Variance Reduced Majorize Minimization algorithms for large scale learning

Gersende Fort
(CNRS, Institut de Mathématiques de Toulouse, France)

Joint work with Eric Moulines (CMAP, Ecole Polytechnique, France) and Hoi-To Wai (Chinese Univ. of Hong Kong, Hong-Kong)

and on going discussions with Florence Forbes (INRIA, France) and Hien Duy Nguyen (Univ. Queensland, Australia)

SMOR Seminar, Univ. of Queensland

In this talk

Motivated by the Large scale Learning setting,

$$\mathrm{argmin}_\theta \mathbb{E}_\pi \left[\ell(X,\theta)\right] \text{ from } (X_i)_i \sim \pi \qquad \mathrm{argmin}_\theta \frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta)$$

solved by a Majorize-Minimization (MM) algorithm

- Part 1. Is it tractable ? no.
- Part 2. Identify the limiting points of MM.
- Part 3. Design a *stochastic* optimization algorithm with the same limiting points: it combines
    - the Stochastic Approximation method Robbins and Monro (1951); book by Benveniste et al. (1990)
      $$\widehat{S}_{n+1} = \widehat{S}_n + \gamma_{n+1} H_{n+1} \qquad H_{n+1}$$
    - a variance reduction technique for the random approximation H_{n+1}.
- Part 4. Explicit bounds of convergence for the `SPIDER MM`.
- Part 5. Numerical illustrations.

I. The Majorize Minimization algorithm in the large scale learning setting

The large scale learning setting

$$\mathrm{argmin}_{\theta \in \mathbb{R}^d} \; (F(\theta) + g(\theta))$$ $g(\theta)$: exact

- "Large batch" learning

$$F(\theta) \stackrel{\mathrm{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, \theta)$$ size n, prohibitive

- Online learning

$$F(\theta) \stackrel{\mathrm{def}}{=} \mathbb{E}_\pi \left[\ell(X, \theta) \right]$$ from a stream of observations $X_i \stackrel{i.i.d.}{\sim} \pi$

The MM algorithm book by K. Lange (2016)

- An iterative algorithm
- Repeat:

$$\theta_t \to \text{majorizing fct} \to \theta_{t+1} \to \text{majorizing fct} \to \cdots$$

- Given θ_t, the majorizing function satisfies

majorize $F(\cdot) \leq G(\cdot;\theta_t)$ $\qquad$ $F(\cdot) + g(\cdot) \leq G(\cdot;\theta_t) + g(\cdot)$

tangent $F(\theta_t) = G(\theta_t;\theta_t)$

- From the majorizing function,

$$\theta_{t+1} \stackrel{\text{def}}{=} \operatorname{argmin}_\theta \left(G(\theta;\theta_t) + g(\theta)\right)$$

- A **descent** property:

$$F(\theta_{t+1})+g(\theta_{t+1}) \leq G(\theta_{t+1};\theta_t)+g(\theta_{t+1}) \leq G(\theta_t;\theta_t)+g(\theta_t) = F(\theta_t)+g(\theta_t)$$

Intractability ?

$$\text{argmin}_\theta \quad \mathsf{MajorizingFct}(\theta;\theta_t)$$

- [Considered here] The *explicit* expression of the majorizing function

$$F(\theta)+g(\theta)=n^{-1}\sum_{i=1}^n \ell(X_i;\theta)+g(\theta)\Longrightarrow \theta\mapsto n^{-1}\sum_{i=1}^n (\quad)+g(\theta)$$
$$F(\theta)+g(\theta)=\mathbb{E}_\pi\left[\ell(X;\theta)\right]+g(\theta)\Longrightarrow \theta\mapsto \mathbb{E}_\pi\left[\quad\right]+g(\theta)$$

- [assumed explicit, here] The optimization step

$$\text{argmin}_\theta \ \mathsf{MajorizingFct}(\theta;\theta_t)$$

Example of MM: EM Dempster et al (1977); book by G. McLachlan and T. Krishnan (2007)

- Inference in latent variable models

$$\ell(X;\theta) \stackrel{\text{def}}{=} -\log \int_{\mathcal{H}} p(X,h;\theta)\, \mathsf{d}\mu(h)$$

- The construction of the majorizing function at the point θ_t:

Example of MM: EM Dempster et al (1977); book by G. McLachlan and T. Krishnan (2007)

- Inference in latent variable models

$$\ell(X;\theta) \stackrel{\text{def}}{=} -\log \int_{\mathcal{H}} p(X,h;\theta)\, \mathsf{d}\mu(h)$$

- The construction of the majorizing function at the point θ_t:

$$\begin{aligned}
\ell(X;\theta) - \ell(X;\theta_t) &= -\log\left(\frac{\int_{\mathcal{H}} p(X,h;\theta)\, \mathsf{d}\mu(h)}{\int_{\mathcal{H}} p(X,h';\theta_t)\, \mathsf{d}\mu(h')}\right) \\
&= -\log\left(\int_{\mathcal{H}} \frac{p(X,h;\theta)}{\int_{\mathcal{H}} p(X,h';\theta_t)\, \mathsf{d}\mu(h')}\, \mathsf{d}\mu(h)\right) \\
&= -\log\left(\int_{\mathcal{H}} \frac{p(X,h;\theta)}{p(X,h;\theta_t)}\; \frac{p(X,h;\theta_t)}{\int_{\mathcal{H}} p(X,h';\theta_t)\, \mathsf{d}\mu(h')}\, \mathsf{d}\mu(h)\right) \\
&\leq -\int_{\mathcal{H}} \log p(X,h;\theta)\; \pi_{\theta_t}(h|X)\, \mathsf{d}\mu(h) + C_t
\end{aligned}$$

Example of MM: EM Dempster et al (1977); book by G. McLachlan and T. Krishnan (2007)

- Inference in latent variable models

$$\ell(X;\theta) \stackrel{\text{def}}{=} -\log \int_{\mathcal{H}} p(X,h;\theta)\, \mathsf{d}\mu(h)$$

- The construction of the majorizing function at the point θ_t:

large batch

$$\theta \mapsto -\frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{H}} \log p(X_i,h;\theta)\ \pi_{\theta_t}(h|X_i)\, \mathsf{d}\mu(h) + g(\theta) + C_t$$

online learning

$$\theta \mapsto \mathbb{E}_{\pi}\left[-\int_{\mathcal{H}} \log p(X,h;\theta)\ \pi_{\theta_t}(h|X)\, \mathsf{d}\mu(h)\right] + g(\theta) + C_t$$

- Intractability: outer sum, inner sum

Example of MM: EM for curved exponential family

- A frequent assumption:

$$\log p(X, h; \theta) = \langle S(X, h), \phi(\theta) \rangle - \psi(\theta)$$

- The majorizing function under this assumption

$$\theta \mapsto g(\theta) + \psi(\theta) - \left\langle \mathbb{E}_\pi \left[\int_{\mathcal{H}} S(X, h) \; \pi_{\theta_t}(h|X) \, \mathsf{d}\mu(h) \right], \phi(\theta) \right\rangle + C_t$$

  in the parametric functional family

$$\theta \mapsto R(\theta) - \langle s, \phi(\theta) \rangle$$

- Under this assumption, the E-step $\equiv$ compute the parameter $''s''$, defined as expectations (outer, inner).

EM, seen in the **surrogate**-space (s-space):
iterative construction of fcts through iterative construction of a parameter "s"

Other examples of MM algorithms

- F is L-smooth $\to$ quadratic surrogate of F $\to$ gradient-type algorithm.

$$\mathrm{argmin}_\theta a + \langle \nabla \xi(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\gamma} \|\theta - \theta_t\|^2 = \theta_t - \gamma \nabla \xi(\theta_t)$$

- Difference of convex functions $\to$ linear surrogate of a concave function

- $\ell(X, \theta) = \inf_h \ell(X, h; \theta)$ $\to$ variational surrogates

In many examples, and assumed HEREAFTER

(large batch) $\quad \text{MajorizingFct}(\theta; \theta_t) = C_t + R(\theta) - \left\langle \frac{1}{n} \sum_{i=1}^n \bar{S}(X_i; \theta_t), \phi(\theta) \right\rangle$

(online) $\quad \text{MajorizingFct}(\theta; \theta_t) = C_t + R(\theta) - \left\langle \mathbb{E}_\pi \left[\bar{S}(X; \theta_t) \right], \phi(\theta) \right\rangle$

Conclusion of Part I.

- MM defines a sequence of surrogate functions $\rightarrow$ MM defines a sequence of parameters "s"

$$\theta \mapsto R(\theta) - \langle s, \phi(\theta) \rangle$$

- In large scale learning: the exact value of "s" is intractable.

Solution ?

- Identify the limiting points of the (exact) MM
- Design a *stochastic* algorithm having the same limiting points.

II. The limiting points of MM

Assumptions

We consider MM algorithms having

- a surrogate function in the family indexed by s:

$$\theta \mapsto R(\theta) - \langle s, \phi(\theta) \rangle$$

At iteration $\#t$ let us write it in the "online learning setting"

$$\theta \mapsto R(\theta) - \left\langle \mathbb{E}_\pi \left[\bar{S}(X, \theta_t) \right], \phi(\theta) \right\rangle$$

- an explicit optimization of this surrogate

$$\mathsf{T}(s) \stackrel{\text{def}}{=} \operatorname{argmin}_\theta \left(R(\theta) - \langle s, \phi(\theta) \rangle \right)$$

Case of EM
Hyp 1: OK when the complete data likelihood is from *the curved exponential family.*
Hyp 2: for convenience.

Fixed points in the surrogate space

$$s_\star : \qquad s_\star = \mathbb{E}_\pi\left[\bar{S}(X, \mathsf{T}(s_\star))\right]$$

MM finds the roots of

$$s \mapsto \mathsf{h}(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi\left[\bar{S}(X, \mathsf{T}(s)) - s\right].$$

The outer expectation is intractable.

Conclusion of Part II.

- Forget the MM scheme
- Keep in mind: algorithm to find the roots of

$$s \mapsto \mathsf{h}(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\bar{S}(X, \mathsf{T}(s)) - s \right].$$

- Replace exact MM with: a **Stochastic Approximation** algorithm designed to find the roots of the *mean field* h.

III. Variance reduction within Stochastic Approximation

Stochastic Approximation algorithms

- Mean field:

$$\mathsf{h}(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\bar{S}(X, \mathsf{T}(s)) - s\right]$$

- Iterative scheme:

$$\widehat{S}_{t+1} = \widehat{S}_t + \gamma_{t+1} \left(\frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \bar{S}(X_i, \mathsf{T}(\widehat{S}_t)) - \widehat{S}_t \right)$$

where $\mathcal{B}_{t+1}$ is a mini-batch of examples of size b

- (large batch) sampled with replacement; b $<< n$
- (online) from the data stream

Review of MM for large scale learning / EM context (1/2)

- **Online-EM.**

  Neal and Hinton, 1998; Cappé and Moulines, 2009); Nguyen et al (2020); Karimi et al (2019a, 2019b).

- (large batch) **iEM.** Incremental EM

  Case $\gamma_t = 1$. Neal and Hinton (1998); Ng and McLachlan (2003); Gunawardana and Byrne (2005); Karimi et al (2019c)

Based on an incremental approximation of

$$\mathsf{h}(\widehat{S}_t) = n^{-1} \sum_{i=1}^{n} S(X_i, \mathsf{T}(\widehat{S}_t)) - \widehat{S}_t.$$

$*$ Init: store for all i, $\sigma_i \stackrel{\text{def}}{=} S(X_i, \mathsf{T}(\widehat{S}_0))$ and compute $\mathsf{h}(\widehat{S}_0)$.
$*$ At iter $\#(t+1)$:
  sample an index I;
  update $\sigma_I \leftarrow S(X_I, \mathsf{T}(\widehat{S}_t))$;
  update the term $\#I$ in the approximation of $\mathsf{h}(\widehat{S}_t)$.

Stochastic Approximation algorithms with Variance Reduction

Variance reduction through control variates

$$\widehat{S}_{t+1} = \widehat{S}_t + \gamma_{t+1} \left(\frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \bar{S}(X_i, \mathsf{T}(\widehat{S}_t)) - \widehat{S}_t + V_{t+1} \right)$$

- V_{t+1} is centered $\rightarrow$ the mean field is not modified.
- V_{t+1} and $\frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \bar{S}(X_i, \mathsf{T}(\widehat{S}_t)) - \widehat{S}_t$ are correlated.

Review of MM for large scale learning / EM context (2/2)

- (large batch) **sEM-vr.** Stochastic EM with Variance Reduction

  Chen et al, 2018. Parallel with "SVRG" by Johnson and Zhang (2013)

- (large batch) **FIEM.** Fast Incremental EM

  Karimi et al, 2019; Fort et al, 2021. Parallel with "SAGA" by Defazio et al (2014).

  The control variate V_{t+1} is defined as in iEM:

  $\ast$ Init: store the σ_i's

  $\ast$ At iter $\#(t+1)$

  sample two indices I, J.

  Update σ_I and the sum $n^{-1} \sum_{i=1} \sigma_i$ by modifying the term $\#I$

  Correlate V_{t+1} to the natural field $S(X_J, \mathsf{T}(\widehat{S}_t)) - \widehat{S}_t$:

$$V_{t+1} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \sigma_i - \sigma_J .$$

(large batch) A novel variance-reduction: SPIDER MM Fort, Moulines, Wai - NeurIPS 2020

Stochastic Path Integrated Differential EstimatoR MM

adapted from: Nguyen et al. (2017), Fang et al. (2018), Wang et al. (2019)

$$V_{t+1} = V_t + \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_t} \bar{S}(X_i, \mathsf{T}(\widehat{S}_{t-1})) - \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t+1}} \bar{S}(X_i, \mathsf{T}(\widehat{S}_{t-1}))$$

- learn zero through an approximation of "$\mathsf{h}(\widehat{S}_{t-1}) - \mathsf{h}(\widehat{S}_{t-1})$".
- correlated to the natural field through $\mathcal{B}_{t+1}$.
- biased ! refresh the control variates regularly.

SPIDER-MM (Stochastic Path Integrated Differential EstimatoR MM)

1: $\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\mathrm{init}} \qquad V_{1,0} = 0 \qquad \mathcal{B}_{1,0} = \{1, \cdots, n\}$
2: **for** $t = 1, \cdots, k_{\mathrm{out}}$ **do**
3: **for** $k = 0, \ldots, \xi_t - 1$ **do**
4: Sample a mini batch $\mathcal{B}_{t,k+1}$ of size b from $\{1, \cdots, n\}$
5: $V_{t,k+1} = V_{t,k} + \mathsf{b}^{-1} \left(\sum_{i \in \mathcal{B}_{t,k}} \bar{S}(X_i, \mathsf{T}(\widehat{S}_{t,k-1})) - \sum_{i \in \mathcal{B}_{t,k+1}} \bar{S}(X_i, \mathsf{T}(\widehat{S}_{t,k-1})) \right)$
6: $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1} \left(\mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \bar{S}(X_i, \mathsf{T}(\widehat{S}_{t,k})) - \widehat{S}_{t,k} + V_{t,k+1} \right)$
7: **end for**
8: $\widehat{S}_{t+1,-1} = \widehat{S}_{t,\xi_t}$
9: $V_{t+1,0} = 0 \qquad \mathcal{B}_{t+1,0} = \{1, \cdots, n\}$
10: $\widehat{S}_{t+1,0} = \widehat{S}_{t+1,-1} + \gamma_{t+1,0} \left(\mathsf{h}(\widehat{S}_{t+1,-1}) + V_{t+1,0} \right)$
11: **end for**

- k_{out} outer loops, the outer $\#t$ is of length ξ_t
- The control variate is refreshed at each *outer loop* $\#t$ (see Line 9)
- A full scan of the examples at each *outer loop* (see Line 9).

Extensions

- The length of the outer loop is a Geometric random variable with expectation ξ_t. Fort, Moulines, Wai - ICASSP 2021
- Avoid the full scan of the examples when starting each outer loop $\rightarrow$ reduction of the computational cost. Fort, Moulines, Wai - ICASSP 2021
- An approximation of $\bar{S}(X_i, \theta)$ Fort, Moulines - SSP 2021

  for example: in EM, $\bar{S}(X_i, \theta)$ is an expectation w.r.t. the a posteriori distribution of the latent variables $\rightarrow$ Monte Carlo approximation.
- A Proximal operator for constrained optimization Fort, Moulines - SSP 2021

$$\widehat{S}_{t,k+1} = \mathrm{Prox}_{\gamma_{t,k+1}\,g}\left(\widehat{S}_{t,k} + \gamma_{t,k+1} H_{t,k+1}\right)$$

  for example: find the roots of h in a compact set.

IV. Convergence analysis of SPIDER MM

Assumptions

1. There exists a continuously differentiable function $W : \mathbb{R}^q \to \mathbb{R}$ such that

$$\nabla W(s) \stackrel{\text{def}}{=} -B(s)\,\mathsf{h}(s) \qquad \mathsf{h}(s) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \bar{S}(X_i, \mathsf{T}(s)) - s$$

   where $B(s)$ is a $q \times q$ positive definite matrix.
   In addition, ∇W is globally Lipschitz with constant $L_{\dot{W}}$,
   and there exist $0 < v_{\min} \leq v_{\max}$ such that the spectrum of $B(s)$ is in $[v_{\min}, v_{\max}]$.

2. For any $i \in \{1, \cdots, n\}$, the function $s \mapsto \bar{S}(X_i, \mathsf{T}(s)) - s$ is globally Lipschitz with constant L_i.

What kind of convergence results ?

- The objective fct: non necessarily convex but $\mathsf{T}(s)$ exists, unique.
- Explicit control of *errors* given a fixed nbr of observations (given a "budget").
- What is "errors"

$$\mathbb{E}\left[\|\mathsf{h}(\widehat{S}_t)\|^2\right]$$

- At time t ? no $\cdots$ at some random time τ ! non convex optim.
- What do we learn from an explicit control ? how *design parameters* scale with n, in order to reach an accuracy ϵ

$$\mathbb{E}\left[\|\mathsf{h}(\widehat{S}_\tau)\|^2\right] \leq \epsilon$$

Convergence in expectation, explicit h_i's

Under the previous assumptions:

(Fort, Moulines, Wai, NeurIPS 2020)

Set $L^2 \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^n L_i^2$. *Fix* $k_{\text{out}}, k_{\text{in}}, \mathsf{b} \in \mathbb{N}_\star$. *Choose* $\alpha \in (0, v_{\min}/\mu_\star(k_{\text{in}}, \mathsf{b}))$ *with*

$$\mu_\star(k_{\text{in}}, \mathsf{b}) \stackrel{\text{def}}{=} v_{\max} \frac{\sqrt{k_{\text{in}}}}{\sqrt{\mathsf{b}}} + \frac{L_{\dot{W}}}{2L}.$$

Run the algorithm with $\xi_t = k_{\text{in}}$ *and* $\gamma_{t,k} \stackrel{\text{def}}{=} \alpha/L$. *Then*

$$\mathbb{E}\left[\|\mathsf{h}\left(\widehat{S}_{\tau,\xi-1}\right)\|^2\right] \leq \left(\frac{1}{k_{\text{in}}} + \frac{\alpha^2}{\mathsf{b}}\right) \frac{1}{k_{\text{out}}} \frac{2L}{\alpha\{v_{\min} - \alpha\mu_\star(k_{\text{in}}, \mathsf{b})\}} \left(\mathbb{E}\left[W(\widehat{S}_{\text{init}})\right] - \min W\right)$$

where (τ, ξ) *is a uniform r.v. on* $\{1, \cdots, k_{\text{out}}\} \times \{0, \cdots, k_{\text{in}} - 1\}$ *indep of* $\{\widehat{S}_{t,k}\}$.

Complexity for ϵ-approximate stationarity

From this **explicit** expression of an upper bound for

$$\mathbb{E}\left[\|\mathsf{h}(\widehat{S}_{\tau,\xi-1})\|^2\right]$$

- in the `non convex` setting
- with a `random stopping rule`
- as a function of $k_{\text{out}}, k_{\text{in}}, \mathsf{b}, n$ and the learning rate γ ($= \gamma_{t,k}$ for any $t, k > 0$)

To reach ϵ-stationarity, the complexity of SPIDER-MM

With: $k_{\text{in}} = \mathsf{b} = O(\sqrt{n}), \quad k_{\text{out}} = O(1/(\epsilon k_{\text{in}}))$

Nbr of optimization steps: $O(1/\epsilon)$
Nbr of $\bar{S}(X_i, \theta)$'s evaluations: $\mathcal{K} = O(\sqrt{n}\,\epsilon^{-1}) \rightarrow$ *state of the art !*

Algorithm	Complexity $\mathcal{K}$
Online-MM	ϵ^{-2}
iMM	$n\,\epsilon^{-1}$
sMM-vr	$n^{2/3}\,\epsilon^{-1}$
FIMM	$n^{2/3}\,\epsilon^{-1} \wedge \sqrt{n}\,\epsilon^{-3/2}$

Sketch of proof

Inside an outer loop $\#t$, then sum along the inner loops $k=0$ to $k=k_{\text{in}}-1$; then sum along the outer loops $t=1$ to $t=k_{\text{out}}$.

- W is Gradient-Lipschitz, and its gradient is a linear function of h

$$
\begin{aligned}
W(\widehat{S}_{t,k+1}) - W(\widehat{S}_{t,k}) &\leq \left\langle \nabla W(\widehat{S}_{t,k}), \widehat{S}_{t,k+1} - \widehat{S}_{t,k} \right\rangle + \frac{L_{\dot{W}}}{2} \|\widehat{S}_{t,k+1} - \widehat{S}_{t,k}\|^2 \\
&\leq -\gamma_{t,k+1} v_{\min} \|H_{t,k+1}\|^2 + \gamma_{t,k+1} \left(\beta^2 v_{\max} + \gamma_{t,k+1} \frac{L_{\dot{W}}}{2} \right) \|H_{t,k+1}\|^2 \\
&+ \frac{\gamma_{t,k+1}}{\beta^2} v_{\max} \|H_{t,k+1} - \mathsf{h}(\widehat{S}_{t,k})\|^2 \qquad \forall \beta > 0; \text{choice: } \beta^2 \propto \gamma_{t,k+1}
\end{aligned}
$$

- Biased field; full scan when refreshing $\rightarrow$ cancel the bias

$$
\mathbb{E}\left[H_{t,k+1} | \mathcal{F}_{t,k}\right] = \mathsf{h}(\widehat{S}_{t,k}) + H_{t,k} - \mathsf{h}(\widehat{S}_{t,k-1}) \qquad \mathbb{E}\left[H_{t,k+1} | \mathcal{F}_{t,0}\right] = 0.
$$

- L^2-error of the field

$$
\mathbb{E}\left[\|H_{t,k+1} - \mathsf{h}(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,0}\right] = \mathbb{E}\left[\|H_{t,k+1} - \mathbb{E}\left[H_{t,k+1}|\mathcal{F}_{t,k}\right]\|^2 | \mathcal{F}_{t,0}\right] + \mathbb{E}\left[\|\underbrace{\mathbb{E}\left[H_{t,k+1}|\mathcal{F}_{t,k}\right] - \mathsf{h}(\widehat{S}_{t,k})}_{H_{t,k} - \mathsf{h}(\widehat{S}_{t,k-1})}\|^2 | \mathcal{F}_{t,0}\right]
$$

- Variance: specific form of $H_{t,k+1}$ $\rightarrow$ difference of h_i's

$$
H_{t,k+1} - \mathbb{E}\left[H_{t,k+1}|\mathcal{F}_{t,k}\right] = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \{\mathsf{h}_i(\widehat{S}_{t,k}) - \mathsf{h}_i(\widehat{S}_{t,k-1})\} - \frac{1}{n} \sum_{i=1}^{n} \{\mathsf{h}_i(\widehat{S}_{t,k}) - \mathsf{h}_i(\widehat{S}_{t,k-1})\}
$$

use: $\|\mathsf{h}_i(\widehat{S}_{t,k}) - \mathsf{h}_i(\widehat{S}_{t,k-1})\|^2 \leq L_i^2 \|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2 = L_i^2 \gamma_{t,k}^2 \|H_{t,k}\|^2$

Assumptions (case: Monte Carlo approximation of $\bar{S}(X_i, \theta)$'s)

In the case

$$\bar{S}(X_i, \mathsf{T}(\widehat{S}_{t,k})) = \int \mathcal{H}(h)\, \pi_{t,k}(h|X_i) \mathsf{d}\mu(h) \approx \frac{1}{m_{t,k+1}} \sum_{r=1}^{m_{t,k+1}} \mathcal{H}(Z_r^{i,t,k})$$

error

$$\eta_{t,k+1} \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_\bullet} \left(\frac{1}{m_{t,k+1}} \sum_{r=1}^{m_{t,k+1}} \mathcal{H}(Z_r^{i,t,k}) - \bar{S}(X_i, \mathsf{T}(\widehat{S}_{t,k})) \right)$$

3. (bias) there exists $C_b \geq 0$ s.t. for any t, k, with probability one

$$\left\| \mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k} \right] \right\| \leq \frac{C_b}{m_{t,k+1}}$$

4. (variance) there exists C_v s.t. for any t, k with probability one

$$\mathbb{E}\left[\left\| \eta_{t,k+1} - \mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k} \right] \right\|^2 | \mathcal{F}_{t,k} \right] \leq \frac{C_v}{M_{t,k+1}}$$

Examples. i.i.d. case: $C_b = 0$; i.i.d. and MCMC cases: $M_{t,k+1} = \mathsf{b}\, m_{t,k+1}$

Convergence in expectation (i.i.d. case)

Fort, Moulines – SSP 2021; i.i.d. case and MCMC case

Choose $\xi_t = k_{\mathrm{in}}$ *and* $\gamma_{t,k} = \gamma$ *where*

$$\gamma \stackrel{\mathrm{def}}{=} \frac{v_{\min}}{L_{\dot{W}} + 2L v_{\max}\sqrt{k_{\mathrm{in}}}/\sqrt{\mathsf{b}}}$$

Then

$$\gamma v_{\min} \mathbb{E}\left[\frac{\|\widehat{S}_{\tau,\xi} - \widehat{S}_{\tau,\xi-1}\|^2}{\gamma^2}\right] \leq \frac{1}{k_{\mathrm{out}}(1+k_{\mathrm{in}})}\left(W(\widehat{S}_{\mathrm{init}}) - \min W\right) + C_1 \frac{v_{\max}}{L}\frac{1}{\sqrt{k_{\mathrm{in}}\mathsf{b}}}\mathbb{E}\left[\frac{k_{\mathrm{in}} - \xi}{m_{\tau,\xi+1}}\right]$$

where (τ, ξ) *is a uniform r.v. on* $\{1, \cdots, k_{\mathrm{out}}\} \times \{0, \cdots, k_{\mathrm{in}}\}$ *indep of* $\{\widehat{S}_{t,k}\}$.

From

$$\widehat{S}_{t,k+1} - \widehat{S}_{t,k} = \gamma_{t,k+1} H_{t,k+1} \neq \gamma_{t,k+1}\,\mathsf{h}(\widehat{S}_{t,k}),$$

a control is then obtained on $\mathbb{E}\left[\|\mathsf{h}(\widehat{S}_{\tau,\xi})\|^2\right]$

Complexity for ϵ-approximate stationarity

From this **explicit** expression of an upper bound for

$$\mathbb{E}\left[\left\|\mathsf{h}(\widehat{S}_{\tau,\xi-1})\right\|^2\right]$$

- in the `non convex` setting
- with a `random stopping rule`
- as a function of $k_{\mathrm{out}}, k_{\mathrm{in}}, \mathsf{b}, n$ and the learning rate γ
- with a Monte Carlo approximation of the $\bar{S}(X_i, \theta)$'s

To reach ϵ-stationarity, the complexity of Perturbed-SPIDER-MM

With: $k_{\mathrm{in}} = \mathsf{b} = O(\sqrt{n}), \quad k_{\mathrm{out}} = O(1/(\epsilon k_{\mathrm{in}})), \quad m_{t,k} = \epsilon^{-1}$

Nbr of optimization steps: $O(1/\epsilon)$
Nbr of $\bar{S}(X_i, \cdot)$'s evaluations: $\mathcal{K} = O(\sqrt{n}\,\epsilon^{-1}) \rightarrow$ *same as SPIDER-MM*
Nbr of Monte Carlo draws: $O(\sqrt{n}/\epsilon^2)$

V. Numerical illustrations
Herafter, MM means EM

SPIDER-EM: state-of-the-art among the incremental EM algorithms

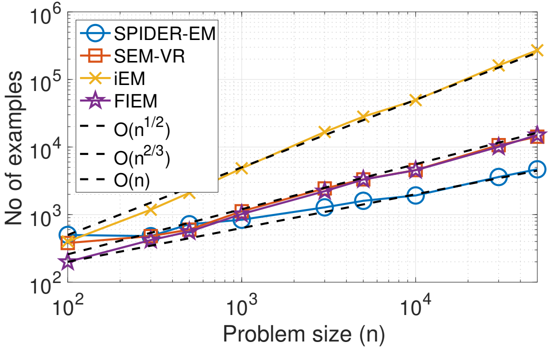

Figure: Nbr of processed examples required to reach convergence, as a function of the problem size n

Estimation of the parameters (1/2)

Case: inference in a mixture of Gaussian distributions (from the MNIST data set). Gaussian mixture models in $\mathbb{R}^{20}$; $G = 12$ components with the same cov matrix; $n = 6\,10^4$ examples

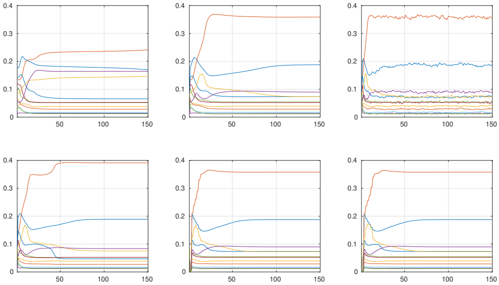

Figure: Evolution of the $L = 12$ iterates $\alpha_k = (\alpha_{k,1}, \ldots, \alpha_{k,L})$ as a function of the number of epochs, for `EM`, `iEM` and `Online EM` on the top from left to right; `FIEM`, `sEM-vr` and `SPIDER-EM` on the bottom from left to right.

Estimation of the parameters (2/2)

Case: inference in a mixture of Gaussian distributions (from the MNIST data set). Gaussian mixture models in $\mathbb{R}^{20}$; $G = 12$ components with the same cov matrix; $n = 6\,10^4$ examples

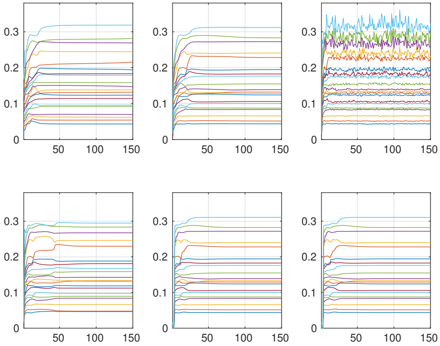

Figure: Evolution of the $p = 20$ eigenvalues of the iterates Σ_k as a function of the number of epochs, for `EM`, `iEM` and `Online EM` on the top from left to right; `FIEM`, `sEM-vr` and `SPIDER-EM` on the bottom from left to right.

Evolution of the objective function

Case: inference in a mixture of Gaussian distributions (from the MNIST data set). Gaussian mixture models in $\mathbb{R}^{20}$; $G = 12$ components with the same cov matrix; $n = 6\,10^4$ examples

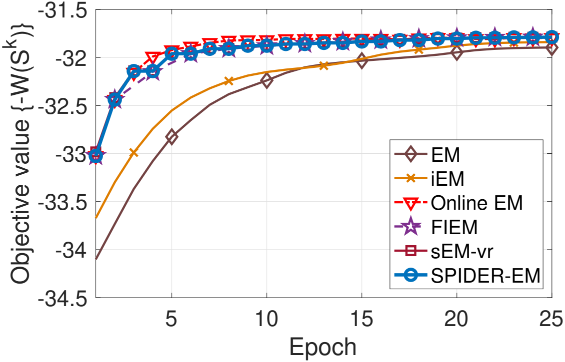

Figure: Evolution of the objective function $-W(\widehat{S}_k)$ vs the number of epochs.

Deterministic or geometric length of the outer loops? Full scan when refreshing ? (1/2)

Case: inference in a mixture of Gaussian distributions (from the MNIST data set). Gaussian mixture models in $\mathbb{R}^{20}$; $G = 12$ components with the same cov matrix; $n = 6\,10^4$ examples

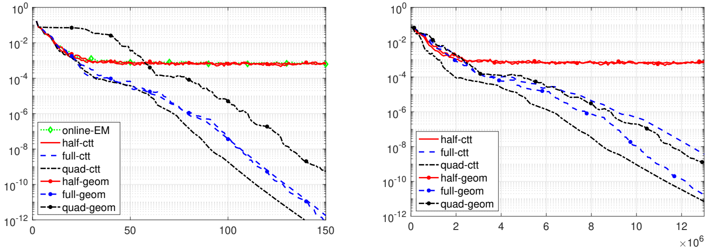

Figure: Quantile of order 0.5 of $\|\mathsf{h}(\widehat{S}_{t,\xi_t})\|^2$ vs the number of epochs (left) and vs the number of $\bar{\mathsf{s}}_i$'s evaluations (right)

Length of each outer loop: either constant (`ctt`) $\xi_t = k_{\text{in}}$, or a geometric r.v. (`geom`) with expectation k_{in}

When refreshing the control variate: use the full data set (`full`), or the half data set (`half`) or a quadratically increasing nbr of examples (`quad`).

Deterministic or geometric length of the inner loops? Full scan when refreshing ? (2/2)

Case: inference in a mixture of Gaussian distributions (from the MNIST data set). Gaussian mixture models in $\mathbb{R}^{20}$; $G = 12$ components with the same cov matrix; $n = 6\,10^4$ examples

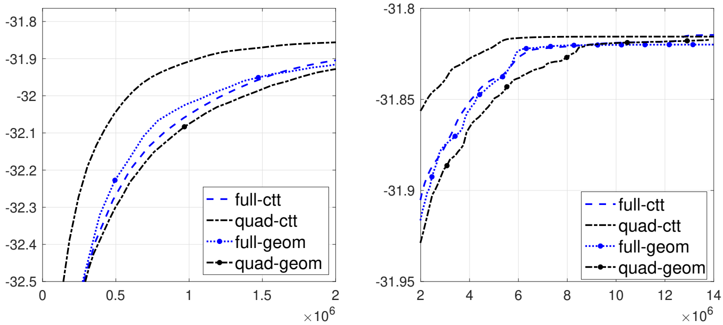

Figure: Evolution of the normalized log-likelihood vs the number of $\bar{s}_i$'s evaluations until $2e6$ (left) and after (right).

Monte Carlo approximations: benefit of variance reduction

Case: Ridge-penalized inference in a logistic regression model (from the MNIST data set). An individual regression vector $Z_i \in \mathbb{R}^{1+50}$ assumed i.i.d. $\mathcal{N}_{51}(\theta, 0.1\, I)$. $n = 24\,989$, 2 classes.

$$\Delta_{t,k+1} \stackrel{\text{def}}{=} \|\widehat{S}_{t,k+1} - \widehat{S}_{t,k}\|^2/\gamma_{t,k+1}^2$$

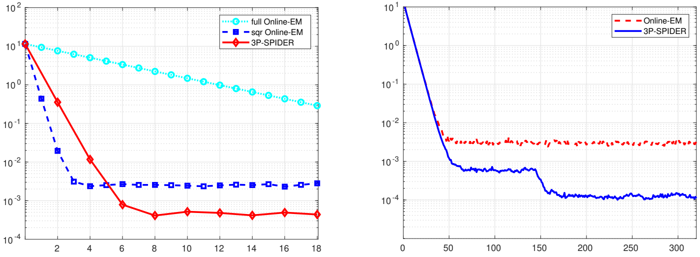

Figure: [left] Monte Carlo estimation of $\mathbb{E}\left[\Delta_{t,k+1}\right]$ vs the number of epochs. Comparison of (Perturbed-Proximal-Preconditioned) 3P-SPIDER-EM and Online-EM when $\mathsf{b} = n$ (case full) and $\mathsf{b} = 10\sqrt{n}$ (case sqr). Monte Carlo approximations with $m_{t,k} = 2\sqrt{n}$. [right] Quantiles 0.75 of $\Delta_{t,k}$ vs the number of epochs, for Online-EM and 3P-SPIDER-EM. For 3P-SPIDER-EM $m_{t,k} = 2\sqrt{n}$ for $t \leq 9$ and $m_{t,k} = 10\sqrt{n}$ for $t \geq 10$.

Monte Carlo approximations: number of points in the Monte Carlo sum

Case: Ridge-penalized inference in a logistic regression model (from the MNIST data set). An individual predictor vector $Z_i \in \mathbb{R}^{1+50}$ assumed i.i.d. $\mathcal{N}_d(\theta, 0.1\, I)$. $n = 24\,989$, 2 classes.

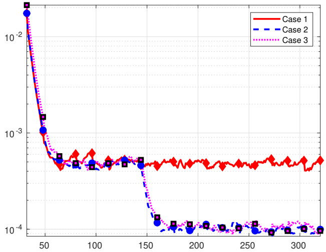

Figure: Monte Carlo estimation of $\mathbb{E}\left[\Delta_{t,k+1}\right]$ vs the number of epochs. `(Perturbed-Proximal-Preconditioned) SPIDER-EM` applied with $\gamma_{t,k} = 0.1$ and $m_{t,k} = 2\sqrt{n}$ in `Case 1`; and with $\gamma_{t,k} = 0.1$ and $m_{t,k} = 2\sqrt{n}$ for $t \leq 10$ and $m_{t,k} = 10\sqrt{n}$ for $t \geq 11$ on `Case 2` and `Case 3`. `Case 2` and `Case 3` differ in the choice of $\gamma_{t,0}$

VI. Bibliography

Results of this talk

- **G. Fort, E. Moulines, H.-T. Wai.** A Stochastic Path Integrated Differential Estimator Expectation Maximization Algorithm. *In Conference Proceedings NeurIPS, 2020.*
- **G. Fort, E. Moulines, H.-T. Wai**. Geom-SPIDER-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization, *ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP):3135–3139.*
- **G. Fort and E. Moulines.** The Perturbed Prox-Preconditioned SPIDER algorithm: non-asymptotic convergence bounds. *Accepted to IEEE Statistical Signal Processing Workshop (SSP 2021).*
- **G. Fort and E. Moulines.** The Perturbed Prox-Preconditioned SPIDER algorithm for EM-based large scale learning. *Accepted to IEEE Statistical Signal Processing Workshop (SSP 2021)*

Other references

- Benveniste, A. and Métivier, M. and Priouret P. Adaptive Algorithms and Stochastic Approxima-tions. Springer Verlag, 1990.
- Cappé, O. and Moulines, E. On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):593–613, 2009.
- Chen, J. Zhu, Y. Teh, and T. Zhang. Stochastic Expectation Maximization with variance reduction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Gar-nett, editors, *Advances in Neural Information Processing Systems 31*, pages 7967–7977. 2018.
- Dempster, A.P. and Laird, N.M. and Rubin, D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977.
- Fang, C. and Li, C. and Lin, Z. and Zhang, T. SPIDER: Near-Optimal Non-Convex Optimization viaStochastic Path-Integrated Differential Estimator. In S.Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 689–699. Curran Associates, Inc., 2018.
- Fort, G. and Gach, P. and Moulines, E. The Fast Incremental Expectation Maximization for finite-sum optimization: asymptotic convergence, *Statistics and Computing*, 2021.
- Karimi, B. and Wai, H.-T., and Moulines, E. and Lavielle, M. On the Global Convergence of (Fast) In-cremental Expectation Maximization Methods. In H. Wallach, H. Larochelle, A. Beygelzimer,F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information ProcessingSystems 32*, pages 2837–2847. Curran Associates, Inc., 2019.
- Neal, R.M. and Hinton, G.E. A View of the EM Algorithm thatJustifies Incremental, Sparse,and other Variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Springer Netherlands, Dordrecht, 1998.
- Nguyen, L.M. and Liu, K. and Scheinberg,K. and Takác M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. *In Proceedings of the 34th International Conference on Machine Learning* - Volume 70, ICML'17, page 2613–2621. 2017
- Robbins, H. and Monro, S.. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*. 22 (3): 400, 1951.
- Wang, Z. and Ji, K. and Zhou, Y. and Liang, Y. and and Tarokh, V. SpiderBoost and Momentum: Faster Stochastic Variance Reduction Algorithms. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2406–2416. 2019.