Tutorial on Gradient methods for non-convex problems

Part 1

Guillaume Garrigos – November 28th – ENS

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?

General results

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Proposition

Let $0 \ll \lambda_k \ll 2/L$, then:

1) $f\left(x^k\right)$ is decreasing
2) if $x^{k_n} \to x_\infty$ then $\nabla f(x_\infty) = 0$
3) **Isolated** local minima are attractive

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

General results

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $C_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Proposition

Let $0 \ll \lambda_k \ll 2/L$, then:

1) $f\left(x^k\right)$ is decreasing
2) if $x^{k_n} \to x_\infty$ then $\nabla f(x_\infty) = 0$
3) **Isolated** local minima are attractive

x^k can have no limit !!

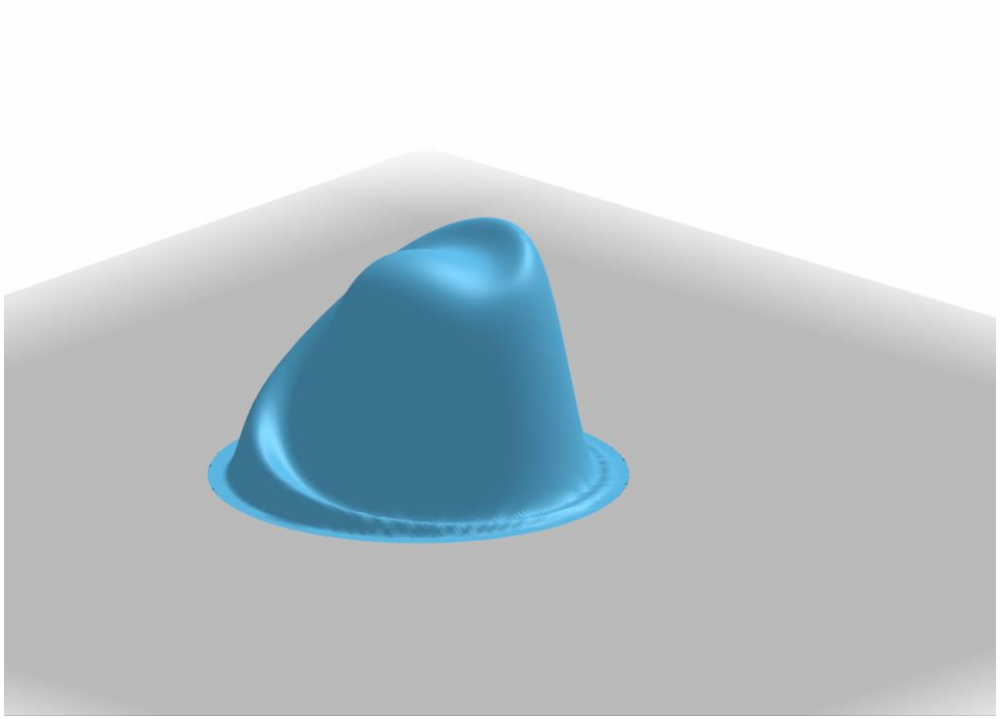
No convergence ≠ Lack of regularity, but rather **wildness**

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

General results

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$



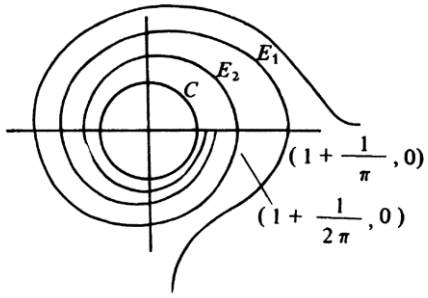


Figure 7

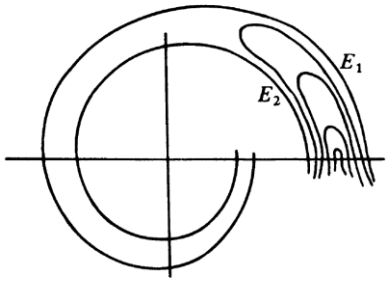


Figure 8

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

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

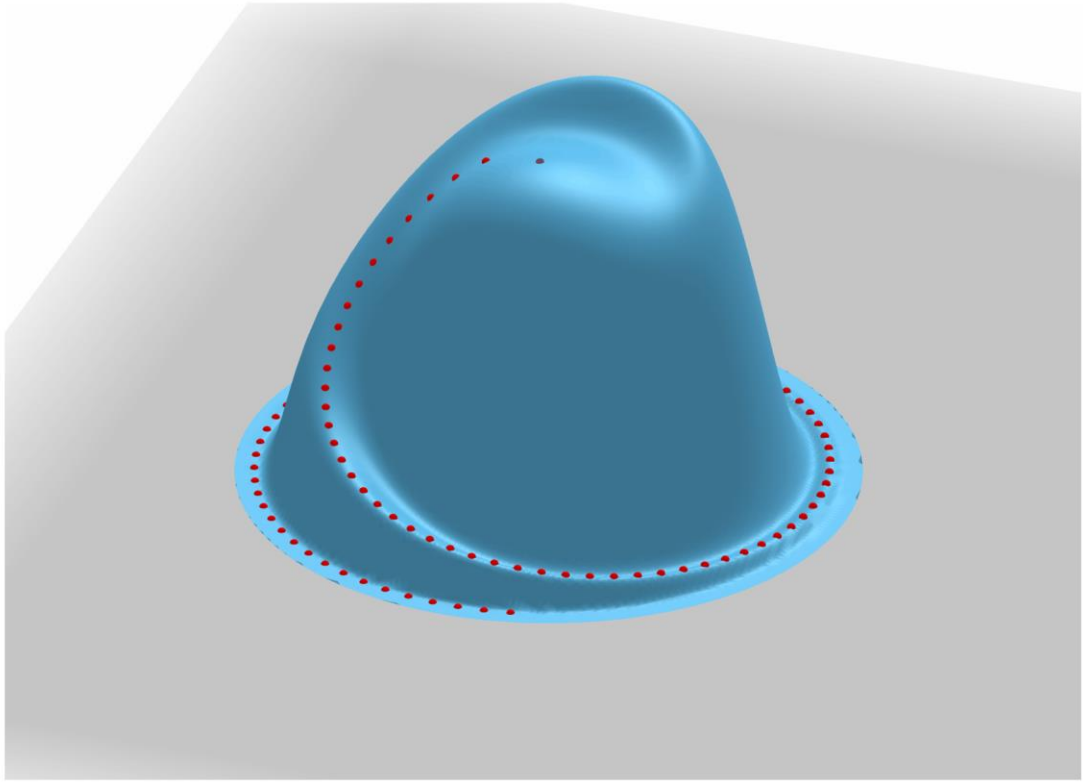
[Ex. 3] Palis, de Melo, Geometric Theory of Dynamical Systems: An Introduction, 1982.
H.B.Curry, The method of steepest descent for nonlinear minimization problems, 1944.

General results

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

$$\dot{x}(t) + \nabla f\left(x(t)\right) = 0$$



[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.

How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \|\dot{x}(t)\|\, dt < \infty$

How to guarantee convergence?

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

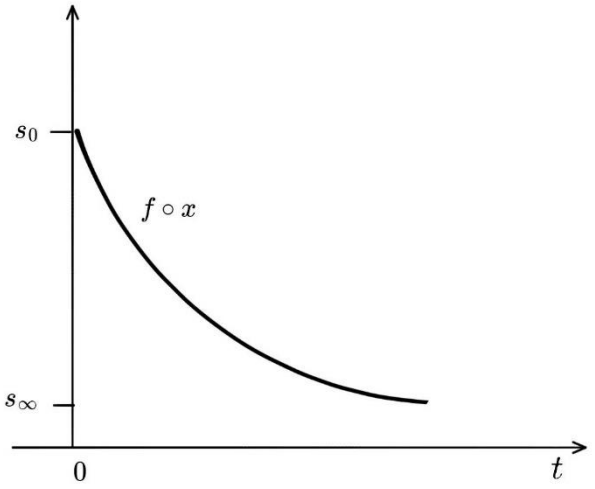
$$x^n := \sum_{k=1}^{n} \frac{(-1)^k}{k} \to -\log(2) \text{ but } \sum \left| x^{n+1} - x^n \right| = \sum \frac{1}{n}$$

How to guarantee convergence?

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

How to guarantee convergence?

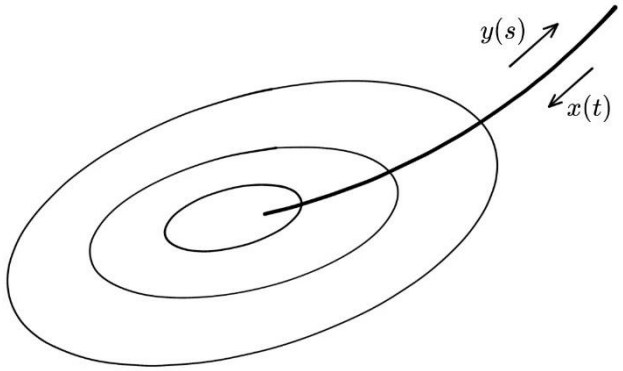
- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \|\dot{x}(t)\|\, dt < \infty$
- Length is invariant up to a reparametrization in time
- We have a natural diffeomorphism $f \circ x : [0, \infty[\to]s_\infty, s_0]$ where $s_0 = f(x_0)$ and $s_\infty = \lim_\infty f(x(t))$



How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \|\dot{x}(t)\|\, dt < \infty$
- Length is invariant up to a reparametrization in time
- We have a natural diffeomorphism $f \circ x : [0, \infty[\to]s_\infty, s_0]$ where $s_0 = f(x_0)$ and $s_\infty = \lim_\infty f(x(t))$
- With $s = f(x(t))$ we can define $y(s) = x\big((f \circ x)^{-1}(s)\big)$ s.t.

$$\dot{y}(s) = \nabla f\big(y(s)\big)\big\|\nabla f\big(y(s)\big)\big\|^{-2}$$



How to guarantee convergence?

- A sufficient condition for $x(t)$ to converge is $\int_0^\infty \|\dot{x}(t)\|\, dt < \infty$
- Length is invariant up to a reparametrization in time
- We have a natural diffeomorphism $f \circ x : [0, \infty[\to]s_\infty, s_0]$ where $s_0 = f(x_0)$ and $s_\infty = \lim_\infty f(x(t))$
- With $s = f(x(t))$ we can define $y(s) = x\big((f \circ x)^{-1}(s)\big)$ s.t.

$$\dot{y}(s) = \nabla f\big(y(s)\big)\big\|\nabla f\big(y(s)\big)\big\|^{-2}$$

- So the length becomes $\int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|}\, ds$

Finite interval !

Ignore $\nabla f\big(y(s)\big) = 0$

How to guarantee convergence?

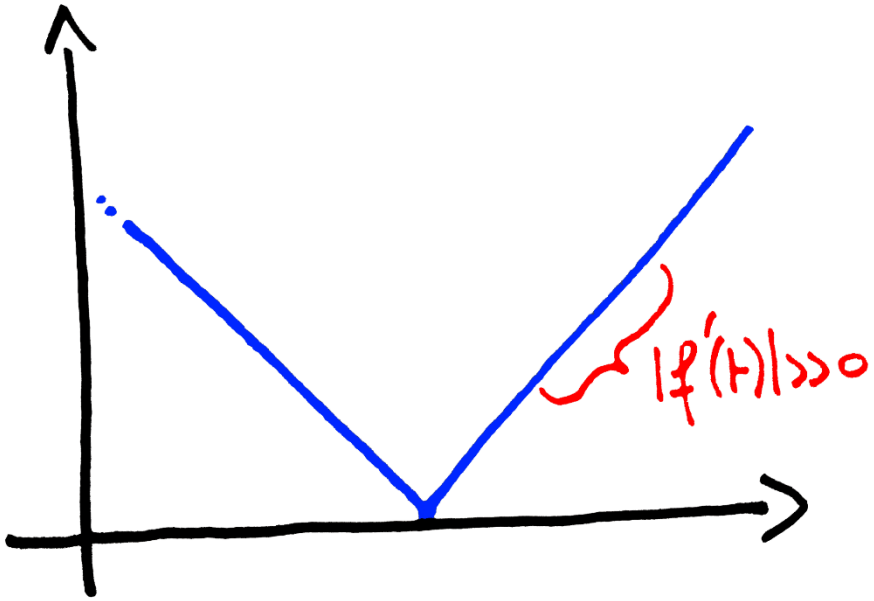
- How to upper bound $\int_0^{\infty} \|\dot{x}(t)\|\, dt = \int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|}\, ds$?

How to guarantee convergence?

- How to upper bound $\int_0^{\infty} \|\dot{x}(t)\| \, dt = \int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|} \, ds$?
- ``Naive'' hypothesis: $\|\nabla f(y)\| \geq C$ i.e. sharpness

How to guarantee convergence?

- How to upper bound $\int_0^{\infty} \|\dot{x}(t)\|\, dt = \int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|}\, ds$?
- ``Naive'' hypothesis: $\|\nabla f(y)\| \geq C$ i.e. sharpness



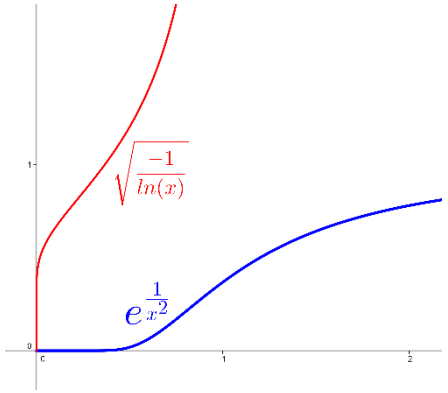
How to guarantee convergence?

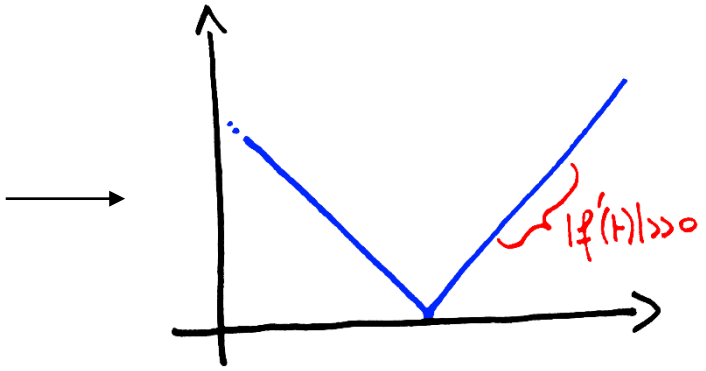
- How to upper bound $\int_0^{\infty} \|\dot{x}(t)\|\, dt = \int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|}\, ds$?
- ``Naive'' hypothesis: $\|\nabla f(y)\| \geq C$ i.e. sharpness
- ``Smart'' hypothesis: $\frac{1}{\|\nabla f(y(s))\|} \leq \varphi'(s)$ with $\varphi \geq 0, \quad \varphi \uparrow$

so the length is $\leq \varphi(s_0) - \varphi(s_\infty) \leq \varphi(s_0)$

How to guarantee convergence?

- How to upper bound $\int_0^{\infty} \|\dot{x}(t)\|\, dt = \int_{s_\infty}^{s_0} \frac{1}{\|\nabla f(y(s))\|}\, ds$?
- ``Naive'' hypothesis: $\|\nabla f(y)\| \geq C$ i.e. sharpness
- ``Smart'' hypothesis: $\frac{1}{\|\nabla f(y(s))\|} \leq \varphi'(s)$ with $\varphi \geq 0, \quad \varphi \uparrow$
  so the length is $\leq \varphi(s_0) - \varphi(s_\infty) \leq \varphi(s_0)$
- In other words $\varphi'\left(f\left(x(t)\right)\right)\left\|\nabla f\left(x(t)\right)\right\| \geq 1$ i.e. $\varphi \circ f$ is sharp:





$$\|\nabla(\varphi \circ f)(x)\| \geq 1$$

The Łojasiewicz property

Definition

We say that f is **Łojasiewicz** at a critical point x^* if

$$\varphi'\big(f(x) - f(x^*)\big)\|\nabla f(x)\| \geq 1,$$

- with $\varphi\colon [0,\infty[\to [0,\infty[$ s.t. $\varphi(0) = 0$, $\varphi \uparrow$, φ concave
- for all $x \in \{\, x' \in \mathbb{B}(x^*, \delta) \mid f(x^*) < f(x') < f(x^*) + r \,\}$

Definition

- f is Łojasiewicz if it is Łojasiewicz at every critical point
- f is p-Łojasiewicz if it is Łojasiewicz at every critical point with $\varphi(s) \simeq s^{1/p}$:

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Theorem (convergence)

Let f be **Łojasiewicz** and $\lambda_k \in]0,2/L[$.

If x^k is **bounded**, then it converges to some critical point x^∞.

Theorem (capture)

Let f be **Łojasiewicz** and $\lambda_k \in]0,2/L[$. For every $x^* \in argmin\, f$, if $x^0 \sim x^*$ then x^k converges to $x^\infty \in argmin\, f$.

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.

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Sketch of proof : show that $\varphi'(s) \geq \|\dot{x}(t)\|$

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Sketch of proof : show that $\varphi'(s) \geq \|\dot{x}(t)\|$

$$\varphi\left(f\left(x^k\right) - f(x^*)\right) - \varphi\left(f\left(x^{k+1}\right) - f(x^*)\right)$$

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Sketch of proof : show that $\varphi'(s) \geq \|\dot{x}(t)\|$

$$\varphi\left(f\left(x^k\right) - f(x^*)\right) - \varphi\left(f\left(x^{k+1}\right) - f(x^*)\right)$$

$$\geq \varphi'(f\left(x^k\right) - f(x^*))(f\left(x^k\right) - f\left(x^{k+1}\right)) \quad \text{because } \varphi \text{ concave}$$

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Sketch of proof : show that $\varphi'(s) \geq \|\dot{x}(t)\|$

$$\varphi\left(f\left(x^k\right) - f(x^*)\right) - \varphi\left(f\left(x^{k+1}\right) - f(x^*)\right)$$

$\geq \varphi'(f\left(x^k\right) - f(x^*))(f\left(x^k\right) - f\left(x^{k+1}\right))$ because φ concave

$\geq \varphi'\left(f\left(x^k\right) - f(x^*)\right) c_{\lambda,L} \left\|x^{k+1} - x^k\right\|^2$ with Descent Lemma

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Sketch of proof : show that $\varphi'(s) \geq \|\dot{x}(t)\|$

$$\varphi\left(f\left(x^k\right) - f(x^*)\right) - \varphi\left(f\left(x^{k+1}\right) - f(x^*)\right)$$

$\geq \varphi'(f\left(x^k\right) - f(x^*))(f\left(x^k\right) - f\left(x^{k+1}\right))$ because φ concave

$\geq \varphi'\left(f\left(x^k\right) - f(x^*)\right) c_{\lambda,L} \left\|x^{k+1} - x^k\right\|^2$ with Descent Lemma

$= \varphi'\left(f\left(x^k\right) - f(x^*)\right) C_{\lambda,L} \left\|x^{k+1} - x^k\right\| \left\|\nabla f\left(x^k\right)\right\|$

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Sketch of proof : show that $\varphi'(s) \geq \|\dot{x}(t)\|$

$$\varphi\left(f\left(x^k\right) - f(x^*)\right) - \varphi\left(f\left(x^{k+1}\right) - f(x^*)\right)$$

$\geq \varphi'(f\left(x^k\right) - f(x^*))(f\left(x^k\right) - f\left(x^{k+1}\right))$ because φ concave

$\geq \varphi'\left(f\left(x^k\right) - f(x^*)\right) c_{\lambda,L} \left\|x^{k+1} - x^k\right\|^2$ with Descent Lemma

$= \varphi'\left(f\left(x^k\right) - f(x^*)\right) C_{\lambda,L} \left\|x^{k+1} - x^k\right\| \left\|\nabla f\left(x^k\right)\right\|$

The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Sketch of proof : show that $\varphi'(s) \geq \|\dot{x}(t)\|$

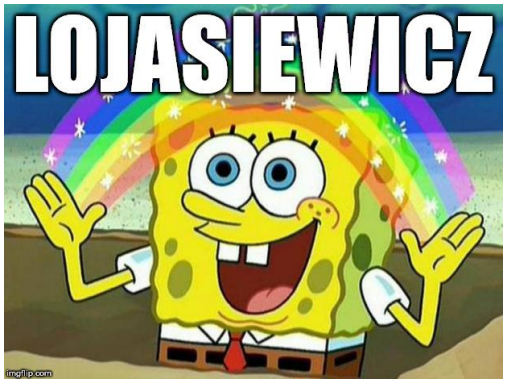
$$\varphi\left(f\left(x^k\right) - f(x^*)\right) - \varphi\left(f\left(x^{k+1}\right) - f(x^*)\right)$$

$\geq \varphi'(f\left(x^k\right) - f(x^*))(f\left(x^k\right) - f\left(x^{k+1}\right))$ because φ concave

$\geq \varphi'\left(f\left(x^k\right) - f(x^*)\right) c_{\lambda,L} \left\|x^{k+1} - x^k\right\|^2$ with Descent Lemma

$= \varphi'\left(f\left(x^k\right) - f(x^*)\right) C_{\lambda,L} \left\|x^{k+1} - x^k\right\| \left\|\nabla f\left(x^k\right)\right\|$

$\geq 1 \cdot C_{\lambda,L} \left\|x^{k+1} - x^k\right\|$ with:



The Łojasiewicz property : convergence

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$x^{k+1} = x^k - \lambda_k \nabla f\left(x^k\right)$$

Theorem (convergence)

Let f be **Łojasiewicz** and $\lambda_k \in]0,2/L[$.

If x^k is **bounded**, then it converges to some critical point x^∞.

Theorem (capture)

Let f be **Łojasiewicz** and $\lambda_k \in]0,2/L[$. For every $x^* \in argmin\, f$, if $x^0 \sim x^*$ then x^k converges to $x^\infty \in argmin\, f$.

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.

The Łojasiewicz property : in practice

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $\mathrm{C}^{1,1}_L$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Examples

- If f is μ-strongly convex, then it is 2-Łojasiewicz with $\mu = \mu$
- If f convex and $\mu\, d(x, \operatorname{argmin} f)^p \leq f(x) - \inf f$

Counter-example

- There is a convex function $f : \mathbb{R}^2 \to \mathbb{R}$ which is not Łojasiewicz

Bolte, Daniilidis, Ley, Mazet, Characterizations of Łojasiewicz inequalities […], 2010.

The Łojasiewicz property : in practice

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Theorem

Any **analytic** function is p-Łojasiewicz at its critical points.

Theorem

- Any **semi-algebraic** function is p-Łojasiewicz at its critical points.
- Any **o-minimal** function is Łojasiewicz.

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.

The Łojasiewicz property : in practice

$f : \mathbb{R}^n \to \mathbb{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Examples of semi-algebraic functions

- Polynomials by parts

The Łojasiewicz property : in practice

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $\mathrm{C}^{1,1}_L$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Examples of semi-algebraic functions

- Polynomials by parts

Theorem (``Tarski-Seidenberg'')

The class of **semi-algebraic** functions is stable under:

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

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

The Łojasiewicz property : in practice

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Examples of semi-algebraic functions

- Polynomials by parts
- $\alpha\|x\|_0 + \|Ax - b\|^2$, $\|x\|_*$, …

Theorem (``Tarski-Seidenberg'')

The class of **semi-algebraic** functions is stable under:

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

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

The Łojasiewicz property : in practice

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Counter-examples of semi-algebraic functions

- Exponential/Logarithmic stuff

The Łojasiewicz property : in practice

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Counter-examples of semi-algebraic functions

- Exponential/Logarithmic stuff

Theorem

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

- Includes the semi-algebraic structure
- Contains the exponential function
- Has the same stability property than the semi-algebraics
- Is also stable by integration (and resolution of 1st order ODEs)

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

The Łojasiewicz property : in practice

$f : \mathbb{R}^n \to \mathbb{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Take-home message:

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

$$\mathbb{R} \to \mathbb{R}$$
$$x \mapsto \sin(x)$$

The Łojasiewicz property : in practice

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Take-home message:

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

$$\mathbb{R} \to \mathbb{R}$$
$$x \mapsto \sin(x)$$

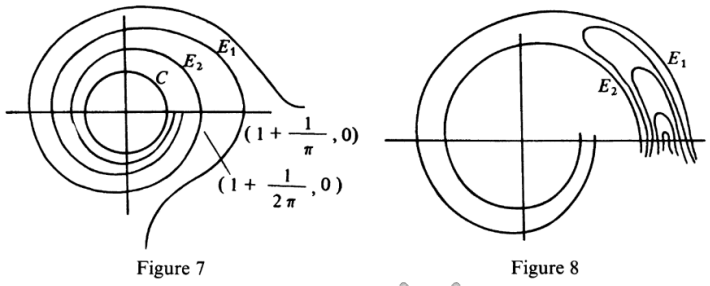
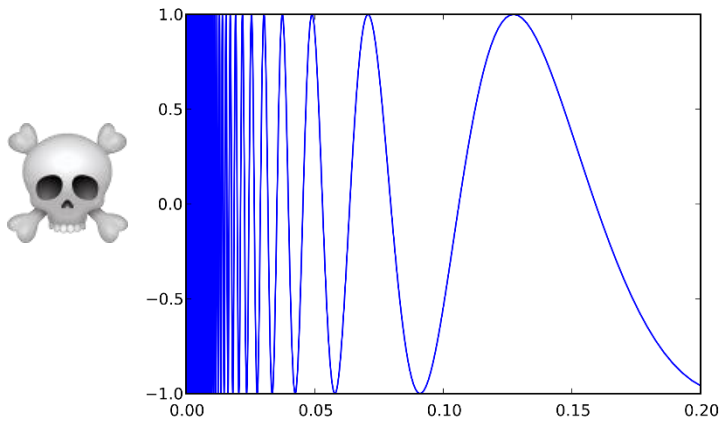


Figure 7

Figure 8

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

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



The Łojasiewicz property : in practice

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Take-home message:

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

$$\mathbb{R} \to \mathbb{R}$$
$$x \mapsto \sin(x)$$

So gradient descent ``always converges'' to a critical point

The Łojasiewicz property : rates for free

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Theorem (p=2)

Let f be globally **2**-Łojasiewicz, $\lambda \in]]0, 2/L[[$, and $x^k \to x^*$.

Then we have **linear** convergence :

$$f(x^{k+1}) - f(x^*) \leq \theta \left(f(x^k) - f(x^*) \right)$$

where $\theta \in [0,1[$, and $\boldsymbol{\theta = 1 - \mu/L}$ if $\lambda = 1/L$.

- If f strongly convex we have $\theta = (1 - \kappa)^2$ for $\lambda = 1/L$.
- If f is [any weak s. convex notion] we have a better θ.
- Rates become asymptotic if local Łojasiewicz only.

Polyak, Gradient methods for the minimisation of functionals, 1963.

The Łojasiewicz property : rates for free

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Theorem (p>2)

Let f be globally **p**-Łojasiewicz, $\lambda \in]]0, 2/L[[$, and $x^k \to x^*$.

Then we have **sublinear** convergence :

$$f(x^k) - f(x^*) = O\left(k^{\frac{-p}{p-2}}\right)$$

- $\frac{p}{p-2} \to +\infty$ when $p \downarrow 2$; $\frac{p}{p-2} \to 1$ when $p \uparrow \infty$
- Rates are matched for $f(x) = x^p$

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.

How to guarantee convergence?

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

Take-home message:

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

$$\mathbb{R} \to \mathbb{R}$$
$$x \mapsto \sin(x)$$

So gradient descent ``always converges'' to a critical point

What about other methods?

How to guarantee convergence?

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

- Nonsmooth 1st-order methods : works the same

Attouch, Bolte, Svaiter, Convergence of descent methods for semi-algebraic and tame problems [...], 2013.

How to guarantee convergence?

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

- Nonsmooth 1st-order methods : works the same
  - Projected gradient
  - Forward Backward
  - Douglas Rachford
  - ADMM
  - Adapts to Maximal Monotone theory (saddle point problems)

Attouch, Bolte, Svaiter, Convergence of descent methods for semi-algebraic and tame problems [...], 2013.

How to guarantee convergence?

$f : \mathrm{R}^n \to \mathrm{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

- Nonsmooth 1st-order methods : works the same
- Inertial (2nd order in time) methods :

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \nabla f\big(x(t)\big) = 0$$

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

Bégout, Bolte, Jendoubi, On damped second-order gradient systems, 2015. + refs within!
Li et al., Convergence Analysis of Proximal Gradient with Momentum for Nonconvex Optimization, 2017.

How to guarantee convergence?

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

- Nonsmooth 1st-order methods : works the same
- Inertial (2nd order in time) methods :
  - Heavy-Ball $(\alpha(t) \equiv \alpha)$ ok
  - Nesterov $(\alpha(t) \sim \alpha/t)$ + Monotone OK pour p=2 global

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \nabla f\big(x(t)\big) = 0$$

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

Bégout, Bolte, Jendoubi, On damped second-order gradient systems, 2015. + refs within!
Li et al., Convergence Analysis of Proximal Gradient with Momentum for Nonconvex Optimization, 2017.

How to guarantee convergence?

$f : \mathbb{R}^n \to \mathbb{R}$ is of class $C_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

- Nonsmooth 1st-order methods : works the same
- Inertial (2^{nd} order in time) methods : ok
- Newton-like (2^{nd} order in space) methods: some results for trust-region methods, Landweber iterations. IDK for Newton/BFGS.

Frankel, Garrigos, Peypouquet, Splitting Methods with Variable Metric [...], 2015. + Absil et al., and many others.

How to guarantee convergence?

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1}$

$$\mu(f(x) - f(x^*))^{p-1} \leq \|\nabla f(x)\|^p$$

- Nonsmooth 1st-order methods : works the same
- Inertial (2nd order in time) methods : ok for Heavy-Ball
- Newton-like (2nd order in space) methods: some results for trust-region methods, Landweber iterations. IDK for Newton/BFGS.
- Stochastic methods: (under global 2-Łojasiewicz)
  - SAGA, SVRG: linear rates
  - SGD: rates $O(\ 1/k\ )$, and linear if vanishing variance
  - SVRG + monotone Nesterov: linear rates

Reddi, Hefny, Sra, Poczos, Smola, Stochastic variance reduction for nonconvex optimization, 2016.
Karimi, Nutini, Schmidt, Linear Convergence of Gradient and Proximal-Gradient […], 2016.
Lei et al., Stochastic Gradient Descent for Nonconvex Learning without Bounded Gradient Assumptions, 2019.

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?

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?

What is the limit? The linear case

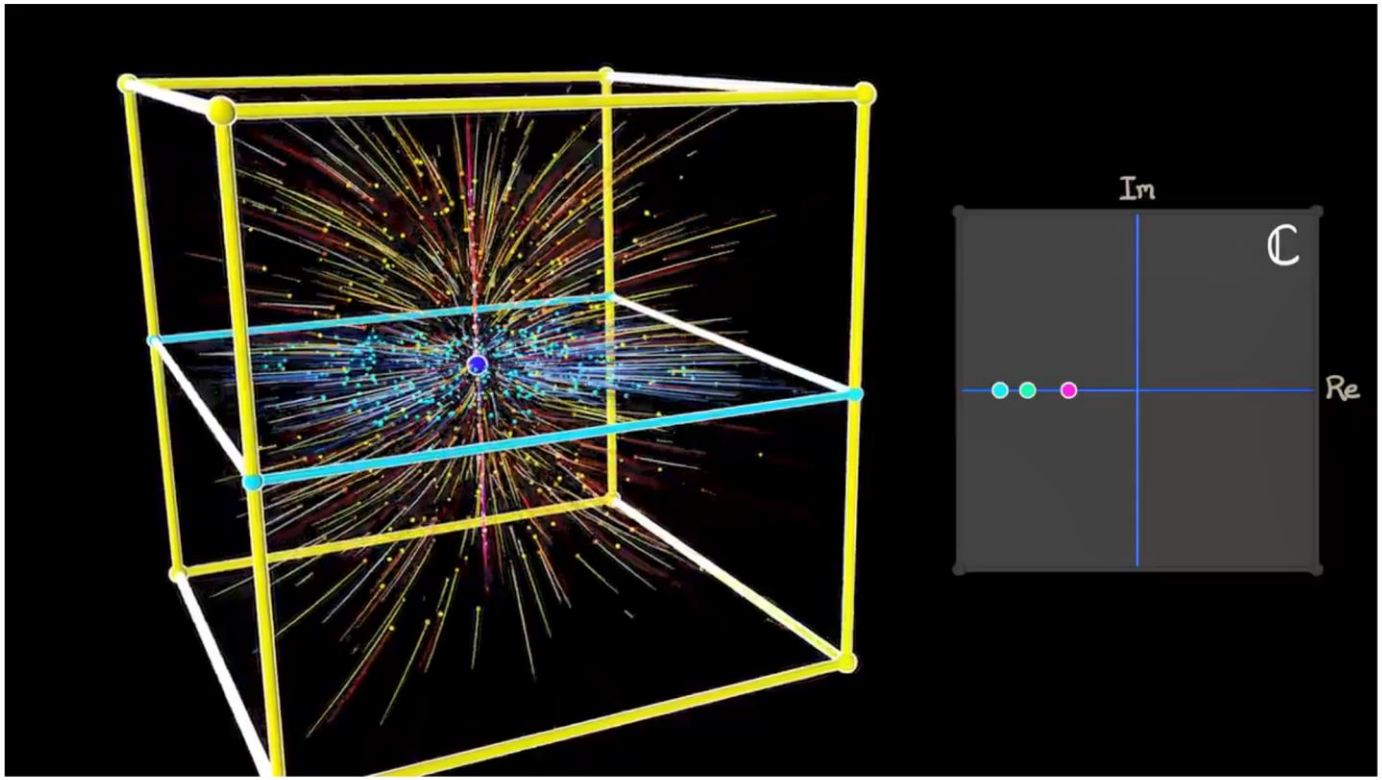
A symmetric operator

$$\dot{x}(t) + Ax(t) = 0$$

What is the limit? The linear case

A symmetric operator

$$\dot{x}(t) + Ax(t) = 0$$



Stolen from Robert Ghrist' Twitter account @robertghrist

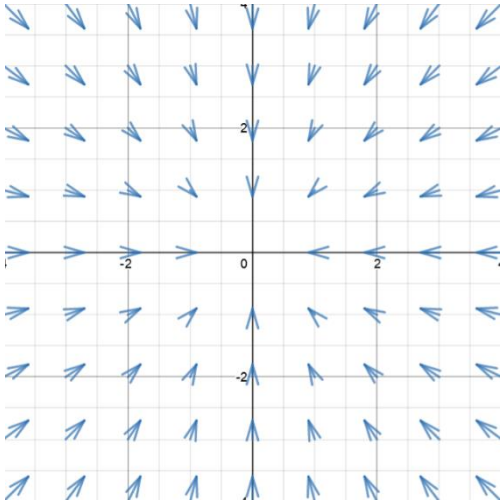
https://twitter.com/robertghrist/status/1185864562299539456

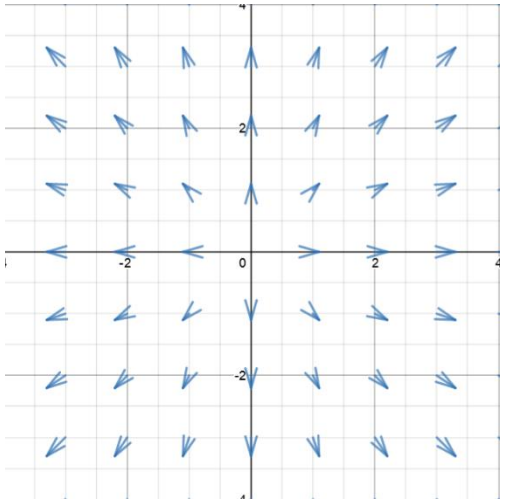
What is the limit? The linear case

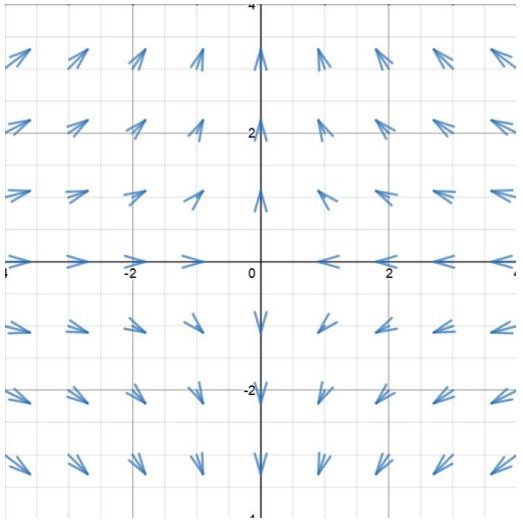
A symmetric operator

$\dot{x}(t) + Ax(t) = 0$

$\dot{x}$







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

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

What is the limit? The linear case

A symmetric operator

$$\dot{x}(t) + Ax(t) = 0$$

Definition

Let $\bar{x}$ be an equilibrium of the system. We define

$$W(\bar{x}) = \{x \mid x(t) \to \bar{x} \text{ with } x(0) = x\}$$

Theorem

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

Corollary

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

What is the limit? The potential case

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class C^2

$$\dot{x}(t) + \nabla f(x(t)) = 0$$

What is the limit? The potential case

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class C^2

$$\dot{x}(t) + \nabla f(x(t)) = 0$$

Theorem (Stable Manifold Lemma)

$W(\bar{x})$ is a submanifold of dimension smaller than the one of

$$\oplus_{\lambda>0}\ E_\lambda\left(\nabla^2 f(\bar{x})\right)$$

Corollary

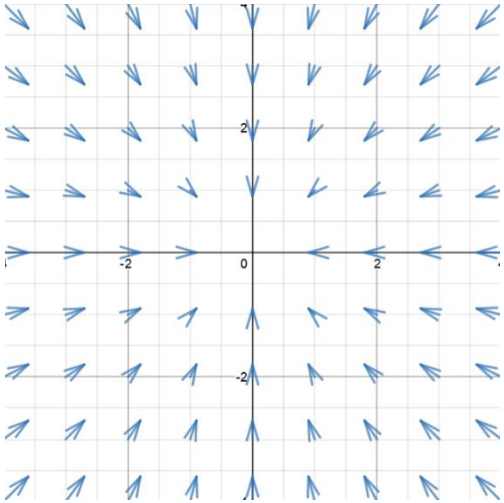
If $\lambda_{min}\left(\nabla^2 f(\bar{x})\right) < 0$, then $W(\bar{x})$ has Lebesgue measure 0.

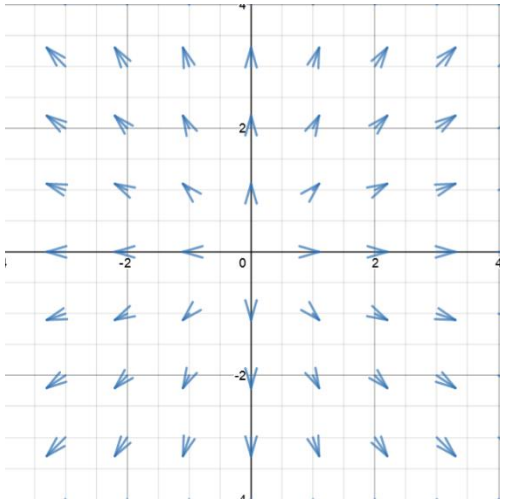
Perron, Die stabilitätsfrage bei differentialgleichungen, 1930. Smale, Differentiable dynamical systems, 1967.

What is the limit? The linear case

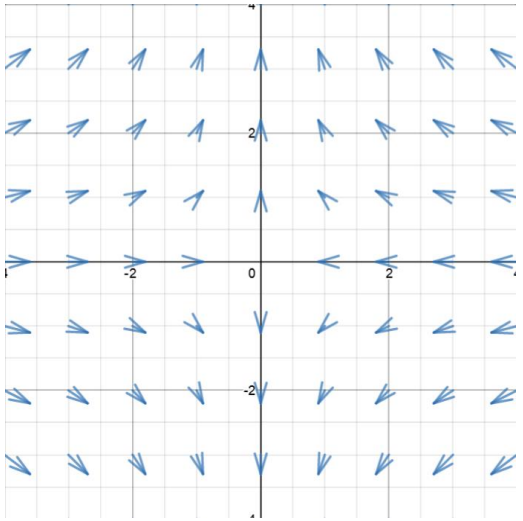
$f : R^n \to R$ is of class C^2

$\dot{x}$





$\dot{x}(t) + \nabla f(x(t)) = 0$



The 3 kinds of critical points:

- The local minima (e.g. $\lambda_{min}\left(\nabla^2 f(\bar{x})\right) > 0$) , **attractive**
- The strict saddles $\lambda_{min}\left(\nabla^2 f(\bar{x})\right) < 0$, **repulsive**
- The degenerated ones (they have $\lambda_{min}\left(\nabla^2 f(\bar{x})\right) = 0$), **???**

What is the limit? The potential case

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class C^2

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla f(x(t)) = 0$$

Goudou, Munier, The gradient and heavy ball with friction dynamical systems: the quasiconvex case, 2007.

What is the limit? The potential case

$f : \mathrm{R}^n \to \mathrm{R}$ is of class C^2

$$\ddot{x}(t) + \alpha \dot{x}(t) + \nabla f(x(t)) = 0$$

Theorem (Stable Manifold Lemma)

$W(\bar{x})$ is a submanifold of dimension smaller than the one of

$$\oplus_{\lambda>0} \; E_\lambda\left(\nabla^2 f(\bar{x})\right)$$

Corollary

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

Goudou, Munier, The gradient and heavy ball with friction dynamical systems: the quasiconvex case, 2007.

What is the limit? The potential case

$\mathrm{f}: \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1} \cap C^2$

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

Theorem

Let $\lambda \in]0,1/L[$. Then $W(\bar{x})$ is a submanifold of dimension smaller than the one of $\bigoplus_{\lambda>0} E_\lambda\left(\nabla^2 f(\bar{x})\right)$

Corollary

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

Corollary

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

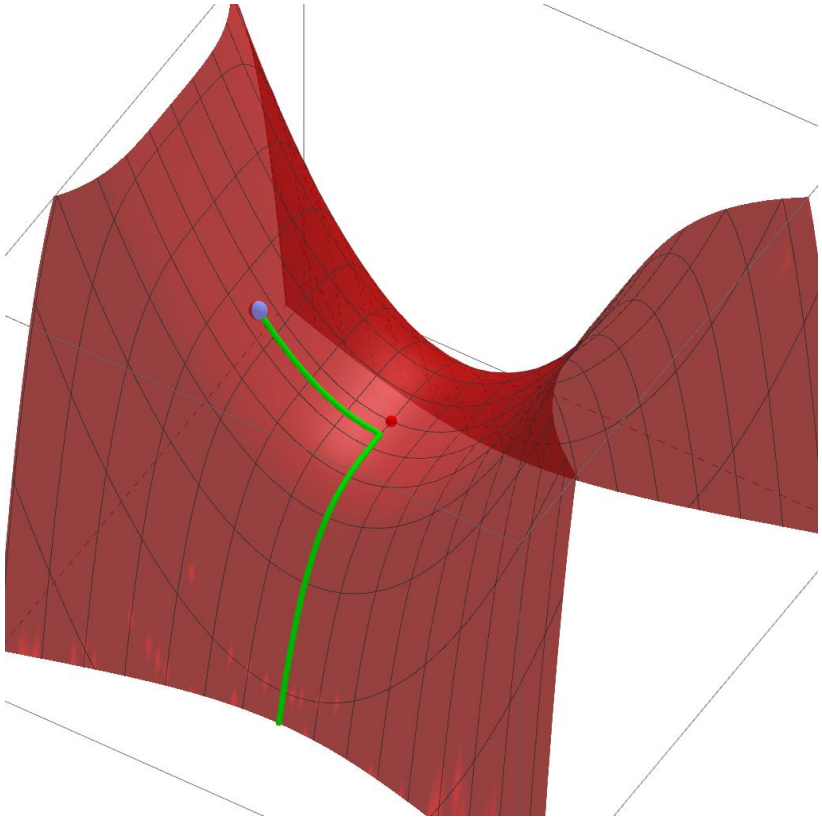
Lee, Simchowitz, Jordan, Recht, Gradient Descent Converges to Minimizers, 2016.

What is the limit? The potential case

$\mathrm{f}: \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1} \cap C^2$

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

It is time now for examples.



What is the limit? The potential case

$\mathrm{f}: \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1} \cap C^2$

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

Corollary

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

Some problems have no degenrated critical points, like the matrix factorization problem a.k.a. two-layer-linear-neural-network

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

Li et al., Symmetry, saddle points, and global geometry of nonconvex matrix factorization, 2016.

What is the limit? The potential case

$\mathrm{f} : \mathrm{R}^{\mathrm{n}} \to \mathrm{R}$ is of class $\mathrm{C}_L^{1,1} \cap C^2$

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

Corollary

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

Corollary

The above result remains true for the noisy gradient method.

The noise here must be isotropic!

Not the case for SGD (proportional to eigenvalues), but proof can be adapted for RKHS learning with a loss s.t. $|\ell''| = O(|\ell'|)$.

Daneshmand et al., Escaping saddles with stochastic gradients, 2018.

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?

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?

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?

Depends strongly on :

- What your problem is
- how you initialize

Yes, this bold statement will be my conclusion

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?

Depends strongly on :

- What your problem is
- how you initialize

Yes, this bold statement will be my conclusion

Any questions ?