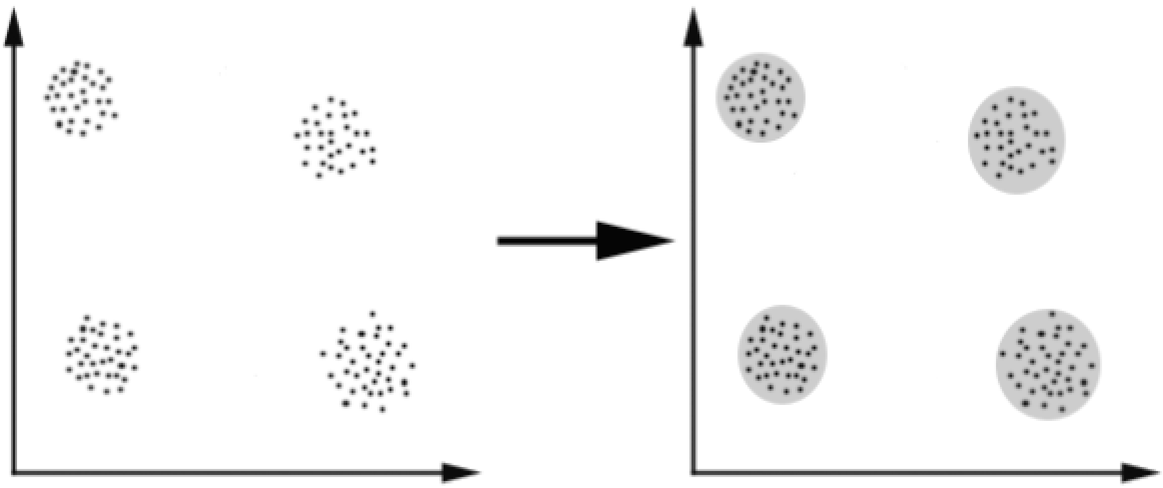
k-means

简介

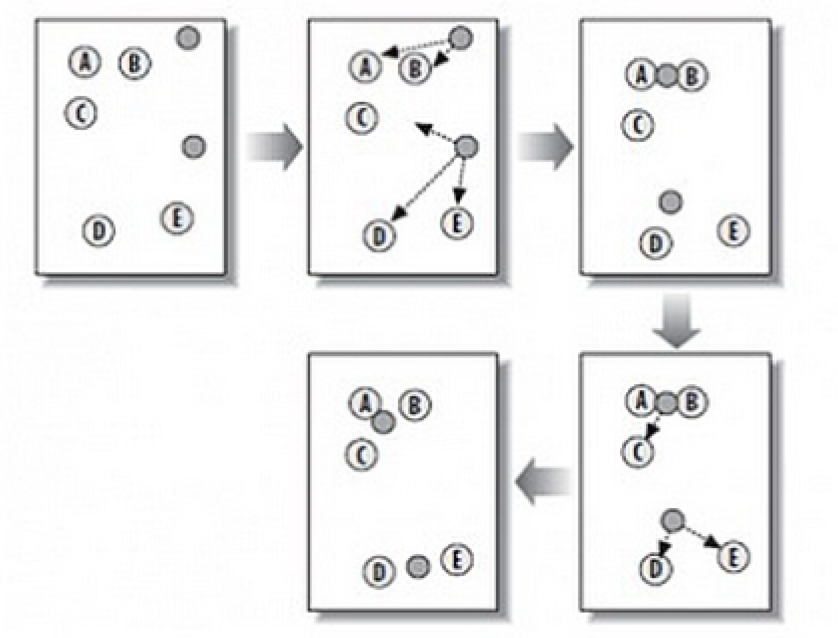
主要思想

> 有四个牧师去郊区布道，一开始牧师们随意选了几个布道点，并且把这几个布道点的情况公告给了郊区所有的居民，于是每个居民到离自己家最近的布道点去听课。听课之后，大家觉得距离太远了，于是每个牧师统计了一下自己的课上所有的居民的地址，搬到了所有地址的中心地带，并且在海报上更新了自己的布道点的位置。牧师每一次移动不可能离所有人都更近，有的人发现A牧师移动以后自己还不如去B牧师处听课更近，于是每个居民又去了离自己最近的布道点……就这样，牧师每个礼拜更新自己的位置，居民根据自己的情况选择布道点，最终稳定了下来。

算法图解一览



K-Means 要解决的问题



K-Means 算法概要

从上图中，我们可以看到，**A, B, C, D, E 是五个在图中点。而灰色的点是我们的种子点，也就是我们用来找点群的点。**有两个种子点，所以k=2。

然后，k-means的算法如下：

1. 随机在图中取k（这里k=2）个种子点。
2. 然后对图中的所有点求到这k个种子点的距离，假如点P_i离种子点S_i最近，那么P_i属于S_i点群。（上图中，我们可以看到A,B属于上面的种子点，C,D,E属于下面中部的种子点）
3. 接下来，我们要移动种子点到属于他的"点群"的中心。（见图上的第三步）
4. 然后重复第2）和第3）步，直到，种子点没有移动（我们可以看到图中的第四步上面的种子点聚合了A,B,C，下面的种子点聚合了D，E）。

归纳下k-means算法的过程就是：

1. 选取初始聚类中心
2. 通过计算距离进行聚类
3. 重新计算聚类中心
4. 重复2-3步直至聚类中心不发生改变(或变化小于一定阈值)或者达到迭代次数上限

损失函数浅探

该算法旨在最小化目标函数，即在上面这些情况下的平方误差函数：

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{j \in S_i} \|x_j - \mu_{S_i}\|^2$$

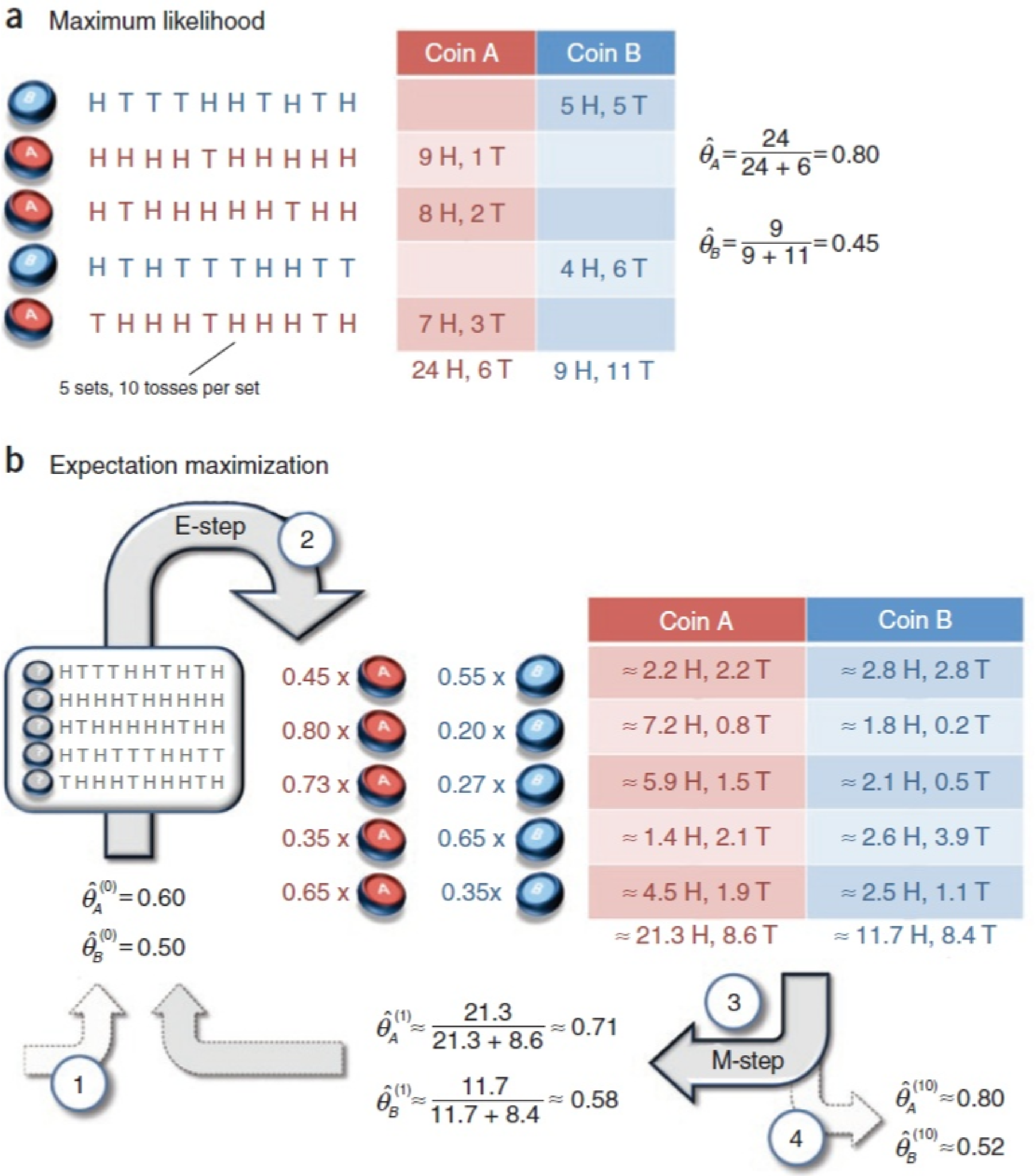
其中$S = S_1, S_2, \ldots, S_k$代表聚类后的k个类别。$\|x_j - \mu_{S_i}\|^2$是我们这里选定的距离公式，用于计算数据点和群集中心的距离。

那么，k-means是否一定会收敛呢？要回答好这个问题需要讲到k-means背后的理论支撑——EM（Expectation Maximum）算法。

Expectation Maximum

问题定义

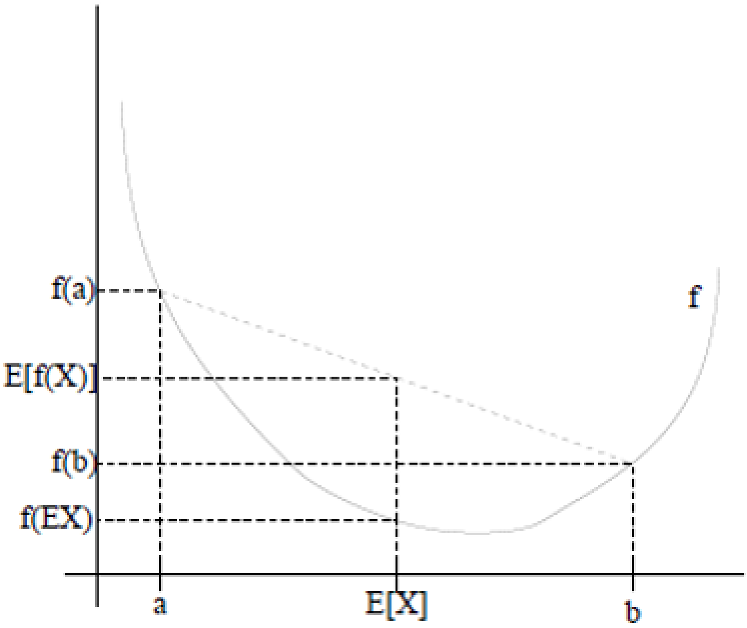
假设有两枚硬币A、B，以相同的概率随机选择一个硬币，进行如下的掷硬币实验：共做 5 次实验，每次实验独立的掷十次，结果如图中 a 所示，例如某次实验产生了H、T、T、T、H、H、T、H、T、H (H代表正面朝上)。a 是在知道每次选择的是A还是B的情况下进行，b是在不知道选择的是A还是B的情况下进行，问如何估计两个硬币正面出现的概率？



在case b中，我们假定A硬币投正面的概率是0.6，B硬币投正面的概率是0.5，然后通过极大似然估计来估计每一组结果是由A硬币投出的概率和由B硬币投出的概率。然后通过估计出来的选择硬币概率反过来重新计算投正面概率。其中每一轮选择的是A硬币还是B硬币我们称之为**隐参数z**，A硬币投正面的概率和B硬币投正面的概率称为**模型参数**θ。

总而言之，EM就是解决带隐参数的参数估计问题的一类算法。

Jensen不等式



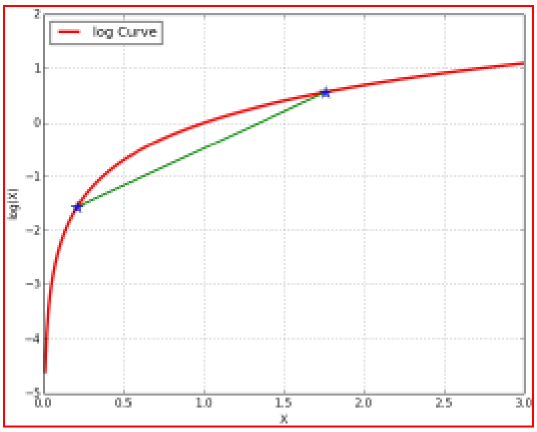
如果f是凸函数，X是随机变量，那么：

$$E[f(X)] \geq f(EX)$$

更特殊的形式：

$$f\left(\frac{x_1+x_2+\cdots+x_n}{n}\right) \leq \frac{f\left(x_1\right)+f\left(x_2\right)+\cdots+f\left(x_n\right)}{n}$$

log函数上的jensen不等式：



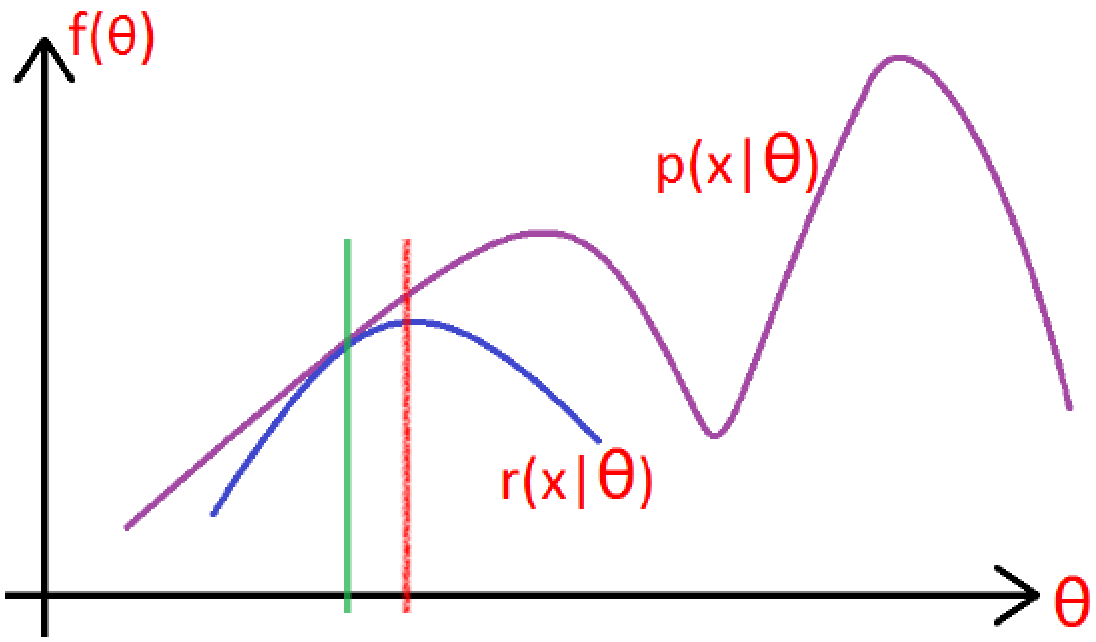
$$E[log(X)] \leq log(EX)$$

算法思路

给定的m个观察样本$\{x^{(1)},\ldots,x^{(m)}\}$，模型的参数为 θ ，我们想找到隐参数z，能使得$p(X,z)$最大。建立似然函数：

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log p(x^{(i)};\theta) \\
&= \sum_{i=1}^{m} \log \sum_{z} p(x^{(i)}, z; \theta)
\end{aligned}
$$

直接计算上述似然函数的最大值比较困难，所以我们希望能够找到一个不带隐变量z的函数$\gamma(x|\theta) \leq l(x, z; \theta)$恒成立，并用$\gamma(x|\theta)$逼近目标函数。



- 在绿色线位置，找到一个函数γ，能够使得该函数最接近目标函数，
  - 固定γ函数，找到最大值，然后更新θ,得到红线；
- 对于红线位置的参数θ：
  - 固定θ，找到一个最好的函数γ，使得该函数更接近目标函数。 重复该过程，直到收敛到局部最大值。

算法过程推导

令Q_i是z相对于样本$x^{(i)}$的一个分布，$Q_i \geq 0$，则：

$$
\begin{aligned}
l(\theta) &= \sum_{i=1}^{m} \log \sum_{z^{(i)}} p\left(x^{(i)}, z^{(i)}; \theta\right) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)} \\
&\geq \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{i}\right)_{z}(i)}{Q_i\left(z^{(i)}\right)}
\end{aligned}
$$

注意这里，$\sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i\left(z^{(i)}\right)}$相当于是一个关于z的函数的期望。简化之后可以表示为：

$$
logE(F(z)) \geq E(logF(z))
$$

其中：

$$F(z)=\frac{p\left(x^{(i)}, z^{(i)} ; \theta\right)}{Q_i\left(z^{(i)}\right)}$$

探究一下等号成立的条件，在jensen不等式中：$\frac{f\left(x_1\right)+f\left(x_2\right)}{2} \geq f\left(\frac{x_1+x_2}{2}\right)$，当且仅当$x_1=x_2$时等号成立。

那么对于上式而言，等号成立的条件应该是$F(z)$恒为常数，即：

$$\frac{p\left(x^{(i)}, z^{(i)} ; \theta\right)}{Q_i\left(z^{(i)}\right)}=C$$

因为Q是z的一个分布，所以应该保证$\sum_z Q_i\left(z^{(i)}\right)=1$，故

$$\begin{aligned}
Q_i\left(z^{(i)}\right) & =\frac{p\left(x^{(i)}, z^{(i)} ; \theta\right)}{\sum_z p\left(x^{(i)}, z^{(i)} ; \theta\right)} \\
& =\frac{p\left(x^{(i)}, z^{(i)} ; \theta\right)}{p\left(x^{(i)} ; \theta\right)}=p\left(z^{(i)} \mid x^{(i)} ; \theta\right)
\end{aligned}$$

所以Q函数是已知观测数据和参数θ的后验概率分布（条件概率分布）。解决了$Q_i\left(z^{(i)}\right)$如何计算的问题之后，就可以迭代求解θ和隐变量z了。

现在我们来回顾一下EM的求解过程，先定一个初始的模型参数θ，然后根据模型参数去求解隐变量z，也就是我们刚刚说的Q，求解Q之后，再反过来通过计算下界函数的最大值来更新θ。

- **E-步：**

$$Q_i\left(z^{(i)}\right):=p\left(z^{(i)} \mid x^{(i)} ; \theta\right)$$

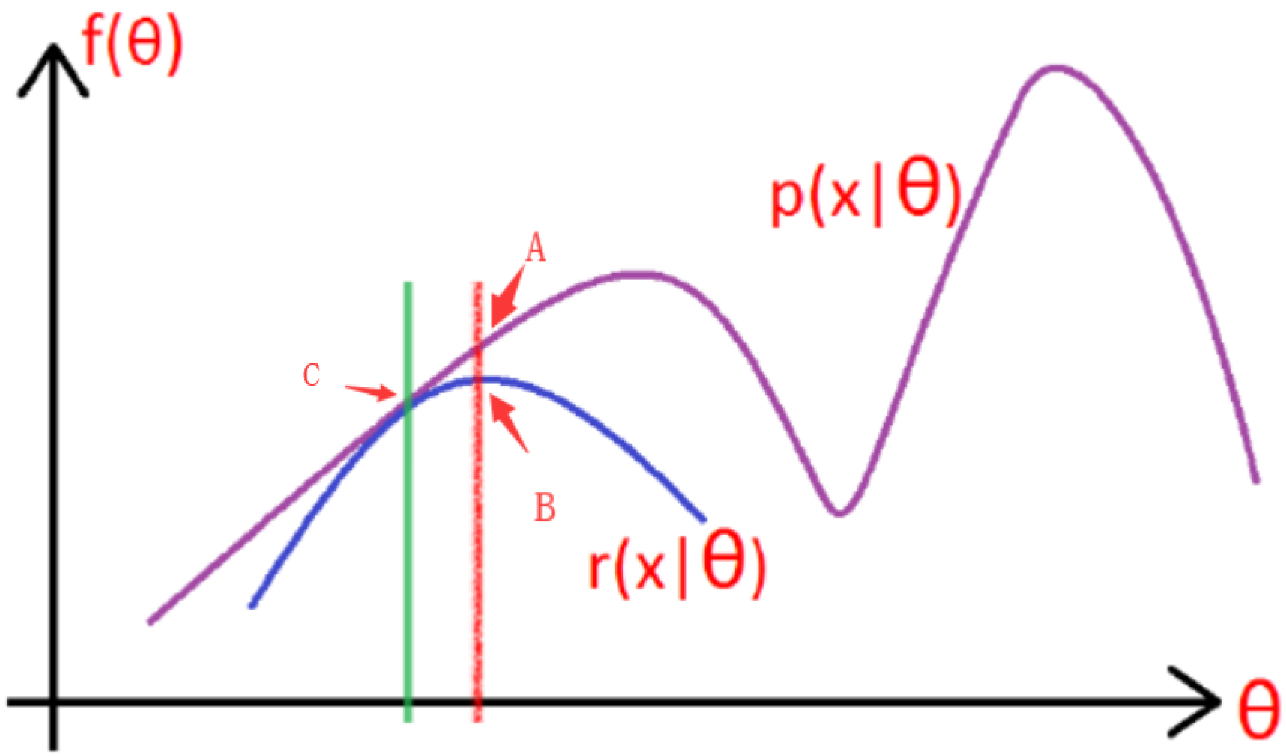
- **M-步：**

$$\theta:=\arg \max _\theta \sum_i \sum_{z^{(i)}} Q_i\left(z^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)} ; \theta\right)}{Q_i\left(z^{(i)}\right)}$$

收敛性证明

$$\begin{aligned}
l\left(\boldsymbol{\theta}_{t+1}\right) & \geq \sum_t \sum_{z_t} Q_{i t}\left(z_i\right) \log \frac{p\left(\mathbf{x}_i, z_i ; \boldsymbol{\theta}_{t+1}\right)}{Q_{i t}\left(z_i\right)} \\
& \geq \sum_i \sum_{z_i} Q_{i t}\left(z_i\right) \log \frac{p\left(\mathbf{x}_i, z_i ; \boldsymbol{\theta}_t\right)}{Q_{i t}\left(z_i\right)} \\
& =l\left(\boldsymbol{\theta}_t\right)
\end{aligned}$$

- 第一个$\geq$：t+1时刻的似然函数必然大于等于t时刻下界函数的最大值
- 第二个$\geq$：t时刻下界函数的最大值必然大于等于t时刻的似然函数（想一想为什么是t时刻的似然函数值？）
- 用图形来解释：A点大于等于B点，B点大于等于C点，所以A点（t+1时刻的似然函数）大于等于C点（t时刻的似然函数）



k-means的收敛性

为了搞清楚k-means的收敛性，我们必须得先搞清楚k-means算法中的模型参数和隐变量。模型参数是簇中心点的位置，隐变量是每个样本属于哪个类别，求解的似然函数是：

$$P\left(x, z \mid \mu_{1}, \mu_{2}, \ldots, \mu_{k}\right) \propto \begin{cases} \exp\left(-\|x-\mu_{z}\|_{2}^{2}\right), \|x-\mu_{z}\|_{2}=\min_{k}\|x-\mu_{k}\|_{2} \\ \quad 0 \quad , \|x-\mu_{z}\|_{2}>\min_{k}\|x-\mu_{k}\|_{2} \end{cases}$$

E步是固定模型参数μ_k（中心点的位置），进而求解隐变量的分布，也就是每个样本属于哪个类别：

$$\gamma_{nk}=\begin{cases} 1, & \text{if } k=\operatorname{argmin}_{j}\|x_{n}-\mu_{j}\|^{2} \\ 0, & \text{otherwise} \end{cases}$$

M步是计算出所有样本所属类别之后，更新模型参数μ_k（中心点的位置）：

$$\mu_{k}=\frac{\sum_{n}\gamma_{nk}x_{n}}{\sum_{n}\gamma_{nk}}$$

通过将k-means “EM"化，就可以通过说明EM收敛性等价到k-means收敛性。

k-means的几个问题

- k个数和位置的选择
- 中心点有其他的更新方式吗？
- 距离公式可以更换吗？