一、前情回顾 04:03

1. 前情回顾：色子与概率

- **概率基础**：以均匀色子为例，每个点数出现概率为1/6，出现特定点数（如4点）的概率可直接计算
- **复合事件概率**：出现≥5点的情况包含5点和6点两个基本事件，概率为2/6=1/3
- **参数表示**：将概率模型中的关键参数记为θ，如均匀色子中$\theta = 1 / 6$

2. 例题1:不均匀色子六点概率问题 06:26

- **模型建立**：假设六点出现概率为θ，其他点数概率均等为$(1 - \theta) / 5$
- **简化模型**：当只关注是否为最大点数时，可简化为二项分布（类似不均匀硬币）
- **概率计算**：观测到六点的概率直接为θ，非六点概率为$1 - \theta$

3. 例题2:抛硬币正反面概率问题 07:16

- **实验设定**：进行n次独立抛掷，正面出现x次，反面$n - x$次
- **联合概率**：具体观测序列（如正正反...正）的概率为$\theta^x(1 - \theta)^{n - x}$
- **充分统计量**：实际只需记录正反面次数，不需记忆具体顺序

4. 逆向思维：根据观测结果求参数 09:40

- **问题转化**：从"已知θ求概率"转为"已知观测数据求最可能θ"
- **合理性原则**：选择使实际观测结果出现概率最大的参数值
- **示例说明**：比较$\theta = 1 / 2$和$\theta = 1 / 3$哪个更可能产生现有观测数据

5. 似然函数与对数似然函数 11:26

- **似然函数定义**：$L(\theta) = P(X|\theta)$，将联合概率视为θ的函数
- **对数变换**：将连乘形式$\prod \theta^{x_i}(1 - \theta)^{1 - x_i}$转为求和形式，简化求导计算
- **函数特性**：对数变换保持单调性，不影响极值点位置

6. 最大化似然函数与极大似然估计 12:22

- **估计原理**：通过求解或$lnL(\theta)$获得参数估计
- **求解方法**：通常对对数似然函数求导并令导数为零
- **直观解释**：选择使当前观测数据"最可能发生"的参数值

7. 似然函数与离散和连续随机变量的关系 12:40

- **离散情形**：似然函数由概率质量函数构造，如$L(\theta) = \prod P(x_i|\theta)$
- **连续情形**：似然函数由概率密度函数构造，如$L(\theta) = \prod f(x_i|\theta)$
- **统一原则**：无论含概率还是密度，都按变量类型选择对应函数构建

8. 数据不完整的情况 14:18

- **完全数据局限**：MLE要求观测数据完整无缺失
- **实际问题**：当存在未观测变量或数据缺失时，MLE直接应用困难
- **解决方向**：需要引入新的方法（如EM算法）处理不完整数据问题

二、em算法介绍 14:52

1. 例题:豆沙口味调查缺失值处理 15:27

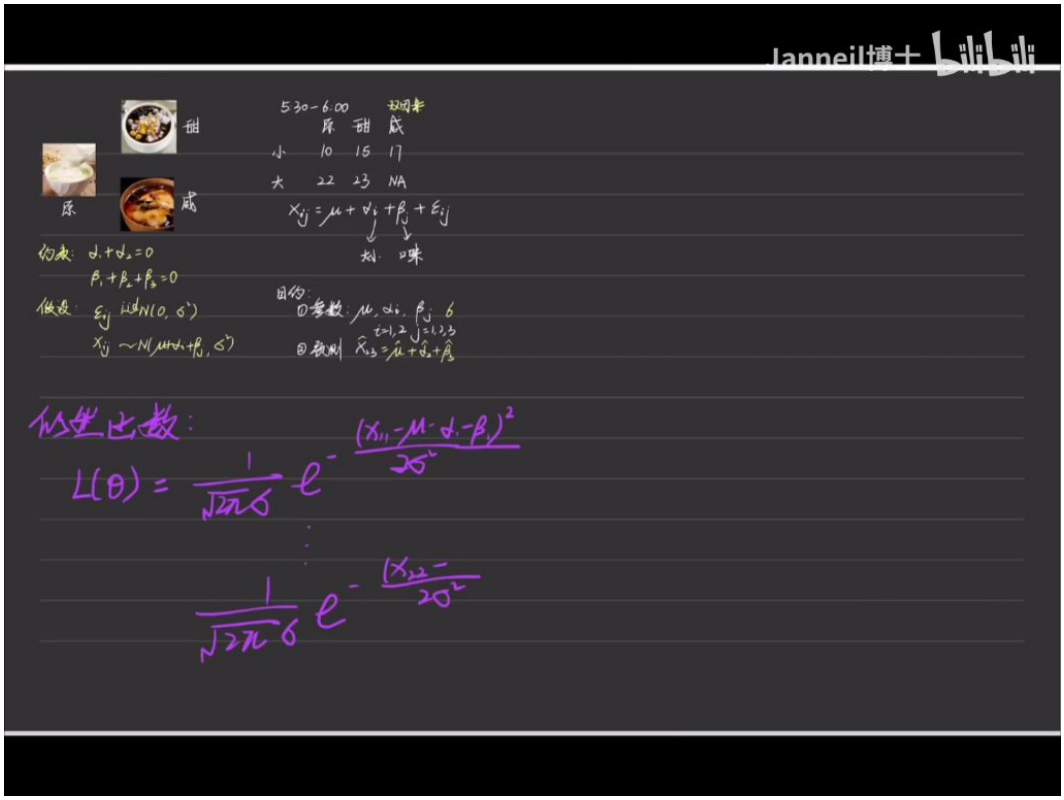
1）数据背景与模型构建

- **调查背景**：研究本地消费者对豆花口味的偏好（原味/甜味/咸味）和碗大小（小碗/大碗）对销量的影响
- **数据缺失**：大碗咸味豆花（x23）的销售数据缺失
- **双因素模型**：构建线性模型$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$，其中：
  - μ：总体均值
  - α_i：碗大小效应（i=1小碗，i=2大碗）
  - β_j：口味效应（j=1原味，j=2甜味，j=3咸味）

2）模型假设与约束条件

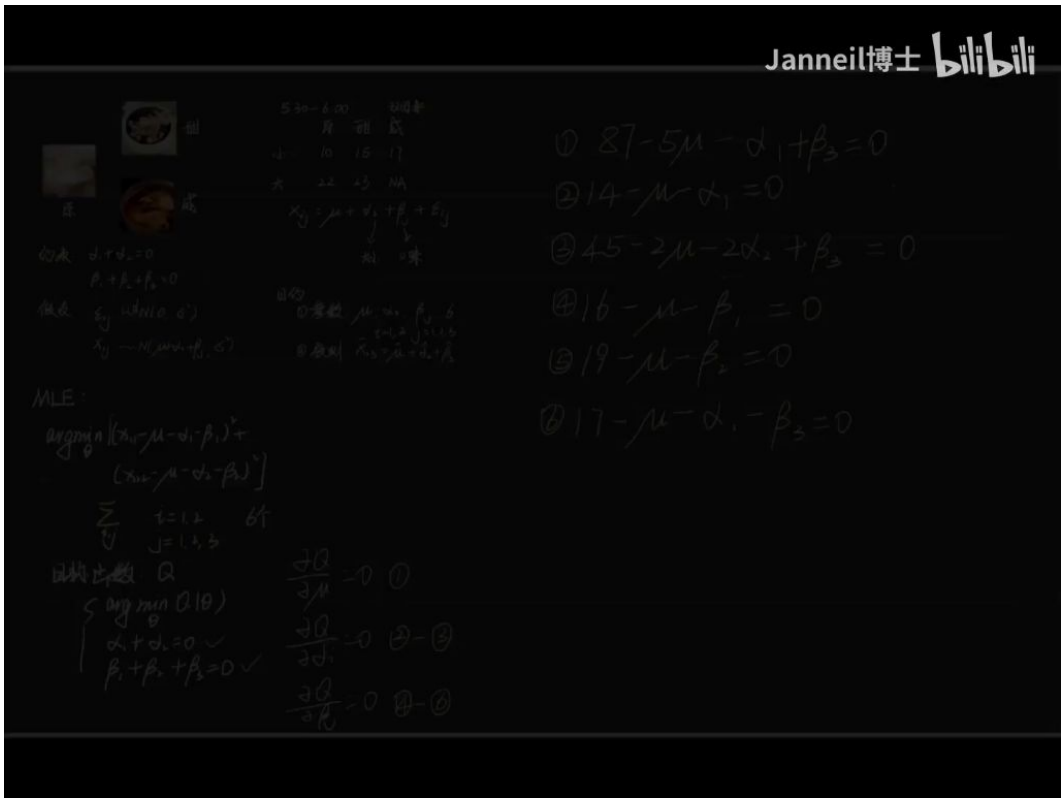
- **参数约束**：
  - $\alpha_1 + \alpha_2 = 0$（碗大小效应相互抵消）
  - $\beta_1 + \beta_2 + \beta_3 = 0$（口味效应相互抵消）
- **分布假设**：
  - 误差项$\epsilon_{ij} \sim N(0, \sigma^2)$且独立同分布
  - 因此$x_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2)$

3）极大似然估计过程



- **似然函数**：基于5个观测值的正态分布概率密度乘积
- **对数似然**：转化为最小化残差平方和$Q = \sum (x_{ij} - \mu - \alpha_i - \beta_j)^2$
- **求解方法**：对6个参数（$\mu,\alpha_1,\alpha_2,\beta_1,\beta_2,\beta_3$）求偏导并考虑约束条件

4）参数估计结果

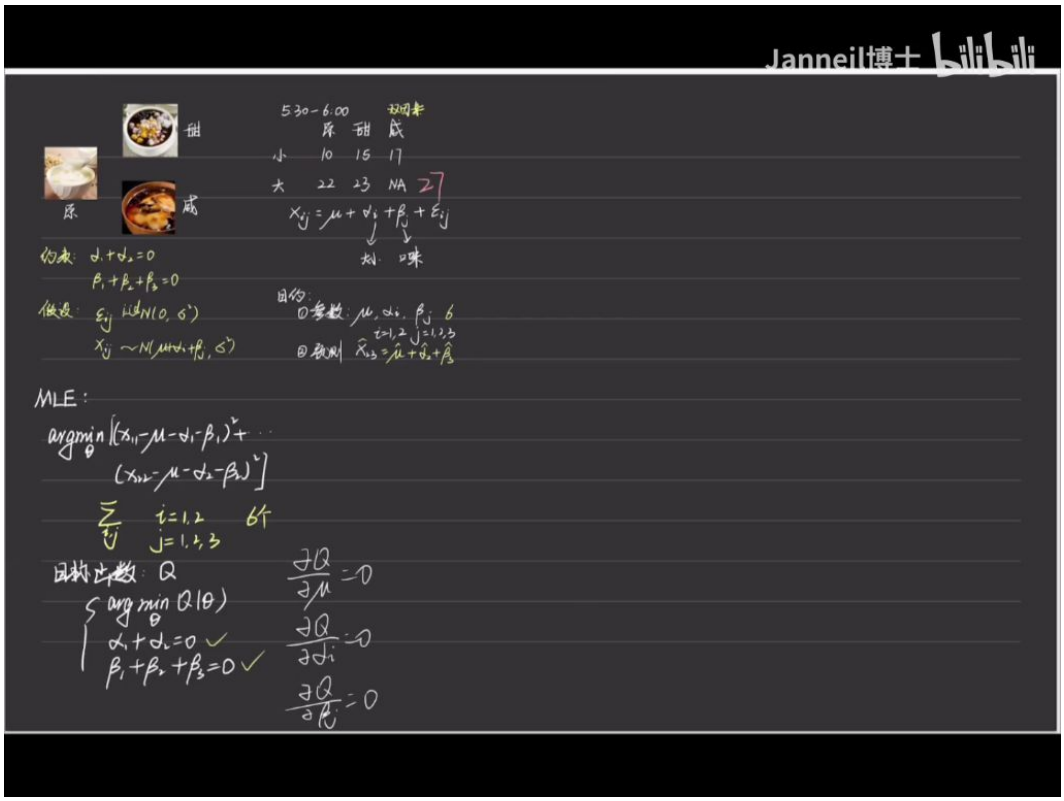


- **估计值**：
  - $\hat{\mu} = 19$

- $\hat{\alpha}_1 = -5, \hat{\alpha}_2 = 5$
- $\beta_1 = -3, \beta_2 = 0, \beta_3 = 3$

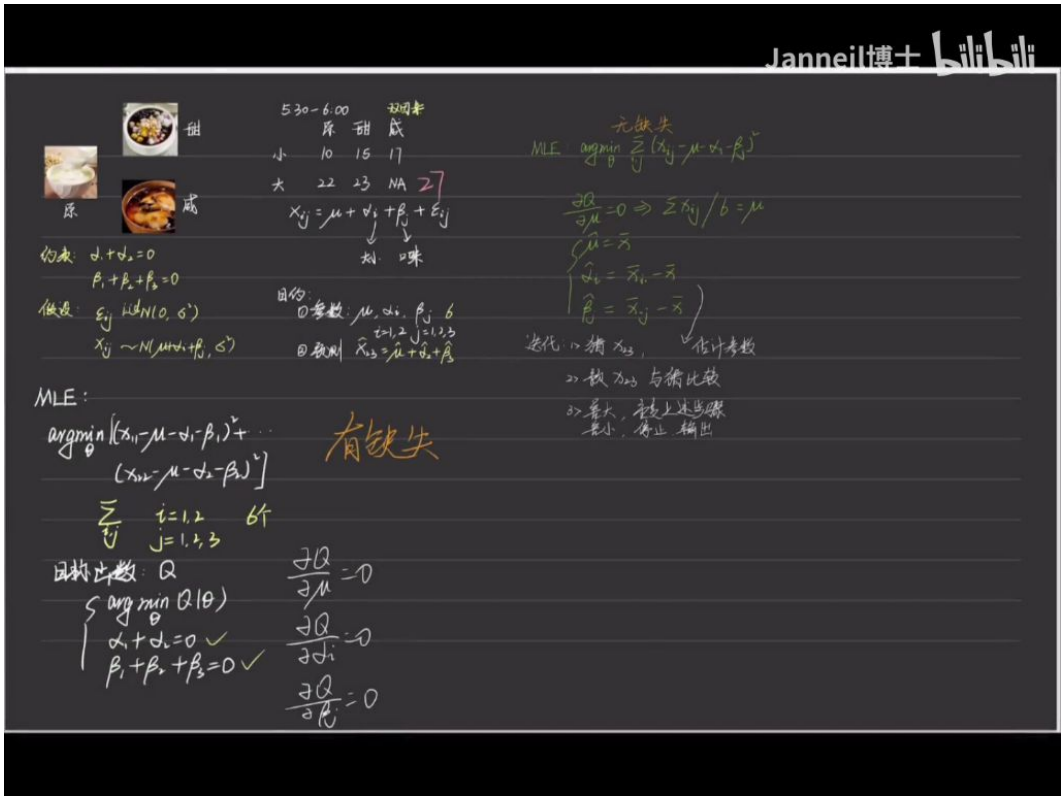
- **缺失值预测**：$\hat{x}_{23} = \hat{\mu} + \hat{\alpha}_2 + \beta_3 = 19 + 5 + 3 = 27$

5）完整数据情况对比



- **无缺失时简化计算**：
  - $\hat{\mu} = \bar{x}_{..}$（总平均值）
  - $\hat{\alpha}_i = \bar{x}_{i.} - \bar{x}_{..}$（行平均值减总平均）
  - $\hat{\beta}_j = \bar{x}_{.j} - \bar{x}_{..}$（列平均值减总平均）
- **迭代补全方法(EM算法)**：
  - 初始猜测缺失值（如用现有数据均值17.4）
  - 用完整数据方法估计参数
  - 用新参数预测缺失值
  - 重复直到收敛（20轮后得$\hat{x}_{23} = 27$）

6）方法比较



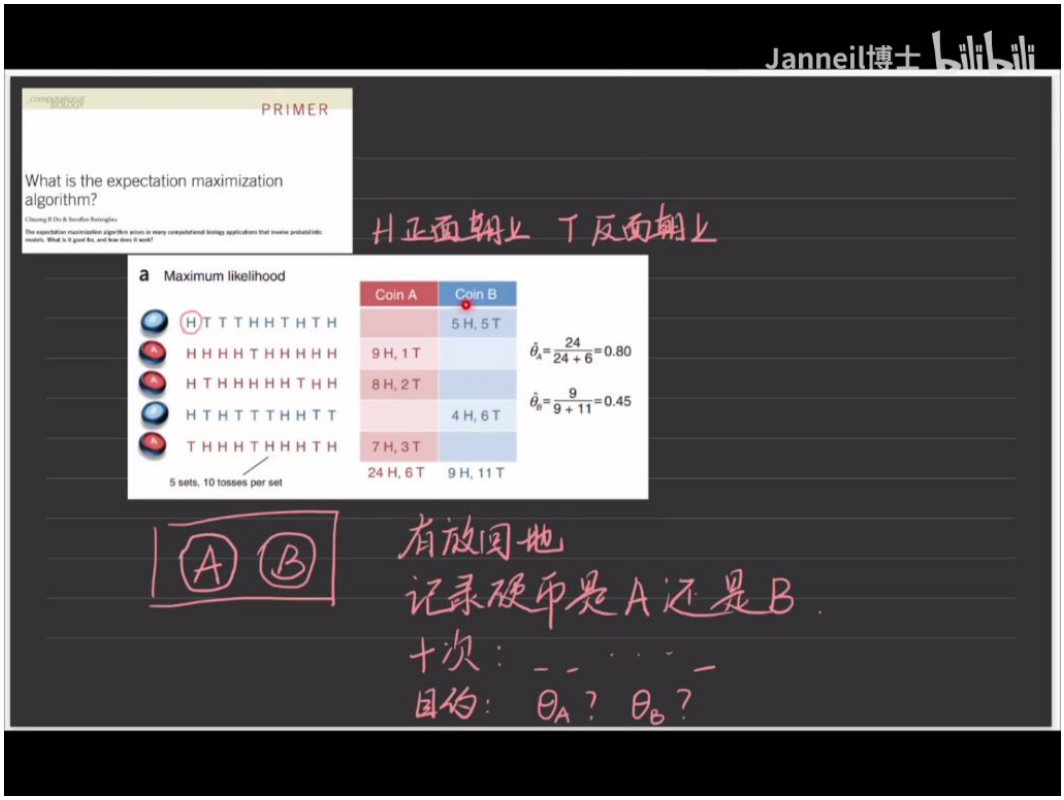
- **极大似然估计**:
  - 优点：直接得到解析解
  - 缺点：缺失值时推导复杂
- **EM算法**:
  - 优点：推导简单，适合计算机实现
  - 缺点：需要多次迭代
- **本质区别**：人工计算复杂度与计算机计算量的权衡

2. 例题:硬币模型 52:18

1）EM算法简介

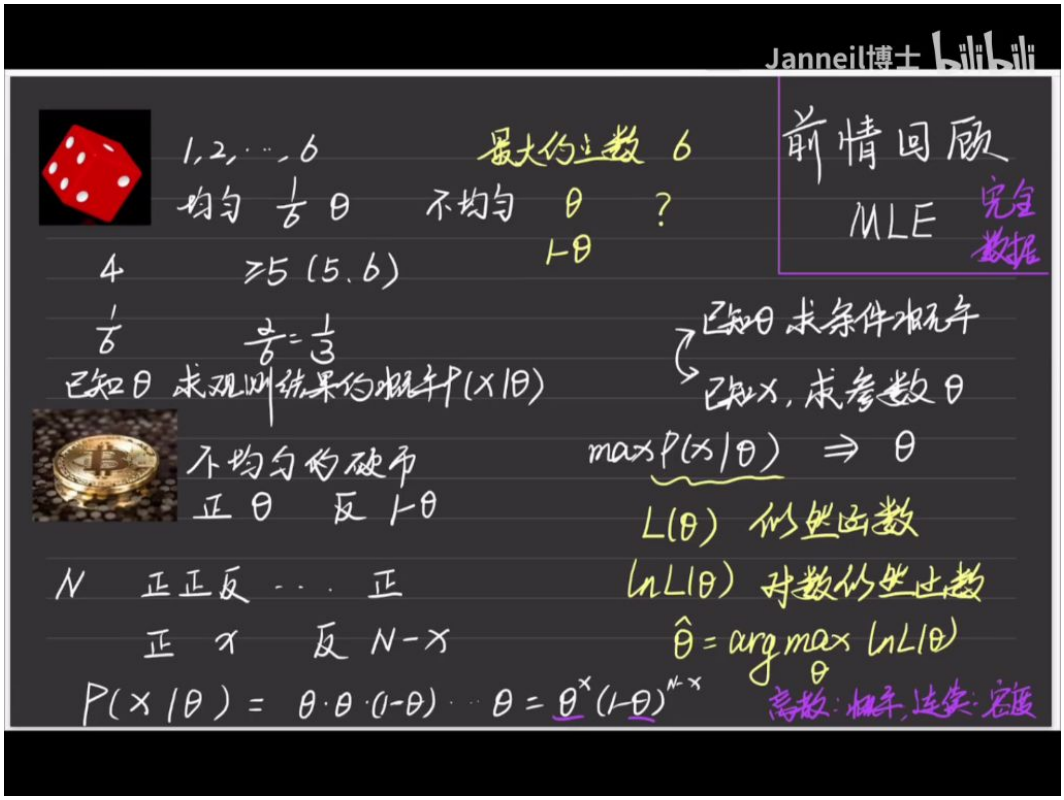
- **基本概念**:　EM(期望最大化)算法是一种处理包含隐变量的概率模型参数的迭代优化方法。
- **应用场景**:　特别适用于数据不完整或存在缺失值的情况，如硬币实验中不知道每次投掷的是哪枚硬币。
- **核心思想**: 通过交替执行E步(期望)和M步(最大化)来逐步优化参数估计。

2）硬币实验案例



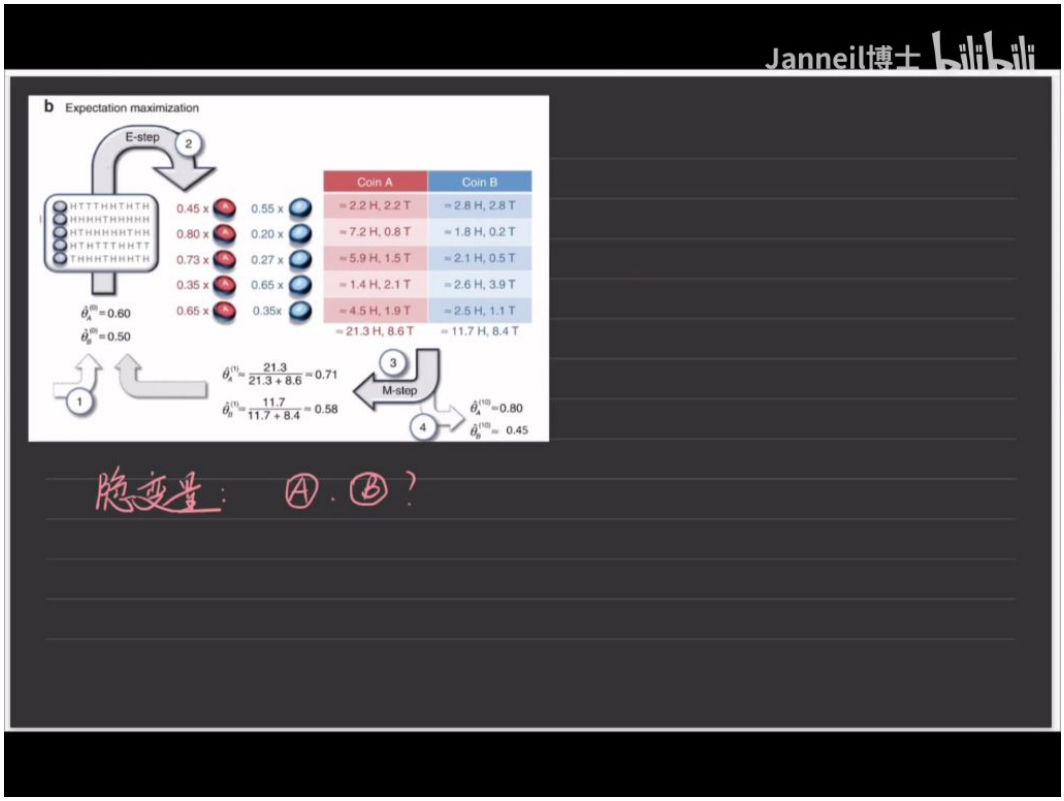
- **实验设计**:
  - 使用两枚不均匀硬币A和B
  - 每次随机抽取一枚硬币投掷10次
  - 共进行5组实验(每组10次投掷)
- **参数定义**:
  - θ_A: 硬币A正面朝上的概率
  - θ_B: 硬币B正面朝上的概率
- **完全数据情况**:
  - 当知道每次投掷的是哪枚硬币时，可直接用频率估计概率
  - $\theta_A = \frac{24}{24+6} = 0.80$
  - $\theta_B = \frac{9}{9+11} = 0.45$

3）最大似然估计回顾



- **似然函数**: 对于n次投掷中x次正面的情况，似然函数为$L(\theta) = \theta^{x}(1 - \theta)^{n - x}$
- **对数似然**: 取对数后求导可得$\theta = \frac{x}{n}$
- **与频率估计的关系**: 最大似然估计结果与简单频率统计结果一致

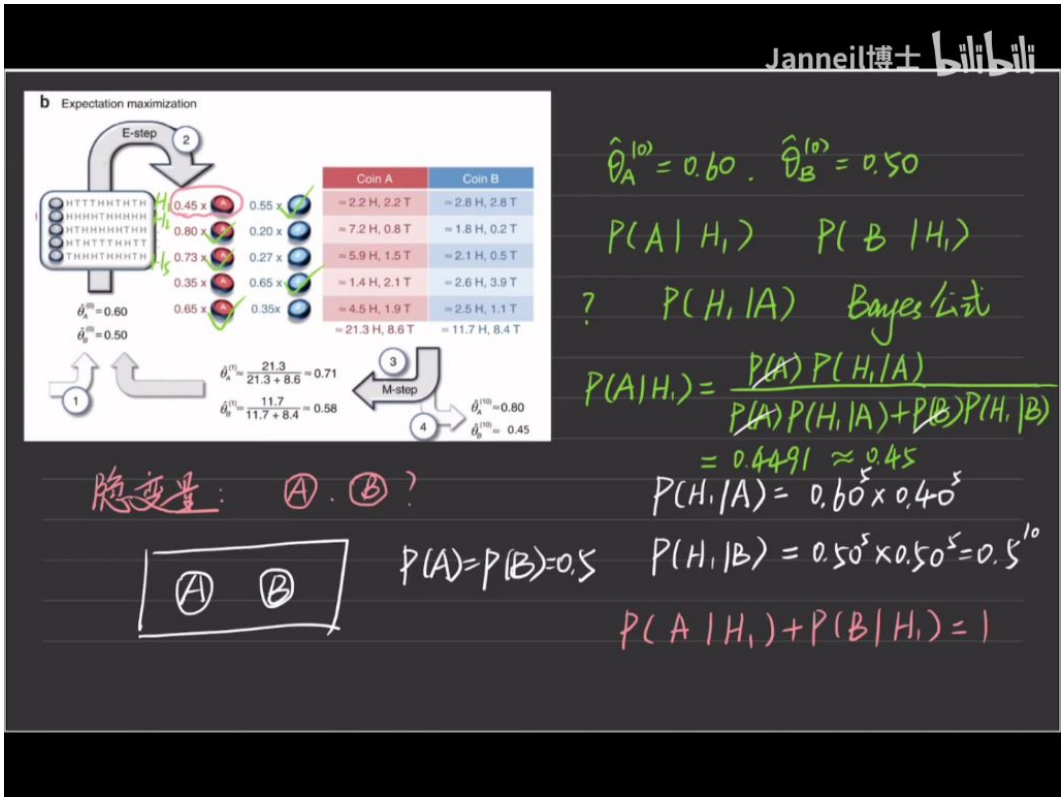
4）隐变量问题



- **关键难点**: 当不知道每次投掷的是哪枚硬币时，存在隐变量(硬币类型)
- **解决方法**:

- 先猜测初始参数值($\theta_A = 0.6, \theta_B = 0.5$)
- 使用贝叶斯公式计算每组数据属于各硬币的概率
- 例：$P(A|H_1) = \frac{P(H_1|A)P(A)}{P(H_1|A)P(A) + P(H_1|B)P(B)} \approx 0.45$

5）EM算法步骤

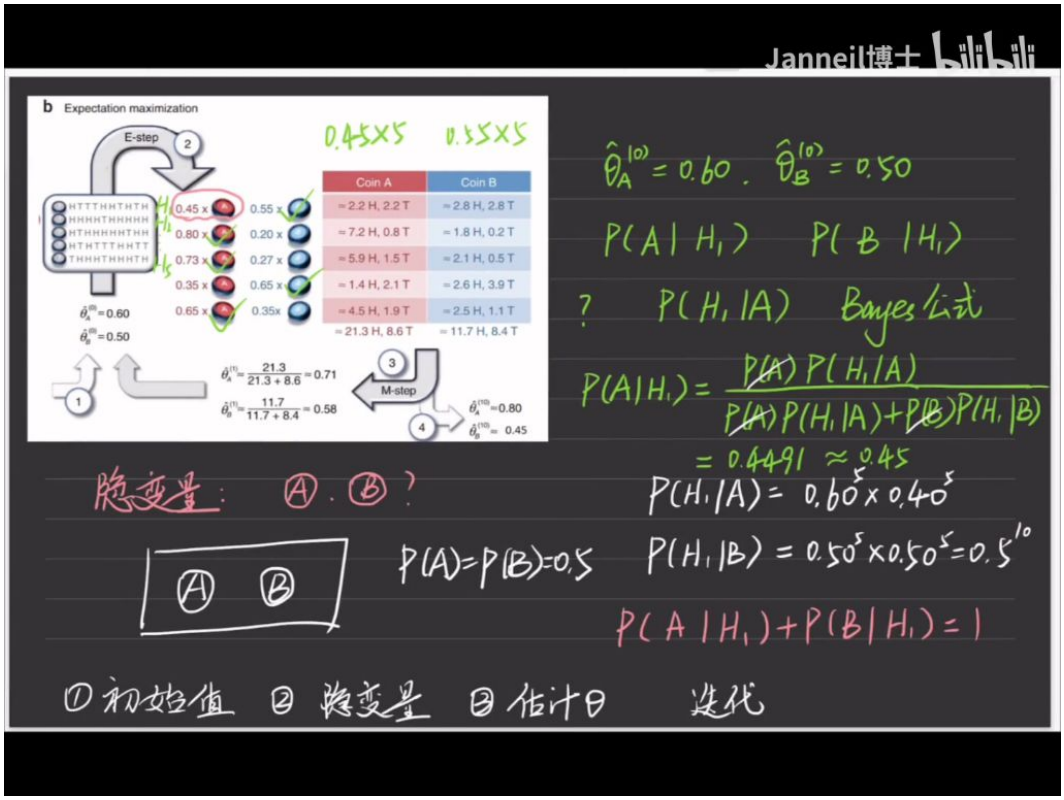


- **E步(期望步)**:
  - 计算隐变量的后验概率
  - 用小数表示"部分归属"(如2.2次正面属于A硬币)
  - 统计各硬币的期望正面/反面次数
- **M步(最大化步)**:
  - 用期望次数重新估计参数
  - $\theta_A^{(1)} = \frac{21.3}{21.3 + 8.6} = 0.71$
  - $\theta_B^{(1)} = \frac{11.7}{11.7 + 8.4} = 0.58$
- **迭代过程**: 重复E步和M步直到收敛(本例10次迭代后收敛到$\theta_A = 0.80, \theta_B = 0.45$)

6）EM算法形式化描述

- **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
- **步骤**:
  - 选择初始参数$\theta^{(0)}$
  - E步: 计算$Q(\theta,\theta^{(i)}) = E_Z[logP(Y,Z|\theta)|Y,\theta^{(i)}]$
  - M步:
  - 重复直到收敛

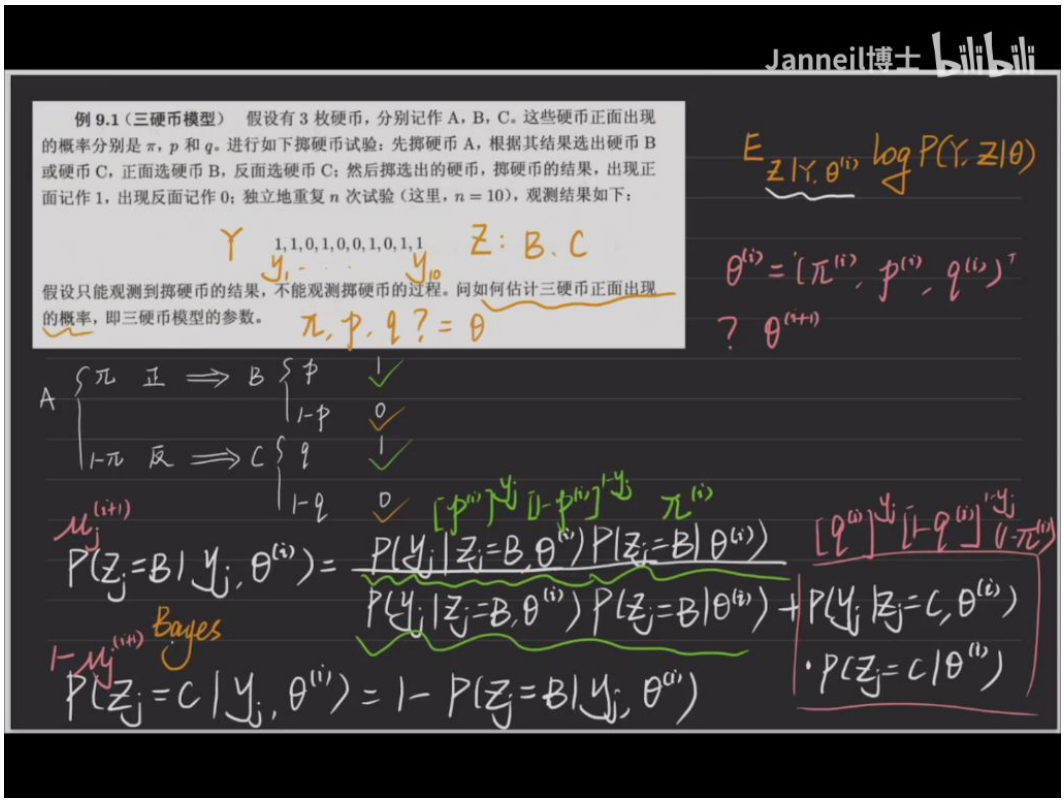
7）算法原理



- **数学基础**: 通过Jensen不等式将对数似然函数转化为可优化的下界
- **收敛性**: 每次迭代保证似然函数不减小，最终收敛到局部最优
- **与完全数据关系**: 当隐变量已知时，EM算法退化为普通最大似然估计

3. em算法 01:13:16

1）例题：三硬币模型 01:27:19

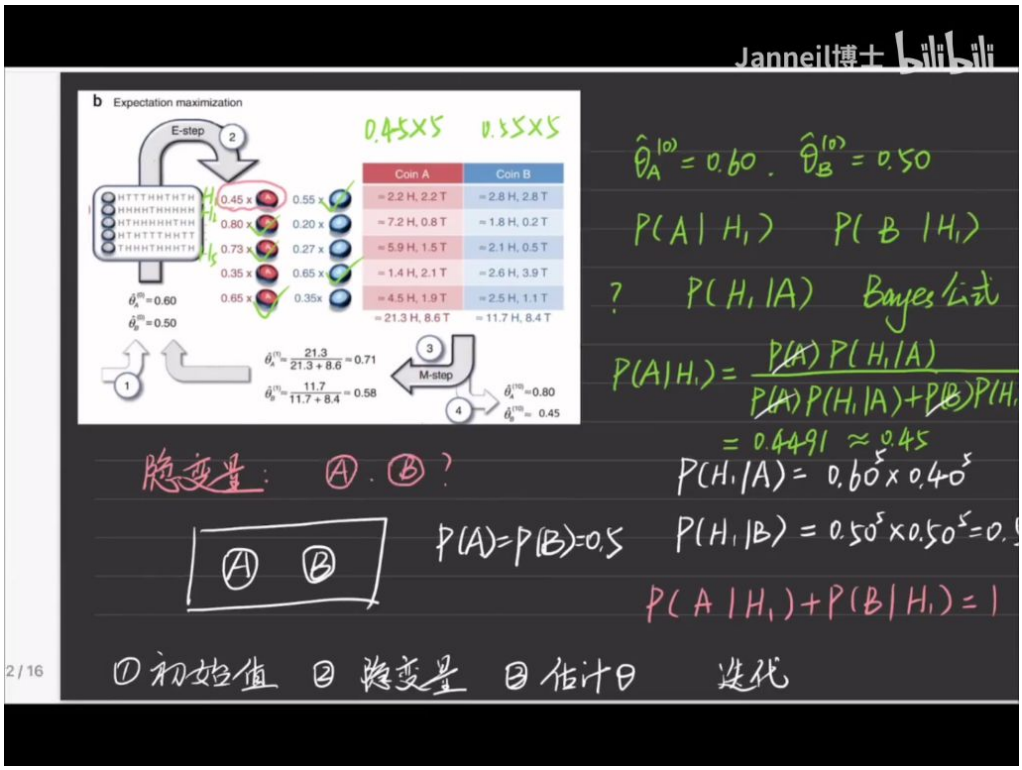


- 模型描述
  - **实验过程**：
    - 先掷硬币A（概率π正面），正面选B（概率p），反面选C（概率q）
    - 掷选出的硬币（B或C），结果1表示正面，0表示反面
    - 独立重复n=10次试验，观测序列：1,1,0,1,0,0,1,0,1,1
  - **问题特点**：只能观测最终结果，不能观测中间过程（A的结果和选择的是B/C）
- 概率计算
  - **单次观测概率**：
    - 出现1的概率：$\pi p + (1 - \pi)q$
    - 出现0的概率：$\pi(1 - p) + (1 - \pi)(1 - q)$

联合概率公式：

$$P(y_j|\theta) = \pi p^{y_j}(1 - p)^{1 - y_j} + (1 - \pi)q^{y_j}(1 - q)^{1 - y_j}$$

  - 其中$\theta = (\pi,p,q)$为待估参数
- EM算法应用



  - E步计算
    - **隐变量定义**：z_j表示第j次试验选择的是B(1)还是C(0)

条件概率计算（贝叶斯公式）：

$$\mu_j^{(i+1)} = P(z_j = B|y_j,\theta^{(i)}) = \frac{\pi^{(i)}p^{(i)y_j}(1 - p^{(i)})^{1 - y_j}}{\pi^{(i)}p^{(i)y_j}(1 - p^{(i)})^{1 - y_j} + (1 - \pi^{(i)})q^{(i)y_j}(1 - q^{(i)})^{1 - y_j}}$$

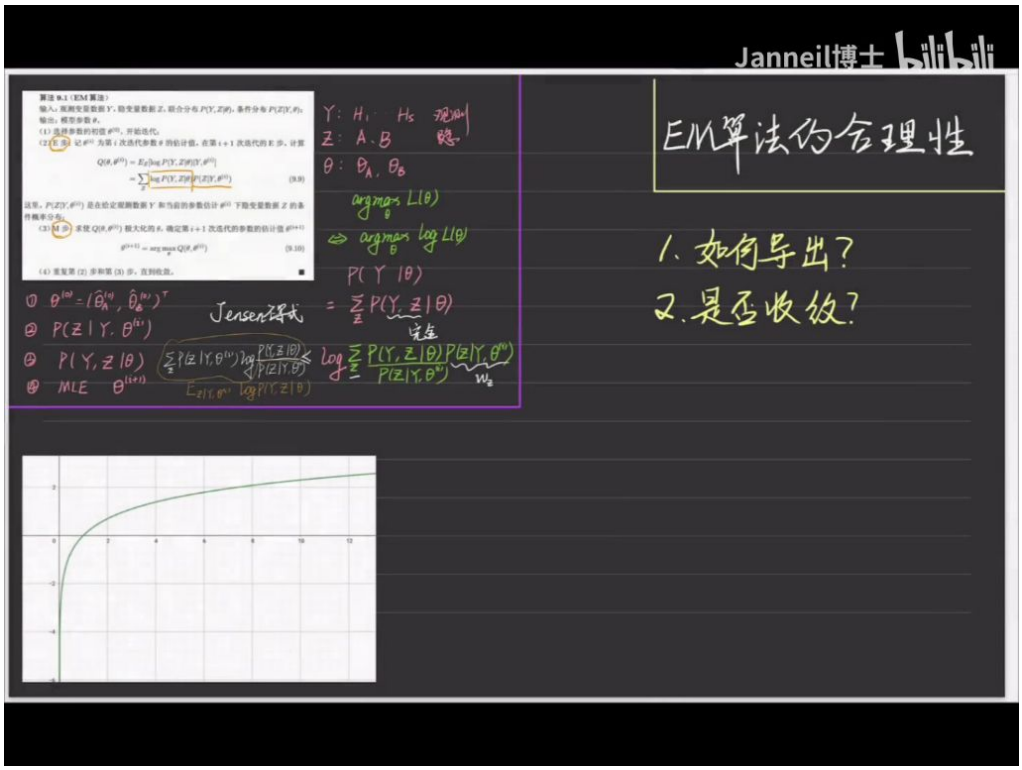
  - M步计算

参数更新公式：

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

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

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

- 迭代结果
  - **初始值1**：$\theta^{(0)} = (0.5,0.5,0.5)$
    - 第一轮：$\theta^{(1)} = (0.5,0.6,0.6)$
    - 第二轮：$\theta^{(2)} = (0.5,0.6,0.6)$（收敛）
  - **初始值2**：$\theta^{(0)} = (0.4,0.6,0.7)$
    - 最终收敛到：$\theta = (0.4064,0.5368,0.6432)$
  - **算法特点**：结果依赖初始值选择，可能收敛到局部最优
- EM算法合理性
  - 

Jensen不等式应用：

$$logE[Y] \geq E[logY]$$

- 用于保证每次迭代似然函数不下降
- **收敛性保证**：
  - E步：计算Q函数（完全数据对数似然的期望）
  - M步：最大化Q函数更新参数
  - 通过Jensen不等式证明每次迭代似然值单调不减

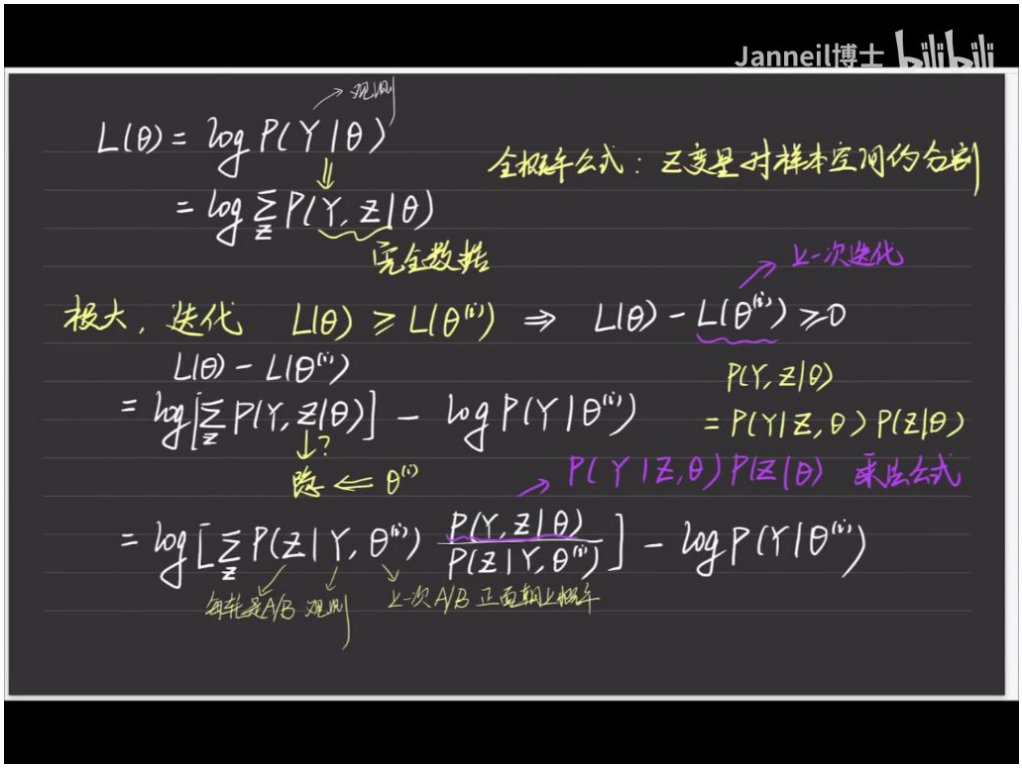
4. em算法的合理性 01:52:18

1）如何导出 01:57:00

- 私人函数与观测结果的关系 01:57:10

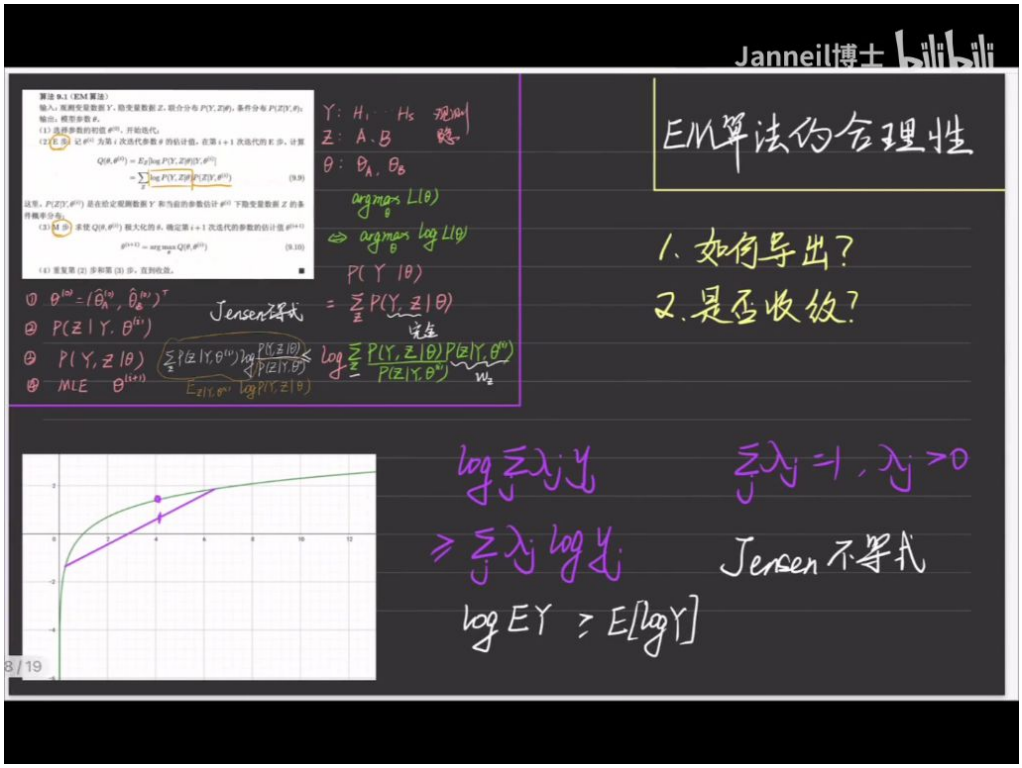
- **对数似然函数**：在参数θ的情况下，所有观测结果y对应的概率$P(y|\theta)$的对数值，目标是使其最大化
- **观测结果特点**：y代表所有观测结果的集合，不是单一变量
- **隐变量问题**：当存在隐变量z（如硬币例子中的A/B硬币选择）时，直接求$P(y|\theta)$不现实

- 引入隐变量z 01:58:03
  - **完全数据构建**：通过引入隐变量z（如每轮投掷的硬币类型）形成完全数据(y,z)
  - **类比说明**：类似Nature论文中A/B硬币的例子，知道具体硬币类型后概率计算更简单
  - **展开原理**：按照隐变量z的所有可能取值进行全概率公式展开
- 全概率公式展开 01:59:04
  - **展开方式**：考虑z所有取值情况下出现观测值y的概率$P(y|\theta) = \sum_z P(y,z|\theta)$
  - **数学表达**：$logP(y|\theta) = log\sum_z P(y,z|\theta)$
  - **空间分割**：隐变量z实际上是对样本空间的一种分割
- 迭代方法与最大化对数似然函数 02:00:02
  - **迭代目标**：保证每次迭代后似然函数值增大，即$L(\theta) - L(\theta^{(i)}) \geq 0$
  - **分解处理**：将$L(\theta)$展开为$logP(y,z|\theta) - logP(y|\theta^{(i)})$
  - **参数关系**：$\theta^{(i)}$是已知的上轮参数，θ是待求的新参数
- 参数与隐变量的处理 02:02:29
  - **信息状态**：
    - y：已知观测数据
    - z：通过$\theta^{(i)}$可推测（半已知）
    - θ：待求参数
  - **贝叶斯应用**：用$\theta^{(i)}$和y推测z的条件概率$P(z|y,\theta^{(i)})$
- 贝叶斯公式的应用 02:03:31
  - **硬币例子对应**：
    - y：每轮观测结果（正/反面）
    - $\theta^{(i)}$：上轮估计的A/B硬币正面概率
    - z：每轮实际使用的硬币类型（A/B）
  - **概率计算**：通过$P(z|y,\theta^{(i)})$估计隐变量分布
- 乘法公式的简易应用 02:06:22



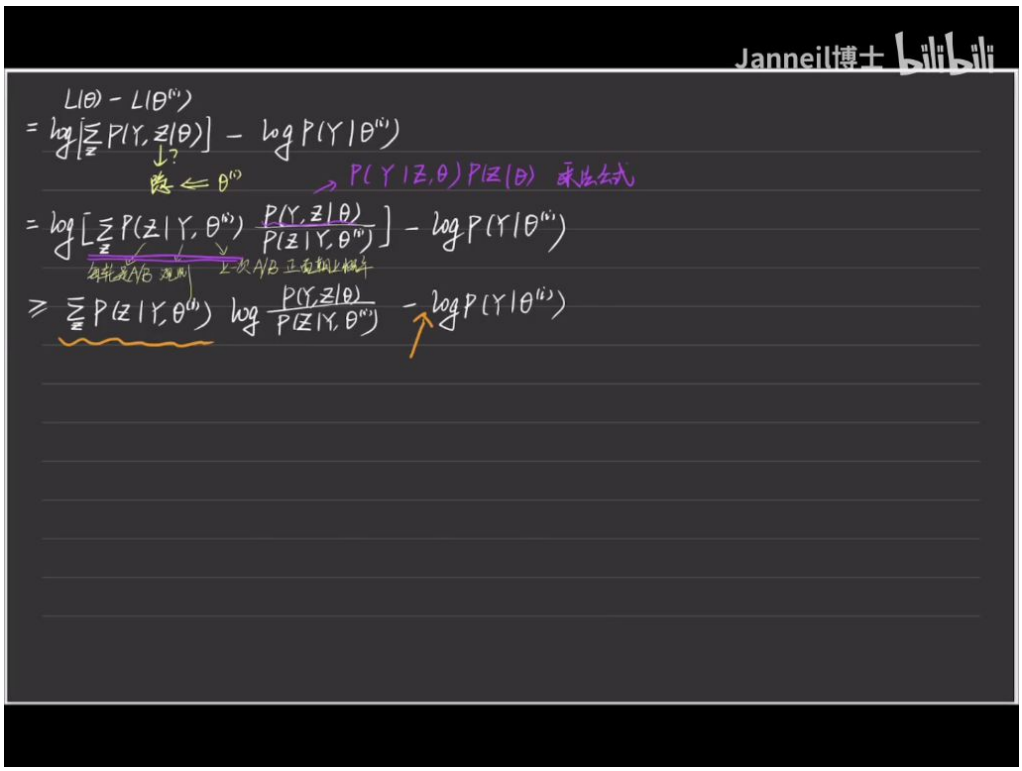
- **联合概率分解**：$P(y,z|\theta) = P(y|z,\theta)P(z|\theta)$
- **条件概率理解**：先求隐变量z的概率，再求在该隐变量下y的条件概率
- **参数条件**：所有概率计算都以参数θ为条件

- EM算法的合理性证明 02:07:24
  - **Jensen不等式应用**：将求和从对数中提出，建立下界
  - **概率分布性质**：$\sum_z P(z|y,\theta^{(i)}) = 1$用于简化表达式
  - **收敛保证**：通过证明$\frac{P(y,z|\theta)}{P(y,z|\theta^{(i)})} \geq 1$确保似然函数单调递增
- EM算法步骤 02:07:43



- **输入**：观测数据Y，隐变量Z，联合分布$P(Y,Z|\theta)$，条件分布$P(Z|Y,\theta)$
- **步骤**：
  - 选择初始参数$\theta^{(0)}$
  - E步：计算$Q(\theta,\theta^{(i)}) = E_Z[logP(Y,Z|\theta)|Y,\theta^{(i)}]$
  - M步：求使Q函数最大化的$\theta^{(i+1)}$
  - 重复直到收敛
- **输出**：最优模型参数θ

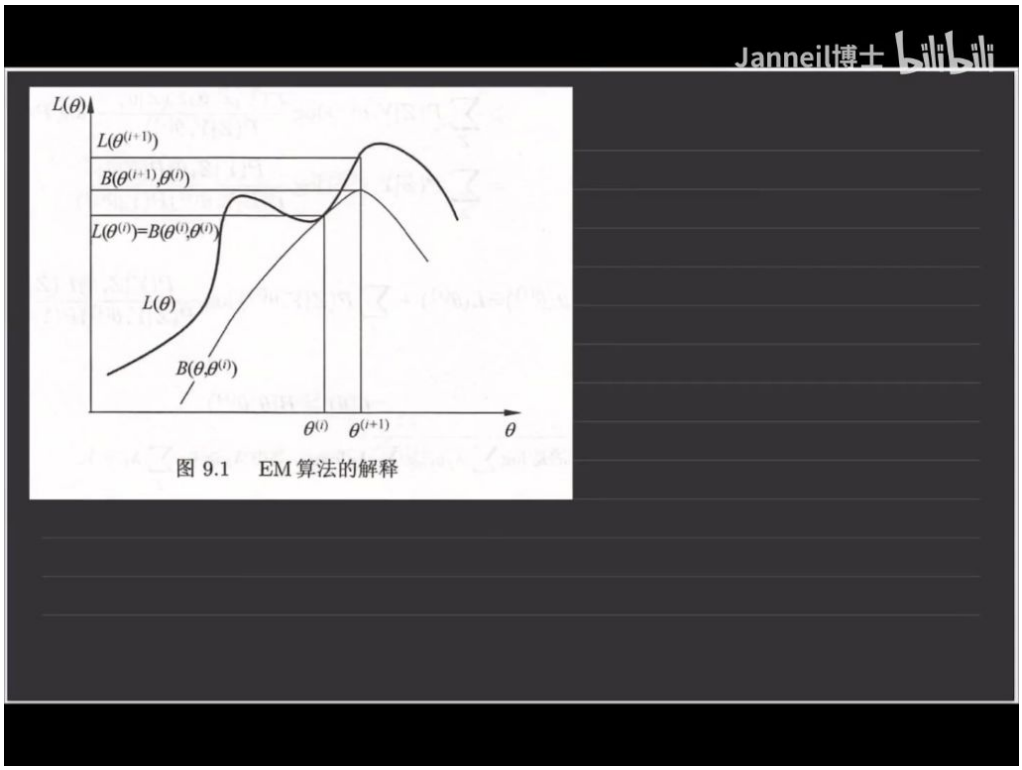
- EM算法的收敛性证明 02:12:47



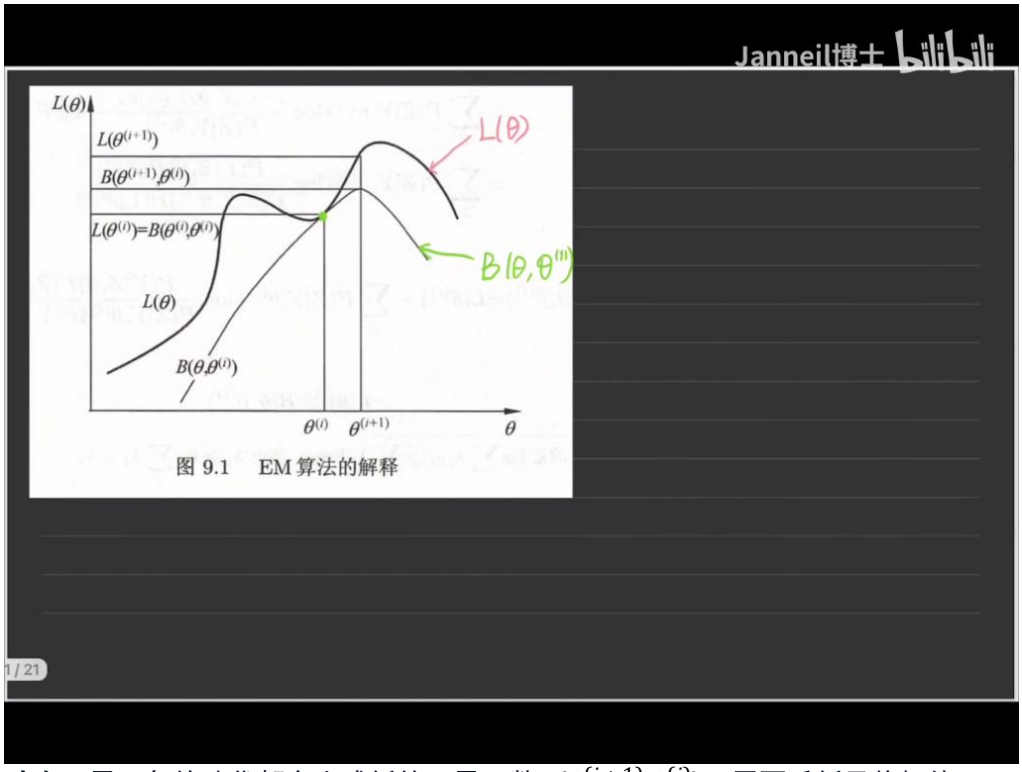
- **收敛条件**: 每次迭代需要满足$l(\theta)$ - $l(\theta^{(t)}) \geq 0$，即新的参数值对应的似然函数必须大于等于前一次参数值对应的似然函数。
- **下界函数构造**：将似然函数$l(\theta)$的下界记为$B(\theta)$，由前次参数的对数似然值$l(\theta^{(t)})$和包含新参数的绿色部分共同组成。

- 最大化下界与参数求解 02:14:21
  - 下界最大化原理
    - **优化策略**: 通过不断增大下界$B(\theta)$来间接实现似然函数$l(\theta)$的最大化。
    - **简化依据**: 由于$l(\theta^{(t)})$是已知常数，最大化下界等价于最大化绿色方框内的表达式。
  - 目标函数化简
    - **分解处理**： 将方框内表达式分解为包含未知参数的部分和完全已知的部分，后者对最大化无贡献可忽略。
    - **最终形式**: 化简后得到粉红色方框内的期望表达式$E[logp(y,z|\theta)]$，即在给定观测数据y和隐变量分布z的条件下，对联合概率对数求期望。
  - EM核心推导
    - **数学本质**： 展示了EM算法中E步（求期望）的数学来源，解释了为什么最终目标函数表现为期望形式。
    - **实现路径**： 通过构造下界、逐步化简，最终将复杂的似然最大化问题转化为可操作的期望最大化问题。

2）是否收敛 02:18:16

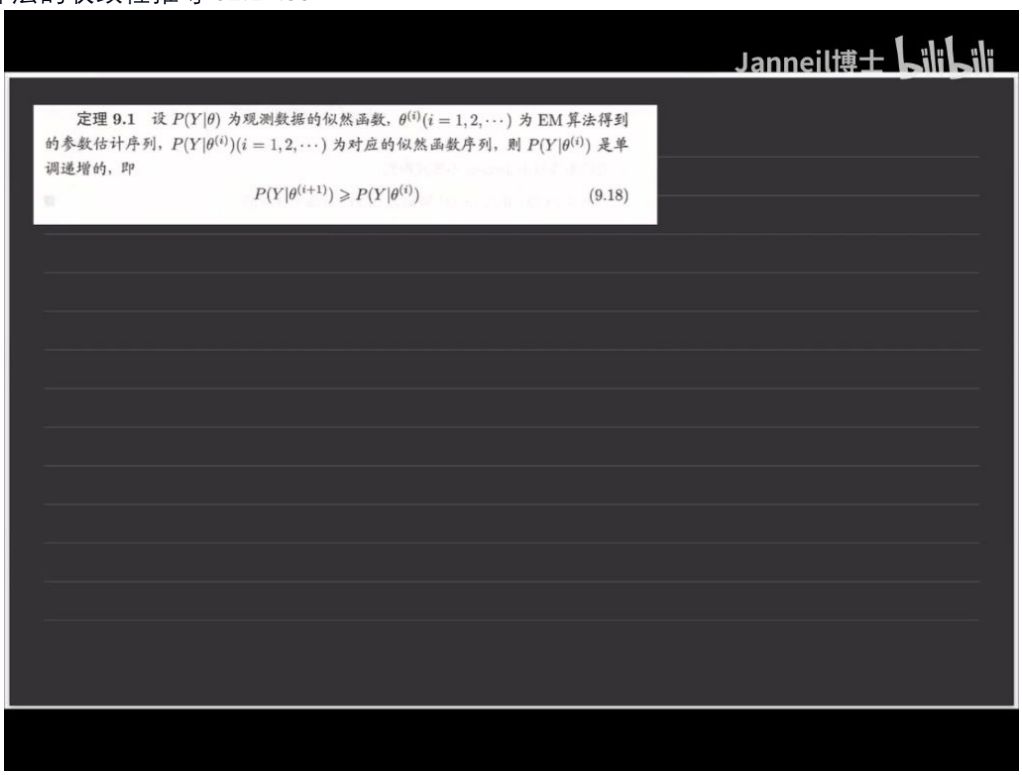
- EM算法的直观图形解释 02:18:19
  - 

  - **对数似然函数**：粗黑线表示$L(\theta)$，即极大似然估计的目标函数，包含隐变量时直接求解困难。
  - **下界函数**：绿色箭头所指的$B(\theta^{(i+1)},\theta^{(i)})$是$L(\theta)$的下界函数，通过迭代优化下界来间接优化原函数。
  - **迭代点性质**：在参数$\theta^{(i)}$处，$L(\theta^{(i)}) = B(\theta^{(i)},\theta^{(i)})$，两函数在该点重合。
- EM算法的下界更新与迭代 02:20:29



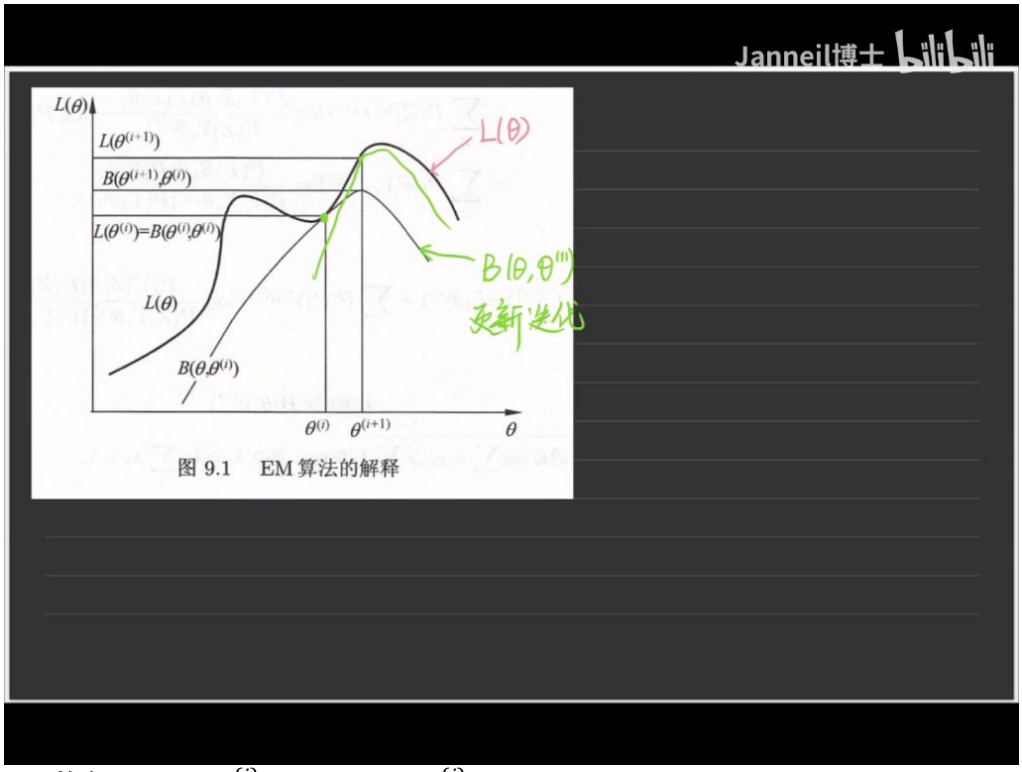
- **动态下界**：每轮迭代都会生成新的下界函数$B(\theta^{(i+1)},\theta^{(i)})$，需要重新寻找极值点。
- **迭代步骤**：
  - 在当前参数$\theta^{(i)}$处建立下界
  - 优化下界函数得到新参数$\theta^{(i+1)}$
  - 在新的参数点建立更高下界
- **图形特征**：每次更新后，似然函数与新下界会在新参数点再次重合，形成阶梯式上升。

- EM算法的收敛性推导 02:21:59



- **定理9.1**：对于观测数据似然函数$P(Y|\theta)$，EM算法产生的参数序列满足$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)})$。
- **对数形式等价**：该定理等价于证明$logP(Y|\theta^{(i+1)}) \geq logP(Y|\theta^{(i)})$。
- **Jensen不等式**：收敛性证明的核心工具，保证了下界函数的有效性。

- EM算法的目标函数与似然函数 02:22:23

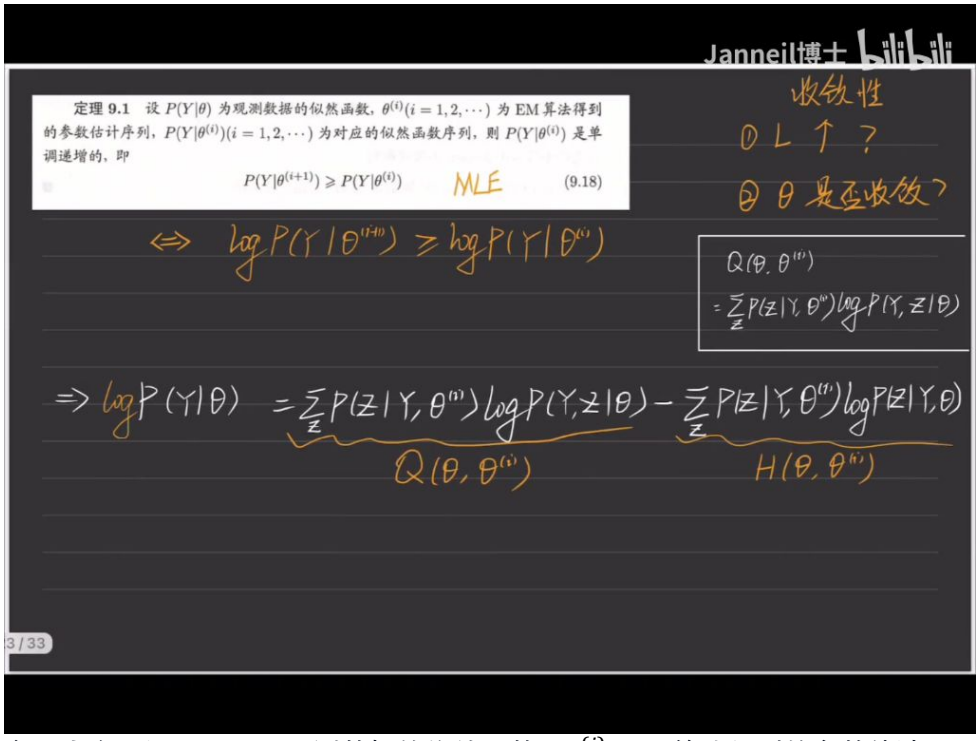


  - **Q函数定义**：$Q(\theta,\theta^{(i)}) = \sum_Z P(Z|Y,\theta^{(i)})logP(Y,Z|\theta)$
  - **单调性保证**：$\theta^{(i+1)}$代入Q函数的值必然大于$\theta^{(i)}$代入值，形成单调递增序列。

条件分布：期望是在隐变量Z的条件概率分布$P(Z|Y,\theta^{(i)})$下求取的。

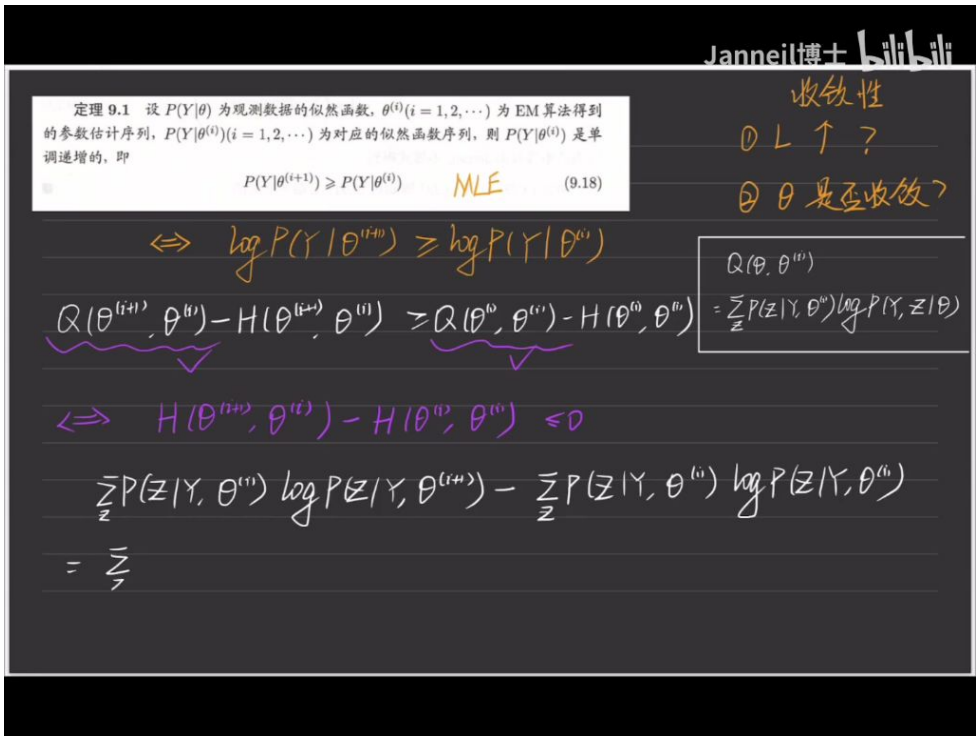
  - 注：所有数学公式均按原始记录保持LaTeX格式，关键概念和迭代过程严格遵循课程讲解内容，未添加任何课程未提及的信息。

- EM算法的收敛性证明 02:26:53
  - 单调递增性证明



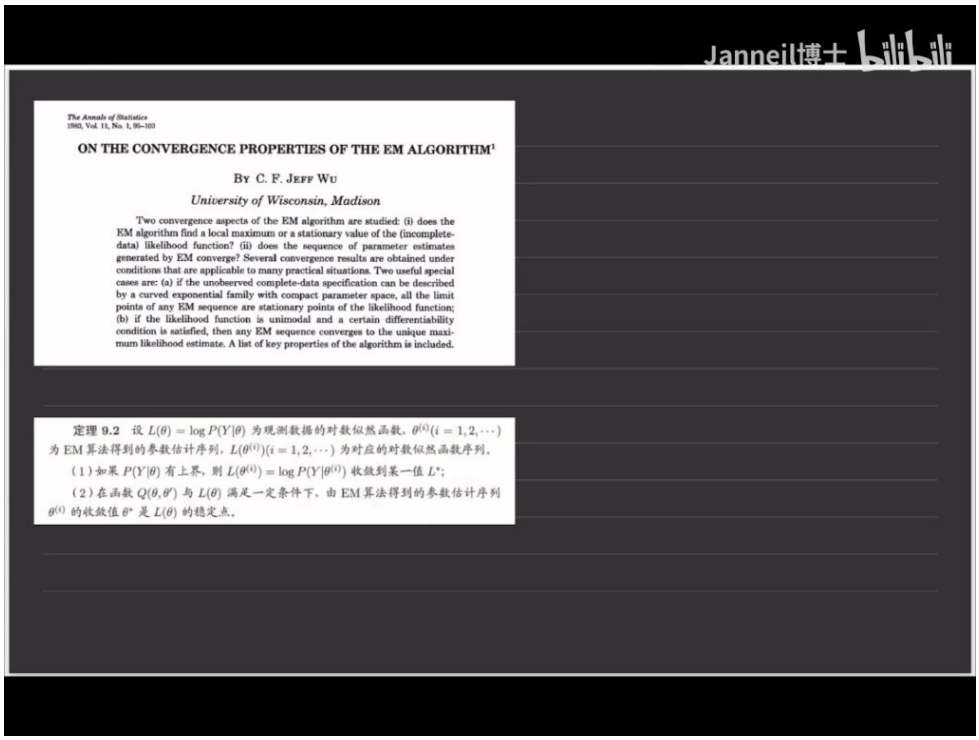
- **定理内容**：设$P(Y|\theta)$为观测数据的似然函数，$\theta^{(i)}$为EM算法得到的参数估计序列，则$P(Y|\theta^{(i)})$是单调递增的，即$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)})$
- **证明思路**：
  - **分解似然函数**：将联合概率分解为$P(Y,Z|\theta) = P(Y|\theta)P(Z|Y,\theta)$
  - **对数变换**：$logP(Y|\theta) = logP(Y,Z|\theta) - logP(Z|Y,\theta)$
  - **引入Q函数**：通过全概率公式展开，定义$Q(\theta,\theta') = \sum_Z P(Z|Y,\theta')logP(Y,Z|\theta)$
  - **H函数定义**：$H(\theta,\theta') = \sum_Z P(Z|Y,\theta')logP(Z|Y,\theta)$
  - **关键不等式**：只需证明$H(\theta^{(i+1)},\theta^{(i)}) - H(\theta^{(i)},\theta^{(i)}) \leq 0$

- 收敛性分析



- **H函数性质**:
  - 通过Jensen不等式证明$\sum_Z P(Z|Y,\theta^{(i)}) log \frac{P(Z|Y,\theta^{(i+1)})}{P(Z|Y,\theta^{(i)})} \leq log \sum_Z P(Z|Y,\theta^{(i+1)}) = log 1 = 0$
  - 由此得出$H(\theta^{(i+1)},\theta^{(i)}) \leq H(\theta^{(i)},\theta^{(i)})$
- **结论**:
  - 对数似然序列$L(\theta^{(i)})$单调递增
  - 似然函数序列$P(Y|\theta^{(i)})$单调递增

- EM算法的收敛性质 02:38:00
  - 基本收敛定理

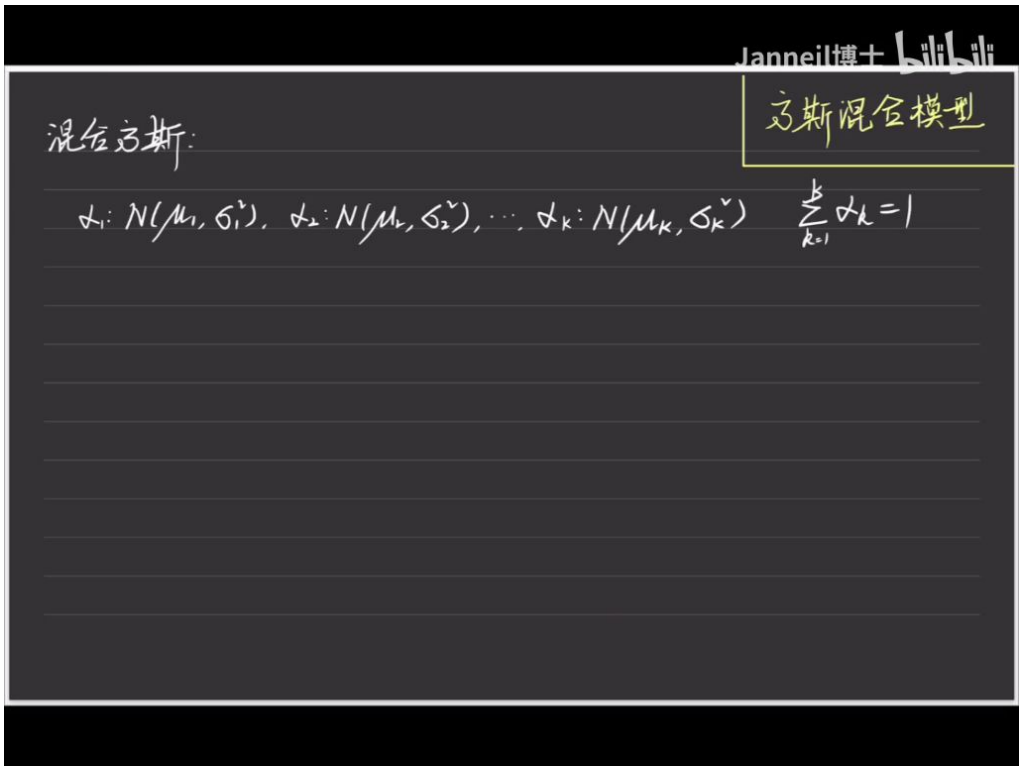


  - **定理9.2**:
    - 设$L(\theta) = log P(Y|\theta)$为对数似然函数，$\theta^{(i)}$为EM算法参数估计序列
    - **第一部分**：如果$P(Y|\theta)$有上界，则$L(\theta^{(i)})$收敛到某一值L^*
    - **第二部分**：在函数$Q(\theta,\theta')$与$L(\theta)$满足一定条件下，参数估计序列$\theta^{(i)}$的收敛值θ^*是$L(\theta)$的稳定点
  - **实际应用**:
    - 由于概率有界性（$P(Y|\theta) \leq 1$），对数似然序列必然收敛
    - 参数估计可能收敛到不同稳定点，需要比较多个初始值的结果
  - 收敛性补充说明
  - **文献支持**:
    - C.F.Jeff Wu在《The Annals of Statistics》上的论文证明了EM和广义EM(GEM)的收敛性
    - 特别情况：(a)当未观测数据属于紧凑参数空间的指数族时，EM序列的极限点是似然函数的稳定点；(b)当似然函数单峰且满足可微条件时，EM序列收敛到唯一MLE
  - **实践建议**:
    - 需要尝试不同初始值以避免局部最优
    - 对于复杂模型，收敛速度可能较慢，需要设置合理的停止准则

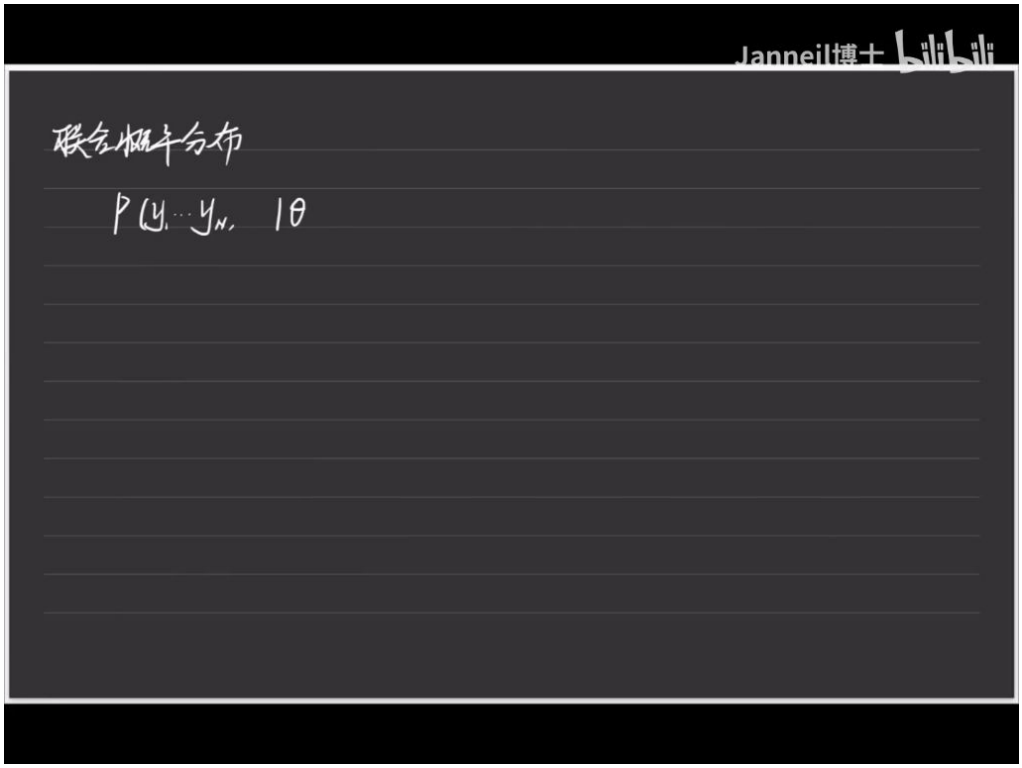
5. 高斯混合模型 02:40:56

1）em算法估计高斯混合模型参数 02:48:52

- 高斯混合模型观测值与隐变量 02:49:03

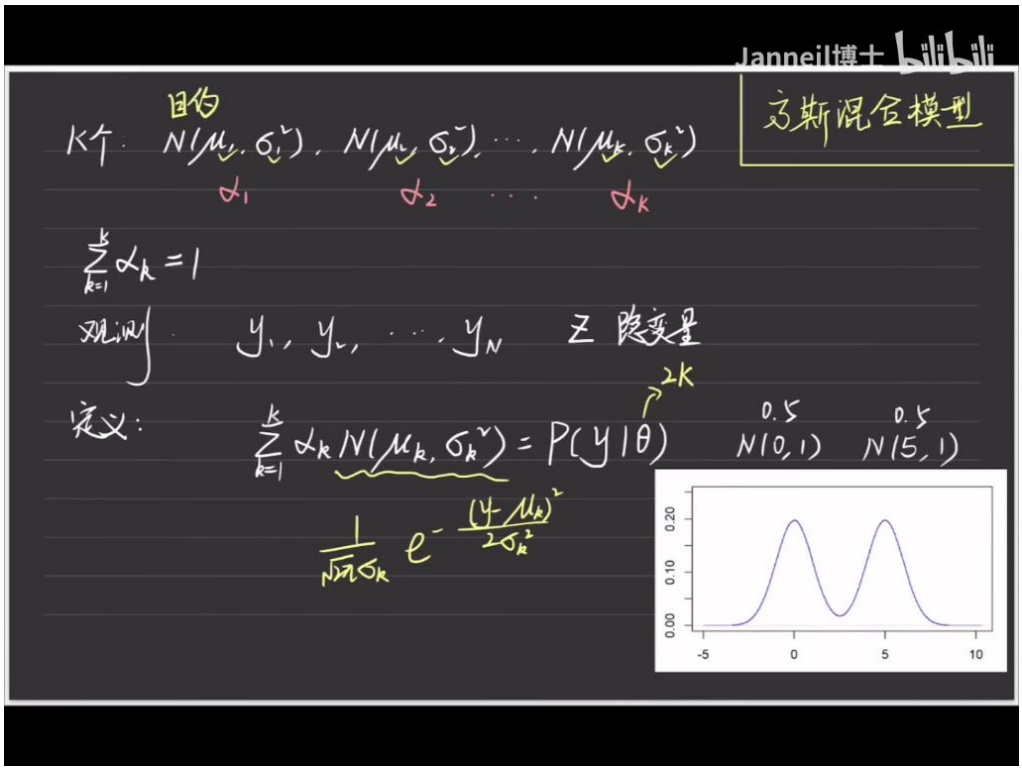


  - **观测值表示**：观测值表示为$y_1,y_2,...,y_N$，共N个观测值，每个观测值来自K个高斯分布中的某一个
  - **隐变量定义**：对于每个观测值y_j，定义K个隐变量$\gamma_{j1},\gamma_{j2},...,\gamma_{jK}$，其中$\gamma_{jk}$ = 1表示y_j来自第k个高斯分布，否则为0
  - **归属确定**：通过这K个隐变量可以完全确定每个观测值的来源分布，γ_{jk}取值为1的项即表示该观测值对应的分布
- 完整数据与参数 02:51:36
  - **完整数据结构**：完整数据包含观测值$y_1,...,y_N$和对应的所有隐变量$\gamma_{11},...,\gamma_{NK}$
  - **参数组成**：参数θ包含三部分：
    - 混合比例：$\alpha_1,...,\alpha_K$
    - 均值：$\mu_1,...,\mu_K$
    - 方差：$\sigma_1^2,...,\sigma_K^2$
  - **参数表示**：完整参数表示为$\theta = (\alpha_1,\mu_1,\sigma_1^2,...,\alpha_K,\mu_K,\sigma_K^2)$
- 联合概率分布 02:54:30



- **联合分布定义**：在参数θ已知情况下观测到完整数据的概率
- **分解形式**：可以分解为N个观测值的连乘形式，每个观测值对应其观测值和隐变量的联合概率
- **连续变量处理**：由于y是连续变量，需要使用概率密度函数而非概率

- 联合概率分布的简化与表达 02:58:59



- **分布归属表示**：使用γ_{jk}作为指示变量，将每个观测值属于各分布的概率表示为$\prod_{k=1}^{K}[\alpha_k \varphi(y_j|\theta_k)]^{\gamma_{jk}}$

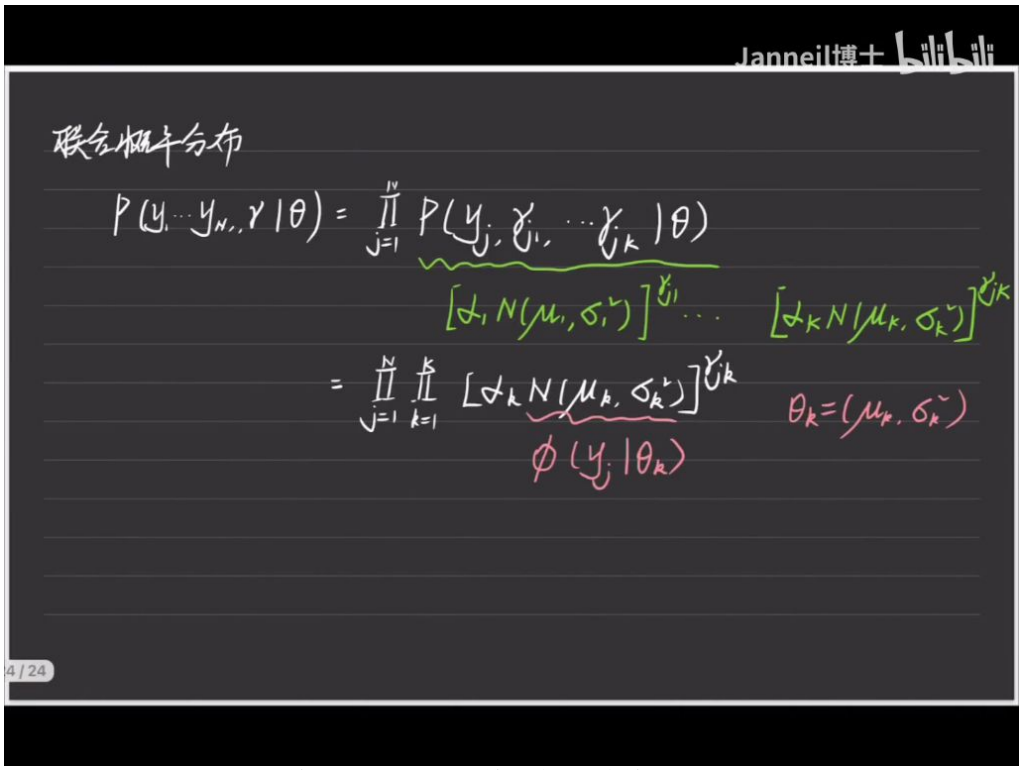
- **概率密度函数**：$\varphi(y_j|\theta_k)$表示第k个高斯分布的概率密度函数，参数为$\theta_k = (\mu_k,\sigma_k^2)$

最终表达式：联合概率分布可简化为：

$$\prod_{k=1}^{K}\alpha_k^{n_k} \times \prod_{j=1}^{N}\prod_{k=1}^{K}[\varphi(y_j|\theta_k)]^{\gamma_{jk}}$$

- 其中$n_k = \sum_{j=1}^{N}\gamma_{jk}$表示来自第k个分布的样本数
- **简化原理**：利用γ_{jk}的指示性质，将连乘分解为混合比例和概率密度两部分

- 对数化处理与似然函数 03:03:15

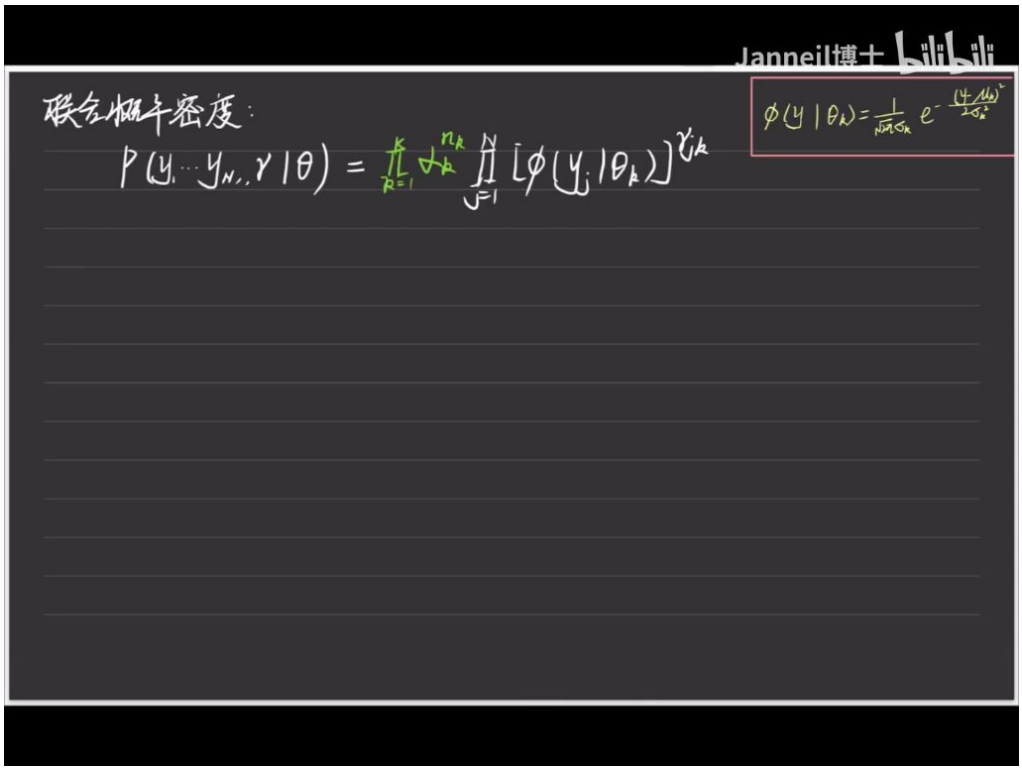


- **对数化优势**：将连乘运算转化为求和运算，简化计算过程。具体表现为将原式中的$\prod_{k=1}^{K}$转换为$\sum_{k=1}^{K}$形式。
- **似然函数构建**：对数化处理后得到的形式即为对数似然函数l，其表达式包含两个主要求和部分：
  - $\sum_{k=1}^{K} n_k ln\alpha_k$：表示各类别先验概率的对数加权和
  - $\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik} lnf(y_i|\theta_k)$：表示观测数据在各类别下的条件概率对数加权和

- 隐变量估计与贝叶斯公式 03:05:03
  - **隐变量性质**：γ_{gk}为二元隐变量，取值只能为0或1，表示样本y_g是否属于第k类。
  - **期望估计法**：
    - **估计原理**：$\hat{\gamma}_{gk} = E[\gamma_{gk}|y,\theta^{(i)}]$
    - **计算展开**：$= 1 \cdot P(\gamma_{gk} = 1|y_g,\theta^{(i)}) + 0 \cdot P(\gamma_{gk} = 0|y_g,\theta^{(i)})$
  - **贝叶斯转换**：

公式应用：通过贝叶斯公式将条件概率转换为可计算形式：

- $P(\gamma_{gk} = 1|y_g,\theta^{(i)}) = \dfrac{f(y_g|\gamma_{gk} = 1,\theta^{(i)})P(\gamma_{gk} = 1|\theta^{(i)})}{\sum_{k=1}^{K} f(y_g|\gamma_{gk} = 1,\theta^{(i)})P(\gamma_{gk} = 1|\theta^{(i)})}$
- **参数解释**：
  - 分子部分：$f(y_g|\mu_k,\sigma_k^2)$表示第k类的高斯密度函数
  - 分母部分：α_k表示上一轮迭代中第k类的混合系数

- 隐变量与样本数量的确定 03:12:15
  - 

  - **隐变量确定**：通过γ_{gk}指示样本是否属于第k类，γ_{gk} = 1表示样本属于第k类
  - **样本计数**：$n_k = \sum_{g=1}^{n} \gamma_{gk}$，统计属于第k类的样本数量
  - **参数依赖**：计算过程需要用到第k类的参数μ_k和σ_k^2
- EM算法：E步回顾 03:14:08
  - **双重目的**：
    - **隐变量补全**：在已知参数$\theta^{(i)}$情况下，通过贝叶斯公式计算γ_{gk}
    - **目标函数构建**：为M步准备优化目标函数$Q(\theta,\theta^{(i)})$

隐变量公式：

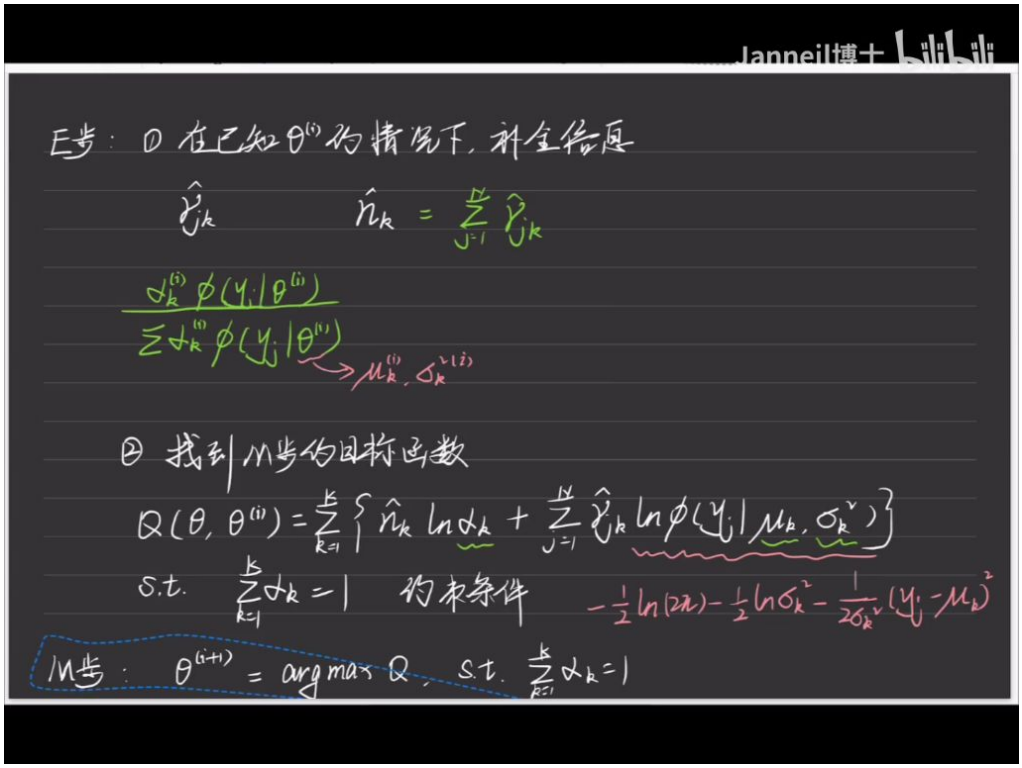
  - $\gamma_{gk} = \frac{\alpha_k p(y_g|\mu_k,\sigma_k^2)}{\sum_{l=1}^{K} \alpha_l p(y_g|\mu_l,\sigma_l^2)}$
  - **参数限定**：计算时只需使用对应第k类的参数μ_k和σ_k^2
- EM算法：M步的目标函数 03:16:19

函数组成：

  - $\sum_{k=1}^{K} n_k ln\alpha_k + \sum_{g=1}^{n} \gamma_{gk} lnp(y_g|\mu_k,\sigma_k^2)$
  - **约束条件**：$\sum_{k=1}^{K} \alpha_k = 1$

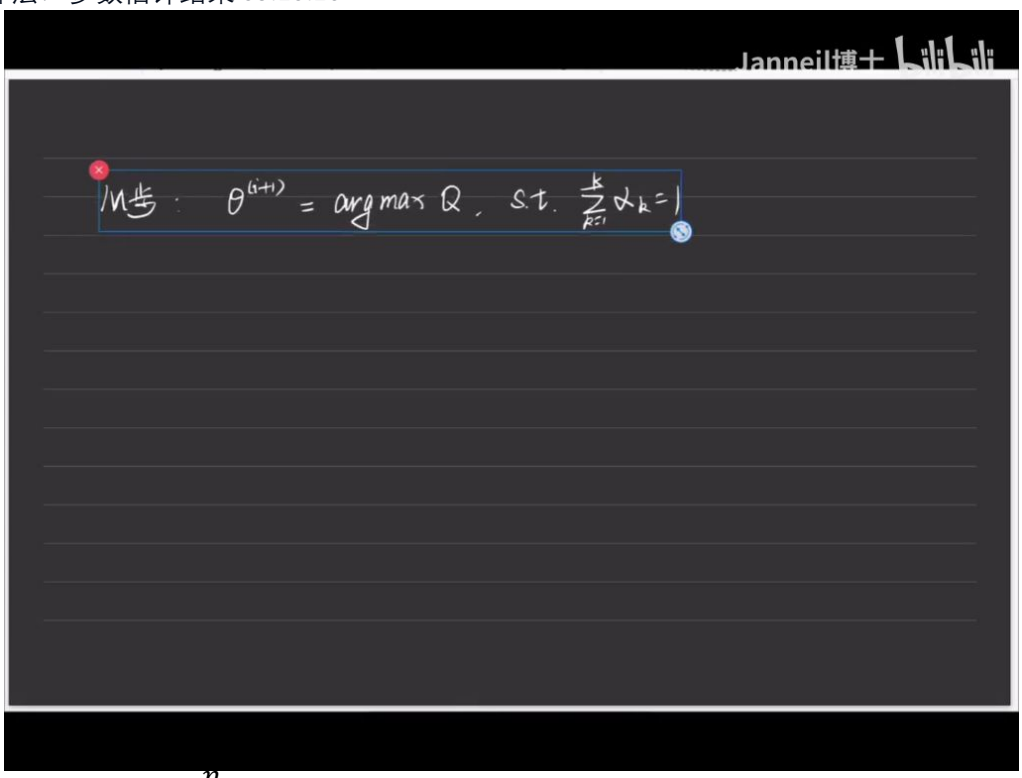
对数展开：

  - $-\frac{1}{2}ln(2\pi) - \frac{1}{2}ln\sigma_k^2 - \frac{1}{2\sigma_k^2}(y_g - \mu_k)^2$
  - **待求参数**：$\alpha_k,\mu_k,\sigma_k^2$
- EM算法：M步求解 03:19:29



- **优化目标**：
- **约束处理**：需考虑$\sum \alpha_k = 1$的约束条件
- **求解方法**：使用拉格朗日乘数法等优化工具

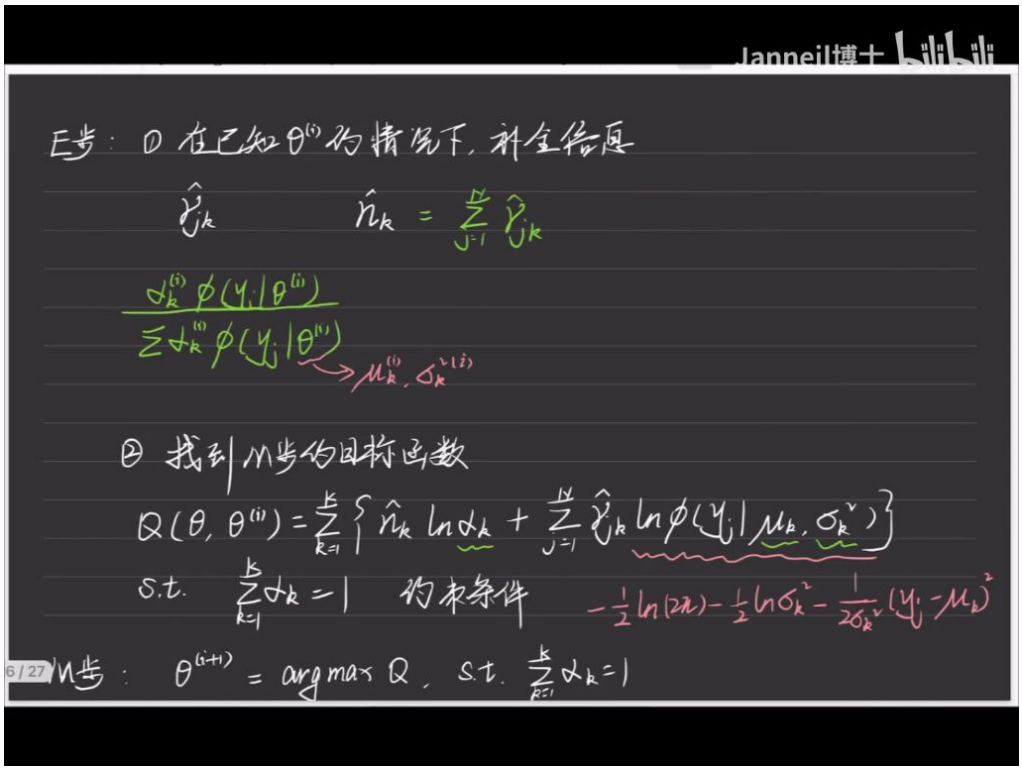
- EM算法：参数估计结果 03:20:26



- **混合系数**：$\hat{\alpha}_k = \frac{n_k}{n}$
- **均值估计**：$\hat{\mu}_k = \frac{\sum_{g=1}^{n} \gamma_{gk} y_g}{n_k}$

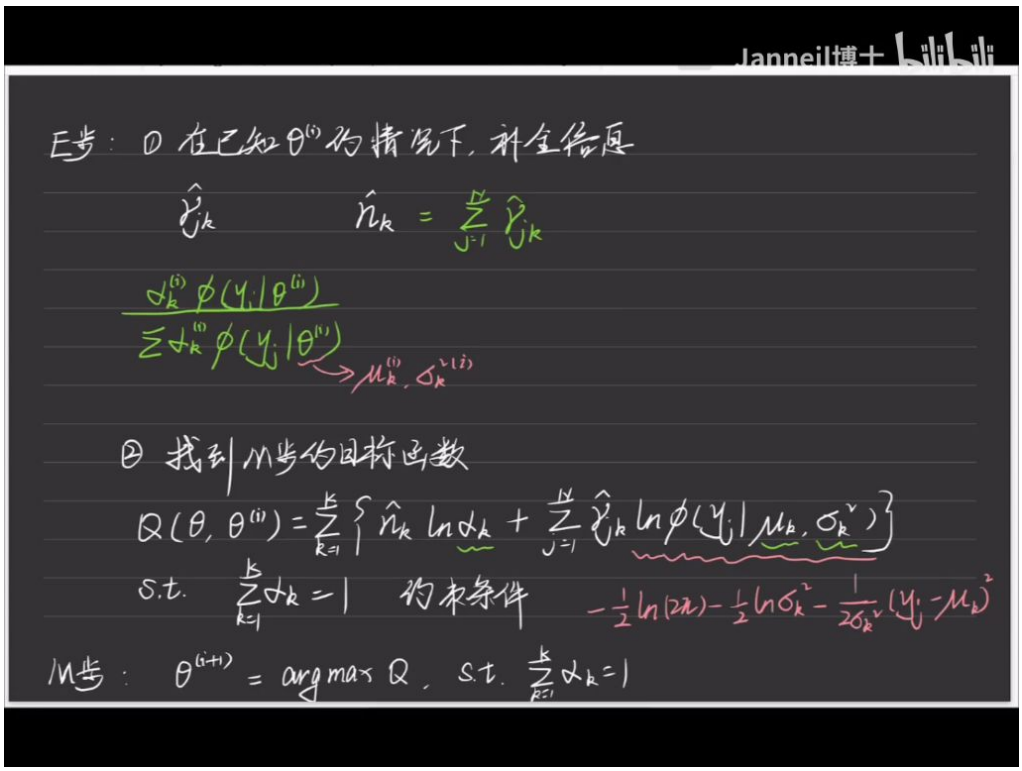
- **方差估计**：$\hat{\sigma}_k^2 = \frac{\sum_{g=1}^{n} \gamma_{gk}(y_g - \hat{\mu}_k)^2}{n_k}$

- EM算法：参数估计结果的理解 03:21:38



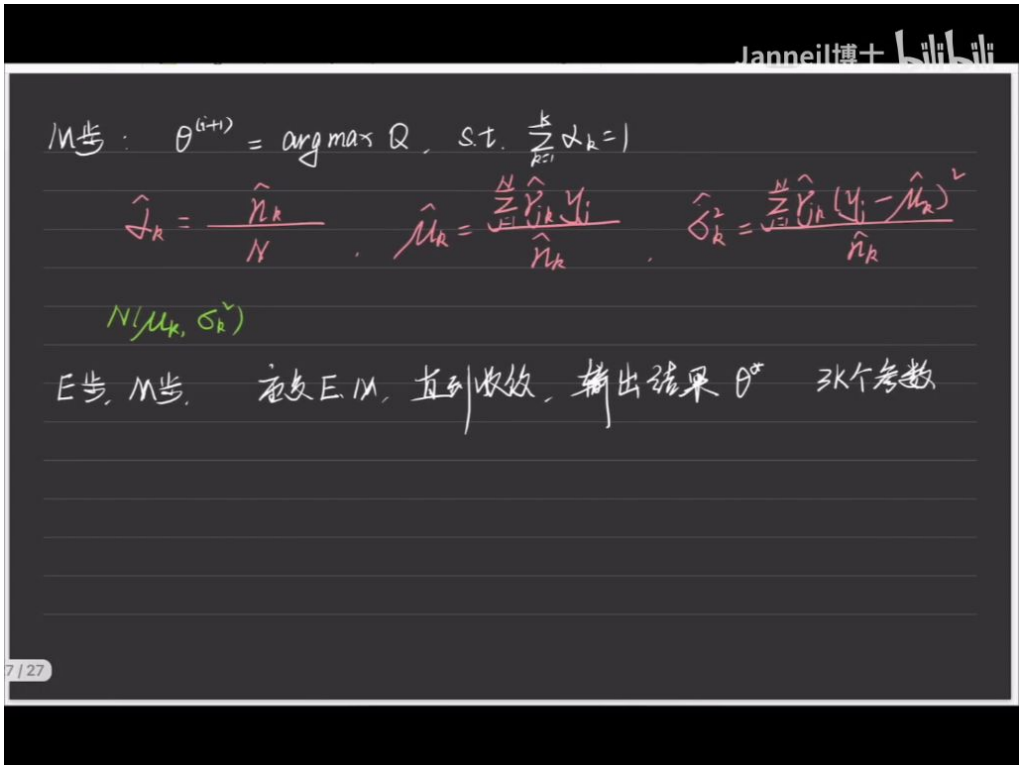
- **混合系数**：表示第k个分布的样本占比，用估计的样本数n_k除以总数n
- **均值估计**：仅使用$\gamma_{gk} = 1$的样本计算加权平均
- **方差估计**：基于已分类样本计算样本方差
- **选择机制**：γ_{gk}作为指示函数筛选属于当前分布的样本

- EM算法：参数求解 03:23:31



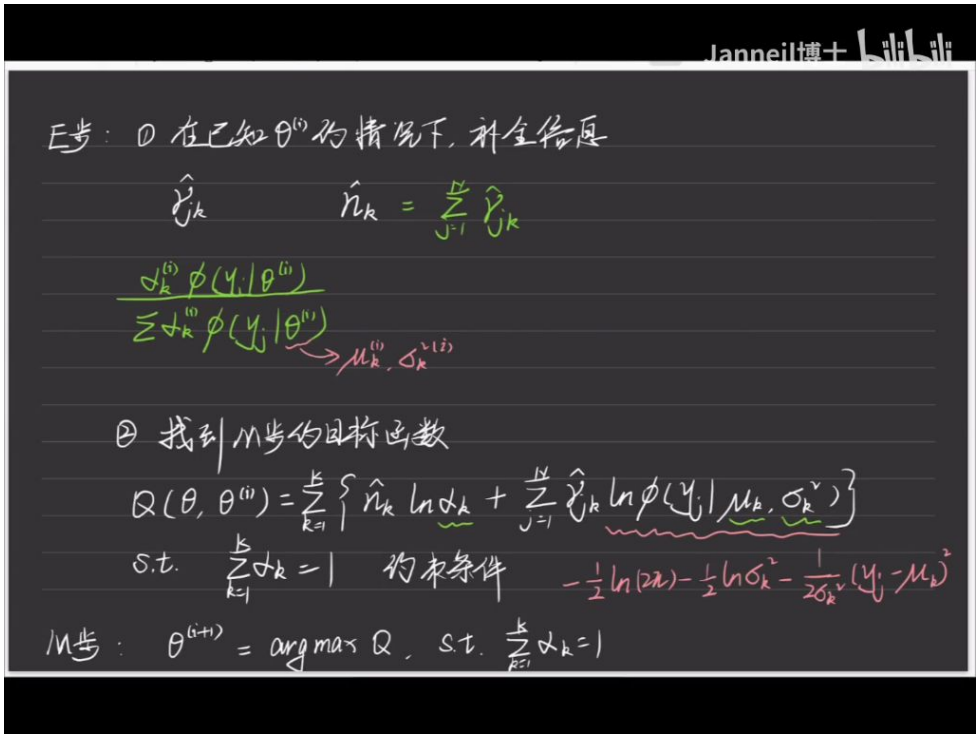
- **迭代过程**：交替执行E步和M步直至收敛
- **输出结果**：最终参数θ^*包含$3K$个参数（K个分布的$\alpha_k,\mu_k,\sigma_k^2$）
- **优化求解**：需要通过带约束的优化问题推导参数估计式

- EM算法：拉格朗日乘子法 03:24:54



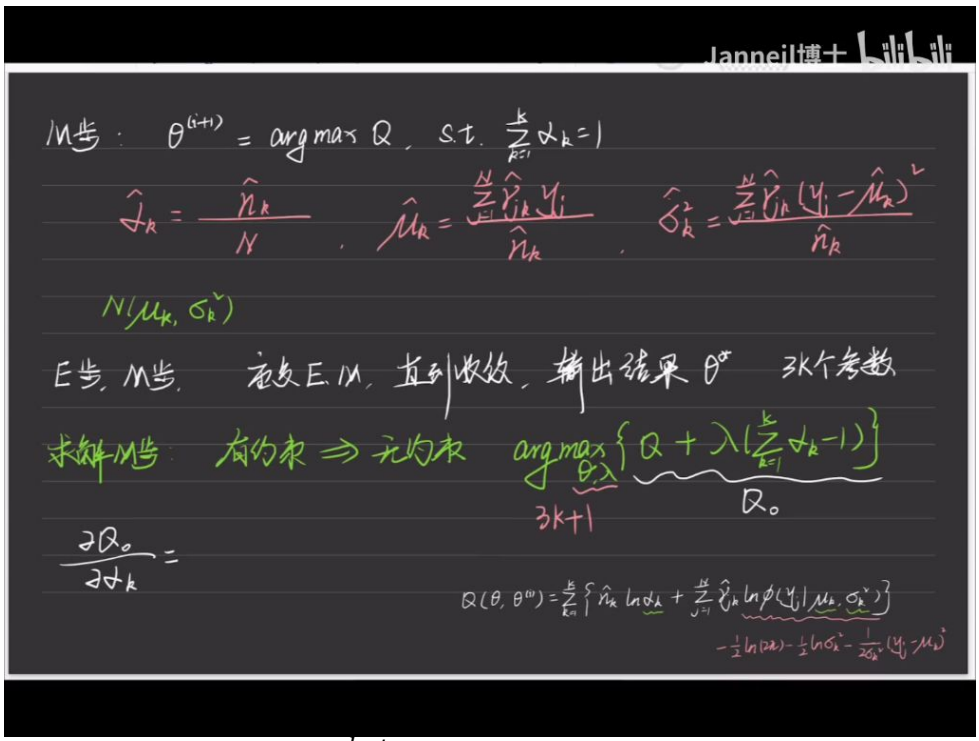
- **约束转换**: 通过引入拉格朗日乘子λ，将有约束优化问题转化为无约束优化问题
- **参数增加**：使用拉格朗日乘子法后，参数总数变为$3k+1$个（包含拉格朗日参数λ）
- **目标函数**: 新的无约束目标函数记为Q_0

- EM算法：拉格朗日乘子法求解 03:26:14
  - 参数α_k的求解



- **求导方法**: 对α_k求偏导，找出包含α_k的项
- **关键项**: 主要包含α_k的项有两处
- **约束条件**: 需要满足$\sum \alpha_k = 1$的约束条件

- 具体推导过程



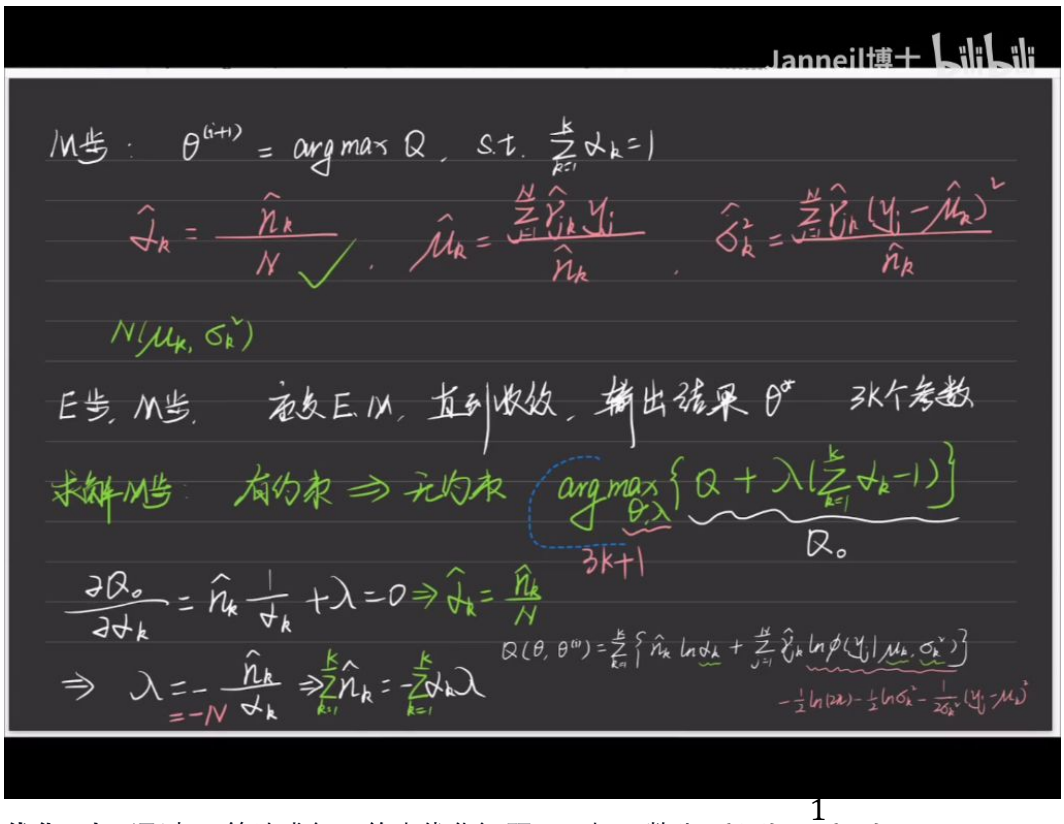
- **偏导结果**: 对α_k求偏导得到$\frac{n_k^{hat}}{\alpha_k} + \lambda = 0$
- **乘子关系**: 推导出$\lambda = -\frac{n_k^{hat}}{\alpha_k}$
- **约束应用**: 利用$\sum \alpha_k = 1$的约束条件，求和后得到$\lambda = -N$

- **最终解**: 代入得到$\alpha_k^{hat} = \frac{n_k}{N}$，与理论结果一致

三、统计学习方法 ***03:29:27***

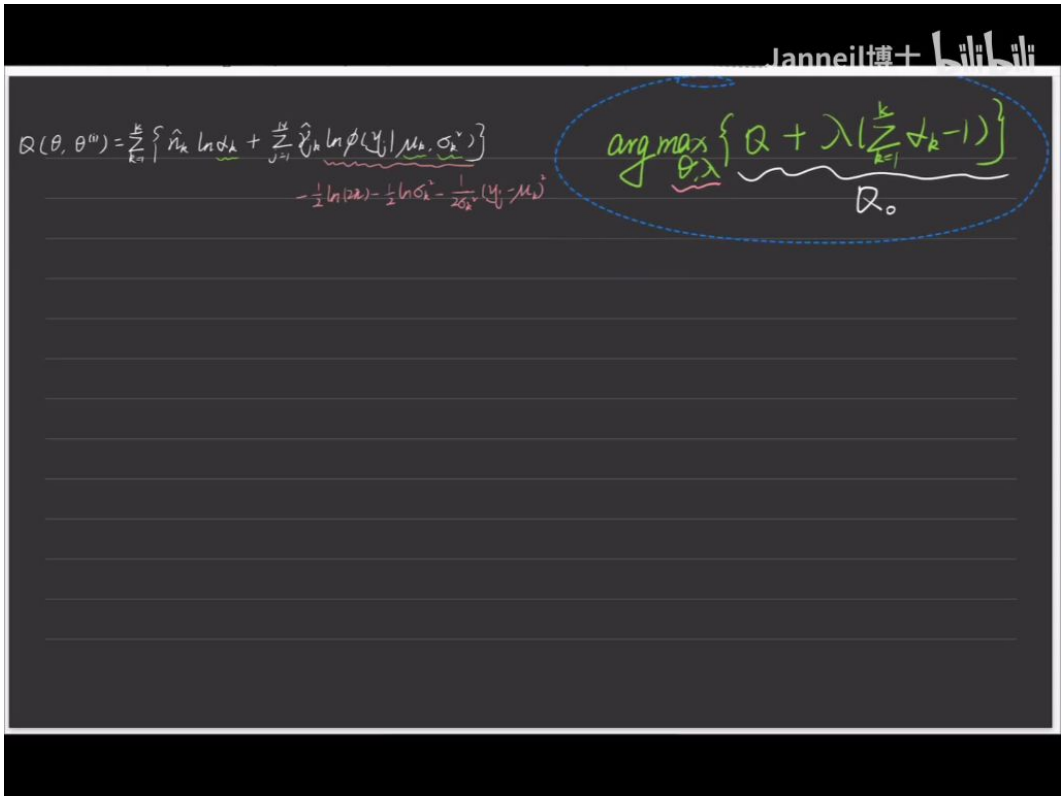
1. 课前回顾

1）无约束问题 03:29:40



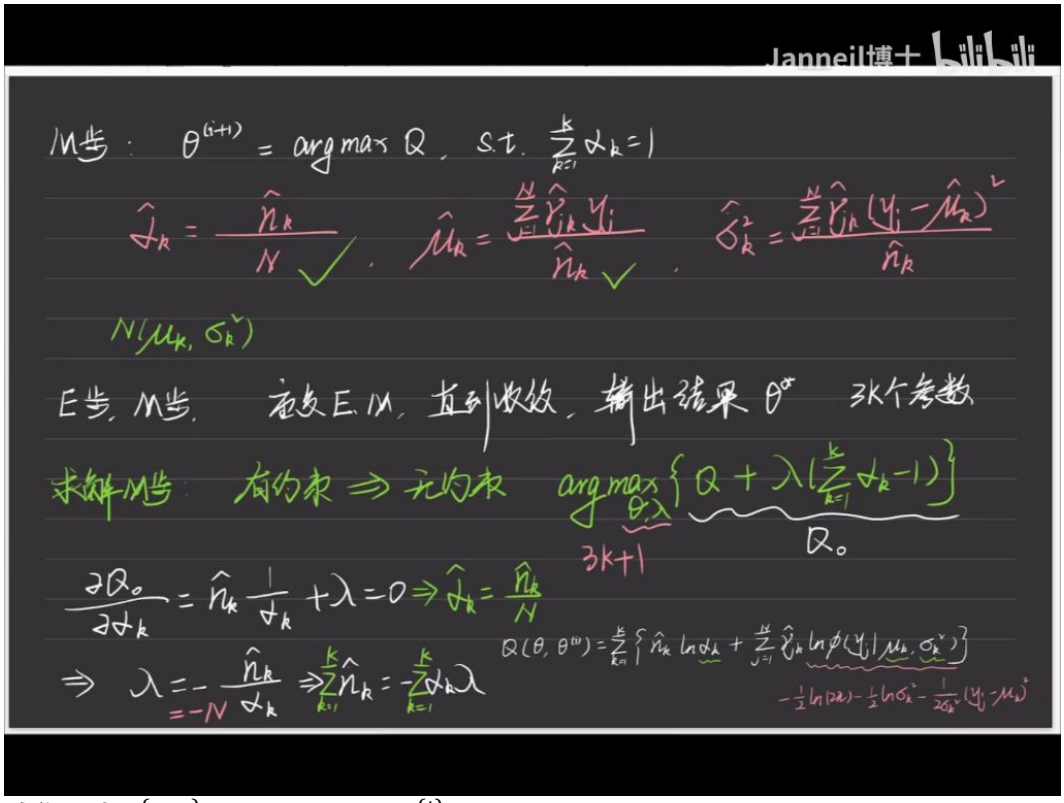
- **优化目标**: 通过EM算法求解无约束优化问题，目标函数为$Q(\theta,\theta) = \frac{1}{2}(x,e)$
- **求解步骤**: 包含E步（期望计算）和M步（参数最大化），迭代直到收敛

2）例题1:求缪开和什么开的平方 03:29:59



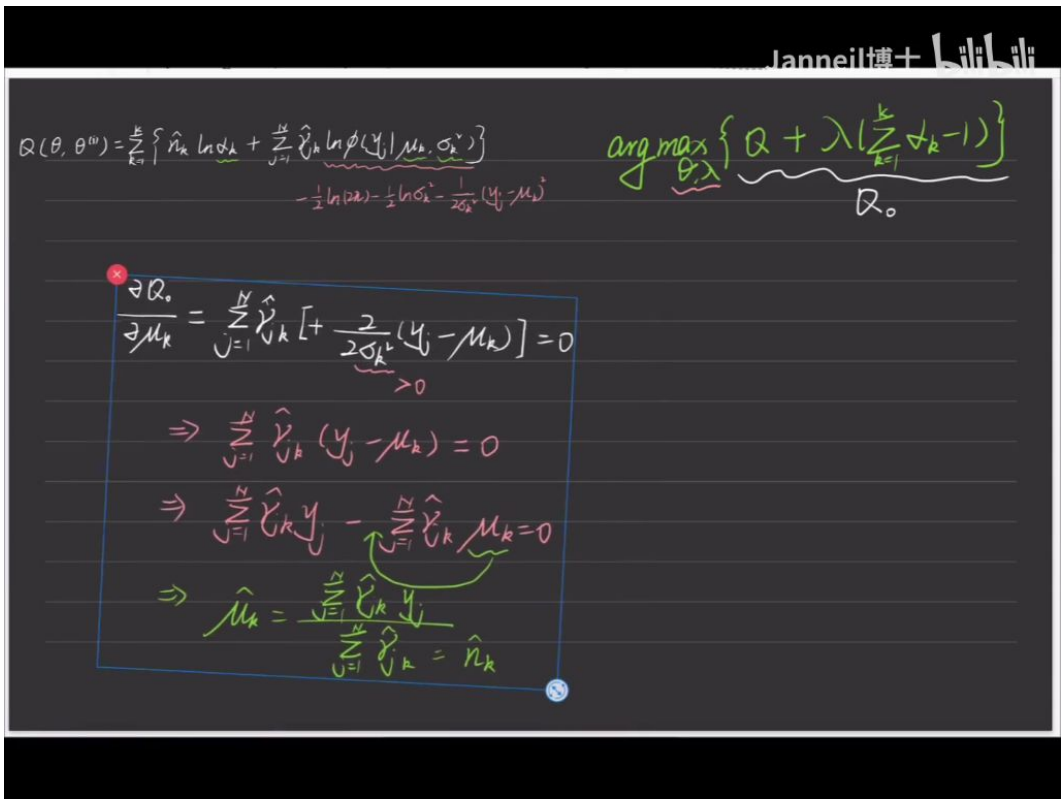
- **参数估计**:
  - **包含项**: 仅需对包含μ_k的项求偏导
  - **偏导过程**: 使用复合函数求导法则，得到$\sum_{g=1}^{n}\gamma_{gk}(y_g-\mu_k)=0$
  - **简化结果**:$\hat{\mu}_k=\frac{\sum_{g=1}^{n}\gamma_{gk}y_g}{\sum_{g=1}^{n}\gamma_{gk}}$
- **数学技巧**: 由于$\sigma_k^2>0$，可简化计算只需令分子部分为零

3）例题2:求无约束问题 03:32:28



- **迭代公式**:$\theta^{(x-x)} = argmax_{\theta}Q(\theta,\theta^{(t)})$
- **收敛条件**: 通过E步和M步交替进行直到参数收敛

4）例题3:求三个门开的平方 03:32:51

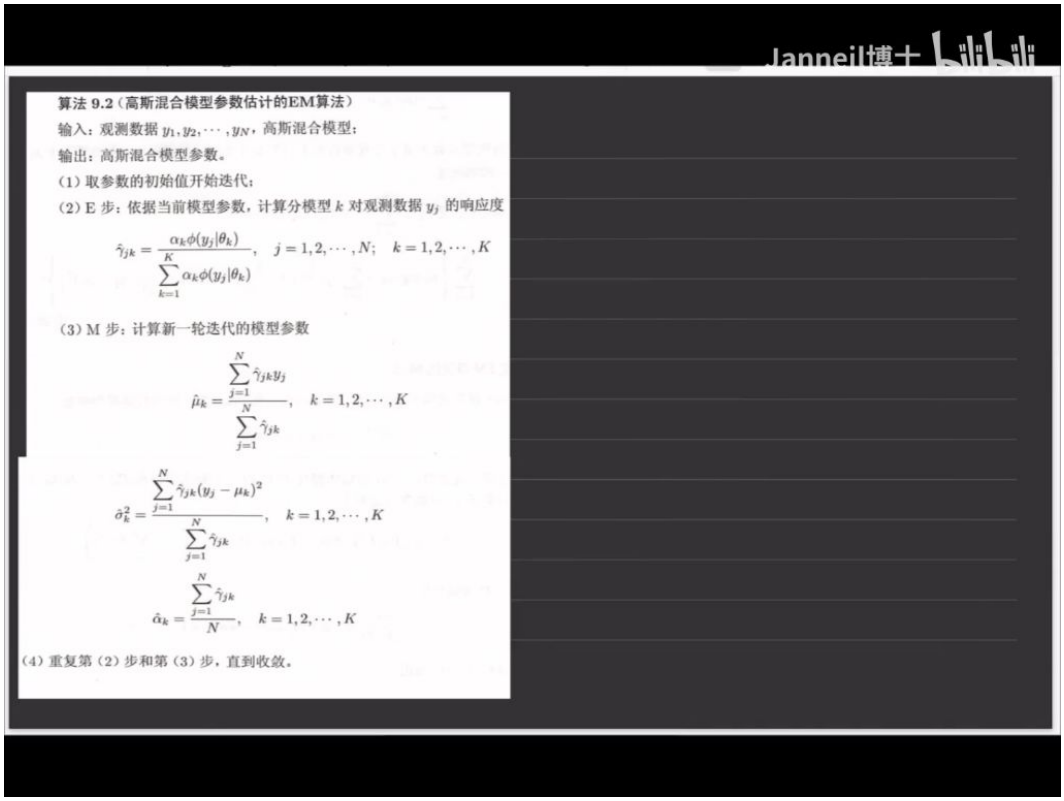


- **估计步骤**:
  - 对σ_k^2求偏导
  - 令偏导数为零并简化

- 得到估计式$\hat{\sigma}_k^2 = \frac{\sum_{g=1}^{n} \gamma_{gk}(y_g - \mu_k)^2}{\sum_{g=1}^{n} \gamma_{gk}}$

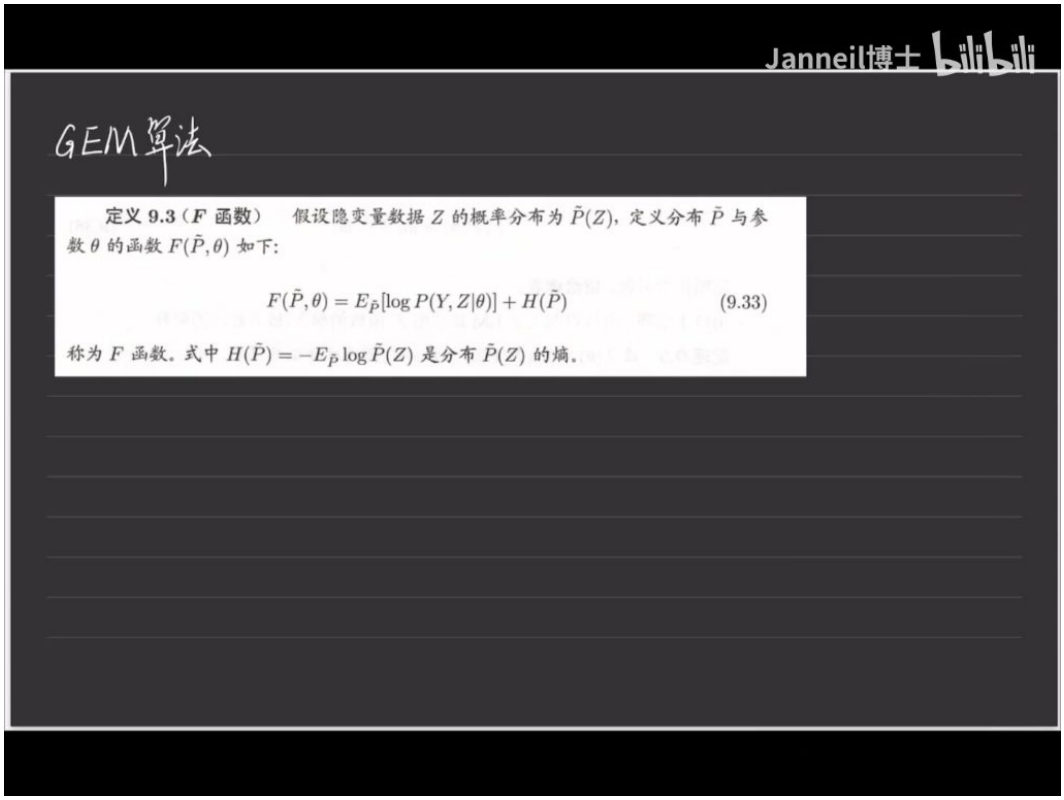
- **统计意义**: 分母$\sum \gamma_{gk}$表示属于第k个分布的样本数

2. 高斯混合模型参数估计的EM算法 03:35:39



- **算法输入**: 观测数据$y_1,...,y_N$和高斯混合模型
- **算法输出**: 模型参数$\{\alpha_k, \mu_k, \sigma_k^2\}_{k=1}^{K}$
- **E步**:
  - **响应度计算**:$\gamma_{jk} = \frac{\alpha_k \varphi(y_j|\theta_k)}{\sum_{k=1}^{K} \alpha_k \varphi(y_j|\theta_k)}$
  - **隐变量估计**: 通过贝叶斯公式补全隐变量信息
- **M步**:
  - **参数更新**:
    - $\hat{\mu}_k = \frac{\sum \gamma_{jk} y_j}{\sum \gamma_{jk}}$
    - $\hat{\sigma}_k^2 = \frac{\sum \gamma_{jk}(y_j - \mu_k)^2}{\sum \gamma_{jk}}$
    - $\hat{\alpha}_k = \frac{\sum \gamma_{jk}}{N}$
- **迭代终止**: 重复E步和M步直到收敛

3. GEM算法

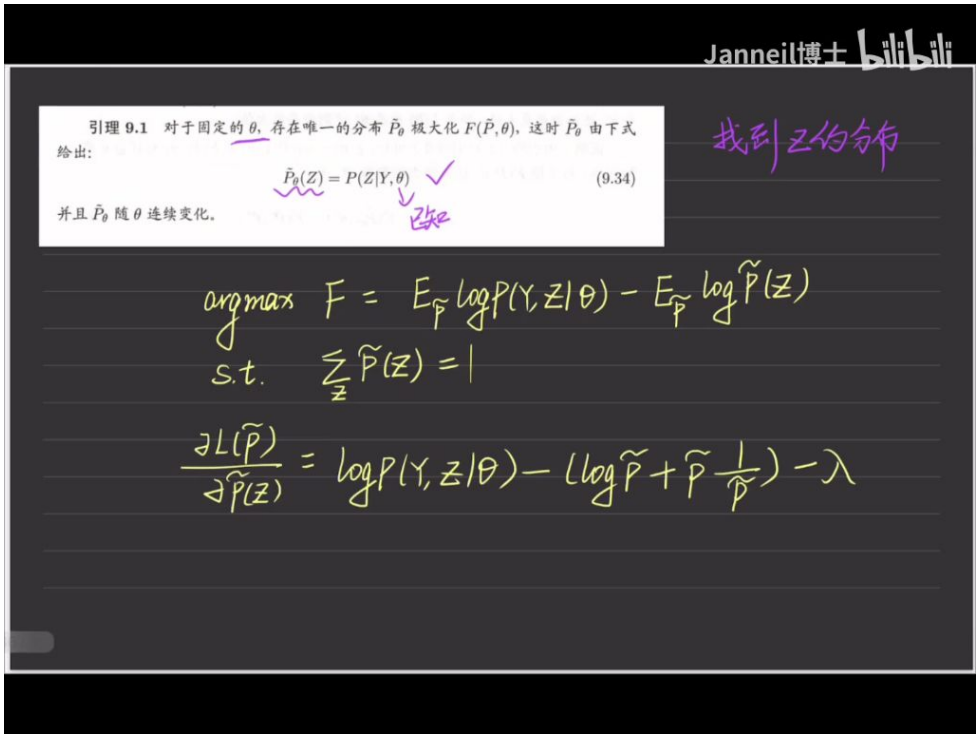


- **F函数构成**:$F(\sim P,\theta) = E_{\sim P}[logP(Y,Z|\theta)] + H(\sim P)$
  - **期望部分**: 与EM算法中的Q函数对应
  - **熵部分**:$H(\sim P) = -E_{\sim P}log\sim P(Z)$
- **算法特点**:
  - **广义形式**: 是EM算法的推广
  - **两步极大化**: 先对隐变量分布极大化，再对参数极大化
- **与EM关系**:
  - 将对数似然分解为Q函数和H函数
  - H函数对应信息熵的概念

4. GEM算法 03:37:46

1）f函数 03:38:47

- 引理9.1 03:45:03
  - 引理9.1的概述

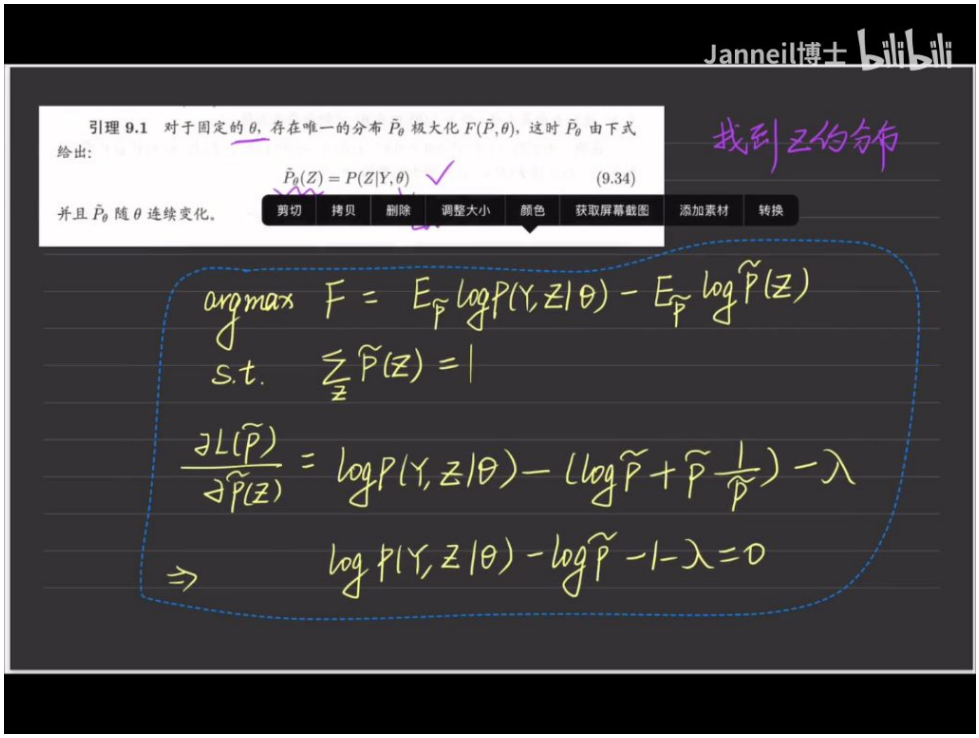


- ■
- ■ **核心结论**：对于固定参数θ，存在唯一分布$\sim P_\theta$使函数$F(\sim P,\theta)$极大化，该分布由$\sim P_\theta(Z) = P(Z|Y,\theta)$给出，且$\sim P_\theta$随$\theta$连续变化。
- ■ **物理意义**：当模型参数θ固定时，可以通过最大化F函数找到隐变量Z的最优分布，这个分布就是给定观测Y和参数θ时的条件分布。

○ 引理9.1的证明思路 03:46:10

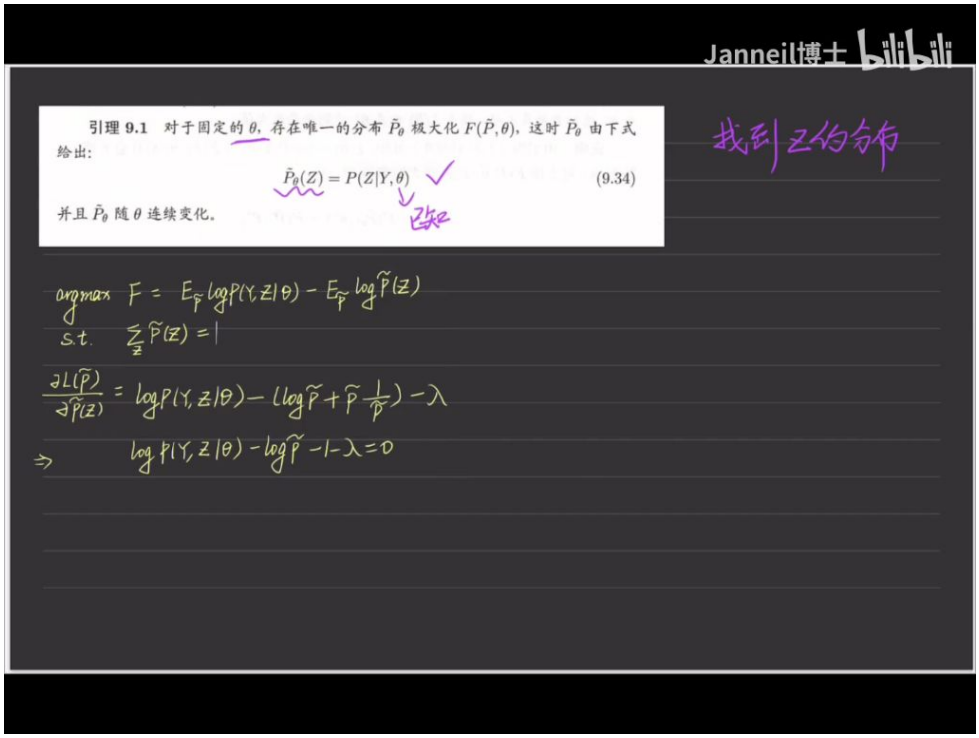
- ■ **优化目标**：将分布求解问题转化为带约束的优化问题，需要同时满足：
  - 目标函数$F(\sim P,\theta)$最大化
  - 概率分布的基本性质$\sum \sim P_\theta(Z) = 1$
- ■ **方法选择**：采用拉格朗日乘子法处理带约束的优化问题，构造泛函进行变分求解。

○ 引理9.1的证明过程 03:47:50



- **函数构造**：
  - F函数分解为两部分：Q函数（对数似然期望）和H函数（熵的负值）
  - 拉格朗日函数形式：$L = E_P[logP(Y,Z|\theta)] - E_P[logP(Z)] + \lambda(1 - \sum P(Z))$
- **求导过程**：
  - 对特定z的概率P(z)求偏导
  - 得到关键方程：$logP(Y,z|\theta) - logP(z) - 1 - \lambda = 0$
- **解的形式**：
  - $P(z) = \frac{P(Y,z|\theta)}{e^{\lambda+1}}$
  - 通过归一化约束确定λ，最终得到$P(z) = P(z|Y,\theta)$

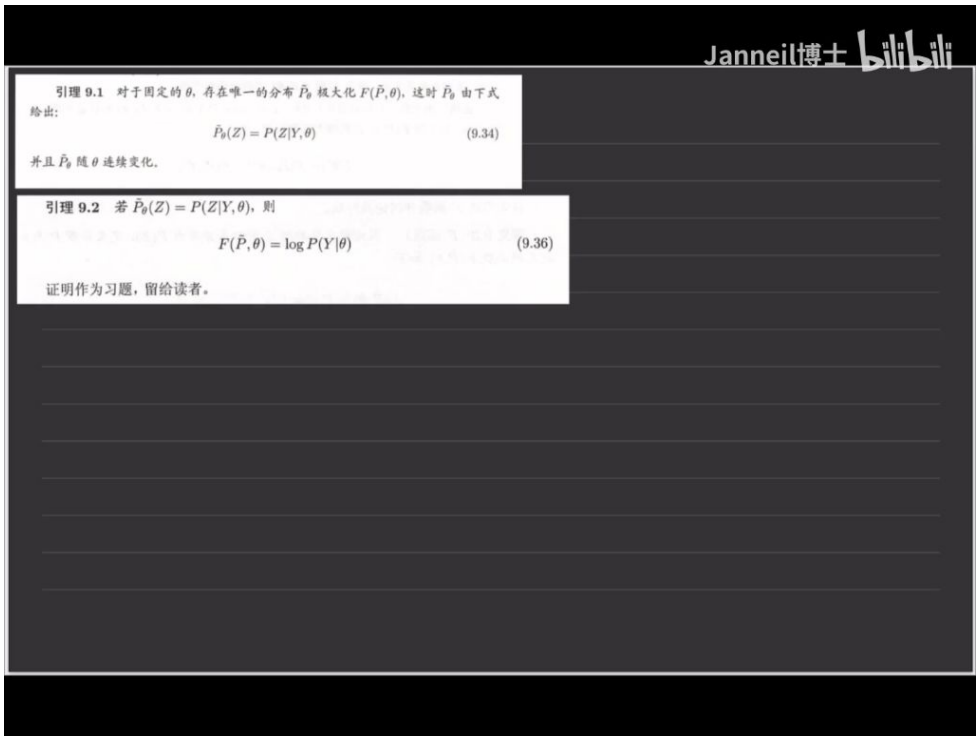
○ 引理9.1的证明收尾与结论 03:54:18



- **归一化处理**：利用概率总和为1的性质，消去拉格朗日乘子λ，最终表达式恰好是条件概率形式。
- **连续性证明**：由于条件概率$P(Z|Y,\theta)$关于θ连续，故~P_θ也随θ连续变化。

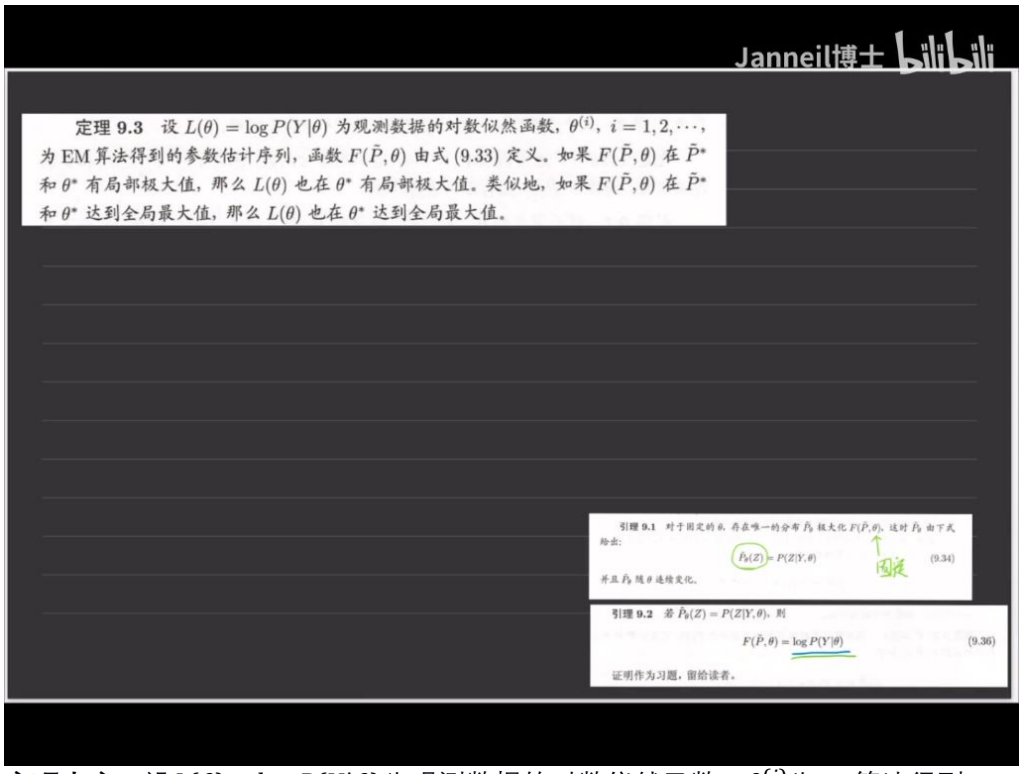
- 引理9.1与EM算法的联系 03:54:34
  - **E步对应**：在EM算法的E步中，正是固定θ求隐变量分布的过程，与引理9.1的结论完全一致。
  - **理论保证**：该引理确保了EM算法E步中隐变量分布求解的唯一性和合理性。

- 引理9.2 03:55:20
  - 引理9.2的背景与核心问题 03:55:24



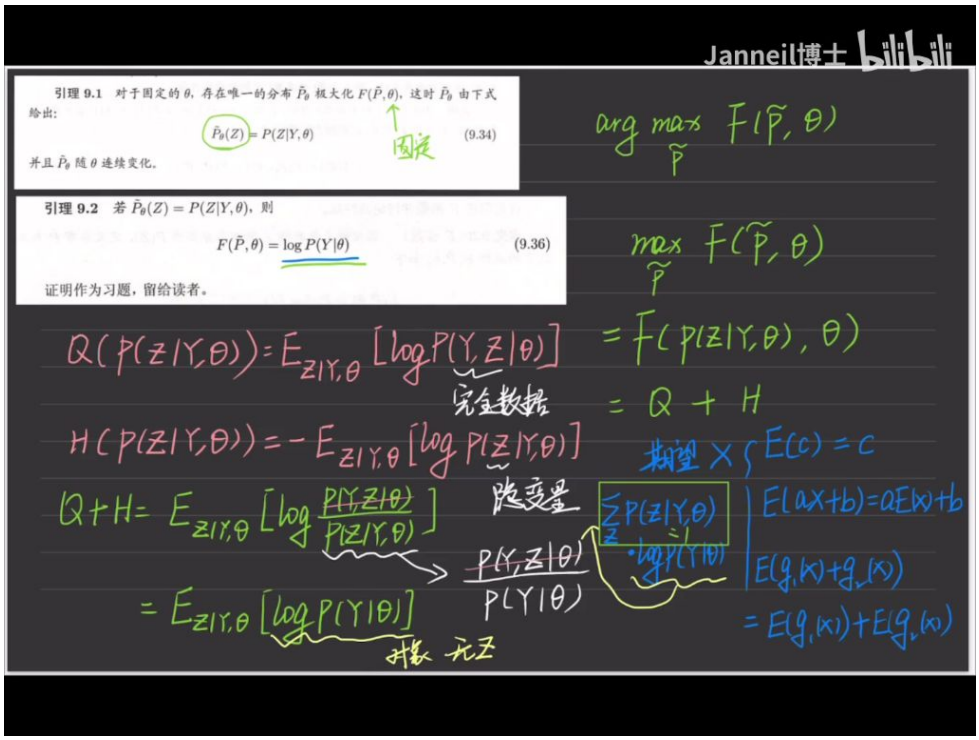
- **命题内容**：当~$P_\theta(Z) = P(Z|Y,\theta)$时，有$F(\sim P,\theta) = logP(Y|\theta)$

        - **意义解读**：说明在最优隐变量分布下，F函数的值等于观测数据的对数边际似然。
    - f函数的构成与期望性质 03:56:16
        - **函数分解**：
            - Q函数：$E[logP(Y,Z|\theta)]$
            - H函数：$-E[logP(Z)]$
        - **期望性质应用**：
            - 线性性：$E[aX+b]=aE[X]+b$
            - 常量性质：常数期望等于其自身
    - q函数与h函数的期望计算 03:56:30
        - **合并处理**：
            - $Q+H=E[log\frac{P(Y,Z|\theta)}{P(Z)}]=E[logP(Y|\theta)]$
        - **关键简化**：
            - 利用条件概率定义：$P(Z|Y,\theta)=\frac{P(Y,Z|\theta)}{P(Y|\theta)}$
            - 分子分母中的$P(Y|\theta)$与分布无关，可提出期望
    - f函数最大值的求解与结论 03:58:15
        - **最终推导**：
            - 由于$logP(Y|\theta)$不依赖Z，根据期望的常量性质，$E[logP(Y|\theta)]=logP(Y|\theta)$
            - 也可以通过展开求和式，利用概率归一性直接验证
        - **结论意义**：建立了F函数与边际似然的直接联系，为EM算法的收敛性证明奠定基础
- 定理9.3 04:03:15



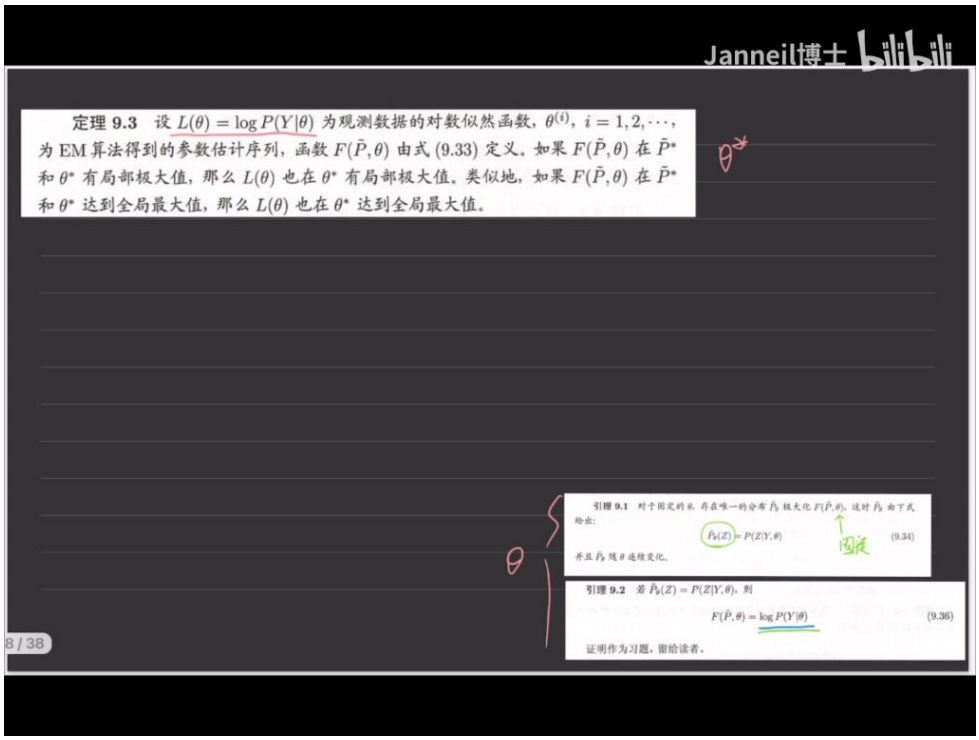
    - **定理内容**：设$L(\theta)=logP(Y|\theta)$为观测数据的对数似然函数，$\theta^{(i)}$为EM算法得到的参数估计序列，函数$F(\sim P,\theta)$由式(9.33)定义。如果$F(\mathcal{P},\theta)$在$\mathcal{P}^*$和θ^*有局部极大值，那么$L(\theta)$也在θ^*有局部极大值；类似地，如果$F(\sim P,\theta)$在$\sim P^*$和θ^*达到全局最大值，那么$L(\theta)$也在θ^*达到全局最大值。

- 引理9.1和引理9.2的内容 04:03:23



- **引理9.1**：对于固定的θ，存在唯一的分布$\mathcal{P}_\theta$极大化$F(\tilde{P},\theta)$，这时$\mathcal{P}_\theta$由$P_\theta(Z) = P(Z|Y,\theta)$给出，并且$\mathcal{P}_\theta$随$\theta$连续变化。
- **引理9.2**：若$\mathcal{P}_\theta(Z) = P(Z|Y,\theta)$，则$F(\tilde{P},\theta) = logP(Y|\theta)$。
- **关系说明**：这两个引理说明了在固定θ的情况下如何通过F函数来获得对数似然函数$L(\theta)$。

- 定理9.3第一部分证明思路 04:05:27

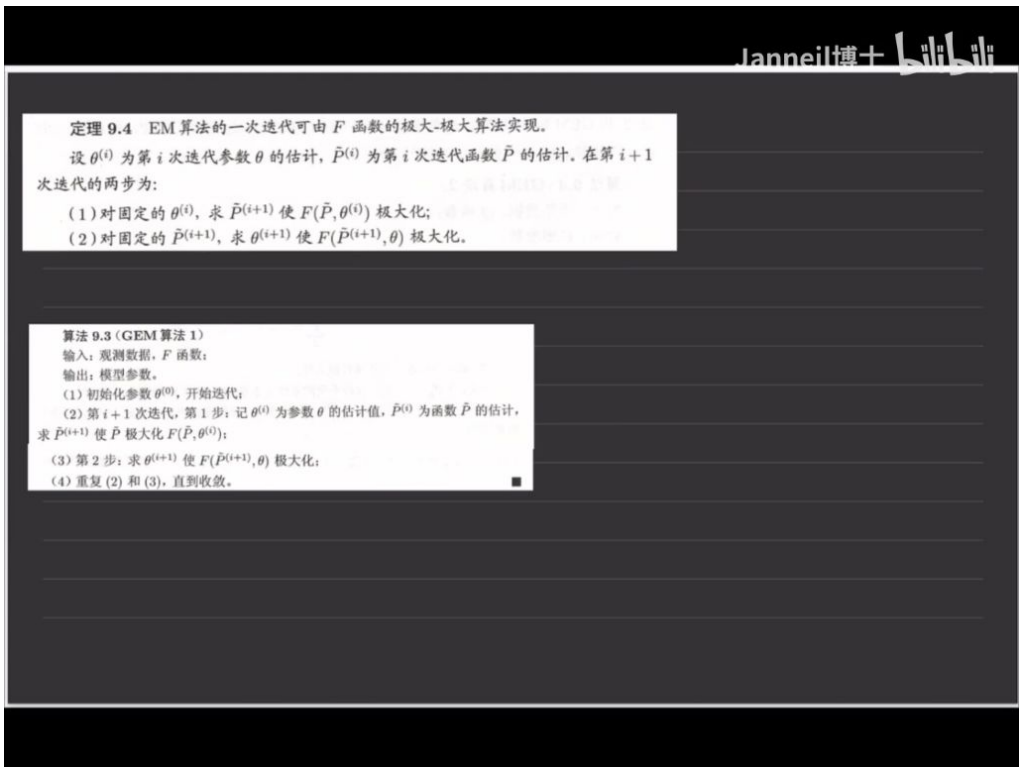


- **证明方法**：采用反证法
  - **假设**：$L(\theta)$在θ^*处未取得局部最大值，即存在θ^{**}使得$L(\theta^{**}) > L(\theta^*)$

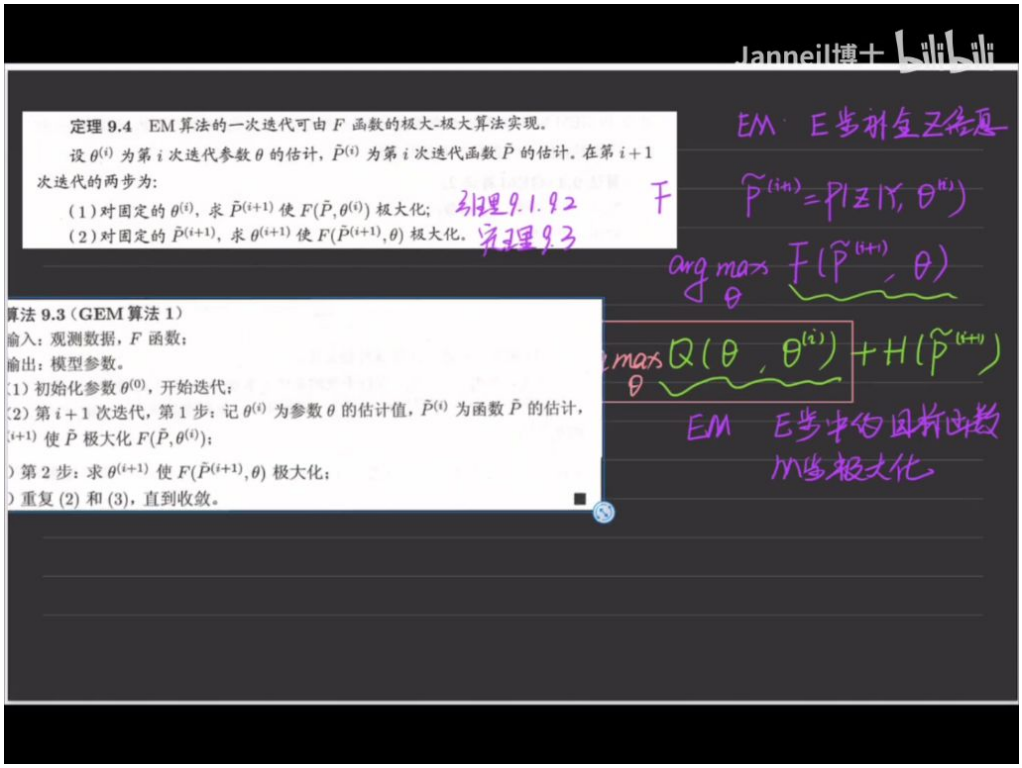
        - **推导**：根据引理9.2，这意味着$F(P^{**},\theta^{**}) > F(P^{*},\theta^{*})$
        - **矛盾**：这与F在(P^{*},θ^{*})有局部极大值的条件矛盾
    - **关键点**：利用P随θ连续变化的性质，说明θ^{**}在θ^{*}的邻域内时，(P^{**},θ^{**})也在(P^{*},θ^{*})的邻域内。
- 定理9.3第二部分证明思路 04:10:56
    - **证明方法**：同样采用反证法
        - **假设**：$L(\theta)$在θ^{*}处未取得全局最大值，即存在θ^{**}使得$L(\theta^{**}) > L(\theta^{*})$
        - **推导**：根据引理9.2，这意味着$F(P^{**},\theta^{**}) > F(P^{*},\theta^{*})$
        - **矛盾**：这与F在(P^{*},θ^{*})有全局极大值的条件矛盾
    - **区别**：与第一部分不同，这里θ^{**}可以在整个定义域内任意取值。
- 定理9.3的总结 04:12:00
    - **核心结论**：F函数的极值性质决定了$L(\theta)$的极值性质
    - **应用意义**：为EM算法中通过最大化F函数来间接最大化$L(\theta)$提供了理论依据
    - **预备知识**：该定理为后续介绍F函数的极大极大算法和广义EM算法奠定了基础

2）EM算法的一次迭代 04:12:43

- 定理9.4 04:12:47

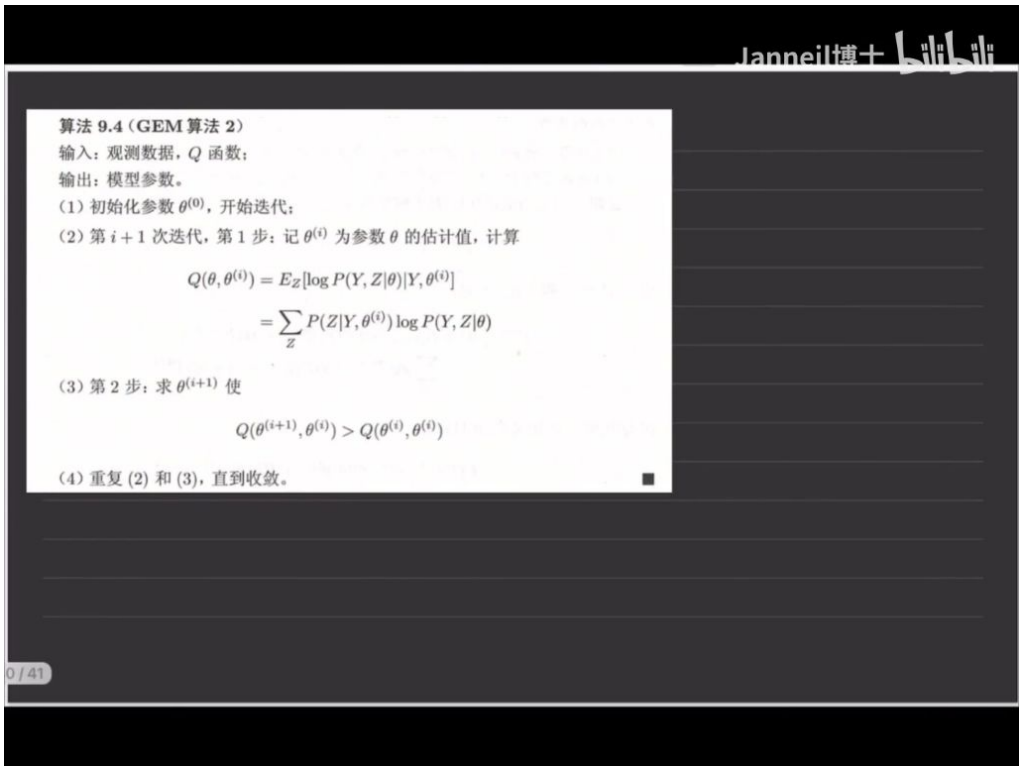


    - **迭代结构**：EM算法每次迭代包含两个步骤，通过F函数的极大-极大算法实现：
        - 固定当前参数估计$\theta^{(i)}$，求使$F(\tilde{P},\theta^{(i)})$极大化的$\sim P^{(i+1)}$
        - 固定$\sim P^{(i+1)}$，求使$F(\tilde{P}^{(i+1)},\theta)$极大化的$\theta^{(i+1)}$
    - **E步本质**：对应引理9.1-9.2，固定参数时求得的$\sim P^{(i+1)}$实际上是隐变量Z的条件概率分布$P(Z|Y,\theta^{(i)})$，这与传统EM算法中"补全隐变量信息"的目标一致。
    - **M步本质**：对应定理9.3，将F函数分解为Q函数和熵函数后，由于熵函数不含待估参数，极大化F函数等价于极大化Q函数。
- GEM算法1 04:16:23



- **算法框架**：
  - 初始化参数$\theta^{(0)}$
  - E步：固定$\theta^{(i)}$，求使$F(\tilde{P},\theta^{(i)})$极大化的$\sim P^{(i+1)}$
  - M步：固定$\sim P^{(i+1)}$，求使$F(\tilde{P}^{(i+1)},\theta)$极大化的$\theta^{(i+1)}$
  - 重复直到收敛
- **与传统EM关系**：该算法通过F函数视角重新解释EM过程，E步对应第一次极大化（补全隐变量），M步对应第二次极大化（更新参数），二者共同构成"极大-极大"过程。

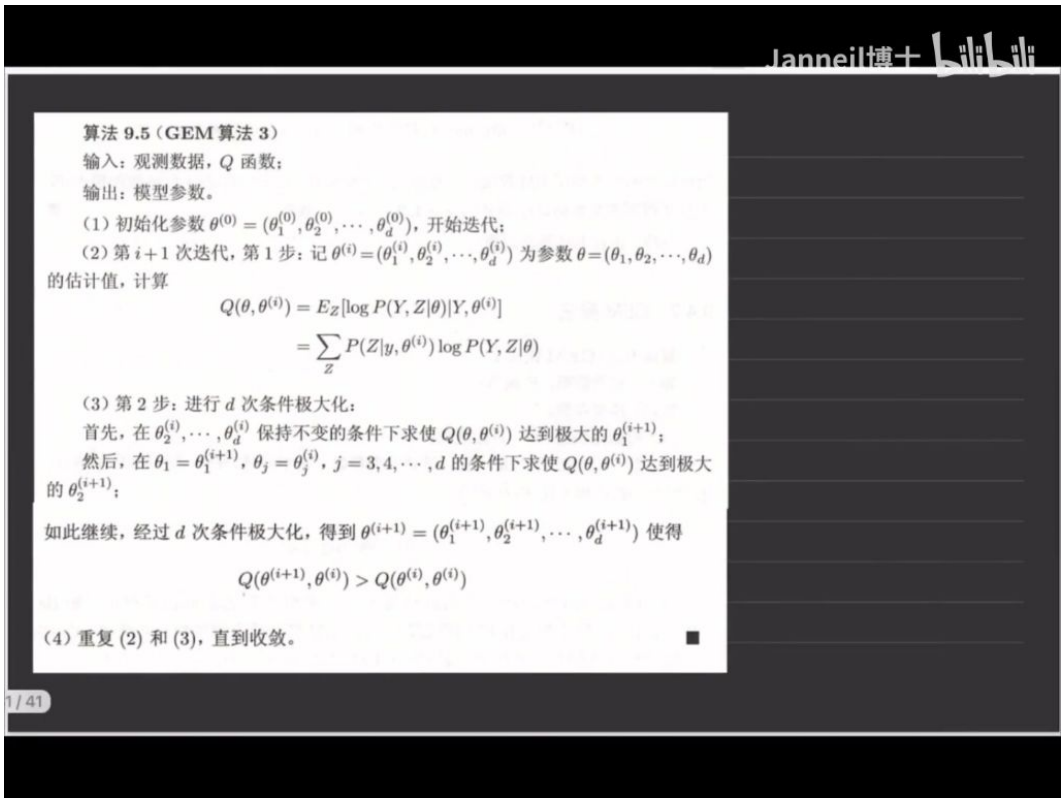
- GEM算法2 04:18:38



- **改进动机**：当参数维度较高（如高斯分布含μ和σ^2）时，直接求Q函数全局极大值困难。
- **核心思想**：退而求其次，只需找到使$Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$的参数即可。
- **算法特点**：
  - 仍保持E步计算Q函数的完整过程
  - M步放宽要求，不要求达到全局最优，只需保证似然函数值单调增加
  - 收敛速度可能较慢，但保证可行性

- GEM算法3 04:20:47
  - **优化策略**：受坐标下降法启发，采用分量更新策略：
    - 每次固定其他所有参数
    - 仅优化单个参数分量使其对应Q函数值增大
    - 轮换更新所有分量完成一轮迭代
  - **优势**：
    - 将高维优化问题分解为系列一维问题
    - 每个子问题求解简单（如高斯分布可分别优化μ和σ^2）
    - 保证整体Q函数值单调上升
  - **理论保证**：当所有分量都完成优化后，新的参数估计必然满足$Q(\theta^{(i+1)},\theta^{(i)}) \geq Q(\theta^{(i)},\theta^{(i)})$

5. 广义EM算法总结



1）算法概述

- **核心思想**：通过引入f函数实现广义EM算法，采用"极大-极大化"策略进行参数估计
- **算法类型**：基于Q函数和坐标下降法的第三种实现方式
- **关键改进**：相比标准EM算法，通过分步条件极大化提高收敛效率

2）GEM算法3具体步骤

- **输入输出**：

- 输入：观测数据Y和Q函数$Q(\theta,\theta^{(i)})$
- 输出：优化后的模型参数θ

- **初始化阶段**：
  - 设置初始参数$\theta^{(0)} = (\theta_1^{(0)},\theta_2^{(0)},\cdots,\theta_d^{(0)})$
  - 开始迭代过程
- **迭代过程**：
  - **E步计算**：
    - 计算期望$Q(\theta,\theta^{(i)}) = E_Z[logP(Y,Z|\theta)|Y,\theta^{(i)}]$
    - 具体形式：$\sum_Z P(Z|Y,\theta^{(i)})logP(Y,Z|\theta)$
  - **M步优化**：
    - 采用d次条件极大化：
      - 固定$\theta_2^{(i)},...,\theta_d^{(i)}$，优化$\theta_1^{(i+1)}$
      - 固定$\theta_1^{(i+1)}$和$\theta_j^{(i)}(j = 3,...,d)$，优化$\theta_2^{(i+1)}$
      - 依次类推完成所有参数更新
    - 确保$Q(\theta^{(i+1)},\theta^{(i)}) > Q(\theta^{(i)},\theta^{(i)})$
  - **收敛判断**：
    - 重复迭代直至满足收敛条件

3）算法特点

- **优化策略**：采用坐标下降法逐个维度优化参数
- **收敛保证**：每次迭代至少保证Q函数不减小
- **计算效率**：相比全局优化，分步优化可能减少单次迭代计算量
- **适用场景**：特别适合高维参数空间的情况

4）相关算法对比

- **算法1**：直接使用f函数的实现
- **算法2**：退而求其次使用Q函数的简化版本
- **算法3**：当前讨论的使用Q函数+坐标下降法的版本

四、知识小结

知识点	核心内容	关键公式/步骤	应用案例	难度系数
EM算法基础概念	期望最大化算法，用于含隐变量或不完整数据的参数估计	$Q(\theta,\theta^i) = E[\log P(Y,Z\|\theta)\|Y,\theta^i]$	硬币抛掷实验（Nature案例）	★★
与极大似然估计关系	EM是极大似然估计的扩展，处理完全数据时退化为MLE	$\hat{\theta} = \arg\max \log P(Y\|\theta)$	不均匀骰子概率估计	★★
E步详解	计算隐变量后验分布，构建Q函数	$\gamma_{jk}^{(i+1)} = P(z_j=k\|y_j,\theta^i)$	三硬币模型隐变量计算	★★★
M步详解	最大化Q函数更新参数	$\theta^{i+1} = \arg\max Q(\theta,\theta^i)$	高斯混合模型参数更新	★★★★

收敛性证明	通过Jensen不等式保证似然函数单调递增	$L(\theta^{i+1}) \geq L(\theta^{i})$	函数下界分析	★★★★
高斯混合模型应用	用EM算法估计多个高斯分布的混合参数	$\alpha_k = n_k/N$, $\mu_k = \Sigma\gamma_{jk}y_j/n_k$	双峰数据分布建模	★★★
GEM算法	广义EM算法，放宽M步要求	$F(p,\theta) = E_p[\log P(Y,Z\|\theta)] + H(p)$	坐标上升法实现	★★★★
算法比较	EM vs GEM：精确最大化 vs 近似优化	表格对比收敛速度/计算复杂度	不同初始化策略对比	★★★
实践注意事项	初始值敏感性/局部最优问题	多随机初始化策略	实际数据缺失处理	★★★★