**马尔可夫链蒙特卡罗算法**

**（Markov Chain Monte Carlo，MCMC）**

MCMC算法是用来在概率空间，通过随机采样估算兴趣参数后验分布的一种方法

# 蒙特卡罗方法

蒙特卡罗方法（Monte Carlo method），也称**统计模拟方法**，是1940年代中期由于科学技术的发展和电子计算机的发明，而提出的一种以概率统计理论为指导的**数值计算方法**，是指使用**随机数**（或更常见的伪随机数）来解决很多计算问题的方法。

## 解决问题分类

通常蒙特卡罗方法所解决的问题可以粗略地分成两类：

一类问题是所求解的问题本身具有内在的随机性，借助计算机的运算能力可以直接模拟这种随机的过程。

另一类问题是本来不具有随机性质的确定性问题可以转化为**某种随机分布的特征数**，比如随机事件出现的**概率**，或者随机变量的**期望值**。通过随机抽样的方法，以随机事件出现的频率估计其概率，或者以抽样的数字特征估算随机变量的数字特征，并将其作为问题的解。这种方法多用于求解复杂的**多维积分问题**。一般蒙特卡罗方法在数学中最常见的应用就是**蒙特卡罗积分。**

## 工作过程

在解决实际问题的时候应用蒙特卡罗方法主要有两部分工作：

* 用蒙特卡罗方法模拟某一过程时，需要产生某种概率分布的随机变量。
* 用统计方法把模型的数字特征估计出来，从而得到实际问题的数值解。

## 解题步骤

蒙特卡罗方法解题过程的三个主要步骤：

1. 构造或描述概率过程
2. 实现从已知概率分布抽样
3. 建立各种估计量

### 中心极限定理

中心极限定理是概率论中的一组定理。中心极限定理说明，**在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于正态分布。**

### 大数定律

大数定律又称大数法则、大数律，是描述相当多次数重复实验的结果的定律。根据这个定律知道，样本数量越多，则其算术平均值就有越高的概率接近期望值。

# 马尔科夫链

## 马尔科夫链的收敛性质：

如果一个非周期的马尔科夫链有状态转移矩阵，并且它的任何两个状态是连通的，那么 与无关，有：

* ，通常称为马尔科夫链的平稳分布。
* 
* 
* 是方程的唯一非负解。

马尔科夫链的状态数可以是有限的，也可以是无限的。因此可以用于连续概率分布和离散概率分布。

## 马尔科夫链的细致平稳条件

如果非周期马尔科夫链的状态转移矩阵和概率分布π(x)对于所有的i,j满足：



　　则称概率分布是状态转移矩阵的平稳分布。

# MCMC采样

目标平稳分布和某一个马尔科夫链状态转移矩阵不满足细致平稳条件，引入一个，使得下式成立。



则



得到对应的马尔科夫链状态转移矩阵



马尔科夫链状态转移矩阵可以通过任意一个马尔科夫链状态转移矩阵乘以得到。我们又一般称之为接受率。取值在[0,1]之间，可以理解为一个概率值，即目标矩阵P可以通过任意一个马尔科夫链状态转移矩阵以一定的接受率获得。

总结下MCMC的采样过程

1. 输入我们任意选定的马尔科夫链状态转移矩阵，目标平稳分布，设定状态转移次数阈值，需要的样本个数
2. 从任意简单概率分布采样得到初始状态值
3. :
4. 从条件概率分布中采样得到样本；
5. 从均匀分布采样；
6. 如果，则接受转移，即，否则不接受转移，即。

样本集即为我们需要的平稳分布对应的样本集。

**Note：可能非常的小，比如0.1，导致我们大部分的采样值都被拒绝转移，采样效率很低。**

# Metropolis-Hastings采样

1. 输入任意选定的马尔科夫链状态转移矩阵，目标平稳分布，设定状态转移次数阈值，需要的样本个数
2. 从任意简单概率分布采样得到初始状态值
3. :
4. 从条件概率分布中采样得到样本；
5. 从均匀分布采样；
6. 如果，则接受转移，即，否则不接受转移，即。

样本集即为我们需要的平稳分布对应的样本集。

M-H采样有两个缺点：一是需要计算接受率，在高维时计算量大。并且由于接受率的原因导致算法收敛时间变长。二是有些高维数据，特征的条件概率分布好求，但是特征的联合分布不好求。

# Gibbs采样

## 二维Gibbs采样

假设是一个二维联合数据分布，观察第一个特征维度相同的两个点和，容易发现下面两式成立：





由于两式的右边相等，因此我们有：



也就是：



观察上式再观察细致平稳条件的公式，我们发现在这条直线上，如果用条件概率分布作为马尔科夫链的状态转移概率，则任意两个点之间的转移满足细致平稳条件！同样的道理，在这条直线上，如果用条件概率分布作为马尔科夫链的状态转移概率，则任意两个点之间的转移也满足细致平稳条件。

基于上面的发现，我们可以这样构造分布的马尔可夫链对应的状态转移矩阵P：



有了上面这个状态转移矩阵，我们很容易验证平面上的任意两点E,F，满足细致平稳条件：



二维Gibbs采样，需要两个维度之间的条件概率

1. 目标平稳分布，设定状态转移次数阈值，需要的样本个数
2. 随机初始化初始状态值
3. :
4. 从条件概率分布中采样得到样本；
5. 从条件概率分布中采样得到样本；

样本集即为我们需要的平稳分布对应的样本集。

## 多维Gibbs采样