SHARP CLOSURE THRESHOLDS FOR TWO-HIDDEN-LAYER RELU NETWORKS

SHUNQI LU

Abstract. We determine the minimum first hidden width needed to approximate the coordinate maximum by a fixed two-hidden-layer ReLU architecture with unrestricted parameters in L_2. In dimension $d \geq 14$, the threshold is $\kappa_d = \binom{d}{2} - \lfloor d^2/4 \rfloor$. The same value holds for ordered L-statistics whose weights have a nonzero discrete second difference. A fan-localization theorem assembles cellwise ridge formulas across a finite hyperplane arrangement, using a fixed finite second layer and output map. A complementary pair-transversal bound turns stable triple junctions into first-layer lower bounds. Its quantitative form is uniform over all weights, biases, and coefficients. The threshold proof is analytic for $d \geq 40$. Smaller dimensions use stored tournament masks for block sizes $7 \leq n \leq 13$ and exact integer existence bounds for $14 \leq n \leq 19$. For the maximum, the critical construction has second width $\exp(O(d \log d))$ and converges in every finite L_p, with an explicit parameter–error rate. The common closure width is asymptotic to $d^2/4$, whereas exact realization requires first width at least $(1/2 - o(1))d^2$. At the critical first width, the constructed architecture therefore has zero approximation-error infimum with no finite-parameter minimizer.

1. Introduction

A fixed neural-network architecture can approximate a function arbitrarily well without representing it at any finite parameter vector. Its realization set need not be closed [22, 14]. We ask how this distinction affects the hidden widths needed to approximate a continuous piecewise-linear function. Throughout the approximation limit, both widths stay fixed; only the parameters vary.

Our principal example is the coordinate maximum

$$\mathrm{MAX}_d(x) = \max_{1 \leq i \leq d} x_i, \qquad x \in [0,1]^d.$$

Let $r_{\mathrm{cl}}(d)$ be the minimum first hidden width for which some fixed finite second width permits convergence to MAX_d in $L_2([0,1]^d)$. For standard two-hidden-layer ReLU networks, without affine input skips, we prove

$$r_{\mathrm{cl}}(d) = \kappa_d, \qquad \kappa_d = \binom{d}{2} - \left\lfloor \frac{d^2}{4} \right\rfloor, \qquad d \geq 14.$$

The proof is analytic for $d \geq 40$. For smaller dimensions, the balanced blocks use stored tournament masks at sizes $7 \leq n \leq 13$ and exact integer existence bounds at sizes $14 \leq n \leq 19$, together with the analytic size-20 block when $d = 39$. The constructed approximating sequence converges in every finite L_p; the distinction between this upper bound and an L_p threshold is made in remark 2.7. Below the L_2 threshold, each fixed finite second width m_2 satisfies the parameter-uniform bound

$$\|N - \mathrm{MAX}_d\|^2_{L_2([0,1]^d)} \geq c_0 d^{-13} (m_2 + 1)^{-10}.$$

2020 *Mathematics Subject Classification.* Primary 68T07; Secondary 41A46, 41A63, 52B11.

Key words and phrases. ReLU networks, fixed-architecture closure, nonlinear approximation, ordered aggregation, width complexity, polyhedral fans.

This lower bound also holds when the second layer and output have affine access to the input.

Two geometric results underlie the threshold. A bounded fan-localization theorem assembles cellwise ridge formulas on a finite hyperplane arrangement. If every activation cell has at least d active rows, a common perturbation makes the local coordinates recoverable; capped Möbius sums then select the correct formula. The complementary lower bound uses stable triple junctions of the target. Each first-layer row selects a coordinate pair. A junction on a triple missed by all selected pairs reduces to shallow approximation of an equilateral three-ray fan. A compact tensor probe gives an error bound uniform over the shallow network's directions, biases, and coefficients. Integrating this local obstruction over a family of transverse disks yields the global lower bound. For exact representation, a separate normalization removes affine channels from the nonlinear count, after which a support graph charges missing coordinate pairs to the remaining neurons. This last step gives the larger width required for exact equality. Figure 1 shows how these three routes combine.

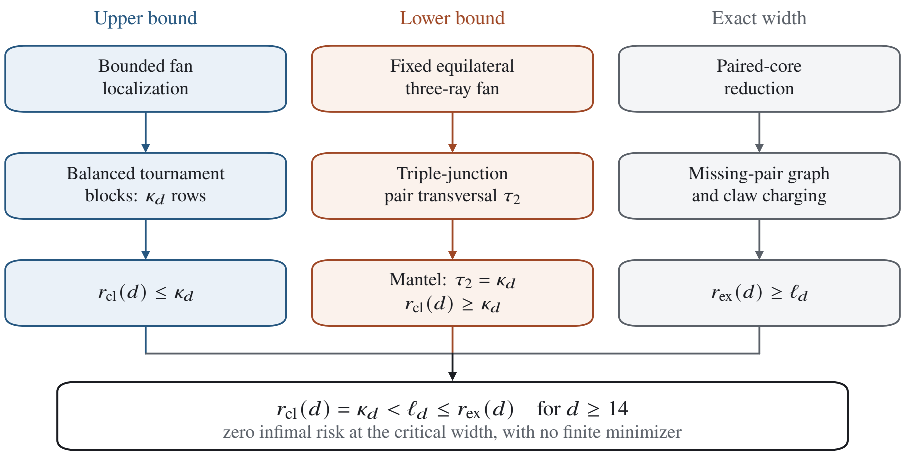

FIGURE 1. The three routes of the argument and the statement they combine into. The left column produces the approximating sequence, the middle column the parameter-uniform obstruction, and the right column the larger width needed for exact equality. Here $\ell_d = \binom{d}{2} - \lfloor 4d/3 \rfloor$.

The maximum is affine on each cell of the braid arrangement $\{x_i = x_j\}_{i<j}$. Its nonsmooth set is the smaller subcomplex

$$\bigcup_{i<j} \{x : x_i = x_j = \max_k x_k\}.$$

Balanced tournament blocks supply the upper construction, while Mantel's theorem supplies a missed triple whenever the first width is below κ_d. The same argument applies to ordered statistics

$$L_\alpha(x) = \sum_{j=1}^{d} \alpha_j x_{(j)}, \qquad x_{(1)} \geq \cdots \geq x_{(d)}.$$

If some discrete second difference of α is nonzero, the first-layer closure threshold is again κ_d for $d \geq 14$. This includes individual order statistics, proper contiguous trimmed sums, and top-k sums. The last are support functions of the hypersimplices $\Delta(d, k)$. Arithmetic weight sequences instead give affine functions plus pairwise absolute-value ridges.

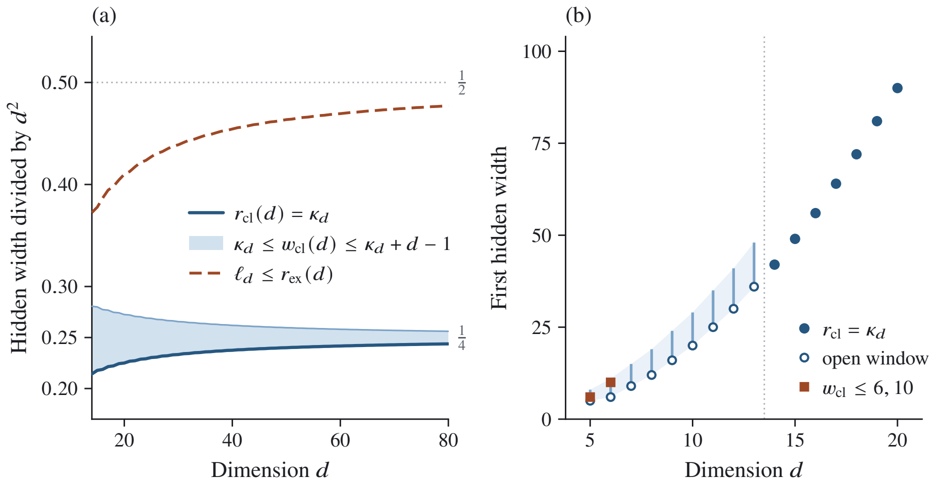

FIGURE 2. Width bounds for the maximum. (a) Normalized widths for $14 \le d \le 80$. Solid: $r_{\mathrm{cl}}(d) = \kappa_d$. Shaded band: $\kappa_d \le w_{\mathrm{cl}}(d) \le \kappa_d + d - 1$. Dashed: the necessary bound $\ell_d = \binom{d}{2} - \lfloor 4d/3 \rfloor \le r_{\mathrm{ex}}(d)$. Dotted horizontal lines mark the limiting scales $1/4$ and $1/2$. (b) The first hidden width itself for $5 \le d \le 20$. Vertical bars and open circles give the interval $\max\{d, \kappa_d\} \le r_{\mathrm{cl}}(d) \le \kappa_d + d - 1$ left open for $5 \le d \le 13$; filled circles give the determined value κ_d for $d \ge 14$. The two squares are the improved common-width bounds $w_{\mathrm{cl}}(5) \le 6$ and $w_{\mathrm{cl}}(6) \le 10$.

The common closure width $w_{\mathrm{cl}}(d)$ bounds both hidden layers. A separate construction gives

$$\max\{d, \kappa_d\} \le w_{\mathrm{cl}}(d) \le \kappa_d + d - 1 \qquad (d \ge 5),$$

so $w_{\mathrm{cl}}(d)/d^2 \to 1/4$. In contrast, every exact finite realization requires

$$m_1 \ge \binom{d}{2} - \left\lfloor \frac{4d}{3} \right\rfloor = \left(\frac{1}{2} - o(1)\right) d^2.$$

This is a necessary bound; general exact realizability remains open. For $d \ge 14$, it exceeds the closure threshold. The critical architecture therefore has a zero, unattained squared-loss infimum under the uniform distribution on the cube. Figure 2 compares the width bounds.

The closest approximation results are due to Safran, Reichman, and Valiant [30], whose lower-bound proof uses top-two supports, Mantel's theorem, and coordinate freezing under parameter bounds. We remove those bounds and determine the first-layer closure threshold. Petersen, Raslan, and Voigtlaender study nonclosed realization sets [22]; Dutta, Safran, and Valiant study approximation of rank statistics [7]. Recent work on exact maximum representations addresses different width and depth resources [28, 1, 27, 38]. Section 8 gives the quantitative comparisons.

The *decoder* is the second hidden layer together with the affine output map; its width is the second hidden width. At the critical first width, its size bound is $\exp(O(d \log d))$. At $d = 14$, the certificate gives widths (42, 16,355,852,100); the result does not provide an efficient decoder. Decoder size, conditioning, and the sharp closure threshold for $5 \le d \le 13$ remain open within the interval $\max\{d, \kappa_d\} \le r_{\mathrm{cl}}(d) \le \kappa_d + d - 1$, with the smaller upper bounds 6 and 10 in dimensions five and six. The cases $d = 3, 4$ are settled by Proposition 2.1. The appendices contain the tensor proof, finite certificates, small-width constructions, and reproducibility details.

2. Main results

Write $\rho(t) = \max\{t, 0\}$, applied coordinatewise. A standard two-hidden-layer ReLU network with input dimension d and hidden widths (m_1, m_2) has the form

$$N(x) = c^\top \rho(B\rho(Ax + a) + b) + c_0, \tag{2.1}$$

where $A \in \mathbb{R}^{m_1\times d}$ and $B \in \mathbb{R}^{m_2\times m_1}$. There are no affine input skips into the second hidden layer or the output. We denote the resulting realization class by $\mathcal{N}^{(d)}_{m_1,m_2}$. We call the second hidden layer together with the affine output map the *decoder*; its width is m_2. Unless stated otherwise, exact equality and function-space closure refer to the restriction to $[0, 1]^d$. For the displayed realization, let $\|\theta(N)\|_\infty$ be the largest absolute entry among A, a, B, b, c, c_0, and write $\mathcal{N}^{(d)}_{m_1,m_2}(R)$ for networks admitting such a realization with $\|\theta(N)\|_\infty \le R$. This is an existence convention: ReLU rescaling means that a function need not have a unique parameter norm.

The closure in this paper is the function-space closure of one fixed architecture. The widths remain fixed while the parameters may diverge. Whenever we use the phrase population risk, it means squared loss under the uniform probability distribution on $[0, 1]^d$; because the cube has unit volume, this is exactly the squared $L_2([0, 1]^d)$ norm used below. Define

$$r_{\mathrm{cl}}(d) = \min\left\{m_1 : \exists m_2 < \infty, \mathrm{MAX}_d \in \overline{\mathcal{N}^{(d)}_{m_1,m_2}}^{L_2([0,1]^d)}\right\}, \tag{2.2}$$

$$w_{\mathrm{cl}}(d) = \min\left\{\max(m_1, m_2) : \mathrm{MAX}_d \in \overline{\mathcal{N}^{(d)}_{m_1,m_2}}^{L_2([0,1]^d)}\right\}, \tag{2.3}$$

$$r_{\mathrm{ex}}(d) = \inf\left\{m_1 : \exists m_2 < \infty, \mathrm{MAX}_d \in \mathcal{N}^{(d)}_{m_1,m_2}\right\}. \tag{2.4}$$

The value in Equation (2.4) is $+\infty$ if exact finite representation does not exist. The quantity r_{cl} permits an arbitrary finite second layer; w_{cl} constrains both hidden widths by the same bound. For a general target $f \in L_2([0, 1]^d)$, we also write

$$r^{(d)}_{\mathrm{cl}}(f) = \inf\left\{m : \exists W < \infty, \quad f \in \overline{\mathcal{N}^{(d)}_{m,W}}^{L_2([0,1]^d)}\right\}. \tag{2.5}$$

We set the infimum of the empty set to $+\infty$. Thus $r_{\mathrm{cl}}(d) = r^{(d)}_{\mathrm{cl}}(\mathrm{MAX}_d)$.

For the lower bound it is useful to enlarge the model. Let

$$\mathcal{S}^{(d)}_{r,W} = \left\{\ell_0(x) + \sum_{s=1}^{W} c_s\rho\left(\ell_s(x) + \sum_{j=1}^{r} B_{sj}\rho(w_j^\top x + b_j)\right)\right\}, \tag{2.6}$$

where each ℓ_s is affine. This class gives the second hidden layer and output free affine access to the input. A lower bound for $\mathcal{S}^{(d)}_{r,W}$ therefore applies to the standard network with the same first width.

2.1. Input dimension.

For completeness, we give a proof of the input-dimension obstruction for the maximum; see also [30].

Proposition 2.1 (Input dimension). *For every $d \ge 1$, the standard architecture satisfies $r_{\mathrm{cl}}(d) \ge d$. In particular,*

$$r_{\mathrm{cl}}(3) = r_{\mathrm{ex}}(3) = 3, \qquad r_{\mathrm{cl}}(4) = r_{\mathrm{ex}}(4) = 4.$$

Proof. Suppose that $N_n \in \mathcal{N}^{(d)}_{m,W}$ converges to MAX_d in $L_2([0, 1]^d)$, with $m < d$. Choose a unit vector $v_n \in \ker A_n$. After passing to a subsequence, $v_n \to v$ with $\|v\|_2 = 1$. Since

$N_n(x + tv_n) = N_n(x)$, every $\varphi \in C_c^\infty((0,1)^d)$ satisfies

$$\int N_n(x)\, v_n \cdot \nabla\varphi(x)\, dx = 0.$$

Passing to the limit gives $\partial_v \mathrm{MAX}_d = 0$ in distributions. On the nonempty open set where coordinate i is the unique maximum, this derivative is the constant v_i. Hence $v_i = 0$ for every i, contradicting $\|v\|_2 = 1$.

For $d = 3$, use the first-layer outputs

$$h_1 = \rho(x_1 - x_2), \qquad h_2 = \rho(x_2), \qquad h_3 = \rho(x_3)$$

and output $\rho(h_1 + h_2 - h_3) + \rho(h_3)$. This is a standard network of widths $(3, 2)$ equal to MAX_3 on the cube. For $d = 4$, use

$$h_1 = \rho(x_1 - x_2), \quad h_2 = \rho(x_2), \quad h_3 = \rho(x_3 - x_4), \quad h_4 = \rho(x_4)$$

and output $\rho(h_1 + h_2 - h_3 - h_4) + \rho(h_3 + h_4)$. This gives an exact standard network of widths $(4, 2)$. The lower bound and $r_{\mathrm{cl}}(d) \le r_{\mathrm{ex}}(d)$ prove the equalities. □

2.2. **The critical width.** Let

$$\kappa_d = \binom{d}{2} - \left\lfloor \frac{d^2}{4} \right\rfloor. \tag{2.7}$$

For even $d = 2n$, this is $n(n-1)$; for odd $d = 2n+1$, it is n^2.

2.3. **Arrangement principles.** We first state the two results that transfer beyond the braid arrangement. The first assembles local ridge formulas across the cells of a finite activation arrangement. Let $K \subset \mathbb{R}^d$ be compact, Lebesgue measurable, and of positive d-dimensional measure. Let $\ell_1, \dots, \ell_m$ be nonzero homogeneous linear forms. Away from their zero hyperplanes, define the nominal active pattern

$$S(x) = \{e \in [m] : \ell_e(x) > 0\}. \tag{2.8}$$

Let Ω be the finite set of patterns attained on full-dimensional cells meeting K. Functions are Lebesgue measurable and identified up to equality almost everywhere on K.

A target f has a *finite ridge formula on K* for this arrangement if, for each $S \in \Omega$, there are affine forms $p_{S,r}$ and coefficients $a_{S,r}$, $1 \le r \le R_S$, such that

$$f(x) = \sum_{r=1}^{R_S} a_{S,r}\rho(p_{S,r}(x)) \quad \text{for a.e. } x \in K \text{ with } S(x) = S. \tag{2.9}$$

Piecewise-affine targets subordinate to the arrangement satisfy this condition because an affine form is the difference of two ReLUs. On the compact set K, the formulas define a bounded representative after assigning arbitrary bounded values on the arrangement walls. Hence such a target belongs to every finite $L_p(K)$.

For $S \in \Omega$, a set $C_S \subseteq S$ is a *downward teaching set* if

$$C_S \subseteq S' \subseteq S, \quad S' \in \Omega \quad \Longrightarrow \quad S' = S. \tag{2.10}$$

The choice $C_S = S$ always works. The teaching set identifies the cell and need not coincide with the d-row subset used to recover affine coordinates.

Theorem 2.2 (Bounded fan localization)**.** *Assume that*

$$|S| \ge d \qquad (S \in \Omega). \tag{2.11}$$

If f has a finite ridge formula on K for the arrangement, then there is a finite M such that

$$f \in \overline{\mathcal{N}^{(d)}_{m,M}}^{L_p(K)} \qquad (1 \le p < \infty). \tag{2.12}$$

One may take

$$M = 2 \sum_{S \in \Omega} R_S 2^{|C_S|}. \tag{2.13}$$

There are constants $R_0, C_p < \infty$, depending on the fixed instance, such that for every $R \geq R_0$ the same architecture contains a network $N_R \in \mathcal{N}^{(d)}_{m,M}(R)$ satisfying

$$\|N_R - f\|^p_{L_p(K)} \leq C_p R^{-1/(d+1)}. \tag{2.14}$$

Here $L_p(K)$ uses d-dimensional Lebesgue measure restricted to K; no surface measure is implicit. The finite union of arrangement hyperplanes has zero d-dimensional measure. Section 3 proves the theorem and records a support-function application for a general finite point configuration.

The second principle extracts lower bounds from stable target junctions. Its transverse model is the equilateral three-ray fan. Let $u_0, u_1, u_2 \in \mathbb{R}^2$ be unit vectors separated by angles $2\pi/3$, and set

$$F_\triangle(y) = \max_{j=0,1,2} u_j^\top y. \tag{2.15}$$

Figure 3 shows this model and its cell structure. Let $\mathcal{R}^{(2)}_W$ be the affine-plus-W-ridge class

$$\left\{ a_0 + a^\top y + \sum_{j=1}^{W} c_j \rho(v_j^\top y + b_j) \right\}.$$

Theorem 2.3 (Fixed equilateral three-ray fan)**.** *There is an absolute constant $c_\triangle > 0$ such that*

$$\inf_{g \in \mathcal{R}^{(2)}_W} \int_{\mathbb{B}^2} |g(y) - F_\triangle(y)|^2 \, dy \geq c_\triangle (W+1)^{-10}. \tag{2.16}$$

The directions, biases, and coefficients in g are arbitrary real numbers.

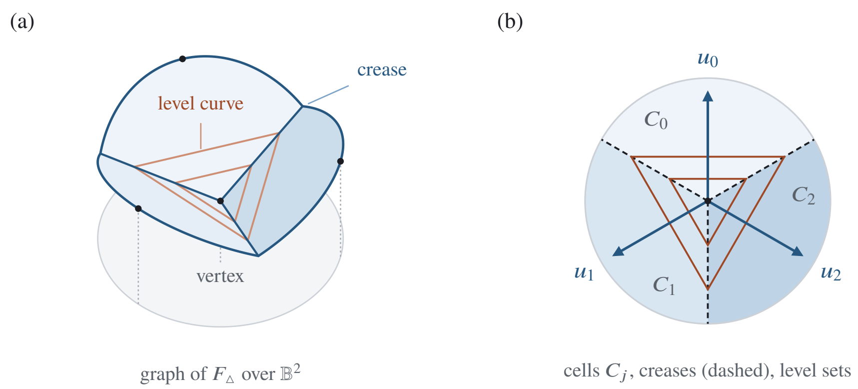

FIGURE 3. The transverse model $F_\triangle = \max_j u_j^\top y$ of Equation (2.15). (a) Its graph over the unit disk: three planar pieces meeting along the creases, with the vertex at the origin, two level curves on the surface, and dotted drop lines to the base. (b) The same function in the plane. The three shaded cells are the regions where the corresponding u_j attains the maximum, the dashed rays are the creases where two of them tie, and the nested triangles are level sets. Theorem 2.3 bounds the distance from $F_\triangle$ to every affine-plus-W-ridge function on this disk.

For $\mathcal{T} \subseteq \binom{[q]}{3}$, define its pair-transversal number

$$\tau_2(\mathcal{T}) = \min\left\{|E| : E \subseteq \binom{[q]}{2}, \quad E \cap \binom{T}{2} \neq \emptyset \text{ for every } T \in \mathcal{T}\right\}. \tag{2.17}$$

Call a coordinate triple T an *s-stable triple junction* of f if there exist a product U of positive-length intervals for the coordinates outside T, a positive-length interval $I \subset (0,1)$, $h_0 > 0$, an isometry $Q : \mathbb{R}^2 \to \mathbf{1}^\perp \subset \mathbb{R}^3$, an orthogonal map R on $\mathbb{R}^2$, and $\gamma \neq 0$ with the following property. For every $z \in U$, $a \in I$, $0 < h \leq h_0$, and $y \in \mathbb{B}^2$, the point determined by $x_{T^c} = z$ and $x_T = a\mathbf{1} + hQy$ lies in the cube and

$$f(z, a\mathbf{1} + hQy) = A_{z,a,h}(y) + h\gamma F_\triangle(Ry), \qquad A_{z,a,h} \in \mathcal{R}_s^{(2)}. \tag{2.18}$$

For $T^c = \emptyset$, the product U is a singleton. Let $\mathcal{J}_s(f)$ be the set of s-stable triple junctions of f.

Proposition 2.4 (Triple-junction transversal obstruction)**.** *Let $q \geq 3$, $s \geq 0$ be an integer, and $f \in L_2([0,1]^q)$. For every finite $\mathcal{T} \subseteq \mathcal{J}_s(f)$,*

$$r_{\mathrm{cl}}^{(q)}(f) \geq \tau_2(\mathcal{T}). \tag{2.19}$$

The same lower bound holds for the enlarged architecture with affine input skips in Equation (2.6)*. Since the ambient set of coordinate triples is finite, one may take $\mathcal{T} = \mathcal{J}_s(f)$.*

For the complete triple hypergraph, Mantel's theorem gives

$$\tau_2\left(\binom{[q]}{3}\right) = \binom{q}{2} - \mathrm{ex}(q, K_3) = \binom{q}{2} - \left\lfloor \frac{q^2}{4} \right\rfloor = \kappa_q. \tag{2.20}$$

Section 4 proves the proposition. Section 5 shows that every non-arithmetic ordered L-statistic exposes the complete triple hypergraph. Theorem 2.3 also gives the following local exact-representation obstruction.

Corollary 2.5 (Shallow MAX_3 obstruction)**.** *No finite shallow ReLU network represents* MAX_3 *exactly on $\mathbb{R}^3$, or on a neighborhood of an interior diagonal point.*

Proof. Translate the diagonal point to the origin and restrict to the plane $\mathbf{1}^\perp$. After a fixed orthogonal change of transverse coordinates, the restriction of MAX_3, modulo an affine term, is $\sqrt{2/3}\,F_\triangle$. A finite shallow network restricts to an element of $\mathcal{R}_W^{(2)}$ for some finite W. Exact equality on a transverse disk would make the left-hand side of Equation (2.16) zero, contrary to Theorem 2.3. ◻

2.4. **Closure thresholds.**

Theorem 2.6 (First-layer closure threshold)**.** *For every $d \geq 40$,*

$$r_{\mathrm{cl}}(d) = \kappa_d. \tag{2.21}$$

The same conclusion holds for every $d \geq 14$ using the stored tournament masks and exact integer existence bounds in Section 3.3. For the upper bound, one fixed standard architecture with first width κ_d and a finite second width has a parameter sequence converging to MAX_d *in $L_p([0,1]^d)$ for every $1 \leq p < \infty$.*

The three proof methods differ only in how they obtain a finite tournament for each block: stored masks for $7 \leq n \leq 13$, exact integer existence bounds for $14 \leq n \leq 19$, and an analytic estimate for $n \geq 20$. The network construction and its limiting identities are the same in all three cases. At $d = 27$ the blocks have sizes $13, 14$, and at $d = 39$ they have sizes $19, 20$, so these dimensions combine adjacent methods.

Remark 2.7 (Dependence on the approximation norm)**.** All closure widths in this paper are defined using L_2. On the unit-volume cube, $\|g\|_2 \le \|g\|_p$ for $2 \le p < \infty$. Thus the L_2 lower bounds, together with convergence of the constructed networks in every finite L_p, also identify the first-layer threshold as κ_d in every L_p with $2 \le p < \infty$, for $d \ge 14$. For $1 \le p < 2$, the construction proves the upper bound κ_d, but the three-ray fan lower bound proved here does not establish a matching L_p obstruction. The same distinction applies to the ordered statistics in Theorem 2.8.

The same threshold holds for ordered L-statistics. For $x \in \mathbb{R}^q$, let $x_{(1)} \ge \cdots \ge x_{(q)}$ be the decreasing rearrangement and define

$$L_\alpha(x) = \sum_{j=1}^{q} \alpha_j x_{(j)}. \tag{2.22}$$

Theorem 2.8 (Ordered-aggregation closure threshold)**.** *Let $q \ge 3$ and $\alpha \in \mathbb{R}^q$. Suppose that for some $r \in \{1, \dots, q-2\}$,*

$$\delta_r = \alpha_r - 2\alpha_{r+1} + \alpha_{r+2} \ne 0. \tag{2.23}$$

There is an absolute constant $c_{\mathrm{ord}} > 0$ such that, for every $m < \kappa_q$, finite W, and $N \in \mathcal{S}^{(q)}_{m,W}$,

$$\|N - L_\alpha\|^2_{L_2([0,1]^q)} \ge c_{\mathrm{ord}}\, \delta_r^2\, 4^{-q} q^{-8} (W+4)^{-10}. \tag{2.24}$$

Parameter size, sign, bias, and bit complexity are unrestricted.

If $q \ge 40$, then the matching upper bound holds analytically and

$$r^{(q)}_{\mathrm{cl}}(L_\alpha) = \kappa_q. \tag{2.25}$$

The same equality holds for every $q \ge 14$ using the stored masks and exact integer existence bounds in Section 3.3. The approximating sequence converges in every finite L_p, with one fixed finite decoder width.

Corollary 2.9 (Rank statistics)**.** *For $1 \le k \le q-1$, put*

$$T_{q,k}(x) = \sum_{j=1}^{k} x_{(j)}. \tag{2.26}$$

For every $q \ge 14$,

$$r^{(q)}_{\mathrm{cl}}(T_{q,k}) = \kappa_q. \tag{2.27}$$

The same threshold holds for every individual order statistic and every proper trimmed sum $\sum_{j=a}^{b} x_{(j)}$, where $1 \le a \le b \le q$ and $(a,b) \ne (1,q)$. In even dimension this includes each of the two middle order statistics $x_{(q/2)}$ and $x_{(q/2+1)}$, as well as their arithmetic mean. In addition,

$$T_{q,k} = h_{\Delta(q,k)}, \qquad \Delta(q,k) = \mathrm{conv}\{\mathbf{1}_S : S \subseteq [q],\ |S| = k\}, \tag{2.28}$$

so Equation (2.27) covers the support functions of all nontrivial hypersimplices. For $2 \le k \le q-2$, these are nonsimplicial polytopes with $\binom{q}{k}$ vertices.

The construction also gives the following finite-parameter rate. Let

$$M_d = \left\lceil 16|A|!|B|! \left(\frac{3}{2}\right)^{d-2} \right\rceil, \qquad |A| = \lceil d/2 \rceil, \quad |B| = \lfloor d/2 \rfloor. \tag{2.29}$$

Corollary 2.10 (Finite-parameter critical-width rate)**.** *For every $d \ge 40$, every $1 \le p < \infty$, and all sufficiently large R, there is an $N_R \in \mathcal{N}^{(d)}_{\kappa_d, M_d}(R)$ such that*

$$\|N_R - \mathrm{MAX}_d\|^p_{L_p([0,1]^d)} \le C_{d,p} R^{-1/(d+1)}. \tag{2.30}$$

The same statement holds for $14 \le d < 40$ *using the finite tournament inputs described above. In particular, the squared* L_2 *error is at most* $C_{d,2}R^{-1/(d+1)}$.

The constants may depend on the dimension. The descent-compressed decoder in Equation (2.29) is $\exp(O(d \log d))$, so no polynomial-size decoder is proved here. Hence Corollary 2.10 quantifies parameter escape in a fixed architecture. Efficient training and rate optimality remain open. The more general arrangement-level statement and proof appear in Section 3. Table 1 summarizes the three widths established in this section, including the two bounds stated next.

Quantity	Lower bound	Upper bound	Range
$r_{\mathrm{cl}}(d)$	κ_d	κ_d	$d \ge 14$
$r_{\mathrm{cl}}(d)$	$\max\{d, \kappa_d\}$	$\kappa_d + d - 1$	$5 \le d \le 13$
$r_{\mathrm{cl}}^{(d)}(L_\alpha)$	κ_d	κ_d	$d \ge 14$
$w_{\mathrm{cl}}(d)$	$\max\{d, \kappa_d\}$	$\kappa_d + d - 1$	$d \ge 5$
$r_{\mathrm{ex}}(d)$	$\binom{d}{2} - \lfloor 4d/3 \rfloor$	—	$d \ge 4$

TABLE 1. Width bounds for standard two-hidden-layer networks. The ordered statistic has $\Delta^2\alpha \ne 0$. The sharp closure upper bounds use stored masks for block sizes $7 \le n \le 13$, exact integer existence bounds for $14 \le n \le 19$, and analytic estimates for $n \ge 20$. The remaining bounds are explicit or analytic. At $d = 5, 6$, the common upper bounds improve to $6, 10$, respectively, and also bound r_{cl}. The exact-width row gives a necessary bound, without asserting existence.

Theorem 2.11 (Quantitative subcritical obstruction)**.** *An absolute constant* $c_0 > 0$ *can be chosen so that the following holds. For every* $d \ge 3$*, every* $r < \kappa_d$*, every finite* W*, and every* $N \in \mathcal{S}_{r,W}^{(d)}$*,*

$$\|N - \mathrm{MAX}_d\|_{L_2([0,1]^d)}^2 \ge c_0 d^{-13}(W+1)^{-10}. \tag{2.31}$$

No restriction is imposed on parameter norms, signs, biases, or bit complexity.

Together, Theorems 2.6 and 2.11 give Equation (2.21). They also give an inverse-polynomial lower bound when the second width is polynomial in d. If $W \le d^C$ with $C \ge 0$, then $W + 1 \le 2d^C$, so Equation (2.31) is at least $c_1 d^{-(13+10C)}$, where $c_1 = 2^{-10}c_0$.

Theorem 2.12 (Common-width bound)**.** *For every* $d \ge 5$*,*

$$\max\{d, \kappa_d\} \le r_{\mathrm{cl}}(d) \le w_{\mathrm{cl}}(d) \le \kappa_d + d - 1. \tag{2.32}$$

Consequently,

$$\lim_{d\to\infty} \frac{w_{\mathrm{cl}}(d)}{d^2} = \frac{1}{4}. \tag{2.33}$$

In dimension five, the explicit bound improves to $w_{\mathrm{cl}}(5) \le 6$*; in dimension six, it improves to* $w_{\mathrm{cl}}(6) \le 10$*.*

In particular, Equation (2.32) records the known interval for the unresolved dimensions $5 \le d \le 13$. For $d = 5$, the lower bound is 5, from Proposition 2.1, whereas $\kappa_5 = 4$; the explicit construction gives $5 \le r_{\mathrm{cl}}(5) \le w_{\mathrm{cl}}(5) \le 6$. For $d = 6$, the corresponding interval is $6 \le r_{\mathrm{cl}}(6) \le w_{\mathrm{cl}}(6) \le 10$; the actual widths $(8, 10)$ also give the sharper first-layer bound $r_{\mathrm{cl}}(6) \le 8$.

Theorem 2.13 (Exact-width bound). *For every $d \geq 4$,*

$$r_{\mathrm{ex}}(d) \geq \binom{d}{2} - \left\lfloor \frac{4d}{3} \right\rfloor . \tag{2.34}$$

We use $r_{\mathrm{ex}}(d) = +\infty$ when no finite exact realization exists, so the inequality also covers that case.

For $d \geq 14$, Theorems 2.6 and 2.13 show that zero infimal risk at the critical first width is not attained by any finite parameter vector in a fixed architecture. They also give the asymptotic separation

$$r_{\mathrm{cl}}(d) = \left(\frac{1}{4} + o(1)\right) d^2, \qquad r_{\mathrm{ex}}(d) \geq \left(\frac{1}{2} - o(1)\right) d^2. \tag{2.35}$$

The second relation is a necessary bound; general exact realizability is not established here. Table 1 collects the three widths.

3. Fan localization

This section proves Theorem 2.2 and applies it to tournament-oriented comparison arrangements. The latter application gives the upper half of Theorem 2.6.

3.1. **Localization.**

The theorem applies to general finite arrangements. For example, let $v_1, \ldots, v_q$ be pairwise distinct and take

$$P = \mathrm{conv}\{v_1, \ldots, v_q\} \subset \mathbb{R}^d, \qquad \ell_{ij}^{\pm}(x) = \pm(v_i - v_j)^\top x \quad (i < j),$$

include both oriented forms, and assume $\binom{q}{2} \geq d$. Off the comparison walls, exactly one orientation of each pair is active, so every cell has $\binom{q}{2}$ active rows. Since the support function

$$h_P(x) = \max_{1 \leq j \leq q} v_j^\top x$$

is affine on each cell, Theorem 2.2 places h_P, for every fixed compact $K \subset \mathbb{R}^d$ and $1 \leq p < \infty$, in the fixed-architecture $L_p(K)$ closure with first width $2\binom{q}{2}$ and some finite decoder width. Thus the result assembles cellwise affine or finite-ridge rules on a finite activation arrangement. The coordinate simplex used later is one optimized instance.

The hypothesis uses the number of active rows; nominal independence is unnecessary. The proof makes these forms simultaneously invertible by a common infinitesimal perturbation.

Figure 4 summarizes the construction. The first layer stores the critical comparison ridges and applies one common full-spark perturbation. Each activation cell then admits a left inverse; a bounded cap and downward Möbius sum retain its local ridge formula.

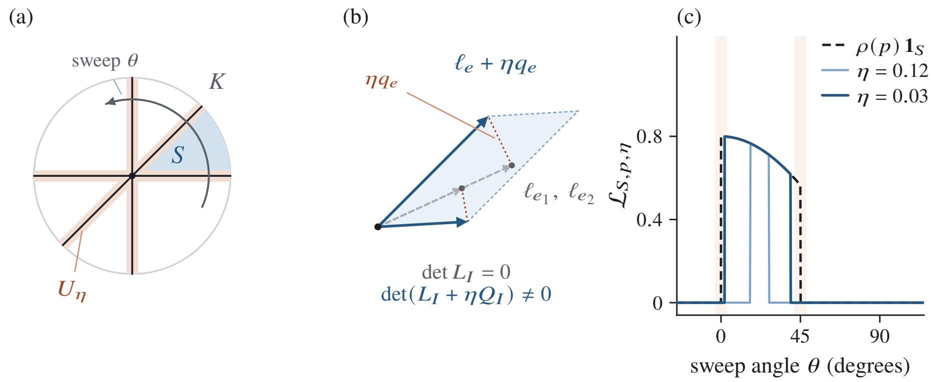

FIGURE 4. Fan localization, drawn in the plane. (a) Three walls of the arrangement cut the compact set K into cells. The shaded sector is one cell S; the shaded strips are the wall neighborhood U_η defined below, whose measure is $O(\eta)$. (b) The common tilt applied below: two selected rows ℓ_{e_1}, ℓ_{e_2} that are parallel, hence useless for recovery, become independent after adding ηq_e, so the corresponding minor is nonzero. The tilt is exaggerated. (c) The Möbius localizer $\mathcal{L}_{S,p,\eta}$ constructed below, along the circular sweep marked in panel (a), for two values of η. Outside the shaded wall strips it already equals its limit $\rho(p)\mathbf{1}_S$.

Proof of Theorem 2.2. We first perturb the shared first layer so that each cell can recover its local affine forms. We then use bounded caps and a Möbius sum to retain exactly the formula of that cell; a final wall-neighborhood estimate quantifies the resulting convergence.

Choose distinct rational numbers $t_1, \ldots, t_m$, for instance $t_e = e$, and define the full-spark forms

$$q_e(x) = x_1 + t_e x_2 + \cdots + t_e^{d-1} x_d. \tag{3.1}$$

For $\eta > 0$, the first hidden layer consists of

$$h_{e,\eta}(x) = \rho(\ell_e(x) + \eta q_e(x)), \qquad e \in [m]. \tag{3.2}$$

Fix $S \in \Omega$ and select any d-element subset $I \subseteq S$. Let L_I and Q_I be the matrices whose rows are the coefficients of ℓ_e and q_e, respectively. The polynomial

$$\det(L_I + \eta Q_I) \tag{3.3}$$

is nonzero: its coefficient of η^d is the nonzero Vandermonde determinant $\det Q_I$. A nonzero polynomial has no positive roots in a sufficiently small punctured interval around zero. Since Ω is finite, one interval works for a selected minor of every pattern.

Consequently, fix an off-wall point x whose nominal activation pattern is S. Every affine form $p(x) = u^\top x + b$ can be recovered as a second-layer preactivation

$$P_{S,p,\eta}(x) = \sum_{e \in I} c_{S,p,e}(\eta) h_{e,\eta}(x) + b = p(x). \tag{3.4}$$

Because all ℓ_e and q_e are homogeneous, the bias b here is exactly the constant term of p, independent of η. For each such fixed x, there is a threshold $\eta_x > 0$ such that the equality holds for $0 < \eta < \eta_x$, once the signs at x agree with the nominal signs. This pointwise sign-stabilization threshold is distinct from the common punctured interval on which all selected minors are nonsingular. Cramer's rule shows, uniformly over the finite set of selected minors, that

$$|c_{S,p,e}(\eta)| = O(\eta^{-d}). \tag{3.5}$$

The exponent d suffices for the estimate below.

It remains to select the correct cell without exposing the input to the second layer. For $B > 0$, define the bounded cap

$$\chi_B(t) = \rho(t) - \rho(t - B), \qquad 0 \le \chi_B(t) \le B. \tag{3.6}$$

Choose $B > 0$ with $B \ge \max_{S,r} \sup_{x\in K} p_{S,r}(x)$. Then $\chi_B(p_{S,r}) = \rho(p_{S,r})$ on K.

For $T \subseteq C_S$, set

$$g_{S,T,\eta}(x) = \sum_{e\in([m]\setminus S)\cup T} h_{e,\eta}(x), \qquad M_\eta = \eta^{-(d+1)}, \tag{3.7}$$

and form

$$\mathcal{L}_{S,p,\eta}(x) = \sum_{T\subseteq C_S} (-1)^{|T|} \chi_B\big(P_{S,p,\eta}(x) - M_\eta g_{S,T,\eta}(x)\big). \tag{3.8}$$

Each summand in Equation (3.8) uses two ordinary second-layer ReLUs. It reads only the first-layer outputs and a bias.

Fix $x \in K$ outside the arrangement walls and let $S' = S(x)$. Its perturbed signs equal its nominal signs for all sufficiently small η. If $S' \not\subseteq S$, an index in $S' \setminus S$ appears in every gate in Equation (3.7), and its first-layer output tends to a fixed positive number. Thus the gate penalty, which grows as $\eta^{-(d+1)}$, dominates the reconstructed preactivation $P_{S,p,\eta}(x) = O(\eta^{-d})$, so every capped term is eventually zero.

It remains to consider $S' \subseteq S$. A gate is zero if and only if $T \cap S' = \varnothing$. For every sufficiently small fixed η, all zero-gate terms have the same capped preactivation $\chi_B(P_{S,p,\eta}(x))$, independently of T. Their total coefficient is

$$\sum_{T\subseteq C_S\setminus S'} (-1)^{|T|} = \mathbf{1}_{\{C_S\subseteq S'\}}. \tag{3.9}$$

If it is zero, the terms cancel for each sufficiently small η. If it is one, Equation (2.10) forces $S' = S$. Then only $T = \varnothing$ is unsuppressed, the selected active rows recover $P_{S,p,\eta}(x) = p(x)$, and the surviving cap is $\chi_B(p(x)) = \rho(p(x))$. Combining the three cases gives

$$\mathcal{L}_{S,p,\eta}(x) \longrightarrow \mathbf{1}_{\{S'=S\}}\rho(p(x)). \tag{3.10}$$

Now define

$$N_\eta(x) = \sum_{S\in\Omega} \sum_{r=1}^{R_S} a_{S,r} \mathcal{L}_{S,p_{S,r},\eta}(x). \tag{3.11}$$

Equation (3.10) and the cellwise ridge formulas give $N_\eta(x) \to f(x)$ for almost every $x \in K$. The function N_η is a fixed finite sum of caps bounded by B. Dominated convergence already proves Equation (2.12). The number of second-layer ReLUs is

$$M = 2\sum_{S\in\Omega} R_S 2^{|C_S|}. \tag{3.12}$$

We quantify this convergence. Finiteness of the selected minors, Cramer's rule, and compactness of K give constants $Q, C_P < \infty$ such that, for all sufficiently small η,

$$\max_{e,x\in K} |q_e(x)| \le Q, \qquad \max_{S,p,x\in K} |P_{S,p,\eta}(x)| \le C_P \eta^{-d}. \tag{3.13}$$

Enlarge C_P to absorb the finitely many fixed biases and the cap height, choose $\Gamma > Q + C_P + 1$, and set

$$U_\eta = \bigcup_{e=1}^{m} \{x \in K : |\ell_e(x)| \le \Gamma\eta\}. \tag{3.14}$$

Outside U_η, perturbed and nominal signs agree, and every active row obeys

$$h_{e,\eta}(x) \ge (\Gamma - Q)\eta.$$

Consequently every nonzero gate in Equation (3.8) satisfies

$$M_\eta g_{S,T,\eta}(x) \geq (\Gamma - Q)\eta^{-d} > C_P \eta^{-d}.$$

All terms intended to be suppressed are therefore exactly zero. The target cell retains its recovered ridge, and every proper-subpattern contribution cancels by the same Möbius sum used above. Hence the stronger finite-η identity

$$N_\eta(x) = f(x) \qquad \text{for a.e. } x \in K \setminus U_\eta \tag{3.15}$$

holds for all sufficiently small η.

The union in Equation (3.14) has volume at most $C_U\eta$. To see this, enclose K in a fixed box and, for each nonzero $\ell_e(x) = a_e^\top x$, integrate in a coordinate satisfying $|a_{e,j}| \geq \|a_e\|_2/\sqrt{d}$. Fubini's theorem bounds the permitted interval in that coordinate by $2\Gamma\eta/|a_{e,j}|$. Since both the network and target have a fixed bounded envelope, it follows that

$$\|N_\eta - f\|_{L_p(K)}^p \leq C_p'\eta. \tag{3.16}$$

Finally, direct inspection of the realization shows

$$\|\theta(N_\eta)\|_\infty \leq C_0\eta^{-(d+1)}. \tag{3.17}$$

The first-layer weights remain bounded; a second-layer coefficient is a sum of an $O(\eta^{-d})$ active-frame coefficient and a gate coefficient of magnitude $M_\eta = \eta^{-(d+1)}$; all cap shifts and output coefficients are fixed. Given sufficiently large R, choose $\eta = (C_0/R)^{1/(d+1)}$. Then Equations (3.16) and (3.17) give Equation (2.14). Applying this result to the two tournament blocks proves Corollary 2.10. □

Capped ReLU differences appear in fixed-width nonclosedness constructions [22], and active-frame reconstruction appears in injective and convex ReLU architectures [25, 11]. Theorem 2.2 combines these ingredients with an exact-support Möbius gate while sharing one first layer across all cells.

3.2. **Tournament blocks.** Split $[d] = A \,\dot\cup\, B$ with $|A| = \lceil d/2 \rceil$ and $|B| = \lfloor d/2 \rfloor$. Orient every edge of each complete block. For an arrow $e = (u, v)$, put

$$\ell_e(x) = x_u - x_v. \tag{3.18}$$

The total number of first-layer rows is

$$|E| = \binom{|A|}{2} + \binom{|B|}{2} = \kappa_d. \tag{3.19}$$

List a total order from the smallest to the largest coordinate, and call an arrow active if its tail has the larger coordinate. A sufficient blockwise condition is that every order of a tournament on n vertices has at least n active arrows. Choose such a tournament independently on A and B. Every product order cell then has at least

$$|A| + |B| = d \tag{3.20}$$

active rows, so Theorem 2.2 applies.

The cell selector can be much smaller than its full active set. For a tournament T and an order $\pi = (\pi_1, \ldots, \pi_n)$ from smallest to largest, let

$$D_T(\pi) = \big\{\{\pi_k, \pi_{k+1}\} : 1 \leq k < n, \text{ the arrow on this pair is active}\big\} \tag{3.21}$$

and define the weighted descent count

$$Z(T) = \sum_{\pi \in S_n} 2^{|D_T(\pi)|}. \tag{3.22}$$

For a product order $\pi = (\pi_A, \pi_B)$, take $C_\pi = D_{T_A}(\pi_A) \,\dot\cup\, D_{T_B}(\pi_B)$. This is a downward teaching set: if another product-order pattern S_τ satisfies $C_\pi \subseteq S_\tau \subseteq S_\pi$, then every pair consecutive in π_A and π_B has the same order in τ: reversing an active consecutive pair violates the first inclusion, while reversing an inactive one violates the second. Consecutive comparisons determine a total order, so $\tau = \pi$. In particular,

$$|C_\pi| \le d - 2. \tag{3.23}$$

On a product order cell, let i_S and j_S be the largest coordinates in A and B. Since the input lies in the unit cube,

$$\mathrm{MAX}_d(x) = \rho(x_{j_S}) + \rho(x_{i_S} - x_{j_S}). \tag{3.24}$$

Thus $R_S = 2$ and $B = 1$ suffice in the localization theorem. We obtain a standard network with first width κ_d and

$$m_2 \le 4Z(T_A)Z(T_B). \tag{3.25}$$

This separates the d-row extraction frame from the smaller teaching set used by the Möbius gate.

3.3. **Tournament existence.** For a random tournament on n vertices and a fixed order, the number of active arrows is $X \sim \mathrm{Binomial}(N, 1/2)$, where $N = \binom{n}{2}$. A union bound over the $n!$ orders gives

$$\Pr\{\text{some order has fewer than } n \text{ active arrows}\} \le n!2^{-N} \sum_{k=0}^{n-1} \binom{N}{k}. \tag{3.26}$$

For the same random tournament and a fixed order, the $n - 1$ adjacent orientations are independent fair bits. Therefore

$$\mathbb{E}_T 2^{|D_T(\pi)|} = \left(\frac{3}{2}\right)^{n-1}, \qquad \mathbb{E}_T Z(T) = n! \left(\frac{3}{2}\right)^{n-1}. \tag{3.27}$$

Markov's inequality makes the probability that $Z(T) \ge 2n!(3/2)^{n-1}$ at most $1/2$.

For $n \ge 20$, Hoeffding's inequality and $n! \le n^n$ bound the right-hand side by

$$\exp\left(n \log n - \frac{(n-1)(n-4)^2}{4n}\right) < 1. \tag{3.28}$$

At $n = 20$, the exponent in Equation (3.28) is less than $-4/5$. Its derivative is negative for $n \ge 20$, and its second derivative is negative there, so a suitable tournament exists for every $n \ge 20$. Balanced blocks now prove the upper bound in Theorem 2.6 for $d \ge 40$ without a finite search certificate.

For $14 \le n \le 19$, the right-hand side of Equation (3.26) is compared with $1/2$ using exact integers and is strictly smaller. The analytic bound is below $e^{-4/5} < 1/2$ for $n \ge 20$. Combining either estimate with Equation (3.27) shows that, for every $n \ge 14$, one tournament simultaneously has at least n active arrows in every order and

$$Z(T) < 2n! \left(\frac{3}{2}\right)^{n-1}. \tag{3.29}$$

For $7 \le n \le 13$, the archive supplies explicit tournaments. Their exact statistics are summarized below; Appendix A lists the orientations.

n	Minimum active arrows	$Z(T)$
7	7	63,945
8	8	756,123
9	10	10,125,785
10	12	150,772,041
11	16	2,508,092,749
12	18	43,971,740,613
13	25	891,525,490,669

TABLE 2. Exact statistics for the archived tournament certificates. The quantity $Z(T) = \sum_{\pi \in S_n} 2^{|D_T(\pi)|}$ is the weighted adjacent-descent sum.

The minimum is verified without sampling. If $D(U)$ is the minimum active-arrow count over orders of the subtournament induced by $U \subseteq [n]$, placing v last gives the exact recurrence

$$D(\varnothing) = 0, \qquad D(U) = \min_{v \in U} \left(D(U \setminus \{v\}) + |\{u \in U \setminus \{v\} : v \to u\}| \right). \tag{3.30}$$

It uses $O(n2^n)$ integer operations. For $n = 7$, a separate enumeration of all 7! orders gives the same minimum. The masks, recurrence, exact binomial comparisons, and analytic tail establish existence for every block size $n \geq 7$, hence every $d \geq 14$. The proof uses stored masks for $7 \leq n \leq 13$, exact integer existence inequalities for $14 \leq n \leq 19$, and the analytic estimate for $n \geq 20$. In dimension terms, both blocks use stored masks for $14 \leq d \leq 26$; $d = 27$ combines a stored mask and an integer existence bound; $28 \leq d \leq 38$ uses two integer existence bounds; $d = 39$ combines an integer bound and the analytic estimate; and $d \geq 40$ is entirely analytic. The values of $Z(T)$ are independently obtained by the weighted Hamilton-path recurrence

$$H(\{v\}, v) = 1, \qquad H(U, v) = \sum_{u \in U \setminus \{v\}} H(U \setminus \{v\}, u)\, 2^{\mathbf{1}_{\{v \to u\}}}, \qquad Z(T) = \sum_{v} H([n], v). \tag{3.31}$$

The table values all satisfy Equation (3.29). Figure 5 displays the certificates and the three regimes of the existence argument.

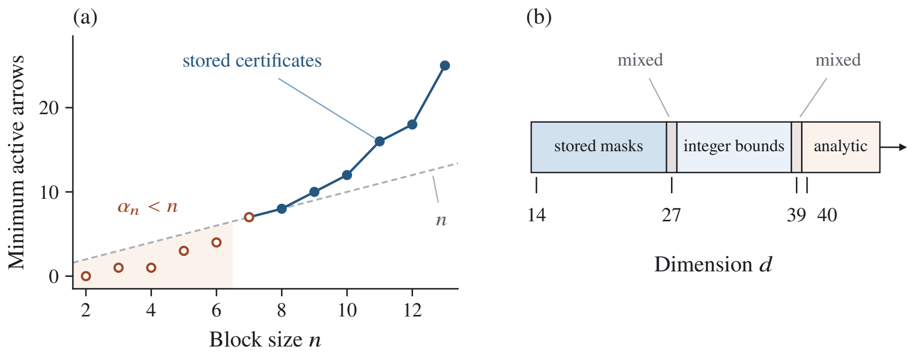

FIGURE 5. Tournament certificates and the regimes of the existence proof. (a) Minimum active-arrow count against block size. Open circles are the optimal values α_n for $2 \le n \le 7$; below the diagonal $\alpha_n < n$, so no balanced two-block arrangement qualifies there. Filled markers are the archived certificates of Table 2, all on or above the diagonal. (b) Which argument supplies the certificate in each range of d: stored masks for $14 \le d \le 26$, exact integer existence inequalities for $28 \le d \le 38$, the analytic estimate of Equation (3.28) for $d \ge 40$, and mixed pairs at $d = 27$ and $d = 39$.

The lower endpoint $d = 14$ reflects a limitation of this application of the active-row hypothesis, rather than a limit on the finite search. Indeed, let α_n be the largest possible minimum active-arrow count among all n-vertex tournaments. Exhaustive integer checks give $\alpha_3, \alpha_4, \alpha_5, \alpha_6 = 1, 1, 3, 4$, and $\alpha_2 = 0$, $\alpha_7 = 7$; details appear in Appendix A.1. Since the two block orders vary independently, their best possible minimum total is $\alpha_{\lceil d/2 \rceil} + \alpha_{\lfloor d/2 \rfloor}$, which is strictly less than d for every $5 \le d \le 13$. In particular, at $d = 13$ even the optimal total is $7 + 4 = 11$, so one block cannot compensate for the other. This excludes the present balanced two-block arrangement from Theorem 2.2 in that range; it does not exclude other localization arguments or architectures.

Combining Equations (3.25) and (3.29) gives the uniform decoder bound

$$m_2 \le \left\lceil 16|A|!|B|! \left(\frac{3}{2}\right)^{d-2} \right\rceil \le \left\lceil \frac{64}{9} \left(\frac{3(d+1)}{4}\right)^{d} \right\rceil. \tag{3.32}$$

This improves the full-active-set exponent from $O(d^2)$ to $O(d \log d)$. For the displayed seven-vertex certificate, $d = 14$ gives the sharper exact count

$$4Z(T_7)^2 = 16{,}355{,}852{,}100. \tag{3.33}$$

4. TRIPLE-JUNCTION BOUNDS

The proof of Theorem 2.11 selects a coordinate triple missed by every first-layer row and freezes the network on a positive-measure family of transverse disks. On each disk the first hidden layer becomes affine, while the target retains an equilateral three-ray junction. Applying Theorem 2.3 on the disks and integrating over their base gives a bound independent of all network parameters. Figure 6 shows the three steps.

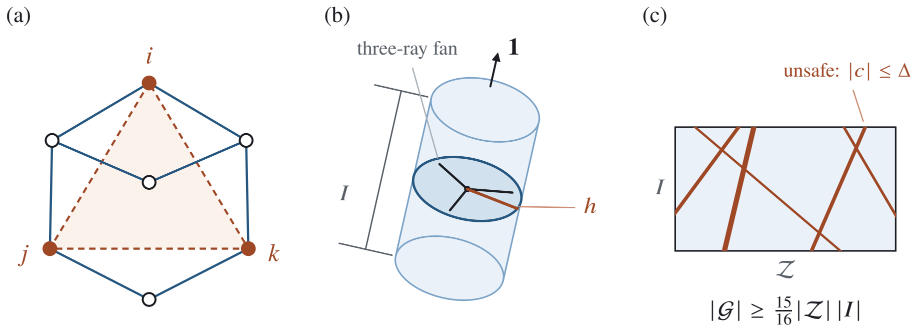

FIGURE 6. The subcritical reduction. (a) The graph of selected top-two coordinate pairs. It has seven vertices and eight edges, fewer than $\kappa_7 = 9$, so some triple $\{i, j, k\}$ carries no selected pair. (b) The oblique cylinder built below over that triple: transverse disks of radius h with axis $\mathbf{1}$, swept as the common shift ranges over I. On each disk the target is the equilateral three-ray fan, whose creases are drawn on the middle disk. (c) The base $\mathcal{Z} \times I$. Each first-layer row is unsafe only on a thin slab $|c| \leq \Delta$; the union of the m_1 slabs leaves the safe set $\mathcal{G}$, on which the network restricts to a shallow ridge sum.

4.1. **Three-ray bound.** The proof of Theorem 2.3 is in Appendix B. We record the steps that make its parameter quantifiers explicit.

For a large odd order k, a compactly supported smooth distributional probe maps a ridge $\rho(v^\top y+b)$ to a scalar multiple of the pure symmetric tensor $v^{\otimes k}$. A width-W shallow network therefore has probe tensor of symmetric rank at most W, regardless of its coefficients and biases. For the equilateral fan, the same probe has an explicit folded-binomial catalecticant. Choosing $k \asymp W^2$, an exact central sign minor and the Eckart–Young theorem give a Frobenius tail of order W^{-3}. The compact radial probe agrees with the Gaussian near the fan vertex, and therefore has the same jet through order $k-3$, while its normalized order-k derivative energy remains controlled. Its jet is still sensitive to the three half-rays. Duality and Parseval then yield the squared-error rate $(W+1)^{-10}$. Writing M_k for the matrix obtained by grouping the tensor slots, the choices and losses form the chain

$$k \asymp (W+1)^2, \qquad \inf_{\operatorname{rank} R \leq W} \|M_k - R\|_F \gtrsim (W+1)^{-3}, \qquad \sigma \asymp k^{-1/2} \asymp (W+1)^{-1}, \tag{4.1}$$

and the scaled compact probe gives

$$\|g - F_\triangle\|_2 \gtrsim \sigma^2 (W+1)^{-3} \asymp (W+1)^{-5}, \qquad \|g - F_\triangle\|_2^2 \gtrsim (W+1)^{-10}. \tag{4.2}$$

The argument estimates distance to a matrix-rank variety uniformly over the noncompact shallow parameter set, and hence remains valid along sequences with diverging directions or output weights. Hermite rank and matricization methods also occur in approximation and learning problems [36, 21, 3]; Appendix B supplies the target-specific growing-order spectrum and the compact localization used here.

Lemma 4.1 (One-dimensional slab estimate)**.** *Let $J \subset \mathbb{R}$ be an interval of length $L > 0$, let $\lambda \neq 0$, let $h \geq 0$ and $B \geq 0$, and suppose*

$$\Phi(t, y) = \lambda t + c + h\psi(y), \qquad \sup_{y \in \mathbb{B}^2} |\psi(y)| \leq B|\lambda|.$$

The set of $t \in J$ for which $\Phi(t, \cdot)$ takes both nonpositive and nonnegative values has length at most $2Bh$, and therefore occupies at most a fraction $2Bh/L$ of J.

Proof. Such a sign change implies $|\lambda t+c| \le h \sup_y |\psi(y)| \le Bh|\lambda|$. The resulting interval in t has length at most $2Bh$. $\square$

4.2. **Transversals.**

Proof of Proposition 2.4. We first find a witnessed triple missed by the selected coordinate pairs. A slab estimate then leaves a common family of disks on which every first-layer activation is affine, so the fixed-fan theorem applies after subtracting the shallow part of the witness. For each nonzero first-layer row, select the two coordinates carrying its largest absolute coefficients, with a fixed tie rule. The resulting simple graph has at most m edges. If $m < \tau_2(\mathcal{T})$, some witnessed triple $T \in \mathcal{T}$ contains no selected edge. Fix one set of witness data $(U_T, I_T, h_T, Q_T, R_T, \gamma_T)$ for each $T \in \mathcal{T}$. The argument below produces an error $\varepsilon_T > 0$ that depends on these data and on m, W, s, while remaining independent of all network parameters. Since $\mathcal{T}$ is finite, $\min_{T\in\mathcal{T}} \varepsilon_T > 0$.

If $q = 3$, a nonempty junction hypergraph has pair-transversal number one, so $m < \tau_2(\mathcal{T})$ means $m = 0$. Fix $0 < h \le h_T$. On each witnessed disk, subtracting $A_{a,h}$ from the network leaves at most $W+s$ shallow ridges. Applying Theorem 2.3 to $h\gamma_T F_\triangle$ gives squared error at least $c_\triangle h^2\gamma_T^2(W+s+1)^{-10}$. Integration over $a \in I_T$, with Jacobian $\sqrt{3}h^2$, gives a positive ambient L_2 lower bound. Assume $q \ge 4$.

Fix the witness cylinder for T. Write a first-row weight as $w = (v, u)$, where $v \in \mathbb{R}^3$ is on T, and let $M = \|v\|_\infty$, $U = \|u\|_\infty$, and $\sigma = \mathbf{1}^\top v$. At most one coordinate of v exceeds U in magnitude, and hence

$$M > 4U \quad\Longrightarrow\quad |\sigma| > M/2. \tag{4.3}$$

If $M = U = 0$, the transverse slope is zero and the row is safe on every fiber. Otherwise, on a transverse disk the preactivation has the form

$$u^\top z + \sigma a + b + h(Q^\top v)^\top y.$$

If $M \le 4U$, condition on an outside coordinate with coefficient U. If $M > 4U$, condition on a and use Equation (4.3). In the first case, $\|Q^\top v\|_2 \le \sqrt{3}M \le 4\sqrt{3}U$; in the second, $\|Q^\top v\|_2 \le \sqrt{3}M < 2\sqrt{3}|\sigma|$. Applying Lemma 4.1 on the selected witness interval shows that the fraction of base points on which this affine form changes sign across the disk is at most $C_T h$, uniformly in the row parameters. The constant depends only on the fixed witness intervals.

Choose

$$h = \min\left\{h_T, \frac{1}{2C_T \max\{1, m\}}\right\}.$$

A union bound leaves at least half the witness-base volume on which every first-layer ReLU is affine or zero throughout the disk. On each such fiber, a two-hidden-layer network with second width W reduces to an element of $\mathcal{R}_W^{(2)}$. After subtracting $A_{z,a,h}$, at most $W+s$ shallow ridges remain. Theorem 2.3 and integration with Jacobian $\sqrt{3}h^2$ give the squared ambient error bound

$$\frac{\sqrt{3}}{2} c_\triangle \gamma_T^2 h^4 |U_T||I_T|(W+s+1)^{-10} > 0,$$

independent of all network parameters. Thus no fixed architecture with $m < \tau_2(\mathcal{T})$ can converge to f in L_2. Affine skips remain affine on every fiber. $\square$

4.3. **Missed triples.** Consider $N \in \mathcal{S}^{(d)}_{m_1,m_2}$ as in Equation (2.6). For each nonzero first-layer vector w_j, choose the two coordinates of largest absolute weight, with a fixed deterministic tie rule. Add the resulting unordered pair to a simple graph G on $[d]$. Duplicates are discarded, so $|E(G)| \le m_1$.

If every triple contains a selected pair, then the complement $\overline{G}$ is triangle-free. Mantel's theorem [19, 2] gives $|E(\overline{G})| \le \lfloor d^2/4 \rfloor$, and therefore

$$|E(G)| \ge \binom{d}{2} - \left\lfloor \frac{d^2}{4} \right\rfloor = \kappa_d. \tag{4.4}$$

When $m_1 < \kappa_d$, there is consequently a triple T containing no selected pair. Relabel so that $T = \{1, 2, 3\}$.

For one first-layer row write $w = (v, u)$, where $v \in \mathbb{R}^3$ is supported on T, and set

$$M = \max_{i \in T} |v_i|, \qquad U = \max_{j \notin T} |u_j|, \qquad s = v_1 + v_2 + v_3. \tag{4.5}$$

Because the two largest coordinates of w are not both in T, at most one coordinate of v exceeds U in magnitude. Hence

$$M > 4U \quad \Longrightarrow \quad |s| \ge M - 2U > M/2. \tag{4.6}$$

4.4. **Oblique cylinders.** Let $Q : \mathbb{R}^2 \to \mathbf{1}^\perp \subset \mathbb{R}^3$ be an isometry. For $d \ge 4$, define

$$\begin{aligned} \mathcal{Z} &= [0, 1 - 1/d]^{d-3}, & I &= \left[1 - \frac{1}{2d}, 1 - \frac{1}{4d}\right], \\ h &= \frac{d^{-3}}{64\sqrt{3}}, & x_{T^c} &= z, \qquad x_T = a\mathbf{1} + hQy, \end{aligned} \tag{4.7}$$

where $(z, a, y) \in \mathcal{Z} \times I \times \mathbb{B}^2$. The cylinder lies in the unit cube, and its three distinguished coordinates exceed all the others. Thus

$$\mathrm{MAX}_d(x) = a + hF_Q(y), \qquad F_Q(y) = \max_{i \le 3} (Qy)_i. \tag{4.8}$$

The three rows of Q form an equilateral triple of radius $\sqrt{2/3}$. Up to an orthogonal map R_0,

$$F_Q(y) = \sqrt{\frac{2}{3}} F_\triangle(R_0 y). \tag{4.9}$$

The absolute Jacobian of $(a, y) \mapsto a\mathbf{1} + hQy$ is $\sqrt{3}h^2$.

On the cylinder, one first-layer preactivation is

$$c(z, a) + h\, v^\top Q y, \qquad c(z, a) = u^\top z + sa + b. \tag{4.10}$$

Put $\Delta = h\|Q^\top v\|_2 \le \sqrt{3}hM$. A base point (z, a) is safe for this neuron if $\Delta = 0$ or $|c(z, a)| > \Delta$. At a safe point, the row's ReLU is affine or zero as a function of y on the entire disk.

Lemma 4.2 (One-row slab estimate)**.** *Under the uniform distribution on $\mathcal{Z} \times I$, each first-layer row satisfies*

$$\Pr\{\textit{not safe}\} \le 16\sqrt{3}\, dh. \tag{4.11}$$

Proof. There is nothing to prove when $\Delta = 0$. Suppose first that $M \le 4U$, so $U > 0$. Condition on a and on all outside coordinates except a coordinate z_j whose coefficient has magnitude U. Its interval has length at least $1/2$, and therefore

$$\Pr\{|c| \le \Delta \mid \text{other variables}\} \le \frac{2\Delta}{U(1 - 1/d)} \le 16\sqrt{3}h. \tag{4.12}$$

If $M > 4U$, condition on z, use Equation (4.6), and recall that $|I| = 1/(4d)$:

$$\Pr\{|c| \le \Delta \mid z\} \le \frac{2\Delta}{|s||I|} \le 16\sqrt{3}\, dh. \tag{4.13}$$

The case $M = U = 0$ belongs to $\Delta = 0$. ◻

Let $\mathcal{G} \subset \mathcal{Z} \times I$ be the set safe for all first-layer rows. Since $m_1 < \kappa_d < d^2/4$, the union bound and the chosen value of h give

$$\frac{|\mathcal{G}|}{|\mathcal{Z}||I|} \geq 1 - 16\sqrt{3}m_1 dh = 1 - \frac{m_1}{4d^2} > \frac{15}{16}. \tag{4.14}$$

Fix $(z, a) \in \mathcal{G}$. Every first-layer ReLU is affine or zero in y throughout the disk. After substitution, the second hidden layer becomes a shallow ridge sum of width at most m_2. Free affine skips in Equation (2.6) restrict to affine terms and preserve the reduction. Subtracting a, rescaling by h, and applying Theorem 2.3 and Equation (4.9) yields

$$\int_{\mathbb{B}^2} |N(x(z,a,y)) - (a + hF_Q(y))|^2\, dy \geq \frac{2}{3} c_\triangle h^2 (m_2+1)^{-10}. \tag{4.15}$$

Integrating over $\mathcal{G}$ and inserting the Jacobian gives

$$\begin{aligned} \|N - \mathrm{MAX}_d\|^2_{L_2([0,1]^d)} &\geq \frac{2\sqrt{3}}{3} c_\triangle h^4 |\mathcal{G}| (m_2+1)^{-10} \\ &\geq \frac{5\sqrt{3}}{8} c_\triangle h^4 |\mathcal{Z}||I| (m_2+1)^{-10}. \end{aligned} \tag{4.16}$$

Now $|\mathcal{Z}| = (1-1/d)^{d-3} \geq 1/4$, $|I| = 1/(4d)$, and $h = d^{-3}/(64\sqrt{3})$. Thus Equation (2.31) holds for every $d \geq 4$ with the explicit choice

$$c_0 = \frac{5\sqrt{3}}{128(64\sqrt{3})^4}\, c_\triangle. \tag{4.17}$$

The exponent 13 comes from the factor h^4, which contributes d^{-12}, and the length of the common-shift interval, which contributes d^{-1}.

When $d = 3$, $m_1 < \kappa_3 = 1$ means $m_1 = 0$, so the skip network is a width-m_2 shallow ridge sum on $\mathbb{R}^3$. Around $x_* = (1/2, 1/2, 1/2)$, write

$$x = x_* + ae_0 + \frac{1}{8}Uy, \qquad e_0 = 3^{-1/2}\mathbf{1}, \qquad a \in [-1/8, 1/8], \quad y \in \mathbb{B}^2,$$

where $U : \mathbb{R}^2 \to e_0^\perp$ is an isometry. Every coordinate differs from $1/2$ by at most $1/(8\sqrt{3}) + 1/8 < 1/2$, so the cylinder lies inside the cube. The interval has length $1/4$, and the absolute Jacobian of $(a, y) \mapsto x$ is $(1/8)^2 = 1/64$, since (e_0, U) is orthogonal. After a fixed transverse rotation R_0,

$$\mathrm{MAX}_3(x) = \frac{1}{2} + \frac{a}{\sqrt{3}} + \frac{1}{8}\sqrt{\frac{2}{3}} F_\triangle(R_0 y).$$

At each fixed a, restriction leaves a two-dimensional shallow ridge sum of width at most m_2. Subtracting the affine term and applying Theorem 2.3, followed by integration, gives

$$\|N - \mathrm{MAX}_3\|^2_{L_2([0,1]^3)} \geq c_\triangle \frac{2}{3}\left(\frac{1}{8}\right)^2 \frac{1}{4}\frac{1}{64}(m_2+1)^{-10} = \frac{c_\triangle}{24576}(m_2+1)^{-10}. \tag{4.18}$$

The choice in Equation (4.17) satisfies $c_0\, 3^{-13} \leq c_\triangle/24576$, so the same constant works for $d = 3$. This completes the proof of Theorem 2.11.

4.5. **Boundary rigidity.** The equality case of Mantel's theorem also records what a critical-width sequence must look like; see Figure 7.

Corollary 4.3 (Critical comparison graph)**.** *Fix $d \geq 3$ and finite W. Suppose $N_n \in \mathcal{S}^{(d)}_{\kappa_d, W}$ and $\|N_n - \mathrm{MAX}_d\|_2 \to 0$. After removing constant first rows, the selected top-two pairs are eventually all distinct, and their graph is*

$$K_A \,\dot\cup\, K_B, \qquad \{|A|, |B|\} = \{\lfloor d/2 \rfloor, \lceil d/2 \rceil\}, \tag{4.19}$$

after passage to a subsequence with a fixed bipartition.

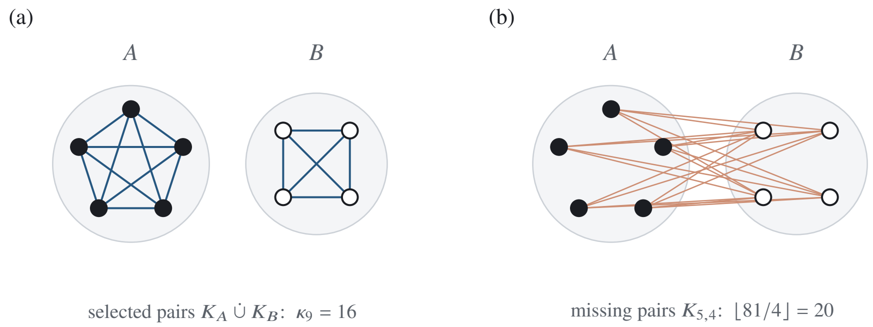

FIGURE 7. The rigid shape of a critical-width sequence at $d = 9$, by Corollary 4.3. (a) The selected top-two pairs must eventually form the disjoint union of two complete graphs on a balanced bipartition, here $K_5 \dot\cup K_4$ with $\kappa_9 = 16$ edges. (b) Its complement is the complete bipartite graph $K_{5,4}$, which has $\lfloor 81/4 \rfloor = 20$ edges and is the unique extremal graph in Mantel's theorem. No other configuration of κ_d rows admits an approximating sequence.

Proof. For $d = 3$, the critical first width is one. A constant first row reduces the network to a shallow sum, which has a uniform positive error by Theorem 2.11 with first width zero. Thus the row is eventually nonconstant; its selected pair gives $K_2 \dot\cup K_1$. Assume $d \geq 4$. If a row is constant or two rows select the same pair, the simple selected graph has fewer than κ_d edges, so Mantel's theorem supplies a missing triple. In the proof of Theorem 2.11, the hypothesis $r < \kappa_d$ is used only to produce such a triple. Once the triple is given, the same cylinder argument applies at $r = \kappa_d$: with $h = d^{-3}/(64\sqrt{3})$, its union bound is $16\sqrt{3}\,\kappa_d d h = \kappa_d/(4d^2) < 1/16$. The fixed-fan lower bound therefore gives a uniform positive error. Thus all κ_d pairs are eventually distinct and every triple is hit. The complement is a triangle-free graph with exactly $\lfloor d^2/4 \rfloor$ edges. Equality in Mantel's theorem makes it a balanced complete bipartite graph, proving Equation (4.19). □

5. ORDERED STATISTICS

The general lower bound is Proposition 2.4, proved in Section 4. We identify the stable-junction hypergraph of an ordered L-statistic and then construct the matching fan-localized approximating sequence. A nonzero second difference of the rank weights isolates a three-ray fan after subtraction of three shallow ridges. For the upper bound, merging two sorted chains gives the required local ridge formulas.

5.1. **Weight curvature.**

Fix r satisfying Equation (2.23) and a coordinate triple T. Place $r-1$ outside coordinates in $[3/4, 1]$, the remaining $q-r-2$ outside coordinates in $[0, 1/4]$, and set

$$x_T = a\mathbf{1} + hQy, \qquad a \in [3/8, 5/8], \qquad y \in \mathbb{B}^2. \tag{5.1}$$

For sufficiently small h, the three coordinates in T occupy global ranks $r, r+1, r+2$. For three numbers $y_{(1)} \geq y_{(2)} \geq y_{(3)}$, the exact identity

$$\begin{aligned} \lambda_1 y_{(1)} + \lambda_2 y_{(2)} + \lambda_3 y_{(3)} &= \lambda_2(y_1 + y_2 + y_3) + (\lambda_2 - \lambda_3)(\max_i y_i - \min_i y_i) \\ &\quad + (\lambda_1 - 2\lambda_2 + \lambda_3)\max_i y_i \end{aligned} \tag{5.2}$$

holds on every order cone. The range term is shallow because

$$\max_i y_i - \min_i y_i = \frac{1}{2}\sum_{i<j} |y_i - y_j|. \tag{5.3}$$

Each absolute value is twice a ReLU plus an affine term, so the range uses three ReLU ridges when affine terms are free. Taking $(\lambda_1, \lambda_2, \lambda_3) = (\alpha_r, \alpha_{r+1}, \alpha_{r+2})$, the last coefficient is δ_r. Since the rows of Q form an equilateral triple, Equations (5.2) and (5.3) give Equation (2.18) with $s = 3$ and $|\gamma| = \sqrt{2/3}|\delta_r|$. The construction works for every T, so

$$\mathcal{J}_3(L_\alpha) = \binom{[q]}{3}. \tag{5.4}$$

Combining Proposition 2.4 and Equation (2.20) proves the qualitative lower bound in Theorem 2.8.

For the displayed quantitative bound, take $h = (128\sqrt{3}\,q^2)^{-1}$, which keeps the middle three coordinates strictly between $1/4$ and $3/4$. For $q \geq 4$, select a missed triple as in Section 4. All base intervals in Equation (5.1) have length $1/4$. If $M \leq 4U$, conditioning on an outside coordinate bounds the unsafe fraction by $2h\sqrt{3}M/(U/4) \leq 32\sqrt{3}h$. If $M > 4U$, conditioning on a and using $|s| > M/2$ gives the smaller bound $16\sqrt{3}h$. The union over $m < \kappa_q < q^2/4$ rows has fraction below $1/16$. The safe base volume is at least $(15/16)4^{-(q-2)}$. When $q = 3$, the hypothesis forces $m = 0$, so every base point is safe and the same volume bound holds.

The cylinder Jacobian is $\sqrt{3}h^2$, and the squared transverse amplitude is $(2/3)h^2\delta_r^2$. Applying Theorem 2.3 after subtracting the three range ridges therefore gives, for all $q \geq 3$,

$$\begin{aligned} \|N - L_\alpha\|^2_{L_2([0,1]^q)} &\geq \frac{2\sqrt{3}}{3} c_\triangle \delta_r^2 h^4 \frac{15}{16} 4^{-(q-2)} (W+4)^{-10} \\ &= c_{\mathrm{ord}} \delta_r^2 4^{-q} q^{-8} (W+4)^{-10}, \end{aligned} \tag{5.5}$$

where an explicit absolute constant is

$$c_{\mathrm{ord}} = \frac{10\sqrt{3}}{(128\sqrt{3})^4}\, c_\triangle = 16 c_0. \tag{5.6}$$

Here c_0 is the choice in Equation (4.17); thus both quantitative bounds use fixed multiples of the same disk constant $c_\triangle$, with all dependence on dimension and weights displayed separately. This proves Equation (2.24).

5.2. **Chain merging.** It remains to prove the upper bound. Split the coordinates into the two tournament blocks from Section 3.2. On one product order cell, write the two decreasing chains as

$$a_1 \geq \cdots \geq a_p, \qquad b_1 \geq \cdots \geq b_s, \qquad p + s = q. \tag{5.7}$$

For the top-k sum, set

$$\ell = \max(0, k - s), \qquad u = \min(k, p), \qquad F_t = \sum_{i=1}^{t} a_i + \sum_{j=1}^{k-t} b_j \quad (\ell \leq t \leq u).$$

Choosing the top k elements from the union of the chains gives $T_{q,k} = \max_t F_t$, and

$$F_{t+1} - F_t = a_{t+1} - b_{k-t} \tag{5.8}$$

is nonincreasing in t. Figure 8 illustrates the split. A discrete concave sequence increases exactly through its positive increments, hence

$$T_{q,k} = F_\ell + \sum_{t=\ell}^{u-1} \rho(a_{t+1} - b_{k-t}). \tag{5.9}$$

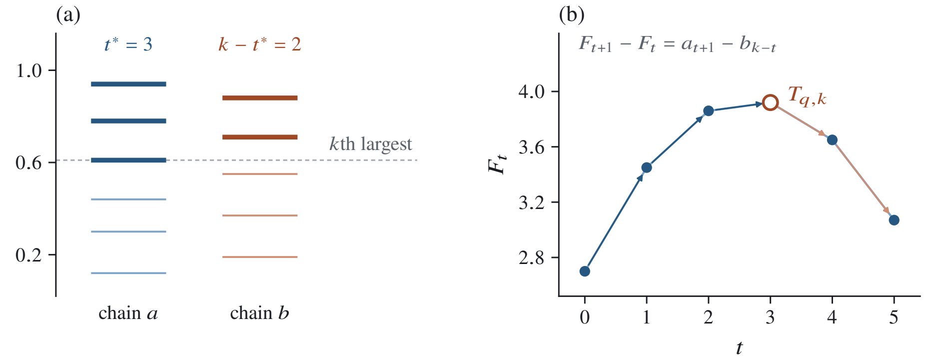

FIGURE 8. Merging two sorted chains on one tournament product cell, for $p = 6$, $s = 5$, and $k = 5$. (a) The two decreasing chains of Equation (5.7). Heavy marks are the entries chosen by the optimal split t^*; the dashed line is the kth largest entry of the union. (b) The partial sums F_t. Their increments Equation (5.8) are nonincreasing, so the sequence is discretely concave and its maximum $T_{q,k}$, circled, is reached through the positive increments only. This is Equation (5.9), a finite ridge formula on the cell.

Thus every top-k sum has a finite ridge formula on every tournament product cell. Every ordered L-statistic then telescopes as

$$L_\alpha = \alpha_q T_{q,q} + \sum_{k=1}^{q-1} (\alpha_k - \alpha_{k+1}) T_{q,k}, \qquad T_{q,q} = \sum_i x_i. \tag{5.10}$$

The active-row count and tournament ranges are unchanged from the maximum case. Applying Theorem 2.2 proves the upper half of Theorem 2.8. For each fixed (q, α), the finitely many local ridge formulas share one finite second hidden-layer width $W_{q,\alpha}$, fixed throughout the approximating sequence. The optimal such width and practical conditioning remain open.

Finally, if every discrete second difference of α vanishes, then $\alpha_j = A + Bj$ is an arithmetic progression. The identity

$$\sum_{i<j} |x_i - x_j| = \sum_{j=1}^{q} (q + 1 - 2j) x_{(j)} \tag{5.11}$$

shows that L_α is already affine plus pairwise ReLU ridges. The non-arithmetic hypothesis is therefore a structural boundary of the triple-junction mechanism; no κ_q lower bound is claimed for the arithmetic subspace.

For $T_{q,k}$, the rank-weight vector is a nonconstant step and therefore has a nonzero second difference whenever $k < q$. Individual order statistics and proper trimmed sums have the same property. The support identity in Equation (2.28) follows from $T_{q,k}(x) = \max_{|S|=k} \sum_{i \in S} x_i$, completing the proof of Corollary 2.9.

6. COMMON WIDTH

The common-width upper bound in Theorem 2.12 uses a quadratic-width construction based on the one-sided tangent gate below.

Lemma 6.1 (Tangent gate)**.** *For $\varepsilon > 0$, let*

$$h_\varepsilon(L, q) = \rho(L - \varepsilon q), \qquad C_\varepsilon(L, q) = \frac{L - h_\varepsilon(L, q)}{\varepsilon}. \tag{6.1}$$

Then $C_\varepsilon(L,q) \le q$ *for all* L, q, *and for fixed* $L \neq 0$,

$$C_\varepsilon(L,q) \longrightarrow \begin{cases} q, & L>0, \\ -\infty, & L<0. \end{cases} \tag{6.2}$$

Proof. If $L - \varepsilon q \ge 0$, then $C_\varepsilon = q$. Otherwise $C_\varepsilon = L/\varepsilon < q$. The limit follows directly. □

Figure 9 illustrates the gate, the two-block selector that the next subsection builds from it, and the diagonal defect that confines the convergence to L_p.

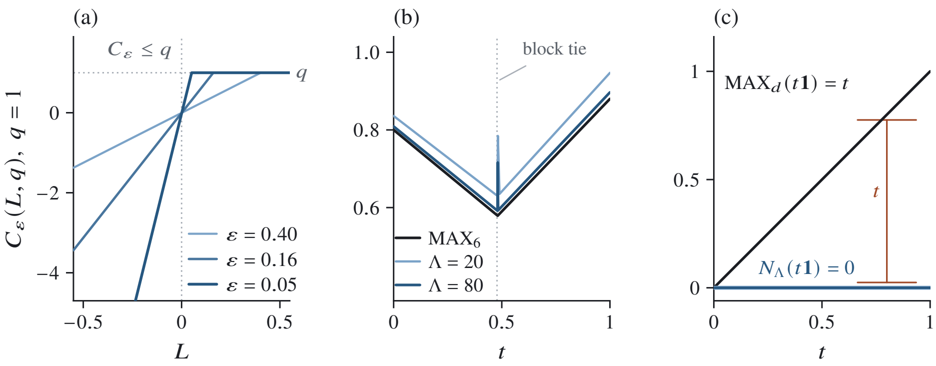

FIGURE 9. The tangent gate and the two-block selector at $d = 6$. (a) The gate $C_\varepsilon(L,q)$ of Equation (6.1) at $q = 1$. It never exceeds q, and as $\varepsilon \downarrow 0$ it tends to q for $L > 0$ and diverges to $-\infty$ for $L < 0$, which is Equation (6.2). (b) The two-block selector N_Λ assembled below, along a segment on which the block winners change. It approaches MAX$_d$ off the block tie, and the gap concentrates near the tie as Λ grows. (c) The diagonal $x = t\mathbf{1}$. Every unit vanishes there, so $N_\Lambda(t\mathbf{1}) = 0$ for all Λ while MAX$_d(t\mathbf{1}) = t$; the convergence is therefore in L_p only, not uniform.

When L is a linear combination of first-layer outputs, Equation (6.1) is a legal second-layer preactivation. A small tilt of one first-layer hyperplane supplies the affine quantity q.

6.1. **Construction.** We retain relative-coordinate channels so that every within-block comparison remains available to the second layer. Tilting selected cycle hinges then recovers the two block offsets, while loser penalties ensure that only the units for the block winners survive in the limit.

Split $[d] = A \dot\cup B$, with $|A| = a \ge 3$ and $|B| = b \ge 3$. Choose anchors $a_0 \in A$, $b_0 \in B$, and write

$$\alpha = x_{a_0}, \quad \beta = x_{b_0}, \quad y_i = x_i - \alpha, \quad z_j = x_j - \beta, \quad \delta = \alpha - \beta. \tag{6.3}$$

The first layer contains $d-2$ always-active relative-coordinate channels

$$\rho(y_i+1) = y_i + 1 \quad (i \neq a_0), \qquad \rho(z_j+1) = z_j + 1 \quad (j \neq b_0), \tag{6.4}$$

and one oriented comparison hinge for every unordered pair within a block. All within-block differences are therefore linear combinations of the first outputs and biases.

Choose a directed Hamiltonian cycle in each block. If s_A, s_B are the successor maps, set

$$L_i = x_i - x_{s_A(i)}, \qquad M_j = x_j - x_{s_B(j)}. \tag{6.5}$$

Tilt the corresponding cycle hinges by

$$H_{i,\varepsilon} = \rho(L_i - \varepsilon\delta), \qquad K_{j,\varepsilon} = \rho(M_j - \varepsilon\beta). \tag{6.6}$$

All other comparison hinges remain untilted.

For a vertex v, use the stored orientation of every incident edge to reconstruct the blockwise loser penalty

$$D_{v,\varepsilon} = \sum_{e=(p,v)} h_{e,\varepsilon} + \sum_{e=(v,q)} \big(h_{e,\varepsilon} - \ell_e\big). \tag{6.7}$$

At $\varepsilon = 0$, this is

$$D_v = \sum_{w \neq v} \rho(x_w - x_v). \tag{6.8}$$

Only two incident cycle edges move, and the tilt forms are bounded by one on the cube. Hence

$$|D_{v,\varepsilon} - D_v| \le 2\varepsilon, \qquad D_{v,\varepsilon} \ge -2\varepsilon. \tag{6.9}$$

Set $\varepsilon = \Lambda^{-2}$. For $j \in B$ and $(i, j) \in A \times B$, use the second-layer units

$$G_{j,\Lambda} = \rho\left(\frac{M_j - K_{j,\varepsilon}}{\varepsilon} + z_j - \Lambda D_{j,\varepsilon}\right), \tag{6.10}$$

$$R_{ij,\Lambda} = \rho\left(\frac{L_i - H_{i,\varepsilon}}{\varepsilon} + y_i - z_j - \Lambda(D_{i,\varepsilon} + D_{j,\varepsilon})\right). \tag{6.11}$$

All displayed affine differences are shorthand for legal linear combinations of Equation (6.4); there is no input skip.

At a point with unique block winners i_* and j_*, their unperturbed penalties vanish, every loser penalty is positive, and the winner's outgoing cycle edge has positive comparison value. Also, $D_{i_*,\varepsilon}, D_{j_*,\varepsilon} = O(\varepsilon)$, so the choice $\varepsilon = \Lambda^{-2}$ gives $\Lambda D_{i_*,\varepsilon}, \Lambda D_{j_*,\varepsilon} = O(\Lambda^{-1}) \to 0$. By Lemma 6.1 and Equation (6.9),

$$G_{j,\Lambda} \to \mathbf{1}_{\{j=j_*\}} x_{j_*}, \qquad R_{ij,\Lambda} \to \mathbf{1}_{\{(i,j)=(i_*,j_*)\}} \rho(x_{i_*} - x_{j_*}). \tag{6.12}$$

Thus

$$N_\Lambda = \sum_{j \in B} G_{j,\Lambda} + \sum_{i \in A, j \in B} R_{ij,\Lambda} \longrightarrow \mathrm{MAX}_d \tag{6.13}$$

away from block ties. The inequality $C_\varepsilon(L, q) \le q$, together with Equation (6.9), bounds every $G_{j,\Lambda}$ and $R_{ij,\Lambda}$ by an absolute constant. Dominated convergence proves the L_2 limit.

For this two-block tangent-gate sequence, used when $d \ge 7$, convergence is not uniform. If $x = t\mathbf{1}$ with $t \in (0, 1]$, then

$$\delta = y_i = z_j = L_i = M_j = 0, \qquad H_{i,\varepsilon} = K_{j,\varepsilon} = D_{i,\varepsilon} = D_{j,\varepsilon} = 0.$$

Consequently, every unit in Equations (6.10) and (6.11) vanishes, so $N_\Lambda(t\mathbf{1}) = 0$ for every Λ, whereas $\mathrm{MAX}_d(t\mathbf{1}) = t$. Hence

$$\sup_{x \in [0,1]^d} |N_\Lambda(x) - \mathrm{MAX}_d(x)| \ge t$$

for every Λ. This sequence establishes convergence in every finite L_p; it does not furnish an exact finite-parameter realization. This defect is panel (c) of Figure 9.

The widths are

$$m_1 = \binom{a}{2} + \binom{b}{2} + d - 2, \qquad m_2 = b(a + 1). \tag{6.14}$$

For odd $d = 2n + 1 \ge 7$, take $(a, b) = (n + 1, n)$ to obtain

$$(m_1, m_2) = (\kappa_d + d - 2, \kappa_d + d - 1). \tag{6.15}$$

For even $d = 2n \geq 8$, the slightly unbalanced split $(a, b) = (n + 1, n - 1)$ gives

$$(m_1, m_2) = (\kappa_d + d - 1, \kappa_d + d - 2). \tag{6.16}$$

In the two smallest remaining dimensions, explicit switching formulas give widths $(6, 4)$ for $d = 5$ and $(8, 10)$ for $d = 6$. Appendix C gives explicit network formulas and verifies their chamberwise limits. This proves the upper bound in Equation (2.32); the lower bound follows from Proposition 2.1 and Theorem 2.11.

7. Exact realization

We explain the support constraints behind Theorem 2.13. Appendix D gives the full proof.

7.1. **Nonlinear cores.** A pair $\rho(r), \rho(-r)$ with opposite coefficients in every next-layer preactivation transmits $\rho(r) - \rho(-r) = r$. Call such units *paired channels*; all unpaired units are *cores*. Paired channels contribute a linear form to every receiving preactivation, so only cores are counted as nonlinear first-layer resources. The support of a first core $\rho(w^\top x)$ is $\operatorname{supp}(w) = \{i : w_i \neq 0\}$.

Proposition 7.1 (Paired-core reduction)**.** *If a standard two-hidden-layer network computes* MAX_d *exactly on the cube, then there is a homogeneous paired-core network with two hidden layers computing* MAX_d *on* $\mathbb{R}^d$*, with at most the original first width in first-layer cores. Every first-core normal* w *satisfies*

$$\mathbf{1}^\top w = 0, \qquad |\operatorname{supp}(w)| \geq 2. \tag{7.1}$$

For each $I \subseteq [d]$ *with* $|I| \geq 3$*, restricting to the coordinates in* I *produces an exact representation of* $\mathrm{MAX}_{|I|}$ *whose first cores are among those with support contained in* I*.*

Proposition D.2 and Lemma D.3 prove the reduction by local homogenization at an interior diagonal point, followed by freezing external coordinates at separated negative values. Paired channels can increase the physical row count. Each retained first core nevertheless descends from a distinct original row, so its count satisfies $m \leq m_1$.

In particular, every three-coordinate set contains the support of a first core. Otherwise the restriction would leave no nonlinear first layer and would give a finite shallow representation of MAX_3, contradicting Corollary 2.5.

7.2. **Missing comparisons.** Let $P \subseteq \binom{[d]}{2}$ be the set of distinct two-coordinate core supports. Define the missing-pair graph

$$H = K_d \setminus P, \qquad e = |E(H)|, \qquad s = m - |P|, \tag{7.2}$$

where m is the number of first cores. The surplus s counts support-three-or-larger cores and duplicate pair supports. Since

$$m = \binom{d}{2} - e + s, \tag{7.3}$$

it is enough to charge missing edges to surplus cores.

Every triangle of H forces a core supported on exactly that triangle: triple coverage supplies a support inside its three vertices, admissibility excludes singleton supports, and the definition of H excludes all three pair supports. Paired channels cannot supply this core because their combined contribution is linear. An induced claw with center v also forces a core supported on three or four of its vertices and containing v (Lemma D.6). The latter constraint follows by differentiating in x_v: the first moment of the resulting signed atomic measure would otherwise express the maximum of the three leaf coordinates as a shallow ReLU sum. Lemma D.5 justifies this step without any sign restriction on the network coefficients. Figure 10 shows the two forced configurations.

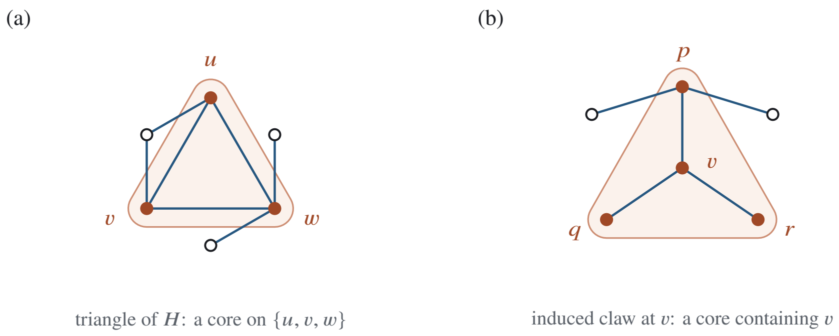

FIGURE 10. The two configurations in the missing-pair graph H of Equation (7.2) that force a surplus core. (a) A triangle $\{u, v, w\}$: triple coverage supplies a core inside these three vertices, and no pair among them is a core support, so the core is supported on exactly the triangle. (b) An induced claw centered at v: differentiating in x_v forces a core containing v and supported on three or four of the four vertices. Counting these witnesses without repetition bounds the surplus s from below.

Let d_v be the degree of v in H, let $\tau(H)$ be the number of triangles, and let κ_v count induced claws centered at v. Triangle cores and claw witnesses are disjoint. A witness for a claw centered at v cannot serve another center, and serves at most $d_v - 2$ claws at that center. Proposition D.7 therefore gives

$$s \geq \tau(H) + \sum_{v:d_v \geq 3} \frac{\kappa_v}{d_v - 2}. \tag{7.4}$$

Applying the local graph inequality in Lemma D.8 to each neighbor graph $H[N_H(v)]$ and summing gives

$$\tau(H) + \sum_{v:d_v \geq 3} \frac{\kappa_v}{d_v - 2} \geq e - \frac{4d}{3}. \tag{7.5}$$

Combining Equations (7.3) to (7.5) yields

$$m \geq \binom{d}{2} - \frac{4d}{3}. \tag{7.6}$$

Integer rounding and $m \leq m_1$ prove Theorem 2.13.

7.3. **Nonattainment.**

Corollary 7.2 (Nonattainment)**.** *For every $d \geq 14$, there is a fixed standard architecture with first width κ_d, finite second width, and zero infimal L_2 risk for* MAX_d*, but no finite parameter vector in that architecture attains the infimum. Every minimizing parameter sequence has unbounded Euclidean norm.*

Figure 11 summarizes the resulting picture.

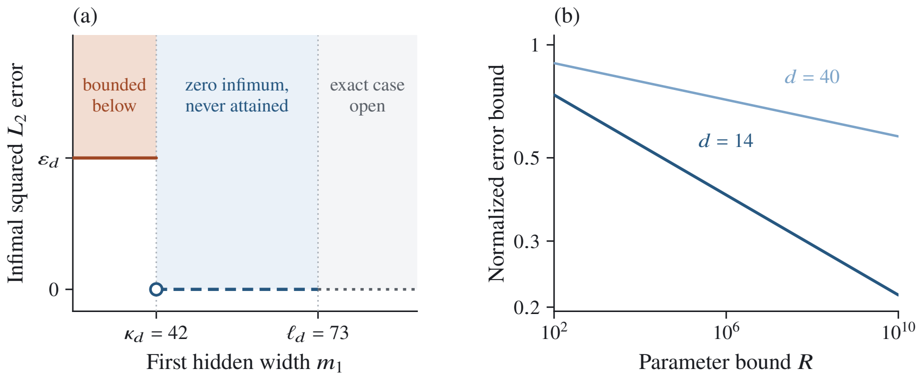

FIGURE 11. Nonattainment at the critical width, drawn at $d = 14$. (a) The infimal squared L_2 risk as a function of the first hidden width. Below κ_d it is bounded away from zero by Theorem 2.11. At κ_d it drops to zero but is not attained, by Corollary 7.2; the open circle marks this jump. Attainment is possible only above $\ell_d = \binom{d}{2} - \lfloor 4d/3 \rfloor$, where the exact question remains open. (b) The rate $R^{-1/(d+1)}$ of Equation (2.14): a minimizing sequence must send its parameter bound R to infinity, and does so slowly in high dimension.

Proof. The upper construction in Section 3 gives zero infimal risk. For $d \geq 14$, the lower bound in Equation (2.34) is strictly larger than κ_d, so no network with that first width represents MAX_d exactly. If a minimizing parameter sequence had a bounded subsequence, that subsequence would remain minimizing, and finite-dimensional compactness would give a further convergent subsequence. The realization map is continuous at finite parameters on the compact cube, so the limit parameter would realize MAX_d, a contradiction. Thus every minimizing sequence eventually leaves every bounded parameter set. □

Thus the proven exact first-width lower bound is asymptotically twice the first-layer closure threshold. The exact minimum and all-dimensional exact two-hidden-layer existence remain open.

8. Discussion

8.1. **Related bounds.** Table 1 fixes the width conventions. Table 3 records the closest numerical results; the paragraphs below state the differences in model and target.

Reference	Target	Notion	Bound
[30]	Maximum	Approximation	$w \leq d(d+1)$
[28]	Maximum	Exact	$w \geq \lfloor d^2/8 - d/4 - 1/2 \rfloor$
[7]	Sorting	Approximation	$w \leq 4d^2$
This paper	Maximum	Closure	$r_{\mathrm{cl}} = \kappa_d$

TABLE 3. Related width bounds for two-hidden-layer ReLU networks. Here w denotes common width; the final row minimizes the first width, allowing one fixed finite second width, and applies for $d \geq 14$. Parameter restrictions for the associated lower bounds are stated below.

The upper selectors and lower-bound geometry of Safran, Reichman, and Valiant [30] are direct antecedents: top-two supports, Mantel's theorem, and coordinate freezing. Their approximation lower bound applies below common width $d^2/5$ with exponentially bounded parameters. We add a coefficient-free rate for the fixed equilateral three-ray fan and an oblique cylinder, giving $c_0 d^{-13}(m_2+1)^{-10}$ with the same absolute constant c_0 as in Equation (2.31) and unrestricted parameters. The upper theorem optimizes first-layer closure width with one fixed decoder.

Exact MAX representation has been studied through extremal graph theory and polyhedral identities [12, 28, 1, 27, 38]. Rueß et al. give exact two-hidden-layer identities through $n = 12$. Our exact result is an all-dimensional first-layer lower bound with unrestricted finite decoder width.

For ordered aggregation, Dutta, Safran, and Valiant [7] give arbitrary-accuracy median and rank-statistic networks, including a quadratic-width sorter. For the median, they prove a positive error lower bound when the first width is at most $d-1$, at arbitrary depth. Their sorter is exact on separated inputs away from the boundary; controlling the exceptional set gives cube approximation. For $q \geq 14$, we determine the minimum first width in a fixed two-hidden-layer closure and prove a parameter-uniform subcritical law. Approximation rates for deep networks and differentiable top-k maps concern complementary resources [34, 32].

Broader approximation theory gives depth–width separations for natural target families [31], dimension-scale width thresholds when depth may grow [16], and wide-to-narrow conversions that spend additional depth [35]. A related unrestricted-parameter depth result is Safran [29]. Its theorems use constructed Lipschitz or CPWL targets, special absolutely continuous distributions, or exact cube equality. Our target is the coordinate maximum or an ordered L-statistic, the distribution is uniform on the cube, and, for $d \geq 14$, the quantity is the sharp first-layer L_2-closure threshold of one depth-three class.

8.2. **Related methods.** Nonclosed realization sets and nonattained objectives have been established for fixed architectures [22, 14]. Fixed-architecture nonclosedness also occurs in Sobolev topologies under activation-dependent hypotheses [17]. Finite-sample ReLU empirical risk can force parameter escape [15], while optimal shallow-ReLU approximants exist under the assumptions of Dereich and Kassing [6]. Maxout units place finite maxima inside the activation [9]; here the maximum remains the target of a standard ReLU network. Paired ReLUs and their symmetries are studied in earlier work [33, 23, 26], while hinging-hyperplane, ridge-function, and braid-arrangement methods provide the geometric setting [37, 24, 10, 5]. Classical arbitrary-ridge approximation gives sharp worst-case L_2 rates for Sobolev classes [18].

General CPWL functions admit modern dimension-independent neuron bounds in the number of pieces [4]. Finite-cost infinite-width shallow CPWL representations reduce to finite width [20], and two-hidden-layer networks weakly represent piecewise-linear finite elements with mesh-dependent widths [13].

The upper proof combines capped ReLU differences, gate penalties, active frame reconstruction, and a downward Möbius localizer [22, 30, 25, 11]. The shallow lower bound uses Hermite probes, tensor rank, and matricization; related tensor methods appear in approximation and learning problems [36, 21, 3]. Appendix B derives the growing-order folded catalecticant and compact localization for the equilateral three-ray fan. The final matrix estimate uses Eckart–Young [8].

8.3. **Open questions.** The critical construction has a nonpolynomial decoder bound. Gradient-based training, finite-precision behavior, and decoder optimality remain open. Fan localization applies to finite-ridge CPWL rules, including max-linear predictors, when each cell exposes enough active rows. Its width and conditioning depend on the arrangement. A related norm-controlled setting exhibits dimension-dependent shallow-network parameter growth that an additional hidden layer can reduce [39].

For the maximum in $5 \le d \le 13$, the known interval is $\max\{d, \kappa_d\} \le r_{\mathrm{cl}}(d) \le w_{\mathrm{cl}}(d) \le \kappa_d + d - 1$, with the improved upper bounds 6 at $d = 5$ and 10 at $d = 6$. The exact first-layer closure threshold within this interval remains open, as does that of general non-arithmetic ordered statistics in dimensions three through thirteen. The finite values of $w_{\mathrm{cl}}(d)$, the best critical decoder, and the norm–width tradeoff are also unknown. The exact theorem supplies an all-dimensional lower bound; general existence remains unresolved. In particular, the finite-dimensional exact constructions cited above do not establish an asymptotic upper bound on $r_{\mathrm{ex}}(d)$. For $1 \le p < 2$, the matching first-layer L_p lower bound also remains unproved here; for $2 \le p < \infty$, remark 2.7 already gives the sharp threshold. Uniform approximation and norm-bounded closure provide further directions.

Appendix A. Tournament certificates

This appendix records the explicit orientations and the finite checks behind Section 3.3. An arc is active when its tail is later in the coordinate order; reversing this convention reverses every order and leaves the minimum unchanged. Vertices are numbered $0, \ldots, n-1$, and bit u of row v's hexadecimal mask is one precisely when $v \to u$; Table 4 stores the seven orientations. The resulting minimum active counts and weighted sums are listed in Table 2.

n	Out-neighbor masks for rows $0, \ldots, n-1$
7	`16, 2c, 58, 31, 62, 45, 0b`
8	`32, 8c, c9, 71, c6, 16, a3, 29`
9	`03a, 1b0, 0a3, 156, 184, 1d8, 017, 049, 0c5`
10	`214, 21d, 328, 3c1, 18c, 21b, 137, 067, 0a3, 1d0`
11	`0cc, 255, 5e0, 476, 485, 153, 790, 22a, 29b, 43d, 1a3`
12	`e52, 0ec, c81, 985, 54e, 05d, c0c, 671, 8e7, 97e, 32a, 4b2`
13	`170a, 095c, 14e9, 0d70, 0aa5, 1443, 1991, 0a2b, 06b4, 106e, 1ad2, 1325, 019a`

Table 4. Stored orientations for the seven tournament certificates. Exact statistics appear in Table 2.

The verifier checks that each unordered pair has exactly one orientation, then evaluates the integer recurrences Equations (3.30) and (3.31). For $n = 7$, independent enumeration of all $7! = 5040$ orders confirms the minimum of seven. All seven stored masks satisfy both required inequalities.

For $14 \le n \le 19$, the proof instead uses an exact integer existence bound. With $N = \binom{n}{2}$, each of the six comparisons is

$$n!\, 2^{-N} \sum_{k=0}^{n-1} \binom{N}{k} < \frac{1}{2} \tag{A.1}$$

checked as $2n! \sum_{k=0}^{n-1} \binom{N}{k} < 2^N$, without floating-point arithmetic. Combining this strict failure bound with the Markov estimate from Equation (3.27) proves existence of an orientation satisfying both desired properties. The argument does not select an orientation in this block-size range.

For $n \ge 20$, Hoeffding's inequality and $n! \le n^n$ give

$$\Pr\{\min_{\pi} X_T(\pi) < n\} \le \exp\left(n \log n - \frac{(n-1)(n-4)^2}{4n}\right) < 1. \tag{A.2}$$

At $n = 20$, use $\log 20 < 3$ to bound the exponent by $-4/5$, so the right-hand side is strictly less than $1/2$. Writing the exponent as $H(n) = n \log n - n^2/4 + 9n/4 - 6 + 4/n$, one has

$$H'(n) = \log n + 13/4 - n/2 - 4/n^2 < 0, \qquad H''(n) = 1/n - 1/2 + 8/n^3 < 0$$

for $n \geq 20$. Thus Equation (A.2) holds for the entire analytic range. Combining it with the same Markov bound gives both simultaneous properties for $n \geq 20$. The exact table values satisfy the same $Z(T)$ bound for $7 \leq n \leq 13$. Thus the explicit masks, exact binomial comparisons, and monotonicity argument together cover every block size $n \geq 7$, including the compressed-decoder requirement.

A.1. **The small-block obstruction.** Write $\alpha_n = \max_T \min_\pi X_T(\pi)$, where the maximum runs over all tournaments on n vertices. Enumerating all $2^{\binom{n}{2}}$ labelled orientations gives

n	2	3	4	5	6
α_n	0	1	1	3	4

The verifier `code/verify_small_tournament_obstruction.py` checks every orientation twice: by Equation (3.30), and independently by comparing its edge bits with those of all transitive orientations. For a fixed order, the number of differing bits is exactly its active count. Thus both checks are exhaustive integer computations.

Every tournament on seven vertices has a vertex with outdegree at most three. Placing it after an optimal order of the remaining six gives $\alpha_7 \leq \alpha_6 + 3 = 7$, while the stored seven-vertex mask attains seven, so $\alpha_7 = 7$. Independent block orders imply that the maximum achievable minimum total active count is $\alpha_{\lceil d/2 \rceil} + \alpha_{\lfloor d/2 \rfloor}$. Its values are

d	5	6	7	8	9	10	11	12	13
$\alpha_{\lceil d/2 \rceil} + \alpha_{\lfloor d/2 \rfloor}$	1	2	2	2	4	6	7	8	11

Each is smaller than d, proving the stated obstruction to the active-row hypothesis for this balanced two-block arrangement.

Appendix B. Three-ray estimate

We prove Theorem 2.3 for the fan in Equation (2.15). The argument has three steps: a compact probe turns each ridge into a rank-one tensor; an explicit matrix minor separates the target tensor from all tensors of rank at most W; and rescaling the probe into the unit disk converts this separation into the claimed L_2 estimate. Tracking the derivative order through the last step is what produces the exponent 10.

B.1. **Tensor localization.** Throughout this proof, $c, C > 0$ denote absolute auxiliary constants that may change between estimates. The resulting fixed constant in Equation (2.16) is denoted by $c_\triangle$.

B.1.1. *A compact probe.* Let $F = F_\triangle$ from Equation (2.15). Its three fan rays are denoted by t_j, and n_j is the corresponding unit normal. The distributional Hessian is

$$D^2 F = \sqrt{3} \sum_{j=0}^{2} n_j n_j^\top \mathcal{H}^1 \restriction_{\mathbb{R}_+ t_j} . \tag{B.1}$$

For a compactly supported C^k function ψ on $\mathbb{R}^2$, define

$$\mathcal{L}_{k,\psi}(h) = \frac{1}{\sqrt{k!}} \int_{\mathbb{R}^2} h(y) D^k \psi(y)\, dy \in \mathrm{Sym}^k(\mathbb{R}^2). \tag{B.2}$$

Lemma B.1 (Tensor probes)**.** *Let $k \geq 2$. For $h \in L_2$,*

$$\|\mathcal{L}_{k,\psi}(h)\|_F \leq \frac{\|D^k \psi\|_{L_2(\mathbb{R}^2; \mathrm{Sym}^k)}}{\sqrt{k!}} \|h\|_2. \tag{B.3}$$

For every affine ridge $r(y) = \rho(w \cdot y + b)$,

$$\mathcal{L}_{k,\psi}(r) = \lambda(w, b) w^{\otimes k} \tag{B.4}$$

for a scalar $\lambda(w, b)$. Hence every matricization of the tensor of a member of $\mathcal{R}_W^{(2)}$ has rank at most W.

Proof. The norm estimate is componentwise Cauchy–Schwarz. If $w = 0$, the ridge is constant and its probe vanishes. For $w \neq 0$, distributionally,

$$D^k \rho(w \cdot y + b) = w^{\otimes k} \rho^{(k)}(w \cdot y + b).$$

Pairing the scalar distribution with ψ leaves only a scalar. Affine terms vanish for $k \geq 2$. No parameter bound is used. □

We use odd orders $k = 2q + 3$. If $u = a_j t_j + b_j n_j$, then

$$\langle D_u^k F, \psi \rangle = -\sqrt{3} \sum_j b_j^2 \int_0^\infty (a_j \partial_t + b_j \partial_s)^{2q+1} \psi(t, 0)\, dt. \tag{B.5}$$

Lemma B.2 (Radial finite-jet localization)**.** *Let $\gamma(y) = (2\pi)^{-1} e^{-\|y\|^2/2}$. If $\psi \in C_c^\infty(\mathbb{R}^2)$ is radial and equals γ near zero, then*

$$\langle D^k F, \psi \rangle = \langle D^k F, \gamma \rangle \tag{B.6}$$

for every odd $k \geq 3$.

Proof. Expand one summand in Equation (B.5) as

$$(a\partial_t + b\partial_s)^{k-2} \psi(t, 0) = \sum_{j=0}^{k-2} \binom{k-2}{j} a^{k-2-j} b^j \partial_t^{k-2-j} \partial_s^j \psi(t, 0).$$

Radiality makes the terms with odd j vanish on $s = 0$. Since $k - 2$ is odd, every surviving term has the positive odd tangential order $m = k - 2 - j$. Compact support (and Gaussian decay for γ) removes the boundary at infinity, so

$$\int_0^\infty \partial_t^m \partial_s^j \psi(t, 0)\, dt = -\partial_t^{m-1} \partial_s^j \psi(0, 0).$$

The derivative on the right has total order $(m - 1) + j = k - 3$. Hence the pairing depends only on the order- at-most-$k - 3$ jet at the origin. Because $\psi = \gamma$ on a neighborhood of zero, these jets agree term by term, proving Equation (B.6). □

B.1.2. *The folded tensor.* Define

$$T_k = \frac{1}{\sqrt{k!}} \langle D^k F, \gamma \rangle \in \mathrm{Sym}^k(\mathbb{R}^2). \tag{B.7}$$

Work in the complexification and set $e_+ = (e_1 - ie_2)/\sqrt{2}$, $e_- = (e_1 + ie_2)/\sqrt{2}$. For $0 \leq p \leq k$, the normalized symmetric basis vector is

$$E_p^{(k)} = \binom{k}{p}^{-1/2} \sum_{\substack{A \subseteq [k] \\ |A| = p}} \bigotimes_{a=1}^{k} f_{\mathbf{1}_{\{a \in A\}}}, \qquad f_1 = e_+, \qquad f_0 = e_-. \tag{B.8}$$

We use the Hermitian inner product that is conjugate-linear in its first slot and linear in its second. These vectors are orthonormal. Put $Z = (X_1 + iX_2)/\sqrt{2}$, whose density is $\pi^{-1} e^{-|z|^2}$, and use the complex Hermite convention

$$H_{p,j}(z, \bar{z}) = (-1)^{p+j} e^{|z|^2} \partial_{\bar{z}}^p \partial_z^j e^{-|z|^2}. \tag{B.9}$$

It satisfies

$$\mathbb{E}\Big[H_{p,j}(Z, \bar{Z}) \overline{H_{p',j'}(Z, \bar{Z})} \Big] = p!\, j!\, \mathbf{1}_{\{p=p',\, j=j'\}}. \tag{B.10}$$

Gaussian integration by parts and Equation (B.8) give, for $j = k - p$,

$$(T_k)_p = \langle E_p^{(k)}, T_k \rangle = \frac{\mathbb{E}[F(X) H_{p,j}(Z, \bar{Z})]}{\sqrt{p! j!}}. \tag{B.11}$$

Lemma B.3 (Folded-binomial coordinates)**.** *For $k = 2q + 3$, after a unitary basis rotation and one common phase,*

$$(T_k)_p = B_k \epsilon_p \sqrt{\binom{k}{p}} \mathbf{1}_{\{2p-k \equiv 0 \pmod 3\}}, \tag{B.12}$$

where

$$B_k = \frac{3\sqrt{6}}{8\pi} \frac{q!}{\sqrt{(2q+3)!}}, \qquad \epsilon_p = (-1)^{\min\{p, k-p\}}. \tag{B.13}$$

Proof. Write $Z = (X_1 + iX_2)/\sqrt{2} = re^{i\theta}$ for a standard Gaussian $X \in \mathbb{R}^2$. Then

$$F(\sqrt{2}Z) = \sqrt{2} r h(\theta), \qquad h(\theta) = \max_j \cos(\theta - 2\pi j/3). \tag{B.14}$$

For nonzero m, direct sector integration gives

$$h_m = \frac{1}{2\pi} \int_0^{2\pi} h(\theta) e^{-im\theta} \, d\theta = \begin{cases} \dfrac{3\sqrt{3}}{2\pi} \dfrac{(-1)^{m/3}}{1-m^2}, & m \in 3\mathbb{Z}, \\ 0, & m \notin 3\mathbb{Z}. \end{cases} \tag{B.15}$$

The profile h is real and even, so $h_{-m} = h_m$. For $p \geq j$, $m = p - j$, use

$$H_{p,j}(z, \bar{z}) = (-1)^j j! z^m L_j^m(|z|^2).$$

When $p + j = k = 2q + 3$,

$$\int_0^\infty r^{m+2} e^{-r^2} L_j^m(r^2) \, dr = \frac{m^2 - 1}{8} \frac{q!}{j!}. \tag{B.16}$$

After $x = r^2$, Rodrigues' formula and j integrations by parts give

$$\frac{1}{2} \int_0^\infty x^{(m+1)/2} e^{-x} L_j^m(x) \, dx = \frac{(-1)^j}{2j!} \left(\frac{1-m}{2} \right)^{\underline{j}} \Gamma\left(\frac{m+3}{2} \right). \tag{B.17}$$

All boundary terms vanish: at infinity this follows from the exponential factor, while at zero the relevant primitive is $O(x^{(m+3)/2})$. This also covers the endpoint $m = 1$. Here $m + 2j = 2q + 3$. Writing $m = 2a + 1$, so $a = q - j + 1$, reduces the right-hand side to $a(a+1)q!/(2j!) = (m^2 - 1)q!/(8j!)$, proving Equation (B.16). This cancels the denominator in Equation (B.15). With the Gaussian and Hermite normalizations,

$$\frac{\mathbb{E}[F(X) H_{p,j}(Z, \bar{Z})]}{\sqrt{p! j!}} = -\frac{3\sqrt{6}}{8\pi} \frac{q!}{\sqrt{p! j!}} (-1)^{j + m/3} \tag{B.18}$$

when $m \in 3\mathbb{Z}$, and it is zero otherwise. A rotation by $\pi/3$ removes $(-1)^{m/3}$. Finally, $1/\sqrt{p!(k-p)!} = \sqrt{\binom{k}{p}}/\sqrt{k!}$; folding the two sides of the center yields Equations (B.12) and (B.13). $\square$

As a low-order normalization check, the coordinate vectors in the orthonormal basis $(E_p^{(k)})_{p=0}^k$, after the same rotation and common phase as in Lemma B.3, are

$$\frac{T_3}{B_3} = (1, 0, 0, 1), \qquad \frac{T_5}{B_5} = (0, -\sqrt{5}, 0, 0, -\sqrt{5}, 0), \qquad \frac{T_7}{B_7} = (0, 0, \sqrt{21}, 0, 0, \sqrt{21}, 0, 0). \tag{B.19}$$

Thus the phase, support congruence, and square-root binomial factor can be checked directly at $k = 3, 5, 7$. With $r = q + 1$ and $s = q + 2$, for each supported coordinate p, the squared coefficients in its image under Equation (B.20) sum to $\binom{k}{p}^{-1}\sum_{i+j=p}\binom{r}{i}\binom{s}{j} = 1$, which independently checks that the catalecticant map below is isometric on these examples.

B.1.3. *A central minor.* Put $r = q + 1$, $s = q + 2$, so $r + s = k$. In normalized symmetric bases, grouping the first r and last s tensor slots defines the comultiplication $\Delta_{r,s} : \mathrm{Sym}^k \to \mathrm{Sym}^r \otimes \mathrm{Sym}^s$. With the normalization in Equation (B.8), one has exactly

$$\Delta_{r,s} E_p^{(k)} = \sum_{\substack{0\le i\le r,\ 0\le j\le s\\ i+j=p}} \sqrt{\frac{\binom{r}{i}\binom{s}{j}}{\binom{k}{p}}}\, E_i^{(r)} \otimes E_j^{(s)}. \tag{B.20}$$

The coefficient counts the binary strings of weight p having weights i and j in the two blocks. Vandermonde's identity says that the squared coefficients in Equation (B.20) sum to one; images corresponding to distinct p are orthogonal. Thus $\Delta_{r,s}$ is an isometry for the Frobenius norms, with no omitted factorial or binomial rescaling. Direct expansion in the normalized bases gives $\Delta_{r,s}(v^{\otimes k}) = v^{\otimes r} \otimes v^{\otimes s}$. Consequently, the matrix of $\Delta_{r,s} T_k$ is

$$M_{ij} = (T_k)_{i+j} \sqrt{\frac{\binom{r}{i}\binom{s}{j}}{\binom{k}{i+j}}}. \tag{B.21}$$

In particular, the normalized Hermite coordinates in Equations (B.11) and (B.12) pass to this catalecticant without a change of norm. By Equation (B.12),

$$M_{ij} = B_k \epsilon_{i+j} \sqrt{\binom{r}{i}\binom{s}{j}}\,\mathbf{1}_{\{i+j\equiv r-1 \pmod 3\}}. \tag{B.22}$$

Lemma B.4 (Catalecticant tail)**.** *For every $W \ge 1$, one can choose $q \asymp W^2$ so that*

$$\inf_{\operatorname{rank} R\le W} \|M - R\|_F \ge cW^{-3}. \tag{B.23}$$

Proof. We select a central square minor whose sign pattern has all singular values bounded below, and then control the two binomial weight factors. Take $n = 2W + 2$, fix $C_0 = 64$, choose $q = \lceil C_0 n^2 \rceil$, and let

$$L = r - 1 - 3(n - 1), \qquad i_0 = \lfloor L/2 \rfloor, \qquad j_0 = L - i_0.$$

Select rows $i_p = i_0 + 3p$ and columns $j_\ell = j_0 + 3\ell$, $0 \le p, \ell < n$. Since $n \ge 4$ and $q = 64n^2$, these indices lie in $[0, r]$ and $[0, s]$, respectively, and are within $2n$ of their respective centers. They satisfy the congruence in Equation (B.22), and $i_p + j_\ell \le r$ exactly when $p + \ell \le n - 1$. Writing $u = p + \ell$, the selected sign is

$$\epsilon_{L+3u} = (-1)^{L+u} \begin{cases} 1, & u \le n - 1, \\ -1, & u > n - 1, \end{cases}$$

because the definition of $\epsilon_t = (-1)^{\min(t,k-t)}$ switches at $t = r$, and k is odd. Multiplying row p by $(-1)^{L+p}$ and column ℓ by $(-1)^\ell$ therefore leaves $+1$ on $p + \ell \le n - 1$ and -1 on its complement. Reversing the columns, $\ell \mapsto n - 1 - \ell$, turns the first condition into $p \le \ell$. Thus the sign matrix is

$$S_{p\ell} = \begin{cases} 1, & p \le \ell, \\ -1, & p > \ell. \end{cases} \tag{B.24}$$

Writing $S = I + K$ gives $K^* = -K$, hence $S^*S = I - K^2 \succeq I$. Every singular value of S is at least one.

All selected binomial indices are within $O(n)$ of their centers. Thus, for $N \in \{r, s\}$ and every selected index u, a local Stirling bound gives

$$\binom{N}{u} \geq c\frac{2^N}{\sqrt{N}}\exp\left(-C\frac{(u-N/2)^2}{N}\right) \geq c2^N q^{-1/2}, \tag{B.25}$$

because $|u - N/2| = O(n) = O(\sqrt{q})$. Also $q!2^q/\sqrt{(2q)!} \asymp q^{1/4}$, while

$$(2q+3)! = (2q)!(2q+1)(2q+2)(2q+3).$$

Together with Equation (B.13), these estimates imply

$$\min_{p,\ell}\sqrt{\binom{r}{i_p}\binom{s}{j_\ell}} \geq c2^{k/2}q^{-1/2}, \qquad |B_k|2^{k/2}q^{-1/2} \geq cq^{-7/4}. \tag{B.26}$$

Thus the selected minor is $B_k D_1 S D_2$, up to signs, and each of its n singular values is at least $cq^{-7/4}$. Restriction preserves a rank-W upper bound. Eckart–Young [8] therefore yields

$$\inf_{\operatorname{rank} R \leq W} \|M - R\|_F \geq c\sqrt{n-W}\, q^{-7/4} \geq cW^{-3},$$

because $q \asymp W^2$. □

Complexification is a relaxation: every real rank-W tensor is a complex rank-W tensor, and all basis changes above are unitary.

B.1.4. *Compact localization.*

Lemma B.5 (Radial cutoff). *For every $k \geq 3$, there is a radial C_c^∞ function χ_k and absolute constants A, C, with $R = A\sqrt{k}$, such that*

$$\begin{gathered} \chi_k = 1 \text{ on } B_R, \qquad \operatorname{supp}(\chi_k) \subset B_{2R}, \\ \|D^j\chi_k\|_\infty \leq (Ck/R)^j \quad (0 \leq j \leq k). \end{gathered} \tag{B.27}$$

For $\psi_k = \gamma\chi_k$,

$$\frac{1}{k!}\int \|D^k\psi_k(y)\|_F^2\, dy \leq C. \tag{B.28}$$

Proof. We first construct a cutoff with controlled derivatives through order k. Gaussian decay then bounds every term introduced by the cutoff, including the full Leibniz sum, on the scale $\sqrt{k!}$. Let $\rho_\epsilon = |B_\epsilon|^{-1}\mathbf{1}_{B_\epsilon}$, take $m = 2k+2$, $\epsilon = R/(16m)$, and set

$$\beta = \rho_\epsilon^{*m} * \zeta_\epsilon, \qquad \chi_k = \mathbf{1}_{B_{3R/2}} * \beta, \tag{B.29}$$

where ζ_ϵ is a radial smooth probability density supported in B_ϵ. The support of β lies in $B_{R/8}$, proving the plateau and support assertions. For a unit direction v, $D_v\rho_\epsilon$ is a signed boundary measure of total variation at most $2/\epsilon$. Put the j ordered derivatives on j distinct disk factors in Equation (B.29); Young's inequality gives $\|D^j\chi_k\|_\infty \leq (2/\epsilon)^j \leq (Ck/R)^j$. Passing from ordered directional derivatives to the Frobenius tensor norm in dimension two costs at most another factor C^j, absorbed into the same absolute constant.

For the uncut Gaussian, Parseval gives

$$\int_{\mathbb{R}^2}\|D^k\gamma\|_F^2 = \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}\|\xi\|^{2k}e^{-\|\xi\|^2}\,d\xi = \frac{k!}{4\pi}. \tag{B.30}$$

We next prove the derivative-tail estimate, including its dependence on the derivative order. Define the probabilists' multivariate Hermite tensor

$$\mathsf{H}_j(y) = (-1)^j\gamma(y)^{-1}D^j\gamma(y).$$

If G is a standard Gaussian vector in $\mathbb{R}^2$, then, in the complexification of $\operatorname{Sym}^j(\mathbb{R}^2)$,

$$\mathsf{H}_j(y) = \mathbb{E}\left[(y+iG)^{\otimes j}\right]. \tag{B.31}$$

To verify this normalization, contract both sides against $u^{\otimes j}$. The exponential generating function of the left-hand side is

$$\frac{\gamma(y-tu)}{\gamma(y)} = \exp\left(t\langle y,u\rangle - \frac{t^2\|u\|^2}{2}\right),$$

which is also $\mathbb{E}\exp(t\langle y+iG,u\rangle)$; equality of the homogeneous coefficients and polarization give Equation (B.31).

The complexified Frobenius norm is a cross norm on pure tensors. Hence, for $j \geq 1$, Minkowski's inequality gives

$$\begin{aligned}\|\mathsf{H}_j(y)\|_F &\leq \mathbb{E}\|y+iG\|^j \leq \mathbb{E}(\|y\|+\|G\|)^j \\ &\leq \left(\|y\| + \left(\mathbb{E}\|G\|^j\right)^{1/j}\right)^j \leq (\|y\|+C\sqrt{j})^j. \end{aligned} \tag{B.32}$$

Here the last inequality is elementary in dimension two: polar integration gives $\mathbb{E}\|G\|^j = 2^{j/2}\Gamma(1+j/2) \leq (C\sqrt{j})^j$. The case $j = 0$ is immediate. Multiplication by $\gamma(y)$ now proves, uniformly for $0 \leq j \leq k$,

$$\|D^j\gamma(y)\|_F \leq C^{j+1}(\|y\|+\sqrt{j})^j e^{-\|y\|^2/2}. \tag{B.33}$$

Take $A \geq \sqrt{2}$. On $\|y\| \geq R = A\sqrt{k}$ and for $j \leq k$, one has $\sqrt{j} \leq \|y\|/A$. Squaring Equation (B.33) and using polar coordinates therefore gives

$$\begin{aligned}\|\mathbf{1}_{\{\|y\|\geq R\}}D^j\gamma\|_2^2 &\leq C(C')^{2j}\int_R^\infty r^{2j+1}e^{-r^2}\,dr, \\ \int_R^\infty r^{2j+1}e^{-r^2}\,dr &= \frac{1}{2}e^{-R^2}\int_0^\infty (R^2+s)^j e^{-s}\,ds \leq R^{2j}e^{-R^2}. \end{aligned} \tag{B.34}$$

Here $(R^2+s)^j \leq R^{2j}e^{js/R^2}$ and $j/R^2 \leq A^{-2} \leq 1/2$, so the last integral is at most $2R^{2j}$, which cancels the prefactor $1/2$. Taking square roots yields, with absolute constants independent of j, k,

$$\|\mathbf{1}_{\{\|y\|\geq R\}}D^j\gamma\|_2 \leq C(CR)^j e^{-R^2/2}. \tag{B.35}$$

For completeness, the norms in Equations (B.33) and (B.35) are the Frobenius norms of the full ordered derivative tensors; thus no multi-index count is suppressed.

It remains to track the Leibniz sum. Since $0 \leq \chi_k \leq 1$, $\chi_k - 1$ vanishes on B_R, and every positive-order derivative of χ_k also vanishes there. Symmetrization is an orthogonal projection for the Frobenius norm, so tensor Leibniz, Equation (B.27), and Equation (B.35) give

$$\begin{aligned}\|D^k((\chi_k-1)\gamma)\|_2 &\leq \sum_{\ell=0}^k \binom{k}{\ell}\left(\frac{Ck}{R}\right)^\ell C(CR)^{k-\ell}e^{-R^2/2} \\ &\leq C\big(C(R+k/R)\big)^k e^{-R^2/2} \leq C(CA\sqrt{k})^k e^{-A^2k/2}. \end{aligned} \tag{B.36}$$

For $\ell = 0$, the displayed cutoff factor means the bound $\|\chi_k - 1\|_\infty \leq 1$. Because $\sqrt{k!} \geq (k/e)^{k/2}$, the ratio of the last expression to $\sqrt{k!}$ is at most

$$C\big(CA\sqrt{e}\,e^{-A^2/2}\big)^k.$$

Fixing one sufficiently large absolute A makes this uniformly bounded for all $k \geq 3$. Thus

$$\|D^k((\chi_k-1)\gamma)\|_2 \leq C\sqrt{k!}. \tag{B.37}$$

The triangle inequality and Equation (B.30) now prove Equation (B.28). □

Completion of Theorem 2.3. By Lemma B.2, the cut and uncut Gaussian pairings agree. Distributional integration by parts therefore gives

$$\mathcal{L}_{k,\psi_k}(F) = (-1)^k T_k. \tag{B.38}$$

This global sign has no effect on any singular value. Choose a fixed $0 < \delta < 1$, put

$$\sigma = \frac{\delta}{2R}, \qquad \widetilde{\psi}_k(y) = \psi_k(y/\sigma), \tag{B.39}$$

Then $\operatorname{supp}(\widetilde{\psi}_k) \subset B_\delta \Subset \mathbb{B}^2$, so the probe remains centered at the fan vertex and no translation is used. Consequently, the final Cauchy–Schwarz step below uses only $L_2(B_\delta)$: no global ridge is assumed to belong to $L_2(\mathbb{R}^2)$. Scaling in two dimensions gives

$$\frac{\|D^k\widetilde{\psi}_k\|_2}{\sqrt{k!}} \le C\sigma^{1-k}, \qquad \frac{\langle D^k F, \widetilde{\psi}_k\rangle}{\sqrt{k!}} = \sigma^{3-k} T_k. \tag{B.40}$$

For $W \ge 1$, take the same $q = \lceil C_0(2W+2)^2\rceil$ as in Lemma B.4, set $k = 2q+3$, and use the corresponding probe above. Combining Lemmas B.1 and B.4 and Equation (B.40) gives, for every $g \in \mathcal{R}_W^{(2)}$,

$$\|g - F\|_2 \ge c\sigma^2(W+1)^{-3}. \tag{B.41}$$

Here $R \asymp W+1$, hence $\sigma \asymp (W+1)^{-1}$, and $\|g - F\|_2 \ge c(W+1)^{-5}$. Squaring proves Equation (2.16). Since $\mathcal{R}_0^{(2)} \subseteq \mathcal{R}_1^{(2)}$, the case $W = 0$ follows from the width-one bound after decreasing the absolute constant. □

Appendix C. Small dimensions

The construction in Section 6 gives the general upper bound. Here we give the smaller standard networks used in dimensions five and six. All displayed linear channels are nonnegative on the cube and are therefore valid ReLU outputs.

C.1. Dimension five.

On $[0,1]^5$, use the six first outputs

$$\begin{aligned} u_3 &= x_3, & u_5 &= x_5,\\ h_{12} &= \rho(x_1 - x_2), & h_{13} &= \rho(x_1 - x_3),\\ h_{23} &= \rho(x_2 - x_3), & h_{45} &= \rho(x_4 - x_5). \end{aligned}$$

Define the following linear combinations of those outputs:

$$B = u_5 + h_{45}, \quad P_1 = u_3 - u_5 + h_{13} - h_{45}, \quad P_2 = u_3 - u_5 + h_{23} - h_{45}.$$

Then the four-neuron second layer and affine output

$$N_\Lambda = \rho(B) + \rho(P_1) + \rho(P_2 - \Lambda h_{12}) - \rho(P_1 - \Lambda h_{12}) \tag{C.1}$$

equal MAX_d when $x_1 \le x_2$. When $x_1 > x_2$, the last two terms vanish in the limit and the first two terms give MAX_d. The gated terms are dominated by their ungated ReLUs, so convergence holds in every finite L_p. This proves $w_{\mathrm{cl}}(5) \le 6$.

C.2. Dimension six.

Use $u_3 = x_3, u_6 = x_6$ and all three within-triple comparison hinges in each of $\{1,2,3\}$ and $\{4,5,6\}$. Thus the first width is eight. Set

$$h = \rho(x_1 - x_2), \quad A_1 = \max(x_1, x_3), \quad A_0 = \max(x_2, x_3),$$

$$k = \rho(x_4 - x_5), \quad B_1 = \max(x_4, x_6), \quad B_0 = \max(x_5, x_6),$$

where every A_r, B_r is linear in the eight first outputs. Abbreviate $[P]_{D,\Lambda} = \rho(P - \Lambda D)$. The ten second neurons enter as

$$\begin{aligned} N_\Lambda = A_1 + \rho(B_1 - A_1) - [A_1 - B_1]_{h,\Lambda} + [A_0 - B_1]_{h,\Lambda} \\ - [B_1 - A_1]_{k,\Lambda} + [B_0 - A_1]_{k,\Lambda} + [B_1 - A_1]_{h+k,\Lambda} - [B_1 - A_0]_{h+k,\Lambda} \\ - [B_0 - A_1]_{h+k,\Lambda} + [B_0 - A_0]_{h+k,\Lambda}. \end{aligned} \tag{C.2}$$

The apparently affine term A_1 is the output of the legal neuron $\rho(A_1) = A_1$. In the four states determined by $h = 0$ or $h > 0$ and $k = 0$ or $k > 0$, the limit of Equation (C.2) is respectively the maximum of the selected pair among A_0, A_1 and B_0, B_1, hence MAX_6. Every second output lies in $[0, 1]$, so dominated convergence applies and $w_{\mathrm{cl}}(6) \leq 10$.

Combining the small constructions with Equation (2.32) for the general construction proves Theorem 2.12.

Appendix D. Exact width

This section proves Theorem 2.13.

We use the following elementary obstruction. A finite one-hidden-layer ReLU function in three variables is affine off a finite union of complete affine hyperplanes. After terms with the same supporting hyperplane are combined, its distributional Hessian has one constant matrix-valued density on each complete hyperplane, away from its intersections with the other hyperplanes. For distinct i, j, k, the Hessian density of MAX_3 on $\{x_i = x_j\}$ is nonzero on the open half

$$x_i = x_j > x_k$$

and zero on the opposite open half $x_i = x_j < x_k$. To compare the two measures, choose one point in each half outside the finitely many intersections with the other ridge hyperplanes, and test in sufficiently small neighborhoods of those points. Ridge hyperplanes distinct from $\{x_i = x_j\}$ make no contribution there. The coincident ridge terms would therefore need the same constant density to be both nonzero and zero, which is impossible. Consequently

$$\mathrm{MAX}_3 \notin \left\{ a_0 + a^\top x + \sum_{j=1}^{W} c_j \rho(v_j^\top x + b_j) : W < \infty \right\}. \tag{D.1}$$

This argument permits arbitrary real coefficients, directions, and biases.

The proof first pairs affine channels without charging them as nonlinear cores, then charges missing pair supports to support-three and support-four cores through triangles and centered induced claws. The next two subsections give the normalization and counting arguments.

D.1. **Paired normalization.** An arbitrary exact network need not be homogeneous, and a ReLU that is strictly active near a diagonal point passes its preactivation there. A literal row count after local homogenization can therefore overstate the number of nonlinear resources by a factor of two. We isolate this degeneracy exactly.

Definition D.1 (Paired-core network)**.** At hidden layer ℓ, a *paired channel* consists of

$$\rho(r_{\ell b}), \qquad \rho(-r_{\ell b}) \tag{D.2}$$

whose two coefficients in every preactivation of the next layer are $(v, -v)$. The pair contributes

$$v\rho(r_{\ell b}) - v\rho(-r_{\ell b}) = v r_{\ell b}. \tag{D.3}$$

Every unpaired hidden neuron is called a *core*. The effective width is $w_{\mathrm{eff}} = \max_\ell |C_\ell|$, where C_ℓ is the set of cores at layer ℓ.

The pair in Equation (D.2) is a standard CReLU object and Equation (D.3) is classical [33, 23, 26]. What matters here is the downstream opposite-column invariant: complete-row duplication at a later layer preserves every earlier pair simultaneously.

Call a global homogeneous paired-core network *admissible* for MAX_d if every first-core normal w satisfies

$$\mathbf{1} \cdot w = 0, \qquad |\operatorname{supp}(w)| \geq 2. \tag{D.4}$$

Proposition D.2 (Paired normalization). *Suppose an ordinary depth-k ReLU network of width W agrees with MAX_d on $[0,1]^d$. Then there is an admissible global homogeneous paired-core network of the same depth, computing MAX_d, with effective width at most W. For $k = 3$, the number m of first-layer cores is at most the original first hidden width m_1.*

Proof. Choose $c \in (0,1)$ outside the finite set of isolated zeros of the non-identically-zero affine functions

$$\beta_i(c) = b_i + c\,\mathbf{1} \cdot w_i \tag{D.5}$$

in the original first layer. Translate $x = y + c\mathbf{1}$ and subtract c at the output. The translated network agrees with $\mathrm{MAX}_d(y)$ on a neighborhood of zero.

Process the hidden layers from input to output. When a row is processed, all earlier hidden biases have already been removed, so its preactivation has the form $r(h) + b$, with $r(h) \to 0$ as $h \to 0$. A row with $b < 0$ is inactive on a sufficiently small common neighborhood and is deleted. A row with $b = 0$ is retained as one core. If $b > 0$, replace the locally affine row by $\rho(r), \rho(-r)$. If its old outgoing coefficient was v, use $v, -v$ and add vb to the receiving bias. Locally,

$$v\rho(r) - v\rho(-r) + vb = v(r + b) = v\rho(r + b). \tag{D.6}$$

Suppose a later complete row has entries $(a, -a)$ on an existing pair of columns. Duplicating that row as $r, -r$ creates entries $(a, -a)$ and $(-a, a)$; both remain in the opposite-column subspace. Thus every previously created pair remains paired through all later processing. Each core descends from one original row in its layer, so no layer contains more than W cores.

This is an effective-core count, not a claim that the displayed normalized network has at most W physical ReLU rows. A locally affine original row may be replaced by two opposite ReLUs. Those auxiliary rows transmit a linear form in the preceding-layer outputs and are excluded from the nonlinear-core count; retained cores still inject into the original rows. In particular, for depth three their number in the first layer is at most the original m_1.

All hidden biases have now vanished. Equality at the origin forces the output bias to vanish as well. The constructed function and MAX_d are positively homogeneous and agree near zero, so the equality is global.

A first-layer row is retained only if $\beta_i(c) = 0$. Genericity of c then forces Equation (D.5) to vanish identically, hence $b_i = 0$ and $\mathbf{1} \cdot w_i = 0$. Delete a zero vector. A nonzero zero-sum vector cannot have singleton support, proving Equation (D.4). □

We next make coordinate restrictions without accidentally creating new first-layer nonlinear resources.

Lemma D.3 (Retained-support principle). *Let an admissible paired-core representation of MAX_d with two hidden layers have first-core supports $S_1, \ldots, S_m$. For every $I \subseteq [d]$, $|I| = r \geq 3$, there is an admissible paired-core representation of MAX_r with two hidden layers and at most*

$$N_I = \#\{a : S_a \subseteq I\} \tag{D.7}$$

first-layer cores. Every such core is the restriction of an original core whose support is contained in I.

Proof. Order the coordinates outside I as $j_1, \ldots, j_q$, fix

$$x_{j_\ell} = -R^\ell, \tag{D.8}$$

and let the coordinates in I range over a bounded positive box. For a core whose support is not contained in I, its frozen contribution is a nonzero polynomial in R. The term with largest ℓ such that $w_{j_\ell} \neq 0$ is its leading term. Because only finitely many cores occur, one sufficiently large common R fixes the sign of every such preactivation throughout the box. The corresponding ReLU is affine or zero. Cores supported inside I remain homogeneous, and the target is MAX_r because the frozen coordinates lie below the box.

Translate an interior diagonal point of the box to zero. Retained zero-sum cores acquire no bias. Collapse the affine first-layer outputs into the second layer, absorbing their constant terms into its biases. Represent each required linear form as an opposite ReLU pair $\rho(r), \rho(-r)$, keep every existing pair through Equation (D.3), and apply the three local cases from Proposition D.2 to the former second layer. Complete-row duplication preserves pairedness. It can add auxiliary physical rows but creates no new first core. Local equality globalizes by homogeneity, proving Equation (D.7). □

Corollary D.4 (Triple coverage)**.** *In every admissible paired-core representation of* MAX_d *with two hidden layers, each three-set* $I \subseteq [d]$ *contains a first-core support.*

Proof. Otherwise Lemma D.3 would remove the first nonlinear layer completely. In each second-layer preactivation, the two terms of every remaining first-layer pair combine to one linear form by Equation (D.3); affine terms created by restriction can likewise be absorbed into that preactivation. This leaves a finite one-hidden-layer representation of MAX_3, in contradiction with Equation (D.1). □

The relative version of the same normalization can also be iterated after freezing external coordinates. The present paper needs only the two-hidden-layer statement above.

D.2. **Support rigidity.** We now prove Theorem 2.13. Work with the admissible homogeneous paired-core representation supplied by Proposition D.2, and let $m \leq m_1$ be the number of its first-layer cores. For each core $\rho(w_a \cdot x)$, write $S_a = \mathrm{supp}(w_a)$.

Let $P \subseteq \binom{[d]}{2}$ be the set of *distinct* two-coordinate supports occurring among the S_a, and define the missing-pair graph

$$H = K_d \setminus P, \qquad p = |P|, \qquad e = |E(H)|, \qquad s = m - p. \tag{D.9}$$

Thus s counts every core of support at least three and every duplicate pair support beyond its first occurrence, and

$$m = \binom{d}{2} - e + s. \tag{D.10}$$

The problem is to charge missing pair supports to the surplus s.

D.2.1. *Coordinate descent.* The following elementary distributional identity prevents an unlimited second hidden layer from hiding a missing coordinate inside arbitrary signed cancellations.

Lemma D.5 (First atomic moment)**.** *Let V be a vector space of continuous functions of y. Suppose*

$$\max\{t, M(y)\} = A(y)t + B(y) + \sum_{\nu=1}^{N} \alpha_\nu \rho(a_\nu t + v_\nu(y)), \qquad v_\nu \in V, \tag{D.11}$$

for all $t \in \mathbb{R}$ and all y, where A, B are continuous, $N < \infty$, and $\alpha_\nu, a_\nu \in \mathbb{R}$ are constants. Then $M \in V$.

Proof. Fix y and take the second distributional derivative with respect to t. The affine term disappears and one obtains an equality of finite signed atomic measures,

$$\delta_{M(y)} = \sum_{\nu:a_\nu\neq 0} \alpha_\nu |a_\nu| \delta_{-v_\nu(y)/a_\nu}. \tag{D.12}$$

Taking first moments gives

$$M(y) = -\sum_{\nu:a_\nu\neq 0} \alpha_\nu \operatorname{sgn}(a_\nu) v_\nu(y) \in V. \tag{D.13}$$

Formally, one may test Equation (D.12) against a compactly supported smooth function that agrees with t on an interval containing all atoms. Because y was arbitrary, Equation (D.13) holds pointwise. □

Every triangle of H forces a core whose support is exactly that triangle: all three pair supports are missing, while Corollary D.4 forces some support inside the three-set and admissibility excludes supports of size at most one. Affine terms and paired channels do not supply an alternative nonlinear resource: they combine to affine first-layer features, which cannot by themselves realize MAX_3 after one further hidden layer. Distinct triangles therefore force distinct support-three cores. Triangle counting alone is insufficient for a linear-deficit theorem; the additional constraint is centered.

For an induced claw $C = (v; \{a, b, c\})$ of H, with center v and leaf set $\{a, b, c\}$, call a first-core support S_j a *center witness* for C if

$$v \in S_j \subseteq \{v, a, b, c\}, \qquad |S_j| \in \{3, 4\}. \tag{D.14}$$

Lemma D.6 (Centered induced-claw witness)**.** *Let $I = \{v, a, b, c\}$ induce a claw in H, with center v. Then some first-core support is a center witness for this claw.*

Proof. Restrict the paired-core representation to I by Lemma D.3. Put $t = x_v$ and $y = (x_a, x_b, x_c)$, and suppose that no retained core contains v. Every retained nonlinear first feature then depends only on y; paired channels transmit only linear forms. Let

$$V = \operatorname{span}\big(\{1, y_a, y_b, y_c\} \cup \{\rho(w_j \cdot y) : j \text{ is a retained core}\}\big). \tag{D.15}$$

After absorbing second-layer paired channels into a term affine in t, every remaining second-layer preactivation has the form $a_\nu t + v_\nu(y)$, with $v_\nu \in V$. Here a_ν is constant because only linear first-layer channels depend on t; all retained nonlinear features depend on y alone. The affine remainder has zero second derivative in t, so neither it nor the paired channels can contribute an additional atom to the measure in Equation (D.12). Applying Lemma D.5 to

$$\max\{t, \mathrm{MAX}_3(y)\} \tag{D.16}$$

would imply $\mathrm{MAX}_3 \in V$. This contradicts the one-hidden-layer obstruction Equation (D.1).

Hence a retained core contains v. It is not a singleton by admissibility. None of the three center–leaf pairs can be its support, because these are edges of H, hence absent from P. Its support is therefore a center-containing three- or four-set. □

For $v \in [d]$, let $d_v = \deg_H(v)$, let $t_v = |E(H[N_H(v)])|$, and let κ_v be the number of independent three-sets in $H[N_H(v)]$. Equivalently, κ_v is the number of induced claws of H centered at v. Denote the number of triangles of H by $\tau(H)$.

Proposition D.7 (Center-sensitive resource certificate)**.** *The surplus in Equation* (D.9) *obeys*

$$s \geq \tau(H) + \sum_{v:d_v\geq 3} \frac{\kappa_v}{d_v - 2}. \tag{D.17}$$

Proof. Reserve one forced support-three core for every triangle of H. Such a core cannot be a center witness for an induced claw, because the claw's four-set is triangle-free.

For each induced claw, choose one center witness supplied by Lemma D.6. If a chosen support-three witness is $S = \{v, a, b\}$, then $H[S]$ is the two-edge path with unique center v: $va, vb \in E(H)$, while $ab \notin E(H)$ because a, b are leaves of the same induced claw. Consequently this support cannot witness a claw with another center, and it belongs to at most $d_v - 2$ claws centered at v, one for each choice of the third leaf. A support-four witness fixes its four-set; if that four-set induces a claw, its degree-three vertex is the unique center. It therefore witnesses only that claw and has capacity one, which is no larger than $d_v - 2$. Thus the sets of eligible witnesses for different centers are disjoint. At least $\kappa_v/(d_v - 2)$ distinct non-triangle cores are needed at each center v, and adding the reserved triangle cores proves Equation (D.17). □

D.2.2. *A graph inequality.* The certificate becomes near-complete only after applying an extremal inequality separately inside every neighbor graph.

Lemma D.8 (Edges versus independent triples)**.** *If a graph J has $r \geq 3$ vertices, q edges, and i independent three-sets, then*

$$\frac{q}{3} + \frac{i}{r-2} \geq \frac{r}{2} - \frac{4}{3}. \tag{D.18}$$

Proof. Every non-independent triple contains an edge, and every edge is contained in $r - 2$ triples. Hence

$$i \geq \binom{r}{3} - q(r-2), \qquad i \geq 0. \tag{D.19}$$

It follows that

$$\frac{q}{3} + \frac{i}{r-2} \geq \max\left\{\frac{q}{3}, \frac{r(r-1)}{6} - \frac{2q}{3}\right\} \geq \frac{r(r-1)}{18}. \tag{D.20}$$

For $r = 3, 4$ and $r \geq 6$, this is at least the right-hand side of Equation (D.18), since

$$\frac{r(r-1)}{18} - \left(\frac{r}{2} - \frac{4}{3}\right) = \frac{(r-4)(r-6)}{18} \geq 0. \tag{D.21}$$

For $r = 5$, Equation (D.18) is equivalent, by integrality, to $q + i \geq 4$. This is immediate for $q \geq 4$; for $q \leq 3$, Equation (D.19) gives $q + i \geq 10 - 2q \geq 4$. □

Proof of Theorem 2.13. Each triangle of H is counted once in $H[N_H(v)]$ at each of its three vertices, so

$$\tau(H) = \frac{1}{3} \sum_v t_v. \tag{D.22}$$

The right-hand side of Equation (D.17) is therefore

$$B(H) = \sum_v \frac{t_v}{3} + \sum_{v : d_v \geq 3} \frac{\kappa_v}{d_v - 2}. \tag{D.23}$$

When $d_v \geq 3$, apply Lemma D.8 to $J = H[N_H(v)]$, with $r = d_v$, $q = t_v$, and $i = \kappa_v$. When $d_v < 3$, the local contribution in Equation (D.23) is nonnegative, whereas $d_v/2 - 4/3 \leq 0$. Summing the vertexwise inequalities gives

$$B(H) \geq \sum_v \left(\frac{d_v}{2} - \frac{4}{3}\right) = e - \frac{4d}{3}. \tag{D.24}$$

Combining Equations (D.10), (D.17) and (D.24) yields

$$m \geq \binom{d}{2} - \frac{4d}{3}. \tag{D.25}$$

Because m and $\binom{d}{2}$ are integers,

$$m \geq \binom{d}{2} - \left\lfloor \frac{4d}{3} \right\rfloor. \tag{D.26}$$

Finally, Proposition D.2 gives $m \leq m_1$. □

Remark D.9 (Sharpness of the graph relaxation)**.** If G is a simple cubic triangle-free graph and $H = L(G)$, then H is four-regular and every neighbor graph is $2K_2$. With $d = |V(H)| = |E(G)|$,

$$e(H) = 2d, \qquad \tau(H) = \frac{2d}{3}, \qquad \kappa_v = 0, \tag{D.27}$$

and hence $e(H) - B(H) = 4d/3$. The standard smallest example is $L(K_{3,3})$, on nine vertices. This shows that $4/3$ is sharp for the passage from Equation (D.17) to the scalar width bound. It does not produce neural coefficients and is not a neural sharpness claim.

There is also a finite-dimensional comparison with the exact constructions of Rueß et al. [27] for $4 \leq d \leq 12$ and Wang–Basu [38] for $4 \leq d \leq 8$. Their pairwise maxima can share the first-layer features $\rho(x_i - x_j)$, $i < j$, using $\max\{x_i, x_j\} = x_j + \rho(x_i - x_j)$, with coordinate linear forms transmitted through paired channels. Thus at most $\binom{d}{2}$ first-layer cores suffice in those constructions, with at most $2d$ additional physical rows for the linear channels on $\mathbb{R}^d$. For the dimensions covered, the difference between this core count and our lower bound is $\lfloor 4d/3 \rfloor$. These bounded-dimensional constructions do not establish an asymptotic $O(d)$ gap for exact two-hidden-layer realization: existence in all dimensions remains open.

Appendix E. Reproducibility

The threshold proof uses stored masks, exact integer existence bounds, and analytic estimates in the block-size ranges collected in Table 5. Once the tournaments are chosen, the network construction and its limiting identities are analytic. Separate regression calculations check formulas already proved in the text.

Evidence	Range	Verification
Small-tournament obstruction	$2 \leq n \leq 6$	Exhaustive integer checks
Tournament masks	$7 \leq n \leq 13$	Subset recurrences
Binomial comparisons	$14 \leq n \leq 19$	Integer arithmetic
Tournament existence	$n \geq 20$	Hoeffding bound
Small networks	$d = 5, 6$	Explicit formulas
Tensor identities	Selected orders	Rational arithmetic

Table 5. Finite and analytic inputs. Here n is a tournament block size. The first row explains a limitation of the balanced construction; the last two rows check formulas whose proofs are given in the text.

E.1. **Finite predicates.** For $7 \leq n \leq 13$, the masks in Table 4 specify every orientation. The verifiers check that each pair is oriented exactly once and that Equations (3.30) and (3.31) give

$$D([n]) \geq n, \qquad Z(T) < 2n! \left(\frac{3}{2}\right)^{n-1}.$$

Both statistics appear in Table 2; verification does not require reproducing a search procedure.

For $14 \leq n \leq 19$, the proof uses the exact integer comparisons in Equation (A.1); for $n \geq 20$, it uses Equation (A.2). In each case the Markov bound for $Z(T)$ yields simultaneous existence. The dimension boundaries, including the mixed cases $d = 27, 39$, are specified in Section 3.3.

E.2. **Verification scope.** The exhaustive check in Appendix A.1 certifies the limitation of the balanced construction and is included in the release audit. Chamberwise evaluations of the networks in Equations (C.1) and (C.2) and selected folded-catalecticant and ordered-aggregation identities are regression checks. They do not replace the proofs.

E.3. **Code availability.** Python verifiers and exact certificates are supplied in the source archive. The accompanying `README.md` describes regeneration, integrity checks, and compilation; full JSON outputs and search inputs remain in the source archive. The verifiers use only the Python standard library.

Funding

The author received no funding for this work.

Conflict of interest

The author declares no conflict of interest.

References

[1] E. Bakaev, F. Brunck, C. Hertrich, J. Stade, and A. Yehudayoff, Better neural network expressivity: Subdividing the simplex. In *Proceedings of the 58th annual ACM symposium on theory of computing*, pp. 500–507, 2026

[2] B. Bollobás, *Modern graph theory*. Graduate Texts in Mathematics 184, Springer, New York, 1998

[3] J. Bruna, The multiscale single-index model: A stylized model for hierarchical feature learning. 2026, `arXiv:2607.03347v1`

[4] K.-L. Chen, H. Garudadri, and B. D. Rao, Improved bounds on neural complexity for representing piecewise linear functions. In *Advances in neural information processing systems*, pp. 7167–7180, 35, Curran Associates, Inc., 2022

[5] P. Das, Piecewise linear functions and neural network expressivity via discriminantal arrangements. 2026, `arXiv:2604.02480v1`

[6] S. Dereich and S. Kassing, On the existence of optimal shallow feedforward networks with ReLU activation. *J. Mach. Learn.* **3** (2024), no. 1, 1–22

[7] A. Dutta, I. Safran, and P. Valiant, The median is easier than it looks: Approximation with a constant-depth, linear-width ReLU network. In *Proceedings of the thirty-ninth conference on learning theory*, pp. 2145–2199, Proceedings of Machine Learning Research 336, 2026

[8] C. Eckart and G. Young, The approximation of one matrix by another of lower rank. *Psychometrika* **1** (1936), no. 3, 211–218

[9] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, Maxout networks. In *Proceedings of the 30th international conference on machine learning*, pp. 1319–1327, Proceedings of Machine Learning Research 28(3), 2013

[10] M. Grillo, C. Hertrich, and G. Loho, Depth-bounds for neural networks via the braid arrangement. In *Advances in neural information processing systems*, pp. 141727–141745, 38, 2025

[11] D. Haider, M. Ehler, and P. Balazs, Convex geometry of ReLU-layers, injectivity on the ball and local reconstruction. In *Proceedings of the 40th international conference on machine learning*, pp. 12339–12350, Proceedings of Machine Learning Research 202, PMLR, 2023

[12] C. Hertrich, A. Basu, M. Di Summa, and M. Skutella, Towards lower bounds on the depth of ReLU neural networks. In *Advances in neural information processing systems*, pp. 3336–3348, 34, 2021

[13] P. Jin, Two-hidden-layer ReLU neural networks and finite elements. *Neural Netw.* **198** (2026), 108559

[14] Q.-T. Le, E. Riccietti, and R. Gribonval, Does a sparse ReLU network training problem always admit an optimum? In *Advances in neural information processing systems*, pp. 65265–65298, 36, 2023

[15] L.-H. Lim, M. Micha lek, and Y. Qi, Best k-layer neural network approximations. *Constr. Approx.* **55** (2022), no. 1, 583–604

[16] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, The expressive power of neural networks: A view from the width. In *Advances in neural information processing systems*, pp. 6231–6239, 30, Curran Associates, Inc., 2017

[17] S. Mahan, E. J. King, and A. Cloninger, Nonclosedness of sets of neural networks in Sobolev spaces. *Neural Netw.* **137** (2021), 85–96

[18] V. E. Maiorov, On best approximation by ridge functions. *J. Approx. Theory* **99** (1999), no. 1, 68–94
[19] W. Mantel, Vraagstuk XXVIII. *Wiskundige Opgaven met de Oplossingen* **10** (1907), 60–61
[20] S. McCarty, Piecewise linear functions representable with infinite width shallow ReLU neural networks. *Proc. Amer. Math. Soc. Ser. B* **10** (2023), 296–310
[21] M. Mondelli and A. Montanari, On the connection between learning two-layer neural networks and tensor decomposition. In *Proceedings of the 22nd international conference on artificial intelligence and statistics*, pp. 1051–1060, Proceedings of Machine Learning Research 89, 2019
[22] P. Petersen, M. Raslan, and F. Voigtlaender, Topological properties of the set of functions generated by neural networks of fixed size. *Found. Comput. Math.* **21** (2021), no. 2, 375–444
[23] H. Petzka, M. Trimmel, and C. Sminchisescu, Notes on the symmetries of 2-layer ReLU-networks. *Proceedings of the Northern Lights Deep Learning Workshop* **1** (2020), 6
[24] A. Pinkus, *Ridge functions*. Cambridge Tracts in Mathematics 205, Cambridge University Press, Cambridge, 2015
[25] M. Puthawala, K. Kothari, M. Lassas, I. Dokmanić, and M. de Hoop, Globally injective ReLU networks. *J. Mach. Learn. Res.* **23** (2022), no. 105, 1–55
[26] P. Ramakrishnan, A complete symmetry classification of shallow ReLU networks. 2026, `arXiv:2604.14037v2`
[27] K. Rueß, G. Averkov, F. Brunck, M. Grillo, C. Hertrich, G. Loho, J. Stade, M. Stargalla, M. Sun, and M. Winter, Shallower ReLU network representations via exact linear algebra. 2026, `arXiv:2607.21651v2`
[28] I. Safran, A depth hierarchy for computing the maximum in ReLU networks via extremal graph theory. In *Proceedings of the thirty-ninth conference on learning theory*, pp. 5721–5742, Proceedings of Machine Learning Research 336, 2026
[29] I. Safran, Every layer counts: An exponential L_2 depth hierarchy for ReLU networks. 2026, `arXiv:2608.23877v2`
[30] I. Safran, D. Reichman, and P. Valiant, How many neurons does it take to approximate the maximum? In *Proceedings of the 2024 annual ACM-SIAM symposium on discrete algorithms*, pp. 3156–3183, 2024
[31] I. Safran and O. Shamir, Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th international conference on machine learning*, edited by D. Precup and Y. W. Teh, pp. 2979–2987, Proceedings of Machine Learning Research 70, PMLR, 2017
[32] M. E. Sander, J. Puigcerver, J. Djolonga, G. Peyré, and M. Blondel, Fast, differentiable and sparse top-k: A convex analysis perspective. In *Proceedings of the 40th international conference on machine learning*, pp. 29919–29936, Proceedings of Machine Learning Research 202, 2023
[33] W. Shang, K. Sohn, D. Almeida, and H. Lee, Understanding and improving convolutional neural networks via concatenated rectified linear units. In *Proceedings of the 33rd international conference on machine learning*, pp. 2217–2225, Proceedings of Machine Learning Research 48, 2016
[34] J. W. Siegel, Optimal approximation rates for deep ReLU neural networks on Sobolev and Besov spaces. *J. Mach. Learn. Res.* **24** (2023), no. 357, 1–52
[35] G. Vardi, G. Yehudai, and O. Shamir, Width is less important than depth in ReLU neural networks. In *Proceedings of the thirty-fifth conference on learning theory*, edited by P.-L. Loh and M. Raginsky, pp. 1249–1281, Proceedings of Machine Learning Research 178, PMLR, 2022
[36] L. Venturi, A. S. Bandeira, and J. Bruna, Spurious valleys in one-hidden-layer neural network optimization landscapes. *J. Mach. Learn. Res.* **20** (2019), no. 133, 1–34
[37] S. Wang and X. Sun, Generalization of hinging hyperplanes. *IEEE Trans. Inform. Theory* **51** (2005), no. 12, 4425–4431
[38] Z. Wang and A. Basu, Representing MAX functions using two-hidden-layer ReLU networks. 2026, `arXiv:2608.25221v1`
[39] S. Wojtowytsch, Optimal bump functions for shallow ReLU networks: Weight decay, depth separation, curse of dimensionality. *J. Mach. Learn. Res.* **25** (2024), no. 27, 1–49

School of Artificial Intelligence, Capital University of Economics and Business
No. 121 Zhangjialukou, Huaxiang, Fengtai District, 100070 Beijing, China
Email address: `32023210246@cueb.edu.cn`