Markov Chain Monte Carlo 01 Sampling Method

Chen Gong

30 December 2019

其实在之前的 Inference Variational 那一节中，我们讲到过一些有关于 Markov Chain Monte Carlo (MCMC) 的知识。也就是我们有一些数据 X，看到这些数据 X，并且有一些隐变量 Z，我们给隐变量一些先验，根据观测数据来推后验知识，也就是 $P(Z|X)$。

但是，很不幸的是 $P(Z|X)$ 的计算非常的复杂，我们大致采用两种思路来解决这个问题，也就是精确推断和近似推断。精确推断无法达到我们想要的结果时，就会采用近似推断的方法。而近似推断中我们又可以分成两大类，即为确定性近似 (VI) 和随机近似 (MCMC)。

Monte Carlo Method 是一种基于采样的随机近似算法。我们的目标是求解后验概率 $P(Z|X)$，其中 Z 为 Latent data，X 为 Observed data。知道分布以后，我们通常的目标是求解：

$$\mathbb{E}_{Z|X}[f(Z)] = \int_Z P(Z|X)f(Z)dZ \approx \frac{1}{N}\sum_{i=1}^{N} f(z_i) \tag{1}$$

然后，问题马上就来了，我们知道了后验分布 $P(Z|X)$，怎么去采样呢？也就是如何通过采样得到 $z^{(1)}, z^{(2)}, \cdots, z^{(N)} \sim P(Z|X)$。那么，我们这一节将要主要介绍三种采样方法，概率分布采样，拒绝采样和重要性采样。

1 概率分布采样

我第一次看到这个概念是在 Distributional Reinforcement Learning 中的 Wesserstein Metric 中。当时，真的把我看得我一脸懵逼，而且作者并没有提到概率分布采样。还有有的文章中，经常省写 c.d.f (概率分布函数)，p.d.f (概率密度函数)，i.i.d (独立同分布)。我觉得我这里有必要提一下。

为什么要有概率分布采样呢？因为我们直接根据概率分布来进行采样非常的复杂。如果我们知道概率分布的具体形式吗？我们可以直接求得概率累积的概率分布函数。由于概率分布函数的值一定是 $[0, 1]$ 之间的。所以，我们可以在均匀概率密度分布 $U(0, 1)$ 上采样，得到 $u^{(i)} \sim U(0, 1)$。然后求 $x^{(i)} \sim cdf^{-1}(u^{(i)})$ 就可以计算得到我们想要的结果。这样就可以采样得到 $\{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$N 个样本点。

虽然，理论上这个方法好像很有效，但是实际上很多情况我们都根本不知道 p.d.f 的具体表现形式。就算知道，很多时候 c.d.f 也并不是那么的好求。所以很多情况下，概率分布采样并没有那么的好求。

2 拒绝采样 (Rejection Sampling)

由于对目标分布 $p(Z)$ 的采样非常的困难，所以我们可以对一个比较简单的分布 $q(Z)$ 进行采样来辅助采样。那么我们具体做法怎么办呢？我们可以设定一个 proposal distribution：$q(Z)$。对于 $\forall z_i$，保

证 $M \cdot q(z^i) \geq p(z^i)$，那么我们为什么要引入 M 呢？这是因为 $\int_Z P(Z)dZ = \int_Z q(Z)dZ = 1$。要使 $q(z^i) \geq p(z^i)$ 是几乎不可能成立的。为了方便描述，我们画图来说明一下：

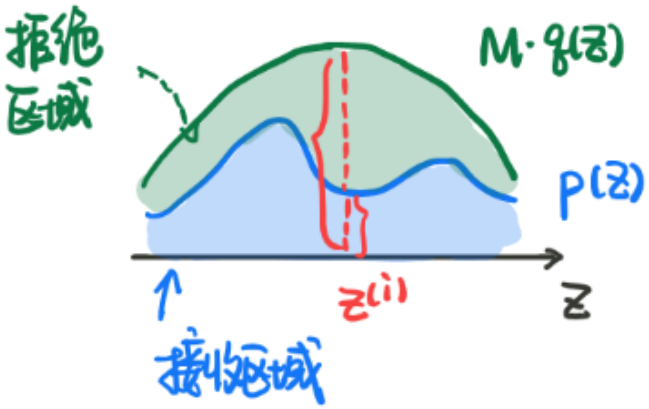

图 1: Rejection Sampling 示意图

在这里我们需要定义一个接受率：$\alpha = \frac{P(z^{(i)})}{M \cdot q(z^{(i)})}$，很显然 $0 \leq \alpha \leq 1$。这个实际就是上图中绿色的部分。

我们来看看具体的步骤：

(1) 首先进行采样 $z^{(i)} \sim q(z)$。

(2)$u \sim U(0,1)$；如果 $u \leq \alpha$，我们就接收 $z^{(i)}$，不然我们就拒绝。

所以，绿色的部分就被我们称为拒绝区域，就是这样来的，所以这个采样方法就是拒绝采样。同样这样的采样方法也有缺点。如果 $M \cdot q(z)$ 比 $p(z)$ 大很多的话，那么我们的采样老是是失败的，这就涉及到一个采样效率低下的问题。而当 $M \cdot q(z) = p(z)$ 的时候，$\alpha = 1$，我们每次采样的结果都是接受的。但是，实际上 $p(z)$ 的分布形式非常的复杂，我们根本就没有办法来得到那么准确的结果，特别是采样 cost 非常高的话，经常性的采样失败带来的损失是很大的。

3 重要性采样 (Importance Sampling)

重要性采样在我们的强化学习 (PPO) 中的应用非常的多。重要性采样并不是直接对概率进行采样，而是对概率分布的期望进行采样。也就是：

$$\begin{aligned} \mathbb{E}_{p(z)}[f(z)] = \int p(z)f(z)dz &= \int \frac{p(z)}{q(z)}q(z)f(z)dz \\ &= \int f(z)\frac{p(z)}{q(z)}q(z)dz \\ &\approx \frac{1}{N}\sum_{i=1}^{N} f(z_i)\frac{p(z_i)}{q(z_i)} \\ & z_i \sim q(z),\ i = 1, 2, \cdots, N \end{aligned} \tag{2}$$

而这里的 $\frac{p(z_i)}{q(z_i)}$ 也就是 Weight，用来平衡不同的概率密度值之间的差距。同样重要性采样也可能会出现一些问题，就是两个分布之间的差距太大了话，总是采样采不到重要的样本，采的可能都是实

际分布概率值小的部分。也就是采样效率不均匀的问题。在这个基础上，我们进一步提出了 Sampling Importance Resampling。

3.1 重要性重采样 (Sampling Importance Resampling)

经过重要性采样后，我们得到了 N 个样本点，以及对应的权重。那么我用权重来作为采样的概率，重新测采样出 N 个样本。也就是如下图所示：

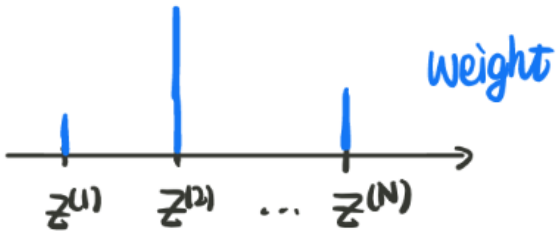

图 2: Sampling Importance Resampling 示意图

通过二次采样可以降低采样不平衡的问题。至于为什么呢？大家想一想，我在这里表达一下自己的看法。$\frac{p(z_i)}{q(z_i)}$ 是 Weight，如果 Weight 比较大的话，说明 $p(z_i)$ 比较大而 $q(z_i)$ 比较的小，也就是我们通过 $q(z_i)$ 采出来的数量比较少。那么我们按权重再来采一次，就可以增加采到重要性样本的概率，成功的弥补了重要性采样带来的缺陷，有效的弥补采样不均衡的问题。

Markov Chain Monte Carlo 02 Markov Chain

Chen Gong

31 December 2019

在上一小节中，我们描述了三种采样方法，也就是概率分布采样法，拒绝采样法和重要性采样法。这三种采样方法在高维情况下的采样效率很低，所以我们需要另找方法。

1 基础概念介绍

首先我们要明确什么是 Random Process，也就是它研究的变量是一个随机变量的序列 $\{x_t\}$。通俗的说就是，随机过程就是一个序列，而这个序列中的每一个元素都是一个随机变量。

而 Markov Chain 就是一个特殊的随机过程，它的时间和状态都是离散的。并且，Markov Chain 需要满足 Markov 性质，也就是未来和过去是无关的。我们用数学的语言表达就是：

$$P(x_{t+1} = x|x_1, x_2, \cdots, x_t) = P(x_{t+1}|x_1, x_2, \cdots, x_{t-m}) \tag{1}$$

上述公式就是一个 m 阶马尔可夫性质。当 $m = 0$ 时，我们就得到了齐次 (一阶) 马尔可夫链，也就是满足：

$$P(x_{t+1} = x|x_1, x_2, \cdots, x_t) = P(x_{t+1}|x_t) \tag{2}$$

而 $P(x_{t+1}|x_t)$ 这个概率我们用什么来表达呢?我们定义 $\mathcal{P}$ 为一个转移矩阵 $[P_{ij}]$,而 $P_{ij} = P(x_{t+1} = j|x_t = i)$。

2 平稳分布 (Stationary Distribution)

一个 Markov Chain 可以用下图来表示：

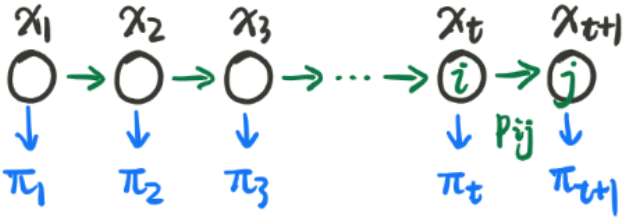

图 1: Markov Chain Model 示意图

此图就是一个时间序列，x_i 就表示在第 i 时刻的状态，而每一个状态都是一个随机变量。而 π_i 描述的就是第 i 个随机变量的分布。对于一个马氏链来讲，它在第 $t+1$ 时刻的概率分布，可以被我们表

达为：

$$\pi_{t+1}(x^*) = \int \pi_t(x) \cdot P(x \mapsto x^*) dx \tag{3}$$

熟悉强化学习的同学就会觉得这个公式非常的熟悉。通俗的讲，他实际上就是在 $t+1$ 时刻所有可能转移到状态 x^* 的概率的和。那么什么是随机分布呢?

假如这里存在一个 π，这里的 π 和前面的 π_t 和 π_{t+1} 都没有一毛钱关系。假如 π 是一个概率分布，那么它可以被我们写成一个无限维向量的形式：

$$\pi = [\pi(1), \pi(2), \cdots, \pi(t), \cdots], \qquad \sum_{i=1}^{\infty} \pi(i) = 1 \tag{4}$$

如果存在式 (4) 使得公式成立：

$$\pi(x^*) = \int \pi(x) \cdot P(x \mapsto x^*) dx \tag{5}$$

我们就称 $\{\pi(k)\}_{k=1}^{\infty}$ 是马氏链 $\{x_k\}$ 的平稳分布。看了数学的描述我相信大部分同学还是不懂这个平稳分布时是个什么东西？用通俗的话讲就是，对于一个马氏链来说，每个时刻都符合一个概率分布，如果每一个时刻的概率的分布都是一样的都是 $\pi(k)$，那么我们就可以称这个马氏链满足一个平稳分布。

那么下一个问题就是我们为什么要引入平稳分布呢？其实，我们想要去求的这个 $P(Z)$，可以被我们看成是一个平稳分布 $\pi(k)$，那么我们就可以通过构建一系列的 $\{x_1, x_2, \cdots, x_t, \cdots\}$ 的马氏链，让它来逼近这个平稳分布。那么我们构建这样的一个马氏链，包括随机变量和转移矩阵，如果它满足平稳分布的条件，确实是可以收敛到平稳分布的。那么，我就可以让构建出来的这个马氏链收敛到平稳分布来求得 $P(Z)$。

既然，已经知道了什么是平稳分布了，那么下一个问题就是，我们需要知道什么样的分布可以称为平稳分布，也就是我们怎样才能构建出一个马氏链让它收敛到一个平稳分布。这里我们需要引入一个条件，也就是 Detailed Balance：

$$\pi(x) \cdot P(x \mapsto x^*) = \pi(x^*) \cdot P(x^* \mapsto x) \tag{6}$$

大家直观的来想想这个公式，为什么满足它就满足了是一个平稳分布呢？其实并不难想到，对于任意两个状态之间，使用概率分布称为转移概率得到的结果都是可逆的，那么这两个状态之间的分布一定是一样的。而说如果一个分布，满足 Detailed Balance 那么它一定可以是一个平稳分布，但是反过来并不能成立。而证明过程也不难，如下所示：

$$\begin{aligned} \int \pi(x) \cdot P(x \mapsto x^*) dx &= \int \pi(x^*) \cdot P(x^* \mapsto x) dx \\ &= \pi(x^*) \underbrace{\int P(x^* \mapsto x)\, dx}_{\sum_{j=1}^{\infty} P_{ij} = 1} \\ &= \pi(x^*) \end{aligned} \tag{7}$$

这样的话，我们就可以吧不用从定义上来证明一个随机过程是马尔可夫链，直接看它满不满足 Detailed Balance 就可以了。而且这个公式中，π 是平稳分布，$P(x \mapsto x^*)$ 是马尔可夫链的状态转移概率，这样就成功的将平稳分布和马尔可夫链结合在了一起。

Markov Chain Monte Carlo 03 Metropolis Hastings Sampling

Chen Gong

01 January 2020

上一节中我们讲解了 Detailed Balance，这是平稳分布的充分必要条件。Detailed Balance 为：

$$\pi(x)P(x \mapsto x^*) = \pi(x^*)P(x^* \mapsto x) \tag{1}$$

这里的 $P(x \mapsto x^*)$ 实际上就是条件概率 $P(z^*|x^*)$，这样写只是便于理解。

首先，我们需要明确一点，我们想要求的是后验概率分布 $P(Z)$，也就是我们推断问题的核心目标。我们求 $P(Z)$ 主要是为了求在 $P(Z)$ 概率分布下的一个相关函数的期望，也就是：

$$\mathbb{E}_{P(Z)}[f(Z)] \approx \frac{1}{N}\sum_{i=1}^{N} f(z^{(i)}) \tag{2}$$

而我们是通过采样来得到 $P(Z) \sim \{z^{(1)}, z^{(2)}, \cdots, z^{(N)}\}$ 样本点。$\pi(x)$ 是最终的平稳分布，可以看成我们这里的 $P(Z)$，下面的问题就是求出概率转移矩阵 P_{ij}，才能满足 Detailed Balance 条件。知道了上面的条件以后，我们每次这样进行采样，$x_1 \sim P(x|x_1)$，$x_2 \sim P(x|x_1)$，$x_3 \sim P(x|x_2)$，$\cdots$，x_N。最终就可以得到我们想要的 N 个样本。

1 Proposal Matrix

那我们怎么来找这个状态转移矩阵 P_{ij} 呢？首先我们可以随机一个状态转移矩阵 Q_{ij}，也就是 Proposal Matrix。

那么肯定是：

$$P(Z)Q(Z^*|Z) \neq P(Z^*)Q(Z|Z^*) \tag{3}$$

那么我们就要想办法找到 Q_{ij} 使得：

$$P(Z)Q(Z^*|Z) = P(Z^*)Q(Z|Z^*) \tag{4}$$

那么，我们怎么来解决这个问题呢？我们可以在左右两边乘上一个因子来解决这个问题。也就是，

$$P(Z)\underbrace{Q(Z^*|Z)\alpha(Z^*, Z)}_{P(Z \mapsto Z^*)} = P(Z^*)\underbrace{Q(Z|Z^*)\alpha(Z, Z^*)}_{P(Z^* \mapsto Z)} \tag{5}$$

而 $\alpha(Z, Z^*)$ 定义为接收率，大小为：

$$\alpha(Z, Z^*) = \min\left(1, \frac{P(Z^*)Q(Z|Z^*)}{P(Z)Q(Z^*|Z)}\right) \tag{6}$$

这样定义就行了？就可以满足 Detailed Balance 吗？我们可以证明一下，

$$\begin{aligned} P(Z)Q(Z^*|Z)\alpha(Z,Z^*) =& P(Z)Q(Z^*|Z)\min\left(1, \frac{P(Z^*)Q(Z|Z^*)}{P(Z)Q(Z^*|Z)}\right) \\ =& \min(P(Z)Q(Z^*|Z), P(Z^*)Q(Z|Z^*)) \\ =& P(Z^*)Q(Z|Z^*)\min\left(\frac{P(Z)Q(Z^*|Z)}{P(Z^*)Q(Z|Z^*)}, 1\right) \\ =& P(Z^*)Q(Z|Z^*)\alpha(Z^*,Z) \end{aligned} \tag{7}$$

那么我们就成功的证明了：

$$P(Z)\underbrace{Q(Z^*|Z)\alpha(Z^*,Z)}_{P(Z\mapsto Z^*)} = P(Z^*)\underbrace{Q(Z|Z^*)\alpha(Z,Z^*)}_{P(Z^*\mapsto Z)} \tag{8}$$

所以，$P(Z)$ 在转移矩阵 $Q(Z|Z^*)\alpha(Z^*,Z)$ 下是一个平稳分布，也就是一个马尔可夫链，通过在这个马尔可夫链中采样就可以得到我们的相应的数据样本点了。实际上这就是大名鼎鼎的 Metropolis-Hastings 采样法。

2 Metropolis-Hastings Sampling

第一步，我们从一个均匀分布中进行采样，$u \sim U(0,1)$；

第二步，从 $Q(Z|Z^{(i-1)})$ 中进行采样得到样本点 Z^*；

第三步，计算接受率，$\alpha = \min\left(1, \frac{P(Z^*)Q(Z|Z^*)}{P(Z)Q(Z^*|Z)}\right)$。注意，这里的 $P(Z) = \frac{\hat{P}(Z)}{Z_p}$。其中 Z_p 指的是归一化因子，几乎非常难以计算，所以一般是未知的。而 $\hat{P}(Z) = \text{likelihood} \times \text{prior}$。所以这里的 $P(Z)$ 和 $P(Z^*)$ 就是 $\hat{P}(Z)$ 和 $\hat{P}(Z^*)$。由于归一化因子被抵消了，干脆就直接写成了 $P(Z)$ 和 $P(Z^*)$。

第四步，如果 $u \leq \alpha$ 时 $Z^i = Z^*$，不然 $Z^i = Z^{(i-1)}$。

这样执行了 N 次，就可以得到 $\{Z^{(1)}, Z^{(2)}, \cdots, Z^{(N)}\}$ 个样本点。

Markov Chain Monte Carlo 04 Gibbs Sampling

Chen Gong

02 January 2020

如果我们要向一个高维的分布 $P(Z) = P(Z_1, Z_2, \cdots, Z_N)$ 中进行采样。那么我们怎么来进行采样呢？我们的思想就是一维一维的来，在对每一维进行采样的时候固定住其他的维度，这就是 Gibbs Sampling。

我们首先规定一个 z_{-i} 是去除 z_i 后的序列，$\{z_1, z_2, \cdots, z_{i-1}, z_{i+1}, \cdots, z_N\}$。

1 A Example

假设 t 时刻，我们获得的样本为 $z_1^{(t)}, z_2^{(t)}, z_3^{(t)}$。

那么 $t+1$ 时刻，我们的采样顺序为：

$$\begin{aligned} z_1^{(t+1)} &\sim P(z_1|z_2^{(t)}, z_3^{(t)}) \\ z_2^{(t+1)} &\sim P(z_2|z_1^{(t+1)}, z_3^{(t)}) \\ z_3^{(t+1)} &\sim P(z_3|z_1^{(t+1)}, z_2^{(t+1)}) \end{aligned} \tag{1}$$

从这个例子中，我们应该可以大致理解固定其他的维度然后进行一维一维采样的意思了。而实际上 Gibbs 是一种特殊的 MH 采样，为什么呢？我们来证明一下。

2 接受率 α 的计算

我们首先回顾一下，MH 采样的方法。我们的目的是从 $Q(Z|Z^{(t)})$ 中采样获得 Z^*，然后计算接受率

$$\alpha = \min\left(1, \frac{P(Z^*)Q(Z|Z^*)}{P(Z)Q(Z^*|Z)}\right) \tag{2}$$

首先我们来看 $Q(Z|Z^{(t)})$：

$$Q(Z|Z^{(t)}) = Q(Z_i, Z_{-i}|Z_i^{(t)}, Z_{-i}^{(t)}) \tag{3}$$

假设我们现在是在对第 i 维进行采样，我们只要关注 $P(Z_i^*|Z_{-i})$。所以，我们可以得到：$Q(Z|Z^{(t)}) = P(Z_i^*|Z_{-i}^{(t)})$。

已经成功的将 $Q(Z|Z^{(t)})$ 做了等价转换以后。那么我们想要求的 α 可以被我们成功的转换成如下的形式：

$$\alpha = \min\left(1, \frac{P(Z_i^*|Z_{-i}^*)P(Z_{-i}^*)P(Z_i|Z_{-i}^*)}{P(Z_i|Z_{-i})P(Z_{-i})P(Z_i^*|Z_{-i})}\right) \tag{4}$$

计算到了这里，我们还是不好进行计算，上面和下面好像还是不好消除。如果我们可以得到 Z_{-i}^{*} 和 Z_{-i} 之间的关系就好了。下面我们会得出一个重要的结论来帮助我们计算 α 的具体值。首先我们来举一个例子：

那么假设当 $t=1$ 的时刻，有一个样本为：$Z_1^{(1)}, Z_2^{(1)}, Z_3^{(1)}$。

当 $t=2$ 的时刻，我们假设先对第一维进行采样就可以得到：$Z_1^{(2)}, Z_2^{(1)}, Z_3^{(1)}$。

很显然 $Z_2^{(1)}, Z_3^{(1)}$ 根本没有发生变化。我们可以得到 $Z_{-1} = Z_{-1}^{*}$。也就是在 Gibbs 采样时，采样前后只关注于一个维度，其他的维度我们都没有关注到。所以就可以得到结论：

$$Z_{-i} = Z_{-i}^{*} \tag{5}$$

那么，我们把这个结论代入到公式 (4) 中，就可以得到：

$$\alpha = \min\left(1, \frac{P(Z_i^{*}|Z_{-i}^{*})P(Z_{-i}^{*})P(Z_i|Z_{-i}^{*})}{P(Z_i|Z_{-i}^{*})P(Z_{-i}^{*})P(Z_i^{*}|Z_{-i}^{*})}\right) = 1 \tag{6}$$

那么计算出接受率为 1，也就是每次都必定被接受。所以，每次从 $Q(Z|Z^{(t)}) = P(Z_i^{*}|Z_{-i})$ 中进行采样得到 Z_i^{*} 即可，一维一维的进行采样就可以采到整个高维的分布，各个维度上的样本。

所以，解释到了这里，大家基本就可以知道 Gibbs Samplings 是 $\alpha = 1$ 的 MH Sampling 的意义了。在 Gibbs Sampling 中 $\alpha = 1$，而且状态转移矩阵 $Q(Z|Z^{(t)}) = P(Z_i^{*}|Z_{-i}^{(t)})$，所以 Gibbs Sampling 就是把目标分布 P 对应的条件概率当作状态转移分布 Q。

这里我们需要额外提醒一下，使用 Gibbs Sampling 是有使用前提的，也就是固定其他维度后的一维分布时方便进行采样的，如果固定其他维度的时候得到的一维分布仍然是非常难进行采样的，那么使用 Gibbs Sampling 也是没有用的。

Markov Chain Monte Carlo 05 Sampling

Cheh Gong

03 January 2020

在前面的章节中，我们已经基本介绍了 Markov Chain Monte Carlo Sampling 的基本概念，基本思路和主要方法。那么这一小节中，我们将主要来介绍一下，什么是采样？我们为什么而采样？什么样的样本是好的样本？以及我们采样中主要会遇到哪些困难？

1 采样的动机

这一小节的目的就是我们要知道什么是采样的动机，我们为什么而采样？

1. 首先第一点很简单，采样本身就是发出常见的任务，我们机器学习中经常需要进行采样来完成各种各样的任务。如果从一个 $P(X)$ 中采出一堆样本。

2. 求和求积分。包括大名鼎鼎的 Monte Carlo 算法。我们求 $P(X)$ 主要是为了求在 $P(X)$ 概率分布下的一个相关函数的期望，也就是：

$$\int P(x)f(x)dx = \mathbb{E}_{P(X)}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \tag{1}$$

而我们是通过采样来得到 $P(X) \sim \{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$ 样本点。

2 什么样的是好样本

既然，我们知道了采样的目的和动机，下一个问题就自然是，同样是采样，什么样的样本就是好样本呢？或者说是采样的效率更高一些。

1. 首先样本的分布肯定是要趋向于原始的目标分布吧，也就是说样本要趋向于高概率选择区域。或者是说，采出来的样本出现的概率和实际的目标分布的概率保持一致。

2. 样本和样本之间是相互独立的。这个就没有那么直观了。大家想一想就知道了，如果我采出来的一堆样本之间都差不多，那么就算采出来了趋向于高概率选择区域的样本，那采样效率太低了，样本中反映的信息量太有限了。

3 实际采样中的困难

实际采样中，采样时困难的，为什么呢？我们这里主要介绍两点：

1. **Partation function is intractable.** 我们的后验分布往往被写成 $P(X) = \frac{1}{Z}\hat{P}(X)$，上面这个 $\hat{P}(X)$ 都比较好求，就是等于 Likelihood $\times$ Prior。而 Z 就是我们要求的归一化常数，它非常的难

以计算，$Z = \int \hat{P}(X)dX$，这几乎就是不可计算的。所以，有很多采样方法就是想要跳过求 $P(X)$ 的过程，来从一个近似的分布中进行采样，当然这个近似的分布采样要比原分布简单。比如：Rejection Sampling 和 Importance Sampling。

2. **The curse of high dimension**. 如果样本空间 $\mathcal{X} \in \mathbb{R}^p$，每个维度都有 K 个状态的话。那么总的样本空间就有 K^p 的状态。要知道那个状态的概率高，就必须要遍历整个样本空间，不然就不知道哪个样本的概率高，如果状态的数量是这样指数型增长的话，全看一遍之后进行采样时不可能的。所以，直接采样的方法是不可行的。

4 采样方法

Rejection Sampling 和 Importance Sampling，都是借助一个 $Q(x)$ 去逼近目标分布 $P(x)$，通过从 $Q(x)$ 中进行采样来达到在 $P(x)$ 中采样的目的，而且在 $Q(x)$ 中采样比较简单。当时如果 $Q(x)$ 和 $P(x)$ 直接的差距太大的话，采样效率会变得很低。

而 MCMC 方法，我们主要介绍了 MH Sampling 和 Gibbs Sampling，我们主要是通过构建一个马氏链去逼近目标分布，具体的描述将在下一节中展开描述。

Markov Chain Monte Carlo 06 Method of MCMC

Chen Gong

04 January 2020

这一小节主要是对前面的补充，希望可以详细的介绍一下 MCMC 原理，将前面的知识点可以顺利的串起来。MCMC 采样中，我们借助了一条马氏链，马氏链的性质，经过若干步以后会收敛到一个平稳分布。马尔可夫链的组成可以大致分成两个部分：

1. 状态空间：$\{1, 2, 3, \cdots, k\}$；
2. 状态转移空间 $Q = [Q_{ij}]_{k\times k}$。

马尔可夫链的模型可以被我们表达为：

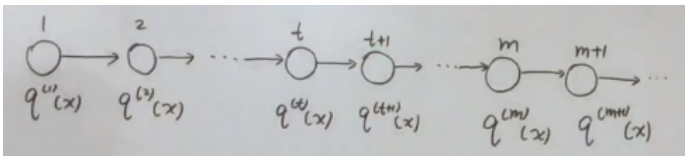

图 1: 马尔可夫链模型抽象图

每一个时间点有一个状态分布，表示当前时间点位于某个状态的概率分布情况，我们表示为 $q^{(t)}(x)$。如果，是在 $t = 1$ 的时间节点，状态的概率分布为 $q^{(1)}(x)$，我们可以用下列表来描述：

x	1	2	3	$\cdots$	k
$q^{(1)}(x)$	q_1^1	q_1^2	q_1^3	$\cdots$	q_1^k

我们假设在 $t = m$ 时刻之后到达了平稳分布状态，那么我们就可以得到：$q^{(m)} = q^{(m+1)} = q^{(m+2)}$。这时的平稳分布就是我们想要的目标分布。相邻时间节点之间的状态转移矩阵为：

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1k} \\ Q_{21} & Q_{22} & \cdots & Q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{k1} & Q_{k2} & \cdots & Q_{kk} \end{bmatrix}_{k\times k} \tag{1}$$

状态转移矩阵描述的是，$Q_{ij} = Q(x^2 = j|x^1 = i)$。描述的是从一个状态转移到另外一个状态的概率。所以，状态转移矩阵的每一行 i 表示为目前状态是 i 时，到其他状态的概率，那么必然有 $\sum_{k=1}^{k} Q_{ik} = 1$。实际上了解强化学习的同学，对于这些概率应该是非常的熟练了，这些都是强化学习的基础。

1 Markov Chain 收敛性介绍

在这一小节中，我们将详细的介绍一下，Markov Chain 中状态转移的过程。并将证明在 Markov Chain 随着迭代一定会收敛到一个平稳分布。

1.1 Markov Chain 状态转移计算

假设在 $t+1$ 时刻，状态是 $x=j$，那么它的分布为所有可能转移到这个状态的概率 i 乘以这个状态的分布 $q^{(t)}(x=i)$，我们用公式表达就是：

$$q^{(t+1)}(x=j)=\sum_{i=1}^{k}q^{(t)}(x=i)Q_{ij} \tag{2}$$

那么，这仅仅是当 $x=j$ 时概率，实际上在 $t+1$ 时刻，可能出现的状态有 k 个，那么 q^{t+1} 的分布，就是将转移到各个状态的概率分别计算出来，也就是如下所示：

$$q^{(t+1)}=\begin{bmatrix}q^{(t+1)}(x=1) & q^{(t+1)}(x=2) & q^{(t+1)}(x=3) & \cdots & q^{(t+1)}(x=k)\end{bmatrix}_{1\times k} \tag{3}$$

而，

$$q^{(t+1)}(x=j)=\sum_{i=1}^{k}q^{(t)}(x=i)Q_{ij} \tag{4}$$

那么，$q^{(t+1)}$ 可以被我们表示为：

$$\begin{aligned}q^{(t+1)}=&\begin{bmatrix}\sum_{i=1}^{k}q^{(t)}(x=i)Q_{i1} & \sum_{i=1}^{k}q^{(t)}(x=i)Q_{i2} & \cdots & \sum_{i=1}^{k}q^{(t)}(x=i)Q_{ik}\end{bmatrix}_{1\times k}\\ =&q^{(t)}\cdot Q\end{aligned} \tag{5}$$

其中，$q^{(t)}=\begin{bmatrix}q^{(t)}(x=1) & q^{(t)}(x=2) & q^{(t)}(x=3) & \cdots & q^{(t)}(x=k)\end{bmatrix}_{1\times k}$。那么，通过这个递推公式，我们可以得到，$q^{(t+1)}=q^{(t)}Q=q^{(t-1)}Q^2=\cdots=q^{(1)}Q^t$。通过上述的描述，详细大家都已经详细的了解了 Markov Chain 中，每个时刻点的状态的分布 $q^{(t)}$ 的计算方法。既然我们知道了每个时间点的概率分布的计算方法，下一个问题就是我们怎么可以知道一定是收敛的呢?

1.2 Markov Chain 收敛性

由于 Q 是一个随机概率矩阵，那么我们可以得到，每个值都是小于 1 的，所以也必然有特征值的绝对值 ≤ 1。为什么呢? 我们可以从特征值的几何意义上好好的想一想，特征值代表变换中方向不变的向量的变化尺度。随机矩阵的变化尺度必然是小于 1 的。所以，我们可以对概率转移矩阵做特征值分解，分解成对角矩阵：

$$Q=A\Lambda A^{-1}\qquad \Lambda=\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_k\end{bmatrix},\qquad |\lambda_i|\leq 1\qquad (i=1,2,\cdots,k) \tag{6}$$

我们假设只有一个 $\lambda_i=1$，则：

$$q^{(t+1)}=q^{(1)}(A\Lambda A^{-1})^t=q^{(1)}A\Lambda^t A^{-1} \tag{7}$$

当 $t \to \infty$ 时，必然有：

$$\Lambda^t = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \tag{8}$$

我们可以假设存在足够大的 M：

$$s.t. \quad \Lambda^M = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \tag{9}$$

所以，

$$\begin{aligned} q^{(m+1)} =& q^{(1)} A \Lambda^m A^{-1} \\ q^{(m+2)} =& q^{(m+1)} A \Lambda A^{-1} \\ =& q^{(1)} A \Lambda^m A^{-1} A \Lambda A^{-1} \\ =& q^{(1)} A \Lambda^{(m+1)} A^{-1} \\ =& q^{(m+1)} \end{aligned} \tag{10}$$

通过上述的证明，我们成功的证明了 $q^{(m+2)} = q^{(m+1)}$。我们用数学的语言来表述一下，也就是当 $t > m$ 时，$q^{(m+1)} = q^{(m+2)} = \cdots = q^{(\infty)}$。这就是平稳分布，我们成功的证明了 Markov Chain 经过足够大的步数 m 之后，一定会收敛到一个平稳分布。于是，这就启发了我们设计一个 Markov Chain，收敛到我们想要采样的分布 $p(x)$。那么。怎么我们才能让它收敛呢？实际上就是由状态转移矩阵 Q 所决定的。我们的核心问题就是设计一个合适的状态转移矩阵 Q。

那么，我们要做的就是设计一个 MCMC，利用 Markov Chain 收敛到一个平稳分布 $q(x)$，使得平稳分布 $\approx$ 目标分布 $p(x)$。也就是当 m 足够大的时候，$q^{(m)}(x) = q^{(m+1)}(x) = q^{(m+2)}(x) = q(x)$。

那么，我们的 Markov Chain 解决了当维度很高的时候，$q(x) \approx p(x)$ 找不到的情况，在 MCMC 中不要显示的去找，而是构建一个 Markov Chain 去近似，跳过了直接去寻找的过程。

这里我们介绍一个概念，也就是从开始到收敛到 m 的这段时期被我们称为 bum-in，中文翻译为燃烧期 (个人觉得非常的难听，所以我从来不用中文的表述形式)。也有说法称这个时间 t 为 Mix-time。当然也不是任何的分布都可以用 MCMC 来进行采样。但是它可以有效的避免我们去寻找 $q(z)$。下面我们将描述一些用 MCMC 采样时遇到的困难的地方。

2 Existing Problem

1. 虽然，我们可以证明出 MCMC 最终可以收敛到一个平稳分布。但是并没有理论来判断 Markov Chain 是否进入了平稳分布，也就是不知道 Markov Chain 什么时候会收敛。

2. Mixing Time 过长，这就是有高维所造成的，维度和维度之间的相关性太强了，$p(x)$ 太过于复杂了。理论上 MCMC 是可以收敛的，但是如果 m 如果实在是太大的话，我们基本就是认为它是不收敛的。实际上，现在有各种各样的 MCMC 算法的变种都是在想办法解决这个 Mixing Time 过长的问题。

3. 我们希望采到的样本之间的样本相互独立，也就是采到的样本之间的相关性越小越好。

这个有关于样本之间独立性的问题,大家可能不太好理解,这是实际在高维分布中我们采用 MCMC 来进行采样很有可能造成样本单一，相关性太强的问题。我们我们来举一个 Mixture Gaussian Distribution 的例子。下图所示是一个 Mixture Gaussian Distribution 的例子：

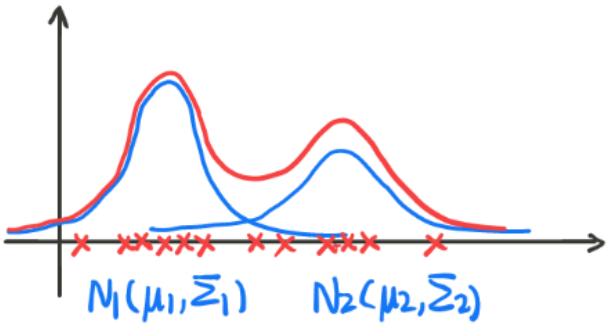

图 2: Mixture Gaussian Distribution 举例

会有一个什么问题呢？就是样本都趋向于一个峰值附近，很有可能会过不了低谷，导致样本都聚集在一个峰值附近。这个问题出现的原因我们可以从能量的角度来解释这个问题。在无向图中，我们常用下列公式来进行表示：

$$P(X) = \frac{1}{Z}\hat{P}(X) = \frac{1}{Z}exp^{-\mathbb{E}(X)} \tag{11}$$

实际上这里的 $\mathbb{E}(X)$ 指的就是能量函数，能量和概率是成反比的，概率越大意味着能量越低，能量越低，越难发生跳跃的现象。所以，采样很容易陷入到一个峰值附近。并且，多峰还可以分为均匀和陡峭，陡峭的情况中，能量差实在是太大了，就很难发生跳跃。就像孙悟空翻出如来佛祖的五指山一样，佛祖的维度很好，孙悟空在翻跟头的时候，一直在一个低维里面不同的打转，根本就跳不出来，就是来自佛祖的降维打击。

所以，在高维情况下，很容易发生在一个峰值附近不停的采样，根本就跳不出来，导致采到的样本的多样性低，样本之间的关联性大，独立性低。