WHEN ACTIVATION PATCHING ADMITS MANY FAITHFUL CIRCUITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation patching can establish faithfulness without determining a unique minimum circuit. In every additive n-component model with minimum r-faithful cardinality $1 \le k < n$, complete order-r evidence admits at most $\binom{n-1}{k}$ minimum circuits. A patched ReLU model with m patching sites attains the bound whenever $0 \le r < m$ and $n-1 \le \binom{m}{r+1}$, giving an exact extremal characterization. For this construction, raising the patch-order bound to $r+1$ makes the full mask uniquely faithful and gives every proper mask faithfulness error at least one. At $k = \lceil n/2 \rceil$, the exact maximum $\binom{n-1}{\lceil n/2 \rceil}$ is $2^{\,n-\frac{1}{2}\log_2 n+O(1)}$. For fixed r, exponential multiplicity persists with per-component fan-in and parameter count bounded independently of n and with uniform response norms through order $r+1$. The extremal family and transition remain stable under component perturbations and bounded departures from additive masking. Exact noiseless deterministic identification of an unknown extremal circuit requires $n-2$ queries of order at most $r+1$ in the worst case for $1 \le k \le n-2$.

1 INTRODUCTION

Circuit-based mechanistic interpretability studies subnetworks or component sets that implement a selected model behavior. Activation patching, path patching, causal tracing, and automated circuit discovery intervene on internal activations and compare the resulting output scores (Wang et al., 2023; Conmy et al., 2023; Bhaskar et al., 2024; Marks et al., 2025; Zhang & Nanda, 2024). Circuit analyses then use faithfulness, completeness, sparsity, inclusion minimality, or minimum cardinality to validate a proposed subnetwork (Wang et al., 2023; Hanna et al., 2024; Hadad et al., 2026). These properties certify a candidate relative to a chosen intervention domain. A structural uniqueness claim requires an additional premise: the admitted interventions must separate that candidate from every other minimum faithful circuit.

Consider a clean–corrupted activation pair with m patching sites and let $r \in \{0, \dots, m-1\}$ be the order bound. A patch is represented by $z \in [0,1]^m$; the nonzero coordinates select clean activations or interpolations toward them. We call $\|z\|_0$ the patch order (intervention size), and complete order-r evidence means the response on every patch with $\|z\|_0 \le r$. Suppose an internal branch adds a constant response and a ReLU term that is zero on every patch of order at most r, yet positive on a chosen order-$(r+1)$ patch. Several branches can then have identical complete order-r signatures and distinct order-$(r+1)$ signatures. Structured zero-sum cancellation among branch responses can make combinatorially many fixed-cardinality branch subsets match the full model on the complete low-order domain. Complete order-r evidence therefore does not distinguish among these circuits, even under exact search and verification on every admitted patch.

We ask how many minimum faithful circuits complete order-r evidence admits and when order $r+1$ excludes every proper candidate.

Our first result gives an exact extremal characterization for additive component models, which include prescribed output-edge groups at a common linear readout and parallel branches with mask-independent additive contributions. Let n be the number of maskable components, let k be the minimum faithful cardinality, and we write $\mathsf{MinFaith}_r$ for the family of minimum order-r faithful

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

circuits. We prove, for every $1 \le k < n$, the sharp bound

$$|\mathsf{MinFaith}_r| \le \binom{n-1}{k}.$$

The case with no omitted nonzero signature is immediate. When the omitted set has positive size at most half of the nonzero signatures, the bound follows by reducing minimum omissions to an intersecting family. When the maximum omission size exceeds half the number of nonzero signatures, a zero-pattern-preserving rational projection reduces the signatures to the large-length zero-sum theorem of Móricz & Nagy (2026, Theorem 1.7 and Lemmas 2.1–2.2). We attain equality in a patched feedforward ReLU model whenever $0 \le r < m$, $1 \le k < n$, and $n - 1 \le \binom{m}{r+1}$. Equality holds on $\{z \in [0,1]^m \mid \|z\|_0 \le r\}$, an uncountable domain for $r \ge 1$ and the singleton $\{0\}$ for $r = 0$. Complete order-r evidence can therefore admit the exact maximum number of minimum faithful circuits, while increasing the patch-order bound by one makes the full mask uniquely faithful; every proper mask then has faithfulness error at least one. At $k = \lceil n/2 \rceil$, the exact low-order maximum satisfies

$$\log_2 |\mathsf{MinFaith}_r| = n - \tfrac{1}{2}\log_2 n + O(1).$$

The extremal construction concentrates cancellation in one component. A distributed construction shows that exponential multiplicity persists without a distinguished component whose parameter count grows with n. With $n = 3q$ and minimum cardinality $k = q$, it has exactly $\binom{2q}{q}$ minimum circuits, all faithful on the complete continuous order-r domain and pairwise distinguishable at order $r + 1$. For fixed r, every component depends on at most two order-$(r + 1)$ ReLU units and has parameter count and response norm bounded independently of q.

The extremal result also persists for masked responses uniformly close to an additive surrogate: faithfulness errors and pairwise separations change by at most twice the additive-surrogate residual radius, preserving the minimum-circuit family and next-order transition in an open sup-norm neighborhood.

For the extremal family, exact noiseless queries from $\mathcal{Z}_{r+1}$ require $n - 2$ deterministic queries in the worst case for $1 \le k \le n - 2$.

2 Related work

Circuit discovery seeks a sparse subgraph that preserves a selected behavior (Conmy et al., 2023; Hanna et al., 2024; Marks et al., 2025). Formal accounts classify circuit queries (Adolfi et al., 2025), certify agreement on continuous input and patching domains, and verify several minimality notions (Hadad et al., 2026). Dataset-level certification stabilizes component selection under bounded edits (Anani et al., 2026), while sequential confidence sequences certify interventional fidelity under adaptive evaluation (Asiaee, 2026). These results provide behavioral or statistical guarantees for a selected candidate. Nikankin et al. (2025) use integer linear programming for circuit edge selection. In contrast, Proposition 20 provides an exact finite-domain uniqueness test and an enumeration procedure for all minimum faithful circuits.

Circuit non-uniqueness is now empirically well established. Small-network enumeration exhibits multiple circuits and causal alignments (Méloux et al., 2025); variance analyses document instability under resampling, rephrasing, sample-wise variation, and other methodological choices (Méloux et al., 2026; Wu et al., 2026); and recent LLM studies recover multiple sparse, low-overlap circuits or dataset-specific mechanisms (Chen et al., 2026; Rai et al., 2026). Circuit claims also vary with the extracted object, pruning threshold, and granularity (Sheng & Fu, 2026; Mahale, 2026). Bayat Makou et al. (2026) find structurally distinct circuits implementing the same computation; under additional linearization and margin assumptions, Chen et al. (2026) derive two faithful low-overlap edge sets. Prior empirical studies therefore document circuit multiplicity and instability, while formal methods provide candidate-specific certification; we determine the exact maximum number of minimum faithful circuits consistent with complete bounded-order evidence and construct patched ReLU models attaining it.

Sun & Toneva (2026), Sahi et al. (2026), and Dhor & Chen (2026) respectively study cross-network interpretive equivalence, circuit-level mechanistic uniqueness and identifiability, and spectral identifiability. Our sharp count concerns minimum faithful masks in a fixed component vocabulary.

Multi-component patching estimands decompose into mediator interactions (Vaidyanathan et al., 2026), whereas we enumerate minimum faithful masks. Related work separates redundancy and synergy in scalar interventional scores (Aghilar et al., 2026). We characterize equality classes of complete binary order-r responses and count the binary component masks in the set of minimum faithful masks. Möbius inversion and interaction indices are classical (Rota, 1964; Harsanyi, 1982; Grabisch et al., 2000; Sundararajan et al., 2020); sparse Möbius recovery uses additional assumptions (Erginbas et al., 2026). We use these tools to characterize bounded-order evidence, while the new structural content is the reduction to intersecting zero-sum omission families, the sharp neural realization, the distributed construction with bounded per-component fan-in, and the approximate-additivity bounds. The universal upper bound combines the Erdős–Ko–Rado theorem (Erdős et al., 1961; Katona, 1972) with the sharp large-length zero-sum count of Móricz & Nagy (2026); Huang (2026) subsequently resolved the remaining fixed-cardinality cases of the Möricz–Nagy zero-sum extremal problem. Our contribution is the reduction to maximum zero-sum omission sets, the extension to signatures in real vector spaces, including bounded function spaces, and a patched neural model attaining the resulting bound for every $1 \le k < n$ whenever $n - 1 \le \binom{m}{r+1}$.

3 Intervention-response functions

We first state the continuous patching domain and then introduce the Möbius coefficients of the binary restriction.

Let $m \in \mathbb{N}$ be the number of labeled patching sites, and let $[m] = \{1, \dots, m\}$. At site i, let u_i^0 and u_i^1 be the corrupted and clean activations. A gate value $z_i \in [0, 1]$ selects the interpolation $u_i(z_i) = (1 - z_i)u_i^0 + z_i u_i^1$. For $z = (z_1, \dots, z_m) \in [0, 1]^m$, we define the patch order as $\|z\|_0$, the number of nonzero coordinates. For $r \in \{0, \dots, m\}$, the complete continuous order-r domain is

$$\mathcal{Z}_r = \{z \in [0, 1]^m \mid \|z\|_0 \le r\}.$$

A scalar response is a function $F : [0, 1]^m \to \mathbb{R}$. For multiple outputs, inputs, or clean–corrupted pairs, exact statements apply coordinatewise; approximate statements require a chosen product norm and the corresponding uniform residual bound.

For the binary restriction, each $S \subseteq [m]$ denotes the gate vector $\mathbf{1}_S \in \{0, 1\}^m$. We write $F(S) = F(\mathbf{1}_S)$ and identify the restriction with a function $F : 2^{[m]} \to \mathbb{R}$. Möbius inversion gives the Möbius coefficients.

Definition 1. *For $T \subseteq [m]$, we let*

$$a_T(F) = \sum_{U \subseteq T} (-1)^{|T|-|U|} F(U).$$

The Möbius degree (interaction degree) is $\deg(F) = \max\{|T| \mid a_T(F) \neq 0\}$*, and the interaction support is* $\operatorname{supp}(F) = \{T \mid a_T(F) \neq 0\}$*, with* $\deg(0) = -\infty$.

Möbius inversion gives, for every $S \subseteq [m]$, $F(S) = \sum_{T \subseteq S} a_T(F)$. Equivalently, the unique multilinear extension of F to $\mathbb{R}^m$ is

$$p_F(z) = \sum_{T \subseteq [m]} a_T(F) \prod_{i \in T} z_i.$$

For $d \in \{0, \dots, m\}$, let $\mathcal{F}_{m,d}$ denote the vector space of degree-at-most-d pseudo-Boolean functions. The dimension of $\mathcal{F}_{m,d}$ is $N_{m,d} = \sum_{j=0}^{d} \binom{m}{j}$.

Definition 2. *For $r \in \{0, \dots, m\}$, let $\mathcal{S}_r = \{S \subseteq [m] \mid |S| \le r\}$. The order-$r$ evidence map is $\mathcal{E}_r(F) = (F(S))_{S \in \mathcal{S}_r}$. We write $F \equiv_r G$ when $\mathcal{E}_r(F) = \mathcal{E}_r(G)$. The order-$r$ truncation is*

$$\tau_r F(S) = \sum_{T \subseteq S,\, |T| \le r} a_T(F).$$

Let $\mathcal{Y}$ be any codomain. An explanation map $\rho : \mathcal{F}_{m,m} \to \mathcal{Y}$ is r-identifiable if $F \equiv_r G$ implies $\rho(F) = \rho(G)$. The definition concerns what the evidence determines, independent of a particular recovery algorithm.

Neural realizability. The pseudo-Boolean response class is realizable by one-hidden-layer ReLU networks on the Boolean cube.

Proposition 3. *Every $F : 2^{[m]} \to \mathbb{R}$ is realized on $\{0,1\}^m$ by a one-hidden-layer ReLU network with one hidden unit for each nonempty $T \in \operatorname{supp}(F)$ and an output bias. If $\deg(F) \le d$, every hidden unit has fan-in at most d.*

Proof sketch. On binary inputs, $\prod_{i\in T} z_i = \mathrm{ReLU}\big(\sum_{i\in T} z_i - |T| + 1\big)$. Substituting these units into the Möbius expansion realizes F, and the degree bound limits every fan-in to d. Thus, the binary-response lower bounds below already hold for shallow piecewise-linear networks. □

4 BOUNDED-ORDER IDENTIFIABILITY

The first theorem characterizes the binary restriction only; Theorem 10 separately proves continuous-domain equality for the ReLU construction.

Theorem 4. *Let $F, G : 2^{[m]} \to \mathbb{R}$ and $d, r \in \{0, \dots, m\}$.*

(i) $F \equiv_r G$ iff $a_T(F) = a_T(G)$ for every $T \subseteq [m]$ with $|T| \le r$.

(ii) On $\mathcal{F}_{m,d}$, the map $\mathcal{E}_r$ has rank $\operatorname{rank}(\mathcal{E}_r|_{\mathcal{F}_{m,d}}) = N_{m,\min\{r,d\}}$, and kernel dimension $N_{m,d} - N_{m,\min\{r,d\}}$.

(iii) τ_r is a complete invariant for $\equiv_r$: an explanation map ρ is r-identifiable iff there is a unique map $\bar\rho$ on $\tau_r(\mathcal{F}_{m,m})$ such that $\rho = \bar\rho \circ \tau_r$.

Proof sketch. For $|T| \le r$, the Möbius formula computes $a_T(F)$ from values $F(U)$ with $U \subseteq T$; conversely, $F(S) = \sum_{T\subseteq S} a_T(F)$ computes every observed value with $|S| \le r$. The two implications prove (i), and counting the observable coordinates proves (ii). Finally, $F \equiv_r G \iff \tau_r F = \tau_r G$, which proves the factorization in (iii). Complete binary order-r responses therefore determine exactly $\tau_r F$, independently of the recovery procedure. □

Corollary 5. *The order-r evidence map is injective on $\mathcal{F}_{m,d}$ iff $r \ge d$. Thus, a Boolean response function of Möbius degree at most d is exactly identifiable from all interventions of order at most r precisely when $r \ge d$.*

Corollary 6. *For every $0 \le r < m$, $\equiv_{r+1} \subsetneq \equiv_r$. The separation is witnessed by $F_0 = 0$ and $F_1(S) = \mathbf{1}[\{1, \dots, r+1\} \subseteq S]$, which are realized by zero and one ReLU hidden units, respectively.*

The hierarchy explains why unlimited observations restricted to patch order at most r cannot recover an order-$(r+1)$ interaction.

5 CIRCUIT IDENTIFIABILITY

We first reduce circuit equivalence to a kernel condition over binary-mask differences, then count zero-sum omissions and realize the sharp count with patched ReLU components. We finally distribute cancellation and establish perturbation stability.

Let a patched neural model have a fixed unmasked response $b(z)$ and $n \in \mathbb{N}$ maskable components with bounded responses $f_1, \dots, f_n : [0,1]^m \to \mathbb{R}$. We assume that the selected behavior score has the additive form

$$Y(z, \boldsymbol{x}) = b(z) + \sum_{j=1}^{n} x_j f_j(z),$$

where $\boldsymbol{x} \in \{0,1\}^n$ is a component mask selecting the component set $C = \{j : x_j = 1\}$, which we call a circuit, and z controls activation patching. Prescribed output-edge groups at a common linear readout and parallel branches with mask-independent additive contributions satisfy the assumption. Since b occurs in both the circuit and full-model response, it cancels from every faithfulness comparison. We therefore subtract b and analyze only the maskable responses.

For circuits $C, D \subseteq [n]$, let $C \triangle D = (C \setminus D) \cup (D \setminus C)$ denote their symmetric difference. For a circuit $C \subseteq [n]$, the circuit response and the full-model response are $F_C = \sum_{j\in C} f_j$ and

$F_{\mathsf{Full}} = \sum_{j=1}^{n} f_j$. The component vocabulary and granularity form part of the model specification. Our formal circuit is a subset of the prescribed vocabulary and need not be a connected subgraph. A component may represent a neuron, attention head, multilayer-perceptron (MLP) block, residual branch, or prescribed edge group, provided that the induced masked score satisfies the additive model above; Section 6 treats uniformly bounded residuals from an additive surrogate.

For a bounded function $G : [0,1]^m \to \mathbb{R}$, let $\|G\|_{\infty,\mathcal{Z}_r} = \sup_{z\in\mathcal{Z}_r} |G(z)|$.

Definition 7. *For $\varepsilon \geq 0$, a circuit C is (r,ε)-faithful when*

$$\|F_C - F_{\mathsf{Full}}\|_{\infty,\mathcal{Z}_r} \leq \varepsilon. \tag{1}$$

Throughout, faithfulness means uniform scalar-response faithfulness on the stated patching domain. We call $(r,0)$-faithfulness r-faithfulness. An r-faithful circuit is inclusion-minimal when no proper subset is r-faithful, and it is minimum-cardinality when $|C|$ is minimum among all r-faithful circuits. We let $\mathsf{MinFaith}_r$ denote the family of minimum-cardinality r-faithful circuits.

Let $\mathcal{V}_r = \ell_\infty(\mathcal{Z}_r)$ be the vector space of bounded real functions on $\mathcal{Z}_r$. For $j \in [n]$, let $\boldsymbol{v}_j^{(r)} = f_j \restriction_{\mathcal{Z}_r} \in \mathcal{V}_r$. The order-$r$ signature operator $A_r : \mathbb{R}^n \to \mathcal{V}_r$ is

$$A_r\boldsymbol{x} = \sum_{j=1}^{n} x_j \boldsymbol{v}_j^{(r)}.$$

Let $\boldsymbol{x}_C \in \{0,1\}^n$ be the indicator of C. For $d \in \mathbb{N}$, let $\mathbf{1}_d \in \mathbb{R}^d$ denote the all-ones vector; we write $\mathbf{1} = \mathbf{1}_n$ in the n-component setting. For $k \in \{0,\dots,n\}$, let

$$\mathcal{D}_C(k) = \{\boldsymbol{x} - \boldsymbol{x}_C \mid \boldsymbol{x} \in \{0,1\}^n,\ \|\boldsymbol{x}\|_0 \leq k\}.$$

Theorem 8. *For $C, D \subseteq [n]$, the following statements hold.*

(i) C is r-faithful iff $A_r\boldsymbol{x}_C = A_r\mathbf{1}$.

(ii) F_C and F_D agree on $\mathcal{Z}_r$ iff $\boldsymbol{x}_C - \boldsymbol{x}_D \in \ker(A_r)$.

(iii) An r-faithful circuit C of cardinality k is the unique minimum-cardinality circuit iff

$$\ker(A_r) \cap \mathcal{D}_C(k) = \{\mathbf{0}\}. \tag{2}$$

Proof sketch. Linearity gives $(F_C - F_{\mathsf{Full}}) \restriction_{\mathcal{Z}_r} = A_r(\boldsymbol{x}_C - \mathbf{1})$ and $(F_C - F_D) \restriction_{\mathcal{Z}_r} = A_r(\boldsymbol{x}_C - \boldsymbol{x}_D)$. These identities prove (i) and (ii). A competing faithful mask of size at most k exists exactly when a nonzero vector $\boldsymbol{x} - \boldsymbol{x}_C$ lies in $\ker(A_r) \cap \mathcal{D}_C(k)$, which proves (iii). The condition therefore separates minimum cardinality from uniqueness of the minimum solution. □

The same condition has a useful zero-sum form. Let

$$\mathcal{H}_r^0 = \{Z \subseteq [n] \mid A_r\boldsymbol{x}_Z = 0\} \quad \text{and} \quad q_r^* = \max\{|Z| \mid Z \in \mathcal{H}_r^0\}.$$

We have $\varnothing \in \mathcal{H}_r^0$, so q_r^* is defined.

Proposition 9. *The minimum faithful cardinality and the minimum faithful circuits are*

$$\min\{|C| \mid C \text{ is } r\text{-faithful}\} = n - q_r^* \quad \text{and} \quad \mathsf{MinFaith}_r = \{[n] \setminus Z \mid Z \in \mathcal{H}_r^0,\ |Z| = q_r^*\}.$$

Inclusion-minimal faithful circuits are the complements of inclusion-maximal members of $\mathcal{H}_r^0$.

Proof sketch. For $Z_C = [n] \setminus C$, additivity gives that C is r-faithful iff $A_r\boldsymbol{x}_{Z_C} = 0$; moreover, $|C| = n - |Z_C|$. Maximizing the zero-sum omission Z_C therefore minimizes C, which proves both formulas. Circuit multiplicity therefore encodes the extremal structure of the zero-sum hypergraph induced by the component signatures. □

Theorem 10. *Let $0 \leq r < m$, $1 \leq k \leq n-1$, and $n - 1 \leq \binom{m}{r+1}$. A feedforward ReLU model with a patched layer of m scalar patching sites and n maskable additive components satisfies:*

(i) $|\mathsf{MinFaith}_r| = \binom{n-1}{k}$;

(ii) every member of $\mathsf{MinFaith}_r$ has k components and is inclusion-minimal;

(iii) any two distinct members of $\mathsf{MinFaith}_r$ *disagree on a binary patch of order* $r+1$*; and*

(iv)

$$\min_{D \subsetneq [n]} \|F_D - F_{\mathsf{Full}}\|_{\infty, \mathcal{Z}_{r+1}} = 1.$$

In particular, $[n]$ *is the unique order-*$(r+1)$ *faithful circuit.*

The equalities in (i) and (ii) hold on the complete continuous domain $\mathcal{Z}_r$*. The model has* $O(m + (n-1)(r+1))$ *nonzero parameters, including the patched interface.*

Proof sketch. For $p = n-1$, choose distinct $(r+1)$-subsets $H_1, \dots, H_p$ and let

$$h_j(z) = \mathrm{ReLU}\left(\sum_{i \in H_j} z_i - r\right), \quad f_j = 1 + h_j \ \ (j \in [p]) \quad \text{and } f_n = -(p-k) - \sum_{j=1}^{p} h_j.$$

The full response is $F_{\mathsf{Full}} = k$, and every h_j vanishes on $\mathcal{Z}_r$. If a components indexed by $[p]$ are selected and $\gamma \in \{0,1\}$ indicates whether component n is selected, the low-order response is $a - \gamma(p-k)$. Equality with k holds only for $(\gamma, a) = (0, k)$ or $(1, p)$; thus, the minimum circuits are exactly the k-subsets of $[p]$. At $z = \mathbf{1}_{H_i}$, we have $h_i = 1$ and $h_j = 0$ for $j \neq i$. The witness separates distinct minimum circuits and gives each proper low-order candidate error at least one at order $r+1$. □

The construction, together with the following universal upper bound, gives an exact extremal characterization for every nontrivial minimum faithful cardinality.

Theorem 11. *Consider any additive model with* n *maskable components on* $\mathcal{Z}_r$*. If the minimum* r*-faithful cardinality is* $1 \le k < n$*, then*

$$|\mathsf{MinFaith}_r| \le \binom{n-1}{k}.$$

Whenever $n - 1 \le \binom{m}{r+1}$*, the model in Theorem 10 attains equality. The upper bound remains valid on any prescribed intervention domain.*

Proof sketch. We remove the zero order-r signatures, which belong to every maximum zero-sum omission. Let N be the number of remaining signatures and $q = N - k$ their maximum zero-sum size. If $q = 0$, the omission family has one member. If $1 \le q \le N/2$, two disjoint maximum omissions would have a larger zero-sum union, so the Erdős–Ko–Rado theorem gives $|\mathsf{MinFaith}_r| \le \binom{N-1}{q-1} = \binom{N-1}{k} \le \binom{n-1}{k}$. If $q > N/2$, a rational linear projection preserving every subset-sum zero pattern reduces the signatures to the sharp zero-sum bound of Móricz & Nagy (2026), which gives the same inequality. The model in Theorem 10 realizes equality. □

Corollary 12. *Assume* $n-1 \le \binom{m}{r+1}$*. For every* $1 \le k < n$*, the exact maximum number of minimum* r*-faithful circuits in the stated additive class is* $\binom{n-1}{k}$*. In particular, at* $k = \lceil n/2 \rceil$*,*

$$\binom{n-1}{\lceil n/2 \rceil} = 2^{\,n - \frac{1}{2}\log_2 n + O(1)}.$$

For the attaining family, every $C \in \mathsf{MinFaith}_r$ has the same response $F_C = k$ on $\mathcal{Z}_r$. Therefore, every recovery procedure whose input is restricted to complete order-r candidate responses receives the same evidence for all $\binom{p}{k}$ candidates. We quantify the additional evidence required after increasing the patch order. Let $p = n-1$, assume $1 \le k \le p-1$, and let an oracle return $F_C(z)$ for an unknown $C \in \mathsf{MinFaith}_r$ and an adaptively chosen $z \in \mathcal{Z}_{r+1}$.

Theorem 13. *In the exact noiseless response-oracle model above, every deterministic adaptive algorithm that identifies every* $C \in \mathsf{MinFaith}_r$ *requires at least* $p - 1 = n - 2$ *queries from* $\mathcal{Z}_{r+1}$ *in the worst case. Querying* $\mathbf{1}_{H_1}, \dots, \mathbf{1}_{H_{p-1}}$ *attains the bound nonadaptively.*

Proof sketch. A query in $\mathcal{Z}_{r+1}$ activates at most one of the distinct $(r+1)$-site features h_j, and therefore reveals at most one membership bit of the unknown k-subset of $[p]$. An adversary preserves two consistent k-subsets until $p-1$ membership bits have been queried. The patches $\mathbf{1}_{H_1}, \dots, \mathbf{1}_{H_{p-1}}$ reveal those bits, and the known cardinality determines membership of index p. □

For $k = n - 1$, $\mathsf{MinFaith}_r$ is a singleton and requires zero queries.

For $r = 1$, $m = 4$, $n = 6$, and $k = 3$, five pairs $H_j \subseteq [4]$ give $\binom{5}{3} = 10$ minimum circuits; Theorem 11 proves optimality, and the patch $\mathbf{1}_{H_j}$ reveals whether component j belongs to the unknown circuit.

The model in Theorem 10 concentrates cancellation in one output group. The next theorem shows that exponential multiplicity persists with uniformly bounded per-component fan-in, parameter count, and response norm.

Theorem 14. *Let $L, p, q \in \mathbb{N}$ satisfy $L \geq 2$, $q < p \leq Lq$, $0 \leq r < m$, and $p \leq \binom{m}{r+1}$. There is a feedforward ReLU model with $n = p + q$ and $k = p - q$ maskable additive components such that:*

(i) the minimum order-r faithful cardinality is k, and $|\mathsf{MinFaith}_r| = \binom{p}{q}$;

(ii) any two distinct members of $\mathsf{MinFaith}_r$ disagree on a binary patch of order $r + 1$; and

(iii) every component depends on at most L order-$(r + 1)$ ReLU units, has $O(L(r + 1))$ nonzero local parameters, and has uniform response norm at most $L + 1$ on $\mathcal{Z}_{r+1}$.

The total model has $O(m + p(r + 1))$ nonzero parameters. In particular, for $p = 2q$, $n = 3q$, $k = n/3$, and

$$|\mathsf{MinFaith}_r| = \binom{2q}{q} = 2^{\frac{2}{3}n - \frac{1}{2}\log_2 n + O(1)}.$$

Proof sketch. Partition the p responses $1 + h_j$ into q blocks of size at most L, and assign block B the local cancellation component $-1 - \sum_{j \in B} h_j$. The full response is $p - q = k$. On $\mathcal{Z}_r$, a mask selecting a of the first p components and b local cancellation components is faithful only if $a - b = k$; its size is $a + b = k + 2b \geq k$. Equality holds exactly for $b = 0$ and $a = k$, which gives $\binom{p}{k} = \binom{p}{q}$ minimum circuits. The order-$(r + 1)$ witnesses separate them, and each cancellation component depends only on the ReLU units in its block. □

For $t \in (0, 1)$, let $H_2(t) = -t \log_2 t - (1 - t) \log_2(1 - t)$ be the binary entropy.

Corollary 15. *For every fixed $\alpha \in (0, 1)$, there is a constant per-component fan-in bound L_α and a sequence of models from Theorem 14 with $k/n \to \alpha$ and*

$$\log_2 |\mathsf{MinFaith}_r| = c_\alpha n - \tfrac{1}{2}\log_2 n + O(1) \quad \textit{and } c_\alpha = \frac{1 + \alpha}{2} H_2\left(\frac{1 - \alpha}{1 + \alpha}\right) > 0.$$

Therefore, exponential ambiguity with bounded per-component fan-in occurs at every limiting circuit-size fraction $\alpha \in (0, 1)$.

Corollary 16. *For the model in Theorem 14, let every component response be perturbed by at most δ uniformly on $\mathcal{Z}_{r+1}$. If*

$$2q\delta \leq \varepsilon, \quad n\delta < 1 - \varepsilon \quad \textit{and} \quad 2k\delta < 1,$$

then exactly $\binom{p}{q}$ circuits are minimum-cardinality (r, ε)-faithful circuits, and any two distinct circuits among them have order-$(r + 1)$ separation at least $1 - 2k\delta$.

Proof sketch. Each intended circuit omits $2q$ components, so its low-order error is at most $2q\delta$. A smaller circuit has unperturbed gap at least 1, whereas the perturbation changes that gap by less than $1 - \varepsilon$; a size-k circuit using a cancellation component has unperturbed gap at least 2. Finally, the order-$(r + 1)$ witness separates two intended circuits by 1, and their symmetric difference contains at most $2k$ perturbed components, giving $1 - 2k\delta$. □

We next replace exact equality by the tolerance criterion in equation 1. Let $(f_1, \ldots, f_n)$ be the component responses constructed in Theorem 10, and let $(\widetilde{f}_1, \ldots, \widetilde{f}_n)$ be perturbed responses. Their circuit and full-model responses are denoted by $\widetilde{F}_C$ and $\widetilde{F}_{\mathsf{Full}}$.

Theorem 17. *Let $p = n - 1$, $0 < \varepsilon < 1$, and $\delta > 0$ satisfy*

$$(p - k + 1)\delta \leq \varepsilon, \quad (p + 1)\delta < 1 - \varepsilon \quad \textit{and} \quad 2k\delta < 1. \tag{3}$$

If $\max_{j\in[n]}\|\widetilde{f}_j - f_j\|_{\infty,\mathcal{Z}_{r+1}} \le \delta$*, then exactly* $\binom{p}{k}$ *circuits are minimum-cardinality* (r,ε)*-faithful circuits. Each distinct pair* C, D *among them satisfies*

$$\|\widetilde{F}_C - \widetilde{F}_D\|_{\infty,\mathcal{Z}_{r+1}} \ge 1 - 2k\delta > 0.$$

The full mask is the only $(r+1,\varepsilon)$*-faithful mask.*

Proof sketch. The proof sums at most $p-k+1$, $p+1$, and $2k$ component perturbations for the three respective bounds. For every $0 < \varepsilon < 1$, equation 3 admits a positive δ. Uniform continuity of a fixed ReLU architecture on the compact domain $\mathcal{Z}_{r+1}$ transfers the response neighborhood in Theorem 17 to an open parameter neighborhood. □

Nontrivial zero-sum cancellation among component signatures is necessary for multiplicity within the additive class. A pointed convex cone $\mathcal{P} \subseteq \mathcal{V}_r$ satisfies $\mathcal{P} \cap (-\mathcal{P}) = \{0\}$.

Proposition 18. *Assume that* $\boldsymbol{v}_j^{(r)} \in \mathcal{P}$ *for every* $j \in [n]$*, where* $\mathcal{P} \subseteq \mathcal{V}_r$ *is a pointed convex cone. Then, the unique minimum-cardinality* r*-faithful circuit is*

$$C_+ = \{j \in [n] \mid \boldsymbol{v}_j^{(r)} \neq 0\}.$$

Proof sketch. If $w_i \in \mathcal{P}$ and $\sum_i w_i = 0$, then $w_1 \in \mathcal{P}$ and $-w_1 = \sum_{i>1} w_i \in \mathcal{P}$, so pointedness gives $w_1 = 0$; iteration gives $w_i = 0$ for every i. Thus, omitted signatures sum to zero exactly when all of them are zero, equivalently $C_+ \subseteq C$, whose unique smallest solution is C_+. □

The proposition therefore covers component responses that are pointwise nonnegative on $\mathcal{Z}_r$. It also applies when a linear $\phi : \mathcal{V}_r \to \mathbb{R}$ is positive on each nonzero signature: the cone $\mathcal{P} = \mathrm{cone}\{\boldsymbol{v}_j^{(r)} \mid j \in [n]\}$ has $\phi(w) > 0$ for every $w \in \mathcal{P} \setminus \{0\}$, hence $\mathcal{P} \cap (-\mathcal{P}) = \{0\}$. Proposition 18 gives a complementary sufficient condition for uniqueness: signatures in a common pointed cone have a unique minimum circuit, while our constructions use structured zero-sum cancellation to attain exponential multiplicity.

6 Approximate additivity

The next result shows that extremal circuit multiplicity is stable under uniformly bounded departures from additive masking.

Ablating a component in a general computation graph can change later component responses, so the mask-to-response map need not be additive. Let $Y : [0,1]^m \times \{0,1\}^n \to \mathbb{R}$ be an arbitrary masked response and let $Y_0(z, \boldsymbol{x}) = b(z) + \sum_{j=1}^n x_j f_j(z)$ be any additive surrogate. We write $R = Y - Y_0$ and assume that R is bounded. For $s \in \{0, \dots, m\}$, we define the order-s additive-surrogate residual radius by $\eta_s = \sup_{z\in\mathcal{Z}_s,\, \boldsymbol{x}\in\{0,1\}^n} |R(z,\boldsymbol{x})|$. For a circuit C, let $e_s(C)$ and $e_s^0(C)$ denote the actual and surrogate faithfulness errors, respectively. Let $d_s(C,D)$ and $d_s^0(C,D)$ denote the corresponding pairwise response distances.

Theorem 19. *For every* $C, D \subseteq [n]$ *and every* $s \in \{0, \dots, m\}$*,*

$$|e_s(C) - e_s^0(C)| \le 2\eta_s \quad \textit{and} \quad |d_s(C,D) - d_s^0(C,D)| \le 2\eta_s.$$

Let $k \in \{0, \dots, n\}$*, let* $\alpha, \beta, \varepsilon \ge 0$*, and let* $\mathcal{A}$ *be a nonempty family of* k*-component masks such that* $e_r^0(C) \le \alpha$ *for* $C \in \mathcal{A}$*, while* $e_r^0(D) \ge \beta$ *for every* $D \notin \mathcal{A}$ *with* $|D| \le k$*. If*

$$\alpha + 2\eta_r \le \varepsilon < \beta - 2\eta_r,$$

then $\mathcal{A}$ *is exactly the family of minimum-cardinality* (r,ε)*-faithful circuits for* Y*. If, for some* $\gamma \ge 0$*, additionally* $d_s^0(C,D) \ge \gamma$ *for distinct* $C, D \in \mathcal{A}$*, then* $d_s(C,D) \ge \gamma - 2\eta_s$*.*

Proof sketch. For a circuit C, let

$$Y(z, \boldsymbol{x}_C) - Y(z, \mathbf{1}) = \big[Y_0(z, \boldsymbol{x}_C) - Y_0(z, \mathbf{1})\big] + \big[R(z, \boldsymbol{x}_C) - R(z, \mathbf{1})\big].$$

The last bracket has uniform norm at most $2\eta_s$ on $\mathcal{Z}_s$. The reverse triangle inequality gives the faithfulness-error bound; replacing the full mask by $\boldsymbol{x}_D$ gives the distance bound. The strict gap between α and β then transfers the stated family. □

Both ReLU constructions have $\alpha = 0$, $\beta = 1$, and $\gamma = 1$ for $s = r + 1$. The sharp construction also gives every proper surrogate mask order-$(r + 1)$ faithfulness error at least one. If

$$2\eta_r \le \varepsilon < 1 - 2\eta_{r+1}, \tag{4}$$

then the actual masked response has the extremal minimum-circuit family at order r and the full mask is the only $(r + 1, \varepsilon)$-faithful mask. Distinct low-order candidates have order-$(r + 1)$ distance at least $1 - 2\eta_{r+1}$. Because $\mathcal{Z}_r \subseteq \mathcal{Z}_{r+1}$, $\eta_r \le \eta_{r+1}$; hence $\eta_{r+1} < 1/4$ makes $\varepsilon = 1/2$ admissible and gives pairwise distance greater than $1/2$. The residual bound permits arbitrary higher-order mask-interaction structure. For $j \in [n]$, let $\boldsymbol{e}_j \in \{0, 1\}^n$ be the j-th standard basis vector. For the canonical surrogate $b(z) = Y(z, \mathbf{0})$ and $f_j(z) = Y(z, \boldsymbol{e}_j) - Y(z, \mathbf{0})$, pointwise Möbius expansion in the mask variables writes $R(z, \boldsymbol{x}) = \sum_{|T|\ge 2} \mu_T(z) \prod_{j\in T} x_j$, where $\mu_T(z)$ is the corresponding mask-interaction coefficient. For the canonical surrogate, R contains only mask interactions of order at least two, and η_s bounds their aggregate contribution. For finite $Q \subseteq \mathcal{Z}_s$, evaluations determine only $\eta_{s,Q} = \max_{z\in Q,\, \boldsymbol{x}\in\{0,1\}^n} |R(z, \boldsymbol{x})| \le \eta_s$. Transfer on $\mathcal{Z}_s$ requires a certified upper bound on η_s, for example from an analytic bound or a Lipschitz bound combined with a finite cover of $\mathcal{Z}_s$.

7 Finite-domain uniqueness test

Let $s \in \{0, \dots, m\}$, $Q = \{z_1, \dots, z_M\} \subseteq \mathcal{Z}_s$, $B_{tj} = f_j(z_t)$, and $\boldsymbol{y} = B\mathbf{1}_n$. The minimum surrogate-faithful cardinality on Q is

$$\min_{\boldsymbol{x}\in\{0,1\}^n} \mathbf{1}_n^\top \boldsymbol{x} \quad \text{subject to} \quad -\varepsilon\mathbf{1}_M \le B\boldsymbol{x} - \boldsymbol{y} \le \varepsilon\mathbf{1}_M.$$

After an optimum $\boldsymbol{x}^*$ of size k^*, impose $\mathbf{1}_n^\top \boldsymbol{x} = k^*$ and the no-good cut $\sum_{j:x_j^*=1}(1 - x_j) + \sum_{j:x_j^*=0} x_j \ge 1$.

Proposition 20. *The first program returns the exact minimum cardinality on Q. The second is infeasible exactly when $\boldsymbol{x}^*$ is the unique optimum; otherwise it returns a distinct optimum. Repeated no-good cuts enumerate the complete minimum family.*

Exact arithmetic certifies rational data; floating-point certification requires certified input and solver-error bounds and strict constraint margins.

Proof sketch. The linear constraints are equivalent to $\|B\boldsymbol{x} - B\mathbf{1}\|_\infty \le \varepsilon$. For binary masks, the no-good left-hand side equals $\|\boldsymbol{x} - \boldsymbol{x}^*\|_1$, which vanishes exactly at $\boldsymbol{x}^*$. The fixed-cardinality second solve therefore gives the stated uniqueness test. □

We use the program to test Y_0 on Q; Theorem 19 transfers the family on Q using $\eta_{s,Q}$, while a $\mathcal{Z}_s$ conclusion requires a certified upper bound on η_s.

8 Discussion and Conclusion

We have distinguished faithfulness, inclusion minimality, minimum cardinality, and identifiability. Complete binary order-r evidence determines $\tau_r F$; Theorems 10 and 11 show that complete continuous evidence on $\mathcal{Z}_r$ can admit the exact maximum $\binom{n-1}{k}$ of minimum faithful circuits. In the sharp construction, complete order-$(r + 1)$ evidence excludes every proper mask, yet identifying an unknown member requires $n - 2$ deterministic queries in the worst case when $1 \le k \le n - 2$. Pointed-cone signatures yield uniqueness; our structured zero-sum constructions attain exponential multiplicity under the stated margins. The bounded-fan-in construction distributes cancellation, and the residual bounds preserve the minimum-circuit family under approximate additivity.

Our worst-case guarantees concern a fixed component vocabulary; continuous order-r evidence is idealized, and extending a finite-domain test to $\mathcal{Z}_s$ requires an independent uniform certificate. Determining the maximum number of minimum faithful circuits under a prescribed per-component fan-in bound and the query complexity beyond the extremal construction remain open problems. Complete bounded-order patching evidence can therefore certify faithfulness without determining a unique minimum circuit; unique recovery requires additional structural information or higher-order interventions.

REPRODUCIBILITY STATEMENT

The formal statements specify their assumptions, and the appendix provides complete proofs of the stated results.

AI USE STATEMENT

Generative AI was used only for language and readability support. The author reviewed the submission and remains fully responsible for its content.

REFERENCES

Federico Adolfi, Martina G Vilas, and Todd Wareham. The computational complexity of circuit discovery for inner interpretability. In *International Conference on Learning Representations*, volume 2025, pp. 11389–11404, 2025.

Potito Aghilar, Sabino Roccotelli, Stanislao Fidanza, Vito Walter Anelli, Sebastiano Stramaglia, and Tommaso Di Noia. The representational limit of scalar interactions: An interventional decomposition, 2026. URL https://arxiv.org/abs/2606.19410.

Alaa Anani, Tobias Lorenz, Bernt Schiele, Mario Fritz, and Jonas Fischer. Certified circuits: Stability guarantees for mechanistic circuits. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IU0c9sP5b5.

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 210–227. PMLR, 2026. URL https://proceedings.mlr.press/v337/asiaee26d.html.

Alireza Bayat Makou, Jingcheng Niu, Subhabrata Dutta, and Iryna Gurevych. Many circuits, one mechanism: Input variation and evaluation granularity in circuit discovery. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Ur42Xx1QPv.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. *Advances in Neural Information Processing Systems*, 37:18506–18534, 2024.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, Dimitris N. Metaxas, Zhaoran Wang, Yutao Yue, and Gerald Penn. All circuits lead to rome: Rethinking functional anisotropy in circuit and sheaf discovery for LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3uC9teMlUt.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Ashim Dhor and Pin-Yu Chen. Intrinsic structure: Spectral identifiability for mechanistic interpretability, 2026. URL https://arxiv.org/abs/2608.10172.

Paul Erdős, Chao Ko, and Richard Rado. Intersection theorems for systems of finite sets. *Quart. J. Math. Oxford Ser.(2)*, 12:313–320, 1961.

Yigit Efe Erginbas, Justin Singh Kang, Elizabeth Polito, and Kannan Ramchandran. Adaptive Sparse Möbius Transforms for Learning Polynomials, 2026. URL https://arxiv.org/abs/2602.06246.

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Timsb74vIY.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

John C Harsanyi. A simplified bargaining model for the n-person cooperative game. In *Papers in game theory*, pp. 44–70. Springer, 1982.

Jing Huang. A Proof of a Conjecture of Móricz and Nagy on Rational-Value Sums, 2026. URL https://arxiv.org/abs/2603.18805.

Gyula OH Katona. A simple proof of the erdös-chao ko-rado theorem. *Journal of Combinatorial Theory, Series B*, 13(2):183–184, 1972.

Ajay Pravin Mahale. Explanation multiplicity: Circuit-level interpretability evidence does not survive defensible analytic variation, 2026. URL https://arxiv.org/abs/2608.13754.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, volume 2025, pp. 32900–32923, 2025.

Maxime Méloux, François Portet, and Maxime Peyrard. Mechanistic interpretability as statistical estimation: A variance analysis. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Z7KqxFpWzN.

Benjamin Móricz and Zoltán Lóránt Nagy. Maximizing the number of rational-value sums or zero-sums. *European Journal of Combinatorics*, 133:104324, 2026.

Yaniv Nikankin, Dana Arad, Itay Itzhak, Anja Reusch, Adi Simhi, Gal Kesten, and Yonatan Belinkov. BlackboxNLP-2025 MIB shared task: Improving circuit faithfulness via better edge selection. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 521–527, 2025.

Daking Rai, Mor Geva, and Ziyu Yao. Data-driven circuit discovery for interpretability of language models, 2026. URL https://arxiv.org/abs/2605.09129.

Gian-Carlo Rota. On the foundations of combinatorial theory: I. theory of möbius functions. In *Classic Papers in Combinatorics*, pp. 332–360. Springer, 1964.

Ali Sahi, Imran Kutianawala, Medhansh Beeram, Siwoo Song, and Aryan Shrivastava. Characterizing mechanistic uniqueness and identifiability through circuit analysis. In *ICLR 2026 Workshop on Representational Alignment (Re^4-Align)*, 2026. URL https://openreview.net/forum?id=cHpZ5DE8wj.

Yang Sheng and Jie Fu. Circuit claims depend on what is extracted and how it is compared, 2026. URL https://arxiv.org/abs/2607.18921.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9lycwRxAOI.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pp. 9259–9268. PMLR, 2020.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching, 2026. URL https://arxiv.org/abs/2606.27510.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Frank Zhengqing Wu, Francesco Tonin, and Volkan Cevher. Demystifying variance in circuit discovery of LLMs, 2026. URL https://arxiv.org/abs/2606.16920.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

APPENDIX: PROOFS

A PROOFS FOR SECTION 3

We first establish Möbius inversion and then prove the ReLU realization.

A.1 PROOF OF LEMMA 21

Lemma 21. *For every $F : 2^{[m]} \to \mathbb{R}$, the coefficients in the Möbius formula in Definition 1 are the unique coefficients satisfying the inversion identity $F(S) = \sum_{T \subseteq S} a_T(F)$.*

Proof. Substitute the Möbius formula in Definition 1 into the right-hand side of the inversion identity $F(S) = \sum_{T \subseteq S} a_T(F)$. For fixed $S \subseteq [m]$,

$$\begin{aligned}\sum_{T \subseteq S} a_T(F) &= \sum_{T \subseteq S} \sum_{U \subseteq T} (-1)^{|T|-|U|} F(U) \\ &= \sum_{U \subseteq S} F(U) \sum_{T : U \subseteq T \subseteq S} (-1)^{|T|-|U|}.\end{aligned}$$

For a fixed $U \subseteq S$, write $T = U \cup W$ with $W \subseteq S \setminus U$. The inner coefficient equals

$$\sum_{W \subseteq S \setminus U} (-1)^{|W|} = (1-1)^{|S \setminus U|}.$$

It is 1 when $U = S$ and 0 otherwise. The entire sum is therefore $F(S)$, proving existence.

For uniqueness, assume $F(S) = \sum_{T \subseteq S} b_T$ for every $S \subseteq [m]$. For each $T \subseteq [m]$, finite-sum rearrangement and the binomial identity above give

$$\begin{aligned}\sum_{U \subseteq T} (-1)^{|T|-|U|} F(U) &= \sum_{U \subseteq T} (-1)^{|T|-|U|} \sum_{V \subseteq U} b_V \\ &= \sum_{V \subseteq T} b_V \sum_{U : V \subseteq U \subseteq T} (-1)^{|T|-|U|} \\ &= \sum_{V \subseteq T} b_V (1-1)^{|T \setminus V|} = b_T.\end{aligned}$$

The left-hand side equals $a_T(F)$, so $b_T = a_T(F)$ for every T. □

A.2 PROOF OF PROPOSITION 3

Proof. For nonempty $T \subseteq [m]$, we define

$$h_T(z) = \mathrm{ReLU}\left(\sum_{i \in T} z_i - |T| + 1 \right).$$

If $z \in \{0,1\}^m$ and $z_i = 1$ for every $i \in T$, the argument of the ReLU is 1, so $h_T(z) = 1$. If at least one z_i with $i \in T$ is zero, then $\sum_{i \in T} z_i \leq |T| - 1$, the argument is nonpositive, and $h_T(z) = 0$. Thus,

$$h_T(z) = \prod_{i \in T} z_i \quad \text{on} \quad \{0,1\}^m.$$

Using the unique multilinear expansion following Definition 1, let the output bias equal $a_\varnothing(F)$, include h_T precisely for the nonempty $T \in \mathrm{supp}(F)$, and give it output weight $a_T(F)$. The network output is

$$a_\varnothing(F) + \sum_{\substack{T \in \mathrm{supp}(F) \\ T \neq \varnothing}} a_T(F) h_T(z) = p_F(z) = F(z)$$

on $\{0,1\}^m$. If $\deg(F) \leq d$, every included T has size at most d, so the corresponding hidden unit has fan-in at most d. □

B Proofs for Section 4

We derive the resolution theorem and both corollaries from Möbius inversion.

B.1 Proof of Theorem 4

Proof. We prove the three claims separately.

Part (i), forward direction. Assume $F \equiv_r G$. Let $T \subseteq [m]$ satisfy $|T| \leq r$. Every $U \subseteq T$ also has $|U| \leq r$, so $F(U) = G(U)$. By the Möbius formula in Definition 1,

$$a_T(F) = \sum_{U \subseteq T} (-1)^{|T|-|U|} F(U) = \sum_{U \subseteq T} (-1)^{|T|-|U|} G(U) = a_T(G).$$

Part (i), reverse direction. Assume $a_T(F) = a_T(G)$ for every $|T| \leq r$. Let $S \in \mathcal{S}_r$. Every subset $T \subseteq S$ has $|T| \leq r$. Lemma 21 gives

$$F(S) = \sum_{T \subseteq S} a_T(F) = \sum_{T \subseteq S} a_T(G) = G(S).$$

Hence, $F \equiv_r G$.

Part (ii). The monomials

$$z_T(z) = \prod_{i \in T} z_i \quad \text{for } |T| \leq d$$

form a basis of $\mathcal{F}_{m,d}$ by the uniqueness in Lemma 21. By part (i), a function $F \in \mathcal{F}_{m,d}$ lies in the kernel of $\mathcal{E}_r$ exactly when

$$a_T(F) = 0 \qquad \text{for all } |T| \leq \min\{r, d\}.$$

The kernel is therefore spanned by the monomials with $r < |T| \leq d$. The kernel dimension is

$$\sum_{j=r+1}^{d} \binom{m}{j} = N_{m,d} - N_{m,\min\{r,d\}},$$

where an empty sum is zero. Rank–nullity and $\dim(\mathcal{F}_{m,d}) = N_{m,d}$ give the asserted rank.

Part (iii), necessity. Assume ρ is r-identifiable. By part (i),

$$\tau_r F = \tau_r G \quad \text{if and only if} \quad F \equiv_r G.$$

Let $\bar{\rho}(\tau_r F) = \rho(F)$. If $\tau_r F = \tau_r G$, then $F \equiv_r G$, so $\rho(F) = \rho(G)$. Thus, $\bar{\rho}$ is well-defined, and by construction $\rho = \bar{\rho} \circ \tau_r$. Since τ_r is surjective onto the image, the factorization uniquely determines $\bar{\rho}$.

Part (iii), sufficiency. Conversely, suppose $\rho = \bar{\rho} \circ \tau_r$. If $F \equiv_r G$, part (i) gives $\tau_r F = \tau_r G$, and hence

$$\rho(F) = \bar{\rho}(\tau_r F) = \bar{\rho}(\tau_r G) = \rho(G).$$

Therefore, ρ is r-identifiable. □

B.2 Proof of Corollary 5

Proof. By Theorem 4(ii), the restriction of $\mathcal{E}_r$ to $\mathcal{F}_{m,d}$ is injective exactly when the kernel has dimension zero:

$$N_{m,d} - N_{m,\min\{r,d\}} = 0.$$

The sequence $N_{m,0}, \ldots, N_{m,m}$ is strictly increasing because $\binom{m}{j} > 0$ for $0 \le j \le m$. Thus, equality holds exactly when $\min\{r, d\} = d$, equivalently $r \ge d$. □

B.3 Proof of Corollary 6

Proof. Since $\mathcal{S}_r \subseteq \mathcal{S}_{r+1}$, equality on $\mathcal{S}_{r+1}$ implies equality on $\mathcal{S}_r$, proving $\equiv_{r+1} \subseteq \equiv_r$.

To prove strictness, let $T_0 = \{1, \ldots, r+1\}$, let $F_0 = 0$, and let

$$F_1(S) = \mathbf{1}[T_0 \subseteq S].$$

If $|S| \le r$, then $T_0 \not\subseteq S$, so $F_1(S) = 0 = F_0(S)$. Hence, $F_0 \equiv_r F_1$. At $S = T_0$, however, $F_1(T_0) = 1 \ne 0 = F_0(T_0)$, so $F_0 \not\equiv_{r+1} F_1$.

Finally, $F_1(z) = \prod_{i \in T_0} z_i$ on the Boolean cube. By the ReLU identity used in Proposition 3, it is realized by the single hidden unit

$$\mathrm{ReLU}\left(\sum_{i=1}^{r+1} z_i - r\right).$$

□

C Proofs for Section 5

We prove the mask characterizations, both constructions, the sharp upper bound, the stability claims, and the pointed-cone criterion.

C.1 Proof of Theorem 8

Proof. For every circuit $C \subseteq [n]$, linearity of restriction and the definitions of F_C and A_r give

$$F_C \restriction_{\mathcal{Z}_r} = \sum_{j \in C} f_j \restriction_{\mathcal{Z}_r} = \sum_{j=1}^{n} (\boldsymbol{x}_C)_j \boldsymbol{v}_j^{(r)} = A_r \boldsymbol{x}_C, \tag{23}$$

and

$$F_{\mathsf{Full}} \restriction_{\mathcal{Z}_r} = \sum_{j=1}^{n} f_j \restriction_{\mathcal{Z}_r} = A_r \mathbf{1}. \tag{24}$$

Equations 23 and 24 give

$$F_C \restriction_{\mathcal{Z}_r} = F_{\mathsf{Full}} \restriction_{\mathcal{Z}_r} \iff A_r \boldsymbol{x}_C = A_r \mathbf{1},$$

which proves part (i). Applying equation 23 to C and D gives

$$F_C \restriction_{\mathcal{Z}_r} = F_D \restriction_{\mathcal{Z}_r} \iff A_r \boldsymbol{x}_C = A_r \boldsymbol{x}_D \iff \boldsymbol{x}_C - \boldsymbol{x}_D \in \ker(A_r),$$

which proves part (ii).

For part (iii), assume that C is r-faithful and $|C| = k$. Suppose first that $\ker(A_r) \cap \mathcal{D}_C(k) = \{\mathbf{0}\}$. Let D be r-faithful with $|D| \le k$. Part (i) gives $A_r \boldsymbol{x}_D = A_r \mathbf{1} = A_r \boldsymbol{x}_C$, and therefore $\boldsymbol{x}_D - \boldsymbol{x}_C \in \ker(A_r)$. Since $\|\boldsymbol{x}_D\|_0 = |D| \le k$, the definition of $\mathcal{D}_C(k)$ gives $\boldsymbol{x}_D - \boldsymbol{x}_C \in \mathcal{D}_C(k)$. The assumed intersection yields $\boldsymbol{x}_D - \boldsymbol{x}_C = \mathbf{0}$, so $D = C$. No faithful circuit has cardinality below k, and C is the unique faithful circuit of cardinality k.

Conversely, assume that C is the unique minimum-cardinality faithful circuit. Let $\boldsymbol{h} \in \ker(A_r) \cap \mathcal{D}_C(k)$. By the definition of $\mathcal{D}_C(k)$, there is $\boldsymbol{x} \in \{0,1\}^n$ such that

$$\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}_C \quad \text{and} \quad \|\boldsymbol{x}\|_0 \le k.$$

Let $D \subseteq [n]$ have indicator $\boldsymbol{x}_D = \boldsymbol{x}$. Since $\boldsymbol{h} \in \ker(A_r)$,

$$A_r \boldsymbol{x}_D = A_r(\boldsymbol{x}_C + \boldsymbol{h}) = A_r \boldsymbol{x}_C = A_r \mathbf{1}.$$

Part (i) makes D faithful. We also have $|D| = \|\boldsymbol{x}\|_0 \le k = |C|$. Unique minimum-cardinality status forces $D = C$, and thus $\boldsymbol{h} = \mathbf{0}$. Therefore, $\ker(A_r) \cap \mathcal{D}_C(k) = \{\mathbf{0}\}$. □

C.2 Proof of Proposition 9

Proof. For $C \subseteq [n]$, let $Z_C = [n] \setminus C$. Theorem 8(i) gives

$$\begin{aligned} C \text{ is } r\text{-faithful} &\iff A_r \boldsymbol{x}_C = A_r \mathbf{1} \\ &\iff A_r(\mathbf{1} - \boldsymbol{x}_C) = 0 \\ &\iff A_r \boldsymbol{x}_{Z_C} = 0 \\ &\iff Z_C \in \mathcal{H}_r^0. \end{aligned}$$

Since $|C| = n - |Z_C|$, minimizing $|C|$ over faithful circuits is equivalent to maximizing $|Z_C|$ over $\mathcal{H}_r^0$. The maximum is q_r^* by the definitions of $\mathcal{H}_r^0$ and q_r^*. Thus, $\min\{|C| \mid C \text{ is } r\text{-faithful}\} = n - q_r^*$, which proves the minimum-cardinality formula in Proposition 9. The map $C \mapsto Z_C$ is a bijection, so its minimum-cardinality circuits correspond exactly to the members of $\mathcal{H}_r^0$ having size q_r^*. The bijection proves the minimum-family formula in Proposition 9.

For inclusion minimality, let C be faithful and $Z_C \in \mathcal{H}_r^0$. A proper faithful subset $D \subsetneq C$ exists if and only if

$$Z_C = [n] \setminus C \subsetneq [n] \setminus D = Z_D \quad \text{and} \quad Z_D \in \mathcal{H}_r^0.$$

Thus, C is inclusion-minimal exactly when Z_C is inclusion-maximal in $\mathcal{H}_r^0$. □

C.3 Proof of Theorem 10

Proof. Let $p = n - 1$. The inequality $p \leq \binom{m}{r+1}$ permits pairwise distinct sets

$$H_1, \ldots, H_p \in \{H \subseteq [m] \mid |H| = r + 1\}. \tag{25}$$

For $z \in [0,1]^m$, let $\ell(z) = 1$, and, for $j \in [p]$, let

$$h_j(z) = \mathrm{ReLU}\left(\sum_{i \in H_j} z_i - r \right). \tag{26}$$

For $j \in [p]$, the first p component responses are

$$f_j = \ell + h_j \qquad (j \in [p]) \tag{27}$$

and component n has response

$$f_n = -(p-k)\ell - \sum_{j=1}^{p} h_j. \tag{28}$$

Equations 27 and 28 give

$$F_{\mathsf{Full}} = \sum_{j=1}^{p} f_j + f_n = p\ell + \sum_{j=1}^{p} h_j - (p-k)\ell - \sum_{j=1}^{p} h_j = k\ell. \tag{29}$$

Faithful circuits. Let $D \subseteq [n]$. For $j \in [p]$, let $\alpha_j \in \{0,1\}$ indicate membership of j in D, and let $\gamma \in \{0,1\}$ indicate membership of n. Let $a = \sum_{j=1}^{p} \alpha_j$. Equations 27 and 28 yield

$$F_D = (a - \gamma(p-k))\,\ell + \sum_{j=1}^{p} (\alpha_j - \gamma) h_j. \tag{30}$$

For $z \in \mathcal{Z}_r$ and $j \in [p]$,

$$\sum_{i \in H_j} z_i \leq |\{i \mid z_i \neq 0\}| = \|z\|_0 \leq r \Longrightarrow h_j(z) = 0. \tag{31}$$

Since $\ell(z) = 1$, equations 29– 31 give

$$D \text{ is } r\text{-faithful} \iff a - \gamma(p-k) = k. \tag{32}$$

The two values of γ give the exhaustive cases

γ	a	selected components in $[p]$	$\lvert D\rvert$
0	k	any k of p	k
1	p	all p	n.

For $\gamma = 0$, equation 32 gives $a = k$. For $\gamma = 1$, it gives $a = k + (p - k) = p$, which gives $\alpha_j = 1$ for every $j \in [p]$. These cases exhaust $\gamma \in \{0, 1\}$. Every faithful circuit is therefore a k-subset of $[p]$ or the full component set. The minimum cardinality is k, and $\binom{p}{k}$ circuits attain it. Removing one component from a minimum circuit changes $a = k$ to $a = k - 1$, violating equation 32. Every minimum circuit is inclusion-minimal. Parts (i) and (ii) follow.

Order-$(r+1)$ separation. Let $C, D \in \mathsf{MinFaith}_r$ be distinct. Both are k-subsets of $[p]$. Choose $i \in C \setminus D$ after interchanging C, D if required. Equations 27 and 26 give

$$F_C - F_D = \sum_{j \in C \setminus D} h_j - \sum_{j \in D \setminus C} h_j. \tag{33}$$

At $z = \mathbf{1}_{H_i}$, pairwise distinctness and equal cardinality in equation 25 give

$$h_i(\mathbf{1}_{H_i}) = 1 \quad \text{and} \quad h_j(\mathbf{1}_{H_i}) = \mathrm{ReLU}(|H_j \cap H_i| - r) = 0 \qquad (j \neq i), \tag{34}$$

because $|H_j \cap H_i| \leq r$ for $j \neq i$. Equations 33 and 34 yield

$$F_C(\mathbf{1}_{H_i}) - F_D(\mathbf{1}_{H_i}) = 1. \tag{35}$$

Because $|H_i| = r + 1$, the order-$(r+1)$ signatures differ. Part (iii) follows.

Order-$(r+1)$ faithfulness. Let $D \subsetneq [n]$, and use a, γ, and $(\alpha_j)_{j \in [p]}$ from equation 30. We prove

$$\|F_D - F_{\mathsf{Full}}\|_{\infty, \mathcal{Z}_{r+1}} \geq 1.$$

The cases $\gamma = 0$ and $\gamma = 1$ are exhaustive.

If $\gamma = 0$ and $a \neq k$, equations 29, 30, and 31 give, for every $z \in \mathcal{Z}_r$,

$$|F_D(z) - F_{\mathsf{Full}}(z)| = |a - k| \geq 1.$$

If $\gamma = 0$ and $a = k$, then $k \geq 1$ gives an index i with $\alpha_i = 1$. Equations 30 and 34 yield

$$F_D(\mathbf{1}_{H_i}) = k + 1 \quad \text{and} \quad F_{\mathsf{Full}}(\mathbf{1}_{H_i}) = k.$$

Thus, the required inequality holds.

Assume $\gamma = 1$. Since $D \subsetneq [n]$, we have $a < p$. For every $z \in \mathcal{Z}_r$, equations 29, 30, and 31 give

$$F_{\mathsf{Full}}(z) - F_D(z) = k - \big(a - (p - k)\big) = p - a \geq 1.$$

The required inequality holds in every case. A k-subset of $[p]$ has order-$(r+1)$ error at least one by the preceding witness. For the reverse inequality, let $z \in \mathcal{Z}_{r+1}$. If $h_j(z) > 0$, then

$$\sum_{i \in H_j} z_i > r.$$

Because $0 \leq z_i \leq 1$, all $r + 1$ coordinates in H_j are nonzero. The support bound on z then gives $\mathrm{supp}(z) = H_j$. Distinctness of the sets $H_1, \ldots, H_p$ implies that at most one order-$(r+1)$ ReLU unit is positive, and every such feature is at most one. For a k-subset $C \subseteq [p]$, we therefore have

$$\|F_C - F_{\mathsf{Full}}\|_{\infty, \mathcal{Z}_{r+1}} = \sup_{z \in \mathcal{Z}_{r+1}} \sum_{j \in C} h_j(z) \leq 1.$$

The witness attains one, so the minimum in part (iv) of Theorem 10 equals one. Part (iv) follows.

Patched ReLU realization. Let the corrupted input produce $u_i^0 = 0$ and the clean input produce $u_i^1 = 1$ at site i. Patching with $z_i \in [0, 1]$ supplies the site value $u_i(z_i) = z_i$. One ReLU layer computes every h_j in equation 26, and a linear readout supplies $\ell = 1$. One maskable additive output group implements each $f_j = \ell + h_j$, and component n implements $f_n = -(p - k)\ell - \sum_j h_j$. The ReLU units use at most

$$p(r + 2)$$

nonzero weights and biases. The $p + 1$ output groups use at most $3p + 1$ nonzero weights. Hence, the total number of nonzero parameters is at most

$$p(r + 2) + 3p + 1 = O\big(p(r + 1)\big).$$

The patched interface uses $O(m)$ additional parameters, giving $O(m + p(r + 1))$ in total. □

C.4 Proof of Theorem 11

Lemma 22. *Let $q, n \in \mathbb{N}$ satisfy $1 \le q \le n/2$. If $\mathcal{A} \subseteq \{Z \subseteq [n] \mid |Z| = q\}$ and $Z \cap Z' \neq \varnothing$ for all $Z, Z' \in \mathcal{A}$, then*

$$|\mathcal{A}| \le \binom{n-1}{q-1}. \tag{36}$$

Proof. We give the cyclic-permutation derivation. A cyclic order of $[n]$ has $(n-1)!$ realizations. Consider the n cyclic intervals of length q in one cyclic order. Any pairwise-intersecting subfamily of these intervals has at most q members. The bound holds for the empty subfamily. For a nonempty subfamily, choose one member and label it $I_0 = \{0, \dots, q-1\}$ in cyclic coordinates. Every length-q interval intersecting I_0 starts at one of

$$-(q-1), \dots, -1, 0, 1, \dots, q-1.$$

For every $i \in \{1, \dots, q-1\}$, the intervals starting at i and $i-q$ are disjoint because $n \ge 2q$. A pairwise-intersecting family contains I_0 and at most one interval from each of these $q-1$ pairs, which proves the claim.

Count pairs (Z, π) for which $Z \in \mathcal{A}$ occurs as a length-q interval in the cyclic order π. For a prescribed Z, contracting the consecutive block gives $q!(n-q)!$ cyclic orders: $q!$ internal orders and $(n-q)!$ cyclic orders of the contracted block and the remaining $n-q$ elements. The interval claim gives

$$|\mathcal{A}|\, q!(n-q)! \le q(n-1)!.$$

Division by $q!(n-q)!$ yields

$$|\mathcal{A}| \le \frac{q(n-1)!}{q!(n-q)!} = \binom{n-1}{q-1},$$

which proves equation 36. □

Lemma 23. *Let $w_1, \dots, w_N$ be nonzero elements of a real vector space, and let $q \in \mathbb{N}$ satisfy $N/2 < q < N$. At most $\binom{N-1}{q-1}$ labeled q-subsets have vector sum zero.*

Proof. We use the following consequence of Móricz & Nagy (2026, Theorem 1.7 and Lemmas 2.1–2.2): if $a_1, \dots, a_N$ are nonzero integers and $N/2 < q < N$, then at most $\binom{N-1}{q-1}$ labeled q-subsets $T \subseteq [N]$ satisfy $\sum_{j \in T} a_j = 0$.

Let W be the finite-dimensional $\mathbb{Q}$-linear span of the w_j. The set of nonzero subset sums

$$\mathcal{S} = \left\{ \sum_{j \in T} w_j \mid T \subseteq [N],\ \sum_{j \in T} w_j \neq 0 \right\}$$

is finite. Let $W^* = \mathrm{Hom}_{\mathbb{Q}}(W, \mathbb{Q})$. For each $s \in \mathcal{S}$, the condition $\phi(s) = 0$ defines a proper $\mathbb{Q}$-linear hyperplane in W^*: extend s to a basis of W, map s to 1, and map every other basis vector to 0. The resulting functional lies outside the hyperplane. A finite union of proper subspaces cannot cover a finite-dimensional vector space over the infinite field $\mathbb{Q}$, so there exists $\phi \in W^*$ such that $\phi(s) \neq 0$ for every $s \in \mathcal{S}$. After multiplication by a common denominator $D \in \mathbb{N}$, let $a_j = D\phi(w_j) \in \mathbb{Z}$. Each singleton sum w_j belongs to $\mathcal{S}$, so $a_j \neq 0$. For every $T \subseteq [N]$, linearity gives

$$\sum_{j \in T} w_j = 0 \implies \sum_{j \in T} a_j = D\phi\left(\sum_{j \in T} w_j \right) = 0.$$

Conversely, if $\sum_{j \in T} a_j = 0$ and $s_T = \sum_{j \in T} w_j \neq 0$, then $s_T \in \mathcal{S}$, while $0 = \sum_{j \in T} a_j = D\phi(s_T)$ contradicts $\phi(s_T) \neq 0$. Therefore,

$$\sum_{j \in T} w_j = 0 \quad \text{if and only if} \quad \sum_{j \in T} a_j = 0.$$

Applying the imported integer theorem to $a_1, \dots, a_N$, whose entries are nonzero and satisfy $N/2 < q < N$, and using the zero-pattern equivalence transfers the bound to the vectors. □

Proof of Theorem 11. For a circuit C, let $Z_C = [n] \setminus C$. Additivity and faithfulness give

$$\sum_{j \in Z_C} \boldsymbol{v}_j^{(r)} = A_r(\mathbf{1} - \boldsymbol{x}_C) = A_r\mathbf{1} - A_r\boldsymbol{x}_C = 0.$$

Let $Z_0 = \{j \mid \boldsymbol{v}_j^{(r)} = 0\}$, write $z = |Z_0|$, and put $N = n - z$. Every maximum-cardinality zero-sum set contains Z_0: otherwise one could add a missing zero signature. Proposition 9 therefore gives a bijection between $\mathsf{MinFaith}_r$ and the zero-sum q-subsets of the N nonzero signatures, where

$$q = (n - k) - z = N - k.$$

Because $k \geq 1$ and q is a cardinality, we have $0 \leq q < N$. No zero-sum subset T of the nonzero signatures has $|T| > q$: otherwise, $T \cup Z_0$ would be a zero-sum omission of size $|T| + z > q + z = n - k = q_r^*$, contradicting the definition of q_r^*. If $q = 0$, the family has one member and the asserted bound holds because $k = N \leq n - 1$, so $1 \leq \binom{n-1}{k}$.

Assume henceforth that $q \geq 1$. If $q \leq N/2$, two disjoint zero-sum q-sets would have a zero-sum union of size $2q > q$, contradicting maximality. Hence, these sets form a pairwise-intersecting q-uniform family. Lemma 22 gives

$$|\mathsf{MinFaith}_r| \leq \binom{N-1}{q-1} = \binom{N-1}{k}.$$

If instead $q > N/2$, the inequalities $q < N$ and $q > N/2$ verify all assumptions of Lemma 23, which gives the same bound. Since $N \leq n$ and $q \geq 1$ implies $k = N - q \leq N - 1$,

$$|\mathsf{MinFaith}_r| \leq \binom{N-1}{k} \leq \binom{n-1}{k}.$$

Theorem 10 constructs $\binom{n-1}{k}$ minimum circuits under the stated condition on m, so the upper bound is attained. □

C.5 Proof of Corollary 12

Proof. Theorems 11 and 10 give the exact maximum $\binom{n-1}{k}$ for every $1 \leq k < n$. Let $k = \lceil n/2 \rceil$. For $n = 2s$,

$$\binom{n-1}{k} = \binom{2s-1}{s} = \frac{1}{2}\binom{2s}{s}.$$

For $n = 2s + 1$,

$$\binom{n-1}{k} = \binom{2s}{s+1} = \frac{s}{s+1}\binom{2s}{s}.$$

Stirling's formula gives $\binom{2s}{s} = \Theta(4^s/\sqrt{s})$. In both parity cases,

$$\binom{n-1}{\lceil n/2 \rceil} = \Theta\left(\frac{2^n}{\sqrt{n}}\right).$$

Taking base-two logarithms proves the asymptotic formula in Corollary 12. □

C.6 Proof of Theorem 14

Proof. Choose distinct $(r+1)$-subsets $H_1, \ldots, H_p \subseteq [m]$, and define

$$h_j(z) = \mathrm{ReLU}\left(\sum_{i \in H_j} z_i - r\right).$$

Write $p = dq + t$ with $d \in \mathbb{N}$ and $0 \leq t < q$. Since $q < p \leq Lq$, we have $d \geq 1$, and either $d < L$ or $(d, t) = (L, 0)$. Partition $[p]$ into t blocks of size $d + 1$ and $q - t$ blocks of size d. All q blocks are nonempty and have size at most L; denote them by $B_1, \ldots, B_q$. The components indexed by $[p]$ have responses $f_j = 1 + h_j$, and the local cancellation components are

$$g_s = -1 - \sum_{j \in B_s} h_j \qquad (s \in [q]).$$

Since the blocks partition $[p]$, the full response is

$$\sum_{j=1}^{p} f_j + \sum_{s=1}^{q} g_s = (p-q) + \sum_{j=1}^{p} h_j - \sum_{s=1}^{q} \sum_{j \in B_s} h_j = p - q = k.$$

Every h_j vanishes on $\mathcal{Z}_r$. Consider a circuit selecting a components f_j and b local cancellation components. The circuit response on $\mathcal{Z}_r$ is the constant $a-b$, so it is faithful exactly when $a-b=k$. The circuit cardinality is then

$$a + b = k + 2b \geq k,$$

with equality exactly when $b=0$ and $a=k$. Hence, the minimum circuits are precisely the k-subsets of $[p]$, and

$$|\mathsf{MinFaith}_r| = \binom{p}{k} = \binom{p}{q}.$$

For distinct minimum circuits C, D, choose $j \in C \setminus D$ after interchanging them if necessary. At $z = \mathbf{1}_{H_j}$, we have $h_j(z) = 1$. For $i \neq j$, distinctness and equal cardinality of H_i, H_j give $|H_i \cap H_j| \leq r$, hence $h_i(z) = 0$. Therefore, $F_C(z) - F_D(z) = 1$, proving order-$(r+1)$ separation.

Each component f_j depends on one order-$(r+1)$ unit, and each local cancellation component depends on at most L such units. For $z \in \mathcal{Z}_{r+1}$, $0 \leq h_j(z) \leq 1$; hence

$$|f_j(z)| = 1 + h_j(z) \leq 2 \quad \text{and} \quad |g_s(z)| \leq 1 + \sum_{j \in B_s} h_j(z) \leq 1 + |B_s| \leq L + 1.$$

The hidden units use $O(p(r+1))$ parameters. Because the blocks partition $[p]$, $\sum_{s=1}^{q} |B_s| = p$. The f_j and local cancellation components therefore use $O(p + q + \sum_{s=1}^{q} |B_s|) = O(p)$ parameters because $q < p$, and the patched interface uses $O(m)$. The derivation proves the claimed local and total complexity bounds. For $p = 2q$, Stirling's formula gives $\binom{2q}{q} = 2^{2q - \frac{1}{2}\log_2 q + O(1)}$, which is the stated expression because $n = 3q$. □

C.7 Proof of Corollary 15

Proof. Let $\alpha \in (0,1)$ and $\lambda = (1+\alpha)/(1-\alpha) > 1$. For all sufficiently large integers t, let $q_t = t$ and $p_t = \lfloor \lambda t \rfloor$. Then $q_t < p_t$, and $p_t \leq L_\alpha q_t$ for $L_\alpha = \lceil \lambda \rceil + 1$. Let $n_t = p_t + q_t$ and $k_t = p_t - q_t$, as in Theorem 14. Also,

$$\frac{k_t}{n_t} = \frac{p_t - q_t}{p_t + q_t} \longrightarrow \alpha.$$

For each t, take $r = 0$ and $m_t = p_t$. Then $p_t = \binom{m_t}{r+1}$, so Theorem 14 provides the required model. Because $p_t = \lambda t + O(1)$, we have

$$\frac{p_t}{n_t} = \frac{1+\alpha}{2} + O(t^{-1}) \quad \text{and} \quad \frac{q_t}{p_t} = \frac{1-\alpha}{1+\alpha} + O(t^{-1}).$$

The binary entropy is continuously differentiable in a neighborhood of $\theta_\alpha = (1-\alpha)/(1+\alpha)$. The mean-value theorem and $q_t/p_t = \theta_\alpha + O(t^{-1})$ give

$$H_2(q_t/p_t) = H_2(\theta_\alpha) + O(t^{-1}).$$

Using $p_t/n_t = (1+\alpha)/2 + O(t^{-1})$, we obtain

$$p_t H_2(q_t/p_t) = c_\alpha n_t + O(1).$$

Stirling's formula now gives

$$\log_2 \binom{p_t}{q_t} = p_t H_2(q_t/p_t) - \tfrac{1}{2}\log_2 p_t + O(1) = c_\alpha n_t - \tfrac{1}{2}\log_2 n_t + O(1),$$

where $\log_2 p_t = \log_2 n_t + O(1)$ because $p_t/n_t \to (1+\alpha)/2 > 0$. Finally, $(1-\alpha)/(1+\alpha) \in (0,1)$, thus $c_\alpha > 0$. □

C.8 Proof of Corollary 16

Proof. Let each perturbed response differ from the corresponding unperturbed response by at most δ on $\mathcal{Z}_{r+1}$. A minimum candidate omits q components f_j and all q local cancellation components. Their unperturbed sum is zero on $\mathcal{Z}_r$. For each component j, let Δ_j denote its response perturbation. The resulting error satisfies

$$\sup_{z\in\mathcal{Z}_r}\left|\sum_{j\notin C}\Delta_j(z)\right| \le \sum_{j\notin C}\sup_{z\in\mathcal{Z}_r}|\Delta_j(z)| \le 2q\delta \le \varepsilon.$$

Let a circuit D select a components f_j and b local cancellation components. On $\mathcal{Z}_r$, the unperturbed full-minus-circuit response is $k-a+b$. If $|D| = a+b < k$, then

$$k-a+b = k-(a+b)+2b \ge 1.$$

For every $z\in\mathcal{Z}_r$, the reverse triangle inequality gives

$$|\widetilde{F}_{\mathsf{Full}}(z) - \widetilde{F}_D(z)| \ge 1-(n-|D|)\delta \ge 1-n\delta > \varepsilon.$$

Thus, no smaller circuit is (r,ε)-faithful. If $|D| = k$ and $b \ge 1$, the unperturbed gap is $2b \ge 2$. For every $z\in\mathcal{Z}_r$, the reverse triangle inequality gives

$$|\widetilde{F}_{\mathsf{Full}}(z) - \widetilde{F}_D(z)| \ge 2b-(n-k)\delta \ge 2-n\delta > 1+\varepsilon > \varepsilon,$$

so the same-size unintended circuit is excluded. Thus, the intended k-subsets of $[p]$ are exactly the minimum faithful circuits.

For two such circuits, the unperturbed witness from the theorem has response difference one. At its witness z,

$$|\widetilde{F}_C(z) - \widetilde{F}_D(z)| \ge |F_C(z)-F_D(z)| - \sum_{j\in C\triangle D}|\Delta_j(z)| \ge 1-2k\delta > 0.$$

□

C.9 Proof of Theorem 17

Proof. Let $\Delta_j = \widetilde{f}_j - f_j$. The perturbation assumption in Theorem 17 gives

$$\|\Delta_j\|_{\infty,\mathcal{Z}_{r+1}} \le \delta \qquad (j\in[n]). \tag{37}$$

The k-subsets of $[p]$. Let $C\subseteq[p]$ satisfy $|C| = k$. The omitted components are $([p]\setminus C)\cup\{n\}$, whose cardinality is $p-k+1$. Their unperturbed sum vanishes on $\mathcal{Z}_r$ by equations 29 and 31. Thus, for $z\in\mathcal{Z}_r$,

$$\begin{aligned}|\widetilde{F}_{\mathsf{Full}}(z) - \widetilde{F}_C(z)| &= |\sum_{j\notin C}\big(f_j(z)+\Delta_j(z)\big)| \\ &= |\sum_{j\notin C}\Delta_j(z)| \le (p-k+1)\delta \le \varepsilon.\end{aligned}$$

Every such C is (r,ε)-faithful.

Circuits of size below k. Let $D\subseteq[n]$ satisfy $|D| < k$. Use a and γ from equation 30. On $\mathcal{Z}_r$,

$$F_{\mathsf{Full}} - F_D = k-a+\gamma(p-k). \tag{38}$$

If $\gamma = 0$, then $a = |D| \le k-1$, so the right-hand side is at least 1. If $\gamma = 1$, then $a+1 = |D| \le k-1$, thus

$$k-a+(p-k) = p-a \ge p-k+2 \ge 1.$$

For every $z\in\mathcal{Z}_r$, equation 37 gives

$$|(\widetilde{F}_{\mathsf{Full}} - \widetilde{F}_D) - (F_{\mathsf{Full}} - F_D)| \le (n-|D|)\delta \le (p+1)\delta.$$

Equations 3 and 38 yield

$$|\widetilde{F}_{\mathsf{Full}}(z) - \widetilde{F}_D(z)| \geq 1 - (p+1)\delta > \varepsilon.$$

No circuit of size below k is (r, ε)-faithful.

Size-k circuits containing component n. Such a circuit has $\gamma = 1$ and $a = k - 1$. Equation 38 gives

$$F_{\mathsf{Full}} - F_D = p - k + 1 \geq 1 \qquad \text{on} \quad \mathcal{Z}_r.$$

For every $z \in \mathcal{Z}_r$, equations 37 and 3 give

$$|\widetilde{F}_{\mathsf{Full}}(z) - \widetilde{F}_D(z)| \geq p - k + 1 - (n-k)\delta \geq 1 - (p+1)\delta > \varepsilon.$$

Therefore, the minimum-cardinality (r, ε)-faithful circuits are exactly the $\binom{p}{k}$ k-subsets of $[p]$.

Separation. Let $C, D \subseteq [p]$ be two distinct k-subsets. Choose $i \in C \setminus D$. Equation 35 gives

$$|F_C(\mathbf{1}_{H_i}) - F_D(\mathbf{1}_{H_i})| = 1.$$

The perturbed difference satisfies

$$\begin{aligned} |\widetilde{F}_C(\mathbf{1}_{H_i}) - \widetilde{F}_D(\mathbf{1}_{H_i})| &\geq 1 - \sum_{j \in C \triangle D} |\Delta_j(\mathbf{1}_{H_i})| \\ &\geq 1 - |C \triangle D|\delta \geq 1 - 2k\delta > 0. \end{aligned}$$

Taking the supremum over $\mathcal{Z}_{r+1}$ proves the separation bound in Theorem 17.

Order-$(r+1)$ faithfulness. Let $D \subsetneq [n]$. Theorem 10(iv) gives a point $z_D \in \mathcal{Z}_{r+1}$ satisfying

$$|F_{\mathsf{Full}}(z_D) - F_D(z_D)| \geq 1.$$

Only omitted components enter the full-minus-circuit response. Equation 37 therefore gives

$$\begin{aligned} |\widetilde{F}_{\mathsf{Full}}(z_D) - \widetilde{F}_D(z_D)| &\geq |F_{\mathsf{Full}}(z_D) - F_D(z_D)| - \sum_{j \notin D} |\Delta_j(z_D)| \\ &\geq 1 - (n - |D|)\delta \\ &\geq 1 - (p+1)\delta > \varepsilon. \end{aligned}$$

No proper circuit is $(r+1, \varepsilon)$-faithful, while the full component set has zero faithfulness error. The asserted uniqueness follows.

For the parameter-neighborhood statement, let θ_0 be the parameter vector of the constructed response mechanism. Let $K = \{\theta : \|\theta - \theta_0\| \leq 1\}$. The joint map $(\theta, z) \mapsto (f_{1,\theta}(z), \ldots, f_{n,\theta}(z))$ is continuous because each coordinate is a finite composition of affine maps and the continuous ReLU map. For a closed ball K centered at θ_0, $K \times \mathcal{Z}_{r+1}$ is compact. Uniform continuity on this product implies that, for every $\delta > 0$, some $0 < \xi \leq 1$ satisfies

$$\|\theta - \theta_0\| < \xi \implies \max_{j \in [n]} \sup_{z \in \mathcal{Z}_{r+1}} |f_{j,\theta}(z) - f_{j,\theta_0}(z)| < \delta.$$

Equivalently, the map

$$\theta \longmapsto \left(f_{1,\theta} \restriction_{\mathcal{Z}_{r+1}}, \ldots, f_{n,\theta} \restriction_{\mathcal{Z}_{r+1}}\right)$$

is continuous in the product uniform norm at θ_0. For any δ satisfying the strict versions of equation 3, continuity gives an open neighborhood of θ_0 on which the perturbation assumption in Theorem 17 holds. The preceding derivation applies throughout that neighborhood. □

C.10 Proof of Proposition 18

Proof. We prove by induction on the number $s \geq 0$ of summands that a zero sum of s elements of $\mathcal{P}$ has only zero summands. The claim is vacuous for $s = 0$. Let $s \geq 1$, assume the claim for $s - 1$, and let $w_1, \ldots, w_s \in \mathcal{P}$ satisfy $\sum_{i=1}^{s} w_i = 0$. Closure of the convex cone under finite sums gives

$$w_1 \in \mathcal{P} \quad \text{and} \quad -w_1 = \sum_{i=2}^{s} w_i \in \mathcal{P}.$$

Thus, $w_1 \in \mathcal{P} \cap (-\mathcal{P}) = \{0\}$. Substitution gives $\sum_{i=2}^{s} w_i = 0$, and the induction hypothesis applied to the remaining $s-1$ summands gives $w_2 = \cdots = w_s = 0$. Therefore,

$$w_1 = \cdots = w_s = 0. \tag{39}$$

Let $C \subseteq [n]$. Theorem 8(i) and the definition of A_r give

$$C \text{ is } r\text{-faithful} \iff \sum_{j \notin C} \boldsymbol{v}_j^{(r)} = 0.$$

Every omitted signature belongs to $\mathcal{P}$. Equation 39 therefore gives

$$C \text{ is } r\text{-faithful} \iff \boldsymbol{v}_j^{(r)} = 0 \text{ for every } j \notin C \iff C_+ \subseteq C.$$

The faithful circuits are exactly the supersets of C_+. Their unique minimum-cardinality member is C_+, proving the stated formula for C_+ in Proposition 18. □

D Proofs for Sections 6 and 7

We establish the nonadditive transfer bounds before verifying the finite-domain optimization certificate.

D.1 Proof of Theorem 19

Proof. For a circuit C, let

$$G_C(z) = Y_0(z, \boldsymbol{x}_C) - Y_0(z, \mathbf{1}) \quad \text{and} \quad E_C(z) = R(z, \boldsymbol{x}_C) - R(z, \mathbf{1}).$$

The actual circuit-to-full difference is $G_C + E_C$, and $\|E_C\|_{\infty, \mathcal{Z}_s} \le 2\eta_s$. The reverse triangle inequality gives

$$|\|G_C + E_C\|_{\infty, \mathcal{Z}_s} - \|G_C\|_{\infty, \mathcal{Z}_s}| \le \|E_C\|_{\infty, \mathcal{Z}_s} \le 2\eta_s,$$

which proves the faithfulness-error bound. Replacing the full mask by $\boldsymbol{x}_D$, let

$$G_{C,D}(z) = Y_0(z, \boldsymbol{x}_C) - Y_0(z, \boldsymbol{x}_D) \quad \text{and} \quad E_{C,D}(z) = R(z, \boldsymbol{x}_C) - R(z, \boldsymbol{x}_D).$$

Then $\|E_{C,D}\|_{\infty, \mathcal{Z}_s} \le 2\eta_s$, and the same reverse triangle inequality gives

$$|d_s(C, D) - d_s^0(C, D)| \le \|E_{C,D}\|_{\infty, \mathcal{Z}_s} \le 2\eta_s.$$

Every $C \in \mathcal{A}$ has actual error at most $\alpha + 2\eta_r \le \varepsilon$. Every $D \notin \mathcal{A}$ with $|D| \le k$ has actual error at least $\beta - 2\eta_r > \varepsilon$. Hence, no circuit smaller than k is faithful and the faithful circuits of size k are exactly $\mathcal{A}$. The pairwise conclusion follows from $d_s(C, D) \ge d_s^0(C, D) - 2\eta_s \ge \gamma - 2\eta_s$.

For the sharp surrogate, Theorem 10(iv) gives $e_{r+1}^0(D) \ge 1$ for every $D \subsetneq [n]$. The first inequality of the theorem yields

$$e_{r+1}(D) \ge 1 - 2\eta_{r+1} > \varepsilon$$

under equation 4. The full mask has zero actual faithfulness error. Thus, it is the unique $(r+1, \varepsilon)$-faithful circuit. □

D.2 Proof of Proposition 20

Proof. For a binary mask $\boldsymbol{x}$, the two linear inequalities are equivalent to $\|B\boldsymbol{x} - B\mathbf{1}_n\|_\infty \le \varepsilon$, exactly the finite-$Q$ faithfulness condition for the additive surrogate Y_0. The objective therefore minimizes circuit cardinality over masks faithful to Y_0 on Q. After fixing the optimum cardinality, the no-good left-hand side satisfies

$$\sum_{j: x_j^* = 1} (1 - x_j) + \sum_{j: x_j^* = 0} x_j = \sum_{j=1}^{n} |x_j - x_j^*| = \|\boldsymbol{x} - \boldsymbol{x}^*\|_1.$$

For binary masks, the quantity is zero exactly when $\boldsymbol{x} = \boldsymbol{x}^*$, so the no-good inequality excludes exactly $\boldsymbol{x}^*$. The second program is therefore feasible precisely when a distinct optimum mask exists. Adding one no-good inequality for each returned optimum enumerates all and only the minimum-cardinality masks faithful to Y_0 on Q. There are finitely many binary masks. Each iteration excludes the returned optimum and no unreturned mask; hence, the procedure terminates, and final infeasibility holds exactly when every optimum has been returned. The constraints contain no remainder term R, so nonadditive conclusions require the finite-domain transfer stated after Proposition 20. □

E PROOF OF THE INTERVENTION-QUERY THEOREM

E.1 PROOF OF THEOREM 13

Proof. Let $z \in \mathcal{Z}_{r+1}$. If $h_j(z) > 0$, then

$$\sum_{i \in H_j} z_i > r. \tag{40}$$

Since $0 \le z_i \le 1$, equation 40 requires at least $r+1$ nonzero coordinates in H_j. The bound $\|z\|_0 \le r+1$ and $|H_j| = r+1$ therefore give

$$\{i \in [m] \mid z_i \neq 0\} = H_j.$$

The sets $H_1, \dots, H_p$ are pairwise distinct, so at most one $h_j(z)$ is positive. For a candidate $C \subseteq [p]$ with $|C| = k$,

$$F_C(z) = k + \sum_{j \in C} h_j(z). \tag{41}$$

If every $h_j(z)$ vanishes, equation 41 returns $F_C(z) = k$ for every candidate. Otherwise, there is a unique $j \in [p]$ with $\lambda = h_j(z) > 0$, and

$$F_C(z) = k + \lambda \mathbf{1}[j \in C].$$

The query z determines j and λ; hence, its response reveals exactly the membership bit $\mathbf{1}[j \in C]$. Therefore, each admitted query reveals at most one adaptive membership bit of the unknown k-subset of $[p]$.

We prove the membership-query lower bound by an adversary. After some queries, let u be the number of unqueried indices and let s be the number of selected indices among them required by the answers and the known cardinality k. The consistent candidates correspond bijectively to the s-subsets of the u unqueried indices and therefore number $\binom{u}{s}$. Initially,

$$u = p, \quad s = k \quad \text{and} \quad 1 \le s \le u - 1.$$

Assume $u \ge 2$ and $1 \le s \le u-1$. When a new index is queried, the adversary answers as follows:

condition	answer	(u', s')
$s \le u-2$	0	$(u-1, s)$
$s = u-1$	1	$(u-1, s-1)$.

The conditions in the two rows are exhaustive because $1 \le s \le u-1$. If $u \ge 3$ and $s \le u-2$, then

$$1 \le s' = s \le u - 2 = u' - 1.$$

If $u \ge 3$ and $s = u-1$, then

$$1 \le s' = s - 1 = u - 2 = u' - 1.$$

Thus, after $t \le p-2$ informative queries, $u = p - t \ge 2$ and $1 \le s \le u-1$, thus $\binom{u}{s} \ge \binom{u}{1} = u \ge 2$. At least two candidates remain, and identification requires at least $p-1$ informative queries. Queries that reveal no membership bit preserve the candidate family, so the worst-case lower bound is $p-1$.

For the upper bound, query $\mathbf{1}_{H_1}, \dots, \mathbf{1}_{H_{p-1}}$. Equations 34 and 41 give

$$F_C(\mathbf{1}_{H_j}) = \begin{cases} k+1, & j \in C, \\ k, & j \notin C. \end{cases}$$

These responses determine membership of the first $p-1$ indices. The known cardinality k determines membership of index p. The nonadaptive upper bound is therefore $p - 1 = n - 2$. □