A STOCHASTIC MAJORIZE-MINIMIZE SUBSPACE ALGORITHM WITH APPLICATION TO FILTER IDENTIFICATION PROBLEMS

Émilie Chouzenoux and Jean-Christophe Pesquet

Laboratoire d'Informatique Gaspard Monge - CNRS
Univ. Paris-Est Marne-la-Vallée, France

25 March 16

Introduction

STOCHASTIC PROBLEM

$$\underset{\boldsymbol{h}\in\mathbb{R}^N}{\text{minimize}} \quad \frac{1}{2}\mathbb{E}(\|\mathbf{y}_n - \mathbf{X}_n^\top \boldsymbol{h}\|^2) + \Psi(\boldsymbol{h})$$

$(\mathbf{X}_n)_{n\geqslant 1}$ random matrices in $\mathbb{R}^{N\times Q}$,
$(\mathbf{y}_n)_{n\geqslant 1}$ random vectors in $\mathbb{R}^Q$ with

$$(\forall n \in \mathbb{N}^*) \qquad \begin{aligned} \mathsf{E}(\|\mathbf{y}_n\|^2) &= \varrho \\ \mathsf{E}(\mathbf{X}_n\mathbf{y}_n) &= \boldsymbol{r} \\ \mathsf{E}(\mathbf{X}_n\mathbf{X}_n^\top) &= \boldsymbol{R}, \end{aligned}$$

and $\Psi\colon \mathbb{R}^N \to \mathbb{R}$ regularization function.

Introduction

NUMEROUS EXAMPLES:

- supervised classification
- inverse problems
- system identification, channel equalization
- linear prediction/interpolation
- echo cancellation, interference removal
- ...

How to solve the problem efficiently when the second-order statistics of $(\mathbf{X}_n, \mathbf{y}_n)_{n \geqslant 1}$ are estimated online, in an adaptive manner ?

Introduction

NUMEROUS EXAMPLES:

- supervised classification
- inverse problems
- system identification, channel equalization
- linear prediction/interpolation
- echo cancellation, interference removal
- ...

How to solve the problem efficiently when the second-order statistics of $(\mathbf{X}_n, \mathbf{y}_n)_{n \geqslant 1}$ are estimated online, in an adaptive manner ? $\Rightarrow$ Majorize-Minimize approach.

Outline

* PROBLEM FORMULATION
  - Stochastic approximation of the criterion
  - Form of the regularization function
* BATCH MAJORIZE-MINIMIZE SUBSPACE ALGORITHM
  - Quadratic tangent majorant
  - Subspace acceleration strategy
  - Convergence results
* STOCHASTIC MAJORIZE-MINIMIZE SUBSPACE ALGORITHM
  - Proposed algorithm
  - Complexity study
  - Convergence results
* APPLICATIONS TO FILTER IDENTIFICATIONS
  - Online estimation of 2D blur
  - Sparse adaptive filtering

Problem formulation

Form of the objective function

Estimate of the objective function at iteration $n \in \mathbb{N}^*$:

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \mathrm{F}_n(\boldsymbol{h}) = \frac{1}{2n}\sum_{k=1}^{n} \|\mathbf{y}_k - \mathbf{X}_k^\top \boldsymbol{h}\|^2 + \Psi(\boldsymbol{h})$$
$$= \frac{1}{2}\rho_n - \mathbf{r}_n{}^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \mathbf{R}_n \boldsymbol{h} + \Psi(\boldsymbol{h})$$

* **Batch case:**

$$\rho_n \equiv \varrho,\ \mathbf{r}_n \equiv \boldsymbol{r} \text{ and } \mathbf{R}_n \equiv \boldsymbol{R}.$$

* **Online case:**

$$\rho_n = \frac{1}{n}\sum_{k=1}^{n} \|\mathbf{y}_k\|^2,\ \mathbf{r}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k \mathbf{y}_k,\ \mathbf{R}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k \mathbf{X}_k^\top.$$

Form of the regularization function

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \Psi(\boldsymbol{h}) = \underbrace{\frac{1}{2}\boldsymbol{h}^\top \boldsymbol{V}_0 \boldsymbol{h} - \boldsymbol{v}_0^\top \boldsymbol{h}}_{\text{elastic net penalization}} + \sum_{s=1}^{S} \psi_s(\|\boldsymbol{V}_s \boldsymbol{h} - \boldsymbol{v}_s\|)$$

where $\boldsymbol{v}_0 \in \mathbb{R}^N$, $\boldsymbol{V}_0 \in \mathbb{R}^{N \times N}$ symmetric positive semi-definite, for every $s \in \{1, \ldots, S\}$, $\boldsymbol{v}_s \in \mathbb{R}^{P_s}$, $\boldsymbol{V}_s \in \mathbb{R}^{P_s \times N}$, $\psi_s : \mathbb{R} \to \mathbb{R}$

$\rightsquigarrow$ ability to take into account linear operators.

Form of the regularization function

$$\boxed{(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \Psi(\boldsymbol{h}) = \underbrace{\frac{1}{2}\boldsymbol{h}^\top \boldsymbol{V}_0 \boldsymbol{h} - \boldsymbol{v}_0^\top \boldsymbol{h}}_{\text{elastic net penalization}} + \sum_{s=1}^{S} \psi_s(\|\boldsymbol{V}_s \boldsymbol{h} - \boldsymbol{v}_s\|)}$$

where $\boldsymbol{v}_0 \in \mathbb{R}^N$, $\boldsymbol{V}_0 \in \mathbb{R}^{N\times N}$ symmetric positive semi-definite, for every $s \in \{1,\dots,S\}$, $\boldsymbol{v}_s \in \mathbb{R}^{P_s}$, $\boldsymbol{V}_s \in \mathbb{R}^{P_s \times N}$, $\psi_s : \mathbb{R} \to \mathbb{R}$

$\rightsquigarrow$ ability to take into account linear operators.

Assumptions on $(\psi_s)_{1\leqslant s\leqslant S}$:

(i) For every $s \in \{1,\dots,S\}$, ψ_s is an even lower-bounded function, which is continuously differentiable, and $\lim_{\substack{t\to 0\\ t\neq 0}} \dot{\psi}_s(t)/t \in \mathbb{R}$,

where $\dot{\psi}_s$ denotes the derivative of ψ_s.

(ii) For every $s \in \{1,\dots,S\}$, $\psi_s(\sqrt{\cdot})$ is concave on $[0,+\infty[$.

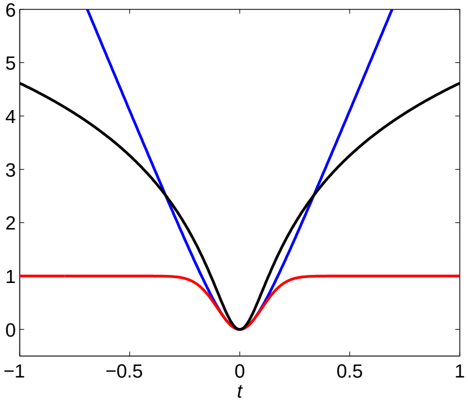
(iii) There exists $\overline{\nu} \in [0,+\infty[$ such that $(\forall s \in \{1,\dots,S\})$ $(\forall t \in [0,+\infty[)$ $0 \leqslant \nu_s(t) \leqslant \overline{\nu}$, where $\nu_s(t) = \dot{\psi}_s(t)/t$.

Examples of functions $(\psi_s)_{1\leqslant s\leqslant S}$

	$\lambda_s^{-1}\psi_s(t)$	Type	Name
Convex	$\lvert t\rvert - \delta_s \log(\lvert t\rvert/\delta_s + 1)$	$\ell_2 - \ell_1$	
	$\begin{cases} t^2 & \text{if } \lvert t\rvert < \delta_s \\ 2\delta_s\lvert t\rvert - \delta_s^2 & \text{otherwise} \end{cases}$	$\ell_2 - \ell_1$	Huber
	$\log(\cosh(t))$	$\ell_2 - \ell_1$	Green
	$(1 + t^2/\delta_s^2)^{\kappa_s/2} - 1$	$\ell_2 - \ell_{\kappa_s}$	
Nonconvex	$1 - \exp(-t^2/(2\delta_s^2))$	$\ell_2 - \ell_0$	Welsch
	$t^2/(2\delta_s^2 + t^2)$	$\ell_2 - \ell_0$	Geman -McClure
	$\begin{cases} 1 - (1 - t^2/(6\delta_s^2))^3 & \text{if } \lvert t\rvert \leqslant \sqrt{6}\delta_s \\ 1 & \text{otherwise} \end{cases}$	$\ell_2 - \ell_0$	Tukey biweight
	$\tanh(t^2/(2\delta_s^2))$	$\ell_2 - \ell_0$	Hyberbolic tangent
	$\log(1 + t^2/\delta_s^2)$	$\ell_2 - \log$	Cauchy
	$1 - \exp(1 - (1 + t^2/(2\delta_s^2))^{\kappa_s/2})$	$\ell_2 - \ell_{\kappa_s} - \ell_0$	Chouzenoux

$(\lambda_s, \delta_s) \in]0, +\infty[^2$, $\kappa_s \in [1, 2]$

Examples of functions $(\psi_s)_{1\leqslant s\leqslant S}$



$\psi_s(t) = (1+\frac{t^2}{\delta^2})^{1/2} - 1$, $\psi_s(t) = \log\left(1+\frac{t^2}{\delta^2}\right)$, $\psi_s(t) = 1-\exp(-\frac{t^2}{2\delta^2})$.

Batch majorize-minimize subspace algorithm

Majorize-Minimize principle

1. **Find a tractable surrogate for F** $\rightsquigarrow$ Majorization step

$\rightsquigarrow$ Quadratic tangent majorant

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \Theta(\boldsymbol{h}, \boldsymbol{h}_n) = F(\boldsymbol{h}_n) + \nabla F(\boldsymbol{h}_n)^\top (\boldsymbol{h} - \boldsymbol{h}_n) + \frac{1}{2}(\boldsymbol{h} - \boldsymbol{h}_n)^\top \boldsymbol{A}(\boldsymbol{h}_n)(\boldsymbol{h} - \boldsymbol{h}_n),$$

where

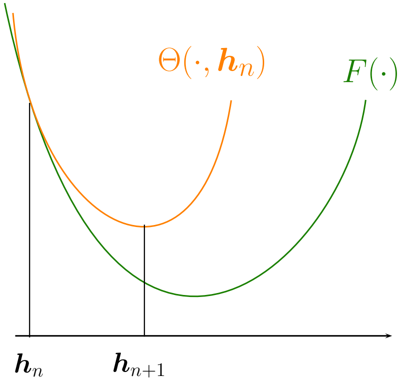
$$\boldsymbol{A}(\boldsymbol{h}) = \boldsymbol{R} + \boldsymbol{V}_0 + \boldsymbol{V}^\top \operatorname{Diag}\big(\boldsymbol{b}(\boldsymbol{h})\big)\boldsymbol{V} \in \mathbb{R}^{N\times N}$$
$$\boldsymbol{V} = [\boldsymbol{V}_1^\top \dots \boldsymbol{V}_S^\top]^\top \in \mathbb{R}^{P\times N}, \quad P = P_1 + \dots + P_S$$

and $\boldsymbol{b}(\boldsymbol{h}) = \big(b_i(\boldsymbol{h})\big)_{1\leqslant i\leqslant P} \in \mathbb{R}^P$ is such that

$$(\forall s \in \{1,\dots,S\})(\forall p \in \{1,\dots,P_s\}) \quad b_{P_1+\dots+P_{s-1}+p}(\boldsymbol{h}) = \nu_s(\|\boldsymbol{V}_s\boldsymbol{h} - \boldsymbol{v}_s\|).$$

Majorize-Minimize principle

1. **Find a tractable surrogate for** $F \rightsquigarrow$ **M**ajorization step



Subspace acceleration

2. Minimize in a subspace $\rightsquigarrow$ Minimization step

$$(\forall n \in \mathbb{N}^*) \qquad \boldsymbol{h}_{n+1} \in \underset{\boldsymbol{h} \in \operatorname{span} \boldsymbol{D}_n}{\operatorname{Argmin}} \; \Theta(\boldsymbol{h}, \boldsymbol{h}_n),$$

with $\boldsymbol{D}_n \in \mathbb{R}^{N \times M_n}$.

- $\operatorname{rank}(\boldsymbol{D}_n) = N \Rightarrow$ half-quadratic algorithm
- M_n small $\Rightarrow$ low-complexity per iteration.

Typical choice:

$$\boldsymbol{D}_n = \begin{cases} [-\nabla F(\boldsymbol{h}_n), \boldsymbol{h}_n, \boldsymbol{h}_n - \boldsymbol{h}_{n-1}] & \text{if } n > 1 \\ [-\nabla F(\boldsymbol{h}_1), \boldsymbol{h}_1] & \text{if } n = 1 \end{cases}$$

$\rightsquigarrow$ **3MG** algorithm

(similar ideas in TWIST, FISTA, NLCG, L-BFGS, ...)

Batch MM subspace algorithm

Initialize $\boldsymbol{D}_0, \boldsymbol{u}_0, \boldsymbol{h}_1 = \boldsymbol{D}_0 \boldsymbol{u}_0$
$\boldsymbol{D}_0^{\boldsymbol{R}} = \boldsymbol{R}\boldsymbol{D}_0, \boldsymbol{D}_0^{\boldsymbol{V}_0} = \boldsymbol{V}_0\boldsymbol{D}_0, \boldsymbol{D}_0^{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{D}_0$

For $n = 1, 2, \ldots$

$\boldsymbol{c}(\boldsymbol{h}_n) = \boldsymbol{r} + \boldsymbol{v}_0 + \boldsymbol{V}^\top \operatorname{Diag}\big(\boldsymbol{b}(\boldsymbol{h})\big)\boldsymbol{v}$
$\boldsymbol{D}_{n-1}^{\boldsymbol{A}} = \boldsymbol{D}_{n-1}^{\boldsymbol{R}} + \boldsymbol{D}_{n-1}^{\boldsymbol{V}_0} + \boldsymbol{V}^\top \operatorname{Diag}\big(\boldsymbol{b}(\boldsymbol{h}_n)\big)\boldsymbol{D}_{n-1}^{\boldsymbol{V}}$
$\nabla F(\boldsymbol{h}_n) = \boldsymbol{D}_{n-1}^{\boldsymbol{A}}\boldsymbol{u}_{n-1} - \boldsymbol{c}(\boldsymbol{h}_n)$
Set $\boldsymbol{D}_n$ using $\nabla F(\boldsymbol{h}_n)$
$\boldsymbol{D}_n^{\boldsymbol{R}} = \boldsymbol{R}\boldsymbol{D}_n, \boldsymbol{D}_n^{\boldsymbol{V}_0} = \boldsymbol{V}_0\boldsymbol{D}_n, \boldsymbol{D}_n^{\boldsymbol{V}} = \boldsymbol{V}\boldsymbol{D}_n$
$\boldsymbol{B}_n = \boldsymbol{D}_n^\top\left(\boldsymbol{D}_n^{\boldsymbol{V}_0} + \boldsymbol{D}_n^{\boldsymbol{R}}\right) + \left(\boldsymbol{D}_n^{\boldsymbol{V}}\right)^\top \operatorname{Diag}\big(\boldsymbol{b}(\boldsymbol{h}_n)\big)\boldsymbol{D}_n^{\boldsymbol{V}}$
$\boldsymbol{u}_n = \boldsymbol{B}_n^\dagger \boldsymbol{D}_n^\top \boldsymbol{c}(\mathbf{h}_n)$
$\boldsymbol{h}_{n+1} = \boldsymbol{D}_n \boldsymbol{u}_n$

Convergence theorem

Let assume that:

1. For every $n \in \mathbb{N}^*$, $\{\nabla F(\boldsymbol{h}_n), \boldsymbol{h}_n\} \subset \operatorname{span} \boldsymbol{D}_n$,
2. $\boldsymbol{R} + \boldsymbol{V}_0$ is a positive definite matrix.

Then, the following hold:

- $\|\nabla F(\boldsymbol{h}_n)\| \to 0$ and $F(\boldsymbol{h}_n) \searrow F(\widetilde{\boldsymbol{h}})$ where $\widetilde{\boldsymbol{h}}$ is a critical point of F.
- If F is convex, then any sequential cluster point of $(\boldsymbol{h}_n)_{n\geqslant 1}$ is a minimizer of F.
- If F is strictly convex, then $(\boldsymbol{h}_n)_{n\geqslant 1}$ converges to the unique (global) minimizer $\widehat{\boldsymbol{h}}$ of F.
- If F satisfies the Kurdyka-Łojasiewicz inequality, then the sequence $(\boldsymbol{h}_n)_{n\geqslant 1}$ converges to a critical point of F.

Stochastic majorize-minimize subspace algorithm

Stochastic approximation of the criterion

Estimate of the objective function at iteration $n \in \mathbb{N}^*$:

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \mathrm{F}_n(\boldsymbol{h}) = \frac{1}{2n} \sum_{k=1}^{n} \|\mathbf{y}_k - \mathbf{X}_k^\top \boldsymbol{h}\|^2 + \Psi(\boldsymbol{h})$$
$$= \frac{1}{2}\rho_n - \mathbf{r}_n{}^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \mathbf{R}_n \boldsymbol{h} + \Psi(\boldsymbol{h})$$

with $\rho_n = \frac{1}{n}\sum_{k=1}^{n} \|\mathbf{y}_k\|^2$, $\mathbf{r}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k \mathbf{y}_k$, and

$$\mathbf{R}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k \mathbf{X}_k^\top.$$

Stochastic approximation of the criterion

Estimate of the objective function at iteration $n \in \mathbb{N}^*$:

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \mathrm{F}_n(\boldsymbol{h}) = \frac{1}{2n} \sum_{k=1}^{n} \|\mathbf{y}_k - \mathbf{X}_k^\top \boldsymbol{h}\|^2 + \Psi(\boldsymbol{h})$$

$$= \frac{1}{2}\rho_n - \mathbf{r}_n{}^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \mathbf{R}_n \boldsymbol{h} + \Psi(\boldsymbol{h})$$

with $\rho_n = \frac{1}{n}\sum_{k=1}^n \|\mathbf{y}_k\|^2$, $\mathbf{r}_n = \frac{1}{n}\sum_{k=1}^n \mathbf{X}_k \mathbf{y}_k$, and

$\mathbf{R}_n = \frac{1}{n}\sum_{k=1}^n \mathbf{X}_k \mathbf{X}_k^\top$.

- **How to make the method adaptive to changes in the input statistics?**

Stochastic approximation of the criterion

Estimate of the objective function at iteration $n \in \mathbb{N}^*$:

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \mathrm{F}_n(\boldsymbol{h}) = \frac{1}{2\overline{\vartheta}_n} \sum_{k=1}^{n} \vartheta^{n-k} \|\mathbf{y}_k - \mathbf{X}_k^\top \boldsymbol{h}\|^2 + \Psi(\boldsymbol{h})$$
$$= \frac{1}{2}\rho_n - \mathbf{r}_n{}^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \mathbf{R}_n \boldsymbol{h} + \Psi(\boldsymbol{h})$$

with $\rho_n = \frac{1}{\overline{\vartheta}_n} \sum_{k=1}^n \vartheta^{n-k} \|\mathbf{y}_k\|^2$, $\mathbf{r}_n = \frac{1}{\overline{\vartheta}_n} \sum_{k=1}^n \vartheta^{n-k} \mathbf{X}_k \mathbf{y}_k$, and

$\mathbf{R}_n = \frac{1}{\overline{\vartheta}_n} \sum_{k=1}^n \vartheta^{n-k} \mathbf{X}_k \mathbf{X}_k^\top$.

$\rightsquigarrow$ **Forgetting factor:**

$$\overline{\vartheta}_n = \sum_{k=0}^{n-1} \vartheta^k = \begin{cases} n & \text{if } \vartheta = 1 \\ \dfrac{1-\vartheta^n}{1-\vartheta} & \text{if } \vartheta \in]0,1[, \end{cases} \quad \text{and} \quad \vartheta \in]0,1].$$

Stochastic MM subspace algorithm

Estimate of the objective function at iteration $n \in \mathbb{N}^*$:

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \mathrm{F}_n(\boldsymbol{h}) = \frac{1}{2\overline{\vartheta}_n} \sum_{k=1}^{n} \vartheta^{n-k} \|\mathbf{y}_k - \mathbf{X}_k^\top \boldsymbol{h}\|^2 + \Psi(\boldsymbol{h}).$$

1. **Find a tractable surrogate for F_n**

$\rightsquigarrow$ Quadratic tangent majorant

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \Theta_n(\boldsymbol{h}, \mathbf{h}_n) = \mathrm{F}_n(\mathbf{h}_n) + \nabla \mathrm{F}_n(\mathbf{h}_n)^\top (\boldsymbol{h} - \mathbf{h}_n) + \frac{1}{2}(\boldsymbol{h} - \mathbf{h}_n)^\top \mathbf{A}_n(\mathbf{h}_n)(\boldsymbol{h} - \mathbf{h}_n),$$

where $\mathbf{A}_n(\boldsymbol{h}) = \mathbf{R}_n + \boldsymbol{V}_0 + \boldsymbol{V}^\top \mathrm{Diag}\left(\boldsymbol{b}(\boldsymbol{h})\right) \boldsymbol{V} \in \mathbb{R}^{N \times N}$.

Stochastic MM subspace algorithm

Estimate of the objective function at iteration $n \in \mathbb{N}^*$:

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \mathrm{F}_n(\boldsymbol{h}) = \frac{1}{2\overline{\vartheta}_n} \sum_{k=1}^{n} \vartheta^{n-k} \|\mathbf{y}_k - \mathbf{X}_k^\top \boldsymbol{h}\|^2 + \Psi(\boldsymbol{h}).$$

1. **Find a tractable surrogate for F_n**
2. **Minimize in a subspace**

$$(\forall n \in \mathbb{N}^*) \qquad \mathbf{h}_{n+1} \in \underset{\boldsymbol{h} \in \operatorname{span} \mathbf{D}_n}{\operatorname{Argmin}} \; \Theta_n(\boldsymbol{h}, \mathbf{h}_n),$$

with $\mathbf{D}_n \in \mathbb{R}^{N \times M_n}$.

$\rightsquigarrow$ Stochastic 3MG algorithm :

$$\mathbf{D}_n = \begin{cases} [-\nabla \mathrm{F}_n(\mathbf{h}_n), \mathbf{h}_n, \mathbf{h}_n - \mathbf{h}_{n-1}] & \text{if } n > 1 \\ [-\nabla \mathrm{F}_n(\mathbf{h}_1), \mathbf{h}_1] & \text{if } n = 1 \end{cases}$$

Stochastic MM subspace algorithm

Estimate of the objective function at iteration $n \in \mathbb{N}^*$:

$$(\forall \boldsymbol{h} \in \mathbb{R}^N) \quad \mathrm{F}_n(\boldsymbol{h}) = \frac{1}{2\overline{\vartheta}_n} \sum_{k=1}^{n} \vartheta^{n-k} \|\mathbf{y}_k - \mathbf{X}_k^\top \boldsymbol{h}\|^2 + \Psi(\boldsymbol{h}).$$

1. **Find a tractable surrogate for F_n**
2. **Minimize in a subspace**
3. **Perform recursive updates of the second-order statistics**

$$(\forall n \in \mathbb{N}^*) \qquad \mathbf{r}_n = \mathbf{r}_{n-1} + \frac{1}{\overline{\vartheta}_n}(\mathbf{X}_n \mathbf{y}_n - \mathbf{r}_{n-1})$$

$$\mathbf{R}_n = \mathbf{R}_{n-1} + \frac{1}{\overline{\vartheta}_n}(\mathbf{X}_n \mathbf{X}_n^\top - \mathbf{R}_{n-1}).$$

Stochastic MM subspace algorithm

$\mathbf{r}_0 = \mathbf{0}, \mathbf{R}_0 = \mathbf{0}$

Initialize $\mathbf{D}_0, \mathbf{u}_0$

$\mathbf{h}_1 = \mathbf{D}_0 \mathbf{u}_0, \mathbf{D}_0^{\mathbf{R}} = \mathbf{0}, \mathbf{D}_0^{\boldsymbol{V}_0} = \boldsymbol{V}_0 \mathbf{D}_0, \mathbf{D}_0^{\boldsymbol{V}} = \boldsymbol{V} \mathbf{D}_0$

For all $n = 1, \dots$

$$\mathbf{r}_n = \mathbf{r}_{n-1} + \frac{1}{\vartheta_n}(\mathbf{X}_n \mathbf{y}_n - \mathbf{r}_{n-1})$$

$$\mathbf{c}_n(\mathbf{h}_n) = \mathbf{r}_n + \boldsymbol{v}_0 + \boldsymbol{V}^\top \operatorname{Diag}\big(\boldsymbol{b}(\mathbf{h}_n)\big)\boldsymbol{v}$$

$$\mathbf{D}_{n-1}^{\mathbf{A}} = \left(1 - \frac{1}{\vartheta_n}\right)\mathbf{D}_{n-1}^{\mathbf{R}} + \frac{1}{\vartheta_n}\mathbf{X}_n(\mathbf{X}_n^\top \mathbf{D}_{n-1}) + \mathbf{D}_{n-1}^{\boldsymbol{V}_0} + \boldsymbol{V}^\top \operatorname{Diag}\big(\boldsymbol{b}(\mathbf{h}_n)\big)\mathbf{D}_{n-1}^{\boldsymbol{V}}$$

$$\nabla \mathrm{F}_n(\mathbf{h}_n) = \mathbf{D}_{n-1}^{\mathbf{A}} \mathbf{u}_{n-1} - \mathbf{c}_n(\mathbf{h}_n)$$

$$\mathbf{R}_n = \mathbf{R}_{n-1} + \frac{1}{\vartheta_n}(\mathbf{X}_n \mathbf{X}_n^\top - \mathbf{R}_{n-1})$$

Set $\mathbf{D}_n$ using $\nabla \mathrm{F}_n(\mathbf{h}_n)$

$$\mathbf{D}_n^{\mathbf{R}} = \mathbf{R}_n \mathbf{D}_n, \mathbf{D}_n^{\boldsymbol{V}_0} = \boldsymbol{V}_0 \mathbf{D}_n, \mathbf{D}_n^{\boldsymbol{V}} = \boldsymbol{V} \mathbf{D}_n$$

$$\mathbf{B}_n = \mathbf{D}_n^\top (\mathbf{D}_n^{\mathbf{R}} + \mathbf{D}_n^{\boldsymbol{V}_0}) + (\mathbf{D}_n^{\boldsymbol{V}})^\top \operatorname{Diag}\big(\boldsymbol{b}(\mathbf{h}_n)\big)\mathbf{D}_n^{\boldsymbol{V}}$$

$$\mathbf{u}_n = \mathbf{B}_n^\dagger \mathbf{D}_n^\top \mathbf{c}_n(\mathbf{h}_n)$$

$$\mathbf{h}_{n+1} = \mathbf{D}_n \mathbf{u}_n$$

Computational complexity

- If $\boldsymbol{V} = \boldsymbol{I}_N$, number of multiplications per iteration:

$$N^2(4M_n + Q)/2,$$

where N length of $\boldsymbol{h}$, Q column dimension of $\mathbf{X}_n$, M_n subspace dimension, and $N \gg \max\{M_n, M_{n-1}, Q\}$.

Computational complexity

- If $\boldsymbol{V} = \boldsymbol{I}_N$, number of multiplications per iteration:

$$N^2(4M_n + Q)/2,$$

where N length of $\boldsymbol{h}$, Q column dimension of $\mathbf{X}_n$, M_n subspace dimension, and $N \gg \max\{M_n, M_{n-1}, Q\}$.

- If $\boldsymbol{V} \neq \boldsymbol{I}_N$, number of multiplications per iteration:

$$N\Big(\underbrace{P(M_n + M_{n-1} + 1)}_{\substack{\text{upper bound on the complexity}\\ \text{induced by linear operators}}} + N(4M_n + Q)/2\Big),$$

where P line dimension of $\boldsymbol{V}$.

Computational complexity

- If $\boldsymbol{V} = \boldsymbol{I}_N$, number of multiplications per iteration:

$$N^2(4M_n + Q)/2,$$

where N length of $\boldsymbol{h}$, Q column dimension of $\mathbf{X}_n$, M_n subspace dimension, and $N \gg \max\{M_n, M_{n-1}, Q\}$.

- If $\boldsymbol{V} \neq \boldsymbol{I}_N$, number of multiplications per iteration:

$$N\Big(\underbrace{P(M_n + M_{n-1} + 1)}_{\substack{\text{upper bound on the complexity} \\ \text{induced by linear operators}}} + N(4M_n + Q)/2\Big),$$

where P line dimension of $\boldsymbol{V}$.

- Further complexity reduction possible by taking into account the structure of $\mathbf{D}_n$, e.g. 3MG, $\boldsymbol{V} = \boldsymbol{I}_N$, $Q = 1 \equiv$ complexity of RLS.

Convergence results: Assumptions

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be the underlying probability space. For every $n \in \mathbb{N}^*$, let $\mathcal{X}_n = \sigma\big((\mathbf{X}_k, \mathbf{y}_k)_{1 \leqslant k \leqslant n}\big)$ be the sub-sigma algebra of $\mathcal{F}$ generated by $(\mathbf{X}_k, \mathbf{y}_k)_{1 \leqslant k \leqslant n}$.

1. $\boldsymbol{R} + \boldsymbol{V}_0$ is a positive definite matrix.
2. $\big((\mathbf{X}_n, \mathbf{y}_n)\big)_{n \geqslant 1}$ is a stationary ergodic sequence and, for every $n \in \mathbb{N}^*$, the elements of $\mathbf{X}_n$ and the components of $\mathbf{y}_n$ have finite fourth-order moments.
3. $(\forall n \in \mathbb{N}^*)$ $\mathsf{E}(\|\mathbf{y}_{n+1}\|^2 \,|\, \mathcal{X}_n) = \varrho$, $\mathsf{E}(\mathbf{X}_{n+1}\mathbf{y}_{n+1} \,|\, \mathcal{X}_n) = \boldsymbol{r}$ and $\mathsf{E}(\mathbf{X}_{n+1}\mathbf{X}_{n+1}^\top \,|\, \mathcal{X}_n) = \boldsymbol{R}$.
4. For every $n \in \mathbb{N}^*$, $\{\nabla \mathrm{F}_n(\mathbf{h}_n), \mathbf{h}_n\} \subset \operatorname{span} \mathbf{D}_n$.
5. $\mathbf{h}_1$ is $\mathcal{X}_1$-measurable and, for every $n \in \mathbb{N}^*$, $\mathbf{D}_n$ is $\mathcal{X}_n$-measurable.

Convergence results: Theorem

- $(\rho_n)_{n\geqslant 1}$, $(\mathbf{R}_n)_{n\geqslant 1}$, and $(\mathbf{r}_n)_{n\geqslant 1}$ converge almost surely to ϱ, $\boldsymbol{R}$ and $\boldsymbol{r}$, respectively.
- The set of cluster points of $(\mathbf{h}_n)_{n\geqslant 1}$ is almost surely a nonempty compact connected set.
- Any element of this set is almost surely a critical point of F.
- If the functions $(\psi_s)_{1\leqslant s\leqslant S}$ are convex, then the sequence $(\mathbf{h}_n)_{n\geqslant 1}$ converges P-a.s. to the unique (global) minimizer of F.

Convergence results : Rate analysis

Assume that Ψ is twice differentiable.
Let $\epsilon \in]0, +\infty[$ be such that $\epsilon \boldsymbol{I}_N \prec \boldsymbol{R} + \boldsymbol{V}_0$.

- There exists $n_\epsilon \in \mathbb{N}^*$ such that, for every $n \geqslant n_\epsilon$,

$$\mathrm{F}_n(\mathbf{h}_{n+1}) - \inf \mathrm{F}_n \leqslant \theta_n \big(\mathrm{F}_n(\mathbf{h}_n) - \inf \mathrm{F}_n\big)$$

where $\theta_n = 1 - (1+\epsilon)^{-1}\widetilde{\theta}_n$, with:

$$\widetilde{\theta}_n = \frac{\big(\nabla \mathrm{F}_n(\mathbf{h}_n)\big)^\top \mathbf{D}_n(\mathbf{D}_n^\top \mathbf{A}_n(\mathbf{h}_n)\mathbf{D}_n)^\dagger \mathbf{D}_n^\top \nabla \mathrm{F}_n(\mathbf{h}_n)}{\big(\nabla \mathrm{F}_n(\mathbf{h}_n)\big)^\top \big(\nabla^2 \mathrm{F}_n(\mathbf{h}_n)\big)^{-1} \nabla \mathrm{F}_n(\mathbf{h}_n)},$$

Let $\underline{\sigma}_n$ (resp. $\overline{\sigma}_n$) denote the minimum (resp. maximum) eigenvalue of the Hessian of F_n at $\mathbf{h}_n$.

- $\theta_n \in [\underline{\theta}_n, \overline{\theta}_n]$ with

$$\begin{cases} \underline{\theta}_n = 1 - (1+\epsilon)^{-1} ||| \big(\mathbf{A}_n(\mathbf{h}_n)\big)^{-\frac{1}{2}} \big(\nabla^2 \mathrm{F}_n(\mathbf{h}_n)\big)^{\frac{1}{2}} |||^2 > 0, \\ \overline{\theta}_n = \frac{1}{1+\epsilon} \left(\left(\frac{\overline{\sigma}_n - \underline{\sigma}_n}{\overline{\sigma}_n + \underline{\sigma}_n} \right)^2 + \epsilon \right) < 1. \end{cases}$$

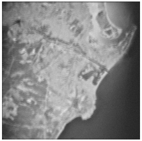
Application to filter identification problems

Application to 2D filter identification

OBSERVATION MODEL

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

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

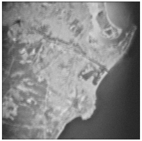


Application to 2D filter identification

OBSERVATION MODEL

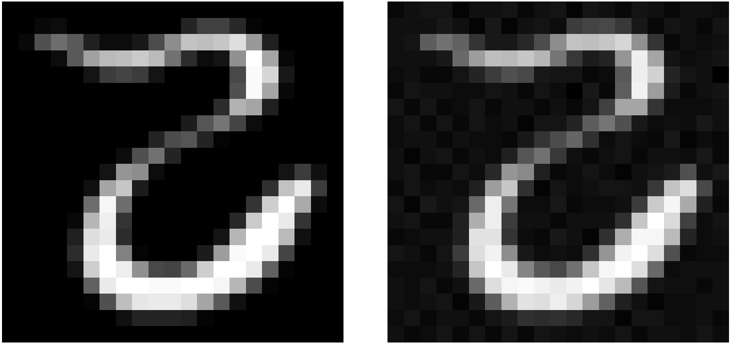
$$\boldsymbol{y} = S(\overline{\boldsymbol{h}})\boldsymbol{x} + \boldsymbol{w}$$

- $\boldsymbol{x} \in \mathbb{R}^L$ large size original image ($L = 4096^2$),
- $\overline{\boldsymbol{h}} \in \mathbb{R}^N$ unknown two-dimensional blur kernel ($N = 21^2$),
- $S(\overline{\boldsymbol{h}})$ Hankel-block Hankel matrix such that $S(\overline{\boldsymbol{h}})\boldsymbol{x} = \boldsymbol{X}\overline{\boldsymbol{h}}$,
- $\boldsymbol{w} \in \mathbb{R}^L$ 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:** $\mathbf{y}_n \in \mathbb{R}^Q$ and $\mathbf{X}_n^\top \in \mathbb{R}^{Q\times N}$: Q lines of $\boldsymbol{y}$ and $\boldsymbol{X}$, $\vartheta = 1$, and Ψ isotropic penalization on the gradient of $\mathbf{h}$, i.e. $S = N$, $(\forall s \in \{1,\dots,S\})$ $P_s = 2$, $\psi_s : u \mapsto \lambda\sqrt{1 + u^2/\delta^2}$, $(\lambda, \delta) \in]0, +\infty[^2$.

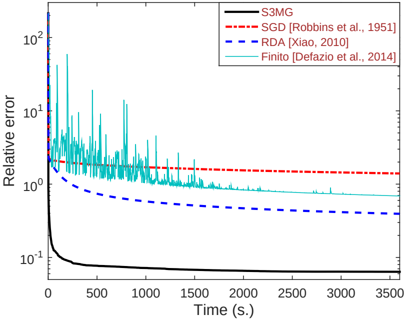
Application to 2D filter identification



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

- The regularization parameters are optimized manually.

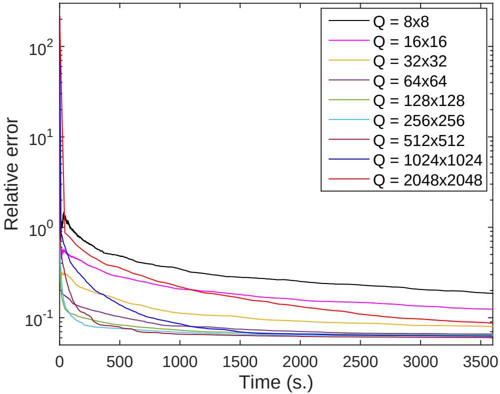
Application to 2D filter identification



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

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

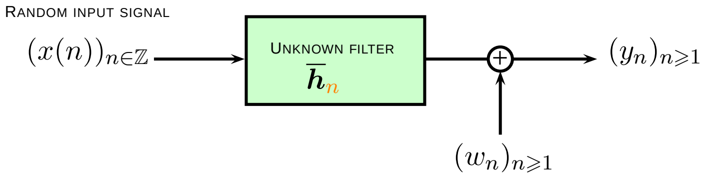
Application to 2D filter identification



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

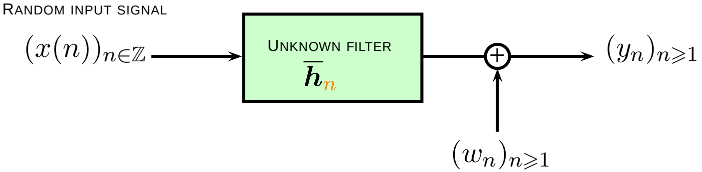
- The best trade-off is obtained for $Q = 256 \times 256$.

Application to sparse adaptive filtering



$(\forall n \in \mathbb{Z}) \quad \overline{\boldsymbol{h}}_n \in \mathbb{R}^N$ **sparse** filter, $y_n \in \mathbb{R}$, $w_n \in \mathbb{R}$.

Application to sparse adaptive filtering



$(\forall n \in \mathbb{Z}) \quad \overline{\boldsymbol{h}}_n \in \mathbb{R}^N$ **sparse** filter, $y_n \in \mathbb{R}$, $w_n \in \mathbb{R}$.

$\Rightarrow$ **Minimization of a penalized MSE criterion:**

$\rightsquigarrow$ $\boldsymbol{X}_n = [x(n-N+1),\dots,x(n)]^\top \in \mathbb{R}^N$;

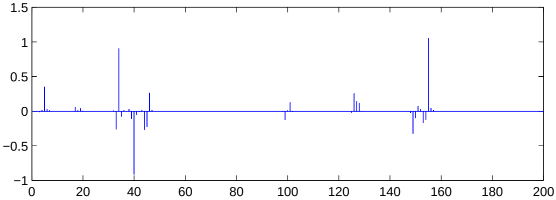
$\rightsquigarrow$ Smoothed ℓ_0 regularization with $S=N$, $\boldsymbol{v}_0 = \boldsymbol{0}$, $\boldsymbol{V}_0 = \boldsymbol{O}_N$, and $(\forall s \in \{1,\dots,N\})$ $P_s = 1$, $\boldsymbol{v}_s = 0$, $\boldsymbol{V}_s \in \mathbb{R}^{1\times N}$ the s-th vector of the canonical basis of $\mathbb{R}^N$ and $\psi_s : u \mapsto \lambda(1-\exp(-u^2/(2\delta^2)))$.

Simulation results

$(\overline{\boldsymbol{h}}_n)$: Time-variant linear system with 200 sparse coefficients,

$(x(n))_n$: Input sequence of 5000 random independent variables uniformly distributed on $\{-1, +1\}$,

$(w_n)_n$: White Gaussian noise with zero mean and variance 0.05.



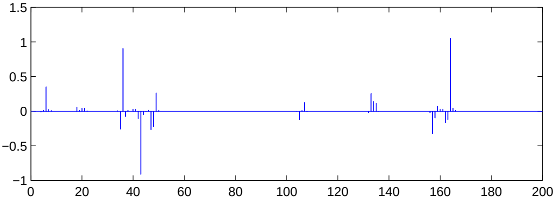
Values of the coefficients of the true sparse filter $\overline{\boldsymbol{h}}_n$ for $1 \leqslant n \leqslant 2500$

Simulation results

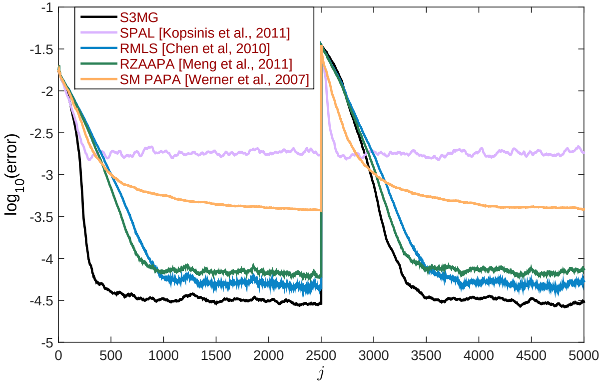
$(\overline{\boldsymbol{h}}_n)$: Time-variant linear system with 200 sparse coefficients,

$(x(n))_n$: Input sequence of 5000 random independent variables uniformly distributed on $\{-1, +1\}$,

$(w_n)_n$: White Gaussian noise with zero mean and variance 0.05.



Values of the coefficients of the true sparse filter $\overline{\boldsymbol{h}}_n$ for $2501 \leqslant n \leqslant 5000$



Estimation error along time, for various sparse adaptive filtering strategies

- For each tested methods, tuning parameters optimized manually.
- S3MG leads to the minimal estimation error, and benefits from good tracking properties.

Conclusion

Solution to online penalized MSE problems?

$\Rightarrow$ Proposition of a novel stochastic MM subspace algorithm.

- ✓ No need to tune up a stepsize (in particular, no decreasing condition on a stepsize);
- ✓ Can be used in an adaptive context thanks to a forgetting factor ;
- ✓ Proven guarantees of convergence ;
- ✓ Analysis of the convergence rate and of the complexity ;
- ✓ Good numerical performance w.r.t. state-of-the art methods.

Some references

E. Chouzenoux and J.-C. Pesquet.
A Stochastic Majorize-Minimize Subspace Algorithm for Online Penalized Least Squares Estimation.
Technical report, 2015. Available at http://arxiv.org/abs/1512.08722.

M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero and S. McLaughlin.
A Survey of Stochastic Simulation and Optimization Methods in Signal Processing.
IEEE Journal of Selected Topics in Signal Processing, Vol. 10, No. 2, pp. 224-241, March 2016.

E. Chouzenoux, A. Jezierska, J.-C. Pesquet and H. Talbot.
A Majorize-Minimize Subspace Approach for $\ell_2 - \ell_0$ *Image Regularization*.
SIAM Journal on Imaging Science, Vol. 6, No. 1, pp. 563-591, 2013.

E. Chouzenoux, J. Idier and S. Moussaoui.
A Majorize-Minimize Strategy for Subspace Optimization Applied to Image Restoration.
IEEE Transactions on Image Processing, Vol. 20, No. 18, pp. 1517-1528, June 2011.

Some references

E. Chouzenoux and J.-C. Pesquet.
A Stochastic Majorize-Minimize Subspace Algorithm for Online Penalized Least Squares Estimation.
Technical report, 2015. Available at http://arxiv.org/abs/1512.08722.

M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero and S. McLaughlin.
A Survey of Stochastic Simulation and Optimization Methods in Signal Processing.
IEEE Journal of Selected Topics in Signal Processing, Vol. 10, No. 2, pp. 224-241, March 2016.

E. Chouzenoux, A. Jezierska, J.-C. Pesquet and H. Talbot.
A Majorize-Minimize Subspace Approach for $\ell_2 - \ell_0$ *Image Regularization*.
SIAM Journal on Imaging Science, Vol. 6, No. 1, pp. 563-591, 2013.

E. Chouzenoux, J. Idier and S. Moussaoui.
A Majorize-Minimize Strategy for Subspace Optimization Applied to Image Restoration.
IEEE Transactions on Image Processing, Vol. 20, No. 18, pp. 1517-1528, June 2011.

Thank you !