointwise
- $\tilde{\varepsilon}_M = p\,\varepsilon_M + (1-p)(1-\varepsilon_M)$

$$= p\,\varepsilon_M + 1 - \varepsilon_M - p + p\,\varepsilon_M$$

$$= 2p\,\varepsilon_M - p + 1 - \varepsilon_M$$

$$= p(2\varepsilon_M - 1) + 1 - \varepsilon_M$$

Now solve for p:

$$p(2\varepsilon_m - 1) + 1 - \varepsilon_m = \hat{\varepsilon}_F = \varepsilon_F$$

$$p(2\varepsilon_m - 1) = \varepsilon_F + \varepsilon_m - 1$$

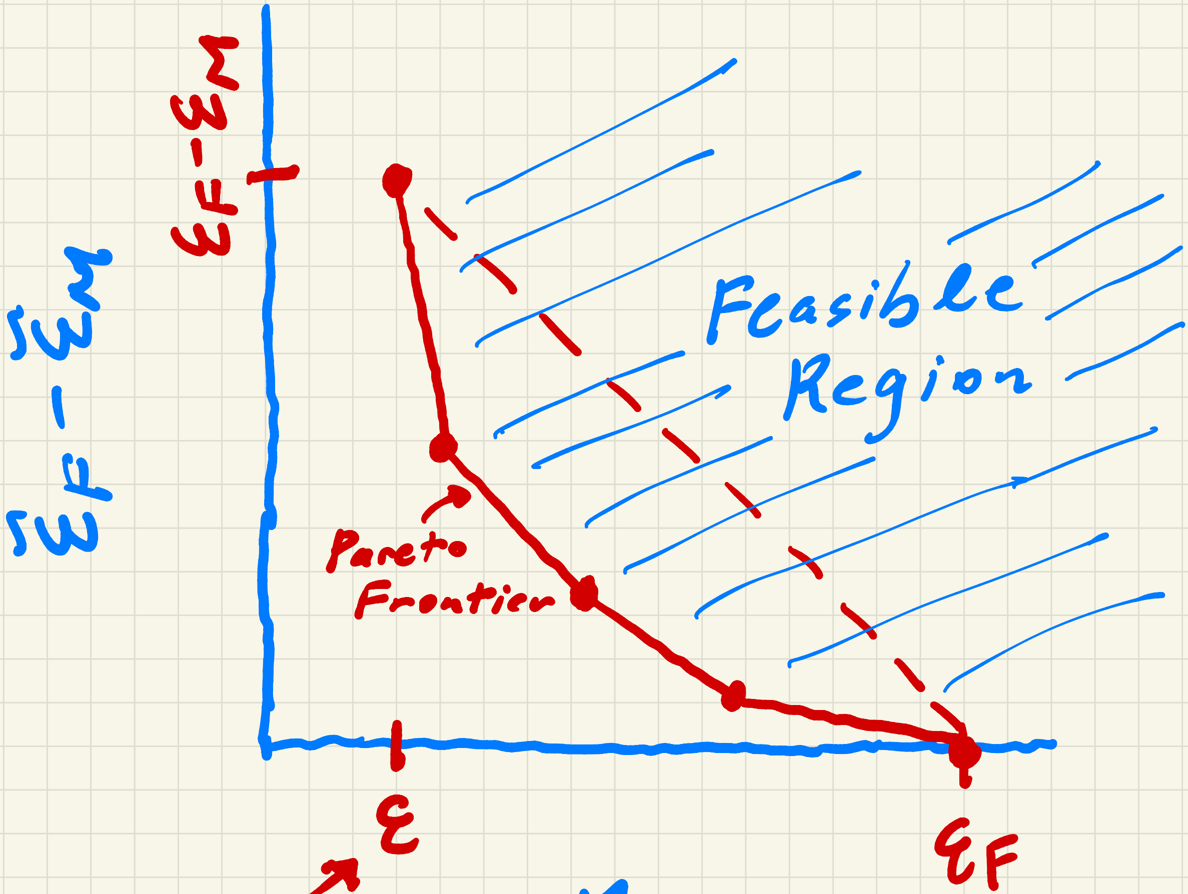
$$p = \frac{\varepsilon_F + \varepsilon_m - 1}{2\varepsilon_m - 1}$$

Sanity Checks:

- $\varepsilon_m, \varepsilon_F \leq 1/2 \Rightarrow$ num & denom ≤ 0
  $\Rightarrow p \geq 0$
- $\varepsilon_m = \varepsilon_F \Rightarrow p = \frac{0}{0} = 1, \ r = 0$
- $\varepsilon_F = 1/2 \Rightarrow p = \dfrac{1/2 + \varepsilon_m - 1}{2\varepsilon_m - 1} = \dfrac{\varepsilon_m - 1/2}{2\varepsilon_m - 1}$

  $= \dfrac{\varepsilon_m - 1/2}{2(\varepsilon_m - 1/2)} = 1/2, \ r = 1/2$

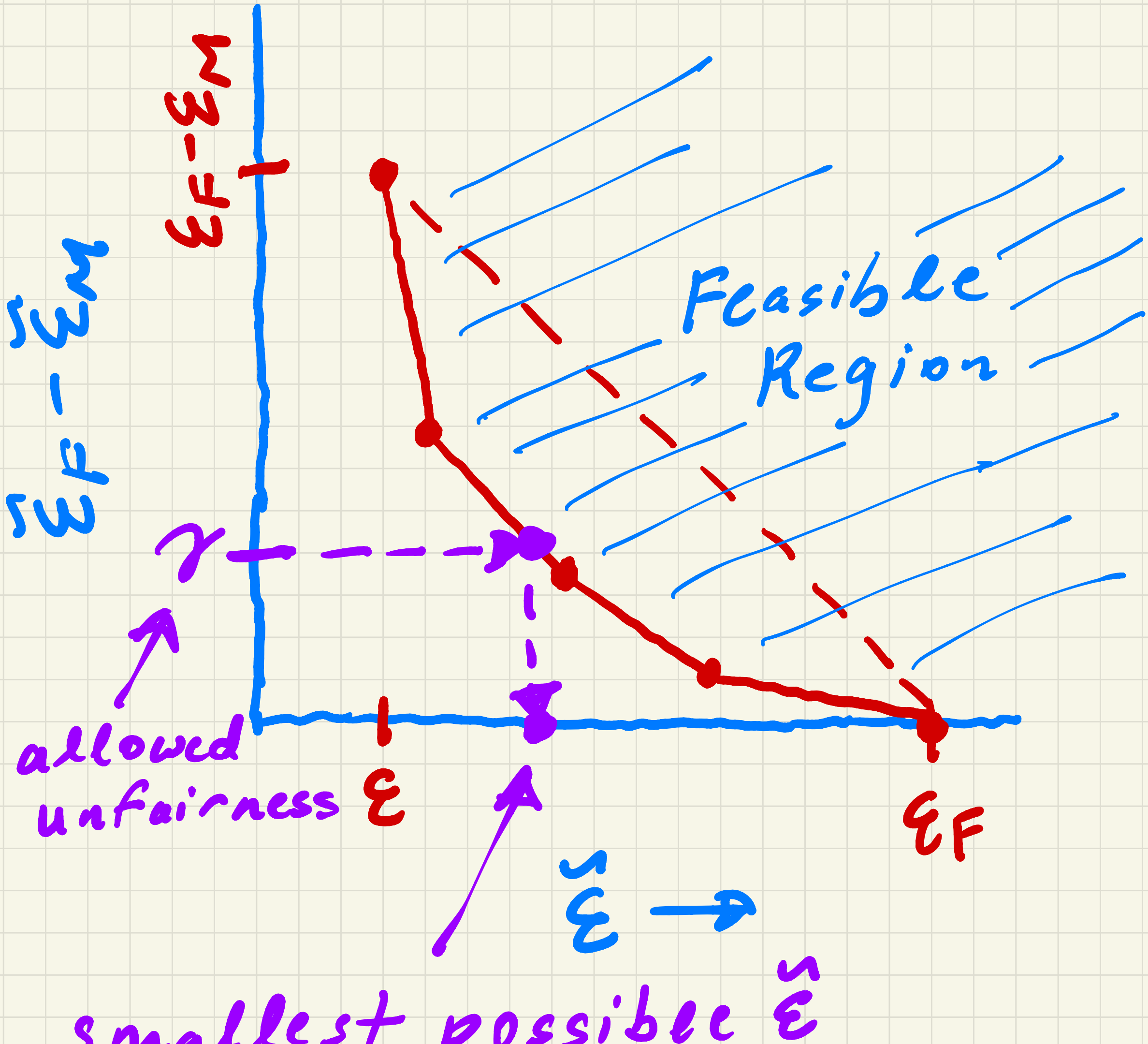
Feasible Region & Pareto Frontier



$\tilde{\varepsilon} \rightarrow$

$\alpha \varepsilon_F + (1-\alpha)\varepsilon_M$
$< \varepsilon_F$

Finding the Frontier



smallest possible $\hat{\epsilon}$ given γ

Algorithm

- Problem of finding $\tilde{h}$ than minimizes $\varepsilon(\tilde{h})$

  subject to

  y-axis $\leq \gamma$

  is a linear program

  in p, q, r, s.

(Framework & result due to Hardt, Price, Srebro.)

Recap

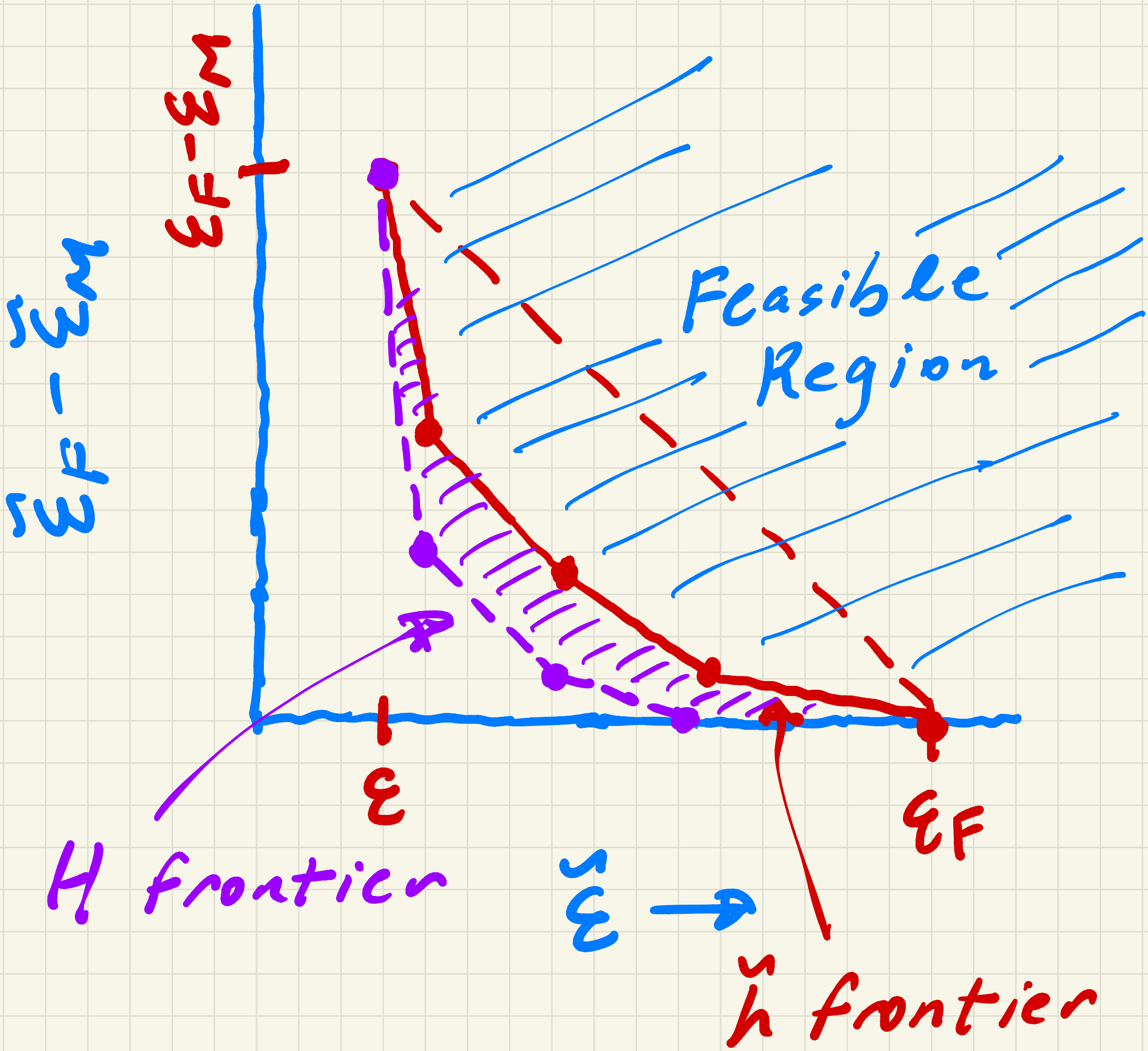
- Introduce randomized $\tilde{h} = \langle p, q, r, s \rangle$ "on top" of h
- Allows different tradeoffs between $\hat{\varepsilon}$ (error) and $\hat{\varepsilon}_F - \hat{\varepsilon}_M$ (unfairness)
- Optimal tradeoffs given by Pareto Frontier
- Fast algorithm
- Generalizes to FP/FN rates, more groups

Any drawbacks?

What More Could We Want?

- Suppose original h came from H
- $h \in H$ chosen to have small error
- But H might contain better error-unfairness tradeoffs than Pareto Frontier of $\tilde{h}$

The H Frontier



Challenges

- H may not be closed under mixtures
- Model space for $\tilde{h}$ was just $\langle p, q, r, s \rangle$, had fast Pareto algo
- In general, even finding low-error $h \in H$ is "hard" even absent fairness
- But ML does have good non-fair heuristics...

"Oracle"/Reduction Approach

- Assume we have a subroutine ("oracle") L for (approx.) minimizing $\varepsilon(h)$ in H (non-fair)
- Assume L handles reweightings of data
- Use repeated calls to L to find H-frontier

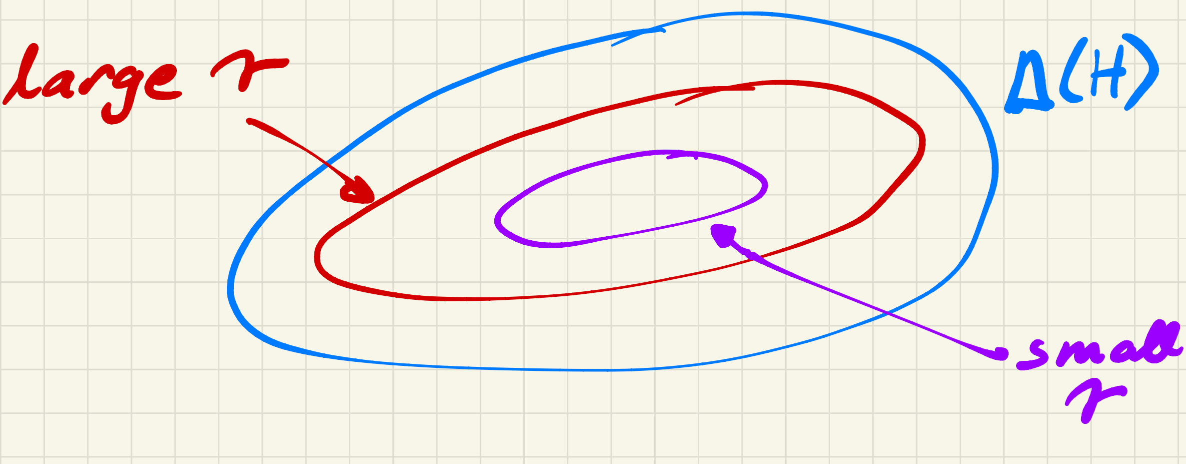
The Set-Up

- L outputs $h \in H$, but we will output mixtures or distributions $\Delta(H)$ (needs to be sparse)
- Want to solve:

$$\text{minimize } \varepsilon(h) \text{ for } h \in \Delta(H)$$

subject to:

$$|\varepsilon_F(h) - \varepsilon_M(h)| \leq \gamma$$



Game Theory Framework

- Algo structured as a
  two-player,
  zero-sum,
  repeated game
- Learner plays some
  $h \in \Delta(H)$ at each round
- Regulator plays a
  distribution on data:
  weight τ on F,
  $1-\tau$ on M

Payoffs

- Payoff to Learner:

  $-[\varepsilon(h) +$ (overall error)

  $\max(0, |\tau\varepsilon_F(h) - (1-\tau)\varepsilon_M(h)| - \gamma)]$

  (fairness violation)

- Payoff to Regulator:

  $+[\text{same}]$

- Important/deep fact:

  Nash equilibrium of game

  = solution ($\Delta(h)$) to

  optimization problem

Algorithm (Sketch)

- Reg. starts by letting τ = wt. of F in S
- Lrn. starts by calling L on $S \to h_1$
- For $t = 1, 2, \ldots T-1$:
  - Reg. updates wts. to improve payoff on $\frac{1}{t}(h_1 + h_2 + \cdots + h_t)$
  - Lrn. calls L on S & current wts $\to h_{t+1}$

Final output: $h = \frac{1}{T}(h_1 + h_2 + \cdots + h_T)$

"Theorem": If L outputs (near) optimal h_t at each round, then after T rounds, h is within $\sim 1/\sqrt{T}$ of opt. soln to:

minimize $\varepsilon(h)$ for $h \in \Delta(H)$
subject to
$$|\varepsilon_F(h) - \varepsilon_M(h)| \leq \gamma$$

- So we have reasonably general/practical algos for training **group fair** models

- What about **individual** fairness guarantees?

- <u>Q</u>: Why not let each x be its own "group"?

- <u>A</u>: E.g. to equalize errors, only rates are 0 (perfect), 1 (opposite) or 1/2 (random)

Better: Metric Fairness

- Posit a distance metric $d(x,x')$ between individuals
- Learn a real-valued model $h(x)$
- Fairness constraint: for every x, x':

$$|h(x)-h(x')| \leq \alpha\, d(x,x')$$

constant

- "Similar individuals treated similarly"

Challenges

- Where/from whom do we get $d(x,x')$?
- Pushes the problem elsewhere?
- $d(x,x')$ closed form?
- Thresholding $h(x)$ may lose fairness
- No practical algos*

Interpolating between group & individual fairness

Rich Subgroup Fairness

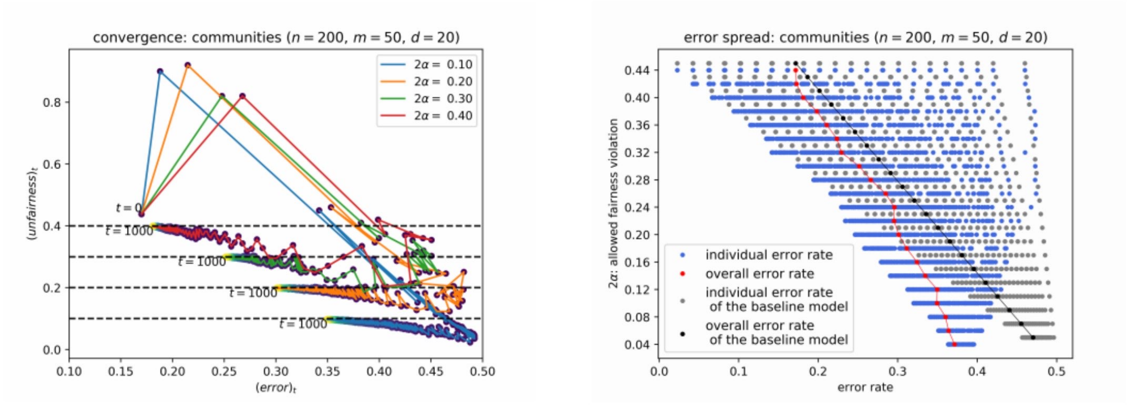
- Suppose we enforce group fairness by race, gender, age, disability, income...
- Might **still** discriminate disabled Hispanic women over 55 making ≤ $40K
- "Fairness gerrymandering"
- Let's enforce group fairness for **all combos*** of protected features

Game Theory II

- Learner still chooses (mixture) model h
- Regulator now finds combos on which h is unfair — itself a learning problem!
- Both players use heuristic/oracle L
- Same kind of theorem — get $\sim 1/\sqrt{T}$ of optimal, on *significant subgroups

Average Individual Fairness

- Suppose we make many decisions about x over time
- E.g. product recs, ads, ..
- Let $\varepsilon_x(h)$ be error rate across decisions for individual x
- Ask that all $\varepsilon_x(h)$ approx. equal
- Game Theory III: Regulator now picks x s.t. $\varepsilon_x(h)$ is largest



(a) convergence plot: communities data set

(b) error spread plot: communities data set

Figure 2: (a) Error-unfairness trajectory plots illustrating the convergence of algorithm **AIF-Learn**. (b) Error-unfairness tradeoffs and individual errors for **AIF-Learn** vs. simple mixtures of the error-optimal model and random classification. Gray dots are shifted upwards slightly to avoid occlusions.

(Subjective) Fairness Elicitation

- What if fairness is complex and subjective?
- Elicitation:
  - "x & x' should receive same outcome"
  - "x should receive at least as good an outcome as x'"
- Game Theory IV: Regulator enforces elicited constraints

5.2 Subjective Fairness Elicitation

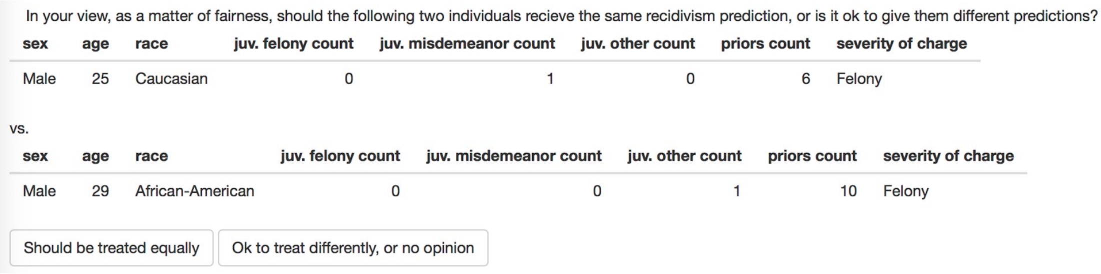
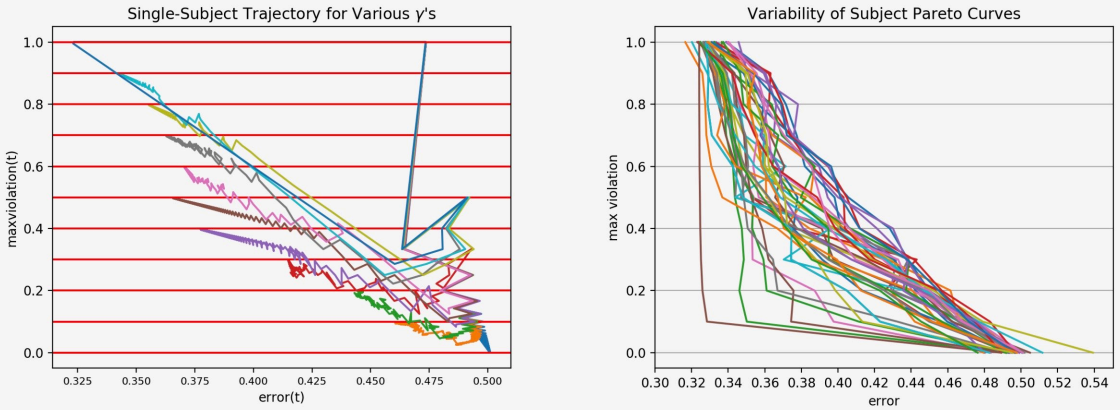


Figure 1: Screenshot of sample subjective fairness elicitation question posed to human subjects.


Single-Subject Trajectory for Various γ's
maxviolation(t)
error(t)
1.0
0.8
0.6
0.4
0.2
0.0
0.325
0.350
0.375
0.400
0.425
0.450
0.475
0.500
Variability of Subject Pareto Curves
max violation
error
1.0
0.8
0.6
0.4
0.2
0.0
0.30
0.32
0.34
0.36
0.38
0.40
0.42
0.44
0.46
0.48
0.50
0.52
0.54

(a) (b)

Other Strengthenings

- Minimax group fairness
  - Make worst group error as small as possible
  - If $\varepsilon_1 \geq \varepsilon_2 \geq \cdots \varepsilon_k$ minimax $\varepsilon_1' = \varepsilon_2' = \cdots = \varepsilon_k'$ equalized then $\varepsilon_1 \leq \varepsilon_1', \ldots, \varepsilon_k \leq \varepsilon_k'$
- Lexicographic group fairness

Fair (Supervised) ML: Recap

- Group fairness notions
- Bolt-on/post-processing
- Pareto frontiers/tradeoffs
- In-processing: reduction to non-fair; game theory
- Individual fairness
- Group $\longleftrightarrow$ Individual

Other Settings

- Fair RL/control
- Fair representations
- Causal approaches
- Fair clustering
- Fair rankings
- Fair labelings

⋮

The Real World

- Open-ended AI services
- Unknown downstream uses
- Data curation/collection
- External/"activist" audits

⋮