A Preconditioned Forward-Backward Approach with Application to Large-Scale Nonconvex Spectral Unmixing Problems

Audrey Repetti, Emilie Chouzenoux, and Jean-Christophe Pesquet

Laboratoire d'Informatique Gaspard Monge - UMR CNRS 8049, Université Paris-Est, France.

ICASSP 2014 - Firenze - 7 May 2014

Motivation

INVERSE PROBLEM: Estimation of an object of interest $\overline{x} \in \mathbb{R}^N$ obtained by minimizing an objective function

$$\boxed{G = F + R}$$

where

- F is a data-fidelity term related to the observation model
- R is a regularization term related to some a priori assumptions on the target solution
  - $\rightsquigarrow$ e.g. an a priori on the smoothness of an image,
  - $\rightsquigarrow$ e.g. a support constraint.

Motivation

INVERSE PROBLEM: Estimation of an object of interest $\overline{x} \in \mathbb{R}^N$ obtained by minimizing an objective function

$$\boxed{G = F + R}$$

where

- F is a data-fidelity term related to the observation model
- R is a regularization term related to some a priori assumptions on the target solution

In the context of large scale problems, how to find an optimization algorithm able to deliver a reliable numerical solution in a reasonable time, with low memory requirement ?

$\Rightarrow$ **Block alternating minimization.**
$\Rightarrow$ **Introduction of a variable metric.**

Minimization problem

Problem

$$\text{Find} \quad \hat{x} \in \operatorname{Argmin}\{G = F + R\},$$

where:

- $F : \mathbb{R}^N \to \mathbb{R}$ is differentiable,

  and has an L-Lipschitz gradient on $\operatorname{dom} R$, i.e.

  $$\big(\forall (x,y) \in (\operatorname{dom} R)^2\big) \quad \|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|,$$

- $R : \mathbb{R}^N \to]-\infty, +\infty]$ is proper, lower semicontinuous.

- G is coercive, i.e. $\lim_{\|x\| \to +\infty} G(x) = +\infty$,

  and is non necessarily convex.

Forward-Backward algorithm

FB Algorithm

Let $x_0 \in \mathbb{R}^N$
For $\ell = 0, 1, \ldots$
$\lfloor\ x_{\ell+1} \in \operatorname{prox}_{\gamma_\ell R}\left(x_\ell - \gamma_\ell \nabla F(x_\ell)\right), \quad \gamma_\ell \in]0, +\infty[.$

- Let $x \in \mathbb{R}^N$. The **proximity operator** is defined by

$$\operatorname{prox}_{\gamma_\ell R}(x) = \underset{y \in \mathbb{R}^N}{\operatorname{Argmin}}\, R(y) + \frac{1}{2\gamma_\ell}\|y - x\|^2.$$

$\rightsquigarrow$ When R is nonconvex:

- Non necessarily uniquely defined.
- Existence guaranteed if R is bounded from below by an affine function.

Forward-Backward algorithm

FB Algorithm

Let $x_0 \in \mathbb{R}^N$
For $\ell = 0, 1, \ldots$
$\lfloor\ x_{\ell+1} \in \mathrm{prox}_{\gamma_\ell R}\left(x_\ell - \gamma_\ell \nabla F(x_\ell)\right), \quad \gamma_\ell \in]0, +\infty[.$

- Let $x \in \mathbb{R}^N$. The **proximity operator** is defined by

  $$\mathrm{prox}_{\gamma_\ell R}(x) = \underset{y \in \mathbb{R}^N}{\mathrm{Argmin}}\, R(y) + \frac{1}{2\gamma_\ell}\|y - x\|^2.$$

  $\rightsquigarrow$ When R is nonconvex:
  - Non necessarily uniquely defined.
  - Existence guaranteed if R is bounded from below by an affine function.

- Slow convergence.

Variable Metric Forward-Backward algorithm

VMFB Algorithm

Let $x_0 \in \mathbb{R}^N$

For $\ell = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} x_{\ell+1} \in \mathrm{prox}_{\gamma_\ell^{-1} A_\ell(x_\ell), R}\left(x_\ell - \gamma_\ell A_\ell(x_\ell)^{-1} \nabla F(x_\ell)\right), \\ \text{with } \gamma_\ell \in]0, +\infty[, \text{ and } A_\ell(x_\ell) \text{ a SPD matrix.} \end{array} \right.$$

- Let $x \in \mathbb{R}^N$. The proximity operator relative to the metric induced by $A_\ell(x_\ell)$ is defined by

$$\mathrm{prox}_{\gamma_\ell^{-1} A_\ell(x_\ell), R}(x) = \underset{y \in \mathbb{R}^N}{\mathrm{Argmin}}\, R(y) + \frac{1}{2\gamma_\ell} \|y - x\|^2_{A_\ell(x_\ell)}.$$

Variable Metric Forward-Backward algorithm

VMFB Algorithm

Let $x_0 \in \mathbb{R}^N$

For $\ell = 0, 1, \ldots$

$x_{\ell+1} \in \operatorname{prox}_{\gamma_\ell^{-1} A_\ell(x_\ell), R}\left(x_\ell - \gamma_\ell A_\ell(x_\ell)^{-1} \nabla F(x_\ell)\right),$

with $\gamma_\ell \in]0, +\infty[$, and $A_\ell(x_\ell)$ a SPD matrix.

- Let $x \in \mathbb{R}^N$. The proximity operator relative to the metric induced by $A_\ell(x_\ell)$ is defined by

$$\operatorname{prox}_{\gamma_\ell^{-1} A_\ell(x_\ell), R}(x) = \underset{y \in \mathbb{R}^N}{\operatorname{Argmin}}\, R(y) + \frac{1}{2\gamma_\ell}\|y - x\|^2_{A_\ell(x_\ell)}.$$

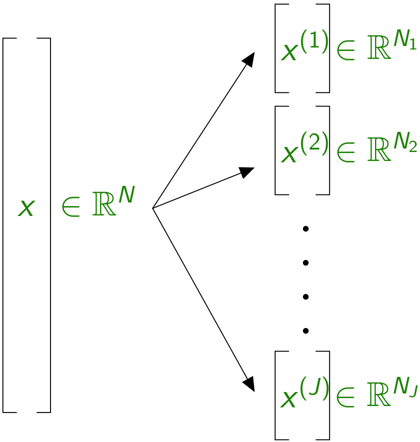
- Convergence is established for a wide class of nonconvex functions G and $(A_\ell(x_\ell))_{\ell \in \mathbb{N}}$ are general SPD matrices in [Chouzenoux *et al.* - 2013]

Block separable structure

- R is an additively block separable function.

Block separable structure

- R is an additively block separable function.



$$N = \sum_{j=1}^{J} N_j$$

Block separable structure

- R is an additively block separable function.

$$R\left(\begin{bmatrix} x \end{bmatrix}\right) = R\left(\begin{matrix} \begin{bmatrix} x^{(1)} \end{bmatrix} \\ \begin{bmatrix} x^{(2)} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} x^{(J)} \end{bmatrix} \end{matrix}\right) = \sum_{j=1}^{J} R_j(x^{(j)})$$

$(\forall j \in \{1, \dots, J\})$ $R_j \colon \mathbb{R}^{N_j} \to]-\infty, +\infty]$ is a lsc, proper function, continuous on its domain and bounded from below by an affine function.

BC Forward-Backward algorithm

BC-FB Algorithm [Bolte *et al.* - 2013]

Let $x_0 \in \mathbb{R}^N$

For $\ell = 0, 1, \ldots$

$$
\left\lfloor
\begin{array}{l}
\text{Let } j_\ell \in \{1, \ldots, J\}, \\
x_{\ell+1}^{(j_\ell)} \in \operatorname{prox}_{\gamma_\ell R_{j_\ell}} \left(x_\ell^{(j_\ell)} - \gamma_\ell \nabla_{j_\ell} F(x_\ell) \right), \quad \gamma_\ell \in]0, +\infty[, \\
x_{\ell+1}^{(\bar{\jmath}_\ell)} = x_\ell^{(\bar{\jmath}_\ell)}.
\end{array}
\right.
$$

- Advantages of a block coordinate strategy:
  - more flexibility,
  - reduce computational cost at each iteration,
  - reduce memory requirement.

BC Variable Metric Forward-Backward algorithm

BC-VMFB Algorithm

Let $x_0 \in \mathbb{R}^N$

For $\ell = 0, 1, \ldots$

$$\left\lfloor \begin{array}{l} \text{Let } j_\ell \in \{1, \ldots, J\}, \\ x_{\ell+1}^{(j_\ell)} \in \mathrm{prox}_{\gamma_\ell^{-1} \boxed{A_{j_\ell}(x_\ell)},\, R_{j_\ell}} \left(x_\ell^{(j_\ell)} - \gamma_\ell \boxed{A_{j_\ell}(x_\ell)}^{-1} \nabla_{j_\ell} F(x_\ell) \right), \\ x_{\ell+1}^{(\bar{\jmath}_\ell)} = x_\ell^{(\bar{\jmath}_\ell)}, \\ \text{with } \gamma_\ell \in]0, +\infty[, \text{ and } \boxed{A_{j_\ell}(x_\ell)} \text{ a SPD matrix.} \end{array} \right.$$

OUR CONTRIBUTIONS:

- How to choose the preconditioning matrices $(A_{j_\ell}(x_\ell))_{\ell \in \mathbb{N}}$?
  $\rightsquigarrow$ Majorize-Minimize principle.
- How to define a general update rule for $(j_\ell)_{\ell \in \mathbb{N}}$?
  $\rightsquigarrow$ Quasi-cyclic rule.

Majorize-Minimize assumption [Jacobson *et al.* - 2007]

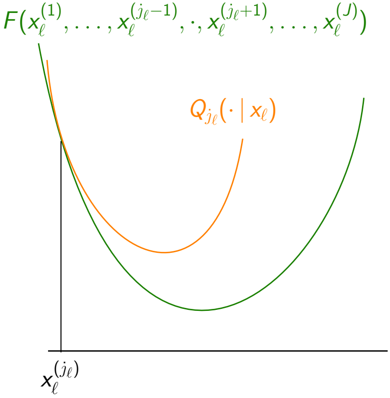
MM Assumption

$(\forall \ell \in \mathbb{N})$ there exists a lower and upper bounded SPD matrix $A_{j_\ell}(x_\ell) \in \mathbb{R}^{N_{j_\ell} \times N_{j_\ell}}$ such that $(\forall y \in \mathbb{R}^{N_{j_\ell}})$

$$Q_{j_\ell}(y \mid x_\ell) = F(x_\ell) + (y - x_\ell^{(j_\ell)})^\top \nabla_{j_\ell} F(x_\ell) + \frac{1}{2}\|y - x_\ell^{(j_\ell)}\|^2_{A_{j_\ell}(x_\ell)},$$

is a *majorant function* on dom R_{j_ℓ} of the restriction of F to its j_ℓ-th block at $x_\ell^{(j_\ell)}$, i.e., $(\forall y \in \operatorname{dom} R_{j_\ell})$

$$F\left(x_\ell^{(1)}, \ldots, x_\ell^{(j_\ell-1)}, y, x_\ell^{(j_\ell+1)}, \ldots, x_\ell^{(J)}\right) \leq Q_{j_\ell}(y \mid x_\ell).$$



Majorize-Minimize assumption [Jacobson *et al.* - 2007]

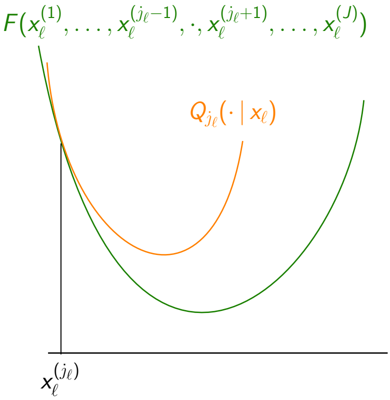
MM Assumption

$(\forall \ell \in \mathbb{N})$ there exists a lower and upper bounded SPD matrix $A_{j_\ell}(x_\ell) \in \mathbb{R}^{N_{j_\ell} \times N_{j_\ell}}$ such that $(\forall y \in \mathbb{R}^{N_{j_\ell}})$

$$Q_{j_\ell}(y \mid x_\ell) = F(x_\ell) + (y - x_\ell^{(j_\ell)})^\top \nabla_{j_\ell} F(x_\ell) + \tfrac{1}{2} \| y - x_\ell^{(j_\ell)} \|^2_{A_{j_\ell}(x_\ell)},$$

is a *majorant function* on $\operatorname{dom} R_{j_\ell}$ of the restriction of F to its j_ℓ-th block at $x_\ell^{(j_\ell)}$, i.e., $(\forall y \in \operatorname{dom} R_{j_\ell})$

$$F\left(x_\ell^{(1)}, \ldots, x_\ell^{(j_\ell - 1)}, y, x_\ell^{(j_\ell+1)}, \ldots, x_\ell^{(J)}\right) \leq Q_{j_\ell}(y \mid x_\ell).$$



dom R is convex and F is L-Lipschitz differentiable $\Rightarrow$ The above assumption holds if $(\forall \ell \in \mathbb{N})$ $A_{j_\ell}(x_\ell) \equiv L\, \mathrm{I}_{N_{j_\ell}}$

Convergence results

Additional assumptions

- G satisfies the Kurdyka-Łojasiewicz inequality [Attouch *et al.* - 2011]:

  For every $\xi \in \mathbb{R}$, for every bounded $E \subset \mathbb{R}^N$, there exist $\kappa, \zeta > 0$ and $\theta \in [0,1)$ such that, for every $x \in E$ such that $|G(x) - \xi| \leq \zeta$,

  $$\big(\forall r \in \partial R(x)\big) \qquad \|\nabla F(x) + r\| \geq \kappa |G(x) - \xi|^{\theta}.$$

  Technical assumption satisfied for a wide class of nonconvex functions

  - semi-algebraic functions
  - real analytic functions
  - ...

Convergence results

Additional assumptions

- G satisfies the Kurdyka-Łojasiewicz inequality [Attouch *et al.* - 2011]:

  For every $\xi \in \mathbb{R}$, for every bounded $E \subset \mathbb{R}^N$, there exist $\kappa, \zeta > 0$ and $\theta \in [0,1)$ such that, for every $x \in E$ such that $|G(x) - \xi| \leq \zeta$,

  $$\big(\forall r \in \partial R(x)\big) \qquad \|\nabla F(x) + r\| \geq \kappa |G(x) - \xi|^{\theta}.$$

  Technical assumption satisfied for a wide class of nonconvex functions

  - semi-algebraic functions
  - real analytic functions
  - ...

  $\rightsquigarrow$ Almost every function you can imagine!

Convergence results

Additional assumptions

- G satisfies the Kurdyka-Łojasiewicz inequality [Attouch *et al.* - 2011]:

  For every $\xi \in \mathbb{R}$, for every bounded $E \subset \mathbb{R}^N$, there exist $\kappa, \zeta > 0$ and $\theta \in [0,1)$ such that, for every $x \in E$ such that $|G(x) - \xi| \leq \zeta$,

  $$\big(\forall r \in \partial R(x)\big) \qquad \|\nabla F(x) + r\| \geq \kappa |G(x) - \xi|^{\theta}.$$

  Technical assumption satisfied for a wide class of nonconvex functions

- Blocks $(j_\ell)_{\ell \in \mathbb{N}}$ updated according to a quasi-cyclic rule, i.e., there exists $K \geq J$ such that, for every $\ell \in \mathbb{N}$, $\{1, \ldots, J\} \subset \{j_\ell, \ldots, j_{\ell+K-1}\}$.

Convergence results

Additional assumptions

- G satisfies the Kurdyka-Łojasiewicz inequality [Attouch *et al.* - 2011]:

  For every $\xi \in \mathbb{R}$, for every bounded $E \subset \mathbb{R}^N$, there exist $\kappa, \zeta > 0$ and $\theta \in [0,1)$ such that, for every $x \in E$ such that $|G(x) - \xi| \leq \zeta$,

  $$\big(\forall r \in \partial R(x)\big) \qquad \|\nabla F(x) + r\| \geq \kappa |G(x) - \xi|^{\theta}.$$

  Technical assumption satisfied for a wide class of nonconvex functions

- Blocks $(j_\ell)_{\ell \in \mathbb{N}}$ updated according to a quasi-cyclic rule, i.e., there exists $K \geq J$ such that, for every $\ell \in \mathbb{N}$, $\{1, \ldots, J\} \subset \{j_\ell, \ldots, j_{\ell+K-1}\}$.

- The step-size is chosen such that:
  - $\exists(\underline{\gamma}, \overline{\gamma}) \in (0, +\infty)^2$ such that $(\forall \ell \in \mathbb{N})$ $\underline{\gamma} \leq \gamma_\ell \leq 1 - \overline{\gamma}$.
  - For every $j \in \{1, \ldots, J\}$, R_j is a convex function and $\exists(\underline{\gamma}, \overline{\gamma}) \in (0, +\infty)^2$ such that $(\forall \ell \in \mathbb{N})$ $\underline{\gamma} \leq \gamma_\ell \leq 2 - \overline{\gamma}$.

Convergence results

Convergence theorem

Let $(x_\ell)_{\ell\in\mathbb{N}}$ be a sequence generated by the BC-VMFB algorithm.

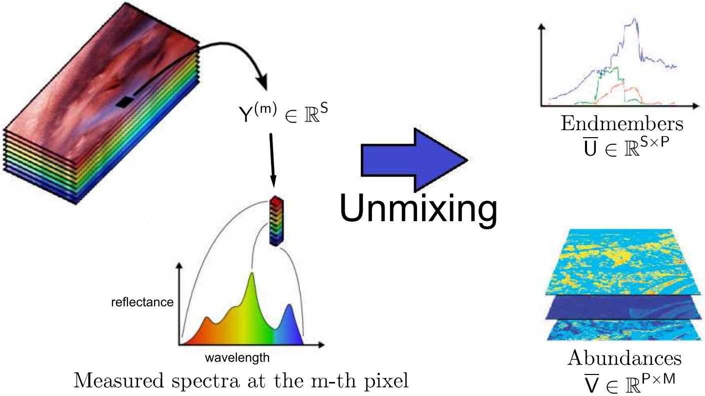
- Global convergence:
  $\rightsquigarrow$ $(x_\ell)_{\ell\in\mathbb{N}}$ converges to a critical point $\widehat{x}$ of G.

  $\rightsquigarrow$ $(G(x_\ell))_{\ell\in\mathbb{N}}$ is a nonincreasing sequence converging to $G(\widehat{x})$.

- Local convergence:
  If $(\exists \upsilon > 0)$ such that $G(x_0) \leq \inf_{x\in\mathbb{R}^N} G(x) + \upsilon$,
  then $(x_\ell)_{\ell\in\mathbb{N}}$ converges to a solution $\widehat{x}$ to the minimization problem.

Spectral unmixing problem

$Y = \left[Y^{(1)}, \dots, Y^{(M)}\right] \in \mathbb{R}^{S \times M}$



$$Y = \overline{U}\,\overline{V} + E$$

Proposed criterion

OBSERVATION MODEL: $Y = \overline{U}\,\overline{V} + E \quad \rightsquigarrow \quad Y = \Omega\overline{T}\,\overline{V} + E$,

with
- $\Omega \in \mathbb{R}^{S\times Q}$ a known spectra library of size $Q \gg P$,
- $\overline{T} \in \mathbb{R}^{Q\times P}$ an unknown matrix assumed to be sparse.

OBJECTIVE: Find estimates of $\overline{T}$ and $\overline{V}$.

Proposed criterion

OBSERVATION MODEL: $Y = \Omega \overline{T}\,\overline{V} + E,$

$$\boxed{\underset{T\in\mathbb{R}^{Q\times P}, V\in\mathbb{R}^{P\times M}}{\text{minimize}} \quad \left(G(T,V) = F(T,V) + R_1(T) + R_2(V)\right),}$$

- $F(T,V) = \frac{1}{2}\|Y - \Omega T V\|_F^2,$

- $$R_1(T) = \sum_{q=1}^{Q}\sum_{p=1}^{P}\left(\iota_{[T_{\min},T_{\max}]}(T^{(q,p)}) + \eta\varphi_\beta(T^{(q,p)})\right),$$
  with φ_β a nonconvex penalization promoting the sparsity, defined in [Chartrand, 2012] for $\beta \in]0,1]$, and $(\eta, T_{\min}, T_{\max}) \in]0,+\infty[^3$.

- $R_2(V) = \iota_{\mathcal{V}}(V),$
  with $\mathcal{V} = \{V \in \mathbb{R}^{P\times M} \mid (\forall m \in \{1,\dots,M\}) \sum_{p=1}^{P} V^{(p,m)} = 1,$
  $(\forall p \in \{1,\dots,P\})(\forall m \in \{1,\dots,M\})\ V^{(p,m)} \geq V_{\min}\},$
  where $V_{\min} > 0$.

Construction of the preconditioning matrices

Let $(T', V') \in \operatorname{dom} R_1 \times \operatorname{dom} R_2$.

$T \mapsto F(T, V') = \frac{1}{2}\|Y - \Omega T V\|_F^2$ is majorized on dom R_1 by

$$Q_1(T \mid T', V') = F(T', V') + \operatorname{tr}\left((T - T')\nabla_1 F(T', V')^\top\right) + \frac{1}{2}\operatorname{tr}\left(((T - T') \odot A_1(T', V'))\,(T - T')^\top\right),$$

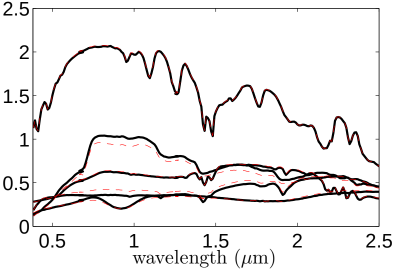
where $A_1(T', V') = ((\Omega^\top \Omega) T' (V' V'^\top)) \oslash T'$.

$V \mapsto F(T', V) = \frac{1}{2}\|Y - \Omega T V\|_F^2$ is majorized on dom R_2 by

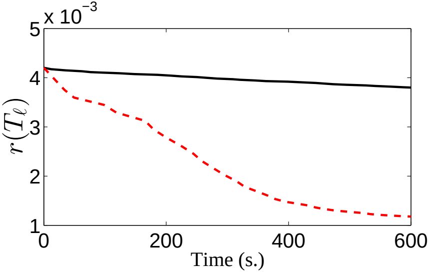
$$Q_2(V \mid T', V') = F(T', V') + \operatorname{tr}\left((V - V')\nabla_2 F(T', V')^\top\right) + \frac{1}{2}\operatorname{tr}\left(((V - V') \odot A_2(T', V'))\,(V - V')^\top\right),$$

where $A_2(T', V') = ((\Omega T')^\top \Omega T' V') \oslash V'$.

Numerical results



- Continuous lines:
  Exact endmembers $\overline{U}$,
- Dashed lines:
  Estimated endmembers $\widehat{U}$.



- Dashed line:
  BC-VMFB algorithm
  [Chouzenoux *et al.* - 2013],
- Continuous line:
  PALM algorithm
  [Bolte *et al.* - 2013].

Conclusion

- Proposition of a new BC-VMFB algorithm for minimizing the sum of
  - a nonconvex smooth function F,
  - a nonconvex non necessarily smooth function R.
- Convergence results both on the iterates and the function values.
- Blocks updated according to a flexible quasi-cyclic rule.
- Acceleration of the convergence thanks to the choice of matrices $(A_{j_\ell}(x_\ell))_{\ell \in \mathbb{N}}$ based on MM principle.

Combining variable metric strategy with a block alternating scheme leads to a significant acceleration in terms of decay of the error on the iterates.

Thank you ! Questions ?

E. Chouzenoux, J.-C. Pesquet and A. Repetti.
Variable Metric Forward-Backward Algortihm for Minimizing the Sum of a Differentiable Function and a Convex Function.
To appear in J. Optim. Theory Appl, 2013.

E. Chouzenoux, J.-C. Pesquet and A. Repetti.
A Block Coordinate Variable Metric Forward-Backward algorithm.
Tech. Rep., 2013. Available on
http://www.optimization-online.org/DB_HTML/2013/12/4178.html.