机器学习介绍

2021年5月20日

抽样与贝叶斯推断

贝叶斯后验分布往往不能被直接计算出来，需要抽样。

Inference

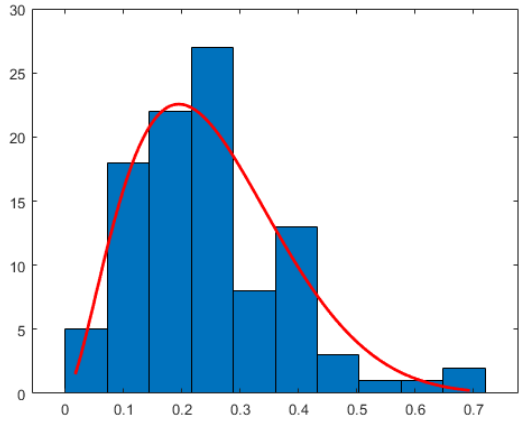
主要目标：系数的完整后验分布

$$Dir(\theta \mid \alpha + x, n_1 + n_2) = \frac{Dir(\theta \mid \alpha, n_1) \times Mu(x \mid n_2, \theta)}{\int Dir(\theta \mid \alpha, n_1) \times Mu(x \mid n_2, \theta) d\theta}$$

似然-先验部分一般可以直接计算

问题：归一化分母部分的积分运算不一定可以计算

解决方案：抽样



- 对后验分布抽样可以代替对后验概率的完整计算，只要抽样频率和概率密度成正比。
- 例如，左图红色线条代表真实分布，蓝色柱子代表抽样，通过正确抽样可以推断分布的统计性质，如均值，概率密度等。
- 已知 $Dir(\theta \mid \alpha + x, n_1 + n_2) \propto Dir(\theta \mid \alpha, n_1) \times Mu(x \mid n_2, \theta)$
- 根据上述正比关系，我们可以直接利用已知的似然和先验来正确抽样。
- 任何对于后验统计性质的计算都可以叫做贝叶斯推断。

抽样与贝叶斯推断

好的抽样频率近似于概率密度。

高尔顿板与正态分布

- 好的抽样：

  对任意两个等距的区间，区间内样本数的比近似于区间中点概率密度的比。

- 高尔顿板可以看成是一维高斯分布的成功抽样，因为：

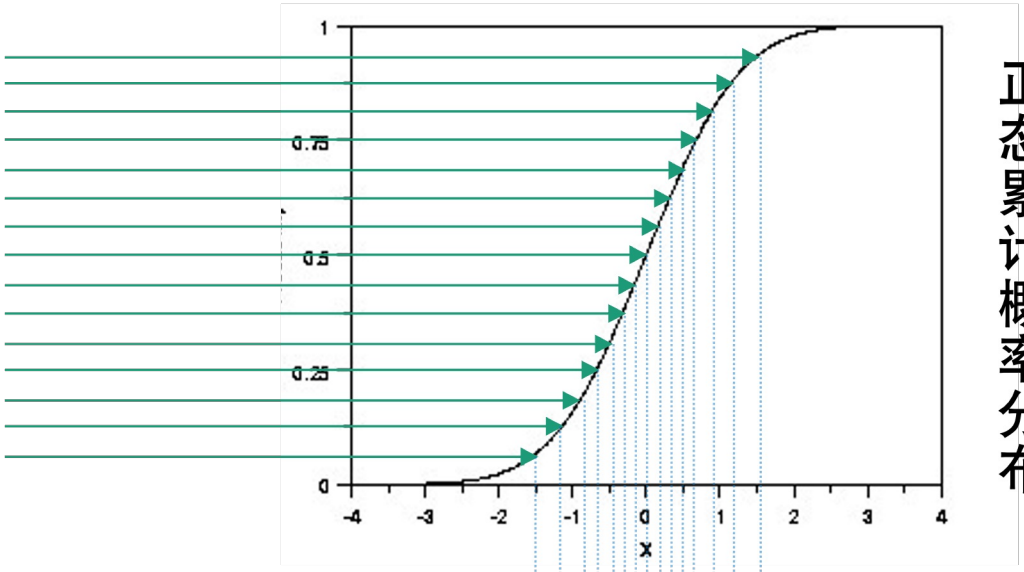
  任意两个格子里小球数量（样本数）的比近似于对应的高斯分布的概率密度比。

- 生成正确抽样的方法有很多种，是贝叶斯学习的重点研究方向。

BOX-MULLER抽样

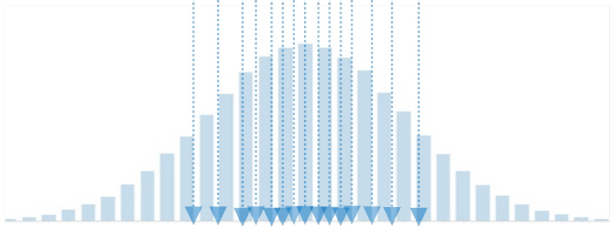
抽样的方法需要创意。

均匀抽样



正态累计概率分布

Box–Muller
采样法



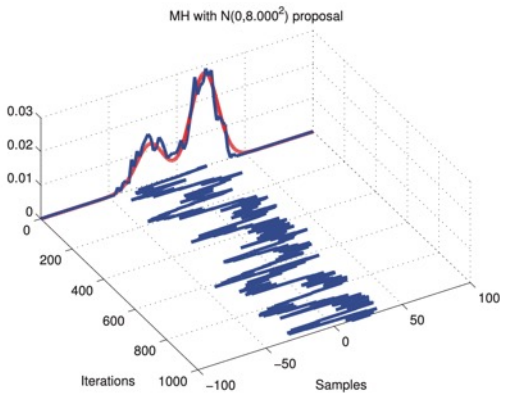
样本密度近似正态分布

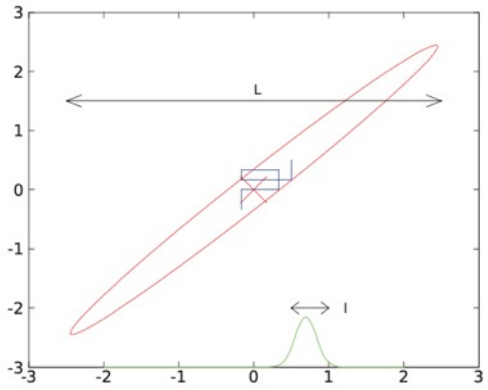
马科夫链蒙地卡罗(MCMC)抽样

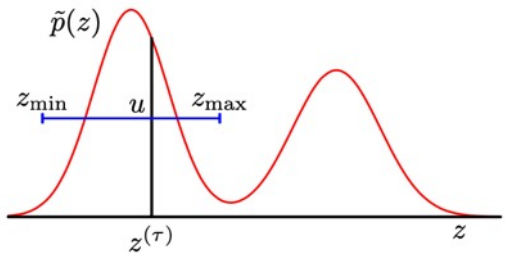
利用马尔科夫链性质达到正确抽样的一类方法

通过基于概率分布的随机游走的方式采样的办法，比较著名的MCMC方法有：

- Metropolis-Hasting
- Gibbs
- Slicing
- ....

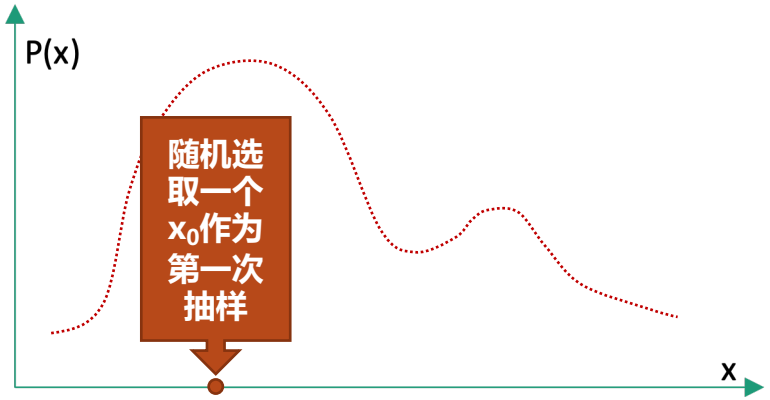




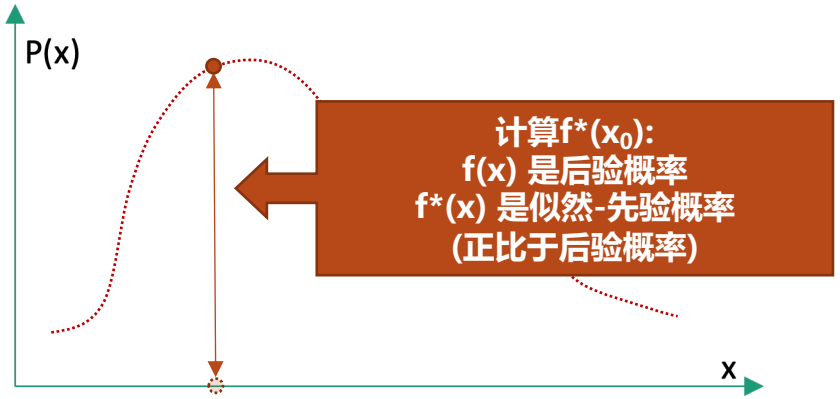


Metropolis-Hasting抽样

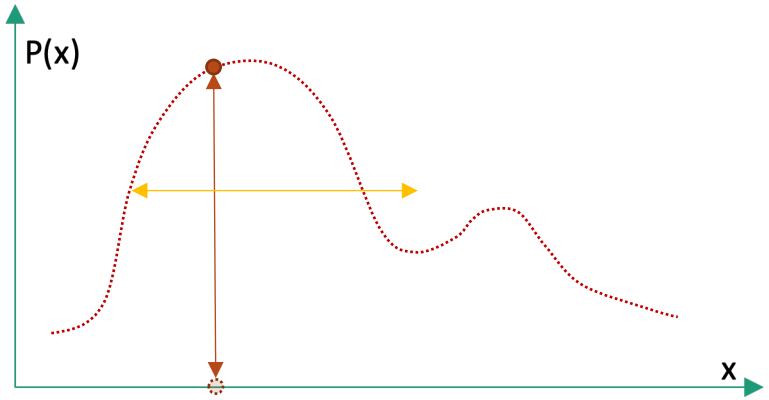
反复横跳。

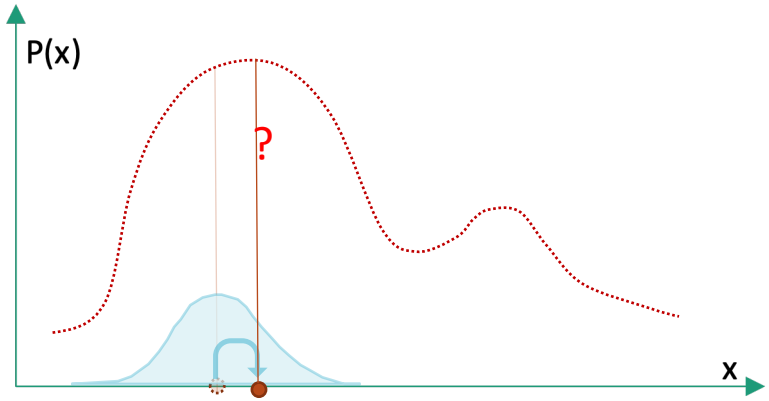


第一步：随机采样x_0



第二步：计算$f(x_0)$

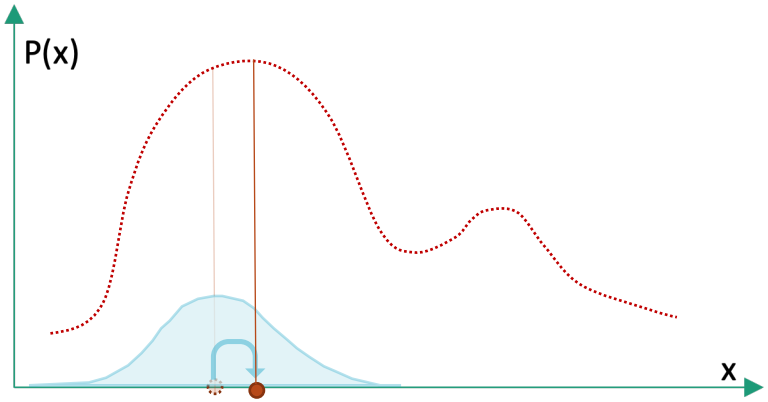




第三步：选取预采样点x_1

Metropolis-Hasting抽样

MH抽样的关键点在于接受概率。



第四步: 决定是否采取预采样点x_1

随机决定是否接受新采样点

随机接受的计算过程：

1. 计算α：

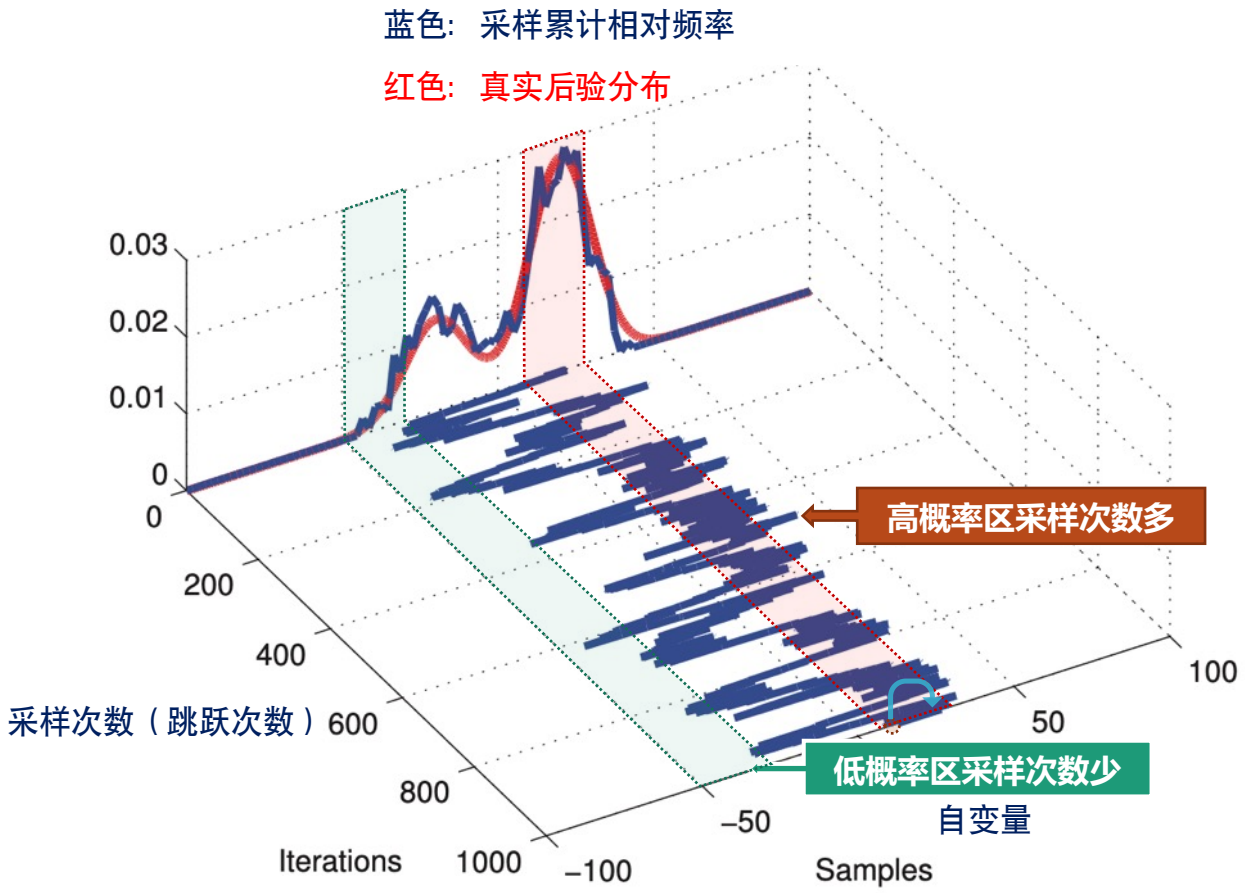
$$\alpha = \frac{f^*(x_1)q(x_1|x_0)}{f^*(x_0)q(x_0|x_1)}$$

$$\alpha = \frac{f^*(x_1)N(x_1|\mu = x_0, \sigma)}{f^*(x_0)N(x_0|\mu = x_1, \sigma)}$$

$$\alpha = \frac{f^*(x_1)}{f^*(x_0)}$$

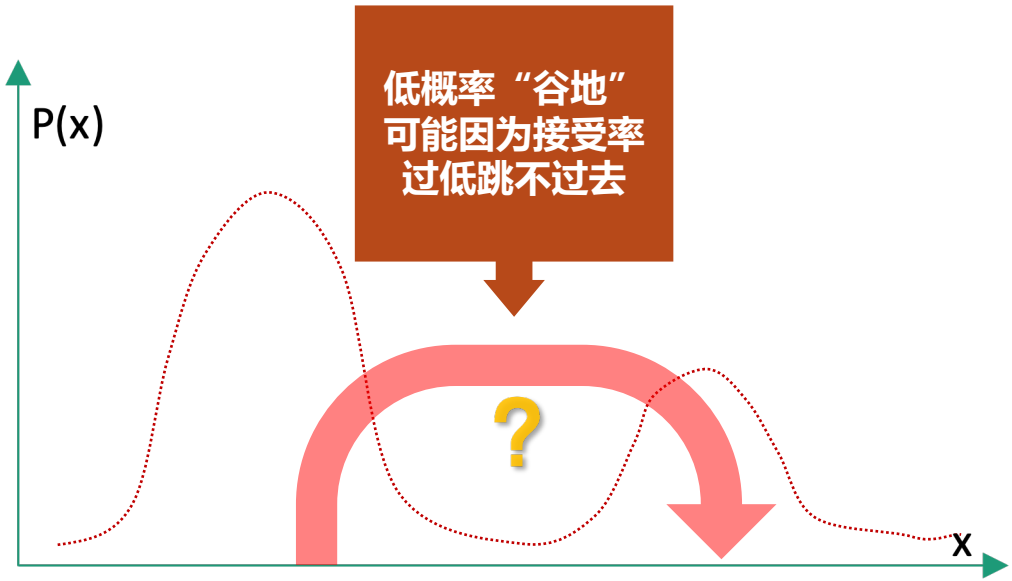
2. 接受采样的概率 r = min（1， α ）
   a. 如接受->x_1即为新的采样点，下次采样从x_1开始跳跃。
   b. 如否定->x_0再次被采样，下次采样从开始x_0跳跃。

Metropolis-Hasting抽样



Metropolis-Hasting抽样

“跳不过去”是MH方法最大的问题。



问题：早期可能采样集中在某个峰值地区，缺乏全局代表性。高纬度分布会加重这个问题。

解决方案：

1. 加大跳跃标准差δ
2. 烧入期Burn-in Period：MCMC需要长时间才能达到稳态，在实践中我们常常会抛弃早期样本，使用后期样本进行分析。
3. 多链Metropolis-Hasting： 从不同点同时开始多条采样。

....

Metropolis-Hasting抽样

MH和梯度下降的联系。

思考题：

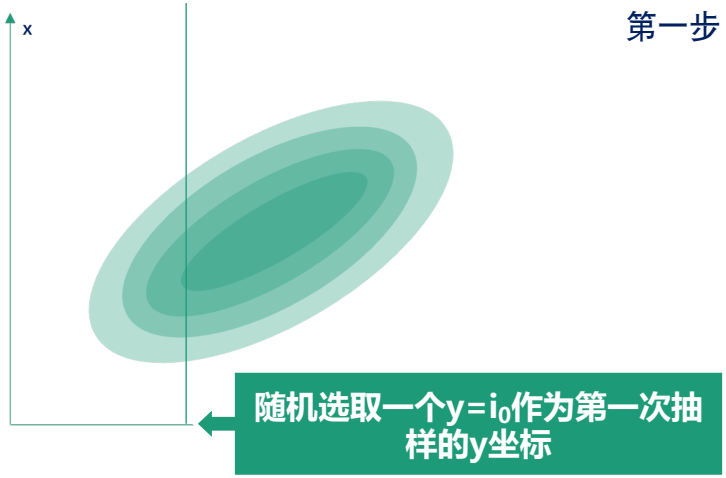
如果只接受向上的跳跃 ($r=1$，$f(x_1)>f(x_0)$； $r=0$，$f(x_1)<f(x_0)$)，这个算法会近似于梯度下降，为什么？

Gibbs抽样

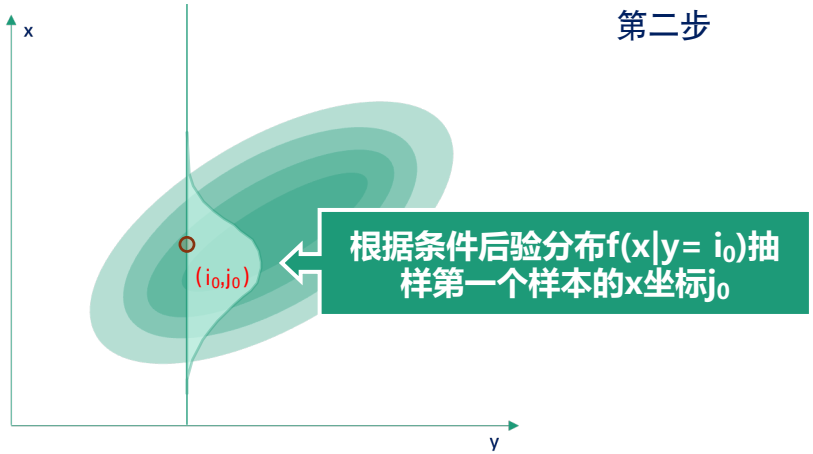
横竖跳跃。

初始设定

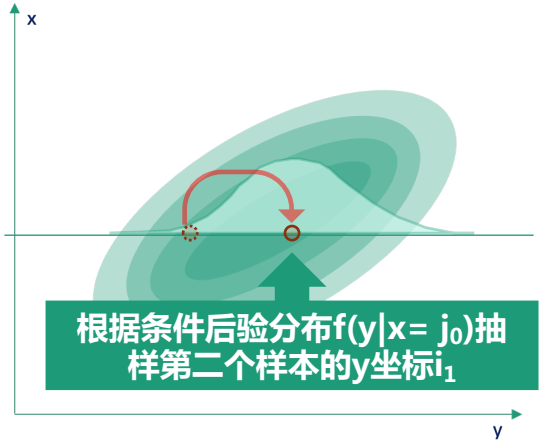
第一步



第二步

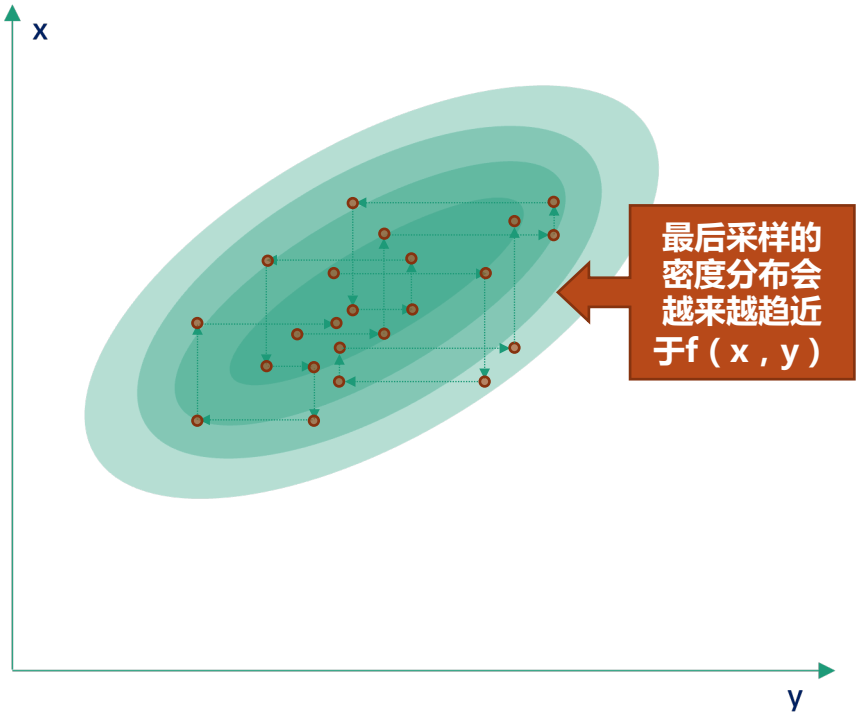


第三步



Gibbs抽样

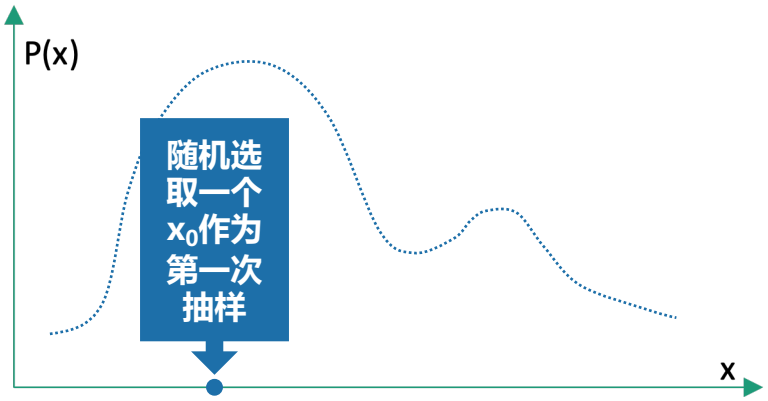
Gibbs相对MH的优劣。



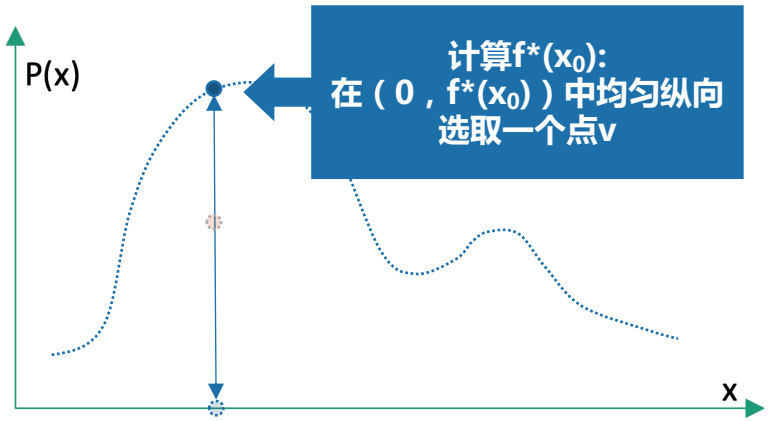
1. 最后采样相对密度会与后验概率密度近似。
2. 同样也需要burn-in period。
3. 需要计算条件概率表达式
4. 不存在接受率的问题，但由于需要计算每个维度的条件概率，不见得采样速度比MH快。

Slicing抽样

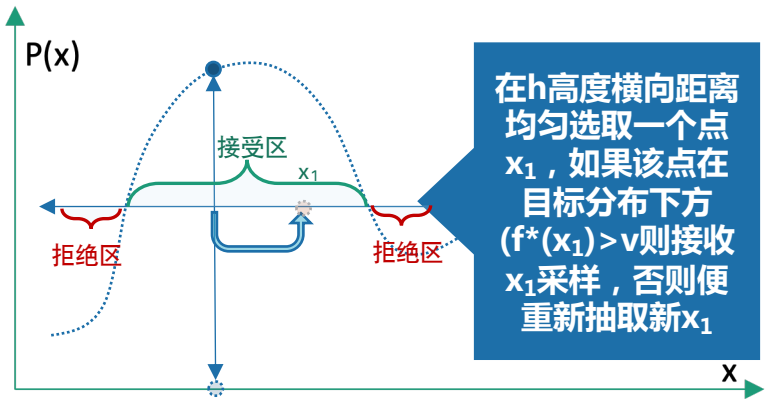
Gibbs相对MH的优劣。



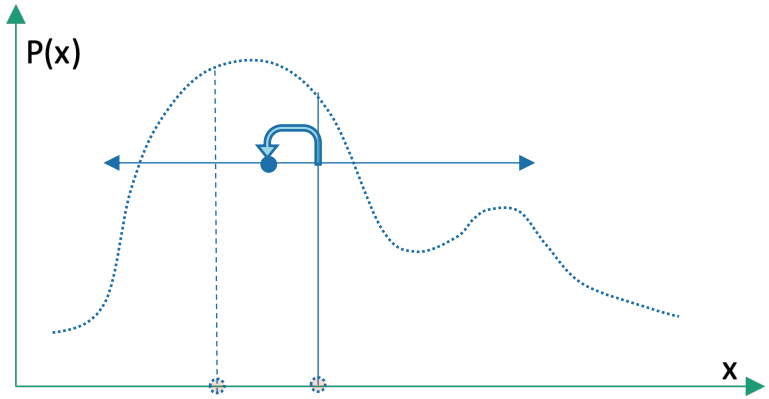
第一步



第二步：计算$f(x_0)$，纵向选择v



第三步：抽样x_1选择接受或拒绝采样

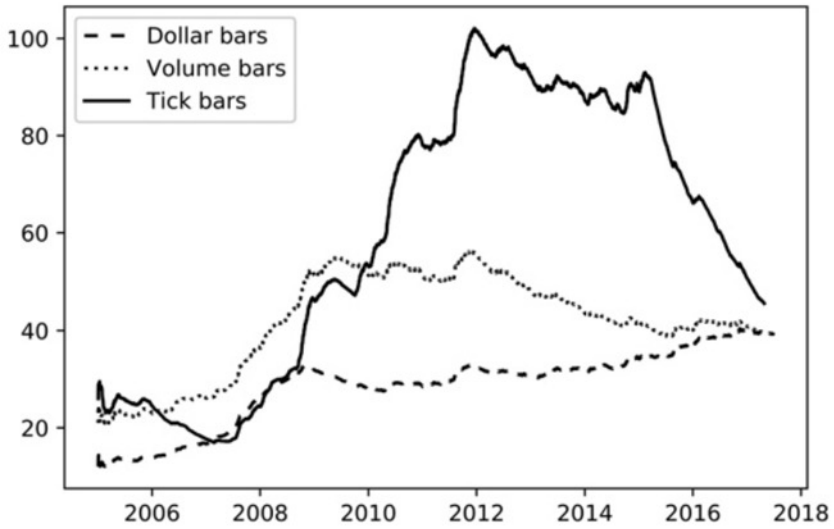


第三步：重复

交易时序数据的基础单位

从Tick、Time Bar、Volume Bar、Money Bar、和基于这些数据的衍生数据中提取训练数据。 *Inference*

1. 逐笔交易（Tick）数据。
2. Time Bars：大部分金融从业人员最常见的K线是以时间为基础的（15MIN，1H，1D...）包括OHLC等等数据。
3. Tick Bars：以固定交易笔数为单位的K线。
4. Volume Bars：以固定的交易量为基础单位的K线，如100手，起始价格是第1手（101手）的交易价格、收尾价格是第100手的交易价格（200手）。
5. Money Bars：以固定的交易金额为基础单位的K线，如100万，起始价格是第1万（101万）的交易价格、收尾价格是第100万的交易价格（200万）。
6. 相关数据的平均值、变化率等等。



美股不同类型k线产生的平均频率

数据来源：<Advances in financial machine learning>