Geom-SPIDER-EM: Faster Variance Reduced Stochastic Expectation Maximization for Nonconvex Finite-Sum Optimization

Gersende Fort (IMT, CNRS, France)
Eric Moulines (CMAP, Ecole Polytechnique, France)
Hoi-To Wai (SEEM, Chinese Univ. of Hong Kong, Hong-Kong)

ICASSP 2021

In this talk

- A novel EM algorithm: `Geom-SPIDER-EM`
- Adapted to the finite sum setting (large number of examples n)
- Stochastic: it combines
  - the stochastic approximation method
  - a variance reduction technique
- Same complexity as `SPIDER-EM` (Fort et al, 2020) – state of the art, among the incremental EM's.

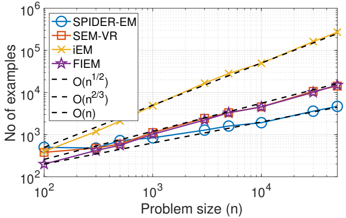


Figure: Nbr of processed examples required to reach convergence, as a function of the problem size n. From Fort et al. (2020, NeurIPS)

Optimization problem: finite sum setting, for curved exponential families

- Solve on $\Theta \subseteq \mathbb{R}^d$ the minimization problem

$$\mathrm{argmin}_{\theta \in \Theta} - \sum_{i=1}^{n} \log \int_{\mathsf{Z}} p_i(z;\theta) \mathsf{d}\mu(z) + \mathsf{R}(\theta), \qquad p_i(z;\theta) > 0$$

- Curved exponential family:

$$- \sum_{i=1}^{n} \log \int_{\mathsf{Z}} h_i(z_i) \exp\left(\langle s_i(z_i), \phi(\theta) \rangle\right) \mathsf{d}\mu(z_i) + \mathsf{R}(\theta)$$

- In computational Statistics: minimization of the (penalized) negative likelihood in *latent variable* models:
  - finite sum setting when the observations are independent.
  - $p_i \equiv p_{\mathbf{Y_i}}(z_i;\theta)$ is the complete data likelihood of the pair $\#i$: (Y_i, Z_i)
  - Curved exponential family: e.g. mixture of curved exponential distributions.

From EM to incremental EM

Objective function:

$$-\sum_{i=1}^{n} \log \int_{\mathsf{Z}} p_i(z;\theta) \mathrm{d}\mu(z_i) + \mathrm{R}(\theta), \qquad p_i(z;\theta) = h_i(z_i)\, \exp\left(\langle s_i(z_i), \phi(\theta)\rangle\right)$$

- **EM algorithm:** Repeat for $t = 0, \ldots$

E-step $\quad \bar{\mathsf{s}}(\theta_t) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{s}}_i(\theta_t) \qquad$ where $\bar{\mathsf{s}}_i(\theta) = \int_{\mathsf{Z}} \mathsf{s}_i(z) \frac{p_i(z;\theta)}{\int p_i(u;\theta)\mathrm{d}\mu(u)} \mathrm{d}\mu(z)$

M-step $\quad \theta_{t+1} = \mathsf{T}\left(\bar{\mathsf{s}}(\theta_t)\right)$

where

$$\mathsf{T}(s) = \mathrm{argmin}_{\theta\in\Theta} \; \mathrm{R}(\theta) - \langle s, \phi(\theta)\rangle$$

E-step $\rightarrow$ sum over n expectations $\rightarrow$ Large computational cost of each EM iteration, when n is large !

• Given a computational budget, what is the best strategy: few iterations of EM or many iterations of *incremental EM* ?

Incremental EM algorithms in the expectation space

- EM: an algorithm in the *expectation space*

$$\theta_{t+1} = \mathsf{T} \circ \bar{\mathsf{s}}(\theta_t) = \mathsf{T} \circ \underbrace{\bar{\mathsf{s}} \circ \mathsf{T}} \circ \bar{\mathsf{s}} \ldots \underbrace{\bar{\mathsf{s}} \circ \mathsf{T}} \circ \bar{\mathsf{s}}(\theta_0)$$

$$S_{t+1} = \bar{\mathsf{s}} \circ \mathsf{T}(S_t) = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathsf{s}}_i \circ \mathsf{T}(S_t)$$

- EM designed to find the roots of

$$\mathsf{h}(s) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \bar{\mathsf{s}}_i \circ \mathsf{T}(s) - s = \mathbb{E}\left[\bar{\mathsf{s}}_I(s) - s + V\right]$$

where $I \sim \mathcal{U}(\{1, \ldots, n\})$ and V is a *control variate* i.e. r.v. correlated with $\bar{\mathsf{s}}_I$ and centered.

- Stochastic Approximation The algorithm

$$\widehat{S}_{t+1} = \widehat{S}_t + \gamma_{t+1} H_{t+1} \qquad \mathbb{E}\left[H_{t+1} | \text{past}_t\right] = \mathsf{h}(\widehat{S}_t)$$

has the **same** limiting set: $\{s : \mathsf{h}(s) = 0\}$.

Variance reduced incremental EM

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

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

- Online-EM (Neal and Hinton, 1998; Cappé and Moulines, 2009). NO variance reduction $(V_{t+1} = 0)$.
- sEM-vr: Stochastic Expectation Maximization with Variance Reduction Chen et al, 2018
- FIEM: Fast Increment Expectation Maximization Karimi et al, 2019; Fort et al, 2021
- SPIDER-EM Fort et al, 2020 and Geom-SPIDER-EM: Stochastic Path Integrated Differential EstimatoR Expectation Maximization

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

$$= V_0 + \sum_{\ell=0}^{t} \left\{ \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_\ell} \bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_{\ell-1}) - \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{\ell+1}} \bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_{\ell-1}) \right\}$$

Geom-SPIDER-EM (Stochastic Path Integrated Differential EstimatoR)

1: $\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\text{init}}$ $\quad$ $\mathsf{S}_{1,0} = \bar{\mathsf{s}} \circ \mathsf{T}(\widehat{S}_{1,-1}) + \mathcal{E}_1$
2: **for** $t = 1, \cdots, k_{\text{out}}$ **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: $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \Big(\bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_{t,k}) - \bar{\mathsf{s}}_i \circ \mathsf{T}(\widehat{S}_{t,k-1}) \Big)$
6: $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1} \left(\mathsf{S}_{t,k+1} - \widehat{S}_{t,k} \right)$
7: **end for**
8: $\widehat{S}_{t+1,-1} = \widehat{S}_{t,\xi_t}$
9: $\mathsf{S}_{t+1,0} = \bar{\mathsf{s}} \circ \mathsf{T}(\widehat{S}_{t+1,-1}) + \mathcal{E}_{t+1}$ $\quad$ $\mathcal{E}_{t+1}$: a possible error
10: $\widehat{S}_{t+1,0} = \widehat{S}_{t+1,-1} + \gamma_{t+1,0} \left(\mathsf{S}_{t+1,0} - \widehat{S}_{t+1,-1} \right)$
11: **end for**

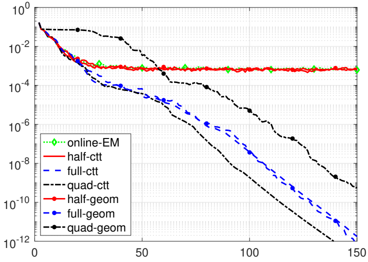
The control variate is refreshed at each *outer loop* $\#t$ (see Line 9)
The length of the outer loop is a Geometric random variable ξ_t

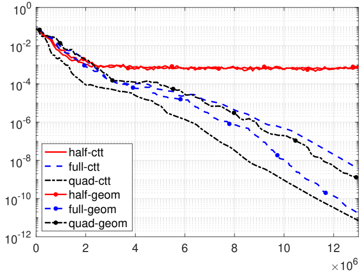
Application: inference in GMM (from the MNIST data set) (1/2)

Gaussian mixture models in $\mathbb{R}^{20}$; $G = 12$ components; $n = 6\,10^4$ examples

Displayed: 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)

Remember: $\mathcal{L} = \{s : \bar{\mathsf{s}} \circ \mathsf{T}(s) - s = 0\}$ is the limiting set of EM in the expectation space.



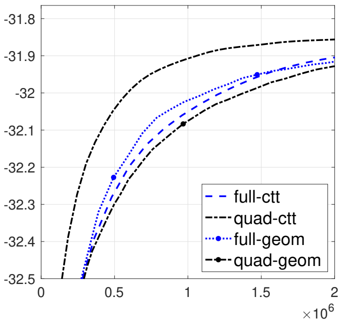


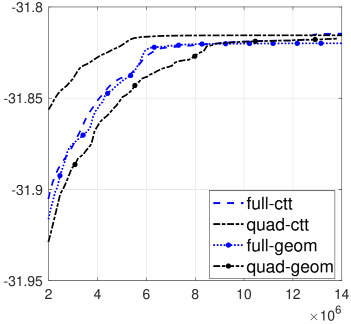
Length of each outer: either constant (`ctt`) $\xi_t = k_{\text{in}}$, or a geometric r.v. (`geom`) with expectation k_{in}

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`).

Application: inference in GMM (from the MNIST data set) (2/2)

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





Complexity for ϵ-approximate stationarity

We provide an **explicit** expression of an upper bound for

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

- in the non convex setting
- at the end of an outer loop $\#\tau$ where τ is sampled unif. in $\{1, \cdots, k_{\text{out}}\}$
- as a function of $k_{\text{out}}, \mathsf{b}, n$ and the learning rate γ ($= \gamma_{t,k}$ for any $t, k > 0$) and the expectation k_{in} of ξ_t.

To reach ϵ-stationarity, the complexity of Geom-SPIDER-EM

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 $\bar{\mathsf{s}}_i$*'s evaluations:* $\mathcal{K} = O(\sqrt{n}\,\epsilon^{-1})$

For Online EM: $\mathcal{K} = O(\epsilon^{-2})$
For sEM-vr: $\mathcal{K} = O(n^{2/3}\,\epsilon^{-1})$
For FIEM: $\mathcal{K} = O(n^{2/3}\,\epsilon^{-1} \wedge \sqrt{n}\epsilon^{-3/2})$
For SPIDER-EM: $\mathcal{K} = O(\sqrt{n}\,\epsilon^{-1})$