Divergence based motivation for online EM and combining hidden variable models

Ehsan Amid and Manfred K. Warmuth

Google Brain, Mountain View
University of California, Santa Cruz

eamid@ucsc.edu, manfred@ucsc.edu

UAI 2020

Batch EM Setup

Given samples $\mathcal{V} = \{v_n\}_{n=1}^{N}$ from data distribution $p_{\mathsf{d}}(v)$

Minimize the **negative log-likelihood** (NLL)

$$\mathcal{L}(\widetilde{\Theta}|\,\mathcal{V}) = -{}^{1}\!/\!{}_{N} \sum_{n} \log \underbrace{p(v_n|\,\widetilde{\Theta})}_{\int_h p(h,v_n|\,\widetilde{\Theta})}$$

where

- v denotes the **visible variable**
- h denotes the **hidden variable**
- $\widetilde{\Theta}$ denotes the **model parameters**

Batch EM Upper-bound

It is easier to minimize an **upper-bound** of NLL

$$\mathrm{U}_{\Theta}(\widetilde{\Theta}|\mathcal{V}) := \mathcal{L}(\widetilde{\Theta}|\,\mathcal{V}) + \underbrace{{}^{1}\!/_{N} \sum_{n} \int_{h} p(h|v_n, \Theta) \log \frac{p(h|v_n, \Theta)}{p(h|v_n, \widetilde{\Theta})}}_{\geq 0}$$

$$= -{}^{1}\!/_{N} \sum_{n} \mathbb{E}_{p(h|\,v_n,\Theta)}\left[\log p(h, v_n|\,\widetilde{\Theta})\right] - \underbrace{{}^{1}\!/_{N} \sum_{n} \mathbb{H}_{\Theta_n}(H|\,v_n)}_{\text{const. wrt } \widetilde{\Theta}}$$

E-Step Calculate the posteriors $p(h|v_n, \Theta)$

M-Step Minimize the upper-bound $\mathrm{U}_{\Theta}(\widetilde{\Theta}|\mathcal{V})$

Batch EM Upper-bound

It is easier to minimize an **upper-bound** of NLL

$$\mathrm{U}_\Theta(\widetilde{\Theta}|\mathcal{V}) := \mathcal{L}(\widetilde{\Theta}|\,\mathcal{V}) \; + \underbrace{{}^1\!/\!_N \sum_n \int_h p(h|v_n,\Theta)\log\frac{p(h|v_n,\Theta)}{p(h|v_n,\widetilde{\Theta})}}_{\geq 0}$$

$$= -\,{}^1\!/\!_N \sum_n \mathbb{E}_{p(h|\,v_n,\Theta)}\left[\log p(h, v_n|\,\widetilde{\Theta})\right] - \underbrace{{}^1\!/\!_N \sum_n \mathbb{H}_{\Theta_n}(H|\,v_n)}_{\text{const. wrt } \widetilde{\Theta}}$$

E-Step Calculate the posteriors $p(h|v_n,\Theta)$

M-Step Minimize the upper-bound $\mathrm{U}_\Theta(\widetilde{\Theta}|\mathcal{V})$

Batch EM Upper-bound

It is easier to minimize an **upper-bound** of NLL

$$\mathrm{U}_{\Theta}(\widetilde{\Theta}|\mathcal{V}) := \mathcal{L}(\widetilde{\Theta}|\mathcal{V}) + \underbrace{{}^{1}\!/\!_{N} \sum_{n} \int_{h} p(h|v_n,\Theta) \log \frac{p(h|v_n,\Theta)}{p(h|v_n,\widetilde{\Theta})}}_{\geq 0}$$

$$= -{}^{1}\!/\!_{N} \sum_{n} \mathbb{E}_{p(h|v_n,\Theta)}\left[\log p(h, v_n|\widetilde{\Theta})\right] - \underbrace{{}^{1}\!/\!_{N} \sum_{n} \mathbb{H}_{\Theta_n}(H|v_n)}_{\text{const. wrt } \widetilde{\Theta}}$$

E-Step Calculate the posteriors $p(h|v_n,\Theta)$

M-Step Minimize the upper-bound $\mathrm{U}_{\Theta}(\widetilde{\Theta}|\mathcal{V})$

Rewriting the EM Upper-bound

Consider the singleton distribution

$$p(h, v|\Theta_n) := \underbrace{\delta_{v_n}(v)}_{\text{Dirac measure at } v_n} \times p(h|v, \Theta)$$

Relative Entropy (RE) divergence between models

$$D_{\mathrm{RE}}(\Theta_n, \widetilde{\Theta}) := \int_{h,v} p(v, h|\Theta_n) \log \frac{p(v, h|\Theta_n)}{p(v, h|\widetilde{\Theta})}$$

$$= -\mathbb{H}_{\Theta_n}(H, V) - \int_{h,v} p(h, v|\Theta_n) \log p(h, v|\widetilde{\Theta})$$

$$= -\underbrace{\mathbb{H}_{\Theta_n}(H, V)}_{\text{const.}} - \mathbb{E}_{p(h|v_n,\Theta)}\left[\log p(h, v_n|\widetilde{\Theta})\right]$$

i.e.

$$\mathrm{U}_{\Theta}(\widetilde{\Theta}|\mathcal{V}) = {}^{1}\!/\!{}_{N} \sum_{n} D_{\mathrm{RE}}(\Theta_n, \widetilde{\Theta}) + \text{const.}$$

EM minimizes sum of RE divergences!

Rewriting the EM Upper-bound

Consider the singleton distribution

$$p(h, v|\Theta_n) := \underbrace{\delta_{v_n}(v)}_{\text{Dirac measure at } v_n} \times p(h|v, \Theta)$$

Relative Entropy (RE) divergence between models

$$D_{\text{RE}}(\Theta_n, \widetilde{\Theta}) := \int_{h,v} p(v, h|\Theta_n) \log \frac{p(v, h|\Theta_n)}{p(v, h|\widetilde{\Theta})}$$

$$= -\mathbb{H}_{\Theta_n}(H, V) - \int_{h,v} p(h, v|\Theta_n) \log p(h, v|\widetilde{\Theta})$$

$$= -\underbrace{\mathbb{H}_{\Theta_n}(H, V)}_{\text{const.}} - \mathbb{E}_{p(h|v_n, \Theta)}\left[\log p(h, v_n|\widetilde{\Theta})\right]$$

i.e.

$$\mathrm{U}_\Theta(\widetilde{\Theta}|\mathcal{V}) = {}^1\!/\!_N \sum_n D_{\text{RE}}(\Theta_n, \widetilde{\Theta}) + \text{const.}$$

EM minimizes sum of RE divergences!

Rewriting the EM Upper-bound

Consider the singleton distribution

$$p(h, v \mid \Theta_n) := \underbrace{\delta_{v_n}(v)}_{\text{Dirac measure at } v_n} \times p(h \mid v, \Theta)$$

Relative Entropy (RE) divergence between models

$$D_{\text{RE}}(\Theta_n, \widetilde{\Theta}) := \int_{h,v} p(v, h \mid \Theta_n) \log \frac{p(v, h \mid \Theta_n)}{p(v, h \mid \widetilde{\Theta})}$$

$$= -\mathbb{H}_{\Theta_n}(H, V) - \int_{h,v} p(h, v \mid \Theta_n) \log p(h, v \mid \widetilde{\Theta})$$

$$= -\underbrace{\mathbb{H}_{\Theta_n}(H, V)}_{\text{const.}} - \mathbb{E}_{p(h \mid v_n, \Theta)}\left[\log p(h, v_n \mid \widetilde{\Theta})\right]$$

i.e.

$$\mathrm{U}_{\Theta}(\widetilde{\Theta} \mid \mathcal{V}) = {}^{1}\!/_{N} \sum_{n} D_{\text{RE}}(\Theta_n, \widetilde{\Theta}) + \text{const.}$$

EM minimizes sum of RE divergences!

Online EM

At iteration t, receive a small batch $\mathcal{V}^t$ of samples

Minimize the NLL upper-bound plus an inertia term

$$\Theta^{t+1} = \arg\min_{\widetilde{\Theta}} \Big\{ \underbrace{\mathrm{U}_{\Theta^t}(\widetilde{\Theta}|\mathcal{V}^t)}_{\text{loss}} + {}^1\!/\eta^t \underbrace{D_{\mathrm{RE}}(\Theta^t, \widetilde{\Theta})}_{\text{inertia}} \Big\} \quad (1)$$

Inertia term keeps Θ^{t+1} close to Θ^t

Both terms have the same form as D_{RE}!

Equivalent to combining $|\mathcal{V}^t| + 1$ models as in batch EM!

Online EM

At iteration t, receive a small batch $\mathcal{V}^t$ of samples

Minimize the NLL upper-bound plus an inertia term

$$\Theta^{t+1} = \arg\min_{\widetilde{\Theta}} \Big\{ \underbrace{\mathrm{U}_{\Theta^t}(\widetilde{\Theta}|\mathcal{V}^t)}_{\text{loss}} + {}^1\!/\eta^t \underbrace{D_{\mathrm{RE}}(\Theta^t, \widetilde{\Theta})}_{\text{inertia}} \Big\} \qquad (1)$$

Inertia term keeps Θ^{t+1} close to Θ^t

Both terms have the same form as D_{RE}!

Equivalent to combining $|\mathcal{V}^t| + 1$ models as in batch EM!

Online EM Implications

(1) **Natural Gradient**

$$D_{\mathsf{RE}}(\Theta^t, \widetilde{\Theta}) \approx 1/2\, \mathrm{d}\widetilde{\Theta}^\top I_{\mathsf{F}}(\Theta^t)\, \mathrm{d}\widetilde{\Theta}$$

$I_{\mathsf{F}}(\Theta^t)$ is the Fisher information matrix

$$\Theta^{t+1} \approx \Theta^t - I_{\mathsf{F}}^{-1}(\Theta^t)\, \nabla \mathcal{L}(\mathcal{V}^t | \Theta^t)$$

(2) **Finite-sample Approximation**

$$D_{\mathsf{RE}}(\Theta^t, \widetilde{\Theta}) \approx \underbrace{-1/N' \sum_{n'} \mathbb{E}_{p(h|v_{n'}, \widetilde{\Theta})} \left[\log p(v_{n'}, h | \widetilde{\Theta}) \right]}_{\mathsf{U}_{\Theta^t}(\widetilde{\Theta}|\mathcal{V}') + \text{const.}} + \text{const.}$$

where the samples $\mathcal{V}' = \{v_{n'}\}_{n'=1}^{N'}$ are drawn from $p(v|\Theta^t)$

Approximate batch EM on $N + N'$ samples!

Online EM Implications

(1) **Natural Gradient**

$$D_{\mathsf{RE}}(\Theta^t, \widetilde{\Theta}) \approx {}^1\!/\!_2 \, \mathrm{d}\widetilde{\Theta}^\top I_{\mathsf{F}}(\Theta^t) \, \mathrm{d}\widetilde{\Theta}$$

$I_{\mathsf{F}}(\Theta^t)$ is the Fisher information matrix

$$\Theta^{t+1} \approx \Theta^t - I_{\mathsf{F}}^{-1}(\Theta^t) \, \nabla \mathcal{L}(\mathcal{V}^t | \Theta^t)$$

(2) **Finite-sample Approximation**

$$D_{\mathsf{RE}}(\Theta^t, \widetilde{\Theta}) \approx \underbrace{- {}^1\!/\!_{N'} \sum_{n'} \mathbb{E}_{p(h|\, v_{n'}, \widetilde{\Theta})} \Big[\log p(v_{n'}, h | \widetilde{\Theta}) \Big]}_{\mathsf{U}_{\Theta^t}(\widetilde{\Theta} | \mathcal{V}') + \text{const.}} + \text{const.}$$

where the samples $\mathcal{V}' = \{v_{n'}\}_{n'=1}^{N'}$ are drawn from $p(v|\,\Theta^t)$

Approximate batch EM on $N + N'$ samples!

Online EM Examples

- **Closed-form Updates**
  - Mixture of Exponential Family
  - Hidden Markov Model
  - Kalman Filter

Note: for models from exponential family, the upper-bound terms involve Bregman divergences

- **Approximate Updates**
  - Compound Dirichlet distribution

Note: updates apply iterative Newton's method

Online EM Examples

- **Closed-form Updates**
  - Mixture of Exponential Family
  - Hidden Markov Model
  - Kalman Filter

Note: for models from exponential family, the upper-bound terms involve Bregman divergences

- **Approximate Updates**
  - Compound Dirichlet distribution

Note: updates apply iterative Newton's method

Combining Models

Given the set of **local** model parameters $\{\Theta^{(m)}\}$, $m \in [M]$

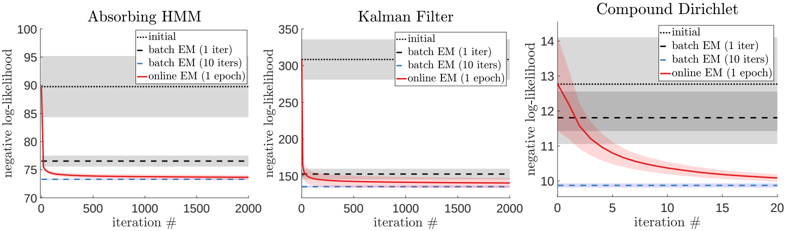
$$\Theta^{(\text{comb})} = \underset{\widetilde{\Theta}}{\arg\min} \sum_{m \in [M]} \alpha_m \, D_{\text{RE}}\big(\Theta^{(m)}, \widetilde{\Theta}\big)$$

where $\alpha_m \geq 0$ is the associated weight for model m

Note: for exponential family models, this corresponds to averaging Complete-data Sufficient Statistics

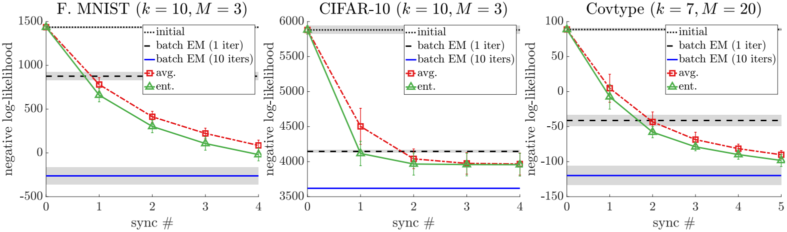
Experiments: Online EM

Online EM on synthetic data



Experiments: Combining Models

Gaussian mixture modeling in a distributed setting



Conclusion

- A unified view of the sample level and model level interpretation of the EM algorithm
- This allows us to:
  - derive updates for more complex models such as HMMs and Kalman filters
  - perform approximate updates (when necessary)
  - combine hidden variable models
- Long term goal: combining larger models such as GANs and autoencoders