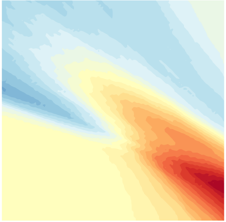
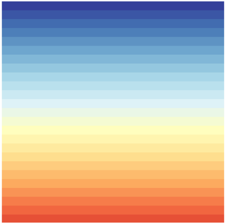
# Too Relaxed to Be Fair

## ICML 2020

**Michael Lohaus**, Michaël Perrot, Ulrike von Luxburg

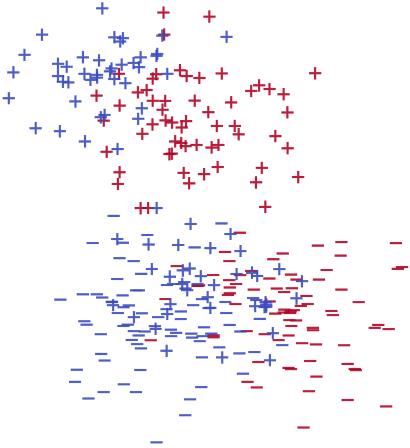




# The Setting: Classification with Fairness

**Given:**

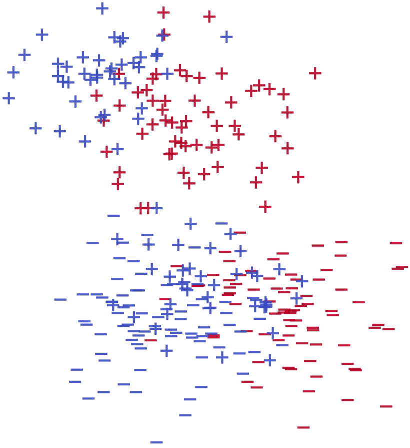
- feature space $\mathcal{X}$,
- class labels $\mathcal{Y} = \{+, -\}$,
- sensitive attributes $\mathcal{S}$={**blue**, **red**},
- a **fairness notion**.



# The Setting: Classification with Fairness

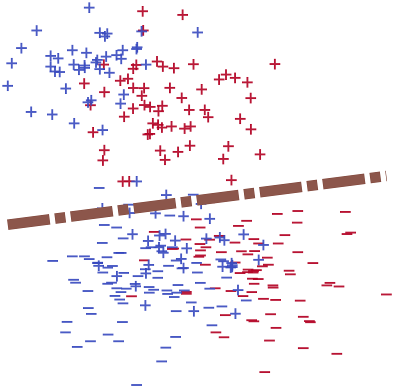
**Given:**

- feature space $\mathcal{X}$,
- class labels $\mathcal{Y} = \{+, -\}$,
- sensitive attributes $\mathcal{S}$={**blue**, **red**},
- a **fairness notion**.



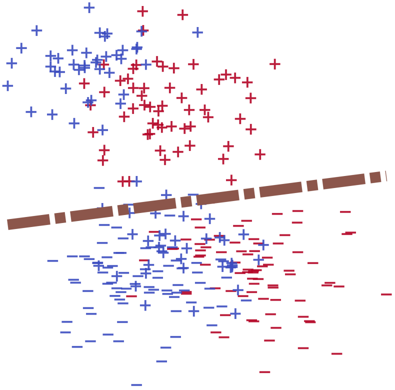
**Goal:** A classifier $h : \mathcal{X} \to \mathcal{Y}$ that is **accurate** while **being fair** with respect to the sensitive attribute.

# The Setting: Classification with Fairness



$$\mathbb{P}\left[f(x)=1 | s=\textbf{red}\right] = 0.50$$
$$\mathbb{P}\left[f(x)=1 | s=\textbf{blue}\right] = 0.31$$

# The Setting: Classification with Fairness



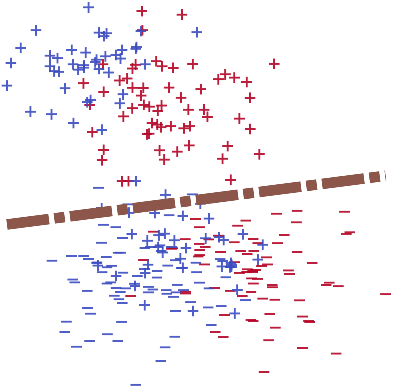
$$\mathbb{P}\left[f(x)=1 | s=\textbf{red}\right] = 0.50$$

$$\mathbb{P}\left[f(x)=1 | s=\textbf{blue}\right] = 0.31$$

**Difference of Demographic Parity:**

$$\mathrm{DDP}(f) = 0.50 - 0.31 = 0.19$$

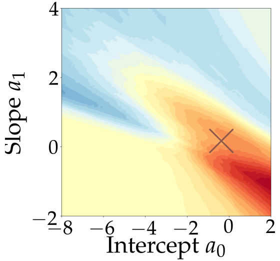
# The Setting: Classification with Fairness



$$\mathbb{P}\left[f(x)=1|s=\textbf{red}\right] = 0.50$$
$$\mathbb{P}\left[f(x)=1|s=\textbf{blue}\right] = 0.31$$

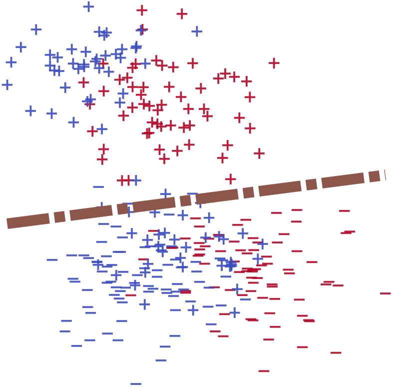
**Difference of Demographic Parity:**

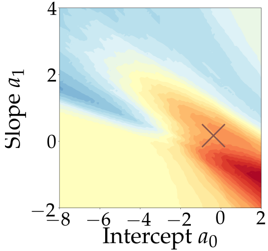
$$\text{DDP}(f) = 0.50 - 0.31 = 0.19$$



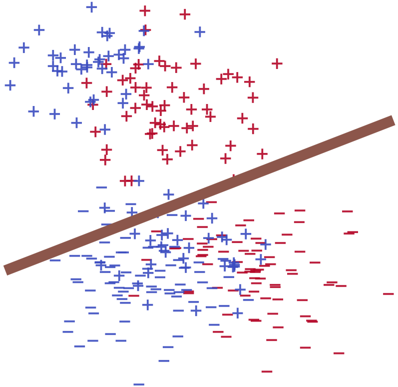
# The Problem: How to achieve fairness?

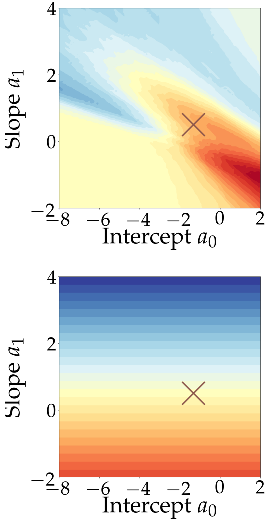
Fairness constraint is **non-convex.**



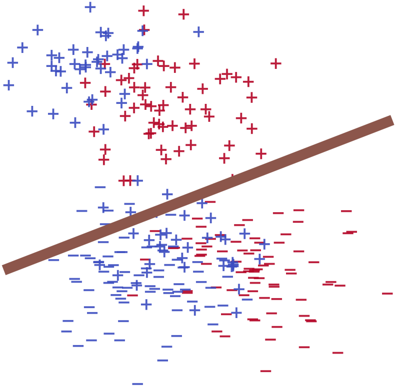


## Possibly a convex relaxations?





**Possibly a convex relaxation?**



$\mathrm{LR}_{\mathrm{DDP}}(g) = 0.$

**But:**

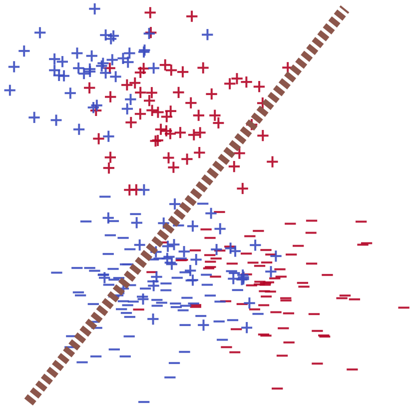
$$\mathbb{P}[g(x) = 1 | s = \textbf{red}] = 0.49$$
$$\mathbb{P}[g(x) = 1 | s = \textbf{blue}] = 0.32$$

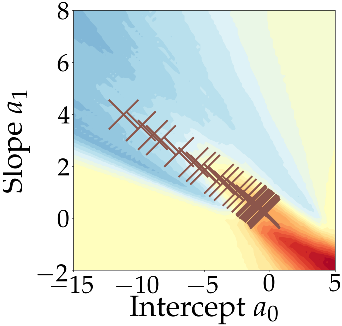
**Difference of Demographic Parity:**

$$\mathrm{DDP}(g) = 0.17$$

# Our Solution: SearchFair

Keep your relaxation and **SearchFair** can guarantee a fair classifier.





# Recent Approaches: Optimization with Fairness Constraint

$$f = \underset{\substack{f \in \mathcal{F} \\ f \text{ is fair}}}{\arg\min}\, L(f) + \beta\Omega(f)\,,$$

with

- convex risk $L(f)$
- a convex regularization $\Omega(f)$
- a fairness constraint.

# Example: Demographic Parity

Measure fairness with **Difference of Demographic Parity**:

$$\mathrm{DDP}(f) = \mathbb{P}\left[f(x)=1 | s=\textbf{red}\right] - \mathbb{P}\left[f(x)=1 | s=\textbf{blue}\right].$$

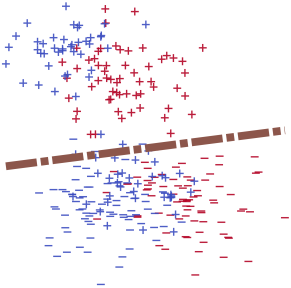
Fairness constraint: $|\mathrm{DDP}(f)| \leq \tau$.

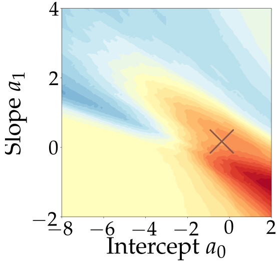
# Example: Demographic Parity

Measure fairness with **Difference of Demographic Parity**:

$$\mathrm{DDP}(f) = \mathbb{P}\left[f(x)=1 | s=\textbf{red}\right] - \mathbb{P}\left[f(x)=1 | s=\textbf{blue}\right].$$

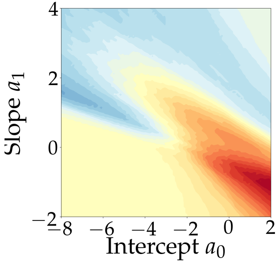
Fairness constraint: $|\mathrm{DDP}(f)| \leq \tau$.



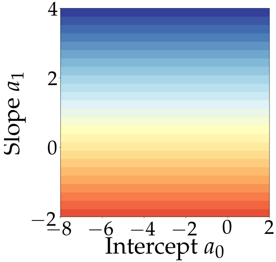



**Difficulty:** Learning a fair classifier with non-convex constraint.

# Linear Relaxation [Donini et al., 2018, Zafar et al., 2017b]

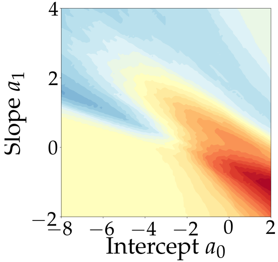



DDP

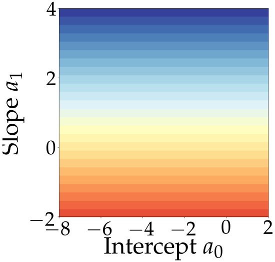



Linear Relaxation $\mathrm{LR}_{\mathrm{DDP}}(f)$

# Linear Relaxation [Donini et al., 2018, Zafar et al., 2017b]




DDP




Linear Relaxation $\mathrm{LR}_{\mathrm{DDP}}(f)$

New convex constraint:

$$|\mathrm{LR}_{\mathrm{DDP}}(f)| \leq \tau.$$

**But:** Fairness is not well approximated.

# Example: Adult dataset

- Label: income $\geq 50,000\$$
- Sensitive attribute: sex

| | | DDP | Linear Relaxation |
|---|---|---|---|
| Unconstrained | | | |
| | linear kernel | 0.25 | 0.86 |
| | RBF kernel | 0.21 | 0.52 |

# Example: Adult dataset

- Label: income $\geq 50,000\$$
- Sensitive attribute: sex

| | | DDP | Linear Relaxation |
|---|---|---|---|
| Unconstrained | | | |
| | linear kernel | 0.25 | 0.86 |
| | RBF kernel | 0.21 | 0.52 |
| Constrained | | | |
| | linear kernel | 0.00 | 0.00 |
| | RBF kernel | 0.20 | 0.02 |

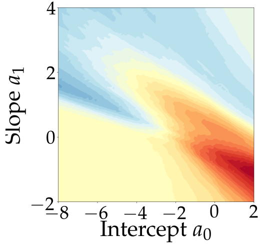
# Example: Adult dataset

- Label: income $\geq 50,000$$
- Sensitive attribute: sex

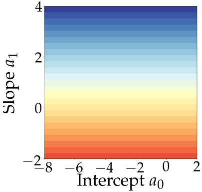
| | | DDP | Linear Relaxation |
|---|---|---|---|
| Unconstrained | | | |
| | linear kernel | 0.25 | 0.86 |
| | RBF kernel | 0.21 | 0.52 |
| Constrained | | | |
| | linear kernel | 0.00 | 0.00 |
| | RBF kernel | 0.20 | 0.02 |

Linear Relaxation is not reliable.

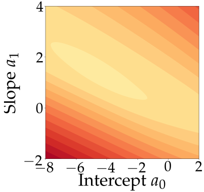
# Other Relaxations [Wu et al., 2019, Zafar et al., 2017a]



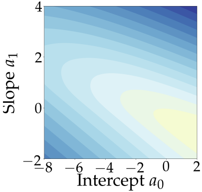

DDP



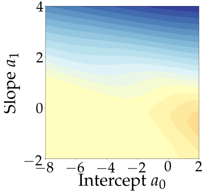

Linear




Wu - Upper




Wu - Lower




Convex-Concave

# Recent Approaches: Optimization with Fairness Constraint

$$f = \underset{\substack{f \in \mathcal{F} \\ |\mathrm{LR}_{\mathrm{DDP}}(f)| \leq \tau}}{\arg\min} \; L(f) + \beta\Omega(f)\,,$$

- $L(f)$ is a convex risk,
- $\Omega(f)$ is a convex regularization term,
- $\beta$ is a trade-off parameter,

# Our Approach: Unconstrained Optimization Problem

$$f(\lambda) = \arg\min_{f \in \mathcal{F}} L(f) + \lambda \mathrm{R}_{\mathrm{DDP}}(f) + \beta\Omega(f)\,,$$

- $L(f)$ is a convex risk,
- $\Omega(f)$ is a strongly convex regularization term,
- $\lambda$ and $\beta$ are trade-off parameters,
- $\mathrm{R}_{\mathrm{DDP}}(f)$ is a convex fairness relaxation.

# Our Approach: Unconstrained Optimization Problem

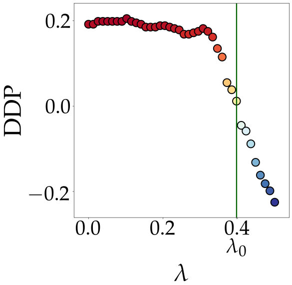
$$f(\lambda) = \arg\min_{f \in \mathcal{F}} L(f) + \lambda \mathrm{R}_{\mathrm{DDP}}(f) + \beta\Omega(f)\,,$$

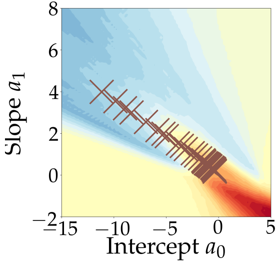
- $L(f)$ is a convex risk,
- $\Omega(f)$ is a strongly convex regularization term,
- $\lambda$ and $\beta$ are trade-off parameters,
- $\mathrm{R}_{\mathrm{DDP}}(f)$ is a convex fairness relaxation.

**Theorem**

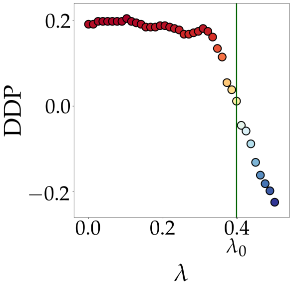
*The function $\lambda \mapsto DDP(f(\lambda))$ is continuous!*

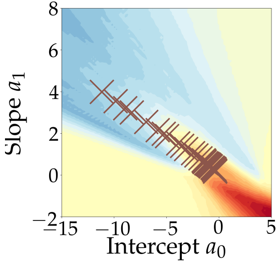
# From theory to algorithm



# From theory to algorithm



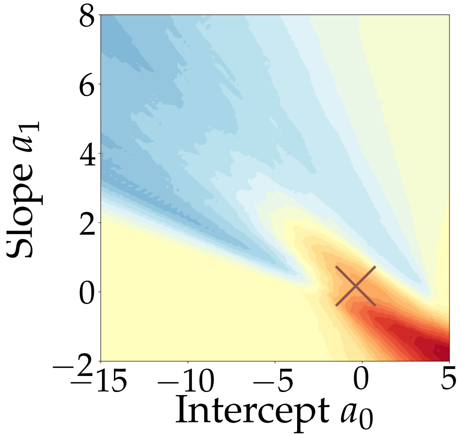




## Corollary

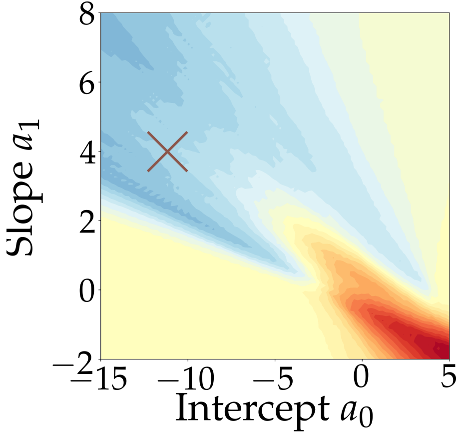
***(i)*** *If there exists* $\lambda_+$ *such that* $DDP(f(\lambda_+)) > 0$,
***(ii)*** *and if there exists* $\lambda_-$ *such that* $DDP(f(\lambda_-)) < 0$,
*then there exists one value* $\lambda_0$ *such that*

$$DDP(f(\lambda_0)) = 0.$$

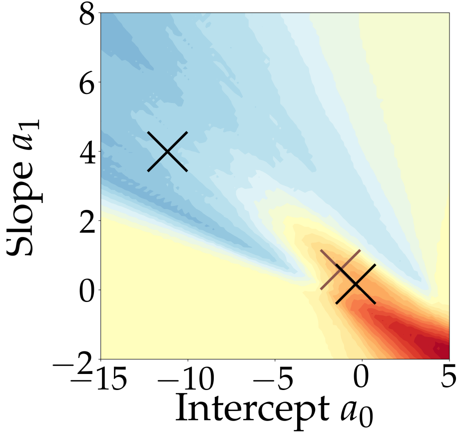
# SearchFair: Using Binary Search



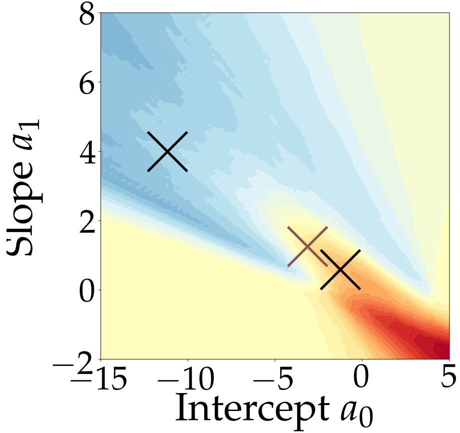
# SearchFair: Using Binary Search



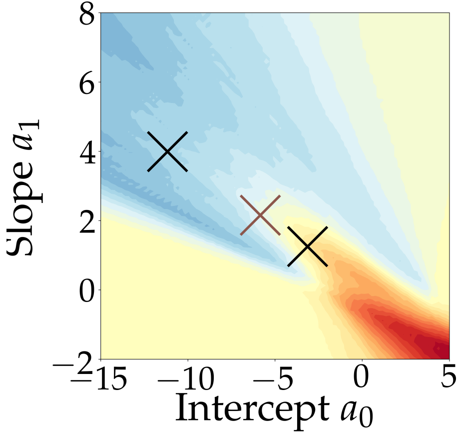
# SearchFair: Using Binary Search



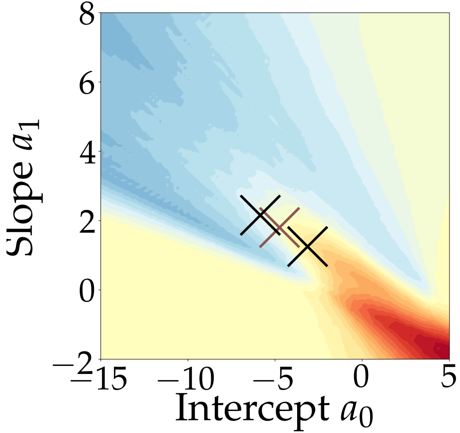
# SearchFair: Using Binary Search



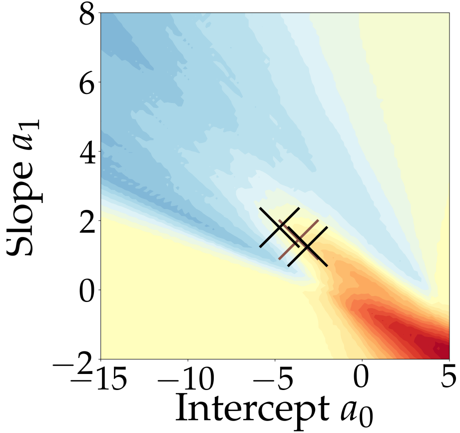
# SearchFair: Using Binary Search



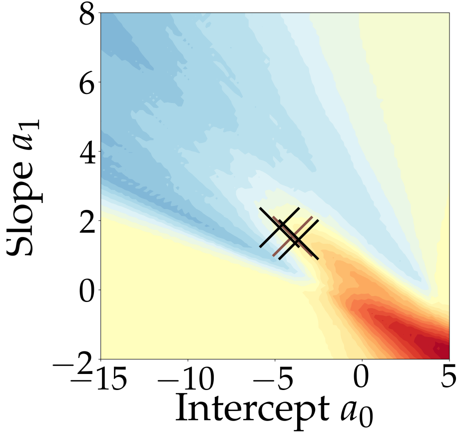
# SearchFair: Using Binary Search



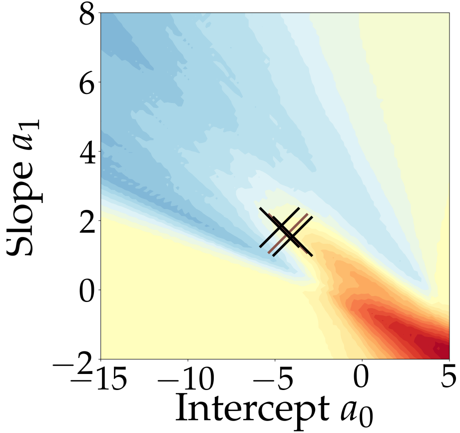
# SearchFair: Using Binary Search



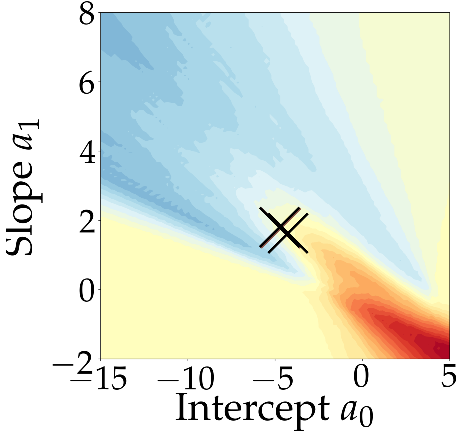
# SearchFair: Using Binary Search



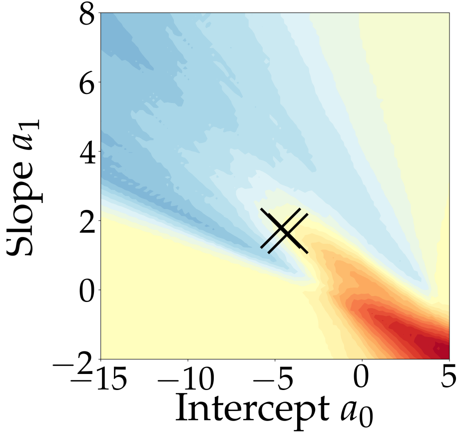
# SearchFair: Using Binary Search



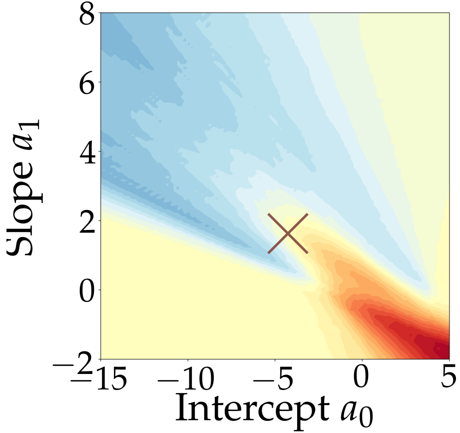
# SearchFair: Using Binary Search



# SearchFair: Using Binary Search

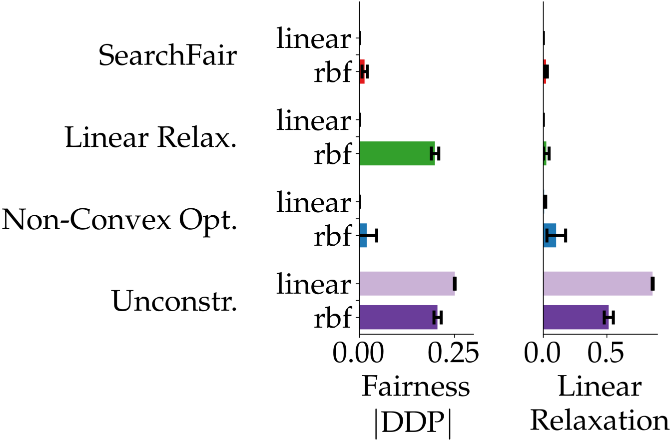


# SearchFair: Using Binary Search



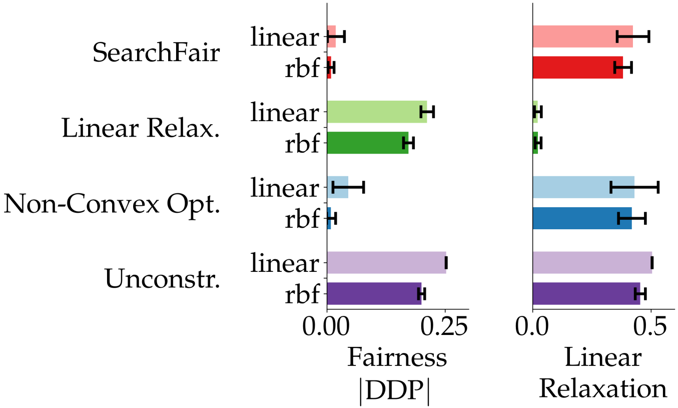
# Results: Adult dataset

- Label: income $\geq 50,000\$$
- Sensitive attribute: sex



# Results: CelebA dataset

- Label: Smiling
- Sensitive attribute: sex



# Example: Equality of Opportunity

**Difference of Equality of Opportunity**:

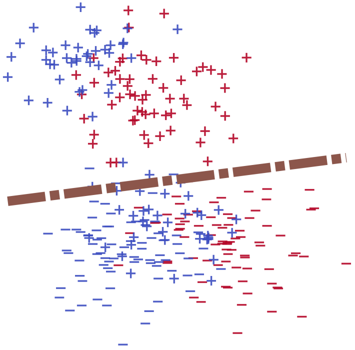
$$\mathrm{DEO}(f) = \mathbb{P}\left[f(x)=1 | s=\textbf{red}, y=+1\right] - \mathbb{P}\left[f(x)=1 | s=\textbf{blue}, y=+1\right].$$

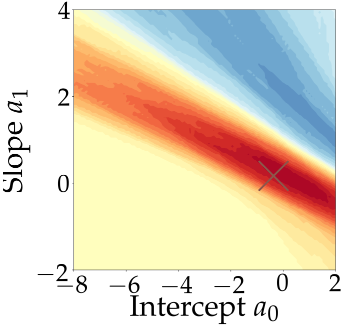
# Example: Equality of Opportunity

**Difference of Equality of Opportunity**:

$$\mathrm{DEO}(f) = \mathbb{P}\left[f(x)=1 | s=\textbf{red}, y=+1\right] - \mathbb{P}\left[f(x)=1 | s=\textbf{blue}, y=+1\right].$$

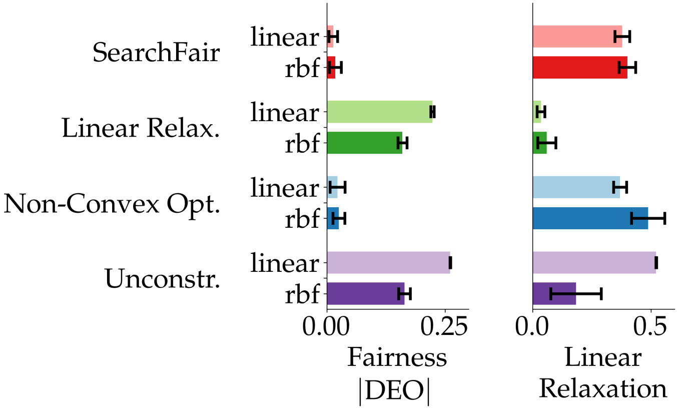
Fairness constraint: $|\mathrm{DEO}(f)| \leq \tau$.





# Results: CelebA dataset

- Label: Smiling
- Sensitive attribute: sex



# Conclusion: Too Relaxed to Be Fair

We found:

- **Convex relaxations cannot reliably learn fair classifiers.**

We propose SearchFair.

- **SearchFair works with many existing relaxations.**
- **SearchFair guarantees a fair solution.**

Try it out!