大数据管理方法与应用
第06章　马尔可夫链蒙特卡罗方法

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

2021 年 4 月 8 日

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**

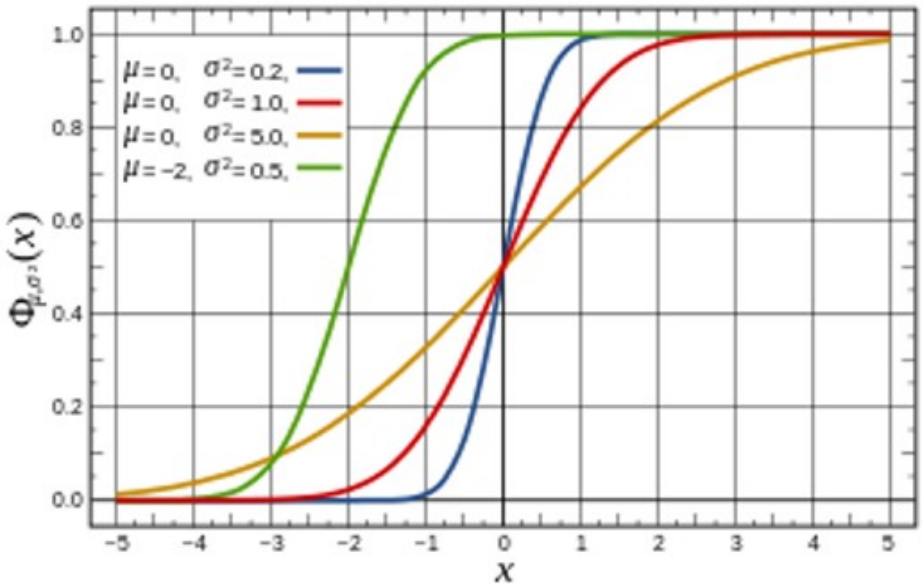
$$x_{n+1} = (ax_n + c) \bmod m$$

  - 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



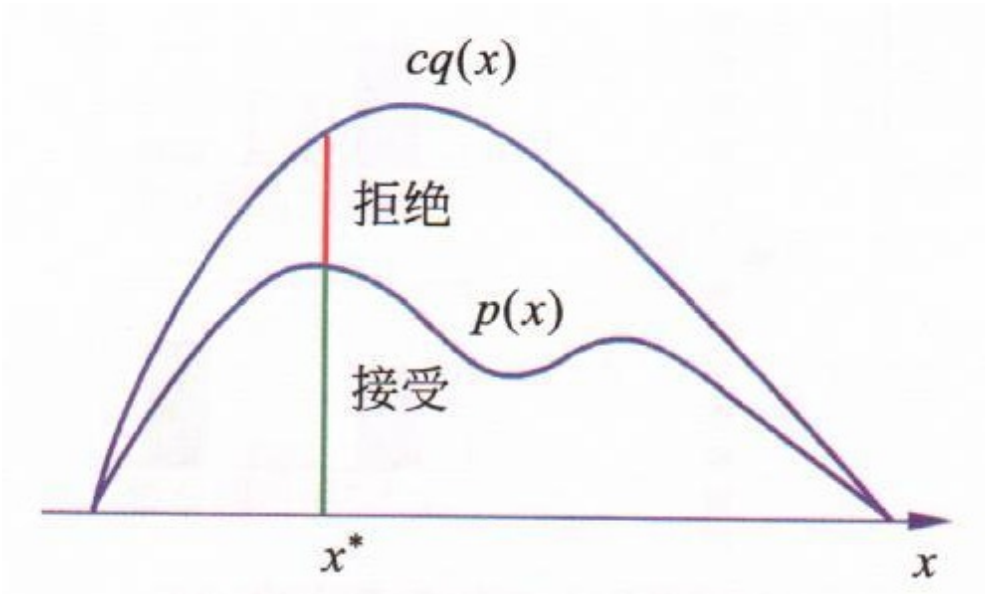
- 在 y 轴上产生（0,1）之间的均匀分布的随机数，水平投影到高斯累积分布函数上，然后垂直向下投影到 x 轴，得到的就是满足高斯分布的样本（随机数）

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 $x_1, x_2, \cdots, x_n$

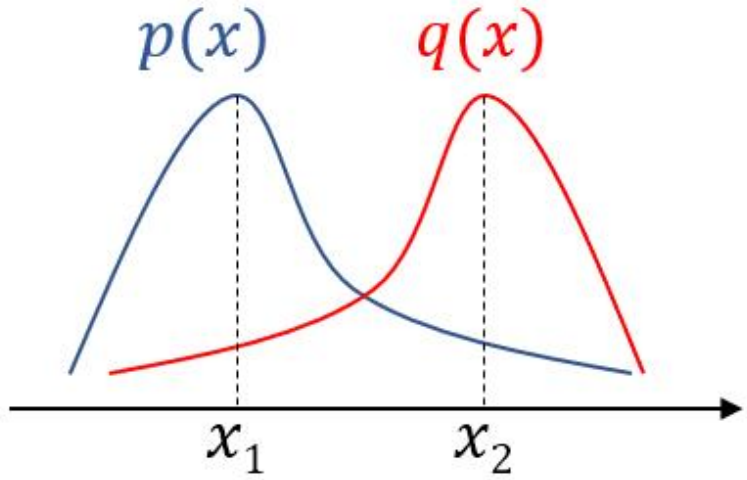
$$
\begin{aligned}
E_{p(x)}[f(x)] &= \int p(x) f(x) \mathrm{d}x \\
&= \int \frac{p(x)}{q(x)} q(x) f(x) \mathrm{d}x \\
&= \int \left(f(x) \frac{p(x)}{q(x)} \right) q(x) \mathrm{d}x \\
&\approx \frac{1}{n} \sum_{i=1}^{n} f(x_i) \frac{p(x_i)}{q(x_i)}
\end{aligned}
$$

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

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