Online EM Algorithm for Hidden Markov Models

Mohamed Benyahia[1,2] Lilian Say[1]

[1]ENS Paris-Saclay

[2]Télécom Paris

January 8, 2025

Introduction

Online version of EM for HMMs

- Allow for a continuous adaptation of the parameters along a potentially infinite data stream.
- Mongillo and Denève's approach [3] in the case of finite-valued observations.

Contributions

- General HMMs, with possibly continuous observations
- Recursive computation of smoothing functionals in E-step
- Convergence of the EM update for HMMs.
- Convergence of the auxiliary quantity in the E-step.

Online EM for HMMs

Idea: Use a recursive approach for the E-Step:

$$S = \frac{1}{n}\mathbb{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \mid Y_{0:n}\right]$$

Assumptions

- Exponential Family Representation:

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

- Explicit M-step:
  $\bar{\theta}(S)$ is the solution to the complete-data maximum likelihood equation

$$\nabla_\theta \psi(\theta) S - \nabla_\theta A(\theta) = 0$$

Online EM for HMMs

Algorithm 1

Iteration: For $n \geq 0$

- E-step: For $x \in \mathcal{X}$

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

$$\hat{\rho}_{n+1}(x) = \sum_{x' \in \mathcal{X}} \left[\gamma_{n+1} s(x', x, Y_{n+1}) + (1 - \gamma_{n+1}) \hat{\rho}_n(x')\right] \hat{r}_n(x' \mid x)$$

- M-step:

$$\hat{\theta}_{n+1} = \bar{\theta}\left(\sum_{x \in \mathcal{X}} \hat{\rho}_{n+1}(x) \hat{\phi}_{n+1}(x)\right)$$

Convergence

Assumptions

- **Finite state space:** X is finite.
- **Compact parameter space:** Θ is compact, and $\theta^* \in \mathring{\Theta}$.
- **Regular transition matrix:** $q_\theta(x, x') \geq \epsilon > 0 \quad \forall\, \theta \in \Theta,\, x, x' \in X$
- **Bounded marginal observation likelihood:** $\bar{g}_\theta(y)$ is bounded, and:

$$\sup_\theta \sup_y \bar{g}_\theta(y) < \infty, \quad \mathbb{E}_{\theta^*}\left[\left|\log \inf_\theta \bar{g}_\theta(Y_0)\right|\right] < \infty,$$

  where $\bar{g}_\theta(y) = \sum_x g_\theta(x, y)$.

- **Smoothness:** ψ, A are continuously differentiable on $\mathring{\Theta}$.

Convergence

Theorem (Convergence of Expected Sufficient Statistics)

$$\frac{1}{n}\mathbb{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \mid Y_{0:n}\right] \xrightarrow{P_{\theta^*}\ a.s.} \mathbb{E}_{\theta^*}\left(\mathbb{E}_\theta\left[s(X_{-1}, X_0, Y_0) \mid Y_{-\infty:\infty}\right]\right).$$

Theorem (Fixed Points of the Limiting EM Algorithm)

The limiting EM update:

$$\theta_{k+1} = \bar{\theta}\left\{\mathbb{E}_{\theta^*}\left(\mathbb{E}_{\theta_k}\left[s(X_{-1}, X_0, Y_0) \mid Y_{-\infty:\infty}\right]\right)\right\},$$

has fixed points that are the stationary points of the limiting likelihood contrast function $c_{\theta^*}(\theta)$*which is defined as:*

$$c_{\theta^*}(\theta) = \mathbb{E}_{\theta^*}\Big[\log p_\theta(Y_0 \mid Y_{-\infty:-1})\Big].$$

Convergence

- Additionally, Fisher's identity provides a key decomposition for the gradient:

$$\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{\nu,\theta}\left[\nabla_\theta \log p_\theta(X_t, Y_t \mid X_{t-1}) \mid Y_{0:n}\right] \xrightarrow{P_{\theta^*}\text{ a.s.}} \nabla_\theta c_{\theta^*}(\theta).$$

Corollary (Convergence of intermediate quantity)

If $\theta_n = \theta$ for all n (Parameter freeze) then:

$$\hat{\rho}_n(x) \xrightarrow{P_{\theta^*}\ a.s.} \mathbb{E}_{\theta^*}\left[\mathbb{E}_\theta\left[s(X_{-1}, X_0, Y_0) \mid Y_{-\infty:\infty}\right]\right], \ \forall x$$

This ensures stability of $\hat{\rho}_n(x)$, even though other terms like $\hat{\phi}_n(x)$ may depend on observations.

Experiments: Markov Chain observed in Gaussian Noise

- $Y_t = X_t + V_t$, where :
  - $X_t \in \mathcal{X} = \{0, 1\}$ is the hidden Markov chain
  - $V_t \sim \mathcal{N}(0, \nu)$

Algorithm 2

Iteration: For $n \geq 0$

- E-step: For $i, j, k \in \mathcal{X}$ and $0 \leq d \leq 2$

$$\hat{\phi}_{n+1}(k) = \frac{\sum_{k'} \hat{\phi}_n(k')\hat{q}_n(k', k)g_{\hat{\theta}_n}(k, Y_{n+1})}{\sum_{k',k''} \hat{\phi}_n(k')\hat{q}_n(k', k'')g_{\hat{\theta}_n}(k'', Y_{n+1})}$$

$$\hat{\rho}^q_{n+1}(i, j, k) = \gamma_{n+1}\delta(j - k)\hat{r}_{n+1}(i|j) + (1 - \gamma_{n+1})\sum_{k'} \hat{\rho}^q_n(i, j, k')\hat{r}_{n+1}(k'|k)$$

$$\hat{\rho}^g_{n+1,d}(i, k) = \gamma_{n+1}\delta(i - k)Y^d_{n+1} + (1 - \gamma_{n+1})\sum_{k'} \hat{\rho}^g_{n,d}(i, k')\hat{r}_{n+1}(k'|k)$$

Experiments: Markov Chain observed in Gaussian Noise

Algorithm 2

- M-step: For $i, j \in \mathcal{X}$ and $0 \leq d \leq 2$

$$\hat{S}^q_{n+1}(i,j) = \sum_{k'=1}^{m} \hat{\rho}_q^{n+1}(i,j,k')\hat{\phi}^{n+1}(k')$$

$$\hat{S}^g_{n+1,d}(i) = \sum_{k'=1}^{m} \hat{\rho}^g_{n+1,d}(i,k')\hat{\phi}_{n+1}(k')$$

$$\hat{q}_{n+1}(i,j) = \frac{\hat{S}^q_n(i,j)}{\sum_{j'} \hat{S}^q_n(i,j')}$$

$$\hat{\mu}_{n+1}(i) = \frac{\hat{S}^g_{n,1}(i)}{S^g_{n,0}(i)}, \quad \hat{\upsilon}_{n+1} = \frac{\sum_{i=1}^m \left(\hat{S}^g_{n,2}(i) - \hat{\mu}^2(i)\hat{S}^g_{n,0}(i)\right)}{\sum_{i=1}^m \hat{S}^g_{n,0}(i)}$$

Experiments: Markov Chain observed in Gaussian Noise

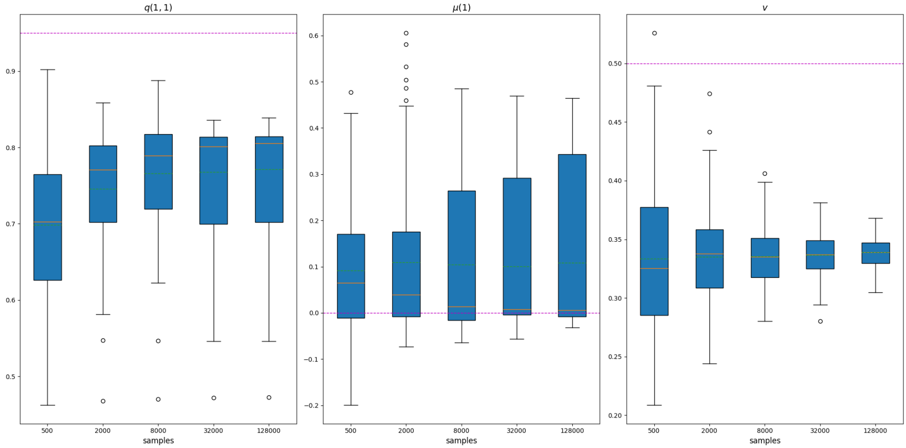

Figure: Estimation results when using the online EM algorithm with $\gamma_n = n^{-0.6}$.

Experiments: Markov Chain observed in Gaussian Noise

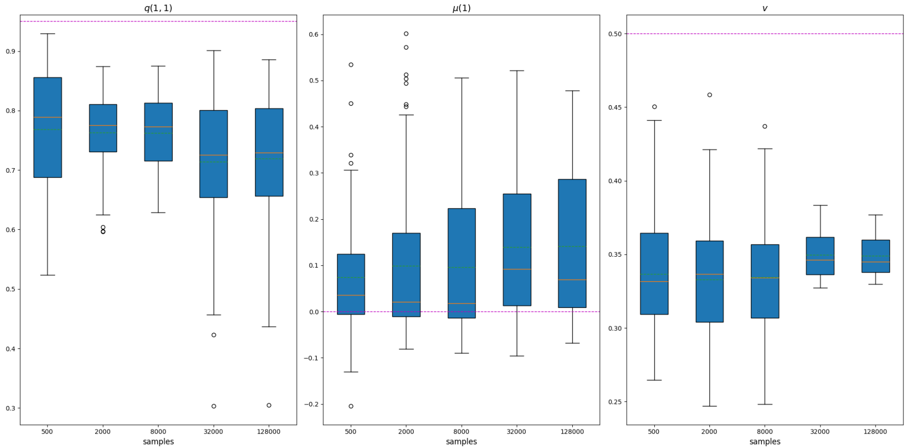

Figure: Estimation results when using the online EM algorithm with $\gamma_n = 0.01$ for $n \leq n_0$ and $\gamma_n = 0.5(n - n_0)^{-1}$ for $n > n_0$, with $n_0 = 10000$.

Experiments: Markov Chain observed in Gaussian Noise

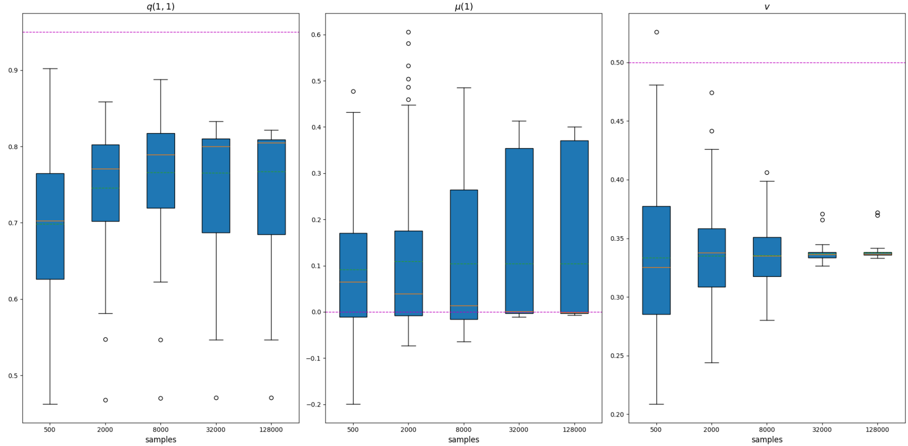

Figure: Estimation results when using the online EM algorithm with $\gamma_n = n^{-0.6}$ with Polyak-Ruppert averaging started after $n = 8000$.

Conclusion

Strengths

- General HMMs
- Better than batch EM for large datasets and small state sets

Limitations

- Exponential Family assumption
- Explicit M-Step :

$$\bar{\theta} : s \mapsto \arg\max_{\theta \in \Theta} \{ \langle \psi(\theta), S \rangle - A(\theta) \}$$

References

[1] O. Cappé, "Online EM Algorithm for Hidden Markov Models," *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 728–749, 2011.

[2] O. Cappé and E. Moulines, "On-Line Expectation–Maximization Algorithm for latent Data Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[3] G. Mongillo and S. Deneve, "Online Learning with Hidden Markov Models," *Neural Computation*, vol. 20, no. 7, pp. 1706–1716, 2008.