Fair and Accurate Regression

Anna Deza[1], Andrés Gómez[2], Alper Atamtürk[1]

[1]Department of Industrial Engineering and Operations Research,
University of California, Berkeley

[2]Department of Industrial and Systems Engineering,
University of Southern California

ICCOPT 2025

Table of Contents

Supervised Learning: Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

$\boldsymbol{x}_i$ n-dimensional feature vector

y_i target label

Train a model f by optimizing loss function over training data

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{m} \mathcal{L}(f(\mathbf{x}_i), y_i).$$

Supervised Learning: Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

$\boldsymbol{x}_i$ n-dimensional feature vector

y_i target label

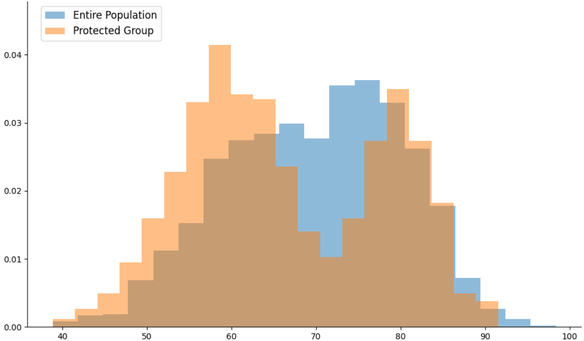
Train a model f by optimizing loss function over training data

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{m} \mathcal{L}(f(\mathbf{x}_i), y_i).$$

$\rightarrow$ What if training data is *biased*? An accurate model may amplify bias in training data ...

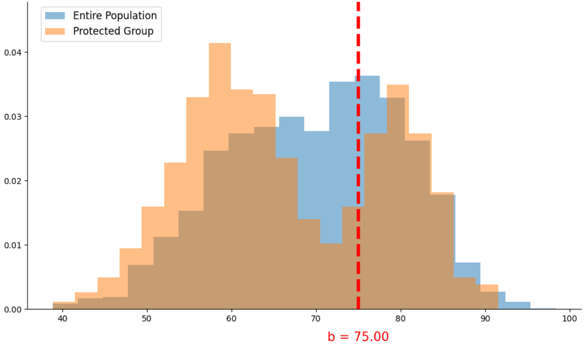
Example: College Admissions

Model predicts scores 0-100 used for college admission process. Fair?



Example: College Admissions

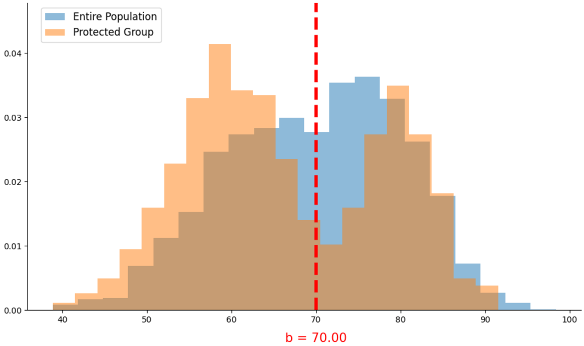
Model predicts scores 0-100 used for college admission process. Fair?



$P(\text{score} > 75) = 0.35$
$P(\text{score} > 75 | \textbf{rural}) = 0.32$

Example: College Admissions

Model predicts scores 0-100 used for college admission process. Fair?

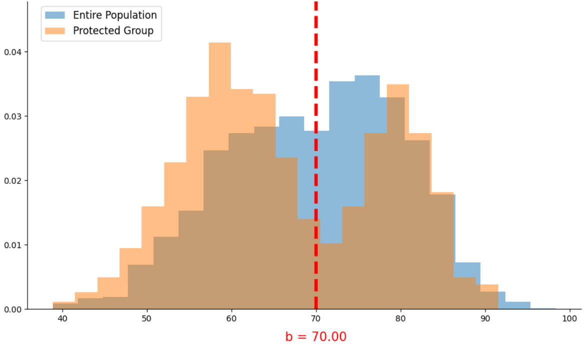


$P(\text{score} > 75) = 0.35$
$P(\text{score} > 75|\textbf{rural}) = \mathbf{0.32}$

$P(\text{score} > 70) = 0.52$
$P(\text{score} > 70|\textbf{rural}) = \mathbf{0.38}$

Example: College Admissions

Model predicts scores 0-100 used for college admission process. Fair?



$P(\text{score} > 75) = 0.35$
$P(\text{score} > 75 | \textbf{rural}) = 0.32$

$P(\text{score} > 70) = 0.52$
$P(\text{score} > 70 | \textbf{rural}) = 0.38$

- Fair Classification: fairness at a single b, vast existing literature
- Fair Regression: fairness at *all* b, far less work $\rightarrow$ **problem we address**

Fairness: Demographic Parity

Suppose population is segmented according to attribute a which we do not believe should influence the prediction $\hat{y} = f(x)$.

Demographic Parity: $\hat{y} \perp a$

Fair Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

$\boldsymbol{x}_i \in \mathbb{R}^n$ feature vector

$y_i \in \mathbb{R}$ target label

$a_i \in \{0, 1\}$ protected class status

Fair Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

$\boldsymbol{x}_i \in \mathbb{R}^n$ feature vector

$y_i \in \mathbb{R}$ target label

$a_i \in \{0, 1\}$ protected class status

Training problem:

$$\min \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i) + \lambda \max_{b \in \mathbb{R}} \left(\hat{P}(\hat{y} > b) - \hat{P}(\hat{y} > b | a = 1) \right)$$

Fair Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

- $\boldsymbol{x}_i \in \mathbb{R}^n$ feature vector
- $y_i \in \mathbb{R}$ target label
- $a_i \in \{0, 1\}$ protected class status

Training problem:

$$\min \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i) + \lambda \max_{b \in \mathbb{R}} \left(\hat{P}(\hat{y} > b) - \hat{P}(\hat{y} > b | a = 1) \right)$$

$$\boxed{\begin{array}{c} P(\hat{y} > b) - P(\hat{y} > b | a = 1) = 0 \\ \forall b \in \mathbb{R} \end{array}} \longrightarrow \boxed{\begin{array}{c} \hat{y} \perp a \\ \text{(resource allocation } \perp \text{ of } a) \end{array}}$$

Fair Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

- $\boldsymbol{x}_i \in \mathbb{R}^n$ feature vector
- $y_i \in \mathbb{R}$ target label
- $a_i \in \{0, 1\}$ protected class status

Training problem:

$$\min \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i) + \lambda \max_{b \in \mathbb{R}} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(\hat{y}_i > b) - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} \mathbb{1}(\hat{y}_i > b) \right)$$

Fair Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

$\boldsymbol{x}_i \in \mathbb{R}^n$ feature vector

$y_i \in \mathbb{R}$ target label

$a_i \in \{0, 1\}$ protected class status

Training problem:

$$\min \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i) + \lambda \max_{b \in \mathbb{R}} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(\hat{y}_i > b) - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} \mathbb{1}(\hat{y}_i > b) \right)$$

Fair Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

$\boldsymbol{x}_i \in \mathbb{R}^n$ feature vector

$y_i \in \mathbb{R}$ target label

$a_i \in \{0, 1\}$ protected class status

Training problem:

$$\min \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i) + \lambda \max_{j \in 1, \ldots \ell} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(\hat{y}_i > b_j) - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} \mathbb{1}(\hat{y}_i > b_j) \right)$$

Discretize $\mathbb{R}$, only consider fairness at ℓ points $b_1 < b_2 < \ldots < b_\ell$

Fair Training Problem

Data: observations $i = 1, \ldots, m$, consisting of:

$\boldsymbol{x}_i \in \mathbb{R}^n$ feature vector

$y_i \in \mathbb{R}$ target label

$a_i \in \{0, 1\}$ protected class status

Training problem:

$$\min \sum_{i=1}^{m} \mathcal{L}_i(\boldsymbol{w}^\top \boldsymbol{x}_i) + \lambda \max_{j \in 1,\ldots\ell} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(\boldsymbol{w}^\top \boldsymbol{x}_i > b_j) - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} \mathbb{1}(\boldsymbol{w}^\top \boldsymbol{x}_i > b_j) \right)$$

Discretize $\mathbb{R}$, only consider fairness at ℓ points $b_1 < b_2 < \ldots < b_\ell$

We will focus on **linear regression**: $\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$, $\mathcal{L}_i(\boldsymbol{w}^\top \boldsymbol{x}_i) = (\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$.

Literature Overview

$$\min \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i) + \lambda \max_{j \in 1,\ldots \ell} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(\hat{y}_i > b_j) - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} \mathbb{1}(\hat{y}_i > b_j) \right)$$

Literature Overview

$$\min \sum_{i=1}^{m} \mathcal{L}(\hat{y}_i, y_i) + \lambda \max_{j \in 1,\ldots \ell} \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(\hat{y}_i > b_j) - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} \mathbb{1}(\hat{y}_i > b_j) \right)$$

Existing methods:

- Convex proxies for fairness: Berk et al. (2017), Do et al. (2022)
- Reduction-based algorthims: Agarwal, Dudik & Wu (2019)
- MIO approach: Ye, Hanasusanto & Xie (2024)

Table of Contents

A First Formulation for Fair Regression

Introduce $z_{ij} \in \{0,1\}$ to model $\mathbb{1}(\boldsymbol{w}^\top \boldsymbol{x}_i > b_j)$

$$
\begin{aligned}
\min \ & \sum_{i=1}^{m} \mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \lambda \max_{j \in 1,\dots\ell} \left(\frac{1}{m} \sum_{i=1}^{m} z_{ij} - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} z_{ij} \right) \\
\text{s.t.} \ & (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j) z_{ij} \geq 0 && j \in [\ell],\ i \in [m] \\
& (\boldsymbol{w}^\top \boldsymbol{x}_j - b_j)(1 - z_{ij}) \leq 0 && j \in [\ell],\ i \in [m] \\
& \boldsymbol{z} \in \{0,1\}^{m \times \ell}.
\end{aligned}
$$

A First Formulation for Fair Regression

Introduce $z_{ij} \in \{0,1\}$ to model $\mathbb{1}(\boldsymbol{w}^\top \boldsymbol{x}_i > b_j)$

$$\begin{aligned}
\min \ & \sum_{i=1}^{m} \mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \lambda \max_{j \in 1,\ldots \ell} \left(\frac{1}{m} \sum_{i=1}^{m} z_{ij} - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} z_{ij} \right) \\
\text{s.t.} \ & (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j) z_{ij} \geq 0 && j \in [\ell],\ i \in [m] && z_{ij} = 1 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i \geq b_j \\
& (\boldsymbol{w}^\top \boldsymbol{x}_j - b_j)(1 - z_{ij}) \leq 0 && j \in [\ell],\ i \in [m] && z_{ij} = 0 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i \leq b_j \\
& \boldsymbol{z} \in \{0,1\}^{m \times \ell}.
\end{aligned}$$

A First Formulation for Fair Regression

Relax $z_{ij} \in \{0,1\}$ to model $\mathbb{1}(\boldsymbol{w}^\top \boldsymbol{x}_i > b_j)$

$$
\begin{aligned}
\min \ & \sum_{i=1}^{m} \mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i, y_i) + \lambda \max_{j \in 1,\ldots \ell} \left(\frac{1}{m} \sum_{i=1}^{m} z_{ij} - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} z_{ij} \right) \\
\text{s.t. } & (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j) z_{ij} \geq 0 \qquad j \in [\ell],\ i \in [m] \qquad z_{ij} = 1 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i \geq b_j \\
& (\boldsymbol{w}^\top \boldsymbol{x}_j - b_j)(1 - z_{ij}) \leq 0 \qquad j \in [\ell],\ i \in [m] \qquad z_{ij} = 0 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i \leq b_j \\
& \boldsymbol{z} \in [0,1]^{m \times \ell}.
\end{aligned}
$$

Uninformative relaxation:

- Continuous relaxation solution is vanilla regression with no fairness
- Why? The closure of the convex hull has no links between $\boldsymbol{w}$ and $\boldsymbol{z}$

Reformulation: Exploiting Objective

$$
\begin{aligned}
\min \sum_{i=1}^{m} s_i + \lambda \max_{j \in 1,\ldots,\ell} & \left(\frac{1}{m} \sum_{i=1}^{m} z_{ij} - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} z_{ij} \right) \\
\text{s.t.}\ & \mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i) \leq s_i, && i \in [m] \\
& (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j) z_{ij} \geq 0, && j \in [\ell],\ i \in [m] \\
& (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j)(1 - z_{ij}) \leq 0, && j \in [\ell],\ i \in [m] \\
& z_{ij} \in \{0, 1\}, && j \in [\ell],\ i \in [m]
\end{aligned}
$$

Reformulation: Exploiting Objective

$$\min \sum_{i=1}^{m} s_i + \lambda \max_{j \in 1,\dots,\ell} \left(\frac{1}{m} \sum_{i=1}^{m} z_{ij} - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} z_{ij} \right)$$

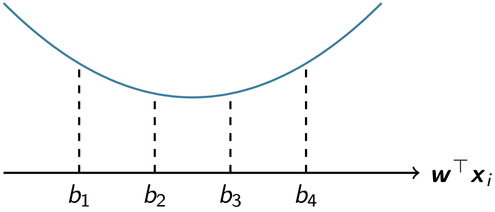
$$\begin{aligned}
\text{s.t.}\quad & \mathcal{L}(\boldsymbol{w}^\top \boldsymbol{x}_i) \le s_i, && && i \in [m] \\
& (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j) z_{ij} \ge 0, && j \in [\ell], && i \in [m] \\
& (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j)(1 - z_{ij}) \le 0, && j \in [\ell], && i \in [m] \\
& z_{ij} \in \{0, 1\}, && j \in [\ell], && i \in [m]
\end{aligned}$$

X_i: single observation ($m = 1$) fair regression

→ Obtain strong reformulation by deriving a **compact extended formulation** for $\mathrm{cl}\ \mathrm{conv}(X_i)$

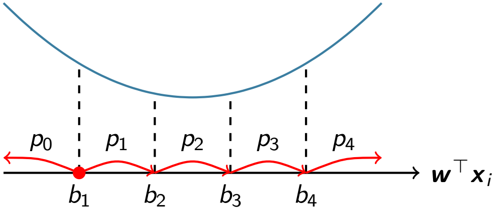
X_i: Epigraph of loss and indicators of prediction i

$$
\begin{aligned}
X_i = \{(\boldsymbol{w}, \boldsymbol{z}, s) \in \mathbb{R}^n \times \{0,1\}^\ell \times \mathbb{R} :\ & \\
& \mathcal{L}_i(\boldsymbol{w}^\top \boldsymbol{x}_i) \leq s \\
& (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j) z_j \geq 0 \qquad j \in [\ell] \\
& (\boldsymbol{w}^\top \boldsymbol{x}_i - b_j)(1 - z_j) \geq 0 \quad j \in [\ell]\}
\end{aligned}
$$



X_i: Epigraph of loss and indicators of prediction i

$$
\begin{aligned}
X_i = &\{(\boldsymbol{w}, \boldsymbol{z}, s) \in \mathbb{R}^n \times \{0,1\}^\ell \times \mathbb{R} : \\
&\mathcal{L}_i(\boldsymbol{w}^\top \boldsymbol{x}_i) \leq s \\
&(\boldsymbol{w}^\top \boldsymbol{x}_i - b_j) z_j \geq 0 \qquad j \in [\ell] \\
&(\boldsymbol{w}^\top \boldsymbol{x}_i - b_j)(1 - z_j) \geq 0 \quad j \in [\ell]\}
\end{aligned}
$$

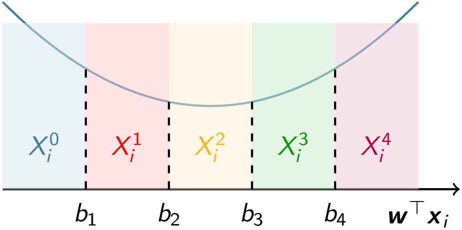


Extended set: $X_i = \{(\boldsymbol{w}, \boldsymbol{z}, s) \in \mathbb{R}^n \times \{0,1\}^\ell \times \mathbb{R} : \exists (p_0, \boldsymbol{p}) \in \mathbb{R}_+^{\ell+1} :$

$$
\begin{aligned}
&\boldsymbol{w}^\top \boldsymbol{x}_i = b_1 + \sum_{j=1}^{\ell} p_j - p_0 \\
&(b_{j+1} - b_j) z_{j+1} \leq p_j \leq (b_{j+1} - b_j) z_j, \; j \in [\ell - 1] \\
&p_0 z_1 = 0, \, p_\ell (1 - z_\ell) = 0 \\
&\mathcal{L}_i(\boldsymbol{w}^\top \boldsymbol{x}_i) \leq s\}
\end{aligned}
$$

cl conv(X_i)

Write as disjunction: $X_i = \bigcup_{j=0}^{\ell} X_i^j$



$$X_i =$$

$$\Big\{ (\boldsymbol{w}, \boldsymbol{z}, s) \in \mathbb{R}^n \times \{0,1\}^\ell \times \mathbb{R} : \exists (p_0, \boldsymbol{p}) \in \mathbb{R}_+^{\ell+1} :$$

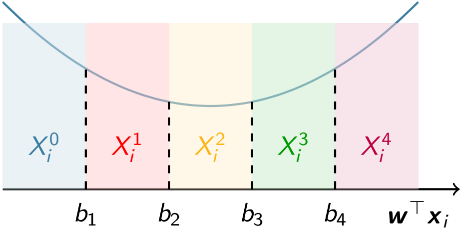
$$\boldsymbol{w}^\top \boldsymbol{x}_i = b_1 + \sum_{j=1}^{\ell} p_j - p_0$$

$$(b_{j+1} - b_j) z_{j+1} \le p_j \le (b_{j+1} - b_j) z_j, \; j \in [\ell - 1]$$

$$p_0 z_1 = 0, \; p_\ell (1 - z_\ell) = 0, \; \mathcal{L}_i(\boldsymbol{w}^\top \boldsymbol{x}_i) \le s \Big\}$$

cl conv(X_i)

Write as disjunction: $X_i = \bigcup_{j=0}^{\ell} X_i^j$



cl conv(X_i) =

$$\Bigg\{(\boldsymbol{w}, \boldsymbol{z}, s) \in \mathbb{R}^n \times [0,1]^\ell \times \mathbb{R} : \exists (p_0, \boldsymbol{p}) \in \mathbb{R}_+^{\ell+1} :$$

$$\boldsymbol{w}^\top \boldsymbol{x}_i = b_1 + \sum_{j=1}^{\ell} p_j - p_0$$

$$(b_{j+1} - b_j) z_{j+1} \leq p_j \leq (b_{j+1} - b_j) z_j, \; j \in [\ell - 1]$$

$$(1 - z_1)\mathcal{L}\left(b_1 - \frac{p_0}{1 - z_1}\right) + \sum_{j=1}^{\ell-1} (z_j - z_{j+1})\mathcal{L}\left(b_i + \frac{p_i - z_{i+1}(b_{i+1} - b_i)}{z_i - z_{i+1}}\right) + z_\ell \mathcal{L}\left(b_\ell + \frac{p_\ell}{z_\ell}\right) \leq s \Bigg\}$$

cl conv(X_i)

Proposition

An extended formulation for cl conv(X_i) is

$$
\Big\{(\boldsymbol{w}, \boldsymbol{z}, s) \in \mathbb{R}^n \times [0,1]^\ell \times \mathbb{R} : \exists (p_0, \boldsymbol{p}) \in \mathbb{R}_+^{\ell+1} :
$$

$$
\boldsymbol{w}^\top \boldsymbol{x}_i = b_1 + \sum_{j=1}^{\ell} p_j - p_0
$$

$$
(b_{j+1} - b_j) z_{j+1} \le p_j \le (b_{j+1} - b_j) z_j, \ \ j \in [\ell - 1]
$$

$$
(1 - z_1)\mathcal{L}\left(b_1 - \frac{p_0}{1 - z_1}\right) + \sum_{j=1}^{\ell-1} (z_j - z_{j+1}) \mathcal{L}\left(b_i + \frac{p_i - z_{i+1}(b_{i+1} - b_i)}{z_i - z_{i+1}}\right) + z_\ell \mathcal{L}\left(b_\ell + \frac{p_\ell}{z_\ell}\right) \le s \Big\}
$$

Strong Reformulation

Applying the strengthening to the fair regression problem:

$$
\begin{aligned}
\min \; & \sum_{i=1}^{m} s_i + \lambda \max_{j \in 1,\ldots,\ell} \left(\frac{1}{m} \sum_{i=1}^{m} z_{ij} - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} z_{ij} \right) \\
\text{s.t. } & (\boldsymbol{w}, \boldsymbol{z}_i, s_i) \in X_i && i \in [m] \\
& \boldsymbol{z}_i \in \{0,1\}^{\ell} && i \in [m]
\end{aligned}
$$

Strong Reformulation

Applying the strengthening to the fair regression problem:

$$\begin{aligned}
\min \sum_{i=1}^{m} s_i + \lambda \max_{j \in 1,\ldots,\ell} & \left(\frac{1}{m} \sum_{i=1}^{m} z_{ij} - \frac{1}{m_1} \sum_{i=1:a_i=1}^{m} z_{ij} \right) \\
\text{s.t. } (\boldsymbol{w}, \boldsymbol{z}_i, s_i) & \in \textbf{cl conv}(\mathbf{X_i}) && i \in [m] \\
\boldsymbol{z}_i & \in \{0,1\}^{\ell} && i \in [m]
\end{aligned}$$

Computational Results on Synthetic Data

$n = 10$ features, $\ell = 40$, $m = \{15, 30, 50, 100\}$, 5 random instances per m

Table: Branch-and-bound performance of `Big-M` vs `Strong` for least squares regression

λ	`Big-M`				`Strong`			
	Relax Gap	End Gap	Time	Nodes	Relax Gap	End Gap	Time	Nodes
0.01	45.3%	2.1%	487	4,010,293	1.9%	0.0%	252	223,581
0.02	60.8%	9.0%	1,654	19,895,563	5.4%	0.7%	908	1,079,684
0.04	73.2%	19.1%	2,040	20,405,258	8.3%	1.9%	950	902,741
0.05	76.6%	28.4%	2,681	29,063,625	11.3%	2.6%	1,212	1,058,541
0.06	78.8%	30.7%	2,759	28,508,133	13.1%	2.9%	1,507	1,447,340
0.08	81.5%	38.1%	3,023	32,654,651	14.0%	4.0%	1,673	1,284,442
0.10	83.4%	47.4%	3,196	35,020,673	16.0%	4.6%	1,803	1,275,695
0.20	87.7%	55.3%	3,147	28,630,313	22.9%	8.6%	1,625	683,419
0.30	89.5%	62.8%	3,295	29,468,759	31.2%	10.8%	1,555	507,315
0.50	91.7%	73.6%	3,600	31,892,472	45.5%	12.7%	1,611	437,835
Avg	**76.9%**	**36.6%**	**2588**	**25,954,974**	**16.9%**	**4.9%**	**1310**	**890,059**

Computational Results on Synthetic Data

$n = 10$ features, $\ell = 40$, $m = \{15, 30, 50, 100\}$, 5 random instances per m

Table: Branch-and-bound performance of `Big-M` vs `Strong` for least squares regression

λ	`Big-M`				`Strong`			
	Relax Gap	End Gap	Time	Nodes	Relax Gap	End Gap	Time	Nodes
0.01	45.3%	2.1%	487	4,010,293	1.9%	0.0%	252	223,581
0.02	60.8%	9.0%	1,654	19,895,563	5.4%	0.7%	908	1,079,684
0.04	73.2%	19.1%	2,040	20,405,258	8.3%	1.9%	950	902,741
0.05	76.6%	28.4%	2,681	29,063,625	11.3%	2.6%	1,212	1,058,541
0.06	78.8%	30.7%	2,759	28,508,133	13.1%	2.9%	1,507	1,447,340
0.08	81.5%	38.1%	3,023	32,654,651	14.0%	4.0%	1,673	1,284,442
0.10	83.4%	47.4%	3,196	35,020,673	16.0%	4.6%	1,803	1,275,695
0.20	87.7%	55.3%	3,147	28,630,313	22.9%	8.6%	1,625	683,419
0.30	89.5%	62.8%	3,295	29,468,759	31.2%	10.8%	1,555	507,315
0.50	91.7%	73.6%	3,600	31,892,472	45.5%	12.7%	1,611	437,835
Avg	**76.9%**	**36.6%**	**2588**	**25,954,974**	**16.9%**	**4.9%**	**1310**	**890,059**

$\rightarrow$ **Relaxation is solved in 0.25s on average.**

Table of Contents

Proposed Methodologies

Three methodologies to obtain *high quality* solutions, balancing **optimality** and **computational costs**.

Proposed Methodologies

Three methodologies to obtain *high quality* solutions, balancing **optimality** and **computational costs**.

1. **Convex relaxation**: Relax integrality constraints of the strong reformulation, solve to optimality and use $\mathbf{w}^*$ solution.
   $\rightarrow$ Solves a single convex problem, efficient and fast with commercial solvers

Proposed Methodologies

Three methodologies to obtain *high quality* solutions, balancing **optimality** and **computational costs**.

1. **Convex relaxation**: Relax integrality constraints of the strong reformulation, solve to optimality and use $\mathbf{w}^*$ solution.
   $\rightarrow$ Solves a single convex problem, efficient and fast with commercial solvers
2. **Coordinate descent**: Initializing with $\mathbf{w}^*$, iteratively improve by optimizing one coordinate at a time.
   $\rightarrow$ We propose an efficient coordinate descent method, motivated by convex-hull representation of X_i

Proposed Methodologies

Three methodologies to obtain *high quality* solutions, balancing **optimality** and **computational costs**.

1. **Convex relaxation**: Relax integrality constraints of the strong reformulation, solve to optimality and use $\mathbf{w}^*$ solution.
   $\rightarrow$ Solves a single convex problem, efficient and fast with commercial solvers
2. **Coordinate descent**: Initializing with $\mathbf{w}^*$, iteratively improve by optimizing one coordinate at a time.
   $\rightarrow$ We propose an efficient coordinate descent method, motivated by convex-hull representation of X_i
3. **Mixed Integer Program**: Use strong convex relaxation with branch-&-bound to solve the exact fair regression problem.
   $\rightarrow$ Substantially improves over MIO methods based on natural big-M formulations, produces better estimators, but currently does not scale

Coordinate Descent: Optimizing a Single Coordinate

Goal: given $\boldsymbol{w}$, improve by updating one coordinate w_k at a time

Coordinate Descent: Optimizing a Single Coordinate

Goal: given $\boldsymbol{w}$, improve by updating one coordinate w_k at a time

---

Recall: Single-Observation Fair Regression

$m = 1$ observations, n features

Have convex hull representation for X_i, where we model indicators

$$\mathbb{1}\left(\boldsymbol{w}^\top \boldsymbol{x} > b_j\right), \quad j = 1, \ldots, \ell$$

Coordinate Descent: Optimizing a Single Coordinate

Goal: given $\boldsymbol{w}$, improve by updating one coordinate w_k at a time

---

Recall: Single-Observation Fair Regression

$m = 1$ observations, n features

Have convex hull representation for X_i, where we model indicators

$$\mathbb{1}\big(\boldsymbol{w}^\top \boldsymbol{x} > b_j\big), \quad j = 1, \ldots, \ell$$

Now: Single-Feature Fair Regression

m observations, $n = 1$ features

$$\mathbb{1}\big(\boldsymbol{w}^\top \boldsymbol{x}_i > b_j\big) = \mathbb{1}\Big(w_k x_{ik} + \sum_{u \neq k} w_u x_{iu} > b_j\Big) = \mathbb{1}\big(w_k > \tilde{b}_{ij}\big),$$

where $\tilde{b}_{ij} \;=\; \dfrac{b_j - \sum_{u \neq k} w_u x_{iu}}{|x_{ik}|} \quad \forall\, j = 1, \ldots, \ell,\; i = 1, \ldots, m.$

Coordinate Descent: Optimizing a Single Coordinate

Goal: given $\boldsymbol{w}$, improve by updating one coordinate w_k at a time

Recall: Single-Observation Fair Regression

$m = 1$ observations, n features

Have convex hull representation for X_i, where we model indicators

$$\mathbb{1}\left(\boldsymbol{w}^\top \boldsymbol{x} > b_j\right), \quad j = 1, \ldots, \ell$$

Now: Single-Feature Fair Regression

m observations, $n = 1$ features

$$\mathbb{1}\left(\boldsymbol{w}^\top \boldsymbol{x}_i > b_j\right) = \mathbb{1}\left(w_k x_{ik} + \sum_{u \neq k} w_u x_{iu} > b_j\right) = \mathbb{1}\left(w_k > \tilde{b}_{ij}\right),$$

where $\tilde{b}_{ij} \; = \; \dfrac{b_j - \sum_{u \neq k} w_u x_{iu}}{|x_{ik}|} \quad \forall\, j = 1, \ldots, \ell,\; i = 1, \ldots, m.$

$\rightarrow$ Equivalent to a *single-observation* fair-regression problem with $m \times \ell$ thresholds, enabling efficient iterations

Table of Contents

Real Data Experiments

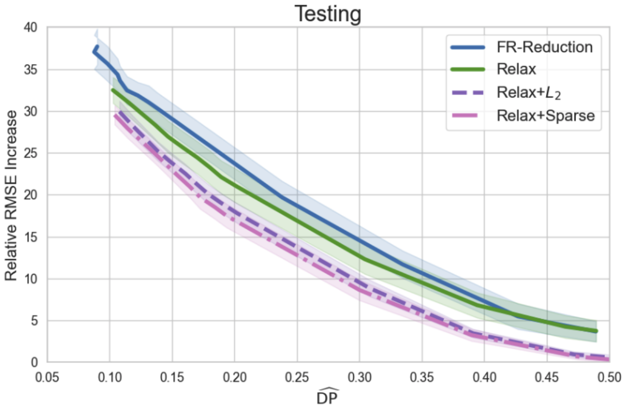
Application: Price of fairness, $\widehat{\mathrm{DP}} = \max_{b\in\mathbb{R}} \left|\hat{P}(\hat{y} > b) - \hat{P}(\hat{y} > b|a = 1)\right|$.

We compare our methodologies with `FR-Reduction` (Agarwal, Dudik & Wu (2019), an algorithm which solves a sequence of weighted classification problem.

Figures present the accuracy-fairness curves attained by different methods on testing data. Accuracy is shown as the relative % increase in loss over a vanilla linear regression.

Results on Real Data: Communities and Crime

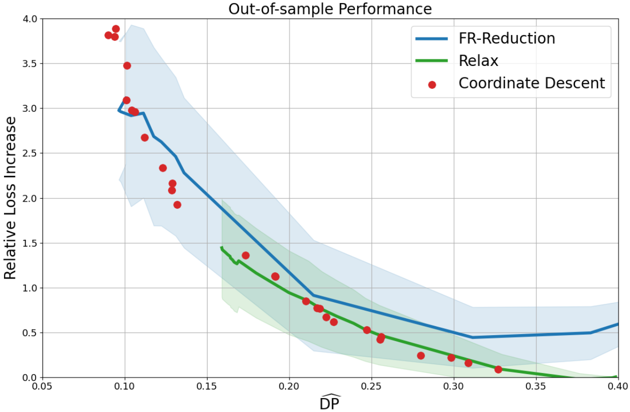
Data: $n = 119, \ m = 1,994, \ \ell = 40$



Average training time reduced from $\approx$ **200s** to **5s**

Results on Real Data: Law School

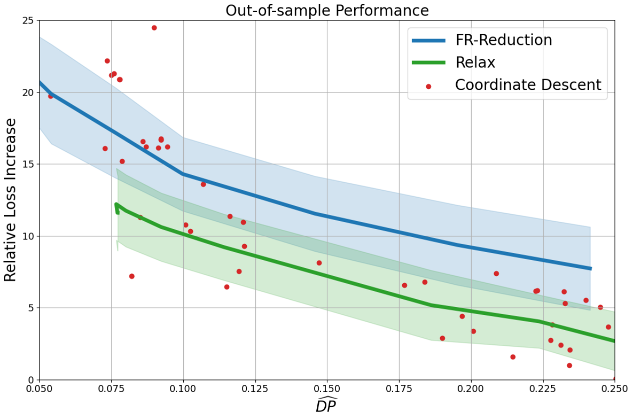
Data: $n = 12,\ m = 2,000,\ \ell = 40$



Average training time reduced from $\approx$ **136s** to **4s** (convex relaxation) / **45s** (coordinate descent)

Results on Real Data: Adult

Data: $n = 103$, $m = 2,000$, $\ell = 40$. Now, we are solving a **logistic regression** problem



Average training time reduced from $\approx$ **1, 417s** to **56s** (convex relaxation)

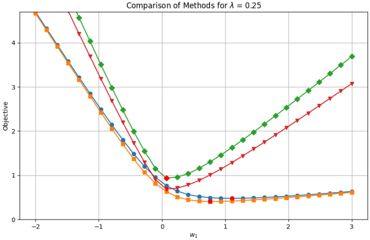
Conclusion

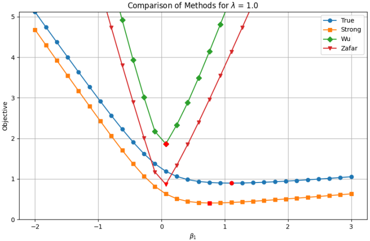
- **Versatile** framework: proposed methods can be adapted for generalized linear regression (ex: logistic regression) and for other popular fairness metrics
- **Key substructure:** piecewise convex function with indicators on intervals
- Relaxation + Coordinate Descent produce models **competitive** with state of the art at a fraction of the time

Our Convex Relaxation vs Convex Approximations

Existing approaches consider convex approximations of the fairness measure

We better capture the shape of true objective by considering the **joint structure of the loss function and fairness**



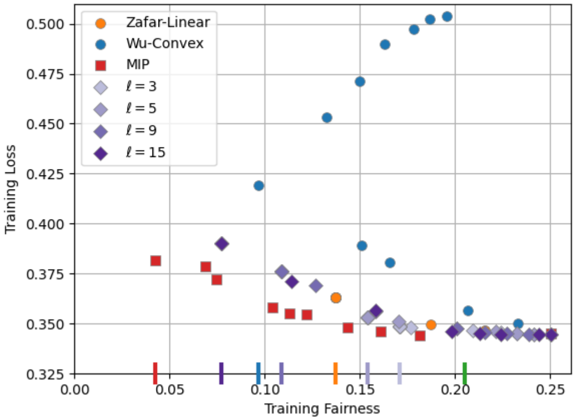


Synthetic Data: Optimality Gaps of Different Methods

λ	relax	CD-0	CD-unfair	CD-relax	MICQO
0.01	7.9	6.3	3.5	2.1	0.0
0.02	13.4	9.0	5.9	4.1	0.7
0.04	19.5	13.6	10.0	7.8	1.9
0.05	24.0	16.1	10.9	8.6	2.6
0.06	24.6	16.3	11.9	9.3	2.9
0.08	26.3	22.2	17.0	13.3	4.0
0.10	32.1	23.9	21.7	15.8	4.6
0.20	44.1	37.1	31.0	24.6	8.6
0.30	50.0	57.5	38.3	28.9	10.8
0.50	56.2	65.4	44.2	36.5	12.7
Avg	**29.8**	**26.7**	**19.4**	**15.1**	**4.9**

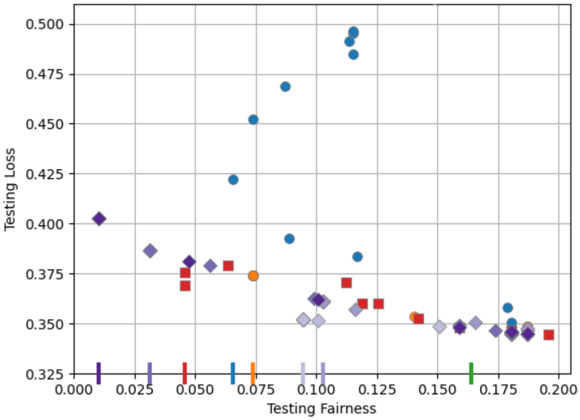
Overcoming 'too relaxed to be fair'

Idea: approximate fair classification $\ell = 1, b_1 = 0$ with a fair regression relaxation that artificially adds levels. This gives us a convex problem that can still produce fair models.



Overcoming 'too relaxed to be fair'

Idea: approximate fair classification $\ell = 1, b_1 = 0$ with a fair regression relaxation that artificially adds levels. This gives us a convex problem that can still produce fair models.



Discretized Fairness vs Exact Fairness Measures

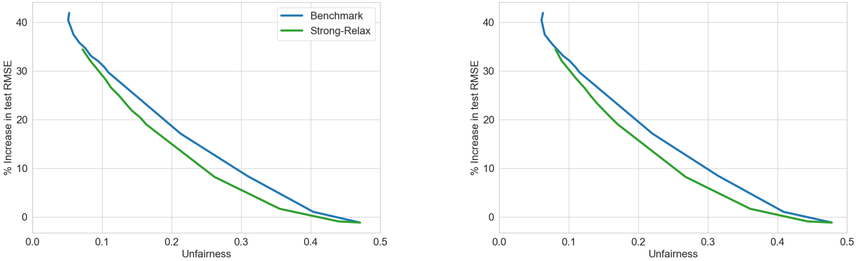


Figure: Training discretized unfairness (left) vs training exact unfairness (right).