Dirichlet 过程及相关随机测度的渐近行为

献给严士健教授 90 华诞

冯水

Department of Mathematics and Statistics, McMaster University, Hamilton, L8S 4K1, Canada

E-mail: shuifeng@mcmaster.ca

收稿日期: 2018-08-28; 接受日期: 2018-12-10; 网络出版日期: 2019-08-27

Natural Sciences and Engineering Research Council of Canada 资助项目

摘要 Dirichlet 过程是取纯原子概率值的随机测度. 它最早由 Ferguson (1973) 提出, 目的是作为 Bayes 非参数统计的先验分布. Pitman 和 Yor 在 20 世纪 90 年代把 Dirichlet 过程推广到了两参数, 从而得到了非常一般的两参数 Dirichlet 过程或 Pitman-Yor 过程. Dirichlet 过程及其两参数推广, 不仅成为 Bayes 非参数统计领域的基本模型, 同时还被广泛应用到生物学、通信、经济学、金融、遗传学、统计物理和机器学习等领域. 近年来在各种应用问题的推动下, 关于 Dirichlet 过程的极限行为方面的研究非常活跃. 本文将介绍这方面的一些研究结果, 主要包括大样本和相关参数趋近边界的极限行为.

关键词 Dirichlet 过程 两参数 Dirichlet 过程 Poisson-Dirichlet 分布 Ewens-Pitman 抽样公式 大偏差 中偏差 中心极限定理

MSC (2010) 主题分类 60F10, 92D10

1 引言

无穷维随机测度是现代概率论的一个非常活跃的研究领域. 这不仅是数学理论本身的自然延伸, 更与现代科技的飞速发展密不可分. Dirichlet 过程是非常重要的一类无穷维随机测度. 它起源于 Bayes 非参数统计方面的研究, 但随后的发展使其在通信、经济学、金融、遗传学、统计物理和机器学习等许多领域找到应用. 近二三十年来, 研究人员得到了各种各样关于 Dirichlet 过程的推广, 其中一个重要推广是 Pitman 和 Yor^[1] 在 20 世纪 90 年代提出并研究的两参数 Dirichlet 过程.

本文将介绍作者及其合作者所得到的与两参数 Dirichlet 过程有关的一些极限结果. 这主要包括大数定律、中心极限定理、大偏差和中偏差. 相应的极限过程既包括样本数增大, 也包括有关参数趋于边界. 大部分结果 (2010 年以前) 的证明都可以在文献 [2] 中找到.

英文引用格式: Feng S. Asymptotic behavior of Dirichlet process and related random measures (in Chinese). Sci Sin Math, 2020, 50: 47–58, doi: 10.1360/N012018-00210

Dirichlet 过程及相关随机测度的渐近行为涉及面非常广, 参考文献也很多. 这里介绍的只是其中一小部分, 所引用的文献不可避免地有很大的局限性. 有兴趣的读者可以进一步参见文献 [3–7] 及其中的文献.

本文后面几节的结构如下: 第 2 节将给出 Dirichlet 过程及两参数推广的定义及相关性质; 第 3 节介绍 Ewens-Pitman 抽样公式和大样本极限结果; 与参数趋于边界有关的极限定理将在第 4 节给出.

2 Dirichlet 过程: 定义和构造

本节给出 Dirichlet 过程的定义并讨论利用平稳独立增量过程的构造. 我们采用的定义来源于文献 [8]. 给定任意一个 Polish 空间 S , 定义

$$\begin{aligned}\Delta &= \left\{ \mathbf{x} = (x_1, x_2, \dots) : x_i \geq 0, \sum_{i=1}^{\infty} x_i = 1 \right\}, \\ \nabla_{\infty} &= \left\{ \mathbf{p} = (p_1, p_2, \dots) : p_1 \geq p_2 \geq \dots \geq 0, \sum_{i=1}^{\infty} p_i = 1 \right\}, \\ M_1(S) &= S \text{ 上的所有概率的集合, 赋予弱收敛拓扑.}\end{aligned}$$

任意给定 $\theta > 0$, 考虑一系列独立同分布 $\text{Beta}(1, \theta)$ 随机变量 $U_1, U_2, \dots$, 引入

$$V_1 = U_1, \quad V_n = (1 - U_1) \cdots (1 - U_{n-1})U_n, \quad n \geq 2,$$

并令 $\mathbf{P}(\theta) = (P_1(\theta), P_2(\theta), \dots)$ 代表 $\mathbf{V}(\theta) = (V_1, V_2, \dots)$ 从大到小的顺序统计量. 给定 $\nu_0 \in M_1(S)$, 设 $\eta_1, \eta_2, \dots$ 为一列独立于 $\mathbf{V}(\theta)$ 的具有同分布 ν_0 的独立随机变量.

定义 2.1 $\mathbf{P}(\theta)$ 的分布称为参数为 θ 的 Poisson-Dirichlet 分布, 记为 $\text{PD}(\theta)$; $\mathbf{V}(\theta)$ 的分布称为 GEM (Griffiths-Engen-McCloskey) 分布, 记作 $\text{GEM}(\theta)$. Dirichlet 过程, 记为 $\mathcal{D}(\theta, \nu_0)$, 是如下随机测度的分布:

$$\Xi_{\theta, \nu_0} = \sum_{i=1}^{\infty} P_i(\theta) \delta_{\eta_i} = \sum_{i=1}^{\infty} V_i(\theta) \delta_{\eta_i},$$

其中 δ_x 代表集中在 x 点的 Dirac 测度. 为了方便起见, 我们也称 Ξ_{θ, ν_0} 为 Dirichlet 过程.

Dirichlet 过程最早出现在文献 [9], 其中给出了严格的构造及在 Bayes 统计中的应用. Dirichlet 过程的后验分布还是 Dirichlet 过程, 只改变了参数. 这个共轭性质使得 Dirichlet 过程成为 Bayes 统计推断领域最重要的先验分布之一. Dirichlet 这个名字是由于它是有限维 Dirichlet 分布的无穷维推广. 对于任意 $m \geq 2$, $\alpha_1, \dots, \alpha_m > 0$, 有限维 Dirichlet 分布 $\mathcal{D}(\alpha_1, \dots, \alpha_m)$ 是空间

$$\Delta_m = \left\{ (x_1, \dots, x_m) : x_i \geq 0, i = 1, \dots, m; \sum_{k=1}^m x_k = 1 \right\}$$

上的连续概率, 其密度函数为

$$\frac{\Gamma(\alpha_1 + \dots + \alpha_m)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m x_i^{\alpha_i - 1}.$$

考虑空间 S 的任意分割 $B_1, \dots, B_m$, 可以证明随机向量 $(\Xi_{\theta, \nu_0}(B_1), \dots, \Xi_{\theta, \nu_0}(B_m))$ 的分布为 $\mathcal{D}(\theta \nu_0(B_1), \dots, \theta \nu_0(B_m))$.

Pitman^[10] 引入了 Poisson-Dirichlet 分布的两参数推广. 其具体定义如下.

给定 $0 \leq \alpha < 1$ 和 $\theta > -\alpha$, 令 $W_1, W_2, \dots$ 为一列独立随机变量, 其中 W_k 服从 $\text{Beta}(1-\alpha, \theta+k\alpha)$ 分布 ($k=1, \dots$). 定义

$$V_1(\alpha, \theta) = W_1, \quad V_n(\alpha, \theta) = (1 - W_1) \cdots (1 - W_{n-1})W_n, \quad n \geq 2,$$

并令 $\mathbf{P}(\alpha, \theta) = (P_1(\alpha, \theta), P_2(\alpha, \theta), \dots)$ 表示 $\mathbf{V}(\alpha, \theta) = (V_1(\alpha, \theta), V_2(\alpha, \theta), \dots)$ 从大到小的顺序统计量. 给定 $\nu_0 \in M_1(S)$, 设 $\eta_1, \eta_2, \dots$ 为一列独立于 $\mathbf{V}(\alpha, \theta)$ 的具有同分布 ν_0 的独立随机变量.

定义 2.2 随机概率 $\mathbf{P}(\alpha, \theta)$ 的分布称为两参数 Poisson-Dirichlet 分布, 记为 $\text{PD}(\alpha, \theta)$; 随机概率 $\mathbf{V}(\alpha, \theta)$ 的分布称为两参数 GEM 分布, 记作 $\text{GEM}(\alpha, \theta)$. 两参数 Dirichlet 或 Pitman-Yor 过程, 记为 $\mathcal{D}(\alpha, \theta, \nu_0)$, 为如下随机测度的分布:

$$\Xi_{\alpha, \theta, \nu_0} = \sum_{i=1}^{\infty} P_i(\alpha, \theta) \delta_{\eta_i} = \sum_{i=1}^{\infty} V_i(\alpha, \theta) \delta_{\eta_i}.$$

与前面类似, 我们也称 $\Xi_{\alpha, \theta, \nu_0}$ 为两参数 Dirichlet 过程.

Dirichlet 过程及其两参数推广可以通过具有平稳独立增量的过程或 Lévy 过程来构造. 这个联系使得许多计算变得更加清晰.

定义 2.3 令 $\mathbb{R}_+ = (0, \infty)$. Λ 代表 $\mathbb{R}_+$ 上的一个测度. Lévy 过程 $\{\tau(t) : t \geq 0\}$ 称作具有 Lévy 测度 Λ 的隶属子 (subordinator), 如果 $\tau_0 = 0$ 并且

$$\mathbb{E}[e^{-\lambda \tau(t)}] = \exp \left\{ t \int_{\mathbb{R}_+} (e^{-\lambda x} - 1) \Lambda(dx) \right\}, \quad \lambda \geq 0.$$

由上述定义可以看出, 隶属子 $\{\tau(t) : t \geq 0\}$ 的行为完全由 Lévy 测度 Λ 决定. 下面几个例子可以用来构造 Dirichlet 过程及其两参数推广.

例 2.4 (Gamma 隶属子) 当 Lévy 测度为

$$\Lambda(dx) = x^{-1} e^{-x} dx, \quad x > 0$$

时, 相应的隶属子称为 Gamma 隶属子. 为明确起见, 我们用 $\{\gamma(t) : t \geq 0\}$ 来表示 Gamma 隶属子. 直接计算可得

$$\begin{aligned} \mathbb{E}[e^{-\lambda \gamma(t)}] &= \exp \left\{ -t \int_0^\infty (1 - e^{-\lambda x}) x^{-1} e^{-x} dx \right\} \\ &= \frac{1}{(1 + \lambda)^t}, \quad \lambda \geq 0. \end{aligned}$$

例 2.5 (稳定隶属子) 任意给定 $\alpha \in (0, 1)$ 和正常数 C , 具有指标 α 的稳定隶属子 $\{\rho(t) : t \geq 0\}$ 的 Lévy 测度为

$$\Lambda(dx) = \frac{C\alpha}{\Gamma(1-\alpha)} x^{-(1+\alpha)} dx, \quad x > 0.$$

利用分部积分公式可以得到

$$\begin{aligned} \mathbb{E}[e^{-\lambda \rho(t)}] &= \exp \left\{ -\frac{C\alpha}{\Gamma(1-\alpha)} t \int_{(0, \infty)} (1 - e^{-\lambda x}) \frac{dx}{x^{1+\alpha}} \right\} \\ &= \exp \left\{ -\frac{C}{\Gamma(1-\alpha)} t \int_{(0, \infty)} \lambda e^{-\lambda x} \frac{dx}{x^\alpha} \right\} \\ &= e^{-C\lambda^\alpha t}, \quad \lambda \geq 0. \end{aligned}$$

例 2.6 (广义 Gamma 隶属子) 广义 Gamma 隶属子 $\{\varrho(t) : t \geq 0\}$ (参见文献 [11, 12]) 的 Lévy 测度为

$$\Lambda(dx) = \Gamma(1 - \alpha)^{-1} x^{-(1+\alpha)} e^{-x} dx, \quad x > 0, \quad 0 < \alpha < 1.$$

相应的 Laplace 变换为

$$\begin{aligned} E[e^{-\lambda \varrho(t)}] &= \exp \left\{ -t \int_{(0, \infty)} (1 - e^{-\lambda x}) \frac{x^{-(1+\alpha)} e^{-x} dx}{\Gamma(1 - \alpha)} \right\} \\ &= \exp \left\{ -\frac{t}{\alpha \Gamma(1 - \alpha)} \int_{(0, \infty)} [(\lambda + 1)e^{-(\lambda+1)x} - e^{-x}] x^{-\alpha} dx \right\} \\ &= \exp \left\{ -\frac{t}{\alpha} [(\lambda + 1)^\alpha - 1] \right\}. \end{aligned}$$

下面定理给出了 Dirichlet 过程及其两参数推广通过隶属子的构造.

定理 2.7 ^[1, 13] 给定 $\nu_0 \in M_1(S)$. 对于 $\theta > 0$, 有

$$\Xi_{\theta, \nu_0}(\cdot) \stackrel{d}{=} \frac{\gamma(\theta \nu_0(\cdot))}{\gamma(\theta)}. \quad (2.1)$$

对于 $\theta = 0, 0 < \alpha < 1$, 有

$$\Xi_{\alpha, 0, \nu_0} \stackrel{d}{=} \frac{\rho(\nu_0(\cdot))}{\rho(1)}. \quad (2.2)$$

最后, 对于 $\theta > 0, 0 < \alpha < 1$, 有

$$\Xi_{\alpha, \theta, \nu_0} \stackrel{d}{=} \frac{\varrho(\alpha \gamma(\theta/\alpha) \nu_0(\cdot))}{\varrho(\alpha \gamma(\theta/\alpha))}, \quad (2.3)$$

其中 $\stackrel{d}{=}$ 代表依分布相等.

3 Ewens-Pitman 抽样公式及渐近结果

考虑一个群体, 其中每个个体都属于某种类. 假设这些类的先验分布为两参数 Dirichlet 过程 $\mathcal{D}(\alpha, \theta, \nu_0)$. 对任意正整数 n , 令 $X_1, \dots, X_n$ 为该群体的一个随机样本. 换句话讲, X_i 代表个体 i 的类而且给定 $\Xi_{\alpha, \theta, \nu_0} = \mu$, $X_1, \dots, X_n$ 为具有同分布 μ 的独立随机变量. 对任意 $i = 1, \dots, n$, 记 $\mathbf{M} = (M_1, \dots, M_n)$, 其中 M_i 代表样本中出现 i 次的类的数量. 令

$$\mathbf{K}_n(\alpha, \theta) = M_1 + \dots + M_n$$

代表在样本中出现的所有不同类的总数. 它的值越大, 表明样本及群体的多样性越强. 显然,

$$n = \sum_{i=1}^n i M_i.$$

引入集合

$$\mathbb{M} = \left\{ \mathbf{m} = (m_1, \dots, m_n) : m_1, \dots, m_n \geq 0, \sum_{i=1}^n i m_i = n \right\}.$$

不难看出, $\mathbf{M}$ 是取值于 $\mathbb{M}$ 的随机向量. 其分布为如下著名的 Ewens-Pitman 抽样公式.

定理 3.1 ^[10,14] 给定 $0 \leq \alpha < 1$ 、 $\theta > -\alpha$ 和任意 $\mathbf{m} \in \mathbb{M}$, 我们有

$$P(\mathbf{M} = \mathbf{m}) = \frac{n!}{(\theta)_n} \prod_{l=0}^{k-1} (\theta + l\alpha) \prod_{i=1}^n \frac{((1-\alpha)_{(i-1)})^{m_i}}{(i!)^{m_i} (m_i!)}, \quad (3.1)$$

其中 $k = \sum_{i=1}^n m_i$. 对任意实数 x 和正整数 n , $(x)_n = x(x+1) \cdots (x+n-1)$. 为方便起见, 记 $(x)_0 = 1$. 特别地, 当 $\alpha = 0$ 时, 我们得到如下的 Ewens 抽样公式:

$$P(\mathbf{M} = \mathbf{m}) = \frac{\theta^k}{(\theta)_n} \frac{n!}{\prod_{i=1}^n i^{m_i} (m_i!)}. \quad (3.2)$$

考虑 $\mathbb{M}$ 中的两个特殊状态 $(n, 0, \dots, 0)$ 和 $(0, 0, \dots, 1)$. 直接计算可得

$$P(\mathbf{M} = (n, 0, \dots, 0)) = P(\mathbf{K}_n(\alpha, \theta) = n) = \frac{\prod_{l=0}^{n-1} (\theta + l\alpha)}{(\theta)_n}$$

和

$$P(\mathbf{M} = (0, 0, \dots, 1)) = P(\mathbf{K}_n(\alpha, \theta) = 1) = \frac{(1-\alpha)_{(n-1)}}{(\theta+1)_{(n-1)}}.$$

当 n 和 α 固定时, 我们有

$$\lim_{\theta \rightarrow \infty} P(\mathbf{M} = (n, 0, \dots, 0)) = \lim_{\theta \rightarrow \infty} P(\mathbf{K}_n(\alpha, \theta) = n) = 1.$$

这说明当 θ 的值增加时, 样本和群体的多样性也增加. 如果 n 和 θ 固定而 α 收敛到 1, 那么,

$$\lim_{\alpha \rightarrow 1} P(\mathbf{M} = (n, 0, \dots, 0)) = \lim_{\alpha \rightarrow 1} P(\mathbf{K}_n(\alpha, \theta) = n) = 1.$$

因此, θ 趋于无穷大和 α 收敛到 1 具有一定的相似性. 这一点在下一节的极限结果中会看得更清楚.

下面介绍的结果是关于 $\mathbf{K}_n(\alpha, \theta)$ 的渐近行为. 相应的极限过程为 θ 和 α 固定, 样本数 n 趋于无穷大.

定理 3.2 设 $\alpha = 0$, $\theta > 0$. 令 $\{\xi_i : i \geq 1\}$ 表示一系列独立 Poisson 随机变量序列, 其中 ξ_i 的均值为 θ/i . 这时, 我们有

$$\mathbf{K}_n(0, \theta) \stackrel{d}{=} \xi_1 + \cdots + \xi_n.$$

同时下列结果成立:

(1) (大数定律)

$$\lim_{n \rightarrow \infty} \frac{\mathbf{K}_n(0, \theta)}{\log n} = \theta, \quad \text{a.s.};$$

(2) (中心极限定理 (参见文献 [15] ($\theta = 1$) 和 [16] (一般 θ)))

$$\sqrt{\log n} \left(\frac{\mathbf{K}_n(0, \theta)}{\log n} - \theta \right) \Rightarrow Z,$$

其中 $\Rightarrow$ 表示依分布收敛, Z 服从标准正态分布;

(3) (参见文献 [17]) 序列 $\{\frac{\mathbf{K}_n(0, \theta)}{\log n} : n \geq 1\}$ 满足大偏差原理, 其速率为 $1/\log n$, 速度函数为

$$I_1(x) = \begin{cases} x \log \frac{x}{\theta} - x + \theta, & x > 0, \\ \theta, & x = 0, \\ +\infty, & x < 0. \end{cases}$$

文献 [18] ($\theta = 1$) 和 [19] (一般 θ) 将中心极限定理推广到了泛函中心极限定理. 对于 $\alpha > 0$, 我们有下面的结果.

定理 3.3 设 $\alpha > 0, \theta > -\alpha$. 令 $S_{\alpha, \theta}$ 为正值连续随机变量, 其密度函数为

$$f_{\alpha, \theta}(s) = \frac{\Gamma(\theta + 1)}{\pi \alpha \Gamma(\frac{\theta}{\alpha} + 1)} s^{\frac{\theta}{\alpha}} \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!} \Gamma(\alpha k + 1) s^{k-1} \sin(\pi \alpha k), \quad s > 0,$$

其中 $\Gamma(\cdot)$ 为 Gamma 函数. 这时我们有下列结果:

(1) (扰动定律^[1])

$$\lim_{n \rightarrow \infty} \frac{\mathbf{K}_n(\alpha, \theta)}{n^\alpha} = S_{\alpha, \theta}, \quad \text{a.s.};$$

(2) (大偏差原理^[17]) 序列 $\{\frac{\mathbf{K}_n(\alpha, \theta)}{n} : n \geq 1\}$ 满足大偏差原理, 其速率为 $1/n$, 速度函数为

$$I_2(x) = \sup_{\lambda} \{\lambda x - \Lambda_{\alpha}(\lambda)\},$$

其中

$$\Lambda_{\alpha}(\lambda) = \begin{cases} -\log[1 - (1 - e^{-\lambda})^{\frac{1}{\alpha}}], & \lambda > 0, \\ 0, & \text{其他;} \end{cases} \quad (3.3)$$

(3) (中偏差原理^[20]) 令序列 $\{\beta_n\}_{n \geq 1}$ 满足

$$\lim_{n \rightarrow +\infty} \frac{\beta_n}{n^{1-\alpha}} = 0, \quad \lim_{n \rightarrow +\infty} \frac{\beta_n}{(\log n)^{1-\alpha}} = +\infty,$$

则序列 $\{(\beta_n n^{\alpha})^{-1} \mathbf{K}_n(\alpha, \theta)\}_{n \geq 1}$ 满足空间 $\mathbb{R}$ 上的大偏差原理, 其速度为 $\beta_n^{-1/(1-\alpha)}$, 相应的速度函数为

$$I_3(x) = \begin{cases} (1 - \alpha) \alpha^{\alpha/(1-\alpha)} x^{1/(1-\alpha)}, & x \geq 0, \\ +\infty, & x < 0. \end{cases}$$

从上面的定理可以看出, $\mathbf{K}_n(\alpha, \theta)$ 关于 n 的阶数非常依赖于 α 的值. 从 $\alpha = 0$ 的对数阶到正 α 的幂阶, 样本和群体的多样性随着 α 的增加而加强. 因此, 随机变量 $S_{\alpha, \theta}$ 称为 α -多样性. 当一个群体的多样性比较强时, 具有正 α 值的两参数 Dirichlet 过程可以作为其分布的自然选择.

与 Ewens-Pitman 抽样公式有关的大样本渐近结果还有很多, 其中包括 $\mathbf{M} = (M_1, \dots, M_n)$ 本身的各种渐近行为, 具体可参见文献 [3, 21–23]. 在 Bayes 非参数统计框架下的研究可参见文献 [24–26].

4 参数边界附近的行为

本节介绍两参数 Dirichlet 过程本身的渐近结果. 极限机制包括 θ 趋于无穷大、 α 趋于 1 和 $\theta \vee \alpha$ 收敛到 0. 对于每一种极限机制, 我们将给出相应的应用背景.

4.1 θ 趋于无穷大

在群体遗传学里, 参数 θ 代表的是群体总体的突变率. 它与群体的有效总量成正比. 因此, 个体的突变率固定时, θ 趋于无穷大对应于群体的有效总量变大. 在这种情形下, 每个个体的作用相对减弱. 与总体有关的一些统计量会呈现 Gauss 的行为.

定理 4.1 设 $0 \leq \alpha < 1$, $\theta > -\alpha$. 两参数 Dirichlet 过程 $\Xi_{\alpha, \theta, \nu_0}$ 满足下列结果:

(1) (大数定律) 在空间 $M_1(S)$ 中,

$$\Xi_{\alpha, \theta, \nu_0} \rightarrow \nu_0 \quad \text{a.s.} \quad \theta \rightarrow \infty.$$

(2) (中心极限型定理 [2, 27]) 令 $\mathbb{D}$ 表示定义在 $[0, 1]$ 上的右连左极函数 (1 点左连续) 空间, 赋予 Skorohod 拓扑. $B(t)$ 为标准 Brown 运动, $\hat{B}(t)$ 为 Brown 桥

$$\hat{B}(t) = B(t) - tB(1), \quad 0 \leq t \leq 1.$$

设 $g(t) = \nu_0([0, t])$, $\mathbb{X}_{\alpha, \theta}(t) = \sqrt{\theta}(\Xi_{\alpha, \theta, \nu_0}([0, t]) - g(t))$, 那么下述依分布收敛结果在 $\mathbb{D}$ 上成立:

$$\mathbb{X}_{\alpha, \theta}(\cdot) \Rightarrow \sqrt{1-\alpha} \hat{B}(g(\cdot)).$$

(3) (大偏差原理 (参见文献 [28, 29]、[30] ($\alpha = 0$) 和 [21] ($\alpha > 0$))) 当 θ 趋于无穷大时, $\{\Xi_{0, \theta, \nu_0} : \theta > 0\}$ 在空间 $M_1(S)$ 上满足大偏差原理, 其速率为 $1/\theta$, 速度函数为

$$I_4(\mu) = \begin{cases} H(\nu_0 | \mu), & \nu_0 \text{ 的支撑包含 } \mu \text{ 的支撑,} \\ +\infty, & \text{其他,} \end{cases}$$

其中 $H(\nu_0 | \mu)$ 为 ν_0 关于 μ 的相对熵.

取 $S = [0, 1]$. 当 $\alpha > 0$ 并且 θ 趋于无穷大时, $\{\Xi_{\alpha, \theta, \nu_0} : \theta > 0\}$ 在空间 $M_1(S)$ 上满足大偏差原理, 其速率为 $1/\theta$, 速度函数为

$$I_5(\mu) = \sup_{f \geq 0, f \in C(S)} \left\{ \frac{1}{\alpha} \log E^{\nu_0}[f^\alpha] + 1 - E^\mu[f] \right\},$$

其中 $C(S)$ 代表 S 上的连续函数集.

注 4.2 结果 (2) 最早出现在文献 [27]. 文献 [2] 给出了不同的证明. 尽管陈述的结果适用于参数的全部区域, 但是 $\alpha = 0$ 和 $\alpha \neq 0$ 两种情形的证明方法不一样.

如果只考虑 $\Xi_{\alpha, \theta, \nu_0}$ 的质量或两参数 Poisson-Dirichlet 分布, 相应的极限结果也很多. 这里只介绍一个应用背景非常强的统计量的渐近结果.

考虑一个含有至多可数类个体的总量, 其中第 i 类个体在总量中的比例是 p_i . 显然, $\mathbf{p} = (p_1, p_2, \dots)$ 属于空间 ∇_∞ . 任意给定整数 $m \geq 2$, 定义 ∇_∞ 上的函数

$$H_m(\mathbf{p}) = \sum_{i=1}^{\infty} p_i^m.$$

不难看出, 如果从总量中任意选取 m 个样本, 那么, 函数 $H_m(\mathbf{p})$ 为所有样本都属于同一类的概率. 函数 $H_m(\mathbf{p})$ 的取值在 0 和 1 之间. 当总量中只有一类个体时, $H_m(\mathbf{p})$ 达到最大值 1. 当总量中类的总数增加并且每一类的比例比较平均时, $H_m(\mathbf{p})$ 值会变小而接近于 0. 因此, 函数 $H_m(\mathbf{p})$ 的取值可以用来刻画一个总量的多样性和平均性. 在群体遗传学中, 函数 $H_2(\mathbf{p})$ 称作纯合子 (homozygosity). 它可以用来估计群体的突变率. 在生态学和经济学中, 函数 $H_2(\mathbf{p})$ 分别称作 Simpson 指数 [31] 和 Herfindahl-Hirschman 指数 [32, 33].

将 $\mathbf{p}$ 改为 $\mathbf{P}(\alpha, \theta)$, 我们得到下面的随机统计量:

$$H_m(\mathbf{P}(\alpha, \theta)) = \sum_{i=1}^{\infty} P_i^m(\alpha, \theta).$$

当 θ 趋于无穷大时, $H_m(\mathbf{P}(\alpha, \theta))$ 的渐近行为非常有趣. 具体地, 我们有下面的定理.

定理 4.3 设 $0 \leq \alpha < 1$, $\theta > -\alpha$. 函数 $H_m(\mathbf{P}(\alpha, \theta))$ 满足下列结果:

(1a) (大数定律)

$$H_m(\mathbf{P}(\alpha, \theta)) \rightarrow 0 \quad \text{a.s.} \quad \theta \rightarrow \infty;$$

(1b) (大数定律)

$$\frac{\theta^{m-1}\Gamma(1-\alpha)}{\Gamma(m-\alpha)} H_m(\mathbf{P}(\alpha, \theta)) \rightarrow 1 \quad \text{a.s.} \quad \theta \rightarrow \infty.$$

(2) (中心极限型定理 [22, 34, 35])

$$\sqrt{\theta} \left[\frac{\theta^{m-1}\Gamma(1-\alpha)}{\Gamma(m-\alpha)} H_m(\mathbf{P}(\alpha, \theta)) - 1 \right] \Rightarrow Z_m,$$

其中 Z_m 为正态随机变量, 均值为 0, 方差为

$$\sigma_{\alpha, m}^2 = \frac{\Gamma(1-\alpha)\Gamma(2m-\alpha)}{\Gamma^2(m-\alpha)} + \alpha - m^2.$$

(3) (中偏差原理 [36, 37]) 当 θ 趋于无穷大时, 对任意 $r \in ((m-1)/(2m-1), 1/2)$, 随机变量族

$$\left\{ \theta^r \left(\frac{\theta^{m-1}\Gamma(1-\alpha)}{\Gamma(m-\alpha)} H_m(\mathbf{P}(\alpha, \theta)) - 1 \right) : \theta > 0 \right\}$$

满足大偏差原理, 速率是 θ^{2r-1} , 速度函数为 $\frac{x^2}{2\sigma_{\alpha, m}^2}$.

(4a) (大偏差原理 [21, 38]) 当 θ 趋于无穷大时, $\{H_m(\mathbf{P}(\alpha, \theta)) : \theta > 0\}$ 在空间 $[0, 1]$ 上满足大偏差原理, 其速率为 $1/\theta$, 速度函数为

$$I_6(x) = \begin{cases} -\log(1 - x^{1/m}), & x < 1, \\ +\infty, & \text{其他}; \end{cases}$$

(4b) (大偏差原理 [39]) 当 θ 趋于无穷大时, $\{\frac{\theta^{m-1}\Gamma(1-\alpha)}{\Gamma(m-\alpha)} H_m(\mathbf{P}(\alpha, \theta)) : \theta > 0\}$ 在空间 $\mathbb{R}$ 上满足大偏差原理, 其速率为 $\theta^{-1/m}$, 速度函数为

$$I_7(x) = \begin{cases} \left[\frac{\Gamma(m-\alpha)(x-1)}{\Gamma(1-\alpha)} \right]^{1/m}, & x \geq 1, \\ +\infty, & \text{其他}. \end{cases}$$

其他有关两参数 Poisson-Dirichlet 分布的结果, 可参见文献 [2, 35, 36, 38].

4.2 α 趋于 1

参数 α 不仅代表稳定隶属子的指标, 还与统计物理中的临界现象联系紧密. 众所周知, 自旋玻璃对应于磁性合金材料的一种亚稳定状态. 它是物理中复杂无序系统的典型例子. 为了理解和刻画自旋玻璃, 研究人员引入了许多数学模型, 其中最简单的要数 Derrida^[40] 引进的随机能量模型.

给定一个大整数 $N \geq 1$, 考虑组态空间 $\Sigma_N = \{-1, +1\}^N$. 系统在组态 $\sigma \in \Sigma_N = \{-1, +1\}^N$ 处的能量称为 Hamilton 函数, 记作 $H_N(\sigma)$. 令 T 代表系统温度, 则刻画系统的 Gibbs 测度具有如下形式:

$$m_{N,T}(\sigma) = Z_N^{-1}(T) \exp \left\{ -\frac{1}{T} H_N(\sigma) \right\},$$

其中 $Z_N(T)$ 为配分函数. Derrida 的随机能量模型假设 $\{H_N(\sigma) : \sigma \in \Sigma_N\}$ 为均值是 0、方差为正比于 N 的独立同分布正态随机变量. 取 $S = [0, 1]$, 并做如下变换:

$$\begin{aligned} \Phi : \Sigma_N &\rightarrow S, \quad \sigma \rightarrow \xi_N(\sigma) = 1 - \sum_{i=1}^N (1 - \sigma_i) 2^{-(i+1)}, \\ \mu_{T,N}(\cdot) &= \sum_{\sigma} m_{N,T}(\sigma) \delta_{\xi_N(\sigma)}(\cdot). \end{aligned}$$

如果对系统取热力学极限 (N 趋于无穷大), 那么有下面的相变结果.

定理 4.4 [5] 设 ν 为 S 上的 Lebesgue 测度, 则存在 $0 < T_c < \infty$, 使得下列结果成立:

(1) 对于 $T \geq T_c$, 有

$$\mu_{T,N}(\cdot) \rightarrow \nu(\cdot) \quad \text{a.s.} \quad N \rightarrow \infty;$$

(2) 对于 $T < T_c$, 有

$$\mu_{T,N}(\cdot) \Rightarrow \Xi_{\alpha,0,\nu}, \quad N \rightarrow \infty,$$

其中 $\alpha = \frac{T}{T_c}$.

从上面的定理可以看出, Dirichlet 过程 $\Xi_{\alpha,0,\nu}$ 中的参数 α 代表低温与临界温度的比值. α 趋于 1 的极限过程所对应的是温度从低升到临界温度的过程. 前面我们讲过, 参数 α 反映了系统的多样性. 这一点与上述结果吻合. 因为温度的升高导致系统的熵增大和系统的多样性增强.

从上述讨论可以看出, 考虑 α 趋于 1 的极限过程不仅仅是有趣的数学问题, 更重要的是它具有非常自然的物理意义.

定理 4.5 [41] 设 $S = [0, 1]$, ν 为 S 上的 Lebesgue 测度.

(1) (大数定律) 在空间 $M_1(S)$ 中,

$$\Xi_{\alpha,0,\nu}(\cdot) \rightarrow \nu(\cdot) \quad \text{a.s.} \quad \alpha \rightarrow 1.$$

(2) 当 α 趋于 1 时, $\{\Xi_{\alpha,0,\nu} : \alpha > 0\}$ 在空间 $M_1(S)$ 上满足大偏差原理, 其速率为 $-1/\log(1 - \alpha)$, 速度函数为

$$I_8(\mu) = \begin{cases} 0, & \mu = \nu, \\ n, & \mu = \sum_{i=1}^n p_i \delta_{x_i} + \left(1 - \sum_{i=1}^n p_i\right) \nu, \quad p_1, \dots, p_n > 0, \\ +\infty, & \text{其他.} \end{cases}$$

速度函数 $I_8(\cdot)$ 的有效定义域 $\{\mu \in M_1(S) : I_8(\mu) < \infty\}$ 只含有具有有限原子的概率, 表明在温度从临界值 T_c 降低的瞬间, 系统的微观结构经历了从完全随机的无原子状态向具有有限个原子的状态转移. 当温度真正降低后, 系统的状态将不再含有连续部分而是变成具有可数原子的离散概率. 这与物理的直观一致. 类似的现象也出现在 Poisson-Dirichlet 分布 $\text{PD}(\alpha, 0)$ 的大偏差原理 (参见文献 [42]).

4.3 $\theta \vee \alpha$ 趋于 0

最后考虑 $\theta \vee \alpha$ 趋于 0 的极限行为. 从前面的讨论可以看到, θ 变小意味着突变率变小, 由此带来群体的多样性下降. 而这一点与 α 变小效果一样. 因此, $\theta \vee \alpha$ 趋于 0 的极限行为对应的是系统多样性的消失. 容易看出, 这时 Dirichlet 过程 $\Xi_{\alpha, \theta, \nu_0}$ 趋于一个随机点上的 Dirac 测度, 而且该随机点的分布为 ν_0 . 有鉴于此, 考虑质量的分布或等价的 Poisson-Dirichlet 分布 $\text{PD}(\alpha, \theta)$ 更自然一些.

为了叙述下面的定理, 需要引入如下的空间.

对任意 $n \geq 1$, 定义

$$\sqcup_n = \left\{ (p_1, \dots, p_n, 0, 0, \dots) \in \nabla_\infty : \sum_{i=1}^n p_i = 1, p_n > 0 \right\},$$

$$\sqcup = \bigcup_{i=1}^{\infty} \sqcup_i.$$

定理 4.6 [37, 43] 令

$$\bar{\nabla}_\infty = \left\{ (p_1, p_2, \dots) : p_1 \geq p_2 \geq \dots \geq 0, \sum_{i=1}^{\infty} p_i \leq 1 \right\}.$$

(1) (大数定律) 在空间 $\bar{\nabla}_\infty$ 中, 我们有如下的依概率收敛结果:

$$\mathbf{P}(\alpha, \theta) \rightarrow (1, 0, \dots), \quad \theta \vee \alpha \rightarrow 0.$$

(2) 当 $\theta \vee \alpha$ 趋于 0 时, $\{\mathbf{P}(\alpha, \theta) : \alpha > 0, \theta > -\alpha\}$ 在空间 $\bar{\nabla}_\infty$ 上满足大偏差原理, 其速率为 $-1/\log(\theta \vee \alpha)$, 速度函数为

$$I_9(\mathbf{p}) = \begin{cases} 0, & \mathbf{p} = (1, 0, \dots), \\ n-1, & \mathbf{p} \in \sqcup_n, \quad n \geq 2, \\ +\infty, & \mathbf{p} \notin \sqcup, \end{cases}$$

为了比较突变和自然选择在群体演进过程中的作用, 文献 [44] 考虑了一类带自然选择的 Poisson-Dirichlet 分布. 作为应用, 文献 [37, 43] 也得到了这类带自然选择的 Poisson-Dirichlet 分布的大偏差结果.

致谢 感谢严士健教授多年的教育、指导和帮助. 衷心祝贺先生 90 华诞.

参考文献

- 1 Pitman J, Yor M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann Probab*, 1997, 25: 855–900
- 2 Feng S. The Poisson-Dirichlet Distribution and Related Topics. Heidelberg: Springer, 2010
- 3 Arratia R, Barbour A D, Tavaré S. Logarithmic Combinatorial Structures: A Probabilistic Approach. EMS Monographs in Mathematics. Zürich: European Mathematical Society, 2003
- 4 Bertoin J. Random Fragmentation and Coagulation Processes. Cambridge Studies in Advanced Mathematics, vol. 102. Cambridge: Cambridge University Press, 2006
- 5 Bovier A. Statistical Mechanics of Disordered Systems: A Mathematical Perspective. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 18. Cambridge: Cambridge University Press, 2006
- 6 Gnedin A, Iksanov A, Marynych A. Λ -coalescents: A survey. *J Appl Probab*, 2014, 51: 23–40

- 7 Pitman J. Combinatorial Stochastic Processes. New York: Springer-Verlag, 2006
- 8 Sethuraman J A. A constructive definition of Dirichlet priors. *Statist Sinica*, 1994, 4: 639–650
- 9 Ferguson T S. A Bayesian analysis of some nonparametric problems. *Ann Statist*, 1973, 1: 209–230
- 10 Pitman J. The Two-Parameter Generalization of Ewens' Random Partition Structure. Technical Report 345. Berkeley: University of California, 1992
- 11 Brix A. Generalized Gamma measures and shot-noise Cox processes. *Adv in Appl Probab*, 1999, 31: 929–953
- 12 Lijoi A, Mena R H, Prünster I. Controlling the reinforcement in Bayesian non-parametric mixture models. *J R Stat Soc Ser B Stat Methodol*, 2007, 69: 715–740
- 13 Kingman J F C. Random discrete distributions. *J R Stat Soc Ser B Stat Methodol*, 1975, 37: 1–22
- 14 Ewens W J. The sampling theory of selectively neutral alleles. *Theoret Popul Biol*, 1972, 3: 87–112
- 15 Goncharov V L. Some facts from combinatorics. *Izvestia Akad Nauk SSSR Ser Mat*, 1944, 8: 3–48
- 16 Shepp L A, Lloyd S P. Ordered cycle lengths in a random permutation. *Trans Amer Math Soc*, 1966, 121: 340–357
- 17 Feng S, Hoppe F M. Large deviation principles for some random combinatorial structures in population genetics and Brownian motion. *Ann Appl Probab*, 1998, 8: 975–994
- 18 Delaurentis J M, Pittel B G. Random permutations and Brownian motion. *Pacific J Math*, 1985, 119: 287–301
- 19 Hansen J C. A functional central limit theorem for the Ewens sampling formula. *J Appl Probab*, 1990, 27: 28–43
- 20 Favaro S, Feng S, Gao F. Moderate deviations for Ewens-Pitman sampling models. *Sankhya A*, 2018, 80: 330–341
- 21 Feng S. Large deviations associated with Poisson-Dirichlet distribution and Ewens sampling formula. *Ann Appl Probab*, 2007, 17: 1570–1595
- 22 Joyce P, Krone S M, Kurtz T G. Gaussian limits associated with the Poisson-Dirichlet distribution and the Ewens sampling formula. *Ann Appl Probab*, 2002, 12: 101–124
- 23 Joyce P, Krone S M, Kurtz T G. When can one detect overdominant selection in the infinite-alleles model? *Ann Appl Probab*, 2003, 13: 181–212
- 24 Favaro S, Feng S. Asymptotics for the conditional number of blocks in the Ewens-Pitman sampling model. *Electron J Probab*, 2014, 19: 1–15
- 25 Favaro S, Feng S. Large deviation principles for the Ewens-Pitman sampling models. *Electron J Probab*, 2015, 20: 1–26
- 26 Favaro S, Feng S, Jenkins P A. Bayesian nonparametric analysis of Kingman's coalescent. *Ann Inst H Poincaré Probab Statist*, 2019, in press
- 27 James L F. Large sample asymptotics for the two-parameter Poisson Dirichlet processes. In: *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, vol. 3. Beachwood: Institute of Mathematical Statistics Collections, 2008, 187–199
- 28 Lynch J, Sethuraman J. Large deviations for processes with independent increments. *Ann Probab*, 1987, 15: 610–627
- 29 Dawson D A, Feng S. Large deviations for the Fleming-Viot process with neutral mutation and selection. *Stochastic Process Appl*, 1998, 77: 207–232
- 30 Dawson D A, Feng S. Large deviations for the Fleming-Viot process with neutral mutation and selection, II. *Stochastic Process Appl*, 2001, 92: 131–162
- 31 Simpson E H. Measurement of diversity. *Nature*, 1949, 163: 688
- 32 Herfindahl O C. Concentration in the U.S. steel industry. PhD Thesis. New York: Columbia University, 1950
- 33 Hirschman A O. National Power and the Structure of Foreign Trade. Berkeley: University of California Press, 1945
- 34 Griffiths R C. On the distribution of allele frequencies in a diffusion model. *Theoret Popul Biol*, 1979, 15: 140–158
- 35 Handa K. The two-parameter Poisson-Dirichlet point process. *Bernoulli*, 2009, 15: 1082–1116
- 36 Feng S, Gao F. Moderate deviations for Poisson-Dirichlet distribution. *Ann Appl Probab*, 2008, 18: 1794–1824
- 37 Feng S, Gao F. Asymptotic results for the two-parameter Poisson-Dirichlet distribution. *Stochastic Process Appl*, 2010, 120: 1159–1177
- 38 Dawson D A, Feng S. Asymptotic behavior of Poisson-Dirichlet distribution for large mutation rate. *Ann Appl Probab*, 2006, 7: 562–582
- 39 Dawson D A, Feng S. Large deviations for homozygosity. *Electron Comm Probab*, 2016, 21: 1–8
- 40 Derrida B. Random-energy model: Limit of a family of disordered models. *Phys Rev Lett*, 1980, 45: 79–82
- 41 Feng S, Gao F, Zhou Y. Limit theorems associated with the Pitman-Yor process. *Adv in Appl Probab*, 2017, 49: 581–602
- 42 Feng S, Zhou Y. Asymptotic behavior of Poisson-Dirichlet distribution and random energy model. In: *XI Symposium on Probability and Stochastic Processes. Progress in Probability*, vol. 69. Basel: Birkhäuser, 2015, 141–155
- 43 Feng S. Poisson-Dirichlet distribution with small mutation rate. *Stochastic Process Appl*, 2009, 119: 2082–2094
- 44 Gillespie J H. The role of population size in molecular evolution. *Theoret Popul Biol*, 1999, 55: 145–156

Asymptotic behavior of Dirichlet process and related random measures

Shui Feng

Abstract Dirichlet process is an atomic-valued random measure. It was introduced by Ferguson (1973) with motivation from Bayesian statistics. Pitman and Yor generalized the process to the two-parameter setting in 1990s. It has since found applications in a wide range of subjects including Bayesian statistics, biology, communication, economy, finance, genetics, and machine learning. Driven by various applications, there has been active study of asymptotic behavior of this class of processes. This paper provides a brief survey on recent work by the author and collaborators.

Keywords Dirichlet process, two-parameter Dirichlet process, Poisson-Dirichlet distribution, Ewens-Pitman sampling formula, large deviations, moderate deviations, central limit theorem

MSC(2010) 60F10, 92D10

doi: 10.1360/N012018-00210