

第三章 随机过程的抽样方法

随机过程的基本概念

- 有限 vs 无限随机变量的集合, 非参数 Bayesian 模型的计算需对随机过程进行抽样, 本章将介绍一些常见随机过程的抽样, 如随机游走, 高斯过程, 泊松过程, Dirichlet process
- 一个随机过程一般记为 $\{X(t) \mid t \in \mathcal{T}\}$
 - ▶ 指标集 (index set) \mathcal{T} 是离散或连续的集合, 如 $\mathcal{T} = \{1, 2, \dots\}$ 或 $\mathcal{T} = [0, \infty)$
 - ▶ 如果 \mathcal{T} 是 \mathbb{R}^d ($d > 1$) 上的一个子集, 称定义在 \mathcal{T} 上的随机过程为 **random field**
- 随机过程 $\{X(t) \mid t \in \mathcal{T}\}$ 的一次实现定义了一个从 \mathcal{T} 到 \mathbb{R} 的函数 $f(\cdot)$, 称函数 $f(\cdot)$ 为该随机过程的一条**样本路径**

随机过程的基本概念

- $\forall t_1, \dots, t_m \in \mathcal{T}$, 称 $(X(t_1), \dots, X(t_m))$ 的联合分布为随机过程 $X(t)$ 的一个有限维分布
 - ▶ 有限维分布不能唯一确定一个随机过程, 比如, $X(t)$ 是定义在指标集 $\mathcal{T} = [0, 1]$ 上的随机过程, 随机抽取 $s \sim \mathbf{U}[0, 1]$, 定义另一个随机过程 $Y(t)$ 如下:

$$Y(t) = \begin{cases} X(s) + 1, & t = s \\ X(t), & t \neq s \end{cases}$$

- 如果需要计算的值涉及随机过程 $X(t)$ 在无限个点 t 处的值, 比如计算 $\mu = E[g(X(\cdot))]$, 可使用 Monte Carlo 方法做近似估计: 先从随机过程生成 n 条样本路径 $X_i(t_{ij}), i = 1, \dots, n$. 假设第 i 条路径有 M_i 个点, 则 μ 可估计为

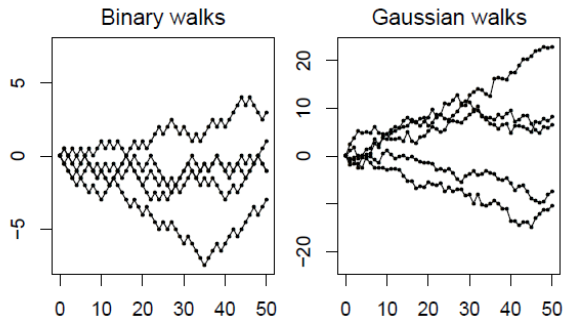
$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n g(X_i(t_{i1}), \dots, X_i(t_{iM_i}))$$

随机游走

随机游走一般具有以下形式：

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

其中 Z_t 是 iid 的随机向量。初始点 X_0 通常取为 0



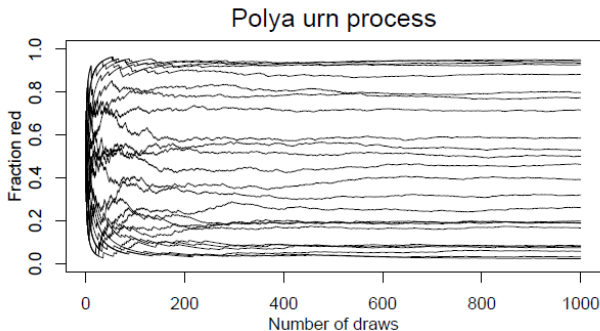
随机游走

- **Pólya's urn process.** 令 R_t 代表 t 时刻的红球数, B_t 代表黑球数, $X_t = (R_t, B_t)$. 初始时刻 $X_0 = (1, 1)$, 增量 Z_t 服从如下分布:

$$Z_t = \begin{cases} (1, 0), & \text{概率} = R_t / (R_t + B_t) \\ (0, 1), & \text{概率} = B_t / (R_t + B_t) \end{cases}$$

$t \rightarrow \infty$ 时, 桶中红球所占的比例 $Y_t = R_t / (R_t + B_t)$ 是多少?

- ▶ 数学家 Pólya 证明 Y_t 的每条样本路径都会收敛到一个值 $Y_\infty \sim \mathbf{U}(0, 1)$



高斯过程

- 任意有限维分布是一个多元正态分布
- 期望函数: $\mu(t) = E[X(t)], t \in \mathcal{T}$
- 协方差函数: $\Sigma(t, s) = \text{Cov}(X(t), X(s)), \forall t, s \in \mathcal{T}$
 - ▶ 对称性: $\Sigma(t, s) = \Sigma(s, t)$
 - ▶ (半) 正定性: $\sum_{i=1}^m \sum_{j=1}^m x_i x_j \Sigma(t_i, t_j) \geq 0, \forall m \geq 1, t_i \in \mathcal{T}, x_i \in \mathbb{R}$
- 为函数插值提供不确定性估计, 非参数 Bayesian 模型常用的先验分布
 - ▶ 假设 $f(\cdot)$ 是高斯过程的一条样本路径, 对 $f(\cdot)$ 的任意有限维分布就有先验信息
 - ▶ 观察到 k 个点的值 $f(t_1), \dots, f(t_k)$ 后, 计算样本路径上任一点 $f(t)$ 的条件期望和方差: $E[f(t) | f(t_1), \dots, f(t_k)], \text{Var}[f(t) | f(t_1), \dots, f(t_k)]$
 - ▶ 在每一点对 $f(t)$ 抽样可以生成一条通过已知点的样本路径

平稳的高斯过程

- **平稳**的随机过程：对任意间隔 Δ , $\forall t \in \mathcal{T}$, $X(t)$ 和 $X(t + \Delta)$ 都是同分布
- 高斯过程的平稳性等价于

$$\begin{aligned}\mu(t + \Delta) &= \mu(t) \equiv \mu(0) \\ \Sigma(t + \Delta, s + \Delta) &= \Sigma(t, s) = \Sigma(t - s, 0) \quad \forall \Delta, \forall t, s \in \mathcal{T}\end{aligned}$$

一些常见的平稳高斯过程的协方差函数：

- **Exponential covariance**

$$\Sigma(t, s) = \sigma^2 \exp(-\theta |t - s|), \quad \theta > 0$$

该过程的样本路径连续但不可导

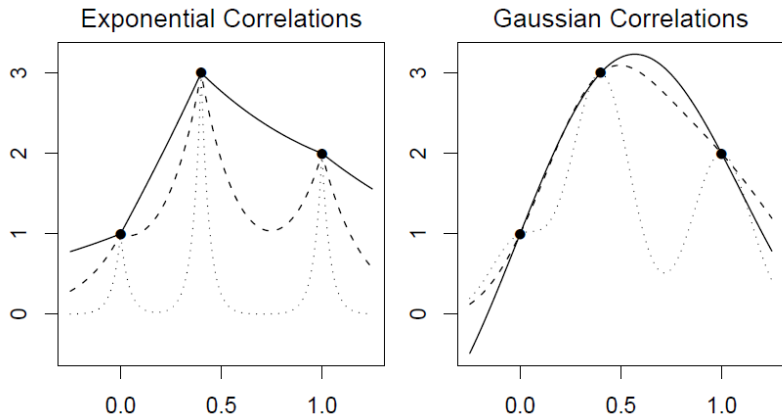
- **Gaussian covariance**

$$\Sigma(t, s) = \sigma^2 \exp(-\theta(t - s)^2), \quad \theta > 0$$

也被称为 squared exponential covariance, 它的样本路径任意阶可导

平稳的高斯过程

Gaussian Process Interpolations

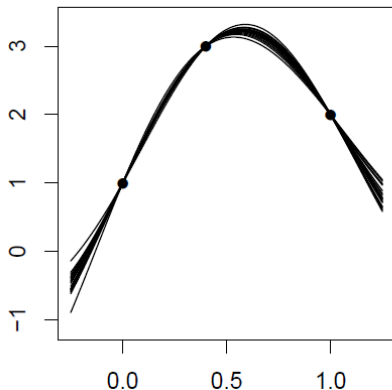


图中的实线，虚线，点线分别对应 $\theta = 1, 5, 25$

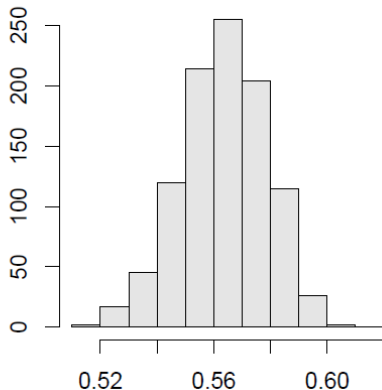
平稳的高斯过程

Gaussian Process Interpolations

Simulated realizations



Sample maxima



Gaussian covariance, $\theta = 1$, $\sigma^2 = 1$

平稳的高斯过程

- **Matérn covariances** $\Sigma(t, s; \nu)$ 由一个平滑度系数 ν 控制, 通过 Bessel function 定义, 当 $\nu = m + 1/2$ 且 $m \in \mathbb{N}$, $\Sigma(t, s; \nu)$ 有解析形式, 前 4 个特例为:

$$\Sigma\left(t, s; \frac{1}{2}\right) = \sigma^2 \exp(-\theta|t - s|)$$

$$\Sigma\left(t, s; \frac{3}{2}\right) = \sigma^2 (1 + \theta|t - s|) \exp(-\theta|t - s|)$$

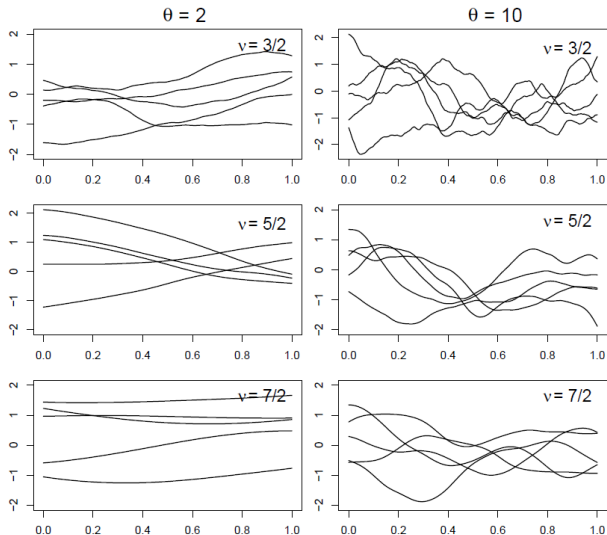
$$\Sigma\left(t, s; \frac{5}{2}\right) = \sigma^2 \left(1 + \theta|t - s| + \frac{1}{3}\theta^2|t - s|^2\right) \exp(-\theta|t - s|)$$

$$\Sigma\left(t, s; \frac{7}{2}\right) = \sigma^2 \left(1 + \theta|t - s| + \frac{2}{5}\theta^2|t - s|^2 + \frac{1}{15}\theta^3|t - s|^3\right) \exp(-\theta|t - s|)$$

- ▶ $\nu = m + 1/2$ 时, Matérn covariance 生成的样本路径有 m 阶导数
- ▶ $\nu \rightarrow \infty$ 时, Matérn covariance 收敛到 Gaussian covariance
- ▶ Matérn covariance 可以提供介于 exponential covariance 和 Gaussian covariance 之间的平滑度

平稳的高斯过程

Matern Process Realizations



布朗运动

- **标准布朗运动** (维纳过程) $B(t)$ 是定义在 $\mathcal{T} = [0, \infty)$ 上的高斯过程, 它有三条性质:

- ① $B(0) = 0$.
- ② 对于任意的 $0 = t_0 < t_1 < \cdots < t_m$, $B(t_i) - B(t_{i-1}) \stackrel{ind}{\sim} N(0, t_i - t_{i-1})$, $i = 1, \dots, m$.
- ③ $B(t)$ 的样本路径在 $[0, \infty)$ 上以概率 1 连续
 - ▶ $B(t)$ 的期望函数 $\mu(t) = 0$, 协方差函数 $\Sigma(t, s) = \min(t, s)$, 不平稳
 - ▶ $B(t)$ 的样本路径连续, 但以概率 1 处处不可导

- 将标准布朗运动记为 $B(\cdot) \sim \text{BM}(0,1)$, 定义一个新的随机过程

$$X(t) = \delta t + \sigma B(t)$$

称 $X(t)$ 是漂移 δ , 方差 σ^2 的布朗运动, 记为 $X(\cdot) \sim \text{BM}(\delta, \sigma^2)$

- ▶ $X(t)$ 的期望函数 $\mu(t) = \delta t$, 协方差函数 $\Sigma(t, s) = \sigma^2 \min(t, s)$
- ▶ 对 $X(\cdot) \sim \text{BM}(\delta, \sigma^2)$ 在 $[0, T]$ 上抽样, 只需先抽取 $B(\cdot) \sim \text{BM}(0,1)$ 在 $[0, 1]$ 上的样本路径, 然后令 $X(t) = \delta t + \sigma \sqrt{TB}(t/T)$

布朗运动

- 如何对标准布朗运动在 $[0,1]$ 上抽样?

- 对于 $[0,1]$ 上的任意一列点, $0 < t_1 < t_2 < \cdots < t_m \leq 1$, 根据定义可得 $B(\cdot)$ 在这些点的样本:

$$B(t_1) = \sqrt{t_1} Z_1,$$

$$B(t_j) = B(t_{j-1}) + \sqrt{t_j - t_{j-1}} Z_j, \quad j = 2, \dots, m$$

其中 $Z_j \stackrel{\text{ind}}{\sim} N(0, 1)$, $j = 1, \dots, m$

- 写为矩阵形式:

$$\begin{pmatrix} B(t_1) \\ B(t_2) \\ \vdots \\ B(t_m) \end{pmatrix} = \begin{pmatrix} \sqrt{t_1} & 0 & \cdots & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \cdots & \sqrt{t_m - t_{m-1}} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}$$

布朗桥

- 称布朗运动 $B(t)$ 在 $[l, r]$ 上给定两端值 $B(l)$ 和 $B(r)$ 的条件分布为一个**布朗桥**
- 标准布朗桥** 是 $[0, 1]$ 上给定 $B(0) = B(1) = 0$ 的标准布朗运动, 记为 $BB(0,1)$
- 对布朗桥抽样可以通过布朗运动的路径得到: 如果 $B(\cdot) \sim BM(0,1)$, 令

$$\tilde{B}(t) = B(t) - tB(1), \quad t \in [0, 1]$$

则 $\tilde{B}(\cdot) \sim BB(0,1)$

- 生成任意两点之间的一条布朗运动的路径: 给定路径的起点 $B(l)$ 和终点 $B(r)$, 可如下产生 $BM(\delta, \sigma^2)$ 在 $[l, r]$ 上的一条路径

$$B(t) = B(l) + \frac{t-l}{r-l}(B(r) - B(l)) + \sigma\sqrt{r-l}\tilde{B}\left(\frac{t-l}{r-l}\right), \quad l \leq t \leq r$$

其中 $\tilde{B}(\cdot) \sim BB(0,1)$

几何布朗运动

- 布朗运动最初描述物体受周围粒子的碰撞在空间中做随机运动，多次微小碰撞的累加效应会趋于一个正态分布
- 股价受各种市场信息的影响不断波动，股价变化常被描述为一种相乘效应或对数尺度上的累加效应，极限分布是对数正态分布，对应的随机过程被称为几何布朗运动
- 股价 S_t 在一个小区间 Δ 上的相对变化 (收益率):

$$\frac{S_{t+\Delta} - S_t}{S_t} = \frac{\Delta S_t}{S_t} \approx \frac{dS_t}{S_t}$$

- 经典的金融模型 (Hull, 2003) 这样描述 S_t 的相对变化:

$$\frac{dS_t}{S_t} = \delta dt + \sigma dB_t$$

其中 $B \sim \text{BM}(0,1)$

- ▶ 该模型假设 S_t 在一个小区间 Δ 上的相对变化

$$\frac{\Delta S_t}{S_t} \sim N(\delta\Delta, \sigma^2\Delta)$$

几何布朗运动

- 称满足

$$dS_t = \delta S_t dt + \sigma S_t dB_t \quad (2)$$

的随机过程 S_t 是一个几何布朗运动, 记为 $S \sim \text{GBM}(S_0, \delta, \sigma^2)$

- 方程(2)是少数几个有解析解的随机微分方程, 解的形式为

$$S_t = S_0 \exp \{ (\delta - \sigma^2/2)t + \sigma B_t \} \quad (3)$$

定理 (Itô's formula)

如果 $dS_t = a(S_t)dt + b(S_t)dB_t$ 且 $f(\cdot)$ 是一个二阶连续可导的函数, 则

$$df(S_t) = \left(f'(S_t)a(S_t) + \frac{1}{2}f''(S_t)b^2(S_t) \right) dt + f'(S_t)b(S_t)dB_t.$$

几何布朗运动

- **亚式看涨期权的定价.** 航空公司最怕遇到油价大幅上涨, 用 S_t 表示 t 时刻的油价, 假设当前时刻的油价 $S_0 = 1$. 如果价格 $S_t > 1.1$, 航空公司就会面临亏损. 有一种亚式看涨期权可以帮助航空公司对冲油价上涨的风险, 如果航空公司购买了该期权, 一年之后会收到以下金额

$$f(S(\cdot)) = \max \left(0, \frac{1}{12} \left(\sum_{j=1}^{12} S_{j/12} \right) - K \right)$$

该看涨期权可以保证航空公司的油价成本不超过 K .

- ▶ 这样一份期权的售价是多少? 理论上该期权在当前时刻的合理价格为 $e^{-rT} E(f(S))$, 其中 T 是距离到期日的时间, r 是无风险利率. 假设油价的波动 S_t 是一个几何布朗运动, 根据(3)可以生成大量 S_t 的样本路径, 每条路径都可以计算一个 f 的值, 这些 f 的独立观察值的平均值会收敛到 $E(f(S))$

泊松点过程

- **点过程**是指某个集合 $S \subset \mathbb{R}^d$ 内的一系列随机点 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\}$
 - ▶ $S = [0, \infty)$: 随机来电的时间、网站访问高峰的时间、台风登陆的时间等
 - ▶ $S \subset \mathbb{R}^d$ ($d \geq 2$): 地震的位置、森林中树的位置、星云中星系的位置等
- 将点过程中点的个数记为 $N(S)$, 对于集合 $A \subset S$, 用 $N(A)$ 表示落在 A 中的点的个数
- **点过程的有限维分布**是 S 上任意 J 个不相交的子集中点的个数的联合分布, 即 $(N(A_1), \dots, N(A_J))$ 的分布, 其中 $A_1, \dots, A_J \subset S$ 且互不相交

定义 (均匀泊松过程)

如果对 S 上任意 J 个不相交的子集 $A_j \subset S$ 且 $\text{vol}(A_j) < \infty, j = 1 \dots, J$, 点列 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\}$ 满足

$$N(A_j) \stackrel{\text{ind}}{\sim} \text{Po}(\lambda \cdot \text{vol}(A_j)), \quad j = 1 \dots, J,$$

称该点列为 S 上一个强度为 λ ($\lambda > 0$) 的**均匀泊松过程**, 记为 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\} \sim \text{PP}(S, \lambda)$.

泊松点过程

定义 (非均匀泊松过程)

如果对 S 上任意 J 个不相交的子集 $A_j \subset S$ 且 $\text{vol}(A_j) < \infty, j = 1 \dots, J$, 点列 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\}$ 满足

$$N(A_j) \stackrel{\text{ind}}{\sim} \text{Po} \left(\int_{A_j} \lambda(\mathbf{s}) d\mathbf{s} \right), \quad j = 1 \dots, J,$$

其中强度函数 $\lambda(\mathbf{s}) \geq 0$, 称该点列为 S 上的一个非均匀泊松过程, 记为 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\} \sim \text{NHPP}(S, \lambda)$.

定理

$\lambda(\mathbf{s}) \geq 0$ 是 S 上的一个强度函数且 $\Lambda(S) = \int_S \lambda(\mathbf{s}) d\mathbf{s} < \infty$. 如果 S 上的点列 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\}$ 满足 $N(S) \sim \text{Po}(\Lambda(S))$, 且给定 $N(S) = n$,

$$P(\mathbf{P}_i \in A) = \frac{1}{\Lambda(S)} \int_A \lambda(\mathbf{s}) d\mathbf{s}, \quad \forall A \subset S, \quad i = 1, \dots, n,$$

则点列 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\} \sim \text{NHPP}(S, \lambda)$.

泊松点过程

推论

对于 S 上一个强度为 λ 的均匀泊松过程, 如果 $\text{vol}(S) < \infty$, 可以如下对其抽样: 首先抽取

$$N(S) \sim \text{Po}(\lambda \cdot \text{vol}(S)),$$

然后在 S 上独立均匀地抽取 $N(S)$ 个点 $\mathbf{P}_i \sim \mathbf{U}(S)$, $i = 1, \dots, N(S)$.

- 练习. 如何在圆盘 $D = \{\mathbf{x} \in \mathbb{R}^2 \mid \mathbf{x}^\top \mathbf{x} \leq 1\}$ 上抽取一系列点 $\{\mathbf{P}_1, \mathbf{P}_2, \dots\} \sim \text{PP}(D, \lambda)$?
- 如果能从 PDF 为 $\rho(\mathbf{s}) \propto \lambda(\mathbf{s})$ 的分布抽样, 就可以对强度为 $\lambda(\mathbf{s})$ 的 NHPP 抽样

$[0, \infty)$ 上的泊松过程

- 假设 $\mathcal{T} = [0, \infty)$ 上的点列按顺序产生, $T_1 < T_2 < \dots$, 定义计数函数

$$N(t) \equiv N([0, t]) = \sum_{i=1}^{\infty} \mathbf{1}(T_i \leq t), \quad 0 \leq t < \infty$$

- $\mathcal{T} = [0, \infty)$ 上的均匀泊松过程 $\text{PP}(\lambda)$ 具有以下三条性质:

- ① $N(0) = 0$
- ② $N(t) - N(s) \sim \text{Po}(\lambda(t-s)), 0 \leq s < t$
- ③ 增量独立: 对任意的 $0 = t_0 < t_1 < \dots < t_m$, $N(t_i) - N(t_{i-1}), i = 1, \dots, m$ 是独立的

- 点列 $\{T_1, T_2, \dots\} \sim \text{PP}(\lambda)$ 的另一重要特性:

$$T_i - T_{i-1} \stackrel{iid}{\sim} \text{Exp}(\lambda), \quad i \geq 1, \quad T_0 = 0.$$

- ▶ $T_0 = 0, T_i = T_{i-1} + E_i/\lambda, i \geq 1$, 其中 $E_i \stackrel{iid}{\sim} \text{Exp}(1)$

$[0, \infty)$ 上的泊松过程

- 如果只需要在一个有界区间 $[0, T]$ 上对 $PP(\lambda)$ 抽样, 一个更简单的抽样方法:

$$N = N(T) \sim \text{Po}(\lambda T)$$

$$S_i \sim \mathbf{U}[0, T], \quad i = 1, \dots, N$$

$$T_i = S_{(i)}, \quad i = 1, \dots, N$$

- $\mathcal{T} = [0, \infty)$ 上的非均匀泊松过程 $NHPP(\lambda)$ 具有以下三条性质:

① $N(0) = 0$

② $N(t) - N(s) \sim \text{Po}\left(\int_s^t \lambda(x) dx\right), \quad 0 \leq s < t$

③ $N(t)$ 的增量独立

- 为方便对 $NHPP(\lambda)$ 抽样, 定义 cumulative rate function:

$$\Lambda(t) = \int_0^t \lambda(s) ds$$

$[0, \infty)$ 上的泊松过程

- 点列 $\{T_1, T_2, \dots\} \sim \text{NHPP}(\lambda)$, 定义随机变量 $Y_i = \Lambda(T_i)$, 证明

$$Y_i \sim \text{PP}(1)$$

- 因此可以如下从 $\text{NHPP}(\lambda)$ 中抽取点列 T_1, T_2, \dots

$$\begin{aligned} Y_i &= Y_{i-1} + E_i, \quad i = 1, 2, \dots \\ T_i &= \Lambda^{-1}(Y_i) \end{aligned} \tag{4}$$

其中 $E_i \stackrel{iid}{\sim} \text{Exp}(1)$, $Y_0 = 0$

- ▶ 如果 $\lim_{t \rightarrow \infty} \Lambda(t) = \infty$, 算法(4)可以一直运行下去
- ▶ 如果 $\lim_{t \rightarrow \infty} \Lambda(t) = M < \infty$, 当 $Y_j > M$ 时, $\Lambda^{-1}(Y_j)$ 不存在, 算法停止

Dirichlet process

- DP 描述的是分布的分布, 每一条样本路径都是一个分布, 有限维分布是一个 Dirichlet 分布
- DP 在非参数 Bayesian 模型中常作为一个未知分布的 prior, DP prior 具有共轭性, 利用 DP 可构造无限维混合分布模型, 即 DP mixture model
- 定义在 $\Omega \subset \mathbb{R}^d$ 上的 DP 是通过一个常数 $\alpha > 0$ 和 Ω 上的一个确定的分布 G (CDF) 定义的, 它的任意有限维分布对应 Ω 的一个有限分割:

$$\Omega = A_1 \cup A_2 \cup \cdots \cup A_m, \quad A_i \cap A_j = \emptyset, \quad i \neq j$$

且满足

$$(F(A_1), \dots, F(A_m)) \sim \text{Dir}(\alpha G(A_1), \dots, \alpha G(A_m))$$

- 如果随机分布 $F \sim \text{DP}(\alpha, G)$ (或 $\text{DP}(\alpha G)$), F 的期望是 G , α 决定了 F 到 G 的平均距离

DP prior

使用 DP prior 的非参数 Bayesian 模型:

$$\begin{aligned}\mathbf{x}_1, \dots, \mathbf{x}_n &\stackrel{iid}{\sim} F \\ F &\sim DP(\alpha, G)\end{aligned}$$

推断观察到数据 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 后 F 的后验分布

- 以 $n = 1$ 为例, 可得 F 的后验分布也是一个 DP

$$F \mid \mathbf{x}_1 \sim DP(\alpha G + \delta_{\mathbf{x}_1})$$

其中 $\delta_{\mathbf{x}_1}$ 是一个退化分布的 CDF, 该分布的所有概率集中在点 \mathbf{x}_1

- 依此类推, 当有 n 个观察值 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 时, F 的后验分布为

$$F \mid \mathbf{x}_1, \dots, \mathbf{x}_n \sim DP\left(\alpha G + \sum_{i=1}^n \delta_{\mathbf{x}_i}\right)$$

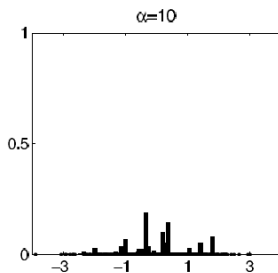
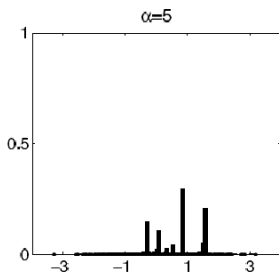
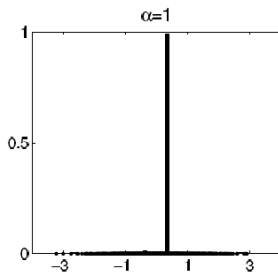
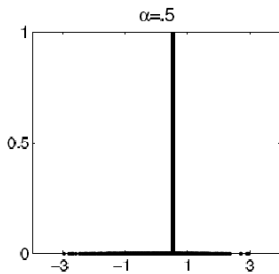
Stick-breaking process

- DP 的定义没有给出该如何对 $F \sim \text{DP}(\alpha, G)$ 抽样, DP 的 **stick-breaking representation** 给出了一种直接建立 DP 样本的方法:

$$F = \sum_{j=1}^{\infty} \pi_j \delta_{\mathbf{x}_j}, \quad \pi_j = \theta_j \prod_{i < j} (1 - \theta_i), \quad \theta_i \stackrel{iid}{\sim} \text{Beta}(1, \alpha), \quad \mathbf{x}_j \stackrel{iid}{\sim} G.$$

- 如果 $F \sim \text{DP}(\alpha, G)$, F 一定是一个离散分布
- 给每个点 \mathbf{x}_j 分配的概率 π_j 从一个 **stick-breaking process** 中产生以保证 $\sum_{j=1}^{\infty} \pi_j = 1$
- 由于 $\theta_j \sim \text{Beta}(1, \alpha)$, $E(\theta_j) = \frac{1}{1 + \alpha}$, 如果 α 很小, 上述过程倾向于给排在前面的点分配较大的概率, 后面的点分到很小的概率

Stick-breaking process



Chinese restaurant process

- CRP 是另一种对 DP 抽样的方法
- 考虑从以下两阶段模型抽样:

$$\begin{aligned} F &\sim \text{DP}(\alpha, G) \\ \mathbf{X}_i &\sim F, \quad i = 1, \dots, n. \end{aligned} \tag{5}$$

- ▶ $n = 1$ 时, \mathbf{X}_1 的边际分布是什么?
- ▶ $n \geq 2$ 时, 依次抽取 $\mathbf{X}_i \sim F \mid \mathbf{X}_1, \dots, \mathbf{X}_{i-1}$

$$F \mid \mathbf{X}_1, \dots, \mathbf{X}_{i-1} \sim \text{DP} \left(\alpha + i - 1, \frac{\alpha G + \sum_{j=1}^{i-1} \delta_{\mathbf{X}_j}}{\alpha + i - 1} \right)$$

此时 $\mathbf{X}_i \sim F \mid \mathbf{X}_1, \dots, \mathbf{X}_{i-1}$ 的边际分布为

$$\mathbf{X}_i \sim (\alpha G + \sum_{j=1}^{i-1} \delta_{\mathbf{X}_j}) / (\alpha + i - 1) \tag{6}$$

Chinese restaurant process

- (6)中的分布是一个混合分布, 可如下进行抽样:

$$\mathbf{x}_i = \begin{cases} \mathbf{Y} \sim G & \text{概率 } \alpha/(\alpha + i - 1) \\ \mathbf{x}_1 & \text{概率 } 1/(\alpha + i - 1) \\ \vdots & \vdots \\ \mathbf{x}_{i-1} & \text{概率 } 1/(\alpha + i - 1). \end{cases} \quad (7)$$

由(7)生成的随机过程被称为Chinese restaurant process (CRP)

Chinese restaurant process



- ▶ 在 CRP 中, 第 n 个顾客到达时期望开设的桌数是 $O(\log n)$
- ▶ 从 CRP 产生的一条样本路径 $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ (n 很大) 的经验分布可近似看作 $DP(\alpha, G)$ 中随机生成的一个分布

DP mixture model

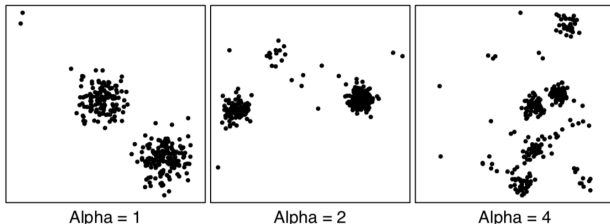
- 在 CRP 模型(5)中再加一层就得到了DP mixture model:

$$\begin{aligned} F &\sim \text{DP}(\alpha, G) \\ \mathbf{X}_1, \dots, \mathbf{X}_n &\sim F \\ \mathbf{Y}_i &\stackrel{\text{ind}}{\sim} H(\cdot \mid \mathbf{X}_i), \quad i = 1, \dots, n. \end{aligned} \tag{8}$$

其中 $\{\mathbf{Y}_i\}_{i=1}^n$ 是观察值, $\{\mathbf{X}_i\}_{i=1}^n$ 和 F 是待估计的参数

- 考察从模型(8)生成的数据 $\{\mathbf{Y}_i\}_{i=1}^n$ 的特点
 - G 为 $N_2(0, 3^2 I)$, $\mathbf{Y}_i \stackrel{\text{ind}}{\sim} N_2(\mathbf{X}_i, 0.4^2 I)$

Dirichlet process mixture samples



DP mixture model

- CRP 模型产生重复值的特点使它很适合描述有聚类特征的数据
- α 越小, 样本 $\{\mathbf{y}_i\}$ 的聚集效应越明显
- DP mixture model (8) 的优点是不需要提前设定聚类的个数, 一般使用 MCMC 方法估计聚类的个数 ($\{\mathbf{x}_i\}$ 取到不同值的个数)
- DP mixture model 允许聚类的个数是无限的, 具有很大灵活度, 可以随着新观察值的加入不断引入新的聚类