

# MCMC

## Markov Chain &

## Monte Carlo

① 基本采样算法

② MCMC

③ 吉布斯采样

④ 面临的困难.

# 基本采样算法

## (Basic Sampling Algorithm)

MCMC 整一章我们关注的点在于基于数值采样的近似推断方法也被称为蒙特卡罗 (Monte Carlo) 方法。

为什么采样是困难的?  $\Rightarrow$  ① partition function is untractable; ② high dimension

一般来说, 我们关注的并非基于未观测变量的后验概率本身, 而是使用该后验概率求期望 (例如变分推断背景中所介绍的)。

所以我们将问题抽象出来, 描述为,

求解某个函数  $f(z)$  基于某个概率分布  $p(x)$  的期望, 即

$$E[f] = \int p(z)f(z)dz = \sum_z p(z)f(z)$$

采样方法就是从  $p(z)$  中, 独立抽取一组样本  $\{z^{(l)} | l=1, 2, \dots, L\}$ , 使用有限的样本去估计期望, 即

$$\hat{f} = \frac{1}{L} \sum_{l=1}^L f(z^{(l)})$$

只要  $z^{(l)}$  是从  $p(z)$  中抽取的, 那么  $E[\hat{f}] = E[f]$ , 同时  $f$  的方差为

$$\text{Var}[\hat{f}] = \frac{1}{L} E[(f - E[f])^2]$$

下面对上述蓝框中的结论进行推导.

$$\textcircled{1} \quad E[\hat{f}] = E[f]$$

$$E[\hat{f}] = \int \hat{f}(z) p(z) dz$$

$$= \int \left( \frac{1}{L} \sum_{l=1}^L f(z^{(l)}) \right) p(z) dz$$

$$= \frac{1}{L} \sum_{l=1}^L E_{z^{(l)}}[f]$$

$$= \frac{1}{L} \sum_{l=1}^L E_{z^{(l)}}[f]$$

由于  $z^{(l)}$  是从  $p(z)$  中抽取的, 故  $z^{(l)}$  与  $z$  同分布

$$\text{即 } E[\hat{f}] = \frac{1}{L} \sum_{l=1}^L E_z[f] = E[f]$$

$$\textcircled{2} \quad \text{Var}[\hat{f}] = \frac{1}{L} E[(f - E[f])^2]$$

$$\text{Var}[\hat{f}] = E[\hat{f}^2] - E[\hat{f}]^2$$

对于后一项, 我们可以 直接使用  $E[\hat{f}]$  进行代替.

$$\text{对于前一项: } E[\hat{f}^2] = \int \hat{f}^2(z) p(z) dz$$

$$= \int \left( \frac{1}{L} \sum_{l=1}^L f(z^{(l)}) \right)^2 p(z) dz$$

$$= \frac{1}{L^2} \int \left( \sum_{l=1}^L f(z^{(l)}) + \sum_{i=1}^L \sum_{j=i+1}^L f(z^{(i)}) f(z^{(j)}) \right) p(z) dz$$

$$= \frac{1}{L^2} \left[ \sum_{l=1}^L \int f(z^{(l)}) p(z^{(l)}) dz^{(l)} + \underbrace{\frac{L(L-1)}{2} \times 2 \int f(z^{(i)}) f(z^{(j)}) p(z^{(i)}) p(z^{(j)}) dz^{(i)} dz^{(j)}}_{\int f(z^{(i)}) p(z^{(i)}) dz^{(i)} \cdot \int f(z^{(j)}) p(z^{(j)}) dz^{(j)}} \right]$$

$$= \frac{1}{L} E[\hat{f}^2] + \frac{L-1}{L} E[\hat{f}]$$

$$\int f(z^{(i)}) p(z^{(i)}) dz^{(i)} \cdot \int f(z^{(j)}) p(z^{(j)}) dz^{(j)}$$

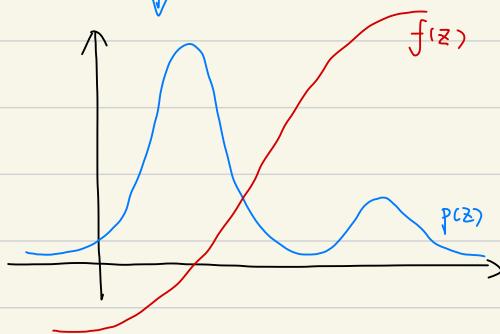
$$\begin{aligned} \text{故 } \text{Var}[\hat{f}] &= \frac{1}{L} E[\hat{f}^2] + \frac{L-1}{L} E[\hat{f}]^2 - E[\hat{f}]^2 \\ &= \frac{1}{L} (E[\hat{f}^2] - E[\hat{f}]^2) \\ &= \underbrace{\frac{1}{L} E[(f - E[f])^2]}_{\text{方差公式}} \end{aligned}$$

注意，这里使用样本估计  $E[\hat{f}]$  和  $\text{Var}[\hat{f}]$ ，估计的精度不依赖于  $L$  的维度，且原则上，数量较少的样本  $L^{(1)}$  可能会达到较高的精度。

$\downarrow$  (10, 20 个独立的样本可以达到高的精度对期望作出估计)

可能是由于数值分析中提到的：计算机中的误差累积所造成。

然而，问题在于样本可能不是独立的，有效样本的数量可能远远小于实际样本的数量，如果  $f(z) =$  较大区域是  $p(z)$  的较小区域，也即说明最终要求的期望是由小概率区域决定的，为了达到足够的精度需要较大的样本。



故，什么叫做样本？大致满足以下两点：

① 样本始于高概率区域；② 样本之间相互独立。

## ① 祖先采样方法 (ancestral sampling approach)

联合概率  $p(z) = p(z_1, z_2, z_3, \dots, z_p)$

如果能用图模型表示，又  $p(z) = \prod_{i=1}^M p(z_i | \text{par}_i)$

其中  $\text{par}_i$  表示与节点  $z_i$  相关联的节点，(Markov Blanket)

所谓祖先采样方法，针对的是无观测变量的有向图，即根据图模型，按收序对  $z$  进行采样。每个  $z_i$  都是从  $p(z_i | \text{par}_i)$  中采样得到的。当采样到  $z_i$  时，要求  $\text{par}_i$  中的所有节点，都已完成初始化。

在对图遍历一次之后，我们会得到一个来自  $p(z)$  的样本。

## ② 遗传采样 (Logic Sampling)

上述是针对无观测变量，而遗传采样针对的是有观测变量的有向图。

方法也很简单。就是按照祖先采样的方式采样。当我们采样到  $z$  且他的值被观测，如果采样值与观测值不相等，则放弃所有采样，从第一个节点重新开始。

实际上，这种方法很少用到，因为随着观测变量数量的增加，以及单个变量可取得的状态，想要得到和观测值一样序列的概率会越来越小（试想，抽 100 次硬币，都观测到正面的概率）。

对于无向图，就不存在遍历一次图就可以采样的方法（可以想一下，无向图最大团内部用这种方法就行不通，因为每个节点互相为相关联的节点，无论从哪个开始采样，总有未初始化的相关节点）。

所以我们会用计算量更大的采样方式。

如后续会介绍的 Gibbs 采样

除了需要从条件概率分布  $p(z|p_{\text{ari}})$  中采样之外，我们也有可能会从边缘概率分布中采样。但是，如果我们已经有了从联合概率分布中采样的方法（从  $p(x, z)$  中采样），得到  $p(x)$  是很容易的，我们只需要忽略每个样本中的  $z$  值即可。

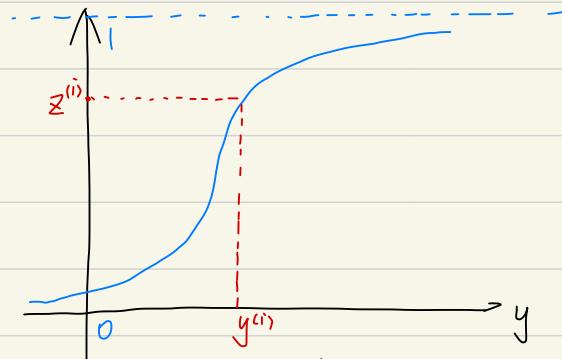
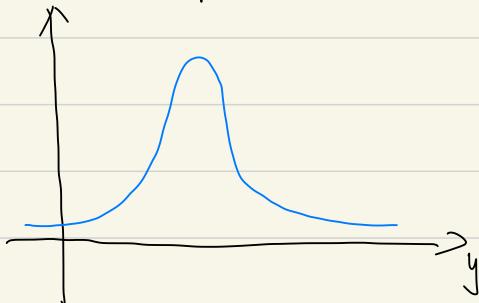
### ③ 标准概率分布采样

利用计算机，我们能很方便地从均匀分布  $U(0, 1)$  中采样。

假设现在， $z \sim U(0, 1)$ ,  $\Rightarrow p(z) = 1$  ( $U(0, 1)$  的概率密度就是 1)

如果我们现在有一个函数  $f(\cdot)$ , 令  $y = f(z)$ , 则可以认为我们在  $U(0, 1)$  上采样，经过  $f(\cdot)$  变换得到  $y$ , 且  $y$  服从概率分布  $p(y)$ 。

那么  $f(\cdot)$  就是  $\int_0^y p(y) dy$  的反函数，即累积密度函数 (cdf) 的反函数。  
直观地理解如下图：



左图为概率密度函数 (pdf), 右图为累积密度函数 (cdf)。

标准概率分布采样二形象化理解就是在右图竖直轴上  $[0, 1]$  区间内均匀采样得到  $z^{(i)}$ ，再根据  $z$  由 cdf 映射到横轴变为  $y^{(i)}$ ，即为一个样本。  
(PRML 中的公式我没怎么看懂，但是看白板推导的视频中应该就是这个意思)

#### ④ 拒绝采样 (Rejection Sampling)

很显然，③中的采样方法要求目标概率分布的不定积分或反函数是可求的，但是，实际应用中大多都是不可求的。所以这里提出另一种解决方法。

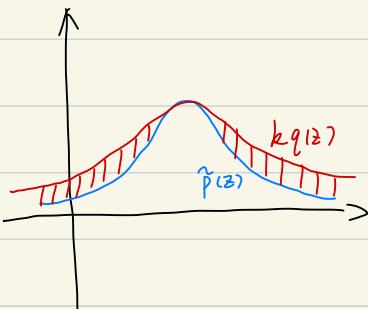
首先，我们考虑单变量的分布，我们假设概率分布如下。

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

其中， $\tilde{p}(z)$  很容易求， $Z_p$  未知。

其次，我们有一个提议分布 (proposal distribution)： $q(z)$ ，这个分布形式简单容易计算，并且可以从中采样。

然后，我们引入一个系数  $k$ ，满足对所有  $z$  的取值都有  $kq(z) \geq \tilde{p}(z)$ 。  
如下图所示



接下来我们进行采样，步骤如下

1. 在  $q(z)$  中采样得到  $z^{(0)}$ 。
2. 在  $[0, kq(z^{(0)})]$  上均匀采样，得到  $u^{(0)}$
3. 若  $u^{(0)}$  落在红色阴影范围内，即  $u^{(0)} > \tilde{p}(z^{(0)})$ ，那么样本被拒绝，反之，被接受。

上述的采样方式就是一般的拒绝采样。

当  $z$  从  $g(z)$  中采样之后，其被接受的概率为  $\frac{\tilde{P}(z)}{kg(z)}$ .

因此，样本被接受的概率为：

$$p(\text{接受}) = \int \left\{ \frac{\tilde{P}(z)}{kg(z)} \right\} g(z) dz = \frac{1}{k} \int \tilde{P}(z) dz$$

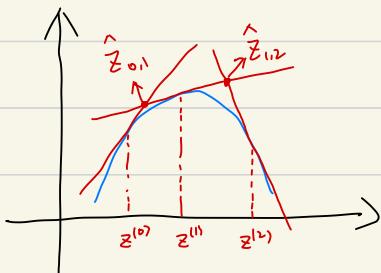
分析上式可知， $p(\text{接受})$  取决于  $\tilde{P}(z)$  的面积和  $k$  的值。 $\tilde{P}(z)$  面积我们不可控，因此，我们只有让  $k$  尽量小， $p(\text{接受})$  才会尽量大。因此一个如  $g(z)$  的形式十分重要，但往往这是很难实现的，所以就有了 调节拒绝采样

(提建议分布)

如果  $p(z)$  是对数凹函数 ( $\ln p(z)$  的导数单调递增)，界限函数的构建是十分简单的。计算  $\ln p(z)$  在某些初始值位置的切线。各个切线相交形成一个分段函数，由于计算的是  $\ln p(z)$  的切线，故分段函数需要加上  $\exp\{\cdot\}$ ，因此，最终形式是一个分段指数簇分布。如下。

$$g(z) = k_i \lambda_i \exp\{-\lambda_i(z - z_{i-1})\}, \hat{z}_{i-1} < z_i \leq \hat{z}_{i+1}$$

其中， $\hat{z}_{i-1}$  表示在  $z^{(i-1)}$  处生成切线与  $z^{(i)}$  处生成切线的交点。这样，我们就可以应用一般的拒绝采样，唯一不同之处在于，如果某个样本被拒绝，我们就将其加入初始值集合，在该处生成一条切线。重新计算  $g(z)$



从左图中感受到，随着采样次数升高， $g(z)$  会越来越接近  $\tilde{P}(z)$

PRML 中还提了一点，不需要满足对数凹函数一节中调节拒绝 Metropolis 采样。

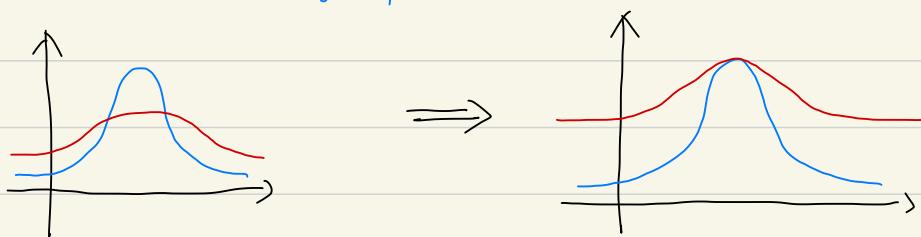
现在，我们考虑高维情形。以高斯分布为例(D维高斯分布)

$$\text{假设 } \tilde{P}(z) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu) \right\}, \quad \Sigma = \sigma_p^2 I$$

$$\text{则 } q(z) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_q|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (z - \mu_q)^T \Sigma_q^{-1} (z - \mu_q) \right\}, \quad \Sigma_q = \sigma_q^2 I$$

虽然为了使  $kq(z) \geq \tilde{P}(z)$  有 k 值存在,  $\sigma_q^2 \geq \sigma_p^2$ , (原因请参照 D=1 一维情形,  
G 越小, 右侧越薄, G 越大, 右侧越厚).

因此为了让  $kq(z) \geq \tilde{P}(z)$ , 我们只需要将  $q(z)$  的极大值放缩到  $\tilde{P}(z)$  极大  
值相同 (原因见下图一维情形)



故 k 就为当  $\Sigma = \Sigma_q = \mu_q$  时, 两者之比值.

$$k = \frac{\frac{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}}{1}}{\frac{(2\pi)^{\frac{D}{2}} |\Sigma_q|^{\frac{1}{2}}}{1}} = \frac{|\Sigma|^{\frac{1}{2}}}{|\Sigma_q|^{\frac{1}{2}}} = \left(\frac{\sigma_q}{\sigma_p}\right)^D \quad p(\text{接受}) = \frac{1}{k} \int \tilde{P}(z) dz.$$

$\downarrow$

$= 1.$   
 $\tilde{P}(z)$  已经归一化

又由于  $\sigma_q \geq \sigma_p$ , 故, k 将以 D 为指数的形式增长 因此  $p(\text{接受}) = \frac{1}{k}$  会随着 D 的增大而迅速减小. (当 D=1000 时,  $\sigma_q$  只比  $\sigma_p$  大一个量级,  $p(\text{接受}) = \frac{1}{2000}$ ).

更加一般的情况, 根本分布可能为多峰甚至有尖峰, 提议分布更困难寻找.

故, 拒绝采样在一维或二维空间中有用, 且不适用于高维空间。

但是对于高维空间中更复杂的算法来说, 它起着子过程的作用。

## ⑤ 重要采样 (Importance Sampling)

我们要从  $p(z)$  中采样的一个原因是计算  $E_{p(z)}[f(z)]$ ，重要采样直接近似了  $E_{p(z)}[f(z)]$ ，本身并没有对  $p(z)$  进行采样。我们回到计算  $E[f]$  上。

我们假设  $p(z)$  很容易计算，于是，我们可以将空间离散化为均匀的格点，用求和代替积分，如下式：

$$E[f] = \int p(z)f(z)dz \approx \sum_{l=1}^N p(z^{(l)})f(z^{(l)})$$

但是，明显的缺点是求和的项会随着维度指数级增长。同时，由于维度灾难加之概率密度的特点，我们感兴趣的概率分布通常将它们的大部分质量限制在空间中的一个很小的区域，因此，均匀格点采样方式效率极其低下。只有很小一部分样本会对上述求和式有巨大的贡献。

因此，这种计算  $E[f]$  的方式行不通，我们再次使用提议分布  $q(z)$ 。

$$E[f] = \int p(z)f(z)dz = \int \frac{p(z)}{q(z)} f(z) \cdot q(z) dz \approx \frac{1}{L} \sum_{l=1}^L \frac{p(z^{(l)})}{q(z^{(l)})} f(z^{(l)})$$

这样，我们只需从  $q(z)$  中采样  $\{z^{(l)} | l=1, 2, \dots, L\}$ ，然后就能计算了（这里默认  $p(z)$  还是很容易进行计算）其中  $r_l = \frac{p(z^{(l)})}{q(z^{(l)})}$  是重要性权重 (Importance Weights)，其修正了由于从不同概率分布中采样所引起的偏差（和拒绝采样不同，这里的每个样本都是被保留的）。

若  $p(z) = \frac{1}{z_p} \tilde{p}(z)$  中,  $\tilde{p}(z)$  是容易计算的, 但  $\frac{1}{z_p}$  不容易算. 这种情形下, 我们令  $\tilde{q}(z) = \frac{1}{z_q} \tilde{p}(z)$ ,  $\tilde{q}(z)$  容易计算. 故

$$E[f] = \frac{z_q}{z_p} \int \frac{\tilde{p}(z)}{\tilde{q}(z)} f(z) g(z) dz$$

$$\approx \frac{z_q}{z_p} \cdot \frac{1}{L} \sum_{l=1}^L \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})} f(z^{(l)})$$

其中,  $\tilde{r}_l = \frac{\tilde{p}(z^{(l)})}{\tilde{q}(z^{(l)})}$ , 至于  $\frac{z_q}{z_p}$  可以通过计算  $\frac{z_p}{z_q}$  得到:

$$\frac{z_p}{z_q} = \int \frac{1}{z_q} \tilde{p}(z) dz \quad \Delta \rightarrow g(z) = \frac{1}{z_q} \tilde{q}(z) \Rightarrow z_q = \frac{\tilde{q}(z)}{g(z)}$$

$$= \int \frac{\tilde{p}(z)}{\tilde{q}(z)} g(z) dz$$

$$= \frac{1}{L} \sum_{l=1}^L \tilde{r}_l \quad \text{故 } \frac{z_q}{z_p} = L \cdot \frac{1}{\sum_{l=1}^L \tilde{r}_l}$$

代入式中, 得:

$$E[f] = \sum_{l=1}^L \frac{\tilde{r}_l}{\sum_{m=1}^L \tilde{r}_m} f(z^{(l)}) = \sum_{l=1}^L w_l f(z^{(l)})$$

$w_l$

与拒绝采样相同, 垂直采样同样依赖于  $g(z)$  和  $p(z)$  二匹配程度。若通过  $g(z)$  采样得到的样本都设在  $f(z)p(z)$  较大的区域中,  $r_l$  和  $r_l f(z^{(l)})$  表面上方差很小, 实际上期望完全估计错误。

故, 垂直采样可能会产生任意错误结果的可能性, 且这种错误无法检测。故  $g(z)$  不应该在  $p(z)$  可能较大的区域中取值很小或为 0. (PRML 对图的垂直采样和仍然加权采样没怎么看懂).

## ⑥ 采样 - 重要性 - 重采样 (Sampling - importance - resampling)

之前在讲拒绝采样的时候，咱们角了权值的重要性，对于大部分  $p(z)$  和  $q(z)$  来说，确定最合适  $k$  是不现实的，也就侧面说明了，想要取得较高的  $p(z)$  接受是比較困难的。

这里，我们借助重要采样。将是否接受样本看作是利用重要采样的权值进行二次采样。具体如下：

①  $L$  个样本  $z^{(1)}, \dots, z^{(L)}$  从  $q(z)$  中抽取。

② 根据重要采样的公式，计算  $w_1, \dots, w_L$ 。

③ 将  $w_1, \dots, w_L$  作为权率，从  $z^{(1)}, \dots, z^{(L)}$  中进行二次采样，得到  $L$  个样本。

多项分布。

当  $L \rightarrow \infty$  时，生成的  $L$  个样本的分布等同于  $p(z)$ 。证明：累积分布函数为：

$$P(z \leq a) = \sum_{z^{(l)} \leq a} w_l = \frac{\sum I(z^{(l)} \leq a) \tilde{p}(z^{(l)}) / \tilde{q}(z^{(l)})}{\sum \tilde{p}(z^{(l)}) / \tilde{q}(z^{(l)})} = \frac{\sum I(z^{(l)} \leq a) \tilde{p}(z^{(l)}) / q(z^{(l)})}{\sum \tilde{p}(z^{(l)}) / q(z^{(l)})} \quad q(z) = \frac{1}{\tilde{q}(z)}$$

$I(z^{(l)} \leq a)$  是示性函数，当  $z^{(l)} \leq a$  时，为 1；反之为 0。

当  $L \rightarrow \infty$  时，我们将上式写成积分。

$$P(z \leq a) = \frac{\int I(z \leq a) \frac{\tilde{p}(z)}{q(z)} dz}{\int \frac{\tilde{p}(z)}{q(z)} dz} = \frac{\int I(z \leq a) \tilde{p}(z) dz}{\int \tilde{p}(z) dz} \underset{Z_p}{=} \int I(z \leq a) p(z) dz$$

等样本近似  $p(z)$  的结果

同样，当  $p(z)$  是  $q(z)$  接近时，效果会更好，当两者相同时，权值  $w_n = \frac{1}{L}$ 。

如果要求  $p(z)$  的各阶矩，只需使用原始样本，因为  $E[f] \approx \sum_{l=1}^L w_l f(z^{(l)})$

重要采样。

(PRML 中还有一小节一采样与EM算法，可以看一下)

# MCMC

Markov Chain: 时间和状态都是离散的随机过程

Monte Carlo Method: 基于随机采样的近似方法.

研究对象是随机变量

是序3·

## 马尔可夫链

一阶马尔可夫链被定义为一系列随机变量  $Z^{(1)}, \dots, Z^{(M)}$ , 使得下述条件独立性质对于  $m \in \{1, \dots, M-1\}$  成立:

$$P(Z^{(m+1)} | Z^{(1)}, \dots, Z^{(m)}) = P(Z^{(m+1)} | Z^{(m)})$$

那么我们只需要给定初始变量的概率分布  $P(Z^{(1)})$ , 以及后续所有变量的条件分布  $P(Z^{(m+1)} | Z^{(m)})$ , 就可以具体化一个马尔可夫链。我们将条件分布用转移概率 (transition probability) 表示. 即,

$$T_m(Z^{(m)}, Z^{(m+1)}) = P(Z^{(m+1)} | Z^{(m)})$$

△ 表示第  $m$  时刻的转移概率.

如果对于所有的  $m$  转移概率都相同, 该马尔可夫链被称为同质的 (homogeneous). 那么对于某一个时刻的边缘概率. 就可以用上一个时刻的边缘概率表示:

$$P(Z^{(m+1)}) = \int P(Z^{(m)}) \cdot T_m(Z^{(m)}, Z^{(m+1)}) dZ^{(m)} = \sum_{Z^{(m)}} P(Z^{(m+1)} | Z^{(m)}) P(Z^{(m)})$$

如果马尔可夫链在每一步都让这个边缘概率分布保持不变, 那么我们

如果转移概率率是恒等变换，任意分布都不变。

称该马尔可夫链是不变的（注意，一个马尔可夫链可以有多个不变的概率分布）

从1. 对于同质的马尔可夫链（转移概率率为 $T(z', z)$ ）来说，若

$$P^*(z) = \sum_{z'} T(z', z) p^*(z') \quad \text{这里 * 主要为了表示两个概率分布是同一个分布}$$

2. 概率分布  $p^*(z)$  不变，且该分布称为平稳分布或均衡分布。  
*equilibrium*

那么什么样的马尔可夫链具有上述的特点？

石角律  $p(z)$  不变的一个充分必要条件是令转移概率率满足细节平衡  
(detailed balance) 性质，如下：

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

且平稳分布就是  $p^*(z)$ ，证明如下：

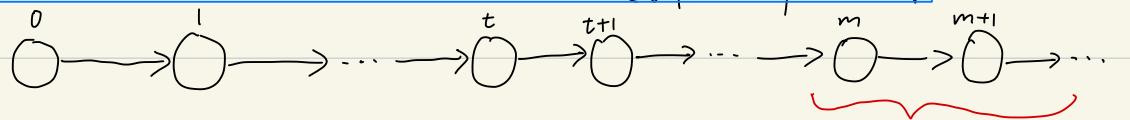
$$\sum_{z'} P^*(z') T(z', z) = \sum_{z'} P^*(z) T(z, z') = P^*(z) \sum_z P(z'|z) = P^*(z)$$

那么，我们要做的，就是使用具有上述特点的马尔可夫链，使最后的平稳分布为采样的目标分布，当  $m \rightarrow \infty$  时， $p(z^{(m)})$  收敛到目标分布，且与初始  $z^{(0)}$  的概率分布无关。这种性质被称为各态历经性 (ergodicity)。

注意：一个具有各态历经性的马尔可夫链只能有一个唯一的平稳分布。且同质的马尔可夫链具有各态历经性。（刚刚说，但PRML上只给出了文献）

PRML中还有用一组基  $B_1, B_2, \dots, B_k$  构建转移概率，可以看看。

下面将展示在板系例中，马尔科夫链一定趋于平稳分布的证明。



令状态转移矩阵（随机矩阵）为：假设状态共有K个取值。

$$Q = \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1K} \\ Q_{21} & Q_{22} & \cdots & Q_{2K} \\ \vdots & \vdots & & \vdots \\ Q_{K1} & Q_{K2} & \cdots & Q_{KK} \end{pmatrix}_{K \times K}, \text{ 其中 } Q_{ij} \text{ 表示从状态 } i \text{ 转移到 } j \text{ 的概率. 即 } T(i, j)$$

$$\text{令 } q^{(t)} = (q^{(t)}(x=1), q^{(t)}(x=2), \dots, q^{(t)}(x=K))_{1 \times K}$$

$$\text{则 } q^{(t+1)}(x=j) = \sum_{i=1}^K q^{(t)}(x=i) Q_{ij} = q^{(t)}(Q_{1j}, Q_{2j}, \dots, Q_{kj})^T$$

$$\text{又 } q^{(t+1)} = \left( \sum_{i=1}^K q^{(t)}(x=i) Q_{i1}, \sum_{i=1}^K q^{(t)}(x=i) Q_{i2}, \dots, \sum_{i=1}^K q^{(t)}(x=i) Q_{iK} \right)$$

$$= q^{(t)} \cdot Q.$$

$$\text{所以 } q^{(t+1)} = q^{(t)} \cdot Q = q^{(t-1)} Q^2 = q^{(0)} \cdot Q^{t+1} \quad (\text{因为同质性, 所以每一时刻均相等})$$

我们将随机矩阵进行相似对角化:  $Q = A \Lambda A^{-1}$ , 且随机矩阵的特征值绝对值  $\leq 1$

(视频中没有解释为什么随机矩阵可以相似对角化且特征值的绝对值  $\leq 1$ )

$$\text{则 } q^{(m)} = q^{(0)} \cdot A \Lambda A^{-1} A \Lambda A^{-1} \cdots A \Lambda A^{-1} = q^{(0)} A \Lambda^m A^{-1}$$

$$q^{(m+1)} = q^{(0)} A \Lambda^{m+1} A^{-1}$$

由于  $|\lambda_i| \leq 1$ . 所以  $\Lambda^m = \Lambda^{m+1}$  ( $\because m$  是足够大时, 只有  $\lambda_i = 1$  的位置被保留, 其余位置均为 0)

所以  $q^{(m)} = q^{(m+1)}$ , 进入平稳分布。

## MH (Metropolis-Hastings) 算法

现在，问题是如何使用 Markov Chain 去进行采样。

关键在于转移概率的构建。

我们知道，随便选取的转移概率不一定满足 detailed balance.

$$\text{即 } p(z) q(z'|z) \neq p(z') q(z|z')$$

如果我们用一个接收率  $A(z, z')$  控制形式，使其能够相等，就可以满足需求。

$$\text{具体表现为: } p(z) q(z'|z) A(z', z) = p(z') q(z|z') A(z, z')$$

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

当然，由于  $p(z)$  可能直接计算，我们用  $\tilde{p}(z)$  来代替，由于存在公式，故可以约去

$$\text{归一化因子，即 } A(z, z') = \min \left\{ 1, \frac{\tilde{p}(z) q(z'|z)}{\tilde{p}(z') q(z|z')} \right\}$$

运用 detailed balance 命定理：

$$\begin{aligned} p(z) q(z'|z) A(z', z) &= \frac{1}{z_p} \min \left\{ \tilde{p}(z) q(z'|z), \tilde{p}(z') q(z|z') \right\} \\ &= \frac{1}{z_p} \min \left\{ \frac{\tilde{p}(z) q(z'|z)}{\tilde{p}(z') q(z|z')}, 1 \right\} \tilde{p}(z') q(z|z') \\ &= p(z') q(z|z') A(z, z') \end{aligned}$$

具体地，MH 算法为：已知  $z^{(t-1)}$ ，则  $z^* \sim q(z^* | z^{(t-1)})$ 。

取  $u \sim U(0,1)$ 。若  $u \leq A(z^*, z^{(t-1)})$ ，则  $z^{(t)} = z^*$ ；反之， $z^{(t)} = z^{(t-1)}$

以概率  $A(z^*, z^{(t-1)})$  搞茎样本。

虽然，这种最后采样出来序列并不是独立的，要想获得独立的样本，

只要设置一个间隔  $m$ ，序列中每隔  $m$  个取一个样本。

(虽然马尔可夫链条件独立性规定了当前样本只依赖前一个样本，似乎依照这个规定，每隔一个样本取一个样本就可以满足独立，但是实际上， $m$  要足够大，才能够保证独立性)。

PRML 中还有说 MH 算法是为了避免随机游走二低效性，这里没怎么看懂，还需要重新看一下。

# 吉布斯采样 (Gibbs Sampling)

吉布斯采样是 MH 算法的一个具体的情形，主要思想是针对多维随机变量，一维一维地进行采样（有点类似坐标上升法）

下面先介绍算法

① 初始化  $\{z_i | i = 1, 2, \dots, M\}$  表示  $Z$  的  $M$  个维度。

② 设置采样周期  $T$ ，对于  $t = 1, 2, \dots, T$ ：

1. 采样  $z_1^{(t+1)} \sim p(z_1 | z_2^{(t)}, z_3^{(t)}, \dots, z_M^{(t)})$

2. 采样  $z_2^{(t+1)} \sim p(z_2 | z_1^{(t+1)}, z_3^{(t)}, \dots, z_M^{(t)})$

:

m. 采样  $z_m^{(t+1)} \sim p(z_m | z_1^{(t+1)}, \dots, z_{m-1}^{(t+1)}, z_{m+1}^{(t+1)}, \dots, z_M^{(t)})$

:

M. 采样  $z_M^{(t+1)} \sim p(z_M | z_1^{(t+1)}, z_2^{(t+1)}, \dots, z_{M-1}^{(t+1)})$

下面，显式地证明，Gibbs Sampling 是 MH 的一个特定的情况：

对于一个 MH 步骤，我们对象是  $Z$ ，条件是  $Z_{-i}$ （表示  $Z$  中除了  $Z_i$  的其他维度）。

则  $p(Z) = p(Z_i | Z_{-i}) p(Z_{-i})$ ,  $q(Z' | Z) = p(Z_i | Z_{-i})$

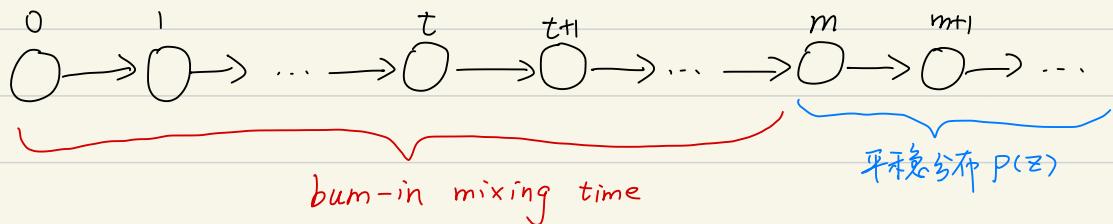
则  $A(Z^*, Z) = \frac{p(Z^* | Z_{-i}^*) p(Z_{-i}^*)}{p(Z_i | Z_{-i}) p(Z_{-i})} \frac{p(Z_i | Z_{-i}^*)}{p(Z_i | Z_{-i})}$  ( $Z_{-i}^* = Z_{-i}$ )

说到底，上面这个等式没太懂。

= 1.  $\Rightarrow$  所以 Gibbs Sampling 就是每个步骤都被接受的 MH

PRML 还有关于 Gibbs 的一些其他讨论，也可以看看。其中有讨论连续样本的依赖性

# 面临的困难 (Problem & Thinking)



MCMC 困难：

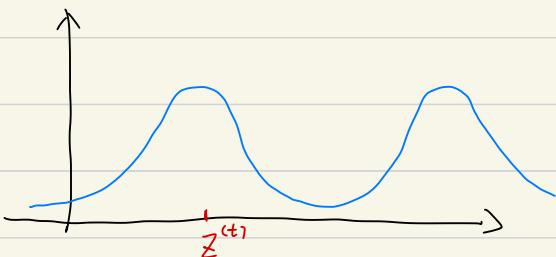
① 理论只保证收敛性，但无法保证何时收敛。  
(也无法检验是否收敛)。

② mixing time 过长  $\leftarrow p(x)$  太复杂  $\leftarrow$  高维以及相关性。  
导致时间成本太大。

③ 样本之间有一定相关性

虽然通过间隔  $m$  可以一定程度上减弱相关性，但是同时也减少了  
可用的样本数。

为何  $p(x)$  太复杂就会导致 mixing time 过长？下面以一个双峰-维高斯分布进  
行解释。



我们看到，假设没我们时刻采样到  $x^{(t)}$  在如图所示位置，那么  $x^{(t)}$  只有很小  
概率可以跨过低概率区域到下一个高  
概率区域（能量都是从高处往低处走，  
概率低，能量大，概率高，能量低，具  
体见能量模型）所以习惯就需要很长  
时间，才有机会跨过去。