IFT 6269: Probabilistic Graphical Models **Fall 2020**

Lecture 12 — October 9

Lecturer: Simon Lacoste-Julien *Scribe: Ismael Martinez, Binulal Narayanan*

Disclaimer: These notes have been quickly proofread by Simon Lacoste-Julien.

12.1 Expectation Maximization

The Expectation-maximization (EM) algorithm is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the models depend on unobserved latent or hidden variables z. Latent variables are variables that are not directly observed but are rather inferred from other variables that are observed. This model is described in Fig. 12.1.

Previous algorithms aimed at estimating the parameter θ that maximized the likelihood of $p(x;\theta)$, where x is the vector of observed variables. In the latent variable model, the probability of the observation x_i is obtained by marginalizing out its corresponding latent variable z_i. We thus wish to maximize the probability

$$\max_{\theta} p(x;\theta) = \sum_{z} p(x,z;\theta).$$

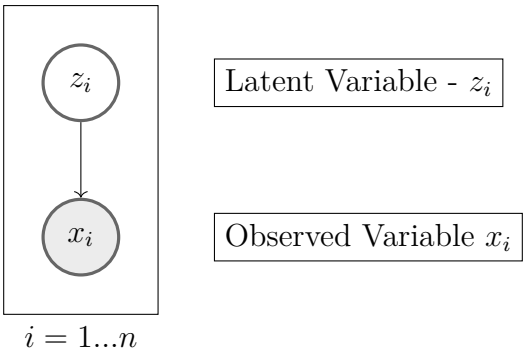

Figure 12.1: Latent variable model

The log likelihood of x given θ is

$$\log(x_{1:n};\theta) = \log\left(\prod_{i=1}^{n} p(x_i;\theta)\right) \tag{12.1}$$

$$= \sum_{i=1}^{n} \log\left(p(x_i;\theta)\right) \tag{12.2}$$

$$= \sum_{i=1}^{n} \log\left(\sum_{z_i} p(x_i, z_i;\theta)\right). \tag{12.3}$$

The sum within the logarithm (12.3) gives a multi-modal distribution, hence a multi-modal optimization problem. This is usually non-convex. Efficiently solving this non-convex optimization problem was the motivation for the EM algorithm.

Options for MLE in latent variable model.

1. Gradient Ascent on non-concave objective.
2. Expectation Maximization (EM).

Block-coordinate ascent is a maximization technique where we divide all function variables into groups, or *blocks*. For two blocks, we iteratively fix the values of the first block and maximize w.r.t. the second block, then fix the second block and maximize w.r.t. the first block.

EM is a block-coordinate ascent method on an auxiliary function which lower bounds $\log\ p(x_{1:n};\theta)$. We attempt to maximize the log-likelihood by maximizing this lower bound over θ. This has a nice interpretation in terms of filling in the missing data, and also the auxiliary function is often concave in θ which yields which yields a nicer optimization problem (and sometimes with closed-form updates). Since we have a distribution over z, the E step of EM can often be interpreted as filling this missing data with *soft values*.

E step – Fill z with "soft values".

M step - Solve the maximization problem w.r.t. θ for the fully observed model.

Comparison with Fisher LDA: We observed x (a Gaussian) and y (the label), and then we did a maximum log-likelihood with convex optimization. Similarly, the E-step fills in "missing values", and the M-step solves the MLE; these steps repeat.

Comparison with K-Mean: We iteratively solved a (hard) E-step which computed the cluster assignments z, followed by an M-step where we recomputed the cluster centroids μ.

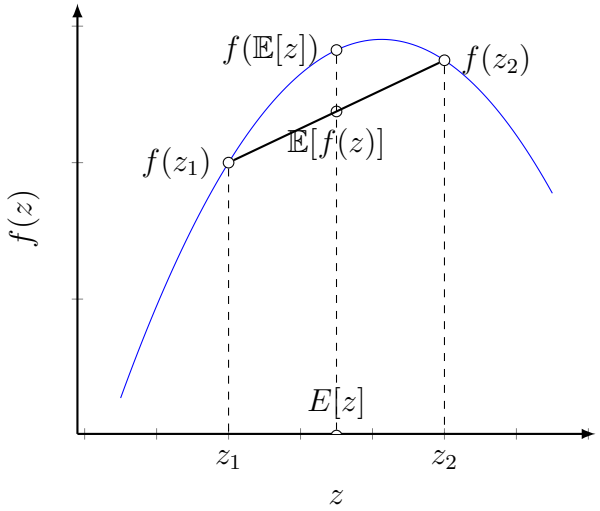

Figure 12.2: Jensen inequality: $\mathbb{E}_q[f(z)] \leq f(\mathbb{E}_q[z])$ when f is concave.

Jensen's Inequality for a concave function[1] is

$$\mathbb{E}_q[f(z)] \leq f\left(\mathbb{E}_q[z]\right)$$

where q is any fixed distribution. This is shown in Fig. 12.2.

Since the logarithm function is concave (see Fig. 12.3), we can use Jensen's inequality to simplify the log-likelihood calculation from (12.3):

$$\log\left(\sum_z p(x,z)\right) = \log\left(\sum_z \frac{q(z)p(x,z)}{q(z)}\right) \tag{12.4}$$

$$= \log\left(\mathbb{E}_q\left[\frac{p(x,z)}{q(z)}\right]\right) \tag{12.5}$$

$$\geq \mathbb{E}_q\left[\log\left(\frac{p(x,z)}{q(z)}\right)\right] \tag{12.6}$$

$$= \sum_z q(z)\ \log\left(\ p_\theta(x,z)\ \right) - \sum_z q(z)\ \log\left(\ q(z)\ \right) \tag{12.7}$$

$$\triangleq \mathcal{L}(q,\theta) \tag{12.8}$$

$$\triangleq \underbrace{\mathbb{E}_q\left[\log(\ p(x,Z;\theta)\ )\right]}_{\textit{Expected complete log-likelihood}} + \underbrace{H(q)}_{\textit{Entropy of } q} \tag{12.9}$$

We have $\log\ p(x;\theta) \geq \mathcal{L}(q,\theta)\ \forall q,\theta$ where q is any distribution over z with bigger support over z than $p(x,z)$, i.e. $p(x,z) \neq 0 \implies q(z) \neq 0$.[2] We thus see that $\mathcal{L}$ is the auxiliary function which lowers bound the log-likelihood, and it depends on both θ and q. From this formulation, we can iterate the maximization of $\mathcal{L}(q,\theta)$ with EM by alternating the maximization of the block of q, and the block of θ.

[1] $f(x)$ is concave $\iff$ $-f(x)$ is convex.

[2] Our convention in the derivation above is that $0/0 = 0$, and we can only allow this when $p(x,z) = 0$.

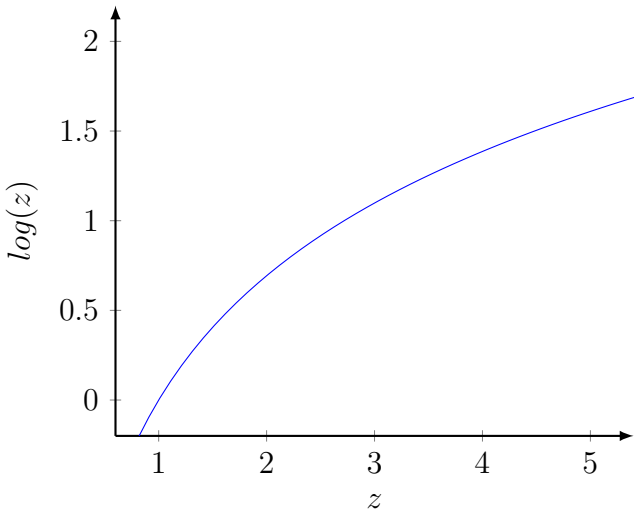

Figure 12.3: The log() function is strictly concave.

EM Algorithm

$$\textbf{E step:} \qquad q_{t+1} \triangleq \arg\max_{q} \mathcal{L}(q, \theta_t) \tag{12.10}$$

$$\implies q_{t+1}(z) = p(z|x;\theta_t) \tag{12.11}$$

$$\textbf{M step:} \qquad \theta_{t+1} \triangleq \arg\max_{\theta} \mathcal{L}(q_{t+1}, \theta) \tag{12.12}$$

$$= \arg\max_{\theta} \mathbb{E}_{q_{t+1}(z)}[\log\ p(x, z;\theta)]. \tag{12.13}$$

The M-step is another maximum likelihood problem, but for complete information! Often when z is a binary variable, we replace z with $\mathbb{E}_q[z]$ is this expression (these are the "soft values" we were referring to earlier).

We now explain how we can derive the closed form solution for q in the E-step as given in (12.11). Jensen's Inequality can be generalized by replacing z by $g(z)$. If we use $f(x) = \log(x)$, then

$$\mathbb{E}_q[\, f(\, g(z)\,)\,] \leq f\,(\,\mathbb{E}_q[\, g(z)\,]\,) \tag{12.14}$$

$$\implies \log\,(\,\mathbb{E}_q[\, g(z)\,]\,) \geq \mathbb{E}_q[\,\log(\, g(z)\,)\,]. \tag{12.15}$$

It turns out that Jensen's inequality is a strict inequality when f is strictly concave, unless the random variable is degenerate. A degenerate distribution means the random variable has only one possible value. Since the logarithm function is strictly concave (Fig. 12.3), the Jensen's inequality will be a strict inequality unless the random variable is degenerate (i.e.

$g(z)$ is a constant).[3] In other words,

$$\begin{aligned} g(z) = \textit{ constant then } \frac{p(x,z)}{q(z)} &= \textit{ constant } \forall z \\ \implies q(z) &\propto p(x,z) \\ \text{i.e. } q^*(z) &= p(z|x;\theta) \\ \text{i.e. } \arg\max_\theta \mathcal{L}(q,\theta_t) &= p(z|x;\theta_t) \text{ and} \\ \mathcal{L}(q_{t+1},\theta_t) &= \log(\, p(x;\theta_t)\,) \end{aligned}$$

Choosing the distribution $q^*(z) = p(z|x;\theta)$ maximizes the auxiliary function w.r.t. q and makes the lower bound tight, i.e.

$$\mathcal{L}(q_{t+1},\theta_t) = \log\ p(x;\theta_t) \geq \mathcal{L}(q,\theta_t)\ \forall\, q \tag{12.16}$$

$$\implies q_{t+1} \text{maximizes}\ \ \mathcal{L}(q,\theta_t) \text{ w.r.t. q} \tag{12.17}$$

12.1.1 Properties of EM

a) Log likelihood is non-decreasing, i.e. $\log\ p(x;\theta_{t+1}) \geq \log\ p(x;\theta_t)$.

Proof: The log likelihood is an upper bound on the auxiliary function.

$$\begin{aligned} \log\ p(x;\theta_{t+1}) &\geq \mathcal{L}(q_{t+1},\theta_{t+1}) \\ &\geq \mathcal{L}(q_{t+1},\theta_t) \quad \text{(By definition of } \theta_{t+1}\text{)} \\ &= \log\ p(x;\theta_t) \quad \text{(Auxiliary function at } (q_{t+1},\ \theta_t) \text{ gives the log likelihood)} \end{aligned}$$

b) Θ_t in EM converges to a stationary point of $\log p(x;\theta)$; i.e.,

$$\nabla_\theta \log\ p(x;\theta) = 0.$$

Like K-Means, initialization is crucial. Usually we do multiple random restarts. For GMM, could use K-Means++ to initialize the means μ_j.

c) By definition, $\mathcal{L}(q,\theta) = \mathbb{E}_q[\frac{\log\ p(x,z;\theta)}{q(z)}]$. Therefore,

$$\begin{aligned} \log p(x;\theta) - \mathcal{L}(q,\theta) &= -\mathbb{E}_q\left[\frac{\log p(x,z;\theta)}{q(z)p(x;\theta))}\right] \\ &= \mathbb{E}_q\left[\frac{\log q(z)}{p(z|x;\theta)}\right] \\ &\triangleq KL(q(\cdot)\|p(\cdot|x;\theta) \quad \textit{(KL Divergence)} \end{aligned}$$

[3] Another possibility is for q to be degenerate. Our derivation assumed that the support of q is including the one for p, so unless p is degenerate (in which case the sum over z collapsed to only one value and everything is trivial), then q cannot be degenerate.

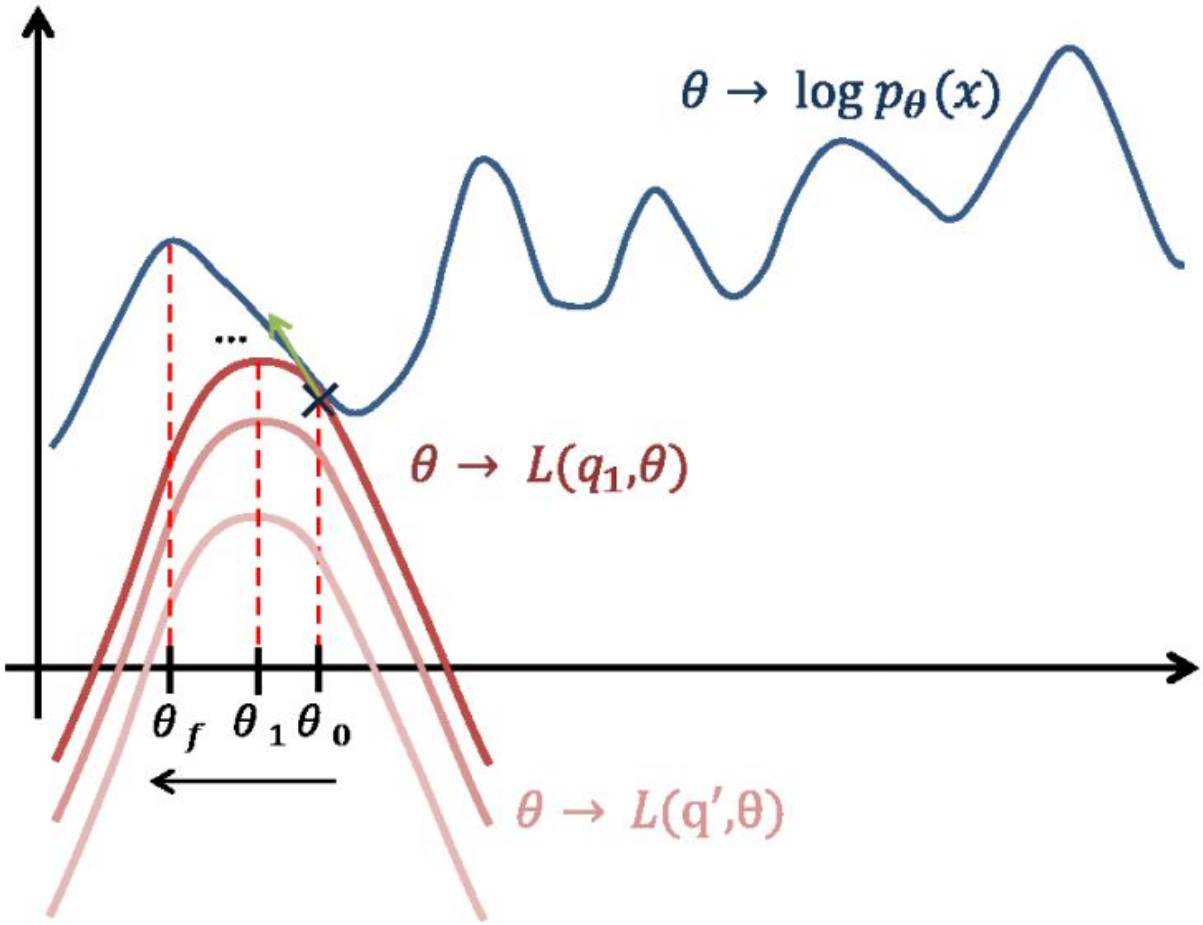

Figure 12.4: EM will iteratively find a parameter that improves the objective function.

Thus the difference between the log-likelihood and the lower bound at a specific θ is given by the KL divergence between q and $p(\cdot|x;\theta)$ (highlighting again why you can make the bound tight by letting $q = p(\cdot|x;\theta)$). We will revisit this KL formulation when we will talk about variational inference. In particular, variational EM replaces the maximization over all q with the maximization over a simpler subset $\mathcal{Q}$, thus giving an approximation of the E-step.

12.1.2 Gaussian Mixture Model

In a latent variable model, we have be a pair (x_i, z_i), $i \in \{1, ..., n\}$ of observed and unobserved nodes respectively (see Fig. 12.1). Suppose $z_i \sim Mult(\pi)$ is one-hot encoded, with $\pi = \pi_1, ..., \pi_k$ and $(x_i, |z_i = j) \sim \mathcal{N}(\mu_j, \Sigma_j)$. Here we have $\theta = (\pi, (\mu_i)_{i=1}^n, (\Sigma_i)_{i=1}^n)$, $x = x_{1:n}$, $z = z_{1:n}$.

From the latent variable model, we can show that

$$p(z|x) = \prod_{i=1}^{n} p(z_i|x) = \prod_{i=1}^{n} p(z_i|x_i).$$

Complete log-likelihood

$$\begin{aligned}
\log p(x, z; \theta) &= \sum_{i=1}^{n} \left[\log \underbrace{p(x_i|z_i;\theta)}_{\text{Gaussian}} + \log \underbrace{p(z_i;\theta)}_{\text{Multinouilli}} \right] \\
&= \sum_{i=1}^{n} \left[\sum_{j=1}^{k} z_{ij} \log \mathcal{N}(x_i|\mu_j, \Sigma_j) + \sum_{j=1}^{k} z_{ij} \log \pi_j \right] \\
\mathbb{E}_q[\log p(x, z; \theta)] &= \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{E}_q[z_{ij}] (\log \mathcal{N}(x_i|\mu_j, \Sigma_j) + \log \pi_j) \\
\mathbb{E}_q[z_{ij}] &= q(z_{ij} = 1) \qquad \text{[marginal distribution]}
\end{aligned}$$

During EM, $q_{t+1}(z) = p(z|x;\theta_t)$.

We define the distribution q as weights $\tau_{ij} \triangleq p(z_{ij} = 1|x_i;\theta_t) = q_{t+1}(z_{ij} = 1)$.

Compute

$$\begin{aligned}
q_{t+1}(z) &\triangleq p(z|x;\theta_t) \\
&= \prod_{i=1}^{n} p(z_i|x_i;\theta_t) \\
\implies q_{t+1}(z_i) &\propto p(x_i|z_i;\theta_t) p(z_i;\theta_t) \\
\tau_{ij}^{(t)} = q_{t+1}(z_{ij} = 1) &= \frac{\pi_j^{(t)} \mathcal{N}(x_i|\mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{j=1}^{k} \pi_l^{(t)} \mathcal{N}(x_i|\mu_l^{(t)}, \Sigma_l^{(t)})} \quad \left(= \frac{p(x_i, z_{ij} = 1|\theta^{(t)})}{p(x_i|\theta^{(t)})} \right)
\end{aligned}$$

E step: Compute $\tau_{ij}^{(t)}$ for $i = 1, ..., n$ using $\theta^{(t)}$.
M step: Compute $\max_{\{\mu_j, \Sigma_j, \pi_j\}} \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{ij}^{(t)} [\log p(x_i|\mu_j, \Sigma_j) + \log \pi_j]$.
The M step yields the updated parameters

$$\hat{\pi}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(t)}}{n}$$
$$\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(t)} x_i}{\sum_{i=1}^{n} \tau_{ij}^{(t)}}$$
$$\hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}^{(t)} (x_i - \hat{\mu}_j^{(t+1)})(x_i - \hat{\mu}_j^{(t+1)})^T}{\sum_{i=1}^{n} \tau_{ij}^{(t)}}$$

Initialize GMM model: e.g.

$\mu_j^{(0)}$from K-Means++
$\Sigma_j^{(0)}$big spherical covariance$\Sigma_j^{(0)} = \underbrace{\sigma^2}_{\text{big}} I$
$\pi_j^{(0)}$: proportions from K-means++.

If you execute EM step in GMM with fixed covariance $\Sigma_j = \sigma^2 I$, and you let $\sigma^2 \to 0$ $\rightsquigarrow$ get K-means algorithm!