Online EM Algorithm for Latent Data Models

[Cappé, Moulines, 2007]

Yingyu Yang, Léon Zheng

Master M2 MVA, Computational Statistics 2019

January 7, 2020

Introduction

Online version of EM algorithm

- Limit of batch EM algorithm: impractical when processing large data sets
- [Titterington, 1984]'s approach to online EM:

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1} I^{-1}(\hat{\theta}_n) \nabla_\theta \log g(Y_{n+1}; \hat{\theta}_n) \tag{1}$$

Contributions

- Stochastic approximation in E-step + maximization in M-step
- Not relying on the complete data information matrix
- Not assuming that the model is *well-specified*
- Convergence to stationary point with optimal rate

Online EM algorithm

Notation: parameter $\theta \in \Theta$, observation Y distributed under π, latent variable X distributed under $f(x,\theta)$, likelihood function $g(y;\theta)$.

Idea: replace the expectation step by a stochastic approximation step

$$\hat{Q}_{n+1}(\theta) = \hat{Q}_n(\theta) + \gamma_{n+1}\left(\mathbb{E}_{\hat{\theta}_n}\left[\log f(X_{n+1};\theta) \middle| Y_{n+1}\right] - \hat{Q}_n(\theta)\right). \qquad (2)$$

SAEM: $X_{k+1} \sim q(x|Y,\hat{\theta}_k)$, $\hat{Q}_{k+1}(\theta) = \hat{Q}_k(\theta) + \gamma_{k+1}\left(\log q(Y,X_{k+1},\theta) - \hat{Q}_k(\theta)\right)$.

Online EM algorithm

Assumption 1:

- exponential model $f(x,\theta) = h(x)\exp\{-\psi(\theta) + \langle S(x), \phi(\theta)\rangle\}$
- $\bar{s}(y;\theta) \triangleq \mathbb{E}_\theta\left[S(X)\middle| Y = y\right]$ can be computed
- for each $s \in \mathcal{S}$, $\bar{\theta}(s) \triangleq \arg\max_{\theta\in\Theta}\{-\psi(\theta) + \langle s, \phi(\theta)\rangle\}$ is unique

Iterations:

$$\begin{aligned} \hat{s}_{n+1} &= \hat{s}_n + \gamma_{n+1}\left(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right) \\ \hat{\theta}_{n+1} &= \bar{\theta}(\hat{s}_{n+1}) \end{aligned} \qquad (3)$$

Consistency

Robbins-Monroe SA procedure: $\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1}\left(h(\hat{s}_n) + \xi_{n+1}\right)$

- mean field $h(s) \triangleq \mathbb{E}_\pi\left[\bar{s}\left(Y; \bar{\theta}(s)\right)\right] - s$
- denote $\Gamma \triangleq \{s \in \mathcal{S} : h(s) = 0\}$, and $\mathcal{L} \triangleq \{\theta \in \Theta : \nabla_\theta \mathrm{KL}\left(\pi \| g_\theta\right) = 0\}$
- if $s^* \in \Gamma$, then $\bar{\theta}(s^*) \in \mathcal{L}$ under Assumption 2 (which includes: for some $p > 2$, $\sup_{s \in \mathcal{K}} \mathbb{E}_\pi\left(\left|\bar{s}\left(Y; \bar{\theta}(s)\right)\right|^p\right)$)
- a Lyapunov function for the mean field h is $w(s) \triangleq \mathrm{KL}\left(\pi \| g_{\bar{\theta}(s)}\right)$

Theorem (Consistency)

Assuming 1, 2, and that, in addition,

- $0 < \gamma_i < 1, \sum_{i=1}^{\infty} \gamma_i = \infty$ *and* $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$
- $\hat{s}_0 \in \mathcal{S}$, $\limsup |\hat{s}_n| < \infty$ *a.s., and* $\liminf d(\hat{s}_n, \mathcal{S}^c) > 0$ *a.s.*
- $w(\Gamma)$ *is nowhere dense.*

Then, $\lim_{n \to \infty} d(\hat{s}_n, \Gamma) = 0$ *and* $\lim_{n \to \infty} d(\hat{\theta}_n, \mathcal{L}) = 0$*, with probability one.*

Rate of convergence

Theorem (Rate of convergence)

Under the assumptions of the previous theorem, let θ^* be a minimum of $\theta \mapsto KL(\pi \| g_\theta)$.
Let $\gamma_n = \gamma_0 n^{-\alpha}$, where $\gamma_0 \in]0, 1[$ when $\alpha \in]\frac{1}{2}, 1[$ and $\gamma_0 > \lambda(\theta^*)$ when $\alpha = 1$.
Then, on the event $\{\lim_{n\to\infty} \hat{\theta}_n = \theta^*\}$, the sequence $\gamma_n^{-\frac{1}{2}} \left(\hat{\theta}_n - \theta^*\right)$ convergences in distribution to $\mathcal{N}\left(0, \Sigma(\theta^*)\right)$, where $\Sigma(\theta^*)$ is the solution of a Lyapunov equation.

Remarks:

- Optimal rate $\mathcal{O}(n^{-\frac{1}{2}})$ for $\alpha = 1$, but with constraint on γ_0.
- Slower convergence for $\alpha < 1$, but without constraint on γ_0.

Experiments: mixture of m Poisson distributions

Likelihood: $g(y;\theta) = \sum_{j=1}^{m} \omega_j \frac{\lambda_j^y}{y!} e^{-\lambda_j}$ where $\theta = (\omega_1, \cdots, \omega_j, \lambda_1, \cdots, \lambda_j)$.

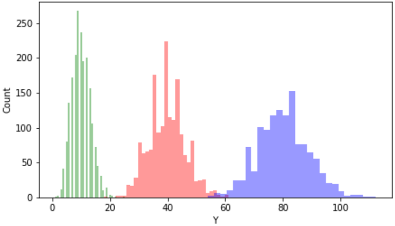

Figure: Distribution of the data set generated by a Poisson mixture with $\omega_1 = 0.45$, $\omega_2 = 0.35$, $\omega_3 = 0.25$, $\lambda_1 = 10$, $\lambda_2 = 40$, $\lambda_3 = 80$.

Implementation of online EM algorithm for Poisson mixture

$$\forall j \in \{1, \cdots, m\},\ \hat{s}_{j,n+1} = \hat{s}_{j,n} + \gamma_{n+1} \left[\begin{pmatrix} \bar{w}_j(Y_{n+1}; \hat{\theta}_n) \\ \bar{w}_j(Y_{n+1}; \hat{\theta}_n) Y_{n+1} \end{pmatrix} - \hat{s}_{j,n} \right]$$
$$\hat{\omega}_{j,n+1} = \hat{s}_{j,n+1},\ \hat{\lambda}_{j,n+1} = \frac{\hat{s}_{j,n+1}(2)}{\hat{s}_{j,n+1}(1)} \tag{4}$$

Experiments: mixture of m Poisson distributions

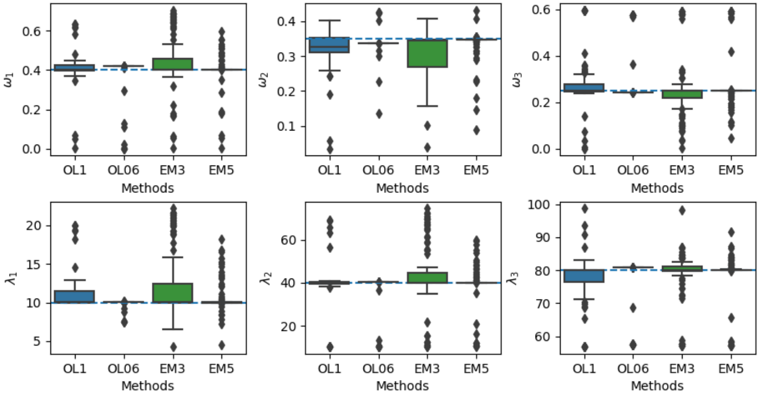

Figure: Methods comparison for 100 random initializations of θ, with 5000 samples. OL1, OL06: online EM, with step size $\gamma_i = 1/i$, $\gamma_i = 1/i^{0.6}$. EM3, EM5: batch EM, 3 and 5 iterations. Ground truth is in dashed line.

Experiments: mixture of m Poisson distributions

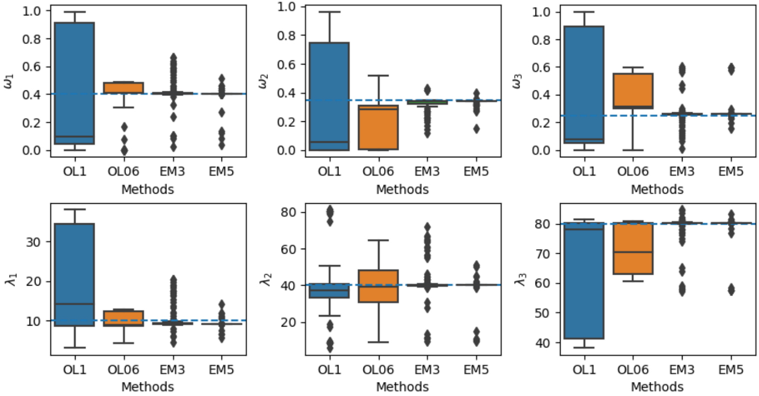

Figure: Methods comparison for 100 random initializations of θ, with 100 samples. OL1, OL06: online EM, with step size $\gamma_i = 1/i$, $\gamma_i = 1/i^{0.6}$. EM3, EM5: batch EM, 3 and 5 iterations. Ground truth is in dashed line.

Experiments: mixture of m Poisson distributions

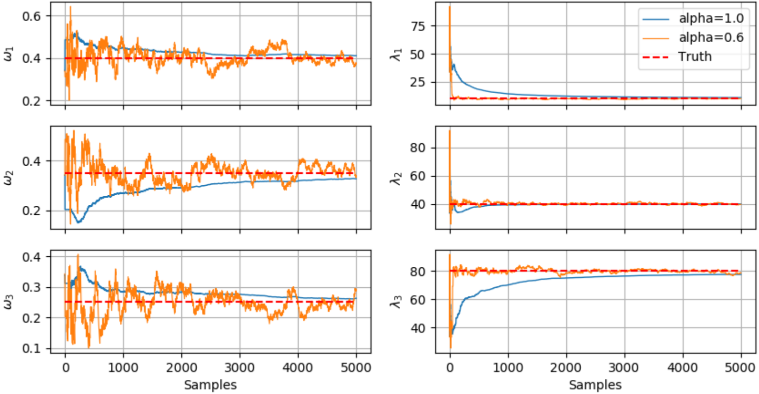

Figure: Parameters trajectories of online EM algorithm on Poisson mixture with 5000 samples, where step size is $\gamma_i = \gamma_0 / i^\alpha$, $\gamma_0 = 1$.

Experiments: mixture of m Poisson distributions

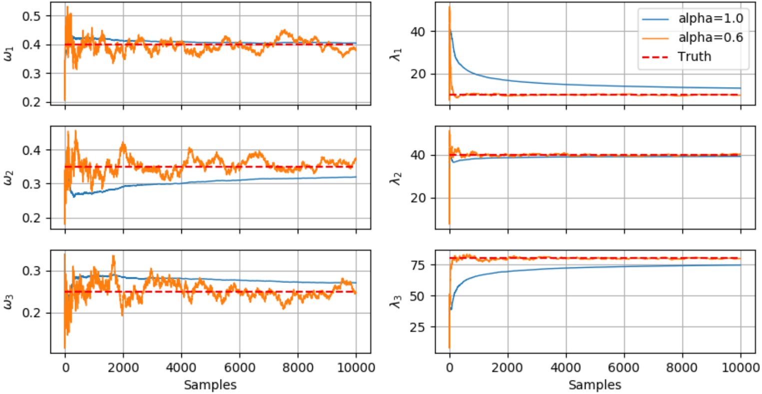

Figure: Parameters trajectories of online EM algorithm on Poisson mixture with 10000 samples, where step size is $\gamma_i = \gamma_0 / i^{\alpha}$, $\gamma_0 = 0.5$.

Experiments: mixture of m Poisson distributions

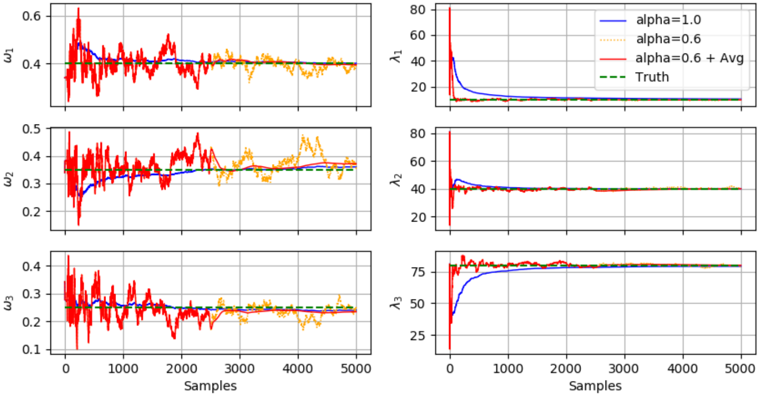

Figure: Parameters trajectories of online EM algorithm on Poisson mixture with 5000 samples, where step size is $\gamma_i = \gamma_0 / i^\alpha$, $\gamma_0 = 1$. We added Polyak-Ruppert averaging for $\alpha = 0.6$ at after 2500 iterations.

Conclusion

Online EM in practice

- Less convergence bias for step size $\gamma_i = \gamma_0 i^{-0.6}$ than for $\gamma_i = \gamma_0 i^{-1}$
- Influence of data size

Limits of online EM

- Limited to simple parametric models
- Still relying on explicit $\bar{\theta} : s \mapsto \arg\max_{\theta \in \Theta} \{-\psi(\theta) + \langle s, \phi(\theta) \rangle\}$

Discussion

- Comparison with SAEM
- When to use online EM algorithm?

References

Cappé, O., Moulines, E. (2007)
"Online EM Algorithm for Latent Data Models"
In *ArXiv*, abs/0712.4273.

Titterington, D. M. (1984)
"Recursive Parameter Estimation Using Incomplete Data"
In *Journal of the Royal Statistical Society. Series B (Methodological)*
Vol. 46, No. 2 (1984), pp. 257-267.