On the Global Convergence of (Fast) Incremental EM Methods

Belhal Karimi, Hoi-To Wai, Eric Moulines, Marc Lavielle

Maximum Likelihood Estimation (MLE)

- We have vectors of data Y that are *observed* and Z that are *latent*
- We assume a probabilistic model on the observations Y, $g(Y, \theta)$
- We can define $f(Z, Y, \theta)$ as the complete data likelihood and $p(Z|Y, \theta)$ as the conditional distribution of Z given Y
- The MLE problem is, given a model $g(Y, \theta)$ and some actual data Y, find the parameter θ which makes the data most likely:

$$\theta^{ML} := \arg\max_{\theta} g(Y, \theta)$$

- This problem is an **optimization problem**, which we could use any imaginable tool to solve
- In practice, it's often **hard** to get expressions for the **derivatives** needed by **gradient** methods
- **Expectation-Maximization (EM)** method is one popular and powerful way of proceeding, but not the only way. **It takes advantage of the latent data to complete the observations.**

Context

Settings and Notations

- Many problems in machine learning pertain to tackling an empirical risk minimization of the form

$$\min_{\theta \in \Theta} \overline{\mathcal{L}}(\theta) := \mathcal{L}(\theta) + \mathrm{R}(\theta) \qquad \text{with} \qquad \mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\theta) := \frac{1}{n}\sum_{i=1}^{n} \left\{-\log g\left(y_i;\theta\right)\right\}$$

- $\{y_i\}_{i=1}^n$ are the observations, Θ is a convex subset of $\mathbb{R}^d$, $\mathrm{R}(\theta)$ is a smooth convex regularization function.

- The objective function $\overline{\mathcal{L}}(\theta)$ is possibly **nonconvex** and is assumed to be **lower bounded** $\overline{\mathcal{L}}(\theta) > -\infty$

Exponential Family

- Latent data model: $\{z_i\}_{i=1}^n$ are not observed

- Complete data likelihood belongs to the curved exponential family:

Sufficient statistics takes values in $\mathsf{S} \subset \mathbb{R}^d$

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

EM Method for Exponential Family

Updates

- **E-step**:

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

where:

$$\bar{\mathbf{s}}_i(\theta) = \int_{\mathbf{Z}} S(z_i, y_i)\, p(z_i | y_i; \theta)\, \mu(\mathrm{d}z_i)$$

- Define the function $L(\cdot;\theta) : \mathsf{S} \to \mathbb{R}$ as:

$$L(s;\theta) := R(\theta) + \psi(\theta) - \langle s | \phi(\theta) \rangle$$

- There exists a function $\bar{\theta} : \mathsf{S} \mapsto \Theta$ such that

$$L(s; \bar{\theta}(s)) \leq L(s;\theta)$$

- **M-step**:

$$\theta = \bar{\theta}(\bar{s}) = \arg\min_{\theta \in \Theta} \{R(\theta) + \psi(\theta) - \langle s | \phi(\theta) \rangle\}$$

Limitations

- Even though the EM has appealing features:
  - Monotone in likelihood
  - Invariant w.r.t. parametrization
  - Numerically stable (well defined set)

- It is not applicable with the sheer size of today's data

- Approaches based on Stochastic Optimization:
  - [Neal and Hinton, 1998]: Incremental EM (iEM)
  - [Cappé and Moulines, 2009]: Online EM (sEM)
  - [Chen+, 2018]: Variance Reduces EM (sEM-VR)

Stochastic Optimization for EM Methods

General Formulation

- Stochastic EM:

sE-step: $\hat{s}^{(k+1)} = \hat{s}^{(k)} - \gamma_{k+1}\left(\hat{s}^{(k)} - \mathcal{S}^{(k+1)}\right)$

where γ_k is the stepsize and $\mathcal{S}^{(k+1)}$ is a proxy for $\bar{s}\left(\boldsymbol{\theta}^{(k)}\right)$

- **M-step**:

$$\theta^{(k+1)} = \bar{\theta}(\hat{s}^{(k+1)}) = \arg\min_{\theta \in \Theta}\{R(\theta) + \psi(\theta) - \left\langle \hat{s}^{(k+1)} | \phi(\theta) \right\rangle\}$$

- We simplify the notations:

$$\bar{s}_i^{(k)} := \bar{s}_i\left(\boldsymbol{\theta}^{(k)}\right) = \int_Z S\left(z_i, y_i\right) p\left(z_i | y_i; \hat{\boldsymbol{\theta}}^{(k)}\right) \mu\left(\mathrm{d}z_i\right)$$

$$\bar{s}^{(k)} := \bar{s}\left(\boldsymbol{\theta}^{(k)}\right) = \frac{1}{n}\sum_{i=1}^{n} \bar{s}_I^{(k)}$$

$\ell(k) := m\lfloor k/m \rfloor$ First iteration number of the current epoch

$$(iEM\ [NH, 1998]) \quad \mathcal{S}^{(k+1)} = \mathcal{S}^{(k)} + \tfrac{1}{n}\left(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\tau_{i_k}^k)}\right) \quad [1]$$

$$(sEM\ [CM, 2009]) \quad \mathcal{S}^{(k+1)} = \bar{\mathbf{s}}_{i_k}^{(k)} \quad [2]$$

$$(sEM-VR\ [CZTZ., 2018]) \quad \mathcal{S}^{(k+1)} = \bar{\mathbf{s}}^{(\ell(k))} + \left(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\ell(k))}\right) \quad [3]$$

$$(fiEM\ [KLMW., 2019]) \quad \mathcal{S}^{(k+1)} = \overline{\mathcal{S}}^{(k)} + \left(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)}\right)$$

$$\overline{\mathcal{S}}^{(k+1)} = \overline{\mathcal{S}}^{(k)} + n^{-1}\left(\bar{\mathbf{s}}_{j_k}^{(k)} - \bar{\mathbf{s}}_{j_k}^{(t_{j_k}^k)}\right). \quad [4]$$

Algorithm 3 sEM algorithms

Initialization: initializations $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathbf{s}}^{(0)} \leftarrow \bar{\mathbf{s}}^{(0)}$, $K_{\max} \leftarrow$ max. iteration number.

Set the terminating iteration number, $K \in \{0, \ldots, K_{\max}-1\}$, as a discrete r.v. with:

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_{\max}-1} \gamma_\ell}. \quad (42)$$

Iteration k: Given the current state of the chain $\psi_i^{(t-1)}$:

1. Draw index $i_k \in [\![1, n]\!]$ uniformly (and $j_k \in [\![1, n]\!]$ for fiEM).
2. Compute the surrogate sufficient statistics $\mathcal{S}^{(k+1)}$ using [1] or [2] or [3] or [4]
3. Compute $\hat{\mathbf{s}}^{(k+1)}$ via the sE-step
4. Compute $\boldsymbol{\theta}^{(k+1)}$ via the M-step

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

Global Convergence

Assumptions

(A1) The function ϕ is smooth and bounded on the interior of Θ , noted $\operatorname{int}(\Theta)$
For all $(\theta, \theta') \in \operatorname{int}(\Theta)$, $\left\|\mathrm{J}_\phi^\theta(\boldsymbol{\theta}) - \mathrm{J}_\phi^\theta(\boldsymbol{\theta}')\right\| \leq \mathrm{L}_\phi \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|$
and $\left\|\mathrm{J}_\phi^\theta(\theta')\right\| \leq C_\phi$

(A2) The conditional distribution is smooth on $\operatorname{int}(\Theta)$

$$\left|p(z|y_i;\boldsymbol{\theta}) - p(z|y_i;\boldsymbol{\theta}')\right| \leq \mathrm{L}_p \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|$$

(A3) The function $\theta \to L(s;\theta) := R(\theta) + \psi(\theta) - \langle s|\phi(\theta)\rangle$ admits a unique global minimum
Also, $\mathrm{J}_\phi^\theta(\overline{\theta}(\mathrm{s}))$ is full rank and $\overline{\theta}(\mathrm{s})$ is L_θ-Lipschitz

Define:

$$\mathrm{B}(\mathrm{s}) := \mathrm{J}_\phi^\theta(\overline{\boldsymbol{\theta}}(\mathrm{s}))\left(\mathrm{H}_L^\theta(\mathrm{s}, \overline{\boldsymbol{\theta}}(\mathrm{s}))\right)^{-1} \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathrm{s}))^\top$$

(A4) $v_{\max} := \sup_{\mathrm{s} \in \mathrm{S}} \|\mathrm{B}(\mathrm{s})\| < \infty$ and $0 < v_{\min} := \inf_{\mathrm{s} \in \mathrm{S}} \lambda_{\min}(\mathrm{B}(\mathrm{s}))$

$$\|B(s) - B(s')\| \leq L_B \|s - s'\|$$

Incremental EM Method

Lemma

Under **(A1)-(A4),** define $e_i(\boldsymbol{\theta};\boldsymbol{\theta}') := Q_i(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{L}_i(\boldsymbol{\theta})$
We have

$$\left\|\nabla e_i(\boldsymbol{\theta};\boldsymbol{\theta}') - \nabla e_i(\overline{\boldsymbol{\theta}};\boldsymbol{\theta}')\right\| \leq \mathrm{L}_e \|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\|$$

where $\mathrm{L}_e := C_\phi C_{\mathrm{Z}} \mathrm{L}_p + C_{\mathrm{S}} \mathrm{L}_\phi$

Theorem

Under **(A1)-(A4)** for the iEM **[1]** for any $K_{\max} \geq 1$

$$\mathbb{E}\left[\left\|\nabla \overline{\mathcal{L}}\left(\boldsymbol{\theta}^{(K)}\right)\right\|^2\right] \leq n \frac{2L_e}{K_{\max}} \mathbb{E}\left[\overline{\mathcal{L}}\left(\boldsymbol{\theta}^{(0)}\right) - \overline{\mathcal{L}}\left(\boldsymbol{\theta}^{(K_{\max})}\right)\right]$$

where L_e is defined above and K is a uniform random variable on $[0, K_{\max} - 1]$ and independent of the $\{i_k\}_{k=0}^{K_{\max}}$

Stochastic EM as Scaled Gradient Methods

- From a (Scaled) Gradients Method point of view, we consider the minimization problem:

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

Lemma

Under **(A1)-(A4),** we have

$$\left\|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i\left(\overline{\boldsymbol{\theta}}\left(\mathbf{s}'\right)\right)\right\| \leq \mathbf{L_s}\left\|\mathbf{s}-\mathbf{s}'\right\|$$

$$\left\|\nabla V(\mathbf{s}) - \nabla V\left(\mathbf{s}'\right)\right\| \leq \mathbf{L_V}\left\|\mathbf{s}-\mathbf{s}'\right\|$$

where $\mathrm{L_s} := C_{\mathrm{Z}}\mathrm{L}_p\mathrm{L}_\theta$ and $\mathrm{L}_V := v_{\max}\left(1+\mathrm{L_s}\right) + \mathrm{L}_B C_{\mathrm{S}}$

Theorem (sEM-VR)

There exists a constant $\mu \in (0,1)$ such that if

$$\overline{L}_v := \max(L_V, L_s) \qquad \gamma = \frac{\mu v_{\min}}{\overline{L}_v n^{2/3}} \qquad m = \frac{n}{2\mu^2 v_{\min}^2 + \mu}$$

Then:

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

Theorem (fiEM)

There exists a constant $\mu \in (0,1)$ such that if

$$\overline{L}_v := \max(L_V, L_s) \qquad \gamma = \frac{v_{\min}}{\alpha\overline{L}_v n^{2/3}} \qquad \alpha := \max(6, 1+4v_{\min})$$

Then:

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

Numerical Applications

Gaussian Mixture Models (GMM)

- Fit a GMM model to a set of n observations
- Each of M components with unit variance
- The complete log likelihood reads:

$$\log f(z_i, y_i; \boldsymbol{\theta}) = \sum_{m=1}^{M} 1_{\{m\}}(z_i)\left[\log(\omega_m) - \mu_m^2/2\right] + \sum_{m=1}^{M} 1_{\{m\}}(z_i)\,\mu_m y_i + \text{ constant}$$

$$\theta := (\omega, \mu) \qquad \omega = \{\omega_m\}_{m=1}^{M-1} \qquad \mu = \{\mu_m\}_{m=1}^{M}$$

- Penalization used:

$$\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{m=1}^{M}\mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$$

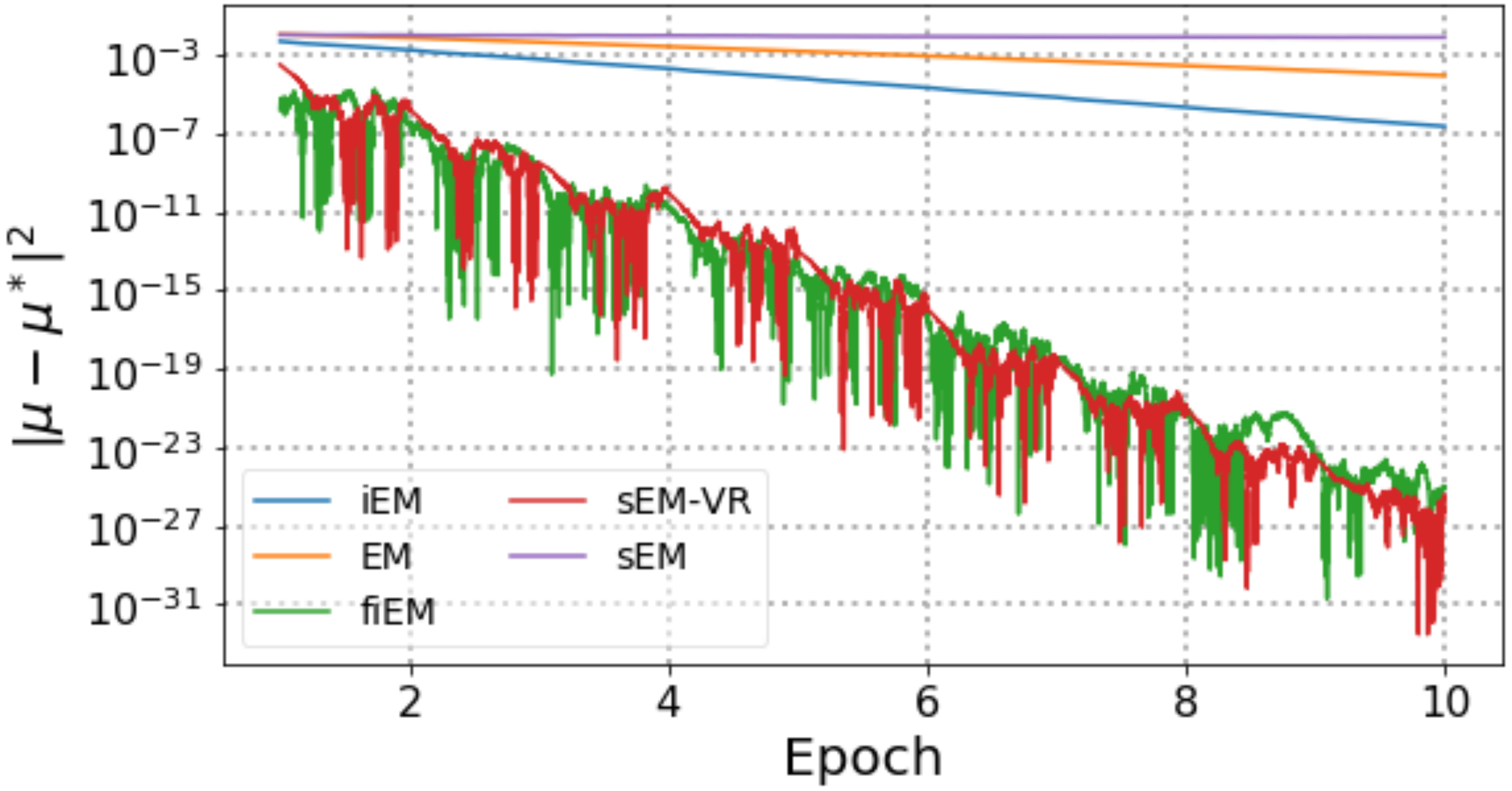
- Numerical: GMM with M=2 and $\mu_1 = -\mu_2 = 0.5$

Experiments

- **Fixed sample size:** size $n = 10^4$ and run to get μ^*

Stepsize for sEM $\gamma_k = 3/(k+10)$

Stepsize for sEM-VR and fiEM prop. to $1/n^{2/3}$



Numerical Applications

Gaussian Mixture Models (GMM)

- Fit a GMM model to a set of n observations
- Each of M components with unit variance
- The complete log likelihood reads:

$$\log f\left(z_i, y_i; \boldsymbol{\theta}\right) = \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\left[\log\left(\omega_m\right) - \mu_m^2/2\right] + \sum_{m=1}^{M} 1_{\{m\}}\left(z_i\right)\mu_m y_i + \text{ constant}$$

$$\theta := (\omega, \mu) \qquad \omega = \{\omega_m\}_{m=1}^{M-1} \qquad \mu = \{\mu_m\}_{m=1}^{M}$$

- Penalization used:

$$\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{m=1}^{M}\mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$$

- Numerical: GMM with M=2 and $\mu_1 = -\mu_2 = 0.5$

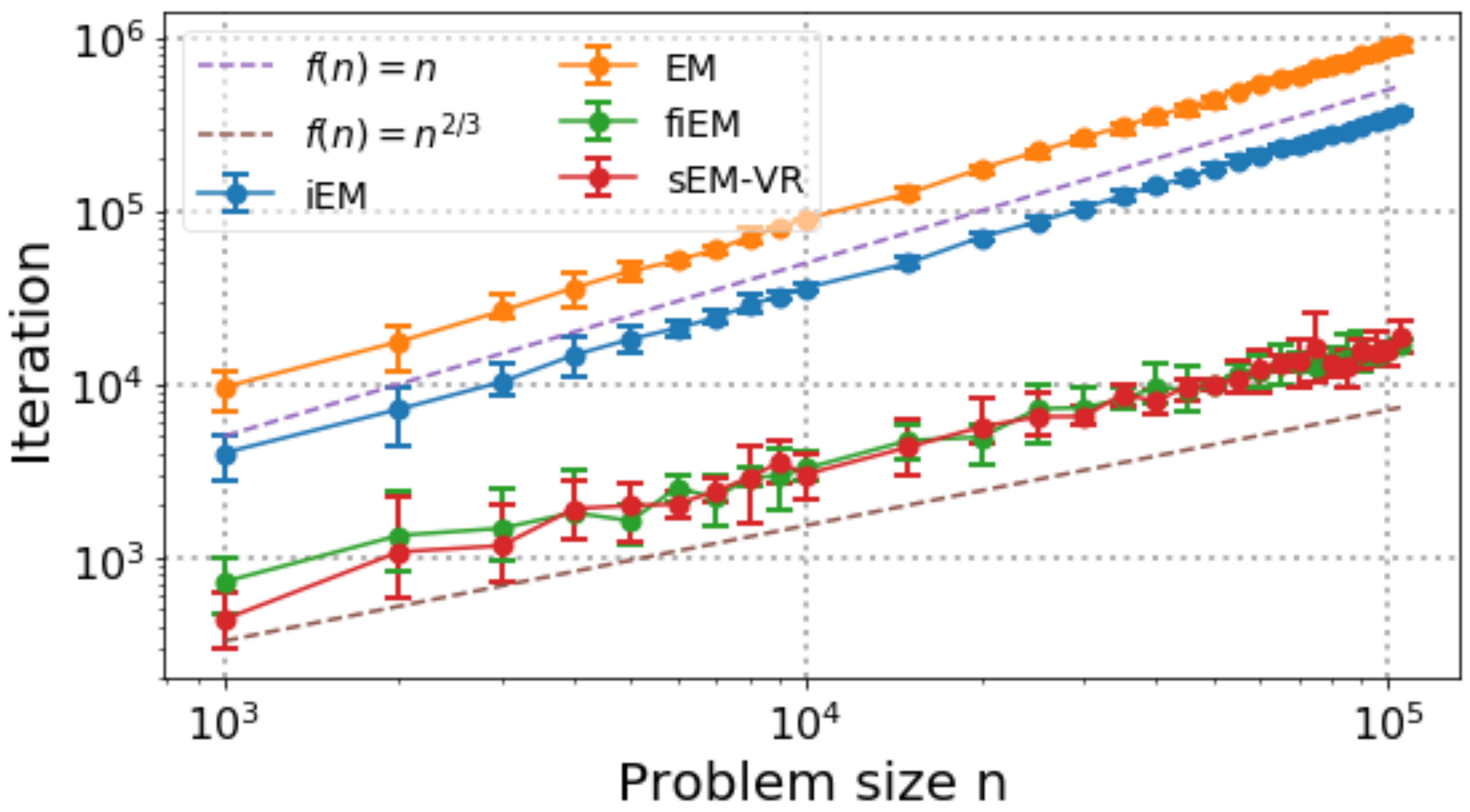
Experiments

- **Fixed sample size:** size $n = 10^4$ and run to get μ^*

Stepsize for sEM $\gamma_k = 3/(k+10)$

Stepsize for sEM-VR and fiEM prop. to $1/n^{2/3}$

- **Varying sample size:** nb. Iterations required to reach a precision of 10^{-3} from $n = 10^3$ to $n = 10^5$



Numerical Applications

Probabilistic Latent Semantic Analysis

- Consider D documents with terms from a vocabulary of size V.
- Data is summarized by a list of tokens

$$\{y_i\}_{i=1}^n \qquad y_i = \left(y_i^{(\mathrm{d})}, y_i^{(\mathrm{w})}\right)$$

- The goal of pLSA is to classify the documents into K topics which is modeled as a latent variable associated with each token $z_i \in [1, K]$

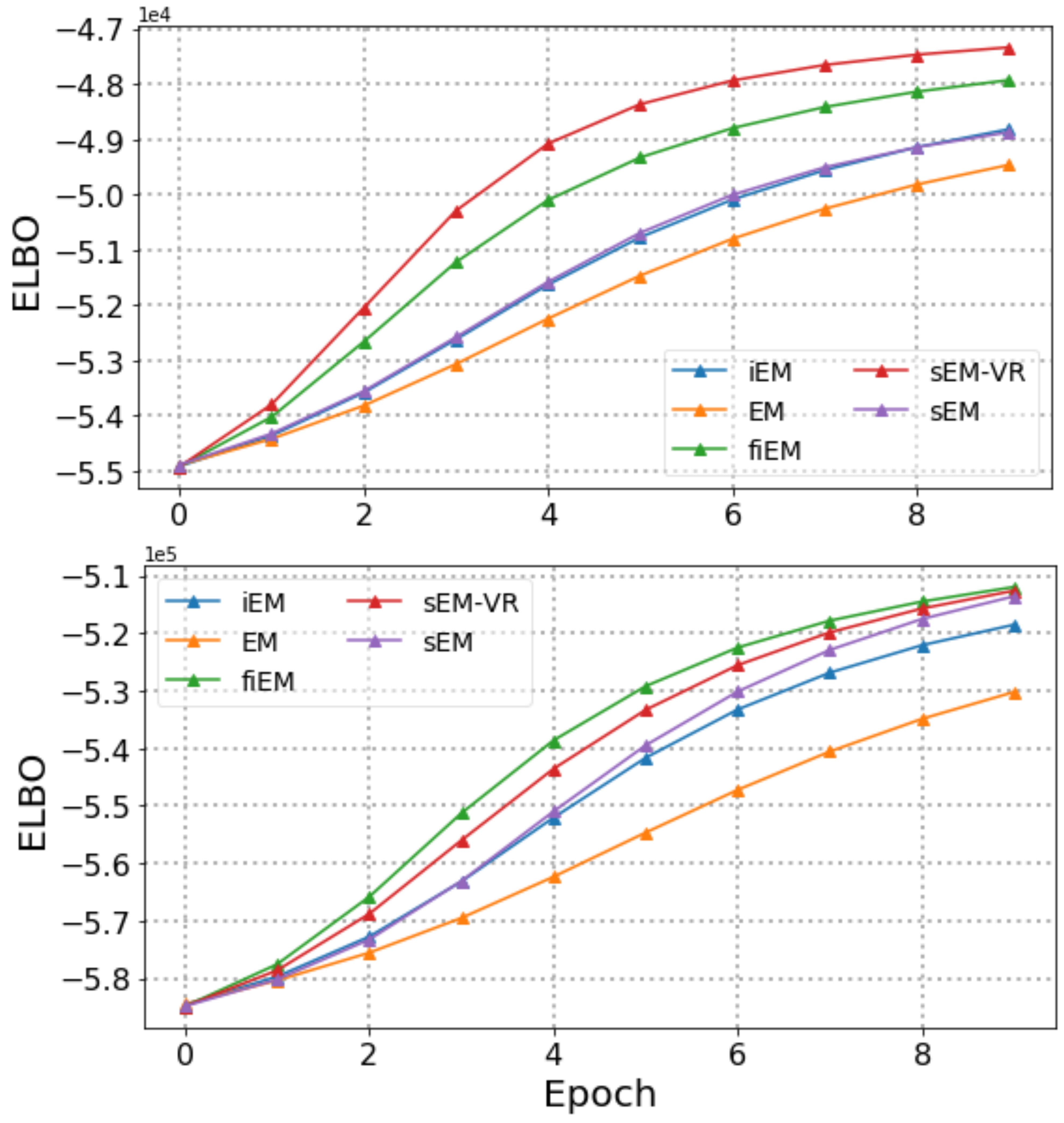
$$\log f(z_i, y_i; \boldsymbol{\theta}) = \sum_{k=1}^{K}\sum_{d=1}^{D} \log(\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})}) \mathbb{1}_{\{k,d\}}(z_i, y_i^{(\mathrm{d})}) + \sum_{k=1}^{K}\sum_{v=1}^{V} \log(\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})}) \mathbb{1}_{\{k,v\}}(z_i, y_i^{(\mathrm{w})})$$

- Penalization used:

$$\mathrm{R}(\boldsymbol{\theta}^{(\mathrm{t|d})}, \boldsymbol{\theta}^{(\mathrm{w|t})}) = -\log \mathrm{Dir}(\boldsymbol{\theta}^{(\mathrm{t|d})}; K, \alpha') - \log \mathrm{Dir}(\boldsymbol{\theta}^{(\mathrm{w|t})}; V, \beta')$$

$$\boldsymbol{\theta} := (\boldsymbol{\theta}^{(\mathrm{t|d})}, \boldsymbol{\theta}^{(\mathrm{w|t})})$$

Experiments



Conclusion

Take-Aways

- We studied the global convergence of stochastic EM Methods
  - Globally (independent of initialization)
  - Non-asymptotic results

- We used a Majorization-Minimization scheme to analyze the incremental EM method

- We interpreted the variance-reduced and the fast incremental method using a scaled gradient scheme to find a stationary point of a well defined Lyapunov function

Thank You !