Fairness Constraints: A Mechanism for Fair Classification

Muhammad Bilal Zafar

with Isabel Valera, Manuel Gomez Rodriguez
and Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS)

Data Driven Decision Making

- Data driven decision making used in
  - Spam classifiers
  - Recommendation systems
  - Bail-granting classifiers
  - Job hiring
  - Loan approval
  - ...
- These tasks work by learning from past data
  - Human decision making follows intuition
  - Data driven approach automates the human decision making

Classification Tasks in Practice

- Given some items along with their features, predict their class labels
  - Making a decision to hire a job applicant
- Classifiers designed to discriminate based on features
- All features are not equal!
  - Non-sensitive features: educational level, work experience, etc.
  - Sensitive features: gender, race, etc.
- Discriminating on sensitive features prohibited

Classifiers and Direct Discrimination

- Historical biases in the data
  - Gender discrimination
  - Racial biases
- Classifiers try to optimize for accuracy
  - Any past human biases will get transferred to the learnt classifiers
  - Bias against females in the past data -> bias against females in the future classifier decisions

Classifiers and Indirect Discrimination

- One might chose not to use sensitive features
- Correlations between sensitive and non-sensitive features [Pedreschi et al.]
  - People from a certain race live in a specific neighborhood
  - Biases can get captured through these correlations
- Discrimination can happen even in the lack of intent

This Talk

- Proposing a framework for fair classification
  - Defining Fairness
  - Introducing fairness constraints and incorporating them into classifiers
  - Evaluation: Analyzing performance trade-offs of fair classifiers

This Talk

- Proposing a framework for fair classification
  - **Defining Fairness**
  - Introducing fairness constraints and incorporating them into classifiers
  - Evaluation: Analyzing performance trade-offs of fair classifiers

Defining Fairness: Motivation

- Doctrine of disparate impact
  - US law concerning employment, housing, etc.
  - "*practices [...] considered discriminatory and illegal if they have a disproportionate adverse impact on persons in a protected class*"
- Does not focus on the inputs or the process but on the outcome!
- What constitutes "disproportionality"?

Applying Doctrine of Disparate Impact: 80% Rule

- The 80% rule
  - If 50% of male applicants get selected for the job, at least 40% of females should also get selected
- A fair system might not always be 80:100
  - In certain scenarios, the prescribed proportion could be 50:100
- Our goal is to enable a range of "fair" proportions

Fair Classifier

A classifier whose output achieves a given proportion of items (in positive class) with different values of sensitive feature

This Talk

- Proposing a framework for fair classification
  - ✓ Defining Fairness
  - **Introducing fairness constraints and incorporating them into classifiers**
  - Evaluation: Analyzing performance trade-offs of fair classifiers

Ensuring Fair Classification

- Classifiers try to optimize certain objectives
  - Logistic regression
  - Support vector machines
- Introduce fairness constraints
  - Optimize the given function *under the constraints*
- **Idea:** Optimize such that ratio of females/ males is greater than 80% threshold

Ensuring Fair Classification

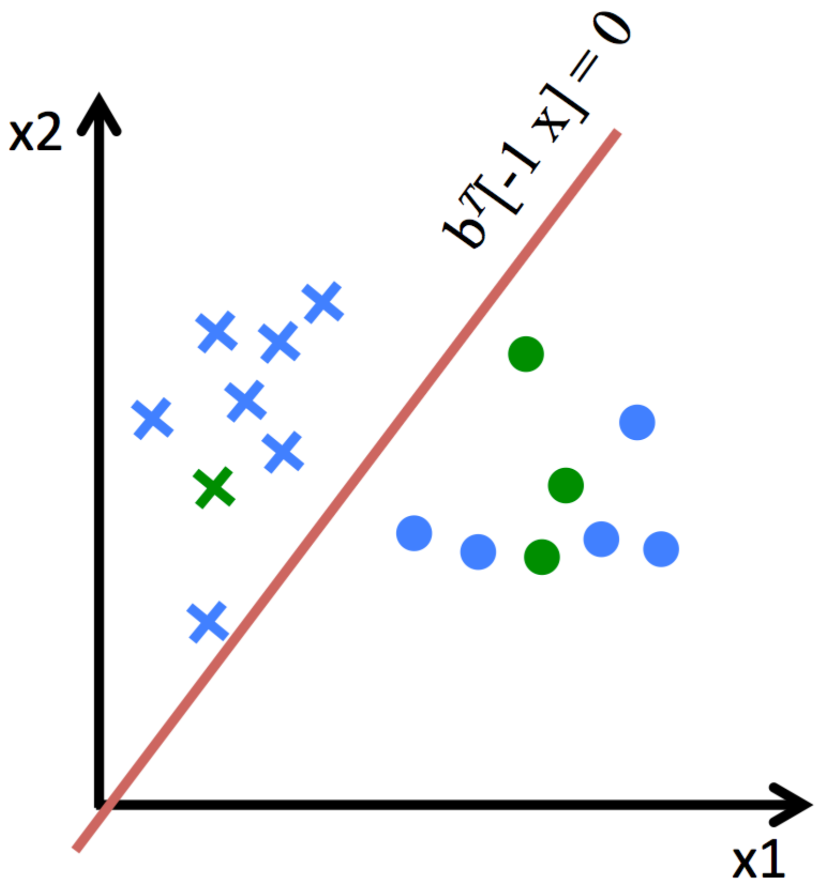
- Classifiers try to optimize certain objectives
  - Logistic regression
  - Support vector machines
- Introduce fairness constraints
  - Optimize the given function *under the constraints*
- **Idea:** Optimize such that ratio of females/ males is greater than 80% threshold
  - Hard to encode and solve these constraints

Fairness Constraints

Key Idea: Limit the cross-covariance between sensitive feature value and distance from decision boundary

An Instance of Implementing Fairness Constraints

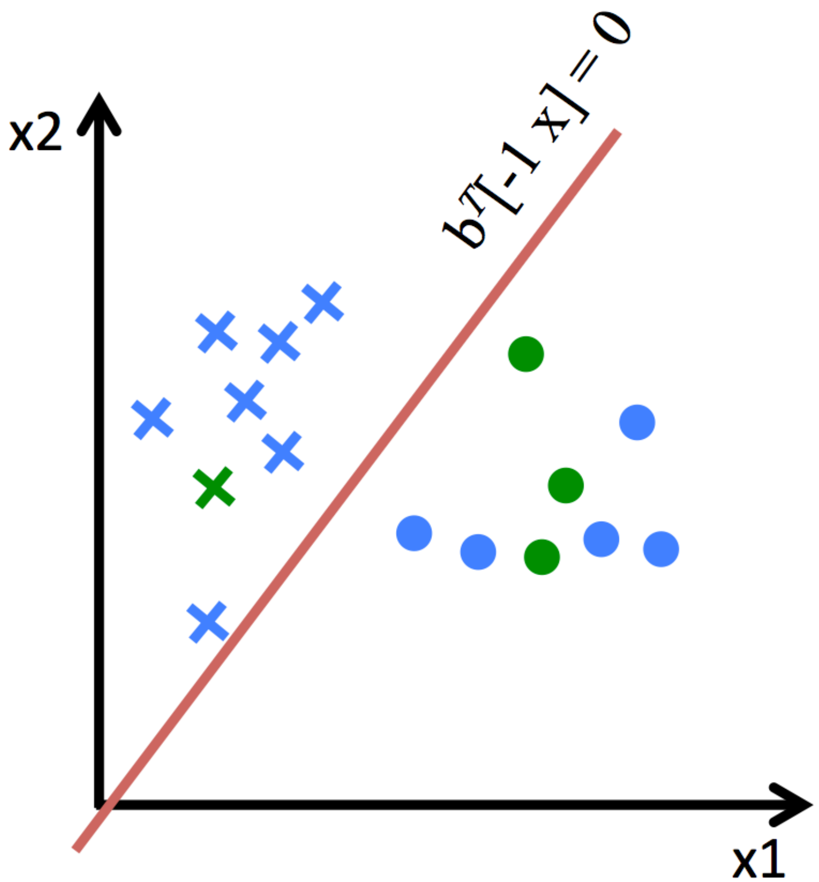
25 : 58
0.43



An Instance of Implementing Fairness Constraints

$$\boxed{\left|\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \bar{\mathbf{z}}\right)\mathbf{b}^T[-1\ \mathbf{x}_i]\right| \leq \mathbf{c}}$$

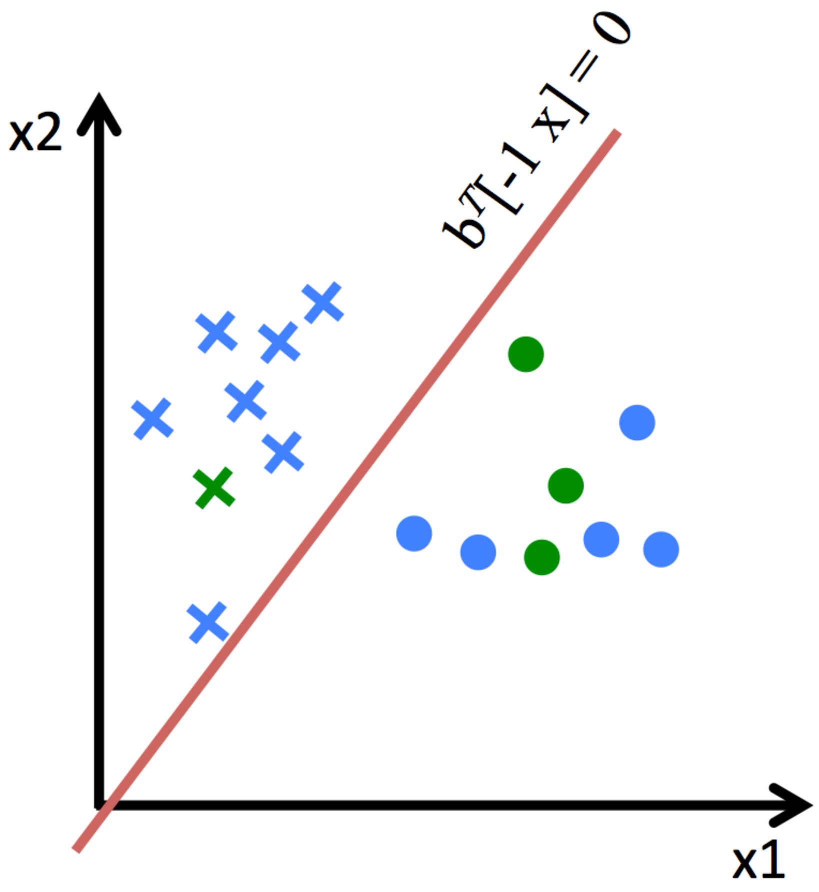
25 : 58
0.43

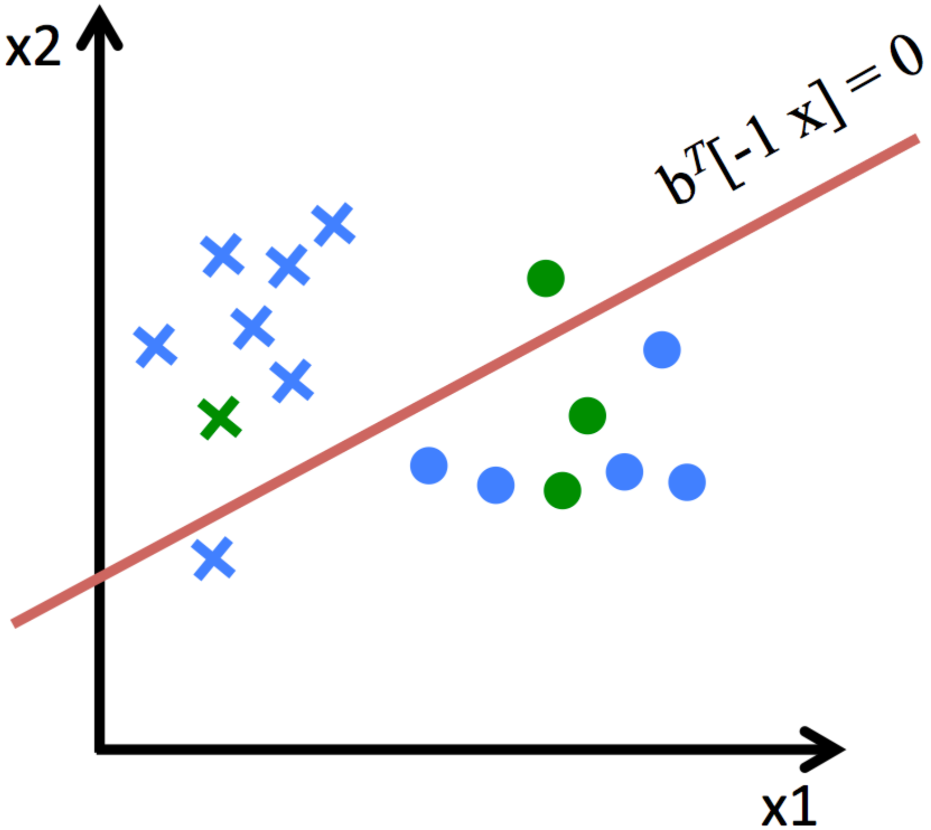


An Instance of Implementing Fairness Constraints

$$\left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T [-1 \ \mathbf{x}_i] \right| \leq \mathbf{c}$$

25 : 58	50 : 50
0.43	1.0





Modifying the Logistic Regression Classifier

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}$$

maximize $\quad \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i)$

Modifying the Logistic Regression Classifier

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}$$

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i) \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T[-1\ \mathbf{x}_i] \geq -\mathbf{c} \end{array}$$

Key point: Possible to solve this problem efficiently

Modifying the Logistic Regression Classifier

$$p(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}$$

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i) \\ \text{subject to} & \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T [-1 \ \mathbf{x}_i] \leq \mathbf{c}, \\ & \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \mathbf{b}^T [-1 \ \mathbf{x}_i] \geq -\mathbf{c} \end{array}$$

Key point: Possible to solve this problem efficiently

Also implemented for SVM and hinge loss classifiers

This Talk

- Proposing a framework for fair classification
  - ✓ Defining Fairness
  - ✓ Introducing fairness constraints and incorporating them into classifiers
  - **Evaluation: Analyzing performance trade-offs of fair classifiers**

Evaluation: Key Questions

- Does introducing cross-covariance (fairness) constraints ensure fair classification?
  - Can one vary the thresholds for cross-covariance to achieve different proportions?
- Do "fairness" constraints lead to optimal performance?

Dataset

- Census income dataset
  - 45,222 subjects
  - 14 features
  - Non-sensitive features: Educational level, number of hour of work per week, etc.
  - Sensitive features: Gender and race
- Prediction task: Whether a person earns >50K$ (positive) or <50K$ (negative) per year

Lack of Fairness in Census Income Dataset

Gender	<50K	>50K
Female	89%	11%
Male	69%	31%

Race	<50K	>50K
American-Indian/Eskimo	88%	12%
Asian/Pacific-Islander	72%	28%
Black	87%	13%
White	74%	26%
Other	87%	13%

Lack of Fairness in Census Income Dataset

Gender	<50K	>50K
Female	89%	**11%**
Male	69%	**31%**

0.35

Race	<50K	>50K
American-Indian/Eskimo	88%	12%
Asian/Pacific-Islander	72%	28%
Black	87%	**13%**
White	74%	**26%**
Other	87%	13%

0.5

Logistic Regression Classifier (without constraints)

- Achieves 81% accuracy
  - Sensitive features not used

Logistic Regression Classifier (without constraints)

- Achieves 81% accuracy
  - Sensitive features not used

Gender	<50K	>50K
Female	94%	6%
Male	80%	20%

0.30

Race	<50K	>50K
American-Indian/Eskimo	95%	5%
Asian/Pacific-Islander	81%	19%
Black	89%	11%
White	83%	17%
Other	92%	8%

0.65

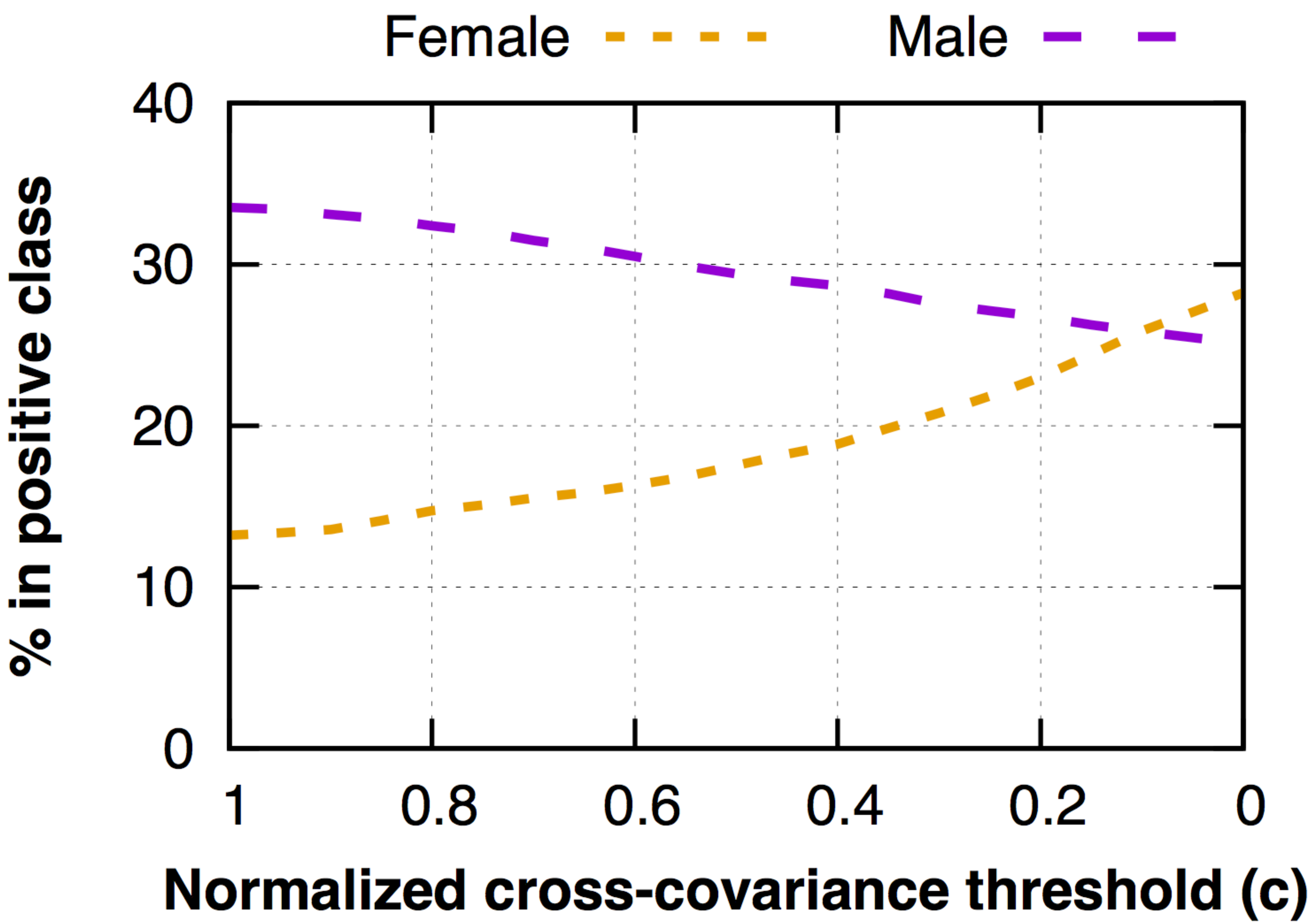
Logistic Regression Classifier (with constraints)

- Introduce cross-covariance constraints

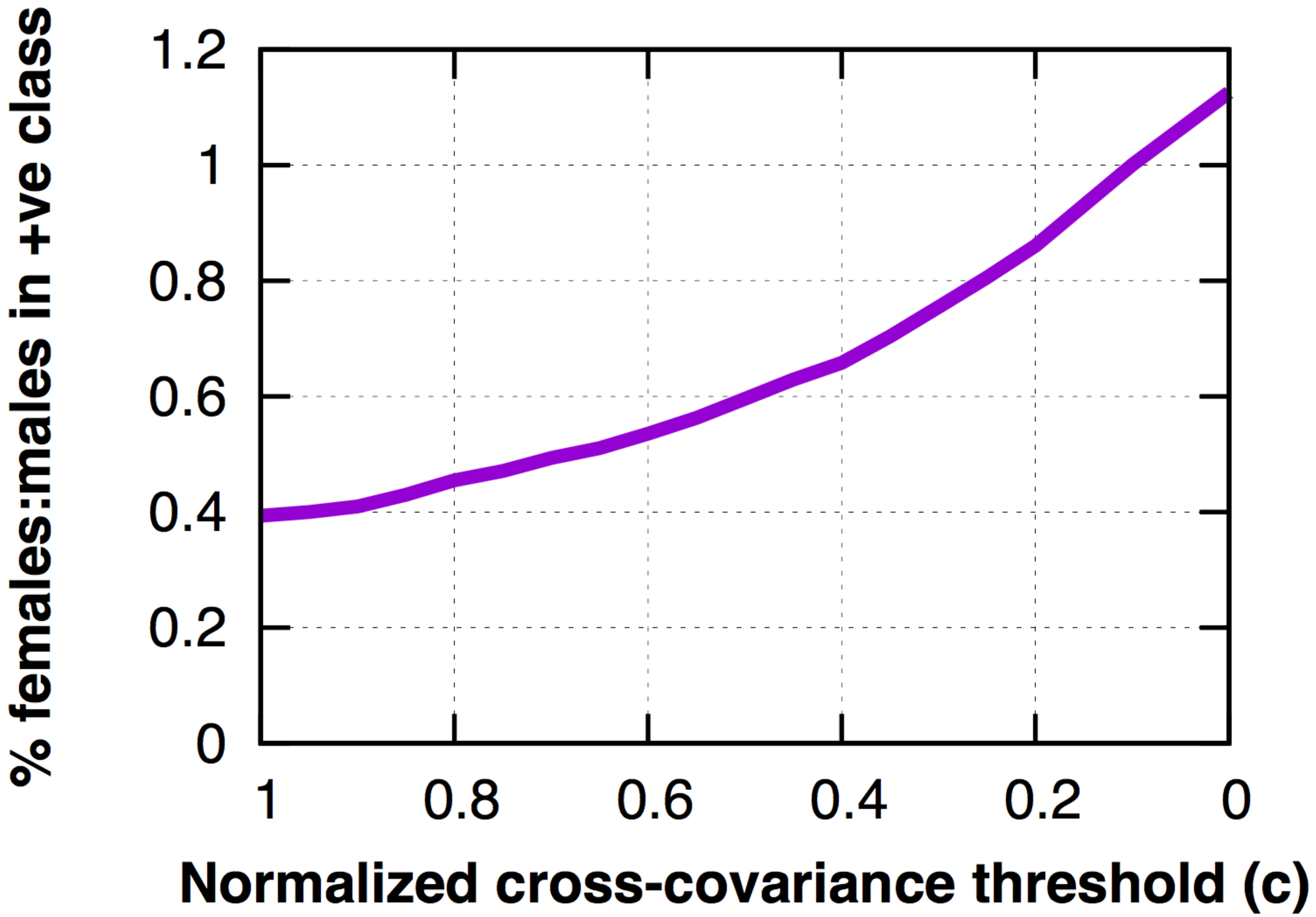
$$\boxed{\left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \mathbf{b}^T [-1 \;\; \mathbf{x}_i] \right| \leq \mathbf{c}}$$

- Vary the fairness threshold (*c*) to achieve different proportions of sensitive feature values

Tightening the Constraints Increases Fairness



Tightening the Constraints Increases Fairness

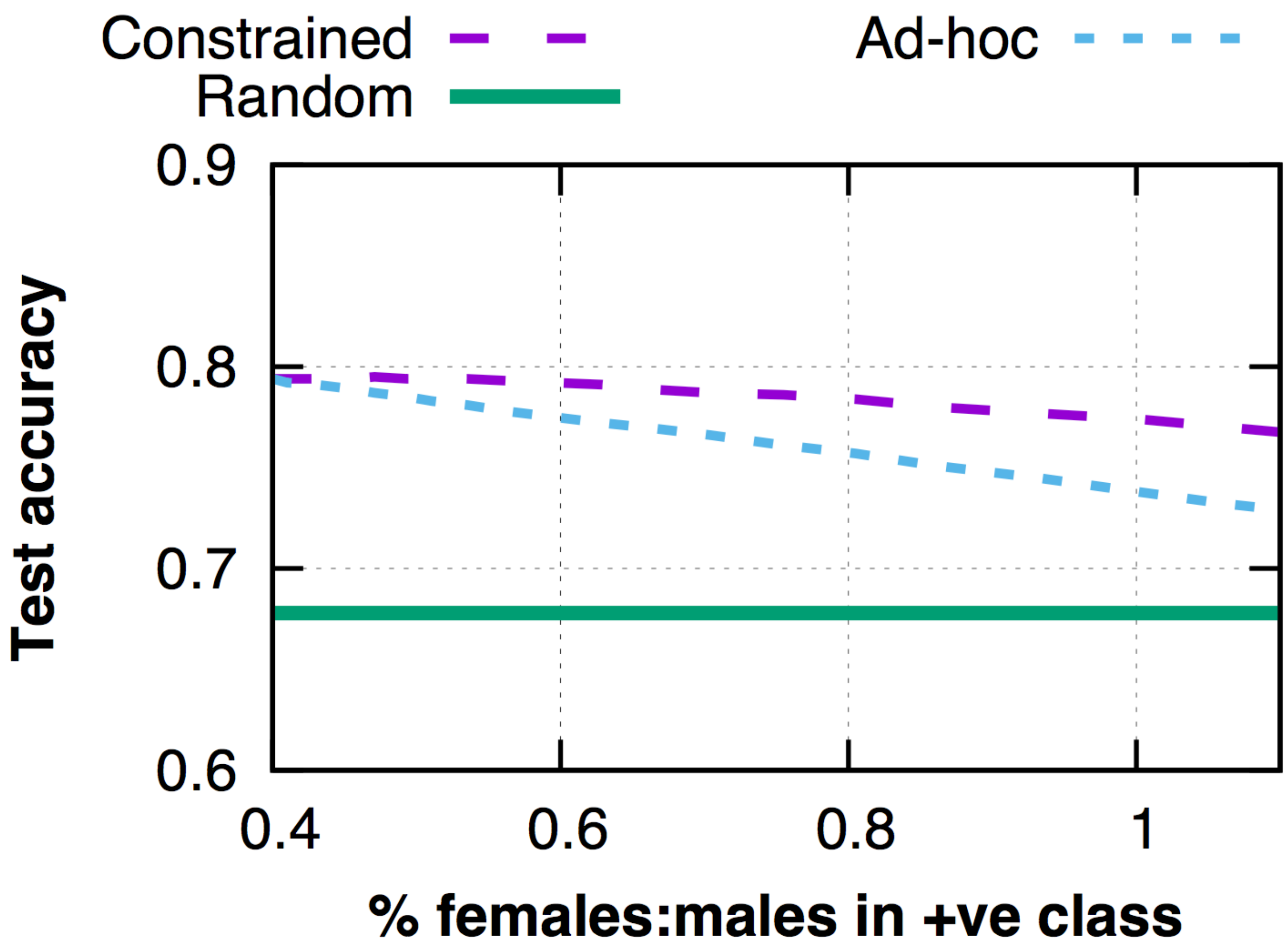


Fairness vs. Accuracy Trade-off

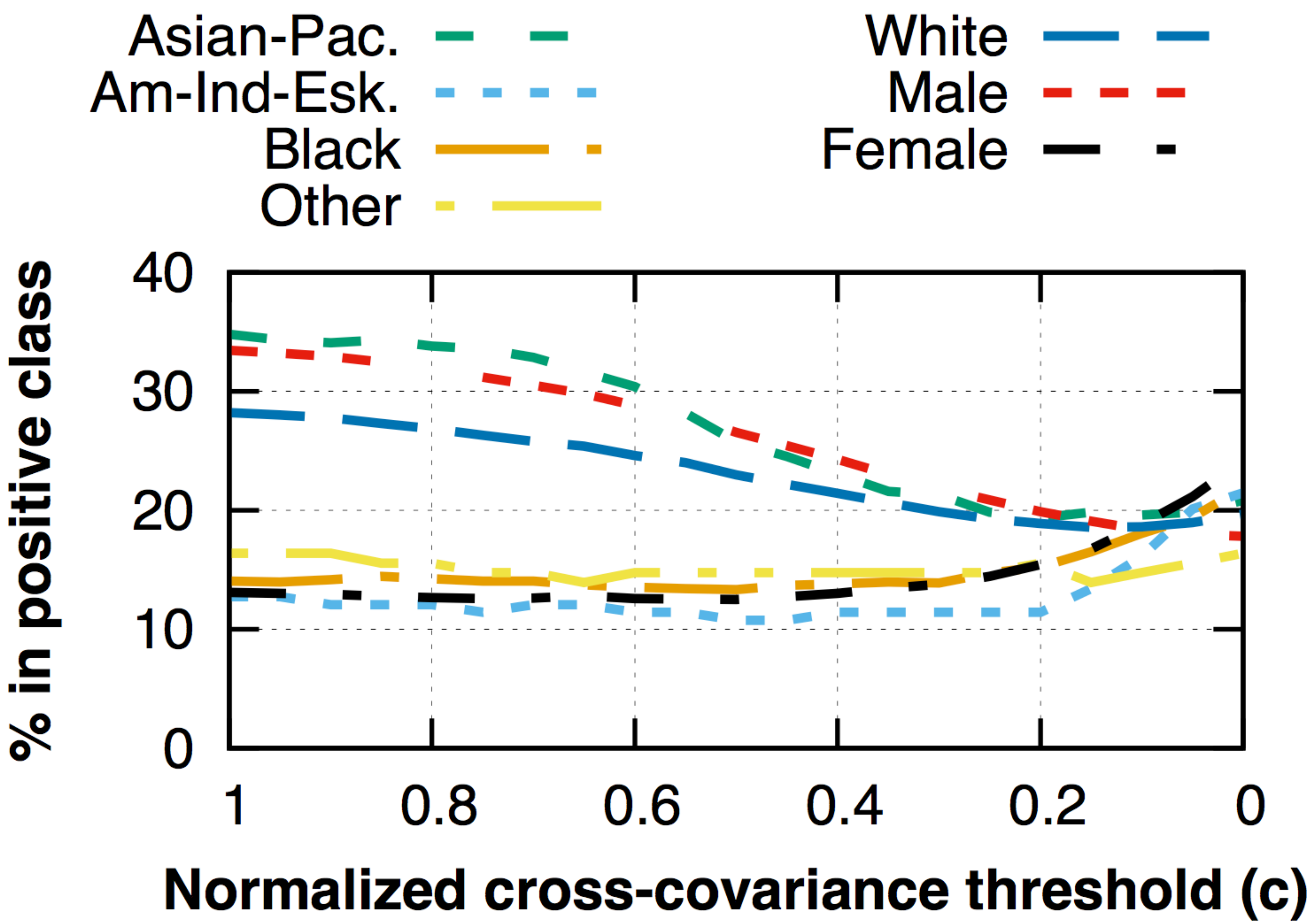
- Random classifier
- Ad-hoc classifier: Switch females to +ve class until fairness ratio is achieved

Fairness vs. Accuracy Trade-off

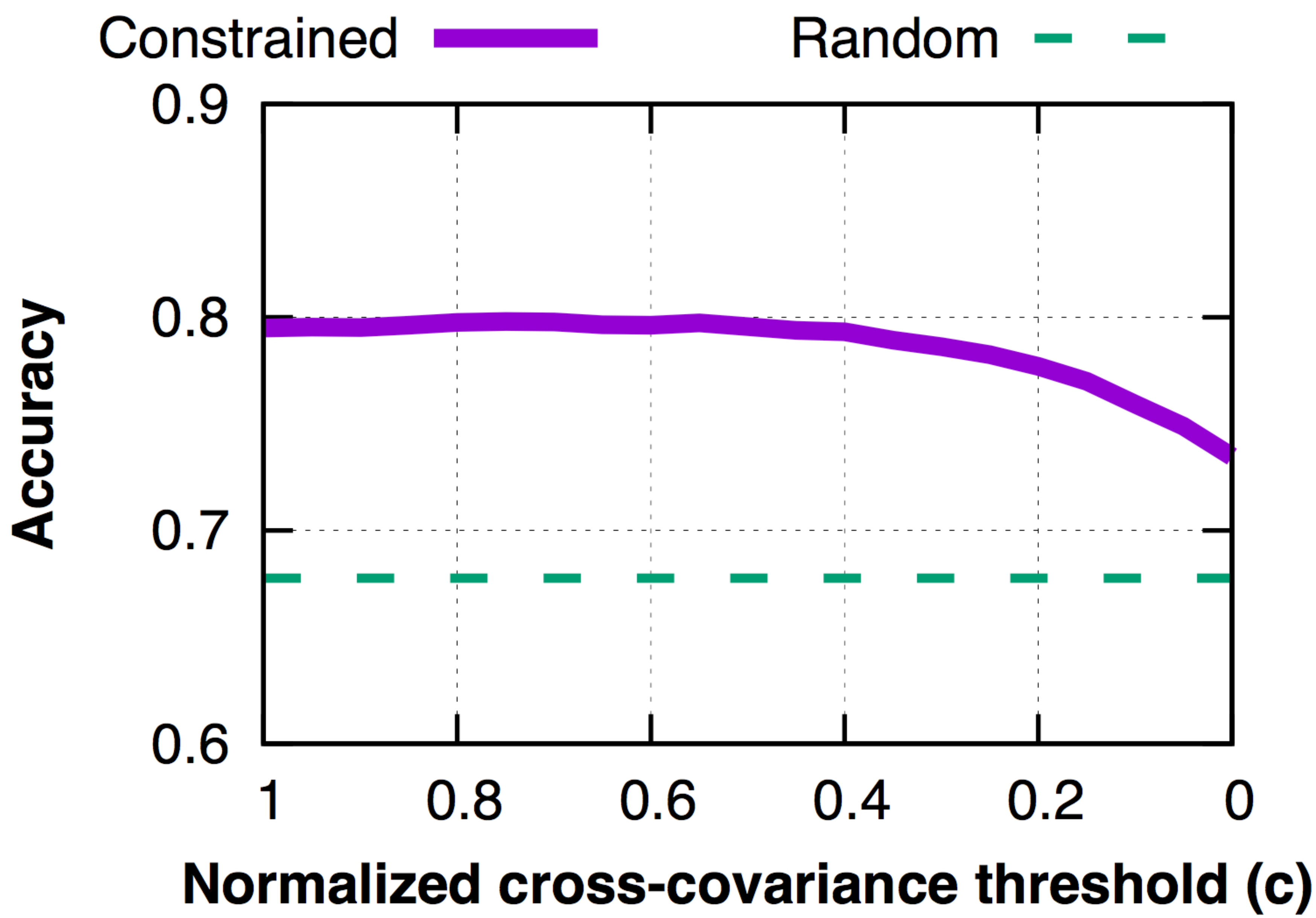
- Random classifier
- Ad-hoc classifier: Switch females to +ve class until fairness ratio is achieved



Introducing Fairness for Multiple Features



Introducing Fairness for Multiple Features



Similar results for SVM and hinge loss classifiers

Advantages of Fairness Constraints

- Do not need the sensitive feature value at decision time
  - Sensitive features are not always available
- Can cater to categorical as well as continuous values of sensitive features
  - Categorical sensitive features with more than two values can also be handled
- Optimality (under given constraints)

Future Directions

- Getting insights into the dataset
  - Analyzing which weights get adjusted while controlling for fairness
  - These changes can be used to better understand the dataset [Chang et al.]
  - Example scenario (hypothetical)

Unconstrained (unfair) classifier

# hours/week	marital status	.....	education
1.3	0.5		1.5

Future Directions

- Getting insights into the dataset
  - Analyzing which weights get adjusted while controlling for fairness
  - These changes can be used to better understand the dataset [Chang et al.]
  - Example scenario (hypothetical)

Unconstrained (unfair) classifier

# hours/week	marital status	.....	education
1.3	0.5		**1.5**

Constrained (fair) classifier

# hours/week	marital status	.....	education
1.2	0.4		**6.1**

Future Directions

- Applications to other domains
  - Spam detection
  - Online ads
- Effectiveness of fairness constraints in different datasets
- Comparison to other techniques

Future Directions

- Applications to other domains
  - Spam detection
  - Online ads
- Effectiveness of fairness constraints in different datasets
- Comparison to other techniques

Questions?