arXiv:2603.20396v1 [cs.AI] 20 Mar 2026

压缩就是你所需要的一切：数学建模

VITALY AKSENOV、EVE BODNIA、MICHAEL H. FREEDMAN 和 MICHAEL MULLIGAN

摘要。人类数学（HM）是人类发现并珍视的数学，它只是形式数学（FM）——所有有效推演的总和——中微不足道的一小部分。我们认为，人类数学的独特之处在于，它可以通过层层嵌套的定义、引理和定理实现压缩。我们用幺半群对此进行建模。数学推演是由原始符号组成的字符串；定义或定理则是一个有名称的子串或宏，其使用可以压缩该字符串。在自由阿贝尔幺半群 A_n 中，一个对数稀疏的宏集合就能使表达力实现指数扩张。在自由非阿贝尔幺半群 F_n 中，即使宏集合的密度达到多项式级，也只能带来线性扩张；要实现超线性扩张，则需要接近最大密度。我们用 MathLib 检验这些模型。MathLib 是一个大型的 Lean 4 数学库，我们将其视为人类数学的替代样本。每个元素都有一个深度（定义嵌套的层数）、一个包装长度（其定义中的词元数量），以及一个展开长度（完全展开所有引用后得到的原始符号数量）。我们发现，展开长度会随着深度和包装长度呈指数增长；而包装长度在所有深度下大致保持恒定。这些结果与 A_n 相符，与 F_n不符，支持这样一种论点：在人类数学所在的形式数学空间中，人类数学所占的部分仅以多项式速度增长，而该空间本身则以指数速度增长。我们还探讨了如何利用在 MathLib 依赖图上测得的压缩率，以及对该图进行的 PageRank 式分析，来量化数学兴趣，并帮助引导自动推理关注人类数学所在的可压缩区域。

1. 引言

本文认为，数学是柔软而有弹性的——这正是它的定义性特征。我们所说的“数学”，并非所有可能的形式演绎，即形式数学（FM）的总和，而是人类数学（HM）：人类会提出、并且当人工智能智能体为我们找到论证时我们会欣赏的那类论证。[1] 所谓“柔软而有弹性”，是指它能够通过层级嵌套的概念——定义、引理和定理——进行压缩。

数学关乎压缩，这一发现并不新鲜。早在 3,000 年前，位值记数法的发明就抢先于我们发现了这一点。考虑 $\mathbb{N}$，即自然数集，其生成集为 {1}。位值记数法引入了额外的符号，也就是宏：“10”表示十个一，“100”表示十个十，以此

日期：2026年3月24日。[1]无论数学如何形式化，我们从哥德尔及其他来源得知，总会存在没有证明的真命题（例如系统的一致性）。皮亚诺算术中极其简单的 Π^0_1 命题，例如哥德巴赫猜想（GC：“每个偶数 > 2 都是两个素数之和”），可能既为真又无法证明。由于我们的讨论以证明这一概念为基础，若情况确实如此，我们就不会将 GC 视为人类数学甚至形式数学的一部分，尽管它令世人感兴趣。还有一个更微妙的问题：假设 GC 有一个证明，但最短的证明长达 10^{100} 行——它算是人类数学的一部分吗？幸好我们不必对此作出裁定；我们的结果还不够精确，无须确定人类数学的确切前沿。此外，更准确地说，人类数学或许可以被视为形式数学上的一种度量，它会逐渐淡出，而非在某个子集边缘突然终止；参见 [BDF]。1

类推。借助对数数量的宏，表达力便能呈指数级扩张。表达力的这种（指数级）扩张，正是记号压缩的另一面。通过创建并运用定义，我们可以在一种原始的、缺乏定义的语言中压缩表达式，从而拓展可达范围。（下文的定理将以扩张的形式表述；非正式讨论中，我们通常会使用压缩这一说法。）

形式数学可被视为一个由公理和句法规则产生的有向超图（DH） [BDF]。DH 记录了完整的演绎空间：每一种可能的证明步骤，其中每条超边都规定了哪些前提被组合起来（图 1，左）。一个证明是 DH 的一个子超图；将其平坦化为线性序列后，它便成为一串原始符号。我们研究有限生成幺半群，将其作为这类字符串的模型：词长衡量大小，而为某个子串命名以便重复使用——即宏——则可对其进行压缩。（带有关系的幺半群能够模拟图灵机 [Pos47], ，因此，尽管这个基本框架很简单，它仍是计算通用的。）最简单的情形是 $A_1 = \mathbb{N}$，即自然数集。为了更一般地研究压缩，我们考虑自由阿贝尔幺半群 A_n 以及自由（非阿贝尔）幺半群 F_n （其中 n 表示生成元的数量）。在 A_n中，生成元可交换，因此只有重数重要。在 F_n中，生成元的顺序很重要，并且不存在关系；由于形式证明是符号串，且顺序至关重要，人们或许会认为 F_n 能够为形式演绎建模。我们将论证，与这种预期相反，人类数学所展现的压缩是 A_n的特征，而非 F_n。

我们的主要理论结果量化了宏在 A_n 和 F_n中的扩张能力。在A_n中，对数数量的宏即可实现指数扩张（定理 1）；而通过华林定理，具有多项式密度（增长指数为 $1/k$）的宏能够实现无限扩张——即每个元素都能以有界长度表示（定理 3）。在 F_n 中，即使宏以多项式速度增长（相对于半径为多项式，相对于体积则为多对数），也只能实现线性扩张（定理 4）。在 F_n 中，超线性扩张需要指数数量的宏（定理 5）；与之相反，在 A_n中，实现指数扩张只需一个对数规模的稀疏宏集。这种差异反映了底层增长率的不同：在 A_n 中，球的增长是多项式级的，而在 F_n中则是指数级的。我们对 A_n 中宏集的研究可以直接扩展到规模大得多的自由幂零幺半群；它们的扩张性质与 A_n 几乎完全相同，根据本文的分析，同样可以作为人类数学（HM）的模型（详情见第 4.3 节）。

我们用 MathLib [Mat20], 这一以 Lean 4 编写的大型数学库来检验这些事实[dMU21] ，其中包含数十万个定义、引理和定理。我们将 MathLib 作为人类数学（HM）的代理。MathLib 可以看作是从完整的演绎超图中提取出的一个有向无环图（DAG）（图 1）。MathLib 的每个元素都是该有向无环图中的一个具名子图，以该元素本身为根，并向下延伸至原始元素。递归展开所有引用，将这个子图扁平化后，会得到一个由原始元素组成的字符串。包装长度统计元素中的词元数量，其

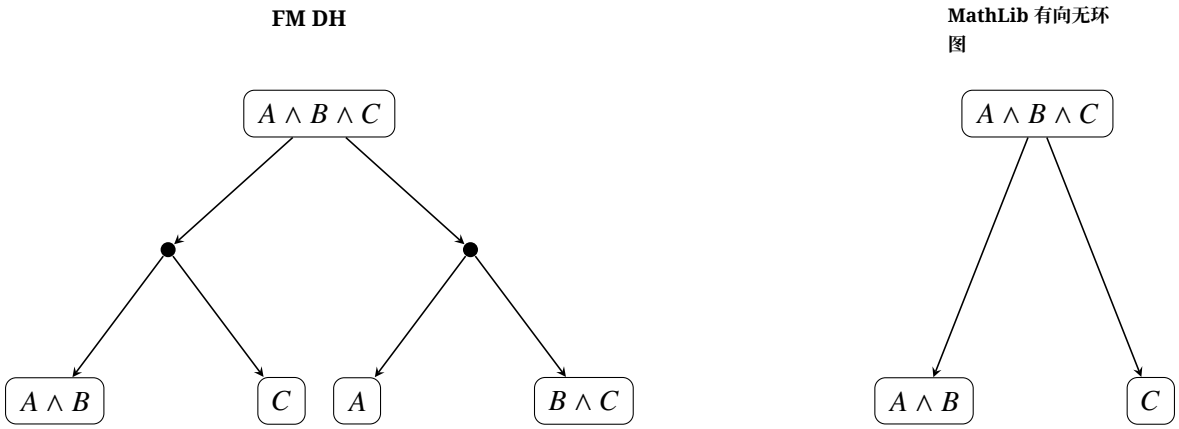


图1. 一个 FM DH 片段（左）以及对应的 MathLib 有向无环图（右），用于通过 $\wedge$-引入推导 $A \wedge B \wedge C$。在 DH 中，实心点表示超边：每条超边将单个推理步骤中使用的前提归为一组。两种证明（通过 $A \wedge B$ 和通过 $B \wedge C$）均有记录。有向无环图选取其中一种证明，并将超边替换为普通边。如果 DH 中的两种证明都记录在有向无环图中，那么给定结论如何由前提推出就会产生歧义。还要注意，上述例子比较特殊，因为超边是可逆的。

定义表达式；展开长度 则统计扁平化字符串中的原始元素数量。我们发现，完全展开后，最长元素约包含 10^{104} 个原始元素——古戈尔（这个数字，而非那家公司）。[2] 我们对 MathLib 的主要观察如下。首先，展开长度随深度（依赖图中到原始元素的最长路径）呈指数增长。其次，包装长度在所有深度上近乎恒定。第三，展开长度随包装长度呈指数增长。MathLib 数据符合 A_n对数密度机制，与我们研究的 F_n 其他机制不符。

我们的核心推断是：在人类数学所处的指数增长空间 FM 中，人类数学是一个增长速度为多项式的稀疏子集。与“人类数学只是 FM 中微不足道的一小部分”这一观察相比，这是一个更强的论断：如果二者都以不同速率指数增长，后一说法同样成立 [Sha48,CT06]。我们提出以 A_n 作为人类数学的模型，以乘积 $\prod F_{n_i}$ 作为 FM 的模型。由于人类数学 $\subset$ FM，这些模型必须体现这种包含关系； A_n 通过将每个生成元映射到一个不同的因子，嵌入 $\prod F_{n_i}$。

我们的玩具模型将人类数学映射到幺半群。在源端和目标端，我们都能清楚地观察到压缩和层次深度：在幺半群中，两者都是从假定的宏集合中推导出来的；在 MathLib 中，两者都可以测量。通过比较，我们得以推断人类数学中一个假想“宏集合”的性质。物理模型常通过定义无法直接观测到的抽象概念来增强解释力——例如电磁学中的矢势、量子力学中的希尔伯特空间。类似地，我们的模型可能不是满射：幺半群中的某些抽象概念——尤其是宏集合本身——在人类数学一侧可能没有直接对应物。我们并未确定一个确切的“宏集合”

[2] MathLib 中的 AlgebraicGeometry.Scheme.exists_hom_hom_comp_eq_comp_of_locallyOfFiniteType；讨论请参见第 3 节。

在 MathLib（或更广义的人类数学 HM）中，存在一个映射到幺半群中宏集合的映射，但我们将此视为一个深奥的开放问题，难度相当于定位人类数学 HM 的使用手册。

位值记数法展现了我们的模型能够捕捉到的数学的两个特征。其一是层级结构：记号、思想、定义和证明以递归的方式相互组合。其二是简约性：我们用于存储新概念的空间有限，因此定义必须经过精心选择，在无限结构中标记出重要节点的同时，又不至于过度消耗我们的记忆能力。人类数学 HM 在可以进行压缩的地方大显身手——它契合我们的思维方式，也顺应了我们与生俱来的惰性，让我们只需付出极少努力，就能在数学领域中大步前进。10 的幂所体现的对数密度，接近这一简约性边界。第 2 节的结果系统地探讨了简约性与扩张之间的权衡，展示了在多个区间内，扩张速率如何取决于宏密度。第 3 节的 MathLib 数据从定量上证实了这种层级结构：对数（展开长度）随深度线性增长，斜率接近每层 1 比特。

如果压缩刻画了人类数学，它也可以用作衡量数学兴趣的指标。简洁的陈述却隐藏着庞大证明主体的元素，体现出较高的演绎压缩；应用定义后压缩幅度显著的元素，则处于定义层级发挥作用的区域。我们将后者称为归约压缩。第 5 节将这些想法发展为定量兴趣度量，并进一步采用类似 PageRank 的改进方法[BP98]，以衡量某个元素在支撑其他高价值数学内容方面所起的作用。我们的目标是让探索形式数学的人工智能智能体获得方向感：留在可以实现压缩的地方。

我们与各种大语言模型合作，完成了第 2 节中定理的证明。下方的符号表示，✅定理已由 Logical Intelligence 开发的定理证明系统 Aleph [Log25]，在 Lean4 中通过形式化验证。

本文其余部分安排如下。第 2 节构建幺半群模型，并证明主要扩张定理。第 3 节介绍 MathLib 分析。第 4 节进一步讨论结果及相关思想。第 5 节探讨自动化数学兴趣度量的未来工作及相关开放问题。附录 A 包含关于 A_n 的补充扩张定理。

2. 幺半群模型

我们研究由 n 个生成元 $G = \{a_1、\ldots\ldots、a_n\}$生成的两个基本幺半群：自由阿贝尔幺半群 A_n和自由幺半群 F_n。在 A_n中，生成元可交换，因此其元素本质上属于 $\mathbb{N}^n$，并采用逐分量加法。在 F_n中，顺序很重要，且没有任何关系；其元素是由 G构成的有限字符串。对于任一幺半群中的元素或词 w，用 $|w|_G$表示其长度：对于 A_n，即系数之和；对于 F_n，即字符串长度。

宏集合 $\{\}$由附加生成元组成，每个生成元都由 $g_i = w_i$ 定义为w_i 一个由 G 中元素构成的词。增广生成集为 $G' = G \cup M$，而 $|w|_{G'}$

表示表示 w 所需的最少 G'生成元数量。从概念上讲，尽管每个$g_i \in M$ 都是一个独立的宏，但集合 M 本身代表一种压缩策略。[3]

我们用扩张函数来量化这种策略的有效性，

$$f_{G'}(s) = \sup\{r \in \mathbb{N} : B_G(r) \subseteq B_{G'}(s)\}.$$

这里，半径为 r的球是 $B_G(r) = \{w : |w|_G \le r\}$，其中 $B_{G'}(s)$的定义类似。由于$G \subseteq G'$，我们有 $|w|_{G'} \le |w|_G$，因此 $B_G(s) \subseteq B_{G'}(s)$。扩张函数度量的是半径为 s的 G'球能够完全覆盖的最大 G半径。

我们的主要结果总结于表1。具体而言，表中给出了$A_1 = \mathbb{N}$的 A_n 结果；将每个宏复制 n 份（每个生成元对应一份），即可推广到一般的 A_n，且渐近扩张率保持不变。在 A_n中，球以多项式速度增长（$|B_G(r)| = \binom{r+n}{n}$），稀疏宏则会带来显著扩张——指数级，甚至无限扩张。多对数行反映的是一个上界（定理 2）；我们没有证明相匹配的下界，因此这类宏的实际扩张可能严格介于指数级与准指数级之间。在 F_n中，球呈指数增长（$|B_G(r)| = \binom{n^{r+1}-1}{n-1}$），对于多项式稠密宏集，扩张是线性的；对于指数增长宏集，扩张则是超线性的。

幺半群	宏 M	密度	扩张 $f_{G'}(s)$	定理
A_1	$\{m^k : m \ge 1\}$	$r^{1/k}$	∞	3
A_1	$\{b^{j^p} : j \ge 1\}$	(对数 $r)^{1/p}$	$\le e^{c \cdot s \log s}$	2
A_1	$\{b^j : j \ge 1\}$	对数 r	$\Theta(b^{cs})$	1
A_1	$\{b^{b^j} : j \ge 0\}$	$\log_b \log r$	$s^{b/(b-1)}$ 到 $s^{(2b-1)/(b-1)}$	6
A_1	有限	$O(1)$	$\Theta(s)$	7
F_n	多项式	r^p	$O(s)$	4
F_n	概率	n^r/对数 r	$\ge e^{c\sqrt{s}}$	5

表1。宏密度与扩张的关系。对于 A_n 满足 $n > 1$的情形，扩张率保持不变。A_n 的扩张性质对自由幂零幺半群同样成立。F_n 多项式结果适用于任意多项式密度和 $n \ge 2$；在常数因子意义下，它适用于任何具有指数增长的有限呈示幺半群。F_n 概率结果给出了一个对数级趋零密度的宏集合（$|M \cap S_r|/|S_r| \sim 1/\log(r) \to 0$），其中 $S_r = \{w : |w|_G = r\}$。

现在我们继续介绍刻画 A_n 和F_n中不同宏集扩张性质的定理。后文不需要这些定理的详细证明。

2.1. 自由阿贝尔幺半群。位值记数法是 A_n 中压缩的典型范例。

[3]在逻辑和计算机科学中，“宏”通常指变换规则（即策略）本身。这里，我们沿用幺半群论惯例，将宏视为所得元素，而集合 M 构成该策略。

定理 1 (位值记数法带来指数扩张)。 对于 A_n 以及任意整数 $b \geq 2$，宏集合 $M = \{b^j a_i : i = 1, \ldots, n,\ j \geq 1\}$ 具有对数密度并满足

$$b^{s/(n(b-1))-1} \leq f_{G'}(s) \leq nb \cdot b^{s/(n(b-1))}$$

对于所有整数 $s \geq 1$。特别地，$f_{G'}(s) = \Theta(b^{s/(n(b-1))})$。

证明。 宏集合 $M = \{g_{i,j} = b^j a_i : i = 1, \ldots, n,\ j \geq 1\}$ 具有对数密度：满足 $|g_{i,j}|_G = b^j \leq r$ 的宏的数量为 $n\lfloor \log_b r \rfloor = O(\log r)$。

下界。 任意元素 $w \in A_n$ 都可以唯一地写成 $w = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n$，其中 $x_i \in \mathbb{N}$。将每个非零 x_i 表示为以 b 为底的形式 $x_i = \sum_{j=0}^{J_i} c_{i,j} b^j$，其中 $c_{i,j} \in \{0, 1, \ldots, b-1\}$，并且 $J_i = \lfloor \log_b x_i \rfloor$，我们有

$$x_i a_i = c_{i,0} a_i + c_{i,1} g_{i,1} + c_{i,2} g_{i,2} + \cdots + c_{i,J_i} g_{i,J_i}.$$

G' 长度为 $x_i a_i$ $\sum_{j=0}^{J_i} c_{i,j} \leq (b-1)(J_i + 1) = (b-1)(\lfloor \log_b x_i \rfloor + 1)$。对于 $w \in B_G(r)$，我们有 $|w|_G = \sum_i x_i \leq r$，因此每个 $x_i \leq r$，于是

$$|w|_{G'} \leq \sum_{i=1}^{n} (b-1)(\lfloor \log_b x_i \rfloor + 1) \leq n(b-1)(\log_b r + 1).$$

因此，当 $s \geq n(b-1)(\log_b r + 1)$ 时，$B_G(r) \subseteq B_{G'}(s)$，由此得到 $f_{G'}(s) \geq b^{s/(n(b-1))-1}$。

上界。 我们给出一个难以压缩的元素。对于任意整数 $k \geq 1$，定义 $w_k = (b^k - 1)(a_1 + \cdots + a_n)$。于是 $|w_k|_G = n(b^k - 1)$。由于 $b^k - 1 = \sum_{j=0}^{k-1} (b-1) b^j$，$|w_k|_{G'} = n(b-1)k$。

现在给定 $s \geq 1$，选择 $k = \lfloor s/(n(b-1)) \rfloor + 1$。那么 $|w_k|_{G'} = n(b-1)k > s$，因此 $w_k \notin B_{G'}(s)$。由于 $w_k \in B_G(n(b^k - 1)) \subseteq B_G(nb^k)$，我们有 $B_G(nb^k) \not\subseteq B_{G'}(s)$，于是

$$f_{G'}(s) < nb^k \leq nb \cdot b^{s/(n(b-1))}.$$

合并这些界可得 $f_{G'}(s) = \Theta(b^{s/(n(b-1))})$。 □

接下来我们证明一个上界：使用多对数个宏时，扩张至多为准指数。

定理 2 （多对数密度带来准指数扩张）。 对于 A_n，令 $M \subseteq A_n$ 为一个具有多对数增长的宏集：

$$|M \cap B_G(r)| \leq c(\log(e + r))^q \quad \textit{for all } r \geq 0,$$

对于某些常数 $c, q > 0$。那么存在一个仅取决于 n, c, q 的常数 $K > 0$，使得 t

$$f_{G'}(s) \leq \exp(Ks \log s) \quad \textit{for all } s \geq 2.$$

证明。 固定 $s \in \mathbb{N}$，并假设 $B_G(r) \subseteq B_{G'}(s)$，即每个长度为 $\leq r$ 的元素都可以表示为至多 s 个来自 $G' = G \cup M$ 的生成元之和。我们将用 s 给出 r 的一个上界。

步骤 1：只有长度为 $\le r$ 的宏才相关。 设 $w \in B_G(r)$，并将其写作 $w = y_1 + \cdots + y_k$，其中 $k \le s$ 且 $y_i \in G'$。每个 y_i 的长度都是 $|y_i|_G \ge 0$，而 $\mathbb{N}^n$ 中的长度可加性给出 $|w|_G = |y_1|_G + \cdots + |y_k|_G$。由于 $|w|_G \le r$，所有 $|y_i|_G \le r$；否则它们的和将超过 r。因此，在 $B_G(r)$中元素的任何表示里，只能出现长度为 $\le r$ 的生成元。

定义 $M_r := M \cap B_G(r)$，使得 $|M_r| \le c(\log(e+r))^q$，并注意，在每个这样的表示中，每个 y_i 都属于 $G \cup M_r$。令

$$t(r) := |G \cup M_r| = n + |M_r| \le (n+c)(\log(e+r))^q.$$

步骤 2：长度为 $\le s$ 的单词数量上界。在大小为 $t(r)$ 的字母表上，长度为$\le s$ 的单词数量至多为

$$N_{\text{words}}(r,s) := \sum_{k=0}^{s} t(r)^k \le (s+1)(n+c)^s(\log(e+r))^{qs}.$$

每个这样的词都表示 A_n 的某个元素。根据假设 $B_G(r) \subseteq B_{G'}(s)$ 以及步骤 1，$B_G(r)$ 的每个元素都至少能由一个这样的词表示。因此 $|B_G(r)| \le N_{\text{words}}(r, s)$。

由于 $|B_G(r)| = \binom{r+n}{n} \ge \frac{r^n}{n!}$，我们有

$$\frac{r^n}{n!} \le (s+1)(n+c)^s(\log(e+r))^{qs}.$$

取对数，当 s 足够大时：

$$n \log r \le (1 + \log(n+c))s + qs \log\log(e+r). \tag{1}$$

步骤 3：界定 r。 我们证明，对于 $K > 2q/n$ 且 s 足够大时，只要 $\log r \ge Ks \log s$，(1) 就不成立。只需考虑 $\log r = Ks \log s$，因为更大的 r 只会使该不等式更加不成立。

对于较大的 s：

$$\log\log(e+r) \le \log 2 + \log(Ks\log s) \le \log(2K) + 2\log s.$$

代入 (1) 得：

$$nKs\log s \le (1 + \log(n+c) + q\log(2K))s + 2qs\log s.$$

除以 $s\log s$：

$$nK \le \frac{1 + \log(n+c) + q\log(2K)}{\log s} + 2q.$$

对于较大的 s右侧趋近于 $2q$，因此取 $K > 2q/n$ 可得矛盾 .

因此，对于所有充分大的 s，$f_{G'}(s) < \exp(Ks\log s)$。对于较小的 s，步骤 2 中的界表明 $f_{G'}(s)$ 是有限的（因为 r 中的多项式增长最终会超过 r 中的多对数增长），所以必要时增大 K，即可得到界 $f_{G'}(s) \le \exp(Ks\log s)$ 对所有 $s \ge 2$ 都成立。□

最后，我们证明多项式密度宏可以产生无限扩张。

定理 3 (多项式密度带来无限扩张)。 对于任意整数 $k \geq 2$，都存在一个宏集 $M \subseteq A_n$，使得：

$$|M \cap B_G(r)| \leq nr^{1/k} \quad \textit{for all } r \geq 1,$$

and

$$f_{G'}(s) = \infty \quad \textit{for all } s \geq ng(k),$$

其中 g（k）是华林常数（即使每个非负整数都能表示为至多 g（k） k 次幂之和的最小整数）。

Proof. 对每个生成元 $a_i \in G$ 和每个 $m \in \mathbb{N}$，定义宏 $g_{i,m} := m^k a_i$。 Let

$$M := \{m^k a_i : i = 1, \ldots, n,\ m \geq 1\}.$$

增长界。 宏 $m^k a_i$具有 G-长度 m^k。对于每个 i，满足$|g_{i,m}|_G = m^k \leq r$的宏数量为 $\lfloor r^{1/k} \rfloor$，因此

$$|M \cap B_G(r)| = n\lfloor r^{1/k} \rfloor \leq nr^{1/k}.$$

无限扩张。 任意元素 $w \in A_n$ 都可以写成 $w = x_1 a_1 + \cdots + x_n a_n$，其中 $x_i \in \mathbb{N}$。根据华林定理，每个 x_i 都是至多 $g(k)$ 个 k 次幂之和：

$$x_i = m_{i,1}^k + \cdots + m_{i,t_i}^k, \quad t_i \leq g(k).$$

Thus

$$x_i a_i = m_{i,1}^k a_i + \cdots + m_{i,t_i}^k a_i,$$

其中每一项$m_{i,j}^k a_i$都属于 M。对所有 i 求和，宏项总数至多为$\sum_{i=1}^n t_i \leq ng(k)$。因此， A_n 中的每个元素都属于 $B_{G'}(ng(k))$，从而得到 $f_{G'}(ng(k)) = \infty$。 □

注记 1。 对于 $k = 2$，拉格朗日四平方定理给出 $g(2) = 4$，因此增长指数为 1/2 的平方数宏集合可达到 $f_{G'}(4) = \infty$。

注记 2。 定理1和定理3中的宏集呈现出某种对偶性：定理 1 使用固定底数的幂 $\{b^j a_i : j \geq 1\}$，而定理 3 使用变化底数的固定幂 $\{m^k a_i : m \geq 1\}$。较稀疏的对数密度集合可实现指数扩张；多项式密度集合可实现无限扩张。

关于 A_n 的其他结果，包括双对数密度和有限宏密度的情形，见附录 A。

2.2. **自由幺半群**。 与 A_n 相比，我们现在证明，对于 F_n，多项式增长宏集合只能实现线性扩张，而超线性扩张则需要指数增长宏集。这反映了底层幺半群的指数增长。

定理 4（多项式密度带来线性扩张）。 对于满足 $n \geq 2$ 的 F_n，设 M 为一个宏集合，对某些常数 $c > 0$ 和 $p \geq 0$，每个 G-长度 $l \geq 2$ 至多包含 cl^p 个宏。

那么存在一个常数 $d = d(n, p, c)$，使得对所有整数 $s \geq 1$:

$$f_{G'}(s) < ds.$$

此外，只需选择一个满足以下条件的整数 $d \geq 3$:

$$n^d > 4e(n+c)d^{p+1}. \tag{2}$$

证明。固定整数 $r, s \geq 1$。考虑恰好具有 G长度 r 的词:

$$S_r := \{w \in F_n : |w|_G = r\}, \qquad |S_r| = n^r.$$

我们将证明，对于适当选择的 d, $|\{w \in S_{ds} : |w|_{G'} \leq s\}| < n^{ds}$，可推出$B_G(ds) \not\subseteq B_{G'}(s)$。

固定 $r = ds$和 $1 \leq k \leq s$。由于 F_n没有关系，任何表示 $w = y_1 \cdots y_k$with $y_i \in G'$ 和 $|w|_G = ds$ 都由以下内容确定:

(1) 将 ds拆分为 k个正部分（$l_1, \ldots, l_k$），其中 $l_i = |y_i|_G$(2) 对每个 i,
在 G' 中选择一个长度为 G的生成元 l_i

对于每个长度 $l \geq 1$, G' 中该长度的生成元至多有 $(n+c)l^p$ 个（当$l = 1$时恰好有 n 个，当 $l > 1$时至多有 cl^p 个）。因此:

$$|\{(y_1, \ldots, y_k) : y_i \in G', |y_i|_G = \ell_i\}| \leq (n+c)^k \prod_{i=1}^{k} \ell_i^p \leq (n+c)^k \left(\frac{ds}{k}\right)^{pk},$$

其中最右侧的不等式由 $\sum_{i=1}^{k} l_i = ds$ 时的算术-几何平均不等式得出:

$$\prod_{i=1}^{k} \ell_i \leq \left(\frac{ds}{k}\right)^k.$$

这样的拆分共有 $\binom{ds-1}{k-1}$ 个，因此:

$$|\{w \in S_{ds} : |w|_{G'} = k\}| \leq \binom{ds-1}{k-1}(n+c)^k \left(\frac{ds}{k}\right)^{pk}.$$

对 $k \leq s$ 求和:

$$|\{w \in S_{ds} : |w|_{G'} \leq s\}| \leq \sum_{k=1}^{s} \binom{ds-1}{k-1}(n+c)^k \left(\frac{ds}{k}\right)^{pk} \leq \binom{ds-1}{s-1} \sum_{k=1}^{s} (n+c)^k \left(\frac{ds}{k}\right)^{pk}.$$

第二个不等式利用了这样一个事实：对于 $d \geq 2$ 和 $1 \leq k \leq s$，有 $\binom{ds-1}{k-1} \leq \binom{ds-1}{s-1}$。

定义:

$$\Sigma_s := \sum_{k=1}^{s} (n+c)^k \left(\frac{ds}{k}\right)^{pk}.$$

写作 $b_k := \left((n+c)d^p \left(\frac{s}{k}\right)^p\right)^k$，可以验证，若 $(n+c)d^p \geq e^p$，则序列 b_k在 k中随 $1 \leq k \leq s$ 递增。由于 $d \geq 3$且 $n + c \geq 2$，我们有 $(n+c)d^p \geq 2 \cdot 3^p > e^p$，因此单调性

条件成立。因此 $\Sigma_s \le s\cdot\, b_s = s\cdot\,((n+c)d^p)^s \le (2(n+c)d^p)^s$。利用 $\binom{ds-1}{s-1} < \binom{ds}{s} \le (ed)^s$：

$$|\{w \in S_{ds} : |w|_{G'} \le s\}| \le (ed)^s (2(n+c)d^p)^s = (2e(n+c)d^{p+1})^s.$$

选择 d，使得 $n^d > 4e(n+c)d^{p+1}$。那么

$$|\{w \in S_{ds} : |w|_{G'} \le s\}| \le (2e(n+c)d^{p+1})^s < (n^d/2)^s < n^{ds}.$$

因此，并非所有长度为 G 的词 ds 都属于 $B_{G'}(s)$，所以 $f_{G'}(s) < ds$。 □

注记 3。当 $p = 0$ 时，条件 (2) 简化为 $n^d > 4e(n+c)d$，从而得到对数密度情形。

下一个结果表明，消失球面密度的宏集合——尽管绝对规模仍呈指数级增长——允许超线性扩张。

定理 5 (概率稀疏宏在 F_n 中产生超线性扩张)。令 F_n 为由 $n \ge 2$ 个生成元 $G = \{a_1, \ldots, a_n\}$ 生成的自由幺半群。存在一个宏集合 $M \subset F_n$，使得

$$\frac{|M \cap S_r|}{|S_r|} \longrightarrow 0 \qquad as\ r \to \infty,$$

and

$$\frac{f_{G'}(s)}{s} \longrightarrow \infty \qquad as\ s \to \infty,$$

其中 $S_r = \{w \in F_n : \ |w|_G = r\}$，$G' = G \cup M$，且 $f_{G'}$ 是扩张函数。更定量地说，存在常数 K、$c > 0$（仅依赖于 n），使得

$$B_G(r) \subseteq B_{G'}\big(K(\log r)^2\big) \quad for\ all\ sufficiently\ large\ r,$$

因此

$$f_{G'}(s) \ge \exp\big(c\sqrt{s}\big) \quad for\ all\ sufficiently\ large\ s.$$

证明。我们将 M 构造为一个小型确定性族（由对数周期词组成）与一个独立随机族的并集。

步骤 0：对数周期词。对于一个长度为 $L = |w|_G$ 的词 $w = b_1 \cdots b_L \in F_n$，若 w 具有周期 d，则称之为如此，当且仅当对所有 $1 \le j \le L-d$，$1 \le d \le L$ 且 $b_j = b_{j+d}$。用 $\mathrm{per}(w)$ 表示满足条件的最小 d。

固定常数 $C > 4\log n$ and $B \ge 2C$（自然对数）。定义确定性宏族

$$P := \Big\{w \in F_n : \ \mathrm{per}(w) \le B\log\big(e + |w|_G\big)\Big\}.$$

步骤 1：密度为 $1/\log l$ 的随机宏。对于每个长度为 $l \ge 2$ 的词 $u \in F_n$，独立地以概率 u 将其纳入随机集合 R

$$p_\ell := \frac{1}{\log(e+\ell)}.$$

令 $M := P \cup R$ 且 $G' := G \cup M$。

步骤 2：消失球面密度。 我们有 $|S_r| = n^r$。首先，计算确定性部分： $P \cap S_r$ 中的任意词都由其周期 $d \le B\log(e+r)$ 和前 d 个字母确定，因此

$$|P \cap S_r| \;\le\; \sum_{d \le B\log(e+r)} n^d \;\le\; n^{B\log(e+r)+1} \;=\; O\big(r^{B\log n}\big).$$

因此 $|P \cap S_r|/n^r \to 0$。

第二， $|R \cap S_r|$服从 $\mathrm{Bin}(n^r,\ p_r)$，其均值为 $\mu_r = n^r/\log(e+r)$。Chernoff界给出

$$\Pr\big(|R \cap S_r| \ge 2\mu_r\big) \;\le\; \exp(-\mu_r/3),$$

且 $\sum_r \exp(-\mu_r/3) < \infty$，因为 μ_r 呈指数增长。由 Borel–Cantelli引理，几乎必然地，对所有充分大的 r， $|R \cap S_r| \le 2n^r/\log(e+r)$。因此，几乎必然地，

$$\frac{|M \cap S_r|}{|S_r|} \;\le\; \frac{|P \cap S_r|}{n^r} + \frac{|R \cap S_r|}{n^r} \;\le\; o(1) + \frac{2}{\log(e+r)} \;\longrightarrow\; 0.$$

步骤 3：一个折半引理（一个宏消耗一半的词）。固定一个长度 r 并令

$$k(r) \;:=\; \lceil C\log(e+r)\rceil.$$

对于一个词 $w = b_1 \cdots b_r \in S_r$， 考虑较长的前置子串族

$$C_r(w) \;:=\; \Big\{ b_i b_{i+1} \cdots b_{i+\ell-1} \;:\; 1 \le i \le k(r),\ \lceil r/2\rceil \le \ell \le r-i+1 \Big\}.$$

$C_r(w)$ 的每个元素长度都介于 $\lceil r/2\rceil$ 和 r 之间。

断言。 对于每个 $w \in S_r$，要么 $C_r(w)$ 包含 P 中的一个元素，要么 $C_r(w)$ 中的所有单词两两不同。

确实，如果两个元素 $C_r(w)$ 相等，它们的长度必定相同，为 l。因此我们会有

$$b_i \cdots b_{i+\ell-1} \;=\; b_j \cdots b_{j+\ell-1} \quad \text{for some } 1 \le i < j \le k(r).$$

设 $d = j - i \le k(r)$。那么词 $u = b_i \cdots b_{i+\ell+d-1}$ 的周期为 d，因此

$$\mathrm{per}(u) \le d \le k(r) \le C\log(e+r) \le B\log\big(e + |u|_G\big),$$

当 r 很大时（因为 $|u|_G \asymp r$ 且 $B \ge 2C$）。因此 $u \in P$，特别是 $C_r(w)$ 包含 P 的一个元素。这就证明了该结论。

现在假设 $C_r(w) \cap P = \emptyset$。那么该结论表明 $C_r(w)$ 是一个由不同的词组成的集合。此外，粗略计数可得

$$|C_r(w)| \;\ge\; k(r)\Big(\lfloor r/2\rfloor - k(r)\Big) \;\ge\; \frac{r}{4}\,k(r)$$

对所有足够大的 r（因为 $k(r) = O(\log r)$）。由于每个 $u \in C_r(w)$ 的长度为 $\ge r/2$，因此我们有$p_{|u|_G} \ge 1/\log(e+r)$。随机选择 R 的过程在不同单词之间相互独立，由此可得

$$\Pr\big(C_r(w) \cap R = \emptyset\big) \;=\; \prod_{u \in C_r(w)} (1 - p_{|u|_G}) \;\le\; \exp\Big(-\sum_{u \in C_r(w)} p_{|u|_G}\Big) \;\le\; \exp\Big(-\frac{|C_r(w)|}{\log(e+r)}\Big).$$

利用 $|C_r(w)| \geq \frac{r}{4}k(r)$ 和 $k(r) = \lceil C\log(e+r)\rceil$，可得

$$\Pr\big(C_r(w) \cap R = \emptyset\big) \;\leq\; \exp\!\Big(-\frac{C}{4}r\Big).$$

因此，对于每个 $w \in S_r$，

$$\Pr\big(C_r(w) \cap M = \emptyset\big) \;\leq\; \exp\!\Big(-\frac{C}{4}r\Big),$$

因为如果 $C_r(w)$ 与 P 相交，那么它当然也与 M相交。

对所有 $w \in S_r$ 使用并集界，

$$\Pr\big(\exists\, w \in S_r \text{ with } C_r(w) \cap M = \emptyset\big) \;\leq\; n^r \exp\!\Big(-\frac{C}{4}r\Big) \;=\; \exp\big(r(\log n - C/4)\big).$$

由于 $C > 4\log n$，右侧关于 r 可求和。根据 Borel–Cantelli引理，以概率 1，存在 r_0，使得对所有 $r \geq r_0$ 每个 词 $w \in S_r$ 都满足 $C_r(w) \cap M \neq \emptyset$。等价地：对于所有足够大的 r，每个长度为-r 的词都有一个长度为 $\geq r/2$ 的宏（确定性或随机），其起点位于该词前 $k(r)$ 个字母之内。

步骤 4：递归解析与 $(\log r)^2$ <style id='5'}>界。 固定 $r \geq r_0$和一个词 $w \in S_r$。根据步骤 3，选择 $1 \leq i \leq k(r)$ 和 $l \geq r/2$，使得

$$u := b_i \cdots b_{i+\ell-1} \in M.$$

于是我们可以写成

$$w \;=\; (b_1 \cdots b_{i-1}) \;\cdot\; u \;\cdot\; (b_{i+\ell} \cdots b_r),$$

其中，前缀 $(b_1 \cdots b_{i-1})$ 使用来自 G 个（填充符）的 $i - 1 \leq k(r)$ 个生成元，而 u 是一个宏。后缀的长度至多为 $r - l \leq r/2$。

对后缀应用相同的步骤，并不断迭代。至多经过 $\lceil \log_2 r \rceil$ 次迭代后，剩余后缀的长度有界，可以用生成元拼写出来。每次迭代中，我们最多使用 $k(r) + 1 = O(\log r)$ 个词元（填充符加一个宏）。因此，存在一个常数 K，使得每个$w \in S_r$ 都有

$$|w|_{G'} \;\leq\; K(\log r)^2$$

对所有足够大的 r 都成立。由于$|w|_{G'}$ 关于 $|w|_G$ 单调，同一界也适用于长度为$\leq r$ 的每个词，即

$$B_G(r) \subseteq B_{G'}\big(K(\log r)^2\big) \quad \text{for all large } r.$$

这就得到了所声称的扩张界。写作 $s = K(\log r)^2$ 意味着 $r = \exp(\sqrt{s}/K)$，因此 $\overline{f_{G'}}(s) \geq \exp(c\sqrt{s})$，其中 $c = 1/\sqrt{K}$。特别地，$f_{G'}(s)/s \to \infty$。□

注 4：宏集合 M 在随机包含过程中几乎必然存在，但未给出明确的构造。

注 5： M 的球面密度满足 $|M \cap S_r|/|S_r| \sim 2/\log r$。就绝对数量而言， $|M \cap S_r| \sim 2n^r/\log r$：每个球面中所占的比例趋于零，但在每个半径上都有指数级数量的宏。这与定理 4 形成对比，后者中每个长度至多有 cl^p 个宏，从而迫使线性扩张。

3. 解读 MathLib 中的数据

现在，我们将第 2 节的结果与 MathLib 进行比较。我们将 MathLib 视为人类数学（HM）的替代指标。当然，MathLib 的结构部分受到 Lean 类型论以及其贡献者社区所做选择的影响。由于 MathLib 仍然规模较小（本研究所用版本约有 500,000 个元素[4]），且发展不均衡，我们在其前沿观察到了显著的有限尺寸效应。这限制了我们推断精确维度性质的能力，但数据与低维行为相符，这种行为比起 F_n，更具 A_n 的特征。

3.1. **构建依赖图**。 我们根据 MathLib 的 Lean 内部表示构建依赖图。顶点包括所有 MathLib 元素：引理、定理、定义、结构和归纳类型，此外还有一个代表类型的合成节点（`Sort`）以及 Lean 核心库中的元素（`And`、`Nat`、`List` 等），这些元素没有进一步的 MathLib 依赖项。

每个 MathLib 元素包含两部分：签名（定理陈述或类型定义）以及可选的主体（证明或定义表达式）。对于每个元素 u，我们统计其他每个元素被引用的次数 v，并据此生成一条从 u 指向 v 的有向边，其权重为该次数。因此，边指向依赖项。每个元素对应的数据由一个引用计数向量构成。请注意，这种构造仅记录每个引用出现的次数，而忽略了引用在元素表达式树中的顺序。例如，两个证明以不同方式应用相同的引理，但各引理的应用次数相同，那么它们在依赖图中的条目就完全相同。

举个简单的例子：

```
引理simple_lemma {A B : Prop} : A/\ B -> B :=函

```

作为参考，其内部表示如下：

```
类型: forallE (A : Sort 0), forallE(B : Sort 0), forallE (_ : app (app And #1)#



```

签名中包含两个`Sort`（用于`A`和`B`的`Prop`类型）以及一个`And`。主体中包含两个`Sort` 、一个`And` 和一个`And.right` 。这些会生成从`simple_lemma`指向相应元素的加权边。

[4]提交 d167cc6dc962ab340507362ea2f4bcfcff44f01b，日期为2025年10月17日。

为了提取这些依赖关系，我们递归遍历每个 MathLib 元素的签名和主体，收集所有 `const` 节点（表示对具名声明的引用）：

```
1 def collectElems (e : Expr) (acc : 哈希映射 Name Nat :=  {}): 哈希映射 Name Nat :=2 match e
```

关键情形是 `.const declName _`，它表示对具名声明的引用；每次出现都会在我们的图中生成一条边。其他情形（`app`表示函数应用、`lam` 表示 λ抽象、`forallE`表示依赖箭头、`letE` 表示 let 绑定）只需递归处理其子表达式。我们还统计`.sort` 的出现次数，将`Sort` （例如，它表示`Prop`）视为原始元素。

所得的图包含少量环（约 60 对相互依赖元素，且全部涉及不安全递归）。由于我们的分析要求图无环，我们将每个强连通分量折叠为单个顶点，使顶点数量从 463,719 减少到 463,661。

3.2. **包装长度与未包装长度，以及深度**。 Lean 核心库元素和`Sort` （原始元素）构成 MathLib 有向无环图的汇点，类似于我们幺半群模型中的生成元 G 。我们将所有其他（非原始）元素视为类似于宏集合M。因此，我们将 MathLib 视为 $G' = G \cup M$。

元素 u 的展开长度 记为 $|u|_G$，即递归展开所有引用后原语的总数。原语的展开长度为 1。对于任何非原始元素，若其边指向元素 v_1、……、 v_k，对应权重为 w_1、……、 w_k:

$$|u|_G = \sum_{i=1}^{k} w_i \cdot |v_i|_G.$$

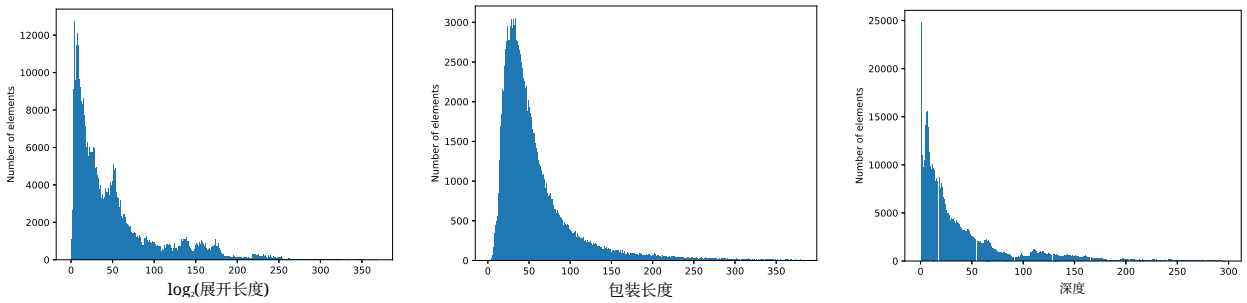
正如记号所示，展开长度直接对应于幺半群模型中的 G-长度。

元素的 包装长度 u 是其词元长度：即 Lean 解析器处理 Lean 定义后得到的词元数量。包装长度对应于 $|u|_{G'\setminus\{u\}}$ 在

幺半群模型中。也可以将包装长度定义为 Lean 内部表示中的引用数量，即依赖图中所有出边的总权重。然而，`simp` 和 `rw` 等策略在 elaboration 过程中会展开成对许多基本元素的内部引用，从而增加某些元素的引用计数，却不会引入深层依赖。按照引用计数指标，这会产生包装长度很大、但展开长度增长缓慢的元素——压缩曲线上的平台期反映的是证明自动化，而非数学内容。词元计数避免了这一问题：无论策略调用在内部生成多少引用，它都只是一个词元。MathLib 中有些元素由 Lean 在内部生成，没有人工编写的源代码；报告包装长度时，我们会省略这些元素。

元素的深度 是其在依赖有向无环图中到原语的最长路径长度，也就是说，到达生成元所需的连续引用扩张步骤的最大次数。原语的深度为 0。

3.3. 包装长度、展开长度和深度的分布。 图 2(a)–(c) 展示了元素按 $\log_2$(展开长度)、包装长度和深度划分的分布。在每种情况下，分布都集中在较低值处，并带有长尾。较大值处的下降反映了有限尺寸效应：MathLib 的覆盖范围并不均匀，我们预计随着库的增长，尾部会延伸并逐渐填充。未来若能追踪 MathLib 随时间的演变，便可测量其前沿如何扩张。



(A) (B) (C)

图 2. MathLib 元素按以下指标的分布：(a) $\log_2$(展开长度)、(b) 包装长度和 (c) 深度。

3.4. 展开长度与包装长度。 图 3 展示了中位数 $\log_2$(展开长度) 与包装长度的关系。$f_{G'}(s)$ 衡量的是 G' 的覆盖范围，而这些数据点反映了 MathLib 压缩策略所实现的 G-长度 $|u|_G$。$\log_2$(展开长度) 与包装长度之间近似线性的关系（斜率为 ≈ 0.4 位/词元）表明存在指数扩张：每增加一个词元，原语数量便会获得近似恒定的乘法增益。第 3.7 节提供了进一步的解读。

包装长度较小时出现的尖峰值得注意。这是由以下几类元素造成的，但不限于这些类型：（1）缩写，例如模块 `Topology.VectorBundle.Basic` 中的 `isOpenMap_proj` ，其包装长度为 7，展开长度为 10^{48}；（2）使用复杂中间定理的“最终”定理，例如模块 `Analysis.Calculus.BumpFunction.Normed` 中的 `integrable`，其包装长度为 9，展开长度为 $2 \cdot 10^{54}$；（3）复杂陈述的特殊情况，例如模块`Analysis.Real.Hyperreal` 中的 `infinitesimal_zero`，其包装长度为 8，展开长度为 $3.6 \cdot 10^{31}$。（超链接位于模块名称下方。）

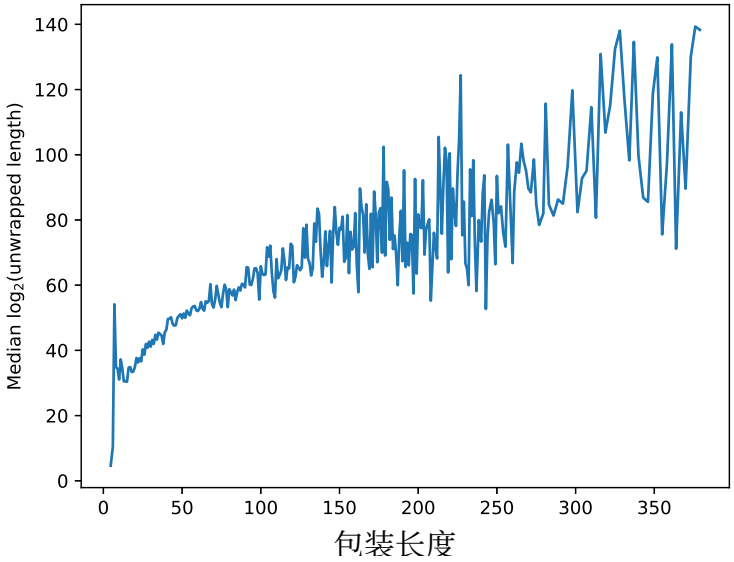


图 3。中位数 $\log_2$(展开长度)与包装长度的关系。每个点表示所有具有该包装长度的元素的中位数。

3.5. **包装长度与深度的关系**。图 4 展示了中位数包装长度随深度的变化。两者的关系近似平坦（至多只有轻微的正斜率）：（忽略大深度处的异常值）在深度 0–300 范围内，中位数包装长度大致保持在 50–120 之间，与深度没有明显的依赖关系。这意味着，单个定义不会随着深度增加而系统性地变长，这让人联想到位值记数法中的 10 的幂宏。对此并不意外，因为定义表达式不太可能变得任意长；相反，我们预计随着包装长度增加，会出现模块化。

3.6. **展开长度与深度**。 图 5 展示了 $\log_2$(展开长度) 中位数随深度的变化。二者关系近似线性，斜率接近 1，这表明展开长度随深度呈指数增长：每增加一层定义，原语数量大致都会获得恒定的乘数增益。最大深度约为 300，这与引言中提到的代数几何条目所达到的最大展开长度约为 10^{104}，一致。

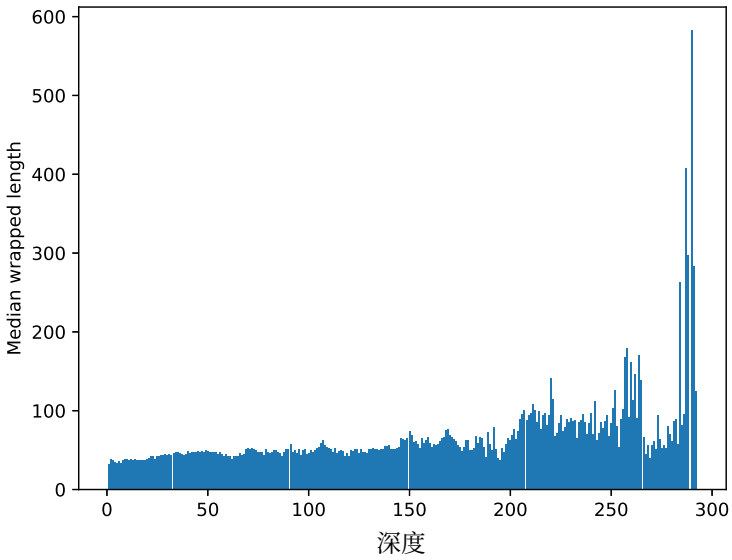


图 4. 包装长度中位数与深度的关系。

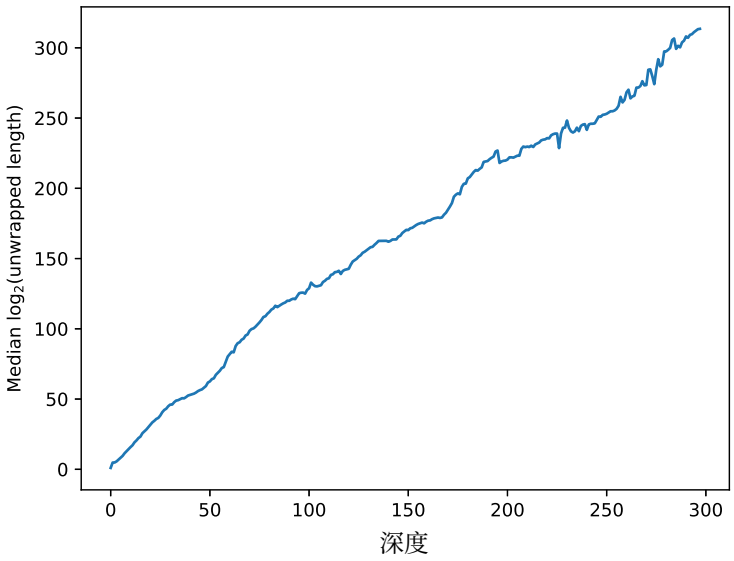


图 5. $\log_2$(展开长度)中位数与深度的关系。近似线性的关系表明，展开长度会随深度呈指数增长。

3.7. 区分不同状态。 为了将 MathLib 的测量结果与第 2 节的幺半群模型联系起来，我们需要建立展开长度、包装长度、深度及其在幺半群中的对应概念之间的关系。展开长度直接对应幺半群中的 G-长度 $|u|_G$：二者都统计完全展开所有引用后原始符号的总数。

幺半群中一个元素的包装长度 $u \in G'$ 为 $|u|_{G'\setminus\{u\}}$：使用除该元素自身以外的所有生成元来表示 u 时所需的最小代价，其中每个生成元（原始元素或

宏）的代价均为 1。这与 MathLib 的惯例相呼应：元素的包装长度就是其令牌计数，也就是使用所有可用的具名元素书写该定义的代价。

幺半群中元素 u 的深度 由 G' 生成，使用最优表示递归定义。[5] 每个原始元素 $g \in G$ 的深度均为 0。对于任意满足 $u \notin G$的元素 u ，计算 u 在 $G'\setminus\{u\}$中的最优表示；若该表示仅使用原始元素，则 $\mathrm{depth}(u) = 1$，否则

$$\mathrm{depth}(u) = 1 + \max\{\mathrm{depth}(v) : v \in M\setminus\{u\},\quad v \text{出现在 } u\text{的最优表示中}\}。$$

因此，深度衡量的是为了以最优方式表示 u所需的最长宏依赖链的长度，链的终点是原始元素。（请注意，第 2 节的扩张定理对宏集合只施加了稠密性条件。）这与 MathLib 深度类似，但并不完全相同：MathLib 深度采用依赖有向无环图中到原始元素的最长路径；MathLib 深度反映作者的选择，而幺半群深度则反映最优压缩下的内在层级结构。

请注意，我们测量的是 MathLib 中所有具名元素的这些量，而不是由它们可能构成的所有表达式。因此，在每种幺半群情形下，我们只考察生成集 $G' = G \cup M$中的元素，而非环境幺半群中的所有元素。

表 2 总结了每种幺半群情形下这三个量之间的关系。“简约性”列表示宏集合的增长速度是否严格慢于环境幺半群：在 A_n中为次多项式宏增长，或在 F_n中为次指数宏增长。简约的宏集在 A_n 中实现指数扩张，但在 F_n中只能实现线性扩张。下面我们依次讨论每一行；在不失一般性的情况下，具体考察 $A_1 = \mathbb{N}$和 F_2 。

情形	$\log\lvert w\rvert_G$ 与深度	$\lvert w\rvert_{G'\setminus\{w\}}$ 与深度	对数 $\lvert w\rvert_G$ 与 $\lvert w\rvert_{G'\setminus\{w\}}$	简约性
A_n，对数稠密性	线性	平坦*	退化*	Yes
A_n，华林	线性	Flat	退化	No
A_n，双对数	指数	双重指数	对数	Yes
F_n，多项式	退化	退化	对数	Yes
F_n，概率	线性	二次	凹 ($\sqrt{\cdot}$)	No

表 2. 各幺半群情形下 $\log\ \lvert w\rvert_G$(log 展开长度)、$\lvert w\rvert_{G'\setminus\{w\}}$(包装长度)与深度之间的预测关系。测量仅限于 $G' = G \cup M$的元素。“退化”表示自变量（即“y 与 x 的关系”中的“x”）有界。星号 $(*)$ 表示，若考虑幺半群中的一般元素，则平坦 $\to$线性，且退化 $\to$ 线性；相关讨论见 3.7.6。

3.7.1. A_n，对数密度宏（定理 *1*）。宏集合为 $M = \{b^j a_i :\ i = 1, \ldots\ldots, n,\ j \geq 1\}$。我们专门考虑带有 $M = \{b^j :\ j \geq 1\}$ 的 A_1。宏 b^j 的 G-长度为 b^j，因此 $\log\lvert b^j\rvert_G = j\log b$。其在 $G'\setminus\{b^j\}$ 中的最优表示使用了 b 个 b^{j-1} 的副本，从而得到包装长度 b，且与

[5]对于一般幺半群，由于缺乏消去律或消去律有限，这一定义无法通过算法计算；该定义适用于 A_n, Nil_n (稍后讨论)和 F_n。

j 无关。对此进行迭代，$\mathrm{depth}(b^j) = j$。因此，$\log |w|_G$ 与深度呈线性关系，包装长度在所有深度上都保持平坦，并且由于包装长度恒定而 $\log |w|_G$ 可自由变化，第三种关系是退化的。

如果考虑 A_1 的一般元素，而不是将范围限定为宏元素，那么第 2 列和第 3 列的情况就会有所不同。一般元素 $x \in \mathbb{N}$ 的包装长度为 $\sim \log x$，深度为$\sim \log x$。因此，包装长度通常随深度线性增长，并且 $\log |x|_G \sim |x|_{G' \setminus \{x\}}$。

3.7.2. A_n，多项式密度宏 / 华林（定理 3）。宏集合为 $M = \{m^k : m \geq 1\}$，其中 $k \geq 2$ 固定（特化为 A_1）。宏 m^k 的长度为 G，因此 $\log |m^k|_G = k \log m$。根据华林定理，m^k 可以表示为至多 $g(k)$ 个严格小于它的整数的 k 次幂之和，因此包装长度至多为 $g(k)$，且与 m 无关。我们不知道每个阶段的最优拆分，但可以界定深度：折半策略（用不超过 $\approx m/2$ 的整数的 k 次幂表示 m^k，然后递归）给出的深度为 $O(\log m)$，并且第 1 列呈线性关系；较慢的递减会带来更大的深度和次线性关系，若深度$\sim m$，则可降至对数关系。所有情况下，包装长度都保持有界，因此第 2 列平坦，第 3 列则退化。

华林和对数密度情形在表 2 的前三列中给出相同的预测，但简约性不同。与对数密度情形相比，对于华林情形下 A_1 的一般元素，第三列仍然是退化的，因为每个 $x \in \mathbb{N}$ 的包装长度至多为 g（k）。

3.7.3. A_n，双对数密度（定理 6）。宏集合为 $M = \{b^{b^j} : j \geq 0\}$（特化为 A_1）。宏 $m_j = b^{b^j}$ 的对数为 $|m_j|_G = b^j$ 对数 b，其随 j = 深度（m_j）呈指数增长。在 $G' \setminus \{m_j\}$ 中，m_j 的最优表示使用 $\lfloor m_j / m_{j-1} \rfloor = b^{b^{j-1}(b-1)}$ 个 m_{j-1} 的副本，因此包装长度为 $\sim b^{b^{j-1}(b-1)}$，其随深度呈双指数增长。消去 j：$\log(\text{wrapped}) \sim b^{j-1}(b-1) \log b \sim \frac{b-1}{b} \log |m_j|_G$，因此 $\log |m_j|_G \sim \frac{b}{b-1} \log(\text{wrapped})$，这表明第三列中存在对数关系。前三列与 MathLib 数据不一致。

3.7.4. F_n，多项式密度宏（定理 4）。宏集合中，每个 G-长度为 l 的长度层最多包含 cl^p 个元素，而 F_n 中该长度共有 n^ℓ 个单词。因此，长度为l的单词中属于宏的比例为 cl^p/n^ℓ，并以指数速度趋于零。对于任意满足 $|m|_G = r$ 的宏 m，其在 $G' \setminus \{m\}$ 中的最优表示包含另一个宏的概率可以忽略：指数稀疏的宏集合无法足够密集地填充 F_n 的指数增长的球面，因而无法维持层级嵌套。因此，几乎所有宏的深度都是 1，其最优表示几乎完全由原语组成。于是，$\log |m|_G$ 与深度的关系，以及封装与深度的关系，都是退化的。由于 $\log |m|_G \approx \log(\text{封装})$，第 3 列的条目是对数关系。同样，前 3 列的条目与 MathLib 不一致。

3.7.5. F_n，概率稀疏宏（定理 5）。这里，宏集合在半径 r 处包含 $\sim 2n^r/\log\ r$ 个元素：数量呈指数增长，但仅占球面的对数级趋零比例（$|M \cap S_r|/|S_r| \sim 2/\log\ r$）。这与多项式情形形成对比，后者的比例呈指数趋于零。绝对密度足以使定理 5 的折半方案发挥作用：对于长度为 r 的几乎每个单词，在其前 $k(r) = O(\log\ r)$ 个位置内，都有一个长度为 $\geq r/2$ 的宏开始出现。层次深度由此发展起来。

对于一个具有 $|m|_G = r$ 的宏 m，折半方案给出深度$(m) \sim \log_2 r$，因此 $\log\ |m|_G \sim$ 深度。对于包装长度：在 $\sim \log_2 r$ 个层级中的每一层，我们至多花费 $k(r_t) = O(\log\ r_t)$ 个原始填充项，加上一个宏，其中 $r_t \leq r/2^t$ 是第 t 层的剩余长度。总包装长度为

$$\text{wrapped} \leq \sum_{t=0}^{\log_2 r} O(\log(r/2^t)) = \sum_{t=0}^{\log_2 r} O(\log r - t) = O((\log r)^2),$$

其中该求和为等差和，包含 $O(\log\ r)$ 项，每项大小为 $O(\log\ r)$。由于深度 $\sim \log\ r$，封装 $= O(\text{深度}^2)$。消去深度：$\log\ |m|_G \sim$ 深度 $\sim \sqrt{\text{封装}}$，从而在第 3 列中得到凹的 $(\sqrt{\cdot})$ 关系。

3.7.6. 总结与“宏集合”的识别。MathLib 数据（图 3、4 和 5）显示：第 1 列中存在近似线性关系，第 2 列中存在近似平坦关系，第 3 列中存在近似线性关系。F_n 情形至少在三列中的两列上与数据不一致。对于宏元素，A_n 对数密度和华林情形都与第 1 列和第 2 列相符，但预测第 3 列会退化。对于对数密度情形中的一般元素（回顾表 2 中的星号），第 2 列和第 3 列都变为线性。MathLib 的第 2 列图近似平坦，但可能具有轻微的正斜率，这与对数密度情形中宏元素和一般元素的混合相一致。这表明 A_n 具有对数密度宏；我们注意到，这也是最简约的情形。

我们并不认为将所有非原语都视为 MathLib 的真正“宏集合”是合理的。MathLib 包含缩写和平凡特化（见第 3.4 节），它们只是调用深层元素，本身几乎没有压缩作用。另一方面，有些元素能带来显著的压缩，例如滤子抽象，它将许多极限定理统一到一个框架中。如何确定正确的“宏集合”是一个核心问题。

优化“宏集合”的一种简单方法，是按依赖图中的入度进行筛选，移除底部 x 百分位数中对应于不同 x的元素。移除 M中的一个元素不会改变未包装长度，但会增加包装长度并降低深度。另一种方法是将“宏集合”限制为定义类元素，例如其结果类型为`Sort`的元素；我们发现，由此得到的宏集合几乎不能带来压缩。还可以将问题表述为一个优化问题：例如，给定一个依赖有向无环图，寻找固定大小为k 的“宏集合”，使总包装长度最小。这些方法或相关方法能否使这三个指标与 A_n 对数密度预测更加一致，仍是一个开放问题。

识别“宏集合”的困难，可能反映了 MathLib 本身作为 HM 代理的局限性。一种能够捕捉 MathLib 留下的隐含关系（例如，一组相关定理是同一模式的实例），或允许像超图那样存储同一定理的多个证明的丰富表示，或许能够得到更令人满意的“宏集合”。寻找一种具有可识别“宏集合”的表示，并不只是图书馆学问题：它可能揭示新的数学结构，并帮助引导自动化智能体。

4. 讨论

4.1. 为什么选择幺半群? 幺半群对证明的序列结构进行建模；对于多个前提同时被使用的依赖结-构， n-范畴和“球面岩浆”提供了替代框架。受“群的增长” [Mil68], 的启发——它本质上独立于生成集——人们或许会预期，形式系统的选择不会显著影响数学的形态。然而，数学的递归性质导致了出人意料地巨大的扭曲：在具有相同证明能力的数学理论的不同表述之间进行改写，可能在非递归意义上效率极低（见 [BDF] 及其中的引用）。如果规则和语法固定，底层的最小证明有向无环图便是一个以公理为基点的离散度量空间。这种度量几何由每个幺半群的凯莱图几何建模，并构成多项式增长和指数增长概念的基础。

4.2. 为什么选择幺半群而非群。群引入了逆元，而逆元使展开分析变得复杂。这一困难有其历史先例：Post [Pos47] 于1947年将停机问题编码进幺半群的字问题中，但将其扩展到群则需要Boone和Novikov再花十年时间研究 [Nov55,Boo59]。这一差距反映了逆元所引入的技术复杂性。

就我们的目的而言，这个问题很具体。在自由群中，任意一个词 w 都可以通过选择几乎互相抵消的宏 m 和 m'，写成长度为 2 的形式： $mm' = w$。只要让 m 和 m'足够长，我们就能对每个词做到这一点，同时使宏密度任意小。这种“消去技巧”使群的扩张问题变得毫无难度。

有人可能会反驳说，肯定前件之类的推理规则（$A \to B,\ \ A \vdash B$）涉及某种消去。不过，可逆逻辑只需付出适度开销就具有普适性 [Ben73], ，因此消去并非计算的必要条件。更务实地说，自由幺半群可以进行精确分析——我们只是在光线充足的地方做研究。

谈到群，值得注意一个李理论类比：包含关系$A_n \hookrightarrow \prod_{i=1}^{n} F_{n_i}$类似于极大环面 $T^{n-1} \hookrightarrow SO(n)$的包含关系。如果将 A_n拓宽为自由幂零幺半群 $Nil_{n,\ k}$ （n 计数同调秩，且 k 计数幂零级），那么包含关系 $Nil_{n,\ k} \hookrightarrow \prod_{i=1}^{n} F_{n_i}$ 让人联想到半单李群的岩泽分解 $G = KAN$。

4.3. 幂零与可解幺半群。 A_n 与 F_n 之间的二分反映了不同的增长率：多项式增长与指数增长。关键在于这种差异，而非交换与非交换。

非负海森堡幺半群展示了多项式增长的情形。[6] 它是幂零且非阿贝尔的，并具有多项式增长。同样，在这里以及所有自由幂零幺半群中，对数密度宏会产生指数扩张，就像 A_n 一样。证明很简单。$Nil_{n,\,k}$ 是一个具有 n 个“基”生成元的幺半群，另有一些次级生成元用来模拟交换，并且只需最少的关系来编码层级为 k 的右嵌套交换子（请记住，这些幺半群中没有逆元）。所有这些生成元各自都生成一个 $\mathbb{N}$ 的副本，而它们的数量关于 n 和 k 是多项式级的。在这些 $\mathbb{N}$ 个方向中的每一个方向上，分别构造宏集。然后取这些 $\mathbb{N}$ 个宏集的并，作为 $Nil_{n,\,k}$ 的宏集。除去多项式因子后，它的扩张函数与每个 $\mathbb{N}$ 中的迷你宏集相同。在具有消去律的幺半群中（即 $ac = bc$ 蕴含 $a = b$，且 $ca = cb$ 蕴含 $a = b$），几何上至关重要的“多项式增长”性质（与群中的情形一样）由格罗莫夫判据刻画[Gro81]：它们是某个幂零幺半群中的有限指数子幺半群。因此，我们已经有了一个充分了解的缓慢增长幺半群类，可用于为人类数学建模。

SOL 格的非负部分展示了指数增长的情形。这个幺半群是可解且非阿贝尔的——其结构比自由幺半群更丰富——但仍具有指数增长。与定理 4 类似的计数论证表明，它不存在扩张超线性的、增长速度为多重对数级的宏。

4.4. 为什么形式数学难以压缩。人类数学容易压缩，这一点从经验中早已熟知，也得到了我们对 MathLib 的分析证实。较不显而易见的是，整体而言，形式数学必然包含大量难以压缩的部分。

我们所说的压缩有其特定含义：归约压缩，即定义的局部替换，而非任意的算法编码。π 的数字说明了两者的区别。数 π 的柯尔莫哥洛夫复杂度很低——一个简短的程序就能计算出它——但目前尚无已知方法能通过模式替换对其数字串进行局部压缩。归约压缩要求找到可命名的重复结构。不过，高度的归约可压缩性确实意味着较低的柯尔莫哥洛夫复杂度，因为定义有向无环图编码了一种简短的重构过程。此外，这一过程的运行时间至多为线性时间，因此归约可压缩元素具有较低的逻辑深度（依照 Bennett [Ben88] 的含义）。因此，归约可压缩性比低柯尔莫哥洛夫复杂度更为严格。

基于这一理解，大多数 FM 的不可压缩性可由标准复杂性假设推出。考虑形如“布尔公式 S 不可满足”的定理。我们认为，典型的 S 这类定理不属于 HM。这类陈述很容易写出，但

[6] 非负海森堡幺半群由 3×3 对角线下方元素为零、对角线元素为一的非负整数矩阵组成。它可以表示为 $\langle a,\ b,\ z \mid ab = baz,\ \ az = za,\ \ bz = zb\rangle$。

对于一般的 S，人们相信其最短证明长度呈指数级增长——本质上，必须检查所有可能的变量赋值。任何定义系统都无法绕过这种穷举搜索。如果这类证明能够被大幅压缩，就会与假设 $P \neq NP$相矛盾。

因此，HM 由 FM 中可压缩的区域构成——也就是定义能够发挥作用的数学领域。这促使我们寻找高度可压缩的幺半群模型（用于建模 HM），同时它们又能嵌入更大的、难以压缩的幺半群中（用于建模 FM）。可压缩性甚至可以作为数学品味的初步近似；第 5 节将进一步探讨这一想法，区分归约压缩（通过定义缩短陈述）和演绎压缩（在给定陈述的情况下缩短证明），并利用 PageRank 将二者结合起来，形成一种自动化的数学兴趣度量。

4.5. **与元胞自动机的比较**。柯尔莫哥洛夫复杂度考虑了任何可能算法的作用，因此不受局部性限制；而归约压缩要求算法在局部发挥作用。元胞自动机（CA）同样在局部发挥作用，但其作用范围限于固定的格或图。在归约压缩中，底层格还会在每一步发生坍缩（其位置标签也会改变）；这种动态过程可称为“坍缩元胞自动机”（CCA），这也是我们对压缩运作方式的理解。

4.6. **一则自指性评论**。 在撰写本文时，我们引入了一个新的数学概念：扩张函数 $f_{G'}(s)$。它推广了加法数论中的一个概念。子集 $S \subseteq \mathbb{N}$ 的加法秩 是使其和集等于 $\mathbb{N}$所需的 S 副本的最少数量；渐近秩 则是使其和集在 $\mathbb{N}$中共尾所需的副本的最少数量。当 S 过于稀疏，以至于任何有限数量的副本都无法覆盖 $\mathbb{N}$时，这些概念便不再适用。我们的扩张函数衡量的是：至多取自 G'的 s 个元素之和，能够覆盖多大的球——当无法实现无限覆盖时，这是一种自然的推广。

我们在研究定义在数学中的作用时，竟然引入了一个定义，这或许恰如其分。定义的形成对于数学实践而言如此自然，以至于即便是分析数学，也需要新的定义。

5. 应用与展望

我们能否赋予人工智能智能体方向感——一种自动化标准，用以判断哪些数学陈述值得关注？历史和文化因素无疑会影响人们对数学兴趣的判断，但在这里，我们试图找出数学表示本身固有的可观测量（另见 [BDF]）。

本文的核心论点提出，压缩性是一个自然的候选指标。这里我们考察两种类型。对于数学语料库中的任意元素 u，无论是定义、引理还是定理，都将其<体 id='2'>归约压缩</体>定义为：

$$T_0(u) = \frac{|S|_G + |B|_G}{|S|_{G'\setminus\{u\}} + |B|_{G'\setminus\{u\}}},$$

	G	$G'\setminus\{u\}$
S	$\lvert S\rvert_G$	$\lvert S\rvert_{G'\setminus\{u\}}$
B	$\lvert B\rvert_G$	$\lvert B\rvert_{G'\setminus\{u\}}$

表3. 签名为 S （陈述）、主体为 B（证明）的元素的四种度量。各行之间的比率衡量归约压缩；$G'\setminus\{u\}$ 列中上下两项之间的比率衡量演绎压缩。

完整元素（签名加主体）的展开长度与包装长度之比。这里，我们使用第 3.7 节中定义的幺半群记号，其中 $\vdash|_G$ 对应展开长度，而 $|\cdot|_{G'}$<样式 id='5'>\{</样式>$_u$<样式 id='8'>} </样式>对应包装长度。表3总结了由此得到的元素的四种度量。具有较大 T_0 （“品味”）的元素存在于定义能发挥显著作用的区域——正是我们所认定的人类数学所在的区域。探索数学的智能体可以在不同区域间探索时，跟踪 T_0 的平均值。这或许能帮助它始终贴近人类数学。尽管这可能会使智能体偏向抽象——而追求抽象本身并非目的——但T_0 能够为智能体提供方向感。

封装后的主体长度与封装后的签名长度之比衡量另一种可压缩性，即演绎压缩[7]：

$$I_0(u) = \frac{|B|_{G'\setminus\{u\}}}{|S|_{G'\setminus\{u\}}}.$$

没有证明体的元素（原始元素、结构声明、归纳类型）具有 $I_0 = 0$；它们仍可通过定义实现较高的 T_0 。具有较大 I_0 （“趣味性”）的元素，其陈述简短而证明冗长，即使经过压缩也是如此。在 MathLib 中，具有较高 I_0 的元素包括代数几何和范畴论中的深层定理；这些领域通过层层抽象，将庞大的未封装表达式压缩成易于处理的陈述。可以想象，标志性结果的形式化版本——费马大定理、庞加莱猜想、奇点消解——会达到极高的压缩比；它们简洁的陈述背后，实则隐藏着庞大的证明机制。

当陈述变得足够长，足以编码逻辑难题时， I_0 就可能被钻空子：元数学构造能为那些不太值得关注的对象产生任意大的数值。例如（并感谢 Sam Buss），断言 k-一致性的定理——即某个形式系统对于 $0 = 1$ 不存在长度少于 k 个符号的证明——其压缩长度可以很短，因为递归函数论允许人们简洁地描述某些极其巨大的整数 k。然而，Pudlák [Pud86,Pud87] 证明，在足够丰富（且一致）的系统中，任何关于 k-一致性的证明都必须具有 $\Omega(k^{1/2})$的长度——任何定义系统都无法绕过这一点。令 $k = \mathrm{BB}(n)$，其中 BB 表示忙海狸函数，就能得到一族定理，其$I_0 \approx \mathrm{BB}^{1/2}(k)/\log\ k$ 极其巨大；至少就单个定理而言，很少有人会认为它有趣，因为

[7]另一种选择是 $|B|_G/|S|_G$，即以 G 而非 $G'\setminus\{u\}$来衡量；后者衡量压缩前的原始证明-陈述比。这样会大幅提高那些陈述看似简单（如费马大定理）、但证明需要复杂额外发展的定理（如椭圆曲线理论）的得分。

这只是一大族类似定理中的一个，它们以 k为参数；单独来看，它们只是逻辑上的奇闻，而非数学的核心内容。

5.1. **PageRank 式的改进**。问题在于，I_0 会平等对待所有高压缩元素，而不考虑它们在更广泛数学结构中的作用。改进方案不仅应纳入某个元素实现的压缩程度，还应考虑它在构建其他高价值元素时的有用性。

Google 的 PageRank算法 [BP98] 提供了一个自然的框架。考虑完整的依赖图，其中的边从每个元素指向其依赖项。在该图上进行随机游走，访问最终会汇聚到原语——即有向无环图的汇点——它们实现的压缩程度较低，而且从我们所寻求的意义上来说，在数学上并不有趣。标准的解决方法是传送：每一步中，游走者以概率 α 沿着一条边前进，以概率 $1-\alpha$ 跳转到一个随机节点。即使采用均匀传送，也可能仅凭图结构识别出有用元素，从而产生非平凡的排名。我们建议让传送偏向高压缩元素。我们将两种压缩度量参数化地组合为 $J_0 = \beta T_0 + (1-\beta) I_0$（先将各项归一化到相近的尺度），其中 $0 < \beta < 1$ 取某个值；并令元素 u 以概率 $J_0(u)/\sum_v J_0(v)$ 被选为传送目的地。由此得到的转移矩阵为

$$P(v,u) = \alpha \cdot \frac{w(u,v)}{W(u)} + (1-\alpha) \cdot \frac{J_0(v)}{Z},$$

其中，$w(u,v)$ 是 u引用 v、$W(u) = \sum_x w(u,x)$ 的次数，是 u 的总引用次数，而 $Z = \sum_x J_0(x)$。

平稳分布 π 满足 $\sum_u P(v,u)\pi(u) = \pi(v)$；标准 PageRank理论（即 Perron–Frobenius定理）保证了其存在性和唯一性，因为每个节点都有正的传送概率。定义

$$I_1(u) = \pi(u).$$

如果元素自身具有较高的压缩性（经常作为传送目的地），或者被经常访问的元素所依赖，那么它们的得分就会很高。这体现了“承重”元素：它们支撑着数学中可压缩的区域。我们预计需要适当调整 α，以避免出现平凡的 π，例如所有质量都集中在公理或有向无环图中最大的 J_0 元素上。

5.2. **一些未决问题**。首先是计算挑战：确定最优压缩需要在各种可能的定义中进行搜索，这在计算上代价高昂。我们的兴趣度量假设定义集合是固定的，但探索 FM 的智能体需要即时提出新的定义。能否以足够高效的方式完成这一过程，从而指导探索?

其次，定义压缩位于一个谱系的极端。一端是柯尔莫哥洛夫复杂度，它允许任意的算法压缩，但不可计算。

定义压缩具有局部性且易于高效验证：应用一个定义只需要检查某些性质是否已经推导出来，整个过程最多在线性时间内运行（用 Bennett 的 [Ben88] 术语来说，它具有较低的“逻辑深度”）。是否存在有用的中间地带——比局部替换更强大、但在计算上仍然易于处理的压缩方法?

最后，我们提出一个关于 MathLib 及类似代码库的经验性问题：对于“有趣”的命题，其证明是脚踏实地（只使用浅层中间引理），还是会借助高度压缩的中间命题一飞冲天? 从物理角度看，要抵达深层定理，必须克服哪些势垒? 第 3 节中的深度与质量分布提供了初步数据，但仍需开展系统研究，将这些指标与人类对趣味性的判断相关联。

引用

[BDF] Maissam Barkeshli、Michael R. Douglas 和 MICHAEL H. FREEDMAN，《人工智能与数学结构》。即将出版。[Ben73] Charles H. Bennett，《计算的逻辑可逆性》，《IBM研究与开发杂志》17（1973），第6期，525–532。[Ben88] ，《逻辑深度与物理复杂性》，《通用图灵机：半世纪回顾》（1988），227–257。[Boo59] William W. Boone，《字问题》，《数学年刊》70（1959），第2期，207–265。[BP98] Sergey Brin 和 Lawrence Page，《大规模超文本网络搜索引擎的剖析》，《计算机网络与ISDN系统》30（1998），第1–7期，107–117。[CT06] Thomas M. Cover 和 Joy A. Thomas，《信息论要义》，第2版，Wiley-Interscience，2006。渐近等分性质和典型集的大小见第3章。[dMU21] Leonardo de Moura 和 Sebastian Ullrich，《Lean 4定理证明器与编程语言》，《自动演绎——CADE 28》，2021年，第625–635页。[Gro81] Mikhael Gromov，《多项式增长群与扩张映射》，《IHÉS数学出版物》53（1981），53–78。[Log25] Logical Intelligence，《Aleph》，2025。[Mat20] Mathlib社区，《Lean数学库》，《第9届ACM SIGPLAN经认证的程序与证明国际会议论文集（CPP 2020）》，2020年，第367–381页。[Mil68] John Milnor，《关于曲率与基本群的注记》，《微分几何杂志》2（1968），第1期，1–7。[Nov55] Pyotr S. Novikov，《论群论中字问题的算法不可解性》，《V.A.斯捷克洛夫数学研究所论文集》44（1955），1–143。[Pos47] Emil L. Post，《Thue问题的递归不可解性》，《符号逻辑杂志》12（1947），第1期，1–11。[Pud86] Pavel Pudlák，《一阶理论中有限主义一致性陈述的证明长度》，《逻辑讨论会 ’84》，1986年，第165–196页。[Pud87] ，《有限主义一致性陈述证明长度的改进界》，《逻辑与组合学》，1987年，第309–331页。[Sha48] Claude E. Shannon，《通信的数学理论》，《贝尔系统技术杂志》27（1948），第3期，379–423。这篇奠基性文献定义了离散信源的熵率和冗余。

附录 A. 补充扩张定理

以下结果完善了表中所概述的 $A_1 = \mathbb{N}$情形 1.

定理 6 (双对数密度给出多项式扩张)。 对于 $A_1 = \mathbb{N}$ 及任意整数 $b \geq 2$，宏集合 $M = \{b^{b^j} : j \geq 0\}$ 具有双对数密度，并满足

$$c_1\, s^{b/(b-1)} \leq f_{G'}(s) \leq c_2\, s^{(2b-1)/(b-1)}$$

对所有 $s \geq 1$成立，其中 c_1、 $c_2 > 0$ 仅取决于 b。

证明。 令 $m_j = b^{b^j}$，其中 $j \geq 0$。长度至多为 r的 G-宏的数量，即满足 $b^{b^j} \leq r$的 j的数量，也就是说，为 $j \leq \log_b \log_b r$。因此， M 具有双对数密度。

的贪心表示法。 不超过 $m_k - 1$ 的最大宏是 m_{k-1}。所使用的副本数量为

$$\left\lfloor \frac{m_k - 1}{m_{k-1}} \right\rfloor = \left\lfloor b^{b^{k-1}(b-1)} - b^{-b^{k-1}} \right\rfloor = b^{b^{k-1}(b-1)} - 1,$$

余数为 $m_{k-1} - 1$。令 $T_k = |m_k - 1|_{G'}$，我们得到

$$T_k = (b^{b^{k-1}(b-1)} - 1) + T_{k-1}, \qquad T_0 = b - 1.$$

首项占主导： $T_k \sim b^{b^{k-1}(b-1)}$。

元素 $m_k - 1$ **最难压缩**。 令 $S_j = \max\{|x|_{G'} : x < m_j\}$。我们断言$S_j = T_j$ 对所有 $j \geq 0$均成立。基础情形 $S_0 = T_0 = b - 1$ 显然成立。对于归纳步骤，假设$S_j = T_j$，并考虑任意满足 $m_j \leq x < m_{j+1}$的 x。将 $x = c_j \cdot m_j + r$ 写作 $0 \leq r < m_j$，则有$|x|_{G'} \leq c_j + T_j$。由于 $c_j \leq b^{b^j(b-1)} - 1$，我们得到 $S_{j+1} \leq T_{j+1}$。当 $x = m_{j+1} - 1$时取等号。

上界。固定 $s \geq b-1$，并令 k 满足 $T_k \leq s < T_{k+1}$。元素 $(s-T_k+1)m_k+(m_k-1)$ 的 G'-长度等于 $(s-T_k+1) + T_k = s+1$，因此它不属于 $B_{G'}(s)$。于是 $f_{G'}(s) < (s+2)m_k$。由递推关系，$T_k \geq b^{b^{k-1}(b-1)} - 1$，因此 $T_k \leq s$ 蕴含 $b^{b^{k-1}(b-1)} \leq s + 1 \leq 2s$。因此， $b^{k-1} \leq \log_b(2s)\, b-1$， 且我们有 $m_k = b^{b \cdot b^{k-1}} \leq (2s)^{b/(b-1)}$。所以 $f_{G'}(s) \leq c_2\, s^{(2b-1)/(b-1)}$。对于 $s < b - 1$，我们只需选取足够大的 c_2。

下界。 如上所述，对于任意 $x = c \cdot m_k + y$，若满足 $0 \leq y < m_k$，则有 $|x|_{G'} \leq c + T_k$。若$c \leq s - T_k$，则 $x \in B_{G'}(s)$，因此 $f_{G'}(s) \geq (s - T_k + 1)m_k - 1$。

情形 $s \leq 2T_k$。 那么 $T_k \geq s/2$，因此 $m_k = b^{b^k} \geq c\, T^{b}{}_{/(b-1)_k} \geq c\,(s/2)$ $^{b/(b-1)} \geq c_1\, s^{b/(b-1)}$。由于$s - T_k + 1 \geq 1$，我们得到 $f_{G'}(s) \geq c\, s^{b/(b-1)}$。

情形 $s > 2T_k$。 那么 $s - T_k + 1 \geq s/2$。由 $s < T_{k+1} \leq C b^{b^k(b-1)}$可得 $b^k \geq \log_b(s/C)b-1$，因此 $m_k = b^{b^k} \geq (s/C)^{1/(b-1)}$。因此 $f_{G'}(s) \geq (s/2)(s/C)^{1/(b-1)} \geq c_1\, s^{b/(b-1)}$。 □

定理7 （有限宏给出线性扩张）。 对于 $A_1 = \mathbb{N}$，设 M 是一个有限宏集合。则 $f_{G'}$（s）$= \Theta$（s）。

证明。 设 $L = \max\{|m|_G : m \in M\}$ 为所有宏中最大的 G-长度。

上界。 任何满足 x with $|x|_{G'} \leq s$ 的元素都是 $G' = G \cup M$ 中至多 s 个生成元之和。每个生成元的 G-长度至多为 L，因此 $|x|_G \leq sL$。于是 $B_{G'}(s) \subseteq B_G(sL)$，从而得到 $f_{G'}(s) \leq sL$。

下界。 由于 $G \subseteq G'$，对于所有 x，我们有 $|x|_{G'} \leq |x|_G$。因此 $B_G(s) \subseteq B_{G'}(s)$，由此可得 $f_{G'}(s) \geq s$。

结合这些界可得 $f_{G'}(s) = \Theta(s)$。 □

注 6。 定理 6 和 7 可扩展到 A_n，其宏集为 $\{b^{b^j} a_i \colon i = 1、\ldots\ldots、n、j \geq 0\}$，以及任意有限的 $M \subseteq A_n$，分别得到相同的渐近扩张率。

Vitaly Aksenov，Logical Intelligence电子邮件地址：
aksenov@logicalintelligence.com

Eve Bodnia , Logical Intelligence
电子邮箱地址: evebodnia@logicalintelligence.com

Michael H. Freedman，Logical Intelligence；数学科学与应用中心，哈佛大学 ，剑桥 ，马萨诸塞州 02138 , USA
电子邮箱地址: michael.freedman@logicalintelligence.com, mfreedman@cmsa.fas.harvard.edu

Michael Mulligan，Logical Intelligence；加利福尼亚大学物理与天文学系，河滨市 ，加利福尼亚州 92521 , USA
电子邮箱地址: michael.mulligan@logicalintelligence.com, michael.mulligan@ucr.edu