Clustering: Models and Algorithms

Shikui Tu

Shanghai Jiao Tong University

2021-03-23

Outline

- **Gaussian Mixture Models (GMM)**
  - **From generation process perspective**

- Expectation-Maximization (EM) for maximum likelihood
  - An alternative view to verify its properties

- A brief history of EM

From distance to probability

distance

likely

$$\| x-\mu \|^2 \longrightarrow \exp\{-\lambda \| x-\mu \|^2\}$$

“The closer, the more likely.”

$$\Big\downarrow$$

Sum or integral to be one

Probability

$$\frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

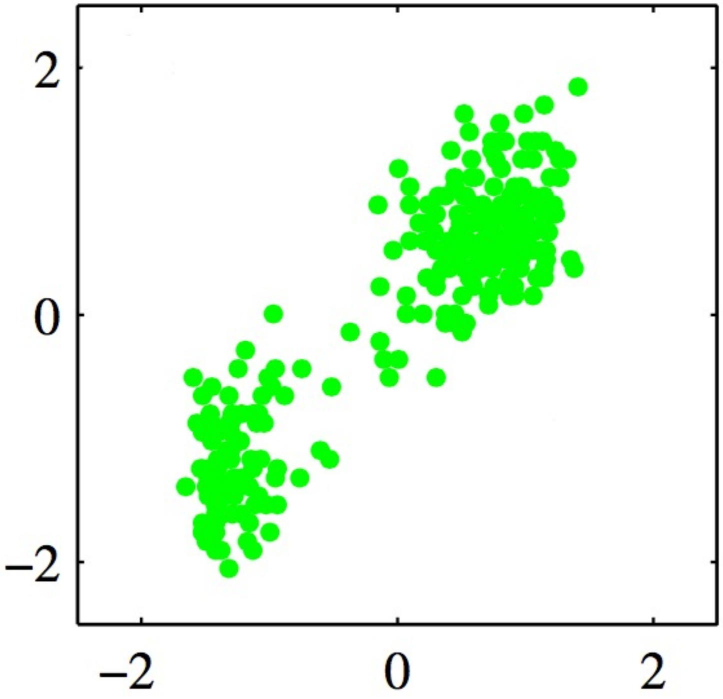
It is more powerful to consider everything in probability framework!

Gaussian distribution with the Mahalanobis distance

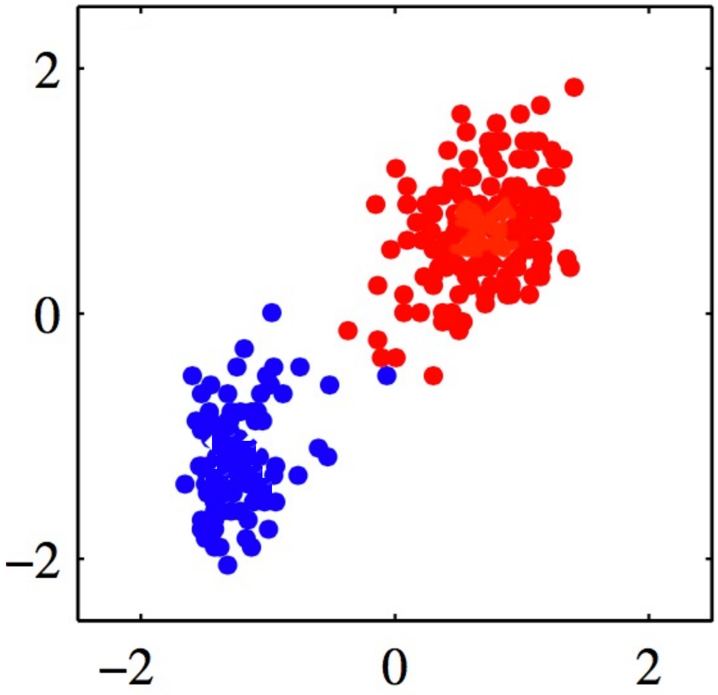
$$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

Review the clustering problem again

We have the following data:

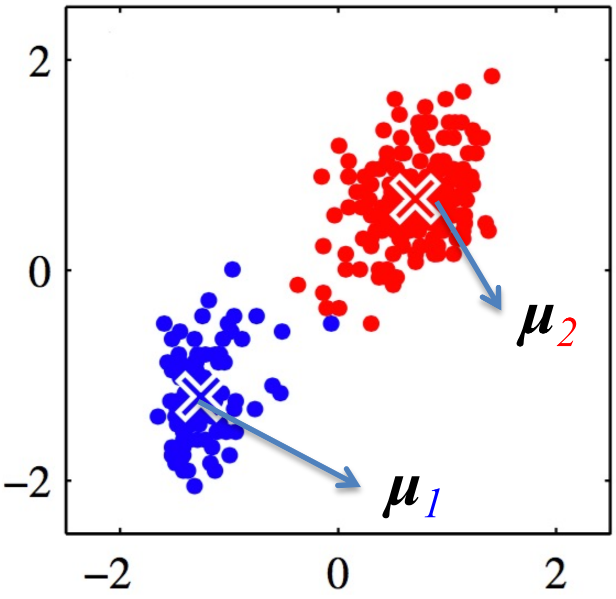


We want to cluster the data into two clusters (red and blue)

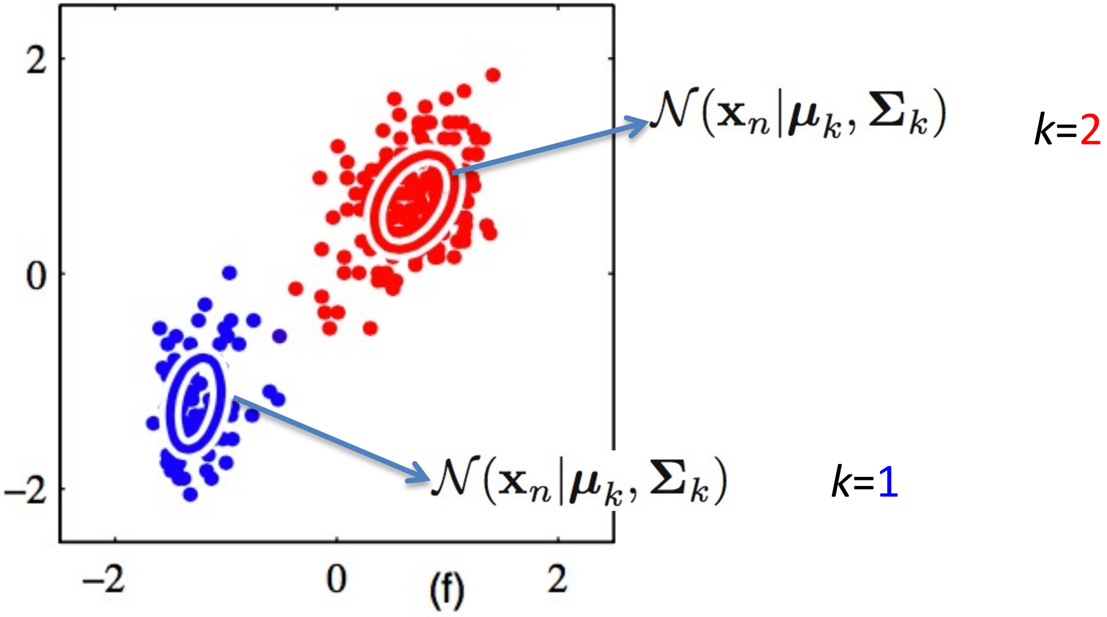


Instead if using $\{\boldsymbol{\mu}_{1,}\ \boldsymbol{\mu}_{2}\}$, each cluster is represented as a Gaussian distribution

K-means

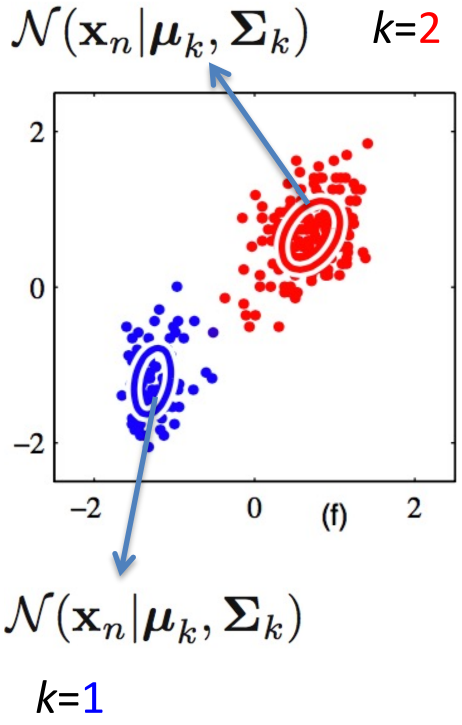


Gaussian Mixture Model (GMM)



$$\mathcal{N}(\mathbf{x}\ |\boldsymbol{\mu}\ ,\boldsymbol{\Sigma}\ ) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Introduce a latent variable



$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ *k*=2

$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=1

We use $z_k = 1$ to indicate a point $\mathbf{x}$ belongs to cluster k

$\mathbf{z} = (z_1, \ldots, z_K)$ $\quad z_k \in \{0, 1\}$ $\quad \sum_k z_k = 1$

A mixing weight for each cluster:

$$\boxed{p(z_k = 1) = \pi_k} \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

prior probability of point belonging to a cluster

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

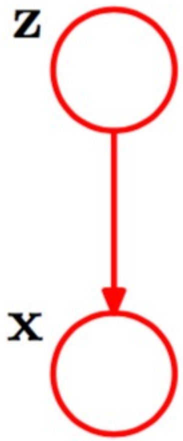
Assume the points in the same cluster follow a **Gaussian distribution**

$$\boxed{p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

Gaussian Mixture Model (GMM)

Generative process

- Randomly sample a **z** from a categorical distribution $[\pi_1, \ldots, \pi_K]$;
- Generate $\boldsymbol{x}$ according to Gaussian distribution $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$



Graphical representation of $\mathbf{p}(\mathbf{x}, \mathbf{z}) = \mathbf{p}(\boldsymbol{z})\mathbf{p}(\mathbf{x}|\mathbf{z})$

So, we get a distribution for the data point **x**:

$$\boxed{p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

From minimizing sum of square distances to finding maximum likelihood

$X = \{x_1, \ldots, x_N\}$

$\pi = \{\pi_1, \ldots, \pi_K\}$

$\mu = \{\mu_1, \ldots, \mu_K\}$

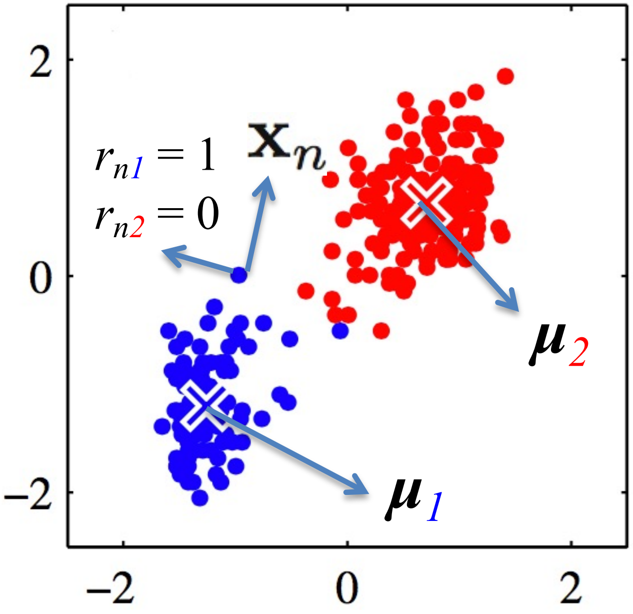
$\Sigma = \{\Sigma_1, \ldots, \Sigma_K\}$

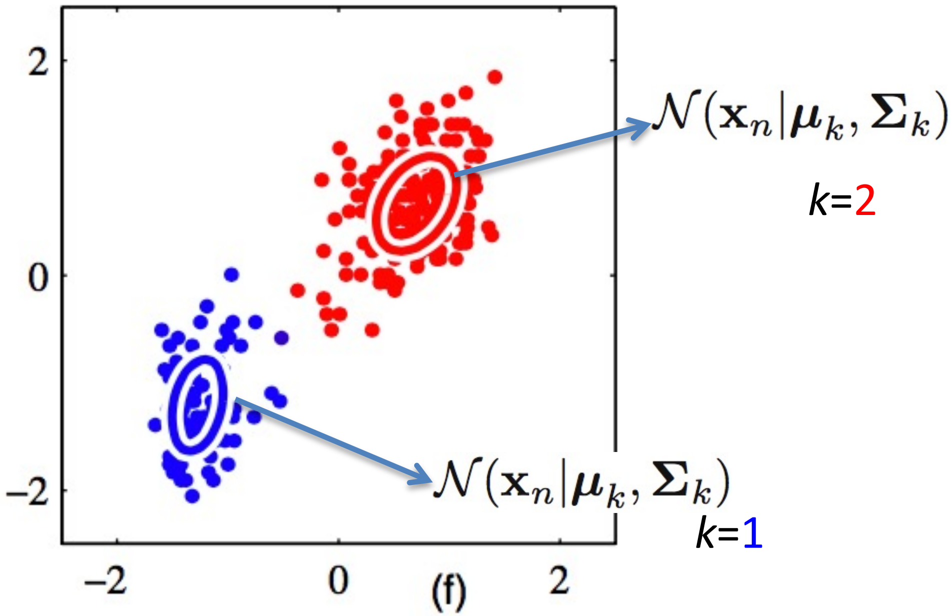
minimize

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

maximize likelihood

$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$





Remember: **The closer the distance, the more likely the probability.**

Outline

- Gaussian Mixture Models (GMM)
  - From generation process perspective

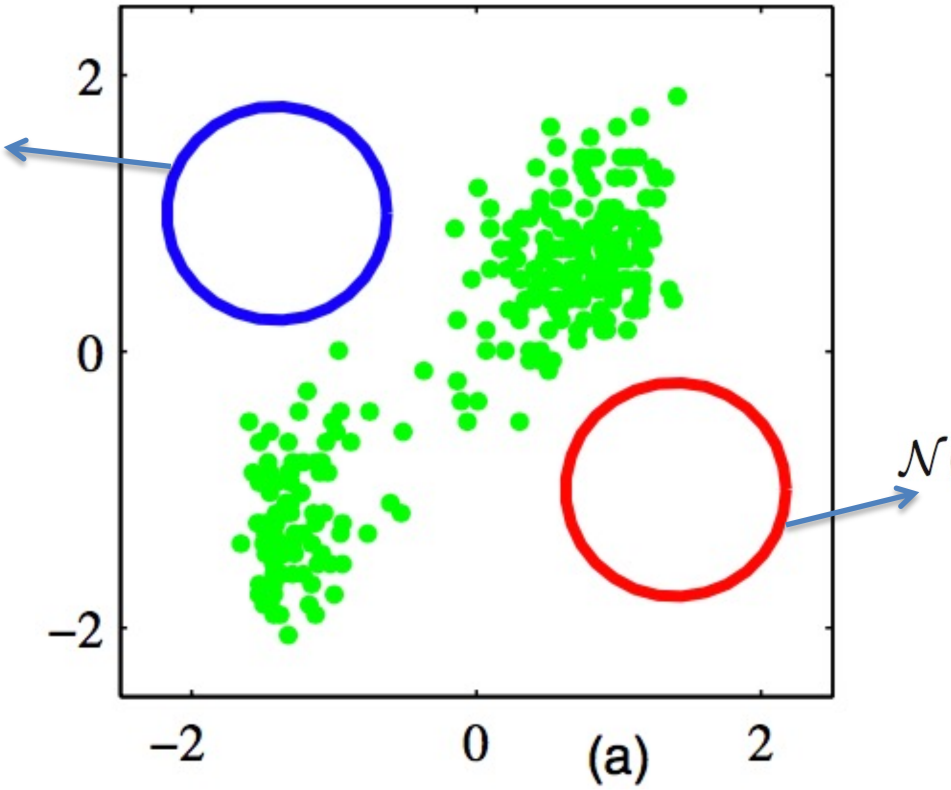
- **Expectation-Maximization (EM) for maximum likelihood**
  - **An alternative view to verify its properties**

- A brief history of EM

Expectation-Maximization (EM) algorithm for maximum likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Initialization



$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=1

$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=2

Details of the EM Algorithm for GMM

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.

2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

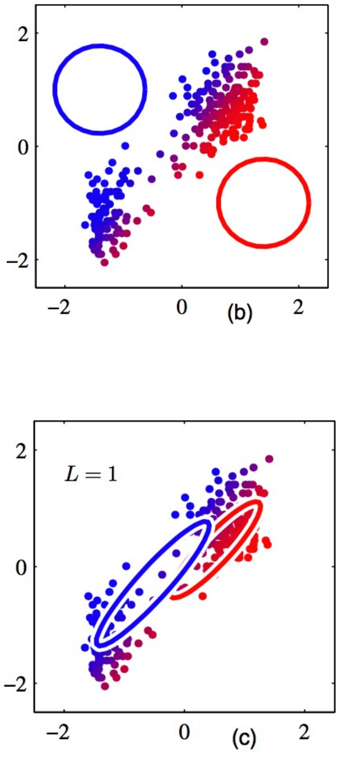
where

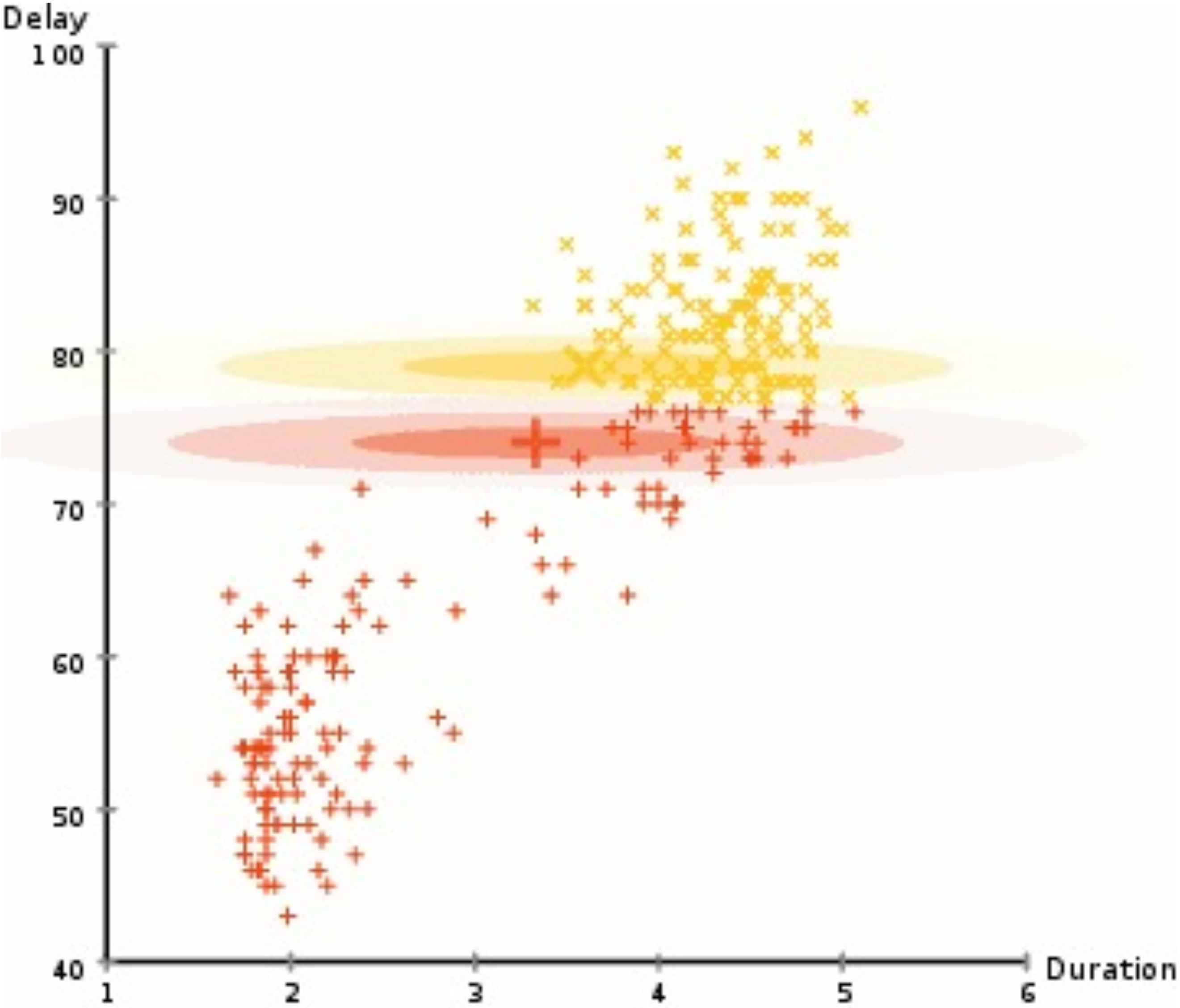
$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}).$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.



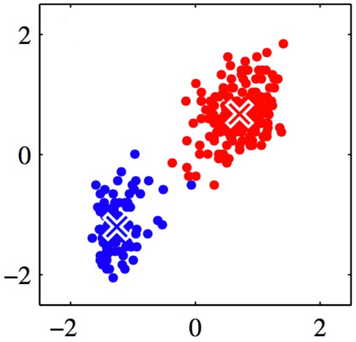

Delay
100
90
80
70
60
50
40
1
2
3
4
5
6
Duration

Discussion Question (1)

- How can EM for GMM degenerate back to k-mean algorithm?

Relation to K-means

$$\|\mathbf{x} - \boldsymbol{\mu}_k\|^2$$

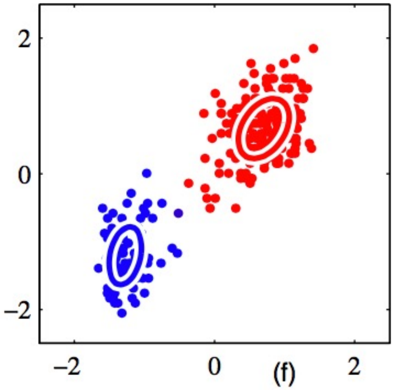


$$\{\mu_k\}$$

One-in-K assignment

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

$$\boldsymbol{\Sigma}_k = \varepsilon \mathbf{I}$$



GMM considers covariance and mixing weights.

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}$$

Soft assignment

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\}}{\sum_j \pi_j \exp\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\}}$$

Discussion Question (2)

- Can you design a variant algorithm between k-mean and EM?

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\boldsymbol{\theta}$, the goal is to maximize the likelihood function $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   where

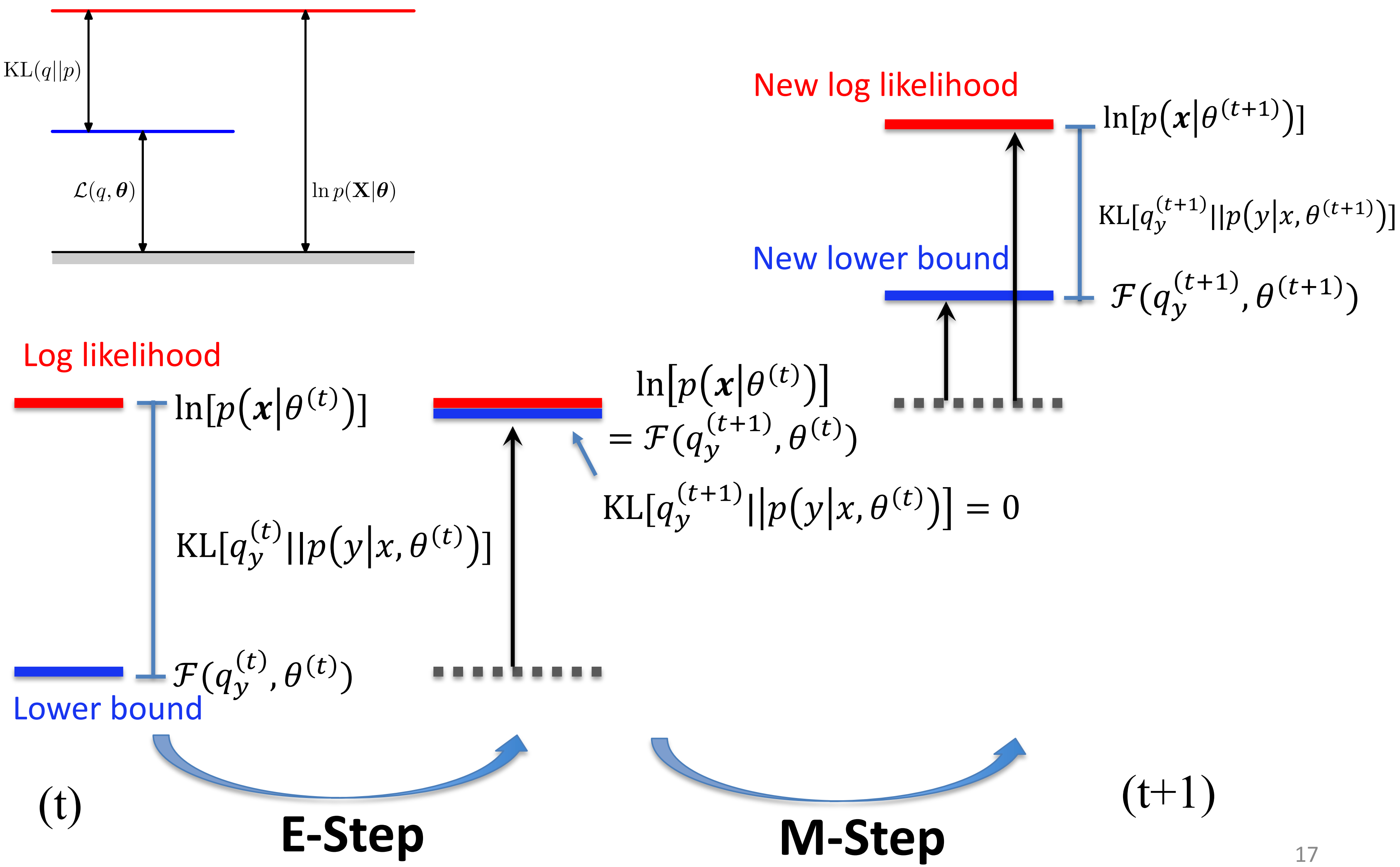
$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$$

   and return to step 2.

EM never decreases the likelihood



$\mathrm{KL}(q||p)$

$\mathcal{L}(q, \boldsymbol{\theta})$

$\ln p(\mathbf{X}|\boldsymbol{\theta})$

New log likelihood

$\ln[p(\boldsymbol{x}|\theta^{(t+1)})]$

$\mathrm{KL}[q_y^{(t+1)}||p(y|x,\theta^{(t+1)})]$

New lower bound

$\mathcal{F}(q_y^{(t+1)},\theta^{(t+1)})$

Log likelihood

$\ln[p(\boldsymbol{x}|\theta^{(t)})]$

$\ln[p(\boldsymbol{x}|\theta^{(t)})]$
$= \mathcal{F}(q_y^{(t+1)},\theta^{(t)})$

$\mathrm{KL}[q_y^{(t+1)}||p(y|x,\theta^{(t)})] = 0$

$\mathrm{KL}[q_y^{(t)}||p(y|x,\theta^{(t)})]$

$\mathcal{F}(q_y^{(t)},\theta^{(t)})$

Lower bound

(t)

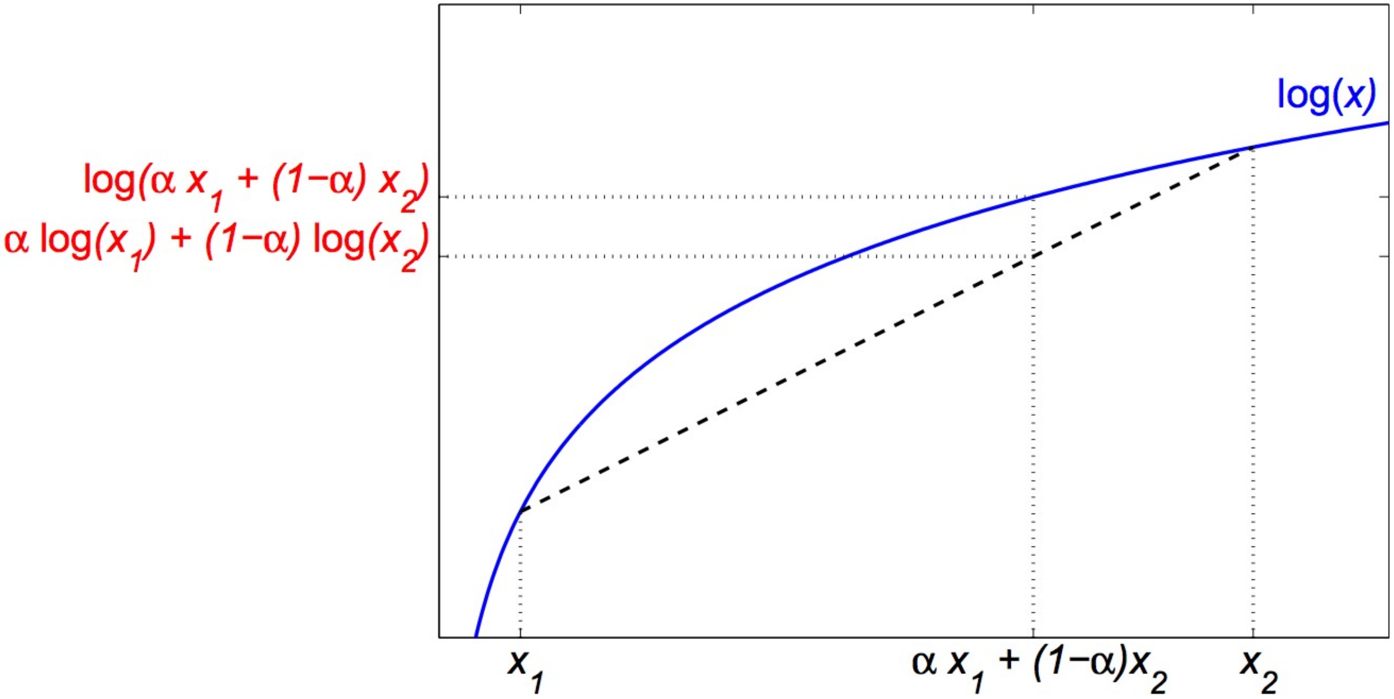
E-Step

M-Step

(t+1)

Why EM never decreases the (log)-likelihood?

Jensen's Inequality



For $\alpha_i \geq 0$, $\sum \alpha_i = 1$ and any $\{x_i > 0\}$

$$\log\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i \log(x_i)$$

Equality if and only if $\alpha_i = 1$ for some i (and therefore all others are 0).

The **KL**$[q(x)\|p(x)]$ is non-negative and zero iff $\forall x: \; p(x) = q(x)$

First let's consider discrete distributions; the Kullback-Liebler divergence is:

$$\textbf{KL}[q\|p] = \sum_i q_i \log \frac{q_i}{p_i}.$$

To find the distribution q which minimizes **KL**$[q\|p]$ we add a Lagrange multiplier to enforce the normalization constraint:

$$E \overset{\text{def}}{=} \textbf{KL}[q\|p] + \lambda\Big(1 - \sum_i q_i\Big) = \sum_i q_i \log \frac{q_i}{p_i} + \lambda\Big(1 - \sum_i q_i\Big)$$

We then take partial derivatives and set to zero:

$$\left.\begin{aligned} \frac{\partial E}{\partial q_i} &= \log q_i - \log p_i + 1 - \lambda = 0 \Rightarrow q_i = p_i \exp(\lambda - 1) \\ \frac{\partial E}{\partial \lambda} &= 1 - \sum_i q_i = 0 \Rightarrow \sum_i q_i = 1 \end{aligned}\right\} \Rightarrow q_i = p_i.$$

Check that the curvature (Hessian) is positive (definite), corresponding to a minimum:

$$\frac{\partial^2 E}{\partial q_i \partial q_i} = \frac{1}{q_i} > 0, \qquad \frac{\partial^2 E}{\partial q_i \partial q_j} = 0,$$

showing that $q_i = p_i$ is a genuine minimum.

At the minimum is it easily verified that **KL**$[p\|p] = 0$.

EM as maximizing a variational lower bound

$E[f(x)] \geq f(E[x])$.

Jensen’s Inequality due to convexity

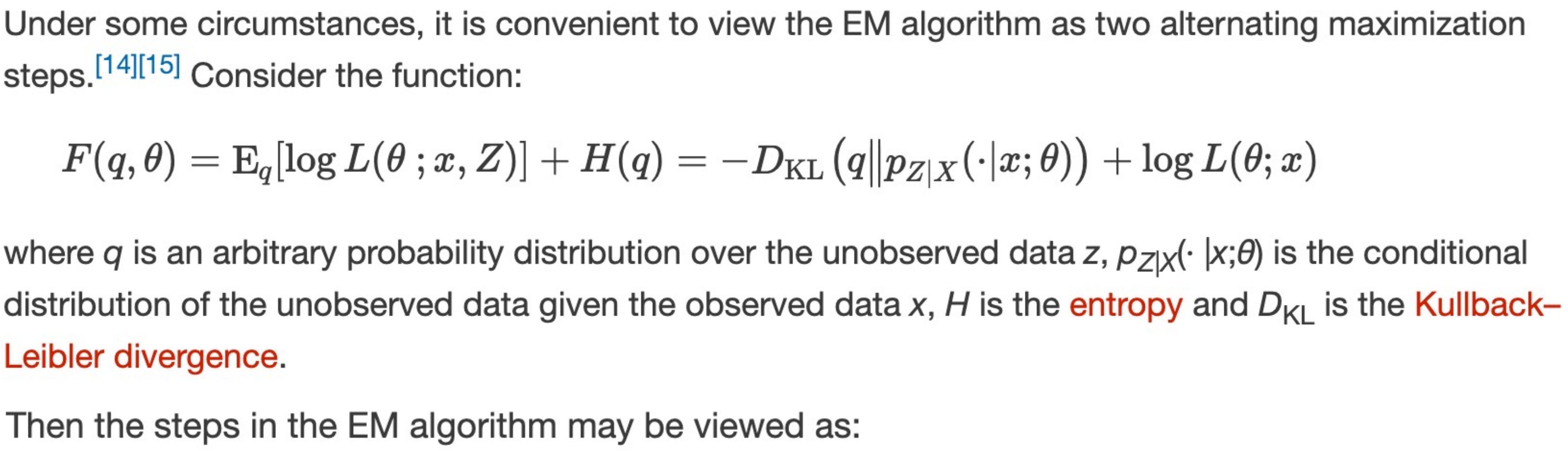
$$\log(P(\mathbf{x}|\theta)) = \log(\sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}|\theta))$$

$$\begin{aligned}
\log(P(\mathbf{x}|\theta)) &= \log(\sum_{\mathbf{y}} q(\mathbf{y})\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}) \\
&\geq E_q[\log(\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(\frac{P(\mathbf{y}|\mathbf{x},\theta)P(\mathbf{x}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(P(\mathbf{x}|\theta)] - E_q[\log(\frac{q(\mathbf{y})}{P(\mathbf{y}|\mathbf{x},\theta)}] \\
&\geq E_q[\log(P(\mathbf{x}|\theta))] - KL(q(\mathbf{y})\|P(\mathbf{y}|\mathbf{x},\theta)) \\
&\geq \log(P(\mathbf{x}|\theta)) - KL(q(\mathbf{y})\|P(\mathbf{y}|\mathbf{x},\theta))
\end{aligned}$$

$$\begin{aligned}
\log(P(\mathbf{x}|\theta)) &\geq E_q[\log(\frac{P(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{y})}] \\
&\geq E_q[\log(P(\mathbf{x},\mathbf{y}|\theta)] - E_q[\log(q(\mathbf{y}))] \\
&\geq E_q[\log(P(\mathbf{x},\mathbf{y}|\theta)] + H(q(\mathbf{y}))
\end{aligned}$$

An alternative view of EM

From wiki

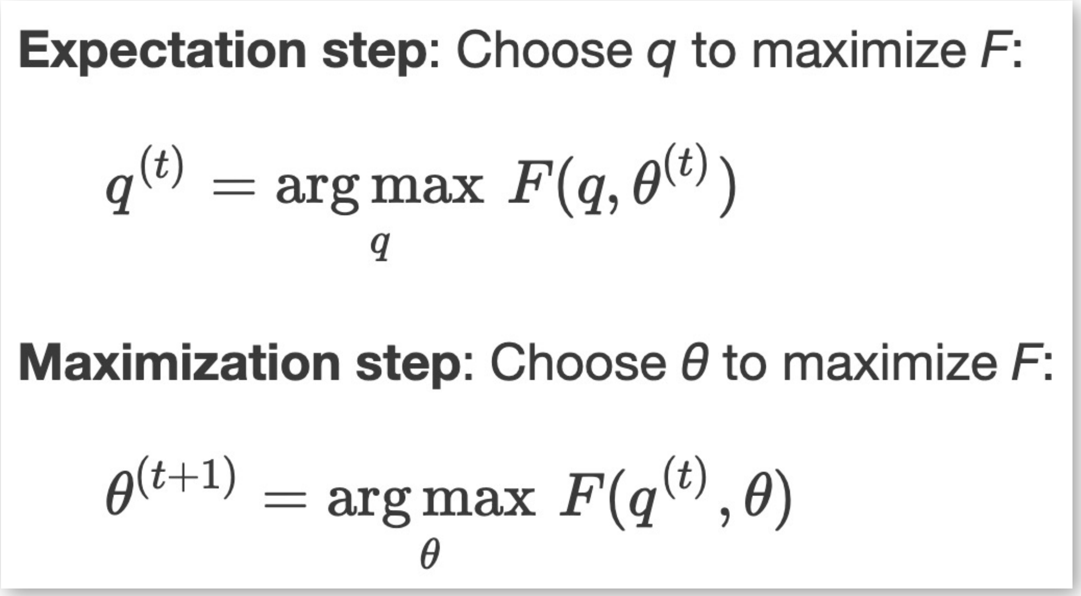


Under some circumstances, it is convenient to view the EM algorithm as two alternating maximization steps.[14][15] Consider the function:

$$F(q,\theta) = \mathrm{E}_q[\log L(\theta\,; x, Z)] + H(q) = -D_{\mathrm{KL}}\big(q \big\| p_{Z|X}(\cdot|x;\theta)\big) + \log L(\theta; x)$$

where q is an arbitrary probability distribution over the unobserved data z, $p_{Z|X}(\cdot\,|x;\theta)$ is the conditional distribution of the unobserved data given the observed data x, H is the entropy and D_{KL} is the Kullback–Leibler divergence.

Then the steps in the EM algorithm may be viewed as:



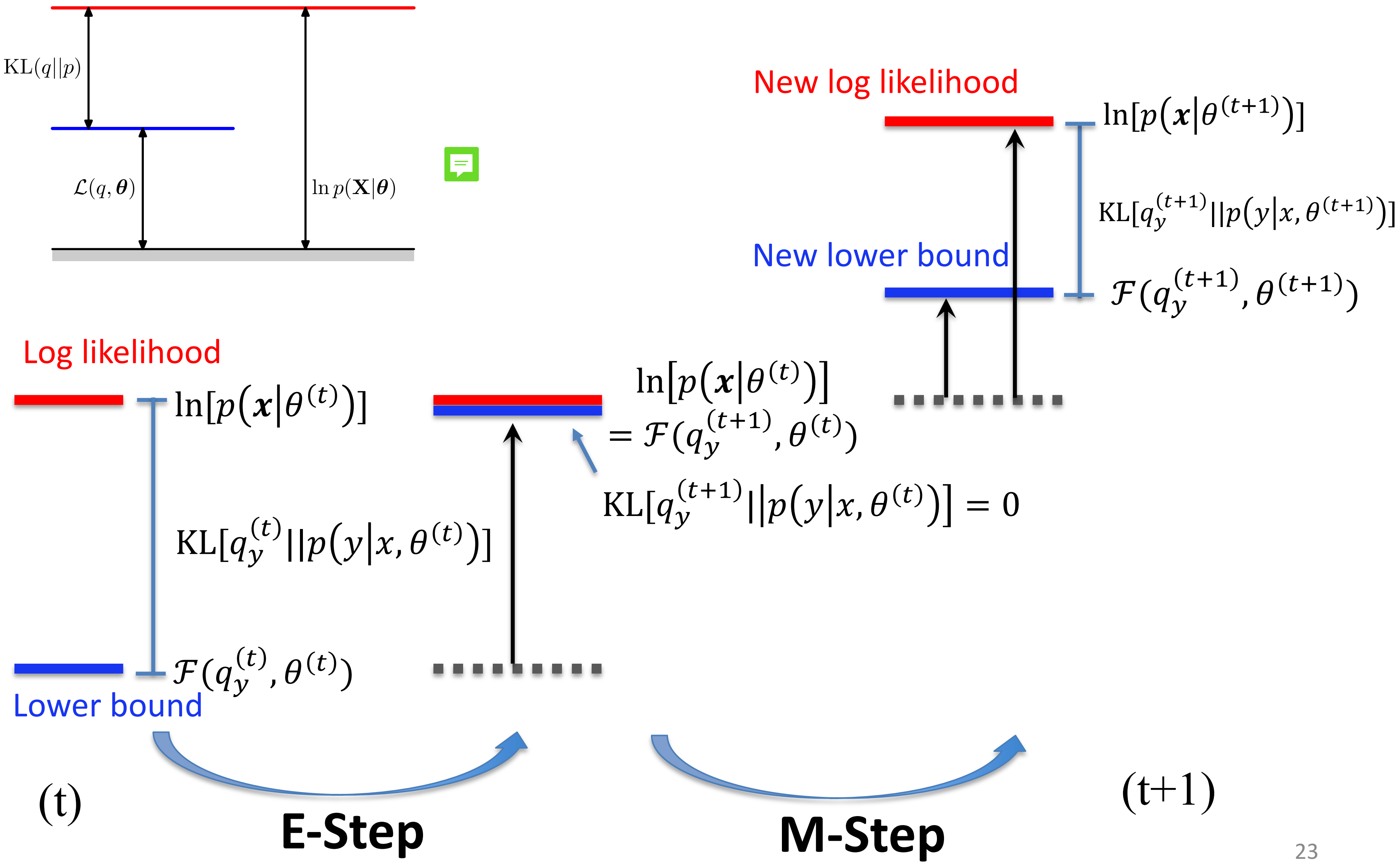
Expectation step: Choose q to maximize F:

$$q^{(t)} = \underset{q}{\arg\max}\ F(q, \theta^{(t)})$$

Maximization step: Choose θ to maximize F:

$$\theta^{(t+1)} = \underset{\theta}{\arg\max}\ F(q^{(t)}, \theta)$$

EM never decreases the likelihood


$\mathrm{KL}(q||p)$
$\mathcal{L}(q, \boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

New log likelihood

$\ln[p(\boldsymbol{x}|\theta^{(t+1)})]$

$\mathrm{KL}[q_y^{(t+1)}||p(y|x, \theta^{(t+1)})]$

New lower bound

$\mathcal{F}(q_y^{(t+1)}, \theta^{(t+1)})$

Log likelihood

$\ln[p(\boldsymbol{x}|\theta^{(t)})]$

$\ln[p(\boldsymbol{x}|\theta^{(t)})]$
$= \mathcal{F}(q_y^{(t+1)}, \theta^{(t)})$

$\mathrm{KL}[q_y^{(t+1)}||p(y|x, \theta^{(t)})] = 0$

$\mathrm{KL}[q_y^{(t)}||p(y|x, \theta^{(t)})]$

$\mathcal{F}(q_y^{(t)}, \theta^{(t)})$

Lower bound

(t)

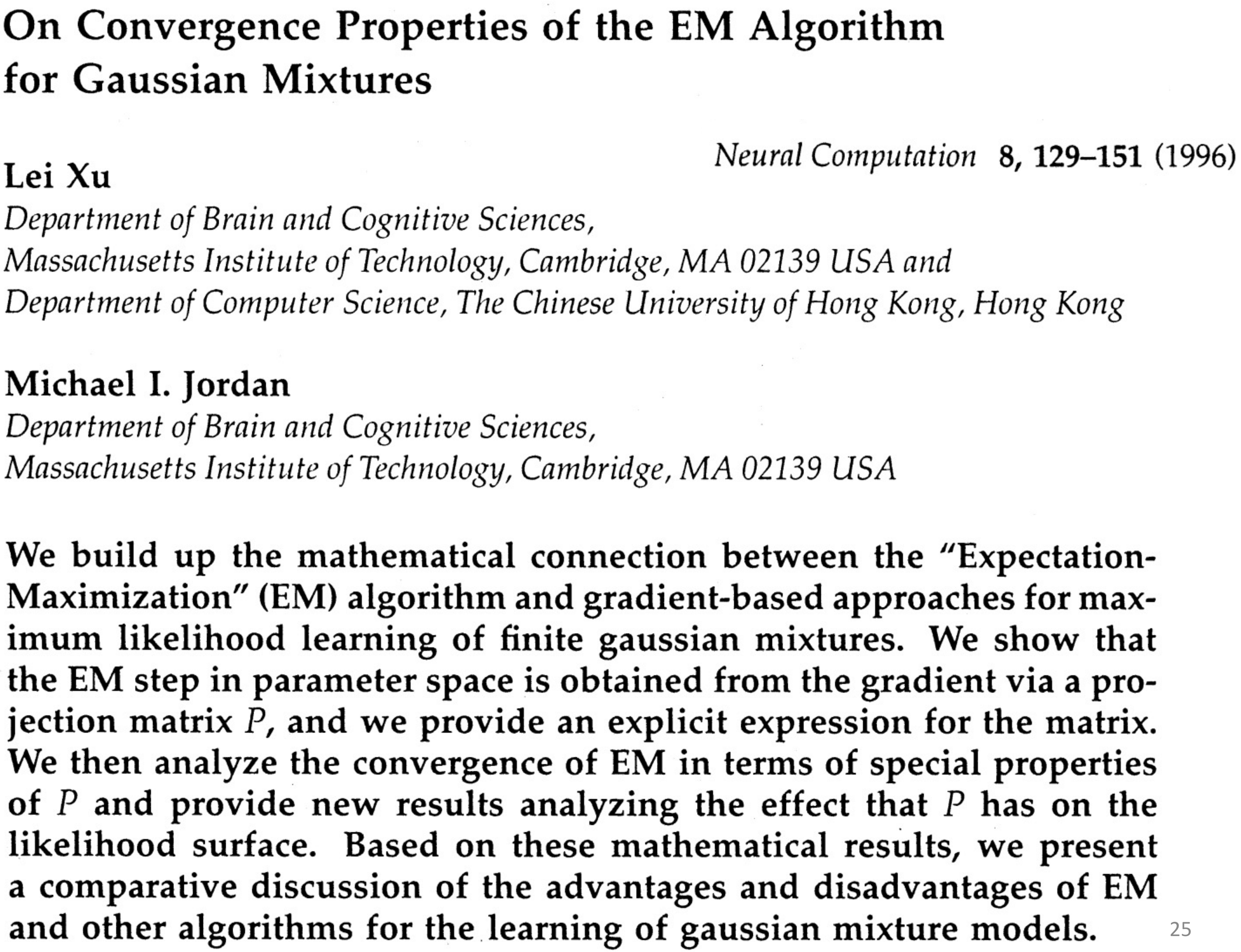
E-Step

M-Step

(t+1)

A brief history of EM

- EM had been "proposed many times in special circumstances" by earlier authors.
- One of the earliest is the gene-counting method for estimating allele frequencies by Cedric Smith (1955).
- A very detailed treatment of the EM method for exponential families was published by Rolf Sundberg in his thesis and several papers (1971,1974,1976).
- The Dempster–Laird–Rubin paper in 1977 generalized the method and sketched a convergence analysis for a wider class of problems. The paper received an enthusiastic discussion at the Royal Statistical Society meeting with Sundberg calling the paper "brilliant".
- The Dempster–Laird–Rubin algorithm was flawed and a correct convergence analysis was published by C. F. Jeff Wu in 1983.
- Xu & Jordan in 1996 built up the mathematical connection between the EM and gradient-based approaches for maximum likelihood learning of GMM.



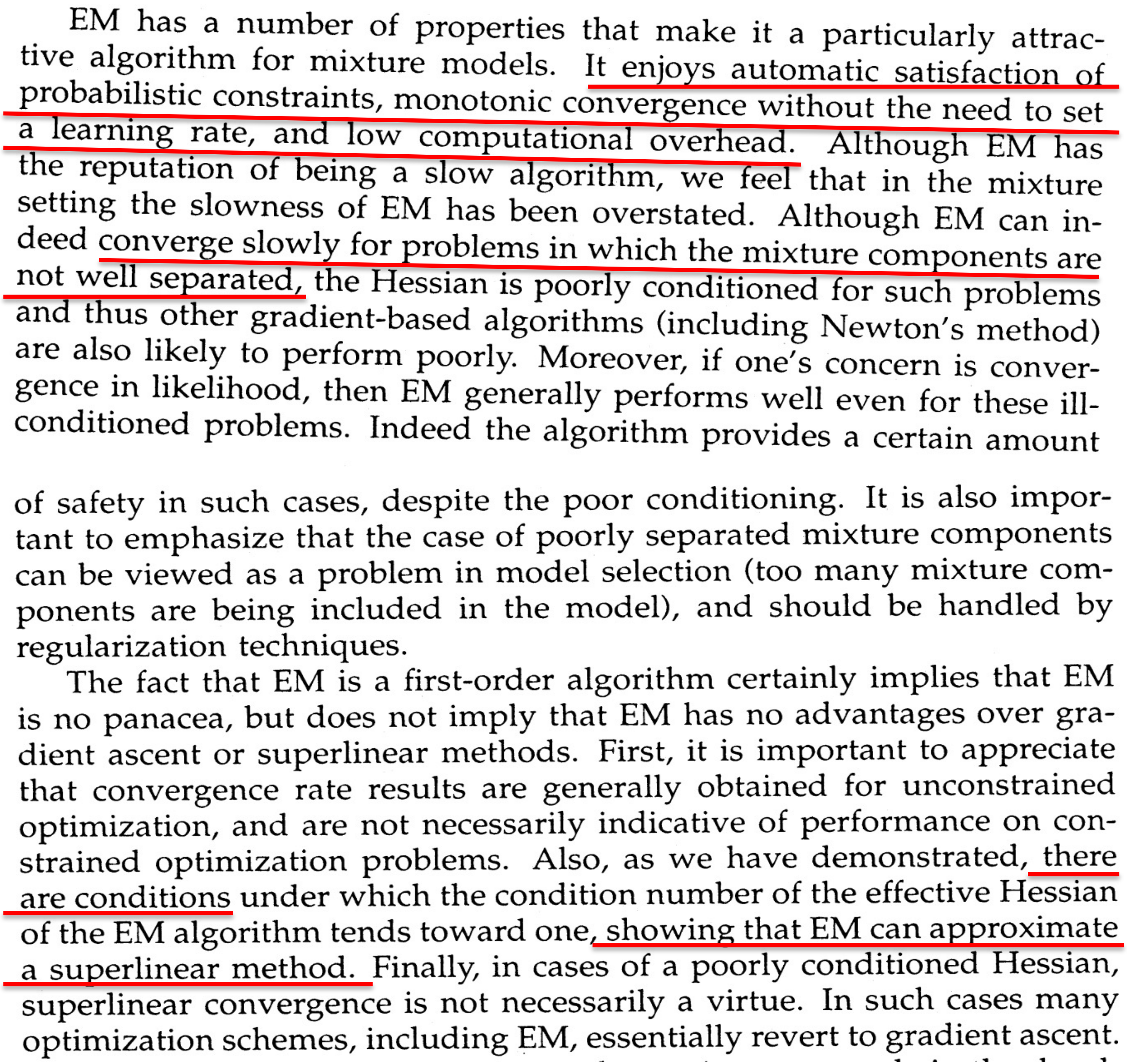
On Convergence Properties of the EM Algorithm for Gaussian Mixtures

Neural Computation **8, 129–151** (1996)

Lei Xu
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, MA 02139 USA and
Department of Computer Science, The Chinese University of Hong Kong, Hong Kong

Michael I. Jordan
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology, Cambridge, MA 02139 USA

We build up the mathematical connection between the "Expectation-Maximization" (EM) algorithm and gradient-based approaches for maximum likelihood learning of finite gaussian mixtures. We show that the EM step in parameter space is obtained from the gradient via a projection matrix P, and we provide an explicit expression for the matrix. We then analyze the convergence of EM in terms of special properties of P and provide new results analyzing the effect that P has on the likelihood surface. Based on these mathematical results, we present a comparative discussion of the advantages and disadvantages of EM and other algorithms for the learning of gaussian mixture models.



EM has a number of properties that make it a particularly attractive algorithm for mixture models. It enjoys automatic satisfaction of probabilistic constraints, monotonic convergence without the need to set a learning rate, and low computational overhead. Although EM has the reputation of being a slow algorithm, we feel that in the mixture setting the slowness of EM has been overstated. Although EM can indeed converge slowly for problems in which the mixture components are not well separated, the Hessian is poorly conditioned for such problems and thus other gradient-based algorithms (including Newton's method) are also likely to perform poorly. Moreover, if one's concern is convergence in likelihood, then EM generally performs well even for these ill-conditioned problems. Indeed the algorithm provides a certain amount of safety in such cases, despite the poor conditioning. It is also important to emphasize that the case of poorly separated mixture components can be viewed as a problem in model selection (too many mixture components are being included in the model), and should be handled by regularization techniques.

The fact that EM is a first-order algorithm certainly implies that EM is no panacea, but does not imply that EM has no advantages over gradient ascent or superlinear methods. First, it is important to appreciate that convergence rate results are generally obtained for unconstrained optimization, and are not necessarily indicative of performance on constrained optimization problems. Also, as we have demonstrated, there are conditions under which the condition number of the effective Hessian of the EM algorithm tends toward one, showing that EM can approximate a superlinear method. Finally, in cases of a poorly conditioned Hessian, superlinear convergence is not necessarily a virtue. In such cases many optimization schemes, including EM, essentially revert to gradient ascent.

EM的九层理解

1. EM 就是 E + M
2. EM 是一种局部下限构造
3. K-Means是一种Hard EM算法
4. 从EM 到 广义EM
5. 广义EM的一个特例是VBEM
6. 广义EM的另一个特例是WS算法
7. 广义EM的再一个特例是Gibbs抽样算法
8. WS算法是VAE和GAN组合的简化版
9. KL距离的统一

http://www.elecfans.com/d/604076.html

Thank you!

Matrix derivatives

$$\left[\frac{\partial \mathbf{x}}{\partial y}\right]_i = \frac{\partial x_i}{\partial y} \qquad \left[\frac{\partial x}{\partial \mathbf{y}}\right]_i = \frac{\partial x}{\partial y_i} \qquad \left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right]_{ij} = \frac{\partial x_i}{\partial y_j}$$

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \tag{69}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \tag{70}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T \tag{71}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T \tag{72}$$

$$\frac{\partial \det(\mathbf{X})}{\partial \mathbf{X}} = \det(\mathbf{X})(\mathbf{X}^{-1})^T \tag{49}$$

$$\frac{\partial \mathbf{Y}^{-1}}{\partial x} = -\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\mathbf{Y}^{-1} \tag{59}$$

http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf