Convergence of Stochastic Gradient Descent for analytic target functions

Sebastian Kassing

University of Bielefeld

Math Machine Learning seminar MPI MIS + UCLA

07/07/2022

joint work with Steffen Dereich

Agenda

I Introduction

II Convergence of $(F(X_n))_{n\in\mathbb{N}_0}$ and $(\nabla F(X_n))_{n\in\mathbb{N}_0}$

III Technical Comment and Motivation

IV Convergence of $(X_n)_{n\in\mathbb{N}_0}$

V Conclusion

I Stochastic gradient descent (SGD)

Aim: Minimise a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$.

Examples:

Empirical error:

$$F(x) = \frac{1}{m} \sum_{i=1}^{m} (\Gamma^x(Y_i) - Z_i)^2$$

I Stochastic gradient descent (SGD)

Aim: Minimise a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$.

Examples:

Empirical error:

$$F(x) = \frac{1}{m} \sum_{i=1}^{m} (\Gamma^x(Y_i) - Z_i)^2$$

Grand truth error:

$$F(x) = \mathbb{E}[(\Gamma^x(Y) - Z)^2]$$

I Stochastic gradient descent (SGD)

Aim: Minimise a differentiable function $F(x) = \frac{1}{m}\sum_{i=1}^{m}(\Gamma^x(Y_i) - Z_i)^2$.

Stochastic gradient descent for ERM: Start at $X_0 \in \mathbb{R}^d$ and consider the dynamical system $(X_n)_{n\in\mathbb{N}}$ given by

$$X_n = X_{n-1} - \gamma_n\Big(\frac{1}{N_n}\sum_{i=1}^{N_n}\nabla F_{I_n^{(i)}}(X_{n-1})\Big),$$

where

- $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals, the step-sizes,
- $(N_n)_{n\in\mathbb{N}}$ is a sequence of natural numbers, the batch-size,
- $F_i(x) = (\Gamma^x(Y_i) - Z_i)^2$
- $(I_n^{(i)})_{n,i\in\mathbb{N}}$ are iid. from $\{1,\dots,m\}$

I Stochastic gradient descent (SGD)

Aim: Minimise a differentiable function $F(x) = \frac{1}{m}\sum_{i=1}^{m}(\Gamma^x(Y_i) - Z_i)^2$.

Stochastic gradient descent for ERM: Start at $X_0 \in \mathbb{R}^d$ and consider the dynamical system $(X_n)_{n\in\mathbb{N}}$ given by

$$\begin{aligned} X_n &= X_{n-1} - \gamma_n\Big(\frac{1}{N_n}\sum_{i=1}^{N_n}\nabla F_{I_n^{(i)}}(X_{n-1})\Big) \\ &= X_{n-1} - \gamma_n\Big(\nabla F(X_{n-1}) - \nabla F(X_{n-1}) + \Big(\frac{1}{N_n}\sum_{i=1}^{N_n}\nabla F_{I_n^{(i)}}(X_{n-1})\Big)\Big) \\ &=: X_{n-1} - \gamma_n\Big(\nabla F(X_{n-1}) + D_n\Big) \end{aligned}$$

I Stochastic gradient descent (SGD)

Aim: Minimise a differentiable function $F(x) = \frac{1}{m}\sum_{i=1}^{m}(\Gamma^x(Y_i) - Z_i)^2$.

Stochastic gradient descent for ERM: Start at $X_0 \in \mathbb{R}^d$ and consider the dynamical system $(X_n)_{n\in\mathbb{N}}$ given by

$$\begin{aligned} X_n &= X_{n-1} - \gamma_n\Big(\frac{1}{N_n}\sum_{i=1}^{N_n}\nabla F_{I_n^{(i)}}(X_{n-1})\Big) \\ &= X_{n-1} - \gamma_n\Big(\nabla F(X_{n-1}) - \nabla F(X_{n-1}) + \Big(\frac{1}{N_n}\sum_{i=1}^{N_n}\nabla F_{I_n^{(i)}}(X_{n-1})\Big)\Big) \\ &=: X_{n-1} - \gamma_n\Big(\nabla F(X_{n-1}) + D_n\Big) \end{aligned}$$

Note that $\mathbb{E}[D_n] = 0$.

I Stochastic gradient descent (SGD)

Aim: Minimise a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$.

Stochastic gradient descent: Consider a dynamical system $(X_n)_{n\in\mathbb{N}_0}$ given by

$$X_n = X_{n-1} - \gamma_n(\nabla F(X_{n-1}) + D_n),$$

where

- $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals, the *step-sizes*,
- $(D_n)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)_{n\in\mathbb{N}}$-adapted sequence of martingale differences, the *perturbation*,
- X_0 is an $\mathcal{F}_0$-measurable random variable, the *initial value*.

I Stochastic gradient descent (SGD)

Aim: Minimise a differentiable function $F : \mathbb{R}^d \to \mathbb{R}$.

Stochastic gradient descent: Consider a dynamical system $(X_n)_{n\in\mathbb{N}_0}$ given by

$$X_n = X_{n-1} - \gamma_n(\nabla F(X_{n-1}) + D_n),$$

where

- $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals, the *step-sizes*,
- $(D_n)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)_{n\in\mathbb{N}}$-adapted sequence of martingale differences, the *perturbation*,
- X_0 is an $\mathcal{F}_0$-measurable random variable, the *initial value*.

Remark:

- If $D_n \equiv 0$, then (X_n) is just classical gradient descent. (Absil et al. '05)
- A step in the SGD-algorithm may be understood as a perturbed Euler step of length γ_n of the ODE

$$\dot{x}_t = -\nabla F(x_t)$$

$\rightsquigarrow$ $t_n = \sum_{k=1}^{n} \gamma_k$ "age" of the SGD

I SGD - historical comments

Introduced by Robbins and Monro in '51

- strong impact in statistics, system control, optimization and machine learning

Classical setting: F has an isolated global minimum x^* and is strongly convex, i.e. there exists a $\lambda > 0$ such that

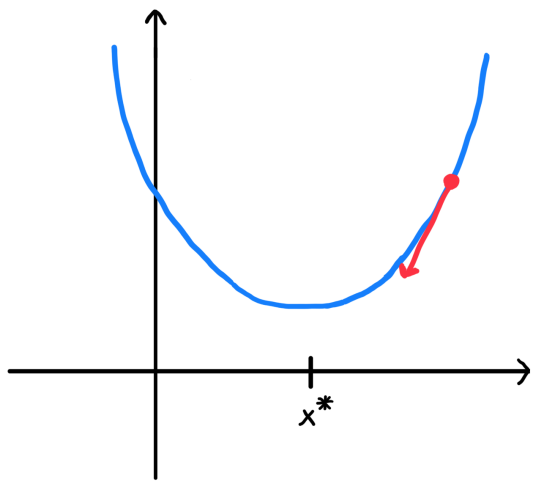
$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\lambda}{2}|y - x|^2$$

for all $x, y \in \mathbb{R}^d$.

Classical step-size conditions:

$$\sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

- example: $(\gamma_n)_{n\in\mathbb{N}} = (C_\gamma n^{-\gamma})_{n\in\mathbb{N}}$ with $C_\gamma > 0$ and $\gamma \in (1/2, 1]$



II Convergence of SGD

SGD:

$$X_n = X_{n-1} - \gamma_n(\nabla F(X_{n-1}) + D_n)$$

First result: Almost sure convergence of

$$\text{(I) } (F(X_n))_{n\in\mathbb{N}_0}, \quad \text{(II) } (\nabla F(X_n))_{n\in\mathbb{N}_0} \quad \text{and} \quad \text{(III) } (X_n)_{n\in\mathbb{N}_0}.$$

II Convergence of SGD

SGD:

$$X_n = X_{n-1} - \gamma_n(\nabla F(X_{n-1}) + D_n)$$

First result: Almost sure convergence of

$$\text{(I) } (F(X_n))_{n\in\mathbb{N}_0}, \quad \text{(II) } (\nabla F(X_n))_{n\in\mathbb{N}_0} \quad \text{and} \quad \text{(III) } (X_n)_{n\in\mathbb{N}_0}.$$

Particular focus on (III)**!**

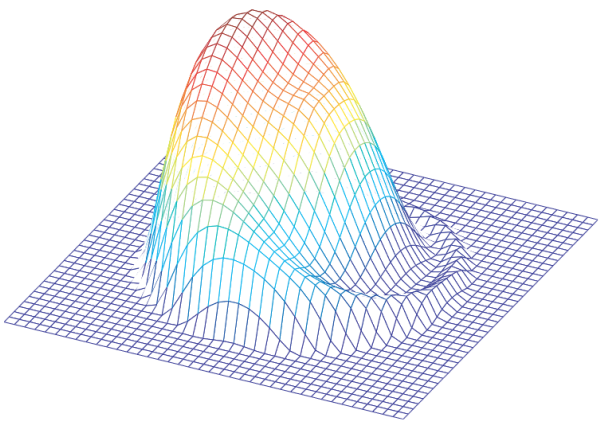


Figure: Absil et al., '05

$$F(r\cos(\varphi), r\sin(\varphi)) =$$

$$\mathbf{1}_{\{r<1\}} e^{-\frac{1}{1-r^2}} \left(1 - \frac{4r^4}{4r^4 + (1-r^2)^4} \sin\left(\varphi - \frac{1}{1-r^2}\right)\right)$$

II Convergence of SGD

We discuss convergence of

$$\text{(I) } (F(X_n))_{n\in\mathbb{N}_0}, \quad \text{(II) } (\nabla F(X_n))_{n\in\mathbb{N}_0} \quad \text{and} \quad \text{(III) } (X_n)_{n\in\mathbb{N}_0}.$$

Convergence of (I) and (II): treated in Walk 92, Gaivoronski '94, Lu, Tseng '94, Grippo '94, Mangasarian, Solodov '94, Bertsekas, Tsitsiklis '00

Convergence of (III) is more subtle.

We will restrict attention on two events:

- the SGD stays $\mathbb{L}$ocal

$$\mathbb{L} = \Big\{ \limsup_{n\to\infty} |X_n| < \infty \Big\}$$

- the $\mathbb{M}$artingale differences (D_n) are of order $(\sigma_n)_{n\in\mathbb{N}}$ for the moment $p \geq 1$

$$\mathbb{M}_\sigma^p = \Big\{ \limsup_{n\to\infty} \sigma_n^{-1} \mathbb{E}[|D_n|^p | \mathcal{F}_{n-1}]^{1/p} < \infty \Big\}.$$

Typically: $\sigma_n = 1$ or $\sigma_n \approx (\text{size of mini-badges})^{-1/2}$

II Convergence of SGD $(F(X_n))$ and $(f(X_n))$

Recall that

$$\mathbb{L} = \Big\{ \limsup_{n\to\infty} |X_n| < \infty \Big\}$$

and

$$\mathbb{M}_\sigma^p = \Big\{ \limsup_{n\to\infty} \sigma_n^{-1} \mathbb{E}[|D_n|^p | \mathcal{F}_{n-1}]^{1/p} < \infty \Big\}.$$

Theorem: (Polyak, Tsypkin '73, Walk '92, . . . , Dereich, K '21+)
Let $p \in (1, 2]$ and suppose that ∇F is locally Lipschitz continuous. If

$$\gamma_n \to 0 \quad , \quad \sum_{n=1}^{\infty} \gamma_n = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} (\gamma_n \sigma_n)^p < \infty,$$

then, on $\mathbb{L} \cap \mathbb{M}_\sigma^p$, almost surely, the limit $(F(X_n))_{n\in\mathbb{N}_0}$ exists and $\lim_{n\to\infty} \nabla F(X_n) = 0$.

Consequence: If the set of critical points does not contain a continuum, then one also gets a.s. convergence of $(X_n)_{n\in\mathbb{N}_0}$ on $\mathbb{L} \cap \mathbb{M}_\sigma^p$.

III Convergence in general setting

Q: Can we hope to generally be able to deduce convergence in the case where the critical points of F contain a continuum?

III Convergence in general setting

Q: Can we hope to generally be able to deduce convergence in the case where the critical points of F contain a continuum?

No! There exist ODE-counterexamples $\rightsquigarrow$ additional assumptions needed!

Counterexample: $\exists$ a C^∞-function $F : \mathbb{R}^2 \to \mathbb{R}$ such that the ODE

$$\dot{x}_t = -\nabla F(x_t)$$

started in an appropriate x_0 spirals infinitely often around a disc:

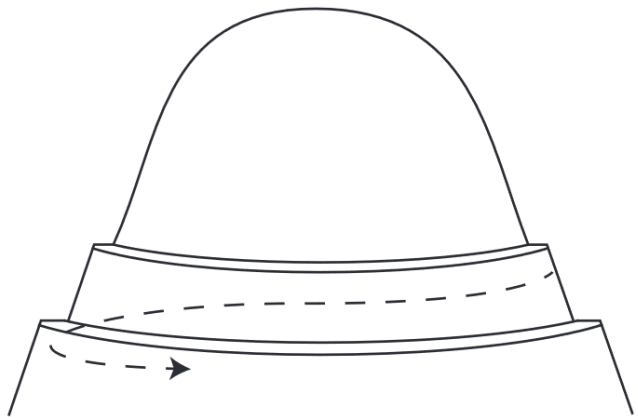


Figure: Colding, Minicozzi II, '15

III ODE idea

From a technical point of view..

The solution to

$$\dot{x}_t = -\nabla F(x_t)$$

satisfies

$$\frac{d}{dt} F(x_t) = -|\nabla F(x_t)|^2$$

so that if $(F(x_t))_{t \geq 0}$ is bounded from below,

$$\int_0^\infty |\nabla F(x_t)|^2 \, dt < \infty.$$

III ODE idea

From a technical point of view..

The solution to

$$\dot{x}_t = -\nabla F(x_t)$$

satisfies

$$\frac{d}{dt} F(x_t) = -|\nabla F(x_t)|^2$$

so that if $(F(x_t))_{t \geq 0}$ is bounded from below,

$$\int_0^\infty |\nabla F(x_t)|^2 \, dt < \infty.$$

But a sufficient criterion for convergence is

$$\int_0^\infty |\nabla F(x_t)| \, dt < \infty. \qquad (*)$$

III ODE idea

From a technical point of view..

The solution to

$$\dot{x}_t = -\nabla F(x_t)$$

satisfies

$$\frac{d}{dt} F(x_t) = -|\nabla F(x_t)|^2$$

so that if $(F(x_t))_{t\geq 0}$ is bounded from below,

$$\int_0^\infty |\nabla F(x_t)|^2 \, dt < \infty.$$

But a sufficient criterion for convergence is

$$\int_0^\infty |\nabla F(x_t)| \, dt < \infty. \qquad (*)$$

Note:

- Need additional assumptions $\rightsquigarrow$ Łojasiewicz-inequality
- Later these will entail that (*) holds.

III The Łojasiewicz-inequality

Def.: A C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy a Łojasiewicz-inequality in $x \in \mathbb{R}^d$ with parameters $\theta \in [\frac{1}{2}, 1)$ and $Ł > 0$ on a neighbourhood U_x of x if for all $y \in U_x$

$$|\nabla F(y)| \geq Ł|F(y) - F(x)|^\theta.$$

Note:

- Łojasiewicz-inequality for $\theta \Rightarrow$ Łojasiewicz-inequality for $\theta' \geq \theta$
- x non-critical point of C^1-function $\Rightarrow$ Łojasiewicz-inequality for $\theta = \frac{1}{2}$

III The Łojasiewicz-inequality

Def.: A C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy a Łojasiewicz-inequality in $x \in \mathbb{R}^d$ with parameters $\theta \in [\frac{1}{2}, 1)$ and $Ł > 0$ on a neighbourhood U_x of x if for all $y \in U_x$

$$|\nabla F(y)| \geq Ł|F(y) - F(x)|^\theta.$$

Examples:

- Parabola: $F(y) = y^2 \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$

III The Łojasiewicz-inequality

Def.: A C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy a Łojasiewicz-inequality in $x \in \mathbb{R}^d$ with parameters $\theta \in [\frac{1}{2}, 1)$ and $Ł > 0$ on a neighbourhood U_x of x if for all $y \in U_x$

$$|\nabla F(y)| \geq Ł|F(y) - F(x)|^\theta.$$

Examples:

- Parabola: $F(y) = y^2 \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$
- $F : \mathbb{R}^d \to [0, \infty)$ is C^2, $M = \nabla F^{-1}(0)$ is a C^1-manifold with

$$\lambda|v|^2 \leq v^\dagger D^2 F(x) v \leq \Lambda|v|^2 \quad \forall x \in M, v \in N_x M$$

with $0 < \lambda \leq \Lambda \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$ (Wojtowytsch 21')

III The Łojasiewicz-inequality

Def.: A C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy a Łojasiewicz-inequality in $x \in \mathbb{R}^d$ with parameters $\theta \in [\frac{1}{2}, 1)$ and $Ł > 0$ on a neighbourhood U_x of x if for all $y \in U_x$

$$|\nabla F(y)| \geq Ł|F(y) - F(x)|^\theta.$$

Examples:

- Parabola: $F(y) = y^2 \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$
- $F : \mathbb{R}^d \to [0, \infty)$ is C^2, $M = \nabla F^{-1}(0)$ is a C^1-manifold with

  $$\lambda|v|^2 \leq v^\dagger D^2 F(x) v \leq \Lambda|v|^2 \quad \forall x \in M, v \in N_x M$$

  with $0 < \lambda \leq \Lambda \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$ (Wojtowytsch 21')
- Monomial: $F(y) = y^p$ ($p \in \{2, \dots\}$); then $|\nabla F(y)| = p|F(y)|^{\frac{p-1}{p}}$
  $\Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{p-1}{p}$

III The Łojasiewicz-inequality

Def.: A C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy a Łojasiewicz-inequality in $x \in \mathbb{R}^d$ with parameters $\theta \in [\frac{1}{2}, 1)$ and $Ł > 0$ on a neighbourhood U_x of x if for all $y \in U_x$

$$|\nabla F(y)| \geq Ł|F(y) - F(x)|^\theta.$$

Examples:

- Parabola: $F(y) = y^2 \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$
- $F : \mathbb{R}^d \to [0, \infty)$ is C^2, $M = \nabla F^{-1}(0)$ is a C^1-manifold with

  $$\lambda|v|^2 \leq v^\dagger D^2 F(x) v \leq \Lambda|v|^2 \quad \forall x \in M, v \in N_x M$$

  with $0 < \lambda \leq \Lambda \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$ (Wojtowytsch 21')
- Monomial: $F(y) = y^p$ ($p \in \{2, \dots\}$); then $|\nabla F(y)| = p|F(y)|^{\frac{p-1}{p}}$
  $\Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{p-1}{p}$
- $F(y) = \exp\{-|y|^{-1}\}$; then $\nabla F(y) = \operatorname{sgn}(y) \exp\{-|y|^{-1}\}|y|^{-2}$
  $\Rightarrow$ no Łojasiewicz-inequality

III The Łojasiewicz-inequality

Def.: A C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ is said to satisfy a Łojasiewicz-inequality in $x \in \mathbb{R}^d$ with parameters $\theta \in [\frac{1}{2}, 1)$ and $Ł > 0$ on a neighbourhood U_x of x if for all $y \in U_x$

$$|\nabla F(y)| \geq Ł|F(y) - F(x)|^\theta.$$

Examples:

- Parabola: $F(y) = y^2 \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$
- $F : \mathbb{R}^d \to [0, \infty)$ is C^2, $M = \nabla F^{-1}(0)$ is a C^1-manifold with

  $$\lambda|v|^2 \leq v^\dagger D^2 F(x) v \leq \Lambda|v|^2 \quad \forall x \in M, v \in N_x M$$

  with $0 < \lambda \leq \Lambda \Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{1}{2}$ (Wojtowytsch 21')
- Monomial: $F(y) = y^p$ ($p \in \{2, \dots\}$); then $|\nabla F(y)| = p|F(y)|^{\frac{p-1}{p}}$
  $\Rightarrow$ Łojasiewicz-inequ. with $\theta = \frac{p-1}{p}$
- $F(y) = \exp\{-|y|^{-1}\}$; then $\nabla F(y) = \operatorname{sgn}(y) \exp\{-|y|^{-1}\}|y|^{-2}$
  $\Rightarrow$ no Łojasiewicz-inequality
- $F : \mathbb{R}^d \to \mathbb{R}$ real analytic function
  $\Rightarrow$ F satisfies in every point a Łojasiewicz-inequality (Łojasiewicz'63)

IV Convergence of SGD for Łojasiewicz-landscapes

Def.: We call a C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ with locally Lipschitz continuous derivative that satisfies for every $x \in \mathbb{R}^d$ a Łojasiewicz-inequality around x a Łojasiewicz-function.

Theorem: (Dereich, K '21+) Let F be a Łojasiewicz-function and $p \geq 2$. Suppose that for $n \in \mathbb{N}$

$$\gamma_n = C_\gamma n^{-\gamma} \quad \text{and} \quad \sigma_n = n^\sigma,$$

where $C_\gamma > 0$, $\gamma \in (\frac{1}{2}, 1]$ and $\sigma \in \mathbb{R}$. If

$$\frac{2}{3}(\sigma + 1) < \gamma \quad \text{and} \quad \frac{1}{2\gamma - \sigma - 1} < p,$$

then, on $\mathbb{L} \cap \mathbb{M}_\sigma^p$, the process (X_n) converges, almost surely, to a critical point of F (possibly a saddle point or a local maximum).

IV Convergence of SGD for Łojasiewicz-landscapes

Def.: We call a C^1-function $F : \mathbb{R}^d \to \mathbb{R}$ with locally Lipschitz continuous derivative that satisfies for every $x \in \mathbb{R}^d$ a Łojasiewicz-inequality around x a Łojasiewicz-function.

Theorem: (Dereich, K '21+) Let F be a Łojasiewicz-function and $p \geq 2$. Suppose that for $n \in \mathbb{N}$

$$\gamma_n = C_\gamma n^{-\gamma} \quad \text{and} \quad \sigma_n = n^\sigma,$$

where $C_\gamma > 0$, $\gamma \in (\frac{1}{2}, 1]$ and $\sigma \in \mathbb{R}$. If

$$\frac{2}{3}(\sigma + 1) < \gamma \quad \text{and} \quad \frac{1}{2\gamma - \sigma - 1} < p,$$

then, on $\mathbb{L} \cap \mathbb{M}_\sigma^p$, the process (X_n) converges, almost surely, to a critical point of F (possibly a saddle point or a local maximum).

ML-application: without mini batches or with constant size batches one has $\sigma = 0$, so γ needs to lie in $(\frac{2}{3}, 1]$ and if $\gamma > \frac{3}{4}$ one can choose $p = 2$.

IV Łojasiewicz-landscapes in deep learning

DL-application: (Dereich, K '21+) Suppose that

- the activation functions and the loss function $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}$ are real analytic,
- that Y and Z are compactly supported $\mathbb{R}^{d_{\text{in}}}$- and $\mathbb{R}$-valued r.v.'s.

IV Łojasiewicz-landscapes in deep learning

DL-application: (Dereich, K '21+) Suppose that

- the activation functions and the loss function $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}$ are real analytic,
- that Y and Z are compactly supported $\mathbb{R}^{d_{\text{in}}}$- and $\mathbb{R}$-valued r.v.'s.

For a fixed architecture $N = (N_0, \dots, N_L)$ with $N_0 = d_{\text{in}}$ and $N_L = 1$ we let

- $\mathcal{P}_N$ be the set of all parameterizations of networks with architecture N and
- for $\Theta \in \mathcal{P}_N$, $\mathcal{R}_N(\Theta, \cdot) : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ be the response function of the Θ-netw.

Then the expected loss

$$F : \mathcal{P}_N \to \mathbb{R},\ \Theta \to \mathbb{E}[\mathcal{L}(\mathcal{R}_N(\Theta, Y), Z)]$$

is analytic.

IV Łojasiewicz-landscapes in deep learning

DL-application: (Dereich, K '21+) Suppose that

- the activation functions and the loss function $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}$ are real analytic,
- that Y and Z are compactly supported $\mathbb{R}^{d_{\text{in}}}$- and $\mathbb{R}$-valued r.v.'s.

For a fixed architecture $N = (N_0, \dots, N_L)$ with $N_0 = d_{\text{in}}$ and $N_L = 1$ we let

- $\mathcal{P}_N$ be the set of all parameterizations of networks with architecture N and
- for $\Theta \in \mathcal{P}_N$, $\mathcal{R}_N(\Theta, \cdot) : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ be the response function of the Θ-netw.

Then the expected loss

$$F : \mathcal{P}_N \to \mathbb{R},\ \Theta \to \mathbb{E}[\mathcal{L}(\mathcal{R}_N(\Theta, Y), Z)]$$

is analytic.

Regression with MSE: take $\mathcal{L}(z, y) = (z - y)^2 \ \to$ analytic

Analytic activation functions: softplus ($t \mapsto \ln(1 + e^t)$), sigmoid ($t \mapsto 1/(1 + e^{-t})$) and the hyperbolic tangent

Example: One may take $(Y, Z) \sim \frac{1}{m} \sum_{k=1}^{m} \delta_{(y_k, z_k)}$ (empirical distribution).

III Why do we care?

Consider convergence to a parabolic minimizer.

Problem: $\gamma_n = C_\gamma n^{-1}$ best choice in terms of asymptotic speed of convergence of stochastic approximation; but

- choice of C_γ is problematic
- age increases only logarithmically $\rightsquigarrow$ if ODE needs time t to get close to a minimum, the SGD needs at least of order e^t steps to get close

III Why do we care?

Consider convergence to a parabolic minimizer.

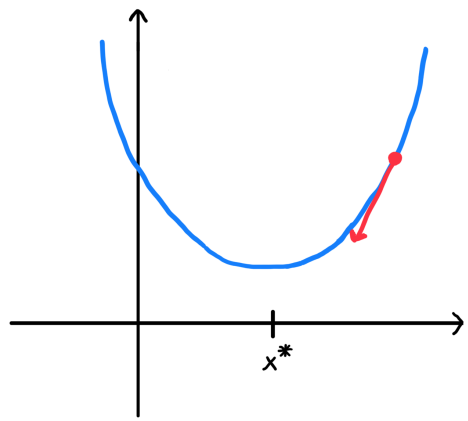
Problem: $\gamma_n = C_\gamma n^{-1}$ best choice in terms of asymptotic speed of convergence of stochastic approximation; but

- choice of C_γ is problematic
- age increases only logarithmically $\rightsquigarrow$ if ODE needs time t to get close to a minimum, the SGD needs at least of order e^t steps to get close

Ruppert-Polyak averaging: (Ruppert '88, Polyak '90, Polyak, Juditsky '92)

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k$$

- fast convergence even for $\gamma_n = C_\gamma n^{-\gamma}$ with $\gamma \in (\frac{1}{2}, 1)$,
- to pass "ODE-distance" t we then need about $t^{1/(1-\gamma)}$ steps.



III Why do we care?

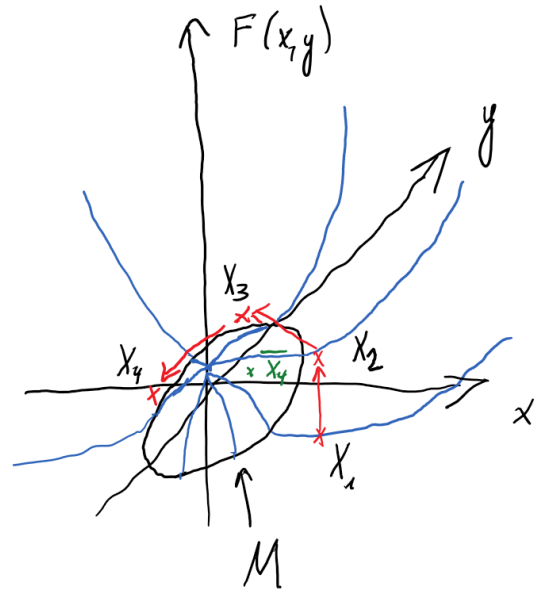
Ruppert-Polyak averaging for stable manifolds: (Dereich, K. '19+)
Assume that the minima of F form a C^3-submanifold M such that for all $x \in M$ and $v \in (T_x M)^\perp \setminus \{0\}$

$$v^\dagger \operatorname{Hess} F(x) v > 0.$$

Then we get a Central Limit Theorem for

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k$$

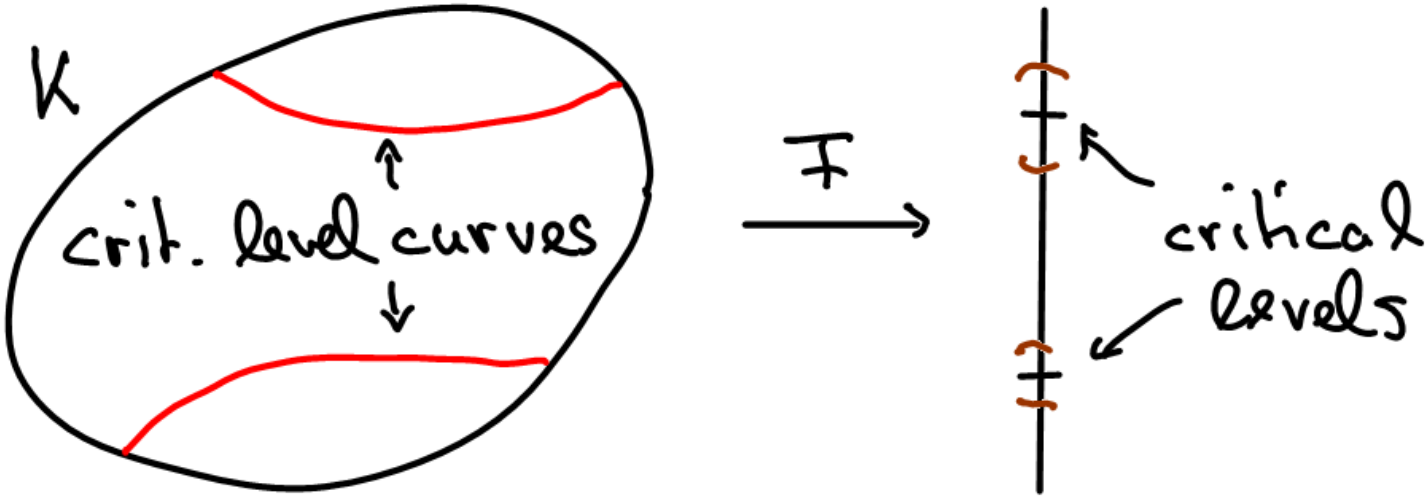
with fast convergence rates even for $\gamma_n = C_\gamma n^{-\gamma}$ with $\gamma \in (\frac{3}{4}, 1)$ **on the set that (X_n) converges!**



$$\sqrt{n}\,(\bar{X}_n - \bar{X}_n^*) \overset{\text{stably}}{\Longrightarrow} \mathcal{N}(0, \Sigma)$$

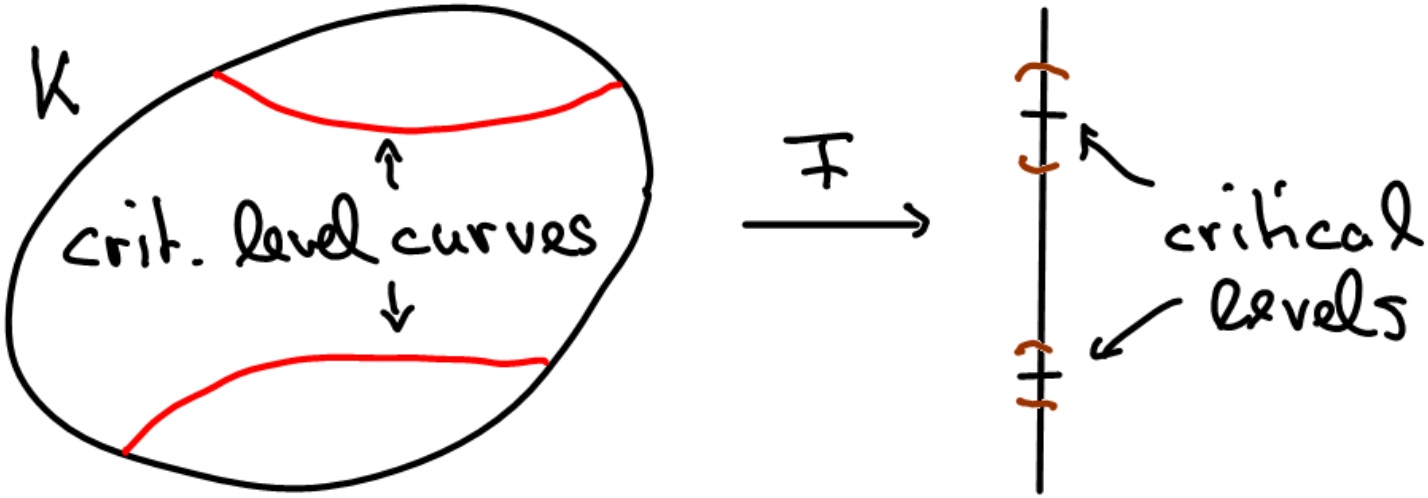
See also: Tripuraneni, Flammarion, Bach, Jordan 2018

V Convergence of SGD: sketch of the proof



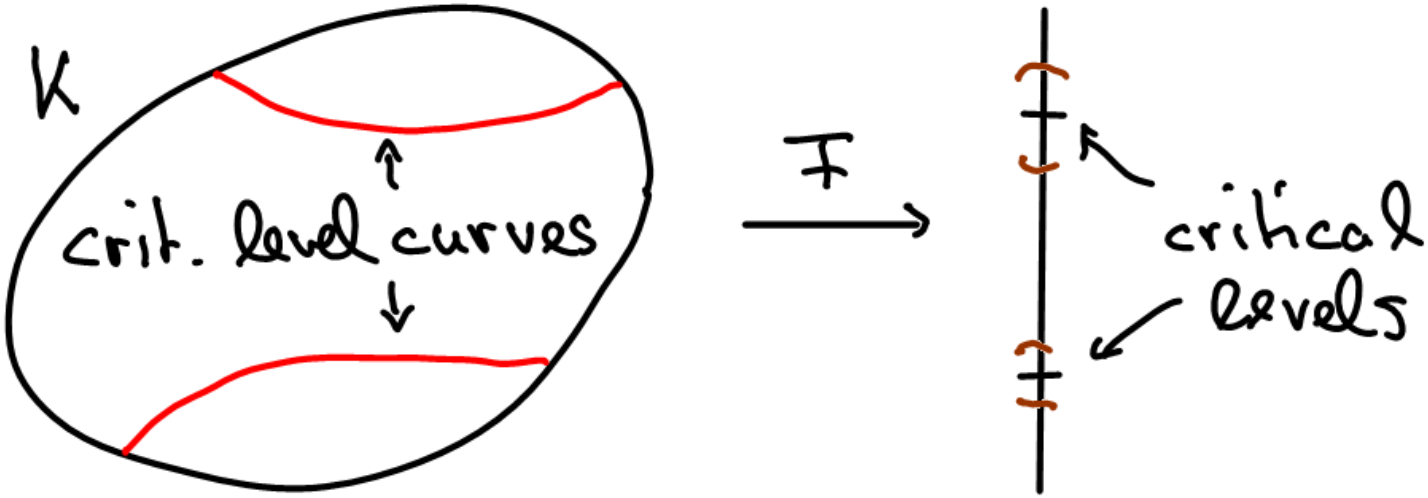
- On each compact set K there is only a finite number of critical levels.

V Convergence of SGD: sketch of the proof



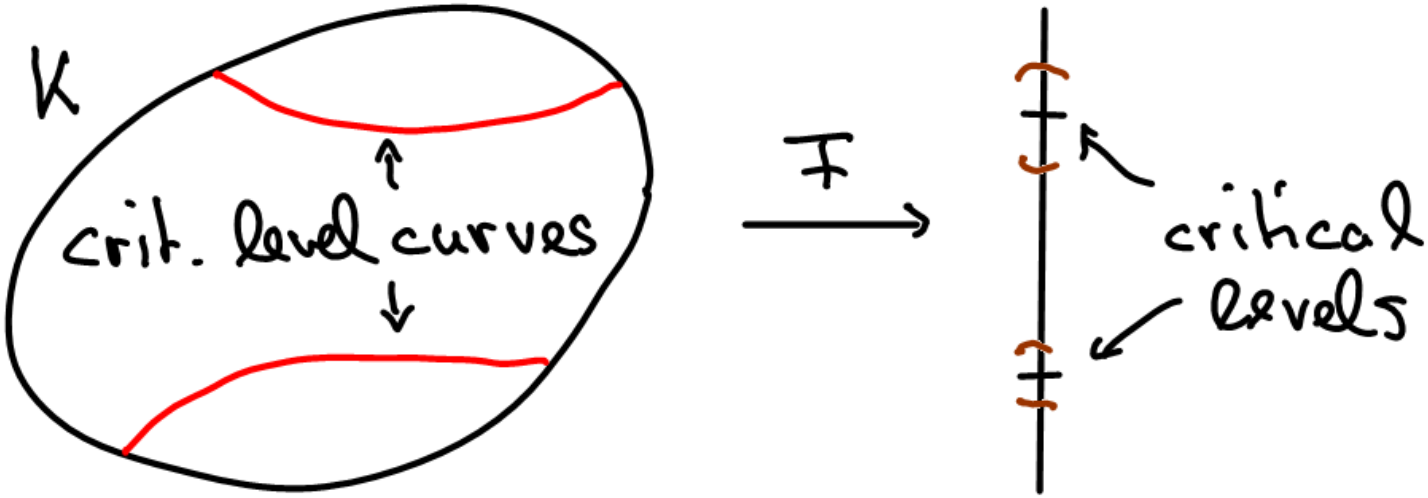
- On each compact set K there is only a finite number of critical levels.
- For each critical level $z \in \mathcal{L}_K := \{F(x) : x \in K \text{ with } \nabla F(x) = 0\}$ a Łojasiewicz-inequality holds on stripe $F^{-1}((F(z) - \varepsilon, F(z) + \varepsilon)) \cap K$.

V Convergence of SGD: sketch of the proof



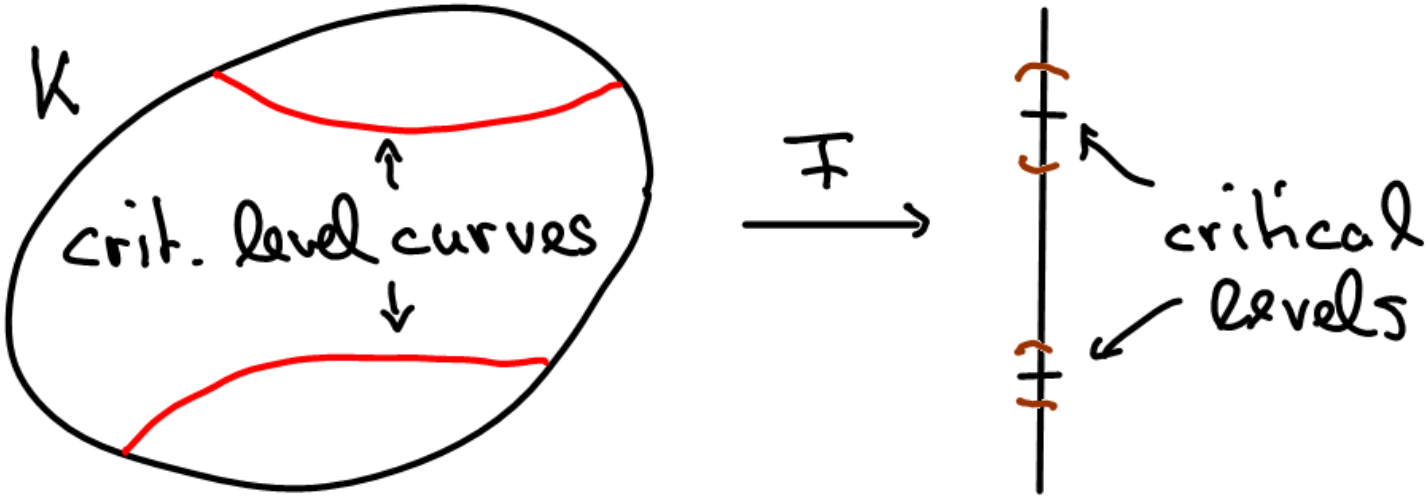
- On each compact set K there is only a finite number of critical levels.
- For each critical level $z \in \mathcal{L}_K := \{F(x) : x \in K \text{ with } \nabla F(x) = 0\}$ a Łojasiewicz-inequality holds on stripe $F^{-1}((F(z) - \varepsilon, F(z) + \varepsilon)) \cap K$.
- If the scheme stays in K, it has to eventually stay in one of the stripes.

V Convergence of SGD: sketch of the proof



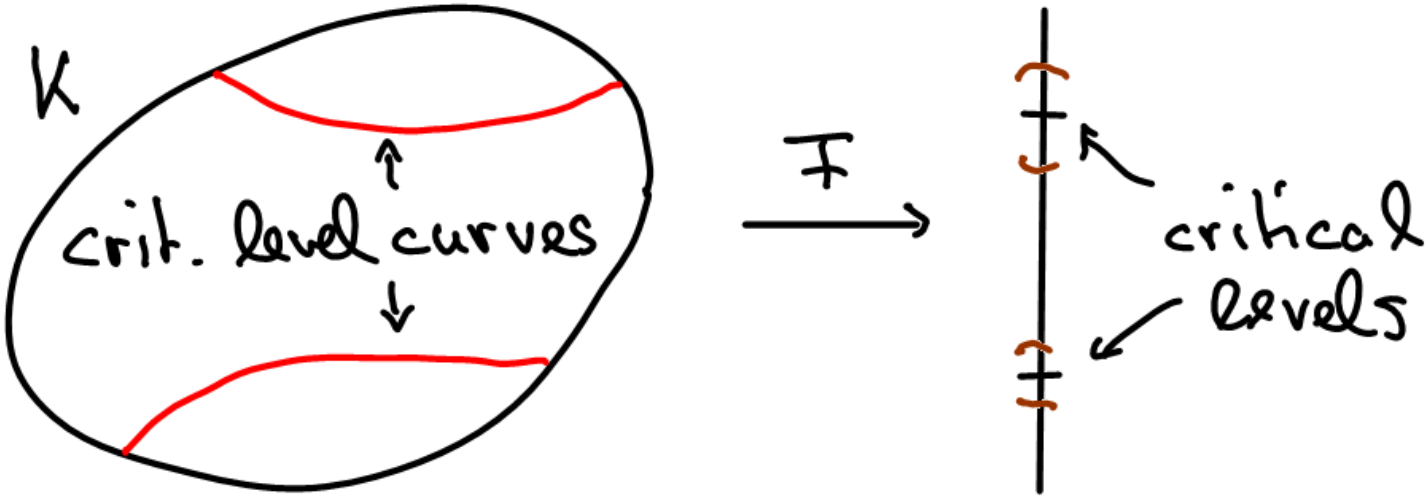
- On each compact set K there is only a finite number of critical levels.
- For each critical level $z \in \mathcal{L}_K := \{F(x) : x \in K \text{ with } \nabla F(x) = 0\}$ a Łojasiewicz-inequality holds on stripe $F^{-1}((F(z) - \varepsilon, F(z) + \varepsilon)) \cap K$.
- If the scheme stays in K, it has to eventually stay in one of the stripes.
- Control $\mathbb{E}[\mathbf{1}\{X \text{ stays in stripe and no lower drop-out}\}(F(X_n) - F(z))]$.

V Convergence of SGD: sketch of the proof



- On each compact set K there is only a finite number of critical levels.
- For each critical level $z \in \mathcal{L}_K := \{F(x) : x \in K \text{ with } \nabla F(x) = 0\}$ a Łojasiewicz-inequality holds on stripe $F^{-1}((F(z) - \varepsilon, F(z) + \varepsilon)) \cap K$.
- If the scheme stays in K, it has to eventually stay in one of the stripes.
- Control $\mathbb{E}[\mathbf{1}\{X \text{ stays in stripe and no lower drop-out}\}(F(X_n) - F(z))]$.
- Similarly to ODE-argument one shows that $\mathbb{E}[\mathbf{1}\{X \text{ stays in stripe and no lower drop-out}\} \sum_{k=1}^{\infty} \gamma_k |\nabla F(X_k)|] < \infty$.

V Convergence of SGD: sketch of the proof



- On each compact set K there is only a finite number of critical levels.
- For each critical level $z \in \mathcal{L}_K := \{F(x) : x \in K \text{ with } \nabla F(x) = 0\}$ a Łojasiewicz-inequality holds on stripe $F^{-1}((F(z) - \varepsilon, F(z) + \varepsilon)) \cap K$.
- If the scheme stays in K, it has to eventually stay in one of the stripes.
- Control $\mathbb{E}[\mathbf{1}\{X \text{ stays in stripe and no lower drop-out}\}(F(X_n) - F(z))]$.
- Similarly to ODE-argument one shows that $\mathbb{E}[\mathbf{1}\{X \text{ stays in stripe and no lower drop-out}\} \sum_{k=1}^{\infty} \gamma_k |\nabla F(X_k)|] < \infty$.
- If a lower drop-out occurs, $(F(X_n))_{n \in \mathbb{N}}$ will with high probability converge to a lower level.

V Conclusion

- SGD converges on $\mathbb{L}$ for Łojasiewicz-functions F under "standard" assumptions on perturbation.

V Conclusion

- SGD converges on $\mathbb{L}$ for Łojasiewicz-functions F under "standard" assumptions on perturbation.
- Limit points may also be saddle points or local maxima.

V Conclusion

- SGD converges on $\mathbb{L}$ for Łojasiewicz-functions F under "standard" assumptions on perturbation.
- Limit points may also be saddle points or local maxima.
- All real analytic functions are Łojasiewicz-functions.

V Conclusion

- SGD converges on $\mathbb{L}$ for Łojasiewicz-functions F under "standard" assumptions on perturbation.
- Limit points may also be saddle points or local maxima.
- All real analytic functions are Łojasiewicz-functions.
- Neural networks with analytic activation and loss function and compactly supported examples (X, Y) are Łojasiewicz-functions.

V Conclusion

- SGD converges on $\mathbb{L}$ for Łojasiewicz-functions F under "standard" assumptions on perturbation.
- Limit points may also be saddle points or local maxima.
- All real analytic functions are Łojasiewicz-functions.
- Neural networks with analytic activation and loss function and compactly supported examples (X, Y) are Łojasiewicz-functions.
- Objective function for which divergence may be observed on $\mathbb{L}$
  $\rightsquigarrow$ continuum of critical points that do not admit a Łojasiewicz-inequality (very restrictive!)

V Conclusion

- SGD converges on $\mathbb{L}$ for Łojasiewicz-functions F under "standard" assumptions on perturbation.
- Limit points may also be saddle points or local maxima.
- All real analytic functions are Łojasiewicz-functions.
- Neural networks with analytic activation and loss function and compactly supported examples (X, Y) are Łojasiewicz-functions.
- Objective function for which divergence may be observed on $\mathbb{L}$
  $\rightsquigarrow$ continuum of critical points that do not admit a Łojasiewicz-inequality (very restrictive!)
- The proofs are based on non-asymptotic estimates which might be of independent interest.