arXiv:2601.04078v1 [math.CO] 7 Jan 2026

序列中的模式

Richard Kenyon

2026年1月8日

摘要

我们研究二进制序列中的模式密度，找出具有固定模式密度的最优极限序列。

1 引言

给定两个有限二进制字 $w \in \{0,1\}^m$ 和 $X \in \{0,1\}^n$，其中 $m \le n$，我们令$N_w(X)$ 表示 w 作为X的非连续子字出现的次数。例如， $N_{10}(0100101) = 4$。我们定义 $\rho_w(X) := N_w(X)/\binom{n}{m}$，即模式密度 w 在 X中。这是从 X 中随机选取一个长度为m 的子序列恰为 w的概率。

长度为 n 的序列 X，其模式密度 ρ_w 最大可能是多少？具有给定密度 ρ_w 的典型序列 X 是什么样的?

这类模式计数问题出现在组合学的多个不同领域。对于二进制序列，最简单的例子是模式 10，它对应于整数分拆：将 X 表示为杨图，其面积为 $N_{10}(X)$（见图1）。一般而言，模式计数与各种整数中的路径之间存在联系

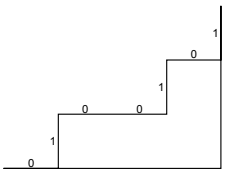


图 1：二进制词可以表示为杨图的边界：一条格路，其中 0 表示向东的一步，1 表示向北的一步。于是， N_{10} 就是杨图下方的面积。在这里， $N_{10}(0100101) = 4$。

海森堡群。对于由.....生成的标准海森堡群

$$M_0 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

正元素乘积的右上角元素为 N_{01}。例如， $M_0M_1M_0M_1M_1$ 的右上角元素为 N_{01} (01011)。见下文第7节。二进制模式计数问题也被证明与 $[0,1]$, 上的某些极值概率测度有关，例如， $[0,1]$ 上的概率测度使 $x(y-x)$最大，其中 x， y 分别是两个独立样本中较小和较大的那个；见下文。

在图的子图密度问题 [10, 1, 8] 以及排列和模式密度问题 [6, 7]中，人们提出了类似的模式密度问题，并对其进行了部分解答。另请参见 [2]，其中给出了一个通用框架。

二进制模式计数的一个具体例子是“BRBR 游戏”：从一副由 52 张扑克牌组成的标准扑克牌组中，均匀随机抽取四张牌，共有 $\binom{52}{4}$ 种可能；抽出的牌保持它们在牌组中的原有顺序。目标是排列原始牌组，使抽到 BRBR 的概率最大，也就是说，第一张和第三张是黑牌，第二张和第四张是红牌。如果牌组经过随机洗牌，获胜概率约为 $1/16$ 。如果牌组按“新牌顺序”排列，也就是先放 13 张黑牌，再放 13 张红牌，接着放 13 张黑牌，最后放 13 张红牌，获胜概率几乎高出 70%：约为 0.1055。怎样排列牌组才能使获胜概率最大?

BRBR 游戏是模式计数的一个特例：对模式 1010 进行优化。有趣的是，可以做得比“新牌顺序”更好；我们将在下文展示，对于大型牌组（其中红牌和黑牌各占一半），最大概率趋近于 $\frac{3}{4e^2}$ ，并刻画相应的最优排列。（对于 52 张牌的牌组，表面上最优的排列具有 ≈ 0.114的获胜概率，见图 2。）这一

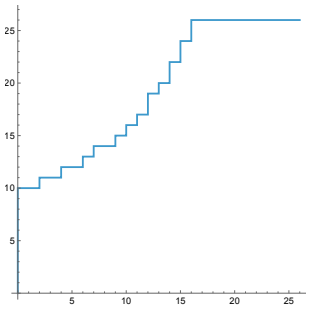


图 2：BRBR 游戏中一个猜想的 52 张牌牌组最优排列。从原点出发，向北的一步表示一张黑牌；向东的一步表示一张红牌。该排列给出 $p \approx 0.114$。与图 5进行比较。

这一非解析的例子展示了一种令人惊讶的现象，即极限形状模式有时可以是图 4 和5所示。这是模式计数模型的一个不寻常特征，也使分析变得颇具挑战性。

本文考虑以下两个一般的模式计数问题。

1. 对于给定的模式集合 w_1、……、 w_k，当 X 长度较大时，允许的密度元组 (ρ_{w_1}(X)……、 ρ_{w_k}(X)) 的“可行集”是什么?

2. 对于较大的 n，长度为 n 且具有给定模式密度 $\rho_{w_i} = \delta_i$ 的典型序列看起来像什么?

我们在两个模式的情形下回答第一个问题，即 w_1，w_2，其中 $w_1 = 1$ 和$w_2 = w$ 是任意的。对于某些短模式集合，我们回答第二个问题；当所有模式均为 $w_k = 1^k0$ 类型时，我们回答这两个问题。我们还讨论模式密度之间的代数关系，并提出一个关于代数独立的模式密度数量的猜想。

模式计数的适当极限对象称为 子勒贝格测度。子勒贝格测度是一个测度 $f(x)\,dx$，定义在 $[0,1]$ 上，并且相对于勒贝格测度绝对连续，其密度有界于 1： $f(x) \in [0,1]$。我们令 Ω 表示子勒贝格测度的集合。二进制序列 $X = (x_1, \ldots\ldots, x_n)$ 具有一个相应的子勒贝格测度 $\mu_X = f_X(x)dx \in \Omega$，其中 f_X 是取 $\{0,1\}$值的阶梯函数，其在 $[\frac{i-1}{n}, \frac{i}{n})$ 上的值为 x_i。

假设有一个序列 $\{X_n\}_{n=1,\ 2,\ \ldots\ldots}$，其 $X_n \in \{0,1\}^n$ 对应的测度 μ_{X_n} 弱收敛到次勒贝格测度 $\mu(x) = f(x)dx$。此时，对于每个模式 w， w 的极限密度 $\lim_{n\to\infty} \rho_w(X_n)$ 都存在，且仅是 μ 的函数。例如

$$\rho_1(\mu) = \int_0^1 f(x)\,dx = F(1)$$

其中 $F(y) = \mu([0,y]) = \int_0^y f(x)\,dx$ 是 μ 的分布函数。另一个例子是

$$\rho_{10}(\mu) = 2\iint_{0\le x<y\le 1} f(x)(1-f(y))\,dx\,dy$$

一般而言，对于长度为 k 的单词 w，

$$\rho_w = k!\int_{0\le x_1<\cdots<x_k\le 1} g(x_1)g(x_2)\ldots g(x_k)dx_1\ \ldots\ dx_k \tag{1}$$

其中

$$g(x_i) = \begin{cases} f(x_i) & w_i = 1 \\ 1-f(x_i) & w_i = 0. \end{cases}$$

本文结构安排如下。在第2 节中，我们证明模式密度刻画了子勒贝格测度，反之亦然。因此，从本质上说，子勒贝格测度是讨论模式密度时恰当的极限对象，正如图子是讨论子图密度时恰当的极限对象一样。在第3节中，我们讨论模式密度之间的各种代数关系，找出长度不超过 5 的模式的所有关系，并对一般长度下独立模式的数量提出猜想。在第4 节中，我们讨论可行域并给出一些例子。我们证明（定理4.1），两个模式 ρ_1 和 ρ_τ的可行域 $E_{1,\ \tau}$ 具有如下形式

$$E_{1,\tau} = \{(\rho_1,\rho_\tau)\ :\ \rho_1 \in [0,1]\ \text{ and }\ 0 \le \rho_\tau \le C_\tau \rho_1^k (1-\rho_1)^l\}$$

对于一个与 ρ_1无关的常数 C_τ。这里 k，l 分别是 τ中 1 和 0 的个数。本节还分析 BRBR 游戏。在第5 节中，我们讨论具有固定模式密度的序列的极限形状。在第6节中，我们讨论模式 1^k0的情形，并求出明确的极限形状。最后，在第7 节中，我们讨论其与上三角矩阵的关系。

致谢。 本研究部分由 Simons 基金会资助项目 327929 资助。我们感谢 Pete Winkler 与我们讨论并由此促成了这项工作。我们感谢 Pavel Galashin 的讨论，以及他指出了与全正性的联系。我们感谢 Mei Yin 和 Persi Diaconis 提供的意见和建议。

2 模式密度与子勒贝格测度

子勒贝格测度空间上自然的距离概念是 Wasserstein 距离，等价地，也可以用分布函数上的 L^1 度量来表示:

$$d_W(\mu_1, \mu_2) = \int_0^1 |F_1(x) - F_2(x)|\, dx$$

其中，$F_i(x) = \mu_i([0, x])$ 是 μ_i 的分布函数。相应的度量拓扑也是测度弱收敛所对应的拓扑。

定理 2.1。 两个子勒贝格测度几乎处处相等，当且仅当它们具有相同的模式密度。更一般地，一列子勒贝格测度在度量 d_W 下收敛，当且仅当其模式密度集合收敛。

证明。 由（1），每个模式密度都是密度 f的积分，因此，若一列密度弱收敛，则模式密度也收敛。反过来，豪斯多夫矩定理 [5] 指出，子勒贝格测度 $f（x）dx$ 由其矩 $m_n = \int x^n f（x）dx$决定；再次利用（1），每个矩 m_n 都是模式密度的有限和 $m_n = \frac{1}{n+1} \sum_{w \in \{0,1\}^n} \rho_{w1\circ}$ —— □

3 代数关系

N_τ之间存在许多代数关系。例如，对于长度为 1 的字符串 τ，对于长度为 n 的字符串 X，我们有:

$$N_0(X) + N_1(X) = n.$$

对于长度为 2 的 τ，我们有以下关系

$$N_{00} = \binom{N_0}{2}$$
$$N_{01} + N_{10} = N_0 N_1$$
$$N_{11} = \binom{N_1}{2}.$$

例如，中间那个等式表示，对于 X 中任意选取的一个 0 和一个 1，要么 0 在 1 之前，要么 1 在 0 之前。这四个等式意味着，给定 n，在长度为 ≤ 2 的字符串 τ 中，只有两个代数独立的模式计数数量，例如 $\{N_1, N_{10}\}$。

对于长度为 3 的模式，我们有类似的线性关系

$$\begin{aligned} N_{000} &= \binom{N_0}{3} \\ N_{001} + N_{010} + N_{100} &= \binom{N_0}{2} N_1 \\ N_{110} + N_{101} + N_{011} &= N_0 \binom{N_1}{2} \\ N_{111} &= \binom{N_1}{3}. \end{aligned} \tag{2}$$

还有另外四个关系

$$N_0 N_{10} = N_{010} + 2N_{100} + N_{10} \tag{3}$$
$$N_0 N_{01} = N_{010} + 2N_{001} + N_{01} \tag{4}$$

及其互补关系（将 0 和 1 互换）。例如，方程（3）可通过在 X 中取一个 10 模式、在 X 中取一个 0，并考虑 0 与 10 模式的 4 种位置关系得出：在其左侧、两个数字之间、其右侧，或位于 0 的上方。

除了长度为 1 和 2 的字符串外，这 8 个关系还给出另外 2 个代数独立数量，例如 N_{100} 和 N_{110}。要证明 N_1、N_{10}、N_{100} 和 N_{110} 代数独立，请注意，字符串 01110001 和 10100110 具有相同的 N_1、N_{10} 和 N_{100}，但 N_{110}不同。

进一步地，对于长度为 4 的情形，我们有 5 个与（2）类似的线性关系

$$\sum_{x_i \in \{0,1\} : \sum x_i = k} N_{x_1 x_2 x_3 x_4} = \binom{N_0}{4-k} \binom{N_1}{k}.$$

在四个数量 N_{0001}、N_{0010}、N_{0100}、N_{1000} 之间，我们还有其他关系

$$\begin{aligned} N_0 N_{001} &= 3N_{0001} + N_{0010} + 2N_{001} \\ N_0 N_{010} &= 2N_{0010} + 2N_{0100} + 2N_{010} \\ N_0 N_{100} &= 3N_{1000} + N_{0100} + 2N_{100} \end{aligned}$$

因此，至多有一个新的代数独立量，暂且记为 N_{1000}。同样，在 N_{0111}、N_{1011}、N_{1101}、N_{1110} 中，至多有一个新的代数独立量，记为 N_{1110}。

对于其余六个量 N_{0011}、.....、 N_{1100}，我们有如下关系

$$N_0N_{011} = 2N_{0011} + N_{0101} + N_{0110} + N_{011}$$
$$N_0N_{101} = N_{0101} + 2N_{1001} + N_{1010} + N_{101}$$
$$N_0N_{110} = N_{0110} + N_{1010} + 2N_{1100} + N_{110}$$

以及另外三个交换0和1的操作。此外

$$N_{10}^2 = 2N_{1010} + 4N_{1100} + 2N_{110} + 2N_{100} + N_{10}.$$

综合这些结果，我们得到

$$N_{1001} = N_1N_{100} + N_{110} - \binom{N_{10}}{2},$$

这是一种令人惊讶的情况：一个长度为4的模式是长度严格小于4的模式的代数函数。总的来说，模去长度为 ≤ 3的模式后，我们发现至多有三个代数独立的长度为4的模式，我们将它们取为 N_{1000}，N_{1100}， N_{1110}。

现在我们断言， N_1， N_{10}， N_{100}， N_{110}， N_{1000}， N_{1100}， N_{1110} 实际上是代数独立的。考虑随 a_1、.....、 a_8变化的序列 $X = 1^{a_1}0^{a_2}\dots1^{a_7}0^{a_8}$ 即可看出这一点。上述每个 $N_\tau(X)$都是 a_i的多项式函数，并且（简单的计算机代数计算表明）以 a_i为自变量的 N_τ向量的雅可比矩阵对一般的 a_i具有满秩。

对于长度为5的序列，类似的计算表明，有6个新的代数独立模式；我们猜想，对于长度为 k 的序列，有$\binom{k-1}{[\ k-1\ \ ()/2]}$ 个新的代数独立模式，因此，对于长度不超过 k 的序列，代数独立模式的数量 $A(k)$可猜想为

$$A(k) = \sum_{j=1}^{k} \binom{j-1}{[\frac{j-1}{2}]}.$$

4 可行域

对于给定的词集 $W = \{w_1, \dots, w_k\}$，通常并非所有密度（$\delta_1, \dots, \delta_k$） $\in[0,1]^k$ 都是可行的，也就是说，能使这些词具有相应密度的子勒贝格测度集合可能为空。我们将 $E_W \subset [0,1]^k$ 定义为可行约束的集合:

$$E_W = \{(\delta_1, \dots, \delta_k) \in [0,1]^k \mid \exists \mu = f(x)dx \in \Omega : (\rho_{w_1}(f), \dots, \rho_{w_k}(f)) = (\delta_1, \dots, \delta_k)\}.$$

例如，对于给定的 $\delta_1 = \rho_1(f)$，当 f 是该函数时，密度 $\delta_{10} = \rho_{10}(f)$ 在值 $2\delta_1(1-\delta_1)$ 处达到最大值

$$f(x) = \begin{cases} 1 & x \leq \rho_1 \\ 0 & \text{else.} \end{cases},$$

而当.....时，在值 0 处达到最小值

$$f(x)=\begin{cases}0 & x\le 1-\rho_1\\ 1 & \text{else},\end{cases}$$

并且可以取 任意中间值（通过对这些值取凸组合得到）。因此

$$E_{\{1,10\}}=\{(\delta_1,\delta_{10})\ :\ 0\le\delta_1\le 1,\quad 0\le\delta_{10}\le 2\delta_1(1-\delta_1)\}.$$

见图3。

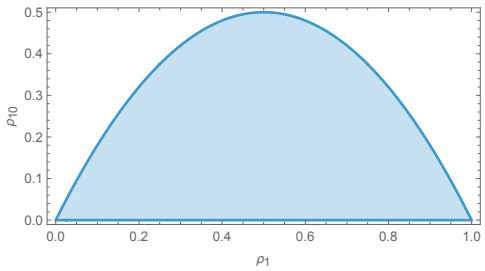


图3：模式 1 和 10 的可行域。

虽然一般来说难以确定 E_W，但有一种较简单的情形是 $W=\{1,w\}$，也就是说，当 W 由两个词组成，其中一个是 1（或 0）时。关于6.1 ，请参见定理；该定理给出了另一类 E_W 具有显式形式的一般情形。

定理 4.1. 对于固定的 $\rho_1=\rho$ ，密度为 ρ_τ 时的可行集是区间$[0,C_\tau\rho^m(1-\rho)^n]$，其中 m、n分别是 τ中1的个数和0的个数，而 C_τ是一个与 ρ无关的常数。

证明。 见定理4.2的证明中的（5） ，下文的证明很容易推广到这种情形。

□

计算 C_τ 的值可能相当复杂。有关某些 τ 取值下的 C_τ ，请参见下表。这些结果的证明在附录中略作说明。

模式	C-值
10	2
1^k0^l	$\binom{k+l}{k}$
$1^k0^l1^m$	$\binom{k+l+m}{k\ l\ m}$,, $\frac{k^k m^m}{(k+m)^{k+m}}$
1010	$\frac{12}{e^2}$
11010	$30e^{-\pi/\sqrt{3}}$
10110	$\frac{20}{9}$
10101	$\frac{30\xi^2}{(1+\xi)^2}$ 其中 $\xi e^{\xi}=e^{-1}$

作为一个特例，我们对 $\tau = 1010$ 的情况进行了完整分析。

定理 4.2。 对于固定的 $\rho_1 = \rho$，密度 ρ_{1010} 在 $\rho_{1010} = \frac{12}{e^2}\rho^2(1-\rho)^2$, 处达到最大值，且唯一使其达到最大值的次勒贝格测度的密度为

$$f(x) = \begin{cases} 1 & x < \frac{\rho}{e} \\ \frac{1}{2}(1 + \frac{e^{1/2}(x+\rho-1)}{\sqrt{4\rho(1-\rho)+e(x+\rho-1)^2}}) & \frac{\rho}{e} < x < 1 - \frac{1-\rho}{e} \\ 0 & 1 - \frac{1-\rho}{e} < x. \end{cases}$$

函数图像见 图 4 和 图 5 。注意，使最大 f-密度 ρ_1 达到最大值的 ρ_{1010}-密度 $\frac{12}{e^2}\rho_1^2(1-\rho_1)^2$ 为 $\rho_1 = \frac{1}{2}$，于是 $\rho_{1010} = \frac{3}{4e^2}$。

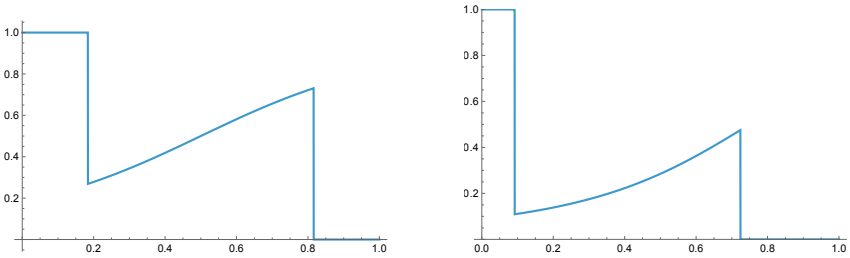


图 4: 密度图 $P_{(x)}$ 中 ρ_{1010}-最大化者在 $\rho_1 = \frac{1}{2}$ 时以及在 $\rho_1 = \frac{1}{4}$ 时的密度图 f。

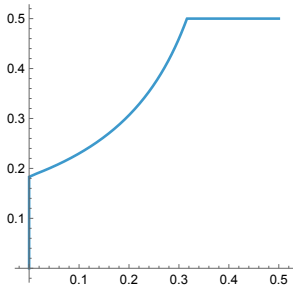


图 5: $\rho_1 = \frac{1}{2}$ 时，作为东北路径的 ρ_{1010}-最大化者图。

证明。 我们要最大化 ρ_{1010}，固定 $\rho = \rho_1$。由（1）可得

$$\rho_{1010} = 24 \int_{0<x<y<z<w<1} f(x)(1-f(y))f(z)(1-f(w))dx\,dy\,dz\,dw.$$

令 $F(x) = \int_0^x f(y)dy$ 为分布函数， $H = F^{-1}$ 为逆分布函数。由于 F 单调不减，H' 被定义为广义函数（H 可能

存在跳跃奇点），并且 $H'(y) \geq 1$，因为 $F'(x) \leq 1$。我们可以通过变量替换 $u = F(x)$ $x = H(u)$ 重写关于 ρ_{1010} 的积分，其中 u 从 0 变化到 ρ，而 x 从 0 变化到 1。因此

$$\begin{aligned}\rho_{1010} &= 24 \int_{x<y<z<w} f(x)(1-f(y))f(z)(1-f(w))dx\,dy\,dz\,dw \\ &= 24 \int_{0<p<u<q<v<\rho} (H'(u)-1)(H'(v)-1)dp\,du\,dq\,dv \\ &= 24 \int_{0<u<v<\rho} u(v-u)(H'(u)-1)(H'(v)-1)\,du\,dv.\end{aligned}$$

令 $h(u) := H'(u) - 1 \geq 0$ 给出

$$\rho_{1010} = 24 \int_0^\rho \int_0^v u(v-u)h(u)h(v)\,du\,dv$$

我们需要在非负的（广义）函数 h 上对其进行最大化，其中 h 在区间 $[0,\rho]$ 上积分为 $1-\rho$。注意尺度不变性：通过将 u, v 按 ρ 进行缩放，我们可以将 h 的定义域改为 $[0,1]$。然后可以按 $\frac{\rho}{1-\rho}$ 进行垂直缩放，使 h 的积分等于 1。换言之，令 $g(x) := \frac{\rho}{1-\rho} h(x\rho)$，则 g 定义在 $[0,1]$, 上，非负，积分为 1，且完全不依赖于 ρ。因此

$$\rho_{1010} = C\rho^2(1-\rho)^2 \tag{5}$$

其中 C 不依赖于 ρ:

$$C := 24 \max_g \int_{0<x<y<1} x(y-x)g(x)g(y)\,dx\,dy \tag{6}$$

最大值取遍所有非负的广义函数 $g \geq 0$ 且其积分为 1，也就是取遍 $[0,1]$上的概率测度。（(6) 的概率解释如下：从 g中抽取两个独立随机样本；令 x 为较小者，令 y 为较大者。哪种分布 g 能使 $x(y-x)$ 的期望值最大？）

为了求得最大值，我们考虑扰动 $g \to g + \varepsilon k$，其中 k 的积分为零。令 $C(g)$ 表示式（6）右端的积分（略去常数因子 24），则有

$$0 = \frac{d}{d\varepsilon} C(g+\varepsilon k)\Big|_{\varepsilon=0} = \int_{0<x<y<1} x(y-x)(g(x)k(y)+k(x)g(y))dx\,dy. \tag{7}$$

如果 a 是 $g(a) > 0$ 的一个点，我们可以选取 $k(u)$ 来逼近狄拉克 δ 函数的导数 $\delta'_a(u)$，经分部积分后可得

$$\begin{aligned}0 &= \int_0^a xg(x)dx + \int_a^1 (y-2a)g(y)dy \\ &= \int_0^1 xg(x)\,dx - 2a(G(1)-G(a))\end{aligned}$$

其中 $G(a) = \int_0^a g(x)\,dx$ 是 g 的分布函数。令 $2c = \int_0^1 xg(x)\,dx$，这表明 $G(a) = G(1) - \frac{c}{a}$，因此当 $g > 0$ 时，$g(a) = \frac{c}{a^2}$。

我们还必须考虑以下可能性：$g \equiv 0$ 在某个区间上（因为在这种情况下，无法通过 $\delta'_a(u)$进行扰动，因为 g 被限制为非负）。假设 g 在某一点 a处为正：$g(a) > 0$，且其右侧某一点 b，即 $b > a$，位于一个 $g \equiv 0$的区间内。接下来考虑一个扰动 $g \mapsto g + \varepsilon k$，其中 k 逼近 $-\delta_a + \delta_b$。方程（7）变为

$$0 = (b-a)\int_0^1 xg(x)dx + \int_a^b (-x^2 + a^2)g(x)dx + (a^2 - b^2)\int_b^1 g(y)dy.$$

将右端视为 b 的函数。这里，中间的积分与 b 无关，因为在 b 附近 g 为零；但此时我们有

$$0 = 2c(b-a) + c_2 + (a^2 - b^2)(G(1) - G(b))$$

这意味着 $G(b)$ 在 b 附近是关于 b 的非常数函数，这与 g 在 b 附近为零这一事实相矛盾。因此，g 不可能在这样的区间上为零。

然而，$g \equiv 0$ 在其左侧没有正的 g 值的区间上是可能的，也就是说，在包含 0 的区间上是可能的。根据对称性（交换模式中的 0 和 1），若对于 $x \in [0, b]$，$g(x) = 0$，那么 g 在 1 处也可以有一个点质量 $b'\delta_1$，其质量为 b'。这使我们考虑形如以下的 g

$$g(x) = \begin{cases} 0 & x < b \\ \frac{c}{x^2} & b < x < 1 \\ b'\delta_1 & x = 1. \end{cases}$$

条件 $\int_0^1 g(x)dx = 1$ 和 $2c = \int_0^1 xg(x)\,dx$ 确定了 b', c 作为 b的函数。然后对 b 取最大值，得到 $b = b' = c = e^{-1}$。对于这些值，我们得到 $C = \frac{12}{e^2}$。解出 f 即可得到陈述中的公式。 □

5 变分原理

现在我们令 n 变大，并考虑 $\{0,1\}^n$ 中具有某些模式 $w_1, \ldots, w_k$的固定密度 ρ_{w_i} 的序列。这样的序列通常是什么样的?

我们可以通过变分原理研究当 $n \to \infty$时这类受约束的随机序列的极限：在满足这些密度约束的条件下，最大化次勒贝格测度的熵。对于次勒贝格测度 $\mu = f(x)\,dx$，我们将熵 $\mathrm{Ent}(f)$ 定义为

$$\mathrm{Ent}(\mu) = \int_0^1 S(f(x))\,dx$$

（其中 $S(p) = -p\log p - (1-p)\log(1-p)$ 是香农熵函数）。熵 $\mathrm{Ent}(f)$ 是密度接近 f 的序列的指数增长率，

其含义如下（见 [11]）。设 $Z_n = \{X \in \{0,1\}^n \mid d_W(\mu_X,\ \mu) \le \varepsilon\}$。那么

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log|Z_n| = \int_0^1 S(f(x))\,dx.$$

对于一组特定的约束条件 $\rho_{w_1} = \delta_1, \ldots, \rho_{w_k} = \delta_k$，可以通过最大化得到极限典型密度

$$\mathcal{E}(f) = \mathrm{Ent}(f) + a_1\rho_{w_1} + \cdots + a_k\rho_{w_k}$$

其中，a_i 是由约束条件确定的拉格朗日乘子。我们将一个优化解称为极限形状。目前我们无法保证优化解的唯一性，尽管我们不知道有任何优化解不唯一的情况。

6 0、10、110、..... 1^k0 模式

首先，让我们求出具有固定模式密度 ρ_0, ρ_{10} 和 ρ_{110} 的最优次勒贝格测度（极限形状）。之后我们再讨论一般情形。使用变量替换 $y = F(x)$ $x = H(y)$，其中 $dx = H'(y)dy$，可得

$$\begin{aligned}\rho_{110} &= 6\int_{0\le x<u<v\le 1} f(x)f(u)(1-f(v))\,dx\,du\,dv\\ &= 6\int_{0\le a<b<y\le\rho_1}(H'(y)-1)da\,db\,dy\\ &= 3\int_0^{\rho_1} y^2(H'(y)-1)\,dy.\end{aligned}$$

对于固定的 ρ_0, ρ_{10} 和 ρ_{110}，要最大化 f，其变分方程为

$$\max_f\left\{\int_0^1 S(f(x))\,dx + A\rho_0 + B\rho_{10} + C\rho_{110}\right\}$$

where A, B, C 是拉格朗日乘子。使用上述变量替换，可得

$$\max_H\left\{\int_0^{\rho_1} S(1/H'(y))H'(y)\,dy + \int_0^{\rho_1}(A + 2By + 3Cy^2)(H'(y)-1)\,dy\right\}.$$

令 $h(y) = H'(y) - 1$。将 $h(y)$ 替换为 $h(y) + \varepsilon\delta'_{y=z}$，并对 ε 求导，取值点为 $\varepsilon = 0$，得到

$$0 = \int_0^{\rho_1} -\log\frac{h(y)}{h(y)+1}\delta'_{y=z}\,dy + \int_0^{\rho_1}(A + 2By + 3Cy^2)\delta'_{y=z}\,dy$$

$$0 = -\frac{h'(z)}{h(z)^2 + h(z)} + 2B + 6Cz.$$

其解为

$$H'(y) = h(y) + 1 = \frac{1}{1 - e^{a+by+cy^2}} \tag{8}$$

对于常数 a，b，c。其中 $b=2B$，$c=3C$取决于 B，C，而且必须选择 a，使得 $H(\rho_1)=1$，具体如下。对于固定的 ρ_1，如果 $a+by+cy^2$ 在 $[0,\rho_1]$中有零点，则积分 $\int_0^{\rho_1} H'(y)dy$ 没有定义；当 $a+by+cy^2>0$时，对于 $y\in[0,\rho_1]$，被积函数为负。因此，我们需要选择 $a<0$，使得对于 $y\in[0,\rho_1]$，$a+by+cy^2<0$。此外，在这一约束下，积分是 a的单调函数，因此存在唯一的 $a<0$，使得 $1=H(\rho_1)=\int_0^{\rho_1} H'(y)dy$。

反过来，给定 a，b，$c\in\mathbb{R}$且满足 $a<0$，我们可以将 ρ_1定义为使 $\int_0^{\rho_1} H'(y)dy=1$成立的第一个正值；由于 $H'(y)>1$，必有 $0<\rho_1<1$。于是，$\rho_0=1-\rho_1$，而 ρ_{10}，ρ_{110}由以下积分定义

$$\rho_{10}=2\int_0^{\rho_1} y(H'(y)-1)\,dy,\qquad \rho_{110}=6\int_0^{\rho_1} y^2(H'(y)-1)\,dy.$$

因此我们看到，极限形状密度具有 (8) 的形式，其中 a，b，c是拉格朗日乘子 A，B，C的函数。此外，从满足 $a<0$的三元组 $(a,\ b,\ c)$ 到三元组 $(A,\ B,\ C)$ 的映射是一个同胚；见下文。

上述相同的计算适用于任意一组模式 ρ_0，ρ_{10}，ρ_{110}、……、ρ_{1^k0}：

定理 6.1。对于一组模式密度 $(\rho_0,\ \rho_{10},\ \rho_{110},\ \ldots\ldots,\ \rho_{1^k0})=(\delta_0,\ \ldots\ldots,\ \delta_k)$，所得的熵最大化函数 $H'(y)$ 具有以下形式

$$H'(y)=\frac{1}{1-e^{p(y)}} \tag{9}$$

对于一个 p次数为 k 、具有实系数且负常数项的多项式。从这类多项式到 E_W（即可行域）内部的映射 $\Phi:\ \mathbb{R}_-\times\mathbb{R}^{k-1}\to E_W$ 是一个同胚。E_W 边界上的密度是取值为 $\{0,1\}$ 的阶梯函数，在至多 $k/2+1$ 个区间上取值为 0 （若该数等于 $k/2+1$，则首尾两段都是取值为 0 的区间）。

例如，假设我们只想固定 ρ_1 和 ρ_{10}。那么我们需要考虑形如以下形式的H

$$H'(y)=\frac{1}{1-e^{a+by}}$$

对于常数 $a<0$ 和 $b\in\mathbb{R}$。这得到了形如 $f(x)=\frac{1}{1+ce^{bx}}$ 的密度 f，其中~~$c=e^a/(1-e^a)$~~，从而重现了 [11] 的结果。E_0,10 的上边界 $\rho_{10}=2\rho_0(1-\rho_0)$ 由阶梯函数密度 $f(x)=\begin{cases}1 & x<\rho_1\\ 0 & \text{else}\end{cases}$ 构成，它们由极限 $a+bx=b(x-\rho_1)$ 在 $b\to\infty$ 时得到；下边界则由阶梯函数密度 $f(x)=\begin{cases}0 & x<\rho_0\\ 1 & \text{else}\end{cases}$’ 构成，它们由极限 $a+bx=b(x-\rho_0)$ 在 $b\to-\infty$ 时得到。另请参见图 6。

证明。事实是 $H'(y)$ 具有 (9) 的形式，这一点可由陈述前的说明得出。令 Φ 为从多项式 p (次数为 $\le k$，且具有负常数项) 到 $(\rho_0,\ldots,\rho_k)$ 的映射。剩下只需证明 Φ 是一个真局部同胚。

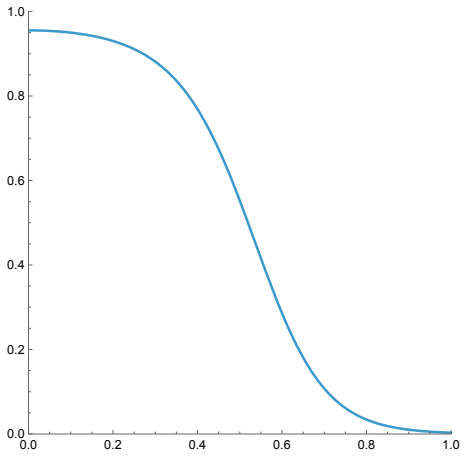


图 6：固定 $\rho_1 = \frac{1}{2}$和 $\rho_{110} = \frac{1}{3}$时的熵最大化密度 $f(x)$。在此情形下，$H'(y) = \frac{1}{1-e^{-a+by^2}}$，其中 $a \approx 3.10795$和 $b \approx 12.42$。

为证明 Φ 是局部同胚，我们证明 Φ 的雅可比矩阵非零：$J(\Phi) \neq 0$。为简化记号，令 $\rho_i := \rho_{1^i 0}$且 $\rho = \rho_{1^0}$。令 $p(y) = a_0 + a_1 y + \cdots + a_k y^k$。首先，对方程 $1 = \int_0^\rho \frac{1}{1-e^{p(y)}}\, dy$ 关于 a_j 求导，得到

$$\frac{1}{1-e^{p(\rho)}} \frac{d\rho}{da_j} = -\int_0^\rho \frac{y^j e^{p(y)}}{(1-e^{p(y)})^2}. \tag{10}$$

现在我们来计算

$$\begin{aligned} \frac{d\rho_i}{da_j} &= \frac{d}{da_j} \int_0^\rho \frac{y^i e^{p(y)}}{1-e^{p(y)}}\, dy \\ &= \int_0^\rho \frac{y^{i+j} e^{p(y)}}{(1-e^{p(y)})^2}\, dy + \frac{\rho^i e^{p(\rho)}}{1-e^{p(\rho)}} \frac{d\rho}{da_j} \\ &= \int_0^\rho \frac{y^j (y^i - \rho^i e^{p(\rho)}) e^{p(y)}}{(1-e^{p(y)})^2}\, dy. \end{aligned} \tag{11}$$

Φ 的微分为

$$D\Phi = \left(\frac{d\rho_i}{da_j} \right)_{i,j=0,\ldots,k}.$$

利用行列式对各列的线性性质和 (11)，我们可以写出

$$J\Phi = \int_0^\rho \cdots \int_0^\rho d\mu(y_0) \ldots d\mu(y_k) \det \begin{vmatrix} 1-c & y_1(1-c) & \ldots & y_k^k(1-c) \\ y_0 - \rho c & y_1(y_1 - \rho c) & \ldots & y_k^k(y_k - \rho c) \\ \ldots & & & \\ y_0^k - \rho^k c & y_1(y_1^k - \rho^k c) & \ldots & y_k^k(y_k^k - \rho^k c) \end{vmatrix}.$$

其中 $c = e^{p(\rho)}$ 和 $d\mu(y) = \frac{e^{p(y)}}{(1-e^{p(y)})_2}\, dy$. 将 $1-c$ 从第 1 行提出，将 y_i^i 从第 $i+1$ 列提出，并进行行变换后，得到一个范德蒙德矩阵，因此

$$J\Phi = (1-c)\int\cdots\int d\mu(y_0)\dots d\mu(y_k) y_1 y_2^2 \dots y_k^k \prod_{i<j}(y_j - y_i).$$

根据引理6.2 下文，$J\Phi > 0$。这证明了 Φ 是一个局部同胚。

现在我们证明 Φ 是适当的：当 $a_0 \to 0$ 或任意 $a_i \to \pm\infty$ 时，相关密度趋向于 E_W 边界上的一个点。首先，如果 a_0 趋于零，那么 $\rho_1 \to 0$。其次，如果任意系数 a_i 趋于 $\pm\infty$ （同时 p 在 $[0,\rho_1]$ 上仍为负），那么 p 在至多 k 个点以外的所有地方都趋于 $-\infty$：拉格朗日插值可根据任意 $k+1$ 个点处的取值唯一确定一个次数为 k 的多项式。事实上，p 保持有界的点至多有 $k/2+1$ 个，因为这些点必然是 p 的局部极大值，或是端点 $p(0)$ 或 $p(\rho_1)$。在这种极限下，H' 的支撑至多包含 $k/2+1$ 个点，并且如果相应的 p 值趋于零，那么这些点处可能会出现一个 δ 函数。这些 δ 函数对应于极限 f取值为 0 的区间；在这些区间之外，$H' = 1$，因此极限 f 必然取值为 1。

□

引理 6.2. 对于区间 $[a,b]$ 上的任意测度 μ，若 μ 的支集不少于 $k+1$ 个点，则有

$$\int_a^b\cdots\int_a^b y_1 y_2^2 \dots y_k^k \prod_{i<j}(y_j - y_i) d\mu(y_0)\dots d\mu(y_k) > 0.$$

证明。对 $y_0,\dots,y_k$ 的所有排列求和，得到

$$\int_a^b\cdots\int_a^b d\mu(y_0)\dots d\mu(y_k)\left(\sum_\sigma (-1)^\sigma y^0_{\sigma(0)} y^1_{\sigma(1)}\cdots y^k_{\sigma(k)}\right)\prod_{i<j}(y_j - y_i).$$

括号中的因子是 $\prod_{i<j}(y_j - y_i)$ 的另一个副本，因此得到

$$\int_a^b\cdots\int_a^b d\mu(y_0)\dots d\mu(y_k)\prod_{i<j}(y_j - y_i)^2$$

这显然是正的。 □

7 上三角矩阵

整数海森堡群 H 是由上三角 3×3 矩阵构成的群

$$M = \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix}$$

其元素为 $a,\ b,\ c \in \mathbb{Z}$. 它由以下矩阵生成

$$M_0 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

给定一个二进制序列 X，都有一个与之对应的矩阵 M_X，它是由生成元构成的一个正字，例如 $M_{01101} = M_0M_1M_1M_0M_1 = \begin{pmatrix} 1 & 2 & 4 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}$. 。 M_X 的超对角线元素分别表示 X中 0 和 1 的个数。右上角元素表示 X中 01 模式的个数。

更一般地，我们可以取一个由更大的矩阵生成的广义海森堡群，其中 M_0 的某些超对角线元素为 1，而 M_1 的其余超对角线元素为 1。例如，若我们取

$$M_0 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

那么在这种情况下

$$M_X = \begin{pmatrix} 1 & N_0 & N_{01} & N_{010} \\ 0 & 1 & N_1 & N_{10} \\ 0 & 0 & 1 & N_0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{12}$$

其中 $N_\tau = N_\tau(X)$。一般来说，如果 M_1 的超对角线元素为 $a_1,\ a_2$、.....、其中 $a_i \in \{0,1\}$，那么 M_X 第一行的元素为 1，$N_{a_1},\ N_{a_1a_2},\ N_{a_1a_2a_3}$，.....。因此，任意模式计数函数 $N_w(X)$ 都可以作为某个此类矩阵乘积的右上角元素出现。

针对一个模式的最优模式计数问题 w 可以表述为：在相应的海森堡群中，求从单位元到右上角元素为 N_w的矩阵子集的最短测地线集合。（关于域 $\mathbb{F}_p$上的相关结果，见 [3]。）

矩阵 M_X 是全非负的：所有子式均非负。通过 LGV 引理可以看出这一点 [9, 4]，见图7。

这种正性对一般的模式集合 W的 E_W 施加了某些限制。不过目前，我们还无法有效地利用这一思路来计算有趣集合 W的E_W 。

8 开放问题

1. 如果我们考虑的是一个含有 k 个字母的字母表，而不是二进制序列，会出现什么新现象吗?

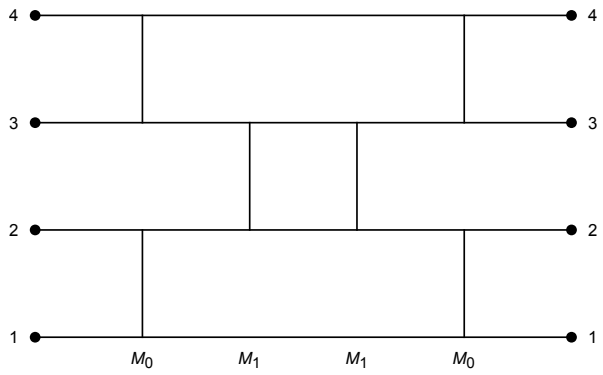


图7： M_{0110}的一个子式的行列式所对应的不相交路径模型。注意，从左侧标记为1的顶点到右侧标记为2、3、4的顶点的东北路径数量分别为 N_0、 N_{01}、 N_{010}。从左侧顶点2到右侧顶点2、3、4的数量分别为1、 N_1、 N_{10}，而从顶点3出发的数量为0、1、 N_0。因此，当 $X = 0110$时，从左侧的1、2、3到右侧的2、3、4的三条不相交东北路径的数量由(12)的右上 $3X3$ 子式给出。

2. 熵在可行域 E_W 的内部是解析函数吗？熵最大化子勒贝格测度在 E_W 的内部总是唯一的吗？3. 对任意 W，可行域 E_W 总是半代数集吗？

9 附录

这里概述表中提到的 C_τ 值的推导过程。

9.1 $1^k0^l1^m$

对于 $\tau = 1^k0^l1^m$，在固定 N_1的情况下，为使 N_τ 最大，最优做法是让所有0连续出现；因此，极限密度 f 必然是如下形式的阶梯函数

$$f(x) = \begin{cases} 1 & x < a \\ 0 & a < x < b \\ 1 & b < x < 1 \end{cases}$$

对于某些满足 $b - a = 1 - \rho_1$ 的 a。密度 ρ_τ 随后为

$$\rho_\tau = (k + l + m)! \frac{a^k}{k!} \frac{(1 - \rho_1)^l}{l!} \frac{(\rho_1 - a)^m}{m!}.$$

当 $a = \frac{k\rho_1}{k+m}$ 时，该值达到最大~~。~~代入即可得到结果。

9.2 11010

对于 $\tau = 11010$，使用定理 4.2 的证明方法可得

$$C_\tau := 5! \max_g \int_{0<x<y<1} x^2(y-x)g(x)g(y)\,dx\,dy.$$

与 1010 的情形一样，我们寻求 g 在区间 $[0,b]$ 上为 0，在 1 处有一个点质量，并且在两者之间解析。考虑扰动 $g \to g + \varepsilon\delta_a'$，得到欧拉-拉格朗日方程

$$0 = \int_0^a x^2 g(x)\,dx + \int_a^1 a(2y-3a)g(y)\,dy. \tag{13}$$

对 a 求三次导数，得到

$$2a^2 g''(a) + 12ag'(a) + 14g(a) = 0. \tag{14}$$

这给出了 g 的解析部分的形式：

$$g(a) = ca^{-5/2}\cos(\frac{\pi}{3} + \frac{\sqrt{3}}{2}\log a)$$

其中常数为 c （这里利用 二阶导数消去了一个积分常数）。其余计算与定理 4.2的证明类似。

9.3 10110

与 11010 的情况相同，只是欧拉-拉格朗日方程为

$$0 = 2\int_0^a x(a-x)g(x)\,dx + \int_a^1 (y-a)(y-3a)g(y)\,dy.$$

对 a 求 3 次导数得到 $g(a) = 0$。不过在这种情况下，我们可以且确实会在内部有一个点质量，因此我们预期 $g = c_1\delta_b + c_2\delta_1$。这意味着 f 是一个取值于 $\{0,1\}$ 的阶梯函数，共有 4 个阶跃。

9.4 10101

与 11010 的情况相同，只是欧拉-拉格朗日方程为

$$0 = 2\int_0^a x(1-2a+x)g(x)\,dx + \int_a^1 (1-y)(y-2a)g(y)\,dy.$$

对 a 求两次导数可得 $2a(1-a)g'(a) + (4-8a)g(a) = 0$，其解为 $g(a) = \frac{c}{a^2(1-a)^2}$，其中 c 为常数。与 ~~1010 的情~~形一样，我们也可以预期在区间 $[0,b]$ 上 g 为零，并且对称地有 $[1-b,1]$。因此我们得到

$$g(x) = \begin{cases} 0 & 0 < x < b \\ \frac{c}{x^2(1-x)^2} & b < x < 1-b \\ 0 & 1-b < x < 1. \end{cases}$$

其中 $c = c(b)$ 由 $\int_0^1 g(x)dx = 1$ 这一事实确定。对 b 积分并进行优化，可得到 $b = \xi/(1+\xi)$ 以及 C_τ 的给定值。

参考资料

[1] Sourav Chatterjee 和 Persi Diaconis。指数随机图模型的估计与理解。《统计年刊》，41(5):2428–2461，2013。[2] L. N. Coregliano 和 A. A. Razborov。稠密组合对象的语义极限。《俄罗斯数学评论》，75(4):627，2020年8月。[3] Persi Diaconis 和 Robert Hough。幺幂矩阵群上的随机游走。*Ann. Sci. ´Ec. Norm. Sup´er. (4)*，54(3):587–625，2021。[4] Ira Gessel 和 G´erard Viennot。二项式行列式、路径与钩长公式。*Adv. in Math.*，58(3):300–321，1985。[5] Felix Hausdorff。求和方法与矩序列 I。《数学杂志》，9(1-2):74–109，1921。[6] Carlos Hoppen、Yoshiharu Kohayakawa、Carlos Gustavo Moreira、Bal´azs R´ath 和 Rudini Menezes Sampaio。排列序列的极限。*Journal of Com-binatorial Theory, Series B*，103(1):93–113，2013。[7] Richard Kenyon、Daniel Kr´aˇl、Charles Radin 和 Peter Winkler。具有固定模式密度的排列。*Random Structures Algorithms*，56(1):220–250，2020。[8] Richard Kenyon、Charles Radin、Kui Ren 和 Lorenzo Sadun。具有边和三角形约束的大型网络的相。《物理学杂志*A*辑》，50(43):435001，22，2017。[9] Bernt Lindstr¨om。论诱导拟阵的向量表示。《伦敦数学会通报》，5:85–90，1973。

[10] L´aszl´oLov´asz.《大型网络和图极限》，美国数学会《学术丛书》系列第*60*卷，美国数学会，罗得岛州普罗维登斯，2012年。

[11] A. M. Vershik。组合分拆的统计力学及其极限构型。《泛函分析及其应用》，30(2):19–39，96，1996年。