Rényi Fair Inference

Sina Baharlouei

Daniel J. Epstein Department of Industrial Engineering

Maher Noueihed

USC → *AUB*

Ahmad Beirami

Meisam Razaviyayn

USC

Machine Learning Algorithms and Fairness

- Machine learning algorithms for automatic decision making
- Not Necessarily Fair!

Propublica Analysis of COMPAS

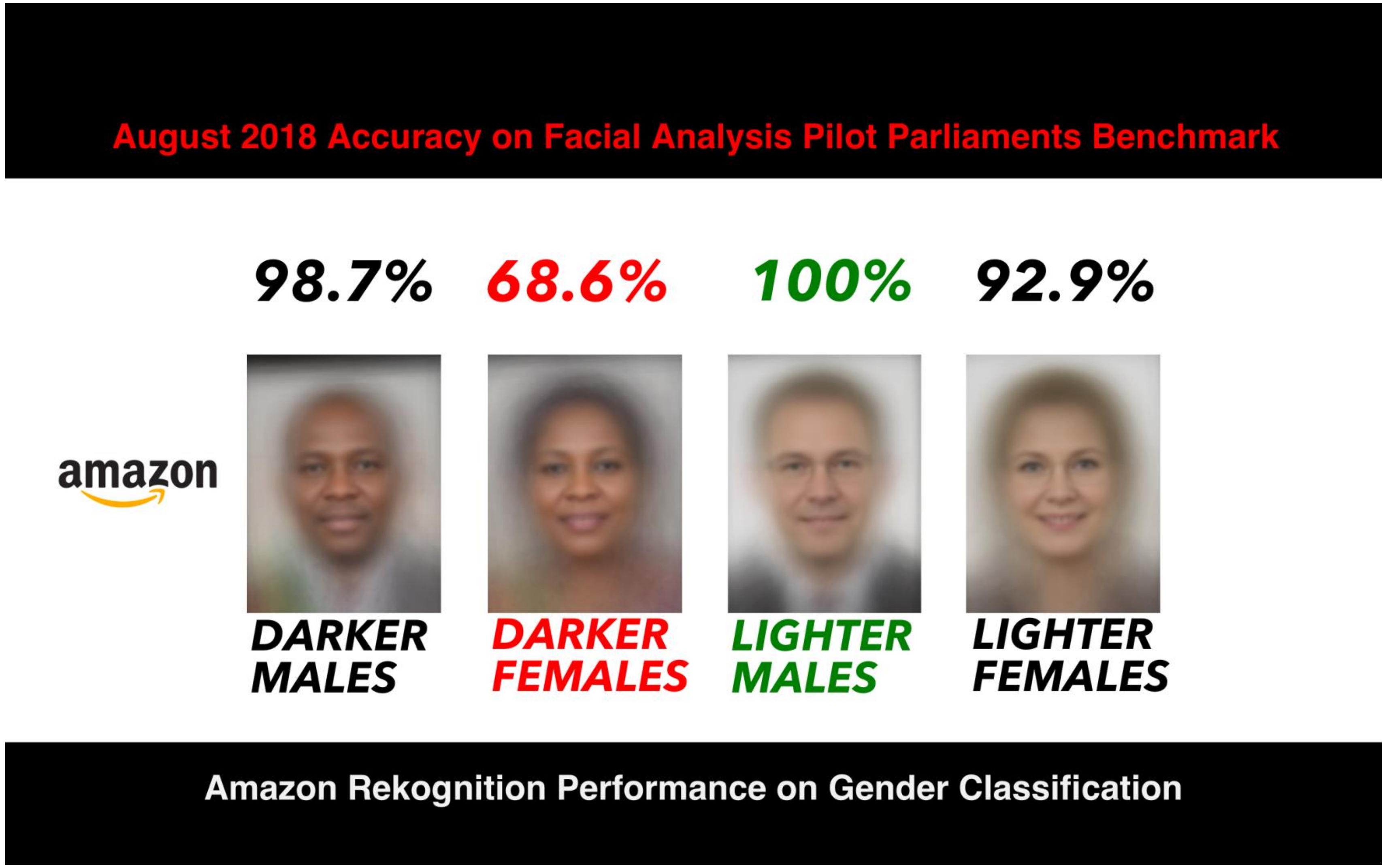
	White	Black
Mislabeled as High-Risk	23.5%	44.9%
Mislabeled as Low-Risk	47.7%	28%

Amazon Recruiting Machine (Reuters, 2018)

- "Woman" Keyword in CV
- Two all-women colleges

Source of Bias Against Protected Groups

- Toxic Historical Data: Machine learning models reflect the toxic training data
- Data Limitation: Low number of samples from protected groups
- Proxies: Features that are highly correlated with sensitive attributes
  - Elimination of sensitive attributes is not enough!



Problem Setup

- Population Risk:

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y)]$$

- Empirical Risk Minimization (ERM):

Feature vector of data i

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}_i), y_i)$$

Predicted label of data i

$$\hat{y} = \mathcal{F}(\boldsymbol{\theta}, \mathbf{x})$$

Actual label of data i

$$\mathbf{x} = (\mathbf{x}^{'}, \mathbf{s})$$

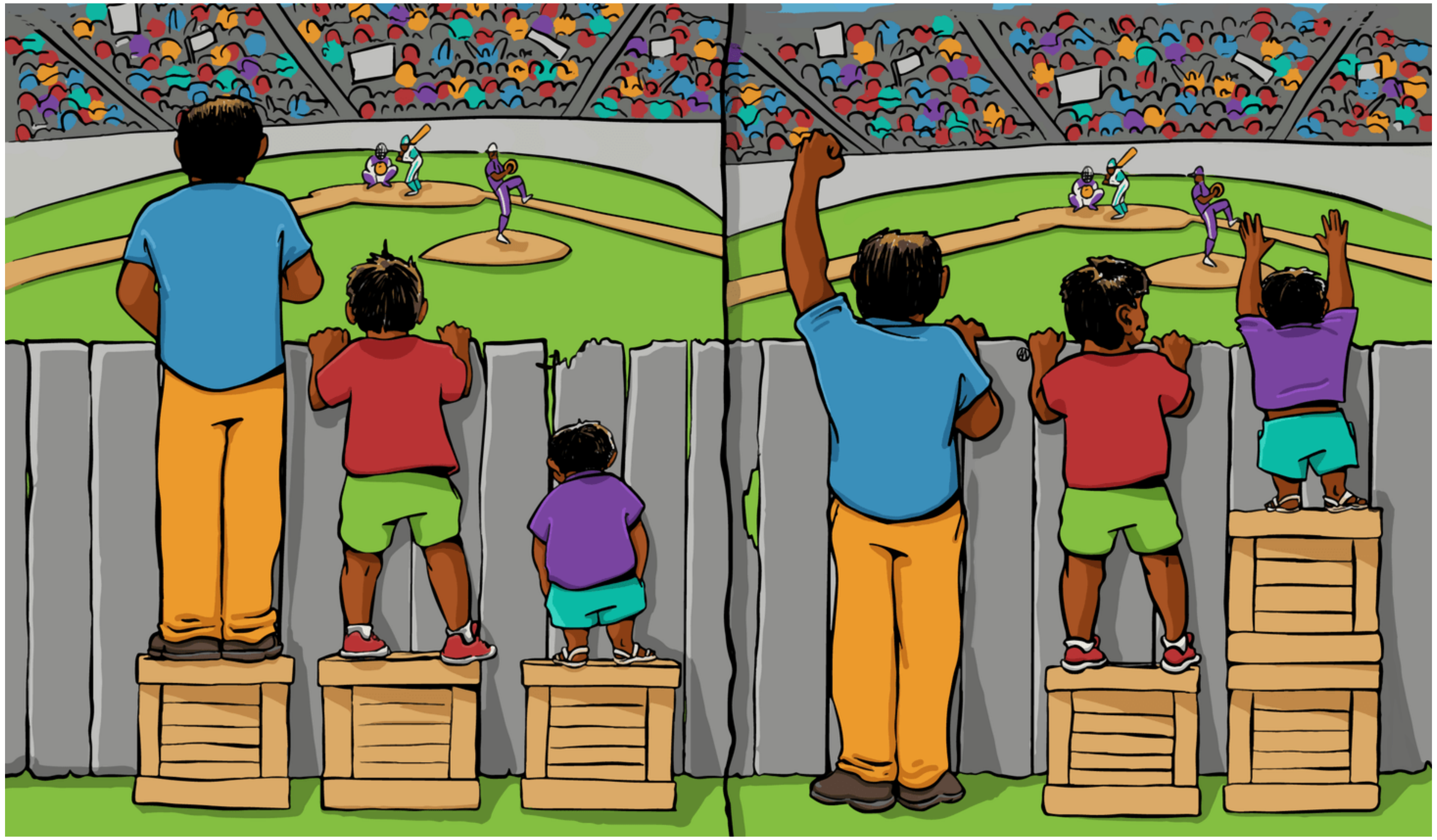
- Minimizing empirical risk

Sensitive attribute(s)

- Satisfying a notion of fairness

Non-sensitive features

Notions of Fairness



Equality of Opportunities

$$s \perp \hat{y} \,|\, y$$

Demographic Parity

$$s \perp \hat{y}$$

- What is the problem with the above picture?

Civil rights act of 1964, title vii, equal employment opportunities. 1964

Hardt, et al. "Equality of opportunity in supervised learning." *NIPS*. 2016.

Fair Empirical Risk Minimization

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y)]$$
$$\text{s.t.} \quad \hat{y} \perp S$$

- Specialization to Fair Logistic Regression:

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{n}\sum_{i=1}^{n} -y_i \log(\boldsymbol{\theta}^T \mathbf{x}_i) + (1 - y_i)\log(1 - \boldsymbol{\theta}^T \mathbf{x}_i)$$
$$\text{s.t.} \quad \hat{y} \perp S$$

- How to handle the constraint in the above problem?

Independence Measures for Fairness

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y)] + \lambda \rho(\hat{y}, \mathbf{s})$$

Pearson Correlation

$$\rho_P(s, \hat{y}) = \frac{\mathrm{Cov}(s, \hat{y})}{\sqrt{\mathrm{Var}(s)\mathrm{Var}(\hat{y})}}$$

✓ Easy to optimize!

✗ Limited to linear correlation

Mutual Information

$$\rho_I(s, \hat{y}) = \sum_{\hat{y} \in \mathcal{Y}} \sum_{s \in \mathcal{S}} p(s, \hat{y}) \log \left(\frac{p(s, \hat{y})}{p_S(s) p_{\hat{Y}}(\hat{y})} \right)$$

✗ Highly non-convex

✓ Capture any correlation!

Zafar, et al. "Fairness constraints: Mechanisms for fair classification." (2015).

Pérez-Suay, Adrián, et al. "Fair kernel learning." (2017).

Kamishima, et al. "Fairness-aware learning through regularization approach." *(2011).*

Song, Jiaming, et al. "Learning controllable fair representations." (2018).

Rényi Correlation

$$\rho_R(s, \hat{y}) = \sup_{f,g} \mathbb{E}[f(s)g(\hat{y})]$$

$$\text{s.t.} \quad \mathbb{E}[f(s)] = \mathbb{E}[g(\hat{y})] = 0$$

$$\mathbb{E}[f^2(s)] = \mathbb{E}[g^2(\hat{y})] = 1$$

✓ Normalized between zero and one.

✓ Zero iff two random variables independent.

✓ Tractable for discrete sensitive attributes.

Hirschfeld, Hermann O. "A connection between correlation and contingency." (1935).

Gebelein, Hans. "Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung." (1941).

Rényi, Alfréd. "On measures of dependence." (1959).

Properties of Rényi Correlation

$$\rho_R(s,\hat{y}) = \sup_{f,g} \mathbb{E}[f(s)g(\hat{y})]$$

- $0 \le \rho_R(s,\hat{y}) \le 1$
- $\rho_R(s,\hat{y}) = 0$ if and only if two random variables are independent.
- $\rho_R(s,\hat{y}) = 1$ iff there exist functions f and g s.t $f(s) = g(\hat{y}) \quad a.s$
- $\rho_R(s,\hat{y}) = \rho_R(f(s), g(\hat{y}))$ for bijective functions f and g.
- If two random variables are jointly Gaussian, then $\rho_R(s,\hat{y}) = |\rho(s,\hat{y})|$

Hirschfeld, Hermann O. "A connection between correlation and contingency." (1935).

Gebelein, Hans. "Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung." (1941).

Rényi, Alfréd. "On measures of dependence." (1959).

General Discrete Sensitive Attribute:

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y)] + \lambda \rho_R^2(\hat{y}, \mathbf{s})$$

Theorem (Witsenhausen (1975)). *Let $s \in \{s_1, \ldots, s_c\}$ and $\hat{y} \in \{\hat{y}_1, \ldots, \hat{y}_d\}$ be two discrete random variables. Then the Rényi coefficient $\rho_R(s, \hat{y})$ is equal to the second largest singular value of the matrix $\mathbf{Q} = [q_{ij}]_{i,j} \in \mathbb{R}^{c \times d}$, where $q_{ij} = \frac{\mathbb{P}(s=s_i, \hat{y}=\hat{y}_j)}{\sqrt{\mathbb{P}(s=s_i)\mathbb{P}(\hat{y}=\hat{y}_j)}}$.*

$$\min_{\boldsymbol{\theta}} \quad \max_{\mathbf{v} \perp \mathbf{v}_1, \|\mathbf{v}\| \leq 1} \quad \left(f_D(\boldsymbol{\theta}, \mathbf{v}) \triangleq \mathbb{E}[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y)] + \lambda \mathbf{v}^T Q_{\boldsymbol{\theta}}^T Q_{\boldsymbol{\theta}} \mathbf{v} \right)$$

- Reminder: $\lambda_{max} = \max_{\|v\|=1} \quad v^T A v$

General Discrete Sensitive Attribute:

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{v}\perp\mathbf{v}_1, \|\mathbf{v}\|\leq 1} \left(f_D(\boldsymbol{\theta}, \mathbf{v}) \triangleq \mathbb{E}[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y)] + \lambda \mathbf{v}^T Q_{\boldsymbol{\theta}}^T Q_{\boldsymbol{\theta}} \mathbf{v} \right)$$

Algorithm Rényi Fair Classifier for Discrete Sensitive Attributes

1: **Input**: $\boldsymbol{\theta}^0 \in \Theta$, step-size η.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: Set $\mathbf{v}^{t+1} \leftarrow \max_{\mathbf{v}\in\perp\mathbf{v}_1, \|\mathbf{v}\|\leq 1} f_D(\boldsymbol{\theta}^t, \mathbf{v})$ by finding the second singular vector of $\mathbf{Q}_{\boldsymbol{\theta}^t}$
4: Set $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}} f_D(\boldsymbol{\theta}^t, \mathbf{v}^{t+1})$
5: **end for**

➢ Converge to an ϵ-stationary solution in $\mathcal{O}(\epsilon^{-4})$ iterations (Jin, et al, 2019).

Jin, Chi, et. al. "Minmax optimization: Stable limit points of gradient descent ascent are locally optimal." (2019).

Binary Sensitive attribute:

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y)] + \lambda \rho_R^2(\hat{y}, s)$$

Theorem (Baharlouei, Nouiehed, Razaviyayn (2019)). *Suppose that $\hat{y} \in \{1, \ldots, c\}$ is a discrete random variable and $s \in \{0,1\}$ is a binary random variable. Let $\tilde{y}$ be the one-hot encoded version of $\hat{y}$. Let $\tilde{s} = s - 1/2$. Then,*

$$\rho_R^2(\hat{y}, s) \triangleq 1 - \frac{\gamma}{\mathbb{P}(s=1)\mathbb{P}(s=0)},$$

where $\gamma \triangleq \min_{\mathbf{w} \in \mathbb{R}^c} \quad \mathbb{E}\left[(\mathbf{w}^T \widetilde{y} - \tilde{s})^2\right].$

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{w}} \quad f_B(\boldsymbol{\theta}, \mathbf{w}) \triangleq \mathbb{E}\Big[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y) - \lambda \sum_{i=1}^{c} w_i^2 \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x}) + \lambda \sum_{i=1}^{c} w_i \tilde{s}_i \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x})\Big]$$

Binary Sensitive Attribute:

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{w}} \quad f_B(\boldsymbol{\theta}, \mathbf{w}) \triangleq \mathbb{E}\Big[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}), y) - \lambda \sum_{i=1}^{c} w_i^2 \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x}) + \lambda \sum_{i=1}^{c} w_i \tilde{s}_i \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x})\Big]$$

Algorithm Rényi Fair Classifier for Binary Sensitive Attributes

Input: $\boldsymbol{\theta}^0 \in \Theta$, step-size η.

for $t = 0, 1, \ldots, T$ **do**

Set $w_i^{t+1} \leftarrow \frac{\sum_{n=1}^{N} \widetilde{s}_n \mathcal{F}_i(\boldsymbol{\theta}^t, \mathbf{x}_n)}{2\sum_{n=1}^{N} \mathcal{F}_i(\boldsymbol{\theta}^t, \mathbf{x}_n)}, \quad \forall i = 1, \ldots, c$

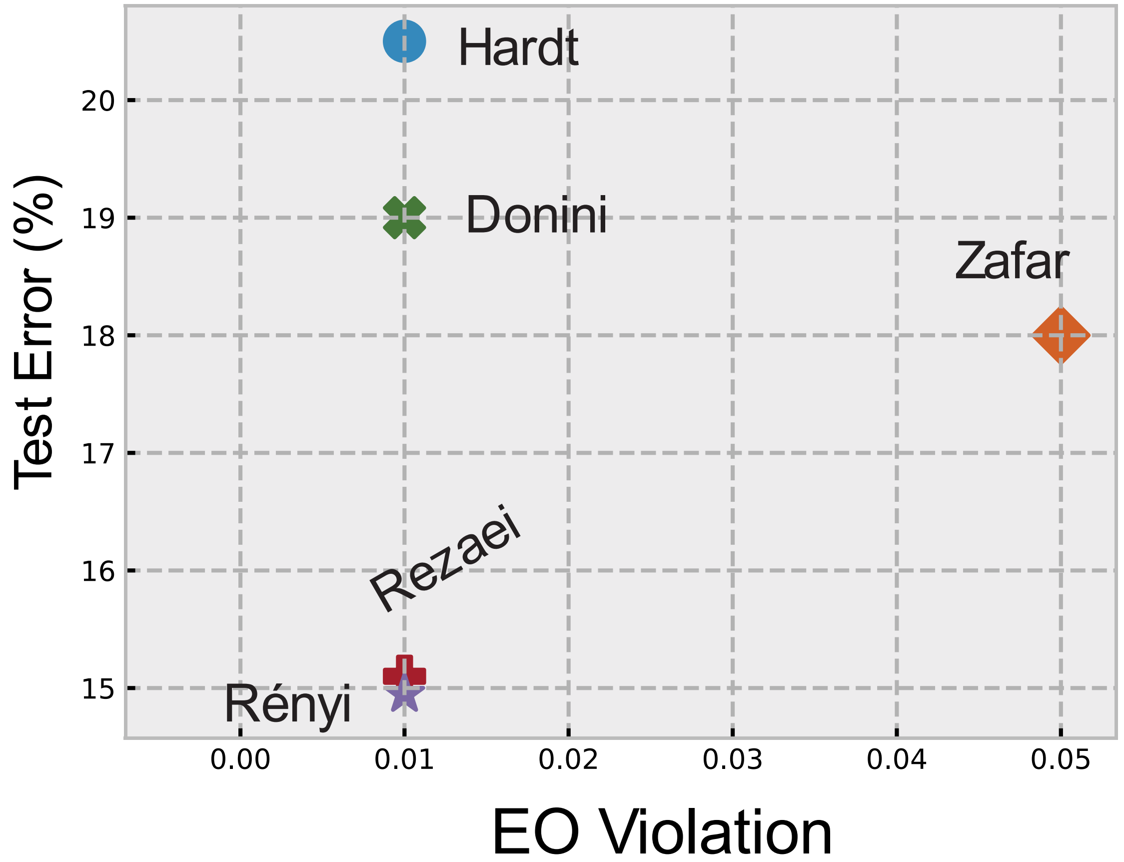
Set $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}} f_B(\boldsymbol{\theta}^t, \mathbf{w}^{t+1})$

end for

Theorem (Baharlouei, Nouiehed, Razaviyayn (2019)). *Suppose that f_B is L_1-gradient Lipschitz. Then the above algorithm computes an ε-stationary solution of the objective function in $\mathcal{O}(\varepsilon^{-2})$ iterations.*

Performance and Fairness (Equality of Opportunity Notion)

- Prediction task: Determine whether a person makes over 50K over a year



$$\text{EO Violation} = \left|\mathbb{P}(\hat{y} = 1 | s = 1, y = 1) - \mathbb{P}(\hat{y} = 1 | s = 0, y = 1)\right|$$

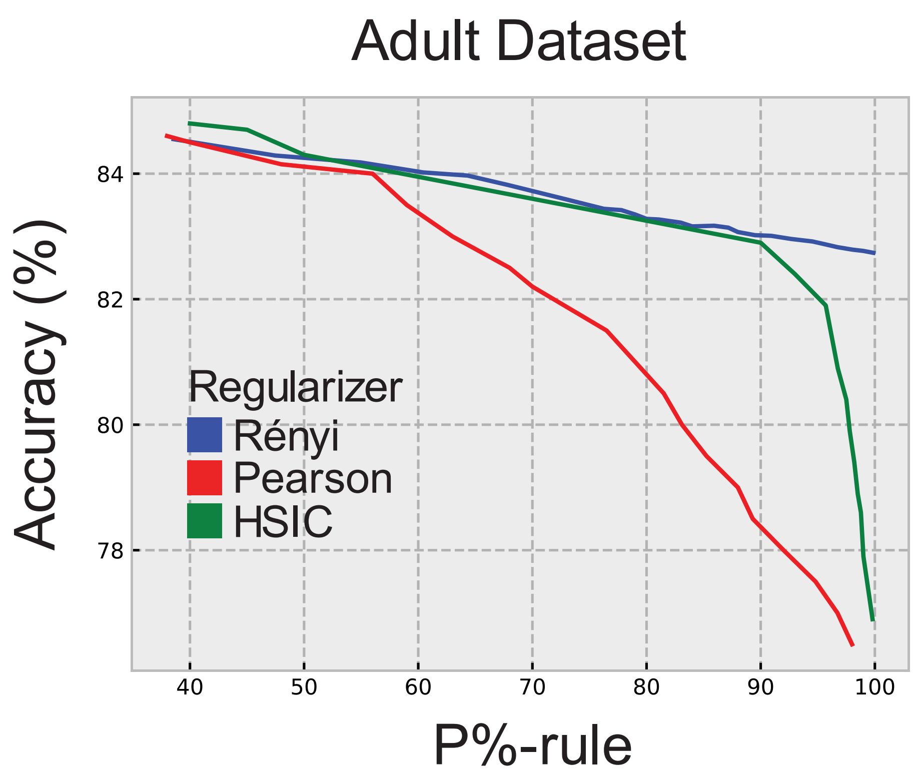
Zafar, et al. "Fairness constraints: Mechanisms for fair classification." (2015).

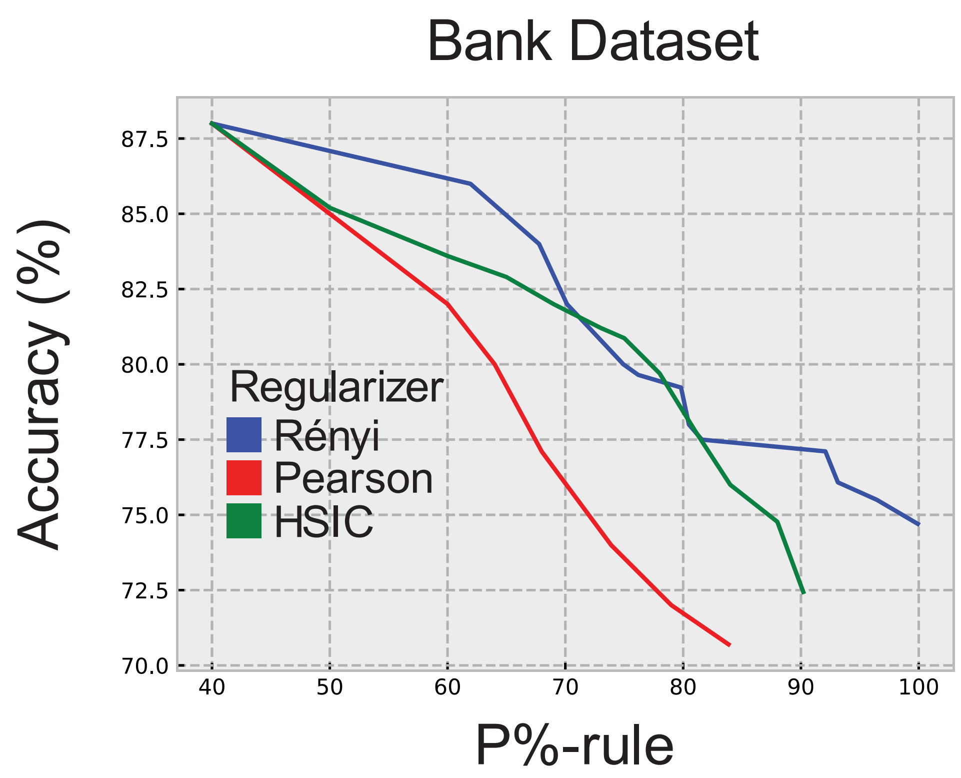
Hardt, et al. "Equality of opportunity in supervised learning." (2016).

Donini, et al. "Empirical risk minimization under fairness constraints." (2018).

Rezaei, et al. "Fair Logistic Regression: An Adversarial Perspective." (2019).

Performance and Fairness (Demographic Parity Notion)





$$p\% = \min\left(\frac{\mathbb{P}(\hat{y}{=}1|s{=}1)}{\mathbb{P}(\hat{y}{=}1|s{=}0)}, \frac{\mathbb{P}(\hat{y}{=}1|s{=}0)}{\mathbb{P}(\hat{y}{=}1|s{=}1)}\right) \times 100$$

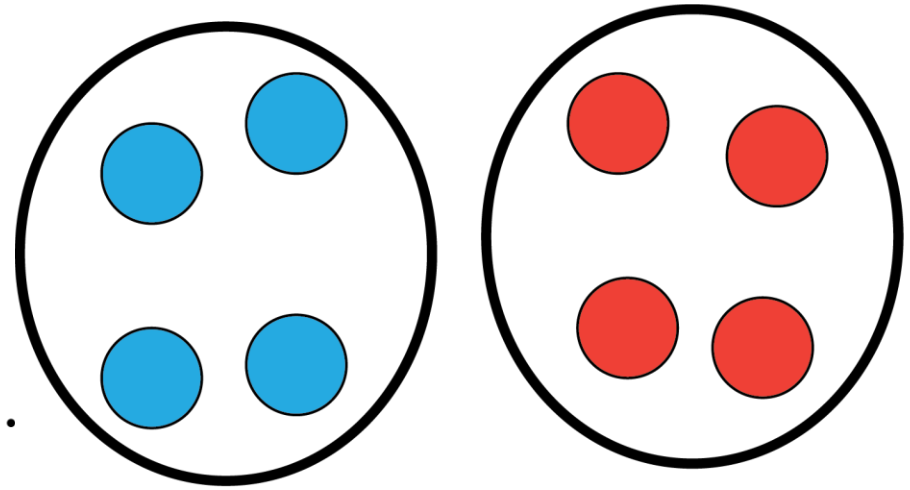
Zafar, et al. "Fairness constraints: Mechanisms for fair classification." (2015).

Pérez-Suay, Adrián, et al. "Fair kernel learning." (2017).

Fair Clustering (K-means)

$$\min_{\mathbf{A},\mathbf{C}} \max_{\mathbf{w}\in\mathbb{R}^K} \overbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} a_{kn}\|\mathbf{x}_n-\mathbf{c}_k\|^2}^{\text{K-means Objective Function}} - \lambda \overbrace{\sum_{n=1}^{N} (\mathbf{a}_n^T\mathbf{w}-s_n)^2}^{\text{Balance of Clusters}}$$

$$\text{s.t.} \quad \textstyle\sum_{k=1}^{K} a_{kn} = 1, \quad \forall\, n, \quad a_{kn} \in \{0,1\}, \quad \forall\, k,n.$$



a) Not balanced clusters

Algorithm Rényi Fair K-means

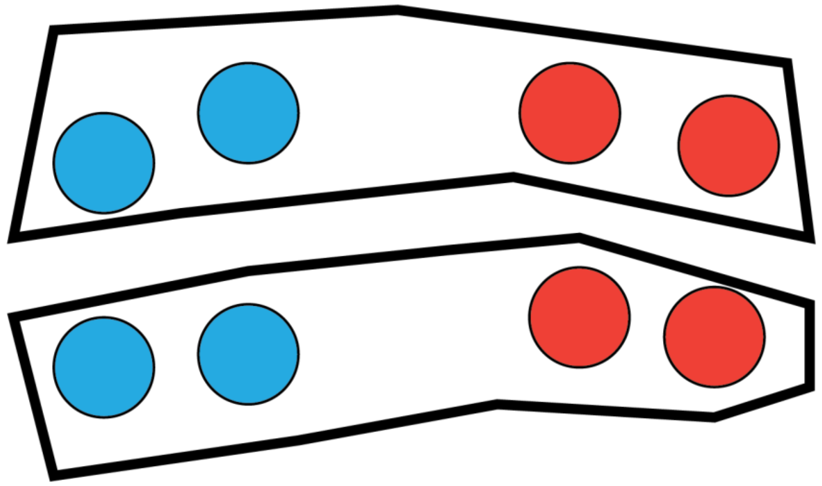
Input: $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ and $\mathbf{S} = \{s_1,\ldots,s_N\}$

Initialize: Random assignment $\mathbf{A}$ s.t. $\sum_{k=1}^{K} a_{kn} = 1\,\forall\, n$; and $a_{kn} \in \{0,1\}$. *Set* $\mathbf{A}_{prev} = \mathbf{0}$.

while $\mathbf{A}_{prev} \neq \mathbf{A}$ **do**

- Set $\mathbf{A}_{prev} = \mathbf{A}$
- **for** $n = 1,\ldots,N$ **do** ▷ Update $\mathbf{A}$
  - $k^* = \arg\min_k \|\mathbf{x}_n-\mathbf{c}_k\|^2 - \lambda(\mathbf{w}_k - s_n)^2$
  - Set $a_{k^*n} = 1$ and $a_{kn} = 0$ for all $k \neq k^*$
  - Set $w_k = \dfrac{\sum_{n=1}^{N} s_n a_{kn}}{\sum_{n=1}^{N} a_{kn}}$, $\forall k = 1,\ldots,K.$ ▷ Update $\mathbf{w}$
- **end for**
- Set $\mathbf{c}_k = \dfrac{\sum_{n=1}^{N} a_{kn}\mathbf{x}_n}{\sum_{n=1}^{N} a_{kn}}$, $\forall k = 1,\ldots,K.$ ▷ Update $\mathbf{c}$

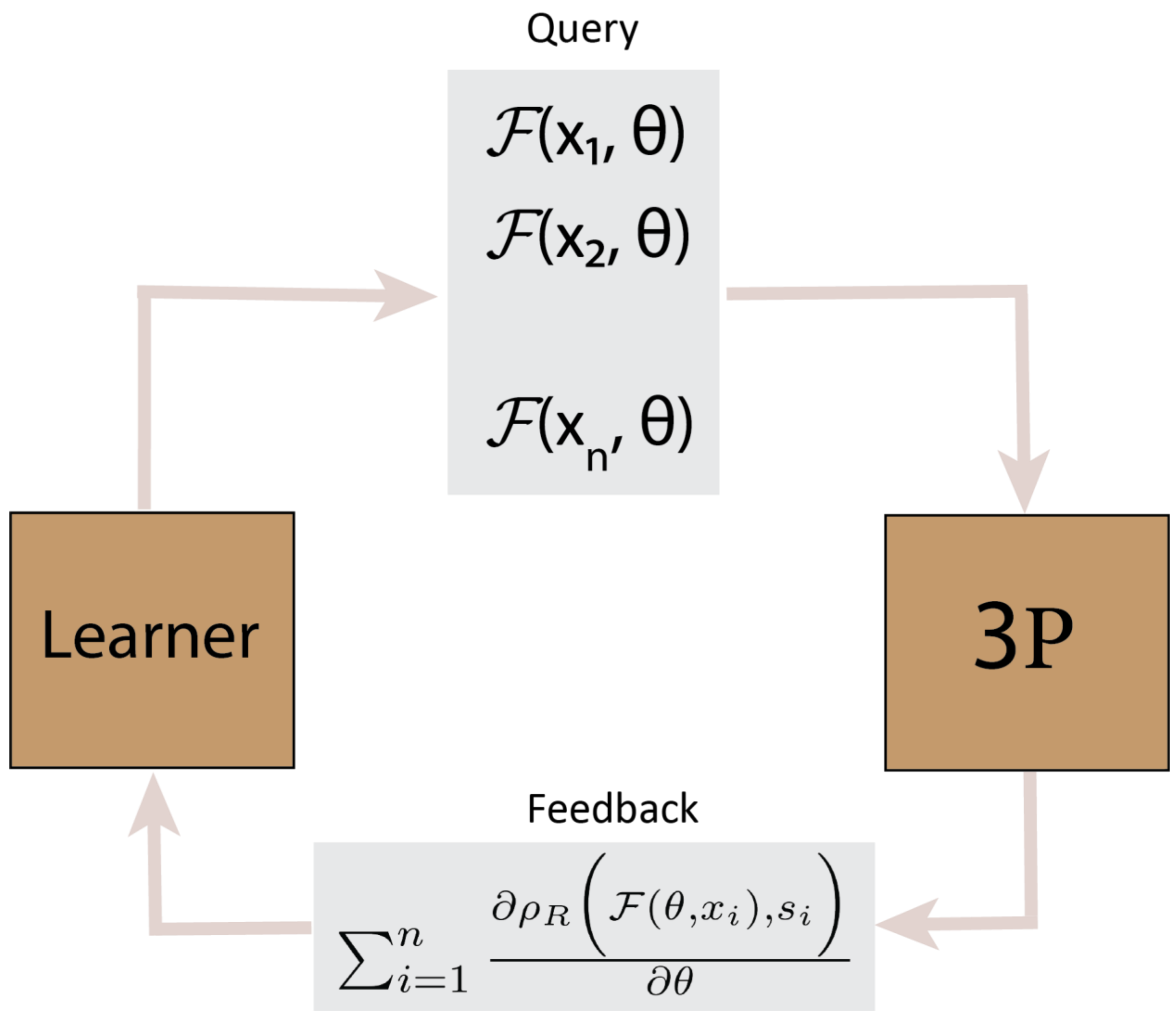
end while



b) Balanced clusters

Extensions and Future Directions:

- Fair Private Learning:



- Extension to Large-scale problems:
  - Unbiased estimator of gradient
  - Smoother function to optimize
- Extension to unsupervised learning problems:
  - Gaussian Mixture Models
  - Principle Component Analysis
- Fair Regression (or continuous sensitive attribute)

References

- Baharlouei, S., Nouiehed, M., Beirami, A., & Razaviyayn, M. (2020). Rényi Fair Inference. In *International Conference on Learning Representations*.