On the Global Convergence of (Fast) Incremental EM Methods

Belhal Karimi\textsuperscript{1,2}, Hoi-To Wai\textsuperscript{3}, Eric Moulines\textsuperscript{2} and Marc Lavielle\textsuperscript{1}

INRIA\textsuperscript{1}, École Polytechnique\textsuperscript{2}, Chinese University of Hong Kong\textsuperscript{3}

Maximum Likelihood Estimation (MLE)

- Given a set of \( n \) observations \( y = (y_i, i \in [n]) \).
- Goal: fitting the parametric model \( g(y; \theta) \).
- Maximum Likelihood Estimation of \( \theta \)
  \[ \theta^* = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log g(y_i; \theta) \]

\( g(y; \theta) = \int_F f(z; \theta) \mu(\text{dz}) \) 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:
  \[ \min_{\theta \in \Theta} \mathcal{L}(\theta) := R(\theta) + \mathcal{L}(\theta) \]
  \[ \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i; \theta) \]

- \( \mathcal{L} \) is possibly nonconvex and lower bounded
- For all \( i \in [n] \), \( f(z; y_i; \theta) \) and \( p(z; \theta) \) denote the complete likelihood and the posterior distribution

Focus on the Regularized Empirical Risk Minimization

Goal: fitting the parametric model

\[ \mathcal{L}(\hat{\theta}(k)) := \mathcal{L}(\hat{\theta}) + \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i; \theta(\hat{\theta})) \]

\( \hat{\theta}(k) \) is \( \mathcal{L} \)-smooth for all \( k \).

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

Stochastic EM (sEM)

- sEM can be interpreted as incremental MM (Maial, 2015) with the upper bound surrogate function:
  \[ Q(\theta; \theta') = - \int_J \log f(y; \theta') \mu(dy) + \mathcal{L}(\theta') \]

Incremental MM: at every iteration \( k \), we obtain \( \hat{\theta}(k+1) = \arg \min_{\theta} \mathbb{E} \mathcal{L}(\theta; \hat{\theta}(k)) \).

Convergence Analysis: with exponential family model, \( Q(\theta; \theta') - \mathcal{L}(\theta) \) is \( L_s \)-smooth for all \( s \).

Theorem (iEM) For any \( K_{\text{max}} \geq 1 \), with exponential family, the global rate:

\[ \mathbb{E}[\| \nabla \mathcal{L}(\hat{\theta}(k)) \|^2] \leq \frac{2 L_s}{K_{\text{max}}} \mathbb{E}[\mathcal{L}(\hat{\theta}(0)) - \mathcal{L}(\hat{\theta}(K_{\text{max}}))]. \]

An incremental MM Scheme

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

- Unlike iEM, the sEM-VR and fiEM methods can be analyzed as scaled gradients methods. Consider:
  \[ \min_{s \in \mathbb{S}} V(s) := \mathcal{L}(\hat{\theta}(s)) = R(\hat{\theta}(s)) + \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\hat{y}(s)) \]

- Variance-reduced scaled gradient: the sE-step update \( \hat{\theta}(k) \) by \( s(k) \) via the sE-step, we can show
  \[ \mathbb{E}[\| \nabla \mathcal{L}(\hat{\theta}(k)) \|^2] \geq \tau V(\hat{\theta}(k)) \]

Convergence Analysis: with exponential family model, \( \mathbb{E}[\| \nabla \mathcal{L}(\hat{\theta}(k)) \|^2] \)

Theorem (sEM-VR) \( \gamma = \min_{k \in \mathbb{S}} \mathbb{E}[V(s)] \) & epoch \( m = \frac{n}{2 \mu \tau V_{\text{min}} + \eta} \),

\[ \mathbb{E}[\| \nabla \mathcal{L}(\hat{\theta}(k)) \|^2] \leq \frac{2 L_s}{m} \mathbb{E}[\| \nabla \mathcal{L}(\hat{\theta}(0)) - \mathcal{L}(\hat{\theta}(K_{\text{max}})) \|^2] \]

Theorem (fiEM) \( \gamma = \min_{k \in \mathbb{S}} \mathbb{E}[V(s)] \) & \( \alpha = \max\{6, 1 + 4\mu m\} \)

\[ \mathbb{E}[\| \nabla \mathcal{L}(\hat{\theta}(k)) \|^2] \leq \frac{2 L_s}{m} \mathbb{E}[\| \nabla \mathcal{L}(\hat{\theta}(0)) - \mathcal{L}(\hat{\theta}(K_{\text{max}})) \|^2] \]

Advances in Neural Information Processing Systems 32, Vancouver, Canada