LVMs (Contd), Expectation Maximization (2)

CS771: Introduction to Machine Learning

Piyush Rai

What is EM Doing?

Maximizing ILL

Assuming $\boldsymbol{Z}$ to be discrete, else replace it by an integral

- The MLE problem was $\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}} \log p(\boldsymbol{X}|\Theta) = \underset{\Theta}{\operatorname{argmax}} \log \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)$

Easier than maximizing ILL

- What EM (and ALT-OPT in a crude way) did is max of CLL: $\Theta_{MLE} = \underset{\Theta}{\operatorname{argmax}} \mathbb{E}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\Theta)]$
- But we did not solve the original problem. Is it okay?
- Assume $p_z = p(\boldsymbol{Z}|\boldsymbol{X}, \Theta)$ and $q(\boldsymbol{Z})$ to be some prob distribution over $\boldsymbol{Z}$, then

Function of a distribution q and parameter Θ

$$\log p(\boldsymbol{X}|\Theta) = \mathcal{L}(q, \Theta) + KL(q||p_z)$$

May verify this identity

- In the above $\mathcal{L}(q, \Theta) = \sum_Z q(Z)\log\left\{\frac{p(X,Z|\Theta)}{q(Z)}\right\}$ and $KL(q||p_z) = -\sum_Z q(\boldsymbol{Z})\log\left\{\frac{p(\boldsymbol{Z}|\boldsymbol{X},\Theta)}{q(\boldsymbol{Z})}\right\}$
- Since KL is always non-negative $\log p(\boldsymbol{X}|\Theta) \geq \mathcal{L}(q, \Theta)$, so $\mathcal{L}(q, \Theta)$ is a lower-bound on ILL
- Thus if we maximize $\mathcal{L}(q, \Theta)$, it will also improve $\log p(\boldsymbol{X}|\Theta)$

What is EM Doing?

- As we saw, $\mathcal{L}(\boldsymbol{q},\boldsymbol{\Theta})$ depends on $\boldsymbol{q}$ and $\boldsymbol{\Theta}$
- Let's maximize $\mathcal{L}(\boldsymbol{q},\boldsymbol{\Theta})$ w.r.t. $\boldsymbol{q}$ with $\boldsymbol{\Theta}$ fixed at $\boldsymbol{\Theta}^{\text{old}}$

Since $\log p(\boldsymbol{X}|\Theta) = \mathcal{L}(q,\Theta) + KL(q||p_z)$ is constant when Θ is held fixed at Θ^{old}

The posterior distribution of Z given older parameters Θ^{old} (will need this posterior to get the expectation of CLL)

$$\hat{q} = \text{argmax}_q \mathcal{L}(q,\Theta^{\text{old}}) = \text{argmin}_q KL(q||p_z) = p_z = p(\boldsymbol{Z}|\boldsymbol{X},\Theta^{\text{old}})$$

- Now let's maximize $\mathcal{L}(q,\Theta)$ w.r.t. Θ with q fixed at $\hat{q} = p_z = p(\boldsymbol{Z}|\boldsymbol{X},\Theta^{\text{old}})$

$$\Theta^{\text{new}} = \text{argmax}_\Theta \mathcal{L}(\hat{q},\Theta) = \text{argmax}_\Theta \sum_Z p(\boldsymbol{Z}|\boldsymbol{X},\Theta^{\text{old}}) \log\left\{\frac{p(\boldsymbol{X},\boldsymbol{Z}|\Theta)}{p(\boldsymbol{Z}|\boldsymbol{X},\Theta^{\text{old}})}\right\}$$

$$= \text{argmax}_\Theta \sum_Z p(\boldsymbol{Z}|\boldsymbol{X},\Theta^{\text{old}}) \log p(\boldsymbol{X},\boldsymbol{Z}|\Theta)$$

$$= \text{argmax}_\Theta \mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\Theta^{\text{old}})}[\log p(\boldsymbol{X},\boldsymbol{Z}|\Theta)]$$

Maximization of expected CLL w.r.t. the posterior distribution of Z given older parameters Θ^{old}

$$= \text{argmax}_\Theta \mathcal{Q}(\Theta,\Theta^{\text{old}})$$

The EM Algorithm in its general form..

- Maximization of $\mathcal{L}(q, \Theta)$ w.r.t. q and Θ gives the EM algorithm (Dempster, Laird, Rubin, 1977)

The EM Algorithm

1. Initialize Θ as $\Theta^{(0)}$, set $t = 1$
2. Step 1: Compute posterior of latent variables given current parameters $\Theta^{(t-1)}$

$$p(\boldsymbol{z}_n^{(t)} | \boldsymbol{x}_n, \Theta^{(t-1)}) = \frac{p(\boldsymbol{z}_n^{(t)} | \Theta^{(t-1)}) p(\boldsymbol{x}_n | \boldsymbol{z}_n^{(t)}, \Theta^{(t-1)})}{p(\boldsymbol{x}_n | \Theta^{(t-1)})} \propto \text{prior} \times \text{likelihood}$$

3. Step 2: Now maximize the expected complete data log-likelihood w.r.t. Θ

$$\Theta^{(t)} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{(t-1)}) = \arg\max_{\Theta} \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n^{(t)} | \boldsymbol{x}_n, \Theta^{(t-1)})} [\log p(\boldsymbol{x}_n, \boldsymbol{z}_n^{(t)} | \Theta)]$$

4. If not yet converged, set $t = t + 1$ and go to step 2.

- Note: If we can take the MAP estimate $\hat{z}_n$ of z_n (not full posterior) in Step 1 and maximize the CLL in Step 2 using that, i.e., do $\text{argmax}_\Theta \sum_{n=1}^{N} [\log p(\boldsymbol{x}_n, \hat{z}_n^{(t)} | \Theta)]$ this will be ALT-OPT

The Expected CLL

- Expected CLL in EM is given by (assume observations are i.i.d.)

$$\mathcal{Q}(\Theta, \Theta^{old}) = \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n,\Theta^{old})}[\log p(x_n, z_n|\Theta)]$$

$$= \sum_{n=1}^{N} \mathbb{E}_{p(z_n|x_n,\Theta^{old})}[\log p(x_n|z_n, \Theta) + \log p(z_n|\Theta)]$$

Was indeed the case of GMM: $p(\boldsymbol{z}_n|\Theta)$ was multinoulli, $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ was Gaussian

- If $p(\boldsymbol{z}_n|\Theta)$ and $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ are exp-family distributions, $\mathcal{Q}(\Theta, \Theta^{\text{old}})$ has a very simple form
- In resulting expressions, replace terms containing $\boldsymbol{z}_n$'s by their respective expectations, e.g.,
  - $\boldsymbol{z}_n$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \widehat{\Theta})}[\boldsymbol{z}_n]$
  - $\boldsymbol{z}_n\boldsymbol{z}_n^\top$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \widehat{\Theta})}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]$
- However, in some LVMs, these expectations are intractable to compute and need to be approximated (beyond the score of CS771)

EM: An Illustration

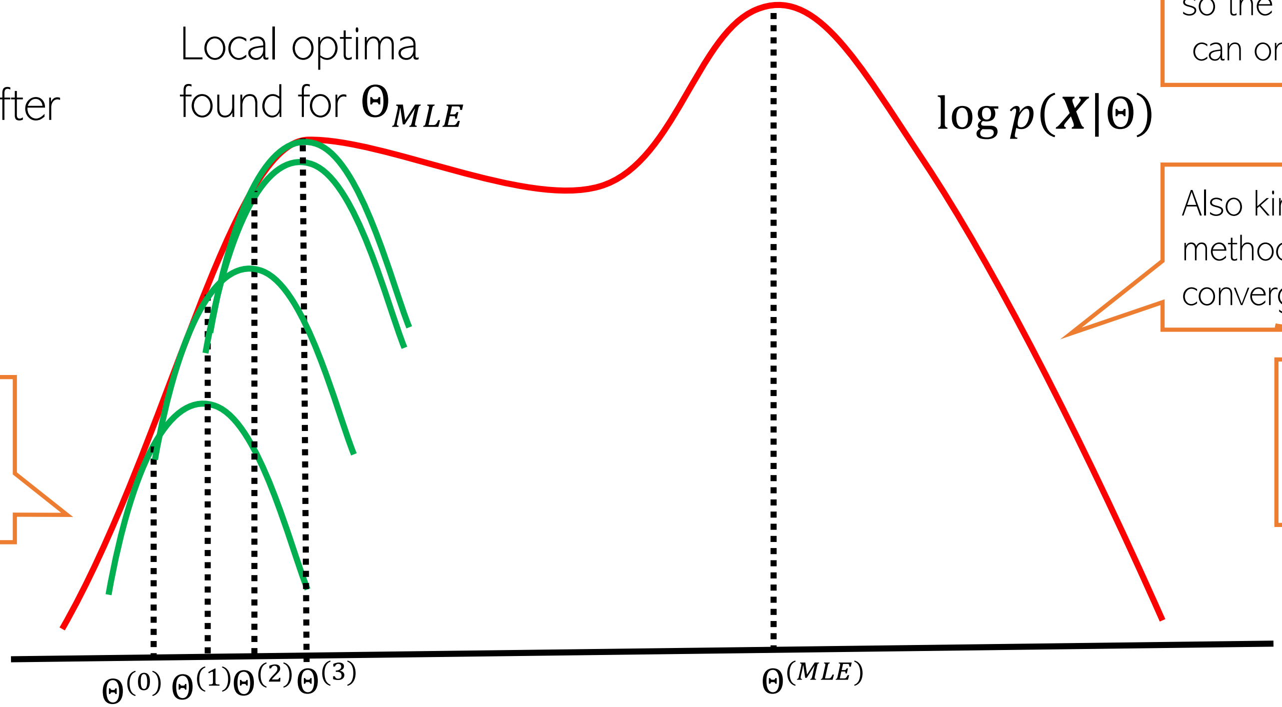
Alternating between them until convergence to some local optima

- As we saw, EM maximizes the lower bound $\mathcal{L}(q, \Theta)$ in two steps
- Step 1 finds the optimal q (call it $\hat{q}$) by setting it the posterior of $\boldsymbol{Z}$ given current Θ
- Step 2 maximizes $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ which gives a new Θ.

Makes $\mathcal{L}(q, \Theta)$ equal to $\log p(\boldsymbol{X}|\Theta)$; thus the curves touch at current Θ

Note that Θ only changes in Step 2 so the objective $\log p(\boldsymbol{X}|\Theta)$ can only change in Step 2

Green curve: $\mathcal{L}(\hat{q}, \Theta)$ after setting q to $\hat{q}$



Good initialization matters; otherwise would converge to a poor local optima

Also kind of similar to Newton's method (and has second order like convergence behavior in some cases)

Unlike Newton's method, we don't construct and optimize a quadratic approximation, but a lower bound

Even though original MLE problem $\text{argmax}_{\Theta} \log p(\boldsymbol{X}|\Theta)$ could be solved using gradient methods, EM often works faster and has cleaner updates

Recap: ALT-OPT vs EM

- ALT-OPT does the following
  1. Initialize $\Theta = \hat{\Theta}$
  2. Estimate $\mathbf{Z}$ as $\hat{\mathbf{Z}} = \arg\max_{\mathbf{Z}} \log p(\mathbf{Z}|\mathbf{X}, \hat{\Theta})$
  3. Estimate Θ as $\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}, \hat{\mathbf{Z}}|\Theta)$
  4. Go to step 2 if not converged

This step could potentially throw away a lot of information about the latent variable $\boldsymbol{Z}$

- EM addresses it using "soft" version of ALT-OPT
  1. Initialize $\Theta = \hat{\Theta}$
  2. Compute the posterior distribution of $\mathbf{Z}$, i.e., $p(\mathbf{Z}|\mathbf{X}, \hat{\Theta})$
  3. Estimate Θ by maximizing the expected CLL $\hat{\Theta} = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \hat{\Theta})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$
  4. Go to step 2 if not converged

ALT-OPT can be seen as as approximation of EM – the posterior $p(\boldsymbol{Z}|\boldsymbol{X}, \Theta)$ is replaced by a point mass at its mode

EM: Some Comments

- The E and M steps may not always be possible to perform exactly. Some reasons
  - Posterior of latent variables $p(\boldsymbol{Z}|\boldsymbol{X},\Theta)$ may not be easy to find and may require approx.
  - Even if $p(\boldsymbol{Z}|\boldsymbol{X},\Theta)$ is easy, expected CLL, i.e., $\mathbb{E}[\log p(\boldsymbol{X},\boldsymbol{Z}|\Theta)]$ may still not be tractable

$$\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)] = \int \log p(\mathbf{X},\mathbf{Z}|\Theta)p(\mathbf{Z}|\mathbf{X},\Theta)d\mathbf{Z}$$

Monte-Carlo EM

..and may need to be approximated, e.g., using Monte-Carlo expectation

Gradient methods may still be needed for this step

  - Maximization of the expected CLL may not be possible in closed form
- EM works even if the M step is only solved approximately (Generalized EM)
- If M step has multiple parameters whose updates depend on each other, they are updated in an alternating fashion - called Expectation Conditional Maximization (ECM) algorithm
- Other advanced probabilistic inference algorithms are based on ideas similar to EM
  - E.g., Variational Bayesian inference a.k.a. Variational Inference (VI)
- EM is also related to non-convex optimization algorithms Majorization-Maximization (MM)