$\langle x, y \rangle \rightarrow$ Training Alg. A $\rightarrow$ Model is Fair based on Some Def

$\uparrow$

Change A

Min $\quad loss(h_\theta(x), y)$

s.t.

$\boxed{\text{Unfairness} \leq \varepsilon}$, for Some Notions of fairness

is not Convex

how to make it Convex / more efficient

[Zafar , et al.]

The overall idea : Change the notion of fairness to make the opt. Convex

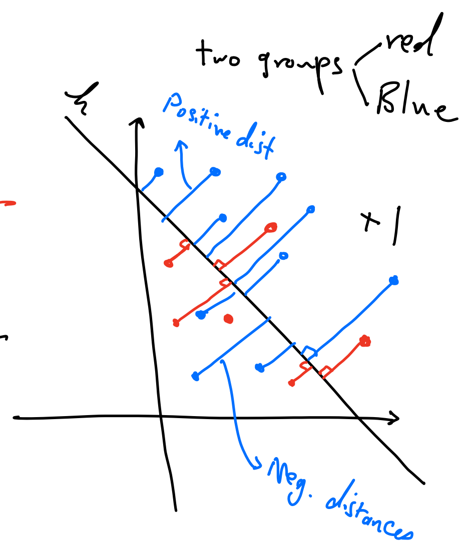
Binary Classification

Binary Sensitive Attribute

Fairness Notion: Demographic Parity

red-ratio = $\frac{2}{6}$

Blue-ratio = $\frac{7}{10}$



Fairness $\underbrace{:}_{\text{based on}}$ The distance of points from the decision Boundary

Independance : $S \perp\!\!\!\perp h_\theta(x)$

$$Cov(S, h_\theta) = 0$$

$Cov(S, d_\theta(x))$ is bounded

↳ distance to the decision boundary

$$|Cov(S, d_\theta(x))| \leq \tau$$

$$Cov(S, d_\theta(x)) = E[(S-\bar{S})\, d_\theta(x)] - \cancelto{0}{E[S-\bar{S}]}\, E[d_\theta(x)]$$

$$= E[(S-\bar{S})\, d_\theta(x)]$$

$$\approx \frac{1}{n}\sum_{i=1}^{n}(S_i-\bar{S})\, d_\theta(x_i) \;?$$

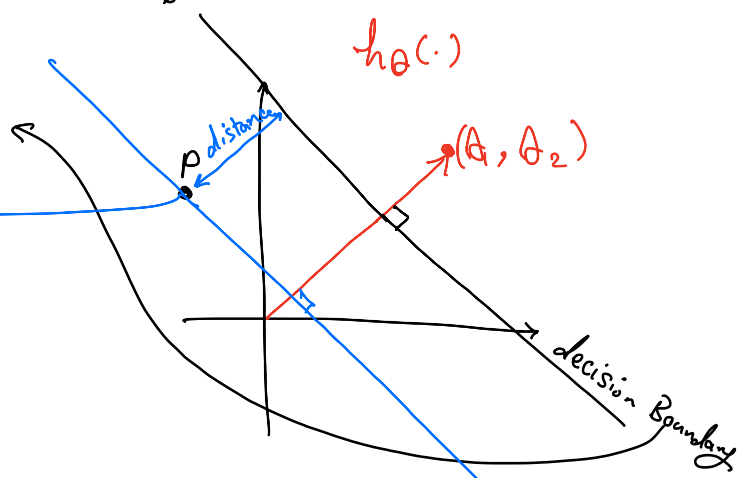
for Linear classifiers

$$h_\theta(x) = \theta^T X = \sum_{j=0}^{m} \theta_j X_{ij}$$

$\theta_1 x_1 + \theta_2 x_2 = -\theta_0$

$\Rightarrow \theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$

$\Rightarrow \theta^T X = 0$



$\theta^T X^p = d_\theta(P)$

$$\Rightarrow \mathrm{Cov}(s, d_\theta(x)) \simeq \frac{1}{n} \sum_{i=1}^{n} (s_i - \bar{s})\, \theta^T x_i$$

$$|\mathrm{Cov}(s, d_\theta(x))| \leq \tau$$

$$\Rightarrow \frac{1}{n} \sum (s_i - \bar{s})\, \theta^T x_i \leq \tau$$

and

$$F \quad \frac{1}{n} \sum (\bar{s} - s_i)\, \theta^T x_i \leq \tau$$

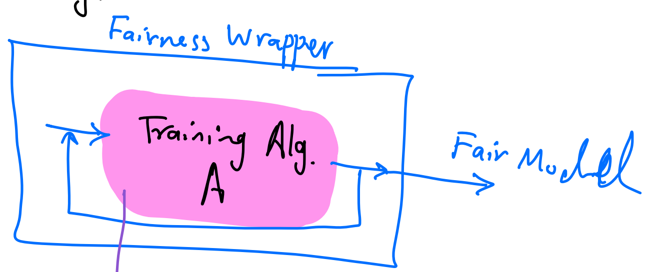
$$\min_\theta \ \mathrm{loss}(h_\theta(x), y)$$

s.t.

$F \leftarrow$ Convex.

[Alekh Agarwal et al.] (A reduction approach to fair classification)

→ They create A Fairness Wrapper on top of Black Box Training Alg.



→ Should work on Several Fairness notions

- " " " Non-Binary Sensitive Attributes
- Not Limited to Binary classification

requirment → The Training Alg. Should accept weighted Samples.

Step 1: Propose a unified Theme for fairness definitions.

✓ group j & definition F,

$$\mu_j(\theta) = E\left[\underbrace{g(X, Y, S, f_\theta)}_{\text{Performance}} \,\middle|\, \mathcal{E}(X, f, S)\right]$$

don't support Sufficiency $S \perp\!\!\!\perp Y \mid f$

e.g., DP: $g_j = f_\theta(X)$, $\mathcal{E}_j = \{S = S_j\}$

$$\mu_j(\theta) = E\left[f_\theta(x) \mid S = S_j\right]$$

$$\mu_j^*(\theta) = E\left[f_\theta(x)\right]$$

$$\mu_j(\theta) - \mu_j^*(\theta) \leq \tau$$

for a set of groups & def.:

$$\mu_j(\theta) - \mu_j^*(\theta) \leq \tau$$

$$\cdots$$

$$\mu_k'(\theta) - \mu_k'^*(\theta) \leq \tau$$

$$\longrightarrow \underset{\text{Matrix}}{M}\,\mu(\theta) \leq \underset{\text{Vector}}{\tau}$$

$$\Rightarrow M\mu - \tau \leq 0$$

Step 2: Optimization Reformulation

Min $\ell(\theta)$

S.t. $M\mu - \tau \leq 0$

Optimize

$$L(\theta, \lambda) = \ell(\theta) - \lambda(M\mu - \tau)$$

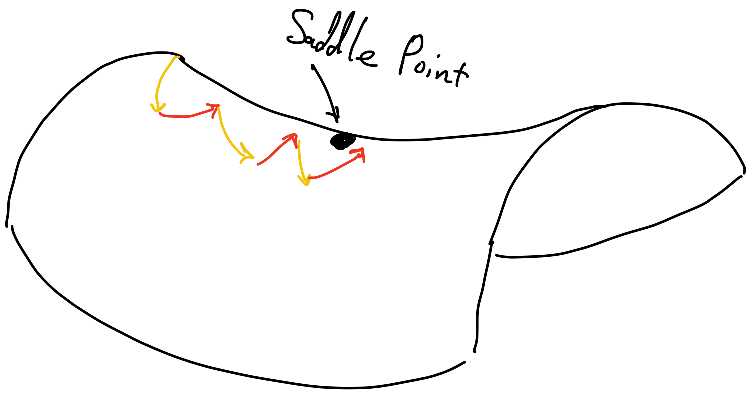
Review: Lagrangian Multipliers
= Dual Form

Max f_θ
s.t.
$ax \leq b$ $\Rightarrow$ $L(\theta, \lambda) = f_\theta - \lambda(ax - b)$

optimize for L

Max $f_\theta - \lambda(ax - b)$

max, min, variables



Opt $L(\theta, \lambda)$

$$\max_{\lambda} \min_{\theta} \; \ell_{\theta}(x) - \lambda (M_{\mu} - \tau)$$

Reweighting

Existing Training Alg.

The # iterations is bounded by $O(n^2)$

Price (Aka Cost) of Fairness

The diff. b/w the opt. value of unfair vs. fair solutions.

e.g.,

(A) $\min L(\theta) = L_A$

(B) $\min L(\theta) = L_B$ s.t. Fairness

$L_B \geq L_A \Rightarrow PoF = L_B - L_A$

Price of Fairness

Pareto – Optimality of Fairness

Suppose there are groups $\{g_1 \cdots g_K\}$.
the Performance of the Alg/model for g_i is Per_i.
If Per_i^U is the Per. of unFair Alg. for g_i
and Per_i^F " " " " Fair " " "

If $\forall g_i \; ; \quad Per_i^U \geq Per_i^F$

and

$\exists g_j : \quad Per_j^U > Per_j^F$

$\Rightarrow$ the Fair Solution is **DOMINATED**
By the unfair one.

Pareto_Optimals is the Set of Solutions that are Not Pareto_dominated

→ AKA skyline