arXiv:2303.08139v1 [math.ST] 14 Mar 2023

广义逆高斯-泊松分布的极限形状

Leonid V. Bogachev*、Ruheyan Nuermaimaiti†和
Jochen Voss‡利兹大学数学学院，英国利兹 *LS2 9JT*

摘要

Sichel 于 20 世纪 70 年代提出的广义逆高斯-泊松分布（GIGP 分布），已被证明是一种灵活的拟合工具，适用于各种频数数据；这些数据统称为物品生产模型。本文确定了来自 GIGP 分布的随机样本经适当缩放后的图示表示（称为杨图）的极限形状（由不完全伽马函数给出）。我们还证明，涨落是渐近正态的；此外，相应的经验随机过程可通过反向时间下经缩放的布朗运动来近似，其非齐次时间尺度由极限形状决定。这里，极限是在生产来源数量趋于无穷大的同时取得的，并且内禀参数所处的范围确保每个来源的平均项目数较大。更确切地说，只有当这种联合增长足够快时，才会收敛到极限形状。在相反的、称为“混沌”的范围内，经验随机过程可通过反向时间下的非齐次泊松过程来近似。本文通过计算机模拟以及信息计量学中的一些经典数据集展示了这些结果。

关键词：计数数据；频数分布；来源/物品模型；广义逆高斯-泊松分布；杨图；极限形状；信息计量学数据

*2020*年数学学科分类：主要分类 62E17；次要分类 62P25

1. 引言

在许多应用场景中，人们会处理计数数据，其形式为在可数多个类别（“盒子”）之一中出现的样本频数，例如用索引$j \in \mathbb{N} = \{1, 2, \dots\}$或 $j \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$标记。通常，将每个盒子 j 中的出现次数解释为多个贡献“来源”之一所产生的相应批量“物品”数是恰当的；例如，落入盒子 $j = 0$ 表示该来源未产生任何物品。于是，观测到的数据集就是向量 (M_j)，其中记录了

*L.V.Bogachev@利兹.ac.uk，ORCID 0000-0002-2365-2621
†通讯作者，mmrn@leeds.ac.uk，ORCID 0000-0002-5764-9949
‡J.Voss@leeds.ac.uk，ORCID 0000-0002-2323-3814

在总数为 $M = \sum_j M_j$ 个来源中，每个盒子里的计数，物品总数为 $N = \sum_j j M_j$。

这类情形在现实生活中的例子多种多样，包括：不同物种（如蝴蝶）的丰度数据，其中将不同物种视为来源，将观测到的计数视为物品；推特上不同账户的关注者数量（物品），其中账户为来源；重复购买数据，其中家庭为来源，购买的单位数量为物品；作者发表的论文数量（论文为物品）；等等。[1] 在最后一个例子中，论文本身也可以充当来源，而引文则是物品。总而言之，这构成了一个有趣的关系三角：作者–论文–引文（*APC*），这是信息计量学 [14]的主要研究主题之一。

处理这类数据时，一个自然的目标是通过拟合合适的分布模型（f_j）来解释观测到的（相对）频数（M_j/M），且该模型最好具有一定的概念基础，并适用于多种情形。当然，在任何现实数据集中，当取值 j 足够大时，数值 M_j 都会降至零；但这与模型预测 $M_j/M \approx f_j$ 并不矛盾，因为理论频数f_j 会随着 $j \to \infty$而趋于零。不过，恰当拟合长尾频率对于理解计数数据中极端值的行为十分重要（例如异常高的引文数）。

理论频率模型的一个著名例子是幂律。Lotka [25] 最初提出这一模型，用以描述化学和物理学领域的出版统计；其依据是经验观察，即样本频数 M_j/M 近似服从形式为 $f_j = c_\alpha j^{-\alpha}$ $(j \in \mathbb{N})$ 的幂律分布，其中指数为 $\alpha > 1$，归一化常数为 [2] $c_\alpha > 0$，且满足 $\sum_{j=1}^{\infty} f_j = 1$ [11, 14]。

将幂律模型拟合到计数数据上的一种直观启发式方法，是查看两个坐标轴均采用对数刻度的频率图（例如直方图），并寻找一条拟合直线，其斜率为 $-\alpha$ [26]。另一种方法 [10], 能够对离散数据进行有效平滑，即使用互补累积频率$F_j = \sum_{\ell \geq j} f_\ell$；同样在双对数图中，良好的拟合对应一条斜率为 $1 - \alpha$ 的直线。更正式地说，可以使用最大似然法或普通最小二乘估计 [26]等标准统计方法拟合该模型。

关于幂律普适性的传统解释基于累积优势原则，也常用“成功带来成功”这一口号来表述。该说法最初是在科学生产力的语境下提出的 [31, 32, 15]（另见较新的综述 [20] ，其中对累积优势进行了批评）。遗憾的是，Lotka 幂律在实际数据建模中的用途往往受到限制，因为它通常只能在较小的计数值范围内很好地拟合数据，因此需要截去较低的值（见 [10] 中的详细讨论），或者截去较高的（长尾）值；后者用拉伸指数律 [24]描述得更好。

其他许多尝试也曾使用负二项分布、修正几何分布、贝塔二项分布等理论分布来拟合各种计数数据集 [21] （讨论及更多参考文献见

[1]更多示例和参考文献请参见，例如 Egghe 的专著 [14] 或 Clauset 等人撰写的综述论文 [10]。[2]归一化常数 c_α 可通过黎曼ζ函数 $c_\alpha^{-1} = \sum_{j=1}^{\infty} j^{-\alpha} = \zeta(\alpha)$ 表示。在实际案例中，幂律指数通常处于范围 $2 < \alpha < 3$ [10]。

[41, 20]），然而，这些分布族都未能充分展现出解释各类计数数据集的“普适性”，往往无法捕捉众数特征和长尾行为等某些特征。

在一系列论文中，Sichel [36, 37, 38, 39, 40, 41] 提出并发展了所谓的广义逆高斯-泊松（*GIGP*）模型，旨在通过考虑来源（例如论文和作者）个体生产力的统计差异，合理刻画物品的产出。更确切地说，假设某个来源按照速率为 λ 的泊松定律产出物品，而该速率本身是随机的，服从特定的广义逆高斯分布（*GIG*）密度 [36, 22]。换言之，GIGP 分布是在 GIG混合密度下的混合泊松分布（参见 [18] 关于长尾计数数据的泊松混合模型综述）。Sichel 将其 GIGP 模型应用于各种用例和多个数据集，包括书面散文中的句长和词频 [38, 39] 、含金刚石矿床中发现的石头数量 [37] 以及科学产出（论文和／或引文） [41]。这些实例展现了 GIGP分布族出色的灵活性和多用途性。

在一项较新的研究中，Yong [49] 提出使用随机整数分拆的组合模型来模拟引文计数数据，其中整数分拆的各组成部分分别表示作者的论文及其对应的引文数。这种方法被认为的主要优势，是能够利用关于经适当缩放的极限形状 的知识；这些杨图展示了（随机）整数分拆中的各个部分，进而可以对某些引文指标进行统计估计，例如 Hirsch [19]提出的 h-指数[3] 。具体而言，Yong 注意到， h-指数在几何上对应于杨图上边界上横纵坐标相等的位置，并利用在两个坐标轴上均按 $\sqrt{N}$缩放时极限形状的显式方程（其中 $N \gg 1$ 为引文总数），提出了 h-指数的一种简单估计， — $h \approx 0.54\sqrt{N}$，随后使用若干数学引文数据集 [49]对其进行了检验。
—

本文将极限形状的概念应用于从 GIGP 分布中抽取的随机样本，并以其期望值 η（以及来源数量 M）作为大参数。在适当归一化后，我们得到极限形状的显式公式，该公式由上不完全伽马函数给出： $\varphi_\nu(x) = \int_x^\infty s^{\nu-1}\,\mathrm{e}^{-s}\,\mathrm{d}s\;(x>0)$，其索引为 GIGP 分布的形状参数 $\nu \geq -1$。在实证数据分析中，这相当于对互补累积频率图进行相应缩放，从而便于快速直观地检验 GIGP 模型的拟合优度；推荐的缩放方式还能提供额外的参考信息。基于渐近置信区间，对误差界进行更细致的分析，也因我们关于极限形状 $\varphi_\nu(x)$ 周围渐近高斯涨落的结果而成为可能。

由预测的极限形状 $\varphi_\nu(x)$ 可知，GIGP 模型的上尾呈现幂律调制的指数衰减，因此显著偏离幂律行为。在大多数实际例子中，由于较大计数值稀少（或缺失），这种差异并不重要；然而，如何充分建模极端值仍是一个值得探讨的问题，而上文提到的拉伸指数模型是一个颇具吸引力的替代方案 [24]。

[3] h 指数定义为作者发表的论文中被引用至少 h 次的论文数量的最大值 h 。

请注意，对于计数数据模型而言，具有非平凡缩放的极限形状这一重要性质并非必然成立；例如，一个显著的例外是幂律分布 $f_j = cj^{-\alpha}$，因为它具有尺度无关性：事实上，对于任意缩放因子 $A > 0$，都有 $A^{\alpha} f_{Ax} \equiv f_x$ $(x > 0)$。

在这方面，我们提及 Nuermaimaiti 等人近期提出并研究的广义幂律（*GPL*）模型。 [29], 该模型旨在弥合计数的小值（在幂律拟合中经常被截断）与幂律型上尾之间的差距。GPL 模型的概念依据同样建立在 Sichel [38], 所采用的混合思想之上，但源生成律（采用几何分布而非泊松分布）和混合密度（采用贝塔分布而非广义逆高斯分布）有所不同。GPL 模型的极限形状存在，且由 $\varphi(x;\alpha) = (1+x)^{1-\alpha}$ $(x \geq 0)$ 给出，其中 $\alpha > 0$ 是 GPL 的形状参数。尽管 GPL 极限形状的长尾具有与上述幂律情形相同的幂律衰减，但关键区别在于 GPL 并非尺度自由的，这确保了相应的缩放并非平凡。

本文其余部分安排如下。在第 2.1 节中，我们定义物品生产的主要模型，并给出基本符号和一些初等关系。第 2.2 节介绍杨图和极限形状的概念，第 2.3 节则通过经典的整数分拆模型加以说明。在第 3.1 节中，我们定义参数为 $\nu \in \mathbb{R}$、 $\alpha > 0$ 和 $0 < \theta < 1$ 的 GIGP 分布；第 3.2 节通过对边界情形 $\alpha = 0$ 的具体规定加以补充；随后在第 3.3 节对所需极限 $\theta \to 1-$ 下 GIGP 期望值进行渐近分析（命题 3.2）。具体而言，该分析解释了我们为何施加限制 $\nu \geq -1$。

完成这些准备后，我们在第 4 节讨论极限形状问题。首先，在第 4.1 节中定义合适的缩放系数 A 和 B（假设 4.2），并陈述主要结果定理 4.1：缩放后的杨图依概率一致收敛至极限形状 0 $\varphi_\nu(x)$ $\nu = 0.5$ $\nu = -0.5$ （） （适用于 $x \geq \delta$，且 $\delta > 0$ 任意） $= \int_x^\infty s^{\nu-1} \mathrm{e}^{-s} \mathrm{d}$ （ ）。文中还通过两个示例情形的计算机模拟对此加以说明，分别为 和 。重要的是，要使该结果成立， y-缩放系数 B 必须趋于无穷（假设 4.3），这进而意味着物品生产模型中的来源数量 M 必须足够大。第 4.3 节首先证明杨图的期望收敛至 φ_ν（x）（定理 4.2）。随后，第 4.4 节建立逐点依概率收敛（定理 4.4）。在第 4.5 节中，我们构造一个反向时间 $t = 1/x$ 上的合适鞅（引理 4.5），从而能够通过应用 Doob–Kolmogorov 次鞅不等式 [48, Sec. 4.6] 来证明定理 4.1。在第 4.7 节中，我们借助将杨图边界解释为经验过程，并利用已有的集中不等式 [8]，给出定理 4.1 的另一种证明。

第 5 节研究随机杨图的涨落，并建立了逐点渐近正态性（定理 5.1），同时在定理 5.2 和定理 5.3 中通过将极限刻画为高斯过程，对这一结果进行了统一。当来源数量 M 增长得不够快，导致无法收敛到极限形状时（称为“混沌”状态），第 6 节将对此进行探讨，并通过计算机模拟说明，泊松近似可以作为确定性极限的合适替代（定理 6.1、6.2 和 6.3）。第 7 节将极限形状结果应用于 Sichel [41] 此前考察过的一些现实频率数据集，并以此作为检验所提出的 GIGP

分布拟合优度的测试平台。本文旨在说明，预测得到的极限形状与适当的（A, B）缩放（基于估计或预先设定的 GIGP 参数）相结合，可作为一种实用的可视化工具，快速而直观地检验 GIGP 模型是否适用。第 8 节总结了本文的主要发现。最后，附录 A 简要汇编了定义 GIGP 分布时涉及的贝塞尔函数 $K_\nu(z)$ 的渐近公式。这些公式适用于小自变量 z 或大阶数 ν，并广泛用于我们的渐近分析。

2. 场景设定

2.1. 物品生产模型

假设有 M 个来源，每个来源生产一批物品，并令 X_i 表示第 i 个来源所生产的一批物品的随机大小（$i = 1, \ldots\ldots, M$）。如果允许输出为空（例如论文的引文），输出大小的取值范围可以是 $j \in \mathbb{N}_0$；也可以是零截断的，此时 $j \in \mathbb{N}$（例如某位作者的论文）。各来源彼此独立，且其随机输出遵循相同的频数分布（f_j）；也就是说，随机变量（X_i）相互独立，并且对每个 $i = 1, \ldots\ldots, M$,

$$\mathsf{P}(X_i = j) = f_j \qquad (j \in \mathbb{N}_0).$$

注 2.1. 为简化记号，我们仍记为 $j \in \mathbb{N}_0$，其中零截断情形包含在内，此时 $f_0 = 0$。

注 2.2. 上文介绍的物品生产模型 可以改述为经典的占位问题：根据概率分布（f_j）[17]，将 M 个粒子独立地分配到无限多个盒子中。

我们假设分布 (f_j) 具有有限均值,

$$\eta := \mathsf{E}(X_i) = \sum_{j=0}^{\infty} j f_j < \infty. \tag{2.1}$$

产出物品的总数（随机变量）等于各项产出的总和,

$$N = \sum_{i=1}^{M} X_i, \tag{2.2}$$

其期望值为

$$\mathsf{E}(N) = \sum_{i=1}^{M} \mathsf{E}(X_i) = M\eta. \tag{2.3}$$

通过在所有可能取值 j 上“扫描”来表示每个 X_i 是很有用的,

$$X_i = \sum_{j=0}^{\infty} j I_{\{X_i = j\}}, \tag{2.4}$$

其中 I_A 表示事件 A 的示性函数（即，当 A 发生时取值为 1，否则取值为 0）。当然，

$$\mathsf{E}\big(I_{\{X_i=j\}}\big) = \mathsf{P}(X_i = j) = f_j \qquad (j \in \mathbb{N}_0). \tag{2.5}$$

考虑 汇总生产中输出大小为 $j \in \mathbb{N}_0$ 的物品 (X_i) 的重数 M_j ，

$$M_j := \#\big\{i \in \{1, \dots, M\} \colon X_i = j\big\} = \sum_{i=1}^{M} I_{\{X_i=j\}} \qquad (j \in \mathbb{N}_0). \tag{2.6}$$

利用 (2.5)，我们可以得到期望

$$\mathsf{E}(M_j) = \sum_{i=1}^{M} \mathsf{E}\big(I_{\{X_i=j\}}\big) = M f_j \qquad (j \in \mathbb{N}_0). \tag{2.7}$$

注意，随机变量 (M_j) 并不相互独立；事实上，它们的和等于来源的数量，

$$\sum_{j=0}^{\infty} M_j = \sum_{j=0}^{\infty} \sum_{i=1}^{M} I_{\{X_i=j\}} = \sum_{i=1}^{M} \sum_{j=0}^{\infty} I_{\{X_i=j\}} = \sum_{i=1}^{M} 1 = M.$$

从重数 M_j 的解释可以明显看出，生产出的物品总数（随机）为

$$N = \sum_{j=0}^{\infty} j M_j. \tag{2.8}$$

利用定义 (2.2) 以及分解式 (2.4) 和 (2.6)，也可以轻易得到相同结果，

$$N = \sum_{i=1}^{M} X_i = \sum_{i=1}^{M} \sum_{j=0}^{\infty} j I_{\{X_i=j\}} = \sum_{j=0}^{\infty} j \sum_{i=1}^{M} I_{\{X_i=j\}} = \sum_{j=0}^{\infty} j M_j.$$

N 的期望值可由（2.7）和（2.1）表示为，

$$\mathsf{E}(N) = \sum_{j=0}^{\infty} j \, \mathsf{E}(M_j) = M \sum_{j=0}^{\infty} j f_j = M\eta, \tag{2.9}$$

当然，这与（2.3）相同。

注释 2.3. 根据公式 (2.3) 和 (2.9)，样本均值 $\hat{}\eta = N/M$ 是期望值 η的无偏估计量，具有一致性和渐近正态性等所有标准性质。该估计量的优点在于它是非参数的，也就是说，无需了解生产输出数据背后的任何分布模型（f_j）。

2.2. 杨图与极限形状

按照产出对来源进行排序很有用，也就是考虑（降序排列的）顺序统计量 $X_{1,M} \geq X_{2,M} \geq \cdots \geq X_{M,M}$；例如，$X_{1,M} = \max_{1 \leq i \leq M}\{X_i\}$ 是 M 个来源中最高的产出得分。产出剖面可以通过杨图 简洁地呈现；该杨图由左对齐、底部对齐、单位高度且长度分别为 $X_{1,M} \geq X_{2,M} \geq \cdots$ 的行块构成，较长的行块位于较低处（见图 1(a)）。特别是，对应于产出值 $j = 0$ 的行块（若允许）会退化为高度各为 1 的垂直区间，并沿纵轴置于杨图其余部分的上方。

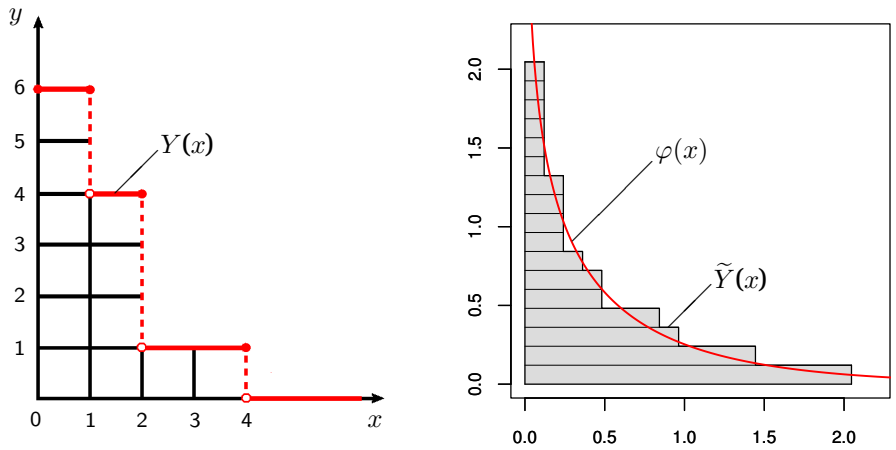


图1：（a）对于 $M = 6$ 个来源和有序输出（$X_{i,M}$），其杨图及边界 $Y(x)$，对应计数 $M_4 = 1$、$M_2 = 3$、$M_1 = 2$。（b）随机占位问题的极限形状 $y = \varphi(x)$（即计数相互独立的情形 M_j，详见第 2.3 节），由方程 $e^{-x\pi/\sqrt{6}} + e^{-y\pi/\sqrt{6}} = 1$ 给出（见（2.30））。阴影区域表示沿两轴按 $\sqrt{n}$ 缩放后的模拟杨图，其中 $n = 100$；水平块对应有序输出 $X_{i,M}$（样本值为 $M = 17$）。

杨图的上边界是（左连续）阶梯函数的图像

$$Y(x) := \sum_{j \geq x} \sum_{i=1}^{M} I_{\{X_i = j\}} = \sum_{j \geq x} M_j \qquad (x \geq 0) \tag{2.10}$$

（见（2.6））。为突出对 x 的依赖，将定义（2.10）改写为以下形式 m

$$Y(x) = \sum_{j=0}^{\infty} M_j \, \mathbf{1}_{[0,j]}(x) \qquad (x \geq 0), \tag{2.11}$$

其中，$\mathbf{1}_D(x)$ 是集合 D 的示性函数（即，当 $x \in D$ 时，$\mathbf{1}_D(x) = 1$；否则，$\mathbf{1}_D(x) = 0$）。

若 $M_0 > 0$，则函数 $Y(x)$ 在 $x = 0$ 处有一个孤立峰；否则， $Y(x)$ 在零点处右连续。原点处的值是来源的总数，

$$Y(0) = \sum_{j=0}^{\infty} M_j = M,$$

wh而图像 $Y(x)$ 下方的面积等于所产生物品的总数 s,

$$\int_0^{\infty} Y(x)\,\mathrm{d}x = \sum_{j=0}^{\infty} M_j \int_0^{\infty} \mathbf{1}_{[0,j]}(x)\,\mathrm{d}x = \sum_{j=0}^{\infty} j\,M_j = N$$

（见（2.11）和（2.8））。

令

$$Z_i(x) := \sum_{j \geq x} I_{\{X_i = j\}} = I_{\{X_i \geq x\}} \qquad (i = 1, \ldots, M), \tag{2.12}$$

公式（2.10）可表示为

$$Y(x) = \sum_{i=1}^{M} \sum_{j \geq x} I_{\{X_i = j\}} = \sum_{i=1}^{M} Z_i(x). \tag{2.13}$$

示性函数 $Z_1(x) \ldots, Z_M(x)$ 是独立同分布的伯努利随机变量；具体而言，

$$\mathsf{P}(Z_i(x) = 1) = \mathsf{P}(X_i \geq x) = \sum_{j \geq x} f_j =: \bar{F}(x), \tag{2.14}$$

$$\mathsf{P}(Z_i(x) = 0) = \mathsf{P}(X_i < x) = \sum_{j < x} f_j =: F(x), \tag{2.15}$$

其中 $\bar{F}(x) + F(x) = 1$ 对所有 $x \geq 0$ 均成立。因此，

$$\mathsf{E}\big(Z_i(x)\big) = \bar{F}(x), \qquad \mathsf{Var}\big(Z_i(x)\big) = \bar{F}(x)\big(1 - \bar{F}(x)\big) = \bar{F}(x)F(x),$$

并且，对于任意 $0 \leq x \leq x'$，

$$\mathsf{Cov}\big(Z_i(x), Z_i(x')\big) = \bar{F}(x') - \bar{F}(x)\,\bar{F}(x') = \bar{F}(x')F(x).$$

由此很容易从（2.13）得出，对于每个 $x \geq 0$，

$$\mathsf{E}\big(Y(x)\big) = M\bar{F}(x), \qquad \mathsf{Var}\big(Y(x)\big) = M\bar{F}(x)F(x). \tag{2.16}$$

并且，对于 $0 \leq x \leq x'$，

$$\begin{aligned} \mathsf{Cov}\big(Y(x), Y(x')\big) &= \sum_{i,i'=1}^{M} \mathsf{Cov}\big(Z_i(x), Z_{i'}(x')\big) \\ &= \sum_{i=1}^{M} \mathsf{Cov}\big(Z_i(x), Z_i(x')\big) = M\bar{F}(x')F(x), \end{aligned} \tag{2.17}$$

通过观察采用缩放系数 A 和 B 的缩放杨图，可以直观地了解生产分布的结构，

$$\widetilde{Y}(x) = \frac{1}{B}\,Y(Ax) = \frac{1}{B}\sum_{j\geq Ax} M_j = \frac{1}{B}\sum_{i=1}^{M} Z_i(Ax) \qquad (x \geq 0). \tag{2.18}$$

目标是寻找一个极限形状 $x \mapsto \varphi(x)$，使得在适当的 $A,\ B \to \infty$条件下，

$$\mathsf{E}\big(\widetilde{Y}(x)\big) \to \varphi(x) \qquad (x > 0), \tag{2.19}$$

~此外，$Y(x)$ 依概率收敛到 $\varphi(x)$，也就是说，对于任意 $\varepsilon > 0$，

$$\mathsf{P}\big(|\widetilde{Y}(x) - \varphi(x)| > \varepsilon\big) \to 0 \qquad (x > 0). \tag{2.20}$$

注 2.4。将（2.19）和（2.20）中的收敛范围限制为 $x > 0$，原因在于某些情况下，$\varphi(0) = \infty$（例如，见图1(b)）。

极限形状这一概念受到随机整数分拆理论中类似主题的启发 [43, 44]。作为说明，第 2.3 节简要回顾了这个经典例子，尽管那里的设定与物品生产模型略有不同。本文针对第 3 节引入的 GIGP 分布研究这一问题。

2.3. 示例：整数分拆的极限形状

为说明极限形状的概念，我们从幂律频数分布的一个基准示例开始，$f_j = j^{-a}/\zeta(a)$ $(j \geq 1)$，其中 $a > 1$。任取满足 $B := M/A^{a-1} \to \infty$ 的 $A \to \infty$；也就是说，$1 \ll A \ll M^{1/(a-1)}$。于是，缩放后的期望杨图边界函数化简为

$$\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{A^{a-1}}{M}\sum_{j\geq Ax} M j^{-a}/\zeta(a) = \frac{1}{\zeta(a)}\sum_{j/A\geq x}\left(\frac{j}{A}\right)^{-a}\frac{1}{A} \tag{2.21}$$

$$\to \frac{1}{\zeta(a)}\int_x^\infty s^{-a}\,\mathrm{d}s = \frac{x^{-(a-1)}}{(a-1)\,\zeta(a)}, \tag{2.22}$$

利用 (2.21) 中的和是 (2.22) 中积分的黎曼积分和。因此，极限形状存在，且由 (2.22) 右侧给出；但如引言中所述，这并无实际用途，因为只要 $A = o(M^{1/(a-1)})$，缩放参数 $A \to \infty$ 就是任意的（这证实了幂律分布具有尺度不变性）。

具有有意义极限形状的频率模型的经典例子来自随机整数分拆理论。在这里，数值 $j = 1$、2、.....被视为整数分拆中的候选部分，相应的重数 M_j 则分别表示部分 j 被使用的次数。特别是，如果 $M_j = 0$，那么数值 j就不会出现在分拆中，并且默认只有有限多个 M_j 非零。总和 $N = \sum_{j=1}^{\infty} j\,M_j$ 给出了被分拆为各部分 j 之和的整数，其中 $M_j > 0$。

这里的标准模型设定不同于第 2.1 节中描述的物品生产模型。具体而言，随机分拆模型并非基于有 M 个独立源、且重数 (M_j) 由公式 (2.6) 表示这一前提，而是假定重数 (M_j) 是服从几何分布的独立随机变量，$M_j \sim \mathsf{Geom}(1-z^j)$ ($j \geq 1$)，即

$$\mathsf{P}(M_j = m) = z^{jm}(1-z^j) \qquad (m \geq 0), \tag{2.23}$$

其期望值为

$$\mathsf{E}(M_j) = \frac{z^j}{1-z^j} \qquad (j \geq 1). \tag{2.24}$$

参数 $z \in (0,1)$ 特意选为

$$z = \mathrm{e}^{-\kappa/\sqrt{n}}, \qquad \kappa := \frac{\pi}{\sqrt{6}} = \sqrt{\zeta(2)}, \tag{2.25}$$

其中 n 是一个外部（较大）参数。

注意，对于任意 $z \in (0,1)$,

$$\mathsf{P}(M_j > 0) = 1 - \mathsf{P}(M_j = 0) = 1 - (1-z^j) = z^j,$$

and

$$\sum_{j=1}^{\infty} \mathsf{P}(M_j > 0) = \sum_{j=1}^{\infty} z^j = \frac{z}{1-z} < \infty.$$

因此，根据博雷尔–坎特利引理（例如，参见 [35, 第 II.10 节，第 255] 页），随机重数序列 (M_j) 中非零项的个数几乎必然有限。

由于 M_j 与几何边缘分布 (2.23) 相互独立，给定重数序列 $M_j = m_j$ ($j \geq 1$)（其中非零项有限）的概率可表示如下,

$$\mathsf{P}(M_j = m_j,\, j = 1, 2, \dots) = \prod_{j=1}^{\infty} z^{j m_j}(1-z^j) = \frac{z^N}{G(z)}, \tag{2.26}$$

其中 $N = \sum_{j=1}^{\infty} j\, m_j$ 和

$$G(z) = \prod_{j=1}^{\infty} \frac{1}{1-z^j} \qquad (0 < z < 1).$$

公式 (2.26) 是所谓玻尔兹曼分布的一个实例，其根源在于统计物理 [2, 45]，并在概率组合学 [1] 和计算领域 [13]有诸多应用。

选择玻尔兹曼分布（2.26）的原因在于，对其进行条件化后，会在相应的子空间上得到均匀分布。具体而言，记 Π_n 为 n的所有整数分拆构成的集合，不难看出，条件

Π_n 中具有特定部分重数的任意分拆的概率，在给定 $N=\sum_{j=1}^{\infty} jM_j = n$的条件下为

$$\mathsf{P}\big(M_j = m_j,\, j\ge 1 \,\big|\, N=\textstyle\sum_j jM_j = n\big) = \dfrac{z^n/G(z)}{(z^n/G(z))\cdot \#\Pi_n} = \dfrac{1}{\#\Pi_n},$$

它是在 Π_n 上的均匀分布。此外，渐近形式 (2.25) 中参数 z 的选取，可由自然的校准条件解释

$$\mathsf{E}(N) = \mathsf{E}\Big(\textstyle\sum_{j=1}^{\infty} jM_j\Big) \sim n \qquad (n\to\infty). \tag{2.27}$$

事实上，利用均值公式 (2.24)，并设参数 z 具有形式 $z=\mathsf{e}^{-\alpha_n}$，其中 $\alpha_n\to 0$，渐近方程 (2.27) 可改写为

$$\mathsf{E}(N) = \sum_{j=1}^{\infty} \frac{j\,\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} = \frac{1}{\alpha_n^2}\sum_{j=1}^{\infty} \frac{\alpha_n j\,\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}\,\alpha_n \sim n. \tag{2.28}$$

注意到 (2.28) 中的和是一个黎曼积分和，因此可得

$$\begin{aligned}\sum_{j=1}^{\infty} \frac{\alpha_n j\,\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}\,\alpha_n &\to \int_0^{\infty} \frac{s\,\mathrm{e}^{-s}}{1-\mathrm{e}^{-s}}\,\mathrm{d}s \\ &= \sum_{\ell=1}^{\infty}\int_0^{\infty} s\,\mathrm{e}^{-\ell s}\,\mathrm{d}s = \sum_{\ell=1}^{\infty}\frac{1}{\ell^2} = \zeta(2) = \frac{\pi^2}{6} = \kappa^2.\end{aligned}$$

将此代入方程 (2.28)，得到 $\alpha_n\sim \kappa/\sqrt{n}$，这与 (2.25) 一致。

现在可以轻松计算分拆模型中的期望极限形状 [44, 5]：令 $A=B=\sqrt{n}$，则对于任意 $x>0$,

$$\begin{aligned}\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{1}{B}\sum_{j\ge Ax}\mathsf{E}(M_j) &= \frac{1}{\sqrt{n}}\sum_{j\ge\sqrt{n}\,x}\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} \\ &\to \frac{1}{\kappa}\int_{\kappa x}^{\infty}\frac{\mathrm{e}^{-u}}{1-\mathrm{e}^{u}}\,\mathrm{d}u = \frac{1}{\kappa}\sum_{\ell=1}^{\infty}\int_{\kappa x}^{\infty}\mathrm{e}^{-\ell u}\,\mathrm{d}u \\ &= \frac{1}{\kappa}\sum_{\ell=1}^{\infty}\frac{1}{\ell}\,\mathrm{e}^{-\ell x} = -\frac{1}{\kappa}\log\left(1-\mathrm{e}^{-\kappa x}\right).\end{aligned} \tag{2.29}$$

因此，极限形状 $y=\varphi(x)$ 由下式给出

$$y = -\kappa^{-1}\log\left(1-\mathrm{e}^{-\kappa x}\right) \qquad (x>0),$$

或者，用一种更对称的形式表示，

$$\mathrm{e}^{-\kappa x}+\mathrm{e}^{-\kappa y} = 1 \qquad (x,y>0), \tag{2.30}$$

其中 $\kappa=\pi/\sqrt{6}$ （见（2.25））。该函数的图像如图1(b)所示。

注意，$\varphi(0) = \infty$。根据（2.29）中的计算，这意味着 M 的期望值增长速度快于 $\sqrt{n}$。更准确地说，我们有

$$\mathsf{E}(M) = \sum_{j=1}^{\infty} \frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} = \sum_{j=1}^{m} \frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} + \frac{1}{\alpha_n} \sum_{j>m}^{\infty} \frac{\alpha_n \mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}}, \tag{2.31}$$

其中 $m = [1/\alpha_n]$ 和 $\alpha_n = -\kappa/\sqrt{n}$（见（2.25））。如前所述，我们可以看到，（2.31）中的最后一个求和收敛到积分 $\int_1^\infty \mathrm{e}^{-\kappa u}(1-\mathrm{e}^{-\kappa u})^{-1}\mathrm{d}u < \infty$。接下来，写作

$$\sum_{j=1}^{m} \frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} = \frac{1}{\alpha_n} \sum_{j=1}^{m} \frac{1}{j} + \sum_{j=1}^{m} \left(\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} - \frac{1}{\alpha_n j} \right),$$

其中 [27, 2.10.8]

$$\sum_{j=1}^{m} \frac{1}{j} \sim \log m \sim -\log \alpha_n$$

and

$$\sum_{j=1}^{m} \left(\frac{\mathrm{e}^{-\alpha_n j}}{1-\mathrm{e}^{-\alpha_n j}} - \frac{1}{\alpha_n j} \right) \sim \frac{1}{\alpha_n} \int_0^1 \left(\frac{\mathrm{e}^{-u}}{1-\mathrm{e}^{-u}} - \frac{1}{u} \right) \mathrm{d}u = O(\alpha_n^{-1}),$$

注意到被积函数在零点处有有限极限。因此，

$$\mathsf{E}(M) \sim \alpha_n^{-1}(-\log \alpha_n) \sim \frac{\sqrt{n}}{2\kappa} \log n = \frac{\sqrt{6n}}{2\pi} \log n \qquad (n \to \infty). \tag{2.32}$$

注记 2.5. 上文讨论的两种不同模型设置——具有独立输出 $X_i (i = 1, \dots, M)$，如物品生产模型（第 2.2 节）；或具有独立重数 $M_j (j \in \mathbb{N}_0)$，如整数分拆随机模型（第 2.3 节）——实际上彼此密切相关，而且在某种意义上是等价的。事实上，对组合结构中的某些参数进行随机化，是一种常用的技术手段 [1]，旨在克服结构约束，例如整数分拆中预先规定各部分之和 [16, 5]。再举一个与物品生产模型直接相关的例子：在占位问题中（见注记 2.2），通常采用所谓的泊松化[1, 7]，将原本（相互依赖的）重数 M_j 分别替换为均值为 Mf_j 的独立泊松随机变量（$j \in \mathbb{N}_0$） [6, 17]。在上述每一种设置中，都可以通过原始问题与随机化版本之间适当的“桥梁”来确保所预期的等价性，例如，整数分拆情况下关于概率 $\mathrm{P}\big(\sum_j jM_j = n\big)$ 渐近行为的局部极限定理 [16, 5]，，或占位问题中的“去泊松化引理” [17, 6]。

3. GIGP 模型

3.1. GIGP 分布

Sichel [36, 41] 引入的广义逆高斯-泊松（*GIGP*）分布形式如下：

$$f_j = \frac{(1-\theta)^{\nu/2}}{K_\nu\big(\alpha(1-\theta)^{1/2}\big)} \cdot \frac{\big(\frac{1}{2}\alpha\theta\big)^j}{j!} K_{\nu+j}(\alpha) \qquad (j \in \mathbb{N}_0), \tag{3.1}$$

其中参数的取值范围为 $\nu \in \mathbb{R}$、 $\alpha > 0$ 和 $0 < \theta < 1$，而 $K_\nu(\cdot)$ 是阶数为 ν [27, 的第二类修正贝塞尔函数§10.25(i), §10.25(ii)]。

如引言中所述，GIGP 模型 (3.1) 是一种混合泊松分布，

$$f_j = \int_0^\infty \frac{\lambda^j \mathrm{e}^{-\lambda}}{j!}\, g(\lambda)\, \mathrm{d}\lambda \qquad (j \geq 0), \tag{3.2}$$

其泊松参数 λ 的混合密度选为广义逆高斯（*GIG*）密度 [36] （另见 [22, 第 284]页）[4]

$$g(\lambda) = \frac{\left(2\,(1-\theta)^{1/2}/\alpha\theta\right)^\nu}{2\,K_\nu\!\left(\alpha\,(1-\theta)^{1/2}\right)}\, \lambda^{\nu-1} \exp\!\left(-\frac{(1-\theta)\,\lambda}{\theta} - \frac{\alpha^2\theta}{4\lambda}\right) \qquad (\lambda > 0). \tag{3.3}$$

(3.3) 中的归一化源自贝塞尔函数的一种积分表示 [27, 10.32.10]。表示式 (3.2) 说明了公式 (3.1) 为何定义了一个概率分布，

$$\sum_{j=0}^\infty f_j = \int_0^\infty \sum_{j=0}^\infty \frac{\lambda^j \mathrm{e}^{-\lambda}}{j!}\, g(\lambda)\, \mathrm{d}\lambda = \int_0^\infty g(\lambda)\, \mathrm{d}\lambda = 1,$$

它还导出了一个有趣的贝塞尔函数恒等式，而特殊函数文献中似乎尚未提及这一恒等式，

$$\sum_{j=0}^\infty \frac{\left(\frac{1}{2}\alpha\theta\right)^j K_{\nu+j}(\alpha)}{j!} = \frac{K_\nu\!\left(\alpha\,(1-\theta)^{1/2}\right)}{(1-\theta)^{\nu/2}}. \tag{3.4}$$

由公式 (3.2)，利用 GIG 密度 (3 3 θ α $\tilde{\theta} = \theta/\,1+\theta$ $\tilde{\alpha} = .$) 的归一化，将参数和分别 $\alpha\sqrt{1+\theta}$ 3 2 替换为 ( ) 和 ，即可轻易得到表达式 (3.1)。此外，公式 (.) 表明，GIGP 分布 (3.1) 的期望值与 GIG 分布 (3.3) 的期望值相同，

$$\begin{aligned}\eta = \sum_{j=0}^\infty j f_j = \int_0^\infty \sum_{j=0}^\infty j\, \frac{\lambda^j \mathrm{e}^{-\lambda}}{j!}\, g(\lambda)\, \mathrm{d}\lambda &= \int_0^\infty \lambda\, g(\lambda)\, \mathrm{d}\lambda \\ &= \frac{\alpha\theta}{2\,(1-\theta)^{1/2}} \cdot \frac{K_{\nu+1}\!\left(\alpha\,(1-\theta)^{1/2}\right)}{K_\nu\!\left(\alpha\,(1-\theta)^{1/2}\right)},\end{aligned} \tag{3.5}$$

wh这里最后一步计算是基于式 (3.3) 中阶数为 $\nu + 1$.[5] 的归一化

正如 Sichel [41, p. 315], 所指出的，频数 (3.1) 满足递推关系

$$f_{j+2} = \frac{(\nu + j + 1)\,\theta}{j+2}\, f_{j+1} + \frac{\alpha^2\theta^2}{4\,(j+2)\,(j+1)}\, f_j \qquad (j \in \mathbb{N}_0),$$

[4]我们沿用 [36] 的命名法。与 [22] 中另一种参数化（θ, ψ, χ）的联系由映射 $\theta \mapsto \nu$, $\psi \mapsto 2(1-\theta)/\theta$, $\chi \mapsto \alpha^2\theta/2$ 给出。[5]利用阶数为 $\nu + 1$ 的恒等式 (3.4)，可直接由定义 (3.1) 得到表达式 (3.5)。

该递推关系可通过对公式 (3.3) 后面提到的积分表示进行分部积分得到。

GIGP 分布 (3.1) 的尾部呈幂几何衰减，这可通过斯特林公式 [27, 5.11.3] 以及大阶贝塞尔函数的渐近行为 (A.7) 得出，从而得到

$$f_j \sim \frac{(1-\theta)^{\nu/2}\left(\frac{1}{2}\alpha\right)^{-\nu}}{2K_\nu\big(\alpha\,(1-\theta)^{1/2}\big)}\, j^{\nu-1}\theta^j \qquad (j\to\infty). \tag{3.6}$$

3.2. 边界情形 $\alpha = 0$

通过极限 $\alpha \to 0+$，还可以将值 $\alpha = 0$ 纳入 GIGP 类。为此，我们需要根据阶数 ν的取值考虑几种情形。具体而言，若 $\nu > 0$，则利用贝塞尔函数的小自变量渐近式（见 (A.2)），由 (3.1) 得到

$$f_j \sim (1-\theta)^\nu\, \frac{\Gamma(\nu+j)\,\theta^j}{\Gamma(\nu)\,j!} = \binom{\nu+j-1}{j}(1-\theta)^\nu\,\theta^j \qquad (j\in\mathbb{N}_0), \tag{3.7}$$

其中 $\Gamma(z) := \int_0^\infty s^{z-1}\mathrm{e}^{-s}\,\mathrm{d}s$ $(z>0)$ 是伽马函数 [27, 5.2.1]。公式 (3.7) 定义了一个负二项分布，其参数为 ν和 θ [21, Sec. 5.1],，期望值为

$$\eta = \frac{\nu\theta}{1-\theta}. \tag{3.8}$$

后一表达式与当 $\alpha \to 0+$ 时式 (3.5) 的极限一致（同样使用 (A.2)）。借助斯特林公式 [27, 5.11.3],，可得式 (3.7) 的尾部行为。

$$f_j \sim \frac{(1-\theta)^\nu\, j^{\nu-1}\theta^j}{\Gamma(\nu)} \qquad (j\to\infty), \tag{3.9}$$

这在形式上与当 $\alpha \to 0+$ 时式 (3.6) 的极限一致。

然而，当 $\nu \le 0$ 时，极限 GIGP 分布退化为 $f_0 = 1$，并且对所有 $j \ge 1$，$f_j = 0$。事实上，当 $\nu = 0$ 时，利用渐近公式 (A.4)，可得

$$f_0 = \frac{K_0(\alpha)}{K_0\big(\alpha\,(1-\theta)^{1/2}\big)} \sim \frac{-\log\alpha}{-\log\big(\alpha\,(1-\theta)^{1/2}\big)} \to 1.$$

当 $\nu < 0$ 时，借助渐近公式 (A.1) 和 (A.2)，我们有

$$f_0 = \frac{(1-\theta)^{\nu/2}\,K_\nu(\alpha)}{K_\nu\big(\alpha\,(1-\theta)^{1/2}\big)} \sim \frac{(1-\theta)^{\nu/2}\,\frac{1}{2}\Gamma(-\nu)\big(\frac{1}{2}\alpha\big)^\nu}{\frac{1}{2}\Gamma(-\nu)\big(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\big)^\nu} = 1.$$

为纠正这种退化，我们改用如下定义的零截断GIGP分布：

$$\mathsf{P}(X_i = j \,|\, X_i \ge 1) = \frac{f_j}{1-f_0} \qquad (j\in\mathbb{N})$$

并取 $\alpha \to 0+$ 时的极限。我们将所得的条件频数记为 $(\check{}f_j)$ $(j\in\mathbb{N})$，相应的期望值记为 $\check{}\eta$。我们将分析范围限制在 $-1 < \nu \le 0$，并分别讨论 $\nu = 0$ 和 $-1 < \nu < 0$ 的情形（关于为何值 $\nu = -1$ 与 $\alpha = 0$不相容，见下文注释3.2）。

注释3.1。不考虑 $\nu < -1$ 且 $\alpha > 0$ 的情形（见命题 3.2(e) 及该命题前的说明）。因此，我们没有必要在此考察极限 $\alpha \to 0$ 。

情形 $\nu = 0$

应用渐近公式 (A.4)，得到

$$1 - f_0 = \frac{K_0\big(\alpha\,(1-\theta)^{1/2}\big) - K_0(\alpha)}{K_0\big(\alpha\,(1-\theta)^{1/2}\big)} \sim \frac{\log(1-\theta)}{\log\alpha},$$

而 (A.2) 和 (A.5) 对于 $j \geq 1$ 给出

$$f_j = \frac{\big(\frac{1}{2}\alpha\theta\big)^j}{j!} \cdot \frac{K_j(\alpha)}{K_0\big(\alpha\,(1-\theta)^{1/2}\big)} \sim \frac{1}{-\log\alpha} \cdot \frac{\theta^j}{j},$$

利用 $\Gamma(j) = (j-1)!$，可得。于是，

$$\frac{f_j}{1-f_0} \sim \check{f}_j := \frac{1}{-\log(1-\theta)} \cdot \frac{\theta^j}{j} \qquad (j \in \mathbb{N}), \tag{3.10}$$

这就是 费希尔对数级数分布 [21, 第 7.1 节。2]。注意，(3.10) 的尾部行为自动呈幂几何形式，与 (3.9) 类似（其中 $\nu = 0$）。该分布的期望值很容易计算，

$$\check{\eta} = \frac{1}{-\log(1-\theta)} \sum_{j=1}^{\infty} \theta^j = \frac{\theta}{(1-\theta)\big(-\log(1-\theta)\big)}. \tag{3.11}$$

情形 $-1 < \nu < 0$

借助渐近公式 (A.6)，我们得到

$$\begin{aligned} 1 - f_0 &= \frac{K_\nu\big(\alpha\,(1-\theta)^{1/2}\big) - (1-\theta)^{\nu/2}\,K_\nu(\alpha)}{K_\nu\big(\alpha\,(1-\theta)^{1/2}\big)} \\ &\sim \frac{\Gamma(\nu+1)}{(-\nu)\,\Gamma(-\nu)} \left(\tfrac{1}{2}\alpha\right)^{-2\nu} \left(1 - (1-\theta)^{-\nu}\right), \end{aligned}$$

并且，对于 $j \geq 1$,

$$f_j \sim \frac{(1-\theta)^{\nu/2}\left(\frac{1}{2}\alpha\theta\right)^j}{j!} \cdot \frac{\frac{1}{2}\Gamma(\nu+j)\left(\frac{1}{2}\alpha\right)^{-\nu-j}}{\frac{1}{2}\Gamma(-\nu)\left(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\right)^{\nu}} \sim \frac{\Gamma(\nu+j)\,\theta^j}{\Gamma(-\nu)\,j!} \left(\tfrac{1}{2}\alpha\right)^{-2\nu}.$$

因此,

$$\frac{f_j}{1-f_0} \sim \check{f}_j := \frac{(-\nu)\,\Gamma(\nu+j)\,\theta^j}{\Gamma(\nu+1)\left(1-(1-\theta)^{-\nu}\right)j!} \qquad (j \in \mathbb{N}). \tag{3.12}$$

这是一个扩展负二项分布 [21, 第 5.12 节2], , 其期望值为

$$\check{\eta} = \frac{(-\nu)\,\theta\,(1-\theta)^{-\nu-1}}{1-(1-\theta)^{-\nu}}. \tag{3.13}$$

利用斯特林公式很容易得到该分布的尾部衰减 (

3 12 [27, .)。. 3],

5 11
$$\check{f}_j \sim \frac{(-\nu)\, j^{\nu-1}\theta^j}{\Gamma(\nu+1)\left(1-(1-\theta)^{-\nu}\right)} \qquad (j\to\infty). \tag{3.14}$$

备注 3.2. 若 $\nu=-1$，那么利用 (A.1)、(A.3) 和 (A.4)，可得

$$1-f_0=\frac{K_1\big(\alpha(1-\theta)^{1/2}\big)-(1-\theta)^{-1/2}K_1(\alpha)}{K_1\big(\alpha(1-\theta)^{1/2}\big)}\sim \tfrac{1}{2}\alpha^2\theta\,(-\log\alpha),$$

and

$$f_1=\frac{(1-\theta)^{-1/2}\left(\frac{1}{2}\alpha\theta\right)K_0(\alpha)}{K_1\big(\alpha(1-\theta)^{1/2}\big)}\sim \tfrac{1}{2}\alpha^2\theta\,(-\log\alpha)\,,$$

因此

$$\frac{f_1}{1-f_0}\sim \check{f}_1=1.$$

ˇ 因此，极限条件分布 (f_j) 似乎是退化的，所有概率质量都集中在 $j=1$。这不适用于建模，因此“角点”情形 $\nu=-1, \alpha=0$ 不予考虑。

3.3. GIGP均值的渐近行为

正如第 2.2 节中的整数分拆示例所示，要使极限形状有意义，杨图的面积必须比组成块的数量增长得更快（见 (2.32)）。在物品生产模型中，这意味着物品总数 $N=\sum_j jM_j$ 应远远大于来源数 $M=\sum_j M_j$。回顾 (2.3)，物品总数的期望为 E(N) $=M\eta$（其中 $\eta=$E(X_i) 是每个来源的物品数的期望，见 (2.1)），这意味着适当的极限状态由 $\eta\to\infty$ 确定。

反过来，从 GIGP均值 η 的表达式 (3.5) 可以推测，当 $\theta\approx 1$ 时，且参数 α 和 ν 保持不变时，便可达到后者。这一点可以利用贝塞尔函数 $K_\nu(z)$ 在 $z\to 0$ 时已知的渐近公式加以验证（参见下文命题 3.2）；下一条引理将这些公式调整为适用于我们的情形。

Le引理 3.1. 当 $\alpha>0$ 和 $\nu\in\mathbb{R}$ 固定时，随着 $\theta\to 1-$，以下渐近行为成立 ，

$$K_\nu\big(\alpha(1-\theta)^{1/2}\big)\sim\begin{cases}\frac{1}{2}\Gamma(\nu)\left(\frac{1}{2}\alpha\right)^{-\nu}(1-\theta)^{-\nu/2} & (\nu>0),\\ \frac{1}{2}\big(-\log(1-\theta)\big) & (\nu=0),\\ \frac{1}{2}\Gamma(-\nu)\left(\frac{1}{2}\alpha\right)^{\nu}(1-\theta)^{\nu/2} & (\nu<0).\end{cases} \tag{3.15}$$

证明。 渐近行为 (3.15) 的首项可直接由 $\nu\neq 0$ 的公式 (A.2)（借助 $\nu<0$ 的公式 (A.1)）以及 $\nu=0$ 的公式 (A.5) 得出， □

利用这个引理，我们可以更精确地刻画当 $1-\theta\to 0+$ 时 GIGP均值的渐近行为。特别是，这项分析表明，当 $\nu\geq -1$ 时，所需的趋于无穷大的增长确实成立；但当 $\nu<-1$ 时则不成立。

命题 3.2. *GIGP* $(\alpha > 0)$ 分布和零截断广义逆高斯泊松分布 $(\alpha = 0)$ 分布的期望值 η 和 $\breve{}$ η 分别具有以下渐近行为，当 $\theta \to 1-$ 时。

(a) $\nu > 0, \quad \alpha \geq 0$:

$$\eta \sim \frac{\nu}{1-\theta}. \tag{3.16}$$

(b) $\nu = 0, \quad \alpha \geq 0$:

$$\eta \sim \frac{1}{(1-\theta)\big(-\log(1-\theta)\big)}, \qquad \breve{\eta} \sim \frac{1}{(1-\theta)\big(-\log(1-\theta)\big)}. \tag{3.17}$$

(c) $-1 < \nu < 0, \quad \alpha \geq 0$:

$$\eta \sim \frac{\Gamma(\nu+1)\big(\frac{1}{2}\alpha\big)^{-2\nu}}{\Gamma(-\nu)\,(1-\theta)^{\nu+1}}, \qquad \breve{\eta} \sim \frac{-\nu}{(1-\theta)^{\nu+1}}. \tag{3.18}$$

(d) $\nu = -1, \quad \alpha > 0$:

$$\eta \sim \big(\tfrac{1}{2}\alpha\big)^2\big(-\log(1-\theta)\big). \tag{3.19}$$

(e) $\nu < -1, \quad \alpha > 0$:

$$\eta \sim \frac{\big(\frac{1}{2}\alpha\big)^2}{-\nu-1}. \tag{3.20}$$

证明。 利用引理3.1中的渐近公式考虑情形 (a)–(e)。

(a) 对于 $\alpha > 0$，利用 (3.15) 中阶数为 ν 和 $\nu+1$ 的第一行，我们有

$$\frac{K_{\nu+1}\big(\alpha(1-\theta)^{1/2}\big)}{K_\nu\big(\alpha(1-\theta)^{1/2}\big)} \sim \frac{\frac{1}{2}\Gamma(\nu+1)\big(\frac{1}{2}\alpha(1-\theta)^{1/2}\big)^{-\nu-1}}{\frac{1}{2}\Gamma(\nu)\big(\frac{1}{2}\alpha(1-\theta)^{1/2}\big)^{-\nu}} = \frac{\nu}{\frac{1}{2}\alpha(1-\theta)^{1/2}}, \tag{3.21}$$

这里还用到了伽马函数的递推性质，$\Gamma(\nu+1) = \nu\,\Gamma(\nu)$ [27, 5.5.1]。将其代入 (3.5) 即得 (3.16)。若 $\alpha = 0$，则由 (3.8) 可直接得到 (3.16)。

(b) 对于 $\alpha > 0$，将 $\nu = 0$ 和 $\nu = 1$ 代入公式 (3.15) 得到

$$\frac{K_1\big(\alpha(1-\theta)^{1/2}\big)}{K_0\big(\alpha(1-\theta)^{1/2}\big)} \sim \frac{\big(\frac{1}{2}\alpha(1-\theta)^{1/2}\big)^{-1}}{-\log(1-\theta)}, \tag{3.22}$$

第一个公式 (3.17) 可由 (3.5) 得到。若 $\alpha = 0$，则该公式立即给出 (3.17) (3.11)
中的第二个公式。

(c) 对于 $\alpha > 0$，利用对称关系 (A.1)，类似于 (3.21)，我们得到

$$\frac{K_{\nu+1}\big(\alpha(1-\theta)^{1/2}\big)}{K_\nu\big(\alpha(1-\theta)^{1/2}\big)} \sim \frac{\Gamma(\nu+1)\big(\frac{1}{2}\alpha(1-\theta)^{1/2}\big)^{-2\nu-1}}{\Gamma(-\nu)},$$

于是，由（3.5）可得（3.18）中的第一个公式。（3.18）中的第二个公式可直接由（3.13）得到。

(d) 对于 由（3.5）结合对称关系（A.1）和渐近比（3.22）可得 .

(e) 再次利用 (A.1) 和 (3.15) 的第一行，并取阶数 $-\nu > 0$ 和 $-\nu - 1 > 0$，可得

$$\frac{K_{\nu+1}\big(\alpha\,(1-\theta)^{1/2}\big)}{K_{\nu}\big(\alpha\,(1-\theta)^{1/2}\big)} \sim \frac{\Gamma(-\nu-1)\big(\frac{1}{2}\alpha\,(1-\theta)^{1/2}\big)}{\Gamma(-\nu)},$$

而 (3 20 3 5 $\Gamma\ z+1\ = z\Gamma\ z$ [27, .) 则由 (.) 得出，这里再次利用递推关系 ( ) ( ) . .1], , 此时取 .。 $z = -\nu - 1$ 5 5

至此，命题 3.2 的证明完成。 □

命题 3.2 描述了期望值 η（当 $\alpha > 0$ 时）或 $\check{}\eta$（当 $\alpha = 0$ 时）随小参数 $1-\theta$ 的变化。对于 GIGP 模型拟合，通过求解渐近方程 (3.16)、(3.17)、(3.18) 和 (3.19)，分别用 η 或 $\check{}\eta$ 表示 $1-\theta$ 很有用。

命题3.3. 在命题 *3.2* 的条件下，以下渐近行为成立。

(a) $\nu > 0, \quad \alpha \geq 0$:

$$1 - \theta \sim \frac{\nu}{\eta}.$$

(b) $\nu = 0,\ \alpha \geq 0$:

$$1 - \theta \sim \frac{1}{\eta \log \eta}, \qquad 1 - \theta \sim \frac{1}{\check{\eta} \log \check{\eta}}.$$

(c) $-1 < \nu < 0,\ \alpha \geq 0$:

$$1 - \theta \sim \left(\frac{\Gamma(\nu+1)\big(\frac{1}{2}\alpha\big)^{-2\nu}}{\Gamma(-\nu)\,\eta} \right)^{1/(\nu+1)}, \qquad 1 - \theta \sim \left(\frac{-\nu}{\check{\eta}} \right)^{1/(\nu+1)}.$$

(d) $\nu = -1, \quad \alpha > 0$:

$$\log(1-\theta) \sim -\frac{4\eta}{\alpha^2}.$$

注记 3.3. 公式 (3.19) 仅给出了 $1-\theta$ 的对数渐近形式，但这已足以满足估计的需要。

4. GIGP 模型中的极限形状

4.1. 缩放系数与主定理

设频数 f_j ($j \in \mathbb{N}_0$) 由 GIGP 分布公式 (3.1) 给出，其参数为 $0 < \theta < 1$、$\nu \geq -1$ 和 $\alpha \geq 0$，但不包括“角点”对 $\nu = -1$、$\alpha = 0$。情形 $\alpha = 0$ 可理解为条件概率 P($X_i = j \mid X_i > 0$) = $f_j/\ 1 - f_0 \quad j \in \mathbb{N} \quad \alpha \to 0+$ 3 2 ( ) ( ) 在 (见第 . 节) 时的极限。

给定由 M 个来源产生的观测到的重数随机向量 (M_j)，我们的目标是研究带有边界（见 (2.18)）的缩放杨图的渐近行为

$$\widetilde{Y}(x) := \frac{Y(Ax)}{B} = \frac{1}{B}\sum_{j\geq Ax} M_j = \frac{1}{B}\sum_{i=1}^{M} Z_i(Ax) \qquad (x \geq 0). \tag{4.1}$$

我们在以下极限状态假设下展开讨论，其中包括缩放系数 A 和 B的指定。

假设 4.1。来源数量很大，$M \to \infty$。此外，假设内禀参数 $\theta \in (0,1)$ 接近其上限 1，即 $\theta \to 1-$，这保证了每个来源的平均项目数很大（见命题 3.2）。

假设 4.2。x-缩放系数 A 选定为

$$A = \frac{1}{-\log\theta} \sim \frac{1}{1-\theta} \to \infty \qquad (\theta \to 1-), \tag{4.2}$$

而 y-缩放系数 B 则根据参数空间 ν 和 α 中的特定区域指定如下：

(a) $\nu > 0, \alpha \geq 0$:

$$B = \frac{M}{\Gamma(\nu)}. \tag{4.3}$$

(b) $\nu = 0, \alpha \geq 0$:

$$B = \frac{M}{-\log(1-\theta)}. \tag{4.4}$$

(c) $-1 \leq \nu < 0, \alpha > 0$:

$$B = \frac{M\left(\frac{1}{2}\alpha\right)^{-2\nu}(1-\theta)^{-\nu}}{\Gamma(-\nu)}. \tag{4.5}$$

(d) $-1 < \nu < 0, \alpha = 0$:

$$B = \frac{M(-\nu)(1-\theta)^{-\nu}}{\Gamma(\nu+1)}. \tag{4.6}$$

假设 4.3. 假设 4.2 中定义的 y-缩放系数 B 很大，$B \to \infty$。对于 $\nu > 0$，根据 (4.3)，这是自动成立的（只要 $M \to \infty$）；但对于 $\nu \leq 0$，我们还必须额外假设：如果 $\nu = 0$，则 $M \gg -\log(1-\theta)$；如果 $\nu < 0$，则 $M \gg (1-\theta)^{\nu}$。

注4.1。假设 4.3 中需要对外部参数 $M \to \infty$ 与内在 GIGP 参数$\theta \to 1-$ 的联合极限行为施加额外条件，这表明，为了使数据呈现出明显的极限形状，来源数量 M 必须足够大。我们将在下文第 6 节阐明相反的情形。

对于 $\nu \geq -1$，考虑函数

$$\varphi_\nu(x) := \int_x^\infty s^{\nu-1}\,\mathrm{e}^{-s}\,\mathrm{d}s \qquad (x>0), \tag{4.7}$$

即 （上）不完全伽马函数 [27, 8.2.2]。以下是我们的主要结果，它证明缩放杨图（见 (4.1) ）依概率收敛到极限形状 $\varphi_\nu(x)$。

定理 4.1。 在假设 *4.1*、*4.2* 和 *4.3* 下，对于任意 $\varepsilon>0$ 和任意 $\delta>0$ ，我们有

$$\mathsf{P}\left(\sup_{x\geq\delta}\left|\widetilde{Y}(x)-\varphi_\nu(x)\right|\geq\varepsilon\right)\to 0. \tag{4.8}$$

定理 4.1 的证明将在下文第 4.3 至 4.7 节中给出。

4.2. 使用计算机模拟进行图示说明

本节通过两个示例说明如何利用计算机模拟数据来近似极限形状，其中 $\nu = 0.5$ 和 $\nu = -0.5$（见图 2 左侧图）。其他参数设置如下： $\alpha = 2$。 $\theta = 0.99$，以及 $M = 1000$。图中将数据表示为式 (2.10) 中定义的杨图 $Y(x)$ 的上边界，并绘出理论 GIGP 互补分布函数 $\bar{F}(x)$（见 (2.14) 和 (3.1)），同时还展示了缩放回原始计数频数后的极限形状，即 $x \mapsto B\,\varphi_\nu(x/A)$，其中 $A = -1/\log\,\theta \doteq 99.49916$ （见 (4.2)），且当 $\nu = 0.5$ 时 $B \doteq 564.1896$，或当 $\nu = -0.5$ 时 $B \doteq 56.41896$ （分别见 (4.3) 和 (4.5)）。在这两种情况下，图像都显示，极限形状与观测值主体部分的拟合非常好。

通过观察式 (4.7)，可以更方便地考察尾部行为：

$$\begin{aligned}\varphi_\nu(x) &= -\int_x^\infty s^{\nu-1}\,\mathrm{d}(\mathrm{e}^{-s})\\ &= x^{\nu-1}\,\mathrm{e}^{-x} + (\nu-1)\int_x^\infty s^{\nu-2}\,\mathrm{e}^{-s}\,\mathrm{d}s \sim x^{\nu-1}\,\mathrm{e}^{-x} \qquad (x\to\infty).\end{aligned}$$

因此，根据式 (4.1) 和 (4.8)，可以预期，当 x 足够大时，

$$y = Y(Ax) \approx B\,\varphi_\nu(x) \approx B\,x^{\nu-1}\,\mathrm{e}^{-x},$$

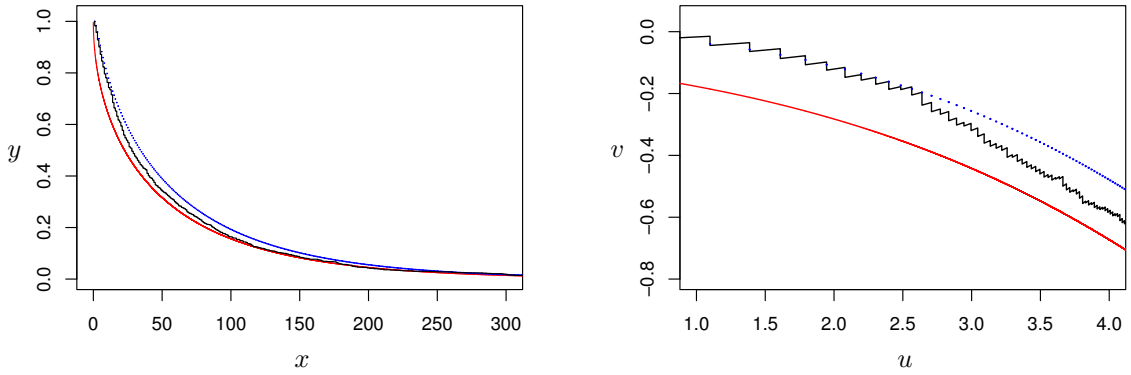
或者，取对数后，

$$\log Y(Ax) + x \approx \log B + (\nu-1)\log x. \tag{4.9}$$

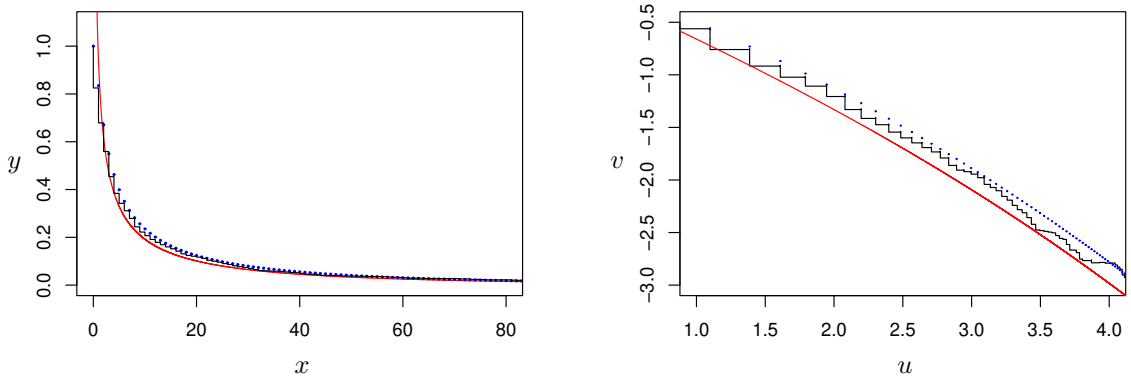
因此，从 $(x,\ y)$ 转换到新坐标

$$u = \log x, \qquad v = \log y + x, \tag{4.10}$$

变换后的数据图预计会接近斜率为 $\nu - 1$ 的直线，理论 GIGP 分布函数的尾部和极限形状的尾部也应如此



(a) $\nu = 0.5$



(b) $\nu = -0.5$

图2：使用通过 GIGP 模型 (3.1) 模拟的 $M = 1000$ 随机值（X_i），展示极限形状近似；模型参数为 $\theta = 0.99$、$\alpha = 2$，且分别对应 (a) $\nu = 0.5$ 或 (b) $\nu = -0.5$。左侧图中，黑色阶梯线表示相应杨图的上边界 $Y(x)$，蓝色点线表示 GIGP 互补分布函数 $\bar{F}$，平滑的红色曲线则表示缩放回原尺度后的极限形状 $x \mapsto B\,\varphi_\nu(x/A)$。右侧图中，尾部以变换坐标 (4.10) 表示，线型和颜色编码相同。

两者相似。图 2（右侧面板）展示了模拟数据的这一特征：在两种情况下，长尾都得到了较好的线性化，分别为 $\nu = 0.5$ 和 $\nu = -0.5$。

即使尚未估计模型参数，也可以使用上述图解法快速直观地检查 GIGP频数模型是否适用：首先尝试调整缩放系数 $A = -1/\log\,\theta$（见（4.2））， 以获得线性化的数据图（从而粗略估计参数 θ）；接着读出拟合斜率（用于估计参数 $\nu - 1$）；然后利用拟合截距（接近 $\log B$，见（4.9）），通过公式（4.3）至（4.6）之一估计参数 α。我们将在第7节中将此方法应用于一些真实数据集。

4.3. 期望杨图的收敛

我们首先证明定理 4.1，说明对于期望杨图，收敛到极限形状 $\varphi_\nu(x)$ 成立。由（2.16）和（2.18），可得

$$\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{M\bar{F}(Ax)}{B}. \tag{4.11}$$

定理 4.2。 在假设 *4.1* 和 *4.2* 下，

$$\mathsf{E}\big(\widetilde{Y}(x)\big) \to \varphi_\nu(x) \qquad (x>0), \tag{4.12}$$

在 $x \geq \delta$ 上一致地，对任意 $\delta > 0$。

备注 4.2。 请注意，定理 4.2 不需要假设 4.3。

单调函数一致收敛的以下实用判据（针对半直线域调整）是众所周知的（例如，参见 [34，第 0.1] 节）。

引理 4.3。 设定义在 $(0,\infty)$ 上的一列单调函数，在$[\delta,\infty)$ 上对任意 $\delta>0$ 一致有界，并逐点收敛到一个连续（单调）函数。那么，对任意 $\delta>0$，该收敛在 $[\delta,\infty)$ 上是一致的。

注意到 (4.7) 中的极限函数 $\varphi_\nu(x)$ 在 (0，∞) 上连续且单调递减，根据引理 4.3，只需证明对每个 $x>0$，逐点收敛 (4.12) 成立。备注 4.3。 在下面的计算中，我们仅考虑级数 $\bar{F}(Ax)$（见 (4.11)）各项的主导渐近项 (3.6)。使用经典的欧拉–麦克劳林求和公式 [27, § 2.10(i)] 以及大阶贝塞尔函数的一致渐近展开 [27, §10.41(ii)]，即可通过控制近似误差进行更仔细的分析。

定理 *4.2* 的证明。 下面的证明根据参数 ν 和 α 的不同子区域分步进行（见假设 4.2）。首先考虑 $\alpha>0$ 的情形，此时 GIGP 分布的支撑集为 $j\in\mathbb{N}_0$；随后转而考虑 $\alpha=0$ 的边界情形，此时支撑集缩减为 $j\in\mathbb{N}$。

- $\alpha>0$

  利用频数的渐近近似（ 3 6 f_j $j \geq Ax \geq .$），由（.）可得 $A\delta \gg 1$ 4 11

  $$\mathsf{E}\big(\widetilde{Y}(x)\big) = \frac{M}{B}\sum_{j\geq Ax} f_j \sim \frac{M\,(1-\theta)^{\nu/2}\left(\frac{1}{2}\alpha\right)^{-\nu}}{2BK_\nu\big(\alpha\,(1-\theta)^{1/2}\big)} \sum_{j\geq Ax} j^{\nu-1}\theta^j. \tag{4.13}$$

  回想一下， $A=(-\log\,\theta)^{-1}\sim(1-\theta)^{-1}$，对于（4.13）中的最后一个求和，我们有

  $$A^{-\nu}\sum_{j\geq Ax} j^{\nu-1}\theta^j = \sum_{j\geq Ax}\left(\frac{j}{A}\right)^{\nu-1}\mathrm{e}^{-j/A}\,\frac{1}{A} \to \int_x^\infty s^{\nu-1}\,\mathrm{e}^{-s}\,\mathrm{d}s = \varphi_\nu(x), \tag{4.14}$$

  显然，将 (4.14) 视为黎曼积分和，它收敛于右侧的积分即可看出这一点。此外，分母在 (4.13) 中的渐近行为可由公式 (3.15) 得出（见引理3.1）。因此，回到 (4.13)，并回顾 A 在 (4.2) 中以及 B 在 (4.3)、(4.4)、(4.5) 中的定义，我们便可轻松得到 (4.12)。

- $\alpha = 0$

  利用尾部近似 (3.9) ($\nu > 0$)、(3.10) ($\nu = 0$) 和 (3.14) ($-1 < \nu < 0$)，由 (4.11) 可得，与 (4.13) 和 (4.14) 类似，

$$\mathsf{E}\big(\widetilde{Y}(x)\big) \sim \frac{MC_\nu(\theta)}{B} \sum_{j \geq Ax} j^{\nu-1}\theta^j \sim \frac{MC_\nu(\theta)A^\nu}{B}\,\varphi_\nu(x), \tag{4.15}$$

  其中 $A \sim (1-\theta)^{-1}$（见 (4.2)），并且

$$C_\nu(\theta) := \begin{cases} (1-\theta)^\nu/\Gamma(\nu) & (\nu > 0), \\ (-\log(1-\theta))^{-1} & (\nu = 0), \\ (-\nu)/\Gamma(\nu+1) & (-1 < \nu < 0). \end{cases}$$

  现在，利用条件 (4.3)、(4.4) 或 (4.6)，不难看出 (4.15) 的右端可化简为 $\varphi_\nu(x)$。

至此，定理 4.2 的证明完成。 □

4.4. 随机杨图的逐点收敛

在讨论更强的定理 4.1（该定理断言依概率一致收敛）之前，我们先给出一个较简单的结论，讨论任意 $x > 0$ 下 $\widetilde{Y}(x)$ 的逐点收敛。

定理 4.4. 在假设 *4.1*、*4.2* 和 *4.3* 下，$Y(x)$与极限形状 $\varphi_\nu(x)$ 之间的均方偏差渐近趋于零，

$$\mathsf{E}\big(\big|\widetilde{Y}(x) - \varphi_\nu(x)\big|^2\big) \to 0, \tag{4.16}$$

对任意 $\delta > 0$，该结果关于 $x \geq \delta$ 一致成立。这意味着依概率收敛，$\widetilde{Y}(x) \xrightarrow{\mathrm{p}} \varphi_\nu(x)$，也就是说，对任意 $\varepsilon > 0$，

$$\sup_{x \geq \delta} \mathsf{P}\big(\big|\widetilde{Y}(x) - \varphi_\nu(x)\big| \geq \varepsilon\big) \to 0. \tag{4.17}$$

证明。 根据均方偏差的标准分解，我们有

$$\mathsf{E}\big(\big|\widetilde{Y}(x) - \varphi_\nu(x)\big|^2\big) = \mathsf{Var}\big(\widetilde{Y}(x)\big) + \big(\mathsf{E}\big(\widetilde{Y}(x)\big) - \varphi_\nu(x)\big)^2. \tag{4.18}$$

利用公式 (2.16) 和 (2.18)，(4.18) 中的方差项可估计如下

$$\begin{aligned} \mathsf{Var}\big(\widetilde{Y}(x)\big) &= \frac{M\bar{F}(Ax)F(Ax)}{B^2} \\ &\leq \frac{M\bar{F}(Ax)}{B^2} \sim \frac{\varphi_\nu(x)}{B} \to 0, \end{aligned} \tag{4.19}$$

根据 (4.11)、(4.12)，以及假设 4.3（它保证 $B \to \infty$）。由定理 4.2，(4.19) 中的收敛在 $[\delta, \infty)$ 上一致，对每个 $\delta > 0$ 均成立。至于 (4.18) 右端的第二项，根据定理 4.2，它在每个区间 $[\delta, \infty)$ 上都渐近趋于零，且一致成立。因此，极限 (4.16) 得证。

最后，依概率收敛（4.17）是由切比雪夫不等式 [35, Sec. II.6, p. 192] 得出的（4.16）的标准推论。 □

4.5. 辅助鞅

回顾定义（2.12），并将 $Z_i(x)$ 解释为一个"时间"$x \in [0,\infty)$ 的随机过程，考虑一个反向时间下的重标度过程 $t = 1/x$,

$$\widetilde{Z}_i(t) := \frac{1 - Z_i(1/t)}{F(1/t)} = \frac{I_{\{X_i<1/t\}}}{F(1/t)} = \begin{cases} 1/F(1/t), & 0 < t < 1/X_i, \\ 0, & 1/X_i \le t < \infty. \end{cases} \tag{4.20}$$

注意，以概率论意义下的概率 1,

$$\lim_{t\to 0+} \widetilde{Z}_i(t) = \frac{I_{\{X_i<\infty\}}}{F(\infty)} = 1,$$

因此，通过令 $\widetilde{Z}_i(0) := 1$，可将 $\widetilde{Z}_i(t)$ 连续延拓到原点。显然,

$$\mathsf{E}\big(\widetilde{Z}_i(t)\big) = \frac{\mathsf{P}(X_i < 1/t)}{F(1/t)} = 1, \tag{4.21}$$

$$\mathsf{Var}\big(\widetilde{Z}_i(t)\big) = \frac{F(1/t)\big(1 - F(1/t)\big)}{F(1/t)^2} = \frac{\bar{F}(1/t)}{F(1/t)}, \tag{4.22}$$

利用 $\bar{F}(1/t) = 1 - F(1/t)$。

令 $\mathcal{F}_i(t) = \sigma\{\widetilde{Z}_i(s) : 0 \le s \le t\}$ $(t \ge 0)$ 表示包含所有事件 $\{Z_i(1/s) = 1\} = \{X_i \ge 1/s\}$ 且满足 $0 \le s \le t$ 的最小σ-代数。考虑乘积σ代数 $\mathcal{F}(t) = \mathcal{F}_1(t) \otimes \cdots \otimes \mathcal{F}_M(t)$，并定义随机过程

$$W(t) := \widetilde{Z}_1(t) + \cdots + \widetilde{Z}_M(t) - M \qquad (t \ge 0). \tag{4.23}$$

由 (4.20) 可知，过程 W 是 *c`adl`ag*（即其样本路径处处右连续且存在左极限）。此外，我们有（见 (4.21) 和 (4.22)）

$$\mathsf{E}\big(W(t)\big) = 0, \qquad \mathsf{Var}\big(W(t)\big) = \frac{M\bar{F}(1/t)}{F(1/t)}, \tag{4.24}$$

因为 the 和式 (4.23) 中的随机变量 $\widetilde{Z}_1(t) \ldots, \widetilde{Z}_M(t)$ 相互独立。

引理 4.5. 过程 $W(t)$ 关于滤过 $\mathcal{F}(t)\,(t \ge 0)$是一个鞅，也就是说，对于任意 $t \ge s \ge 0$，我们有 $\mathsf{E}\big(W(t)\,|\,\mathcal{F}(s)\big) = W(s)$，其成立的概率为1,

证明. 令 $t \ge s \ge 0$。根据定义 (4.23) 及 (X_i) 的独立性，我们有

$$\mathsf{E}\big(W(t)\,|\,\mathcal{F}(s)\big) = \sum_{i=1}^{M} \mathsf{E}\big(\widetilde{Z}_i(t)\,|\,\mathcal{F}_i(s)\big) - M, \tag{4.25}$$

因此，只需证明

$$\mathsf{E}\big(\widetilde{Z}_i(t)\,|\,\mathcal{F}_i(s)\big) = \widetilde{Z}_i(s). \tag{4.26}$$

根据定义，σ-代数 $\mathcal{F}_i(s)$ 中包含的信息与如下形式的阈值事件有关：对于所有 $0 \le r \le s$，$\{X_i \ge 1/r\}$

鉴于式 (4.26) 中的目标随机变量通过 $Z_i(1/t) = I_{\{X_i \geq 1/t\}}$ 表示，这实质上相当于了解 $X_i \geq 1/s$ 是否成立。

首先假设 $X_i < 1/s$，即 $\widetilde{Z}_i(s) = 1/F(1/s)$（见 (4.20)）。于是

$$\begin{aligned}\mathsf{E}\big(\widetilde{Z}_i(t) \,|\, X_i < 1/s\big) &= \frac{\mathsf{P}(X_i < 1/t \,|\, X_i < 1/s)}{F(1/t)} \\ &= \frac{\mathsf{P}(X_i < 1/t)}{F(1/t)\,\mathsf{P}(X_i < 1/s)} \\ &= \frac{1}{F(1/s)} = \widetilde{Z}_i(s).\end{aligned}$$

~类似地，如果 $X_i \geq 1/s \geq 1/t$，那么 $Z_i(s) = 0$，并且

$$\begin{aligned}\mathsf{E}\big(\widetilde{Z}_i(t) \,|\, X_i \geq 1/s\big) &= \frac{\mathsf{P}(X_i < 1/t \,|\, X_i \geq 1/s)}{F(1/t)} \\ &= \frac{\mathsf{P}(X_i < 1/t, X_i \geq 1/s)}{F(1/t)\,\mathsf{P}(X_i \geq 1/s)} \\ &= 0 = \widetilde{Z}_i(s).\end{aligned}$$

因此，关系 (4.26) 成立。结合 (4.25) 以及定义 (4.23)，引理 4.5 的证明即告完成。 □

鞅 $W(t)$ 将在下一节中用于估计函数 $\widetilde{Y}(x)$ 与 $\varphi_\nu(x)$ 之间的一致距离。

4.6. 随机杨图的一致收敛

定理 4.4 中建立的逐点收敛可以加强为远离原点时的一致收敛，从而得到上文作为定理 4.1 陈述的主要结果。

定理 *4.1* 的证明。注意

$$\sup_{x \geq \delta} \big|\widetilde{Y}(x) - \varphi_\nu(x)\big| \leq \sup_{x \geq \delta} \big|\widetilde{Y}(x) - \mathsf{E}\big(\widetilde{Y}(x)\big)\big| + \sup_{x \geq \delta} \big|\mathsf{E}\big(\widetilde{Y}(x) - \varphi_\mu(x)\big)\big|,$$

其中 $\mathsf{E}\big(\widetilde{Y}(x)\big) = M\bar{F}(Ax)/B$（见 (4.11)）。因此，根据定理 4.2，只需考虑这些偏差

$$\sup_{x \geq \delta} \left| \widetilde{Y}(x) - \frac{M\bar{F}(Ax)}{B} \right| \geq \varepsilon.$$

利用定义 (4.20)、(4.23) 以及关系 $\bar{F}(x) = 1 - F(x)$，我们得到

$$\begin{aligned}\widetilde{Y}(x) - \frac{M\bar{F}(Ax)}{B} &= \frac{1}{B}\left(\sum_{i=1}^{M} Z_i(Ax) - M\bar{F}(Ax)\right)\\ &= -\frac{1}{B}\left(\sum_{i=1}^{M}\bigl(1 - Z_i(Ax)\bigr) - MF(Ax)\right)\\ &= -\frac{F(Ax)}{B}\left(\sum_{i=1}^{M}\widetilde{Z}_i(1/Ax) - M\right)\\ &= -\frac{F(Ax)}{B}\,W(1/Ax).\end{aligned}$$

由于 $0 \le F(Ax) \le 1$ 且 $t := 1/Ax \in [0, 1/A\delta]$，这意味着

$$\sup_{x\ge\delta}\left|\widetilde{Y}(x) - \frac{M\bar{F}(Ax)}{B}\right| \le \frac{1}{B}\sup_{t\le 1/A\delta}\bigl|W(t)\bigr|.$$

因此，将 Doob–Kolmogorov 次鞅不等式（参见例如 [48, 定理 6.16，第 101] 页）应用于鞅 $W(t)$（参见引理 4.5），并利用公式 (4.24)，可得

$$\begin{aligned}\mathsf{P}\left(\sup_{x\ge\delta}\left|\widetilde{Y}(x) - \frac{M\bar{F}(Ax)}{B}\right| \ge \varepsilon\right) &\le \mathsf{P}\left(\sup_{t\le 1/A\delta}\bigl|W(t)\bigr| \ge B\varepsilon\right)\\ &\le \frac{\mathsf{Var}\bigl(W(1/A\delta)\bigr)}{B^2\varepsilon^2}\\ &= \frac{M\bar{F}(A\delta)}{F(A\delta)\,B^2\varepsilon^2} \sim \frac{\varphi_\nu(\delta)}{B\,\varepsilon^2} \to 0, \qquad (4.27)\end{aligned}$$

其中，最后一步用到了 $F(A\delta) \to 1$（因为 $A \to \infty$），以及极限 (4.19)（其中 $x = \delta$）。由此完成定理 4.1 的证明。 □

4.7. 杨图边界作为经验过程

随机函数 $Y(x)$ 由 (2.13) 给出，可视为一个经验过程 [8, 第 11], 章，它由一族测试函数 $(g_x,\, x \ge 0)$ 作用于独立随机样本 $(X_1, \dots, X_M)$ 而确定，定义如下：

$$g_x(j) := \mathbf{1}_{[x,\infty)}(j) \qquad (j \in \mathbb{N}_0). \qquad (4.28)$$

也就是说，利用 (2.12) 和 (2.13)，我们可以写成

$$Y(x) = \sum_{i=1}^{M} Z_i(x) = \sum_{i=1}^{M} \mathbf{1}_{[x,\infty)}(X_i) = \sum_{i=1}^{M} g_x(X_i).$$

这种方法的优点在于，经验过程上确界的方差具有精确的上界。

请注意,

$$\sup_{x\ge\delta}\left|\widetilde{Y}(x)-\frac{M\bar F(Ax)}{B}\right| \le \frac{1}{B}\sup_{x\ge\delta}\sum_{i=1}^{M}\bigl|g_{Ax}(X_i)-\bar F(Ax)\bigr| = \frac{1}{B}\sup_{x\ge\delta}\sum_{i=1}^{M}\bigl|\tilde g_x(X_i)\bigr|, \tag{4.29}$$

其中

$$\tilde g_x(j) := g_{Ax}(j)-\bar F(Ax) \qquad (j\in\mathbb{N}_0). \tag{4.30}$$

根据 [8, 定理 11 1，第 314, 316], 页

$$\mathsf{Var}\left(\sup_{x\ge\delta}\sum_{i=1}^{M}\bigl|\tilde g_x(X_i)\bigr|\right) \le \sum_{i=1}^{M}\mathsf{E}\left(\sup_{x\ge\delta}\bigl|\tilde g_x(X_i)\bigr|^2\right). \tag{4.31}$$

利用 (4.28) 和 (4.30)，我们得到

$$\begin{aligned}\bigl|\tilde g_x(X_i)\bigr|^2 &\le \bigl(Z_i(Ax)+\bar F(Ax)\bigr)^2\\ &= Z_i(Ax)^2+2\,Z_i(Ax)\bar F(Ax)+\bar F(Ax)^2\\ &\le Z_i(Ax)+2\,Z_i(Ax)+\bar F(Ax),\end{aligned}$$

因此

$$\sup_{x\ge\delta}\bigl|g_{Ax}(X_i)\bigr|^2 = 3\,Z_i(A\delta)+\bar F(A\delta)$$

并且（见 (4.31)）

$$\mathsf{E}\left(\sup_{x\ge\delta}\bigl|g_{Ax}(X_i)\bigr|^2\right) \le 3\,\mathsf{E}\bigl(Z_i(A\delta)\bigr)+\bar F(A\delta) = 4\,\bar F(A\delta).$$

回到 (4.29)，结合切比雪夫不等式可得,

$$\begin{aligned}\mathsf{P}\left(\sup_{x\ge\delta}\left|\widetilde{Y}(x)-\frac{M\bar F(Ax)}{B}\right|\ge\varepsilon\right) &\le \mathsf{P}\left(\sup_{x\ge\delta}\sum_{i=1}^{M}\bigl|\tilde g_x(X_i)\bigr|\ge B\varepsilon\right)\\ &\le \frac{4M\bar F(A\delta)}{B^2\varepsilon^2}\sim\frac{4\varphi_\nu(\delta)}{B\,\varepsilon^2}\to 0,\end{aligned}$$

使用 th 在 $B\to\infty$。因此，这为定理 4.1 提供了另一种证明（参见（4.27））。

5. 随机杨图的涨落

~回想一下， $Y(x)$ 是独立指示变量 $Z_i(Ax) = I_{\{X_i\ge Ax\}}, i = 1, \ldots\ldots, M$ 的（归一化）和（见（4.1）），因此自然可以预期 $\widetilde{Y}(x)$ 渐近正态，其均值为 $\mathsf{E}\bigl(\widetilde{Y}(x)\bigr) = M\bar F(Ax)/B \sim \varphi_\nu(x)$，方差为 $M\bar F(Ax)F(Ax)/B^2 \sim \varphi_\nu(x)/B$（见（4.11）和（4.12））。然而，标准的中心极限定理不能直接适用¯，因为“成功”概率 $\mathsf{P}(Z_i(Ax) = 1) = F(Ax)$ 并非常数（而且它趋于 0），所以我们必须使用特征函数法重新证明这一结论。

定理 5.1。 在假设 *4.1*、*4.2* 和 *4.3* 下，对于任意 $x>0$,

$$\Upsilon(x):=\sqrt{\frac{B}{\varphi_\nu(x)}}\left(\widetilde{Y}(x)-\frac{M\bar{F}(Ax)}{B}\right)\xrightarrow{\mathrm{d}}\mathcal{N}(0,1), \tag{5.1}$$

其中 $\mathcal{N}(0,1)$ 为标准正态分布（即均值为零、方差为单位方差），而 $\xrightarrow{\mathrm{d}}$ 表示依分布收敛。

证明。 代入 (4.1)，可将 (5.1) 的左侧改写为

$$\Upsilon(x)=\frac{1}{\sqrt{B\,\varphi_\nu(x)}}\sum_{i=1}^{M}\big(Z_i(Ax)-\bar{F}(Ax)\big). \tag{5.2}$$

(5.2) 的特征函数为

$$\psi(t;x):=\mathsf{E}\big(\mathrm{e}^{\mathrm{i}t\Upsilon(x)}\big)=\mathrm{e}^{-\mathrm{i}\tilde{t}M\bar{F}(Ax)}\Big(1+\bar{F}(Ax)\big(\mathrm{e}^{\mathrm{i}\tilde{t}}-1\big)\Big)^{M}, \tag{5.3}$$

其中

$$\tilde{t}=\frac{t}{\sqrt{B\,\varphi_\nu(x)}},\qquad t\in\mathbb{R}. \tag{5.4}$$

取对数函数 $\mathbb{C}\setminus\{0\}\ni z\mapsto\log\, z\in\mathbb{C}$ 的主值分支（即满足 $\log\,1=0$），我们可将 (5.3) 改写为

$$\log\psi(t;x)=-\mathrm{i}\tilde{t}M\bar{F}(Ax)+M\log(1+w), \tag{5.5}$$

其中

$$w:=\bar{F}(Ax)\big(\mathrm{e}^{\mathrm{i}\tilde{t}}-1\big). \tag{5.6}$$

由于 $A\to\infty$ 和 $B\to\infty$ （根据假设 4.2 和 4.3），我们有 $\tilde{t}\to0$ 和 $w\to0$，因此

$$\log(1+w)=w-\tfrac{1}{2}w^2+O(|w|^3).$$

因此，对 $\mathrm{e}^{\mathrm{i}\tilde{t}}=1+\mathrm{i}\tilde{t}-\frac{1}{2}\tilde{t}^2+O(\tilde{t}^3)$ 进行泰勒展开，并代入 (5.4) 和 (5.6)，公式 (5.5) 可展开如下，

$$\log\psi(t;x)=-\frac{M\bar{F}(Ax)F(Ax)\,t^2}{2B\,\varphi_\nu(x)}+O\left(\frac{M\bar{F}(Ax)}{B^{3/2}}\right)\to-\frac{t^2}{2},$$

利用 $M\bar{F}(Ax)/B\sim\varphi_\nu(x)$ 和 $F(Ax)\to1$ 对任意 $x>0$ 均成立。因此，$\psi(t;x)\to\mathrm{e}^{-t^2/2}$，这正是正态分布 $\mathcal{N}(0,1)$ 的特征函数，如所述。 □

类似地，定理 5.1 可以推广到有限维收敛。

定理 5.2. 在假设 *4.1*、*4.2* 和 *4.3* 下，式 *(5.1)* 和 *(5.2)* 中定义的随机过程 $(\Upsilon(x)\ x>0)$ 按照有限维分布收敛的意义，收敛于均值为零、协方差函数为下式的高斯随机过程 $(\Xi(x)\ x>0)$

$$K(x,x'):=\mathsf{Cov}\big(\Xi(x),\Xi(x')\big)=\sqrt{\frac{\varphi_\nu(x')}{\varphi_\nu(x)}}\qquad(0<x\le x').$$

证明. 证明过程与定理 5.1 类似，通过对任意有限数组 $0 < x_1 \le \cdots \le x_m$ ($m \in \mathbb{N}$) 的多元特征函数进行渐近分析，

$$\psi(t_1, \dots, t_m; x_1, \dots, x_m) = \mathsf{E}\left(\mathrm{i} \sum_{k=1}^{m} t_k \Upsilon(x_k) \right) \qquad (t_1, \dots, t_m \in \mathbb{R}).$$

极限协方差函数很容易计算：利用 (5.2) 和 (2.17)，对于$0 < x \le x'$，我们有：

$$\mathsf{Cov}\big(\Upsilon(x), \Upsilon(x')\big) = \frac{M \bar{F}(Ax') F(Ax)}{B \sqrt{\varphi_\nu(x)\, \varphi_\nu(x')}} \to \sqrt{\frac{\varphi_\nu(x')}{\varphi_\nu(x)}},$$

正如所要求的。 □

我们通过指出，极限高斯过程 $\Xi(x)$ 可以在分布意义上表示为标准布朗运动 (B_t, $t \ge 0$)（即均值为零，协方差函数 $\mathsf{Cov}(B_t,\ B_{t'}) = \min\{t, t'\}$），以此结束本节。

定理 5.3。 以下分布表示成立，

$$\Xi(x) \overset{\mathrm{d}}{=} \frac{B_{\varphi_\nu(x)}}{\sqrt{\varphi_\nu(x)}} \qquad (x > 0). \tag{5.7}$$

特别地，反向时间下的重标度过程 $t = 1/x$，定义为

$$\widetilde{\Xi}(t) := \sqrt{\varphi_\nu(1/t)}\, \Xi(1/t) \qquad (t \ge 0), \tag{5.8}$$

具有独立增量。

证明。 只需注意到，等式 (5.7) 右侧过程的协方差函数为（对任意 $0 < x \le x'$）

$$\frac{1}{\sqrt{\varphi_\nu(x)\, \varphi_\nu(x')}}\, \mathsf{Cov}\big(B_{\varphi_\nu(x)}, B_{\varphi_\nu(x')}\big) = \frac{1}{\sqrt{\varphi_\nu(x)\, \varphi_\nu(x')}}\, \varphi_\nu(x') = K(x, x'),$$

利用 $\varphi_\nu(x') \le \varphi_\nu(x)$。最后，过程 (5.8) 的增量独立性直接源于布朗运动 B_t 的同一性质。 □

6. “混沌”机制下的泊松近似

本节考虑假设 4.3 不成立的情形，此时 y-缩放系数 B 有界（这只可能发生在 $\nu \le 0$ 的情况下，见公式 (4.3) 至 (4.6)）。我们称这种情形为混沌，因为随机变量 $\widetilde{Y}(x)$ 不会收敛到极限形状 $\varphi_\nu(x)$（参见定理 4.4），尽管期望值 $\mathsf{E}\big(\widetilde{Y}(x)\big) = M\bar{F}(Ax)/B \to \varphi_\nu(x)$ 是收敛的（定理 4.2）。导致这一失败的根本原因是：尽管 $Y(x)$ 是独立伯努利变量 $Z_i(Ax)$ 的归一化和 $= I_{\{X_i \ge Ax\}}$（见 (4.1)），但成功概率 $\mathsf{P}(X_i \ge Ax) = \bar{F}(Ax)$ 趋于零，而项数 M 的增长速度不足以抵消这一点（见注 4.1）。

6.1. 一维分布

我们首先研究 $Y(x)$ 的一维分布，即在给定点 $x>0$ 处的分布。¯为便于理解，考虑一种典型情形：若 $B=1$，则 $MF(Ax)\to\varphi_\nu(x)$；根据经典的泊松分布“小数定律”，[46], 二项分布的和 $Y(x)=Z_1(Ax)+\cdots+Z_M(Ax)$ 渐近接近参数为 $\lambda=\varphi_\nu(x)$ 的泊松分布。也就是说，和 $\widetilde{Y}(x)$ 不会趋于某个确定常数（如大数定律所述），而是由于持续存在的“微小”随机性，在不作任何归一化的情况下具有非退化的（泊松）近似。下面将这一观察推广如下。

定理6.1。 假设假设 *4.1* 和 *4.2* 成立，但假设 *4.3* 不成立，因此 $B=O(1)$。则随机变量 $Y(Ax)$ 在 $x>0$¯下的分布可由参数为 $MF(Ax)\sim B\varphi_\nu(x)$ 的泊松分布近似，并且全变差距离（或柯尔莫哥洛夫一致距离）中的相应误差有界于 $O(M^{-1})=o(1)$。

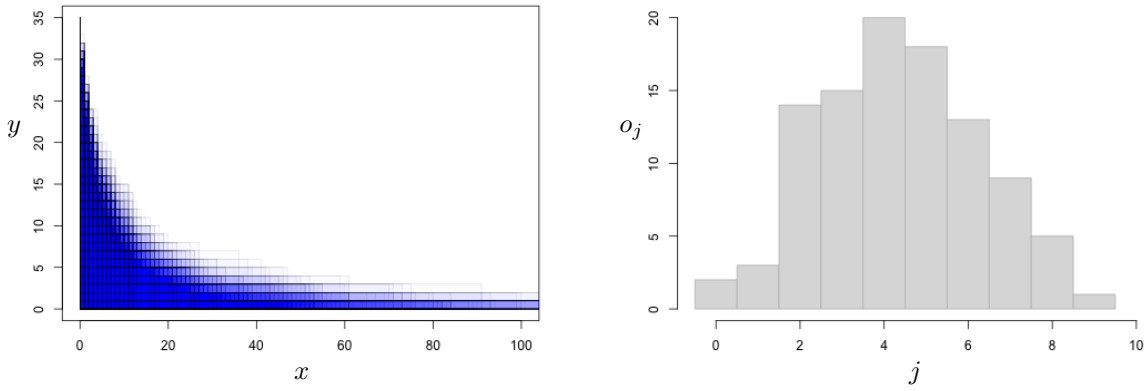
证明。 这是一个众所周知的近似结果的直接推论：独立伯努利试验中成功的总次数服从二项分布，其成功概率为 p，并可由参数为 $\lambda=np$ 的泊松分布近似，误差有界于 $\sigma^2=np^2$（参见例如 [3, 28]）。在本例中，$\lambda=M\bar{F}(Ax)$ 和 $\sigma^2=M\bar{F}(Ax)^2\sim\big(B\varphi_\nu(x)\big)^2/M=O(1/M)=o(1)$。 □

定理6.1中所述的泊松近似通过图 3展示，其中使用了从参数为 $\nu=-0.5$、$\alpha=2$ 和 $\theta=0.99$ 的 GIGP 分布 (3.1) 中抽取的 100 个模拟样本（每个样本大小为 $M=35$）。根据 (4.5) 计算得到的 y 缩放系数为 $B\doteq 1.974664$，这证实该情形处于混沌状态（即不满足假设 4.3）。x 缩放系数 (4.2) 在此情形下为 $A\doteq 99.49916$。图 3 左侧面板展示了叠加在一起的样本杨图，并以蓝色透明阴影呈现，因此颜色较深处表示出现频率较高。正如预期，不存在收敛到确定性极限形状的情形，但蓝色杨图逐渐显现出的“典型”边界清楚地表明，极限中会形成一条预期曲线。

在图 3 的右侧面板中，我们选取一个试验值 $x_0=0.2$，并绘制随机值 $Y(Ax_0)$ 的观测频数直方图，其中 $Ax_0\doteq 19.89983$。目测结果表明，该直方图与均值为 $M\bar{F}(Ax_0)\doteq 4.342498$ 的泊松分布相当吻合。Pearson χ^2 检验也证实了这一点：组距分别标记为从 0 到 9 的 j 值，相应的观测频数为 o_j。由于期望频数 $e_0\doteq 1.3004$ 和 $e_9\doteq 3.340067$ 小于 5，我们遵循一项常见建议，分别将组距 $j=0$ 和 $j=9$ 与 $j=1$ 和 $j=8$ 合并。计算得到分组后的 χ^2 统计量为 1.972246，自由度为 $10-2-1=7$，p 值为 96.14%，因此拟合优度检验轻松通过。

6.2. 有限维分布

定理6.1的结果可以推广到有限维分布。令 $Y(\infty)=0$ 较为方便，因为以概率 1，$Y(x)\to 0$，当 $x\to\infty$ 时。



(a) (b)

图3：混沌状态下泊松统计的示意图。左图展示了从 GIGP 模型 (3.1) 中生成的 100 个随机样本（每个样本大小为 $M = 35$ ）对应的杨图叠加结果，参数为 $\nu = -0.5$、$\alpha = 2$ 和 $\theta = 0.99$。右图展示了随机值 $Y(Ax_0)$ 的观测频数 o_j 的直方图，其中 $Ax_0 \doteq 19.89983$。为便于参照，给出了近似泊松分布的均值 $M\bar{F}(Ax_0) \doteq 4.342498$。

定理 6.2. 在定理*6.1*的假设下，对于任意有限数组 $0 < x_1 < \cdots < x_k < x_{k+1} = \infty$，增量 $\{Y(Ax_i) - Y(Ax_{i+1}) \ \ i = 1, \ldots\ldots, \ k\}$ 渐近独立，其边缘分布分别可由参数为 $\lambda_i = M\big(\bar{F}(Ax_i) - \bar{F}(Ax_{i+1})\big)$ 的泊松分布近似，误差（全变差）上界为 $O(M^{-1}) = o(1)$。

注记 6.1. 最左侧的增量 $Y(0) - Y(Ax_1) = M - Y(Ax_1)$ 不包含在陈述中，因为它可以通过其他增量（线性地）表示，

$$Y(0) - Y(Ax_1) = M - \sum_{i=1}^{k} \big(Y(Ax_i) - Y(Ax_{i+1})\big).$$

定理 *6.2* 的证明。 注意，增量的联合分布为多项分布，参数为 M（试验次数），概率为 $p_i = F(Ax_i) - F(Ax_{i+1})$，分别对应于属于区间 $[Ax_i, Ax_{i+1})$ 和 $(i = 1, \ldots, k)$ 的结果。根据一般的泊松逼近定理（见 [28]），即可得到结论。特别地，记 $p := p_1 + \cdots + p_k$，则已知全变差误差的上界为 $O(Mp^2)$。在我们的情形下，$p = \sum_{i=1}^{k} \big(\bar{F}(Ax_i) - \bar{F}(Ax_{i+1})\big) = \bar{F}(Ax_1)$，估计式 $O(M^{-1})$ 便可如定理 6.1 中那样直接得到。

□

注 6.2。 为了理解定理 6.2 中增量渐近独立性的缘由，考察相邻增量之间的协方差在极限下的情形会很有帮助。

消失　事实上，对于任意 $0 < x < x' < x'' \le \infty$，借助 (2.17) 可得　，

$$\begin{aligned}
&\operatorname{Cov}\big(Y(Ax) - Y(Ax'),\, Y(Ax') - Y(Ax'')\big) \\
&\quad = M\bar{F}(Ax')\,F(Ax) - M\bar{F}(Ax'')\,F(Ax) \\
&\qquad - M\bar{F}(Ax')\,F(Ax') + M\bar{F}(Ax'')\,F(Ax') \\
&\quad \to \varphi_\nu(x') - \varphi_\nu(x'') - \varphi_\nu(x') + \varphi_\nu(x'') = 0,
\end{aligned}$$

再次利用定理 4.2 中的收敛性。

定理 6.2 的有限维近似可以统一表述为一个在反向时间下考察的适当泊松过程。具体而言，考虑一个非齐次泊松分布过程 $(\xi_t,\, t \ge 0)$，　$\xi_0 = 0$，其积分速率函数为 $\Lambda(t) = MF(A/t) \sim B\,\varphi_\nu(1/t)$，也就是说，$\mathrm{E}(\xi_t) = \Lambda(t)$。

定理 6.3。 在定理 *6.1* 和定理 *6.2* 的假设下，过程 $(Y(Ax)\ \ 0 < x \le \infty)$ 的分布可由随机过程$(\xi_{1/x},\, 0 < x \le \infty)$的分布近似。

因此，不同于第 4 节和第 5 节的“常规”情形（在该情形下，随机杨图收敛到极限形状 $\varphi_\nu(x)$），在本节所研究的混沌情形中，函数 $\varphi_\nu(x)$ 的作用是确定泊松近似的（积分）速率。

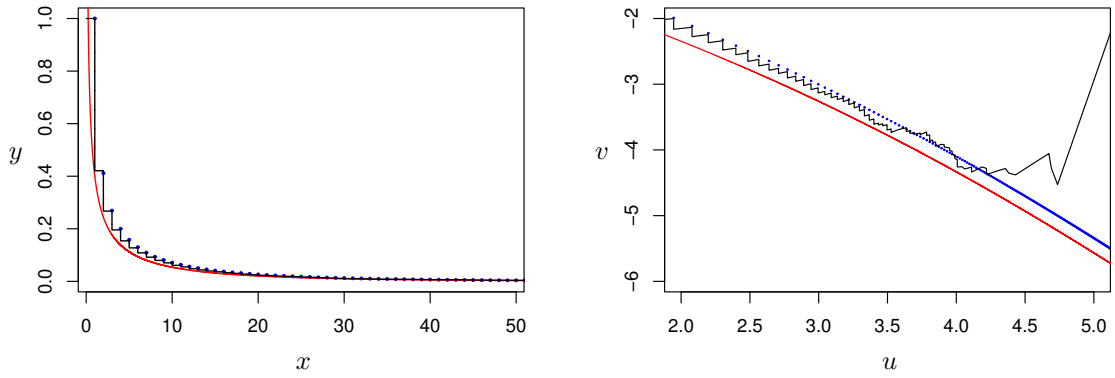
7. 真实数据示例

在本节中，我们将考察理论极限形状 $\varphi_\nu(x)$ 与 Sichel [41]此前研究的一些真实数据集的吻合程度。

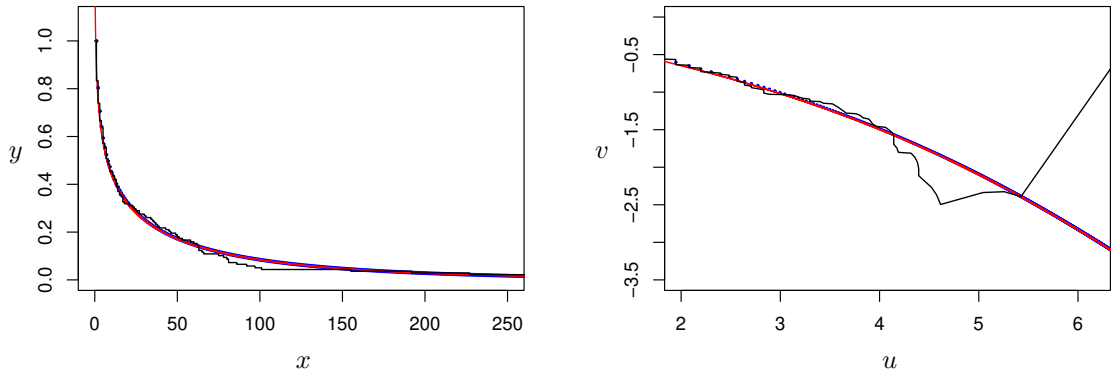
7.1. Lotka 的数据集：作者生产力

我们从 Lotka 在其开创性论文 [25], 中考察的一个经典数据集开始，该数据集统计了 1907–1916 年间在《化学文摘》发表论文的作者及其论文（物品）数量。这一数据集通常被视为计数幂律统计的基准示例 [10, 25], ，但 Sichel [41, （第316–317页及表 2]）认为，具有预先设定参数 $\nu = -0.5$、 $\alpha = 0$ 并估计 $\theta = 0.96876$ 的 GIGP 模型对数据的拟合更好（在 Pearson χ^2检验中， p值为 91.8%）。这两个模型的显著区别当然在于其长尾行为：分别呈幂律和幂几何形式。

为了从图形上考察拟合优度，与第 4.2 节类似，我们首先绘制经验杨图 $Y(x)$，并将其与拟合得到的 GIGP 互补分布函数 $\bar{F}(x)$ 进行比较，同时与缩放回原始坐标后的理论极限形状进行对照，即 $x \mapsto B\,\varphi(x/A)$。这里， $M = 6891$，由 (4.2) 和 (4.6) 计算得到的缩放系数分别为 $A \doteq 31.5076$ 和 $B \doteq 343.5839$。尽管 A 的值并不特别大（因为 θ 并非极其接近 1），但 B 的较大值证实，数据具备较好的极限形状近似基础。



(a) 洛特卡数据 [25, 41]: GIGP 参数 $\nu = -0.5, \alpha = 0$ （预先设定）， $\theta = 0.96876$ （估计）



(b) 陈的数据 [9, 41]: GIGP 参数 $\nu = 0, \alpha = 0$ （预先设定）， $\theta = 0.99369$ （估计）。

图 4：GIGP 模型对真实数据集的拟合。黑色曲线表示数据，蓝色虚线表示拟合得到的 GIGP 互补分布函数，平滑的红线描绘了缩放还原后的极限形状图像。右侧面板显示了这些图在变换坐标 (4.10) 中的尾部版本。

图 4(a) 左图显示，GIGP 模型和缩放回原始坐标后的极限形状 $x \mapsto B\,\varphi(x/A)$ 都能很好地拟合大部分数据。对图 4(a) 右图中的尾部进行目视检查（采用变换坐标 (4.10)）可确认，对于适度大的观测值 j，拟合效果良好；但在区域 $u \geq 4.5$ 中，对应于数值 $j \geq \mathrm{e}^{4.5} \approx 90$，与 GIGP 预测的偏差逐渐增大。这表明，洛特卡数据尾部的极大值需要采用不同的拟合模型，例如拉伸指数近似 [24]。顺便一提，仔细查看洛特卡数据集中的极端值后会发现，大于 90 的数值寥寥无几，分别为 95、107、109、114 和 346。令人惊讶的是，最大值 346 似乎是一个真正的离群值，比第二大值高出三倍！有趣的是，这一纪录属于埃米尔·阿布德哈尔登教授，他是一位多产且颇具争议的瑞士生物化学家和生理学家，活跃于 20 世纪上半叶 [47]。由于这一数据点相当异常，或许应将其从统计分析中剔除。

7.2. Chen 的数据集：期刊使用情况

作为第二个真实数据示例，我们重新考察 Chen [9], 的数据集。Sichel [41, （第318–319 页及表 4] ）曾为检验 GIGP 模型而研究这组数据。数据记录了1971年麻省理工学院科学图书馆中物理学期刊的使用次数（物品数），按从书架取出供阅读或复印的每册卷数（来源）统计。涉及的来源总数（即曾被借取的卷册数）为 $M = 138$。Sichel 使用预先设定的值 $\nu = 0$、 $\alpha = 0$ 和估计得到的 $\theta = 0.99369$ 拟合了 GIGP 模型。他通过 Pearson χ^2 检验拟合优度，观察到相当高的 p 值，为 31.2%，因此没有显示出显著的不匹配。

为了用我们的方法进一步检验拟合情况，我们像第7.1节一样，将经验杨图的边界 $\bar{}$（ Y）与拟合的 GIGP 函数 F（x）以及缩放回原尺度的极限形状 $B\,\varphi$（x/A）绘制在一起（见图4(b)左图），其中根据式（4.2）和（4.4）计算的缩放系数分别为 $A \doteq 157.9781$ 和 $B \doteq 27.24247$。值得指出的是，与第7.1节所考察的 Lotka 的数据集不同，这里的 x 缩放系数 A 值相当大，而 y 缩放参数 B 的值相对较小。

乍一看，图 4(b)（左图）中的图形似乎符合 GIGP 模型；然而，不难注意到，在 $x = 100$ 附近明显偏离了理论预测。查看变换坐标中的尾部图 (4.10) 后，这一点便得到证实（见图 4(b) 右图），其中 $x = 100$ 对应于 $u = \log\ 100 \doteq 4.60517$。为了通过统计检验判断这些偏差是否显著，我们可以利用定理 5.1 得出的 $Y(x)$ 的渐近正态性。具体而言，令 $x = 100$，并按照公式 (5.1) 进行标准化，我们计算得到 $\Upsilon(100/A)$ $\doteq -3.413073$，其 p 值极小，仅为 0.032%。因此，该偏差高度显著，这意味着 GIGP 模型拟合得并不准确，至少对于从约 $x = 70$ 开始的中等偏大取值而言如此。

8. 结论

本文研究了 GIGP 分布这一类分布下项目生产剖面的渐近收敛性。该剖面通过杨图进行可视化并经过适当缩放，最终收敛于极限形状。GIGP 分布由 Sichel [36] 提出，并已成功应用于多个计数数据集，包括信息计量学中常见的应用场景 [41]。当生产来源的数量很大时 $M \gg 1$，极限得以实现；但同时还需满足一项自然假设，即每个来源的物品数渐近趋于很大，从而进入参数区域 $\theta \approx 1$，其中 $\theta \in (0,1)$ 是 GIGP 的一个尺度参数。

所得极限形状族已被确定为不完全伽马函数 $\varphi_\nu(x) = \int_x^\infty s^{\nu-1}\,\mathrm{e}^{-s}\,\mathrm{d}s$，其中 $\nu \geq -1$ 是 GIGP 形状参数（表示为贝塞尔函数 $K_\nu(z)$ 的阶数，该函数参与 GIGP 概率分布的定义）。合适的 x 缩放系数具有普适性，为 $A = -1/\log\,\theta \to \infty$；但 y 尺度 B（与来源数量 M 成正比）取决于 ν 的符号。特别地，由此可知， $\nu > 0$ 时 $B \to \infty$ （这保证了向极限形状收敛），但在 $\nu \leq 0$ 时这一点可能不成立，例如因为 M 不够大。对于后一种情况（称为“混沌”情形），我们证明产出剖面可由一个合适的泊松过程近似，其速率定义为极限形状 $\varphi_\nu(x)$ 的函数。

在正则情形（即收敛到极限形状成立的情形）下，我们还证明了涨落的渐近正态性。

我们通过计算机模拟（使用 $\nu = 0.5$ and$\nu = -0.5$）展示了理论结果：在正则情形下，极限形状与经验数据高度吻合，同时也证实了混沌状态下的泊松统计。我们还提出了一种简单的数据变换，可将分布的尾部线性化，从而有助于通过可视化诊断判断 GIGP 模型是否适用。

将我们的方法应用于现实数据集时，可以为提取有关 GIGP 模型拟合的有用信息提供一种新颖途径。Lotka 作者生产力数据是其中一个经典例子 [25], ，拟合得到的 GIGP 模型与数据高度吻合，这一点可通过我们的图形图直观地看出。然而，对适当增强后的上尾部进行绘图后，可以发现数据与 GIGP 模型存在一定偏离，而诸如 Pearson χ^2-检验或基于概率分布间一致距离的 Kolmogorov–Smirnov检验 [41] 等常用的汇总拟合优度工具并未捕捉到这一偏离。识别出的极端值可以追溯至原始数据并加以解释，尤其是能够标记出第7.1节提到的一个非常特殊的“离群值”。

此外，在我们使用陈的数据研究期刊使用情况的第二个示例中， [9], 我们的方法发现，在某个计数范围内，数据频率图与理论 GIGP 预测存在明显偏差。凭借我们关于涨落渐近正态性的结果，我们证明了这些偏差具有高度显著性。这与传统工具给出的相当可信的拟合优度“通过”结论形成对比， [41], 再次展现了我们方法的优越性。我们预计，本文提出的方法也可能适用于更广泛的各种计数数据场景。

A. 贝塞尔函数的渐近公式

以下列出贝塞尔函数 $K_\nu(z)$ 的一些有用性质，包括自变量 z 和阶数 ν 处于不同范围时的一些渐近公式。为便于使用，我们在此汇集这些性质，并参考 NIST手册 [27]。

引理 A.1 ([27], 10.27.3)。 对于任意 ν 和 z,

$$K_{-\nu}(z) = K_\nu(z). \tag{A.1}$$

引理 A.2. 令 ν 为固定值，且 $z \to 0+$。

(a) ([27], 10.30.2) 若 $\nu > 0$ 则

$$K_\nu(z) \sim \tfrac{1}{2}\Gamma(\nu)\left(\tfrac{1}{2}z\right)^{-\nu}. \tag{A.2}$$

(b)（借助 10.25.2 的 [27], 10.31.1）若 $\nu = 1$ 则

$$K_1(z) = z^{-1} + \tfrac{1}{2}z\log z + O(z). \tag{A.3}$$

(c)（借助 10.25.2 的 [27], 10.31.2）若 $\nu = 0$ 则

$$K_0(z) = -\log\left(\tfrac{1}{2}z\right) - \gamma + O(z^2 \log z). \tag{A.4}$$

其中 $\gamma = 0.5772\ldots$ 是欧拉常数（*[27], 5.2.3*）。特别地，

$$K_0(z) \sim -\log z. \tag{A.5}$$

(d)（借助 5.5.3，由 [27], 10.27.4 和 10.25.2 得出）对于 $-1 < \nu < 0$,

$$K_\nu(z) = \tfrac{1}{2}\Gamma(-\nu)\left(\tfrac{1}{2}z\right)^{\nu} + \frac{\Gamma(\nu+1)}{2\nu}\left(\tfrac{1}{2}z\right)^{-\nu} + O(z^{\nu+2}). \tag{A.6}$$

引理 A.3 ([27], 10.41.2)。 如果 $z \neq 0$ 是固定的且 $\nu \to +\infty$，那么

$$K_\nu(z) \sim \sqrt{\frac{\pi}{2\nu}}\left(\frac{\mathrm{e}z}{2\nu}\right)^{-\nu}. \tag{A.7}$$

参考文献

[1] Arratia, R., Barbour, A.D., Tavar´e, S. 对数组合结构：一种概率方法。EMS 数学专著。欧洲数学会：苏¨黎世，2003年；doi:10.4171/000。[2] Auluck, F.C., Kothari, D.S. 统计力学与整数分拆。剑桥哲学学会会刊**42** (1946年)，272–277；doi:10.1017/S030500410002303。[3] Barbour, A.D., Holst, L., Janson, S. 泊松近似。牛津概率论研究丛书，**第2卷**。牛津科学出版物。克拉伦登出版社：牛津，1992年。[4] Barndorff-Nielsen, O.E., Blæsild, P., Seshadri, V. 具有广义逆高斯边际分布的多元分布及相关的泊松混合。加拿大统计学杂志**20** (1992年)，109–120；doi:10.2307/3315462。[5] Bogachev, L.V. 等权部分随机整数分拆乘法系综极限形状的统一推导。随机结构与算法**47** (2015年)，227–266；doi:10.1002/rsa.20540。[6] Bogachev, L.V., Gnedin, A.V., Yakubovich, Yu.V. 占用箱子数量的方差。应用数学进展**40** (2008年)，401–432；doi:10.1016/j.aam.2007.05. 002。[7] Borisov, I., Jetpisbaev, M. 一类加性统计量的泊松化原理。数学**10** (2022年)，4084；doi:10.3390/math10214084[8] Boucheron, S., Lugosi, G., Massart, P. 集中不等式：独立性的非渐近理论。牛津大学出版社：牛津，2013年；doi:10.1093/acprof:oso/9780199535255.001.0001。

[9] Chen, C.-C. 大型学术研究图书馆中物理学期刊的使用模式。美国信息科学学会期刊23 (1972年)，254–270；doi:10.1002/asi.4630230405。[10] Clauset, A., Shalizi, C.R., Newman, M.E.J. 经验数据中的幂律分布。SIAM评论51 (2009年)，661–703；doi:10.1137/070710111。[11] Coile, R.C. 科学生产力的洛特卡频数分布。美国信息科学学会期刊28 (1977年)，366–370；doi:10.1002/asi.4630280610。[12] Cram´er, H. 统计数学方法。普林斯顿数学丛书，第9卷。普林斯顿大学出版社：新泽西州普林斯顿，1946年；https://www.jstor.org/stable/j.ctt 1bpm9r4。[13] Duchon, P., Flajolet, P., Louchard, G., Schaeffer, G. 用于随机生成组合结构的玻尔兹曼抽样器。组合学、概率与计算13 (2004年)，577–625；doi:10.1017/S0963548304006315。[14] Egghe, L. 信息生产过程中的幂律：洛特卡信息计量学。图书馆与信息科学，第5卷。爱思唯尔学术出版社：阿姆斯特丹，2005年；Emerald：宾利，2005年；doi:10.1108/S1876-0562(2005)0000005004。[15] Egghe, L., Rousseau, R. 广义成功孕育成功原则及其导出的时变信息计量分布。美国信息科学学会期刊46 (1995年)，426–445；doi:10.1002/(SICI)1097-4571(199507)46:6¡426::AID-ASI3>3.0.CO;2-B。[16] Fristedt, B. 大整数随机分拆的结构。美国数学会汇刊337 (1993年)，703–735；doi:10.1090/S0002-9947-1993-1094553-1。[17] Gnedin, A., Hansen, B., Pitman, J. 关于无限多个盒子的占位问题：一般渐近性质与幂律。概率调查4 (2007年)，146–171；doi:10.1214/07-PS092。[18] Gupta, R.C., Ong, S.H. 用泊松混合分析长尾计数数据。统计学通讯：理论与方法34 (2005年)，557–573；doi:10.1081/STA-200052144。[19] Hirsch, J.E. 衡量个人科学研究产出的指标。美国国家科学院院刊102 (2005年)，16569–16572；doi:10.1073/pnas.0507655102。[20] Huber, J.C. 生成洛特卡定律的新模型。美国信息科学学会期刊53 (2002年)，209–219；doi:10.1002/asi.10025。[21] Johnson, N.L., Kemp, A.W., Kotz, S. 单变量离散分布，第3版。威利概率与数理统计丛书。威利出版社：纽约，2005年；doi:10.1002/ 0471715816。[22] Johnson, N.L., Kotz, S., Balakrishnan, N. 单变量连续分布，第1卷，第2版。威利概率与数理统计丛书。威利出版社：纽约，1994年。

[23] Karlin, S. 某些无限瓮方案的中心极限定理。《数学与力学杂志》17（1967），373–401；https://www.jstor.org/stable/24902077。[24] Laherr`ere, J., Sornette, D. 自然界和经济中的拉伸指数分布：“肥尾”与特征尺度。《欧洲物理学杂志B辑》2（1998），525–539；doi:10.1007/ s100510050276。[25] Lotka, A.J. 科学生产力的频数分布。《华盛顿科学院杂志》16（1926），317–323；http://www.jstor.org/stable/24529203。[26] Nicholls, P.T. 齐夫参数的估计。《美国信息科学学会期刊》38（1987），443– 445；doi:10.1002/(SICI)1097-4571(198711)38:6<443::AID-ASI4>3.0.CO;2-E；勘误：Kraft, D.H. 同上，39（1988），287；doi:10.1002/(SICI)1097-4571(198807)39:4<287:: AID-ASI10>3.0.CO;2-T。[27] Olver, F.W.J. 等编，《NIST数学函数手册》。美国国家标准与技术研究院，美国商务部：Washing- ton, DC；剑桥大学出版社：剑桥，2010；https://www.cambridge.org/ catalogue/catalogue.asp?isbn =9780521192255；在线版：NIST数学函数数字图书馆，1.1.8版，发布日期：2022年12月15日；https://dlmf.nist. gov（访问日期：2023年1月8日）。[28] Novak, S.Y. 泊松近似。《概率调查》16（2019），228–276；doi:10. 1214/18-PS318。[29] Nuermaimaiti, R., Bogachev, L.V., Voss, J. 引文的广义幂律模型。载于《第18届科学计量学与信息计量学国际会议（ISSI 2021）论文集》（鲁汶，2021），Gl¨a nzel, W. 等编。国际科学计量学与信息计量学会（I.S.S.I.），2021，第843–848页；https://issi2021.org/proceedings/。[30] Pittel, B. 关于随机费雷尔图的一种可能形状。《应用数学进展》18（1997），432–488；doi:10.1006/aama.1996.0523。[31] Price, D.J.S. 科学论文网络。《科学》149（1965），510–515；doi:10.1126/ science.149.3683.510。[32] Price, D.J.S. 文献计量及其他累积优势过程的一般理论。《美国信息科学学会期刊》27（1976），292–306；doi:10.1002/asi.4630270505。[33] Redner, S. 你的论文有多受欢迎？引文分布的实证研究。《欧洲物理学杂志B辑》4（1998），131–134；doi:10.1007/s100510050359。[34] Resnick, S.I. 《极端值、正则变差与点过程》。应用概率：应用概率信托丛书，第4卷。施普林格：纽约，1987；doi:10. 1007/978-0-387-75953-1。[35] Shiryaev, A.N. 《概率论》，第2版。数学研究生教材，第95卷。施普林格：纽约，1996；doi:10.1007/978-1-4757-2539-1。

[36] Sichel, H.S. 一类特别适合表示长尾频数数据的离散分布。载于《第三届数学统计学研讨会论文集》，Laubscher, N.F. 编。科学与工业研究委员会（CSIR）：比勒陀利亚，1971年，第51–97页。[37] Sichel, H.S. 含金刚石矿床的统计估值。《南非矿业学会期刊》73（1973），235–243；https://hdl.handle.net/10520/AJA0038223X 302。[38] Sichel, H.S. 一种表示书面散文句子长度的分布。《皇家统计学会会刊A辑》137（1974），25–34；doi:10.2307/2345142。[39] Sichel, H.S. 词频分布定律。《美国统计学会期刊》70（1975），542–547；https://doi:10.1080/01621459.1975.10482469。[40] Sichel, H.S. 重复购买与广义逆高斯-泊松分布。《皇家统计学会会刊C辑》31（1982），193–204；doi:10.2307/2347993。[41] Sichel, H.S. 一种真正有效的文献计量分布。《美国信息科学学会期刊》36（1985），314–321；doi:10.1002/asi.4630360506。[42] Sichel, H.S. 用广义逆高斯-泊松分布对物种丰度频数和物种-个体函数建模。《南非统计学期刊》31（1997），13–37；https://hdl.handle.net/10520/AJA0038271X 555。[43] Vershik, A.M. 渐近组合学与代数分析。载于《国际数学家大会论文集》（Z ¨urich，1994），第2卷，Chatterji, S.D. 编。Birkh ¨auser：巴塞尔，1995，第1384–1394页；doi:10.1007/978-3-0348-9078-6 133。[44] Vershik, A.M. 组合划分的统计力学及其极限形状。《函数分析及其应用》30（1996），90–105；doi:10.1007/BF02509449。[45] Vershik, A.M. 从自然数划分理论的角度看量子理想气体能量的极限分布。《俄罗斯数学调查》52（1997），379– 386；doi:10.1070/RM1997v052n02ABEH001782。[46] Whitaker, L. 关于小数定律。《生物计量学》10（1914），36–71；doi: 10.1093/biomet/10.1.36。[47] 维基百科，自由百科全书。“埃米尔·阿布德哈登”。在线网址：https://en.wikipedia.org/wiki/Emil Abderhalden（访问日期：2023年3月10日）。[48] Yeh, J.《鞅与随机分析》。多元分析丛书，第1卷。世界科学出版社：新加坡，1995；doi:10.1142/2948。[49] Yong, A. 对赫希引文指数的批判：一个组合费米问题。《美国数学会通报》61（2014），1040–1050；doi:10.1090/noti1164。