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

Gersende Fort[1], Eric Moulines[2] and Hoi-To Wai[3]

CNRS, France[1], École Polytechnique, France[2], Chinese University of Hong Kong, Hong-Kong[3]

Contribution

- A novel EM algorithm
- Adapted to the finite sum setting
- Stochastic: it combines Stochastic Approximation and variance reduction techniques
- Same complexity as `SPIDER-EM` (Fort et al, 2020b) – state of the art, among the incremental EM's.

Optimization problem

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

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

- Curved exponential family:

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

- In computational Statistics: minimization of the (penalized) negative likelihood in *latent variable* models.

From EM to incremental EM

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

$$\text{E-step} \qquad \bar{\mathrm{s}}(\theta_t) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathrm{s}}_i(\theta_t)$$

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

where

$$\bar{\mathrm{s}}_i(\theta) = \int_Z \mathrm{s}_i(z) \frac{p_i(z;\theta)}{\int p_i(u;\theta)\mathrm{d}\mu(u)} \mathrm{d}\mu(z)$$

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

- EM: an algorithm in the *expectation space* (Delyon et al, 1999)

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

- EM designed to find the roots of

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

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

- Stochastic Approximation The algorithm

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

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

Variance reduced incremental EM

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

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

- 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, 2020a
- SPIDER-EM Fort et al, 2020b and Geom-SPIDER-EM: Stochastic Path Integrated Differential EstimatoR Expectation Maximization

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

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

1: $\widehat{S}_{1,0} = \widehat{S}_{1,-1} = \widehat{S}_{\mathrm{init}} \qquad \mathrm{S}_{1,0} = \bar{\mathrm{s}} \circ \mathrm{T}(\widehat{S}_{1,-1}) + \mathcal{E}_1$
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: $\mathrm{S}_{t,k+1} = \mathrm{S}_{t,k} + \mathrm{b}^{-1}\sum_{i\in\mathcal{B}_{t,k+1}}\left(\bar{\mathrm{s}}_i \circ \mathrm{T}(\widehat{S}_{t,k}) - \bar{\mathrm{s}}_i \circ \mathrm{T}(\widehat{S}_{t,k-1})\right)$
6: $\widehat{S}_{t,k+1} = \widehat{S}_{t,k} + \gamma_{t,k+1}\left(\mathrm{S}_{t,k+1} - \widehat{S}_{t,k}\right)$
7: **end for**
8: $\widehat{S}_{t+1,-1} = \widehat{S}_{t,\xi_t}$
9: $\mathrm{S}_{t+1,0} = \bar{\mathrm{s}} \circ \mathrm{T}(\widehat{S}_{t+1,-1}) + \mathcal{E}_{t+1}$ $\qquad \mathcal{E}_{t+1}$: an error (e.g. part of the data set)
10: $\widehat{S}_{t+1,0} = \widehat{S}_{t+1,-1} + \gamma_{t+1,0}\left(\mathrm{S}_{t+1,0} - \widehat{S}_{t+1,-1}\right)$
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 (Li et al, 2020; Horvath et al., 2020)

Complexity for ϵ-approximate stationarity

We provide an **explicit** expression of an upper bound for $\qquad \mathbb{E}\left[\|\mathrm{h}(\widehat{S}_{\tau,\xi_\tau})\|^2\right]$

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

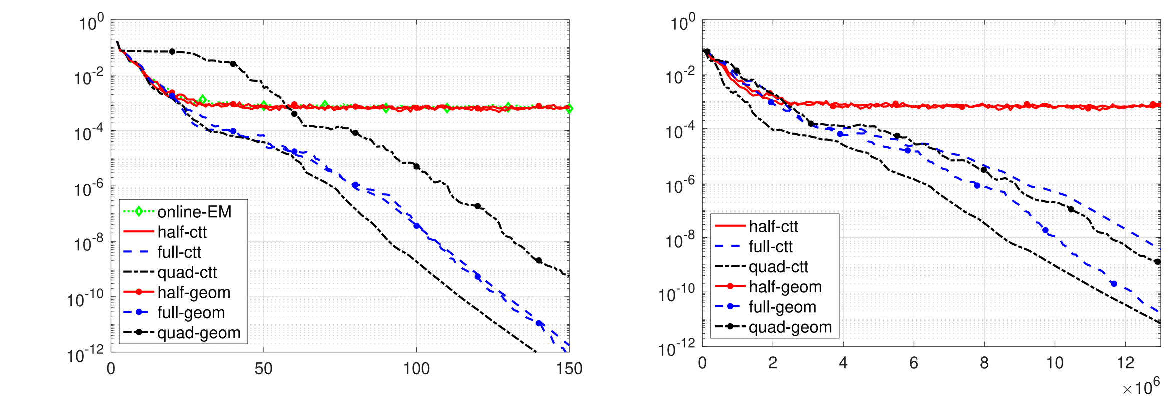
With: $k_{\mathrm{in}} = \mathrm{b} = O(\sqrt{n})$, $\quad k_{\mathrm{out}} = O(1/(\epsilon k_{\mathrm{in}}))$
Nbr of optimization steps: $O(1/\epsilon)$ $\qquad$ Nbr of $\bar{\mathrm{s}}_i$'s evaluations: $\mathcal{K} = O(\sqrt{n}\,\epsilon^{-1})$

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

Inference in Gaussian Mixture Models (from the MNIST data set)

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

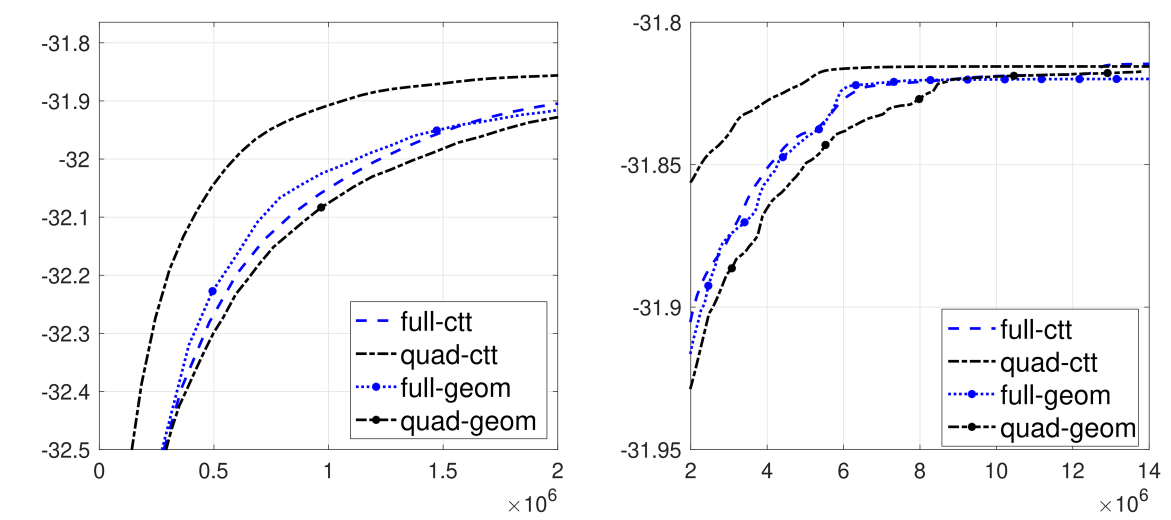
▶ Displayed: quantile of order 0.5 of $\|\mathrm{h}(\widehat{S}_{t,\xi_t})\|^2$ vs the number of epochs (left) and vs the number of $\bar{\mathrm{s}}_i$'s evaluations (right)



Length of each outer: either constant (`ctt`) $\xi_t = k_{\mathrm{in}}$, 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`).

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



References

O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):593–613, 2009.

J. Chen, J. Zhu, Y.W. Teh, and T. Zhang. Stochastic expectation maximization with variance reduction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7967–7977. 2018.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a Stochastic Approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999.

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

G. Fort, E. Moulines, and P. Gach. Fast Incremental Expectation Maximization for finite-sum optimization: nonasymptotic convergence. Technical report, HAL 02617725v2, 2020a.

G. Fort, E. Moulines, and H.T. Wai. A Stochastic Path-Integrated Differential EstimatoR Expectation Maximization Algorithm. In *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc., 2020b.

S. Horvath, L. Lei, P. Richtarik, and M.I. Jordan. Adaptivity of Stochastic Gradient Methods for Nonconvex Optimization. Technical report, arXiv 2002.05359, 2020.

B. Karimi, H.-T. Wai, E. Moulines, and M. Lavielle. On the Global Convergence of (Fast) Incremental Expectation Maximization Methods. In *NeurIPS 32*. 2019.

Z. Li, H. Bao, X. Zhang, and P. Richtrik. PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization. Technical report, arXiv 2008.10898, 2020.

R. M. Neal and G. E. Hinton. *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*, pages 355–368. Springer Netherlands, Dordrecht, 1998.

ICASSP 2021, Paper #3455