Latent Variable Models and Expectation Maximization

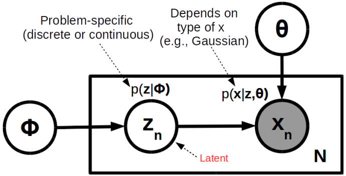
Piyush Rai

Introduction to Machine Learning (CS771A)

September 27, 2018

Recap: Latent Variable Models

- Assume each observation $\boldsymbol{x}_n$ to be associated with a "local" latent variable $\boldsymbol{z}_n$



- Parameters of $p(\boldsymbol{x}|\boldsymbol{z},\theta)$ and $p(\boldsymbol{z}|\phi)$ are collectively referred to as "global" parameters
- For brevity, we usually refer to the global parameters θ and ϕ as $\Theta = (\theta, \phi)$
- A Gaussian mixture model is an example of such a model
  - $\boldsymbol{z}_n \in \{1, \ldots, K\}$ with $p(\boldsymbol{z}_n|\phi) = \text{multinoulli}(\pi_1, \ldots, \pi_K)$
  - $\boldsymbol{x}_n \in \mathbb{R}^D$ with $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta) = \mathcal{N}(\boldsymbol{x}|\mu_{\boldsymbol{z}_n}.\Sigma_{\boldsymbol{z}_n})$
  - Here $\Theta = (\phi, \theta) = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$
- Given data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, the goal is to estimate the parameters Θ or latent variable $\mathbf{Z}$ or both (note: we can usually estimate Θ given $\mathbf{Z}$, and vice-versa)

Why Estimation is Difficult in LVMs?

- Suppose we want to estimate parameters Θ. If we knew both $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$ then we could do

$$\Theta_{MLE} = \arg\max_{\Theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) = \arg\max_{\Theta} \sum_{n=1}^{N} [\log p(\boldsymbol{z}_n|\phi) + \log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)]$$

- Simple to solve (usually closed form) if $p(\boldsymbol{z}_n|\phi)$ and $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)$ are "simple" (e.g., exp-fam. dist.)
- However, in LVMs where $\boldsymbol{z}_n$ is "hidden", the MLE problem will be the following

$$\Theta_{MLE} = \arg\max_{\Theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\Theta) = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta)$$

- The form of $p(\boldsymbol{x}_n|\Theta)$ may not be simple since we need to sum over unknown $\boldsymbol{z}_n$'s possible values

$$p(\boldsymbol{x}_n|\Theta) = \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) \quad \text{... or if } \boldsymbol{z}_n \text{ is continuous:} \quad p(\boldsymbol{x}_n|\Theta) = \int p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) d\boldsymbol{z}_n$$

- The summation/integral may be intractable + may lead to complex expressions for $p(\boldsymbol{x}_n|\Theta)$, in fact almost never an exponential family distribution. MLE for Θ won't have closed form solutions!

An Important Identity

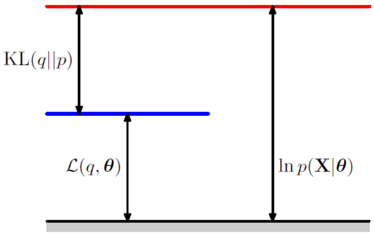
- Define $p_z = p(\mathbf{Z}|\mathbf{X}, \Theta)$ and let $q(\mathbf{Z})$ be some distribution over $\mathbf{Z}$
- Assume discrete $\mathbf{Z}$, the identity below holds for any choice of the distribution $q(\mathbf{Z})$

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

$$\mathcal{L}(q, \Theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \right\}$$

$$\text{KL}(q||p_z) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \Theta)}{q(\mathbf{Z})} \right\}$$

(Exercise: Verify the above identity)



- Since $\text{KL}(q||p_z) \geq 0$, $\mathcal{L}(q, \Theta)$ is a lower-bound on $\log p(\mathbf{X}|\Theta)$

$$\log p(\mathbf{X}|\Theta) \geq \mathcal{L}(q, \Theta)$$

- Maximizing $\mathcal{L}(q, \Theta)$ will also improve $\log p(\mathbf{X}|\Theta)$. Also, as we'll see, it's easier to maximize $\mathcal{L}(q, \Theta)$

Maximizing $\mathcal{L}(q, \Theta)$

- Note that $\mathcal{L}(q, \Theta)$ depends on two things $q(\mathbf{Z})$ and Θ. Let's do ALT-OPT for these
- First recall the identity we had: $\log p(\mathbf{X}|\Theta) = \mathcal{L}(q, \Theta) + \text{KL}(q||p_z)$ with

$$\mathcal{L}(q, \Theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \right\} \quad \text{and} \quad \text{KL}(q||p_z) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \Theta)}{q(\mathbf{Z})} \right\}$$

- Maximize $\mathcal{L}$ w.r.t. q with Θ fixed at Θ^{old}: Since $\log p(\mathbf{X}|\Theta)$ will be a constant in this case,

$$\hat{q} = \arg\max_q \mathcal{L}(q, \Theta^{old}) = \arg\min_q \text{KL}(q||p_z) = p_z = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$$

- Maximize $\mathcal{L}$ w.r.t. Θ with q fixed at $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$

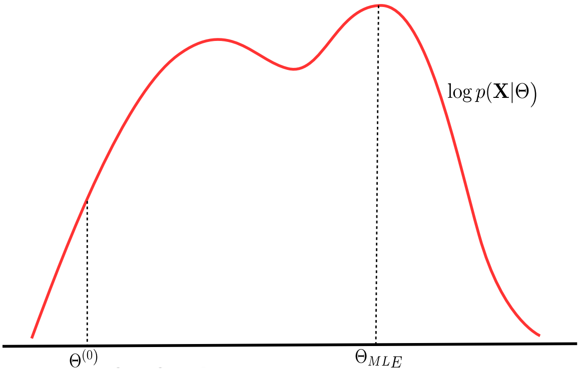
$$\Theta^{new} = \arg\max_\Theta \mathcal{L}(\hat{q}, \Theta) = \arg\max_\Theta \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{p(\mathbf{Z}|\mathbf{X}, \Theta^{old})} = \arg\max_\Theta \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \Theta^{old}) \log p(\mathbf{X}, \mathbf{Z}|\Theta)$$

.. therefore, $\boxed{\Theta^{new} = \arg\max_\theta \mathcal{Q}(\Theta, \Theta^{old})}$ where $\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$

- $\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$ is known as expected complete data log-likelihood (CLL)

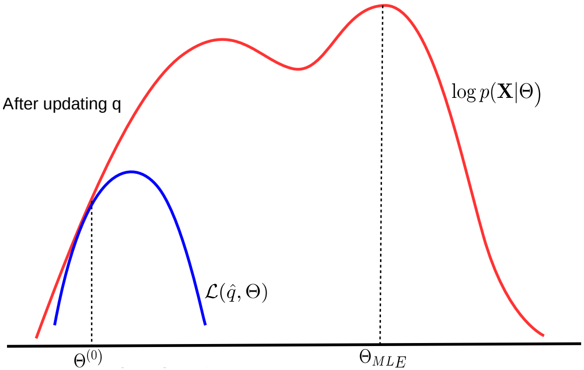
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)



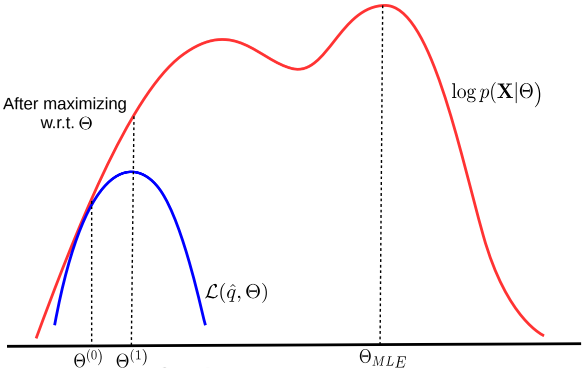
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)



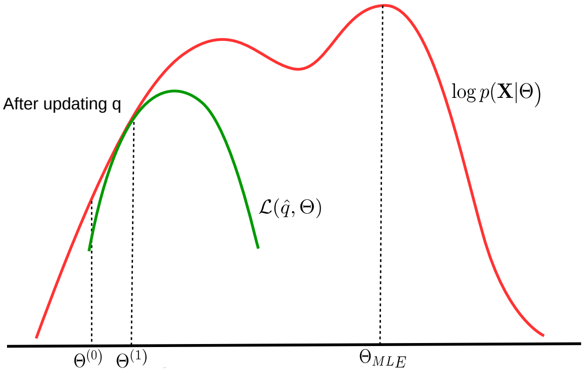
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)



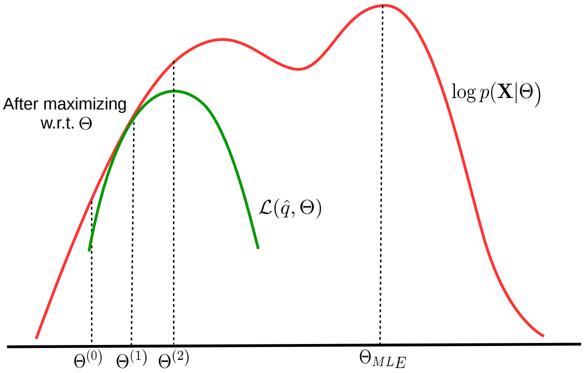
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)



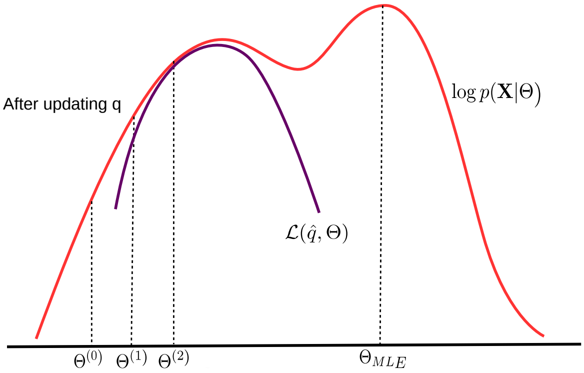
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)



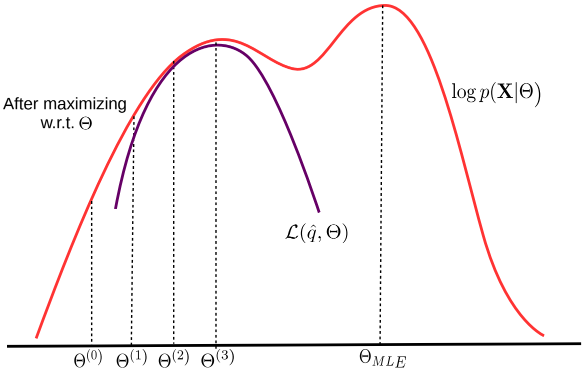
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)



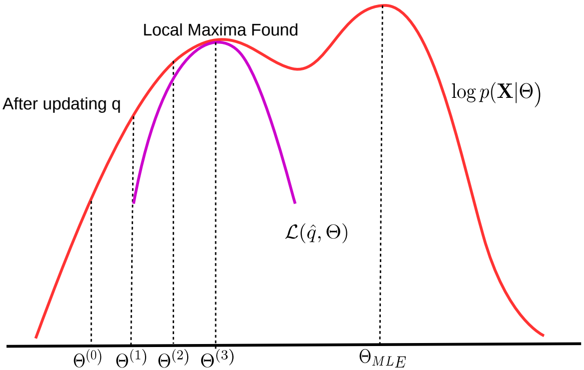
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)



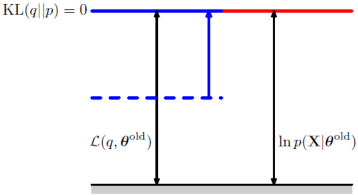
What's Going On: A Visual Illustration..

- Step 1: We set $\hat{q} = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$, $\mathcal{L}(\hat{q}, \Theta)$ touches $\log p(\mathbf{X}|\Theta)$ at Θ^{old}
- Step 2: We maximize $\mathcal{L}(\hat{q}, \Theta)$ w.r.t. Θ (equivalent to maximizing $\mathcal{Q}(\Theta, \Theta^{old})$)

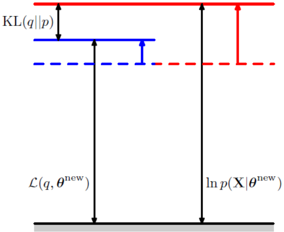


What's Going On: Another Illustration

- The two-step alternating optimzation scheme we saw can never decrease $p(\mathbf{X}|\Theta)$ (good thing)
- To see this consider both steps: (1) Optimize q given $\Theta = \Theta^{old}$; (2) Optimize Θ given this q



(Step 1)



(Step 2)

- Step 1 keeps Θ fixed, so $p(\mathbf{X}|\Theta)$ obviously can't decrease (stays unchanged in this step)
- Step 2 maximizes the lower bound $\mathcal{L}(q, \Theta)$ w.r.t Θ. Thus $p(\mathbf{X}|\Theta)$ can't decrease!

The Expectation Maximization (EM) Algorithm

The ALT-OPT of $\mathcal{L}(q, \Theta)$ that we saw leads to 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 $\hat{\boldsymbol{z}}_n$ of $\boldsymbol{z}_n$ (not full posterior) in Step 1 and maximize the CLL in Step 2 using that estimate, i.e., do $\arg\max_{\Theta} \sum_{n=1}^{N} \log p(\boldsymbol{x}_n, \hat{\boldsymbol{z}}_n^{(t)} | \Theta)$, this will be identical to ALT-OPT

Writing Down the Expected CLL

- Deriving the EM algorithm for any model requires finding the expression of the expected CLL

$$\begin{aligned}\mathcal{Q}(\Theta, \Theta^{old}) &= \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta)] \\ &= \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta) + \log p(\boldsymbol{z}_n|\Theta)]\end{aligned}$$

- If $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ and $p(\boldsymbol{z}_n|\Theta)$ are exp-family distributions, expected CLL will have a simple form
- Finding the expression for the expected CLL in such cases is fairly straightforward
  - First write down the expressions for $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \Theta)$ and $p(\boldsymbol{z}_n|\Theta)$ and simplify as much as possible
  - In the resulting expressions, replace all terms containing $\boldsymbol{z}_n$'s by their respective expectations, e.g.,
    - $\boldsymbol{z}_n$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\boldsymbol{z}_n]$, i.e., the posterior mean of $\boldsymbol{z}_n$
    - $\boldsymbol{z}_n\boldsymbol{z}_n^\top$ replaced by $\mathbb{E}_{p(\boldsymbol{z}_n|\boldsymbol{x}_n,\Theta^{old})}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]$
    - .. and so on..
- The expected CLL may not always be computable and may need to be approximated

EM for Gaussian Mixture Model

EM for Gaussian Mixture Model

- Let's first look at the CLL. Similar to generative classification with Gaussian class-conditionals

$$\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}[\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)] \qquad \text{(we've seen how we get this)}$$

- The expected CLL $\mathcal{Q}(\Theta, \Theta^{old})$ will be

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbb{E}[z_{nk}][\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\mu_k, \Sigma_k)]$$

.. where the expectation is w.r.t. the current posterior of $\boldsymbol{z}_n$, i.e., $p(\boldsymbol{z}_n|\boldsymbol{x}_n, \Theta^{old})$

- In this case, we only need $\mathbb{E}[z_{nk}]$ which can be computed as

$$\mathbb{E}[z_{nk}] = \gamma_{nk} = 0 \times p(z_{nk} = 0|\boldsymbol{x}_n, \Theta^{old}) + 1 \times p(z_{nk} = 1|\boldsymbol{x}_n, \Theta^{old}) = p(z_{nk} = 1|\boldsymbol{x}_n)$$

$$\propto p(z_{nk} = 1)p(\boldsymbol{x}_n|z_{nk} = 1) \qquad \text{(from Bayes Rule)}$$

$$\text{Thus} \quad \mathbb{E}[z_{nk}] \propto \pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \text{(Posterior prob. that } \boldsymbol{x}_n \text{ is generated by } k\text{-th Gaussian)}$$

- Note: We can finally normalize $\mathbb{E}[z_{nk}]$ as $\mathbb{E}[z_{nk}] = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)}$ since $\sum_{k=1}^{K} \mathbb{E}[z_{nk}] = 1$

EM for Gaussian Mixture Model

EM for Gaussian Mixture Model

1. Initialize $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ as $\Theta^{(0)}$, set $t = 1$
2. E step: compute the expectation of each $\boldsymbol{z}_n$ (we need it in M step)

$$\mathbb{E}[z_{nk}^{(t)}] = \gamma_{nk}^{(t)} = \frac{\pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})}{\sum_{\ell=1}^K \pi_\ell^{(t-1)} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_\ell^{(t-1)}, \boldsymbol{\Sigma}_\ell^{(t-1)})} \quad \forall n, k$$

3. Given "responsibilities" $\gamma_{nk} = \mathbb{E}[z_{nk}]$, and $N_k = \sum_{n=1}^N \gamma_{nk}$, re-estimate Θ via MLE

$$\boldsymbol{\mu}_k^{(t)} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk}^{(t)} \boldsymbol{x}_n$$

$$\boldsymbol{\Sigma}_k^{(t)} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk}^{(t)} (\boldsymbol{x}_n - \boldsymbol{\mu}_k^{(t)})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{(t)})^\top$$

$$\pi_k^{(t)} = \frac{N_k}{N}$$

4. Set $t = t + 1$ and go to step 2 if not yet converged

Another Example: (Probabilistic) Dimensionality Reduction

- Let's consider a latent factor model for dimensionality reduction (will revisit this later)

$$p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}, \sigma^2) = \mathcal{N}(\mathbf{W}\boldsymbol{z}_n, \sigma^2\mathbf{I}_D) \qquad p(\boldsymbol{z}_n) = \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$$

- A low-dim $\boldsymbol{z}_n \in \mathbb{R}^K$ mapped to high-dim $\boldsymbol{x}_n \in \mathbb{R}^D$ via a projection matrix $\mathbf{W} \in \mathbb{R}^{D\times K}$
- The complete data log-likelihood for this model will be

$$\log p(\mathbf{X}, \mathbf{Z}|\mathbf{W}, \sigma^2) = \log \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\mathbf{W}, \sigma^2) = \log \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}, \sigma^2)p(\boldsymbol{z}_n) = \sum_{n=1}^{N}\{\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}, \sigma^2) + \log p(\boldsymbol{z}_n)\}$$

- Plugging in the expressions for $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \mathbf{W}, \sigma^2)$ and $p(\boldsymbol{z}_n)$ and simplifying (exercise)

$$CLL = -\sum_{n=1}^{N}\left\{\frac{D}{2}\log\sigma^2 + \frac{1}{2\sigma^2}||\boldsymbol{x}_n||^2 - \frac{1}{\sigma^2}\boldsymbol{z}_n^\top\mathbf{W}^\top\boldsymbol{x}_n + \frac{1}{2\sigma^2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^\top\mathbf{W}^\top\mathbf{W}) + \frac{1}{2}\text{tr}(\boldsymbol{z}_n\boldsymbol{z}_n^\top)\right\}$$

- Expected CLL will require replacing $\boldsymbol{z}_n$ by $\mathbb{E}[\boldsymbol{z}_n]$ and $\boldsymbol{z}_n\boldsymbol{z}_n^\top$ by $\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]$
  - These expectations can be obtained from the posterior $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ (easy to compute due to conjugacy)
- The M step maximizes the expected CLL w.r.t. the parameters ($\mathbf{W}, \sigma^2$ in this case)

The EM Algorithm: Some Comments

- The E and M steps may not always be possible to perform exactly. Some reasons
  - The posterior of latent variables $p(\mathbf{Z}|\mathbf{X}, \Theta)$ may not be easy to find
    - Would need to approximate $p(\mathbf{Z}|\mathbf{X}, \Theta)$ in such a case
  - Even if $p(\mathbf{Z}|\mathbf{X}, \Theta)$ is easy, the expected CLL, i.e., $\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$ may still not be tractabe

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

    .. which can be approximated, e.g., using Monte-Carlo expectation (called Monte-Carlo EM)
  - 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 (VB) inference