（请大家预习）

# 第3章
# 概率密度估计—参数估计（第2讲）

## Estimation on PDF: Parameter Estimation

张 燕 明

ymzhang@nlpr.ia.ac.cn

people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)

多模态人工智能系统实验室 中科院自动化所

助教： 杨 奇 ( yangqi2021@ia.ac.cn )

张 涛 ( zhangtao2021@ia.ac.cn )

# 公式太多，怎么办？

- 注重宏观思维
  - 先整体，后局部，再回到整体；先简化，再回去
  - 理解概念最重要！
    - 特征空间、符号、公式的物理意义，形成直觉
    - 高维空间物理意义如何理解：简化到低维，再推广到高维
  - 注重不同方法之间的区别和联系(共性)
  - 理解概念的基础上再去了解细节和数学证明
    - 对主要的方法理解其原理、过程和结论
    - 复杂的数学证明过程可忽略，记住结论即可
      - 简单的情况要清楚细节，如高斯密度函数的最大似然估计求解过程
      - 高斯混合密度的最大似然估计(EM算法)了解主要步骤(E-step, M-step)
      - 低维空间和简单模型能写出详细过程，高维或复杂模型则不要求
  - 数学分析（形式化）和证明的能力对创新研究很重要，但不可能（没有精力）把所有细节都搞懂
    - 善于利用已有概念、原理和结论，理解和会用是基础

# 上次课主要内容回顾

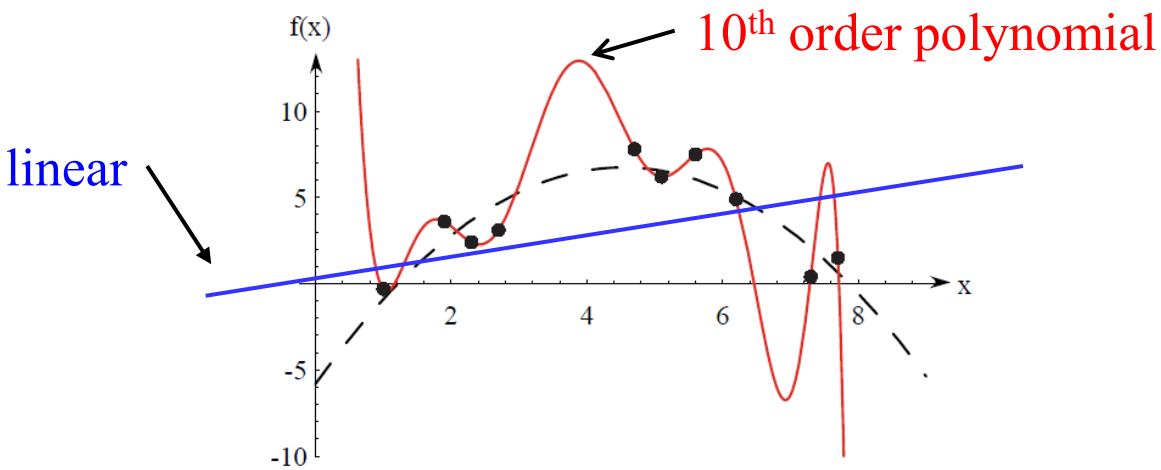
- 离散特征贝叶斯决策
  - 特征独立性假设（朴素贝叶斯）
- 复合模式分类

$$P(\omega_1,\omega_2,\ldots,\omega_n \mid \mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)=\frac{p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n \mid \omega_1,\omega_2,\ldots,\omega_n)P(\omega_1,\omega_2,\ldots,\omega_n)}{p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)}$$

- 分类与特征的关系
  - 特征给定，贝叶斯分类错误率是分类的理论最优值，与具体分类器、优化算法的选择无关
  - 特征对于分类性能至关重要：增加具有判别性的特征，可以增加类别之间的可分性，改善贝叶斯分类错误率
  - 但增加特征可能带来过拟合风险以及计算、存储开销

# 上次课主要内容回顾

- 过拟合(overfitting)
  - 现象：模型在训练数据上性能极好，在测试数据上性能不佳
  - 一般是由于模型过度复杂、训练数据不足造成的，可通过降维、正则化、降低模型复杂度等方式缓解
  - 过拟合的反面是欠拟合(underfitting)，是由于模型拟合能力不足导致的性能不佳




过拟合/欠拟合的背后是泛化性和模型选择问题，是PRML的一个核心问题（Chapter 8）

# 上次课主要内容回顾

- 最大似然参数估计(Maximum Likelihood)
  - 参数θ固定但未知，估计参数的全部信息来自数据

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} p(D\,|\,\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \prod_{i=1}^{n} p(\mathbf{x}_i\,|\,\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} \ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})$$

$p(\mathbf{x}_i|\boldsymbol{\theta})$：样本似然，$p(D|\boldsymbol{\theta})$：数据集似然

  - 求解：

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0, \quad \text{or} \quad \frac{\partial H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

一般通过梯度上升法迭代求解；个别简单问题可解析求解。

# 上次课主要内容回顾

- 贝叶斯估计
  - 将参数θ视作随机变量
  - 参数先验$p(\boldsymbol{\theta})$：与观测数据无关的先验知识
  - 参数后验：融合先验知识与数据信息

$$p(\boldsymbol{\theta}\,|\,D)=\frac{p(D\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}=\frac{p(D\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}}p(D\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

  - 数据后验：

$$p(\mathbf{x}\,|\,D)=\int_{\boldsymbol{\theta}}p(\mathbf{x},\boldsymbol{\theta}\,|\,D)d\boldsymbol{\theta}=\boxed{\int_{\boldsymbol{\theta}}p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}\,|\,D)d\boldsymbol{\theta}}$$

  - 最大后验估计(Maximum A Posterior)：

$$\hat{\boldsymbol{\theta}}_{MAP}=\arg\max_{\boldsymbol{\theta}\in\Theta}p(\boldsymbol{\theta}\,|\,D)=\arg\max_{\boldsymbol{\theta}\in\Theta}\prod_{i=1}^{n}p(\mathbf{x}_i\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})=\arg\max_{\boldsymbol{\theta}\in\Theta}\boxed{\sum_{i=1}^{n}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})+\ln p(\boldsymbol{\theta})}$$

$$\hat{\boldsymbol{\theta}}_{ML}=\arg\max_{\boldsymbol{\theta}\in\Theta}p(D\,|\,\boldsymbol{\theta})=\arg\max_{\boldsymbol{\theta}\in\Theta}\prod_{i=1}^{n}p(\mathbf{x}_i\,|\,\boldsymbol{\theta})=\arg\max_{\boldsymbol{\theta}\in\Theta}\boxed{\sum_{i=1}^{n}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})}$$

ML和MAP的优化目标仅相差$\ln p(\boldsymbol{\theta})$

# 提 纲

- 第3章：参数估计

  (贝叶斯分类的参数法、半参数法)

  – 期望最大法(EM)
    - 一般情况
    - EM for Gaussian Mixture
    - EM for incomplete data
  – 隐马尔可夫模型(HMM)
    - HMM的表示
    - HMM的学习
    - HMM的解码

# 3.8 期望最大化 (Expectation-Maximization, EM)

- 数据属性完整
  - 数据是完整的，即不缺少任何特征属性，可直接应用最大似然估计、贝叶斯估计等方法实现参数估计。
- 数据属性不完整
  - 典型情形1： 一些样本的属性缺失

| | Day1 | Day2 | Day3 | Day4 | Day5 |
|---|---|---|---|---|---|
| 气温 | 22° C | 20° C | 19° C | 23° C | 27° C |
| 相对湿度 | 58% | 46% | * | * | 65% |
| 风力 | 4级 | 3级 | * | 3级 | 2级 |
| 空气质量 | 53 | * | * | 98 | 47 |

$\mathbf{x}_1=(22,0.58,4,53)$ $\mathbf{x}_2=(20,0.46,3,*)$ $\mathbf{x}_3=(19,*,*,*)$ $\mathbf{x}_4=(23,*,3,98)$ $\mathbf{x}_5=(27,0.65,2,47)$

# 3.8 期望最大化 (Expectation-Maximization, EM)

- 数据属性不完整
  - 典型情形2：一些隐藏属性不可观测

病人：（症状描述，血常规，体温，血压，心/肝/肺等器官的状态）

导师：（教育背景，文章数，引用量，头衔，学术水平/人品）

可观测　　不可观测

✓ 隐藏属性又称为隐变量(latent variables)，对于所有样本均不可见

✓ 隐变量模型是PR/ML中非常重要的一类模型，如HMM, MRF, VAE, Diffusion Model

EM算法是对属性不完整数据进行参数估计的一种有效方法

# 3.8.1 EM 算法：一个例子

- 有A, B, C三枚硬币，掷出正面的概率分别为$p_A, p_B$, 0.5。重复5次如下实验：掷1次C；若C为正面，则掷10次A；若C为背面，则掷10次B。正面用1表示，背面用0表示。问题：基于实验结果估计$p_A, p_B$

# 3.8.1 EM 算法：一个例子

- 有A, B, C三枚硬币，掷出正面的概率分别为$p_A, p_B$, 0.5。重复5次如下实验：掷1次C；若C为正面，则掷10次A；若C为背面，则掷10次B。正面用1表示，背面用0表示。问题：基于实验结果估计$p_A, p_B$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | C=1 (A) | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |

# 3.8.1 EM 算法：一个例子

- 有A, B, C三枚硬币，掷出正面的概率分别为$p_A, p_B$, 0.5。重复5次如下实验：掷1次C；若C为正面，则掷10次A；若C为背面，则掷10次B。正面用1表示，背面用0表示。问题：基于实验结果估计$p_A, p_B$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | C=1 (A) | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 第2次 | C=1 (A) | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

# 3.8.1 EM 算法：一个例子

- 有A, B, C三枚硬币，掷出正面的概率分别为$p_A, p_B$, 0.5。重复5次如下实验：掷1次C；若C为正面，则掷10次A；若C为背面，则掷10次B。正面用1表示，背面用0表示。问题：基于实验结果估计$p_A, p_B$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | C=1 (A) | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 第2次 | C=1 (A) | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 第3次 | C=0 (B) | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 第4次 | C=1 (A) | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 第5次 | C=0 (B) | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |

# 3.8.1 EM 算法：一个例子

- 根据二项分布的极大似然估计

| | | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | C=1 (A) | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | C=1 (A) | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | C=0 (B) | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | C=1 (A) | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | C=0 (B) | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

二项分布的极大似然估计： $\hat{p} = \dfrac{\#\text{正面次数}}{\#\text{总投掷次数}}$

$$\hat{p}_A = \frac{\#\text{正面次数}}{\#\text{总投掷次数}} = \frac{3+2+3}{10+10+10} = 0.267$$

$$\hat{p}_B = \frac{\#\text{正面次数}}{\#\text{总投掷次数}} = \frac{6+8}{10+10} = 0.7$$

# 3.8.1 EM 算法：一个例子

- 实验2：有A, B, C三枚硬币，掷出正面的概率分别为$p_A, p_B$, 0.5。重复5次如下实验：掷1次C；若C为正面，则掷10次A；若C为背面，则掷10次B。但是，C不可观测。问题：基于实验结果估计$p_A, p_B$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |

- 实验2
  - 设随机变量 $x$ 表示硬币A/B的正面或背面。$x=1$，表示正面；$x=0$，表示背面。（可观测）
  - 设随机变量 $z$ 表示硬币C的正面或背面。（不可观测）
    - $z=1$，表示正面（对应硬币A）
    - $z=0$，表示背面（对应硬币B）
  - $P(z=1)=0.5;\ P(z=0)=0.5$
  - 记 $\boldsymbol{\theta}=(p_A, p_B)^T \in R^2$
  - 考虑第 1 次试验，观测到10个 $x_i$, 于是有

$$
\begin{aligned}
P(X_1) &= P(x_1, x_2, \ldots, x_{10} \mid \boldsymbol{\theta}) \\
&= \sum_z P(x_1, x_2, \ldots, x_{10}, z \mid \boldsymbol{\theta}) \\
&= P(x_1, x_2, \ldots, x_{10} \mid z=1, \boldsymbol{\theta}) P(z=1) + P(x_1, x_2, \ldots, x_{10} \mid z=0, \boldsymbol{\theta}) P(z=0) \\
&= P(z=1) \prod_{i=1}^{10} P(x_i \mid z=1) + P(z=0) \prod_{i=1}^{10} P(x_i \mid z=0) \\
&= 0.5 \prod_{i=1}^{10} p_A^{x_i} (1-p_A)^{1-x_i} + 0.5 \prod_{i=1}^{10} p_B^{x_i} (1-p_B)^{1-x_i}
\end{aligned}
$$

- 实验2
  - 考虑掷了5次硬币$C$, 一共有5组实验(每组10次），此时联合概率为：

$$P\left(X_1,X_2,X_3,X_4,X_5\right)=P(x_1,x_2,\ldots,x_{10}\mid\boldsymbol{\theta})\times P(x_{11},x_{12},\ldots,x_{20}\mid\boldsymbol{\theta})$$
$$\times P(x_{21},x_{22},\ldots,x_{30}\mid\boldsymbol{\theta})\times P(x_{31},x_{32},\ldots,x_{40}\mid\boldsymbol{\theta})$$
$$\times P(x_{41},x_{42},\ldots,x_{50}\mid\boldsymbol{\theta})$$

$$\ln\left(P\left(X_1,X_2,X_3,X_4,X_5\right)\right)=\ln\left(\underline{0.5\prod_{i=1}^{10}p_A^{\,x_i}(1-p_A)^{1-x_i}+0.5\prod_{i=1}^{10}p_B^{\,x_i}(1-p_B)^{1-x_i}}\right)$$
$$+\ln\left(\underline{0.5\prod_{i=11}^{20}p_A^{\,x_i}(1-p_A)^{1-x_i}+0.5\prod_{i=11}^{20}p_B^{\,x_i}(1-p_B)^{1-x_i}}\right)$$
$$+\cdots$$

✓ 最大似然估计：求导数时$p_A$和$p_B$不能分离，无解析解，需要迭代更新！

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

第一步：随机初始化$p_A, p_B$。 比如，令$p_A = 0.46, p_B = 0.61$

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

第一步：随机初始化$p_A, p_B$。 比如，令$p_A = 0.46, p_B = 0.61$

第二步，根据第一次掷$C$后10次试验， 猜一个$C$的取值。

✓ 基于$p_A = 0.46, p_B = 0.61$，猜C 的取值为 正面(对应硬币A) ----为什么？

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

第一步：随机初始化$p_A, p_B$。 比如，令$p_A = 0.46, p_B = 0.61$

第二步，根据第一次掷$C$后10次试验， 猜一个$C$的取值。

- ✓ 基于$p_A = 0.46, p_B = 0.61$，猜$C$ 的取值为 正面(对应硬币A) ----为什么？
- ✓ 如果$C$为正面，则掷10次硬币A，联合概率为 $\mathbf{0.46^3 * (1-0.46)^7 = 0.0013}$
- ✓ 如果$C$为背面，则掷10次硬币B，联合概率为 $\mathbf{0.61^3 * (1-0.61)^7 = 0.0003}$

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

第一步：随机初始化$p_A, p_B$。 比如，令$p_A = 0.46, p_B = 0.61$

第二步，根据第一次掷$C$后10次试验， 猜一个$C$的取值。

- ✓ 基于$p_A = 0.46, p_B = 0.61$，猜$C$ 的取值为 正面(对应硬币A) ----为什么？
- ✓ 如果$C$为正面，则掷10次硬币A，联合概率为 $\mathbf{0.46^3 * (1-0.46)^7 = 0.0013}$
- ✓ 如果$C$为背面，则掷10次硬币B，联合概率为 $\mathbf{0.61^3 * (1-0.61)^7 = 0.0003}$

第三步：根据猜测的$C$更新 $p_A$： $p_A = 3/10 = 0.3, p_B = 0.61$

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

**第四步：根据第二次掷$C$后10次试验， 猜一个$C$的取值 。**

- ✓ 如果$C$为正面，则掷10次硬币A，联合概率为 $\mathbf{0.3^2 * (1-0.3)^8 = 0.0052}$
- ✓ 如果$C$为背面，则掷10次硬币B，联合概率为 $\mathbf{0.61^3 * (1-0.61)^7 = 0.0003}$
- ✓ 所以猜$C$ 的取值为 正面(对应硬币A) 。

**第五步：根据猜测的$C$更新$p_A$：** $p_A = 5/20 = 0.25, p_B = 0.61$

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

**第六步：根据第三次掷$C$后10次试验， 猜一个$C$的取值 。**

- ✓ 如果$C$为正面，则掷10次硬币A，联合概率为 $\mathbf{0.25^6 * (1\text{-}0.25)^4 < 0.0001}$
- ✓ 如果$C$为背面，则掷10次硬币B，联合概率为 $\mathbf{0.61^6 * (1\text{-}0.61)^4 = 0.0012}$
- ✓ 所以猜$C$ 的取值为 背面(**对应硬币B**) 。

**第七步：根据猜测的$C$更新$p_B$：** $p_A = 0.25, p_B = 6/10 = 0.6$

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

**第八步：根据第四次掷$C$后10次试验， 猜一个$C$的取值 。**

- ✓ 如果$C$为正面，则掷10次硬币A，联合概率为 $\mathbf{0.25^3 * (1\text{-}0.25)^7 > 0.002}$
- ✓ 如果$C$为背面，则掷10次硬币B，联合概率为 $\mathbf{0.6^3 * (1\text{-}0.6)^7 < 0.0004}$
- ✓ 所以猜$C$ 的取值为 正面(**对应硬币A**) 。

**第九步：根据猜测的$C$更新$p_A$ ：** $p_A = 8/30 = 0.267, p_B = 0.6$

# 3.8.1 EM 算法：一个例子

- 试验2

| | | | | | | | | | | | #正面 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 第1次 | | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3 |
| 第2次 | | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |
| 第3次 | | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 6 |
| 第4次 | | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 |
| 第5次 | | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 8 |

第十步：根据第五次掷$C$后10次试验， 猜一个$C$的取值 。

- ✓ 如果$C$为正面，则掷10次硬币A，联合概率为 $(8/30)^8 * (1-8/30)^2 < 0.00002$
- ✓ 如果$C$为背面，则掷10次硬币B，联合概率为 $0.6^8 * (1-0.6)^2 = 0.027$
- ✓ 所以猜$C$ 的取值为 背面(对应硬币B) 。

第十一步：根据猜测的$C$更新 $p_B$：$p_A$=0.267, $p_B$ =14/20 = 0.7

# 3.8.1 EM 算法

- 对上述方法的总结
  - 每次迭代包含两个基本步骤：

(1) 基于当前参数$p_{\mathrm{A}}, p_{\mathrm{B}}$，猜隐变量$C$的取值

(2) 基于猜到的隐变量$C$取值，更新参数$p_{\mathrm{A}}, p_{\mathrm{B}}$

这就是EM算法的基本框架

  - 每次猜出的C可能不准确，可以做两点升级：
    - 不猜C的值，而是估计C的后验分布
    - 多次迭代，反复估计C的后验分布和分布参数$p_A, p_B$

# 3.8.1 EM 算法

- 期望最大化 (Expectation Maximization, EM)算法
  - Dempster-Laird-Rubin算法
    - Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), pp.1-38.
  - EM是一类通过迭代实现参数估计的优化算法
    - 作为最大似然法的替代，用于对包含隐变量(latent variable)或缺失数据(incomplete data)的概率模型进行参数估计。

# 3.8.1 EM 算法

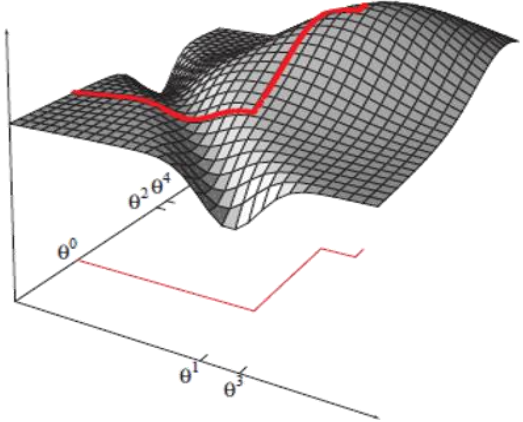
- EM算法解决的问题：包含隐变量的概率密度参数估计
  - 观测变量：$\mathbf{x}$
  - 隐含变量：$\mathbf{z}$
  - 任务： 给定数据集$\{\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_n\}$，估计观测数据概率密度的参数
- EM算法的基本概念
  - 观测数据： $\{\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_n\}$ （不完全数据）
  - 隐含数据： $\{\mathbf{z}_1, \mathbf{z}_2,\ldots, \mathbf{z}_n\}$
  - 完全数据：$\{(\mathbf{x}_1, \mathbf{z}_1), (\mathbf{x}_2, \mathbf{z}_2), \ldots, (\mathbf{x}_n, \mathbf{z}_n)\}$
  - 观测数据的概率密度 ：$p(\mathbf{x} \mid \boldsymbol{\theta})$
  - 完全数据的联合概率密度： $p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta})$
  - 观测数据的对数似然：

$$\ln \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \boldsymbol{\theta})$$

  - 完全数据的对数似然：

$$\ln \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta})$$

# 3.8.1 EM 算法




- EM算法步骤
  - 初始化 $\boldsymbol{\theta}^{old}$
  - Repeat
    - E step: 基于当前$\boldsymbol{\theta}^{old}$，估计每个样本隐变量的后验分布$p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})$
    - M step: 基于当前所估计的$p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta}^{old})$ 更新参数$\boldsymbol{\theta}$：

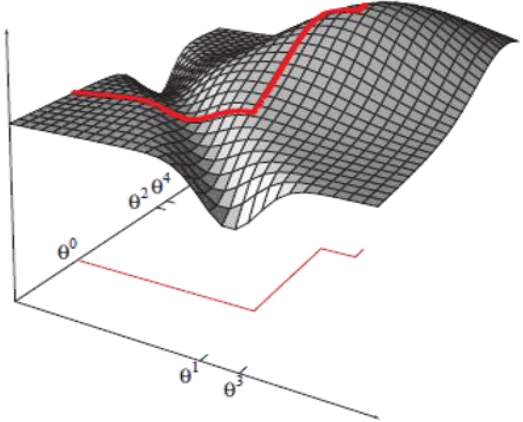
$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_i E_{p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})} \left[ \underline{\ln\left(p\left(x_i, z_i \mid \boldsymbol{\theta}\right)\right)} \right]$$

$$= \sum_i \sum_{z_i} p(z_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{old}) \ln\left(p\left(x_i, z_i \mid \boldsymbol{\theta}\right)\right)$$

    - $t=t+1$

完全数据的对数似然

# 3.8.1 EM 算法




- EM算法步骤（另一种等价表述）
  - 初始化 $\boldsymbol{\theta}^{old}$
  - Repeat
    - E(Expectation) step: 基于当前$\boldsymbol{\theta}^{old}$ 和样本，估计隐变量的后验分布$p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})$，并计算$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$：

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{old}\right)=\sum_i E_{p(\mathbf{z}_i|\mathbf{x}_i,\boldsymbol{\theta}^{old})}\left[\ln\left(p\left(x_i,z_i\mid\boldsymbol{\theta}\right)\right)\right]$$

$$=\sum_i\sum_{z_i} p(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old})\ \ln\left(p\left(x_i,z_i\mid\boldsymbol{\theta}\right)\right)$$

    - M(Maximization) step: 更新参数$\boldsymbol{\theta}$：

$$\boldsymbol{\theta}^{new}=\arg\max_{\boldsymbol{\theta}}\ Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$$

    - $t=t+1$

# 3.8.2 EM for Gaussian mixture model

- 混合密度模型
  - 混合密度模型由 $K$ 个不同成分组成
  - 每个成分对应一个概率密度函数： $p(\mathbf{x}|\,\theta_k)$
  - 每个成分的权重为$\pi_k$， $k = 1, 2, \ldots, K$，且满足：$\sum_{k=1}^{K}\pi_k = 1 \quad \forall k : \pi_k \geq 0$
  - 称以下密度函数为混合密度模型：

$$p(\mathbf{x}\,|\,\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{k=1}^{K}\pi_k\, p(\mathbf{x}\,|\,\theta_k)$$

其中，$\boldsymbol{\theta}=\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2,\ldots, \boldsymbol{\theta}_K\}$，$\boldsymbol{\pi}=\{\pi_1, \pi_2,\ldots, \pi_K\}$是混合密度模型的参数。

- 混合密度模型的参数估计
  - 已知样本集$D=\{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$，且样本是从以上混合密度函数中独立抽取的。通过$D$估计( $\boldsymbol{\pi}, \boldsymbol{\theta}$ )。

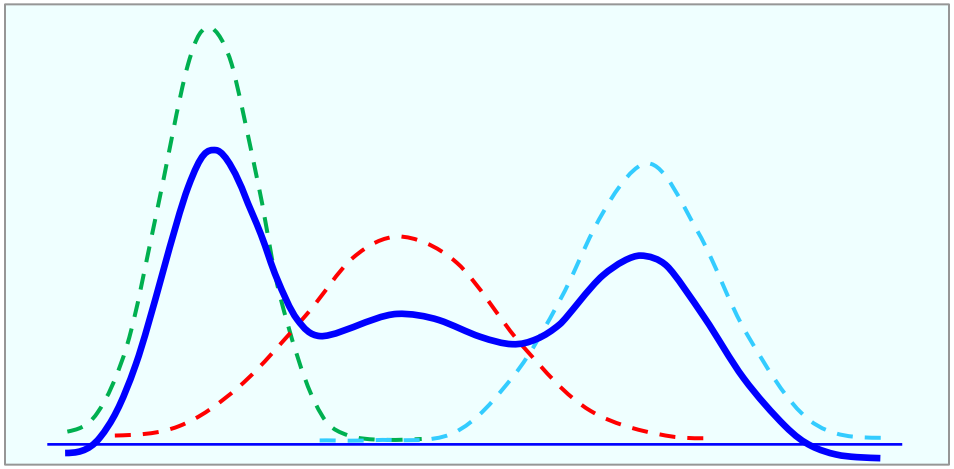
# 3.8.2 EM for Gaussian mixture model

- 称成分密度为高斯密度的混合模型为高斯混合模型：

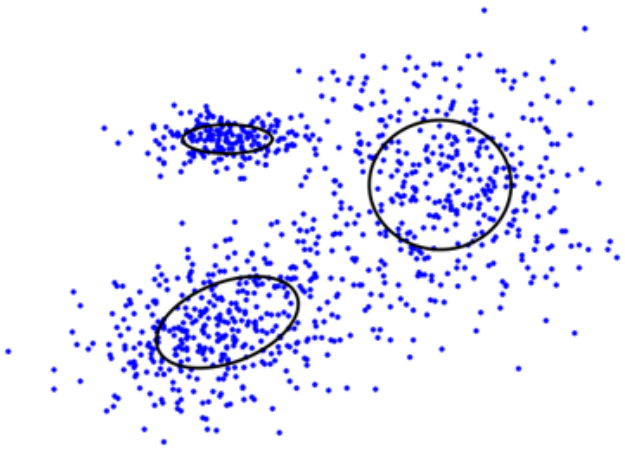
$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}\,|\,\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \pi_k \boxed{\mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

高斯密度函数

- 举例：



三个一维高斯分布的混合



三个二维高斯分布的混合

混合密度模型，可以表示复杂的分布

# 3.8.2 EM for Gaussian mixture model

- Gaussian mixture model (GMM)：

$$\sum_{k=1}^{K}\pi_k = 1 \quad \forall k : \pi_k \geq 0$$

$$p(\mathbf{x}\,|\,\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}\,|\,\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

  – GMM中的参数：

权重参数：$\pi_k$，成分参数：$\mu_k$ , $\Sigma_k$ $(k=1, 2, \ldots, K)$

- 参数估计：Maximum Likelihood (ML)

$$\max L(\theta) = \ln \prod_{i=1}^{n} p(\mathbf{x}_i) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\nabla_{\pi_k} L(\theta) = 0, \quad \nabla_{\mu_k} L(\theta) = 0, \quad \nabla_{\Sigma_k} L(\theta) = 0$$

可通过梯度上升迭代求解，但不能解析求解

# 3.8.2 EM for Gaussian mixture model

- 换个角度分析混合密度模型：引入隐变量 $z$ 用于指示不同的成分密度，$z \in \{1, \ldots, K\}$

假设：$P(z=k)=\pi_k$，　$p(\mathbf{x} \mid z=k)=\mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

则观测变量$\boldsymbol{x}$和隐变量$z$的联合分布：

$$p(\mathbf{x}, z=k)=p(\mathbf{x} \mid z=k) P(z=k)=\pi_k \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

观测变量$\mathbf{x}$的边缘分布：

$$p(\mathbf{x})=\sum_{z=1}^{K} p(\mathbf{x}, z)=\sum_{z=1}^{K} \pi_z \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$$

因此，可以将混合密度模型看成包含隐变量$z$的概率密度函数，从而用EM算法估计参数。

# 3.8.2 EM for Gaussian mixture model

– Incomplete data X, complete data {X, Z}

Missing the latent value $z$ for each sample $z \in \{1, 2, \ldots, K\}$

– Expectation of complete data log-likelihood

1. Choose an initial set of parameters for $\boldsymbol{\theta}^{old}$
2. Do

   E-step: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old})$

   M-step: Update parameters: $\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{old}) \ln\left(p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\right)$$

   If convergence condition is not satisfied

$$\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$$

3. End

(C.M. Bishop, *Pattern Recognition and Machine Learning*, 2006)

# 3.8.2 EM for Gaussian mixture model

– E Step: 固定当前估计的参数$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, 对每个样本求 $P(z_i|\mathbf{x}_i, \boldsymbol{\theta}^{old})$ ， $i = 1, 2, \ldots, n$

$$P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) = \frac{p(\mathbf{x}_i, z_i \mid \boldsymbol{\theta}^{old})}{p(\mathbf{x}_i \mid \boldsymbol{\theta}^{old})} = \frac{\pi_{z_i}\mathcal{N}\left(\mathbf{x}_i, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}\right)}{\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}$$

$(z_i = 1, 2, \ldots, or, K)$

当前估计值

- M Step: 固定$\{P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old})\}$, 通过 $\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$ 更新参数$\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$

$$
\begin{aligned}
Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) &= \sum_i \sum_{z_i=1:K} P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ \ln\Big(p\big(\mathbf{x}_i, z_i \,|\, \boldsymbol{\theta}\big)\Big) \\
&= \sum_i \sum_{z_i=1:K} P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ \ln\Big(\pi_{z_i} \mathcal{N}(\mathbf{x}_i \,|\, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})\Big) \\
&= \sum_i \sum_{z_i=1:K} P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ \Big(\ln \pi_{z_i} + \ln \mathcal{N}(\mathbf{x}_i \,|\, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})\Big) \\
&= \sum_i \sum_{z_i=1:K} \Big(P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ln \pi_{z_i} + P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ln \mathcal{N}(\mathbf{x}_i \,|\, \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})\Big) \\
&= \sum_i \sum_{z_i=1:K} P(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ln \pi_{z_i} + \sum_{k=1:K} \sum_i P(z_i = k|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ln \mathcal{N}(\mathbf{x}_i \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned}
$$

$$
\frac{\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})}{\partial \boldsymbol{\mu}_k} = \mathbf{0}, \quad \frac{\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})}{\partial \boldsymbol{\Sigma}_k} = \mathbf{0}, \quad k = 1, 2, \cdots, K
$$

$$
\arg\max_{\pi} \sum_i \sum_{z_i=1:K} p(z_i|\mathbf{x}_i,\boldsymbol{\theta}^{old}) \ \ln\Big(\pi_{z_i}\Big) \quad \text{s.t.} \ \sum_{k=1:K} \pi_k = 1
$$

# 3.8.2 EM for Gaussian mixture model

成分权重：
$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} P(z_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})$$

成分均值：
$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^{n} P(z_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})\mathbf{x}_i}{\sum_{i=1}^{n} P(z_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})}$$

成分协方差矩阵：
$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{n} P(z_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T}{\sum_{i=1}^{n} P(z_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{old})}$$

（$\hat{\boldsymbol{\theta}}$记录所有的未知参数）

$$P(z_i = k \mid \mathbf{x}_i, \boldsymbol{\theta}^{old}) = \frac{p(\mathbf{x}_i, z_i = k \mid \boldsymbol{\theta}^{old})}{p(\mathbf{x}_i \mid \boldsymbol{\theta}^{old})} = \frac{p(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)\hat{\pi}_k}{\sum_{j=1}^{K} p(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)\hat{\pi}_j}$$

（所以要迭代）

- 在极端情况下，即当样本 $\mathbf{x}_i$ 来自于一个成分时，其后验概率 $\hat{P}(\omega_k \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})$ 为 1，否则就为零，此时有：

**后验分布变成one-hot:**

$$P(z_i \mid \mathbf{x}_i, \hat{\boldsymbol{\mu}}) = \begin{cases} 1, & \mathbf{x}_i \in \omega_k \\ 0, & \mathbf{x}_i \notin \omega_k \end{cases}$$

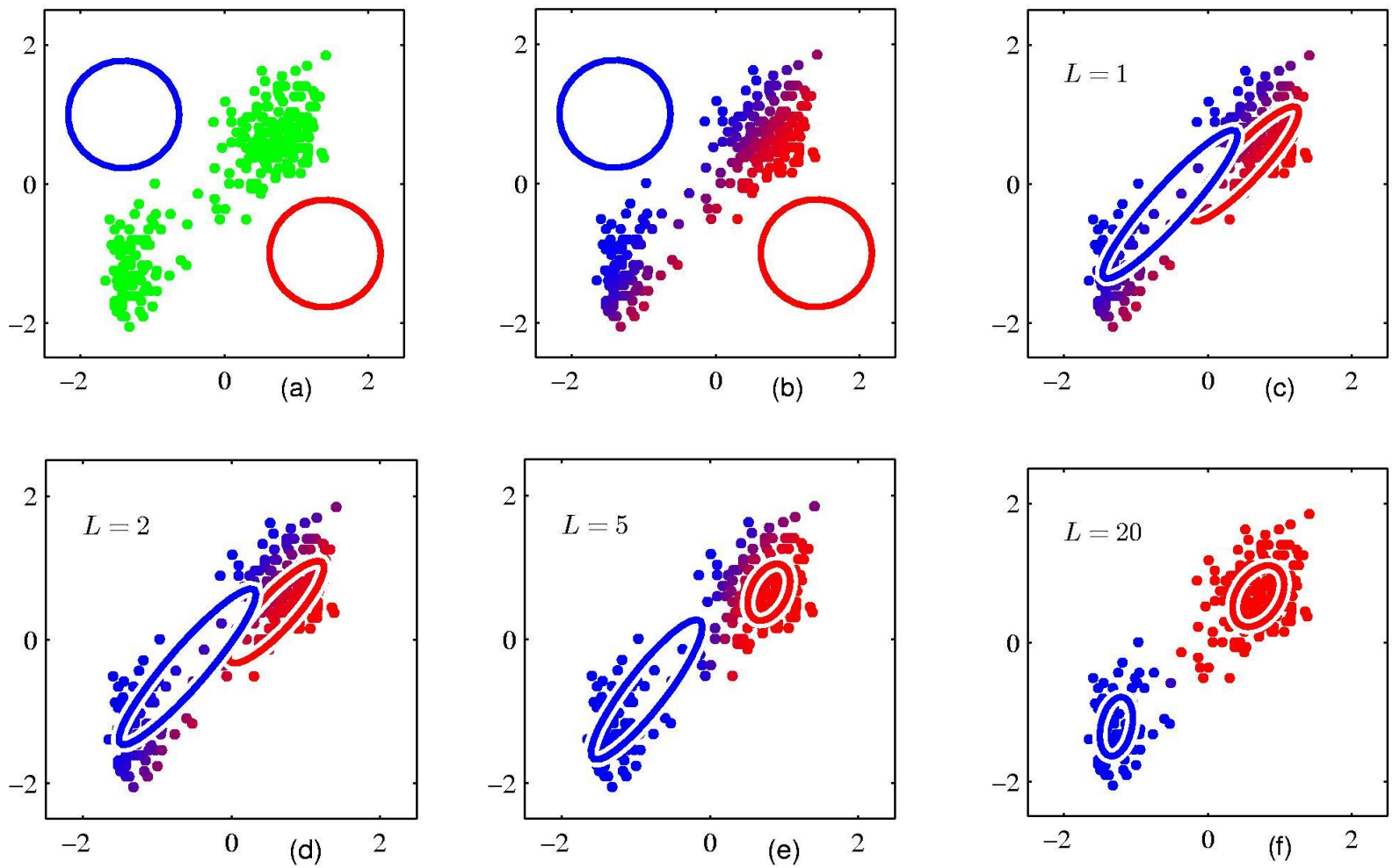
$$\pi_k = \frac{n_k}{n},$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i=1}^{n_i} \mathbf{x}_i^{(k)},$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (\mathbf{x}_i^{(k)} - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i^{(k)} - \hat{\boldsymbol{\mu}}_k)^T$$

上标 $(k)$ 表示属于第$k$个成分的样本，$n_k$表示属于第$k$个成分样本总数。

# 3.8.2 EM for Gaussian mixture model




An example, from (C.M. Bishop, *Pattern Recognition and Machine Learning*, 2006. Figure 9.8)

# 3.8.3 EM：数据缺失情况下的参数估计

- 数据缺失情况下的参数估计
  - Good features, bad (missing) features

$$\mathbf{x}_i = \{\mathbf{x}_{ig}, \mathbf{x}_{ib}\}$$

  - 已知参数值 $\boldsymbol{\theta}^{old}$ 情况下估计新参数值$\boldsymbol{\theta}$
    - 将缺失数据视为隐变量，对其求期望(marginalize)

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) = \sum_i E_{p(\mathbf{x}_{ib}|\mathbf{x}_{ig}, \boldsymbol{\theta}^{old})}\left[\ln\left(p\left(\mathbf{x}_{ig}, \mathbf{x}_{ib} \mid \boldsymbol{\theta}\right)\right)\right]$$

（回顾）：

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) = \sum_i E_{p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{old})}\left[\ln\left(p\left(x_i, z_i \mid \boldsymbol{\theta}\right)\right)\right]$$

# 3.8.3 EM：数据缺失情况下的参数估计

- EM算法
  - Initialize $\boldsymbol{\theta}^{old}$, $t = 0$
  - Do $t \leftarrow t+1$
    - E step: Evaluate $p(\mathbf{x}_{ib}|\mathbf{x}_{ig}, \boldsymbol{\theta}^{old})$, compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{old}\right)=\sum_{i}E_{p(\mathbf{x}_{ib}|\mathbf{x}_{ig},\boldsymbol{\theta}^{old})}\left[\ln\left(p\left(\mathbf{x}_{ig},\mathbf{x}_{ib}\,|\,\boldsymbol{\theta}\right)\right)\right]$$

    - M step: $\boldsymbol{\theta} \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$
  - Until convergence (比如目标函数前后两次差异很小，或前后两次参数变化很小)
  - Return $\boldsymbol{\theta}$

# 3.8.3 EM：数据缺失情况下的参数估计

- 一个例子： EM for a 2D Gaussian

$$D=\{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4\}=\left\{\begin{pmatrix}0\\2\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}2\\2\end{pmatrix},\begin{pmatrix}*\\4\end{pmatrix}\right\}$$

parameters $\boldsymbol{\theta}=(\mu_1,\mu_2,\sigma_1,\sigma_2)^T$ initially $\boldsymbol{\theta}^{old}=(0,0,1,1)^T$

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\mid\boldsymbol{\theta})+E_{p(x_{41}|x_{42}=4,\boldsymbol{\theta}^{old})}\left[\ln p(\mathbf{x}_4\mid\boldsymbol{\theta})\right]$$

$$=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\mid\boldsymbol{\theta})+\int_{-\infty}^{+\infty}\ln p(\mathbf{x}_4\mid\boldsymbol{\theta})p(x_{41}\mid x_{42}=4,\boldsymbol{\theta}^{old})dx_{41}$$

($x_{41}$表示第4个样本的第一维特征)

（接上页）

($x_{41}$表示第4个样本的第一维特征)

$$Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{old}\right)=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})+\int_{-\infty}^{+\infty}\ln p(\mathbf{x}_4\,|\,\boldsymbol{\theta})\;p(x_{41}\,|\,x_{42}=4,\boldsymbol{\theta}^{old})dx_{41}$$

$$=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})+\int_{-\infty}^{+\infty}\ln p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}\middle|\,\boldsymbol{\theta}\right)\frac{p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}\middle|\,\boldsymbol{\theta}^{old}\right)}{\int_{-\infty}^{+\infty}p\left(\begin{pmatrix}x'_{41}\\4\end{pmatrix}\middle|\,\boldsymbol{\theta}^{old}\right)dx'_{41}}dx_{41}$$

$$=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})+\frac{1}{\alpha}\int_{-\infty}^{+\infty}\ln p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}\middle|\,\boldsymbol{\theta}\right)p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}\middle|\,\boldsymbol{\theta}^{old}\right)dx_{41}$$

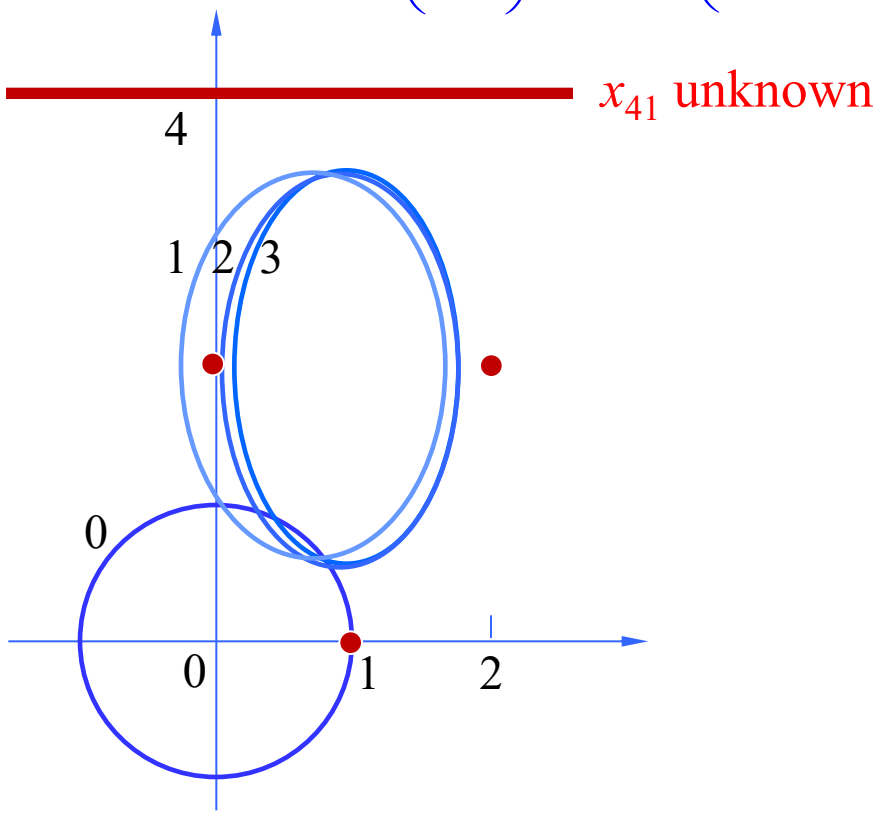
$$=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})+\frac{1}{\alpha}\int_{-\infty}^{+\infty}\ln p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}\middle|\,\boldsymbol{\theta}\right)\frac{1}{2\pi\begin{vmatrix}1&0\\0&1\end{vmatrix}^{1/2}}\exp\left(-\frac{1}{2}\left(x_{41}^2+4^2\right)\right)dx_{41}$$

$$=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\,|\,\boldsymbol{\theta})-\frac{1+\mu_1^2}{2\sigma_1^2}-\frac{(4-\mu_2)^2}{2\sigma_2^2}-\ln\left(2\pi\sigma_1\sigma_2\right)$$

$$\max Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{old}\right)=\sum_{i=1}^{3}\ln p(\mathbf{x}_i\mid\boldsymbol{\theta})-\frac{1+\mu_1^2}{2\sigma_1^2}-\frac{(4-\mu_2)^2}{2\sigma_2^2}-\ln\left(2\pi\sigma_1\sigma_2\right)$$

Then we can get $\boldsymbol{\theta} = (0.75, 2.0, 0.938, 2.0)^T$

After 3 iterations, we obtain $\boldsymbol{\mu}=\begin{pmatrix}1.0\\2.0\end{pmatrix}$, $\boldsymbol{\Sigma}=\begin{pmatrix}0.667 & 0\\0 & 2.0\end{pmatrix}$




If $\mathbf{x}_4=(1,4)^T$, we can obtain $\boldsymbol{\theta} = (1.0, 2.0, 0.5, 2.0)^T$

# 3.8.4 EM算法的优化目标（损失函数）

- 对任意分布$q(\mathbf{z})$：

$$p(\mathbf{x}\mid\boldsymbol{\theta})=\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})}=\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}\frac{q(\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})}$$

$$\ln p(\mathbf{x}\mid\boldsymbol{\theta})=\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}+\ln\frac{q(\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})}$$

等式两边乘以$q(\mathbf{z})$

$$q(\mathbf{z})\ln p(\mathbf{x}\mid\boldsymbol{\theta})=q(\mathbf{z})\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}+q(\mathbf{z})\ln\frac{q(\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})}$$

将取不同$\mathbf{z}$值的等式相加

$$\boxed{\sum_{\mathbf{z}}q(\mathbf{z})\ln p(\mathbf{x}\mid\boldsymbol{\theta})}=\boxed{\sum_{\mathbf{z}}q(\mathbf{z})\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}}+\boxed{\sum_{\mathbf{z}}q(\mathbf{z})\ln\frac{q(\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})}}$$

$$\boxed{\ln p(\mathbf{x}\mid\boldsymbol{\theta})}=\boxed{E_{q(\mathbf{z})}\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}}+\boxed{E_{q(\mathbf{z})}\ln\frac{q(\mathbf{z})}{p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})}}$$

$$=E_{q(\mathbf{z})}\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}+\mathrm{KL}\left(q(\mathbf{z})\,\|\,p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})\right)$$

Kullback-Leibler Divergence

# 3.8.4 EM算法的优化目标

- 对任意分布$q(\mathbf{z})$，存在如下等式:

$$\ln p(\mathbf{x}\mid\boldsymbol{\theta}) = E_{q(\mathbf{z})}\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})} + \mathrm{KL}\left(q(\mathbf{z})\,\|\,p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})\right)$$

✓ $\mathrm{KL}(p\|q)$：称为KL距离或KL散度，衡量相同事件空间中两个概率分布之间的差异。KL距离恒大于零；当且仅当$\forall\mathbf{z}, p(\mathbf{z}) = q(\mathbf{z})$时，有$\mathrm{KL}(p(\mathbf{z})\|q(\mathbf{z}))=0$。

✓ 因为KL距离恒大于零，因此，$E_{q(\mathbf{z})}\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}$ 是$\ln p(\mathbf{x}|\boldsymbol{\theta})$的下界，当且仅当$q(\mathbf{z})=p(\mathbf{z}|\mathbf{x},\boldsymbol{\theta})$时，有$E_{q(\mathbf{z})}\ln\frac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})} = \ln p(\mathbf{x}\mid\boldsymbol{\theta})$

# 3.8.4 EM算法的优化目标

- 令 $L(q,\boldsymbol{\theta}) \equiv E_{q(\mathbf{z})} \ln \dfrac{p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})}{q(\mathbf{z})}$
- E Step: 固定θ，最大化 $\max\limits_{q} L(q,\boldsymbol{\theta})$

$$\because\ L(q,\boldsymbol{\theta}) = \ln p(\mathbf{x}\mid\boldsymbol{\theta}) - \mathrm{KL}\big(q(\mathbf{z}) \,\|\, p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})\big)$$

$$\therefore\ \max_{q} L(q,\boldsymbol{\theta}) \Leftrightarrow \min_{q} \mathrm{KL}\big(q(\mathbf{z}) \,\|\, p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})\big)$$

$$\Rightarrow q(\mathbf{z}) = p(\mathbf{z}\mid\mathbf{x},\boldsymbol{\theta})$$

- M Step: 固定$q(\mathbf{z})$，最大化 $\max\limits_{\boldsymbol{\theta}} L(q,\boldsymbol{\theta})$

$$\max_{\boldsymbol{\theta}} L(q,\boldsymbol{\theta}) \Leftrightarrow \max_{\boldsymbol{\theta}} E_{q(\mathbf{z})} \ln p(\mathbf{x},\mathbf{z}\mid\boldsymbol{\theta})$$

✓ 因此，EM是在通过坐标轮替法最大化$L(q, \theta)$。

✓ $L(q, \theta)$是ln p(**x**| θ)的下界，也被称为证据下界(Evidence Lower BOund, ELBO)，广泛用作VAE, Diffusion model的训练损失函数。

Check (C.M. Bishop, *Pattern Recognition and Machine Learning*, 2006. Chapter 9.4) for more detail.

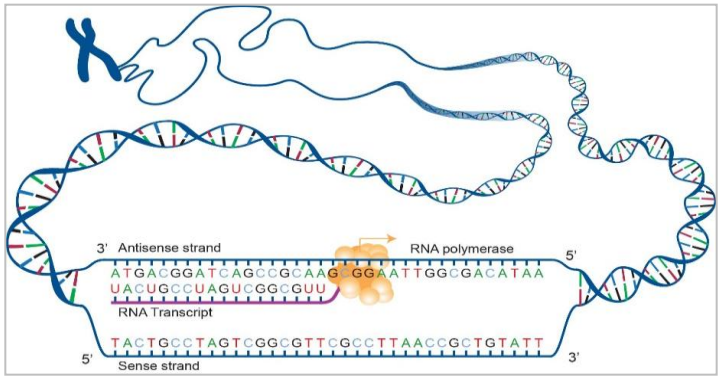
# 3.9 隐马尔可夫模型 (Hidden Markov Model, HMM)

- 时间序列数据

语音、语言

金融数据

视频--动作




DNA序列

# 3.9 隐马尔可夫模型

- 时间序列数据
  - $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, 其中：
    - $n$ 为序列长度
    - $\mathbf{x}_t \in R^d$ 是 $X$ 在第 $t$ 时刻的观测数据
  - 与单模式分类不同，$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$不满足独立假设，观测数据间具有很强的相关性
  - 核心问题：如何对序列数据表示、学习和推理
    - 首先需要引入关于数据分布和时间依赖关系的概率模型，即如何表示：

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$$

# 3.9 隐马尔可夫模型

- 对 $P(X)$ 的假定
  - 方法1：不对数据做任何独立性假设，直接对条件分布 $p(\mathbf{x}_t|\ \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1})$建模（即$\mathbf{x}_t$和它的全部历史相关）

联合分布：$$p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)=p(\mathbf{x}_1)\prod_{t=2}^{n}\boxed{p(\mathbf{x}_t\mid\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{t-1})}$$ （乘法公式）

极强的灵活性、通用性，但模型复杂、数据需求量大

  - 方法2：假设$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$相互独立，只对边缘分布$p(\mathbf{x}_t)$建模：

联合分布：$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n)=\prod_{t=1}^{n}\boxed{p(\mathbf{x}_t)}$$

极差的灵活性、通用性，但模型简单、数据需求量低

如何平衡灵活性和复杂度？

- 对 $P(X)$ 的假定
  - 方法3：
    - 一阶马尔可夫假设：$\mathbf{x}_{t\text{-}1}$ 已知时，$\mathbf{x}_t$ 与 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t\text{-}2}\}$ 独立

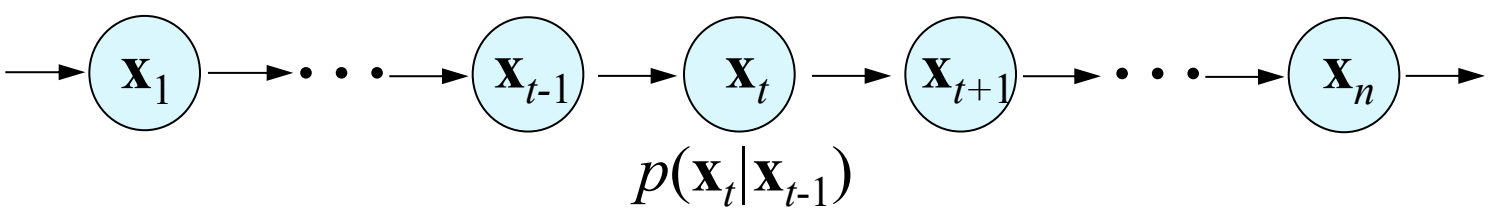
$$\forall t \quad p\left(\mathbf{x}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}\right) = p\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)$$

给定当前信息($\mathbf{x}_{t-1}$)，过去信息($\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-2}\}$)对于预测将来($\mathbf{x}_t$)是无作用的

- 一阶马尔可夫链(first-order Markov chains)

联合分布：

$$p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) = p\left(\mathbf{x}_1\right) \prod_{t=2}^{n} p\left(\mathbf{x}_t \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}\right) = p(\mathbf{x}_1) \prod_{t=2}^{n} p\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)$$




- 因此，只需对$p(\mathbf{x}_t \mid \mathbf{x}_{t\text{-}1})$ 进行建模。这就是构建隐马模型的出发点。

# 3.9.1 马尔可夫链

- 时齐、离散的一阶马尔可夫链 (first-order Markov chain)
  - 一阶马氏链的联合分布

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n)=p(\mathbf{x}_1)\prod_{t=2}^{n}p(\mathbf{x}_t\mid\mathbf{x}_{t-1})$$

  - 离散马氏链： $\mathbf{x}_t\in\{1,2,\ldots,K\}$，$K$为状态数
  - 时齐(time-homogenous)马氏链：转移概率$p(\mathbf{x}_t\mid\mathbf{x}_{t-1})$只与状态有关，与时间 $t$ 无关
  - 初始状态分布：$P(\mathbf{x}_1)=\pi\in R^K$
  - 状态转移概率: $P(\mathbf{x}_t\mid\mathbf{x}_{t-1})=\mathbf{A}\in R^{K\times K}$

$$P(\mathbf{x}_t=j\mid\mathbf{x}_{t-1}=i)=A_{ij}$$

从状态$i$ 转移到状态$j$ 的概率

非负，行和等于1

$$\mathbf{A}=\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1K} \\ A_{21} & A_{22} & \cdots & A_{1K} \\ \cdots & \cdots & \cdots & \cdots \\ A_{K1} & A_{K2} & \cdots & A_{KK} \end{pmatrix}$$

状态转移矩阵

- 一个例子

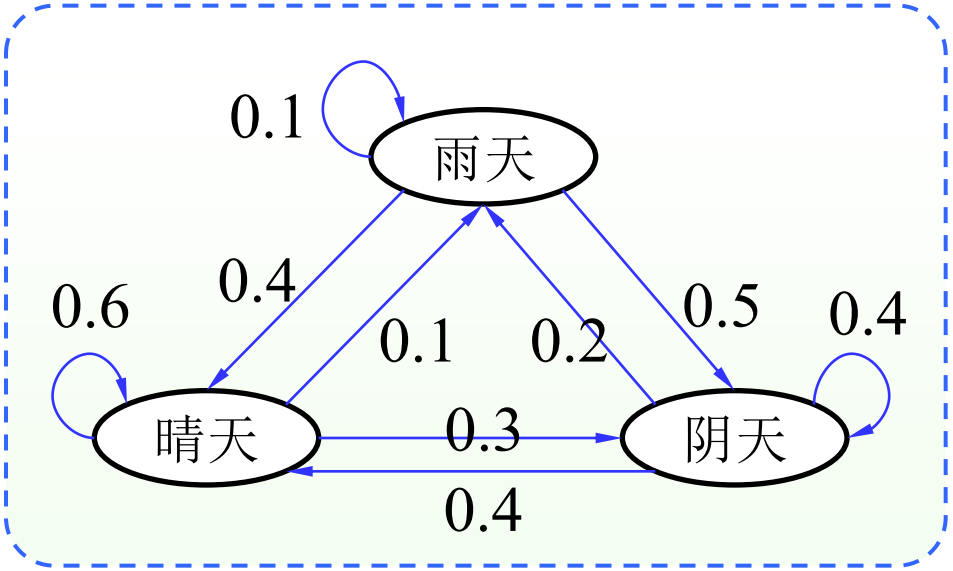
$x_t \in \{$“雨天”，“晴天”，“阴天”$\}, K=3$

初始状态分布：$\pi = [0.1, 0.6, 0.3]$

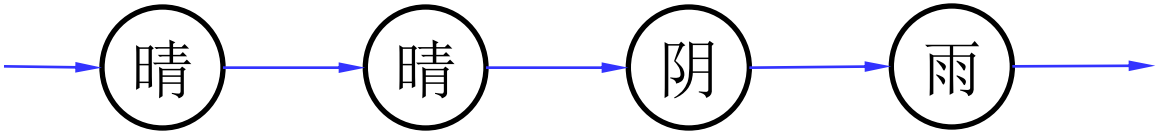
状态转移概率 $\boldsymbol{A}$

| | | 明天 | | |
|---|---|---|---|---|
| | | 雨 | 晴 | 阴 |
| 今天 | 雨 | 0.1 | 0.4 | 0.5 |
| | 晴 | 0.1 | 0.6 | 0.3 |
| | 阴 | 0.2 | 0.4 | 0.4 |

状态转移图




采样（序列生成）：



- 一个例子

$x_t \in$ {“雨天”, “晴天”, “阴天”}, $K$=3

初始状态分布：$\pi = [0.1, 0.6, 0.3]$

状态转移概率：$A = \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.1 & 0.6 & 0.3 \\ 0.2 & 0.4 & 0.4 \end{bmatrix}$

已知第t天是雨天，第t+2天是晴天的概率？

$$p(x_t) = [1, 0, 0]^T$$

$$p(x_{t+1}) = A^T p(x_t) = [0.1, 0.4, 0.5]^T$$

$$p(x_{t+2}) = A^T p(x_{t+1}) = [0.15, 0.48, 0.37]^T$$

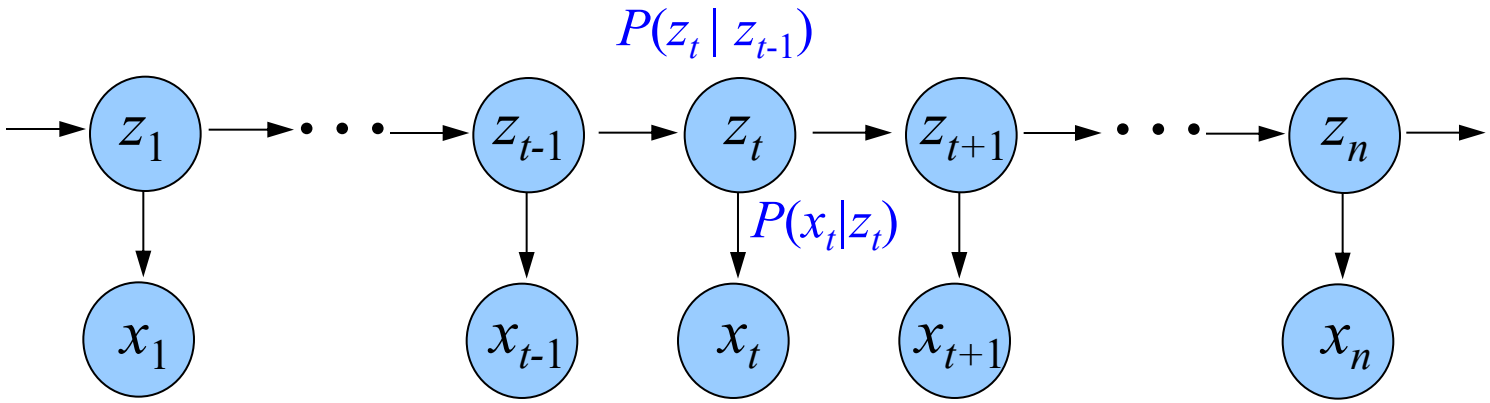
全概率公式

# 3.9.2隐马尔可夫模型简介

- Hidden Markov Model(HMM)的基本思想
  - 观测序列由一个不可见的马尔可夫链生成
  - HMM的随机变量可分为两组：
    - 状态变量$\{z_1, z_2, \ldots, z_n\}$：构成一阶、离散、静态马尔可夫链。用于描述系统内部的状态变化，通常是隐藏的，不可被观测的。其中，$z_t$表示第 $t$ 时刻系统的状态。
    - 观测变量$\{x_1, x_2, \ldots, x_n\}$：其中，$x_t$表示第 $t$ 时刻的观测变量，通过条件概率$p(x_t \mid z_t)$由状态变量$z_t$生成；根据具体问题，$x_t$可以是离散或连续，一维或多维。
  - 主要用于时序数据建模，在CV、NLP、语音识别中有诸多应用

# 3.9.2 HMM简介

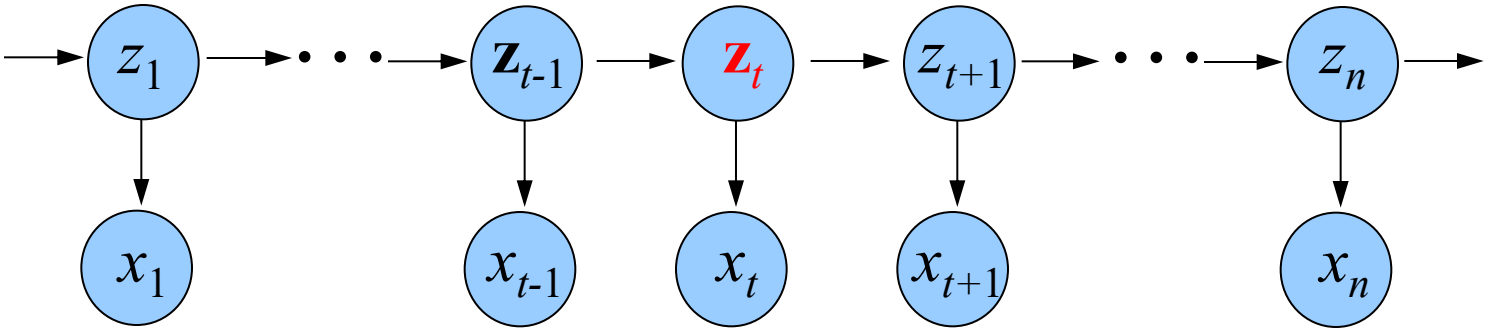
- HMM的图结构




上图中，箭头表示依赖关系。通过图中的路径以及连通性，可以推导出变量间的（条件）独立性。深入内容可以学习概率图模型/因果推断中相关知识。

# 3.9.2 HMM简介

- HMM中的条件独立性




**状态独立性：**

$$p(\mathbf{z}_t \mid \mathbf{z}_1, \mathbf{x}_1, \mathbf{z}_2, \mathbf{x}_2, \cdots, \mathbf{z}_{t\text{-}1}, \mathbf{x}_{t\text{-}1}) = p(\mathbf{z}_t \mid \mathbf{z}_{t\text{-}1})$$

当$\mathbf{z}_{t\text{-}1}$已知， $\mathbf{z}_t$与前 $t$-2个状态和观测变量独立，即$\{\mathbf{z}_t\}$构成马尔可夫链。

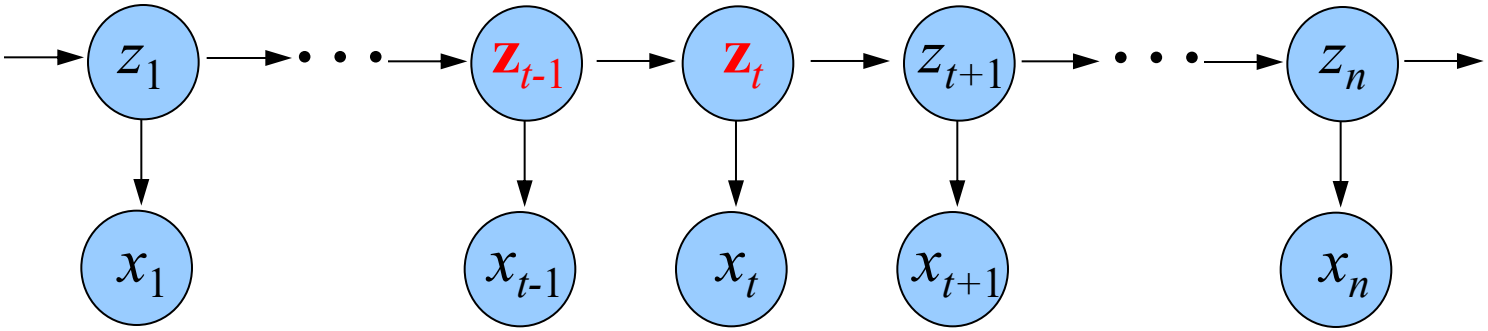
**观测独立性：**

$$p(\mathbf{x}_t \mid \mathbf{z}_1, \mathbf{x}_1, \mathbf{z}_2, \mathbf{x}_2, \cdots, \mathbf{z}_t, \cdots, \mathbf{z}_n, \mathbf{x}_n) = p(\mathbf{x}_t \mid \mathbf{z}_t)$$

当$\mathbf{z}_t$已知，$\mathbf{x}_t$与其它所有变量独立，不依赖于任何状态和观测。

# 3.9.2 HMM简介

- HMM中的条件独立性




$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n \mid \mathbf{z}_t) = p(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1} \mid \mathbf{z}_t)\, p(\mathbf{x}_t \mid \mathbf{z}_t)\, p(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{z}_t)$$

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n \mid \mathbf{z}_{t-1},\mathbf{z}_t) = p(\mathbf{x}_1,\ldots,\mathbf{x}_{t-2} \mid \mathbf{z}_{t-1})\, p(\mathbf{x}_{t-1} \mid \mathbf{z}_{t-1})\, p(\mathbf{x}_t \mid \mathbf{z}_t)\, p(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{z}_t)$$

---

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1} \mid \mathbf{x}_t,\mathbf{z}_t) = p(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1} \mid \mathbf{z}_t)$$

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1} \mid \mathbf{z}_{t-1},\mathbf{z}_t) = p(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1} \mid \mathbf{z}_{t-1})$$

---

$$p(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{x}_t,\mathbf{z}_t) = p(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{z}_t)$$

$$p(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{z}_t,\mathbf{z}_{t+1}) = p(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{z}_{t+1})$$

- HMM的联合分布

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n)=p(\mathbf{z}_1)\prod_{t=2}^{n}p(\mathbf{z}_t\mid\mathbf{z}_{t-1})\prod_{t=1}^{n}p(\mathbf{x}_t\mid\mathbf{z}_t)$$

初始状态概率　状态转移概率　发射概率

- HMM基本要素，三组参数 θ=(π , A, B)：
  - 初始状态概率向量 $\pi\in R^K$：

$$\pi_k=P(z_1=k),\quad 1\le k\le K$$

  - 状态转移概率矩阵 $\mathbf{A}\in R^{K\times K}$：

$$A_{i,j}=P(z_t=j\mid z_{t-1}=i),\quad 1\le i,j\le K$$

  - 发射概率矩阵 $\mathbf{B}\in R^{K\times M}$：　为简洁起见, 考虑离散的观测变量

$$B_{i,j}=P(x_t=j\mid z_t=i),\quad 1\le i\le K,1\le j\le M$$

✓ 发射概率的选取与具体问题相关。常见包括：高斯、混合高斯、多项分布等。HMM的学习与推理算法与发射概率形式无关。

- 一个例子

$z_t \in$ {“雨”，“晴”，“阴”}

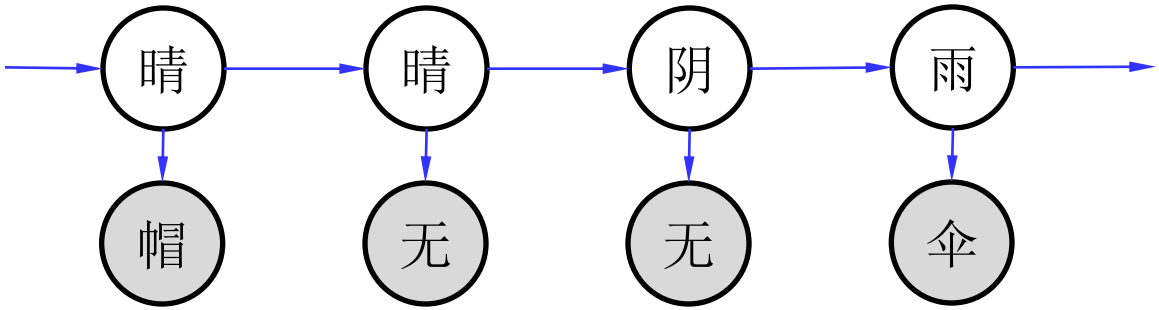
$x_t \in$ {“打伞”，“戴帽”，“无伞无帽”}

明天

| 今天 | 雨 | 晴 | 阴 |
|---|---|---|---|
| 雨 | 0.1 | 0.4 | 0.5 |
| 晴 | 0.1 | 0.6 | 0.3 |
| 阴 | 0.2 | 0.4 | 0.4 |

| | 打伞 | 戴帽 | 无伞无帽 |
|---|---|---|---|
| 雨 | 0.8 | 0.1 | 0.1 |
| 晴 | 0.1 | 0.4 | 0.5 |
| 阴 | 0 | 0.2 | 0.8 |

采样：




隐状态序列（不可观测）

观测序列

✓ 已知观测序列，如何估计状态转移概率和发射概率？**HMM的学习问题**

✓ 已知观测序列，如何推断天气？**HMM的解码问题**

- 三个基本问题
  - 给定模型[$\mathbf{A},\mathbf{B},\boldsymbol{\pi}$]，如何有效计算其产生观测序列$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$的概率$P(\mathbf{x}|\mathbf{A},\mathbf{B},\boldsymbol{\pi})$? 即评估模型与观测数据的匹配程度。
    - 许多任务需要根据以往的观测序列来预测当前时刻最有可能的观测值。
  - 给定模型[$\mathbf{A},\mathbf{B},\boldsymbol{\pi}$]和观测序列$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$，如何找到与此观测序列相匹配的状态序列$\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2,\ldots, \mathbf{z}_n\}$？即根据观测序列推断出隐藏的模型状态。(解码问题）
    - 在语言识别中，观测值为语音信号，隐藏状态为对应的文本内容，目标就是观测信号推断最有可能的状态。
  - 给定观测序列$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$，如何调整模型参数[$\mathbf{A},\mathbf{B},\boldsymbol{\pi}$]使该序列出现的概率$P(\mathbf{x}\ |\mathbf{A},\mathbf{B},\boldsymbol{\pi})$最大？即如何模型使其能够最好地描述观测数据。(参数估计—学习问题）
    - 在大多数实际应用中，人工指定参数已变得不可行，需要根据训练样本学习最优模型。

# 3.9.3 HMM参数学习

- 参数学习的任务
  - 通过拟合观测序列，确定HMM中的参数，即 $\theta=(\pi, \mathbf{A}, \mathbf{B})$
- EM算法步骤
  - **E Step:** 对给定的 $\theta$ , 估计：

$$q(\mathbf{z}_1,\ldots,\mathbf{z}_n) \equiv p_{\boldsymbol{\theta}}\left(\mathbf{z}_1,\ldots,\mathbf{z}_n \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)$$

  - **M Step：** 用估计出的$q(\mathbf{z}_1, \ldots, \mathbf{z}_n)$，更新 $\theta$：

$$\theta = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{z}} q(\mathbf{z}_1,\ldots,\mathbf{z}_n) \ln p_{\boldsymbol{\theta}}(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n)$$

  - E步和M步迭代运行，直至收敛

M step：更新 θ

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n)=p(\mathbf{z}_1)\prod_{t=2}^{n}p(\mathbf{z}_t\mid\mathbf{z}_{t-1})\prod_{t=1}^{n}p(\mathbf{x}_t\mid\mathbf{z}_t)$$

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})=\sum_{\mathbf{z}}q(\mathbf{z}_1,\ldots,\mathbf{z}_n)\ln p_{\boldsymbol{\theta}}(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n)$$

$$=\sum_{\mathbf{z}}q(\mathbf{z}_1,\ldots,\mathbf{z}_n)\left(\ln p_{\boldsymbol{\theta}}(\mathbf{z}_1)+\sum_{t=2}^{n}\ln p_{\boldsymbol{\theta}}(\mathbf{z}_t\mid\mathbf{z}_{t-1})+\sum_{t=1}^{n}\ln p_{\boldsymbol{\theta}}(\mathbf{x}_t\mid\mathbf{z}_t)\right)$$

$$=\sum_{\mathbf{z}}q(\mathbf{z}_1,\ldots,\mathbf{z}_n)\ln p_{\boldsymbol{\theta}}(\mathbf{z}_1)+\sum_{\mathbf{z}}q(\mathbf{z}_1,\ldots,\mathbf{z}_n)\sum_{t=2}^{n}\ln p_{\boldsymbol{\theta}}(\mathbf{z}_t\mid\mathbf{z}_{t-1})$$

$$+\sum_{\mathbf{z}}q(\mathbf{z}_1,\ldots,\mathbf{z}_n)\sum_{t=1}^{n}\ln p_{\boldsymbol{\theta}}(\mathbf{x}_t\mid\mathbf{z}_t)$$

$$=\sum_{z_1=1}^{K}q(\mathbf{z}_1)\ln p_{\boldsymbol{\theta}}(\mathbf{z}_1)+\sum_{t=2}^{n}\sum_{\mathbf{z}_{t-1},\mathbf{z}_t=1}^{K}q(\mathbf{z}_{t-1},\mathbf{z}_t)\ln p_{\boldsymbol{\theta}}(\mathbf{z}_t\mid\mathbf{z}_{t-1})$$

$$+\sum_{t=1}^{n}\sum_{\mathbf{z}_t=1}^{K}q(\mathbf{z}_t)\ln p_{\boldsymbol{\theta}}(\mathbf{x}_t\mid\mathbf{z}_t)$$

（接上页）

M step：更新 θ

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n)=p(\mathbf{z}_1)\prod_{t=2}^{n}p(\mathbf{z}_t\mid\mathbf{z}_{t-1})\prod_{t=1}^{n}p(\mathbf{x}_t\mid\mathbf{z}_t)$$

$$\sum_{\mathbf{z}}q(\mathbf{z}_1,\ldots,\mathbf{z}_n)\ln p_{\boldsymbol{\theta}}(\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{z}_1,\ldots,\mathbf{z}_n)$$

$$=\sum_{z_1=1}^{K}q(\mathbf{z}_1)\ln p_{\boldsymbol{\theta}}(\mathbf{z}_1)+\sum_{t=2}^{n}\sum_{\mathbf{z}_{t-1},\mathbf{z}_t=1}^{K}q(\mathbf{z}_{t-1},\mathbf{z}_t)\ln p_{\boldsymbol{\theta}}(\mathbf{z}_t\mid\mathbf{z}_{t-1})$$

$$+\sum_{t=1}^{n}\sum_{\mathbf{z}_t=1}^{K}q(\mathbf{z}_t)\ln p_{\boldsymbol{\theta}}(\mathbf{x}_t\mid\mathbf{z}_t)$$

$$=\boxed{\sum_{z_1=1}^{K}q(\mathbf{z}_1)\ln\pi_{\mathbf{z}_1}}+\boxed{\sum_{t=2}^{n}\sum_{\mathbf{z}_{t-1},\mathbf{z}_t=1}^{K}q(\mathbf{z}_{t-1},\mathbf{z}_t)\ln A_{\mathbf{z}_{t-1},\mathbf{z}_t}}$$

只含$A$

只含π

$$\boxed{+\sum_{t=1}^{n}\sum_{\mathbf{z}_t=1}^{K}q(\mathbf{z}_t)\ln B_{\mathbf{z}_t,\mathbf{x}_t}}$$

只含$B$

M step：更新 θ

用拉格朗日乘子法优化以下问题，可得：

$$\boldsymbol{\pi}: \quad \arg\max_{\pi} \sum_{\mathbf{z}_1=1}^{K} q(\mathbf{z}_1) \ln \pi_{\mathbf{z}_1}, \quad \text{s.t.} \ \sum_{k=1}^{K} \pi_k = 1$$

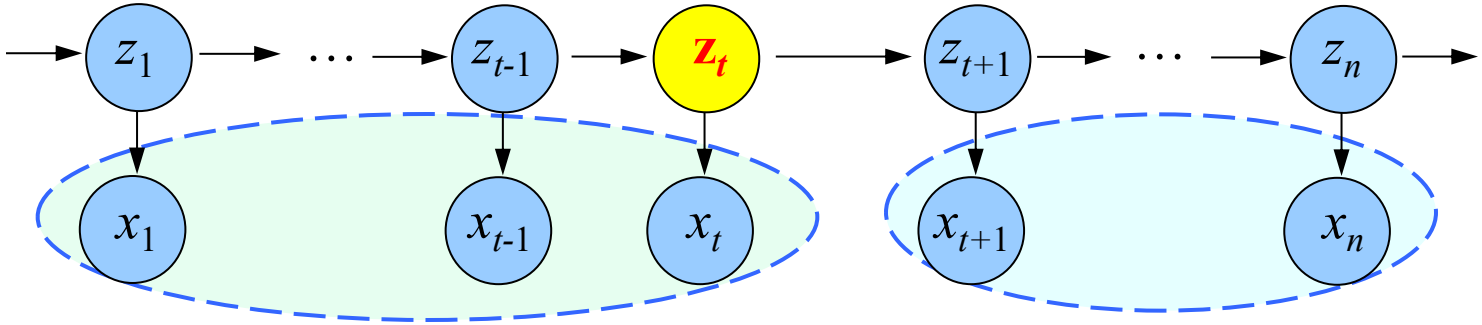
$$\Rightarrow \pi_k = q(\mathbf{z}_1 = k)$$

$$\mathbf{A}: \quad \arg\max_{\mathbf{A}} \sum_{t=2}^{n} \sum_{\mathbf{z}_{t-1},\mathbf{z}_t=1}^{K} q(\mathbf{z}_{t-1},\mathbf{z}_t) \ln A_{\mathbf{z}_{t-1},\mathbf{z}_t}, \quad \text{s.t.} \ \forall i \ \sum_{j=1}^{K} A_{i,j} = 1$$

$$\Rightarrow \mathbf{A}_{i,j} = \frac{\sum_{t=2}^{n} q(\mathbf{z}_{t-1} = i, \mathbf{z}_t = j)}{\sum_{t=2}^{n} \sum_{k=1}^{K} q(\mathbf{z}_{t-1} = i, \mathbf{z}_t = k)}$$

$$\mathbf{B}: \quad \arg\max_{B} \sum_{t=1}^{n} \sum_{\mathbf{z}_t=1}^{K} q(\mathbf{z}_t) \ln B_{\mathbf{z}_t,\mathrm{x}_t}, \quad \text{s.t.} \ \forall i \ \sum_{j=1}^{M} B_{i,j} = 1$$

$$\Rightarrow \mathbf{B}_{i,j} = \frac{\sum_{t=1}^{n} I\{x_t == j\} q(\mathbf{z}_t = i)}{\sum_{t=1}^{n} q(\mathbf{z}_t = i)}$$




E Step: 对给定的 θ，估计 $q(\mathbf{z}_1,\ldots,\mathbf{z}_n) \equiv p_{\boldsymbol{\theta}}\left(\mathbf{z}_1,\ldots,\mathbf{z}_n \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)$

根据上页推导，只需估计:

$$q(\mathbf{z}_t) \equiv p_{\boldsymbol{\theta}}\left(\mathbf{z}_t \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right), \qquad q(\mathbf{z}_{t-1},\mathbf{z}_t) \equiv p_{\boldsymbol{\theta}}\left(\mathbf{z}_{t-1},\mathbf{z}_t \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)$$

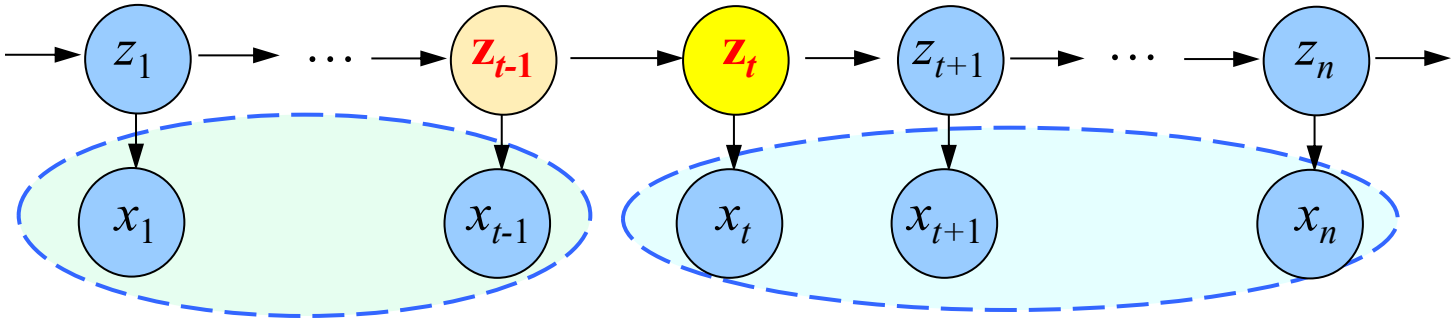
以下省略 θ：

$$q(\mathbf{z}_t) = \frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{z}_t\right)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)} = \frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n \mid \mathbf{z}_t\right)p(\mathbf{z}_t)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

HMM的条件独立性

$$= \frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t \mid \mathbf{z}_t\right)p\left(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{z}_t\right)p(\mathbf{z}_t)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

$$= \frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t,\mathbf{z}_t\right)p\left(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n \mid \mathbf{z}_t\right)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$




以下省略 θ：

$$q(\mathbf{z}_{t-1},\mathbf{z}_t)=p(\mathbf{z}_{t-1},\mathbf{z}_t\mid\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)=\frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{z}_{t-1},\mathbf{z}_t\right)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

$$=\frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\mathbf{z}_t\mid\mathbf{z}_{t-1}\right)p(\mathbf{z}_{t-1})}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

$$=\frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{t-1}\mid\mathbf{z}_{t-1}\right)p\left(\mathbf{x}_t,\ldots,\mathbf{x}_n,\mathbf{z}_t\mid\mathbf{z}_{t-1}\right)p(\mathbf{z}_{t-1})}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

HMM的条件独立性

$$=\frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{t-1},\mathbf{z}_{t-1}\right)p\left(\mathbf{x}_t,\ldots,\mathbf{x}_n,\mathbf{z}_t\mid\mathbf{z}_{t-1}\right)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

$$=\frac{p\left(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{z}_{t-1}\right)p(\mathbf{z}_t\mid\mathbf{z}_{t-1})p(\mathbf{x}_t\mid\mathbf{z}_t)p\left(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n\mid\mathbf{z}_t\right)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

# 前向-后向算法 (forward-backward algorithm)

$$q(\mathbf{z}_t)=\frac{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t,\mathbf{z}_t\right)p\left(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n\mid\mathbf{z}_t\right)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

$$q(\mathbf{z}_{t-1},\mathbf{z}_t)=\frac{p\left(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{z}_{t-1}\right)p(\mathbf{z}_t\mid\mathbf{z}_{t-1})p(\mathbf{x}_t\mid\mathbf{z}_t)p\left(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n\mid\mathbf{z}_t\right)}{p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)}$$

前向概率：$\alpha_t(\mathbf{z}_t)\equiv p\left(\mathbf{x}_1,\ldots,\mathbf{x}_t,\mathbf{z}_t\right)$

后向概率：$\beta_t(\mathbf{z}_t)\equiv p\left(\mathbf{x}_{t+1},\ldots,\mathbf{x}_n\mid\mathbf{z}_t\right)$

递归求解

观测概率：$p\left(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\right)=\sum_{\mathbf{z}_n=1}^{K}\alpha_n(\mathbf{z}_n)$ HMM的第一个基本问题

$$q(\mathbf{z}_t)=\frac{\alpha(\mathbf{z}_t)\beta(\mathbf{z}_t)}{\sum_{\mathbf{z}_n=1}^{K}\alpha_n(\mathbf{z}_n)}$$

$$q(\mathbf{z}_{t-1},\mathbf{z}_t)=\frac{\alpha_{t-1}\left(\mathbf{z}_{t-1}\right)p(\mathbf{z}_t\mid\mathbf{z}_{t-1})p(\mathbf{x}_t\mid\mathbf{z}_t)\beta_t\left(\mathbf{z}_t\right)}{\sum_{\mathbf{z}_n=1}^{K}\alpha_n(\mathbf{z}_n)}$$

## 从前向后计算，$t=1, \ldots, n$

$$\alpha_t(\mathbf{z}_t) \equiv p(\mathbf{x}_1,\ldots,\mathbf{x}_t,\mathbf{z}_t) = \sum_{\mathbf{z}_{t-1}} p(\mathbf{x}_1,\ldots,\mathbf{x}_t,\mathbf{z}_{t-1},\mathbf{z}_t)$$

$$= \sum_{\mathbf{z}_{t-1}} p(\mathbf{x}_1,\ldots,\mathbf{x}_t,\mathbf{z}_t \mid \mathbf{z}_{t-1})\, p(\mathbf{z}_{t-1})$$

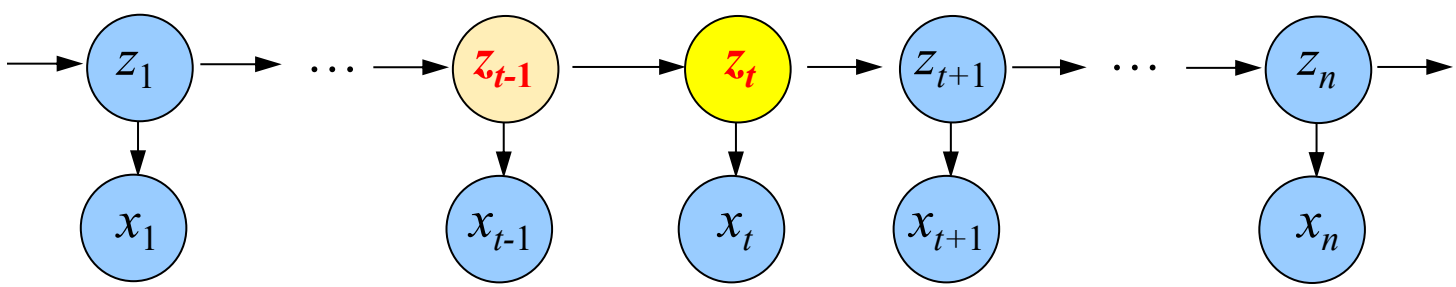
$$= \sum_{\mathbf{z}_{t-1}} p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{t-1} \mid \mathbf{z}_{t-1})\, p(\mathbf{x}_t,\mathbf{z}_t \mid \mathbf{z}_{t-1})\, p(\mathbf{z}_{t-1})$$

HMM的
条件独立性

$$= \sum_{\mathbf{z}_{t-1}} p(\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{z}_{t-1})\, p(\mathbf{x}_t,\mathbf{z}_t \mid \mathbf{z}_{t-1})$$

$$= \sum_{\mathbf{z}_{t-1}} \alpha_{t-1}(\mathbf{z}_{t-1})\, p(\mathbf{z}_t \mid \mathbf{z}_{t-1})\, p(\mathbf{x}_t \mid \mathbf{z}_t)$$

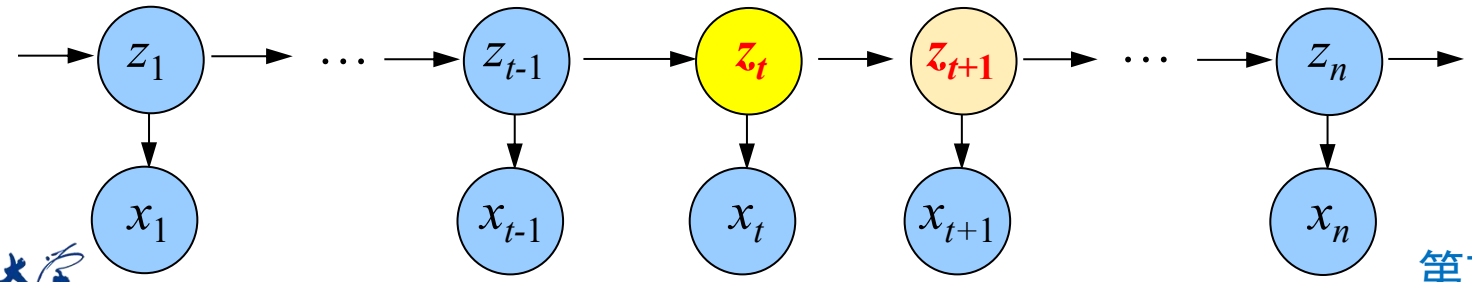
$$= p(\mathbf{x}_t \mid \mathbf{z}_t) \sum_{\mathbf{z}_{t-1}} \alpha_{t-1}(\mathbf{z}_{t-1})\, p(\mathbf{z}_t \mid \mathbf{z}_{t-1})$$



## 从后向前计算，$t = n, n\text{-}1, \ldots, 1$

$$
\begin{aligned}
\beta_t(\mathbf{z}_t) &\equiv p\left(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n \mid \mathbf{z}_t\right) \\
&= \sum_{\mathbf{z}_{t+1}} p\left(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n, \mathbf{z}_{t+1} \mid \mathbf{z}_t\right) \\
&= \sum_{\mathbf{z}_{t+1}} p\left(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n \mid \mathbf{z}_t, \mathbf{z}_{t+1}\right) p\left(\mathbf{z}_{t+1} \mid \mathbf{z}_t\right) \\
&= \sum_{\mathbf{z}_{t+1}} p\left(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_n \mid \mathbf{z}_{t+1}\right) p\left(\mathbf{z}_{t+1} \mid \mathbf{z}_t\right) \\
&= \sum_{\mathbf{z}_{t+1}} p\left(\mathbf{x}_{t+2}, \ldots, \mathbf{x}_n \mid \mathbf{z}_{t+1}\right) p\left(\mathbf{x}_{t+1} \mid \mathbf{z}_{t+1}\right) p\left(\mathbf{z}_{t+1} \mid \mathbf{z}_t\right) \\
&= \sum_{\mathbf{z}_{t+1}} \beta_{t+1}(\mathbf{z}_{t+1}) p\left(\mathbf{x}_{t+1} \mid \mathbf{z}_{t+1}\right) p\left(\mathbf{z}_{t+1} \mid \mathbf{z}_t\right)
\end{aligned}
$$

HMM的
条件独立性



# 3.9.4 HMM的解码

- HMM的解码问题(复合模式分类)
  - 在实际问题中，状态变量通常有明确的含义。如语音识别中，$z_t$表示语音信号$x_t$对应的文本。因此，经常需要根据观测序列推断状态序列。
  - 对给定的HMM模型θ=(π , A, B)和观测序列$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$，求解：

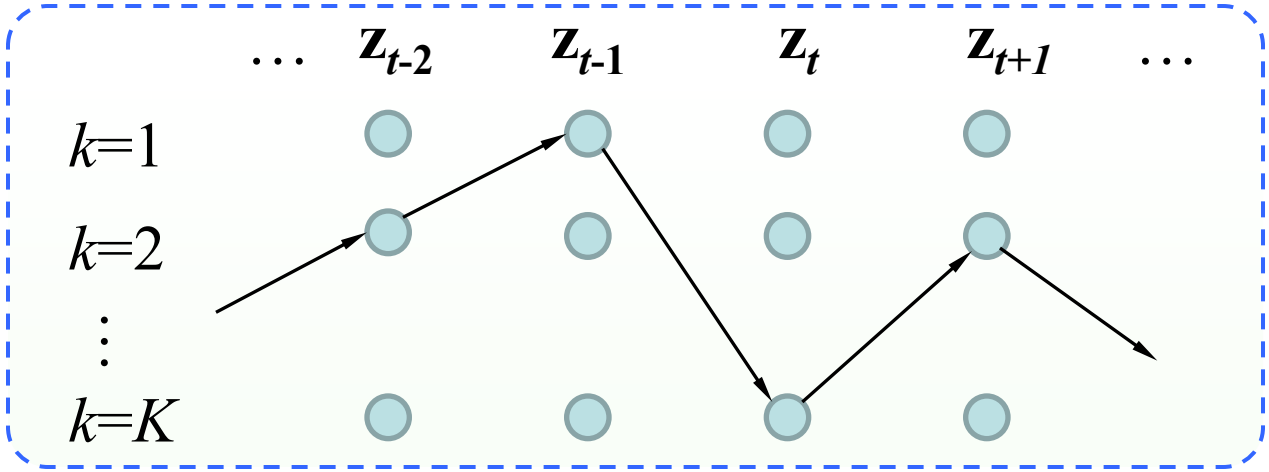
$$\mathbf{z}^* = \arg\max_{\mathbf{z}} p_\theta\left(\mathbf{z}_1, \ldots, \mathbf{z}_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right)$$

    - $z^*$是最大后验概率对应的状态序列，也称为最优状态路径
    - 这对应单模式分类问题中的最大后验概率决策，$z_t$对应$x_t$的类别；但与单模式分类中对$x_t$独立解码不同，HMM需要联合解码
    - $\underset{\mathbf{z}}{\operatorname{argmax}}\, p_\theta(z_1, \ldots, z_n | x_1, \ldots, x_n) = \underset{\mathbf{z}}{\operatorname{argmax}}\, p_\theta(z_1, \ldots, z_n, x_1, \ldots, x_n)$

$$= \underset{\mathbf{z}}{\operatorname{argmax}} \ln p_\theta(z_1, \ldots, z_n, x_1, \ldots, x_n)$$

# 3.9.4 HMM的解码

- 状态路径：$\mathbf{z}_1,\ldots,\mathbf{z}_n$




- 对于给定的HMM模型和观测序列$\{\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n\}$，不同状态路径对应不同的后验概率：$p_\theta\left(\mathbf{z}_1,\ldots,\mathbf{z}_n \mid \mathbf{x}_1,\ldots,\mathbf{x}_n\right)$
- 共有$K^n$条可能的状态路径，对应$K^n$个概率值
- 直接计算这些概率，然后选出$z^*$的复杂度为$O(K^n)$

# 3.9.4 HMM的解码

- HMM的解码算法：维特比算法 (Viterbi, 1967)
  - 最优子问题：寻找以状态$z_t$结束的前$t$步最优状态路径

$$w_t(\mathbf{z}_t) \equiv \max_{\mathbf{z}_1,\ldots,\mathbf{z}_{t-1}} \ln p_\theta(\mathbf{x}_1,\ldots,\mathbf{x}_t,\mathbf{z}_1,\ldots,\mathbf{z}_{t-1},\mathbf{z}_t) \in R^K$$

  - 动态规划算法

For $\mathbf{z}_1 = 1, \ldots, K$: $\quad w_1(\mathbf{z}_1) = \ln p(\mathbf{z}_1) + \ln p(\mathbf{x}_1|\mathbf{z}_1)$

For $t = 2, \ldots, n$:

For $\mathbf{z}_t = 1, \ldots, K$:

$$w_t(\mathbf{z}_t) = \ln p(\mathbf{x}_t \mid \mathbf{z}_t) + \max_{\mathbf{z}_{t-1}\in\{1,\ldots,K\}} \left\{ w_{t-1}(\mathbf{z}_{t-1}) + \ln p(\mathbf{z}_t \mid \mathbf{z}_{t-1}) \right\}$$

  - 计算复杂度：$O(nK^2)$

# 下次课内容

- 第4章 非参数法
  - 密度估计
  - Parzen窗方法
  - K近邻估计
  - 最近邻规则
  - 距离度量

# 致谢

- PPT由向世明老师提供

# Thank All of You!

## (Questions?)

张燕明

ymzhang@nlpr.ia.ac.cn

people.ucas.ac.cn/~ymzhang

模式分析与学习课题组(PAL)

中科院自动化研究所·多模态人工智能系统实验室