Fair Classification

$$\max_{M} \; \mathop{\mathbb{E}}_{x\sim V}\left[\mathop{\mathbb{E}}_{o\sim M_x}\left[u(x,o)\right]\right] \leftarrow \text{Maximize utility}$$

$$\text{s.t.} \quad \|M_x - M_y\| \le d(x,y) \leftarrow \text{s.t. fairness constraint}$$

Classifier $M = [M_{x_1}, \ldots, M_{x_i}, \ldots, M_{x_n}]$

Input $i \rightarrow x_i$

outputs = $|A|$

Notation

$M_x(o) = \Pr[M_x = o]$ ← scalar

$u(x,o)$ = scalar utility function (given to us)

$p(x)$ = Probability of drawing $x \in V$ (user/input vector)

(1) Show that objective is linear:

$$\mathop{\mathbb{E}}_{x\sim V}\left[\mathop{\mathbb{E}}_{o\sim M_x}\left[u(x,o)\right]\right] = \sum_{x\in V} p(x)\left(\sum_{o\in A} M_x(o)\cdot u(x,o)\right)$$

$$= \sum_{x\in V}\sum_{o\in A} M_x(o)\cdot \underbrace{p(x)\cdot u(x,o)}$$

Do not depend on M!
$\Rightarrow$ objective is linear in M

(2) Show that constraint is linear:

$$\|M_x - M_y\| \le d(x,y) \Rightarrow \tfrac{1}{2}\sum_{o\in A} |M_x(o) - M_y(o)| \le d(x,y)$$

(scalar)

We can write this as follows, with vector $q \in \{-1,1\}^{|A|}$:

$$\tfrac{1}{2}\sum_{o\in A} q(o)\left(M_x(o) - M_y(o)\right) \le d(x,y), \quad \forall\, x,y,q\in\{-1,1\}^{|A|}$$

{ Each of these is linear in M
... but now we have $2^{|A|}$ inequalities for each initial one!

$\Rightarrow$ The fair classification optimization is a linear program.