PROBABLY APPROXIMATELY CORRECT CONSTRAINED LEARNING

Luiz F. O. Chamon and Alejandro Ribeiro

Electrical and Systems Engineering, University of Pennsylvania

e-mail: luizf@seas.upenn.edu, aribeiro@seas.upenn.edu

Penn UNIVERSITY of PENNSYLVANIA

CONTRIBUTIONS

1. *What does it mean to learn under constraints?*
   Define PAC constrained (PACC) learning
2. *When (if at all) is it possible to learn under constraints?*
   Constrained learning is as hard as unconstrained learning (PACC $\Leftrightarrow$ PAC)
3. *Is there a practical constrained learning rule?*
   Dual empirical learning and primal-dual algorithms (under mild conditions)

INTRODUCTION

Learning is ubiquitous, but has serious shortcomings

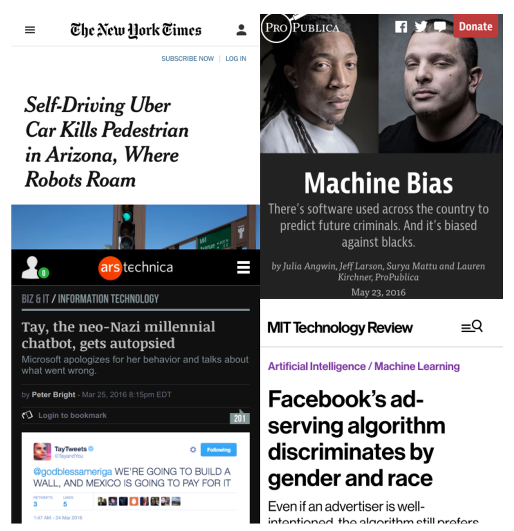
SIEMENS

Predictive maintenance

citi PayPal

Fraud detection

"Smart" applications



FUJITSU

Quality control

Pfizer

Drug discovery

Problem (Constrained learning)

Train a CNN that is robust to input perturbations

$$\begin{aligned}&\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{minimize}} && \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\Big[\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)\Big]\\ &\text{subject to} && \mathbb{E}_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\mathfrak{A}}\Big[\ell\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}),\tilde{y}\big)\Big]\leq c\end{aligned} \quad \approx \quad ?$$

Unconstrained alternatives

- **Learning** (PAC learning theory)

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\Big[\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)\Big]\approx\frac{1}{N}\sum_{n=1}^{N}\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n),y_n\big)$$

  - May not satisfy the requirements

- **Regularized learning** (PAC learning theory)

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{minimize}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\Big[\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)\Big]+\lambda\,\mathbb{E}_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\mathfrak{A}}\Big[\ell\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}),\tilde{y}\big)\Big]$$

$$\approx\frac{1}{N}\sum_{n=1}^{N}\Big[\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n),y_n\big)+\lambda\,\ell\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_n),\tilde{y}_n\big)\Big]$$

  - For what λ is the solution feasible? Does it generalize or is it dataset-dependent?

CONSTRAINED LEARNING THEORY

1. What is constrained learning?

Definition (PACC learnability)

$\mathcal{H}$ is PACC learnable if for every ϵ,δ and every distributions $\mathfrak{D}_k$, we can obtain $f_{\boldsymbol{\theta}^\dagger}\in\mathcal{H}$ from $N_{\mathcal{H}}(\epsilon,\delta)$ samples that is, with probability $1-\delta$,

near-optimal ($\Rightarrow$ PAC learning)

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{D}}\Big[\ell\big(f_{\boldsymbol{\theta}^\dagger}(\boldsymbol{x}),y\big)\Big]\leq P^\star+\epsilon$$

approximately feasible

$$\mathbb{E}_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\mathfrak{A}}\Big[\ell\big(f_{\boldsymbol{\theta}^\dagger}(\tilde{\boldsymbol{x}}),\tilde{y}\big)\Big]\leq c+\epsilon$$

2. When (if at all) is it possible to learn under constraints?

Theorem 1

$\mathcal{H}$ is PACC learnable $\Leftrightarrow$ $\mathcal{H}$ is PAC learnable

3. Is there a practical constrained learning rule?

Theorem 2

Let f be ν-universal, i.e., for each $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\gamma\in[0,1]$ there exists $\boldsymbol{\theta}$ such that $\mathbb{E}\Big[\big|\gamma f_{\boldsymbol{\theta}_1}(\boldsymbol{x})+(1-\gamma)f_{\boldsymbol{\theta}_2}(\boldsymbol{x})-f_{\boldsymbol{\theta}}(\boldsymbol{x})\big|\Big]\leq\nu$ and ℓ be convex, bounded, and M-Lipschitz continuous. Then $\hat{D}^\star$ is a (near-)PACC learner, i.e., if $\boldsymbol{\theta}^\dagger$ achieves $\hat{D}^\star$, then with probability $1-\delta$,

$$\mathbb{E}\Big[\ell\big(f_{\boldsymbol{\theta}^\dagger}(\boldsymbol{x}),y\big)\Big]\leq P^\star+\epsilon_0+\epsilon \qquad \mathbb{E}\Big[\ell\big(f(\boldsymbol{\theta}^\dagger,\tilde{\boldsymbol{x}}),\tilde{y}\big)\Big]\leq c+\epsilon$$

$$\epsilon_0=\Big(2+\big\|\boldsymbol{\lambda}_p^\star\big\|_1\Big)M\nu \qquad \epsilon=B\sqrt{\frac{1}{N}\bigg[1+\log\bigg(\frac{4(m+2)(2N)^{d_{\text{VC}}}}{\delta}\bigg)\bigg]}$$

A primal-dual algorithm

Minimize the primal: $$\boldsymbol{\theta}^+\approx\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{argmin}}\ \frac{1}{N}\sum_{n=1}^{N}\Big[\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n),y_n\big)+\lambda\ell\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_n),\tilde{y}_n\big)\Big]$$

$$=\boldsymbol{\theta}-\eta\left[\nabla_{\boldsymbol{\theta}}\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n),y_n\big)+\lambda\nabla_{\boldsymbol{\theta}}\ell\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_n),\tilde{y}_n\big)\right]$$

Update the dual: $$\lambda^+=\left[\lambda+\eta\left(\frac{1}{N}\sum_{n=1}^{N}\ell\big(f_{\boldsymbol{\theta}^+}(\tilde{\boldsymbol{x}}_n),\tilde{y}_n\big)-c\right)\right]_+$$

Theorem 3

$\boldsymbol{\theta}^+$ is a ρ-approximate minimizer $\Rightarrow$ converges to a ρ-neighborhood of $\hat{D}^\star_\theta$ (under mild conditions)

MAIN RESULTS

$$P^\star=\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\ \mathbb{E}\Big[\ell_0\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)\Big] \quad \text{s. to}\ \mathbb{E}\Big[\ell_i\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}),\tilde{y}\big)\Big]\leq c$$

$$\xleftrightarrow{\text{Thm. 1}}\quad \min_{\boldsymbol{\theta}\in\mathbb{R}^p}\ \frac{1}{N}\sum_{n=1}^{N}\ell_0\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n),y_n\big) \quad \text{s. to}\ \frac{1}{N}\sum_{n=1}^{N}\ell_i\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_n),\tilde{y}_n\big)\leq c$$

Thm. 2

$$\hat{D}^\star=\max_{\lambda\geq0}\ \min_{\boldsymbol{\theta}\in\mathbb{R}^p}\ \frac{1}{N}\sum_{n=1}^{N}\Big(\ell_0\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n),y_n\big)+\lambda\big[\ell\big(f_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}_n),\tilde{y}_n\big)-c\big]\Big)$$

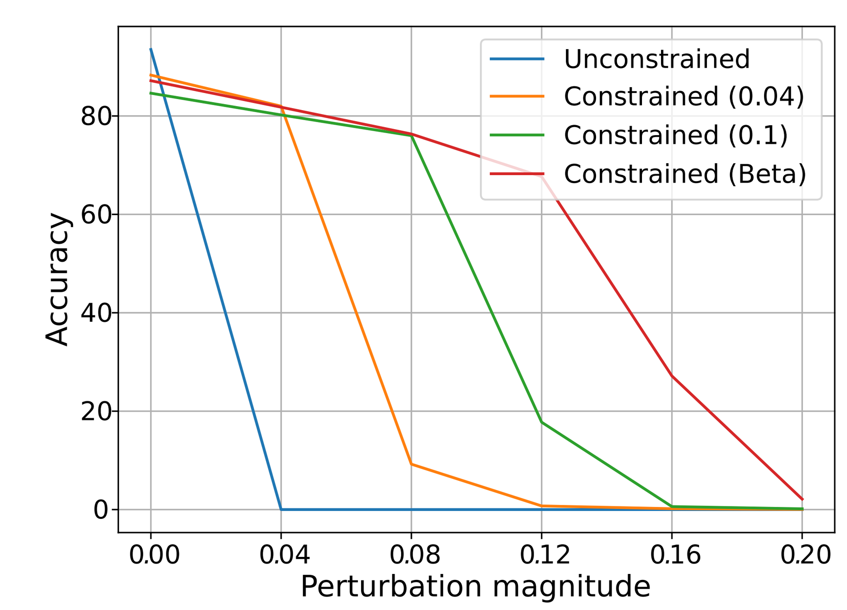
In the paper...

- Multiple constraints, different losses
- Pointwise constraints: $\ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)\leq c$, $\mathfrak{D}$-a.e.

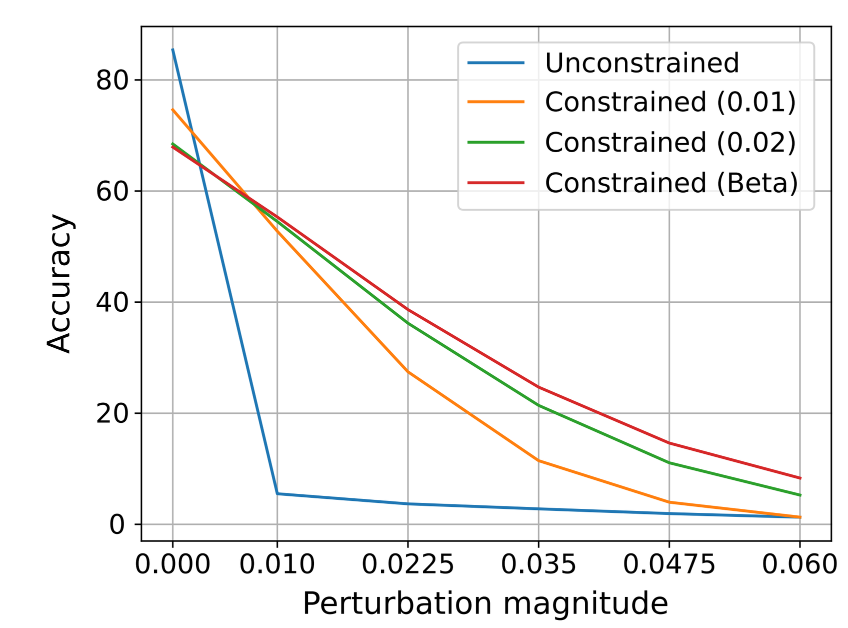
APPLICATION

Robustness

FMNIST

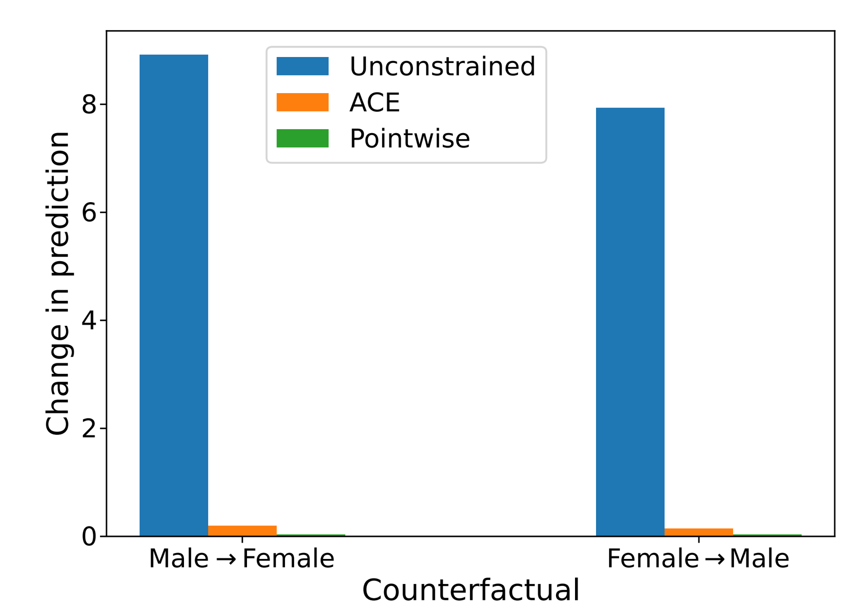


CIFAR-10



And also...

Fairness



Safe RL [PCCR, NeurIPS'19]