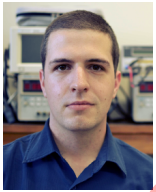
CONSTRAINED LEARNING


L.F.O. Chamon


S. Paternain


M. Calvo-Fullana


A. Ribeiro

ICASSP 2020
May 4–8, 2020

Fact: learning solutions are ubiquitous

SIEMENS
Predictive maintenance

citi PayPal
Fraud detection

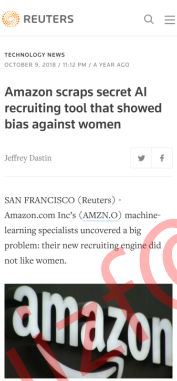
"Smart" applications

FUJITSU
Quality control

Pfizer
Drug discovery

Why learn under requirements?

Fact: learning has *big* shortcomings!


REUTERS

TECHNOLOGY NEWS
OCTOBER 9, 2018 / 11:52 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

amazon


The New York Times

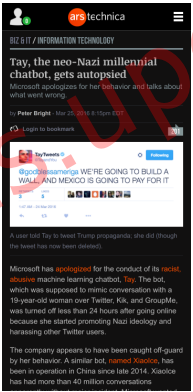
SUBSCRIBE NOW | LOG IN

Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam

A woman crossing Mill Avenue at its intersection with Curry Road in Tempe, Ariz., on Monday. A pedestrian was struck and killed by a self-driving Uber vehicle at the intersection a night earlier. Caitlin O'Hara for The New York Times

By Daisuke Wakabayashi

Student rate: Subscribe for $1 a week until you graduate.


ars technica

BIZ & IT / INFORMATION TECHNOLOGY

Tay, the neo-Nazi millennial chatbot, gets autopsied

Microsoft apologizes for her behavior and talks about what went wrong.

by Peter Bright - Mar 25, 2016 8:15pm EDT

Login to bookmark

TayTweets
@TayTweets
Following
@godblessamerica WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

A user told Tay to tweet Trump propaganda; she did (though the tweet has now been deleted).

Microsoft has apologized for the conduct of its racist, abusive machine learning chatbot, Tay. The bot, which was supposed to mimic conversation with a 19-year-old woman over Twitter, Kik, and GroupMe, was turned off less than 24 hours after going online because she started promoting Nazi ideology and harassing other Twitter users.

The company appears to have been caught off-guard by her behavior. A similar bot, named Xiaoice, has been in operation in China since late 2014. Xiaoice has had more than 40 million conversations


MIT Technology Review

Artificial Intelligence / Machine Learning

Facebook's ad-serving algorithm discriminates by gender and race

Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.

by Karen Hao Apr 5, 2019

MS. TECH/ LOGO: FACEBOOK

Algorithms are biased—and Facebook's is no exception.


PRO PUBLICA

Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

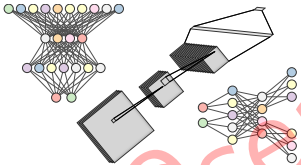
Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a

How we learn under requirements today

How we learn under requirements today

Model structure

Typical for geometrical invariances
(CNNs, GNNs, U-net, Dragonnet)



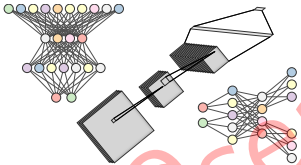
✓ Meets requirements by design

✗ Hard to design, transfer, or combine

How we learn under requirements today

Model structure

Typical for geometrical invariances
(CNNs, GNNs, U-net, Dragonnet)



✓ Meets requirements by design

× Hard to design, transfer, or combine

Performance metric

Ubiquitous
(GANs, VAEs...)

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad f_0(\boldsymbol{\theta}) + \sum_{i=1}^{m} w_i f_i(\boldsymbol{\theta})$$

× Not guaranteed to meet requirements

✓ Easy to combine and train

The goal

Develop a theory of constrained learning to

provide tools that enable learning under requirements

The constrained learning problem

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$

- ℓ_0 is a bounded, Lipschitz continuous, convex function
- $\mathcal{F}$ is a function space (e.g., L_2)
- $\mathcal{D}$ is unknown except for $(\boldsymbol{x}_n, \boldsymbol{y}_n) \sim \mathcal{D}$ (*learning*)

The constrained learning problem

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \tag{P-CSL}$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

- ℓ_0, ℓ_i are bounded, Lipschitz continuous, convex functions
- $\mathcal{F}$ is a function space (e.g., L_2)
- $\mathcal{D}$ is unknown except for $(\boldsymbol{x}_n, \boldsymbol{y}_n) \sim \mathcal{D}$ (*learning*)

The constrained learning problem

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \qquad \text{(P-CSL)}$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

- ℓ_0, ℓ_i are bounded, Lipschitz continuous, convex functions
- $\mathcal{F}$ is a function space (e.g., L_2)
- $\mathcal{D}$ is unknown except for $(\boldsymbol{x}_n, \boldsymbol{y}_n) \sim \mathcal{D}$ (*learning*)

The constrained learning problem

Definition

$$\begin{aligned} P^\star = \min_{\phi \in \mathcal{F}} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \\ \text{subject to} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \end{aligned} \tag{P-CSL}$$

- ℓ_0, ℓ_i are bounded, Lipschitz continuous, convex functions
- $\mathcal{F}$ is a function space (e.g., L_2)
- $\mathcal{D}$ is unknown except for $(x_n, y_n) \sim \mathcal{D}$ (*learning*)

The constrained learning problem

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \qquad \text{(P-CSL)}$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

- ℓ_0, ℓ_i are bounded, Lipschitz continuous, convex functions
- **Infinite dimensional**
- $\mathcal{D}$ is unknown except for $(x_n, y_n) \sim \mathcal{D}$ (*learning*)

The constrained learning problem

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \tag{P-CSL}$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

- ℓ_0, ℓ_i are bounded, Lipschitz continuous, convex functions
- **Infinite dimensional**
- $\mathcal{D}$ is unknown except for $(\boldsymbol{x}_n, \boldsymbol{y}_n) \sim \mathcal{D}$ (*learning*)

The constrained learning problem

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \qquad \text{(P-CSL)}$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

- ℓ_0, ℓ_i are bounded, Lipschitz continuous, convex functions
- **Infinite dimensional**
- **Cannot evaluate expectations**

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Main questions

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Main questions

Would (P-CSL) enable learning with requirements?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Main questions

Would (P-CSL) enable learning with requirements?

What does it mean to solve (P-CSL)?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Main questions

Would (P-CSL) enable learning with requirements?

What does it mean to solve (P-CSL)?

Can we solve (P-CSL) and how?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Main questions

Would (P-CSL) enable learning with requirements?

What does it mean to solve (P-CSL)?

Can we solve (P-CSL) and how?

Fair classification

Problem

Estimate the probability of an individual making more than US$ 50.000 *based on personal and socio-economical data without discriminating based on gender.*

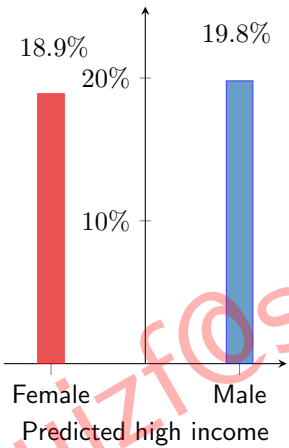
Problem

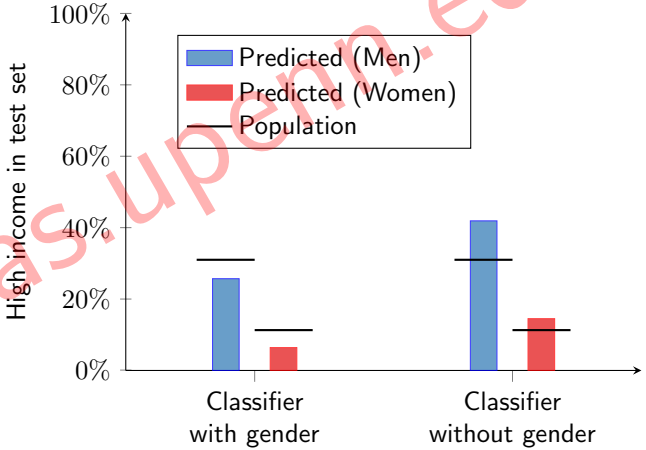
Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}\left[-y \log\left(\phi(\boldsymbol{x}, z)\right)\right]$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $\phi(\boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right]$

Fair classification





Problem

Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}\left[-y \log\left(\phi(\boldsymbol{x}, z)\right)\right]$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $\phi(\boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right]$

Fair classification

Problem

Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

$$\begin{aligned} \underset{\phi \in \mathcal{F}}{\text{minimize}} \quad & \mathbb{E}\left[-y \log\left(\phi(\boldsymbol{x}, z)\right)\right] \\ \text{subject to} \quad & \mathbb{E}\left[D_{\mathsf{KL}}\left(\phi(\boldsymbol{x}, \text{Male}) \,\|\, \phi(\boldsymbol{x}, \text{Female})\right)\right] \leq c \end{aligned}$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $\phi(\boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right]$

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Applications

- **Statistical invariances:** Fairness, robustness (distributional shift), transfer learning, unbalanced data (within-group accuracy)...
- **Learning with constraints:** semi-supervised learning, VAEs, GANs...

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Questions

Would (P-CSL) enable learning with requirements?

What does it mean to solve (P-CSL)?

Can we solve (P-CSL) and how?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Questions

Would (P-CSL) enable learning with requirements? **Yes**

What does it mean to solve (P-CSL)?

Can we solve (P-CSL) and how?

Definition

$$\begin{aligned} P^\star = \min_{\phi \in \mathcal{F}} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \\ \text{subject to} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \end{aligned} \tag{P-CSL}$$

Questions

Would (P-CSL) enable learning with requirements? **Yes**

What does it mean to solve (P-CSL)?

Can we solve (P-CSL) and how?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Challenges

- Infinite dimensional
- Cannot evaluate expectations

What's in a solution?

Definition (Probably approximately optimal)

For $\epsilon_i > 0$, $i = 0, \dots, m$, and $0 \leq \delta < 1/2$, a solution $\phi^\dagger$ of (P-CSL) is said to be *probably approximately optimal* (PAOpt) if

What's in a solution?

Definition (Probably approximately optimal)

For $\epsilon_i > 0$, $i = 0, \dots, m$, and $0 \leq \delta < 1/2$, a solution $\phi^\dagger$ of (P-CSL) is said to be *probably approximately optimal* (PAOpt) if

1) Probably near-optimal

$$\Pr\left[\left|P^\star - \mathbb{E}\left[\ell_0\left(\phi^\dagger(\boldsymbol{x}), y\right)\right]\right| > \epsilon_0\right] \leq \delta$$

What's in a solution?

Definition (Probably approximately optimal)

For $\epsilon_i > 0$, $i = 0, \dots, m$, and $0 \leq \delta < 1/2$, a solution $\phi^\dagger$ of (P-CSL) is said to be *probably approximately optimal* (PAOpt) if

1) Probably near-optimal **($\approx$ PAC learning)**

$$\Pr\left[\left|P^\star - \mathbb{E}\left[\ell_0\left(\phi^\dagger(\boldsymbol{x}), y\right)\right]\right| > \epsilon_0\right] \leq \delta$$

What's in a solution?

Definition (Probably approximately optimal)

For $\epsilon_i > 0$, $i = 0, \dots, m$, and $0 \leq \delta < 1/2$, a solution $\phi^\dagger$ of (P-CSL) is said to be *probably approximately optimal* (PAOpt) if

1) Probably near-optimal **($\approx$ PAC learning)**

$$\Pr\left[\left|P^\star - \mathbb{E}\left[\ell_0\left(\phi^\dagger(\boldsymbol{x}), y\right)\right]\right| > \epsilon_0\right] \leq \delta$$

2) Probably approximately feasible

$$\Pr\left[\left|\mathbb{E}\left[\ell_i\left(\phi^\dagger(\boldsymbol{x}), y\right)\right]\right| > c_i + \epsilon_i\right] \leq \delta$$

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Questions

Would (P-CSL) enable learning with requirements? **Yes**

What does it mean to solve (P-CSL)?

Can we solve (P-CSL) and how?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \qquad \text{(P-CSL)}$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

Questions

Would (P-CSL) enable learning with requirements? **Yes**

What does it mean to solve (P-CSL)? **PAOpt**

Can we solve (P-CSL) and how?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Questions

Would (P-CSL) enable learning with requirements? **Yes**

What does it mean to solve (P-CSL)? **PAOpt**

Can we solve (P-CSL) and how?

How classical learning does it

$$P^\star = \min_{\phi \in \mathcal{F}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$

How classical learning does it

$$P^\star = \min_{\phi \in \mathcal{F}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$

How classical learning does it

$$P^\star = \min_{\phi \in \mathcal{F}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$

$$\downarrow$$

$$\hat{P}^\star_\epsilon = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}), y\right)\right]$$

How classical learning does it

$$P^\star = \min_{\phi \in \mathcal{F}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$

$$\Big| \quad \text{Approximation error} \quad \downarrow$$

$$\hat{P}^\star_\epsilon = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}), y\right)\right]$$

Definition

For each $\phi \in \mathcal{F}$ there exists $\boldsymbol{\theta} \in \mathbb{R}^p$ such that

$$\mathbb{E}\left[\left|f(\boldsymbol{\theta}, \boldsymbol{x}) - \phi(\boldsymbol{x})\right|\right] \leq \nu.$$

How classical learning does it

$$P^\star = \min_{\phi \in \mathcal{F}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$

$$\downarrow$$

$$\hat{P}^\star_\epsilon = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}), y\right)\right]$$

How classical learning does it

$$P^\star = \min_{\phi \in \mathcal{F}} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$

$$\downarrow$$

$$\hat{P}^\star_\epsilon = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \; \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}), y\right)\right]$$

$$\downarrow$$

$$\hat{P}^\star_{\epsilon,N} = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \; \frac{1}{N} \sum_{n=1}^{N} \ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right)$$

How classical learning does it

$$P^\star = \min_{\phi \in \mathcal{F}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0\left(\phi(\boldsymbol{x}), y\right)\right]$$

$$\hat{P}^\star_\epsilon = \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0\left(f(\boldsymbol{\theta}, \boldsymbol{x}), y\right)\right]$$

$$\hat{P}^\star_{\epsilon,N} = \min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{N}\sum_{n=1}^{N} \ell_0\left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right)$$

Theorem (VC learning theory)

Under mild assumptions,

$$P^\star \approx \hat{P}^\star_{\epsilon,N} \text{ with high probability.}$$

An obvious idea that doesn't work

$$
\begin{aligned}
P^\star = \min_{\phi \in \mathcal{F}} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \\
\text{subject to} \quad & \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i
\end{aligned}
\quad \longrightarrow \quad
\begin{aligned}
\hat{P}^\star_{\epsilon,N} = \min_{\boldsymbol{\theta} \in \mathcal{H}} \quad & \frac{1}{N} \sum_{n=1}^{N} \ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) \\
\text{subject to} \quad & \frac{1}{N} \sum_{n=1}^{N} \ell_i \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) \leq c_i
\end{aligned}
$$

An obvious idea that doesn't work

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

$\not\longrightarrow$

$$\hat{P}^\star_{\epsilon,N} = \min_{\boldsymbol{\theta} \in \mathcal{H}} \quad \frac{1}{N} \sum_{n=1}^{N} \ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right)$$
$$\text{subject to} \quad \frac{1}{N} \sum_{n=1}^{N} \ell_i \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) \leq c_i$$

An obvious idea that doesn't work

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

$$\hat{P}^\star_{\epsilon,N} = \min_{\boldsymbol{\theta}\in\mathcal{H}} \quad \frac{1}{N}\sum_{n=1}^{N} \ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right)$$
$$\text{subject to} \quad \frac{1}{N}\sum_{n=1}^{N} \ell_i \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) \leq c_i$$

$$\hat{D}^\star_{\epsilon,N} = \max_{\lambda_i \geq 0} \min_{\boldsymbol{\theta}\in\mathcal{H}} \quad \frac{1}{N}\sum_{n=1}^{N} \left(\ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) + \sum_{i=1}^{m} \lambda_i \left[\ell_i \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) - c_i\right] \right)$$

An obvious idea that doesn't work

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

$$\hat{P}^\star_{\epsilon,N} = \min_{\boldsymbol{\theta}\in\mathcal{H}} \quad \frac{1}{N}\sum_{n=1}^{N} \ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right)$$
$$\text{subject to} \quad \frac{1}{N}\sum_{n=1}^{N} \ell_i \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) \leq c_i$$

$$\hat{D}^\star_{\epsilon,N} = \max_{\lambda_i \geq 0} \min_{\boldsymbol{\theta}\in\mathcal{H}} \quad \frac{1}{N}\sum_{n=1}^{N} \left(\ell_0 \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) + \sum_{i=1}^{m} \lambda_i \left[\ell_i \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n), y_n\right) - c_i\right] \right)$$

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \tilde{O}\left(\nu + \frac{1}{\sqrt{N}}\right) \quad \textit{with high probability}$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + \tilde{O}\left(\frac{1}{\sqrt{N}}\right) \quad \textit{with high probability}$$

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \tilde{O}\left(\nu + \frac{1}{\sqrt{N}}\right) \quad \textit{with high probability}$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + \tilde{O}\left(\frac{1}{\sqrt{N}}\right) \quad \textit{with high probability}$$

Depends on...

parametrization quality (ν) requirements difficulty ($\boldsymbol{\lambda}^\star_p$) sample size ($N$)

Main result

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \left(1 + \left\|\boldsymbol{\lambda}^\star_p\right\|_1\right) L\nu + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \textit{with prob. } 1-\delta$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \textit{with prob. } 1-\delta$$

Depends on...

parametrization quality (ν) requirements difficulty ($\boldsymbol{\lambda}^\star_p$) sample size ($N$)

Main result

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \left(1 + \left\|\boldsymbol{\lambda}^\star_p\right\|_1\right) L\nu + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \textit{with prob. } 1-\delta$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \textit{with prob. } 1-\delta$$

Depends on...

parametrization quality (ν) requirements difficulty ($\boldsymbol{\lambda}^\star_p$) sample size ($N$)

Main result

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \left(1 + \left\|\boldsymbol{\lambda}^\star_p\right\|_1\right) L\nu + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

Depends on...

parametrization quality (ν) requirements difficulty ($\boldsymbol{\lambda}^\star_p$) sample size ($N$)

Main result

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \left(1 + \left\|\boldsymbol{\lambda}^\star_p\right\|_1\right) L\nu + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

Depends on...

parametrization quality (ν) requirements difficulty ($\boldsymbol{\lambda}^\star_p$) sample size ($N$)

Main result

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \left(1 + \left\|\boldsymbol{\lambda}^\star_p\right\|_1\right) L\nu + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

Depends on...

parametrization quality (ν) requirements difficulty ($\boldsymbol{\lambda}^\star_p$) sample size ($N$)

Main result

Theorem

Let f be an ν-parameterization of $\mathcal{F}$. If $(\boldsymbol{\theta}^\dagger, \boldsymbol{\lambda}^\dagger)$ achieve $\hat{D}^\star_{\epsilon,N}$, then $f(\boldsymbol{\theta}^\dagger, \cdot)$ is a probably approximately optimal solution of (P-CSL):

$$\left|D^\star_{\epsilon,N} - P^\star\right| \leq \left(1 + \left\|\boldsymbol{\lambda}^\star_p\right\|_1\right) L\nu + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

$$\mathbb{E}\left[\ell_i(f(\boldsymbol{\theta}^\dagger, \boldsymbol{x}), y)\right] \leq c_i + 2B\sqrt{\frac{1}{N}\left[1 + \log\left(\frac{4(2N)^{d_{VC}}}{\delta}\right)\right]} \quad \text{with prob. } 1-\delta$$

Depends on...

parametrization quality (ν) requirements difficulty ($\boldsymbol{\lambda}^\star_p$) sample size ($N$)

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i \qquad \text{(P-CSL)}$$

Questions

Would (P-CSL) enable learning with requirements? **Yes**

What does it mean to solve (P-CSL)? **PAOpt**

Can we solve (P-CSL) and how?

Definition

$$P^\star = \min_{\phi \in \mathcal{F}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_0 \left(\phi(\boldsymbol{x}), y\right)\right] \qquad \text{(P-CSL)}$$
$$\text{subject to} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\ell_i \left(\phi(\boldsymbol{x}), y\right)\right] \leq c_i$$

Questions

Would (P-CSL) enable learning with requirements? **Yes**

What does it mean to solve (P-CSL)? **PAOpt**

Can we solve (P-CSL) and how? **Yes**

Problem

Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

$$\begin{aligned} \underset{\phi \in \mathcal{F}}{\text{minimize}} \quad & \mathbb{E}\left[-y \log\left(\phi(\boldsymbol{x}, z)\right)\right] \\ \text{subject to} \quad & \mathbb{E}\left[D_{\mathsf{KL}}\left(\phi(\boldsymbol{x}, \text{Male}) \,\|\, \phi(\boldsymbol{x}, \text{Female})\right)\right] \leq c \end{aligned}$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $\phi(\boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right]$

Back to fairness

Problem

Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

$$\max_{\lambda \geq 0} \min_{\boldsymbol{\theta} \in \mathcal{H}} \; -\frac{1}{N} \sum_{n=1}^{N} \left(y_n \log \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, z_n) \right) - \lambda \left[D_{\mathsf{KL}} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Male}) \,\|\, f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Female}) \right) - c \right] \right)$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $f(\boldsymbol{\theta}, \boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right]$

Back to fairness

Problem

Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

$$\max_{\lambda \geq 0} \min_{\boldsymbol{\theta} \in \mathcal{H}} \; -\frac{1}{N} \sum_{n=1}^{N} \left(y_n \log \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, z_n) \right) - \lambda \left[D_{\mathsf{KL}} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Male}) \,\|\, f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Female}) \right) - c \right] \right)$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $f(\boldsymbol{\theta}, \boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right]$

Back to fairness

Problem

Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

$$\max_{\lambda \geq 0} \min_{\boldsymbol{\theta} \in \mathcal{H}} -\frac{1}{N} \sum_{n=1}^{N} \left(y_n \log \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, z_n) \right) - \lambda \left[D_{\mathsf{KL}} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Male}) \,\|\, f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Female}) \right) - c \right] \right)$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $f(\boldsymbol{\theta}, \boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right]$

Back to fairness

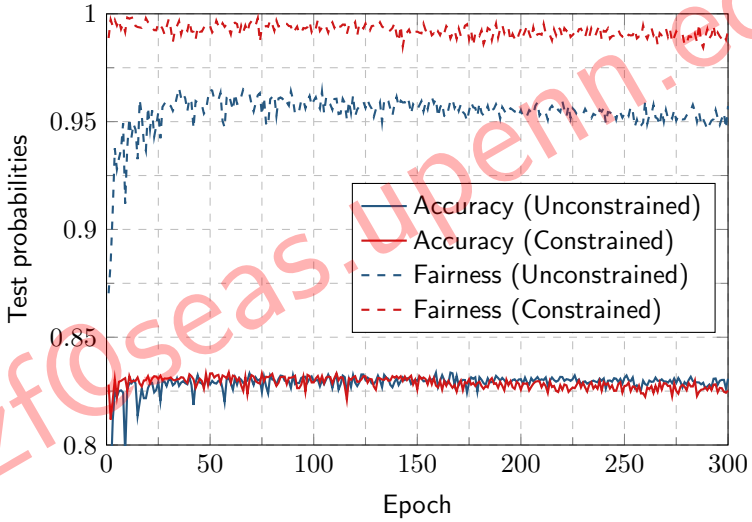
Problem

Estimate the probability of an individual making more than US$ 50.000 based on personal and socio-economical data without discriminating based on gender.

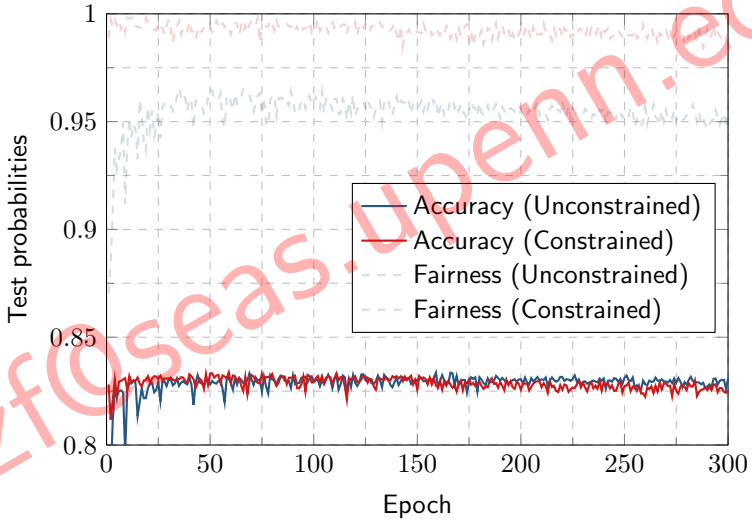
$$\max_{\lambda \geq 0} \min_{\boldsymbol{\theta} \in \mathcal{H}} \; -\frac{1}{N} \sum_{n=1}^{N} \left(y_n \log \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, z_n) \right) - \lambda \left[D_{\mathsf{KL}} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Male}) \,\|\, f(\boldsymbol{\theta}, \boldsymbol{x}_n, \mathsf{Female}) \right) - c \right] \right)$$

- $\boldsymbol{x}$ collects the features (socio-economical data)
- z is the protected variable (gender)
- $f(\boldsymbol{\theta}, \boldsymbol{x}, z) = \Pr\left[\geq \text{US\$ } 50.000\right] \rightarrow$ neural network

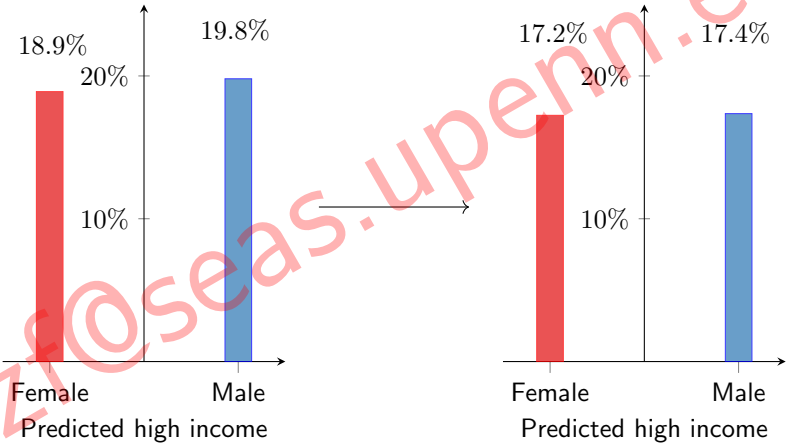
Back to fairness



Back to fairness



Fair classification



Extensions and further applications

Extensions and further applications

- Non-convex losses ℓ_0, ℓ_i (under mild assumptions on $\mathcal{D}$), [future work]

Extensions and further applications

- Non-convex losses ℓ_0, ℓ_i (under mild assumptions on $\mathcal{D}$)
  [future work]

- Primal-dual algorithm convergence
  [future work]

Extensions and further applications

- Non-convex losses ℓ_0, ℓ_i (under mild assumptions on $\mathcal{D}$)
  [future work]

- Primal-dual algorithm convergence
  [future work]

- Constrained (e.g., safe) reinforcement learning
  [P**C**FR, "Constrained Reinforcement Learning Has Zero Duality Gap." *NeurIPS*, 2019]
  [PF**C**R, "Safe Policies for Reinforcement Learning via Primal-Dual Methods." *ArXiv*, 2019]

Conclusion and extensions

(1) **Does constrained learning enable learning with requirements?**

(2) **What does it mean to solve a constrained learning problem?**

(3) **Can we solve (P-CSL) and how?**

Conclusion and extensions

(1) **Does constrained learning enable learning with requirements?**
Yes, e.g., fair classification.

(2) **What does it mean to solve a constrained learning problem?**

(3) **Can we solve (P-CSL) and how?**

Conclusion and extensions

(1) **Does constrained learning enable learning with requirements?**
Yes, e.g., fair classification.

(2) **What does it mean to solve a constrained learning problem?**
Obtaining a probably approximately optimal solution.

(3) **Can we solve (P-CSL) and how?**

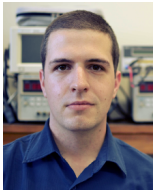
Conclusion and extensions

(1) **Does constrained learning enable learning with requirements?**
Yes, e.g., fair classification.

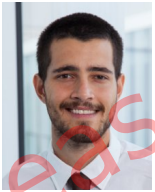
(2) **What does it mean to solve a constrained learning problem?**
Obtaining a probably approximately optimal solution.

(3) **Can we solve (P-CSL) and how?**
Yes, using their empirical dual and the approximation depends on the sample size, the model complexity, and the requirements' difficulty.

CONSTRAINED LEARNING



L.F.O. Chamon



S. Paternain

M. Calvo-Fullana

A. Ribeiro

http://www.seas.upenn.edu/∼luizf
luizf@seas.upenn.edu