Latent Variables

- A system with observed data **X**
  - may be far easier to understand in terms of additional variables **Z** corresponding to **X**,
  - but they are not observed (**latent**).
- For example, in a mixture of Gaussians,
  - For a single sample **x**, the latent variable **z** specifies which Gaussian generated the sample **x**.
  - The *responsibility* is the **posterior** p(**z**|**x**).

Latent Variables

- A system with observed variables **X**
  - may be easier to understand with latent variables **Z**, but they are not observed (**latent**).
- Notations:
  - We denote the set of all observed data by **X**, in which the n^{th} row represents $x^T_{n.}$
  - Similarly we denote the set of all latent variables by **Z**, with a corresponding row z^T_n.
  - Note: we use lowercase symbol for single sample (**x**), matrix symbol for all data (**X**).

Learning a Latent Variable Model

- We find model parameters by maximizing the log-likelihood of observed data $\log p(\mathbf{X} \mid \theta)$.
- If we had complete data {**X**, **Z**}, we could easily maximize the *complete* data likelihood $p(\mathbf{X}, \mathbf{Z} \mid \theta)$.
- Unfortunately, with *incomplete* data (**X** only), we must marginalize over **Z**, so

$$\log p(\mathbf{X} \mid \theta) = \log \left[\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \theta) \right]$$

(the sum inside the log makes it hard.)

The EM Algorithm in General

- Expectation-Maximization (EM) is a general recipe for finding the parameters that maximize the (log-) likelihood of *latent* variable models
- To find a parameter θ that maximizes the likelihood $p(\mathbf{X} \mid \theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \theta)$, the EM algorithm first introduces a new (variable) distribution $q(\mathbf{Z})$ over the latent variables.
- A lower bound $\mathcal{L}(q, \theta)$ for the log-likelihood $\log p(\mathbf{X} \mid \theta)$ is established based on q and θ.
- Then, $q(\mathbf{Z})$ and θ are alternatingly updated (keeping the other fixed) so that $\mathcal{L}(q, \theta)$ is maximized (similar to coordinate ascent) until convergence.

The EM Algorithm in General

- Our goal is to maximize $p(\mathbf{X} \mid \theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \theta)$
- For ***any distribution*** $q(\mathbf{Z})$ over latent variables:

$$\begin{aligned}
\log p(\mathbf{X} \mid \theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X} \mid \theta) \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{p(\mathbf{Z} \mid \mathbf{X}, \theta)} \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})} \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \theta)} \\
&= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})} + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \theta)} \\
&= \mathcal{L}(q, \theta) + KL(q(\mathbf{Z}) \| p(\mathbf{Z} | \mathbf{X}, \theta)) \\
&\geq \mathcal{L}(q, \theta)
\end{aligned}$$

Note: KL Divergence

Let *p* and *q* be probability distributions of a random variable Z.

$$KL(q \,\|\, p) = \mathbb{E}_{z \sim q(z)} \left[\log \frac{q(z)}{p(z)} \right] = \sum_z q(z) \log \frac{q(z)}{p(z)}$$

$$= -\textstyle\sum_z q(z) \log p(z) + \sum_z q(z) \log q(z)$$

This is one way to measure the **dissimilarity** of two probability distributions.

Remarks: (note: the first can be proved using Jensen's inequality)

- $KL(q \,\|\, p) \geq 0, \quad \text{with equality iff } p = q.$
- $KL(q \,\|\, p) \neq KL(p \,\|\, q) \quad \text{in general}$

Background note: Jensen's Inequality

- If f is convex, then for any $\theta_i \;\; \text{s.t.} \;\; 0 \leq \theta_i \leq 1 \; (\forall i),$ $\theta_1 + \theta_2 + \cdots + \theta_k = 1$

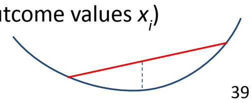
$$f(\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k) \leq \theta_1 f(x_1) + \cdots + \theta_k f(x_k)$$

- It can be seen as a generalization of the definition of convex function:

$$f \text{ is convex} \iff f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y) \text{ for all } 0 \leq \theta \leq 1$$

- Jensen's inequality can be written in expectation form (think of θ_i as probability mass for different outcome values x_i)

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$



Background note: Jensen's Inequality

- If f is convex, then for any $\theta_i \;\; \text{s.t.} \;\; 0 \leq \theta_i \leq 1 \; (\forall i),$ $\theta_1 + \theta_2 + \cdots + \theta_k = 1$

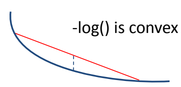
$$f(\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k) \leq \theta_1 f(x_1) + \cdots + \theta_k f(x_k)$$

- Jensen's inequality can be written in expectation form

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

- To show $KL(q \,\|\, p)$ is non-negative for any p, q, plug in *f*(...) = -log (...) and the following:

$$\theta_i = q(z), x_i = \frac{p(z)}{q(z)}$$



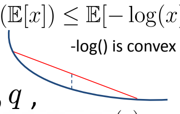
$$-\log(\mathbb{E}[x]) \leq \mathbb{E}[-\log(x)]$$

Non-negativity of KL divergence

- Jensen's inequality can be written in expectation form for a convex function f

$$f(\mathbb{E}[x]) \le \mathbb{E}[f(x)]$$

$$-\log(\mathbb{E}[x]) \le \mathbb{E}[-\log(x)]$$



- To show $KL(q \,\|\, p)$ is non-negative for any p, q, plug in *f*(...) = -log (...) and the following: $\theta_i = q(z), x_i = \frac{p(z)}{q(z)}$

$$KL(q\|p) = \sum_z q(z) \log(\frac{q(z)}{p(z)})$$
$$= \sum_z q(z) \left(-\log(\frac{p(z)}{q(z)})\right)$$
$$\ge -\log \left(\underbrace{\sum_z q(z) \frac{p(z)}{q(z)}}_{=\sum_z p(z)=1} \right)$$
$$= 0$$

Jensen's inequality for -log():

$$-\log(\mathbb{E}[x]) \le \mathbb{E}[-\log(x)]$$

i.e., plugin

$$-\log(\sum_i \theta_i x_i) \le \sum_i \theta_i(-\log(x_i))$$

with $\theta_i = q(z), x_i = \frac{p(z)}{q(z)}$

The EM Algorithm in a nutshell

- We have shown that: [variational lower bound]

$$\log p(\mathbf{X} \mid \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})} + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \theta)}$$
$$= \mathcal{L}(q, \theta) + KL(q(\mathbf{Z}) \,\|\, p(\mathbf{Z}|\mathbf{X}, \theta))$$
$$\ge \mathcal{L}(q, \theta)$$ Evidence Lower bound (ELBO) or variational lower bound

with equality holding if and only if $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$

- **EM algorithm**:
  Repeat alternating optimization until convergence:
  - E-step: for fixed θ, find *q* that maximizes $\mathcal{L}(q, \theta)$
  - M-step: for fixed *q*, find θ that maximizes $\mathcal{L}(q, \theta)$

* E: expectation
* M: maximization

The EM Algorithm: E-step

- We have shown that: [variational lower bound]

$$\log p(\mathbf{X} \mid \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})} + \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} \mid \mathbf{X}, \theta)}$$
$$= \mathcal{L}(q, \theta) + KL(q(\mathbf{Z}) \,\|\, p(\mathbf{Z}|\mathbf{X}, \theta))$$
$$\ge \mathcal{L}(q, \theta)$$ Evidence Lower bound (ELBO) or variational lower bound

with equality holding if and only if $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$

- **(E-step)** For a fixed θ, which q maximizes $\mathcal{L}(q, \theta)$?

⇒ $p(\mathbf{Z}|\mathbf{X}, \theta)$, because all other q would make $\mathcal{L}(q, \theta)$ strictly less than $\log p(\mathbf{X} \mid \theta)$

The EM Algorithm: M-step

- We also note that for a fixed q, the $\mathcal{L}(q, \theta)$ term can be decomposed into two terms:

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})}$$
$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z} \mid \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})$$

  - (1) A weighted sum of log *p*(**X**, **Z**|θ).
    This is tractable and can be optimized w.r.t θ
  - (2) Entropy of *q*(**Z**) which is independent of θ since *q* is fixed.

- **(M-step)** Thus, when *q* is fixed, we can find θ that maximizes $\mathcal{L}(q, \theta)$.

The EM Algorithm: summary

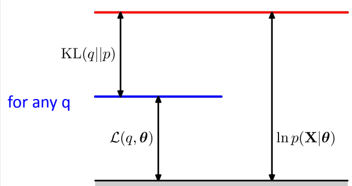
- Initialize parameters θ randomly
- Repeat until convergence:
  (optimize $\mathcal{L}(q, \theta)$ w.r.t. *q* and θ alternatingly.)
  - "E-step": Set $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \theta)$ compute posterior → optimal q(Z)!
  - "M-step": Update θ via the following maximization

$$\text{argmax}_\theta \; \mathcal{L}(q, \theta) = \text{argmax}_\theta \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\theta)$$

use *q*(Z) as (factional) pseudo-counts and maximize the "data completion" log-likelihood

- Note we have assumed that $p(\mathbf{Z} \mid \mathbf{X}, \theta)$ is tractable (i.e., find exact posterior $p(\mathbf{Z} \mid \mathbf{X}, \theta)$). Q. What if it is not?

Visualize the Decomposition



$$\log p(\mathbf{X} \mid \theta) = \mathcal{L}(q, \theta) + KL(q\|p)$$

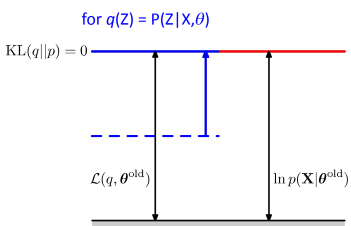
$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})}$$

- Note: $KL(q\|p) \ge 0$
  - with equality only when *q*=*p*.
- Thus, $\mathcal{L}(q, \theta)$ is a lower bound on $\log p(\mathbf{X} \mid \theta)$

  which EM tries to maximize.

Visualize the E-Step

- E-step: for fixed θ, find *q* that maximizes $\mathcal{L}(q, \theta)$



$$\log p(\mathbf{X} \mid \theta) = \mathcal{L}(q, \theta) + KL(q\|p)$$

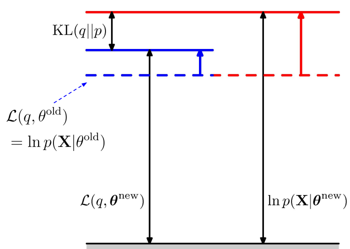
$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})}$$

- E-Step changes $q(\mathbf{Z})$ to maximize $\mathcal{L}(q, \theta)$
- So maximized when

$$KL(q\|p) = 0$$
$$q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \theta)$$

Visualize the M-Step

- M-step: for fixed *q*, find θ that maximizes $\mathcal{L}(q, \theta)$



$$\log p(\mathbf{X} \mid \theta) = \mathcal{L}(q, \theta) + KL(q\|p)$$

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})}$$

- Holding $q(\mathbf{Z})$ constant; increase $\mathcal{L}(q, \theta)$
- Updating θ will make $\log p(\mathbf{X} \mid \theta)$ increase!

$$\ln p(\mathbf{X}|\theta^{\text{new}}) \ge \ln p(\mathbf{X}|\theta^{\text{old}})$$

- But now $p \ne q$
- so $KL(q\|p) > 0$

The EM Algorithm: Multiple data-points

- Variational lower bound for a single example $\mathbf{x}$:

$$\log p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}|\theta)}{q(\mathbf{z})} + KL(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}, \theta))$$
$$\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}|\theta)}{q(\mathbf{z})}$$

- Lower bound on the log-likelihood of the *entire* training data $\mathcal{D} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$:

$$\log p(\mathcal{D}|\theta) = \sum_n \log p(\mathbf{x}^{(n)}|\theta) = \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})} + \sum_n KL(q^{(n)}(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}^{(n)}, \theta))$$
$$\geq \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})}$$

Note that different $q^{(n)}$ is used for each n

The EM Algorithm: Multiple data-points

$$\log p(\mathcal{D}|\theta) = \sum_n \log p(\mathbf{x}^{(n)}|\theta) = \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})} + \sum_n KL(q^{(n)}(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}^{(n)}, \theta))$$
$$\geq \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x}^{(n)}|\theta)}{q^{(n)}(\mathbf{z})}$$

- Initialize random parameters θ
- Repeat until convergence:
  - "**E-step**": Set $q^{(n)}(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}^{(n)}, \theta)$, for each training sample n.
  - "**M-step**": Update θ via the following maximization:

$$\arg\max_\theta \; \sum_n \sum_{\mathbf{z}} q^{(n)}(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}^{(n)} \mid \theta)$$