采样算法

在实际问题中我们会遇到这样的问题，假设得到了一个联合概率分布，需利用这个联合概率分布来计算某几维随机变量的边缘分布，或者观测到一部分随机变量的取值，而要去推理（Inference）其它未观测到的随机变量的取值的时候，往往出现两个问题：首先，这个联合分布可能并没有显式的表达式，因此给用积分求边缘分布的方法带来了困难；其次，即便联合分布有显式的表达式，有时候也会因为计算量过于庞大而难以实现。退而求其次的方法就是寻找一个近似解，而蒙特卡洛采样就是在这样一个背景下诞生的。虽然蒙特卡罗方法得到的是近似解，但是理论上该方法在采样足够多次时，得到的近似解能够以任意的精度接近精确解。

本文主要介绍几种常用的蒙特卡洛（MC）采样算法以及经典的马尔科夫链蒙特卡洛算法（MCMC）。

1. 蒙特卡洛采样（Monte Carlo Sampling）

1.1 简单随机变量的采样与变换方法（Transformation Method）

最简单的一维分布就是均匀分布，$X \sim U(0,1)$ 表示随机变量 X 在区间 $(0,1)$ 上服从均匀分布，计算机非常容易模拟产生这样的随机变量。以区间 $(0,1)$ 上的均匀分布为基础，可以模拟很多较简单的一维随机变量，比如任意区间上的均匀分布、标准正态分布、Dirichlet 分布、指数分布等等，其基本思想就是对均匀分布的随机变量 X 进行变换。令 $Y = f(X)$，则只要设计合适的变换函数 f，就可以实现想要的随机变量 Y，使其具有特定的分布。比如要想 $Y \sim U(0,a)$，只需令：$Y = f(X) = aX$ 即可；一维标准正态分布也可以通过对均匀分布进行变换得到，这里不再缀余。

一维离散分布也可以由均匀分布经过变换得到。比如离散随机变量 Y，可取值的集合为 $\{y_1, y_2, y_3, y_4\}$，对应的概率分别为 $\{p_1, p_2, p_3, p_4\}$。假设对均匀分布 $X \sim U(0,1)$ 进行一次采样，得到 X^s，则可以通过如下变换得到随机变量 Y 的样本 Y^s：

$$Y^s=\begin{cases} y_1 & X^s\in[0,p_1] \\ y_2 & X^s\in\left[p_1,\sum_{i=1}^{2}p_i\right] \\ y_3 & X^s\in\left[\sum_{i=1}^{2}p_i,\sum_{i=1}^{3}p_i\right] \\ y_4 & X^s\in\left[\sum_{i=1}^{3}p_i,1\right] \end{cases} \tag{1}$$

下面我们给出利用数学变换对目标分布进行采样的方法（Transformation Method）。假设我们需要采样的随机变量$Y\sim p(y)$，随机变量$Z\sim U(0,1)$，我们需要找到一个变换$Y=f(Z)$，从而使得随机变量 Y 的分布为$p(y)$。令 Y 的累积分布函数为$h(y)$，则有：

$$\begin{aligned} &\int_{-\infty}^{y}p(w)dw=h(y) \\ &=P(Y\le y)=P(f(Z)\le y) \\ &=P\left(Z\le f^{-1}(y)\right)=\int_{0}^{f^{-1}(y)}1dz \\ &=f^{-1}(y) \\ &=z \end{aligned} \tag{2}$$

即：

$$Y=h^{-1}(Z) \tag{3}$$

从上式可知，只需知道随机变量 Y 的累积分布函数$h(y)$的表达式，就可以通过变换实现想要的分布，但要想得到累积分布函数$h(y)$，需要首先对概率密度函数 $p(y)$进行积分。实际问题中，随机变量 Y 的概率密度函数可能比较复杂，或者根本没有解析表达式，因此对其积分有时候难以实现，这就是变换方法的最大缺点。

对于高维随机变量的采样，应用较多的是对有向图模型（贝叶斯网络）的采样。由于有向图模型中各个随机变量的独立性非常明确，给定某个节点的父节点的取值之后，该节点与其父节点对应的其它所有子节点相互独立，并且任意子节点的取值只和其所有直接父节点取值相关。因此采样可以首先对根节点进行采样（是一个一维随机变量的采样），得到根节点的采样值之后，依次对其子节点进

行条件分布的采样（仍然是一个一维随机变量的采样），以此类推，实现对整个图的一次采样。

对于已经有部分变量观测到的贝叶斯网络的采样，仍然可以按照上述方法进行，只要在完成一次完整采样之后，判断观测到的那些随机变量的取值和采样得到的是否相同，如果相同，则保留本次采样结果，如果不同，则抛弃本次采样结果，重新采样。该方法有一个很明显的缺陷，就是效率可能会非常低。因为那些观测到的随机变量取值发生的概率可能非常小，这就意味着采样得到的大部分样本都将被抛弃。

对于一般的高维随机变量，由于我们不知道各个随机变量之间的独立性关系，因此不能再用对贝叶斯网络的采样方法进行采样了。

1.2 拒绝性采样（Rejection Sampling）

实际中许多随机变量的分布不是一个简单的分布，甚至没有解析的表达式，如果要对这些随机变量进行采样，用前面介绍的变换方法显然不行，但拒绝性采样可以解决这一问题。

假设随机变量的分布密度函数为 $X \sim p(x)$，其中 $p(x)$ 要么非常复杂，要么根本没有解析表达式；假设一个分布密度函数 $q(x)$ 具有比较简单的表达式，并且满足 $kq(x) \geq p(x)$，其中 k 为常数。则首先可以按照分布 $q(x)$ 产生一个样本 x_0，其次按照均匀分布 $U(0, kq(x_0))$ 产生一个样本 u_0，如果 $u_0 > p(x_0)$，则拒绝本次采样，否则保留本次采样得到的样本。

如果按照这样的采样步骤，那么坐标轴上的任意一点 x 对应的非归一化概率密度为：

$$p'(X = x) = q(x)\frac{p(x)}{kq(x)} = \frac{p(x)}{k} = \frac{1}{k}p(x) \qquad (4)$$

因此得到的样本确实服从分布 $p(x)$。对于从分布 $q(x)$ 产生的任意样本，它被接受的概率为：

$$P(Accept) = \int \frac{p(x)}{kq(x)} q(x)\,dx = \frac{1}{k}\int p(x)\,dx = \frac{1}{k} = \frac{\int p(x)\,dx}{\int kq(x)\,dx} \qquad (5)$$

即样本被接受的概率等于概率分布 $p(x)$ 所覆盖的面积与 $kq(x)$ 所覆盖的面积之

比，因此要想提高样本被接受的概率，就要寻找一个分布 $q(x)$，使得它的形状尽量的接近 $p(x)$，这样才能保证常数 k 尽可能的小。

1.3 重要性采样（Importance Sampling）

实际中也可能遇到利用采样的方法近似某个随机变量函数期望的问题，即：

$$E\left(f\left(X\right)\right)=\int f\left(x\right)p\left(x\right)dx\approx\frac{1}{N}\sum_{i=1}^{N}f\left(x^{i}\right) \tag{6}$$

然而很不幸的是概率密度函数 $p(x)$ 要么非常复杂，要么没有解析表达式，导致采样无法进行。但是我们可以对上式稍做变化：

$$E\left(f\left(X\right)\right)=\int f\left(x\right)\frac{p\left(x\right)}{q\left(x\right)}q\left(x\right)dx\approx\frac{1}{N}\sum_{i=1}^{N}f\left(x^{i}\right)\frac{p\left(x^{i}\right)}{q\left(x^{i}\right)} \tag{7}$$

其中 $q(x)$ 是一个分布密度函数，并且具有较简单的解析表达式。我们把 $r=p(x)/q(x)$ 称为重要性加权系数（Importance Weights），因为它校正了从错误分布 $q(x)$ 采样导致的偏移（bias）。重要性采样和拒绝性采样主要区别为：重要性采样会抛弃某些样本，而该方法接受所有的样本。

2. 马尔科夫链蒙特卡洛采样（Markov Chain Monte Carlo Sampling）

2.1 马氏链及其平稳分布

给定一个状态空间，并且假设状态转移具有齐次一阶马尔可夫性质：

$$P\left(X_{t+1}=x_{t+1}\mid X_{t}=x_{t},X_{t-1}=x_{t-1},X_{t-2}=x_{t-2}\cdots\right)=P\left(X_{t+1}=x_{t+1}\mid X_{t}=x_{t}\right) \tag{8}$$

即下一时刻的状态取值只和当前时刻的状态取值有关，和更早的时刻没有关系。对于一个所有状态连通的非周期马尔科夫链，有如下定理成立。

马氏链定理：如果一个非周期的齐次马氏链具有转移概率矩阵 P，并且任意两个状态连通（状态并不要求有限个），则任意两个状态 i，j 之间的极限转移概率 $P_{i,j}$ 存在，并且与状态 i 无关，即：

$$\lim_{n\to\infty} P^n = \begin{bmatrix} \pi(1) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \pi(1) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{9}$$

其中向量 $\boldsymbol{\pi}=[\pi(1),\pi(2),\cdots,\pi(j),\cdots]$ 称为马尔科夫链的平稳分布，并且：

$$\pi(j)=\sum_{i=0}^{\infty}\pi(i)P_{ij},\ \lim_{n\to\infty}P_{ij}^{n}=\pi(j) \tag{10}$$

因此假定马尔科夫链的初始状态为 $\boldsymbol{\pi}_0=[\pi_0(1),\pi_0(2),\cdots,\pi_0(j),\cdots]$，显然经过足够多步的状态转移之后：

$$\pi_n(j)=\sum_{i=1}^{\infty}\pi_0(i)P_{ij}^{n} \tag{11}$$

当 $n\to\infty$ 时，对上式两边取极限，得到 $\pi_n(j)=\pi(j)$。显然当初始分布恰好是平稳分布时，有：

$$\boldsymbol{\pi}P=\boldsymbol{\pi} \tag{12}$$

实际上，上式可以用来计算一个马尔科夫链的平稳分布 $\boldsymbol{\pi}$。

马氏链定理存在的意义在于：对于一个具有一阶马尔科夫链性质的状态转移空间，不论初始状态的分布如何，经过足够多次的状态转移之后，各个状态出现的概率服从的分布会收敛到平稳分布，并且这个平稳分布唯一存在。因此我们有这样一个想法：能否构造一个马尔科夫链，使得它的状态空间对应某个随机变量（一维或者多维）所有可能的取值空间，而平稳分布恰好是这个随机变量的概率密度函数（或者分布列）。那么我们就可以从马氏链的任意状态出发，经过足够多步的状态转移之后，各个状态出现的概率进入平稳分布，此时只要不断运行马尔科夫链，就得到了我们想要的分布的样本。

这个想法在 1953 年被 Metropolis 想到，并设计出了 Metropolis 算法，该算法是一个普适性的采样算法，是一系列 MCMC 算法的基础。下面介绍两种常用的 MCMC 算法：Metropolis-Hastings 算法和 Gibbs Sampling 算法。

2.2 Metropolis-Hastings 算法

上一小节简单介绍了马尔科夫链的平稳分布，在引入本小节的算法之前，我们需要介绍判断马尔科夫链的细致平稳条件。

细致平稳条件：如果状态连通的非周期马尔科夫链的转移矩阵 P 和分布 $\boldsymbol{\pi}=[\pi(1),\pi(2),\cdots,\pi(j),\cdots]$ 满足：

$$\pi(i)P_{ij}=\pi(j)P_{ji} \quad \forall i,j \tag{13}$$

则 $\boldsymbol{\pi}=[\pi(1),\pi(2),\cdots,\pi(j),\cdots]$ 是马氏链的平稳分布。

证明非常简单，由于平稳分布的唯一性，我们只需要验证细致平稳条件满足 $\boldsymbol{\pi}P=\boldsymbol{\pi}$ 即可。对上式两端对 i 求和有：

$$\begin{aligned}
&\sum_{i=1}^{\infty}\pi(i)P_{ij}=\sum_{i=1}^{\infty}\pi(j)P_{ji}\\
&\sum_{i=1}^{\infty}\pi(i)P_{ij}=\pi(j)\sum_{i=1}^{\infty}P_{ji}\\
&\sum_{i=1}^{\infty}\pi(i)P_{ij}=\pi(j)\\
&\boldsymbol{\pi}P=\boldsymbol{\pi}
\end{aligned} \tag{14}$$

我们以一维随机变量的采样为例，假设我们要用马尔科夫链的平稳分布来逼近概率分布函数 $p(x)$，并假定我们已经有一个转移矩阵为 Q 的马尔科夫链，令 $q(i,j)$ 表示状态 i 到状态 j 的一步转移概率。显然在通常情况下：

$$p(i)q(i,j)\neq p(j)q(j,i) \tag{15}$$

即细致平稳条件不成立，也就是说任意给定一个马尔科夫链，$p(x)$ 一般不会是这个马氏链的平稳分布。但是我们可否对该马氏链进行改造，使得细致平稳条件成立呢？譬如我们引入一个因子 $\alpha(i,j)$，从而使得：

$$p(i)q(i,j)\alpha(i,j)=p(j)q(j,i)\alpha(j,i) \tag{16}$$

最直接的想法就是令：

$$\alpha(i,j)=p(j)q(j,i),\quad \alpha(j,i)=p(i)q(i,j) \tag{17}$$

于是我们就把一个原来具有转移概率矩阵 Q 的普通马氏链，改造成了一个具有转移矩阵 Q' 的马氏链，而 Q' 恰好满足细致平稳条件，因此改造后的马氏链的平稳分布就是 $p(x)$。

人们一般把因子 $\alpha(i,j)$ 理解为接受率（或者拒绝率），即在原来的马尔科夫链上，从状态 i 以概率 $q(i,j)$ 跳转到状态 j 的时候，我们会以 $\alpha(i,j)$ 的概率接收这次转移，因此得到的新的马氏链的转移概率就变成了 $q(i,j)\alpha(i,j)$。因此 MCMC 实质上也是一种拒绝性采样。

以上 MCMC 算法存在一个小小的问题，就是转移矩阵为 Q 的马氏链在转移的过程中接受率 $\alpha(i,j)$ 可能会非常小，这样会导致马氏链在转移过程中可能原地踏步，这使得马尔科夫链遍历所有的状态需要花费的时间可能太长，收敛速度太慢。我们发现：

$$\begin{aligned} p(i)q(i,j)\alpha(i,j) &= p(j)q(j,i)\alpha(j,i) \\ p(i)q(i,j)\left[k\alpha(i,j)\right] &= p(j)q(j,i)\left[k\alpha(j,i)\right] \end{aligned} \tag{18}$$

于是我们可以将等式两边的接受率同时放大 k 倍，使得：

$$\max\left(k\alpha(i,j), k\alpha(j,i)\right) = 1 \tag{19}$$

这样不仅满足了细致平稳条件，而且增大了状态转移的接受率。因此我们可以取：

$$\alpha(i,j) = \begin{cases} \dfrac{p(j)q(j,i)}{p(i)q(i,j)} & if\ \ p(j)q(j,i) < p(i)q(i,j) \\ 1 & elsewise \end{cases} \tag{20}$$

基于上述分析和假设，下面给出 Metropolis-Hastings 算法的算法流程。

（I） 初始化马氏链的初始状态 $X_0 = x_0$

（II） 对 $t = 0,1,2,\cdots$ 循环以下过程进行采样：

- 第 t 个时刻马氏链的状态为 $X_t = x_t$，采样 $y \sim q(x \mid x_t)$
- 从均匀分布采样 $u \sim Uniform(0,1)$
- 如果 $u < \alpha(x_t, y)$，则接受本次状态转移 $x_t \to y$，即 $X_{t+1} = y$
- 否则不接受该次状态转移，令 $X_{t+1} = x_t$

需要注意的是，该算法需要截取转移矩阵Q'对应的马氏链达到平稳（收敛）之后的输出作为采样值。

对于多为随机变量，同样可以按照上述算法实现采样，只不过此时状态空间的一个状态不再是一维变量，而是一个向量，细致平稳条件变成了：

$$p(\mathbf{x})q(\mathbf{x}\to\mathbf{y})\alpha(\mathbf{x}\to\mathbf{y})=p(\mathbf{y})q(\mathbf{y}\to\mathbf{x})\alpha(\mathbf{y}\to\mathbf{x}) \tag{21}$$

但是即便对接受率α进行放大，也不能保证任意$\alpha(i,j)$取值等于 1，实际上许多α的取值会很小，所以算法的收敛性能会比较差，我们需要找到一种方法，使得在高维情况下，接受率α=1。

2.3 Gibbs Sampling 算法

Gibbs 采样算法也是一种 MCMC，适用于对高维随机变量的采样，在采样过程中，能够实现接受率α=1。

我们先以二维情形为例，假设二维随机变量(X,Y)，服从分布$p(x,y)$，考虑两个点(x_1,y_1)和(x_1,y_2)，它们的x坐标相同。我们有：

$$\begin{aligned} p(x_1,y_1)p(y_2\mid x_1)&=p(x_1)p(y_1\mid x_1)p(y_2\mid x_1)\\ p(x_1,y_2)p(y_1\mid x_1)&=p(x_1)p(y_2\mid x_1)p(y_1\mid x_1)\end{aligned} \tag{22}$$

所以有：

$$p(x_1,y_1)p(y_2\mid x_1)=p(x_1,y_2)p(y_1\mid x_1) \tag{23}$$

基于以上等式我们发现，如果固定二维随机变量中X的取值$X=x_1$，那么对于任意两个状态(x_1,y_1)和(x_1,y_2)，如果取转移概率为$p(y_2\mid x_1)$和$p(y_1\mid x_1)$，则这两个状态的相互转移满足细致平稳条件。

我们构造如下的状态转移概率矩阵：

$$\begin{aligned} Q\big((x,y_1)\to(x,y_2)\big)&=p(y_2\mid x)\\ Q\big((x_1,y)\to(x_2,y)\big)&=p(x_2\mid y)\\ Q\big((x_1,y_1)\to(x_2,y_2)\big)&=0\end{aligned} \tag{24}$$

则对于由(X,Y)构成的状态空间中的任意两个状态(a,b)和(c,d)，有：

$$p\big((a,b)\big)Q\big((a,b)\to(c,d)\big)=p\big((c,s)\big)Q\big((c,d)\to(a,b)\big) \qquad (25)$$

即矩阵 Q 满足细致平稳条件。因此由矩阵 Q 决定的马尔科夫链将收敛到平稳分布 $p(x,y)$。可以发现，以上构造的马尔科夫链的状态转移每一步都可以完成转移，并没有接受率，因此接受率为 1。另外需要注意的一点是，假设发生了一次有效的状态转移，由矩阵 Q 的性质可知，相邻两个转移状态之间只会有一个维度的取值发生变化，其它维度的取值不变，因此一段可能的状态转移为：

$$\big(x_0,y_0\big)\to\big(x_0,y_1\big)\to\big(x_1,y_1\big)\to\big(x_1,y_2\big)\to\big(x_2,y_2\big)\to\cdots \qquad (26)$$

显然上述二维情形的结论很容易推广到高维情形，这里不再缀余。下面给出 Gibbs 采样在一般情况下的算法流程。

（I） 假设为 N 维随机变量 $\mathbf{X}$。首先随机产生一个初始状态 $\mathbf{X}=\mathbf{X}^0$。

（II） 对 $t=1,2,3,\cdots$ 循环：

对 $n=1,2,\cdots,N$，循环采样：

$$X_n^{t+1}\sim p(x_n\mid X_{-n}^t)$$

对上述算法有一点补充说明：对高维随机变量的采样（内循环）不一定要按照坐标顺序轮流采样；最一般的情形是：在时刻 t，可以在 N 维坐标中任取一维坐标，然后按照条件概率进行状态转移，得到新的状态，继续时刻 $t+1$ 的状态转移。