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, Sepetember 10th, Italia

The Optimization Model

Minimize a "block-nonsmooth" function

$$\min \left\{ \Psi(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + g(\mathbf{y}) + H(\mathbf{x},\mathbf{y}) \ : \ \mathbf{x} \in \mathbb{R}^m,\ \mathbf{y} \in \mathbb{R}^m \right\}$$

Assumption

(i) (i1) *f, g proper lsc real-extended-valued functions.*

(i2) *$H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ 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

$$\left\{ \begin{array}{l} A \approx XY, \\ X \in \mathcal{F} \subset \mathbb{R}^{m \times r} \\ Y \in \mathcal{G} \subset \mathbb{R}^{r \times n} \end{array} \right.$$

Using "merit functions" $\rightarrow$

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

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

$$L_1(Y) \equiv \left\| YY^T \right\|_F, \; L_2(X) \equiv \left\| X^T X \right\|_F$$

$$(M) \qquad \text{minimize}\left\{\Psi(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + g(\mathbf{y}) + H(\mathbf{x},\mathbf{y}) \;:\; \mathbf{x}\in\mathbb{R}^m,\ \mathbf{y}\in\mathbb{R}^m\right\}$$

Most optimization methods are about improving a given "state" $(\mathbf{x}^0, \mathbf{y}^0)$ 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: $\mathbf{x}, \mathbf{y}$ decomposition
- we decompose within the objective through the "smooth+nonsmooth" decomposition technique (aka forward-backward)

Space decomposition

$$(M) \qquad \text{minimize}\left\{\Psi\left(\mathbf{x},\mathbf{y}\right) := f\left(\mathbf{x}\right) + g\left(\mathbf{y}\right) + H\left(\mathbf{x},\mathbf{y}\right) \; : \; \mathbf{x} \in \mathbb{R}^{m}, \; \mathbf{y} \in \mathbb{R}^{m}\right\}$$

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

$$x^{k+1} \in \operatorname{argmin} \Psi\left(\mathbf{x}, y^{k}\right); \qquad y^{k+1} \in \operatorname{argmin} \Psi\left(x^{k+1}, \mathbf{y}\right).$$

“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

$$(M) \qquad \text{minimize}\left\{\Psi\left(\mathbf{x},\mathbf{y}\right) := f\left(\mathbf{x}\right) + g\left(\mathbf{y}\right) + H\left(\mathbf{x},\mathbf{y}\right) \ : \ \mathbf{x} \in \mathbb{R}^m, \ \mathbf{y} \in \mathbb{R}^m\right\}$$

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

$$x^{k+1} \in \text{argmin}\left\{\Psi\left(\mathbf{x}, y^k\right) + \lambda_k \|\mathbf{x} - x^k\|^2\right\}; \quad y^{k+1} \in \text{argmin}\{\Psi\left(x^{k+1}, \mathbf{y}\right) + \nu_k \|\mathbf{y} - y^k\|^2\}.$$

"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:

$$(P) \qquad \min \left\{ \sigma(\mathbf{u}) + h(\mathbf{u}) : \ \mathbf{u} \in \mathbb{R}^n \right\}, \quad h \in \text{ smooth, i.e. } C^{1,1}, \ \sigma \text{ nonsmooth.}$$

$$(P_k) \qquad u^{k+1} \in \operatorname{argmin} \left\{ \sigma(\mathbf{u}) + h(u^k) + \langle \mathbf{u} - u^k, \nabla h(u^k) \rangle + \frac{1}{2.\text{step}} \|\mathbf{u} - u^k\|^2 : \mathbf{u} \in \mathbb{R}^d \right\}.$$

- 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).......]

Proximal Map for Nonconvex Functions

Let $\sigma : \mathbb{R}^n \to (-\infty, +\infty]$ 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:

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

Proposition (Well-definedness of proximal maps)

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

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

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

the set-valued projection operator onto X.

A synthetic expression for the Proximal-Forward Backward Scheme

$$(P) \qquad \min\left\{h(\mathbf{u})+\sigma(\mathbf{u}):\ u\in\mathbb{R}^n\right\}, \qquad h\in \text{ smooth, i.e. } C^{1,1}$$

$$(P_k) \qquad u^{k+1}\in \operatorname{argmin}\{h(u^k)+\langle \mathbf{u}-u^k,\nabla h(u^k)\rangle+\frac{c_k}{2}\|\mathbf{u}-u^k\|^2+\sigma(\mathbf{u})\}.$$

Reformulation of PFB:

$$\boxed{u^{k+1}\in \operatorname{prox}^{\sigma}_{c_k}\left(x^k-\frac{1}{c_k}\nabla h\left(x^k\right)\right).}$$

- **ACTUAL DECOMPOSITION** for **"simple" nonsmooth** $\sigma(\cdot)$, i.e., **easy prox**.
- ∇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 !!

Decomposition....

In our case ∇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)

$\boxed{\textbf{PALM}}$

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

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

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

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

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

Main computational step: Computing the prox of a "simple" function.

PALM is Well Defined

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.

- **Fréchet Subdifferential** :
  $x^* \in \hat{\partial}\sigma(x)$ means that

$$\sigma(u) \geq \sigma(x \ \ ) + \langle x^* \ \ , u - x \ \ \rangle + o(\|u - x \ \ \|)$$

PALM is Well Defined

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

$$(x_k, x^*) \to (x, x^*) \text{ s.t. } \sigma(x_k) \to \sigma(x)$$

PALM is Well Defined

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

  $$(x_k, x^*) \to (x, x^*) \text{ s.t. } \sigma(x_k) \to \sigma(x)$$

- ⋆ $x \in \mathbb{R}^d$ is a **critical point** of σ 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

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

- 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***

$$\boxed{\textit{limit points = critical points}} \quad \&$$

$$\boxed{\textit{slow motion for large times: } \sum \|z^{k+1} - z^k\|^2 < +\infty}$$

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

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

- 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***

  $\boxed{\textit{limit points = critical points}}$ ***&***

  $\boxed{\textit{slow motion for large times: } \sum \|z^{k+1} - z^k\|^2 < +\infty}$

- **What about actual convergence ?...Or convergence rate??**

An abstract convergence theorem

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

(i) **Sufficient decrease property:** There exists ρ_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 ρ_2 such that

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

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

(i) **Sufficient decrease property:** There exists ρ_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 ρ_2 such that

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

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

Let Φ be a semi-algebraic function and ζ^k a descent sequence for Φ. If ζ^k is bounded then it converges to a critical point ζ^* of f. Besides

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

with $\gamma > 0$.

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 Ψ.

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 ????

Sharpness: A Geometric Snapshot at KL

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

$$\|\partial \mathbf{f}(\mathbf{x})\|_- \quad \geq \quad \mathbf{c}$$

$$\text{i.e.} \qquad \min\left\{\|\xi\| : \; \xi \in \partial f(x)\right\} \quad \geq \quad c.$$

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

Basic Examples: $f(x) = \|x\|$ or more generally $f(x) = \|Ax\|$.
Many works on sharpness starting around 1970 :

- in Optimization: Polyak, Rockafellar, Burke, Kiwiel....

Sharpness: A Geometric Snapshot at KL

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

$$\|\partial \mathbf{f}(\mathbf{x})\|_- \quad \geq \quad \mathbf{c}$$

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

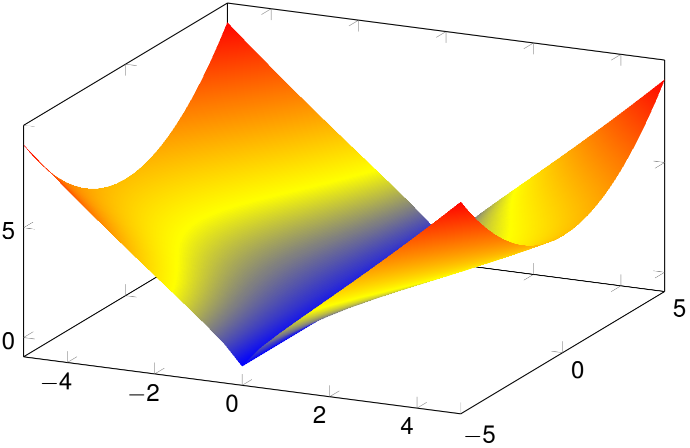
$\forall\, x \in [r_0 < f < r_1]$.

Basic Examples: $f(x) = \|x\|$ or more generally $f(x) = \|Ax\|$.
Many works on sharpness starting around 1970 :

- in Optimization: Polyak, Rockafellar, Burke, Kiwiel.... $\Longrightarrow$ **Excellent convergence properties**

Sharpness: example

Nonconvex illustration with a continuum of minimizers



Sharpness II

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".

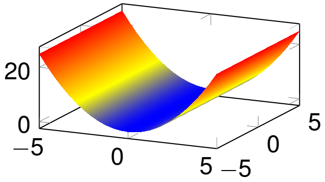
Sharpness II

Question: Can we measure "lack"/"Default" of sharpness on a slice $[0 < f < r_0]$?

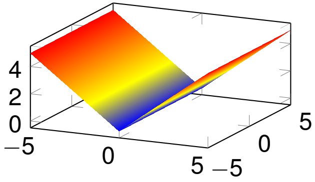
A possible approach is to look at mappings of the form $\varphi \circ f$ where $\varphi : [0, r_0) \to \mathbb{R}_+$ is used to "make f sharp".

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

Flat function Q



Sharp reparameterization $\varphi \circ Q$



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. $\forall \bar{u} \in \operatorname{dom} f$

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

such that for all

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

we have

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

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

"KL + Basic descent properties $\Rightarrow$ PALM Converges"

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

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

(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

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

An illustration: Sparse Nonnegative Matrix Factorization

Consider the problem

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

The overall sparsity measure of a matrix defined by

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

$$\min\left\{\frac{1}{2}\|A - XY\|_F^2 : X \geq 0,\ \|X\|_0 \leq s,\ \&\ Y \geq 0,\ \|Y\|_0 \leq t\ \right\}$$

Sparse Constraints in Nonnegative Matrix Factorization

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

$$\mathrm{prox}_1^f (U) = \mathrm{argmin} \left\{ \frac{1}{2} \|X - U\|_F^2 : X \geq 0, \|X\|_0 \leq s \right\} = T_s (P_+ (U))$$

where T_s is defined by

$$T_s (U) := \mathrm{argmin}_{V \in \mathbb{R}^{m \times n}} \left\{ \|U - V\|_F^2 : \ \|U\|_0 \leq s \right\}.$$

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 $\left\{\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$:

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

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

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

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

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

Let $\left\{\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$ be a sequence generated by PALM-Sparse NMF. If $\inf_{k\in\mathbb{N}}\left\{\left\|X^k\right\|_F, \left\|Y^k\right\|_F\right\} > 0$. Then, $\left\{\left(X^k, Y^k\right)\right\}_{k\in\mathbb{N}}$ converges to a critical point (X^*, Y^*) of the Sparse NMF.

Bolte, J., Sabach, S. and Teboulle, M., Proximal alternating linearized minimization for nonconvex and nonsmooth problems, *Mathematical Programming Series A*. Published online !

That's all folks !