马尔可夫链蒙特卡罗方法
Markov Chain Monte Carlo

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

计算机是如何产生随机数的？

- **计算机本身无法产生真正的随机数！**
- 但是计算机可以根据一定的算法产生伪随机数(pseudo-random numbers)
- 最古老最简单的莫过于**线性同余生成器**

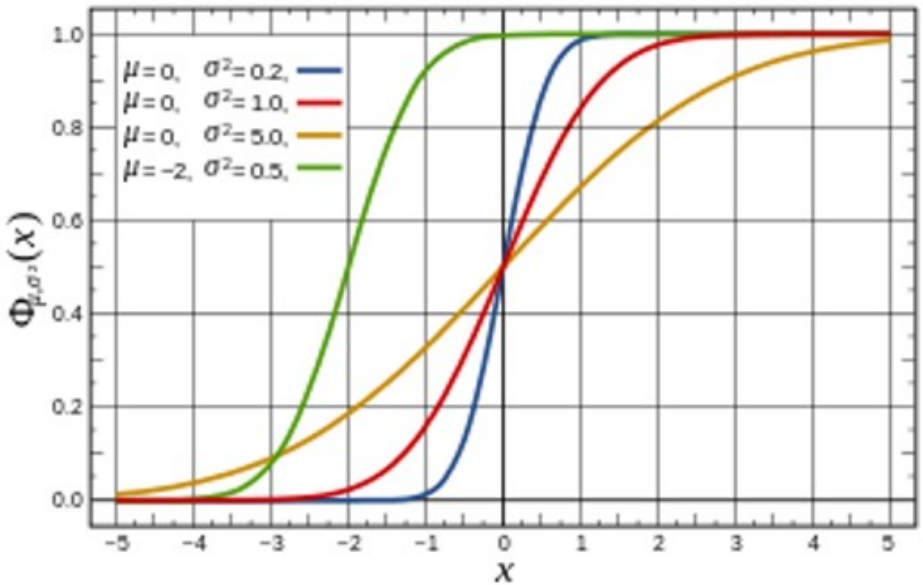
$$x_{n+1} = (ax_n + c) \bmod m$$

  - 可以产生满足均匀分布的随机数
  - 其中a和c是一些数学知识推导出的合适的常数
  - 这种算法产生的下一个随机数完全依赖现在的随机数的大小，而且当随机数序列足够大的时候，随机数将出现重复子序列的情况
- 有一些先进的满足均匀分布的随机数的生成算法
  - 比如python数值运算库numpy用的是Mersenne Twister
- 不管算法如何发展，这些都不是本质上的随机数

Anyone who considers arithmetic methods of producing random digits is, of course, in a state of sin.
——John von Neuman

计算机是如何产生随机数的？

- 如何产生满足其他分布（比如高斯分布）的随机数呢？
  - 上述过程亦被称为**采样**（Sampling）
- 第一种方法：**逆变换采样法**（Inverse transform sampling，亦被称为Smirnov transform）
- 以一维高斯分布举例，该采样方法的原理是利用高斯分布的累积分布函数(cumulative distribution function, CDF)来处理，过程如下



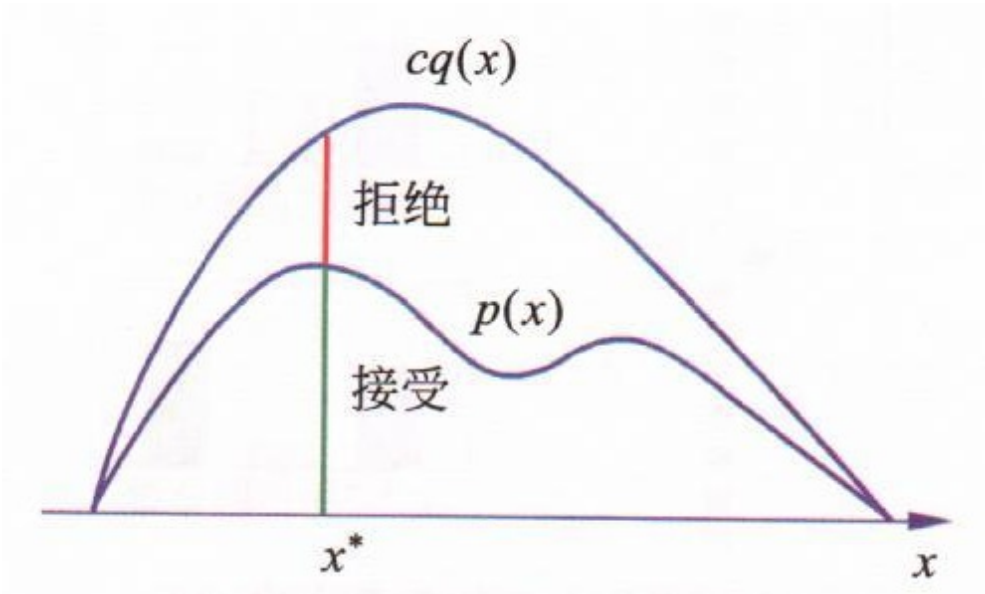
- 在 y 轴上产生（0,1）之间的均匀分布的随机数，水平投影到高斯累积分布函数上，然后垂直向下投影到 x 轴，得到的就是满足高斯分布的样本（随机数）

计算机是如何产生随机数的？

- 前面的例子展示了利用逆变换采样方法采样得到满足高斯分布的样本（随机数）的过程
- 虽然这种方法的思想很巧妙，但对于很多复杂的概率分布却无能为力
- （1）一些分布的CDF计算不出来（无法用公式表示），导致该方法无能为力
- （2）高维情形下很难获得PDF的表达式，只能得到变量之间的条件概率分布
  - 例如，不知道二维分布$p(x, y)$的具体表达式，但容易得到$p(x \mid y)$和$p(y \mid x)$

计算机是如何产生随机数的？

- 第二种方法：**接受-拒绝采样法**（Accept-reject sampling）
- 基本思想：假设 $p(x)$ 不可以直接抽样。找一个可以直接抽样的分布，称为建议分布/提议分布（proposal distribution）。假设 $q(x)$ 是建议分布的概率密度函数，并且有 $q(x)$ 的 c 倍一定大于等于 $p(x)$, 其中 $c > 0$ 。按照 $q(x)$ 进行抽样，假设得到结果是 x^*，再按照 $\frac{p(x^*)}{cq(x^*)}$ 的比例随机决定是否接受 x^*



- 接受-拒绝法实际是按照 $p(x)$ 的涵盖面积（或涵盖体积）占 $cq(x)$ 的涵盖面积（或涵盖体积）的比例进行抽样

计算机是如何产生随机数的？

- 接受-拒绝采样法的算法实现

输入：抽样的目标概率分布的概率密度函数 $p(x)$；

输出：概率分布的随机样本 $x_1, x_2, \cdots, x_n$。

参数：样本数 n

（1）选择概率密度函数为 $q(x)$ 的概率分布，作为建议分布，使其对任一 x 满足 $cq(x) \geqslant p(x)$，其中 $c > 0$。

（2）按照建议分布 $q(x)$ 随机抽样得到样本 x^*，再按照均匀分布在 $(0, 1)$ 范围内抽样得到 u。

（3）如果 $u \leqslant \dfrac{p(x^*)}{cq(x^*)}$，则将 x^* 作为抽样结果；否则，回到步骤 (2)。

（4）直至得到 n 个随机样本，结束。

计算机是如何产生随机数的？

- 接受-拒绝采样法的优点是容易实现，缺点是效率可能不高
- 关键在于找到提议分布 $q(x)$ 使得 $p(x)$ 与 $cq(x)$ 很相似
- 如果 $p(x)$ 的涵盖体积占 $cq(x)$ 的涵盖体积的比例很低，就会导致拒绝的比例很高，抽样效率很低
- 一般是在高维空间进行抽样，即使 $p(x)$ 与 $cq(x)$ 很接近，两者涵盖体积的差异也可能很大（与我们在三维空间的直观理解不同）

计算机是如何产生随机数的？

- 第三种方法：**重要性采样法**（Importance sampling），主要用于数学期望估计和积分计算
- 假设有随机变量 x，取值 $x \in \mathcal{X}$，其概率密度函数为 $p(x)$，$f(x)$ 为定义在 $\mathcal{X}$ 上的函数，目标是求函数 $f(x)$ 关于密度函数 $p(x)$ 的数学期望 $E_{p(x)}[f(x)]$
- 针对这个问题，可以按照概率分布 $p(x)$ 独立地抽取 n 个样本 $x_1, x_2, \cdots, x_n$

$$
\begin{aligned}
E_{p(x)}\left[f\left(x\right)\right] &= \int p\left(x\right) f\left(x\right) \mathrm{d}x \\
&\approx \frac{1}{n} \sum_{i=1}^{n} f\left(x_i\right) \\
&= \widehat{f}_n
\end{aligned}
$$

用样本均值 $\widehat{f}_n$ 作为数学期望 $E_{p(x)}[f(x)]$ 的近似值

计算机是如何产生随机数的？

- 理论依据：根据大数定律可知，当样本容量增大时，样本均值以概率 1 收敛于数学期望：

$$\hat{f}_n \to E_{p(x)}[f(x)], \quad n \to \infty$$

这样就得到了数学期望的近似计算方法

$$E_{p(x)}[f(x)] \approx \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

计算机是如何产生随机数的？

- 如果概率分布$p(x)$无法直接抽样，可以采用重要性采样法：找一个可以容易采样的提议分布$q(x)$，按照概率分布 $q(x)$ 独立地抽取 n 个样本 $x_1, x_2, \cdots, x_n$

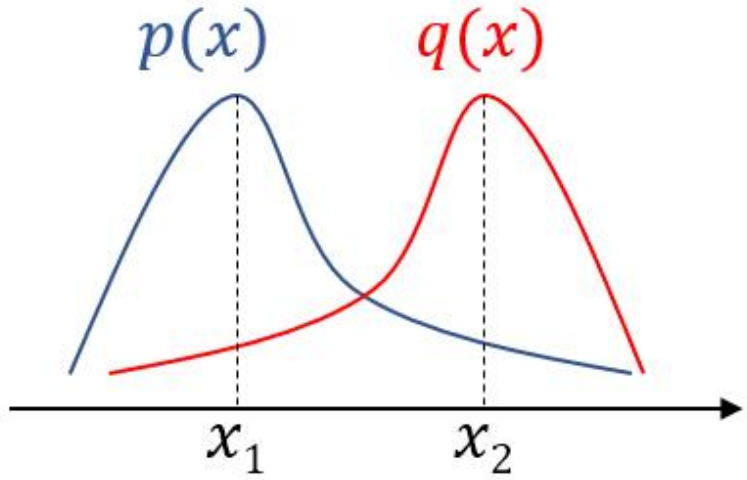
$$
\begin{aligned}
E_{p(x)}[f(x)] &= \int p(x) f(x) \mathrm{d}x \\
&= \int \frac{p(x)}{q(x)} q(x) f(x) \mathrm{d}x \\
&= \int \left(f(x) \frac{p(x)}{q(x)} \right) q(x) \mathrm{d}x \\
&\approx \frac{1}{n} \sum_{i=1}^{n} f(x_i) \frac{p(x_i)}{q(x_i)}
\end{aligned}
$$

其中$\frac{p(x_i)}{q(x_i)}$可视为样本x_i的权重(weight)/重要性(importance)

计算机是如何产生随机数的？

- 重要性采样法的实际效果严重依赖于两个概率分布$p(x)$和$q(x)$的相似程度
- 如果$p(x)$和$q(x)$差异很大，会导致采样得到的部分样本的权重过大，而另一部分样本的权重过小，最终的得到的近似结果主要受那些权重较大的样本的影响，导致计算结果不精确

计算机是如何产生随机数的？



- 采样是根据概率分布$q(x)$进行的，采样得到的大部分样本分布在高概率密度区域（如x_2），少部分分布在低概率密度区域（如x_1）
- 由于$p(x)$和$q(x)$差异很大，来自高概率密度区域的样本权重很小（如$\frac{p(x_2)}{q(x_2)}$），对计算结果的影响就很小，而来自低概率密度区域的样本权重很大（如$\frac{p(x_1)}{q(x_1)}$），对计算结果的影响就很大
- 造成的结果就是少量的大权重样本（如x_1）主要影响近似计算的结果，使得多次近似计算的结果区别很大，计算结果不稳定

计算机是如何产生随机数的？

- 该思想可以用于计算积分：假设有一个函数 $h(x)$，目标是计算该函数的积分

$$\begin{aligned}\int_{\mathcal{X}} h(x)\mathrm{d}x &= \int_{\mathcal{X}} \frac{h(x)}{p(x)} p(x)\,\mathrm{d}x \\ &= E_{p(x)}\left[\frac{h(x)}{p(x)}\right] \\ &\approx \frac{1}{n}\sum_{i=1}^{n} \frac{h(x_i)}{p(x_i)}\end{aligned}$$

$x_1, x_2, \cdots, x_n$是按照概率分布 $p(x)$ 独立抽取的 n 个样本

计算机是如何产生随机数的？

- **例1：** 使用重要性采样法计算定积分 $\int_0^1 e^{-x^2/2}\mathrm{d}x$
- **解：** 将 $h(x) = e^{-x^2/2}$ 视为

$$h(x) = f(x)p(x)$$

其中 $f(x) = e^{-x^2/2}$, $p(x) = 1$. 也就是说，假设随机变量 x 在 (0,1) 区间遵循均匀分布

- 在 (0,1) 区间按照均匀分布抽取 10 个随机样本 $x_1, x_2, \cdots, x_{10}$. 计算样本的函数均值 $\hat{f}_{10}$

$$\hat{f}_{10} = \frac{1}{10}\sum_{i=1}^{10} \mathrm{e}^{-x_i^2/2} = 0.832$$

也就是积分的近似. 随机样本数越大，计算就越精确

计算机是如何产生随机数的？

- **例2：**使用重要性采样法计算定积分 $\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) \mathrm{d}x$
- **解：**将 $h(x) = x \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$ 视为

$$h(x) = f(x)p(x)$$

其中 $f(x) = x$, 而 $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$ 是标准正态分布的密度函数

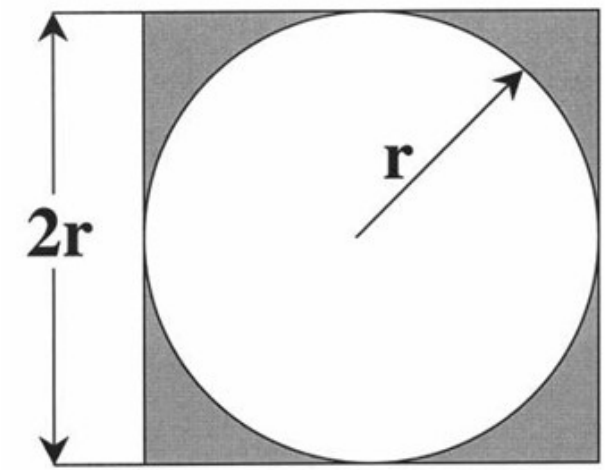
- 按照标准正态分布在区间 $(-\infty, \infty)$ 抽样 $x_1, x_2, \cdots, x_n$, 取其平均值，就得到要求的积分值. 当样本增大时, 积分值趋于 0

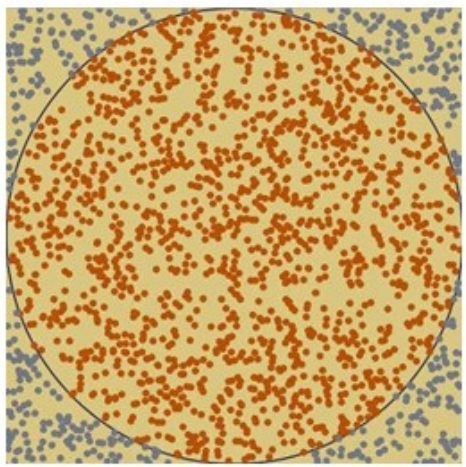
计算机是如何产生随机数的？

- 上述利用重要性采样法计算数学期望和积分的过程亦被称为蒙特卡罗方法（Monte Carlo method），或统计模拟方法（statistical simulation method），是一种通过从概率模型的随机抽样进行近似数值计算的方法
  - 该方法被广泛应用在物理、热力学、金融等领域的科学计算中
- 蒙特卡罗是摩纳哥的一个城市，以赌博闻名于世。蒙特卡罗方法作为一种计算方法，是由S.M.乌拉姆和J.冯诺依曼在20世纪40年代中叶为研制核武器的需要而提出的

π的计算

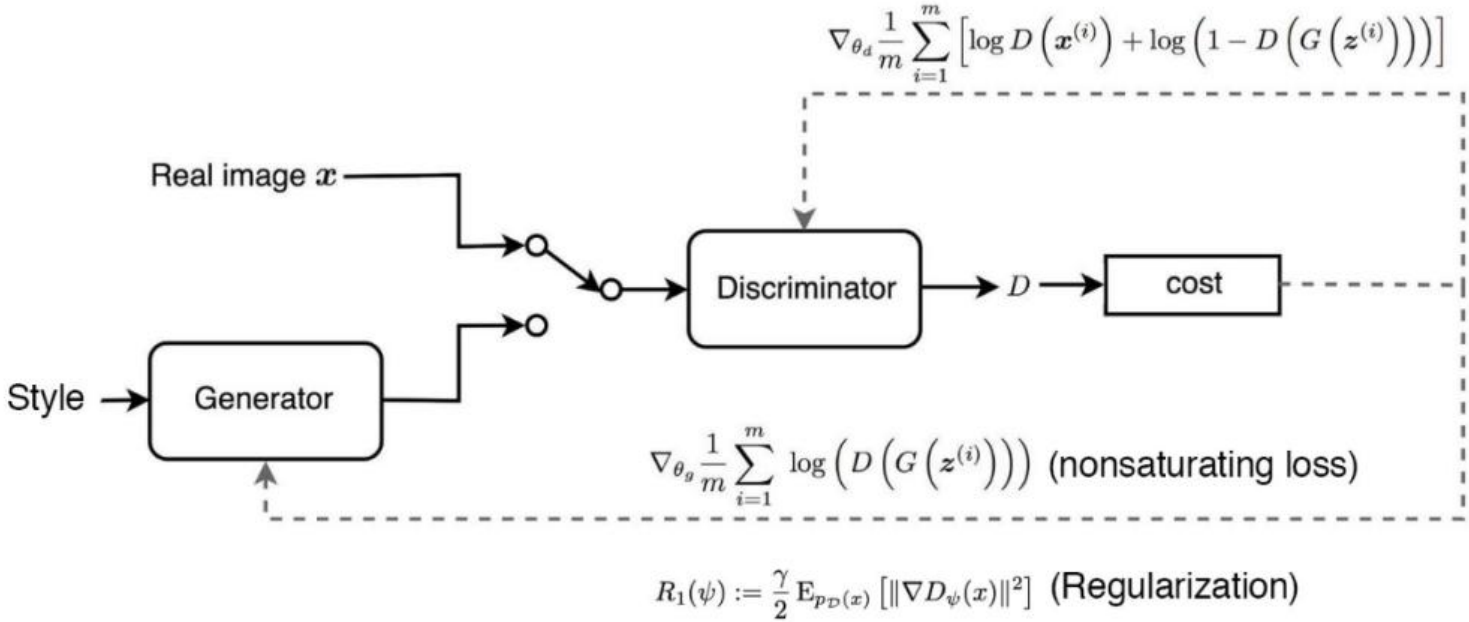
$$\frac{Area\ of\ Circle}{Area\ of\ Square} = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4}$$





小结一下

- **（1）机器学习中为什么要采样？**
- 动机1：采样本身就是常见的任务。例如在深度生成模型(如VAE, GAN)中，采样可以实现图片、语音、文本的自动生成

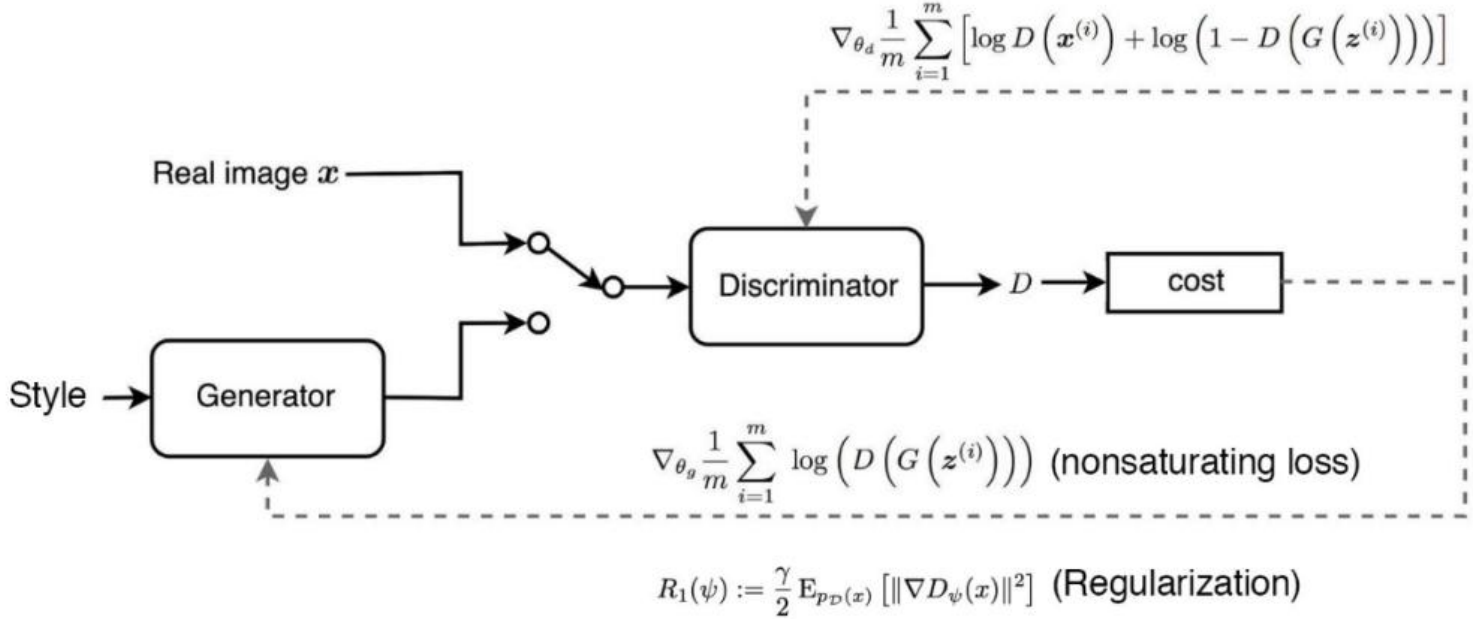


小结一下

小结一下

小结一下

- **（1）机器学习中为什么要采样？**
- 动机1：采样本身就是常见的任务。例如在深度生成模型(如VAE, GAN)中，采样可以实现图片、语音、文本的自动生成



- 动机2：求和或者求积分

小结一下

- **（2）什么才是好样本？**
- 1. 样本趋向于高概率密度区域
- 2. 样本之间相互独立
- 上述两个要素可以通过“选举”的例子来理解

小结一下

- **（3）为什么采样是困难的？**
- 1. PDF已知的情况下，CDF没有显示表达式，无法用逆变换采样法采样
- 2. 维度诅咒（curse of dimensionality）
  - 2.1. 离散分布：高维情形下的状态数呈指数级增长，无法枚举各个状态下的概率密度（PMF）
  - 2.2. 连续分布：接受-拒绝采样法和重要性采样法都需要找到和目标分布相似的提议分布，高维情形下这是非常困难的
- 3. PDF未知，仅知道变量之间的条件概率分布，无法计算PDF
  - 例如，不知道二维分布$p(x, y)$的具体表达式，仅知道$p(x \mid y)$和$p(y \mid x)$。因为$p(x, y) = p(x \mid y)\, p(y)$，所以要计算$p(x, y)$需要知道$p(y)$，而计算$p(y)$需要知道$p(x, y)$（因为$p(y) = \int p(x, y)\, \mathrm{d}\, x$），导致了“鸡生蛋”还是“蛋生鸡”的问题

马尔可夫链

- **随机过程：**考虑一个随机变量的序列

$$X = \{X_0, X_1, \cdots, X_t, \cdots\}$$

这里 X_t 表示时刻 t 的随机变量，$t = 0, 1, 2, \cdots$ 。每个随机变量 $X_t (t = 0, 1, 2, \cdots)$ 的取值集合相同，称为状态空间，表示为 $\mathcal{S}$。随机变量可以是离散的，也可以是连续的，以上随机变量的序列构成随机过程 (stochastic process)

马尔可夫链

- **马尔可夫链：**假设在时刻 0 的随机变量 X_0 遵循概率分布 $P(X_0) = \pi_0$, 称为初始状态分布。
  在某个时刻 $t \geqslant 1$ 的随机变量 X_t 与前一个时刻的随机变量 X_{t-1} 之间有条件分布 $P(X_t \mid X_{t-1})$, 如果 X_t 只依赖于 X_{t-1}, 而不依赖于过去的随机变量$\{X_0, X_1, \cdots, X_{t-2}\}$, 这一性质称为马尔可夫性, 即

$$P(X_t \mid X_0, X_1, \cdots, X_{t-1}) = P(X_t \mid X_{t-1}), \quad t = 1, 2, \cdots$$

  具有马尔可夫性的随机序列 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$ 称为马尔可夫链 (Markov chain), 或马尔可夫过程 (Markov process)。条件概率分布 $P(X_t \mid X_{t-1})$ 称为马尔可夫链的转移概率分布。转移概率分布决定了马尔可夫链的特性
- 马尔可夫性的直观解释是“未来只依赖于现在(假设现在已知)，而与过去无关”。这个假设在许多应用中是合理的

马尔可夫链

- 若转移概率分布 $P(X_t \mid X_{t-1})$ 与 t 无关，即

$$P(X_{t+s} \mid X_{t-1+s}) = P(X_t \mid X_{t-1}), \quad t = 1, 2, \cdots; \quad s = 1, 2, \cdots$$

则称该马尔可夫链为时间齐次的马尔可夫链(time homogenous Markov chain)

- 我们用到的都是时间齐次的马尔可夫链

离散状态马尔可夫链

- 离散状态马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$, 随机变量 $X_t (t = 0, 1, 2, \cdots)$ 定义在离散空间 $\mathcal{S}$
- 若马尔可夫链在时刻 $(t-1)$ 处于状态 j, 在时刻 t 移动到状态 i, 将转移概率记作

$$p_{ij} = (X_t = i \mid X_{t-1} = j), \quad i = 1, 2, \cdots; \quad j = 1, 2, \cdots$$

满足

$$p_{ij} \geqslant 0, \quad \sum_i p_{ij} = 1$$

离散状态马尔可夫链

- 马尔可夫链的转移概率 p_{ij} 可以由矩阵表示，即

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

称为马尔可夫链的转移概率矩阵, 转移概率矩阵 P 满足条件 $p_{ij} \geqslant 0, \sum_i p_{ij} = 1$

离散状态马尔可夫链

- 考虑马尔可夫链 $X=\{X_0, X_1, \cdots, X_t, \cdots\}$ 在时刻 $t(t=0,1,2,\cdots)$ 的概率分布，称为时刻 t 的状态分布，记作

$$\boldsymbol{\pi}(t)=\begin{bmatrix} \pi_1(t) \\ \pi_2(t) \\ \vdots \end{bmatrix}$$

其中 $\pi_i(t)$ 表示时刻 t 状态为 i 的概率 $P(X_t=i)$,

$$\pi_i(t)=P(X_t=i), \quad i=1,2,\cdots$$

离散状态马尔可夫链

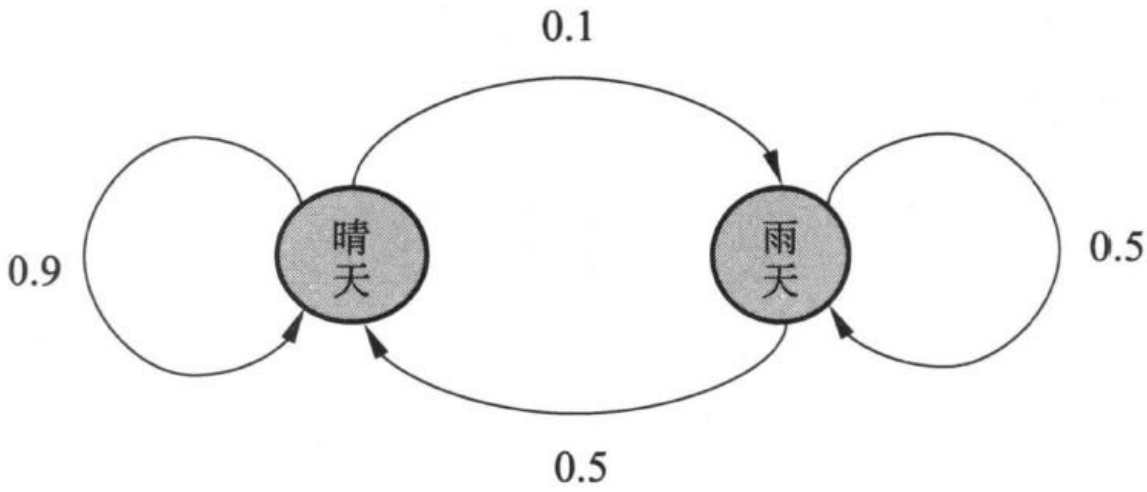
- 特别地，马尔可夫链的初始状态分布可以表示为

$$\boldsymbol{\pi}(0) = \begin{bmatrix} \pi_1(0) \\ \pi_2(0) \\ \vdots \end{bmatrix}$$

  其中 $\pi_i(0)$ 表示时刻 0 状态为 i 的概率 $P(X_0 = i)$
- 通常初始分布 $\boldsymbol{\pi}(0)$ 的向量只有一个分量是 1, 其余分量都是 0, 表示马尔可夫链从一个具体状态开始

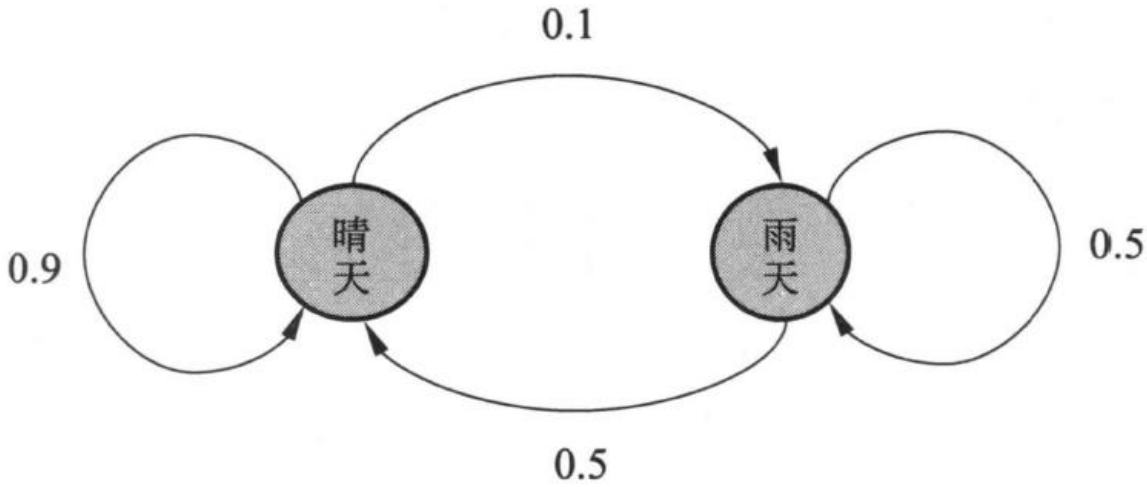
离散状态马尔可夫链

- 有限离散状态的马尔可夫链可以由有向图表示
- 结点表示状态，边表示状态之间的转移，边上的数值表示转移概率



- 从一个初始状态出发，根据有向边上定义的概率在状态之间随机跳转 (或随机转移)，就可以产生状态的序列
- 马尔可夫链实际上是刻画随时间在状态之间转移的模型，假设未来的转移状态只依赖于现在的状态，而与过去的状态无关

离散状态马尔可夫链



- 转移矩阵

$$\boldsymbol{P} = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{bmatrix}$$

离散状态马尔可夫链

- 如果第一天是晴天的话，其天气概率分布 (初始状态分布）如下:

$$\pi(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

- **问题：如何计算以后的天气概率分布（状态分布）?**

离散状态马尔可夫链

- 马尔可夫链 X 在时刻 t 的状态分布，可以由在时刻 $(t-1)$ 的状态分布以及转移概率分布决定

$$\boldsymbol{\pi}(t) = \boldsymbol{P}\boldsymbol{\pi}(t-1)$$

这是因为

$$\begin{aligned}
\pi_i(t) &= P\left(X_t = i\right) \\
&= \sum_m P\left(X_t = i \mid X_{t-1} = m\right) P\left(X_{t-1} = m\right) \\
&= \sum_m p_{im}\pi_m(t-1)
\end{aligned}$$

离散状态马尔可夫链

- 马尔可夫链在时刻 t 的状态分布，可以通过递推得到。事实上，由

$$\pi(t) = \boldsymbol{P}\pi(t-1) = \boldsymbol{P}(\boldsymbol{P}\pi(t-2)) = \boldsymbol{P}^2\pi(t-2)$$

递推得到

$$\pi(t) = \boldsymbol{P}^t\pi(0)$$

上式说明，马尔可夫链的状态分布由初始分布和转移概率分布决定

离散状态马尔可夫链

- 对于前面的天气例子，可以计算第二天、第三天及之后的天气概率分布（状态分布）

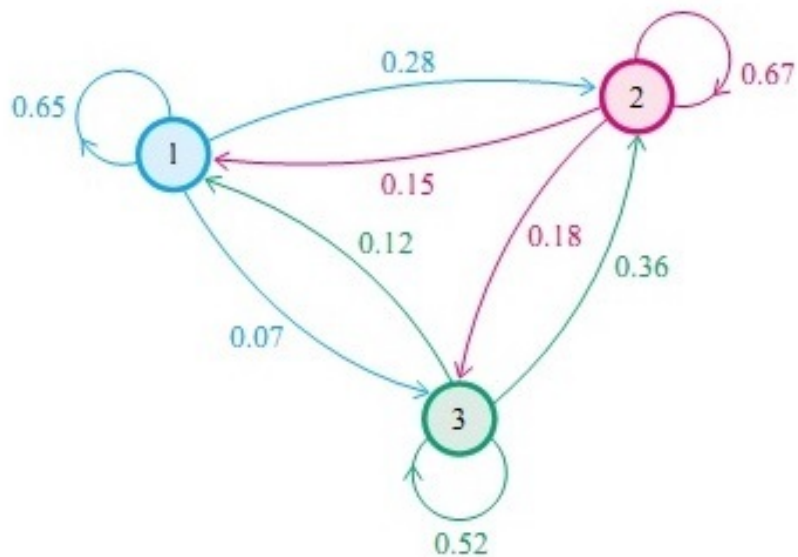
$$\pi(1) = \boldsymbol{P}\pi(0) = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix}$$

$$\pi(2) = \boldsymbol{P}^2\pi(0) = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{bmatrix}^2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.86 \\ 0.14 \end{bmatrix}$$

离散状态马尔可夫链

- **例：**社会学家经常把人按其经济状况分成3类：下层(lower-class)、中层(middle-class)、上层(upper-class)，分别用1、2、3代表这三个阶层。社会学家们发现决定一个人的收入阶层的最重要的因素就是其父母的收入阶层。如果一个人的收入属于下层类别，那么他的孩子属于下层收入的概率是0.65，属于中层收入的概率是0.28，属于上层收入的概率是0.07。事实上，从父代到子代，收入阶层的变化的转移概率如下

		父代		
	状态	1	2	3
子代	1	0.65	0.15	0.12
	2	0.28	0.67	0.36
	3	0.07	0.18	0.52



离散状态马尔可夫链

- 状态转移矩阵

$$\boldsymbol{P}=\begin{bmatrix}0.65 & 0.15 & 0.12\\ 0.28 & 0.67 & 0.36\\ 0.07 & 0.18 & 0.52\end{bmatrix}$$

- 假设当前这一代人处在下层、中层、上层的人的比例是概率分布向量$\boldsymbol{\pi}(0)=(\pi_1(0),\pi_2(0),\pi_3(0))^{\mathrm{T}}$，那么他们的子女的分布比例将是

$$\boldsymbol{\pi}(1)=\boldsymbol{P}\boldsymbol{\pi}(0),$$

他们的孙子代的分布比例将是

$$\boldsymbol{\pi}(2)=\boldsymbol{P}\boldsymbol{\pi}(1)=\boldsymbol{P}^2\boldsymbol{\pi}(0),$$

$$\cdots\cdots$$

第n代子孙的分布比例将是

$$\boldsymbol{\pi}(n)=\boldsymbol{P}\boldsymbol{\pi}(n-1)=\cdots=\boldsymbol{P}^n\boldsymbol{\pi}(0)$$

离散状态马尔可夫链

- 假设初始概率分布为$\pi(0) = (0.21, 0.68, 0.11)^T$，那么我们可以计算前$n$代人的分布状况如下

第n代人	下层	中层	上层
0	0.210	0.680	0.110
1	0.252	0.554	0.194
2	0.270	0.512	0.218
3	0.278	0.497	0.225
4	0.282	0.490	0.226
5	0.285	0.489	0.225
6	0.286	0.489	0.225
7	0.286	0.489	0.225
8	0.289	0.488	0.225
9	0.286	0.489	0.225
10	0.286	0.489	0.225
...	...	...	...

- 我们发现从第7代人开始，这个分布就稳定不变了，这是偶然现象吗？

离散状态马尔可夫链

- 我们换一个初始概率分布$\pi(0) = (0.75, 0.15, 0.10)^T$试试看，继续计算前$n$代人的分布状况如下

第n代人	下层	中层	上层
0	0.75	0.15	0.1
1	0.522	0.347	0.132
2	0.407	0.426	0.167
3	0.349	0.459	0.192
4	0.318	0.475	0.207
5	0.303	0.482	0.215
6	0.295	0.485	0.220
7	0.291	0.487	0.222
8	0.289	0.488	0.225
9	0.286	0.489	0.225
10	0.286	0.489	0.225
...	...	...	...

- 我们发现，到第9代人的时候，分布又收敛了

离散状态马尔可夫链

- 最为奇特的是，两次给定不同的初始概率分布，最终都收敛到概率分布

$$\boldsymbol{\pi} = (0.286, 0.489, 0.225)^{\mathrm{T}}$$

这说明收敛的行为和初始概率分布$\boldsymbol{\pi}(0)$无关。这说明这个收敛行为主要是由状态转移矩阵$\boldsymbol{P}$决定的

离散状态马尔可夫链

- **问题1：**对于任意的马尔可夫链，是否一定能收敛到同一个概率分布？
- **问题2：**当满足什么条件时，马尔可夫链才能收敛？

离散状态马尔可夫链

- **平稳分布：**设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，其状态空间为 $\mathcal{S}$，转移概率矩阵为 $\boldsymbol{P} = (p_{ij})$，如果存在状态空间 $\mathcal{S}$ 上的一个分布

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \end{bmatrix}$$

使得

$$\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\pi}$$

则称 $\boldsymbol{\pi}$ 为马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$ 的平稳分布

- 直观上，如果马尔可夫链的平稳分布存在，那么以该平稳分布作为初始分布，面向未来进行随机状态转移，之后任何一个时刻的状态分布都是该平稳分布

离散状态马尔可夫链

- **引理：**给定一个马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，状态空间为 $\mathcal{S}$，转移概率矩阵为 $\boldsymbol{P} = (p_{ij})$，则分布 $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 为 X 的平稳分布的充分必要条件是 $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 是下列方程组的解:

$$
\begin{aligned}
&x_i = \textstyle\sum_j p_{ij} x_j, \quad i = 1, 2, \cdots \\
&x_i \geqslant 0, \quad i = 1, 2, \cdots \\
&\textstyle\sum_i x_i = 1
\end{aligned}
$$

离散状态马尔可夫链

- **例：**考虑具有下列转移概率矩阵的马尔可夫链，求其平稳分布

$$\boldsymbol{P} = \begin{bmatrix} 1/2 & 1/2 & 1/4 \\ 1/4 & 0 & 1/4 \\ 1/4 & 1/2 & 1/2 \end{bmatrix}$$

离散状态马尔可夫链

- **解：**设平稳分布为 $\pi = (x_1, x_2, x_3)^{\mathrm{T}}$，根据上述引理有

$$
\begin{aligned}
x_1 &= \tfrac{1}{2}x_1 + \tfrac{1}{2}x_2 + \tfrac{1}{4}x_3 \\
x_2 &= \tfrac{1}{4}x_1 + \tfrac{1}{4}x_3 \\
x_3 &= \tfrac{1}{4}x_1 + \tfrac{1}{2}x_2 + \tfrac{1}{2}x_3 \\
x_1 &+ x_2 + x_3 = 1 \\
x_i &\geqslant 0, \quad i = 1, 2, 3
\end{aligned}
$$

解方程组，得到唯一的平稳分布

$$\pi = (2/5, 1/5, 2/5)^{\mathrm{T}}$$

离散状态马尔可夫链

- **例：**考虑具有下列转移概率矩阵的马尔可夫链，求其平稳分布

$$\boldsymbol{P} = \begin{bmatrix} 1 & 1/3 & 0 \\ 0 & 1/3 & 0 \\ 0 & 1/3 & 1 \end{bmatrix}$$

- **解：**这个马尔可夫链的平稳分布并不唯一，
$\boldsymbol{\pi} = (3/4, 0, 1/4)^{\mathrm{T}}$, $\boldsymbol{\pi} = (2/3, 0, 1/3)^{\mathrm{T}}$等皆为其平稳分布

- **马尔可夫链可能存在唯一平稳分布，无穷多个平稳分布，或不存在平稳分布**

连续状态马尔可夫链

- 连续状态马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，随机变量 $X_t (t = 0, 1, 2, \cdots)$ 定义在连续状态空间 $\mathcal{S}$，转移概率分布由概率转移核或转移核（transition kernel）表示
- 设 $\mathcal{S}$ 是连续状态空间，对任意的 $x \in \mathcal{S}, A \subset \mathcal{S}$, 转移核 $P(x, A)$ 定义为

$$P(x, A) = \int_A p(x, y) \mathrm{d}y$$

其中 $p(x, \cdot)$ 是概率密度函数，满足 $p(x, \cdot) \geqslant 0$, $P(x, \mathcal{S}) = \int_{\mathcal{S}} p(x, y) \mathrm{d}y = 1$。

- 转移核 $P(x, A)$ 表示从 $x \sim A$ 的转移概率

$$P\left(X_t = A \mid X_{t-1} = x\right) = P(x, A)$$

有时也将概率密度函数 $p(x, \cdot)$ 称为转移核

连续状态马尔可夫链

- 若马尔可夫链的状态空间 $\mathcal{S}$ 上的概率分布 $\pi(x)$ 满足条件

$$\pi(y) = \int p(x, y)\pi(x)\mathrm{d}x, \quad \forall y \in \mathcal{S}$$

则称分布 $\pi(x)$ 为该马尔可夫链的平稳分布。等价地，

$$\pi(A) = \int P(x, A)\pi(x)\mathrm{d}x, \quad \forall A \subset \mathcal{S}$$

或简写为

$$\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\pi}$$

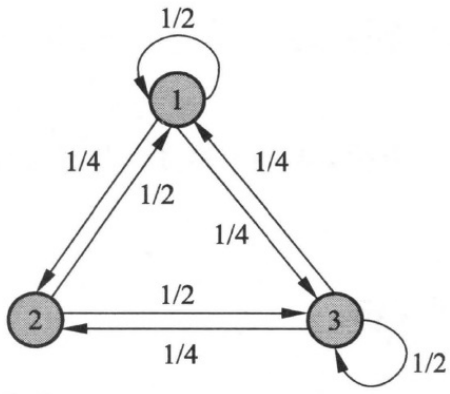
马尔可夫链的性质

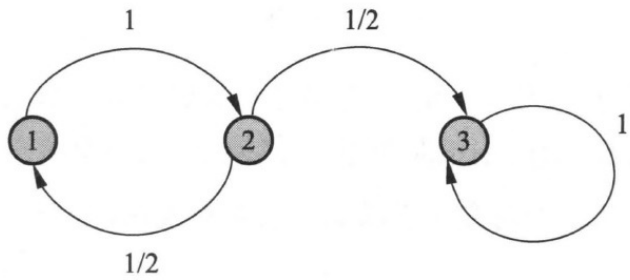
- 1.（**不可约**）设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$, 状态空间为$\mathcal{S}$, 对于任意状态 $i, j \in \mathcal{S}$, 如果存在一个时刻 $t(t > 0)$ 满足

$$P(X_t = i \mid X_0 = j) > 0$$

  也就是说，时刻 0 从状态 j 出发，时刻 t 到达状态 i 的概率大于 0, 则称此马尔可夫链 X 是不可约的 (irreducible), 否则称马尔可夫链是可约的 (reducible)
- 直观上，一个不可约的马尔可夫链，从任意状态出发，当经过充分长时间后，可以到达任意状态

马尔可夫链的性质



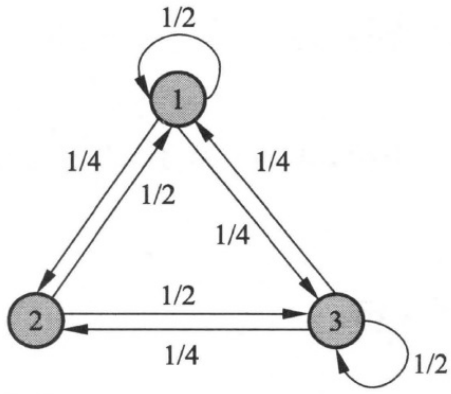


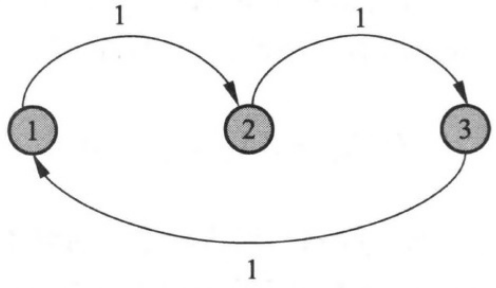
- 左图中的马尔可夫链是不可约的，右图中的马尔可夫链是可约的
- 右图中的马尔可夫链的平稳分布是 $\pi = (0 \quad 0 \quad 1)^{\mathrm{T}}$ 。此马尔可夫链，转移到状态 3 后，就在该状态上循环跳转，不能到达状态 1 和状态 2，最终停留在状态 3

马尔可夫链的性质

- 2.（**非周期**）设有马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$, 状态空间为 $\mathcal{S}$, 对于任意状态 $i \in \mathcal{S}$, 如果时刻 0 从状态 i 出发， t 时刻返回状态 i 的所有时间长 $\{t : P(X_t = i \mid X_0 = i) > 0\}$ 的最大公约数是 1, 则称此马尔可夫链 X 是非周期的 (aperiodic)，否则称马尔可夫链是周期的 (periodic)
- 直观上，一个非周期性的马尔可夫链, 不存在一个状态，从这一个状态出发, 再返回到这个状态时所经历的时间长呈一定的周期性

马尔可夫链的性质





- 左图中的马尔可夫链是非周期的，右图中的马尔可夫链是周期的
- 右图中的马尔可夫链的平稳分布是 $\boldsymbol{\pi} = (1/3 \quad 1/3 \quad 1/3)^{\mathrm{T}}$。此马尔可夫链从每个状态出发，返回该状态的时刻都是 3 的倍数，$\{3,6,9\}$，具有周期性，最终停留在每个状态的概率都为 $1/3$

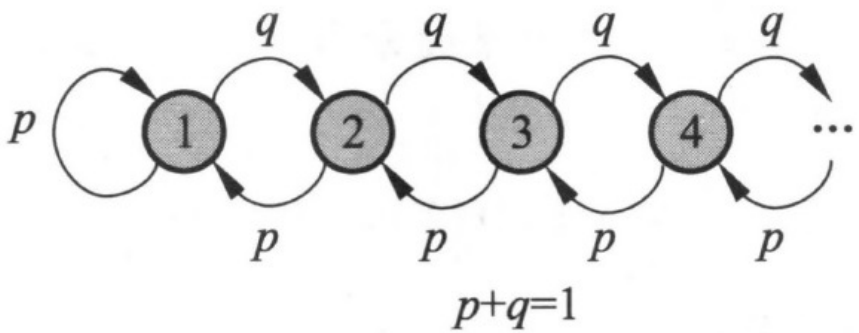
马尔可夫链的性质

- （**定理**）不可约且非周期的有限状态马尔可夫链，有唯一平稳分布存在

马尔可夫链的性质

- 3.（**正常返**）设有马尔可夫链 $X=\{X_0, X_1, \cdots, X_t, \cdots\}$, 状态空间为 $\mathcal{S}$, 对于任意状态 $i, j \in \mathcal{S}$, 定义概率 p_{ij}^t 为时刻 0 从状态 j 出发，时刻 t 首次转移到状态 i 的概率，即 $p_{ij}^t = P\left(X_t=i, X_s \neq i, s=1,2, \cdots, t-1 \mid X_0=j\right), t=1,2, \cdots$ 。若对所有状态 i, j 都满足 $\lim_{t\to\infty} p_{ij}^t > 0$, 则称马尔可夫链 X 是正常返的（positive recurrent)
- 直观上，一个正常返的马尔可夫链，其中任意一个状态，从其他任意一个状态出发，当时间趋于无穷时，首次转移到这个状态的概率不为 0

马尔可夫链的性质



转移概率矩阵

$$\begin{bmatrix} p & p & 0 & 0 & \\ q & 0 & p & 0 & \cdots \\ 0 & q & 0 & p & \\ 0 & 0 & q & 0 & \\ \vdots & & & & \ddots \end{bmatrix}$$

- 当 $p > q$ 时，平稳分布是

$$\pi_i = \left(\frac{q}{p}\right)^i \left(\frac{p-q}{p}\right), \quad i = 1, 2, \cdots$$

  当时间趋于无穷时，转移到任何一个状态的概率不为 0, 马尔可夫链是正常返的
- 当 $p \leqslant q$ 时, 不存在平稳分布，马尔可夫链不是正常返的

马尔可夫链的性质

- （**定理**）不可约、非周期且正常返的马尔可夫链，有唯一平稳分布存在

马尔可夫链的性质

- （**遍历定理**）设有马尔可夫链 $X=\{X_0, X_1, \cdots, X_t, \cdots\}$，状态空间为 $\mathcal{S}$, 若马尔可夫针 X 是不可约、非周期且正常返的，则该马尔可夫链有唯一平稳分布 $\boldsymbol{\pi}=(\pi_1, \pi_2, \cdots)^{\mathrm{T}}$，并且转移概率的极限分布是马尔可夫链的平稳分布

$$\lim_{t\to\infty} P(X_t = i \mid X_0 = j) = \pi_i, \quad i = 1, 2, \cdots; \quad j = 1, 2, \cdots$$

马尔可夫链的性质

- （接上页）
  若 $f(X)$ 是定义在状态空间上的函数，$E_\pi[|f(X)|] < \infty$, 则

$$P\left\{\hat{f}_t \to E_\pi[f(X)]\right\} = 1$$

这里

$$\hat{f}_t = \frac{1}{t}\sum_{s=1}^{t} f\left(x_s\right),$$

而$E_\pi[f(X)] = \sum_i f(i)\pi_i$ 是 $f(X)$ 关于平稳分布 $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots)^{\mathrm{T}}$ 的数学期望。等式$P\left\{\hat{f}_t \to E_\pi[f(X)]\right\} = 1$ 表示

$$\hat{f}_t \to E_\pi[f(X)], \quad t \to \infty$$

几乎处处成立或以概率 1 成立（无偏性）

马尔可夫链的性质

- 遍历定理的直观解释：满足相应条件（不可约、非周期、正常返）的马尔可夫链，当时间趋于无穷时，马尔可夫链的状态分布趋近于平稳分布，随机变量的函数的样本均值以概率 1 收敛于该函数的数学期望

- 遍历定理的三个条件: 不可约、非周期、正常返，保证了当时间趋于无穷时达到任意一个状态的概率不为 0

马尔可夫链的性质

- 理论上并不知道经过多少次迭代，马尔可夫链的状态分布才能接近于平稳分布，在实际应用遍历定理时，取一个足够大的整数 m ，经过 m 次迭代之后认为状态分布就是平稳分布，这时计算从第 $m+1$ 次迭代到第 n 次迭代的均值，即

$$\hat{E}f = \frac{1}{n-m}\sum_{i=m+1}^{n} f(x_i)$$

  称为遍历均值

- 这就是马尔可夫链蒙特卡罗法的计算目的

马尔可夫链的性质

- （**可逆马尔可夫链**）设有马尔可夫链 $X=\{X_0, X_1, \cdots, X_t, \cdots\}$, 状态空间为 $\mathcal{S}$, 转移概率矩阵为 P, 如果有状态分布 $\boldsymbol{\pi}=(\pi_1, \pi_2, \cdots)^{\mathrm{T}}$, 对于任意状态 $i, j \in \mathcal{S}$, 对任意一个时刻 t 满足

$$P\left(X_t=i \mid X_{t-1}=j\right) \pi_j=P\left(X_{t-1}=j \mid X_t=i\right) \pi_i, \quad i, j=1,2, \cdots$$

或简写为

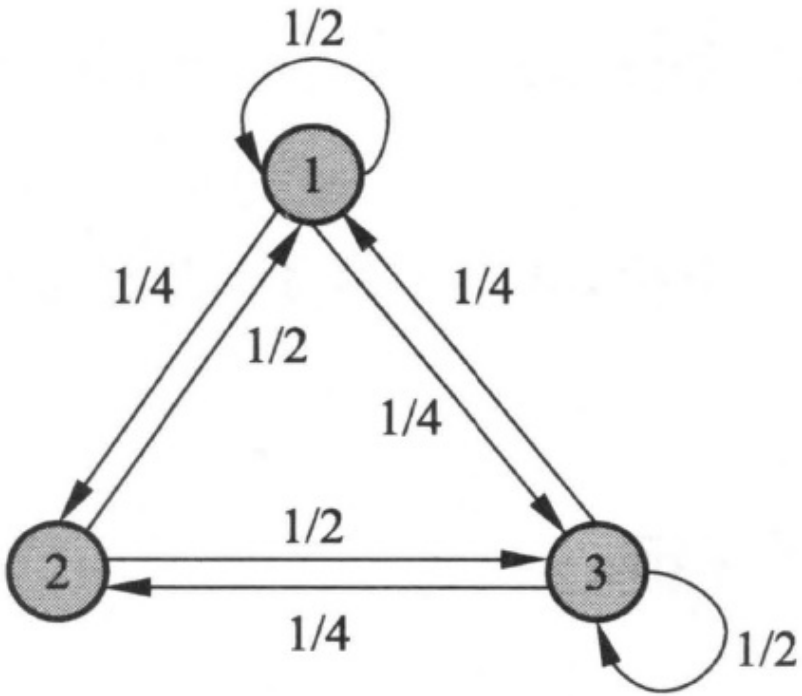
$$p_{ij} \pi_j=p_{ji} \pi_i, \quad i, j=1,2, \cdots$$

则称此马尔可夫链 X 为可逆马尔可夫链 (reversible Markov chain)，上式称为细致平衡方程 (detailed balance equation) /细致平稳条件

马尔可夫链的性质

- 直观上，如果有可逆的马尔可夫链，那么以该马尔可夫链的平稳分布作为初始分布，进行随机状态转移，无论是面向未来还是面向过去，任何一个时刻的状态分布都是该平稳分布

马尔可夫链的性质



转移概率矩阵

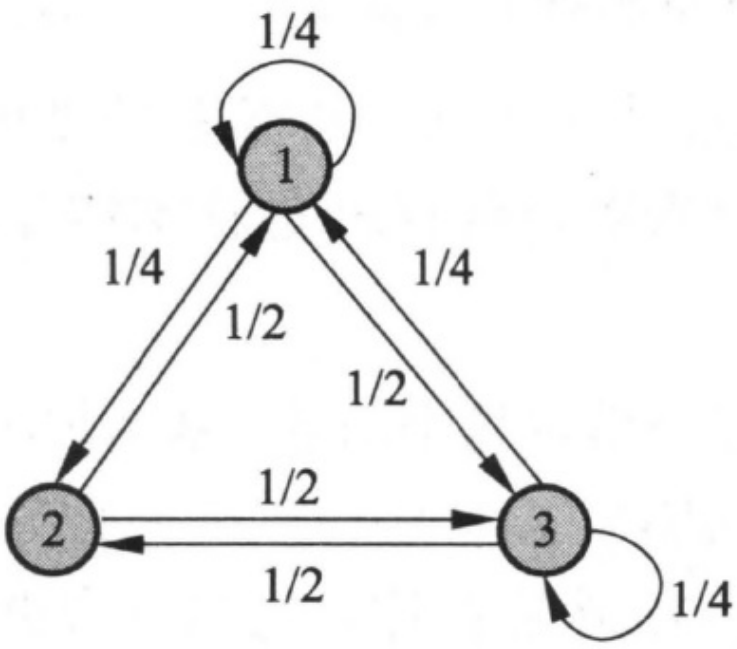
$$P = \begin{bmatrix} 1/2 & 1/2 & 1/4 \\ 1/4 & 0 & 1/4 \\ 1/4 & 1/2 & 1/2 \end{bmatrix}$$

平稳分布

$$\pi = (2/5 \quad 1/5 \quad 2/5)^{\mathrm{T}}$$

- 该马尔可夫链是可逆的

马尔可夫链的性质



转移概率矩阵

$$\begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 1/4 & 0 & 1/2 \\ 1/2 & 1/2 & 1/4 \end{bmatrix}$$

平稳分布

$$\pi = (8/25 \quad 7/25 \quad 2/5)^{\mathrm{T}}$$

- 该马尔可夫链是不可逆的

马尔可夫链的性质

- （**定理**）满足细致平衡方程的状态分布 π 就是该马尔可夫链的平稳分布。即

$$\boldsymbol{P\pi = \pi}$$

- （**证明**）事实上

$$(P\pi)_i = \sum_j p_{ij}\pi_j = \sum_j p_{ji}\pi_i = \pi_i \sum_j p_{ji} = \pi_i, \quad i = 1, 2, \cdots$$

马尔可夫链的性质

- 上述定理说明，可逆马尔可夫链（满足细致平衡方程）一定有唯一平稳分布，给出了一个马尔可夫链有平稳分布的充分条件（不是必要条件）
- 也就是说，可逆马尔可夫链满足遍历定理的条件
- 这就是马尔可夫链蒙特卡罗法的理论依据

马尔可夫链蒙特卡罗法

- 假设目标是对一个概率分布进行随机抽样，或者是求函数关于该概率分布的数学期望
  - 可以采用传统的蒙特卡罗法，如接受-拒绝法、重要性抽样法，也可以使用马尔可夫链蒙特卡罗法 (Markov Chain Monte Carlo, MCMC)
- MCMC更适合于随机变量是多元的、密度函数是非标准形式的、随机变量各分量不独立等情况

马尔可夫链蒙特卡罗法

- 假设多元随机变量 x, 满足 $x \in \mathcal{X}$, 其概率密度函数为 $p(x), f(x)$ 为定义在 $x \in \mathcal{X}$ 上的函数
- 目标是获得概率分布 $p(x)$ 的样本集合，以及求函数 $f(x)$ 的数学期望 $E_{p(x)}[f(x)]$
- **为什么要采样？为什么要计算数学期望？**

马尔可夫链蒙特卡罗法

- MCMC在统计学习，特别是贝叶斯学习中，起着重要的作用。主要是因为MCMC可以用在概率模型的学习和推理上
- 假设观测数据由随机变量 $y \in \mathcal{Y}$ 表示，模型由随机变量 $x \in \mathcal{X}$ 表示，贝叶斯学习通过贝叶斯定理计算给定数据条件下模型的后验概率，并选择后验概率最大的模型
- 后验概率

$$p(x \mid y) = \frac{p(x)p(y \mid x)}{\int_{\mathcal{X}} p\left(x'\right) p\left(y \mid x'\right) \mathrm{d}x'}$$

马尔可夫链蒙特卡罗法

- **MCMC与统计学习**
- 贝叶斯学习中经常需要进行三种积分运算: 归范化（normalization）、边缘化（marginalization）、数学期望（expectation）
- 后验概率计算中需要归范化计算:

$$\int_{\mathcal{X}} p\left(x'\right) p\left(y \mid x'\right) \mathrm{d}x'$$

- 如果有隐变量 $z \in \mathcal{Z}$, 后验概率的计算需要边缘化计算:

$$p(x \mid y) = \int_{\mathcal{Z}} p(x, z \mid y) \mathrm{d}z$$

马尔可夫链蒙特卡罗法

- 如果有一个函数 $f(x)$, 可以计算该函数的关于后验概率分布的数学期望:

$$E_{P(x|y)}[f(x)] = \int_{\mathcal{X}} f(x)p(x \mid y)\mathrm{d}x$$

- 当观测数据和模型都很复杂的时候，以上的积分计算变得困难。MCMC为这些计算提供了一个通用的有效解决方案

马尔可夫链蒙特卡罗法

- **基本想法：**
- 在随机变量 x 的状态空间 $\mathcal{S}$ 上定义一个满足遍历定理的马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，使其平稳分布就是抽样的目标分布 $p(x)$
- 在这个马尔可夫链上进行随机游走，每个时刻得到一个样本
- 根据遍历定理，当时间趋于无穷时，样本的分布趋近平稳分布，样本的函数均值趋近函数的数学期望
- 所以，当时间足够长时 (时刻大于某个正整数 m)，在之后的时间 (时刻小于等于某个正整数 $n, n > m$) 里随机游走得到的样本集合$\{x_{m+1}, x_{m+2}, \cdots, x_n\}$ 就是目标概率分布的抽样结果，得到的函数均值（遍历均值）就是要计算的数学期望值

$$\hat{E}f = \frac{1}{n-m} \sum_{i=m+1}^{n} f(x_i)$$

到时刻 m 为止的时间段称为燃烧期

马尔可夫链蒙特卡罗法

- **基本步骤：**
- （1）在随机变量 x 的状态空间 $\mathcal{S}$ 上构造一个满足遍历定理的马尔可夫链，使其平稳分布为目标分布 $p(x)$
- （2）从状态空间的某一点 x_0 出发，用构造的马尔可夫链进行随机游走，产生样本序列 $x_0, x_1, \cdots, x_t, \cdots$
- （3）应用马尔可夫链的遍历定理，确定正整数 m 和 n $(m < n)$, 得到样本集合 $\{x_{m+1}, x_{m+2}, \cdots, x_n\}$, 求得函数 $f(x)$ 的均值（遍历均值）

$$\hat{E}f = \frac{1}{n-m} \sum_{i=m+1}^{n} f(x_i)$$

就是MCMC的计算公式

马尔可夫链蒙特卡罗法

- 这里有几个重要问题:
- （1）如何定义马尔可夫链，保证MCMC的条件成立？
- （2）如何确定收敛步数 m, 保证样本抽样的无偏性？
- （3）如何确定迭代步数 n, 保证遍历均值计算的精度？

马尔可夫链蒙特卡罗法

- **问题1：如何定义马尔可夫链，保证MCMC的条件成立？**
- 连续变量的时候，需要定义转移核函数; 离散变量的时候，需要定义转移矩阵。一个方法是定义特殊的转移核函数或者转移矩阵，构建可逆马尔可夫链，这样可以保证遍历定理成立
- 常用的MCMC方法包括Metropolis-Hastings 算法、吉布斯抽样，分别采用不同的转移核函数或者转移矩阵
- 由于这个马尔可夫链满足遍历定理，随机游走的起始点并不影响得到的结果，即从不同的起始点出发，都会收敛到同一平稳分布

马尔可夫链蒙特卡罗法

- **问题2：如何确定收敛步数 m，保证样本抽样的无偏性？**
- MCMC的收敛性的判断通常是经验性的，比如，在马尔可夫链上进行随机游走，检验遍历均值是否收敛
- 具体地，每隔一段时间取一次样本，得到多个样本以后，计算遍历均值，当计算的均值稳定后，认为马尔可夫链已经收敛
- 再比如，在马尔可夫链上并行进行多个随机游走，比较各个随机游走的遍历均值是否接近一致

马尔可夫链蒙特卡罗法

- **问题3：如何确定迭代步数 n, 保证遍历均值计算的精度？**
- 可以在判断马尔可夫链是否收敛的同时判断遍历均值计算的精度
- 通常取10000个样本

Metropolis-Hastings算法

- Metropolis-Hastings (MH) 算法是MCMC方法中最具代表性的算法
- **基本想法：**假设要抽样的概率分布为 $p(x)$。在随机变量 x 的状态空间 $\mathcal{S}$ 上定义一个满足遍历定理的马尔可夫链 $X = \{X_0, X_1, \cdots, X_t, \cdots\}$，使其平稳分布就是抽样的目标分布 $p(x)$
- 为此，我们希望构建一个可逆的马尔可夫链，使得细致平衡方程成立

$$p\left(x\right) q\left(x, x'\right) = p\left(x'\right) q\left(x', x\right)$$

$q\left(x, x'\right)$ 是马尔可夫链的转移核，这样的话 $p(x)$ 就是该可逆的马尔可夫链的平稳分布，从而遍历定理成立

Metropolis-Hastings算法

$$p(x)\, q(x, x') = p(x')\, q(x', x)$$

- 常见的容易抽样的分布（如正态分布、t分布）通常无法满足上述细致平衡方程
- **问题：能否对转移核 $q(x, x')$ 进行改造，使得细致平衡方程成立呢？**

Metropolis-Hastings算法

- 为此，我们可以分别引入 $\alpha(x, x')$ 和 $\alpha(x', x)$，使得细致平衡方程成立

$$p(x)\, q(x, x')\, \alpha(x, x') = p(x')\, q(x', x)\, \alpha(x', x)$$

- $\alpha(x, x')$ 和 $\alpha(x', x)$ 分别取什么值呢？最简单的，按照对称性，我们可以取

$$\begin{aligned} \alpha(x, x') &= p(x')\, q(x', x) \\ \alpha(x', x) &= p(x)\, q(x, x') \end{aligned}$$

这样一来，我们就有

$$p(x)\, q(x, x') \underbrace{p(x')\, q(x', x)}_{\alpha(x, x')} = p(x')\, p(x) \underbrace{p(x)\, q(x, x')}_{\alpha(x', x)}$$

成立

Metropolis-Hastings算法

$$p(x)\underbrace{q(x,x')\alpha(x,x')}_{p(x,x')} = p(x')\underbrace{q(x',x)\alpha(x',x)}_{p(x',x)}$$

- 如果将

$$p(x,x') = q(x,x')\alpha(x,x')$$

和

$$p(x',x) = q(x',x)\alpha(x',x)$$

看成是新的转移核，那么我们就构造了一个满足细致平衡方程的可逆马尔可夫链，其平稳分布就是抽样的目标分布$p(x)$:

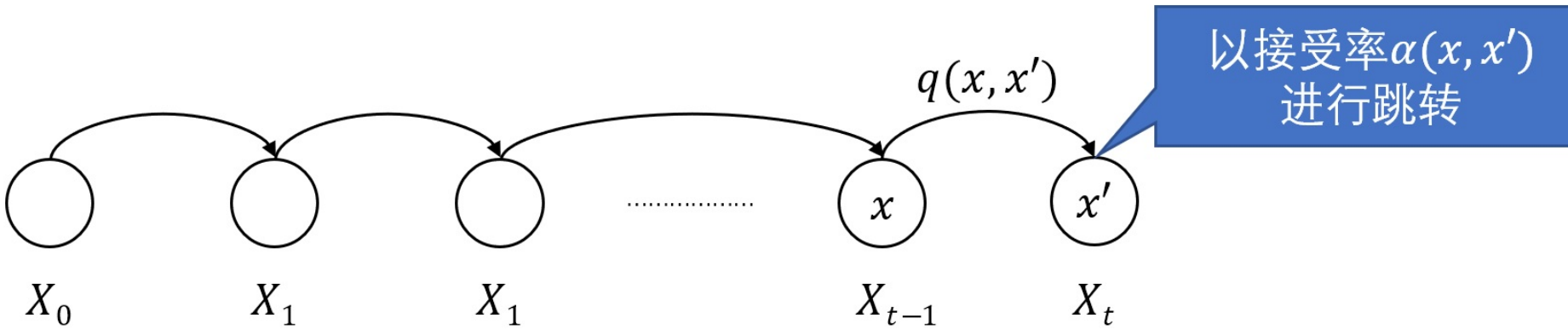
$$p(x)p(x,x') = p(x')p(x',x)$$

Metropolis-Hastings算法

- 如何理解新的马尔可夫链的转移核?

$$p\left(x, x'\right) = q\left(x, x'\right)\alpha\left(x, x'\right)$$

- 引入的 $\alpha\left(x, x'\right)$ 称为接受率，转移核 $p\left(x, x'\right)$ 可以看成是在原来的马尔可夫链上，从状态x以概率$q\left(x, x'\right)$跳转到状态x'时，我们以 $\alpha\left(x, x'\right)$ 的概率接受这个转移，所以新的马尔可夫链的转移核为 $p\left(x, x'\right) = q\left(x, x'\right)\alpha\left(x, x'\right)$



Metropolis-Hastings算法

- **问题：如何实现接受率$\alpha(x, x')$?**
- 从[0,1]区间按均匀分布随机抽取一个数 u
  如果 $u \leqslant \alpha(x, x')$，则接受新状态 x';
  否则停留在原状态 x

Metropolis-Hastings算法

- 引入的接受率 $\alpha(x, x')$ 和 $\alpha(x', x)$ 可能偏小，这会导致马尔可夫链在原地踏步，拒绝大量的跳转，从而花费更长的时间收敛到平稳分布 $p(x)$
- 有没有办法提高接受率呢？
- 假设 $\alpha(x, x') = 0.1$, $\alpha(x', x) = 0.2$, 此时满足细致平衡方程

$$p(x)\, q(x, x') \times 0.1 = p(x')\, q(x', x) \times 0.2$$

上式两边同时扩大5倍，改写为

$$p(x)\, q(x, x') \times 0.5 = p(x')\, q(x', x) \times 1$$

上式仍满足细致平衡方程

- 这启发我们可以把接受率 $\alpha(x, x')$ 和 $\alpha(x', x)$ 同比例放大，使得两者中最大的一个变为1，从而提高了接受率

Metropolis-Hastings算法

- 因此，常用的接受率定义为

$$\alpha\left(x, x'\right) = \min\left\{1, \frac{p\left(x'\right) q\left(x', x\right)}{p(x) q\left(x, x'\right)}\right\}$$

$$\alpha\left(x', x\right) = \min\left\{1, \frac{p\left(x\right) q\left(x, x'\right)}{p(x') q\left(x', x\right)}\right\}$$

Metropolis-Hastings算法

- **基本原理：**假设要抽样的概率分布为 $p(x)$。MH算法采用转移核为 $p\left(x, x'\right)$ 的马尔可夫链：

$$p\left(x, x'\right) = q\left(x, x'\right)\alpha\left(x, x'\right)$$

  - 建议分布 $q\left(x, x'\right)$ 是另一个马尔可夫链的转移核，并且 $q\left(x, x'\right)$ 是不可约的，即其概率值恒不为 0，同时是一个容易抽样的分布
  - 接受分布 $\alpha\left(x, x'\right)$ 是

$$\alpha\left(x, x'\right) = \min\left\{1, \frac{p\left(x'\right)q\left(x', x\right)}{p(x)q\left(x, x'\right)}\right\}$$

Metropolis-Hastings算法

- 转移核为 $p(x, x')$ 的马尔可夫链上的随机游走以以下方式进行：
- 如果在时刻$(t-1)$ 处于状态 x, 即 $x_{t-1}=x$, 则先按建议分布 $q(x, x')$ 抽样产生一个候选状态 x', 然后按照接受分布 $\alpha(x, x')$ 抽样决定是否接受状态 x'
  - 以概率 $\alpha(x, x')$ 接受 x', 决定时刻 t 转移到状态 x'
  - 而以概率 $1-\alpha(x, x')$ 拒绝 x', 决定时刻 t 仍停留在状态 x
- 具体地，从区间 (0,1) 上的均匀分布中抽取一个随机数 u, 决定时刻 t 的状态

$$x_t=\begin{cases} x', & u \leqslant \alpha(x, x') \\ x, & u>\alpha(x, x') \end{cases}$$

Metropolis-Hastings算法

- **定理：**由转移核

$$
\begin{aligned}
p\left(x, x'\right) &= q\left(x, x'\right) \alpha\left(x, x'\right) \\
&= q\left(x, x'\right) \min\left\{1, \frac{p\left(x'\right) q\left(x', x\right)}{p\left(x\right) q\left(x, x'\right)}\right\}
\end{aligned}
$$

构成的马尔可夫链是可逆的，即

$$
p(x) p\left(x, x'\right) = p\left(x'\right) p\left(x', x\right)
$$

并且 $p(x)$ 是该马尔可夫链的平稳分布

- 上述定理说明，转移核为 $p\left(x, x'\right)$ 的马尔可夫链是可逆马尔可夫链（满足遍历定理），其平稳分布就是 $p(x)$ ，即要抽样的目标分布

Metropolis-Hastings算法

$$
\begin{aligned}
p\left(x, x'\right) &= q\left(x, x'\right) \alpha\left(x, x'\right) \\
&= q\left(x, x'\right) \min\left\{1, \frac{p\left(x'\right) q\left(x', x\right)}{p\left(x\right) q\left(x, x'\right)}\right\}
\end{aligned}
$$

- **证明：** 若 $x = x'$，则显然 $p(x)p\left(x, x'\right) = p\left(x'\right) p\left(x', x\right)$ 成立
  若 $x \neq x'$，则

$$
\begin{aligned}
p(x)p\left(x, x'\right) &= p(x)q\left(x, x'\right) \min\left\{1, \frac{p\left(x'\right) q\left(x', x\right)}{p(x)q\left(x, x'\right)}\right\} \\
&= \min\left\{p(x)q\left(x, x'\right), p\left(x'\right) q\left(x', x\right)\right\} \\
&= p\left(x'\right) q\left(x', x\right) \min\left\{\frac{p(x)q\left(x, x'\right)}{p\left(x'\right) q\left(x', x\right)}, 1\right\} \\
&= p\left(x'\right) p\left(x', x\right)
\end{aligned}
$$

Metropolis-Hastings算法

- 进一步，由 $p(x)p\left(x,x'\right)=p\left(x'\right)p\left(x',x\right)$ 可知，

$$\begin{aligned}\int p(x)p\left(x,x'\right)\mathrm{d}x &= \int p\left(x'\right)p\left(x',x\right)\mathrm{d}x \\ &= p\left(x'\right)\int p\left(x',x\right)\mathrm{d}x \\ &= p\left(x'\right)\end{aligned}$$

上式说明，$p(x)$ 是马尔可夫链的平稳分布

Metropolis-Hastings算法

- **建议分布：**
- 建议分布 $q(x, x')$ 有多种可能形式，这里介绍两种常用形式

- 第一种形式，假设建议分布是对称的，即对任意的 x 和 x' 有

$$q(x, x') = q(x', x)$$

这样的建议分布称为 Metropolis 选择，也是 Metropolis-Hastings 算法最初采用的建议分布。这时，接受分布 $\alpha(x, x')$ 简化为

$$\alpha(x, x') = \min\left\{1, \frac{p(x')}{p(x)}\right\}$$

Metropolis-Hastings算法

- Metropolis 选择的一个特例是 $q(x, x')$ 取条件概率分布 $p(x' \mid x)$, 定义为多元正态分布，其均值是 x ，其协方差矩阵是常数矩阵
- Metropolis 选择的另一个特例是令 $q(x, x') = q(|x - x'|)$, 这时算法称为随机游走 Metropolis 算法。例如,

$$q(x, x') \propto \exp\left(-\frac{(x' - x)^2}{2}\right)$$

- Metropolis 选择的特点是当 x' 与 x 接近时，$q(x, x')$ 的概率值高，否则 $q(x, x')$ 的概率值低。状态转移在附近点的可能性更大

Metropolis-Hastings算法

- 第二种形式称为独立抽样。假设 $q(x, x')$ 与当前状态 x 无关，即 $q(x, x') = q(x')$ 。建议分布的计算按照 $q(x')$ 独立抽样进行。此时，接受分布 $\alpha(x, x')$ 可以写成

$$\alpha(x, x') = \min\left\{1, \frac{p(x')\, q(x)}{p(x)\, q(x')}\right\}$$

- 独立抽样实现简单, 但可能收敛速度慢, 通常选择接近目标分布 $p(x)$ 的分布作为建议分布 $q(x)$

Metropolis-Hastings算法

算法（Metropolis-Hastings 算法）

输入：抽样的目标分布的密度函数 $p(x)$，函数 $f(x)$；

输出：$p(x)$ 的随机样本 $x_{m+1}, x_{m+2}, \cdots, x_n$，函数样本均值 f_{mn}；

参数：收敛步数 m，迭代步数 n。

（1）任意选择一个初始值 x_0

（2）对 $i = 1, 2, \cdots, n$ 循环执行

（a）设状态 $x_{i-1} = x$，按照建议分布 $q(x, x')$ 随机抽取一个候选状态 x'。

（b）计算接受概率

$$\alpha(x, x') = \min\left\{1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\right\}$$

（c）从区间 $(0, 1)$ 中按均匀分布随机抽取一个数 u。

若 $u \leqslant \alpha(x, x')$，则状态 $x_i = x'$；否则，状态 $x_i = x$。

（3）得到样本集合 $\{x_{m+1}, x_{m+2}, \cdots, x_n\}$

计算

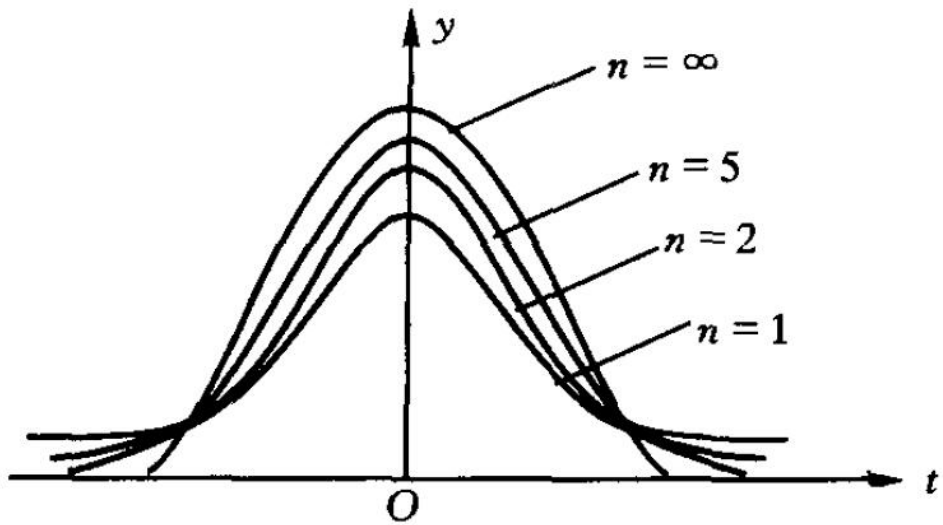
$$f_{mn} = \frac{1}{n-m}\sum_{i=m+1}^{n} f(x_i)$$

Metropolis-Hastings算法

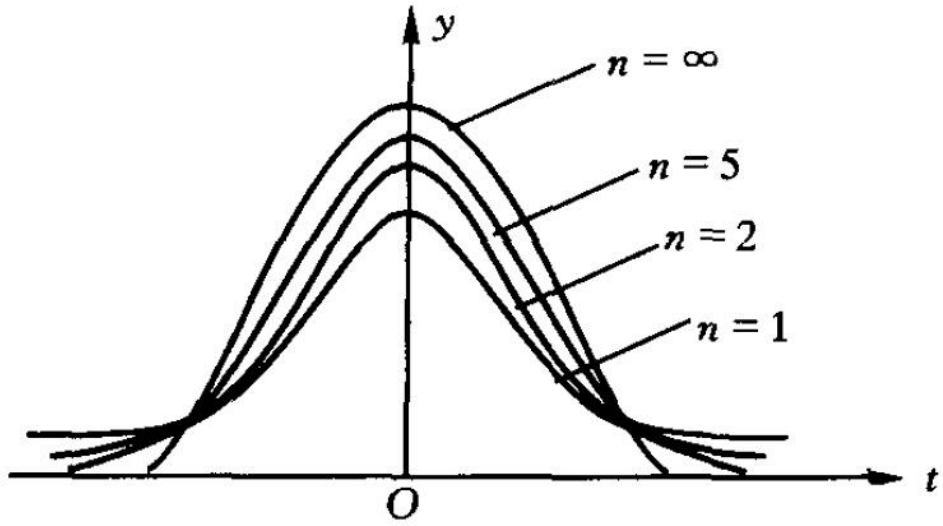
- **实例1：**使用MH算法对目标分布t分布进行随机采样
  如果随机变量 T 具有概率密度

$$t(x;n)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}\quad(-\infty<x<+\infty)$$

则称 T 服从自由度为 n 的 t 分布，记为 $T\sim t(n)$。t 分布又称为学生(student)分布



Metropolis-Hastings算法



- 分布的概率密度 $t(x;n)$ 关于 $x=0$ 对称, 并且当 $|x| \to +\infty$ 时单调下降地趋于 0
- 可以证明 $\lim_{n\to\infty}(x;n) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$, 即当自由度 $n \to +\infty$ 时, 自由度为 n 的 t 分布收敛于标准正态分布 $N(0,1)$
- 一般来说, 当 $n > 30$ 时, t分布与标准正态分布就非常接近了, 但对较小的 n 值, t分布与正态分布之间有较大差异

Metropolis-Hastings算法

- **实例2：**掷硬币试验，掷出 n 次，设随机变量 X 表示正面向上的次数，因此随机变量 X 服从二项分布 $B(n, \theta)$，θ 是硬币正面向上的概率，概率分布如下

$$p(X = x \mid \theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n$$

其中 x 表示观测到正面向上的次数。使用贝叶斯参数估计，假设参数 θ 的先验分布为参数 $a = 1, b = 1$ 的Beta分布。使用MH算法对参数 θ 进行后验估计

Metropolis-Hastings算法

- 参数 θ 的先验分布: 参数 $a = 1, b = 1$ 的Beta分布

$$\text{Beta}(\theta \mid a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}$$

- 似然函数: 观测到正面向上的次数为 x

$$p(X = x \mid \theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x}$$

- 参数 θ 的后验分布: 参数为 $a = 1 + x$ 和 $b = 1 + n - x$ 的Beta分布

Metropolis-Hastings算法

- **问题：**假设我们不知道Beta分布是二项分布的共轭先验，能否直接使用MH算法对参数 θ 进行后验估计呢？

Metropolis-Hastings算法

- 应用贝叶斯公式，得到

$$
\begin{aligned}
p(\theta|x) &= \frac{p(\theta)\,p(x|\theta)}{p(x)} \\
&= \frac{p(\theta)\,p(x|\theta)}{\int p(\theta)\,p(x|\theta)\,\mathrm{d}\theta} \\
&\propto p(\theta)\,p(x|\theta)
\end{aligned}
$$

对于不同的 θ，归一化因子 $p(x)$ 是共有的，所以不同 θ 的后验概率密度由先验概率密度 $p(\theta)$ 和似然函数 $p(x|\theta)$ 决定

Metropolis-Hastings算法

- **使用MH算法从参数 θ 的后验分布中进行随机采样：**
- 目标分布 $p(\theta|x)$
- 使用独立抽样作为建议分布

$$\theta' \sim \text{Beta}(a = 1, b = 1)$$

- 接受分布

$$\begin{aligned}
\alpha(\theta, \theta') &= \min\left\{1, \frac{p(\theta'|x)\, q(\theta)}{p(\theta|x)\, q(\theta')}\right\} \\
&= \min\left\{1, \frac{p(\theta'|x)}{p(\theta|x)}\right\} \\
&= \min\left\{1, \frac{p(\theta')\, p(x|\theta')}{p(\theta)\, p(x|\theta)}\right\} \\
&= \min\left\{1, \frac{p(x|\theta')}{p(x|\theta)}\right\}
\end{aligned}$$

Metropolis-Hastings算法

$$\alpha\left(\theta,\theta'\right) = \min\left\{1, \frac{p\left(x|\theta'\right)}{p\left(x|\theta\right)}\right\}$$

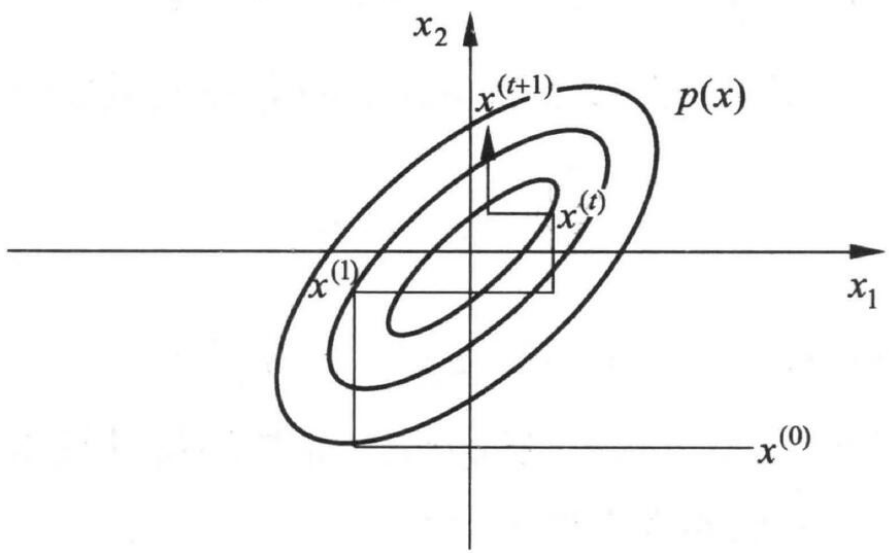
$$= \min\left\{1, \frac{\left(\begin{array}{c} n \\ x \end{array}\right)\theta'^{x}(1-\theta')^{n-x}}{\left(\begin{array}{c} n \\ x \end{array}\right)\theta^{x}(1-\theta)^{n-x}}\right\}$$

$$= \min\left\{1, \frac{\theta'^{x}(1-\theta')^{n-x}}{\theta^{x}(1-\theta)^{n-x}}\right\}$$

$$\log\alpha\left(\theta,\theta'\right) = \min\left\{0, x\log\theta' + (n-x)\log\left(1-\theta'\right) - x\log\theta - (n-x)\log\left(1-\theta\right)\right\}$$

Metropolis-Hastings算法

- **单分量Metropolis-Hastings算法**
- 在 Metropolis-Hastings 算法中，通常需要对多元变量分布进行抽样，有时对多元变量分布的抽样是困难的
- 可以对多元变量的每一变量的条件分布依次分别进行抽样，从而实现对整个多元变量的一次抽样，这就是单分量 Metropolis-Hastings (single-component Metropolis-Hastings) 算法



- 不一定是对单个维度依次采样，而是可以对多元随机变量的各个分块依次采样

吉布斯算法

- 本节叙述MCMC法的常用算法**吉布斯抽样（Gibbs Sampling）**，可以认为是 MetropolisHastings 算法的特殊情况，但是更容易实现，因而被广泛使用

吉布斯算法

- **满条件分布**
- MCMC法的目标分布通常是多元联合概率分布 $p(x) = p(x_1, x_2, \cdots, x_k)$, 其中 $x = (x_1, x_2, \cdots, x_k)^{\mathrm{T}}$ 为 k 维随机变量
- 如果条件概率分布 $p(x_I \mid x_{-I})$ 中所有 k 个变量全部出现, 其中 $x_I = \{x_i, i \in I\}, x_{-I} = \{x_i, i \notin I\}, I \subset K = \{1, 2, \cdots, k\}$, 那么称这种条件概率分布为满条件分布（full conditional distribution）

吉布斯算法

- 满条件分布有以下性质: 对任意的 $x, x' \in \mathcal{X}$ 和任意的 $I \subset K$, 有

$$p\left(x_I \mid x_{-I}\right) = \frac{p(x)}{\int p(x)\mathrm{d}x_I} \propto p(x)$$

吉布斯算法

- **例：**设 x_1 和 x_2 的联合概率分布的密度函数为

$$p(x_1, x_2) \propto \exp\left\{-\frac{1}{2}(x_1 - 1)^2 (x_2 - 1)^2\right\}$$

求其满条件分布

- 由满条件分布的定义有

$$\begin{aligned} p(x_1 \mid x_2) &\propto p(x_1, x_2) \\ &\propto \exp\left\{-\frac{1}{2}(x_1 - 1)^2 (x_2 - 1)^2\right\} \\ &\propto N\left(1, (x_2 - 1)^{-2}\right) \end{aligned}$$

这里 $N\left(1, (x_2 - 1)^{-2}\right)$ 是均值为 1，方差为 $(x_2 - 1)^{-2}$ 的正态分布，这时 x_1 是变量，x_2 是参数

吉布斯算法

- 同样可得

$$
\begin{aligned}
p\left(x_2 \mid x_1\right) &\propto p\left(x_1, x_2\right) \\
&\propto \exp\left\{-\frac{1}{2}\left(x_2-1\right)^2\left(x_1-1\right)^2\right\} \\
&\propto N\left(1,\left(x_1-1\right)^{-2}\right)
\end{aligned}
$$

这里 $N\left(1,\left(x_1-1\right)^{-2}\right)$ 是均值为 1，方差为 $\left(x_1-1\right)^{-2}$ 的正态分布，这时 x_2 是变量，x_1 是参数

吉布斯算法

- **基本原理：**从联合概率分布定义满条件概率分布，依次对满条件概率分布进行抽样，得到样本的序列
- 可以证明这样的抽样过程是在一个马尔可夫链上的随机游走，每一个样本对应着马尔可夫链的状态，平稳分布就是目标的联合分布
- 整体成为一个马尔可夫链蒙特卡罗法，燃烧期之后的样本就是联合分布的随机样本

吉布斯算法

- 假设多元变量的联合概率分布为 $p(x) = p\left(x_1, x_2, \cdots, x_k\right)$
- 吉布斯抽样从一个初始样本 $x^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \ldots, x_k^{(0)}\right)^{\mathrm{T}}$ 出发，不断进行迭代，每一次迭代得到联合分布的一个样本 $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}\right)^{\mathrm{T}}$
- 样本序列 $\left\{x^{(0)}, x^{(1)}, \cdots, x^{(n)}\right\}$

吉布斯算法

- 在每次迭代中，依次对 k 个随机变量中的一个变量进行随机抽样
- 如果在第 i 次迭代中，对第 j 个变量进行随机抽样，那么抽样的分布是满条件概率分布 $p\left(x_j \mid x_{-j}^{(i)}\right)$，这里 $x_{-j}^{(i)}$ 表示第 i 次迭代中，变量 j 以外的其他变量

吉布斯算法

- 设在第 $(i-1)$ 步得到样本 $\left(x_1^{(i-1)}, x_2^{(i-1)}, \cdots, x_k^{(i-1)}\right)^{\mathrm{T}}$，在第 i 步，首先对第一个变量按照以下满条件概率分布随机抽样

$$p\left(x_1 \mid x_2^{(t-1)}, \cdots, x_k^{(t-1)}\right)$$

得到 $x_1^{(i)}$，之后依次对第 j 个变量按照以下满条件概率分布随机抽样

$$p\left(x_j \mid x_1^{(i)}, \cdots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \cdots, x_k^{(i-1)}\right), \quad j = 2, \cdots, k-1$$

得到 $x_j^{(i)}$，最后对第 k 个变量按照以下满条件概率分布随机抽样

$$p\left(x_k \mid x_1^{(i)}, \cdots, x_{k-1}^{(i)}\right)$$

得到 $x_k^{(i)}$，于是得到整体样本 $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)}\right)^{\mathrm{T}}$

吉布斯算法

- 吉布斯抽样是单分量 Metropolis-Hastings 算法的特殊情况
- 定义建议分布是当前变量 $x_j, j = 1, 2, \cdots, k$ 的满条件概率分布

$$q\left(x, x'\right) = p\left(x'_j \mid x_{-j}\right)$$

这时，接受概率 $\alpha = 1$,

$$\begin{aligned}\alpha\left(x, x'\right) &= \min\left\{1, \frac{p\left(x'\right) q\left(x', x\right)}{p(x) q\left(x, x'\right)}\right\} \\ &= \min\left\{1, \frac{p\left(x'_{-j}\right) p\left(x'_j \mid x'_{-j}\right) p\left(x_j \mid x'_{-j}\right)}{p\left(x_{-j}\right) p\left(x_j \mid x_{-j}\right) p\left(x'_j \mid x_{-j}\right)}\right\} = 1\end{aligned}$$

这里用到 $p\left(x_{-j}\right) = p\left(x'_{-j}\right)$ 和 $p\left(\cdot \mid x_{-j}\right) = p\left(\cdot \mid x'_{-j}\right)$

吉布斯算法

- 转移核就是满条件概率分布

$$p\left(x, x'\right) = p\left(x'_j \mid x_{-j}\right)$$

也就是说依次按照单变量的满条件概率分布 $p\left(x'_j \mid x_{-j}\right)$ 进行随机抽样，就能实现单分量 Metropolis-Hastings 算法

- 吉布斯抽样对每次抽样的结果都接受，没有拒绝，这一点和一般的 Metropolis-Hastings 算法不同

吉布斯算法

算法 (吉布斯抽样)

输入：目标概率分布的密度函数 $p(x)$，函数 $f(x)$；

输出：$p(x)$ 的随机样本 $x_{m+1}, x_{m+2}, \cdots, x_n$，函数样本均值 f_{mn}；

参数：收敛步数 m，迭代步数 n。

(1) 初始化。给出初始样本 $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \cdots, x_k^{(0)})^{\mathrm{T}}$。

(2) 对 i 循环执行

设第 $(i-1)$ 次迭代结束时的样本为 $x^{(i-1)} = (x_1^{(i-1)}, x_2^{(i-1)}, \cdots, x_k^{(i-1)})^{\mathrm{T}}$，则第 i 次迭代进行如下几步操作：

$$
\begin{cases}
\text{(1) 由满条件分布 } p(x_1 | x_2^{(i-1)}, \cdots, x_k^{(i-1)}) \text{ 抽取 } x_1^{(i)} \\
\quad \vdots \\
\text{(j) 由满条件分布 } p(x_j | x_1^{(i)}, \cdots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \cdots, x_k^{(i-1)}) \text{ 抽取 } x_j^{(i)} \\
\quad \vdots \\
\text{(k) 由满条件分布 } p(x_k | x_1^{(i)}, \cdots, x_{k-1}^{(i)}) \text{ 抽取 } x_k^{(i)}
\end{cases}
$$

得到第 i 次迭代值 $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \cdots, x_k^{(i)})^{\mathrm{T}}$。

(3) 得到样本集合

$$\{x^{(m+1)}, x^{(m+2)}, \cdots, x^{(n)}\}$$

(4) 计算

$$f_{mn} = \frac{1}{n-m} \sum_{i=m+1}^{n} f(x^{(i)})$$

吉布斯算法

- **例：**用吉布斯抽样从以下二元正态分布中抽取随机样本

$$x = (x_1, x_2)^{\mathrm{T}} \sim p(x_1, x_2)$$

$$p(x_1, x_2) = \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

吉布斯算法

- **解：**条件概率分布为一元正态分布

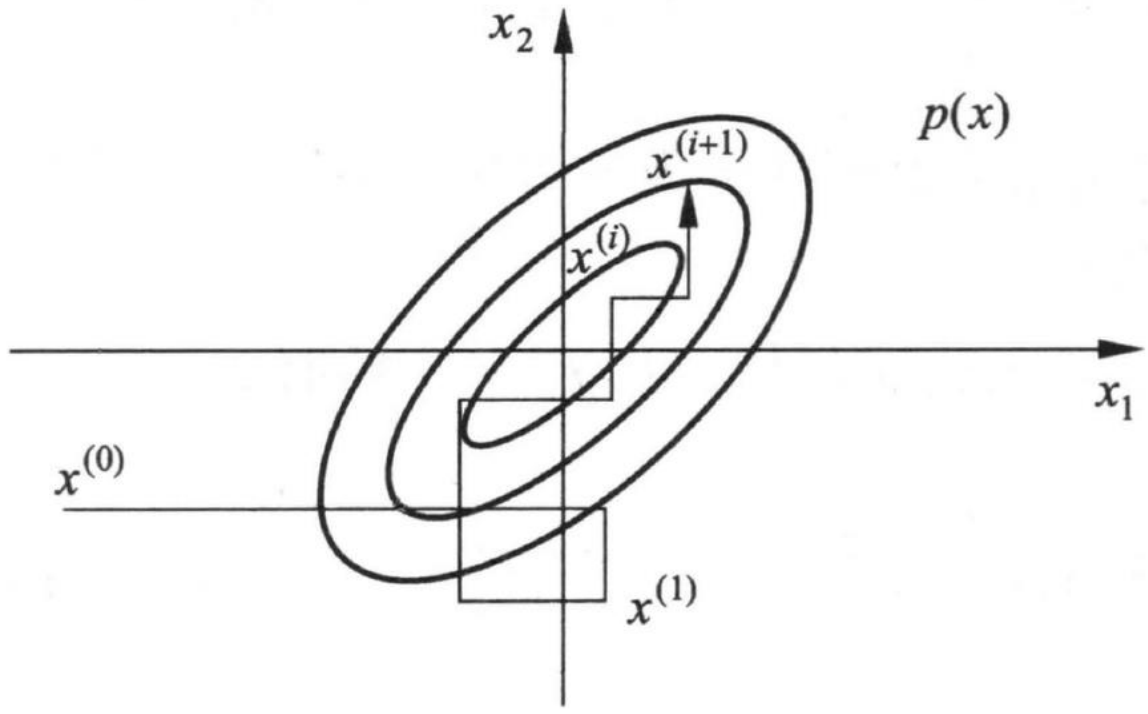
$$p\left(x_1 \mid x_2\right)=\mathcal{N}\left(\rho x_2,\left(1-\rho^2\right)\right)$$
$$p\left(x_2 \mid x_1\right)=\mathcal{N}\left(\rho x_1,\left(1-\rho^2\right)\right)$$

- 假设初始样本为 $x^{(0)}=\left(x_1^{(0)}, x_2^{(0)}\right)$, 通过吉布斯抽样, 可以得到以下样本序列:

迭代次数	对 x_1 抽样	对 x_2 抽样	产生样本
1	$x_1 \sim N(\rho x_2^{(0)}, (1-\rho^2))$, 得到 $x_1^{(1)}$	$x_2 \sim N(\rho x_1^{(1)}, (1-\rho^2))$, 得到 $x_2^{(1)}$	$x^{(1)}=(x_1^{(1)}, x_2^{(1)})^{\mathrm{T}}$
	$\vdots$	$\vdots$	$\vdots$
i	$x_1 \sim N(\rho x_2^{(t-1)}, (1-\rho^2))$, 得到 $x_1^{(t)}$	$x_2 \sim N(\rho x_1^{(t)}, (1-\rho^2))$, 得到 $x_2^{(t)}$	$x^{(t)}=(x_1^{(t)}, x_2^{(t)})^{\mathrm{T}}$
	$\vdots$	$\vdots$	$\vdots$

- 得到的样本集合 $\left\{x^{(m+1)}, x^{(m+2)}, \cdots, x^{(n)}\right\}, m<n$ 就是二元正态分布的随机抽样

吉布斯算法



- 单分量 Metropolis-Hastings 算法和吉布斯抽样的不同之处在于，在前者算法中，抽样会在样本点之间移动，但其间可能在某一些样本点上停留（由于抽样被拒绝）；而在后者算法中，抽样会在样本点之间持续移动

吉布斯算法

- **实例1：**使用吉布斯抽样推断下列只有截距项的简单回归模型的参数

$$y_i = \mu + e_i; \quad e_i \sim \mathcal{N}(0, \sigma^2)$$

因此，模型的似然函数为

$$p\left(y \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \mu)^2}{\sigma^2}\right)$$

- 假设参数 μ 和 σ^2 具有独立的先验，μ 的先验分布是均值为 η 方差为 τ^2 的一元正态分布 $p(\mu) = \mathcal{N}\left(\eta, \tau^2\right)$，而 σ^2 的先验分布是参数为 a 和 b 的逆伽马（Inverse Gamma）分布 $p\left(\sigma^2\right) = \mathrm{IG}(a, b)$

吉布斯算法

- 对于似然函数为正态分布的模型

$$p\left(y \mid \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right)$$

- 当方差 σ^2 已知时，均值 μ 的正态先验 $p(\mu) = \mathcal{N}\left(\eta, \tau^2\right)$ 与似然函数共轭
- 当均值 μ 已知时，方差 σ^2 的逆伽马（Inverse Gamma）先验 $p\left(\sigma^2\right) = \mathrm{IG}(a, b)$ 与似然函数共轭
- 但是，均值 μ 和方差 σ^2 的联合先验 $p\left(\mu, \sigma^2\right) = p(\mu)p\left(\sigma^2\right)$ 不是似然函数的共轭先验
- 参考：

  `https://en.wikipedia.org/wiki/Conjugate_prior`

吉布斯算法

- 由于均值 μ 和方差 σ^2 的联合先验 $p\left(\mu, \sigma^2\right) = p(\mu)p\left(\sigma^2\right)$ 不是似然函数的共轭先验，所以均值 μ 和方差 σ^2 的后验分布

$$p\left(\mu, \sigma^2 \mid y\right) = \frac{p\left(y \mid \mu, \sigma^2\right) p(\mu)p\left(\sigma^2\right)}{\iint p\left(y \mid \mu, \sigma^2\right) p(\mu)p\left(\sigma^2\right) d\mu d\sigma^2}$$

  没有闭式解 (closed-form solution)

吉布斯算法

- 使用吉布斯抽样：
- 均值 μ 的满条件分布

$$p\left(\mu \mid \sigma^2, \{y_i\}\right) = \mathcal{N}\left(\hat{\mu}, v_\mu\right)$$

其中 $\hat{\mu} = v_\mu \left[\tau^{-2}\eta + n\sigma^{-2}\bar{y}\right]$, $v_\mu^{-1} = \tau^{-2} + n\sigma^{-2}$, $\bar{y}$ 是数据 $y_{i=1}^n$ 的均值

- 方差 σ^2 的满条件分布

$$p\left(\sigma^2 \mid \mu, \{y_i\}\right) = \mathrm{IG}\left(a_1, b_1\right)$$

其中 $a_1 = a + n/2$, $b_1 = \left(\frac{1}{2}\sum_{i=1}^n \left(y_i - \mu\right)^2 + b^{-1}\right)^{-1}$