Nonsmooth, nonconvex optimization, implications for deep-learning

S. Gerchinovitz[1], F. Malgouyres[1], **E. Pauwels**[2] & N. Thome[3]

[1] Institut de Mathématiques de Toulouse, Université Toulouse 3 Paul Sabatier.
[2] Institut de recherche en informatique de Toulouse, Université Toulouse 3 Paul Sabatier.
[3] Centre d'étude et de recherche en informatique et communication, Conservatoire national des arts et métiers.

1 Fev. 2021

Acknowledgements

Jérôme Bolte (Toulouse School of Economics):

Plan

1. Introduction
2. Convergence to local minima for Morse-Functions
3. On the structure of deep learning training loss
4. Convergence to critical points for tame functions
5. Approaching critical point with noise
6. Extensions to nonsmooth settings

Training a deep network

Finite dimensional optimization problem

$$\min_{\mathbf{w},\mathbf{b}} \frac{1}{n}\sum_{i=1}^{n} L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$$

- $((x_i, y_i))_{i=1}^n$: training set in $\mathcal{X} \times \mathcal{Y}$.
- L loss.
- $(\mathbf{w}, \mathbf{b})$ network parameters (linear maps and offset).
- $f_{\mathbf{w},\mathbf{b}} : \mathcal{X} \mapsto \mathcal{Y}$ neural network.

Training a deep network

Finite dimensional optimization problem

$$\min_{\mathbf{w},\mathbf{b}} \frac{1}{n} \sum_{i=1}^{n} L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$$

- $((x_i, y_i))_{i=1}^{n}$: training set in $\mathcal{X} \times \mathcal{Y}$.
- L loss.
- $(\mathbf{w}, \mathbf{b})$ network parameters (linear maps and offset).
- $f_{\mathbf{w},\mathbf{b}} : \mathcal{X} \mapsto \mathcal{Y}$ neural network.

Notations:

$$\begin{aligned} F \colon \mathbb{R}^p &\mapsto \mathbb{R} \\ \theta &\mapsto \frac{1}{n} \sum_{i=1}^{n} l_i(\theta) \end{aligned} \tag{P}$$

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \ldots n$.

Main question

$$\min_{\theta \in \mathbb{R}^p} \qquad F(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_i(\theta) \tag{1}$$

Compositional structure of deep network: Computing a (stochastic)-gradient of F has a cost comparable to evaluating F.

Main question

$$\min_{\theta \in \mathbb{R}^p} \qquad F(\theta) = \frac{1}{n} \sum_{i=1}^{n} l_i(\theta) \tag{1}$$

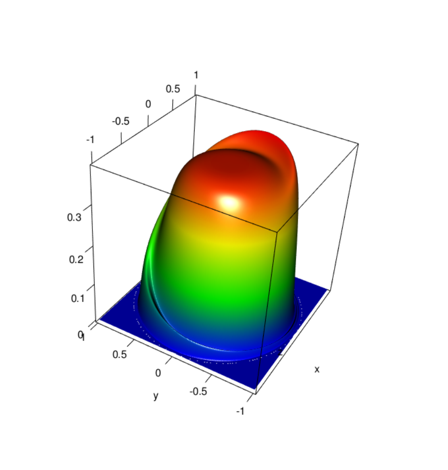
Compositional structure of deep network: Computing a (stochastic)-gradient of F has a cost comparable to evaluating F.

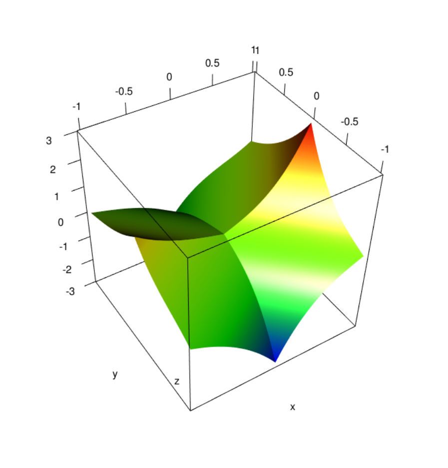
Deep nets are trained with variants of gradient descent.

$$\begin{aligned} \theta_{k+1} &= \theta_k - \alpha_k \nabla F(\theta_k) \\ \alpha_k &> 0 \end{aligned} \tag{GD}$$

Long term behaviour for this recursion?

Non convexity, non smoothness





Roadmap: longterm behavior of gradient descent

Main difficulty: The objective term is not convex and may be not smooth.

Roadmap: longterm behavior of gradient descent

Main difficulty: The objective term is not convex and may be not smooth.

Long history in mathematics.
Foundations from two fields:

- Smooth dynamical systems Poincaré, Hadamard, Lyapunov, Hirsch, Smale, Shub, Hartman, Grobman, Thom . . .
- Favorable geometric structure of F (semi-algebraic/tame geometry). Łojasiewicz, Hironaka, Grothendiek, van den Dries, Shiota. . .

Roadmap: longterm behavior of gradient descent

Main difficulty: The objective term is not convex and may be not smooth.

Long history in mathematics.
Foundations from two fields:

- Smooth dynamical systems Poincaré, Hadamard, Lyapunov, Hirsch, Smale, Shub, Hartman, Grobman, Thom ...
- Favorable geometric structure of F (semi-algebraic/tame geometry). Łojasiewicz, Hironaka, Grothendiek, van den Dries, Shiota...

Program for today:

- Convergence to second order critical point for Morse functions (60's).
- Favorable structure of deep learning landscapes (60's).
- Convergence to critical points under Łojasiewicz assumption (60's).
- Approaching critical point with stochastic subgradient (ODE method, 70's).

Plan

1. Introduction
2. Convergence to local minima for Morse-Functions
3. On the structure of deep learning training loss
4. Convergence to critical points for tame functions
5. Approaching critical point with noise
6. Extensions to nonsmooth settings

Main idea

Smooth dynamical systems

$$\dot{x} = S(x) \text{ (flow)}$$
$$x_{k+1} = T(x_k) \text{ (discrete)}$$

$S, T\colon \mathbb{R}^p \mapsto \mathbb{R}^p$, local diffeomorphisms (differentiable with differentiable inverse).

Main idea

Smooth dynamical systems

$$\dot{x} = S(x) \text{ (flow)}$$
$$x_{k+1} = T(x_k) \text{ (discrete)}$$

$S, T \colon \mathbb{R}^p \mapsto \mathbb{R}^p$, local diffeomorphisms (differentiable with differentiable inverse).

Long term behaviour: convergence, bifurcation, chaos . . .

Main idea

Smooth dynamical systems

$$\dot{x} = S(x) \text{ (flow)}$$
$$x_{k+1} = T(x_k) \text{ (discrete)}$$

$S, T\colon \mathbb{R}^p \mapsto \mathbb{R}^p$, local diffeomorphisms (differentiable with differentiable inverse).

Long term behaviour: convergence, bifurcation, chaos . . .

Generic results: Nonlinear dynamics behave similarly as their linear approximations.

Main idea

Smooth dynamical systems

$$\dot{x} = S(x) \text{ (flow)}$$
$$x_{k+1} = T(x_k) \text{ (discrete)}$$

$S, T \colon \mathbb{R}^p \mapsto \mathbb{R}^p$, local diffeomorphisms (differentiable with differentiable inverse).

Long term behaviour: convergence, bifurcation, chaos ...

Generic results: Nonlinear dynamics behave similarly as their linear approximations.

Lemma: Let F be $\mathcal{C}^2$, if ∇F is L-Lipschitz, then the gradient mapping $T \colon x \to x - \alpha \nabla F(x)$ is a diffeomorphism $0 < \alpha < 1/L$.

The gradient mapping is a diffeormorphism

Constructive proof:

- For any $x \in \mathbb{R}^p$, the Jacobian $\nabla T = I - \alpha \nabla^2 F(x)$ is positive definite (exercise). We have a local diffeomorphism as a consequence of implicit function theorem.

The gradient mapping is a diffeormorphism

Constructive proof:

- For any $x \in \mathbb{R}^p$, the Jacobian $\nabla T = I - \alpha \nabla^2 F(x)$ is positive definite (exercise). We have a local diffeomorphism as a consequence of implicit function theorem.
- Explicitely, for any $x, y \in \mathbb{R}^p$ such that $T(x) = T(y)$,

$$\|x - y\| = \alpha \|\nabla F(x) - \nabla F(y)\| < L\alpha \|x - y\|, \qquad \text{hence } x = y.$$

The gradient mapping is a diffeormorphism

Constructive proof:

- For any $x \in \mathbb{R}^p$, the Jacobian $\nabla T = I - \alpha \nabla^2 F(x)$ is positive definite (exercise). We have a local diffeomorphism as a consequence of implicit function theorem.
- Explicitely, for any $x, y \in \mathbb{R}^p$ such that $T(x) = T(y)$,

$$\|x - y\| = \alpha \|\nabla F(x) - \nabla F(y)\| < L\alpha \|x - y\|, \qquad \text{hence } x = y.$$

- Explicit inverse: solution to the strictly convex problem,

$$\text{prox}_{-\alpha F} \colon z \mapsto \arg\min_{y \in \mathbb{R}^p} -\alpha F(y) + \frac{1}{2}\|y - z\|_2^2$$
$$x = \text{prox}_{-\alpha F}(z) \Leftrightarrow z = x - \alpha \nabla F(x).$$

Quizz: linear isomorphisms

Convergence to 0?

Quizz: linear isomorphisms

Convergence to 0?

- $x_0 \in \mathbb{R}$, $a \in \mathbb{C}$, $a \neq 0$, $x_{k+1} = a x_k$.

Quizz: linear isomorphisms

Convergence to 0?

- $x_0 \in \mathbb{R}$, $a \in \mathbb{C}$, $a \neq 0$, $x_{k+1} = ax_k$.
- $x_0 \in \mathbb{R}^p$, $D \in \mathbb{R}^{p \times p}$, diagonal, no zero entry, $x_{k+1} = Dx_k$.

Quizz: linear isomorphisms

Convergence to 0?

- $x_0 \in \mathbb{R}$, $a \in \mathbb{C}$, $a \neq 0$, $x_{k+1} = a x_k$.
- $x_0 \in \mathbb{R}^p$, $D \in \mathbb{R}^{p \times p}$, diagonal, no zero entry, $x_{k+1} = D x_k$.
- $x_0 \in \mathbb{R}^p$, $M \in \mathbb{R}^{p \times p}$ diagonalisable over $\mathbb{C}$, $x_{k+1} = M x_k$.

Quizz: linear isomorphisms

Convergence to 0?

- $x_0 \in \mathbb{R}$, $a \in \mathbb{C}$, $a \neq 0$, $x_{k+1} = ax_k$.
- $x_0 \in \mathbb{R}^p$, $D \in \mathbb{R}^{p\times p}$, diagonal, no zero entry, $x_{k+1} = Dx_k$.
- $x_0 \in \mathbb{R}^p$, $M \in \mathbb{R}^{p\times p}$ diagonalisable over $\mathbb{C}$, $x_{k+1} = Mx_k$.

Symmetric real matrix: If $M \in \mathbb{R}^{p\times p}$, no eigenvalue such that $|\lambda| = 1$, one can set

$$\mathbb{R}^p = E_s \oplus E_u$$

- E_s is the stable space of M:
  - all x such that $M^k x \underset{k\to\infty}{\to} 0$.
  - eigenspace corresponding to eigenvalues $|\lambda| < 1$.

Quizz: linear isomorphisms

Convergence to 0?

- $x_0 \in \mathbb{R}$, $a \in \mathbb{C}$, $a \neq 0$, $x_{k+1} = ax_k$.
- $x_0 \in \mathbb{R}^p$, $D \in \mathbb{R}^{p \times p}$, diagonal, no zero entry, $x_{k+1} = Dx_k$.
- $x_0 \in \mathbb{R}^p$, $M \in \mathbb{R}^{p \times p}$ diagonalisable over $\mathbb{C}$, $x_{k+1} = Mx_k$.

Symmetric real matrix: If $M \in \mathbb{R}^{p \times p}$, no eigenvalue such that $|\lambda| = 1$, one can set

$$\mathbb{R}^p = E_s \oplus E_u$$

- E_s is the stable space of M:
  - all x such that $M^k x \underset{k \to \infty}{\to} 0$.
  - eigenspace corresponding to eigenvalues $|\lambda| < 1$.
- E_u is the unstable space of M:
  - all x such that $M^{-k} x \underset{k \to \infty}{\to} 0$.
  - eigenspace corresponding to eigenvalues $|\lambda| > 1$.

Quizz: linear isomorphisms

Convergence to 0?

- $x_0 \in \mathbb{R}$, $a \in \mathbb{C}$, $a \neq 0$, $x_{k+1} = ax_k$.
- $x_0 \in \mathbb{R}^p$, $D \in \mathbb{R}^{p \times p}$, diagonal, no zero entry, $x_{k+1} = Dx_k$.
- $x_0 \in \mathbb{R}^p$, $M \in \mathbb{R}^{p \times p}$ diagonalisable over $\mathbb{C}$, $x_{k+1} = Mx_k$.

Symmetric real matrix: If $M \in \mathbb{R}^{p \times p}$, no eigenvalue such that $|\lambda| = 1$, one can set

$$\mathbb{R}^p = E_s \oplus E_u$$

- E_s is the stable space of M:
  - all x such that $M^k x \underset{k \to \infty}{\to} 0$.
  - eigenspace corresponding to eigenvalues $|\lambda| < 1$.
- E_u is the unstable space of M:
  - all x such that $M^{-k} x \underset{k \to \infty}{\to} 0$.
  - eigenspace corresponding to eigenvalues $|\lambda| > 1$.

If $\dim(E_u) > 0$, then the divergence behaviour is generic (for almost every x).

Quizz: linear isomorphisms

Convergence to 0?

- $x_0 \in \mathbb{R}$, $a \in \mathbb{C}$, $a \neq 0$, $x_{k+1} = ax_k$.
- $x_0 \in \mathbb{R}^p$, $D \in \mathbb{R}^{p\times p}$, diagonal, no zero entry, $x_{k+1} = Dx_k$.
- $x_0 \in \mathbb{R}^p$, $M \in \mathbb{R}^{p\times p}$ diagonalisable over $\mathbb{C}$, $x_{k+1} = Mx_k$.

Symmetric real matrix: If $M \in \mathbb{R}^{p\times p}$, no eigenvalue such that $|\lambda| = 1$, one can set

$$\mathbb{R}^p = E_s \oplus E_u$$

- E_s is the stable space of M:
  - all x such that $M^k x \underset{k\to\infty}{\to} 0$.
  - eigenspace corresponding to eigenvalues $|\lambda| < 1$.
- E_u is the unstable space of M:
  - all x such that $M^{-k} x \underset{k\to\infty}{\to} 0$.
  - eigenspace corresponding to eigenvalues $|\lambda| > 1$.

If $\dim(E_u) > 0$, then the divergence behaviour is generic (for almost every x).

Extension to any square matrix using Jordan normal form.

Stable manifold theorem

Idea dates back to Hadamard, Lyapunov and Perron. This is a difficult result.

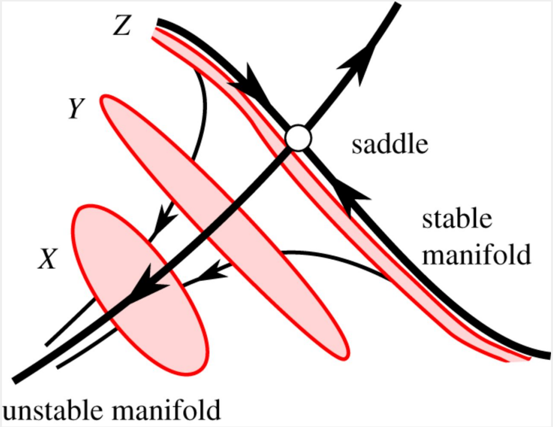
Theorem (*e.g.* Schub's book [S1987]): Let $T\colon \mathbb{R}^p \to \mathbb{R}^p$ be a local diffeomorphism $\bar{x}$ a fixed point of T such that $\nabla T(\bar{x})$ does not have any eigenvalue on the unit circle and at least one eigenvalue of modulus > 1.

Then there exists a neighborhood U of $\bar{x}$ such that

$$W^s(T,\bar{x}) = \{x_0 \in U, T^n(x_0) \to \bar{x}, n \to \infty\},$$
$$W^u(T,\bar{x}) = \{x_0 \in U, T^n(x_0) \to \bar{x}, n \to -\infty\},$$

are differentiable manifolds tangent to the stable and unstable spaces of $\nabla T(\bar{x})$. In particular, $W^s(T,\bar{x})$ has dimension $< n$.

With a picture



Obayashi *et al.* (2016). Formation mechanism of a basin of attraction for passive dynamic walking induced by intrinsic hyperbolicity. Proceedings of the Royal Society A.

Convergence to local minima on Morse functions

Assume that $F\colon \mathbb{R}^p \mapsto \mathbb{R}$ is $\mathcal{C}^2$, with L-lipschitz gradient. Assume that $\bar{x} \in \mathbb{R}^p$ satisfies.

$$\begin{aligned}
&\nabla F(\bar{x}) = 0 \\
&\nabla^2 F(\bar{x}) \qquad \text{has no null eigenvalue} \\
&\nabla^2 F(\bar{x}) \qquad \text{has at least one strictly negative eigenvalue.}
\end{aligned}$$

Assume that if x_0 is taken randomly ($\ll$ Lebesgue, *e.g.* Gaussian) and $(x_k)_{k\in\mathbb{N}}$ is given by gradient descent starting at x_0 with $\alpha < 1/L$. Then with respect to the random choice of the initialization.

$$\mathbb{P}\left[x_k \to \bar{x}\right] = 0$$

Convergence to local minima on Morse functions

Assume that $F\colon \mathbb{R}^p \mapsto \mathbb{R}$ is $\mathcal{C}^2$, with L-lipschitz gradient. Assume that $\bar{x} \in \mathbb{R}^p$ satisfies.

$$\begin{aligned} &\nabla F(\bar{x}) = 0 \\ &\nabla^2 F(\bar{x}) \qquad \text{has no null eigenvalue} \\ &\nabla^2 F(\bar{x}) \qquad \text{has at least one strictly negative eigenvalue.} \end{aligned}$$

Assume that if x_0 is taken randomly ($\ll$ Lebesgue, *e.g.* Gaussian) and $(x_k)_{k\in\mathbb{N}}$ is given by gradient descent starting at x_0 with $\alpha < 1/L$. Then with respect to the random choice of the initialization.

$$\mathbb{P}[x_k \to \bar{x}] = 0$$

Proof: The gradient mapping $T\colon x \mapsto x - \alpha\nabla F(x)$ satisies hypotheses of the stable manifold theorem. If $x_k \to \bar{x}$, this means that after a finite number of steps K, $x_k \in U$ for all $k \geq K$ which implies that $x_k \in W^s(T,\bar{x})$ for all $k \geq K$. Hence

$$\left\{x_0 \in \mathbb{R}^p,\, T^k(x_0) \underset{k\to\infty}{\to} \bar{x}\right\} = \cup_{K\in\mathbb{N}} T^{-K}(W^s(T,\bar{x}))$$

$W^s(T,\bar{x})$ has Lebesgue measure 0, images of zero measure sets by diffeomorphism have measure 0 and countable union of measure 0 set is of measure 0.

Extension: Gradient Descent Only Converges to Minimizers

Lee, Simchowitz, Jordan, Recht [LSJR2016]: drop the full rank assumption on the Hessian.

Plan

1 Introduction

2 Convergence to local minima for Morse-Functions

3 On the structure of deep learning training loss

4 Convergence to critical points for tame functions

5 Approaching critical point with noise

6 Extensions to nonsmooth settings

Deep learning training loss

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \ldots n$.

$$\begin{aligned} F \colon \mathbb{R}^p &\mapsto \mathbb{R} \\ \theta &\mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \end{aligned} \tag{P}$$

Consider $L\colon (\hat{y}, y) = (\hat{y} - y)^2$ or $L\colon (\hat{y}, y) = |\hat{y} - y|$ and a Relu network: activation function is the positive part $\max(0, \cdot)$.

Deep learning training loss

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$F \colon \mathbb{R}^p \mapsto \mathbb{R}$$
$$\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \tag{P}$$

Consider $L\colon (\hat{y}, y) = (\hat{y} - y)^2$ or $L\colon (\hat{y}, y) = |\hat{y} - y|$ and a Relu network: activation function is the positive part $\max(0, \cdot)$.

Then F has a highly favorable structure: it is .

Deep learning training loss

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$F \colon \mathbb{R}^p \mapsto \mathbb{R}$$
$$\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \qquad \text{(P)}$$

Consider $L\colon (\hat{y}, y) = (\hat{y} - y)^2$ or $L\colon (\hat{y}, y) = |\hat{y} - y|$ and a Relu network: activation function is the positive part $\max(0, \cdot)$.

Then F has a highly favorable structure: it is “piecewise” polynomial.

Semi-algebraic sets and functions (SA)

SA set in $\mathbb{R}^p$: Union of finitely many solution sets of systems of the form.

$$\{x \in \mathbb{R}^p,\, P(x) = 0, Q_1(x) > 0, \ldots, Q_l(x) > 0\}$$

for some polynomials functions $P, Q_1, \ldots Q_l$ over $\mathbb{R}^p$.

Semi-algebraic sets and functions (SA)

SA set in $\mathbb{R}^p$: Union of finitely many solution sets of systems of the form.

$$\{x \in \mathbb{R}^p,\, P(x) = 0, Q_1(x) > 0, \ldots, Q_l(x) > 0\}$$

for some polynomials functions $P, Q_1, \ldots Q_l$ over $\mathbb{R}^p$.

SA map $\mathbb{R}^p \to \mathbb{R}^{p'}$: A map $F \colon \mathbb{R}^p \mapsto \mathbb{R}^{p'}$ whose graph

$$\mathrm{graph}_f = \left\{(x, z) \in \mathbb{R}^{p+p'}, z = F(x)\right\}$$

is SA.

Semi-algebraic sets and functions (SA)

SA set in $\mathbb{R}^p$: Union of finitely many solution sets of systems of the form.

$$\{x \in \mathbb{R}^p,\, P(x) = 0, Q_1(x) > 0, \ldots, Q_l(x) > 0\}$$

for some polynomials functions $P, Q_1, \ldots Q_l$ over $\mathbb{R}^p$.

SA map $\mathbb{R}^p \to \mathbb{R}^{p'}$: A map $F \colon \mathbb{R}^p \mapsto \mathbb{R}^{p'}$ whose graph

$$\mathrm{graph}_f = \left\{(x, z) \in \mathbb{R}^{p+p'}, z = F(x)\right\}$$

is SA.

SA set in $\mathbb{R}$: Union of finitely many intervals.

Semi-algebraic sets and functions (SA)

SA set in $\mathbb{R}^p$: Union of finitely many solution sets of systems of the form.

$$\{x \in \mathbb{R}^p, \, P(x) = 0, Q_1(x) > 0, \dots, Q_l(x) > 0\}$$

for some polynomials functions $P, Q_1, \dots Q_l$ over $\mathbb{R}^p$.

SA map $\mathbb{R}^p \to \mathbb{R}^{p'}$: A map $F \colon \mathbb{R}^p \mapsto \mathbb{R}^{p'}$ whose graph

$$\mathrm{graph}_f = \left\{(x, z) \in \mathbb{R}^{p+p'}, z = F(x)\right\}$$

is SA.

SA set in $\mathbb{R}$: Union of finitely many intervals.

Properties: Closed under union, intersection, complementation, product.

SA functions: examples

- Polynomials: $P(x)$
- "Piecewise polynomials": $P(x)$ if $x > 0$, $Q(x)$ otherwise
- Rational functions: $1/P(x)$
- Rational powers: $P(x)^q$, $q \in \mathbb{Q}$.
- Absolute value: $\|\cdot\|_1$.
- $\|\cdot\|_0$ pseudo-norm.
- Rank of matrices
- ...

Tarski-Seidenberg Theorem

Theorem: Let $A \subset \mathbb{R}^{p+1}$ be a SA and π be the projection on the first p coordinates, then $\pi(A)$ is SA:

$$\{x \in \mathbb{R}^p, \exists y \in \mathbb{R}, (x,y) \in A\} \qquad \text{is SA.}$$

It can be described by finitely many polynomial inequalities in x only.

Tarski-Seidenberg Theorem

Theorem: Let $A \subset \mathbb{R}^{p+1}$ be a SA and π be the projection on the first p coordinates, then $\pi(A)$ is SA:

$$\{x \in \mathbb{R}^p, \exists y \in \mathbb{R}, (x, y) \in A\} \qquad \text{is SA.}$$

It can be described by finitely many polynomial inequalities in x only.

Eliminate existential quantifier.

Tarski-Seidenberg Theorem

Theorem: Let $A \subset \mathbb{R}^{p+1}$ be a SA and π be the projection on the first p coordinates, then $\pi(A)$ is SA:

$$\{x \in \mathbb{R}^p, \exists y \in \mathbb{R}, (x, y) \in A\} \qquad \text{is SA.}$$

It can be described by finitely many polynomial inequalities in x only.

Eliminate existential quantifier. Eliminate also universal quantifier $\pi(A)^c$ is SA

$$\pi(A)^c = \{x \in \mathbb{R}^p, \forall y \in \mathbb{R}, (x, y) \in A^c\}$$

Recursively, eliminate a finite number of quantifier on variables.

Tarski-Seidenberg Theorem

Theorem: Let $A \subset \mathbb{R}^{p+1}$ be a SA and π be the projection on the first p coordinates, then $\pi(A)$ is SA:

$$\{x \in \mathbb{R}^p, \exists y \in \mathbb{R}, (x,y) \in A\} \qquad \text{is SA.}$$

It can be described by finitely many polynomial inequalities in x only.

Eliminate existential quantifier. Eliminate also universal quantifier $\pi(A)^c$ is SA

$$\pi(A)^c = \{x \in \mathbb{R}^p, \forall y \in \mathbb{R}, (x,y) \in A^c\}$$

Recursively, eliminate a finite number of quantifier on variables.

Consequences: Any set or function described with a first order formula, with real variables, SA objects, addition, multiplication, equality and inequality signs, is SA.

Tarski-Seidenberg Theorem

Theorem: Let $A \subset \mathbb{R}^{p+1}$ be a SA and π be the projection on the first p coordinates, then $\pi(A)$ is SA:

$$\{x \in \mathbb{R}^p, \exists y \in \mathbb{R}, (x,y) \in A\} \qquad \text{is SA.}$$

It can be described by finitely many polynomial inequalities in x only.

Eliminate existential quantifier. Eliminate also universal quantifier $\pi(A)^c$ is SA

$$\pi(A)^c = \{x \in \mathbb{R}^p, \forall y \in \mathbb{R}, (x,y) \in A^c\}$$

Recursively, eliminate a finite number of quantifier on variables.

Consequences: Any set or function described with a first order formula, with real variables, SA objects, addition, multiplication, equality and inequality signs, is SA.

- The image and pre-image of SA maps.
- The interior, closure and boundary of SA sets.
- The derivatives of a differentiable SA functions.
- The set of non continuity, non differentiability points of SA functions.

Tarski-Seidenberg Theorem

Theorem: Let $A \subset \mathbb{R}^{p+1}$ be a SA and π be the projection on the first p coordinates, then $\pi(A)$ is SA:

$$\{x \in \mathbb{R}^p, \exists y \in \mathbb{R}, (x,y) \in A\} \qquad \text{is SA.}$$

It can be described by finitely many polynomial inequalities in x only.

Eliminate existential quantifier. Eliminate also universal quantifier $\pi(A)^c$ is SA

$$\pi(A)^c = \{x \in \mathbb{R}^p, \forall y \in \mathbb{R}, (x,y) \in A^c\}$$

Recursively, eliminate a finite number of quantifier on variables.

Consequences: Any set or function described with a first order formula, with real variables, SA objects, addition, multiplication, equality and inequality signs, is SA.

- The image and pre-image of SA maps.
- The interior, closure and boundary of SA sets.
- The derivatives of a differentiable SA functions.
- The set of non continuity, non differentiability points of SA functions.

For more: Michel Coste's Introduction to semi-algebraic geometry [C2002].

Semi-algebraic sets and functions are “not pathological”

Semi-algebraic sets and functions are “not pathological”

Univariate SA functions:

- Have left and right limits.
- Are continuous except at finitely many points.
- Are C^k except at finitely many points.
- Are nicely structured (piecewise constant, increasing or decreasing).

Semi-algebraic sets and functions are “not pathological”

Univariate SA functions:

- Have left and right limits.
- Are continuous except at finitely many points.
- Are C^k except at finitely many points.
- Are nicely structured (piecewise constant, increasing or decreasing).

Higher dimension:

- SA functions are C^k on a dense open set.
- True for all restrictions of SA functions to SA sets.
- SA sets have well defined integral dimension.
- Full measure and dense open are equivalent.
- Stratification . . .

Example: Morse-Sard theorem

Theorem: Let $f : \mathbb{R} \mapsto \mathbb{R}$ be SA differentiable, then the set of critical values of f is finite:

$$\mathrm{crit}_f = f\left(\left\{x \in \mathbb{R}, f'(x) = 0\right\}\right)$$

Example: Morse-Sard theorem

Theorem: Let $f : \mathbb{R} \mapsto \mathbb{R}$ be SA differentiable, then the set of critical values of f is finite:

$$\mathrm{crit}_f = f\left(\left\{x \in \mathbb{R}, f'(x) = 0\right\}\right)$$

Proof: Setting $C = \{x \in \mathbb{R}, f'(x) = 0\}$, f' is SA, C is semialgebraic and there is $m \in \mathbb{N}$ and intervals $J_1, \ldots, J_m$ such that $C = \cup_{i=1}^m J_i$.

Example: Morse-Sard theorem

Theorem: Let $f : \mathbb{R} \mapsto \mathbb{R}$ be SA differentiable, then the set of critical values of f is finite:

$$\mathrm{crit}_f = f\left(\left\{x \in \mathbb{R}, f'(x) = 0\right\}\right)$$

Proof: Setting $C = \{x \in \mathbb{R}, f'(x) = 0\}$, f' is SA, C is semialgebraic and there is $m \in \mathbb{N}$ and intervals $J_1, \ldots, J_m$ such that $C = \cup_{i=1}^m J_i$.

For $i = 1, \ldots m$, J_i is an interval, $f' = 0$ is continuous on J_i, for any $a, b \in J_i$, we have

$$f(b) - f(a) = \int_a^b f'(t)dt = 0.$$

Hence f is constant on J_i for all $i = 1 \ldots m$ and $f(C)$ has at most m values.

Example: Morse-Sard theorem

Theorem: Let $f\colon \mathbb{R} \mapsto \mathbb{R}$ be SA differentiable, then the set of critical values of f is finite:

$$\mathrm{crit}_f = f\left(\left\{x \in \mathbb{R}, f'(x) = 0\right\}\right)$$

Proof: Setting $C = \{x \in \mathbb{R}, f'(x) = 0\}$, f' is SA, C is semialgebraic and there is $m \in \mathbb{N}$ and intervals $J_1, \ldots, J_m$ such that $C = \cup_{i=1}^m J_i$.

For $i = 1, \ldots m$, J_i is an interval, $f' = 0$ is continuous on J_i, for any $a, b \in J_i$, we have

$$f(b) - f(a) = \int_a^b f'(t)dt = 0.$$

Hence f is constant on J_i for all $i = 1 \ldots m$ and $f(C)$ has at most m values.

Feature of this theory: Some results have simple short proof but rely on a deep technical construction.

Extension to o-minimal structure (van den Dries, Shiota)

o-minimal structure, axiomatic definition: $\mathcal{M} = \cup_{p \in \mathbb{N}} \mathcal{M}_p$, where each $\mathcal{M}_p$ is a family of subsets of $\mathbb{R}^p$

Extension to o-minimal structure (van den Dries, Shiota)

o-minimal structure, axiomatic definition: $\mathcal{M} = \cup_{p \in \mathbb{N}} \mathcal{M}_p$, where each $\mathcal{M}_p$ is a family of subsets of $\mathbb{R}^p$ such that

- if $A, B \in \mathcal{M}_p$ then so does $A \cup B$, $A \cap B$ and $\mathbb{R}^p \setminus A$.
- if $A \in \mathcal{M}_p$ and $B \in \mathcal{M}_p'$, then $A \times B \in \mathcal{M}_{p+p'}$
- each $\mathcal{M}_p$ contains the semi-algebraic sets in $\mathbb{R}^p$.
- if $A \in \mathcal{M}_{p+1}$, denoting π the projection on the first p coordinates, $\pi(A) \in \mathcal{M}_p$.
- M_1 consists of all finite unions intervals.

Extension to o-minimal structure (van den Dries, Shiota)

o-minimal structure, axiomatic definition: $\mathcal{M} = \cup_{p\in\mathbb{N}}\mathcal{M}_p$, where each $\mathcal{M}_p$ is a family of subsets of $\mathbb{R}^p$ such that

- if $A, B \in \mathcal{M}_p$ then so does $A \cup B$, $A \cap B$ and $\mathbb{R}^p \setminus A$.
- if $A \in \mathcal{M}_p$ and $B \in \mathcal{M}'_p$, then $A \times B \in \mathcal{M}_{p+p'}$
- each $\mathcal{M}_p$ contains the semi-algebraic sets in $\mathbb{R}^p$.
- if $A \in \mathcal{M}_{p+1}$, denoting π the projection on the first p coordinates, $\pi(A) \in \mathcal{M}_p$.
- M_1 consists of all finite unions intervals.

Tame function: A function which graph is an element of an o-minimal structure.

Extension to o-minimal structure (van den Dries, Shiota)

o-minimal structure, axiomatic definition: $\mathcal{M} = \cup_{p \in \mathbb{N}} \mathcal{M}_p$, where each $\mathcal{M}_p$ is a family of subsets of $\mathbb{R}^p$ such that

- if $A, B \in \mathcal{M}_p$ then so does $A \cup B$, $A \cap B$ and $\mathbb{R}^p \setminus A$.
- if $A \in \mathcal{M}_p$ and $B \in \mathcal{M}_p'$, then $A \times B \in \mathcal{M}_{p+p'}$
- each $\mathcal{M}_p$ contains the semi-algebraic sets in $\mathbb{R}^p$.
- if $A \in \mathcal{M}_{p+1}$, denoting π the projection on the first p coordinates, $\pi(A) \in \mathcal{M}_p$.
- M_1 consists of all finite unions intervals.

Tame function: A function which graph is an element of an o-minimal structure.

Example: Semialgebraics sets (Tarski-Seidenberg), exp-definable sets (Wilkie), restriction of analytic functions to bounded sets (Gabrielov).

Extension to o-minimal structure (van den Dries, Shiota)

o-minimal structure, axiomatic definition: $\mathcal{M} = \cup_{p \in \mathbb{N}} \mathcal{M}_p$, where each $\mathcal{M}_p$ is a family of subsets of $\mathbb{R}^p$ such that

- if $A, B \in \mathcal{M}_p$ then so does $A \cup B$, $A \cap B$ and $\mathbb{R}^p \setminus A$.
- if $A \in \mathcal{M}_p$ and $B \in \mathcal{M}_p'$, then $A \times B \in \mathcal{M}_{p+p'}$
- each $\mathcal{M}_p$ contains the semi-algebraic sets in $\mathbb{R}^p$.
- if $A \in \mathcal{M}_{p+1}$, denoting π the projection on the first p coordinates, $\pi(A) \in \mathcal{M}_p$.
- M_1 consists of all finite unions intervals.

Tame function: A function which graph is an element of an o-minimal structure.

Example: Semialgebraics sets (Tarski-Seidenberg), exp-definable sets (Wilkie), restriction of analytic functions to bounded sets (Gabrielov).

Consequences: Many results which hold for semi-algebraic sets actually hold for tame functions.

For more: van den Dries and Miller [VdD1998, VdDM1996], Shiota [S1995], Coste's introduction to o-minimal geometry [C2000].

Deep learning training loss

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$
\begin{aligned}
F \colon \mathbb{R}^p &\mapsto \mathbb{R} \\
\theta &\mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta)
\end{aligned}
\tag{P}
$$

$L(\cdot) = (\cdot)^2$ or $L(\cdot) = |\cdot|$ and a Relu network: F is semi-algebraic. More generally for any semi-algebraic L and activation functions.

Deep learning training loss

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$\begin{aligned} F \colon \mathbb{R}^p &\mapsto \mathbb{R} \\ \theta &\mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \end{aligned} \tag{P}$$

$L(\cdot) = (\cdot)^2$ or $L(\cdot) = |\cdot|$ and a Relu network: F is semi-algebraic. More generally for any semi-algebraic L and activation functions.

For most choices of L and activation functions, F is tame (sigmoid, logistic loss . . .).

Plan

1 Introduction

2 Convergence to local minima for Morse-Functions

3 On the structure of deep learning training loss

4 Convergence to critical points for tame functions

5 Approaching critical point with noise

6 Extensions to nonsmooth settings

Introduction

- $F\colon \mathbb{R}^p \mapsto \mathbb{R}$ is $\mathcal{C}^1$ with L-Lipschitz gradient

Introduction

- $F : \mathbb{R}^p \mapsto \mathbb{R}$ is $\mathcal{C}^1$ with L-Lipschitz gradient
- $\alpha \in (0, 1/L]$.

$$x_{k+1} = x_k - \alpha \nabla F(x_k)$$

Introduction

- $F : \mathbb{R}^p \mapsto \mathbb{R}$ is $\mathcal{C}^1$ with L-Lipschitz gradient
- $\alpha \in (0, 1/L]$.

$$x_{k+1} = x_k - \alpha \nabla F(x_k)$$

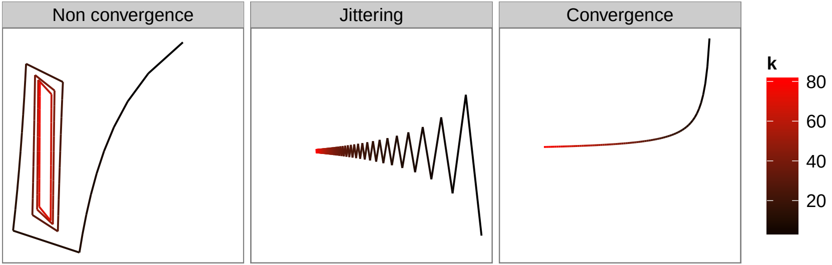
- Convergence of the iterates?

Introduction

- $F : \mathbb{R}^p \mapsto \mathbb{R}$ is $\mathcal{C}^1$ with L-Lipschitz gradient
- $\alpha \in (0, 1/L]$.

$$x_{k+1} = x_k - \alpha \nabla F(x_k)$$

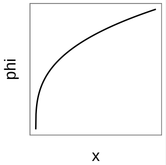
- Convergence of the iterates?



KL property (Łojasiewicz 63, Kurdyka 98)

Desingularizing functions on $(0, r_0)$

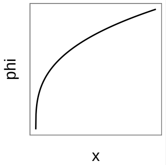
- $\varphi \in \mathcal{C}([0, r_0), \mathbb{R}_+)$,
- $\varphi \in \mathcal{C}^1(0, r_0)$, $\varphi' > 0$,
- φ concave and $\varphi(0) = 0$.



KL property (Łojasiewicz 63, Kurdyka 98)

Desingularizing functions on $(0, r_0)$

- $\varphi \in \mathcal{C}([0, r_0), \mathbb{R}_+)$,
- $\varphi \in \mathcal{C}^1(0, r_0)$, $\varphi' > 0$,
- φ concave and $\varphi(0) = 0$.



Definition

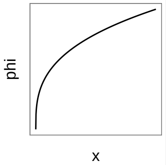
Let $F : \mathbb{R}^p \mapsto \mathbb{R}$ be $\mathcal{C}^1$. F has the KL property at $\bar{x}$ ($F(\bar{x}) = 0$) if there exists $\varepsilon > 0$ and a desingularizing function φ such that,

$$\|\nabla(\varphi \circ F)(x)\|_2 = \phi' \circ F(x) \|F(x)\|_2 \geq 1, \qquad \forall x, \ ||x - \bar{x}|| < \varepsilon, \ 0 < F(x) < \varepsilon.$$

KL property (Łojasiewicz 63, Kurdyka 98)

Desingularizing functions on $(0, r_0)$

- $\varphi \in \mathcal{C}([0, r_0), \mathbb{R}_+)$,
- $\varphi \in \mathcal{C}^1(0, r_0)$, $\varphi' > 0$,
- φ concave and $\varphi(0) = 0$.



Definition

Let $F\colon \mathbb{R}^p \mapsto \mathbb{R}$ be $\mathcal{C}^1$. F has the KL property at $\bar{x}$ ($F(\bar{x}) = 0$) if there exists $\varepsilon > 0$ and a desingularizing function φ such that,

$$\|\nabla(\varphi \circ F)(x)\|_2 = \phi' \circ F(x) \|F(x)\|_2 \geq 1, \qquad \forall x,\ \|x - \bar{x}\| < \varepsilon,\ 0 < F(x) < \varepsilon.$$

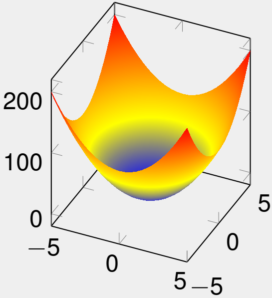
Theorem

KL inequality holds for

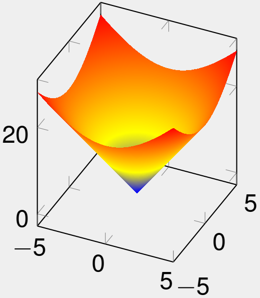
- *differentiable semi-algebraic functions (Łojasiewicz 1963 [Ł1963]).*
- *differentiable tame functions (Kurdyka 1998, [K1998]).*
- *nonsmooth tame functions (Bolte-Daniilidis-Lewis-Shiota 2007 [BDLS2007]).*

Illustration F and $\varphi \circ F$

F and $\varphi \circ F$



Parameterize with φ sharpens the function
$\longrightarrow$



KL inequality examples

Trivial outside critical points: If $\nabla F(\bar{x}) \neq 0$ then one can take φ as multiplication by a small positive constant.

KL inequality examples

Trivial outside critical points: If $\nabla F(\bar{x}) \neq 0$ then one can take φ as multiplication by a small positive constant.

Univariate analytic functions: $F\colon x \mapsto \sum_{i=l}^{+\infty} a_i x^i$ with $l \geq 1$. F is differentiable, around 0

$$|f'| \geq c|f|^{\theta}, \qquad c > 0, \qquad \theta = 1 - \frac{1}{l}.$$

Original form of Łojasiewicz's gradient inequality, corresponds to $\varphi\colon t \mapsto \frac{(1-\theta)}{c} t^{1-\theta}$.

KL inequality examples

Trivial outside critical points: If $\nabla F(\bar{x}) \neq 0$ then one can take φ as multiplication by a small positive constant.

Univariate analytic functions: $F: x \mapsto \sum_{i=l}^{+\infty} a_i x^i$ with $l \geq 1$. F is differentiable, around 0

$$|f'| \geq c|f|^{\theta}, \qquad c > 0, \qquad \theta = 1 - \frac{1}{l}.$$

Original form of Łojasiewicz's gradient inequality, corresponds to $\varphi: t \mapsto \frac{(1-\theta)}{c} t^{1-\theta}$.

μ-strongly convex functions: x^* realises the minimum of F.

$$F(x^*) \geq F(x) + \langle \nabla F(x), x^* - x \rangle + \frac{\mu}{2}\|x - x^*\|_2^2 \qquad \forall x \in \mathbb{R}^p$$

$$\geq F(x) - \frac{1}{2\mu}\|\nabla F(x)\|^2$$

$$2\mu(F(x) - F(x^*)) \leq \|\nabla F(x)\|^2$$

$\theta = 1/2$, $\varphi(\cdot) = \sqrt{\cdot}/\mu$

Non convex examples: $\theta = 1/2$, $\varphi(\cdot) = \sqrt{\cdot}/\mu$

Non convex examples: $\theta = 1/2$, $\varphi(\cdot) = \sqrt{\cdot}/\mu$

Quadratics: $F\colon x \to \frac{1}{2}(x-b)^T A(x-b)$, A symetric, $b \in \mathbb{R}^p$:
μ smallest non zero positive eigenvalue of A and $-A$.

$$\left.\begin{array}{c} \mu A \preceq A^2 \\ -\mu A \preceq A^2 \end{array}\right\} 2\mu(F(x) - F^*) \leq ||\nabla F(x)||^2$$

Non convex examples: $\theta = 1/2$, $\varphi(\cdot) = \sqrt{\cdot}/\mu$

Quadratics: $F\colon x \to \frac{1}{2}(x-b)^T A(x-b)$, A symetric, $b \in \mathbb{R}^p$:
μ smallest non zero positive eigenvalue of A and $-A$.

$$\left.\begin{array}{c} \mu A \preceq A^2 \\ -\mu A \preceq A^2 \end{array}\right\} 2\mu(F(x) - F^*) \leq ||\nabla F(x)||^2$$

Morse functions: F $\mathcal{C}^2$ such that $\nabla F(x) = 0$ implies $\nabla^2 F(x)$ is non singular.

Non convex examples: $\theta = 1/2$, $\varphi(\cdot) = \sqrt{\cdot}/\mu$

Quadratics: $F\colon x \to \frac{1}{2}(x-b)^T A(x-b)$, A symetric, $b \in \mathbb{R}^p$:
μ smallest non zero positive eigenvalue of A and $-A$.

$$\left.\begin{array}{c} \mu A \preceq A^2 \\ -\mu A \preceq A^2 \end{array}\right\} 2\mu(F(x) - F^*) \leq ||\nabla F(x)||^2$$

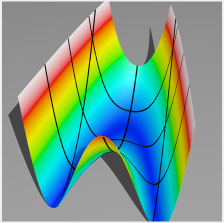
Morse functions: F $\mathcal{C}^2$ such that $\nabla F(x) = 0$ implies $\nabla^2 F(x)$ is non singular.

Non convex example:

$$F\colon x \to \left(\frac{x_1}{2} - x_2^2\right)^2$$

$$\nabla F(x) = \begin{pmatrix} \left(\frac{x_1}{2} - x_2^2\right) \\ -4x_2\left(\frac{x_1}{2} - x_2^2\right) \end{pmatrix}$$

$$(F(x) - F^*) \leq ||\nabla F(x)||^2.$$



Convergence under KL assumption

Theorem (Absil, Mahony, Andrews 2005 [AMA2005])

Let $F : \mathbb{R}^p \mapsto \mathbb{R}$ be bounded bellow, $\mathcal{C}^1$ with L-Lipschitz gradient and satisfy KL property. Assume that $x_0 \in \mathbb{R}^p$ is such that $\{x \in \mathbb{R}^p, F(x) \leq F(x_0)\}$ is compact and for all $k \in \mathbb{N}$

$$x_{k+1} = x_k - \alpha \nabla F(x_k),$$

with $\alpha = 1/L$. Then $x_k \underset{k \to \infty}{\to} \bar{x}$ where $\nabla F(\bar{x}) = 0$ and $\sum_{k \in \mathbb{N}} \|x_{k+1} - x_k\|$ is finite.

Note that KL assumption holds automatically if F is tame which is the case for deep network training losses.

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Ω: set of accumulation points of $(x_k)_{k\in\mathbb{N}}$ (non empty). Ω compact, F constant on Ω (=0), $\mathrm{dist}(x_k, \Omega) \underset{k\to\infty}{\to} 0$. F satisfy KL property at each $x \in \Omega$.

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Ω: set of accumulation points of $(x_k)_{k\in\mathbb{N}}$ (non empty). Ω compact, F constant on Ω (=0), $\mathrm{dist}(x_k, \Omega) \underset{k\to\infty}{\to} 0$. F satisfy KL property at each $x \in \Omega$. Construct ϕ, globally desingularizing.

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Ω: set of accumulation points of $(x_k)_{k\in\mathbb{N}}$ (non empty). Ω compact, F constant on Ω (=0), $\mathrm{dist}(x_k, \Omega) \underset{k\to\infty}{\to} 0$. F satisfy KL property at each $x \in \Omega$. Construct ϕ, globally desingularizing.

Claim: $\exists\, \varepsilon > 0$ and ϕ, s.t. $\|\nabla \phi \circ F(y)\| \geq 1$, $\forall y$, $\mathrm{dist}(y, \Omega) < \varepsilon$, $0 < F(y) < \varepsilon$.

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Ω: set of accumulation points of $(x_k)_{k\in\mathbb{N}}$ (non empty). Ω compact, F constant on Ω (=0), $\mathrm{dist}(x_k, \Omega) \underset{k\to\infty}{\to} 0$. F satisfy KL property at each $x \in \Omega$. Construct ϕ, globally desingularizing.

Claim: $\exists\, \varepsilon > 0$ and ϕ, s.t. $\|\nabla \phi \circ F(y)\| \geq 1$, $\forall y$, $\mathrm{dist}(y, \Omega) < \varepsilon$, $0 < F(y) < \varepsilon$.
For each $x \in \Omega$, consider $\varepsilon_x > 0$ and function φ_x, given by KL property. They induce an open cover of Ω which is compact, extract a finite collection of such open sets which cover Ω.

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Ω: set of accumulation points of $(x_k)_{k\in\mathbb{N}}$ (non empty). Ω compact, F constant on Ω (=0), $\mathrm{dist}(x_k, \Omega) \underset{k\to\infty}{\to} 0$. F satisfy KL property at each $x \in \Omega$. Construct ϕ, globally desingularizing.

Claim: $\exists\, \varepsilon > 0$ and ϕ, s.t. $\|\nabla \phi \circ F(y)\| \geq 1$, $\forall y$, $\mathrm{dist}(y, \Omega) < \varepsilon$, $0 < F(y) < \varepsilon$.
For each $x \in \Omega$, consider $\varepsilon_x > 0$ and function φ_x, given by KL property. They induce an open cover of Ω which is compact, extract a finite collection of such open sets which cover Ω.

We have $x_i \in \Omega$, $\varepsilon_i > 0$, for $i = 1, \ldots n$, open balls $B(x_i, \varepsilon_i)$ cover Ω.

$$0 < \qquad \varepsilon_0 := \min_{y\in\mathbb{R}^p} \mathrm{dist}(y, \Omega), \text{ such that } \|y - x_i\| \geq \varepsilon_i, i = 1 \ldots m$$

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Ω: set of accumulation points of $(x_k)_{k\in\mathbb{N}}$ (non empty). Ω compact, F constant on Ω (=0), $\mathrm{dist}(x_k, \Omega) \underset{k\to\infty}{\to} 0$. F satisfy KL property at each $x \in \Omega$. Construct ϕ, globally desingularizing.

Claim: $\exists\, \varepsilon > 0$ and ϕ, s.t. $\|\nabla \phi \circ F(y)\| \geq 1$, $\forall y$, $\mathrm{dist}(y, \Omega) < \varepsilon$, $0 < F(y) < \varepsilon$.
For each $x \in \Omega$, consider $\varepsilon_x > 0$ and function φ_x, given by KL property. They induce an open cover of Ω which is compact, extract a finite collection of such open sets which cover Ω.

We have $x_i \in \Omega$, $\varepsilon_i > 0$, for $i = 1, \ldots n$, open balls $B(x_i, \varepsilon_i)$ cover Ω.

$$0 < \qquad \varepsilon_0 := \min_{y\in\mathbb{R}^p} \mathrm{dist}(y, \Omega), \text{ such that } \|y - x_i\| \geq \varepsilon_i, i = 1 \ldots m$$

Setting $\varepsilon = \min_{i=0,\ldots,m} \varepsilon_i$, $\varphi = \sum_{i=1}^{m} \varphi_i$, φ is a global desingularizing function on

$$U = \{y \in \mathbb{R}^p, \mathrm{dist}(y, \Omega) < \varepsilon, 0 < F(y) < \varepsilon\}.$$

Proof sketch of convergence under KL assumption

From the descent Lemma, we have

$$\sum_{i=0}^{k} \|\nabla F(x_i)\|_2^2 \leq 2L(F(x_0) - F(x_{k+1})) \leq 2L(F(x_0) - F^*).$$

The sum converges and $\nabla F(x_k) \underset{k\to\infty}{\to} 0$, all accumulation points are critical points.

Ω: set of accumulation points of $(x_k)_{k\in\mathbb{N}}$ (non empty). Ω compact, F constant on Ω (=0), $\mathrm{dist}(x_k,\Omega) \underset{k\to\infty}{\to} 0$. F satisfy KL property at each $x \in \Omega$. Construct ϕ, globally desingularizing.

Claim: $\exists\, \varepsilon > 0$ and ϕ, s.t. $\|\nabla\phi \circ F(y)\| \geq 1$, $\forall y$, $\mathrm{dist}(y,\Omega) < \varepsilon$, $0 < F(y) < \varepsilon$.
For each $x \in \Omega$, consider $\varepsilon_x > 0$ and function φ_x, given by KL property. They induce an open cover of Ω which is compact, extract a finite collection of such open sets which cover Ω.

We have $x_i \in \Omega$, $\varepsilon_i > 0$, for $i = 1,\dots n$, open balls $B(x_i,\varepsilon_i)$ cover Ω.

$$0 < \qquad \varepsilon_0 := \min_{y\in\mathbb{R}^p} \mathrm{dist}(y,\Omega), \text{ such that } \|y - x_i\| \geq \varepsilon_i, i = 1\dots m$$

Setting $\varepsilon = \min_{i=0,\dots,m} \varepsilon_i$, $\varphi = \sum_{i=1}^m \varphi_i$, φ is a global desingularizing function on

$$U = \{y \in \mathbb{R}^p, \mathrm{dist}(y,\Omega) < \varepsilon, 0 < F(y) < \varepsilon\}.$$

$x_k \in U$ for all $k \geq K$ for some $K \in \mathbb{N}$. Discard the first terms so that $K = 0$.

Proof sketch of convergence under KL assumption

For all $k \in \mathbb{N}$,

$$F(x_{k+1})) \leq F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| \qquad \text{descent Lemma}$$

Proof sketch of convergence under KL assumption

For all $k \in \mathbb{N}$,

$$F(x_{k+1})) \leq F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| \qquad \text{descent Lemma}$$

$$\varphi(F(x_{k+1}))) \leq \varphi\left(F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\|\right) \qquad \phi \text{ increasing}$$

Proof sketch of convergence under KL assumption

For all $k \in \mathbb{N}$,

$$
\begin{aligned}
F(x_{k+1})) &\leq F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{descent Lemma} \\
\varphi(F(x_{k+1}))) &\leq \varphi\left(F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\|\right) && \phi \text{ increasing} \\
&\leq \varphi(F(x_k))) - \phi'(F(x_k))\frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{concavity}
\end{aligned}
$$

Proof sketch of convergence under KL assumption

For all $k \in \mathbb{N}$,

$$
\begin{aligned}
F(x_{k+1})) &\leq F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{descent Lemma} \\
\varphi(F(x_{k+1}))) &\leq \varphi\left(F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\|\right) && \phi \text{ increasing} \\
&\leq \varphi(F(x_k))) - \phi'(F(x_k))\frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{concavity} \\
&= \varphi(F(x_k))) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla \phi \circ F(x_k)\|
\end{aligned}
$$

Proof sketch of convergence under KL assumption

For all $k \in \mathbb{N}$,

$$
\begin{aligned}
F(x_{k+1})) &\leq F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{descent Lemma} \\
\varphi(F(x_{k+1}))) &\leq \varphi\left(F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\|\right) && \phi \text{ increasing} \\
&\leq \varphi(F(x_k))) - \phi'(F(x_k))\frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{concavity} \\
&= \varphi(F(x_k))) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla \phi \circ F(x_k)\| && \\
&\leq \varphi(F(x_k))) - \frac{1}{2}\|x_{k+1} - x_k\|_2 && \text{desingularizing}
\end{aligned}
$$

Proof sketch of convergence under KL assumption

For all $k \in \mathbb{N}$,

$$
\begin{aligned}
F(x_{k+1})) &\leq F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{descent Lemma}\\
\varphi(F(x_{k+1}))) &\leq \varphi\left(F(x_k) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\|\right) && \phi \text{ increasing}\\
&\leq \varphi(F(x_k))) - \phi'(F(x_k))\frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla F(x_k)\| && \text{concavity}\\
&= \varphi(F(x_k))) - \frac{1}{2}\|x_{k+1} - x_k\|_2 \|\nabla \phi \circ F(x_k)\| && \\
&\leq \varphi(F(x_k))) - \frac{1}{2}\|x_{k+1} - x_k\|_2 && \text{desingularizing}
\end{aligned}
$$

Finally, $\sum_{k\in\mathbb{N}} \|x_{k+1} - x_k\|_2 \leq 2\phi(F(x_0))$. The sequence is Cauchy and converges.

Generalizations

Idea dates back to Łojasiewicz in the 60's [Ł1963]. Nonsmooth KL inequality [BDLS2007], results and proof techniques extends to many algorithms:

- Forward-backward or proximal gradient for smooth + non smooth.
- Projected gradient for smooth under constraint.
- Proximal point algorithm.
- Block alternating variants
- Inertial variants
- Sequential convex programs for composite objectives
- ...

Also provides convergence rates. See for example [BA2009, ABS2013, BST2014].

Plan

1. Introduction
2. Convergence to local minima for Morse-Functions
3. On the structure of deep learning training loss
4. Convergence to critical points for tame functions
5. Approaching critical point with noise
6. Extensions to nonsmooth settings

Stochastic gradient

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$\begin{aligned} F\colon \mathbb{R}^p &\mapsto \mathbb{R} \\ \theta &\mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \end{aligned} \tag{P}$$

Gradient sampling. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1, \dots, n\}$.

$$\begin{aligned} \theta_{k+1}|\theta_k &= \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \\ \alpha_k &> 0 \end{aligned} \tag{SG}$$

Stochastic gradient

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$\begin{aligned} F \colon \mathbb{R}^p &\mapsto \mathbb{R} \\ \theta &\mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \end{aligned} \tag{P}$$

Stochastic gradient. Let $(M_k)_{k\in\mathbb{N}}$ be a martingale difference sequence.

$$\begin{aligned} \theta_{k+1} | \mathrm{past} &= \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \\ \mathbb{E}\left[M_{k+1} | \mathrm{past}\right] &= 0 \\ \alpha_k &> 0 \end{aligned} \tag{SG}$$

Stochastic gradient

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$\begin{aligned} F \colon \mathbb{R}^p &\mapsto \mathbb{R} \\ \theta &\mapsto \frac{1}{n} \sum_{i=1}^{n} l_i(\theta) \end{aligned} \tag{P}$$

Stochastic gradient. Let $(M_k)_{k \in \mathbb{N}}$ be a martingale difference sequence.

$$\begin{aligned} \theta_{k+1} | \text{past} &= \theta_k - \alpha_k (\nabla F(\theta_k) + M_{k+1}) \\ \mathbb{E}\left[M_{k+1} | \text{past}\right] &= 0 \\ \alpha_k &> 0 \end{aligned} \tag{SG}$$

Stochastic approximation: Robbins and Monro 1951 [RM1951].

The ODE method

Averaging out noise: vanishing step size, $\sum_{k \in \mathbb{N}} \alpha_k = +\infty$, $\sum_{k \in \mathbb{N}} \alpha_k^2 < +\infty$.

The ODE method

Averaging out noise: vanishing step size, $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$.

Differentiable F (Ljung 1977 [L1977]): The sequence $(\theta_k)_{k\in\mathbb{N}}$ behaves in the limit as solutions to the differential equation

$$\dot{\theta} = -\nabla F(\theta) \tag{GS}$$

The ODE method

Averaging out noise: vanishing step size, $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$.

Differentiable F **(Ljung 1977 [L1977]):** The sequence $(\theta_k)_{k\in\mathbb{N}}$ behaves in the limit as solutions to the differential equation

$$\dot{\theta} = -\nabla F(\theta) \tag{GS}$$

Gradient flow: ∇F Lipschitz, then the flow is locally Lipschitz, given by

$$\begin{aligned} S\colon \mathbb{R}^p \times \mathbb{R} &\to \mathbb{R}^p \\ (x,t) &\mapsto \theta(t) \qquad \text{solution of (GS) with } \theta(0) = x. \end{aligned}$$

The ODE method

Averaging out noise: vanishing step size, $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$.

Differentiable F **(Ljung 1977 [L1977]):** The sequence $(\theta_k)_{k\in\mathbb{N}}$ behaves in the limit as solutions to the differential equation

$$\dot{\theta} = -\nabla F(\theta) \tag{GS}$$

Gradient flow: ∇F Lipschitz, then the flow is locally Lipschitz, given by

$$\begin{aligned} S\colon \mathbb{R}^p \times \mathbb{R} &\to \mathbb{R}^p \\ (x,t) &\mapsto \theta(t) \qquad \text{solution of (GS) with } \theta(0) = x. \end{aligned}$$

Developments: Benaïm [B1996], Kushner and Yin [KY1997] ....

Piecewise affine interpolated process

Gradient sampling. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\ldots,n\}$.

$$\theta_{k+1} = \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Piecewise affine interpolated process

Gradient sampling. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\ldots,n\}$.

$$\theta_{k+1} = \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Interpolated process: $\tau_0 = 0$, $\tau_n = \sum_{k=1}^{n} \alpha_k$ for $n \geq 1$ (time).

Piecewise affine interpolated process

Gradient sampling. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\ldots,n\}$.

$$\theta_{k+1} = \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Interpolated process: $\tau_0 = 0$, $\tau_n = \sum_{k=1}^{n} \alpha_k$ for $n \geq 1$ (time).
Define $w\colon \mathbb{R}_+ \to \mathbb{R}^p$, affine interpolation such that $w(\tau_n) = \theta_n$, $n \in \mathbb{N}$.

Piecewise affine interpolated process

Gradient sampling. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\dots,n\}$.

$$\theta_{k+1} = \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \qquad\qquad \alpha_k > 0 \qquad\qquad \text{(SG)}$$

Interpolated process: $\tau_0 = 0$, $\tau_n = \sum_{k=1}^n \alpha_k$ for $n \geq 1$ (time).
Define $w\colon \mathbb{R}_+ \to \mathbb{R}^p$, affine interpolation such that $w(\tau_n) = \theta_n$, $n \in \mathbb{N}$.
For all $n \in \mathbb{N}$ and $0 \leq s < \alpha_{n+1}$

$$w(\tau_n + s) = \theta_n \left(1 - \frac{s}{\alpha_{n+1}}\right) + \frac{s}{\alpha_{n+1}} \theta_{n+1}.$$

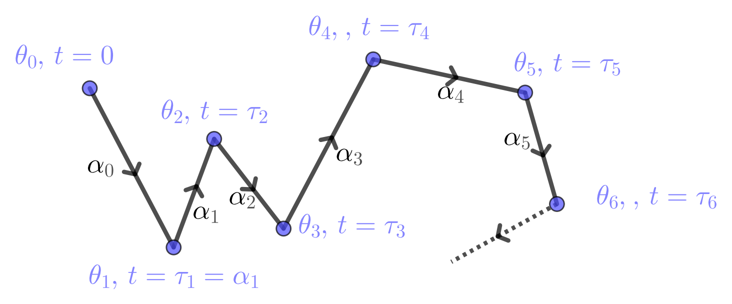
Piecewise affine interpolated process

Gradient sampling. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\ldots,n\}$.

$$\theta_{k+1} = \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Interpolated process: $\tau_0 = 0$, $\tau_n = \sum_{k=1}^{n} \alpha_k$ for $n \geq 1$ (time).
Define $w\colon \mathbb{R}_+ \to \mathbb{R}^p$, affine interpolation such that $w(\tau_n) = \theta_n$, $n \in \mathbb{N}$.
For all $n \in \mathbb{N}$ and $0 \leq s < \alpha_{n+1}$

$$w(\tau_n + s) = \theta_n \left(1 - \frac{s}{\alpha_{n+1}}\right) + \frac{s}{\alpha_{n+1}} \theta_{n+1}.$$



A result of Benaim: flow attracts interpolated process

$$\theta_{k+1}|\text{past} = \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \qquad \mathbb{E}[M_{k+1}|\text{past}] = 0, \qquad \alpha_k > 0 \qquad \text{(SG)}$$

A result of Benaim: flow attracts interpolated process

$$\theta_{k+1}|\text{past} = \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \qquad \mathbb{E}[M_{k+1}|\text{past}] = 0, \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Bounded conditional variance: $\exists M \geq 0$ such that $\sup_{k\in\mathbb{N}} \mathbb{E}\left[M_k^2|\text{past}\right] \leq M$.

A result of Benaim: flow attracts interpolated process

$$\theta_{k+1}|\mathrm{past} = \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \qquad \mathbb{E}[M_{k+1}|\mathrm{past}] = 0, \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Bounded conditional variance: $\exists M \geq 0$ such that $\sup_{k\in\mathbb{N}} \mathbb{E}\left[M_k^2|\mathrm{past}\right] \leq M$.

Step size: $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$

A result of Benaim: flow attracts interpolated process

$$\theta_{k+1}|\mathrm{past} = \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \qquad \mathbb{E}[M_{k+1}|\mathrm{past}] = 0, \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Bounded conditional variance: $\exists M \geq 0$ such that $\sup_{k\in\mathbb{N}} \mathbb{E}\left[M_k^2|\mathrm{past}\right] \leq M$.

Step size: $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty \Rightarrow \sum_{i=0}^{k} \alpha_k M_{k+1}$ converges a.s.

A result of Benaim: flow attracts interpolated process

$$\theta_{k+1}|\mathrm{past} = \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \qquad \mathbb{E}[M_{k+1}|\mathrm{past}] = 0, \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Bounded conditional variance: $\exists M \geq 0$ such that $\sup_{k\in\mathbb{N}} \mathbb{E}\left[M_k^2|\mathrm{past}\right] \leq M$.

Step size: $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty \Rightarrow \sum_{i=0}^{k} \alpha_k M_{k+1}$ converges a.s. (Martingale convergence, square summable increments, Durret Exercise 5.4.8).

A result of Benaim: flow attracts interpolated process

$$\theta_{k+1}|\mathrm{past} = \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \qquad \mathbb{E}[M_{k+1}|\mathrm{past}] = 0, \qquad \alpha_k > 0 \qquad \text{(SG)}$$

Bounded conditional variance: $\exists M \geq 0$ such that $\sup_{k\in\mathbb{N}} \mathbb{E}\left[M_k^2|\mathrm{past}\right] \leq M$.

Step size: $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty \Rightarrow \sum_{i=0}^{k} \alpha_k M_{k+1}$ converges a.s. (Martingale convergence, square summable increments, Durret Exercise 5.4.8).

Theorem (Benaim 1996 [B1996])

Assume that there is $C > 0$ such that $\sup_k \|\theta_k\| \leq C$ almost surely. Then for all $T > 0$,

$$\lim_{t\to\infty} \sup_{0\leq s\leq T} \|w(t+s) - S(w(t), s)\| = 0$$

A result of Benaim: flow attracts interpolated process

$$\theta_{k+1}|\text{past} = \theta_k - \alpha_k(\nabla F(\theta_k) + M_{k+1}) \qquad \mathbb{E}[M_{k+1}|\text{past}] = 0, \qquad \alpha_k > 0 \qquad \text{(SG)}$$

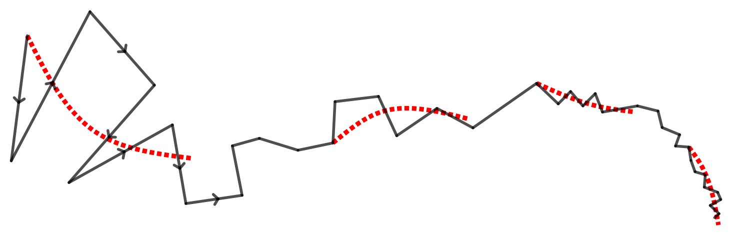
Bounded conditional variance: $\exists M \geq 0$ such that $\sup_{k\in\mathbb{N}} \mathbb{E}\left[M_k^2|\text{past}\right] \leq M$.

Step size: $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty \Rightarrow \sum_{i=0}^{k} \alpha_k M_{k+1}$ converges a.s.
(Martingale convergence, square summable increments, Durret Exercise 5.4.8).

Theorem (Benaim 1996 [B1996])

Assume that there is $C > 0$ such that $\sup_k \|\theta_k\| \leq C$ almost surely. Then for all $T > 0$,

$$\lim_{t\to\infty} \sup_{0\leq s\leq T} \|w(t+s) - S(w(t),s)\| = 0$$



Consequence: descent in the limit

Lemma

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\|\theta_k - \bar{\theta}\| \leq \varepsilon \qquad \forall k = k_i, \ldots, l_i$$
$$F(\theta_{l_i}) \leq F(\bar{\theta}) - \delta.$$

Consequence: descent in the limit

Lemma

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\|\theta_k - \bar{\theta}\| \leq \varepsilon \qquad \forall k = k_i, \ldots, l_i$$
$$F(\theta_{l_i}) \leq F(\bar{\theta}) - \delta.$$

Proof: Choose $T > 0$ and γ the solution to $\dot{\theta} = -\nabla F(\theta)$ with $\gamma(0) = \bar{\theta}$ on $[0, T]$, such that $\|\gamma(s) - \bar{\theta}\| < \varepsilon$ for all $s \in [0, T]$.

Consequence: descent in the limit

Lemma

Let $(k_i)_{i \in \mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i \to \infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\|\theta_k - \bar{\theta}\| \leq \varepsilon \qquad \forall k = k_i, \ldots, l_i$$
$$F(\theta_{l_i}) \leq F(\bar{\theta}) - \delta.$$

Proof: Choose $T > 0$ and γ the solution to $\dot{\theta} = -\nabla F(\theta)$ with $\gamma(0) = \bar{\theta}$ on $[0, T]$, such that $\|\gamma(s) - \bar{\theta}\| < \varepsilon$ for all $s \in [0, T]$.

$$F(\gamma(T)) = F(\bar{\theta}) + \int_0^T \langle \dot{\gamma}(s), \nabla F(\gamma(s)) \rangle \, ds$$

Consequence: descent in the limit

Lemma

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\|\theta_k - \bar{\theta}\| \leq \varepsilon \qquad \forall k = k_i, \ldots, l_i$$
$$F(\theta_{l_i}) \leq F(\bar{\theta}) - \delta.$$

Proof: Choose $T > 0$ and γ the solution to $\dot{\theta} = -\nabla F(\theta)$ with $\gamma(0) = \bar{\theta}$ on $[0, T]$, such that $\|\gamma(s) - \bar{\theta}\| < \varepsilon$ for all $s \in [0, T]$.

$$F(\gamma(T)) = F(\bar{\theta}) + \int_0^T \langle \dot{\gamma}(s), \nabla F(\gamma(s)) \rangle \, ds = F(\bar{\theta}) - \underbrace{\int_0^T \|\nabla F(\gamma(s))\|^2 ds}_{\delta > 0}$$

Consequence: descent in the limit

Lemma

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\begin{aligned} \|\theta_k - \bar{\theta}\| &\leq \varepsilon \qquad \forall k = k_i, \ldots, l_i \\ F(\theta_{l_i}) &\leq F(\bar{\theta}) - \delta. \end{aligned}$$

Proof: Choose $T > 0$ and γ the solution to $\dot{\theta} = -\nabla F(\theta)$ with $\gamma(0) = \bar{\theta}$ on $[0, T]$, such that $\|\gamma(s) - \bar{\theta}\| < \varepsilon$ for all $s \in [0, T]$.

$$F(\gamma(T)) = F(\bar{\theta}) + \int_0^T \langle \dot{\gamma}(s), \nabla F(\gamma(s)) \rangle \, ds = F(\bar{\theta}) - \underbrace{\int_0^T \|\nabla F(\gamma(s))\|^2 ds}_{\delta > 0}$$

Set l_i the largest index l such that $\tau_l \leq \tau_{k_i} + T$.

Consequence: descent in the limit

Lemma

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\|\theta_k - \bar{\theta}\| \leq \varepsilon \qquad \forall k = k_i, \dots, l_i$$
$$F(\theta_{l_i}) \leq F(\bar{\theta}) - \delta.$$

Proof: Choose $T > 0$ and γ the solution to $\dot{\theta} = -\nabla F(\theta)$ with $\gamma(0) = \bar{\theta}$ on $[0, T]$, such that $\|\gamma(s) - \bar{\theta}\| < \varepsilon$ for all $s \in [0, T]$.

$$F(\gamma(T)) = F(\bar{\theta}) + \int_0^T \langle \dot{\gamma}(s), \nabla F(\gamma(s)) \rangle \, ds = F(\bar{\theta}) - \underbrace{\int_0^T \|\nabla F(\gamma(s))\|^2 ds}_{\delta > 0}$$

Set l_i the largest index l such that $\tau_l \leq \tau_{k_i} + T$. As $i \to \infty$

$$\max_{k=k_i,\dots,l_i} \min_{s\in[0,T]} \|\theta_k - \gamma(s)\| \to 0 \qquad \text{Benaim + continuous flow}$$

Consequence: descent in the limit

Lemma

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\|\theta_k - \bar{\theta}\| \leq \varepsilon \qquad \forall k = k_i, \dots, l_i$$
$$F(\theta_{l_i}) \leq F(\bar{\theta}) - \delta.$$

Proof: Choose $T > 0$ and γ the solution to $\dot{\theta} = -\nabla F(\theta)$ with $\gamma(0) = \bar{\theta}$ on $[0, T]$, such that $\|\gamma(s) - \bar{\theta}\| < \varepsilon$ for all $s \in [0, T]$.

$$F(\gamma(T)) = F(\bar{\theta}) + \int_0^T \langle \dot{\gamma}(s), \nabla F(\gamma(s)) \rangle \, ds = F(\bar{\theta}) - \underbrace{\int_0^T \|\nabla F(\gamma(s))\|^2 ds}_{\delta > 0}$$

Set l_i the largest index l such that $\tau_l \leq \tau_{k_i} + T$. As $i \to \infty$

$$\max_{k=k_i,\dots,l_i} \min_{s\in[0,T]} \|\theta_k - \gamma(s)\| \to 0 \qquad \text{Benaim + continuous flow}$$

$$\tau_{l_i} - \tau_{k_i} \to T \qquad \text{vanishing steps}$$

Consequence: descent in the limit

Lemma

Let $(k_i)_{i\in\mathbb{N}}$ be a subsequence, $\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta}$, with $\nabla F(\bar{\theta}) \neq 0$. Then for any $\varepsilon > 0$, there exists $\delta, T > 0$, a subsequence $l_i \geq k_i$, $i \in \mathbb{N}$, such that, for large enough i

$$\|\theta_k - \bar{\theta}\| \leq \varepsilon \qquad \forall k = k_i, \ldots, l_i$$
$$F(\theta_{l_i}) \leq F(\bar{\theta}) - \delta.$$

Proof: Choose $T > 0$ and γ the solution to $\dot{\theta} = -\nabla F(\theta)$ with $\gamma(0) = \bar{\theta}$ on $[0, T]$, such that $\|\gamma(s) - \bar{\theta}\| < \varepsilon$ for all $s \in [0, T]$.

$$F(\gamma(T)) = F(\bar{\theta}) + \int_0^T \langle \dot{\gamma}(s), \nabla F(\gamma(s)) \rangle \, ds = F(\bar{\theta}) - \underbrace{\int_0^T \|\nabla F(\gamma(s))\|^2 ds}_{\delta > 0}$$

Set l_i the largest index l such that $\tau_l \leq \tau_{k_i} + T$. As $i \to \infty$

$$\max_{k=k_i,\ldots,l_i} \min_{s\in[0,T]} \|\theta_k - \gamma(s)\| \to 0 \qquad \text{Benaim + continuous flow}$$

$$\tau_{l_i} - \tau_{k_i} \to T \qquad \text{vanishing steps}$$

$$F(\theta_{l_i}) \to F(\gamma(T)) = F(\bar{\theta}) - \delta \qquad \text{Benaim + continuity of } F.$$

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar{\theta}$ critical.

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar{\theta}$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar{\theta}$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Lemma

Let Ω be the set of limit point of $(F(\theta_k))_{k\in\mathbb{N}}$. Ω is an interval contained in F^*.

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar{\theta}$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Lemma

Let Ω be the set of limit point of $(F(\theta_k))_{k\in\mathbb{N}}$. Ω is an interval contained in F^*.

Proof: Ω is a compact interval (exercise). $\min_{t\in\Omega} t \in F^*$. Assume not singleton.

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar{\theta}$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Lemma

Let Ω be the set of limit point of $(F(\theta_k))_{k\in\mathbb{N}}$. Ω is an interval contained in F^*.

Proof: Ω is a compact interval (exercise). $\min_{t\in\Omega} t \in F^*$. Assume not singleton.
Suppose $\bar{f} \in \mathrm{int}(\Omega) \setminus F^*$, then there is $f_2 > \bar{f}$, $f_2 \in \Omega$.

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar\theta$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Lemma

Let Ω be the set of limit point of $(F(\theta_k))_{k\in\mathbb{N}}$. Ω is an interval contained in F^*.

Proof: Ω is a compact interval (exercise). $\min_{t\in\Omega} t \in F^*$. Assume not singleton.
Suppose $\bar f \in \mathrm{int}(\Omega) \setminus F^*$, then there is $f_2 > \bar f$, $f_2 \in \Omega$.
There exists subsequences $(k_i)_{i\in\mathbb{N}}$, $(m_i)_{i\in\mathbb{N}}$, such that

$$F(\theta_{k_i}) \leq \bar f \qquad F(\theta_{m_i}) \geq f_2 \qquad \bar f \leq F(\theta_k) \leq f_2, \qquad \forall k = k_i + 1, \ldots m_i - 1$$

$$\theta_{k_i} \underset{i\to\infty}{\to} \bar\theta \qquad f(\bar\theta) = \bar f < f_2.$$

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar{\theta}$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Lemma

Let Ω be the set of limit point of $(F(\theta_k))_{k\in\mathbb{N}}$. Ω is an interval contained in F^*.

Proof: Ω is a compact interval (exercise). $\min_{t\in\Omega} t \in F^*$. Assume not singleton.
Suppose $\bar{f} \in \mathrm{int}(\Omega) \setminus F^*$, then there is $f_2 > \bar{f}$, $f_2 \in \Omega$.
There exists subsequences $(k_i)_{i\in\mathbb{N}}$, $(m_i)_{i\in\mathbb{N}}$, such that

$$F(\theta_{k_i}) \leq \bar{f} \qquad F(\theta_{m_i}) \geq f_2 \qquad \bar{f} \leq F(\theta_k) \leq f_2, \qquad \forall k = k_i + 1, \dots m_i - 1$$
$$\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta} \qquad f(\bar{\theta}) = \bar{f} < f_2.$$

Then $\nabla F(\bar{\theta}) \neq 0$. Choose $\varepsilon > 0$ such that

$$\max_{\|\bar{\theta}-y\|\leq\varepsilon} f(y) < f_2.$$

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar\theta$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Lemma

Let Ω be the set of limit point of $(F(\theta_k))_{k\in\mathbb{N}}$. Ω is an interval contained in F^*.

Proof: Ω is a compact interval (exercise). $\min_{t\in\Omega} t \in F^*$. Assume not singleton.
Suppose $\bar f \in \mathrm{int}(\Omega) \setminus F^*$, then there is $f_2 > \bar f$, $f_2 \in \Omega$.
There exists subsequences $(k_i)_{i\in\mathbb{N}}$, $(m_i)_{i\in\mathbb{N}}$, such that

$$F(\theta_{k_i}) \leq \bar f \qquad F(\theta_{m_i}) \geq f_2 \qquad \bar f \leq F(\theta_k) \leq f_2, \qquad \forall k = k_i+1,\dots m_i - 1$$
$$\theta_{k_i} \underset{i\to\infty}{\to} \bar\theta \qquad f(\bar\theta) = \bar f < f_2.$$

Then $\nabla F(\bar\theta) \neq 0$. Choose $\varepsilon > 0$ such that

$$\max_{\|\bar\theta - y\| \leq \varepsilon} f(y) < f_2.$$

Descent in the limit: for i large enough,
$\exists l_i$, $k_i \leq l_i$, $F(\theta_{l_i}) < \bar f$ and $\|\theta_k - \bar\theta\| \leq \varepsilon$, $k = k_i \dots, l_i$.

Consequence: limit values and critical points

$\liminf_{k\to\infty} F(\theta_k)$ critical value of F. Corresponding accumulation points $\bar{\theta}$ critical.
Set $F^* = \{F(\theta), \nabla F(\theta) = 0, \|\theta\| \leq C\}$ the critical values of F (closed).

Lemma

Let Ω be the set of limit point of $(F(\theta_k))_{k\in\mathbb{N}}$. Ω is an interval contained in F^*.

Proof: Ω is a compact interval (exercise). $\min_{t\in\Omega} t \in F^*$. Assume not singleton.
Suppose $\bar{f} \in \mathrm{int}(\Omega) \setminus F^*$, then there is $f_2 > \bar{f}$, $f_2 \in \Omega$.
There exists subsequences $(k_i)_{i\in\mathbb{N}}$, $(m_i)_{i\in\mathbb{N}}$, such that

$$F(\theta_{k_i}) \leq \bar{f} \qquad F(\theta_{m_i}) \geq f_2 \qquad \bar{f} \leq F(\theta_k) \leq f_2, \qquad \forall k = k_i + 1, \dots m_i - 1$$

$$\theta_{k_i} \underset{i\to\infty}{\to} \bar{\theta} \qquad f(\bar{\theta}) = \bar{f} < f_2.$$

Then $\nabla F(\bar{\theta}) \neq 0$. Choose $\varepsilon > 0$ such that

$$\max_{\|\bar{\theta}-y\|\leq\varepsilon} f(y) < f_2.$$

Descent in the limit: for i large enough,
$\exists l_i$, $k_i \leq l_i$, $F(\theta_{l_i}) < \bar{f}$ and $\|\theta_k - \bar{\theta}\| \leq \varepsilon$, $k = k_i \dots, l_i$. $l_i \leq m_i$, contradiction.

Corollary for deep learning

Stochastic gradient. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\ldots,n\}$.

$$\theta_{k+1}|\theta_k = \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \qquad \text{(SG)}$$
$$\alpha_k > 0$$

Conditioning on $(\theta_k)_{k\in\mathbb{N}}$ being bounded, almost surely, $F(\theta_k)$ converges and any accumulation point $\bar{\theta}$ satisfies $\nabla F(\bar{\theta}) = 0$.

Corollary for deep learning

Stochastic gradient. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\ldots,n\}$.

$$\theta_{k+1}|\theta_k = \theta_k - \alpha_k \nabla l_{i_k}(\theta_k) \quad \text{(SG)}$$
$$\alpha_k > 0$$

Conditioning on $(\theta_k)_{k\in\mathbb{N}}$ being bounded, almost surely, $F(\theta_k)$ converges and any accumulation point $\bar{\theta}$ satisfies $\nabla F(\bar{\theta}) = 0$.

Main ingredients:

- Common neural networks are tame (semialgebraic). F is definable.
- **Definable Morse-Sard theorem:** the set F^* of critical values of F is finite.
- $\Omega \subset F^*$ is an interval. It is a singleton. $F(\theta_k)$ converges.
- Accumulation points are critical, otherwise descent in the limit implies Ω not singleton.

Plan

1 Introduction

2 Convergence to local minima for Morse-Functions

3 On the structure of deep learning training loss

4 Convergence to critical points for tame functions

5 Approaching critical point with noise

6 Extensions to nonsmooth settings

Stochastic subgradient

$\theta = (\mathbf{w}, \mathbf{b})$, $l_i(\theta) = L(f_{\mathbf{w},\mathbf{b}}(x_i), y_i)$, $i = 1 \dots n$.

$$F \colon \mathbb{R}^p \mapsto \mathbb{R}$$
$$\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \tag{P}$$

Stochastic subgradient method. Let $(M_k)_{k\in\mathbb{N}}$ be a martingale difference sequence.

$$\theta_{k+1}|\text{past} = \theta_k - \alpha_k(v + M_{k+1}) \tag{SG}$$
$$\mathbb{E}\left[M_{k+1}|\text{past}\right] = 0$$
$$v \in \partial F(\theta_k)$$
$$\alpha_k > 0$$

∂ is a suitable generalization of the gradient.

Subgradients: $F : \mathbb{R}^p \mapsto \mathbb{R}$ Lipschitz continuous

Convex: only for F convex, global lower tangent.

$$\partial_{\text{conv}} F(x) = \left\{ v \in \mathbb{R}^p, F(y) \geq F(x) + v^T(y - x), \forall y \in \mathbb{R}^p \right\}.$$

Subgradients: $F : \mathbb{R}^p \mapsto \mathbb{R}$ Lipschitz continuous

Convex: only for F convex, global lower tangent.

$$\partial_{\mathrm{conv}} F(x) = \left\{ v \in \mathbb{R}^p,\, F(y) \geq F(x) + v^T(y-x),\, \forall y \in \mathbb{R}^p \right\}.$$

Fréchet: local lower tangent.

$$\partial_{\mathrm{Frechet}} F(x) = \left\{ v \in \mathbb{R}^p,\, \liminf_{y \to x,\, y \neq x} \frac{F(y) - F(x) - v^T(y-x)}{\|y-x\|} \geq 0 \right\}.$$

Subgradients: $F : \mathbb{R}^p \mapsto \mathbb{R}$ Lipschitz continuous

Convex: only for F convex, global lower tangent.

$$\partial_{\mathrm{conv}} F(x) = \left\{ v \in \mathbb{R}^p,\, F(y) \geq F(x) + v^T(y - x), \forall y \in \mathbb{R}^p \right\}.$$

Fréchet: local lower tangent.

$$\partial_{\mathrm{Frechet}} F(x) = \left\{ v \in \mathbb{R}^p,\, \lim_{y \to x,} \inf_{y \neq x} \frac{F(y) - F(x) - v^T(y - x)}{\|y - x\|} \geq 0 \right\}.$$

Limiting: sequential closure.

$$\partial_{\lim} F(x) = \left\{ v \in \mathbb{R}^p, \exists (y_k, v_k)_{k \in \mathbb{N}},\, y_k \underset{k \to \infty}{\to} x,\, v_k \underset{k \to \infty}{\to} v,\, v_k \in \partial_{\mathrm{Frechet}} F(y_k), k \in \mathbb{N} \right\}.$$

Subgradients: $F : \mathbb{R}^p \mapsto \mathbb{R}$ Lipschitz continuous

Convex: only for F convex, global lower tangent.

$$\partial_{\mathrm{conv}} F(x) = \left\{ v \in \mathbb{R}^p,\, F(y) \geq F(x) + v^T(y-x), \forall y \in \mathbb{R}^p \right\}.$$

Fréchet: local lower tangent.

$$\partial_{\mathrm{Frechet}} F(x) = \left\{ v \in \mathbb{R}^p,\, \liminf_{y \to x,\, y \neq x} \frac{F(y) - F(x) - v^T(y-x)}{\|y-x\|} \geq 0 \right\}.$$

Limiting: sequential closure.

$$\partial_{\mathrm{lim}} F(x) = \left\{ v \in \mathbb{R}^p,\, \exists (y_k, v_k)_{k \in \mathbb{N}},\, y_k \underset{k \to \infty}{\to} x,\, v_k \underset{k \to \infty}{\to} v,\, v_k \in \partial_{\mathrm{Frechet}} F(y_k), k \in \mathbb{N} \right\}.$$

Clarke: convex closure.

$$\partial_{\mathrm{Clarke}} F(x) = \mathrm{conv}(\partial_{\mathrm{lim}} F(x)).$$

Subgradients: $F\colon \mathbb{R}^p \mapsto \mathbb{R}$ Lipschitz continuous

Example: $F\colon x \mapsto -|x|$.

$$\begin{aligned}
\partial_{\text{conv}} F(0) &= \emptyset \\
\partial_{\text{Frechet}} F(0) &= \emptyset \\
\partial_{\lim} F(0) &= \{-1, 1\} \\
\partial_{\text{Clarke}} F(0) &= [-1, 1].
\end{aligned}$$

0 is a local maximum, it is critical only for for the most general notion of subgradient which we have seen . . .

- $\partial_{\text{Frechet}} F(x) \subset \partial_{\lim} F(x) \subset \partial_{\text{Clarke}} F(x)$ for all x.
- Fermat rule: $\bar{x}$ is a local minimum of F if and only if $0 \in \partial_{\text{Frechet}} F(\bar{x})$.

We will work with Clarke subgradients.

Absolute continuity

Definition (Absolutely continuous map)

Let $g\colon I \mapsto \mathbb{R}^p$ be AC on an interval I. This means that for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any collection of pairwise disjoint sub intervals of I, $\{[u_k, v_k]\}_{k\in\mathbb{N}}$, we have

$$\sum_{k\in\mathbb{N}} |u_k - v_k| \leq \delta \Rightarrow \sum_{k\in\mathbb{N}} \|g(u_k) - g(v_k)\| \leq \varepsilon.$$

Lipschitz functions are AC and composition of Lipschitz and AC functions are AC.

Equivalently there exists a Lebesgue integrable function $y\colon I \to \mathbb{R}^p$ and $a \in I$ such that for all $t \in I$,

$$f(t) = f(a) + \int_a^t y(s)ds.$$

Most importantly: AC functions are differentiable almost everywhere and are the integral of their derivative.

Differential inclusion solutions

F Lipschitz: gradient ODE replaced by a differential inclusion.

Definition

Given $\theta_0 \in \mathbb{R}^p$, a solution of the problem

$$\dot{\theta} \in -\partial F(\theta), \qquad \theta(0) = \theta_0,$$

is any Absolutely Continuous map $\theta \colon \mathbb{R} \mapsto \mathbb{R}^p$, such that $\frac{d}{dt}\theta(t) \in -\partial F(\theta(t))$ for almost all t and $\theta(0) = \theta_0$.

Example: absolute value.

Differential inclusion solutions

F Lipschitz: gradient ODE replaced by a differential inclusion.

Definition

Given $\theta_0 \in \mathbb{R}^p$, a solution of the problem

$$\dot{\theta} \in -\partial F(\theta), \qquad \theta(0) = \theta_0,$$

is any Absolutely Continuous map $\theta \colon \mathbb{R} \mapsto \mathbb{R}^p$, such that $\frac{d}{dt}\theta(t) \in -\partial F(\theta(t))$ for almost all t and $\theta(0) = \theta_0$.

Example: absolute value.

Theorem: Properties of ∂F ensures existence of solution (not unique).

Chain rule along absolutely continuous curves

If $F\colon \mathbb{R}^p \to \mathbb{R}$, and $\theta\colon \mathbb{R} \to \mathbb{R}^p$ are C^1, then $\frac{d}{dt}F(\theta(t)) = \left\langle \nabla F(\theta(t)), \dot{\theta}(t) \right\rangle$.

Chain rule along absolutely continuous curves

If $F\colon \mathbb{R}^p \to \mathbb{R}$, and $\theta\colon \mathbb{R} \to \mathbb{R}^p$ are C^1, then $\frac{d}{dt}F(\theta(t)) = \left\langle \nabla F(\theta(t)), \dot{\theta}(t) \right\rangle$.

Lemma (Convex chain rule, Lyapunov function)

If F is convex and $\theta\colon \mathbb{R} \to \mathbb{R}^p$ is absolutely continuous, then for almost all $t \in \mathbb{R}$,

$$\frac{d}{dt}F(\theta(t)) = \left\langle v, \dot{\theta} \right\rangle \qquad \forall v \in \partial F(\theta(t)).$$

Chain rule along absolutely continuous curves

If $F\colon \mathbb{R}^p \to \mathbb{R}$, and $\theta\colon \mathbb{R} \to \mathbb{R}^p$ are C^1, then $\frac{d}{dt}F(\theta(t)) = \left\langle \nabla F(\theta(t)), \dot{\theta}(t) \right\rangle$.

Lemma (Convex chain rule, Lyapunov function)

If F is convex and $\theta\colon \mathbb{R} \to \mathbb{R}^p$ is absolutely continuous, then for almost all $t \in \mathbb{R}$,

$$\frac{d}{dt}F(\theta(t)) = \left\langle v, \dot{\theta} \right\rangle \qquad \forall v \in \partial F(\theta(t)).$$

If in addition θ is solution to $\dot{\theta} \in -\partial F(\theta)$, then for almost all $t \in \mathbb{R}^+$,

$$\frac{d}{dt}F(\theta(t)) = -\mathrm{dist}(0, \partial F(\theta(t)))^2.$$

Example: ℓ_1 norm. See Brézis 1973 [B1973].

Chain rule along absolutely continuous curves

If $F\colon \mathbb{R}^p \to \mathbb{R}$, and $\theta\colon \mathbb{R} \to \mathbb{R}^p$ are C^1, then $\frac{d}{dt}F(\theta(t)) = \left\langle \nabla F(\theta(t)), \dot\theta(t) \right\rangle$.

Lemma (Convex chain rule, Lyapunov function)

If F is convex and $\theta\colon \mathbb{R} \to \mathbb{R}^p$ is absolutely continuous, then for almost all $t \in \mathbb{R}$,

$$\frac{d}{dt}F(\theta(t)) = \left\langle v, \dot\theta \right\rangle \qquad \forall v \in \partial F(\theta(t)).$$

If in addition θ is solution to $\dot\theta \in -\partial F(\theta)$, then for almost all $t \in \mathbb{R}^+$,

$$\frac{d}{dt}F(\theta(t)) = -\mathrm{dist}(0, \partial F(\theta(t)))^2.$$

Example: ℓ_1 norm. See Brézis 1973 [B1973].

Remark: not true in general: there are 1-Lipschitz F such that:

$$\partial^c F \text{ is the unit ball everywhere.}$$

Chain rule for convex functions

F is locally Lipschitz, and x is AC so that the composition is also AC and we may choose t_0 such that both θ and $F \circ \theta$ are differentiable. Let $\theta_0 = \theta(t_0)$ and $\dot{\theta}_0 = \dot{\theta}(t_0)$, we have

$$\theta(t_0 + h) = \theta_0 + h\dot{\theta}_0 + o(h)$$
$$\theta(t_0 - h) = \theta_0 - h\dot{\theta}_0 + o(h)$$

For any $v \in \partial F(\theta_0)$, it holds that $\langle v, y - \theta_0 \rangle \leq F(y) - F(\theta_0)$ for all $y \in \mathbb{R}^p$. Now imposing $h > 0$, we have

$$\frac{\langle v, \theta(t_0 + h) - \theta_0 \rangle}{h} = \left\langle v, \dot{\theta}_0 \right\rangle + o(1) \leq \frac{F(\theta(t_0 + h)) - F(\theta_0)}{h} \underset{h \to 0}{\to} \frac{d}{dt}(F \circ \theta)(t_0).$$

On the other hand, still considering h positive

$$\frac{\langle v, \theta(t_0 - h) - \theta_0 \rangle}{-h} = \left\langle v, \dot{\theta}_0 \right\rangle + o(1) \geq \frac{F(\theta(t_0 - h)) - F(\theta_0)}{-h} \underset{h \to 0}{\to} \frac{d}{dt}(F \circ \theta)(t_0).$$

This proves the first identity. $\dot{\theta}_0 \in \partial F(\theta_0)$ and it is “orthogonal” to $\partial F(\theta_0)$ so that it is the minimum norm element.

Corollary for deep learning

Stochastic subgradient. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\dots,n\}$.

$$\begin{aligned}\theta_{k+1}|\mathrm{past} &= \theta_k - \alpha_k(v + M_{k+1}) \\ \mathbb{E}\left[M_{k+1}|\mathrm{past}\right] &= 0 \\ v &\in \partial F(\theta_k) \\ \alpha_k &> 0\end{aligned} \tag{SG}$$

Conditioning on $(\theta_k)_{k\in\mathbb{N}}$ being bounded, almost surely, $F(\theta_k)$ converges and any accumulation point $\bar\theta$ satisfies $0 \in \partial F(\bar\theta)$.

Corollary for deep learning

Stochastic subgradient. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1,\dots,n\}$.

$$\begin{aligned} \theta_{k+1}|\text{past} &= \theta_k - \alpha_k(v + M_{k+1}) \\ \mathbb{E}[M_{k+1}|\text{past}] &= 0 \\ v &\in \partial F(\theta_k) \\ \alpha_k &> 0 \end{aligned} \tag{SG}$$

Conditioning on $(\theta_k)_{k\in\mathbb{N}}$ being bounded, almost surely, $F(\theta_k)$ converges and any accumulation point $\bar{\theta}$ satisfies $0 \in \partial F(\bar{\theta})$.

Proof arguments: Same idea as in the smooth case

- The differential inclusion flow attracts the dynamics[BHS2005].
- F definable, chain rule [DDKL2018], using variational stratification of [BDLS2007].
- $\to$ Descent in the limit.
- Ω, the accumulation values of $F(\theta_k)$ form a closed interval in F^*.
- F is definable, nonsmooth Morse Sard [BDLS2007], critical values are finite.

Opening

Nonsmooth functions satisfying a chain rule with Clarke subdifferential are called *path differentiable*. In this case Clarke subgradient is called *conservative*.

Opening

Nonsmooth functions satisfying a chain rule with Clarke subdifferential are called *path differentiable*. In this case Clarke subgradient is called *conservative*.

- Differential calculus and backpropagation.
- Strong geometric interpretation.
- Various extensions: implicit functions, abstract integrals, ODE flows, complexity.
- Ongoing . . .

P.A. Absil, R. Mahony, B. Andrews, B. (2005).
Convergence of the iterates of descent methods for analytic cost functions.
SIAM Journal on Optimization, 16(2), 531–547.

H. Attouch, J. Bolte and B.F. Svaiter (2013).
Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods.
Mathematical Programming, 137(1-2), 91–129.

J.P. Aubin and A. Cellina (1984).
Differential inclusions: set-valued maps and viability theory. Springer Science & Business Media.

M. Benaim (1996).
A dynamical system approach to stochastic approximations.
SIAM Journal on Control and Optimization, 34(2), 437–472.

M. Benaïm (1999).
Dynamics of stochastic approximation algorithms.
Séminaire de probabilités XXXIII (pp. 1-68). Springer, Berlin, Heidelberg.

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

J. Bolte, A. Daniilidis, A. Lewis and M. Shiota (2007).
Clarke subgradients of stratifiable functions.
SIAM Journal on Optimization, 18(2), 556-572.

H. Attouch and J. Bolte (2009).
On the convergence of the proximal algorithm for nonsmooth functions involving analytic features.
Mathematical Programming, 116(1-2), 5-16.

J. Bolte, S. Sabach and M. Teboulle (2014).
Proximal alternating linearized minimization or nonconvex and nonsmooth problems.
Mathematical Programming, 146(1-2), 459–494.

H. Brézis (1973).
Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert (Vol. 5). Elsevier.

F.H. Clarke, Y.S. Ledyaev, R.J. Stern and P.R. Wolenski (1998).
Nonsmooth analysis and control theory.
Springer Science & Business Media.

M. Coste (2000).
An introduction to o-minimal geometry.
Pisa: Istituti editoriali e poligrafici internazionali.

M. Coste (2002).
An introduction to semialgebraic geometry.
Pisa: Istituti editoriali e poligrafici internazionali.

D. Davis, D. Drusvyatskiy, S. Kakade and J. D. Lee (2018).
Stochastic subgradient method converges on tame functions.
arXiv preprint arXiv:1804.07795.

L. Van den Dries and C.Miller (1996).
Geometric categories and o-minimal structures.
Duke Math. J, 84(2), 497-540.

L. Van den Dries, (1998).
Tame topology and o-minimal structures (Vol. 248). Cambridge university press.

K. Kurdyka (1998).
On gradients of functions definable in o-minimal structures.
Annales de l'institut Fourier 48(3)769–784.

H. Kushner and G.G. Yin (1997). Stochastic approximation and recursive algorithms and applications. Springer Science & Business Media.

J.D. Lee and M. Simchowitz and M.I. Jordan and B. Recht (2016).
Gradient descent only converges to minimizers.
Conference on Learning Theory (pp. 1246–1257).

L. Ljung (1977).
Analysis of recursive stochastic algorithms.
IEEE transactions on automatic control, 22(4), 551–575.

S. Łojasiewicz (1963).
Une propriété topologique des sous-ensembles analytiques réels.
Les équations aux dérivées partielles, 117, 87–89.

H. Robbins and S. Monro (1951).
A Stochastic Approximation Method.
The Annals of Mathematical Statistics, 22(3), 400–407.

M. Shiota (1995).
Geometry of subanalytic and semialgebraic sets. Springer Science & Business Media.

M. Shub (1987).
Global stability of dynamical systems. Springer Science & Business Media.