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 \( \bar{x} \in \mathbb{R}^N \) obtained by minimizing an objective function

\[
G = F + R
\]

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
  - e.g. an a priori on the smoothness of an image,
  - e.g. a support constraint.
Motivation

**Inverse problem**: Estimation of an object of interest $\mathbf{x} \in \mathbb{R}^N$ obtained by minimizing an objective function $G = F + R$

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

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

where:

- $F: \mathbb{R}^N \rightarrow \mathbb{R}$ is differentiable, and has an $L$-Lipschitz gradient on $\text{dom} \ R$, i.e.
  \[(\forall (x, y) \in (\text{dom} \ R)^2) \quad \|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|,\]

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

- $G$ is coercive, i.e. \(\lim_{\|x\| \rightarrow +\infty} G(x) = +\infty\), and is non necessarily convex.
Forward-Backward algorithm

**FB Algorithm**

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

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

\[ x_{\ell+1} \in \text{prox}_{\gamma_\ell R} (x_\ell - \gamma_\ell \nabla F(x_\ell)), \quad \gamma_\ell \in ]0, +\infty[. \]

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

\[
\text{prox}_{\gamma_\ell R}(x) = \text{Argmin}_{y \in \mathbb{R}^N} R(y) + \frac{1}{2\gamma_\ell} \| y - x \|^2.
\]

~~ 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 \( \ell = 0, 1, \ldots \)

\[
\begin{align*}
    x_{\ell+1} &\in \text{prox}_{\gamma\ell R} (x_\ell - \gamma\ell \nabla F(x_\ell)), & \gamma\ell &\in ]0, +\infty[. \\
\end{align*}
\]

\[\Rightarrow\] Let \( x \in \mathbb{R}^N \). The **proximity operator** is defined by

\[
\text{prox}_{\gamma\ell R}(x) = \text{Argmin}_{y \in \mathbb{R}^N} R(y) + \frac{1}{2\gamma\ell} \|y - x\|^2.
\]

\[\sim\] When \( R \) is nonconvex:

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

\[\Rightarrow\] Slow convergence.
Variable Metric Forward-Backward algorithm

**VMFB Algorithm**

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

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

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

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

$$\text{prox}_{\frac{1}{\gamma_{\ell}} A_{\ell}(x_{\ell}), R} (x) = \text{Argmin}_{y \in \mathbb{R}^N} R(y) + \frac{1}{2\gamma_{\ell}} \| y - x \|^2_{A_{\ell}(x_{\ell})}.$$
Variable Metric Forward-Backward algorithm

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

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

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

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

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

- 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.

$$
\begin{align*}
\mathbf{x} \in \mathbb{R}^N \\
\mathbf{x}^{(1)} \in \mathbb{R}^{N_1} \\
\mathbf{x}^{(2)} \in \mathbb{R}^{N_2} \\
\vdots \\
\mathbf{x}^{(J)} \in \mathbb{R}^{N_J}
\end{align*}
$$

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

- \( R \) is an additively block separable function.

\[
R(x) = \sum_{j=1}^{J} R_j(x^{(j)})
\]

\((\forall j \in \{1, \ldots, J\})\) \( R_j: \mathbb{R}^{N_j} \rightarrow ]-\infty, +\infty[\) 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 $\ell = 0, 1, \ldots$

1. Let $j_\ell \in \{1, \ldots, J\}$,
2. $x_{\ell+1}^{(j_\ell)} \in \text{prox}_{\gamma_\ell R_{j_\ell}} \left( x_\ell^{(j_\ell)} - \gamma_\ell \nabla_{j_\ell} F(x_\ell) \right)$, $\gamma_\ell \in ]0, +\infty[$,
3. $x_{\ell+1}^{(\overline{j}_\ell)} = x_\ell^{(\overline{j}_\ell)}$.

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 \( \ell = 0, 1, \ldots \)

Let \( j_\ell \in \{1, \ldots, J\} \),

\[
\begin{align*}
    & x_{\ell + 1}^{(j_\ell)} \in \operatorname{prox}_{\gamma_{\ell}^{-1}\left( x_\ell(x_{\ell}^{(j_\ell)}) - \gamma_{\ell} A_{j_\ell}(x_{\ell})^{-1} \nabla_{j_\ell} F(x_{\ell}) \right)}, \\
    & x_{\ell + 1}^{(\overline{j_\ell})} = x_{\ell}^{(\overline{j_\ell})}, \\
    & x_{\ell + 1} = x_{\ell}^{(j_\ell)},
\end{align*}
\]

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

**Our contributions:**

- How to choose the preconditioning matrices \((A_{j_\ell}(x_{\ell}))_{\ell \in \mathbb{N}}\)?
  \(\leadsto\) Majorize-Minimize principle.

- How to define a general update rule for \((j_\ell)_{\ell \in \mathbb{N}}\)?
  \(\leadsto\) Quasi-cyclic rule.
Majorize–Minimize assumption

[Jacobson et al. - 2007]

\( \forall \ell \in \mathbb{N} \) 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 \( \text{dom } R_{j_{\ell}} \) of the restriction of \( F \) to its \( j_{\ell} \)-th block at \( x_{\ell}^{(j\ell)} \), i.e., \( \forall y \in \text{dom } R_{j_{\ell}} \)

\[ F(x_{\ell}^{(1)}, \ldots, x_{\ell}^{(j\ell-1)}, y, x_{\ell}^{(j\ell+1)}, \ldots, x_{\ell}^{(J)}) \leq Q_{j\ell}(y \mid x_{\ell}). \]
Majorize-Minimize assumption

**MM Assumption**

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

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

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

\[(\forall y \in \text{dom } R_{\ell\ell})\]

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

\[\text{dom } R \text{ is convex and } F \text{ is } L\text{-Lipschitz differentiable} \quad \Rightarrow \quad \text{The above assumption holds if } (\forall \ell \in \mathbb{N}) A_{\ell\ell}(x_{\ell}) \equiv L \mathbf{1}_{N_{\ell\ell}}\]
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 $|G(x) - \xi| \leq \zeta$,

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

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 $|G(x) - \xi| \leq \zeta$,

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

  Technical assumption satisfied for a wide class of nonconvex functions
  - semi-algebraic functions
  - real analytic functions
  - ...

  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 $|G(x) - \xi| \leq \zeta$,
  \[ (\forall r \in \partial R(x)) \quad \|\nabla F(x) + r\| \geq \kappa |G(x) - \xi|^{\theta}. \]
  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 $|G(x) - \xi| \leq \zeta$,
  
  \[
  (\forall r \in \partial R(x)) \quad \|\nabla F(x) + r\| \geq \kappa |G(x) - \xi|^\theta.
  \]

  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:**
  \(\Rightarrow (x_\ell)_{\ell \in \mathbb{N}}\) converges to a critical point \(\hat{x}\) of \(G\).
  \(\Rightarrow (G(x_\ell))_{\ell \in \mathbb{N}}\) is a nonincreasing sequence converging to \(G(\hat{x})\).

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

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

\[ Y^{(m)} \in \mathbb{R}^{S} \]  

Unmixing

Measured spectra at the m-th pixel

\[ Y = \overline{U} \overline{V} + E \]
Proposed criterion

**Observation model:** \( Y = \bar{U} \bar{V} + E \leadsto Y = \bar{\Omega} \bar{T} \bar{V} + E, \)

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

**Objective:** Find estimates of \( \bar{T} \) and \( \bar{V}. \)
Proposed criterion

**Observation model:** $Y = \Omega \bar{T} \bar{V} + E$, 

\[
\min_{T \in \mathbb{R}^{Q \times P}, V \in \mathbb{R}^{P \times M}} \quad G(T, V) = F(T, V) + R_1(T) + R_2(V),
\]

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

- $R_1(T) = \sum_{q=1}^{Q} \sum_{p=1}^{P} \left( \iota_{[T_{\min}, T_{\max}]}(T(q,p)) + \eta \varphi_{\beta}(T(q,p)) \right),$ 
  with $\varphi_{\beta}$ 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 $\mathcal{V} = \{ V \in \mathbb{R}^{P \times M} \mid (\forall m \in \{1, \ldots, M\}) \sum_{p=1}^{P} V^{(p,m)} = 1,$
  \[
  (\forall p \in \{1, \ldots, P\})(\forall m \in \{1, \ldots, M\}) V^{(p,m)} \geq V_{\min},
  \]
  where $V_{\min} > 0.$
Construction of the preconditioning matrices

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

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

\[ Q_1(T | T', V') = F(T', V') + \text{tr} \left( (T - T') \nabla_1 F(T', V')^\top \right) \]
\[ + \frac{1}{2} \text{tr} \left( ((T - T') \odot A_1(T', V')) (T - T')^\top \right), \]

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

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

\[ Q_2(V | T', V') = F(T', V') + \text{tr} \left( (V - V') \nabla_2 F(T', V')^\top \right) \]
\[ + \frac{1}{2} \text{tr} \left( ((V - V') \odot A_2(T', V')) (V - V')^\top \right), \]

where \( A_2(T', V') = ((\Omega T')^\top \Omega T' V') \odot V'. \)
Numerical results

- Continuous lines: Exact endmembers $\bar{T}$,
- Dashed lines: Estimated endmembers $\hat{T}$.

- 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 Algorithm for Minimizing the Sum of a Differentiable Function and a Convex Function.*

E. Chouzenoux, J.-C. Pesquet and A. Repetti.
*A Block Coordinate Variable Metric Forward-Backward algorithm.*