

Online EM Algorithms for Mixtures (and HMMs)

Olivier Cappé

Télécom ParisTech & CNRS

ICMS Mixture Estimation and Applications Workshop
March 5, 2010

- 1 Goals and Assumptions
- 2 Limiting EM Behaviour
- 3 The Proposed Online EM Algorithm
- 4 Discussion
- 5 What About HMMs? (Trailer)

Online Estimation for Missing Data Models

Based on (C & Moulines, 2009) and (C, 2009)

Goals

- 1 Maximum likelihood estimation, or
 - 1' Competitive with maximum likelihood estimation when $\#obs.$ is large
- 2 Good scaling (performance vs. computational cost) as $\#obs.$ increases
- (3) Process data on-the-fly (no storage)

Latent Data Model Observations Y_t , missing data X_t
(Curved) Exponential Family Model

$$p_{\theta}(x_t, y_t) = \exp(\langle s(x_t, y_t), \psi(\theta) \rangle - A(\theta))$$

with respect to some measure¹

Explicit Complete-Data ML

$$S \mapsto \bar{\theta}(S) = \arg \max_{\theta} \langle S, \psi(\theta) \rangle - A(\theta)$$

is available in closed-form

IID Data (Y_t) is an iid. process with marginal π
(not necessarily equal to $f_{\theta_{\star}}$)

¹ f_{θ} denotes the marginal in y_t

The (Usual) Expectation-Maximisation Algorithm

The k -th EM Iteration (From n Observations)

E-Step

$$S_{n,k} = \frac{1}{n} \sum_{t=1}^n \mathbb{E}_{\theta_{k-1}} [s(X_t, Y_t) | Y_t]$$

E-Step

$$\theta_k = \bar{\theta}(S_{n,k})$$

Can be fully reparameterised in the domain of **sufficient statistics**

$$S_{n,k} = \frac{1}{n} \sum_{t=1}^n \mathbb{E}_{\bar{\theta}(S_{n,k-1})} [s(X_t, Y_t) | Y_t]$$

The Limiting EM Recursion

By letting n tend to infinity, one obtains the two equivalent updates:

Sufficient Statistics Update

$$S_k = \mathbb{E}_\pi \left(\mathbb{E}_{\bar{\theta}(S_{k-1})} [s(X_0, Y_0) | Y_0] \right)$$

Parameter Update

$$\theta_k = \bar{\theta} \left\{ \mathbb{E}_\pi \left(\mathbb{E}_{\theta_{k-1}} [s(X_0, Y_0) | Y_0] \right) \right\}$$

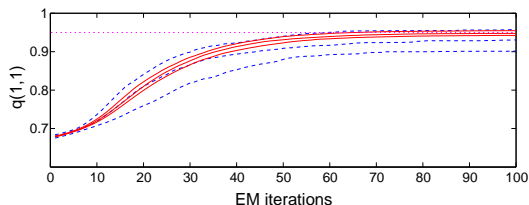
Using usual EM arguments, these updates² are such that

- 1** The Kullback-Leibler divergence $D(\pi | f_{\theta_k})$ is monotonically decreasing with k
- 2** Stationary points of the mapping are such that $\nabla_\theta D(\pi | f_\theta) = 0$

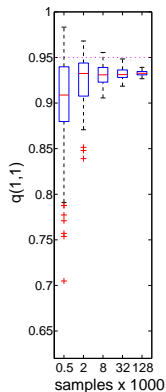
²Used in [C *et al.*, 2008] for adaptive importance sampling

Batch EM Is Not Efficient for Large Data Records

see also (Neal & Hinton, 1999)



Above: Estimated quartiles for different numbers of EM iterations from 500 or 8000 observations; Right: Using a fixed number of 50 EM iterations for various data sizes



- We try to locate the solutions of

$$\mathbb{E}_\pi \left(\mathbb{E}_{\bar{\theta}(S)} [s(X_0, Y_0) | Y_0] \right) - S = 0$$

- Viewing $\mathbb{E}_{\bar{\theta}(S)} [s(X_n, Y_n) | Y_n]$ as a noisy observation of $\mathbb{E}_\pi \left(\mathbb{E}_{\bar{\theta}(S)} [s(X_0, Y_0) | Y_0] \right)$, this is exactly the usual **Stochastic Approximation** (or **Robbins-Monro**) setup:

$$S_n = S_{n-1} + \gamma_n \left(\mathbb{E}_{\bar{\theta}(S_{n-1})} [s(X_n, Y_n) | Y_n] - S_{n-1} \right)$$

where (γ_n) is a sequence of decreasing positive stepsizes

Online EM Algorithm

Stochastic E-Step

$$S_n = (1 - \gamma_n)S_{n-1} + \gamma_n E_{\theta_{n-1}} [s(X_n, Y_n) | Y_n]$$

M Step

$$\theta_n = \bar{\theta}(S_n)$$

Analysis

(C & Moulines, 2009)

Under $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 < \infty$ (and other appropriate assumptions)

- The estimate θ_n converges to one of the roots of $\nabla_{\theta} D(\pi | f_{\theta}) = 0$
- The algorithm is asymptotically equivalent to

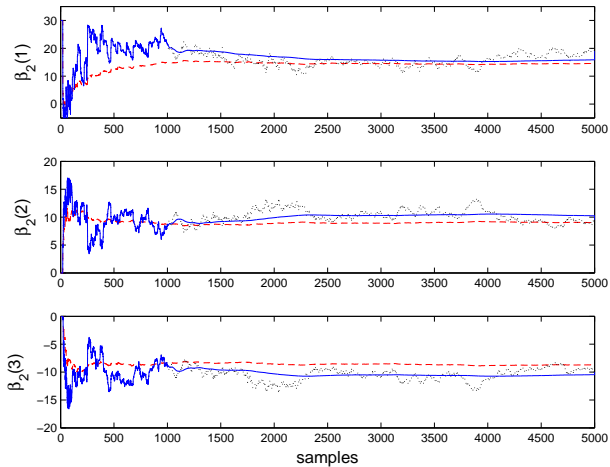
$$\theta_n = \theta_{n-1} + \gamma_n I_{\pi}^{-1}(\theta_{n-1}) \nabla_{\theta} \log f_{\theta_{n-1}}(Y_n)$$

where $I_{\pi}(\theta) = -\mathbb{E}_{\pi} (\mathbb{E}_{\theta} [\nabla_{\theta}^2 \log f_{\theta}(X_0) | Y_0])$

- For a well specified model ($\pi = f_{\theta_{\star}}$) and under **Polyak-Ruppert averaging**³ θ_n is Fisher efficient

³ $\tilde{\theta}_n = 1/(n - n_0) \sum_{t=n_0+1}^n \theta_t$, with $\gamma_n = n^{-\alpha}$ and $\alpha \in (1/2, 1)$

Illustration of Polyak-Ruppert Averaging



(Titterington, 1984) Proposes a gradient algorithm

$$\theta_n = \theta_{n-1} + \gamma_n I^{-1}(\theta_{n-1}) \nabla_{\theta} \log f_{\theta_{n-1}}(Y_n)$$

It is asymptotically equivalent to the algorithm
(previously described) for flat models ($\psi \equiv 1$)

(Neal & Hinton, 1999) Describe the algorithm with $\gamma_n = 1/n$
(equivalent up to first batch tour only)

(Sato, 2000; Sato & Ishii, 2000) Describe the algorithm and
provide some analysis in the flat model case and for
mixtures of Gaussian

How Does This Work in Practice?

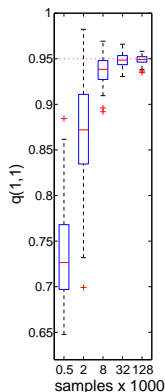
Fine But don't use $\gamma_n = 1/n$

Simulations in (C & Moulines, 2009) using mixtures of regression

Large Scale Experiments on Real Data in (Liang & Klein, 2009), where the use of **mini-batch** blocking was found useful:

- Apply the proposed algorithm considering $Y_{mk+1}, Y_{mk+2} \dots Y_{m(k+1)}$ as one observation

Mini-batch blocking is useful in dealing with mixture-like models with infrequent components



Pros and Cons

The Good

- Easy
- Can be used for ML estimation from a batch of observations (analysis correct for random scans)
- Robust wrt. to stepsize selection (note that scale is fixed due to the use of convex combinations)
- Handles parameter constraints nicely (only requires that \mathcal{S} be closed under convex combinations with expected sufficient statistics)
- Compatible with use of Monte Carlo draws from $X_n|Y_n$ rather than computation of $E_{\theta_{n-1}} [s(X_n, Y_n) | Y_n]$

The Bad

- Needs that $\bar{\theta}$ be available
- Not optimal for short (say, less than 1000 observations) data records

Why Working in \mathcal{S} Rather Than in Θ ?

Our original motivation was the Direction Of Arrival (DOA) model used in array processing, which is somewhat related to probabilistic PCA

- (X_t, Y_t) are jointly Gaussian with covariance matrix

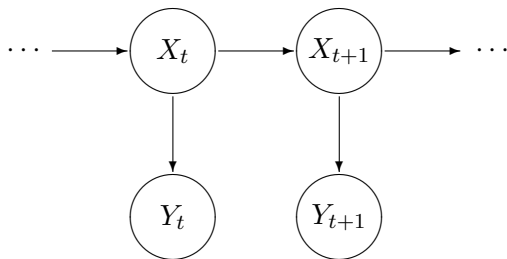
$$\mathbf{\Gamma}(\alpha, \omega, v) = \alpha \mathbf{a}(\omega) \mathbf{a}'(\omega) + v \mathbf{I}$$

M-Step

$$\bar{\theta}(\mathcal{S})|_{\omega} = \arg \min_{\omega} \left\{ \min_{\alpha, v} \log |\mathbf{\Gamma}(\alpha, \omega, v)| + \text{trace} (\mathbf{\Gamma}(\alpha, \omega, v)^{-1} \mathcal{S}) \right\}$$

Extension to Hidden Markov Models

When (X_t, Y_t) is assumed to follow an HMM, the law of the process is fully determined by $p_{\theta}(x_t, y_t | x_{t-1})$



We assume as previously that

$$p_{\theta}(x_t, y_t | x_{t-1}) = \exp(\langle s(x_{t-1}, x_t, y_t), \psi(\theta) \rangle - A(\theta))$$

wrt. to some measure

The Limiting EM Algorithm for HMMs

The EM update from n observations is now

$$S_{n,k} = \frac{1}{n} \sum_{t=1}^n \mathbb{E}_{\bar{\theta}(S_{k-1})} [s(X_{t-1}, X_t, Y_t) | Y_{1:n}]$$

Assuming (Y_t) to be strictly stationary with distribution π and (strong) forgetting properties on the model p_θ (C, 2009), the limiting EM recursion becomes

$$S_k = \mathbb{E}_\pi \left(\mathbb{E}_{\bar{\theta}(S_{k-1})} [s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty}] \right)$$

The Smoothing Dilemma

How to get (approximate) noisy observations of

$$E_{\pi} (E_{\theta} [s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty}])$$

without storing whole trajectories?

A natural idea is to consider computable

- limited-memory approximations $E_{\theta} [s(X_{t-1}, X_t, Y_t) | Y_{t-l:t+d}]$
- fixed-lag smoothing approximations $E_{\theta} [s(X_{t-1}, X_t, Y_t) | Y_{-\infty:t+d}]$

but the corresponding limiting mappings

$$E_{\pi} (E_{\theta} [s(X_{-1}, X_0, Y_0) | Y_{t-l:t+d}]) \text{ or}$$

$E_{\pi} (E_{\theta} [s(X_{-1}, X_0, Y_0) | Y_{-\infty:t+d}])$ may have different convergence behaviour and spurious stationary points

Our Proposal

(C, 2009)

Use the recursive smoothing trick of (Zeitouni & Dembo, 1988), (Elliot *et al.*, 1995):

$$\begin{aligned} E_{\theta} [s(X_{t-1}, X_t, Y_t) | Y_{1:n}, X_n = x] = \\ \sum_{x'} E_{\theta} [s(X_{t-1}, X_t, Y_t) | Y_{1:n-1}, X_{n-1} = x'] \\ P_{\theta} (X_{n-1} = x' | Y_{1:n-1}, X_n = x) \end{aligned}$$

for $t \leq n - 1$ (assuming discrete \mathcal{X})

The resulting algorithm appears to work well in practice; it generalises the algorithm of (Mongillo & Denève, 2008) but can no more be analysed using standard stochastic approximation theory

Online EM Algorithm for HMMs

Compute, for $x \in \mathcal{X}$,

$$\hat{\phi}_n(x) = \frac{\sum_{x' \in \mathcal{X}} \hat{\phi}_{n-1}(x') q_{\hat{\theta}_{n-1}}(x', x) g_{\hat{\theta}_{n-1}}(x, Y_n)}{\sum_{x', x'' \in \mathcal{X}^2} \hat{\phi}_{n-1}(x') q_{\hat{\theta}_{n-1}}(x', x'') g_{\hat{\theta}_{n-1}}(x'', Y_n)}$$

$$\hat{\rho}_n(x) = \sum_{x' \in \mathcal{X}} \{ \gamma_n s(x', x, Y_n) + (1 - \gamma_n) \hat{\rho}_{n-1}(x') \} \frac{\hat{\phi}_{n-1}(x') q_{\hat{\theta}_{n-1}}(x', x)}{\sum_{x'' \in \mathcal{X}} \hat{\phi}_{n-1}(x'') q_{\hat{\theta}_{n-1}}(x'', x)}$$

Update the parameter according to

$$\hat{\theta}_n = \bar{\theta} \left(\sum_{x \in \mathcal{X}} \hat{\rho}_n(x) \hat{\phi}_n(x) \right)$$

- Cappé, O. (2009). Online EM algorithm for hidden Markov models. Preprint.
- Cappé, O., Douc, R., Guillin, A., Marin, J-M. & Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459.
- Cappé, O. & Moulines, E. (2009). On-line expectation-maximization algorithm for latent data models. *J. Roy. Statist. Soc. B*, 71(3):593-613.
- Elliott, R. J., Aggoun, L., & Moore, J. B. (1995). *Hidden Markov Models: Estimation and Control*. Springer, New York.
- Liang, P. & Klein, D. (2009). Online EM for Unsupervised Models. *Proceedings NAACL Conference*.
- Mongillo, G. & Denève, S. (2008). Online learning with hidden Markov models. *Neural Computation*, 20(7):1706-1716.
- Neal, R. M. & Hinton, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in graphical models*, pages 355–368. MIT Press, Cambridge, MA, USA.
- Sato, M. (2000). Convergence of on-line EM algorithm. In *Proceedings of the International Conference on Neural Information Processing*, vol. 1, pages 476–481.
- Sato, M. & Ishii, S. (2000). On-line EM algorithm for the normalized Gaussian network. *Neural Computation*, 12:407-432.
- Zeitouni, O. & Dembo, A. (1988). Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes. *IEEE Trans. Inform. Theory*, 34(4).