Equilibrium Computation and Machine Learning

Constantinos (a.k.a. “Costis”) Daskalakis
EECS & CSAIL, MIT


Max Fishelson (MIT)


Noah Golowich (MIT)


Stratis Skoulakis (SUTD)


Manolis Zampetakis (UC Berkeley)

A Motivating Question

vs

How is it that ML models beat humans in Go and Poker, but can't enter highways?

Equilibrium Problems in Machine Learning

Past Decade:

Exciting Progress in
Deep Learning
speech/image recognition
text generation
translation
…

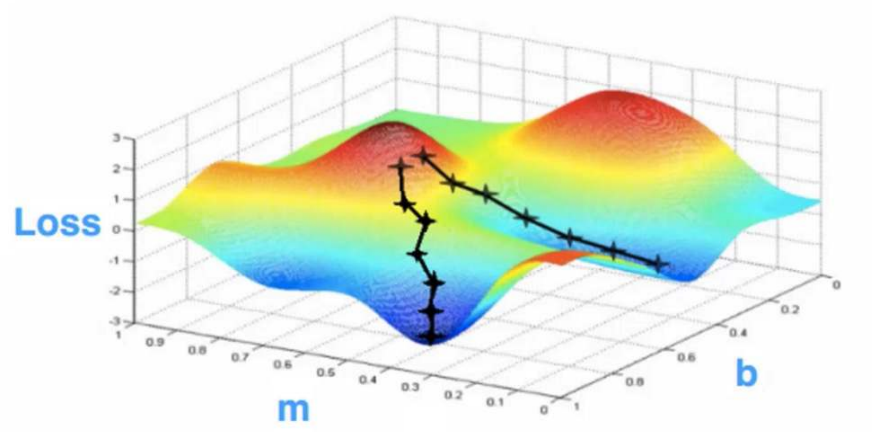
$\approx$

Single-Agent Optimization

$$\boxed{\min_x f(x)}$$

f: non-convex

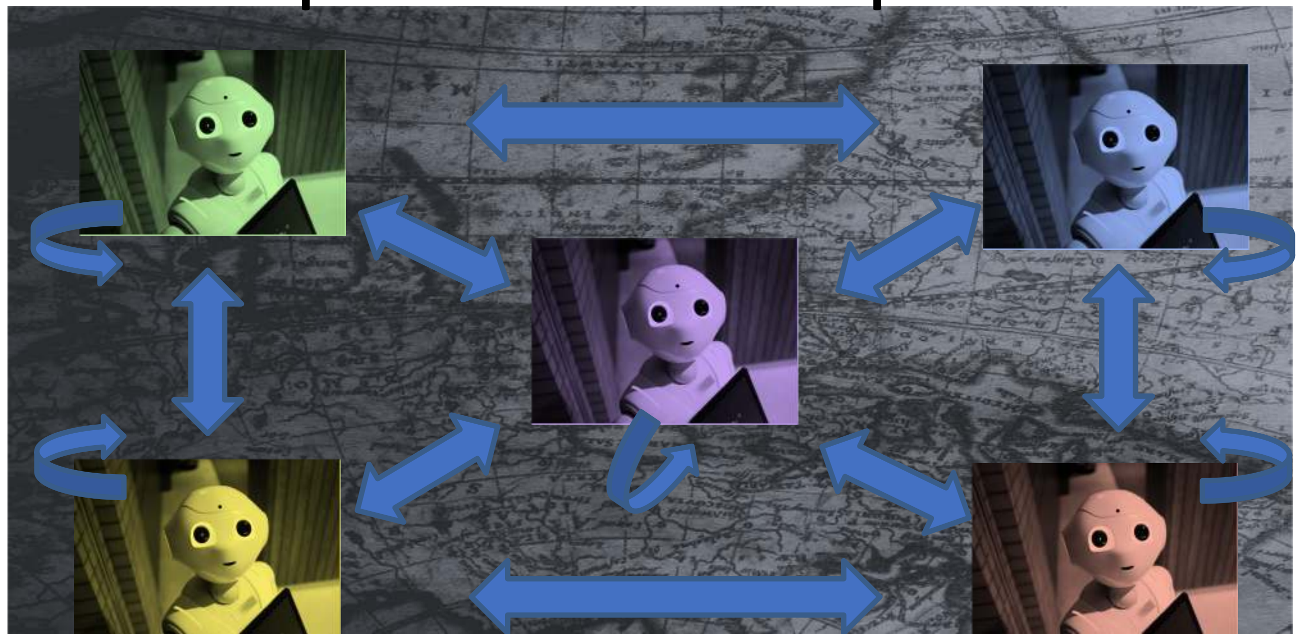
(+ models, learning objectives
hardware, data, …)



Empirical Finding: Gradient Descent (GD) and its variants *discover local minima* which generalize well

Equilibrium Computation

Now:



Practical Experience: GD vs GD (vs GD…) have a hard time converging, let alone to something meaningful

<u>How *deep* (no pun intended) is this issue?</u>

Training Oscillations and/or Garbage Solutions: already in two-agent min-max settings

$$\min_x \max_y f(x, y)$$

typically f is not convex/concave; and x, y multidimensional

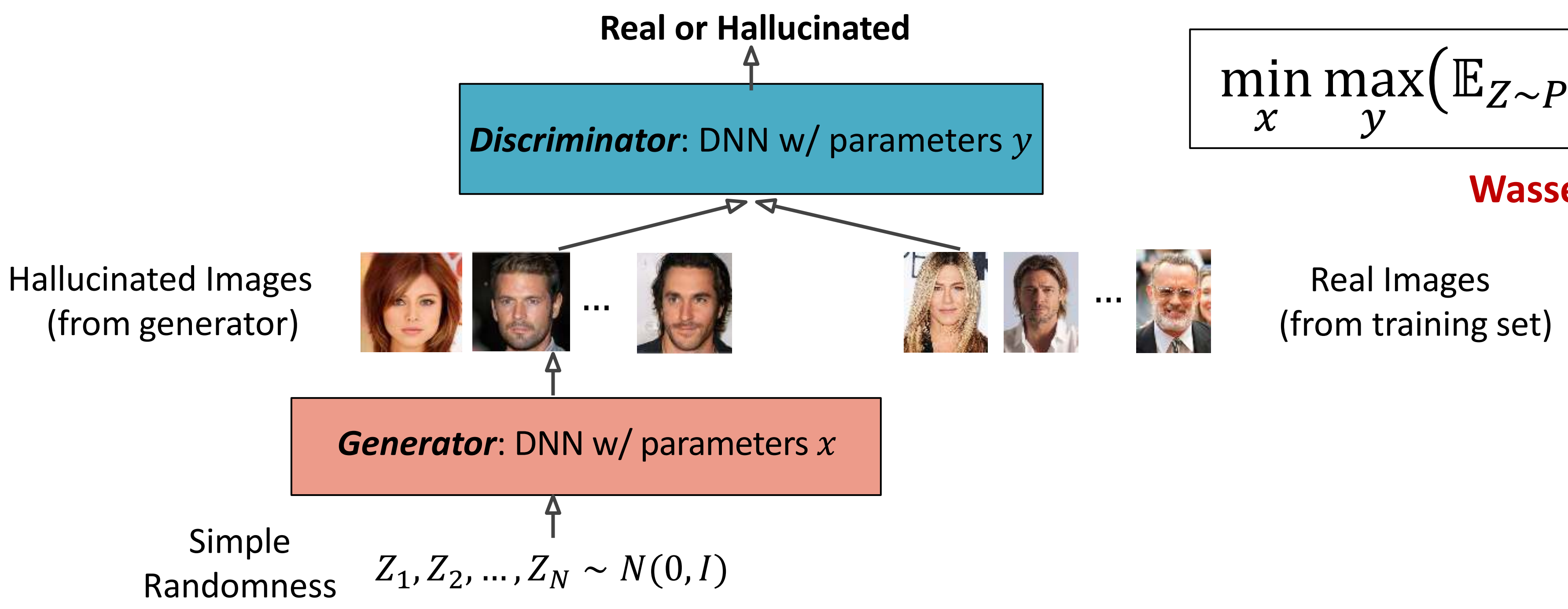
e.g. **GANs**, robust classification, 2-agent RL

Gradient Descent-Ascent (GDA) Dynamics:

$$x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t)$$
$$y_{t+1} = y_t + \eta \cdot \nabla_y f(x_t, y_t)$$

Generative Adversarial Nets (GANs) [Goodfellow et al'14]: $Z \sim \mathcal{N}(0, I) \longrightarrow \boxed{G_x(\cdot)} \longrightarrow$  $\sim P_{\text{interesting}}$

How? Set up a ***zero-game*** between a player tuning the parameters x of a *"Generator"* DNN and a player tuning the parameters y of a *"Discriminator"* DNN:



$$\min_x \max_y \left(\mathbb{E}_{Z \sim P_{real}} [D_y(Z)] - \mathbb{E}_{Z \sim N(0,I)} [D_y(G_x(Z))] \right)$$

Wassertsein GAN [Arjovsky-Chintala-Bottou'17]

Training Oscillations and/or Garbage Solutions: already in two-agent min-max settings

$$\boxed{\min_x \max_y f(x, y)}$$

typically f is not convex/concave; and x, y multidimensional

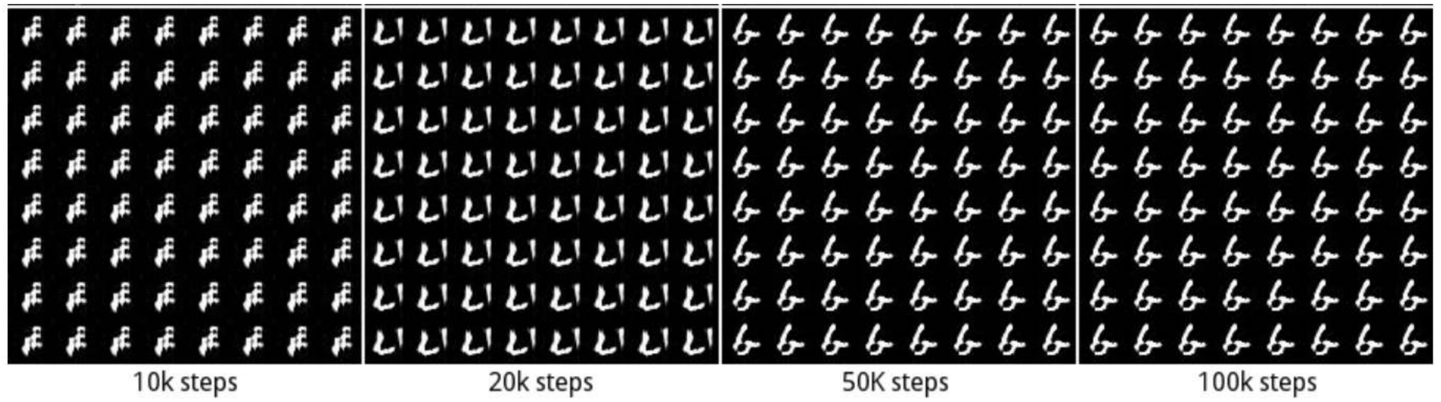
e.g. **GANs**, robust classification, 2-agent RL

Gradient Descent-Ascent (GDA) Dynamics:

$$x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t)$$
$$y_{t+1} = y_t + \eta \cdot \nabla_y f(x_t, y_t)$$

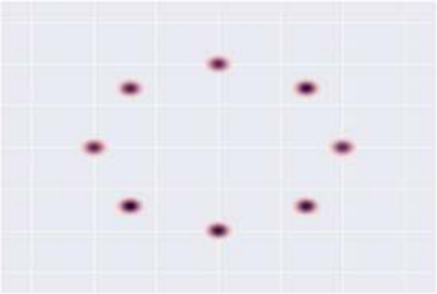
- GAN training on MNIST:

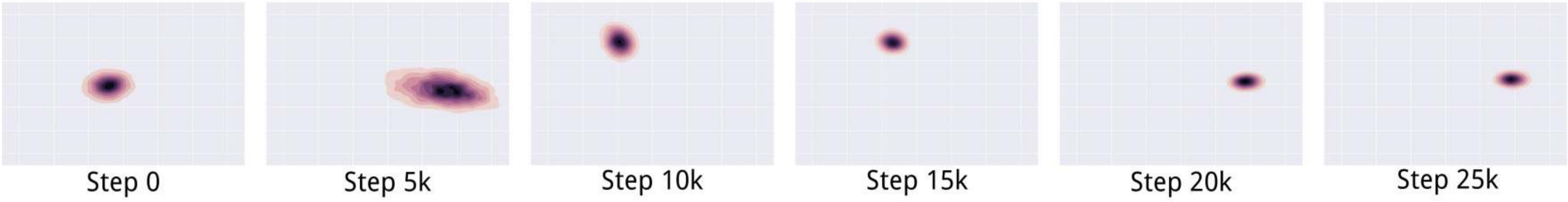
Target:



- GAN training on mixture of Gaussians:

Target:





pictures from **[Metz et al ICLR'17]**

Training Oscillations:
even for Gaussian data/bilinear objectives

- **True distribution:** isotropic Normal distribution, namely $X \sim \mathcal{N}\left(\begin{bmatrix}3\\4\end{bmatrix}, I_{2\times 2}\right)$

- **Generator architecture:** $G_{\boldsymbol{\theta}}(Z) = Z + \boldsymbol{\theta}$ (adds input Z to internal params)

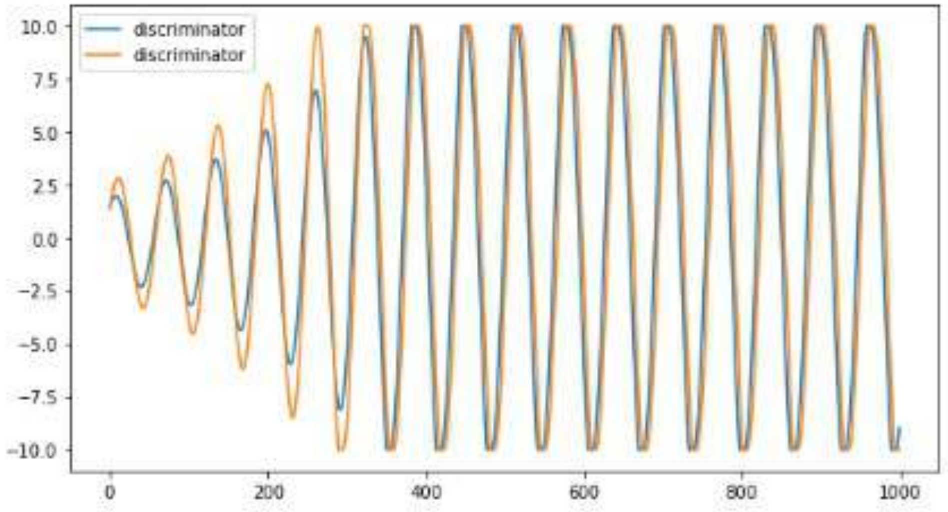
Z, θ, w: 2-dimensional

- **Discriminator architecture:** $D_{\boldsymbol{w}}(\cdot) = \langle \boldsymbol{w}, \cdot \rangle$ (linear projection)

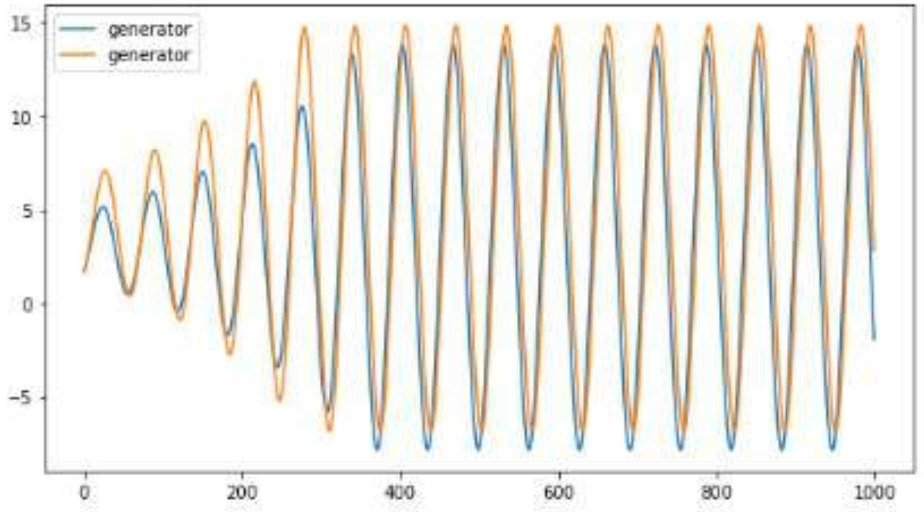
- **Wasserstein-GAN objective:** $\min_{\boldsymbol{\theta}} \max_{\boldsymbol{w}} \mathbb{E}_X[D_{\boldsymbol{w}}(X)] - \mathbb{E}_Z[D_{\boldsymbol{w}}(G_{\boldsymbol{\theta}}(Z))]$
(infinite samples)

$$= \min_{\boldsymbol{\theta}} \max_{\boldsymbol{w}} \boldsymbol{w}^{\mathrm{T}} \cdot \left(\begin{bmatrix}3\\4\end{bmatrix} - \boldsymbol{\theta}\right)$$

convex-concave function

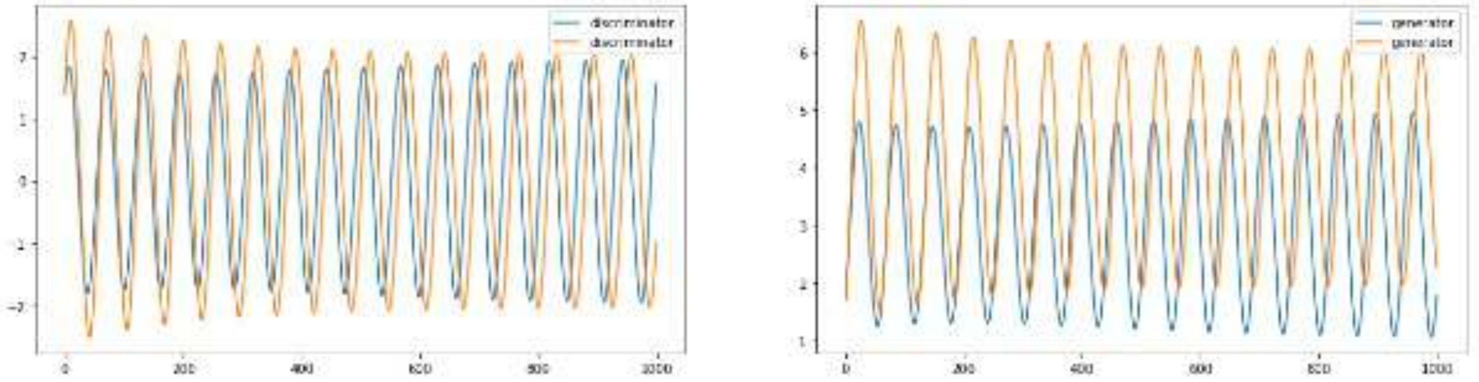


Gradient Descent Dynamics

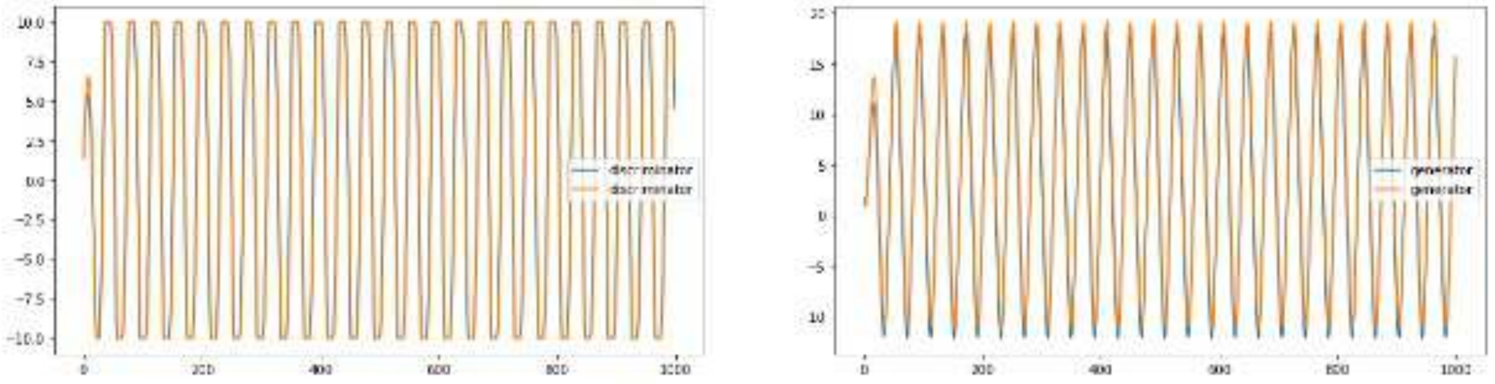


from **[Daskalakis, Ilyas, Syrgkanis, Zeng ICLR’18]**

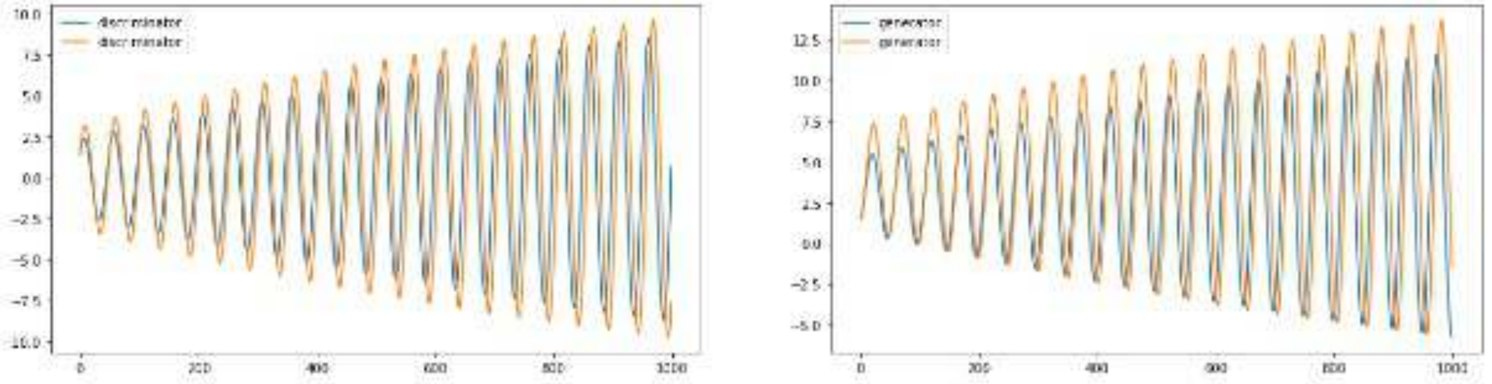
Training Oscillations:
persistence for variants of Gradient Descent/Ascent



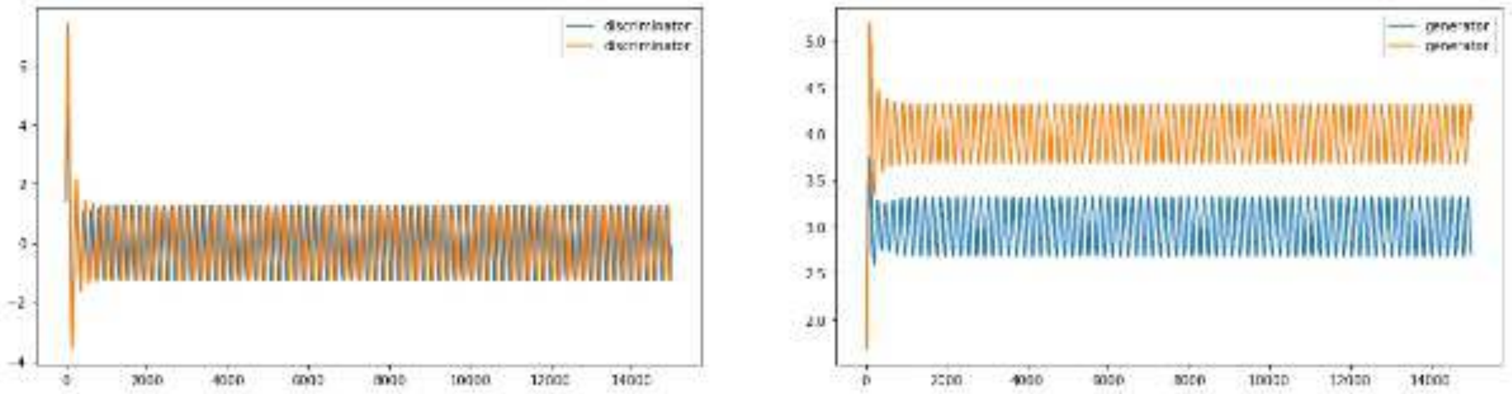
(a) GD dynamics with a gradient penalty added to the loss. $\eta = 0.1$ and $\lambda = 0.1$.



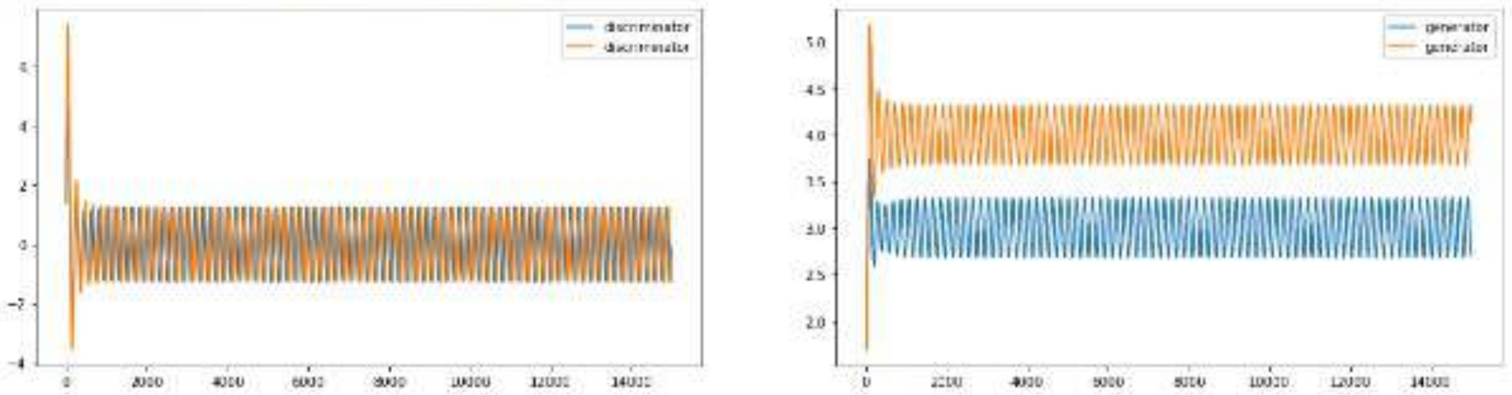
(b) GD dynamics with momentum. $\eta = 0.1$ and $\gamma = 0.5$.



(c) GD dynamics with momentum and gradient penalty. $\eta = .1$, $\gamma = 0.2$ and $\lambda = 0.1$.



(d) GD dynamics with momentum and gradient penalty, training generator every 15 training iterations of the discriminator. $\eta = .1$, $\gamma = 0.2$ and $\lambda = 0.1$.



(e) GD dynamics with Nesterov momentum and gradient penalty, training generator every 15 training iterations of the discriminator. $\eta = .1$, $\gamma = 0.2$ and $\lambda = 0.1$.

from **[Daskalakis, Ilyas, Syrgkanis, Zeng ICLR'18]**

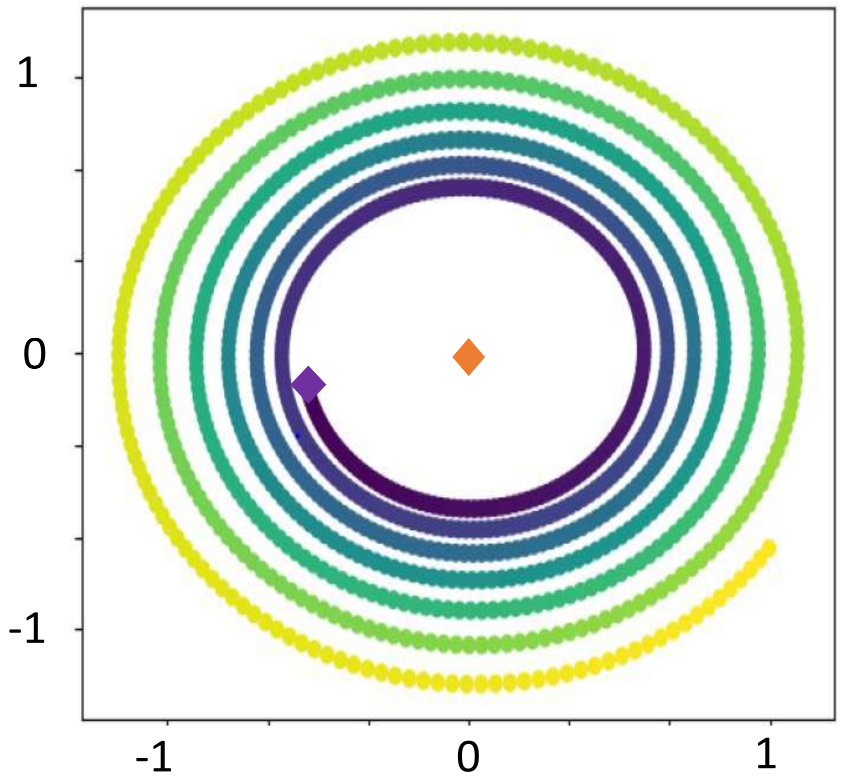
Training Oscillations:
the simplest oscillating min-max example

$$\min_x \max_y f(x, y)$$

Gradient Descent-Ascent (GDA) Dynamics:

$$x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t)$$
$$y_{t+1} = y_t + \eta \cdot \nabla_y f(x_t, y_t)$$

$$f(x, y) = x \cdot y$$



$$x_{t+1} = x_t - \eta \cdot y_t$$
$$y_{t+1} = y_t + \eta \cdot x_t$$

◆ : initialization

◆ : min-max equilibrium

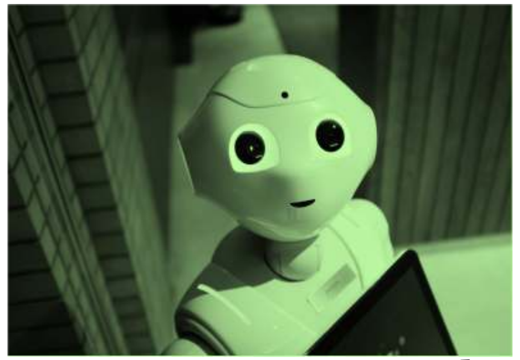
What gives?

- Training oscillations/garbage solutions arise:
  - even in two-agent, min-max settings
  - even when the objective is convex-concave, low-dimensional
  - even when the objective is perfectly known

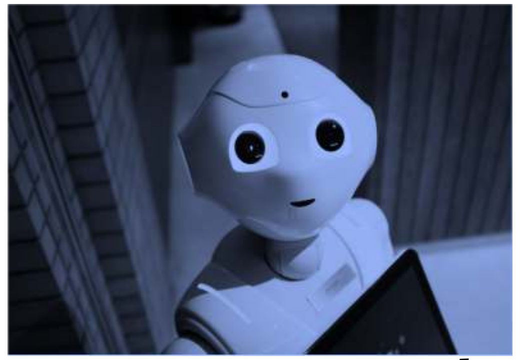
What gives?

- Training oscillations/garbage solutions arise:
  - even in two-agent, min-max settings
  - even when the objective is convex-concave, low-dimensional
  - even when the objective is perfectly known
- So good luck when:
  - the objective needs to be learned besides optimized
  - the objective is nonconvex-nonconcave, high-dimensional
  - the setting is multi-agent, multi-objective

Broad Focus: Equilibrium Learning

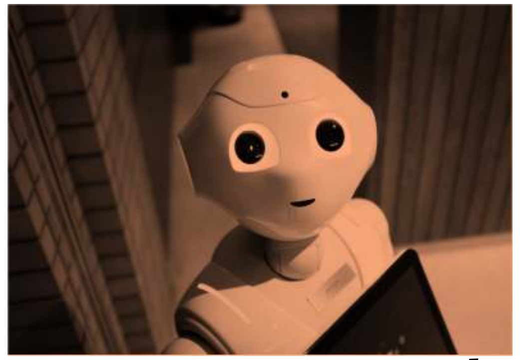


action: $x_1 \in \mathbb{R}^{d_1}$
goal: min $f_1(x_1, \dots, x_n)$



action: $x_2 \in \mathbb{R}^{d_2}$
goal: min $f_2(x_1, \dots, x_n)$

…



action: $x_n \in \mathbb{R}^{d_n}$
goal: min $f_n(x_1, \dots, x_n)$

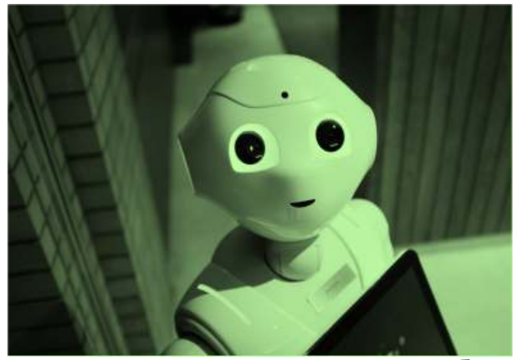
Sources of tension:

- x_{-i} may be imposing constraints on feasible x_i
- each f_i depends on the whole $\vec{x}$, yet
  - $f_1, \dots, f_n$ may be misaligned
  - players may be uncoordinated in choosing actions and may have partial observability of actions/payoffs/information of others

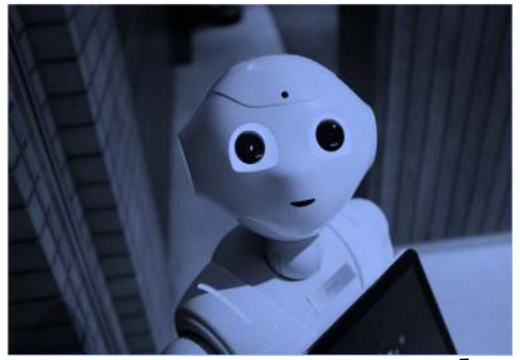
Game theory:

- offers *solution concepts*, such as Nash or correlated equilibrium, to predict what might reasonably happen
- but *is GD or variants going to get there*?

Broad Focus: Equilibrium Learning

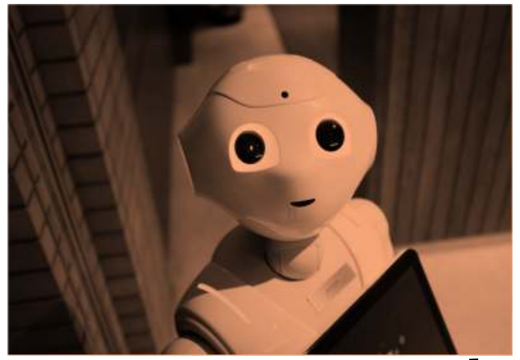


action: $x_1 \in \mathbb{R}^{d_1}$
goal: min $f_1(x_1, \ldots, x_n)$



action: $x_2 \in \mathbb{R}^{d_2}$
goal: min $f_2(x_1, \ldots, x_n)$

…



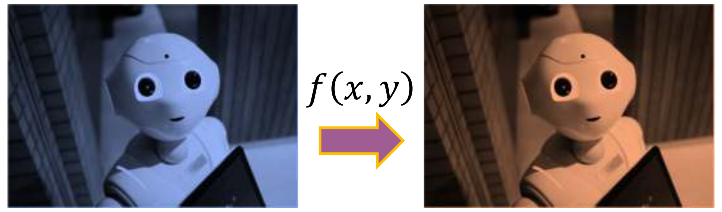
action: $x_n \in \mathbb{R}^{d_n}$
goal: min $f_n(x_1, \ldots, x_n)$

Main Question: *When each agent uses Gradient Descent (or some other learning dynamics), will the strategy profile converge to some Nash, correlated equilibrium, or other meaningful solution concept?*

Important consideration: is f_i *convex* in x_i (***convex game***) or not (***nonconvex game***) ?

- without convexity even equilibrium existence is at risk!
- even *with* convexity, Nash equilibrium is intractable **[Daskalakis-Goldberg-Papadimitriou'06, Chen-Deng'06]** so consider alternatives such as (coarse) correlated equilibrium / minimizing regret / …

Main Focus: Min-Max Optimization



$$\min_x \max_y f(x,y)$$

s.t. $\quad (x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- f: Lipschitz, L-smooth (i.e. ∇f is L-Lipschitz)
- constraint set S: convex, compact

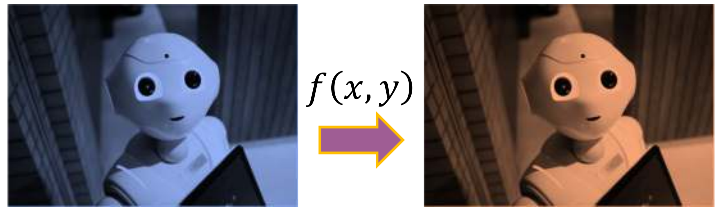
I will view the game as ***simultaneous***

sequential games are also important in GT and ML
and no harder computationally
c.f. **[Jin-Netrapali-Jordan ICML'20] [Mangoubi-Vishnoi STOC'21]**

Main Focus: Minimization vs Min-Max Optimization

$$\min_x f(x)$$
s.t. $x \in S \subset \mathbb{R}^d$

vs



$$\min_x \max_y f(x,y)$$
s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

(I view the game as ***simultaneous***)

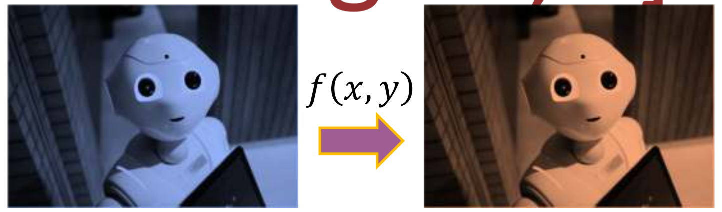
- f: Lipschitz, L-smooth (i.e. ∇f is L-Lipschitz)
- constraint set S: convex, compact

Minimization vs Min-Max Optimization

the classical setting [von Neumann'28, Dantzig'48,…]

$$\min_x f(x)$$
s.t. $x \in S \subset \mathbb{R}^d$

- f: **convex**, L-smooth
- constraint set: convex, compact

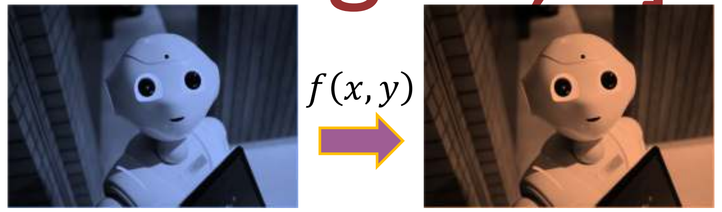


$$\min_x \max_y f(x, y)$$
s.t. $(x, y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- f: **convex in x & concave in y**, L-smooth
- constraint set: convex, compact

Minimization vs Min-Max Optimization

the classical setting [von Neumann’28, Dantzig’48,…]



$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$

$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- f: **convex**, L-smooth
- constraint set: convex, compact

- f: **convex in x & concave in y**, L-smooth
- constraint set: convex, compact

Theorem [standard]
First-order methods find approximate minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, L, diameter of S.

Theorem [standard]
First-order methods find approximate min-max equilibria, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, L, diameter of S.

$$f(x^*) \leq f(x) + \varepsilon, \forall x \in S$$

$$f(x^*,y) - \varepsilon \leq f(x^*,y^*) \leq f(x,y^*) + \varepsilon$$

$\forall y$ s.t. $(x^*,y) \in S$

$\forall x$ s.t. $(x,y^*) \in S$

Minimization vs Min-Max Optimization

the classical setting [von Neumann'28, Dantzig'48,…]

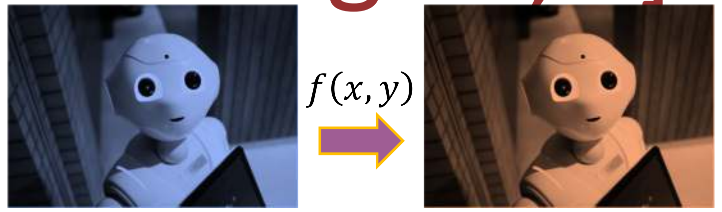
$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$

- f: **convex**, L-smooth
- constraint set: convex, compact

Theorem [standard]
First-order methods find approximate minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, L, diameter of S.

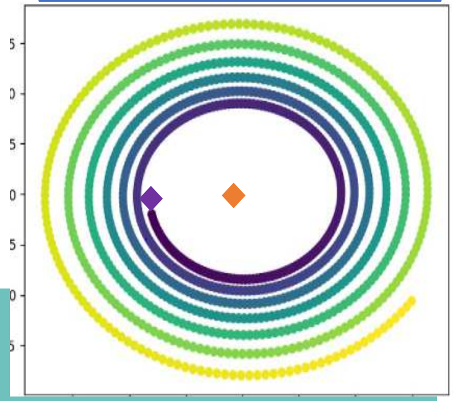
$$f(x^*) \leq f(x) + \varepsilon, \forall x \in S$$



$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

$f(x,y) = x \cdot y$



- f: **convex in x & concave in y**, L-smooth
- constraint set: convex, compact

Theorem [standard]
First-order methods find approximate min-max equilibria, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, L, diameter of S.

Training oscillations of GDA here not due to computational intractability, but are feature of training method; *can they be removed?*

Minimization vs Min-Max Optimization

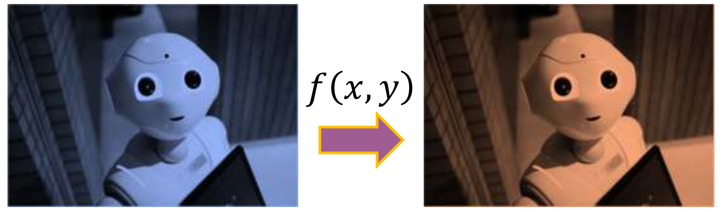
the modern setting

$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$

➢ f: Lipschitz, L-smooth

➢ constraint set S: convex, compact



$$\min_x \max_y f(x, y)$$

s.t. $(x, y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

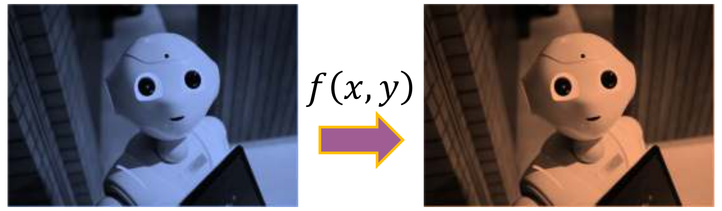
(I view the game as ***simultaneous***)

?

Minimization vs Min-Max Optimization

the *modern* *setting*

$$\min_x f(x)$$
s.t. $\quad x \in S \subset \mathbb{R}^d$



$$\min_x \max_y f(x,y)$$
s.t. $\quad (x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

(I view the game as ***simultaneous***)

- ➢ f: Lipschitz, L-smooth
- ➢ constraint set S: convex, compact

it's intractable (NP-hard) to find global optima

& global optima may not even exist in the RHS

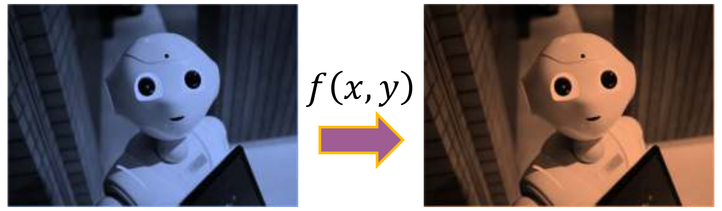
but, <u>how about *local* optima</u>?

Minimization vs Min-Max Optimization

the modern setting

$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$



$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- f: Lipschitz, L-smooth
- constraint set S: convex, compact

$B_\delta(x^*)$={x s.t. $\|x - x^*\| \le \delta\}$

$B_\delta(y^*)$={y s.t. $\|y - y^*\| \le \delta\}$

Def: (ε,δ)-*local* minimum

$$f(x^*) \le f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Def: (ε,δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*, y) - \varepsilon \le f(x^*, y^*) \le f(x, y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*, y) \in S$

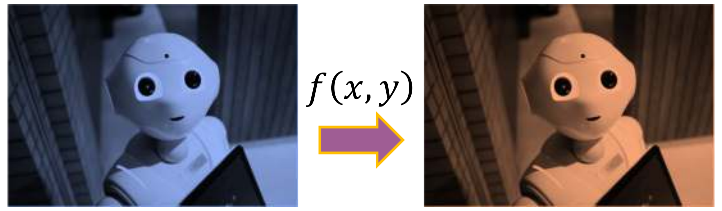
$\forall x \in B_\delta(x^*)$ s.t. $(x, y^*) \in S$

Minimization vs Min-Max Optimization

the modern setting

$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$



$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- constraint set S: convex, compact

$\boxed{B_\delta(x^*) = \{x \text{ s.t. } \|x - x^*\| \leq \delta\}}$

$\boxed{B_\delta(y^*) = \{y \text{ s.t. } \|y - y^*\| \leq \delta\}}$

Def: (ε, δ)-*local* minimum

$$f(x^*) \leq f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Def: (ε, δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*, y) - \varepsilon \leq f(x^*, y^*) \leq f(x, y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*, y) \in S$

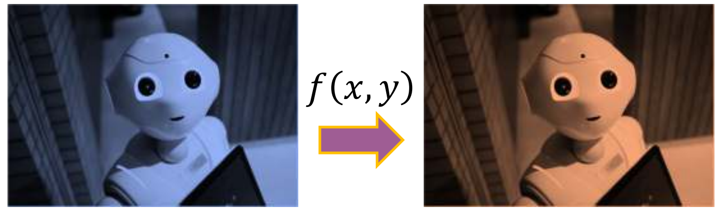
$\forall x \in B_\delta(x^*)$ s.t. $(x, y^*) \in S$

Minimization vs Min-Max Optimization

the modern setting

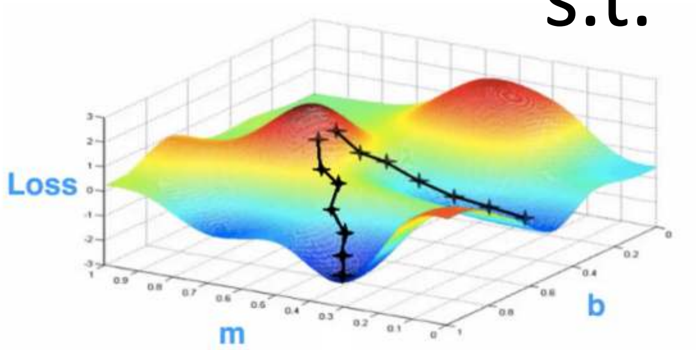
$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$



$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$



- ➢ f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- ➢ constraint set S: convex, compact

$B_\delta(x^*)$={x s.t. $\|x - x^*\| \le \delta$}

$B_\delta(y^*)$={y s.t. $\|y - y^*\| \le \delta$}

Def: (ε,δ)-*local* minimum

$$f(x^*) \le f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Theorem [folklore]

If $\delta \le \sqrt{2\varepsilon/L}$, first-order methods find (ε,δ)-*local* minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, smoothness of f.

Def: (ε,δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

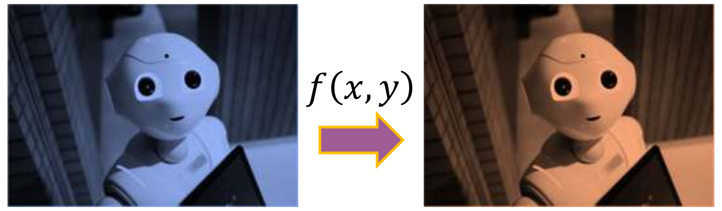
$$f(x^*,y) - \varepsilon \le f(x^*,y^*) \le f(x,y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*,y) \in S$

$\forall x \in B_\delta(x^*)$ s.t. $(x,y^*) \in S$

Minimization vs Min-Max Optimization

the modern setting

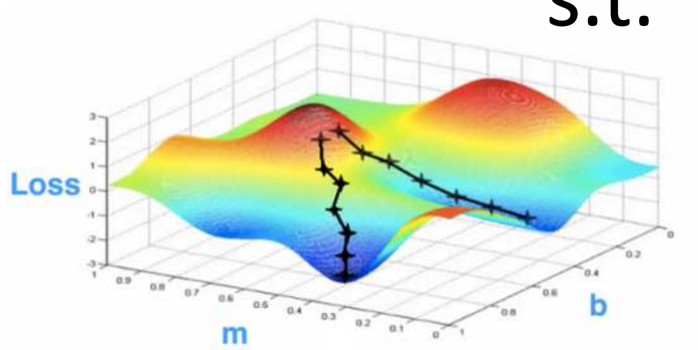


$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$

$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$



- f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- constraint set S: convex, compact

$B_\delta(x^*)=\{x \text{ s.t. } \|x - x^*\| \le \delta\}$

$B_\delta(y^*)=\{y \text{ s.t. } \|y - y^*\| \le \delta\}$

Def: (ε, δ)-*local* minimum

$$f(x^*) \le f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Theorem [folklore]

If $\delta \le \sqrt{2\varepsilon/L}$, first-order methods find (ε, δ)-*local* minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, smoothness of f.

(for larger δ existence holds, but problem becomes **NP**-hard)

Def: (ε, δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*,y) - \varepsilon \le f(x^*,y^*) \le f(x,y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*,y) \in S$

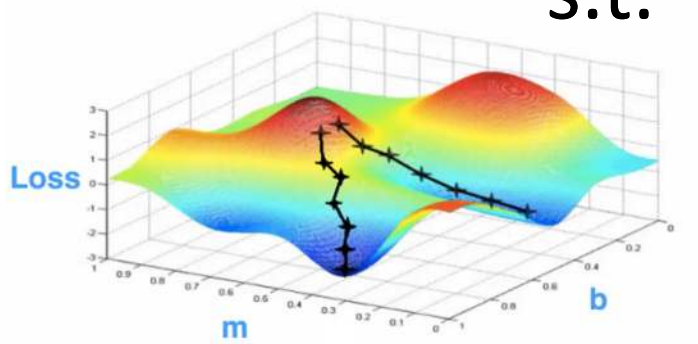
$\forall x \in B_\delta(x^*)$ s.t. $(x,y^*) \in S$

Minimization vs Min-Max Optimization

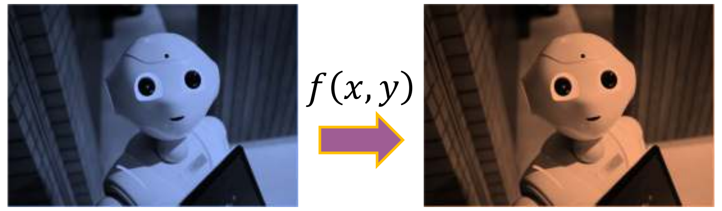
the modern setting

$$\min_x f(x)$$

s.t. $\quad x \in S \subset \mathbb{R}^d$



- ➢ f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- ➢ constraint set S: convex, compact



$$\min_x \max_y f(x,y)$$

s.t. $\quad (x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

$B_\delta(x^*)$={x s.t. $\|x - x^*\| \leq \delta$}

$B_\delta(y^*)$={y s.t. $\|y - y^*\| \leq \delta$}

Def: (ε, δ)-*local* minimum

$$f(x^*) \leq f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Theorem [folklore]

If $\delta \leq \sqrt{2\varepsilon/L}$, first-order methods find (ε, δ)-*local* minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, smoothness of f.

(for larger δ existence holds, but problem becomes **NP**-hard)

Def: (ε, δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*, y) - \varepsilon \leq f(x^*, y^*) \leq f(x, y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*, y) \in S$ $\qquad \forall x \in B_\delta(x^*)$ s.t. $(x, y^*) \in S$

exist for small enough $\delta \leq \sqrt{2\varepsilon/L}$

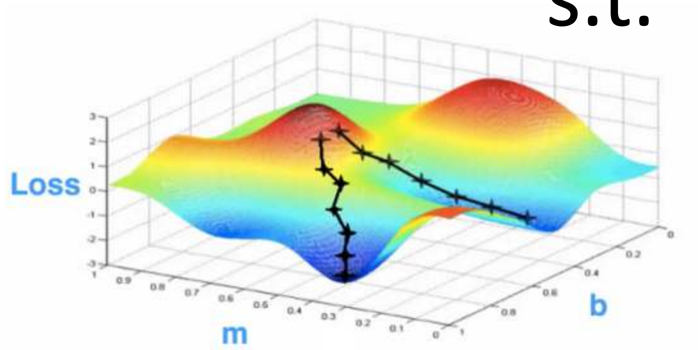
complexity ????

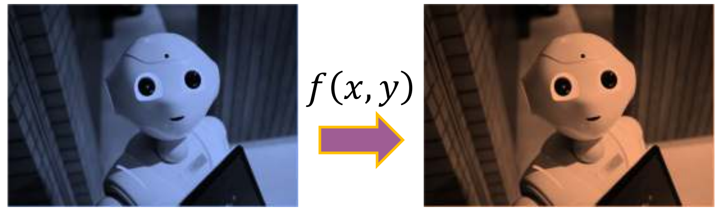
Minimization vs Min-Max Optimization

the modern setting

$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$





$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- ➢ f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- ➢ constraint set S: convex, compact

$B_\delta(x^*)$={x s.t. $\|x - x^*\| \le \delta$}

$B_\delta(y^*)$={y s.t. $\|y - y^*\| \le \delta$}

Def: (ε,δ)-*local* minimum

$$f(x^*) \le f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Theorem [folklore]

If $\delta \le \sqrt{2\varepsilon/L}$, first-order methods find (ε,δ)-*local* minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, smoothness of f.

(for larger δ existence holds, but problem becomes **NP**-hard)

Def: (ε,δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*,y) - \varepsilon \le f(x^*,y^*) \le f(x,y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*,y) \in S$ $\qquad$ $\forall x \in B_\delta(x^*)$ s.t. $(x,y^*) \in S$

exist for small enough $\delta \le \sqrt{2\varepsilon/L}$

complexity ????

Training oscillations here could be due to computational intractability; ***are they?***

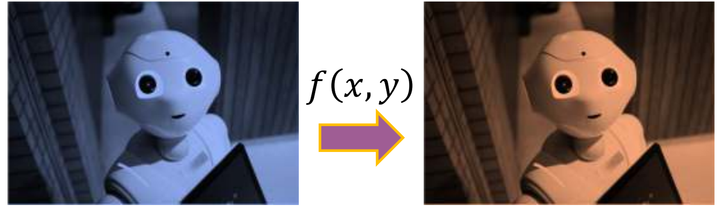
Menu

- **Motivation**
- Convex Games
  - remove training oscillations?
- Nonconvex Games
  - are oscillations inherent/reflective of intractability?
- Conclusions

Menu

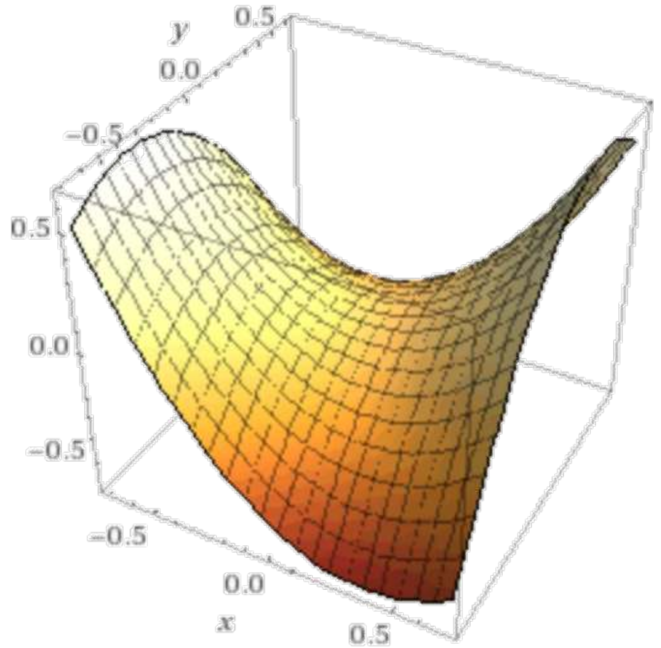
- **Motivation**
- **Convex Games**
  - **remove training oscillations?**
- Nonconvex Games
  - are oscillations inherent/reflective of intractability?
- Conclusions

Convex *Two-Player Zero-Sum* Games *theoretical bearings*

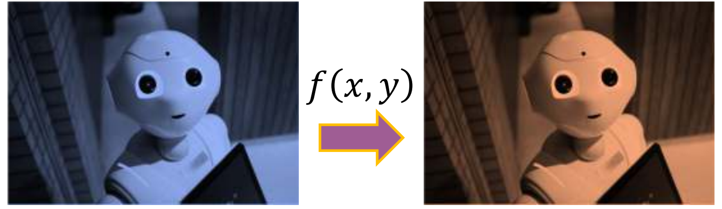


f: convex in x
& concave in y

- **[von Neumann 1928]:** If $X \subset \mathbb{R}^n, Y \subset \mathbb{R}^m$ are compact and convex, and $f: X \times Y \to \mathbb{R}$ is continuous and convex-concave (i.e. $f(x,y)$ is convex in x for all y and is concave in y for all x), then
$$\min_{x \in X} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in X} f(x,y)$$
- Min-max optimal point (x,y) is essentially unique (unique if f is strictly convex-concave, o.w. a convex set of solutions); value always unique
- E.g. $f(x,y) = x^2 - y^2 + x \cdot y$



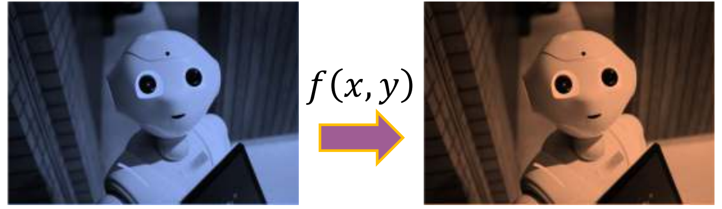
Convex *Two-Player Zero-Sum* Games *theoretical bearings*



f: convex in x
& concave in y

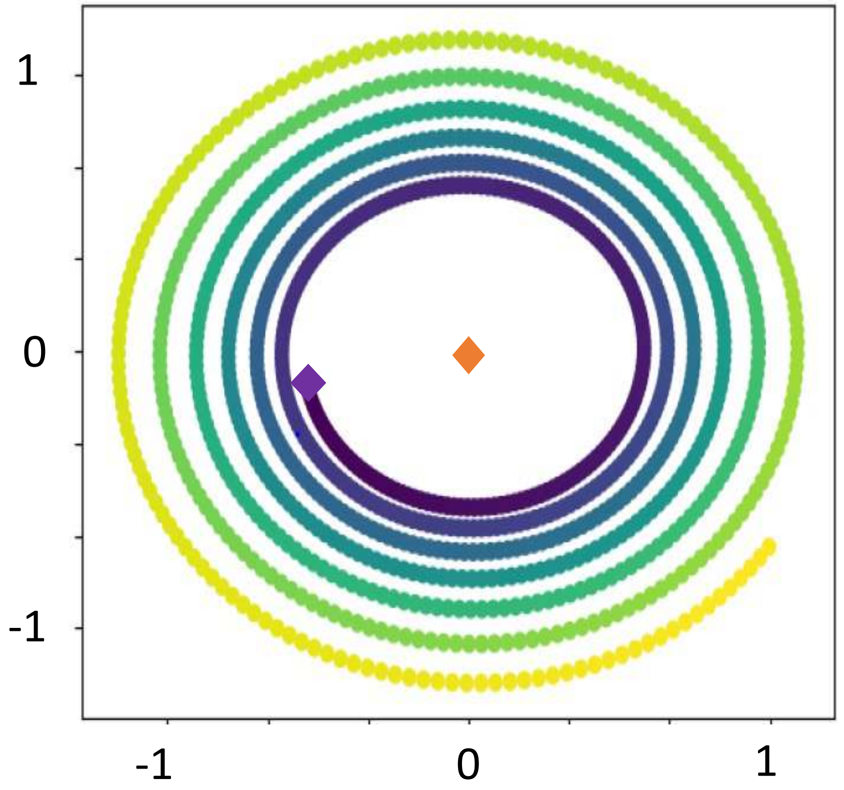
- **[von Neumann 1928]:** If $X \subset \mathbb{R}^n, Y \subset \mathbb{R}^m$ are compact and convex, and $f: X \times Y \to \mathbb{R}$ is continuous and convex-concave (i.e. $f(x,y)$ is convex in x for all y and is concave in y for all x), then
$$\min_{x \in X} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in X} f(x,y)$$
- Min-max optimal point (x,y) is essentially unique (unique if f is strictly convex-concave, o.w. a convex set of solutions); value always unique
- Min-max points = equilibria of zero-sum game where min player pays max player $f(x,y)$
- von Neumann: "*As far as I can see, there could be no theory of games … without that theorem … I thought there was nothing worth publishing until the Minimax Theorem was proved*"
- When f is bilinear, i.e. $f(x,y) = x^{\mathrm{T}} A y + b^T x + c^T y$ and X, Y polytopes
    - **[von Neumann-Dantzig 1947, Adler IJGT'13]:** Minmax $\Leftrightarrow$ strong LP duality
    - min-max solutions can be found w/ Linear Programming and vice versa
- General convex-concave objectives: equivalence to strong convex duality
- **[Blackwell'56, Hannan'57,…]:** if min and max run *no-regret online learning* procedures (e.g. online gradient descent) then behavior will "converge" to equilibrium!

Convex *Two-Player Zero-Sum* Games
so what's the issue with GDA non-convergence?



f: convex in x
& concave in y

- E.g. $f(x,y) = x \cdot y$

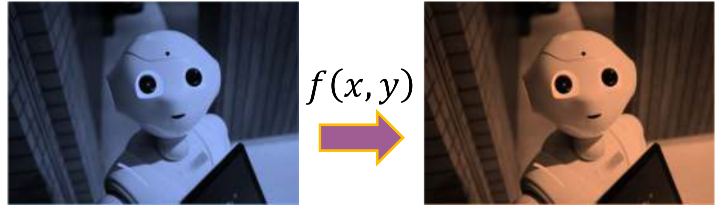


$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \end{aligned}}$$

◆ : start

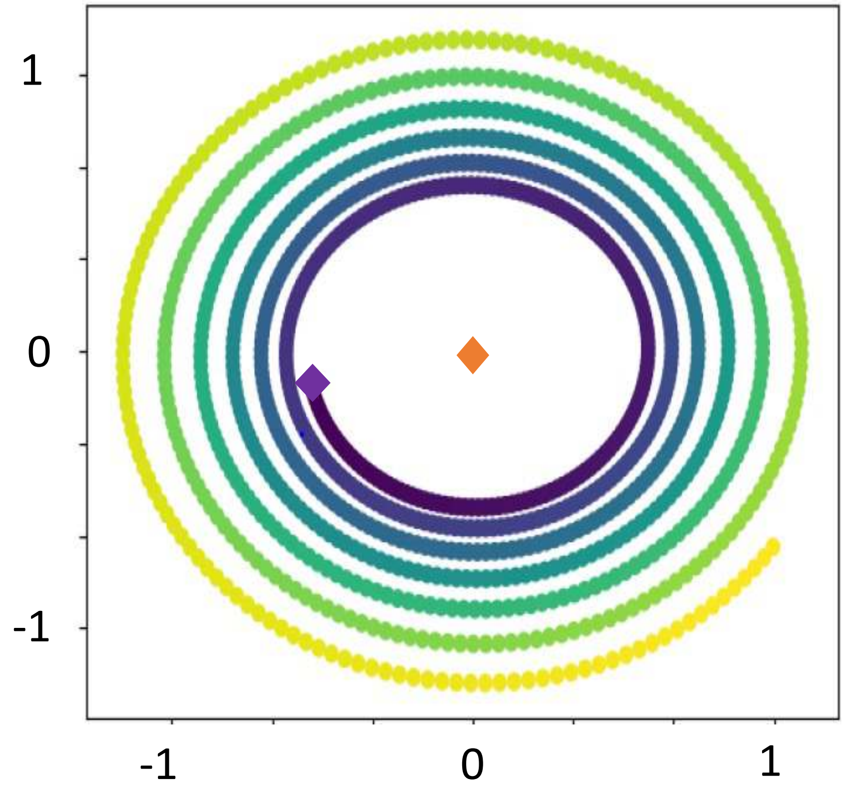
◆ : min-max equilibrium

Convex *Two-Player Zero-Sum* Games so what's the issue with GDA non-convergence?



f: convex in x
& concave in y

- E.g. $f(x, y) = x \cdot y$



$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \end{aligned}}$$

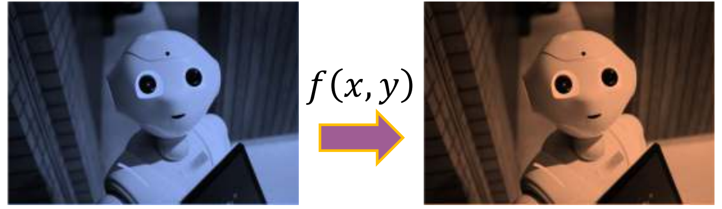
◆ : start

◆ : min-max equilibrium

$$\frac{1}{T}\sum_{t=1}^{T}(x_t, y_t) \to (x^*, y^*)$$

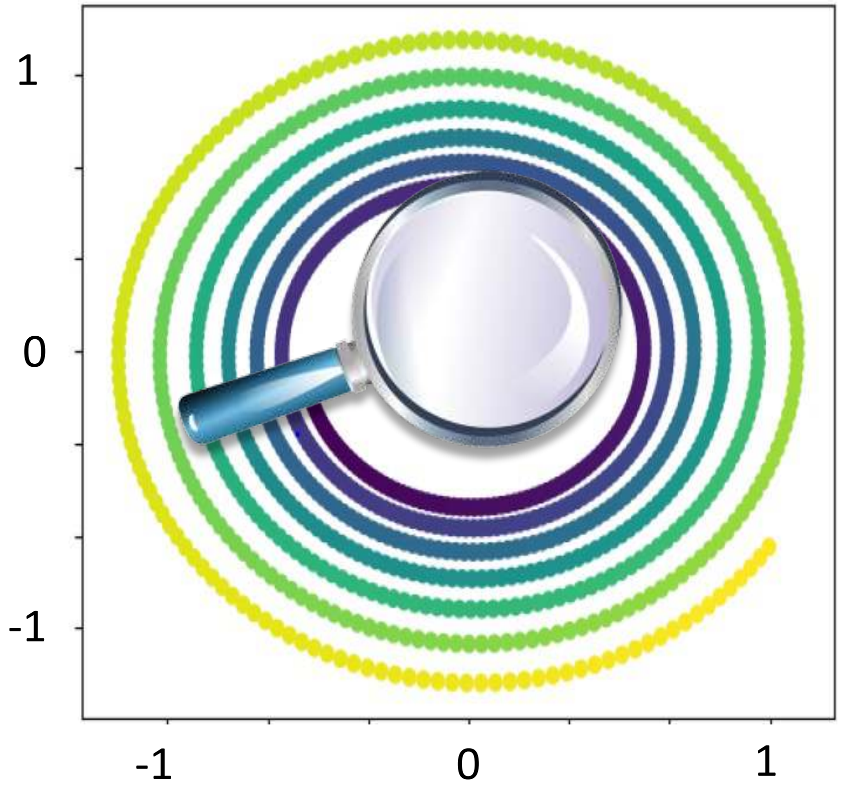
(typical of no-regret learners)

Convex *Two-Player Zero-Sum* Games so what's the issue with GDA non-convergence?



f: convex in x
& concave in y

- E.g. $f(x, y) = x \cdot y$



$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \end{aligned}}$$

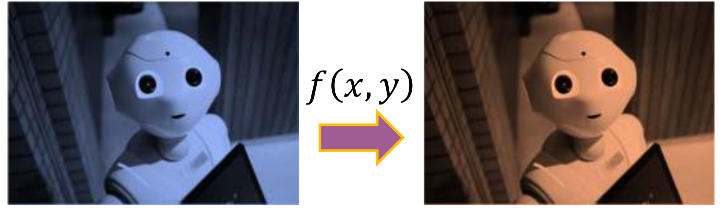
◆ : start

◆ : min-max equilibrium

$$\frac{1}{T}\sum_{t=1}^{T}(x_t, y_t) \to (x^*, y^*)$$

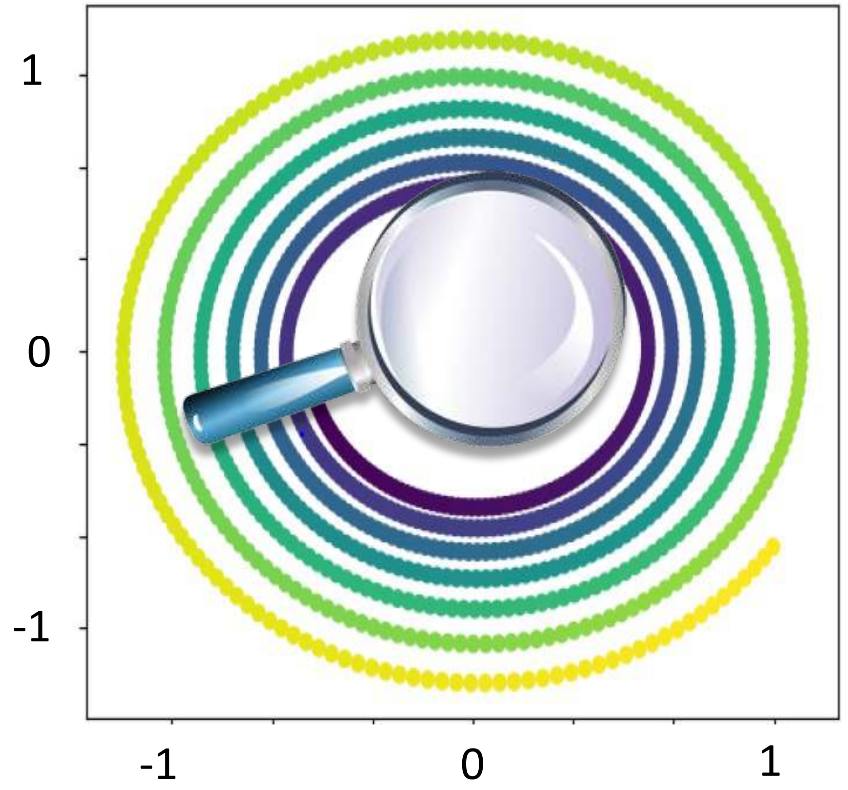
(typical of no-regret learners)

Convex *Two-Player Zero-Sum* Games so what's the issue with GDA non-convergence?



f: convex in x
& concave in y

- E.g. $f(x,y) = x \cdot y$



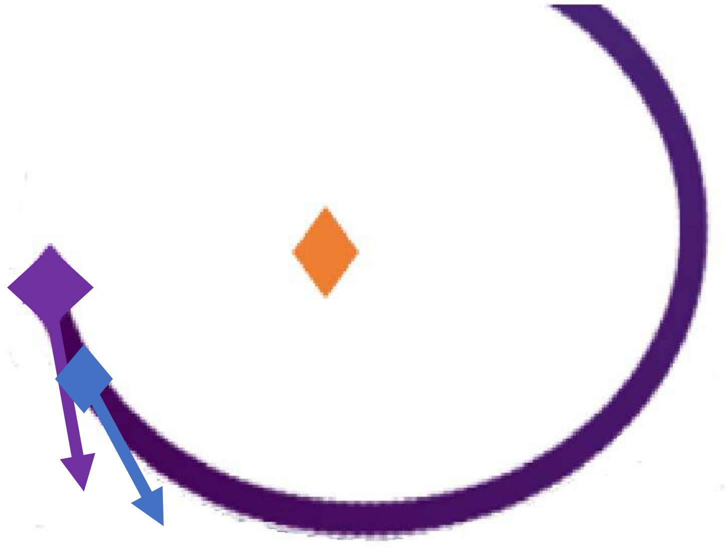
$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \end{aligned}}$$

◆ : start

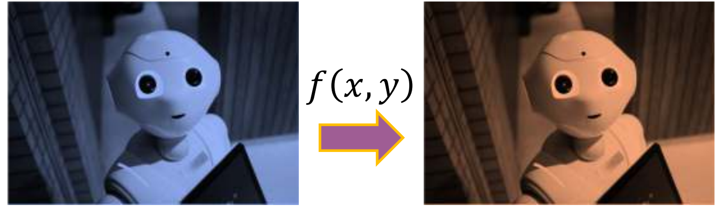
◆ : min-max equilibrium

$$\frac{1}{T}\sum_{t=1}^{T}(x_t, y_t) \to (x^*, y^*)$$

(typical of no-regret learners)

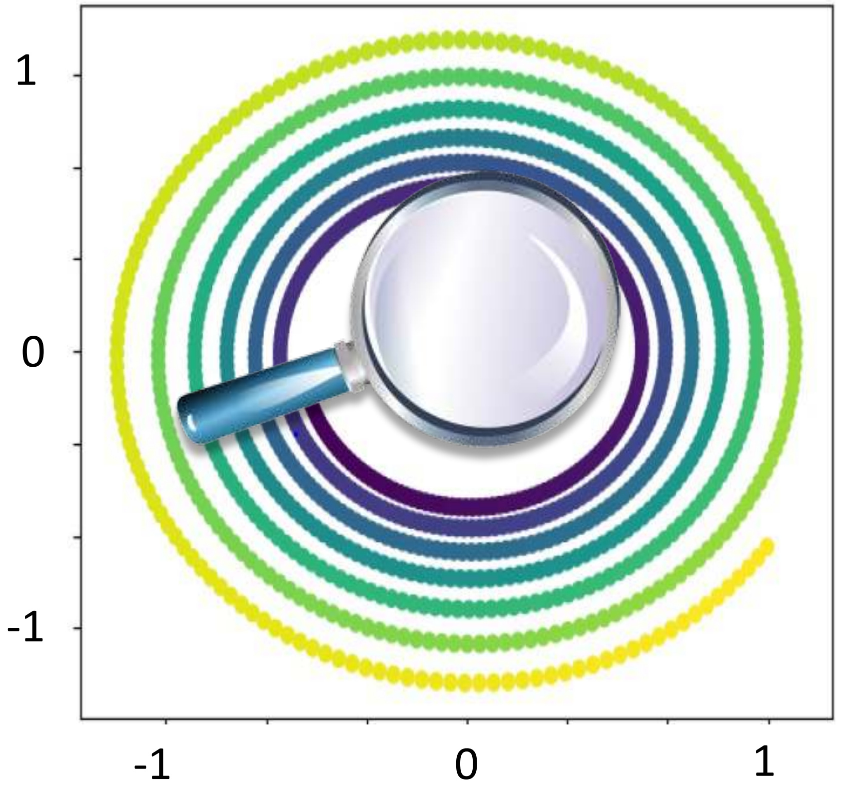


Convex *Two-Player Zero-Sum* Games
so what's the issue with GDA non-convergence?



f: convex in x
& concave in y

- E.g. $f(x,y) = x \cdot y$



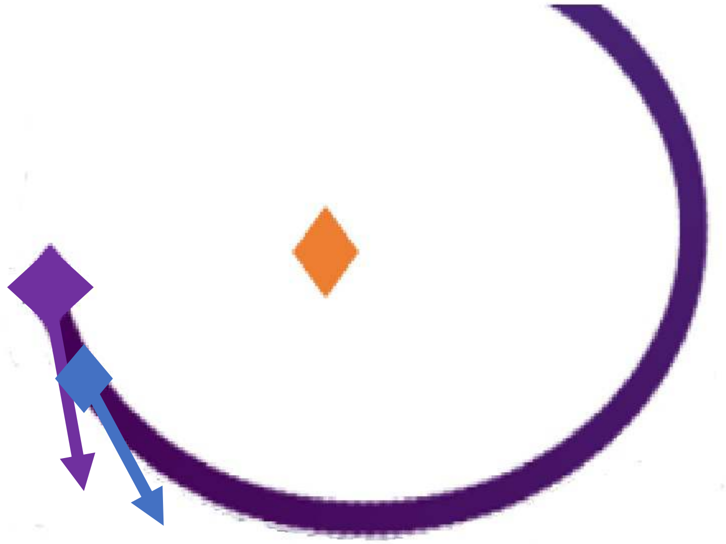
$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \end{aligned}}$$

◆ : start

◆ : min-max equilibrium

$$\frac{1}{T}\sum_{t=1}^{T}(x_t, y_t) \to (x^*, y^*)$$

(typical of no-regret learners)



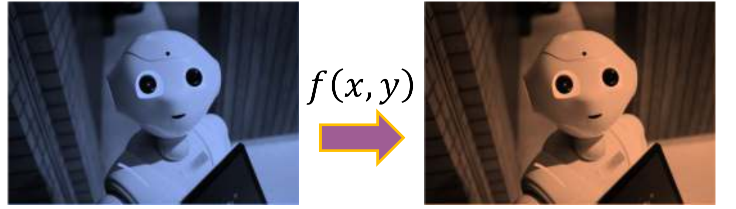
Correcting the momentum?

[Daskalakis, Ilyas, Syrgkanis, Zeng ICLR'18]

$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ &\quad + \boldsymbol{\eta/2 \cdot \nabla_x f(x_{t-1}, y_{t-1})} \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \\ &\quad - \boldsymbol{\eta/2 \cdot \nabla_y f(x_{t-1}, y_{t-1})} \end{aligned}}$$

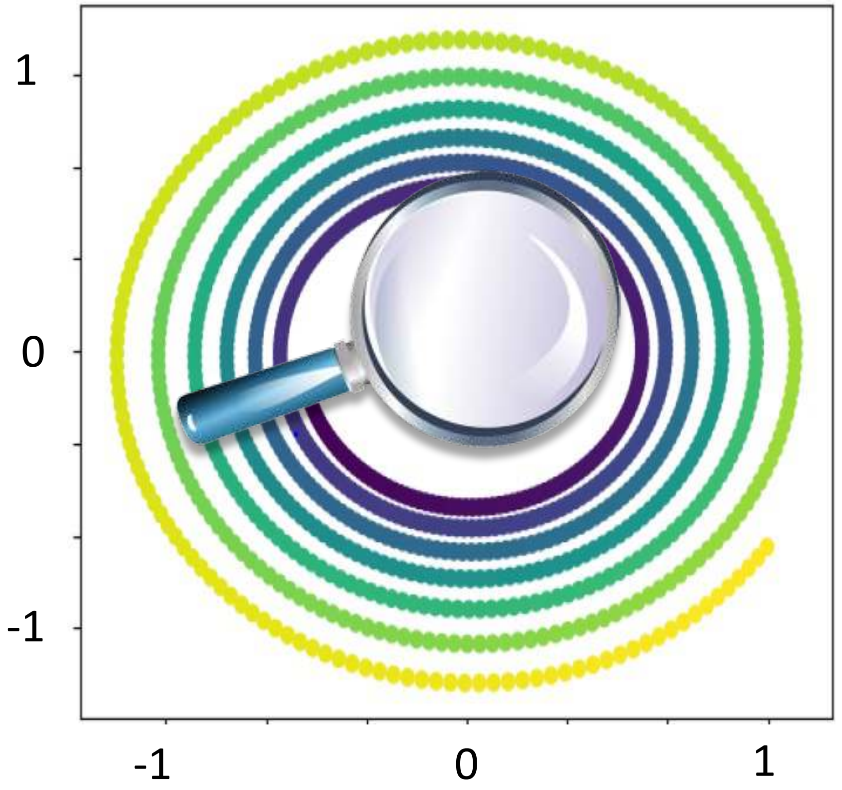
[Popov'80]

Convex *Two-Player Zero-Sum* Games so what's the issue with GDA non-convergence?



f: convex in x & concave in y

- E.g. $f(x,y) = x \cdot y$



$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \end{aligned}}$$

◆ : start

◆ : min-max equilibrium

$$\frac{1}{T}\sum_{t=1}^{T}(x_t, y_t) \to (x^*, y^*)$$

(typical of no-regret learners)

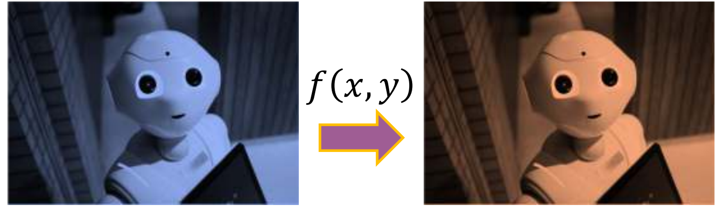
Correcting the momentum?

[Daskalakis, Ilyas, Syrgkanis, Zeng ICLR'18]

$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ &\quad + \eta/2 \cdot \nabla_x f(x_{t-1}, y_{t-1}) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \\ &\quad - \eta/2 \cdot \nabla_y f(x_{t-1}, y_{t-1}) \end{aligned}}$$

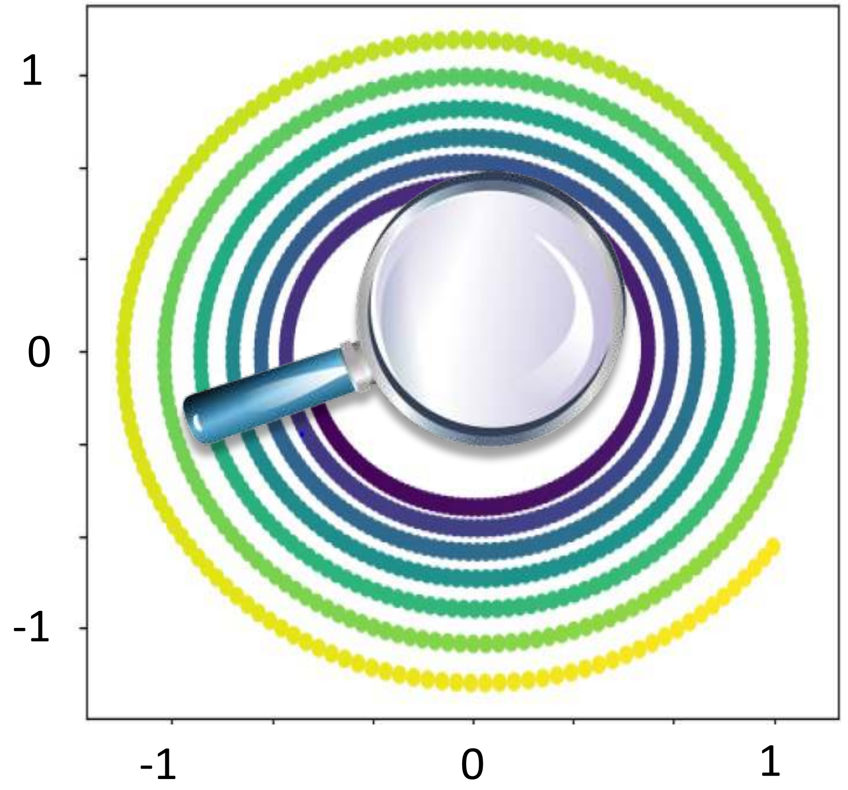
[Popov'80]

Convex *Two-Player Zero-Sum* Games
so what's the issue with GDA non-convergence?



f: convex in x
& concave in y

- E.g. $f(x,y) = x \cdot y$



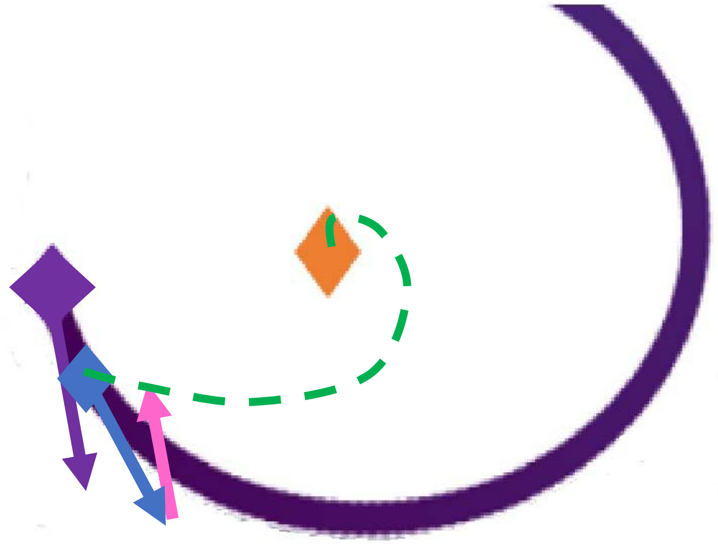
$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \end{aligned}}$$

◆ : start

◆ : min-max equilibrium

$$\frac{1}{T}\sum_{t=1}^{T}(x_t, y_t) \to (x^*, y^*)$$

(typical of no-regret learners)



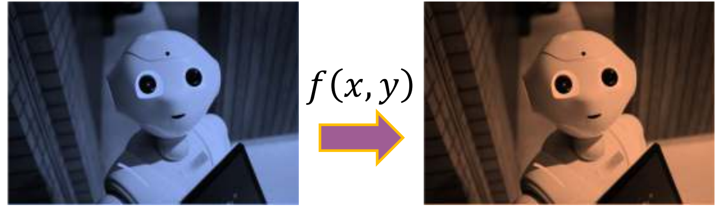
Correcting the momentum?

[Daskalakis, Ilyas, Syrgkanis, Zeng ICLR'18]

$$\boxed{\begin{aligned} x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ &\qquad + \eta/2 \cdot \nabla_x f(x_{t-1}, y_{t-1}) \\ \\ y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \\ &\qquad - \eta/2 \cdot \nabla_y f(x_{t-1}, y_{t-1}) \end{aligned}}$$

[Popov'80]

Convex *Two-Player Zero-Sum* Games correcting the momentum



f: convex in x
& concave in y

Optimistic GDA [Popov'80]

$$x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ + \boldsymbol{\eta/2 \cdot \nabla_x f(x_{t-1}, y_{t-1})}$$

$$y_{t+1} = y_t + \eta \cdot \nabla_y f(x_t, y_t) \\ - \boldsymbol{\eta/2 \cdot \nabla_y f(x_{t-1}, y_{t-1})}$$

Extra-Gradient Method [Korpelevich'76]

$$\boldsymbol{x_{t+1/2}} = x_t - \eta \cdot \nabla_x f(x_t, y_t)$$

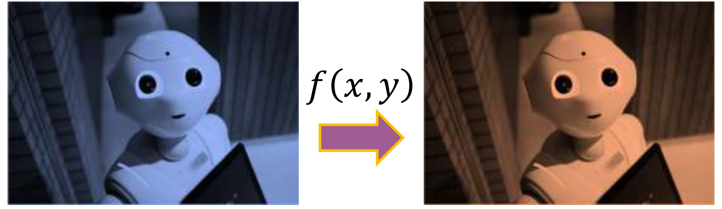
$$x_{t+1} = x_t - \eta \cdot \nabla_x f\left(\boldsymbol{x_{t+1/2}}, \boldsymbol{y_{t+1/2}}\right)$$

$$\boldsymbol{y_{t+1/2}} = y_t + \eta \cdot \nabla_y f(x_t, y_t)$$

$$y_{t+1} = y_t + \eta \cdot \nabla_y f\left(\boldsymbol{x_{t+1/2}}, \boldsymbol{y_{t+1/2}}\right)$$

- **[Korpelevich'76, Popov'80, Facchinei-Pang'03]:** Asymptotic ***last-iterate*** convergence results for Optimistic GDA, Extra-Gradient, Mirror-Prox, and related methods when f *is* ***convex-concave***

Convex *Two-Player Zero-Sum* Games
correcting the momentum



f: convex in x
& concave in y

Optimistic GDA [Popov'80]

$$x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t) + \boldsymbol{\eta/2 \cdot \nabla_x f(x_{t-1}, y_{t-1})}$$

$$y_{t+1} = y_t + \eta \cdot \nabla_y f(x_t, y_t) - \boldsymbol{\eta/2 \cdot \nabla_y f(x_{t-1}, y_{t-1})}$$

Extra-Gradient Method [Korpelevich'76]

$$\boldsymbol{x_{t+1/2}} = x_t - \eta \cdot \nabla_x f(x_t, y_t)$$

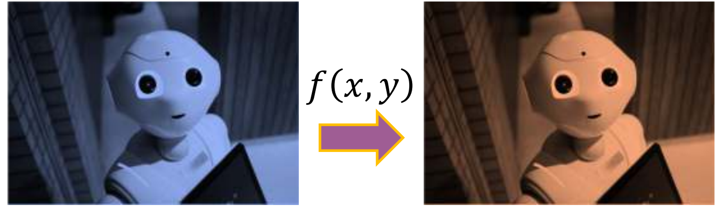
$$x_{t+1} = x_t - \eta \cdot \nabla_x f\left(\boldsymbol{x_{t+1/2}}, \boldsymbol{y_{t+1/2}}\right)$$

$$\boldsymbol{y_{t+1/2}} = y_t + \eta \cdot \nabla_y f(x_t, y_t)$$

$$y_{t+1} = y_t + \eta \cdot \nabla_y f\left(\boldsymbol{x_{t+1/2}}, \boldsymbol{y_{t+1/2}}\right)$$

- **[Korpelevich'76, Popov'80, Facchinei-Pang'03]:** Asymptotic ***last-iterate*** convergence results for Optimistic GDA, Extra-Gradient, Mirror-Prox, and related methods when f *is* ***convex-concave***
- Rates?
  - unconstrained setting: quite clear understanding **[Tseng'95, Daskalakis-Ilyas-Syrgkanis-Zeng ICLR'18, Liang-Stokes AISTATS'19, Gidel et al AISTATS'19, Mokhtari et al '19, Liang-Stokes AISTATS'19, Mokhtari et al '19, Azizian et al AISTATS'20, Golowich-Pattathil- Daskalakis-Ozdaglar COLT'20, Golowich-Pattathil-Daskalakis NeurIPS'20,…]**
  - constrained setting: mostly unclear **[Korpelevich'76;Tseng'95;Daskalakis-Panageas'19;Lee-Luo-Wei-Zhang'20]**

Convex *Two-Player Zero-Sum* Games correcting the momentum



f: convex in x
& concave in y

Optimistic GDA [Popov'80]

$$x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t) \\ \quad + \boldsymbol{\eta/2 \cdot \nabla_x f(x_{t-1}, y_{t-1})}$$

$$y_{t+1} = y_t + \eta \cdot \nabla_y f(x_t, y_t) \\ \quad - \boldsymbol{\eta/2 \cdot \nabla_y f(x_{t-1}, y_{t-1})}$$

Extra-Gradient Method [Korpelevich'76]

$$\boldsymbol{x_{t+1/2}} = x_t - \eta \cdot \nabla_x f(x_t, y_t)$$

$$x_{t+1} = x_t - \eta \cdot \nabla_x f\left(\boldsymbol{x_{t+1/2}}, \boldsymbol{y_{t+1/2}}\right)$$

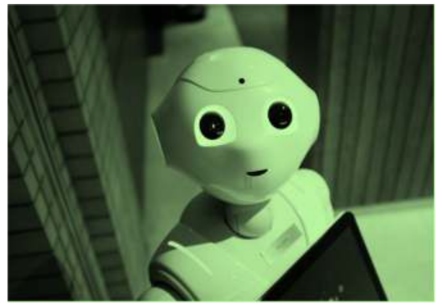
$$\boldsymbol{y_{t+1/2}} = y_t + \eta \cdot \nabla_y f(x_t, y_t)$$

$$y_{t+1} = y_t + \eta \cdot \nabla_y f\left(\boldsymbol{x_{t+1/2}}, \boldsymbol{y_{t+1/2}}\right)$$

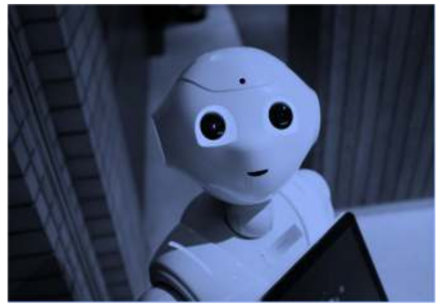
- **[Korpelevich'76, Popov'80, Facchinei-Pang'03]:** Asymptotic ***last-iterate*** convergence results for Optimistic GDA, Extra-Gradient, Mirror-Prox, and related methods when f *is **convex-concave***
- Rates?
  - unconstrained setting: quite clear understanding **[Tseng'95, Daskalakis-Ilyas-Syrgkanis-Zeng ICLR'18, Liang-Stokes AISTATS'19, Gidel et al AISTATS'19, Mokhtari et al '19, Liang-Stokes AISTATS'19, Mokhtari et al '19, Azizian et al AISTATS'20, Golowich-Pattathil- Daskalakis-Ozdaglar COLT'20, Golowich-Pattathil-Daskalakis NeurIPS'20,…]**
  - constrained setting: mostly unclear **[Korpelevich'76;Tseng'95;Daskalakis-Panageas'19;Lee-Luo-Wei-Zhang'20]**
- **interesting question:** Fast, last-iterate convergence rates in constrained case?
  - ➢ match $O\left(\frac{1}{\sqrt{T}}\right)$ rates (w/ mild dimension-dependence) known for average-iterate convergence of no-regret learning methods

Convex *Multi-Player* Games

the further benefits of negative momentum

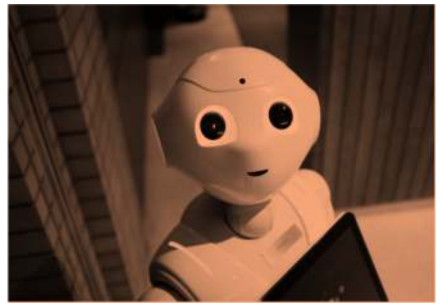


action: x_1
goal: min $f_1(\vec{x})$
f_1: convex in x_1



action: x_2
goal: min $f_2(\vec{x})$
f_2: convex in x_2

…

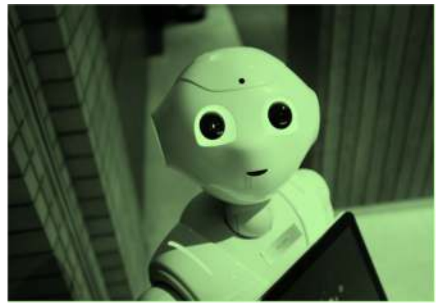


action: x_n
goal: min $f_n(\vec{x})$
f_n: convex in x_n

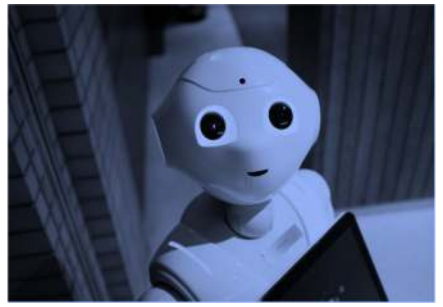
- Nash equilibria are generally intractable **[Daskalakis-Goldberg-Papadimitriou’06, Chen-Deng’06]** but (coarse) correlated equilibria are quite generally tractable **[Papadimitriou-Roughgarden’08, Jiang-LeytonBrown’11]**
- A generic way to converge to (coarse) correlated equilibria is via no-regret learning
  - e.g. Online Gradient Descent, Multiplicative-Weights-Updates, Follow-The-Regularized-Leader
  - No-regret learning is heavily used in Libratus and recent successes in Poker, e.g. **[Brown-Ganzfried-Sandholm’15, Brown-Sandholm’17, Farina-Kroer-Sandholdm’21]**
- Standard no-regret learners have hindsight regret $\boldsymbol{O\left(\sqrt{T}\right)}$ in T rounds $\leftrightarrow$ $\boldsymbol{O\left(1/\sqrt{T}\right)}$ rate of convergence of empirical play to (coarse) Correlated Equilibria
- Better rates?

Convex *Multi-Player* Games

the further benefits of negative momentum

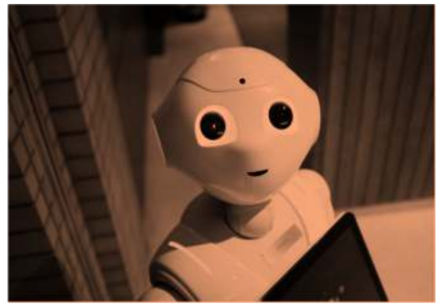


action: x_1
goal: min $f_1(\vec{x})$
f_1: convex in x_1



action: x_2
goal: min $f_2(\vec{x})$
f_2: convex in x_2

…



action: x_n
goal: min $f_n(\vec{x})$
f_n: convex in x_n

- Standard no-regret learners have hindsight regret $O(\sqrt{T})$ in T rounds $\leftrightarrow$ $O(1/\sqrt{T})$ rate of convergence of empirical play to (coarse) Correlated Equilibria
- Better rates?
- Use of *negative momentum* leads to better rates:
  - **[Rakhlin-Sridharan'13, Syrgkanis-Agarwal-Luo-Schapire'15]:** $O(T^{1/4})$ regret in multi-player general-sum games
  - **[Chen-Peng'20]:** $O(T^{1/6})$ regret in 2-player general-sum games
  - **[Daskalakis-Deckelbaum-Kim'11, Hsieh-Antonakopoulos-Mertikopoulos'21]:** $\mathrm{poly}(\log T)$ regret in 2-player zero-sum games
- **[Daskalakis-Fishelson-Golowich'21]:** $\mathrm{poly}(\log T)$ regret in multi-player general-sum games 🔥
  - i.e. optimal $\tilde{O}(1/T)$ convergence of empirical play to *coarse* correlated equilibria!
  - **[Anagnostides-Daskalakis-Fishelson-Golowich-Sandholm'21]:** ditto for no internal-regret learning, no swap-regret learning, thus $\tilde{O}(1/T)$ convergence of empirical play to correlated equilibria!

Menu

- **Motivation**
- **Convex Games**
  - **training oscillations can be removed using negative momentum**
- Nonconvex Games
  - are oscillations inherent/reflective of intractability?
- Conclusions

Menu

- **Motivation**
- **Convex Games**
  - **training oscillations can be removed using negative momentum**
- **Nonconvex Games**
  - **are oscillations inherent/reflective of intractability?**
- Conclusions

Menu

- **Motivation**
- **Convex Games**
  - **training oscillations can be removed using negative momentum**
- **Nonconvex Games**
  - **are oscillations inherent/reflective of intractability?**
    - **an experiment**
- Conclusions

Negative Momentum: in the Wild?

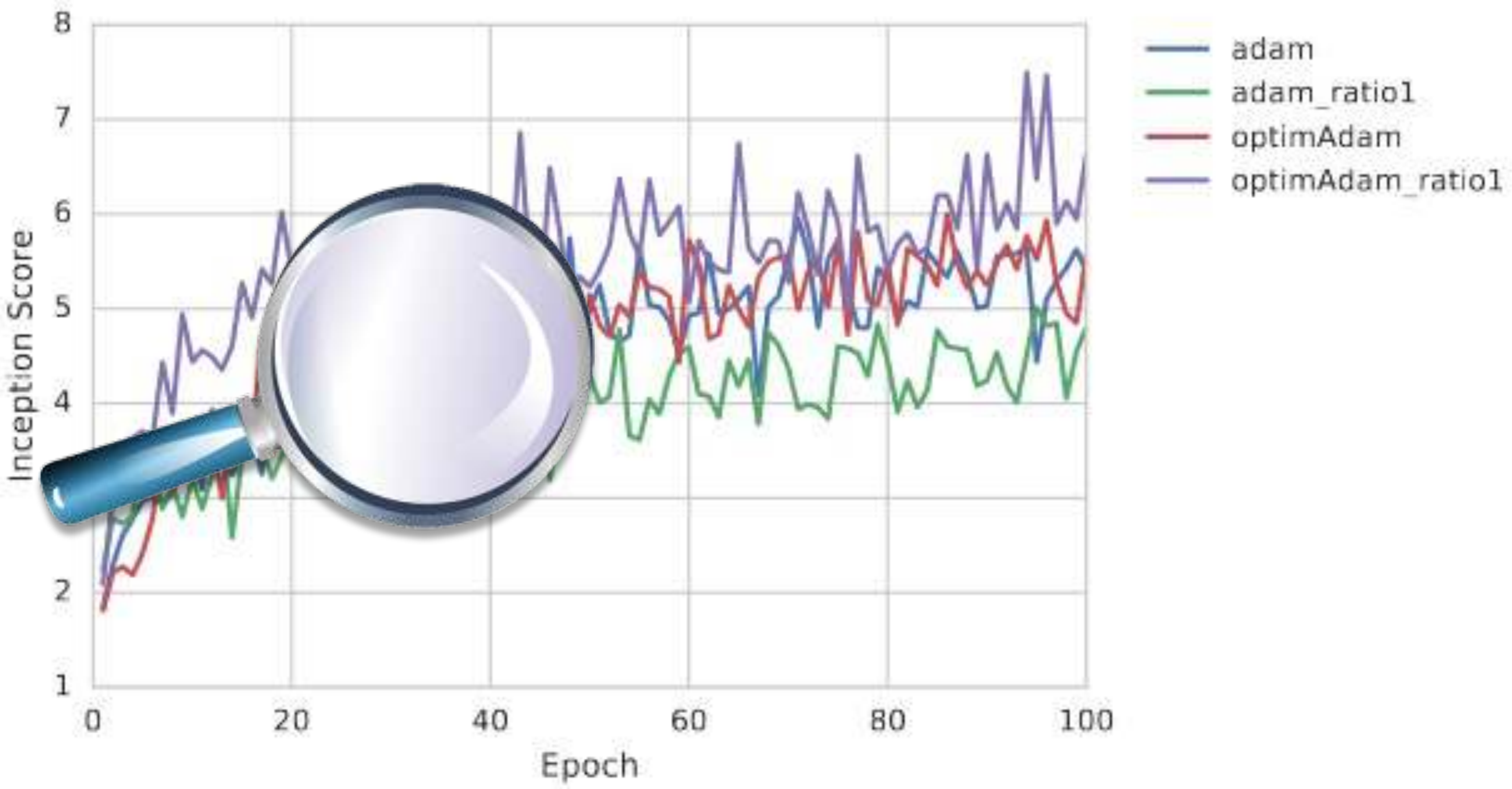
- Is negative momentum helpful, outside of the convex-concave setting?

- **[Daskalakis-Ilyas-Syrgkanis-Zeng ICLR'18]:** Optimistic Adam
  - ***Adam***, a variant of stochastic gradient descent with momentum and per-parameter adaptive learning rates, proposed by **[Kingma-Ba ICLR'15]**, has found wide adoption in deep learning, although it doesn't always converge, even in simple convex settings **[Reddi-Kale-Kumar ICLR'18]**

- In any event, ***Optimistic Adam*** is the right adaptation of Adam to "undo some of the past gradients," i.e. have negative momentum

Optimistic Adam, on CIFAR10

- Compare **Adam** and **Optimistic Adam**, trained on CIFAR10, in terms of Inception Score
- No fine-tuning for **Optimistic Adam**; used same hyper-parameters for both algorithms as suggested in Gulrajani et al. (2017)



Optimistic Adam, on CIFAR10

- Compare **Adam** and **Optimistic Adam**, trained on CIFAR10, in terms of Inception Score
- No fine-tuning for **Optimistic Adam**; used same hyper-parameters for both algorithms as suggested in Gulrajani et al. (2017)

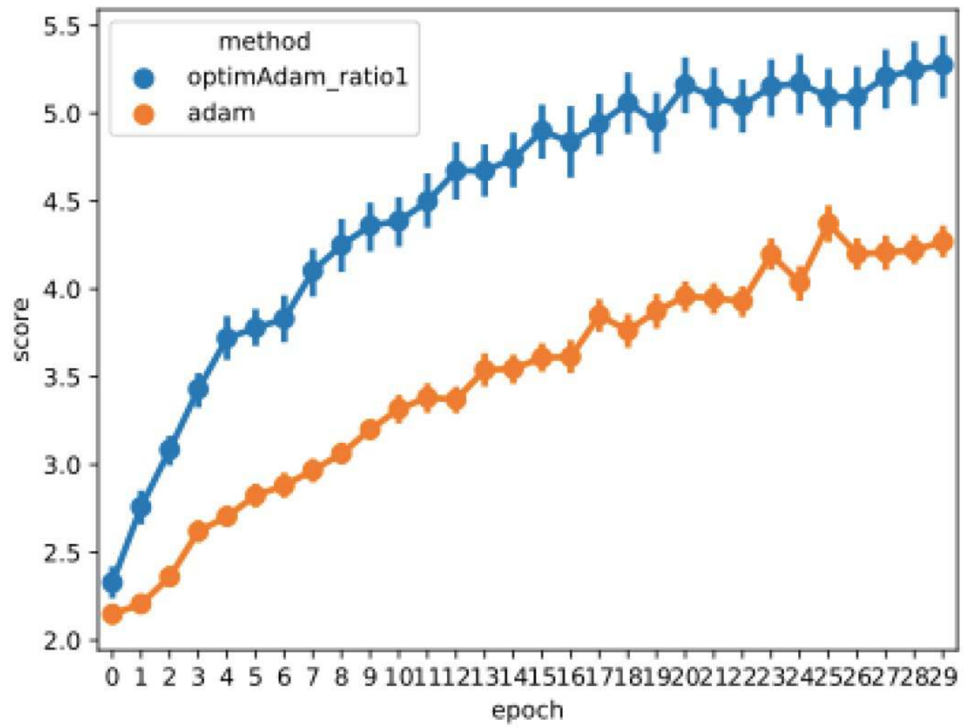


Figure 14: The inception scores across epochs for GANs trained with Optimistic Adam (ratio 1) and Adam (ratio 5) on CIFAR10 (the two top-performing optimizers found in Section 6, with 10%-90% confidence intervals. The GANs were trained for 30 epochs and results gathered across 35 runs.

(b) Sample of images from Generator of Epoch 94, which had the highest inception score.

- Further evidence in favor of negative momentum methods by **[Yadav et al. ICLR'18, Gidel et al. AISTATS'19, Chavdarova et al. NeurIPS'19]**

Decreasing Momentum Trend

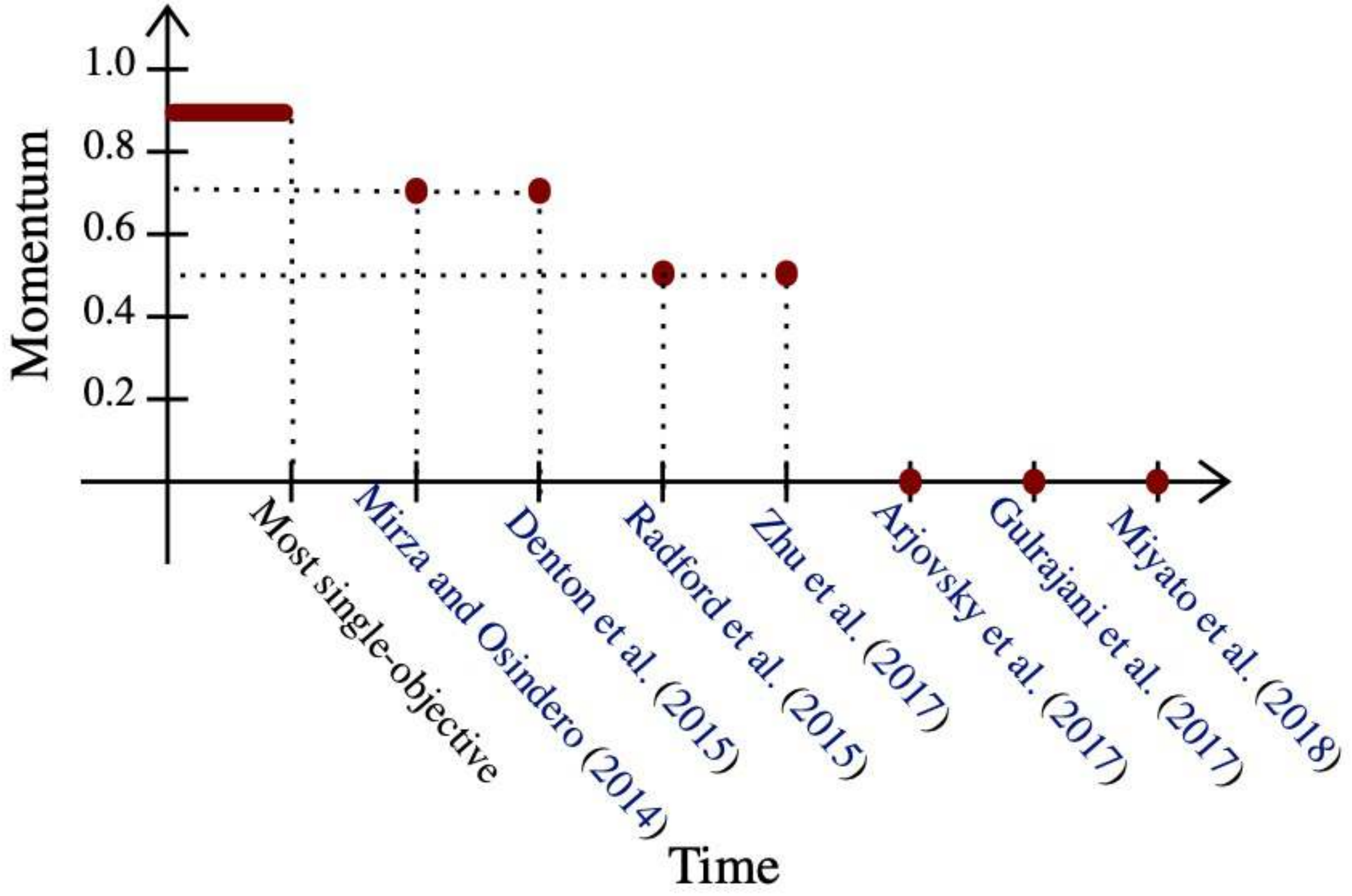


Figure 1: Decreasing trend in the value of momentum used for training GANs across time.

[Gidel et al. AISTATS'19]

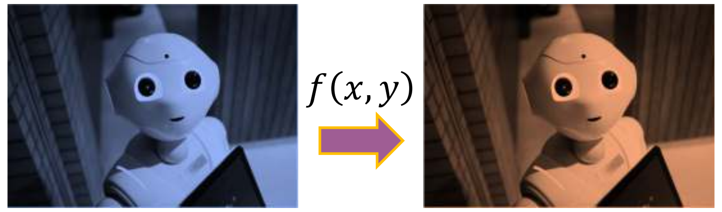
Menu

- **Motivation**
- **Convex Games**
  - **training oscillations can be removed using negative momentum**
- **Nonconvex Games**
  - **are oscillations inherent/reflective of intractability?**
    - **an experiment**
    - theoretical understanding
- Conclusions

Menu

- **Motivation**
- **Convex Games**
  - **training oscillations can be removed using negative momentum**
- **Nonconvex Games**
  - **are oscillations inherent/reflective of intractability?**
    - **an experiment**
    - **theoretical understanding**
- Conclusions

Nonconvex-Nonconcave Objectives



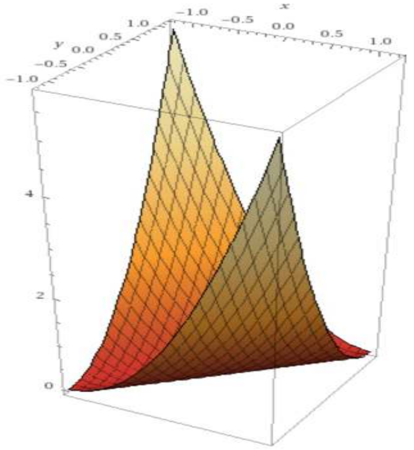
$$\min_x \max_y f(x,y)$$

s.t. $\quad (x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- If $f(x,y)$ is not convex-concave, von Neumann's theorem breaks
- For some f: $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x,y) \neq \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x,y)$

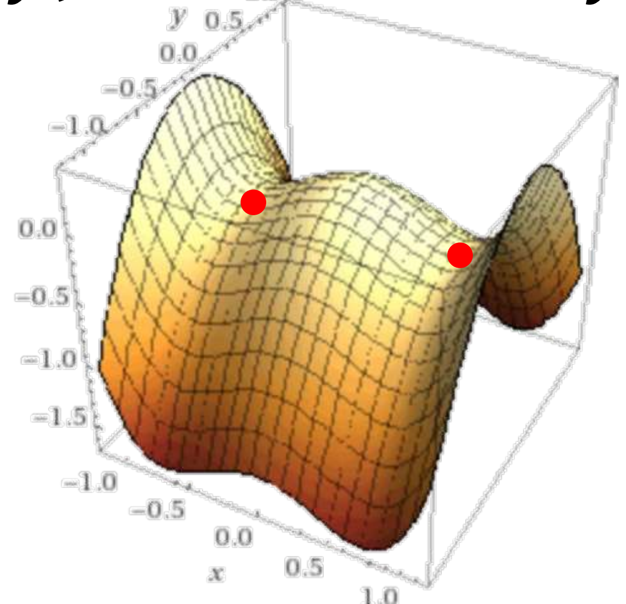
  (both are well-defined when f is continuous and $\mathcal{X}$ and $\mathcal{Y}$ are convex and compact)
- If the game is sequential, the order matters!
- For other f: equality holds but there are multiple, disconnected solutions

$f(x,y) = (x-y)^2$



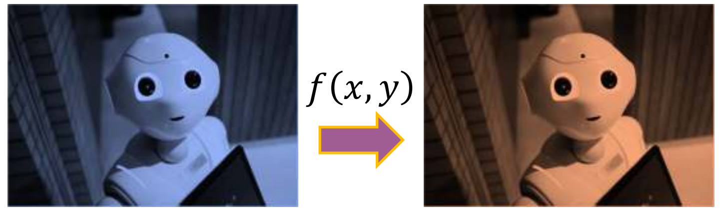
$$\min_{x \in [-1,1]} \max_{y \in [-1,1]} f(x,y) \neq \max_{y \in [-1,1]} \min_{x \in [-1,1]} f(x,y)$$

$f(x,y) = x^4 - x^2 - y^2$



Minimization vs Min-Max Optimization

non-convex setting



$$\min_x f(x)$$
s.t. $x \in S \subset \mathbb{R}^d$

$$\min_x \max_y f(x,y)$$
s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

- f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- constraint set S: convex, compact

$B_\delta(x^*)$={x s.t. $\|x - x^*\| \leq \delta$}

$B_\delta(y^*)$={y s.t. $\|y - y^*\| \leq \delta$}

Def: (ε, δ)-*local* minimum

$$f(x^*) \leq f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Def: (ε, δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*, y) - \varepsilon \leq f(x^*, y^*) \leq f(x, y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*, y) \in S$

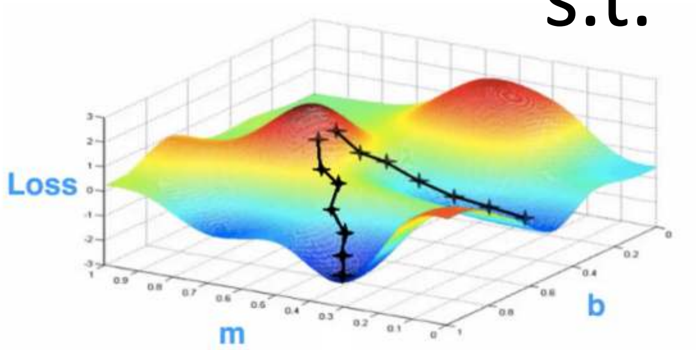
$\forall x \in B_\delta(x^*)$ s.t. $(x, y^*) \in S$

Minimization vs Min-Max Optimization

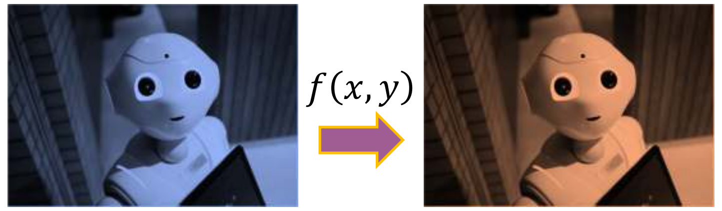
non-convex setting

$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$



- ➢ f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- ➢ constraint set S: convex, compact



$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$

$B_\delta(x^*)$=$\{x$ s.t. $\|x - x^*\| \le \delta\}$

$B_\delta(y^*)$=$\{y$ s.t. $\|y - y^*\| \le \delta\}$

Def: (ε, δ)-*local* minimum

$$f(x^*) \le f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Theorem [folklore]

If $\delta \le \sqrt{2\varepsilon/L}$, first-order methods find (ε, δ)-*local* minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, smoothness of f.

(for larger δ existence holds, but problem becomes **NP**-hard)

Def: (ε, δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*, y) - \varepsilon \le f(x^*, y^*) \le f(x, y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*, y) \in S$ $\quad\quad$ $\forall x \in B_\delta(x^*)$ s.t. $(x, y^*) \in S$

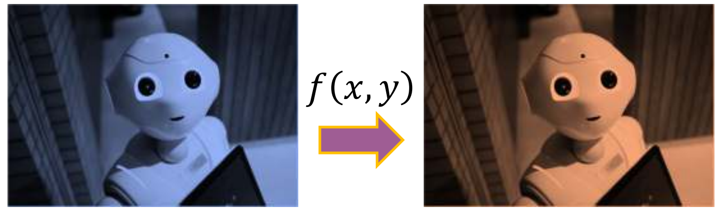
exist for small enough $\delta \le \sqrt{2\varepsilon/L}$

complexity ????

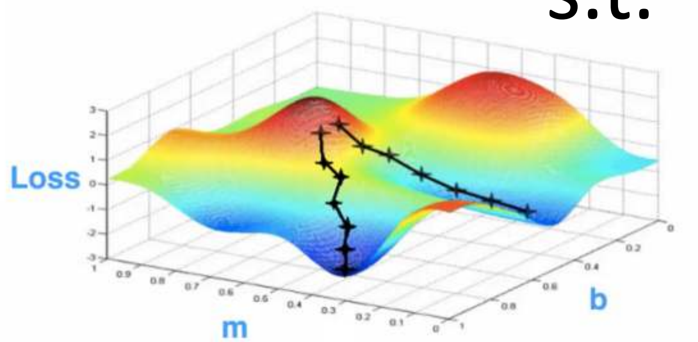
Minimization vs Min-Max Optimization

non-convex setting

$$\min_x f(x)$$
$$\text{s.t.} \quad x \in S \subset \mathbb{R}^d$$



$$\min_x \max_y f(x,y)$$
$$\text{s.t.} \quad (x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$$



- ➢ f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- ➢ constraint set S: convex, compact

$B_\delta(x^*)=\{x \text{ s.t. } \|x - x^*\| \le \delta\}$

$B_\delta(y^*)=\{y \text{ s.t. } \|y - y^*\| \le \delta\}$

Def: (ε,δ)-*local* minimum

$$f(x^*) \le f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Theorem [folklore]

If $\delta \le \sqrt{2\varepsilon/L}$, first-order methods find (ε,δ)-*local* minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, smoothness of f.

(for larger δ existence holds, but problem becomes **NP**-hard)

Def: (ε,δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*,y) - \varepsilon \le f(x^*,y^*) \le f(x,y^*) + \varepsilon$$

$\forall y \in B_\delta(y^*)$ s.t. $(x^*,y) \in S$ $\qquad \forall x \in B_\delta(x^*)$ s.t. $(x,y^*) \in S$

Theorem [Daskalakis-Skoulakis-Zampetakis STOC'21] 🔥

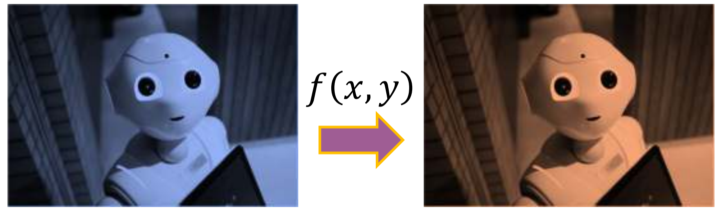
First-order methods need a number of queries to f or ∇f that is ***exponential*** in at least one of $\frac{1}{\varepsilon}, L,$ or dimension to find (ε,δ)-local min-max equilibria, even when $\delta \le \sqrt{2\varepsilon/L}$ (the regime in which they are guaranteed to exist).

Minimization vs Min-Max Optimization

non-convex setting

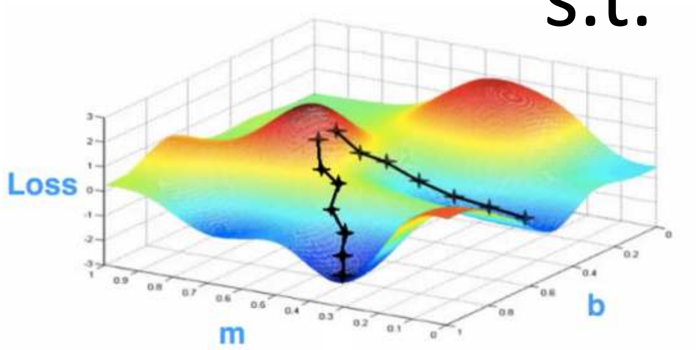
$$\min_x f(x)$$

s.t. $x \in S \subset \mathbb{R}^d$



$$\min_x \max_y f(x,y)$$

s.t. $(x,y) \in S \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$



- f: Lipschitz, L-smooth, $f(x) \in [0,1]$
- constraint set S: convex, compact

$B_\delta(x^*)$={x s.t. $\|x - x^*\| \le \delta$}

$B_\delta(y^*)$={y s.t. $\|y - y^*\| \le \delta$}

Def: (ε,δ)-*local* minimum

$$f(x^*) \le f(x) + \varepsilon, \forall\, x \in B_\delta(x^*) \cap S$$

Theorem [folklore]

If $\delta \le \sqrt{2\varepsilon/L}$, first-order methods find (ε,δ)-*local* minima, in #steps/queries to f or ∇f that are polynomial in $1/\varepsilon$, smoothness of f.

(for larger δ existence holds, but problem becomes **NP**-hard)

Def: (ε,δ)-*local* min-max equilibrium [Daskalakis-Panageas'18, Mazumdar-Ratliff'18]

$$f(x^*,y) - \varepsilon \le f(x^*,y^*) \le f(x,y^*) + \varepsilon$$

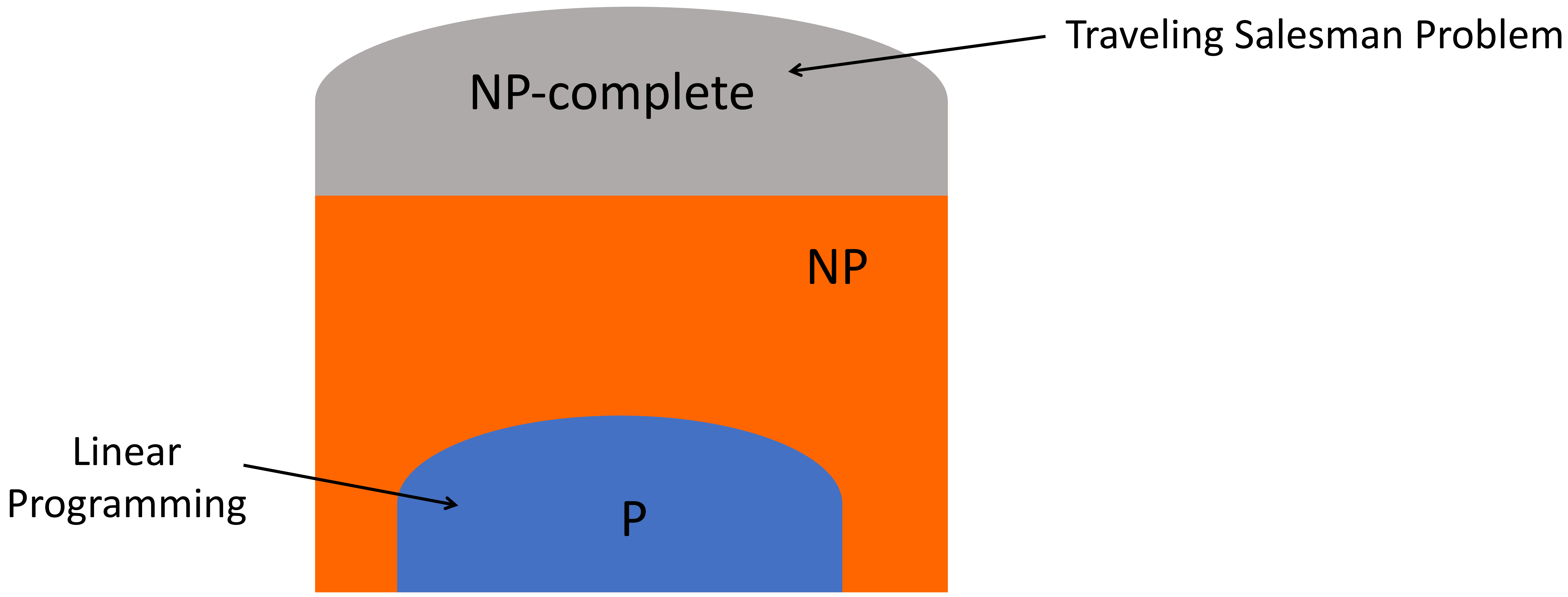
$\forall y \in B_\delta(y^*)$ s.t. $(x^*,y) \in S$ $\forall x \in B_\delta(x^*)$ s.t. $(x,y^*) \in S$

Theorem [w/ Skoulakis-Zampetakis STOC'21]

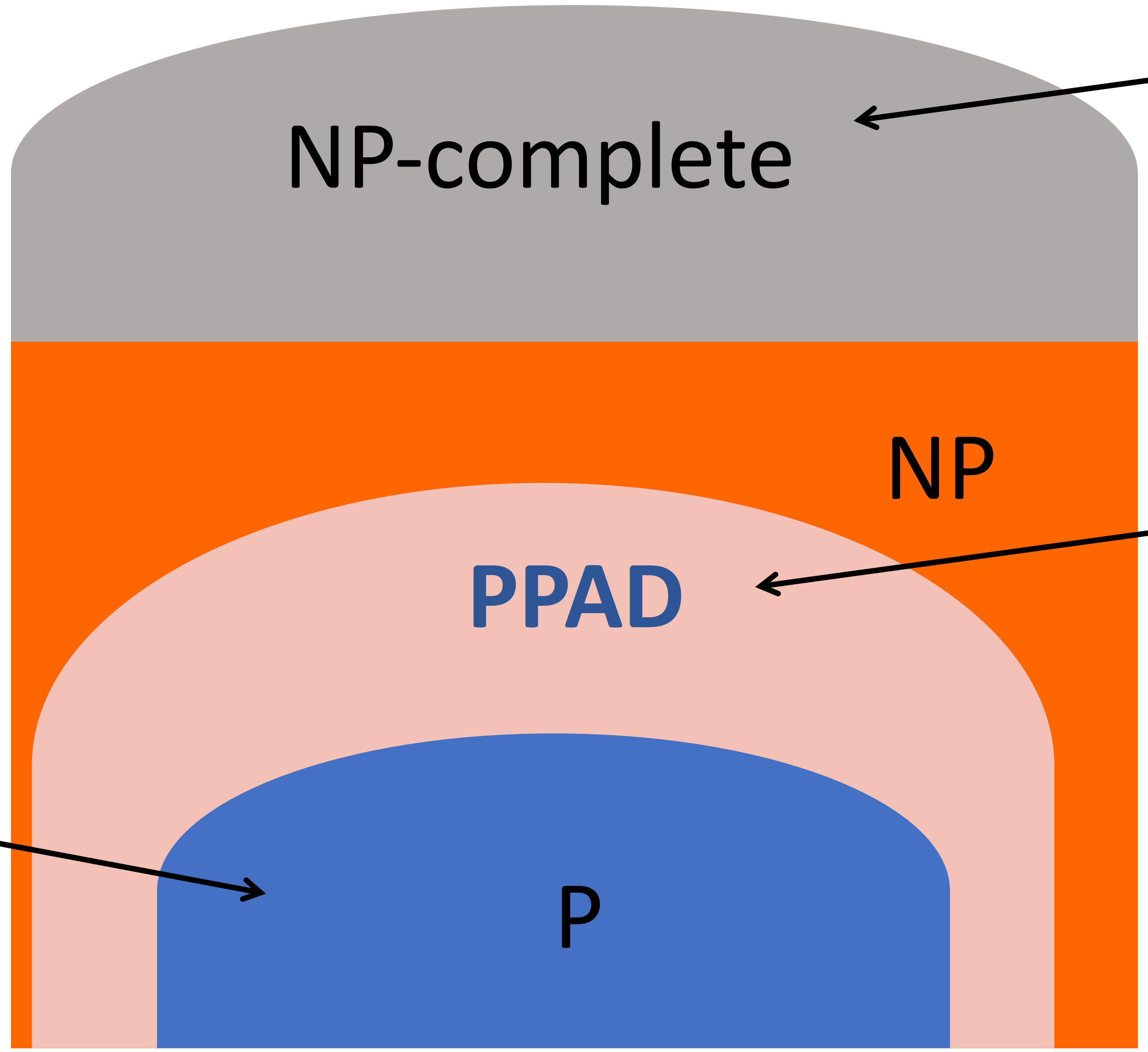
Computing (ε,δ)-local min-max equilibria, for $\delta \le \sqrt{2\varepsilon/L}$, is **PPAD**-complete.

Corollary: *Any* algorithm (first-order, second-order, whatever) takes ***super-polynomial*** time, unless **P**=**PPAD**.

The Complexity of Local Min-Max Equilibrium



The Complexity of Local Min-Max Equilibrium

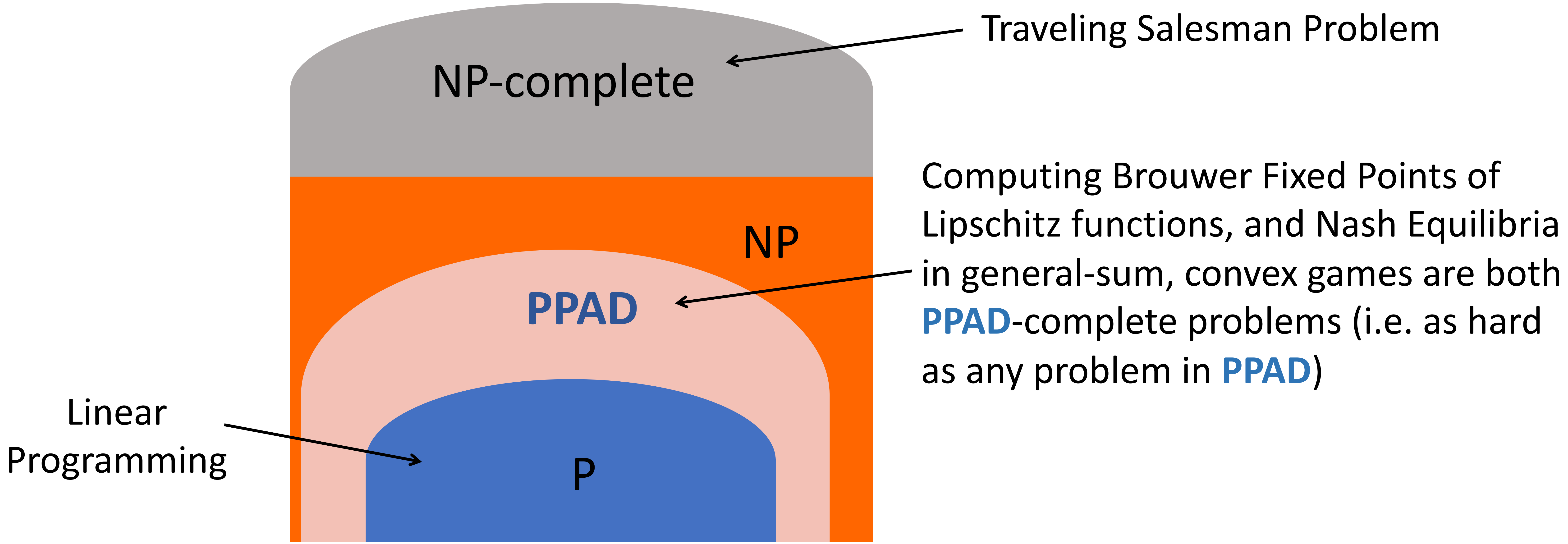


Traveling Salesman Problem

Computing Brouwer Fixed Points of Lipschitz functions, and Nash Equilibria in general-sum, convex games are both **PPAD**-complete problems (i.e. as hard as any problem in **PPAD**)

Linear Programming

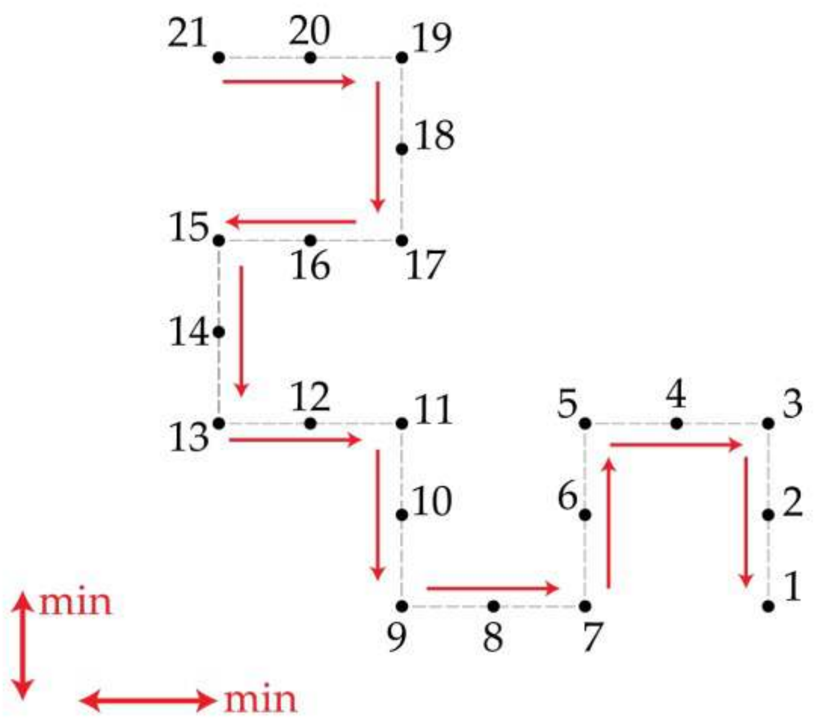
The Complexity of Local Min-Max Equilibrium

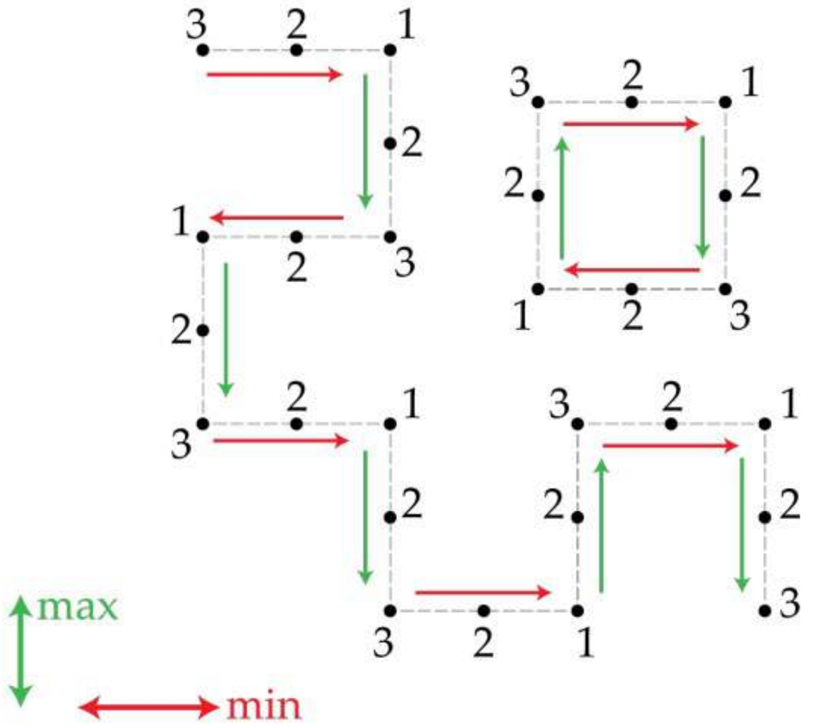


[Daskalakis-Skoulakis-Zampetakis STOC'21]: Computing local min-max equilibria in nonconvex-nonconcave zero-sum games is exactly as hard as (i) computing Brouwer fixed points of Lispchitz functions, (ii) computing Nash equilibrium in general-sum convex games, (iii) at least as hard as any other problem in **PPAD**.

Min-Min vs Min-Max – what's the difference?

Consider a long path of better-response dynamics in a min-min (i.e. fully cooperative) game and a min-max (i.e. fully competitive) game





function value decreases along better-response path, thus: (i) moving along better-response path makes progress towards (local) minimum

(ii) function value along ε-step better-response

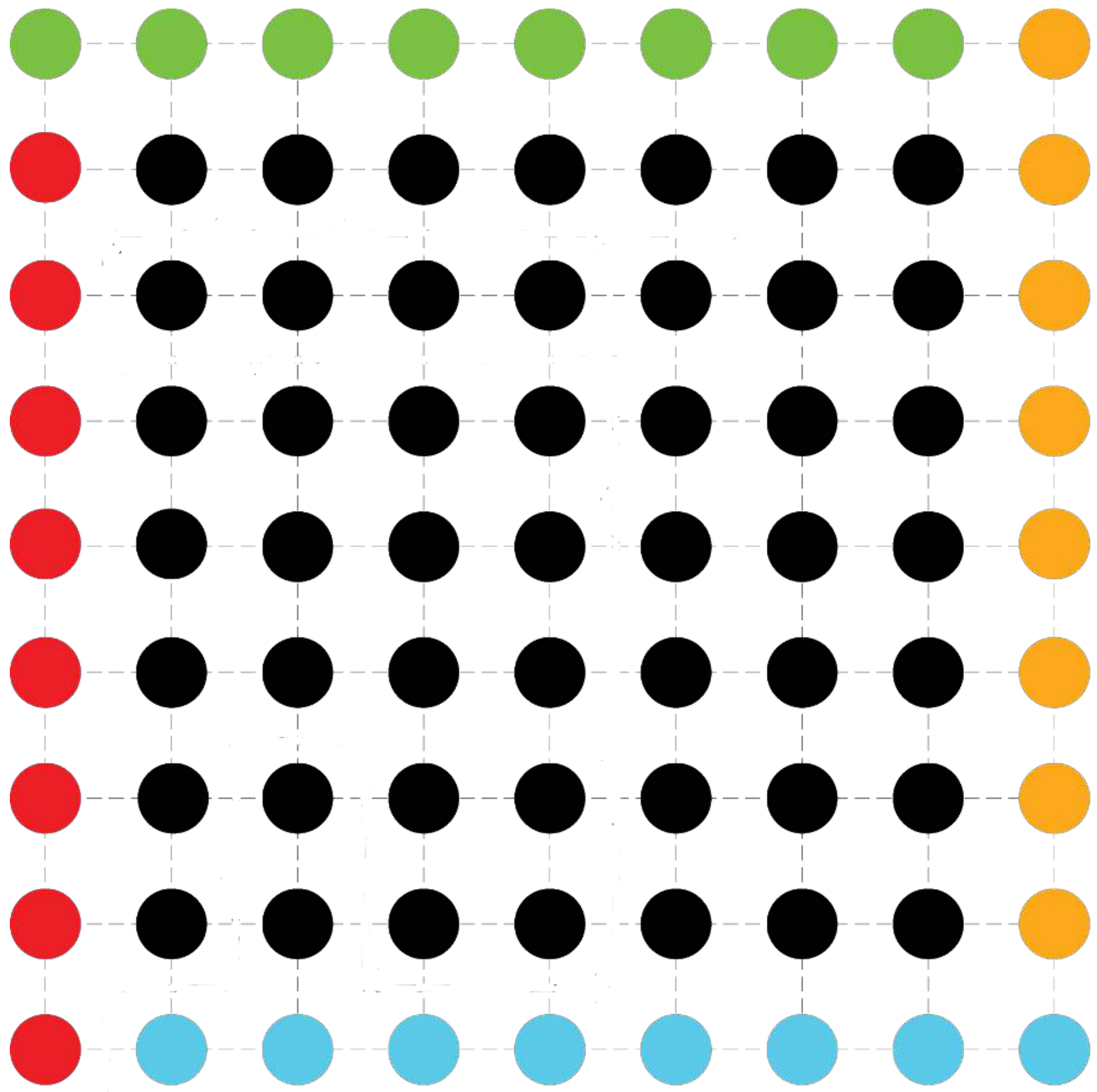
better-response paths may be cyclic :S

querying function value along non-cyclic ε-step better-response path does not reveal information about how far the end of the path is!

to implement this, we appeal to the complexity-theoretic machinery of PPAD and its tight relationship to Brouwer fixed point computation

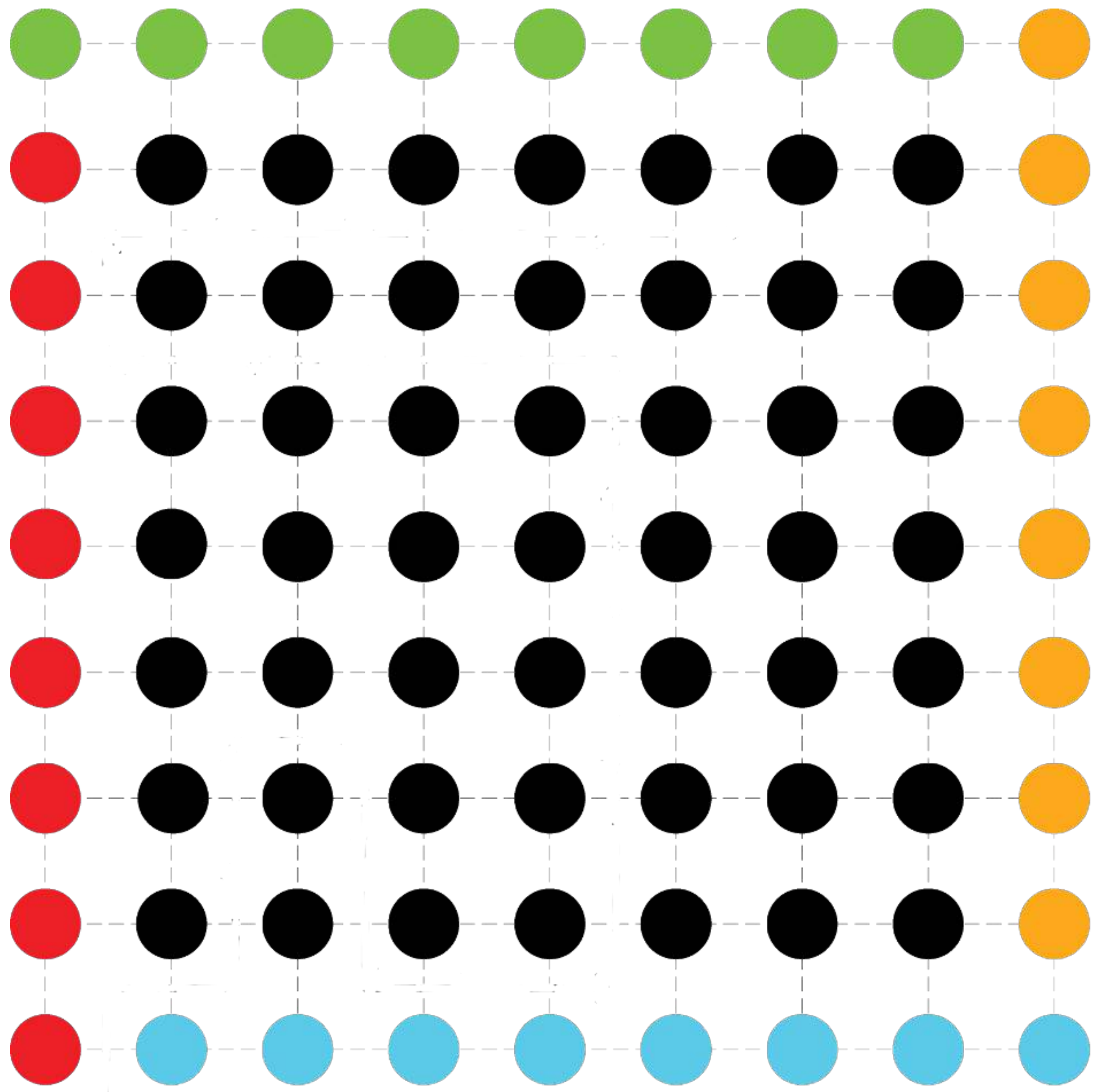
to turn this intuition into an intractability proof, hide exponentially long best-response path within ambient space s.t. no easy to find local min-max equilibria in ambient space

The Topological Nature of Local Min-Max



(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

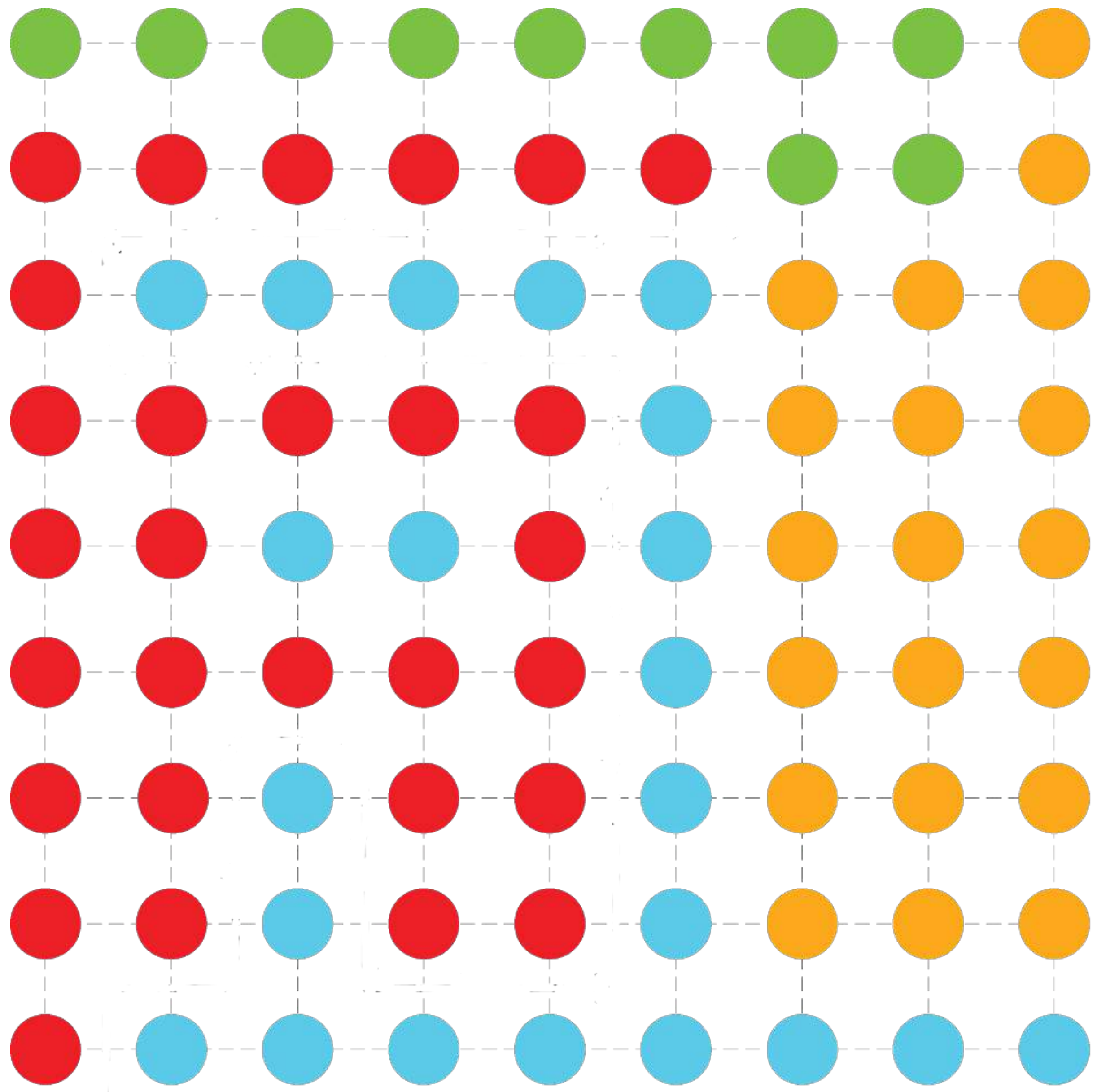
The Topological Nature of Local Min-Max



Note that **red** and **yellow** is an interesting pair, as is **blue** and **green** (all other pairs appear somewhere on the boundary)

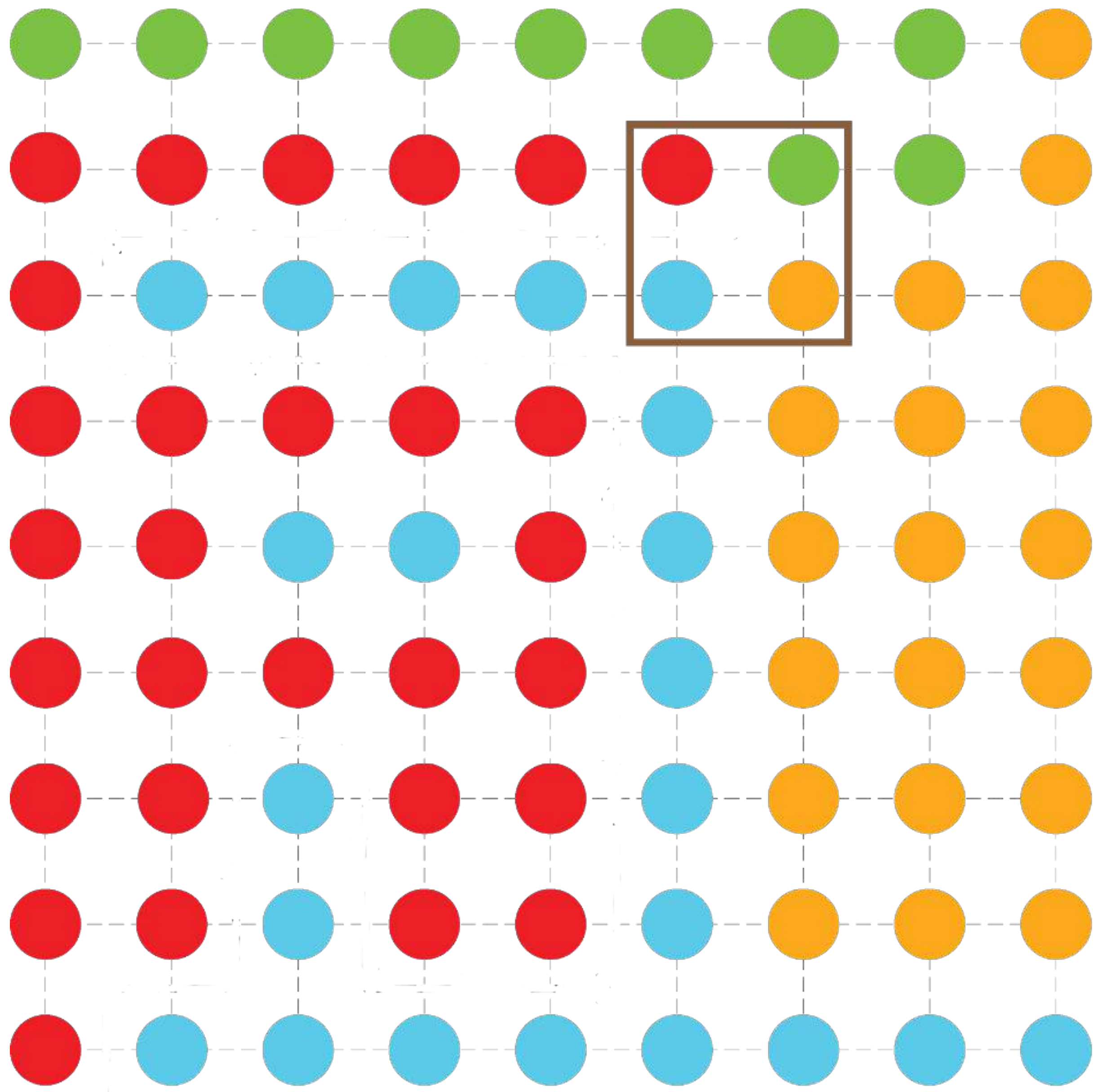
(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

The Topological Nature of Local Min-Max



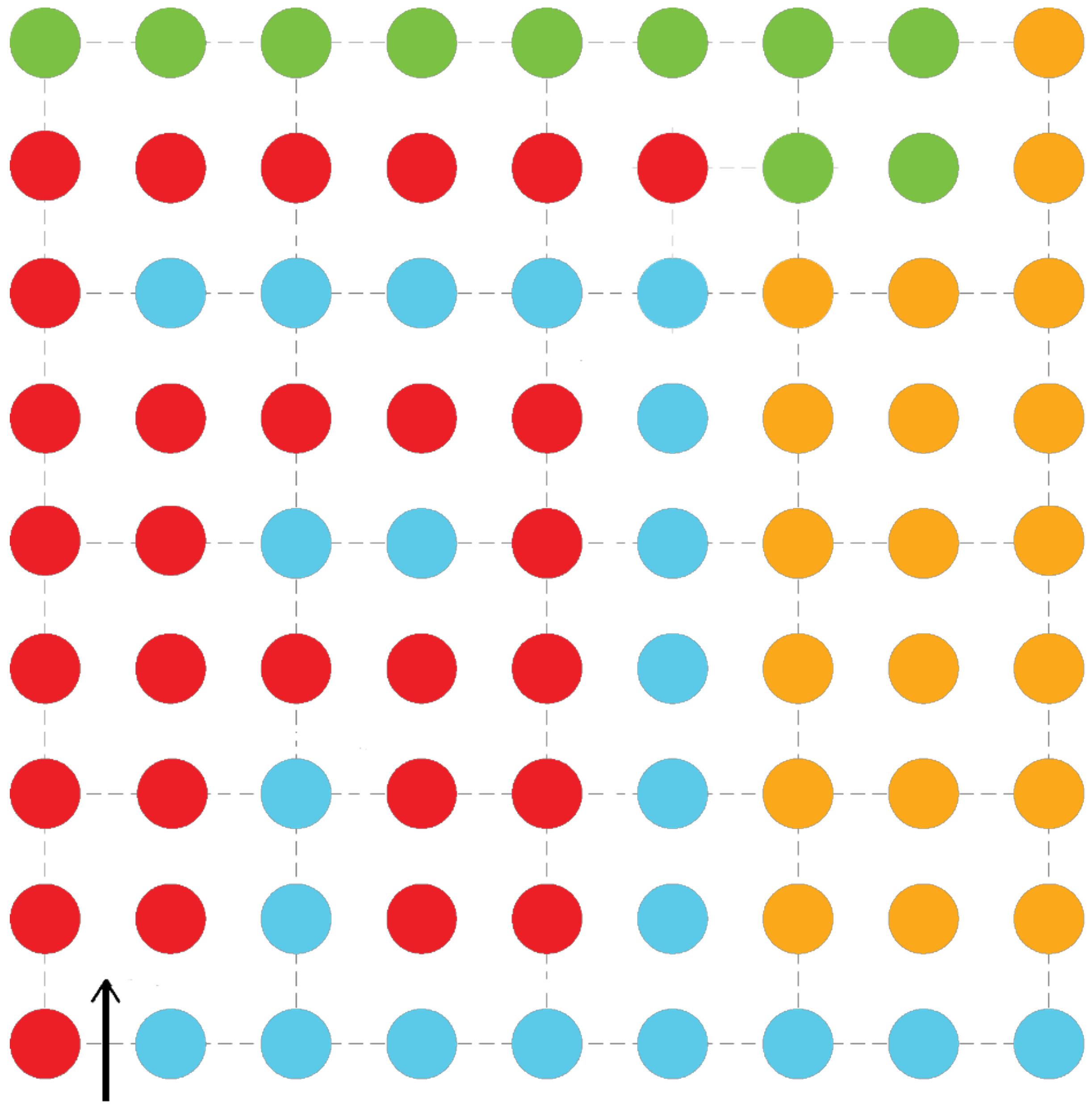
(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

The Topological Nature of Local Min-Max



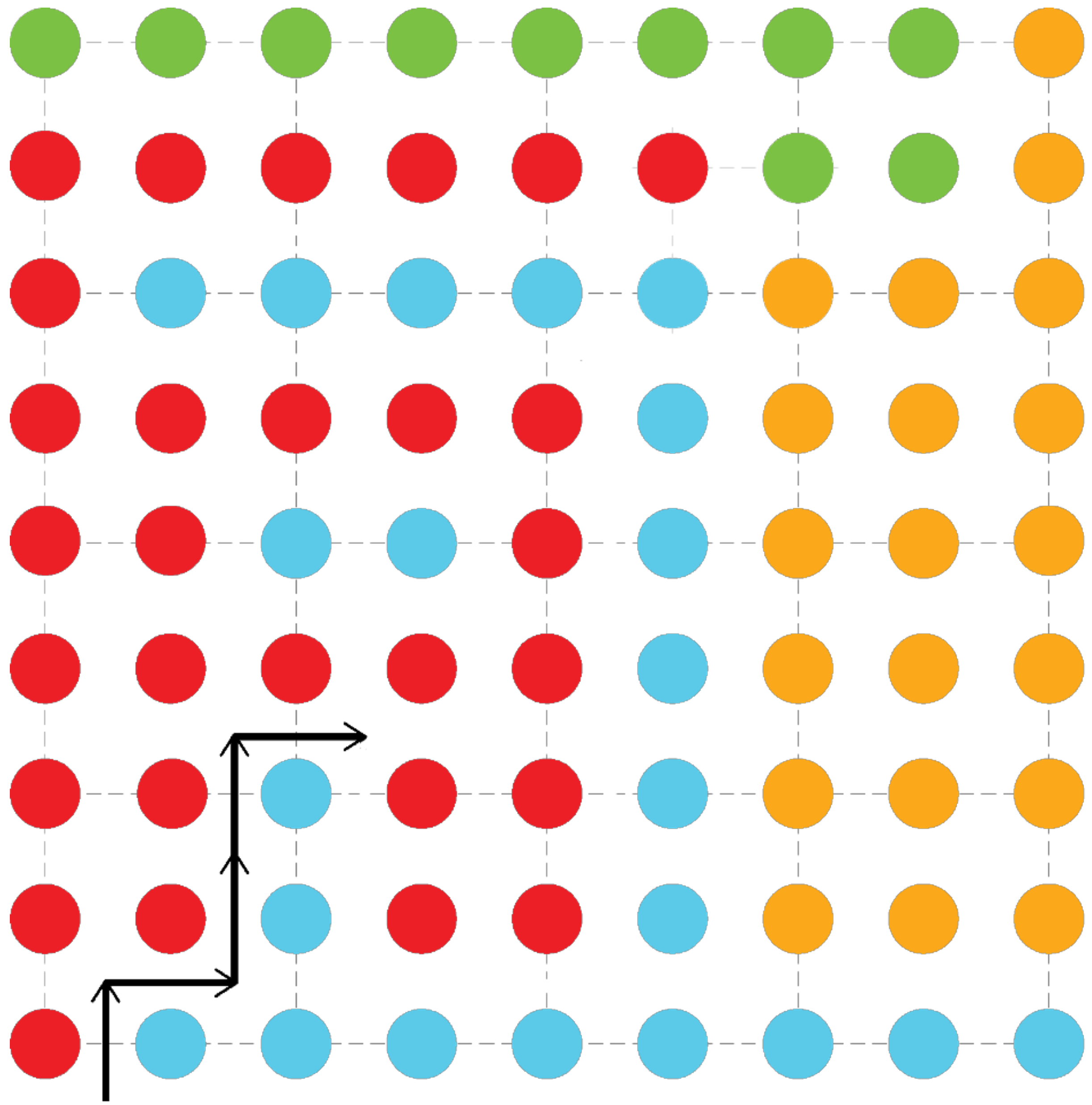
(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

The Topological Nature of Local Min-Max



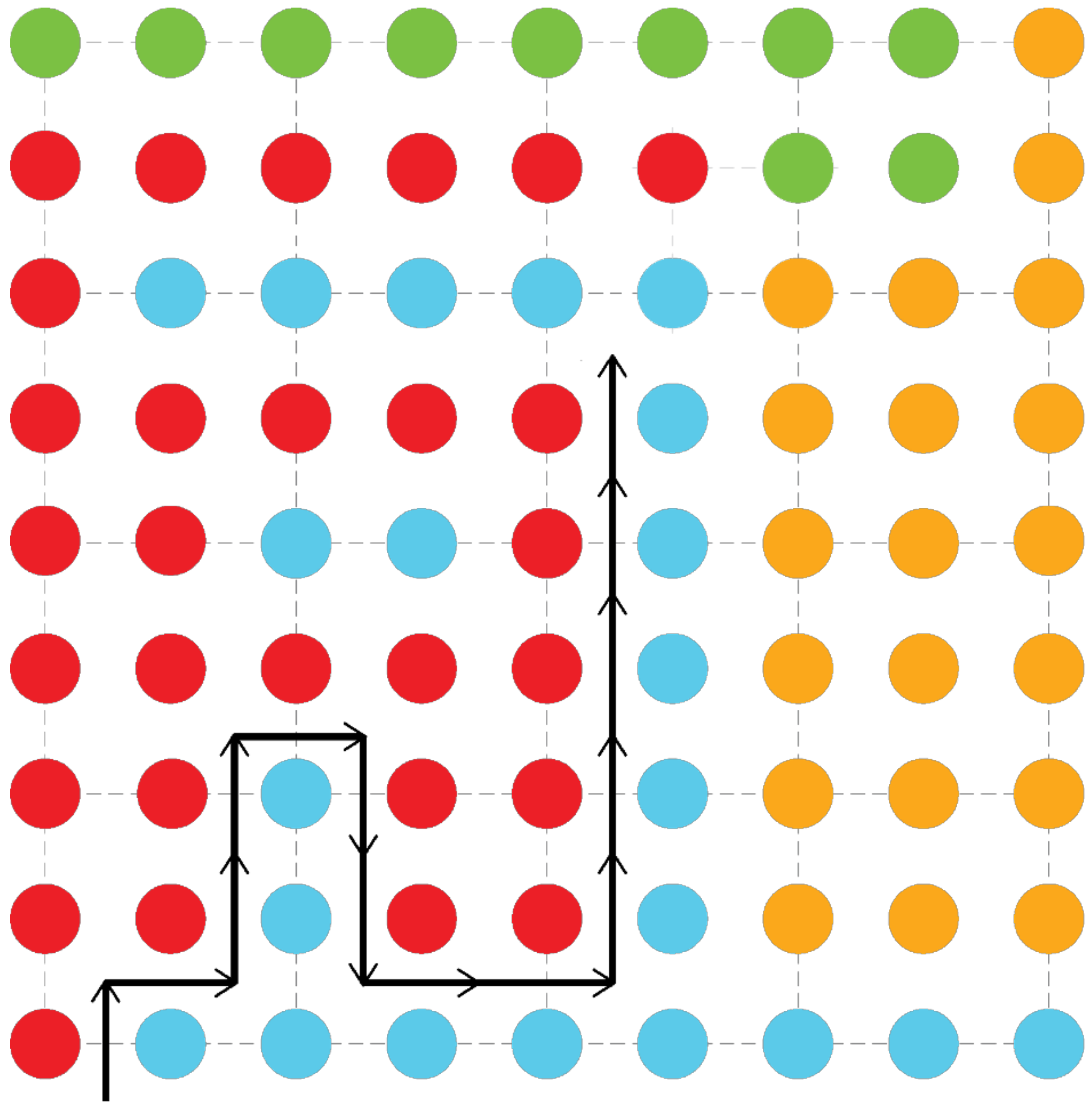
(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

The Topological Nature of Local Min-Max



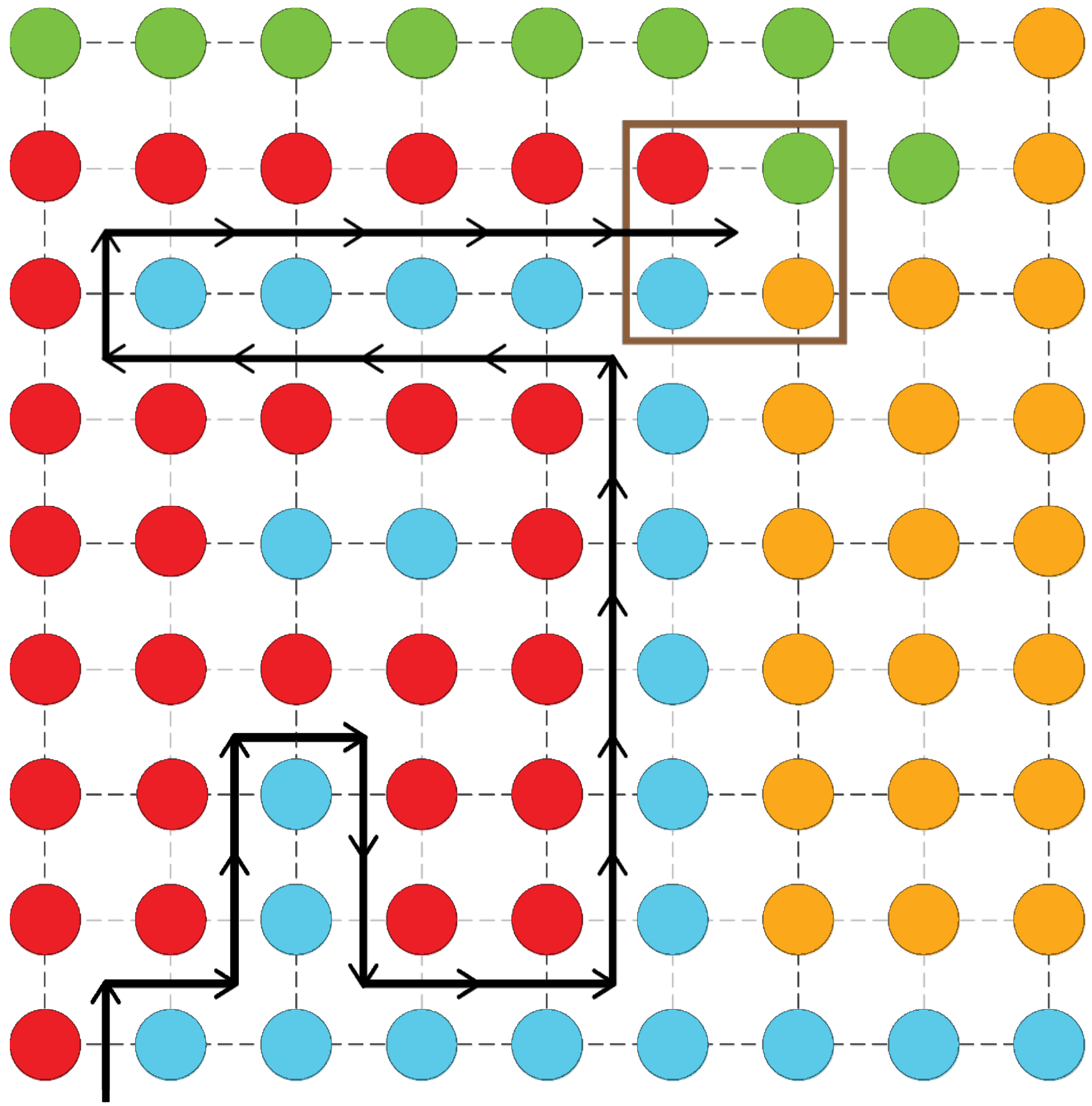
(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

The Topological Nature of Local Min-Max



(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

The Topological Nature of Local Min-Max



(variant of) Sperner's Lemma: No matter how the internal vertices are colored, there must exist a square containing both **red** and **yellow** or both **blue** and **green**.

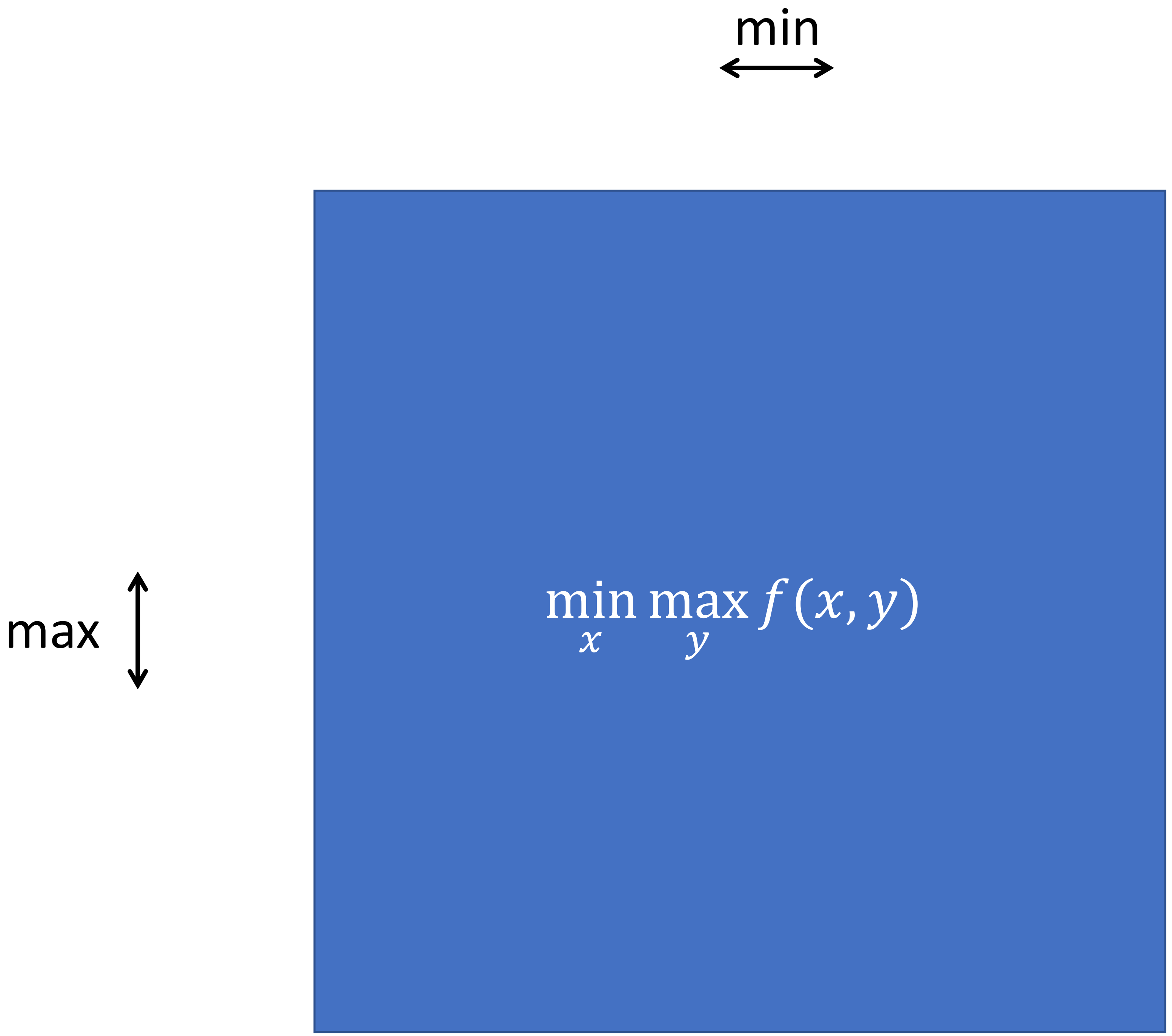
The Topological Nature of Local Min-Max

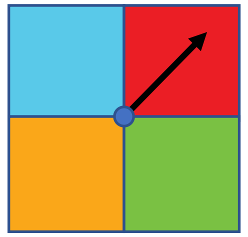
min

max

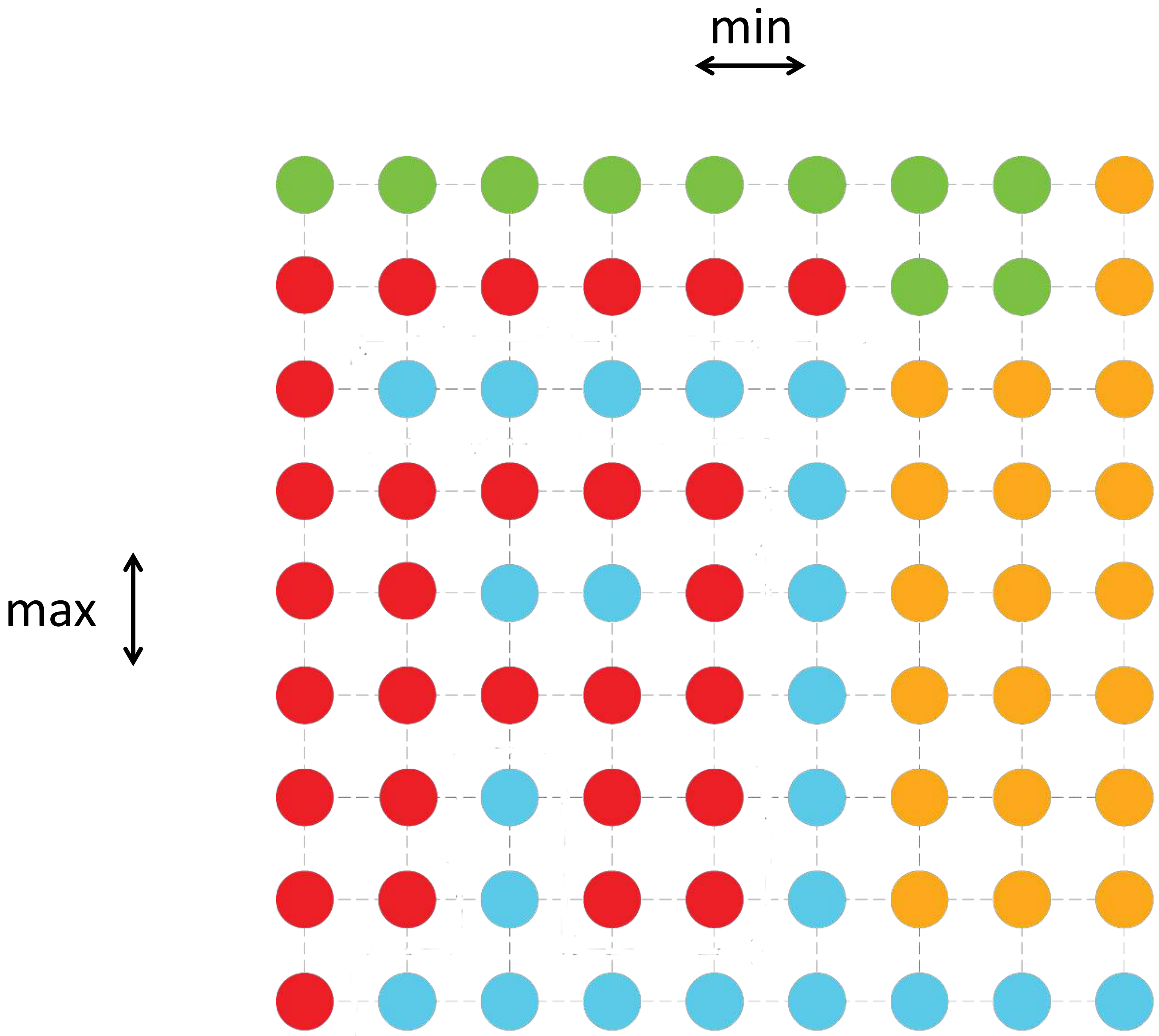
$$\min_x \max_y f(x, y)$$

The Topological Nature of Local Min-Max

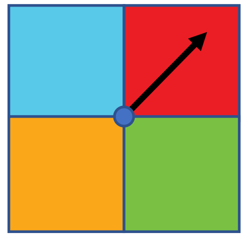


Local Min-Max to Sperner: color grid according to the direction of $V(x,y) = \underbrace{(-\nabla_x f(x,y), \nabla_y f(x,y))}_{\text{local best-response direction}}$ using 

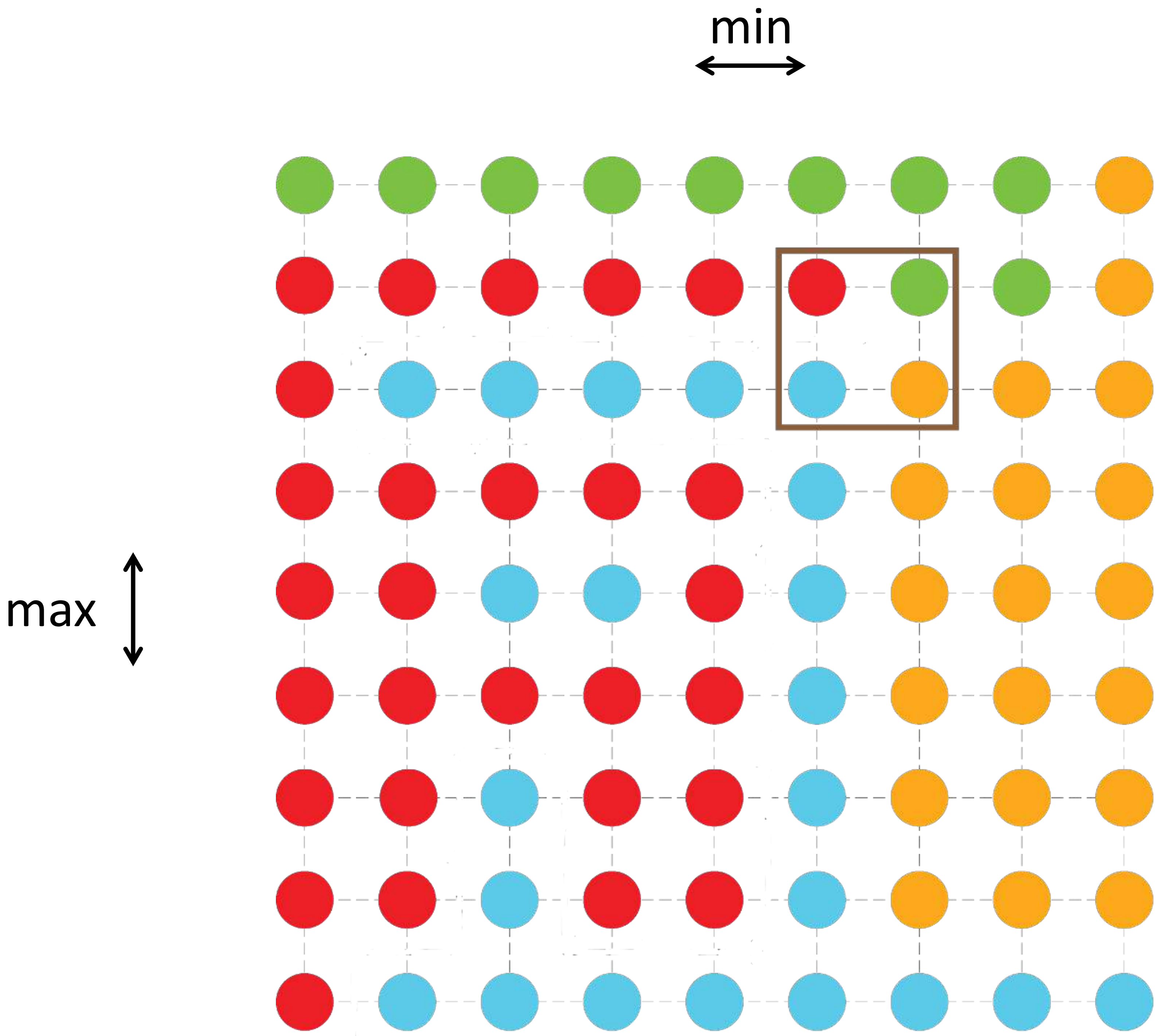
The Topological Nature of Local Min-Max



Local Min-Max to Sperner: color grid according to the direction of $V(x,y) = \underbrace{(-\nabla_x f(x,y), \nabla_y f(x,y))}_{\text{local best-response direction}}$ using

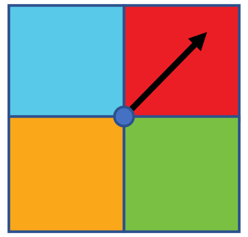


The Topological Nature of Local Min-Max

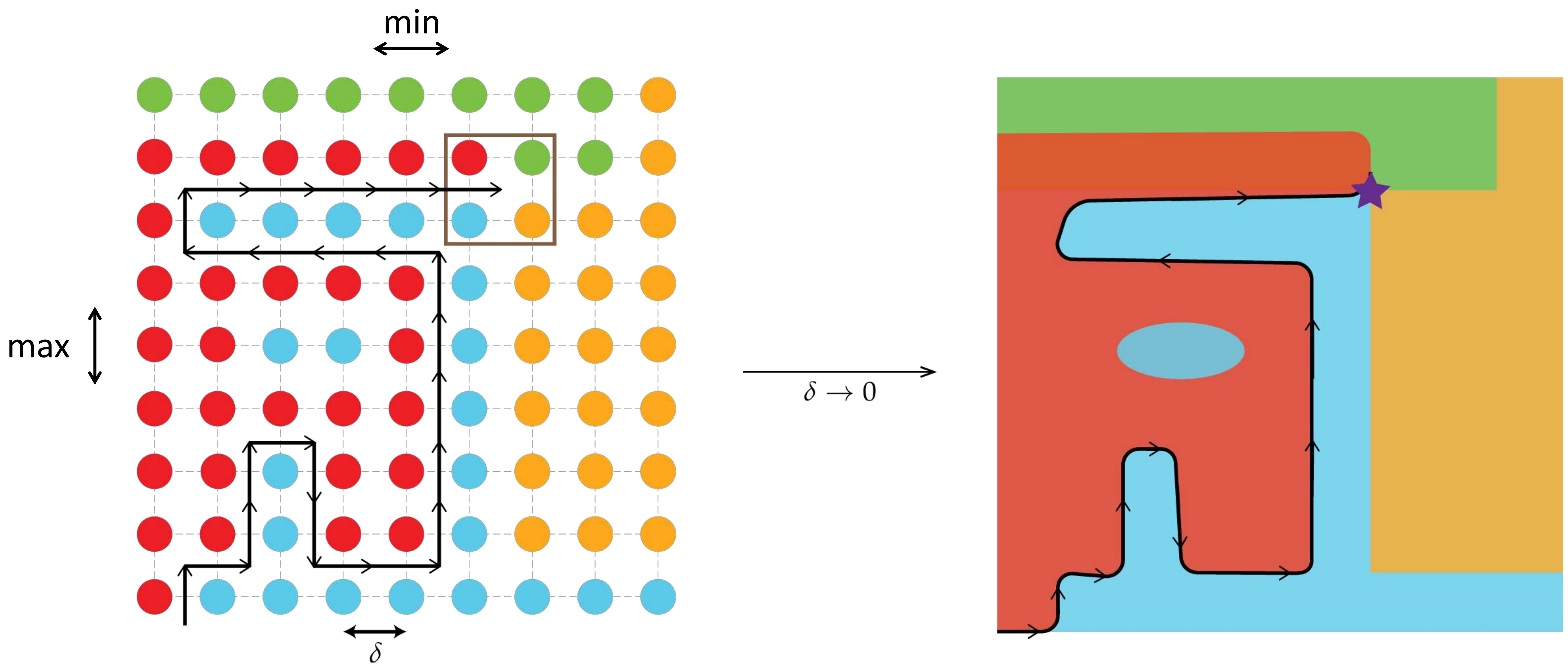


When **red** meets **yellow** or **blue** meets **green** that's a local min-max!

meeting guaranteed by Sperner!

Local Min-Max to Sperner: color grid according to the direction of $V(x,y) = \underbrace{(-\nabla_x f(x,y), \nabla_y f(x,y))}_{\text{local best-response direction}}$ using 

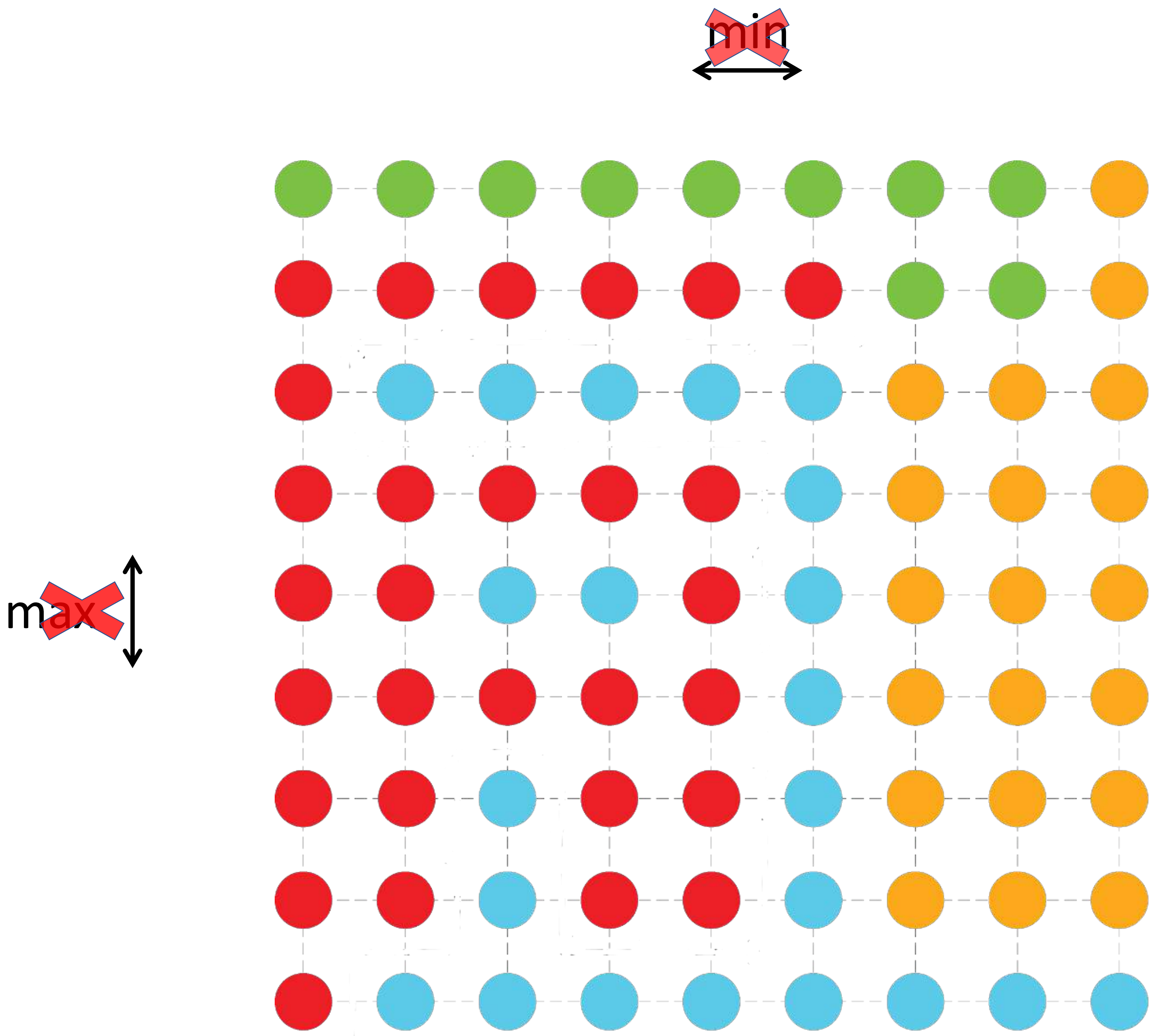
The Topological Nature of Local Min-Max



Local Min-Max to Sperner: taking limits, gives rise to second-order method with *guaranteed asymptotic convergence* to local min-max equilibria **[Daskalakis-Golowich-Skoulakis-Zampetakis'2?]**

- ➢ related to follow-the-ridge method of **[Wang-Zhang-Ba ICLR'19]** which exhibits only local convergence

The Topological Nature of Local Min-Max



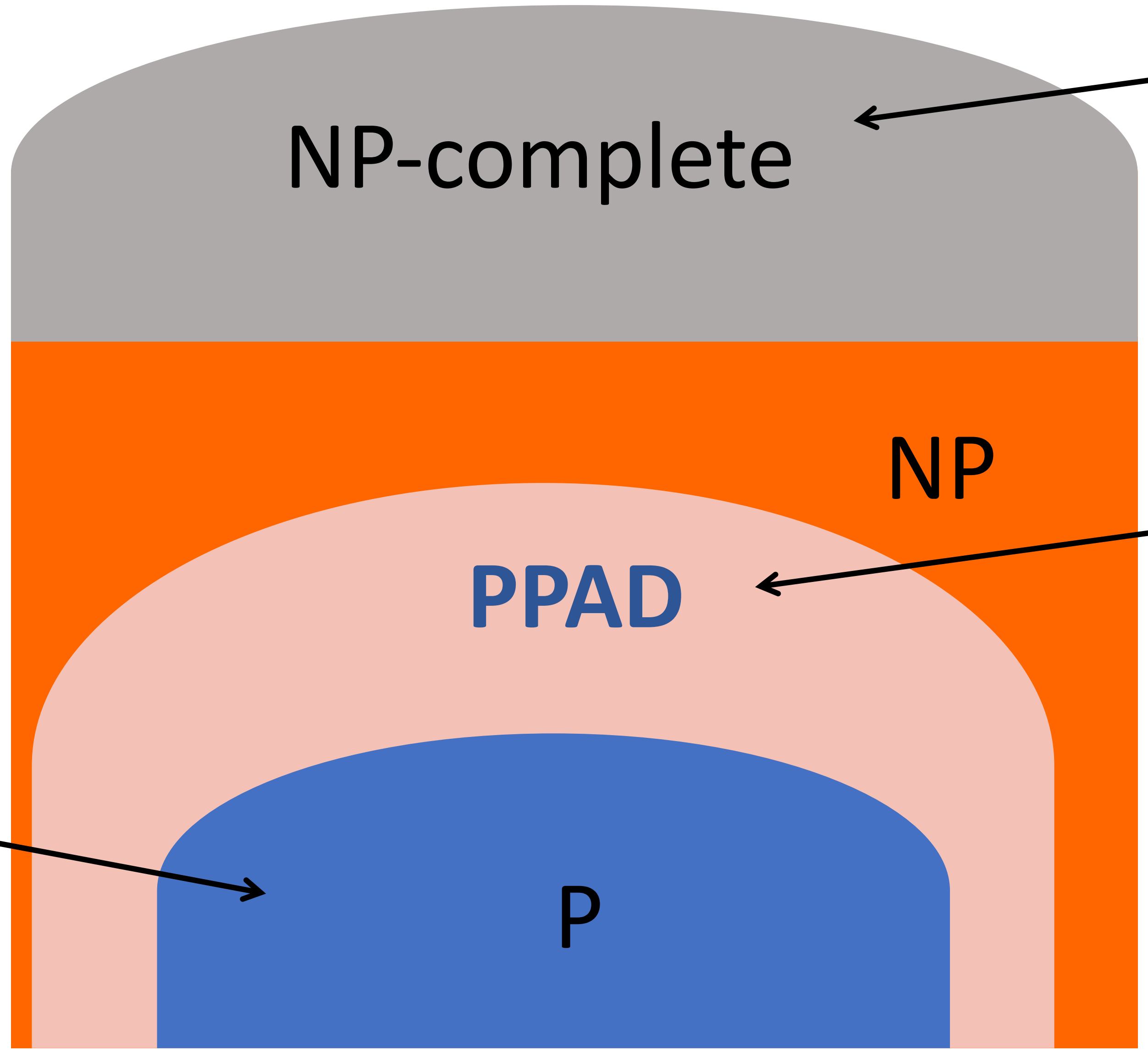
Roughly max chooses “squares” and min chooses “doors” and is penalized/rewarded according to the colors/orientation of the door inside the square

Complication: pass to continuum…

Sperner to Local Min-Max: go in the reverse

- given colors of any Sperner instance, construct $f(x, y)$ such that local min-max eq $\leftrightarrow$ well-colored squares
- implies local min-max is PPAD-complete because Sperner is.

The Complexity of Local Min-Max Equilibrium



Traveling Salesman Problem

Computing Brouwer Fixed Points of Lipschitz functions, and Nash Equilibria in general-sum, convex games are both **PPAD**-complete problems (i.e. as hard as any problem in **PPAD**)

Linear Programming

[Daskalakis-Skoulakis-Zampetakis STOC'21]: Computing local min-max equilibria in nonconvex-nonconcave zero-sum games is exactly as hard as (i) computing Brouwer fixed points of Lispchitz functions, (ii) computing Nash equilibrium in general-sum convex games, (iii) at least as hard as any other problem in **PPAD**.

Menu

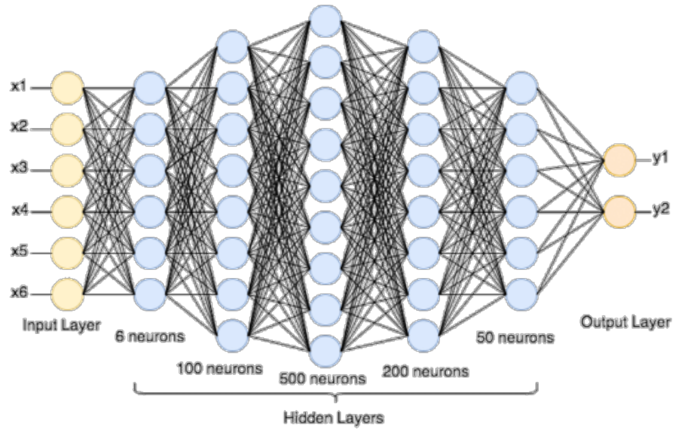
- **Motivation**
- **Convex Games**
  - **training oscillations can be removed using negative momentum**
- **Nonconvex Games**
  - **are oscillations inherent/reflective of intractability?**
    - **an experiment**
    - **theoretical understanding**
  - **main result: intractability of nonconvex-nonconcave min-max**
  - **oscillations can be removed but only asymptotic convergence, in general**
  - **impressionistic proof vignette**
- Conclusions

Menu

- **Motivation**
- **Convex Games**
  - **training oscillations can be removed using negative momentum**
- **Nonconvex Games**
  - **are oscillations inherent/reflective of intractability?**
    - **an experiment**
    - **theoretical understanding**
  - **main result: intractability of nonconvex-nonconcave min-max**
  - **oscillations can be removed but only asymptotic convergence, in general**
  - **impressionistic proof vignette**
- **Philosophical Corollary and Conclusions**

Philosophical Corollary (my opinion, debatable)

- Cannot base multi-agent deep learning on:



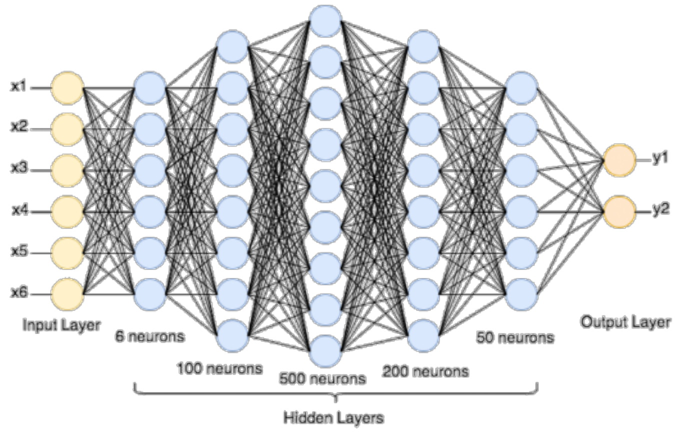
semi-agnostic

$+ \ \theta_{t+1} \leftarrow \theta_t - \nabla_\theta(f(\theta_t)) \ +$

$+$

Philosophical Corollary (my opinion, debatable)

- Cannot base multi-agent deep learning on:



semi-agnostic

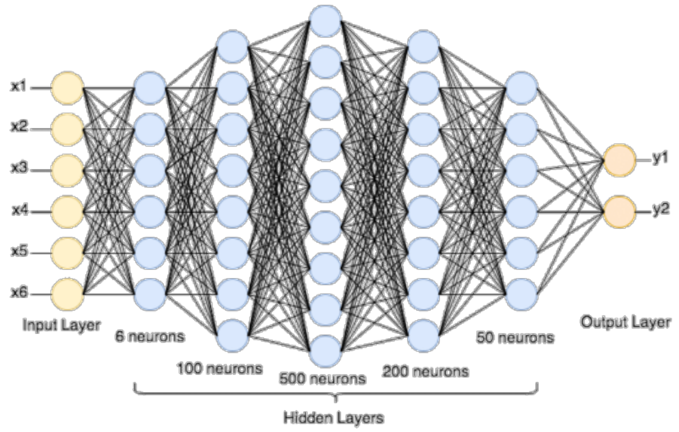
+ $\theta_{t+1} \leftarrow \theta_t - \nabla_\theta(f(\theta_t))$ +

+

- Instead will need a lot more work on (i) modeling the setting, (ii) choosing the learning model, (iii) deciding what are meaningful optimization objectives and solutions, (iv) designing the learning/optimization algorithm

Philosophical Corollary (my opinion, debatable)

- Cannot base multi-agent deep learning on:



semi-agnostic

$+ \;\; \theta_{t+1} \leftarrow \theta_t - \nabla_\theta(f(\theta_t)) \;+$

$+$

- Instead will need a lot more work on (i) modeling the setting, (ii) choosing the learning model, (iii) deciding what are meaningful optimization objectives and solutions, (iv) designing the learning/optimization algorithm

Then we might have some more successes, like AlphaGo and Libratus (which are certainly not "blindfolded GD" but use game-theoretic understanding Monte-Carlo tree search/regret minimization)

Conclusions

- Min-max optimization and equilibrium computation are intimately related to the foundations of Economics, Game Theory, Mathematical Programming, and Online Learning Theory
- They have also found profound applications in Statistics, Complexity Theory, and many other fields
- Applications in Machine Learning pose big challenges due to the dimensionality and non-convexity of the problems *(as well as the entanglement of decisions with learning)*
- I expect such applications to explode, going forward, as ML turns more to multi-agent learning applications, and (indirectly) as ML models become more complex and harder to interpret

Conclusions

- In non-convex settings, even local equilibria are generally intractable (PPAD-hardness, and first-order optimization oracle lower bounds) even in two-player zero-sum games
- **Challenge (wide open):** Identify gradient-based (or other first-order/light-weight) methods for *equilibrium learning* in multi-player games (with state)
- **Baby Challenge (wide open):** Two-player zero-sum games: $\min_x \max_y f(x, y)$
  - identify asymptotically convergent methods in general settings c.f. **[Daskalakis-Golowich-Skoulakis-Zampetakis'21]**
  - identify special cases w/ structure, enabling fast convergence to (local notions of) equilibrium
    - two-player zero-sum RL settings **[Daskalakis-Foster-Golowich NeurIPS'20]**
      - min-max theorem holds (thanks Shapley!), yet objective is not convex-concave
    - (coarse) correlated equilibrium in multi-player RL
    - non-monotone variational inequalities **[Dang-Lang'15, Zhou et al NeurIPS'17, Lin et al'18, Malitsky'19, Mertikopoulos et al ICLR'19, Liu et al ICLR'20, Song et al NeurIPS'20, J. Diakonikolas-Daskalakis-Jordan AISTATS'21]**

Thank you!