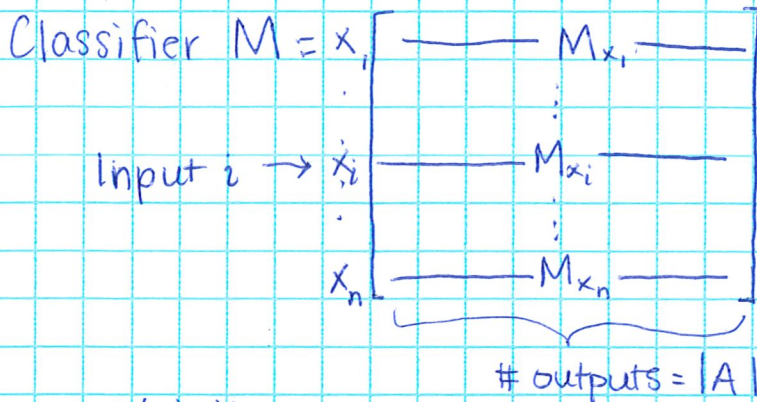


# Fair Classification

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

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



Notation

$$M_x(o) = \Pr[M_x = o] \leftarrow \text{scalar}$$

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

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

① Show that objective is linear:

$$\begin{aligned} \mathbb{E}_{x \sim V} \left[ \mathbb{E}_{o \sim M_x} [U(x, o)] \right] &= \sum_{x \in V} p(x) \left( \sum_{o \in A} M_x(o) \cdot U(x, o) \right) \\ &= \sum_{x \in V} \sum_{o \in A} M_x(o) \cdot \underbrace{p(x) \cdot U(x, o)} \end{aligned}$$

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

② Show that constraint is linear:

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

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

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

↑ 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.