Expectation Maximization (Contd.)

Piyush Rai

Probabilistic Machine Learning (CS772A)

September 5, 2017

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^{K}$ randomly, or using K-means

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^{K}$ randomly, or using K-means
- For $t = 1, 2, ..$ or until convergence (e.g., when $\log p(\boldsymbol{x}|\Theta)$ ceases to increase)

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^K$ randomly, or using K-means
- For $t = 1, 2, ..$ or until convergence (e.g., when $\log p(\boldsymbol{x}|\Theta)$ ceases to increase)
  - **E Step:** Given $\Theta^{(t-1)}$, compute expectation $\mathbb{E}[z_{nk}]$ (same as posterior probability of $z_{nk} = 1$), $\forall n, k$

$$\gamma_{nk}^{(t)} = \mathbb{E}[z_{nk}^{(t)}] \propto \pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)})$$

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^K$ randomly, or using K-means
- For $t = 1, 2, ..$ or until convergence (e.g., when $\log p(\boldsymbol{x}|\Theta)$ ceases to increase)
  - **E Step:** Given $\Theta^{(t-1)}$, compute expectation $\mathbb{E}[z_{nk}]$ (same as posterior probability of $z_{nk} = 1$), $\forall n, k$

$$\gamma_{nk}^{(t)} = \mathbb{E}[z_{nk}^{(t)}] \propto \pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)}) \qquad \text{(basically depends on posterior } p(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(t-1)}))$$

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^K$ randomly, or using K-means
- For $t = 1, 2, ..$ or until convergence (e.g., when $\log p(\boldsymbol{x}|\Theta)$ ceases to increase)
  - **E Step:** Given $\Theta^{(t-1)}$, compute expectation $\mathbb{E}[z_{nk}]$ (same as posterior probability of $z_{nk} = 1$), $\forall n, k$

    $$\gamma_{nk}^{(t)} = \mathbb{E}[z_{nk}^{(t)}] \propto \pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)}) \qquad \text{(basically depends on posterior } p(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(t-1)})\text{)}$$

  - **M Step:** Given $\gamma_{nk}^{(t)}$, maximize expected CLL w.r.t. Θ. Results in the following updates

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^K$ randomly, or using K-means
- For $t = 1, 2, ..$ or until convergence (e.g., when $\log p(\boldsymbol{x}|\Theta)$ ceases to increase)
  - **E Step:** Given $\Theta^{(t-1)}$, compute expectation $\mathbb{E}[z_{nk}]$ (same as posterior probability of $z_{nk} = 1$), $\forall n, k$

$$\gamma_{nk}^{(t)} = \mathbb{E}[z_{nk}^{(t)}] \propto \pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)}) \qquad \text{(basically depends on posterior } p(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(t-1)}))$$

  - **M Step:** Given $\gamma_{nk}^{(t)}$, maximize expected CLL w.r.t. Θ. Results in the following updates

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

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^K$ randomly, or using K-means
- For $t = 1, 2, ..$ or until convergence (e.g., when $\log p(\boldsymbol{x}|\Theta)$ ceases to increase)
  - **E Step:** Given $\Theta^{(t-1)}$, compute expectation $\mathbb{E}[z_{nk}]$ (same as posterior probability of $z_{nk} = 1$), $\forall n, k$

$$\gamma_{nk}^{(t)} = \mathbb{E}[z_{nk}^{(t)}] \propto \pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)}) \qquad \text{(basically depends on posterior } p(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(t-1)}))$$

  - **M Step:** Given $\gamma_{nk}^{(t)}$, maximize expected CLL w.r.t. Θ. Results in the following updates

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

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

Recap: EM for Learning GMM

- Initialize the parameters $\Theta^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{k=1}^K$ randomly, or using K-means
- For $t = 1, 2, ..$ or until convergence (e.g., when $\log p(\boldsymbol{x}|\Theta)$ ceases to increase)
  - **E Step:** Given $\Theta^{(t-1)}$, compute expectation $\mathbb{E}[z_{nk}]$ (same as posterior probability of $z_{nk} = 1$), $\forall n, k$

$$\gamma_{nk}^{(t)} = \mathbb{E}[z_{nk}^{(t)}] \propto \pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)}) \qquad \text{(basically depends on posterior } p(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(t-1)}))$$

  - **M Step:** Given $\gamma_{nk}^{(t)}$, maximize expected CLL w.r.t. Θ. Results in the following updates

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

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

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

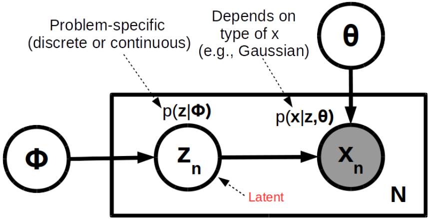
The General EM Algorithm

Parameter Estimation in Latent Variable Models

- Consider a latent variable model with joint distribution

$$p(\mathbf{X}, \mathbf{Z}|\Theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)p(\boldsymbol{z}_n|\phi)$$

- All model parameters collectively denoted by $\Theta = (\theta, \phi)$, $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$

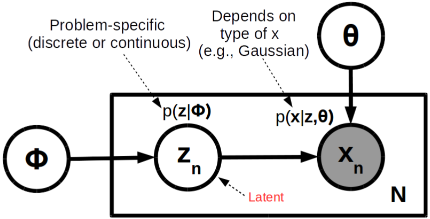


Parameter Estimation in Latent Variable Models

- Consider a latent variable model with joint distribution

$$p(\mathbf{X}, \mathbf{Z}|\Theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\Theta) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)p(\boldsymbol{z}_n|\phi)$$

- All model parameters collectively denoted by $\Theta = (\theta, \phi)$, $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$



- Goal: Estimate the model parameters Θ via MLE/MAP (sometimes also the latent variales $\boldsymbol{z}_n$'s)

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta)$$

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta)$$

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) d\mathbf{Z}$$

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)$ and $p(\boldsymbol{z}_n|\phi)$ are exponential family distributions

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)$ and $p(\boldsymbol{z}_n|\phi)$ are exponential family distributions
- Question: Is MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ a sensible thing to do?

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)$ and $p(\boldsymbol{z}_n|\phi)$ are exponential family distributions
- Question: Is MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ a sensible thing to do? Yes :-) Will see justification shortly

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X},\mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n,\theta)$ and $p(\boldsymbol{z}_n|\phi)$ are exponential family distributions
- Question: Is MLE on $\log p(\mathbf{X},\mathbf{Z}|\Theta)$ a sensible thing to do? Yes :-) Will see justification shortly
- Note: Since $\mathbf{Z}$ is latent, we will actually do MLE on the expectation of $\log p(\mathbf{X},\mathbf{Z}|\Theta)$

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)$ and $p(\boldsymbol{z}_n|\phi)$ are exponential family distributions
- Question: Is MLE on $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$ a sensible thing to do? Yes :-) Will see justification shortly
- Note: Since $\mathbf{Z}$ is latent, we will actually do MLE on the expectation of $\log p(\mathbf{X}, \mathbf{Z}|\Theta)$

$$\boxed{\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)] = \mathbb{E}[\log p(\mathbf{X}|\mathbf{Z}, \theta) + \log p(\mathbf{Z}|\phi)]} \quad \text{(called exp. complete data log-lik.)}$$

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X},\mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n,\theta)$ and $p(\boldsymbol{z}_n|\phi)$ are exponential family distributions
- Question: Is MLE on $\log p(\mathbf{X},\mathbf{Z}|\Theta)$ a sensible thing to do? Yes :-) Will see justification shortly
- Note: Since $\mathbf{Z}$ is latent, we will actually do MLE on the expectation of $\log p(\mathbf{X},\mathbf{Z}|\Theta)$

$$\boxed{\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)] = \mathbb{E}[\log p(\mathbf{X}|\mathbf{Z},\theta) + \log p(\mathbf{Z}|\phi)]} \quad \text{(called exp. complete data log-lik.)}$$

where expectation is taken w.r.t. an "optimal" distribution of $\mathbf{Z}$ (will see shortly what this distr. is)

Parameter Estimation in Latent Variable Models

- Goal: Estimate the model parameters Θ via MLE/MAP

$$\hat{\Theta} = \arg\max_{\Theta} \log p(\mathbf{X}|\Theta) = \arg\max_{\Theta} \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) \quad \text{or} \quad \arg\max_{\Theta} \log \int_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) d\mathbf{Z}$$

- Getting closed-form MLE/MAP in such models can be difficult because of the log-sum/integral
- **An idea:** Let's do MLE on $\log p(\mathbf{X},\mathbf{Z}|\Theta)$ instead since it usually has a much simpler expression
  - .. simpler especially when $p(\boldsymbol{x}_n|\boldsymbol{z}_n,\theta)$ and $p(\boldsymbol{z}_n|\phi)$ are exponential family distributions
- Question: Is MLE on $\log p(\mathbf{X},\mathbf{Z}|\Theta)$ a sensible thing to do? Yes :-) Will see justification shortly
- Note: Since $\mathbf{Z}$ is latent, we will actually do MLE on the expectation of $\log p(\mathbf{X},\mathbf{Z}|\Theta)$

$$\boxed{\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)] = \mathbb{E}[\log p(\mathbf{X}|\mathbf{Z},\theta) + \log p(\mathbf{Z}|\phi)]} \quad \text{(called exp. complete data log-lik.)}$$

where expectation is taken w.r.t. an "optimal" distribution of $\mathbf{Z}$ (will see shortly what this distr. is)

- This procedure is basically the Expectation Maximization (EM) algorithm for latent variable models

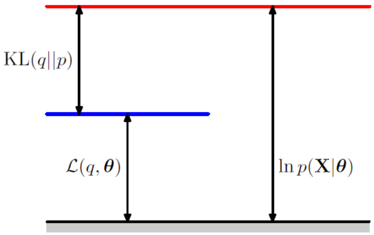
EM in a Nutshell

- Define $p_z = p(\mathbf{Z}|\mathbf{X}, \Theta)$. The identity below holds for any choice of the distribution q

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

$$\begin{aligned}
\mathcal{L}(q, \Theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \right\} \\
\mathrm{KL}(q||p_z) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \Theta)}{q(\mathbf{Z})} \right\}
\end{aligned}$$

(Exercise: Verify the above identity)



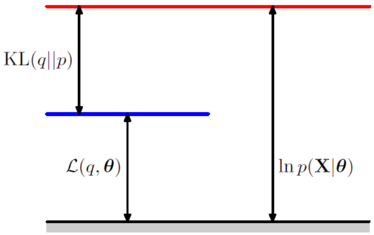
EM in a Nutshell

- Define $p_z = p(\mathbf{Z}|\mathbf{X}, \Theta)$. The identity below holds for any choice of the distribution q

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

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

(Exercise: Verify the above identity)



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

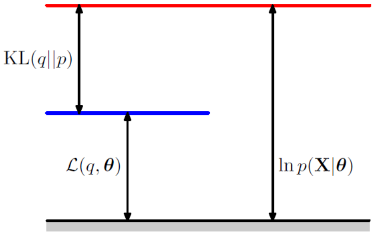
EM in a Nutshell

- Define $p_z = p(\mathbf{Z}|\mathbf{X},\Theta)$. The identity below holds for any choice of the distribution q

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

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

(Exercise: Verify the above identity)



- Since $\text{KL}(q||p_z) \geq 0$, $\mathcal{L}(q,\Theta)$ is a lower-bound on $\log p(\mathbf{X}|\Theta)$
  - **Consequence:** Maximizing $\mathcal{L}(q,\Theta)$ will also make $\log p(\mathbf{X}|\Theta)$ go up

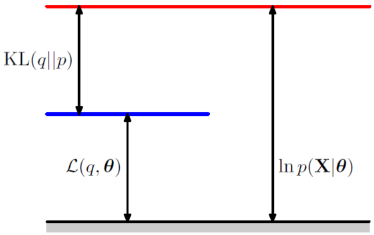
EM in a Nutshell

- Define $p_z = p(\mathbf{Z}|\mathbf{X},\Theta)$. The identity below holds for any choice of the distribution q

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

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

(Exercise: Verify the above identity)



- Since $\text{KL}(q||p_z) \geq 0$, $\mathcal{L}(q,\Theta)$ is a lower-bound on $\log p(\mathbf{X}|\Theta)$
  - **Consequence:** Maximizing $\mathcal{L}(q,\Theta)$ will also make $\log p(\mathbf{X}|\Theta)$ go up
- Therefore, to do MLE on $\log p(\mathbf{X}|\Theta)$, we can maximize $\mathcal{L}(q,\Theta)$ w.r.t. q and Θ

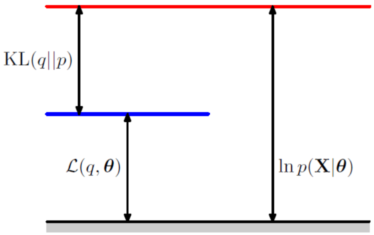
EM in a Nutshell

- Define $p_z = p(\mathbf{Z}|\mathbf{X},\Theta)$. The identity below holds for any choice of the distribution q

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

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

(Exercise: Verify the above identity)



- Since $\text{KL}(q||p_z) \geq 0$, $\mathcal{L}(q,\Theta)$ is a lower-bound on $\log p(\mathbf{X}|\Theta)$
  - **Consequence:** Maximizing $\mathcal{L}(q,\Theta)$ will also make $\log p(\mathbf{X}|\Theta)$ go up
- Therefore, to do MLE on $\log p(\mathbf{X}|\Theta)$, we can maximize $\mathcal{L}(q,\Theta)$ w.r.t. q and Θ
  - Maximizing $\mathcal{L}(q,\Theta)$ jointly w.r.t. q and Θ isn't possible, so we alternate between:

EM in a Nutshell

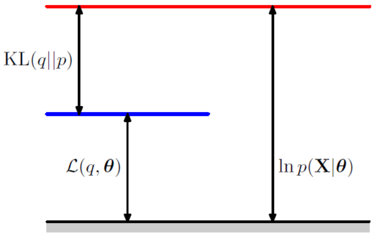
- Define $p_z = p(\mathbf{Z}|\mathbf{X},\Theta)$. The identity below holds for any choice of the distribution q

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

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

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

(Exercise: Verify the above identity)



- Since $\text{KL}(q||p_z) \geq 0$, $\mathcal{L}(q,\Theta)$ is a lower-bound on $\log p(\mathbf{X}|\Theta)$
  - **Consequence:** Maximizing $\mathcal{L}(q,\Theta)$ will also make $\log p(\mathbf{X}|\Theta)$ go up
- Therefore, to do MLE on $\log p(\mathbf{X}|\Theta)$, we can maximize $\mathcal{L}(q,\Theta)$ w.r.t. q and Θ
  - Maximizing $\mathcal{L}(q,\Theta)$ jointly w.r.t. q and Θ isn't possible, so we alternate between: (1) Given Θ, find the optimal q that maximizes $\mathcal{L}(q,\Theta)$

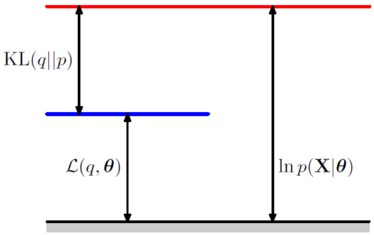
EM in a Nutshell

- Define $p_z = p(\mathbf{Z}|\mathbf{X}, \Theta)$. The identity below holds for any choice of the distribution q

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

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

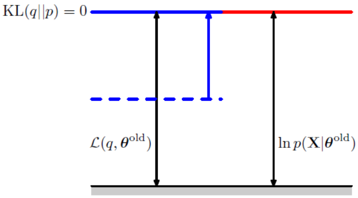
(Exercise: Verify the above identity)



- Since $\text{KL}(q||p_z) \geq 0$, $\mathcal{L}(q, \Theta)$ is a lower-bound on $\log p(\mathbf{X}|\Theta)$
  - **Consequence:** Maximizing $\mathcal{L}(q, \Theta)$ will also make $\log p(\mathbf{X}|\Theta)$ go up
- Therefore, to do MLE on $\log p(\mathbf{X}|\Theta)$, we can maximize $\mathcal{L}(q, \Theta)$ w.r.t. q and Θ
  - Maximizing $\mathcal{L}(q, \Theta)$ jointly w.r.t. q and Θ isn't possible, so we alternate between: (1) Given Θ, find the optimal q that maximizes $\mathcal{L}(q, \Theta)$, and (2) Given q, find the optimal Θ that maximizes $\mathcal{L}(q, \Theta)$

Given Θ, which q maximizes $\mathcal{L}(q, \Theta)$?

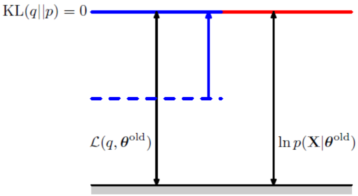
$$\boxed{\log p(\mathbf{X}|\Theta) = \mathcal{L}(q, \Theta) + \mathrm{KL}(q||p_z)}$$



- Since Θ (and thus $\log p(\mathbf{X}|\Theta)$) doesn't change in this step, the sum $\mathcal{L}(q, \Theta) + \mathrm{KL}(q||p_z)$ is fixed

Given Θ, which q maximizes $\mathcal{L}(q, \Theta)$?

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

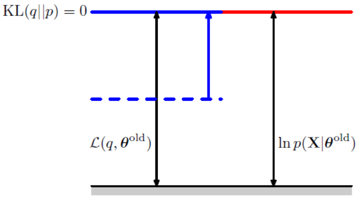


- Since Θ (and thus $\log p(\mathbf{X}|\Theta)$) doesn't change in this step, the sum $\mathcal{L}(q, \Theta) + \mathrm{KL}(q||p_z)$ is fixed
- Therefore, with Θ fixed to Θ^{old}, maximizing the lower bound $\mathcal{L}(q, \Theta^{old})$ w.r.t. q, i.e.,

$$\boxed{\hat{q} = \arg\max_q \mathcal{L}(q, \Theta^{old})}$$

Given Θ, which q maximizes $\mathcal{L}(q, \Theta)$?

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



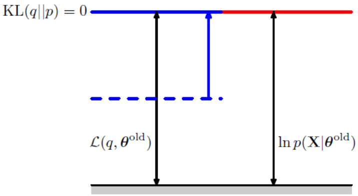
- Since Θ (and thus $\log p(\mathbf{X}|\Theta)$) doesn't change in this step, the sum $\mathcal{L}(q, \Theta) + \mathrm{KL}(q||p_z)$ is fixed
- Therefore, with Θ fixed to Θ^{old}, maximizing the lower bound $\mathcal{L}(q, \Theta^{old})$ w.r.t. q, i.e.,

$$\boxed{\hat{q} = \arg\max_q \mathcal{L}(q, \Theta^{old})}$$

.. is equivalent to finding q for which $\mathrm{KL}(q||p_z) = 0$

Given Θ, which q maximizes $\mathcal{L}(q, \Theta)$?

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



- Since Θ (and thus $\log p(\mathbf{X}|\Theta)$) doesn't change in this step, the sum $\mathcal{L}(q, \Theta) + \text{KL}(q||p_z)$ is fixed
- Therefore, with Θ fixed to Θ^{old}, maximizing the lower bound $\mathcal{L}(q, \Theta^{old})$ w.r.t. q, i.e.,

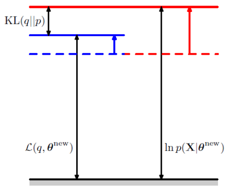
$$\boxed{\hat{q} = \arg\max_q \mathcal{L}(q, \Theta^{old})}$$

.. is equivalent to finding q for which $\text{KL}(q||p_z) = 0$, i.e., $\hat{q} = p_z = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$

Given q, which Θ maximizes $\mathcal{L}(q,\Theta)$?

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

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

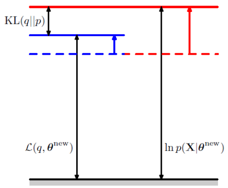


- With q fixed at $\hat{q} = p(\mathbf{Z}|\mathbf{X},\Theta^{old})$, we want to maximize $\mathcal{L}(\hat{q},\Theta)$ w.r.t. Θ, where

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

Given q, which Θ maximizes $\mathcal{L}(q,\Theta)$?

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



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

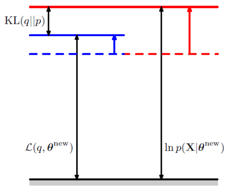
- With q fixed at $\hat{q} = p(\mathbf{Z}|\mathbf{X},\Theta^{old})$, we want to maximize $\mathcal{L}(\hat{q},\Theta)$ w.r.t. Θ, where

$$\begin{aligned} \mathcal{L}(\hat{q},\Theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^{old}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^{old}) \log p(\mathbf{Z}|\mathbf{X},\Theta^{old}) \\ &= \mathcal{Q}(\Theta,\Theta^{old}) + \text{const} \end{aligned}$$

.. where $\mathcal{Q}(\Theta,\Theta^{old}) = \mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)]$ is the exp. complete data log-lik

Given q, which Θ maximizes $\mathcal{L}(q, \Theta)$?

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



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

- With q fixed at $\hat{q} = p(\mathbf{Z}|\mathbf{X},\Theta^{old})$, we want to maximize $\mathcal{L}(\hat{q},\Theta)$ w.r.t. Θ, where

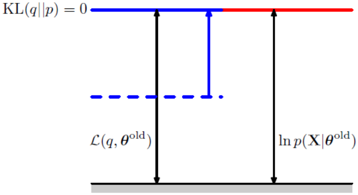
$$\begin{aligned} \mathcal{L}(\hat{q},\Theta) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^{old}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta^{old}) \log p(\mathbf{Z}|\mathbf{X},\Theta^{old}) \\ &= \mathcal{Q}(\Theta,\Theta^{old}) + \text{const} \end{aligned}$$

.. where $\mathcal{Q}(\Theta,\Theta^{old}) = \mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)]$ is the exp. complete data log-lik

- Therefore the optimal Θ is $\boxed{\Theta^{new} = \arg\max_{\Theta} \mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)]}$

EM Iterations never decrease $\log p(\mathbf{X}|\Theta)$

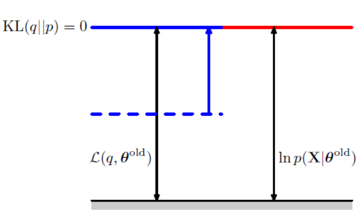
- Step 1: Setting $q = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ makes KL zero, and $\mathcal{L}(q, \Theta^{old})$ becomes equal to $\log p(\mathbf{X}|\Theta^{old})$



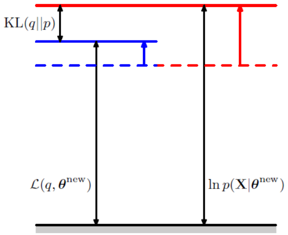
(Step 1)

EM Iterations never decrease $\log p(\mathbf{X}|\Theta)$

- Step 1: Setting $q = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ makes KL zero, and $\mathcal{L}(q, \Theta^{old})$ becomes equal to $\log p(\mathbf{X}|\Theta^{old})$



(Step 1)

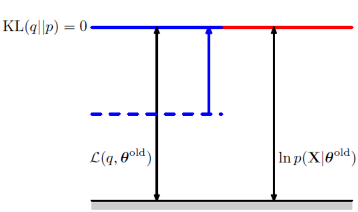


(Step 2)

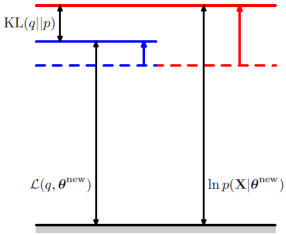
- Step 2: Maximizing $\mathcal{L}(q, \Theta) = \mathcal{Q}(\Theta, \Theta^{old})$ w.r.t. Θ will lead to

EM Iterations never decrease $\log p(\mathbf{X}|\Theta)$

- Step 1: Setting $q = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ makes KL zero, and $\mathcal{L}(q, \Theta^{old})$ becomes equal to $\log p(\mathbf{X}|\Theta^{old})$



(Step 1)

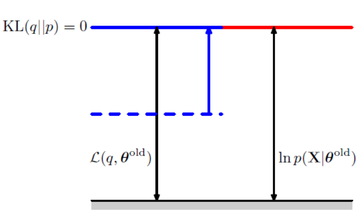


(Step 2)

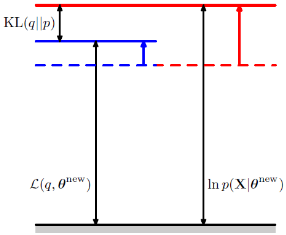
- Step 2: Maximizing $\mathcal{L}(q, \Theta) = \mathcal{Q}(\Theta, \Theta^{old})$ w.r.t. Θ will lead to
  - $\mathcal{L}(q, \Theta)$ increasing, if not already at the optima

EM Iterations never decrease $\log p(\mathbf{X}|\Theta)$

- Step 1: Setting $q = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ makes KL zero, and $\mathcal{L}(q, \Theta^{old})$ becomes equal to $\log p(\mathbf{X}|\Theta^{old})$



(Step 1)

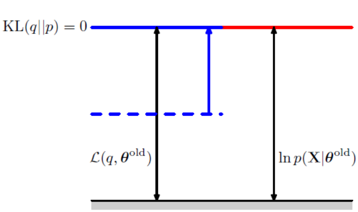


(Step 2)

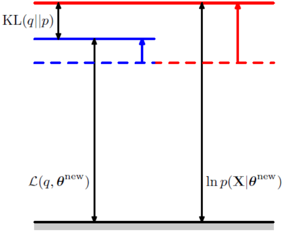
- Step 2: Maximizing $\mathcal{L}(q, \Theta) = \mathcal{Q}(\Theta, \Theta^{old})$ w.r.t. Θ will lead to
  - $\mathcal{L}(q, \Theta)$ increasing, if not already at the optima
  - $\mathrm{KL}(q||p_z) > 0$ again because $q \neq p(\mathbf{Z}|\mathbf{X}, \Theta^{new})$

EM Iterations never decrease $\log p(\mathbf{X}|\Theta)$

- Step 1: Setting $q = p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ makes KL zero, and $\mathcal{L}(q, \Theta^{old})$ becomes equal to $\log p(\mathbf{X}|\Theta^{old})$



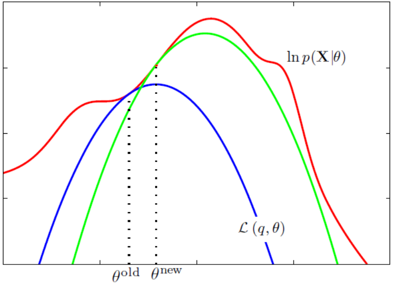
(Step 1)



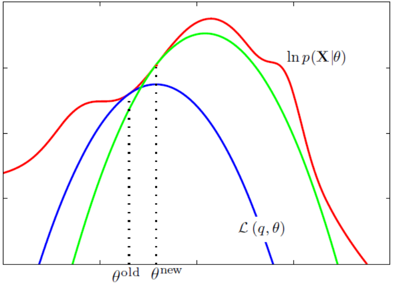
(Step 2)

- Step 2: Maximizing $\mathcal{L}(q, \Theta) = \mathcal{Q}(\Theta, \Theta^{old})$ w.r.t. Θ will lead to
  - $\mathcal{L}(q, \Theta)$ increasing, if not already at the optima
  - $\mathrm{KL}(q||p_z) > 0$ again because $q \neq p(\mathbf{Z}|\mathbf{X}, \Theta^{new})$
  - As a result, ensures that $\log p(\mathbf{X}|\Theta^{new}) = \mathcal{L}(q, \Theta^{new}) + \mathrm{KL}(q||p_z) \geq \log p(\mathbf{X}|\Theta^{old})$

EM: A View in the Parameter (Θ) Space

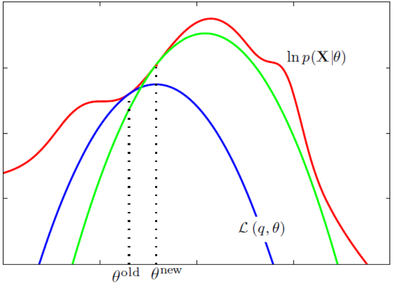


EM: A View in the Parameter (Θ) Space



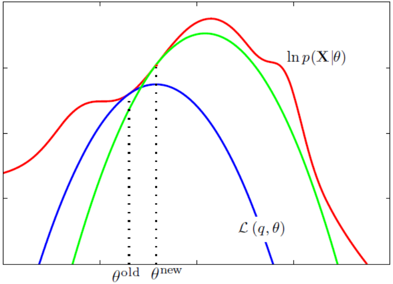
- E-step: Update of q makes the $\mathcal{L}(q, \Theta)$ curve touch the $\log p(\mathbf{X}|\Theta)$ curve at Θ^{old}

EM: A View in the Parameter (Θ) Space



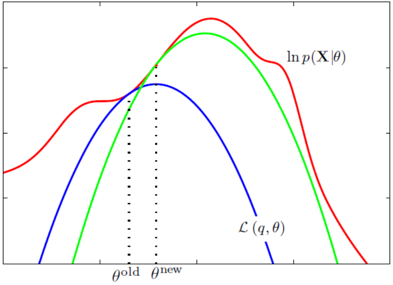
- E-step: Update of q makes the $\mathcal{L}(q, \Theta)$ curve touch the $\log p(\mathbf{X}|\Theta)$ curve at Θ^{old}
- M-step gives the maxima Θ^{new} of $\mathcal{L}(q, \Theta^{old})$

EM: A View in the Parameter (Θ) Space



- E-step: Update of q makes the $\mathcal{L}(q,\Theta)$ curve touch the $\log p(\mathbf{X}|\Theta)$ curve at Θ^{old}
- M-step gives the maxima Θ^{new} of $\mathcal{L}(q,\Theta^{old})$
- Next E-step readjusts $\mathcal{L}(q,\Theta^{old})$ curve (green) to meet $\log p(\mathbf{X}|\Theta)$ curve again, now at Θ^{new}

EM: A View in the Parameter (Θ) Space



- E-step: Update of q makes the $\mathcal{L}(q,\Theta)$ curve touch the $\log p(\mathbf{X}|\Theta)$ curve at Θ^{old}
- M-step gives the maxima Θ^{new} of $\mathcal{L}(q,\Theta^{old})$
- Next E-step readjusts $\mathcal{L}(q,\Theta^{old})$ curve (green) to meet $\log p(\mathbf{X}|\Theta)$ curve again, now at Θ^{new}
- This continues until a local maxima of $\log p(\mathbf{X}|\Theta)$ is reached

An Alternate View of What EM Does

- Consider the ‘incomplete” data log likelihood

$$\log p(\mathbf{X}|\Theta) \quad = \quad \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta)$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) \quad = \quad \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})}$$

An Alternate View of What EM Does

- Consider the ‘incomplete” data log likelihood

$$\log p(\mathbf{X}|\Theta) \quad = \quad \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)}$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})}$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1)$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta}
\end{aligned}
$$

An Alternate View of What EM Does

- Consider the ‘incomplete” data log likelihood

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality.

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1)$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1)$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}$$

- If we <u>set</u> $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})}$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\log p(\mathbf{X}|\Theta) = \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)}$$

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1)$$

$$\log p(\mathbf{X}|\Theta) \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}}$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta)
$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$
\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta)
\end{aligned}
$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we <u>set</u> $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$
\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) = \log p(\mathbf{X}|\Theta)
\end{aligned}
$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i), \text{ if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) = \log p(\mathbf{X}|\Theta)
\end{aligned}$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, we have

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we <u>set</u> $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$
\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) = \log p(\mathbf{X}|\Theta)
\end{aligned}
$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, we have

$$
\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$
\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}
$$

- If we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$
\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) = \log p(\mathbf{X}|\Theta)
\end{aligned}
$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, we have

$$
\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.} = \mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)] + \text{const.}
$$

An Alternate View of What EM Does

- Consider the 'incomplete" data log likelihood

$$\begin{aligned}
\log p(\mathbf{X}|\Theta) &= \log \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\Theta) = \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(where } q(\mathbf{Z}) \text{ can be any distribution)} \\
&\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} \quad \text{(concave } f\text{, Jensen's Ineq.: } f(\textstyle\sum \lambda_i x_i) \geq \sum \lambda_i f(x_i)\text{, if } \sum_i \lambda_i = 1) \\
\log p(\mathbf{X}|\Theta) &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) - \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log q(\mathbf{Z})}_{\text{doesn't depend on } \Theta} = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.}
\end{aligned}$$

- If we <u>set</u> $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, the above inequality becomes equality. To see this, note that

$$\begin{aligned}
\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z}|\Theta)}{q(\mathbf{Z})} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log \frac{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}\, p(\mathbf{X}|\Theta)}{\cancel{p(\mathbf{Z}|\mathbf{X},\Theta)}} = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X}|\Theta) \\
&= \log p(\mathbf{X}|\Theta) \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) = \log p(\mathbf{X}|\Theta)
\end{aligned}$$

- Thus for $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\Theta)$, we have

$$\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\Theta) \log p(\mathbf{X},\mathbf{Z}|\Theta) + \text{const.} = \mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)] + \text{const.}$$

- Thus ILL $\log p(\mathbf{X}|\Theta)$ is tightly lower-bounded by expected CLL $\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\Theta)]$ which EM maximizes

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**

- **M (Maximization) step:**

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
- **M (Maximization) step:**

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior distribution

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X}, \Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

- **M (Maximization) step:**

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior distribution

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

  - Note: Expected CLL $\mathcal{Q}(\Theta, \Theta^{old})$ usually has a simple expression when $p(\boldsymbol{x}|\boldsymbol{z}, \theta)$ and $p(\boldsymbol{z}|\phi)$ are exp. family distributions - can simply write down the log joint probability expression $\log p(\boldsymbol{x}, \boldsymbol{z}|\Theta)$ and simply replace any occurrence of a term with $\boldsymbol{z}$ by its expectation (we already saw it in GMM)
- **M (Maximization) step:**

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior distribution $p(\mathbf{Z}|\mathbf{X},\Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior distribution

$$\mathcal{Q}(\Theta,\Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X},\mathbf{Z}|\Theta)]$$

  - Note: Expected CLL $\mathcal{Q}(\Theta,\Theta^{old})$ usually has a simple expression when $p(\boldsymbol{x}|\boldsymbol{z},\theta)$ and $p(\boldsymbol{z}|\phi)$ are exp. famility distributions - can simply write down the log joint probability expression $\log p(\boldsymbol{x},\boldsymbol{z}|\Theta)$ and simply replace any occurrence of a term with $\boldsymbol{z}$ by its expectation (we already saw it in GMM)

- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. Θ

$$\Theta^{new} = \arg\max_{\Theta} \mathcal{Q}(\Theta,\Theta^{old}) \qquad \text{(if doing MLE)}$$

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior distribution

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

  - Note: Expected CLL $\mathcal{Q}(\Theta, \Theta^{old})$ usually has a simple expression when $p(\boldsymbol{x}|\boldsymbol{z}, \theta)$ and $p(\boldsymbol{z}|\phi)$ are exp. famility distributions - can simply write down the log joint probability expression $\log p(\boldsymbol{x}, \boldsymbol{z}|\Theta)$ and simply replace any occurrence of a term with $\boldsymbol{z}$ by its expectation (we already saw it in GMM)
- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. Θ

$$\Theta^{new} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{old}) \qquad \text{(if doing MLE)}$$

$$\Theta^{new} = \arg\max_{\Theta} \{\mathcal{Q}(\Theta, \Theta^{old}) + \log p(\Theta)\} \qquad \text{(if doing MAP)}$$

The Expectation Maximization (EM) Algorithm

Initialize the parameters: Θ^{old}. Then alternate between these steps:

- **E (Expectation) step:**
  - Compute the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \Theta^{old})$ over latent variables $\mathbf{Z}$ using Θ^{old}
  - Compute the expected complete data log-likelihood w.r.t. *this* posterior distribution

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}_{p(\mathbf{Z}|\mathbf{X},\Theta^{old})}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$$

  - Note: Expected CLL $\mathcal{Q}(\Theta, \Theta^{old})$ usually has a simple expression when $p(\boldsymbol{x}|\boldsymbol{z}, \theta)$ and $p(\boldsymbol{z}|\phi)$ are exp. famility distributions - can simply write down the log joint probability expression $\log p(\boldsymbol{x}, \boldsymbol{z}|\Theta)$ and simply replace any occurrence of a term with $\boldsymbol{z}$ by its expectation (we already saw it in GMM)
- **M (Maximization) step:**
  - Maximize the expected complete data log-likelihood w.r.t. Θ

$$\Theta^{new} = \arg\max_{\Theta} \mathcal{Q}(\Theta, \Theta^{old}) \qquad \text{(if doing MLE)}$$

$$\Theta^{new} = \arg\max_{\Theta} \{\mathcal{Q}(\Theta, \Theta^{old}) + \log p(\Theta)\} \qquad \text{(if doing MAP)}$$

- If the incomplete log-lik $p(\mathbf{X}|\Theta)$ not yet converged then set $\Theta^{old} = \Theta^{new}$ and go to the E step.

EM: An Example of what “Expected CLL” looks like

- Already seen GMM. Let’s consider a latent factor model for dimensionality reduction (next class)

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

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

- A linear Gaussian model: Low-dim $\boldsymbol{z}_n \in \mathbb{R}^K$ mapped to high-dim $\boldsymbol{x}_n \in \mathbb{R}^D$ via $\mathbf{W} \in \mathbb{W}^{D\times K}$

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

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

$$\log p(\mathbf{X}, \mathbf{Z}|\mathbf{W}, \sigma^2)$$

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

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

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

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

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

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

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

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

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

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

- A linear Gaussian model: Low-dim $\boldsymbol{z}_n \in \mathbb{R}^K$ mapped to high-dim $\boldsymbol{x}_n \in \mathbb{R}^D$ via $\mathbf{W} \in \mathbb{W}^{D\times K}$
- 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

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

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

- A linear Gaussian model: Low-dim $\boldsymbol{z}_n \in \mathbb{R}^K$ mapped to high-dim $\boldsymbol{x}_n \in \mathbb{R}^D$ via $\mathbf{W} \in \mathbb{W}^{D\times K}$
- 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

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

- Expected CLL will require replacing $\boldsymbol{z}_n$ by $\mathbb{E}[\boldsymbol{z}_n]$ and $\boldsymbol{z}_n\boldsymbol{z}_n^\top$ by $\mathbb{E}[\boldsymbol{z}_n\boldsymbol{z}_n^\top]$

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

- A linear Gaussian model: Low-dim $\boldsymbol{z}_n \in \mathbb{R}^K$ mapped to high-dim $\boldsymbol{x}_n \in \mathbb{R}^D$ via $\mathbf{W} \in \mathbb{W}^{D\times K}$
- 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

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

- 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 easily obtained from the posterior $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ (computed in E step)

EM: An Example of what "Expected CLL" looks like

- Already seen GMM. Let's consider a latent factor model for dimensionality reduction (next class)

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

- A linear Gaussian model: Low-dim $\boldsymbol{z}_n \in \mathbb{R}^K$ mapped to high-dim $\boldsymbol{x}_n \in \mathbb{R}^D$ via $\mathbf{W} \in \mathbb{W}^{D\times K}$
- 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

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

- 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 easily obtained from the posterior $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ (computed in E step)
- The M step maximizes the expected CLL w.r.t. the parameters ($\mathbf{W}$ in this case)

Online or Incremental EM

- Needn't compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ for every $\boldsymbol{x}_n$ in each EM iteration (computational/storage efficiency)

Online or Incremental EM

- Needn't compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ for every $\boldsymbol{x}_n$ in each EM iteration (computational/storage efficiency)
  - Recall that the expected CLL is often a sum over all data points

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{n=1}^{N} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]$$

Online or Incremental EM

- Needn't compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ for every $\boldsymbol{x}_n$ in each EM iteration (computational/storage efficiency)
  - Recall that the expected CLL is often a sum over all data points

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{n=1}^{N} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]$$

  - Can compute this quantity recursively using small minibatches of data

$$\mathcal{Q}_t = (1-\gamma_t)\mathcal{Q}_{t-1} + \gamma_t \left[\sum_{n=1}^{N_t} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]\right]$$

  .. where $\gamma_t = (1+t)^{-\kappa}, 0.5 < \kappa \leq 1$ is a decaying learning rate

Online or Incremental EM

- Needn't compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ for every $\boldsymbol{x}_n$ in each EM iteration (computational/storage efficiency)
  - Recall that the expected CLL is often a sum over all data points

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{n=1}^{N} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]$$

  - Can compute this quantity recursively using small minibatches of data

$$\mathcal{Q}_t = (1-\gamma_t)\mathcal{Q}_{t-1} + \gamma_t \left[\sum_{n=1}^{N_t} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]\right]$$

  .. where $\gamma_t = (1+t)^{-\kappa}, 0.5 < \kappa \leq 1$ is a decaying learning rate
  - Requires computing $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ only for data in current mini-batch (computational/storage efficiency)

Online or Incremental EM

- Needn't compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ for every $\boldsymbol{x}_n$ in each EM iteration (computational/storage efficiency)
  - Recall that the expected CLL is often a sum over all data points

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{n=1}^{N} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]$$

  - Can compute this quantity recursively using small minibatches of data

$$\mathcal{Q}_t = (1-\gamma_t)\mathcal{Q}_{t-1} + \gamma_t \left[\sum_{n=1}^{N_t} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]\right]$$

    .. where $\gamma_t = (1+t)^{-\kappa}, 0.5 < \kappa \leq 1$ is a decaying learning rate
  - Requires computing $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ only for data in current mini-batch (computational/storage efficiency)
  - MLE on above $\mathcal{Q}_t$ akin to simple recursive updates for Θ. E.g., something like this for GMM

$$\mu_k^{(t)} = (1-\gamma_t)\mu_k^{(t-1)} + \gamma_t \frac{1}{N_t}\sum_{n=1}^{N_t} \gamma_{nk}\boldsymbol{x}_n \quad \text{(update for } k\text{-th Gaussian's mean)}$$

Online or Incremental EM

- Needn't compute $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ for every $\boldsymbol{x}_n$ in each EM iteration (computational/storage efficiency)
  - Recall that the expected CLL is often a sum over all data points

$$\mathcal{Q}(\Theta, \Theta^{old}) = \mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta) = \sum_{n=1}^{N} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]$$

  - Can compute this quantity recursively using small minibatches of data

$$\mathcal{Q}_t = (1-\gamma_t)\mathcal{Q}_{t-1} + \gamma_t \left[\sum_{n=1}^{N_t} \mathbb{E}[\log p(\boldsymbol{x}_n|\boldsymbol{z}_n, \theta)] + \mathbb{E}[\log p(\boldsymbol{z}_n|\phi)]\right]$$

  .. where $\gamma_t = (1+t)^{-\kappa}, 0.5 < \kappa \leq 1$ is a decaying learning rate

  - Requires computing $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ only for data in current mini-batch (computational/storage efficiency)
  - MLE on above $\mathcal{Q}_t$ akin to simple recursive updates for Θ. E.g., something like this for GMM

$$\mu_k^{(t)} = (1-\gamma_t)\mu_k^{(t-1)} + \gamma_t \frac{1}{N_t} \sum_{n=1}^{N_t} \gamma_{nk} \boldsymbol{x}_n \quad \text{(update for } k\text{-th Gaussian's mean)}$$

- Note: The above is only a sketchy description of the procedure. I will provide a reference.

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression

$$p(\boldsymbol{y}|\mathbf{X}, \lambda, \beta) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression
  $$p(\boldsymbol{y}|\mathbf{X}, \lambda, \beta) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$
  - As an alternative, can treat $\boldsymbol{w}$ as a latent variable and β, λ as parameters and use EM to learn these

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression

$$p(\boldsymbol{y}|\mathbf{X},\lambda,\beta) = \int p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$

  - As an alternative, can treat $\boldsymbol{w}$ as a latent variable and β,λ as parameters and use EM to learn these
  - E step computes posterior $p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda)$ assuming β,λ fixed from the previous M step

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression

$$p(\boldsymbol{y}|\mathbf{X},\lambda,\beta) = \int p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$

  - As an alternative, can treat $\boldsymbol{w}$ as a latent variable and β, λ as parameters and use EM to learn these
  - E step computes posterior $p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda)$ assuming β, λ fixed from the previous M step
  - M step maximizes $\mathbb{E}[\log p(\boldsymbol{y},\boldsymbol{w}|\mathbf{X},\beta,\lambda)] = \mathbb{E}[\log p(\boldsymbol{y}|\boldsymbol{w},\mathbf{X},\beta) + \log p(\boldsymbol{w}|\lambda)]$ w.r.t. λ, β

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression

$$p(\boldsymbol{y}|\mathbf{X}, \lambda, \beta) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$

  - As an alternative, can treat $\boldsymbol{w}$ as a latent variable and β, λ as parameters and use EM to learn these
  - E step computes posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ assuming β, λ fixed from the previous M step
  - M step maximizes $\mathbb{E}[\log p(\boldsymbol{y}, \boldsymbol{w}|\mathbf{X}, \beta, \lambda)] = \mathbb{E}[\log p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta) + \log p(\boldsymbol{w}|\lambda)]$ w.r.t. λ, β
    - This requires using expectations of quantities like $\boldsymbol{w}$ and $\boldsymbol{w}\boldsymbol{w}^\top$ which can be obtained easily from the posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ which we compute in the E step

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression

$$p(\boldsymbol{y}|\mathbf{X}, \lambda, \beta) = \int p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$

  - As an alternative, can treat $\boldsymbol{w}$ as a latent variable and β, λ as parameters and use EM to learn these
  - E step computes posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ assuming β, λ fixed from the previous M step
  - M step maximizes $\mathbb{E}[\log p(\boldsymbol{y}, \boldsymbol{w}|\mathbf{X}, \beta, \lambda)] = \mathbb{E}[\log p(\boldsymbol{y}|\boldsymbol{w}, \mathbf{X}, \beta) + \log p(\boldsymbol{w}|\lambda)]$ w.r.t. λ, β
    - This requires using expectations of quantities like $\boldsymbol{w}$ and $\boldsymbol{w}\boldsymbol{w}^\top$ which can be obtained easily from the posterior $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda)$ which we compute in the E step
- Exercise: Try some of these by yourself. The books (MLAPP/PRML) contain such examples

EM: Some Other Examples

- Mixture of (multivariate) Bernoulli distributions (if each $\boldsymbol{x}$ is a binary vector)
- Mixture of (multivariate) multinoulli distributions (if each $\boldsymbol{x}$ is a categorical vector)
- Hyperparameter estimation in probabilistic models (an alternative to MLE-II)
  - We've already seen MLE-II where we did MLE on marginal likelihood, e.g., for linear regression

$$p(\boldsymbol{y}|\mathbf{X},\lambda,\beta) = \int p(\boldsymbol{y}|\mathbf{X},\boldsymbol{w},\beta)p(\boldsymbol{w}|\lambda)d\boldsymbol{w}$$

  - As an alternative, can treat $\boldsymbol{w}$ as a latent variable and β, λ as parameters and use EM to learn these
  - E step computes posterior $p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda)$ assuming β, λ fixed from the previous M step
  - M step maximizes $\mathbb{E}[\log p(\boldsymbol{y},\boldsymbol{w}|\mathbf{X},\beta,\lambda)] = \mathbb{E}[\log p(\boldsymbol{y}|\boldsymbol{w},\mathbf{X},\beta) + \log p(\boldsymbol{w}|\lambda)]$ w.r.t. λ, β
    - This requires using expectations of quantities like $\boldsymbol{w}$ and $\boldsymbol{w}\boldsymbol{w}^\top$ which can be obtained easily from the posterior $p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y},\beta,\lambda)$ which we compute in the E step
- Exercise: Try some of these by yourself. The books (MLAPP/PRML) contain such examples
- Next Class: EM for latent factor models (dimensionality reduction) and mixture of LFMs

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step
- Good convergence properties (in some cases converges like second-order optimization methods)

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step
- Good convergence properties (in some cases converges like second-order optimization methods)
- Other advanced probabilistic inference algorithms are based on ideas similar to EM

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step
- Good convergence properties (in some cases converges like second-order optimization methods)
- Other advanced probabilistic inference algorithms are based on ideas similar to EM
  - E.g., Variational Bayesian (VB) inference

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step
- Good convergence properties (in some cases converges like second-order optimization methods)
- Other advanced probabilistic inference algorithms are based on ideas similar to EM
  - E.g., Variational Bayesian (VB) inference
- Note: The E and M steps may not always be possible to perform exactly

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step
- Good convergence properties (in some cases converges like second-order optimization methods)
- Other advanced probabilistic inference algorithms are based on ideas similar to EM
  - E.g., Variational Bayesian (VB) inference
- Note: The E and M steps may not always be possible to perform exactly
  - Approximate inference methods may be needed in such cases

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step
- Good convergence properties (in some cases converges like second-order optimization methods)
- Other advanced probabilistic inference algorithms are based on ideas similar to EM
  - E.g., Variational Bayesian (VB) inference
- Note: The E and M steps may not always be possible to perform exactly
  - Approximate inference methods may be needed in such cases
- EM works even if the M step is only solved approximately (Generalized EM)

EM: Some Comments

- A general framework for parameter estimation (MLE/MAP) in latent variable models
- Widely used for problems with "missing data", e.g., missing features, missing labels (SSL)
  - "Missing" parts can be treated as latent variables $\boldsymbol{z}$ and estimated in the E step
- Good convergence properties (in some cases converges like second-order optimization methods)
- Other advanced probabilistic inference algorithms are based on ideas similar to EM
  - E.g., Variational Bayesian (VB) inference
- Note: The E and M steps may not always be possible to perform exactly
  - Approximate inference methods may be needed in such cases
- 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