Stochastic Majorize-Minimization algorithms for large scale learning

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse, France

Journées Statistiques du Sud, 2022.

In collaboration with

- Aymeric Dieuleveut, Ecole Polytechnique, CMAP, France
- Eric Moulines, Ecole Polytechnique, CMAP, France
- Geneviève Robin, CNRS, LaMME, France
- Hoi-To Wai, Chinese Univ. of Hong-Kong, Hong-Kong

Partly funded by

Fondation Simone et Cino Del Duca, Project OpSiMorE

I. Theme and Motivation

General theme: Stochastic Optimization

- Composite objective function

$$\mathrm{Argmin}_{\mathbb{R}^d} \left(f(\theta) + g(\theta) \right) \qquad\qquad f(\theta) := \mathbb{E}_{Z \sim \pi}\left[\ell(Z, \theta) \right]$$

- the function $f : \mathbb{R}^d \to \mathbb{R}$

- *mean value* of a loss over the examples (observations, data)
- $\mathbb{E}_\pi[\cdot]$ **can not** be explicitly evaluated
- possibly **non convex**, continuously differentiable

- the function $g : \mathbb{R}^d \to (0, +\infty]$

- a regularization / penalization term, constraints
- explicit evaluation of $g(\theta)$
- $\Theta := \{\theta \in \mathbb{R}^d : g(\theta) < \infty\}$.
- convex, proper, lower semi-continuous

Computational Statistical Learning

$$\mathrm{Argmin}_{\mathbb{R}^d} \left(f(\theta) + g(\theta) \right) \qquad f(\theta) := \mathbb{E}_{Z \sim \pi} \left[\ell(Z, \theta) \right]$$

- Large batch learning: empirical loss, finite sum setting

$$\pi := \frac{1}{N} \sum_{j=1}^{N} \delta_{Z_j} \qquad f(\theta) = \frac{1}{N} \sum_{j=1}^{N} \ell(Z_j, \theta)$$

- Online learning: expected loss

data stream$\{Z_j, j \geq 0\} \sim \pi$ $\qquad f(\theta) = \int \ell(z, \theta) \pi(\mathrm{d}z)$

- Examples of loss functions

- quadratic $\|Z - \Xi(\theta)\|^2$
- linear regression, quadratic loss $\|Y - X\theta\|^2$, $Z = (Y, X)$
- negative log-likelihood $-\log \mathrm{like}(Z; \theta)$

What we intend to do

Table of contents

- Among the family of **Majorize-Minimization** optimization algorithm

- what is MM ?
- why MM ?

- Stochastic Variance reduction techniques for large scale learning

- why "stochastic" is required ?
- A **mirror** definition of MM and its stochastic version
- Why "variance reduction"
- A novel Variance-Reduced MM

- Federated Learning

- what is FL ?
- Stochastic MM in the FL setting

.

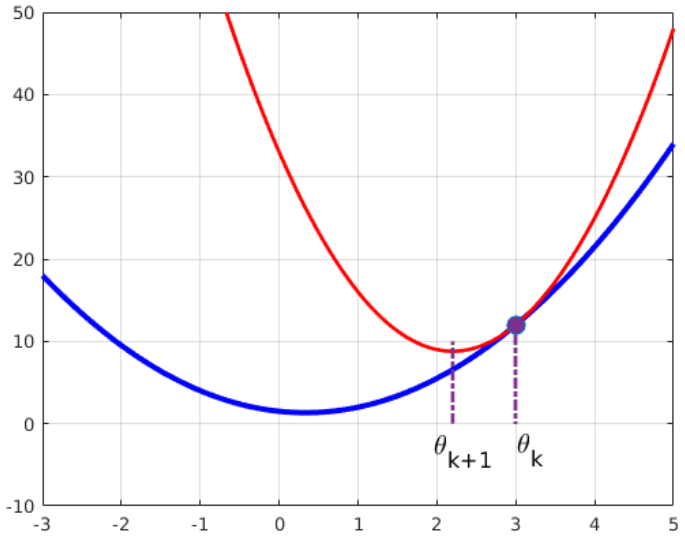
Optimization tool:
Majorize-Minimization

.

See e.g. K. Lange "Optimization", Springer-Verlag, 2013.

The Majorize-Minimization algorithm

Iterative algorithm $\{\theta_k, k \geq 0\}$



- Given the current value θ_k

- define a majorizing function (red)
- tangent at θ_k
- minimizer of the majorizing function $\rightarrow \theta_{k+1}$

- Key property:

$$f(\theta_{k+1}) \leq f(\theta_k)$$

Proof:

$$\begin{aligned} f(\theta_{k+1}) &\leq M(\theta_{k+1};\theta_k) && M(\cdot;\theta_k) \text{ majorizes } f \\ &\leq M(\theta_k,;\theta_k) && \theta_{k+1} \text{ minimizer of } M(\cdot;\theta_k) \\ &= f(\theta_k) && \text{tangent property: } f(\theta_k) = M(\theta_k;\theta_k) \end{aligned}$$

- MM will be used to majorize f; and we then deduce a majorizing function for g.

Example 1: gradient-based algorithms

Assume that the loss function f is smooth with L_f-Lipschitz gradient on Θ.

- Ex. In large scale learning: $\nabla f(\theta) = \mathbb{E}_\pi\left[G(Z,\theta)\right]$

- **Majorizing function of f, tangent at $\tau \in \Theta$:**

$$\begin{aligned} f(\theta) &\leq f(\tau) + \langle \nabla f(\tau), \theta - \tau\rangle + \frac{L_f}{2}\|\theta - \tau\|^2 \\ &\leq C_\tau + \frac{1}{2\gamma}\|\theta\|^2 - \langle s_\tau, \theta\rangle \qquad s_\tau := \frac{1}{\gamma}\tau - \nabla f(\tau) \qquad \gamma \leq 1/L_f \end{aligned}$$

- **Minimization step.**

$$\begin{aligned} \mathsf{T}(s_\tau) &= \mathrm{prox}_{\gamma\, g}(\gamma\, s_\tau) \\ &= \mathrm{prox}_{\gamma\, g}\left(\tau - \gamma \nabla f(\tau)\right) \\ &= \mathrm{prox}_{\gamma\, g}\left(\tau - \gamma\, \mathbb{E}_\pi\left[G(Z,\tau)\right]\right) \end{aligned}$$

Gradient-based algorithms are among MM algorithms

With quadratic majorizing fcts.

Example 2: Expectation Maximization algorithms Dempster et al. (1977); Wu (1983)

Assume that the loss function:

$$\ell(Z,\theta) := -\log \int_{\mathcal{H}} p(Z,h,\theta)\ \nu(\mathsf{d}h), \qquad \log p(Z,h,\theta) = \langle S(Z,h), \phi(\theta)\rangle - \psi(\theta)$$

- Ex. negative log-likelihood + Latent variable model + complete data model in the curved exponential family.

- **Majorizing function of f, tangent at $\tau \in \Theta$. (E-step)**

Jensen's inequality:

$$\begin{aligned}\ell(Z,\theta) &\leq C_\tau - \int_{\mathcal{H}} \log p(Z,h,\theta)\ \nu(\mathsf{d}h|Z,\tau) \\ &\leq C_\tau + \psi(\theta) - \left\langle \int_{\mathcal{H}} S(Z,h)\ \nu(\mathsf{d}h|Z,\tau), \phi(\theta)\right\rangle .\end{aligned}$$

- **Optimization step. (M-step)**

$$\mathsf{T}(s_\tau) = \mathrm{argmin}_\theta \ \ g(\theta) + \psi(\theta) - \langle s_\tau, \phi(\theta)\rangle$$

$$s_\tau := \mathbb{E}_\pi \left[\int_{\mathcal{H}} S(Z,h)\ \nu(\mathsf{d}h|Z,\tau)\right]$$

EM algorithms are among MM algorithms

Majorizing fct **not** quadratic in g$^{\mathrm{al}}$

Example 3: Variational Surrogates

Assume that the loss function

$$\ell(Z,\theta) = \min_{h \in \mathcal{H}} \tilde{\ell}(Z,h,\theta), \qquad \tilde{\ell}(Z,h,\theta) := \psi(\theta) - \langle S(Z,h), \phi(\theta) \rangle$$

Set $\mathsf{M}(Z,\theta) := \mathrm{argmin}_{h \in \mathcal{H}} \tilde{\ell}(Z,h,\theta)$.

- **Majorizing function of f, at τ.**

$$\ell(Z,\theta) \leq C_\tau + \psi(\theta) - \langle S(Z, \mathsf{M}(Z,\tau)), \phi(\theta) \rangle$$

- **Optimization step.**

$$\mathsf{T}(s_\tau) = \mathrm{argmin}_\theta \;\; g(\theta) + \psi(\theta) - \langle s_\tau, \phi(\theta) \rangle$$

$$s_\tau := \mathbb{E}_{Z \sim \pi} \left[S(Z, \mathsf{M}(Z,\tau)) \right]$$

- Example:

Dictionary Learning

$$f(\theta) = \mathbb{E}_\pi \left[\min_{h \in \mathcal{H}} \|Z - \theta h\|^2 \right]$$

Z: observations $d \times 1$

θ: dictionary $d \times K$

h: code $K \times 1$

All these problems share

- A parametric majorizing function of f at $\tau \in \Theta$, of the form

$$\theta \mapsto C_\tau + \psi(\theta) - \left\langle \mathbb{E}_\pi \left[\bar{S}(Z,\tau)\right], \phi(\theta)\right\rangle$$

- A parametric majorizing function of $f+g$ at $\tau \in \Theta$, of the form

$$\theta \mapsto C_\tau + g(\theta) + \psi(\theta) - \left\langle \mathbb{E}_\pi \left[\bar{S}(Z,\tau)\right], \phi(\theta)\right\rangle$$

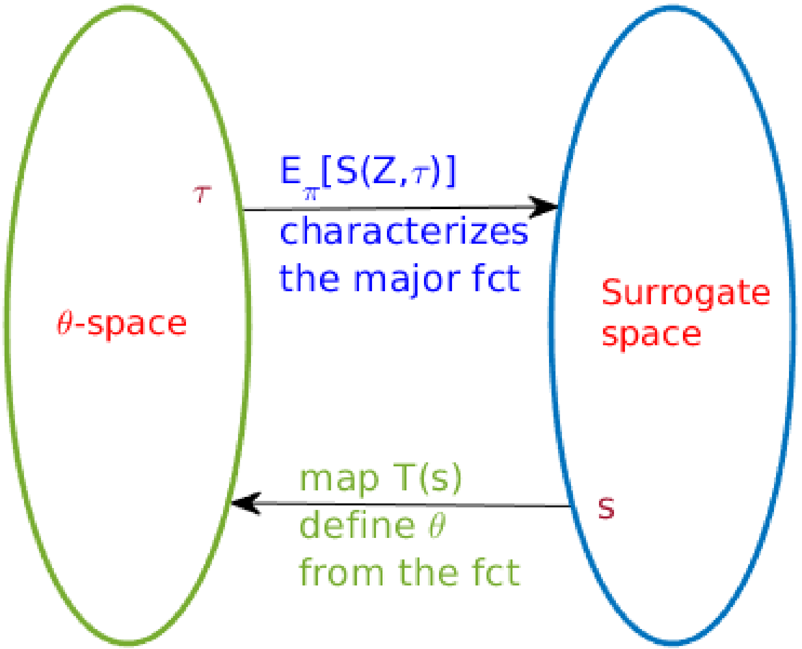
- The iterative process

$$\cdot \to \theta_k \to \mathbb{E}_\pi \left[\bar{S}(Z,\theta_k)\right] \xrightarrow{\mathsf{T}} \theta_{k+1} \to \mathbb{E}_\pi \left[\bar{S}(Z,\theta_{k+1})\right] \xrightarrow{\mathsf{T}} \theta_{k+2} \to \cdots$$

or equivalently

$$\cdot \to \theta_k \to \underbrace{\mathbb{E}_\pi \left[\bar{S}(Z,\theta_k)\right]}_{s_k} \xrightarrow{\mathsf{T}} \underbrace{\theta_{k+1}}_{\mathsf{T}(s_k)} \to \underbrace{\mathbb{E}_\pi \left[\bar{S}(Z,\mathsf{T}(s_k))\right]}_{s_{k+1}} \xrightarrow{\mathsf{T}} \underbrace{\theta_{k+2}}_{\mathsf{T}(s_{k+1})} \to \cdots$$

A unifying point of view Dieuleveut, F., Wai (2022)



- **... "the surrogate-space, in the foreground ! the θ-space in the background"**
- **the θ-space is the** *mirror*
- Novel approaches for (i) large scale learning, (ii) federated learning.

A specific structure for the majorizing functions

$$\mathrm{Argmin}_{\mathbb{R}^d}\left(f(\theta)+g(\theta)\right) \qquad f(\theta):=\mathbb{E}_{Z\sim\pi}\left[\ell(Z,\theta)\right]$$

We assume hereafter

- Hyp. **MM-1** There exist $\psi:\mathbb{R}^d\to\mathbb{R}$, $\phi:\mathbb{R}^d\to\mathbb{R}^q$, $\bar{S}:\mathbb{R}^p\times\mathbb{R}^d\to\mathcal{S}\subseteq\mathbb{R}^q$ s.t.

$$\forall\tau,\quad\forall\theta\in\Theta:\qquad f(\theta)\leq f(\tau)+\psi(\theta)-\psi(\tau)-\left\langle\mathbb{E}_\pi\left[\bar{S}(Z,\tau)\right],\phi(\theta)-\phi(\tau)\right\rangle$$

i.e. for any $\tau\in\Theta$, there exists a majorizing function for f, tangent at τ

The majorizing function of $(f+g)$ is in a *parametric* family of functions

$$\theta\mapsto C_\tau+g(\theta)+\psi(\theta)-\langle s_\tau,\phi(\theta)\rangle \qquad \text{where}\ \ s_\tau:=\mathbb{E}_\pi\left[\bar{S}(Z,\tau)\right]$$

- Hyp. **MM-2** For any $s\in\mathcal{S}$,

$$\mathsf{T}(s):=\mathrm{argmin}_\theta\left(g(\theta)+\psi(\theta)-\langle s,\phi(\theta)\rangle\right)$$

exists and is unique.

Stationary points in the θ-space / stationary points in the surrogate space

- Under regularity assumptions, for any $s \in \mathcal{S}$, there exists $p_s \in \partial g(\mathsf{T}(s))$,

$$p_s + \nabla f(\mathsf{T}(s)) = -J_\phi(\mathsf{T}(s))\ \mathsf{h}(s) \qquad \mathsf{h}(s) := \mathbb{E}_\pi\left[\bar{S}(Z, \mathsf{T}(s))\right] - s.$$

Dieuleveut, F., Wai (2022)

If $s_\star \in \mathcal{S}$ satisfies $\mathsf{h}(s_\star) = 0$ then $\theta_\star := \mathsf{T}(s_\star)$ is a stationary point of $f + g$.

And conversely.

- **We will design algorithms**

- in the s-space,

- targeting a zero of $\mathsf{h}(s)$

- $\mathsf{h}(s)$ **is an untractable expectation (w.r.t.** π**) !**

Stochastic surrogate MM

and
Variance reduction

The `Stochastic surrogate MM` algorithm Dieuleveut, F., Wai (2022)

- Based on Stochastic Approximation Robbins and Monro (1951); Benveniste et al. (1990): given learning rates $\{\gamma_k, k \geq 0\}$

$$\hat{S}_{k+1} = \hat{S}_k + \gamma_{k+1} H_{k+1} \qquad H_{k+1} \approx \mathsf{h}(\hat{S}_k) = \mathbb{E}_\pi \left[\bar{S}(Z, \mathsf{T}(\hat{S}_k)) \right] - \hat{S}_k .$$

```
Algorithm 1: Stochastic Surrogate MM (StoSur-MM)
Input: $k_{\max} > 0$, $\hat{S}_0 \in \mathcal{S}$
Result: StoSur-MM sequence $\{\hat{S}_k, k \leq k_{\max}\}$ and its mirror
        $\{\mathsf{T}(\hat{S}_k), k < k_{\max}\}$
for $k = 0, \ldots, k_{\max} - 1$ do
    Compute $\mathsf{T}(\hat{S}_k)$
    Sample $\mathsf{S}_{k+1}$, a random oracle for $\mathbb{E}_\pi[\bar{S}(Z, \mathsf{T}(\hat{S}_k))]$
    Set $\hat{S}_{k+1} = \hat{S}_k + \gamma_{k+1}(\mathsf{S}_{k+1} - \hat{S}_k)$
```

- **Examples of oracles:**

- Large batch learning:

$$\mathsf{S}_{k+1} = \frac{1}{\mathsf{b}} \sum_{j \in \mathcal{B}_{k+1}} \bar{S}(Z_j, \mathsf{T}(\hat{S}_k))$$

- Online learning:

$$\mathsf{S}_{k+1} = \bar{S}(Z_{k+1}, \mathsf{T}(\hat{S}_k))$$

- In the Gdt Case, known as "Stochastic Gdt"

In the EM case, known as "Online EM" Cappé and Moulines (2009)

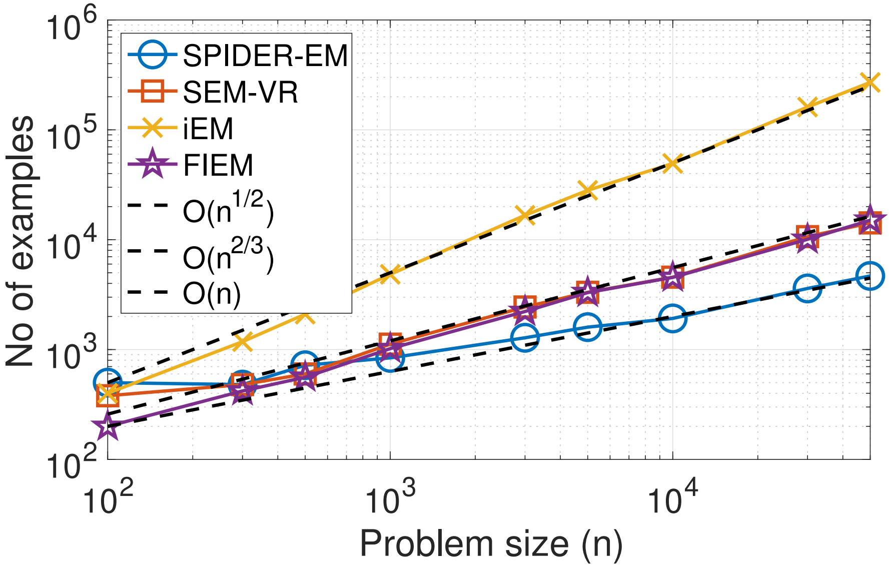
Large batch learning: Variance reduction within Stochastic Approximation (SA)

- Control variate

- If $\mathsf{h}(s) = \mathbb{E}[H]$ and $\mathbb{E}[V] = 0$, then $\mathsf{h}(s) = \mathbb{E}[H + V] \rightarrow$ many possibilities for the definition of SA
- "If U and V are negatively correlated, there exists c such that $\mathrm{Var}(U + cV) < \mathrm{Var}(U)$".

- Examples: Incremental Gdt / Incremental EM

- $\ldots$ / incremental EM Neal and Hinton (1998); Ng and McLachlan (2003)
- SVRG Johnson and Zhang (2013) / SEM-VR Chen et al. (2018)
- SAGA Defazio et al. (2014) / FIEM Karimi et al. (2019); F., Gach, Moulines (2021)
- SPIDER Fang et al. (2018)/ SPIDER-EM F., Moulines, Wai (2020)



Nbr of processed examples required to reach convergence, as a fct of the problem size n (F., Moulines, Wai, 2020)

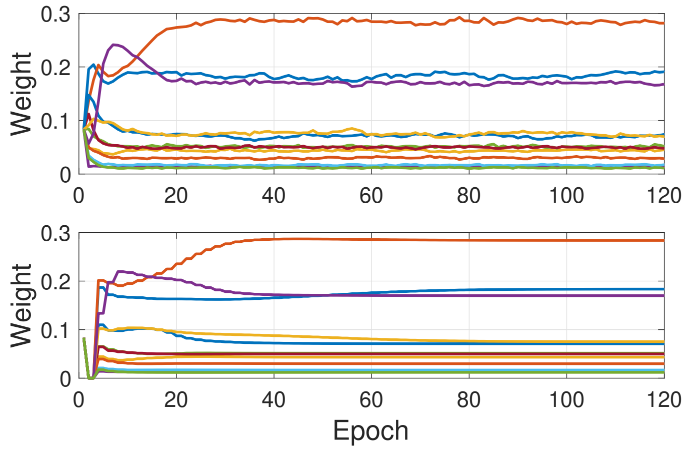
Focus on SPIDER EM F., Moulines, Wai (2020)

```
Algorithm 2: Stochastic Path-Integrated Differential EstimatoR - EM
Data: $k_{\text{in}} \in \mathbb{N}_\star$, $k_{\text{out}} \in \mathbb{N}_\star$, $\hat{S}_{\text{init}} \in \mathcal{S}$, $\{\gamma_{t,k+1}, t \geq 1, k \geq 0\}$ positive sequence.
Result: The SPIDER-EM sequence: $\hat{S}_{t,k}, t = 1, \ldots, k_{\text{out}}$ and $k = 0, \ldots, k_{\text{in}} - 1$
$\hat{S}_{1,0} = \hat{S}_{1,-1} = \hat{S}_{\text{init}}$,    $\mathsf{S}_{1,0} = N^{-1} \sum_j \bar{S}(Z_j, \mathsf{T}(\hat{S}_{1,-1}))$
for $t = 1, \ldots, k_{\text{out}}$ do
    for $k = 0, \ldots, k_{\text{in}} - 2$ do
        Sample a mini-batch $\mathcal{B}_{t,k+1}$ in $\{1, \ldots, N\}$ of size $\mathsf{b}$, with or without replacement
        $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \mathsf{b}^{-1} \sum_{j \in \mathcal{B}_{t,k+1}} \left\{ \bar{S}(Z_j, \mathsf{T}(\hat{S}_{t,k})) - \bar{S}(Z_j, \mathsf{T}(\hat{S}_{t,k-1})) \right\}$
        $\hat{S}_{t,k+1} = \hat{S}_{t,k} + \gamma_{t,k+1} \left( \mathsf{S}_{t,k+1} - \hat{S}_{t,k} \right)$
    $\hat{S}_{t+1,-1} = \hat{S}_{t,k_{\text{in}}-1}$
    $\mathsf{S}_{t+1,0} = N^{-1} \sum_j \bar{S}(Z_j, \mathsf{T}(\hat{S}_{t+1,-1}))$
    $\hat{S}_{t+1,0} = \hat{S}_{t,k_{\text{in}}-1} + \gamma_{t,k_{\text{in}}} (\mathsf{S}_{t+1,0} - \hat{S}_{t,k_{\text{in}}-1})$
```

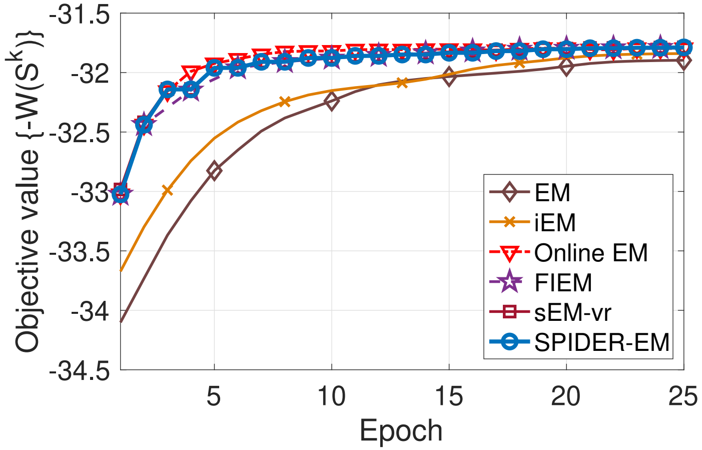
$$\mathsf{S}_{t,k+1} = \frac{1}{\mathsf{b}} \sum_{j \in \mathcal{B}_{t,k+1}} \bar{S}(Z_j, \mathsf{T}(\hat{S}_{t,k})) + \underbrace{\mathsf{S}_{t,k}}_{\approx \mathbb{E}_\pi\left[\bar{S}(Z, \mathsf{T}(\hat{S}_{t,k-1}))\right]} - \frac{1}{\mathsf{b}} \sum_{j \in \mathcal{B}_{t,k+1}} \bar{S}(Z_j, \mathsf{T}(\hat{S}_{t,k-1}))$$

- natural random field (red) and the control variate (blue): correlated through $\mathcal{B}_{t,k+1}$
- unfortunately, biased approximation $\mathsf{S}_{t,k+1}$ → restart every k_{in} iterations in order to cancel the bias.

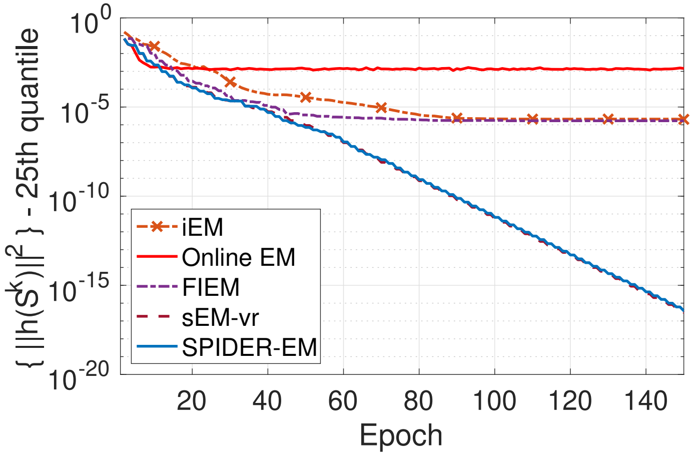
SPIDER EM for inference of Gaussian mixture (MNIST data set) F, Moulines, Wai (2020)



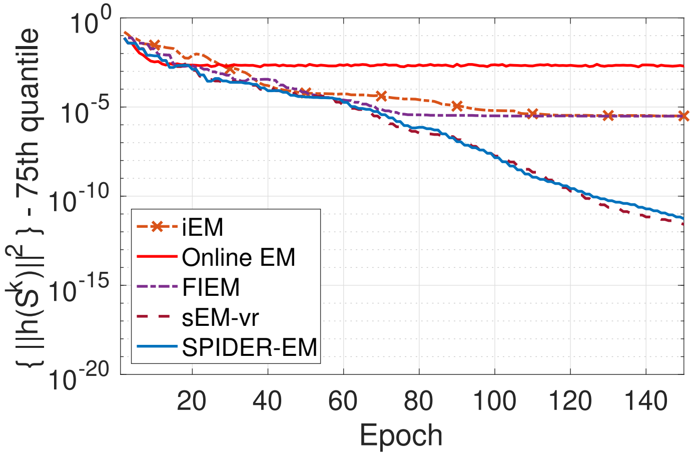
(top) Online EM, (bottom) SPIDER EM. Estimation of the weights, vs the nbr of epoch



Objective fct value vs the nbr of epoch



Quantile 0.25 of the distribution of $\|\mathsf{h}(\hat{S}_k)\|^2$



Quantile 0.75 of the distribution of $\|\mathsf{h}(\hat{S}_k)\|^2$

What kind of control of convergence ?

- From a Lyapunov inequality,

$$(f+g)(\mathsf{T}(\hat{S}_{k+1})) \leq (f+g)(\mathsf{T}(\hat{S}_k)) - (D_{k+1})^2 + R_{k+1}$$

- Almost-sure convergence when $k \to \infty$ (Robbins-Siegmund lemma)
- Explicit (upper) bounds $\Sigma_{k=0}^{k_{\max}-1} D_{k+1}^2 \leq (f+g)(\mathsf{T}(\hat{S}_0)) - \min(f+g) + \Sigma_{k=0}^{k_{\max}-1} R_{k+1}$
- Difficulties: it is **not** a gradient-SA algorithm

- Bounds for

- ϵ-stationarity (algo designed to find the roots of h)
- non convex optimization

$$\mathbb{E}\left[\|\mathsf{h}(\hat{S}_\tau)\|^2\right] \leq \bullet \qquad \tau \text{ uniform on } \{1, \cdots, k_{\max}\}$$

- scaling of the design parameters (k_{in}, k_{out}, b) and the learning rate as a fct of N (data set size) and the accuracy level ϵ.

ϵ-stationarity for SPIDER-EM

- Theorem F., Moulines, Wai (2020): Explicit control of the mean error $\|\mathsf{h}(\hat{S}_{\cdot})\|^2$

Set $W(s) := (f+g)(\mathsf{T}(s))$. /* Lyapunov fct in the s-space; smooth ($L_{\dot{W}}$); $\nabla W(s) = -B(s)\mathsf{h}(s)$ */

Set $L^2 := N^{-1} \sum_{i=1}^{N} L_i^2$. /* Lipschitz constants related to $s \mapsto \bar{S}(Z_i, \mathsf{T}(s))$ */

Fix $k_{\mathrm{out}}, k_{\mathrm{in}}, \mathsf{b} \in \mathbb{N}_\star$. Choose $\alpha \in (0, v_{\min}/\mu_\star(k_{\mathrm{in}}, \mathsf{b}))$ with

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

/*spectrum of $B(s)$ in $[v_{\min}, v_{\max}]$ uniformly in s */

Run the algorithm with $\xi_t = k_{\mathrm{in}}$ and $\gamma_{t,k} := \alpha/L$. Then

$$\mathbb{E}\left[\|\mathsf{h}\left(\hat{S}_{\tau,\xi-1}\right)\|^2\right] \leq \left(\frac{1}{k_{\mathrm{in}}} + \frac{\alpha^2}{\mathsf{b}}\right) \frac{1}{k_{\mathrm{out}}} \frac{2L}{\alpha\{v_{\min} - \alpha\mu_\star(k_{\mathrm{in}}, \mathsf{b})\}} \left(\mathbb{E}\left[W(\hat{S}_0)\right] - \min W\right)$$

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

Complexity analysis

- How to choose k_{in}, k_{out}, b in order to reach ϵ-stationarity, with an optimal computational cost ?

- ϵ-stationarity for non-convex optimization

$$\mathbb{E}\left[\|\mathsf{h}(\hat{S}_\tau)\|^2\right] \leq \epsilon \qquad \text{random stopping time } \tau$$

- Computational cost: *(i)* Nbr of processed example, *(ii)* Nbr of optimization steps.

- Result:

$$k_{\text{in}} = \mathsf{b} = O(\sqrt{n}), \quad k_{\text{out}} = O\left(\frac{1}{(\epsilon k_{\text{in}})}\right)$$

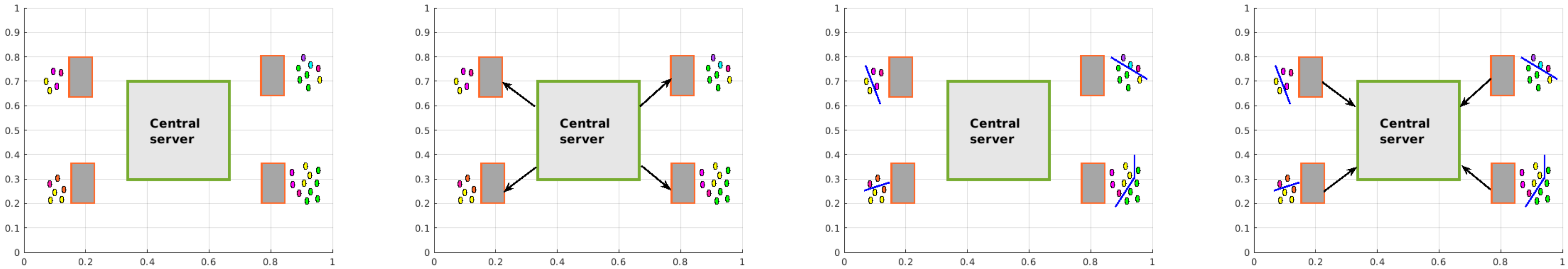
- Nbr of optimization steps: $O(1/\epsilon)$
- Nbr of processed examples: $O(\sqrt{n}\,\epsilon^{-1})$

Algorithm	Complexity (proc. ex.)
Online-EM	ϵ^{-2}
i-EM	$n\,\epsilon^{-1}$
sEM-vr	$n^{2/3}\,\epsilon^{-1}$
FIEM	$n^{2/3}\,\epsilon^{-1} \wedge \sqrt{n}\,\epsilon^{-3/2}$
SPIDER-EM	$\sqrt{n}\ \epsilon^{-1}$

SPIDER-EM – state of the art !

Federated learning through MM

Federated Learning



- The central server coordinates the participation of the local devices/clients/workers
- Local training data sets, **never** uploaded to the server
- FL reduces privacy and security risks

- Global model maintained by the central server: sent to the devices
- Each worker computes an update of the global model
- Only this update is communicated to the central server; aggregation by the central server

●Methodological developments under the 'constraints'

- Local data sets, heterogeneous, unbalanced
- Partial participation of the clients (charged devices, plugged-in, free wi-fi connection, $\cdots$)
- Massively distributed: large nbr devices w.r.t. the size of the local data sets
- Communication cost $>>$ Computational cost: compression

Federated MM: the data

- n local workers with their own data set ($\rightarrow$ distribution π_i)

- Local worker $\#i$: has a local objective function

$$\theta \mapsto f_i(\theta) := \mathbb{E}_{\pi_i}\left[\ell(Z,\theta)\right]$$

and the central server targets

$$\theta \mapsto f(\theta) := \sum_{i=1}^{n} \mu_i \, \mathbb{E}_{\pi_i}\left[\ell(Z,\theta)\right], \qquad \mu_i \text{ weights, unbalanced local data sets } (= 1/n \text{ for online learning})$$

- The central server runs a stochastic surrogate MM, by aggregation of local oracles provided by the local agents

$$\mathsf{S}_{k+1,i} \approx \mathbb{E}_{\pi_i}\left[\bar{S}(Z, \mathsf{T}(\hat{S}_k))\right] \longrightarrow \mathsf{S}_{k+1} \approx \mathbb{E}_{\pi}\left[\bar{S}(Z, \mathsf{T}(\hat{S}_k))\right]$$

and then, perform the optimization step: $\mathsf{T}(\hat{S}_{k+1})$ and broadcast it to the local agents.

Federated MM: partial participation and quantization

- At each iteration, a local worker $\#i$ is active with probability $p \in (0,1)$.

- If active at iteration $\#(k+1)$, the local worker $\#i$

- learns an oracle $\mathsf{S}_{k+1,i}$ of the parameter $\mathbb{E}_{\pi_i}\left[\bar{S}(Z, \mathsf{T}(\hat{S}_k))\right]$ characterizing the local majorizing function

- quantize the information before sending it to the central server: $\mathrm{Quant}(\cdot)$ see e.g. Alistarh et al. (2018), Horvath et al. (2019)

- the quantized information is **not the oracle**, but the difference with the previously sent information $V_{k,i}$

$$\Delta_{k+1,i} := \mathsf{S}_{k+1,i} - \hat{S}_k - V_{k,i}$$

- it updates this control variate $V_{\cdot,i}$'s adapted from Mishchenko et al., 2019

$$V_{k+1,i} = V_{k,i} + \alpha \; \mathrm{Quant}(\Delta_{k+1,i})$$

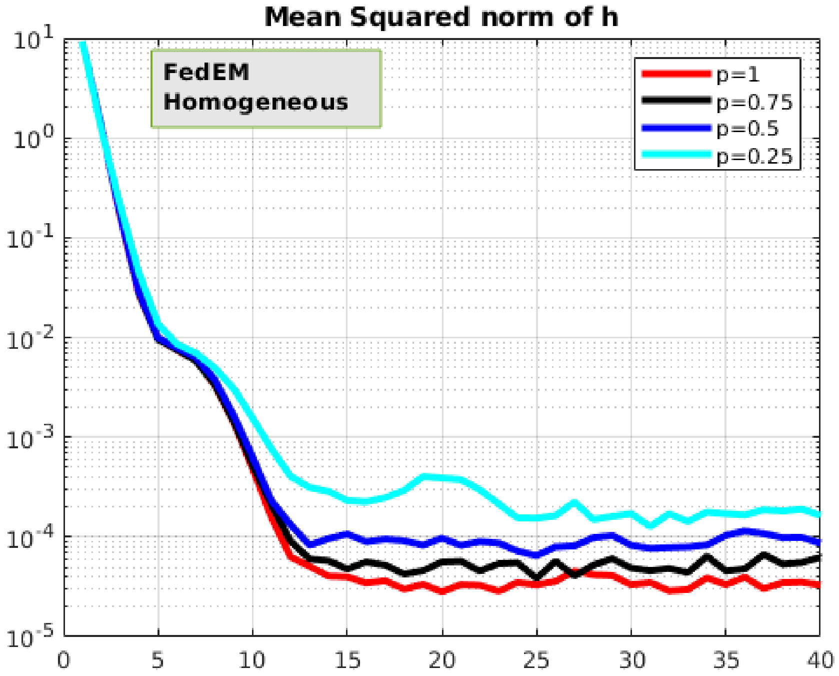
Federated MM Dieuleveut, F., Moulines, Robin (2021) and Dieuleveut, F., Wai (2022)

Algorithm 3: Federated MM with partial participation (PP)

```
Data: $k_{\max} \in \mathbb{N}^\star$; for $i \in [n]^\star$, $V_{0,i} \in \mathbb{R}^q$; $\hat{S}_0 \in \mathbb{R}^q$; a positive sequence $\{\gamma_{k+1}, k \in [k_{\max} - 1]\}$; $\alpha > 0$ and $p \in (0,1)$.
Result: The sequence: $\{\hat{S}_k, k \in [k_{\max}]\}$ and its mirror $\mathsf{T}(\hat{S}_k)$
Set $V_0 = n^{-1} \sum_{i=1}^n V_{0,i}$
for $k = 0, \ldots, k_{\max} - 1$ do
    Sample $\mathcal{A}_{k+1}$, the set of active workers    /* each worker active with probability $p$, independently */
    for $i \in \mathcal{A}_{k+1}$ do
        Sample $\mathsf{S}_{k+1,i}$, an approximation of $\mathbb{E}_{\pi_i}\left[\bar{S}(Z, \mathsf{T}(\hat{S}_k))\right]$
        Set $\Delta_{k+1,i} = \mathsf{S}_{k+1,i} - \hat{S}_k - V_{k,i}$
        Set $V_{k+1,i} = V_{k,i} + \alpha\,\mathrm{Quant}(\Delta_{k+1,i})$.    /* $0 < \alpha \leq 1/(1+\omega)$. $V_k$: control variate */
        Send $\mathrm{Quant}(\Delta_{k+1,i})$ to the central server    /* random, unbiased, $\mathbb{E}[\|\mathrm{Quant}(x) - x\|^2] \leq \omega\|x\|^2$ */
    for $i \notin \mathcal{A}_{k+1}$ do
        Set $V_{k+1,i} = V_{k,i}$ (no update)
    Set $H_{k+1} = V_k + p^{-1} \sum_{i \in \mathcal{A}_{k+1}} \mu_i \mathrm{Quant}(\Delta_{k+1,i})$    /* compensate for (i) the $V_k$'s and (ii) the PP */
    Set $\hat{S}_{k+1} = \hat{S}_k + \gamma_{k+1} H_{k+1}$
    Set $V_{k+1} = V_k + \alpha \sum_{i \in \mathcal{A}_{k+1}} \mu_i \mathrm{Quant}(\Delta_{k+1,i})$    /* learn the aggregated control variates */
    Send $\hat{S}_{k+1}$ and $\mathsf{T}(\hat{S}_{k+1})$ to the $n$ workers
```

Toy example: inference of a $\mathbb{R}^2$-valued Gaussian mixture model with 2 components (1/2)

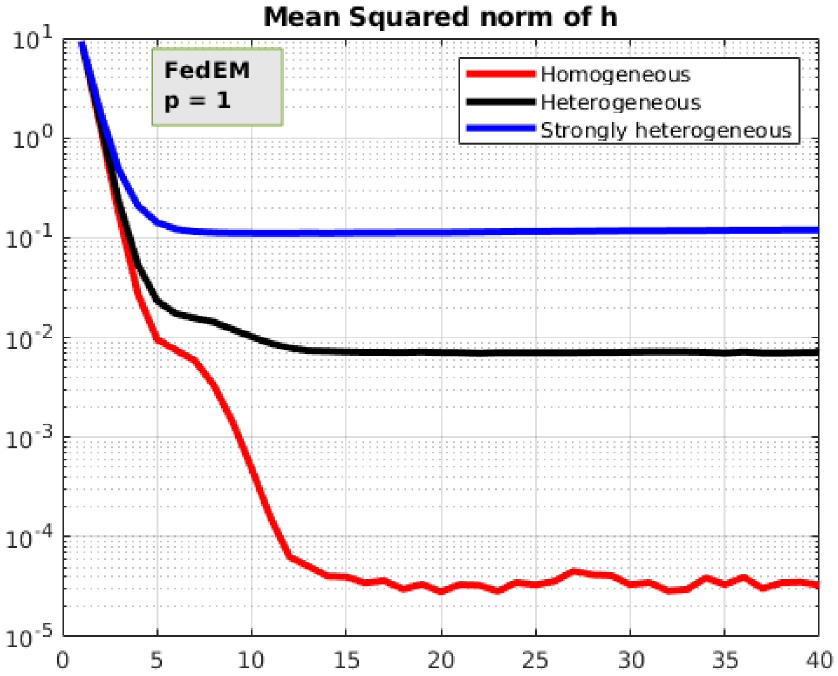
- Robustness to partial participation



$k \mapsto \mathbb{E}\left[\|h(\hat{S}_k)\|^2\right]$ vs the nbr of epochs.

Estimated by Monte Carlo

- Robustness to heterogeneity



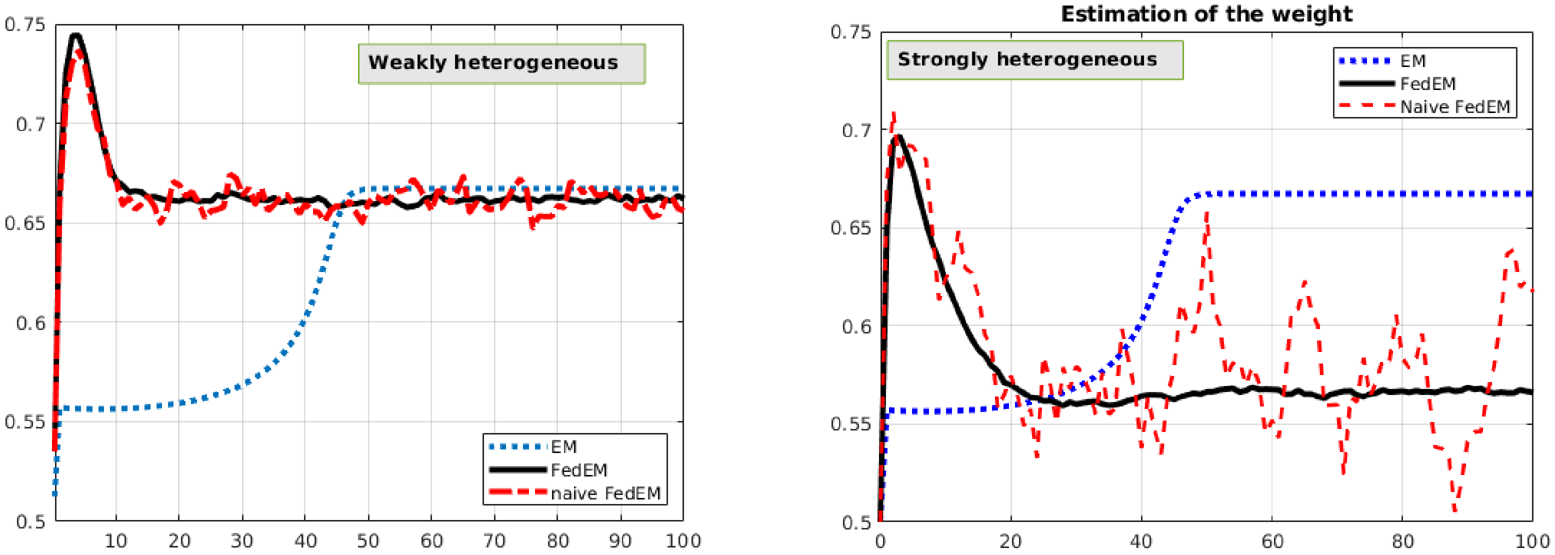
$k \mapsto \mathbb{E}\left[\|h(\hat{S}_k)\|^2\right]$ vs the nbr of epochs.

Estimated by Monte Carlo

Toy example: inference of a $\mathbb{R}^2$-valued Gaussian mixture model with 2 components (2/2)

- Federated MM vs naive-Federated MM ?

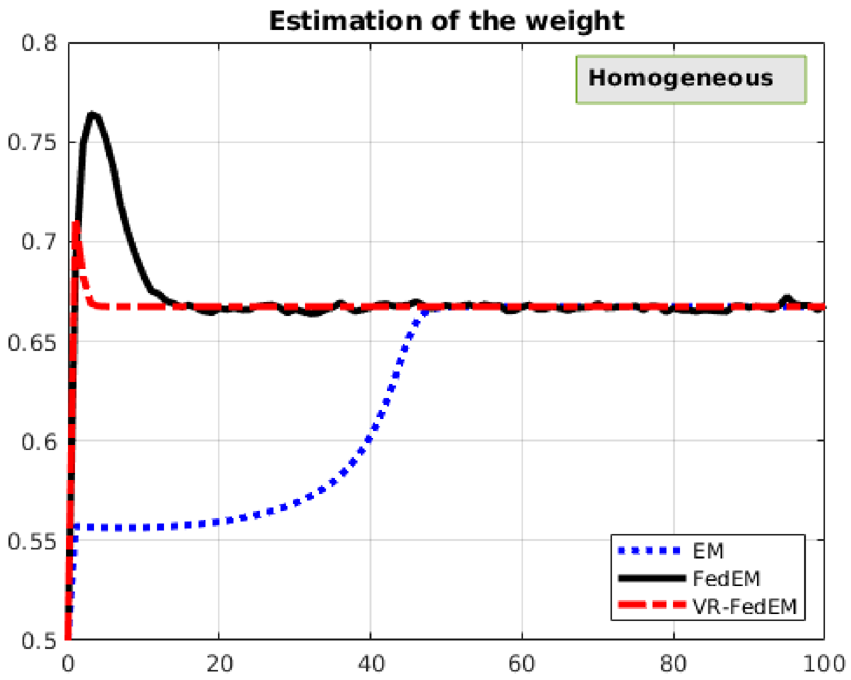
In naive-Federated MM: remove the variables $V_{\cdot c}$'s – i.e. the control variates introduced to control the variance of the quantization step.

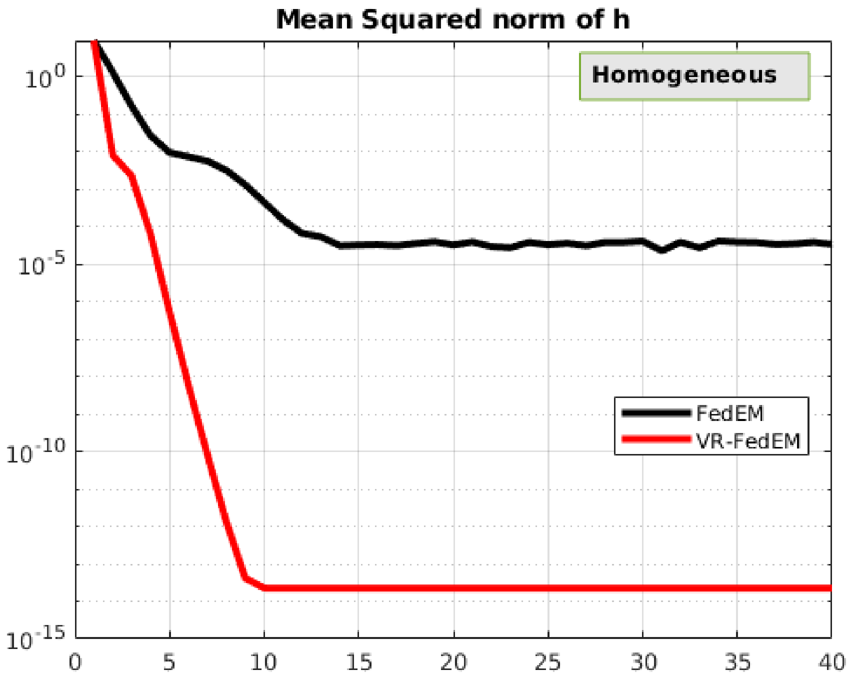


Estimation of the weight vs the nbr epoch; Case "homogeneous" and case "strongly heterogeneous"

Variance Reduction in Federated MM

- Mix the SPIDER variance reduction idea and the federated learning algorithm
- Case of EM (toy example, to follow)





Estimation of the weight vs the nbr epoch

$k \mapsto \mathbb{E}\left[\|h(\hat{S}_k)\|^2\right]$ vs the nbr of epochs.

Estimated by Monte Carlo

Bounds for convergence

- Theorem (adapted from) Dieuleveut, F., Moulines, Robin (2021): Explicit control of the mean error $\|\mathsf{h}(\hat{S}_\cdot)\|^2$

Run the algorithm with $\alpha := (1+\omega)^{-1}$ /* $\mathbb{E}[\|\mathrm{Quant}(x) - x\|^2] \leq \omega \|x\|^2$ */

and $\gamma_k = \gamma \in (0, \gamma_{\max}]$, where

$$\gamma_{\max} := \frac{v_{\min}}{2L_{\dot{W}}} \wedge \frac{p\sqrt{n}}{2\sqrt{2}L(1+\omega)\sqrt{\omega + (1-p)(1+\omega)/p}}.$$

Set $\sigma^2 := n \,\Sigma_{i=1}^n \, \mu_i^2 \, \sigma_i^2$. /* σ_i^2: variance of the oracles at agent $\#i$ */

Denote by τ the uniform random variable on $\{0, \cdots, k_{\max} - 1\}$. Then,

$$v_{\min}\left(1 - \gamma \frac{L_{\dot{W}}}{v_{\min}}\right) \mathbb{E}\ \left[\|\mathsf{h}(\hat{S}_\tau)\|^2\right] \leq \frac{\left(W(\hat{S}_0) - \min W\right)}{\gamma k_{\max}} + \frac{\gamma}{k_{\max}} 2L_{\dot{W}} \frac{\omega}{\alpha} \sum_{i=1}^n \mu_i^2 \|V_{0,i} - \mathsf{h}_i(\hat{S}_0)\|^2 + \gamma L_{\dot{W}} \frac{1 + 5\left(\omega + (1-p)(1+\omega)/p\right)}{n} \sigma^2.$$

Complexity analysis in the case $p = 1$

Given an accuracy level ϵ, how to choose the design parameters in order to minimize the number of optimization ?

- The number of optimization is $k_{\max}$ chosen in order to reach the accuracy level ϵ:

$$\mathcal{K}_{\text{opt}}(\epsilon) = O\left(\frac{1}{\epsilon^2}\frac{(1+\omega)\sigma^2}{n}\right) \vee O\left(\frac{1}{\epsilon\,\gamma_{\max}}\right)$$

1st term is leader iff $\epsilon << \gamma_{\max}(1+\omega)\sigma^2/n$ (high noise regime)

- **Compression effect:**

γ impacted by compression iff $n << \omega^3$.

On $\mathcal{K}_{\text{opt}}$: see table

	Complexity regime:	$\frac{(1+\omega)\sigma^2}{n\epsilon^2}$	$\frac{1}{\gamma_{\max}\epsilon}$
$\gamma_{\max}$ **regime:**	**E.g. case when**	High noise σ^2, small ϵ	Low σ^2 larger ϵ
$\frac{v_{\min}}{2L_{\dot{W}}}$	large ratio n/ω^3	$\times\omega$	$\times 1$
$\frac{\sqrt{n}}{2\sqrt{2}L(1+\omega)\sqrt{\omega}}$	low ratio n/ω^3	$\times\omega$	$\times\omega^{3/2}/\sqrt{n}$

Conclusion

Contributions

- A unifying point of view: surrogate MM covering popular algorithms.

- "surrogate in the foreground": efficient approach to *(i)* summarize the data through "sufficient statistics", *(ii)* in FL: interpretation of the quantity aggregated at the central server.

- Novel algorithms for large scale learning (large batch, online); and federated learning.

- Explicit bound of convergence, complexity analysis for the Stochastic Surrogate MM, possibly with variance reduction, and Federated Surrogate MM.

- Non convex optimization. Algorithms and some theoretical results, for the non smooth optimization case.

Open theoretical analyses

- The map T: when no closed form.
- The constraints $s \in \mathcal{S} \rightarrow$ Stochastic Approximation restricted to a structured subset of $\mathbb{R}^d$.
- Other strategies for variance reduction in Federated Learning

Bibliography

Talk based on the publications:

G. Fort, E. Moulines, and H.-T. Wai.
A Stochastic Path Integral Differential EstimatoR Expectation Maximization Algorithm.
In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 16972–16982. Curran Associates, Inc., 2020.

A. Dieuleveut, G. Fort, E. Moulines and G. Robin.
Federated-EM with heterogeneity mitigation and variance reduction.
In A. Beygelzimer and Y. Dauphin and P. Liang and J. Wortman Vaughan editors, Advances in Neural Information Processing Systems, 2021.

Other references

D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli.
The convergence of sparsified gradient methods.
In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31, pages 5973–5983. Curran Associates, Inc., 2018.

A. Benveniste, M. Métivier, and P. Priouret.
Adaptive Algorithms and Stochastic Approximations.
Springer Verlag, 1990.

Cappé, O. and Moulines, E.
On-line Expectation Maximization algorithm for latentdata models.
J Roy Stat Soc B Met 71(3):593–613, 2009.

Chen, J. and Zhu, J. and Teh, Y. and Zhang, T.
Stochastic Expectation Maximization withVariance Reduction.
In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems31, Curran Associates, Inc., pp 7967–7977; 2018.

Defazio, A. and Bach, F. and Lacoste-Julien, S.
SAGA: A Fast Incremental GradientMethod With Support for Non-Strongly Convex Composite Objectives.
In: Ghahra-mani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds)Advances inNeural Information Processing Systems 27, Curran Associates, Inc., pp 1646–1654; 2014.

A. Dempster, N. Laird, and D. Rubin.
Maximum Likelihood from Incomplete Data via the EM Algorithm.
J. Roy. Stat. Soc. B Met., 39(1):1–38, 1977.

Fang, C. and Li, C. and Lin, Z. and Zhang, T.
SPIDER: Near-Optimal Non-Convex Optimiza-tion via Stochastic Path-Integrated Differential Estimator.
In: Bengio S, WallachH, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in NeuralInformation Processing Systems 31, Curran Associates, Inc., pp 689–69; 2018.

Fort, G. and Gach, P. and Moulines, E.
The Fast Incremental Expectation Maximization for finite-sum optimization: asymptotic convergence.
Statistics and Computing, 31, 2021.

Horváth, S. and Kovalev, D. and Mishchenko, D. and Stich, S. and Richtárik, P.
Stochastic distributed learning with gradient quantization and variance reduction.
arXiv preprint arXiv:1904.05115, 2019.

Johnson, R. and Zhang, T.
Accelerating Stochastic GradientDescent using PredictiveVariance Reduction.
In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Wein-berger KQ (eds) Advances in Neural Information Processing Systems 26, CurranAssociates, Inc., pp 315–323; 2013.

Karimi, B. and Wai, H.T. and Moulines, E. and Lavielle, M.
On the GlobalConvergence of(Fast) Incremental Expectation Maximization Methods.
In:Wallach H, LarochelleH, Beygelzimer A, d'Alchée Buc F, Fox E, Garnett R (eds) Advances in NeuralInformation Processing Systems 32, Curran Associates, Inc., pp 2837–2847; 2019.

K. Lange.
MM Optimization Algorithms.
SIAM-Society for Industrial and Applied Mathematics, 2016.