

DSC291: Machine Learning with Few Labels

Supervised/Unsupervised Learning

Zhiting Hu

Lecture 12, April 26, 2024

UC San Diego

HALICIOĞLU DATA SCIENCE INSTITUTE

Recap: KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$

$$\text{KL}(q(\mathbf{x}) \parallel p(\mathbf{x})) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \approx \mathbb{E}_q \left[\log \frac{q(\mathbf{x})}{p(\mathbf{x})} \right]$$

- a.k.a. Relative entropy
- $\text{KL} \geq 0$ (Jensen's inequality)
- Intuitively:
 - If q is high and p is high, then we are happy (i.e. low KL divergence)
 - If q is high and p is low then we pay a price (i.e. high KL divergence).
 - If q is low then we don't care (i.e. also low KL divergence, regardless of p)
- not a true "distance":
 - not commutative (symmetric) $\text{KL}(p \parallel q) \neq \text{KL}(q \parallel p)$
 - doesn't satisfy triangle inequality

Recap: Supervised Learning

- Model to be learned $p_{\theta}(\mathbf{x})$
- Observe **full** data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$
 - e.g., \mathbf{x}_i includes both input (e.g., image) and output (e.g., image label)
 - \mathcal{D} defines an empirical data distribution $\tilde{p}(\mathbf{x})$
 - $\mathbf{x} \sim \mathcal{D} \Leftrightarrow \mathbf{x} \sim \tilde{p}(\mathbf{x})$

$$\min_{\theta} \mathbb{E}_{x \sim \tilde{p}(x)} [-\log p_{\theta}(x)]$$

loss function

empirical risk

- Maximum Likelihood Estimation (MLE)
 - The most classical learning algorithm

$$\min_{\theta} - \mathbb{E}_{x \sim \tilde{p}(x)} [\log p_{\theta}(x)]$$

- **Question:** Show that MLE is minimizing the KL divergence between the empirical data distribution and the model distribution

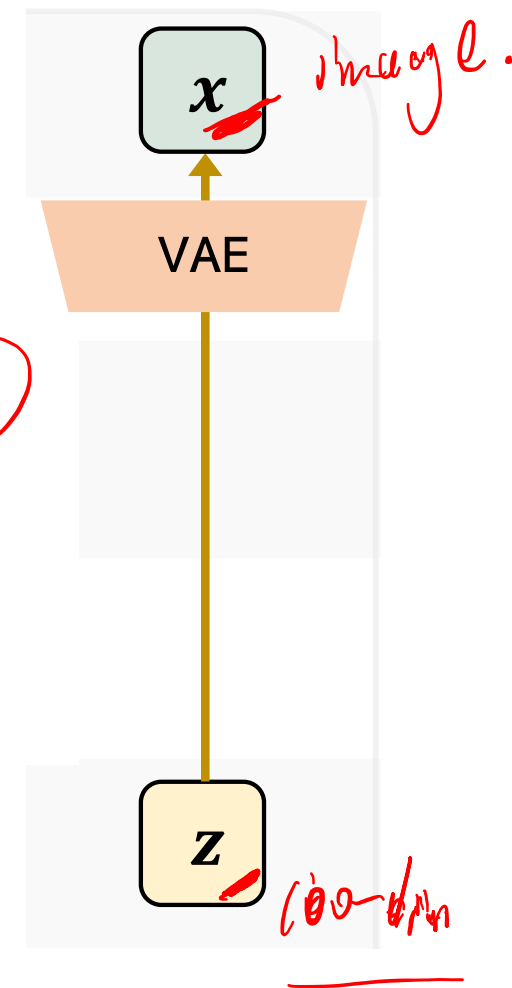
$$\min_{\theta} \text{KL}(\tilde{p}(x) \parallel p_{\theta}(x)) = -\mathbb{E}_{\tilde{p}(x)} [\log p_{\theta}(x)] + H(\tilde{p}(x))$$

Cross entropy

Unsupervised Learning

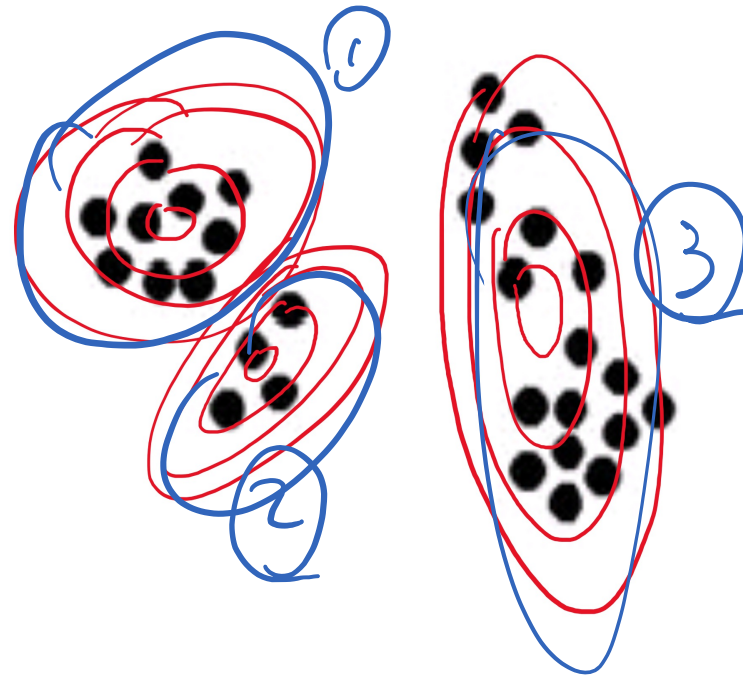
- Each data instance is partitioned into two parts:
 - observed variables x
 - latent (unobserved) variables z
- Want to learn a model $p_{\theta}(x, z)$

Label $\in (0, 12 \dots 100)$



Latent (unobserved) variables

- A variable can be unobserved (latent) because:
 - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
 - e.g., speech recognition models, mixture models, ...



Latent (unobserved) variables

- A variable can be unobserved (latent) because:
 - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
 - e.g., speech recognition models, mixture models, ...

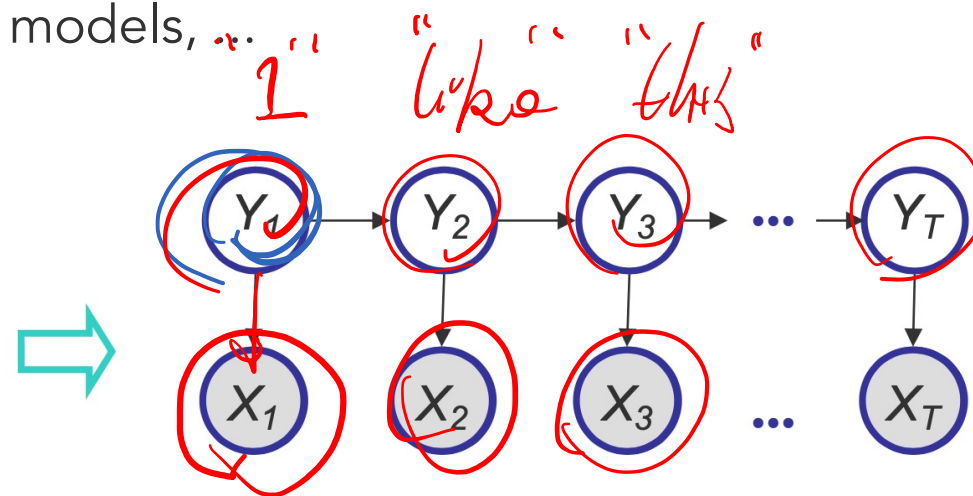
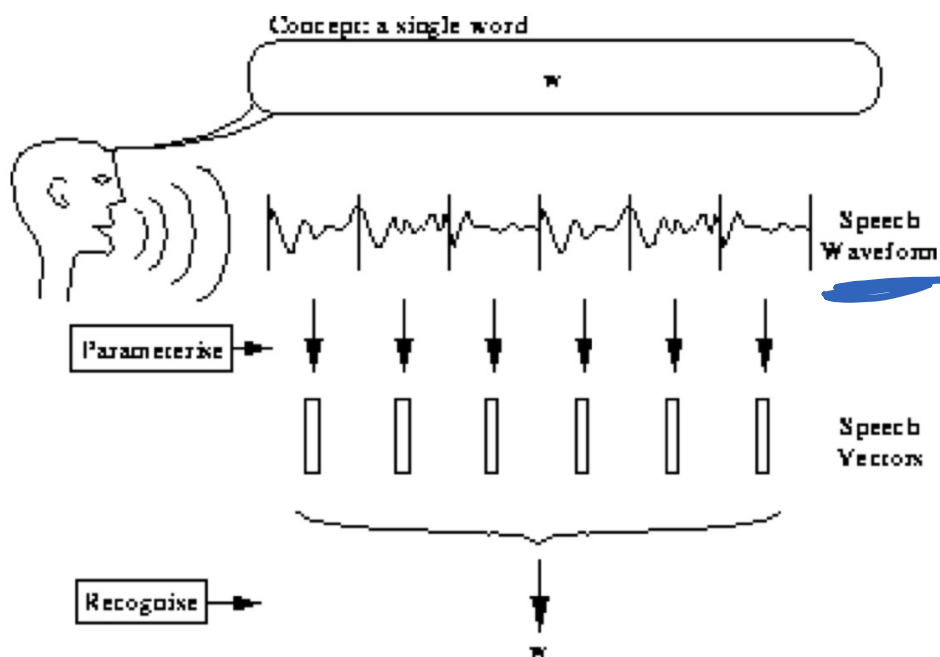


Fig. 1.2 Isolated Word Problem

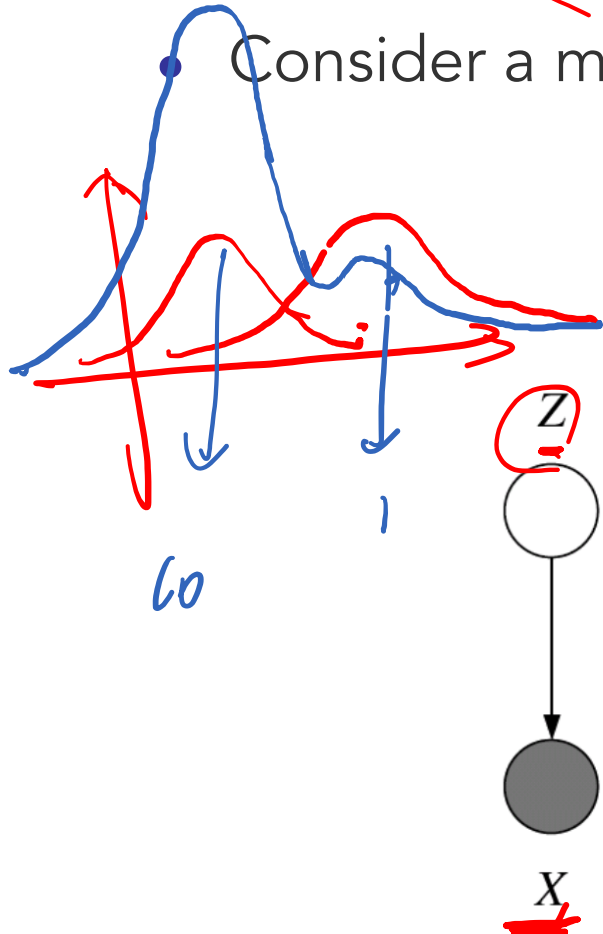
Latent (unobserved) variables

- A variable can be unobserved (latent) because:
 - imaginary quantity: meant to provide some simplified and abstractive view of the data generation process
 - e.g., speech recognition models, mixture models, ...
 - a real-world object (and/or phenomena), but difficult or impossible to measure
 - e.g., the temperature of a star, causes of a disease, evolutionary ancestors ...
 - a real-world object (and/or phenomena), but sometimes wasn't measured, because of faulty sensors, etc.
- Discrete latent variables can be used to partition/cluster data into sub-groups
- Continuous latent variables (factors) can be used for dimensionality reduction (e.g., factor analysis, etc.)

Example: Gaussian Mixture Models (GMMs)

Mixture of Experts
(MoE)

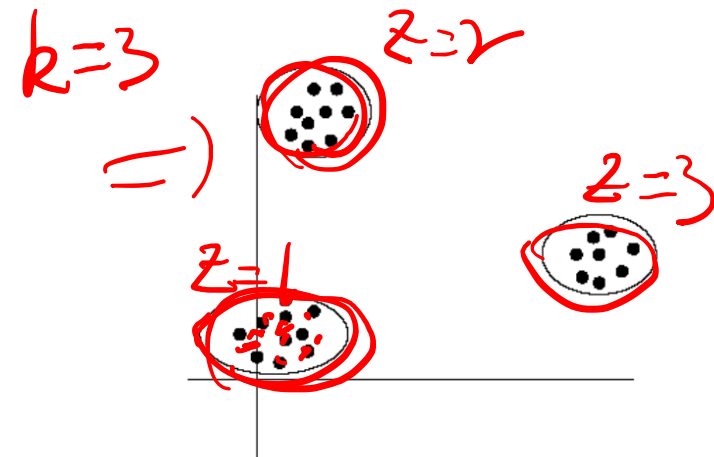
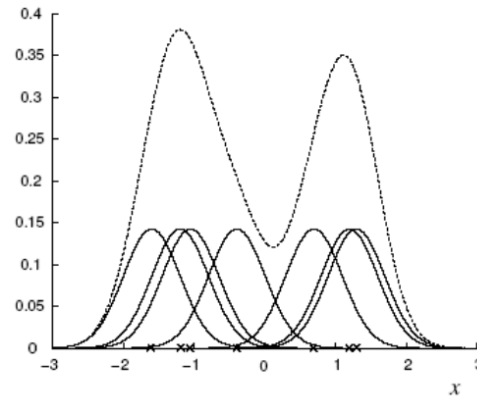
Consider a mixture of K Gaussian components:



$$p(x_n | \mu, \Sigma) = \sum_k \pi_k N(x, | \mu_k, \Sigma_k)$$

π_k
mixture proportion
mixture component

$\sum_{k=1}^K \pi_k = 1$



- This model can be used for unsupervised clustering.
 - This model (fit by AutoClass) has been used to discover new kinds of stars in astronomical data, etc.

$z_n \rightarrow x_n$

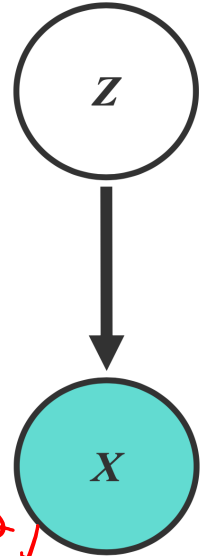
Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components:

- Z is a latent class indicator vector:

$$p(z_n) = \text{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

$\sum_k \pi_k = 1$
 $\pi_1 = 0.3$, $\pi_2 = 0.4$, $\pi_3 = 0.3$
 $k=1, 2, 3$



- X is a conditional Gaussian variable with a class-specific mean/covariance

$$p(x_n | z_n^k = 1, \mu, \Sigma) = N(x_n : \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right\}$$

- The likelihood of a sample:

Parameters to be learned:

$$p(x_n | \mu, \Sigma) = \sum_k p(z^k = 1 | \pi) p(x_n | z^k = 1, \mu, \Sigma)$$

$$= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k} \right) = \sum_k \pi_k N(x_n | \mu_k, \Sigma_k)$$

mixture proportion $\rightarrow \pi_k$
 mixture component $\rightarrow \mu_k, \Sigma_k$

Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components: $p(x_n | \mu, \Sigma) = \sum_k \pi_k N(x_n | \mu_k, \Sigma_k)$
- Recall MLE for completely observed data

- Data log-likelihood:

$$\ell(\theta; D) = \log \prod_n p(z_n, x_n) = \log \prod_n p(z_n | \pi) p(x_n | z_n, \mu, \sigma)$$

$$= \sum_n \log \prod_k \pi_k^{z_n^k} + \sum_n \log \prod_k N(x_n; \mu_k, \sigma)^{z_n^k}$$

$$= \sum_n \sum_k z_n^k \log \pi_k - \sum_n \sum_k z_n^k \frac{1}{2\sigma^2} (x_n - \mu_k)^2 + C$$

- MLE:

$$\hat{\pi}_{k,MLE} = \arg \max_{\pi} \ell(\theta; D),$$

$$\hat{\mu}_{k,MLE} = \arg \max_{\mu} \ell(\theta; D)$$

$$\hat{\sigma}_{k,MLE} = \arg \max_{\sigma} \ell(\theta; D)$$

$$\Rightarrow \hat{\mu}_{k,MLE} = \frac{\sum_n z_n^k x_n}{\sum_n z_n^k}$$

- What if we do not know z_n ?

Why is Learning Harder?

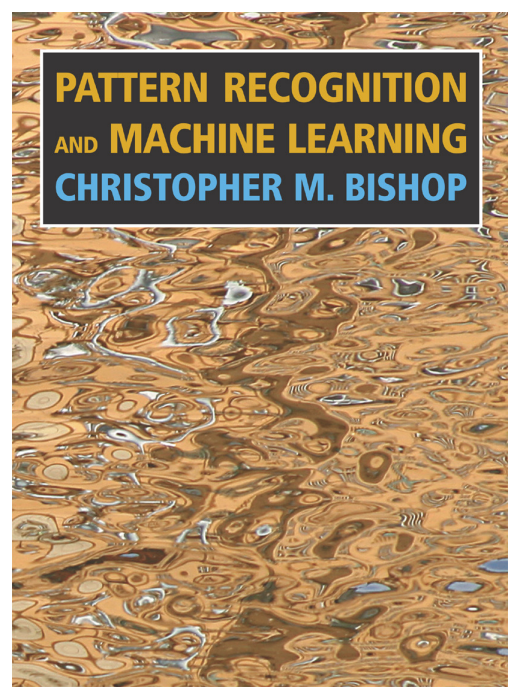
- **Complete log likelihood:** if both \mathbf{x} and \mathbf{z} can be observed, then

$$\ell_c(\theta; \mathbf{x}, \mathbf{z}) = \log p(\mathbf{x}, \mathbf{z}|\theta) = \log p(\mathbf{z}|\theta_z) + \log p(\mathbf{x}|\mathbf{z}, \theta_x)$$

- Decomposes into a sum of factors, the parameter for each factor can be estimated separately
- But given that \mathbf{z} is not observed, $\ell_c(\theta; \mathbf{x}, \mathbf{z})$ is a random quantity, cannot be maximized directly
- **Incomplete (or marginal) log likelihood:** with \mathbf{z} unobserved, our objective becomes the log of a marginal probability:

$$\ell(\theta; \mathbf{x}) = \log p(\mathbf{x}|\theta) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta)$$

- All parameters become coupled together
- In other models when \mathbf{z} is complex (continuous) variables (as we'll see later), marginalization over \mathbf{z} is intractable.



Expectation Maximization (EM)

- 9 Mixture Models and EM 423**
 - 9.1 *K*-means Clustering 424
 - 9.1.1 Image segmentation and compression 428
 - 9.2 Mixtures of Gaussians 430
 - 9.2.1 Maximum likelihood 432
 - 9.2.2 EM for Gaussian mixtures 435
 - 9.3 An Alternative View of EM 439
 - 9.3.1 Gaussian mixtures revisited 441
 - 9.3.2 Relation to *K*-means 443
 - 9.3.3 Mixtures of Bernoulli distributions 444
 - 9.3.4 EM for Bayesian linear regression 448
 - 9.4 The EM Algorithm in General 450
 - Exercises 455

This class →

Expectation Maximization (EM)

- For any distribution $q(\mathbf{z}|\mathbf{x})$, define **expected complete log likelihood**:

$$\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] = \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$$

- A deterministic function of θ
- Inherit the factorizability of $\ell_c(\theta; \mathbf{x}, \mathbf{z})$
- Use this as the surrogate objective
- Does maximizing this surrogate yield a maximizer of the likelihood?
 - We can show that:

$$\ell(\theta; \mathbf{x}) \geq \mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] + H(q)$$

Expectation Maximization (EM)

- For any distribution $q(\mathbf{z}|\mathbf{x})$, define **expected complete log likelihood**:

$$\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] = \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$$

- **Question**: show that $\ell(\theta; \mathbf{x}) \geq \mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] + H(q)$

Expectation Maximization (EM)

- For any distribution $q(\mathbf{z}|\mathbf{x})$, define **expected complete log likelihood**:

$$\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] = \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$$

- **Question**: show that $\ell(\theta; \mathbf{x}) \geq \mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] + H(q)$

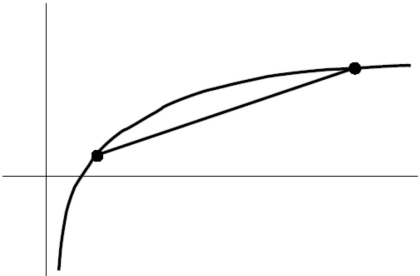
Expectation Maximization (EM)

- For any distribution $q(\mathbf{z}|\mathbf{x})$, define **expected complete log likelihood**:

$$\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] = \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$$

$$\begin{aligned} \ell(\theta; \mathbf{x}) &= \log p(\mathbf{x}|\theta) \\ &= \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta) \end{aligned}$$

Jensen's inequality



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

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

Evidence Lower Bound (ELBO)

$$= \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta) - \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log q(\mathbf{z}|\mathbf{x})$$

$$= \mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] + H(q)$$

Expectation Maximization (EM)

- For any distribution $q(\mathbf{z}|\mathbf{x})$, define **expected complete log likelihood**:

$$\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] = \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$$

- Conclusion-1:

$$\ell(\theta; \mathbf{x}) \geq \mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] + H(q) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] \quad (\text{ELBO})$$

- **Question**: show that

$$\ell(\theta; \mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] + \text{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}, \theta))$$

Expectation Maximization (EM)

- For any distribution $q(\mathbf{z}|\mathbf{x})$, define **expected complete log likelihood**:

$$\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] = \sum_{\mathbf{z}} q(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$$

- **Conclusion-1**:

$$\ell(\theta; \mathbf{x}) \geq \mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] + H(q) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] \quad (\text{ELBO})$$

- **Question**: show that

$$\ell(\theta; \mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] + \text{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}, \theta))$$

- Since KL divergence is non-negative, this is another way to prove **Conclusion-1**

Lower Bound and Free Energy

- For fixed data \mathbf{x} , define a functional called the (variational) free energy:

$$F(q, \theta) = -\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] - H(q) \geq -\ell(\theta; \mathbf{x})$$

- The EM algorithm is coordinate-decent on F
 - At each step t :

- E-step: $q^{t+1} = \arg \min_q F(q, \theta^t)$

- M-step: $\theta^{t+1} = \arg \min_{\theta} F(q^{t+1}, \theta)$

E-step: minimization of $F(q, \theta)$ w.r.t q

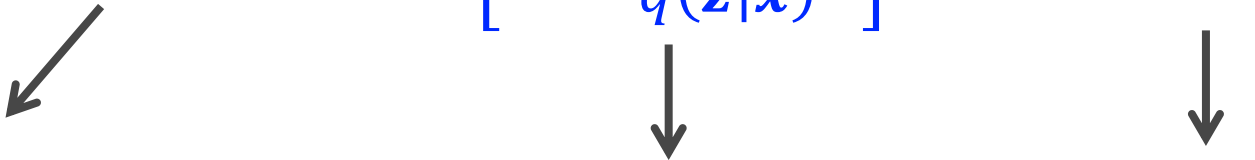
- Claim:

$$q^{t+1} = \operatorname{argmin}_q F(q, \theta^t) = p(\mathbf{z}|\mathbf{x}, \theta^t)$$

- This is the posterior distribution over the latent variables given the data and the current parameters.

- Proof (easy): recall

$$\ell(\theta^t; \mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta^t)}{q(\mathbf{z}|\mathbf{x})} \right] + \text{KL}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x}, \theta^t))$$



Independent of q $-F(q, \theta^t)$ ≥ 0

- $F(q, \theta^t)$ is minimized when $\text{KL}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x}, \theta^t)) = 0$, which is achieved only when $q(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x}, \theta^t)$

M-step: minimization of $F(q, \theta)$ w.r.t θ

- Note that the free energy breaks into two terms:

$$F(q, \theta) = -\mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] - H(q) \geq -\ell(\theta; \mathbf{x})$$

- The first term is the expected complete log likelihood and the second term, which does not depend on q , is the entropy.
- Thus, in the M-step, maximizing with respect to θ for fixed q we only need to consider the first term:

$$\theta^{t+1} = \operatorname{argmax}_{\theta} \mathbb{E}_q[\ell_c(\theta; \mathbf{x}, \mathbf{z})] = \operatorname{argmax}_{\theta} \sum_{\mathbf{z}} q^{t+1}(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$$

- Under optimal q^{t+1} , this is equivalent to solving a standard MLE of fully observed model $p(\mathbf{x}, \mathbf{z}|\theta)$, with \mathbf{z} replaced by its expectation w.r.t $p(\mathbf{z}|\mathbf{x}, \theta^t)$

EM Algorithm: Quick Summary

- Observed variables \mathbf{x} , latent variables \mathbf{z}
- To learn a model $p(\mathbf{x}, \mathbf{z}|\theta)$, we want to maximize the marginal log-likelihood

$$\ell(\theta; \mathbf{x}) = \log p(\mathbf{x}|\theta) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta)$$

- But it's too difficult
- EM algorithm:
 - maximize a lower bound of $\ell(\theta; \mathbf{x})$
 - Or equivalently, minimize an upper bound of $-\ell(\theta; \mathbf{x})$
- Key equation:

$$\ell(\theta; \mathbf{x}) = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right]}_{\text{Evidence Lower Bound (ELBO)}} + \text{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}, \theta))$$

$$= \underbrace{-F(q, \theta)}_{\text{Variational free energy}} + \text{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}, \theta))$$

Variational free energy

EM Algorithm: Quick Summary

- The EM algorithm is coordinate-descent on $F(q, \theta)$
 - E-step: $q^{t+1} = \arg \min_q F(q, \theta^t) = p(\mathbf{z}|\mathbf{x}, \theta^t)$
 - the posterior distribution over the latent variables given the data and the current parameters
 - M-step: $\theta^{t+1} = \arg \min_{\theta} F(q^{t+1}, \theta) = \operatorname{argmax}_{\theta} \sum_{\mathbf{z}} q^{t+1}(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$

$$\begin{aligned} \ell(\theta; \mathbf{x}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] + \text{KL}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x}, \theta)) \\ &= -F(q, \theta) + \text{KL}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}|\mathbf{x}, \theta)) \end{aligned}$$

Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components:

- Z is a latent class indicator vector:

$$p(z_n) = \text{multi}(z_n : \pi) = \prod_k (\pi_k)^{z_n^k}$$

- X is a conditional Gaussian variable with a class-specific mean/covariance

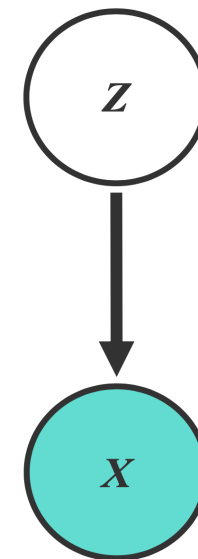
$$p(x_n | z_n^k = \mathbf{1}, \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2} (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)\right\}$$

- The likelihood of a sample:

$$\begin{aligned} p(x_n | \mu, \Sigma) &= \sum_k p(z^k = \mathbf{1} | \pi) p(x, | z^k = \mathbf{1}, \mu, \Sigma) \\ &= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k} \right) = \sum_k \pi_k N(x, | \mu_k, \Sigma_k) \end{aligned}$$

mixture proportion

mixture component



Example: Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components
- The expected complete log likelihood

$$\begin{aligned}\mathbb{E}_q [\ell_c(\boldsymbol{\theta}; x, z)] &= \sum_n \mathbb{E}_q [\log p(z_n | \pi)] + \sum_n \mathbb{E}_q [\log p(x_n | z_n, \mu, \Sigma)] \\ &= \sum_n \sum_k \mathbb{E}_q [z_n^k] \log \pi_k - \frac{1}{2} \sum_n \sum_k \mathbb{E}_q [z_n^k] \left((x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log |\Sigma_k| + C \right)\end{aligned}$$

- E-step: computing the posterior of z_n given the current estimate of the parameters (i.e., π, μ, Σ)

$$p(z_n^k = 1 | x, \mu^{(t)}, \Sigma^{(t)}) = \frac{\pi_k^{(t)} N(x_n, | \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_i \pi_i^{(t)} N(x_n, | \mu_i^{(t)}, \Sigma_i^{(t)})}$$

$\nearrow p(z_n^k = 1, x, \mu^{(t)}, \Sigma^{(t)})$
 $\searrow p(x, \mu^{(t)}, \Sigma^{(t)})$

Example: Gaussian Mixture Models (GMMs)

- E-step: computing the posterior of z_n given the current estimate of the parameters (i.e., π, μ, Σ)

$$\begin{aligned} p(z^k = 1 \mid \mathbf{x}) &= \frac{p(z^k = 1)p(\mathbf{x} \mid z^k = 1)}{p(\mathbf{x})} \\ &= \frac{p(z^k = 1)p(\mathbf{x} \mid z^k = 1)}{\sum_{j=1}^K p(z^j = 1)p(\mathbf{x} \mid z^j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} \mid \mu_j, \Sigma_j)} \\ &:= \gamma_k \end{aligned}$$

Example: Gaussian Mixture Models (GMMs)

- M-step: computing the parameters given the current estimate of z_n
 - Once we have $q^{t+1}(z^k|x) = p(z^k|x, \theta^t) = \gamma^k$, we can compute the expected likelihood:

$$\begin{aligned}\theta^{t+1} &= \operatorname{argmax}_{\theta} \sum_k q^{t+1}(z^k = 1|x) \log p(x, z^k = 1|\theta) \\ &= \mathbb{E}_{q^{t+1}} [\log (p(\mathbf{x}, z | \boldsymbol{\theta}))] \\ &= \sum_k \gamma_k (\log p(z^k = 1|\boldsymbol{\theta}) + \log P(\mathbf{x} | z^k = 1, \boldsymbol{\theta})) \\ &= \sum_k \gamma_k \log \pi_k + \sum_k \gamma_k \log \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)\end{aligned}$$

- We need to fit K Gaussians, just need to weight examples by γ_k

Example: Gaussian Mixture Models (GMMs)

- M-step: computing the parameters given the current estimate of z_n

$$\pi_k^* = \arg \max \langle l_c(\boldsymbol{\theta}) \rangle, \quad \Rightarrow \quad \frac{\partial}{\partial \pi_k} \langle l_c(\boldsymbol{\theta}) \rangle = 0, \forall k, \quad \text{s.t.} \quad \sum_k \pi_k = 1$$
$$\Rightarrow \quad \pi_k^* = \frac{\sum_n \langle z_n^k \rangle_{q^{(t)}}}{N} = \frac{\sum_n \tau_n^{k(t)}}{N} = \frac{\langle n_k \rangle}{N}$$

$$\mu_k^* = \arg \max \langle l(\boldsymbol{\theta}) \rangle, \quad \Rightarrow \quad \mu_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} x_n}{\sum_n \tau_n^{k(t)}}$$

$$\Sigma_k^* = \arg \max \langle l(\boldsymbol{\theta}) \rangle, \quad \Rightarrow \quad \Sigma_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} (x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_n \tau_n^{k(t)}}$$

Fact:

$$\frac{\partial \log |\mathbf{A}^{-1}|}{\partial \mathbf{A}^{-1}} = \mathbf{A}^T$$

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{A}} = \mathbf{x} \mathbf{x}^T$$

EM Algorithm for GMM: Quick Summary

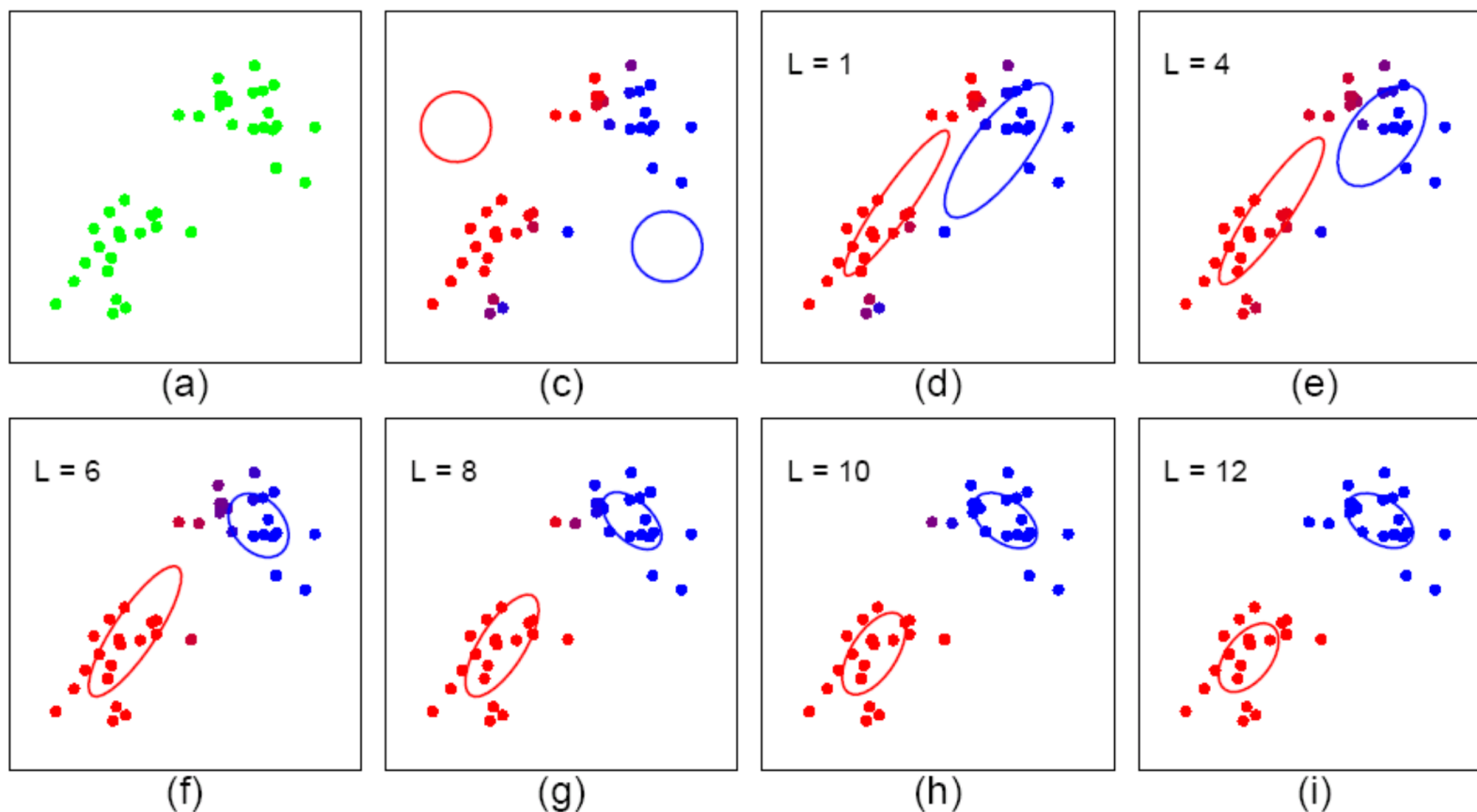
- Initialize the means μ_k , covariances Σ_k and mixing coefficients π_k
- Iterate until convergence:
 - E-step: Evaluate the posterior given current parameters

$$p(z^k = 1 \mid \mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} \mid \mu_j, \Sigma_j)} := \gamma_k$$

- M-step: Re-estimate the parameters given current posterior
 - Initialize the means μ_k , covariances Σ_k and mixing coefficients π_k
 - Iterate until convergence:
 - E-step: Evaluate the posterior given current parameters
 - M-step: Re-estimate the parameters given current posterior

Example: Gaussian Mixture Models (GMMs)

- Start: “guess” the centroid μ_k and covariance Σ_k of each of the K clusters
- Loop:



Summary: EM Algorithm

- A way of maximizing likelihood function for latent variable models. Finds MLE of parameters when the original (hard) problem can be broken up into two (easy) pieces
 - Estimate some “missing” or “unobserved” data from observed data and current parameters.
 - Using this “complete” data, find the maximum likelihood parameter estimates.

Summary: EM Algorithm

- The EM algorithm is coordinate-decent on $F(q, \theta)$

- E-step: $q^{t+1} = \arg \min_q F(q, \theta^t) = p(\mathbf{z}|\mathbf{x}, \theta^t)$

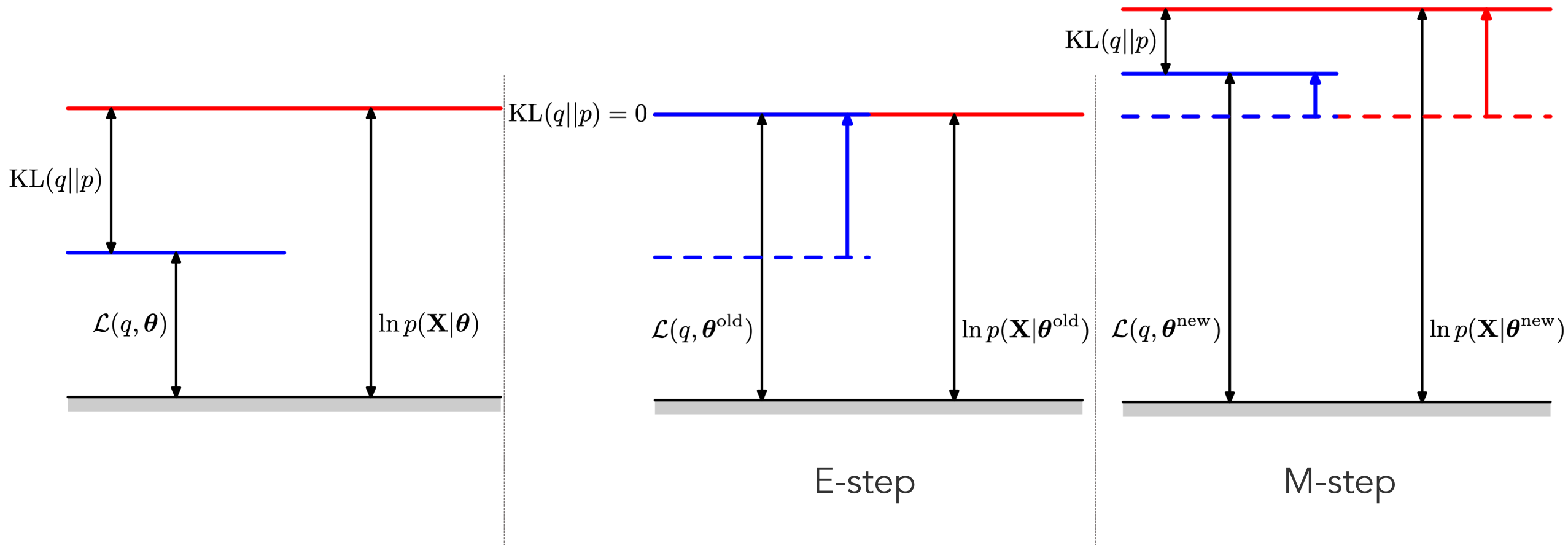
- M-step: $\theta^{t+1} = \arg \min_{\theta} F(q^{t+1}, \theta) = \operatorname{argmax}_{\theta} \sum_{\mathbf{z}} q^{t+1}(\mathbf{z}|\mathbf{x}) \log p(\mathbf{x}, \mathbf{z}|\theta)$

$$\begin{aligned} \ell(\theta; \mathbf{x}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] + \text{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}, \theta)) \\ &= -F(q, \theta) + \text{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}, \theta)) \end{aligned}$$

- Limitation: need to be able to compute $p(\mathbf{z}|\mathbf{x}, \theta)$, not possible for more complicated models --- solution: Variational inference

Each EM iteration guarantees to improve the likelihood

$$\ell(\theta; \mathbf{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] + \text{KL}(q(\mathbf{z}|\mathbf{x}) || p(\mathbf{z}|\mathbf{x}, \theta))$$



EM Variants

- Sparse EM
 - Do not re-compute exactly the posterior probability on each data point under all models, because it is almost zero.
 - Instead keep an “active list” which you update every once in a while.
- Generalized (Incomplete) EM:
 - It might be hard to find the ML parameters in the M-step, even given the completed data. We can still make progress by doing an M-step that improves the likelihood a bit (e.g. gradient step).

Questions?