Fast Incremental Expectation Maximization: how many iterations for an ϵ-stationary point ?

Gersende Fort

CNRS & Institut de Mathématiques de Toulouse, France

CIRM "Optimization for Machine Learning", March 2020.

Based on a joint work with

- Pierre Gach - (IMT, Univ. Paul Sabatier, France)
- Eric Moulines - (CMAP, Ecole Polytechnique, France)

Acknowledgments:

- *Fondation Simone et Cino Del Duca* - **Op**timisation et **Si**mulation **Mo**nte Ca**r**lo : **E**ntrelacements
- *Labex CIMI* - **S**tochastic **D**escent Algorithms **with** **M**arkovian **I**nputs

.

Optimization pbm for Statistical Learning

.

An optimization problem occurring, for example,
in large scale Statistical inference

The optimization problem

$$\text{argmin}_{\theta \in \Theta} F(\theta), \qquad F(\theta) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\theta) + R(\theta)$$

with

- $\Theta \subseteq \mathbb{R}^d$

- n is large
- $\mathcal{L}_i : \Theta \to \mathbb{R}$ is not explicit and of the form

$$\mathcal{L}_i(\theta) \stackrel{\text{def}}{=} -\log \int_Z \exp\left(\langle s_i(z), \phi(\theta) \rangle\right) \, \mathrm{d}\mu(z)$$

- **No convexity** assumptions
- (for the cvg analysis) Regularity properties of $\mathcal{L}_i, R$

Motivation: Statistical inference (1/3)

$$\operatorname{argmin}_{\theta\in\Theta} F(\theta), \qquad F(\theta) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) + R(\theta), \qquad \mathcal{L}_i(\theta) \overset{\text{def}}{=} -\log \int_{\mathsf{Z}} \exp\left(\langle s_i(z), \phi(\theta)\rangle\right) \, \mathrm{d}\mu(z)$$

- n observations $Y_1, \ldots, Y_n$: a parameteric statistical model indexed by $\theta \in \Theta$
- $R(\theta)$: a penalty term, a regularization term, a prior (in a Bayesian setting solved by MAP)
- A loss function associated to each Y_i: $\mathcal{L}_i(\theta)$
- $\exp(-\mathcal{L}_i)$ is a "sum" over **latent** variables $z \in \mathsf{Z}$. For example,

$$\mathcal{L}_i(\theta) = \underbrace{-}_{\text{negative}} \underbrace{\log}_{\text{log-}} \underbrace{\int_{\mathsf{Z}} \ldots \mathrm{d}\mu}_{\text{likelihood of } Y_i}$$

Example: mixture models (2/3)

- data: $y_1, \cdots, y_n$ modeled as iid from: $\sum_{\ell=1}^{L} \omega_\ell \, \mathcal{N}(\mu_\ell, 1)$ $\qquad \theta = (\omega_{1:L}, \mu_{1:L})$
- Equivalently: $z_i \in \{1, \cdots, L\}$ s.t. $\mathcal{L}(Y_i | Z_i = \ell) \sim \mathcal{N}(\mu_\ell, 1)$ $\qquad \mathcal{L}(Z_i) \sim (\omega_\ell)_\ell$.
- Joint distribution of (Y_i, Z_i) $\qquad \omega_\ell / \omega_L = \exp(\alpha_\ell)$

$$\sum_{\ell=1}^{L} 1_{z_i=\ell} \, \omega_\ell \, \exp\left(-(y_i - \mu_\ell)^2/2\right) = \sum_{\ell=1}^{L} 1_{z_i=\ell} \, \exp\left(\ln \omega_\ell - (y_i - \mu_\ell)^2/2\right)$$

$$= \exp\left(\sum_{\ell=1}^{L} 1_{z_i=\ell} \left\{ \ln \omega_\ell - (y_i - \mu_\ell)^2/2 \right\} \right)$$

$$= \exp\left(\sum_{\ell=1}^{L-1} 1_{z_i=\ell} \left\{ \alpha_\ell - (y_i - \mu_\ell)^2/2 + (y_i - \mu_L)^2/2 \right\} - \ln(1 + \sum_{\ell=1}^{L-1} \exp(\alpha_\ell)) - (y_i - \mu_L)^2/2 \right)$$

$$s_i(z) \stackrel{\text{def}}{=} (1_{z=1}, \ldots, 1_{z=L-1}, \quad y_i 1_{z=1}, \ldots, y_i 1_{z=L-1}, \quad y_i, \quad -1),$$

$$\phi(\theta) \stackrel{\text{def}}{=} \left(\alpha_1 - \frac{\mu_1^2 - \mu_L^2}{2}, \ldots, \alpha_{L-1} - \frac{\mu_{L-1}^2 - \mu_L^2}{2}, \mu_1, \ldots, \mu_{L-1}, \mu_L, \ln(1 + \sum_{\ell=1}^{L-1} \exp(\alpha_\ell)) + \mu_L^2/2 \right)$$

Example: Logistic regression (3/3)

- data: $y_1, \cdots, y_n \in \{0,1\}^n$ modeled as indep from
- $\mathcal{L}(Y_i|Z_i) \sim \text{Bern}\left(\frac{1}{1+\exp(-\eta_i(z_i))}\right)$ $\qquad$ $\mathcal{L}(Z_i) \sim \mathcal{N}_p(\theta, \mathrm{I})$.
- The joint distribution of (Y_i, Z_i)

$$\left(\frac{\exp(\eta_i(z_i))}{1+\exp(\eta_i(z_i))}\right)^{Y_i} \left(\frac{1}{1+\exp(\eta_i(z_i))}\right)^{1-Y_i} \exp\left(-\|z_i-\theta\|^2/2\right) = \frac{\exp(Y_i\,\eta_i(z_i))}{1+\exp(\eta_i(z_i))} \exp\left(-\|z_i-\theta\|^2/2\right).$$

- The likelihood of Y_i

$$\int_{\mathbb{R}^p} \frac{\exp(Y_i\,\eta_i(z_i))}{1+\exp(\eta_i(z_i))} \exp\left(-\|z_i-\theta\|^2/2\right)\, \mathrm{d}z = \int_Z \exp(\langle s_i(z), \phi(\theta)\rangle)\, \mathrm{d}z$$

$$s_i(z) = \left(y_i\eta_i(z) - \ln(1+\exp(\eta_i(z))) - \|z\|^2/2,\ 1,\ z'\right)$$
$$\phi(\theta) = \left(1,\ -\|\theta\|^2/2,\ \theta'\right)$$

Which numerical tool ?

$$\text{argmin}_{\theta\in\Theta} F(\theta), \qquad F(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) + R(\theta),$$

$$\mathcal{L}_i(\theta) \stackrel{\text{def}}{=} -\log \int_{\mathsf{Z}} \exp\left(\langle s_i(z), \phi(\theta)\rangle\right)\ \mu(\mathrm{d}z).$$

An algorithmic solution designed for
- large n: rare computations of a sum over n terms allowed,
- non convex setting

Solution: Expectation Maximization-based methods: Dempster et al. (1977), Wu (1983)

.

II- Expectation Maximization (EM) algorithms

.

EM algorithm: its derivation for this optim pbm,
its intractability,
an alternative.

EM: A Majorize-Minimization algorithm (1/2)

$$\text{Argmin}_\theta F(\theta) \qquad F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) + R(\theta) \qquad \mathcal{L}_i(\theta) = -\log \int_{\mathsf{Z}} \exp(\langle s_i(z), \phi(\theta)\rangle \ \mathrm{d}\mu(z)\,.$$

- The surrogate function at the current point $\theta_k \in \Theta$, (Jensen's inequality)

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) \leq \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta_k) + \langle \bar{s}(\theta_k), \phi(\theta_k)\rangle - \langle \bar{s}(\theta_k), \phi(\theta)\rangle$$

where

$$\bar{s}(\theta_k) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \bar{s}_i(\theta_k), \qquad \bar{s}_i(\theta_k) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} s_i(z)\, \frac{\exp(\langle s_i(z), \phi(\theta_k)\rangle)}{\exp(-\mathcal{L}_i(\theta_k))}\, \mu(\mathrm{d}z).$$

- [E-step] Compute $\bar{s}(\theta_k)$
- [M-step] Minimize the majorizing function (under hyp: unique argmin)

$$\theta_{k+1} = \mathsf{T} \circ \bar{s}(\theta_k) \stackrel{\text{def}}{=} \text{Argmin}_{\theta\in\Theta} \{-\langle \bar{s}(\theta_k), \phi(\theta)\rangle + \mathsf{R}(\theta)\}\,.$$

EM in the Statistic-space (2/2)

	E	M	E	M	
θ_k	$\bar{s}(\theta_k)$	$\theta_{k+1} = \mathsf{T} \circ \bar{s}(\theta_k)$	$\bar{s} \circ \mathsf{T} \circ \bar{s}(\theta_k)$	$\theta_{k+2} = \mathsf{T} \circ \bar{s} \circ \mathsf{T} \circ \bar{s}(\theta_k)$	$F(\theta)$
	s_k	$\mathsf{T}(s_k)$	$s_{k+1} = \bar{s} \circ \mathsf{T}(s_k)$	$\mathsf{T} \circ \bar{s} \circ \mathsf{T}(s_k)$	$V(s)$

We will see EM-based algorithms as evolving in the "s-space"; the objective function: $V = F \circ \mathsf{T}$

- If convergence, to the roots of

$$h(s) \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(s) - s.$$

Under assumptions, the roots of h are the roots of $\dot{V}$. Do Stochastic EM avoid traps ? not the topic today

EM is designed to find the roots of h

From EM to Stochastic Approximation within EM

- Each iteration of EM requires

$$s_{k+1} = \frac{1}{n} \sum_{i=1}^{n} \bar{s}_i(\mathsf{T}(s_k))$$

It does not answer the specifications for large scale learning.

- It is designed to find the zeros of h and h is intractable:

$$h(s) = \frac{1}{n} \sum_{i=1}^{n} \bar{s}_i \circ \mathsf{T}(s) - s = \mathbb{E}\left[\bar{s}_I \circ \mathsf{T}(s) - s\right] \qquad I \sim \mathcal{U}(\{1, \cdots, n\})$$

what about *Stochastic Approximation within EM approaches ?*

among the many "Stochastic EM" algorithms [SEM] Celeux & Diebolt (1985); [MCEM] Wei & Tanner (1990), Fort & Moulines (2003); [SAEM] Delyon, Lavielle & Moulines (1999); Kuhn & Lavielle (2004); [Online EM] Cappé & Moulines (2009), [Incremental EM] Neal & Hinton (1999) ...

Stochastic Approximation (SA) within EM:

EM $\rightarrow$ designed to find the roots of $h(s) \stackrel{\text{def}}{=} \bar{s} \circ \mathsf{T}(s) - s$

- What is SA ?

Solve $h(s) = 0$ when $h(s) \stackrel{\text{def}}{=} \mathbb{E}\left[H(U, s)\right]$ by:

$$\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} \underbrace{H(U_{k+1}, \widehat{S}^k)}_{\text{approx. of } h(\widehat{S}^k)}, \qquad U_{k+1} \sim U \text{ or "almost"}$$

- Here, possibly two sources of intractability

$$h(s) = \frac{1}{n} \sum_{i=1}^{n} \bar{s}_i \circ \mathsf{T}(s) - s = \frac{1}{n} \sum_{i=1}^{n} \int_{\mathsf{Z}} s_i(z)\, \pi(\mathrm{d}z; \mathsf{T}(s)) - s$$

Delyon et al. (1999), the second intractability only with i.i.d. U'_k.

Extensions to MCMC sampling, first by Lavielle & Kuhn (2004)

1st idea: SA

$$h(s) = \frac{1}{n}\sum_{i=1}^{n} \bar{s}_i \circ \mathsf{T}(s) - s = \mathbb{E}\left[\bar{s}_J \circ \mathsf{T}(s) - s\right] \qquad J \sim \mathcal{U}([[n]])$$

Instead of the EM iterations $s_{k+1} = h(s_k) = \bar{s} \circ \mathsf{T}(s_k)$ do:

Data: $\widehat{S}^0 \in \mathcal{S}$, $\gamma_k \in (0, \infty)$ for $k \geq 1$
Result: The SA sequence: $\widehat{S}^k, k = 0, \ldots,$
for $k \geq 1$ **do**
    Sample J_{k+1} uniformly on $[[n]]$;
    $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}\left(\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k\right)$

Fast Incremental EM (FIEM) [FIEM] Karimi et al. (2019); [SAGA] Defazio et al. (2014)

$$h(s) = \mathbb{E}\left[\bar{s}_I \circ \mathsf{T}(s) - s + V_{k+1}\right] \qquad I \sim \mathcal{U}([[n]]), \quad \mathbb{E}[V_{k+1}] = 0$$

Data: $\widehat{S}^0 \in \mathcal{S}$, $\gamma_k \in (0, \infty)$ for $k \geq 1$

Result: The FIEM sequence: $\widehat{S}^k, k = 0, \ldots,$

$\mathsf{S}_{0,i} = \bar{s}_i \circ \mathsf{T}(\widehat{S}^0)$ for all $i \in [[n]]$;

$\widetilde{S}^0 = n^{-1} \sum_{i=1}^n \mathsf{S}_{0,i}$;

for $k \geq 1$ **do**

* Sample I_{k+1} uniformly on $[[n]]$;
* $\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \neq I_{k+1}$;
* $\mathsf{S}_{k+1,I_{k+1}} = \bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$;
* $\widetilde{S}^{k+1} = \widetilde{S}^k + n^{-1}\left(\mathsf{S}_{k+1,I_{k+1}} - \mathsf{S}_{k,I_{k+1}}\right)$;

Sample J_{k+1} uniformly on $[[n]]$;

$\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}\left(\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) - \widehat{S}^k - \left\{\mathsf{S}_{k+1,J_{k+1}} - \widetilde{S}^{k+1}\right\}\right)$

* iterative comput. of the sum:

$$\widetilde{S}^{k+1} = \frac{1}{n}\sum_{i=1}^n \bar{s}_i \circ \mathsf{T}(\widehat{S}^{\text{last time when } \#i \text{ sampled}})$$

.

III- Non asymptotic convergence bounds

.

Non asymptotic $\rightarrow$ a maximal number of iterations
How to define the estimate of the solution ?
"Convergence bounds": in which sense ?
How to choose the stepsize sequence ?

Non asymptotic convergence bounds: which ones ?

- The user chooses a maximal length of the path: $K_{\max}$.

K is a random stopping time in the range $\{0, \ldots, K_{\max} - 1\}$. Ghadimi & Lan (2013)

- mean "distance" to the roots of the gradient of $V \overset{\text{def}}{=} F \circ \mathsf{T}$

$$\inf_{k \leq K_{\max}} \mathbb{E}\left[\|\dot{V}(\widehat{S}^k)\|^2\right] \leq \mathsf{E}_g \overset{\text{def}}{=} \mathbb{E}\left[\|\dot{V}(\widehat{S}^K)\|^2\right]$$

- mean "distance" to the roots of h

$$\mathsf{E}_h \overset{\text{def}}{=} \mathbb{E}\left[\|h(\widehat{S}^K)\|^2\right]$$

- Under the stated assumptions

$$\frac{1}{v_{\max}^2} \mathsf{E}_g \leq \mathsf{E}_h$$

Assumptions

A1 $\Theta \subseteq \mathbb{R}^d$ is an open set. $(\mathsf{Z}, \mathcal{Z})$ is a measurable space and μ is a σ-finite positive measure on $\mathcal{Z}$. The functions $\phi : \Theta \to \mathbb{R}^q$, $s_i : \mathsf{Z} \to \mathbb{R}^q$ for all $i \in [[n]]$ and $\mathsf{R} : \Theta \to \mathbb{R}$ are measurable functions. Finally, for any $\theta \in \Theta$ and $i \in [[n]]$, $-\infty < \mathcal{L}_i(\theta) < \infty$.

A2 For all $\theta \in \Theta$ and $i \in [[n]]$, the expectation $\bar{s}_i(\theta)$ exists. For any $s \in \mathbb{R}^q$, $\mathrm{Argmin}_{\theta \in \Theta} \left(- \langle s, \phi(\theta) \rangle + \mathsf{R} \right.$ is a (non empty) singleton denoted by $\{\mathsf{T}(s)\}$.

A3 ϕ, R are C^1 on Θ. T is C^1 on $\mathbb{R}^q$.
For any $s \in \mathbb{R}^q$, $B(s) \stackrel{\text{def}}{=} \nabla(\phi \circ \mathsf{T})(s)$ is a symmetric $q \times q$ matrix and there exist $0 < v_{min} \leq v_{max} < \infty$ such that for all $s \in \mathcal{S}$, the spectrum of $B(s)$ is in $[v_{min}, v_{max}]$.
For any $i \in [[n]]$, $\bar{s}_i \circ \mathsf{T}$ is globally Lipschitz on $\mathbb{R}^q$ with constant L_i.
$s \mapsto B^T(s) \left(\bar{s} \circ T(s) - s \right)$ is globally Lipschitz on $\mathbb{R}^q$ with constant $L_{\dot{V}}$.

Corollary: R is C^1; $\mathcal{L}_i$ is C^1; $\dot{V}$ is Lipschitz; $\|\dot{V}(s)\| \leq v_{\max} \|h(s)\|$.

Sketch of proof, $V \overset{\text{def}}{=} F \circ \mathsf{T}$

An upper bound of $\mathbb{E}\left[\|h(\widehat{S}^K)\|^2\right]$ when $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1} H_{k+1}$

- A Taylor expansion, first order + Gradient Lipschitz

$$V(\widehat{S}^{k+1}) \leq V(\widehat{S}^k) + \gamma_{k+1} \left\langle H_{k+1}, \dot{V}(\widehat{S}^k) \right\rangle + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \|H_{k+1}\|^2$$

  - The expectation

$$\mathbb{E}\left[V(\widehat{S}^{k+1})\right] \leq \mathbb{E}\left[V(\widehat{S}^k)\right] + \gamma_{k+1} \mathbb{E}\left[\left\langle h(\widehat{S}^k), \dot{V}(\widehat{S}^k) \right\rangle\right] + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \mathbb{E}\left[\|H_{k+1}\|^2\right]$$

  - The assumption (Lyapunov contraction)

$$\mathbb{E}\left[V(\widehat{S}^{k+1})\right] \leq \mathbb{E}\left[V(\widehat{S}^k)\right] - \gamma_{k+1} v_{\min} \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] + \gamma_{k+1}^2 \frac{L_{\dot{V}}}{2} \mathbb{E}\left[\|H_{k+1}\|^2\right]$$

  - A sum from $k = 0$ to $k = K_{\max} - 1$,

$$v_{\min} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1} \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] \leq \mathbb{E}\left[V(\widehat{S}^0)\right] - \mathbb{E}\left[V(\widehat{S}^{K_{\max}})\right] + \frac{L_{\dot{V}}}{2} \sum_{k=0}^{K_{\max}-1} \gamma_{k+1}^2 \mathbb{E}\left[\|H_{k+1}\|^2\right]$$

  - Few pages later:

$$\sum_{k=0}^{K_{\max}-1} \alpha_k \, \mathbb{E}\left[\|h(\widehat{S}^k)\|^2\right] \leq \mathbb{E}\left[V(\widehat{S}^0)\right] - \mathbb{E}\left[V(\widehat{S}^{K_{\max}})\right]$$

Result 1 (Gach, F., Moulines-2020)

Assume A1 to A3 and set $L^2 \overset{\text{def}}{=} n^{-1} \sum_{i=1}^n L_i^2$. Choose $\mu \in (0,1)$.

Let K be a $\{0, \ldots, K_{\max} - 1\}$-valued uniform r.v.
Run FIEM with a constant step size

$$\gamma_\ell = \frac{\sqrt{C}}{n^{2/3} L}$$

where $C \in (0,1)$ is the unique solution of "an equation depending upon $v_{\min}, L, L_{\dot{V}}, \mu$"

Then, for any $n \geq 2$ and $K_{\max} \geq 1$

$$\mathsf{E}_h \leq \frac{n^{2/3}}{K_{\max}} \frac{L}{\sqrt{C}(1-\mu) v_{\min}} \mathbb{E}\left[V(\widehat{S}^0) - V(\widehat{S}^{K_{\max}})\right] .$$

Corollaries of Result 1

$$\mathsf{E}_h \leq \frac{n^{2/3}}{K_{\max}} \frac{L}{\sqrt{C}(1-\mu)v_{\min}} \mathbb{E}\left[V(\widehat{S}^0) - V(\widehat{S}^{K_{\max}})\right] \qquad \gamma_k = \frac{\sqrt{C}}{n^{2/3}L}$$

- Dependence upon n and $K_{\max}$: as in Karimi et al. (2019)
- Constant stepsize $O(n^{-2/3})$ as in Karimi et al. (2019)
- Precision ε: $K_{\max} = M\, n^{2/3}\varepsilon^{-1}$
- $C \in (0,1)$ is explicit $\rightarrow$ "definition" of γ_k $\rightarrow$ improves on previous results

$$\sqrt{C}\frac{L_{\dot{V}}}{2L}\left(\frac{1}{n^{2/3}} + \frac{C}{1/2 - Cn^{-1/3}}\left(\frac{1}{n}+2\right)\right) = \mu v_{\min}$$

when $n \to \infty$ $\quad C^{3/2} = \frac{L}{2L_{\dot{V}}}\mu v_{\min}$

Result 2 (Gach, F., Moulines-2020)

Assume A1 to A3 and set $L^2 \overset{\text{def}}{=} n^{-1} \sum_{i=1}^n L_i^2$. Choose $\mu \in (0,1)$.

Let K be a $\{0, \dots, K_{\max} - 1\}$-valued uniform r.v.
Run FIEM with a constant step size

$$\gamma_\ell = \frac{\sqrt{C}}{n^{1/3} K_{\max}^{1/3} L}$$

where $C > 0$ is the unique solution of "an equation depending upon $v_{\min}, L, L_{\dot{V}}, \mu$"

Then, for any $n \geq 1$ and $K_{\max} \geq 1$

$$\mathsf{E}_h \leq \frac{n^{1/3}}{K_{\max}^{2/3}} \frac{L}{\sqrt{C}(1-\mu) v_{\min}} \mathbb{E}\left[V(\widehat{S}^0) - V(\widehat{S}^{K_{\max}})\right] .$$

Corollaries of Result 2

$$\mathsf{E}_h \leq \frac{n^{1/3}}{K_{\max}^{2/3}} \frac{L}{\sqrt{C}(1-\mu)v_{\min}} \mathbb{E}\left[V(\widehat{S}^0) - V(\widehat{S}^{K_{\max}})\right] \qquad \gamma_\ell = \frac{\sqrt{C}}{n^{1/3}K_{\max}^{1/3}L}.$$

- Dependence upon n and $K_{\max}$: new result.
- Constant stepsize $O(n^{-1/3}K_{\max}^{-1/3})$
- Precision ε: $K_{\max} = M\sqrt{n}\,\varepsilon^{-3/2}$ $\rightarrow$ stepsize $= O(\sqrt{\epsilon}/\sqrt{n})$
- $C \in (0,1)$ is explicit $\rightarrow$ "definition" of γ_k

Result 3 (Gach, F., Moulines-2020)

Let K be a $\{0, \dots, K_{\max} - 1\}$-valued r.v. with weights $p_0, \dots, p_{K_{\max}-1}$, $\inf_k p_k > 0$.

Let $C \in (0, 1)$ be the unique solution of "an equation depending upon $v_{\min}, L, L_{\dot{V}}, \mu$" Run FIEM with

$$\gamma_{k+1} \stackrel{\text{def}}{=} \frac{g_k}{n^{2/3} L} , \quad g_k \stackrel{\text{def}}{=} F_{n,C}^{-1} \left(\frac{p_k}{\max_\ell p_\ell} \frac{v_{\min} \sqrt{C}}{2L} \frac{1}{n^{2/3}} \right) ;$$

$$F_{n,C} : \quad x \mapsto \frac{1}{L n^{2/3}} x \left(v_{\min} - x f_n(C) \right) , \qquad f_n(C) \stackrel{\text{def}}{=} \frac{L_{\dot{V}}}{2L} \left(\frac{1}{n^{2/3}} + \frac{C}{1/2 - C n^{-1/3}} \left(\frac{1}{n} + 2 \right) \right) .$$

For any $n \geq 2$ and $K_{\max} \geq 1$, we have

$$\mathsf{E}_h \leq n^{2/3} \max_k p_k \frac{2L}{\sqrt{C} v_{\min}} \mathbb{E} \left[V(\widehat{S}^0) - V(\widehat{S}^{K_{\max}}) \right] .$$

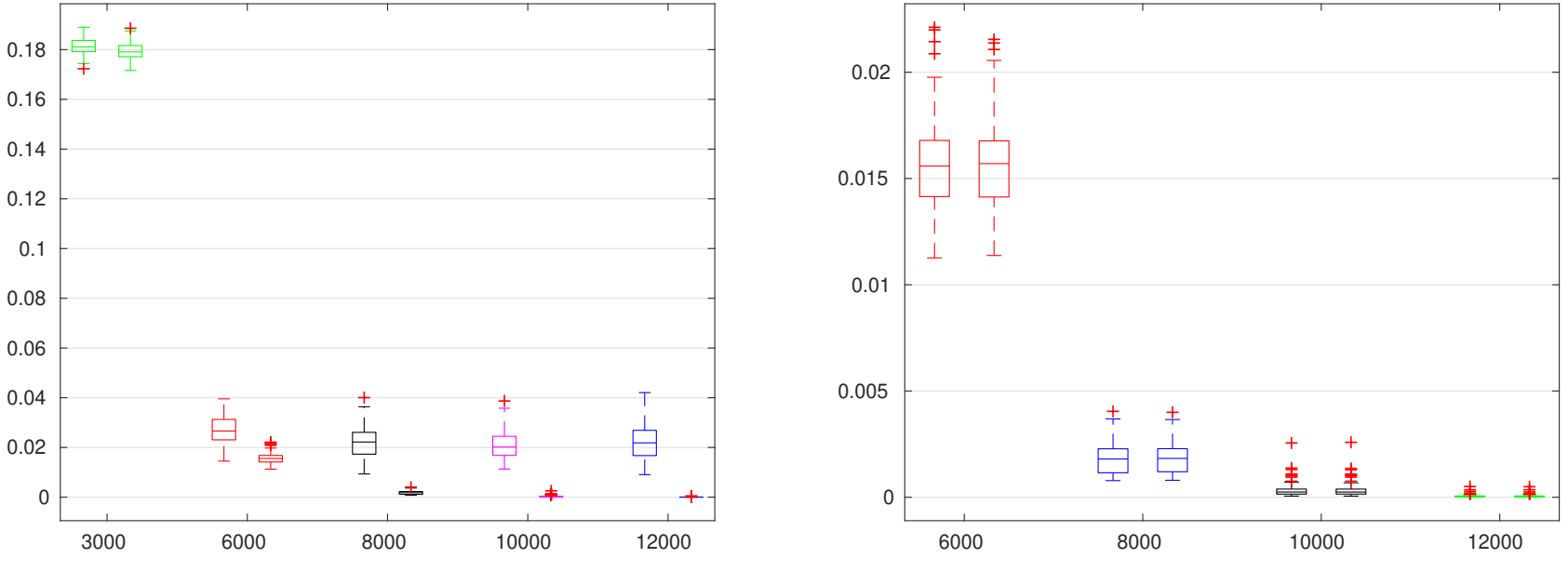
- Optimal sampling strategy : $p_k = 1/K_{\max}$ $\quad \rightarrow$ same as Result 1.
- New result.

On a toy example (1/4)

- Toy example: $\text{Argmin}_\theta F(\theta) = \theta_\star$ is explicit.

- Compare SA and FIEM (left) through boxplots of $\|\theta_k - \theta_\star\|$ for different k; same thing for FIEM and `FIEM-opt` in the convergence phase (right)

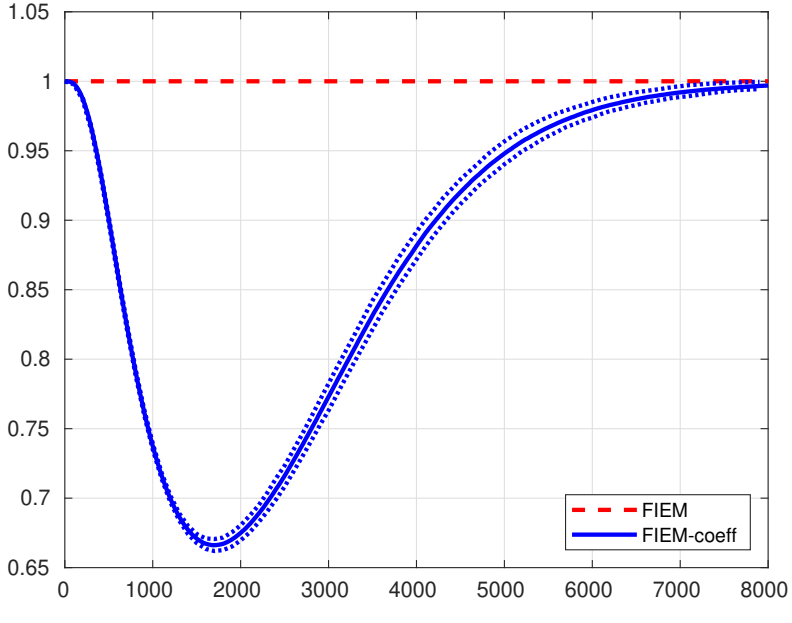
`FIEM-opt` $\widehat{S}^{k+1} = \widehat{S}^k + \gamma_{k+1}\left(T_{k+1} + \lambda_{k+1} V_{k+1}\right)$

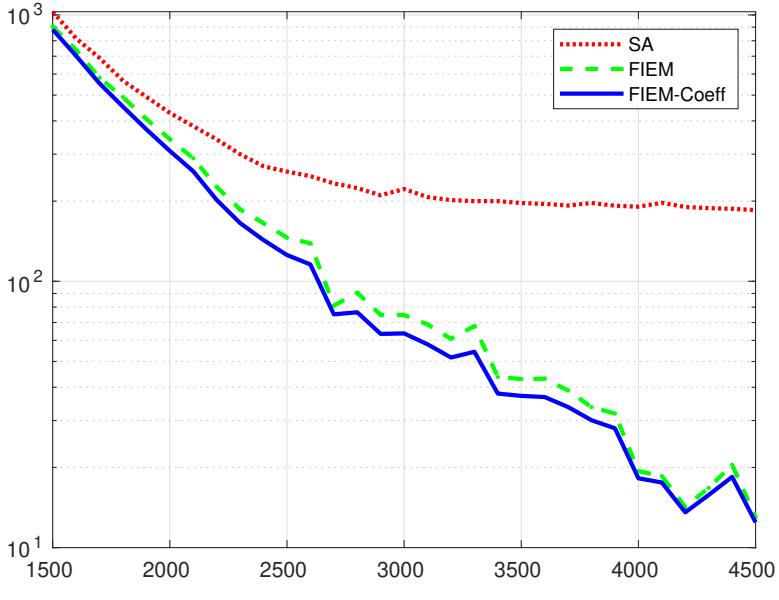


On a toy example (2/4)

- FIEM and `FIEM-opt`

(left) The coefficient $k \mapsto \lambda_{k+1}$;
(right) The L^2-moment of the field estimated by Monte Carlo: $\mathbb{E}\left[\|H_{k+1}\|^2\right]$.

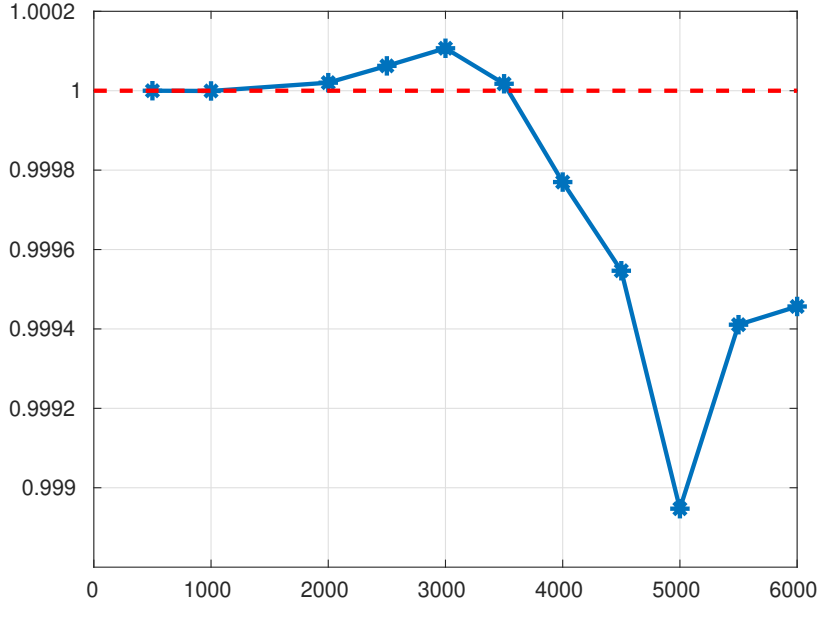


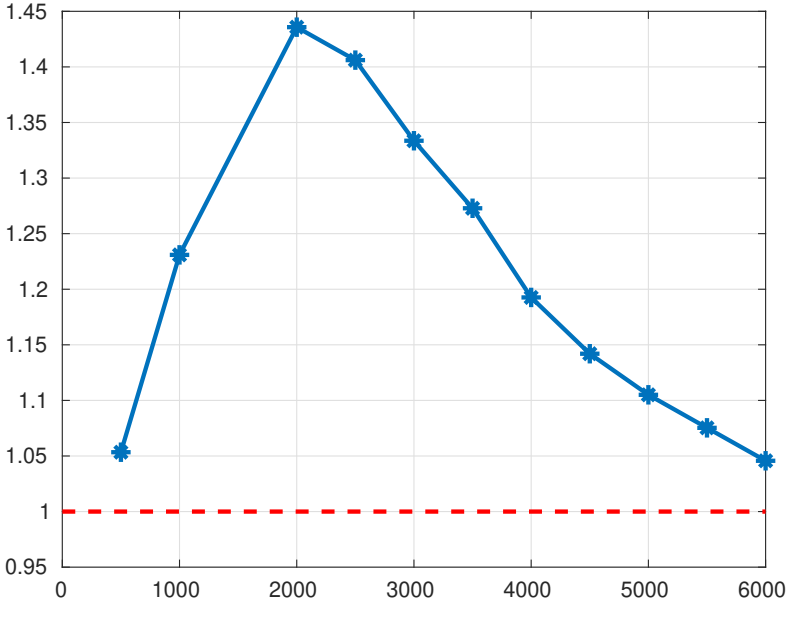


On a toy example (3/4)

- FIEM and `FIEM-opt` - Monte Carlo evaluation of $\mathbb{E}\left[\|\theta_k - \theta_\star\|\right]$ and $\mathrm{std}(\|\theta_k - \theta_\star\|)$.

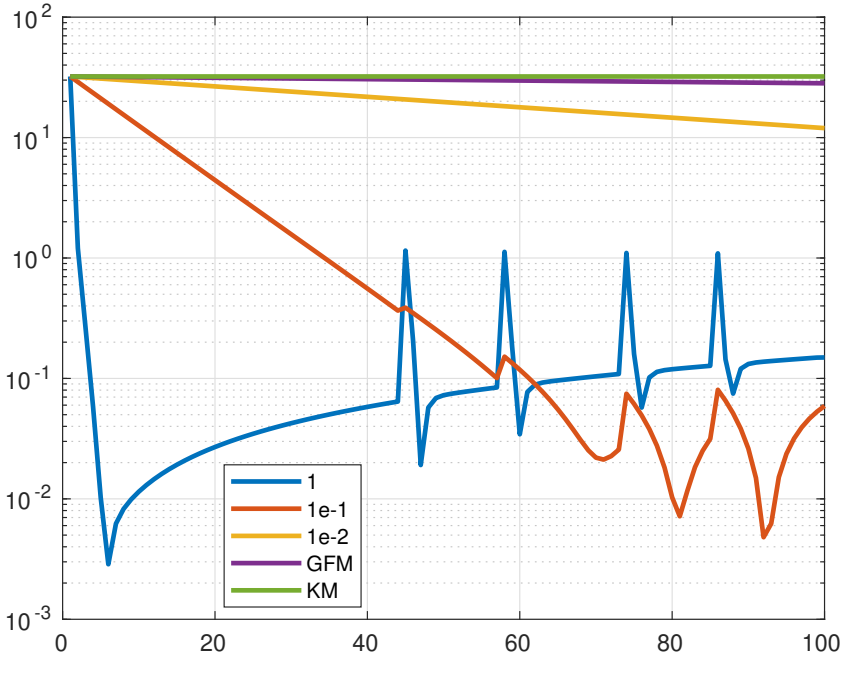
(left) ratio of the expectations FIEM / `FIEM-opt`
(right) ratio of the std FIEM / `FIEM-opt`

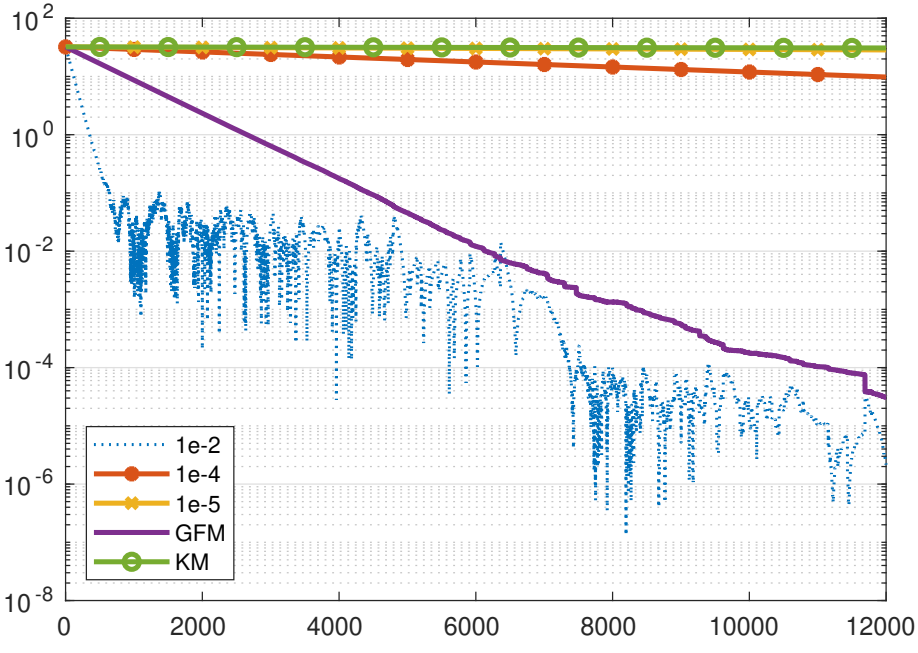




On a toy example (4/4)

- For FIEM, constant stepsize γ, observe $k \mapsto \|\theta_k - \theta_\star\|$





.

IV- What happens when $\overline{s}_i$ is not explicit ?

.

The perturbed case: when the expectation are approximated,
which conditions on the approximations in order to have the same rates as in the "exact" case?

The Perturbed FIEM algorithm

Data: $\widehat{S}^0 \in \mathcal{S}$, $\gamma_k \in (0, \infty)$ for $k \geq 1$

Result: The FIEM sequence: $\widehat{S}^k, k = 0, \ldots,$

$\mathsf{S}_{0,i} = \tilde{s}_i$, an approximation of $\bar{s}_i \circ \mathsf{T}(\widehat{S}^0)$ for all $i \in [[n]]$;

$\widetilde{S}^0 = n^{-1} \sum_{i=1}^n \mathsf{S}_{0,i}$;

for $k \geq 1$ **do**

* Sample I_{k+1} uniformly on $[[n]]$;
* $\mathsf{S}_{k+1,i} = \mathsf{S}_{k,i}$ for $i \neq I_{k+1}$;
* $\mathsf{S}_{k+1,I_{k+1}} = \check{\mathsf{S}}_{k+1}$ an approximation of $\bar{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$;
* $\widetilde{S}^{k+1} = \widetilde{S}^k + n^{-1} \left(\mathsf{S}_{k+1,I_{k+1}} - \mathsf{S}_{k,I_{k+1}} \right)$;

Sample J_{k+1} uniformly on $[[n]]$;

Compute $\tilde{s}_{k+1}$, an approximation of $\bar{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k)$;

$\widehat{S}^{k+1} = \widehat{S}^k - \gamma_{k+1} \left(\tilde{s}_{k+1} - \widehat{S}^k - \left\{ \mathsf{S}_{k+1,J_{k+1}} - \widetilde{S}^{k+1} \right\} \right)$

Assumptions (on the perturbations)

There exist positive sequences $\{m_k, k \geq 0\}$ and $\{\overline{m}_k, k \geq 0\}$, positive numbers $M^{(1)}$ and $M^{(2)}$ and $M_\nu^{(2)} \geq 0$ such that for all $k \geq 0$, the approximations $\tilde{s}_{k+1}$ and $\check{\mathsf{S}}_{k+1}$ satisfy

$$\mathbb{E}[\|\check{\mathsf{S}}_{k+1} - \overline{s}_{I_{k+1}} \circ \mathsf{T}(\widehat{S}^k)\|^2] \leq \frac{M^{(1)}}{\overline{m}_{k+1}},$$

$$\mathbb{E}[\| \mathbb{E}\left[\tilde{s}_{k+1} - \overline{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k) | J_{k+1}, I_{k+1}, \widehat{S}^k\right] \|^2] \leq \frac{M_\nu^{(2)}}{m_{k+1}^2},$$

$$\mathbb{E}[\|\tilde{s}_{k+1} - \overline{s}_{J_{k+1}} \circ \mathsf{T}(\widehat{S}^k)\|^2] \leq \frac{M^{(2)}}{m_{k+1}}\,.$$

Well $\cdots$ in a Monte Carlo approximation

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\frac{1}{\overline{m}_{k+1}} \sum_{\ell=1}^{\overline{m}_{k+1}} s_i(Z_{\ell,k}) - \overline{s}_i \circ \mathsf{T}(\widehat{S}^k)\|^2] \leq \frac{\mathbb{E}\left[\mathsf{C}(\widehat{S}^k)\right]}{\overline{m}_{k+1}} \leq \frac{M^{(1)}}{\overline{m}_{k+1}},$$

Convergence result (Gach, F., Moulines (2020))

$$
\begin{aligned}
\mathsf{E}_h \leq & \frac{n^{2/3}}{K_{\max}} \mathcal{C}_0 \\
& + \frac{\mathcal{C}_1}{n^{2/3}} \{1 \wedge \frac{n}{K_{\max}}\} \mathbb{E}\left[\varepsilon^{(0)}\right] && \text{error when initializing} \\
& + \frac{\mathcal{C}_1}{n^{5/3}} \{1 \wedge \frac{n}{K_{\max}}\} \sum_{k=0}^{K_{\max}-1} \mathbb{E}\left[\|\eta_{k+1}^{(1)}\|^2\right] && \text{Error: on the control variate} \\
& + \mathcal{C}_2 \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}\left[\|\mathbb{E}\left[\eta_{k+1}^{(2)} | \mathcal{F}_k\right]\|^2\right] && \eta^{(2)}\text{: second error} \\
& + \frac{\mathcal{C}_1}{2(1+\nu)} \frac{1}{n^{2/3} K_{\max}} \sum_{k=0}^{K_{\max}-1} \left(\mathbb{E}\left[\|\eta_{k+1}^{(2)}\|^2\right] + \mathbb{E}\left[\|\mathbb{E}\left[\eta_{k+1}^{(2)} | \mathcal{F}_{k+3/4}\right]\|^2\right] \right);
\end{aligned}
$$

In the case of a Monte Carlo approximation

For a precision ε,

- with an **unbiased** Monte Carlo approximation (ex. i.i.d.)

$+ \; K_{\max} = M n^{2/3} \; \varepsilon^{-1}$

$+ \; m = n^{-2/3} \varepsilon^{-1} \qquad \overline{m} = (n^{-2/3} \varepsilon^{-1}) \wedge (n^{-1} \varepsilon^{-2})$

- with a **biased** Monte Carlo approximation (ex. Markov chain Monte Carlo)

$+ \; K_{\max} = M n^{2/3} \; \varepsilon^{-1}$

$+ \; m = (n^{-2/3} \varepsilon^{-1}) \vee \varepsilon^{-1/2} \qquad \overline{m} = (n^{-2/3} \varepsilon^{-1}) \wedge (n^{-1} \varepsilon^{-2})$