DSC291: Machine Learning with Few Labels

Unsupervised Learning

Zhiting Hu

Lecture 14, May 3, 2024

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Recap: Expectation Maximization (EM)

- Supervised MLE is easy: $\max_{\theta} \ \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $\max_{\theta} \ \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_{z} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe only $\boldsymbol{x}$
- EM, intuitively:

E-step: $q(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|\boldsymbol{x}, \theta)$

We don't actually observe q, let's estimate it

M-step: $\max_{\theta} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$

Let's "pretend" we also observe $\boldsymbol{z}$ *(its distribution)*

Recap: Expectation Maximization (EM)

- Supervised MLE is easy: $\max_{\theta} \ \ell_c(\theta; \boldsymbol{x}, \boldsymbol{z}) = \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe both $\boldsymbol{x}$ and $\boldsymbol{z}$
- Unsupervised MLE is hard: $\max_{\theta} \ \ell(\theta; \boldsymbol{x}) = \log p(\boldsymbol{x}|\theta) = \log \sum_{z} p(\boldsymbol{x}, \boldsymbol{z}|\theta)$
  - Observe only $\boldsymbol{x}$
- EM, intuitively:

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

We don't actually observe q, let's estimate it

M-step: $\max_{\theta} \mathbb{E}_{q^{t+1}(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$

Let's "pretend" we also observe $\boldsymbol{z}$ *(its distribution)*

This is an iterative process

Recap: Expectation Maximization (EM)

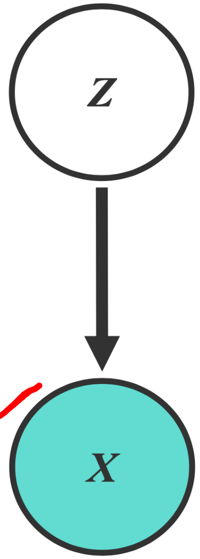
- The EM algorithm is coordinate-decent on $F(q, \theta)$
  - E-step: $q^{t+1} = \arg\min_q F\left(q, \theta^t\right) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$
    - the posterior distribution over the latent variables given the data and the current parameters
  - M-step: $\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right) = \text{argmax}_\theta\ \mathbb{E}_{q^{t+1}}[\log p(\boldsymbol{x}, \boldsymbol{z}|\theta)]$

ELBO

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

$$= -F(q, \theta) + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

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

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

  - The likelihood of a sample:

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

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

mixture proportion

mixture component

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)

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

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:

$$\theta^{t+1} = \text{argmax}_\theta \sum_k q^{t+1}(z^k = 1|x) \log p(x, z^k = 1|\theta) =$$

$$\mathbb{E}_{q^{t+1}} \left[\log\left(p\left(\boldsymbol{x}, z \mid \boldsymbol{\theta}\right)\right)\right]$$

$$= \sum_k \gamma_k \left(\log p\left(z^k = 1|\boldsymbol{\theta}\right) + \log P\left(\boldsymbol{x} \mid z^k = 1, \boldsymbol{\theta}\right)\right)$$

$$= \sum_k \gamma_k \log \pi_k + \sum_k \gamma_k \log \mathcal{N}\left(\boldsymbol{x}; \mu_k, \Sigma_k\right)$$

  - 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

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

$$\Rightarrow \quad \pi_k^* = \sum_n \langle z_n^k \rangle_{q^{(t)}} \Big/ N = \sum_n \tau_n^{k(t)} \Big/ N = \langle n_k \rangle \Big/ N$$

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

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

Fact:

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

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

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 Bayes' rule

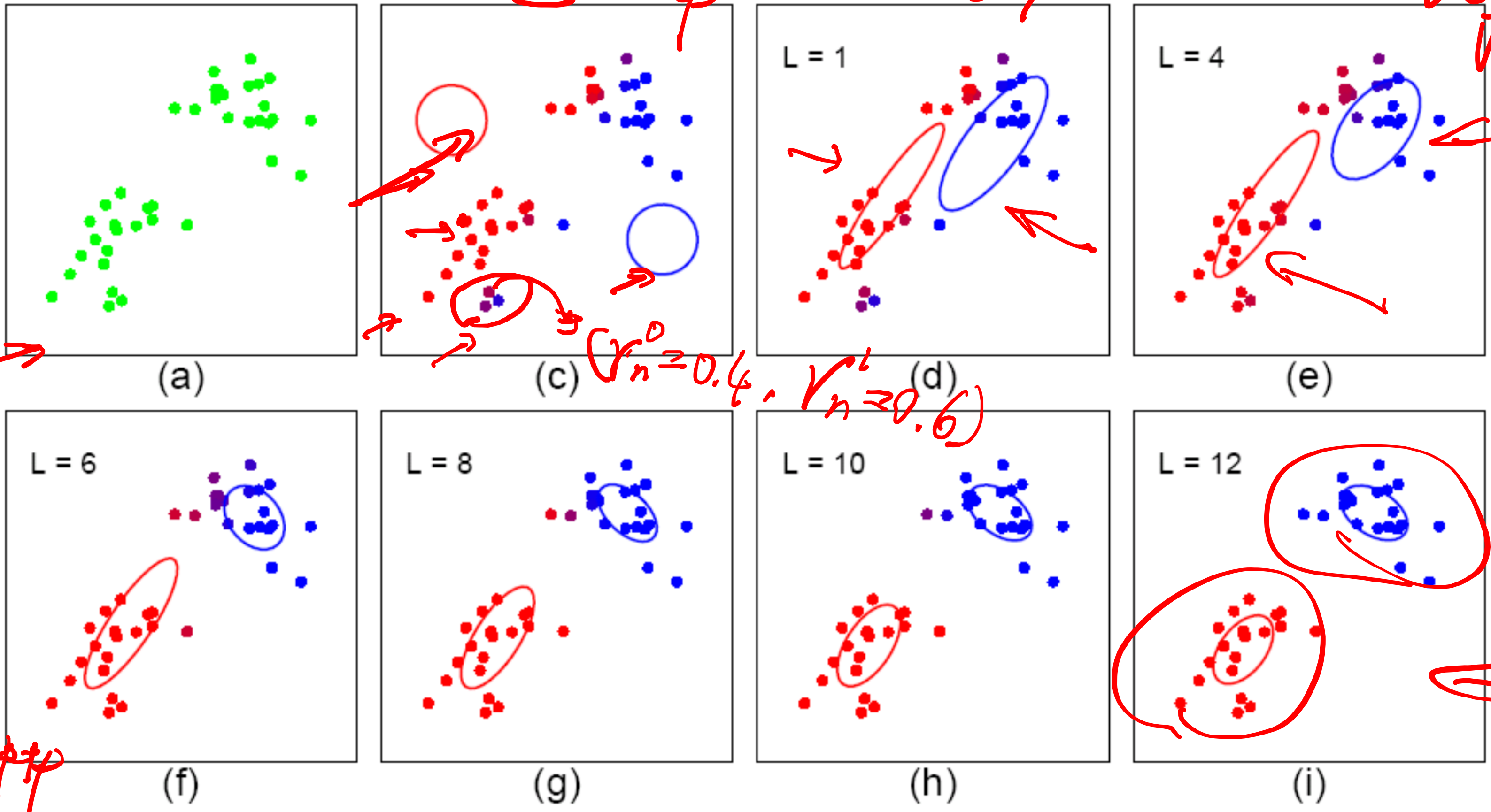
$$p(z^k = 1 \mid \boldsymbol{x}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x} \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x} \mid \mu_j, \Sigma_j)} := \gamma_k$$

γ_n^k

  - 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: $q^{t+1} = \arg\min_q F\left(q, \theta^t\right) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$
  - M-step: $\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right) = \text{argmax}_\theta \sum_z q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$

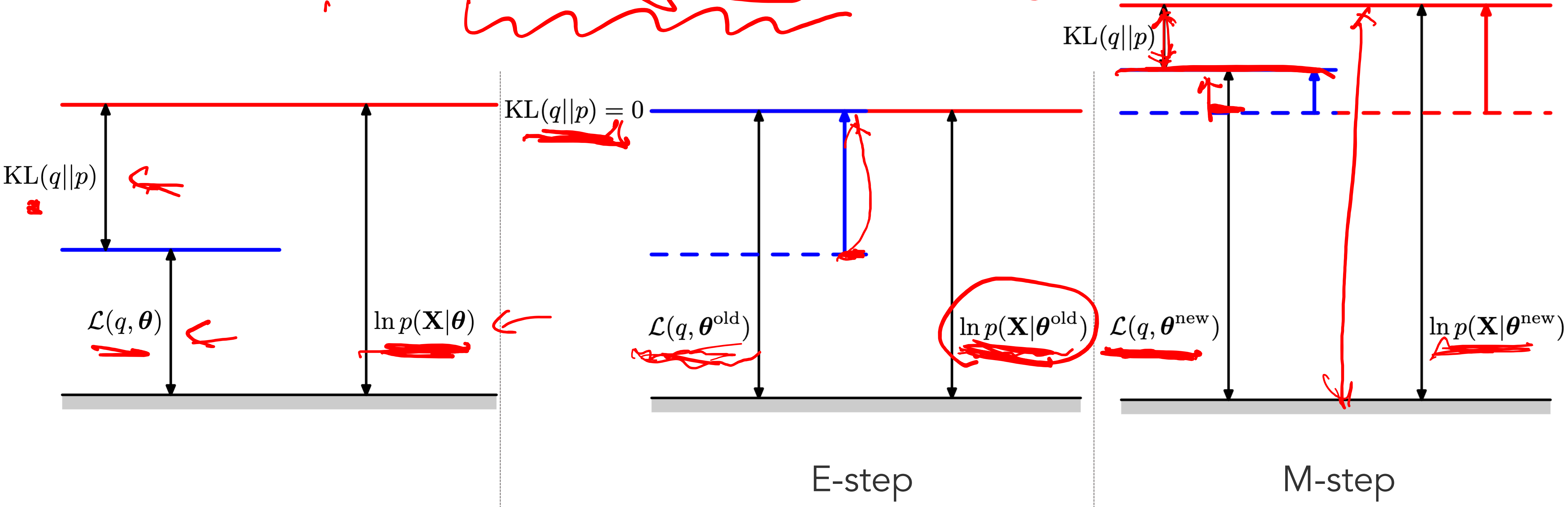
$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \text{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

$$= -F(q, \theta) + \text{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

ELBO ≤ likelihood

Each EM iteration guarantees to improve the likelihood

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



[PRML, Chap 9.4]

Summary: EM Algorithm

- The EM algorithm is coordinate-decent on $F(q, \theta)$
  - E-step: $$q^{t+1} = \arg\min_q F\left(q, \theta^t\right) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^t)$$
  - M-step: $$\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right) = \text{argmax}_\theta \sum_z q^{t+1}(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{x}, \boldsymbol{z}|\theta)$$

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

$$= -F(q, \theta) + \text{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

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

Variational Inference

Content adapted from CMU 10-708 Spring 2017

Inference

- Given a model, the goals of inference can include:
  - Computing the likelihood of observed data $p(\boldsymbol{x}^*)$
  - Computing the marginal distribution over a given subset of variables in the model $p(\boldsymbol{x}_A)$
  - Computing the conditional distribution over a subsets of nodes given a disjoint subset of nodes $p(\boldsymbol{x}_A | \boldsymbol{x}_B)$
  - Computing a mode of the density (for the above distributions) $\text{argmax}_{\boldsymbol{x}}\, p(\boldsymbol{x})$
  - ….

Variational Inference

- Observed variables $\boldsymbol{x}$, latent variables $\boldsymbol{z}$
- Variational (Bayesian) inference, a.k.a. **variational Bayes**, is most often used to approximately infer the conditional distribution over the latent variables given the observations (and parameters)
  - i.e., the **posterior distribution** over the latent variables

$$p(\boldsymbol{z}|\boldsymbol{x},\theta) = \frac{p(\boldsymbol{z},\boldsymbol{x}|\theta)}{\sum_z p(\boldsymbol{z},\boldsymbol{x}|\theta)}$$

Motivating Example

- Why do we often need to use an approximate inference methods (such as variational Bayes) to compute the posterior distribution?

- It’s because we cannot directly compute the posterior distribution for many interesting models
  - I.e. the posterior density is in an intractable form (often involving integrals) which cannot be easily analytically solved.

- As a motivating example, we will try to compute the posterior for a (Bayesian) mixture of Gaussians.

Bayesian mixture of Gaussians

- The mean μ_k is treated as a (latent) random variable

$$\mu_k \sim \mathcal{N}(0, \tau^2) \text{ for } k = 1, \ldots, K$$

- For each data $i = 1, \ldots, n$

$$z_i \sim \mathrm{Cat}(\pi).$$

$$x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2).$$

- We have
  - observed variables $x_{1:n}$
  - latent variables $\mu_{1:k}$ and $z_{1:n}$
  - parameters $\{\tau^2, \pi, \sigma^2\}$
- $p(x_{1:n}, z_{1:n}, \mu_{1:k} | \tau^2, \pi, \sigma^2) = \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})$

Bayesian mixture of Gaussians

- We can write the posterior distribution as

$$p(\mu_{1:K}, z_{1:n} | x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i | z_i, \mu_{1:K})}$$

- The numerator can be computed for any choice of the latent variables
- The problem is the denominator (the marginal probability of the observations)
  - This integral cannot easily be computed analytically
- We need some approximation..

Variational Inference

The main idea behind variational inference:

- Choose a family of distributions over the latent variables $z_{1:m}$ with its own set of variational parameters ν , i.e.

$$q(z_{1:m}|\nu)$$

- Then, we find the setting of the parameters that makes our approximation q closest to the posterior distribution.
  - This is where optimization algorithms come in.

- Then we can use q with the fitted parameters in place of the posterior.
  - E.g. to form predictions about future data, or to investigate the posterior distribution over the hidden variables, find modes, etc.

Variational Inference

- We want to minimize the KL divergence between our approximation $q(\boldsymbol{z}|\boldsymbol{x})$ and our posterior $p(\boldsymbol{z}|\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x})\ ||\ p(\boldsymbol{z}|\boldsymbol{x}))$$

  - But we can't actually minimize this quantity w.r.t q because $p(\boldsymbol{z}|\boldsymbol{x})$ is unknown

Evidence Lower Bound (ELBO)

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

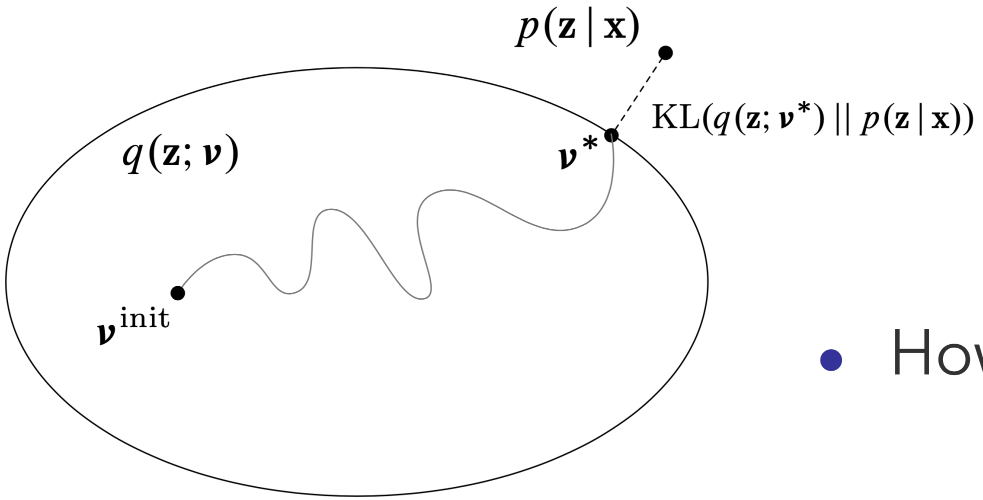
- The ELBO is equal to the negative KL divergence up to a constant $\ell(\theta;\boldsymbol{x})$
- We maximize the ELBO over q to find an "optimal approximation" to $p(\boldsymbol{z}|\boldsymbol{x})$

Variational Inference

- Choose a family of distributions over the latent variables $\boldsymbol{z}$ with its own set of variational parameters ν , i.e. $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{v})$
- We maximize the ELBO over q to find an "optimal approximation" to $p(\boldsymbol{z}|\boldsymbol{x})$

$$\text{argmax}_{\nu}\ \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})}\right]$$

$$= \text{argmax}_{\nu}\ \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})}[\log p(\boldsymbol{x},\boldsymbol{z}|\theta)] - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})}[\ \log q(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{v})\ ]$$



- How do we choose the variational family $q(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{v})$?

Questions?