Expectation Maximization

Avinava Dubey

Preliminary [2]

- Convex Functions:-

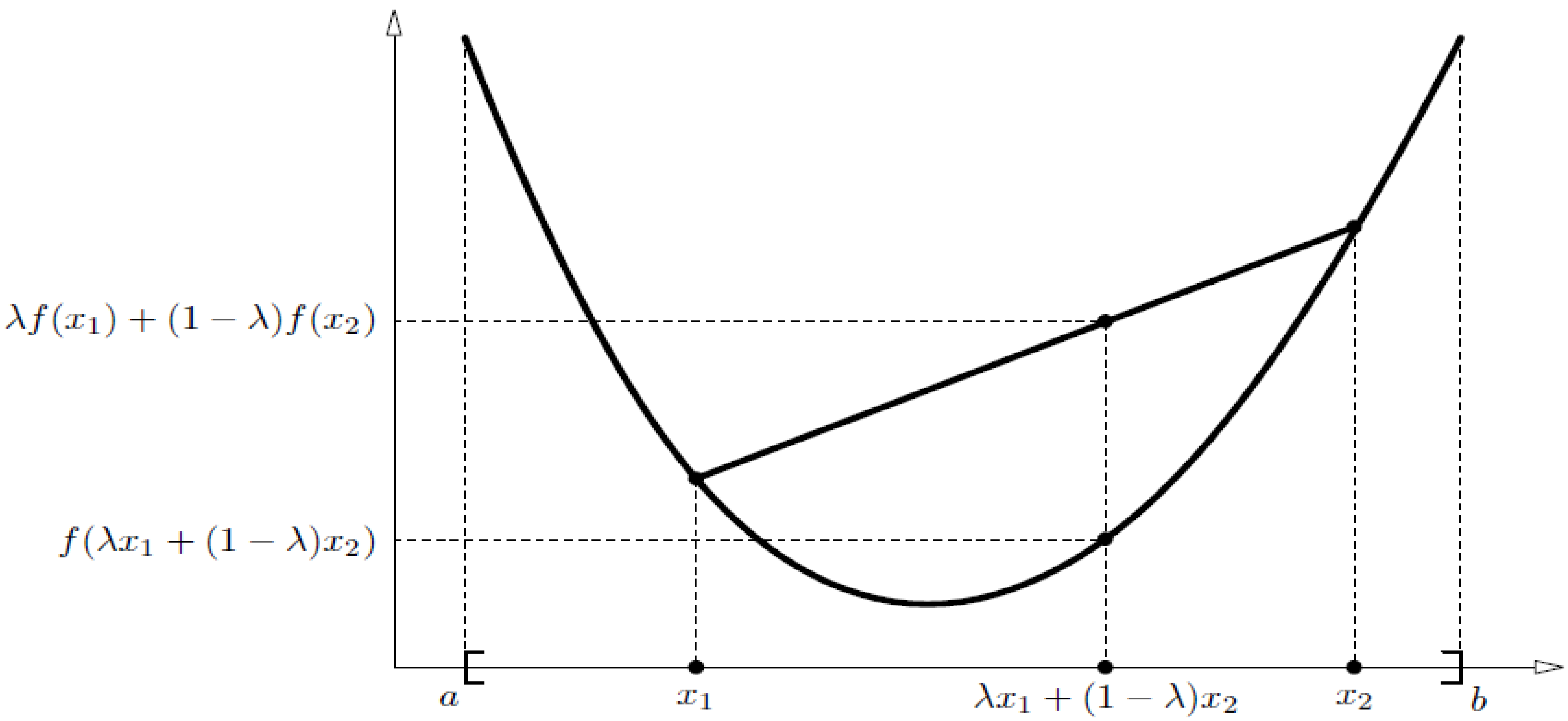
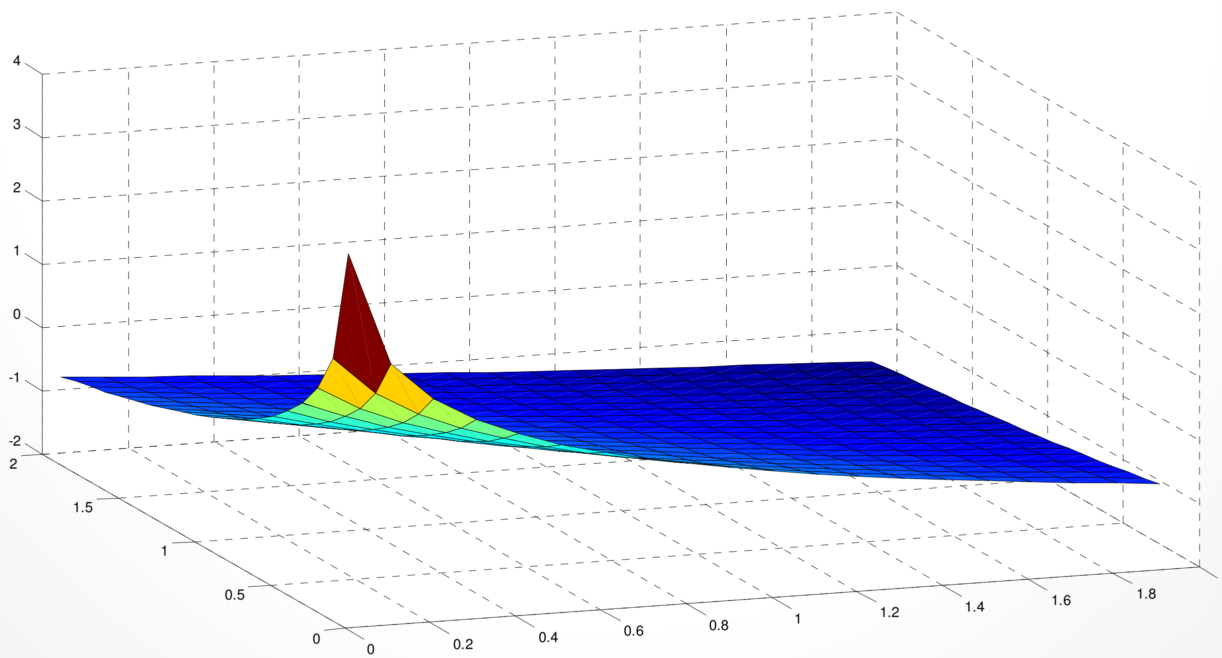


Figure 1: f is *convex* on $[a, b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a, b], \quad \lambda \in [0, 1]$.

Convex Fn properties

- f is concave if –f is convex
- If f(x) is twice differentiable and f''(x) >= 0 then f(x) is convex
- -ln(x) is convex on the interval (0,inf)



Contd

- Jensen's inequality

Theorem 2 (Jensen's inequality) *Let f be a convex function defined on an interval I. If $x_1, x_2, \ldots, x_n \in I$ and $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$,*

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i\right) \\ &= f\left(\lambda_{n+1} x_{n+1} + (1 - \lambda_{n+1}) \frac{1}{1 - \lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i\right) \end{aligned}$$

Contd

$$
\begin{aligned}
&\leq \lambda_{n+1} f\left(x_{n+1}\right)+\left(1-\lambda_{n+1}\right) f\left(\frac{1}{1-\lambda_{n+1}} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&= \lambda_{n+1} f\left(x_{n+1}\right)+\left(1-\lambda_{n+1}\right) f\left(\sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right)+\left(1-\lambda_{n+1}\right) \sum_{i=1}^{n} \frac{\lambda_i}{1-\lambda_{n+1}} f\left(x_i\right) \\
&= \lambda_{n+1} f\left(x_{n+1}\right)+\sum_{i=1}^{n} \lambda_i f\left(x_i\right) \\
&= \sum_{i=1}^{n+1} \lambda_i f\left(x_i\right)
\end{aligned}
$$

- Note that since –ln(x) is convex we have

$$
\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i).
$$

Three coin Example [1]

- We observe a series of coin tosses generated in the following way:
- A person has three coins.
  - Coin 0: probability of Head is λ
  - Coin 1: probability of Head p
  - Coin 2: probability of Head q
- Consider the following coin-tossing scenarios:

Estimation Problems

- Scenario I: Toss one of the coins six times.

  Observing HHHTHT

  Which coin is more likely to produce this sequence? Suppose we know the probability of H for each coin.
- Scenario II: Toss coin 0. If Head – toss coin 1; o/w – toss coin 2

  Observing the sequence HHHHT, THTHT, HHHHT, HHTTH

  produced by Coin 0 , Coin1 and Coin2

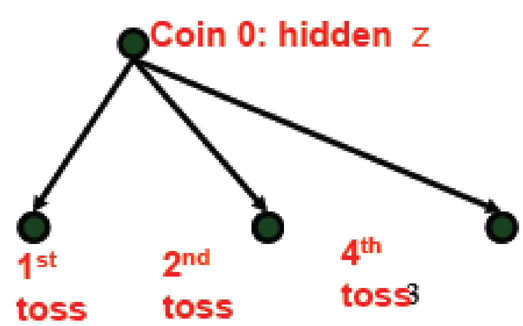
  Estimate most likely values for p, q, λ (the probability of H in each coin)
- Scenario III: Toss coin 0. If Head – toss coin 1; o/w – toss coin 2

  Observing the sequence HHHT, HTHT, HHHT, HTTH

  produced by Coin 1 and/or Coin 2

  Estimate most likely values for λ, p, q.

The label of the first toss (z) is hidden, we want to estimate the most likely hypothesis θ = (λ,p,q) under hidden z.



Key Intuition

- If we knew which of the data points (HHHT), (HTHT), (HTTH) came from Coin1 and which from Coin2, there was no problem.

- Recall that the "simple" estimation is the ML estimation:
- Assume that you toss a (p,1-p) coin m times and get k Heads m-k Tails.

$$\log[P(D|p)] = \log [\, p^k (1-p)^{m-k} \,] = k \log p + (m-k) \log (1-p)$$

- To maximize, set the derivative w.r.t. p equal to 0:

$$d/dp \, \{\log P(D|p)\} = k/p - (m-k)/(1-p) = 0$$

- Solving this for p, gives: $p=k/m$

Key Intuition

- Since we do not know which of the data points (HHHT), (HTHT), (HTTH) came from Coin1 and which from Coin2, we use an iterative approach for estimating (λ,p,q).

Derivation [2]

- Log likelihood:- $L(\theta) = \ln \mathcal{P}(\mathbf{X}|\theta).$

$$\mathcal{P}(\mathbf{X}|\theta) = \sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta).$$

- We wish to find θ iterative such that $L(\theta) > L(\theta_n)$

Where θ_n is previous iterations θ value.

- The difference can be written as

$$L(\theta) - L(\theta_n) = \ln \mathcal{P}(\mathbf{X}|\theta) - \ln \mathcal{P}(\mathbf{X}|\theta_n).$$

$$= \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n).$$

Derivation

- Note that $\ln \sum_{i=1}^{n} \lambda_i x_i \geq \sum_{i=1}^{n} \lambda_i \ln(x_i)$

$$
\begin{aligned}
L(\theta) - L(\theta_n) &= \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \cdot \frac{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \ln \left(\sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&\geq \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right) - \ln \mathcal{P}(\mathbf{X}|\theta_n) \\
&= \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \left(\frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)\mathcal{P}(\mathbf{X}|\theta_n)} \right) \\
&\triangleq \Delta(\theta|\theta_n).
\end{aligned}
$$

Derivation

- So far $L(\theta) \geq \underbrace{L(\theta_n) + \Delta(\theta|\theta_n)}$

$$l(\theta|\theta_n) \quad \Rightarrow \quad L(\theta) \geq l(\theta|\theta_n)$$

$$\begin{aligned}
l(\theta_n|\theta_n) &= L(\theta_n) + \Delta(\theta_n|\theta_n) \\
&= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n) \ln \frac{\mathcal{P}(\mathbf{X}|\mathbf{z}, \theta_n)\mathcal{P}(\mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n)\mathcal{P}(\mathbf{X}|\theta_n)} \\
&= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n) \ln \frac{\mathcal{P}(\mathbf{X}, \mathbf{z}|\theta_n)}{\mathcal{P}(\mathbf{X}, \mathbf{z}|\theta_n)} \\
&= L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X}, \theta_n) \ln 1 \\
&= L(\theta_n),
\end{aligned}$$

Intuition

- In EM we optimize $l(\theta|\theta_n)$

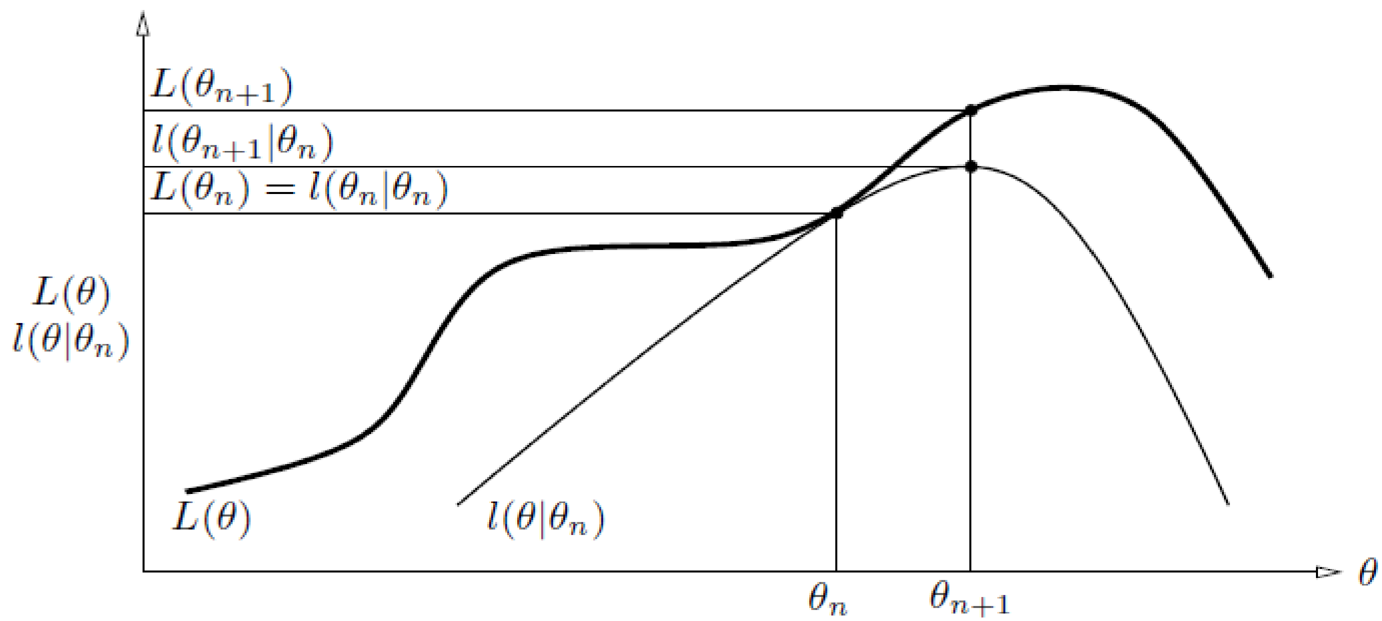


Figure 2: Graphical interpretation of a single iteration of the EM algorithm: The function $l(\theta|\theta_n)$ is bounded above by the likelihood function $L(\theta)$. The functions are equal at $\theta = \theta_n$. The EM algorithm chooses θ_{n+1} as the value of θ for which $l(\theta|\theta_n)$ is a maximum. Since $L(\theta) \geq l(\theta|\theta_n)$ increasing $l(\theta|\theta_n)$ ensures that the value of the likelihood function $L(\theta)$ is increased at each step.

-

Derivation

- Formally we have

$$
\begin{aligned}
\theta_{n+1} &= \arg\max_{\theta} \{ l(\theta|\theta_n) \} \\
&= \arg\max_{\theta} \left\{ L(\theta_n) + \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta)}{\mathcal{P}(\mathbf{X}|\theta_n)\mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n)} \right\} \\
& \quad \text{Now drop terms which are constant w.r.t. } \theta \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X}|\mathbf{z},\theta)\mathcal{P}(\mathbf{z}|\theta) \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \frac{\mathcal{P}(\mathbf{X},\mathbf{z},\theta)}{\mathcal{P}(\mathbf{z},\theta)} \frac{\mathcal{P}(\mathbf{z},\theta)}{\mathcal{P}(\theta)} \right\} \\
&= \arg\max_{\theta} \left\{ \sum_{\mathbf{z}} \mathcal{P}(\mathbf{z}|\mathbf{X},\theta_n) \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \right\} \\
&= \arg\max_{\theta} \left\{ \mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_n} \{ \ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta) \} \right\}
\end{aligned}
$$

Algorithm

- E-step Find the conditional expectation,

$$\mathrm{E}_{\mathbf{Z}|\mathbf{X},\theta_n}\{\ln \mathcal{P}(\mathbf{X},\mathbf{z}|\theta)\}$$

- Maximize wrt θ

Convergence

- Intuition
  - At each iteration the objective is non-decreasing
  - The log-likelihood is bounded above

- It should converge but at a local minima

Three Coin Estimation Problems

- Scenario I: Toss one of the coins six times.

  Observing HHHTHT

  Which coin is more likely to produce this sequence? Suppose we know the probability of H for each coin.
- Scenario II: Toss coin 0. If Head – toss coin 1; o/w – toss coin 2

  Observing the sequence HHHHT, THTHT, HHHHT, HHTTH

  produced by Coin 0 , Coin1 and Coin2

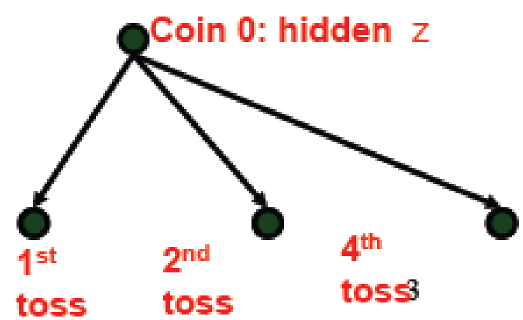
  Estimate most likely values for p, q, λ (the probability of H in each coin)
- Scenario III: Toss coin 0. If Head – toss coin 1; o/w – toss coin 2

  Observing the sequence HHHT, HTHT, HHHT, HTTH

  produced by Coin 1 and/or Coin 2

  Estimate most likely values for λ, p, q.

The label of the first toss (z) is hidden, we want to estimate the most likely hypothesis θ = (λ,p,q) under hidden z.



EM

$$P(D^i,1|\lambda,p,q) = \lambda p^{h_i}(1-p)^{m-h_i}$$

$$P(D^i,0|\lambda,p,q) = (1-\lambda)q^{h_i}(1-q)^{m-h_i}$$

z_i is an indicator variable

$$P(D^i,z_i|\lambda,p,q) = [\lambda p^{h_i}(1-p)^{m-h_i}]^{z_i}[(1-\lambda)q^{h_i}(1-q)^{m-h_i}]^{(1-z_i)}$$

$$= \lambda^{z_i} p^{z_i h_i}(1-p)^{z_i(m-h_i)}(1-\lambda)^{1-z_i} q^{(1-z_i)h_i}(1-q)^{(1-z_i)(m-h_i)}$$

$$\log P(D^i,z_i|\lambda,p,q) = z_i\log\lambda + z_i h_i \log p + z_i(m-h_i)\log(1-p) + (1-z_i)\log(1-\lambda) + (1-z_i)h_i\log q + (1-z_i)(m-h_i)\log(1-q)$$

$$P(D,z|\lambda,p,q) = \prod_i P(D^i,z_i|,p,q)$$

$$\log P(D,z|\lambda,p,q) = \sum_i \log P(D^i,z_i|\lambda,p,q)$$

$$E[X+Y] = E[X] + E[Y]$$

$$E[z_i] = P_i$$

$$E[\log P(D,z|\lambda,p,q)] = E[\sum_i \log P(D^i,z_i|\lambda,p,q)] = \sum_i E[\log P(D^i,z_i|\lambda,p,q)]$$

$$= \sum_i E[z_i\log\lambda + z_i h_i\log p + z_i(m-h_i)\log(1-p) + (1-z_i)\log(1-\lambda) + (1-z_i)h_i\log q + (1-z_i)(m-h_i)\log(1-q)]$$

$$= \sum_i P_i\log\lambda + P_i h_i\log p + P_i(m-h_i)\log(1-p) + (1-P_i)\log(1-\lambda) + (1-P_i)h_i\log q + (1-P_i)(m-h_i)\log(1-q)]$$

EM

- Suppose $(\tilde{\lambda},\tilde{p},\tilde{q})$ is the current estimate of parameters.
- What is the probability P(z) given $(\tilde{\lambda},\tilde{p},\tilde{q})$ and D?
- Suppose there were m coin tosses and h heads in D^i. Given the current parameters,

$$P_i = P(z_i = 1 \mid D^i) = P(\text{Coin1} \mid D^i) = \frac{P(D^i \mid \text{Coin1})\, P(\text{Coin1})}{P(D^i)} =$$

$$= \frac{\tilde{\lambda}\tilde{p}^{h_i}(1-\tilde{p})^{m-h_i}}{\tilde{\lambda}\tilde{p}^{h_i}(1-\tilde{p})^{m-h_i} + (1-\tilde{\lambda})\tilde{q}^{h}(1-\tilde{q})^{m-h_i}}$$

$$E[Y] = \sum_{y_i} y_i P(Y = y_i)$$

$$E[z_i] = 1 \times P(D_i \text{ was obtained from Coin 1}) + 0 \times P(D_i \text{ was obtained from Coin 2}) = P_i$$

EM

$$\frac{dE}{d\lambda} = \sum \left(\frac{P_i}{\lambda} - \frac{1-P_i}{1-\lambda} \right) = 0 \qquad \lambda = \frac{\sum P_i}{n}$$

$$\frac{dE}{dp} = \sum P_i \left(\frac{h_i}{p} - \frac{m-h_i}{1-p} \right) = 0 \quad \Rightarrow \quad p = \frac{\sum P_i \frac{h_i}{m}}{\sum P_i}$$

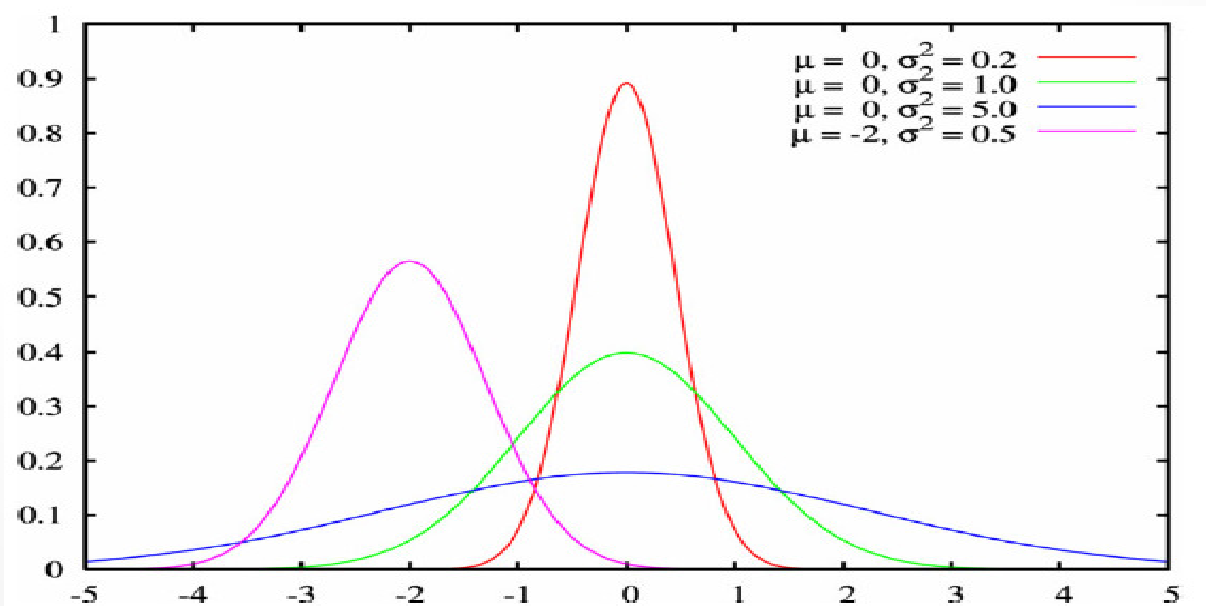
$$\frac{dE}{dq} = \sum_{i=1}^{n} (1-P_i) \left(\frac{h_i}{q} - \frac{m-h_i}{1-q} \right) = 0 \quad \Rightarrow \quad q = \frac{\sum (1-P_i) \frac{h_i}{m}}{\sum (1-P_i)}$$

Example 2 GMM [3]

- Gaussian Distribution

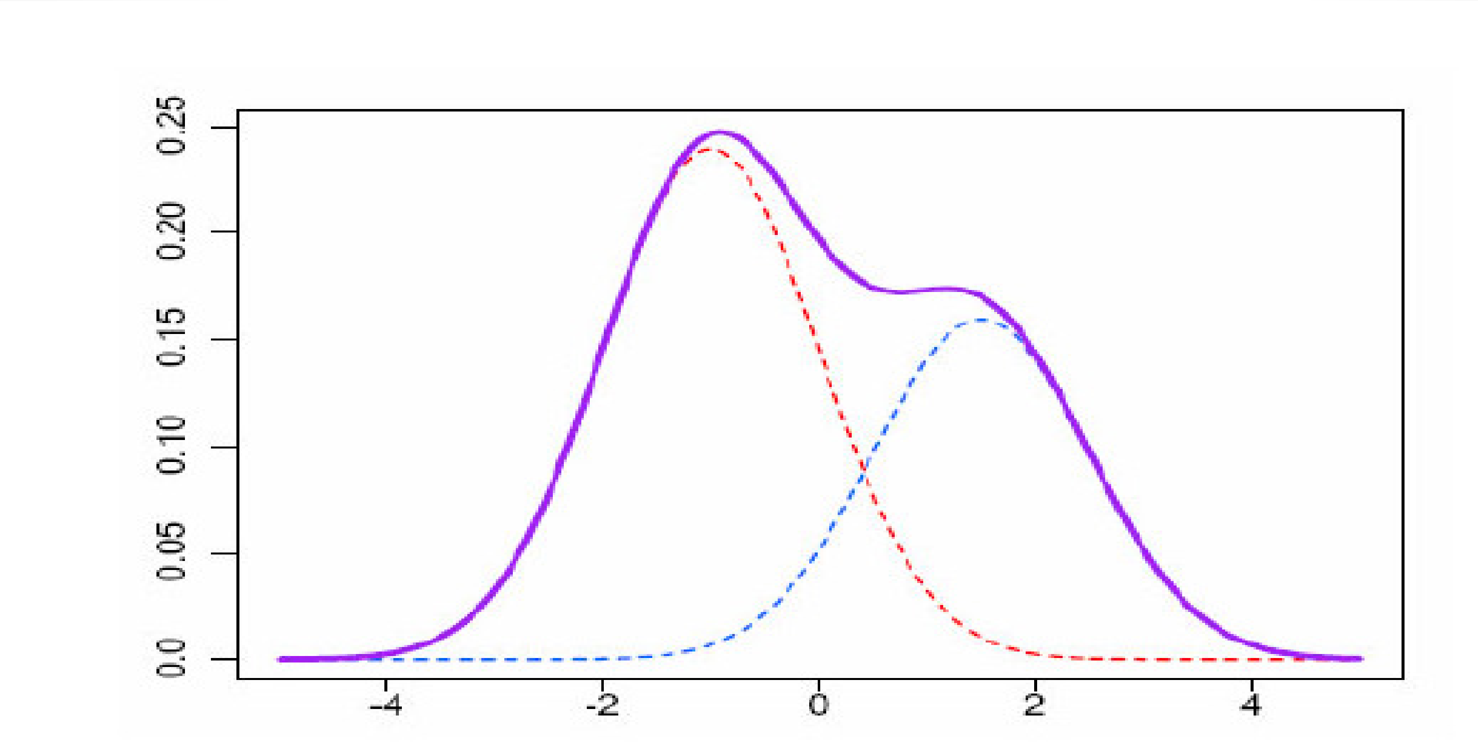
$$G_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Mixture of Gaussian can model arbitrary distributions

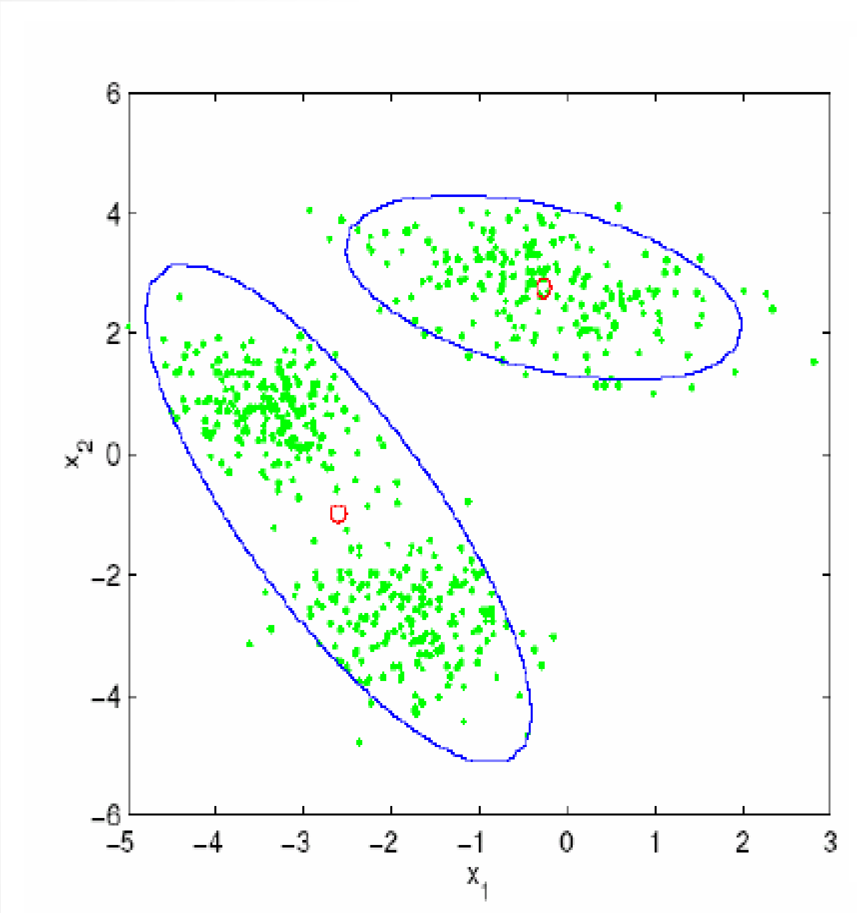


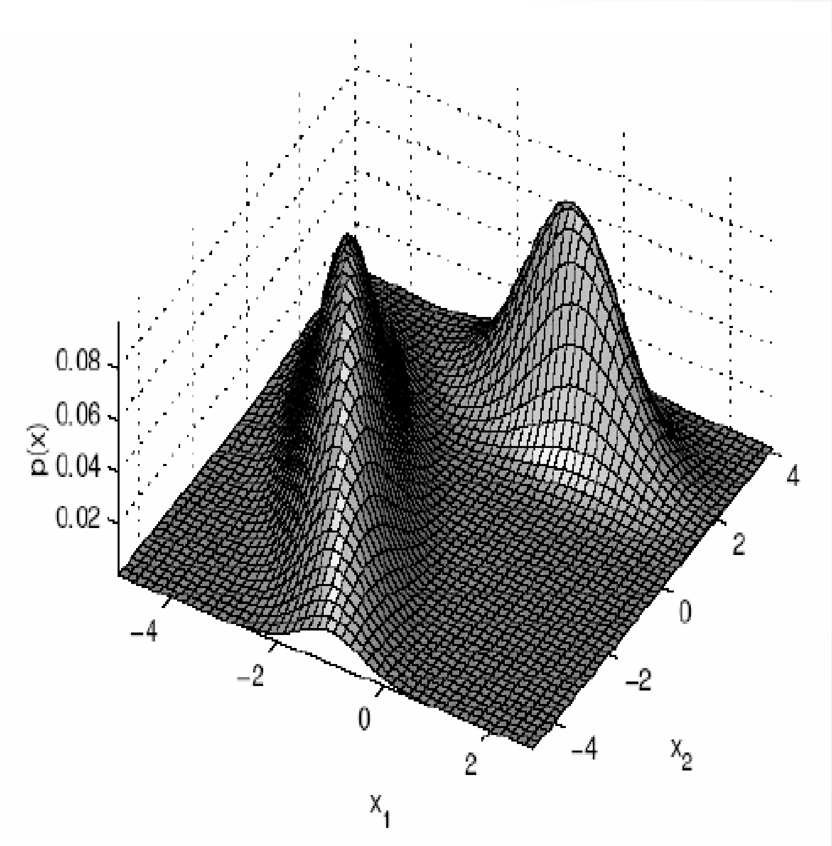
GMM

- An example of two mixtures:-

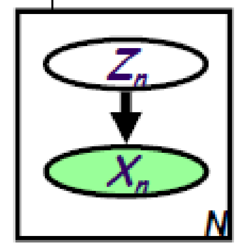


GMM





EM algorithm for GMM



- E.g., A mixture of K Gaussians:

  - Z is a latent class indicator vector

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

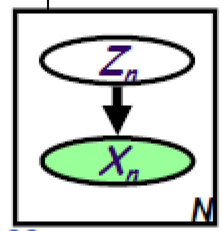
  - X is a conditional Gaussian variable with a class-specific mean/covariance

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

  - The likelihood of a sample:

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

How is EM derived?



- A mixture of K Gaussians:
  - Z is a latent class indicator vector

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

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

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

  - The likelihood of a sample:

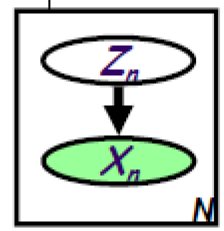
$$p(x_n|\mu,\Sigma) = \sum_k p(z_n^k = 1 \mid \pi) p(x, \mid z_n^k = 1, \mu, \Sigma)$$
$$= \sum_{z_n} \prod_k \left((\pi_k)^{z_n^k} N(x_n : \mu_k, \Sigma_k)^{z_n^k}\right) = \sum_k \pi_k N(x, \mid \mu_k, \Sigma_k)$$

- The "complete" likelihood

$$p(x_n, z_n^k = 1|\mu,\Sigma) = p(z_n^k = 1 \mid \pi) p(x, \mid z_n^k = 1, \mu, \Sigma) = \pi_k N(x, \mid \mu_k, \Sigma_k)$$
$$p(x_n, z_n|\mu,\Sigma) = \prod_k \left[\pi_k N(x, \mid \mu_k, \Sigma_k)\right]^{z_n^k}$$

But this is itself a random variable! Not good as objective function

How is EM derived?



- The complete log likelihood:

$$\ell(\boldsymbol{\theta}; D) = \log \prod_n p(z_n, x_n) = \log \prod_n p(z_n \mid \pi) p(x_n \mid z_n, \mu, \sigma)$$

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

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

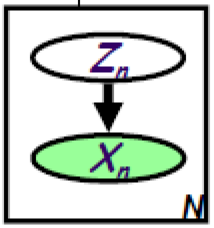
- The expected complete log likelihood

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

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

E-step

- We maximize $\langle l_c(\theta) \rangle$ iteratively using the following iterative procedure:

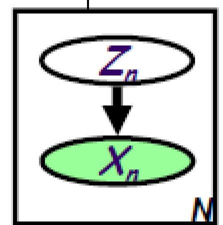


  – Expectation step: computing the expected value of the sufficient statistics of the hidden variables (i.e., z) given current est. of the parameters (i.e., π and μ).

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

  - Here we are essentially doing **inference**

M-step



- We maximize $\langle l_c(\theta)\rangle$ iteratively using the following iterative procedure:
  - Maximization step: compute the parameters under current results of the expected value of the hidden variables

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

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

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

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

Fact:

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

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

- This is isomorphic to MLE except that the variables that are hidden are replaced by their expectations (in general they will by replaced by their corresponding "sufficient statistics")

Example 3 HMM

- Observation space
  - Alphabetic set: $C = \{c_1, c_2, \cdots, c_K\}$
  - Euclidean space: $\mathbb{R}^d$
- Index set of hidden states

  $I = \{1, 2, \cdots, M\}$
- Transition probabilities between any two states

  $p(y_t^j = 1 \mid y_{t-1}^i = 1) = a_{i,j}$,

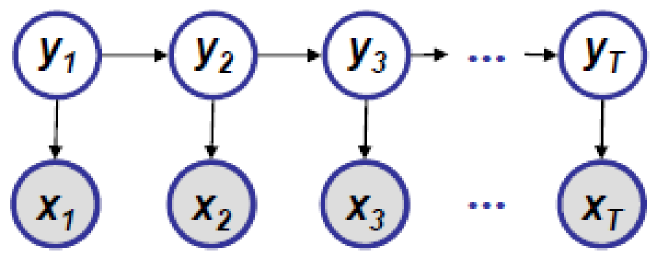
  or $p(y_t \mid y_{t-1}^i = 1) \sim \text{Multinomial}(a_{i,1}, a_{i,2}, \ldots, a_{i,M}), \forall i \in I.$
- Start probabilities

  $p(y_1) \sim \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_M).$
- Emission probabilities associated with each state

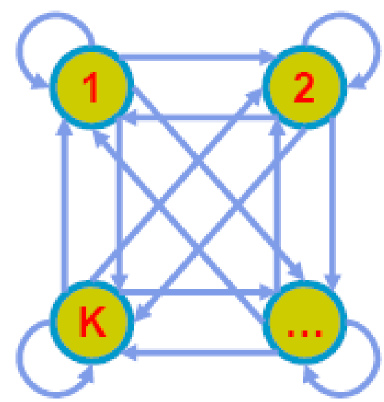
  $p(x_t \mid y_t^i = 1) \sim \text{Multinomial}(b_{i,1}, b_{i,2}, \ldots, b_{i,K}), \forall i \in I.$

  or in general:

  $p(x_t \mid y_t^i = 1) \sim f(\cdot \mid \theta_i), \forall i \in I.$



Graphical model



State automata

The Baum Welch algorithm

- The complete log likelihood

$$\ell_c(\theta; \mathbf{x}, \mathbf{y}) = \log p(\mathbf{x}, \mathbf{y}) = \log \prod_n \left(p(y_{n,1}) \prod_{t=2}^{T} p(y_{n,t} \mid y_{n,t-1}) \prod_{t=1}^{T} p(x_{n,t} \mid x_{n,t}) \right)$$

- The expected complete log likelihood

$$\langle \ell_c(\theta; \mathbf{x}, \mathbf{y}) \rangle = \sum_n \left(\langle y_{n,1}^i \rangle_{p(y_{n,1}|\mathbf{x}_n)} \log \pi_i \right) + \sum_n \sum_{t=2}^{T} \left(\langle y_{n,t-1}^i y_{n,t}^j \rangle_{p(y_{n,t-1}, y_{n,t}|\mathbf{x}_n)} \log a_{i,j} \right) + \sum_n \sum_{t=1}^{T} \left(x_{n,t}^k \langle y_{n,t}^i \rangle_{p(y_{n,t}|\mathbf{x}_n)} \log b_{i,k} \right)$$

- EM
  - The **E** step

$$\gamma_{n,t}^i = \langle y_{n,t}^i \rangle = p(y_{n,t}^i = 1 \mid \mathbf{x}_n)$$

$$\xi_{n,t}^{i,j} = \langle y_{n,t-1}^i y_{n,t}^j \rangle = p(y_{n,t-1}^i = 1, y_{n,t}^j = 1 \mid \mathbf{x}_n)$$

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

$$\pi_i^{ML} = \frac{\sum_n \gamma_{n,1}^i}{N} \qquad a_{ij}^{ML} = \frac{\sum_n \sum_{t=2}^{T} \xi_{n,t}^{i,j}}{\sum_n \sum_{t=1}^{T-1} \gamma_{n,t}^i} \qquad b_{ik}^{ML} = \frac{\sum_n \sum_{t=1}^{T} \gamma_{n,t}^i x_{n,t}^k}{\sum_n \sum_{t=1}^{T-1} \gamma_{n,t}^i}$$

EM summary

- Nice method to get to local optimum solution
- Guaranteed to converge, never decrease likelihood.
- Some problem may require time consuming inference.

- [1] http://www.cs.ucsb.edu/~ambuj/Courses/bioinformatics/EM.pdf
- [2] http://www.seanborman.com/publications/EM_algorithm.pdf
- [3] Class lecture notes