Fairness

Christos Dimitrakakis

November 8, 2024

Fairness

What is it?

Fairness

What is it?

- Meritocracy.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.
- Equal treatment.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.
- Equal treatment.
- Non-discrimination.

Meritocracy

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Meritocracy

Example 1 (College admissions)

- Student A has a grade 4/5 from Gota Highschool.
- Student B has a grade 5/5 from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with 4+ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?
- Admit randomly?

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

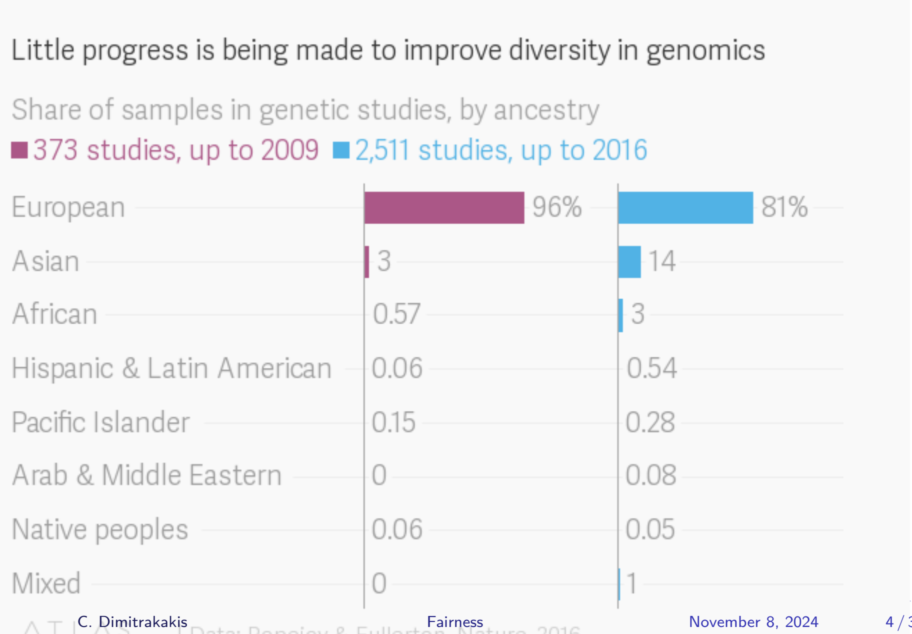
- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?
- Admit randomly?
- Use prediction of individual academic performance?

Proportional representation

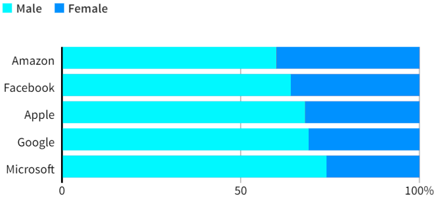


Hiring decisions

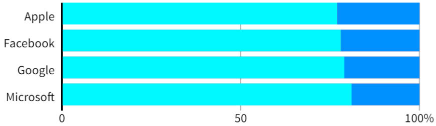
Dominated by men

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

GLOBAL HEADCOUNT



EMPLOYEES IN TECHNICAL ROLES



Fairness and information

Example 3 (College admissions data)

School	Male	Female
A	62%	82%
B	63%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%
Average	*45%*	*38%*

Fairness

What is it?

Fairness

What is it?

- Meritocracy.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.
- Equal treatment.

Fairness

What is it?

- Meritocracy.
- Proportionality and representation.
- Equal treatment.
- Non-discrimination.

Meritocracy

Meritocracy

Example 4 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Meritocracy

Example 4 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 5 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

Meritocracy

Example 4 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 5 (Additional information)

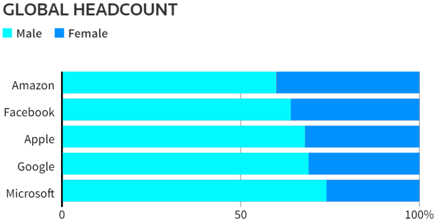
- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

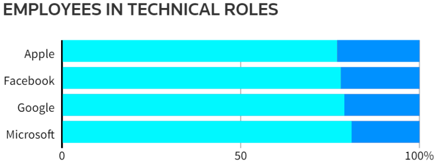
We still don't know how a specific student will do!

Hiring decisions

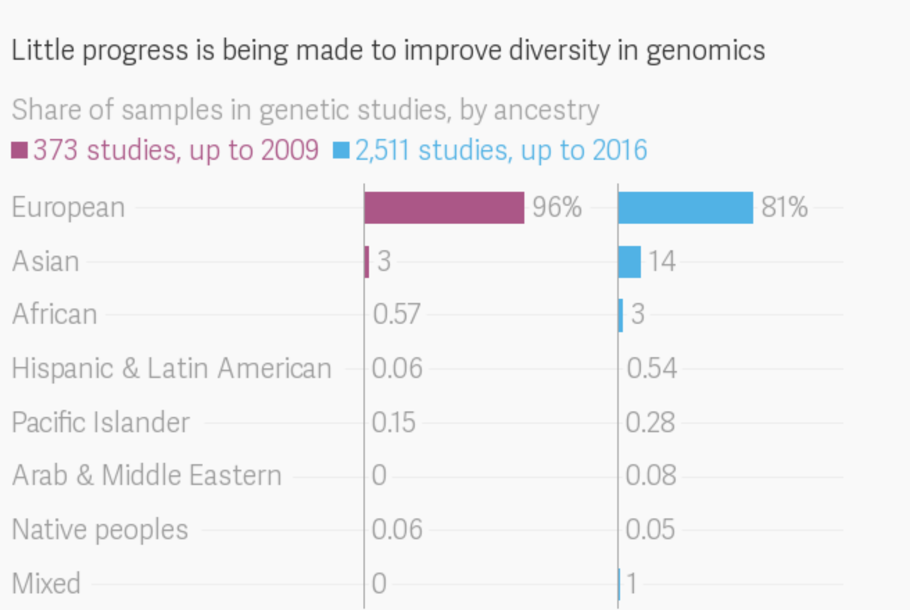
Dominated by men

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.





Group fairness and proportionality



Solutions

- Admit everybody?

Solutions

- Admit everybody?
- Admit randomly?

Solutions

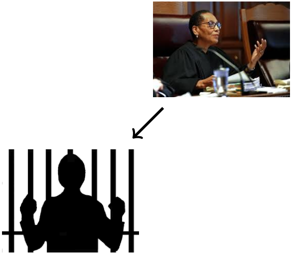
- Admit everybody?
- Admit randomly?
- Use prediction of individual academic performance?

Solutions

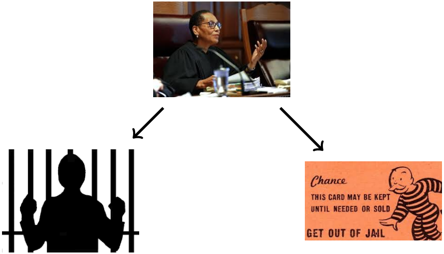
- Admit everybody?
- Admit randomly?
- Use prediction of individual academic performance?
- Should we take into account group membership or other population information?

Bail decisions

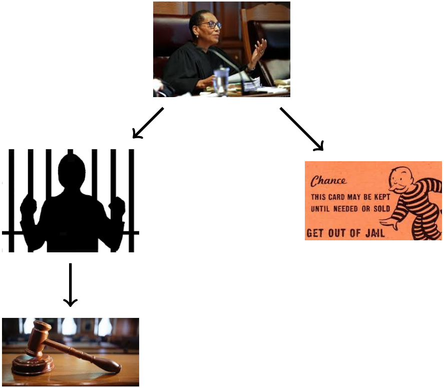
Bail decisions



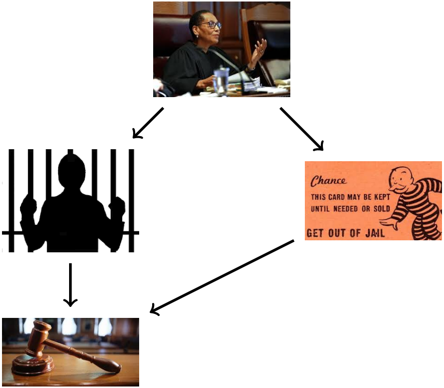
Bail decisions



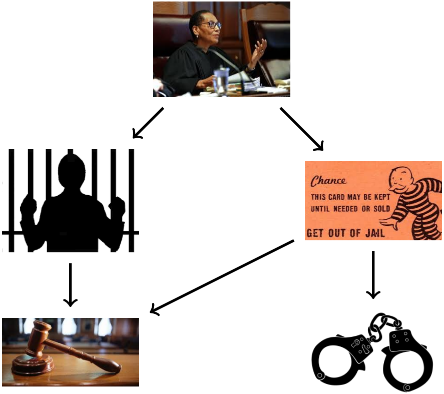
Bail decisions



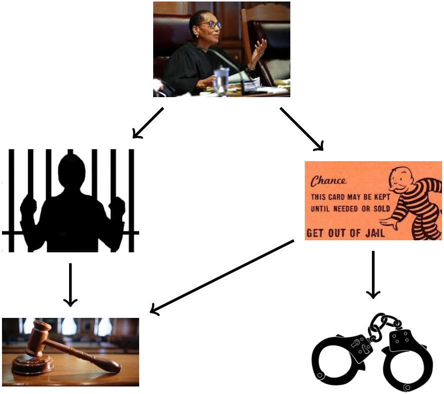
Bail decisions

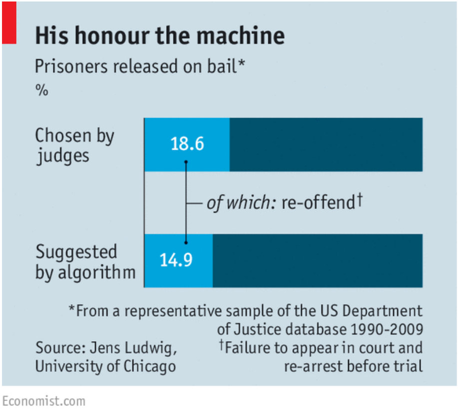


Bail decisions



Bail decisions





Demographic parity and equality of opportunity.

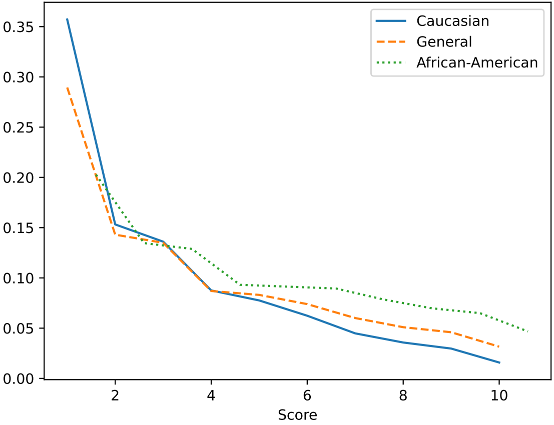


Figure: Apparent bias in risk scores towards black versus white defendants.

$$\mathbb{P}^{\pi}_{\theta}(a_t | z_t) = \mathbb{P}^{\pi}_{\theta}(a_t). \tag{3.1}$$

Calibration.

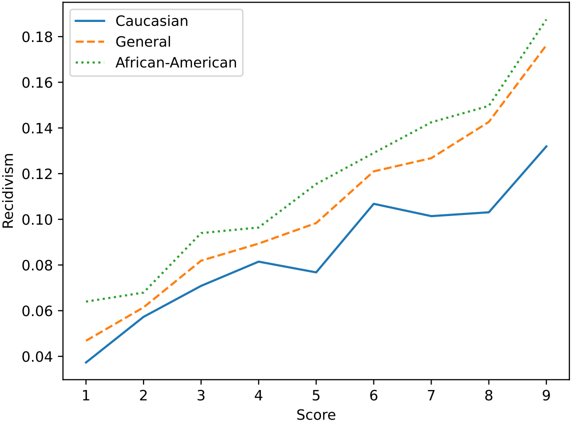
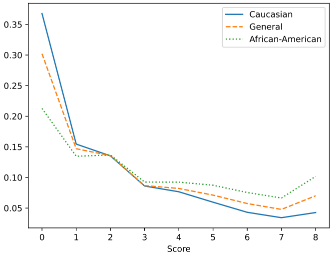


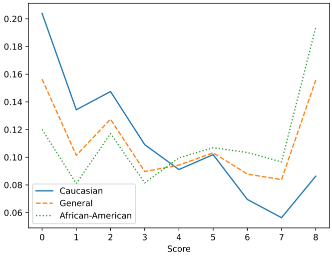
Figure: Recidivism rates by risk score.

$$\mathbb{P}^{\pi}_{\theta}(y_t | a_t, z_t) = \mathbb{P}^{\pi}_{\theta}(y_t | a_t). \tag{3.2}$$

Balance



(a) No recidivism



(b) Recidivism

Figure: Score breakdown based on recidivism rates.

$$\mathbb{P}_\theta^\pi(a_t | y_t, z_t) = \mathbb{P}_\theta^\pi(a_t | y_t). \tag{3.3}$$

Graphical models and independence

- Why is it not possible to be fair in all respects?

Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.

Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.

Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.

Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.

- θ: environment parameters (latent variable)

Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.

- θ: environment parameters (latent variable)
- π: policy of the decision maker (decision variable)

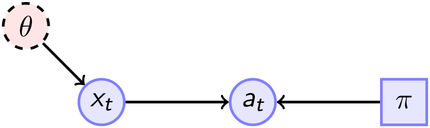
Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.

- θ: environment parameters (latent variable)
- π: policy of the decision maker (decision variable)
- x_t: observation

Graphical models and independence

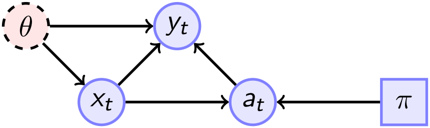
- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.



- θ: environment parameters (latent variable)
- π: policy of the decision maker (decision variable)
- x_t: observation
- a_t: action taken

Graphical models and independence

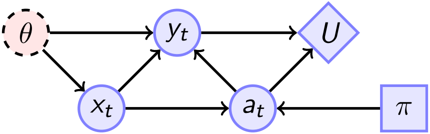
- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.



- θ: environment parameters (latent variable)
- π: policy of the decision maker (decision variable)
- x_t: observation
- a_t: action taken
- y_t: outcome

Graphical models and independence

- Why is it not possible to be fair in all respects?
- Different notions of conditional independence.
- Can only be satisfied rarely simultaneously.



- θ: environment parameters (latent variable)
- π: policy of the decision maker (decision variable)
- x_t: observation
- a_t: action taken
- y_t: outcome
- U: utility

Example 6 (Classification)

- y: labels
- a: label prediction
- $U(a, y) = \mathbb{I}\{a = y\}$.
- Here the outcome is independent of the action:

$$\mathbb{P}_\theta^\pi(y_t \mid a_t, x_t, z_t) = \mathbb{P}_\theta^\pi(y_t \mid x_t, z_t).$$

Example 7 (Regression)

In this setting we can also relax our framework to work with expectations instead of probabilities. For example, calibration and balance can be written as the conditions:

$$\mathbb{E}^{\pi}_{\theta}(y_t | a_t, z_t) = \mathbb{E}^{\pi}_{\theta}(y_t | a_t), \qquad \mathbb{E}^{\pi}_{\theta}(a_t | y_t, z_t) = \mathbb{E}^{\pi}_{\theta}(a_t | y_t),$$

respectively.

Measuring and optimising fairness and utility.

Expected utility.

Let us write out expected utility, with $\mathbb{P}^\pi_\theta(\cdot) \equiv \mathbb{P}(\cdot \mid \theta, \pi)$

$$\mathbb{E}^\pi_\theta(U) = \sum_{x,z,y} \mathbb{P}^\pi_\theta(y, a, x, z) U(a, y).$$

The y_t, a_t, x_t, z_t is already sampled from $\mathbb{P}^\pi_\theta$, so we can approximate the expected utility of the historical policy with

$$\widehat{E}_n(U) = \sum_{t=1}^n U(a_t, y_t), \qquad a_t, y_t \sim \mathbb{P}^\pi_\theta .$$

Model specification

For simplicity, assume $z = x_1$, i.e. one of the features. Then

$$\mathbb{P}^\pi_\theta(y, a, x, z) = \underbrace{\mathbb{P}_\theta(y \mid a, x)}_{\text{outcome distribution}} \underbrace{\pi(a \mid x)}_{\text{policy}} \underbrace{\mathbb{P}_\theta(x)}_{\text{feature distribution}}$$

Using the model, we can estimate the expected utility of any policy.

Deviation from balance.

$$\mathbb{P}_\theta^\pi(a|y,z), \qquad \mathbb{P}_\theta^\pi(a|y).$$

for all values of y, z. Let us first look at the total variation distance:

$$\| \mathbb{P}_\theta^\pi(a|y,z) - \mathbb{P}_\theta^\pi(a|y)\|_1 = \sum_a \| \mathbb{P}_\theta^\pi(a|y,z) - \mathbb{P}_\theta^\pi(a|y)\|_1.$$

$$\hat{P}_n(a|y,z) = \sum_t \frac{\mathbb{I}\left\{a_t = a \wedge y_t = y \wedge z_t = i\right\}}{\mathbb{I}\left\{y_t = y \wedge z_t = i\right\}}$$

We can then plug those into our original measure:

$$\| \mathbb{P}_\theta^\pi(a|y,z) - \mathbb{P}_\theta^\pi(a|y)\|_1 \approx \sum_a \|\hat{P}_n(a|y,z) - \hat{P}_n(a|y)\|_1.$$

Formalisation of the problem

Unconstrained optimisation.

Let $\lambda \in [0, 1]$:

$$\max_{\pi} (1 - \lambda) U(\pi, \theta) - \lambda F(\pi, \theta)$$

Constrained fairness.

Let $\epsilon \geq 0$:

$$\max_{\pi \in \Pi} U(\pi, \theta)$$
$$\text{s.t.} F(\pi, \theta) \leq \epsilon.$$

Constrained utility.

Let $\epsilon \geq 0$:

$$\max_{\pi \in \Pi} F(\pi, \theta)$$
$$\text{s.t.} U(\pi, \theta) \geq \max_{\pi'} U(\pi', \theta) - \epsilon$$

Individual fairness

Main variables for individual fairness.

- $x_i \in \mathbb{R}^d$: Individual features.
- $w_i \in \mathbb{R}$: Individual worth.
- $u_i : \mathcal{A} \to \mathbb{R}$: Individual utility.
- $a \in \mathcal{A}$: Action taken by the decision maker.
- π: The decision maker's policy for making decisions.

Meritocracy for given utilities and worths.

Definition 8

A **decision** a is **fair** if, for all i, j, we have $w_i > w_j \Rightarrow u_i(a) \geq u_j(a)$.

Example 9

w		$a = 0$	$a = 1$
1	u_1	0	1
0	u_2	1	0

Definition 10

A **policy** π is **fair** if, for all i, j it holds that: $\mathbb{E}_\pi[u_i|w] > \mathbb{E}[u_j|w] \Rightarrow w_i > w_j$.

Example 11 (Ranking and top-k cohort selection.)

- n individuals
- i-th individual receiving an evaluation w_i.
- $a_i = 1$ if we select an individual, and 0 otherwise.
- Ranked selection algorithm π: first rank the individuals so that $w_i > w_{i+1}$. We then can assign $a_i = 1$ for all $i \leq k$.

Example 12

Matching problems.

- $u_1(x) > u_1(y) > u_1(z)$,
- $u_2(z) > u_2(y) > u_2(x)$
- $u_3(z) > u_3(x) > u_3(y)$

and $w_1 > w_2 > w_3$. A stable matching is then $a_1 = x, a_2 = z, a_3 = y$.

Treating similar individuals similarly.

- $d(x_i, x_j)$: Distance between individuals i, j.
- $D[\pi(a_i|x_i), \pi(a_j|x_j)]$: Distance between policy decisions.
- Utility function $U(\pi)$.

Our goal is to find a decision rule π maximising $U(\pi)$, so that it is Lipschitz-smooth with respect to D, d:

$$D[\pi(a_i|x_i), \pi(a_j|x_j)] \leq d(x_i, x_j)$$

.

Connection to meritocracy.

Consider a decision maker that is choosing applicants to hire $(a = 1)$. For any two applicants with respective features x_i, x_j, we need to assume that they both prefer being hired to not being hired. On the other hand, the DM prefers to hire applicants that result in higher utility. In particular, hiring applicant x results in utility $u(x)$. Let us now define the utility of a policy, when applied to a single applicant, as

$$\mathbb{E}^\pi(U|x) = \pi(a = 1|x)u(x).$$

Trivially, our policy should be such that we tend to hire people with higher utility to us. If the utility function is L-Lipschitz, then $|u(x) - u(x')| \leq Ld(x, x')$. Now consider the case where x' can get an unfair advantage through nepotism. However, if our policy is smooth, it requires $\pi(a = 1|x) \geq \pi(a = 1|x') - d(x, x') \geq \pi(a = 1|x') - |u(x) - u(x')|/L$. In addition, since our policy must be a maximising policy, if $u(x) > u(x')$, then $\pi(a = 1|x) \geq \pi(a = 1|x')$.

If we need to fix a policy before seeing the population, then we must use a parametrised policy that can give us a probability distribution over a for any possible $\boldsymbol{x} \in \mathcal{X}$ where $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$. A simple parametrisation is a linear-softmax network,[1] with $\boldsymbol{\theta}_a \in \mathbb{R}^n$ for all actions a, given by:

$$\pi(a \mid \boldsymbol{x}) = \frac{e^{\boldsymbol{\theta}_a^\top x}}{\sum_{a' \in \mathcal{A}} e^{\boldsymbol{\theta}_{a'}^\top \boldsymbol{x}}}.$$

This is not really difficult to work with, but we have a potential problem when considering the smoothness constraints. If we let $d(\boldsymbol{x}, \boldsymbol{x}') = \|bx - \boldsymbol{x}'\|$ then the Lipschitz condition is in fact a condition on the gradient:

$$\nabla_{\boldsymbol{x}} \pi(a \mid \boldsymbol{x}) \leq 1, \qquad \forall \boldsymbol{x} \in \mathcal{X}$$

In particular, we can no longer make it so

Smoothness and parity.

Can group fairness be satisfied in this setting? To define groups, we select $S, T \subset \mathcal{X}$. We can say a policy satisfies approximate parity between the groups if:

Definition 13

ϵ-parity

$$\|\pi(a|x \in S) - \pi(a|x \in T)\|_1 \leq \epsilon.$$

We can also similarly define the worst-case bias for some action:

Definition 14

Bias for some action a:

$$\max_\pi \pi(a|x \in S) - \pi(a|x \in T) \qquad \text{s.t. } \pi \text{ is Lipschitz}$$