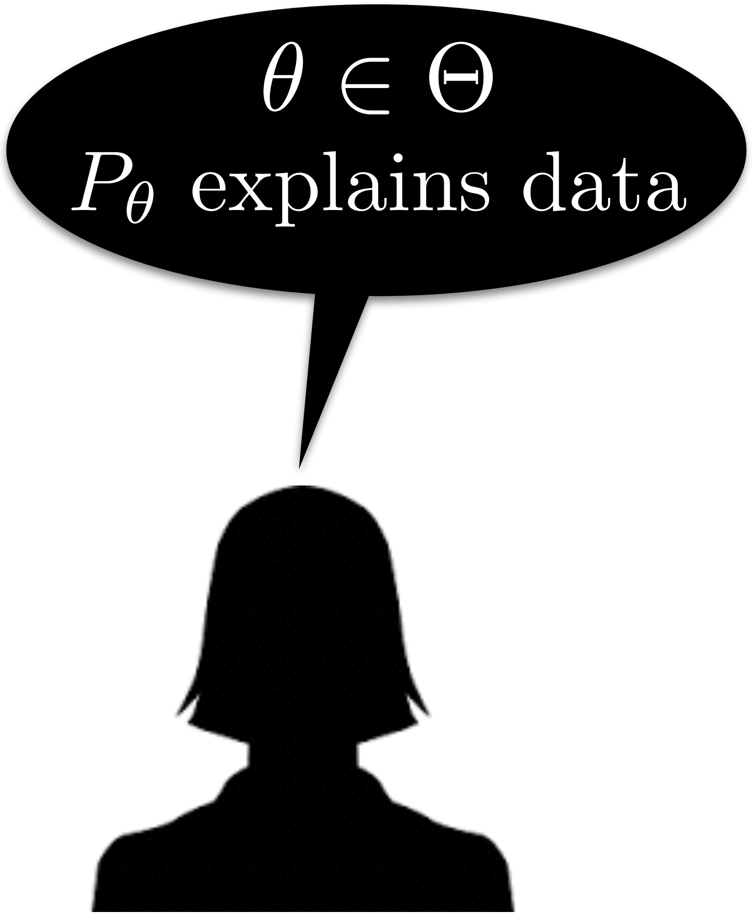
# Machine Learning for Data Science (CS4786)
# Lecture 16

Probabilistic Modeling and EM Algorithm

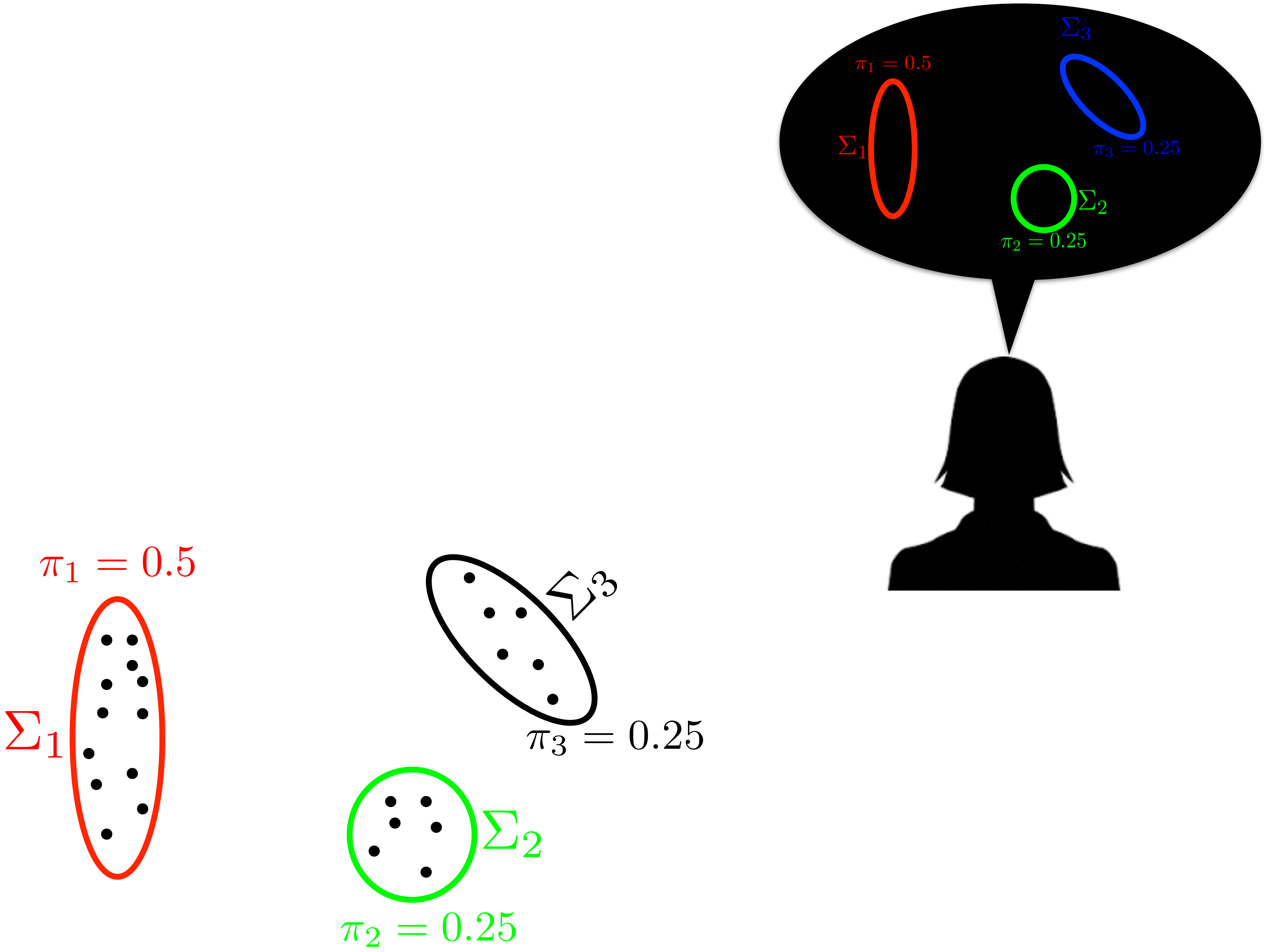
Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2017fa/

# Probabilistic Model




Data: $\mathbf{x}_1 \ldots, \mathbf{x}_n$

# Probabilistic Model



# Examples

- Gaussian Mixture Model
  - Each $\theta$ consists of mixture distribution $\pi = (\pi_1, \ldots, \pi_K)$, means $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K$
  - At time $t$ we generate a new tree as follows:

$$c_t \sim \pi, \qquad x_t \sim N(\mu_{c_t}, \Sigma_{c_t})$$

# Probabilistic Models

- Set of models $\Theta$ consists of parameters s.t. $P_\theta$ for each $\theta \in \Theta$ is a distribution over data.
- Learning: Estimate $\theta^* \in \Theta$ that best models given data

# Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

# Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

Reasoning:

- One of the models in $\Theta$ is the correct one

# Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

Reasoning:

- One of the models in $\Theta$ is the correct one
- Given data we pick the one that best explains the observed data

# Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

Reasoning:

- One of the models in $\Theta$ is the correct one
- Given data we pick the one that best explains the observed data
- Equivalently pick the maximum likelihood estimator,

$$\theta_{MLE} = \mathrm{argmax}_{\theta \in \Theta} \log P_\theta(x_1, \ldots, x_n)$$

# Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

Reasoning:

- One of the models in $\Theta$ is the correct one
- Given data we pick the one that best explains the observed data
- Equivalently pick the maximum likelihood estimator,

$$\theta_{MLE} = \text{argmax}_{\theta \in \Theta} \log P_\theta(x_1, \ldots, x_n)$$

Often referred to as frequentist view

# Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

$$\theta_{MLE} = \mathrm{argmax}_{\theta\in\Theta} \underbrace{\log P_\theta(x_1, \ldots, x_n)}_{\text{Likelihood}}$$

- A priori all models are equally good, data could have been generated by any one of them

# Maximum A Posteriori

Say you had a prior belief about models provided by $P(\theta)$

Pick $\theta \in \Theta$ that is most likely given data

# Maximum A Posteriori

Say you had a prior belief about models provided by $P(\theta)$

Pick $\theta \in \Theta$ that is most likely given data

Reasoning:

- Models are abstractions that capture our belief

# Maximum A Posteriori

Say you had a prior belief about models provided by $P(\theta)$

Pick $\theta \in \Theta$ that is most likely given data

Reasoning:

- Models are abstractions that capture our belief
- We update our belief based on observed data

# Maximum A Posteriori

Say you had a prior belief about models provided by $P(\theta)$

Pick $\theta \in \Theta$ that is most likely given data

Reasoning:

- Models are abstractions that capture our belief
- We update our belief based on observed data
- Given data we pick the model that we believe the most

# Maximum A Posteriori

Say you had a prior belief about models provided by $P(\theta)$

Pick $\theta \in \Theta$ that is most likely given data

Reasoning:

- Models are abstractions that capture our belief
- We update our belief based on observed data
- Given data we pick the model that we believe the most
- Pick $\theta$ that maximizes $\log P(\theta | x_1, \ldots, x_n)$

# Maximum A Posteriori

Say you had a prior belief about models provided by $P(\theta)$

Pick $\theta \in \Theta$ that is most likely given data

Reasoning:

- Models are abstractions that capture our belief
- We update our belief based on observed data
- Given data we pick the model that we believe the most
- Pick $\theta$ that maximizes $\log P(\theta|x_1, \ldots, x_n)$

I want to say : Often referred to as Bayesian view

# Maximum A Posteriori

Say you had a prior belief about models provided by $P(\theta)$

Pick $\theta \in \Theta$ that is most likely given data

Reasoning:

- Models are abstractions that capture our belief
- We update our belief based on observed data
- Given data we pick the model that we believe the most
- Pick $\theta$ that maximizes $\log P(\theta|x_1, \ldots, x_n)$

I want to say : Often referred to as Bayesian view

There are Bayesian and there Bayesians

# Maximum A Posteriori

Pick $\theta \in \Theta$ that is most likely given data

Maximize a posteriori probability of model given data

$$\theta_{MAP} = \text{argmax}_{\theta \in \Theta} P(\theta | x_1, \ldots, x_n)$$

# The Bayesian Choice

Don't pick any $\theta^* \in \Theta$

- Model is simply an abstraction
- We have a prosteriori distribution over models, why pick one $\theta$?

$$P(X|\text{data}) = \sum_{\theta \in \Theta} P(X, \theta|\text{data}) = \sum_{\theta \in \Theta} P(X|\theta) P(\theta|\text{data})$$

# Latent Variables and Expectation Maximization (EM)

# EXAMPLE: GAUSSIAN MIXTURE MODEL

MLE: $\theta = (\mu_1, \ldots, \mu_K), \pi, \Sigma$

$$P_\theta(x_1, \ldots, x_n) = \prod_{t=1}^{n} \left( \sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{(2 * 3.1415)^2 |\Sigma_i|}} \exp\left(-(x_t - \mu_i)^\top \Sigma_i (x_t - \mu_i)\right) \right)$$

Find $\theta$ that maximizes $\log P_\theta(x_1, \ldots, x_n)$

# MLE for GMM

Let us consider the one dimensional case,

$$\log P_\theta(x_{1,\ldots,n}) = \sum_{t=1}^{n} \log\left(\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2 * 3.1415\sigma_i^2}} \exp\left(-(x_t - \mu_i)^2/\sigma_i^2\right)\right)$$

# MLE FOR GMM

Say by some magic you knew cluster assignments, then

$$\log P_\theta((x_t, c_t)_{1,\ldots,n}) = \sum_{t=1}^{n} \log\left(\frac{\pi_{c_t}}{\sqrt{2 * 3.1415\sigma_{c_t}^2}} \exp\left(-\frac{(x_t - \mu_{c_t})^2}{2\sigma_{c_t}^2}\right)\right)$$

$$= \sum_{t=1}^{n} \left(\log(\pi_{c_t}) - \log(2 * 3.1415 * \sigma_{c_t}^2) - \frac{(x_t - \mu_{c_t})^2}{2\sigma_{c_t}^2}\right)$$

# Latent Variables

- We only observe $x_1, \ldots, x_n$, cluster assignments $c_1, \ldots, c_n$ are not observed
- Finding $\theta \in \Theta$ (even for 1-d GMM) that directly maximizes Likelihood or A Posteriori given $x_1, \ldots, x_n$ is hard!
- Given latent variables $c_1, \ldots, c_n$, the problem of maximizing likelihood (or a posteriori) became easy

Can we use latent variables to device an algorithm?

# Expectation Maximization Algorithm

- For demonstration we shall consider the problem of finding MLE (MAP version is very similar)

# EXPECTATION MAXIMIZATION ALGORITHM

- For demonstration we shall consider the problem of finding MLE (MAP version is very similar)
- Initialize $\theta^{(0)}$ arbitrarily, repeat unit convergence:

(E step) For every $t$, define distribution $Q_t$ over the latent variable $c_t$ as:

$$Q_t^{(i)}(c_t) = P(c_t | x_t, \theta^{(i-1)})$$

(M step)

$$\theta^{(i)} = \mathrm{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t} Q_t^{(i)}(c_t) \log P(x_t, c_t | \theta)$$

# Example: EM for GMM

- E step: For every $k \in [K]$,

$$Q_t^{(i)}(c_t = k) = P\left(c_t = k | x_t, \theta^{(i-1)}\right) = P\left(x_t | c_t = k, \theta^{(i-1)}\right) \times P\left(c_t = k | \theta^{(i-1)}\right)$$

$$\propto \underbrace{\phi\left(x_t; \mu_k^{(i-1)}, \Sigma_k^{(i-1)}\right)}_{\text{gaussian p.d.f.}} \times \pi_k^{(i-1)}$$

# Example: EM for GMM

- E step: For every $k \in [K]$,

$$Q_t^{(i)}(c_t = k) = P\left(c_t = k | x_t, \theta^{(i-1)}\right) = P\left(x_t | c_t = k, \theta^{(i-1)}\right) \times P\left(c_t = k | \theta^{(i-1)}\right)$$
$$\propto \underbrace{\phi\left(x_t; \mu_k^{(i-1)}, \Sigma_k^{(i-1)}\right)}_{\text{gaussian p.d.f.}} \times \pi_k^{(i-1)}$$

- M step: Given $Q_1, \ldots, Q_n$, we need to find

$$\begin{aligned}
\theta^{(i)} &= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log P(x_t, c_t = k | \theta) \\
&= \underset{\theta}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \left(\log P(x_t | c_t = k, \theta) + \log P(c_t = k | \theta)\right) \\
&= \underset{\pi, \mu_{1,\ldots,K}, \Sigma_{1,\ldots,K}}{\operatorname{argmax}} \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(k) \left(\log \phi\left(x_t; \mu_k, \Sigma_k\right) + \log \pi_k\right)
\end{aligned}$$

# Example: EM for GMM

For every $k \in [K]$, the maximization step yields,

$$\mu_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) x_t}{\sum_{t=1}^n Q_t(k)}, \quad \Sigma_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k) \left(x_t - \mu_k^{(i)}\right)\left(x_t - \mu_k^{(i)}\right)^\top}{\sum_{t=1}^n Q_t(k)}$$

$$\pi_k^{(i)} = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

# WHY SHOULD EM WORK?

A very high level view:

- Performing E-step will never decrease log-likelihood (or log a posteriori)

# WHY SHOULD EM WORK?

A very high level view:

- Performing E-step will never decrease log-likelihood (or log a posteriori)
- Performing M-step will never decrease log-likelihood (or log a posteriori)

# Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$ :

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n)$$

# Why should EM work?

Steps to show that $\log \mathrm{Lik}(\theta^{(i)}) \geq \log \mathrm{Lik}(\theta^{(i-1)})$ :

$$\log P_{\theta^{(i)}}(x_1, \ldots, x_n) \geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i)}(c_t) \log \left( \frac{P_{\theta^{(i)}}(x_t, c_t)}{Q^{(i)}(c_t)} \right)$$

# Why should EM work?

- Likelihood never decreases
- So whenever we converge we converge to a local optima
- However problem is non-convex and can have many local optimal
- In general no guarantee on rate of convergence
- In practice, do multiple random initializations and pick the best one!

# EM IN GENERAL

- There was nothing special about GMM or clustering problems
- EM can be used as a general strategy for any problem with latent/missing/unobserved variables
- The MAP version only involves an extra prior term over $\theta$ multiplied to the likelihood
- In general probabilistic models with observed and latent variables can be represented succinctly as graphical models.
  **Next time ...**