

Enforcing Fair Predicted Scores on Intervals of Percentiles by Difference-of-Convex Constraints

Yutian He¹, Yankun Huang², Yao Yao³, Qihang Lin⁴

¹Department of Mathematics, University of Iowa; ²Department of Information Systems, Arizona State University

³School of Mathematics, University of Minnesota Twin Cities; ⁴Department of Business Analytics, University of Iowa

1

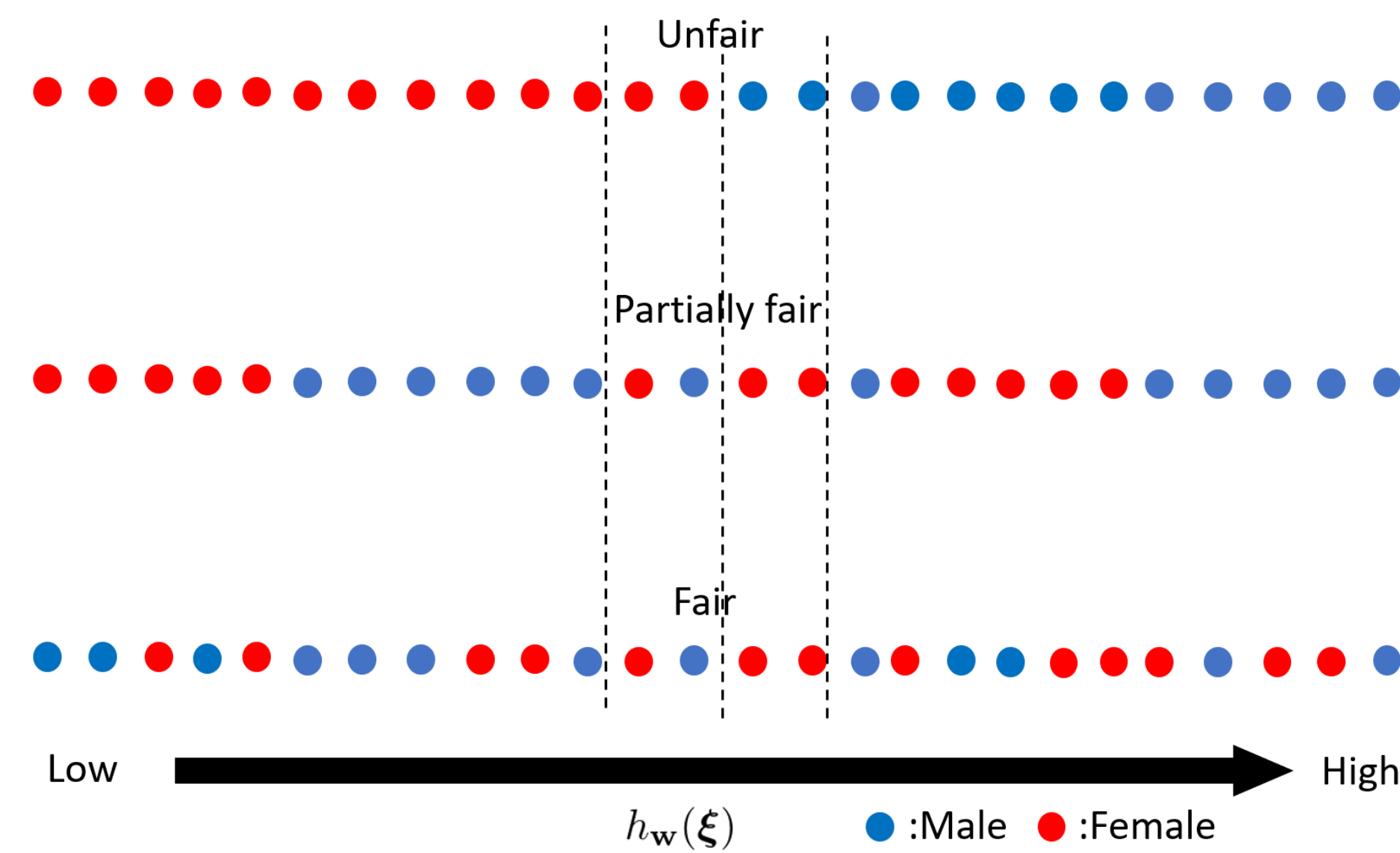
Introduction

Preliminary and notation

- $(\xi, \zeta, \gamma) \in \mathbb{R}^{d+2}$: Data point.
 - $\xi \in \mathbb{R}^d$: Feature vector.
 - $\zeta \in \mathbb{R}$: Target variable.
 - $\gamma \in \mathcal{G} := \{1, 2, \dots, G\}$: Sensitive attribute (sex, race, etc.).
- $h_{\mathbf{w}}: \mathbb{R}^d \rightarrow \mathbb{R}$: Predictive model.
 - Parameterized by $\mathbf{w}$.
 - $h_{\mathbf{w}}(\xi)$: Prediction score for ξ .

Goal: How to ensure $h_{\mathbf{w}}$ is fair across groups defined by γ ?

Suppose top 40% ~ 60% applicants scored by $h_{\mathbf{w}}$ will be admitted.



In some case, it suffices to ensure the fairness within certain score region (e.g., top 40%~60%), known as **partial fairness**.

2

Fairness and Partial Fairness

Statistical Parity

Definition

$h_{\mathbf{w}}$ satisfies *statistical parity* (also called distributional parity) if the distribution of $h_{\mathbf{w}}(\xi)$ is independent of γ , which means

$$\begin{aligned} & \text{SP}_{k,k'}(\mathbf{w}) \\ &:= \max_{\theta} |\Pr(h_{\mathbf{w}}(\xi) > \theta | \gamma = k) - \Pr(h_{\mathbf{w}}(\xi) > \theta | \gamma = k')| \\ &= 0 \end{aligned}$$

for any k and k' in $\mathcal{G}$.

Partial Statistical Parity

Definition

$h_{\mathbf{w}}$ satisfies *partial statistical parity with respect to interval* $\mathcal{I} := [\alpha, \beta] \subset [0, 1]$ if

$$\begin{aligned} & \text{SP}_{k,k'}^{\mathcal{I}}(\mathbf{w}) \\ &:= \max_{\theta} \left| \Pr(h_{\mathbf{w}}(\xi) > \theta | \gamma = k, \bar{F}_{\mathbf{w},k}(h_{\mathbf{w}}(\xi)) \in \mathcal{I}) \right. \\ & \quad \left. - \Pr(h_{\mathbf{w}}(\xi) > \theta | \gamma = k', \bar{F}_{\mathbf{w},k'}(h_{\mathbf{w}}(\xi)) \in \mathcal{I}) \right| \\ &= 0 \end{aligned}$$

for any k and k' in $\mathcal{G}$.

Optimization Problem

$$\min_{\mathbf{w}} f_0(\mathbf{w}) \text{ s.t. } \text{DP}_{k,k'}^{\mathcal{I}}(\mathbf{w}) \leq \kappa, \forall k, k' \in \mathcal{G},$$

- $f_0(\mathbf{w})$: Loss function.
- $\kappa \in \mathbb{R}_+$: Tolerance level.

Lemma

$\text{SP}_{k,k'}^{\mathcal{I}}(\mathbf{w}) \leq \kappa, \forall k, k' \in \mathcal{G}$ if and only if, for any $p \in [\alpha, \beta - \kappa(\beta - \alpha)]$, there exists $\theta_p \in \mathbb{R}$ such that

$$\begin{aligned} & \Pr(h_{\mathbf{w}}(\xi) > \theta_p | \gamma = k) \geq p, \forall k \in \mathcal{G}, \\ & \Pr(h_{\mathbf{w}}(\xi) > \theta_p | \gamma = k) \leq p + \kappa(\beta - \alpha), \forall k \in \mathcal{G}. \end{aligned}$$

Training data:

$$\mathcal{D}_k = \{(\xi_i^k, \zeta_i^k, k)\}_{i=1}^{n_k}, \quad k = 1, 2, \dots, G.$$

1. Constraints with empirical distribution:

$$\begin{aligned} & \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{1}_{h_{\mathbf{w}}(\xi_i^k) > \theta_p} \geq p, \\ & \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{1}_{h_{\mathbf{w}}(\xi_i^k) > \theta_p} \leq p + \kappa(\beta - \alpha), \\ & \forall k \in \mathcal{G}, \forall p \in [\alpha, \beta - \kappa(\beta - \alpha)]. \end{aligned}$$

2. Approximate $\mathbf{1}_{x>0}$ by a continuous surrogate $\sigma(x)$.
3. Approximate $[\alpha, \beta - \kappa(\beta - \alpha)]$ by a finite set $\hat{\mathcal{I}}$.

Partial Statistical Parity

$$\begin{aligned} & \min_{\mathbf{w}, (\theta_p)_{p \in \hat{\mathcal{I}}}} f_0(\mathbf{w}) \\ & \text{s.t. } \frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\xi_i^k) - \theta_p) \geq p, \forall p \in \hat{\mathcal{I}}, k \in \mathcal{G}, \\ & \quad \frac{1}{n_k} \sum_{i=1}^{n_k} \sigma(h_{\mathbf{w}}(\xi_i^k) - \theta_p) \leq p + \kappa(\beta - \alpha), \forall p \in \hat{\mathcal{I}}, k \in \mathcal{G}, \end{aligned}$$

- $\sigma(x) = \max\{\min\{x + 0.5, 1\}, 0\}$.
- $\sigma(x) = \exp(x)/(1 + \exp(x))$.

3

Inexact Difference-of-Convex Algorithm

DC Constrained Optimization Problem

$$\begin{aligned} & \min_{\mathbf{w}} f_0(\mathbf{w}) := f_0^+(\mathbf{w}) - f_0^-(\mathbf{w}) \\ & \text{s.t. } f_i(\mathbf{w}) := f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}) \leq 0, \quad i \in [n] \end{aligned}$$

- f_i^+ and f_i^- for $i = 0, \dots, n$ are real-valued, convex functions.

Convex subproblem

$$\begin{aligned} & \hat{\mathbf{w}}^{(k+1)} \in \arg \min_{\mathbf{v}} g_0(\mathbf{v}) := f_0^+(\mathbf{v}) - \hat{f}_0^-(\mathbf{v}; \mathbf{w}^{(k)}) \\ & \text{s.t. } g_i(\mathbf{v}) := f_i^+(\mathbf{v}) - \hat{f}_i^-(\mathbf{v}; \mathbf{w}^{(k)}) \leq 0, \quad i \in [m], \end{aligned}$$

where $\hat{f}_i^-(\mathbf{v}; \mathbf{w}) := f_i^-(\mathbf{w}) + (\mathbf{f}_i^-)^T(\mathbf{v} - \mathbf{w}) \leq f_i^-(\mathbf{v})$ and $\mathbf{f}_i^- \in \partial f_i^-(\mathbf{w})$.

Algorithm Inexact Difference-of-Convex Algorithm (IDCA)

1. **Input:** A feasible $\mathbf{w}^{(0)}$, the number of iterations K , precision ϵ_k , $k = 0, 1, \dots, K-1$.
2. **for** $k = 0, 1, \dots, K-1$ **do**
3. Find $\mathbf{w}^{(k+1)}$ as an ϵ_k -optimal solution of the subproblem, i.e.

$$g_0(\mathbf{w}^{(k+1)}) - g_0(\hat{\mathbf{w}}^{(k+1)}) \leq \epsilon_k, \quad g_i(\mathbf{w}^{(k+1)}) \leq \epsilon_k, \quad i \in [m]$$

4. **end for**

Algorithm Switching Subgradient (SSG) Method: $\text{SSG}(\mathbf{v}^{(0)}, \epsilon, T)$

1. **Input:** Infeasibility tolerance $\epsilon > 0$, initial solution $\mathbf{v}^{(0)}$ with $g_i(\mathbf{v}^{(0)}) \leq \epsilon$ for $i \in [m]$, and number of iterations T .
2. $\mathcal{T} \leftarrow \emptyset$
3. **for** $t = 0$ to $T-1$ **do**
4. **if** $g(\mathbf{v}^{(t)}) \leq \epsilon$ **then**
5. Compute $\mathbf{g}_0^{(t)} \in \partial g_0(\mathbf{v}^{(t)})$
6. $\mathbf{v}^{(t+1)} \leftarrow \mathbf{v}^{(t)} - \frac{\epsilon}{\|\mathbf{g}_0^{(t)}\|^2} \mathbf{g}_0^{(t)}$
7. $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$
8. **else**
9. Compute $\mathbf{g}^{(t)} \in \partial g(\mathbf{v}^{(t)})$
10. $\mathbf{v}^{(t+1)} \leftarrow \mathbf{v}^{(t)} - \frac{g(\mathbf{v}^{(t)})}{\|\mathbf{g}^{(t)}\|^2} \mathbf{g}^{(t)}$
11. **end if**
12. **end for**
13. **Output:** $\mathbf{v}^{(\tau)}$ with $\tau = \arg \min_{t \in \mathcal{T}} g_0(\mathbf{v}^{(t)})$

4

Complexity Theorem

Definition

A point $\mathbf{w}$ is an ϵ -KKT point if there exist $\mathbf{f}_i^+ \in \partial f_i^+(\mathbf{w})$, $\mathbf{f}_i^- \in \partial f_i^-(\mathbf{w})$, and $\lambda_i \geq 0$ for $i \in [m]$ satisfying

$$\|\mathbf{f}_0^+ - \mathbf{f}_0^- + \sum_{i=1}^m \lambda_i (\mathbf{f}_i^+ - \mathbf{f}_i^-)\| \leq \epsilon, \quad (1)$$

$$f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}) \leq \epsilon, \quad i \in [m], \quad (2)$$

$$|\lambda_i (f_i^+(\mathbf{w}) - f_i^-(\mathbf{w}))| \leq \epsilon, \quad i \in [m]. \quad (3)$$

Definition

A point $\mathbf{w}$ is a *nearly ϵ -KKT point* if there exist $\hat{\mathbf{w}}$, $\mathbf{f}_i^+ \in \partial f_i^+(\hat{\mathbf{w}})$, $\mathbf{f}_i^- \in \partial f_i^-(\mathbf{w})$, and $\lambda_i \geq 0$ for $i \in [m]$ satisfying $\|\hat{\mathbf{w}} - \mathbf{w}\| \leq \epsilon$, (1), (2) and (3).

Assumption

The following statements hold.

- A. Functions in $\{f_i^+\}_{i=0}^m$ and $\{f_i^-\}_{i=0}^m$ are real-valued and μ -strongly convex on $\mathbb{R}^n$.
- B. There exists constant $\nu > 0$ such that, if $f_i(\mathbf{w}) \leq 0$ for $i \in [m]$, there exists $\mathbf{v} \in \mathcal{W}$ such that $f_i^+(\mathbf{v}) - \hat{f}_i^-(\mathbf{v}; \mathbf{w}) \leq -\nu, \forall i \in [m]$.
- C. $f_{\text{lb}} := \min_{\mathbf{w}} f_0(\mathbf{w}) \geq -\infty$.
- D. $f_i^+(\mathbf{w})$ and $f_i^-(\mathbf{w})$ are M -Lipchitz continuous for any $i \in [m]$.

Proposition (Theorem 3.2.3 in Nesterov (2018))

If $f_i(\mathbf{w}^{(k)}) \leq 0$ for $i \in [m]$ and $\mathbf{w}^{(k+1)} = \text{SSG}(\mathbf{w}^{(k)}, \epsilon_k, T_k)$ with $T_k \geq \frac{M^2}{\epsilon_k \mu} \ln \left(\frac{\mu \|\mathbf{w}^{(k)} - \hat{\mathbf{w}}^{(k+1)}\|^2}{\epsilon_k} \right)$, it holds that

$$g_0(\mathbf{w}^{(k+1)}) - g_0(\hat{\mathbf{w}}^{(k+1)}) \leq \epsilon_k, \quad g_i(\mathbf{w}^{(k+1)}) \leq \epsilon_k, \quad i \in [m].$$

Theorem

Suppose $K = \left\lceil \max \left\{ 1, 4M^2, \frac{8M^4}{\mu\nu} \right\} \frac{f_0(\mathbf{w}^{(0)}) - f_{\text{lb}}}{\mu\epsilon^2} \right\rceil$, $\epsilon_k = \frac{\mu}{8} \min \left\{ 1, \frac{1}{4M^2}, \frac{8M^4}{\mu\nu} \right\} \cdot \epsilon^2$, and set $\mathbf{w}^{(k+1)} = \text{SSG}(\mathbf{w}^{(k)}, \epsilon_k, T_k)$ along with $T_k = \frac{M^2}{\epsilon_k \mu} \ln \left(\frac{\mu \|\mathbf{w}^{(k)} - \hat{\mathbf{w}}^{(k+1)}\|^2}{\epsilon_k} \right)$. IDCA finds a nearly ϵ -KKT point with an oracle complexity of $\tilde{O}(\epsilon^{-4})$.

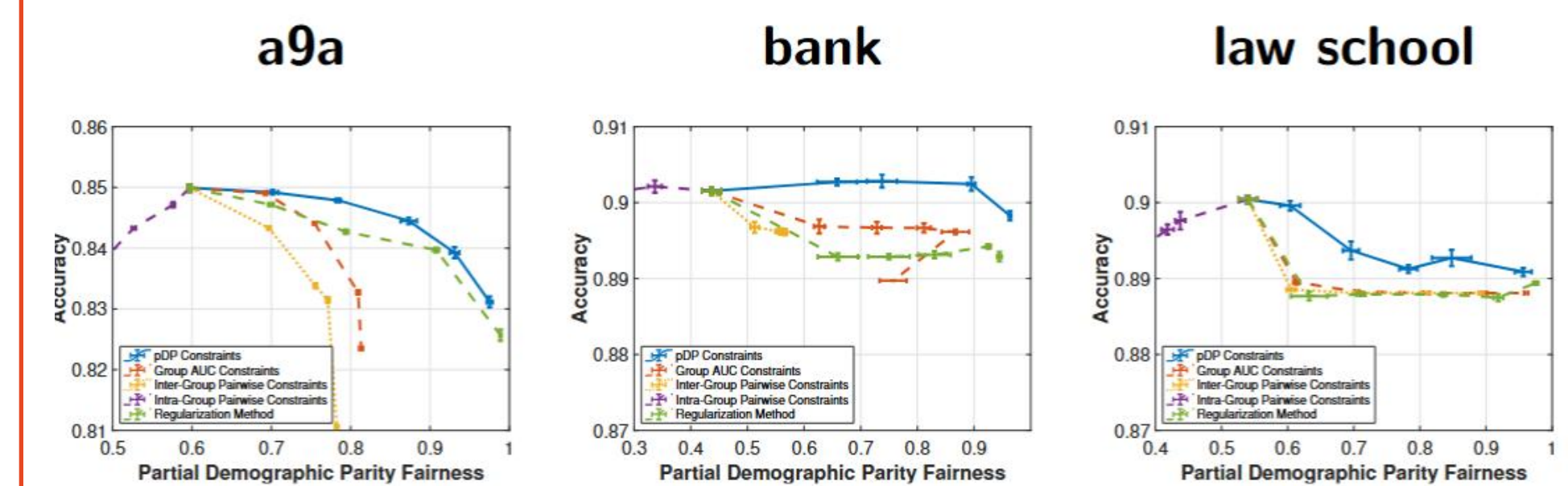
4

Numerical Experiment

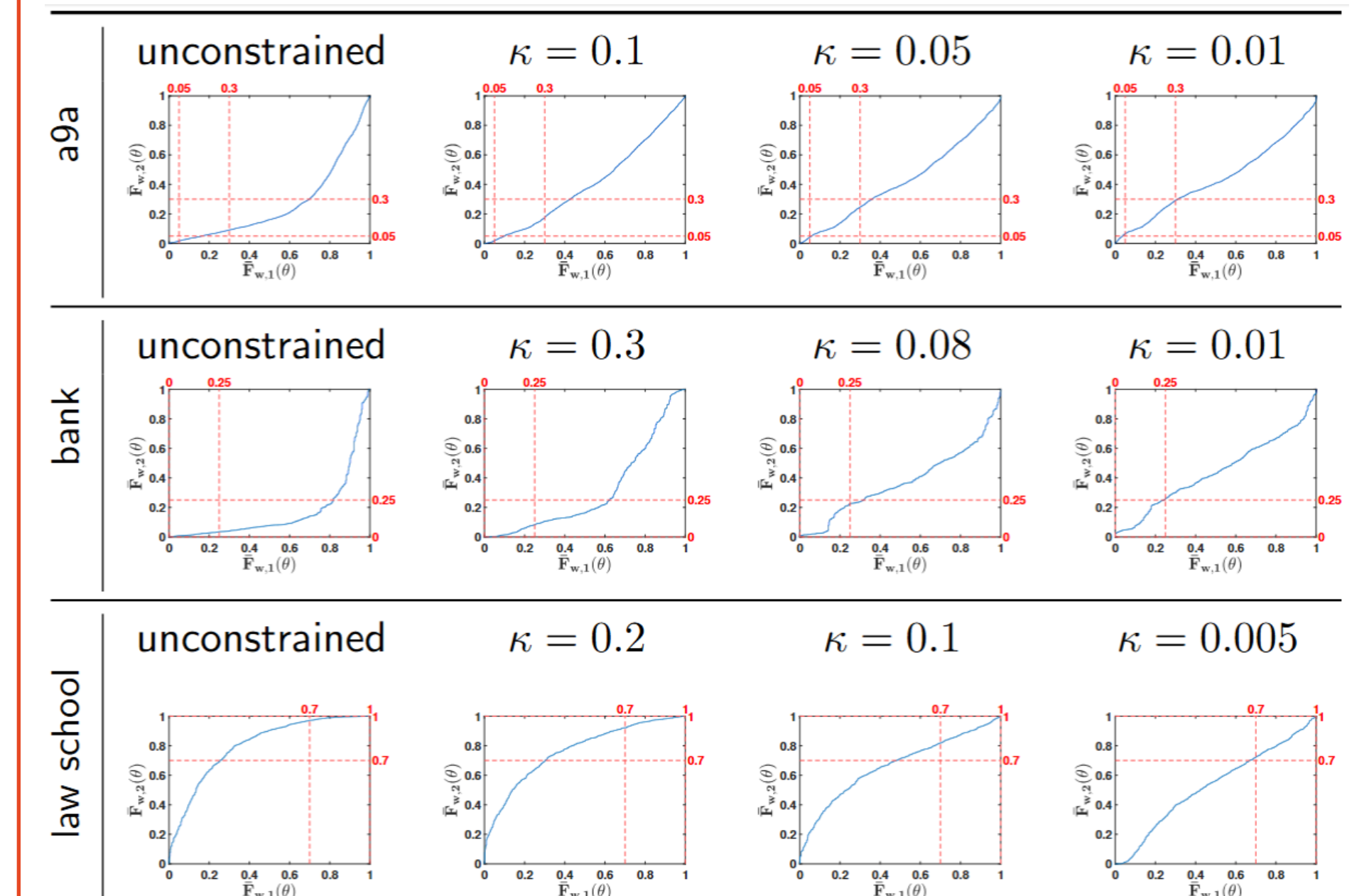
Dataset	#Instances	#Attributes	Class Label	Sensitive Attribute
a9a	48,842	123	Income	Gender
bank	41,188	54	Subscription	Age
law school	20,798	12	Passing exam	Race

- Group AUC Fairness (Yao et al, 2023, Vogel et al, 2021)
- Inter-Group Pairwise Fairness (Kallus et al, 2019, Beutel et al, 2019)
- Intra-Group Pairwise Fairness (Beutel et al, 2019).
- Post-processing method (Kallus et al, 2019)
- Regularization method: $\min_{\mathbf{w}} f_0(\mathbf{w}) + \lambda \max_{i \in [m]} [f_i(\mathbf{w})]_+$

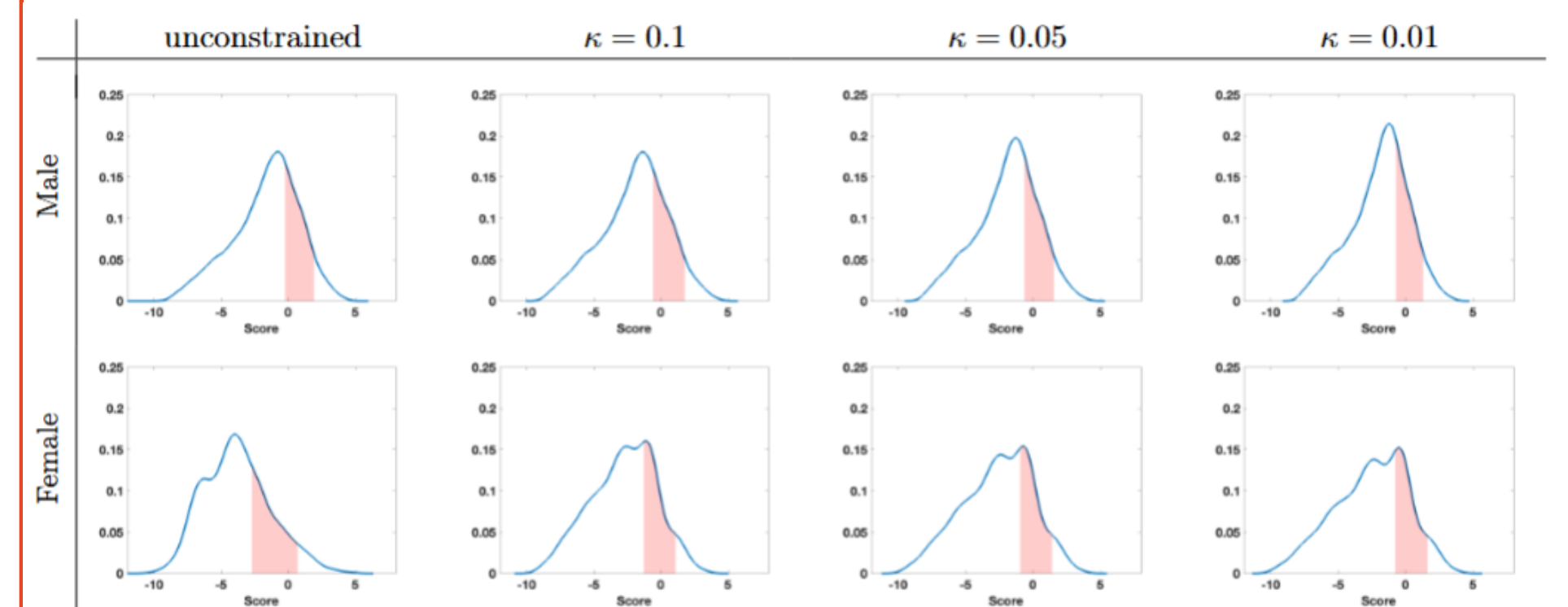
Pareto frontiers for classification accuracy vs pDP



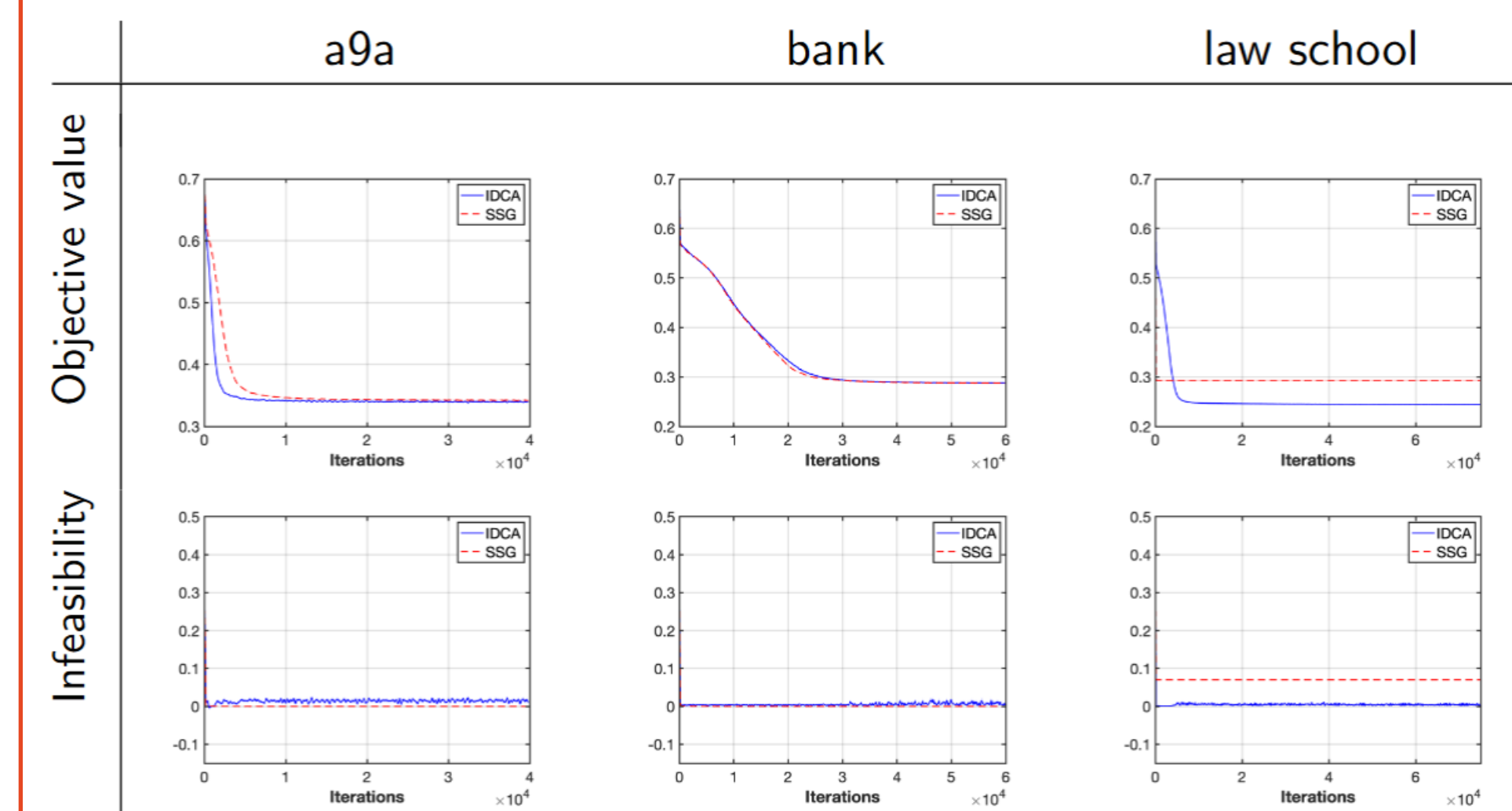
Predicted Positive Rates vs Threshold θ



Distributions of predicted scores of different sensitive groups on a9a



Optimization Performance of ICDA



5

References