Neyman-Pearson and equal opportunity: when efficiency meets fairness in classification

J. Fan, X. Tong, Y. Wu, and S. Yao

Department of Data Sciences and Operations
University of Southern California

Concern for bias in algorithm

RISK & COMPLIANCE JOURNAL

AI Hiring Tools Can Violate Disability Protections, Government Warns

Justice Department, Equal Employment Opportunity Commission say companies whose use of AI tools leads to discrimination could face legal trouble

BUSINESS

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

A loan example

What is the goal of credit card issuing algorithm?

Societal concerns (fairness):

- Applicants from every race/gender should have the same probability of receiving credit cards.
- Applicants with the same profile should have the same outcome despite their race/gender.
- Etc.

Institutional interests (efficiency):

- Control the financial risks.
- Maximize profit.

How do people balance these goals?

A loan example

What profile a bank may rely on:

$$\underbrace{\text{payment history, annual income,..., credit score,}}_{\text{neutral attributes } X} \quad \underbrace{\text{gender}}_{\text{sensitive attribute } S} \quad , \underbrace{\text{default or not}}_{Y}$$

- In this project, $S \in \{a, b\}$ and $Y \in \{0, 1\}$
- $Y = 1$ implies "non-default" and $Y = 0$ implies "default".

When a new applicant walks into the bank:

$$\underbrace{\text{payment history, annual income,..., credit score,}}_{\text{known } X} \underbrace{\text{gender}}_{\text{known } S}, \underbrace{\text{default or not}}_{\text{prediction goal } Y}$$

- Bankers want an algorithm $\phi(X, S)$ to give them Y.
- The resources to build such a $\phi(X, S)$ is past records $(X_1, S_1, Y_1), (X_2, S_2, Y_2), \ldots$

Fairness in machine learning

In this scenario, what is algorithmic fairness?

- What about not using sensitive attributes in ϕ at all?

Common criteria in statistical sense? [Barocas et al.(2019)]

- Demographic Parity (Independence):

$$\mathbb{P}(\hat{\phi}(X,S)=1 \mid S=a)=\mathbb{P}(\hat{\phi}(X,S)=1 \mid S=b)$$

- Sufficiency:

$$\mathbb{P}(Y=1 \mid \hat{\phi}(X,S)=1, S=a)=\mathbb{P}(Y=1 \mid \hat{\phi}(X,S)=1, S=b)$$
$$\mathbb{P}(Y=1 \mid \hat{\phi}(X,S)=0, S=a)=\mathbb{P}(Y=1 \mid \hat{\phi}(X,S)=0, S=b)$$

- Equalized odds:

$$\mathbb{P}(\hat{\phi}(X,S)\neq Y \mid Y=1, S=a)=\mathbb{P}(\hat{\phi}(X,S)\neq Y \mid Y=1, S=b)$$
$$\mathbb{P}(\hat{\phi}(X,S)\neq Y \mid Y=0, S=a)=\mathbb{P}(\hat{\phi}(X,S)\neq Y \mid Y=0, S=b)$$

Fairness in machine learning

How do statisticians train a fair algorithm?

- Pre-processing methods, e.g., Re-weighting method. [Han et al. (2022)]
- In-processing methods, e.g., Optimization with penalty term. [Domini et al. (2018)]
- Post-processing methods, e.g., Calibration. [Hardt et al. (2016)]

Notations

Type I error: $R_0(\hat{\phi}) = \mathbb{P}(\hat{\phi}(X, S) = 1 \mid Y = 0)$

- Probability of issuing credit card to unqualified applicants.

Type II error: $R_1(\hat{\phi}) = \mathbb{P}(\hat{\phi}(X, S) = 0 \mid Y = 1)$

- Probability of NOT issuing credit card to qualified applicants.

Both $\mathbb{P}$ are taken conditional on $\hat{\phi}$. Both quantities are random since $\hat{\phi}$ is random.

Notations

Type I/II error conditional on sensitive attribute:

$$R_y^s(\hat{\phi}) = \mathbb{P}\left(\hat{\phi}(X,S) \neq Y \mid Y = y, S = s\right),$$

Type I error disparity: $L_0(\hat{\phi}) = |R_0^a(\hat{\phi}) - R_0^b(\hat{\phi})|$

- Difference of probabilities that unqualified applicants from group a and b are given credit cards.

Type II error disparity: $L_1(\hat{\phi}) = |R_1^a(\hat{\phi}) - R_1^b(\hat{\phi})|$

- Difference of probabilities that qualified applicants from group a and b are NOT given credit cards.

Both disparities are random.

Equal Opportunity

The equalized odds condition(Hardt et al., (2016))

$$L_0(\hat{\phi}) = L_1(\hat{\phi}) = 0$$

- Qualified and unqualifed applicants from each group have the same probability to get the same outcome.

The equalized odds condition means that candidates from different social groups have the same probability to obtain both types of outcomes.
This condition is very stringent.

Equal Opportunity

The equal opportunity (EO) condition(Hardt et al., (2016))

$$L_1(\hat{\phi}) = 0\,.$$

We care about type II error disparity because it is the concern of the society.

- $L_0 = 0$ implies default applicants from different groups are equally likely to get credit cards.
- $L_1 = 0$ implies non-default from different groups are equally likely to not get credit cards.
- People will fight for what they deserve.

This condition is also impossible because $\hat{\phi}$ is random.

Equal Opportunity

A less stringent EO condition:

$$L_1(\hat{\phi}) \leq \varepsilon\,.$$

- This criterion addresses the fairness concern of the society.
- A good fair algorithm also needs to be useful, or efficient.

Neyman-Pearson classification

What do banks need from a default detection algorithm?

- Reduce financial risk: Do not classify too many "default" (0) as "non-default" (1). That is, avoid issuing credit card to unqualified applicants to control the financial risk.
  - Reduce $R_0(\hat{\phi})$

- Increase profit: Identify as many "non-default" as possible. That is, issue as many cards as possible to qualified applicants.
  - Reduce $R_1(\hat{\phi})$

However, type I/II errors trade-off is not uncommon.

Neyman-Pearson classifcation

How to balance the two goals?

- Financial risk is more crucial than profit.
- Banks often have hard constraint for financial risk.

Banks can control the financial risk at certain level, and then maximizing the profit.

NP and NP-EO paradigm

The Neyman-Pearson classification paradigm (Cannon et al. [2002], Scott and Nowak [2005], Rigollet and Tong [2011], Tong et al. [2018], Yao et al. [2022])

$$\phi^* \in \underset{\phi: R_0(\phi) \leq \alpha}{\arg\min}\ R_1(\phi)$$

The NP-EO paradigm:

$$\phi^* \in \underset{\substack{\phi: R_0(\phi) \leq \alpha \\ L_1(\phi) \leq \varepsilon}}{\arg\min}\ R_1(\phi)$$

- Institutional efficiency: $R_0(\phi) \leq \alpha$, $\arg\min R_1(\phi)$
- Societal fairness: $L_1(\phi) \leq \varepsilon$

Does this classifier exists? If so, what does it look like?

NP-EO oracle classifier

The Neyman-Pearson Equal-Opportunity classification paradigm can be formmulated as follows:

Theorem 1

Under mild continuity conditions for X, the NP-EO oracle classifier exists and has the form

$$\mathbb{1}\left\{\frac{f(X \mid S=a, Y=1)}{f(X \mid S=a, Y=0)} > c_a^*\right\}\mathbb{1}\{S=a\} + \mathbb{1}\left\{\frac{f(X \mid S=b, Y=1)}{f(X \mid S=b, Y=0)} > c_b^*\right\}\mathbb{1}\{S=b\}$$

Here, f is the density function of X.

NP-EO umbrella algorithm

We have established the NP-EO oracle classifier:

$$\mathbb{1}\left\{\frac{f(X \mid S=a, Y=1)}{f(X \mid S=a, Y=0)} > c_a^*\right\}\mathbb{1}\{S=a\} + \mathbb{1}\left\{\frac{f(X \mid S=b, Y=1)}{f(X \mid S=b, Y=0)} > c_b^*\right\}\mathbb{1}\{S=b\}$$

How do we train an NP-EO classifier, especially if we want to use certain classification algorithms, e.g., logistic regression, neural networks, etc? That is, we want to train a classifier of the form

$$\hat{\phi}(X,S) = \mathbb{1}\left\{\hat{T}^a(X) > \hat{c}_a\right\}\mathbb{1}\{S=a\} + \mathbb{1}\left\{\hat{T}^b(X) > \hat{c}_b\right\}\mathbb{1}\{S=b\},$$

where $\hat{T}^a$, $\hat{T}^b$ are scoring functions trained by user-specified classification algorithms. Moreover, $\hat{c}_a$ and $\hat{c}_b$ need to be determined by data.

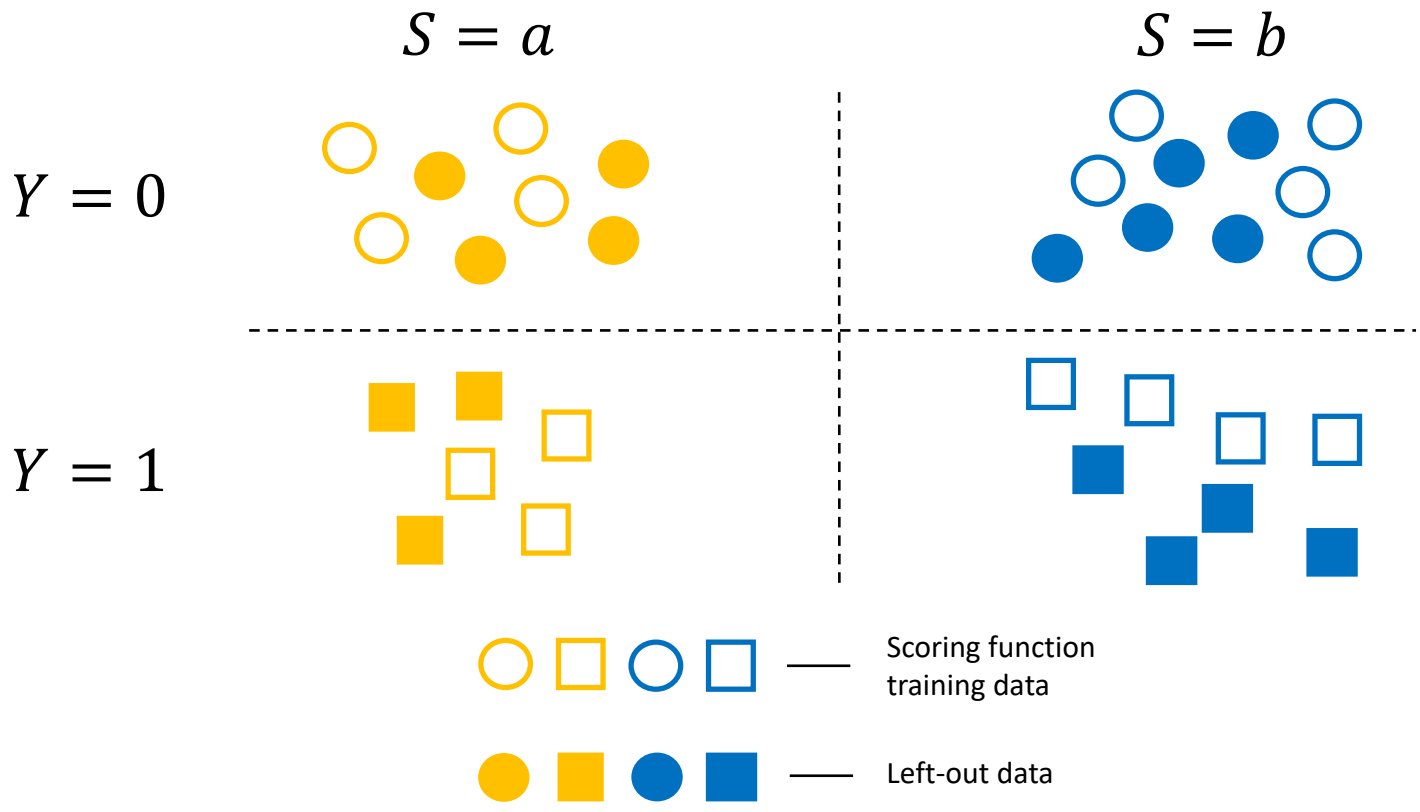
NP-EO umbrella algorithm

Since $\hat{\phi}$ is trained from data, $R_0(\hat{\phi}) \leq \alpha$ and $L_1(\hat{\phi}) \leq \varepsilon$ almost surely cannot be achieved. Instead, we seek high probability versions.

- high probability NP condition: $\mathbb{P}(R_0(\hat{\phi}) > \alpha) \leq \delta$.
- high probability EO condition: $\mathbb{P}(L_1(\hat{\phi}) > \varepsilon) \leq \gamma$.
- δ, γ are user-specified.

Here, $\mathbb{P}$ is taken with respect to randomness of data.

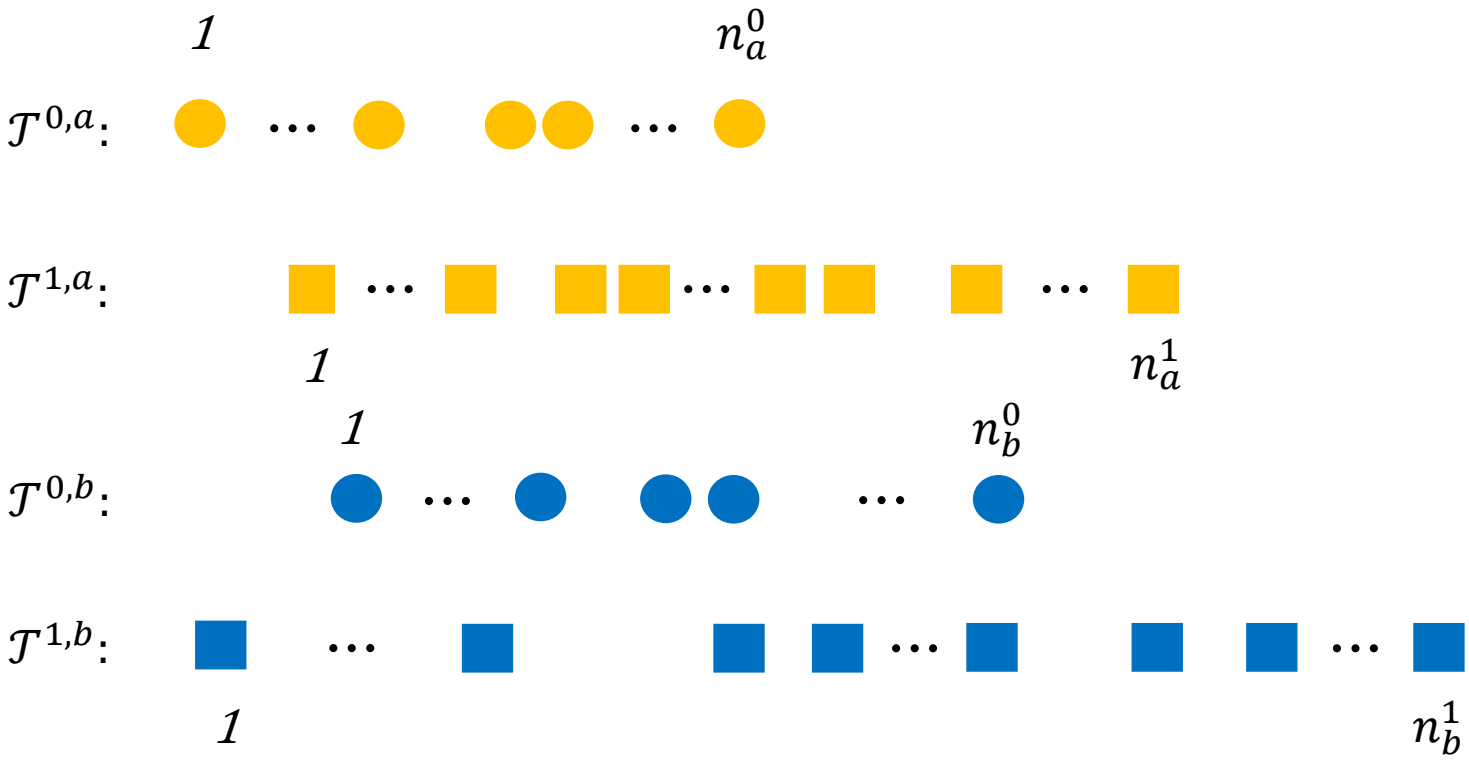
NP-EO umbrella algorithm



- The scoring function training is used to train a scoring function $\hat{T}$.
- $\hat{T}^a(\cdot) = \hat{T}(\cdot, a)$, $\hat{T}^b(\cdot) = \hat{T}(\cdot, b)$.
- Apply $\hat{T}$ to all corresponding parts of left-out data.

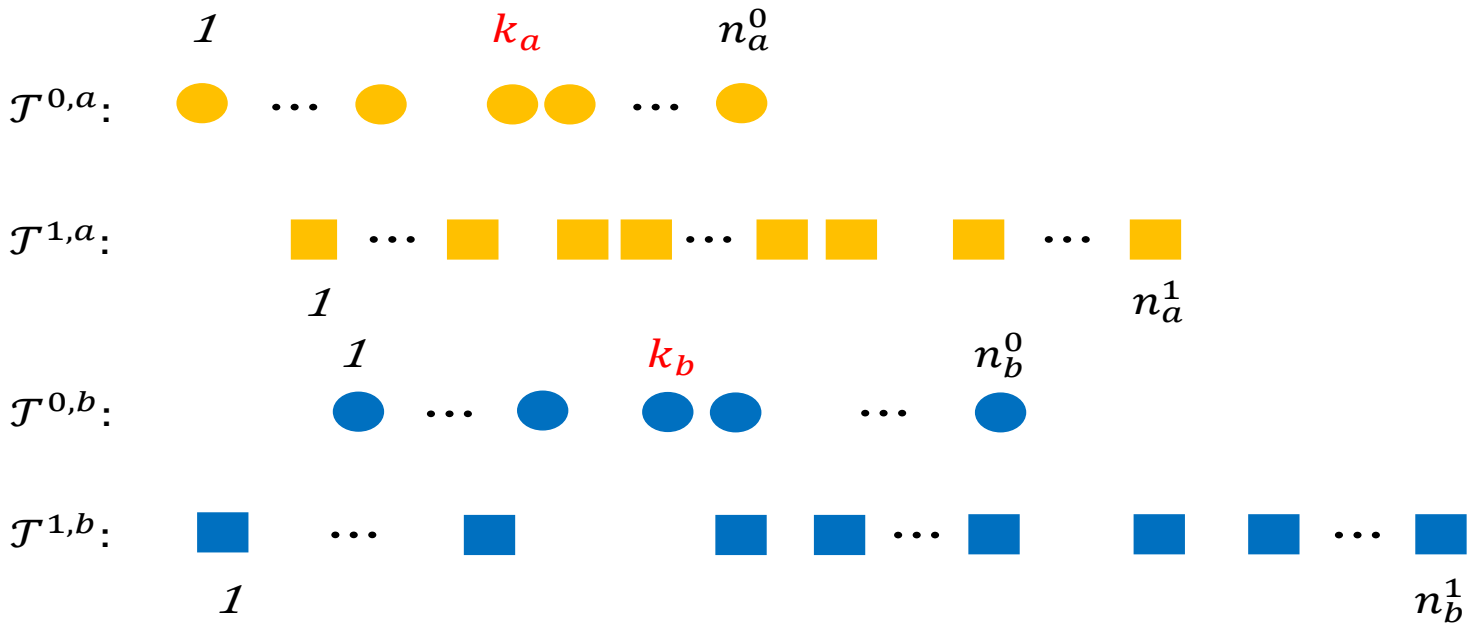
NP-EO umbrella algorithm

Applying $\hat{T}$ to left-out data:



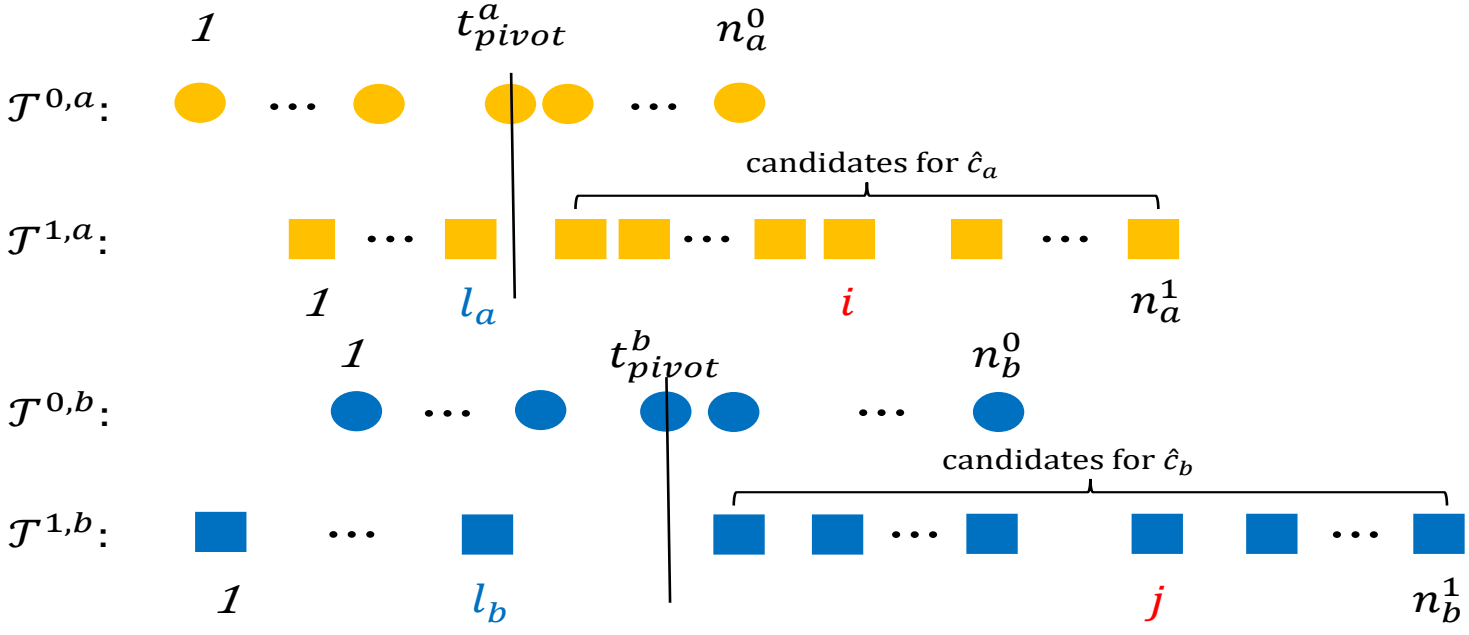
We plan to find $\hat{c}_a$ and $\hat{c}_b$ among these candidates.

NP-EO umbrella algorithm



- **NP umbrella algorithm** [Tong et al., (2018)]: select k such that the scoring based classifier that uses the k^{th} order statistic among $Y = 0$ sample of size n satisfies high probability NP condition at any level α, δ.
- $R_0 = R_0^a \mathbb{P}(S = a \mid Y = 0) + R_b^a \mathbb{P}(S = b \mid Y = 0)$. If R_0^a and R_0^b can be controlled separately, R_0 can also be controlled.

NP-EO umbrella algorithm



- Let $r_1^a(i)$ be $R_1^a(\hat{\phi})$ if we select the i^{th} order statistic in $\mathcal{T}^{1,a}$ as $\hat{c}_a$ in $\hat{\phi}$.
- Let $r_1^b(j)$ be $R_1^b(\hat{\phi})$ if we select the j^{th} order statistic in $\mathcal{T}^{1,b}$ as $\hat{c}_b$ in $\hat{\phi}$
- $i > l_a$, $j > l_b$.

NP-EO umbrella algorithm

The EO violation rate:

$$\mathbb{P}(L_1(\hat{\phi}) > \delta) = \mathbb{E}_{\hat{T}}\mathbb{E}_{I_a,I_b}\underbrace{\mathbb{P}(|r_1^a(i) - r_1^b(j)| > \delta \mid I_a, I_b)}_{\leq \gamma}$$

- Conditional on $\hat{T}$, quantities involving a are independent of quantities involving b.
- For every i, j, we want to approximate the distributions of $r_1^a(i) \mid I_a$ and $r_1^b(j) \mid I_b$.
- The distribution of $r_1^a(i)$ can be approximated by $Z + (1 - Z)B$ where Z is a normal distribution and B is a Beta distribution. The parameters in the distributions of Z and B are known function of i, I_a and n_a^1.
- The distribution of $r_1^b(j)$ can be approximated analogously.
- Select i, j such that EO violation rate is smaller than γ.
- For all feasible pairs of i, j, select the one that minimizes the empirical type II error.

NP-EO umbrella algorithm

Theorem 2

Under certain regularity conditions, the classifier $\hat{\phi}$ trained by NP-EO umbrella algorithm satisfies

1. $\mathbb{P}(R_0(\hat{\phi}) > \alpha) \leq \delta$.
2. $\mathbb{P}(L_1(\hat{\phi}) > \varepsilon) \leq \gamma + r(n_a^1, n_b^1)$, where $r(n_a^1, n_b^1)$ converges to 0 as n_a^1, n_b^1 go to infinity.

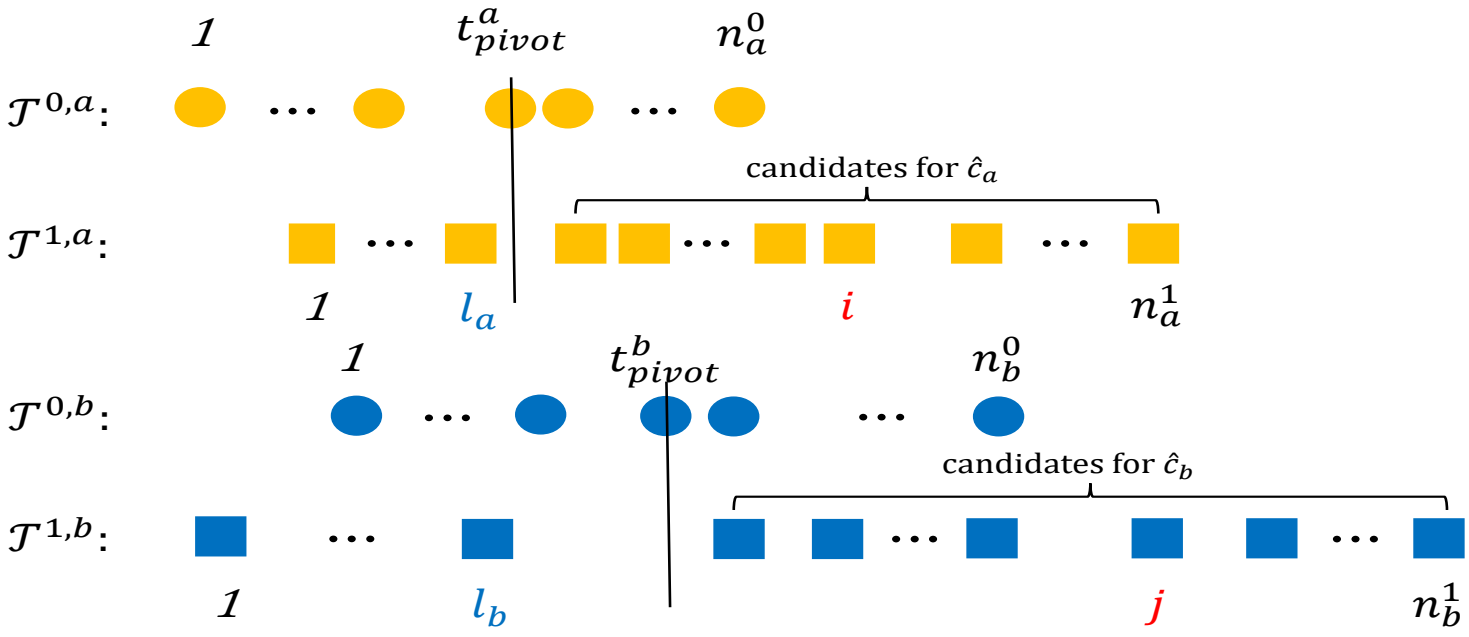
NP-EO umbrella algorithm

A bit of simulation: let X be Gaussian with different means conditional on different values of Y and S. Set $\alpha = 0.05$, $\delta = 0.05$, $\gamma = 0.2$ and $\varepsilon = 0.05$.

	average of type I errors	average of type II errors	NP violation rate	EO violation rate
NP-EO	0.012	0.480	0	0.046

- The NP-EO umbrella algorithm is able to satisfy NP and EO condition with desired high probability.
- Recall that $R_0(\phi^*_{\text{NP-EO}}) = \alpha$. However, the NP violation rate for NP-EO umbrella algorithm is 0. This contradicts the property of NP oracle classifier philosophically.

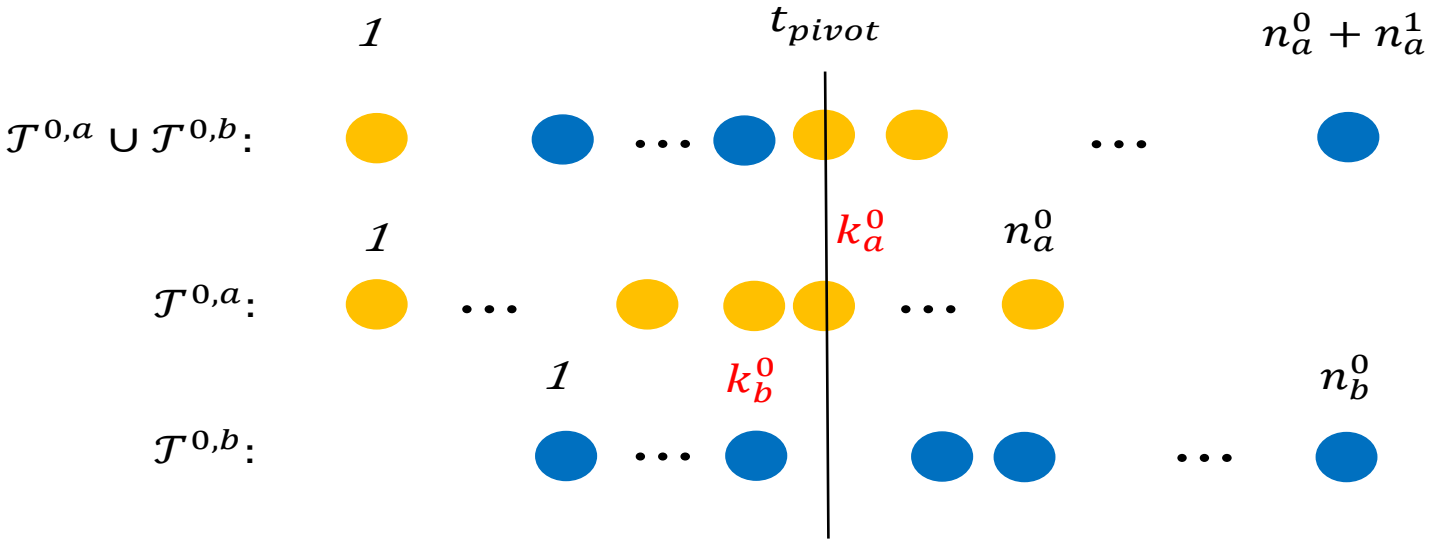
NP-EO umbrella algorithm - modified approach



- Note that both pivots are selected by NP umbrella algorithm, which ensures the high probability NP condition.
- Both selected thresholds are larger than the pivots to guarantee high probability NP condition.

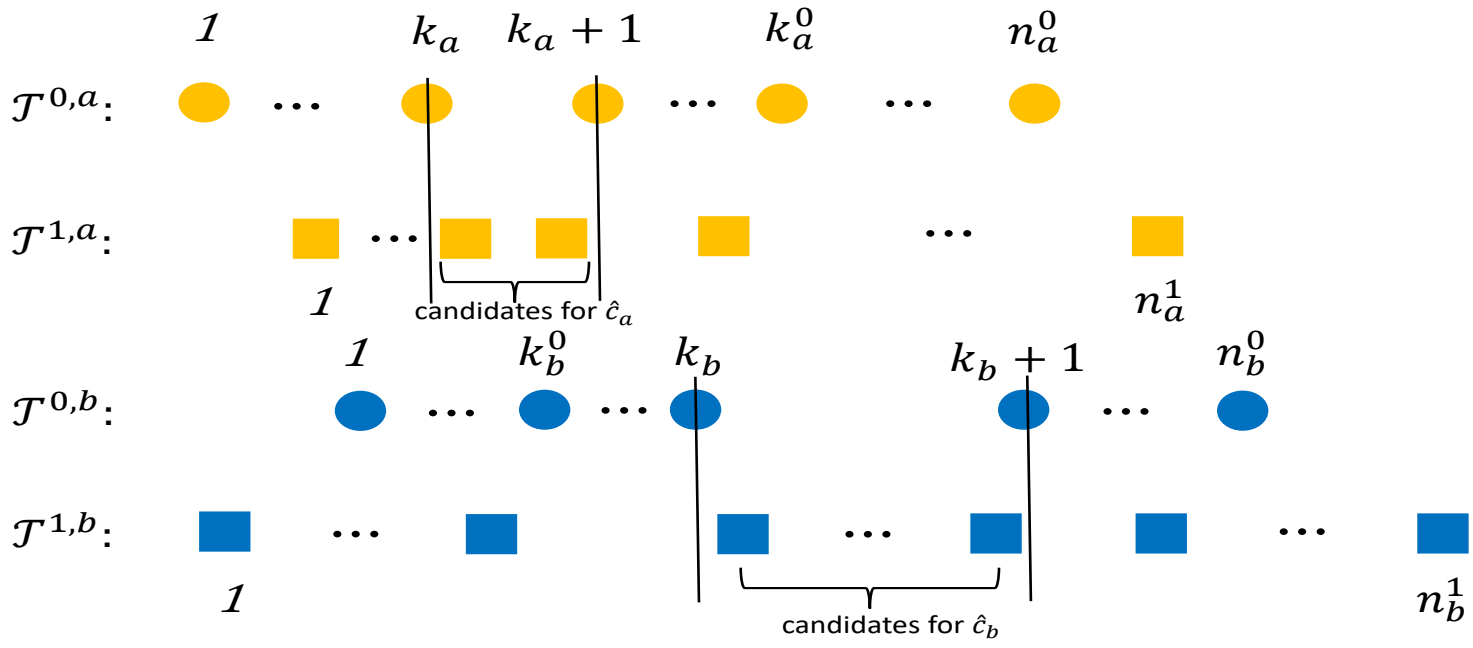
What if we relax this guarantee?

NP-EO umbrella algorithm - modified approach



- Apply NP umbrella algorithm to $\mathcal{T}^{0,a} \cup \mathcal{T}^{0,b}$ and select t_{pivot}. That is, t_{pivot} as a threshold controls R_0 with high probability.
- The two separate pivots as thresholds achieve the same empirical R_0.

NP-EO umbrella algorithm - modified approach



- We look at all k_a, k_b such that $k_a + k_b = k_a^0 + k_b^0$. That is, the empirical R_0 is not changed.
- $\hat{c}_a$ and $\hat{c}_b$ are selected to satisfy high probability EO condition using the similar approximation approach.
- k_a can be smaller than k_a^0 (or k_b can be smaller than k_b^0). This allows us to select smaller thresholds.

NP-EO umbrella algorithm - modified approach

Theorem 3

Under certain regularity conditions, the classifier $\hat{\phi}$ trained by modified NP-EO umbrella algorithm satisfies

1. $\mathbb{P}(R_0(\hat{\phi}) > \alpha) \leq \delta + r_0(n_a^0, n_b^0)$.
2. $\mathbb{P}(L_1(\hat{\phi}) > \varepsilon) \leq \gamma + r_1(n_a^1, n_b^1)$, where $r_0(n_a^0, n_b^0)$ and $r_1(n_a^1, n_b^1)$ converge to 0 as sample sizes go to infinity.

NP-EO umbrella algorithm - modified approach

Simulation: let X be Gaussian with different means conditional on different values of Y and S. Set $\alpha = 0.05$, $\delta = 0.05$, $\gamma = 0.2$ and $\varepsilon = 0.05$.

	average of type I errors	average of type II errors	NP violation rate	EO violation rate
NP-EO	0.012	0.480	0	0.046
NP-EO modified	0.039	0.387	0.033	0.029

Numerical Analysis

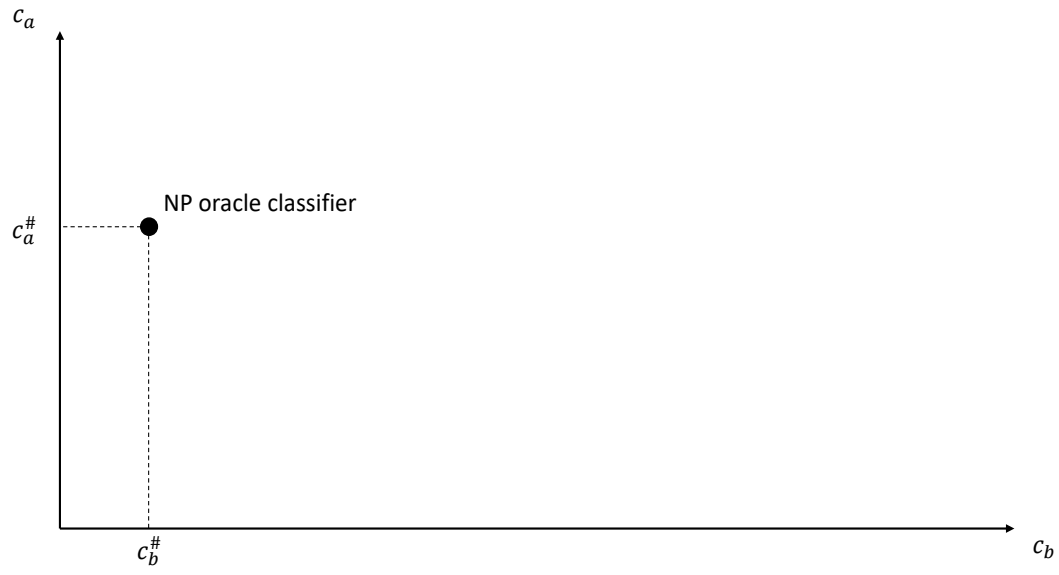
Real data analysis: We use the credit card dataset, where sensitive attribute is gender, neutral attributes are payment history and othe demographic features. $\alpha = 0.1$, $\delta = 0.1$, $\varepsilon = 0.05$, $\gamma = 0.1$.

	average of type I errors	average of type II errors	NP violation rate	EO violation rate
NP-EO	0.081	0.720	0.033	0.034
NP-EO modified	0.089	0.701	0.114	0.054
NP	0.088	0.700	0.111	0.482
classic	0.633	0.059	1	0

- If classic paradigm is used, then with probability 1 (the goal is $\delta = 0.1$), the bank fails to control the financial risk under (type I error) 0.1.
- If NP paradigm is used, then with probability 0.482 (the goal is $\gamma = 0.1$), the bank fails to keep the equal opportunity disparity under 0.05.
- If NP-EO paradigm is used, then the probability of NOT able to control financial risk under $\alpha = 0.1$ is lower than $\delta = 0.1$, and the probability of NOT able to control the fairness disparity under $\varepsilon = 0.05$ is lower than $\gamma = 0.1$.

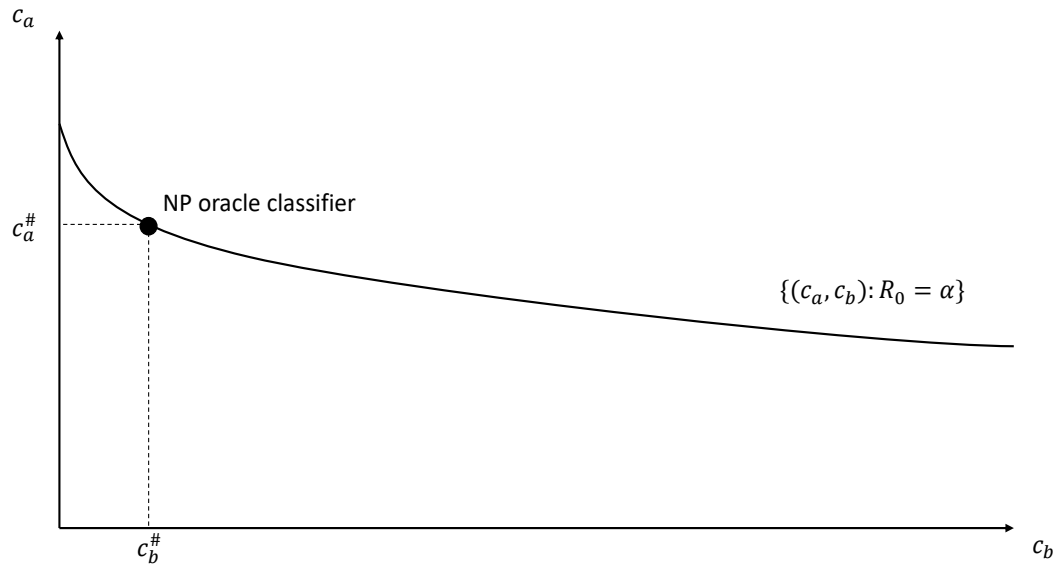
Thank You!

NP-EO oracle classifier



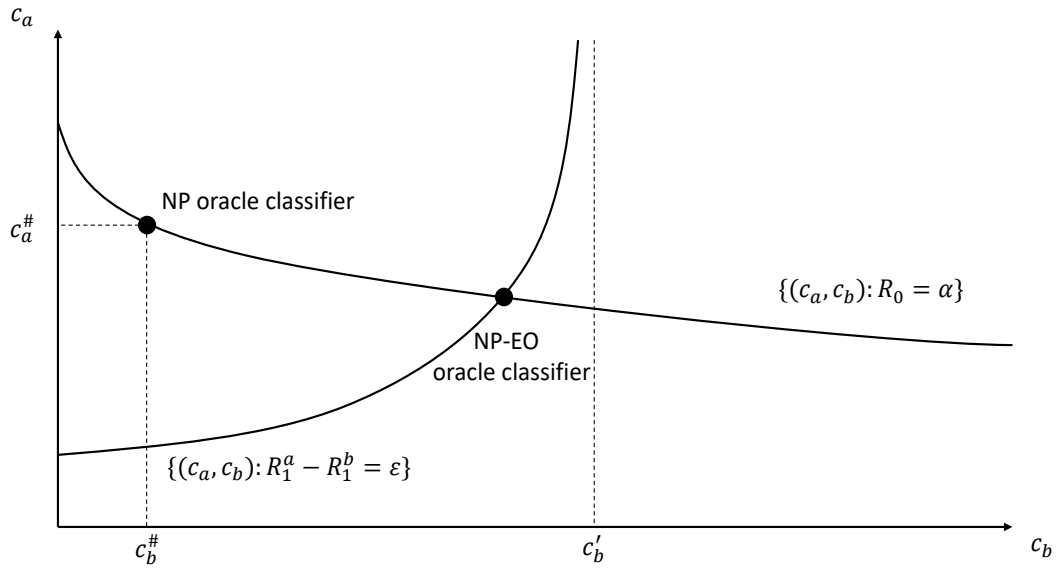
Every point (c_a, c_b) in the first quadrant represents a classifier with threshold pair (c_a, c_b)

NP-EO oracle classifier



- $R_0 = \mathbb{P}\left(\frac{f(X|S=a,Y=1)}{f(X|S=a,Y=0)} > c_a \mid S = a, Y = 0\right) \mathbb{P}(S = a \mid Y = 0)$
  $+ \mathbb{P}\left(\frac{f(X|S=b,Y=1)}{f(X|S=b,Y=0)} > c_b \mid S = b, Y = 0\right) \mathbb{P}(S = b \mid Y = 0).$
- NP oracle classifier achieves $R_0 = \alpha$.

NP-EO oracle classifier



- $R_1^a - R_1^b = \mathbb{P}\left(\frac{f(X|S=a,Y=1)}{f(X|S=a,Y=0)} \leq c_a \mid S = a, Y = 0\right)$
  $-\mathbb{P}\left(\frac{f(X|S=b,Y=1)}{f(X|S=b,Y=1)} \leq c_b \mid S = b, Y = 1\right)$
- $R_1^b = 1 - \varepsilon$ at c_b'.
- $R_1^a(\phi_{\text{NP}}^*) - R_1^b(\phi_{\text{NP}}^*) > \varepsilon$.

NP-EO umbrella algorithm

Let $t^{1,a}$ be an arbitrary element in $\mathcal{T}^{1,a}$ and $t^{1,a}_{(i)}$ be the i^{th} order statistic in $\mathcal{T}^{1,a}$. Furthermore, let $T^{1,a} = \hat{T}^a(X) \mid (S = a, Y = 1)$. Given $\hat{T}$, we have:

- $\mathbb{P}(T^{1,a} \leq t^{1,a} \mid t^{1,a})$ is uniformly distributed.
- $r^a_1(i) = \mathbb{P}(T^{1,a} \leq t^{1,a}_{(i)} \mid t^{1,a})$ is Beta distributed.
- $l_a = \sum_{t^{1,a} \in \mathcal{T}^{1,a}} \mathbb{1}\{t^{1,a} \leq t^a_{pivot}\}$.
- $r^a_1(i) \mid l_a$ has the same distribution as

$$F^a \mid l_a + (1 - F^a \mid l_a) Beta(n^1_a - l_a, n^1_a - i + 1)\,,$$

where the distribution of $F^a = \mathbb{P}(t^{1,a} \leq t^a_{\text{pivot}} \mid t^a_{\text{pivot}})$.

NP-EO umbrella algorithm

The distribution of $F^a \mid I_a$ is still unknown.
Recall that $I_a = \sum_{t^{1,a} \in \mathcal{T}^{1,a}} \mathbb{1}\{t^{1,a} \leq t^a_{pivot}\}$ and $F^a = \mathbb{P}(t^{1,a} \leq t^a_{\text{pivot}} \mid t^a_{\text{pivot}})$.

- I_a is the sum of independent $Bernoulli(F_a)$ random variables.
- I_a / n^1_a is the maximum likelihood estimator of F^a.

So how do estimate the distribution of $F^a \mid I_a$?

- A naive solution: I_a / n^1_a. However, this ignores the randomness of t^a_{pivot}.
- A more precise solution: Bernstein-von Mises theorem. Since the distribution of $F^a \mid I_a$ is the posterior distribution, it is "close to" a normal distribution, where n^1_a is the cardinality of $\mathcal{T}^{1,a}$.