Optimization for Machine Learning

Chapter IV : Nonconvex Optimization

Guillaume Garrigos

1st Semester 2022-2023

What can we expect?

- Does my algorithm converge? $x_\infty := \lim_{k \to +\infty} x^k$ exists?
- What is the nature of the limit x_∞?
  Global/local minima? Saddle point?

I : Convergence of the methods

I : Convergence of the methods

1 : General results

General results

$f : \mathbb{R}^N \longrightarrow \mathbb{R}$ is of class $C_L^{1,1}$ $\qquad x^{k+1} = x^k - \lambda \nabla f(x^k)$

Proposition

Let $\lambda \in]0, \frac{2}{L}[$, then

1. $f(x^k)$ is decreasing
2. if $x^{k_n} \to x^*$ then $\nabla f(x^*) = 0$
3. **Isolated** local minima are attractive : $(\exists \delta > 0)(\forall x^0 \in \mathbb{B}(x^*, \delta)) \quad x^k \to x^*$

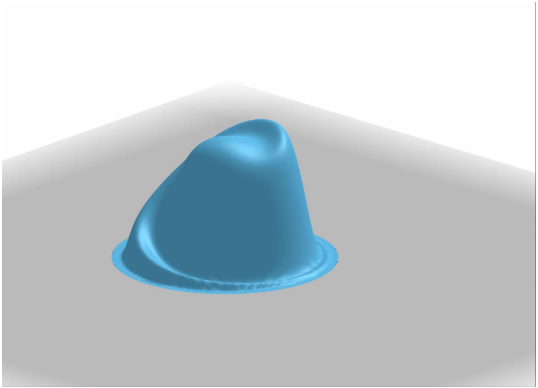
- x^k can have **no limit** !
- No convergence $\neq$ lack of regularity, it's a matter of **wilderness**

[Pro 1.2.3, 1.2.5 & Ex. 1.2.18] Bertsekas, Nonlinear Programming, 1999.

General results

$f : \mathbb{R}^N \longrightarrow \mathbb{R}$ is of class $C_L^{1,1}$

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$



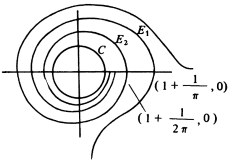


Figure 7

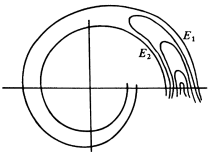


Figure 8

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

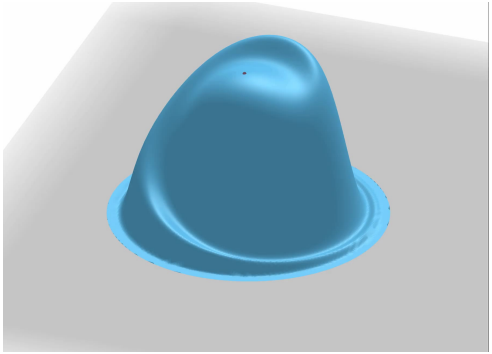
$$f(r\cos\theta, r\sin\theta) = \begin{cases} e^{1/(r^2-1)}, & \text{if } r < 1; \\ 0, & \text{if } r = 1; \\ e^{-1/(r^2-1)}\sin(1/(r-1)-\theta), & \text{if } r > 1. \end{cases}$$

[Ex. 3] Palis, de Melo, Geometric Theory of Dynamical Systems: An Introduction, 1982.
H.B.Curry, The method of steepest descent for nonlinear minimization problems, 1944.

General results

$f : \mathbb{R}^N \longrightarrow \mathbb{R}$ is of class $C_L^{1,1}$

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$



[Ex. 3] Palis, de Melo, Geometric Theory of Dynamical Systems: An Introduction, 1982.
H.B.Curry, The method of steepest descent for nonlinear minimization problems, 1944.

We need to do an **hypothesis** :

1. this hypothesis must guarantee the convergence of the dynamic
2. this hypothesis must be satisfied by most day-to-day functions

How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \dot{x}(t)\ \mathrm{d}t < +\infty$

How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \dot{x}(t)\ \mathrm{d}t < +\infty$
  - It is a classic result that "Finite Length" implies convergence
  - Converse is not true (but tricky):

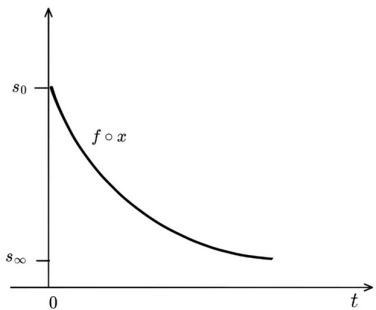
$$x^k := \sum_{n=1}^{k} \frac{(-1)^n}{n} \longrightarrow -\ln(2) \text{ but } \sum_{k=1}^{\infty} \|x^{k+1} - x^k\| = \sum_{k=1}^{\infty} \frac{1}{k} = +\infty.$$

How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \dot{x}(t)\ \mathrm{d}t < +\infty$
- Length is invariant up to a reparametrization in time

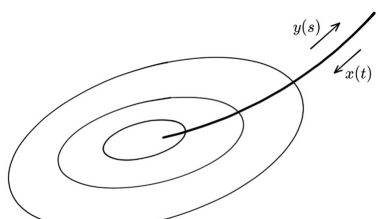
How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \dot{x}(t)\,\mathrm{d}t < +\infty$
- Length is invariant up to a reparametrization in time
- We have a natural diffeomorphism $f \circ x : [0, +\infty[\longrightarrow]s_\infty, s_0]$, where $s_0 = f(x(0))$ and $s_\infty = \lim_{t \to +\infty} f(x(t))$



How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \dot{x}(t)\,\mathrm{d}t < +\infty$
- Length is invariant up to a reparametrization in time
- We have a natural diffeomorphism $f \circ x : [0, +\infty[\longrightarrow]s_\infty, s_0]$, where $s_0 = f(x(0))$ and $s_\infty = \lim_{t \to +\infty} f(x(t))$
- With $s := f(x(t))$ we can define $y(s) := x((f \circ x)^{-1}(s))$ such that $\dot{y}(s) = \frac{\nabla f(y(s))}{\|\nabla f(y(s))\|^2}$



How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \dot{x}(t)\ \mathrm{d}t < +\infty$
- Length is invariant up to a reparametrization in time
- We have a natural diffeomorphism $f \circ x : [0, +\infty[\longrightarrow]s_\infty, s_0]$, where $s_0 = f(x(0))$ and $s_\infty = \lim_{t \to +\infty} f(x(t))$
- With $s := f(x(t))$ we can define $y(s) := x((f \circ x)^{-1}(s))$ such that $\dot{y}(s) = \frac{\nabla f(y(s))}{\|\nabla f(y(s))\|^2}$
- So this length becomes $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|}\ \mathrm{d}t$

How to guarantee convergence?

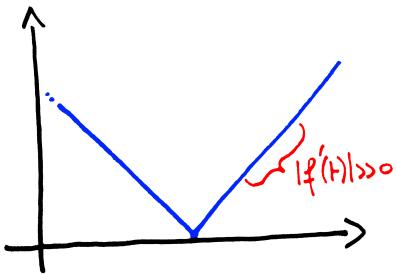
- A sufficient condition for $x(t)$ to converge is $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, \mathrm{d}t < +\infty$

How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, \mathrm{d}t < +\infty$
- How to upper bound this?

How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, \mathrm{d}t < +\infty$
- How to upper bound this?
- Naive hypothesis : $\|\nabla f(y)\| \geqslant C$ i.e. f is *sharp*

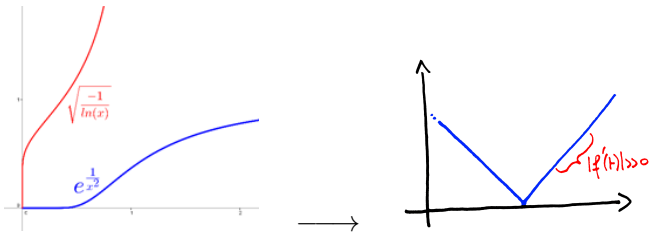


How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, \mathrm{d}t < +\infty$
- How to upper bound this?
- Naive hypothesis : $\|\nabla f(y)\| \geqslant C$ i.e. f is *sharp*
- 'Smart' hypothesis : $\frac{1}{\|\nabla f(y(s))\|} \leqslant \varphi'(s)$ so that $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, \mathrm{d}t \leqslant \varphi(s_0) - \varphi(s_\infty)$

How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, \mathrm{d}t < +\infty$
- How to upper bound this?
- Naive hypothesis : $\|\nabla f(y)\| \geqslant C$ i.e. f is *sharp*
- 'Smart' hypothesis : $\frac{1}{\|\nabla f(y(s))\|} \leqslant \varphi'(s)$ so that $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, \mathrm{d}t \leqslant \varphi(s_0) - \varphi(s_\infty)$
- In other words we need $\varphi'(f(x(t)))\|\nabla f(x(t))\| \geqslant 1$, i.e. $\|\nabla(\varphi \circ f)(x)\| \geqslant 1$



I : Convergence of the methods

2 : The Lojasiewicz inequality

Definition

We say that $f \in C^1(\mathbb{R}^N)$ is **Lojasiewicz** at a critical point x^* if

$$\varphi'(f(x) - f(x^*))\|\nabla f(x)\| \geqslant 1$$

- for all x such that $\|x - x^*\| < \delta$, $f(x) - f(x^*) < r$
- for some $\varphi : [0, +\infty[\to [0, +\infty[$ such that $\varphi(0) = 0$, increasing, concave (e.g. $\sqrt[p]{t}$)

Definition

- f is Lojasiewicz if it is Lojasiewicz at every critical point
- f is p-Lojasiewicz if at every critical point it is Lojasiewicz with $\varphi(t) \sim \sqrt[p]{t}$:

$$(\exists \mu > 0) \quad \mu(f(x) - f(x^*))^{p-1} \leqslant \frac{1}{p}\|\nabla f(x)\|^p$$

Convergence

Theorem (Convergence)

Let $f \in C_L^{1,1}(\mathbb{R}^N)$ be **Lojasiewicz**, and $\lambda \in]0, \frac{2}{L}[$.
If x^k is bounded, then it converges to some x^* such that $\nabla f(x^*) = 0$.

Theorem (Capture)

Let $f \in C_L^{1,1}(\mathbb{R}^N)$ be **Lojasiewicz**, and $\lambda \in]0, \frac{2}{L}[$.
For each $x^* \in \operatorname{argmin} f$, there exists $\delta > 0$ such that $x^0 \in \mathbb{B}(x^*; \delta)$ guarantees that it converges to some $x^\infty \in \operatorname{argmin} f$.

Łojasiewicz. Sur les trajectoires du gradient d'une fonction analytique, 1984.
Absil, Mahony, Andrews. Convergence of the Iterates of Descent Methods for Analytic Cost Functions, 2005.

Convergence

Proof: We wanna show something like $\varphi'(s) \geqslant \|\dot{x}(t)\|$, with $s = f(x(t))$

$$
\begin{aligned}
& \varphi(f(x^k) - \inf f) - \varphi(f(x^{k+1}) - \inf f) & \\
\geqslant\ & \varphi'(f(x^k) - \inf f)(f(x^k) - f(x^{k+1})) & \text{because } \varphi \text{ concave} \\
\geqslant\ & \varphi'(f(x^k) - \inf f) c_{\lambda,L} \|x^{k+1} - x^k\|^2 & \text{with Descent Lemma} \\
\geqslant\ & \varphi'(f(x^k) - \inf f) C_{\lambda,L} \|\nabla f(x^k)\| \|x^{k+1} - x^k\| & \\
\geqslant\ & 1 \cdot C_{\lambda,L} \|x^{k+1} - x^k\| & \text{with}
\end{aligned}
$$

So $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\| \leqslant C_{\lambda,L}^{-1} \varphi(f(x^0) - \inf f) < +\infty$.

The sequence has finite length, therefore it converges.

Rates for free

Theorem (Rates for $p = 2$, Gradient Descent)

Let f be *globally* 2-Lojasiewicz : $(\forall x \in \mathbb{R}^N) \quad \mu(f(x) - \inf f) \leqslant \frac{1}{2}\|\nabla f(x)\|^2$.
Then we have linear convergence : $f(x^k) - \inf f = O(\theta^{2k})$ with $\theta \in [0, 1[$.
If $\lambda = \frac{1}{L}$, we have $\theta = \sqrt{1 - \frac{\mu}{L}}$.

- With strong convexity we had a better $\theta = 1 - \frac{\mu}{L}$
- Global Lojasiewicz is a lot to ask (every critical point is a minimizer)
- Rates become asymptotic if local Łojasiewicz only

Polyak, Gradient methods for the minimisation of functionals, 1963. [Theorem 4]

Rates for free

Theorem (Rates for $p > 2$, Gradient Descent)

Let f be *globally* p-Lojasiewicz : $(\forall x \in \mathbb{R}^N) \quad \mu(f(x) - \inf f)^{p-1} \leqslant \frac{1}{p}\|\nabla f(x)\|^p$.
Then we have sublinear convergence : $f(x^k) - \inf f = O(\frac{1}{k^{\frac{p}{p-2}}})$.

- $\frac{p}{p-2} \to +\infty$ when $p \downarrow 2$; $\frac{p}{p-2} \to 1$ when $p \uparrow +\infty$
- Rates are matched when $f(x) = |x|^p$
- Same remark about local/global Lojasiewicz

Attouch, Bolte, On the convergence of the proximal algorithm for nonsmooth functions [...], 2009.
Chouzenoux, Pesquet, Repetti, A block coordinate variable metric forward-backward algorithm, 2014.

Proof: (Case $p = 2$)

$$
\begin{array}{rcll}
 & f(x^k) - f(x^{k+1}) & & \\
\geqslant & \left(\frac{1}{\lambda} - \frac{L}{2}\right) \|x^{k+1} - x^k\|^2 & \text{with Descent Lemma} \\
= & \left(\frac{1}{\lambda} - \frac{L}{2}\right) \lambda^2 \|\nabla f(x^k)\|^2 & \text{from the definition of } x^{k+1} \\
\geqslant & \left(\frac{1}{\lambda} - \frac{L}{2}\right) \lambda^2 2\mu (f(x^k) - \inf f) & \text{with Lojasiewicz}
\end{array}
$$

So $f(x^{k+1}) - \inf f \leqslant \theta^2 (f(x^k) - \inf f)$ with

$$
\begin{aligned}
\theta^2 &= 1 - \left(\frac{1}{\lambda} - \frac{L}{2}\right) \lambda^2 2\mu \\
&= 1 - 2\lambda\mu + L\lambda^2\mu \in [0, 1[\text{ if } \lambda \in]0, \frac{2}{L}[\\
&= 1 - \frac{\mu}{L} \text{ if } \lambda = \frac{1}{L}
\end{aligned}
$$

Proof: (Case $p > 2$)

$$
\begin{aligned}
& f(x^k) - f(x^{k+1}) & \\
\geqslant \quad & \left(\tfrac{1}{\lambda} - \tfrac{L}{2}\right) \|x^{k+1} - x^k\|^2 & \text{with Descent Lemma} \\
= \quad & \left(\tfrac{1}{\lambda} - \tfrac{L}{2}\right) \lambda^2 \|\nabla f(x^k)\|^2 & \text{from the definition of } x^{k+1} \\
\geqslant \quad & \left(\tfrac{1}{\lambda} - \tfrac{L}{2}\right) \lambda^2 (p\mu(f(x^k) - \inf f))^{2\frac{p-1}{p}} & \text{with Lojasiewicz}
\end{aligned}
$$

So $f(x^{k+1}) - \inf f \leqslant (f(x^k) - \inf f) - C_{p,\mu,\lambda}(f(x^k) - \inf f)^\alpha$ with $\alpha = 2\frac{p-1}{p} \in]1, 2[$.

It is a not-so-easy[1] exercise to prove that this inequality implies

$$f(x^k) - \inf f = O\left(\frac{1}{k^{\frac{1}{\alpha-1}}}\right)$$

which is exactly what we need.

[1] Li, Mordukhovich, Hölder Metric Subregularity with Applications to Proximal Point Method, 2012. [Lemma 7.1]

Works also for SGD

Theorem (Complexity for $p = 2$, Stochastic Gradient Descent)

Let f be *globally* 2-Lojasiewicz and $\mathcal{L}$-smooth with respect to the sampling. Assume that λ is small enough. Then

$$\mathbb{E}\left[f(x^k) - \inf f\right] \leqslant (1 - \mu\lambda)^k (f(x^0) - \inf f) + \lambda c \Delta^*,$$

where $\Delta^* = \inf f - \sum_{i=1}^m \inf f_i$, $c > 0$.

- Δ^* can be see as a variance term, like σ^* (actually $\sigma^* \leqslant 2\mathcal{L}\Delta^*$)
- We can guarantee a $O(1/k)$ rate by taking $\lambda_k = O(1/k)$.
- Complexity is also like in the strongly convex case.
- Global Lojasiewicz is a lot to ask (every critical point is a minimizer)

Gower, Sebbouh, and Loizou, *SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation*, 2020.

We need to do an **hypothesis** :

1. this hypothesis must guarantee the convergence of the dynamic 
2. this hypothesis must be satisfied by most day-to-day functions

I : Convergence of the methods

3 : The Lojasiewicz property in practice

Big classes of Lojasiewicz functions

Theorem

1. Every **analytic** function is p-Lojasiewicz at its critical points
2. Every **semi-algebraic** function is p-Lojasiewicz at its critical points
3. Every **o-minimal** function is Lojasiewicz

Łojasiewicz, Ensembles semi-analytiques, 1965.
Kurdyka, On gradients of functions definable in o-minimal structures, 1998.
Bolte, Daniilidis, Lewis, Shiota, Clarke Subgradients of Stratifiable Functions, 2007.

Semi-algebraic functions

Definition

A function $F : \mathbb{R}^N \longrightarrow \mathbb{R}^M$ is **semi-algebraic** if its graph can be described with a finite number of polynomial (in)equalities:

$$graph(f) = \{z \in \mathbb{R}^N \times \mathbb{R}^M \mid P_1(z) \leqslant 0, \ldots, P_r(z) \leqslant 0\}$$

Coste, An Introduction to O-minimal Geometry, 2000.

Semi-algebraic functions

Definition

A function $F : \mathbb{R}^N \longrightarrow \mathbb{R}^M$ is **semi-algebraic** if its graph can be described with a finite number of polynomial (in)equalities:

$$graph(f) = \{z \in \mathbb{R}^N \times \mathbb{R}^M \mid P_1(z) \leqslant 0, \dots, P_r(z) \leqslant 0\}$$

Example (semi-algebraic functions)

- Polynomials : Linear maps, quadratic functions $f(x) = \|\Phi x - y\|^2$
- Polynomials by parts : $f(x) = \|x\|_1$, ReLU

Coste, An Introduction to O-minimal Geometry, 2000.

Semi-algebraic functions

Theorem (Tarski-Seidenberg)

The class of semi-algebraic functions is stable under:

- addition, multiplication, division, sup, inf
- restriction, composition, inverse f^{-1}
- derivative

Semi-algebraic functions

Theorem (Tarski-Seidenberg)

The class of semi-algebraic functions is stable under:

- addition, multiplication, division, sup, inf
- restriction, composition, inverse f^{-1}
- derivative

Example (semi-algebraic functions)

- Lasso : $f(x) = \|x\|_1 + \|\Phi x - y\|^2$
- Neural Networks with ReLU activation : $f(x) = (\sigma \circ A_\ell \circ \cdots \circ \sigma \circ A_1)(x)$

Semi-algebraic functions

Example (Counter-example)

Exponential/logarithmic stuff isn't semi-algebraic

Semi-algebraic functions

Example (Counter-example)

Exponential/logarithmic stuff isn't semi-algebraic

Theorem

There exists a class of functions (**o-minimal** structure) which:

- includes the semi-algebraic structure
- contains the exponential function
- has the same stability properties as in the Tarski-Seidenberg theorem (sum, etc.)
- is also stable by integration

Speissegger, The Pfaffian closure of an o-minimal structure, 1999.

Big classes of Lojasiewicz functions

Theorem

1. Every **analytic** function is p-Lojasiewicz at its critical points
2. Every **semi-algebraic** function is p-Lojasiewicz at its critical points
3. Every **o-minimal** function is Lojasiewicz

Łojasiewicz, Ensembles semi-analytiques, 1965.
Kurdyka, On gradients of functions definable in o-minimal structures, 1998.
Bolte, Daniilidis, Lewis, Shiota, Clarke Subgradients of Stratifiable Functions, 2007.

A more reasonable local result

Theorem

Suppose that the functions are locally Lipschitz, and semi-algebraic (or o-minimal). Take vanishing stepsizes $\lambda_t \in \ell^2(\mathbb{N}) \setminus \ell^1(\mathbb{N})$.
If the trajectory $(x^t)_{t\in\mathbb{N}}$ is bounded a.s., then

1. $f(x^t)$ converges
2. Every limit point of the sequence is a critical point.

Rk: Convergence of the iterates? Unknown/unaware.

Davis, Drusvyatskiy, Kakade, Lee, *Stochastic Subgradient Method Converges on Tame Functions*, 2020.
Bolte, Pauwels, *Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning*, 2021.

A more reasonable local result

We have similar (less good) results about inertial methods

Theorem (Li et al., 2017)

If $f \in C_L^{1,1}(H)$ is locally p-Lojasiewicz at its critical points, and appopriate assumptions on the parameters, then both Heavy Ball and Nesterov's method converge to critical points.

Take-home message:

Virtually **any function** you can think about is Łojasiewicz, as long as it does not involve

$$\begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \sin(x) \end{array}$$

Take-home message:

Virtually **any function** you can think about is Łojasiewicz, as long as it does not involve

$$\begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \sin(x) \end{array}$$

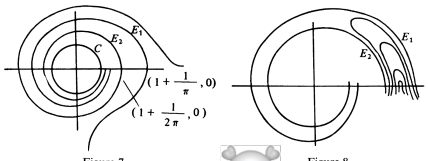
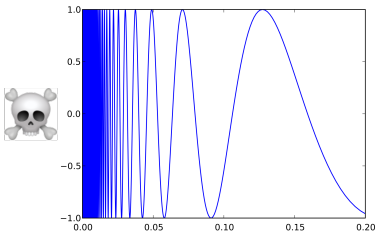


Figure 7

Figure 8

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(r\cos\theta, r\sin\theta) = \begin{cases} e^{1/(r^2-1)}, & \text{if } r < 1; \\ 0, & \text{if } r = 1; \\ e^{-1/(r^2-1)} \sin(1/(r-1) - \theta), & \text{if } r > 1. \end{cases}$$



Take-home message:

Virtually **any function** you can think about is Łojasiewicz, as long as it does not involve

$$\begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \sin(x) \end{array}$$

So gradient descent *"always converges"* to a critical point

What can we expect?

- Does my algorithm converge? $x_\infty := \lim_{k \to +\infty} x^k$ exists?

- What is the nature of the limit x_∞?
  Global/local minima? Saddle point?

II : Finding global minimizers

II : Finding global minimizers

1 : Avoiding saddle points

Let's consider the three likely outcomes for the limit point:

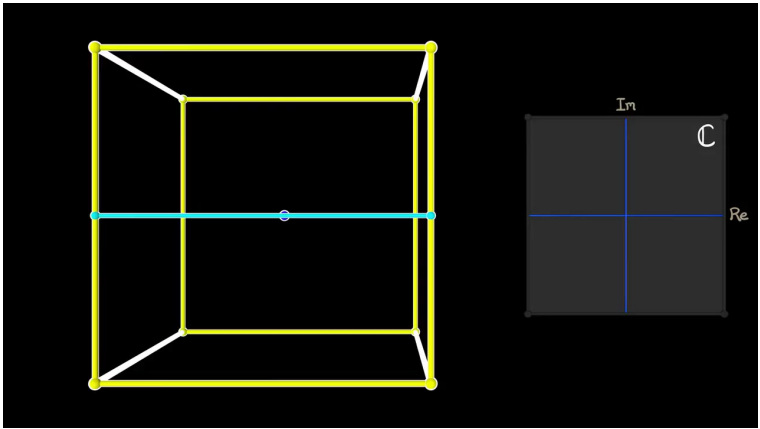
1. x^* is a global minimizer : $(\forall x \in \mathbb{R}^N) \quad f(x^*) \leqslant f(x)$
2. x^* is a local minimizer : $(\exists \delta > 0)(\forall x \in \mathbb{B}(x^*, \delta)) \quad f(x^*) \leqslant f(x)$
3. x^* is a saddle point : $(\forall \delta > 0)(\exists x_-, x^+ \in \mathbb{B}(x^*, \delta)) \quad f(x_-) \leqslant f(x^*) \leqslant f(x_+)$

Our algorithm is **local** so there no way we can differentiate 1. and 2.

The question is : can we avoid saddle points?

$f(x) = \frac{1}{2}\langle Ax, x\rangle$

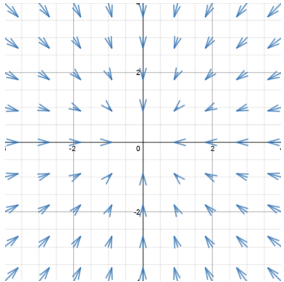
$\dot{x}(t) = -\nabla f(x(t))$



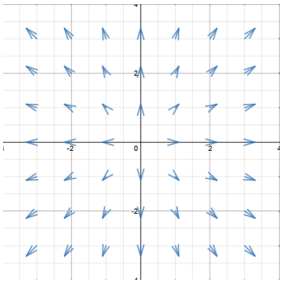
Stolen from Robert Ghrist's Twitter account @robertghrist

https://twitter.com/robertghrist/status/1185864562299539456

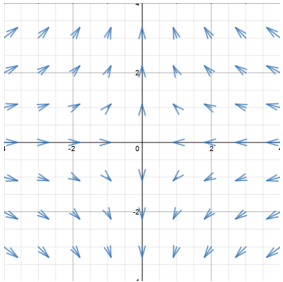
$f(x) = \frac{1}{2}\langle Ax, x\rangle$ $\dot{x}(t) = -\nabla f(x(t))$



$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$



$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

- Positive eigs. are attractive, negative eigs. are repulsive.
- Converging to the saddle point requires starting from $E_{+1}(A)$

$f(x) = \frac{1}{2}\langle Ax, x\rangle$ $\dot{x}(t) = -\nabla f(x(t))$

Definition

Let x^* be an equilibrium of the system. We define

$$W(x^*) := \{x \in \mathbb{R}^N \mid x(0) = x \quad \text{and} \quad \lim x(t) = x^*\}$$

$f(x) = \frac{1}{2}\langle Ax, x\rangle$ $\dot{x}(t) = -\nabla f(x(t))$

Definition

Let x^* be an equilibrium of the system. We define

$$W(x^*) := \{x \in \mathbb{R}^N \mid x(0) = x \quad \text{and} \quad \lim x(t) = x^*\}$$

Theorem

$$W(x^*) \simeq \bigoplus_{\lambda>0} E_\lambda(A)$$

Corollary

If $\lambda_{min}(A) < 0$, then $W(x^*)$ has Lebesgue measure 0

$f(x) \in C^2(\mathbb{R}^N)$ $\dot{x}(t) = -\nabla f(x(t))$

Theorem (Stable Manifold Lemma)

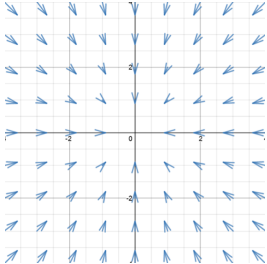
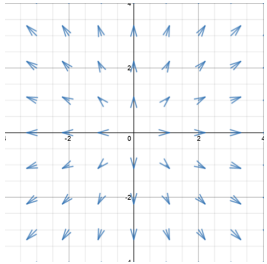
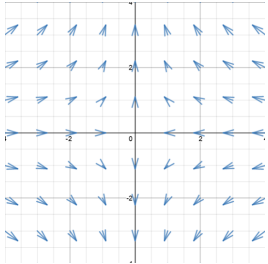
$W(x^*)$ is a submanifold of dimension smaller than the one of

$$\bigoplus_{\lambda>0} E_\lambda(\nabla^2 f(x^*))$$

Corollary

If $\lambda_{min}(\nabla^2 f(x^*)) < 0$, then $W(x^*)$ has Lebesgue measure 0

$f(x) \in C^2(\mathbb{R}^N)$ $\dot{x}(t) = -\nabla f(x(t))$

Rk: The 3 kinds of critical points:

- the local minima (e.g. $\lambda_{min}(\nabla^2 f(x^*)) > 0$) : **attractive**
- The strict saddles $\lambda_{min}(\nabla^2 f(x^*)) < 0$: **repulsive**
- the degenerated ones (in particular $\lambda_{min}(\nabla^2 f(x^*)) = 0$) : **complicated**

$f(x) \in C^{1,1}(\mathbb{R}^N)$ $\quad$ $x^{k+1} = x^k - \lambda \nabla f(x^k)$

Theorem

If $\lambda < \frac{1}{L}$ then $W(x^*)$ is a submanifold of dimension smaller than $\oplus_{\lambda>0} E_\lambda(\nabla^2 f(x^*))$

Corollary

If $\lambda_{min}(\nabla^2 f(x^*)) < 0$, then $W(x^*)$ has Lebesgue measure 0

Lee, Simchowitz, Jordan, Recht, Gradient Descent Converges to Minimizers, 2016.

$f(x) \in C^{1,1}(\mathbb{R}^N)$ $\quad$ $x^{k+1} = x^k - \lambda \nabla f(x^k)$

Theorem

If $\lambda < \frac{1}{L}$ then $W(x^*)$ is a submanifold of dimension smaller than $\oplus_{\lambda>0} E_\lambda(\nabla^2 f(x^*))$

Corollary

If $\lambda_{min}(\nabla^2 f(x^*)) < 0$, then $W(x^*)$ has Lebesgue measure 0

Corollary

If f has no degenerated critical points and is Lojasiewicz, then x^k converges a.s. to a local minima with random initialization

Lee, Simchowitz, Jordan, Recht, Gradient Descent Converges to Minimizers, 2016.

$f(x) \in C^2(\mathbb{R}^N)$ $\qquad$ $x^{t+1} = x^t - \lambda_t \nabla f_{i_t}(x^t) = x^t - \lambda_t(\nabla f(x^t) + \xi^t)$

What if the noise ξ^t maintains us on the attractive space $E_{\geqslant 0}(x^*)$?

Theorem

Assume that $\lambda_t \in \ell^2(\mathbb{N}) \setminus \ell^1(\mathbb{N})$, and that x^* is a strict saddle. Assume essentially that $\liminf \mathbb{E}\left[\|\text{proj}(\xi^t; E_{<0}(x^*))\| \mid x^t\right] > 0$. Then $\mathbb{P}(x^t \to x^*) = 0$.

Rk: Can be adapted if f non smooth (hard).
Similar results if η^t follows some isotropic law (but irrelevant for SGD?).

Pemantle, *Nonconvergence to unstable points in urn models and stochastic approximations*, 1990.
Brandière, Duflo, *Les algorithmes stochastiques contournent-ils les pieges?*, 1996.
Bianchi, Hachem, Schechtman, *Stochastic subgradient descent escapes active strict saddles*, 2021.

II : Finding global minimizers

2 : Degenerated saddles do not exist ?

The generic argument

Theorem

Generically, locally Lipschitz semi-algebraic functions have no degenerated saddles.

Davis, Drusvyatskiy, Jiang, *Subgradient methods near active manifolds: saddle point avoidance, [...]*, 2021. Corollary 3.2.2.

The generic argument

Theorem

Generically, locally Lipschitz semi-algebraic functions have no degenerated saddles.

- Devil's in the details : given f, for a.e. $v \in \mathbb{R}^p$, the tilted function $f - \langle v, \cdot \rangle$ has no degenerated saddles
- Those functions also have a finite number of local min.
- Could be ok for underparametrized NN, not true for overparametrized NN (see later)

Davis, Drusvyatskiy, Jiang, *Subgradient methods near active manifolds: saddle point avoidance, [...]*, 2021. Corollary 3.2.2.

The generic argument

Theorem

Generically, locally Lipschitz semi-algebraic functions have no degenerated saddles.

Example

The matrix factorization problem a.k.a. two-layer-linear-neural-network

$$\min_{X \in \mathbb{R}^{d \times r}} f(X) = \|X^\top X - A\|_F^2$$

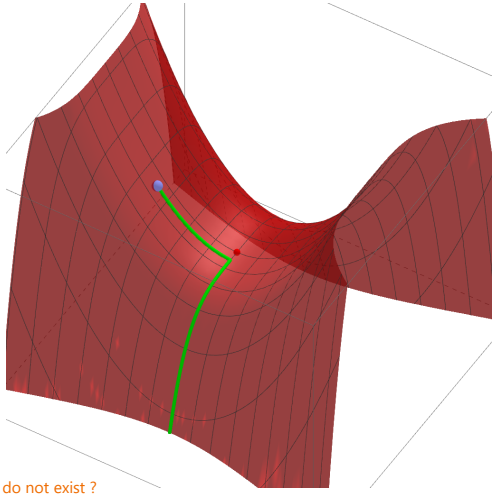
has no degenerated saddles.

Davis, Drusvyatskiy, Jiang, *Subgradient methods near active manifolds: saddle point avoidance, [...]*, 2021. Corollary 3.2.2.

$f(x) \in C^2(\mathbb{R}^N)$

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

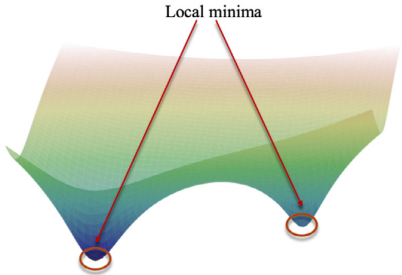
It is time now for some examples



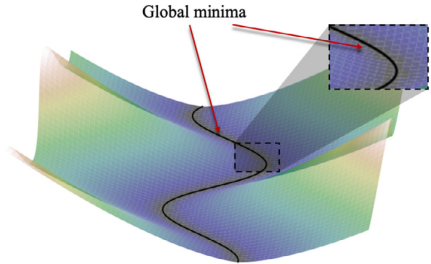
II : Finding global minimizers

3 : The landscape of overparametrized Neural Networks

People try to prove "overparametrized $\Rightarrow$ no spurious local minima"



(a) Loss landscape of under-parameterized models



(b) Loss landscape of over-parameterized models

Liu, Zhu, Belkin, *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks*, 2022.

Notations

$$\min_{w\in\mathbb{R}^p} f(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w), \quad f_i(w) := \ell(\phi(w;x_i);y_i) = \frac{1}{2}(\phi(w;x_i)-y_i)^2.$$

Neural Network : $\phi(w;x_i) = (\sigma_L \circ W_L \circ \dots \sigma_1 \circ W_1)(x_i)$, with $\sigma_L = id$, and **width** m.

We note $\phi_i(w) := \phi(w;x_i)$, $\Phi(w) := (\phi_i(w))_{i=1}^n$, so we want $\Phi(w) \simeq y$

- Optimization : $\nabla f(w) \in \mathbb{R}^p$, $\nabla^2 f(w) \in \mathbb{R}^{p,p}$
- Neural Network : $D\Phi(w) \in \mathbb{R}^{n,p}$, $D^2\Phi(w) = (\nabla^2\phi_i(w))_{i=1}^n \in \mathbb{R}^{n,p,p}$
- **Tangent Kernel** : $K(w) = D\Phi(w)D\Phi(w)^\top \in \mathbb{R}^{n,n}$

Assumption (Regular NN)

Φ is L_Φ-Lipschitz continuous and twice differentiable (sigmoid, tanh)

Wide NN are almost linear

Theorem (Transition to linearity)

Let Φ be the NN described before, and $\bar{w} \sim \mathcal{N}(0,1)$. With high probability:

$$(\forall R > 0)(\forall w \in \mathbb{B}(\bar{w}, R)), \text{ with high probability } \|D^2\Phi(w)\|_{\infty,2,2} = O\left(\frac{R^{3L}}{\sqrt{m}}\right)$$

- When the width $m \to +\infty$ we roughly have $D^2\Phi(w) \equiv 0$, meaning Φ is linear.
- Φ linear means $f(w) = \|\Phi(w) - y\|^2$ is quadratic $\Rightarrow$ convex and 2-Lojasiewicz

Liu, Zhu, Belkin, *On the linearity of large non-linear models: when and why the tangent kernel is constant*, 2021. Theorem 3.2.
Liu, Zhu, Belkin, *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks*, 2022. Theorem 5.

Uniformly elliptic Tangent Kernel $\Rightarrow$ 2-Lojasiewicz loss

Theorem

Let $\Omega \subset \mathbb{R}^p$ such that $\lambda_{min}(K(w)) \geqslant \mu > 0$ on Ω, where $K(w) = D\Phi(w)D\Phi(w)^\top$.
Then f is 2-Lojasiewicz on Ω :

$$(\forall w \in \Omega) \quad \mu(f(w) - \inf f) \leqslant \frac{1}{2}\|\nabla f(w)\|^2.$$

Rk: 2-Lojasiewicz is called Polyak-Lojasiewicz in the ML literature (+ subtleties)

Proof: We have $f(w) = (1/2)\|\Phi(w) - y\|^2$ so

$$\begin{aligned} \|\nabla f(w)\|^2 &= \|D\Phi(w)^\top(\Phi(w) - y)\|^2 = \langle K(w)(\Phi(w) - y), (\Phi(w) - y)\rangle \\ &\geqslant \mu\|\Phi(w) - y)\|^2 = 2\mu f(w) \geqslant 2\mu(f(w) - \inf f). \end{aligned}$$

Liu, Zhu, Belkin, *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks*, 2022. Theorem 1.

Wide NN $\Rightarrow$ 2-Lojasiewicz on balls

Theorem

Let $\bar{w} \sim \mathcal{N}(0,1)$, and suppose that $\lambda_{min}(K(\bar{w})) \geqslant \bar{\mu} > 0$. For all $\mu < \bar{\mu}$, $R > 0$:

$$\text{if } m = O\left(\frac{L_\Phi^2 n R^{6L+2}}{(\bar{\mu}-\mu)^2}\right), \text{ then } f \text{ is } 2\text{-Lojasiewicz on } \mathbb{B}(\bar{w}, R).$$

Wide NN $\Rightarrow$ 2-Lojasiewicz on balls

Theorem

Let $\bar{w} \sim \mathcal{N}(0,1)$, and suppose that $\lambda_{min}(K(\bar{w})) \geqslant \bar{\mu} > 0$. For all $\mu < \bar{\mu}$, $R > 0$:

$$\text{if } m = O\left(\frac{L_\Phi^2 n R^{6L+2}}{(\bar{\mu}-\mu)^2}\right), \text{ then } f \text{ is } 2\text{-Lojasiewicz on } \mathbb{B}(\bar{w}, R).$$

- Hiding a **lot** of details here
- $K(\bar{w}) \succ 0$ is true for infinite networks, so true with h.p. for large ones.
- A bounded sequence of iterates could fit inside $\mathbb{B}(\bar{w}, R)$.

Liu, Zhu, Belkin, *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks*, 2022. Theorem 4.

Wide NN $\Rightarrow$ 2-Lojasiewicz on balls

Theorem

Let $\bar{w} \sim \mathcal{N}(0,1)$, and suppose that $\lambda_{min}(K(\bar{w})) \geqslant \bar{\mu} > 0$. For all $\mu < \bar{\mu}$, $R > 0$:

$$\text{if } m = O\left(\frac{L_\Phi^2 n R^{6L+2}}{(\bar{\mu}-\mu)^2}\right), \text{ then } f \text{ is } 2\text{-Lojasiewicz on } \mathbb{B}(\bar{w}, R).$$

Proof: For $w \in \mathbb{B}(\bar{w}, R)$ we have

$$|\lambda_{min}(K(w)) - \lambda_{min}(K(\bar{w}))| \leqslant \|K(w) - K(\bar{w})\| \leqslant 2L_\Phi \|D\Phi(w) - D\Phi(\bar{w})\|$$

$$\leqslant \quad 2L_\Phi\sqrt{n} \sup_{\mathbb{B}(\bar{w},R)} \|D^2\Phi(\cdot)\|_{\infty,2,2} \, \|w - \bar{w}\| \lesssim 2L_\Phi\sqrt{n} \, \frac{R^{3L}}{\sqrt{m}} \, R \simeq \bar{\mu} - \mu.$$

So $\lambda_{min}(K(w)) \geqslant \lambda_{min}(K(\bar{w})) - (\bar{\mu} - \mu) \geqslant \mu$.

Convergence of algorithms : GD

Theorem 6 *(Local PL* condition $\Rightarrow$ existence of a solution + fast convergence). Suppose the system $\mathcal{F}$ is $L_{\mathcal{F}}$-Lipschitz continuous and $\beta_{\mathcal{F}}$-smooth. Suppose the square loss $\mathcal{L}(\mathbf{w})$ satisfies the μ-PL* condition in the ball $B(\mathbf{w}_0, R) := \{\mathbf{w} \in \mathbb{R}^m : \|\mathbf{w} - \mathbf{w}_0\| \leq R\}$ with $R = \frac{2L_{\mathcal{F}}\|\mathcal{F}(\mathbf{w}_0)-\mathbf{y}\|}{\mu}$. Then we have the following:*

(a) Existence of a solution: There exists a solution (global minimizer of $\mathcal{L}$) $\mathbf{w}^* \in B(\mathbf{w}_0, R)$, such that $\mathcal{F}(\mathbf{w}^*) = \mathbf{y}$.

(b) Convergence of GD: Gradient descent with a step size $\eta \leq 1/(L_{\mathcal{F}}^2 + \beta_{\mathcal{F}}\|\mathcal{F}(\mathbf{w}_0) - \mathbf{y}\|)$ converges to a global solution in $B(\mathbf{w}_0, R)$, with an exponential (a.k.a. linear) convergence rate:

$$\mathcal{L}(\mathbf{w}_t) \leq \left(1 - \kappa_{\mathcal{F}}^{-1}(B(\mathbf{w}_0, R))\right)^t \mathcal{L}(\mathbf{w}_0), \tag{28}$$

where the condition number $\kappa_{\mathcal{F}}(B(\mathbf{w}_0, R)) = \frac{1}{\eta\mu}$.

Convergence of algorithms : SGD

Theorem 7. *Given $0 < \delta < 1$, assume each $\ell_i(\mathbf{w})$ is β-smooth and $\mathcal{L}(\mathbf{w})$ satisfies the μ-PL* condition in the ball $B(\mathbf{w}_0, R)$ with $R = \frac{2n\sqrt{2\beta\mathcal{L}(\mathbf{w}_0)}}{\mu\delta}$. Then, with probability $1-\delta$, SGD with mini-batch size $s \in \mathbb{N}$ and step size $\eta \leq \frac{n\mu}{n\beta(n^2\beta+\mu(s-1))}$ converges to a global solution in the ball $B(\mathbf{w}_0, R)$, with an exponential convergence rate:*

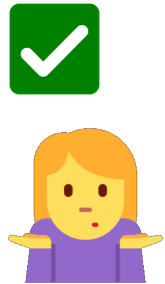
$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t)] \leq \left(1 - \frac{\mu s\eta}{n}\right)^t \mathcal{L}(\mathbf{w}_0). \tag{29}$$

Convergence with fixed stepsize because model is overparametrized : no variance term.

What can we expect?

- Does my algorithm converge? $x_\infty := \lim_{k \to +\infty} x^k$ exists?
- What is the nature of the limit x_∞?
  Global/local minima? Saddle point?
  Depends strongly on:
  - What your problem is
  - how you initialize

  No good answers in general!
- (Absurdly) Large Neural Networks enjoy nice geometries, need to bridge the gap with practice



End of the Chapter IV. End of the class.