Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 12: Learning from Partially-Observed Data

March 6, 2025

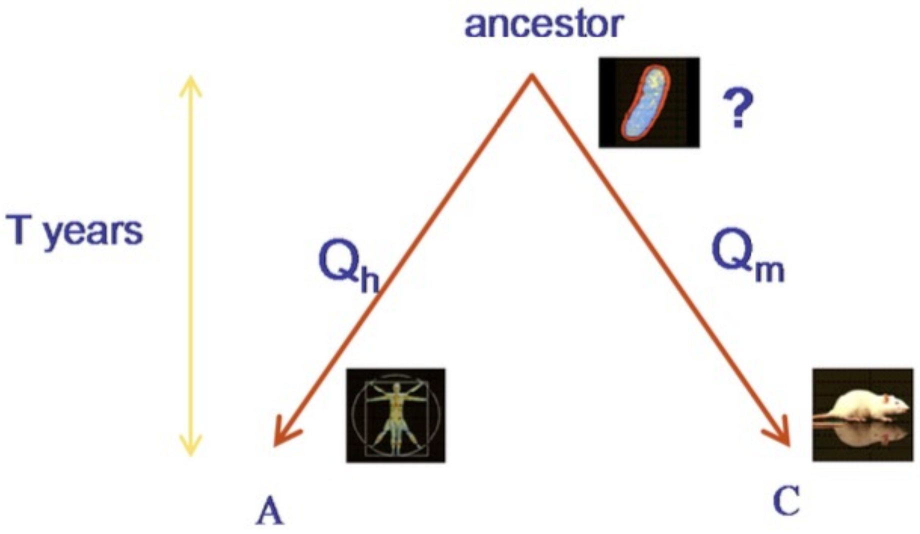
Reading: See course homepage

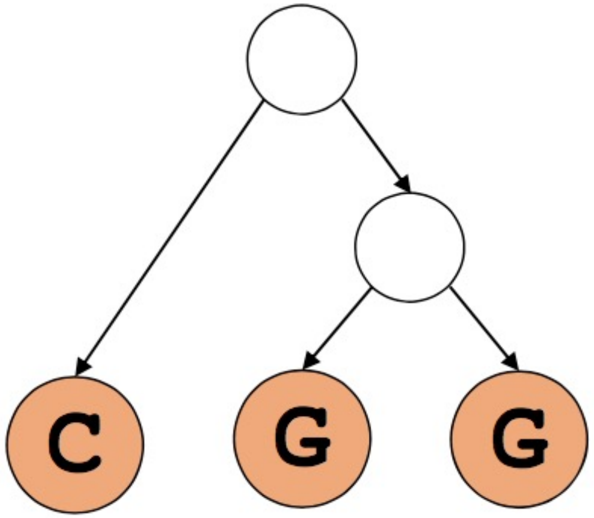
Today

- Partially-Observed GMs
- Expectation-Maximization
  - K-Means Clustering

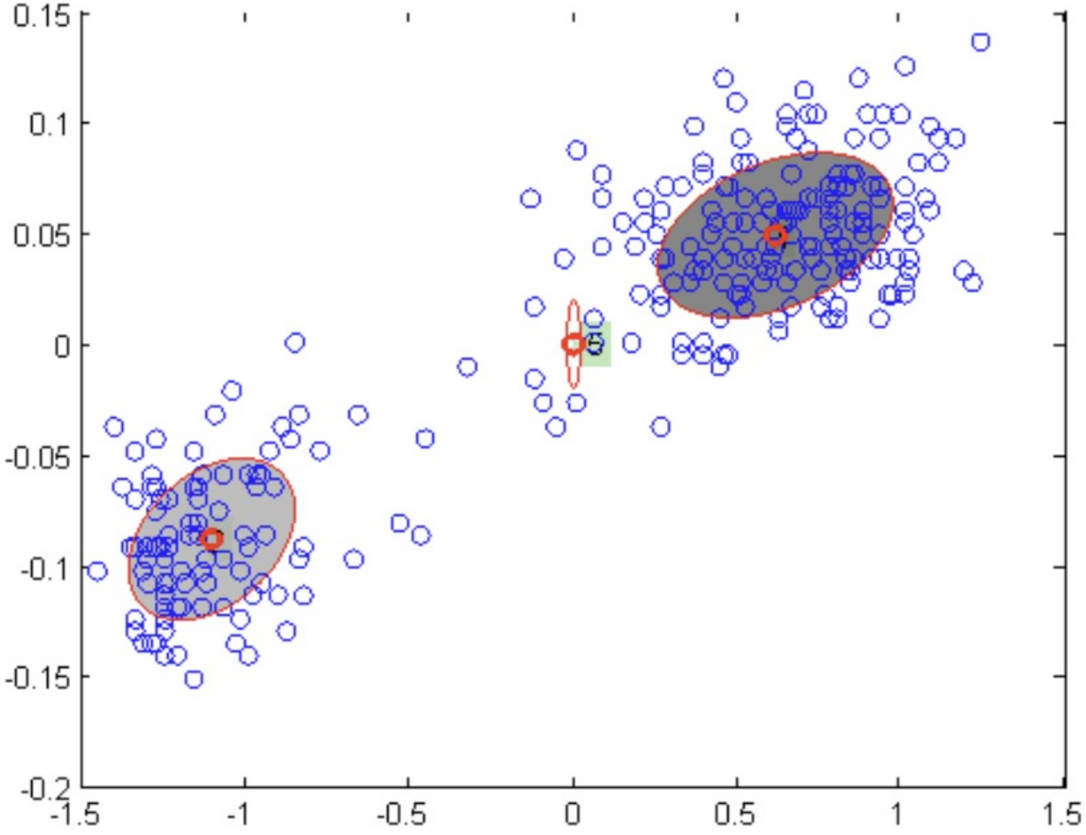
Partially Observed GMs

Partially-Observed GMs





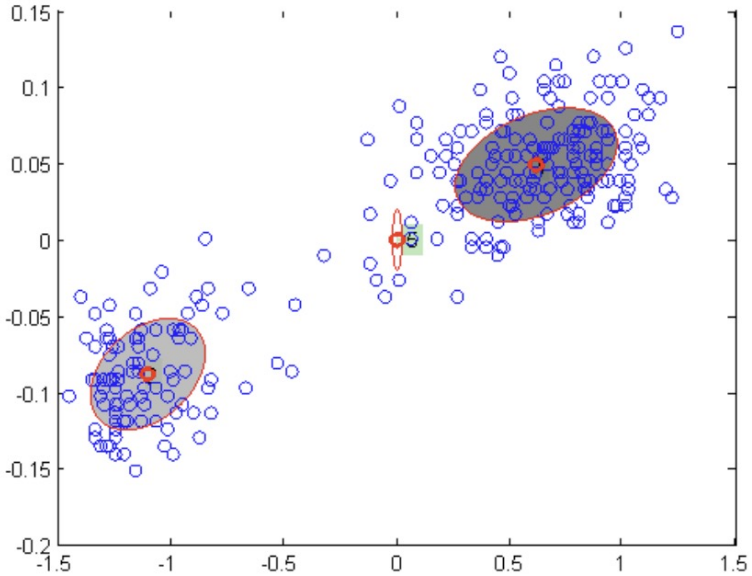
Partially-Observed GMs

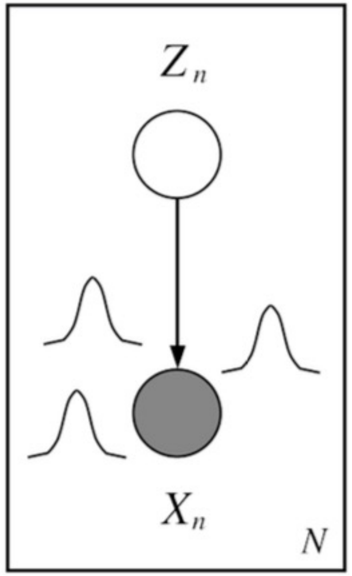


Mixture Models

Partially-Observed GMs: Mixture models

- A density model $p(x)$ may be multi-modal
- Can we model it as a mixture of uni-modal distributions?





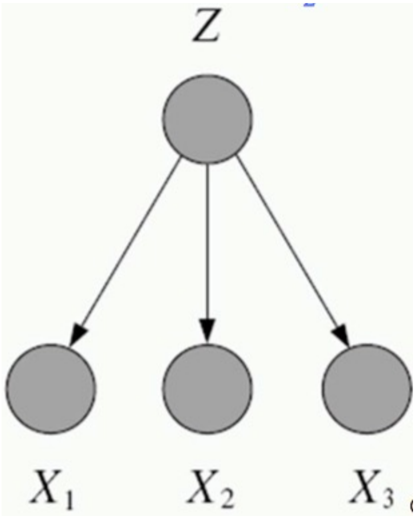
Unobserved Variables

- A variable can be unobserved (latent) because:
  - It is difficult or impossible to measure
    - e.g. Causes of a disease, evolutionary ancestors
  - It is only sometimes measured
    - e.g. faulty sensors
  - It is an imaginary quantity meant to provide some simplified but useful view of the data generation process
    - e.g. Mixture assignments
- Discrete latent variables can be used for as cluster assignments
- Continuous latent variables can be used for dimensionality reduction

Why is learning with latent variables harder?

- In fully-observed IID settings, the log-likelihood decomposes into a sum of local terms:

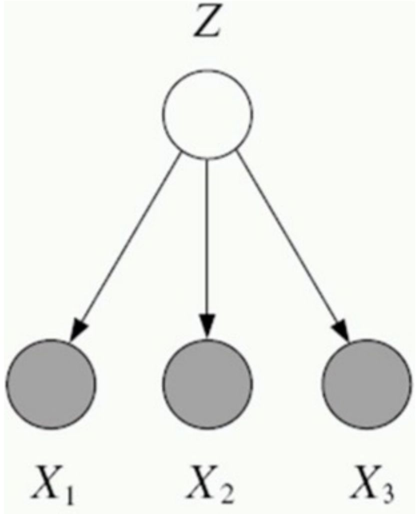
$$\ell_c(\theta; D) = \log p(x, z \mid \theta) = \log p(z \mid \theta_z) + \log p(x \mid z, \theta_x)$$



- With latent variables, all parameters become coupled via marginalization

$$\ell_c(\theta; D) = \log \sum_z p(x, z \mid \theta) = \log \sum_z p(z \mid \theta_z) p(x \mid z, \theta_x)$$

Sum over z is inside log



Strategy:

1. Guess value of Z
2. Apply MLE to estimate best model parameters based on Z
3. Inference most likely Z based on MLE parameter estimates
4. Return to step 2 until Z stops changing

Expectation-Maximization

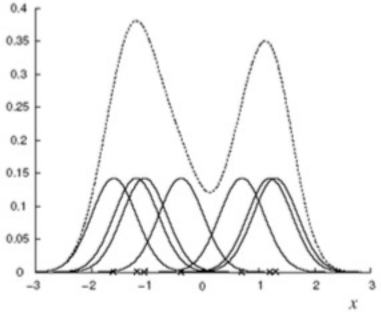
Gaussian Mixture Models (GMMs)

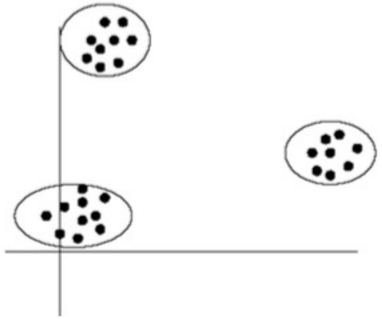
- Consider a mixture of K Gaussian components:

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

mixture proportion

mixture component





- This model can be used for unsupervised clustering

Gaussian Mixture Models (GMMs)

- Consider a mixture of K Gaussian components:
  - Z is a latent class indicator

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

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

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

  - Likelihood

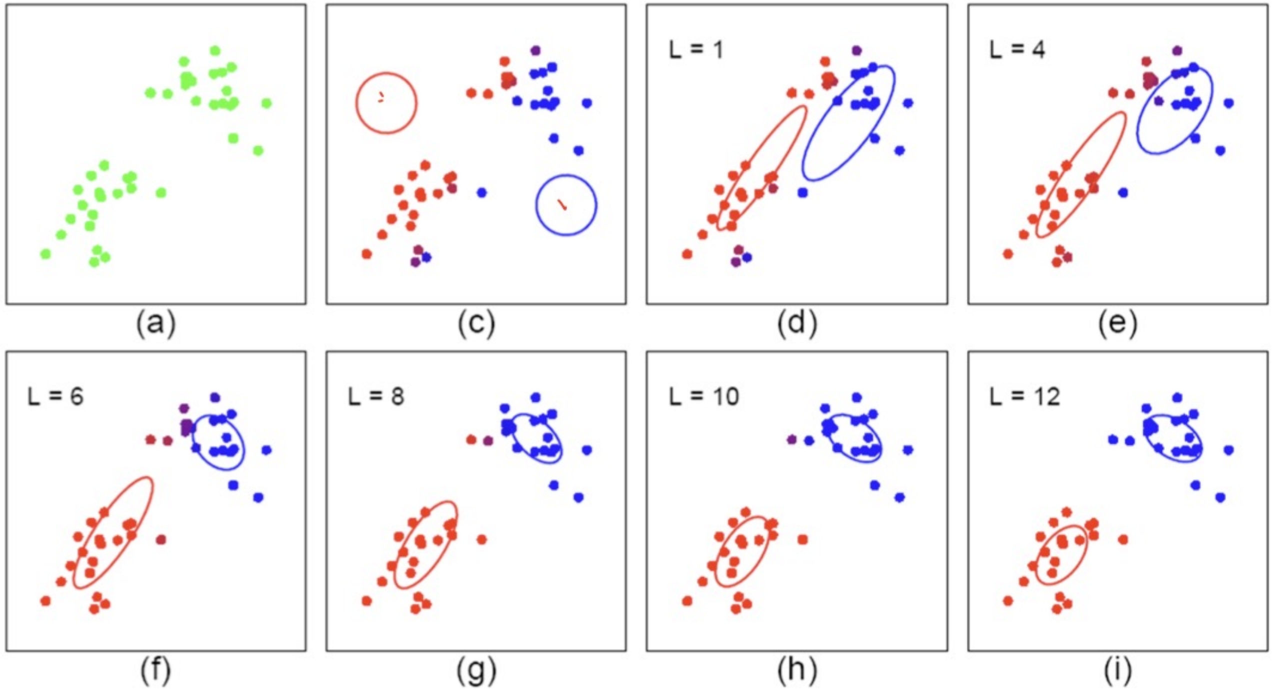
mixture component

mixture proportion

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

Expectation-Maximization for GMMs

- Start
  - Guess the value of centroids μ_k and covariances Σ_k of each of the K clusters
- Loop



Towards Expectation-Maximization

- Start from MLE for completely-observed data:

$$\begin{aligned}\boldsymbol{\ell}(\boldsymbol{\theta}; D) &= \log \prod_n p(z_n, x_n) = \log \prod_n p(z_n \mid \pi) p(x_n \mid z_n, \mu, \sigma) \\ &= \sum_n \log \prod_k \pi_k^{z_n^k} + \sum_n \log \prod_k N(x_n; \mu_k, \sigma)^{z_n^k} \\ &= \sum_n \sum_k z_n^k \log \pi_k - \sum_n \sum_k z_n^k \tfrac{1}{2\sigma^2} (x_n - \mu_k)^2 + C\end{aligned}$$

- Gives nice MLE estimators:

$$\begin{aligned}\hat{\pi}_{k,MLE} &= \arg\max_{\pi} \boldsymbol{\ell}(\boldsymbol{\theta}; D), \\ \hat{\mu}_{k,MLE} &= \arg\max_{\mu} \boldsymbol{\ell}(\boldsymbol{\theta}; D) \\ \hat{\sigma}_{k,MLE} &= \arg\max_{\sigma} \boldsymbol{\ell}(\boldsymbol{\theta}; D)\end{aligned}$$

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

We don’t know z

Towards Expectation-Maximization

- Likelihood for unobserved z:

mixture component

mixture proportion

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

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

- The expected log-likelihood is then:

Expectation over current p(z|x)

$$\langle \ell_c(\theta; x, z) \rangle = \sum_n \langle \log p(z_n | \pi) \rangle_{p(z|x)} + \sum_n \langle \log p(x_n | z_n, \mu, \Sigma) \rangle_{p(z|x)}$$

$$= \sum_n \sum_k \langle z_n^k \rangle \log \pi_k - \frac{1}{2} \sum_n \sum_k \langle z_n^k \rangle \left((x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) + \log|\Sigma_k| + C \right)$$

Expectation-Maximization algorithm

- **E-step**:
  - Compute the expected value of the sufficient statistics of the hidden variables under current estimates of parameters

- **M-step**:
  - Using the current expected value of the hidden variables, compute the parameters that maximize the likelihood.

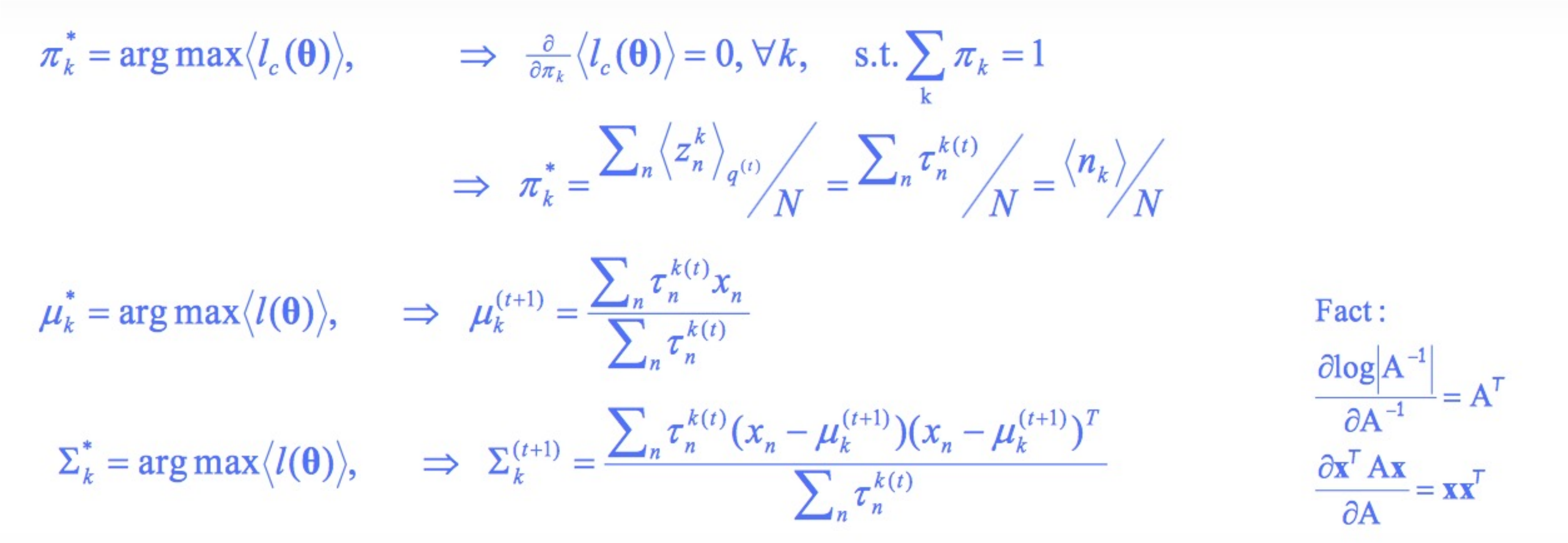
Expectation-Maximization for our GMM

- **E-step**:
  - Compute the expected value of the sufficient statistics of the hidden variables under current estimates of parameters

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

Expectation-Maximization algorithm

- **M-step**:
  - Using the current expected value of the hidden variables, compute the parameters that maximize the likelihood.



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

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

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

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

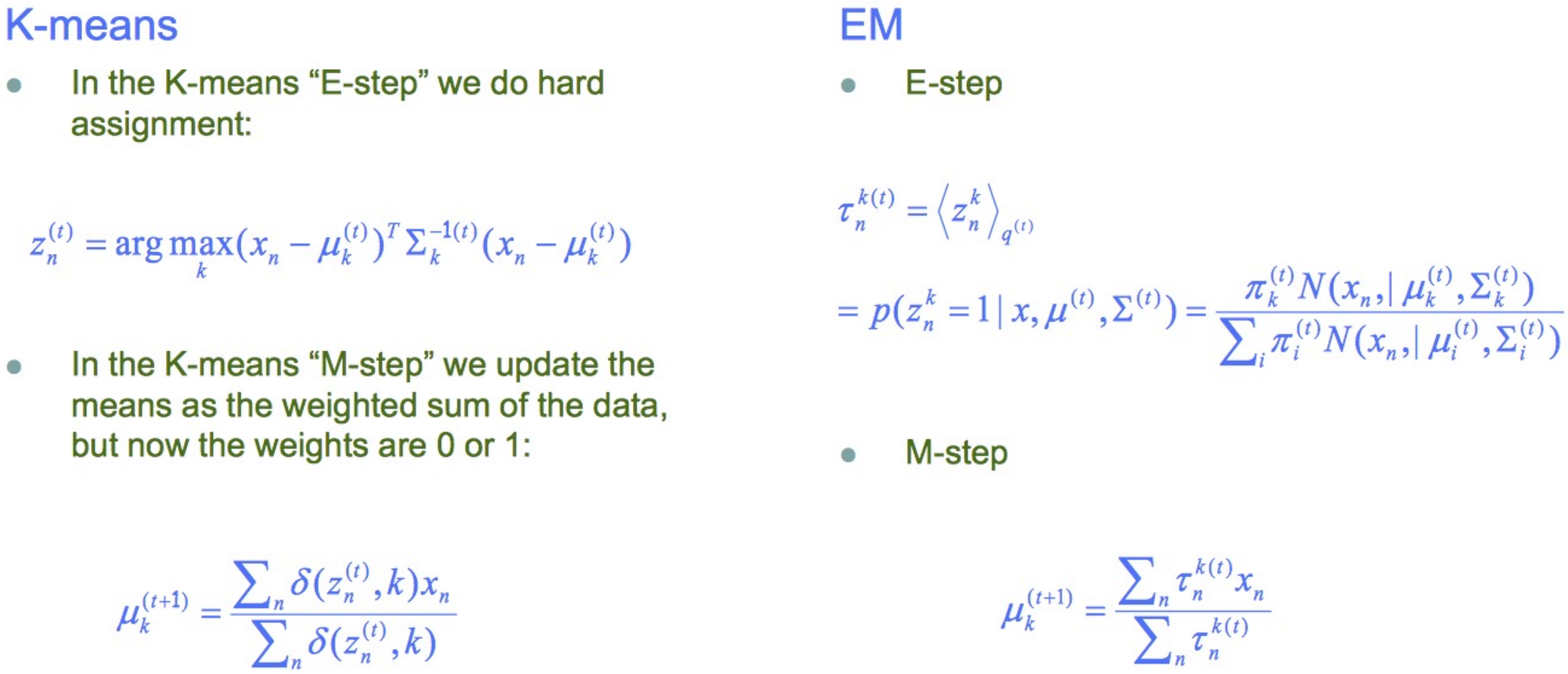
Fact :

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

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

K-Means vs EM

- K-means clustering algorithm is hard-assignment version of EM for mixture of Gaussians



K-means

- In the K-means "E-step" we do hard assignment:

$$z_n^{(t)} = \arg\max_k (x_n - \mu_k^{(t)})^T \Sigma_k^{-1(t)} (x_n - \mu_k^{(t)})$$

- In the K-means "M-step" we update the means as the weighted sum of the data, but now the weights are 0 or 1:

$$\mu_k^{(t+1)} = \frac{\sum_n \delta(z_n^{(t)}, k) x_n}{\sum_n \delta(z_n^{(t)}, k)}$$

EM

- E-step

$$\tau_n^{k(t)} = \left\langle z_n^k \right\rangle_{q^{(t)}}$$

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

- M-step

$$\mu_k^{(t+1)} = \frac{\sum_n \tau_n^{k(t)} x_n}{\sum_n \tau_n^{k(t)}}$$

Why does EM work? (Approximation view)

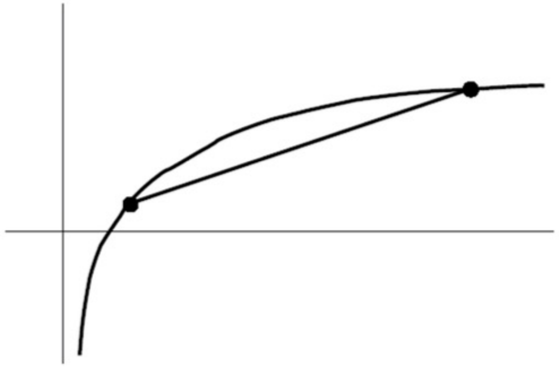
- For a distribution $q(z)$ define the **expected complete log-likelihood**

$$\langle \ell_c(\theta; x, z) \rangle_q \overset{\text{def}}{=} \sum_z q(z \mid x, \theta) \log p(x, z \mid \theta)$$

- The expected complete log-likelihood is a **lower-bound** on the log-likelihood:

$$\begin{aligned} \ell(\theta; x) &= \log p(x \mid \theta) \\ &= \log \sum_z p(x, z \mid \theta) \\ &= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} \\ &\geq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \quad \Rightarrow \quad \ell(\theta; x) \geq \langle \ell_c(\theta; x, z) \rangle_q + H_q \end{aligned}$$

Jensen's inequality



Why does EM work?

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

$$\log p(z|x,\theta) = \log p(\, x,z \mid \theta \,) - \log p(x \mid \theta)$$

$$\mathrm{E}_{z\sim q}[\log p(z|x,\theta)] = E_{z\sim q}[\log p(\, x,z \mid \theta \,)] - \log p(x \mid \theta)$$

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

$$E_{z\sim q}[\log p(z \mid x,\theta)] = E_{z\sim q}[\log q(z \mid x)] - KL(q(\, z \mid x \,) \,||\, p(\, z \mid x,\theta \,))$$

$$E_{z\sim q}[\log p(\, x,z \mid \theta \,)] - \log p(x \mid \theta) = E_{z\sim q}[\log q(z \mid x)] - KL(q(\, z \mid x \,) \,||\, p(\, z \mid x,\theta \,))$$

$$\log p(x \mid \theta) = E_{z\sim q}[\log p(\, x,z \mid \theta \,)] - E_{z\sim q}[\log q(z \mid x)] + KL(q(\, z \mid x \,) \,||\, p(\, z \mid x,\theta \,))$$

$$\log p(x \mid \theta) = E_{z\sim q}[\log p(\, x,z \mid \theta \,)] + H(q) + KL(q(\, z \mid x \,) \,||\, p(\, z \mid x,\theta \,))$$

EM: Let $q_t(\, z \mid x \,) = p(z \mid x,\theta_t)$. Then at convergence:

$$\log p(x \mid \theta) = E_{z\sim q_t}[\log p(\, x,z \mid \theta \,)] + H(q_t) + 0$$

$$Q(\theta',\theta_t) = E_{z\sim p(z|\theta_t)}[\log p(\, x,z \mid \theta' \,)]$$

$$\theta_{t+1} = \mathrm{argmax}_{\theta'} Q(\theta',\theta_t)$$

Foreshadowing Variational Inference

$$\log p(x \mid \theta) = E_{z\sim q}[\log p(x, z \mid \theta)] + H(q) + KL(q(z \mid x) \mid\mid p(z \mid x, \theta))$$

EM: Let $q_t(z \mid x) = p(z \mid x, \theta_t)$.
Max $p(x \mid \theta)$ by iterating:

$$Q(\theta', \theta_t) = E_{z\sim p(z|\theta_t)}[\log p(x, z \mid \theta')]$$
$$\theta_{t+1} = \text{argmax}_{\theta'} Q(\theta', \theta_t)$$

Variational Inference: Let $q(z \mid x)$ be some family that's easier to optimize.

$$\log p(x \mid \theta) \geq \underbrace{E_{z\sim q}[\log p(x, z \mid \theta)] + H(q)}_{\text{“ELBO”: Evidence Lower Bound}}$$

equivalently,

$$\text{ELBO} = \log p(x \mid \theta) - KL(q(z \mid x) || p(z, x\theta))$$

What's the implication of the entropy term H_q?

Another example of EM: Baum-Welch for HMMs

- The **E** step

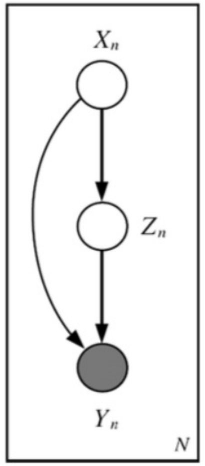
$$\gamma^{i}_{n,t} = \left\langle y^{i}_{n,t} \right\rangle = p(y^{i}_{n,t} = 1 \,|\, \mathbf{x}_n)$$

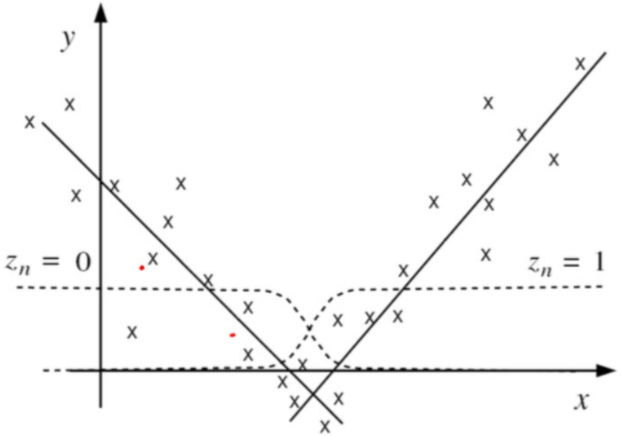
$$\xi^{i,j}_{n,t} = \left\langle y^{i}_{n,t-1} y^{j}_{n,t} \right\rangle = p(y^{i}_{n,t-1} = 1, y^{j}_{n,t} = 1 \,|\, \mathbf{x}_n)$$

- The **M** step ("symbolically" identical to MLE)

$$\pi_i^{ML} = \frac{\sum_n \gamma^{i}_{n,1}}{N} \qquad a_{ij}^{ML} = \frac{\sum_n \sum_{t=2}^{T} \xi^{i,j}_{n,t}}{\sum_n \sum_{t=1}^{T-1} \gamma^{i}_{n,t}} \qquad b_{ik}^{ML} = \frac{\sum_n \sum_{t=1}^{T} \gamma^{i}_{n,t} x^{k}_{n,t}}{\sum_n \sum_{t=1}^{T-1} \gamma^{i}_{n,t}}$$

Another example of EM: Mixture of Linear Experts





- We will model P(Y|X) using different experts, each responsible for different regions of the input space.
  - Latent variable Z chooses expert using softmax $P(z^k = 1|x) = \mathrm{Softmax}(\xi^T x)$
  - Each expert can be a linear regression model: $P(y|x, z^k = 1) = \mathcal{N}(y; \theta_k^T x, \sigma_k^2)$

$$P(z^k = 1|x, y, \theta) = \frac{p(z^k = 1|x) p_k(y|x, \theta_k, \sigma_k^2)}{\sum_j p(z^j = 1|x) p_j(y|x, \theta_j, \sigma_j^2)}$$

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON