机器学习
Machine Learning

北京航空航天大学计算机学院智能识别与图像处理实验室
IRIP Lab, School of Computer Science and Engineering, Beihang University
黄 迪　刘庆杰

2018年秋季学期
Fall 2018

课前回顾

基本采样法（Basic Sampling）

- **思想**：从基本概率分布中产生新变量的分布

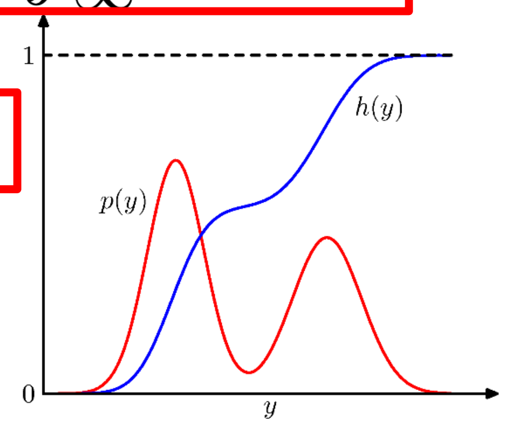
➢ **均匀分布（Uniform distribution）**：$p(z)=1 \quad z\in(0,1)$

➢ **产生非均匀分布**：$p(y), \quad y=f(z)$

$$\left.\begin{aligned} p(y) &= p(z)\left|\frac{dz}{dy}\right| \\ p(z) &= 1 \end{aligned}\right\} \Rightarrow z = \boxed{h(y) \equiv \int_{\infty}^{y} p(\hat{y})d\hat{y}}$$

累积分布函数（CDF）

$$\Rightarrow \quad y = h^{-1}(z)$$



基本采样法（Basic Sampling）

- **练习：指数分布** $p(y) = \lambda \exp(-\lambda y) \quad y \in [0, \infty)$

求 $y = f(z)$

$$h(y) = \int_{-\infty}^{y} p(\hat{y}) d\hat{(y)} = 1 - \exp(\lambda y)$$

$$y = h^{-1}(z) = \lambda^{-1} \ln(1 - z)$$

- **多变量分布形式：**

$$p(y_1, \ldots, y_M) = p(z_1, \ldots, z_M) \left| \frac{\partial(z_1, \ldots, z_M)}{\partial(y_1, \ldots, y_M)} \right|$$

基本采样法（Basic Sampling）

- **高斯分布**：

$$z_1, z_2 \sim \text{uniform}(-1, 1)$$

$$p(z_1, z_2) = \frac{1}{\pi}, \quad (z_1 + z_2)^2 \leq 1$$

$$\boxed{\begin{aligned} y_1 &= z_1 \left(\frac{-2 \ln z_1}{r^2} \right)^{1/2} \\ y_2 &= z_2 \left(\frac{-2 \ln z_2}{r^2} \right)^{1/2} \end{aligned}} \qquad r^2 = (z_1 + z_2)^2$$

Box-Muller变换

$$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right|$$

$$= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right]$$

基本采样法（Basic Sampling）

- 高斯分布：

$$\boxed{\begin{aligned} y_1 &= \sqrt{-2\ln z_1}\cos(2\pi z_2) \\ y_2 &= \sqrt{-2\ln z_1}\sin(2\pi z_2) \end{aligned}}$$

$$z_1, z_2 \sim \mathrm{u}$$

$$p(z_1, z_2) = \frac{1}{\pi}, \quad (z_1 + z_2)^2 \leq 1$$

$$\boxed{\begin{aligned} y_1 &= z_1 \left(\frac{-2\ln z_1}{r^2} \right)^{1/2} \\ y_2 &= z_2 \left(\frac{-2\ln z_2}{r^2} \right)^{1/2} \end{aligned}} \qquad r^2 = (z_1 + z_2)^2$$

Box-Muller变换

$$p(y_1, y_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(y_1, y_2)} \right|$$

$$= \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2) \right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2) \right]$$

基本采样法（Basic Sampling）

- 高斯分布：

$$z_1, z_2 \sim \mathrm{u}$$

$$\boxed{\begin{aligned} y_1 &= \sqrt{-2\ln z_1}\cos(2\pi z_2) \\ y_2 &= \sqrt{-2\ln z_1}\sin(2\pi z_2) \end{aligned}}$$

$$p(z \quad (z_1 + z_2)^2 \leq 1$$

不适用于积分及反函数难求的分布!

$$y_2$$

$$r^2 = (z_1 + z_2)^2$$

Box-Muller变换

$$p(y_1, y_2) = p(z_1, z_2)\left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right|$$

$$= \left[\frac{1}{\sqrt{2\pi}}\exp(-y_1^2/2)\right]\left[\frac{1}{\sqrt{2\pi}}\exp(-y_2^2/2)\right]$$

拒绝采样（Rejection Sampling）

- 对一个很复杂的分布 $p(z)$ 进行采样，不能给出其具体的解析形式，但是每个z可以估算其比例

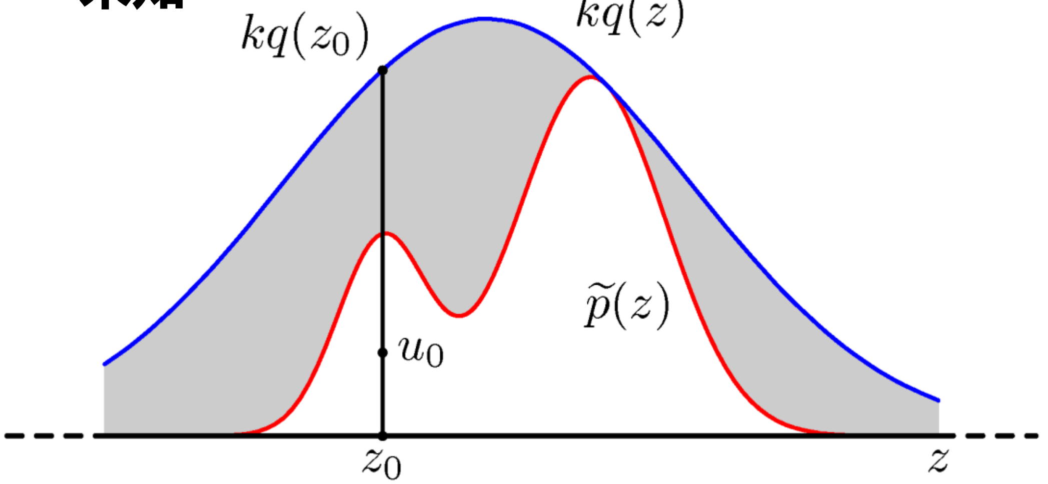
$$p(z) = \frac{1}{Z_p}\tilde{p}(z)$$

已知

未知

- 用 $q(z)$ 去逼近 $\tilde{p}(z)$

$$kq(z) \geq \tilde{p}(z)$$



$q(z)$ 称为建议分布proposal distribution

拒绝采样（Rejection Sampling）

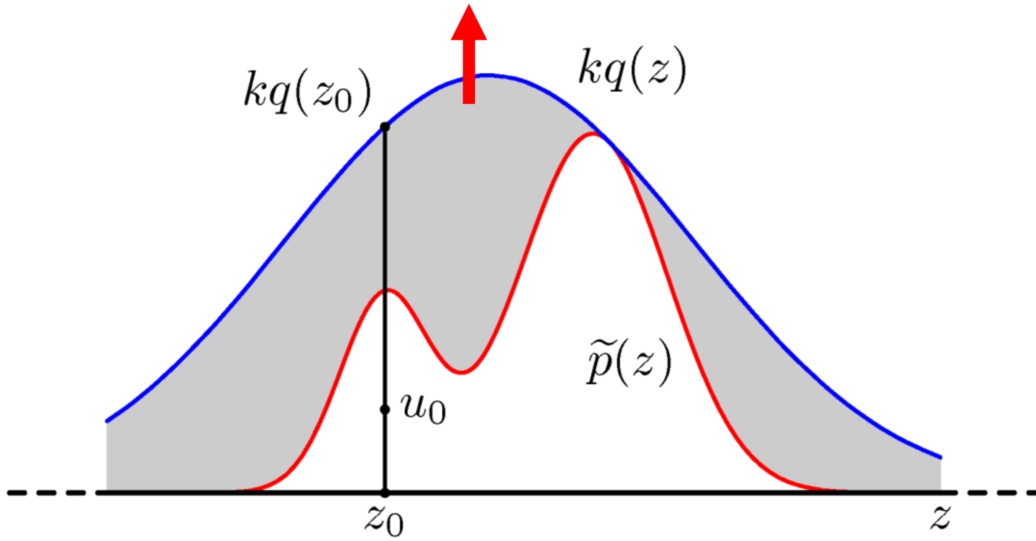
- 采样步骤

1、从 $q(z)$ 中产生 z_0

2、从 $[0, kq(z_0)]$ 均匀分布中产生 μ_0

3、如果 $\mu_0 > \tilde{p}(z_0)$ 拒绝采样，否则接受采样

被拒绝的概率



$$p(\text{accept}) = \int \{\tilde{p}(z)/kq(z)\}q(z)dz = \frac{1}{k}\int \tilde{p}(z)dz$$

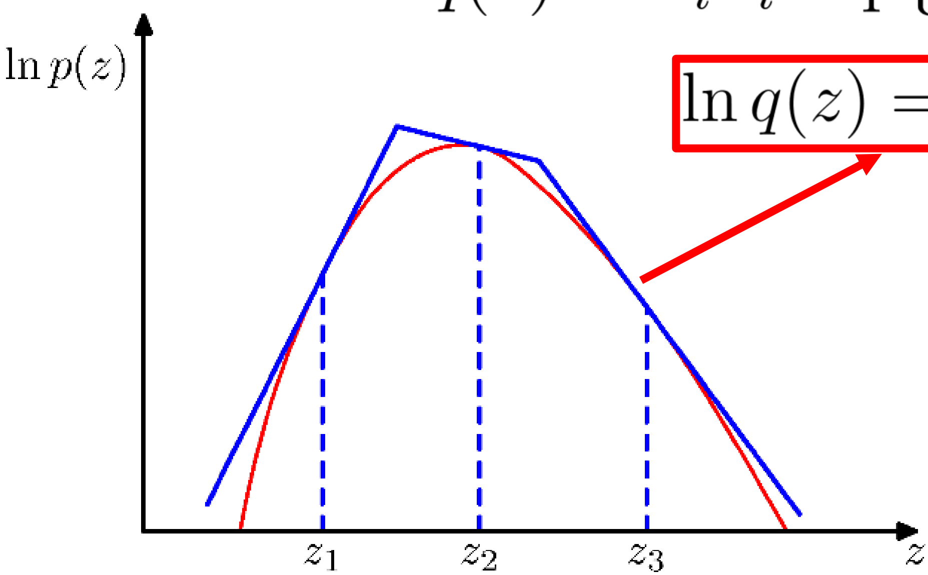
拒绝采样又称为接受-拒绝采样（Acceptance-Rejection Sampling）

自适应拒绝采样 (Adaptive Rejection Sampling)

- 实际应用中，往往很难找到合适的 $q(z)$
- 特别地，当 $p(z)$ 为Log凸函数时，可采用ARS

$$q(z) = k_i \lambda_i \exp\{-\lambda_i(z - z_{i-1})\} \quad z_{i-1} < z \le z_i$$

$$\boxed{\ln q(z) = C - \lambda(z - z_{i-1}), \quad z_{i-1} < z \le z_i}$$



在log域执行拒绝采样

- 如果满足，接受
- 如果拒绝，重新逼近

重要性采样（Importance Sampling）

- 对于个变量 $z \sim p(z)$
- 一个关于z 函数 $f(z)$，预测 $f(z)$ 的期望值：

$$\mathbb{E}(f) = \int f(z)p(z)dz$$

- $p(z)$ 很复杂，但可以估算每个出现 z 的概率 $p(z)$

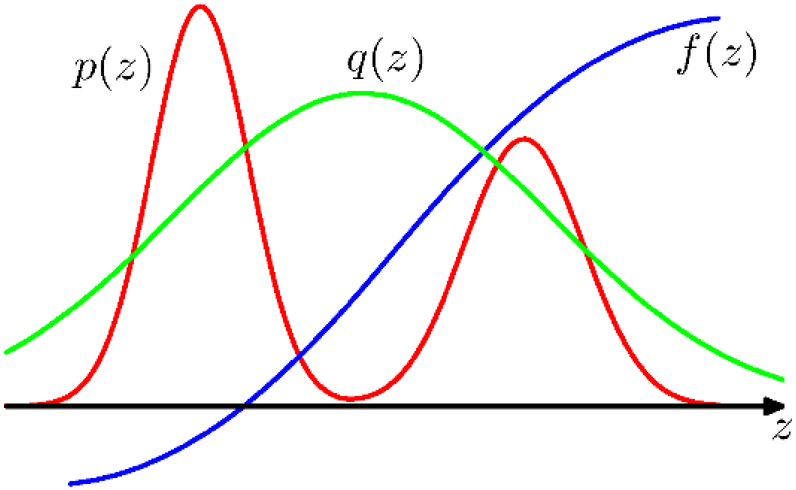
- 要估计 $\mathbb{E}(f) = \int f(z)p(z)dz$
- 可以按如下方式计算-将空间网络化

$$\boxed{\mathbb{E}(f) \approx \sum_{l=1}^{L} p(z^{(l)})f(z^{(l)})} \rightarrow \text{维数灾难}$$

重要性采样（Importance Sampling）

- 与拒绝采样类似，借助一个容易采样的建议分布 $q(z)$

$$\begin{aligned}\mathbb{E}(f) &= \int f(z)p(z)dz \\ &= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\ &\approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^{(l)})}{q(z^{(l)})}f(z^{(l)})\end{aligned}$$



注意：很多时候不知道准确的$p(z)$，而只知道其比例，即 $p(z) = \frac{1}{Z_p}\tilde{p}(z)$

重要性采样（Importance Sampling）

- 同样地，我们也希望 $q(z)$ 具有类似的性质，即

$$q(z) = \frac{1}{Z_q}\tilde{q}(z)$$

- 再看 $\mathbb{E}(f)$

$$\begin{aligned}\mathbb{E}(f) &= \int f(z)p(z)dz \\ &= \frac{Z_q}{Z_p}\int f(z)\frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)dz \\ &\approx \boxed{\frac{Z_q}{Z_p}}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l f(z^{(l)})\end{aligned}$$

其中 $\tilde{r}_l = \dfrac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}$

重要性采样（Importance Sampling）

$$
\begin{aligned}
\frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(z)dz \\
&= \int \frac{\tilde{p}(z)}{\tilde{q}(z)} q(z)dz \\
&\approx \frac{1}{L}\sum_{l=1}^{L}\tilde{r}_l
\end{aligned}
$$

$$
\mathbb{E}(f) \approx \sum_{l=1}^{L} w_l f(z^{(l)})
$$

$$
w_l = \frac{\tilde{r}_l}{\sum_m \tilde{r}_m} = \frac{\tilde{p}(z^{(l)})/q(z^{(l)})}{\sum_m \tilde{p}(z^{(m)})/q(z^{(m)})}
$$

蒙特卡罗方法（ Monte Carlo Method）

注：不能准确知道 $p(z)$ ，而只知道其比例 $\tilde{p}(z)$

$$p(z) = \frac{1}{Z_p}\tilde{p}(z)$$

- 首先产生一个采样点 $\mathbf{z}^{(\tau)}$
- 根据建议概率 $q(\mathbf{z}|\mathbf{z}^{(\tau)})$ 产生新的采样点
- 依次类推，产生马尔可夫链 $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$

- 要求 $q(\mathbf{z}|\mathbf{z}^{(\tau)})$ 尽可能简单，便于产生采样点；
- 有一个准则去决定是接受还是拒绝产生的采样点

Metropolis采样法

- **建议概率：**

$$q(\mathbf{z}_A|\mathbf{z}_B) = q(\mathbf{z}_B|\mathbf{z}_A)$$

- **接受概率：**

$$A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\mathbf{z}^*)}{\tilde{p}(\mathbf{z}^{(\tau)})}\right)$$

- **在(0,1)的均匀分布上获得采样点** u **；**
- **如果** $u < A(\mathbf{z}^*, \mathbf{z}^{(\tau)})$ **则接受** $\mathbf{z}^{(\tau+1)} = \mathbf{z}^*$
- **否则拒绝；**

$$\tau \rightarrow \infty \quad \mathrm{z}^{(\tau)} \rightarrow p(\mathbf{z})$$

马尔可夫链（Markov Chain）

- **一阶马尔可夫链（First Order Markov Chain）**

$$p(z^{(m+1)}|z^{(1)},...,z^{(m)}) = p(z^{(m+1)}|z^{(m)})$$

- **高阶马尔可夫（High Order Markov Chain）**

$$p(z^{(m+1)}|z^{(1)},...,z^{(m)}) = p(z^{(m+1)}|z^{(m)},...,z^{(m-n)})$$

- **转移概率（Transition Probabilities）**

$$T(z^{(m)},z^{(m+1)}) \equiv p(z^{(m+1)}|z^{(m)})$$

- **转移概率矩阵**

$$T = \begin{bmatrix} T(1,1),T(1,2),...,T(1,m) \\ T(2,1),T(2,2),...,T(2,m) \\ \vdots \quad\quad \ddots \quad\quad \vdots \\ T(m,1),T(m,2),...,T(m,m) \end{bmatrix}$$

- **如果对所有的$z^{(m)}$都有相同的转移概率T_m，则称为齐次马尔可夫（Homogeneous Markov）**

马尔可夫链（Markov Chain）

- **一个状态的边缘分布可以表示为**

$$p(z^{(m+1)}) = \sum_{z^{(m)}} p(z^{(m+1)}|z^{(m)})p(z^{(m)})$$

- **平稳性（Stationary,或不变性 Invariant）**

$$p^*(z) = \sum_{z'} T(z',z)p^*(z')$$

- **细致平稳（Detailed balance）**

$$p^*(z)T(z,z') = p^*(z')T(z',z)$$

$$\sum_{z'} p^*(z')T(z',z) = \sum_{z'} p^*(z)T(z,z') = p^*(z)\sum_{z'} p(z'|z) = p^*(z)$$

Metropolis-Hastings 方法

- **思想**：对于需要采样的一分布 $p(z)$，构造一个转移矩阵为 T 的马尔可夫链，使它的平稳分布恰好为 $p(z)$
- 假设有一个转移矩阵 $Q(z,z')=q(z'|z)$，$q(z)$ 为容易采样的分布
- 通常情况下，该转移矩阵难以满足细致平稳条件

$$p(z)q(z'|z)\neq p(z')q(z|z')$$

- 引入 $a(z,z')$ 使

$$p(z)q(z'|z)a(z,z')=p(z')q(z|z')a(z',z)$$

其中：

$$\left.\begin{aligned}a(z,z')&=p(z')q(z|z')\\a(z',z)&=p(z)q(z'|z)\end{aligned}\right\}\text{接受率}$$

Metropolis-Hastings 方法

- **步骤：**

  1、 初始化马尔可夫链状态 $z = z_0$

  2、 对 $\tau = 1, 2, \ldots$ **循环以下过程采样**

  1）第τ**个时刻马尔可夫链状态为** $z = z^{(\tau)}$ **采样** $z^* = q(z|z^{(\tau)})$

  2）从均匀分布中采样 $u \sim \text{uniform}[0, 1]$

  3）如果 $\boxed{u < a(z^{(\tau)}, z^*) = p(z^*)q(z^{(\tau)}|z^*)}$ **则接受** $z^{(\tau+1)} = z^*$

  4）否则 $z^{(\tau+1)} = z^{(\tau)}$

如果$a(z^{(\tau)}, z)$过小，则采样效率较低！

Metropolis-Hastings 方法

- **在细致平稳条件两边乘以因子*C***

$$p(z)q(z'|z)a(z,z')\cdot C = p(z')q(z|z')a(z',z)\cdot C$$

细致平稳条件并没有打破！！！

- **同比例放大** $a(z,z')$, $a(z',z)$ **使最大的为1，令**

$$A(z,z') = \min\left\{1, \frac{p(z)q(z'|z)}{p(z')q(z|z')}\right\}$$

$$= \min\left\{1, \frac{\tilde{p}(z)q(z'|z)}{\tilde{p}(z')q(z|z')}\right\} \quad p(z) = \frac{1}{Z_p}\tilde{p}(z)$$

Metropolis-Hastings 方法

- **步骤:**

  1、 初始化马尔可夫链状态 $z = z_0$

  2、 对 $\tau = 1, 2, \ldots$ **循环以下过程采样**

  1）第 τ **个时刻马尔可夫链状态为** $z = z^{(\tau)}$ **采样** $z^* = q(z|z^{(\tau)})$

  2）从均匀分布中采样 $u \sim \text{uniform}[0, 1]$

  3）如果 $u < A(z^{(\tau)}, z^*) = \min\left\{1, \frac{\tilde{p}(z^*)q(z'|z^*)}{\tilde{p}(z')q(z^*|z')}\right\}$ **则接受** $z^{(\tau+1)} = z^*$

  4）否则 $z^{(\tau+1)} = z^{(\tau)}$

吉布斯采样（Gibbs Sampling）

- 一种特殊的M-H采样算法
- 针对多元分布进行采样 $p(\mathbf{z}) = p(z_1, \ldots, z_M)$

每次只改变一个维度上的值，保持其他维度不变

$$p(z_1, z_2, z_3)$$

首先：初始化 $(z_1^{(0)}, z_2^{(0)}, z_3^{(0)})$

在第 $\mathcal{T}$步，假设已经产生了 $(z_1^{(\tau)}, z_2^{(\tau)}, z_3^{(\tau)})$

→ 根据 $p(z_1|z_2^{(\tau)}, z_3^{(\tau)})$ 产生 $z_1^{(\tau+1)}$

→ 根据 $p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)})$ 产生 $z_2^{(\tau+1)}$

→ 根据 $p(z_3|z_1^{(\tau+1)}, z_2^{(\tau+1)})$ 产生 $z_3^{(\tau+1)}$

与M-H的关系

- **建议概率：**

$$q_k(\mathbf{z}^*|\mathbf{z}) = p(z_k^*|\mathbf{z}_{\setminus k})$$

- **接受概率：**

$$A_k(z^*, z^{(\tau)}) = \min\left\{1, \frac{\tilde{p}(z^*)q_k(z^{(\tau)}|z^*)}{\tilde{p}(z^{(\tau)})q_k(z^*|z^{(\tau)})}\right\}$$

$$p(\mathbf{z}) = p(z_k|\mathbf{z}_{\setminus k})p(\mathbf{z}_{\setminus k}) \qquad \mathbf{z}^*_{\setminus k} = \mathbf{z}_{\setminus k}$$

$$A(\mathbf{z}^*, \mathbf{z}) = \frac{p(\mathbf{z}^*)q_k(\mathbf{z}|\mathbf{z}^*)}{p(\mathbf{z})q_k(\mathbf{z}^*|\mathbf{z})} = \frac{p(z_k^*|\mathbf{z}^*_{\setminus k})p(\mathbf{z}^*_{\setminus k})p(z_k|\mathbf{z}^*_{\setminus k})}{p(z_k|\mathbf{z}_{\setminus k})p(\mathbf{z}_{\setminus k})p(z_k^*|\mathbf{z}_{\setminus k})} = 1$$

Slice Sampling

- **Metropolis 方法的缺点：**
  - **步长太短：走得太慢（可能随机散步）**
  - **步长太长：拒绝率很好，效率较差；**
- **SLICE采样可以自适应调整步长**
  - **将 Z 空间扩展成 (z, u) 空间**

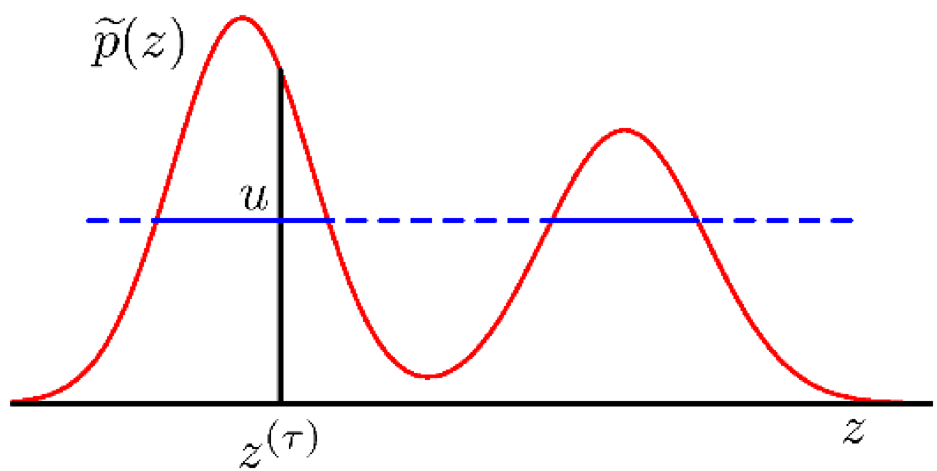
$$\hat{p}(z,u) = \begin{cases} 1/Z_p & \text{if } 0 \le u \le \tilde{p}(z) \\ 0 & \text{otherwise} \end{cases}$$

$$\int \hat{p}(z,u)du = \int_0^{\tilde{p}(z)} \frac{1}{Z_p} du = \frac{\tilde{p}(z)}{Z_p} = p(z)$$

Slice Sampling

$$\hat{p}(z,u) = \begin{cases} 1/Z_p & \text{if } 0 \leq u \leq \tilde{p}(z) \\ 0 & \text{otherwise} \end{cases} \quad \int \hat{p}(z,u)du \int_0^{\tilde{p}(z)} \frac{1}{Z_p} du = \frac{\tilde{p}(z)}{Z_p} = p(z$$

- **第一步：给定**z**，在** $0 \leq u \leq \tilde{p}(z)$**范围内均匀分布产生** u
- **第二步：给定**u**，在**$\{z : \tilde{p}(z) > u\}$**范围内均匀分布产生** z



(a)

第十一讲：概率图模型

Chapter 11: Probabilistic Graphical Models

概率图模型

- 概率图模型（Probabilistic Graphical Model）

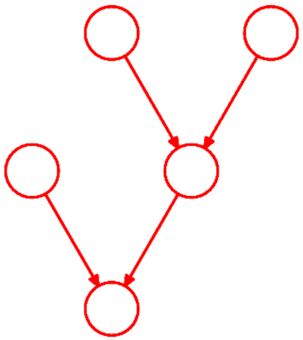
概率论

$$p(X) = \sum_{Y} p(X, Y)$$

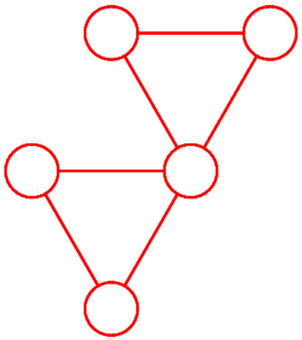
$$p(X, Y) = p(Y|X)p(X)$$

$$p(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

图论



有向图

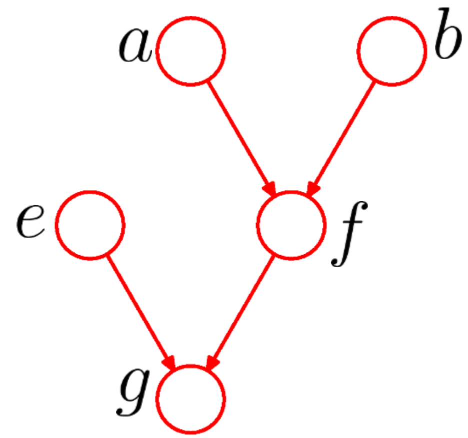


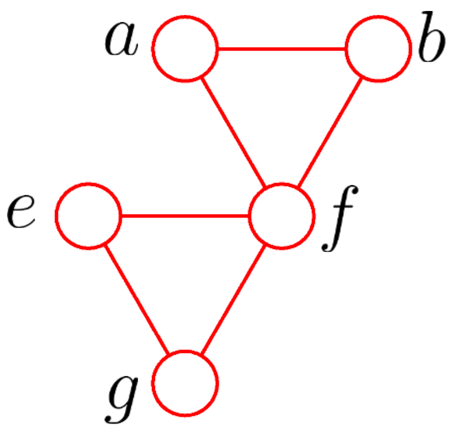
无向图

概率论 + 图论 = 概率图模型

概率图模型

- **概率图模型（Probabilistic Graphical Model）**

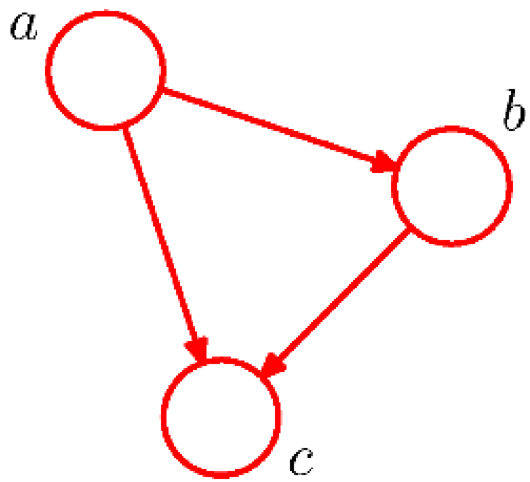




- **结点：随机变量或一组随机变量**
- **连接弧：随机变量之间的关系**

概率图模型

- **贝叶斯网络（Bayesian network）**
  - **有向无环图（Directed Acyclic Graph, DAG）**

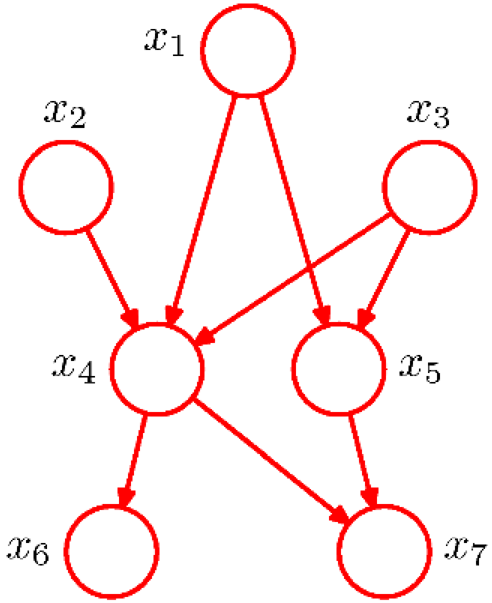


$$p(a,b,c) = p(c|a,b)p(a,b) = p(c|a,b)p(b|a)p(a)$$

$$p(x_1,, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$$

概率图模型

- **贝叶斯网络（Bayesian network）**

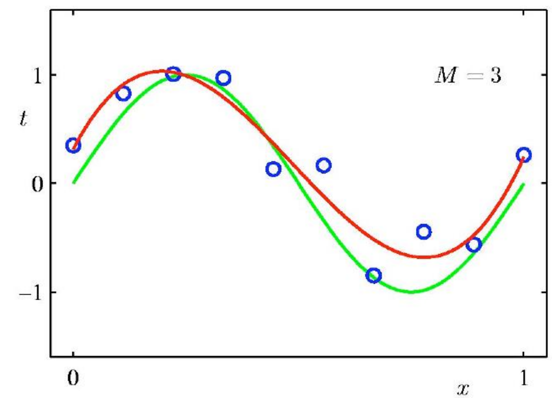


$$
\begin{aligned}
p(x_1, \ldots, x_7) \quad = \quad & p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\
& p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)
\end{aligned}
$$

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

概率图模型

- 例子：多项式拟合

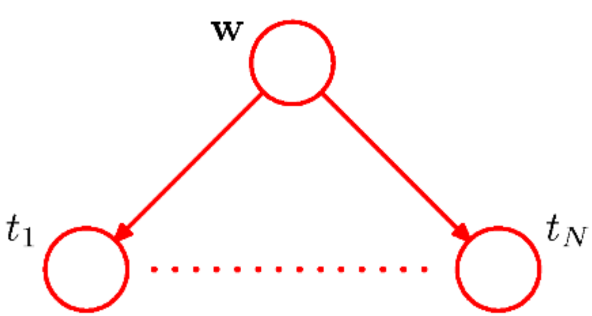


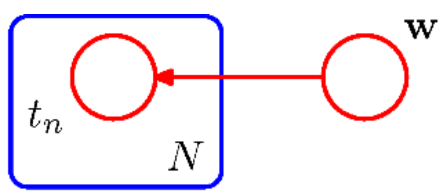
$$p(\mathbf{t},\mathbf{w}) = p(\mathbf{w})\prod_{n=1}^{N} p(t_n|y(\mathbf{w},x_n))$$

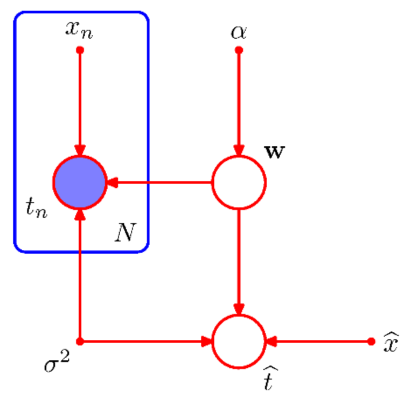
$$p(\mathbf{t},\mathbf{w}|\mathbf{x},\alpha,\sigma^2) = p(\mathbf{w}|\alpha)\prod_{n=1}^{N} p(t_n|\mathbf{w},x_n,\sigma^2)$$

$$y(x,\mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

$$p(\widehat{t},\mathbf{t},\mathbf{w}|\widehat{x},\mathbf{x},\alpha,\sigma^2) = \left[\prod_{n=1}^{N} p(t_n|x_n,\mathbf{w},\sigma^2)\right] p(\mathbf{w}|\alpha)p(\widehat{t}|\widehat{x},\mathbf{w},\sigma^2)$$







条件独立
（Conditional independence）

- **三个变量** a, b, c

$$p(a|b,c) = p(a|c)$$

称在**给定*c*的条件下，*a*与*b*条件独立**

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}$$

$$p(a,b|c) = p(a|c)p(b|c) \Rightarrow a \perp\!\!\!\perp b \mid c$$

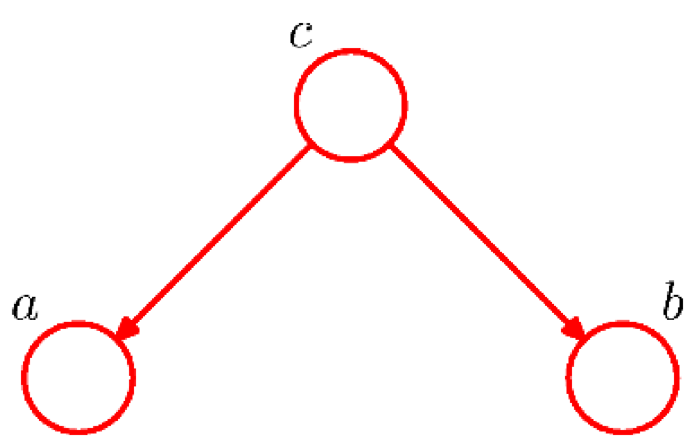
条件独立
（Conditional independence）

$$p(a,b,c) = p(a|c)p(b|c)p(c)$$

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c)$$

$$\stackrel{?}{=} p(a)p(b)$$

$$\boxed{a \not\perp\!\!\!\perp b \mid \emptyset}$$



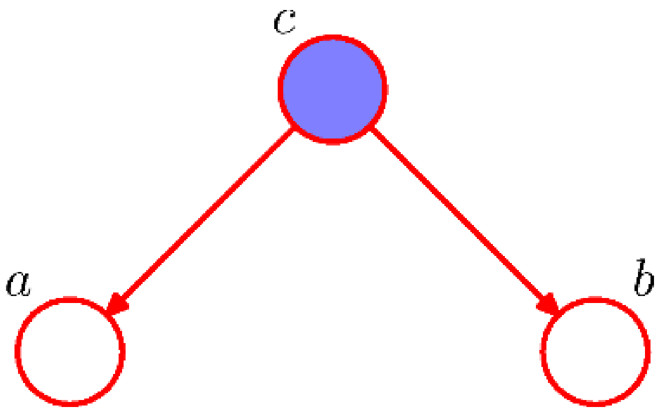
尾尾相连（tail-to-tail)

条件独立
（Conditional independence）

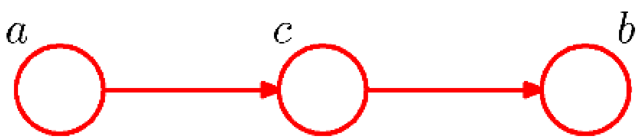
$$p(a, b|c) = \frac{p(a, b, c)}{p(c)}$$

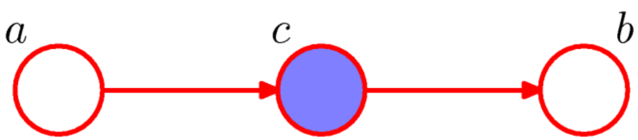
$$= p(a|c)p(b|c)$$

$$\boxed{a \perp\!\!\!\perp b \mid c}$$



条件独立（Conditional independence）





头尾相连（head-to-tail）

$$p(a,b,c) = p(a)p(c|a)p(b|c)$$

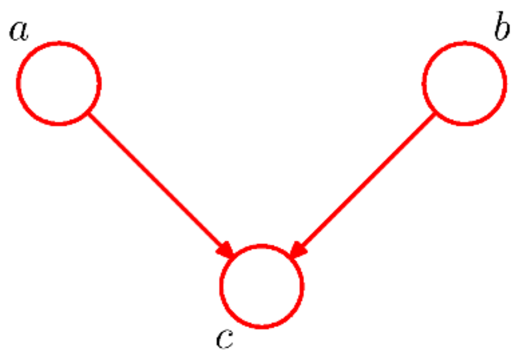
$$p(a,b) = p(a)\sum_c p(c|a)p(b|c)$$
$$= p(a)p(b|a)$$

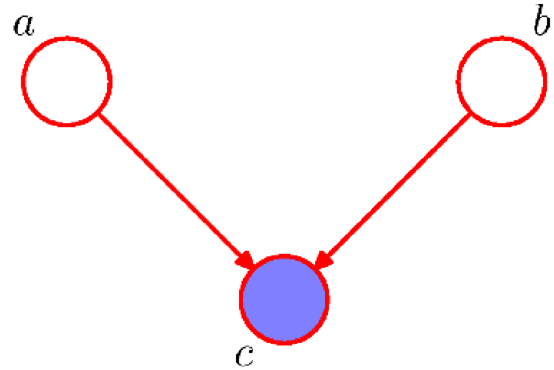
$$a \not\perp\!\!\!\perp b \mid \emptyset$$

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)}$$
$$= \frac{p(a)p(c|a)p(b|c)}{p(c)}$$
$$= p(a|c)p(b|c)$$

$$a \perp\!\!\!\perp b \mid c$$

条件独立
(Conditional independence)





头头相连（head-to-head)

$$p(a,b,c) = p(a)p(b)p(c|a,b)$$

$$p(a,b) = p(a)p(b)$$

$$a \perp\!\!\!\perp b \mid \emptyset$$

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)}$$

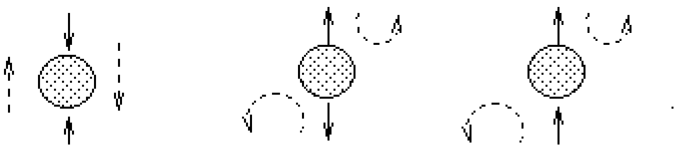
$$a \not\!\perp\!\!\!\perp b \mid c$$

“D-分离”（d-separation）

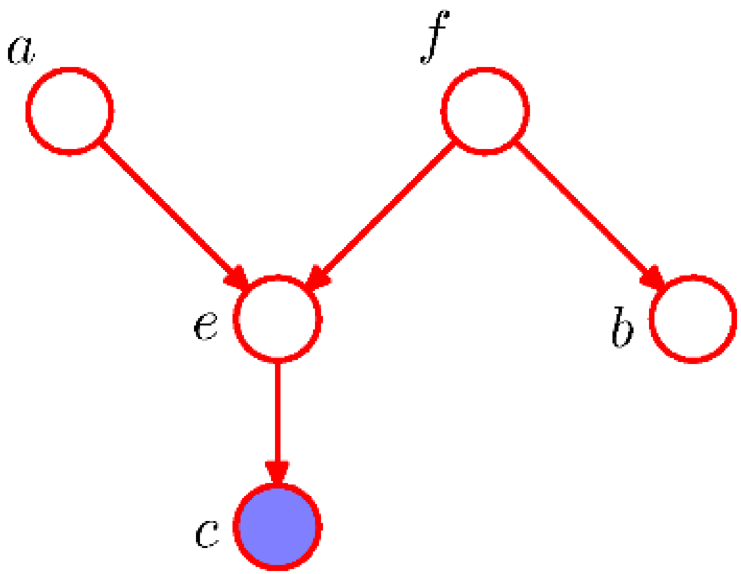
$$A \perp\!\!\!\perp B \mid C$$

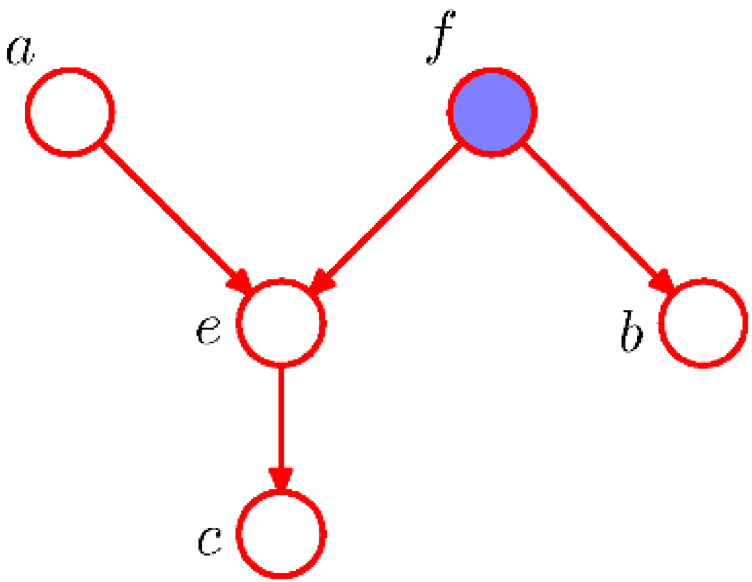
看A与B相连的每条路径有没有都被阻隔（blocked）

- C中的结点满足“头尾相连”或“尾尾相连”；
- “头头相连”的节点和它的任何后裔节点都不在C中



“D-分离”
(d-separation)

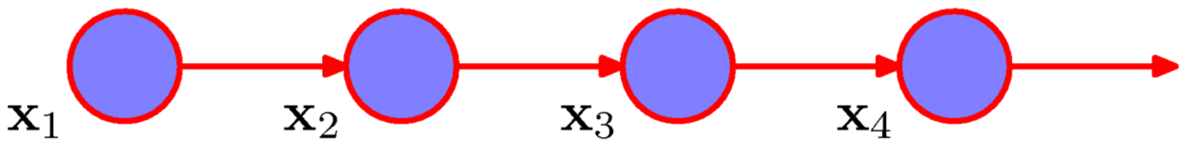




$$a \not\perp\!\!\!\perp b \mid c \qquad\qquad a \perp\!\!\!\perp b \mid f$$

隐马尔可夫模型（Hidden Markov Models）

- 马尔可夫链



$$p(\mathbf{x}_n|\mathbf{x}_1,\ldots,\mathbf{x}_{n-1}) = p(\mathbf{x}_n|\mathbf{x}_{n-1})$$

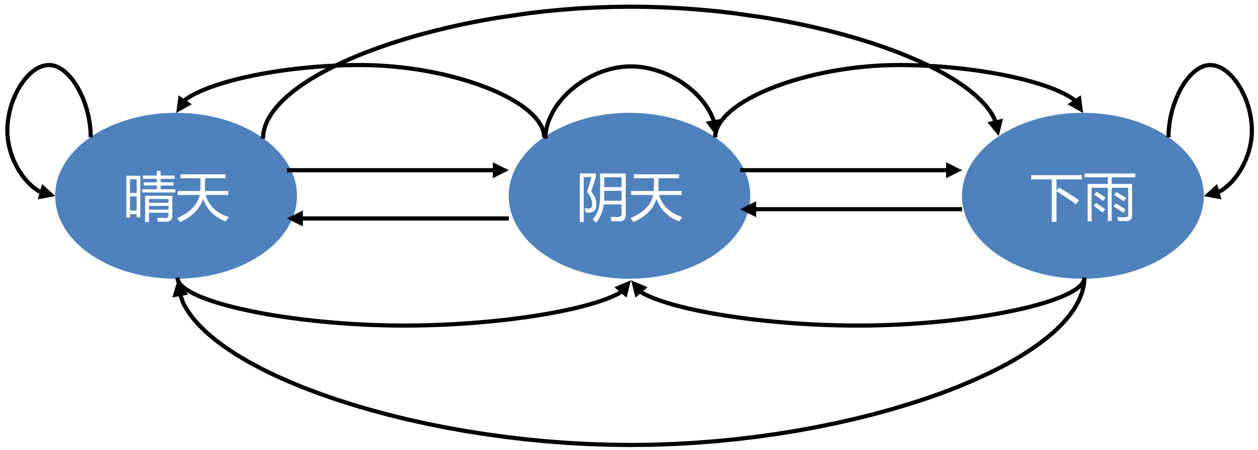
$$p(\mathbf{x}_1,\ldots,\mathbf{x}_n) = p(\mathbf{x}_1)\prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1})$$

如果一个过程的“将来”仅依赖“现在”而不依赖“过去”，则此过程具有马尔可夫性,或称此过程为马尔可夫过程。

$$X(t+1) = f(X(t))$$

隐马尔可夫模型
（Hidden Markov Models）

- **时间**和**状态**都离散的马尔可夫过程称为马尔可夫链



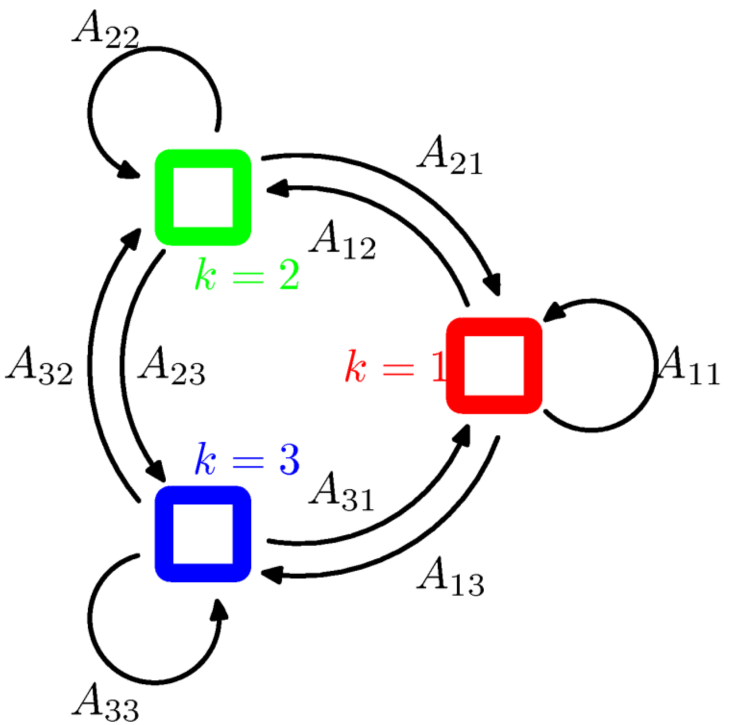
隐马尔可夫模型
（Hidden Markov Models）

- **转移概率**

$$A_{jk} \equiv p(\mathbf{z}_{nk} = 1 | \mathbf{z}_{n-1,j} = 1)$$

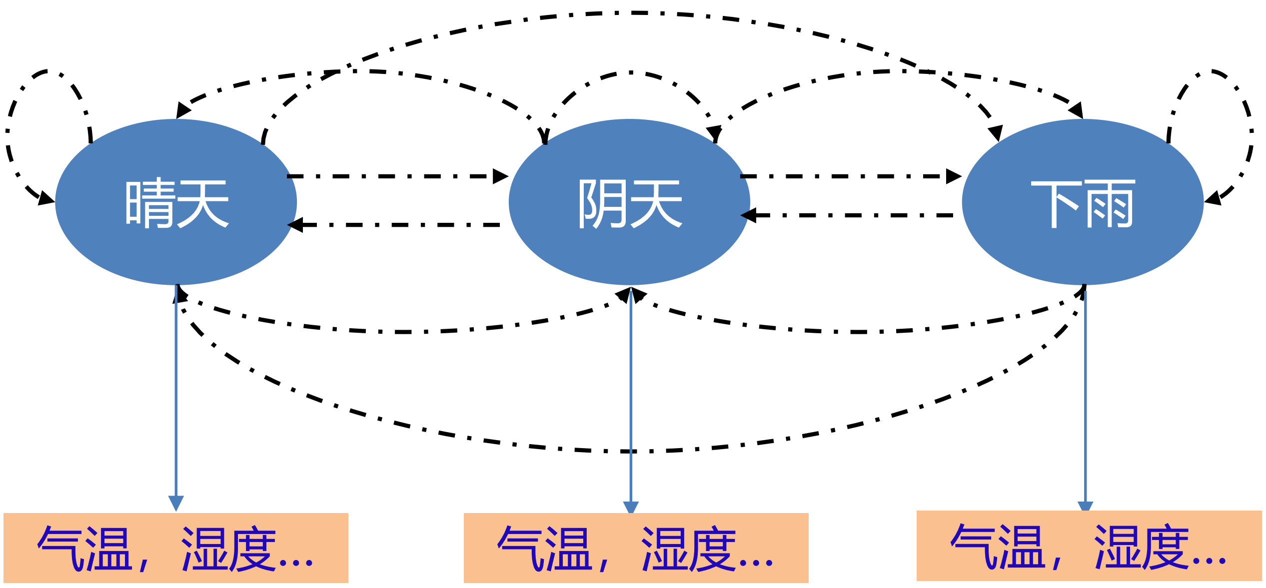
$$0 \leq A_{jk} \leq 1$$

$$\sum_k A_{jk} = 1$$



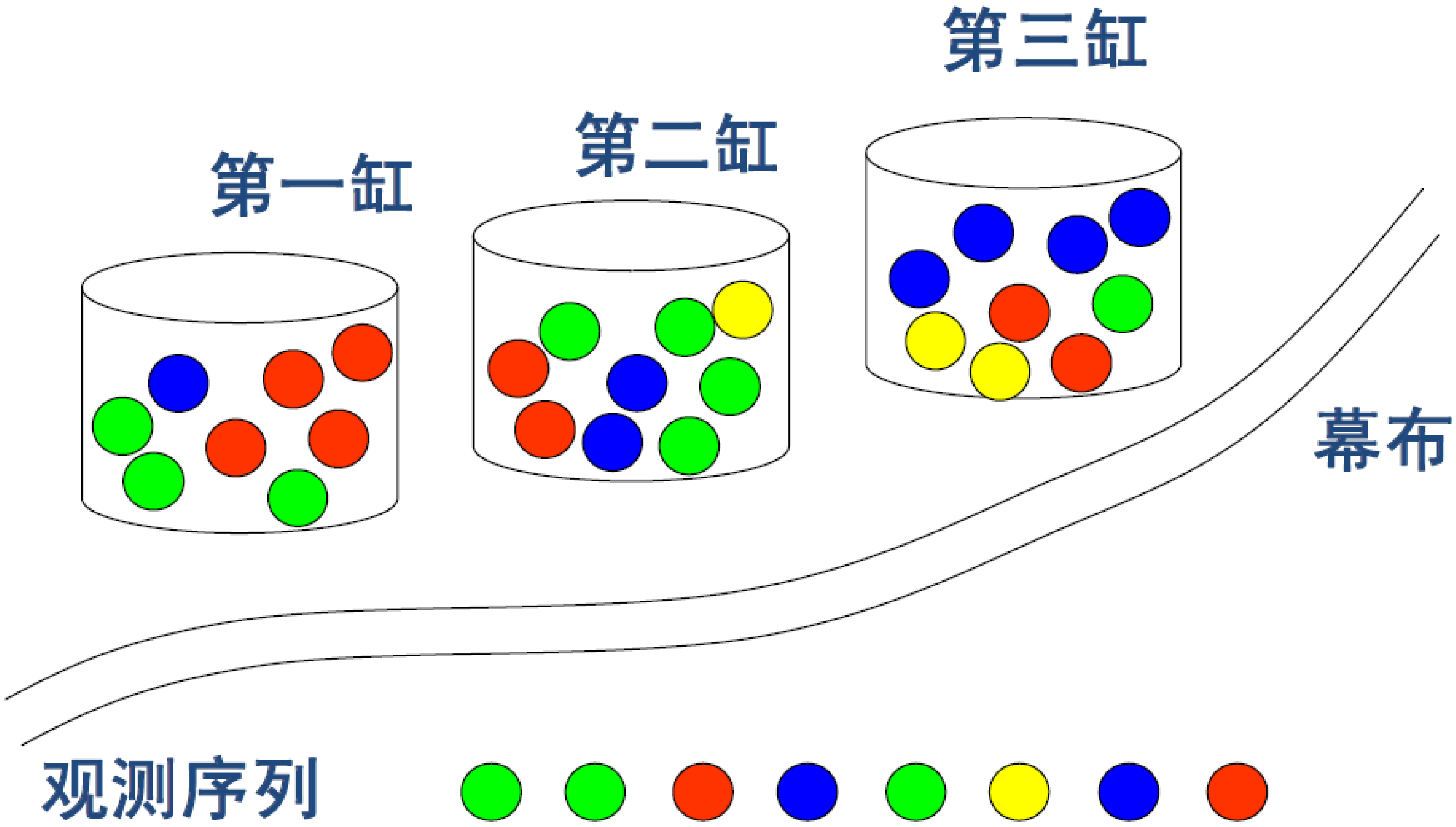
隐马尔可夫模型（Hidden Markov Models）

- 状态序列（state sequence）不可见的过程为隐马尔可夫过程



- 只能得到对状态的观测序列（observation sequence）

隐马尔可夫模型
（Hidden Markov Models）



隐马尔可夫模型（Hidden Markov Models）

- **假设有N个缸，每个缸中装有很多个彩球**

缸	1	2	3	…
颜色1	m_{11}	m_{12}	m_{13}	
颜色2	m_{21}	m_{22}	m_{23}	…
颜色3	m_{31}	m_{32}	m_{33}	
		…		

- **按照以下方式实验**
  - 随机从N个缸中选取一个
  - 从中随机抽出一个球，记录其颜色O_1，并放回
  - 以概率p_i转移到缸i

  重复以上步骤

最后得到一个球的观测序列 $O = \{O_1, O_2, ..., O_T\}$

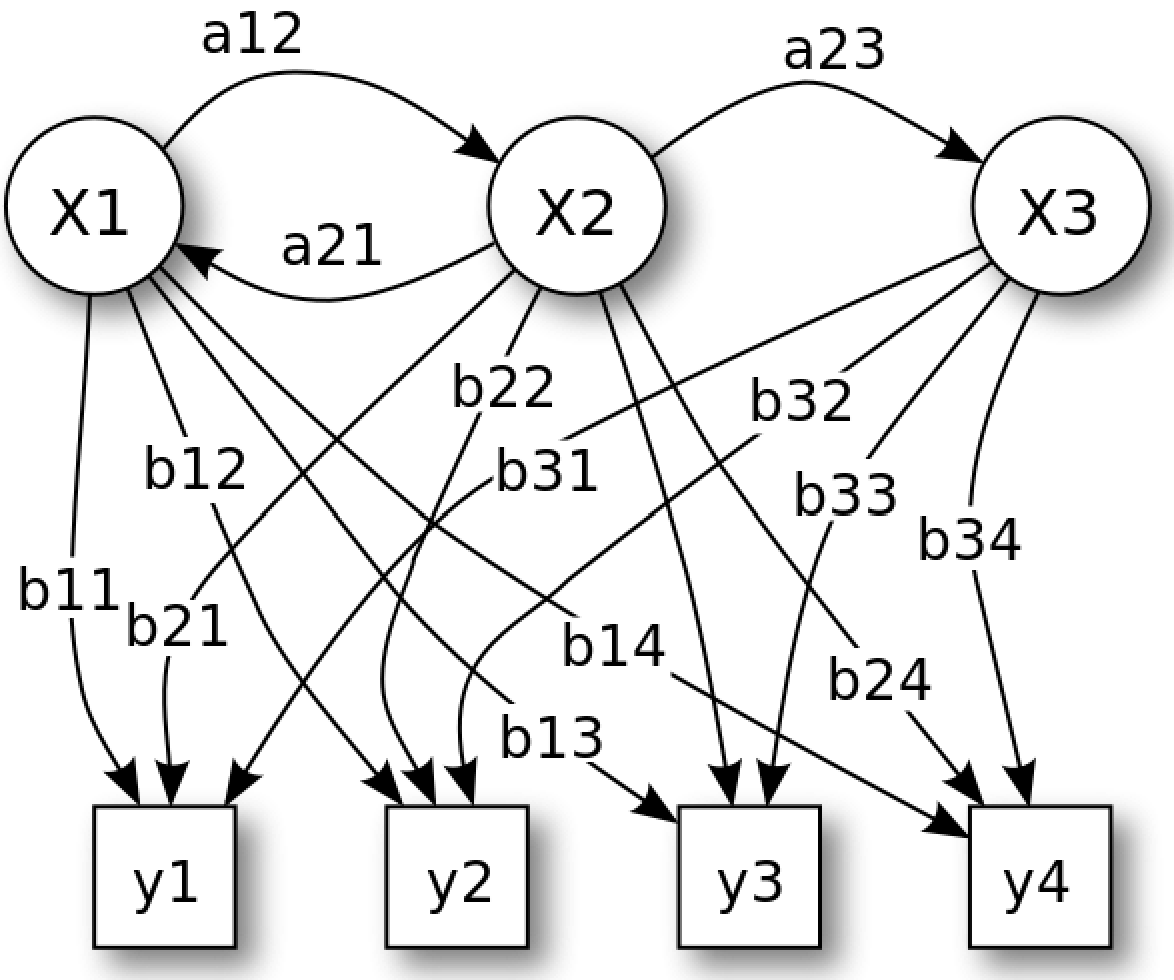
隐马尔可夫模型
（Hidden Markov Models）

- **HMM的状态是不确定或不可见的，只有通过观测序列的随机过程才能表现出来**

- **观察到的事件与状态并不是一一对应，而是通过一组概率分布相联系**

- **HMM是一个双重随机过程**
  - 马尔可夫随机：状态之间的转移是随机的，且具有马尔可夫性，状态之间的转移用**转移概率**描述。
  - 一般随机过程：状态生成某种观测是随机的，用**观测概率**描述

隐马尔可夫模型（Hidden Markov Models）



隐马尔可夫模型
（Hidden Markov Models）

- **HMM的模型用** (N, M, π, A, B) **五元组来表示，或简写为** $\lambda = (\pi, A, B)$

参数	含义	实例
N	状态数目	缸的数目
M	观测值数目	球的颜色数目
π	初始状态概率分布	初始时选择某口缸的概率
A	与时间无关的状态转移概率矩阵	在某个缸选后之后跳转到另一个缸的概率
B	状态生成观测的概率矩阵	每个缸的颜色数目

- **HMM两个基本假设**
  - 齐次马尔可夫假设
  - 观测独立假设

隐马尔可夫模型
（Hidden Markov Models）

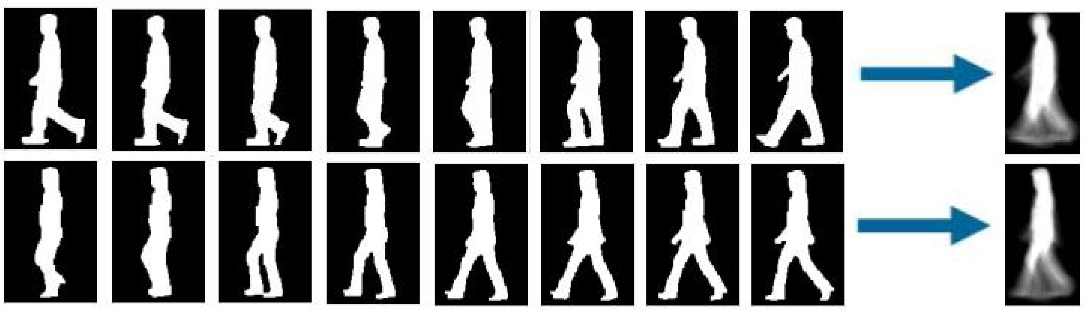
- **HMM的三个基本问题**
  - **概率计算问题**：给定观测序列 前向-后向算法 ..., $O_T\}$，以及模型$\lambda=(\pi,A,B)$，如何计算

  - **预测问题（又叫解码问题）**：给定观测序列$O=\{O_1,O_2,...,O_T\}$，以及模型 维特比（Viterbi）算法 如何选择一个状态序列$S=(q_1,q_2,...,q_T)$，使 合理的解释观测序列

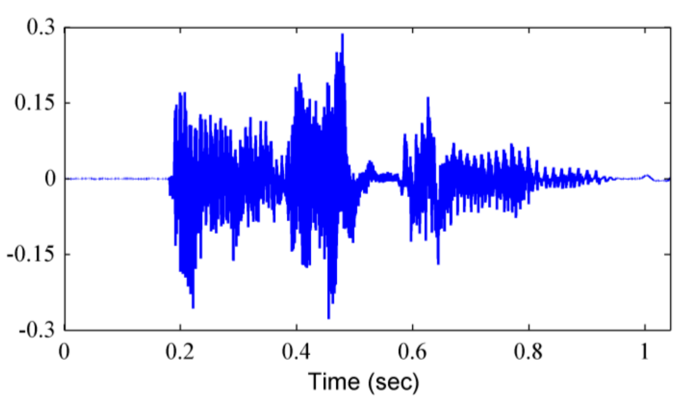
  - **学习问题**：给定观测序列$O=$ Baum-Welch算法 $\}$，估计模型参数$\lambda=(\pi,A,B)$，使得在该 列出现的概率最大

隐马尔可夫模型（Hidden Markov Models）

- **HMM的应用**
  - 步态识别:



  - 语音识别:



“Bayes”

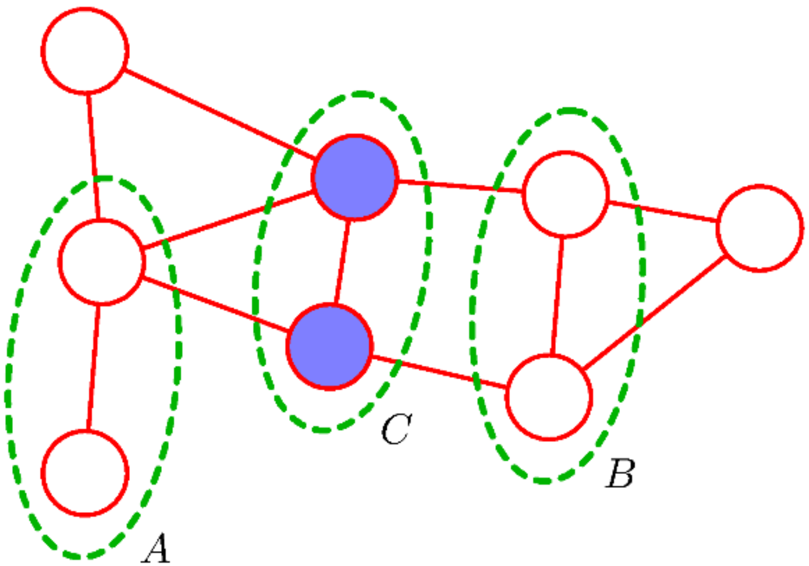
Markov Random Fields

- 马尔可夫随机场（Markov Random Field, Markov Network or Undirected Graphical Model）

  - 如果A，B之间每条路径存在至少 一个节点在C中；

或者

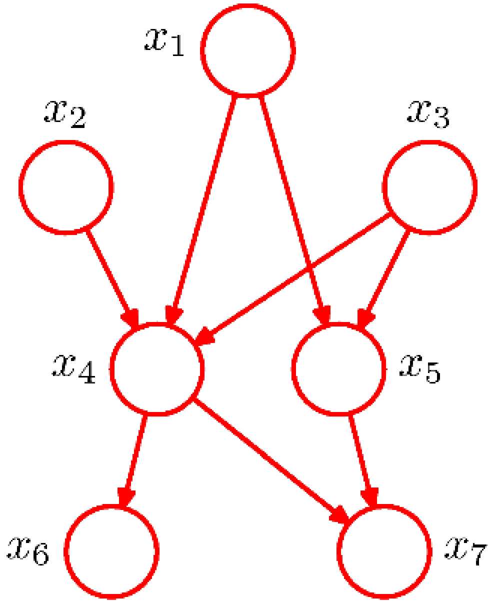
  - 如果去掉C中的所有的节点，A和B没有连通路径



$$A \perp\!\!\!\perp B \mid C$$

因式分解

- 贝叶斯网络（Bayesian network）

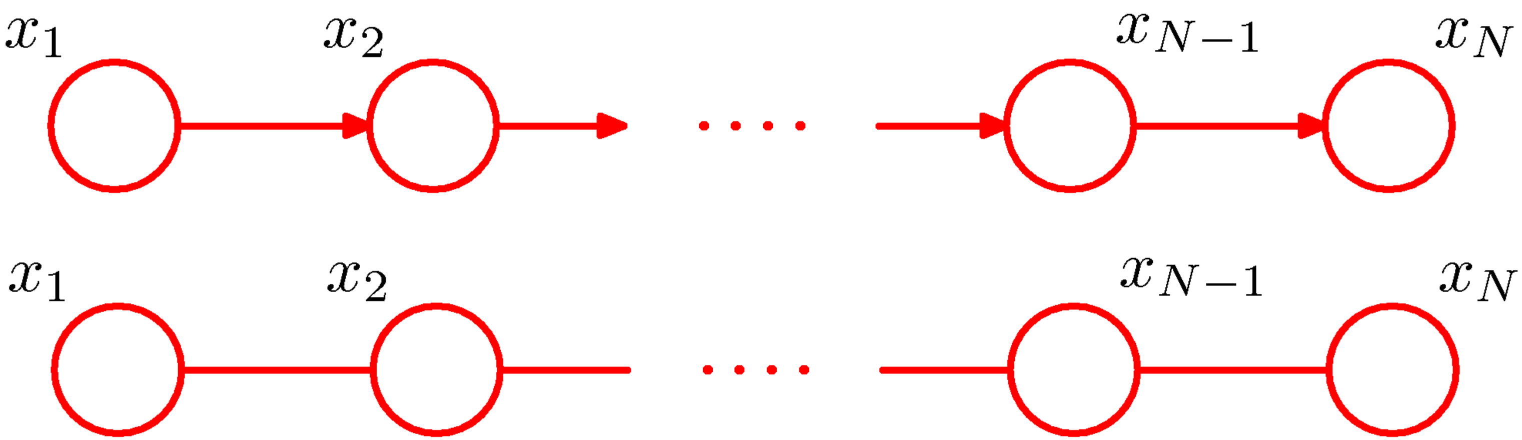


$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

因式分解（Factorization）

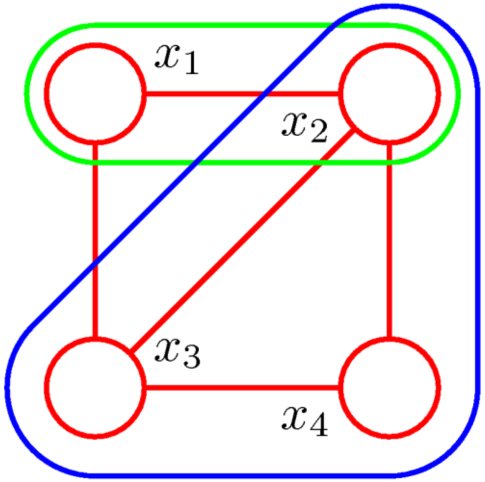
- **马尔可夫随机场**



$$p(x_i, x_j | \mathbf{x}_{\backslash\{i,j\}}) = p(x_i | \mathbf{x}_{\backslash\{i,j\}}) p(x_j | \mathbf{x}_{\backslash\{i,j\}})$$

因式分解（Factorization）

- **马尔可夫随机场**



团(*clique*)

$$\{x_1, x_2\}$$

$$\{x_2, x_3\}$$

$$\{x_3, x_4\}$$

$$\{x_4, x_2\}$$

$$\{x_1, x_3\}$$

$$\{x_1, x_2\, x_3\}$$

$$\{x_2\, x_3, x_4\}$$

最大团(maximal clique)

势函数（potential functions）

- 对于一团C，C中的元素 $\mathbf{x}_C$，定义函数$\psi_C(\mathbf{x}_C)$ 为 $\mathbf{x}_C$ 的势（能）函数（potential functions）

- 团C的联合分布概率为

$$p(\mathbf{x}) = \frac{1}{Z}\prod_C \psi_C(\mathbf{x}_C) \quad \psi_C(\mathbf{x}_C) \geq 0$$

Z 为归一化常量，又叫配分函数（partition function ）

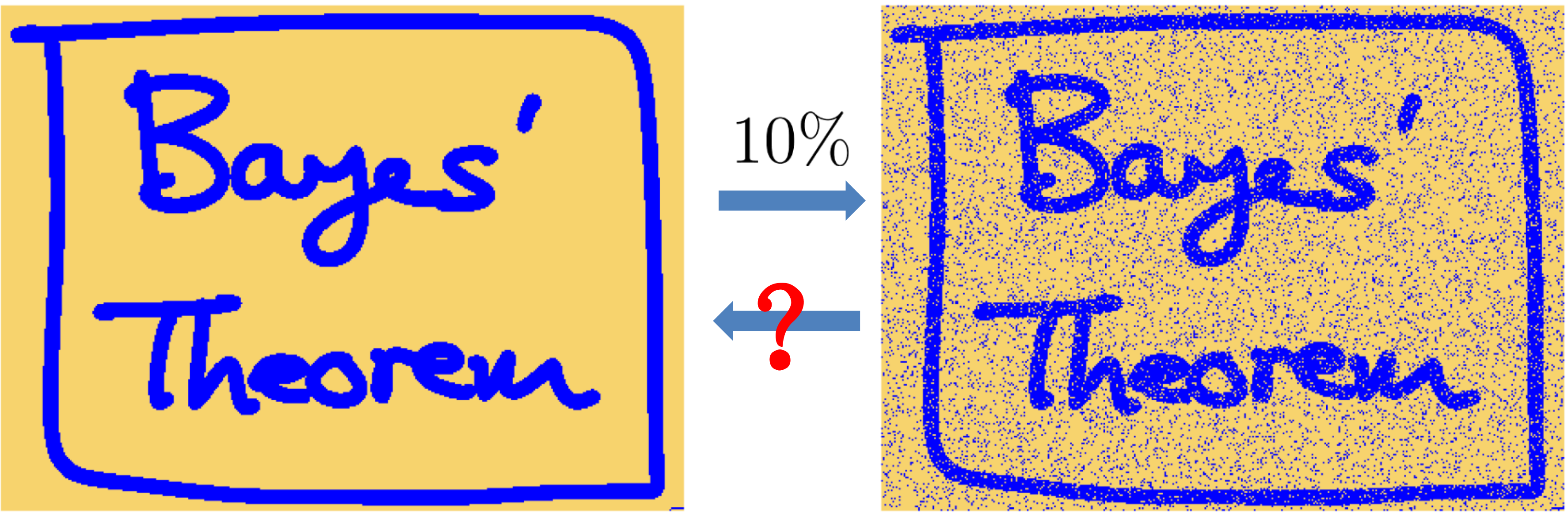
$$Z = \sum_{\mathbf{x}}\prod_C \psi_C(\mathbf{x}_C)$$

$$\boxed{\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\}}$$

波尔兹曼分布

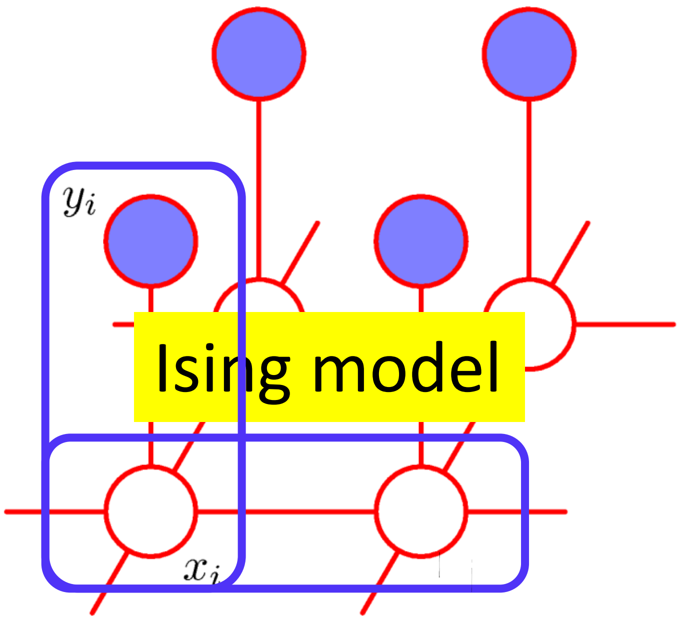
能量函数

例子-图像去噪



$\{-1, +1\}$

例子-图像去噪



$$p(\mathbf{x},\mathbf{y}) = \frac{1}{Z}\exp\{-E(\mathbf{x},\mathbf{y})\}$$

$$\{x_i, y_i\}$$

$$\psi_C(\mathbf{x}_C) = \exp\{-\boxed{E(\mathbf{x}_C)}\}$$

$$-\eta x_i y_i$$

$$\{x_i, x_j\}$$

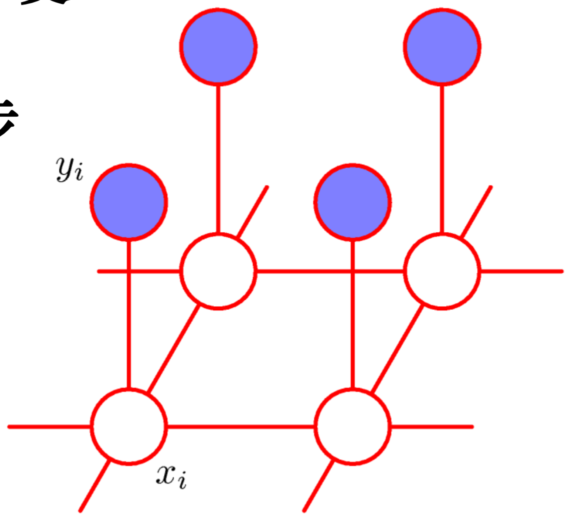
$$\psi_C(\mathbf{x}_C) = \exp\{-\boxed{E(\mathbf{x}_C)}\}$$

$$-\beta x_i x_j$$

$$E(\mathbf{x},\mathbf{y}) = h\sum_i x_i - \beta\sum_{\{i,j\}} x_i x_j - \eta\sum_i x_i y_i$$

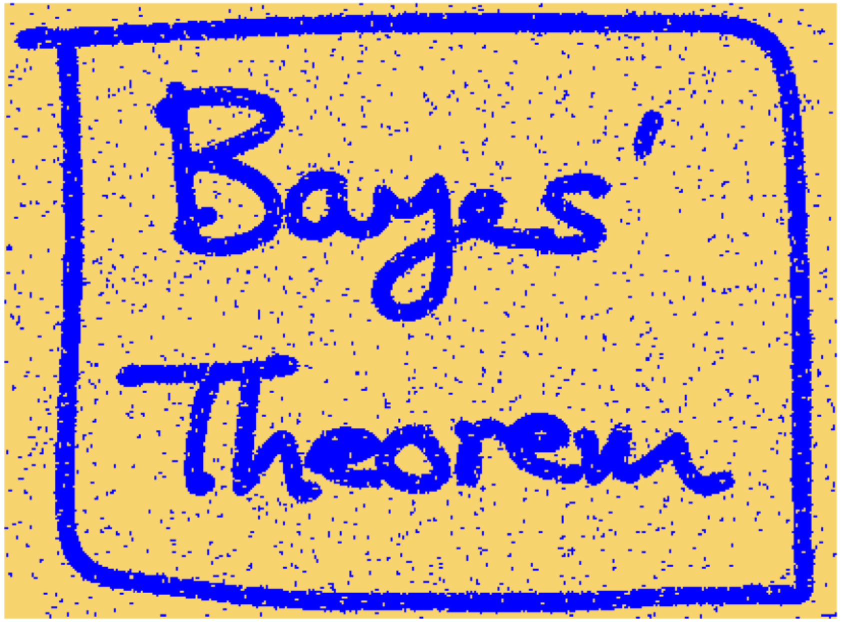
例子-图像去噪

- 梯度下降法（iterated conditional modes, ICM, Kittler1984）
  - 1. 初始化：$x_i = y_i \quad \text{for all} \quad i$
  - 2. 逐像素遍历图像，尝试改变 x_i 的值 $\begin{cases} x_i = +1 \rightarrow x_i = -1 \\ x_i = -1 \rightarrow x_i = +1 \end{cases}$
  - 3. 如果能量降低则改变，如果能量升高则不变
  - 4. 达到终止条件：停止；否则，回到第二步

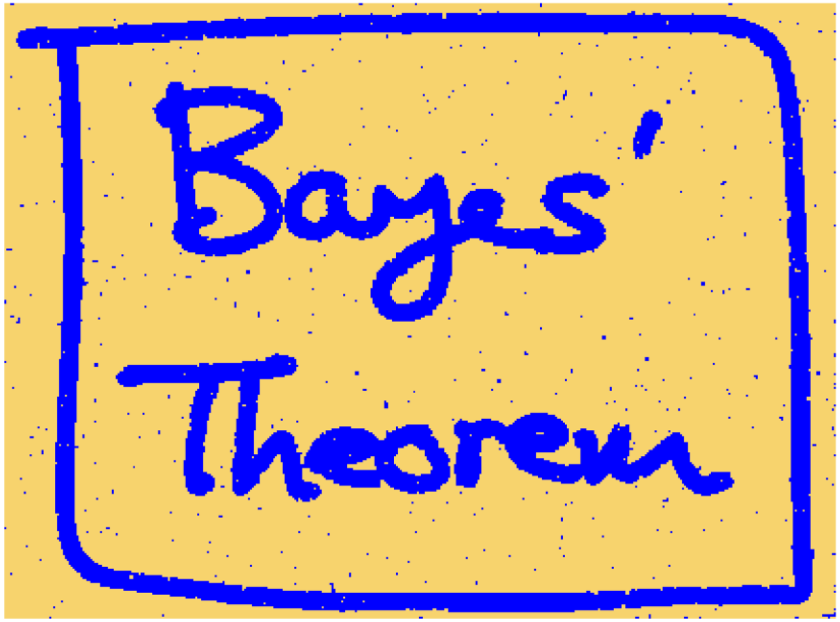


- 高效！但有可能陷入局部极值。

例子-图像去噪



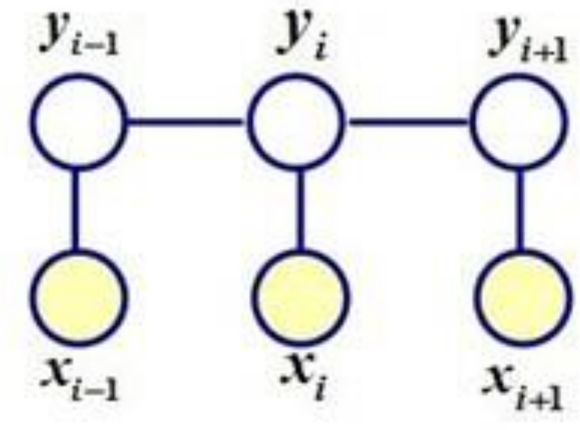
ICM

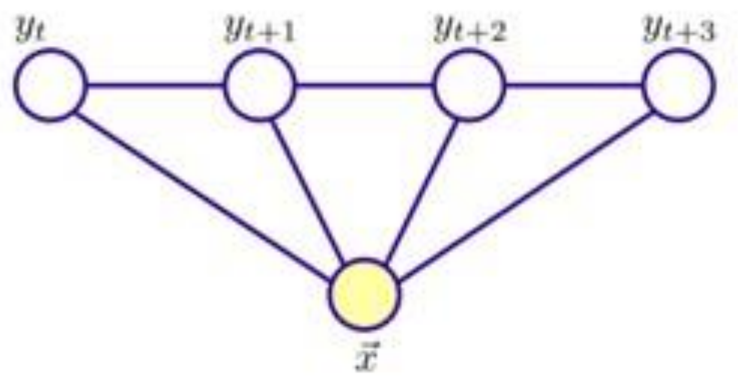


Graph Cut

条件随机场（Conditional Random Fields）

- 条件随机场（Conditional Random Fields，CRFs）模型最早由Lafferty等人于2001年（ICML2001）提出的。
- CRF是在给定随机变量*X*（或*X*的观测）条件下，随机变量*Y*的马尔可夫场。





条件随机场
（Conditional Random Fields）

- **CRF定义**

X与Y是随机变量，是在给定X的条件下的条件概率分布，若随机变量Y构成一个由无向图G =（*V*, *E*）表示的马尔可夫场，即

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$$

对任意结点v成立，则称条件概率分布$p(Y|X)$ 为条件随机场. 式中$w \sim v$表示在图G中与结点v有边连接的所有结点w,$w \neq v$表示结点v以外的所有结点。Y_v, Y_w 为结点 v, w 对应的随机变量。

条件随机场
（Conditional Random Fields）

- **线性链条随机场（Linear-chain CRF）**

令 $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ **表示观测序列**
是 $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ **状态序列变量，有：**

$$p(\mathbf{y}|\mathbf{x}, \lambda) \propto \exp\left(\sum_j \lambda_j \mathbf{t}_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k \mathbf{s}_k(y_i, x, i)\right)$$

$\mathbf{t}_j(y_{i-1}, y_i, x, i)$ **是定义在边上的特征函数，称为转移特征函数，依赖当前和前一个位置。**

$\mathbf{s}_k(y_i, x, i)$ **是定义在结点上的特征函数，称为状态特征函数，依赖于当前位置。**

条件随机场
（Conditional Random Fields）

- 两个特征函数可以统一为：$f_j(y_{i-1}, y_i, x, i)$

  则有：

$$p(\mathbf{y}|\mathbf{x}, \lambda) = \frac{1}{\mathbf{Z}(\mathbf{x})} \exp\left(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)\right)$$

其中：

$$\mathbf{Z}(\mathbf{x}) = \sum_{j} \exp\left(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)\right)$$

条件随机场（Conditional Random Fields）

- 特征函数为布尔函数，当满足特征条件时为1，否则为0.
  其中：

$$f_j(y_{i-1}, y_i, x, i) = \begin{cases} 1 & T(\text{condition}) \\ 0 & \text{otherwise} \end{cases}$$

- CRF的概率计算：前向-后向算法
- CRF的参数学习：改进的迭代尺度法（IIS），梯度下降法，拟牛顿法
- CRF的预测算法：维特比算法