Majorization-Minimization Algorithms for Large Scale Data Processing

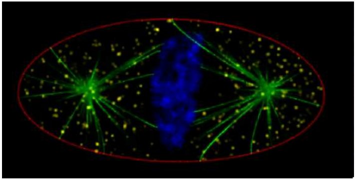
Emilie CHOUZENOUX

Center for Visual Computing, CentraleSupelec, INRIA Saclay

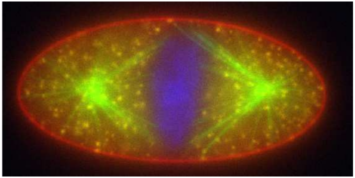
IFPEN

6 December 2017

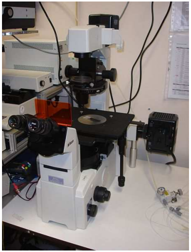
Inverse problems and large scale optimization



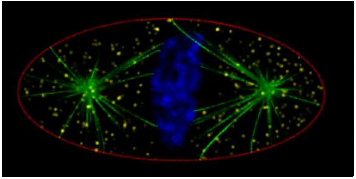
Original image



Degraded image

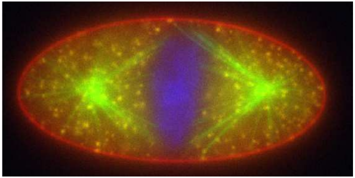


Inverse problems and large scale optimization



Original image

$\overline{\boldsymbol{x}} \in \mathbb{R}^N$



Degraded image

$\boldsymbol{y} = \mathcal{D}(\boldsymbol{H}\overline{\boldsymbol{x}}) \in \mathbb{R}^M$

- $\boldsymbol{H} \in \mathbb{R}^{M \times N}$: matrix associated with the degradation operator.
- $\mathcal{D} \colon \mathbb{R}^M \to \mathbb{R}^M$: noise degradation.

How to find a good estimate of $\overline{\boldsymbol{x}}$ from the observations $\boldsymbol{y}$ and the model $\boldsymbol{H}$ in the context of large scale processing?

Inverse problems and large scale optimization

Variational approach:

An image estimate $\hat{\boldsymbol{x}} \in \mathbb{R}^N$ is generated by minimizing (iteratively)

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = f(\boldsymbol{H}\boldsymbol{x}) + \Psi(\boldsymbol{x})$$

with $f : \mathbb{R}^M \to \mathbb{R}$, $\Psi : \mathbb{R}^N \to \mathbb{R}$.

Inverse problems and large scale optimization

Variational approach:

An image estimate $\hat{x} \in \mathbb{R}^N$ is generated by minimizing (iteratively)

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = f(\boldsymbol{H}\boldsymbol{x}) + \Psi(\boldsymbol{x})$$

with $f : \mathbb{R}^M \to \mathbb{R}$, $\Psi : \mathbb{R}^N \to \mathbb{R}$.

$\Rightarrow$ In the context of maximum a posteriori estimation :

$*$ $f \circ \boldsymbol{H}$: Data fidelity term related to the acquisition model;

Example: Least squares function

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad f(\boldsymbol{H}\boldsymbol{x}) = \|\boldsymbol{H}\boldsymbol{x} - \boldsymbol{y}\|^2$$

Inverse problems and large scale optimization

Variational approach:

An image estimate $\hat{\boldsymbol{x}} \in \mathbb{R}^N$ is generated by minimizing (iteratively)

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = f(\boldsymbol{H}\boldsymbol{x}) + \Psi(\boldsymbol{x})$$

with $f : \mathbb{R}^M \to \mathbb{R}$, $\Psi : \mathbb{R}^N \to \mathbb{R}$.

$\Rightarrow$ In the context of maximum a posteriori estimation :

$*$ $f \circ \boldsymbol{H}$: Data fidelity term related to the acquisition model;

$*$ Ψ : Regularization function.

Example: Sparsity prior (analysis)

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad \Psi(\boldsymbol{x}) = \|\boldsymbol{F}\boldsymbol{x}\|_1$$

with $\boldsymbol{F} \in \mathbb{R}^{P \times N}$, $P \geq N$, a frame decomposition operator.

Inverse problems and large scale optimization

Variational approach:

An image estimate $\hat{\boldsymbol{x}} \in \mathbb{R}^N$ is generated by minimizing (iteratively)

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F(\boldsymbol{x}) = f(\boldsymbol{H}\boldsymbol{x}) + \Psi(\boldsymbol{x})$$

with $f : \mathbb{R}^M \to \mathbb{R}$, $\Psi : \mathbb{R}^N \to \mathbb{R}$.

$\Rightarrow$ In the context of maximum a posteriori estimation :

$*$ $f \circ \boldsymbol{H}$: Data fidelity term related to the acquisition model;

$*$ Ψ : Regularization function.

▶ **Choosing an efficient iterative minimization strategy depends on the properties of (f, Ψ).**

A unified framework: Majorize-Minimize principle

PROBLEM: Find $\hat{\boldsymbol{x}} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathbb{R}^N} F(\boldsymbol{x})$

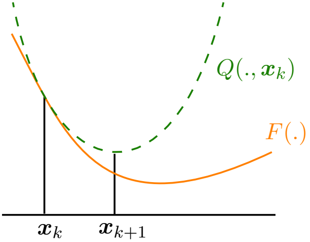
For all $\boldsymbol{x}' \in \mathbb{R}^N$, let $Q(.,\boldsymbol{x}')$ a ***tangent majorant*** of F at $\boldsymbol{x}'$ i.e.,

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad Q(\boldsymbol{x},\boldsymbol{x}') \geqslant F(\boldsymbol{x}) \quad \text{and} \quad Q(\boldsymbol{x}',\boldsymbol{x}') = F(\boldsymbol{x}')$$

MM algorithm:

$$(\forall k \in \mathbb{N})$$

$$\boldsymbol{x}_{k+1} \in \operatorname{Argmin}_{\boldsymbol{x}\in\mathbb{R}^N} Q(\boldsymbol{x},\boldsymbol{x}_k)$$



⋆ Quadratic majorants $\rightsquigarrow$ **tractable inner minimization step**

Outline

* MAJORIZE-MINIMIZE MEMORY GRADIENT ALGORITHM
  - Majorize-Minimize principle
  - Subspace acceleration
  - Convergence properties
  - Block parallel 3MG algorithm
  - Stochastic 3MG algorithm
* VARIABLE METRIC FORWARD-BACKWARD ALGORITHM
  - Majorize-Minimize preconditioning
  - Block alternating extension
  - Application to phase retrieval

Majorize-Minimize Memory Gradient algorithm

Majorize-Minimize subspace algorithm [Chouzenoux *et al.*, 2013]

⋆ Minimize differentiable and nonconvex function F on $\mathbb{R}^N$.

At each iteration $k \in \mathbb{N}$:

❶ Build a quadratic majorant function $Q(\cdot, \boldsymbol{x}_k)$ of F at $\boldsymbol{x}_k$:

$$Q(\boldsymbol{x}, \boldsymbol{x}_k) = F(\boldsymbol{x}_k) + (\boldsymbol{x} - \boldsymbol{x}_k)^\top \nabla F(\boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^\top \boldsymbol{A}_k(\boldsymbol{x} - \boldsymbol{x}_k)$$

❷ Minimize it within the subspace spanned by the columns of a matrix $\boldsymbol{D}_k \in \mathbb{R}^{N \times M_k}$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{D}_k(\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k)^\dagger \boldsymbol{D}_k^\top \nabla F(\boldsymbol{x}_k)$$

✘ MM algorithm : $\text{rank}(\boldsymbol{D}_k) = N \rightsquigarrow$ Large computational cost.

Majorize-Minimize subspace algorithm [Chouzenoux *et al.*, 2013]

⋆ Minimize differentiable and nonconvex function F on $\mathbb{R}^N$.

At each iteration $k \in \mathbb{N}$:

1. Build a quadratic majorant function $Q(\cdot, \boldsymbol{x}_k)$ of F at $\boldsymbol{x}_k$:

$$Q(\boldsymbol{x}, \boldsymbol{x}_k) = F(\boldsymbol{x}_k) + (\boldsymbol{x} - \boldsymbol{x}_k)^\top \nabla F(\boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^\top \boldsymbol{A}_k(\boldsymbol{x} - \boldsymbol{x}_k)$$

2. Minimize it within the subspace spanned by the columns of a matrix $\boldsymbol{D}_k \in \mathbb{R}^{N \times M_k}$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{D}_k(\boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k)^\dagger \boldsymbol{D}_k^\top \nabla F(\boldsymbol{x}_k)$$

✘ MM algorithm : rank$(\boldsymbol{D}_k) = N \rightsquigarrow$ Large computational cost.

☞ 3MG algorithm : $M_k = 2$ and $\boldsymbol{D}_k = [\nabla F(\boldsymbol{x}_k) \mid \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$.

Other examples of subspace construction

Subspace name	**Set of directions D_k**
Memory gradient	$[-\nabla F(\boldsymbol{x}_k) \,\vert\, \boldsymbol{d}_{k-1}]$
Supermemory gradient	$[-\nabla F(\boldsymbol{x}_k) \,\vert\, \boldsymbol{d}_{k-1} \,\vert\, \dots \,\vert\, \boldsymbol{d}_{k-m}]$
Gradient subspace	$[-\nabla F(\boldsymbol{x}_k) \,\vert\, -\nabla F(\boldsymbol{x}_{k-1}) \,\vert\, \dots \,\vert\, -\nabla F(\boldsymbol{x}_{k-m})]$
Nemirovski subspace	$\left[-\nabla F(\boldsymbol{x}_k) \,\vert\, \boldsymbol{x}_k - \boldsymbol{x}_1 \,\vert\, \sum_{i=0}^{k} \omega_i \nabla F(\boldsymbol{x}_k)\right]$
Sequential subspace	$\left[-\nabla F(\boldsymbol{x}_k) \,\vert\, \boldsymbol{x}_k - \boldsymbol{x}_1 \,\vert\, \sum_{i=0}^{k} \omega_i \nabla F(\boldsymbol{x}_k) \,\vert\, \boldsymbol{d}_{k-1} \,\vert\, \dots \,\vert\, \boldsymbol{d}_{k-m}\right]$
Quasi-Newton subspace	$[-\nabla F(\boldsymbol{x}_k) \,\vert\, \boldsymbol{\delta}_{k-1} \,\vert\, \dots \vert \boldsymbol{\delta}_{k-m} \,\vert\, \boldsymbol{d}_{k-1} \,\vert\, \dots \,\vert\, \boldsymbol{d}_{k-m}]$

For all $k \in \mathbb{N}$, $(\omega_i)_{1 \le i \le k} \in \mathbb{R}^N$, $\boldsymbol{d}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$, $\boldsymbol{\delta}_k = \nabla F(\boldsymbol{x}_{k+1}) - \nabla F(\boldsymbol{x}_k)$.

3MG algorithm

Initialize $\boldsymbol{x}_0 \in \mathbb{R}^N$
For $k = 0, 1, 2, \ldots$
 Compute $\nabla F(\boldsymbol{x}_k)$
 If $k = 0$
  $\boldsymbol{D}_k = -\nabla F(\boldsymbol{x}_0)$
 Else
  $\boldsymbol{D}_k = [-\nabla F(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}_{k-1}]$
 $\boldsymbol{S}_k = \boldsymbol{D}_k^\top \boldsymbol{A}_k \boldsymbol{D}_k$
 $\boldsymbol{u}_k = \boldsymbol{S}_k^\dagger \boldsymbol{D}_k^\top \nabla F(\boldsymbol{x}_k)$
 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{D}_k \boldsymbol{u}_k$

$\rightsquigarrow$ Low computational cost since $\boldsymbol{S}_k$ is of dimension $M_k \times M_k$, with $M_k \in \{1, 2\}$.

$\rightsquigarrow$ Complexity reductions possible by taking into account the structures of F and $\boldsymbol{D}_k$.

Link between MM-subspace and other approaches

- When F is quadratic and $F \equiv Q$, 3MG is equivalent to the famous linear conjugate gradient.
- More generally, 3MG can be viewed as a special instance of a nonlinear conjugate gradient method with closed forms for stepsize/conjugacy parameters.
- MM-subspace, with Quasi-Newton direction set, is similar to a low memory BFGS algorithm with a specific combination of memory directions and closed form stepsize parameter.
- MM-subspace, associated with directions spanning the whole space $\mathbb{R}^N$, is equivalent to a half-quadratic approach.

Convergence theorem for 3MG

Let assume that:

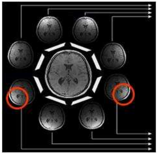
1. $F : \mathbb{R}^N \to \mathbb{R}$ is a coercive, differentiable function.
2. There exists $(\underline{\nu}, \overline{\nu}) \in]0, +\infty[^2$ such that $(\forall k \in \mathbb{N})$ $\underline{\nu}\,\mathbf{Id} \preceq \boldsymbol{A}_k \preceq \overline{\nu}\,\mathbf{Id}$,

Then, the following hold:

- $\|\nabla F(\boldsymbol{x}_k)\| \to 0$ and $F(\boldsymbol{x}_k) \searrow F(\widehat{\boldsymbol{x}})$ where $\widehat{\boldsymbol{x}}$ is a critical point of F.
- If F is convex, any sequential cluster point of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is a minimizer of F.
- If F is strongly convex, then $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges to the unique (global) minimizer $\widehat{\boldsymbol{x}}$ of F
- If F satisfies the Kurdyka-Łojasiewicz inequality, then the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ converges to a critical point of F.

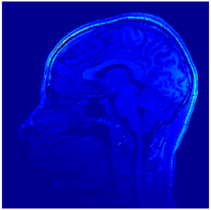
Application to parallel MRI [Florescu *et al.* - 2014]

Challenges:

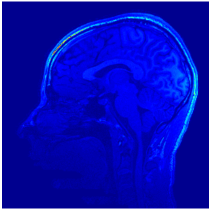


- Parallel acquisition and compressive sensing
- Complex-valued signals

Results:

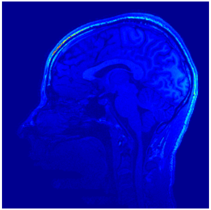


Original



3MG - convex

SNR = 20.05 dB



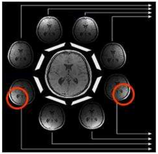
3MG - nonconvex

SNR = 20.27 dB

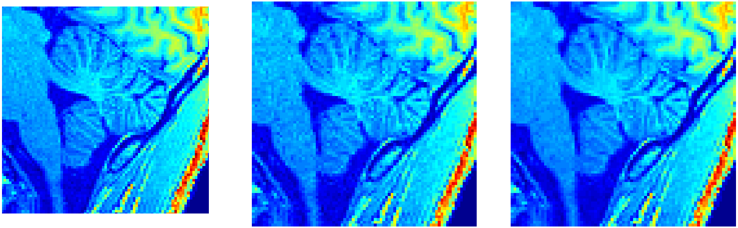
Application to parallel MRI [Florescu *et al.* - 2014]

Challenges:

- Parallel acquisition and compressive sensing
- Complex-valued signals



Results:



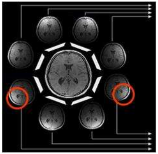
Original (zoom)

3MG - convex
SNR = 20.05 dB

3MG - nonconvex
SNR = 20.27 dB

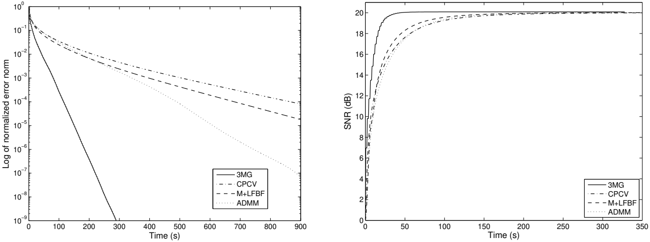
Application to parallel MRI [Florescu *et al.* - 2014]

Challenges:



- Parallel acquisition and compressive sensing
- Complex-valued signals

Results:



Convergence speed of 3MG, compared with several proximal-based algorithms

3MG in high dimensional problems

3MG algorithm outperforms state-of-the arts optimization algorithms in many image processing applications.

Problem: **Computational issues with very large-size problems.**

Main reasons: High computational time; High storage cost.

Large value of N

$\Downarrow$

Block parallel approach

Large value of M

$\Downarrow$

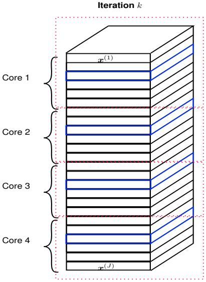
Online approach

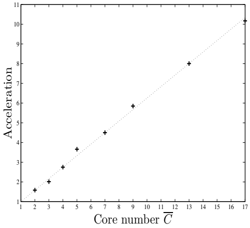
Parallel 3MG algorithm [Cadoni *et al.*, 2016]

How to make 3MG algorithm efficient for parallel implementation ?

At each iteration $k \in \mathbb{N}$:

1. Choose a subset of block indexes $\mathcal{S}_k \subset \{1, \ldots, J\}$.
2. Update the selected blocks using a 3MG step performed in parallel thanks to a block-diagonal MM metric.





- Application to 3D image deblurring with space-variant PSF *(CNRS OPTIMISM project)*.
- SPMD implementation on Matlab Parallel Toolbox.
- Great potential for parallelization.

Stochastic 3MG algorithm

STOCHASTIC PROBLEM

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \quad \left(F(\boldsymbol{x}) = \frac{1}{2}\mathbb{E}(\|\mathbf{y}_j - \mathbf{h}_j^\top \boldsymbol{x}\|^2) + \Psi(\boldsymbol{x}) \right)$$

⋆ The second-order statistics of $(\mathbf{h}_j, \mathbf{y}_j)_{j\geq 1}$ are estimated online in an adaptive manner.

NUMEROUS APPLICATIONS:

- supervised classification
- inverse problems
- system identification
- linear prediction/interpolation
- echo cancellation
- channel equalization

How to find a fast and flexible stochastic optimization algorithm with theoretical convergence guarantees ?

?

Stochastic 3MG algorithm [Chouzenoux and Pesquet, 2017]

At each iteration $j \in \mathbb{N}^*$:

1. Build an estimate of the objective function:

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad F_j(\boldsymbol{x}) = \frac{1}{2j} \sum_{k=1}^{j} \|\mathbf{y}_k - \mathbf{h}_k^\top \boldsymbol{x}\|^2 + \Psi(\boldsymbol{x})$$

2. Construct a quadratic majorant for F_j.
3. Minimize in a memory gradient subspace.
4. Perform recursive updates of the second-order statistics.

✓ CONVERGENCE GUARANTEES on the sequence $(\mathbf{x}_j)_{j\geq 1}$.

✓ REDUCED COMPLEXITY thanks to recursive update scheme.

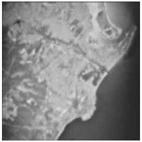
✓ CONVERGENCE RATE ANALYSIS in stochastic and batch case ([Chouzenoux and Pesquet, 2016]).

Application to 2D filter identification [Chouzenoux *et al.* - 2014]

OBSERVATION MODEL

$$\boldsymbol{y} = S(\overline{\boldsymbol{x}})\boldsymbol{h} + \boldsymbol{w}$$

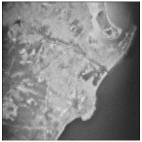
- $\boldsymbol{h} \in \mathbb{R}^L$ large size original image ($L = 4096^2$),
- $\overline{\boldsymbol{x}} \in \mathbb{R}^N$ unknown two-dimensional blur kernel ($N = 21^2$),
- $S(\overline{\boldsymbol{x}})$ Hankel-block Hankel matrix such that $S(\overline{\boldsymbol{x}})\boldsymbol{h} = \boldsymbol{H}\overline{\boldsymbol{x}}$,
- $\boldsymbol{w} \in \mathbb{R}^L$ realization of white $\mathcal{N}(0, 0.03^2)$ noise (BSNR = 25.7 dB)
- $\boldsymbol{y} \in \mathbb{R}^L$ blurred and noisy image.



Application to 2D filter identification [Chouzenoux *et al.* - 2014]

OBSERVATION MODEL

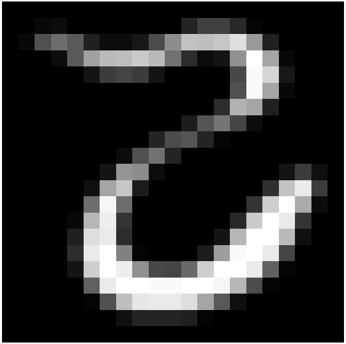
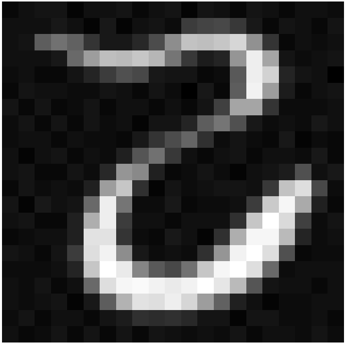
$$\boldsymbol{y} = S(\overline{\boldsymbol{x}})\boldsymbol{h} + \boldsymbol{w}$$



- $\boldsymbol{h} \in \mathbb{R}^L$ large size original image ($L = 4096^2$),
- $\overline{\boldsymbol{x}} \in \mathbb{R}^N$ unknown two-dimensional blur kernel ($N = 21^2$),
- $S(\overline{\boldsymbol{x}})$ Hankel-block Hankel matrix such that $S(\overline{\boldsymbol{x}})\boldsymbol{h} = \boldsymbol{H}\overline{\boldsymbol{x}}$,
- $\boldsymbol{w} \in \mathbb{R}^L$ realization of white $\mathcal{N}(0, 0.03^2)$ noise (BSNR = 25.7 dB)
- $\boldsymbol{y} \in \mathbb{R}^L$ blurred and noisy image.

$\Rightarrow$ **Minimization of a penalized MSE criterion:** $\mathbf{y}_k \in \mathbb{R}^Q$ and $\mathbf{h}_k^\top \in \mathbb{R}^{Q\times N}$: Q lines of $\boldsymbol{y}$ and $\boldsymbol{H}$, $\vartheta = 1$, and Ψ isotropic penalization on the gradient of $\mathbf{x}$ ($\sim$ smoothed version of total variation prior).

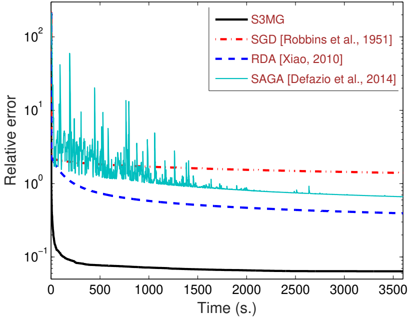
Application to 2D filter identification

Original blur kernel 21×21. Estimated blur kernel, relative error 0.064.

- The regularization parameters are optimized manually.

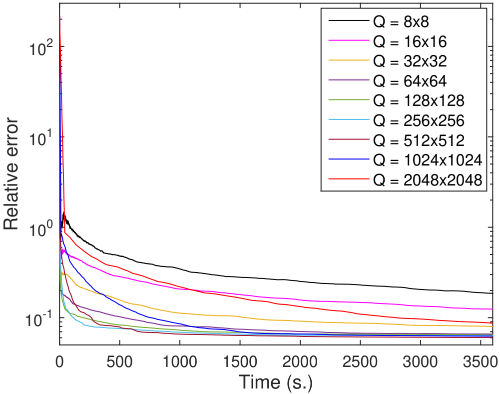
Application to 2D filter identification



Comparison of stochastic 3MG algorithm, SGD algorithm with decreasing stepsize $\propto j^{-1/2}$, and SAGA/RDA algorithms with constant stepsizes.

- The stepsize values in SGD/SAGA/RDA methods are optimized manually.
- The S3MG algorithm leads to a faster convergence.

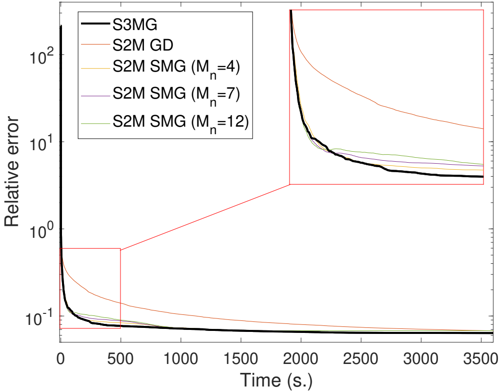
Application to 2D filter identification



Effect of the minibatch size Q on the convergence speed of S3MG.

- The best trade-off is obtained for $Q = 256 \times 256$.

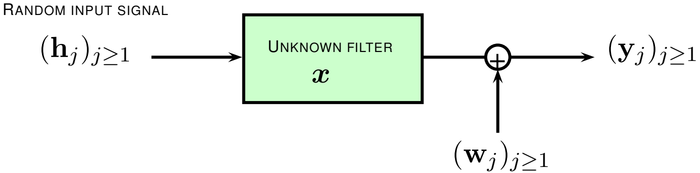
Application to 2D filter identification

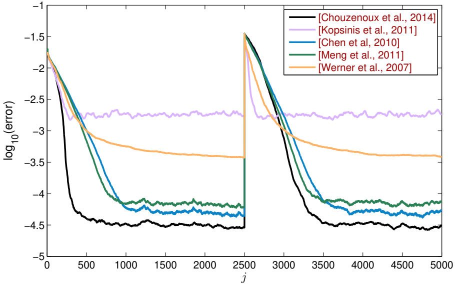


Effect of the choice of the subspace on the convergence speed.

- The best trade-off is obtained for memory gradient subspace.

Application to sparse adaptive filtering





- $\boldsymbol{x}$: sparse linear filter with abrupt change at $j = 2500$.
- S3MG algorithm with forgetting factor and smoothed ℓ_0 penalty.
- Minimal estimation error, and good tracking properties.

Variable metric forward-backward algorithm

Variable metric FB algorithm [Chouzenoux *et al.*, 2014]

⋆ Minimize $F = f_1 + f_2$ with f_1 Lipschitz-differentiable and f_2 non smooth.

⇒ Forward-backward: gradient steps on f_1 and proximal steps on f_2:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \text{prox}_{\theta_k f_2} \left(\boldsymbol{x}_k - \theta_k \nabla f_1(\boldsymbol{x}_k) \right).$$

✘ slow convergence in practice.

Variable metric FB algorithm [Chouzenoux *et al.*, 2014]

⋆ Minimize $F = f_1 + f_2$ with f_1 Lipschitz-differentiable and f_2 non smooth.

⇒ Forward-backward: gradient steps on f_1 and proximal steps on f_2:

☛ Use MM framework to propose an efficient variable metric strategy:

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \mathrm{prox}_{\theta_k^{-1}\boldsymbol{A}_k, f_2}\left(\boldsymbol{x}_k - \theta_k \boldsymbol{A}_k^{-1} \nabla f_1(\boldsymbol{x}_k)\right).$$

✓ CONVERGENCE of the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ to a critical point of F under KL assumption.

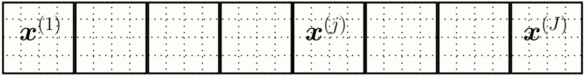
✓ ROBUSTNESS TO ERRORS in the computation of the proximity operator within the metric.

✓ EFFICIENT CONSTRUCTION of the preconditioning matrices thanks to the MM framework.

Block alternating strategy

The vector of unknowns x is partitioned into **block subsets**.
At each iteration, **one** or **several blocks** are updated.

$\boldsymbol{x} =$



PRACTICAL ADVANTAGES:

- ✓ Control of memory for large scale image processing (eg, 3D, video).
- ✓ Flexibility of alternating scheme suitable to blind/unmixing problems.
- ✓ A first step towards parallel and distributed implementation.

How to find efficient and reliable block alternating schemes for nonconvex and/or non differentiable optimization problems ?

?

Block coordinate VMFB algorithm [Chouzenoux *et al.*, 2016]

⋆ Minimize $F = f_1 + f_2$ with f_1 smooth and f_2 non differentiable.

At each iteration $k \in \mathbb{N}$:

1. Choose a block index $j_k \in \{1, \ldots, J\}$ according to a quasi-cyclic rule.
2. Perform a gradient step on the restriction of f_1 to block j_k, using a MM preconditioner.
3. Perform a proximal step on the restriction of f_2 to block j_k, within the MM metric.

✓ CONVERGENCE GUARANTEES on the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ under KL assumption.

✓ EXPERIMENTAL VALIDATION in numerous applications of image/signal processing (eg, phase retrieval, spectral unmixing, blind deconvolution).

Application to phase retrieval

OBSERVATION MODEL:

We observe measurements $\boldsymbol{y} \in [0,+\infty)^S$ through

$$\boldsymbol{y} = |\boldsymbol{H}\overline{\boldsymbol{v}}| + \boldsymbol{w}.$$

- $\overline{\boldsymbol{v}} \in \mathbb{R}^M$ $\rightsquigarrow$ original unknown image
- $\boldsymbol{H} \in \mathbb{C}^{S\times M}$ $\rightsquigarrow$ degradation operator
- $\boldsymbol{w} \in [0,+\infty)^S$ $\rightsquigarrow$ additive noise.

Objective: Produce an estimate $\widehat{\boldsymbol{v}}$ of the target image $\overline{\boldsymbol{v}}$ from the observed measurements $\boldsymbol{y}$.

Application fields:

- Crystallography [Harrison *et al.* - 1993]
- Phase contrast tomography [Bauschke *et al.* - 2005]
- Coherent diffraction imaging [Shechtman, *et al.* - 2013]

Application to phase retrieval

OBSERVATION MODEL:

We observe measurements $\boldsymbol{y} \in [0, +\infty)^S$ through

$$\boldsymbol{y} = |\boldsymbol{H}\overline{\boldsymbol{v}}| + \boldsymbol{w}.$$

- $\overline{\boldsymbol{v}} \in \mathbb{R}^M$ $\rightsquigarrow$ original unknown image
- $\boldsymbol{H} \in \mathbb{C}^{S \times M}$ $\rightsquigarrow$ degradation operator
- $\boldsymbol{w} \in [0, +\infty)^S$ $\rightsquigarrow$ additive noise.

What happens if $\overline{v}$ is complex?

$$\overline{\boldsymbol{v}} = \overline{\boldsymbol{v}}_{\mathcal{R}} + \mathrm{i}\, \overline{\boldsymbol{v}}_{\mathcal{I}}$$

$$\rightsquigarrow \quad \boldsymbol{y} = \left| (\boldsymbol{H}_{\mathcal{R}} + \mathrm{i}\, \boldsymbol{H}_{\mathcal{I}})\, (\overline{\boldsymbol{v}}_{\mathcal{R}} + \mathrm{i}\, \overline{\boldsymbol{v}}_{\mathcal{I}}) \right| + \boldsymbol{w}$$

$$\rightsquigarrow \quad \boldsymbol{y} = \Bigg| \underbrace{[\boldsymbol{H}_{\mathcal{R}} + \mathrm{i}\, \boldsymbol{H}_{\mathcal{I}} \mid -\boldsymbol{H}_{\mathcal{I}} + \mathrm{i}\, \boldsymbol{H}_{\mathcal{R}}]}_{\text{Complex}} \underbrace{\begin{bmatrix} \overline{\boldsymbol{v}}_{\mathcal{R}} \\ \overline{\boldsymbol{v}}_{\mathcal{I}} \end{bmatrix}}_{\text{Real}} \Bigg| + \boldsymbol{w}$$

State of the art

- Alternating projections methods:
  [Gerchberg *et al.* - 1972] [Fienup - 1972] [Bauschke *et al.* - 2002]

- Convex relaxations based on SDP programming:
  - ⇝ PhaseLift algorithm [Candés *et al.* - 2013]
  - ⇝ PhaseCut algorithm [Waldspurger *et al.* - 2013]

- Regularized approaches assuming that $\overline{v}$ is sparse in a given dictionary:
  - ⇝ SPD programming [Fogel *et al.* - 2013]
  - ⇝ Alternating projections [Mukherjee *et al.* - 2012]
  - ⇝ Greedy algorithm [Shechtman *et al.* - 2013]

Proposed method

Synthesis approach: Let $\boldsymbol{W} \in \mathbb{R}^{M \times N}$, $M \leq N$, be a given frame synthesis operator such that $\widehat{\boldsymbol{v}} = \boldsymbol{W}\widehat{\boldsymbol{x}}$.

The frame coefficient vector $\widehat{\boldsymbol{x}} \in \mathbb{R}^N$ is estimated by minimizing $f_1 + f_2$ where

- f_1 is a smooth nonconvex data fidelity term,

$$(\forall \boldsymbol{x} \in \mathbb{R}^N) \quad f_1(\boldsymbol{x}) = \sum_{s=1}^{S} \varphi^{(s)}([\boldsymbol{HWx}]^{(s)}), \qquad \text{where}$$

$$(\forall u \in \mathbb{C}) \quad \varphi^{(s)}(u) = \frac{1}{2}\left(|u|^2 + (z^{(s)})^2\right) - z^{(s)}\left(|u|^2 + \delta^2\right)^{1/2},$$

with $\delta \in (0, +\infty)$.

- f_2 is a block separable regularization function.

Construction of the preconditioning matrices

At iteration $k \in \mathbb{N}$, let j_k be the chosen index in $\{1, \dots, J\}$ and let $\boldsymbol{x}_k$ be the k-th iterate generated by the BC-VMFB algorithm.

The majorization condition is fulfilled by the diagonal matrix

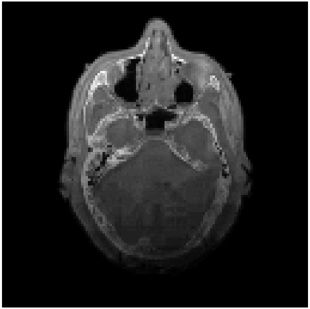
$$\boldsymbol{A}_{j_k} = \mathrm{Diag}\left(\boldsymbol{\Omega}_{j_k}^{\top} 1_S\right)$$

where 1_S is the unit vector on $\mathbb{R}^S$ and the elements of $\boldsymbol{\Omega}_{j_k} \in \mathbb{R}^{S \times N_{j_k}}$ are given by

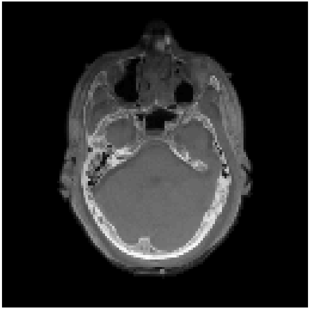
$$\Omega^{(s,n)} = \left|[\boldsymbol{H}\boldsymbol{W}]_{\mathcal{R}}^{(s,n)}\right| \sum_{n'=1}^{N} \left|[\boldsymbol{H}\boldsymbol{W}]_{\mathcal{R}}^{(s,n')}\right| + \left|[\boldsymbol{H}\boldsymbol{W}]_{\mathcal{I}}^{(s,n)}\right| \sum_{n'=1}^{N} \left|[\boldsymbol{H}\boldsymbol{W}]_{\mathcal{I}}^{(s,n')}\right|.$$

Simulation results

- **Complex valued original image:**
  $\overline{\boldsymbol{v}} \in \mathbb{C}^M$ with $M = 128 \times 128$



Real part $\overline{\boldsymbol{v}}_{\mathcal{R}} \in \mathbb{R}^M$



Imaginary part $\overline{\boldsymbol{v}}_{\mathcal{I}} \in \mathbb{R}^M$

Simulation results

▶ **Observation matrix:**

$\boldsymbol{H} \in \mathbb{C}^{S \times M}$ is the composition of

- a projection matrix modeling $S = 23400$ Radon projections from
  - 128 parallel acquisition lines,
  - 180 angles regularly distributed on $[0, \pi)$,
- a complex-valued blur operator.

⇝ Reminiscent of the phase contrast tomography model from [Davidoiu *et al.* - 2012].

Simulation results

- **Observation matrix:**

$\boldsymbol{H} \in \mathbb{C}^{S \times M}$ is the composition of

- a projection matrix modeling $S = 23400$ Radon projections from
  - 128 parallel acquisition lines,
  - 180 angles regularly distributed on $[0, \pi)$,
- a complex-valued blur operator.

$\rightsquigarrow$ Reminiscent of the phase contrast tomography model from [Davidoiu *et al.* - 2012].

- **Synthesis frame operator:**

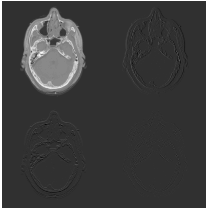
$\boldsymbol{W} \in \mathbb{C}^{M \times N}$, $N = 8M$, is such that $\boldsymbol{x} = (x^{(n)})_{1 \le n \le N}$ is the concatenation of an overcomplete Haar decomposition of $\boldsymbol{v}_{\mathcal{R}}$ (resp. $\boldsymbol{v}_{\mathcal{I}}$) for one resolution level.

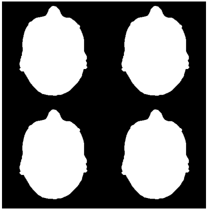
Simulation results

- **Regularization function:**

f_2 is the sum, for $p \in \{1, \ldots, 4M\}$, of

$$\rho^{(p)}(u^{(p)}) = \begin{cases} \vartheta_p \|u^{(p)} - \omega_p\|_2^{\kappa_p} & \text{if } p \notin \mathbb{E}, \\ 0 & \text{if } p \in \mathbb{E} \text{ and } u^{(p)} = 0, \\ +\infty & \text{otherwise,} \end{cases}$$





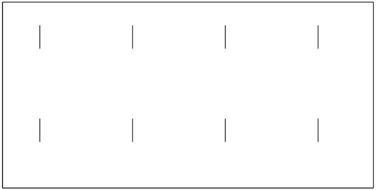
where

- $u^{(p)} \in \mathbb{R}^2$ is the p-th pair of frame coefficients corresponding to the real and imaginary parts of the image,
- $\mathbb{E}$ is the object background,
- $\kappa_p = 1$, $\vartheta_p = \vartheta^d \in (0, +\infty)$ for the detail subbands, and $\kappa_p = 2$, $\vartheta_p = \vartheta^a \in (0, +\infty)$ for the approximation subbands,
- $\omega_p \in \mathbb{R}^2$ controls the mean value of $u^{(p)}$.

Simulation results

- **Definition of blocks:**

For every j, $\boldsymbol{x}^{(j)} \in \mathbb{R}^{8Q}$ gathers 8 blocks from the approximation and detail subbands of both real and imaginary parts.



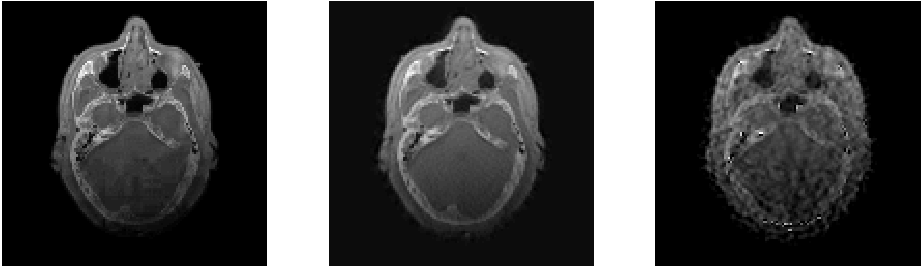
Indices of a block $\boldsymbol{x}^{(j)}$ for $Q = 32$.

⇝ At each iteration $k \in \mathbb{N}$, j_k is randomly chosen so that each block is updated at least once per J iterations.

Simulation results

Real part

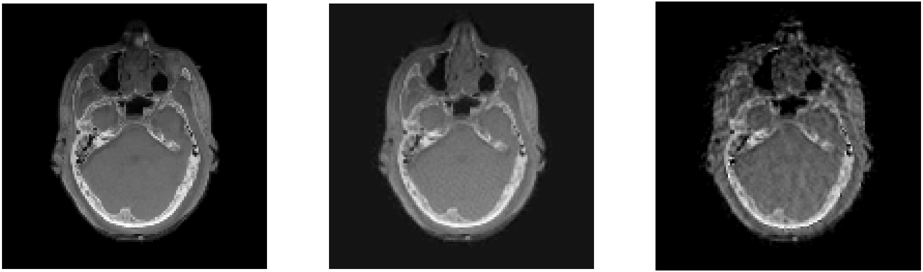
- Original image $\overline{v}_{\mathcal{R}}$
- Reconstructed image $\widehat{v}_{\mathcal{R}}$ with BC-VMFB Algorithm: SNR $= 21.27$ dB.
- Reconstructed image $\widehat{v}_{\mathcal{R}}$ with the ℓ_0-regularized Fienup algorithm from [Mukherjee *et al.* - 2012]: SNR $= 14.45$ dB.



Simulation results

Imaginary part

- Original image $\overline{\boldsymbol{v}}_{\mathcal{I}}$
- Reconstructed image $\widehat{\boldsymbol{v}}_{\mathcal{I}}$ with BC-VMFB Algorithm: SNR $= 21.27$ dB.
- Reconstructed image $\widehat{\boldsymbol{v}}_{\mathcal{I}}$ with the ℓ_0-regularized Fienup algorithm from [Mukherjee *et al.* - 2012]: SNR $= 14.45$ dB.



Conclusion

MM algorithms allow to solve efficiently optimization problems of image/signal processing.

Several extensions are proposed for *very large scale problems* :

⇝ Block Parallel 3MG

⇝ Stochastic 3MG

⇝ Block-coordinate VMFB

More to come, with ANR MajIC project.

THANK YOU !

Some references

- S. Cadoni, E. Chouzenoux, J.-C. Pesquet, C. Chaux.
  A Block Parallel Majorize-Minimize Memory Gradient Algorithm
  in *Proc. IEEE Conf. Image Process.* (ICIP 2016), pp. 3194–3198, Phoenix, AZ, 25-28 Sep. 2016. Best paper award finalist.
- E. Chouzenoux, J. Idier, S. Moussaoui
  A Majorize-Minimize strategy for subspace optimization applied to image restoration
  in *IEEE Trans. Image Process.*, vol.20, no.18, pp. 1517-1528, 2011.
- E. Chouzenoux, A. Jezierska, J.-C. Pesquet, H. Talbot
  A Majorize-Minimize Subspace Approach for ℓ_2-ℓ_0 Image Regularization
  in *SIAM J. Imag. Sci.*, vol.6, no.1, pp. 563-591,2013.
- E. Chouzenoux, J.-C. Pesquet and A. Repetti
  A Block Coordinate Variable Metric Forward-Backward Algorithm
  in *J. Global Optim.*, vol. 66, no. 3, pp. 457485, 2016
- E. Chouzenoux and J.-C. Pesquet
  Convergence Rate Analysis of the Majorize-Minimize Subspace Algorithm
  in *IEEE Sig. Process. Lett.*, vol. 23, no. 9, pp. 1284-1288, Sep. 2016
- E. Chouzenoux and J.-C. Pesquet
  A Stochastic Majorize-Minimize Subspace Algorithm for Online Penalized Least Squares Estimation
  in *IEEE Trans. Sig. Proc.*, vol. 65, no. 18, pp. 4770-4783, 2017.