神经加性模型的可证明解释

Shahaf Bassan[1*] **Yizhak Yisrael Elboher**[1*] **Tobias Ladner**[2*] **Volkan Sahin**[2] **Jan Kˇretínský**[3] **Matthias Althoff**[2] **Guy Katz**[1]
[1] 耶路撒冷希伯来大学，以色列，[2] 慕尼黑工业大学，德国，[3] 马萨里克大学，捷克共和国 联系方式：shahaf.bassan@mail.huji.ac.il, yizhak.elboher@mail.huji.ac.il, tobias.ladner@tum.de

arXiv:2602.17530v1 [cs.LG] 19 Feb 2026

摘要

尽管神经网络事后解释方法取得了显著进展，但其中许多仍依赖启发式方法，且缺乏可证明保证。一种获得具有可证明保证的解释的关键方法，是找出输入特征的基数最小子集，该子集本身足以通过证明 确定预测结果。然而，对于标准神经网络，这项任务在计算上往往难以实现，因为其需要的验证查询数量在最坏情况下会随输入特征数量呈指数增长，且每个查询都是 NP 困难的。在这项工作中，我们表明，对于神经加性模型（NAM），这一较新且更具可解释性的神经网络家族，我们可以高效地生成具有此类保证的解释。我们提出了一种针对 NAM 的新型模型专用算法，仅使用数量随输入特征数量呈对数级增长 的验证查询，即可生成可证明基数最小的解释；此前，需对每个小型单变量神经加性模型组件执行并行化预处理步骤，其运行时间随所需精度呈对数级增长。我们的算法不仅使获取基数最小解释成为可能，其表现甚至优于现有的、旨在寻找较宽松的子集最小 解释变体的算法——后者可能更大、信息量更少，但计算起来更容易——尽管我们的算法解决的是困难得多的任务。实验表明，与以往算法相比，我们的方法能够提供可证明比现有工作更小的解释，并大幅缩短计算时间。此外，我们还表明，与通常用于解释 NAM 的标准基于采样的技术相比，我们生成的可证明解释具有其无法实现的优势。

1 引言

人们提出了多种方法来解释神经网络的预测。经典的加法特征归因方法——例如 LIME（Ribeiro 等人，2016）、SHAP（Lundberg 和 Lee，2017）以及 IG（Sundararajan 等人，2017）——假设实例周围的局部区域内近似呈线性行为。其他方法，如 Anchors（Ribeiro 等人，2018）和 SIS（Carter 等人，2019），旨在找出一个（近乎）充分的输入特征子集——本文将其称为解释——该子集能够决定预测。Anchors 和 SIS 依赖概率采样，且无法保证充分性；而近期研究表明，神经网络验证工具可以作为生成可证明充分解释的基础（Wu 等人，2023年；Bassan 和 Katz，2023年；La Malfa 等人，2021年；Izza 等人，2024年），因此这类方法在安全关键领域尤为重要（Marques- Silva 和 Ignatiev，2022年）。在这一背景下，通常更倾向于更小的充分解释，因为最小性被视为另一项关键的可解释性属性（Ignatiev 等人，2019年；Carter 等人，2019年；Darwiche与Hirth，2020年；Ribeiro 等人，2018年；Barceló 等人，2020b）。

然而，尽管这类解释非常理想，但为标准神经网络生成这类解释在计算上出了名地困难（Barceló 等人，2020b）。尤其是，获取基数最小 解释最坏情况下需要进行指数级 次神经网络验证查询（Barceló 等人，2020b；Ignatiev 等人，2019；Bassan 和 Katz，2023），而每次

*同等贡献。

验证都是 NP 困难的（Katz 等人，2017；Sälzer & Lange，2021），因此即便是玩具示例，这项任务也难以实现（Ignatiev 等人，2019）。因此，现有方法侧重于子集最小 解释（Wu 等人，2023；Bassan 和 Katz，2023；Bassan 等人，2025b），但这类解释的大小通常并非最优，可能非常大，因而不如基数最小的解释信息丰富。此外，即便是这些方法，也仍仅适用于相对较小的模型，因为它们依然需要线性数量的验证查询（Wu 等人，2023；Bassan 等人，2025c）。

我们的贡献。 由于已证明，计算基数最小充分解释对于通用神经网络而言难以处理，因此自然会产生一个问题：某些具有更易解释结构的神经网络架构，能否实现对此类解释的高效计算？尽管即使对于二值化神经网络或仅含单个隐藏层的神经网络等简化模型，这项任务仍颇具挑战（Adolfi 等，2025；Barceló 等，2020a；Sälzer & Lange，2021），但我们在本文中表明，对于另一类此前在这一背景下尚未探索的神经网络架构，这项任务变得可处理：神经加性模型（*NAM*）（Agarwal 等，2021）。近年来，NAM 已成为广泛采用且备受关注的架构（Agarwal 等，2021；Radenovic 等人，2022；Bechler-Speicher 等，2024；Kim 等，2024；Zhang 等人，2024）。NAM 通过在输入特征上施加加性结构，在保留神经网络表达能力的同时，支持可解释的逐特征贡献。

我们的 NAM 专用算法与先前算法的比较。不同于现有算法：对于子集最小解释，它们需要的验证查询次数随输入特征数量线性增长；对于基数最小解释，查询次数则呈指数增长。我们的方法利用神经加性模型的加性结构，仅通过对数级数量的验证查询，即可计算出可证明基数最小的解释子集。为此，我们引入了高度并行化的预处理步骤，其中每项操作都独立地处理 NAM 的一个小型单变量分量，从而大幅提升整体效率。因此，与通常用于神经网络的标准算法相比，我们的方法能够以高得多的效率生成既可证明充分又基数最小的解释。采用最先进的验证器进行的实验表明，我们的算法生成解释的速度显著更快，且得到的子集明显小于以往方法。

我们的可证明 NAM 解释与标准采样式解释的比较。通常，人们通过对输入点进行采样并将各单变量函数的行为可视化来解释神经加性模型（Agarwal，2021年；Radenovic 等人，2022年）。得益于其加性结构，这些模型可以逐一检查各特征的贡献。我们表明，纯粹基于采样的方法可能会得出误导性的解释，而我们的可证明充分解释从设计上避免了这一问题，凸显了它们在安全关键领域中的重要性。

总而言之，我们的工作通过两种方式推动了具有可证明保证的解释： (i) 提出了首个针对 NAM 的可认证解释方法，提升了其在安全关键场景中的可信度；(ii) 通过高效地生成可证明的解释，NAM 不同于通用神经网络，为实现可解释架构铺平了道路，使这类保证能够大规模推导。未来的一项核心挑战是设计能够兼顾高表达能力、准确性与高效可证明解释的模型，而我们认为，这项工作是朝这一方向迈出的重要第一步。

2 预备知识

2.1 符号约定

我们用小写字母表示标量，用黑体小写字母表示向量，并用花体字体表示集合。向量$\mathbf{x}$的第 i个分量记为$\mathbf{x}_i$。对于 $n \in \mathbb{N}$，令 $[n] := \{1, \ldots, n\}$。

2.2 神经网络验证

神经网络验证旨在验证神经网络的某些输入输出关系。对于神经网络 $f\colon\ \mathbb{R}^n \to \mathbb{R}^c$，神经网络验证器通过形式化证明不存在满足以下条件的输入$\mathbf{x} \in \mathbb{R}^n$：输入规范 $\psi_{\text{in}}(\mathbf{x})$ 与不安全输出规范 $\psi_{\text{out}}(f(\mathbf{x}))$ 同时成立。尽管这是一个 NP 困难问题（Katz 等人，2017；Sälzer & Lange, 2021），近年来这些工具的可扩展性已迅速提升（Kaulen 等人，2025）。

2.3 神经加性模型（NAM）

用于回归任务的神经加性模型（NAM） f，其中 $f\colon\ \mathbb{R}^n \to \mathbb{R}$，定义如下：

$$f(\mathbf{x}) := \beta_0 + \sum_{i=1}^{k} f_i(\mathbf{x}_i), \tag{1}$$

其中，每个 $f_i\colon\ \mathbb{R} \to \mathbb{R}$ 都是一个单变量神经网络，$\beta_0 \in \mathbb{R}$ 是截距，f 表示完整的神经加性模型。对于 二分类，我们假设额外应用一个阶跃函数：$f(\mathbf{x})\colon =\text{step}(\beta_0 + \sum_{i=1}^{k} f_i(\mathbf{x}_i))$，其中 $\text{step}(z) = 1$ if $z \geq 0$，而 $\text{step}(z) = 0$ otherwise。在具有 c 个类别的<多分类>设置中，类别 $j \in [c]$ 的对数几率由 $f_j(\mathbf{x})\colon = \beta_{j,\,0} + \sum_{i=1}^{k} f_{j,\,i}(\mathbf{x}_i)$给出，并且模型预测 $f(\mathbf{x})\colon = \arg\max_{j\in[c]} f_j(\mathbf{x})$。

在本文中，我们针对以下三种设置开发算法——（i）回归、（ii）二分类和（iii）多分类——但为清晰起见，正文主要介绍二分类情形，其他设置的扩展见附录E。

3 神经网络的可证明充分解释

我们首先回顾为通用神经网络开发的标准算法，这些算法能够识别可证明最小充分解释。值得注意的是，我们关注针对特定输入$\mathbf{x} \in \mathbb{R}^n$的事后充分解释，即针对输出 $f(\mathbf{x})$的解释；“事后”表示解释是在模型训练完成后生成的。

充分解释。解释分类器决策的一种常见方法，是识别输入特征的子集 $\mathcal{S} \subseteq [n]$，使得将这些特征固定为其特定取值后，能够保证预测保持不变。具体而言，这些方法保证，对于补集 $\bar{\mathcal{S}} := [n] \setminus \mathcal{S}$中的任意潜在赋值，分类结果都保持一致。在经典设定中，补集 $\bar{\mathcal{S}}$中的特征可以取任意可能的特征值（Ignatiev 等人，2019；Darwiche与Hirth，2020；Bassan 和 Katz，2023），而更可行、泛化性更强的版本则将 $\mathcal{S}$ 的可能赋值限制在有界 ϵ_p-区域内（Wu 等人，2023；La Malfa 等人，2021；Izza 等人，2024）。我们用$(\mathbf{x}_{\mathcal{S}};\, \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \in \mathbb{R}^n$ 表示一种赋值，其中 $\mathcal{S}$ 的特征取向量$\mathbf{x} \in \mathbb{R}^n$ 中的值，而$\bar{\mathcal{S}}$的特征取另一个向量 $\tilde{\mathbf{x}} \in \mathbb{R}^n$ 在该 ϵ_p-区域内的值。

定义 1 (充分解释)。 给定一个神经网络 f、输入 $\boldsymbol{x} \in \mathbb{R}^n$、扰动半径 $\epsilon_p \in \mathbb{R}_+$和子集 $\mathcal{S} \subseteq [n]$，若满足以下条件，我们称 $\mathcal{S}$ 是关于查询 $\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p\rangle$在以$\boldsymbol{x}$ 为中心、半径为 $\epsilon_p \in \mathbb{R}_+$ 的 ℓ_p-范数球 $B_p^{\epsilon_p}$上的充分解释：

$$\forall \tilde{\boldsymbol{x}} \in B_p^{\epsilon_p}(\boldsymbol{x})\colon \quad f(\boldsymbol{x}_{\mathcal{S}}; \tilde{\boldsymbol{x}}_{\bar{\mathcal{S}}}) = f(\boldsymbol{x}), \qquad \textit{with } B_p^{\epsilon_p}(\boldsymbol{x}) := \{\tilde{\boldsymbol{x}} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_p \leq \epsilon_p\}.$$

当且仅当 $\mathcal{S}$ 构成关于查询$\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p\rangle$的充分解释时，我们定义 $\text{suff}(f, \boldsymbol{x}, \mathcal{S}, \epsilon_p) = 1$；否则，$\text{suff}(f, \boldsymbol{x}, \mathcal{S}, \epsilon_p) = 0$。

定义 1 可以表述为一个神经网络验证查询（第 2.2 节）。此前的研究已提出这一方法，并采用这些技术来验证特定子集的充分性（Wu 等人，2023年；Bassan 和 Katz，2023年；La Malfa 等人，2021年；Izza 等人，2024年）。

最小解释。 显然，选择整个输入集作为子集 $\mathcal{S}$，即设置$\mathcal{S} := [n]$，可以得到一个充分解释。尽管如此，文献中的主流共识是，较小的子集往往信息量更大或更有意义（Ribeiro 等人，2018年；Carter 等人，2019年；Barceló 等人，2020b；Ignatiev 等人，2019年）。因此，人们十分关注如何找出不仅充分、还满足某种最小性概念的子集。我们重点关注两种具体的最小性准则：基数最小性和子集最小性。

定义 2 (最小充分解释)。 给定一个神经网络 f、一个输入 $\boldsymbol{x} \in \mathbb{R}^n$，以及一个关于 $\langle f$ 的充分解释子集 $\mathcal{S} \subseteq [n]$，$\boldsymbol{x}, \mathcal{S}, \epsilon_p\rangle$，在半径为 ϵ_p 的 $B_p^{\epsilon_p}$ 上，则：

1. 我们称 $\mathcal{S}$ 是关于 $\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p\rangle$ 的基数最小充分解释（*Barceló* 等人，*2020a*；*Bassan* 等人，*2024*），当且仅当不存在关于 $\langle f, \boldsymbol{x}, \mathcal{S}', \epsilon_p\rangle$ 的充分解释 $\mathcal{S}'$，且满足 $|\mathcal{S}'| < |\mathcal{S}|$。

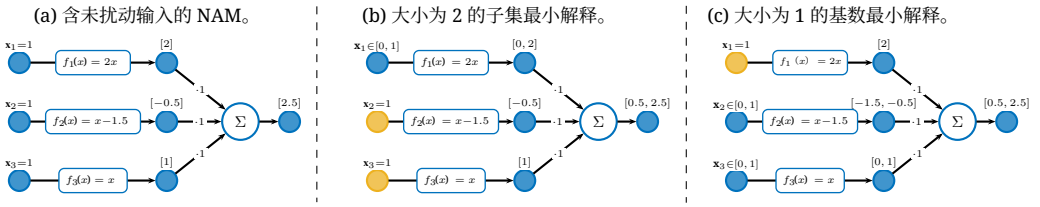

图 1：NAM 的子集最小解释与基数最小解释的比较。(b) 和 (c) 中的两个解释（以黄色显示）都是最小的，因为扰动任何额外特征都可能导致整体输出变为负值。

> *2.* 我们称 $\mathcal{S}$ 是关于 $\langle f, \boldsymbol{x}\ \mathcal{S}, \epsilon_p \rangle$ 的子集最小充分解释（*Arenas* 等人，*2022*年；*Ignatiev* 等人，*2019*年），当且仅当任意 $\mathcal{S}' \subset \mathcal{S}$ 都不是关于$\langle f, \boldsymbol{x}\ \mathcal{S}', \epsilon_p \rangle$ 的充分解释。

最小充分解释也可以使用神经网络验证器来确定。该过程需要执行多次验证查询，以确保子集的最小性。算法 1 概述了这样的过程（Ignatiev 等人，2019；Wu 等人，2023；Bassan 和 Katz，2023）。该算法从一个涵盖整个特征集 $[n]$ 的解释 $\mathcal{S}$开始，并迭代地尝试从 $\mathcal{S}$中排除特征 i，每次检查 $\mathcal{S} \setminus \{i\}$ 是否仍然充分。如果 $\mathcal{S} \setminus \{i\}$ 仍然充分，则移除特征 i；否则将其保留在解释中。重复此过程，直到获得子集最小充分解释。NAM 的这种子集最小解释与基数最小解释一起在图 1 中进行了可视化展示。

算法 1 贪心包含关系极小解释搜索

输入： 神经网络 f：$\mathbb{R}^n \to \mathbb{R}^c$，输入 $\mathbf{x} \in \mathbb{R}^n$，扰动半径 $\epsilon_p \in \mathbb{R}_+$

```
S ← [n]
对于每个 特征 i ∈ [n]执行                                  ▷suff(f, x S, ε_p) 成立
    如果 suff（f, x S \ {i}, ε_p）则
        S ← S \ {i}
    结束条件判断
结束循环
返回 S                                  ▷ S 是关于 ⟨f子集最小解释, x S, ε_p⟩
```

4 神经加性模型的可证明基数最小充分解释

虽然对于一般神经网络而言，生成基数最小充分解释在计算上开销很大，但本节提出了一种针对神经加性模型的高效算法。由于采用了单变量函数，神经加性模型通常被认为具有很强的可解释性，但只有通过形式化验证才能获得充分保证，避免得出误导性结论（图 2）。我们的算法包含两个主要阶段：(i) 在预处理步骤中，我们根据单变量函数 f_i 为每个特征 $i \in [n]$ 计算一个“重要性”区间，以获得全序。(ii) 这使我们能够在排序后的区间上进行二分搜索，从而确定基数最小充分解释。

所有完整证明均见附录 C。

4.1 阶段 1 — 并行区间重要性排序

本小节概述我们的算法的第一阶段，即确定将在后续阶段使用的特征重要性排序。我们采用与（Barceló 等，2020a）类似的“重要性”概念，该概念最初用于二元线性阈值模型。具体而言，我们称特征 $i \in [n]$ 比特征 $j \neq i$ 更“重要”，如果扰动特征 i 会造成

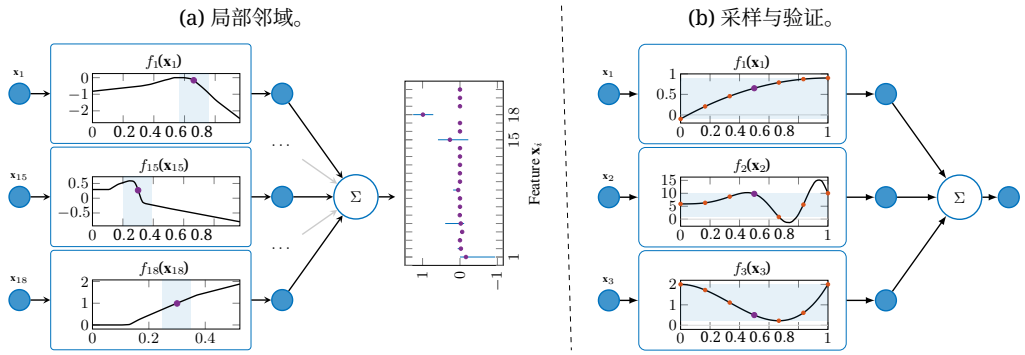

图 2：神经加性模型中的充分解释：(a) 用户必须检查输入的邻域，才能正确理解；例如，用户可能会错误地认为，FICO HELOC 数据集中的特征 1 单独决定了正向输出，但特征 15 或 18 的微小变化就可能改变分类结果。(b) 如果未进行验证，连续邻域的输出可能会产生误导，因为采样可能会漏掉极值。例如，用户可能会根据采样错误地认为，除特征 1 外的所有特征都只有正向输出，因此固定特征 1 总会使整个神经加性模型产生正向输出；然而，特征 2 的输出可能变得足够负，从而改变分类结果。

最终预测 $f(\mathbf{x})$ 的更大变化，大于扰动特征 j。具体而言，我们衡量为翻转分类而朝向决策边界的偏移。得益于 NAM 的加性结构，该分析可针对每个单变量分量 f_i: $\mathbb{R} \to \mathbb{R}$ 独立进行，从而直接比较它们各自的重要性。

不失一般性，假设我们的二元分类器预测 $f(\mathbf{x}) = 1$，这意味着$\beta_0 + \sum_{i=1}^{k} f_i(\mathbf{x}_i) \geq 0$（情况 $f(\mathbf{x}) = 0$ 具有对称性）。在这种情况下，我们分别扰动每个单变量分量的输入 f_i，并测量整体预测降低了多少，即朝决策边界移动的程度。因此，对于每个 $i \in [n]$，我们希望找到

$$\mathbf{x}_i^* = \underset{\tilde{\mathbf{x}}_i \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_i)}{\arg\min} f_i(\tilde{\mathbf{x}}_i). \tag{2}$$

然而，与（Barceló 等人，2020b）研究的二元线性阈值模型不同，精确确定这些最小值通常在计算上不可行（Katz 等人，2017）。幸运的是，我们无需找到精确最小值，只需找到界 $[l_i, u_i] \subset \mathbb{R}$，使得$l_i \leq f_i(\mathbf{x}_i^*) \leq u_i$，且精度足以确定输入特征的全序。该过程概述于算法 2。

算法 2 在 NAM 的每个单变量分量上并行运行。对于每个分量，算法首先发出一个不完全验证查询，以获取初始下界$f_i(\tilde{\mathbf{x}}_i)$，记为 l_i 和 u_i，并在域 $\tilde{\mathbf{x}}_i \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_i)$ 上进行评估。随后，每个线程独立使用验证查询执行二分搜索，迭代细化 $[l_i, u_i]$，逐步逼近真实下界 $f_i(\mathbf{x}_i^*)$。为量化未扰动输出 $f_i(\mathbf{x}_i)$ 与这些界之间的偏差，我们定义相对差值 Δl_i 和 Δu_i。每个并行线程完成一次迭代后，算法都会评估是否可以对所有配对 $(\Delta l_1, \Delta u_1) \ldots, (\Delta l_n, \Delta u_n)$ 进行完整且不重叠的排序。如果尚无法实现这样的排序，就继续迭代细化这些界。最后，返回依据所确定重要性得到的排序。

虽然最初计算得到的界限 α_i, β_i 涵盖了每个分量$i \in [n]$的整个输出域，但二分搜索会逐渐收紧界限，使其围绕最小值 $f_i(\mathbf{x}_i^*)$（对于$f(\mathbf{x}) = 0$，则围绕最大值）展开；因此，界限不再覆盖整个域。由于 $f_i(\mathbf{x}_i^*)$ 只是一个标量，因此可以使用完备验证器确定这个全序。

这种互不重叠的排序为通过相应的分量函数 $f_i(\mathbf{x}_i)$ 扰动单个特征所产生的影响提供了严格的度量。这种排序构成了算法下一阶段的基础，该阶段旨在识别一个经严格证明的基数最小充分子集。以下命题将这第一步形式化：

命题 1. 给定一个 *NAM* f、一个输入$\boldsymbol{x} \in \mathbb{R}^n$ 和一个扰动半径 $\epsilon_p \in \mathbb{R}_+$，令算法 *2* 根据输入特征的重要性返回一个全序列表。则以下结论成立：

算法 2并行区间重要性排序

输入： NAM f，输入 $\mathbf{x} \in \mathbb{R}^n$，扰动半径 $\epsilon_p \in \mathbb{R}_+$

对于每个特征 $i \in [n]$并行**执行**
  提取初始界 α_i，β_i 对于 $f_i(\tilde{\mathbf{x}}_i)$，使得 $\tilde{\mathbf{x}}_i \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_i)$
  $l_i \leftarrow \alpha_i$，$u_i \leftarrow \beta_i$
  当True **时执行**
    $m_i \leftarrow \frac{l_i+u_i}{2}$
    若验证($\forall \tilde{\mathbf{x}}_i \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_i)$ $f_i(\tilde{\mathbf{x}}_i) \geq m_i$) **则**
      $l_i \leftarrow m_i$
    else
      $u_i \leftarrow m_i$
    结束若
    $\Delta l_i \leftarrow f_i(\mathbf{x}_i) - l_i$； $\Delta u_i \leftarrow f_i(\mathbf{x}_i) - u_i$
    若对所有 $j \neq i$均成立： $\Delta u_i \geq \Delta l_j$ 或 $\Delta u_j \geq \Delta l_i$ **则**
      中断
    结束 if
  结束 while
结束 for
返回 arg sort([$(\Delta l_1, \Delta u_1)$ ……，$(\Delta l_n, \Delta u_n)$])，按升序排列

对于任何包含特征 i 的充分解释 $\mathcal{S}$，以及列表排序中满足$i \prec j$ 的任何特征 $j \notin \mathcal{S}$，集合 $\mathcal{S} \setminus \{i\} \cup \{j\}$ 也是一个充分解释。

复杂度。算法 2 的复杂度取决于使用的 ρ并行处理器数量，其中每个处理器独立执行二分搜索。此二分搜索会迭代地根据 Δu_i 和 Δl_i 界进行划分，并在两个不同单变量分量的界不再重叠时终止。给定初始上界与下界之间的间隔 $\alpha_i - f_i(\mathbf{x}_i)$，对于每个分量 f_i，以及分量 f_i 的精度（由其与排序结果中相邻特征之间的最小间隔定义），即 $\xi_i := \min\{|\hat{\Delta} l_{i+1} - \hat{\Delta} u_i|, |\hat{\Delta} l_i - \hat{\Delta} u_{i-1}|\}$，其中 $\hat{\Delta} l_\square$，$\Delta u_\square$ 表示最后一次迭代中的界。现在可以证明，神经网络验证器调用次数受一个（并行化的）对数项所界定，具体表述如下命题。局限性和优化将在附录 D 中进一步讨论。

命题 2。给定 ρ 个并行处理器，算法 *2* 执行的神经网络验证器调用总次数为 $T_p(n) = \mathcal{O}\Big(\big(\frac{n}{p}\big)\log\big(\max^{i\in[n]}(\frac{\beta_i-\alpha_i}{\xi_i})\big)\Big) \xrightarrow[p\to n]{} \mathcal{O}\Big(\log\big(\max^{i\in[n]}(\frac{\beta_i-\alpha_i}{\xi_i})\big)\Big)$，每次调用均针对一个单变量分量 $f_i(\cdot)$，其中 $\xi_i := \min\{|\hat{\Delta} l_{i+1} - \hat{\Delta} u_i|, |\hat{\Delta} l_i - \hat{\Delta} u_{i-1}|\}$。

边界情况 $\xi_i \to 0$。虽然算法 2 关于单变量分量最优值之间精度差距的增长呈对数级，但理论上该差距可能任意小。我们在附录 D.1 中讨论这一边界情况，并证明即使 ξ_i 趋近于零，只要模型具有 ReLU 激活，算法仍保证在与模型编码大小呈线性关系的步数内终止。这仍比通用神经网络已知的指数级最坏情况行为有了显著改进。关键的是，我们的实证结果表明，实践中该参数绝不会接近零，因此能够实现高效的排序阶段。附录 F.5 中提供了对该参数的详细消融分析。

4.2 阶段2 — 特征　　基于推导出的特征　　选择区间

在本小节中，我们将描述我们的算法的第二部分：给定由算法 2 得到的区间排序，该部分可以获得可证明的基数最小充分解释。与二元线性模型（Barceló 等人，2020b）不同，在二元线性模型中，每个特征的精确最大和最小贡献是已知的，因此可以直接选择排名前k 个特征，而神经加性模型并不具备这种便利。对于神经加性模型，我们只能获得每个单变量分量的界，因此必须显式证明固定某些特征确实是充分的。然而，对于该全序，我们可以应用二分搜索 来获得解释，从而使验证查询次数相对于输入特征数量呈对数级。然而，为简化该算法的表述，我们将首先介绍一种朴素贪心方法，其运行步数为

线性数量的步骤，类似于（Barceló 等人，2020b）所提出的方法，然后再介绍二分搜索方法。该朴素方法如算法 3 所示。

算法 3贪心基数最小线性解释搜索

输入： NAM f，输入 $\mathbf{x} \in \mathbb{R}^n$，扰动半径 $\epsilon_p \in \mathbb{R}_+$

$\mathcal{S} \leftarrow [n]$
对于每个特征 $i \in [n]$，按算法 2 排序**执行**　　▷suff($f, \mathbf{x}\ \mathcal{S}, \epsilon_p$) 成立
　　若 suff(f，$\mathbf{x}\ \mathcal{S} \setminus \{i\}$，$\epsilon_p$) **则**
　　$\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$
　　结束 if
结束 for
返回 $\mathcal{S}$　　▷ $\mathcal{S}$ 是关于 $\langle f, \mathbf{x}\ \mathcal{S}, \epsilon_p \rangle$ 的基数最小的解释

算法 3 与算法 1 的操作十分相似：它首先将解释 $\mathcal{S}$ 初始化为完整的特征集 $[n]$，然后迭代地移除特征并更新 $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$，直到得到一个极小解释。然而，与仅保证收敛到子集最小解释的算法 1 不同，算法 3 收敛于更具挑战性的目标：找到一个基数最小的充分解释。这一更强的保证得益于算法 2 计算出的全序 $(\hat{}\Delta l_i, \hat{}\Delta u_i)_{i=1}^{n}$，该全序按照特征的重要性对其进行排序。由此得到以下命题：

命题 3。 给定一个 *NAM* f、输入 $\boldsymbol{x} \in \mathbb{R}^n$ 和扰动半径 $\epsilon_p \in \mathbb{R}_+$，算法 *3* 执行 $\mathcal{O}(n)$ 次查询，并返回一个基数最小 充分解释。这与算法 *1* 形成对比，后者仅能保证返回一个子集最小 充分解释。

通过用二分搜索策略（算法 4）替代线性排序，可以显著增强算法 3。关键的是，朴素的未排序方法（算法 1）无法采用这一步，因为它无法保证收敛到基数最小解释，甚至可能无法得到子集最小 解释。这是因为，在任意特征排序中，非充分子集可能在多个位置变为充分，使得二分搜索不可靠。相比之下，算法 2 的预处理步骤会对特征进行结构化排序，使算法 4 仅使用对数级 数量的查询，就能可靠地收敛到基数最小 充分解释。

算法 4 贪心基数最小对数解释搜索**输入：** NAM f，输入 $\mathbf{x} \in \mathbb{R}^n$，扰动半径 $\epsilon_p \in \mathbb{R}_+$**1**: $F \leftarrow$ 特征的全序（算法 2）**2**: $l \leftarrow 1$； $u \leftarrow n$**3**: **当** $l \neq u$ **执行** **4** $m \leftarrow \lfloor \frac{l+u}{2} \rfloor$: : **5** **如果** suff(f，$\mathbf{x}\ \{F[1], \ldots, F[m]\}$，$\epsilon_p$) **则****6**: $l \leftarrow m$**7**: **否则****8**: $u \leftarrow m - 1$**9**: **结束如果****10**: **结束当****11**: $\mathcal{S} \leftarrow \{F[1], \ldots F[m]\}$**12**: **返回** $\mathcal{S}$　　▷ $\mathcal{S}$ 是一个基数最小解释，关于 $\langle f, \mathbf{x}\ \mathcal{S}, \overline{\quad} \epsilon_p \rangle$

命题 4. 给定一个 *NAM* f、一个输入 $\boldsymbol{x} \in \mathbb{R}^n$ 以及一个扰动半径 $\epsilon_p \in \mathbb{R}_+$，算法 *4* 执行 $\mathcal{O}(\log(n))$ 次查询，并返回一个基数最小充分解释。

总体复杂度结果。 结合算法 2 和算法 4，我们可以为 $\langle f, \mathbf{x}\ \epsilon_p \rangle$ 得到一个基数最小解释。该统一算法显著提升了效率，将最坏情况下所需的验证查询次数从指数级降至仅对数级的（并行）查询次数。这些查询的第一部分通过对模型的单变量分量 f_i 执行验证查询来进行；这些分量规模要小得多，因此比直接向 f 发起查询更容易验证。其复杂度界由以下定理形式化给出：

表 1：平均解释规模和计算时间的比较。

	乳腺癌		CREDIT		FICO HELOC	
方法	大小 (↓)	时间 [秒](↓)	大小 (↓)	时间 [秒](↓)	大小 (↓)	时间 [秒](↓)
Ours	$\mathbf{4.00}\pm\mathbf{4.24}$	$\mathbf{35.60}\pm\mathbf{1.34}$	$\mathbf{3.76}\pm2.62$	$\mathbf{132.67}\pm\mathbf{36.76}$	$\mathbf{5.59}\pm1.80$	317.92 ± 222.07
字典序	16.58 ± 5.44	634.92 ± 77.23	12.42 ± 6.45	473.63 ± 128.38	15.60 ± 7.53	$\mathbf{146.16}\pm188.07$
敏感性	16.27 ± 5.57	636.79 ± 87.44	3.82 ± 1.84	407.93 ± 126.63	9.45 ± 5.90	250.09 ± 148.44

定理 *1.* 运行算法 *2* 和算法 *4* 可得到一个基数最小充分解释，所需查询次数为$T_p(n) = \mathcal{O}\left(\binom{n}{p}log(max^{i\in[n]}(\frac{\beta_i-\alpha_i}{\xi_i})\right) \xrightarrow[p\to n]{} \mathcal{O}\left(log(max^{i\in[n]}(\frac{\beta_i-\alpha_i}{\xi_i})\right)$次查询，针对 $f_i(\cdot)$ 个分量，另加 $\mathcal{O}(log\ n)$ 次针对 $f(\cdot)$ 的查询。相比之下，标准算法要获得基数最小解释，需要 $\mathcal{O}(2^n)$ 次验证查询，针对 $f(\cdot)$；而要获得子集最小解释，也需要 $\mathcal{O}(n)$ 次验证查询，针对 $f(\cdot)$。

如前所述，我们的复杂度分析以单变量分量 ξ_i 的最小值之间的间隔为参数；在实践中，这个间隔通常不可忽略，不过某些 ξ_i 可能较小。为解决这一问题，我们证明，依次运行算法 2 和算法 4 所需的时间与模型的编码大小成线性关系。我们还证明了匹配的查询界：该任务至多可通过线性数量的查询解决，而任何算法都至少需要关于 f 编码大小的对数数量的查询。

定理 2。（非正式表述。）设 f 是一个 *NAM*，$\boldsymbol{x} \in \mathbb{R}^n$ 是一个输入，且 $\epsilon_p \in \mathbb{R}$。计算 $\langle f, \boldsymbol{x}\ \epsilon_p\rangle$ 的基数最小充分解释的大小，在最坏情况下需要向 *NP* 预言机进行至多 $\mathcal{O}(m)$ 次查询（论断 *1*），且至少需要 $\mathcal{O}(\log(m))$ 次此类查询（引理 *5*），其中 m 是 f 的编码大小。此外，依次运行算法 *2* 和算法 *4*，至多经过 $\mathcal{O}(m)$ 次查询后即可终止（命题 *5*）。

5 评估

实验设置。我们使用 α-β-CROWN 作为后端验证器，实现了主要的算法方法（先执行算法 2，再执行算法 4）。α-β-CROWN 是当前神经网络验证领域的最先进方法（Wang 等，2021；Zhou 等，2024；Kotha 等，2023；Kaulen 等人，2025；Chiu 等人，2025）。我们在 NAM 领域广泛使用的四个表格数据基准上进行了大量实验（Agarwal 等，2021；Radenovic 等人，2022）：(i) 乳腺癌，(ii) CREDIT，(iii) FICO HELOC；这些基准均在安全关键领域中具有重要地位。我们采用了神经加性模型文献中先前工作的相同模型架构（Agarwal 等，2021；Radenovic 等人，2022）。评估细节、其他实验和消融研究见附录 F。

5.1 我们的算法与以往算法

我们首先将结果与文献中为获得可证明最小充分解释而提出的先前算法进行比较。由于我们的方法首次针对强得多的基数最小充分解释这一概念，任何朴素的基线方法——穷举所有 2^n 输入子集，验证其充分性，并选出基数最小的子集来计算此类解释——都无法在合理的超时时限内完成。因此，我们将方法与可扩展性更强的任务进行比较：使用先前工作采用的标准贪心算法（Wu 等人，2023；Bassan 等人，2025b；Izza 等人，2024；Ignatiev 等人，2019；La Malfa 等人，2021）（算法 1），寻找子集最小解释这一较弱的最小性概念。由于子集最小解释取决于特征排序，我们考虑两种设置：(i) 对特征进行基本的字典序排序；(ii) 遵循先前方法（Wu 等人，2023；Bassan 等人，2025b；Izza 等人，2024；Wu 等人，2024b），采用更复杂的反向敏感度排序。

表 1 中的结果表明，与先前算法相比，我们提出的算法在计算时间和解释大小方面均取得了显著改进。除了比标准的子集最小解释算法进一步减小解释大小（这是必然的，因为我们采用了更强的最小性定义）之外，我们的方法还大幅提升了运行时间，

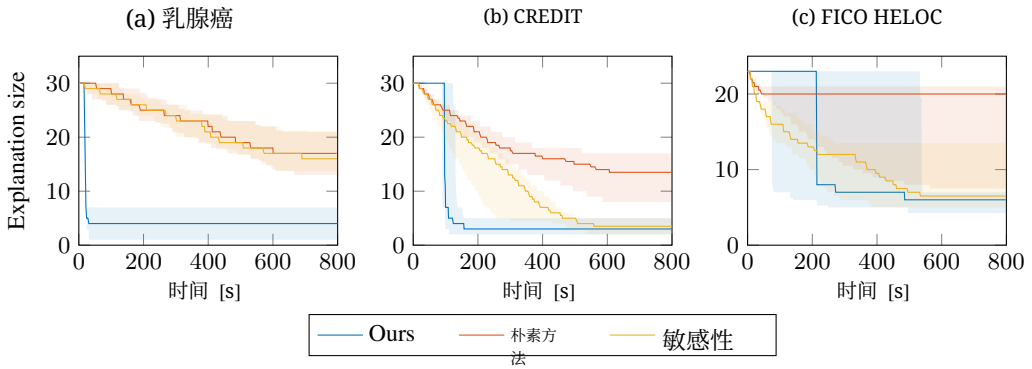

图 3：所有数据集的解释大小随时间的变化。

表 2：与纯基于采样的方法的比较。

	CREDIT			FICO HELOC		
方法	大小 (↓)	时间 [s] (↓)	充分性 [%] (↑)	大小 (↓)	时间 [s](↓)	充分性 [%] (↑)
Ours	3.76±2.62	132.67±36.76	**100.00**	5.59±1.80	317.92±222.07	**100.00**
采样	**2.67**±3.10	**5.10±0.26**	31.37	**3.04**±3.05	**6.89±2.57**	25.49

尽管解决的是难得多的任务。我们 NAM 专用算法的优势在于，它只需对单变量分量 f_i 执行对数数量的并行化查询，而非线性数量的查询；这些分量的验证速度快得多。

5.2 解释随时间的演进

为进一步评估我们的算法相较于以往方法的优势，我们分析了该方法生成的解释大小，并将其与子集极小方法进行比较，同时追踪其随时间的演变。图3展示了这项分析。结果表明，尽管子集最小方法收敛缓慢，且常常停滞于局部极小值，但我们的方法虽然因算法 2 中的预处理步骤（该步骤仅通过对数数量的并行化查询对特征进行排序）而较晚启动，却在排序完成后迅速超越了它们。此时，我们的方法所需的查询显著更少，只需像算法 4 那样对已排序的特征进行二分搜索。第二阶段不仅快得多，而且可以证明能够达到基数最小解释，即全局最优解；子集最小方法则无法做到这一点。

5.3 与纯采样方法的比较

NAM 通常被认为具有很强的可解释性，因为每个特征对应的单变量函数 f_i 便于进行简单的可视化（如图 2 所示）。最常见的做法是对相应的特征域进行采样，以获得每个单变量函数的良好近似。然而，我们在这项实验中表明，通过采样进行离散化及由此得到的解释可能会产生误导。采样过程中遗漏的峰值和其他极值可能导致解释不充分，而这在安全关键领域可能是致命的。图 2b 抽象地展示了一个极端情况，但我们也在实践中观察到了解释不充分的现象。为此，我们用 1,000 个等间距样本代替每个验证查询。生成解释后，我们使用 α, β-CROWN（表 2）检验其充分性：在 CREDIT 和 FICO HELOC 数据集上，通过采样得到的解释有一半以上无法通过验证。相比之下，我们的所有解释在构造时就保证了充分性。

6 相关工作

形式化可解释人工智能。 我们的工作与形式化可解释人工智能领域相关（Marques-Silva，2023），该领域致力于寻找具有可证明保证的解释。以往研究已针对多种模型开发了充分解释，例如

决策树（Huang 等人，2021）、线性模型（Marques-Silva 等人，2020；Subercaseaux 等人，2025）和树集成（Izza 与 Marques-Silva，2021；Ignatiev 等人，2022；Audemard 等人，2022b；2023）。其他相关研究则致力于为神经网络获得最小充分解释（La Malfa 等人，2021；Wu 等人，2023；Izza 等人，2024；Bassan 等人，2023；2025b），这些研究依赖神经网络验证查询。尽管近年来此类验证器的可扩展性不断提高，计算这类解释的成本仍然高昂，通常需要进行大量（线性或指数级）验证查询。我们的方法着眼于结构更易解释的神经网络族，尤其关注神经加性模型（NAMs），以此迈出降低这一成本的一步。

与我们的研究设定更接近的是 Harzli 等人（2023）的工作，该研究探讨了另一类可解释模型——单调 神经网络的基数最小充分解释。在该研究中，单调性简化了充分性的认证，而最小性仍然具有挑战性。相比之下，对于NAM，认证充分性 要困难得多，但由于加性结构，最小性 则更容易。

与我们的工作相关的还有 Barceló 等人（2020b）的研究。该研究表明，在众多结果中，利用排序和选择特征等思路，可以在多项式时间内为二进制输入上的线性（感知机）模型计算基数最小充分解释。然而，对于连续输入上的神经加性模型——其表达能力强得多——要获得同类解释，则会带来新的算法挑战：（i）使用可扩展的神经网络验证方法（Wang 等，2021）计算并细化每个单变量分量的界；（ii）并行认证单变量 分量，而非整个网络；（iii）采用更高效的选择阶段，利用这些外界、对分量排序，并进行对数 搜索，从而更高效地获得基数最小充分解释。

神经加性模型（NAMs）。NAMs 将广义加性模型（*GAMs*）（Hastie，2017年；Nelder 和 Wedderburn，1972）这一经典且可解释的机器学习模型家族（Caruana 等，2015；Zhong 等，2023；Liu 等，2022；Bordt 和 von Luxburg，2023；Enouen 和 Liu，2025；Chen 等，2020）中的每个单变量分量替换为神经网络，从而兼具可解释性与表达能力。NAMs 最初由 Agarwal 等，2021 提出，在表格任务上取得了有竞争力的准确率，并应用于医疗保健和 COVID-19 建模。后续研究提出了改进其训练和架构的可能方向（Radenovic 等人，2022；Chang 等，2021；Bouchiat 等，2024；Xu 等，2023），并提出了其他变体（Bechler-Speicher 等，2024；Jiao 等，2024）及应用（Thielmann 等，2024）。

7 局限性

与所有能够获得可证明最小且充分的解释的方法一样，我们的方法依赖于调用神经网络验证查询，而这类查询目前还无法扩展到最先进的模型。不过，近年来神经网络验证取得了快速进展（Kaulen 等人，2025；Chiu 等人，2025；Wu 等人，2024a），我们的方法也将随之提升可扩展性。重要的是，我们的方法有两项重大改进：（i）将查询数量从指数级（或在放宽的任务中从线性级）降至对数级；（ii）作用于单变量 分量f_i，由于经认证的模型规模较小，且在设计上比整个大型模型更具可解释性，因此验证成本低得多 f。这些优势使我们的算法对神经加性模型更加实用；我们也证明，该算法确实能在标准基准上高效地产生解释，而以往的算法在这些基准上会失效。我们的方法还依赖于完备验证器，而这类验证器通常带有数值容差和超时限制。关于这一点的更详细讨论见 D 节。

8 结论

可证明的最小且充分的解释是可解释性领域极具价值的目标，因为它们能对忠实性和简洁性提供可验证的保证。然而，对于标准神经网络而言，这项任务在计算上难以承受，需要进行指数级数量的验证查询。我们提出了一种 NAM 专用算法，将复杂度从指数级降低至对数级并行化查询，大幅提升了速度并显著缩小了解释大小。此外，我们还表明，这些解释揭示了基于采样的方法无法捕捉的 NAM 洞见。因此，我们的工作使可证明的解释在实践中切实可行，并为将其拓展至其他可解释神经网络类别开辟了道路。

致谢

Elboher、Bassan 和 Katz 的工作部分由欧盟（欧洲研究委员会、VeriDeL，101112713）资助。但文中表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧洲研究委员会执行局的观点。欧盟和资助机构均不对此承担责任。Elboher、Bassan 和 Katz 的工作还获得了以色列科学基金会的资助（资助编号 558/24）。Ladner、Sahin 和 Althoff 的工作由德国研究联合会（德国研究联合会，DFG）通过资助项目 AL 1185/33-1 提供支持。但文中表达的观点和意见仅代表作者本人，并不一定反映欧盟或任何其他资助机构的观点。

参考文献

Federico Adolfi、Martina Vilas 和 Todd Wareham。《内部可解释性的电路发现的计算复杂性》。载于 第 *13* 届国际学习表征会议（*ICLR*）论文集，2025年。

Rishabh Agarwal、Levi Melnick、Nicholas Frosst、Xuezhou Zhang、Ben Lengerich、Rich Caruana 和 Geoffrey E Hinton。《神经加性模型：使用神经网络的可解释机器学习》。载于第 *35* 届神经信息处理系统会议（*NeurIPS*）论文集，2021年。

David Alvarez Melis 和 Tommi Jaakkola。《借助自解释神经网络实现稳健的可解释性》。载于 第 *32* 届神经信息处理系统会议（*NeurIPS*）论文集，2018年。

Guy Amir、Shahaf Bassan 与 Guy Katz。难以解释：论分布内模型解释的计算难度。载于第*27*届欧洲人工智能会议论文集（*ECAI*），2024年。

Alaa Anani、Tobias Lorenz、Mario Fritz 和 Bernt Schiele。《通过随机平滑实现像素级认证解释》。载于第*42*届机器学习国际会议（*ICML*）论文集，2025年。Marcelo Arenas、Pablo Barceló、Miguel Romero Orth 和 Bernardo Subercaseaux。《关于决策树概率解释的计算》。载于第*35*届神经信息处理系统进展国际会议（*NeurIPS*）论文集，第28695–28707页，2022年。S. Arora 和 B. Barak。《计算复杂性：现代方法》。剑桥大学出版社，2009年。G. Audemard、S. Bellart、L. Bounia、F. Koriche、J. Lagniez 和 P. Marquis。《布尔分类器的计算可理解性》。载于第*18*届知识表示与推理原理国际会议（*KR*）论文集，第74–86页，2021年。Gilles Audemard、Steve Bellart、Louenas Bounia、Frédéric Koriche、Jean-Marie Lagniez 和 Pierre Marquis。《决策树和随机森林的优选溯因解释》。载于第*31*届人工智能国际联合会议（*IJCAI*）论文集，第643–650页，2022a。Gilles Audemard、Steve Bellart、Louenas Bounia、Frédéric Koriche、Jean-Marie Lagniez 和 Pierre Marquis。《以复杂性换取随机森林解释的稀疏性》。载于第*36*届*AAAI*人工智能会议论文集，第5461–5469页，2022b。Gilles Audemard、Jean-Marie Lagniez、Pierre Marquis 和 Nicolas Szczepanski。《提升树的溯因解释计算》。载于人工智能与统计学国际会议（*AISTATS*）论文集，第4699–4711页，2023年。Steve Azzolin、Sagar Malhotra、Andrea Passerini 和 Stefano Teso。《超越拓扑自解释GNN：形式化可解释性视角》。载于第*42*届机器学习国际会议（*ICML*）论文集，2025年。P. Barceló、M. Monet、J. Pérez 和 B. Subercaseaux。《从计算复杂性的视角审视模型可解释性》。载于第*34*届神经信息处理系统会议（*NeurIPS*）论文集，第15487–15498页，2020a。

Pablo Barceló、Mikaël Monet、Jorge Pérez 和 Bernardo Subercaseaux。《从计算复杂性的视角审视模型可解释性》。第*33*届神经信息处理系统进展国际会议（*NeurIPS*）论文集，第15487–15498页，2020b。Shahaf Bassan 和 Guy Katz。《迈向形式化XAI：对神经网络最小解释进行形式化近似》。载于第*29*届系统构造与分析工具和算法国际会议（*TACAS*）论文集，第187–207页，2023年。Shahaf Bassan、Guy Amir、Davide Corsi、Idan Refaeli 和 Guy Katz。《在反应式系统中形式化解释神经网络》。载于第*23*届计算机辅助设计形式化方法国际会议（*FMCAD*）论文集，第1–13页，2023年。Shahaf Bassan、Guy Amir 和 Guy Katz。《局部与全局可解释性：计算复杂性视角》。载于第*41*届机器学习国际会议（*ICML*）论文集，第3133–3167页，2024年。Shahaf Bassan、Guy Amir、Meirav Zehavi 和 Guy Katz。《什么让集成（不可）解释？》。载于第*42*届机器学习国际会议（*ICML*）论文集，2025a。Shahaf Bassan、Yizhak Yisrael Elboher、Tobias Ladner、Matthias Althoff 和 Guy Katz。《解释的快与慢：可证明解释的抽象与细化》。载于第*42*届机器学习国际会议（*ICML*）论文集，2025b。Shahaf Bassan、Ron Eliav 和 Shlomit Gur。《简要解释自己！具有简洁充分理由的自解释神经网络》。载于第*13*届学习表征国际会议（*ICLR*）论文集，2025c。Shahaf Bassan、Shlomit Gur、Sergey Zeltyn、Konstantinos Mavrogiorgos、Ron Eliav 和 Dimosthenis Kyriazis。《用于业务流程监控的自解释神经网络》。2025d。技术报告。https://arXiv:2503.18067。

Shahaf Bassan、Michal Moshkovitz 和 Guy Katz。《加性模型解析：一种计算复杂性方法》。发表于 第 *39* 届神经信息处理系统会议论文集（*NeurIPS*），2025e。

Shahaf Bassan、Xuanxiang Huang和Guy Katz。统一形式化解释：复杂性理论视角。载于第*14*届学习表征国际会议（*ICLR*）论文集，2026年。Maya Bechler-Speicher、Amir Globerson和Ran Gilad-Bachrach。可理解且有效的图神经加性网络。第*38*届神经信息处理系统会议（*NeurIPS*）论文集，第90552–90578页，2024年。Guy Blanc、Jane Lange和Li-Yang Tan。可证明高效、简洁且精确的解释。载于第*35*届神经信息处理系统会议（*NeurIPS*）论文集，第6129–6141页，2021年。Guy Blanc、Caleb Koch、Jane Lange和Li-Yang Tan。寻找反事实的查询最优算法。载于第*39*届机器学习国际会议（*ICML*）论文集，第2075–2090页，2022年。Sebastian Bordt和Ulrike von Luxburg。从Shapley值到广义加性模型再返回。载于第*26*届人工智能与统计学国际会议（*AISTATS*）论文集，第709–745页，2023年。Kouroche Bouchiat、Alexander Immer、Hugo Yèche、Gunnar Ratsch和Vincent Fortuin。用贝叶斯原理改进神经加性模型。载于第*41*届机器学习国际会议（*ICML*）论文集，第4416–4443页，2024年。Ryma Boumazouza、Fahima Cheikh-Alili、Bertrand Mazure和Karim Tabia。ASTERYX：一种与模型无关的基于SAT的符号解释与基于分数的解释方法。载于第*30*届*ACM*信息与知识管理国际会议（*CIKM*）论文集，第120–129页，2021年。Ryma Boumazouza、Raya Elsaleh、Melanie Ducoffe、Shahaf Bassan和Guy Katz。FAME：面向神经网络的形式化抽象最小解释。载于第*14*届学习表征国际会议（*ICLR*）论文集，2026年。

Louenas Bounia。用简洁可靠的概率解释增强布尔决策树的可理解性。载于第*20*届知识型系统中的信息处理与不确定性管理国际会议（*IPMU*）论文集，第1174卷，第205–218页，2024年。Louenas Bounia。使用次模优化近似基于树的模型的最小规模溯因路径解释。载于第*41*届人工智能中的不确定性会议（*UAI*）论文集，2025年。Louenas Bounia和Frederic Koriche。通过超模最小化近似概率解释。载于第*39*届人工智能中的不确定性国际会议（*UAI*）论文集，第216–225页，2023年。Christopher Brix、Stanley Bak、Taylor T. Johnson和Haoze Wu。第五届神经网络国际验证竞赛（VNN-COMP 2024）：总结与结果。2024年。技术报告。https://arXiv:2412.19985。Brandon Carter、Jonas Mueller、Siddhartha Jain和David Gifford。是什么让你这样做？用充分输入子集理解黑盒决策。载于第*22*届人工智能与统计学国际会议（*AISTATS*）论文集，第567–576页，2019年。Rich Caruana、Yin Lou、Johannes Gehrke、Paul Koch、Marc Sturm和Noemie Elhadad。医疗保健可理解模型：预测肺炎风险与医院30天再入院。载于第*21*届*ACM*知识发现与数据挖掘国际会议（*SIGKDD*）论文集，第1721–1730页，2015年。Chun-Hao Chang、Rich Caruana和Anna Goldenberg。NODE-GAM：面向可解释深度学习的神经广义加性模型。载于第*9*届学习表征国际会议（*ICLR*）论文集，2021年。Hong Chen、Yingjie Wang、Feng Zheng、Cheng Deng和Heng Huang。稀疏模态加性模型。*IEEE*神经网络与学习系统汇刊，32(6):2373–2387，2020年。Hong-Ming Chiu、Hao Chen、Huan Zhang和Richard Y Zhang。SDP-CROWN：借助半定规划紧性实现神经网络验证的高效界传播。载于第*42*届机器学习国际会议（*ICML*）论文集，2025年。Daniel Dadush、Bento Natura和László A. Végh。重访线性规划的Tardos框架：使用近似求解器获得更快的精确解。载于第*61*届计算机科学基础年度研讨会（*FOCS*）论文集，第931–942页，2020年。Adnan Darwiche。面向可解释人工智能的逻辑。载于第*38*届*ACM/IEEE*计算机科学中的逻辑研讨会（*LICS*）论文集，第1–11页，2023年。Adnan Darwiche和Auguste Hirth。论决策背后的原因。载于第*24*届欧洲人工智能会议（*ECAI*）论文集，第712–720页，2020年。Adnan Darwiche和Chunxi Ji。论必要且充分解释的计算。载于第*36*届*AAAI*人工智能会议论文集，第5582–5591页，2022年。Alessandro De Palma、Greta Dolcetti和Caterina Urban。面向神经网络的更快经验证的解释。2025年。技术报告。https://arXiv:2512.00164。Ruediger Ehlers。分段线性前馈神经网络的形式化验证。载于第*15*届验证与分析自动化技术国际研讨会（*ATVA*）论文集，第269–286页。Springer，2017年。James Enouen和Yan Liu。InstaSHAP：可解释加性模型即时解释Shapley值。载于第*13*届学习表征国际会议（*ICLR*）论文集，2025年。Thomas Fel、Mélanie Ducoffe、David Vigouroux、Rémi Cadène、Mikael Capelle、Claire Nicodème和Thomas Serre。别骗我！基于经验证扰动分析的稳健高效可解释性。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第16153–16163页，2023年。

Claudio Ferrari、MarkNiklas Mueller、Nikola Jovanovi´c 和 MartinVechev。《通过多神经元松弛引导的分支定界实现完全验证》。载于 第十届国际学习表征会议（*ICLR*）论文集，2022年。

Timon Gehr、Matthew Mirman、Dana Drachsler-Cohen、Petar Tsankov、Swarat Chaudhuri 和 Martin Vechev。《AI2：通过抽象解释实现神经网络的安全性与鲁棒性认证》。载于 *IEEE*安全与隐私研讨会论文集（*S&P*），第3–18页。IEEE，2018年。

Sven Gowal、Krishnamurthy Dvijotham、Robert Stanforth、Rudy Bunel、Chongli Qin、Jonathan Uesato、Relja Arandjelovic、Timothy Mann 和 Pushmeet Kohli。《区间界传播在训练可验证鲁棒模型中的有效性》。2018年。技术报告。https://arXiv:1810.12715。

Itamar Hadad、Shahaf Bassan 和 Guy Katz。《形式化机制可解释性：具有可证明保证的自动化电路发现》。载于第 *14* 届国际学习表征会议（*ICLR*）论文集，2026年。

Ouns El Harzli、Bernardo Cuenca Grau 和 Ian Horrocks。《单调神经网络的基数最小解释》。载于第 *32* 届国际人工智能联合会议（*IJCAI*）论文集，第 3677–3685 页，2023年。

Trevor J. Hastie。《广义加性模型》。载于*S*语言中的统计模型，第 249–307 页。劳特利奇出版社，2017年。

XuanxiangHuang、YacineIzza、Alexey Ignatiev 和 JoaoMarques-Silva。《高效解释基于图的分类器》。载于第 *18* 届知识表示与推理原理国际会议（*KR*）论文集，2021年。

Alexey Ignatiev、Nina Narodytska 和 Joao Marques-Silva。基于溯因的解释：机器学习模型。在 *AAAI* 人工智能会议论文集，第 1511–1519 页，2019年。

Alexey Ignatiev、Yacine Izza、Peter Stuckey 和 Joao Marques-Silva。使用最大可满足性高效解释树集成。在第 *36* 届 *AAAI* 人工智能会议论文集，第 3776–3785 页，2022年。

Yacine Izza 和 Joao Marques-Silva。使用 SAT 解释随机森林。在第 *30* 届人工智能国际联合会议（*IJCAI*）论文集，2021年。

YacineIzza、Xuanxiang Huang、AntonioMorgado、JordiPlanes、AlexeyIgnatiev 和 Joao Marques-Silva。距离受限解释：理论基础与高效实现。在第 *21* 届知识表示与推理原理国际会议（*KR*）论文集，第 475–486 页，2024年。

Yining Jiao、Carlton J. Zdanski、Julia S. Kimbell、Andrew Prince、Cameron Worden、Samuel Kirse、ChristopherRutter、Benjamin Shields、William Dunn、Jisan Mahmud 等。《NAISR：用于可解释形状表示的三维神经加性模型》。载于第 *12* 届国际学习表征会议论文集（*ICLR*），2024年。

Helen Jin、Anton Xue、Weiqiu You、Surbhi Goel 和 Eric Wong。《特征归因的概率稳定性保证》。载于第 *39* 届神经信息处理系统会议论文集（*NeurIPS*），2025年。Guy Katz、Clark Barrett、David L. Dill、Kyle Julian 和 Mykel J. Kochenderfer。《Reluplex：一种用于验证深度神经网络的高效 SMT 求解器》。载于第 *29* 届计算机辅助验证国际会议论文集（*CAV*），第 97–117 页，2017年。

Guy Katz、Derek A. Huang、Duligur Ibeling、Kyle Julian、Christopher Lazarus、Rachel Lim、Parth Shah、Shantanu Thakoor、Haoze Wu、Aleksandar Zelji´c、David L. Dill、Mykel J. Kochenderfer 和 Clark Barret。用于深度神经网络验证与分析的 Marabou 框架。载于第*29*届国际计算机辅助验证会议论文集（*CAV*），第443–452页，2019年。

Konstantin Kaulen、Tobias Ladner、Stanley Bak、Christopher Brix、Hai Duong、Thomas Flinkow、Taylor T. Johnson、Lukas Koller、Edoardo Manino、ThanhVu H. Nguyen 和 Haoze Wu。第六届国际神经网络验证竞赛（VNN-COMP 2025）：总结与结果。2025年。技术报告。https://arXiv:2512.19007。Young Kyung Kim、Juan Matias Di Martino 和 Guillermo Sapiro。通过统计多模态分析推广神经加性模型。机器学习研究汇刊（*TMLR*），2024年。Suhas Kotha、Christopher Brix、J. Zico Kolter、Krishnamurthy Dvijotham 和 Huan Zhang。可证明地界定神经网络原像。载于第*37*届神经信息处理系统会议论文集（*NeurIPS*），第80270–80290页，2023年。Mark W. Krentel。优化问题的复杂性。载于第*18*届*ACM*计算理论年度研讨会论文集（*STOC*），第69–76页，1986年。

Emanuele La Malfa、Agnieszka Zbrzezny、Rhiannon Michelmore、Nicola Paoletti 和 Marta Kwiatkowska。《关于自然语言处理模型的有保证最优鲁棒解释》。载于人工智能国际联合会议（*IJCAI*）论文集，第 2658–2665 页，2021年。

Faezeh Labbaf、Tomáš Kolárik、Martin Blicha、Grigory Fedyukovich、Michael Wand 和 Natasha Sharygina。《神经网络分类的空间解释》。载于第 *37* 届计算机辅助验证国际会议（*CAV*）论文集，第 287–303 页，2025年。

Chu Min Li 和 Felip Manya。《最大可满足性、硬约束与软约束》。载于可满足性手册，第 613–631 页。IOS出版社，2009年。

Jiachang Liu、Chudi Zhong、Margo Seltzer 和 Cynthia Rudin。《广义线性模型与加性模型的快速稀疏分类》。发表于第*25*届人工智能与统计学国际会议（*AISTATS*）论文集，第9304页，2022年。Scott M. Lundberg 和 Su-In Lee。《解释模型预测的统一方法》。发表于第*30*届神经信息处理系统进展国际会议（*NeurIPS*）论文集，2017年。E manuele La Malfa、Agnieszka Zbrzezny、Rhiannon Michelmore、Nicola Paoletti 和 Marta Kwiatkowska。《自然语言处理模型的保证最优鲁棒解释》。发表于国际人工智能联合会议（*IJCAI*）论文集，2021年。J. Marques-Silva、T. Gerspacher、M. Cooper、A. Ignatiev 和 N. Narodytska。《以多项式时间与延迟解释朴素贝叶斯及其他线性分类器》。发表于第*33*届神经信息处理系统进展国际会议（*NeurIPS*）论文集，第20590–20600页，2020年。Joao Marques-Silva。《机器学习中的逻辑可解释性》。发表于第*18*届推理网络暑期国际学校论文集：因果关系、解释与声明式知识，第24–104页，2023年。Joao Marques-Silva 和 Alexey Ignatiev。《通过形式化可解释人工智能实现可信人工智能》。发表于第*36*届*AAAI*人工智能会议论文集，第12342–12350页，2022年。Joao Marques-Silva 和 Alexey Ignatiev。《没有灵丹妙药：可解释机器学习模型必须得到解释》。发表于《人工智能前沿》，第6卷：1128212，2023年。Joao Marques-Silva、Thomas Gerspacher、Martin Cooper、Alexey Ignatiev 和 Nina Narodytska。《单调分类器的解释》。发表于第*38*届机器学习国际会议（*ICML*）论文集，第7469–7479页，2021年。Reda Marzouk 和 Colin De La Higuera。《马尔可夫分布下SHAP解释的可处理性》。发表于第*41*届机器学习国际会议（*ICML*）论文集，第34961–34986页，2024年。Reda Marzouk、Shahaf Bassan 和 Guy Katz。《SHAP遇上张量网络：具有并行性的可证明可处理解释》。发表于第*39*届神经信息处理系统会议（*NeurIPS*）论文集，2025a年。

Reda Marzouk、Shahaf Bassan、Guy Katz 和 De la Higuera。《众多Shapley值的计算可处理性》。发表于第*28*届人工智能与统计学国际会议（*AISTATS*）论文集，2025b年。Mark N. Müller、Gleb Makarchuk、Gagandeep Singh、Markus Püschel 和 Martin Vechev。《PRIMA：通过可扩展凸包近似实现通用且精确的神经网络认证》。发表于第*6*届*ACM*编程语言会议（*POPL*）论文集，第1–33页，2022年。John Ashworth Nelder 和 Robert WM Wedderburn。《广义线性模型》。发表于《英国皇家统计学会期刊》 *A* 辑：社会统计学，135(3):370–384，1972年。Filip Radenovic、Abhimanyu Dubey 和 Dhruv Mahajan。《用于可解释性的神经基模型》。发表于第*36*届神经信息处理系统会议（*NeurIPS*）论文集，第8414–8426页，2022年。M. Ribeiro、S. Singh 和 C. Guestrin。《我为什么应该相信你？解释任意分类器的预测》。发表于第*22*届知识发现与数据挖掘国际会议（*SIGKDD*）论文集，第1135–1144页，2016年。M. Ribeiro、S. Singh 和 C. Guestrin。《Anchors：高精度模型无关解释》。发表于第*32*届*AAAI*人工智能会议论文集，2018年。Marco Sälzer 和 Martin Lange。《即使是最简单的神经网络，可达性问题也是NP完全的》。发表于第*15*届可达性问题国际会议（*RP*）论文集，第149–164页，2021年。Zhouxing Shi、Qirui Jin、Zico Kolter、Suman Jana、Cho-Jui Hsieh 和 Huan Zhang。《采用分支定界法验证具有一般非线性函数的神经网络》。发表于第*31*届系统构造与分析工具和算法国际会议（*TACAS*）论文集，第315–335页，2025年。A. Shih、A. Choi 和 A. Darwiche。《解释贝叶斯网络分类器的符号方法》。发表于第*27*届国际人工智能联合会议（*IJCAI*）论文集，第5103–5111页，2018年。Gagandeep Singh、Timon Gehr、Markus Püschel 和 Martin Vechev。《用于认证神经网络的抽象域》。发表于第*3*届*ACM*编程语言会议（*POPL*）论文集，第1–30页，2019年。Jules Soria、Zakaria Chihani、Julien Girard-Satabin、Alban Grastien、Romain Xu-Darme 和 Daniela Cancila。《基于原型的网络的形式化溯因潜在解释》。发表于第*40*届*AAAI*人工智能会议论文集，2025年。Bernardo Subercaseaux、Marcelo Arenas 和 Kuldeep Meel。《线性模型的概率解释》。发表于第*39*届*AAAI*人工智能会议论文集，第20655–20662页，2025年。Mukund Sundararajan、Ankur Taly 和 Qiqi Yan。《深度网络的公理化归因》。发表于第*34*届机器学习国际会议（*ICML*）论文集，第3319–3328页，2017年。Anton Frederik Thielmann、René-Marcel Kruse、Thomas Kneib 和 Benjamin Säfken。《位置、尺度与形状的神经加性模型：一种超越均值的可解释神经回归框架》。发表于第*27*届人工智能与统计学国际会议（*AISTATS*）论文集，第1783–1791页，2024年。Vincent Tjeng、Kai Y. Xiao 和 Russ Tedrake。《使用混合整数规划评估神经网络的鲁棒性》。发表于第*7*届学习表征国际会议（*ICLR*）论文集，2019年。Jorge R. Vera。《有限精度算术下线性规划的复杂性》。发表于《数学规划》，80(1):91–123，1998年。Shiqi Wang、Huan Zhang、Kaidi Xu、Xue Lin、Suman Jana、Cho-Jui Hsieh 和 J. Zico Kolter。《Beta-Crown：通过逐神经元分割约束实现高效界传播，以验证神经网络的鲁棒性》。发表于第*35*届神经信息处理系统进展会议（*NeurIPS*）论文集，第29909–29921页，2021年。Haoze Wu、Omri Isac、Aleksandar Zelji´c、Teruhiro Tagomori、Matthew Daggitt、Wen Kokke、Idan Refaeli、Guy Amir、Kyle Julian、Shahaf Bassan 等。《Marabou 2.0：一种通用的神经网络形式化分析器》。发表于第*36*届计算机辅助验证国际会议（*CAV*）论文集，第249–264页，2024a年。

Min Wu、Haoze Wu 和 Clark Barrett。《Verix：迈向深度神经网络的可验证可解释性》。第 *36* 届神经信息处理系统进展国际会议论文集（*NeurIPS*），2023年。

Min Wu、Xiaofu Li、Haoze Wu 和 Clark Barrett。《面向不正确性和分布外检测的更优经验证的解释》。2024b。技术报告。https://arXiv: 2409.03060。

Shiyun Xu、Zhiqi Bu、Pratik Chaudhari 和 Ian J. Barnett。《稀疏神经加性模型：通过组稀疏实现特征选择的可解释深度学习》。发表于第 *27* 届欧洲机器学习与数据库知识发现联合会议论文集（*ECML PKDD*），第 343–359 页，2023年。

Anton Xue、Rajeev Alur 和 Eric Wong。《乘法平滑下特征归因的稳定性保证》。发表于第 *37* 届神经信息处理系统进展会议论文集（*NeurIPS*），第 62388–62413 页，2023年。

Weiqiu You、Helen Qu、Marco Gatti、BhuvneshJain 和 Eric Wong。Sum-of-parts：通过端到端学习特征组的自归因神经网络。载于第*42*届国际机器学习会议（*ICML*）论文集，2025年。

Jinqiang Yu、Alexey Ignatiev、Peter Stuckey、Nina Narodytska 和 Joao Marques-Silva。排除不可能的情况，剩下的无论是什么都必定为真：在形式化解释中提取和应用背景知识。载于第*37*届*AAAI*人工智能会议论文集，第4123–4131页，2023年。

Huan Zhang、Tsui-Wei Weng、Pin-Yu Chen、Cho-Jui Hsieh 和 Luca Daniel。《采用一般激活函数的神经网络鲁棒性高效认证》。载于第 *32* 届神经信息处理系统会议（*NeurIPS*）论文集，第 4939–4948 页，2018年。Huan Zhang、Hongge Chen、Chaowei Xiao、Sven Gowal、Robert Stanforth、Bo Li、Duane Boning 和 Cho Jui Hsieh。《迈向可验证鲁棒神经网络的稳定高效训练》。载于第 *8* 届国际学习表征会议（*ICLR*）论文集，2020年。Wei Zhang、Brian Barr 和 John Paisley。《高斯过程神经加性模型》。载于第 *38* 届 *AAAI* 人工智能会议论文集，第 15 期，第 16865–16872 页，2024年。Chudi Zhong、Zhi Chen、Jiachang Liu、Margo Seltzer 和 Cynthia Rudin。《探索并交互于优质稀疏广义加性模型集合》。第 *37* 届神经信息处理系统会议（*NeurIPS*）论文集，第 56673–56699 页，2023年。Duo Zhou、Christopher Brix、Grani A Hanasusanto 和 Huan Zhang。《利用分支定界推断割平面的可扩展神经网络验证》。载于第 *38* 届神经信息处理系统会议（*Neurips*）论文集，2024年。

附录

附录包含论文中提及的所有证明、优化方法、额外设置和补充实验：

附录 A介绍神经网络验证与形式化可解释人工智能的扩展背景。**附录 B** 给出了证明中将用到的形式化内容。**附录 C** 包含命题 1 至 4 的证明。**附录 D** 讨论了重要性排序的理论基础与实际优化。**附录 E**介绍了多分类和回归任务的扩展。**附录 F**包含所有实验细节和消融研究。**附录 G** 包含大语言模型使用情况披露。

神经网络验证与形式化可解释人工智能的扩展背景

神经网络验证。 神经网络验证可为神经网络在连续输入区域上的行为提供可证明保证。经典的基于 SMT 和混合整数线性规划的方法（Katz 等人，2017；Wu 等人，2024a；Katz 等人，2019；Tjeng 等人，2019；Ehlers，2017）将 ReLU 网络和规范编码为逻辑约束或混合整数约束，从而提供精确保证，但只能扩展到中小型模型。抽象解释方法（Singh 等人，2019；Gehr 等人，2018；Ferrari 等人，2022；Müller 等人，2022）则逐层传播过近似，能够快速生成鲁棒性证书，但并不完备。基于线性松弛的界传播取得了重大突破，其中尤以 CROWN（Zhang 等人，2018）及其扩展方法（Wang 等，2021；Chiu 等人，2025；Zhou 等，2024；Shi 等人，2025）为代表。这些方法可计算紧致的对偶线性界，既可作为可扩展但不完备的验证器，也可作为精确搜索过程中的强松弛。现代分支定界（BaB）框架基于这些松弛方法，实现了大规模的完备验证。特别是，α-β-CROWN 及相关变体如今已在 VNN-COMP（Brix 等人，2024）中占据主导地位，可处理参数量达数百万的模型。近期进展包括更紧致的松弛方法（例如半定规划混合方法（Chiu 等人，2025））、切平面增强（Zhou 等，2024），以及对非 ReLU 非线性函数的支持（Shi 等人，2025）。如今，验证方法通常能够为中等规模的卷积神经网络和残差网络认证鲁棒性，但对于 Transformer、高度复杂的架构，以及更丰富的时序或关系规范，仍面临重大挑战。

形式化可解释人工智能。 我们的工作属于一项名为形式化可解释人工智能 （形式化可解释人工智能）的研究领域（Marques-Silva & Ignatiev, 2022），该领域研究具有可证明保证的解释（Yu et al., 2023; Darwiche & Ji, 2022; Darwiche, 2023; Shih et al., 2018; Azzolin et al., 2025; Audemard et al., 2021）。由于生成此类有保证的解释通常在计算上难以处理（Barceló 等，2020a; Bassan 等人，2026; Marzouk et al., 2025a;b; Marzouk & De La Higuera, 2024; Amir et al., 2024; Blanc et al., 2021; 2022），许多研究将重点放在限制更严格的模型类别上（Marques-Silva & Ignatiev, 2023），包括决策树（Bounia, 2025; Arenas 等人, 2022; Bounia, 2024; Bounia & Koriche, 2023）、单调模型（Marques-Silva et al., 2021; Harzli et al., 2023）和树集成（Audemard et al., 2023; 2022a; Ignatiev 等人, 2022; Bassan et al., 2025a; Boumazouza et al., 2021）。与我们的工作相关的是一些针对神经网络生成此类解释的框架（Malfa et al., 2021; Bas- san et al., 2023; Hadad et al., 2026; Wu 等人，2023; Labbaf et al., 2025; De Palma et al., 2025; Fel et al., 2023; Soria et al., 2025）。鉴于这一问题具有很高的计算复杂性，一些方法尝试通过抽象（Bassan 等人，2025b; Boumazouza et al., 2026）、平滑技术（Xue et al., 2023; Ji n et al., 2025; Anani et al., 2025）或自解释框架（Alvarez Melis & Jaakkola, 2018; Bassan et al., 2025c;d; You et al., 2025）来降低其难度。另一项相关研究探讨了线性模型的可证明保证解释（Marques-Silva et al., 2020; Subercaseaux et al., 2025）；线性模型是加性模型的一个受限子类。此外，Bassan et al. (2025e) 的理论分析研究了为这类模型生成此类解释的计算复杂性。

B 形式化

B.1 计算复杂度类

我们假设读者熟悉标准复杂性类多项式时间（PTIME）和非确定性多项式时间（NP、coNP）；入门介绍请参见 Arora & Barak (2009)。NP 和 coNP 刻画了判定问题的复杂性——这类问题的输出仅为“是”或“否”。相比之下，FP 是 PTIME 的函数对应概念，用于描述这样一类函数问题：其输出可以是任意值，且能够在多项式时间内计算。
我们假设读者熟悉多项式时间（PTIME）和非确定性多项式时间（NP、coNP）的基本复杂度类（介绍见 Arora & Barak (2009)）。NP 和 coNP 是处理判定 问题的复杂度类，即确定具有是/否答案的问题的复杂度。FP 复杂度类是 PTIME 的函数 变体（也就是说，它返回一个函数作为输出，而非是/否答案），描述的是可在多项式时间内求解的函数问题类。

复杂度类*OptP* 由这样的函数组成：对 NP 机器所有接受路径上的输出值取最大值（或最小值）即可计算这些函数。或者，如果存在一个多项式 $p(\cdot)$ 和一个可在多项式时间内计算的函数 $G\colon \Sigma^* \times \{0,1\}^{p(n)} \to \mathbb{Z}$，我们称其为值函数，使得对于每个输入 $X \in \Sigma^*$：函数 $F\colon \Sigma^* \to \mathbb{Z}$ 属于 OptP

$$F(X) = \max_{Y \in \{0,1\}^{p(|X|)}} G(X, Y), \tag{3}$$

其中，Y 是一个长度为多项式的见证，函数 G 可在确定性多项式时间内计算，且整数输出 $G(X, Y)$ 在 $n = |X|$ 中有多项式界。若解 $F(X)$ 有界于 $z(n)$，则称该问题属于 OptP$[z(n)]$。

函数多项式时间$^{\text{NP}}[z(n)]$类由所有可在多项式时间内、使用至多 $z(n)$ 次 NP 预言机查询计算的函数问题构成。Krentel（1986年）的开创性工作证明了 OptP$[z(n)] \subseteq$ FP$^{\text{NP}}[z(n)]$,，此外还证明，只要 $z(n)$满足 Krentel（1986年）中定义的“平滑性”条件，函数多项式时间$^{\text{NP}}[z(n)]$ 中的每个函数都可以通过求解一个适当的 OptP$[z(n)]$ 优化问题，再进行多项式时间后处理来计算。

对于函数问题或优化问题，通常采用度量归约 （Krentel，1986年）。形式化地，给定某个 f，$g\colon \Sigma^* \to \mathbb{N}$，从 f 到 g 的度量归约是一对可在多项式时间内计算的函数 $T_1\colon \Sigma^* \times \mathbb{N} \to \mathbb{N}$，使得对所有 $\mathbf{x} \in \Sigma^*$，都有$f(\mathbf{x}) = T_2(\mathbf{x}\ g(T_1(\mathbf{x})))$。

B.2 单变量神经网络的形式化

在本小节中，我们将对分析中的单变量神经网络组件 $f_i\colon \mathbb{R} \to \mathbb{R}$ 进行形式化。设有一个包含 κ层的单变量神经网络组件 f，其输入为$\mathbf{x} \in \mathbb{R}^{n_0}$。$\mathbf{y} := f_i(\mathbf{x}) \in \mathbb{R}$的输出计算如下：

$$\mathbf{h}^0 := \mathbf{x}, \qquad \mathbf{h}^k := L_k(\mathbf{h}^{k-1}), \qquad \mathbf{y} = \mathbf{h}^\kappa, \quad k \in [\kappa],$$

其中，我们用 $L_k\colon \mathbb{R}^{n_{k-1}} \to \mathbb{R}^{n_k}$ 表示第 k 层的运算，其定义为

$$L_k(\mathbf{h}^{k-1}) := \sigma(W_k \mathbf{h}^{k-1} + \mathbf{b}^k),$$

权重矩阵记为 $W_k \in \mathbb{Q}^{n_k \times n_{k-1}}$，偏置向量 $\mathbf{b}^k \in \mathbb{Q}^{n_k}$，激活函数$\sigma\colon \mathbb{R}^{n_k} \to \mathbb{R}^{n_k}$，神经元数量为 $n_k \in \mathbb{N}$。按照标准惯例，我们假设所有权重和偏置都是以二进制编码的有理数（因为它们对应于以固定机器精度存储的训练参数），并且可以写成 $\frac{p}{q}$ 的形式，其中 p，$q \in \mathbb{N}$。此外，我们假设 σ 是标准 ReLU 激活函数 ReLU$(\mathbf{x}) := \max(0, \mathbf{x})$。NAM 是一种回归或分类模型，其核心计算形式为 $f_1(\mathbf{x}_1) + \cdots + f_n(\mathbf{x}_n) + \beta_0$，其中每个 f_i 都是一个单变量神经网络，且 $\beta_0 \in \mathbb{Q}$。我们将单变量分量 f_i 的大小 定义为其所有权重和偏置的总位长度，并将完整模型 f 的大小定义为所有单变量分量 f_i 的大小之和，再加上全局偏置项 β_0的位长度。

尽管我们采用标准前馈设定来表述该模型，但这并不失一般性：任意连接方式的模型都可以通过插入中间层并添加零权重连接，转换为等价的前馈网络，使计算图变为无环图，同时不改变其函数。此外，ReLU 网络可以轻松表示加法、标量乘法和整流函数 $\max(0, \mathbf{x})$ 等基本代数运算，因此能够表示这些运算的任意组合。因此，就我们的目的而言，将讨论范围限制在前馈 ReLU 架构上完全具有一般性。

C 证明

C.1 命题 1 的证明

命题 1。 给定一个 *NAM* f、一个输入 $\boldsymbol{x} \in \mathbb{R}^n$ 以及一个扰动半径 $\epsilon_p \in \mathbb{R}_+$，令算法 *2* 根据输入特征的重要性返回一个完整的输入特征列表排序。则以下结论成立：对于任何包含特征 i 的充分解释 $\mathcal{S}$，以及列表排序中满足$i \prec j$ 的任意特征 $j \notin \mathcal{S}$，集合 $\mathcal{S} \setminus \{i\} \cup \{j\}$ 也是一个充分解释。

证明。我们回顾一下，在算法 2 中我们假设 $f(\mathbf{x})$ 产生正预测，即被分类为 1。因此，最终的界列表 $[(\Delta l_1, \Delta u_1)、(\Delta l_2, \Delta u_2).....、(\Delta l_n, \Delta u_n)]$ 的推导方式是：取每个 $f_i(\tilde{}\mathbf{x}_i)$ 的可能最小值，并计算 $f_i(\mathbf{x}_i) - f_i(\tilde{}\mathbf{x}_i)$ 的下界和上界。我们给出的证明同样适用于预测为负的情况：在这种情况下，我们改为取 $f_i(\tilde{}\mathbf{x}_i)$ 的最大值，并对 $f_i(\tilde{}\mathbf{x}_i) - f_i(\mathbf{x}_i)$ 求界。该情况的详细讨论将留待后文。由于我们假设 $f(\mathbf{x}) \geq 0$，条件$\mathcal{S}$是相对于 $\langle f、\mathbf{x}\ \epsilon_p \rangle$ 的充分解释，意味着：

$$\begin{aligned} \forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}).\quad & f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \geq 0 \iff \\ & \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \geq 0. \end{aligned} \tag{4}$$

$f(\mathbf{x}_{\mathcal{S}};\ \tilde{}\mathbf{x}_{\bar{\mathcal{S}}})$ 的值是通过将 $\mathcal{S}$ 中的特征固定为 $\mathbf{x}$，并将互补特征 $\bar{}$features $\mathcal{S}$ 扰动为来自 $\tilde{}\mathbf{x}$ 的取值而获得的。由于这一构造，以及 NAM f 的加性形式可以表示为 $f(\mathbf{x}) := \sum_{t\in[n]} f_t(\mathbf{x}_t)$，我们可以得出以下结论：

$$\begin{aligned} \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} l_t \leq (f_t(\mathbf{x}_t) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} \sum_{t\in\bar{\mathcal{S}}} f_t(\tilde{\mathbf{x}}_t)) \leq \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} u_t \iff \\ \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} l_t \leq \sum_{t\in\bar{\mathcal{S}}} (f_t(\mathbf{x}_t) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t)) + \sum_{t\in\mathcal{S}} (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t)) \leq \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} u_t \iff \\ \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} l_t \leq \sum_{t\in[n]} f(\mathbf{x}) - \sum_{t\in\mathcal{S}} f_t(\mathbf{x}_t) - \sum_{t\in\bar{\mathcal{S}}} \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\tilde{\mathbf{x}}_t) \leq \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} u_t \iff \\ \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} l_t \leq (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})) \leq \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} u_t. \end{aligned} \tag{5}$$

根据我们先前在公式 4 中的假设，我们知道 $\sum_{t\in\bar{\mathcal{S}}} \hat{}\Delta u_t \geq 0$。现在定义 $\mathcal{S}' := \mathcal{S} \cup \{j\} \setminus \{i\}$。按照之前相同的推理过程，我们得到：

$$\begin{aligned} \sum_{t\in\bar{\mathcal{S}'}} \hat{\Delta} l_t \leq (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}'}})) \leq \sum_{t\in\bar{\mathcal{S}'}} \hat{\Delta} u_t \iff \\ \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} l_t + \hat{\Delta} l_j - \hat{\Delta} l_i \leq (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}'}})) \leq \sum_{t\in\bar{\mathcal{S}}} \hat{\Delta} u_t + \hat{\Delta} u_j - \hat{\Delta} u_i. \end{aligned} \tag{6}$$

$\hat{}\hat{}\hat{}\hat{}\hat{}\hat{}$ 由于我们假设 $i \prec j$ 按照 $[(\Delta l_1, \Delta u_1)、(\Delta l_2, \Delta u_2) \ldots、(\Delta l_n, \Delta u_n)]$ 的排序，且 $\hat{}\hat{}\hat{}\hat{}$ 各界互不相交，因此可得 $\Delta l_j - \Delta l_i \geq 0$ 和 $\Delta u_j - \Delta u_i \geq 0$。由此，我们得出，$(f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'};\ \tilde{}\mathbf{x}_{\bar{\mathcal{S}'}}))$ 的上下界均小于 $(f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}};\ \tilde{}\mathbf{x}_{\bar{\mathcal{S}}}))$ 的对应值。这进而意味着：

$$(f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'})) - (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})) \leq 0 \iff \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \leq 0. \tag{7}$$

此外，由于根据公式 4，$\min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})$ 非负，因此可得：

$$\min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \geq \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \geq 0 \implies \forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \geq 0. \tag{8}$$

这证明了 $\mathcal{S}'$ 构成了关于 $\langle f, \mathbf{x}\ \epsilon_p \rangle$的充分解释，从而结束了这部分证明。

现在我们转向对称情形，其中 $f(\mathbf{x}) < 0$。在这种设定下，算法 2 通过对每个 $f_i(\tilde{\mathbf{x}}_i)$ 取最大可行值，并为 $f_i(\tilde{\mathbf{x}}_i) - f_i(\mathbf{x}_i)$ 推导相应的上界和下界，从而对称地应用。 $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2) \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$ 现在按重要性降序排列，而非升序。根据假设 $f(\mathbf{x}) < 0$, 关于 $\langle f, \mathbf{x}\ \epsilon_p \rangle$，要求$\mathcal{S}$构成充分解释，可表示为：

$$\forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) < 0 \iff \max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) < 0. \tag{9}$$

与前述情形类似，利用 NAM f 的加性结构，可将其写作 $\hat{f}(\mathbf{x}) := \sum_{t \in [n]} f_t(\mathbf{x}_t)$；结合 $f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})$ 的定义以及界 Δl_i 和 Δu_i，我们可以推导出以下一系列陈述：

$$\begin{aligned}
\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} l_t \leq \sum_{t \in \bar{\mathcal{S}}} (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t) - f_t(\mathbf{x}_t)) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} u_t \iff \\
\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} l_t \leq \sum_{t \in \bar{\mathcal{S}}} (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t) - f_t(\mathbf{x}_t)) + \sum_{t \in \mathcal{S}} (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t)) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} u_t \iff \\
\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} l_t \leq \sum_{t \in \bar{\mathcal{S}}} \max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t) + \sum_{t \in \mathcal{S}} f_t(\mathbf{x}_t) - \sum_{t \in [n]} f_t(\mathbf{x}_t) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} u_t \iff \\
\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} l_t \leq (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - f(\mathbf{x})) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} u_t.
\end{aligned} \tag{10}$$

由于我们知道 $\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} u_t \geq 0$，并按照之前的定义令 $\mathcal{S}' := \mathcal{S} \cup \{j\} \setminus \{i\}$，现在可以推导出：

$$\begin{aligned}
\sum_{t \in \bar{\mathcal{S}}'} \hat{\Delta} l_t \leq (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})) \leq \sum_{t \in \bar{\mathcal{S}}'} \hat{\Delta} u_t \iff \\
\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} l_t + \hat{\Delta} l_i - \hat{\Delta} l_j \leq (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta} u_t + \hat{\Delta} u_i - \hat{\Delta} u_j.
\end{aligned} \tag{11}$$

$\hat{\Delta}\hat{\Delta}\hat{\Delta}\hat{\Delta}\hat{\Delta}\hat{\Delta}$ 与之前一样，由于我们假设 $i \prec j$ 在 $[(\Delta l_1, \Delta u_1)、(\Delta l_2, \Delta u_2) \ldots、(\Delta l_n, \Delta u_n)]$, 的排序中，并且已知这些界互不相交，再结合列表按值递减排序这一假设，可知 $\Delta l_i - \Delta l_j \geq 0$ 和 $\Delta u_i - \Delta u_j \geq 0$。因此，我们得到不同的结果：也就是说，$\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} \left(f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})\right)$ 的上下界都严格大于 $\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} \left(f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - f(\mathbf{x})\right)$。这又意味着：

$$
\begin{aligned}
(\max_{\tilde{\mathbf{x}}\in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})) - (\max_{\tilde{\mathbf{x}}\in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - f(\mathbf{x})) \le 0 \iff \\
\max_{\tilde{\mathbf{x}}\in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - \max_{\tilde{\mathbf{x}}\in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \le 0.
\end{aligned} \tag{12}
$$

此外，由于公式 9 保证 $\max_{\tilde{\mathbf{x}}\in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}})$ 为负，因此我们得到：

$$
\begin{aligned}
\max_{\tilde{\mathbf{x}}\in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \le \max_{\tilde{\mathbf{x}}\in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) < 0 \implies \\
\forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}'};\tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) < 0.
\end{aligned} \tag{13}
$$

这表明 $\mathcal{S}'$ 是 $\langle f, \mathbf{x}\ \epsilon_p\rangle$的充分解释。据此， $f(\mathbf{x})$的否定情形已得到解决，再结合肯定情形，证明即告完成。

□

C.2 命题 2 的证明

命题 2. 给定 ρ 并行处理器，算法 *2* 总共执行 $T_p(n) = \mathcal{O}\Big(\binom{n}{p}\log\big(\max^{i\in[n]}\binom{\beta_i-\alpha_i}{\xi_i}\big)\Big) \xrightarrow[p\to n]{} \mathcal{O}\Big(\log\big(\bar{\max}^{i\in[n]}\binom{\beta_i-\alpha_i}{\xi_i}\big)\Big)$ 次验证器调用，每次作用于一个$f_i(\cdot)$分量，其中 $\xi_i := \min\{|\hat{\Delta}l_{i+1} - \hat{\Delta}u_i|, |\hat{\Delta}l_i - \hat{\Delta}u_{i-1}|\}$。

证明。 当不存在任何两个单变量函数 f_i 和 f_j 具有重叠边界时，算法终止。由于该过程依赖二分搜索，每个阶段都会将当前区间一分为二。初始时，特征 i 的上界与下界之间的间距恰好为 $\alpha_i - f_i(\mathbf{x}_i)$。特征 i 所达到的精度受两个距离中较小者限制：一个是它与排序中紧邻其上方的特征界限（$i+1$）之间的距离，另一个是它与紧邻其下方的特征界限（$i-1$）之间的距离。因此，我们将特征 i的整体精度记为 ξ_i：

$$
\xi_i := \min\{|\hat{\Delta}l_{i+1} - \hat{\Delta}u_i|, |\hat{\Delta}l_i - \hat{\Delta}u_{i-1}|\}. \tag{14}
$$

总体而言，考虑到二分搜索过程在每次迭代时都会将区间一分为二，我们将针对单个特征 i 执行的划分次数定义为 k_i：

$$
\begin{aligned}
\frac{\beta_i - \alpha_i}{2^{k_i}} \le \xi_i \iff \\
k_i \le \mathcal{O}\big(\log(\frac{\beta_i - \alpha_i}{\xi_i})\big).
\end{aligned} \tag{15}
$$

因此，执行划分次数最多的特征记为 $k_{\max}$，其为：

$$
k_{max} \le \mathcal{O}\Big(\max_{i\in[n]}\big(\log(\frac{\beta_i - \alpha_i}{\xi_i})\big)\Big). \tag{16}
$$

因此，每个特征 $i \in [n]$最多需要 k_{max} 次验证查询。由此，总工作负载的上界为 $n \cdot k_{max}$；将其分配到 ρ 个线程后，可得以下并行复杂度结果 $T_p(n)$：

$$
T_\rho(n) \le \mathcal{O}\big((\frac{n}{\rho}) \cdot k_{max}\big) \le \mathcal{O}\Big(\big(\frac{n}{\rho}\big)\log\big(\max_{i\in[n]}(\frac{\beta_i - \alpha_i}{\xi_i})\big)\Big). \tag{17}
$$

证明至此完成。

□

$$T_\rho(n) \leq \mathcal{O}\big((\frac{n}{\rho}) \cdot k_{max}\big) \leq \mathcal{O}\Big(\big(\frac{n}{\rho}\big) \max\big(\log\big(\max_{i\in[n]}(\frac{\beta_i - \alpha_i}{\xi_i})\big), \xi'\big)\Big). \tag{18}$$

C.3 命题 3 的证明

命题 3. 给定一个 *NAM* f、一个输入 $\boldsymbol{x} \in \mathbb{R}^n$ 和一个扰动半径 $\epsilon_p \in \mathbb{R}_+$，算法 *3* 执行 $\mathcal{O}(n)$ 次查询，并返回一个基数最小的充分解释。这与算法 *1* 形成对比，后者仅保证返回一个子集极小的充分解释。

证明。 我们首先注意到，该算法以迭代方式运行，对查询 suff(f, $\mathbf{x}$ $\mathcal{S} \setminus \{i\}$, ϵ_p)进行 $|n|$ 次调用。每次此类查询都可以使用神经网络验证器进行编码，这意味着该算法总共需要调用 $\mathcal{O}(n)$ 次。现在我们转而证明，算法 3 确实能够针对 $\langle f, \mathbf{x}\ \epsilon_p\rangle$ 生成一个基数最小充分解释。首先，让我们证明算法 3 提供了一个有效的充分解释。该结论是直接的，因为最后检查的条件是：suff(f, $\mathbf{x}$ $\mathcal{S} \setminus \{i\}$, ϵ_p)，且在满足该条件后，$\mathcal{S}$ 会被更新为 $\mathcal{S} \setminus \{i\}$ 并返回。因此，根据定义，所返回子集的充分性得以满足。

我们现在将证明，生成的集合 $\mathcal{S}$ 是针对 $\langle f, \mathbf{x}\ \epsilon_p\rangle$ 的基数最小充分解释。令 $1 \leq \ell \leq n$ 表示在算法 3 中最后一次尝试从 $\mathcal{S}$ 中移除的特征。那么，对于 $\mathcal{S}' := \mathcal{S} \setminus \{\ell\}$，可得：suff($f$, $\mathbf{x}$ $\mathcal{S}'$, ϵ_p) 不成立，这意味着 $\mathcal{S}'$ 不是 $\langle f, \mathbf{x}\ \epsilon_p\rangle$ 的充分解释。我们首先证明一个有助于证明该命题的引理：

引理 *1*。给定一个 *NAM* f，令算法 *2* 返回排序后的不相交配对列表：$\hat{}\hat{}\hat{}\hat{}\hat{}\hat{}$[$(\Delta l_1, \Delta u_1)$, $(\Delta l_2, \Delta u_2)$ . . . , $(\Delta l_n, \Delta u_n)$]。那么，以下结论成立：如果 $\mathcal{S}$ 表示按 [$(\Delta l_1, \Delta u_1)$, $(\Delta l_2, \Delta u_2)$ . . . , $(\Delta l_n, \Delta u_n)$] 排序后的前 $\hat{}\hat{}\hat{}\hat{}\hat{}\hat{}$]$|\mathcal{S}|$ 个特征，并且它不是关于 $\langle f, x\ \epsilon_p\rangle$ 的充分解释，那么任何大小为 $|\mathcal{S}|$ 的子集 $\mathcal{S}' \subseteq [n]$ 也不是关于 $\langle f, x\ \epsilon_p\rangle$ 的充分解释。

证明。 我们首先注意到，$\mathcal{S} \subseteq [n]$ 相对于$\langle f, \mathbf{x}\ \epsilon_p\rangle$并非充分解释。假设存在一个与 $\mathcal{S}$ 基数相同的 $\mathcal{S}' \neq \mathcal{S}$，它相对于 $\langle f, \mathbf{x}\ \epsilon_p\rangle$是充分解释，这将导致矛盾。由于 $\mathcal{S}$ 和 $\mathcal{S}'$ 的大小相同，我们可以根据排序 $\hat{}\hat{}\hat{}\hat{}\hat{}\hat{}$[$(\Delta l_1, \Delta u_1)$, $(\Delta l_2, \Delta u_2)$ . . . , $(\Delta l_n, \Delta u_n)$] 中的位置，将 $\mathcal{S}'$ 中的每个特征与 $\mathcal{S}$ 中的一个特征对应起来。

根据定义，$\mathcal{S}$ 由该排序中排名最高的 $|\mathcal{S}|$个特征组成。因此，在这种映射下，$\mathcal{S}'$ 中的每个特征都会被映射到 $\mathcal{S}$中排名严格更高或相同的特征。现在考虑一系列替换操作：每一步都将 $\mathcal{S}'$ 中的一个特征替换为 $\mathcal{S}$中与之对应且等价或排名更高的特征。命题 1保证每次替换都能保持充分性，因为这是用“排名更高”的特征替换“排名更低”的特征。重复这一过程，最终会将 $\mathcal{S}'$ 转变为 $\mathcal{S}$，同时始终保持充分性。因此，$\mathcal{S}$ 本身必定是相对于 $\langle f, \mathbf{x}\ \epsilon_p\rangle$ 的充分解释——这与其并非充分解释这一初始假设相矛盾。至此，证明完成。

□

根据引理 1，由于 $\mathcal{S}' \setminus \{\ell\}$ 中的特征具有最高的 $|\mathcal{S}'| = |\mathcal{S}| - 1$排序，因此大小为 $|\mathcal{S}'|$ 的任何子集 $\mathcal{S}'' \subseteq [n]$ 都不是充分解释。为了完成证明的其余部分，我们现在将用到另一个引理：

引理 2. 设存在某个 f、$\boldsymbol{x}$ 和 ϵ_p。那么，如果 $\mathcal{S} \subseteq [n]$ 不是关于$\langle f$、$\boldsymbol{x}$ $\epsilon_p\rangle$的充分解释，则任何 $\mathcal{S}' \subseteq \mathcal{S}$都不是关于 $\langle f$、$\boldsymbol{x}$ $\epsilon_p\rangle$的充分解释。

证明。 如果 $\mathcal{S} \subseteq [n]$ 不是关于 $\langle f$、$\mathbf{x}$ $\epsilon_p\rangle$的充分解释，那么：

$$\exists \mathbf{z} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}}; \mathbf{z}_{\bar{\mathcal{S}}}) \neq f(\mathbf{x}). \tag{19}$$

反设存在某个 $\mathcal{S}' \subseteq \mathcal{S}$ 是充分解释。换言之：

$$\forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) = f(\mathbf{x}). \tag{20}$$

然而，考虑向量$\mathbf{z}'$：将$\mathcal{S}'$中的特征固定为$\mathbf{x}$，将$\mathcal{S}\setminus\mathcal{S}'$中的特征也固定为$\mathbf{x}$，并根据$\mathbf{z}$设置所有其余坐标。根据此前由公式 20 得出的推论，我们必须有：$f(\mathbf{x}_{\mathcal{S}'};\mathbf{z}'_{\bar{\mathcal{S}}}) \neq f(\mathbf{x})$，这与$\mathcal{S}'$充分的假设矛盾（公式 19）。

□

我们现在继续完成该命题证明的剩余部分。由于根据证明的前一部分可知，不存在大小为$|\mathcal{S}'| = |\mathcal{S}| - 1$、关于$\langle f, \mathbf{x}\ \epsilon_p\rangle$的解释，我们现在可以利用引理 2 的结果得出：这些大小为$|\mathcal{S}'|$的子集的子集也都不是充分解释，这意味着不存在任何大小小于或等于$|\mathcal{S}'|-1$、且是$\langle f, \mathbf{x}\ \epsilon_p\rangle$的充分解释的解释，从而证明$\mathcal{S}$是基数最小充分解释，证明至此结束。

□

C.4 命题 4 的证明

命题 4. 给定 *NAM* f、输入$\boldsymbol{x}\in\mathbb{R}^n$以及扰动半径$\epsilon_p\in\mathbb{R}_+$，算法 *3* 执行$\mathcal{O}(\log(n))$次查询，并返回一个基数最小 充分解释。

^^^^^^证明。 我们将证明，给定特征排序$[(\Delta l_1,\Delta u_1),(\Delta l_2,\Delta u_2)\ldots,(\Delta l_n,\Delta u_n)]$，恰好存在一个索引$i\in[n]$，使得$\mathcal{S}=[i+1]$是充分解释，而$\mathcal{S}'=[i]$不是。此外，这一结论对任意排序都成立。我们的证明源自一个与引理 2 密切相关的引理，我们将在下文重述并证明该引理：

引理 3。 设有某个f、$\boldsymbol{x}$和ϵ_p。那么，如果$\mathcal{S}\in[n]$是关于$\langle f$、$\boldsymbol{x}\ \epsilon_p\rangle$的充分解释，则任何满足$\mathcal{S}\subseteq\mathcal{S}'$也是关于$\langle f$、$\boldsymbol{x}\ \epsilon_p\rangle$的充分解释的$\mathcal{S}'$都成立。

证明。 这可直接由引理 2 得出。由于已知$\mathcal{S}$是充分解释，反设存在某个$\mathcal{S}'\subseteq[n]$满足$\mathcal{S}\subseteq\mathcal{S}'$，使得$\mathcal{S}'$不是充分解释。根据引理 2，这意味着不存在任何子集$\mathcal{S}''\subseteq\mathcal{S}'$能成为充分解释——这与$\mathcal{S}\subseteq\mathcal{S}'$是充分解释这一事实相矛盾。

□

为结束该命题的证明，考虑按照特征的排序依次遍历这些特征。我们从空集$\emptyset$开始，并测试它是否是关于$\langle f, \mathbf{x}\ \epsilon_p\rangle$的充分解释。如果不是，我们就逐个添加特征：先添加$\{1\}$，然后是$\{1,2\}$，接着是$\{1,2,3\}$，依此类推。最终，我们会遇到某个特征$i\in[n]$，使得$[i]$关于$\langle f, \mathbf{x}\ \epsilon_p\rangle$是充分的。根据引理 3，任何满足$[i]\subseteq\mathcal{S}$的$\mathcal{S}$也都是充分的。因此，从不充分到充分的唯一转变发生在$[i-1]$和$[i]$之间，并且在变为充分之前，不可能在之后的某个索引$j>i$处，$[j]$又恢复为不充分。

既然我们已经证明了这一结论，那么算法 4 中的二分搜索——它在找到第一个满足$[i]$充分而$[i-1]$不充分的特征i时停止——将返回与迭代式“朴素”算法 3 相同的子集。算法 3 以贪心方式逐步遍历特征，并输出$[i]$。此外，命题 3 表明，算法 3 总能收敛到一个基数最小解释。因此，算法 4 也必定收敛到同一个基数最小解释，但只需进行$\mathcal{O}(\log(n))$次充分性检查，而非$\mathcal{O}(n)$次。至此，证明完成。

□

C.5 求解神经加性模型中基数最小充分解释的计算复杂性结果

在本小节中，我们研究求解神经加性模型的基数最小充分解释所涉及的计算复杂性。为清晰起见，我们首先对所讨论的计算问题进行正式定义并命名。

NAM 的最小充分解释（MSE-NAM）。
输入。 一个 NAM $f\colon\ \mathbb{R}^n\to\{0,1\}$，一个输入$\mathbf{x}\in\mathbb{R}^n$，以及$\mathbf{x}$周围的$\epsilon\in\mathbb{Q}$扰动。

输出。 使 $\mathcal{S} \subseteq [n]$ 成为关于 $\langle f, \mathbf{x}\ \epsilon\rangle$的充分解释的最小 $k \in \mathbb{N}$，满足 $|\mathcal{S}| = k$。

证明分为成员资格部分和困难性部分。我们首先证明成员资格。

论断 1. *MSE-NAM* 问题属于函数多项式时间$^{NP}[\mathcal{O}(N)]$,，其中 N 表示模型大小的编码。

证明。 我们通过展示该问题可先经由一次 OptP$[\mathcal{O}(N)]$ 计算求解，再进行多项式时间的后处理，从而证明其属于函数多项式时间$^{\mathrm{NP}}[\mathcal{O}(N)]$；两步结合便可得出该问题属于函数多项式时间$^{\mathrm{NP}}[\mathcal{O}(N)]$ (Krentel, 1986)。我们先给出证明的第一部分：

引理 *4.* 考虑一个 *ReLU* 神经网络 f: $\mathbb{R}^n \to \mathbb{R}^n$、一个输入 $x \in \mathbb{R}^n$和一个半径 $\epsilon \in \mathbb{Q}$。将 x 周围的 ϵ-球定义为 $B_\epsilon(x) := \{z \in \mathbb{R}^n: x_i - \epsilon \le z_i \le x_i + \epsilon\}$ 对所有 $i \in [n]$成立。那么，计算 $max_{z \in B_\epsilon}(x)(f(z)_1 + \cdots + f(z)_n)$ 的问题属于 *OptP*$[\mathcal{O}(N)]$。

证明。 我们通过将本问题表述为 Sälzer & Lange（2021年）所提出的神经网络可达性问题的一个特例来证明该引理。在其表述中，支持一大类输入规格 ψ_{in} 和输出规范 ψ_{out}。就我们的目的而言，我们将其框架限制为一个特别简单的实例：每个输入特征$\mathbf{x}_i$ 受到 ψ_{in}内的有理数上下界 $l_i, u_i \in \mathbb{Q}$ 约束，且每个输出特征$\mathbf{y}_i$ 受到 ψ_{out}内同类界的约束。也就是说，我们只考虑形如 $l_i \le \mathbf{x}_i \le u_i$ 和$l_i \le \mathbf{y}_i \le u_i$的规范，这些规范在 Sälzer & Lange（2021年）的表述中得到直接支持。现在我们着手正式定义神经网络可达性问题。

神经网络可达性（NNReach）(Sälzer & Lange, 2021)。

输入。 具有 ReLU 激活模型 f: $\mathbb{R}^n \to \mathbb{R}^n$、输入规范 $\psi_{\text{in}}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$以及输出规范 $\psi_{\text{out}}(\mathbf{y}_1, \ldots, \mathbf{y}_d)$。**输出。** 是，如果存在$\mathbf{x} \in \mathbb{R}^n$ 使得 $\psi_{\text{in}}(\mathbf{x})$与 $\psi_{\text{out}}(f(\mathbf{x}))$为真。

Sälzer & Lange（2021年）证明了 NNReach 问题属于 NP。由于无法在连续域中直接猜测输入赋值作为见证，他们的方法转而猜测每个神经元的激活模式（激活/未激活）。一旦固定这些激活状态，网络就会成为纯线性系统，而可达性问题也就转化为一个关于所有输入变量 $\mathbf{x}_1, \ldots, \mathbf{x}_n$、输出变量 $\mathbf{y}_1, \ldots, \mathbf{y}_d$以及对应于神经元运算的额外辅助变量的线性规划（LP）。随后可以在多项式时间内求解该 LP，从而证明其属于 NP。

现在，我们将在特定语境下描述 NNReach 问题。我们现在有一个模型 f: $\mathbb{R}^n \to \mathbb{R}^n$（即输出向量$\mathbf{y}$的大小为 n）。我们将输入规范 ψ_{in} 设为集合 $\mathbf{x}_i - \epsilon \le \mathbf{x}_i \le \mathbf{x}_i + \epsilon$，其中每个$\mathbf{x}_i$ 均满足该条件，并且对$\mathbf{y}$不设置输出规范。根据 (Sälzer & Lange, 2021) 的结果，我们知道该问题属于 NP。现在，我们可以定义映射函数 $G(X, Y)$（依据附录 B 中复杂度类 OptP 的定义），使得$Y \in \{0,1\}^{p(|X|)}$ 是 (Sälzer & Lange, 2021) 用于证明 NNReach 问题属于 NP 的见证赋值（每个 ReLU 的激活状态），$\{0,1\}^{p(|X|)}$ 表示线性规划中所有对激活/未激活神经元的 0-1 赋值构成的集合，而函数 G定义了与 (Sälzer & Lange, 2021) 中相同的线性规划，同时求解使$\mathbf{y}_1 + \mathbf{y}_2 + \ldots \mathbf{y}_n$ 的值最大（即所有输出值之和）这一任务。对于单个见证，结果将给出由该见证猜测的激活状态 Y所确定的定义域中某个特定的 线性区域内、以$\mathbf{x}$为中心的 ϵ-球区域中的最大值。然而，通过取

$$F(X) = \max_{Y \in \{0,1\}^{p(|X|)}} G(X, Y), \tag{21}$$

$F(X)$将给出前述所有最大值中的最大值，这些最大值是在由猜测神经元激活状态所确定的、以全部以$\mathbf{x}$ 为中心的 ϵ-球区域内取得的。综上可知，$F(X) = \max_{\mathbf{z} \in B_\epsilon(\mathbf{x})}(f(\mathbf{z})_1 + f(\mathbf{z})_2 + \ldots + f(\mathbf{z})_n)$，这表明该问题属于 OptP。现在，由于这个最优解是一个由模型参数定义的线性规划的解——我们知道它是有限的，且受程序参数精度的限制。假设每个单变量模型 f_i 中的每个权重和偏置均编码为有理数 $\frac{p_j}{q_j}$，满足q_j，$p_j \in \mathbb{N}$，并以二进制给出。令 p' 和 q' 分别表示总体模型 f中任意位置出现的权重/偏置的最大分子和分母，并定义 $d := \max(|p'|, |q'|)$。对最优线性规划解的大小，一个直接的上界是 $\mathcal{O}(d^m)$，其中 m是

f中的权重和偏置数量，它们可以用 $\mathcal{O}(m \cdot \log(d))$ 位编码，并且该数量关于 N尤其呈线性。这表明该问题尤其属于 OptP$[\mathcal{O}(N)]$。

□

在证明优化步骤属于 OptP$[\mathcal{O}(N)]$, 之后，我们现在通过描述所需的多项式后处理，完成原问题属于 FP$^{\text{NP}}[\mathcal{O}(N)]$ 的成员资格证明。首先，我们注意到，对任意 ReLU 激活神经网络f：$\mathbb{R}^n \to \mathbb{R}^n$（引理 4）所证明的结果，同样适用于形如 $\beta_o + f_1(\mathbf{x}_1) + \ldots + f_n(\mathbf{x}_n)$ 的网络；也就是说，这正是 NAM 模型的内在结构，其中每个 f_i 本身都是一个单变量 ReLU 网络。这是因为，按照相同的形式化定义，这样的和仍然是一个神经网络。在这种结构中，对完整表达式 $\beta_o + f_1(\mathbf{x}_1) + \ldots + f_n(\mathbf{x}_n)$ 求最大值或最小值，可归结为对每个单变量分量 $f_i(\mathbf{x}_i)$ 分别求最大值或最小值。对每个单变量分量执行最大化后，即可得到特征的全序——等价地，算法 2 中所有的界都“坍缩”为离散的单点——从而可以应用算法 3 或算法 4 来计算基数最小充分解释（如命题 3 所证），并返回其大小 k。由于整个过程由一次 OptP$[\mathcal{O}(N)]$ 计算和一个多项式时间后处理步骤组成，因此该问题整体属于 FP$^{\text{NP}}[\mathcal{O}(N)]$。至此，成员资格证明完成。

□

引理 5。 *MSE-NAM* 问题是 *OptP*$[\mathcal{O}(\log(N))]$-难的（在度量归约下）。

证明。 为证明其困难性，我们的主要归约将从最大可满足性（最大可满足性）问题出发；已知该问题是 OptP$[\mathcal{O}(\log(N))]$-完全的（Krentel，1986年）。在形式化介绍该问题之前，我们首先回顾合取范式可满足性（SAT）的相关背景概念：

考虑一组布尔变量 $\mathbf{X}_1$、.....、$\mathbf{X}_n \in \{0,1\}$。令 $m \in \mathbb{N}$ 表示子句的数量。对于每个 $1 \le i \le m$，我们将子句 C_i 定义如下：

$$C_i := \mathbf{X}_{i,1} \vee \mathbf{X}_{i,2} \vee \ldots \mathbf{X}_{i,k_i}, \tag{22}$$

其中，每个文字 i、j都是 $\mathbf{X}_{i,j} \in \{\mathbf{X}_1$、.....、$\mathbf{X}_n$、$\neg\mathbf{X}_1$、.....、$\neg\mathbf{X}_n\}$ 中的一个元素，而 $k_i \in \mathbb{N}$ 表示出现在子句 C_i中的文字数量。完整的合取范式公式 ϕ 定义为：

$$\phi := C_1 \wedge C_2 \wedge \ldots \wedge C_m. \tag{23}$$

例如，考虑以下合取范式公式 ϕ：

$$\begin{aligned}\phi := (\mathbf{X}_{1,1} \vee \ldots \vee \mathbf{X}_{1,k_1}) \wedge (\mathbf{X}_{2,1} \vee \ldots \vee \mathbf{X}_{2,k_2}) \wedge \ldots \wedge (\mathbf{X}_{m,k_1} \vee \ldots \vee \mathbf{X}_{m,k_m}) = \\ (\mathbf{X}_7 \vee \neg\mathbf{X}_2 \vee \mathbf{X}_3 \vee \mathbf{X}_5) \wedge (\mathbf{X}_2) \wedge (\mathbf{X}_3 \vee \neg\mathbf{X}_4).\end{aligned} \tag{24}$$

如果存在一个布尔赋值$\alpha \in \{0,1\}^n$，赋给变量 $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$，使子句 C_i 或整个公式 ϕ的值为真，我们就称其为可满足。例如，在上面的例子中，将 $\mathbf{X}_5 = 1$、$\mathbf{X}_2 = 1$和$\mathbf{X}_4 = 0$赋值后，三个子句 C_1、 C_2和 C_3都得到满足，因此公式 ϕ也得到满足。这引出了 SAT 问题的经典定义：

合取范式可满足性（SAT）。I

合取范式可满足性 ϕ。**输出**。是，如果 ϕ可满足；否

A Boolean formula

nput。

则为否。

在我们的设定中，我们还依赖这些问题的优化版本，具体而言是最大可满足性问题（Li 和 Manya，2009），该问题要求找到一个赋值，使满足的子句数量最多，等价地：

最大可满足性（MaxSAT）。

输入。 一个合取范式（CNF）布尔公式 ϕ。

输出。 可同时满足的子句的最大数量，即

$$\max_{\alpha \in \{0,1\}^n} \sum_{i=1}^{m} \mathbf{1}_{\{C_i \text{ is satisfiable by } \alpha\}}. \tag{25}$$

在给出证明之前，我们先形式化若干后续所需的辅助问题。首先回顾 Sälzer & Lange（2021）提出的神经网络可达性（NNReach）问题，

我们此前曾用它证明问题属于 NP。该研究表明，NNReach 是 NP 困难的。此外，他们证明，即使对于只有一个 输出的网络，以及非常简单的输入规范，这种困难性依然成立；具体而言，每个输入特征$\mathbf{x}_i$ 仅受界 $0 \leq \mathbf{x}_i \leq 1$ 约束（他们论文中的定理 3）。值得注意的是，(Sälzer & Lange, 2021) 中的所有归约都先证明二进制输入域 $\{0,1\}^n$上的困难性，再将构造扩展到连续域 $[0,1]^n$。因此，在这两种设定下，问题都是 NP 困难的。基于此，我们将他们研究中考虑的神经网络可达性问题的受限版本称为受限神经网络可达性（*R-NNReach*）；Sälzer & Lange（2021）已证明该问题是 NP 困难的。

受限神经网络可达性 (R-NNReach)。
输入。 一个神经网络 $f\colon \{0,1\}^n \to \mathbb{R}$，输入规范 $\psi_{\text{in}}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ 设为：$\psi_{\text{in}} := \bigwedge_{i=1}^{n} \mathbf{x}_i \geq 0 \wedge \mathbf{x}_i \leq 1$ 以及输出规范 $\psi_{\text{out}}(\mathbf{y})$，设为：$\psi_{\text{out}} := \mathbf{y} = r$ 对于某个 $r \in \mathbb{Q}$。**输出**。 是 如果存在一个 $\mathbf{x} \in \mathbb{R}^n$ 使得 $\psi_{\text{in}}(\mathbf{x})$ 且 $\psi_{\text{out}}(f(\mathbf{x}))$ 为真。

We w还将介绍该问题的一个额外变体，以支持我们的最大化
证明相关

最大化受限神经网络可达性（MR-NNReach）。
输入。一个神经网络 $f\colon \{0,1\}^n \to \mathbb{R}$，一个输入规范 $\psi_{\text{in}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \bigwedge_{i=1}^{n} 0 \leq \mathbf{x}_i \leq 1$

输出。使 $\psi_{\text{in}}(\mathbf{x})$ 为真的 $f(\mathbf{x})$ 的最大值。
我们将把证明的其余部分分为四个核心部分：

1. **引理 6** 说明，NNReach 问题，特别是其受限变体 R-NNReach（最初通过从合取范式-SAT 的归约得到（Sälzer & Lange, 2021；Katz 等人，2017）），可以扩展为支持从最大可满足性问题进行归约。具体而言，该引理证明，最大可满足性问题可以在多项式时间内归约到 MR-NNReach。 2. **引理 7**说明，通过加入一个额外的技术构件（类似于（Sälzer & Lange, 2021）中的构造，但需要进行非平凡的技术调整），我们可以将困难性结果扩展到连续域。具体而言，这表明在考虑函数 $f\colon [0,1]^n \to \mathbb{R}$ 时，归约仍然成立。 3. **引理 8**表明，ReLU 网络可以实现二进制分解过程，并且这一构造可以叠加在我们的模型之上。进而，这表明，即使对于 f_i 这一形式为 $f_i\colon \{0,1,\ldots,2^n-1\} \to \mathbb{R}$ 的单变量网络，从最大可满足性问题到 MR-NNReach 的归约仍然成立。**引理 9** 随后将这一归约进一步扩展到完全连续情形，即扩展到形式为 $f_i\colon \mathbb{R} \to \mathbb{R}$ 的函数。

4. **引理 10** 通过构造一个包含 m^2 个公式 ϕ 单变量副本的 NAM，将最大可满足性问题归约为 MSE-NAM；这些副本由引理 9 构造（其中m 是原始合取范式中的子句数）。这种构造使我们能够恢复精确的 MaxSAT 最优值，从而证明该问题是 OptP[$\mathcal{O}$(log N)]-困难的。

引理 6. 最大可满足性问题可以在多项式时间内归约为 *MR-NNReach*。

证明。 我们首先重新审视 Sälzer & Lange, 2021 年的 NP 困难性证明，该证明对 Katz 等人，2017 年的原始论证作出了技术性修正。这一修正处理了将归约从布尔域扩展到连续情形时出现的微妙问题——我们稍后会再讨论这一点。暂且忽略这一区别，这两种归约实际上完全相同。我们先介绍这一构造。核心思路是将一个含有 m 个子句的合取范式公式 ϕ归约为一个 ReLU 激活神经网络 $f\colon \{0,1\}^n \to \mathbb{R}$，具体形式化如下：

$$y := m - \sum_{i=1}^{m} \max(0, 1 - \sum_{j=0}^{k_i} f_i(\mathbf{x}_{ij})) \tag{26}$$

其中 $f_j(\mathbf{x}_{j,k}) = \mathbf{X}_{jk}$ 如果 $\mathbf{X}_{j,k}$ 在 C_j 中正向出现，而 $f_j(\mathbf{x}_{j,k}) = 1 - \mathbf{X}_{j,k}$ 如果它负向出现。随后可以直接构造相应的神经网络，因为线性变换和 $\max(0,\cdot)$ 运算都可以很容易地用 ReLU 单元表示。这

这一归约已在 (Katz 等人，2017) 和 (Sälzer & Lange, 2021) 中得到明确证明。如其中所示，我们可以立即得出结论：对于任意从 $\alpha \in \{0,1\}^n$ 到 ϕ的布尔赋值，我们有：

$$\max(0, 1 - \sum_{j=0}^{k_i} f_i(\mathbf{x}_{ij})) = 0 \iff C_i \text{ is satisfied by } \alpha. \tag{27}$$

由此可以直接证明，施加输出约束 $\psi_{\text{out}} := y = m$ 当且仅当 ϕ 中的每个子句都得到满足。进一步可以直接证明，当设置输出约束 ψ_{out} 为 $y = m$ 时，公式得到满足当且仅当 ϕ 中的所有子句都得到满足。此外，请注意，在此构造中，输出值 y 已经编码了满足的子句数量。因此，同一归约可以立即得到从 最大可满足性 到*MR-NNReach* 问题的多项式时间归约。

□

引理 7. 最大可满足性问题可以在多项式时间内归约到 *MR-NNReach*，其中 *MR-NNReach* 定义在形式为 $f\colon [0,1]^n \to \mathbb{R}$ 的神经网络上。

证明。 现在，我们将归约从二进制输入域 $\{0,1\}^n$ 转移到连续区域$[0,1]^n$。为此，我们采用一种与 Bool* 构件类似的技术，这些构件由 (Sälzer & Lange, 2021) 提出。首先，我们简要回顾 Bool* 构件在标准 CNF-SAT 归约中的作用，具体来说，它们如何使我们能够将困难性证明从离散布尔域迁移到连续域。然后，我们将说明如何将这一构造调整到我们的设定中。

在从 SAT 到 R-NN-Reach 问题的标准归约中，Sälzer & Lange, 2021 的作者提出了以下构造。

$$y := \sum_{i=1}^{n} (\max(0, \frac{1}{2} - \mathbf{x}_i) + \max(0, \mathbf{x}_i - \frac{1}{2})) - \sum_{i=1}^{m} \max(0, 1 - \sum_{j=0}^{k_i} f_i(\mathbf{x}_{ij})) + (m - \frac{n}{2}) \tag{28}$$

项 $+m-\frac{n}{2}$ 在他们的表述中被省略，但在此处将其纳入并不影响该归约的正确性，并且稍后将有助于简化我们的阐述。如 (Sälzer & Lange, 2021) 所示，表达式 $\mathbf{z}_i = \max(0, \frac{1}{2} - \mathbf{x}_i) + \max(0, \mathbf{x}_i - \frac{1}{2})$ 的求值结果为：

$$\mathbf{z}_i = \begin{cases} \frac{1}{2} - \mathbf{x}_i & \text{if } \mathbf{x}_i < \frac{1}{2}, \\ \mathbf{x}_i - \frac{1}{2} & \text{otherwise} \end{cases} \tag{29}$$

因此，我们有 $\max(0, \frac{1}{2} - \mathbf{x}_i) + \max(0, \mathbf{x}_i - \frac{1}{2}) = 0$ 当且仅当 $\mathbf{x}_i \in \{0,1\}$。因此，该构件迫使每个二元特征取值为 0 或 1。在实践中，(Sälzer & Lange, 2021) 通过施加输出约束 ψ_{out} 要求 $\mathbf{y} = m$，在其归约中实现了这一点。这是因为表达式：

$$-\sum_{i=1}^{m} \max(0, 1 - \sum_{j=0}^{k_i} f_i(\mathbf{x}_{ij})),$$

我们称之为“最大子句优化项”的这一项始终为非正。因此，为了强制满足 $y = m$，必须使以下项成立：

$$\sum_{i=1}^{n} (\max(0, \frac{1}{2} - \mathbf{x}_i) + \max(0, \mathbf{x}_i - \frac{1}{2})),$$

我们将其称为“二元值正则项”，其值始终恰好为 $\frac{n}{2}$。如前文所述，当且仅当每个$\mathbf{x}_i$ 在 $\{0,1\}$ 中取严格的二进制值时，才会出现这种情况。

现在，我们旨在进行类似的转换，从二元输入转向连续域，以便将最大可满足性 归约为我们的问题。对于最大可满足性，请注意不存在输出约束；目标只是最大化满足的子句数量。直观而言，我们希望最大化最大子句优化项，同时确保二元值正则项

保持恰好为 $\frac{n}{2}$；与之前一样，当且仅当所有变量$\mathbf{x}_i$的取值严格属于 $\{0,1\}$ 时，这一条件才成立。

然而，使用式（30）中描述的 (Sälzer & Lange, 2021) 构造时，同时最大化所有分量在某些情况下可能会优先增大最大子句优化项，而牺牲二元正则项。这可能导致输入变量偏离严格的二进制值。为避免这种情况，我们需要增大二元正则项的“权重”，使得最大化该项始终严格更有利。一种实现方法是最大化以下形式的项:

$$y := C \cdot \sum_{i=1}^{n} (\max(0, \frac{1}{2} - \mathbf{x}_i) + \max(0, \mathbf{x}_i - \frac{1}{2})) - \sum_{i=1}^{m} \max(0, 1 - \sum_{j=0}^{k_i} f_i(\mathbf{x}_{ij})) + (m - \frac{C \cdot n}{2}) \quad (30)$$

现在我们可以设定 $C := |\phi| + 1$。注意，只需调整相应的权重，下面的项仍可通过神经网络实现。采用这一 C 选择后，对于 $\mathbf{y}$，我们可以得到如下性质:

命题 2。 *当 $\boldsymbol{x}$ 取自 $[0,1]^n$ 时，$\boldsymbol{y}$ 的最大值等于 $\boldsymbol{y}$ 在 $\boldsymbol{x}$ 取自 $\{0,1\}^n$ 时的最大值。*

证明。 反设，为了导出矛盾，假设 y 的最大可达值是在某个输入 $\mathbf{x}$ 处取得的，且该输入至少有一个坐标满足 $\mathbf{x}_i \in (0,1)$。特别地，假设 $\mathbf{x}_i \geq \frac{1}{2}$。那么，根据公式 29，我们有:

$$\max(0, \frac{1}{2} - \mathbf{x}_i) + \max(0, \mathbf{x}_i - \frac{1}{2}) = \mathbf{x}_i - \frac{1}{2} \quad (31)$$

现在我们证明，将 $\mathbf{x}_i$ 的值提高到恰好为 1，会严格增大 $\mathbf{y}$的值。假设其当前值为 $\mathbf{x}_i = 1 - \epsilon$。将其提高到 1，会使二进制值正则化项增加$\epsilon \cdot C = \epsilon \cdot (|\phi| + 1)$。重要的是，无论这一变化如何影响最大子句优化项，即使导致 y减少，该减少也不会超过 $(|\phi| + 1) \cdot \epsilon$，因为此项包含的加数严格少于 $|\phi| + 1$ 个。因此，净效应是 y 增加，这意味着 y 的当前值并非最大值。

现在假设 $\mathbf{x}_i < \frac{1}{2}$。在这种情况下，公式 29 意味着:

$$\max(0, \frac{1}{2} - \mathbf{x}_i) + \max(0, \mathbf{x}_i - \frac{1}{2}) = \frac{1}{2} - \mathbf{x}_i \quad (32)$$

现在我们证明，将$\mathbf{x}_i$ 的值减小至恰好为 0，会使 y 的值增加。设当前值为$\mathbf{x}_i := \epsilon$。将其减小至 0，会使二进制值正则化项增加 $\epsilon \cdot C = \epsilon \cdot (|\phi| + 1)$。如前所述，尽管这一变化可能影响最大子句优化项，甚至可能使 y 减小，但其减小幅度至多为 $(|\phi| + 1) \cdot \epsilon$，因为该项包含的分量严格少于 $|\phi| + 1$。因此，y 的净值必定增加，这与其最大性矛盾。

□

引理 8. *设有一个输入区域：$D_n := [0, 2^n - 1] \cap \mathbb{Z}$。可以在多项式时间内构造一个多项式规模的 ReLU 神经网络 f: $D_n \to \{0,1\}^n$，使得 $f(x) = \boldsymbol{bin}(x)$，其中 $\boldsymbol{bin}(x)$ 是 x 的二进制分解向量。*

证明。 我们首先使用一个多项式规模的 ReLU 网络，将单个离散值 $x \in \{0, 1, \ldots, 2^n - 1\}$ 映射为其对应的二进制表示。此步骤执行经典的二进制分解过程，如算法 5 所述。

我们构造神经网络，直接实现算法中的数学运算。从提供给 f 的输入整数 x 开始，网络经过 n 次迭代，每次迭代由三个连续的层组成。在第 i 次迭代中，值 r^i、q^i 和 b^i 编码了算法所维护的相应中间值。每次迭代中的值 b^i 都通过残差连接传递到输出层向量 $\mathbf{o}$ 的第 i 个分量 $\mathbf{o}_i$。形式上，

算法 5二进制分解

输入： 整数 x，位数 n，其中 $0 \le x < 2^n$

$r \leftarrow x$
对每个 $i = 0$ 至 $n-1$ **执行**
  $q \leftarrow r/2^{n-1-i}$
  如果 $q \ge 1$**则**
    $\mathbf{b}_i \leftarrow 1$
    $r \leftarrow r - 2^{n-1-i}$
  else
    $\mathbf{b}_i \leftarrow 0$
  结束条件判断
结束每次循环
返回 $(\mathbf{b}_0, \ldots, \mathbf{b}_{n-1})$

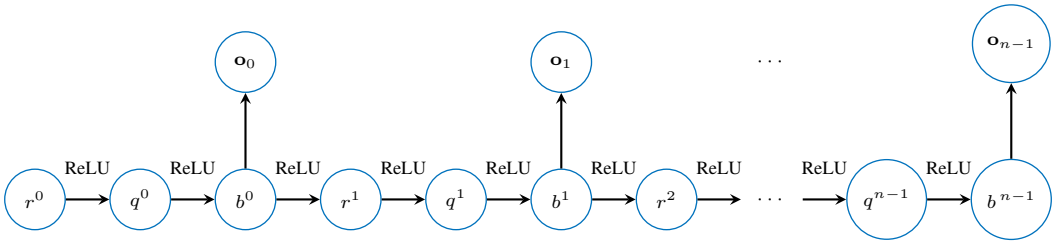

图 4：式 33 中定义的二进制分解的计算过程，输入为 $\mathbf{x}$，输出向量 $\mathbf{o}_0, \ldots, \mathbf{o}_{n-1}$ 表示该二进制分解。

构造定义如下：

$$\begin{aligned} q^i &= \max(0, \frac{r^i}{2^{n-1-i}} - 1), \\ b^i &= \max(0, 2^n \cdot q^i) - \max(0, (2^n \cdot q^i) - 1), \\ r^{i+1} &= r^i - 2^{n-1-i} \cdot \max(0, b^i), \\ \mathbf{o}_i &= b^i, \end{aligned} \tag{33}$$

其中 $r^0 = x$ 是提供给 f的输入。我们在图 4 中说明了这一构造。尽管该构造未经优化，它仍具有多项式规模，并且可以使用 ReLU 激活实现。特别地，所有所需运算——精确的 ReLU 约束、加法、乘法与减法——都可以通过标准的神经元与权重配置来实现。

根据归纳法，该神经网络精确模拟了该算法。在第 $i = 0$ 次迭代中，模型计算 $\mathbf{r}^0_{2^{n-1}} = \frac{x}{2^{n-1}}$。若 $x \ge 2^{n-1}$，则 q^i 会取得范围为 $[\frac{1}{2^n}, 1]$ 的正值（下界由假设 x 是 $\{0, 1, \ldots, 2^{n-1}\}$ 中的一个离散值可得）。若 $x < 2^{n-1}$，则将 q^i 设为 0。当 $q^i = 0$ 时，相应的位 b^i 也设为 0，并直接映射到输出 $\mathbf{o}_0$。当 q^i 为正时——由于离散性，它必然处于 $[\frac{1}{2^n}, 1]$ 内——将 b^i 设为 1。随后，按照算法的规定精确更新 r^1 的值，并对每次迭代递归地重复相同的过程。所有步骤都与标准二进制分解算法相对应，其正确性可由此直接得出。唯一的修改涉及 b^i 的计算；只要 q^i 的值始终大于 $\frac{1}{2^n}$，该计算就仍然有效。由于假设输入在范围 $\{0, \ldots, 2^{n-1}\}$ 内是离散的，因此这一条件始终成立，并确保所有值 b^i 和 r^i 在整个计算过程中都保持离散。因此，最终输出$\mathbf{o}_0, \ldots, \mathbf{o}_{n-1}$ 正确表示了 x 的二进制分解。

□

断言 3. 可以在多项式时间内构造一个 *ReLU* 神经网络 $f\colon [0,1] \to \mathbb{R}^n$，其具有多项式规模，使得 $\{0,1\}^n \subseteq f([0,1]) \subseteq [0,1]^n$，其中我们记 $f([0,1]) := \{f(x) \mid x \in [0,1]\}$。

证明。我们按如下方式构造神经网络 f。首先，对输入层 x 应用线性变换，将其乘以 $2^n - 1$。这会缩放输入，使得在第一个隐藏层中，我们得到一个取值范围恰为 $[0, 2^n - 1]$ 的单个神经元 $\mathbf{y}$。接着，将该神经元y 输入引理 8 中的构造。该引理证明，定义域 $\{0, 1, \ldots, 2^n - 1\}$ 中的每个离散值都会映射为相应的 n 位二进制表示。由于这些二进制表示涵盖整个空间 $\{0, 1\}^n$，因此，当输入限制为满足 $y \in \{0, 1, \ldots, 2^n - 1\}$ 的情况时，我们的网络 f 的像已包含所有二进制向量。此外，由于 y 的实际定义域是完整区间 $[0, 2^n - 1]$ ——它严格包含所有离散值 $\{0, 1, \ldots, 2^n - 1\}$ ——可知 $\{0, 1\}^n$ 仍属于整体模型 f 的像。

对于证明的第二部分，即证明 $f([0, 1]) \subseteq [0, 1]^n$，我们注意到，在引理 8 的构造中，每个输出神经元 $\mathbf{o}_i$ 都会接收 $b^i = \max(0, 2^n \cdot q^i) - \max(0, (2^n \cdot q^i) - 1)$ 的值。根据构造，对于任意实值 q^i，该表达式的取值始终位于区间 $[0, 1]$ 内。因此，每个输出坐标都属于 $[0, 1]$,，证明得证。

□

引理 9. 最大可满足性问题可在多项式时间内归约为 *MR-NNReach*，其中 *MR-NNReach* 定义在形式为 $f\colon \mathbb{R} \to \mathbb{R}$ 的神经网络上。

证明。在本证明中，我们展示，针对形式为 $f\colon [0, 1]^n \to \mathbb{R}$ 的神经网络建立的从最大可满足性问题到 MR-NNReach 的归约（引理 7），在目标模型形式为 $f\colon \mathbb{R} \to \mathbb{R}$ 时仍然成立。由于 MR-NNReach 问题包含输入约束 ψ_{in}，要求每个输入变量 $\mathbf{x}_i$ 满足 $0 \leq \mathbf{x}_i \leq 1$，因此只需考虑定义域为 $[0, 1]$ 的模型。利用断言 3，我们构造一个神经网络 $f\colon [0, 1] \to \mathbb{R}^n$，其像满足 $\{0, 1\}^n \subseteq f([0, 1]) \subseteq [0, 1]^n$。接下来，我们应用引理 7 中的构造，得到一个模型 $f'\colon [0, 1]^n \to \mathbb{R}$，其在 $[0, 1]^n$ 上的最大值恰好对应于最大可满足性问题实例的最优值。重要的是，该归约明确保证，$f'(\mathbf{x})$ 的任何最大化点都必须满足 $\mathbf{x} \in \{0, 1\}^n$，尽管定义域是 $[0, 1]^n$。现在我们定义 $g := f' \circ f$，并为清晰起见，将各分量记为 $g_f := f$ 和 $g_{f'} := f'$，使得 $g = g_{f'} \circ g_f$。引理 7 中引入的约束仍适用于内部分量 $g_{f'}$，确保其最大值仅在布尔输入上取得。此外，断言 3 保证 g_f 的值域包含所有布尔向量 $\{0, 1\}^n$，并排除 $[0, 1]^n$ 之外的任何点。因此，$g_{f'}$ 与 g_f 组合后可达到的最大值，与 f' 在 $[0, 1]^n$ 上可达到的最大值完全相同。因此，该归约对于形式为 $f\colon \mathbb{R} \to \mathbb{R}$ 的神经网络仍然成立，证明得证。

□

引理 10. 存在从最大可满足性到 *MSE-NAM* 的度量归约。

证明。给定合取范式公式 ϕ，我们首先应用引理 9 将其归约为单变量神经网络$f_i\colon \mathbb{R} \to \mathbb{R}$。我们任取一个值 $\mathbf{x}_i \in [0, 1]$。为简单起见，假设 $f_i(\mathbf{x}_i) = r$。我们可以通过附加一个权重为 1、偏置为 $-r$的额外最终层来“规范化”此值，从而得到修改后的模型 f'，使得 $f'_i(\mathbf{x}_i) = 0$（不过，为了方便记号，我们仍记作 f）。请注意，在此变换下，f 的最优 值也会平移 r。此外，我们注意到，通过附加一个权重为 -1的额外层，可以“镜像”该单变量分量，从而得到 $-f_i$。接着，我们构造一个由 m^2 个单变量分量 $-f_i$组成的 NAM f。现在，我们通过重复 $\mathbf{x}_i$ 值 m^2 次，构造向量 $\mathbf{x} \in \{0, 1\}^{m^2}$。我们令 $\epsilon = [0, 1]$，并将 NAM 的截距定义为 $\beta_0 := m^2$。至此，对应的 MSE-NAM 实例构造完成。现在我们观察到：

$$-\sum_{i=1}^{m^2} f_i(\mathbf{x}_i) + \beta_0 = 0 + m^2 > 0 \tag{34}$$

换句话说，这意味着 $f(\mathbf{x})$ 获得了正标签，即 $f(\mathbf{x}) = 1$。假设某个解释 $\mathcal{S}$ 是大小为 k 的充分解释。要使 $\mathcal{S}$ 确实符合充分解释的条件，它必须

满足以下条件：

$$-\sum_{i\in\mathcal{S}} f_i(\mathbf{x}_i) - \sum_{i\in\bar{\mathcal{S}}} \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) + \beta_0 \geq 0 \iff -\sum_{i\in\bar{\mathcal{S}}} \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) + \beta_0 \geq 0 \iff$$
$$-|\bar{\mathcal{S}}| \cdot \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) + m^2 \geq 0 \iff \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) \leq \frac{m^2}{|\bar{\mathcal{S}}|} = \frac{m^2}{m^2-k} \tag{35}$$

这意味着，当且仅当满足以下条件时，存在大小为 k 的充分解释：

$$(m^2-k)\cdot \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) \leq m^2 \tag{36}$$

这也意味着，当且仅当满足以下条件时，不存在大小为 k' 的充分解释：

$$(m^2-k')\cdot \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) > m^2 \iff m^2-k' > \frac{m^2}{\max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i)} \tag{37}$$

我们还注意到，$\max_{\mathbf{x}_i\in[0,1]} f_i(\mathbf{x}_i) \leq m-r \leq m$，因为最多只能满足 m 个子句。因此，可得：

$$\frac{1}{\max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i)} \geq \frac{1}{m} \tag{38}$$

此外，将公式 37 应用于任意一个不是大小为 k'的充分解释的子集，可得：

$$m^2-k' > \frac{m^2}{\max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i)} \iff m^2-k' > m \iff m^2-k' \geq m+1 \tag{39}$$

综上所述，我们已经证明，当且仅当 $m^2-k' \geq m+1$时，不存在大小为 k' 的充分解释。现在，假设解释 $\mathcal{S}$ 的大小为 k，且它是基数最小的。这意味着 $\mathcal{S}$ 满足公式 36，而且每个大小为 $k-1$ 的解释都不是充分解释。因此，当设定 $k' := k-1$ 时，公式 37 成立，由此得到 $m^2-(k-1) \geq m+1$，这等价于 $m^2-k \geq m$。因此，我们得到以下结论：

$$\frac{m^2}{m^2-(k-1)} < \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) \leq \frac{m^2}{m^2-k} \tag{40}$$

此外，以下关系可由简单代数运算直接得到：

$$\frac{m^2}{m^2-k} - \frac{m^2}{m^2-(k-1)} = m^2\cdot\frac{m^2-k+1-(m^2-k)}{(m^2-k)\cdot(m^2-k+1)} = \frac{m^2}{(m^2-k)\cdot(m^2-k+1)} \tag{41}$$

因此，我们得到：

$$\frac{m^2}{m^2-(k-1)} < \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) \leq \frac{m^2}{m^2-(k-1)} + \frac{m^2}{(m^2-k)\cdot(m^2-k+1)} \tag{42}$$

由于我们知道 $m^2-k \geq m$，因此可立即推出 $m^2-k+1 > m$。综上，我们得到：

$$\frac{m^2}{m^2-(k-1)} < \max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) < \frac{m^2}{m^2-(k-1)} + 1 \tag{43}$$

总之，在我们的构造下，$\max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i)$ 始终被限制在一个宽度严格小于 1 的区间内。根据引理 9，我们已经知道，$\max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i)$ 等于 ϕ 减去 r （归一化步骤中减去的量）后的 MaxSAT 值。由于 MaxSAT 最优值是整数，因此可直接得出，$\max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i) + r$ 也是整数。此外，区间 $[\frac{m^2}{m^2-(k-1)}+r, \frac{m^2}{m^2-(k-1)}+r+1]$ 必须恰好包含一个整数，因为其中包含 $\max_{\mathbf{z}_i\in[0,1]} f_i(\mathbf{z}_i)$，且其长度严格小于 1。因此，给定将 ϕ 编码到 NAM 实例和 MSE-NAM 问题中的结果，我们可以构造一个简单的多项式时间函数，根据 k 的值计算该区间内唯一的整数，从而恢复 ϕ 中可满足子句的最优数量。

□

D 关于重要性排序的理论与实践

D.1 参数 ξ_i 的理论最坏情况行为

我们回顾前面的命题 2，其中我们已证明，我们的算法执行的总步数等于：

$$\frac{\beta_i - \alpha_i}{2^{k_i}} \leq \xi_i \iff k_i \leq \mathcal{O}\big(\log(\frac{\beta_i - \alpha_i}{\xi_i})\big). \tag{44}$$

一个自然的问题是，当 ξ_i 趋近于 0，或区间 $\beta_i - \alpha_i$ 变得非常大时会发生什么。接下来我们将证明，在这种情况下，只要每个单变量模型中的权重$w_{i,\,j}$和偏置 $b_{i,\,j}$ 都以有理数 $\frac{p_{i,\,j}}{q_{i,\,j}}$ 表示，且 $q_{i,\,j},\ p_{i,\,j} \in \mathbb{N}$以二进制给出，我们的算法仍能保证在与模型表示大小成线性关系的查询次数内终止。

命题 5。 设 f 是一个具有 *ReLU* 激活的 *NAM*。那么，运行算法 *2*，随后运行算法 *3* 或算法 *4* 的过程，至多在与模型 f 的编码大小成线性关系的查询次数内，向神经网络验证器发出查询后终止。

证明。 我们分三个阶段证明该命题：(i) 首先，我们证明对于每个 $i \in [n]$，第 i个单变量函数的最小值（即 $min_{z \in B_\epsilon(\mathbf{x}_i)}\big(f_i(\mathbf{z}_i)\big)$ 其中$B_\epsilon(\mathbf{x}_i)$ 是以 $\mathbf{x}_i$ 为中心的 ϵ-球）的位长度有限，且不超过 f_i 的编码大小。(ii) 其次，我们证明验证界参数 α_i, β_i 的位长度也不超过 f_i 的编码大小。(iii) 最后，我们利用这一结论证明，算法 2 至多经过相对于模型编码大小呈线性数量的神经网络验证器查询后便会终止。由此可以直接得出，运行该算法后再运行算法 3 或算法 4，也必定会在至多线性数量的查询后终止。我们先证明结论的第一部分：

引理 11. 考虑一个单变量 *ReLU* 神经网络组件 f_i：$\mathbb{R} \to \mathbb{R}$、一个输入 $\boldsymbol{x}_i \in \mathbb{Q}$ 以及一个半径 $\epsilon \in \mathbb{Q}$。将围绕 $\boldsymbol{x}_i$ 的 ϵ-球定义为 $B_\epsilon(\boldsymbol{x}_i) = \{\boldsymbol{z}_i \in \mathbb{R} : \boldsymbol{x}_i - \epsilon \leq \boldsymbol{z}_i \leq \boldsymbol{x}_i + \epsilon\}$，对所有 $i \in [n]$ 成立。则 $\min_{\boldsymbol{z}_i \in B_\epsilon(\boldsymbol{x})}(f_i(\boldsymbol{z}_i))$ 是有限的，其位长度以 f_i 的编码大小为界。

证明。 我们针对某个一般的单变量模型 f_i 证明该论断。不过为便于表述，我们将定义 f_i 的一个副本“g”，即 g：$\mathbb{R} \to \mathbb{R}$，使得 $g := f_i$，以避免重复使用下标 i。我们将特别证明 $\min_{z^0 \in B_\epsilon(x)}(g(z^0))$ 是有限的，且其位长度以 g 的编码大小为界。我们在 z^0 中使用下标 0，以表示它是 ReLU 模型中的“0”输入层（即输入层）。

由于我们的 NAM 使用 ReLU 激活函数，最小值，即使 $\min_{z^0 \in B_\epsilon(x)}(g(z^0))$ 达到最小时的值，是某个线性规划的解，其约束由模型的权重和偏置直接决定。这一观察是神经网络验证中的标准基石。为完整起见，我们在此给出更明确的形式化表述。正如 Sälzer & Lange（2021）所示，一旦“猜测”出每个 ReLU 约束的激活模式（激活/非激活状态），验证查询就会化简为一个线性规划，其约束由模型参数定义。因此，最小值 $\min_{z^0 \in B_\epsilon(x)}(g(z^0))$ 就是由对激活模式的“猜测”所导出的线性规划问题的最小值。

更具体地说，假设我们对每个 ReLU 约束对应的所有激活状态都作出“猜测”：$\alpha_{j,k} \in \{0, 1\}$，其中该约束是第 k 层的第 j 个变量，那么可以定义以下线性规划：最小化 $g(z^0)$，满足输入规格：$x - \varepsilon \leq z^0 \leq x + \varepsilon$；由 $\mathbf{h}^k = W^k \mathbf{z}^{k-1} + \mathbf{b}^k$ 指定的线性规格，对所有 $k \in \kappa$ 均成立，其中 $\kappa \in \mathbb{N}$ 表示 g 中的层数；以及通过在 $\alpha_{j,k} = 1$ 时令 $\mathbf{z}_j^k = \mathbf{h}_{kj}$ 且 $\mathbf{h}_{kj} \geq 0$，或在 $\alpha_{j,k} = 0$ 时令 $\mathbf{z}_j^k = 0$ 且 $\mathbf{h}_{kj} \leq 0$ 给出的 ReLU 规格，其中 $\mathbf{z}^k$ 表示第 k 层的输出变量。在线性规划中被最小化的输出 $g(z^0)$，就是变量 $\mathbf{z}^\kappa$ 的值（即最后一个输出层的值）。

众所周知，线性规划的任何解都具有有限精度（Vera, 1998；Dadush 等，2020），对解的精度进行粗略估计，其上界为 $\mathcal{O}(n \cdot L)$，其中 n 为变量的数量，L 为任意输入系数的最大位长度。因此，同理，

在我们描述的线性规划中，若 m_i 表示 f_i 中的权重和偏置总数，并令 d 表示模型中任意参数（即权重/偏置）的最大位长度，也就是说，2^d本身表示这种编码的大小，那么最大值本身可以用 $\mathcal{O}(m_i \cdot d)$ 位进行编码，其大小与 f_i的编码大小呈线性关系。

由此立即可知，该线性规划的最优值——即在固定线性区域内取得的最小值——是有限的，且受该规划参数的数值精度所界定。此外，由至少一种激活模式 $\alpha_{j,k}$所确定的某个线性区域上的最小值，即 $\min_{z^0 \in B_\epsilon(x)}(g(z^0))$，同样适用相同的位长度界，因此 $\min_{z^0 \in B_\epsilon(x)}(g(z^0))$ 也满足这一界。 □

根据引理 11，ξ_i（如方程 2 中所定义）表示连续域上两个最小单变量值之间的最小间隔，该值也具有有限精度。具体而言，其位长度至多为 $\max_{i \in \{1,\dots,n\}} (m_i \cdot \log(d))$。现在我们可以继续证明的其余部分：

引理 12. 给定单变量模型 f_i的权重、偏置和输入区间的最大位长度均不超过 d_i，则输出区间的界 α_i 和 β_i 均为有限值，其位长度至多为 $m_i \cdot d_i$，其中 m_i 是 f_i中的参数数量。

证明。 参数 β_i 和 α_i 可以通过神经网络验证中开发的多种界传播技术获得。即使采用区间界传播（IBP）这类“朴素”方法（Zhang 等人，2020；Gowal 等人，2018）——该方法逐层传播界，在每一步计算一个外近似，并利用 ReLU 激活的单调性——所得界在最坏情况下的位长度仍与模型大小呈线性关系（尽管非常宽松）。

形式化地，对于每个 $k \in [\kappa]$，其中 κ 是单变量模型的层数，第 k层中传播所得区间的值，记为 α^k，β^k，可通过区间算术，利用第 k层的权重和偏置，由第 $(k-1)$ 层的区间 α^{k-1}，β^{k-1}计算得到：

$$\alpha^k = \min_{\alpha^{k-1} \le z \le \beta^{k-1}} \big(\mathrm{ReLU}(W^k z + b^k)\big), \qquad \beta^k = \max_{\alpha^{k-1} \le z \le \beta^{k-1}} \big(\mathrm{ReLU}(W^k z + b^k)\big),$$

其中 $\alpha^0 = x_i - \epsilon$ 和 $\beta^0 = x_i + \epsilon$。由于 $\mathrm{ReLU}(x) = \max(0, x)$，激活函数不会增加位长度。一次乘法最多可使位长度增加 d_i，一次加法最多可使其增加 1。因此，计算第 k 层中的 $W^k z + b^k$——该计算先对 α^{k-1}，β^{k-1} 与位长度最大为 d_i 的数逐元素相乘，再求和并加上偏置——最多会使位长度增加 d_i（由乘法导致），总计最多增加 m_i（由各层中的加法导致）。因此，α^k，β^k 的位长度最多比 α^{k-1}，β^{k-1}大 $d_i + \ell_k$，其中 ℓ_k 是第 k 层中的参数数量。

令 L表示 f_i中的层数。于是，我们可以给出输出区间值位长度的上界：

$$d_i \cdot L + \sum_{k \in [\kappa]} \ell_k \;\le\; d_i \cdot m_i + \sum_{k \in [\kappa]} \ell_k \;\le\; d_i \cdot m_i + m_i = m_i(d_i + 1).$$

总体而言，传播的输入界的位长度至多为 $\mathcal{O}(m_i \cdot d_i)$。 □

现在，利用引理 11 和引理 12，我们可以完成命题 5 的证明。具体而言，在算法 2 的循环每次迭代中，第 5 行将验证问题的界减半，从而使近似最小值的位长度减少 1。接着，第 12 行检查每一对单变量网络 f_i，$f_j(i,\ j \in [n])$，判断 f_i 的下界是否大于或等于 f_j 的上界，或反之。根据引理 11，f_i 和 f_j 的最小值的位长度界分别为 $m_i \cdot d$，$m_j \cdot d$ 位。此外，根据引理 12，α_i，β_i 的位长度上界为 $m_i \cdot (d+1)$。

因此，项 $\beta_i - \alpha_i$和 ξ_i均可使用至多 $\mathcal{O}(m \cdot d)$ 位来表示。因此，对该数量进行二分搜索至多需要 $\mathcal{O}(m \cdot d)$ 步。具体而言，在 $m \cdot d$ 中经过线性数量的步骤后，第 12 行的条件必须得到满足：对于每一对 $i,\ j \in [n]$，要么 f_i 的下界将与 f_j的上界相匹配，要么 f_j 的下界将与 f_i 的上界相匹配——包括真实最小值相等的情况。

换言之，我们执行 n 个并行线程，并且在最大单变量函数的编码大小上经过至多线性数量的迭代（该编码大小本身与总函数大小呈线性关系）后，命题 5 保证终止。此外，由于算法 3 和算法 4 分别在 n 上以线性时间或对数时间运行（n 同样与模型编码大小呈线性关系），因此整个过程至多经过模型编码大小上线性数量的查询后终止。

□

我们注意到，由于该界限关于模型编码大小呈线性关系，因此与通用神经网络中观察到的指数级最坏情况复杂度相比（Barceló 等人，2020b），实现了显著改进。

与验证器和机器精度的关系。 在实践中，几乎不可能达到这一精度的最坏情况界限（实证证据见 F.5 节）。假设情况下，即使是验证器或机器精度也会超过该参数。此外，我们的实验表明，训练后的 NAM 中的 ξ_i 值实际上很少恰好位于决策边界上。因此，在我们的实证评估中，决定运行时间的有效精度参数为 ξ_i。正如我们的实验所示——在这些实验中，我们高效地为常用 NAM 计算基数最小解释——以及 F.5 节中针对该参数的消融研究所示，这使总体运行时间得到了显著改善。

D.2 重要性排序的实用优化

在实践中分析算法 2 的行为时，我们注意到，在某些边缘情况下，该算法会进行过多的不必要验证器调用。这里简要介绍这些情况，以及我们基于启发式方法所做的（可靠）优化。与 4.1 节一样，我们考虑 $f(\mathbf{x}) = 1$ 的情形，这要求我们为每个特征 $i \in [n]$ 找到精确最小值 $f_i(\mathbf{x}_i^*)$ 的界 $[l_i, u_i]$。$f(\mathbf{x}) = 1$ 的情形与此对称。我们考虑的两种边缘情况涉及精确最小值等于（或接近）其中一个界的情形：

下界接近精确最小值。请注意，当我们通过神经网络验证调用迭代地细化界时，验证器有时会连同验证结果一起给出一个反例 $\mathbf{x}'$。我们可以将 $\min(f_i(\mathbf{x}'_i)\ m)$ 用作最小值 u_i 的新上界（算法 2，第 9 行）。我们观察到，在实践中，这能加快排序过程，尤其是因为验证器返回的反例 x' 往往接近甚至就是精确最小值。对于采用 ReLU 激活的 NAM，这种方法在 $[0, u_i]$ 的情况下实践效果良好，因为 $\mathbf{x}'_i$ 只需进入包含 $\mathbf{x}_i^*$ 的同一分段线性区域，该区域会通过 f_i 映射为 0。

接近精确最小值的上界。有时，上界 u_i 已经是（或接近）精确上界。因此，与其反复运行验证查询，直到 l_i收敛到 u_i，不如直接测试对于某个较小的 $\delta \in \mathbb{R}_+$，是否仍能达到 $u_i - \delta$。一种适合切换到此测试的启发式方法是：如果验证器得出结论，认为 $f(\mathbf{x}_i^*) \in [m_i, u_i]$，却无法返回能够证明这一点的反例（因为鉴于它非常接近 $f(\mathbf{x}_i^*)$，验证器必须或多或少恰好命中 u_i）。

这两种优化通常能将命题 2 中所述的验证器调用次数降至很少。实践中，我们发现，通常每个特征只需调用验证器 3 次，即可确定全序。

E 扩展至其他设置

E.1 扩展至多分类

对于多分类，我们假设获胜类别为 $t = f(\mathbf{x}) = \arg\max_{j \in [c]} f_j(\mathbf{x})$（第 2.3 节）。于是，我们可以区分两种情况：(i) 获胜类别与所有类别对比的解释，以及 (ii) 所有成对的获胜类别与单个类别对比的解释。两种情况都可以化简为一个二分类任务，其中 (i) 中的新二分类网络为 $f(\mathbf{x}) = f_t(\mathbf{x}) - \max_{j \neq t} f_j(\mathbf{x})$，而 (ii) 中的新二分类网络为 $\hat{f}_j(\mathbf{x}) = f_t(\mathbf{x}) - f_j(\mathbf{x})\ \ j \in [n]$，其中类别 1 对应于原始获胜类别 t。然后，我们可以应用算法 2 和算法 4 来生成相应的解释。

E.2 回归扩展

假设我们要找到子集 $\mathcal{S}$，使其足以确定预测结果始终会比原始输出大 某个偏差值 $\delta \in \mathbb{R}_+$（对于小于原始输出的预测，要找到同样保证的结果则正好相反）。算法的第一部分将与二分类中结果为正的情况完全相同（另一种用例则对应结果为负的情况）。

算法6 回归：并行区间重要性排序 输入：NAM f，输入 $\mathbf{x} \in \mathbb{R}^n$，扰动半径 $\epsilon_p \in \mathbb{R}_+$1：并行处理每个特征 $i \in [n]$ 2：为 $f_i(\tilde{\mathbf{x}}_i)$ 提取初始界 α_i, β_i，使得 $\tilde{\mathbf{x}}_i \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_i)$ 3：$l_i \leftarrow \alpha_i, u_i \leftarrow \beta_i$ 4：while True do 5：$m_i \leftarrow \frac{l_i+u_i}{2}$ 6：如果 verify($\forall \tilde{\mathbf{x}}_i \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_i)$ $f_i(\tilde{\mathbf{x}}_i) \geq m_i$)，则 7：$l_i \leftarrow m_i$ 8：否则 9：$u_i \leftarrow m_i$ 10：end if 11：$\Delta l_i \leftarrow f_i(\mathbf{x}_i) - l_i$; $\Delta u_i \leftarrow f_i(\mathbf{x}_i) - u_i$ 12：如果对所有 $i \neq j$ 均成立：$\Delta u_i \geq \Delta l_j$ 或 $\Delta u_j \geq \Delta l_i$，则 13：跳出 14：end if 15：end while 16：end for 17：按升序返回 arg sort([$(\Delta l_1, \Delta u_1) \ldots,$ $(\Delta l_n, \Delta u_n)$])

现在我们继续介绍回归情形下算法的第二部分：

算法7回归：贪心基数最小线性解释搜索

输入： NAM f，输入 $\mathbf{x} \in \mathbb{R}^n$，扰动半径 $\epsilon_p \in \mathbb{R}_+$，输出偏差 $\delta \in \mathbb{R}_+$

```
S ← [n]
对于每个 特征 i ∈ [n]，按算法6排序执行                ▷ suff(f, x S, δ, ε_p) 成立
    如果suff(f, x S \ {i}, δ, ε_p) 则
        S ← S \ {i}
    结束如果
结束循环
返回 S                                  ▷ S 是一个关于 ⟨f, x δ, ε_p⟩的基数最小解释
```

这与我们用于二分类的算法相同（算法 3），但现在对*suff* 的评估变为检查固定特征子集 $\mathcal{S}$ 是否能确保输出比原始输入高出的幅度小于 δ。基于该结果，这同样可以扩展到我们的对数版本（算法 4）：

F 实验详情与消融研究

F.1 数据集与实验详情

数据集。我们在四个广泛使用的基准数据集上评估了我们的方法（Agarwal 等，2021；Radenovic 等人，2022），涵盖分类和回归任务。乳腺癌数据集包含 569 个样本，每个样本有 30 个数值特征，用于将肿瘤分类为恶性或良性。CREDIT 数据集包含 1,000 个样本，每个样本有 20 个属性，用于评估贷款偿还概率。FICO HELOC 数据集包含 10,459 个样本，每个样本有 23 个金融和人口统计特征，用于预测信用worthiness。这些数据集使我们能够综合评估方法在不同问题类型、输入

算法 8回归：贪心基数最小对数解释搜索

输入： NAM f，输入 $\mathbf{x} \in \mathbb{R}^n$，扰动半径 $\epsilon_p \in \mathbb{R}_+$，输出偏差 $\delta \in \mathbb{R}_+$

$F \leftarrow$ 特征的全序（算法 6）
$l \leftarrow 1$； $u \leftarrow n$
当 $l \neq u$**时执行**
  $m \leftarrow \lfloor \frac{l+u}{2} \rfloor$
  如果充分$(f, \mathbf{x}\ \{F[1], \ldots, F[m]\}, \delta, \epsilon_p)$ **则**
    $l \leftarrow m$
  else
    $u \leftarrow m - 1$
  结束如果
结束当
$\mathcal{S} \leftarrow \{F[1], \ldots F[m]\}$
返回 $\mathcal{S}$ ▷ $\mathcal{S}$ 是一个关于 $\langle f, \mathbf{x}\ \delta, \epsilon_p \rangle$的基数最小解释

表 3：在所有数据集上改变扰动半径 ϵ。

	乳腺癌		CREDIT		FICO HELOC	
ϵ	大小 (↓)	时间 [s] (↓)	大小（↓）	时间 [s]（↓）	大小（↓）	时间 [s]（↓）
0.01	4.00±4.24	35.60±1.34	—	—	—	—
0.1	4.45±3.98	115.08±73.48	1.79±1.07	76.76±88.21	2.77±1.62	131.31±141.90
0.2	6.29±3.50	143.81±121.27	2.80±2.07	97.16±34.45	3.88±1.58	136.82±77.97
0.5	9.76±2.75	146.01±64.76	3.76±2.62	132.67±36.76	5.59±1.80	317.92±222.07

维度和领域特征上的性能与稳健性。**模型**。 我们在前三个数据集上训练了 3 个二元分类 NAM，在最后一个数据集上训练了一个回归 NAM。我们遵循（Agarwal 等，2021；Radenovic 等人，2022）中的标准架构，并为每个特征训练一个具有大小为（64,64, 32）的隐藏层的网络。乳腺癌、CREDIT 和 FICO HELOC 模型的准确率分别为 97.37%、94.92% 和 69.02%。

评估。 所有结果均为 50 个样本的平均值，超时时间为 600 秒；除非另有说明，扰动是相对于归一化输入的 ℓ_∞-范数而言，且使用扰动半径 $\epsilon = 0.5$。我们过滤掉了平凡样本，例如所有特征都被作为解释返回的情况。对于乳腺癌数据集，我们使用 $\epsilon = 0.01$，以便与局部策略进行有意义的比较。实验在配备 64GB 内存和第 13 代 Intel(R) Core(TM) i7-1365U 的 Ubuntu 24.04 计算机上运行。CREDIT 实验在一台配备 Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz 和两块 40GB NVIDIA A100-PCIE 的 Ubuntu 24.04 计算机上运行。除非另有说明，所有实验均限制使用 32 个 CPU 线程。图中展示了中位数，并以阴影区域表示 25/75% 分位数。

F.2 消融扰动半径

US

在形式化可解释人工智能中，生成的解释在很大程度上取决于所选的扰动半径 ϵ。在表 3 中，我们展示了解释的大小和生成时间如何随 ϵ 变化。通常，解释大小和生成时间会随着 ϵ增加，这符合预期，因为先前获得的最小解释确实会变得不足，而且随着 ϵ 增加，验证查询也会变得更难求解。请注意，对于 $\epsilon = 0.01$，在 CREDIT 和 FICO 上解释并非平凡的情况下，找到的样本数量不足以计算可靠的平均值。

F.3 随时间推移已处理特征的数量

在本实验中，我们展示了我们的方法在初始排序阶段之后如何更快地处理特征，从而获得基数最小解释，因为它只需对数数量的验证查询即可做到这一点。这甚至优于那些获得子集最小方法的方法，其中包括对特征进行排序所需的时间（图 5）。

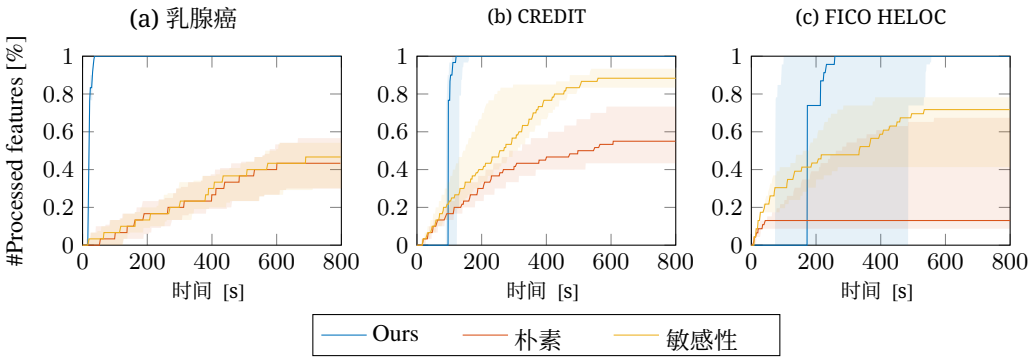

图 5：随时间变化的已处理特征数量。

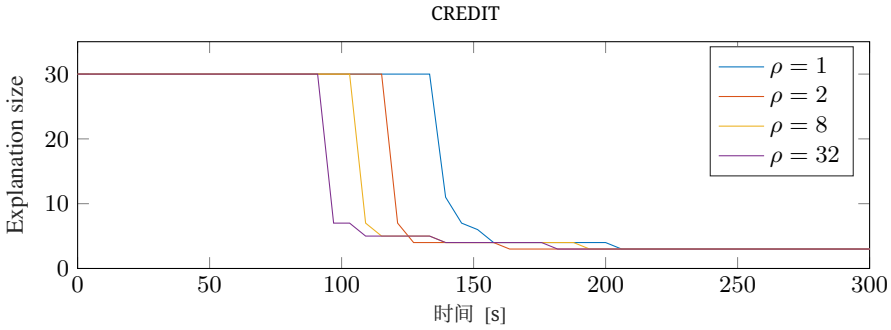

图6：使用不同数量的处理器生成解释的对比 ρ。

F.4 理解并行化：消融处理器数量

影响我们方法运行时间的一个关键因素——尤其是算法 2 中对单变量组件重要性的排序——是用于并行化对数二分搜索的处理器数量。为了评估这一因素的影响，我们通过改变处理器数量进行了消融研究。

图3展示了并行化对解释大小和计算时间的影响。具体而言，随着处理器数量 ρ 增加，按照重要性对特征进行排序（算法 2）所需的时间可以缩短，因为界限细化可以并行化。相比之下，后续的解释生成（算法 4）几乎不受处理器数量 ρ 的影响。重要的是，即使只有单个处理器（即不进行并行化），我们的算法仍能计算出基数最小解释——这比子集最小解释更难求得——因此从设计上改善了解释大小和计算时间（参见图3）。

F.5 近同特征的消融研究

我们的方法依赖于对特征进行高效的重要性排序（4.1 节）。本小节进一步分析了有效排序所需的精度（命题 2）。

我们首先在图7中展示数据集中每种所需精度的出现频率。图8可视化了这些精度对排序时间的影响。这些结果表明，所需精度并非任意地小（最小约为 10^{-8}），并且在实践中，排序时间还受到验证器等其他因素的影响。这意味着，在实践中，基于参数 ξ_i 进行排序非常有效，这与我们在为神经加性模型计算基数最小解释时观察到的强劲实证表现一致，并有助于解释这一表现。

为了分离该效应，我们使用一个合成神经网络开展了消融研究，该网络具有几乎相同的特征网络 f_i、$i \in [2]$，其中仅将 f_2 最后一层中的偏置略微移位。因此，较大的偏差移位对应于计算得到的界限之间没有重叠，而较小的偏差移位

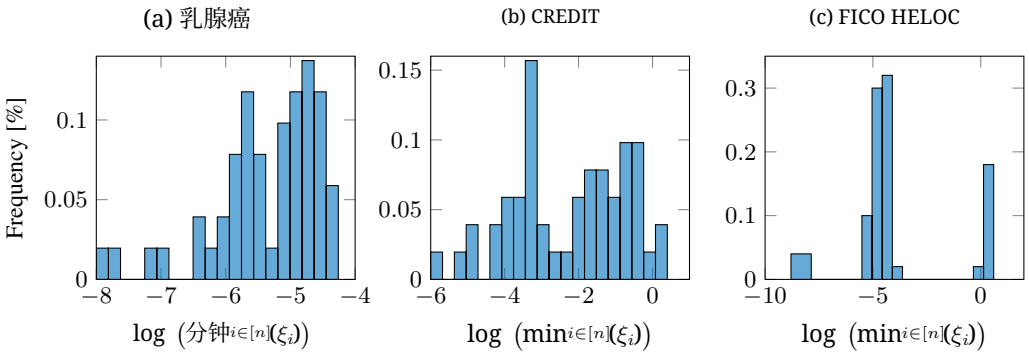

图 7：根据特征重要性对特征进行排序所需精度的频率。

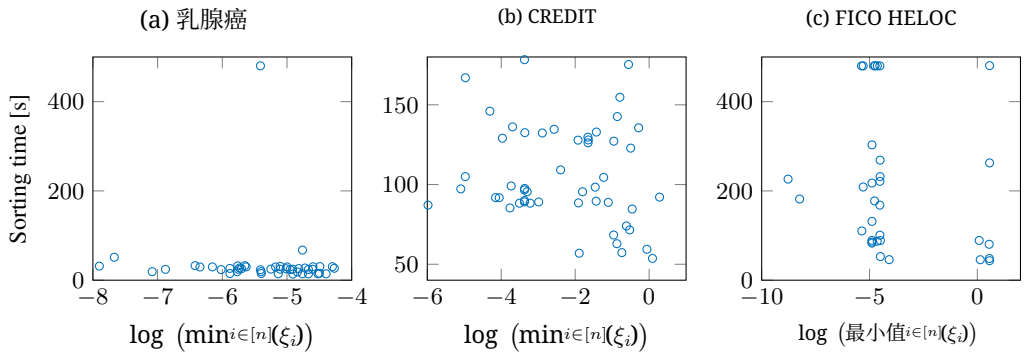

图 8：所需精度对排序时间的影响。

则会产生几乎相同的界。我们使用与在真实数据集上训练的网络相同的网络架构，并且在本研究中不应用 D.2 节中讨论的任何优化。我们在表 4 中记录了不同偏差值的排序时间，偏差值从 1（即使对于初始界也没有重叠）降至 10^{-7}（几乎完全重叠）。我们观察到，由于二分搜索，排序时间仅随偏差移位的数量级（对应于 ξ_i）线性增加，从而证实了我们在命题 2 中的结果。

G 披露：使用大型语言模型（LLM）

大型语言模型（LLM）仅作为写作辅助工具，用于润色语言和提升表达效果。它未参与研究构思、研究设计、分析或结果解读。这些工作均由作者独立完成。

表 4：近乎相同的特征对合成示例中排序时间的影响。

偏差移位	排序时间 [秒]
1	0.08
0.1	7.58
0.01	12.71
0.001	16.24
0.0001	20.95
0.00001	28.37
0.000001	31.54
0.0000001	36.57