MIN-MAX OPTIMIZATION FROM A DYNAMICAL SYSTEMS VIEWPOINT

Panayotis Mertikopoulos

French National Center for Scientific Research (CNRS)

Laboratoire d'Informatique de Grenoble (LIG)

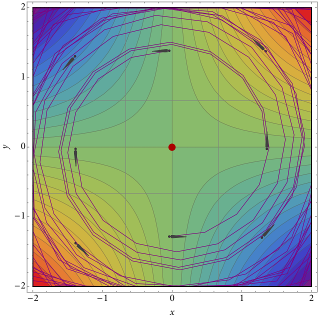
Criteo AI Lab

⟨ Adversarial Approaches in ML | UC Berkeley | February 23, 2022 ⟩

Outline

Bilinear min-max problems

A playground for adversarial approaches: $\min_{a \leq x_1 \leq b} \max_{a \leq x_2 \leq b} f(x_1, x_2) = x_1 x_2$

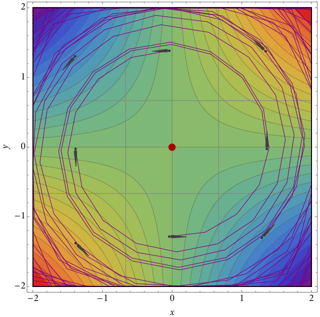


(a) Vanilla gradient

$$X_{n+1} = X_n - \gamma_n V_n$$
$$V \leftarrow \mathrm{oracle}(\partial_1 f, -\partial_2 f)$$

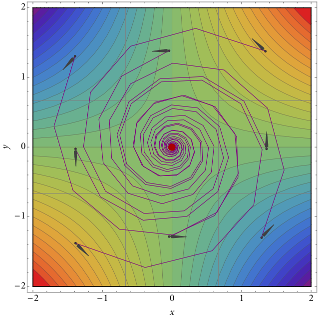
Bilinear min-max problems

A playground for adversarial approaches: $\min_{a \leq x_1 \leq b} \max_{a \leq x_2 \leq b} f(x_1, x_2) = x_1 x_2$



(a) Vanilla gradient

$$X_{n+1} = X_n - \gamma_n V_n$$
$$V \leftarrow \operatorname{oracle}(\partial_1 f, -\partial_2 f)$$

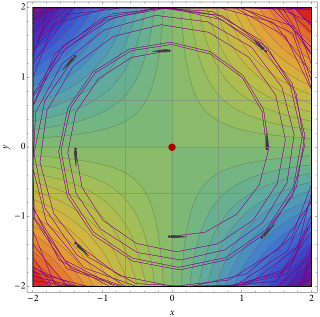


(b) Extra-gradient [Korpelevich, 1976]

$$\begin{aligned} X_{n+1/2} &= X_n - \gamma_n V_n \\ X_{n+1} &= X_n - \gamma_n V_{n+1/2} \end{aligned} \tag{EG}$$

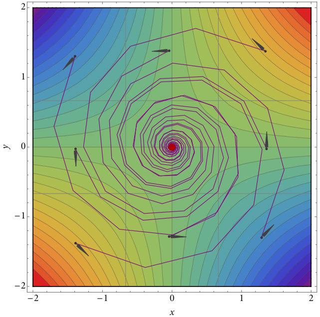
Bilinear min-max problems

A playground for adversarial approaches: $\min_{a \le x_1 \le b} \max_{a \le x_2 \le b} f(x_1, x_2) = x_1 x_2$



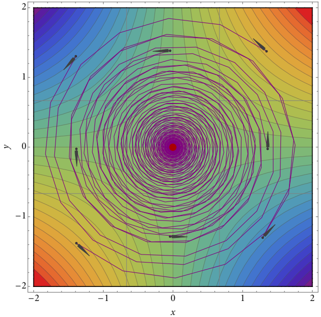
(a) Vanilla gradient

$$X_{n+1} = X_n - \gamma_n V_n$$
$$V \leftarrow \text{oracle}(\partial_1 f, -\partial_2 f)$$



(b) Extra-gradient [Korpelevich, 1976]

$$\begin{aligned} X_{n+1/2} &= X_n - \gamma_n V_n \\ X_{n+1} &= X_n - \gamma_n V_{n+1/2} \end{aligned} \tag{EG}$$

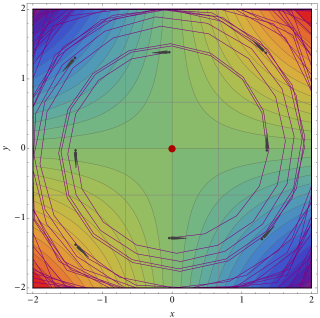


(c) Optimistic gradient [Popov, 1980]

$$\begin{aligned} X_{n+1/2} &= X_n - \gamma_n V_{n-1/2} \\ X_{n+1} &= X_n - \gamma_n V_{n+1/2} \end{aligned} \tag{OG}$$

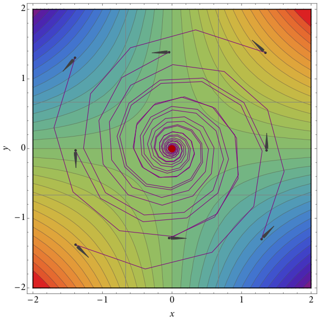
Bilinear min-max problems

A playground for adversarial approaches: $\displaystyle \min_{a \le x_1 \le b} \max_{a \le x_2 \le b} f(x_1, x_2) = x_1 x_2$



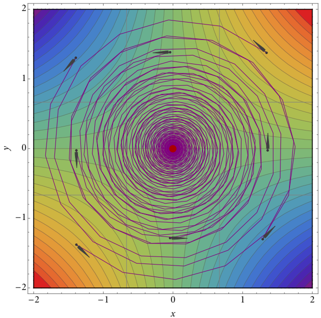
(a) Vanilla gradient

$$X_{n+1} = X_n - \gamma_n V_n$$
$$V \leftarrow \text{oracle}(\partial_1 f, -\partial_2 f)$$



(b) Extra-gradient [Korpelevich, 1976]

$$\begin{aligned} X_{n+1/2} &= X_n - \gamma_n V_n \\ X_{n+1} &= X_n - \gamma_n V_{n+1/2} \end{aligned} \qquad \text{(EG)}$$



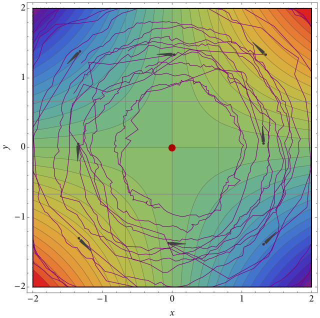
(c) Optimistic gradient [Popov, 1980]

$$\begin{aligned} X_{n+1/2} &= X_n - \gamma_n V_{n-1/2} \\ X_{n+1} &= X_n - \gamma_n V_{n+1/2} \end{aligned} \qquad \text{(OG)}$$

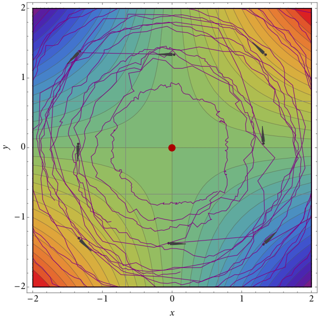
Improved properties of (EG) / (OG) $\Longrightarrow$ huge literature + testing ground for new algorithms

Stochastic min-max problems

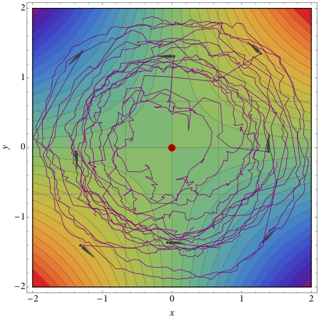
The stochastic world is **different:**



(a) Stochastic gradient



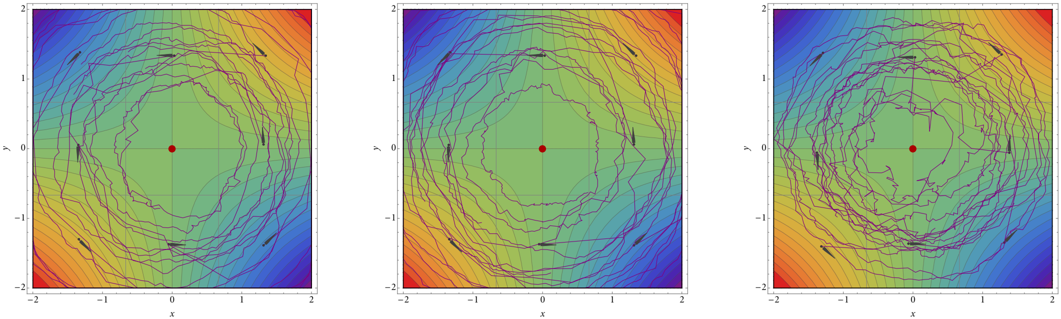
(b) Stochastic EG [Juditsky et al., 2011]



(c) Stochastic OG [Gidel et al., 2019]

Stochastic min-max problems

The stochastic world is **different:**



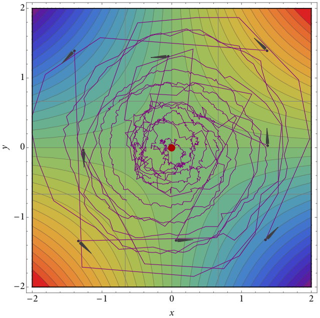
(a) Stochastic gradient (b) Stochastic EG [Juditsky et al., 2011] (c) Stochastic OG [Gidel et al., 2019]

Noise mitigation mechanisms:

- Iterate averaging [For convex-concave problems]
- Variance reduction [Chavdarova et al., 2019]
- Double step-size policies [Hsieh et al., 2020; Diakonikolas et al., 2021]

Stochastic min-max problems

The stochastic world is **different:**



(d) Double step-size EG

Double step-size extra-gradient

$$\begin{aligned} X_{n+1/2} &= X_n - \gamma_n V_n \\ X_{n+1} &= X_n - \eta_n V_{n+1/2} \end{aligned} \tag{DSEG}$$

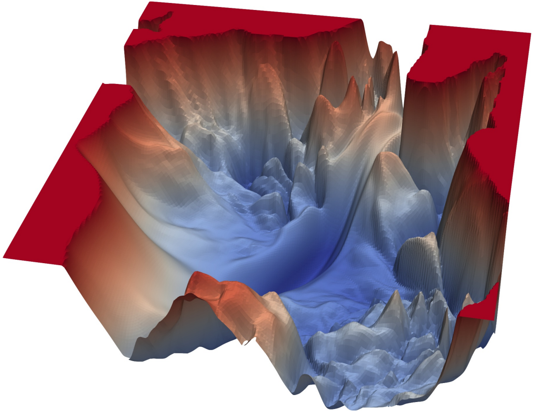
where $\eta_n/\gamma_n \to 0$

"Explore aggressively, update conservatively"

Noise mitigation mechanisms:
- Double step-size policies [Hsieh et al., 2020]

Training landscape

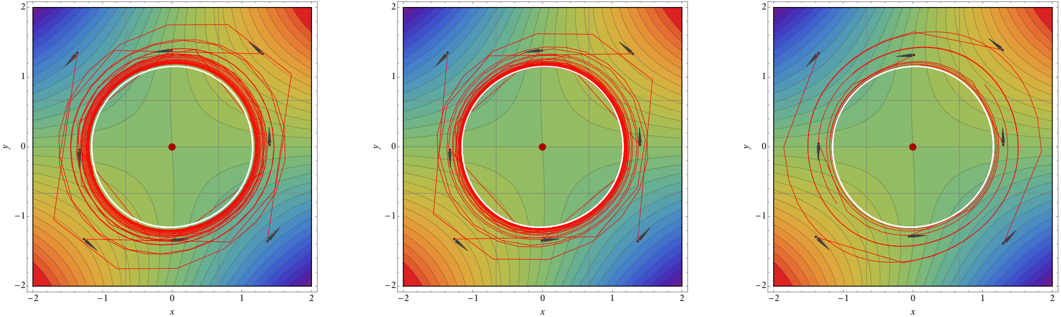
A deep learning loss landscape



[Source: Li et al., 2018]

Out of the bilinear sandbox

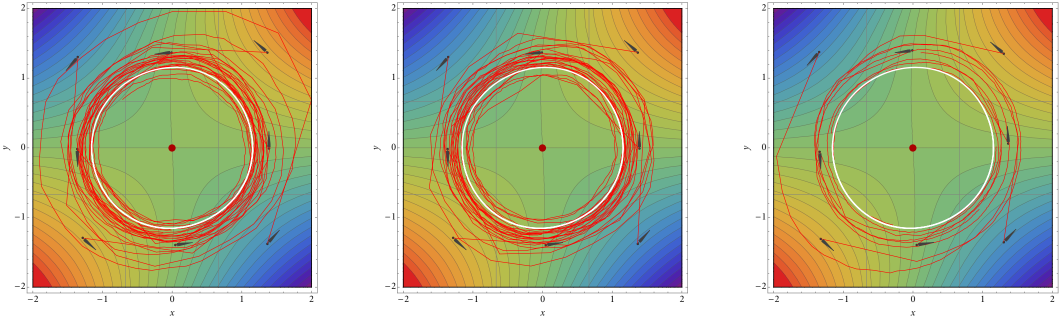
The non-monotone world is **fundamentally different:** $\min_{a \le x_1 \le b} \max_{a \le x_2 \le b} f(x_1, x_2) = x_1 x_2 + \varepsilon(x_2^2/2 - x_2^4/4)$ $[\varepsilon \approx 0]$



(a) Vanilla gradient (b) Extra-gradient (c) Double step-size extra-gradient

Out of the bilinear sandbox

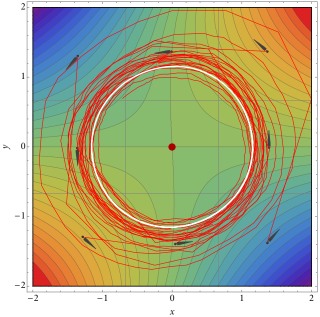
The non-monotone world is **fundamentally different:** $\min_{a \le x_1 \le b} \max_{a \le x_2 \le b} f(x_1, x_2) = x_1 x_2 + \varepsilon(x_2^2/2 - x_2^4/4)$ $[\varepsilon \approx 0]$



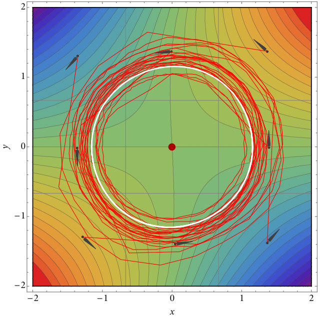
(a) Vanilla gradient (stoch.) (b) Extra-gradient (stoch.) (c) Double step-size extra-gradient (stoch.)

Out of the bilinear sandbox

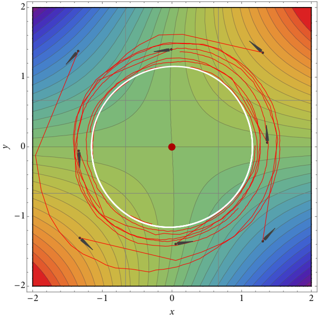
The non-monotone world is **fundamentally different:** $\min_{a\le x_1\le b} \max_{a\le x_2\le b} f(x_1,x_2) = x_1x_2 + \varepsilon(x_2^2/2 - x_2^4/4)$ [$\varepsilon \approx 0$]



(a) Vanilla gradient (stoch.)



(b) Extra-gradient (stoch.)



(c) Double step-size extra-gradient (stoch.)

- Different methods **no longer lead to different outcomes,** even for arbitrarily small ε [Here: $\varepsilon = 10^{-2}$]
- **Stochasticity is not important** in the long run [Converge to same limit cycle]

Out of the bilinear sandbox

The non-monotone world is **fundamentally different:** $\min_{a \le x_1 \le b} \max_{a \le x_2 \le b} f(x_1, x_2) = x_1 x_2 + \varepsilon(x_2^2/2 - x_2^4/4)$ [$\varepsilon \approx 0$]

Why does this happen?

- Different methods **no longer lead to different outcomes,** even for arbitrarily small ε [Here: $\varepsilon = 10^{-2}$]
- **Stochasticity is not important** in the long run [Converge to same limit cycle]

Overview

What is the long-run behavior of first-order methods in non-linear min-max games?

Overview

What is the long-run behavior of first-order methods in non-linear min-max games?

In minimization problems:

- ✓ First-order (= gradient-based) algorithms converge to critical points
- ✓ Saddle points are avoided (one way or another)

Overview

What is the long-run behavior of first-order methods in non-linear min-max games?

In minimization problems:

- ✓ First-order (= gradient-based) algorithms converge to critical points
- ✓ Saddle points are avoided (one way or another)

In min-max problems / games:

- ✍ Do gradient methods converge to critical points?
- ✍ What are the possible limit sets?

Overview

What is the long-run behavior of first-order methods in non-linear min-max games?

In minimization problems:

- ✓ First-order (= gradient-based) algorithms converge to critical points
- ✓ Saddle points are avoided (one way or another)

In min-max problems / games:

- ✍ Do gradient methods converge to critical points?
- ✍ What are the possible limit sets?

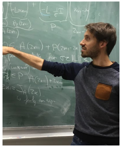
Dynamical systems viewpoint: from discrete to continuous time and back

About


V. Cevher


Y.-P. Hsieh


Y.-G. Hsieh


F. Iutzeler


J. Malick


C. Papadimitriou


G. Piliouras


Z. Zhou

- Hsieh, M & Cevher, ***The limits of min-max optimization algorithms: convergence to spurious non-critical sets,*** ICML 2021
- Hsieh, Iutzeler, Malick & M, ***Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling,*** NeurIPS 2020
- M, Papadimitriou & Piliouras, ***Cycles in adversarial regularized learning,*** SODA 2018
- M, Hsieh & Cevher, ***Online learning in games: A unified view through the lens of stochastic approximation,*** forthcoming
- M & Zhou, ***Learning in games with continuous action sets and unknown payoff functions,*** Mathematical Programming, vol. 173, pp. 465-507, Jan. 2019

Outline

Mathematical formulation

Minimization problems

$$\min_{x \in \mathcal{X}} f(x) \tag{Opt}$$

Min-max / Saddle-point problems

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} f(x_1, x_2) \tag{SP}$$

Mathematical formulation

Minimization problems (stochastic)

$$\min_{x \in \mathcal{X}} f(x) = \mathbb{E}_\theta[F(x;\theta)] \tag{Opt}$$

Min-max / Saddle-point problems (stochastic)

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} f(x_1, x_2) = \mathbb{E}_\theta[F(x_1, x_2;\theta)] \tag{SP}$$

Problem formulation

Main difficulties:

- No convex structure [technical assumptions later]
- Difficult to manipulate f in closed form [black-box oracle methods]

Problem formulation

Main difficulties:

- No convex structure [technical assumptions later]
- Difficult to manipulate f in closed form [black-box oracle methods]

Critical points:

$$\text{Find } x^* \text{ such that } v(x^*) = 0 \tag{FOS}$$

where $v(x)$ is the problem's ***defining vector field***

- **Gradient field** for (Opt):

$$v(x) = \nabla f(x)$$

- **Individual gradient field** for (SP):

$$v(x) = (\nabla_{x_1} f(x_1, x_2), -\nabla_{x_2} f(x_1, x_2))$$

[Notation: $x \leftarrow (x_1, x_2)$, $\mathcal{X} \leftarrow \mathcal{X}_1 \times \mathcal{X}_2$]

Assumptions

Blanket assumptions

- ***Unconstrained problems:***

$$\mathcal{X} = \text{finite-dimensional Euclidean space}$$

- ***Existence of solutions:***

$$\operatorname{crit}(f) = \{x^* \in \mathcal{X} : v(x^*) = 0\} \text{ is nonempty}$$

- ***Lipschitz continuity:***

$$|f(x') - f(x)| \leq G\|x' - x\| \quad \text{for all } x, x' \in \mathcal{X} \tag{LC}$$

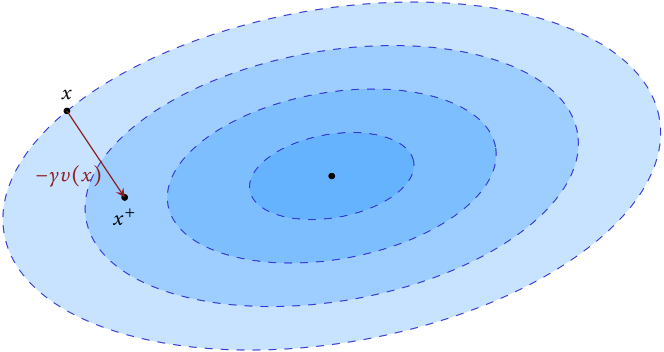
- ***Lipschitz smoothness:***

$$\|v(x') - v(x)\| \leq L\|x' - x\| \quad \text{for all } x, x' \in \mathcal{X} \tag{LS}$$

Algorithms I: Gradient descent

Gradient descent (+/ascent): [Arrow et al., 1958]

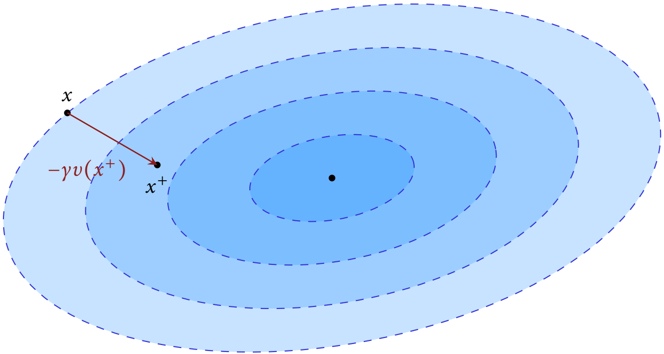
$$X_{n+1} = X_n - \gamma_n v(X_n) \tag{GD}$$



Algorithms II: Proximal point method

Proximal point method: [Martinet, 1970; Rockafellar, 1976]

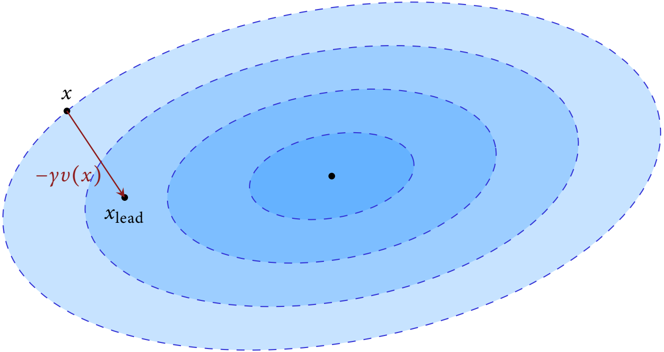
$$X_{n+1} = X_n - \gamma_n v(X_{n+1}) \tag{PPM}$$



Algorithms III: Extra-gradient

Extra-gradient: [Korpelevich, 1976]

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

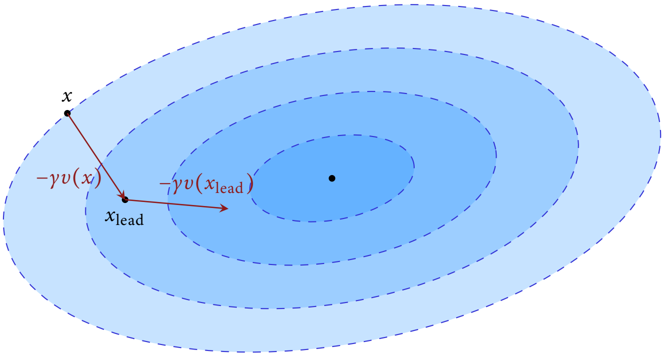


Algorithms III: Extra-gradient

Extra-gradient:

[Korpelevich, 1976]

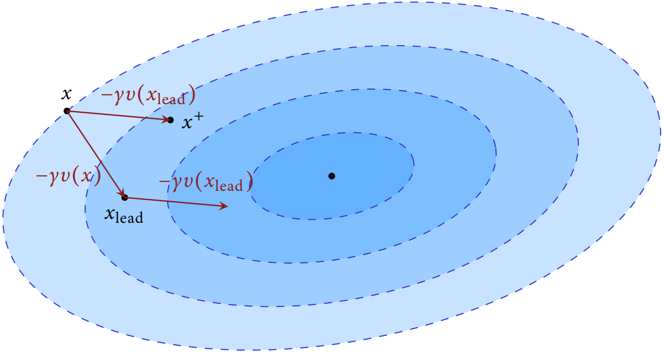
$$X_{n+1/2} = X_n - \gamma_n v(X_n) \qquad X_{n+1} = X_n - \gamma_n v(X_{n+1/2}) \tag{EG}$$



Algorithms III: Extra-gradient

Extra-gradient: [Korpelevich, 1976]

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

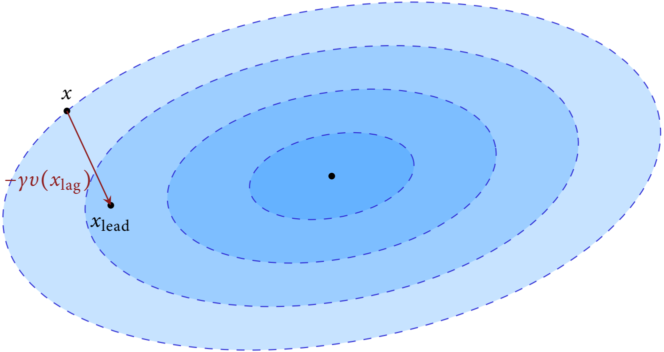


Algorithms IV: Optimistic gradient

Optimistic gradient:

[Popov, 1980; Rakhlin & Sridharan, 2013]

$$X_{n+1/2} = X_n - \gamma_n v(X_{n-1/2}) \qquad X_{n+1} = X_n - \gamma_n v(X_{n+1/2}) \tag{OG}$$

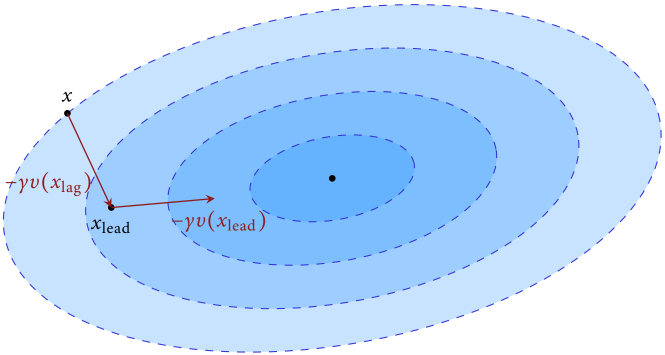


Algorithms IV: Optimistic gradient

Optimistic gradient:

[Popov, 1980; Rakhlin & Sridharan, 2013]

$$X_{n+1/2} = X_n - \gamma_n v(X_{n-1/2}) \qquad X_{n+1} = X_n - \gamma_n v(X_{n+1/2}) \tag{OG}$$

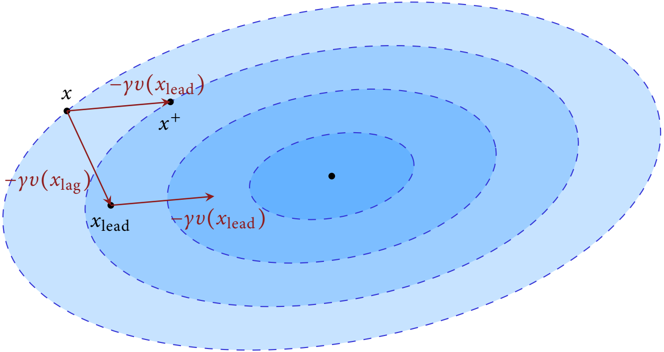


Algorithms IV: Optimistic gradient

Optimistic gradient:

[Popov, 1980; Rakhlin & Sridharan, 2013]

$$X_{n+1/2} = X_n - \gamma_n v(X_{n-1/2}) \qquad X_{n+1} = X_n - \gamma_n v(X_{n+1/2}) \tag{OG}$$



Algorithms V: more than you can shake a stick at...

Variants for min-max problems:

- *Alternating algorithms:*

$$\begin{aligned} \text{Player 1:} \quad X_{1,n+1} &= X_{1,n} - \gamma_n v(X_{1,n}, X_{2,n}) \\ \text{Player 2:} \quad X_{2,n+1} &= X_{2,n} - \gamma_n v(X_{1,n+1}, X_{2,n}) \end{aligned} \tag{GD$_{\text{alt}}$}$$

(+ variants for extra/optimistic/...)

Algorithms V: more than you can shake a stick at...

Variants for min-max problems:

- ***Alternating algorithms:***

$$\begin{aligned} \text{Player 1:} \quad & X_{1,n+1} = X_{1,n} - \gamma_n v(X_{1,n}, X_{2,n}) \\ \text{Player 2:} \quad & X_{2,n+1} = X_{2,n} - \gamma_n v(X_{1,n+1}, X_{2,n}) \end{aligned} \tag{GD$_{\text{alt}}$}$$

(+ variants for extra/optimistic/...)

- $k:1$ ***algorithms:***

$$\begin{aligned} \text{Player 1:} \quad & X^{(1)}_{1,n+1} = X_{1,n} - \gamma_n v(X_{1,n}, X_{2,n}) \\ \text{Player 1:} \quad & X^{(2)}_{1,n+1} = X^{(1)}_{1,n+1} - \gamma_n v(X^{(1)}_{1,n+1}, X_{2,n}) \\ & \cdots \\ \text{Player 1:} \quad & X_{1,n+1} = X^{(k-1)}_{1,n+1} - \gamma_n v(X^{(k-1)}_{1,n+1}, X_{2,n}) \\ \text{Player 2:} \quad & X_{2,n+1} = X_{2,n} - \gamma_n v(X_{1,n+1}, X_{2,n}) \end{aligned} \tag{GD$_{k:1}$}$$

(practical implementation of two-time-scale methods)

Algorithms V: more than you can shake a stick at...

Variants for min-max problems:

- ***Alternating algorithms:***

$$\begin{aligned} \text{Player 1:} \quad & X_{1,n+1} = X_{1,n} - \gamma_n v(X_{1,n}, X_{2,n}) \\ \text{Player 2:} \quad & X_{2,n+1} = X_{2,n} - \gamma_n v(X_{1,n+1}, X_{2,n}) \end{aligned} \tag{GD$_{\text{alt}}$}$$

  (+ variants for extra/optimistic/...)

- $k:1$ ***algorithms:***

$$\begin{aligned} \text{Player 1:} \quad & X_{1,n+1}^{(1)} = X_{1,n} - \gamma_n v(X_{1,n}, X_{2,n}) \\ \text{Player 1:} \quad & X_{1,n+1}^{(2)} = X_{1,n+1}^{(1)} - \gamma_n v(X_{1,n+1}^{(1)}, X_{2,n}) \\ & \cdots \\ \text{Player 1:} \quad & X_{1,n+1} = X_{1,n+1}^{(k-1)} - \gamma_n v(X_{1,n+1}^{(k-1)}, X_{2,n}) \\ \text{Player 2:} \quad & X_{2,n+1} = X_{2,n} - \gamma_n v(X_{1,n+1}, X_{2,n}) \end{aligned} \tag{GD$_{k:1}$}$$

  (practical implementation of two-time-scale methods)

- Chambolle-Pock; step-size scaling; variance reduction; ...

The Robbins-Monro template

Generalized Robbins-Monro algorithm

$$X_{n+1} = X_n - \gamma_n [v(X_n) + U_n + b_n] \tag{RM}$$

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

Examples

- Gradient descent (+/ascent): $b_n = 0$
- Proximal point method (det.): $U_n = 0$, $b_n = v(X_{n+1}) - v(X_n)$
- Extra-gradient: $b_n = v(X_{n+1/2}) - v(X_n)$
- Optimistic gradient: $b_n = v(X_{n+1/2}) - v(X_n)$
- Single-point stochastic approximation (stoch.): $U_n = (d/\varepsilon) f(\hat{X}_n) W_n - v_\varepsilon(X_n)$, $b_n = v_\varepsilon(X_n) - v(X_n)$ where
$$f_\varepsilon(x) = \frac{1}{\operatorname{vol}(\mathbb{B}_\delta)} \int_{\mathbb{B}_\delta} f(x + \varepsilon z)\, dz$$
- …

Outline

Mean dynamics in continuous time

Characteristic property of Robbins-Monro (RM) schemes

$$\frac{X_{n+1} - X_n}{\gamma_n} = -v(X_n) + Z_n$$

Mean dynamics

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

Stochastic approximation

Basic idea: *if γ_n is "small", the errors wash out and* "$\lim_{t\to\infty}(\mathrm{RM}) = \lim_{t\to\infty}(\mathrm{MD})$"

Stochastic approximation

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

$\Longrightarrow$ **ODE method of stochastic approximation**

[Ljung, 1977; Benveniste et al., 1990; Kushner & 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):
$$\lim_{t\to\infty} \sup_{0\le h\le T} \|X(t+h) - \Phi_h(X(t))\| = 0$$
where $\Phi_s(x)$ denotes the position at time s of an orbit of (MD) starting at x
- ***Long run:*** $X(t)$ tracks (MD) with arbitrary accuracy over windows of arbitrary length

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

Stochastic approximation criteria

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

(Base)

- f is ***Lipschitz continuous and smooth:***

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

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

- f is ***weakly coercive:*** $\langle v(x), x\rangle \geq 0$ for sufficiently large x

(Impl)

- $b_n \to 0$ with probability 1
- $\mathbb{E}[\sum_n \gamma_n \|b_n\|] < \infty$
- $\mathbb{E}[\sum_n \gamma_n^2 (1 + \|U_n\|^2)] < \infty$

Proposition (Benaïm & Hirsch, 1996)

➥ **Assume:** (Base) + (Impl)

✍ ***Then:*** X_n *is an APT of* (MD) *with probability* 1

APT criteria: explicit

Explicit algorithmic criteria:

(Expl)

- Black-box oracle:

$$\mathsf{V}(x) = v(x) + \mathsf{err}(x)$$

- Oracle returns unbiased gradients with finite mean square error

$$\mathbb{E}[\mathsf{V}(x)] = v(x) \qquad \mathbb{V}[\mathsf{V}(x)] \leq \sigma^2$$

NB: unbiasedness at query point **does not mean** $b_n = 0$ in (RM)

- $A/n \leq \gamma_n \leq B/\sqrt{n(\log n)^{1+\varepsilon}}$ for some $A, B, \varepsilon > 0$

Proposition (Hsieh, M & Cevher, 2021)

➥ **Assume: (Base) + (Expl)**

✍ ***Then:*** *the sequence X_n generated by any of the Algorithms I-V is an APT*

Outline

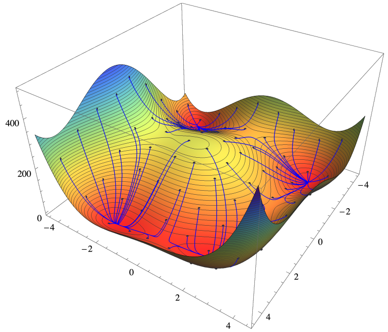
Convergence of gradient flows

Gradient flow of a function $f\colon \mathcal{X} \to \mathbb{R}$

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

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

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



Single- vs. multi-agent setting

In minimization problems:

✓ RM methods onverge to the problem's critical set [Ljung, 1977; Kushner & Yin, 1997; Benaïm & Hirsch, 1996]

✓ RM methods avoid spurious, saddle-point manifolds [Pemantle, 1990; Ge et al., 2015; Lee et al., 2019; M et al., 2020]

Single- vs. multi-agent setting

In minimization problems:

- ✓ RM methods onverge to the problem's critical set [Ljung, 1977; Kushner & Yin, 1997; Benaïm & Hirsch, 1996]
- ✓ RM methods avoid spurious, saddle-point manifolds [Pemantle, 1990; Ge et al., 2015; Lee et al., 2019; M et al., 2020]

Does this intuition carry over to min-max optimization problems?

Do min-max algorithms:

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

Min-max dynamics

The main issue:

- ✓ **Minimization problems:** (MD) is a gradient flow
- ✗ **Min-max problems:** (MD) can be arbitrarily complicated

Min-max dynamics

The main issue:

✓ **Minimization problems:** (MD) is a gradient flow

✗ **Min-max problems:** (MD) can be arbitrarily complicated

An assorted zoology of stationary sets

- **Invariant:** image of $\mathcal{S}$ under (MD) = $\mathcal{S}$ [$\Phi_t(\mathcal{S}) = \mathcal{S}$ for all t]
- **Attracting:** invariant + attracts all nearby orbits of (MD)
- **Internally chain transitive:** invariant + (MD) restricted on $\mathcal{S}$ contains no proper attractors

Examples

Some examples (more later):

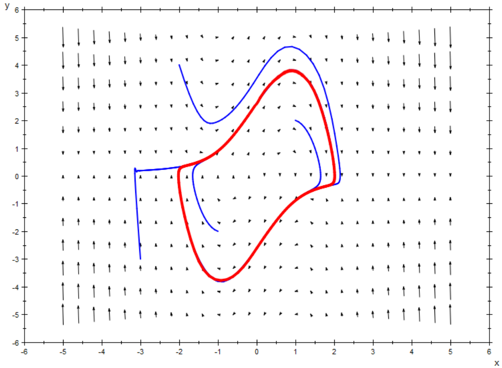


Figure: An attracting limit cycle (the Van Der Pol osillator)

Examples

Some examples (more later):

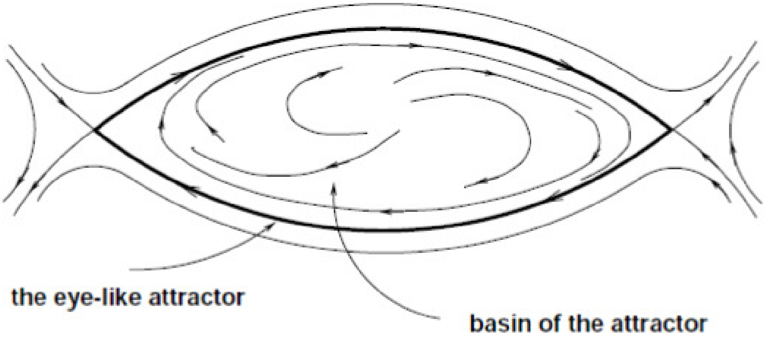


Figure: An attracting heteroclinic cycle (Bowen's eye)

Convergence to ICT sets

Theorem (Benaïm, 1999; implicit)

➥ **Assume:** (Base) + (Impl)

✍ ***Then:*** *with probability* 1*, the sequence* X_n *generated by* (RM) *converges to an ICT set of* (MD)

Theorem (Hsieh, M & Cevher, 2021; explicit)

➥ **Assume:** (Base) + (Expl)

✍ ***Then:*** *with probability* 1*, the sequence* X_n *generated by any of the Algs. I-V converges to an ICT set of* (MD)

Avoidance of unstable points and periodic orbits

Generically, any ICT set $\mathcal{S}$ possesses ***stable*** and ***unstable*** manifolds:

- **Stable manifold:** invariant + all trajectories starting here **converge** to $\mathcal{S}$
- **Unstable manifold:** invariant + all trajectories starting here **diverge** from $\mathcal{S}$
- **Unstable point / periodic orbit:** possesses a nontrivial unstable manifold

Avoidance of unstable points and periodic orbits

Generically, any ICT set $\mathcal{S}$ possesses ***stable*** and ***unstable*** manifolds:

- **Stable manifold:** invariant + all trajectories starting here **converge** to $\mathcal{S}$
- **Unstable manifold:** invariant + all trajectories starting here **diverge** from $\mathcal{S}$
- **Unstable point / periodic orbit:** possesses a nontrivial unstable manifold

Theorem (Hsieh, M & Cevher, 2021)

➥ **Assume:**

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

$$\mathbb{E}[\langle U, z\rangle^+] \geq c \quad \textit{for all unit vectors } z \in \mathbb{S}^{d-1}, x \in \mathcal{X}$$

- $\gamma_n \propto 1/n^p$ *for some* $p \in (0, 1]$

✍ ***Then:*** $\mathbb{P}(X_n \textit{ converges to an unstable point / periodic orbit}) = 0$

Minimization vs. min-max optimization

Qualitatively similar landscape (??)

- Components of critical points $\leftrightsquigarrow$ ICT sets
- Avoidance of strict saddles $\leftrightsquigarrow$ avoidance of unstable periodic orbits

Is there a fundamental difference between min and min-max problems?

Minimization vs. min-max optimization

Qualitatively similar landscape (??)

- Components of critical points $\leftrightsquigarrow$ ICT sets ✗
- Avoidance of strict saddles $\leftrightsquigarrow$ avoidance of unstable periodic orbits ✓

Is there a fundamental difference between min and min-max problems?

Non-gradient problems may have spurious ICT sets!

Outline

Bilinear games redux

Bilinear min-max games

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

Mean dynamics:

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

Bilinear games redux

Bilinear min-max games

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

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)

Periodic orbits

Roadblock: the energy may be a constant of motion

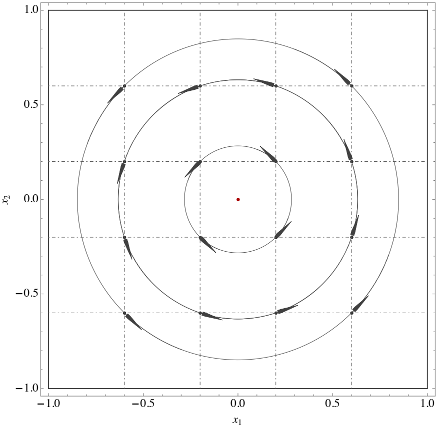
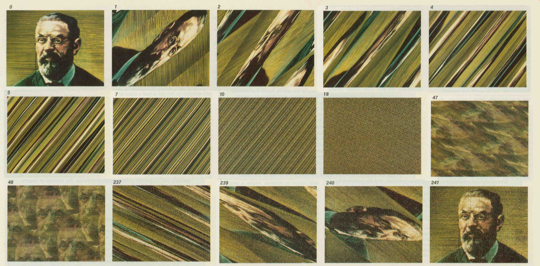


Figure: Hamiltonian flow of $f(x_1, x_2) = x_1 x_2$

Poincaré recurrence

Definition (Poincaré, 1890's)

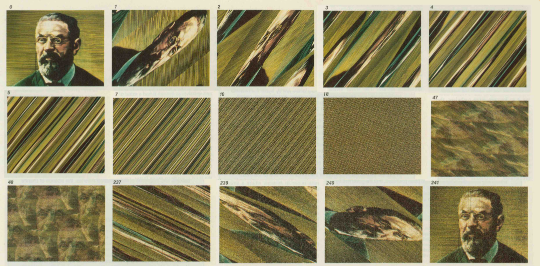
A system is **Poincaré recurrent** if almost all solution trajectories return ***infinitely close*** to their starting point ***infinitely often***



Poincaré recurrence

Definition (Poincaré, 1890's)

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



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

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

Behavior of gradient descent

Vanilla gradient:

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

Behavior of gradient descent

Vanilla gradient:

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

Energy no longer a constant:

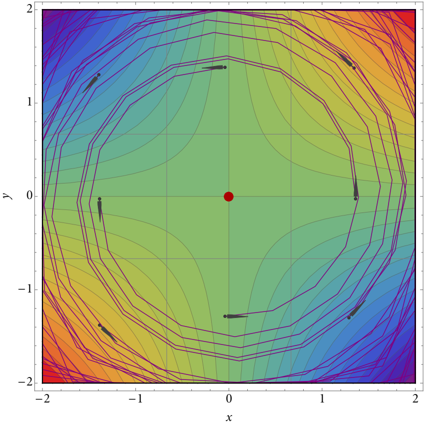
$$\frac{1}{2}\|X_{n+1} - x^*\|^2 = \frac{1}{2}\|X_n - x^*\|^2 + \gamma_n \underbrace{\cancel{\langle v(X_n), X_n - x^* \rangle}}_{\text{from (MD)}} + \frac{1}{2}\underbrace{\gamma_n^2 \|v(X_n)\|^2}_{\text{discretization error}}$$

...even worse

Behavior of gradient descent

Vanilla gradient:

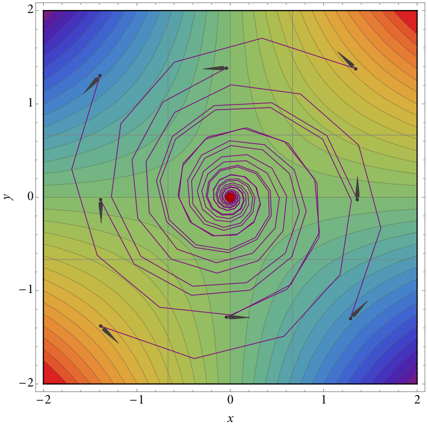
$$X_{n+1} = X_n - \gamma_n v(X_n)$$



Behavior of extra-gradient

Extra-gradient:

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



Recap

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]

Recap

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 mean dynamics!

The stochastic case

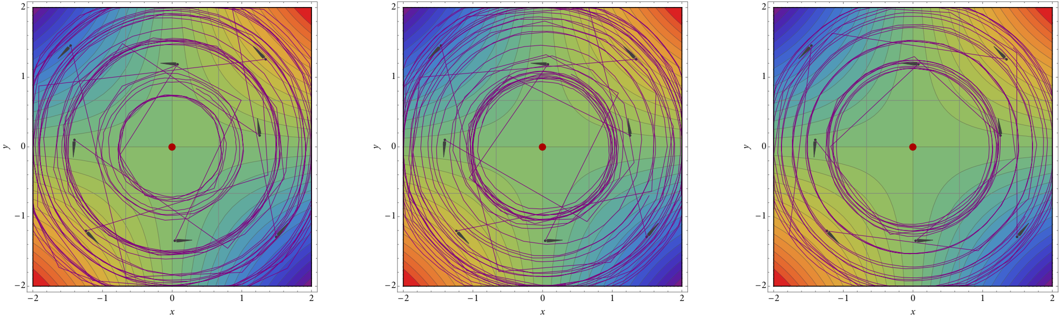


Figure: Behavior of (GD), (EG) and (OG) with stochastic first-order oracle feedback

Proposition (Hsieh, M & Cevher, 2021)

Under **(Base)** + **(Expl)***, all Algs. I-V converge to a (possibly random) periodic orbit*

[But see also Chavdarova et al., 2019; Hsieh et al., 2020]

The degeneracy issue

Degeneracy of ICTs

- The state space "foliates" into disjoint periodic orbits [Every point is recurrent]
- All periodic orbits are Lyapunov stable [Nearby initializations remain nearby]
- None of these orbits is attracting [No "preferred" outcome]

➥ **Long-run behavior difficult to predict!**

The degeneracy issue

Degeneracy of ICTs

- The state space "foliates" into disjoint periodic orbits [Every point is recurrent]
- All periodic orbits are Lyapunov stable [Nearby initializations remain nearby]
- None of these orbits is attracting [No "preferred" outcome]

➥ **Long-run behavior difficult to predict!**

How common is this situation?

The Kupka-Smale theorem

Systems with the structure of bilinear games are **rare:**

Theorem (Kupka, 1963)

Let $\mathcal{V} = C^2(\mathbb{R}^d; \mathbb{R}^d)$ be the space of C^2 vector fields on $\mathbb{R}^d$ endowed with the Whitney topology. Then the set of vector fields with a non-trivial recurrent set is ***meager*** *(in the Baire category sense).*

The Kupka-Smale theorem

Systems with the structure of bilinear games are **rare:**

Theorem (Kupka, 1963)

*Let $\mathcal{V} = C^2(\mathbb{R}^d; \mathbb{R}^d)$ be the space of C^2 vector fields on $\mathbb{R}^d$ endowed with the Whitney topology. Then the set of vector fields with a non-trivial recurrent set is **meager** (in the Baire category sense).*

Theorem (Smale, 1963)

For any vector field $v \in \mathcal{V}$, the following properties are generic (in the Baire category sense):

- *All closed orbits are **hyperbolic***
- *Heteroclinic orbits are **transversal** (i.e., stable and unstable manifolds intersect transversally)*

The Kupka-Smale theorem

Systems with the structure of bilinear games are **rare:**

Theorem (Kupka, 1963)

*Let $\mathcal{V} = C^2(\mathbb{R}^d; \mathbb{R}^d)$ be the space of C^2 vector fields on $\mathbb{R}^d$ endowed with the Whitney topology. Then the set of vector fields with a non-trivial recurrent set is **meager** (in the Baire category sense).*

Theorem (Smale, 1963)

For any vector field $v \in \mathcal{V}$, the following properties are generic (in the Baire category sense):

- *All closed orbits are **hyperbolic***
- *Heteroclinic orbits are **transversal** (i.e., stable and unstable manifolds intersect transversally)*

TL;DR

➥ Non-attracting periodic orbits are **non-generic** (they occur negligibly often)

Convergence to attractors

Attractors $\rightsquigarrow$ natural solution concepts for non-minimization problems

Theorem (Hsieh, M & Cevher, 2021; implicit)

➥ **Assume:** (Base) + (Impl); *$\mathcal{S}$ is an attractor; X_n is generated by* (RM)

✍ ***Then:*** *for all $\alpha > 0$, there exists a neighborhood $\mathcal{U}$ of $\mathcal{S}$ such that*

$$\mathbb{P}(X_n \text{ converges to } \mathcal{S} \mid X_1 \in \mathcal{U}) \geq 1 - \alpha$$

Theorem (Hsieh, M & Cevher, 2021; explicit)

➥ **Assume:** (Base) + (Expl); *$\mathcal{S}$ is an attractor; X_n is generated by any of Algorithms I-V*

✍ ***Then:*** *for all $\alpha > 0$, there exists a neighborhood $\mathcal{U}$ of $\mathcal{S}$ such that*

$$\mathbb{P}(X_n \text{ converges to } \mathcal{S} \mid X_1 \in \mathcal{U}) \geq 1 - \alpha$$

Foliations are fragile

Consider again the "**almost bilinear**" game

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

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

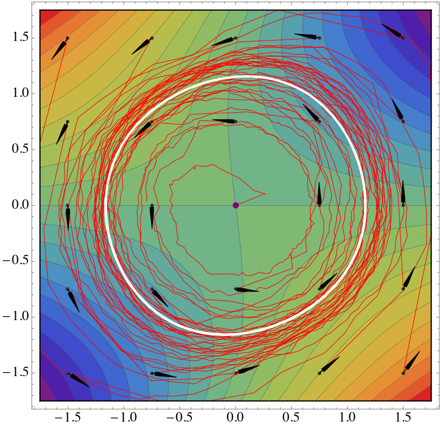
Properties:

- Unique critical point at the origin
- **Unstable under** (MD)
- (MD) attracted to unique, stable limit cycle from almost all initial conditions

[Hsieh, M & Cevher, 2021]

Spurious attractors in almost bilinear games

Trajectories of (RM) converge to a spurious limit cycle with **no critical points**



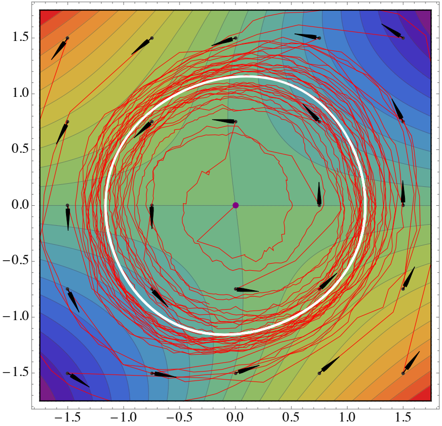


Figure: Left: stochastic gradient descent (SGD); right: stochastic extra-gradient

Forsaken solutions

Another almost bilinear game

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

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

Properties:

- **Unique (local) min-max point near the origin**
- **Two isolated non-constant periodic orbits:**
  - One **unstable**, shielding critical point, but small
  - One stable, attracts all trajectories of (MD) outside small basin

[Hsieh, M & Cevher, 2021]

Forsaken solutions in almost bilinear games

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

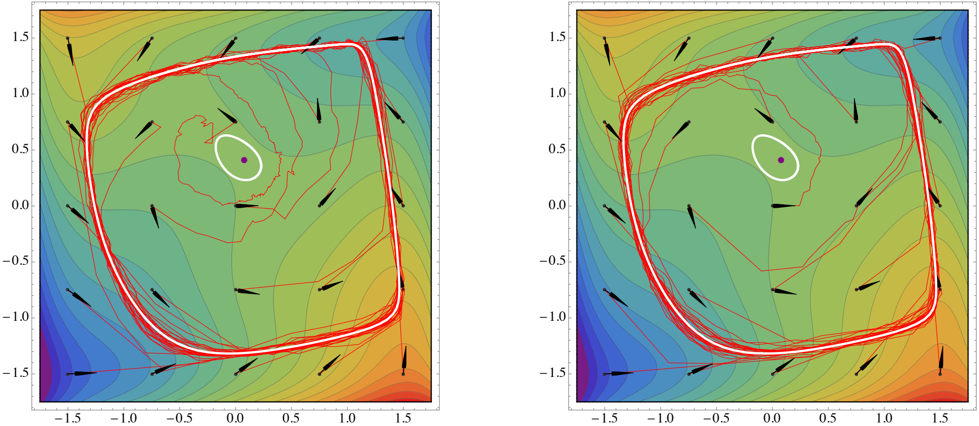


Figure: Left: stochastic gradient descent; right: stochastic extra-gradient

Conclusions

Minimization and min-max optimization are fundamentally different:

- First-order min-max methods may have limit points that are **neither stable nor stationary**
- Bilinear games **may not be representative** case studies for min-max optimization
- **Cannot avoid spurious, non-equilibrium sets** with positive probability
- **Different approaches needed** (mixed-strategy learning, multiple-timescales,...)

Conclusions

Minimization and min-max optimization are fundamentally different:

- First-order min-max methods may have limit points that are **neither stable nor stationary**
- Bilinear games **may not be representative** case studies for min-max optimization
- **Cannot avoid spurious, non-equilibrium sets** with positive probability
- **Different approaches needed** (mixed-strategy learning, multiple-timescales,...)

Many open questions:

- How to detect spurious cycles in a real system?
- Is there ***any*** first-order method that converges only to critical points?
- What about finite games (where bilinear games are no longer fragile)?
- Which equilibria are stable under first-order methods?
- ...

References I

Arrow, K. J., Hurwicz, L., and Uzawa, H. *Studies in linear and non-linear programming*. Stanford University Press, 1958.

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

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

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

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

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

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

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. Reducing noise in GAN training with variance reduced extragradient. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Diakonikolas, J., Daskalakis, C., and Jordan, M. I. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *AISTATS '21: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

References II

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

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. In *ICLR '19: Proceedings of the 2019 International Conference on Learning Representations*, 2019.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

Juditsky, A., Nemirovski, A. S., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1): 17–58, 2011.

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

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

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1):311–337, February 2019.

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

References III

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

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

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

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

Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. 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.

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

Popov, L. D. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845-848, 1980.

Rakhlin, A. and Sridharan, K. Optimization, learning, and games with predictable sequences. In *NIPS '13: Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2013.

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