Learning via non-convex min-max optimization

Meisam Razaviyayn

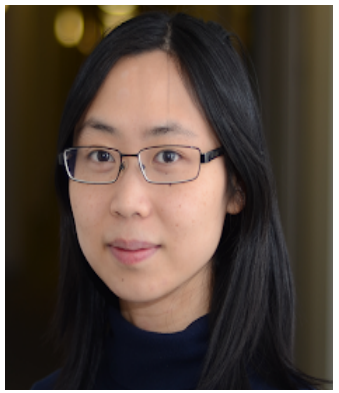
University of Southern California

Maher Nouiehed

USC → AUB

Tianjian Haung
USC

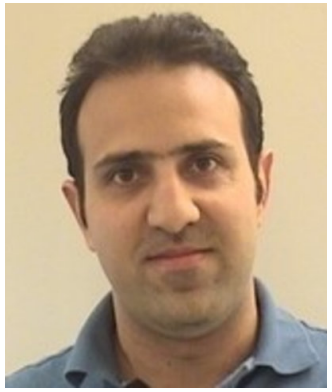
Maziar Sanjabi
Facebook AI



Sze-chuan Suen
USC

Andrew Lowy

USC

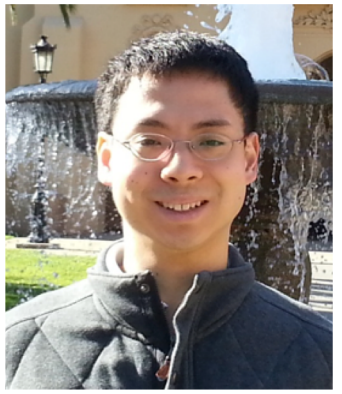


Ahmad Beirami
Facebook AI

Sina Baharlouei
USC

Babak Barazandeh

USC → Splunk

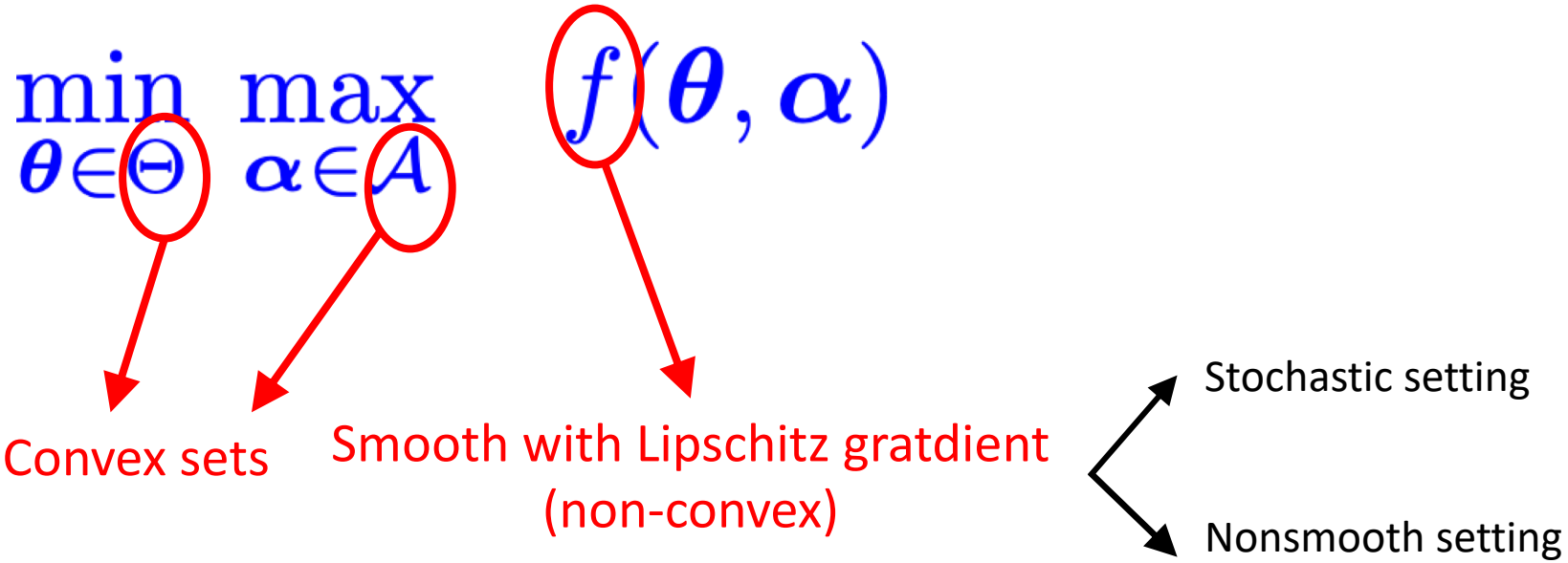


Jason Lee

Princeton

Dmitrii Ostrovskii
USC

Non-convex min-max games/optimizations



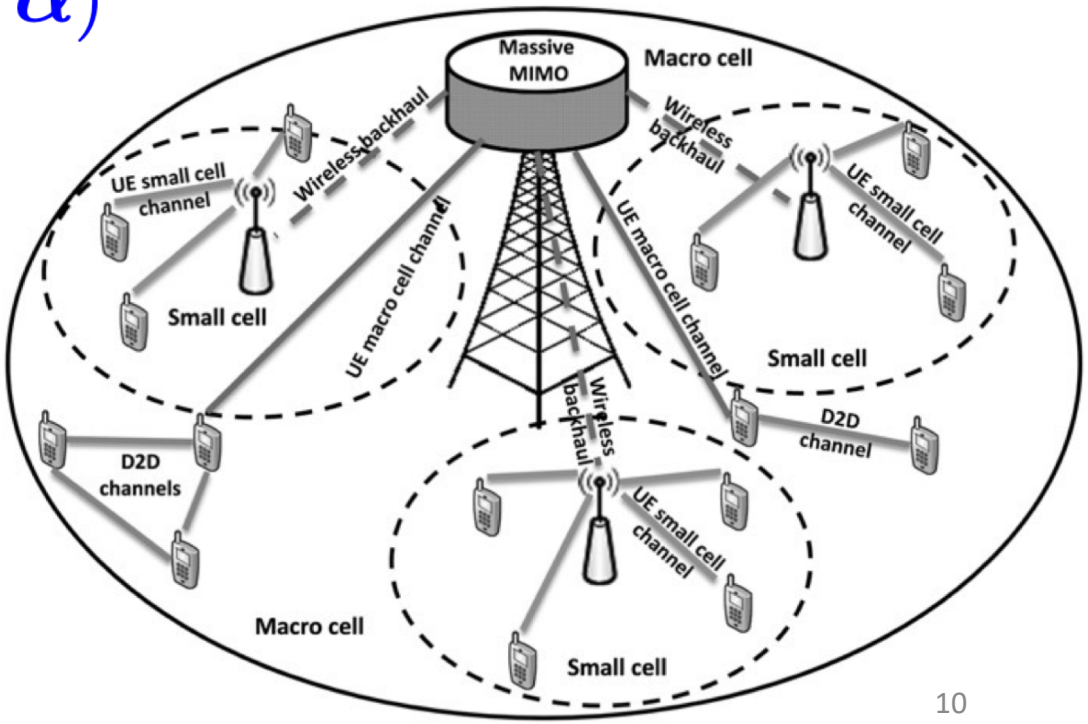
- Why is this problem important? Applications?
- Why is it challenging?
- Some algorithms and discussions

- ✓ $f(\theta,\alpha)$ is (strongly) concave in α
- ✓ Small coupling between two variables: $\nabla^2_{\theta\alpha} f(\theta,\alpha)$ is small
- ✓ The radius of $\mathcal{A}$ is small

Application 1: Min-max problems and robustness

- Design a system with a robust performance against changes in certain parameters
- Design for nominal value: $\min_{\boldsymbol{\theta}\in\Theta} \quad f(\boldsymbol{\theta}, \boldsymbol{\alpha}_0)$
- Robust design: $\min_{\boldsymbol{\theta}\in\Theta} \quad \max_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|\leq\delta} \quad f(\boldsymbol{\theta}, \boldsymbol{\alpha})$
- Massive MIMO application

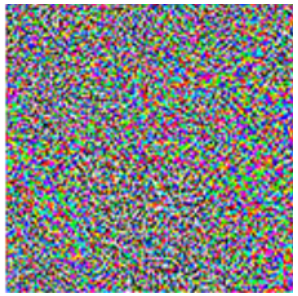
$$\min_{\mathbf{w}} \quad \max_{\mathbf{H}\in\mathcal{H}} \quad \ell(\mathbf{w}, \mathbf{H})$$



Application 1: Min-max problems and robustness

- Adversarial attacks to neural networks

$+\,0.007\times$



$=$

"panda"
57.7% confidence

"gibon"
99.3% confidence

$$\min_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta},\boldsymbol{\alpha}_0) \longrightarrow \min_{\boldsymbol{\theta}\in\Theta} \max_{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|\le\delta} f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

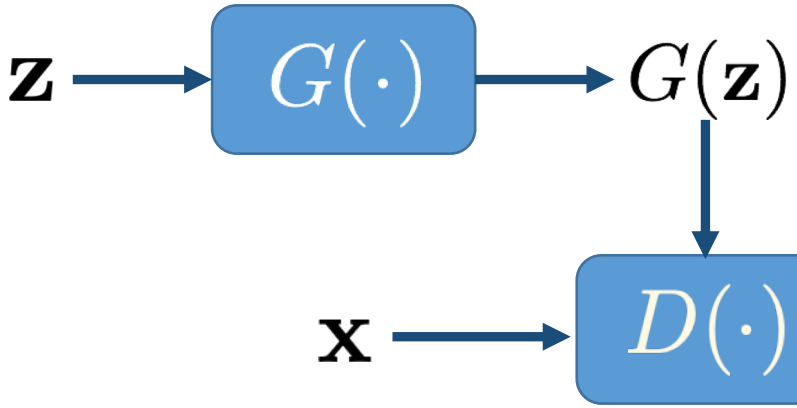
Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.

Application 2: Min-max and GANs

Goal: Generate samples that look like real samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathbb{P}_x$

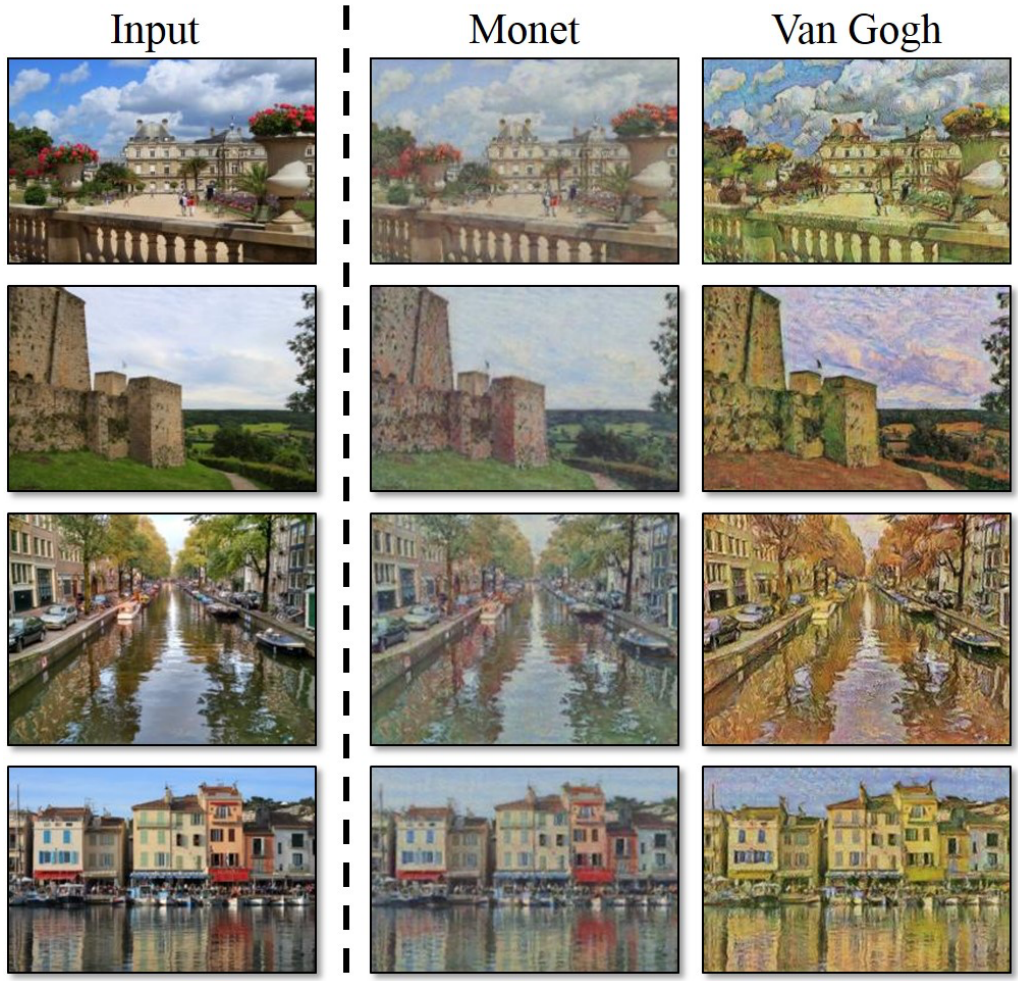
Neural Network

We need $G(\mathbf{z})$ to have the same distribution as $\mathbb{P}_x$



Goodfellow et al 2014

Are they different?



https://junyanz.github.io/CycleGAN/

- The two neural networks are playing a zero-sum game

Application 2: Min-max and GANs

- MMD GANs

$$\min_{G} \max_{D} \left\| \mathbb{E}[D(G(\mathbf{z}))] - \mathbb{E}[D(\mathbf{x})] \right\|$$

- Jensen-Shannon GANs:

$$\min_{G} \max_{D \in \mathbb{D}} \mathbb{E}_{\mathbf{x}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}} \log (1 - D(G(\mathbf{z})))$$

$\mathbb{D} =$ set of all functions with range $(0, 1)$

- Wasserstein GANs:

$$\min_{G} \max_{\gamma} \mathbb{E}_{\mathbf{x}} [\gamma(\mathbf{x})] - \mathbb{E}_{\mathbf{z}} [\gamma(G(\mathbf{z}))]$$

$$\text{s.t.} \quad \gamma(\mathbf{x}) - \gamma(\mathbf{y}) \leq \|\mathbf{x} - \mathbf{y}\|_2, \forall \mathbf{x}, \mathbf{y}$$

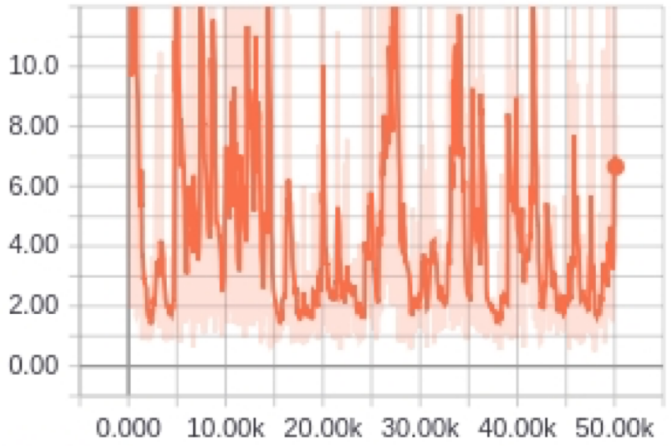
All are non-convex min-max problems!

Why are non-convex min-max problems challenging?

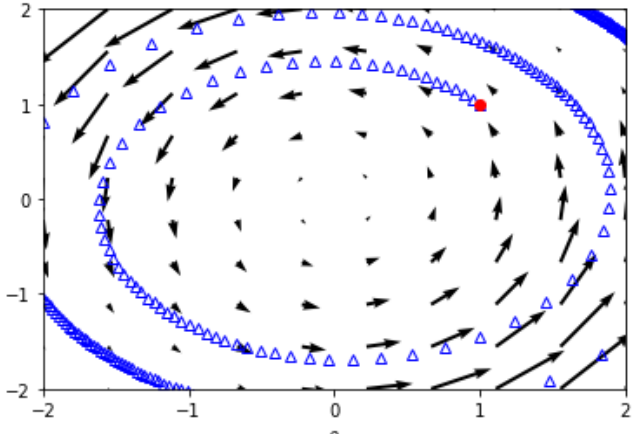
$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta}, \boldsymbol{\alpha})$$

- What should we do? Gradient descent/ascent?

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t, \boldsymbol{\alpha}^t)$$
$$\boldsymbol{\alpha}^{t+1} = \boldsymbol{\alpha}^t + \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t, \boldsymbol{\alpha}^t)$$



Iterates trajectory for $f(\theta, \alpha) = \theta\alpha$



$$\min_{\boldsymbol{\beta}\in\mathcal{B}} h(\boldsymbol{\beta})$$

- Apply (projected) gradient descent:
  - Objective function improves over iterates
  - It is not exhaustive search
  - Convergence to certain stationarity concepts
  - Iteration complexity lower- and upper- bounds

- Even more: what should we compute?

What is a reasonable solution?

$$\min_{\boldsymbol{\theta}\in\Theta}\max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

Bilevel/Stackelberg viewpoint

$$\min_{\boldsymbol{\theta}\in\Theta}\left(g(\boldsymbol{\theta}) \triangleq \max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta},\boldsymbol{\alpha})\right)$$

Find a θ^* which is a reasonable/stationary point of $g(\cdot)$

Find a $\epsilon-$ stationary point of $g(\cdot)$

$$\|\nabla \tilde{g}(\boldsymbol{\theta})\| \leq \epsilon$$

Nash viewpoint

$$\min_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta},\boldsymbol{\alpha})$$
$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

Find a point$(\bar{\theta},\bar{\alpha})$ so that $\bar{\theta}$ and $\bar{\alpha}$ are stationary/reasonable points of their own utility

Find a $(\epsilon_\theta,\epsilon_\alpha)-$ First-order Nash equilibrium of $f(\cdot,\cdot)$

$$\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta},\boldsymbol{\alpha})\| \leq \epsilon_{\boldsymbol{\theta}} \qquad \|\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\theta},\boldsymbol{\alpha})\| \leq \epsilon_{\boldsymbol{\alpha}}$$

What is a reasonable solution?

Bilevel/Stackelberg viewpoint

$$\min_{\boldsymbol{\theta}\in\Theta} \left(g(\boldsymbol{\theta}) \triangleq \max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta},\boldsymbol{\alpha}) \right)$$

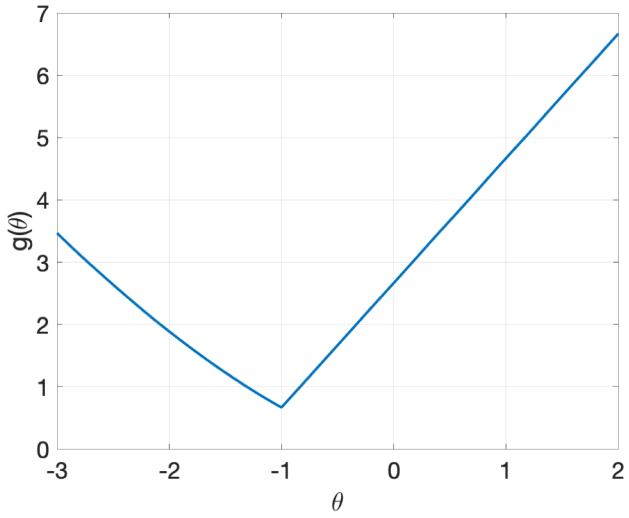
Find a θ^* which is a reasonable/stationary point of $g(\cdot)$

Nash viewpoint

$$\min_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta},\boldsymbol{\alpha}) \qquad \max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

Find a point$(\bar{\theta},\bar{\alpha})$ so that $\bar{\theta}$ and $\bar{\alpha}$ are stationary/reasonable points of their own utility

- The two solution concepts are generally different:
  - Example: $\min_{\theta} \max_{-2\le\alpha\le 2} \theta\alpha + \frac{1}{3}\alpha^3$
  - Bilevel optimal solution: $\theta^* = -1$
  - First-order Nash equilibrium: $(\bar{\theta},\bar{\alpha}) = (0,0)$



- In some cases (e.g. nonconvex-concave setting) the two solution concepts coincide

Iteration complexity

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad f(\boldsymbol{\theta}, \boldsymbol{\alpha})$$

for $t = 1, 2, \ldots$ do

$$\boldsymbol{\alpha}^{t+1} \approx \arg\max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad f(\boldsymbol{\theta}^t, \boldsymbol{\alpha})$$

$$\boldsymbol{\theta}^{t+1} = \left[\boldsymbol{\theta}^t - \gamma \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t, \boldsymbol{\alpha}^{t+1})\right]_+$$

Apply K steps of projected gradient ascent on $\boldsymbol{\alpha}$

$$K \approx \mathcal{O}(\log(\epsilon^{-1}))$$

Need $\mathcal{O}(\epsilon^{-2})$ iterations on $\boldsymbol{\theta}$

Theorem [Nouiehed, Huang, Sanjabi, Lee, Razaviyayn 2018]: Assume $f(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is strongly concave in $\boldsymbol{\alpha}$. Then, the algorithm requires $O(\epsilon^{-2} \log \epsilon^{-1})$ gradient evaluations for computing ϵ −stationary.

- Optimal rate up to logarithmic factors
- Can be obtained under Polyak-Łojasiewicz (PL) condition
  - Requires establishing Danskin's-type result under PL assumption

Strongly convex composite with affine ✓

Relaxing the strong convexity assumption?

Non-convex-concave scenario

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta},\boldsymbol{\alpha}) \quad \Rightarrow \quad \min_{\boldsymbol{\theta}\in\Theta} g(\boldsymbol{\theta})$$

- Assume $f(\boldsymbol{\theta},\boldsymbol{\alpha})$ is concave in $\boldsymbol{\alpha}$ (but not strongly concave)
- $g(\cdot)$ is no longer differentiable
- Smoothify $g(\cdot)$ $\quad g_\lambda(\boldsymbol{\theta}) \triangleq \max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta},\boldsymbol{\alpha}) - \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2$
- **Algorithm:**

$$\boldsymbol{\theta}^{t+1} \approx \left[\boldsymbol{\theta}^t - \gamma \nabla_{\boldsymbol{\theta}} g_\lambda(\boldsymbol{\theta}^t)\right]_+$$

Danskin's Theorem →

for $t = 1, 2, \ldots$ do

$$\boldsymbol{\alpha}^{t+1} \approx \arg\max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta}^t,\boldsymbol{\alpha}) - \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2$$

$$\boldsymbol{\theta}^{t+1} = \left[\boldsymbol{\theta}^t - \gamma \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t, \boldsymbol{\alpha}^{t+1})\right]_+$$

Iteration complexity

$$\min_{\boldsymbol{\theta}\in\Theta}\ \max_{\boldsymbol{\alpha}\in\mathcal{A}}\ f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

Algorithm:

for $t = 1, 2, \ldots$ do

$$\boldsymbol{\alpha}^{t+1} \approx \arg\max_{\boldsymbol{\alpha}\in\mathcal{A}}\ f(\boldsymbol{\theta}^t, \boldsymbol{\alpha}) - \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2$$

$$\boldsymbol{\theta}^{t+1} = \left[\boldsymbol{\theta}^t - \gamma\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t, \boldsymbol{\alpha}^{t+1})\right]_+$$

→ Solve with Nesterov's algorithm

Theorem [Nouiehed, Huang, Sanjabi, Lee, Razaviyayn 2018]: Assume $f(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is concave in $\boldsymbol{\alpha}$. Then, the above algorithm requires

- $O(\epsilon^{-3.5}\log\epsilon^{-1})$ gradient evaluations for computing (ϵ, ϵ) –first order NE (Nash viewpoint)
- $O(\epsilon^{-4})$ gradient evaluations for computing ϵ – stationary point (Stackelberg viewpoint)

- Why does it become so much slower compared to the nonconvex-strongly concave setting?

Why do we observe significant rate drop?

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

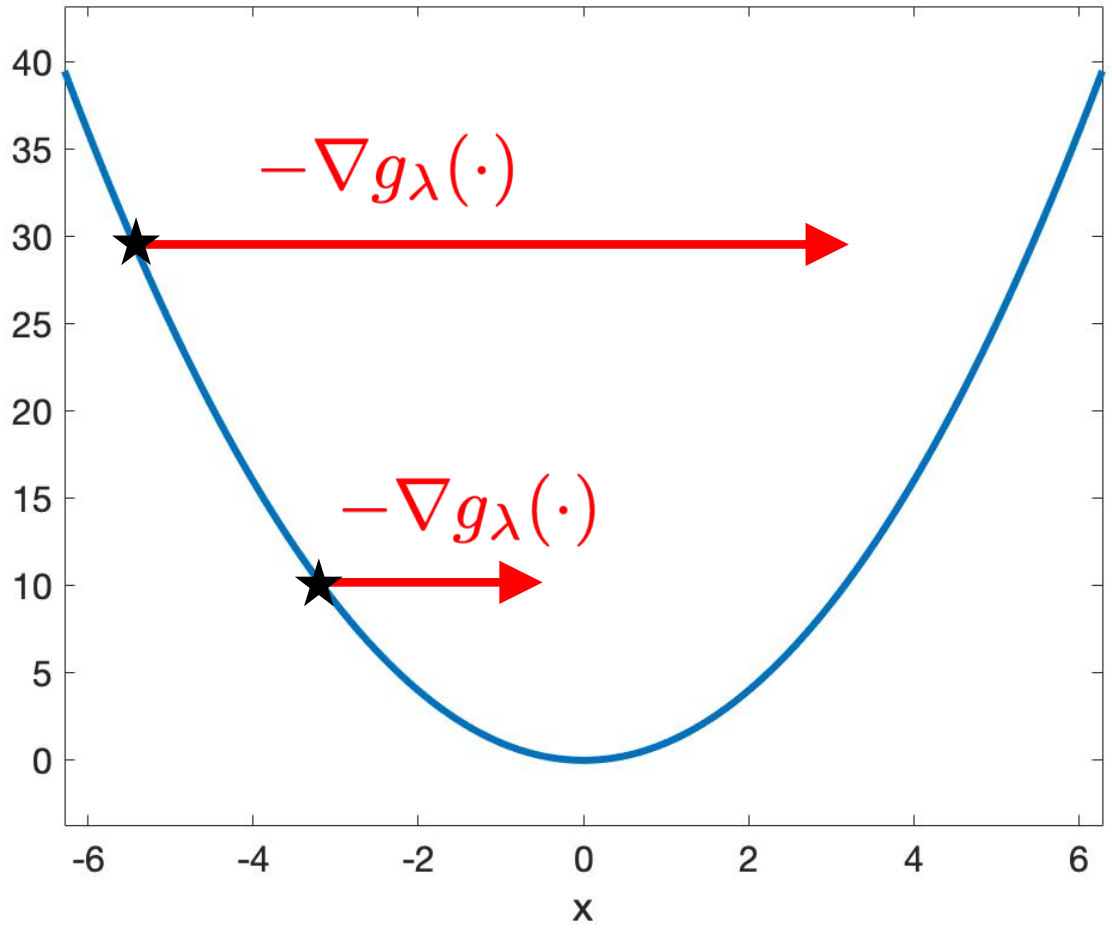
$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\alpha}} \; f(\boldsymbol{\theta},\boldsymbol{\alpha}) - \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2 \qquad \lambda \approx \epsilon$$

$$\min_{\boldsymbol{\theta}} \; g_\lambda(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \nabla g_\lambda(\boldsymbol{\theta}^t) \qquad \eta \approx \epsilon$$

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \nabla g_\lambda(\boldsymbol{\theta}^{t+1})$$

$$\boldsymbol{\theta}^{t+1} = \arg\min_{\boldsymbol{\theta}\in\Theta} \; g_\lambda(\boldsymbol{\theta}) + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2$$



Iteration complexity

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\alpha}} \; f(\boldsymbol{\theta}, \boldsymbol{\alpha}) - \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2$$

$$\min_{\boldsymbol{\theta}} \; g_\lambda(\boldsymbol{\theta})$$

Proximal Point Algorithm:

$$\boldsymbol{\theta}^{t+1} = \arg\min_{\boldsymbol{\theta}\in\Theta} \; g_\lambda(\boldsymbol{\theta}) + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2$$

$$(\boldsymbol{\theta}^{t+1}, \boldsymbol{\alpha}^{t+1}) \approx \arg\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \; f(\boldsymbol{\theta}, \boldsymbol{\alpha}) - \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2 + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2$$

$$= \arg\max_{\boldsymbol{\alpha}\in\mathcal{A}} \min_{\boldsymbol{\theta}\in\Theta} \; f(\boldsymbol{\theta}, \boldsymbol{\alpha}) - \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2 + \frac{1}{2\eta}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2$$

Theorem [Ostrovskii, Lowy, and Razaviyayn 2020]: Assume $f(\boldsymbol{\theta}, \boldsymbol{\alpha})$ is concave in $\boldsymbol{\alpha}$. Then, the above algorithm requires $O(\epsilon^{-2.5}\log\epsilon^{-1})$ gradient evaluations for computing ϵ −first-order NE.

[Lin, Jin, and Jordan 2020]

What can we do in nonconvex-nonconcave regime?

Relation to closely related works

$$\min_{\boldsymbol{\theta}\in\Theta}\ \max_{\boldsymbol{\alpha}\in\mathcal{A}}\ f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

[Thekumparampil-Jain-Netrapalli 2019]

- Only for the unconstrained case of $\Theta = R^d$
- A bit more complex (extra-gradient + Nesterov)
- Does not work for non-Euclidean proximal geometries
- Weaker stationary notion (Slower rate of convergence)

[Lin-Jin-Jordan 2020]

- Going from Nash to Stackelberg is for unconstrained setting $\Theta = R^d$
- Does not work for non-Euclidean proximal geometries
- Weaker stationary notion (Slower rate of convergence)

Difference in stationary notion: $\min_{z\in\mathcal{Z}} h(z)$

$\epsilon-$ stationary of the first type: $\ell\left\|P_{\mathcal{Z}}\left(\bar{z}-\frac{1}{\ell}\nabla h(\bar{z})\right)-\bar{z}\right\|\leq\epsilon$ (1)

$\epsilon-$ stationary of the second type: $-2\ell\min_{z\in\mathcal{Z}}\left[\langle\nabla h(\bar{z}),z-\bar{z}\rangle+\frac{\ell}{2}\|z-\bar{z}\|^2\right]\leq\epsilon^2$ (2)

Theorem: $\epsilon-$ stationary of the second type is a stronger notion, i.e., if a point satisfies (2), it also satisfies (1). Moreover, there exists a problem for which a given feasible point $\bar{\boldsymbol{z}}$ is $\epsilon-$stationary point of the first type, but it is not $\epsilon'-$stationary point of the second type for any $\epsilon' < \sqrt{\epsilon}$.

[Kong-Monteiro 2019],[Zhao 2020]

Can we go beyond nonconvex-concave setting?

Two other relatively easy cases

$$\min_{\theta\in\Theta}\ \max_{\alpha\in\mathcal{A}}\quad f(\theta,\alpha)$$

Example 1: $\min_{\theta}\max_{\alpha} f_1(\theta)+f_2(\alpha)$

Example 2: $\min_{\theta}\ \max_{\alpha\in\{\alpha_0\}} f(\theta,\alpha)$

Theorem [Ostrovskii, Barazandeh, and Razaviyayn 2021]: When $\min\left\{L_{\theta\alpha}D_\alpha, D_\alpha\sqrt{L_{\alpha\alpha}L_{\theta\theta}}\right\}\lesssim\epsilon$

Then instead of solving $\min_{\theta\in\Theta}\ \max_{\alpha\in\mathcal{A}}\ f(\theta,\alpha)$, we can solve

$$\min_{\theta\in\Theta}\ \max_{\alpha\in\mathcal{A}}\ f(\theta,\alpha_0)+\langle\nabla_\alpha f(\theta,\alpha_0),\alpha-\alpha_0\rangle$$ Concave in α

- Resulting in efficient algorithms when radius of $\mathcal{A}$ is small, or when the coupling is small.
- Can be generalized to higher-order approximations and lead to efficient algorithms
  - Ostrovskii, Barazandeh, and Razaviyayn, "Nonconvex-Nonconcave Min-Max Optimization with a Small Maximization Domain," arXiv 2110.03950, *2021*.
- Application: Defense against adversarial attacks in neural networks

Extensions to zeroth order methods

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} f(\boldsymbol{\theta}, \boldsymbol{\alpha})$$

Zhongruo Wang
UC Davis

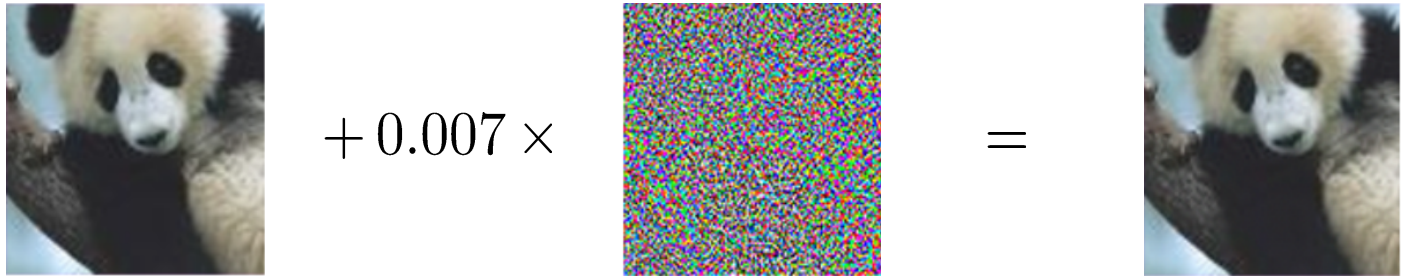
Krishnakumar Balasubramanian
UC Davis

Shiqian Ma
UC Davis

Z. Wang, K. Balasubramanian, S. Ma, and M. Razaviyayn, "Zeroth-Order Algorithms for Nonconvex Minimax Problems with Improved Complexities," *arXiv preprint arXiv:2001.07819, 2020*

Are these results/algorithms useful in practice?

Training robust neural networks



$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{x}_i) \quad \longrightarrow \quad \min_{\mathbf{w}} \sum_{i=1}^{n} \max_{\|\boldsymbol{\delta}\| \le \epsilon} \ell(\mathbf{w}, \mathbf{x}_i + \boldsymbol{\delta})$$

[Madry et al. 2017]: **Repeat**:

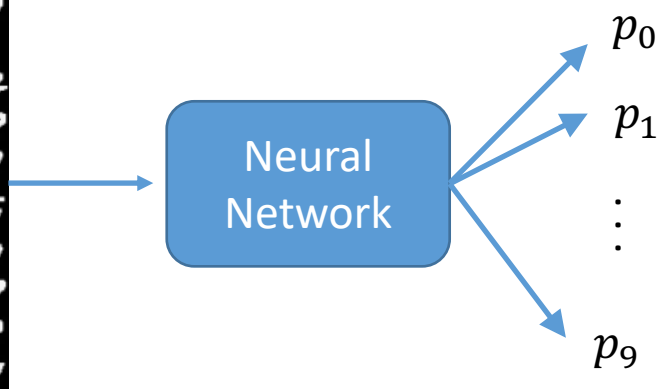
- Apply multi-steps of gradient ascent on $\boldsymbol{\delta}$ (reinitialize multiple times and pick the best)
- Perform one step of gradient descent on $\mathbf{w}$

- No theoretical convergence guarantee, not scalable, and requires heavy tuning to work
- Can we apply our theory and algorithm?

Training robust neural networks

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \max_{\|\boldsymbol{\delta}\| \le \epsilon} \ell(\mathbf{w}, \mathbf{x}_i + \boldsymbol{\delta})$$

- Idea: approximate the maximization with a concave function



$$\min_{\mathbf{w}} \sum_{i=1}^{n} \max \left\{ \ell\big(\mathbf{w}, \mathbf{x}_i + d_0(\mathbf{w}, \mathbf{x}_i)\big), \ldots, \ell\big(\mathbf{w}, \mathbf{x}_i + d_9(\mathbf{w}, \mathbf{x}_i)\big) \right\}$$

$$\nabla_{\mathbf{x}} p_0(\mathbf{w}, \mathbf{x}_i) - \nabla_{\mathbf{x}} p_c(\mathbf{w}, \mathbf{x}_i) \qquad \nabla_{\mathbf{x}} p_9(\mathbf{w}, \mathbf{x}_i) - \nabla_{\mathbf{x}} p_c(\mathbf{w}, \mathbf{x}_i)$$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \left[\max_{\mathbf{t} \in \mathcal{P}} \sum_{k=0}^{9} t_k \ell\big(\mathbf{w}, \mathbf{x}_i + d_k(\mathbf{w}, \mathbf{x}_i)\big) \right]$$

Non-convex in $\boldsymbol{w}$, but concave in $\boldsymbol{t}$

Numerical results

[1] Madry et al. "Towards deep learning models resistant to adversarial attacks." *ICLR 2017*

[2] Zhang et al. “Theoretically principled trade-o between robustness and accuracy” *ICML 2019*.

→ No theoretical convergence guarantee

	Regular Performance	Performance under FGSM attack			Performance under PGD attack		
---	----	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$
[1]	**98.58%**	96.09%	94.82%	89.84%	94.64%	91.41%	78.67%
[2]	97.21%	96.19%	96.17%	96.14%	95.01%	94.36%	94.11%
Proposed	98.20%	**97.04%**	**96.66%**	**96.23%**	**96.00%**	**95.17%**	**94.22%**

FGSM attack: Goodfellow, Shlens, and Szegedy, “Explaining and harnessing adversarial examples,” *arXiv:1412.6572* (2014).

PGD attack: Kurakin, Goodfellow, and Bengio, “Adversarial Machine Learning” at Scale, ICLR 2016.

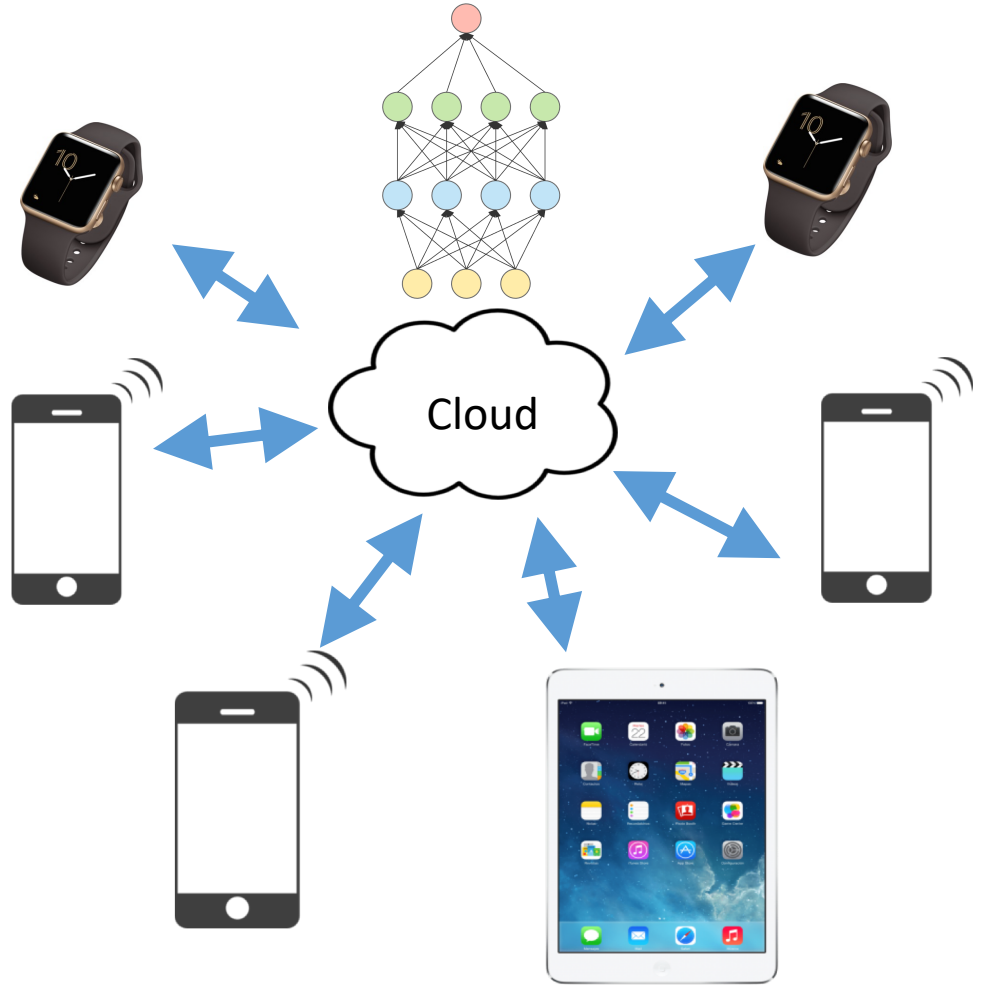
Min-max and fairness among users in learning

- Designing a machine learning model that works for everyone

$$\min_{\mathbf{w}} \max\{\ell_1(\mathbf{w}), \ldots, \ell_k(\mathbf{w})\}$$

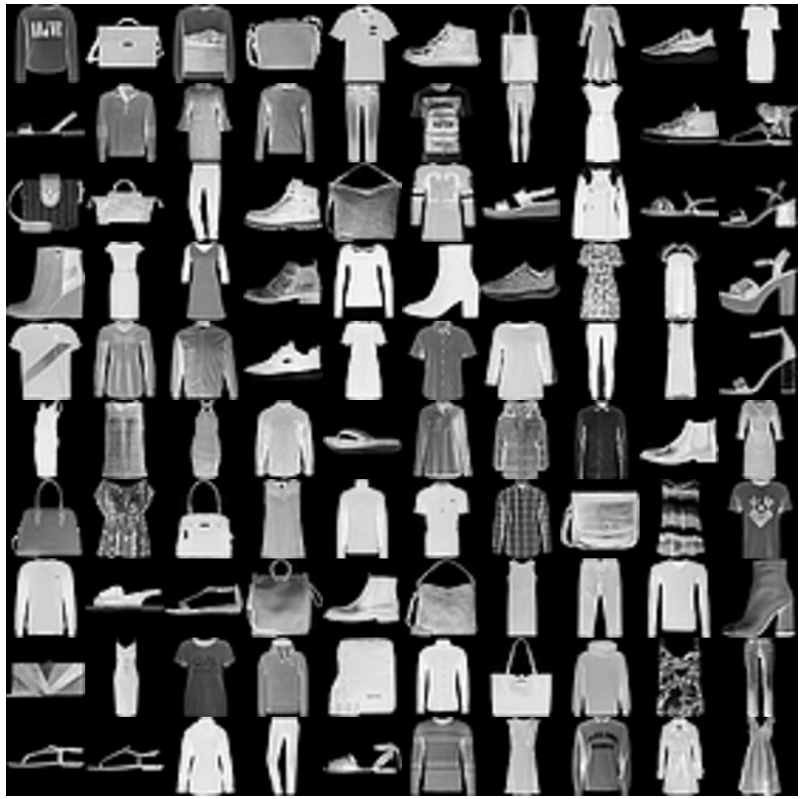
$$\downarrow$$

$$\min_{\mathbf{w}} \max_{\mathbf{t}\in\mathcal{P}} \sum_{i=1}^{k} t_i \ell_i(\mathbf{w})$$



Mohri et al. "Agnostic federated learning." *arXiv:1902.00146* (2019).

Numerical results



- Fair performance among different categories of data

$$\min_{\mathbf{w}} \max\{\ell_1(\mathbf{w}), \ell_2(\mathbf{w}), \ell_3(\mathbf{w})\}$$

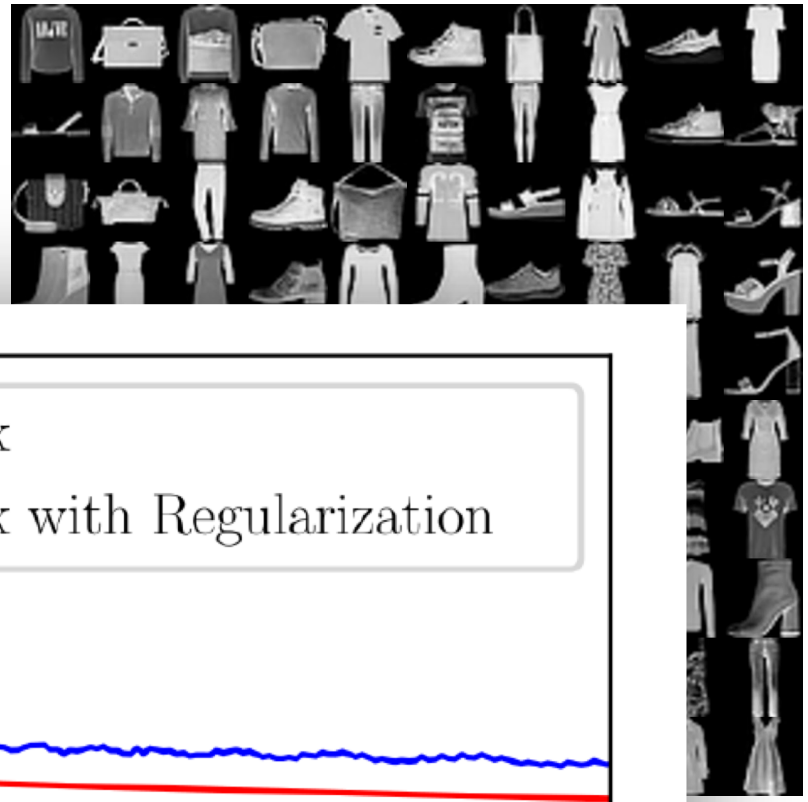
$$\min_{\mathbf{w}} \max_{\mathbf{t}\in\mathcal{P}} \sum_{i=1}^{3} t_i \ell_i(\mathbf{w})$$

Average performance over 100 training:

	T-shirt/Top	Coat	Shirt
Normal Training	84.1 ±1.8%	86.4 ±2.1%	**70.6 ±3.7%**
Min-max no regularizer	75.4 ±1.5%	**71.6 ±3.0%**	73.3 ±1.9%
Min-max with regularizer	76.3 ±1.4%	**73.9 ±2.8%**	74.8 ±1.6%

- Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn, "Solving a class of non-convex min-max games using iterative first order methods," arXiv:1902.08297, *accepted in NeurIPS 2019*.
- Mohri et al. "Agnostic federated learning." *arXiv:1902.00146* (2019).

Numerical results



- Fair performance among different categories of data

Min-max with regularizer	76.3 ±1.4%	**73.9 ±2.8%**	74.8 ±1.6%	**73.4 ±2.4%**

Mohri et al. "Agnostic federated learning." *arXiv:1902.00146* (2019).

Fair learning

- Discriminatory behaviors in human decisions and machine learning models:
  - [Bickel et al., 1975]: Sex bias in graduate admissions in Berkeley
  - [Datta et al. 2015]: Google's online advertising showed high-income jobs ads to men more than to women.
  - [Sweeney 2013]: ads for arrest records shows up on searches for distinctively black names.
  - Amazon's recruitment engine has bias against women*
- Different reasons such as old data human bias
- Regulated domains: employment, housing, education, …
- Designing *discrimination-free* machine learning models
- **Goals**:
  - Make $\hat{y}$ and s independent
  - Keep $\hat{y}$ close to y

* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Imposing fairness via regularization

- **Goals**:
  - Make $\hat{\boldsymbol{y}}$ and $\boldsymbol{s}$ independent
  - Keep $\hat{\boldsymbol{y}}$ close to $\boldsymbol{y}$

$$\min_{\boldsymbol{\theta}} \; \mathbb{E}[\ell(\mathbf{y}, \hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}))] + \lambda \rho(\hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{s})$$

Keep classification error small

Imposing fairness

Different correlation measures: Mutual information [*Kamishima et al. 2011*], false positive/negative rates [*Bechavod & Ligett 2017*], equalized odds [*Donini et al. 2018*], Pearson correlation coefficient [*Zaffar et al. 2015, 2017*], Hilbert Schmidt independence criterion [*Pérez-Suay et al. 2017*]

- Either do not have convergence guarantees or cannot guarantee statistical independence

Rényi Fair Inference

- **Goals**:
  - Make $\hat{\boldsymbol{y}}$ and $\boldsymbol{s}$ independent
  - Keep $\hat{\boldsymbol{y}}$ close to $\boldsymbol{y}$

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}[\ell(\mathbf{y}, \hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}))] + \lambda \rho(\hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{s})$$

- **Use Rényi (maximal) correlation**

$$\rho(A, B) = \sup_{f,g} \quad \mathbb{E}[f(A)g(B)]$$
$$\text{s.t.} \quad \mathbb{E}[f(A)] = \mathbb{E}[g(B)] = 0, \quad \mathbb{E}[f^2(A)] = \mathbb{E}[g^2(B)] = 1$$

- **Rényi Fair Inference** [*Bahrlouei, Nouiehed, Beirami, Razaviyayn, ICLR 2020*]

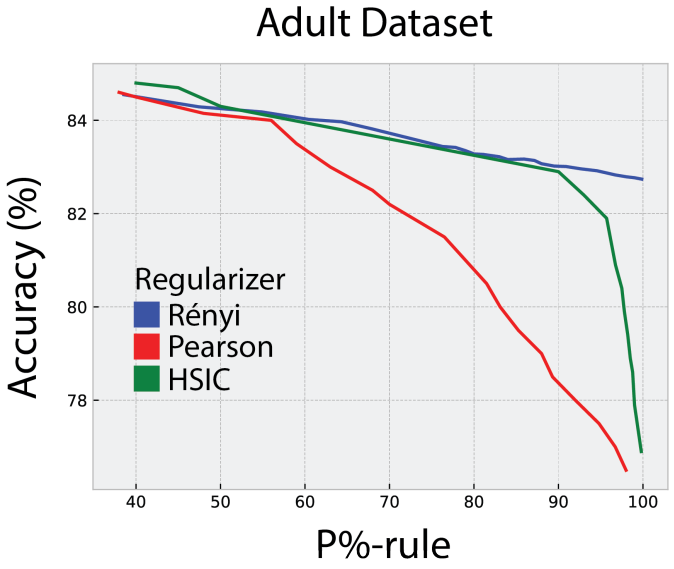
$$\min_{\boldsymbol{\theta}} \max_{f,g} \quad \mathbb{E}[\ell(\mathbf{y}, \hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}))] + \lambda \mathbb{E}[f(\hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}))g(\mathbf{s})]$$
$$\text{s.t.} \quad \mathbb{E}[f(\hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}))] = \mathbb{E}[g(\mathbf{s})] = 0, \quad \mathbb{E}[f^2(\hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}))] = \mathbb{E}[g^2(\mathbf{s})] = 1$$

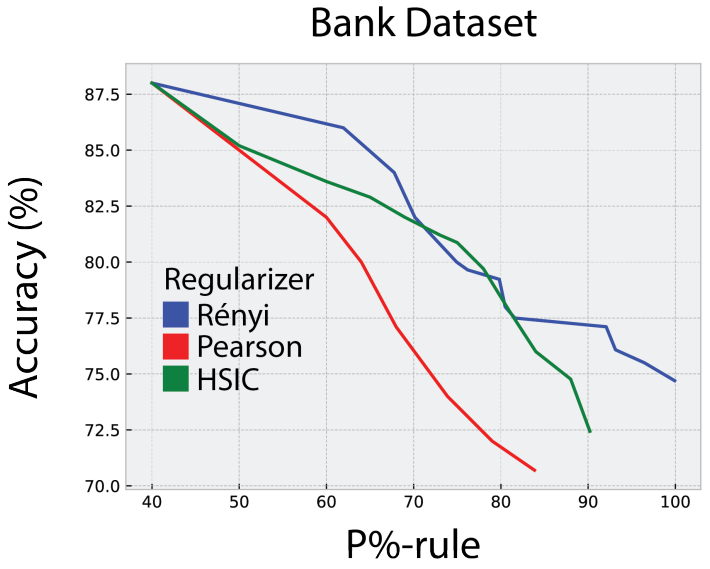
- Can be solved for discrete random variables

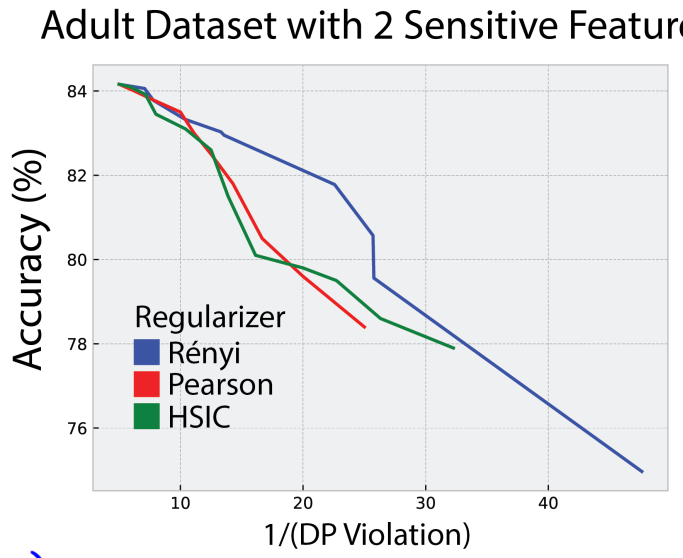
Numerical Experiments

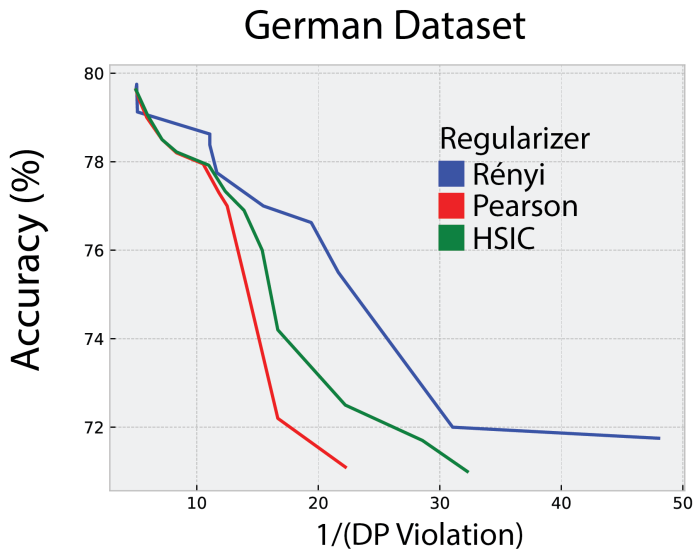
- Pearson correlation coefficient
  - [*Zaffar et al. 2015, 2017*]
- Hilbert Schmidt Independence Criterion
  - [*Pérez-Suay et al. 2017*]
- Rényi Fair Inference
  - [*Baharlouei et al. 2019*]

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}[\ell(\mathbf{y}, \hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}))] + \lambda \rho(\hat{\mathbf{y}}_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{s})$$







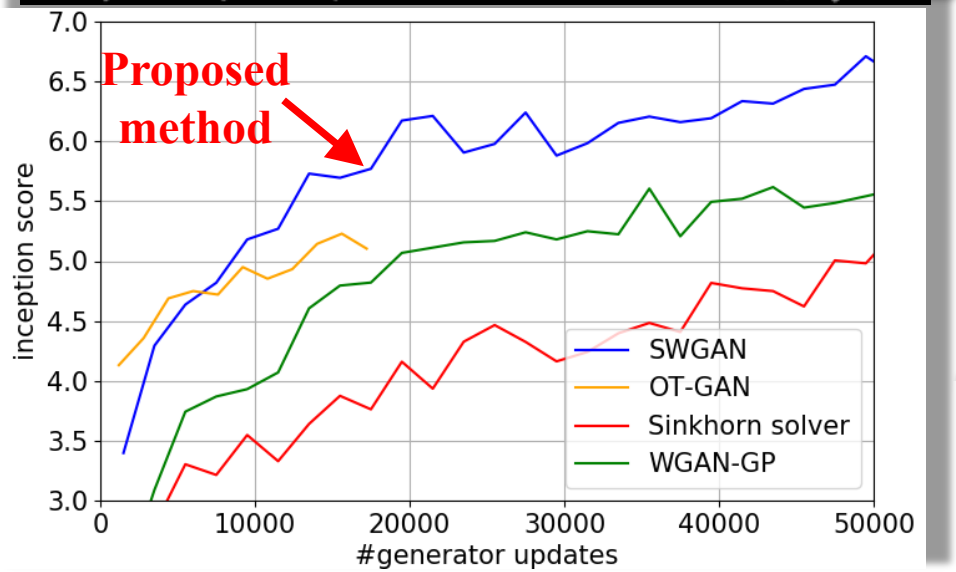


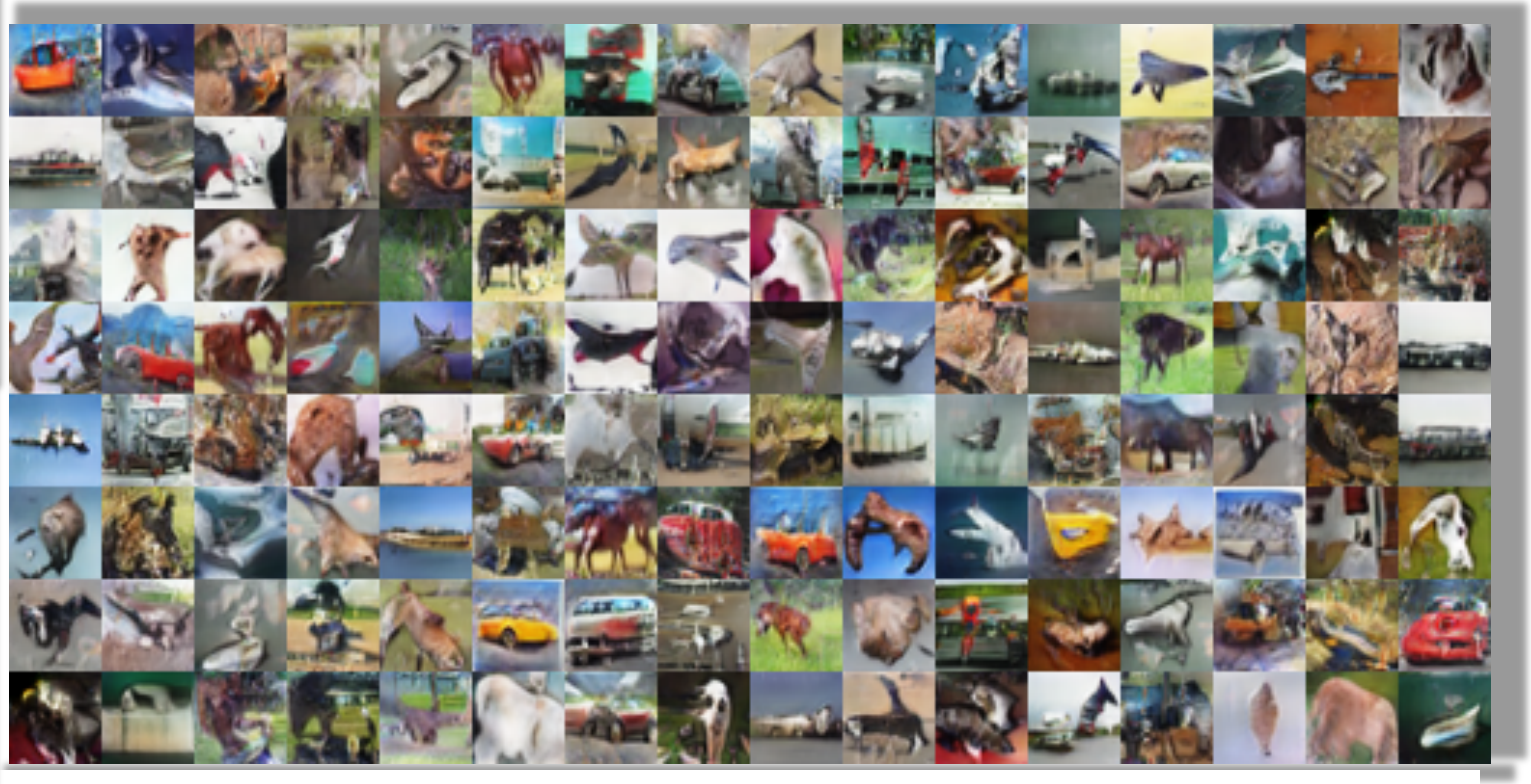
$$p\% = \min\left\{\frac{\mathbb{P}(\hat{Y}=1|S=1)}{\mathbb{P}(\hat{Y}=1|S=0)}, \frac{\mathbb{P}(\hat{Y}=1|S=0)}{\mathbb{P}(\hat{Y}=1|S=1)}\right\}$$

$$\text{DP Violation} = \max_{a,b} |\mathbb{P}(\hat{Y}=1|S=a) - \mathbb{P}(\hat{Y}=1|S=b)|$$

Extension to stochastic setting and applications in training GANs

- Sanjabi, Ba, Razaviyayn, Lee. "On the convergence and robustness of training GANs with regularized optimal transport," *NeurIPS* 2018





Summary

- Non-convex min-max problems appear in many modern applications
- Non-convex min-max problems are challenging
- Special cases can be solved *efficiently*
- Still many open problems

A long history

$$\min_{\boldsymbol{\theta}\in\Theta}\ \max_{\boldsymbol{\alpha}\in\mathcal{A}}\ f(\boldsymbol{\theta},\boldsymbol{\alpha})$$

- Using monotone operator:
  - [Sibony'70], [Korpelevich'76], [Nemirovski'04], [Martinet'70], [Rockafellar'76], [Di-Sun'99], [Juditsky-Nemirovsky'16], …
- Weak Monotonicity
  - [Davis-Grimmer'17], [Davis-Drusvyatskiy'18], [Zhang-He'18], [Lin et al'18], …
- More general VI's
  - [Facchinei-Pang'03], [Monteiro-Svaiter'10], [Nesterov'07], [Dong-Lan'14], …
- Stochastic VI's
  - [Juditsky-Nemirovski-Tauvel '11], [Koshal-Nedic-Shanbag'13], [Rosasco-Villa-Vũ'14], [Balamurugan-Bach'16], …
- Bilinear convex-concave
  - [Arrow-Hurwicz-Uzawa'58, Zhu-Chan'08], [Chambolle-Pock'11&16], [Chen-Lan-Ouyang'14], [Dong-Lan'14, Chambolle et al'17], [Wang-Xiao'17], …
- Convex-Concave saddle points
  - [Tseng'08], [He and Monterio'17], [Hamedani-Jalilzadeh-Aybat-Shanbhag'18], …

Other recent results for min-max regimes

- [Ioan Bot-Böhm 2021]
- [Jamali-Rad and Szabó 2021]
- [Ouyang-Xu 2021]
- [Anagnostidi, Lucchi, and Diouane 2021]
- [Huang, Gao, and Huang 2021]
- [Yoon and Ryu 2021]
- [Han, Xie, and Zhang 2021]
- [Vladislav et al 2021]
- [Mangoubi and Vishnoi 2021]
- [Zhang et al 2020]
- [Tran-Dinh et al 2020]
- [Yang, Kiavash, and He 2020]
- [Lin, Jin, Jordan 2020]
- [Lu, Tsaknakis, and Hong 2019]
- [Gidel, Hemmat, Pezeshki, Huang, Lepriol, Lacoste-Julien, and Mitligkas 2018]
- [Gidel, Jebara, and Lacoste-Julien 2018]
- [Lu, Tsaknakis, Hong, Chen 2019]
- [Mokhtari, Ozdaglar, Pattathil 2019]
- [Daskalakis and Panageas 2019]
- [Thekumparampil, Jain, Netrapalli, and Oh 2019]
- [Jin, Netrapalli, and Jordan 2019]
- [Lin, Jin, Jordan 2019]
- [Letcher, Balduzzi, Racaniere, Martens, Foerster, Tuyls, and Graepel 2019]
- [Lin, Liu, Rafique, Yang 2018]
- [Hameani, Jalilzadeh, Aybat, Shanbhag 2018]
- [Rafique, Liu, Lin, and Yang 2018]
- [Sinha, Namkoong, and Duchi 2018]
- [Mescheder, Geiger, and Nowozin 2018]
- [Daskalakis and Panageas 2018]
- And many other recent works…

References

- M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. "On the convergence and robustness of training GANs with regularized optimal transport," *NeurIPS* 2018.
- M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn, "Solving a class of non-convex min-max games using iterative first-order methods," *NeurIPS 2019*.
- D. Ostrovskii, A. Lowy, and M. Razaviyayn, "Efficient search of first-order Stationary points in nonconvex-concave smooth min-max problems," *SIOPT, 2021*.
- S. Baharlouei, M. Nouiehed, and M. Razaviyayn. "Rènyi fair inference," *ICLR 2020.*
- B. Barazandeh and M. Razaviyayn, "Solving non-convex non-differentiable min-max games using proximal gradient method," *ICASSP 2020.*
- Z. Wang, K. Balasubramanian, S. Ma, and M. Razaviyayn, "Zeroth-order algorithms for nonconvex minimax problems with improved complexities," *arXiv:2001.07819, 2020.*
- M. Razaviyayn, T. Huang, S. Lu, M. Nouiehed, M. Sanjabi, and M. Hong. "Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances." *IEEE Signal Processing Magazine, 2020.*
- D. Ostrovskii, B. Barazandeh, and M. Razaviyayn, "Nonconvex-Nonconcave Min-Max Optimization with a Small Maximization Domain," arXiv: 2110.03950, *2021*.

- Codes are available at https://github.com/optimization-for-data-driven-scienc