## 0、引例

一个袋子中有红球和白球，被抽到的概率分别为p和1-p，放回抽样得到观测序列为：x1,x2,x3...xn,求p的似然估计。

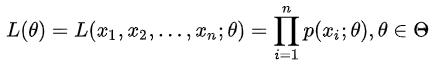
l(p) = log

然后求函数的极值点，可以解出：p=|{xi=红}|/n，也就是说：利用最大似然方法，估计值就等于红色朝上的次数/总的实验次数。

## 1、极大似然估计

假设一致分布的模型，但是模型的参数不知道，根据观测序列来估计模型的方法，极大似然估计就是极大化观测序列出现概率求模型参数。

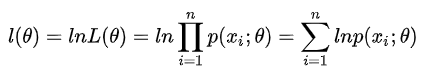
假设观测序列为X，待估计的参数为θ。



目标就是：



然后：



然后求函数的零点即为所求。

## 2、Jensen不等式

定义：如果函数时凹函数(二阶导>=0)，那么E(f(x)) >= f(E(x))。函数的期望大于等于期望的函数。

应用1：KL散度(相对熵)非负



## 3、EM算法

### 3.1 EM应用案例：

观测序列：x1,x2,...xn；总的先验参数θ=(π,p,q)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 红色 先验 | 白色 先验 | 红色 联合 | 白色 联合 |
| 袋子1（π） | P | 1-p | πp | π(1-p) |
| 袋子2（1-π） | q | 1-q | (1-π)q | (1-π)(1-q) |

假如我们已知所有的先验参数(π0,P0,q0)，那么可以计算后验概率表为：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 后验概率  (条件概率密度) | x1 | x2 | ...... | xn |
| P(袋子1|红色) | =a1 | a2 | ...... | an |
| P(袋子1|百色) | =b1 |  |  |  |
| P(袋子2|红色) | =c1 |  |  |  |
| P(袋子2|白色) | =d1 |  |  |  |
| Sum | 1 | 1 | 1 | 1 |

问题1：x1取自第一个袋子的概率是多少？

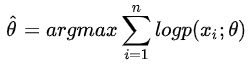
答案1：a1+b1

模型参数的估计值：(xi=1代表实际为红色)

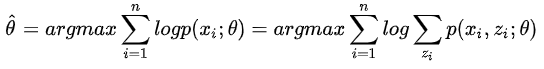
思想：将上一轮的后验作为下一轮的先验。迭代到收敛。

**3.2 EM理论**

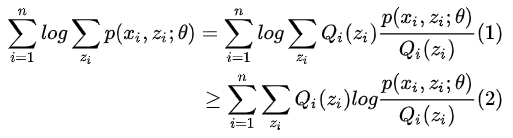
对于n个观测x1,x2,...xn，找出样本的模型参数θ, 极大化模型分布的对数似然函数如下：



加入隐变量后为：

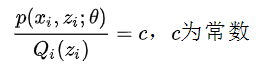


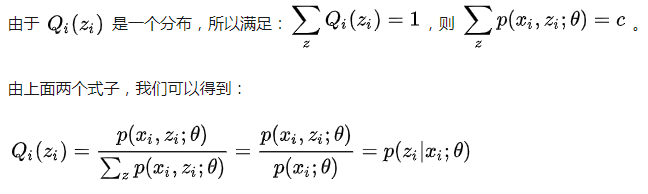
由于包含和的对数以及因变量，所以无法直接给出解析解，尝试jensen缩放：



这在数学上是成立的，至于Q(zi)是什么，后面会分析。

至此，我们找到了一个难解函数的下界，但是下界增大，是原函数增大既不必要也不充分条件，所以有必要进一步增加约束：就是两个函数至少有一个交点，那么这对Q有什么要求呢？根据jensen不等式，相等意味着变量为常数，即：

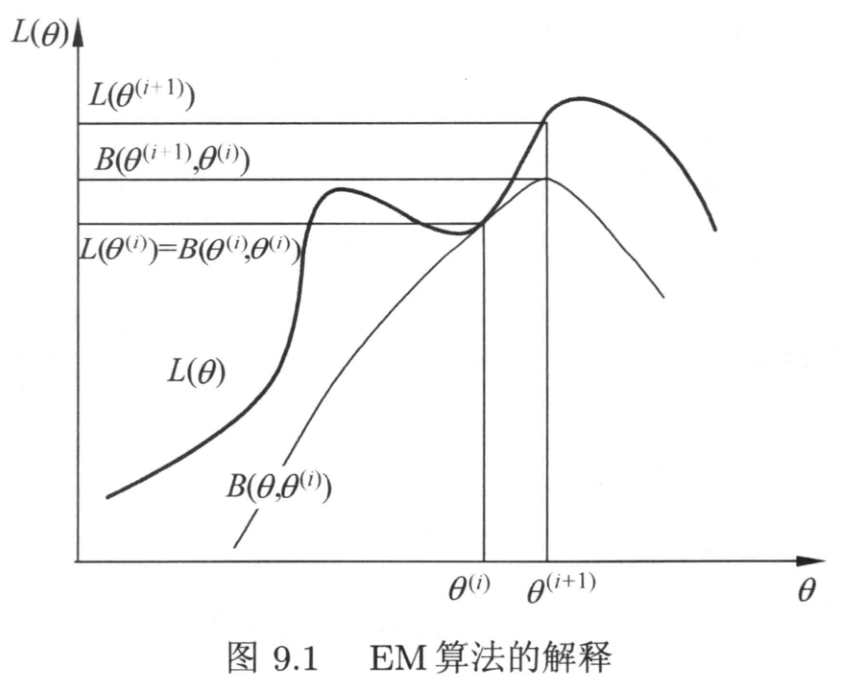




为什么不直接用联合概率？因为不知道zi是什么。**至此，小结一下：**

1、当Q为隐变量的后验概率的时，序列似然概率的下界与之至少交于一点

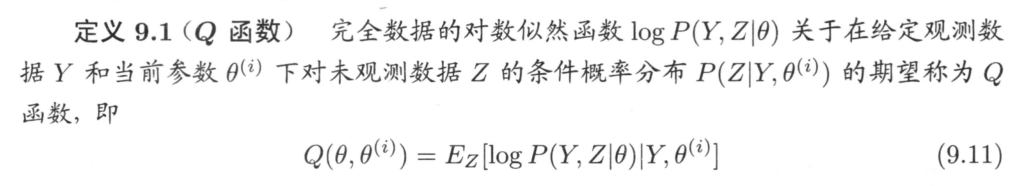
2、下界函数可以直接优化求解

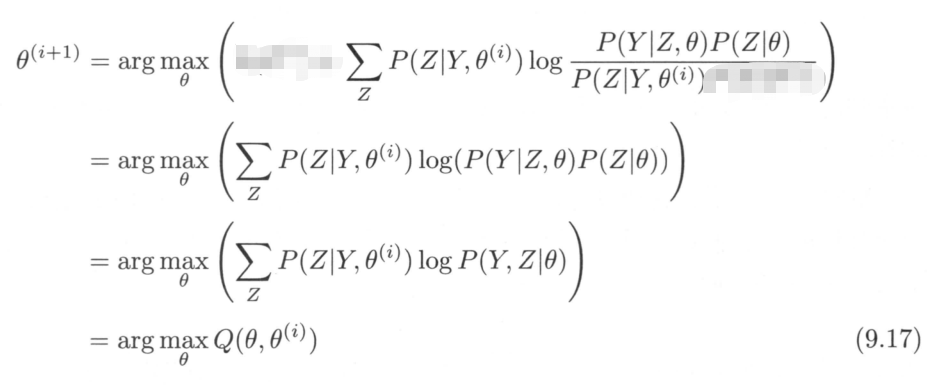


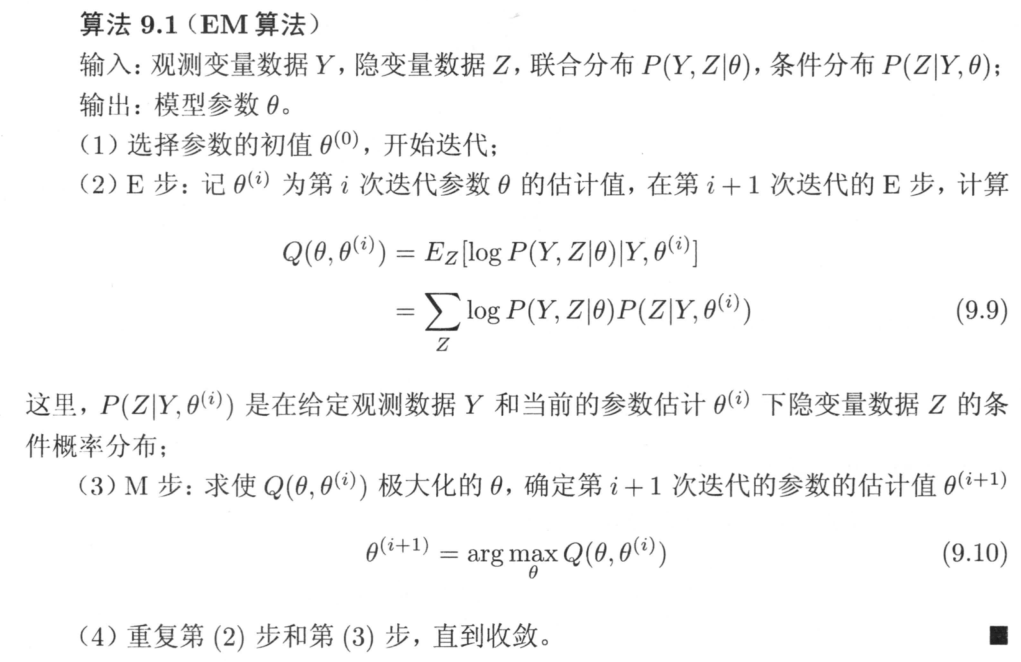
可以看到，下界函数收敛代表EM算法会收敛，单数不一定是极值点，更不用说最值点。

### 3.3 Q函数与EM导出

下面是小蓝书上的，符号有一点不一样，Y对应前面分析的X(观测序列)







最后，理论层融合工程层面，θi+1是什么？

基于公式，求似然函数的极值，这个是可以计算出来的，但是实际应用的时候很简单， 从3.1的案例可以看出，更新的方法就是，逐个的遍历观测序列然后就可以在线更新各个公式的分子分母，便利完毕，参数就更新完毕了。

参考1: <https://zhuanlan.zhihu.com/p/40991784>(EM算法详解)

参考2：《统计学习导论》