GAMES, DYNAMICS & OPTIMIZATION

Panayotis Mertikopoulos

French National Center for Scientific Research (CNRS)

Laboratoire d'Informatique de Grenoble (LIG)

Criteo AI Lab

CoreLab, NTUA – October 5, 2020

Outline

Overview

What is the long-run behavior of first-order methods in optimization / games?

Overview

What is the long-run behavior of first-order methods in optimization / games?

In optimization:

- Do first-order (= gradient-based) algorithms converge to critical points?
- Are local minimizers selected?

Overview

What is the long-run behavior of first-order methods in optimization / games?

In optimization:

- Do first-order (= gradient-based) algorithms converge to critical points?
- Are local minimizers selected?

In games:

- Do gradient methods converge to Nash equilibrium?
- Are all Nash equilibria created equal?

Overview

What is the long-run behavior of first-order methods in optimization / games?

In optimization:

- Do first-order (= gradient-based) algorithms converge to critical points?
- Are local minimizers selected?

In games:

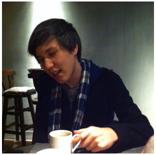
- Do gradient methods converge to Nash equilibrium?
- Are all Nash equilibria created equal?

Dynamics: from discrete to continuous and back again

About

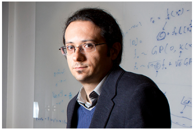
N. Hallak

A. Kavis



Y. -P. Hsieh

C. Papadimitriou



V. Cevher

G. Piliouras

Z. Zhou

- M, Papadimitriou & Piliouras, *Cycles in adversarial regularized learning,* SODA 2018
- M & Zhou, *Learning in games with continuous action sets and unknown payoff functions,* Mathematical Programming, vol. 173, pp. 465–507, Jan. 2019
- M, Hallak, Kavis & Cevher, *On the almost sure convergence of stochastic gradient descent in non-convex problems,* NeurIPS 2020
- Hsieh, M & Cevher, *The limits of min-max optimization algorithms: convergence to spurious non-critical sets,* `https://arxiv.org/abs/2006.09065`

Outline

Overview

From flows to algorithms

From algorithms to flows

Flows in games

Monotone games

Spurious limits

Basic problem

$$\text{minimize}_{x\in\mathbb{R}^d} \quad f(x)$$

- f non-convex [technical assumptions later]
- f unknown/difficult to manipulate in closed form [low precision methods]
- Single-player game: calculate best responses [more in second part]

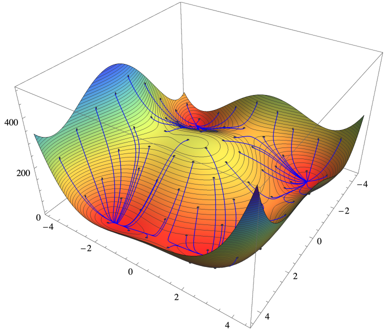
Gradient flows

Gradient flow of a function $f\colon \mathbb{R}^d \to \mathbb{R}$

$$\dot{x}(t) = -\nabla f(x(t)) \tag{GF}$$

Main property: f is a (strict) *Lyapunov function* for (GF)

$$df/dt = -\|\nabla f(x(t))\|^2 \leq 0 \qquad \text{w/ equality iff } \nabla f(x) = 0$$



Convergence of gradient flows

Blanket assumptions

- *Lipschitz smoothness:*

$$\|\nabla f(x') - \nabla f(x)\| \leq L\|x' - x\| \quad \text{for all } x, x' \in \mathbb{R}^d \tag{LS}$$

- *Bounded sublevels:*

$$L_c \equiv \{x \in \mathbb{R}^d : f(x) \leq c\} \quad \text{is bounded for all } c < \sup f \tag{sub}$$

Theorem

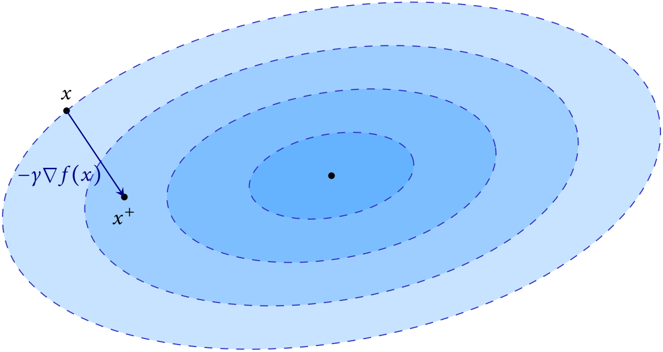
- **Assume:** (LS), (sub)
- **Then:** $x(t)$ *converges to* $\operatorname{crit}(f) \equiv \{x^* \in \mathbb{R}^d : \nabla f(x^*) = 0\}$

[NB: **setwise**, not pointwise convergence, cf. Palis Jr. and de Melo, 1982]

From flows to algorithms: gradient descent

Forward Euler (explicit) $\Longrightarrow$ **gradient descent (GD)** [Cauchy, 1847]

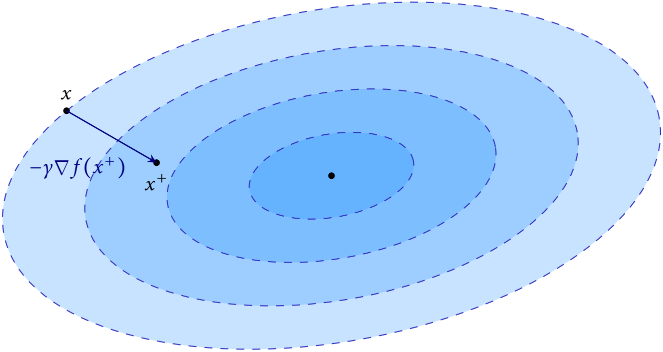
$$X_{n+1} = X_n - \gamma_n \nabla f(X_n) \tag{GD}$$



From flows to algorithms: proximal gradient

Backward Euler (implicit) $\Longrightarrow$ proximal gradient (PG) [Martinet, 1970]

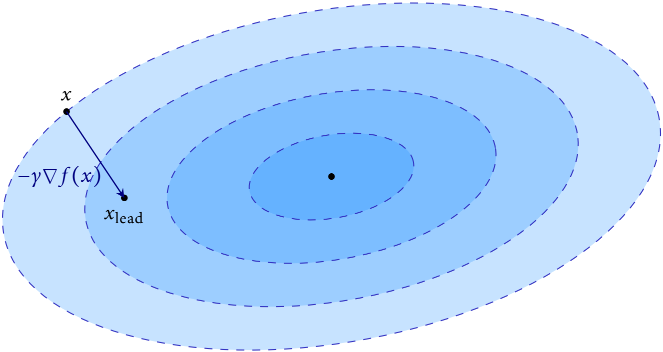
$$X_{n+1} = X_n - \gamma_n \nabla f(X_{n+1}) \qquad \text{(PG)}$$



From flows to algorithms: extra-gradient

Midpoint Runge-Kutta (explicit) $\Longrightarrow$ **extra-gradient (EG)** [Korpelevich, 1976]

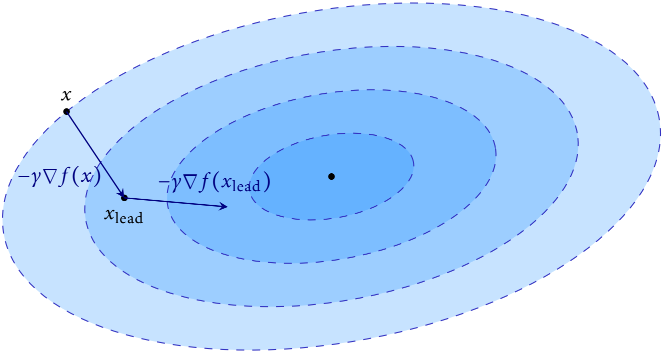
$$X_{n+1/2} = X_n - \gamma_n \nabla f(X_n) \qquad X_{n+1} = X_n - \gamma_n \nabla f(X_{n+1/2}) \tag{EG}$$



From flows to algorithms: extra-gradient

Midpoint Runge-Kutta (explicit) $\Longrightarrow$ **extra-gradient (EG)** [Korpelevich, 1976]

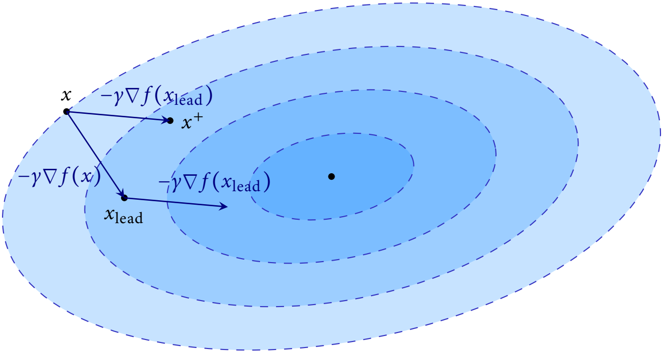
$$X_{n+1/2} = X_n - \gamma_n \nabla f(X_n) \qquad X_{n+1} = X_n - \gamma_n \nabla f(X_{n+1/2}) \tag{EG}$$



From flows to algorithms: extra-gradient

Midpoint Runge-Kutta (explicit) $\Longrightarrow$ **extra-gradient (EG)** [Korpelevich, 1976]

$$X_{n+1/2} = X_n - \gamma_n \nabla f(X_n) \qquad X_{n+1} = X_n - \gamma_n \nabla f(X_{n+1/2}) \tag{EG}$$



Stochastic gradient feedback

In many applications, perfect gradient information is unavailable / too costly:

- **Machine learning:**
  $f(x) = \sum_{i=1}^{N} f_i(x)$ and only a batch of $\nabla f_i(x)$ is computable per iteration
- **Control / Engineering:**
  $f(x) = \mathbb{E}[F(x;\omega)]$ and only $\nabla F(x;\omega)$ can be observed for a random ω
- **Game Theory / Bandit Learning:**
  Only $f(x)$ is observable

Stochastic gradient feedback

In many applications, perfect gradient information is unavailable / too costly:

- **Machine learning:**
  $f(x) = \sum_{i=1}^{N} f_i(x)$ and only a batch of $\nabla f_i(x)$ is computable per iteration
- **Control / Engineering:**
  $f(x) = \mathbb{E}[F(x;\omega)]$ and only $\nabla F(x;\omega)$ can be observed for a random ω
- **Game Theory / Bandit Learning:**
  Only $f(x)$ is observable

Stochastic first-order oracle (SFO) feedback:

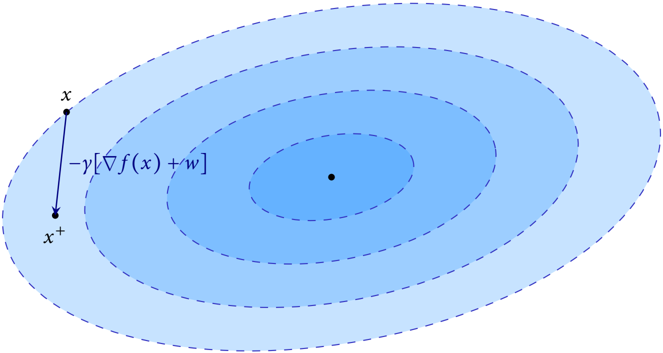
$$X_n \mapsto \underbrace{V_n}_{\text{feedback}} = \underbrace{\nabla f(X_n)}_{\text{gradient}} + \underbrace{Z_n}_{\text{noise}} + \underbrace{b_n}_{\text{bias}} \tag{SFO}$$

where Z_n is “zero-mean” and b_n is “small” (more later)

Stochastic gradient descent

Noisy Euler (explicit) $\implies$ stochastic gradient descent (SGD)

$$X_{n+1} = X_n - \gamma_n [\nabla f(X_n) + \underbrace{W_n}_{\text{noise}}] \tag{SGD}$$



Example: zeroth-order feedback

Given $f\colon \mathbb{R} \to \mathbb{R}$, estimate $f'(x)$ at target point $x \in \mathbb{R}$

$$f'(x) \approx \frac{f(x+\delta) - f(x-\delta)}{2\delta}$$

Pick $u = \pm 1$ with probability $1/2$. Then:

$$\mathbb{E}[f(x+\delta u)u] = \frac{1}{2} f(x+\delta) - \frac{1}{2} f(x-\delta)$$

$\implies$ Estimate $f'(x)$ with a single query of f at $\hat{x} = x + \delta u$

Example: zeroth-order feedback

Given $f: \mathbb{R} \to \mathbb{R}$, estimate $f'(x)$ at target point $x \in \mathbb{R}$

$$f'(x) \approx \frac{f(x+\delta) - f(x-\delta)}{2\delta}$$

Pick $u = \pm 1$ with probability $1/2$. Then:

$$\mathbb{E}[f(x+\delta u)u] = \frac{1}{2} f(x+\delta) - \frac{1}{2} f(x-\delta)$$

$\implies$ Estimate $f'(x)$ with a single query of f at $\hat{x} = x + \delta u$

Algorithm Simultaneous perturbation stochastic approximation [Spall, 1992]

1: Draw u uniformly from $\mathbb{S}^d$
2: **Query** $\hat{x} = x + \delta u$
3: **Get** $\hat{f} = f(\hat{x})$
4: **Set** $V = (d/\delta)\hat{f}u$

The Robbins-Monro template

Generalized Robbins-Monro:

$$X_{n+1} = X_n - \gamma_n[\nabla f(X_n) + Z_n + b_n] \tag{RM}$$

with $\sum_n \gamma_n = \infty$, $\gamma_n \to 0$, and $\mathbb{E}[Z_n \mid X_n, \dots, X_1] = 0$

Examples

- Gradient descent (det.): $Z_n = 0,\ b_n = 0$
- Proximal gradient (det.): $Z_n = 0,\ b_n = \nabla f(X_{n+1}) - \nabla f(X_n)$
- Extra-gradient (det.): $Z_n = 0,\ b_n = \nabla f(X_{n+1/2}) - \nabla f(X_n)$
- Stochastic gradient descent (stoch.): Z_n = zero-mean, $b_n = 0$
- SPSA (stoch.): $Z_n = (d/\delta) f(\hat{X}_n) U_n - \nabla f_\delta(X_n),\ b_n = \nabla f_\delta(X_n) - \nabla f(X_n)$ where

$$f_\delta(x) = \frac{1}{\mathrm{vol}(\mathbb{B}_\delta)} \int_{\mathbb{B}_\delta} f(x + \delta u)\, du$$

- ...

Outline

From algorithms to flows

Basic idea: *if γ_n is "small", the noise washes out and "$\lim_{t\to\infty}(\text{RM}) = \lim_{t\to\infty}(\text{GF})$"*

From algorithms to flows

Basic idea: *if γ_n is "small", the noise washes out and "$\lim_{t\to\infty}$ (RM) = $\lim_{t\to\infty}$ (GF)"*

$\implies$ ODE method of stochastic approximation

[Ljung, 1977; Benveniste et al., 1990; Kushner and Yin, 1997; Benaïm, 1999]

- **Virtual time:** $\tau_n = \sum_{k=1}^{n} \gamma_k$
- **Virtual trajectory:** $X(t) = X_n + \dfrac{t - \tau_n}{\tau_{n+1} - \tau_n}(X_{n+1} - X_n)$
- **Asymptotic pseudotrajectory (APT):**

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} \|X(t+h) - \Phi_h(X(t))\| = 0$$

where $\Phi_s(x)$ denotes the position at time s of an orbit of (GF) starting at x

- **Long run:** $X(t)$ tracks (GF) with arbitrary accuracy over windows of arbitrary length

[Benaïm and Hirsch, 1995, 1996; Benaïm, 1999; Benaïm et al., 2005, 2006]

Stochastic approximation criteria

When is a sequence generated by (RM) an APT?

(A)
- X_n is bounded
- f is *Lipschitz continuous and smooth:*

$$|f(x') - f(x)| \leq G\|x' - x\| \tag{LC}$$

$$\|\nabla f(x') - \nabla f(x)\| \leq L\|x' - x\| \tag{LS}$$

(B)
- $\mathbb{E}[\sum_n \gamma_n^2 \|Z_n\|^2] < \infty$
- $\sup_n \mathbb{E}[\|Z_n\|^q] < \infty$ and $\sum_n \gamma_n^{1+q/2} < \infty$
- Z_n sub-Gaussian and $\gamma_n = o(1/\log n)$

(C)
- $\sum_n \gamma_n b_n = 0$ with probability 1

Proposition (Benaïm, 1999; Hsieh, **M** & Cevher, 2020)

- Assume: *any of* (A); *any of* (B); (C)
- Then: X_n *is an APT of* (GF) *with probability* 1

Convergence of APTs

Theorem (Benaïm and Hirsch, 1995, 1996)

- **Assume:** X_n *is a bounded APT of* (GF)
- **Then:** X_n *converges to* $\operatorname{crit}(f)$ *with probability* 1

Convergence of APTs

Theorem (Benaïm and Hirsch, 1995, 1996)

- **Assume:** X_n *is a bounded APT of* (GF)
- **Then:** X_n *converges to* $\operatorname{crit}(f)$ *with probability* 1

Theorem (Ljung, 1977; Benaïm, 1999)

- **Assume:** (LC), (LS), (sub); $\sup_n \|X_n\| < \infty$
- **Then:** X_n *converges* (a.s.) *to a component of* $\operatorname{crit}(f)$ *where* f *is constant*

Boundedness: implicit, algorithm-dependent assumption; **non-verifiable!**

Can boundedness be dropped?

Key obstacle: infinite plains of vanishing gradients [think $f(x) = -\exp(-x^2)$]

Can boundedness be dropped?

Key obstacle: infinite plains of vanishing gradients [think $f(x) = -\exp(-x^2)$]

Countered if *gradient sublevel sets* do not extend to infinity

$$M_\varepsilon \equiv \{x \in \mathbb{R}^d : \|\nabla f(x)\| \leq \varepsilon\} \quad \text{is bounded for some } \varepsilon > 0 \tag{Gsub}$$

[standard under regularization]

Can boundedness be dropped?

Key obstacle: infinite plains of vanishing gradients [think $f(x) = -\exp(-x^2)$]

Countered if *gradient sublevel sets* do not extend to infinity

$$M_\varepsilon \equiv \{x \in \mathbb{R}^d : \|\nabla f(x)\| \leq \varepsilon\} \quad \text{is bounded for some } \varepsilon > 0 \tag{Gsub}$$

[standard under regularization]

Proposition (M, Hallak, Kavis & Cevher, 2020)

- **Assume:** (LC), (LS), (sub), (Gsub)
- **Then:** *for all $\varepsilon > 0$, there exists some $\tau = \tau(\varepsilon)$ such that, for all $t \geq \tau$:*
  (a) $f(x(t)) \leq f(x(0)) - \varepsilon$; *or*
  (b) $x(t)$ *is within ε-distance of* $\operatorname{crit}(f)$

In words: (GF) either descends f by a uniform amount, or it is already near-critical

Can boundedness be dropped?

Proposition

- Assume: (LC), (LS), (sub), (Gsub); *any of* (B); (C)
- Then: *With probability* 1*, a subsequence of* X_n *converges to* $\operatorname{crit}(f)$

Can boundedness be dropped?

Proposition

- **Assume:** (LC), (LS), (sub), (Gsub); *any of* (B); (C)
- **Then:** *With probability* 1, *a subsequence of* X_n *converges to* $\operatorname{crit}(f)$

Theorem (M, Hallak, Kavis & Cevher, 2020)

- **Assume:** (LC), (LS), (sub), (Gsub); *any of* (B); (C)
- **Then:** *With probability* 1, X_n *converges to a (possibly random) component of* $\operatorname{crit}(f)$ *over which* f *is constant*

Are all critical points desirable?

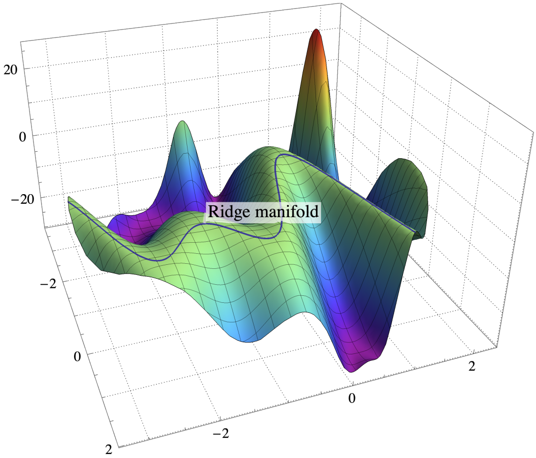


Figure: A hyperbolic ridge manifold, typical of ResNet loss landscapes [Li et al., 2018]

Are traps avoided?

Hyperbolic saddle (isolated non-minimizing critical point)

$$\lambda_{\min}(\mathrm{Hess}(f(x^*))) < 0, \quad \det(\mathrm{Hess}(f(x^*))) \neq 0$$

$\implies$ (GF) is linearly unstable near x^*

$\implies$ convergence to x^* unlikely

Are traps avoided?

Hyperbolic saddle (isolated non-minimizing critical point)

$$\lambda_{\min}(\mathrm{Hess}(f(x^*))) < 0, \quad \det(\mathrm{Hess}(f(x^*))) \neq 0$$

$\Longrightarrow$ (GF) is linearly unstable near x^*

$\Longrightarrow$ convergence to x^* unlikely

Theorem (Pemantle, 1990)

- Assume:
  - x^* *is a hyperbolic saddle point*
  - Z_n *is finite (a.s.) and uniformly exciting*

$$\mathbb{E}[\langle Z, u\rangle^+] \geq c \quad \textit{for all unit vectors } u \in \mathbb{S}^{d-1}, x \in \mathbb{R}^d$$

  - $\gamma_n \propto 1/n$
- Then: $\mathbb{P}(\lim_{n\to\infty} X_n = x^*) = 0$

Are non-hyperbolic traps avoided?

Strict saddle

$$\lambda_{\min}(\mathrm{Hess}(f(x^*))) < 0$$

Are non-hyperbolic traps avoided?

Strict saddle

$$\lambda_{\min}(\mathrm{Hess}(f(x^*))) < 0$$

Theorem (Ge et al., 2015)

- **Given:** *confidence level* $\zeta > 0$
- **Assume:**
  - f *is bounded and satisfies* (LS)
  - $\mathrm{Hess}(f(x))$ *is Lipschitz continuous*
  - *for all* $x \in \mathbb{R}^d$: (a) $\|\nabla f(x)\| \geq \varepsilon$; *or* (b) $\lambda_{\min}(\mathrm{Hess}(f(x))) \leq -\beta$; *or* (c) x *is* δ*-close to a local minimum* x^* *of* f *around which* f *is* α*-strongly convex*
  - Z_n *is finite* (a.s.) *and contains a component uniformly sampled from the unit sphere; also,* $b_n = 0$
  - $\gamma_n \equiv \gamma$ *with* $\gamma = \mathcal{O}(1/\log(1/\zeta))$
- **Then:** *with probability at least* $1 - \zeta$, *SGD produces after* $\mathcal{O}(\gamma^{-2}\log(1/(\gamma\zeta)))$ *iterations a point which is* $\mathcal{O}(\sqrt{\gamma}\log(1/(\gamma\zeta)))$*-close to* x^* (*and hence away from any strict saddle*)

Are non-hyperbolic traps avoided *always*?

Theorem (**M**, Hallak, Kavis & Cevher, 2020)

- Assume:
  - f *satisfies* (LC) *and* (LS)
  - Z_n *is finite* (a.s.) *and uniformly exciting*

$$\mathbb{E}[\langle Z, u\rangle^+] \geq c \quad \text{for all unit vectors } u \in \mathbb{S}^{d-1}, x \in \mathbb{R}^d$$

  - $\gamma_n \propto 1/n^p$ *for some* $p \in (0,1]$
- Then: $\mathbb{P}(X_n \text{ converges to a set of strict saddle points}) = 0$

Proof.

Use Pemantle (1990) + differential geometric arguments of Benaïm and Hirsch (1995). □

Outline

Single- vs. multi-agent setting

In single-agent optimization, first-order iterative schemes

- Converge to the problem's set of critical points
- Avoid spurious, non-minimizing critical manifolds

Single- vs. multi-agent setting

In **single-agent optimization**, first-order iterative schemes

- Converge to the problem's set of critical points
- Avoid spurious, non-minimizing critical manifolds

Does this intuition carry over to games?

Do **multi-agent** learning algorithms

- Converge to unilaterally stable/stationary points?
- Avoid spurious, non-equilibrium points?

Online decision processes

Agents called to take repeated decisions with **minimal** information:

for $n \geq 0$ **do**

Choose **action** X_n [focal agent choice]

Incur **loss** $\ell_n(X_n)$ [depends on all agents]

end for

Driving question: *How to choose "good" actions?*

- Unknown world: no beliefs, knowledge of the game, etc.
- Minimal information: feedback often limited to incurred losses

N-player games

The game

- Finite set of **players** $i \in \mathcal{N} = \{1, \ldots, N\}$
- Each player selects an **action** from a closed convex set $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$
- Loss of player i given by **loss function** $\ell_i \colon \mathcal{X} \equiv \prod_i \mathcal{X}_i \to \mathbb{R}$

Examples

- Finite games (mixed extensions)
- Divisible good auctions (Kelly)
- Traffic routing
- Power control/allocation problems
- Cournot oligopolies
- ...

Nash equilibrium

Nash equilibrium

Action profile $x^* = (x_1^*, \dots, x_n^*) \in \mathcal{X}$ that is **unilaterally stable**

$$\ell_i(x_i^*; x_{-i}^*) \leq \ell_i(x_i; x_{-i}^*) \quad \text{for every player } i \in \mathcal{N} \text{ and every deviation } x_i \in \mathcal{X}_i$$

- **Local Nash equilibrium:** local version [stable under local deviations]
- **Critical point:** unilateral stationarity [x_i^* is stationary for $\ell_i(\cdot, x_{-i}^*)$]

Nash equilibrium

Nash equilibrium

Action profile $x^* = (x_1^*, \ldots, x_n^*) \in \mathcal{X}$ that is unilaterally stable

$$\ell_i(x_i^*; x_{-i}^*) \leq \ell_i(x_i; x_{-i}^*) \quad \text{for every player } i \in \mathcal{N} \text{ and every deviation } x_i \in \mathcal{X}_i$$

- **Local Nash equilibrium:** local version [stable under local deviations]
- **Critical point:** unilateral stationarity [x_i^* is stationary for $\ell_i(\cdot, x_{-i}^*)$]

Individual loss gradients

$$V_i(x) = \nabla_{x_i} \ell_i(x_i; x_{-i})$$

$\Longrightarrow$ individually steepest variation

Nash equilibrium

Nash equilibrium

Action profile $x^* = (x_1^*, \dots, x_n^*) \in \mathcal{X}$ that is unilaterally stable

$$\ell_i(x_i^*; x_{-i}^*) \leq \ell_i(x_i; x_{-i}^*) \quad \text{for every player } i \in \mathcal{N} \text{ and every deviation } x_i \in \mathcal{X}_i$$

- **Local Nash equilibrium:** local version [stable under local deviations]
- **Critical point:** unilateral stationarity [x_i^* is stationary for $\ell_i(\cdot, x_{-i}^*)$]

Individual loss gradients

$$V_i(x) = \nabla_{x_i} \ell_i(x_i; x_{-i})$$

$\Longrightarrow$ individually steepest variation

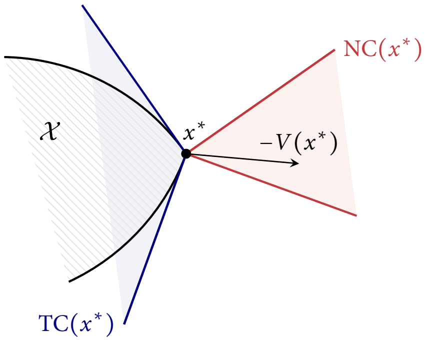
Variational characterization

If x^* is a (local) Nash equilibrium, then

$$\langle V_i(x^*), x_i - x_i^* \rangle \geq 0 \quad \text{for all } i \in \mathcal{N}, x_i \in \mathcal{X}_i$$

Intuition: $\ell_i(x_i; x_{-i}^*)$ weakly increasing along all rays emanating from x_i^*

Geometric interpretation



At Nash equilibrium, individual descent directions are outward-pointing

First-order algorithms in games

Individual gradient field $V(x) = (V_1(x), \dots, V_N(x))$, $x = (x_1, \dots, x_N)$

- **Individual gradient descent:**

$$X_{n+1} = X_n - \gamma_n V(X_n)$$

- **Extra-gradient:**

$$X_{n+1/2} = X_n - \gamma_n \nabla \ell(X_n) \qquad X_{n+1} = X_n - \gamma_n \nabla \ell(X_{n+1/2})$$

- ...

Mean dynamics:

$$\dot{x}(t) = -V(x(t)) \tag{MD}$$

$\implies$ no longer a gradient system

Outline

The dynamics of min-max games

Bilinear min-max games (saddle-point problems)

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} \quad L(x_1, x_2) = (x_1 - b_1)^\top A (x_2 - b_2) \tag{SP}$$

[no constraints: $\mathcal{X}_1 = \mathbb{R}^{d_1}$, $\mathcal{X}_2 = \mathbb{R}^{d_2}$]

Mean dynamics:

$$\dot{x}_1 = -A(x_2 - b_2) \qquad \dot{x}_2 = A^\top (x_1 - b_1)$$

The dynamics of min-max games

Bilinear min-max games (saddle-point problems)

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} \quad L(x_1, x_2) = (x_1 - b_1)^\top A (x_2 - b_2) \tag{SP}$$

[no constraints: $\mathcal{X}_1 = \mathbb{R}^{d_1}$, $\mathcal{X}_2 = \mathbb{R}^{d_2}$]

Mean dynamics:

$$\dot{x}_1 = -A(x_2 - b_2) \qquad \dot{x}_2 = A^\top (x_1 - b_1)$$

Energy function:

$$E(x) = \frac{1}{2}\|x_1 - b_1\|^2 + \frac{1}{2}\|x_2 - b_2\|^2$$

Lyapunov property:

$$\frac{dE}{dt} \leq 0 \quad \text{w/ equality if } A = A^\top$$

$\Longrightarrow$ distance to solutions (weakly) decreasing along (MD)

Cycles

Roadblock: the energy might be a constant of motion [Hofbauer et al., 2009]

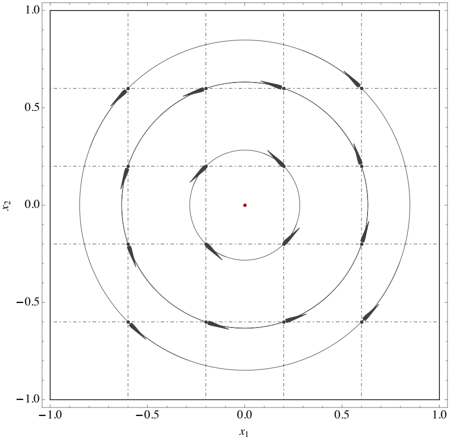
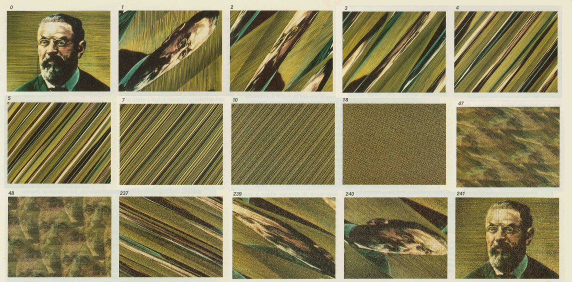


Figure: Hamiltonian flow of $L(x_1, x_2) = x_1 x_2$

Poincaré recurrence

Definition (Poincaré, 1890's)

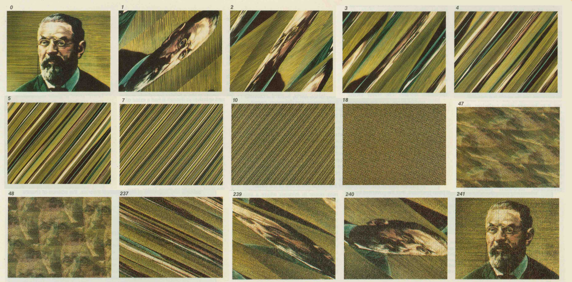
A dynamical system is **Poincaré recurrent** if almost all solution trajectories return *infinitely close* to their starting point *infinitely many times*



Poincaré recurrence

Definition (Poincaré, 1890's)

A dynamical system is **Poincaré recurrent** if almost all solution trajectories return *infinitely close* to their starting point *infinitely many times*



Theorem (**M**, Papadimitriou, Piliouras, 2018; unconstrained version)

(MD) *is Poincaré recurrent in all bilinear min-max games that admit an equilibrium*

Learning in min-max games: gradient descent

Individual gradient descent:

$$X_{n+1} = X_n - \gamma_n V(X_n)$$

Learning in min-max games: gradient descent

Individual gradient descent:

$$X_{n+1} = X_n - \gamma_n V(X_n)$$

Energy no longer a constant:

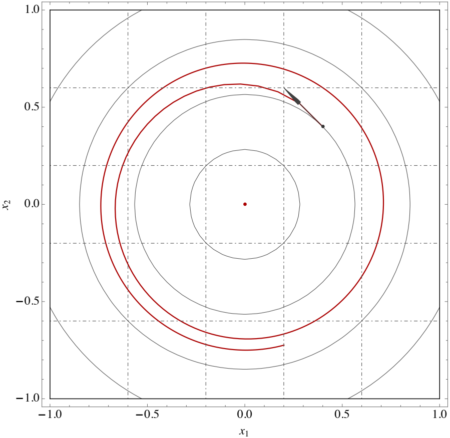
$$\frac{1}{2}\|X_{n+1} - x^*\|^2 = \frac{1}{2}\|X_n - x^*\|^2 + \gamma_n \underbrace{\cancel{\langle V(X_n), X_n - x^* \rangle}}_{\text{from (MD)}} + \frac{1}{2} \underbrace{\gamma_n^2 \|V(X_n)\|^2}_{\text{discretization error}}$$

...even worse

Learning in min-max games: gradient descent

Individual gradient descent:

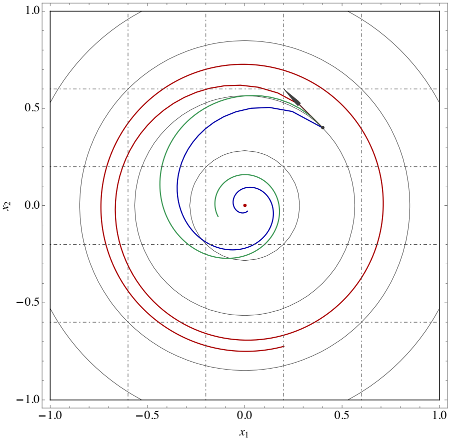
$$X_{n+1} = X_n - \gamma_n V(X_n)$$



Learning in min-max games: extra-gradient

Extra-gradient:

$$X_{n+1/2} = X_n - \gamma_n \nabla \ell(X_n) \qquad X_{n+1} = X_n - \gamma_n \nabla \ell(X_{n+1/2})$$



Learning in min-max games

Long-run behavior of min-max learning algorithms:

- ▶ **Mean dynamics:** Poincaré recurrent (periodic orbits)
- ✗ **Individual gradient descent:** divergence (outward spirals)
- ✓ **Extra-gradient:** convergence (inward spirals)

Learning in min-max games

Long-run behavior of min-max learning algorithms:

- ▸ **Mean dynamics:** Poincaré recurrent (periodic orbits)
- ✗ **Individual gradient descent:** divergence (outward spirals)
- ✓ **Extra-gradient:** convergence (inward spirals)

Different outcomes despite same underlying dynamics!

Monotonicity and strict monotonicity

Bilinear games are special cases of **monotone games:**

$$\langle V(x') - V(x), x' - x \rangle \geq 0 \quad \text{for all } x, x' \in \mathcal{X} \tag{MC}$$

[$\Longrightarrow$ **strictly monotone** if (MC) is strict for $x \neq x'$]

Monotonicity and strict monotonicity

Bilinear games are special cases of **monotone games:**

$$\langle V(x') - V(x), x' - x \rangle \geq 0 \quad \text{for all } x, x' \in \mathcal{X} \tag{MC}$$

[$\implies$ **strictly monotone** if (MC) is strict for $x \neq x'$]

Equivalently: $H(x) \succcurlyeq 0$ where H is the game's **Hessian matrix:**

$$H_{ij}(x) = \frac{1}{2}\nabla_{x_j}\nabla_{x_j}\ell_i(x) + \frac{1}{2}(\nabla_{x_i}\nabla_{x_j}\ell_j(x))^\top$$

Monotonicity and strict monotonicity

Bilinear games are special cases of **monotone games:**

$$\langle V(x') - V(x), x' - x \rangle \geq 0 \quad \text{for all } x, x' \in \mathcal{X} \tag{MC}$$

[$\Longrightarrow$ **strictly monotone** if (MC) is strict for $x \neq x'$]

Equivalently: $H(x) \succcurlyeq 0$ where H is the game's **Hessian matrix:**

$$H_{ij}(x) = \frac{1}{2}\nabla_{x_j}\nabla_{x_j}\ell_i(x) + \frac{1}{2}(\nabla_{x_i}\nabla_{x_j}\ell_j(x))^\top$$

Examples: bilinear games (not strict), Kelly auctions, Cournot markets, routing, …

Nomenclature:

- **Diagonal strict convexity** [Rosen, 1965]
- Stable games [Hofbauer and Sandholm, 2009]
- Contractive games [Sandholm, 2015]
- Dissipative games [Sorin and Wan, 2016]

Convergence to equilibrium

Different behavior under strict monotonicity:

$$\frac{1}{2}\|X_{n+1} - x^*\|^2 = \frac{1}{2}\|X_n - x^*\|^2 - \gamma_n \underbrace{\langle V(X_n), X_n - x^* \rangle}_{> 0 \text{ if } X_n \text{ not Nash}} + \frac{1}{2} \underbrace{\gamma_n^2 \|V(X_n)\|^2}_{\text{discretization error}}$$

Can the drift overcome the discretization error?

Convergence to equilibrium

Different behavior under **strict** monotonicity:

$$\frac{1}{2}\|X_{n+1}-x^*\|^2 = \frac{1}{2}\|X_n - x^*\|^2 - \gamma_n \underbrace{\langle V(X_n), X_n - x^*\rangle}_{>0 \text{ if } X_n \text{ not Nash}} + \frac{1}{2}\underbrace{\gamma_n^2 \|V(X_n)\|^2}_{\text{discretization error}}$$

Can the drift overcome the discretization error?

Theorem (M & Zhou, 2019)

- **Assume:** *strict monotonicity; any of* (A); *any of* (B); (C)
- **Then:** *any generalized Robbins-Monro learning algorithm converges to the game's (unique) Nash equilibrium with probability* 1

> In strictly monotone games, gradient methods $\rightsquigarrow$ Nash equilibrium

Outline

Almost bilinear games

Consider the "almost bilinear" game

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} \quad L(x_1, x_2) = x_1 x_2 + \varepsilon \phi(x_2)$$

where $\varepsilon > 0$ and $\phi(x) = (1/2)x^2 - (1/4)x^4$

Properties:

- Unique critical point at the origin
- Not Nash; unstable under (MD)
- (MD) attracted to unique, stable limit cycle from almost all initial conditions

[Hsieh, **M** & Cevher, 2020]

Spurious limits in almost bilinear games

Trajectories of (RM) converge to a spurious cycle that contains no critical points

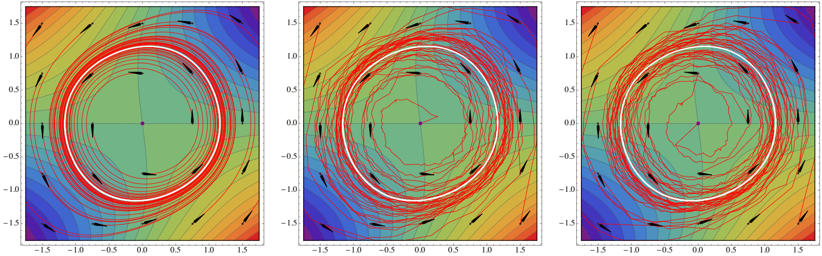


Figure: Left: (MD); center: SGD; right: stochastic extra-gradient (SEG)

Forsaken solutions

Another almost bilinear game

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} \quad L(x_1, x_2) = x_1 x_2 + \varepsilon[\phi(x_1) - \phi(x_2)]$$

where $\varepsilon > 0$ and $\phi(x) = (1/4)x^2 - (1/2)x^4 + (1/6)x^6$

Properties:

- Unique critical point at the origin
- Local Nash equilibrium; stable under (MD)
- Two isolated periodic orbits:
  - One unstable, shielding equilibrium, but small
  - One stable, attracts all trajectories of (MD) outside small basin

[Hsieh, **M** & Cevher, 2020]

Forsaken solutions in almost bilinear games

With high probability, (RM) forsakes the game's unique (local) equilibrium

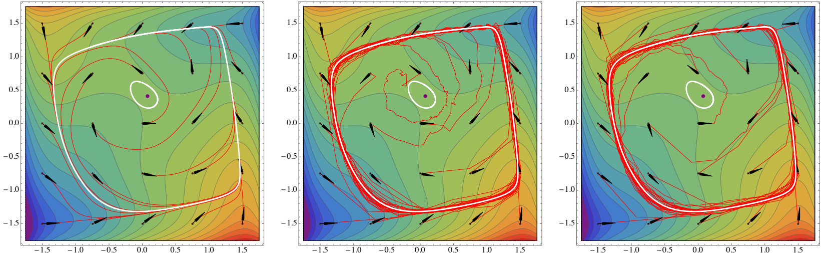


Figure: Left: (MD); center: SGD; right: SEG

The limits of gradient-based learning in games

Limit cycles $\implies$ internally chain transitive (ICT) = invariant, no proper attractors

Examples of ICT sets

- $V = \nabla \ell \implies$ components of critical points
- $L(x_1, x_2) = x_1 x_2 \implies$ any annular region centered on $(0, 0)$
- Almost bilinear $\implies$ isolated periodic orbits + unique stationary point

The limits of gradient-based learning in games

Limit cycles $\Longrightarrow$ internally chain transitive (ICT) = invariant, no proper attractors

Examples of ICT sets

- $V = \nabla \ell \Longrightarrow$ components of critical points
- $L(x_1, x_2) = x_1 x_2 \Longrightarrow$ any annular region centered on $(0, 0)$
- Almost bilinear $\Longrightarrow$ isolated periodic orbits + unique stationary point

Theorem (Hsieh, **M** & Cevher, 2020)

- Assume: *any of* (A); *any of* (B); (C)
- Then:
  - X_n *converges to an ICT of* (MD) *with probability* 1
  - (RM) *converges to attractors of* (MD) *with arbitrarily high probability*

Conclusions

In contrast to single-agent problems (optimization), game-theoretic learning

- May have limit points that are neither stable nor stationary
- Cannot avoid spurious, non-equilibrium points with positive probability
- Different approach needed (mixed-strategy learning, multiple-timescales…)

Conclusions

In contrast to single-agent problems (optimization), game-theoretic learning

- May have limit points that are neither stable nor stationary
- Cannot avoid spurious, non-equilibrium points with positive probability
- Different approach needed (mixed-strategy learning, multiple-timescales…)

What about finite games?

- Limit cycles may still appear
- Which Nash equilibria are stable under no-regret learning?

[stay tuned to CoreLab FM ☺]

References I

M. Benaïm. Dynamics of stochastic approximation algorithms. In J. Azéma, M. Émery, M. Ledoux, and M. Yor, editors, *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer Berlin Heidelberg, 1999.

M. Benaïm and M. W. Hirsch. Dynamics of Morse-Smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6):1005–1030, December 1995.

M. Benaïm and M. W. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions, part II: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points – Online stochastic gradient for tensor decomposition. In *COLT '15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

J. Hofbauer and W. H. Sandholm. Stable games and their dynamics. *Journal of Economic Theory*, 144 (4):1665–1693, July 2009.

J. Hofbauer, S. Sorin, and Y. Viossat. Time average replicator and best reply dynamics. *Mathematics of Operations Research*, 34(2):263–269, May 2009.

References II

Y.-P. Hsieh, P. Mertikopoulos, and V. Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. `https://arxiv.org/abs/2006.09065`, 2020.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

H. J. Kushner and G. G. Yin. *Stochastic approximation algorithms and applications*. Springer-Verlag, New York, NY, 1997.

H. Li, Z. Xu, G. Taylor, C. Suder, and T. Goldstein. Visualizing the loss landscape of neural nets. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Autom. Control*, 22(4):551–575, August 1977.

B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis*, 4(R3):154–158, 1970.

P. Mertikopoulos and Z. Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1-2):465–507, January 2019.

P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *SODA '18: Proceedings of the 29th annual ACM-SIAM Symposium on Discrete Algorithms*, 2018.

References III

P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

J. Palis Jr. and W. de Melo. *Geometric Theory of Dynamical Systems*. Springer-Verlag, 1982.

R. Pemantle. Nonconvergence to unstable points in urn models and stochastic aproximations. *Annals of Probability*, 18(2):698–712, April 1990.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Optimization*, 14(5):877–898, 1976.

J. B. Rosen. Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica*, 33(3):520–534, 1965.

W. H. Sandholm. Population games and deterministic evolutionary dynamics. In H. P. Young and S. Zamir, editors, *Handbook of Game Theory IV*, pages 703–778. Elsevier, 2015.

S. Sorin and C. Wan. Finite composite games: Equilibria and dynamics. *Journal of Dynamics and Games*, 3(1):101–120, January 2016.

J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control*, 37(3):332–341, March 1992.