DSC291: Machine Learning with Few Labels

Unsupervised Learning

Zhiting Hu
Lecture 13, April 29, 2024

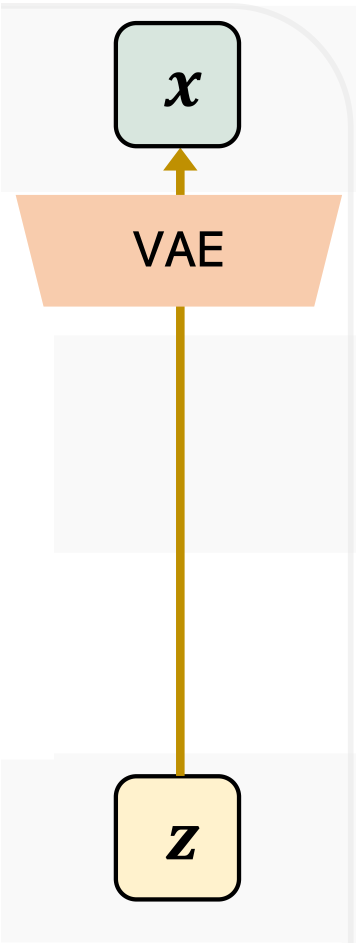
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Logistics

- 04/29: today, lecture #13
- 05/01: no class
- 05/03: Zoom (zoom link on Piazza)
- 05/06: Zoom
- 05/08: Zoom
- 05/10: Zoom
- 05/13 and future: in-person, lecture + paper presentations

Recap: Unsupervised Learning

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



[Content adapted from CMU 10-708]

Recap: Why is Learning Harder?

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

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

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

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

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

Recap: Expectation Maximization (EM)

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

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

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

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

intractable

Expectation Maximization (EM)

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

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

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

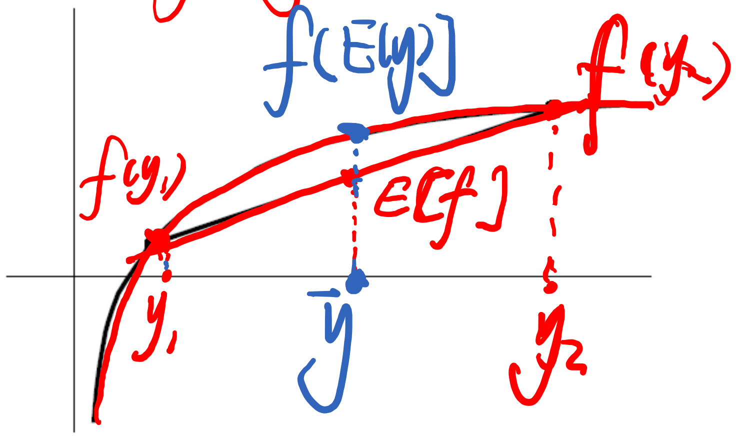
- **Hint**: a useful inequality: Jensen's inequality
  - If f is convex:

$$\underline{\mathbb{E}_{p(\boldsymbol{y})}[f(\boldsymbol{y})]} \geq \underline{f(\mathbb{E}_{p(\boldsymbol{y})}[\boldsymbol{y}])}$$

f : concave

$- - - \leq - -$

$f(y) = \log y$



Expectation Maximization (EM)

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

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

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

$$\ell(\theta;x) = \log \sum_z p(x,z|\theta)$$

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

$$= \log E_{q(z|x)}\left[\frac{p(x,z|\theta)}{q(z|x)}\right]$$

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

$$= E_q[\log p(x,z|\theta)] \checkmark$$

$$+ \underbrace{E_q[-\log q(z|x)]}_{H(q)}$$

Expectation Maximization (EM)

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

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

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

Expectation Maximization (EM)

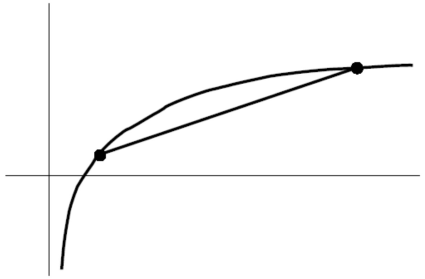
- For any distribution $q(\boldsymbol{z}|\boldsymbol{x})$, define expected complete log likelihood:

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

$$\begin{aligned}
\ell(\theta;\underline{x}) &= \log p(x \mid \theta) \\
&= \log \sum_z p(x,z \mid \theta) \\
&= \log \sum_z q(z \mid x) \frac{p(x,z \mid \theta)}{q(z \mid x)} \\
&\geq \underline{\sum_z q(z \mid x) \log \frac{p(x,z \mid \theta)}{q(z \mid x)}} \checkmark \\
&= \sum_z q(z \mid x) \log p(x,z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x) \\
&= \mathbb{E}_q[\ell_c(\theta;\boldsymbol{x},\boldsymbol{z})] + H(q) \checkmark
\end{aligned}$$

Evidence Lower Bound (ELBO) → data.

Jensen's inequality



Expectation Maximization (EM)

Bayes' rule: $p(z|x,\theta) = \frac{p(x,z|\theta)}{p(x|\theta)}$

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

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

- Conclusion-1:

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

- **Question**: show that

☆ $$\underset{D}{\ell(\theta;\boldsymbol{x})} = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \overset{\geq 0}{\mathrm{KL}\left(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x},\theta)\right)}$$

$$= E_q\left[\log \frac{p(x,z|\theta)}{q}\right] + E_q\left[\log \frac{q}{p(z|x,\theta)}\right] = E_q\left[\log \frac{p(x,z|\theta)}{\cancel{q}} \cdot \frac{\cancel{q}}{p(z|x,\theta)}\right]$$

$$= E_q\left[\log \frac{p(x,z|\theta)}{p(z|x,\theta)}\right] = E_q[\log p(x|\theta)] = \log p(x|\theta)$$

Expectation Maximization (EM)

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

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

- Conclusion-1:

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

= ELBO

||

Variational

(ELBO) free energy

- **Question**: show that

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

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

Lower Bound and Free Energy

$\min F(q, \theta) \geq -\ell(\theta; x)$

$q(z|\theta x)$

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

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

- The EM algorithm is coordinate-decent on F

$q, \theta.$

  - At each step t:
    - E-step: $$q^{t+1} = \arg\min_q F\left(q, \theta^t\right)$$
    - M-step: $$\theta^{t+1} = \arg\min_\theta F\left(q^{t+1}, \theta^t\right)$$

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

$F(q, \theta^t)$

$= E$

☆

- **Question**: show that that optimal solution of E-step is

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

  - I.e., the posterior distribution over the latent variables given the data and the current parameters.

- **Hint:** use the fact

Evidence Lower Bound (ELBO)

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

const.

$q = p$

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

Variational free energy

$\text{KL} = 0$

iff $q = p$

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

- Claim:

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

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

- Proof (easy): recall

$$\underbrace{\ell(\theta^t;\boldsymbol{x})}_{\text{Independent of } q} = \underbrace{\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z}|\theta^t)}{q(\boldsymbol{z}|\boldsymbol{x})}\right]}_{-F(q,\theta^t)} + \underbrace{\mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x},\theta^t)\big)}_{\geq 0}$$

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

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

- Note that the free energy breaks into two terms:

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

  - 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:

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

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

EM Algorithm: Quick Summary

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

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

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

Evidence Lower Bound (ELBO)

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

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

Variational free energy

EM Algorithm: Quick Summary

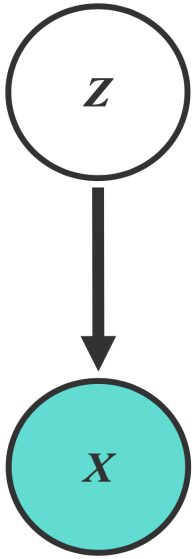
- The EM algorithm is coordinate-decent on $F(q, \theta)$
  - ○ E-step: $q^{t+1} = \arg\min_{q} F\left(q, \theta^{t}\right) = p(\boldsymbol{z}|\boldsymbol{x}, \theta^{t})$
    - the posterior distribution over the latent variables given the data and the current parameters

$\mathbb{E}_q[\log p(x, z|\theta)]$

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

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

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

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

  - The likelihood of a sample:

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

mixture proportion

mixture component

Example: Gaussian Mixture Models (GMMs)

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

$$\mathbb{E}_q\left[\ell_c(\boldsymbol{\theta}; x, z)\right] = \sum_n \mathbb{E}_q\left[\log p\left(z_n \mid \pi\right)\right] + \sum_n \mathbb{E}_q\left[\log p\left(x_n \mid z_n, \mu, \Sigma\right)\right]$$

$$= \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \log \pi_k - \frac{1}{2} \sum_n \sum_k \mathbb{E}_q\left[z_n^k\right] \left((x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log |\Sigma_k| + C \right)$$

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

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

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

$$p(x \mid \mu^{(t)}, \Sigma^{(t)})$$

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)

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

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:

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

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

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

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

  - 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

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

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

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

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

Fact:

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

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

EM Algorithm for GMM: Quick Summary

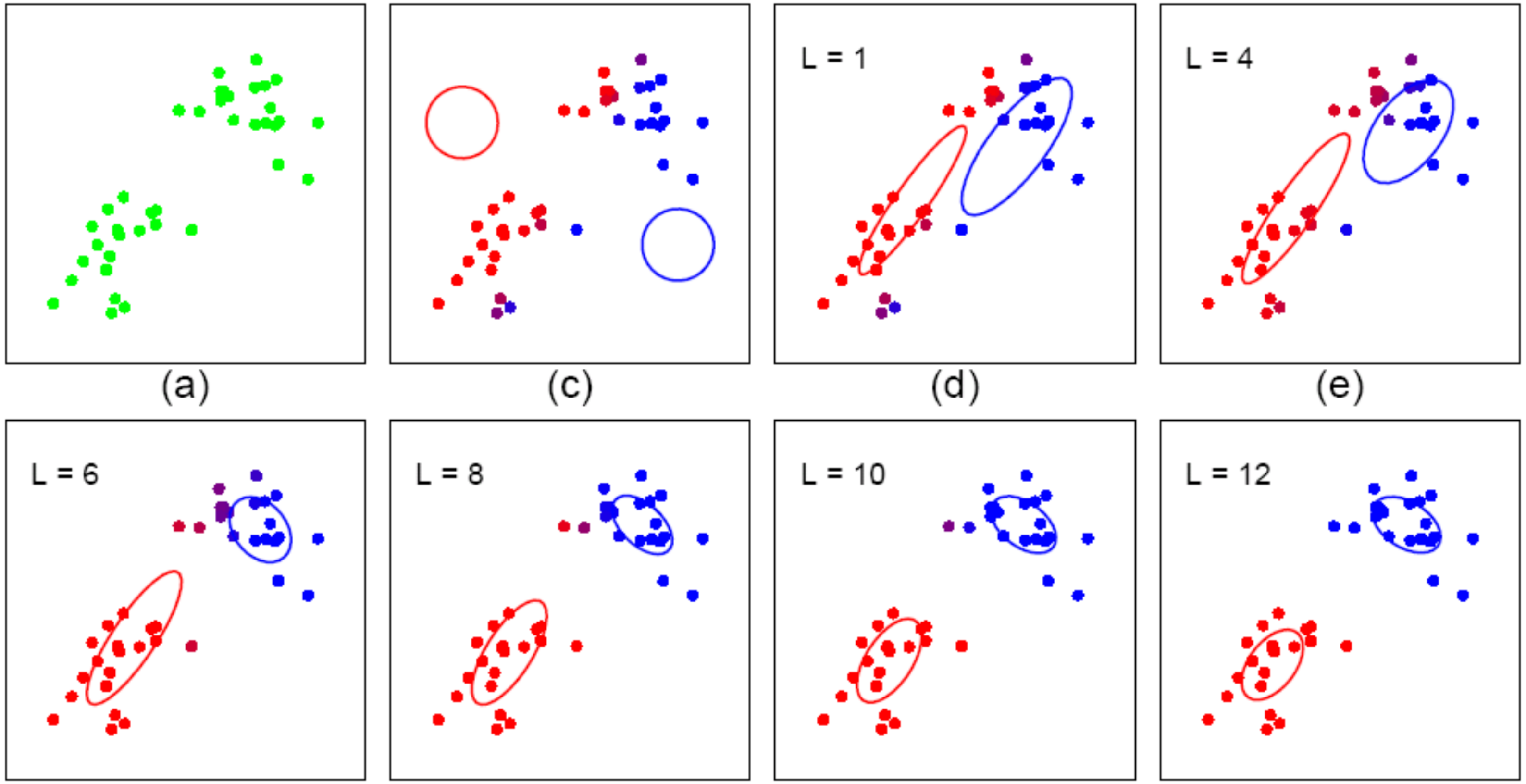
- Initialize the means μ_k , covariances Σ_k and mixing coefficients π_k
- Iterate until convergence:
  - E-step: Evaluate the posterior given current parameters

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

  - 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:



(a) (c) (d) (e)

(f) (g) (h) (i)

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

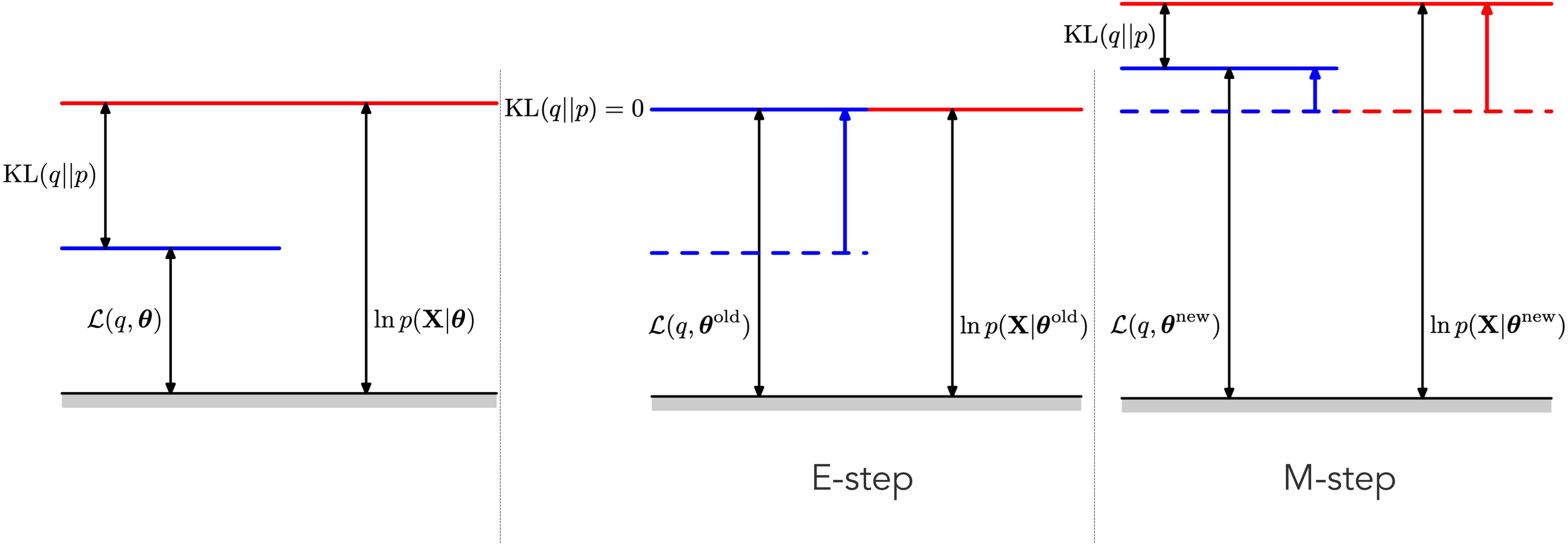
$$\ell(\theta; \boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z}|\theta)}{q(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathrm{KL}\big(q(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z}|\boldsymbol{x}, \theta)\big)$$

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

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

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]

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?