Optimization tools for deep-learning

S. Gerchinovitz[1], F. Malgouyres[2], **E. Pauwels**[3] & N. Thome[4]

[1] IRT Saint Exupéry, Toulouse
[2] Institut de Mathématiques de Toulouse, Université Toulouse 3 Paul Sabatier.
[3] Toulouse School of Economics, Université Toulouse 1 Capitole.
[4] Centre d'étude et de recherche en informatique et communication, Conservatoire national des arts et métiers.

Acknowledgements

Jérôme Bolte (Toulouse School of Economics):

We are looking for students!

Plan

Training a deep network

Finite dimensional optimization problem

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

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

Notations:

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

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

Optimizer zoo

TensorFlow

- Adadelta (> 2010)
- Adagrad (> 2010)
- Adam (> 2010)
- AdamW (> 2010)
- Adamax (> 2010)
- Ftrl (> 2010)
- Nadam (> 2010)
- RMSprop (> 2010)
- SGD (1951)

PyTorch

- Adadelta (> 2010)
- Adagrad (> 2010)
- Adam (> 2010)
- AdamW (> 2010)
- SparseAdam (> 2010)
- Adamax (> 2010)
- Averaged SGD (90's)
- LBFGS (70's)
- RMSprop (> 2010)
- Rprop, signs (90's)
- SGD (1951)

Main question

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

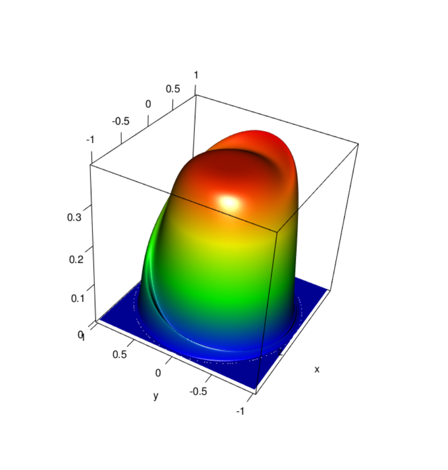
Compositional structure of deep network: Computing a (stochastic)-gradient of F has a cost comparable to evaluating F.

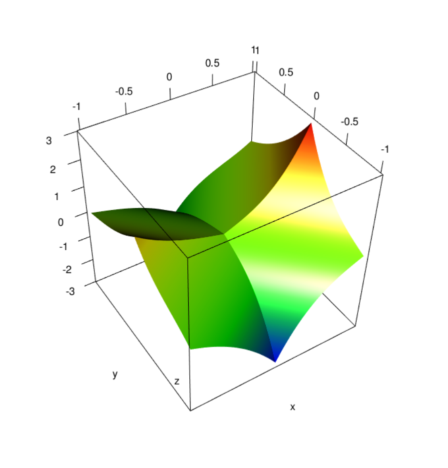
Deep nets are trained with variants of gradient descent.

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

Long term behaviour for this recursion?

Non convexity, non smoothness





Roadmap: longterm behavior of gradient descent

Main difficulty: The objective term is not convex, $(a, b) \mapsto ab$ is not convex, and may be not smooth. .

Long history in mathematics.
Foundations from two fields:

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

Program for today:

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

Plan

Main idea

Smooth dynamical systems

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

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

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

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

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

The gradient mapping is a diffeormorphism

Constructive proof:

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

$$\|x - y\| = \alpha \|\nabla F(x) - \nabla F(y)\| \leq L\alpha \|x - y\|, \qquad L\alpha < 1 \text{ hence } x = y.$$

- Explicit inverse: solution to the strictly convex problem,

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

Quizz: linear isomorphisms

Convergence to 0?

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

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

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

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

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

Extension to any square matrix using Jordan normal form.

Stable manifold theorem

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

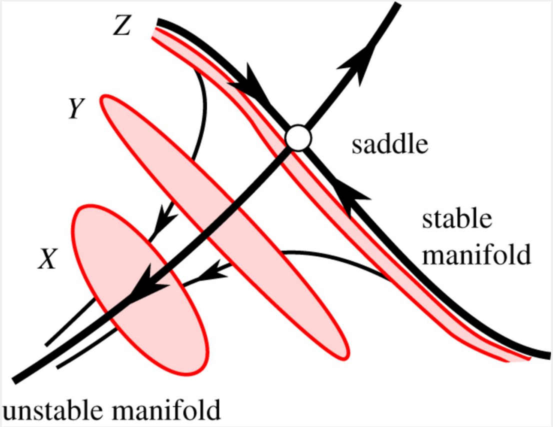
Theorem (*e.g.* Schub's book [S1987]): Let $T\colon \mathbb{R}^p \to \mathbb{R}^p$ be a local diffeomorphism $\bar{x}$ a fixed point of T such that $\nabla T(\bar{x})$ does not have any eigenvalue on the unit circle and at least one eigenvalue of modulus > 1.

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

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

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

With a picture



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

Convergence to local minima on Morse functions

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

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

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

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

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

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

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

Extension: Gradient Descent Only Converges to Minimizers

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

Plan

Deep learning training loss

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

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

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

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

Semi-algebraic sets and functions (SA)

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

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

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

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

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

is SA.

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

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

SA functions: examples

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

Tarski-Seidenberg Theorem

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

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

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

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

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

Recursively, eliminate a finite number of quantifier on variables.

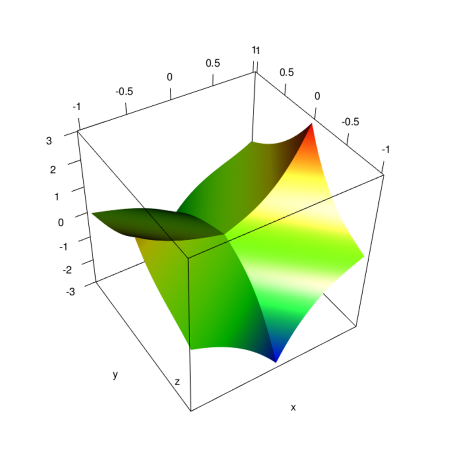
First order formula: quantification on real variables (not on sets), SA sets and functions, equality and inequality signs: $\left\{x, \forall y \leq 1, \exists z > 0, x^2 + y^2 + z = 1\right\}$.

Consequences: Any set or function described with a first order formula is SA.

- Image of SA map F: $\quad \mathrm{Im}F = \{y, \exists x, y = F(x)\}$
- Interior of SA set S: $\quad \mathrm{int}S = \{x, \exists \varepsilon > 0, \forall y \in B_{\varepsilon,x}, y \in S\}$.
- Derivatives of SA function f:
  $f'(x) = \left\{l, \forall \varepsilon > 0, \exists \delta > 0, \forall y \in B_{\delta,x}, |f(y) - f(x) - l(y-x)| \leq \varepsilon |y-x|\right\}$.

A lot more: Michel Coste's Introduction to o-minimal geometry [C2002].

Semi-algebraic sets and functions are "not pathological"



Univariate SA functions:

- Left and right limits.
- Continuous except at finitely many points.
- C^k except at finitely many points.
- Nicely structured (piecewise constant, increasing or decreasing)

Example: Morse-Sard theorem

Theorem: Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be SA differentiable, then critical values of f are finite:

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

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

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

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

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

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

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

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

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

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

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

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

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

Deep learning training loss

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

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

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

For most choices of L and activation functions, F is tame (sigmoid, logistic loss . . .).
$\rightarrow$ lots of qualitative properties

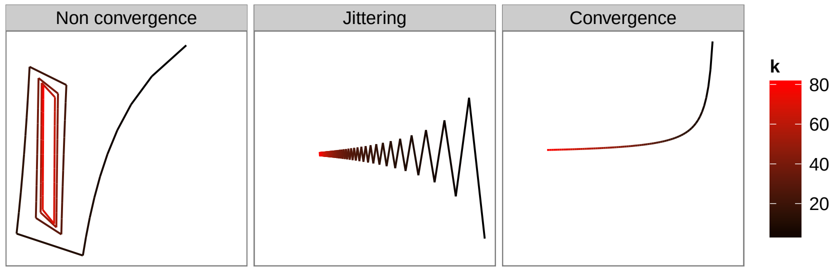
Plan

Introduction

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

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

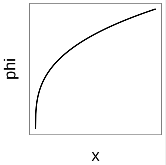
- Convergence of the iterates?



KL property (Łojasiewicz 63, Kurdyka 98)

Desingularizing functions on $(0, r_0)$

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



Definition

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

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

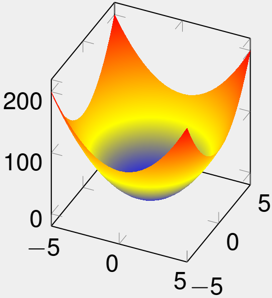
F has the KL property if this is satisfied for all $\bar{x}$.

Theorem (KL inequality holds for)

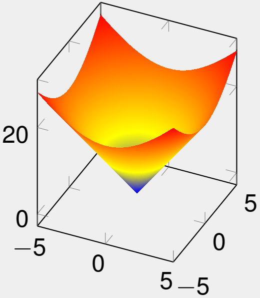
- *differentiable semi-algebraic functions (Łojasiewicz 1963 [Ł1963]).*
- *differentiable tame functions (Kurdyka 1998, [K1998]).*
- *nonsmooth tame functions (Bolte-Daniilidis-Lewis-Shiota 2007 [BDLS2007]).*

Illustration F and $\varphi \circ F$

F and $\varphi \circ F$



Parameterize with φ sharpens the function
$\longrightarrow$



KL inequality examples

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

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

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

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

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

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

$$\geq F(x) + \min_y \langle \nabla F(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2 = F(x) - \frac{1}{2\mu}\|\nabla F(x)\|^2$$

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

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

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

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

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

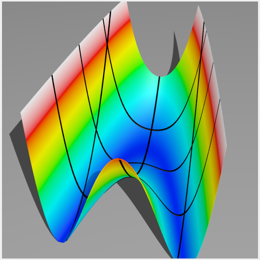
Morse functions: F $\mathcal{C}^2$ such that $\nabla F(x) = 0$ implies $\nabla^2 F(x)$ is non singular.

Non convex example: $F(x) = \|H(x)\|^2$,
$H\colon \mathbb{R}^p \to \mathbb{R}^n$, J_H surjective.

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

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

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



Convergence under KL assumption

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

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

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

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

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

Mexican hat function: Convergence of the sequence may not occur (without KL) even for C^∞ losses.

Proof sketch of convergence under KL assumption

Descent Lemma: For any $y, x \in \mathbb{R}^p$, $F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$.

Accumulation points are critical: From the descent Lemma, we have
$F(x_{k+1}) \leq F(x_k) - \frac{1}{2L}\|\nabla F(x_k)\|^2$,
$F(x_k)$ decreases and converge, say to $\bar{F} = 0$

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

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

KL function: Fix $\bar{x}$ a critical point with $F(\bar{x}) = 0$ (*e.g.* accumulation point).
There is $\varepsilon > 0$ and ϕ desingularizing function such that

$$\begin{cases} \|x - \bar{x}\|_2 < \varepsilon \\ 0 < F(x) \end{cases} \qquad \Rightarrow \qquad \|\nabla \phi \circ F(x)\|_2 = \phi'(F(x))\|\nabla F(x)\|_2 \geq 1$$

Note: if $F(x_k) = 0$ for some k, the sequence is stationary. Say $F(x_k) > 0$ for all k

Proof sketch: the trap mechanism

Assume $\|x_k - \bar{x}\| + 2\phi(F(x_k))) < \varepsilon$ (*e.g.* $\bar{x}$ accumulation point, k large enough)

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

By recursion, for any $K \geq k$

$$
\|x_{K+1} - \bar{x}\|_2 \leq \|x_k - \bar{x}\|_2 + \sum_{i=k}^{K} \|x_{i+1} - x_i\| < \|x_k - \bar{x}\|_2 + 2(\phi(F(x_k)) - \phi(F(x_{K+1}))) < \varepsilon.
$$

$\sum_{i=k}^{+\infty} \|x_{i+1} - x_i\|$ is finite, the sequence is Cauchy and converges.

Variant: KL 1/2 and linear convergence

Assume $\|x-\bar{x}\|_2 < \varepsilon, F(x) > 0 \quad \Rightarrow \quad \|\nabla F(x)\| \geq \frac{1}{\varphi'(F(x_i))}$

Assume furthermore, that $\varphi(t) = 2\sqrt{\frac{t}{c}}$, for some $c > 0$,

then if $\|x_k - \bar{x}\| + 2\varphi(F(x_k))) < \varepsilon$, for all $i \geq k$

$$\|\nabla F(x_i)\|^2 \geq \frac{1}{\varphi'(F(x_i))^2} = cF(x_i) > 0$$

We obtain

$$\begin{aligned} F(x_{i+1}) &\leq F(x_i) - \frac{1}{2L}\|\nabla F(x_i)\|^2 && \text{descent Lemma} \\ &\leq F(x_i)(1 - \frac{c}{2L}) && \text{KL } 1/2. \end{aligned}$$

which is linear convergence.

Remark:

- Does not require $F(\bar{x}) = 0$.
- If $\varepsilon = +\infty$, $\|\nabla F(x)\|^2 \geq cF(x)$ for all x, then $F(x_k)$ converges globally linearly to 0.

Data interpolation and overparameterization

Least squares loss: $H\colon \mathbb{R}^p \to \mathbb{R}^n$ C^2, $y \in \mathbb{R}^n$.

$$\min_{x\in\mathbb{R}^p} F(x) \quad := \quad \|H(x) - y\|^2$$

$$\frac{\|\nabla F(x)\|^2}{F(x)} = 2\frac{\|J_H(x)^T(H(x)-y)\|^2}{\|H(x)-y\|^2} \geq 2\lambda_{\min}(J_H(x)J_H(x)^T)$$

$$J_H J_H^T = \sum_{i=1}^{p} \left(\frac{\partial H}{\partial x_i}\right)\left(\frac{\partial H}{\partial x_i}\right)^T \in \mathbb{R}^{n\times n}$$

Overparameterization intuition: For $p \gg n$, J_H is surjective, $J_H J_H^T$ is invertible.
Assume $0 = H(0)$, then for any y,

$$\exists R, c > 0,\ \|\nabla F(x)\|^2 \geq cF(x),\ \forall x,\ \|x\| \leq R.$$

For y and x_0 small enough, $4\sqrt{\frac{F(x_0)}{c}} + \|x_0\| < R \quad \to$ trap, global linear convergence.

Main challenge: Estimate R and c.

Classical trap argument, revisited in a deep learning context

Two global convergence results

$$\exists R, c > 0, \|\nabla F(x)\|^2 \geq cF(x), \forall x, \|x\| \leq R.$$

GRADIENT DESCENT PROVABLY OPTIMIZES OVER-PARAMETERIZED NEURAL NETWORKS

Simon S. Du* Xiyu Zhai* Barnabás Póczos Aarti Singh

One hidden layer: Random initialization x.
For large p, quantitative gradient domination, controlled w.h.p. c, R.
Trap mechanism, linear convergence. Many extensions.

Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot Franck Gabriel Clément Hongler

Infinite width limit: Gradient domination in function space. NTK: $J_H J_H^T$, stabilizes and remains constant during training.

Lazy training: both theories suggest that the length of the trajectory is very small.

Non exhaustive historical landmarks

1963. Lojasiewicz: analyticity $\Rightarrow$ Lojasiewicz gradient inequality. Trap mechanism.

1963. Polyak: gradient domination (exponent $1/2$) implies linear convergence of GD.
Geometers mention Lojasiewicz's inequality.
Russian optimizers mention gradient dominated function.

1998. Kurdyka, generalizes Lojasiewicz arguments to o-minimal structures.

2005. Absil, Mahony, Andrews, Lojasiewicz's trap for GD on analytic losses.

2005's. Bolte, Daniilidis, Lewis, Shiota, Kurdyka's argument for nonsmooth losses.
Introduce the name Kurdyka-Lojasiewicz inequality.

2010's. Bolte *et. al.* Extend the trap argument to many algorithm. Connection with error bounds. Convergence rates. Complexity for convex optimization ...

2015. Karimi, Nutini, Schmidt. Revisit Polyak's arguments in an ML context.
Introduce the name Polyak-Lojasiewicz inequality for
global gradient domination with power $1/2$.

Since: trap argument ubiquitous in global training of overparameterized networks.
under the name "PL" inequality.

Plan

Stochastic gradient

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

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

Gradient sampling. $(i_k)_{k\in\mathbb{N}}$ *iid* RVs uniform on $\{1, \dots, n\}$. Stochastic gradient. Let $(M_k)_{k\in\mathbb{N}}$ be a martingale difference sequence.

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

Stochastic approximation: Robbins and Monro 1951 [RM1951].

The ODE method

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

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

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

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

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

(**Exercise:** If F is bounded below, then the gradient flow is well defined for all $t > 0$)

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

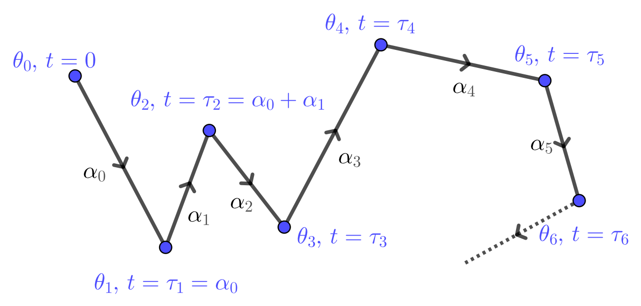
Piecewise affine interpolated process

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

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

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

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



A result of Benaim: flow attracts interpolated process

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

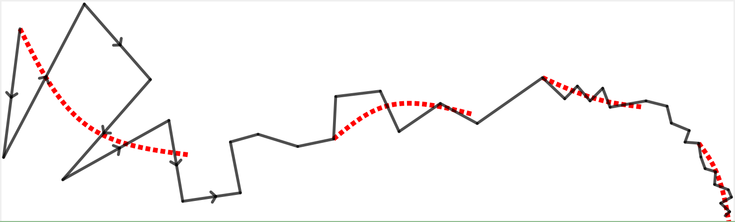
Bounded conditional variance: $\exists M \geq 0$ such that $\sup_{k\in\mathbb{N}} \mathbb{E}\left[M_{k+1}^2|\text{past}\right] \leq M$.

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

Theorem (Benaim 1996 [B1996])

Assume that there is $C > 0$ such that $\sup_k \|\theta_k\| \leq C$ almost surely. Then for all $T > 0$: Set $y_t\colon s \to w(t+s)$, $0 \leq s \leq T$, $\mathrm{dist}(y_t, \mathrm{flow}) \underset{t\to\infty}{\to} 0$, sup norm on $[0, T]$.

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



Consequence: descent in the limit

Lemma

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

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

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

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

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

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

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

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

Consequence: limit values and critical points

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

Lemma

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

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

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

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

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

Descent in the limit: $\exists \delta > 0$, $(l_i)_{i\in\mathbb{N}}$, $k_i \leq l_i$ and for i large enough,
$F(\theta_{l_i}) \leq \bar{f} - \delta$ and $\|\theta_k - \bar{\theta}\| \leq \varepsilon$, $k = k_i, \dots, l_i$. $l_i \leq m_i$, contradiction.

Corollary for deep learning

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

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

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

Main ingredients:

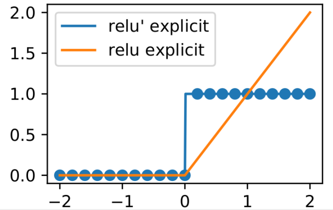
- Common neural networks are tame (semialgebraic). F is definable.
- **Definable Morse-Sard theorem:** the set F^* of critical values of F is finite.
- $\Omega \subset F^*$ is an interval. It is a singleton. $F(\theta_k)$ converges.
- Accumulation points are critical, otherwise descent in the limit implies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

Plan

Nonsmoothness is needed

Differentiate programs: `if ... then ...` or `while`

```
def myRelu(x):
    if x<=0:
        return 0
    else:
        return x
```



Massive practice: elementary functions: relu, maxpool, sort, implicit layers
Ex: 75% of `torchvision` models.

Stochastic subgradient

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

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

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

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

∂ is a suitable generalization of the gradient.

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

Convex: only for F convex, global lower tangent.

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

Fréchet: local lower tangent.

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

Limiting: sequential closure.

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

Clarke: convex closure.

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

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

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

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

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

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

We will work with Clarke subgradients.

Fundamental theorem of Lebesgue integral

Theorem

Univariate locally lipschitz (absolutely continuous) functions are differentiable almost everywhere and are the integral of their derivative: $\forall t, a$

$$f(t) = f(a) + \int_a^t f'(s)ds.$$

Differential inclusion solutions

F Lipschitz: gradient ODE replaced by a differential inclusion.

Definition

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

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

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

Example: absolute value.

Theorem (*e.g.* Filippov, Aubin Celina): Convexity and continuity (upper semi-continuity) properties of ∂F ensures existence of solution (not unique).

Chain rule along Lipschitz curves

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

Lemma (Brézis' convex chain rule, Lyapunov function)

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

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

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

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

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

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

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

Chain rule for convex functions

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

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

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

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

On the other hand, still considering h positive

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

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

Corollary for deep learning

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

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

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

Proof arguments: Same idea as in the smooth case

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

Opening

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

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

Plan

Main question

$$\min_{\theta \in \mathbb{R}^p} \qquad F(\theta) = \frac{1}{n}\sum_{i=1}^{n} l_i(\theta) \tag{2}$$

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

Deep nets are trained with variants of gradient descent. Long term behaviour:

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

- Avoidance of strict saddle points, stable manifold.
- Rigidity structure of neural net training losses (semi-algebraic).
- Sequential convergence of GD under with KL inequality. Global convergence for least squares.
- The ODE method for stochastic approximation.
- Extension to nonsmooth losses.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Du, Zhai, Poczos and Singh.
Gradient Descent Provably Optimizes Over-parameterized Neural Networks. ICLR 2019.

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

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

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

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

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

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

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

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