On the Global Convergence of (Fast) Incremental EM Methods

Belhal Karimi[1,2], Hoi-To Wai[3], Eric Moulines[2] and Marc Lavielle[1]

INRIA[1], École Polytechnique[2], Chinese University of Hong Kong[3]

Maximum Likelihood Estimation (MLE)

- Given a set of n observations $y = (y_i, i \in [n])$.
- Goal: fitting the parametric model $g(\cdot, \boldsymbol{\theta})$.
- **Maximum Likelihood Estimation of $\boldsymbol{\theta}$**

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log g(y_i, \boldsymbol{\theta})$$

- $g(y, \boldsymbol{\theta}) = \int_Z f(z, y, \boldsymbol{\theta}) \mu(dz)$ is a parametric model with a latent variable z – the function is generally intractable.
- We use the **EM** algorithm which takes advantage of the latent structure.

Settings and Notation

- *Regularized Empirical Risk Minimization*:

$$\min_{\boldsymbol{\theta} \in \Theta} \overline{\mathcal{L}}(\boldsymbol{\theta}) := \mathrm{R}(\boldsymbol{\theta}) + \mathcal{L}(\boldsymbol{\theta})$$

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \{ -\log g(y_i; \boldsymbol{\theta}) \}$$

- $\overline{\mathcal{L}}$ is possibly **nonconvex** and lower bounded
- For all $i \in [n]$, $f(z_i, y_i, \boldsymbol{\theta})$ and $p(z_i|y_i, \boldsymbol{\theta})$ denote the complete likelihood and the posterior distribution
- Focus on the **Exponential Family Distribution**:

$$f(z_i, y_i, \boldsymbol{\theta}) = h(z_i, y_i) \exp\left(\langle S(z_i, y_i) \,|\, \phi(\boldsymbol{\theta}) \rangle - \psi(\boldsymbol{\theta}) \right)$$

EM for Exponential Family

- We define the following EM-related operations.

★ **E-operation**: for any $\boldsymbol{\theta} \in \Theta$,

$$\overline{\mathbf{s}}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \overline{\mathbf{s}}_i(\boldsymbol{\theta})$$

also, define $\overline{\mathbf{s}}_i(\boldsymbol{\theta}) := \int_Z S(z_i, y_i) p(z_i|y_i; \boldsymbol{\theta}) \mu(\mathrm{d}z_i)$.

★ **M-operation**: for any $\hat{\mathbf{s}}$,

$$\hat{\boldsymbol{\theta}} := \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}) := \arg\min_{\boldsymbol{\theta} \in \Theta} \left\{ \mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \langle \hat{\mathbf{s}} \,|\, \phi(\boldsymbol{\theta}) \rangle \right\}$$

Batch EM Algorithm: given $\hat{\boldsymbol{\theta}}^{(0)}$, set $k = 0$,

1. (E-step) $\hat{\mathbf{s}}^{(k+1)} = \overline{\mathbf{s}}(\hat{\boldsymbol{\theta}}^{(k)})$
2. (M-step) $\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$

For large n, the E-step is computationally expensive!

General Formulation of Stochastic EM (sEM)

Idea: We replace the **E-step** with a **stochastic/incremental E-step (sE-step)** that looks at 1 sample only.

★ **sE-step** (general form): w/ const. stepsize γ,

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} - \gamma(\hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)})$$

- iEM: $\mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} + \frac{1}{n}(\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\tau_{i_k}^k)})$
- sEM-VR: $\mathcal{S}^{(k+1)} = \overline{\mathbf{s}}^{(\ell(k))} + (\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\ell(k))})$
- fiEM: $\mathcal{S}^{(k+1)} = \overline{\mathcal{S}}^{(k)} + (\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(t_{i_k}^k)})$ and
  $\overline{\mathcal{S}}^{(k+1)} = \overline{\mathcal{S}}^{(k)} + n^{-1}(\overline{\mathbf{s}}_{j_k}^{(k)} - \overline{\mathbf{s}}_{j_k}^{(t_{j_k}^k)})$

- Set the termination number $K \sim \mathcal{U}\{1, 2, \ldots, K_{\max}\}$
- For $k > 0$:
  - Draw index $i_k \in [n]$ uniformly (and $j_k \in [n]$ for fiEM).
  - Compute the surrogate sufficient statistics $\mathcal{S}^{(k+1)}$.
  - Compute $\hat{\mathbf{s}}^{(k+1)}$ via the sE-step (**see the left**).
  - Compute $\hat{\boldsymbol{\theta}}^{(k+1)}$ via the M-step (**same as batch EM**):

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$$

- **Return**: $\hat{\boldsymbol{\theta}}^{(K)}$.

- **Prior works**: iEM–Neal & Hinton, 1998; sEM-VR–Chen et al., 2018 (local convergence); fiEM–inspired by SAGA.

How to analyze their global convergence? And which algorithm is faster?

iEM as an incremental MM Scheme

- iEM can be interpreted as **incremental MM** (Mairal, 2015) with the **upper bound surrogate function**:

$$Q_i(\boldsymbol{\theta}; \boldsymbol{\theta}') := -\int_Z \{\log f(z_i, y_i; \boldsymbol{\theta}) - \log p(z_i|y_i; \boldsymbol{\theta}')\} \, p(z_i|y_i; \boldsymbol{\theta}') \mu(\mathrm{d}z_i), \quad \text{where} \quad Q_i(\boldsymbol{\theta}; \boldsymbol{\theta}') \geq \mathcal{L}_i(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta}.$$

- **Incremental MM**: at every iteration k, we obtain $\hat{\boldsymbol{\theta}}^{(k+1)} = \arg\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\tau_i^k)})$.
- **Convergence Analysis**: with exponential family model, $Q_i(\boldsymbol{\theta}; \boldsymbol{\theta}') - \mathcal{L}_i(\boldsymbol{\theta})$ **is L_e-smooth for all i**:

Theorem (iEM) For any $K_{\max} \geq 1$, $K \sim \mathcal{U}([0, K_{\max} - 1])$ independent of the $\{i_k\}_{k=0}^{K_{\max}}$, we have the **global rate**:

$$\mathbb{E}[\|\nabla \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(K)})\|^2] \leq n \frac{2\,\mathrm{L}_e}{K_{\max}} \mathbb{E}[\overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(0)}) - \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(K_{\max})})], \tag{1}$$

sEM-VR/fiEM are Scaled Gradient Methods and Faster than iEM

- Unlike iEM, the sEM-VR and fiEM methods can be analyzed as **scaled gradients** methods. Consider:

$$\min_{\mathbf{s} \in \mathsf{S}} V(\mathbf{s}) := \overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) = \mathrm{R}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})).$$

- **Variance-reduced scaled gradient**: the sE-step update $\hat{\mathbf{s}}^{(k)}$ by $\hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)}$, we can show

$$\langle \mathbb{E}[\hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)}] \,|\, \nabla V(\hat{\mathbf{s}}^{(k)}) \rangle \geq \upsilon_1 \|\mathbb{E}[\hat{\mathbf{s}}^{(k)} - \mathcal{S}^{(k+1)}]\|^2 \geq \upsilon_2 \|\nabla V(\hat{\mathbf{s}}^{(k)})\|^2 \quad \text{for some } \upsilon_1, \upsilon_2 > 0.$$

$\therefore$ the sEM-VR/fiEM methods are **variance-reduced, scaled gradient** updates of the sufficient statistics.

- **Convergence Analysis**: with exponential family model, **the function $V(\mathbf{s})$ is $\overline{L}_v$-smooth**,

Theorem (sEM-VR) $\gamma = \frac{\mu \upsilon_{\min}}{L_v n^{2/3}}$ & epoch $m = \frac{n}{2\mu^2 \upsilon_{\min}^2 + \mu}$:

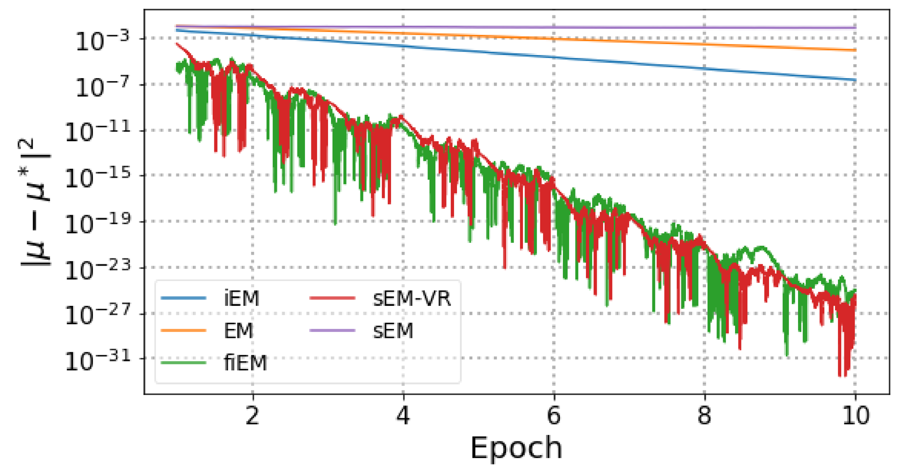
$$\mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(K)})\|^2] \leq n^{\frac{2}{3}} \frac{2\overline{L}_v}{\mu K_{\max}} \frac{\upsilon_{\max}^2}{\upsilon_{\min}^2} \mathbb{E}[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})].$$

Theorem (fiEM) $\gamma = \frac{\upsilon_{\min}}{\alpha L_v n^{2/3}}$ & $\alpha = \max\{6, 1 + 4\upsilon_{\min}\}$:

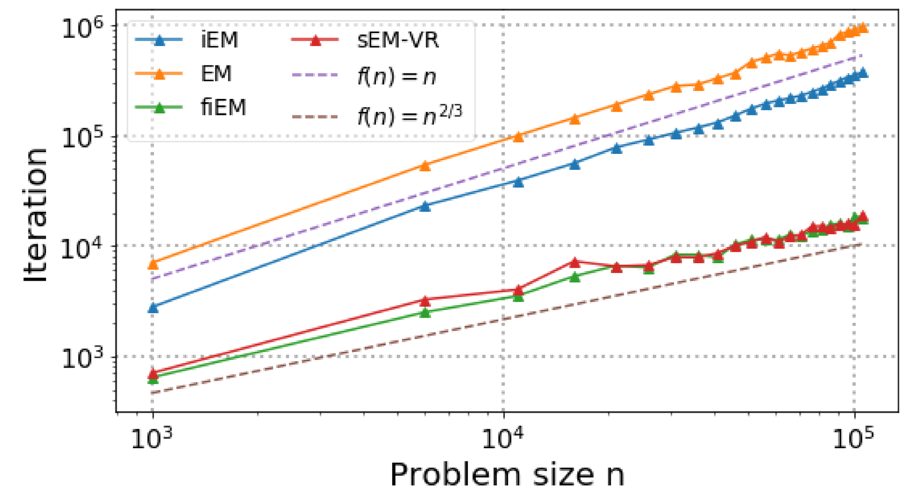
$$\mathbb{E}[\|\nabla V(\hat{\mathbf{s}}^{(K)})\|^2] \leq n^{\frac{2}{3}} \frac{\alpha^2 \overline{L}_v \upsilon_{\max}^2}{K_{\max} \upsilon_{\min}^2} \mathbb{E}[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\max})})].$$

Fitting a Gaussian Mixture Model

- **Goal**: fitting a GMM with a penalization.
- Faster rate for sEM-VR and fiEM ($n = 10^5$)



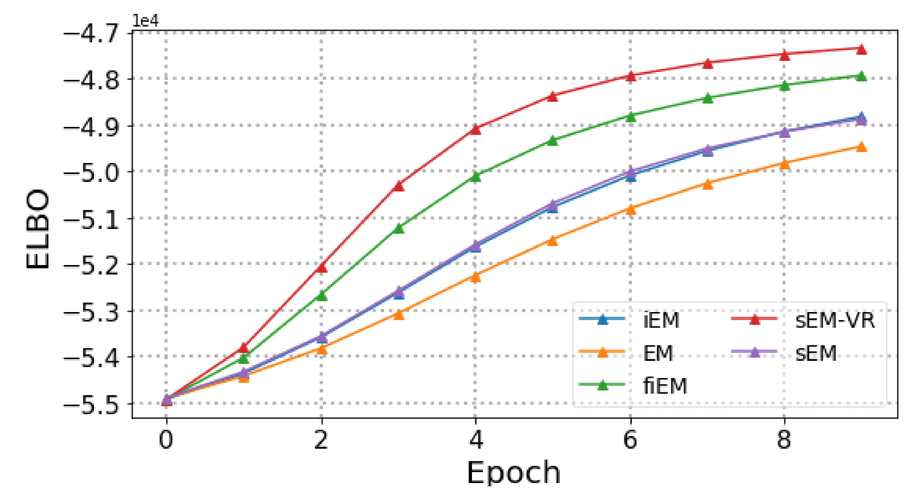
- Iteration number to reach ϵ-accuracy vs. n.



- Reveal the linear $[\mathcal{O}(n/\epsilon)$ for iEM$]$ and sublinear $[\mathcal{O}(n^{\frac{2}{3}}/\epsilon)$ for sEM-VR/fiEM$]$ rates

Probabilistic Latent Semantic Analysis (PLSA)

- **Goal**: Classifying D docs into K topics
- FAO (UN Food and Agriculture Organization) datasets.



References

Jianfei Chen, Jun Zhu, Yee Whye Teh, and Tong Zhang. Stochastic expectation maximization with variance reduction. In *NeurIPS*, 2018.

Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 2015.

Radford M Neal and Geoffrey E Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

Advances in Neural Information Processing Systems 32, Vancouver, Canada