Proximal Alternating Linearized Minimization for Semi-algebraic Problems

Jérôme Bolte

TSE, Université Toulouse I Capitole

Joint work with

Shoham Sabach, Göttingen University
Marc Teboulle, Tel Aviv University

Sestri Levante, September 10th, Italia
The Optimization Model

Minimize a "block-nonsmooth" function

\[ \min \{ \psi (x, y) := f(x) + g(y) + H(x, y) : x \in \mathbb{R}^m, y \in \mathbb{R}^m \} \]

Assumption

(i) \( f, g \) proper lsc real-extended-valued functions.

(ii) \( H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R} \) is a \( C^1 \) function

(ii) \( \nabla_x H(\cdot, y) \) is \( L_1(y) \) Lipschitz continuous and \( \nabla_y H(x, \cdot) \) is \( L_2(x) \) Lipschitz continuous.

- the **Lipschitz constant** of \( \nabla_x H(\cdot, y) \) **depends on** \( y \); denoted \( L_1(y) \). Same with \( x \)...
- **NO convexity**
- Why two blocks of variables ? Simplicity !!
Target problems: Matrix Factorization, Blind Deconvolution or Dictionary Learning...

E.g. Matrix factorization: $A \in \mathbb{R}^{m \times n}$, $r \in \mathbb{N}$ fixed. Find $X \in \mathbb{R}^{m \times r}$ and $Y \in \mathbb{R}^{r \times n}$ such that

$$A \approx XY,$$

$$X \in \mathcal{F} \subset \mathbb{R}^{m \times r},$$

$$Y \in \mathcal{G} \subset \mathbb{R}^{r \times n}.$$

Using “merit functions” →

$$\min \left\{ \frac{1}{2} \| A - XY \|_F^2 : X \in \mathcal{F}, Y \in \mathcal{G} \right\}.$$

Introducing the indicator functions $f = i_{\mathcal{F}}$ and $g = i_{\mathcal{G}}$, we note that the problem matches our assumption with

$$L_1(Y) \equiv \| YY^T \|_F, \quad L_2(X) \equiv \| X^T X \|_F$$
Most optimization methods are about improving a given “state” \((x^0, y^0)\) by decomposing into simple & well adapted problems.

Indeed, we hardly know nothing about “exact” problem solving, save perhaps

- solving linear systems
- solving some linear/quadratic problems

Newton’s method, Cauchy gradient, Gauss-Seidel, Banach-Picard and their modern variants...

**HERE:** In view of the structure of our problem two strategies seem possible

- we decompose the variable space: \(x, y\) decomposition
- we decompose within the objective through the “smooth+nonsmooth” decomposition technique (aka forward-backward)
Space decomposition

\[(M) \quad \text{minimize} \left\{ \psi(x, y) := f(x) + g(y) + H(x, y) : x \in \mathbb{R}^m, y \in \mathbb{R}^m \right\} \]

**Alternating Minimization/Coordinate Descent “à la Gauss-Seidel”**

\[x^{k+1} \in \text{argmin} \psi(x, y^k) ; \quad y^{k+1} \in \text{argmin} \psi(x^{k+1}, y) .\]

“Simple” but...

- well-posedness issues
- Only convergence of subsequences can be derived...and under restrictive convexity-like assumptions... [Auslender (71), Powell (73),..., Grippo-Sciandrone (00)].
- **INCOMPLETE DECOMPOSITION in general:** Nested scheme involving possibly difficult subproblems
Space decomposition: proximal variant

\[
(M) \quad \text{minimize} \left\{ \psi(x, y) := f(x) + g(y) + H(x, y) : x \in \mathbb{R}^m, y \in \mathbb{R}^m \right\}
\]

Prox regularization of the "Gauss-Seidel" method (Auslender 92)

\[
\begin{align*}
\chi^{k+1} & \in \arg\min \left\{ \psi \left( \chi, y^k \right) + \lambda_k \| \chi - \chi^k \|_2 \right\}; \\
y^{k+1} & \in \arg\min \left\{ \psi \left( \chi^{k+1}, y \right) + \nu_k \| y - y^k \|_2 \right\}.
\end{align*}
\]

“Quite simple” also and furthermore

- well-posed under very weak assumptions
- Convergence can be derived for semi-algebraic functions...and under restrictive assumptions... [Attouch, Bolte, Redont, Soubeyran 2011].

INCOMPLETE DECOMPOSITION in general: nested scheme involving possibly difficult subproblems
Smooth + Nonsmooth decomposition

The Proximal-Forward Backward Scheme / Proximal Gradient
Forget about blocks: one vectorial variable

Composite smooth + simple nonsmooth model:

\[
(P) \quad \min \{ \sigma(u) + h(u) : u \in \mathbb{R}^n \}, \quad h \in \text{smooth, i.e. } C^{1,1}, \sigma \text{ nonsmooth}.
\]

\[
(P_k) \quad u^{k+1} \in \arg\min \left\{ \sigma(u) + h(u^k) + \langle u - u^k, \nabla h(u^k) \rangle + \frac{1}{2}\text{.step} \|u - u^k\|^2 : u \in \mathbb{R}^d \right\}.
\]

- Origin: [Passty (79), Lions-Mercier (79)…] / see also gradient projection
- Convex case well understood, convergence and complexity [ Lions-Mercier (79), Combettes-Wajs (05), Nesterov (07), Beck-Teboulle (09)……]
Let $\sigma : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ be a proper and lsc function. Given $x \in \mathbb{R}^n$ and $t > 0$, the proximal map is a point-to-set map defined by:

$$\operatorname{prox}_t^\sigma (x) := \arg\min \left\{ \sigma (u) + \frac{t}{2} \| u - x \|^2 : u \in \mathbb{R}^n \right\} \subset \mathbb{R}^n.$$ 

**Proposition (Well-definedness of proximal maps)**

*If $\sigma$ is bounded from below then $\operatorname{prox}_t^\sigma (x)$ is nonempty and compact for each positive $t$.*

**Remark** $X$ is a subset of $\mathbb{R}^n$, then:

$$\operatorname{prox}_t^{i_X} = \operatorname{proj}_X$$

the set-valued projection operator onto $X$. 

Proximal Map for Nonconvex Functions
A synthetic expression for the Proximal-Forward Backward Scheme

\[
(P) \quad \min \ \{ h(u) + \sigma(u) : u \in \mathbb{R}^n \}, \quad h \in \text{smooth, i.e. } C^{1,1}
\]

\[
(P_k) \quad u^{k+1} \in \arg\min \{ h(u^k) + \langle u - u^k, \nabla h(u^k) \rangle + \frac{c_k}{2} \| u - u^k \|^2 + \sigma(u) \}.
\]

Reformulation of PFB:

\[
u^{k+1} \in \text{prox}_{\sigma c_k} \left( x^k - \frac{1}{c_k} \nabla h(x^k) \right).
\]

- **ACTUAL DECOMPOSITION** for “simple” nonsmooth \( \sigma(\cdot) \), i.e., easy prox.
- \( \nabla h \) must be globally Lipschitz else the local model is not relevant
- **Nonconvex case:** Convergence of the whole sequence to a critical point!
  Very recent in [Attouch-B.-Svaiter (12)], under “some” assumption...More on this soon...
- Let’s apply it to our problem!!
In our case $\nabla H$ is not globally Lipschitz

- local model "are not relevant"
- stepsizes may either accumulate to zero and/or not lead to critical points

Idea combine

- space decomposition
- smooth+nonsmooth decomposition
The Algorithm: Proximal Alternating Linearization Minimization (Palm)

1. Take $\gamma_1 > 1$, set $c_k = \gamma_1 L_1(y^k)$ and compute

$$x^{k+1} \in \text{prox}_{c_k} \left( x^k - \frac{1}{c_k} \nabla_x H(x^k, y^k) \right).$$

2. Take $\gamma_2 > 1$, set $d_k = \gamma_2 L_2(x^{k+1})$ and compute

$$y^{k+1} \in \text{prox}_{d_k} \left( y^k - \frac{1}{d_k} \nabla_y H(x^{k+1}, y^k) \right).$$

Stepsizes $c_k^{-1}, d_k^{-1}$ are in $\left[ 0, 1/L_2(y^k) \right] \cap \left[ 0, 1/L_1(x^{k+1}) \right]$.

Main computational step: Computing the prox of a “simple” function.
Extra assumptions:

\[
\inf_{\mathbb{R}^n \times \mathbb{R}^m} \Psi > -\infty, \quad \inf_{\mathbb{R}^n} f > -\infty \quad \text{and} \quad \inf_{\mathbb{R}^m} g > -\infty.
\]

Quick Recall on Nonsmooth Analysis – [Rockafellar-Wets (98)] Let \( \sigma : \mathbb{R}^d \rightarrow (-\infty, +\infty] \) be a proper and lower semicontinuous function.

\( \star \) Fréchet Subdifferential:
\( x^* \in \partial \sigma(x) \) means that

\[
\sigma(u) \geq \sigma(x) + \langle x^* , u - x \rangle + o(\|u - x\|)
\]
Extra assumptions:

\[
\inf_{\mathbb{R}^n \times \mathbb{R}^m} \Psi > -\infty, \quad \inf_{\mathbb{R}^n} f > -\infty \quad \text{and} \quad \inf_{\mathbb{R}^m} g > -\infty.
\]

Quick Recall on Nonsmooth Analysis – [Rockafellar-Wets (98)] Let \( \sigma : \mathbb{R}^d \to (-\infty, +\infty] \) be a proper and lower semicontinuous function.

- **Limiting Subdifferential**: 
  \( x^* \in \partial \sigma(x) \) means that

\[
\sigma(u) \geq \sigma(x_k) + \langle x_k^*, u - x_k \rangle + o(\|u - x_k\|)
\]

with

\[
(x_k, x^*) \to (x, x^*) \quad \text{s.t.} \quad \sigma(x_k) \to \sigma(x)
\]
Extra assumptions:

\[ \inf_{\mathbb{R}^n \times \mathbb{R}^m} \Psi > -\infty, \quad \inf_{\mathbb{R}^n} f > -\infty \quad \text{and} \quad \inf_{\mathbb{R}^m} g > -\infty. \]

Quick Recall on Nonsmooth Analysis – [Rockafellar-Wets (98)] Let \( \sigma : \mathbb{R}^d \rightarrow (-\infty, +\infty] \) be a proper and lower semicontinuous function.

- **Limiting Subdifferential**:
  \( x^* \in \partial \sigma(x) \) means that

  \[ \sigma(u) \geq \sigma(x_k) + \langle x_k^*, u - x_k \rangle + o(\|u - x_k\|) \]

  with

  \( (x_k, x^*) \rightarrow (x, x^*) \) s.t. \( \sigma(x_k) \rightarrow \sigma(x) \)

- \( x \in \mathbb{R}^d \) is a **critical point** of \( \sigma \) if \( \partial \sigma(x) \ni 0 \).
Back to PALM: Convergence analysis

Set $z^k = (x^k, y^k)$ then:

(i) **Sufficient decrease property:** There exists $\rho_1 > 0$ such that

$$
\rho_1 \| z^{k+1} - z^k \|^2 \leq \Psi(z^k) - \Psi(z^{k+1}), \quad \forall k = 0, 1, \ldots.
$$

(ii) **A subgradient lower bound for the iterates gap:** There exists $\rho_2 > 0$, such that

$$
\| w^k \| \leq \rho_2 \| z^k - z^{k-1} \|, \quad w^k \in \partial \Psi(z^k), \quad \forall k = 0, 1, \ldots.
$$

These two steps are typical for *many many descent* type algorithms and lead to the fact that $\{ z^k \}_{k \in \mathbb{N}}$ bounded implies

\[
\text{limit points} = \text{critical points} \quad \& \quad \text{slow motion for large times: } \sum \| z^{k+1} - z^k \|^2 < +\infty
\]
Set $z^k = (x^k, y^k)$ then:

(i) **Sufficient decrease property:** There exists $\rho_1 > 0$ such that

$$\rho_1 \|z^{k+1} - z^k\|^2 \leq \psi(z^k) - \psi(z^{k+1}), \quad \forall k = 0, 1, \ldots.$$

(ii) **A subgradient lower bound for the iterates gap:** There exists $\rho_2 > 0$, such that

$$\|w^k\| \leq \rho_2 \|z^k - z^{k-1}\|, \quad w^k \in \partial \psi(z^k), \quad \forall k = 0, 1, \ldots.$$

- These two steps are typical for many many descent type algorithms and lead to the fact that $\{z^k\}_{k \in \mathbb{N}}$ bounded implies

  - **limit points = critical points**
  - **slow motion for large times:** $\sum \|z^{k+1} - z^k\|^2 < +\infty$

- What about actual convergence?...Or convergence rate??
A sequence $\zeta_k$ is called a gradient-like descent sequence for $\Phi : \mathbb{R}^n \to (-\infty, +\infty]$ if

(i) **Sufficient decrease property:** There exists $\rho_1$ such that

$$\rho_1 \|\zeta^{k+1} - \zeta^k\|^2 \leq \Phi(\zeta^k) - \Phi(\zeta^{k+1}), \quad \forall k = 0, 1, \ldots.$$ 

(ii) **A subgradient lower bound for the iterates gap:** Assume that $\{\zeta^k\}_{k \in \mathbb{N}}$ is bounded. There exists $\rho_2$ such that

$$\|w^k\| \leq \rho_2 \|\zeta^k - \zeta^{k-1}\|, \quad w^k \in \partial \Phi(\zeta^k), \quad \forall k = 0, 1, \ldots.$$
An abstract convergence theorem

A sequence $\zeta_k$ is called a gradient-like descent sequence for $\Phi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ if

(i) **Sufficient decrease property:** There exists $\rho_1$ such that

$$\rho_1\|\zeta^{k+1} - \zeta^k\|^2 \leq \Phi(\zeta^k) - \Phi(\zeta^{k+1}), \quad \forall k = 0, 1, \ldots$$

(ii) **A subgradient lower bound for the iterates gap:** Assume that $\{\zeta^k\}_{k \in \mathbb{N}}$ is bounded. There exists $\rho_2$ such that

$$\|w^k\| \leq \rho_2\|\zeta^k - \zeta^{k-1}\|, \quad w^k \in \partial \Phi(\zeta^k), \quad \forall k = 0, 1, \ldots$$

**Theorem (B-Sabach-Teboulle / Attouch-B-Svaiter)**

Let $\Phi$ be a semi-algebraic function and $\zeta_k$ a descent sequence for $\Phi$. If $\zeta^k$ is bounded then it converges to a critical point $\zeta^*$ of $f$. Besides

$$\|\zeta^k - \zeta^*\| \leq C \, k^{-\gamma}$$

with $\gamma > 0$. 

Jérôme Bolte (Toulouse Capitole)  Algorithm for Nonconvex and Nonsmooth Minimization
Convergence of PALM

Theorem (B.–Sabach–Teboulle, 13)

Under basic assumptions and assuming \( f, g, H \) real semi-algebraic.

Any bounded PALM sequence \( \{ z^k \}_{k \in \mathbb{N}} \) converges to a critical point \( z^* = (x^*, y^*) \) of \( \Psi \).

Moreover there exists \( \gamma > 0, C > 0 \) such that

\[
\| z^k - z^* \| \leq C k^{-\gamma}
\]

- Are there many semi-algebraic functions? **Ubiquitous in applications...**
- What is behind these results? **Any semi-algebraic function is a KL function**

........ but what is a KL function ?????
Definition (Sharpness)

A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is called sharp on the slice

$$[r_0 < f < r_1] := \left\{ x \in \mathbb{R}^d : r_0 < f(x) < r_1 \right\},$$

if there exists $c > 0$ such that

$$\|\partial f(x)\|_{\text{ }} \geq c$$

i.e. $\min \{\|\xi\| : \xi \in \partial f(x)\} \geq c.$

$\forall x \in [r_0 < f < r_1].$

Basic Examples: $f(x) = \|x\|$ or more generally $f(x) = \|Ax\|.$

Many works on sharpness starting around 1970:

- in Optimization: Polyak, Rockafellar, Burke, Kiwiel....
A function $f : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ is called sharp on the slice $[r_0 < f < r_1] := \{ x \in \mathbb{R}^d : r_0 < f(x) < r_1 \}$, if there exists $c > 0$ such that
\[
\|\partial f(x)\|_{\infty} \geq c
\]
i.e. \[\min \{\|\xi\| : \xi \in \partial f(x)\} \geq c.\]
\forall x \in [r_0 < f < r_1].

**Basic Examples:** $f(x) = \|x\|$ or more generally $f(x) = \|Ax\|$.

Many works on sharpness starting around 1970:
- in Optimization: Polyak, Rockafellar, Burke, Kiwiel.... $\Rightarrow$ Excellent convergence properties
Nonconvex illustration with a continuum of minimizers
Question: Can we measure “lack”/“Default” of sharpness on a slice $[0 < f < r_0]$?

A possible approach is to look at mappings of the form $\varphi \circ f$ where $\varphi : [0, r_0) \to \mathbb{R}_+$ is used to “make $f$ sharp”.

Quad. forms: $Q(x) = \frac{1}{2} \langle Ax, x \rangle$.

Choose $\varphi(s) = \sqrt{\lambda_{\text{max}}(A)} \sqrt{s}$ and $(\varphi \circ Q)$ is sharp.

Locally, around a critical level set of $f$, does there exist a reparameterization $\varphi \circ f$ which is sharp?

Notions: KL property / KL functions.

Jérôme Bolte (Toulouse Capitole)  Algorithm for Nonconvex and Nonsmooth Minimization
Question: Can we measure “lack”/“Default” of sharpness on a slice $[0 < f < r_0]$?

A possible approach is to look at mappings of the form $\varphi \circ f$ where $\varphi : [0, r_0) \rightarrow \mathbb{R}_+$ is used to “make $f$ sharp”.

Quad. forms: $Q(x) = \frac{1}{2} \langle Ax, x \rangle$, choose $\varphi(s) = \frac{1}{\sqrt{\lambda_{\text{max}}(A)}} \sqrt{s}$ and $(\varphi \circ Q)$ is sharp

Locally, around a critical level set of $f$, does there exist a reparameterization $\varphi \circ f$ which is sharp?

Notions: KL property / KL functions.
Are there Many Functions Satisfying KL?

Yes... and for a very broad class of functions!....

- Łojasiewicz, 1968 Real analytic functions $f : \Omega \to \mathbb{R}^n$ have this property around each points of their domain
- Kurdyka, 1998 Functions $C^1$ and definable in an o-minimal structure

**Theorem (Bolte-Daniilidis-Lewis (2006))**

*Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper lsc function. If $f$ is real semi-algebraic then it satisfies the KL property at any point of $\mathbb{R}^d$, i.e.*

$$\forall \bar{u} \in \text{dom } f$$

- $\exists \eta \in (0, +\infty]$,
- a neighborhood $U$ of $\bar{u}$
- a concave increasing function $\varphi \in C^1(0, \eta) \cap C^0[0, \eta)$, such that $\varphi(0) = 0$

such that for all

$$u \in U \cap [f(\bar{u}) < f(u) < f(\bar{u}) + \eta],$$

we have

$$\|\partial(\varphi \circ (f(\cdot) - f(\bar{u}))(u))\|_{-} \geq 1.$$ 

In other word any real semi-algebraic function is KL. $^a$

$^a$The result is actually valid for tame functions, i.e. functions definable in an o-minimal structure
To sum up

(i) **Sufficient decrease property:** There exists $\rho_1 > 0$ such that

$$\rho_1 \| z^{k+1} - z^k \|^2 \leq \Psi(z^k) - \Psi(z^{k+1}), \quad \forall k = 0, 1, \ldots.$$

(ii) **A subgradient lower bound for the iterates gap:** There exists $\rho_2 > 0$, such that

$$\| w^k \| \leq \rho_2 \| z^k - z^{k-1} \|, \quad w^k \in \partial \Psi(z^k), \quad \forall k = 0, 1, \ldots.$$

(iii) **is KL**

*Theorem (B.–Sabach–Teboulle, 13)*

**Under the above assumptions (+f, g, H bounded from below)**

Any bounded PALM sequence $\{z^k\}_{k \in \mathbb{N}}$ converges to a critical point $z^* = (x^*, y^*)$ of

$$\Psi = f + g + H.$$

Moreover there exists $\gamma > 0, C = C(z^0) > 0$ such that

$$\| z^k - z^* \| \leq C k^{-\gamma}$$
Consider the problem

\[
\begin{aligned}
A &\approx XY, \\
X &\text{ is sparse in } \mathbb{R}_+^{m \times r}, \\
Y &\text{ is sparse in } \mathbb{R}_+^{r \times n}
\end{aligned}
\]

The overall sparsity measure of a matrix defined by

\[
\|X\|_0 = \text{ number of nonzero entries in } X = \#\{(i,j) : X_{ij} \neq 0\}
\]

\[
\min \left\{ \frac{1}{2} \|A - XY\|_F^2 : X \geq 0, \|X\|_0 \leq s, \& \ Y \geq 0, \|Y\|_0 \leq t \right\}
\]
To apply PALM all we need is to compute the **prox of**

\[ f := i_{X \geq 0} + i_{\|X\|_0 \leq s}. \]

**Proposition (Proximal map formula for \( f \))**

Let \( U \in \mathbb{R}^{m \times n} \). Then

\[
\text{prox}_1^f (U) = \arg\min \left\{ \frac{1}{2} \|X - U\|_F^2 : X \geq 0, \|X\|_0 \leq s \right\} = T_s (P_+ (U))
\]

where \( T_s \) is defined by

\[
T_s (U) := \arg\min_{V \in \mathbb{R}^{m \times n}} \left\{ \|U - V\|_F^2 : \|U\|_0 \leq s \right\}.
\]

Computing \( T_s \) simply requires determining the \( s \)-th largest numbers of \( mn \) numbers. This can be done in \( O(mn) \) time, and zeroing out the proper entries in one more pass of the \( mn \) numbers.
PALM for Sparse NMF

1. Initialization: Select random nonnegative $X^0 \in \mathbb{R}^{m \times r}$ and $Y^0 \in \mathbb{R}^{r \times n}$.

2. For each $k = 0, 1, \ldots$ generate a sequence $\{(X^k, Y^k)\}_{k \in \mathbb{N}}$:

   2.1. Take $\gamma_1 > 1$, set $c_k = \gamma_1 \|Y^k (Y^k)^T\|_F$ and compute

   $$U^k = X^k - \frac{1}{c_k} (X^k Y^k - A) (Y^k)^T; \quad X^{k+1} \in \text{prox}_{R_1}^{c_k} (U^k) = T_\alpha (P_+ (U^k)).$$

   2.2. Take $\gamma_2 > 1$, set $d_k = \gamma_2 \|X^{k+1} (X^{k+1})^T\|_F$ and compute

   $$V^k = Y^k - \frac{1}{d_k} (X^{k+1})^T (X^{k+1} Y^k - A); \quad Y^{k+1} \in \text{prox}_{R_2}^{d_k} (V^k) = T_\beta (P_+ (V^k)).$$

Applying our main Theorem we thus get the desired global convergence result:

Let $\{(X^k, Y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by PALM-Sparse NMF. If

$$\inf_{k \in \mathbb{N}} \left\{ \|X^k\|_F, \|Y^k\|_F \right\} > 0.$$

Then, $\{(X^k, Y^k)\}_{k \in \mathbb{N}}$ converges to a critical point $(X^*, Y^*)$ of the Sparse NMF.

**That’s all folks!**