FROM ROBBINS-MONRO TO ARTIFICIAL INTELLIGENCE

70 YEARS OF STOCHASTIC APPROXIMATION & THE ROAD AHEAD

Παναγιώτης Μερτικόπουλος

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Τμήμα Μαθηματικών

⟨ Γενικό Σεμινάριο | ΕΚΠΑ, Τμήμα Μαθηματικών | 30 Μαρτίου, 2023 ⟩

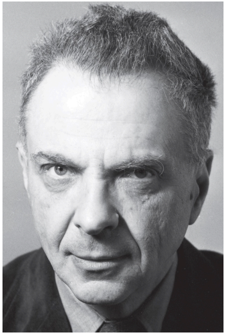
Outline

1. Background & motivation
2. The classical theory
3. Applications to minimization problems
4. Applications to min-max problems

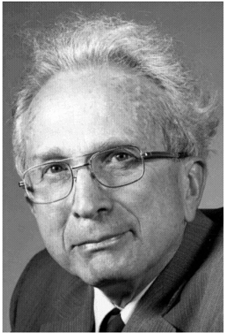
Stochastic approximation: from the 1950's...

Stochastic approximation

Find a root of a nonlinear system involving unknown functions, accessible only via noisy evaluations




Herbert Robbins & Sutton Monro

Jack Kiefer & Jacob Wolfowitz

Figure: The pioneers of the theory of stochastic approximation

...to the 2020's

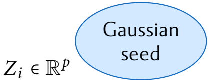
Which person is fake?

...to the 2020's

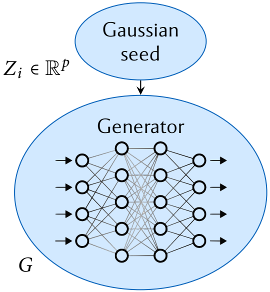
Which person is fake?

☞ https://thispersondoesnotexist.com

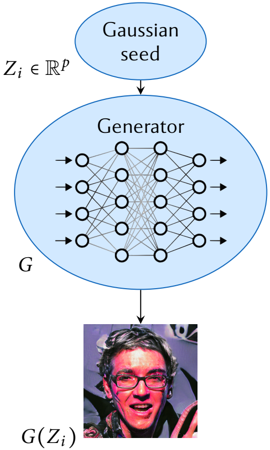
Generative adversarial networks



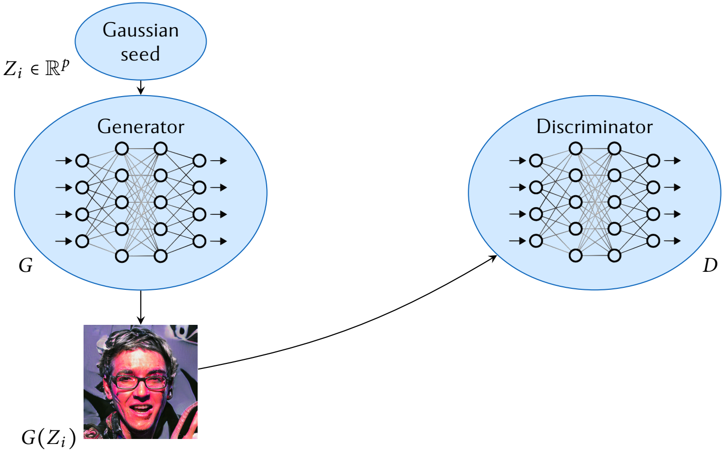
Generative adversarial networks



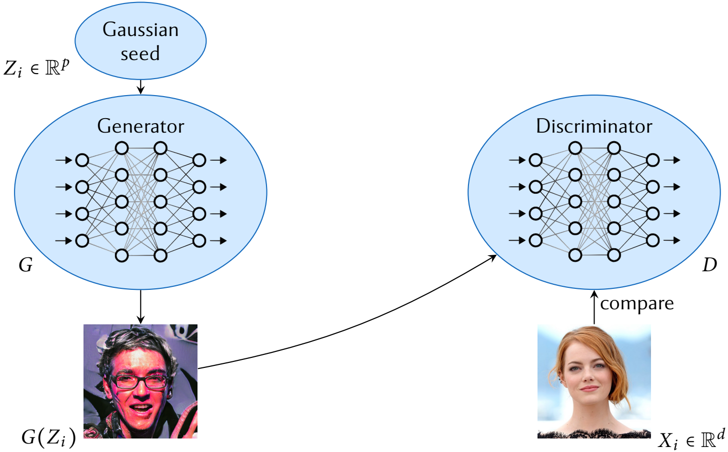
Generative adversarial networks



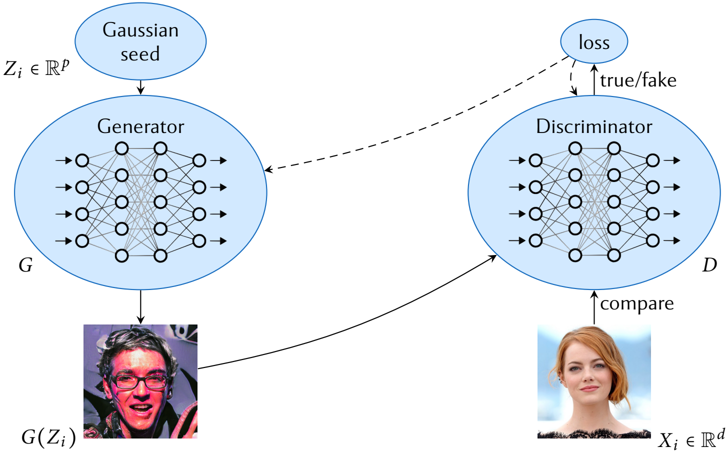
Generative adversarial networks



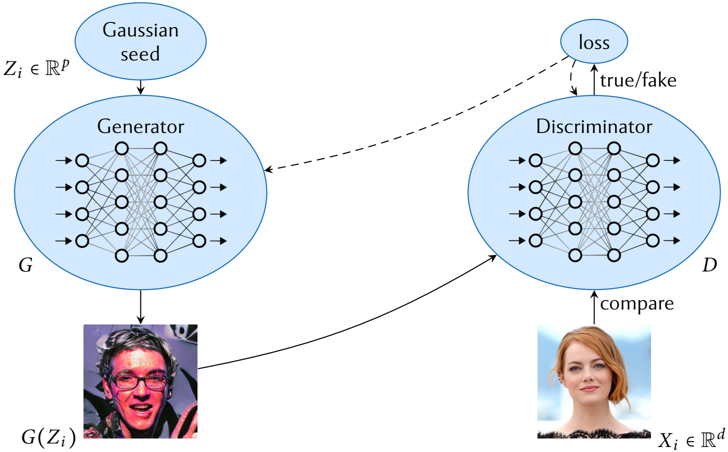
Generative adversarial networks



Generative adversarial networks



Generative adversarial networks



Model likelihood: $$L(G, D) = \prod_{i=1}^{N} D(X_i) \times \prod_{i=1}^{N} (1 - D(G(Z_i)))$$

GAN training

How to find good generators ($G \in \mathcal{G}$) and discriminators ($D \in \mathcal{D}$)?

Discriminator: maximize (log-)likelihood estimation

$$\max_{D \in \mathcal{D}} \log L(G, D)$$

Generator: minimize the resulting divergence

$$\min_{G \in \mathcal{G}} \max_{D \in \mathcal{D}} \log L(G, D)$$

Training a GAN $\iff$ solving a min-max problem

Loss surfaces

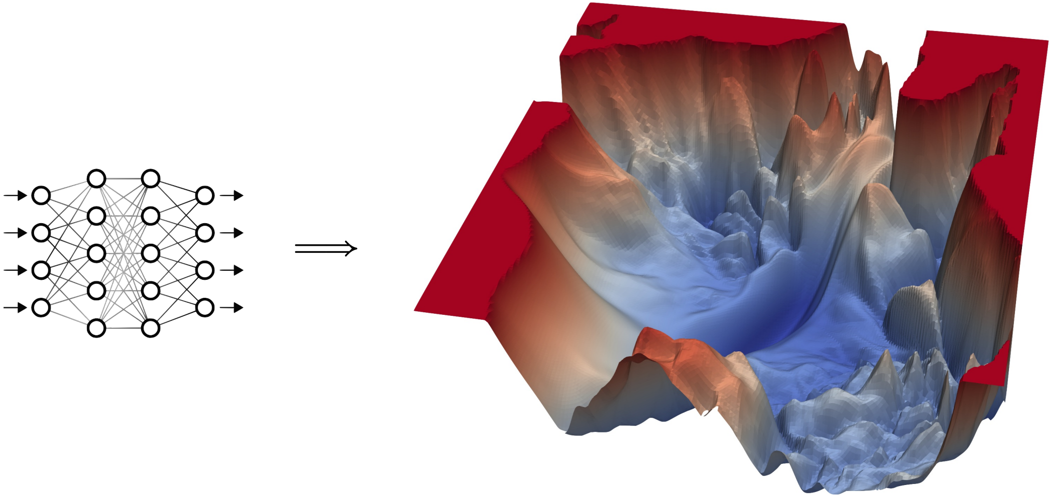


Figure: The loss landscape of a deep neural network [Li et al., 2018].

Mathematical formulation

Minimization problems

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

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)

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

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

Focus on ***critical points:***

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

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

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

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

- **Hamiltonian field** for (SP):

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

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:***

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

- ***Lipschitz continuity:***

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

- ***Lipschitz smoothness:***

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

Overview

Stochastic approximation

$$X_{n+1} = X_n - \gamma_n \hat{g}_n \tag{SA}$$

where $\hat{g}_n$, $n = 1, 2, \ldots$, is a "*stochastic approximation*" of $g(X_n)$ and $\gamma_n > 0$ is a "*step-size*" parameter.

Overview

Stochastic approximation

$$X_{n+1} = X_n - \gamma_n \hat{g}_n \tag{SA}$$

where $\hat{g}_n$, $n = 1, 2, \ldots$, is a "*stochastic approximation*" of $g(X_n)$ and $\gamma_n > 0$ is a "*step-size*" parameter.

Main question: what is the long-run behavior of X_n?

Overview

Stochastic approximation

$$X_{n+1} = X_n - \gamma_n \hat{g}_n \tag{SA}$$

where $\hat{g}_n$, $n = 1, 2, \dots$, is a "*stochastic approximation*" of $g(X_n)$ and $\gamma_n > 0$ is a "*step-size*" parameter.

Main question: what is the long-run behavior of X_n?

In minimization problems:

- ✓ First-order (= gradient-based) algorithms converge to critical points
- ✓ Are non-minimizers avoided?

Overview

Stochastic approximation

$$X_{n+1} = X_n - \gamma_n \hat{g}_n \tag{SA}$$

where $\hat{g}_n$, $n = 1, 2, \ldots$, is a "*stochastic approximation*" of $g(X_n)$ and $\gamma_n > 0$ is a "*step-size*" parameter.

Main question: what is the long-run behavior of X_n?

In minimization problems:

- ✓ First-order (= gradient-based) algorithms converge to critical points
- ✓ Are non-minimizers avoided?

In min-max problems / games:

- ☞ Do gradient methods converge to critical points?
- ☞ Are non-equilibrium sets avoided?

Overview

Stochastic approximation

$$X_{n+1} = X_n - \gamma_n \hat{g}_n \tag{SA}$$

where $\hat{g}_n$, $n = 1, 2, \ldots$, is a "*stochastic approximation*" of $g(X_n)$ and $\gamma_n > 0$ is a "*step-size*" parameter.

Main question: what is the long-run behavior of X_n?

In minimization problems:

- ✓ First-order (= gradient-based) algorithms converge to critical points
- ✓ Are non-minimizers avoided?

In min-max problems / games:

- ☞ Do gradient methods converge to critical points?
- ☞ Are non-equilibrium sets avoided?

Dynamical systems: from discrete to continuous time and back

Outline

Stochastic approximation algorithms

Stochastic approximation template

$$X_{n+1} = X_n - \gamma_n \hat{g}_n \tag{SA}$$

where:

- $X_n \in \mathbb{R}^d$ is the ***state*** of the method at epoch $n = 1, 2, \dots$
- $\gamma_n > 0$ is a variable ***step-size*** parameter
- $\hat{g}_n \in \mathbb{R}^d$ is a ***stochastic approximation*** of $g(X_n)$

Blanket assumptions

1. ***Step-size sequence:*** γ_n is of the form γ/n^p # $\gamma > 0$, $p \in [0, 1]$
2. ***Random error:*** $U_n = \hat{g}_n - \mathbb{E}[\hat{g}_n \mid \mathcal{F}_n]$ is bounded as $\mathbb{E}[\|U_n\|^q \mid \mathcal{F}_n] \leq \sigma_n^q$ # $q \geq 2$
3. ***Systematic error:*** $b_n = \mathbb{E}[\hat{g}_n \mid \mathcal{F}_n] - g(X_n)$ is bounded as $\mathbb{E}[\|b_n\| \mid \mathcal{F}_n] \leq B_n$

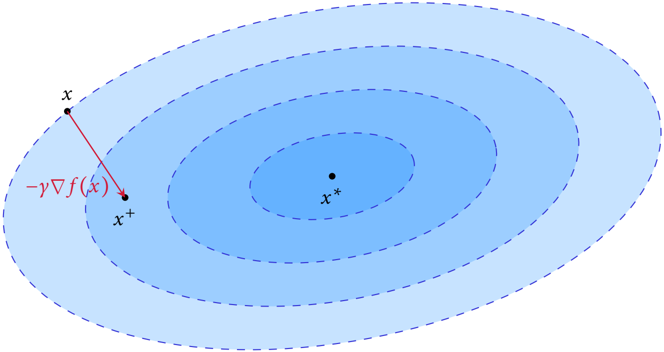
where:

- $B_n, \sigma_n \geq 0$ are **deterministic** sequences
- $\mathcal{F}_n = \mathcal{F}(X_1, \dots, X_n)$ is the ***history*** of X_n

Methods, I: Gradient descent

Gradient descent [Cauchy, 1847]

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

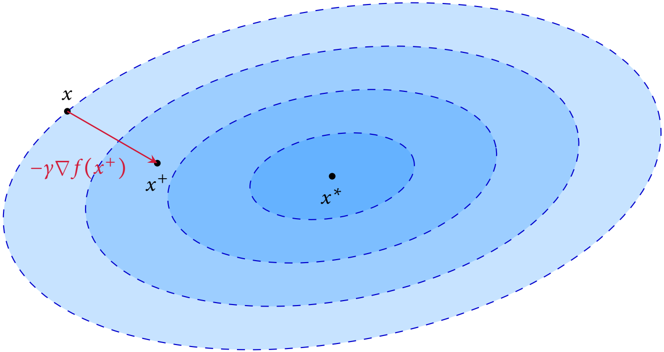


- ✓ **Potential:** $g = \nabla f$
- ✓ **Deterministic:** $\sigma_n = 0$
- ✓ **No offset:** $B_n = 0$

Methods, II: Proximal gradient

Proximal gradient [Martinet, 1970; Rockafellar, 1976]

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



✓ **Potential:**
$$g = \nabla f$$

✓ **Deterministic:**
$$\sigma_n = 0$$

⚠ **Offset:**
$$B_n = \mathcal{O}(\gamma_n)$$

Oracle 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

- **Reinforcement Learning / Control:**

  $f(x) = \mathbb{E}[F(x;\theta)]$ and only $\nabla F(x;\theta)$ can be observed for a random θ

- **Game Theory / Bandits:**

  Only $f(x)$ is observable

Stochastic first-order oracle

A ***stochastic first-order oracle (SFO)*** is a random field $G(x;\theta)$ with the following properties

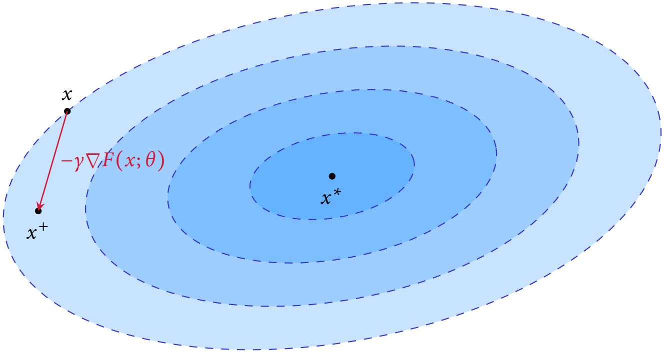
1. ***Unbiasedness:*** $\mathbb{E}_\theta[G(x;\theta)] = g(x)$
2. ***Finite variance:*** $\mathbb{E}_\theta[\|G(x;\theta) - g(x)\|^2] \leq \sigma^2$

⚠ **Special case:** if $g = \nabla f$, then G is called a ***stochastic gradient*** of f

Methods, III: Stochastic gradient descent

Stochastic gradient descent [Robbins & Monro, 1951; Ljung, 1978; Bertsekas & Tsitsiklis, 2000]

$$X_{n+1} = X_n - \gamma_n \nabla F(X_n; \theta_n) \tag{SGD}$$



✓ **Potential:**
$g = \nabla f$

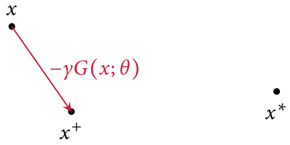
⚠ **Stochastic:**
$\sigma_n = \mathcal{O}(1)$

✓ **No offset:**
$B_n = 0$

Methods, IV: Robbins-Monro

The Robbins-Monro algorithm [Robbins & Monro, 1951]

$$X_{n+1} = X_n - \gamma_n G(X_n; \theta_n) \tag{RM}$$



⚠ **Non-potential:** general g

⚠ **Stochastic:** $\sigma_n = \mathcal{O}(1)$

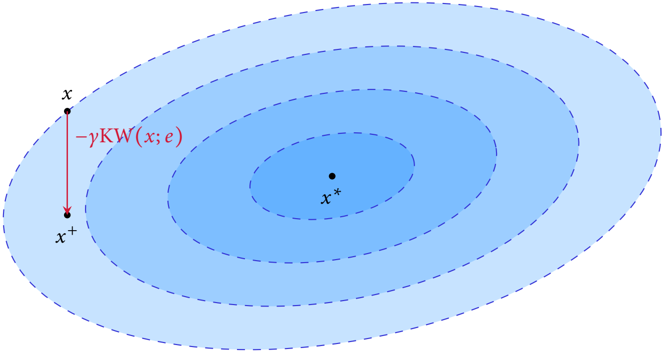
- **No offset:** $B_n = 0$

Methods, V: Kiefer-Wolfowitz

The Kiefer-Wolfowitz algorithm [Kiefer & Wolfowitz, 1952]

$$X_{n+1} = X_n - \gamma_n \frac{f(X_n + \delta_n e_n) - f(X_n - \delta_n e_n)}{2\delta_n} e_n \tag{KW}$$

where $e_n \sim \mathrm{unif}\{e_1, \dots, e_d\}$ is a ***random direction*** and δ_n is the ***width*** of the finite difference quotient



- **Potential:**
  $g = \nabla f$
- ⚠ **Stochastic:**
  $\sigma_n = \mathcal{O}(1)$
- ⚠ **Offset:**
  $B_n = \mathcal{O}(\delta_n)$

From algorithms to flows

Characteristic property of SA schemes

$$\frac{X_{n+1} - X_n}{\gamma_n} = -g(X_n) + Z_n \approx -g(X_n) \quad \text{“on average”}$$

Mean dynamics

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

Asymptotic pseudotrajectories

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

The ODE method

[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:***

$$\lim_{t\to\infty} \sup_{0\le h\le T} \|X(t+h) - \phi_h(X(t))\| = 0 \tag{APT}$$

where $\phi_t(x)$ denotes the position at time t of an orbit of (MD) starting at x

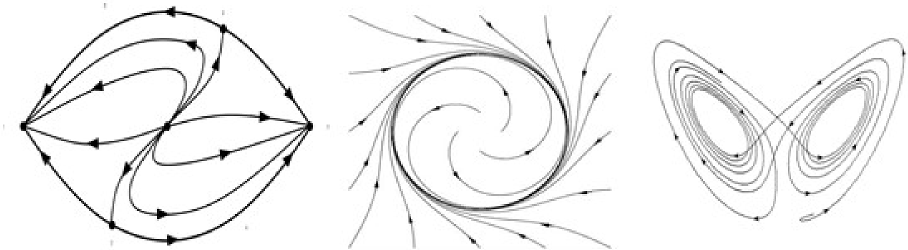
- ***Long run:*** $X(t)$ tracks (MD) with arbitrary accuracy over windows of arbitrary length

Benaïm & Hirsch, 1995, 1996; Benaïm, 1999

Internally chain transitive sets

Stationary sets: an assorted zoology

- **Invariant:** image of $\mathcal{S}$ under (MD) = $\mathcal{S}$ # $\phi_t(\mathcal{S}) = \mathcal{S}$ for all t
- **Attractor:** invariant + attracts uniformly all nearby orbits of (MD)
- **Internally chain transitive:** invariant + contains no proper attractors

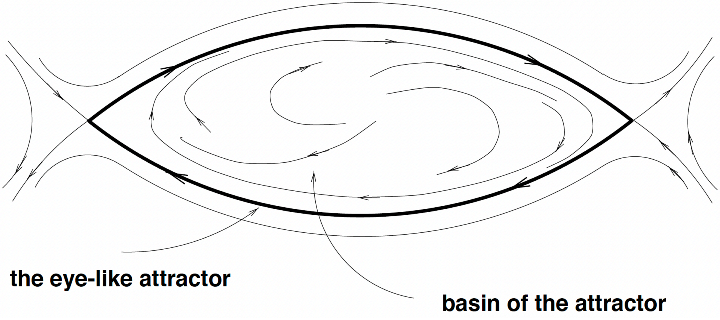


a) Stable and unstable fixed points   b) A limit cycle   c) A chaotic Lorenz attractor

Internally chain transitive sets

Stationary sets: an assorted zoology

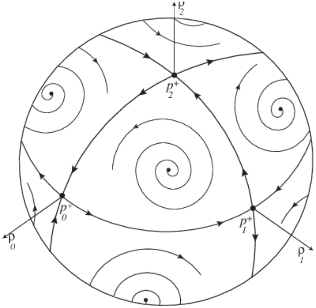
- **Invariant:** image of $\mathcal{S}$ under (MD) = $\mathcal{S}$ # $\phi_t(\mathcal{S}) = \mathcal{S}$ for all t
- **Attractor:** invariant + attracts uniformly all nearby orbits of (MD)
- **Internally chain transitive:** invariant + contains no proper attractors



Internally chain transitive sets

Stationary sets: an assorted zoology

- **Invariant:** image of $\mathcal{S}$ under (MD) = $\mathcal{S}$ $\quad$ # $\phi_t(\mathcal{S}) = \mathcal{S}$ for all t
- **Attractor:** invariant + attracts uniformly all nearby orbits of (MD)
- **Internally chain transitive:** invariant + contains no proper attractors



The limit set theorem

How does the long-run behavior of an APT relate to that of (MD)?

Theorem (Benaïm & Hirsch, 1996)

Let $X(t)$ be a bounded APT of (MD) *and let*

$$\mathcal{L}(X) = \{x \in \mathcal{X} : X(t_n) \to x \text{ for some } t_n \to \infty\}$$

denote the set of limit points of X. Then:

- $\mathcal{L}(X)$ *is an ICT set of* (MD)
- *If $\mathcal{S}$ is ICT, there exists some APT of* (MD) *such that $\mathcal{L}(X) = \mathcal{S}$*

The limit set theorem

How does the long-run behavior of an APT relate to that of (MD)?

Theorem (Benaïm & Hirsch, 1996)

Let $X(t)$ be a bounded APT of (MD) *and let*

$$\mathcal{L}(X) = \{x \in \mathcal{X} : X(t_n) \to x \text{ for some } t_n \to \infty\}$$

denote the set of limit points of X. Then:

- *$\mathcal{L}(X)$ is an ICT set of* (MD)
- *If $\mathcal{S}$ is ICT, there exists some APT of* (MD) *such that $\mathcal{L}(X) = \mathcal{S}$*

Two basic questions:

Q1. *When is an SA sequence an APT of* (MD)*?*

Q2. *What are the ICT sets of* (MD)*?*

Stochastic approximation criteria

Is a stochastic approximation sequence an APT of (MD)?

(A) g is ***subcoercive:***

$$\langle g(x), x \rangle \geq 0 \quad \text{for sufficiently large } x$$

(B) The noise and offset parameters of (SA) satisfy:

- $\lim_{n\to\infty} B_n = 0$
- $\sum_n \gamma_n B_n < \infty$
- $\sum_n \gamma_n^2 \sigma_n^2 < \infty$

Proposition (Benaïm & Hirsch, 1996; Hsieh et al., 2021)

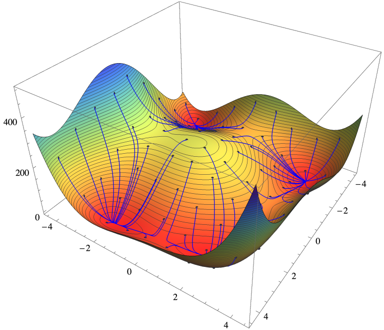
Assume: *(A) + (B)*

Then: *X_n is a bounded APT of (MD) with probability 1*

Outline

Convergence in minimization problems

- **Problem:** minimize $f(x) = \mathbb{E}_\theta[F(x;\theta)]$
- **Drift:** $g = \nabla f$
- **Key property:** $df/dt = -\|\nabla f(x(t))\|^2 \leq 0$ w/ equality iff $\nabla f(x) = 0$



Theorem (Bertsekas & Tsitsiklis, 2000; M, Hallak, Kavis & Cevher, 2020)

Assume: (A) + (B)

Then: X_n *converges with probability* 1 *to a component of* $\operatorname{crit}(f)$ *where* f *is constant.*

Are all critical points desirable?

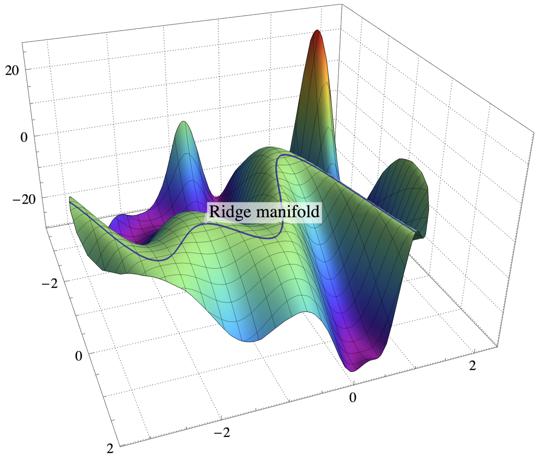


Figure: A hyperbolic ridge manifold, typical of ResNet loss landscapes [Li et al., 2018]

Are traps avoided?

Hyperbolic saddle (isolated non-minimizing critical point)

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

$\implies$ the flow is **linearly unstable** near x^*

$\implies$ convergence to x^* **unlikely**

Are traps avoided?

Hyperbolic saddle (isolated non-minimizing critical point)

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

$\Longrightarrow$ the flow is **linearly unstable** near x^*

$\Longrightarrow$ convergence to x^* **unlikely**

Theorem (Pemantle, 1990)

Assume:

- *x^* is a hyperbolic saddle point*
- $b_n = 0$
- *U_n is uniformly bounded (a.s.) and uniformly exciting*

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

- $\gamma_n \propto 1/n$

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

Escape from non-hyperbolic traps

Strict saddles

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

Escape from non-hyperbolic traps

Strict saddles

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

Theorem (Ge et al., 2015)

Given: *tolerance level* $\zeta > 0$

Assume:

- f *is bounded and satisfies* (LS)
- $\mathrm{Hess}(f(x))$ *is Lipschitz continuous*
- *for all* $x \in \mathcal{X}$: **(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*
- $b_n = 0$
- U_n *is uniformly bounded* (a.s.) *and contains a component uniformly sampled from the unit sphere*
- $\gamma_n \equiv \gamma$ *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^*

Are non-hyperbolic traps avoided almost surely?

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

Assume:

- *Conditions **(B)***
- U_n *is uniformly bounded (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: $\mathbb{P}(X_n \textit{ converges to a set of strict saddle points}) = 0$

Outline

Minimization vs. min-max optimization

In minimization problems:

- ✓ RM methods converge to the problem's critical set
- ✓ RM methods avoid spurious, non-minimizing critical manifolds

Minimization vs. min-max optimization

In minimization problems:

✓ RM methods converge to the problem's critical set

✓ RM methods avoid spurious, non-minimizing critical manifolds

Do these properties carry over to min-max optimization problems?

Minimization vs. min-max optimization

In minimization problems:

- ✓ RM methods converge to the problem's critical set
- ✓ RM methods avoid spurious, non-minimizing critical manifolds

Do these properties 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

Mean dynamics

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

✓ **Minimization problems:** (MD) is a gradient flow # $g = \nabla f$

✗ **Min-max problems:** (MD) can be arbitrarily complicated # non-potential g

Min-max dynamics

Mean dynamics

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

✓ **Minimization problems:** (MD) is a gradient flow # $g = \nabla f$

✗ **Min-max problems:** (MD) can be arbitrarily complicated # non-potential g

Theorem (Hsieh et al., 2021)

Assume:

- *Conditions **(B)***
- *U_n is uniformly bounded (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: $\mathbb{P}(X_n \text{ converges to an unstable point / periodic orbit}) = 0$

Toy example: bilinear problems

Bilinear min-max problems

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

Mean dynamics:

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

Toy example: bilinear problems

Bilinear min-max problems

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

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**

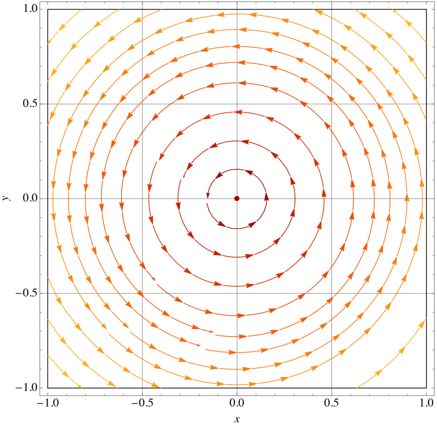
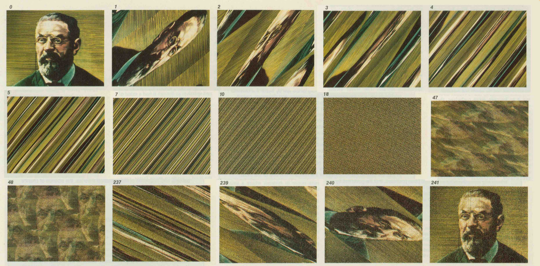


Figure: Hamiltonian flow of $f(x_1, x_2) = x_1 x_2$

Poincaré recurrence

Definition (Poincaré, 1890's)

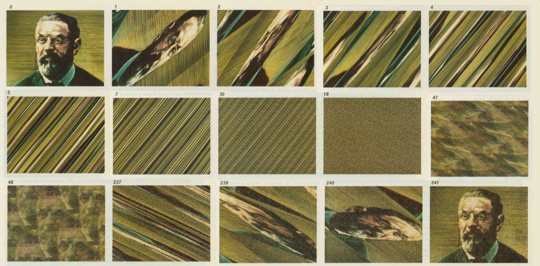
A system is **Poincaré recurrent** if almost every orbit returns ***infinitely close*** to its starting point ***infinitely often***



Poincaré recurrence

Definition (Poincaré, 1890's)

A system is **Poincaré recurrent** if almost every orbit returns ***infinitely close*** to its starting point ***infinitely often***



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

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

The stochastic case

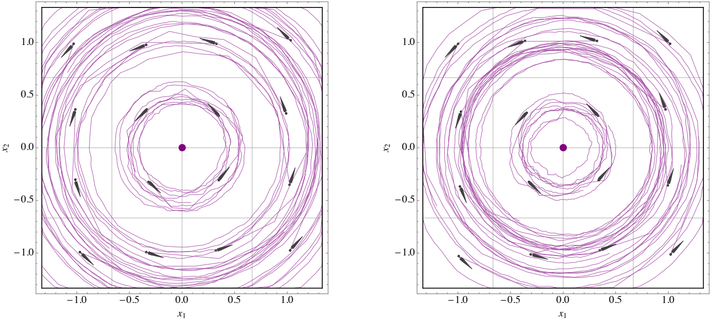


Figure: Behavior of gradient and extra-gradient methods with stochastic feedback

Under (A) + (B), first-order methods converge to a (random) periodic orbit

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

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 $g \in \mathcal{V}$, 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)*

TLDR: non-attracting periodic orbits are **non-generic** (they occur negligibly often)

Convergence to attractors

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

Theorem (Hsieh et al., 2021)

Assume: *$\mathcal{S}$ is an attractor of* (MD) + *Conditions* ***(B)***

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

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

Minimization vs. min-max optimization

Qualitatively similar landscape (??)

- Avoidance of strict saddles $\leftrightsquigarrow$ avoidance of unstable invariant sets

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

Minimization vs. min-max optimization

Qualitatively similar landscape (??)

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

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

Non-gradient problems may have spurious invariant sets!

"spurious" $\implies$ contains no critical points

Almost bilinear games

Consider 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)

✗ **All RM algorithms attracted to spurious limit cycle from almost all initial conditions**

Hsieh et al., 2021

Spurious attractors in almost bilinear games

RM algorithms converge to a spurious limit cycle with **no critical points**

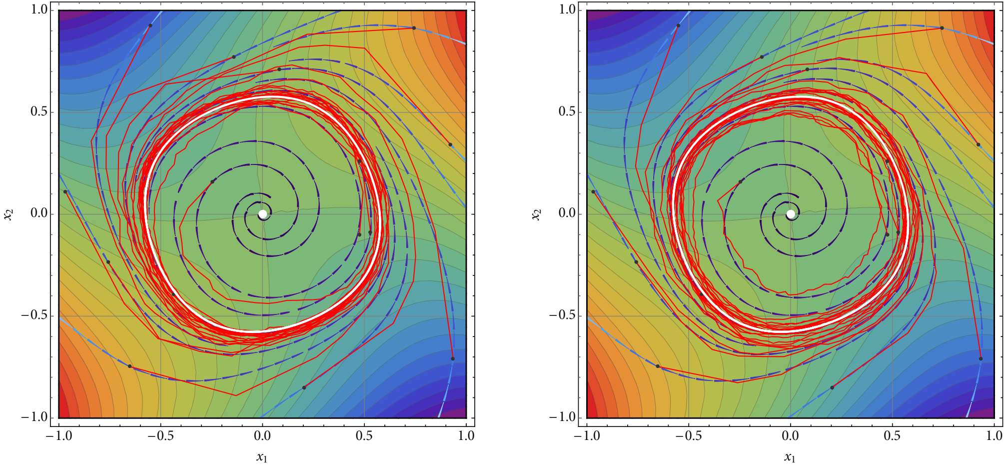


Figure: Convergence to a spurious attractor. Left: stochastic gradient descent; right: stochastic extra-gradient

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 critical point near the origin**
- Stable under (MD), but **not a local min-max**
- **Two isolated periodic orbits:**
  - One **unstable**, shielding critical point, but small
  - One **stable**, attracts all trajectories of (MD) outside small basin

Hsieh et al., 2021

Forsaken solutions in almost bilinear games

With high probability, all Robbins-Monro (RM) algorithms forsake the game's unique (local) equilibrium

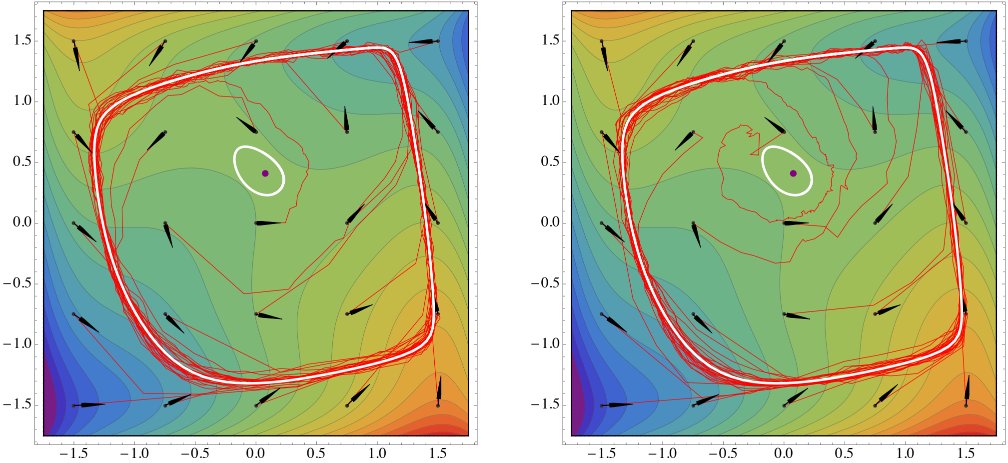


Figure: Convergence to a spurious attractor. Left: stochastic gradient descent; right: stochastic extra-gradient

Conclusions

Minimization and min-max optimization problems are fundamentally different:

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

Conclusions

Minimization and min-max optimization problems are fundamentally different:

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

Many open questions:

- What about second-order methods?
- Applications to finite games (where bilinear games are no longer fragile)?
- Which equilibria are stable under first-order methods for learning in games?
- ...

References I

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.

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

Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627-642, 2000.

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.

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.

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.

Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462-466, 1952.

Kupka, I. Contribution à la théorie des champs génériques. *Contributions to Differential Equations*, 2:457-484, 1963.

References II

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

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.

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

Ljung, L. Strong convergence of a stochastic approximation algorithm. *Annals of Statistics*, 6(3):680-696, 1978.

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

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

Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400-407, 1951.

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

Smale, S. Stable manifolds for differential equations and diffeomorphisms. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 17 (1-2):97-116, 1963.