

Markov Chain & Monte Carlo (MCMC)

MCMC 是一种基于采样的随机近似方法.



$$P(z|x) \rightarrow E_{z|x}[f(z)] = \int P(z|x) \cdot f(z) dz \approx \frac{1}{N} \sum_{i=1}^N f(z_i)$$

其中 z_i 是从概率分布 $P(z|x)$ 中提取的样本, $z^{(1)}, z^{(2)}, \dots, z^{(N)} \sim P(z|x)$.

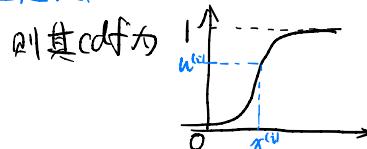
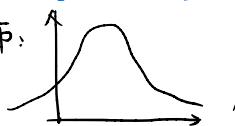
由于 $P(z|x)$ 可能会非常复杂, 因此核心问题是如何从这个后验分布中去采样.

一些基本的采样方法:

1. 概率分布采样.

在计算机中很容易从均匀分布 $U(0,1)$ 中进行采样, 因此, 对于一个分布 $p(x)$, 能否求得其 cdf (分布函数), 则很容易和 $(0,1)$ 产生关联.

例如 $p(x)$ 为高斯分布:



采样步骤:

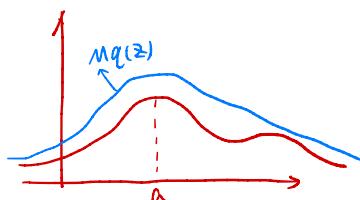
① 生成 0 到 1 之间的伪随机数 $u^{(i)}$, $u^{(i)} \sim U(0,1)$

② 根据 $u^{(i)}$ 和 cdf 的反函数取得采样点 $x^{(i)}$, $x^{(i)} = \text{cdf}^{-1}(u^{(i)})$

缺点: 大部分 pdf 难以求得 cdf.

2. 拒绝采样 (Rejection Sampling)

我们希望能够得到服从分布 $P(z)$ 的样本, 所谓服从分布 $P(z)$ 的样本, 即对于 $P(z)$ 的图像:



设 $q(z)$ 为更简单的分布, 找到一个常数 M , 这个 M 应该尽可能接近 1, 使得对 $\forall z^{(i)}$, 有 $Mq(z^{(i)}) \geq P(z^{(i)})$, $q(z)$ 被称为提议分布.

因此，可以从 $q(z)$ 进行采样来获得样本 $z^{(i)}$ 。

设以为接收率， $\alpha = \frac{p(z^{(i)})}{Mq(z^{(i)})}$, $0 \leq \alpha \leq 1$.

采样步骤如下：

- ① 从提议分布抽样得到样本 $z^{(i)}$, $z^{(i)} \sim q(z)$
- ② 从均匀分布 $U(0,1)$ 抽样, 得到 u , $u \sim U(0,1)$
- ③ 若 $u \leq \alpha$, 则接收样本 $z^{(i)}$, 否则拒绝.

缺点：如果 $p(z)$ 面积占比低，采样效率可能较低

3. 重要性采样

直接对 $E_{p(z)}[f(z)]$ 进行采样。

3) 从提议分布 $q(z)$

$$\begin{aligned} E_{p(z)}[f(z)] &= \int p(z) f(z) dz = \int \frac{p(z)}{q(z)} \cdot q(z) f(z) dz \\ &= \int f(z) \cdot \frac{p(z)}{q(z)} \cdot q(z) dz \\ &\approx \frac{1}{N} \sum_{i=1}^N f(z_i) \cdot \frac{p(z_i)}{q(z_i)} \rightarrow \text{weight}. \end{aligned}$$

$z^{(i)} \sim q(z)$, $i = 1, \dots, N$, DP $z^{(i)}$ 是来自 $q(z)$ 的采样。

重要性采样的变形：Sampling - Importance - Resampling

其原理为：首先和上面一样进行采样，然后在采样出来的 N 个样本中重新采样。

重新采样时按每个样本点的 weight 作为概率值进行采样。

拒绝采样和重要性采样都要求提议分布 $q(z)$: ①尽可能地和 $p(z)$ 接近；② $q(z)$ 是简单，易于采样的，因此，选好 $q(z)$ 是关键。

二. 马尔可夫链.

随机过程研究的是随机变量序列而不是单个随机变量.

马尔可夫链: 时间和状态都是离散的.

齐次马尔可夫链(一阶马尔可夫链)

$X = \{X_1, X_2, \dots, X_t\}$ 其中 X_t 表示 t 时刻的随机变量, 并且每个随机变量的取值空间相同. 如果 X_t 只依赖于 X_{t-1} , 而不依赖于 $\{X_1, X_2, \dots, X_{t-1}\}$, 则称这一性质为 马尔可夫性. 即:

$$P(X_t | X_1, X_2, \dots, X_{t-1}) = P(X_t | X_{t-1}), \quad t=1, 2, \dots$$

具有马尔可夫性的随机序列为 $X = \{X_1, X_2, \dots, X_t\}$ 称为 马尔可夫链 (Markov Chain), 或 马尔可夫过程 (Markov Process), 条件概率分布 $P(X_t | X_{t-1})$ 称为 马尔可夫链的转移概率分布.

当转移概率与 t 无关, 即不同时刻的转移概率均相同, 则称该马尔可夫链为 时间齐次的马尔可夫链 (Time Homogenous Markov Chain):

$$P(X_{t+s} | X_{t-s}) = P(X_t | X_{t-s}), \quad t=1, 2, \dots, s=1, 2, \dots$$

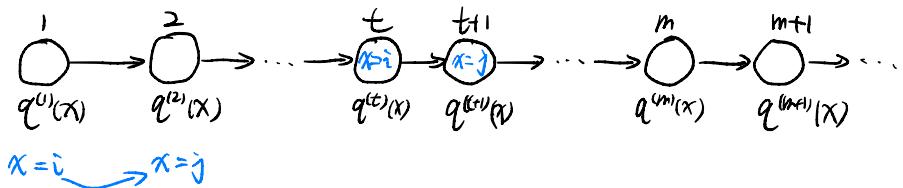
设矩阵 P 为转移矩阵, 马尔可夫链在 t 时刻处于状态 i , 在 $t+1$ 时刻处于状态 j , 则转移概率记作 $p_{ij} = P(X_{t+1}=j | X_t=i)$

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & & \vdots \\ \vdots & & & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix} \quad \text{转移矩阵每行之和为 1.}$$

MCMC 是一策方法, 由于拒绝采样和重要性采样的 $q(x)$ 较难选取, 假设过强, 因此而引出了 MCMC.

★ MCMC 的原理: 借助马尔可夫链的性质: 在经过若干时间步后, 马尔可夫链会收敛到平稳分布, 我们希望这个平稳分布是 $P(x)$ 或接近 $P(x)$, 这样在这个平稳分布上进行采样即可. 因此我们需要构建一个转移矩阵, 使得 Markov Chain 收敛后达到 $P(x)$.

马尔可夫链平稳分布的解释：



设状态空间为 $\{1, 2, \dots, k\}$,

每一个时刻都有对应自己的概率分布，它们都不尽相同，但到达某一时刻 m 后，每个时刻的分布会变成相同的，即达到平稳分布。

在每个时刻下都可以有 k 个状态，状态转移矩阵（随机矩阵）为：

$$Q = \begin{pmatrix} Q_{11} & \cdots & Q_{1k} \\ \vdots & & \vdots \\ Q_{k1} & & Q_{kk} \end{pmatrix}_{k \times k}$$

$$q^{(t+1)}(x=j) = \sum_{i=1}^k q^{(t)}(x=i) \cdot Q_{ij}$$

$$\text{令 } Q^{(t+1)} = (q^{(t+1)}(x=1) \ q^{(t+1)}(x=2) \ \cdots \ q^{(t+1)}(x=k))$$

$$\text{而 } q^{(t+1)}(x=j) = \sum_{i=1}^k q^{(t)}(x=i) \cdot Q_{ij}$$

$$\therefore Q^{(t+1)} = \left(\sum_{i=1}^k q^{(t)}(x=i) \cdot Q_{i1} \ \sum_{i=1}^k q^{(t)}(x=i) \cdot Q_{i2} \ \cdots \ \sum_{i=1}^k q^{(t)}(x=i) \cdot Q_{ik} \right)_{k \times k}$$

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

$$\therefore Q^{(t+1)} = Q^{(t)} \cdot Q = Q^{(t)} \cdot Q^t$$

又转移矩阵的特征值 ≤ 1

$$\therefore Q = A \Lambda A^{-1}, \quad \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, |\lambda_i| \leq 1$$

不妨设只有一个 λ 为1，则：

$$q^{(t+1)} = Q^{(t)} \cdot A \Lambda A^{-1} \cdot A \Lambda A^{-1} \cdots A \Lambda A^{-1}$$

$$= Q^{(t)} \cdot A \Lambda^t A^{-1}$$

存在足够的 m ，使得 $\Lambda^m = \begin{pmatrix} 0 & & \\ & \ddots & \\ & & 0 \end{pmatrix}$

$$q^{(m+1)} = Q^{(t)} \cdot A \Lambda^m A^{-1}$$

$$q^{(m+2)} = Q^{(t)} \cdot A \Lambda^m A^{-1} = Q^{(t)} \cdot A \Lambda^{m+1} A^{-1} = Q^{(t)} \cdot A \Lambda^m A^{-1} = Q^{(m+1)}$$

因此当 $t \geq m$ 时， $q^{(m+1)} = q^{(m+2)} = \cdots = q^{(\infty)} = \cdots$

基于这个性质，我们就可以使马尔可夫链达到一个逼近 $\pi(x)$ 的平稳分布，对平稳分布采样即可。

如何构造一个这样的马氏链？

(~~粗略地~~) Detailed Balance = $\pi(x) P(x \rightarrow x^*) = \pi(x^*) P(x^* \rightarrow x)$ x 和 x^* 是两个不同的状态

它是平稳分布的充分必要条件，即如果一个分布 π 满足 DB，则它是平稳分布。

$$\begin{aligned} \int \pi(x) P(x \rightarrow x^*) dx &= \int \pi(x^*) P(x^* \rightarrow x) dx \\ &= \pi(x^*) \int P(x^* \rightarrow x) dx = \pi(x^*) \end{aligned}$$

因此，如果分布 π 不能够满足 Detailed Balance 条件，那么这个分布一定是平稳分布。

MH 算法 (Metropolis - Hastings)

由细致平衡条件可知，我们需要知道概率转移矩阵，但这个转移矩阵是无法直接求得的，因此，我们选取一个转移矩阵 Q 作为提议矩阵，此时：

$$p(z) \cdot Q(z \mapsto z^*) \neq p(z^*) \cdot Q(z^* \mapsto z)$$

在等式两边同乘一个因子，使得：

$$\begin{aligned} p(z) \cdot \underbrace{Q(z \mapsto z^*)}_{p(z)} \cdot \alpha(z, z^*) &= p(z^*) \cdot \underbrace{Q(z^* \mapsto z)}_{\alpha(z^*, z)} \\ p(z) \cdot p(z \mapsto z^*) &= p(z^*) \cdot p(z^* \mapsto z) \end{aligned}$$

其中 $\alpha(z, z^*) = \min\left(1, \frac{p(z^*) \cdot Q(z^* \mapsto z)}{p(z) \cdot Q(z \mapsto z^*)}\right)$

证明：

$$\begin{aligned} p(z) \cdot Q(z \mapsto z^*) \cdot \alpha(z, z^*) &= p(z) \cdot Q(z \mapsto z^*) \cdot \min\left(1, \frac{p(z^*) \cdot Q(z^* \mapsto z)}{p(z) \cdot Q(z \mapsto z^*)}\right) \\ &= \min(p(z) \cdot Q(z \mapsto z^*), p(z^*) \cdot Q(z^* \mapsto z)) \\ &= p(z^*) \cdot Q(z^* \mapsto z) \cdot \underbrace{\min\left(1, \frac{p(z) \cdot Q(z \mapsto z^*)}{p(z^*) \cdot Q(z^* \mapsto z)}\right)}_{\alpha(z^*, z)} \\ &= p(z^*) \cdot Q(z^* \mapsto z) \cdot \alpha(z^*, z) \end{aligned}$$

Metropolis - Hastings:

$$u \sim U(0, 1)$$

$$z^* \sim Q(z | z^{(i-1)})$$

$$\alpha = \min\left(1, \frac{p(z^*) \cdot Q(z^* \mapsto z)}{p(z) \cdot Q(z \mapsto z^*)}\right)$$

$$\text{if } u \leq \alpha, z^{(i)} = z^*$$

$$\text{else } z^{(i)} = z^{(i-1)}$$

一个常见的 Markov Chain 转移矩阵 Q 通过一定的接受率就可以得到目标转移矩阵 P 。

Gibbs 算法.

假设 $p(z)$ 维度很高, $p(z) = p(z_1, z_2, \dots, z_m)$

$$z_i \sim p(z_i | z_{-i})$$

$$\rightarrow z_1, z_2, \dots, z_{i-1}, z_{i+1}, \dots, z_m$$

假设 $p(z) = p(z_1, z_2, z_3)$ 维度为 3 维.

采样步骤:

① 定义初始值: $z_1^{(0)}, z_2^{(0)}, z_3^{(0)}$

② $t+1$ 时刻:

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

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

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

$$\frac{p(z^*) \cdot Q(z^* \mapsto z_1)}{p(z) \cdot Q(z \mapsto z^*)}$$

$$= \frac{p(z_1^* | z_{-1}) \cdot p(z_{-1}^*) \cdot p(z | z^*)}{p(z_1 | z_{-1}) \cdot p(z_{-1}) \cdot p(z^* | z)}$$

由于 z_1^* 对于 z_{-1} , 都是以 $-i$ 为条件进行采样, 因此:

$$= \frac{p(z_1^* | z_{-1}) \cdot p(z_{-1}^*) \cdot p(z_i | z_{-i}^*)}{p(z_1 | z_{-1}) \cdot p(z_{-1}) \cdot p(z_i^* | z_{-i})}$$

又由于 z_i^* 和 z_i 都是剩余分量, 二者相等, 则上式为 1.

即接受率为 1.

MCMC 回顾

采样的动机？

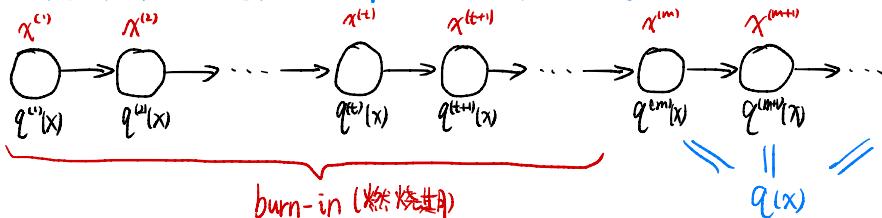
- ① 样本本身就是常见的任务
- ② 求和或求积分

什么是好的样本？

- ① 样本趋向于高概率区域
- ② 样本之间相互独立

样本是困难的（高维带来的）

平稳分布只与转移矩阵有关，与初始状态无关。



达到平稳分布之前的时期叫做燃烧期，所花费的时间叫做混合时间 (mixing time)

MCMC 问题：

- ① 理论只保证收敛性，但无法知道何时收敛。
- ② mixing time 可能由于 $p(x)$ 的复杂、维度高及维度之间的相关性而过长。
- ③ 由于 Markov Chain 本身特性导致样本之间有一定的相关性。