2026年2月3日：1–47, 2026

可认证的布尔推理具有普适性

Wenhao Li*
CHRISWENHAO.LI@MAIL.UTORONTO.CA加拿大多伦多大学机械与工业工程系；

Anastasis Kratsios* KRATSIOA@MCMASTER.CA
加拿大麦克马斯特大学数学系及*Vector*研究所；意大利比萨高等师范学校恩尼奥·德·乔治数学研究中心。

Hrad Ghoukasian GHOUKASH@MCMASTER.CA
加拿大麦克马斯特大学数学系及 *Vector*研究所；

Dennis Zvigelsky DENNIS.ZVIGELSKY@MCMASTER.CA
加拿大麦克马斯特大学数学系及 *Vector*研究所；

arXiv:2602.05120v1 [cs.CC] 4 Feb 2026

摘要

智能体系统的激增使人工智能的推理能力成为当代机器学习的焦点。尽管已知存在能够推理任何布尔任务的神经网络 $f\colon \{0,1\}^B \to \{0,1\}$，也就是说，它们能够模拟由扇入 2、扇出 1 的门构成的布尔电路，但研究一再表明，经过训练的模型远未达到这些理论理想。这引出了一个问题：能否展示一种深度学习模型，使其经认证始终能够推理，并且能够通用地推理任何布尔任务？此外，理想情况下，这种模型应只需少量参数便能解决简单的布尔任务。

我们通过展示一种深度学习架构，对布尔电路上的分布进行参数化，并肯定地回答了这一问题。该架构保证，对于每种参数配置，采样得到的几乎必然是有效的布尔电路（因而具有内在的电路级证书）。随后，我们证明了一个普适性定理：对于任意布尔 f，都存在一种参数配置，使得采样得到的电路以任意高的概率计算 f。当 f 是一个 $\mathcal{O}(\log B)$-junta 时，所需参数数量随输入维度 B 线性增长。

在一个与我们的设定相符、受控的真值表补全基准上，实验结果表明，所提出的架构能够可靠地训练，并取得较高的完全匹配准确率，同时保留了预测的结构：每个内部单元在 $\{0,1\}^B$ 上都是布尔值。匹配的多层感知机基线取得了相近的准确率，但只有约 10% 的隐藏单元具有布尔表示；也就是说，它们在布尔立方体上仅取两个值。

关键词：可认证推理、人工智能安全、人工智能推理、布尔电路、通用近似信息，整数可解释性。

* 共同贡献。通讯作者。

© 2026 W. Li、A. Kratsios、H. Ghoukasian 和 D. Zvigelsky。

1. 引言

随着能够跨越较长时间跨度行动的代理模型近期迅速兴起，例如 Shinn 等人（2023）；Wang 等人（2024），以及能够协调工具和结构化搜索的基于人工智能的推理系统，例如 Shen 等人（2023）；Schick 等人（2023）；Yao 等人（2023a,b），还有用于改进解决方案或自举推理依据的迭代式自我改进技术，例如 Madaan 等人（2023）；Wang 等人（2023）；Zelikman 等人（2022）；Chen 等人（2025）；Song 等人（2025），（所谓的）人工智能推理已成为现代机器学习流水线中的常见做法，并日益影响终端用户的决策。人工智能推理模型的这波热潮，建立在深度学习与复杂性理论之间早已确立的联系之上，包括：针对递归／循环网络的图灵机视角 Siegelmann（2012）；P´erez 等人（2019）；Chung 和 Siegelmann（2021）；Mali 等人（2021）；Bournez 等人（2025）；针对非递归架构（例如多层感知机和 Transformer）的布尔值视角 Maass 等人（1991）；Siu 等人（1993）；Merrill 等人（2022）；Merrill 和 Sabharwal（2023）；Chiang（2024），以及一般电路复杂性视角 Lowd 和 Domingos（2008）；Kratsios 等（2025）；此外，还有与数理逻辑的联系 Blumer 等人（1989）；Karpinski 和 Macintyre（1997）；Margulies 和 Morton（2016）；Cook 和 Nguyen（2010）。总的来说，这些结果表明，存在这样的深度学习模型：原则上可以对其进行配置，使其能够推理任意任务；这里的推理指精确模拟某种算法，或等价地，计算某个（布尔值）函数的电路。

相比之下，越来越多的证据表明，大多数在现实世界中训练的人工智能无法完成复杂的多步推理任务 Keysers et al. (2020)；Wei et al. (2022年)；Zhao et al. (2024年)、算术（Cobbe et al., 2021）或更复杂的数学推理（Hendrycks et al., 2021）、初等演绎逻辑 Berglund and et al. (2023年)、归纳推理 Olsson et al. (2022年)；Zhao et al. (2024年)，以及其他若干种虚假推理炫耀 Song et al. (2024年)。实验证据日益表明，经过训练的深度学习流水线会成为对数量极其庞大的训练实例的高效记忆器 Bender et al. (2021)；Wu et al. (2024年)；Shojaee et al. (2025年)。

即便看似无害的训练数据污染Zou etal. (2025年);Chaudhariet al.(2026年)或对抗性提示攻击 Liu et al.(2024年); Rajeev et al. (2025年)已知会严重破坏这些模型的推理。这进一步加重了如下证据：人工智能经训练 于数据 往往无法 进行推理，而且即便能够推理，其思维链也可能被对抗性提示注入破坏。因此，存在非常现实的**人工智能安全关切**；即，如何保证现实世界中的人工智能（而非那些在数学真空中配置的人工智能，例如 Kratsios et al. (2025年))能够可认证地推理。这就引出一个自然的问题：

> 如何构建一种神经网络架构，使其始终具备可认证、普适且高效的电路级解释，同时在现代命题推理基准上保持实际竞争力？

更准确地说，我们寻求一种满足以下条件的架构：

（i）**可认证推理：** 对于网络参数的每一种选择，该模型都会在“简单且可解释”的布尔电路上诱导出有意义的分布，即门具有扇入 2 和扇出 1 的布尔电路。特别地，从该分布中进行的任何一次采样，几乎必然都会得到一个有效且可解释的布尔电路。（ii）**普适性：** 当神经元数量足够多时，可以对参数进行配置，使模型以任意高的概率采样出形如（i）的电路，并计算 B 位上的任意布尔函数。（iii）**效率：** 如果 $f: \{0,1\}^B \to \{0,1\}$ 是稀疏的，那么实现（ii）只需要 $\mathcal{O}(B)$ 个参数。（iv）**竞争性表现：** 该模型能够可靠地训练，并在命题逻辑任务上达到最先进的性能。

我们的主要贡献是构建了一个简单的神经网络模型，能够肯定地满足理论上的期望 (i)–(iii)，同时通过实验验证 (iv)。

1.1. 贡献

对于布尔任务，我们对基于推理的人工智能安全基础领域中的这一敏感问题给出了肯定的回答。我们的解决方案引入了可训练的随机布尔电路，其门的扇入为 2、扇出为 1；也就是说，对于任何参数配置（见图1，并在（11）中作了形式化定义），我们证明模型几乎必然会采样得到一个有意义的布尔电路（定理7）。随后我们表明，只要神经元数量足够多，就可以选择模型参数，使得采样得到的电路通常能以任意高的概率计算任意给定的布尔函数（定理8）；也就是说，该模型具有通用性。对于足够稀疏的布尔函数（即所谓的 junta 函数），可训练参数的数量可以与输入位数成线性关系。

我们的所有结果都经过了实验验证：在补全随机采样的布尔命题的真值表时，多层感知机（MLP）虽然表现出相当的推理能力，但这些模型中大多数（约 80%）神经元都可被验证为并未计算任何布尔门，这使人们对其推理能力产生怀疑（参见定义14）；而我们的人工智能则不会出现这种情况。

1.2 论文结构

第2节汇总了表述和阐释本文结果所需的符号、背景知识和术语。第3 节介绍我们的模型，并阐明推动各项架构创新的理论原则。第4 节给出主要结果，即一项通用计算定理（定理8）、其定量的 $\mathcal{O}$(对数(B))-junta 版本（定理9），以及模型的可认证推理结构，该结构对所有参数配置均有保证（定理7）。随后，我们在第5节通过实验评估这一具有理论依据的架构，并将其与经典深度学习模型进行比较。所有证明及更多实验细节均列于附录。

2. 预备知识

2.1. 设定与记号

以下我们固定一个概率空间 $(\Omega\ \mathcal{F}, \mathbb{P})$，所有随机变量均定义在该空间上。所有随机变量均以粗体表示。我们用 $\mathbb{N}$ 表示非负整数集，用 $\mathbb{N}_+$ 表示正整数集；对于每个 $N \in \mathbb{N}$，记 $\mathbb{N}_N$ 定义为 $= \{n \in \mathbb{N}:\ n \geq N\}$，$[N]$ 定义为$= \{n \in \mathbb{N}:\ n \leq N\}$，以及 $[N]_+$ 定义为$= [N] \cap [1, \infty)$。

对于任意 $n,\ m \in \mathbb{N}_+$，每个 $i \in [n]_+$ 与 $j \in [m]_+$，以及每个 $n \times m$ 矩阵 A，我们将 $A_{i:}$ 记为 A 的第 i^{th}行，并将 $A_{:j}$ 记为其第 j^{th} 列。所有向量均为列向量。对于任意矩阵（或向量），$\|A\|_0$ 统计 A中非零项的个数。由全部扇入 2 且扇出 1 的二元布尔门构成的集合，参见第 A.2 节，记为 $\mathbb{G}^\star$。

固定 $N \in \mathbb{N}_+$，$\mathbf{1}_N \in \mathbb{R}^N$ 表示全 1 向量。我们将 N-单纯形记为 Δ_N 定义$= \{w \in [0,1]^N:\ \mathbf{1}_N^\top w = 1\}$，其相对内部为 $\dot{\Delta}_N$定义$= \{w \in (0,\ 1)^N:\ \mathbf{1}_N^\top w = 1\}$。我们将 *softmax*函数 $\mathsf{SM}_N:\ \mathbb{R}^N \to \Delta_N$ 定义为把任意 $x \in \mathbb{R}^N$ 映射到 $\mathsf{SM}_N(x)$ 定义$= \left(e^{x_n}/\sum_{m=1}^N e^{x_m}\right)_{n=1}^N$。当维数可由上下文确定时，我们简写为 SM。

2.2. 扇入 2、扇出 1 的布尔电路与 AC^0 类

有向无环图 有向图 $G = (V,\ E)$ 是由一个（非空）集合 V（其中包含节点）和一组有序对 $E \subseteq V^2 \setminus \{(v,\ v)\}_{v \in V}$构成的，这些有序对表示有向边。

给定两个节点 v、$w \in V$，v与 w之间的有向路径是一个序列 (v, v_1)、(v_1, v_2)、……、$(v_k, w) \in E$，它“连接 v与 w”，而 $k+1$ 是这条有向路径的长度。如果任意两个不同节点之间都存在有向路径，则称这样的 G 为连通的。若不存在从某个节点到其自身的有向路径，则称 G 为无环。如果 G 连通且无环，我们就称其为连通有向无环图。若两个节点 u、$v \in V$ 之间存在一条从 u 出发并以 v为终点的有向路径，则称它们是偏序的，记作 $u \lesssim v$；参见（Bang-Jensen 和 Gutin，2008，命题 2.1.3）。 G 中不可比较顶点的最大集合的基数称为 G的宽度。 G的深度是 G中最长有向路径的长度。*

相比之下，无向图 $G_u = (V,\ E_u)$ 是由一个（非空）节点集合 V和一组表示边的无序节点对 $E_u \subseteq V^2 \setminus \{(v,\ v)\}_{v \in V}$构成的。若对任意不同的 $v_1,\ v_2 \in V$，都存在一条从 v_1 到 v_2的唯一路径，则称无向图为树。如果一个有向无环图的底层无向图是一棵树，则称其为有向树。没有有向路径终止于其中的节点集合称为根节点／输入节点，而没有有向路径始于其中的节点集合称为叶节点／输出节点。二叉树是一种树，其中每个非根节点 $v \in V$恰好有两个不同的节点 $u,\ w \in V$，使得 $(u,\ v)$，$(w,\ v) \in E$。

布尔电路布尔门 g 的扇入为 2、扇出为 1，是“简单”的基本布尔计算，即与二叉（有向）树结构相容的布尔函数；更准确地说，$g:\ \{0,1\}^2 \to \{0,1\}$；例如，与、或，或提升后的非门 $\mathsf{NOT}(x_1,\ x_2)$ 定义 $= 1 - x_1$。此类门恰有 16 种；参见附录 3 中的 A。

*。对于任意多层感知机计算图底层的有向图而言，这些宽度和深度概念与多层感知机中通常所说的宽度和深度一致。

固定 B、Δ、$\Upsilon \in \mathbb{N}_{+}$。本文中，具有 B 个输入位、深度 Δ 和宽度 Υ 的布尔电路，指一个三元组 $\mathcal{C} = (V, E, g_\cdot)$，其中有向无环图 $G = (V, E)$ 具有 B 个根/输入节点 $v_{\text{in:1}}, \ldots, v_{\text{in:}B} \in V$ 和一个终端/输出节点 $v_{\text{out}} \in V$，并满足：每个非根节点的入度恰为 2；但初始位提升通道除外，该通道用于选择输入位及其否定。位提升通道由映射 $x \mapsto \Pi(x, \mathbf{1}_B - x)$ 形式化，其中 Π 是一个以独热向量为行的矩阵。既非输入节点、也非位提升通道输出的所有节点，称为计算节点，记为 V_{comp} 定义= $V \setminus \{v_{\text{in:}b}\}_{b=1}^{B}$。此外，$g_\cdot = (g_v)_{v \in V_{\text{comp}}}$ 是一个由扇入为 2、扇出为 1 的布尔门构成的族。所有此类布尔电路的集合† 记为 $\text{CIRCUIT}_{\Delta,\Upsilon}^{B}$。由于 $\mathcal{C}$ 的深度为 Δ，其顶点集可分层为 $V = \bigsqcup_{l=0}^{\Delta} V_l$，其中$V_0 = \{v_{\text{in:}b}\}_{b=1}^{B}$, $V_\Delta = \{v_{\text{out}}\}$，使得每条有向边 $(u, v) \in E$ 都满足 $u \in V_l$ 和 $v \in V_{l+1}$，其中某个 $l \in \{0, 1, \ldots, \Delta - 1\}$ 成立。此外，由于 $\mathcal{C}$ 的扇入为 2，每个满足 $l \geq 2$ 的非输入顶点 $v \in V_l$ 都恰有两个前驱节点 $v_{-1}, v_{-2} \in V_{l-1}$，且满足 (v_{-1}, v)、$(v_{-2}, v) \in E$。

布尔推理作为布尔函数的计算采用当前人工智能推理理论的视角 Merrill 等人（2022年）；Chiang （2024年），我们将通过“布尔任务”进行推理理解为：由布尔电路计算一个布尔函数，其形式为 $\cup_{\Delta,\, B,\, \Upsilon \in \mathbb{N}_{+}}$电路$_{\Delta,\, \Upsilon}^{B}$；后者按照电路复杂性理论的标准含义定义，参见 Jukna （2012年），并建立在布尔电路与一阶命题逻辑之间的经典对应关系之上，参见 Cook 和 Nguyen（2010）。固定选择 $B^{\uparrow} \in \mathbb{N}_{+}$ （允许重复的）个输入变量，这些变量取自 $\{x_b\}_{b=1}^{B}$ 及其否定 $\{(1 + x_b) \bmod 2\}_{b=1}^{B}$，其中 $B^{\uparrow} \leq \Upsilon$，等价地由一个 $B^{\uparrow} \times 2B$ 行随机矩阵 Π编码。于是，每个布尔电路 $\mathcal{C} \in$ 电路$_{\Delta,\, \Upsilon}^{B}$ 都会归纳出一个布尔函数 $f_{\mathcal{C}}\colon \{0,1\}^B \to \{0,1\}$ 其递归定义如下：

$$
\begin{aligned}
f_{\mathcal{C}}(x) &\stackrel{\text{def}}{=} f_{\Delta}(x) \\
f_l(x) &\stackrel{\text{def}}{=} \left(g_v\left(f_{l-1}(x)_{v_{-1}}, f_{l-1}(x)_{v_{-2}}\right)\right)_{v \in V_l}, \qquad l = 2, \ldots, \Delta, \\
f_1(x) &\stackrel{\text{def}}{=} x_1 \\
x_1 &\stackrel{\text{def}}{=} \Pi \begin{pmatrix} x \\ \mathbf{1}_B - x \end{pmatrix},
\end{aligned}
\tag{1}
$$

其中 v_{-1} 和 v_{-2} 指的是向量 $f_{l-1}(x)$中对应于 v的两个父节点的索引。我们称布尔函数 $f\colon \{0,1\}^B \to \{0,1\}$ 是可计算的，如果存在一个布尔电路 $\mathcal{C}$，使其诱导的函数 $f_{\mathcal{C}}$ 满足

$$
f_{\mathcal{C}}(x) = f(x) \qquad \text{for every } x \in \{0,1\}^B.
$$

在非一致计算模型／电路复杂性理论中，存在若干<STYLE>复杂性类，例如 AC^k、NC^k、TC^k 等，适用于 $k \in \mathbb{N}$。它们描述的并非哪些布尔函数可计算或不可计算，而是通过限制计算这些函数的电路深度，对计算一个给定的可计算布尔函数有多困难进行分类；这取决于这些电路所用门的“强大”程度

†。该类中允许使用任何扇入 2、扇出 1 的布尔门；也可以考虑仅使用与、或和非门的子类，这不会实质性地改变理论。

<STYLE>所允许的门。例如，AC^k <STYLE>允许使用无界扇入的与、或和非门；NC^k<STYLE>则限制为有界扇入的与 /<STYLE>或 / <STYLE>非门，对应于高度可并行化的电路族；而 TC^k <STYLE>允许使用更强大的门，例如阈值门（多数门），这些门无法通过更初等的布尔门集合进行计算；例如参见 <STYLE>H˚<STYLE>astad<STYLE>（<STYLE>1986<STYLE>），其中给出了多项式时间的相关结果。

我们的模型能够同时确保一组关键的新性质：它几乎必然会从 CIRCUIT 中采样一个随机元素$\overset{B}{\Delta, \Upsilon}$ 对于给定的 Δ， Υ，并且 B (定理7)，每个这样的电路都可以由我们的通用模型采样（定理8）。此外，我们可以通过可微优化训练其权重和偏置，引导它趋向任何给定的目标电路（第 5节）。

2.3. 潜在布尔推理的测试

在继续之前，我们先问自己：“非布尔函数在潜在层面上执行布尔推理，究竟意味着什么？”如果我们能够将这一点形式化，就能轻松检验标准神经网络（例如多层感知机）中的神经元是否在执行某种潜在形式的布尔推理。无论我们的定义是什么，它都应与推理探针背后的理念一致，例如（Alain 和 Bengio，2016；Craven 和 Shavlik，1995；Dai 等人，2022年），并且对于我们的理想化推理器——即布尔电路——始终成立。

我们认为，函数 f 从布尔立方体 $\{0,1\}^B$ 映射到某个欧几里得空间，如果 f 的每个组件都可理解为向 x 的输入赋予布尔值，那么它就具有“布尔神经表示”，也称为一种解释；因此，每个组件都可解释为一个神经元。当然，若要实现这一点， f 的每个潜在神经元（即其中的每个组件）赋予布尔立方体元素的不同取值不能超过两个；否则就必须存在两个以上的真值，这是不可能的。

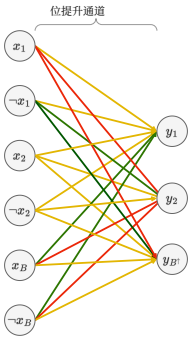
定义 1（布尔神经元可表示性）令 $B_{in}, B_{out} \in \mathbb{N}_+$。若函数 $f : \{0,1\}^{B_{in}} \to \mathbb{R}^{B_{out}}$ 满足以下条件，则称其具有布尔神经表示：对每个 $i \in [B_{out}]_+$，都存在一个单射映射 $\mathcal{R}_i : \mathcal{X}_i \to \{0,1\}$；其中 $\mathcal{X}_i \overset{\text{def}}{=} p_i \circ f(\{0,1\}^{B_{in}})$，且 p_i 表示i^{th} 坐标投影。若存在这样的 $\mathcal{R}_1, \ldots, \mathcal{R}_{B_{out}}$，则称 f 具有布尔神经元可表示性（*BNR*）；任何映射 $\mathcal{R} : \mathbb{R}^{B_{out}} \to \{0,1\}^{B_{out}}$ 都称为 f 的布尔神经表示，若

$$\mathcal{R}(f(x)) = \big(\mathcal{R}_i(f(x)_i)\big)_{i=1}^{B_{out}} \qquad \textit{for all } x \in \{0,1\}^{B_{in}}.$$

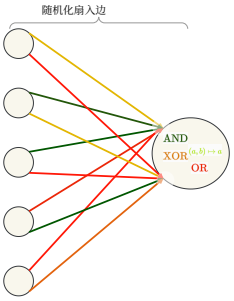
根据构造，每个布尔函数都是 BNR。更一般地， $\{0,1\}$值函数，例如具有受神经科学启发的阈值激活函数的多层感知机，阈值激活函数，例如McCulloch 和 Pitts (1943)，其 BNR 密度也达到 1（见定义14）。相比之下，即使是简单的现代神经元，例如 ReLU 神经元，也很容易不满足 BNR 条件（见例3 和命题12）。

3. 随机布尔电路模型

我们的随机布尔电路模型（图1）由三个组件构成：关于扇入 2、扇出 1 布尔门的参数化分布；关于边对的分布；以及一个随机位提升通道。



(a)随机位提升通道：我们模型的第一层采用随机位提升通道（参见（3））， 从 $B^\uparrow$ 个B 输入位及其否定中进行选择（允许重复）。



(b)随机化布尔神经元：从前一层的节点中恰好选取两条边（扇入 2）（参见（4））；随后随机选择一个布尔门（参见（7）），并计算输出。

图 1：随机布尔电路模型的构建模块：我们的前馈模型定义了一个布尔电路上的分布（定理7），其参数可以设置为以任意高的概率采样到任何所需的电路（定理8）。

阶段 1。 SBC 使用随机位提升通道来选择最相关的 $B^\uparrow$ 位及其否定，并将其传递到网络的其余部分。

阶段 2。 后续每一层中的每个节点都会随机选择恰好两个输入，并应用随机的扇入 2、扇出 1 布尔值计算。

参数控制内容： 其中 红色 边（相应地，门）表示被选中的可能性较低的边，而 绿色 边（相应地，门）表示更有可能被绘制的边。发生这种选择的概率由位提升通道和神经元的可训练权重参数决定。

现在，我们严格定义随机布尔电路模型的组件。

3.1. 模型组件：随机位提升通道

设 $B, B^\uparrow \in \mathbb{N}_+$，并设 $W \in \mathbb{R}^{B^\uparrow \times 2B}$。对每个 $x \in \mathbb{R}^B$，将随机位提升通道 $\mathcal{L}(\cdot|W)$：$\mathbb{R}^B \to \mathbb{R}^{B^\uparrow}$ 定义为

$$\mathcal{L}(x|W) \stackrel{\text{def}}{=} \mathrm{SM}(W)\left(\begin{pmatrix} I_B \\ -I_B \end{pmatrix} x + \begin{pmatrix} \mathbf{0}_B \\ \mathbf{1}_B \end{pmatrix}\right) \tag{2}$$

其中 Softmax 函数 SM 按行作用于矩阵 W。我们用 $\mathcal{L}(\cdot|W)$ 来参数化以下随机变量的分布。具体而言，设 $\mathbf{X}^{\uparrow:W}$ 为取值于 $C(\{0,1\}^B, \{0,1\}^{\uparrow B})$ 的随机变量，其第 i^{th} 个坐标 $\mathbf{X}_i^{\uparrow:W}$ 的分布对任意 $x \in \{0,1\}^B$ 由下式给出

$$\mathbf{X}_i^{\uparrow:W}(x) \sim \sum_{j=1}^{B} \left(\mathrm{SM}_{2B}(W_{i:})\right)_j \delta_{x_j} + \sum_{j=1}^{B} \left(\mathrm{SM}_{2B}(W_{i:})\right)_{j+B} \delta_{\neg x_j}. \tag{3}$$

3.2. 模型组件：随机化布尔神经元

3.2.1. Edges: Trainable Random Fan-in Edges

对于每个 $B \in \mathbb{N}_+$ 以及每个温度参数 $\eta > 0$，定义运算 $\mathrm{S}_\eta: \mathbb{R}^B \times \mathbb{R}^B \to \mathbb{R}^B \times \mathbb{R}^B$，将任意 $(W_1, W_2) \in \mathbb{R}^B \times \mathbb{R}^B$ 映射为

$$\mathrm{S}_\eta\left(W_1, W_2\right) \stackrel{\text{def}}{=} \Big(\mathrm{SM}(\eta W_1), \mathrm{SM}\left(\eta(1 - \mathrm{SM}(\eta W_1)) \odot \mathrm{SM}(\eta W_2)\right)\Big) \tag{4}$$

其中，$\odot$ 表示 $\mathbb{R}^B$中向量对的逐分量（Hadamard）乘积。对于 $i = \{1, 2\}$，$S_\eta(W_1, W_2)_i$ 表示给定布尔二输入门在输入 i 时于 Δ_B 中的分布。

3.2.2. Gates: Trainable Stochastic Boolean Gates

令 $\bar{x}_1, \ldots, \bar{x}_4$ 为 $\{0,1\}^2$ 中的元素，按 (0,0), (0, 1),(1, 0),(1, 1)的顺序排列。对于每个真值赋值 $Z \in \{0,1\}^4$，定义二次插值函数（注意它是可微的） $\tilde{\sigma}: \{0,1\}^2 \to \{0,1\}$，对每个$x \in \{0,1\}^2$ 定义为

$$\tilde{\sigma}(x|Z) \stackrel{\text{def}}{=} \sum_{i=1}^{4} \varsigma(x - \bar{x}_i) Z_i, \quad \text{where} \quad \varsigma(x) \stackrel{\text{def}}{=} \begin{cases} c\, e^{1/(4\|x\|_2^2 - 1)} & \text{if } \|x\|_2 < \frac{1}{2} \\ 0 & \text{if } \|x\|_2 \geq \frac{1}{2} \end{cases} \tag{5}$$

其中 $c \stackrel{\text{定义}}{=} e$。将其叠加起来，我们把（5）扩展为 $\bar{\sigma}$： $\{0,1\}^2 \to \{0,1\}^{16}$ 将任意 $\{0,1\}^2$ 映射到

$$\bar{\sigma}(x) \stackrel{\text{def}}{=} \left(\tilde{\sigma}(x|Z)\right)_{Z \in \{0,1\}^4}. \tag{6}$$

现在，我们构造一个取值于 $\mathbb{R}^{16}$ 的非线性函数，其坐标简单地就是 $(\tilde{\sigma}(x|Z))$，其中 Z 遍历四维布尔超立方体的不同取值。具体来说，我们考虑非线性函数 σ： $\mathbb{R}^2 \to \mathbb{R}^{16}$，对于任意 $x \in \mathbb{R}^2$，定义为

$$\sigma(x) = \left(\tilde{\sigma}(x|Z^{(i)})\right)_{i=1}^{16}, \tag{7}$$

其中 $Z^{(1)} = (0,\ 0,\ 0,\ 0)$……，$Z^{(16)} = (1,\ 1,\ 1,\ 1)$ 是 $\{0,1\}^4$ 的一个枚举。

考虑为任意可训练参数 $W_\Sigma \in \mathbb{R}^{16}$和任意输入 $x \in \mathbb{R}^2$ 定义的随机布尔门选择器： $\Sigma(x \mid W_\Sigma)$ 定义为 $= \mathrm{SM}_{16}(W_\Sigma)^\top \sigma(x)$。该量是……的逐点均值，

任意随机变量 $\boldsymbol{\sigma}^{W_\Sigma} : \Omega \to C^\infty(\mathbb{R}^2, \mathbb{R})$，对任意结果 $\omega \in \Omega$ 定义为

$$\begin{aligned} \boldsymbol{\sigma}^{W_\Sigma}(\omega)(x) &\stackrel{\text{def}}{=} \sum_{i=1}^{16} X_i^{W_\Sigma}(\omega)\, \tilde{\sigma}(x|Z^{(i)}) \\ \text{s.t. } \mathbb{P}(X^{W_\Sigma} = e_i) &\stackrel{\text{def}}{=} \left(\mathrm{SM}_{16}(W_\Sigma)\right)_i \end{aligned} \tag{8}$$

对于每个 $i \in [16]_+$，其中 x 是 $\mathbb{R}^2$中的任意元素。

3.2.3. RANDOM CIRCUIT NETWORK RECURSION

固定深度参数 $\Delta \in \mathbb{N}_+$、宽度 $(B_{in}$、……、$B_{out}) \in \mathbb{N}_+^{\Delta+1}$以及温度参数$\eta \geq 0$。对于每一层 $l \in [\Delta - 1]_+$ 以及每组节点权重 $W^{(l:\sigma)} \in \mathbb{R}^{B_{l+1} \times 16}$，我们将**随机布尔门**/激活函数 $\boldsymbol{\sigma}^{W^{(l:\sigma)}}$: $\mathbb{R}^{B_{l+1}} \times \mathbb{R}^{B_{l+1}} \to \mathbb{R}^{B_{l+1}}$ 定义为将每个 $(x, y) \in \mathbb{R}^{B_{l+1}} \times \mathbb{R}^{B_{l+1}}$ 映射到

$$\boldsymbol{\sigma}^{W^{(l:\sigma)}}(x, y) \stackrel{\text{def}}{=} \left(\boldsymbol{\sigma}^{W_j^{(l:\sigma)}}(x_j, y_j)\right)_{j=1}^{B_{l+1}}. \tag{9}$$

然后，$\boldsymbol{\sigma}^{W^{(l:\sigma)}}$ 将随机布尔电路中实现我们模型的每个节点上的门指定为由可训练矩阵 $W^{(l:\sigma)}$参数化。在每一层 $l \in [\Delta - 1]_+$，随机布尔电路中的边由 $B_{l+1} \times B_l$定义

$$\boldsymbol{E}^{(l:1)} \stackrel{\text{def}}{=} \begin{bmatrix} \mathrm{SM}(\eta W_1^{(l:1)}) \\ \vdots \\ \mathrm{SM}(\eta W_{B_{l+1}}^{(l:1)}) \end{bmatrix} \text{ and } \boldsymbol{E}^{(l:2)} \stackrel{\text{def}}{=} \begin{bmatrix} \mathrm{SM}\left(\eta\big(1 - \mathrm{SM}(\eta W_1^{(l:1)})\big) \odot \mathrm{SM}(\eta\, W_1^{(l:2)})\right) \\ \vdots \\ \mathrm{SM}\left(\eta\big(1 - \mathrm{SM}(\eta W_{B_{l+1}}^{(l:1)})\big) \odot \mathrm{SM}(\eta\, W_{B_{l+1}}^{(l:2)})\right) \end{bmatrix}, \tag{10}$$

其中 $\odot$ 表示向量对上的分量式（Hadamard）乘积，作用于 $\mathbb{R}^{B_l}$。令 $\boldsymbol{E}^{(l:1:e)}$和 $\boldsymbol{E}^{(l:2:e)}$ 表示相应的采样邻接矩阵，其行分布由（10）指定（即对每个 $j \in [B_{l+1}]_+$，j-th 行$\mathbf{E}^{(l:1:e)}$ 和$\mathbf{E}^{(l:2:e)}$ 中的行分别从 $\{e_1, \ldots, e_{B_l}\}$中抽取独热向量，其均值分别由 j^{th}行 $E^{(l:1)}$和 $E^{(l:2)}$给出）。现在我们可以正式定义 SBC 模型。

定义 2（随机布尔电路模型）固定 B_{in}、$B^\uparrow$、Δ、$B_{out} \in \mathbb{N}_+$。随机布尔电路是一个映射 $\hat{} \ \boldsymbol{f} : \Omega \times \{0,1\}^{B_{in}} \to \{0,1\}^{B_{out}}$，通过迭代映射任意$x \in \{0,1\}^{B_{in}}$ 至[‡]来定义。

$$\begin{aligned} \hat{\boldsymbol{f}}(x) &\stackrel{\text{def}}{=} \boldsymbol{x}^{(\Delta)} \\ \boldsymbol{x}^{(l+1)} &\stackrel{\text{def}}{=} \boldsymbol{\sigma}^{W^{(l:\sigma)}}\left(\boldsymbol{E}^{(l:1:e)}\, \boldsymbol{x}^{(l)}, \boldsymbol{E}^{(l:2:e)}\, \boldsymbol{x}^{(l)}\right) \qquad \textit{for } l = 1, \ldots, \Delta - 1 \\ \boldsymbol{x}^{(1)} &\stackrel{\text{def}}{=} \mathbf{X}^{\uparrow:W}(x), \end{aligned} \tag{11}$$

‡. 按照概率论的标准惯例，我们省略随机布尔门$\{\boldsymbol{\sigma}^{W_{(l:\sigma)}}\}_{l=1}^{\Delta-1}$、比特提升通道的随机输出$\mathbf{X}^{\uparrow:W}$以及随机邻接矩阵 $\{\boldsymbol{E}^{(l:j:e)}\}^{\Delta-1,\ l=1\ ,\ j=1}$ 对结果 $\omega \in \Omega$的依赖关系。

其中，随机矩阵 $\boldsymbol{E}^{(l:1:e)}$ 和 $\boldsymbol{E}^{(l:2:e)}$ <style id="7">如 (10)所示，随机布尔门 $\boldsymbol{\sigma}^{W^{(l:\sigma)}}$ 如 (9)所示，随机位提升通道 $\mathbf{X}^{\uparrow:W}$如 (3)所示。对于 $l=1$、.....、 Δ，这些对象的维度应使得定义 $\hat{}$ $\boldsymbol{f}$ 的复合对于每个 $(\omega, x) \in \Omega \times \{0,1\}^{B_{in}}$均有定义。

3.3. SBC模型各组件的说明

现在，我们简要概述所提出的随机布尔电路模型各组件的关键可解释性和表达能力特性，从而说明各组件的参数化方式。

对于任意权重矩阵 W，随机位提升通道 $\mathcal{L}(\cdot \mid W)$ 始终定义一个取值于 $B^{\uparrow}$ 维布尔立方体的随机函数。此外，在训练过程中可以选择 W，使得 $\{0,1\}^{B}$ 中任意所需的变量子集或其否定可重复或省略任意次数。

命题 3（随机位提升通道） 设 $B, B^{\uparrow} \in \mathbb{N}_{+}$，设 $W \in \mathbb{R}^{B^{\uparrow} \times 2B}$，并设$\mathbf{X}^{\uparrow:W}$在 (3) 中定义。则，$\mathbf{X}^{\uparrow:W}(x) \in \{0,1\}^{B^{\uparrow}}$ $\mathbb{P}$-几乎必然。

此外，如果 W 是二元矩阵，且每行只有一个非零项，那么：对于每个 $0 < \delta \leq 1$，都存在某个 $\eta > 0$，使得

$$\mathbb{P}\left(\left(\forall i \in [B]_{+}\right) \#\left\{X_j^{\uparrow:\eta W} = x_i\right\}_{j \in [B^{\uparrow}]_{+}} = \|W_{:i}\|_0 \text{ and } \#\left\{X_j^{\uparrow:\eta W} = \neg\, x_i\right\}_{j \in [B^{\uparrow}]_{+}} = \|W_{:i+B}\|_0 \right) \geq 1-\delta.$$

接下来，我们证明 SBC 中的每个神经元恰好从前一层选择一对入边，并恰好计算 16 个扇入2扇出1布尔门中的一个。

命题 4（仅且任意两条边可被采样） 设 $B \in \mathbb{N}_{+}$。对全部 $\eta > 0$,

（*i*）非奇异概率可解释性：$\mathrm{S}\eta$ 的值域为 $\Delta_B \times \Delta_B$,

（*ii*）光滑性：映射$\mathrm{S}\eta$ 是（C^{∞}-）光滑的。

此外，对每个 $0 < \varepsilon < \frac{1}{2}$，存在温度阈值 $\eta_{\varepsilon} > 0$，使得：对每个$i \in [B]_{+}$,

(iii) ***No Doubled Edges:*** $\sup_{\substack{W_2 \in \mathbb{R}^B \\ \exists\, j \neq i:\ (W_2)_j > (W_2)_i}} \mathrm{S}_{\eta_{\varepsilon}}\left(e_i, W_2\right)_{i:2} \in [0, \varepsilon].$

（*iv*）恢复：若 $j \in [B]_{+}$且 $j \neq i$，则 $\left\|\mathrm{S}_{\eta_{\varepsilon}}\left(e_i,\ e_j\right) - \left(e_i,\ e_j\right)\right\|_1 < \varepsilon$。

特别是，如果$(\mathbf{E}_{i:1}, \mathbf{E}_{i:2}) \sim S_{\eta_{\varepsilon}}(e_i, e_j)$，那么 $\mathbb{P}\big((\mathbf{E}_{i:1}, \mathbf{E}_{i:2}) = (e_i, e_j)\big) \geq 1-\varepsilon$。

我们证明随机布尔门是有意义的；也就是说，它们属于 $\mathbb{G}^{\star}$，并且我们可以将随机布尔门设置为 $\mathbb{G}^{\star}$ 中的特定门，且概率可以任意接近 1。

命题 5（随机布尔门：采样布尔门） 对于每个 $W \in \mathbb{R}^{16}$，$\boldsymbol{\sigma}^{W}|_{\{0,1\}^2} \in \mathbb{G}^{\star}$。

命题 6（随机布尔门：通用采样布尔门）对于每个失败概率 $\delta \in (0,1]$ 以及每个 $i \in [16]_+$，如果 $\eta \geq \ln\left(\frac{15}{\delta} - 15\right)$ 那么

$$\mathrm{TV}\left(\mathrm{Law}(\boldsymbol{\sigma}^{\eta e_i}|_{\{0,1\}^2}), \delta_{g_i}\right) \leq \delta. \tag{12}$$

尤其是，$\mathbb{P}\left(\boldsymbol{\sigma}^{\eta e_i}|_{\{0,1\}^2} = g_i\right) \geq 1 - \delta$。

4. 主要结果

现在，我们已准备好介绍第一个主要结果。该结果证明 随机布尔电路的任何可能参数配置，都会产生一个分布，其取值为扇入 2、扇出 1 的布尔电路。因此，从此类配置中抽取的样本几乎必然是形式 (i) 的此类布尔电路；见第 2 页。

定理 7（可认证的布尔推理）固定一个深度 $\Delta \in \mathbb{N}_+$、一个宽度 $\Upsilon \in \mathbb{N}_+$、位$B_{in} = B_0$、……、$B_\Delta = B_{out} = 1 \in \mathbb{N}_+$。那么，对于任意矩阵 W，$\{W^{(l:\sigma)}\}_{l=1}^{\Delta-1}$ 以及 $\{\mathbf{E}^{(l:1:e)}\}_{l=1}^{\Delta-1}$，$\{\mathbf{E}^{(l:2:e)}\}_{l=1}^{\Delta-1}$，如 (11) 所示，（随机）函数$\hat{}$ $\boldsymbol{f}$ 满足

$$\mathbb{P}\left(\hat{\boldsymbol{f}} \in \mathrm{CIRCUIT}_{\Delta,\Upsilon}^{B_{in}}\right) = 1.$$

在保证随机布尔电路几乎必然会对形式为 (i) 的布尔电路进行采样后，我们现在证明它具有通用性：只要模型使用足够多的神经元，就可以配置其参数，使其以任意高的概率采样任意指定的布尔电路。

定理 8（通用布尔推理）设 $B_{in} \in \mathbb{N}_+$ 且 $f\colon \{0,1\}^{B_{in}} \to \{0,1\}$。对于每个失败概率 $0 < \delta \leq 1$，都存在一个$\hat{}$ $\boldsymbol{f}\colon \Omega \times \{0,1\}^B \to \{0,1\}$，其具有如下表示： (11)

$$\mathbb{P}\left(\hat{\boldsymbol{f}}(x) = f(x)\ \forall x \in \{0,1\}^{B_{in}}\right) \geq 1 - \delta. \tag{13}$$

定理8 是多层感知机经典通用逼近定理在连续函数逼近方面的直接类比，例如Funahashi (1989)；Cybenko (1989)；Hornik 等人 (1989)；Pinkus(1999)；Kidger 和 Lyons (2020)，因为它提供了一种存在性保证，且不受表示（或逼近）真实函数的“难度”影响。自然，人们会想知道是否存在能够实现较优速率的实例；我们现在确认这是可能的。

4.1. 高效推理

事实上，与经典深度学习模型的定量通用逼近结果类似，尽管显式逼近率在最坏情况下是最优的（参见 Yarotsky (2018)），但往往并非最有说服力的结果。相反，核心问题在于参数数量能否随着问题的内在自由度有利地扩展。在经典神经网络

逼近理论中，这种有利的扩展通常是在对目标函数施加结构性假设后获得的，尤其是要求其具备某种形式的光滑性 Yarotsky(2017)；Suzuki (2019)；Lu 等人(2021)；G¨uhring 和 Raslan(2021)；Siegel (2023)。在我们的布尔设定中，与之类似的“良好结构”概念是稀疏性，具体而言，目标函数是一个 S-junta，且满足 $S \ll B$；也就是说，目标函数只依赖于总计 S 个输入位中的 B个。现在我们证明，当稀疏参数 S 随输入数量 B呈对数级增长时，定理 8 中的随机布尔电路所需的神经元数量为多项式级；在某些情况下，除对数因子外甚至为线性级。

定理 *9*（针对 $\mathcal{O}(\log(B))$*-Junta* 函数的高效布尔推理）设 $B, S \in \mathbb{N}_+$，$C > 0$，其中 $S \leq C\log(B)$，并设 $f: \{0,1\}^B \to \{0,1\}$ 是一个 S*-junta*。对于每个 $0 < \delta \leq 1$，都存在一个 $\hat{f}: \Omega \times \{0,1\}^B \to \{0,1\}$，其深度为 Δ、宽度为 Υ，且神经元数量至多为 N，满足 *(13)*；其中

$$\Delta \in \mathcal{O}\big(\log(B) + \log(\log(B))\big), \ \Upsilon \in \mathcal{O}(B), \ \textit{and} \ N, B^{\uparrow} \in \mathcal{O}\big(B^p \log(B)\big)$$

其中 $p = C > 0$。特别地，若 $0 < C \leq 1$，则 $N, B^{\uparrow} \in \mathcal{O}\big(B \text{ 对数}(B)\big)$。

在验证了理论上的预期目标 (i)–(iii) 后，我们现在确认随机布尔电路模型在实验中的表现 (iv)。

5. 实验验证

我们通过实验验证两项主张：(i) 所提出的模型可以经过训练，在命题推理任务上达到有竞争力的准确率；(ii) 其内部计算可以得到忠实的电路级解释。

设置。 每个实例都是一个布尔函数 $f\colon \{0,1\}^B \to \{0,1\}$ 由句法公式指定。我们使用完整真值表（全部 2^B 种赋值）进行训练，并评估 精确匹配 （EM），即预测的真值表必须在所有 2^B 个输入上都与目标一致。我们在三种控制架构或参数预算的匹配机制下，将模型与 ReLU 多层感知机基线进行比较（神经元匹配和两种参数匹配变体）。数据集生成、匹配规则和训练细节见附录 D。

预测性能。 表 1报告了五个随机种子下的精确匹配结果。在这一匹配机制下，匹配后的多层感知机取得了略高的精确匹配率；尽管如此，我们的模型在采用电路结构化计算的同时，仍保持了竞争力。由于精确匹配是在完整真值表上计算的，报告的差异反映的是完全正确与否，而非部分得分指标。

逐神经元布尔性（BNR）。我们衡量布尔神经元可表示性（BNR），其在定义1 中引入，并在命题15中得到刻画。直观而言，若一个单元在布尔立方体上至多取两个不同的值，从而可以单射地重新编码为 $\{0,1\}$，则该单元是 BNR。对于连续基线（ReLU 多层感知机），我们通过对每个单元在完整真值表上枚举，并应用（i）舍入后二值检查（$\text{BNR}_{\text{exact}}$）和（ii）基于容差的双聚类检查（$\text{BNR}_{\varepsilon}$），来计算实用诊断；见附录D.6。

MLP匹配 平均 MLP参数		MLP精确匹配	SBC精确匹配
神经元	368	0.987 ± 0.002	
软参数	683	0.992 ± 0.001	0.971 ± 0.003
总参数	1154	0.999 ± 0.002	

表 1：真值表评估中的精确匹配（EM）（5 个随机种子下的均值 ± 标准差）。SBC精确匹配在标准SBC设置下报告一次；MLP精确匹配针对每种匹配机制分别展示。

模型 / 匹配	$\mathrm{BNR}_{\mathrm{exact}}$(L1)	$\mathrm{BNR}_{\mathrm{exact}}$(全部)	$\mathrm{BNR}_{\varepsilon}$(L1)	$\mathrm{BNR}_{\varepsilon}$(全部)
SBC	1.000	1.000	1.000	1.000
神经元	0.114 ± 0.003	0.209 ± 0.002	0.115 ± 0.004	0.209 ± 0.002
软参数	0.107 ± 0.002	0.199 ± 0.001	0.107 ± 0.002	0.199 ± 0.001
总参数	0.105 ± 0.002	0.195 ± 0.002	0.105 ± 0.002	0.196 ± 0.002

表 2：BNR 诊断（5 个随机种子的均值 ± 标准差）。 “全部”表示对隐藏层的平均值。形式化定义和数值计算程序详见附录 B.1和附录 D.6。

可认证的可解释性差距。表2 显示出明显的分离：我们的模型对全部内部单元均达到 BNR$= 1$，而匹配的 MLP 基线则表现出较低的逐神经元布尔性（在第一隐藏层约为 10%，在各隐藏层上平均为 $\approx 20\%$）。这体现了预期的证书级区别：我们的模型在设计上实现逐神经元布尔计算，而标准 MLP 激活在布尔立方体上通常并非二值。完整的实验方案细节（包括匹配机制、解码和评估）见附录D。

更多门可恢复性探测和门类型直方图见附录E.2 和附录E.3；它们为解读 BNR 和精确匹配提供了补充背景。

6. 结论

我们提出了随机布尔电路（SBC），这是一种可训练的分布，其定义在扇入 2、扇出 1 的布尔电路上。该模型对全部参数均提供可认证推理（定理7），并且通过适当选择参数实现通用布尔推理（定理8）。在稀疏情形下，当目标函数是一个 $\mathcal{O}$（对数 B）-junta 时，这种普适性只需环境维度的近线性扩展即可实现（定理9）。

实验验证了随机布尔电路架构，表明它能够表示多种布尔函数，同时几乎必然采样出扇入 2、扇出 1 的布尔电路。该模型实现了完美的神经元级布尔性（ BNR$= 1$），而匹配的 ReLU 多层感知机虽能达到相似的准确率，却表现出较低的 BNR，这引发了人们对经典架构中电路级推理能力的担忧。

未来工作

未来的工作将探索覆盖更丰富电路类和程序类、且具有可比结构性保证的可训练分布。

致谢

A. Kratsios、H. Ghoukasian 和 D. Zvigelsky 感谢加拿大自然科学与工程研究理事会探索资助（编号 RGPIN-2023-04482 和 DGECR-2023-00230）提供的经费支持。A. Kratsios 还获得了 Bando PRIN 2022 项目的支持，该项目名为“Qnt4Green - 绿色债券市场的定量方法：风险评估、代理问题与政策激励”，项目代码为 2022JRY7EF，CUP 编号为 E53D23006330006，资金由欧盟“下一代欧盟”计划 M4c2 提供。

参考文献

Guillaume Alain 和 Yoshua Bengio。使用线性分类器探针理解中间层。*arXiv* 预印本 *arXiv:1610.01644*，2016年。

Samy Badreddine、Artur d’Avila Garcez、Luciano Serafini 和 Michael Spranger。逻辑张量网络。人工智能，303:103649，2022年。

Jørgen Bang-Jensen 和 Gregory Z Gutin。《有向图：理论、算法与应用》。施普林格科学与商业媒体，2008年。

Emily M. Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。随机鹦鹉的危害：语言模型会不会太大？载于*ACM*公平性、问责制与透明度会议（*FAccT*）论文集，第610–623页，2021年。网址https://arxiv.org/abs/2101.11770。

Leo Berglund 等。逆转诅咒：用“a是b”训练的大语言模型未能学会“b是a”。*arXiv*预印本 *arXiv:2309.12288*，2023年。网址https://arxiv.org/abs/2309.12288。

Anselm Blumer、Andrzej Ehrenfeucht、David Haussler 和 Manfred KWarmuth。可学习性与Vapnik–Chervonenkis维数。《美国计算机学会会刊》（*JACM*），36(4):929–965，1989年。

Olivier Bournez、Johanne Cohen 和 Adrian Wurm。《神经网络的通用一致逼近定理》。载于国际计算机科学数学基础研讨会论文集（*MFCS*），法国帕莱索，2025年。

Harsh Chaudhari、Ethan Rathbun、Hanna Foerster、Jamie Hayes、Matthew Jagielski、Milad Nasr、Ilia Shumailov 和 Alina Oprea。利用语言模型中的推理模式进行间接定向投毒，2026年。网址 https://openreview.net/forum?id=hSLopCTOtT。

Justin Chih-Yao Chen、Archiki Prasad、Swarnadeep Saha、Elias Stengel-Eskin 和 Mohit Bansal。Magicore：用于推理的多智能体迭代式由粗到细精炼。载于 *EMNLP* 会议论文集，2025年。网址 https://aclanthology.org/2025.emnlp-main.1660/。

David Chiang. 统一 TC 中的 Transformer^0。 *arXiv* 预印本，arXiv:2409.13629，2024。Y Choi、Antonio Vergari 和 Guy Van den Broeck。概率电路：一种用于可处理概率模型的统一框架。 加州大学洛杉矶分校。网址：*http://starai. cs. ucla. edu/papers/ProbCirc20. pdf*，第 6 页，2020。Stephen Chung 和 Hava Siegelmann。有界精度循环神经网络的图灵完备性。 神经信息处理系统进展，34：28431–28441，2021。Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano 等。训练验证器以求解数学文字题。载于*arXiv* 预印本 *arXiv:2110.14168*，2021。Stephen A. Cook 和 Phuong Nguyen。 证明复杂性的逻辑基础。逻辑视角。剑桥大学出版社，剑桥，2010。ISBN 9780521517225。Mark Craven 和 Jude Shavlik。提取训练后网络的树状结构表示。载于 D. Touretzky、M.C. Mozer 和 M. Hasselmo 编，神经信息处理系统进展，第 8 卷。麻省理工学院出版社，1995。网址 https://proceedings.neurips.cc/paper_files/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf。George Cybenko。通过 S 型函数的叠加进行逼近。 控制、信号与系统数学，2(4)：303–314，1989。Damai Dai、Li Dong、Yaru Hao、Zhifang Sui、Baobao Chang 和 Furu Wei。预训练 Transformer 中的知识神经元。载于计算语言学协会第 *60* 届年会论文集（第 *1* 卷：长篇论文），第 8493–8502 页，爱尔兰都柏林，2022。计算语言学协会。Kenji Funahashi。神经网络对连续映射的近似实现。神经网络，2(3)：183–192，1989。doi：10.1016/0893-6080(89)90003-8。Ingo G¨uhring 和 Mones Raslan。光滑空间中具有可编码权重的神经网络的逼近率。 神经网络，134：107–130，2021。Johan H˚astad。小深度电路的近乎最优下界。载于 第 *18* 届 *ACM* 计算理论年度研讨会论文集（*STOC*），第 6–20 页。ACM，1986。Dan Hendrycks、Collin Burns、Saurav Kadavath、Akul Arora、Steven Basart、Eric Tang、Dawn Song 和 Jacob Steinhardt。使用 MATH 数据集衡量数学问题求解能力。载于*arXiv* 预印本 *arXiv:2103.03874*，2021。

Kurt Hornik、Maxwell Stinchcombe 和 Halbert White。多层前馈网络是通用逼近器。 神经网络，2(5)：359–366，1989。Stasys Jukna。 布尔函数复杂性：进展与前沿，第 27 卷，属于 算法与组合学。Springer，柏林、海德堡，2012。ISBN 978-3-642-24507-5。Marek Karpinski 和 Angus Macintyre。S 型及一般 Pfaffian 神经网络 VC 维数的多项式界。 计算机与系统科学杂志，54(1)：169–176，1997。Daniel Keysers、Nathanael Sch¨arli、Nathan Scales、Hylke Buisman、Daniel Furrer、Sergey Kashubin、Nikola Momchev、Danila Sinopalnikov、Lukasz Stafiniak、Tomas Tihon 等。衡量组合泛化：一种基于现实数据的综合方法。载于 国际学习表征会议（*ICLR*），2020。网址：https://arxiv.org/abs/1912.09713。Patrick Kidger 和 Terry Lyons。深窄网络的通用逼近。载于 学习理论会议，第 2306–2327 页。PMLR，2020。Anastasis Kratsios、Dennis Zvigelsky 和 Bradd Hart。量化人工智能推理的局限：算法的系统性神经网络表示。 *arXiv* 预印本 *arXiv:2508.18526*，2025。Yupei Liu、Yuqi Jia、Runpeng Geng、Jinyuan Jia 和 Neil Zhenqiang Gong。形式化提示注入攻击与防御措施并对其进行基准测试。载于 第 *33* 届 *USENIX* 安全研讨会（*USENIX Security 24*），第 1831–1847 页，2024。Daniel Lowd 和 Pedro Domingos。学习算术电路。载于第二十四届人工智能不确定性会议论文集，第 383–392 页，2008。Jianfeng Lu、Zuowei Shen、Haizhao Yang 和 Shijun Zhang。光滑函数的深度网络逼近。 *SIAM* 数学分析杂志，53(5)：5465–5506，2021。Wolfgang Maass、Georg Schnitger 和 Eduardo D. Sontag。S 型电路与布尔阈值电路的计算能力比较。载于 第 *32* 届计算机科学基础年度研讨会论文集（*FOCS*），第 767–776 页，1991。Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang 等。自我精炼：利用自我反馈进行迭代精炼。 神经信息处理系统进展，36：46534–46594，2023。Ankur Arjun Mali、Alexander G Ororbia II 和 C Lee Giles。一种神经状态下推自动机。*IEEE* 人工智能汇刊，1(3)：193–205，2021。Robin Manhaeve、Sebastijan Dumancic、Angelika Kimmig、Thomas Demeester 和 Luc De Raedt。DeepProbLog：神经概率逻辑编程。 神经信息处理系统进展，31，2018。

Susan Margulies 和 Jason Morton。通过代数模型和 Pfaffian 电路在多项式时间内求解# CSP 问题。符号计算杂志，74:152–180，2016 年。Warren S. McCulloch 和 Walter Pitts。神经活动中内在观念的逻辑演算。数学生物物理学通报，5(4):115–133，1943 年。William Merrill 和 Ashish Sabharwal。并行性权衡：对数精度 Transformer 的局限性。计算语言学协会汇刊，2023 年。William Merrill、Ashish Sabharwal 和 Noah A. Smith。饱和 Transformer 是常数深度阈值电路。计算语言学协会汇刊，2022 年。Catherine Olsson、Nelson Elhage、Neel Nanda、Nicholas Joseph、Nova DasSarma、Tom Henighan、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen 等。上下文学习与归纳头。arXiv 预印本 arXiv:2209.11895，2022 年。网址 https://arxiv.org/abs/2209.11895。Allan Pinkus。神经网络中 MLP 模型的逼近理论。剑桥应用与计算数学专著。剑桥大学出版社，1999 年。Hoifung Poon 和 Pedro Domingos。和积网络：一种新的深度架构。载于 2011 IEEE 国际计算机视觉会议研讨会（ICCV 研讨会），第 689–690 页。IEEE，2011 年。Jorge P´erez、Javier Marinkovi´c 和 Pablo Barcel´o。论现代神经网络架构的图灵完备性。载于国际学习表征会议，2019 年。网址 https://openreview.net/forum?id=HyGBdo0qFm。Meghana Arakkal Rajeev、Rajkumar Ramamurthy、Prapti Trivedi、Vikas Yadav、Oluwanifemi Bamgbose、Sathwik Tejaswi Madhusudhan、James Zou 和 Nazneen Rajani。猫会混淆推理 LLM：面向推理模型的查询无关对抗触发器。载于第二届语言建模会议，2025 年。网址 https://openreview.net/forum?id=VrEPiN5WhM。Timo Schick、Jane Dwivedi-Yu、Roberto Dess`ı、Roberta Raileanu、Maria Lomeli、Luke Zettlemoyer、Nicola Cancedda 和 Thomas Scialom。Toolformer：语言模型能够学会使用工具。载于神经信息处理系统进展（NeurIPS），2023 年。网址 https://arxiv.org/abs/2302.04761。Yongliang Shen、Lianmin Chen、Xinyu Tang、Zhilin Li、Hui Wang、Rui Qian、Chenliang Li、Wenbo Jiang、Yujie Xiong、Wei Zhang、Zhuang Liu 和 Peng Liu。HuggingGPT：在 Hugging Face 中利用 ChatGPT 及其伙伴解决人工智能任务。载于神经信息处理系统进展（NeurIPS），2023 年。网址 https://arxiv.org/abs/2303.17580。Noah Shinn、Beck Labash、Ashwin Gopinath 和 Karthik R. Narasimhan。Reflexion：采用语言强化学习的语言智能体。载于神经信息处理系统进展（NeurIPS），2023 年。网址 https://arxiv.org/abs/2303.11366。

Parshin Shojaee、Seyed Iman Mirzadeh、Keivan Alizadeh、Maxwell Horton、Samy Bengio 和 Mehrdad Farajtabar。思考的幻觉：通过问题复杂度的视角理解推理模型的优势与局限性。载于 第三十九届神经信息处理系统年会，2025 年。网址 https://openreview.net/forum?id=YghiOusmvw。

Jonathan W. Siegel。深度 ReLU 神经网络在 Sobolev 和 Besov 空间上的最优逼近率。机器学习研究期刊，24(357):1–52，2023年。网址http://jmlr.org/papers/v24/23-0025.html。

Hava T Siegelmann。神经网络与模拟计算：超越图灵极限。施普林格科学与商业媒体，2012年。

Kai-Yeung Siu、Jehoshua Bruck、Thomas Kailath 和 Thomas Hofmeister。用于除法及相关问题的深度高效神经网络。*IEEE* 信息论汇刊，39(3): 946–956，1993年。doi: 10.1109/18.256501。

Peiyang Song、Pengrui Han 和 Noah D. Goodman。大语言模型推理失败综述。*arXiv* 预印本 *arXiv:2410.08944*，2024年。网址 https://arxiv.org/abs/2410.08944。

Zhao Song、Song Yue 和 Jiahao Zhang。思考并非幻觉：通过工具增强克服推理模型的局限性。*arXiv* 预印本 *arXiv:2507.17699*，2025年。

Taiji Suzuki。深度 ReLU 网络在 Besov 空间和混合光滑 Besov 空间中学习的自适应性：最优速率与维数灾难。载于国际学习表征会议，2019年。网址 https://openreview.net/forum?id=H1ebTsActm。

Guanzhi Wang、Yu Wang、Xinyi Liu、Shibin Liu、Yizhou Chen、Yiming Zhang、Jian Zhao 和 Tong Zhang。Voyager：一种具有大语言模型的开放式具身智能体。载于 国际学习表征会议（*ICLR*），2024年。网址https://arxiv.org/abs/2305.16291。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc V. Le、Ed H. Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。自洽性提升语言模型中的思维链推理。载于国际学习表征会议（*ICLR*），2023年。网址https://openreview.net/forum?id=1PL1NIMMrw。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le和Denny Zhou。思维链提示激发大语言模型中的推理。载于神经信息处理系统进展（*NeurIPS*），2022年。网址https://arxiv.org/abs/2201.11903。

Zhaofeng Wu、Linlu Qiu、Alexis Ross、Ekin Aky¨urek、Boyuan Chen、Bailin Wang、Najoung Kim、Jacob Andreas 和 Yoon Kim。《推理还是复述？通过反事实任务探索语言模型的能力与局限——

性。计算语言学协会年会论文集（*ACL*）。计算语言学协会，2024年。Shunyu Yao、Dian Yu、Jeffrey Zhao、Izhak Shafran、Karthik R. Narasimhan 和 Yuan Cao。《思维树：借助大语言模型进行审慎的问题求解》。载于神经信息处理系统进展（*NeurIPS*），2023a。网址：https://arxiv.org/abs/2305.10601。Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik R. Narasimhan 和 Yuan Cao。《React：在语言模型中协同推理与行动》。载于国际学习表征会议（*ICLR*），2023b。网址：https://arxiv.org/abs/2210.03629。Dmitry Yarotsky。《深层 ReLU 网络逼近的误差界》。神经网络，94：103–114，2017年。Dmitry Yarotsky。《超深 ReLU 网络对连续函数的最优逼近》。载于学习理论会议，第639–649页。PMLR，2018年。Eric Zelikman、Yuhuai Wu 和 Noah D Goodman。《Star：自我教学推理器》。载于*NIPS*论文集，第22卷，2022年。Jun Zhao、Yifan Li、Zhen Zhang 和 Yiming Chen。《Mathtrap：通过陷阱问题评估大语言模型的数学推理能力》。载于自然语言处理实证方法会议论文集（*EMNLP*），2024年。网址：https://aclanthology.org/2024.emnlp-main.915。Wei Zou、Runpeng Geng、Binghui Wang 和 Jinyuan Jia。{PoisonedRAG}：针对大语言模型 {检索增强} 生成的知识污染攻击。载于第*34*届 *USENIX* 安全研讨会（*USENIX Security 25*），第3827–3844页，2025年。

附录 x A. 扇入 2、扇出 1 的全部 16 种布尔门表

表3 列出了扇入 2、扇出 1 时所有可能的布尔门。

			$Z^{(i)}$			
门编号	Gate	表达式	(0,0)	(0,1)	(1,0)	(1,1)
g_1	常量假	0	0	0	0	0
g_2	AND	$A \wedge B$	0	0	0	1
g_3	A 与 非 B	$A \wedge \neg B$	0	0	1	0
g_4	PROJ_A	A	0	0	1	1
g_5	非 A 与 B	$\neg A \wedge B$	0	1	0	0
g_6	PROJ_B	B	0	1	0	1
g_7	XOR	$A \oplus B$	0	1	1	0
g_8	OR	$A \vee B$	0	1	1	1
g_9	NOR	$\neg(A \vee B)$	1	0	0	0
g_{10}	XNOR	$A \leftrightarrow B$	1	0	0	1
g_{11}	非 B	$\neg B$	1	0	1	0
g_{12}	A 或 非 B	$A \vee \neg B$	1	0	1	1
g_{13}	非 A	$\neg A$	1	1	0	0
g_{14}	非 A 或 B	$\neg A \vee B$	1	1	0	1
g_{15}	NAND	$\neg(A \wedge B)$	1	1	1	0
g_{16}	常真值	1	1	1	1	1

表 3: 全部初等布尔门的枚举；我们将集合记为 $\mathbb{G}^\star \stackrel{\text{def}}{=} \{g_i\}_{i=1}^{16}$.

A.1. 布尔门的初等性质

A.2. 布尔逻辑门和二元逻辑门

注意，有两个 0 元二元门，即常真值$_{\text{定义}} = 1$ 或假 $\stackrel{\text{def}}{=} 0$ 门。此外，“真正的”一元门只能有两个：恒等门会将任意$x \in \{0,1\}$ 映射到其自身，而否定门 NOT 会将 x 映射为 $1 + x$ 模 (2)。我们的第一个引理指出，每个 0 元门和每个一元门都可以通过与投影 π 复合，实现在 $\mathbb{G}^\star$ 中的某个二元门上：$\{0,1\}^2 \ni (x_1, x_2) \to x_1 \in \{0,1\}$。

引理 10（一元门的嵌入）存在两个常值二元门，分别为 g_1和 g_{16}，见表 3。此外，设 g：$\{0,1\} \to \{0,1\}$ 是一个一元布尔门。那么，存在一个二元布尔门 $g^\uparrow$，使得 $g^\uparrow = g \circ \pi$。因此，我们只需研究二元布尔门，并将对常值（零元）门和一元布尔门的研究纳入其中。

证明 [引理 10] 的证明。根据定义，表 g_1 和 g_{16} 中的门3 是常值门。我们还有 $g_4 = \text{恒等} \circ \pi$ 和 $g_{13} = \text{非} \circ \pi$；一元常真门和常假门也可分别通过与$g_{16}$ 复合 g_1 得到 π。 ■

我们将所有标准布尔门的集合 $\mathbb{G}^{\text{Bool}} \overset{\text{def}}{=} \{\text{与、或、非}\}$ 记作 $\mathbb{G}^\star$的子集；然而，正如下一个结果所示，还存在恰好另外 13 个未包含在 $\mathbb{G}^{\text{Bool}}$ 中的可能二元布尔门，它们列于表 3。

引理11（$\mathbb{G}^\star$的枚举）恰好存在 16 个 2元二元布尔门。此外，这些门列于表3中。

证明 [引理 11] 每个布尔门都是一个映射 $g\colon \{0,1\}^2 \to \{0,1\}$。由于 $\#\{0,1\} = 2$，因此 $\#\{0,1\}^2 = 4$；由此 $\{0,1\}^{\{0,1\}^2}$ 的基数为 $2^4 = 16$。对于第二个陈述，注意表 3中的每个门都有不同的真值赋值。 ■

附录B. 布尔神经元可表示性（BNR）

现在，我们进一步补充一些关于实值函数的布尔神经表示性质的细节。

例1 设 $a > 0$，映射 $f\colon \{0,1\}^2 \ni x \to a \cdot x \in \mathbb{R}^2$ 具有布尔神经表示性质。

例2 映射 $f\colon \{0,1\}^2 \ni x \to (x_1 + x_2, x_1) \in \mathbb{R}^2$ 不具有布尔神经表示性质，因为 $f((0,0))_1 = 0$, $f((0,1))_1 = 1$ 和 $f((1,1))_1 = 2$；这意味着不可能为这三个不同的值赋予两个不同的真值。

例*3*（*ReLU*神经元不一定具有布尔神经表示性质）设 $2 \le B \in \mathbb{N}_+$。设 $0 < a_1 < a_2$ 并定义 $a \overset{\text{def}}{=} (a_i)_{i=1}^B$。因此，$\mathit{ReLU}(a^\top 0) = 0 < \mathit{ReLU}(a^\top e_1) = a_1 < \mathit{ReLU}(a^\top e_2) = a_2$，所以 $f(\cdot) \overset{\text{def}}{=} \mathit{ReLU}(a^\top \cdot + 0)$ 不满足命题 *15 (ii)*；由此 f 不具有布尔神经表示性质。

命题12（随机ReLU神经元以高概率不具备BNR性质）令 $3 \le B \in \mathbb{N}_+$，并令$a \overset{\text{定义为}}{=} (a_i)_{i=1}^B \in \mathbb{R}^B$ 具有独立同分布坐标，其中每个 a_i 均取自关于0对称的无原子分布。定义 $f(\cdot) \overset{\text{定义为}}{=} \text{ReLU}(a^\top \cdot + 0)$。那么，以至少 $1 - \frac{B+2}{2^B}$ 的概率，存在互不相同的指标 i、j、$k \in [B]_+$，使得

$$\text{ReLU}(a^\top e_i) = 0 < \text{ReLU}(a^\top e_j) = a_j < \text{ReLU}(a^\top e_k) = a_k,$$

因此 $f(\cdot)$ 不满足命题*15* （*ii*）；因此 f 不具备*BNR*性质。

证明 [命题 12] 的证明：由于每个 a_i 都是从关于 0 对称的无原子分布中独立同分布抽取的，我们有

$$\mathbb{P}(a_i > 0) = \mathbb{P}(a_i < 0) = \tfrac{1}{2}, \qquad \mathbb{P}(a_i = 0) = 0,$$

此外，由无原子性，$\mathbb{P}(\exists i \neq j : a_i = a_j) = 0$，因此几乎必然，值 a_1、……、a_B 两两不同。

我们有 $a^\top e_i = a_i$ 因此 $\mathrm{ReLU}(a^\top e_i) = \max\{a_i, 0\}$。令 $N_+ \overset{\text{定义}}{=} |\{i \in [B]_+ : a_i > 0\}|$，于是 $N_+ \sim$ 二项分布$(B, \frac{1}{2})$。若 $2 \leq N_+ \leq B-1$，则存在互不相同的 i，j，$k \in [B]_+$ 满足 $a_i < 0 < a_j < a_k$（注意，a_1，……，a_B 几乎必然互不相同），因此

$$\mathrm{ReLU}(a^\top e_i) = 0 < \mathrm{ReLU}(a^\top e_j) = a_j < \mathrm{ReLU}(a^\top e_k) = a_k.$$

因此

$$\mathbb{P}\big(\exists\, i, j, k \text{ distinct s.t. } 0 = \mathrm{ReLU}(a^\top e_i) < \mathrm{ReLU}(a^\top e_j) < \mathrm{ReLU}(a^\top e_k)\big) \;\geq\; \mathbb{P}(2 \leq N_+ \leq B-1).$$

最后，

$$\mathbb{P}(2 \leq N_+ \leq B-1) = 1 - \mathbb{P}(N_+ = 0) - \mathbb{P}(N_+ = 1) - \mathbb{P}(N_+ = B) = 1 - \frac{B+2}{2^B}.$$

因此，以至少 $1 - \frac{B+2}{2^B}$ 的概率，上述不等式成立。因此，$f(\cdot)$ 不满足命题 15(ii)，因此 $f(\cdot)$ 不具备 BNR。 ■

以下给出了 BNR 性质的一个简单刻画，可用作高效判据，以验证某个神经元是否可由布尔神经表示（性质 (ii)），并将其作为 BNR 的规范选择。

命题 13（BNR的刻画——命题 15 的扩展版本） 设 B_{in}，$B_{out} \in \mathbb{N}_+$ 且 $f : \{0,1\}^{B_{in}} \to \mathbb{R}^{B_{out}}$。则以下条件等价：

(i) f 是*BNR*，

(ii) 对每个 $i \in [B_{out}]_+$，$\#\{f(x)_i : x \in \{0,1\}^{B_{in}}\} \leq 2$，

(iii) 映射 $\mathcal{R}_f : \mathbb{R}^{B_{out}} \to \{0,1\}^{B_{out}}$ 将任意 $x \in \mathbb{R}^{B_{out}}$ 映射到

$$\mathcal{R}_f(x) \overset{\text{def}}{=} \big(I(x_i = f(\mathbf{0})_i)\big)_{i=1}^{B_{out}},$$

其中 $\mathbf{0} \overset{\text{定义为}}{=} (0 \ldots 0) \in \{0,1\}^{B_{in}}$，是 f 的布尔神经表示。

当 f 是*BNR*时，我们将 $\mathcal{R}_f$ 称为典范*BNR*。

定义14（BNR密度） 令 $L \in \mathbb{N}_+$，并令 $B_1, \ldots, B_{L+1} \in \mathbb{N}_+$。对于每个 $l \in [L]_+$，令 $f_l : \mathbb{R}^{B_l} \to \mathbb{R}^{B_{l+1}}$，并考虑复合映射 $f \overset{\text{定义为}}{=} f_L \circ \cdots \circ f_1$。定义集合 $\mathcal{D}_0^f \overset{\text{定义为}}{=} \{0,1\}^{B_1}$，并对每个 $l \in [L]_+$，递归定义

$$\mathcal{D}_l^f \overset{\text{def}}{=} f_l(\mathcal{D}_{l-1}^f).$$

f 的 *BNR*密度为 $\mathrm{BNR}(f) \overset{\text{定义为}}{=} \sum_{l=1}^{L} \sum_{i=1}^{B_{l+1}} \frac{I\left(\#\{(f_l(x))_i : x \in \mathcal{D}_{l-1}^f\} \leq 2\right)}{LB_{l+1}}$。

例 4（布尔函数（和门）的 BNR 密度为 1） 设 $L = 1$，$B_1 = B_{in}$，且 $B_2 = B_{out}$。对于每个 $f : \{0,1\}^{B_{in}} \to \{0,1\}^{B_{out}}$ 我们有 $\mathrm{BNR}(f) = 1$。

B.1. 关于 BNR 的证明

证明 [命题的证明 13] 假设 (i)，并且对于每个 $i \in [B_{out}]_+$ 写作

$$S_i \stackrel{\text{def}}{=} \{f(x)_i : x \in \{0,1\}^{B_{in}}\}.$$

如果存在某个 $i \in [B_{out}]_+$，使得 $\#S_i \geq 3$，那么根据鸽巢原理，映射 $\mathcal{R}_i : S_i \to \{0,1\}$ 不可能是单射，这与 BNR 的定义矛盾。这就证明了 (ii)。接下来，假设 (ii) 成立。固定 $i \in [B_{out}]_+$。由于 $\#S_i \leq 2$，S_i 中至多有一个元素不同于 $f(\mathbf{0})_i$。因此，谓词 $x_i = f(\mathbf{0})_i$ 将 S_i 中的元素区分开来，这意味着 $(\mathcal{R}_f)_i\big|_{S_i}$ 是单射。由于 i 是任意的，$\mathcal{R}_f$ 是 f 的布尔神经表示，从而证明了 (iii)。最后，(iii)⇒(i) 的蕴含关系显然成立。 ■

例 5 映射 $f\colon \{0,1\}^2 \ni x \to (x_1 + x_2, x_1) \in \mathbb{R}^2$ 不是 *BNR*，因为 $f((0,0))_1 = 0$、$f((0,1))_1 = 1$ 和 $f((1,1))_1 = 2$；这意味着不可能给这三个不同的值赋予两个不同的真值。

下面给出了布尔神经表示性质的一种简单刻画，可作为高效判据，用于验证某个神经元是否可由布尔神经表示（性质 (ii)），并将其作为 BNR 的标准选择。

命题 15（BNR的刻画） 设 B_{in}、$B_{out} \in \mathbb{N}_+$ 和 $f\colon \{0,1\}^{B_{in}} \to \mathbb{R}^{B_{out}}$。那么，以下条件等价：

(i) f 是 *BNR*,

(ii) *For every* $i \in [B_{out}]_+$, $\#\{f(x)_i : x \in \{0,1\}^{B_{in}}\} \leq 2$.

附录 C. 证明

我们的证明将采用 Kratsios 等人 (2025年) 的手术方法变体。具体而言，我们首先展示一个由以下门构成的最坏情形布尔电路 {与、或、PROJ_B、非}，它可以实现任意给定的 S-junta。随后，我们交换这些门——这就是Kratsios 等人 (2025年) 的手术技术——得到执行相同计算但现在具有高概率的网络。综合起来，这将实现通用推理。

命题 16（初等布尔门的光滑紧支撑穷竭） 对于任意 $x \in \{0,1\}^2$，我们有

$$\sigma(x) = \big(g_i(x)\big)_{i=1}^{16}$$

其中，对每个 $i \in [16]_+$ g_i，它都是表 3中的初等布尔门。此外，σ 是（C^∞-）光滑且紧支撑的。

首先，我们确认，按照 $\boldsymbol{\sigma}^W$ 定义的任何随机布尔门在采样时几乎必然是有效布尔门。

证明 [命题的证明 16] 由于 ς 是多项式函数 $x \mapsto \|2x\|^2$ 与1光滑紧支撑突起函数 $x \mapsto I_{|x|<1} e^{-|x|^2-1}$ 的复合，因此它是光滑的。由于 $\tilde{\sigma}(\cdot|Z)$是这些映射的平移的线性组合，所以每个 $\tilde{\sigma}(\cdot|Z)$ 也是如此，其中 $Z \in \{0,1\}^4$。由于 σ 是紧支撑光滑函数的直和，因此它也光滑且紧支撑。

根据构造，对于每个 $\bar{x}_i \in \{0,1\}^2$ 和每个 $x \in \{0,1\}^2$，我们有 $\varsigma(x-\bar{x}_i) = 1$ 当且仅当$x = \bar{x}_i$，否则 $\varsigma(x-\bar{x}_i) = 0$。因此，对于每个 $x \in \{0,1\}^2$ 和每个 $i \in [4]_+$

$$\tilde{\sigma}(x|Z) = \sum_{i=1}^{4} \varsigma(x-\bar{x}_i) Z_i = \sum_{i=1}^{4} I_{x=\bar{x}_i} Z_i. \tag{14}$$

根据插值恒等式（(14)），我们有：对每个 $i \in [4]_+$，$\tilde{\sigma}(x|Z) = Z_i$ 当且仅当$x = x_i$。

■

证明 [命题的证明 5] 对任意 $W \in \mathbb{R}^{16}$，根据 $\boldsymbol{\sigma}^W$ 的定义，我们有 $\boldsymbol{\sigma}^W \in \{\tilde{\sigma}(\cdot|Z_i)\}_{i\in[16]_+}$。于是该结果由命题16得出。 ■

在以下证明中，我们用 $\mathcal{P}(\mathcal{X})$ 表示度量空间 $\mathcal{X}$ 上的博雷尔概率测度空间。我们用全变差度量 TV 来度量化 $\mathcal{P}(\mathcal{X})$，该度量赋予任意概率测度 $\mathbb{P}$，$\mathbb{Q} \in \mathcal{P}(\mathcal{X})$

$$\mathrm{TV}(\mathbb{P},\mathbb{Q}) \stackrel{\text{def}}{=} \sup_A |\mathbb{P}(A) - \mathbb{Q}(A)|$$

其中上确界取遍 $\mathcal{X}$ 的全部博雷尔子集 A。对于定义在同一概率空间上且分布相同的两个随机变量，我们记作 $X \sim Y$。

证明 [命题 6] 的证明。固定 $\eta > 0$，留待事后确定。由命题16，$\boldsymbol{\sigma}^W|_{\{0,1\}^2} \sim \mathbb{P}^W \underset{\text{定义}}{=} \sum_{i=1}^{16} \mathrm{SM}(W)_i \delta_{g_i}$ 其中，$W = \eta e_i$。特别地，配备度量 TV 的 $\mathbb{P}^W \in \mathcal{P}\big(\{g_i\}_{i\in[16]_+}\big)$ 与 $(\Delta_{16}, \frac{1}{2}\|\cdot\|_1)$ 等距，等距映射为 $\varphi\colon \Delta_{16} \to \mathcal{P}(\mathbb{G}^\star)$ 将任意 $w \in \Delta_{16}$ 映为 $\mathbb{P}_w \underset{\text{定义}}{=} \sum_{i=1}^{16} w_i \delta_{g_i}$。因此，并由 $\mathbb{P}^W$ 的定义，我们有，$\mathrm{TV}(\mathbb{P}^W, \delta_{g_i}) = \frac{1}{2}\|\mathrm{SM}(\eta e_i) - e_i\|_1$。现在，由于 $\|\mathrm{SM}(\eta e_i) - e_i\|_1 = \frac{2}{(G-1)\, e^\eta + (G-1)}$；其中 $G = 16$，则，

$$\mathrm{TV}(\mathbb{P}^{\eta e_i}, \delta_{g_i}) = \frac{1}{2}\|\mathrm{SM}(\eta e_i) - e_i\|_1 = \frac{15}{e^\eta + 15}. \tag{15}$$

因此，如果 $\eta \geq \ln\left(\frac{15}{\delta} - 15\right)$ 则 (15) 意味着 $\mathrm{TV}(\mathbb{P}^{\eta e_i}, \delta_{g_i}) \leq \delta$。根据全变差距离的定义，我们有

$$\mathbb{P}\big(\boldsymbol{\sigma}^{\eta e_i}|_{\{0,1\}^2} = g_i\big) = 1 - \mathrm{TV}(\mathbb{P}^{\eta e_i}, \delta_{g_i}) = 1 - \frac{1}{2}\|\mathrm{SM}(\eta e_i) - e_i\|_1 \geq 1 - \delta$$

完成我们的证明。 ■

例 6 令 $B = 2$ 和 $B^{\uparrow} = 3$。任取一个 $x \in \{0,1\}^2$。对于每个 $\eta > 0$，定义 $W^{\eta} \in \mathbb{R}^{B^{\uparrow} \times 2B}$ 为 $W^{\eta} = \eta \left(e_1, e_3, e_2\right)^{\top}$，其中 (e_1, e_2, e_3, e_4) 表示 $\mathbb{R}^4$ 的标准基。于是，当 $\eta \to \infty$ 时，随机变量 $(\mathbf{X}^{\uparrow:W^{\eta}})_{\eta>0}$ 趋于取值为 $(x_1, \neg x_1, x_2)\mathbb{P}$ 的退化随机变量，几乎处处成立。

证明 [命题 3] 第一个断言显然。对于第二个断言，论证与命题6的证明几乎完全类似，只需作相应修改。 ■

证明 [命题 4] 现在注意到，性质（i）成立，因为 SM 的值域为 Δ_B。

性质（ii）成立，因为 SM 是解析的，哈达玛积、乘法和常数加法也是如此；此外，(C^{∞}-)光滑函数的复合与拼接也满足该性质。

为证明（iii），我们先证明 SM 的两个性质。首先证明 SM 是单调的。也就是说，给定 $W \in \mathbb{R}^B$，满足 $(W)_i < (W)_j$，其中 $i,\ j \in B$ 且 $\eta > 0$,

$$\mathrm{SM}(\eta W)_i < \mathrm{SM}(\eta W)_j. \tag{16}$$

这由 SM 的定义以及 e 的单调性可得：

$$\mathrm{SM}(\eta W)_i = \frac{e^{\eta (W)_i}}{\sum_{k=1}^{B} e^{\eta (W)_k}} < \frac{e^{\eta (W)_j}}{\sum_{k=1}^{B} e^{\eta (W)_k}} = \mathrm{SM}(\eta W)_j. \tag{17}$$

关于第二个性质，再次假设 $(W)_i < (W)_j$。接下来证明，对于全部 $\varepsilon > 0$，存在某个 $\eta > 0$，使得

$$\mathrm{SM}(\eta W)_i < \epsilon. \tag{18}$$

现在由 SM 的定义可得

$$\mathrm{SM}(\eta W)_i = \frac{e^{\eta (W)_i}}{\sum_{k=1}^{B} e^{\eta (W)_k}} < \frac{e^{\eta (W)_i}}{e^{\eta (W)_j}} = e^{-\eta ((W)_j - (W)_i)}. \tag{19}$$

令 $e^{-\eta ((W)_j - (W)_i)} < \varepsilon$ 可得

$$\eta > \frac{\ln (1/\varepsilon)}{(W)_j - (W)_i}. \tag{20}$$

现在转向 (iii)：注意，SM 的定义与 (16) 必然要求

$$\eta (1 - \mathrm{SM}(\eta e_i)_i)\, \mathrm{SM}(\eta (W_2))_i < \eta (1 - \mathrm{SM}(\eta e_i)_j)\, \mathrm{SM}(\eta (W_2))_j, \tag{21}$$

因此由 (18)，给定 $\varepsilon > 0$，存在 $\eta^{\star:1} > 0$ 使得对全部 $\eta > \eta^{\star:1}$ 该断言成立。

现在我们证明 (iv)。对于每个 $k \in [B]_+ \setminus \{i\}$,

$$\lim_{\eta \to \infty} \mathrm{SM}(\eta e_i)_k = \lim_{\eta \to \infty} \frac{1}{(B-1)e^{\eta 0} + e^{\eta}} = \lim_{\eta \to \infty} \frac{1}{(B-1) + e^{\eta}} = 0. \tag{22}$$

由于 SM 的值域属于 B-单纯形，因此

$$\lim_{\eta\to\infty} \mathrm{SM}(\eta e_i)_i = \lim_{\eta\to\infty} 1 - \sum_{k>1}^{B} \mathrm{SM}(\eta e_i)_k = \lim_{\eta\to\infty} 1 - 0 = 1. \tag{23}$$

合起来（23）得出 并且（22）意味着存在某个 $\eta^{\star:2} > 0$ 使得：对于所有 $\eta \geq \eta^{\star:2}$ 都有

$$\| \mathrm{SM}(\eta e_i) - e_i \|_1 < \tfrac{\varepsilon}{2}. \tag{24}$$

现在取 $w = e_j$ 使得 $j \neq i$。化简 $\mathrm{SM}\left(\eta\big((1 - \mathrm{SM}(\eta e_i)) \odot \mathrm{SM}(\eta e_j)\big)\right)_j$ 后，我们发现它等于

$$\frac{\exp\left(\eta \frac{e^\eta(e^\eta+B-2)}{(e^\eta+(B-1))^2}\right)}{\exp\left(\eta \frac{B-1}{(e^\eta+(B-1))^2}\right) + \exp\left(\eta \frac{e^\eta(e^\eta+B-2)}{(e^\eta+(B-1))^2}\right) + (B-2)\exp\left(\eta \frac{e^\eta+B-2}{(e^\eta+(B-1))^2}\right)}. \tag{25}$$

现在令（25）中的 η 趋于无穷大，得到

$$\lim_{\eta\to\infty} \mathrm{SM}\left(\eta\big((1 - \mathrm{SM}(\eta e_i)) \odot \mathrm{SM}(\eta e_j)\big)\right)_j \tag{26}$$

$$= \lim_{\eta\to\infty} \frac{\exp\left(\eta \frac{e^\eta(e^\eta+B-2)}{(e^\eta+(B-1))^2}\right)}{\exp\left(\eta \frac{B-1}{(e^\eta+(B-1))^2}\right) + \exp\left(\eta \frac{e^\eta(e^\eta+B-2)}{(e^\eta+(B-1))^2}\right) + (B-2)\exp\left(\eta \frac{e^\eta+B-2}{(e^\eta+(B-1))^2}\right)} = 1. \tag{27}$$

由于 SM 连续且映射到 Δ_B 这一闭集，因此其极限也必属于B-单纯形。然而，如果 $\lim_{\eta\to\infty} \mathrm{SM}\left(\eta\big((1 - \mathrm{SM}(\eta e_i)) \odot \mathrm{SM}(\eta e_j)\big)\right)_j$ 且极限的每个坐标都非负，则有

$$\mathrm{SM}\left(\eta\big((1 - \mathrm{SM}(\eta e_i)) \odot \mathrm{SM}(\eta e_j)\big)\right)_i$$

对每个 $i \in [B]_+ \setminus \{j\}$都成立。因此，对于所有 $\varepsilon > 0$，存在某个 $\eta^{\star:3} > 0$ 使得，若 $\eta > \eta^{\star}$ 定义= 最大值$\{\eta^{\star:1}, \eta^{\star:2}, \eta^{\star:3}\}$ 则有

$$\left\| \mathrm{SM}\left(\eta\big((1 - \mathrm{SM}(\eta e_i)) \odot \mathrm{SM}(\eta e_j)\big)\right) - e_j \right\|_1 < \tfrac{\varepsilon}{2}. \tag{28}$$

结合（24）与（28）可得（ii），从而完成我们的证明。

令 $(\mathbf{E}_{i:1}, \mathbf{E}_{i:2}) \sim S_\eta(e_i, e_j)$。定义分歧集 $A^\star$ 如下：

$$A^\star \stackrel{\text{def}}{=} \{\omega \in \Omega : (\mathbf{E}_{i:1}(\omega), \mathbf{E}_{i:2}(\omega)) \neq (e_i, e_j)\}.$$

注意， $A^\star$ 是可测的，因为 $\mathbf{E}^{i:j}$ 对 $j \in \{1,2\}$ 可测，而且 $\mathbb{R}^2 \ni (x,y) \mapsto I_{x=y} \in \{0,1\}$ 是博雷尔集。现在，对于最后一个结论，根据全变差距离的对偶表示以及 (ii)，我们有

$$\begin{aligned} \left\| \operatorname{Law}\left((\mathbf{E}_{i:1}, \mathbf{E}_{i:2})\right)(A^\star) - \delta_{(e_i,e_j)}(A^\star) \right\| &\le \sup_A \left\| \operatorname{Law}\left((\mathbf{E}_{i:1}, \mathbf{E}_{i:2})\right)(A) - \delta_{(e_i,e_j)}(A) \right\| \\ &= \left\| \operatorname{Law}\left((\mathbf{E}_{i:1}, \mathbf{E}_{i:2})\right) - \delta_{(e_i,e_j)} \right\|_{TV} \\ &= \left\| (\mathbf{E}_{i:1}, \mathbf{E}_{i:2}) - (e_i, e_j) \right\|_1 \le \varepsilon \end{aligned} \tag{29}$$

其中，上确界取遍所有可测集合 A。因此，对（29）两边取补集即可得到最终结论。

■

证明 [定理 7] 的证明。回顾一下，（11）中的递归以 $l = 1,\ldots,\Delta-1$ 为索引，因此对于每个 $l \in \{1,\ldots,\Delta-1\}$，层表示满足 $\boldsymbol{x}^{(l)} \in \{0,1\}^{B_l}$ 且 $\boldsymbol{x}^{(l+1)} \in \{0,1\}^{B_{l+1}}$.相应地，对于每个 $l \in \{1,\ldots,\Delta-1\}$ 和 $j \in \{1,2\}$，随机邻接矩阵 $\boldsymbol{E}^{(l:j:e)}$ 是一个$B_{l+1} \times B_l$ 矩阵，其各行几乎必然为独热向量；等价地，

$$\boldsymbol{E}^{(l:j:e)}_{i:} \in \{e_1,\ldots,e_{B_l}\} \quad \forall i \in [B_{l+1}]_+, \ \mathbb{P}\text{-a.s.}$$

相比之下，从原始输入层 $l = 0$ 到提升层 $l = 1$ 的转换，是由随机位提升通道 $\mathbf{X}^{\uparrow:W}$ 生成的（而非由矩阵 $\boldsymbol{E}^{(l:j:e)}$生成）。

固定 $\Delta \in \mathbb{N}_+$、$\Upsilon \in \mathbb{N}_+$ 和宽度 $B_{in} = B_0$、B_1、……、$B_\Delta = B_{out} = 1$，其中 $B_l \le \Upsilon$ 对于 $l \in [\Delta]$ 成立。令 $\hat{} \boldsymbol{f}$ 由式（11）定义。我们证明，对于 ω 中的 $\mathbb{P}$-几乎处处情形，存在 $\mathcal{C}_\omega \in$ 电路$^{B_{in}}_{\Delta,\Upsilon}$ 使得$\hat{} \boldsymbol{f}(\omega, x) = f_{\mathcal{C}_\omega}(x)$ 对所有 $x \in \{0,1\}^{B_{in}}$ 均成立。

步骤 1：一个使全部模块均为布尔值的概率为 1 的事件。模型中的所有随机对象都定义在一个固定的概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 上。具体而言：提升输入映射的取值属于有限集合 $\{0,1\}^{B_1}$。此外，对于每个 $l \in \{1,\ldots,\Delta-1\}$ 和 $j \in \{1,2\}$，每个邻接矩阵 $\boldsymbol{E}^{(l:j:e)}$ 的行的取值都属于有限集合 $\{e_1,\ldots,e_{B_l}\}$，且每个门限制都属于有限集合 $\mathbb{G}^\star$。因此，下文定义中涉及的所有随机变量均为有限值，任何由其取值约束定义的事件都是 $\mathcal{F}$-可测的。

令 $\Omega_{\text{lift}} \subseteq \Omega$ 为命题 3 中的事件，使得对于每个 $\omega \in \Omega_{\text{lift}}$ 和每个输入$x \in \{0,1\}^{B_{in}}$,

$$\mathbf{X}^{\uparrow:W}(\omega)(x) \in \{0,1\}^{B_1}.$$

由于提升输入映射是有限值的，$\Omega_{\text{lift}} \in \mathcal{F}$，且命题 3 推出 $\mathbb{P}(\Omega_{\text{lift}}) = 1$。

令 $\Omega_{\text{gate}} \subseteq \Omega$ 为命题 5 中的事件，使得对于每个 $\omega \in \Omega_{\text{gate}}$、每个 $l \in \{1,\ldots,\Delta-1\}$ 以及每个节点索引 $j \in \{1,\ldots,B_{l+1}\}$，采样激活映射的限制

$$\boldsymbol{\sigma}^{W_j^{(l:\sigma)}}(\omega)\big|_{\{0,1\}^2}$$

属于 $\mathbb{G}^\star$，即在 $\{0,1\}^2$上定义了一个有意义的布尔门。由于该限制的取值属于有限集合 $\mathbb{G}^\star$，事件 Ω_{gate} 是 $\mathcal{F}$- 可测的。根据命题 5，我们有 $\mathbb{P}(\Omega_{\text{gate}}) = 1$。

最后，对于每个 $l \in \{1, \dots, \Delta - 1\}$，随机邻接矩阵 $\boldsymbol{E}^{(l:1:e)}$ 和 $\boldsymbol{E}^{(l:2:e)}$ 为层 $l+1$中的每个节点恰好选取层 l 中的一个前驱节点，即每一行都是独热的。令 $\Omega_{\text{edge}} \subseteq \Omega$ 表示满足以下条件的事件：对于每个 $l \in \{1, \dots, \Delta - 1\}$、$j \in \{1, 2\}$ 以及每个 $i \in [B_{l+1}]_+$，

$$\boldsymbol{E}_{i:}^{(l:j:e)}(\omega) \in \{e_1, \dots, e_{B_l}\}.$$

由于每个邻接矩阵行的取值都属于有限集合，$\Omega_{\text{edge}} \in \mathcal{F}$。根据邻接矩阵的假设，$\mathbb{P}(\Omega_{\text{edge}}) = 1$。

定义

$$\Omega_{\text{all}} \stackrel{\text{def}}{=} \Omega_{\text{lift}} \cap \Omega_{\text{gate}} \cap \Omega_{\text{edge}}.$$

于是 $\Omega_{\text{all}} \in \mathcal{F}$，且 $\mathbb{P}(\Omega_{\text{all}}) = 1$。在证明余下部分中，固定任意 $\omega \in \Omega_{\text{all}}$。我们构造一个电路 $\mathcal{C}_\omega \in \text{CIRCUIT}_\Delta^{B_{in},\,\Upsilon}$，用于计算 $x \mapsto \hat{\boldsymbol{f}}(\omega,\ x)$。

Step 2：构造实现电路 $\mathcal{C}_\omega$。 我们构建一个分层有向无环图，其层 s

$$V_0 = \{v_{0,1}, \dots, v_{0,B_0}\},\ V_1 = \{v_{1,1}, \dots, v_{1,B_1}\},\ \dots,\ V_\Delta = \{v_{\Delta,1}\}, \tag{30}$$

使得层 V_0 对应于输入 $x \in \{0,1\}^{B_0}$，而对于 $l \in \{1, \dots, \Delta\}$，层 V_l 对应于模型状态 $\boldsymbol{x}^{(l)} \in \{0,1\}^{B_l}$。

令 $\text{val}_{\mathcal{C}_\omega}(v_{l,i};\ x)$ 表示分层有向无环图中第 i 个节点在第 l 层的输出值，它是在输入 x处、诱导布尔函数 $f_{\mathcal{C}_\omega}$ 中计算得到的（对应于节点 $v_{l,i}$，如式 (30) 所示）。特别地，$\text{val}_{\mathcal{C}_\omega}(v_{\Delta,1};\ x)$ 对应于式 (1) 中诱导布尔函数 $f_{\mathcal{C}_\omega}(x) = f_\Delta(x)$ 的输出值。接下来，我们从输入层开始定义这些值。

输入值。 对于 $x \in \{0,1\}^{B_{in}}$，将原始输入节点处的值定义为

$$\text{val}_{\mathcal{C}_\omega}(v_{0,i}; x) \stackrel{\text{def}}{=} x_i \qquad \forall i \in \{1, \dots, B_0\}. \tag{31}$$

提升层的值。 对于 $x \in \{0,1\}^{B_{in}}$，将 $v_{1,\ j}$ 处的值定义为

$$\text{val}_{\mathcal{C}_\omega}(v_{1,j}; x) \stackrel{\text{def}}{=} \boldsymbol{x}_j^{(1)}(\omega, x) = \mathbf{X}^{\uparrow:W}(\omega)(x)_j \in \{0,1\} \qquad \forall j \in \{1, \cdots, B_1\}, \tag{32}$$

使用 $\omega \in \Omega_{\text{lift}}$。

提升边（第 $0 \to 1$ 层）。 固定 $\omega \in \Omega_{\text{lift}}$。根据提升通道 $\mathbf{X}^{\uparrow:W}$ 的定义，存在一个（随机的）二进制矩阵 $W(\omega) \in \{0,1\}^{B_1 \times 2B_0}$，使得对于每个输入 $x \in \{0,1\}^{B_0}$，通过选取扩展输入向量 $\begin{pmatrix} x \\ \mathbf{1}_{B_0} - x \end{pmatrix} \in \{0,1\}^{2B_0}$的坐标，可以得到提升表示 $\boldsymbol{x}^{(1)}(\omega,\ x)$。特别地，对于每个 $j \in \{1, \dots, B_1\}$，$W(\omega)$ 的第 j 行会选取唯一的索引 $q(j;\omega) \in \{1, \dots, 2B_0\}$。

因此，对每个 $j \in \{1, \dots, B_1\}$，将相应的原始索引 $i(j;\ \omega) \in \{1, \dots, B_0\}$ 定义为

$$i(j;\omega) \stackrel{\text{def}}{=} \begin{cases} q(j;\omega), & q(j;\omega) \le B_0, \\ q(j;\omega) - B_0, & q(j;\omega) > B_0. \end{cases}$$

然后，我们通过添加有向边来记录底层的原始坐标

$$\big(v_{0,i(j;\omega)},\, v_{1,j}\big).$$

此处的边标识原始索引 $i(j;\omega)$。

请注意，第 V_1 层的值仍由（32）定义；从 V_0 到 V_1 的边编码了由 W（ω）所诱导的提升后的比特与原始输入坐标之间的对应关系。

边。对于每个 $l \in \{1,\dots,\Delta-1\}$ 和每个节点索引 $j \in \{1,\dots,B_{l+1}\}$，由于 $\omega \in \Omega^{\text{edge}}$，$\boldsymbol{E}^{(l:1:e)}(\omega)$ 的第 j 行是独热的，因此它选择唯一的索引 $p_1(l,\ j;\omega) \in \{1,\dots,B_l\}$；类似地，$\boldsymbol{E}^{(l:2:e)}(\omega)$ 的第 j 行选择唯一的 $p_2(l,\ j;\omega) \in \{1,\dots,B_l\}$，添加有向边

$$\big(v_{l,p_1(l,j;\omega)},\, v_{l+1,j}\big), \qquad \big(v_{l,p_2(l,j;\omega)},\, v_{l+1,j}\big).$$

因此，对于 $l \geq 2$ 的 V_l 层中的每个节点，其扇入都恰好为 2，边仅出现在相邻层之间，并且从 V_0 到 V_1 的边编码了由提升映射所诱导的对应关系。

门。对于每个 $l \in \{1,\dots,\Delta-1\}$ 和每个 $j \in \{1,\dots,B_{l+1}\}$，将布尔门指派给节点 $v_{l+1,\ j}$

$$g_{l+1,j} \overset{\text{def}}{=} \boldsymbol{\sigma}^{W_j^{(l:\sigma)}}(\omega)\big|_{\{0,1\}^2} \in \mathbb{G}^\star,$$

使用 $\omega \in \Omega^{\text{gate}}$。

实现电路的定义。与 $\omega \in \Omega_{\text{all}}$ 关联的实现布尔电路是三元组 $\mathcal{C}_\omega \text{ 定义为}= (V,\ E,\ g.)$，其中：

- The vertex set is $V \overset{\text{def}}{=} \bigsqcup_{l=0}^{\Delta} V_l$ with V_l defined as in (30).
- 边集为

$$E \overset{\text{def}}{=} \Big(\bigcup_{j=1}^{B_1}\{(v_{0,i(j;\omega)},\, v_{1,j})\}\Big) \cup \Big(\bigcup_{l=1}^{\Delta-1}\bigcup_{j=1}^{B_{l+1}}\big\{(v_{l,p_1(l,j;\omega)},\, v_{l+1,j}),(v_{l,p_2(l,j;\omega)},\, v_{l+1,j})\big\}\Big).$$

- 门指派为 $g. \text{ 定义为}= \{\, g_{l+1,j} :\ \{0,1\}^2 \to \{0,1\}\,\}_{l=1,\ \dots,\ \Delta-1;\ \ j=1,\ \dots,\ B_{l+1}}$.

这定义了一个深度为 Δ、宽度至多为 $\max_{0\leq l\leq\Delta} B_l \leq \Upsilon$ 的布尔电路 $\mathcal{C}_\omega = (V,\ E,\ g.)$。

步骤 3：证明 $\mathcal{C}_\omega$ 计算 $\hat{}\ \boldsymbol{f}(\omega,\cdot)$（深度归纳法）。我们证明，对于每个$l \in \{1,\dots,\Delta\}$、每个输入 $x \in \{0,1\}^{B_{in}}$ 以及每个节点索引 $j \in \{1,\dots,B_l\}$，

$$\text{val}_{\mathcal{C}_\omega}(v_{l,j};x) = \boldsymbol{x}_j^{(l)}(\omega,x). \tag{33}$$

基本情形（$l=1$）。根据输入层和提升后的输入通道的定义 l,

$$\text{val}_{\mathcal{C}_\omega}(v_{1,j};x) \overset{\text{def}}{=} \boldsymbol{x}_j^{(1)}(\omega,x) = \mathbf{X}^{\uparrow:W}(\omega)(x)_j \in \{0,1\} \qquad \forall j \in \{1,\cdots,B_1\}$$

因此，（33）对 $l=1$ 成立（见（32））。

归纳假设。假设对于某个 $l \in \{1,\dots,\Delta-1\}$，式（33）对层 l中的全部节点成立，即

$$\mathrm{val}_{\mathcal{C}_\omega}(v_{l,j};x) = \boldsymbol{x}_j^{(l)}(\omega,x) \quad \forall j \in \{1,\dots,B_l\},\ \forall x \in \{0,1\}^{B_{in}}.$$

归纳步骤。固定任意节点索引 $j \in \{1,\dots,B_{l+1}\}$ 和输入 $x \in \{0,1\}^{B_{in}}$。根据 $\mathcal{C}_\omega$的构造，

$$\mathrm{val}_{\mathcal{C}_\omega}(v_{l+1,j};x) = g_{l+1,j}\Big(\mathrm{val}_{\mathcal{C}_\omega}(v_{l,p_1(l,j;\omega)};x), \mathrm{val}_{\mathcal{C}_\omega}(v_{l,p_2(l,j;\omega)};x)\Big).$$

根据归纳假设，

$$\mathrm{val}_{\mathcal{C}_\omega}(v_{l,p_i(l,j;\omega)};x) = \boldsymbol{x}^{(l)}_{p_i(l,j;\omega)}(\omega,x), \qquad i=1,2.$$

由于 $g_{l+1,\ j} = \boldsymbol{\sigma}^{W_j^{(l:\sigma)}}(\omega)\big|_{\{0,1\}^2}$，且 $\boldsymbol{E}^{(l:1:e)}(\omega)$ 和 $\boldsymbol{E}^{(l:2:e)}(\omega)$ 的独热行恰好选择了前驱节点 p_1（l，j；ω）和 p_2（l，j；ω，递归（11）可推出

$$\boldsymbol{x}_j^{(l+1)}(\omega,x) = \boldsymbol{\sigma}^{W_j^{(l:\sigma)}}(\omega)\Big(\boldsymbol{x}^{(l)}_{p_1(l,j;\omega)}(\omega,x), \boldsymbol{x}^{(l)}_{p_2(l,j;\omega)}(\omega,x)\Big).$$

Thus

$$\mathrm{val}_{\mathcal{C}_\omega}(v_{l+1,j};x) = \boldsymbol{x}_j^{(l+1)}(\omega,x).$$

由于 j 是任意的，式（33）对层 $l+1$中的全部节点成立。归纳证明完成。

步骤 4：最终结论。将 $l=\Delta$ 代入式（33），并回顾 $B_\Delta = 1$，对于每个输入$x \in \{0,1\}^{B_{in}}$，

$$\hat{\boldsymbol{f}}(\omega,x) = \boldsymbol{x}^{(\Delta)}(\omega,x) = \mathrm{val}_{\mathcal{C}_\omega}(v_{\Delta,1};x),$$

因此，$\hat{\ }\boldsymbol{f}(\omega,\ \cdot)$ 由深度为 Δ、宽度为 $\le \Upsilon$ 的布尔电路 $\mathcal{C}_\omega$ 计算得到。因此，对于每个 $\omega \in \Omega_{\text{all}}$，$\hat{\ }\boldsymbol{f}(\omega,\ \cdot) \in$ 电路$_{\Delta,\ \Upsilon}^{B_{in}}$。由于 $\mathbb{P}(\Omega_{\text{all}}) = 1$，可得

$$\mathbb{P}\big(\hat{\boldsymbol{f}} \in \mathrm{CIRCUIT}_{\Delta,\Upsilon}^{B_{in}}\big) = 1.$$

证明完毕。 ■

C.1. 树构造

我们从布尔函数的一种最坏情形表示开始：将其表示为一棵树，其叶节点是布尔立方体中的文字、它们的否定，以及每个文字可能重复出现的副本。此外，该电路由一棵满二叉树实现，因此具有所需的精确多部结构：每个内部节点都是一个扇入为2、扇出为1的与门、或门或 PROJ_A 门。这样便避免了电路表示带来的复杂性：其计算图可能不同于我们的神经网络的计算图，或者在内部节点处混合一元门和二元门，从而形成不兼容的计算结构。

Lemma 17（布尔函数的完全二叉树表示s）
设 $B \in \mathbb{N}_+$ 是一个布尔函数 $f: \{0,1\}^B \to \{0,1\}$ 存在整数 $B^{\uparrow}$、$\Delta \in \mathbb{N}_+$ 以及一个计算 f 的布尔电路 $\mathcal{C}$，满足：

(i) $\mathcal{C}$ 的底层有向无环图是一棵深度为 Δ 的有根完全二叉树，

(ii) $B^{\uparrow}$ 的叶节点标记为 $\{x_1, \dots, x_B, \neg x_1, \dots, \neg x_B\}$中的文字，允许重复；即对于每个 $\tilde{b} \in [B^{\uparrow}]_+$，都有

$$x_{\tilde{b}} \in \{x_b\}_{b=1}^{B} \cup \{\neg x_b\}_{b=1}^{B},$$

(iii) $\mathcal{C}$ 中的每个门都属于 $\{\text{或}、\text{与}、\text{投影}_A\} = \{g_8, g_2, g_4\}$。

此外，门的数量恰好为 $B^{\uparrow} - 1 \le 2B2^B - 1$，且

$$B^{\uparrow} \le 2B2^B, \qquad \Delta \le \lceil \log_2(B2^B) \rceil \le B + \lceil \log_2 B \rceil. \tag{34}$$

现在，我们将前面的结果进一步推广到在布尔意义上可能“稀疏”的函数。

定义 18（S-Junta）若存在输入 x、$y \in \{0,1\}^B$，它们仅在第 i 个坐标上不同，且满足f（x）$\neq f$（y），则称布尔函数 f：$\{0,1\}^B \to \{0,1\}$ 依赖于第 i个变量。等价地，我们称第 i 个变量对 f 是相关的。若 f 至多有 s 个相关变量，则称 f 是一个 s-junta。

显然，定义在 B 个比特上的 B-Junta，只不过是定义在 $\{0,1\}^B$ 上的任意布尔函数。然而，在谱系的另一端，则是独裁者函数；其定义如下。

例 7（独裁者函数）设 $B \in \mathbb{N}_+$。若存在某个 $b \in [B]_+$，使得对每个 $x \in \{0,1\}^B$，f：$\{0,1\}^B \to \{0,1\}$ 都满足以下条件之一：f（x）$= x_b$，或 f（x）$= \neg x_b$，则称布尔函数 f：$\{0,1\}^B \to \{0,1\}$ 为独裁者函数。独裁者函数恰好是1-*junta*。

下面这种等价的 junta 表述对我们更方便。

引理 19（S-Junta 函数作为嵌入置换）设 $B, S \in \mathbb{N}_+$ 满足 $S \le B$。布尔函数 f: $\{0,1\}^B \to \{0,1\}$ 是一个 *S-junta*，当且仅当存在互不相同的 $(i_1, \dots, i_S) \in [B]_+$ 以及某个f_S: $\{0,1\}^S \to \{0,1\}$，使得

$$f(x) = f_S\big(x_{i_1}, \dots, x_{i_S}\big) \quad \textit{for all } x \in \{0,1\}^B.$$

等价地，令 $\Pi \in \{0,1\}^{S \times B}$ 为选择矩阵，对每个 $j \in [S]_+$ 都有 $(\Pi x)^j = x_{i_j}$，并且对每个 $x \in \{0,1\}^B$，我们有

$$f(x) = f_S(\Pi x). \tag{35}$$

引理 17和19 的直接推论给出了以下更强的结果。

引理 20（在提升文字上构造低维完全布尔电路）设 $S, B \in \mathbb{N}_+$ 且 $S \le B$，并设 $f\colon \{0,1\}^B \to \{0,1\}$ 是一个 *S-junta*。那么，存在一个 $B^{\uparrow} \in \mathbb{N}_+$、一个随机矩阵 $\Pi \in \{0,1\}^{B^{\uparrow} \times B}$、一个布尔函数 $f_S\colon \{0,1\}^{B^{\uparrow}} \to \{0,1\}$、整数 $B^{\uparrow}, \Delta \in \mathbb{N}_+$ 以及一个布尔电路 $\mathcal{C}$，使得

$$f = f_S \circ (\Pi \cdot) \tag{36}$$

对于每个 $x \in \{0,1\}^B$，以下条件均成立：

(i) $\mathcal{C}$ 计算 f_S，

(ii) $\mathcal{C}$ 的底层有向无环图是一棵深度为 Δ的有根完全二叉树，

(iii) $B^{\uparrow}$的叶节点标记为 $\{x_1, \dots, x_B, \neg x_1, \dots, \neg x_B\}$中的文字，允许重复；也就是说，对于每个 $\tilde{b} \in [B^{\uparrow}]_+$，我们有

$$x_{\tilde{b}} \in \{x_b\}_{b=1}^{B} \cup \{\neg x_b\}_{b=1}^{B},$$

(iv) $\mathcal{C}$ 中的每个门都属于 {或、与、投影$_A$} $= \{g_8, g_2, g_4\}$。

门的数量恰为 $B^{\uparrow} - 1 \le 2S2^S - 1$，且 $B^{\uparrow} \le 2S2^S$和 $\Delta \le S + \lceil \text{对数}_2\, S \rceil$。此外，若存在某个 $C \in \mathbb{N}$，使得 $S \le C\lceil \log(B) \rceil$，则

$$S \le C\lceil \log(B) \rceil \tag{37}$$

那么存在某些 $p, c \in \mathbb{N}_+$，它们与 B 无关，使得 $B^{\uparrow} \le c\, B^p\, \lceil \text{对数}_2(B) \rceil$。

C.1.1. Proofs of Results from Subsection C.1

证明 [引理的证明 17] 如果 $f = 0$ 是恒假函数，那么我们可以取公式x_1与$\neg x_1$，并直接跳到下文的填充步骤，此时 $N_{\text{lit}} = 2 \le B2^B$；参见（39）。因此，以下我们假设 $f \ne 0$。

步骤 1：文字出现次数有界的析取范式。 将 f 写成规范析取范式（CDN）。对于每个 $a = (a_1, \dots, a_B) \in \{0,1\}^B$，其中 $f(a) = 1$，定义 $T_a\colon \{0,1\}^B \to \{0,1\}$ 对每个 $x \in \{0,1\}^B$，令

$$T_a(x) \stackrel{\text{def}}{=} \bigwedge_{i=1}^{B} \ell_i^{(a)}(x), \text{ where } \ell_i^{(a)}(x) \stackrel{\text{def}}{=} \begin{cases} x_i, & a_i = 1, \\ \neg x_i, & a_i = 0 \end{cases} \tag{38}$$

对每个 $i \in [B]_+$。那么，对于每个 $x \in \{0,1\}^B$，$f(x) = \bigvee_{\{a\colon\, f(a)=1\}} T_a(x)$。令 T 为 f 的满足赋值数。那么 $1 \le T \le 2^B$，且每个 T_a 恰好使用 B 个文字，因此文字出现次数总计 N_{lit} 满足

$$N_{\text{lit}} \stackrel{\text{def}}{=} TB \le B2^B. \tag{39}$$

因此，我们得到一个关于 {与，或} 的公式，其文字叶节点至多有 N_{lit} 个。

步骤 2：转换为二叉树。
注：现在，我们将第1 步中的公式替换为一个等价的二叉公式，其深度我们可以显式地界定。

固定 a，其中有 f 个 $a = 1$。我们声称，合取 $T_a(x) = \bigwedge_{i=1}^{B} \ell_i^{(a)}(x)$ 可以由一个电路计算，该电路的底层有向无环图是一棵深度至多为 $\lceil \log_2 B \rceil$ 的与门二叉树，其叶节点为文字 $\ell_1^{(a)}(x), \ldots, \ell_B^{(a)}(x)$。为此，我们逐层定义 T_a 的电路：对于 0th 层，节点是输入文字 $z_1^{(0)}, \ldots, z_B^{(0)}$，其中对每个 $i \in [B]_+$，我们设定 $z_i^{(0)} \stackrel{\text{def}}{=} \ell_i^{(a)}(x)$。

当层 $k \in \mathbb{N}_+$ 足够大时，我们迭代进行。假设第 $k-1$ 层有 m_{k-1} 个节点，并记作 $z_1^{(k-1)}, \ldots, z_{m_{k-1}}^{(k-1)}$。现在，对每个 $j \in [\lfloor m_{k-1}/2 \rfloor]_+$，我们引入一个输入为 $z_{2j-1}^{(k-1)}$ 和 $z_{2j}^{(k-1)}$ 的与门，并将其输出记作 $z_j^{(k)} \stackrel{\text{def}}{=} z_{2j-1}^{(k-1)} \wedge z_{2j}^{(k-1)}$。若 m_{k-1} 为偶数，则过程结束；否则，恰有一个剩余节点 $z_{m_{k-1}}^{(k-1)}$ 尚未配对。在这种情况下，设定 $z_{m_k}^{(k)} \stackrel{\text{def}}{=} z_{m_{k-1}}^{(k-1)}$；其中 $m_k \stackrel{\text{def}}{=} \lfloor m_{k-1}/2 \rfloor + 1$。因此，

$$m_k \le \left\lceil \frac{m_{k-1}}{2} \right\rceil. \tag{40}$$

由于 $m_0 = B$，我们的递归得出了

$$m_k \le \left\lceil \frac{B}{2^k} \right\rceil. \tag{41}$$

因此，对于 $K \stackrel{\text{定义}}{=} \lceil \log_2 B \rceil$，我们有 $m_K \le \left\lceil \frac{B}{2^{\lceil \log_2 B \rceil}} \right\rceil = 1$。特别是，经过至多 K 层后，恰好剩下一个节点 $z_1^{(K)}$。

根据构造，每个 $z_j^{(k)}$ 都是文字 $\{\ell_i^{(a)}(x)\}_{i=1}^{B}$某个子集的合取，而这些子集构成了第 k 层所有节点上 $\{1, \ldots, B\}$ 的一个划分。因此，最终节点 $z_1^{(K)}$ 等于

$$z_1^{(K)} = \bigwedge_{i=1}^{B} \ell_i^{(a)}(x) = T_a(x). \tag{42}$$

此外，该构造中的每个内部节点的扇入均为 2，因此底层有向无环图是一棵深度至多为 $K = \lceil \log_2 B \rceil$ 的二叉树。
或者，如果我们有如下形式的析取

$$f(x) = \bigvee_{\{a:\ f(a)=1\}} T_a(x).$$

共有 T 个这样的项。完全采用上述相同的构造，但将与替换为或，并将输入 $z_j^{(0)}$ 替换为 $T_a(x)$，即可得到一棵计算 $\bigvee_{\{a:\ f(a)=1\}} T_a(x)$ 的或门二叉树，其深度至多为 $\lceil \log_2 T \rceil$。

综上所述，将每个 T_a 对应的与树与顶层或树组合起来，我们得到一个计算 f 的二叉公式树 T_0，满足：

(a) T_0 的叶节点恰好是步骤 1 中析取范式的文字出现次数 $\ell_i^{(a)}$ (x)，因此其总数为

$$N_{\text{lit}} \le B\,2^B, \tag{43}$$

(b) 由于 $T \le 2^B$，因此 T_0 的深度至多为

$$\lceil \log_2 B \rceil + \lceil \log_2 T \rceil \le \lceil \log_2 B \rceil + B. \tag{44}$$

这个 T_0 就是我们将在下一步补齐的二叉树。

步骤 3：通过投影门转换为完全二叉树

$$\Delta \stackrel{\text{def}}{=} B + \left\lceil \log_2 B \right\rceil \text{ and } B^{\uparrow} \stackrel{\text{def}}{=} 2^{\Delta}. \tag{45}$$

因此，T_0 的深度记作 depth(T_0)，满足 depth(T_0) $\le \Delta$，并且 T_0 的每个叶节点都位于某个深度 $\ell \le \Delta$。对于 T_0 中位于深度为 $\ell < \Delta$ 的某一层、且满足 $\ell \in \mathbb{N}$ 的任意叶节点 v，我们依次将 v 替换为一个新的内部门 u，并将其标记为 PROJ$_A$；然后，可以将 v 设为 u 的左子节点，并添加一个新叶节点 w 作为 u 的右子节点，其标签可为 $\{x_1, \dots, x_B, \neg x_1, \dots, \neg x_B\}$ 中的任意文字。由于 PROJ$_A$(x, y) $= x$ 会忽略其第二个输入，因此 u 处的值等于 v 处的值。这会使该分支的深度增加 1，并使叶节点总数增加 1。

对深度为 $< \Delta$ 的每个叶节点重复此操作，直到全部叶节点都位于深度 Δ，即可得到一棵新树 T_1，满足：1）每个内部节点都有两个子节点（原有的与、或门以及新添加的 PROJ$_A$ 门）；2）每个叶节点都恰好位于深度 Δ。因此，T_1 是一棵深度为 Δ 的完全二叉树。任何有限的满二叉树（即每个内部节点的出度均为 2）若深度为 Δ，则恰有 2^{Δ} 个叶节点。因此，$B^{\uparrow} = 2^{\Delta}$，且 T_1 的根仍计算 f，所有叶节点都像 (ii) 中那样标记为文字，并且所有门都像 (iii) 中那样属于 $\{$或、与、PROJ$_A\}$。最后，根据对 Δ 的选择，

$$B^{\uparrow} = 2^{\Delta} = 2^{B+\lceil \log_2 B \rceil} = 2^B\, 2^{\lceil \log_2 B \rceil} \le 2^B (2B) = 2B2^B.$$

现在，由于 $\log_2(B2^B) = B + \log_2 B$，那么我们还发现 $\Delta = B + \lceil \log_2 B \rceil \le \lceil \log_2(B2^B) \rceil$.

步骤 4：最终计数。接下来需要计算门的数量。令 I 表示 T_1的内部节点数，即那些标有门的节点。对边采用两种方式计数，可得

$$2I = \#\{\text{edges}\} = I + B^{\uparrow} - 1.$$

因此，$B^{\uparrow} = I + 1$。由此，$I = B^{\uparrow} - 1 \le 2B2^B - 1$，这就是 $\mathcal{C}$的门数。 ■

证明 [引理 19] 的证明：如果 f 是一个 S-junta，则存在一组 $J \subseteq [B]_+$ 相关变量，使得$|J| \le S$。如有必要，将 J扩充为 $\{i_1, \dots, i_S\}$。定义 f_S：$\{0,1\}^S \to \{0,1\}$，令 $f_S(z_1, \dots, z_S)$ 定义为 $= f(x)$，其中任意 x 满足 $x_{i_j} = z_j$；这是良定义的，因为 f 不依赖于$\{i_1, \dots, i_S\}$之外的坐标。于是，对所有 x，都有 $f(x) = f_S(x_{i_1}, \dots, x_{i_S})$。反过来，如果 $f(x) = f_S(x_{i_1}, \dots, x_{i_S})$，那么改变 $\{i_1, \dots, i_S\}$ 之外的任何坐标都不会改变 $f(x)$，因此 f至多有 S 个相关变量；所以，f 是一个 S-junta。 ■

证明 [引理 20] 除最后一项结论外，其余结论都可直接由引理 17 和 19推出。对于最后一项结论，注意，如果存在某个 $C \in \mathbb{N}$ 使得 $S \leq C\lceil$对数$(B)\rceil$ 那么由（34）可知

$$B^{\uparrow} \leq S2^{S} \leq C\lceil \log_2 B\rceil\, 2^{C\lceil \log_2 B\rceil} \leq C\, 2^{C}\, B^{C}\, \lceil \log_2 B\rceil.$$

设定 $p \overset{\text{def}}{=} C$ 和 $c \overset{\text{def}}{=} C2^{C}$ 即可得到结果。 ■

现在我们可以证明通用逼近保证了。我们证明定理9 ，因为取8 时，它意味着 $S = B$。

证明 [定理 8] 的证明。固定失败概率 $\delta \in (0, 1]$。

步骤0 - 设置：

固定温度超参数 $\eta > 0$，其值将在之后确定。由于 $f\colon \{0,1\}^{B} \to \{0,1\}$是一个 S-Junta 函数，因此应用引理20，我们可以将 f 表示为（36）中的形式。现在，电路$G = (V, E)$的计算图 $\mathcal{C}$ 计算 f_S，按照引理20的记号，是一棵深度为 $\Delta \in \mathbb{N}_{+}$的有根完全二叉树。因此，我们可以将 V 分解为

$$V = \bigsqcup_{l=0}^{\Delta} V_l \text{ where } V_l = \{v_{l:i}\}_{i=1}^{2^{\Delta-l}} \tag{46}$$

并且我们可以将 E 划分为

$$E = \bigsqcup_{l=1}^{\Delta} E_l \text{ where } E_l = \bigsqcup_{i=1}^{2^{\Delta-l}} \left\{(v_{l-1:i_1}, v_{l:i}), (v_{l-1:i_2}, v_{l:i})\right\}_{i=1}^{2^{\Delta-l}} \tag{47}$$

其中，对于每个 $i \in [2^{\Delta-l}]_{+}$，索引 i_1，$i_2 \in [2^{\Delta-l+1}]_{+}$ 在（47）中互不相同且唯一。此外，不失一般性，我们可以安排 $v_{l:i}$ 图中的边 E G 互不交叉，因为树是平面的。

此外，由于 $\mathcal{C}$ 中的门只标记 $\sqcup_{l=1}^{\Delta} V_l$ 中的顶点，而不标记 V_0 （这些只是引理20 （iii）所述的变量）。我们将标记顶点 $v_{l:i}$ 的门记为 $g_{l:i}$，并且不失一般性地，将 $g_{l:i} \in \{g_k\}_{k=1}^{K}$ （比如 $g_{l:i} = g_k$）与$\{e_k\}_{k=1}^{16}$ 中唯一的标准基向量对应，该向量仅在索引 k对应的位置上为 1。

步骤1 - 表示：

考虑$\boldsymbol{f}$的表示为（11），其深度为（2^{Δ}，$2^{\Delta-1}$，..，2_0）$= (2^{\Delta-l})_{l=0}^{\Delta} \in \mathbb{N}_{+}^{\Delta+1}$；现在我们确定其内部参数。现在，对于每个 $l \in [\Delta]_{+}$ 以及每个 $i \in [2^{\Delta-l}]_{+}$，令

$$W^{(l:\sigma)} \overset{\text{def}}{=} \left(\eta\, g_{l:i}\right)_{i=1}^{2^{\Delta-l}},\ W^{(l:1)} \overset{\text{def}}{=} \eta \oplus_{i=1}^{2^{\Delta-l}} e_{i_1}, \text{and } W^{(l:2)} \overset{\text{def}}{=} \eta \oplus_{i=1}^{2^{\Delta-l}} e_{i_2}. \tag{48}$$

现在应用命题 6和 4并取并集界，$\hat{}\ \boldsymbol{f}(x, \omega) = f(x$对布尔立方体 $x \in \{0,1\}^{B^{\uparrow}}$ 中的每个元素都成立的概率至少为

$$1 - \sum_{l=1}^{\Delta} \sum_{i=1}^{2^{\Delta-l}} \tilde{\delta}(\eta) \tag{49}$$

~ ~其中 $\delta(\eta) \in (0,1]$ 连续依赖于 $\eta > 0$；且 $\lim_{\eta \uparrow \infty} \delta(\eta) = 0$。现在，应用命题 3，我们可以选取 $W \in \mathbb{R}^{B^{\uparrow} \times 2B}$，因此 $X^{\uparrow:W}$ 如式（3）中所定义，从而 $X^{\uparrow:W}$ 与 Π 一致的概率至少为

$$1 - \bar{\delta}(\eta) \tag{50}$$

– ¯对于某个依赖于 η 、连续变化的 $\delta(\eta)$，且满足 $\lim_{\eta \uparrow \infty} \delta(\eta) = 0$。因此，我们可以选择足够大的 η，使得

$$\max\left\{\bar{\delta}(\eta), \tilde{\delta}(\eta)\right\} \leq \delta\left(\max\left\{2, 2^{\Delta} - 1\right\}\right)^{-1}$$

注意到 $\sum_{l=1}^{\Delta} \sum_{i=1}^{2^{\Delta-l}} 1 = 2^{\Delta} - 1$。再应用一次并集界，我们对 η 的选择以及（49）和（50）中的下界即可得出结论。 ■

附录 D. 实验设置与实现细节

本附录记录了用于获得正文中报告的数值结果的实验方案及面向实现的细节。第 3–4节中的理论结果不受这些选择的影响；此处的目标是：(i) 确保可复现性；(ii) 精确描述用于连续基线的评估和可解释性探针。

D.1. 任务族与数据集生成

每个实例都是一个布尔函数 $f\colon \{0,1\}^B \to \{0,1\}$，由句法公式指定（积之和形式并带有额外宏；参见 src/ubcircuit/tasks.py）。对于每个实例，我们都在完整真值表上进行训练和评估，即全部 2^B 种赋值。我们报告精确匹配 (EM)，即预测的真值表在全部 2^B 个输入上均与目标真值表相同。

数据集提供的 结构代理。每一行数据集都提供生成时的代理指标

$$S_{\text{base}} \quad \text{and} \quad L_{\text{base}},$$

其中 S_{base} 是生成过程中使用的宽度代理（例如项数/元数代理），而 L_{base} 是启发式深度估计。这些信息被视为有参考价值，但并非特权信息：基准公式并未进行代数化简，而且多个语法形式不同的公式在逻辑上可能等价。

D.2. 模型与匹配机制

SBC（我们的模型）。我们在全文中均使用16门版本（表3），其中包含二输入门和概率路由。训练完成后，我们通过对学习得到的分类分布逐层取argmax决策来解码离散电路，从而得到（i）解码电路表达式和（ii）门使用直方图（附录D.7）。

MLP 基线。我们使用在相同真值表上训练的 ReLU 多层感知机。我们考虑三种匹配机制：

- **神经元匹配：** MLP 隐藏层宽度设为 $S_{\text{模型}}$，深度设为 $L_{\text{模型}}$（与 SBC 的“形状”匹配）。
- **参数匹配（软）：** 选择 MLP 宽度，使可训练参数数量为 $\leq$ SBC可训练参数数量，且二者具有相同的（B, $S_{\text{模型}}$ $L_{\text{模型}}$）。
- **参数匹配（总计）：** 选择 MLP 宽度，使可训练参数数量等于 $\leq$ SBC 可训练参数加上固定的基元计数代理（将每个基元门基元素视为固定组件）。

D.3. 宽度/深度缩放规则

为研究超出数据集代理指标的额外“预算”所产生的影响，我们应用以下缩放规则：

$$S_{\text{model}} = \text{Clamp}\big(\text{Op}_S(S_{\text{base}}; k_S)\big), \qquad L_{\text{model}} = \text{Clamp}\big(\text{Op}_L(L_{\text{base}}; k_L)\big),$$

其中，Op 可以是恒等、加法或乘法，而截断函数会施加预设边界（例如 $S_{\min} \leq S_{\text{model}} \leq S_{\max}$）。在 SBC 与 MLP 的主要比较中，我们采用加法设置

$$S_{\text{model}} = S_{\text{base}} + S_{\text{add}}, \qquad L_{\text{model}} = L_{\text{base}} + L_{\text{add}},$$

使用 $S_{\text{add}} = 10$ 和 $L_{\text{add}} = 0$，除非另有说明。附录 E.6 中的扫描图展示了 S_{add} 和L_{add}。

D.4. 实验中使用的训练系统

本节介绍我们在实验中使用的具体训练系统。这些选择是用于优化可微分电路模型的实现层面稳定措施；它们并非普适性证明所必需。

D.4.1. DEPTHSTACK ARCHITECTURE.

我们的实现（src/ubcircuit/modules.py）构建了一个深度为L_{model} 的堆栈：(i) 可选的 BitLifting 映射 $B \mapsto B_{\text{eff}}$，(ii) 首个通用层，将 $B_{\text{eff}} \to 2$ 个线路映射为$B_{\text{eff}} > 2$（当$B_{\text{eff}} > 2$时进行成对选择），(iii) $L_{\text{model}} - 2$ 个中间推理层，将 $2 \to 2$ 个线路映射为 $2 \to 2$ 个线路，以及 (iv) 最终推理层，将 $2 \to 1$ 个线路映射为输出。每层包含 S_{model} 个并行布尔单元（“头”），并通过按行计算的 softmax 混合器 L_ℓ 将各头的输出路由到下一层的线路。

D.4.2. Gate interpolants for the 16-gate head

我们的实现使用一个可微特征头 $\sigma_{16}:\ [0,1]^2 \to [0,1]^{16}$ 它同时计算表 3 中的 16 个二输入布尔门。对于每个门 g，其真值表向量为 $Z^{(g)} \in \{0,1\}^4$ 按顶点 $(0,0)$、$(0,1)$、$(1,0)$、$(1,1)$ 排序，我们定义一个插值函数

$$\sigma(x \mid Z^{(g)}) \;=\; \sum_{c\in\{00,01,10,11\}} Z_c^{(g)}\,\phi_c(x), \qquad \sigma_{16}(x) \;=\; \big(\sigma(x \mid Z^{(g)})\big)_{g=1}^{16},$$

其中 $\{\phi_c\}$ 是一组（依模式而定的）角点基函数，满足 $\phi_c(\bar{x}_{c'}) = \delta_{c,\,c'}$ 在四个布尔角点 $\bar{x}_{00}$，$\bar{x}_{01}$，$\bar{x}_{10}$，$\bar{x}_{11}$处。

模式 拉格朗日 **（多重线性）**。令 $\bar{x}_{00} = (0,0)$、$\bar{x}_{01} = (0,1)$、$\bar{x}_{10} = (1,0)$、$\bar{x}_{11} = (1,1)$。在 $[0,1]^2$ 上定义双线性拉格朗日基：

$$\phi_{00}(x) = (1-x_1)(1-x_2), \quad \phi_{01}(x) = (1-x_1)x_2, \quad \phi_{10}(x) = x_1(1-x_2), \quad \phi_{11}(x) = x_1x_2.$$

该基满足 $\phi_{ij}(\bar{x}_{kl}) = \delta_{(ij),\ (kl)}$ 并为每个门生成精确的多重线性扩展。

模式 径向基函数 **（归一化高斯核）**。令 $w_c(x) = \exp(-\|x-\bar{x}_c\|_2^2/(2s^2))$，其中带宽为$s>0$，角点为 $\bar{x}_c$。通过归一化定义单位分解基：

$$\phi_c(x) \;=\; \frac{w_c(x)}{\sum_{c'} w_{c'}(x)}.$$

这会得到一个光滑的（C^∞）插值函数，其锐度由 s 控制。

模式 隆起函数**（紧支撑 C^∞ 核）**。令 $w_c(x) = \psi(\|x-\bar{x}_c\|_2/r)$，其中 $r>0$ 是半径，ψ 是一个标准的光滑紧支撑鼓包函数，满足 $\psi(t) = 0$，其中 $t \ge 1$。我们再次进行归一化：

$$\phi_c(x) \;=\; \frac{w_c(x)}{\sum_{c'} w_{c'}(x)}.$$

在我们的代码中，sigma16.radius 控制 r。当支撑半径的选取使得在每个角点处$\bar{x}_c$ 所有其他核都为零时（例如 $r<1$，因为角点之间的距离为 ≥ 1），归一化基满足 $\phi_c(\bar{x}_{c'}) = \delta_{c,\ c'}$。

D.4.3. Layerwise bandwidth schedule.

对于径向基函数以及 sigma16 使用的内部带宽参数，我们可以选择使用线性调度，跨层逐渐调整锐度

$$s_\ell \;=\; s_{\text{start}} + \frac{\ell}{L_{\text{model}}-1}\,(s_{\text{end}} - s_{\text{start}}), \qquad \ell = 0,\dots,L_{\text{model}}-1.$$

匹配实现参数`sigma16.s start` 与`sigma16.send`。我们在表`sigma16.mode`中报告了11上的消融实验结果。

D.4.4. Pair selection and MI-based priors.

对于 $B_{\text{eff}} > 2$，第一层使用带有两个行softmax矩阵 PL、PR $\in \mathbb{R}^{S_{\text{model}} \times B_{\text{eff}}}$ 的可微配对选择器，这些矩阵定义了关于左侧/右侧输入的分类分布。当 pair.route=mi soft 时，我们根据真值表计算实例特定的高互信息配对（通过 src/ubcircuit/mi.py 中的 top mi pairs），并将其转换为先验 PL_{prior}、$\text{PR}_{\text{prior}} \in \Delta_{B_{\text{eff}}}^{S_{\text{model}} \times B_{\text{eff}}}$（如果启用提升，则扩展$B_{\text{eff}}$）。这些先验会以 pair.prior strength 指定的强度作为对数概率偏置注入。当 pair.route=mi hard 时，使用互信息选出的配对作为固定离散配对（FixedPairSelector）。当 pair.route= learned 时，不使用任何先验。

D.4.5. Repulsive right-pick (optional).

为抑制一个头内部的自配对，配对选择器可以选择性地将右侧选择与左侧选择关联起来。如果 pair.repel=true，实现支持：(i) 对数空间排斥（pair.mode=log 或 hard-log），将 $\eta \log(1 - \text{PL})$ 加到右侧 logits 中，并在 hard-log 模式下选择性地屏蔽左侧 argmax 坐标；以及 (ii) 概率空间排斥（pair.mode=mul 或 hard-mul），将右侧概率乘以 $(1 - \text{PL})$，并在 hard-mul 模式下选择性地将左侧 argmax 坐标硬置为零。我们报告的最佳配置使用互信息先验（pair.route=mi soft），并禁用排斥。

D.4.6. Asynchronous temperature annealing.

每一层使用温度 $\tau_\ell(t)$ 来控制门选择和路由的锐度。我们采用方向为 $\in$ {自上而下，自下而上}、调度方式为$\in$ {线性，余弦}的异步调度。训练初期，较大的 τ_ℓ 会使选择保持柔和（梯度更平滑）；后期，$\tau_\ell \downarrow T_{\min}$ 则使分布接近独热，从而实现稳定的最大值解码。

D.4.7. Sigma16 bandwidth schedule.

使用 16门头时，我们可以选择跨层逐渐调整门特征带宽（sigma16.s start→ sigma16.s end），方法是调用 DepthStack.set layer taus and bandwidths。这是插值函数上的逐层带宽（而非 softmax 温度）。

D.4.8. Objective, optimizer, and early stopping.

所有方法均使用完整真值表和二元交叉熵进行训练。SBC 优化

$$\mathcal{J} = \mathcal{L}_{\text{BCE}} + \mathcal{R}_{\text{ent}} + \mathcal{R}_{\text{div-units}} + \mathcal{R}_{\text{div-rows}} + \mathcal{R}_{\text{const16}}.$$

主要实验中使用 RMSProp。早停会在预热后监控精确匹配（可配置最小步数、检查频率和耐心检查次数）。

D.4.9. Regularizers $\mathcal{R}_\bullet$ used in experiments.

在实现中（src/ubcircuit/utils.py，regularizersbundle），正则化项根据前向传播返回的逐层诊断计算。对于每一层 ℓ,

模型会暴露：(i) 路由行 $L_\ell \in \Delta_S^{B_{\ell+1}\times S}$（对头进行行softmax），以及 (ii) 头门控logits $\{W^{(\ell:s)}\}_{s=1}^{S}$（每个 $W^{(\ell:\,s)} \in \mathbb{R}^{16}$ 当门控集为=16时）。令 τ_ℓ 为当前优化步骤使用的层温度。

熵项。 代码对每个路由行以及每个头的门控分布的香农熵求和：

$$\mathcal{R}_{\text{ent}} = \lambda_{\text{ent}} \sum_{\ell} \left(\sum_{k=1}^{B_{\ell+1}} H(L_\ell[k,:]) + \sum_{s=1}^{S} H\Big(\mathrm{SM}(W^{(\ell:s)}/\tau_\ell)\Big) \right), \qquad H(q) \triangleq -\sum_i q_i \log q_i.$$

头多样性项。 令 $p^{(\ell:s)} = \mathrm{SM}(W^{(\ell:s)}/\tau_\ell) \in \Delta_{16}$。头多样性惩罚是成对余弦相似度之和：

$$\mathcal{R}_{\text{div-units}} = \lambda_{\text{div-units}} \sum_{\ell} \sum_{1\le s<s'\le S} \cos\Big(p^{(\ell:s)}, p^{(\ell:s')}\Big), \qquad \cos(u,v) = \frac{\langle u, v\rangle}{\|u\|_2 \|v\|_2}.$$

行多样性项。 路由多样性惩罚是各路由行之间成对余弦相似度之和：

$$\mathcal{R}_{\text{div-rows}} = \lambda_{\text{div-rows}} \sum_{\ell} \sum_{1\le k<k'\le B_{\ell+1}} \cos\Big(L_\ell[k,:], L_\ell[k',:]\Big).$$

常数门控惩罚（*16*门头）。 启用时（lam const16>0），代码会对在全部非最终层上选择假和真施加惩罚：

$$\mathcal{R}_{\text{const16}} = \lambda_{\text{const16}} \sum_{\ell<L_{\text{model}}-1} \sum_{s=1}^{S} \Big(p^{(\ell:s)}_{\texttt{FALSE}} + p^{(\ell:s)}_{\texttt{TRUE}}\Big),$$

并使用 $\mathcal{R}_{\text{const16}} = 0$ 在最终层上。（在已报告的最佳设置中，我们设定 $\lambda_{\text{const16}} = 0$。）

D.4.10. Decoding and evaluation.

解码是事后进行的：训练后，我们在每一层通过 arg max 选择门和路由，以获得一个离散电路和一个符号表达式。EM 根据模型的真值表预测计算（阈值为 0.5），与符号字符串相等无关。

D.5. 证明层面构造与实验实例化

第 3–4 节中的理论研究了一个随机电路值模型，并证明了通用可表示性和电路结构化实现。实验系统使用具体的可微参数化和确定性解码协议实例化了相同的电路模板。对于解释数值结果而言，相关的主要差异总结于表 4 中。特别是在实践中使用了若干训练期启发式方法（例如温度调度和正则化），以稳定优化并促进清晰的最大值解码；存在性证明并不要求这些方法。

组件	证明级陈述	实验实例化（含参考文献）
输入提升	随机位提升通道 $\mathbf{X}^{\uparrow:W}$ 是包含（第 3.1节）。	已实现（BitLifting），but 最佳性能 设置 禁用 提升 (提升.use=假)。
门族	对全部 16 输入门（表 3）。	使用 16门头（门控集=16），其中训练期间的光滑 σ_{16} 插值以及评估时的最大值解码（见第 D.4.2 以及 D.4.10）。
边选择	抑制去-的结构化选择生成配对（见第3.2.1）。	第一层使用 PairSelector 和学习得到的选取，或 互信息驱动先验（软互信息/硬互信息） and 可选的 排斥 模式 (对数/乘法, 硬-*)（第 D.4.4 节）。
解码	电路实现是针对每个输出结果 ω （定理 7）。	事后取最大值解码会产生单一的区分-离散电路表达式和门控直方图 (章节D.4.10 和D.7)。
优化辅助措施	普适性/存在性证明不需要——性证明。	异步温度退火和正则化项用于稳定优化和清晰解码 (节 D.4.8)。

表 4：证明层面构造与实验中使用的训练系统之间的高层次差异。

D.6. 连续基线的可解释性度量

BNR（附录 B.1）针对布尔值映射定义。多层感知机的隐藏单元取实数值，因此我们通过在完整真值表上枚举每个单元，计算操作性 BNR 诊断和额外的门可恢复性探测。

BNR 诊断（舍入与容差）。令 $u : \{0,1\}^B \to \mathbb{R}$ 为一个在全部 2^B 个输入上进行评估的隐藏单元。我们报告两个操作性变体：

- $\mathbf{BNR}_{\text{exact}}$ **（舍入）**。固定一个舍入精度 d （我们使用 $d = 6$）。将 u 声明为 BNR if

$$\#\big\{\mathrm{Round}_d(u(x)) : x \in \{0,1\}^B\big\} \le 2.$$

- $\mathbf{BNR}_\varepsilon$ **（双簇容差）**。固定 $\varepsilon > 0$ （我们使用 $\varepsilon = 10^{-3}$）。将多重集 $\{u(x)\}$ 按其中位数分成两半，分别取每半的中位数作为中心 c_0, c_1，并在以下情况下判定为 u ε-BNR

$$\max_{x\in\{0,1\}^B} \min\big\{|u(x) - c_0|, |u(x) - c_1|\big\} \le \varepsilon.$$

对于包含 H 个单元的层，我们报告满足各项判据的单元比例。对于 L 层多层感知机，我们同时报告第一层的数值（L1）以及隐藏层的平均值（记为“全部”）。SBC 的构造本身就是布尔值，因此所有内部单元均达到 BNR$= 1$ 。

门探测的确定性二值化。为了探测多层感知机单元的布尔决策模式，我们通过在全零赋值处进行阈值化，将实值单元映射为布尔函数：

$$t(u) \triangleq u(\mathbf{0}), \qquad \hat{u}(x) \triangleq \mathbb{I}\{u(x) \geq t(u)\} \in \{0,1\}.$$

此二值化仅用于下文的门可恢复性探测。

第一隐藏层上的原语可恢复性（输入原语）。对于每个第一层二值化单元 $\hat{u}$，定义

$$\mathrm{bestAcc}_{\mathrm{in}}(u) \triangleq \max_{p \in \mathcal{P}_{\mathrm{in}}} \Pr_{x \sim \mathrm{Unif}(\{0,1\}^B)} \big[\hat{u}(x) = p(x)\big],$$

其中 $\mathcal{P}_{\mathrm{in}}$ 包含文字 $\{a_i, \neg a_i\}_{i=1}^{B}$ 以及表 3 中的 16 个二输入门，它们作用于有序输入对（a_i,a_j）（其中 $i \neq j$）。我们报告

$$\mathrm{primHit}_{\mathrm{in}} \triangleq \Pr_u[\mathrm{bestAcc}_{\mathrm{in}}(u) = 1], \qquad \mathrm{primBest}_{\mathrm{in}} \triangleq \mathbb{E}_u[\mathrm{bestAcc}_{\mathrm{in}}(u)].$$

这里**primHit**是精确恢复率（不允许误差），而**primBest** 则对最佳匹配准确率取平均。

逐层原语可恢复性（全部层）。令 $\hat{z}_1^{(\ell-1)}, \ldots, \hat{z}_{H_{\ell-1}}^{(\ell-1)}$ 表示第 $\ell-1$ 层中的二值化单元。对于第 ℓ 层中的二值化单元 $\hat{u}_j^{(\ell)}$，定义

$$\mathrm{bestAcc}_{\mathrm{layer}}^{(\ell)}(u_j) \triangleq \max_{\substack{i \neq k \\ g \in \mathbb{G}^\star}} \Pr_{x \sim \mathrm{Unif}(\{0,1\}^B)} \Big[\hat{u}_j^{(\ell)}(x) = g(\hat{z}_i^{(\ell-1)}(x), \hat{z}_k^{(\ell-1)}(x))\Big].$$

我们定义逐层汇总指标

$$\mathrm{primHit}_{\mathrm{layer}}^{(\ell)} \triangleq \Pr_j[\mathrm{bestAcc}_{\mathrm{layer}}^{(\ell)}(u_j) = 1], \qquad \mathrm{primBest}_{\mathrm{layer}}^{(\ell)} \triangleq \mathbb{E}_j[\mathrm{bestAcc}_{\mathrm{layer}}^{(\ell)}(u_j)],$$

并报告它们在各隐藏层上的平均值（记为“全部”）。

D.7. 门直方图诊断

我们将门类型直方图记录为描述性诊断：

- **SBC-all：**在最大值解码下，跨全部单元和层的门直方图。
- **SBC-path：**仅限于解码后的读出路径的门直方图。
- **MLP-all：**来自逐层可恢复性探测的精确门匹配直方图，并跨层聚合。

这些直方图并非正确性证书；它们有助于解释哪些基本操作在解码计算中占主导地位。（在我们的实验中，多层感知机精确匹配恢复往往由诸如真之类的常量类型行为主导，因此我们将较高的 **primBest** 与**primHit** 以及门控直方图结合起来进行解释。）

MLP匹配 平均 MLP参数		MLP精确匹配	SBC精确匹配
神经元	368	0.987 ± 0.002	
软参数	683	0.992 ± 0.001	0.971 ± 0.003
总参数	1154	0.999 ± 0.002	

表 5：真值表评估上的精确匹配（EM）（5个随机种子上的均值 ± 标准差）。SBC精确匹配仅报告一次（标准SBC设置），MLP精确匹配则针对每种匹配机制分别给出。

模型 / 匹配	$\mathrm{BNR}_{\mathrm{exact}}$(L1)	$\mathrm{BNR}_{\mathrm{exact}}$(全部)	$\mathrm{BNR}_{\varepsilon}$(L1)	$\mathrm{BNR}_{\varepsilon}$(全部)
SBC	1.000	1.000	1.000	1.000
神经元	0.114 ± 0.003	0.209 ± 0.002	0.115 ± 0.004	0.209 ± 0.002
软参数	0.107 ± 0.002	0.199 ± 0.001	0.107 ± 0.002	0.199 ± 0.001
总参数	0.105 ± 0.002	0.195 ± 0.002	0.105 ± 0.002	0.196 ± 0.002

表 6：BNR 诊断（5 个随机种子的均值 ± 标准差）。“全部”表示对隐藏层取平均。SBC 按构造即满足 BNR，因此 BNR$\equiv 1$。

附录 E. 补充结果与诊断

E.1. 精确匹配与 BNR

我们总结 (i) 通过精确匹配（EM）衡量的任务正确性，以及 (ii) 通过 BNR 衡量的逐神经元布尔值性。EM 在每个实例的完整真值表上计算（第 D.1 节）；BNR 的定义见附录 D.6。表 5 报告了各匹配机制下 MLP 基线的 EM，同时仅报告一次 SBC EM，采用全文统一使用的固定标准SBC设置。表 6 报告了 BNR 诊断：对于 SBC，BNR$\equiv 1$ 按构造即成立（定理 7）；对于多层感知机，则在隐藏激活上进行事后测量。

要点。 在各种匹配机制下，MLP 的 EM 均高于 SBC（表 5），而 SBC 保证逐神经元布尔值性（表 6）。对于多层感知机，严格的二值性并不常见，尤其是在早期层；在此匹配机制下，基于容差的 $\mathrm{BNR}_{\varepsilon}$ 与取整后的 $\mathrm{BNR}_{\mathrm{exact}}$ 密切吻合。

E.2. 原语可恢复性

BNR 衡量隐藏激活是否严格为两个取值；它并不关心某个单元的行为是否类似于特定的布尔门。为探查多层感知机中类似门的决策模式，我们报告确定性二值化后的扇入为 2 的基本操作可恢复性（附录D.6）。表7总结了两种情况下的可恢复性：(i) 相对于输入位的基本操作（L1）；(ii) 逐层相对于二值化的前一层单元（全部）。

多层感知机匹配	$\text{primHit}_{\text{in}}$(L1)	$\text{primBest}_{\text{in}}$(L1)	$\text{primHit}^{\text{layer}}$(全部)	$\text{primBest}^{\text{layer}}$(全部)
神经元	0.463 ± 0.004	0.945 ± 0.001	0.552 ± 0.002	0.953 ± 0.000
软参数	0.466 ± 0.003	0.941 ± 0.000	0.559 ± 0.002	0.953 ± 0.000
总参数	0.477 ± 0.003	0.939 ± 0.001	0.570 ± 0.002	0.953 ± 0.000

表 7：确定性二值化后的二输入原语可恢复性（5 个随机种子上的均值 ± 标准差）。primHit 表示精确原语匹配的单元所占比例；primBest 表示最佳匹配准确率的平均值。“输入”评估第一层单元与输入位上的基本操作的匹配情况；“层”评估相对于二值化的前一层单元的门可恢复性，并对所有隐藏层取平均。

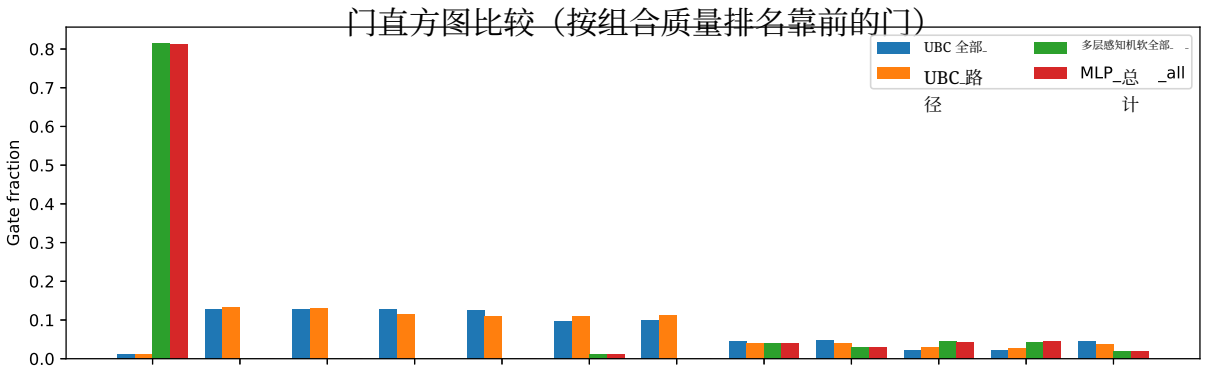


图 2：不同匹配机制下的门直方图：SBC-all、SBC-path 和 MLP-all（精确恢复的门）。

要点。基本操作可恢复性可能显著高于严格的 BNR（对比表6和7），这表明许多多层感知机单元在二值化后会产生类似门的布尔模式，即使其实值幅度仍呈多级分布。

E.3. 门直方图（SBC 与 MLP）

为帮助理解表7，我们展示恢复的计算中占主导地位的门类型。图 2 比较了每种匹配机制下的：(i) SBC-all（全部解码单元），(ii) SBC-path（解码后的读出路径），以及 (iii) MLP-all（逐层原语探测恢复的精确门，按层汇总）（附录D.7）。

要点。 尽管表 7 报告了较高的 `primBest`值，图2 显示，在此情形下，多层感知机的精确命中直方图主要由常量真 门主导。因此，应结合门直方图来解读原语可恢复性：较高的最佳匹配得分可能反映的是大量平凡（常量）门，而非非恒等

设置	标签令牌	SBC令牌
SBC（主）	11.80 ± 0.00	9.80 ± 0.15

表8：标签表达式与解码后的SBC表达式的平均令牌复杂度（各随机种子上的均值 ±标准差）。标签表达式在未经代数化简的情况下生成；在逻辑等价形式中，解码后的SBC表达式往往更短。

路径	排斥	mode	eta	lam_const16	EM	n
软互信息	假	log	2	0	0.8651 ± 0.0124	20
软互信息	假	log	2	0.005	0.8578 ± 0.0126	20
软互信息	假	log	2	0.001	0.8557 ± 0.0126	20
软互信息	True	硬对数	2	0.001	0.8435 ± 0.0006	19
软互信息	True	log	2	0.001	0.8435 ± 0.0006	20
软互信息	True	log	2	0.005	0.8435 ± 0.0004	18
软互信息	True	硬对数	2	0.005	0.8435 ± 0.0004	20
软互信息	True	硬对数	2	0	0.8432 ± 0.0012	20

表 9：完整网格上的前8种配置（随机种子之间的均值 ± 标准差）。

二输入逻辑的广泛恢复。这促使我们将 BNR（表 6）作为主要的逐神经元可解释性证书，并将原语探针用作辅助诊断。

E.4. 表达式复杂度

我们还报告了一种粗略的表达式复杂度代理指标，基于词元计数（变量出现次数加布尔运算符）计算，适用于标签公式和解码后的 *SBC*表达式。由于数据集公式是在未进行代数化简的情况下生成的，因此较短的解码表达式只能被理解为在逻辑等价表示之间存在简洁性偏好，而非语义优势（表 8）。

E.5. 如何选择最终超参数设置（消融路径）

我们通过分阶段消融选定了最终实验配置，并保持数据集与评估协议固定（完整真值表训练和EM评估）。各阶段如下：

阶段I：粗粒度超参数网格（包括提升）。我们首先在路由策略、排斥设置、优化器选择以及提升开关（以及其他调节项）上运行了广泛网格，并按精确匹配对配置进行排序。表9 以及表10 报告了排名靠前的配置（跨随机种子汇总），并说明了我们选择后续阶段所用基线超参数设置的理由。在此阶段，提升（`lifting.use`）被视为可调选项。

第二阶段：16 门插值器消融（固定第一阶段的设置）。固定表现最佳的第一阶段超参数后，我们对训练期间使用的可微 16 门插值器进行了消融实验

路径	排斥	mode	eta	lam_const16	EM	n
学习得到的	True	硬对数	2	0.005	0.7496 ± 0.0056	20
学习得到的	True	硬对数	2	0	0.7457 ± 0.0055	18
学习得到的	True	硬对数	2	0.001	0.7408 ± 0.0078	19
学习得到的	True	log	2	0.005	0.7391 ± 0.0045	20
学习得到的	True	log	2	0	0.7367 ± 0.0061	18
学习得到的	假	log	2	0	0.7366 ± 0.0055	18
学习得到的	假	log	2	0.005	0.7357 ± 0.0044	15
学习得到的	假	log	2	0.001	0.7296 ± 0.0079	18

表 10：lift完整网格上的前8种配置（跨随机种子的均值 ±标准差）。

σ_{16} 模式	EM	行准确率
rbf	0.8721 ± 0.0024	0.9908 ± 0.0004
bump	0.8477 ± 0.0044	0.9873 ± 0.0006
拉格朗日	0.5694 ± 0.0178	0.9552 ± 0.0022

表 11：在 mode cmp 测试集上对 σ_{16} 门插值函数选择进行消融实验（随机种子间的均值 ± 标准差）。

：`sigma16.mode` ∈ {径向基函数、隆起函数、拉格朗日}。表11 总结了结果，表明在我们的实验条件下，径向基函数表现最佳；因此，后续实验中我们固定`sigma16.mode`=径向基函数。

第三阶段：围绕数据集代理指标的宽度/深度预算扫描。 最后，在固定全部其他超参数（包括 sigma16.mode=径向基函数）的情况下，我们相对于数据集提供的代理指标扫描了额外的宽度/深度预算：

$$S_{\text{model}} = S_{\text{base}} + S_{\text{add}}, \qquad L_{\text{model}} = L_{\text{base}} + L_{\text{add}}.$$

图 3 表明，增大 S_{add} 可显著提高精确匹配，而增大 L_{add} 在该情形下往往会降低精确匹配。基于此次扫描，我们选择 $S_{\text{add}} = 10$ 和 $L_{\text{add}} = 0$ 作为正文中 SBC 与 MLP 的比较所使用的最终设置。

E.6. 宽度/深度扫描：额外预算的影响

附录 F：其他相关工作

神经符号推理与概率电路。 神经符号方法将神经组件与符号结构或基于逻辑的推理相结合。逻辑张量网络在训练过程中确保一阶约束得到可微分地满足 Badreddine 等人(2022年，而 DeepProbLog 等神经概率逻辑编程系统则将神经谓词嵌入概率程序，以实现联合学习与推理 Manhaeve 等人(2018)。概率电路，

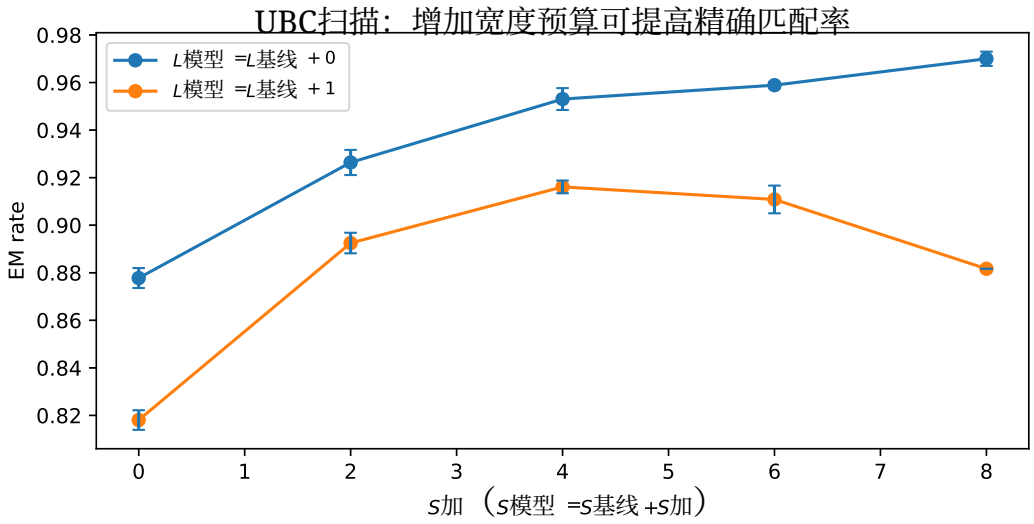


图3：SBC扫描：精确匹配率与 S_{add} 的关系，针对不同的 L_{add}，其中 $S_{\text{model}} = S_{\text{base}} + S_{\text{add}}$ 和 $L_{\text{model}} = L_{\text{base}} + L_{\text{add}}$。

包括和积网络（SPNs），将分布表示为结构化复合，并且在满足可处理结构条件时可进行高效的精确推断 Poon 和 Domingos(2011)；Choi 等人 (2020)。我们的方法在理念上与之最为接近，因为它内在地构建了电路级表示，而非事后提取表示；我们专门针对扇入为2、扇出为1的布尔电路，并获得逐神经元布尔性的显式证书。将该构造扩展到更丰富的电路族（例如，可变扇入/扇出或其他门集合）是未来研究的自然方向。