

09 EM算法

隐变量、证据下界与交替优化

■ 课堂主线 | 第 1—4 部分

- 1 GMM 与隐变量问题：区分可见观测与缺失的分量标签
- 2 ELBO 与标准 EM：理解 E 步贴紧、M 步抬高的下界机制
- 3 GMM 计算与边界：计算责任度和参数更新，识别退化风险
- 4 变分 EM：精确后验不可处理时，转向受限族中的近似 E 步

■ 课后完整路径 | 接续第 5 部分

- 补全必修推导与生成模型接口；正文的“见第 5 部分”指出接续位置

1 GMM 与隐变量问题

观测值可见，分量标签不可见

■ 模型假设

- 数据由两个高斯分量混合生成，每个样本只来自其中一个分量

■ 观测与隐变量

- x_i 是第 i 个观测值； $z_i \in \{1,2\}$ 是未记录的分量标签

■ 每个分量的参数

- μ_k 决定第 k 个高斯的中心， σ_k^2 决定其宽窄

■ 分量所占比例

- $\pi_k = p(z_i = k)$ 是样本来自分量 k 的先验概率， $\pi_1 + \pi_2 = 1$

■ 待估参数

- $\theta = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ ；本讲示例使用 $x = (0,2,8,10)$

■ 生成过程

- 先按概率 π_k 选择分量 $z = k$, 再从 $\mathcal{N}(\mu_k, \sigma_k^2)$ 中生成 x

■ 两分量 GMM 的密度

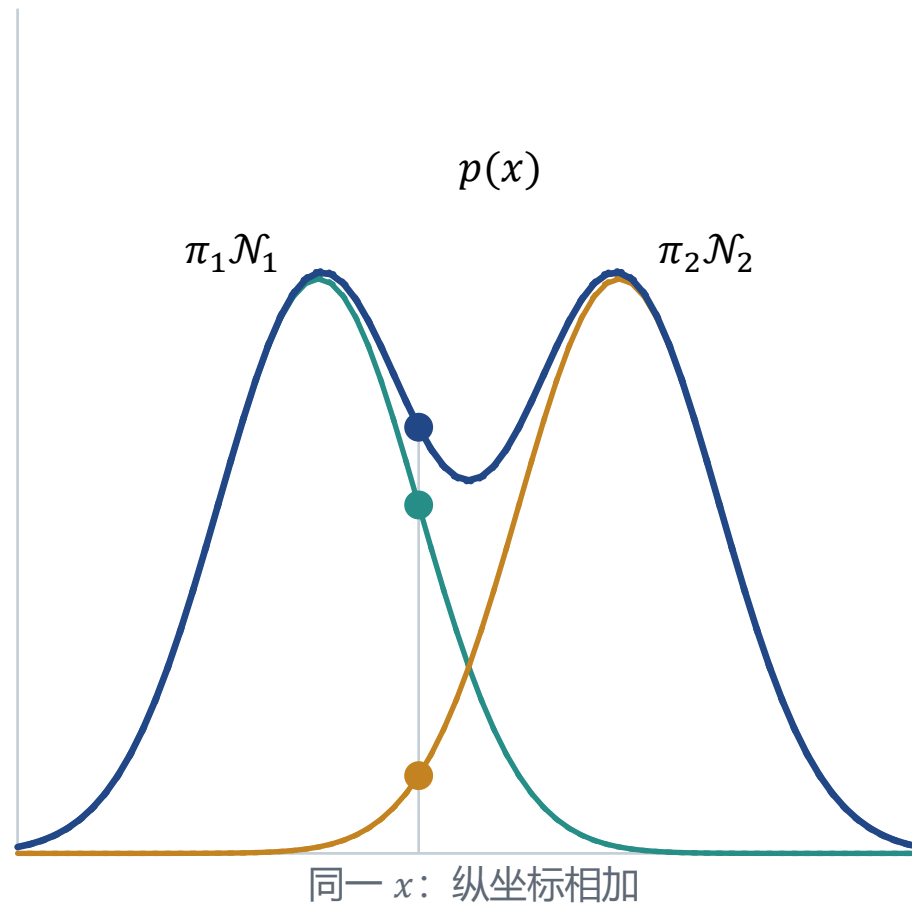
- $p(x) = \pi_1 \mathcal{N}(x | \mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(x | \mu_2, \sigma_2^2)$

■ 公式的直接含义

- 在每个 x 处, 先用 π_k 缩放第 k 条高斯曲线的高度, 再把两个高度相加

■ 图形读法

- 两条分量曲线分别由各自的均值和方差决定, 混合密度曲线表示最终观测到的分布



- **完整数据**

- 同时观察 (x_i, z_i)

- **完整数据对数似然**

- $\ell_c(\theta) = \sum_i \log p_\theta(x_i, z_i)$

- **GMM 中的结果**

- 已知分量标签后，可以分别统计各分量的样本数、均值和方差

- **真正的困难**

- 训练数据没有记录 z_i

- 单个样本的观测似然

- $p_{\theta}(x_i) = \sum_{z_i} p_{\theta}(x_i, z_i)$

- 数据集的观测对数似然

- $\ell(\theta) = \sum_i \log \sum_{z_i} p_{\theta}(x_i, z_i)$

- 连续隐变量

- 对 z_i 的求和相应改为积分

- 目标不变

- 仍然希望选择使观测数据对数似然较大的参数 θ

■ 为什么交替

- 分量参数决定样本的后验软权重，参数更新又需要这些权重给完整数据状态加权；后验加权梯度见第 5 部分

■ E 步

- 固定当前参数，计算样本级的后验软权重 $r_{ik} = p(z_i = k \mid x_i, \theta^t)$

■ M 步

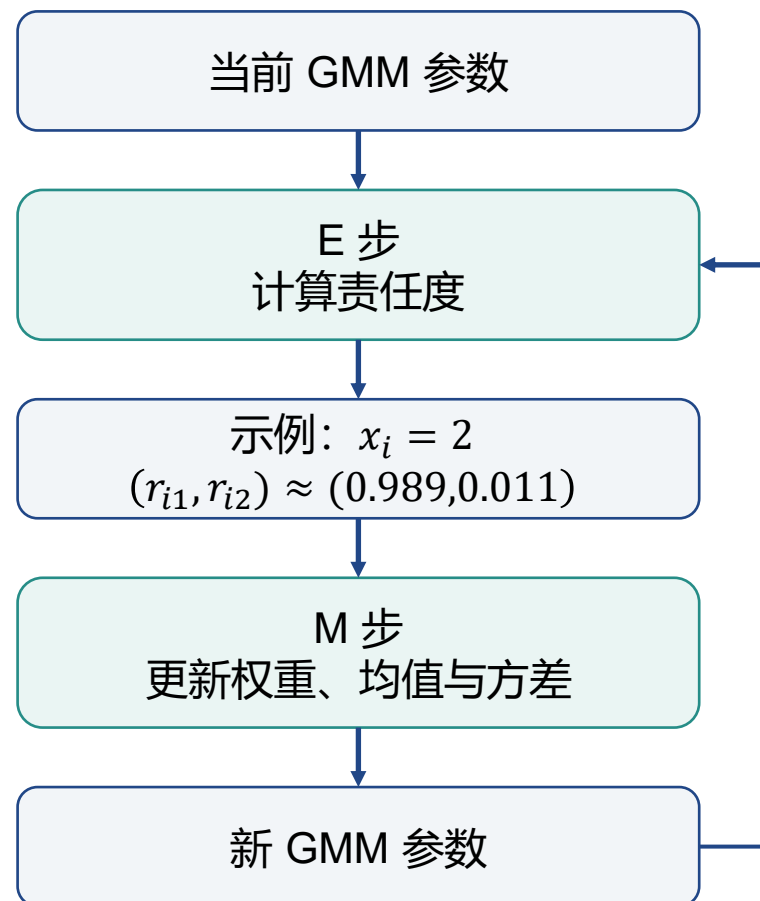
- 固定 r_{ik} ，重新估计混合权重 π_k 、均值和方差

■ 两种“权重”不同

- r_{ik} 针对“样本 i 属于分量 k ”； π_k 针对“总体中分量 k 占多少”

■ 重复

- 新参数返回下一轮 E 步



2 ELBO 与标准 EM

构造下界，再解释 E 步贴紧与 M 步抬高

为什么引入辅助分布 q

■ 当前障碍

- 隐状态 z 不可见，观测似然中的“对数外求和”让分量归属与参数更新耦合

■ 先回到 GMM

- 一般写作 $q(z)$ ；在 GMM 中，每个样本都有分量标签分布 $q_i(z_i)$ ，其中 $q_i(k)$ 是分量 k 的候选软解释，精确 E 步选 $q_i(k) = r_{ik}$

■ 需要一种软解释

- 一般记为 $q(z)$ ：各项之和为 1；联合概率为正时 $q(z) > 0$ ，但不要求等于真实后验

■ 在求和中乘除同一项

- $p_{\theta}(x) = \sum_z q(z) \frac{p_{\theta}(x, z)}{q(z)}$

■ 关键转折：改写为期望

- $p_{\theta}(x) = E_q \left[\frac{p_{\theta}(x, z)}{q(z)} \right]$

■ 从观测证据开始

- $\log p_{\theta}(x) = \log E_q \left[\frac{p_{\theta}(x,z)}{q(z)} \right]$

■ 对数凹性给出下界

- $\log p_{\theta}(x) \geq E_q \left[\log \frac{p_{\theta}(x,z)}{q(z)} \right]$

■ 定义证据下界

- $\mathcal{L}(q, \theta) = \sum_z q(z) \log \frac{p_{\theta}(x,z)}{q(z)}$

■ 所得关系

- 对任何满足条件的 q , 都有 $\mathcal{L}(q, \theta) \leq \log p_{\theta}(x)$

■ 课后推导

- 第 5 部分给出 Jensen 的加权形式、代入条件和逐步推导

■ 加入后验关系

- $p_{\theta}(z | x) = p_{\theta}(x, z) / p_{\theta}(x)$

■ KL 间隙的中间形式

- $D_{KL}(q \| p_{\theta}(z | x)) = \log p_{\theta}(x) - \mathcal{L}(q, \theta)$

■ 移项得到

- $\log p_{\theta}(x) = \mathcal{L}(q, \theta) + D_{KL}(q(z) \| p_{\theta}(z | x))$

■ KL 的性质

- $D_{KL} \geq 0$, 因此 ELBO 不超过真实对数证据

■ 间隙的含义

- q 越接近当前后验, 下界越接近真实目标

■ 等式依据

- 把后验定义代入 KL, 再展开对数; 第 5 部分补全这段代数及 ELBO 的两项组成

■ 贴紧选择

- $q(z) = p_{\theta}(z | x)$

■ KL 间隙消失

- $D_{KL}(q \| p_{\theta}(z | x)) = 0$

■ 当前参数处相等

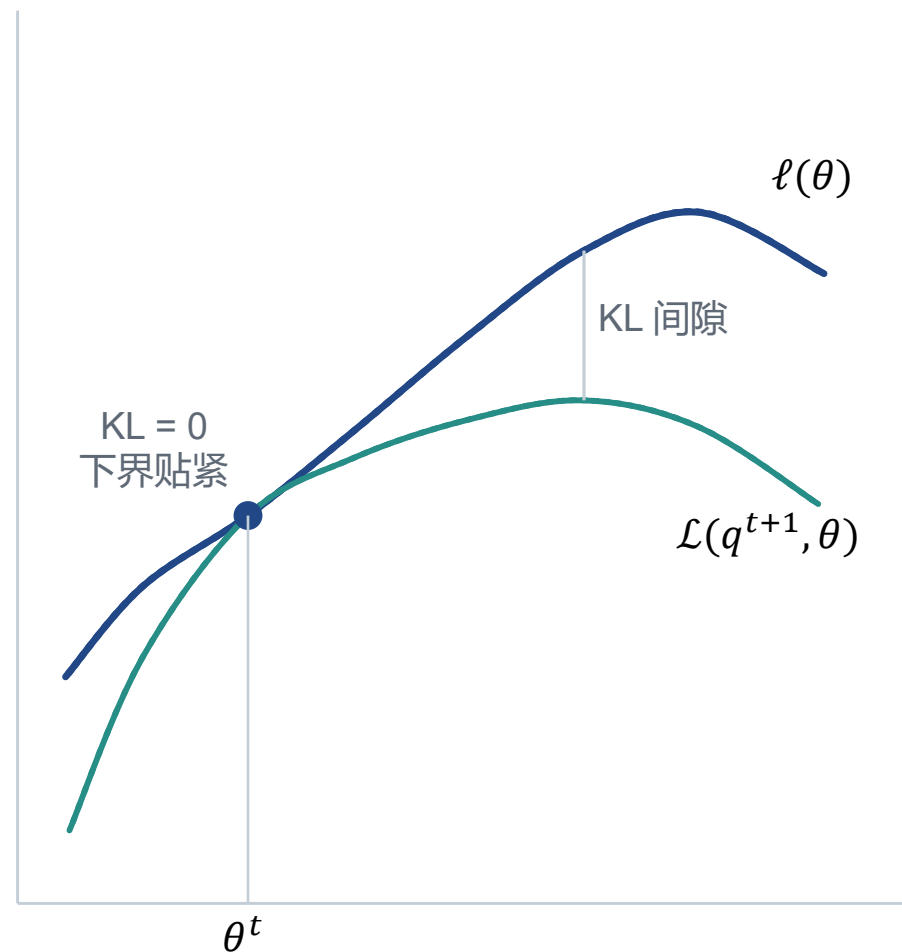
- $\mathcal{L}(q, \theta) = \log p_{\theta}(x)$

■ 回到 GMM

- 对样本 x_i , $q_i(k) = r_{ik} = p_{\theta}(z_i = k | x_i)$

■ 从单个样本到数据集

- 独立样本的目标逐样本相加: $\mathcal{L}_{\text{data}}(q, \theta) = \sum_{i=1}^n \mathcal{L}_i(q_i, \theta)$, $q(z_{1:n}) = \prod_{i=1}^n q_i(z_i)$



■ 初始化

- 选择模型参数 θ^0

■ 精确 E 步

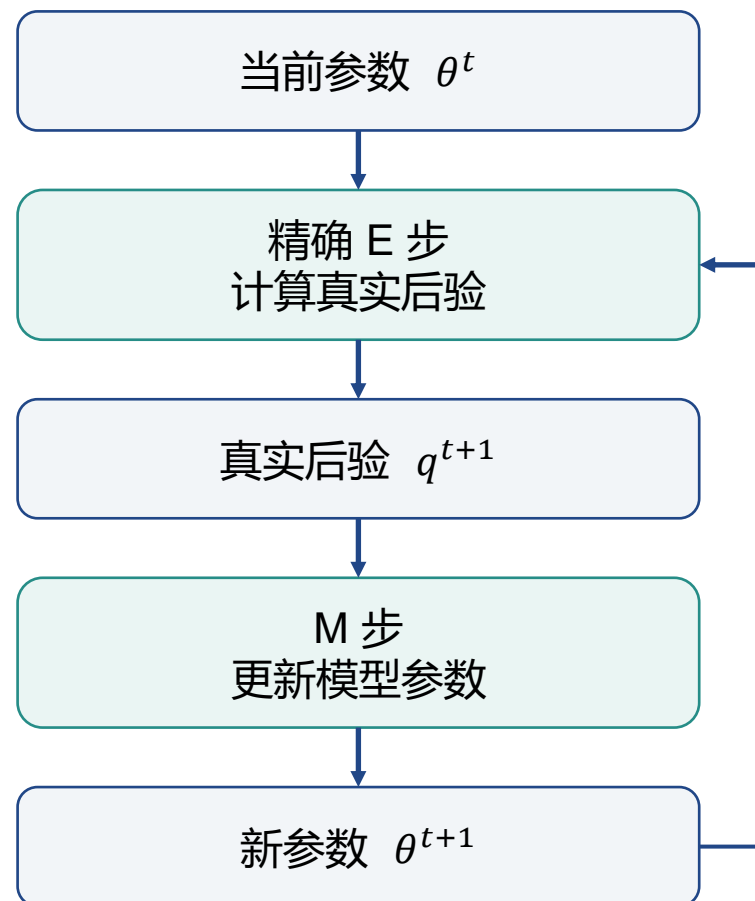
- 固定 θ^t , 计算当前真实后验: $q^{t+1}(z) = p_{\theta^t}(z | x)$

■ 标准 M 步

- 固定 q^{t+1} , 选择使 ELBO 最大的参数: $\theta^{t+1} = \operatorname{argmax}_{\theta} \mathcal{L}(q^{t+1}, \theta)$

■ 广义 EM 的边界

- 若 M 步只保证 Q 函数或 ELBO 不下降, 而不要求达到最大值, 则称为广义 EM



■ 旧参数处

- 精确 E 步选择当前后验，使 ELBO 与观测对数似然相等

■ 固定当前后验

- M 步把参数从 θ^t 更新到 θ^{t+1} ，ELBO 不下降

■ 新参数处

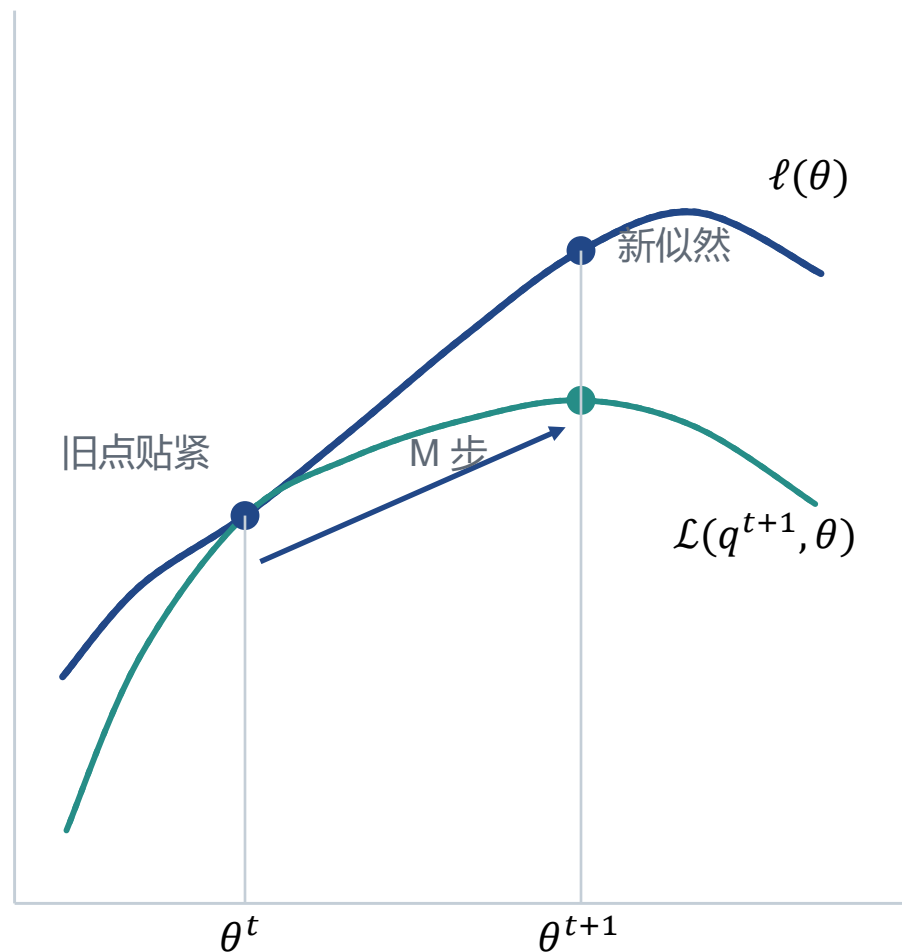
- 观测对数似然仍不低于对应 ELBO

■ 观察顺序

- 旧点贴紧，沿下界移动，再读取新似然

■ 证明接续

- 第 5 部分把这三段图形关系写成完整不等式链，这是本讲必修证明



3 GMM 一轮计算与实践边界

用责任度更新参数，再检查初始化与退化风险

- 观测数据

- $x = (0, 2, 8, 10)$

- 初始混合权重

- $\pi_1 = \pi_2 = 0.5$

- 初始均值与方差

- $\mu_1 = 2, \mu_2 = 8; \sigma_1^2 = \sigma_2^2 = 4$

- 本轮任务

- 计算责任度，再更新权重、均值和方差

■ E 步公式

- $$r_{ik} = p(z_i = k \mid x_i, \theta^t) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)}{\sum_j \pi_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2)}$$

■ 样本 $x_i = 2$

- 分量 1 的密度中心就在 2, 分量 2 的均值距离为 6
- 代入当前参数得到 $r_{i1} \approx 0.989$, $r_{i2} \approx 0.011$

■ 读取方式

- 两个责任度总和为 1, 但它们是软分配而不是已观察标签

样本	分量 1 责任度	分量 2 责任度	主要归属
0	0.999	0.001	分量 1
2	0.989	0.011	分量 1
8	0.011	0.989	分量 2
10	0.001	0.999	分量 2

■ 这里的 Q 是什么

- 固定 E 步得到的责任度后，对完整数据对数联合概率取后验加权平均，作为 M 步目标；一般形式见第 5 部分

■ 固定 E 步责任度，展开完整数据目标

- $Q(\theta, \theta^t) = \sum_i \sum_k r_{ik} [\log \pi_k + \log \mathcal{N}(x_i | \mu_k, \sigma_k^2)]$

■ 责任度成为软计数

- 对分量 k , $N_k = \sum_i r_{ik}$: 每个样本按后验概率部分计入

■ 分别优化参数

- 权重由软计数决定；均值是责任度加权平均；方差是责任度加权离差

■ 本页只建立对应关系

- 后面给出一轮具体更新，不展开完整求导

■ 每个分量的软计数

- $N_k = \sum_i r_{ik}$: 每个样本按责任度部分计入分量 k

■ 本例结果

- 由对称性得到 $N_1 \approx N_2 \approx 2$

■ 统计含义

- N_k 是分量 k 承担的有效样本数, 也是权重和均值更新的分母

■ 混合权重

- $\pi_k^{new} = \frac{N_k}{n}$
- 本例仍约为 0.5 与 0.5

■ 均值更新

- $\mu_k^{new} = \frac{\sum_i r_{ik} x_i}{N_k}$

■ 本例结果

- $\mu_1^{new} \approx 1.04, \mu_2^{new} \approx 8.96$

■ 更新公式

- $$(\sigma_k^2)^{new} = \frac{\sum_i r_{ik}(x_i - \mu_k^{new})^2}{N_k}$$

■ 更新顺序

- 方差公式使用本轮刚更新的均值

■ 本例结果

- 两个分量的新方差都约为 1.28

■ 含义

- 远离新均值且责任度较大的样本贡献更大的加权离差

■ 更新前

- 均值为 2 与 8，方差均为 4

■ E 步结果

- 左侧样本主要支持分量 1，右侧样本主要支持分量 2

■ M 步结果

- 均值约为 1.04 与 8.96
- 方差约为 1.28，混合权重保持接近相等

■ 下一轮

- 使用新参数重新计算责任度，而不是沿用旧责任度

■ 初始化

- 不同初始化可能使 EM 收敛到不同驻点或局部最优，通常需要多次初始化并比较最终观测对数似然
- K-means 可以提供候选起点，但不保证更快或更优

■ 停止条件

- 标准 EM 检查观测对数似然，变分 EM 检查 ELBO
- 也可检查参数变化和最大迭代次数

■ 数值计算

- 责任度计算常使用对数域方法减少下溢

■ 退化路径

- 一个分量的均值靠近单个样本，协方差趋近于零

■ 似然后果

- 该样本处的高斯密度可以不断增大，使最大似然问题失去有限最优解

■ 常用保护

- 对协方差加入正则项或设置方差下限

■ 边界结论

- EM 的单调性不替代模型约束和数值保护

4 变分 EM 与生成模型接口

精确后验不可处理时，在受限分布族中近似 E 步

- **标准 EM 的要求**

- E 步必须计算 $p_{\theta}(z | x)$

- **失效条件**

- 后验的归一化、期望或表示本身不可处理

- **不能继续做的事**

- 无法令任意候选 q 精确等于真实后验

- **需要改变的对象**

- 把 q 限制在可计算的分布族中，并直接优化 ELBO

■ 受限集合

- 选择能够计算和优化的分布族 Q
- 花体 Q 表示候选分布族；普通大写 $Q(\theta, \theta^t)$ 表示 M 步目标函数

■ 变分 E 步

- $q^{t+1} = \operatorname{argmax}_{q \in Q} \mathcal{L}(q, \theta^t)$

■ 输出

- 得到可处理的近似后验，而不要求等于真实后验

■ 直接边界

- 若真实后验不属于 Q ，KL 间隙通常不能降为零

■ 变分 E 步

- 固定 θ , 在 Q 中更新 q 以提高 ELBO

■ M 步

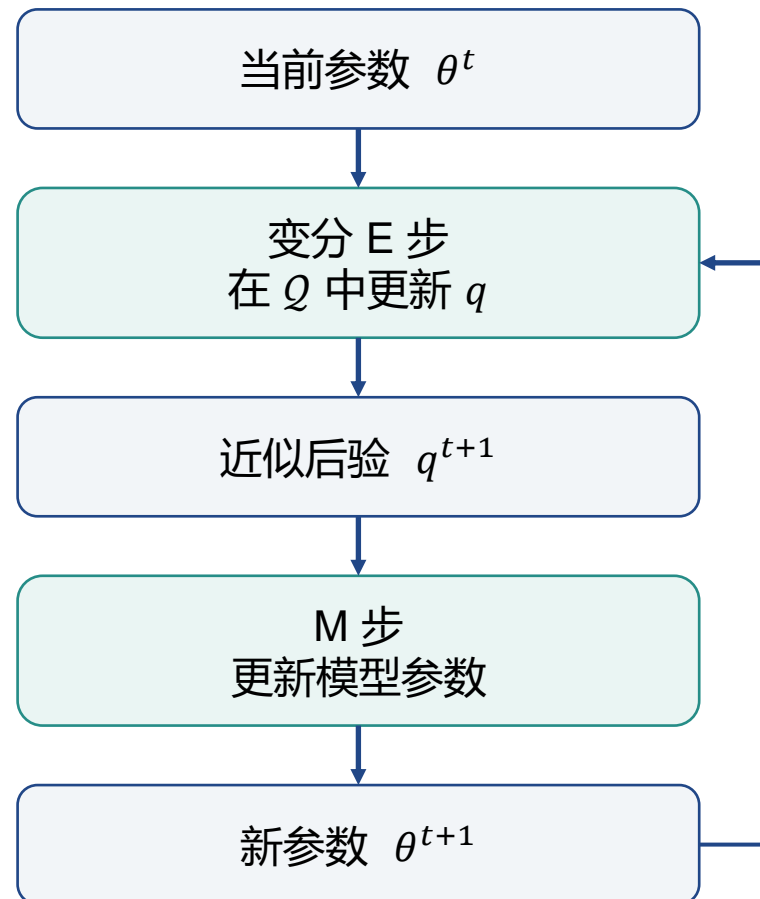
- 固定 q , 更新生成模型参数 θ 以提高 ELBO

■ 循环目标

- 两步都作用于同一个 $\mathcal{L}(q, \theta)$

■ 与标准 EM 的联系

- 若 Q 包含真实后验且 E 步达到它, 就恢复精确贴紧



对照项	标准 EM	变分 EM
E 步输出	当前真实后验	受限族中的近似后验
KL 间隙	当前参数处为 0	通常大于 0
直接保证不下降的目标	观测对数似然	ELBO
是否保证真实似然不下降	是	通常不保证
成立条件	E 步精确; M 步使 Q 函数不下降	两个更新步骤均使同一 ELBO 不下降
都不保证	全局最优或参数序列必然收敛	全局最优或真实似然逐步提高

- **问题**：隐变量让观测对数似然出现对数外求和
- **下界**：Jensen 给出 ELBO，真实目标与它相差非负 KL
- **标准 EM**：精确 E 步贴紧下界，M 步提高固定下界
- **GMM**：责任度形成软计数，支持权重、均值和方差更新
- **推广与边界**：变分 EM 近似 E 步并提高 ELBO；两种 EM 都不保证全局最优
- **课后继续第 5 部分**：补全推导、单调性证明和生成模型接口，供课后阅读和复习

5 必修推导与生成模型接口（课后阅读）

课后接续第 2 部分的下界机制、第 3 部分的加权更新和第 4 部分的推广

■ 回看第 1 部分的交替困难

- 对数外求和可求导，但完整状态的贡献权重随当前参数变化

■ 观测似然的梯度

- $\nabla_{\theta} \log p_{\theta}(x_i) = \sum_k r_{ik}(\theta) \nabla_{\theta} \log p_{\theta}(x_i, z_i = k)$

■ 当前后验权重

- $r_{ik}(\theta) = p_{\theta}(z_i = k \mid x_i) = \frac{p_{\theta}(x_i, z_i=k)}{p_{\theta}(x_i)}$

■ 这说明什么

- 参数决定样本的软归属，而分量参数的更新又需按这份软归属加权

■ 回看第 2 部分的 ELBO 下界

- 这里补全从改写后的观测证据到下界的不等式依据

■ 加权平均的条件

- 设 $a_z > 0$, $q(z) \geq 0$, 且 $\sum_z q(z) = 1$

■ 对数凹性

- $\log(\sum_z q(z)a_z) \geq \sum_z q(z)\log a_z$

■ 本讲的代入

- 取 $a_z = p_\theta(x, z)/q(z)$, 左边是 $\log p_\theta(x)$, 右边是 ELBO; 联合概率为正的状态需有 $q(z) > 0$

- 回看第 2 部分的下界定义

- 展开 $\log[p_\theta(x, z)/q(z)]$ 得到两项

- 展开结果

- $\mathcal{L}(q, \theta) = E_q[\log p_\theta(x, z)] + H(q)$

- 期望完整数据对数联合概率

- 评价模型怎样解释由 q 加权的完整数据状态

- 辅助分布的熵

- $H(q) = -E_q[\log q(z)]$, 保留下界中的分布项; 固定 q 做 M 步时它对 θ 是常数

- **代入后验定义：** 用 $p_{\theta}(z | x) = p_{\theta}(x, z)/p_{\theta}(x)$ 展开 KL

- $D_{KL}(q \| p_{\theta}(z | x)) = \sum_z q(z) \log \frac{q(z)p_{\theta}(x)}{p_{\theta}(x, z)}$

- **拆开证据项并识别 ELBO**

- $D_{KL} = \log p_{\theta}(x) - \sum_z q(z) \log \frac{p_{\theta}(x, z)}{q(z)}$

- $\log p_{\theta}(x) = \mathcal{L}(q, \theta) + D_{KL}(q \| p_{\theta}(z | x))$

- **回看第 2 部分的下界贴紧**

- 固定旧参数 θ^t , 让候选分布等于当前真实后验

- **更新规则**

- $q_i^{t+1}(z_i) = p_{\theta^t}(z_i | x_i)$

- **直接结果**

- 在 θ^t 处 KL 间隙为零, ELBO 等于观测对数似然

- **GMM 中的输出**

- 对每个样本计算各分量责任度, 而不是把分量标签当作已观测

- 回看第 3 部分的 GMM 加权目标

- 固定精确 E 步得到的 q^{t+1} , 只更新模型参数 θ

- ELBO 参数更新

- $\theta^{t+1} = \operatorname{argmax}_{\theta} \mathcal{L}(q^{t+1}, \theta)$

- 一般 Q 函数

- $Q(\theta, \theta^t) = E_{p(z|x, \theta^t)} [\log p_{\theta}(x, z)]$

- 为什么等价

- 熵 $H(q^{t+1})$ 固定且与 θ 无关, 因此最大化 ELBO 等价于最大化 Q ; q 是分布, Q 是函数

- 旧参数处贴紧: $\log p_{\theta^t}(x) = \mathcal{L}(q^{t+1}, \theta^t)$
- M 步提高固定下界: $\mathcal{L}(q^{t+1}, \theta^{t+1}) \geq \mathcal{L}(q^{t+1}, \theta^t)$
- 新似然不低于下界: $\log p_{\theta^{t+1}}(x) \geq \mathcal{L}(q^{t+1}, \theta^{t+1})$
- 串联得到似然不下降: $\log p_{\theta^{t+1}}(x) \geq \log p_{\theta^t}(x)$

■ 回看第 4 部分的变分 EM

- 精确后验不可处理时，用可计算分布近似它，并共同优化 ELBO

■ 生成模型

- $p_{\theta}(x, z)$ 描述隐变量与观测的联合关系

■ 近似推断模型

- $q_{\phi}(z | x)$ 给出可计算的后验近似

■ 共同目标与本讲边界

- ELBO 依赖生成参数 θ 与推断参数 ϕ ；摊销推断、随机梯度和 VAE 实现留待后续课程

