期望最大化算法

Expectation-Maximum Algorithm

- 三硬币模型: 硬币A、B、C, 正面概率$\theta = (\pi, p, q)$
- A正面时选B, 反面选C
- 得到结果：1，1，0，1，0，0，1，0，1，1
- 问题：只能看结果，不能看中间过程，估算θ

  解:

$$
\begin{aligned}
P(y|\theta) &= \sum_z P(y, z|\theta) = \sum_z P(z|\theta)P(y|z, \theta) \\
&= \pi p^y(1-p)^{1-y} + (1-\pi)q^y(1-q)^{1-y}
\end{aligned}
$$

这里随机变量y时观测变量，表示一次试验观测的结果是1或0;
随机变量z是隐变量，表示未观测到的掷硬币A的结果。

- 观测数据(不完全数据incomplete data): $Y = (y_1, y_2, \ldots, y_n)^T$
- 未观测数据(隐变量): $Z = (z_1, z_2, \ldots, z_n)^T$
- Y和Z连在一起称为完全数据(complete data)
- 似然函数: $P(Y|\theta) = \sum_z P(Z|\theta)P(Y|Z,\theta)$
- 即: $P(Y|\theta) = \prod_{j=1}^{n} \pi p^{y_j}(1-p)^{1-y_j} + (1-\pi)q^{y_j}(1-q)^{1-y_j}$
- 极大似然估计:

$$\hat{\theta} = \arg\max_{\theta} \log P(Y|\theta)$$

- 该问题没有解析解，需要使用 EM迭代法。
- EM算法通过迭代法计算$L(\theta) = \log P(Y|\theta)$的极大似然估计, 每次迭代包含两步:
  - E步，计算期望
  - M步，极大化

选取初值: $\theta^{(0)}=(\pi^{(0)}, p^{(0)}, q^{(0)})$

第i步的估计值: $\theta^{(i)}=(\pi^{(i)}, p^{(i)}, q^{(i)})$

EM算法第$i+1$次迭代:

E步：计算在模型参数下$\pi^{(i)}, p^{(i)}, q^{(i)}$观测数据$y_j$来自硬币B的概率：

$$\mu_j^{(i+1)}=\frac{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}+(1-\pi^{(i)})(q^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}}$$

M步：计算模型参数的新估计值:

$\pi^{(i+1)}=\frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)}, p^{(i+1)}=\frac{\sum_{j=1}^{n}\mu_j^{(i+1)}y_j}{\sum_{j=1}^{n}\mu_j^{(i+1)}},$

$q^{(i+1)}=\frac{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})y_j}{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})}$

如果取初值: $\pi^{(0)} = 0.5, p^{(0)} = 0.5, q^{(0)} = 0.5$

对$y_j = 1$与$y_j = 0$均有$\mu_j^{(1)} = 0.5$

利用迭代公式，得: $\pi^{(1)} = 0.5, p^{(1)} = 0.6, q^{(1)} = 0.6$

$\mu_j^{(2)} = 0.5, j = 1, 2, \ldots, 10$

继续迭代，得: $\pi^{(2)} = 0.5, p^{(2)} = 0.6, q^{(2)} = 0.6$

得到模型参数的极大似然估计： $\hat{\pi} = 0.5, \hat{p} = 0.6, \hat{q} = 0.6$

如果取初值: $\pi^{(0)} = 0.4, p^{(0)} = 0.6, q^{(0)} = 0.7,$

按照上述方法计算可得: $\hat{\pi} = 0.4064, \hat{p} = 0.5368, \hat{q} = 0.6432$

EM算法

- 输入: 观测变量数据Y，隐变量数据Z，联合分布$P(Y,Z|\theta)$, 条件分布$P(Z|Y,\theta)$
- 输出: 模型参数θ
  (1)选择参数的初值$\theta^{(0)}$, 开始迭代;
  (2)E步: 记$\theta^{(i)}$为第i次迭代参数θ的估计值，在第$i+1$次迭代的E步，给定观测数据Y和当前参数估计$\theta^{(i)}$, 计算期望:
  $$\begin{aligned} Q(\theta,\theta^{(i)}) &= E_{Z\sim P(Z|Y,\theta^{(i)})}[\log P(Y,Z|\theta)] \\ &= \int_Z P(Z|Y,\theta^{(i)})\log P(Y,Z|\theta)dZ \end{aligned}$$
  (3)M步: 通过极大化$Q(\theta,\theta)$，确定第$i+1$次迭代的参数估计值$\theta^{(i+1)}$
  $$\theta^{(i+1)} = \arg\max_\theta Q(\theta,\theta^{(i)})$$
  (4)重复第(2)步和第(3)步，直到收敛。

关于EM算法的几点说明：

(1) 参数的初值可以任意选择，但需注意EM算法对初值是敏感的

(2) E步求$Q(\theta, \theta^{(i)})$. Q函数式中Z是未观测数据，Y是观测数据。

(3) 每次迭代实际在求Q函数及其极大，实际上可使似然函数$P(Y|\theta)$增大或达到局部极值.

(4) 停止迭代的条件，一般是对较小的正数ϵ_1, ϵ_2, 若满足

$$\|\theta^{(i+1)} - \theta^{(i)}\| < \epsilon_1 \text{ 或 } \|Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)})\| < \epsilon_2$$

则停止迭代.

EM算法的导出(从KL Divergence角度进行分析)

根据极大似然估计的思想，目标是极大化观测数据(Y)关于参数θ的对数似然函数，即极大化

$$\begin{aligned} L(\theta) &= \log P(Y|\theta) \\ &= \log \int_Z P(Y,Z|\theta)dZ \\ &= \log \int_Z P(Y|Z,\theta)P(Z|\theta)dZ \end{aligned}$$

由于Z是隐藏变量，该方案不可行。
考虑

$$\begin{aligned} && \log P(Y|\theta) &= \log P(Y,Z|\theta) - \log(Z|Y,\theta) \\ && &= \log \frac{P(Y,Z|\theta)}{Q(Z)} - \log \frac{P(Z|Y,\theta)}{Q(Z)} \\ \Rightarrow && \int_Z Q(Z) \log P(Y|\theta)dZ &= \int_Z Q(Z) \log \frac{P(Y,Z|\theta)}{Q(Z)}dZ - \int_Z Q(Z) \log \frac{P(Z|Y,\theta)}{Q(Z)}dZ \\ && \log P(Y|\theta) &= ELBO + KL(Q(Z), P(Z|Y,\theta)) \\ \Rightarrow && \log P(Y|\theta) &\geq ELBO \end{aligned}$$

当且仅当$P(Z|X,\theta) = Q(Z)$时，等号成立。

对下界$ELBO$(Evidence Lower Bound)进行优化，使其尽可能的变大

$$\hat{\theta} = \arg\max_{\theta} ELBO = \arg\max_{\theta} \int_Z Q(Z) \log \frac{P(Y, Z|\theta)}{Q(Z)} dZ$$

取$Q(Z) = P(Z|Y, \theta^{(t)})$, 则:

$$\begin{aligned}
\theta^{(t+1)} \quad & = \arg\max_\theta \int_Z Q(Z) \log \tfrac{P(Y, Z|\theta)}{Q(Z)} dZ \\
& = \arg\max_\theta \int_Z P(Z|Y, \theta^{(t)}) \log \tfrac{P(Y, Z|\theta)}{P(Z|Y, \theta^{(t)})} dZ \\
& = \arg\max_\theta \int_Z P(Z|Y, \theta^{(t)}) \log P(Y, Z|\theta) dZ \\
& \quad - \int_Z P(Z|Y, \theta^{(t)}) \log P(Z|Y, \theta^{(t)}) dZ \\
& = \arg\max_\theta \int_Z P(Z|Y, \theta^{(t)}) \log P(Y, Z|\theta) dZ \\
& = \arg\max_\theta E_{Z \sim P(Z|Y, \theta^{(t)})} [\log P(Y, Z|\theta)]
\end{aligned}$$

EM算法的导出(从Jensen Inequality的角度进行分析)

- Jensen不等式: 对于一个凸函数f, 都有函数值的期望大于等于期望的函数值，即:

$$E(f(X)) \geq f(E(X))$$

等号成立当且仅当(1)如果f是严格凸函数时，有$X = C$;
(2)如果f不是严格凸函数时，有$X = C$或在该区间内f是仿射函数。
利用Jensen不等式得到:

$$\begin{aligned} \log P(Y|\theta) &= \log \int_Z P(Y,Z|\theta) dZ \\ &= \log \int_Z Q(Z) \frac{P(Y,Z|\theta)}{Q(Z)} dZ \\ &= \log E_{Z\sim Q(Z)}[\frac{P(Y,Z|\theta)}{Q(Z)}] \\ &\geq E_{Z\sim Q(Z)}[\log \frac{P(Y,Z|\theta)}{Q(Z)}] \end{aligned}$$

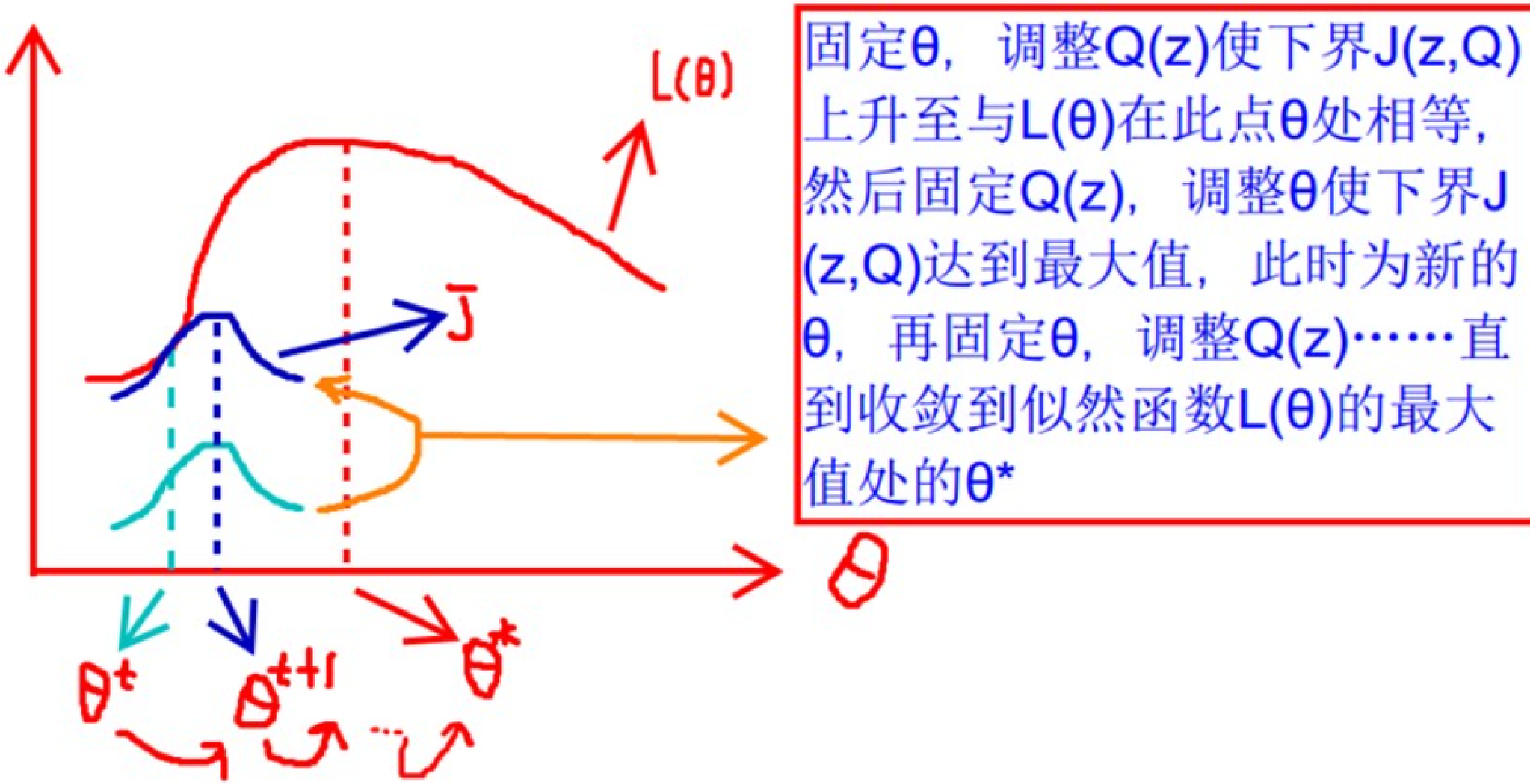
等号成立当且仅当$\log \frac{P(Y,Z|\theta)}{Q(Z)} = C$, 即$Q(Z) = P(Z|Y,\theta)$

于是，我们考虑迭代计算，取$Q(Z)=P(Z|Y,\theta^{(t)})$
因此，有:

$$E_{Z\sim Q(Z)}[\log \frac{P(Y,Z|\theta)}{Q(Z)}] \quad = E_{Z\sim P(Z|Y,\theta^{(t)})}[\log \frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(t)})}]$$

则:

$$\begin{aligned}\theta^{(t+1)} \quad &= \arg\max_\theta \int_Z P(Z|Y,\theta^{(t)}) \log \frac{P(Y,Z|\theta)}{P(Z|Y,\theta^{(t)})} dZ \\ &= \arg\max_\theta \int_Z P(Z|Y,\theta^{(t)}) \log P(Y,Z|\theta) dZ \\ &\quad - \int_Z P(Z|Y,\theta^{(t)}) \log P(Z|Y,\theta^{(t)}) dZ \\ &= \arg\max_\theta \int_Z P(Z|Y,\theta^{(t)}) \log P(Y,Z|\theta) dZ\end{aligned}$$


L(θ)
J̄
θ
θ^t
θ^{t+1}
θ*
固定θ，调整Q(z)使下界J(z,Q)上升至与L(θ)在此点θ处相等，然后固定Q(z)，调整θ使下界J(z,Q)达到最大值，此时为新的θ，再固定θ，调整Q(z)……直到收敛到似然函数L(θ)的最大值处的θ*

EM算法的收敛性

目标: 当$\theta^{(t)} \to \theta^{(t+1)}$时，有$\log P(Y|\theta^{(t)}) \leq \log p(Y|\theta^{(t+1)})$
考虑:

$$\log P(Y|\theta) \quad = \log P(Y, Z|\theta) - \log(Z|Y, \theta)$$

同时对两边求关于$P(Z|Y, \theta^{(t)})$的期望
左边:

$$\begin{aligned} E_{Z\sim P(Z|Y,\theta^{(t)})}[\log P(Y|\theta)] \quad &= \int_Z P(Z|Y, \theta^{(t)}) \log P(Y|\theta) dZ \\ &= \log P(Y|\theta) \int_Z P(Z|Y, \theta^{(t)}) dZ \\ &= \log P(Y|\theta) \cdot 1 = \log P(Y|\theta) \end{aligned}$$

引入$Q(\theta, \theta^{(t)}) = \int_Z P(Z|Y, \theta^{(t)}) \log P(Y, Z|\theta) dZ$
$H(\theta, \theta^{(t)}) = \int_Z P(Z|Y, \theta^{(t)}) \log P(Z|Y, \theta) dZ$

则右边$= Q(\theta, \theta^{(t)}) - H(\theta, \theta^{(t)})$

显然$Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)})$

又:

$$
\begin{aligned}
& H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) \\
= & \int_Z P(Z|Y, \theta^{(t)}) \log P(Z|Y, \theta^{(t+1)}) dZ - \int_Z P(Z|Y, \theta^{(t)}) \log P(Z|Y, \theta^{(t)}) dZ \\
= & \int_Z P(Z|Y, \theta^{(t)}) \log \frac{P(Z|Y, \theta^{(t+1)})}{P(Z|Y, \theta^{(t)})} dZ \\
= & -KL(P(Z|Y, \theta^{(t)}) \| P(Z|Y, \theta^{(t+1)})) \leq 0
\end{aligned}
$$

因此，$\log P(Y|\theta^{(t)}) \leq \log p(Y|\theta^{(t+1)})$.

广义的EM算法

根据 $\log P(Y|\theta) = ELBO + KL(Q\|P) \geq ELBO$

其中： $ELBO = \int_Z Q(Z) \log \frac{P(Y,Z|\theta)}{Q(Z)} dZ = L(Q,\theta)$
$KL(Q\|P) = \int_Z Q(Z) \log \frac{Q(Z)}{P(Z|Y,\theta)} dZ$

EM思想: 固定$\theta^{(t)}$, 计算后验$P(Z|Y,\theta^{(t)})$，然后通过极大化ELBO得到新的$\theta^{(t+1)}$

如果后验$P(Z|Y,\theta^{(t)})$很难计算，可以通过极大化ELBO找到一个合适的Q,即:

$$\hat{Q}(Z) = \arg\max_Q L(Q,\theta)$$

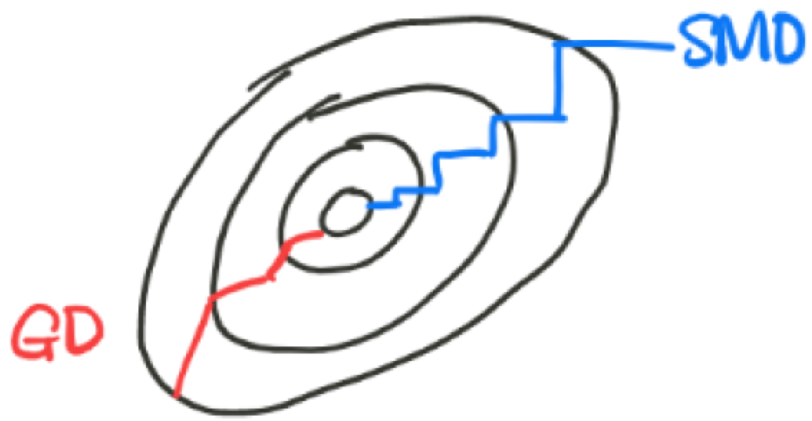
广义的EM算法为：

E-Step: $Q^{(t+1)} = \arg\max_Q L(Q, \theta^{(t)})$

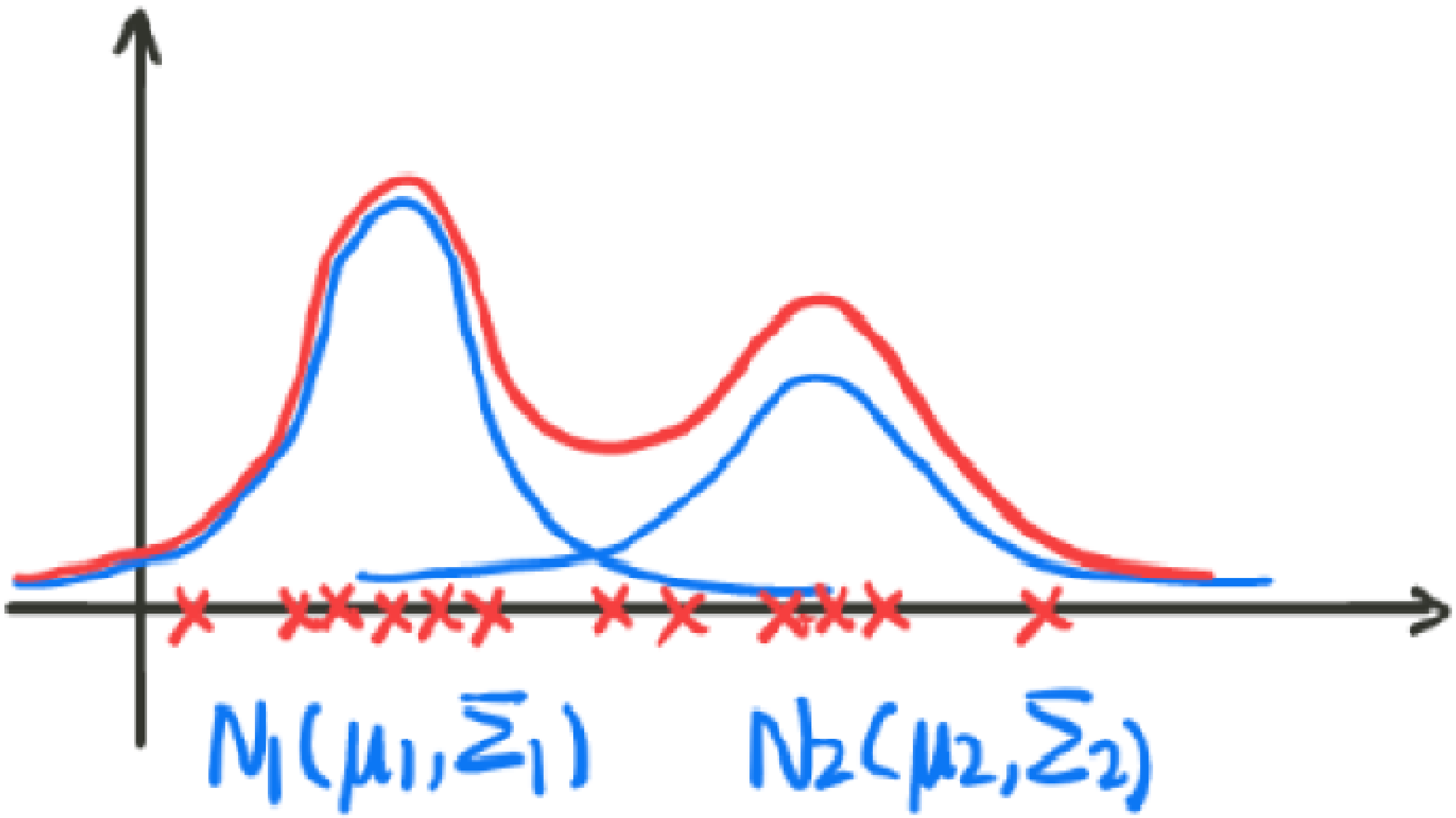
M-Step: $\theta^{(t+1)} = \arg\max_\theta L(Q^{(t+1)}, \theta)$

$$\begin{aligned} L(Q,\theta) &= \int_Z Q(Z) \log \frac{P(Y,Z|\theta)}{Q(Z)} dZ \\ &= E_Q[\log P(X,Z|\theta)] - E_Q[\log Q] \end{aligned}$$

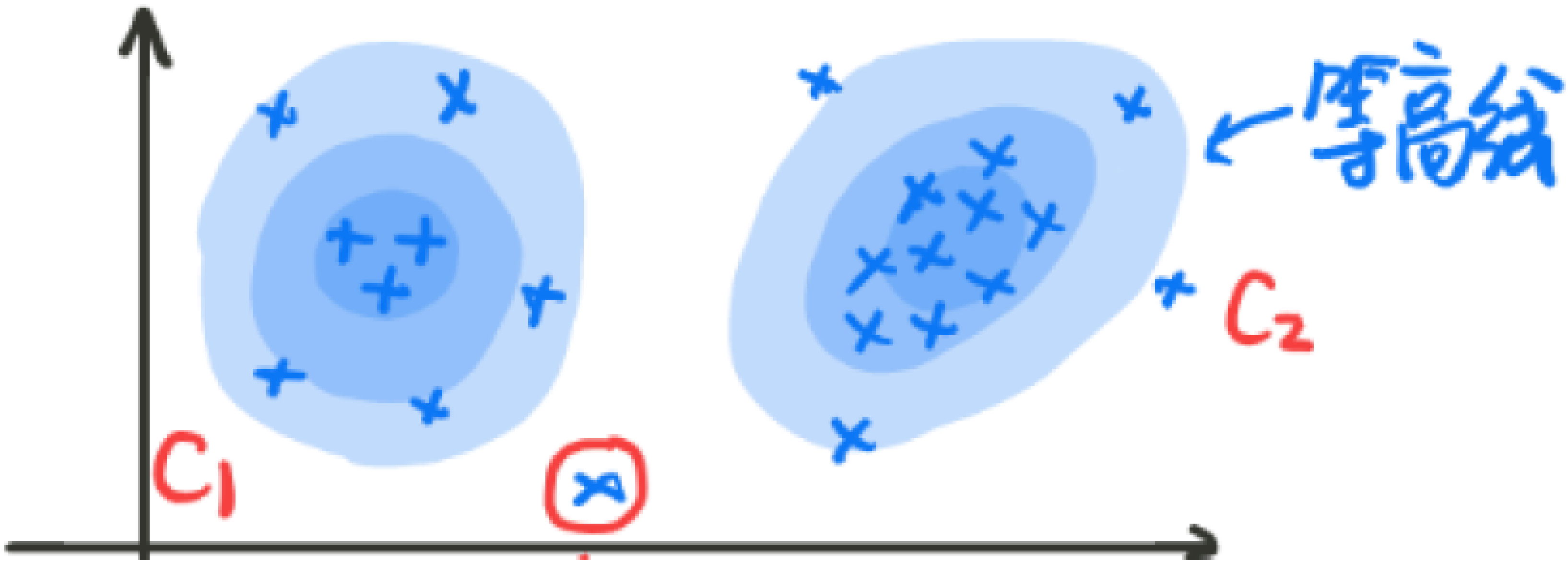
从这里可以看出EM是广义EM的一种特殊情况，也可以看成坐标上升法。



高斯混合模型(Gaussian Mixture Model, GMM)

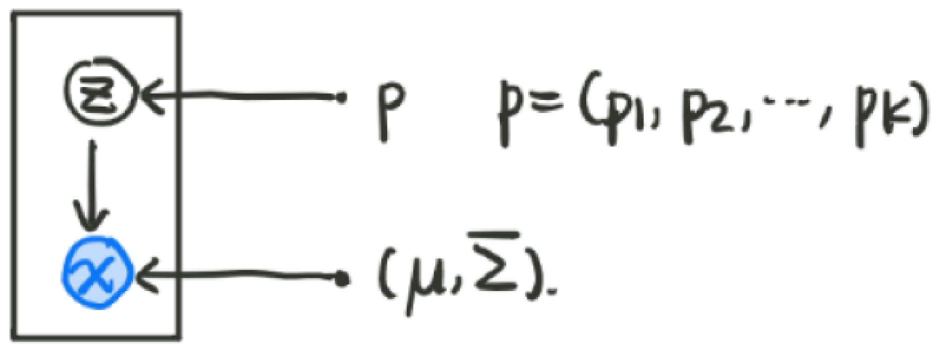


$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(x|\mu_k, \sigma_k), \qquad \sum_{k=1}^{K} \alpha_k = 1$$



引入隐变量Z, 根据其取值，决定选哪个高斯分布。

Z	C_1	C_2	$\cdots$	C_k
$P(Z)$	P_1	P_2	$\cdots$	P_k



我们根据一个离散的随机变量Z来选择是选取那个高斯分布，利用这个高斯分布$\mathcal{N}(\mu;\sigma)$来采样
得到我们想要的样本点。而且，离散随机变量Z符合一个离散分布$p = (p_1, p_2, \cdots, p_k)$。

我们想使用极大似然估计来求解GMM的最优参数结果。

$$X = (x_1, x_2, \cdots, x_N)$$

$$(X, Z) = \{(x_1, z_1), (x_2, z_2), \cdots, (x_N, z_N)\}$$

$$\theta = \{P_1, \cdots, P_k, \mu_1, \cdots, \mu_k, \Sigma_1, \cdots, \Sigma_k\}$$

$$\begin{aligned} P(X) &= \sum\nolimits_Z P(X, Z) \\ &= \sum\nolimits_{k=1}^K P(X, Z = C_k) \\ &= \sum\nolimits_{k=1}^K P(Z = C_k) \cdot P(X|Z = C_k) \\ &= \sum\nolimits_{k=1}^K P_k \cdot \mathcal{N}(X|\mu_k, \Sigma_k) \end{aligned}$$

其中，P_k也就是数据点去第k个高斯分布的概率。

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_\theta \log P(X) \\
&= \arg\max_\theta \log \prod_{i=1}^{N} P(x_i) \\
&= \arg\max_\theta \sum_{i=1}^{N} \log P(x_i) \\
&= \arg\max_\theta \sum_{i=1}^{N} \log \sum_{k=1}^{K} P_k \cdot \mathcal{N}(x_i | \mu_k, \Sigma_k)
\end{aligned}
$$

直接使用MLE 求解GMM，无法得到解析解。

EM算法的表达式，：

$$\theta^{(t+1)} = \arg\max_\theta \underbrace{\mathbb{E}_{P(Z|X,\theta^{(t)})}\left[\log P(X,Z|\theta)\right]}_{Q(\theta,\theta^{(t)})}$$

- E-Step:

$$
\begin{aligned}
Q(\theta,\theta^{(t)}) &= \int_Z \log P(X,Z|\theta)\cdot P(Z|X,\theta^{(t)})dZ \\
&= \sum_Z \log \prod_{i=1}^N P(X_i,Z_i|\theta)\cdot \prod_{i=1}^N P(Z_i|X_i,\theta^{(t)})dZ \\
&= \sum_{Z_1,\cdots,Z_N} \sum_{i=1}^N \log P(X_i,Z_i|\theta)\cdot \prod_{i=1}^N P(Z_i|X_i,\theta^{(t)})dZ \\
&= \sum_{Z_1,\cdots,Z_N} [\log P(X_1,Z_1|\theta) + \log P(X_2,Z_2|\theta) + \cdots + \log P(X_N,Z_N|\theta)] \\
&\quad \cdot \prod_{i=1}^N P(Z_i|X_i,\theta^{(t)})dZ \\
&= \sum_{i=1}^N \sum_{Z_i} \log P(X_i,Z_i|\theta)\cdot P(Z_i|X_i,\theta^{(t)}) \\
&= \sum_{i=1}^N \sum_{Z_i} \log P_{Z_i}\cdot \mathcal{N}(X_i|\mu_{Z_i},\Sigma_{Z_i})\cdot P(Z_i|Xs_i,\theta^{(t)})
\end{aligned}
$$

- M-Step: $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{(t)})$

考虑计算$P_K^{(t+1)}$

$$\begin{cases} \arg\max_{P_k} \sum_{k=1}^K \sum_{i=1}^N \log P_k \cdot P(Z_i = C_k | X_i, \theta^{(t)}) \\ s.t. \quad \sum_{k=1}^K P_k = 1 \end{cases}$$

使用拉格朗日算子法，我们可以写成：

$$\mathcal{L}(P, \lambda) = \sum_{k=1}^K \sum_{i=1}^N \log P_k \cdot P(Z_i = C_k | X_i, \theta^{(t)}) + \lambda(\sum_{k=1}^K P_k - 1)$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}(P,\lambda)}{\partial P_k} = & \quad \textstyle\sum_{i=1}^{N} \frac{1}{P_k} \cdot P(Z_i = C_k | X_i, \theta^{(t)}) + \lambda = 0 \\
\Rightarrow & \quad \textstyle\sum_{i=1}^{N} P(Z_i = C_k | X_i, \theta^{(t)}) + P_k \lambda = 0 \\
\xRightarrow{k=1,\cdots,K} & \quad \textstyle\sum_{i=1}^{N} \underbrace{\displaystyle\sum_{k=1}^{K} P(Z_i = C_k | X_i, \theta^{(t)})}_{1} + \underbrace{\displaystyle\sum_{k=1}^{K} P_k}_{1} \lambda = 0 \\
\Rightarrow & \quad N + \lambda = 0
\end{aligned}
$$

所以，我们可以轻易的得到$\lambda = -N$，所以有

$$
P_K^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} P(Z_i = C_k | X_i, \theta^{(t)})
$$