Paper review: Statistical guarantees for the EM: From population to sample-based analysis

Botao Hao

Department of Statistics
Purdue University

April 9, 2015

Contribution

- Resolve the gap between the global minimizer and the local optima

$$\|M_n(\theta^{t-1}) - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa} \varepsilon_M^{unif}(n, \delta) \tag{1}$$

- Provide non-asymptotic guarantees for EM algorithms
- Apply their general theory for three specific models

$$\begin{cases} \text{Gaussian mixture model} \\ \text{Mixture of regression model} \\ \text{Linear regression with missing covariates} \end{cases}$$

Outline

- Review for EM and Gradient EM Algorithm
- General convergence result for EM
- Guarantees for population-version EM
- Guarantees for sample-based EM
- Some related works

Expectation-maximization algorithm

- The pair (Y, Z) has a joint density function f_{θ^*}, where Y is observed and Z is missing.
- Compute some $\hat{\theta} \in \Omega$ maximizing $g_\theta(y)$, where

$$g_\theta(y) = \int_{\mathcal{Z}} f_\theta(y, z) dz \tag{2}$$

- $k_\theta(z|y)$ denotes the conditional density of z given y. We have a lower bound:

$$\underbrace{L(\theta')}_{log(g_{\theta'}(y))} \geq \underbrace{\int_{\mathcal{Z}} k_\theta(z|y) log f_{\theta'}(y, z) dz}_{Q(\theta'|\theta)} - \int_{\mathcal{Z}} k_\theta(z|y) log k_{\theta'}(z|y) dz \tag{3}$$

Expectation-maximization algorithm

Standard EM updates:

E-step: Calculate $Q(\theta'|\theta)$

M-step: Compute the maximizer $\theta_{t+1} = argmax_{\theta' \in \Omega} Q(\theta'|\theta_t)$

We let $\theta_{t+1} = M(\theta_t)$, which is a mapping $M : \Omega \to \Omega$

Generalized EM updates:

M-step: Choose θ_{t+1} such that $Q(\theta_{t+1}|\theta_t) \geq Q(\theta_t|\theta_t)$

Gradient EM updates:

M-step: Update $\theta_{t+1} = \theta_t + \alpha \nabla Q(\theta_t|\theta_t)$

We let $G(\theta) = \theta + \alpha \nabla Q(\theta|\theta)$

Population version versus sample version

- Population version:

$$Q(\theta'|\theta) = \int_{\mathcal{Y}} (\int_{\mathcal{Z}} k_\theta(z|y) log f_{\theta'}(y,z) dz) g_{\theta^*}(y) dy \tag{4}$$

- Sample-based version:

$$Q_n(\theta'|\theta) = \frac{1}{n} \sum_{i=1}^{n} (\int_{\mathcal{Z}} k_\theta(z|y) log f_{\theta'}(y,z) dz) \tag{5}$$

General convergence results for EM

Wu(1983) established some of the most general convergence results for the EM algorithm.

- Suppose $Q(\theta'|\theta)$ is continuous in both θ' and θ. Then all the limit points of any θ_t of an EM algorithm are stationary points of L.
- Additionally, suppose $sup_{\Phi' \in \Omega} Q(\Phi'|\Phi) > Q(\Phi|\Phi)$. Then all the limit points are local maxima.
- Suppose that $L(\theta')$ is *unimodal* in Ω with θ^* being the only stationary point. Then θ_t converges to the unique maximizer θ^*.

Guarantees for population-level EM

Definition (*Self-consistency*)

$$\theta^* = argmax_{\theta \in \Omega} Q(\theta|\theta^*)$$

Definition (*Contractive Mapping*)

A contractive mapping on a metric space($\mathcal{M}$, d) is a function f from $\mathcal{M}$ to itself, such that for all x and y in $\mathcal{M}$

$$d(f(x), f(y)) \leq k \cdot d(x, y), 0 \leq k < 1$$

- Add some convexity and smoothness conditions on $q(\cdot) = Q(\cdot|\theta^*)$
- Then, the population operators are contractive on a ball containing the fixed point θ^*, where $B_2(r; \theta^*) = \{\theta \in \Omega | \|\theta - \theta^*\|_2 \leq r\}$

Two key conditions

Condition ($\lambda-$strongly concave)

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \leq -\frac{\lambda}{2} \|\theta_1 - \theta_2\|_2^2 \tag{6}$$

for all pairs (θ_1, θ_2) *in a neighborhood of* θ^*.

Condition (First-order Stability(FOS))

The function$Q(\cdot|\theta), \theta \in \Omega$ *satisfy condition FOS(γ) over* $B_2(r; \theta^*)$ *if*

$$\|\nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta)\|_2 \leq \gamma \|\theta - \theta^*\|_2$$

$$B_2(r; \theta^*) := \{\theta \in \Omega | \|\theta - \theta^*\|_2 \leq r\}$$

Actually, FOS is the Lipschitz condition for $\nabla Q(M(\theta)|\cdot)$.

Characterizations condition

Proposition (First order optimality condition(Bubeck 2014))

Let f be convex and X a closed convex set on which f is differentiable. Then

$$x^* \in argmin_{x \in X} f(x) \tag{7}$$

if and only if one has

$$\nabla f(x^*)^T (x^* - y) \leq 0, \forall y \in X \tag{8}$$

Condition (First-order optimality)

$$\left\langle \nabla Q(\theta^*|\theta^*), \theta' - \theta^* \right\rangle \leq 0 \tag{9}$$

$$\left\langle \nabla Q(M(\theta)|\theta), \theta' - M(\theta) \right\rangle \leq 0 \tag{10}$$

$\theta^* = argmax_{\theta \in \Omega} Q(\theta|\theta^*)$ $M(\theta) = argmax_{\theta' \in \Omega} Q(\theta'|\theta)$

Population version EM theorem

Theorem

For some radius $r > 0$ and pair (γ, λ) such that $0 \leq \gamma < \lambda$, suppose that the function $Q(\cdot|\theta^*)$ is λ-strongly concave, and that the $FOS(\lambda)$ condition holds on the ball $B_2(r; \theta^*)$. Then the population EM operator M is contractive over $B_2(r; \theta^*)$, in particular with

$$\|M(\theta) - \theta^*\|_2 \leq \frac{\gamma}{\lambda}\|\theta - \theta^*\|_2, \forall \theta \in B_2(r; \theta^*) \tag{11}$$

$$\|\theta^t - \theta^*\|_2 \leq (\frac{\gamma}{\lambda})^t \|\theta^0 - \theta^*\|_2 \tag{12}$$

Guarantees for sample-based EM

- Recall $M_n(\theta) := argmax_{\theta' \in R^d} Q_n(\theta'|\theta)$
- For a given sample size n, tolerance parameter $\delta \in (0,1)$ and any *fixed* $\theta \in B_2(r;\theta^*)$, we have

$$\|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon_M(n,\delta) \tag{13}$$

  with probability at least 1-δ
- A stronger condition: uniform upper bound

$$sup_{\theta \in B_2(r;\theta^*)} \|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon_M^{unif}(n,\delta) \tag{14}$$

  with probability at least 1-δ
- For specific models, we have a close form for $\varepsilon_M^{unif}(n,\delta)$

Main Result

Theorem

Suppose that the population EM operator $M : \Omega \rightarrow \Omega$ is contractive with parameter $\kappa \in (0,1)$ on the ball $B_2(r;\theta^*)$, and the initial vector θ^0 belongs to $B_2(r;\theta^*)$.

If the sample size n is large enough to ensure that

$$\varepsilon_M^{unif}(n,\delta) \leq (1-\kappa)r \tag{15}$$

then the EM iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy the bound

$$\|\theta^t - M(\theta^{t-1}) + M(\theta^{t-1}) - \theta^*\|_2 \leq \underbrace{\kappa^t\|\theta^0 - \theta^*\|_2}_{Optimization\ Error} + \underbrace{\frac{1}{1-\kappa}\varepsilon_M^{unif}(n,\delta)}_{Statistical\ Error} \tag{16}$$

with probability at least 1-δ

Decomposition of the error!

A similar idea for Bayes risk (Martin Wainwright's lecture)

- $R(\hat{g}_n) = P[Y \neq sign(\hat{g}_n(X) - \frac{1}{2})]$, corresponding to the true error probability of our classifier.
- $R^* = inf_{g \in F} P[Y \neq sign(g(X) - \frac{1}{2})]$, corresponding to the best rule.
- Decomposition:

$$R(\hat{g}_n) - R^* = \underbrace{\{R(\hat{g}_n) - inf_{g \in \mathcal{F}_{\epsilon_n}} R(g)\}}_{Estimation\ error} + \underbrace{\{inf_{g \in \mathcal{F}_{\epsilon_n}} R(g) - inf_{g \in \mathcal{F}} R(g)\}}_{Approximation\ error}$$

where $\mathcal{F}_{\epsilon_n}$ is an $\epsilon_n - covering$ of $\mathcal{F}$.

Stopping Rule

Consider any positive integer T such that

$$T \geq log_{\frac{1}{\kappa}} \frac{(1-\kappa)\|\theta^0 - \theta^*\|_2}{\varepsilon_M^{unif}(n,\delta)} \tag{17}$$

Then the *Optimization Error* is dominated by *Statistical Error*.
However, this T is not computable based only on data. It just provides a iteration complexity growth ratio.

Graph illustration

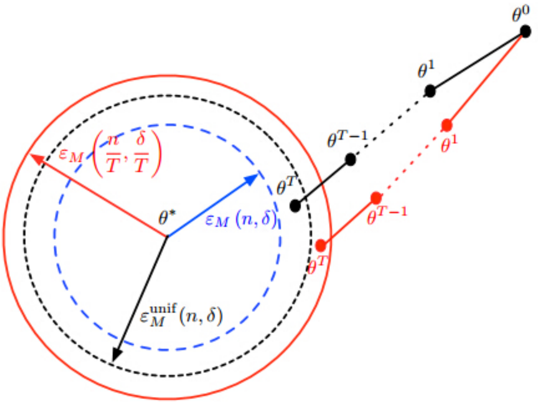

Figure: An illustration of Main Theorem

Sample-splitting version of the EM algorithm

- Divide the full data set into T subsets of size $[n/T]$
- Perform the updates $\theta^{t+1} = M_{T/n}(\theta^t)$
- *Use a fresh subset of samples at each iteration*
- Do we need to consider *the divide and conquer* procedure?

Gradient EM algorithm(Population version)

Definition 1 ($\mu - smooth$)

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \geq -\frac{\mu}{2} \|\theta_1 - \theta_2\|_2^2 \tag{18}$$

Definition 2 (Gradient Stability(GS))

$$\|\nabla Q(\theta|\theta^*) - \nabla Q(\theta|\theta)\|_2 \leq \gamma \|\theta - \theta^*\|_2 \tag{19}$$

for all $\theta \in B_2(r; \theta^*)$

Gradient Decent v.s Gradient EM

Gradient Decent $T(\theta) := \theta + \alpha \nabla Q(\theta|\theta^*)$

Gradient EM $G(\theta) := \theta + \alpha \nabla Q(\theta|\theta)$

- In the standard optimization theory, the gradient operator $T : \Omega \to \Omega$ with step size $\alpha = \frac{2}{\mu+\lambda}$ is contractive over $B_2(r;\theta^*)$,

$$\|T(\theta) - \theta^*\|_2 \leq (\frac{\mu - \lambda}{\mu + \lambda})\|\theta - \theta^*\|_2 \tag{20}$$

We use the condition $\lambda -$ *strong* concavity and $\mu -$ *smooth*.

- The population gradient EM operator G with step size $\alpha = \frac{2}{\mu+\lambda}$ is contractive over $B_2(r;\theta^*)$,

$$\|G(\theta) - \theta^*\|_2 \leq (1 - \frac{2\lambda - 2\gamma}{\mu + \lambda})\|\theta - \theta^*\|_2 \tag{21}$$

We use the condition $\lambda -$ *strong* concavity, $\mu -$ *smooth* and GS.

Main result for Gradient EM algorithm

Theorem

Suppose that the population gradient EM operator $G : \Omega \to \Omega$ is contractive with parameter $\kappa \in (0, 1)$ on the ball $B_2(r; \theta^*)$, and the initial vector θ^0 belongs to $B_2(r; \theta^*)$.
If the sample size n is large enough to ensure that

$$\varepsilon_G^{unif}(n, \delta) \leq (1 - \kappa) r \tag{22}$$

then the EM iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy the bound

$$\|\theta^t - \theta^*\|_2 \leq \underbrace{\kappa^t \|\theta^0 - \theta^*\|_2}_{\textit{Optimization Error}} + \underbrace{\frac{1}{1-\kappa} \varepsilon_G^{unif}(n, \delta)}_{\textit{Statistical Error}}, \tag{23}$$

with probability at least 1-δ

Some related works

- Two-stage alternating minimization(Yi, Caramanis, Sanghavi(2014))
- High dimensional EM (Wang, Gu, Ning, Liu(2014))

Two-stage alternating minimization

- Only consider the mixed linear regression model

$$y_i = \langle x_i, \beta_1^* \rangle z_i + \langle x_i, \beta_2^* \rangle (1 - z_i) + w_i \tag{24}$$

- SVD a algorithm for initialization step
- Using $O(klog^2 k)$ samples, with high probability their initialization procedure returns $\beta_1^{(0)}$, $\beta_2^{(0)}$ which are within a constant distance of the true β_1^*, β_2^*.

$$max\{\|\beta_1^{(t)} - \beta_1^*\|_2, \|\beta_2^{(t)} - \beta_2^*\|_2\} \leq \tilde{c} min\{p_1, p_2\}\|\beta_1^* - \beta_2^*\|_2 \tag{25}$$

High dimensional EM

- Attach a truncation step to the expectation step and maximization step
- The iterative solution sequence $\{\beta^{(t)}\}_{t=0}^{T}$ satisfies

$$\|\beta^{(t)} - \beta^*\|_2 \leq \underbrace{\Delta_1 \rho^{t/2}}_{\textit{Optimization Error}} + \underbrace{\Delta_2 \sqrt{\frac{s^* logd}{n}}}_{\textit{Statistical Error}} \tag{26}$$

  with high probability
- Add some high-dimensional inferences

References

- Balakrishnan, S., Wainwright, M. J. and Yu, B. (2014). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *arXiv preprint arXiv:1408.2156.*
- X. Yi, C. Caramanis, and S. Sanghavi. (2014) Alternating minimization for mixed linear regression. *arXiv preprint arXiv:1310.3745*
- S. Bubeck. (2014) Theory of convex optimization for machine learning. *arXiv preprint arXiv:1405.4980.*
- Z. Wang, Q. Gu, Y. Ning, H. Liu. (2014) High Dimensional Expectation-Maximization Algorithm: Statistical Optimization and Asymptotic Normality *arXiv preprint arXiv:1412.8729v2.*
- A. Andresen, V. Spokoiny. (2015) Two convergence results for an alternation maximization procedure. *arXiv preprint arXiv:1501.01525*

Thank you!