A semi-algebraic look at first-order methods

Jérôme Bolte

Université de Toulouse / TSE

Nesterov's 60th birthday, Les Houches, 2016

Start with a reasonable FOM (some splitting method, gradient projection method, alternating projection method, Lagrangian method...):

- ✓ **Criticality / necessary optimality conditions** (discrete Lasalle's invariance principle called "Zangwill's theorem")
- ? But what about convergence guarantees for the iterates?
- ? Decrease rates for the iterates/value?

Answer: well-designed notions of piecewise smoothness? Does not work !

Smooth counter-examples: Palis-De Melo, an extension Absil et al.

There exist

- $f : \mathbb{R}^2 \to \mathbb{R}$ C^∞ coercive nonnegative with $\nabla f^{-1}(\{0\}) = \operatorname{argmin} f$ =unit disk
- a bounded gradient sequence with constant step $s > 0$ (as small as we want):

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

such that

$$f(x_k) \downarrow \min f = 0, \ \nabla f(x_k) \to 0$$

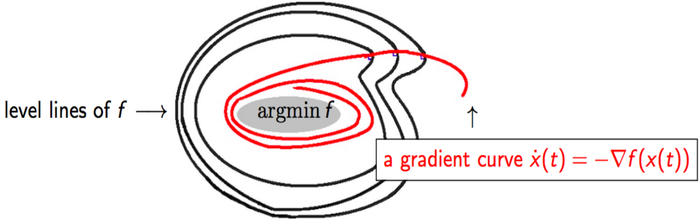
but with this awful property

The set of limit-points of x_k is the unit circle

Smooth counter-examples: Palis-De Melo, an extension Absil et al.

Easier to understand with $\dot{x}(t) = -\nabla f(x(t))$

Idea: Spiraling bump yields spiraling curves



A function that yields similar results:

$$f(r,\theta) = \exp(1-r^2)\left(1 - \frac{r^4}{r^4 + (1-r^2)^4}\right)\sin\left(\theta - \frac{1}{1-r^2}\right)$$

Source of counter-examples: oscillations

Oscillations in Optimization:

- Oscillations $\Longrightarrow$ gradient direction=very bad predictors
- arbitrarily bad rates/complexity for any fixed dimension...

– Even worst behaviors for nonsmooth functions

– Same awful behaviors for more complex methods: e.g. Forward-Backward

Solutions to the **oscillation issue**? Alternative framework?

Solutions to these bad behaviors?

Topological approach: avoid wild oscillations. Use semi-algebraic (or definable) functions

Too vague !!

Ad hoc approach: study deformations of a class of nonsmooth functions "sharp function"

Too specific ??

Semi-algebraic objects

- Defined by finitely many polynomials
- Easy to recognize (Tarski-Seidenberg, quantifier elimination)
- Very stable (image/pre-image, derivation, composition, subdifferentiation...)
- Oscillations are controlled: monotonicity lemma/"finiteness of the number of connected components.

  Take $A \subset \mathbb{R}^p \times \mathbb{R}^n$ and $A_x = \{y \in \mathbb{R}^p : (x, y) \in A\}$.

  There exists $N \in \mathbb{N}$ such that $\mathrm{cc}(A_x) \leq N$, $\forall x \in \mathbb{R}^p$

Concrete examples

- Polynomials $\frac{1}{2}||Ax-b||^2$, $(A,B)\to\frac{1}{2}||AB-M||^2$
- max or min of polynomials
- rank, ℓ^p norms (p rational or $p=\infty$/infinity norm, $p=0$/zero norm)
- standard cones: $\mathbb{R}^n_+$, Lorenz cone, SDP...

What about non semi-algebraic problems?

- analytic functions, e.g., $\log\det$
- ℓ^p norm with p arbitrary
- ...

a similar theory holds:

o-minimality van den Dries / Shiota: global subanalyticity, Dirichlet series, log-exp structure...

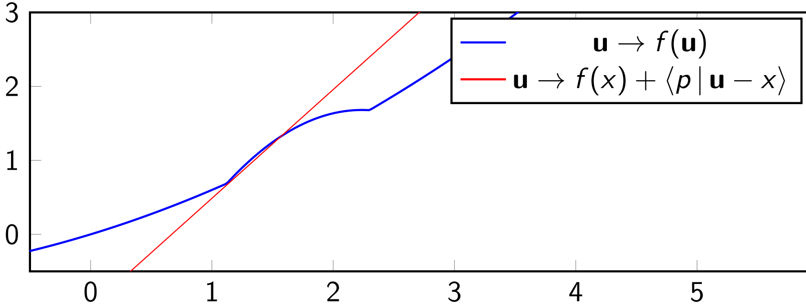
Fréchet subdifferential

$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lsc proper.

First-order formulation: Let x in dom f

(Fréchet) subdifferential : $p \in \hat{\partial} f(x)$ iff

$$f(\mathbf{u}) \geq f(x) + \langle p \,|\, \mathbf{u} - x \rangle + o(||\mathbf{u} - x||), \ \forall \mathbf{u} \in \mathbb{R}^n.$$



Subdifferential

Definition (Mordhukovich 76)

It is denoted by ∂f and defined through:
$x^* \in \partial f(x)$ iff $(x_k, x_k^*) \to (x, x^*)$ such that $f(x_k) \to f(x)$ and

$$f(u) \geq f(x_k) + \langle x_k^* \mid u - x_k \rangle + o(\|u - x_k\|).$$

Example: $f(x) = \|x\|_1$, $\partial f(0) = B_\infty = [-1, 1]^n$
Set

$$\|\partial f(x)\|_- = \min\{\|x^*\| : x^* \in \partial f(x)\}$$

Properties (Critical point)

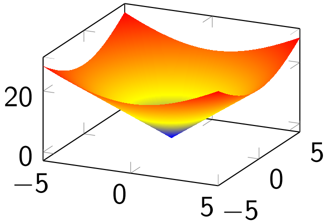
Fermat's rule if f has a minimizer at x then $\partial f(x) \ni 0$.
Conversely when $0 \in \partial f(x)$, the point x is called critical.

An elementary remedy to “gradient oscillation”: Sharpness

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called *sharp on the slice* $[r_0 < f < r_1] := \{x \in \mathbb{R}^n : r_0 < f(x) < r_1\}$, if there exists $c > 0$

$$\boxed{\|\partial f(x)\|_- \geq c > 0, \quad \forall x \in [r_0 < f < r_1]}$$

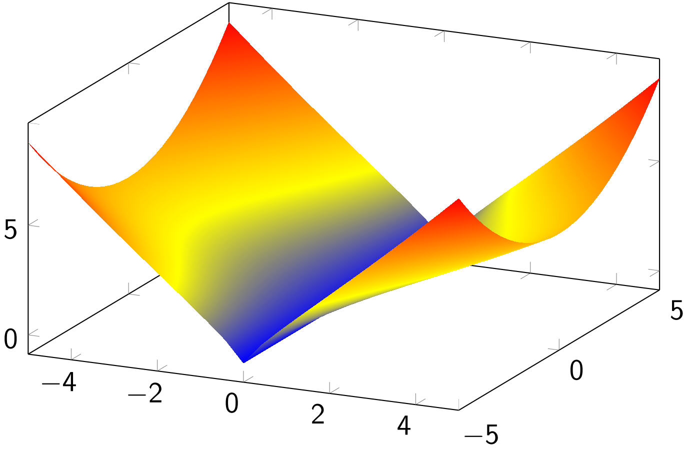
Basic example $f(x) = \|x\|$



Many works since 78, Rockafellar, Polyak, Ferris, Burke and many many others...

Sharpness: example

Nonconvex illustration with a continuum of minimizers



Finite convergence

Why is sharpness a remedy?

Slopes towards the "minimizers" overcome frankly other "parasite slopes"

- **Gradient curves reach "the valley" within a finite time.**
- **Easy to see the phenomenon on proximal descent**

  Prox descent = formal setting for implicit gradient

$$x^+ = x - \mathrm{step}.\, \partial f(x^+)$$

Prox operator:

$\mathrm{prox}_f^s x = \mathrm{argmin}\{f(u) + \frac{1}{2s}||u - x||^2 : u \in \mathbb{R}^n\}$

(Moreau)

$$x^+ = \mathrm{prox}_f^{\mathrm{step}}(x)$$

When f is sharp, convergence occurs in finite time !!

Proof

Assume f is sharp, then if x^{k+1} is non critical $f(x^{k+1}) \leq f(x^k) - \delta$.

Write (assume "step" is constant)

- $f(x^{k+1}) \leq f(x^k) - \|x^{k+1} - x^k\|^2$
- $x^{k+1} - x^k \in step.\,\partial f(x^{k+1})$ thus $\|x^{k+1} - x^k\| \geq c.step$
- $f(x^{k+1}) \leq f(x^k) - (c.step)^2$
  Set $\delta = (c.step)^2$.

Measuring the default of sharpness

0 is a critical value of f (true up to a translation).
Set $[0 < f < r_0] := \{x \in \mathbb{R}^n : 0 < f(x) < r_0\}$ and **assume that there are no critical points in $[0 < f < r_0]$.**

f **has the KL property on** $[0 < f < r_0]$ if there exists a function $g : [0 < f < r_0] \to \mathbb{R} \cup \{+\infty\}$ whose sublevel sets are those of f, *ie* $[f \leq r]_{r \in (0,r_0)}$, and such that

$$||\partial g(x)||_- \geq 1 \text{ for all } x \text{ in } [0 < f < r_0].$$

$\boxed{\textbf{EX}}$ **(Concentric "ellipsoids")** $f(x) = \frac{1}{2}\langle Ax, x\rangle$ **and** $g(x) = \sqrt{f}$

Formal KL property

Formally :

Desingularizing functions on $(0, r_0)$:
$\varphi \in C([0, r_0), \mathbb{R}_+)$**, concave,** $\varphi \in C^1(0, r_0)$**,** $\varphi' > 0$ **and** $\varphi(0) = 0$**.**

Definition

f has the KL property on $[0 < f < r_0]$ if there exists a desingularizing function φ such that

$$||\partial(\varphi \circ f)(x)||_- \geq 1, \; \forall x \in [0 < f < r_0].$$

Local version : replace $[0 < f < r_0]$ by the intersection of $[0 < f < r_0]$ with a closed ball.

Some results in the "smooth world"

- **Theorem [Łojasiewicz 1963/Kurdyka 98]** If f is analytic/"o-minimal", it is KL with $\varphi(s) = Ks^{1-\theta}$.
- Gradient dominated functions of **Polyak 63**, for f convex

$$\|\nabla f(x)\|^2 \geq K(f(x) - \min f)$$

KL with $\frac{1}{\sqrt{K}}\sqrt{s}$. Many ideas were present...

- Y. **Ol'Khoovsky** (1972) analyzed the gradient method (Absil, Mahony, Andrews 05 similar results independently)
- In 2006 **Polyak & Nesterov 06**, introduced

$$\|\nabla f(x)\|^p \geq K(f(x) - \min f)$$

for complexity purposes of second/third order methods. This is exactly Łojasiewicz inequality when $p > 1$.

Going beyond smoothness/analyticity?

Theorem (B-Daniilidis-Lewis 2006)

Nonsmooth semialgebraic/subanalytic case 2006: *Take $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ lower semicontinuous and semialgebraic then f has KL property around each point.*

Many many functions satisfy KL inequality see:

– B-Daniilidis-Lewis-Shiota 2007
– B-Daniilidis-Ley-Mazet 2010

In Optimization see also but also Attouch, B., Redont 2010, B-Sabach-Teboulle 2014...

Brief “technical” comments on KL

The ingredients

- A nonsmooth Sard’s theorem: finiteness of critical values
- “Amenability to sharpness” is equivalent to the fact that there exists a talweg (“path in the valley”) of finite length a result from B., Daniilidis, Ley, Mazet.
- In the semi-algebraic world, desingularization functions are of the form $\varphi(s) = cs^{1-\theta}$ this is Puiseux Lemma.

How all this impacts optimization?

Descent methods at large (?P. Tseng?)

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function; $a, b > 0$.
Let x^k be a sequence in dom f such that

Sufficient decrease condition

$$f(x^{k+1}) + a\|x^{k+1} - x^k\|^2 \leq f(x^k); \quad \forall k \geq 0$$

Relative error condition For each $k \in \mathbb{N}$, there exists $w^{k+1} \in \partial f(x^{k+1})$ such that

$$\|w^{k+1}\| \leq b\|x^{k+1} - x^k\|;$$

Convergence theorem

Theorem (Attouch-B-Svaiter 2012 / B. Sabach-Teboulle 13)

Let f be a KL function and x^k a descent subgradient sequence for f.
If x^k is bounded then it converges to a critical point of f.

Corollary

Let f be a coercive semi-algebraic function and x^k a descent sequence for f.
Then x^k converges to a critical point of f.

Rate of convergence

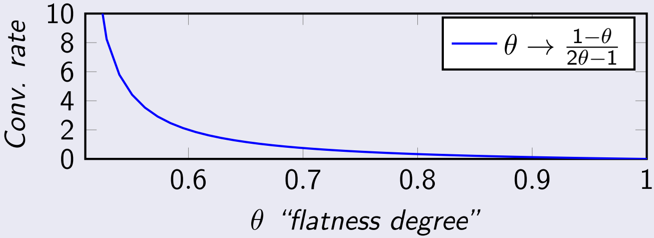
Assume that $\varphi(s) = cs^{1-\theta}$ with $c > 0$, $\theta \in [0,1)$.

Theorem

(i) If $\theta \in (0, \frac{1}{2}]$ then there exist $d > 0$ and $Q \in [0,1)$ such that $\|x^k - x^\infty\| \leq d\, Q^k$,

(ii) If $\theta \in (\frac{1}{2}, 1)$ then there exists $c_1, c_2 > 0$ such that

$$\|x^k - x^\infty\| \leq c_1\, k^{-\frac{1-\theta}{2\theta-1}},\;\; f(x_k) - f(x_\infty) \leq c_2\, k^{-\frac{1}{2\theta-1}} = o\left(\frac{1}{k}\right)$$



Forward-backward splitting algorithm

From now on all objects are assumed to be SA

Minimizing nonsmooth+smooth structure: $f = g + h$

with $\begin{cases} h \in C^1 \text{ and } \nabla h \quad L - \text{Lipschitz continuous} \\ \text{g lsc bounded from below + prox is easily computable} \end{cases}$

Forward-backward splitting (Lions-Mercier 79): Let γ_k be such that $0 < \underline{\gamma} < \gamma_k < \overline{\gamma} < \frac{1}{L}$

$$x^{k+1} \in \operatorname{prox}_{\gamma_k g}\left(x^k - \gamma_k \nabla h(x^k)\right).$$

Theorem

If the problem is coercive x^k is a converging sequence

Gradient projection

C a semi-algebraic set, f a semi-algebraic function.

$$x^{k+1} \in P_C\left(x^k - \gamma_k \nabla f(x^k)\right)$$

Theorem

If the sequence x^k is bounded it converges to a critical point of the problem $\min_C f$

Example: Inverse problems with sparsity constraints:

$$\min\left\{\frac{1}{2}\|Ax - b\|^2 : x \in C\right\}$$

C=sparsity constraint/rank constraint/simple constraints

von Neumann alternating projections?

Assumptions: C semi-algebraic closed and D semi-algebraic and convex.
Example: D=Hankel matrices and C=matrices of rank lower than r.

$$x^{k+1} \in P_C(y_k), \quad y_{k+1} = P_D(x_k).$$

Careful oscillations are possible... A circle for C and its center $D = \{\text{center}\}$
A solution under-relax:

Theorem

Bounded sequences of the form $x^{k+1} \in P_C\Big(\epsilon x_k + (1-\epsilon)P_D(x_k)\Big)$ *are converging.*

More subtle results by Noll-Rondepierre in the same category based on nonsmooth KL

Averaged projections with underrelaxation

$$x^{k+1} \in (1-\theta)\, x^k + \theta \left(\frac{1}{p} \sum_{i=1}^{p} P_{F_i}(x^k) \right), \theta \in]0,1[$$

Theorem (Averaged projection method)

$F_1, \ldots, F_p$ be closed semi-algebraic which satisfy $\bigcap_{i=1}^p F_i \neq \emptyset$. If x^0 is sufficiently close to $\bigcap_{i=1}^p F_i$, then x^k converges to a feasible point $\bar{x}$, i.e. *such that*

$$\bar{x} \in \bigcap_{i=1}^{p} F_i.$$

Alternating versions of prox algorithm

$F : \mathbb{R}^{m_1} \times \ldots \times \mathbb{R}^{m_p} \to \mathbb{R} \cup \{+\infty\}$ lsc semi-algebraic

Structure of F :

$$F(x_1, \ldots, x_p) = \sum_{i=1}^{p} f_i(x_i) + Q(x_1, \ldots, x_p)$$

f_i are proper lsc and Q is C^1.

Gauss-Seidel method/ Prox **(Auslender 1993)**

$$x_1^{k+1} \in \underset{\mathbf{u} \in \mathbb{R}^{m_1}}{\operatorname{argmin}} \quad F(\mathbf{u}, x_2^k, \ldots, x_p^k)$$

$$\ldots$$

$$x_p^{k+1} \in \underset{\mathbf{u} \in \mathbb{R}^{m_p}}{\operatorname{argmin}} \quad F(x_1^{k+1}, \ldots, x_{p-1}^{k+1}, \mathbf{u})$$

Alternating versions of prox algorithm

$F : \mathbb{R}^{m_1} \times \ldots \times \mathbb{R}^{m_p} \to \mathbb{R} \cup \{+\infty\}$ lsc semi-algebraic

Structure of F :

$$F(x_1, \ldots, x_p) = \sum_{i=1}^{p} f_i(x_i) + Q(x_1, \ldots, x_p)$$

f_i are proper lsc and Q is C^1.

Gauss-Seidel method/ Prox **(Auslender 1993)**

$$x_1^{k+1} \in \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^{m_1}} \quad F(\mathbf{u}, x_2^k, \ldots, x_p^k) + \frac{1}{2\mu^1}\|\mathbf{u} - x_k^1\|^2$$

$$\ldots$$

$$x_p^{k+1} \in \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^{m_p}} \quad F(x_1^{k+1}, \ldots, x_{p-1}^{k+1}, \mathbf{u}) + \frac{1}{2\mu^p}\|\mathbf{u} - x_k^p\|^2$$

Proximal Gauss-Seidel

$$F(x_1, \ldots, x_p) = \sum_{i=1}^{p} f_i(x_i) + Q(x_1, \ldots, x_p)$$

f_i are proper lsc and Q is C^1.

$$x_1^{k+1} \in \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^{m_1}} \quad f_1(\mathbf{u}) + Q(\mathbf{u}, \ldots, x_p) + \frac{1}{2\mu^1}\|\mathbf{u} - x_k^1\|^2$$

$$x_p^{k+1} \in \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^{m_p}} \quad f_p(\mathbf{u}) + Q(x_1, \ldots, \mathbf{u}) + \frac{1}{2\mu^p}\|\mathbf{u} - x_k^p\|^2$$

Theorem

Bounded sequences of the prox Gauss-Seidel method converge

Proximal alternating linearized method: PALM (B. Sabach, Teboulle)

$F(x) = Q(x_1, x_2) + f_1(x_1) + f_2(x_2)$
(B., Sabach, Teboulle, 14) linearization idea:

$$x_1^{k+1} \in \underset{\mathbf{u} \in \mathbb{R}^{m_1}}{\operatorname{argmin}} \quad f_1(\mathbf{u}) + \langle \nabla Q(\mathbf{u}, x_2) | u - x_1 \rangle + \frac{1}{2\mu^1} \|\mathbf{u} - x_k^1\|^2$$

$$x_2^{k+1} \in \underset{\mathbf{u} \in \mathbb{R}^{m_2}}{\operatorname{argmin}} \quad f_2(\mathbf{u}) + \langle \nabla Q(x_1, \mathbf{u}) | u - x_1 \rangle + \frac{1}{2\mu^2} \|\mathbf{u} - x_k^2\|^2$$

$$x_1^{k+1} \in \operatorname{prox}_{\frac{1}{\mu_1^k} f_1} \left(x_1^k - \frac{1}{\mu_1^k} \nabla_{x_1} Q(x_1^k, x_2^k) \right),$$

$$x_2^{k+1} \in \operatorname{prox}_{\frac{1}{\mu_2^k} f_1} \left(x_1^k - \frac{1}{\mu_2^k} \nabla_{x_2} Q(x_1^{k+1}, x_2^k) \right).$$

Good choice of steps?

Proximal alternating linearized method

$\forall x_1$, $Q(x_1,\cdot)$ is C^1 with $L_2(x_1)>0$ Lipschitz continuous gradient (same for x_2 with $L_1(x_2)$)

$$x_1^{k+1} \in \operatorname{prox}_{\frac{1}{L_1(x_2^k)} f_1}\left(x_1^k - \frac{1}{L_1(x_2^k)}\nabla_{x_1} Q(x_1^k, x_2^k)\right).$$
$$x_2^{k+1} \in \operatorname{prox}_{\frac{1}{L_2(x_1^k)} f_1}\left(x_1^k - \frac{1}{L_2(x_1^k)}\nabla_{x_2} Q(x_1^{k+1}, x_2^k)\right).$$

Example: Several applications in sparse NMF, blind deconvolution (Pesquet, Repetti...), dictionary learning (Gribonval, Malgouyres..). General structure: M a fixed matrix, r, s integers,
min $\left\{\frac{1}{2}\|AB-M\|^2 : \|A\|_0 \le r, \|B\|_0 \le s\right\}$

Theorem

Bounded sequences of PALM converge

Going further in this line?

- Complex problems à la SQP (with Pauwels)
- Lagrangian methods (with Sabach-Teboulle)
- **Alternating methods have a different nature (with Pauwels-Ngambou)** but there are works by Li, Noll-Rondepierre, Druviatskii-Ioffe-Lewis....
- Fast methods???

A flavour of this complications: A simple SQP/SCQP method

We wish to solve

$$\min\{f(x) : x \in \mathbb{R}^n, f_i(x) \leq 0\}$$

which can be written $\min f + i_C$ with $C = [f_i \leq 0, \forall i]$.

We assume:

∇f is L_f Lipschitz
$\forall i$, ∇f_i is L_{f_i} Lipschitz continuous.

Prox operator here are out of reach: too complex !!

Moving balls methods

Classical Sequential Quadratic Programming (SQP) idea, replace functions by some simple approximations.

Moving balls method (Auslender-Shefi-Teboulle, Math. Prog., 2010)

$$\min_{x \in \mathbb{R}^n} \; f(x_k) + f'(x_k)(x - x_k) + \frac{L_f}{2}||x - x_k||^2$$

$$f_1(x_k) + f_1'(x_k)(x - x_k) + \frac{L_{f_1}}{2}||x - x_k||^2 \leq 0$$

$$\ldots$$

$$f_m(x_k) + f_m'(x_k)(x - x_k) + \frac{L_{f_m}}{2}||x - x_k||^2 \leq 0$$

Bad surprise: x_k is not a descent sequence for $f + i_C$.

Moving balls methods

Introduce

$$\mathrm{val}(x) = \min_{y\in\mathbb{R}^n} \; f(x) + f'(x)(y-x) + \frac{L_f}{2}||y-x||^2$$

$$f_1(x) + f_1'(x)(y-x) + \frac{L_{f_1}}{2}||y-x||^2 \leq 0$$

$$\ldots$$

$$f_m(x) + f_m'(x)(y-x) + \frac{L_{f_m}}{2}||y-x||^2 \leq 0$$

Good surprise: x_k is a descent sequence for val.

Theorem

Assume Mangasarian-Fromovitz qualification condition. If x_k is bounded it converges to a KKT point of the original problem.