Game theory, min-max optimization and modern machine learning

Ioannis Mitliagkas and collaborators

TUC Colloquium - March 2021

min-max formulations are everywhere

(more generally game-theoretic formulations) of increasing importance in modern ML

still a lot to explore in terms of:

1. Applications
2. Methods

Structure of my talk

1. Applications

2. Methods

3. Open questions/discussion

Structure of my talk

1. **Applications**
2. Methods
3. Open questions/discussion

APPLICATIONS

Generative Adversarial Networks

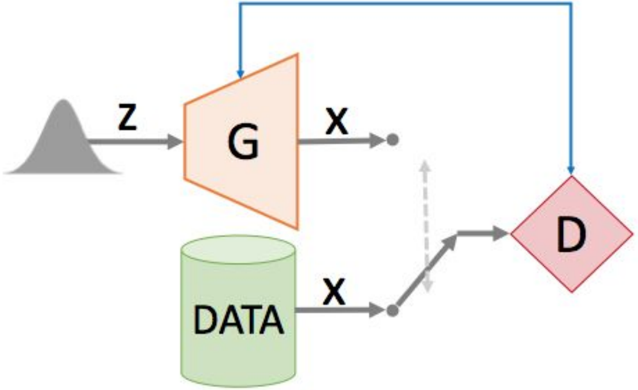
Both differentiable

Generator network, G

Given latent code, z, produces sample G(z)

Discriminator network, D

Given sample x or G(z), estimates probability it is real



$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim\mathbb{P}_x}[\log D(x)] + \mathbb{E}_{z\sim\mathbb{P}_z}[\log(1 - D(G(z)))]$$

Generative Adversarial Networks

Min-max = optimize worst case

"uncertainty set": random realization ω drawn from set Ω

$$\max_{x \in X} \min_{\omega \in \Omega} f(x, \omega).$$

- Operations research (planning for worst-case demand)
- Telecommunications (beamforming)
- Policy design

Lately: Connections to causality

“Robust to many environments” ~= “Causal” understanding

Invariant risk minimization [Arjovsky et al, 2019]

“Max robustness” ≈ Causality [Buhlmann, 2018]

Exciting modern **ML** applications

1. Out-of-distribution generalization
    a. Meta: studying generalization
2. Performative prediction
3. Fairness in ML

Exciting modern **ML** applications

1. Out-of-distribution generalization
   a. **Meta: studying generalization**
2. Performative prediction
3. Fairness in ML

Robust generalization measures

Goal:

- Use a robust prediction framework to evaluate generalization measures
  - i.e. good measure predicts generalization error in a wide variety of interesting settings
- Spoiler: No existing measure in literature is robustly predictive!
- Collaboration with UofT Stats/Vector, ElementAI
- NeurIPS 2020

In Search of Robust Measures of Generalization

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, Daniel Roy

Robust generalization measures

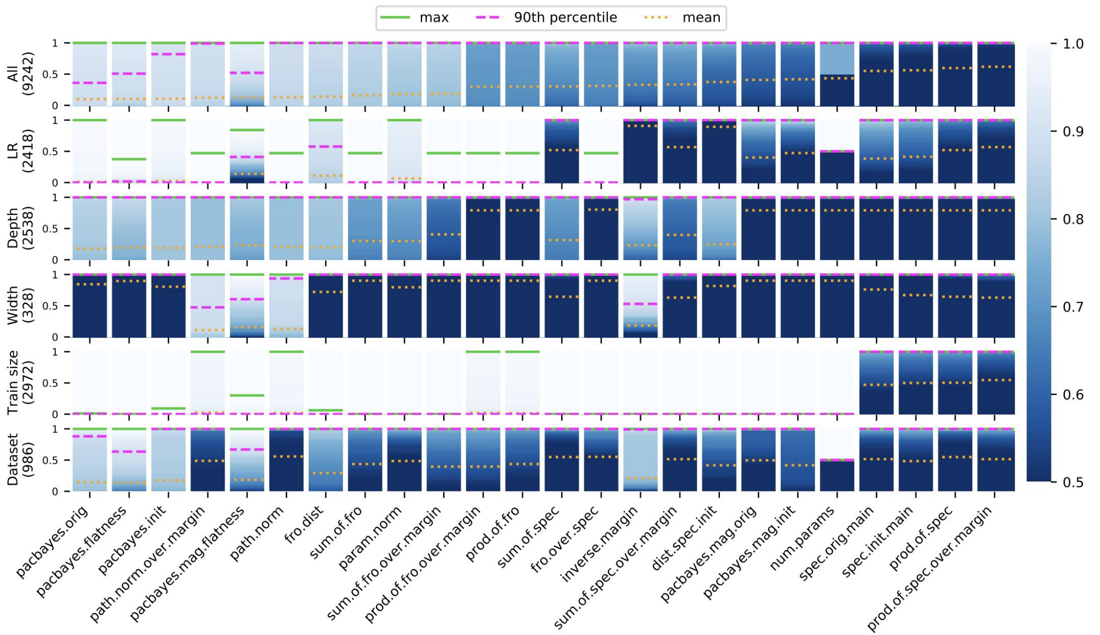


Figure 1: Cumulative distribution of the sign-error across subsets of environments for each generalization measure. The measures are ordered based on the mean across 'All' environments. A completely *white* bar indicates that the measure is perfectly robust, whereas a *dark blue* bar indicates that it completely fails to be robust.

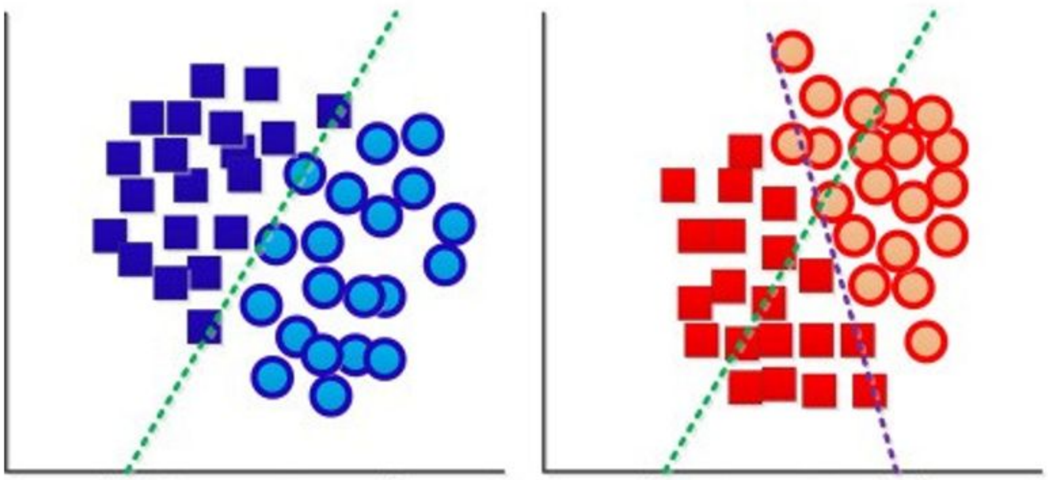
Beyond I.I.D. generalization (classic, in-distribution)

i.i.d. quantities

=

Ποσότητες
Ανεξάρτητες και
Ομοίως
Κατανεμημένες

Ταξινόμηση χωρίς ΠΑΟΚ



(a) Source Domain (b) Target Domain

No I.I.D assumption

- Performance degrades outside the training distribution
  - Major challenge to deployment of ML models!

- Need better **out-of-distribution (OOD) generalization**!

- Humans are doing better in many regards for OOD generalization

Out-of-distribution generalizaton

- Domain Adaptation
- Domain Generalization
- Adversarial Machine Learning

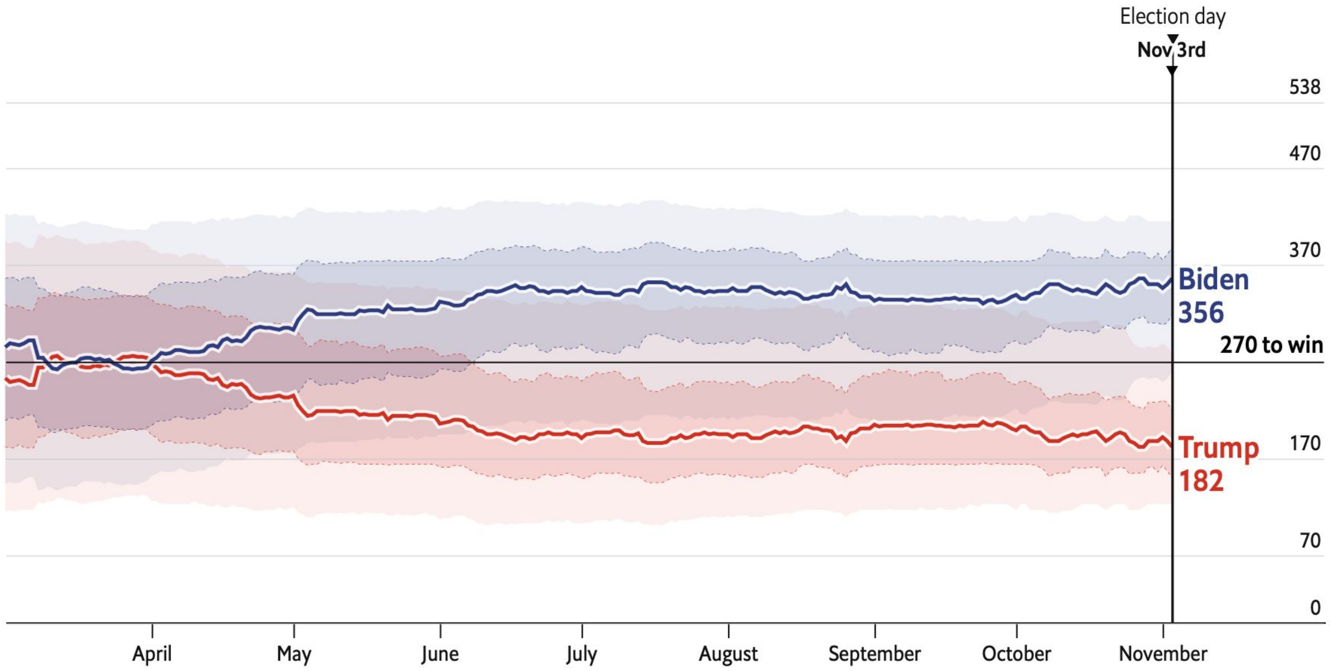
Adversarial target-invariant representation learning for domain generalization

Isabela Albuquerque, João Monteiro, Mohammad Darvishi,
Tiago Falk, Ioannis Mitliagkas

Exciting modern **ML** applications

1. Out-of-distribution generalization
   a. Meta: studying generalization
2. **Performative prediction**
3. Fairness in ML

Elections!!

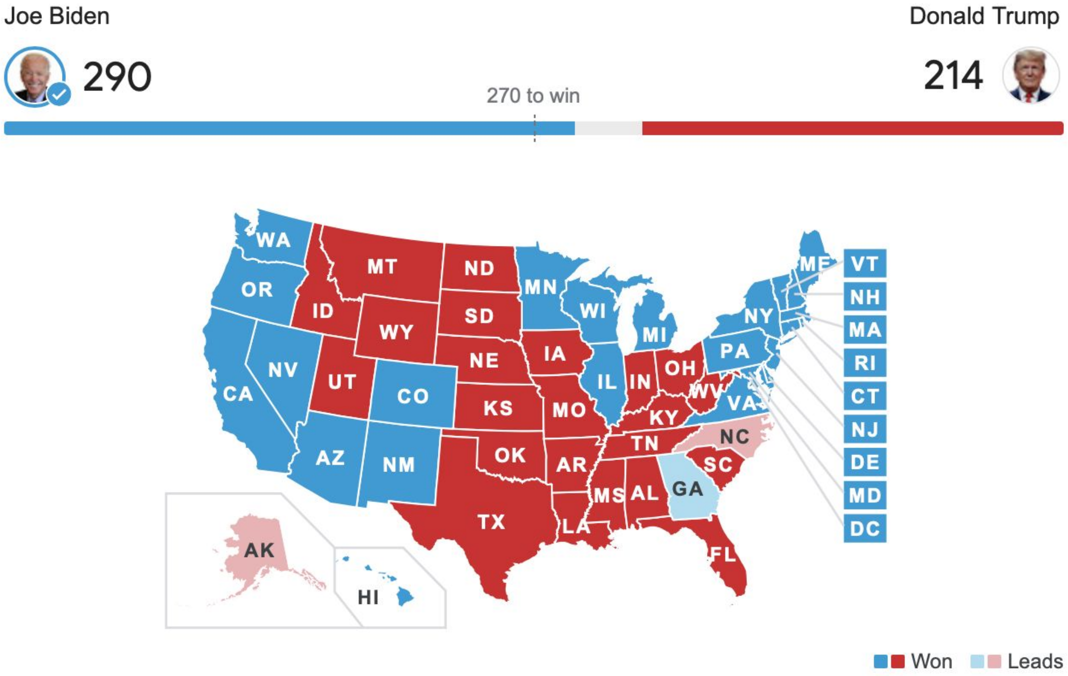


Economist

Predicting elections

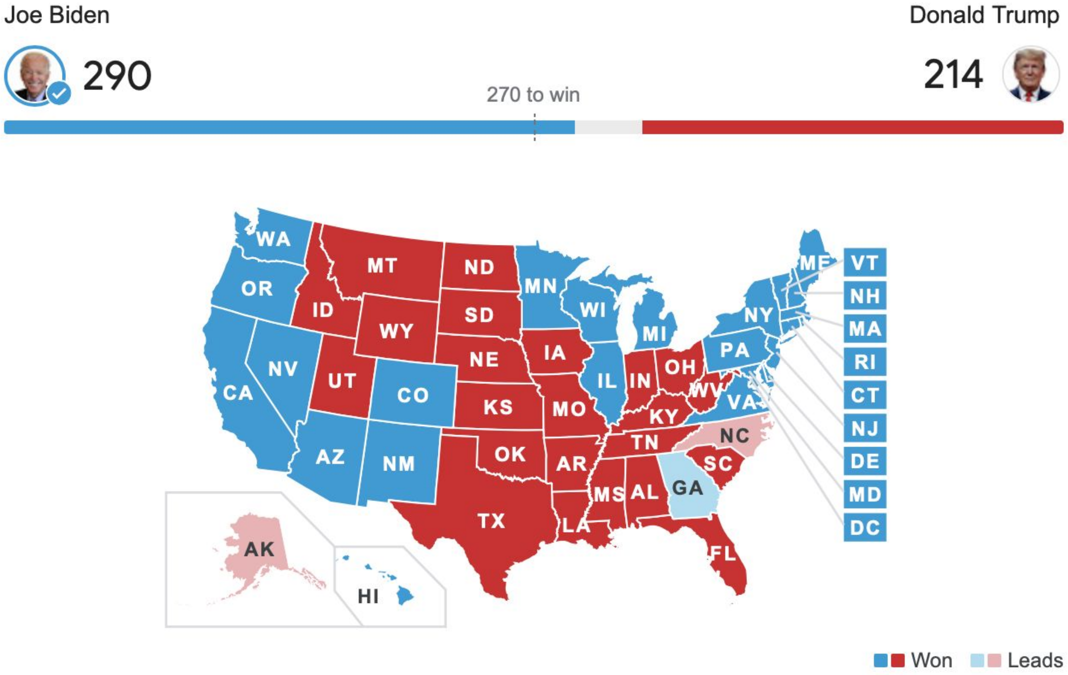
FiveThirtyEight

The result (Nov 5th)



Google

Why?


Joe Biden
290
270 to win
Donald Trump
214
WA
OR
CA
NV
ID
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
OH
KY
TN
MS
AL
GA
FL
SC
NC
WV
VA
PA
NY
ME
VT
NH
MA
RI
CT
NJ
DE
MD
DC
AK
HI
Won
Leads

Google

Why?

1. Polling in modern era is much harder

2. Closeted voters

3. **“Underdog effect”**

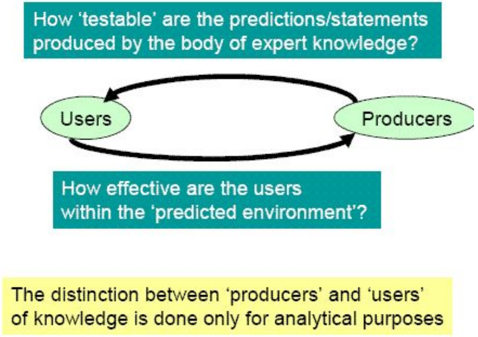
Underdog effect

An underdog effect, on the other hand, could penalise the leading candidate. This is because supporters think it's a done deal and don't mobilise to vote (resting on their laurels) or because the supporters of the trailing candidate are motivated by the idea of losing (a back-to-the-wall effect).

A feedback loop from prediction back to the data distribution!

PERFORMATIVITY

The concept of **performativity**, as developed in **economic** sociology (**Callon, 1998**;MacKenzie et al., 2007), directs our attention to the role of expert bodies of knowledge (e.g., theories, formulae, models) in the functioning of the **economy** and organizational life.



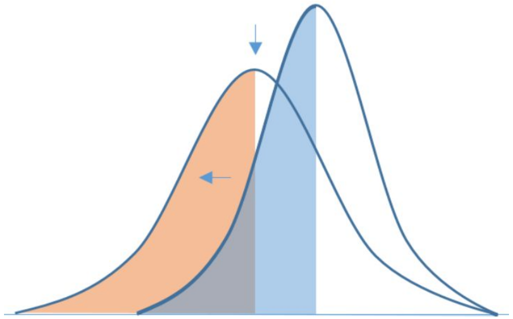
well-studied phenomenon in policy-making

but neglected in supervised learning.

Performative Prediction

Juan C. Perdomo* Tijana Zrnic* Celestine Mendler-Dünner Moritz Hardt

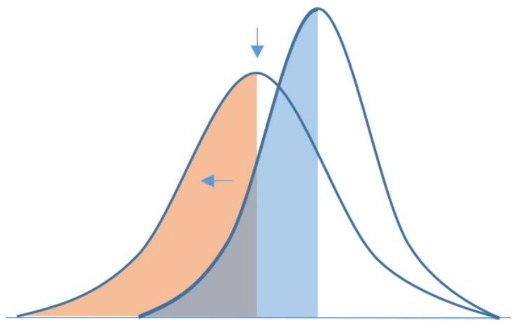
"Predictions that support decisions,
may influence the outcome they aim to predict."



Ok, that's one more example of out-of-distribution generalization!

Performative Prediction

Special structure!
- Causality
- Game-theoretic formulation



Stackelberg equilibria identified as
"performative optima"

Understanding of game theory and developing the right methodology
→ CRITICAL

Exciting modern **ML** applications

1. Out-of-distribution generalization
    a. Meta: studying generalization
2. Performative prediction
3. **Fairness in ML**

Fairness in ML

The Disparate Equilibria of Algorithmic Decision Making when Individuals Invest Rationally

Lydia T. Liu
University of California, Berkeley

Ashia Wilson
Microsoft Research

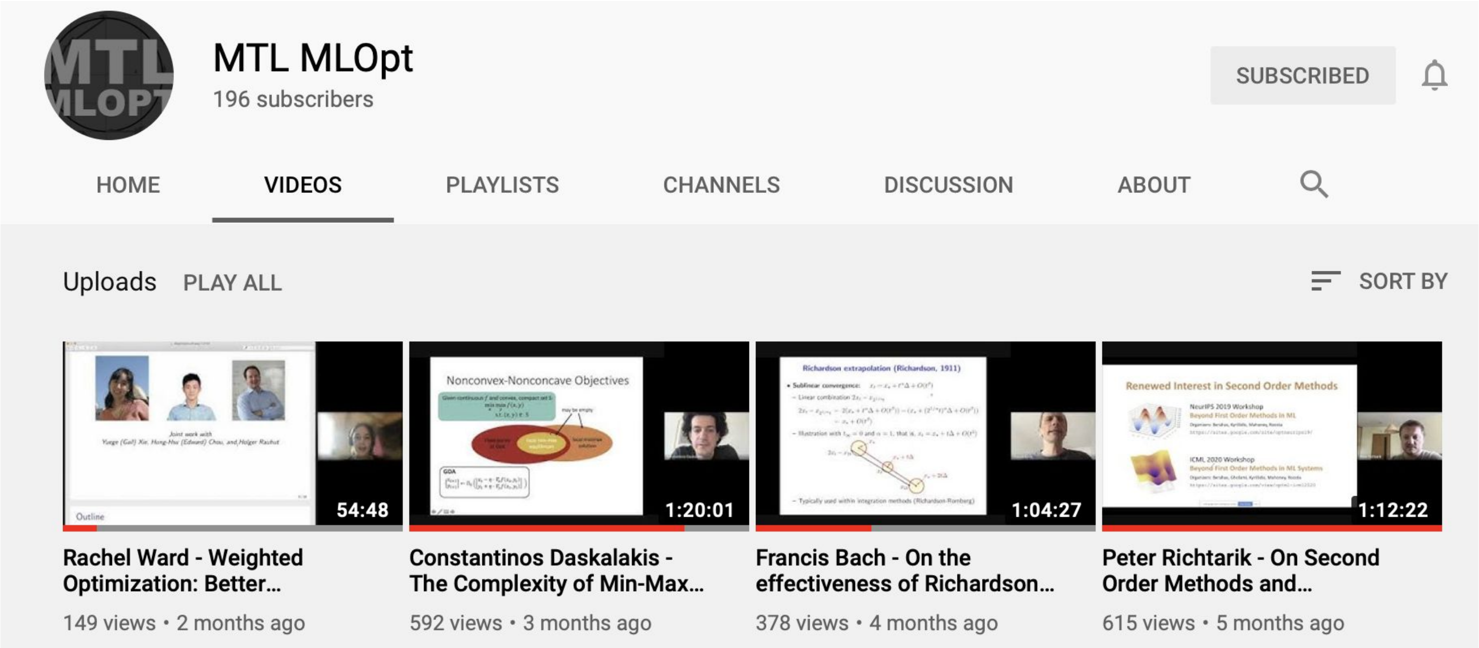
Nika Haghtalab
Cornell University

Adam Tauman Kalai
Microsoft Research

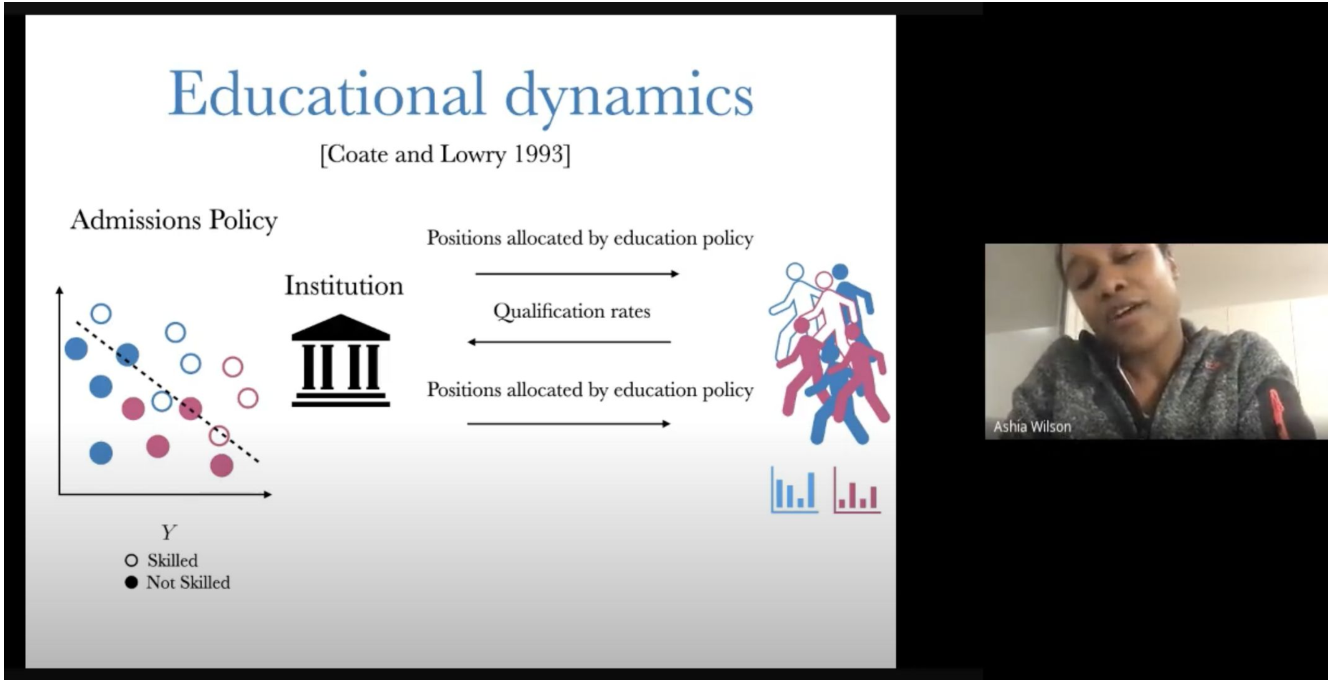
Christian Borgs
Microsoft Research

Jennifer Chayes
Microsoft Research

mtl-mlopt.github.io



Fairness in ML (Ashia Wilson)



Fairness in ML

Model

✴ S. Coate and G. C. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 83(5):1220–1240, 1993.

Institution utility maximizing

Chooses qualification parameter θ to accept candidates according to

$$\arg\max_{\theta} \ \text{profit} \cdot \mathbb{P}(\hat{Y}_\theta = 1, Y = 1) - \text{cost} \cdot \mathbb{P}(\hat{Y}_\theta = 1, Y = 0)$$

Individuals utility maximizing

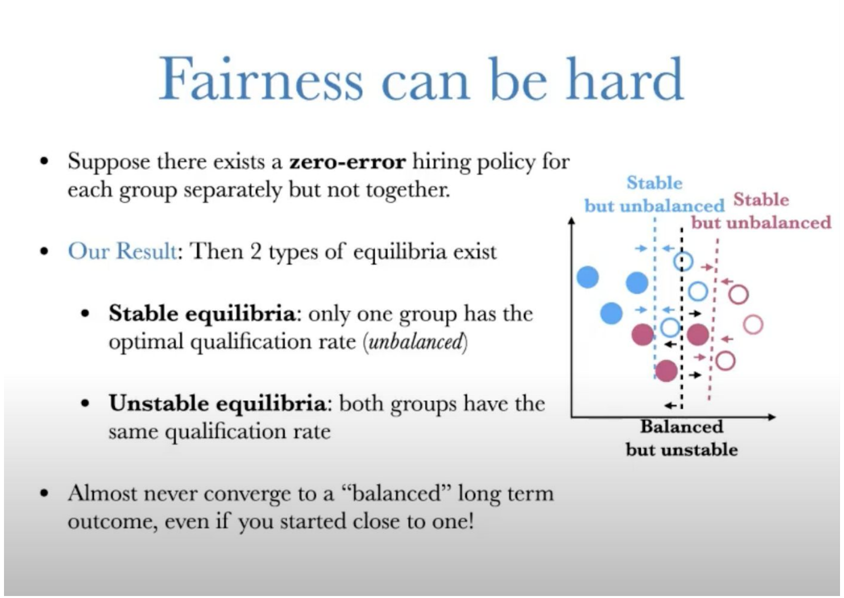
Individuals acquire label $Y = 1$ if

$$\text{wage} \ \underbrace{(\mathbb{P}(Y_\theta = 1 \mid Y = 1) - \mathbb{P}(Y_\theta = 1 \mid Y = 0))}_{\text{expectation investment will payoff (function of error)}} - \text{cost to invest} \geq 0$$

Fairness is hard

- Stable equilibria not balanced
- Balanced states are not stable

- Exciting questions



“Minimization to current AI is what min-max optimization is to future AI”

--Costis Daskalakis

Structure of my talk

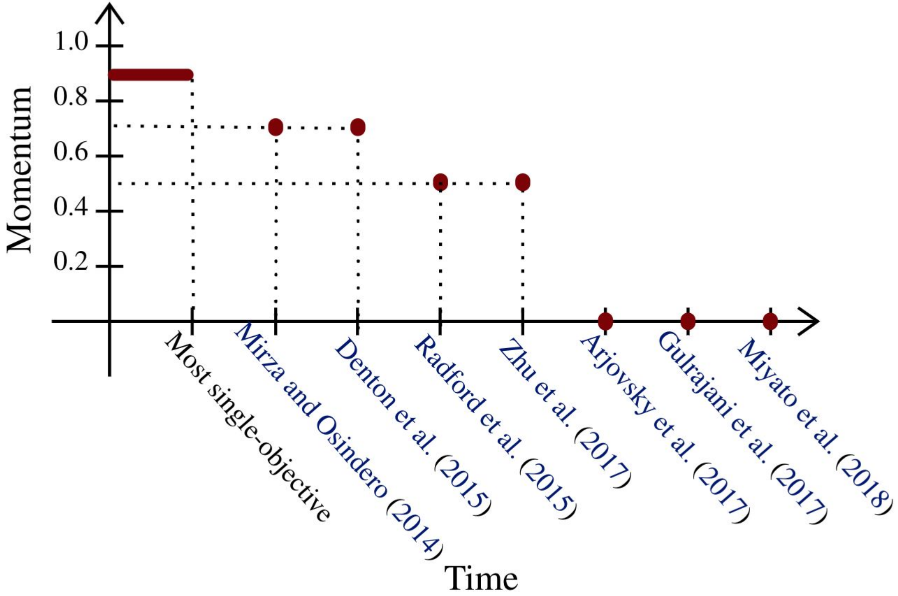
1. Applications
2. **Methods**
3. Open questions/discussion

METHODS

Negative Momentum for Improved Game Dynamics

along with Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Gabriel Huang, Remi Lepriol, Simon Lacoste-Julien

Trend in GAN literature



Start with optimization dynamics

Optimization

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta})$$

Smooth, differentiable cost function, L

→ Looking for stationary (fixed) points (gradient is 0)

→ Gradient descent

Optimization

Ferenc Huszar

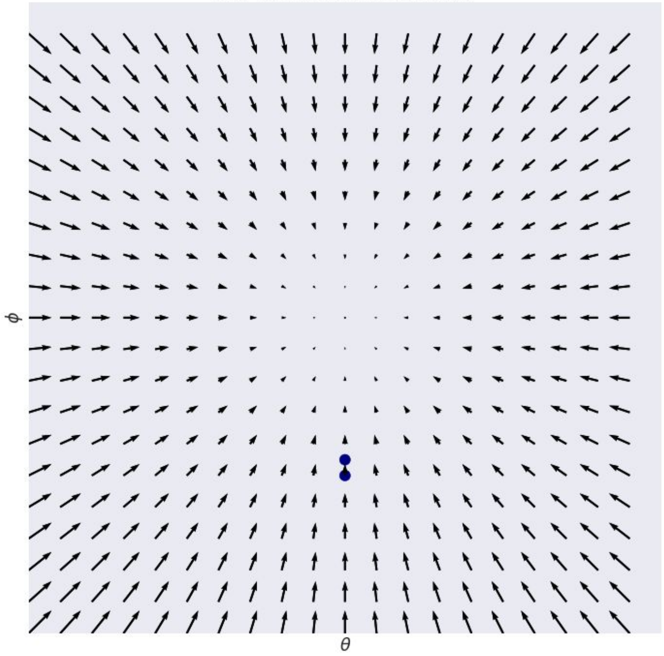
$$v(\boldsymbol{\theta}) = \nabla \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta})$$

Conservative vector field

→

Straightforward dynamics

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{v}(\boldsymbol{\theta}_t)$$



Gradient descent

$$\boldsymbol{v}(\boldsymbol{\theta}) = \nabla \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta})$$

Conservative vector field

→

Straightforward dynamics

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{v}(\boldsymbol{\theta}_t)$$

Fixed-point analysis

$$F_\eta(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta \boldsymbol{v}(\theta)$$

Jacobian of operator

$$\nabla F_\eta(\boldsymbol{\theta}) = I - \eta \underline{\nabla \boldsymbol{v}(\theta)}$$

Hessian of objective, L

Local convergence

Theorem 1 (Prop. 4.4.1 Bertsekas [1999]). *If the spectral radius* $\rho_{\max} \overset{def}{=} \rho(\nabla F_\eta(\boldsymbol{\omega}^*)) < 1$*, then, for* $\boldsymbol{\omega}_0$ *in a neighborhood of* $\boldsymbol{\omega}^*$*, the distance of* $\boldsymbol{\omega}_t$ *to the stationary point* $\boldsymbol{\omega}^*$ *converges at a linear rate of* $\mathcal{O}\big((\rho_{\max} + \epsilon)^t\big)$, $\forall \epsilon > 0$.

Eigenvalues of op. Jacobian

$$\lambda(\nabla F_\eta(\boldsymbol{\theta})) = 1 - \eta\lambda(\nabla \boldsymbol{v}(\theta))$$

If ρ(θ*)=max |λ(θ*)|<1, then fast local convergence

Jacobian of operator

$$\nabla F_\eta(\boldsymbol{\theta}) = I - \eta\underline{\nabla \boldsymbol{v}(\theta)}$$

Hessian of objective, L
Symmetric, real-eigenvalues

Games

Nash Equilibrium

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}^*)$$

$$\boldsymbol{\varphi}^* \in \arg\min_{\boldsymbol{\varphi} \in \boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi})$$

Smooth, differentiable L
→ Looking for local Nash eq

→ Gradient descent
- → **Simultaneous**
- → **Alternating**

Game dynamics

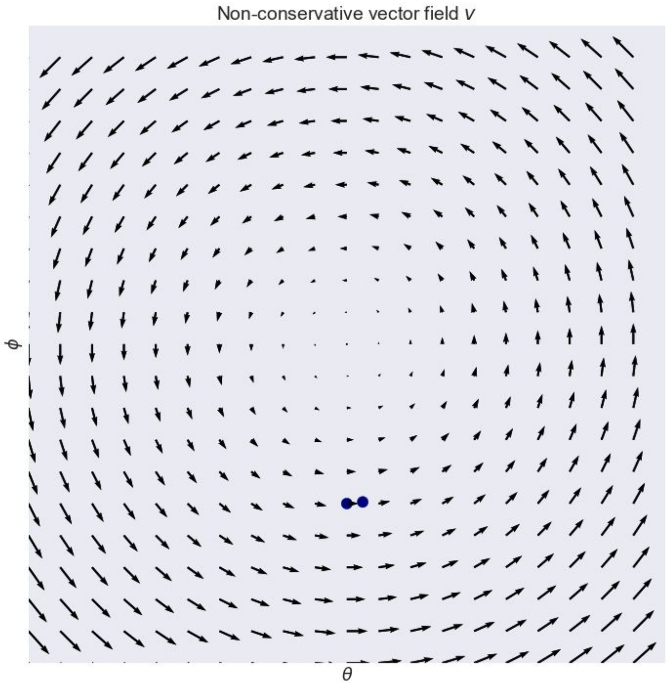
$$\boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta}) := \begin{bmatrix} \nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \\ \nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \end{bmatrix}$$

Non-conservative vector field

→

Rotational dynamics

$$F_\eta(\boldsymbol{\varphi}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\varphi} & \boldsymbol{\theta} \end{bmatrix}^\top - \eta\, \boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta})$$

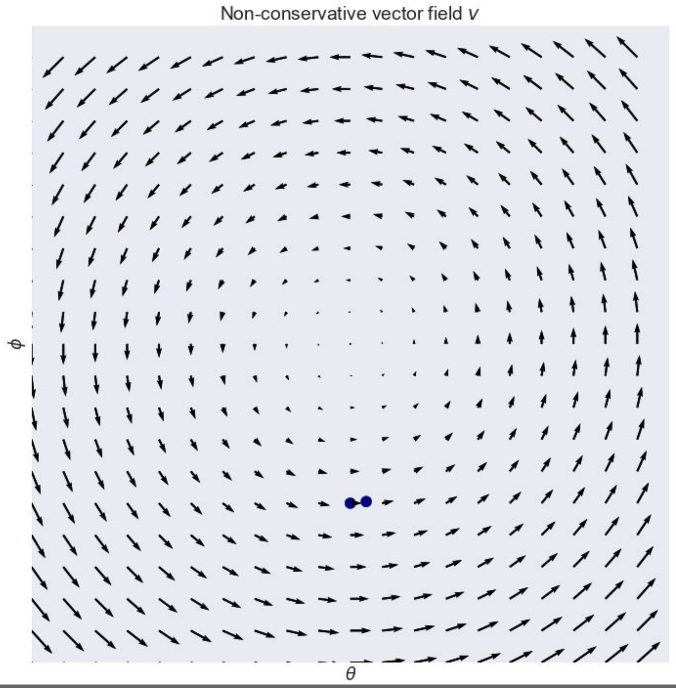


Game dynamics under gradient descent

$$F_\eta(\boldsymbol{\varphi}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} [\boldsymbol{\varphi} \quad \boldsymbol{\theta}]^\top - \eta\, \boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta})$$

Jacobian is non-symmetric, with complex eigenvalues → Rotations in decision space

Games demonstrate rotational dynamics.



Bilinear game

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} \ \boldsymbol{\theta}^\top \boldsymbol{A} \boldsymbol{\varphi}$$

Method	β	Bounded	Converges
Simultaneous	$\beta \in \mathbb{R}$	✗	✗
Alternated	>0	✗	✗
	0	✓	✗
	<0	✓	✓

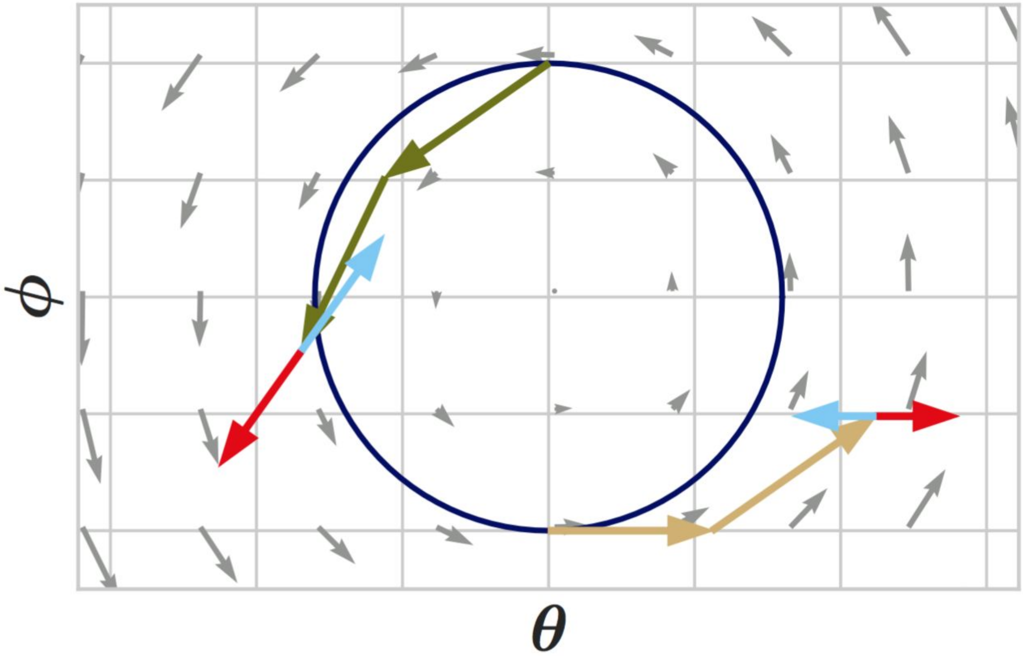
"Proof by picture"

Gradient descent

- Simultaneous
- Alternating

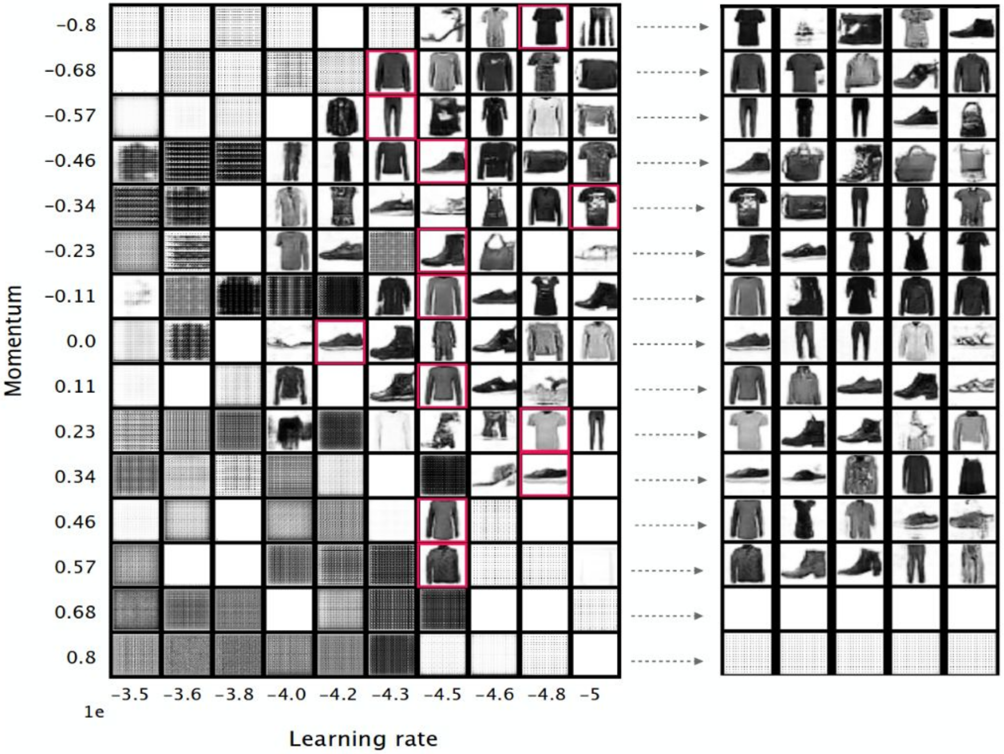
Momentum

- Positive
- Negative



What happens in practice ?

Fashion MNIST:



What happens in practice ?

CIFAR-10:

Negative Momentum

To sum up:

- Negative momentum seems to improve the behaviour due to “bad” eigenvalues.
- Optimal for a class of games
- Empirically optimal on “saturating” GANs

Mila

Université de Montréal

Linear Lower Bounds and Conditioning of Differentiable Games

$$\kappa = \sqrt{\frac{L_1 L_2 + L_{12}^2}{\mu_1 \mu_2 + \mu_{12}^2}}$$

Adam Ibrahim, Waïss Azizian, Gauthier Gidel, Ioannis Mitliagkas

A Tight and Unified Analysis of Extragradient for a Whole Spectrum of Differentiable Games

along with Waïss Azizian, Gauthier Gidel, Simon Lacoste-Julien

Accelerating Smooth Games by Manipulating Spectral Shapes

along with Waïss Azizian, Damien Scieur,
Simon Lacoste-Julien, Gauthier Gidel

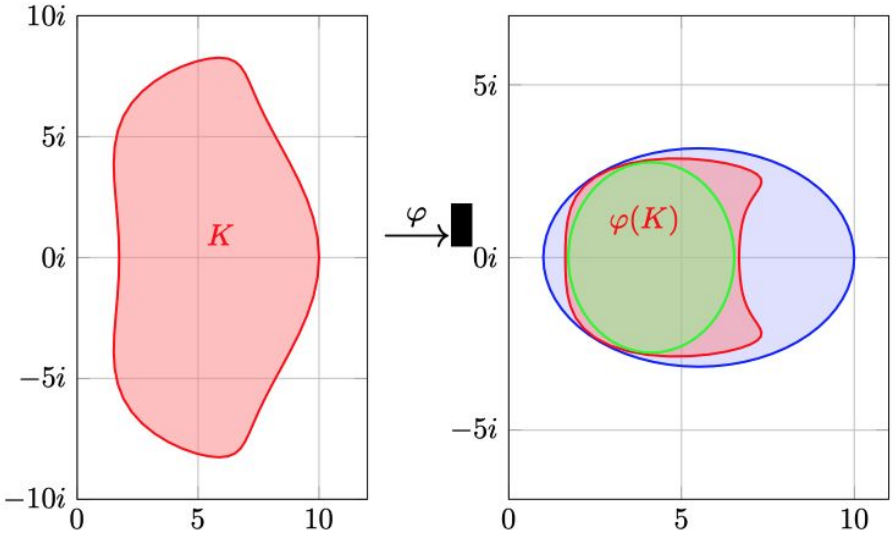


Figure 1: Transformation of the spectral shape K (in red from left to right) by the extragradient operator $\varphi : \lambda \mapsto \lambda(1 - \eta\lambda)$. Any ellipse (e.g. in blue) that contains the transformed red shape $\varphi(K)$ provides a upper convergence bound using extragradient with Polyak momentum (with step-size and momentum that depends on the ellipse parameters). Any ellipse included in it (e.g. in green) provides a lower bound. See §3.4.

Mila

Université de Montréal

Stochastic **Hamiltonian** Gradient Methods for Smooth Games

Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, Ioannis Mitliagkas

LEAD: Least-Action Dynamics for Min-Max Optimization

Reyhane Askari Hemmat, Amartya Mitra, Guillaume Lajoie,
Ioannis Mitliagkas

Structure of my talk

1. Applications

2. Methods

3. **Open questions/discussion**

Multi-objective training of Generative Adversarial Networks

Isabella Albuquerque, Joao Monteiro, T. Doan, B. Considine, T. Falk, I. Mitliagkas

Structure of my talk

1. Applications

2. Methods

3. **Open questions/discussion**

OPEN QUESTIONS AND DISCUSSION

Optimal methods

1. Convex-concave
2. Stochastic
3. Constrained
4. Non-convex, non-concave

Notions of equilibria

- What's the point of Nash equilibria?
- LOLA (Foerster, 2019)
  - Hints to Pareto semiorder of solutions
- Performative optima
- Stackelberg equilibria
- Domain specific?

Notions of equilibria

- What's the point of Nash equilibria?
- LOLA (Foerster, 2019)
    - Hints to Pareto semiorder of solutions
- Performative optima
- Stackelberg equilibria
- Domain specific?

Goldmine

ML ∩ Game theory ∩ Causality

Growing interest in ML and numerical optimization community

Thank you kindly