ON THE EXACT MINIMUM WIDTH FOR UNIVERSAL APPROXIMATION WITH DEEP RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

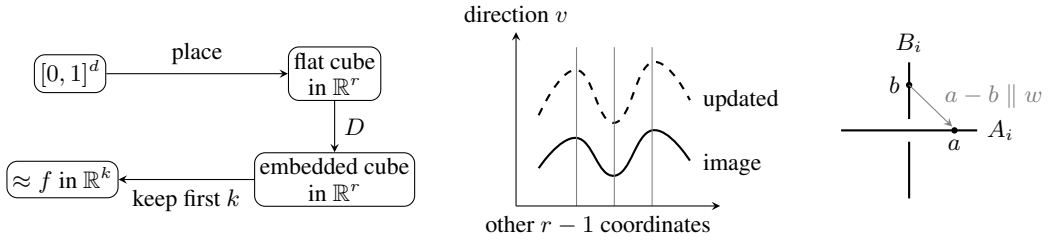
How narrow can a deep network be while still approximating every continuous map from $[0,1]^d$ to $\mathbb{R}^k$? For ReLU networks, the exact minimum width has been known only in a few cases. We prove that it is exactly k whenever $k \geq 2d+2$, improving the previous upper bound by one, and show that the same holds for GELU, SiLU, Softplus, and other activations that approximate ReLU. Our proof builds on a recent diffeomorphism approach that separates the minimum width into a geometric quantity $w(d,k)$ and the additional width needed to approximate diffeomorphisms. We show that one extra neuron suffices for every continuous nonaffine activation with a nonzero derivative at some point, giving the general upper bound $w(d,k)+1$. For ReLU, this extra neuron is necessary when approximating diffeomorphisms on full-dimensional regions, but universal approximation only requires approximation on the lower-dimensional image of $[0,1]^d$. We exploit this distinction to remove the extra neuron when $k \geq 2d+2$, where $w(d,k) = k$. For $k = 2d+1$, we construct embeddings for which the geometric argument fails even after small perturbations. More generally, for all $k \leq 2d+1$, the minimum width is either $w(d,k)$ or $w(d,k)+1$, and we conjecture that it is always $w(d,k)$.

1 INTRODUCTION

Classical results show that neural networks with one hidden layer approximate every continuous function on a compact set when the width is unbounded (Cybenko, 1989; Hornik et al., 1989; Funahashi, 1989; Pinkus, 1999). Deep networks raise a complementary question: if the depth is arbitrary but every hidden layer has at most m neurons, what is the least m for which universal approximation still holds? This *minimum width* measures how much information each layer must carry. It has been studied for many activations and norms, including continuous nonpolynomial activations (Kidger & Lyons, 2020), monotone activations (Johnson, 2019), Leaky-ReLU (Cai, 2023; Li et al., 2023; Yang et al., 2025), squashable activations such as the sigmoid (Shin et al., 2025), and nonaffine polynomials (Hwang & Kang, 2026). We write $W_\sigma(d,k)$ for the minimum width of networks with activation σ that uniformly approximate every continuous map $[0,1]^d \to \mathbb{R}^k$, and focus on ordinary ReLU networks (Lu et al., 2017; Hanin & Sellke, 2017). For scalar outputs, $W_{\mathrm{ReLU}}(d,1) = d+1$ (Hanin & Sellke, 2017), and a construction that keeps a copy of the input in d neurons gives $W_{\mathrm{ReLU}}(d,k) \leq d+k$ (Hanin & Sellke, 2017; Hanin, 2019). For vector outputs the exact value is known only in a few low-dimensional cases (Park et al., 2021; Kim et al., 2024). When $k > 2d$, for instance, it was only known that $W_{\mathrm{ReLU}}(d,k)$ is k or $k+1$ (Hwang, 2025b).

The minimum width is closely tied to diffeomorphisms, that is, smooth bijections whose inverses are also smooth. For Leaky-ReLU networks, Li et al. (2023) lift the input to a higher-dimensional space, deform it by a diffeomorphism generated by a flow, and approximate this diffeomorphism by a deep network. Hwang (2025b) turned this idea into a general framework based on the following geometric quantity. Let $w(d,k)$ be the least r such that every continuous map $[0,1]^d \to \mathbb{R}^k$ can be approximated uniformly by a smooth embedding of the cube into $\mathbb{R}^r$ (a smooth map without self-intersections or singular points) followed by keeping the first k coordinates. A target is first approximated in this way; the embedding is then extended to a diffeomorphism of $\mathbb{R}^r$, which is approximated by a network. This last step needs width r plus an activation-dependent number of extra neurons: none for Leaky-ReLU, one for ReLU, and two for activations that are C^1 near a point with nonzero derivative (Hwang, 2025b, Lemma 4.1 and Corollary 4.6). This paper asks how many of these extra neurons are really needed.

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



(a) diffeomorphism approach **(b)** update along v **(c)** $r = 2d + 1$

Figure 1: Overview. (a) The diffeomorphism approach approximates $f : [0,1]^d \to \mathbb{R}^k$ in three steps: place the cube flat in the first d coordinates of $\mathbb{R}^r$, apply a diffeomorphism D of $\mathbb{R}^r$ (a smooth bijection with smooth inverse), and keep the first k coordinates. A network must then approximate D. For ReLU this needs width $r + 1$ if D must be approximated on a full-dimensional region (Theorem 3), but only width r on the embedded cube when $r \geq 2d + 2$ (Theorem 9). (b) The key step for $r \geq 2d + 2$. If every line parallel to a direction v (gray) meets the current image of the cube at most once, a width-r network can keep the other $r - 1$ coordinates fixed and move each point of the image along v (dashed). (c) For $r = 2d + 1$, some embedded cubes contain, for each coordinate i, two nearly flat pieces A_i and B_i such that every direction w is parallel to $a - b$ for some i, $a \in A_i$, and $b \in B_i$, even after small perturbations of the embedding. Then, for every direction, some line in that direction meets the image twice, so step (b) is unavailable.

For ReLU, the answer depends on where the diffeomorphism must be approximated. We prove that on a full-dimensional region of $\mathbb{R}^r$ the extra neuron is necessary: if ReLU networks of width r approximate a diffeomorphism there arbitrarily well, the diffeomorphism must be affine. The reason is that a ReLU network of width r has full rank only where all of its neurons are active, and there it is affine. For universal approximation, however, the diffeomorphism only needs to be approximated on the embedded cube, which has dimension d. If $r \geq 2d + 2$, there is a direction along which the embedded cube never overlaps itself. We show that a network of width r can then copy the other $r - 1$ coordinates unchanged and update only the one along this direction, so no extra neuron is needed. Figure 1 illustrates both cases.

Contributions. Our results concern ordinary fully connected networks with affine biases, coordinatewise hidden activation, affine output, and arbitrary finite depth.

1. We prove that width $r + 1$ networks approximate every C^2 diffeomorphism of $\mathbb{R}^r$ for every continuous nonaffine activation differentiable with nonzero derivative at one point. Consequently,
$$W_\sigma(d, k) \leq w(d, k) + 1 \leq \min\{d + k + 1, \max\{2d + 1, k\} + 1\}. \tag{1}$$
The construction covers all nonaffine polynomial activations, to which proofs based on the density of one-hidden-layer networks do not apply. This reduces the two extra neurons needed in Hwang (2025b, Lemma 4.1) to one under a weaker hypothesis.
2. We prove that if ReLU networks of width r converge uniformly to a homeomorphism of $\mathbb{R}^r$ on an r-dimensional convex region, then the limit is affine on that region. Hence approximating a nonaffine C^2 diffeomorphism of $\mathbb{R}^r$ on such a region requires width exactly $r + 1$.
3. We prove that restrictions of diffeomorphisms of $\mathbb{R}^r$ to compact subsets of finitely many d-dimensional affine subspaces of $\mathbb{R}^r$, such as finite unions of d-dimensional simplices, are approximable by width-r ReLU networks for $r \geq 2d+2$. This closes an infinite family of minimum-width problems:
$$\boxed{W_{\mathrm{ReLU}}(d, k) = k \qquad (k \geq 2d + 2).} \tag{2}$$
The same exact value holds for every activation that emulates ReLU on compact sets (Corollary 10), including GELU, SiLU, Softplus, and ReLU6.
4. For $r = 2d + 1$ we construct a smooth embedding of the cube in $\mathbb{R}^r$ such that neither the embedding nor any continuous map uniformly close to it has a direction along which it never overlaps itself (Proposition 11). Hence the construction of item 3 does not extend to $k = 2d + 1$. This is not a lower bound for general ReLU networks of width $2d + 1$.

Every result of this paper has been formalized in Lean 4, with the five theorems cited from the literature as explicit hypotheses; the code and a blueprint with detailed proofs are provided as supplementary material (see the reproducibility statement).

Relation to prior results. *General activations.* Kidger & Lyons (2020, Proposition 4.9) proved $W_\sigma(d,k) \le d+k+1$ for continuous nonpolynomial activations that are continuously differentiable near a point with nonzero derivative, and Hwang & Kang (2026, Corollary 18) proved the same bound for nonaffine polynomials. Hwang (2025a, Theorem 4.1) showed that width $r+1$ suffices to approximate diffeomorphisms of $\mathbb{R}^r$ together with their derivatives, for nonpolynomial $C^{1,1}$ activations. Our Theorem 1 requires only differentiability at one point, includes all nonaffine polynomial activations, and gives $W_\sigma(d,k) \le w(d,k)+1$. Since always $w(d,k) \le d+k$ (embed the cube as the graph of a smooth approximation of the target), this bound is never worse than $d+k+1$, and it is strictly better whenever $w(d,k) < d+k$, for example for all $k \ge 2d+1$, where $w(d,k) = k$.

ReLU and related activations. Table 1 lists the known bounds on $W_{\mathrm{ReLU}}(d,k)$ for $k > d$. Theorem 9 settles the range $k \ge 2d+2$, where the best previous upper bound was $k+1$. For Leaky-ReLU with a fixed slope $\beta \ne 1$, the minimum width equals $w(d,k)$ for all d and k (Hwang, 2025b, Corollary 4.6; Yang et al., 2025, Theorem 14); in particular it equals k for $k > 2d$, including the case $k = 2d+1$ that remains open for ReLU. Fixed ELU and SELU with $\alpha \ne 1$ behave the same way (Remark 21). For residual networks whose branches have a single hidden neuron, Zhou et al. (2026) show that the minimum width is k for $k \ge 2d+1$; there the skip connection copies the state, so the extra neuron discussed above is not needed.

Other work. Li et al. (2023) proposed a closed formula for the minimum width of Leaky-ReLU networks; Hwang (2025b, Remark 4.17) found a counterexample, and the authors withdrew the formula in the third arXiv version of their paper (February 2024). Related questions include injective approximation by ReLU networks of unrestricted width (Puthawala et al., 2022), topological obstructions for narrow networks from linking numbers (Ren & Lim, 2026) and from degree theory (Yang, 2025), and a lower bound $\max\{d,k\}+1$ for continuous monotone activations when $k \le 2d$ (Rochau et al., 2025), which extends the ReLU bound of Kim et al. (2024).

Table 1: Bounds on the minimum width $W_{\mathrm{ReLU}}(d,k)$ of ReLU networks for output dimension $k > d$. The last two rows follow from the rows above them; whether the minimum width equals the lower or the upper bound there is open.

Range	Bound	Source
all d, k	$W_{\mathrm{ReLU}}(d,k) \ge k$	Cai (2023)
all d, k	$W_{\mathrm{ReLU}}(d,k) \le d+k$	Hanin & Sellke (2017)
all d, k	$W_{\mathrm{ReLU}}(d,k) \le \max\{2d+1, k\}+1$	Hwang (2025b)
$d < k \le 2d$	$W_{\mathrm{ReLU}}(d,k) \ge k+1$	Kim et al. (2024)
$k \ge 2d+2$	$W_{\mathrm{ReLU}}(d,k) = k$	this paper
$k \ge 2d+2$	$W_\sigma(d,k) = k$ for σ emulating ReLU	this paper
$k = 2d+1$	$k \le W_{\mathrm{ReLU}}(d,k) \le k+1$	open
$k = 2d$, $d \ge 2$	$k+1 \le W_{\mathrm{ReLU}}(d,k) \le k+2$	open

2 PRELIMINARIES

We present some definitions and notation that will be used later in the paper.

An activation $\sigma : \mathbb{R} \to \mathbb{R}$ acts coordinatewise on vectors. A *σ-network* of depth $L \ge 1$ is a map $F = A_{L+1} \circ \sigma \circ A_L \circ \cdots \circ \sigma \circ A_1 : \mathbb{R}^d \to \mathbb{R}^k$ with affine $A_\ell : \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$, $n_0 = d$, and $n_{L+1} = k$; its *width* is $\max_{1 \le \ell \le L} n_\ell$. Write $\|f\|_{C(X)} = \sup_{x \in X} \|f(x)\|$, where $\|\cdot\|$ is the maximum norm on $\mathbb{R}^k$; since all norms on $\mathbb{R}^k$ are equivalent, no result depends on this choice. The *minimum width* $W_\sigma(d,k)$ is the least m such that for every continuous $f : [0,1]^d \to \mathbb{R}^k$ and $\varepsilon > 0$ some σ-network F of width at most m has $\|F - f\|_{C([0,1]^d)} < \varepsilon$. Networks of width m *approximate* $G : \mathbb{R}^r \to \mathbb{R}^r$ *uniformly on compact sets* if for every compact K and $\varepsilon > 0$ one of them has $\|F - G\|_{C(K)} < \varepsilon$.

A C^s *diffeomorphism* of $\mathbb{R}^r$ ($1 \le s \le \infty$) is a bijection $D : \mathbb{R}^r \to \mathbb{R}^r$ such that D and D^{-1} are of class C^s; *smooth* means C^∞. A *smooth embedding* of the cube is an injective map $e : [0,1]^d \to \mathbb{R}^r$ that extends to a C^∞ map on a neighborhood of $[0,1]^d$ whose derivative is injective at every point of $[0,1]^d$; we write $\mathrm{Emb}^\infty([0,1]^d, \mathbb{R}^r)$ for the set of such maps. Let $q_{d,r}(x) = (x, 0)$ for $d \le r$, and let $p_{r,k} : \mathbb{R}^r \to \mathbb{R}^k$ keep the first k coordinates. We call $q_{d,r}([0,1]^d)$ the *flat cube* in $\mathbb{R}^r$.

Following Hwang (2025b), define

$$w(d,k) := \min\big\{r \ge \max\{d,k\} :\ p_{r,k} \circ \mathrm{Emb}^\infty([0,1]^d, \mathbb{R}^r) \text{ is dense in } C([0,1]^d, \mathbb{R}^k)\big\}. \tag{3}$$

Embedding the cube as the graph $x \mapsto (g(x), x)$ of a smooth approximation g of the target, and the density of smooth embeddings in $\mathbb{R}^r$ for $r \ge 2d+1$ (Hirsch, 1976, Chapter 2, Theorems 2.6 and 2.13), give

$$w(d,k) \le \min\{d+k, \max\{2d+1, k\}\}. \tag{4}$$

Also $w(d,k) \ge d+1$: every coordinate of a smooth embedding of the cube into $\mathbb{R}^d$ has nonvanishing gradient and hence no interior local maximum, so it cannot approximate $x \mapsto -\|x-c\|_2^2$ uniformly, where c is the centre of the cube.

For ReLU, $w(d,k) \le W_{\mathrm{ReLU}}(d,k) \le d+k$. The upper bound follows from the construction of Hanin & Sellke (2017), which keeps a copy of the input in d neurons; the lower bound is Hwang (2025b, Theorem 4.5), valid for every continuous increasing activation, and Appendix A proves it directly, for every continuous activation that is a uniform limit on compact intervals of smooth increasing functions (Proposition 12).

3 One extra neuron suffices to approximate diffeomorphisms

This section proves that, for every continuous nonaffine activation with a nonzero derivative at some point, networks with one neuron more than the dimension approximate every C^2 diffeomorphism, and deduces the bound $W_\sigma(d,k) \le w(d,k) + 1$.

Theorem 1 (Diffeomorphisms with one extra neuron)**.** *Let σ be continuous and nonaffine. If σ is differentiable at some α with $\sigma'(\alpha) \ne 0$, then σ-networks of width $r+1$ approximate every C^2 diffeomorphism $\mathbb{R}^r \to \mathbb{R}^r$ uniformly on compact sets. Consequently, $W_\sigma(d,k) \le w(d,k) + 1$.*

Proof sketch. The proof has five steps. Steps 1–4 approximate a diffeomorphism that changes only one coordinate, called a *single-coordinate transformation*; Step 5 builds general diffeomorphisms from such maps. Until Step 4, we allow the r neurons that copy the state to use the identity instead of σ; the remaining neuron, which uses σ, does the computation.

Step 1: ridge functions are dense. On every compact set, finite sums of functions $z \mapsto c\,\sigma^{\circ q}(a^\top z + b)$, where $\sigma^{\circ q}$ applies σ q times, approximate every continuous function (Lemma 13). For nonpolynomial σ, $q = 1$ suffices (Leshno et al., 1993). If σ is a polynomial of degree at least two, repeated application raises the degree without bound; exact Taylor expansion then yields every power $(a^\top z)^j$, hence every exponential $e^{a^\top z}$, and exponentials are dense by the Stone–Weierstrass theorem.

Step 2: one network block adds one ridge term. Write $z = (x, y) \in \mathbb{R}^{r-1} \times \mathbb{R}$ and let h be one such ridge term. The map

$$S_{h,\delta}(z) = z + \delta h(z)\, e_r \tag{5}$$

is computed exactly with width $r+1$: r neurons copy z, the extra neuron computes $h(z)$, and the next affine layer adds $\delta h(z)$ to y (Lemma 14). Blocks can be chained, so the same extra neuron is reused for every ridge term.

Step 3: a single-coordinate transformation is a chain of such updates. Let $T(x,y) = (x, \tau(x,y))$ with $\partial_y \tau > 0$. Moving from y to $\tau(x,y)$ along the straight path $(1-t)y + t\tau(x,y)$ in m equal time steps is exactly a sequence of m updates of the current last coordinate ζ, of the form $\zeta \mapsto \zeta + \frac{1}{m} v(t,x,\zeta)$, for an explicit Lipschitz function v (Lemma 15). Approximating v by ridge sums on finitely many time intervals and applying the updates (5) with a small step reproduces T approximately; a discrete Gronwall estimate keeps the accumulated error small (Lemma 16).

Step 4: the copying neurons can use σ. Since $\sigma'(\alpha) \ne 0$, the rescaled activation $s \mapsto (\sigma(\alpha + \lambda s) - \sigma(\alpha))/(\lambda \sigma'(\alpha))$ tends to the identity uniformly on bounded intervals as $\lambda \to 0$. The rescaling is

affine, so it is absorbed into the neighboring affine layers, and every identity neuron can be replaced by σ without changing the width (Lemma 17).

Step 5: general diffeomorphisms and the width bound. On a compact set, every C^2 diffeomorphism of $\mathbb{R}^r$ is a finite composition of affine maps and single-coordinate transformations of the form in Step 3 (Teshima et al., 2020, Theorem 4); for $r = 1$ the map itself, possibly after a reflection, has this form. Composing the approximations of the factors proves the first claim of Theorem 1. For the second, a smooth embedding of the cube that approximates the target after projection is, up to a small error, a diffeomorphism of $\mathbb{R}^{w(d,k)}$ applied to the flat cube (Palais, 1960, Theorem B and Corollary 2), so approximating that diffeomorphism with width $w(d, k) + 1$ approximates the target.

Appendix B proves each step in full, with explicit error bounds.

4 ReLU NEEDS THE EXTRA NEURON ON FULL-DIMENSIONAL REGIONS

Theorem 1 approximates diffeomorphisms of $\mathbb{R}^r$ with one neuron more than the dimension. This section shows that, for ReLU, the extra neuron cannot be removed when the diffeomorphism must be approximated on a full-dimensional region: ReLU networks of width r can converge on an r-dimensional convex region only to maps that are affine there (Theorem 3). Hence the exact width for this task is $r + 1$ (Corollary 4). We write $\rho(t) = \max\{t, 0\}$; the depth and parameters of the networks may vary along a sequence.

The key fact is that a ReLU network of width r on $\mathbb{R}^r$ can have full rank only where all of its neurons are active, and there it is affine.

Lemma 2 (The all-active region). *Let $N : \mathbb{R}^r \to \mathbb{R}^r$ be a ReLU network of width at most r, and let $Q \subset \mathbb{R}^r$ be compact. There is a possibly empty open convex set $P_N \subset \mathbb{R}^r$ such that N agrees with one affine map on P_N and*

$$C_N(Q) := N(Q \setminus P_N)$$

is compact and nowhere dense in $\mathbb{R}^r$.

Theorem 3 (Rigidity at width r). *Let $r \geq 1$, let $U \subset \mathbb{R}^r$ be nonempty, bounded, open, and convex, and put $Q = \overline{U}$. Let $F : \mathbb{R}^r \to \mathbb{R}^r$ be a homeomorphism. If ReLU networks $N_j : \mathbb{R}^r \to \mathbb{R}^r$ of width at most r satisfy*

$$\|N_j - F\|_{C(Q)} \longrightarrow 0,$$

then $F|_U$ is affine.

Proof idea. By the Baire category theorem, the points $x \in U$ whose images $F(x)$ avoid every set $C_{N_j}(Q)$ of Lemma 2 form a dense subset of U. For such x and large j, Brouwer's fixed point theorem gives a preimage of $F(x)$ under N_j near x, and this preimage must lie in the all-active region P_{N_j}, where N_j is affine. Since P_{N_j} is convex, N_j maps each convex combination of such preimages to the same convex combination of their images. Letting $j \to \infty$ gives $F(\lambda x + (1 - \lambda)x') = \lambda F(x) + (1 - \lambda)F(x')$ for these points, and by density and continuity for all $x, x' \in U$; so F is affine on U. The proofs are in Appendix C.

If F is not affine on U, the theorem shows that the approximation error of ReLU networks of width at most r on Q is bounded below by a positive constant. Together with Theorem 1, this gives the exact width.

Corollary 4 (Exact width for nonaffine diffeomorphisms). *Let $r \geq 1$ and let $F : \mathbb{R}^r \to \mathbb{R}^r$ be a nonaffine C^2 diffeomorphism. There is a bounded open convex set U such that, on $Q = \overline{U}$, the least hidden width at which ReLU networks can approximate $F|_Q$ uniformly is exactly $r + 1$.*

In particular, for

$$F_r(x_1, \ldots, x_r) = (x_1 + x_1^3, x_2, \ldots, x_r),$$

the exact approximation width on $[-1, 1]^r$ is $r + 1$.

Proof. Because F is nonaffine, it is nonaffine on some bounded open ball U; otherwise its affine formulas on overlapping balls would make it globally affine. Theorem 3 rules out uniform approximation on $\overline{U}$ at width at most r. Theorem 1, applied to the ReLU activation, gives uniform approximation of every C^2 diffeomorphism of $\mathbb{R}^r$ at width $r + 1$. This proves exactness. The displayed map F_r is a

smooth diffeomorphism of $\mathbb{R}^r$ because $t \mapsto t + t^3$ has strictly positive derivative and is onto; it is nonaffine on the interior of the stated cube. □

The corollary concerns approximation on a full-dimensional region. On a lower-dimensional set such as an embedded cube, the set $C_N(Q)$ need not prevent injectivity, so Theorem 3 gives no lower bound for universal approximation. Section 5 shows that, in high dimensions, the extra neuron is indeed unnecessary there.

Remark 5 (Clipped activations)**.** *The proof of Theorem 3 uses only two properties of ReLU: it is affine with nonzero slope on an open interval $I \neq \mathbb{R}$ and constant on each component of $\mathbb{R} \setminus \overline{I}$. Any continuous activation with this structure, such as ReLU6, hard-tanh, and hard-sigmoid, satisfies Lemma 2 with P_N the set on which every preactivation lies in I. Hence Theorem 3 holds for these activations, and, since Theorem 1 applies to them as well, Corollary 4 holds for all of them.*

5 Exact minimum width for $k \geq 2d + 2$

This section proves the main result: $W_{\mathrm{ReLU}}(d, k) = k$ for $k \geq 2d + 2$ (Theorem 9). By Section 4, ReLU needs the extra neuron to approximate a diffeomorphism on a full-dimensional region. Universal approximation, however, needs the diffeomorphism only on the embedded cube. We show that there the extra neuron is unnecessary once $r \geq 2d + 2$. The reason is that, for some direction v, every line parallel to v meets the image of the cube at most once (Figure 1(b)). A width-r network can then keep the other $r - 1$ coordinates unchanged and move each point along v.

We call $M \subset \mathbb{R}^r$ *d-flat-covered* if it lies in the union of finitely many affine subspaces of dimension at most d; for instance, a finite union of d-dimensional simplices. For a nonzero $v \in \mathbb{R}^r$, let $\pi_v : \mathbb{R}^r \to \mathbb{R}^{r-1}$ be any surjective linear map with $\ker \pi_v = \mathrm{span}(v)$. We call v a *good direction* for $M \subset \mathbb{R}^r$ if $\pi_v|_M$ is injective, that is, if every line parallel to v meets M at most once. This condition does not depend on the choice of π_v. We also call v a good direction for a map g if $\pi_v \circ g$ is injective.

5.1 Updating one direction at width r

The update along a good direction uses the following consequence of the max–min construction of Hanin & Sellke (2017, Propositions 2 and 3). If $Y \subset \mathbb{R}^n$ is compact and $h \in C(Y, \mathbb{R})$, then, for every $\varepsilon > 0$, there is a ReLU network whose hidden width is $n + 1$ and whose final hidden state on Y has the form

$$y \longmapsto (y + c, \widehat{h}(y) + C), \qquad \|\widehat{h} - h\|_{C(Y)} < \varepsilon,$$

for fixed shifts $c \in \mathbb{R}^n$ and $C \in \mathbb{R}$. Thus the network passes its n inputs through unchanged, up to the shifts c, and computes an approximation of h in one extra neuron. Their construction has exactly this form; only its final output map, which reads off the last coordinate, must be replaced by an affine map of the whole hidden state. Their printed proof of the density step (Proposition 3) treats input dimension two in detail; Appendix G gives a complete proof in every dimension.

Lemma 6 (Graph update)**.** *Let $M \subset \mathbb{R}^r$ be compact, and let v be a good direction for M. For every continuous map*

$$S(z) = z + v\phi(z), \qquad z \in M,$$

and every $\varepsilon > 0$, there is a ReLU network $B : \mathbb{R}^r \to \mathbb{R}^r$ of hidden width r such that $\|B - S\|_{C(M)} < \varepsilon$.

Proof. Rescaling v and ϕ, assume that $\|v\| = 1$. Choose a linear right inverse $J : \mathbb{R}^{r-1} \to \mathbb{R}^r$ of π_v. Because M is compact and $\pi_v|_M$ is a continuous injection into a Hausdorff space, it is a homeomorphism onto $Y := \pi_v(M)$. Hence there is a continuous height $g : Y \to \mathbb{R}$ for which

$$M = \{Jy + vg(y) : y \in Y\}.$$

On this graph, S replaces g by the continuous height

$$s(y) := g(y) + \phi(Jy + vg(y)).$$

Apply the scalar max–min construction above with input dimension $r-1$ to obtain $\widehat{s}$ satisfying $\|\widehat{s} - s\|_{C(Y)} < \varepsilon$. Precompose its first affine map with π_v, pass the $r-1$ coordinates $y = \pi_v(z)$ through unchanged, and replace its scalar output projection by

$$(y', t') \longmapsto J(y' - c) + v(t' - C).$$

The resulting network B has hidden width $(r-1)+1 = r$ and, on M, equals $Jy + v\widehat{s}(y)$ with $y = \pi_v(z)$, whereas $S(z) = Jy + vs(y)$. Hence $\|B - S\|_{C(M)} \le \|\widehat{s} - s\|_{C(Y)} < \varepsilon$. □

A deep network applies Lemma 6 many times, each time to the image of the cube under the blocks built so far. Lemma 7 below provides a good direction for this image as long as it is d-flat-covered, and this property is preserved: every ReLU network maps d-flat-covered sets to d-flat-covered sets. Indeed, for $u \in \mathbb{R}^n$ let $S_{p(u)}$ be the diagonal matrix with entry 1 for the coordinates with $u_i \ge 0$ and 0 otherwise, so that $\mathrm{ReLU}(u) = S_{p(u)}u$. For an affine $A : \mathbb{R}^m \to \mathbb{R}^n$ and an affine subspace E this gives $\mathrm{ReLU}(A(E)) \subset \bigcup_{p\in\{0,1\}^n} S_p A(E)$, a union of 2^n affine subspaces of dimension at most $\dim E$. An affine map also sends each affine subspace to an affine subspace of at most the same dimension, and a network is a finite composition of such maps.

5.2 Good directions exist when $r \ge 2d+2$

For $x \ne y$ in M, call $(x-y)/\|x-y\|$ a secant direction. A nonzero v is a good direction precisely when neither $v/\|v\|$ nor its negative is a secant direction.

Lemma 7 (Good directions are dense)**.** *Let $M \subset \mathbb{R}^r$ be d-flat-covered. If $r \ge 2d+2$, then the good directions for M are dense in $\mathbb{R}^r$: every nonempty open subset of $\mathbb{R}^r$ contains a good direction for M.*

Proof. Let $E_1, \dots, E_N$ be affine subspaces of dimension at most d covering M, and let V_{ab} be the linear span of $E_a - E_b = \{x - y : x \in E_a,\ y \in E_b\}$. If E_a and E_b are nonempty, fix $p \in E_a$ and $q \in E_b$. The identity $x - y = (x-p) - (y-q) + (p-q)$ places V_{ab} in the sum of the two directions and $\mathrm{span}(p-q)$, so

$$\dim V_{ab} \le 2d + 1 < r.$$

If $x \ne y$ in M and $x - y = tv$, then $t \ne 0$, and $v \in V_{ab}$ for pieces $E_a \ni x$ and $E_b \ni y$. Hence every nonzero vector outside the finite union of the proper linear subspaces V_{ab} is a good direction. Each V_{ab} is closed with empty interior in $\mathbb{R}^r$, so the complement of their finite union is open and dense. □

The strict inequality $2d + 1 < r$ in this proof is the only place where the assumption $r \ge 2d+2$ is used.

5.3 Diffeomorphisms on flat-covered sets

We next combine Lemmas 6 and 7 with the finite factorization theorem of Teshima et al. (2020, Theorem 4). Their theorem states that, on a prescribed compact subset of its domain, a C^2 diffeomorphism agrees with a finite composition of invertible affine maps and C^2 single-coordinate transformations. After an affine coordinate permutation, a nonaffine factor has the form

$$T(z) = z + e_i\psi(z).$$

Theorem 8 (Width-r approximation on flat-covered sets)**.** *Let $M_0 \subset \mathbb{R}^r$ be compact and d-flat-covered and suppose $r \ge 2d+2$. If $D : \mathbb{R}^r \to \mathbb{R}^r$ is a smooth diffeomorphism, then, for every $\varepsilon > 0$, there is a width-r ReLU network $N : \mathbb{R}^r \to \mathbb{R}^r$ such that*

$$\|N - D\|_{C(M_0)} < \varepsilon.$$

Proof. On M_0, write $D = F_m \circ \cdots \circ F_1$, where each F_j is an invertible affine map or a single-coordinate transformation, using the factorization above. Affine factors are implemented exactly.

We build one network block for each factor. Suppose the blocks built so far map M_0 onto M; this is the image under the network, which only approximates the image under the exact factors. The set M is compact, and it is d-flat-covered because ReLU networks preserve this property. Consider a factor $T(z) = z + e_i\psi(z)$. By Lemma 7, choose a good direction v for M, as close to e_i as desired. Lemma 6, applied on M with $\phi = \psi$, gives a width-r block B approximating $S_v(z) = z + v\psi(z)$ on M. On every compact set Q, $\|S_v - T\|_{C(Q)} \le \|\psi\|_{C(Q)}\|v - e_i\|$. We can therefore repeat the construction for every factor.

We now bound the error. Let $\widehat{z}$ be the image of a point of M_0 under the blocks built so far, and z its image under the corresponding exact factors. On a compact neighborhood Q of these points, the triangle inequality gives

$$\|B(\widehat{z}) - T(z)\| \le a_j + b_j + \omega_j(\|\widehat{z} - z\|),$$

where the graph error a_j and the direction error $b_j = \|\psi\|_{C(Q)}\|v - e_i\|$ can be chosen arbitrarily small, and ω_j is a modulus of continuity of T on Q. Because the factorization is finite, the errors a_j and b_j can be chosen successively so that all network images stay in the prescribed neighborhoods and make the final error less than ε.

Finally, concatenating the blocks does not increase width: the affine output of one block is composed with the affine input of the next. The invertible affine factors are absorbed into these affine maps. If every factor is affine, a width-r ReLU network represents D exactly on M_0: it shifts all coordinates to be positive on this compact set, applies ReLU, which then acts as the identity, and shifts back. Thus the final composition is a ReLU network of hidden width r. □

5.4 From flat-covered sets to all targets

We now apply Theorem 8 to the flat cube $q_{d,k}([0,1]^d)$, with $q_{d,k}(x) = (x, 0)$ as in Section 2, and the diffeomorphism approach gives the main result. The embedding is constructed on a closed ball containing the cube, because Palais's theorem concerns embeddings of a ball.

Theorem 9 (Exact minimum width for $k \ge 2d + 2$)**.** *For all $d, k \ge 1$ with $k \ge 2d + 2$,*

$$W_{\mathrm{ReLU}}(d, k) = k.$$

Proof. Fix $f \in C([0,1]^d, \mathbb{R}^k)$ and $\varepsilon > 0$, and let $\mathbb{B} \subset \mathbb{R}^d$ be a closed ball containing the cube. Choose a flat torus $T^d = \mathbb{R}^d/L\mathbb{Z}^d$, with L so large that its quotient map π is injective on a neighborhood of $\mathbb{B}$, and extend $f \circ \pi^{-1}$ from $\pi([0,1]^d)$ to a continuous map $T^d \to \mathbb{R}^k$ by Tietze's theorem. Since T^d is compact and $k \ge 2d + 1$, smooth maps are dense in $C(T^d, \mathbb{R}^k)$ and smooth embeddings are dense among them (Hirsch, 1976, Chapter 2, Theorems 2.6 and 2.13); composing such an embedding with π gives a smooth embedding $g : \mathbb{B} \to \mathbb{R}^k$, defined as for the cube in Section 2, with $\|g - f\|_{C([0,1]^d)} < \varepsilon/2$. Palais's theorem, with no orientation condition because $d < k$, supplies a diffeomorphism D of $\mathbb{R}^k$ satisfying $g = D \circ q_{d,k}$ on $\mathbb{B}$ (Palais, 1960, Theorem B).

Apply Theorem 8 to the flat cube $M_0 = q_{d,k}([0,1]^d)$, which lies in the d-dimensional subspace $q_{d,k}(\mathbb{R}^d)$, and this D. Precomposing the resulting width-k network with $q_{d,k}$ gives a width-k ReLU network $N : \mathbb{R}^d \to \mathbb{R}^k$ satisfying $\|N - g\|_{C([0,1]^d)} < \varepsilon/2$, and hence $\|N - f\|_{C([0,1]^d)} < \varepsilon$. This proves $W_{\mathrm{ReLU}}(d, k) \le k$.

For the reverse inequality, choose a target taking affinely independent values at $k + 1$ input points. A network whose last hidden width is below k takes all its values in one proper affine subspace. Since the set of affinely dependent $(k + 1)$-tuples is closed, such networks stay a positive distance from this target. Hence $W_{\mathrm{ReLU}}(d, k) \ge k$. □

For $k \ge 2d + 2$, the best previous upper bound was $k + 1$ (Hwang, 2025b). Theorem 9 removes the extra neuron and matches the lower bound k.

For all d and k, Theorem 1 and the bounds in Section 2 give $w(d, k) \le W_{\mathrm{ReLU}}(d, k) \le \min\{d + k, w(d, k) + 1\}$.

Activations that emulate ReLU. A continuous $\sigma : \mathbb{R} \to \mathbb{R}$ *emulates ReLU on compact sets* if $\beta_n \circ \sigma \circ \alpha_n \to \mathrm{ReLU}$ uniformly on compact intervals for some affine maps α_n, β_n of $\mathbb{R}$. This holds when $\sigma(y)/y$ has a nonzero limit as $y \to +\infty$ and tends to zero as $y \to -\infty$ (GELU, SiLU, Mish, Softplus, ELU, CELU, SELU), and by exact rescaling for ReLU6, hard-tanh, and hard-sigmoid (Lemma 20).

Corollary 10 (ReLU-emulating activations)**.** *If σ emulates ReLU on compact sets, then $W_\sigma(d,k) \le W_{\mathrm{ReLU}}(d,k)$ for all $d, k \ge 1$; in particular $W_\sigma(d,k) = k$ for all $k \ge 2d+2$.*

The proof (Appendix E) replaces every ReLU in a fixed network by $\beta_n \circ \sigma \circ \alpha_n$ and shows, layer by layer, that the error tends to zero on compact sets, as Kim et al. (2024) do for L^p approximation. For the activations listed, the previous best bound for $k \ge 2d+2$ was $k+1$. Fixed ELU and SELU with $\alpha \ne 1$ satisfy $W_\sigma(d,k) = w(d,k)$ for all d and k (Remark 21).

The case $k = 2d+1$. Lemma 7 fails for $r = 2d+1$, and not only because the dimension count is inconclusive: some embeddings have no good direction, and neither has any continuous map uniformly close to them. Let $\mathbb{D}^d$ denote the closed unit ball of $\mathbb{R}^d$; smooth embeddings of $\mathbb{D}^d$ are defined as for the cube.

Proposition 11 (Embeddings with no good direction)**.** *Let $d \ge 1$ and $r = 2d+1$. There are a smooth embedding $E : \mathbb{D}^d \to \mathbb{R}^r$ and a number $\eta > 0$ such that every continuous $E' : \mathbb{D}^d \to \mathbb{R}^r$ satisfying $\|E' - E\|_{C(\mathbb{D}^d)} < \eta$ has the following property: for every nonzero $v \in \mathbb{R}^r$, there are distinct $a, b \in \mathbb{D}^d$ with $E'(a) - E'(b) \in \mathrm{span}(v)$, so that $\pi_v \circ E'$ is not injective. The embedding E may moreover be chosen of the form $E = D \circ q_{d,r}|_{\mathbb{D}^d}$ for a smooth diffeomorphism $D : \mathbb{R}^r \to \mathbb{R}^r$.*

The proof (Appendix D) places in the image of E, up to a small uniform error, $2r$ flat d-dimensional patches A_i, B_i such that, for every direction v, the differences $a - b$ with $a \in A_i$, $b \in B_i$ fill, for some i, a $2d$-dimensional cube that contains a multiple of v and is transversal to v (Figure 1(c)); Brouwer's fixed point theorem then yields the two points, for E and for every map uniformly close to it. The same holds with the cube in place of $\mathbb{D}^d$. Hence the proof of Theorem 9 does not extend to $k = 2d+1$. This does not rule out other width-$(2d+1)$ ReLU networks, so the minimum width remains open: $2d+1 \le W_{\mathrm{ReLU}}(d, 2d+1) \le 2d+2$, and in particular $3 \le W_{\mathrm{ReLU}}(1,3) \le 4$. For $d = 1$, Appendix F shows that width three approximates every curve in $\mathbb{R}^3$ with an injective linear projection to a plane, and also an explicit polygonal arc that has no such projection, even after small perturbations (Proposition 24). So Proposition 11 does not show that width four is needed, even for embedded arcs. For ReLU-emulating activations, Corollary 10 gives only the upper bound $2d+2$ at $k = 2d+1$. By contrast, the minimum width there is exactly k for Leaky-ReLU with a fixed slope, for fixed ELU and SELU with $\alpha \ne 1$, and for residual networks whose branches have a single hidden neuron (Section 1).

6 Limitations and conclusion

Our results concern expressivity with arbitrary finite depth; they give no depth rates, training guarantees, or parameter-efficiency bounds. The case $k = 2d+1$ remains open for ReLU, and Theorem 3 gives no lower bound for approximation on a lower-dimensional set such as the embedded cube. For general nonmonotone activations, equation (1) is an upper bound only: the known lower bounds beyond $\max\{d, k\}$ are proved for monotone activations, and a nonmonotone activation can fold the image of the cube.

The main conclusion is that, for ReLU, approximating a diffeomorphism of $\mathbb{R}^r$ on a full-dimensional region costs one extra neuron, whereas approximating it on an embedded d-dimensional cube costs none when $r \ge 2d+2$. This yields an exact minimum-width family and isolates the unresolved case $k = 2d+1$. Every exactly known value satisfies $W_{\mathrm{ReLU}}(d,k) = w(d,k)$: $k = 1$, $(d,k) \in \{(1,2),(2,2)\}$, and $k \ge 2d+2$. We conjecture that $W_{\mathrm{ReLU}}(d,k) = w(d,k)$ for all d, k, that is, that the extra neuron is never needed on the embedded cube. The conjecture would follow if, for every $r \ge d+1$, ReLU networks of width r approximated every C^2 diffeomorphism of $\mathbb{R}^r$ uniformly on the flat cube, and it would give $W_{\mathrm{ReLU}}(d, 2d) = W_{\mathrm{ReLU}}(d, 2d+1) = 2d+1$.

AI USE STATEMENT

We used generative AI tools to search and summarize the literature, to assist in writing proofs, and to draft and edit text. Generative AI tools also assisted in the Lean 4 formalization of all technical results of this paper. We did not use generative AI tools to generate data or interpret experimental results. We reviewed all AI-assisted work and verified every citation against its original source and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

All results are mathematical theorems. The networks, the minimum width, and the geometric width are defined in Section 2, and every result states its assumptions in full. Complete proofs are given in the main text and in Appendices A–G, which also contain the additional results for $k = 2d + 1$. The supplementary material contains a Lean 4 formalization of every lemma, proposition, theorem, and corollary of the paper, together with a blueprint that gives a detailed written proof of each result and names its Lean counterpart. The formalization uses only Lean's standard axioms. Five results cited from the literature are not formalized and enter as explicit hypotheses of the Lean theorems that use them: the density of smooth embeddings (Hirsch, 1976), the extension theorem of Palais (1960), the factorization theorem of Teshima et al. (2020), the density theorem of Leshno et al. (1993), and, for Remark 21 only, the minimum width of Leaky-ReLU networks (Hwang, 2025b; Yang et al., 2025). Brouwer's fixed point theorem and the density of max–min strings (Appendix G) are proved in Lean.

REFERENCES

Yongqiang Cai. Achieve the minimum width of neural networks for universal approximation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.11395.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. doi: 10.1007/BF02551274.

Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989. doi: 10.1016/0893-6080(89)90003-8.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10):992, 2019. doi: 10.3390/math7100992.

Boris Hanin and Mark Sellke. Approximating continuous functions by ReLU nets of minimal width. arXiv preprint arXiv:1710.11278, 2017. Version 2, March 2018.

Morris W. Hirsch. *Differential Topology*, volume 33 of *Graduate Texts in Mathematics*. Springer, 1976. doi: 10.1007/978-1-4684-9449-5.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

Geonho Hwang. Optimal minimum width for the universal approximation of continuously differentiable functions by deep narrow MLPs. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://openreview.net/forum?id=GAa7Gy7FW7.

Geonho Hwang. Minimum width for deep, narrow MLP: A diffeomorphism approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://openreview.net/forum?id=S2fswKCN9T. Camera-ready version; arXiv:2308.15873.

Geonho Hwang and Myungjoo Kang. Universal approximation property of fully convolutional neural networks with zero padding. *Applied and Computational Harmonic Analysis*, 82:101833, 2026. doi: 10.1016/j.acha.2025.101833.

Jesse Johnson. Deep, skinny neural networks are not universal approximators. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGgSsAcFQ.

Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327, 2020.

Namjun Kim, Chanho Min, and Sejun Park. Minimum width for universal approximation using ReLU networks on compact domain. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.10402.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. doi: 10.1016/S0893-6080(05)80131-5.

Li'ang Li, Yifei Duan, Guanghua Ji, and Yongqiang Cai. Minimum width of Leaky-ReLU neural networks for uniform universal approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19460–19470, 2023. URL https://proceedings.mlr.press/v202/li23g.html.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Richard S. Palais. Extending diffeomorphisms. *Proceedings of the American Mathematical Society*, 11(2):274–277, 1960. doi: 10.1090/S0002-9939-1960-0117741-0.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.08859.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999. doi: 10.1017/S0962492900002919.

Michael Puthawala, Konik Kothari, Matti Lassas, Ivan Dokmanić, and Maarten de Hoop. Globally injective ReLU networks. *Journal of Machine Learning Research*, 23(105):1–55, 2022. URL http://jmlr.org/papers/v23/21-0282.html.

Junyu Ren and Lek-Heng Lim. Low-dimensional topology of deep neural networks. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. arXiv:2606.31856.

Dennis Rochau, Robin Chan, and Hanno Gottschalk. New advances in universal approximation with neural networks of minimal width. arXiv preprint arXiv:2411.08735, 2025.

Jonghyun Shin, Namjun Kim, Geonho Hwang, and Sejun Park. Minimum width for universal approximation using squashable activation functions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3362–3373, 2020.

Xiao-Song Yang. Dimensionality reduction and width of deep neural networks based on topological degree theory. arXiv preprint arXiv:2511.06821, 2025.

Xiao-Song Yang, Qi Zhou, and Xuan Zhou. Minimum width of deep narrow networks for universal approximation. arXiv preprint arXiv:2511.06837, 2025.

Qi Zhou, Xuan Zhou, and Xiao-Song Yang. Minimum block width for universal approximation by residual neural networks with inner width one. arXiv preprint arXiv:2607.04597, 2026.

A THE GEOMETRIC LOWER BOUND FOR RELU

This appendix proves the lower bound $w(d,k) \le W_{\mathrm{ReLU}}(d,k)$ of Section 2. We state it for a class of activations that also contains those of Remark 21.

Proposition 12 (Geometric lower bound)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous and, on every compact interval, a uniform limit of C^∞ functions $s : \mathbb{R} \to \mathbb{R}$ with $s' > 0$ on $\mathbb{R}$. Then $w(d,k) \le W_\sigma(d,k)$ for all $d, k \ge 1$.*

ReLU satisfies the hypothesis with the approximants $s_\delta(t) = \frac{1}{2}(t + \sqrt{t^2+\delta^2})$, $\delta \downarrow 0$. So do Leaky-ReLU with a fixed positive slope and ELU (Remark 21); indeed, every continuous nondecreasing σ does, because mollifying σ and adding δt gives such approximants. Below, s_δ denotes an approximant of σ with $\sup|s_\delta - \sigma| \to 0$ on the compact interval in question.

Proof. Suppose width-m σ-networks are dense in $C([0,1]^d, \mathbb{R}^k)$. Then $m \ge k$ by the last paragraph of the proof of Theorem 9, which applies to every activation, and $m \ge d$: if the first hidden layer A had fewer than d outputs, some $v \ne 0$ in the kernel of its linear part would give $F(p) = F(p+tv)$ for the centre p of the cube and $\|tv\|_\infty = \frac{1}{2}$, so no network could approximate $x \mapsto (\|x-p\|_\infty, 0, \dots, 0)$ within $\frac{1}{4}$.

Call $G : \mathbb{R}^d \to \mathbb{R}^j$ ($j \le m$) *embedding-approximable* if for every compact K and $\eta > 0$ there are a smooth injective immersion $E : \mathbb{R}^d \to \mathbb{R}^m$ and a linear $\pi : \mathbb{R}^m \to \mathbb{R}^j$ with $\|\pi \circ E - G\|_{C(K)} < \eta$. The identity is embedding-approximable, through $x \mapsto (x, 0)$. Suppose G is, and let $A(y) = Py + a$ map to $\mathbb{R}^{j'}$, $j' \le m$. For every $\theta > 0$ there is an invertible $m \times m$ matrix M whose first j' rows are within θ of the rows of $P\pi$, because $\det\big([P\pi; 0] + \lambda I\big)$ is a monic polynomial in λ. Put $E' = s_\delta(ME + (a, 0))$, with s_δ applied coordinatewise; it is a smooth injective immersion, because $s'_\delta > 0$. Let π' keep the first j' coordinates. On K,

$$\big\|\pi' E' - \sigma(AG)\big\| \le \sup|s_\delta - \sigma| + \omega\big(\theta \|E\|_{C(K)}\big) + \omega\big(\|P\|\,\|\pi E - G\|_{C(K)}\big),$$

where ω is a modulus of continuity of σ on a compact interval; all three terms can be made small, by choosing first E and π, then θ, and then δ. Hence $\sigma(AG)$ is embedding-approximable, and by induction over the layers so is the composition g of the hidden layers of every width-m network $F = B \circ g$, where B is the final affine map.

Fix f and ε, a network $F = B \circ g$ within $\varepsilon/2$ of f, and E, π with $\|B\pi E - Bg\| < \varepsilon/4$ on the cube. Write $B\pi E(x) = LE(x) + c$ with $L : \mathbb{R}^m \to \mathbb{R}^k$. Choose an invertible M whose first k rows are close enough to L, and put $T(z) = Mz + (c, 0)$. Then $p_{m,k} \circ T \circ E$ is within ε of f on the cube, and $T \circ E$ is a smooth injective immersion of $\mathbb{R}^d$; its restriction to the cube is a smooth embedding of the cube in the sense of Section 2. Density of width-m σ-networks thus implies density of such projected embeddings, and the definition (3) forces $m \ge w(d,k)$. Unlike the argument of Hwang (2025b, Theorem 4.5), this proof needs no padding of hidden layers and no intermediate Leaky-ReLU network. □

B PROOF OF THEOREM 1: ONE EXTRA NEURON SUFFICES

This appendix proves Theorem 1, following the five steps of Section 3. Theorem 18 is its first claim and Corollary 19 its second. The extra neuron does all the computation and is reused at every step, while the other r neurons copy the current state. The main point is to approximate a single-coordinate transformation by a chain of ridge updates (5). Because the straight-line interpolation τ_t below is affine in t, the transformation is exactly a composition of explicit Euler steps of the field (6). For a polynomial σ, sums of ridge functions $\sigma(a^\top z + b)$ have bounded degree and are not dense; we therefore use the iterates $\sigma^{\circ q}$, whose degree is unbounded.

We temporarily allow *hybrid* networks: in each hidden layer, at most one coordinate has activation σ, and all remaining coordinates have the identity activation. Let $\mathcal{H}_r^\sigma$ denote the maps $\mathbb{R}^r \to \mathbb{R}^r$ represented by such networks with hidden width at most $r+1$. As usual, all affine maps may have biases, and the output layer is affine. We write $\sigma^{\circ q}$ for the q-fold composition of σ.

B.1 DENSITY OF ITERATED RIDGE FUNCTIONS

Lemma 13 (Iterated ridges). *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous and nonaffine. For every $r \geq 1$, the linear space*

$$\mathcal{A}_\sigma(r) := \operatorname{span}\{z \mapsto \sigma^{\circ q}(a^\top z + b) : q \geq 1,\ a \in \mathbb{R}^r,\ b \in \mathbb{R}\}$$

is dense in $C(K)$ for every compact $K \subset \mathbb{R}^r$.

Proof. If σ is nonpolynomial, the claim with $q = 1$ is the classical ridge-function universal approximation theorem Leshno et al. (1993).

Suppose instead that σ is a polynomial of degree $m \geq 2$. Fix $q \geq 1$ and put $\rho = \sigma^{\circ q}$, whose degree is $N = m^q$. Choose $b_0 \in \mathbb{R}$ such that $\rho^{(j)}(b_0) \neq 0$ for $0 \leq j \leq N$; this is possible because these are finitely many nonzero polynomials. Given $a \in \mathbb{R}^r$, choose $N + 1$ distinct scalars $\lambda_0, \ldots, \lambda_N$. Taylor's identity gives

$$\rho(b_0 + \lambda_\ell a^\top z) = \sum_{j=0}^{N} \frac{\rho^{(j)}(b_0)}{j!} \lambda_\ell^j (a^\top z)^j, \qquad 0 \leq \ell \leq N.$$

The coefficient matrix is the product of an invertible Vandermonde matrix and an invertible diagonal matrix. Hence $(a^\top z)^j$ belongs to the span of these $N + 1$ allowed ridges for every $0 \leq j \leq N$, and, letting q grow, for every j. On K the partial sums of $\sum_j (a^\top z)^j / j!$ converge uniformly to $e^{a^\top z}$, so every exponential ridge lies in the closure of $\mathcal{A}_\sigma(r)$. The span of $\{e^{a^\top z} : a \in \mathbb{R}^r\}$ is a subalgebra of $C(K)$ that contains the constants ($a = 0$) and separates points ($a = x - y$), because $e^{a^\top z} e^{b^\top z} = e^{(a+b)^\top z}$. By the Stone–Weierstrass theorem it is dense, and hence $\mathcal{A}_\sigma(r)$ is dense. □

The next lemma shows that a hybrid network of width $r + 1$ computes a ridge update exactly. Write $z = (x, y) \in \mathbb{R}^{r-1} \times \mathbb{R}$ and let e_r be the last standard basis vector.

Lemma 14 (Exact hybrid ridge update). *For*

$$h(z) = c\,\sigma^{\circ q}(a^\top z + b), \qquad S_{h,\delta}(z) = z + \delta e_r h(z),$$

the map $S_{h,\delta}$ is represented exactly by a member of $\mathcal{H}_r^\sigma$. Every finite composition of these updates is also represented at hidden width $r + 1$.

Proof. Use r identity neurons to copy z and the remaining neuron to compute $s_1 = \sigma(a^\top z + b)$. Copy z unchanged while setting $s_{j+1} = \sigma(s_j)$ for $j < q$. The affine output map sends (z, s_q) to $z + \delta c s_q e_r$.

No additional neuron is needed between two updates. If the next ridge term has parameters (a^+, b^+), the affine map after the current value s_q can simultaneously form

$$z^+ = z + \delta c s_q e_r, \qquad a^{+\top} z^+ + b^+ = a^{+\top} z + \delta c\, a_r^+ s_q + b^+.$$

The identity coordinates output z^+, while the one nonlinear coordinate outputs $\sigma(a^{+\top} z^+ + b^+)$. Iteration proves the claim. □

B.2 APPROXIMATING A SINGLE-COORDINATE TRANSFORMATION

A single-coordinate transformation (after a coordinate permutation) has the form

$$T(x, y) = (x, \tau(x, y)).$$

We assume that τ is C^1, that $\partial_y \tau > 0$, and that $\tau(x, \cdot)$ is onto $\mathbb{R}$ for every x. For $t \in [0, 1]$ put $\tau_t(x, y) = (1 - t)y + t\tau(x, y)$. Its derivative in y is $(1 - t) + t\partial_y \tau > 0$, and it tends to $\pm\infty$ at $\pm\infty$, so $\tau_t(x, \cdot)$ is a bijection of $\mathbb{R}$. For $\zeta \in \mathbb{R}$ let $u = u_t(x, \zeta) := \tau_t(x, \cdot)^{-1}(\zeta)$ and set

$$v(t, x, \zeta) = \tau(x, u) - u. \tag{6}$$

We also write $v(t, w)$ for $w = (x, \zeta) \in \mathbb{R}^r$.

Lemma 15 (Single-coordinate transformations are exact Euler compositions)**.** *For every $m \geq 1$ and $z_0 = (x, y)$, the explicit Euler states $z_k = z_{k-1} + \frac{1}{m} e_r v(\frac{k-1}{m}, z_{k-1})$ satisfy $z_k = (x, \tau_{k/m}(x, y))$; in particular $z_m = T(z_0)$. Moreover, for every compact Q there is Λ with $|v(t, w) - v(t', w')| \leq \Lambda(|t - t'| + \|w - w'\|)$ for $t, t' \in [0, 1]$ and $w, w' \in Q$.*

Proof. If $z_{k-1} = (x, \tau_t(x, y))$ with $t = (k-1)/m$, then the preimage in (6) is $u = y$, so the last coordinate of z_k is $\tau_t(x, y) + \frac{1}{m}(\tau(x, y) - y) = \tau_{t+1/m}(x, y)$, because τ_t is affine in t. Induction proves the identity.

For the estimate, choose a compact box X containing the x-components of Q and a bound W for the last coordinates. Since $\tau(x, \pm Y) \to \pm\infty$ as $Y \to \infty$, monotonically in Y, and X is compact, some $Y > W$ satisfies $\pm\tau(x, \pm Y) > W$ on X; then all preimages u for $(x, \zeta) \in Q$ and $t \in [0, 1]$ lie in $[-Y, Y]$. On the compact convex set $R = X \times [-Y, Y]$, let $c = \min(1, \min_R \partial_y \tau) > 0$, let Λ_0 be a Lipschitz constant of τ, and let $\Psi = \max_R |\tau - y|$. Let $w = (x, \zeta)$ and $w' = (x', \zeta')$ in Q. If $y = u_t(x, \zeta)$ and $y' = u_{t'}(x', \zeta')$, then

$$\tau_t(x, y) - \tau_t(x, y') = (\zeta - \zeta') + (t' - t)(\tau - y)(x', y') + t\big(\tau(x', y') - \tau(x, y')\big),$$

and the mean value theorem gives $c|y - y'| \leq (1 + \Lambda_0)\|w - w'\| + \Psi|t - t'|$. Since $\tau - y$ is $(\Lambda_0 + 1)$-Lipschitz on R, the bound for v follows. □

Lemma 16 (Approximation by hybrid networks)**.** *Let σ be continuous and nonaffine, and let $T(x, y) = (x, \tau(x, y))$ satisfy the assumptions above. For every compact $K \subset \mathbb{R}^r$ and $\varepsilon > 0$ there is $H \in \mathcal{H}_r^\sigma$ with $\sup_{z \in K} \|H(z) - T(z)\| < \varepsilon$.*

Proof. The tube $\mathcal{T} = \{(x, \tau_t(x, y)) : (x, y) \in K,\ 0 \leq t \leq 1\}$ is compact and contains all exact Euler states of Lemma 15. Let Q be its closed 1-neighborhood, let Λ be the constant of Lemma 15 on Q, and choose $\eta, \kappa > 0$ with $(2\eta + \kappa)e^\Lambda < \min(1, \varepsilon)$. Choose an integer P with $\Lambda/P < \eta$ and, by Lemma 13, finite sums $g_i = \sum_{j=1}^{M_i} h_{ij}$ of iterated ridges with $\sup_Q |g_i - v(i/P, \cdot)| < \eta$ for $0 \leq i < P$.

For a step size $\delta > 0$, the composition $\Psi_{i,\delta} = S_{h_{iM_i},\delta} \circ \cdots \circ S_{h_{i1},\delta}$ is represented exactly by a hybrid network, by Lemma 14. Let B bound all $|h_{ij}|$ on the closed 1-neighborhood Q' of Q. By uniform continuity of the finitely many h_{ij} on Q' there is $\delta_0 > 0$ such that, for $0 < \delta \leq \delta_0$ and $z \in Q$, the intermediate states stay within $\delta M_i B < 1$ of z and

$$\big\|\Psi_{i,\delta}(z) - z - \delta e_r g_i(z)\big\| \leq \delta\kappa. \tag{7}$$

This is the step that avoids any Lipschitz assumption on the ridge terms or on σ.

Let $m = P\mu$ with an integer μ and $1/m \leq \delta_0$, put $\delta = 1/m$, and apply at step k the map $\Psi_{i(k),\delta}$ with $i(k) = \lfloor (k-1)/\mu \rfloor$. Let w_k be the resulting states, z_k the exact Euler states, and $e_k = \|w_k - z_k\|$. If $w_{k-1} \in Q$, then (7) and Lemma 15 give

$$e_k \leq e_{k-1} + \delta\big(\eta + \Lambda/P + \Lambda e_{k-1}\big) + \delta\kappa \leq (1 + \Lambda\delta)e_{k-1} + \delta(2\eta + \kappa). \tag{8}$$

By induction, $e_k \leq k\delta(2\eta + \kappa)e^{\Lambda k\delta} \leq (2\eta + \kappa)e^\Lambda < 1$, so w_k stays within distance one of $\mathcal{T}$ and (8) remains valid. At $k = m$ the composed hybrid network H satisfies $\|H(z) - T(z)\| = e_m < \varepsilon$. □

B.3 From hybrid to ordinary networks

Rescaling the activation near α replaces every identity neuron by a σ neuron.

Lemma 17 (Identity replacement)**.** *Suppose that σ is continuous, differentiable at α, and $\sigma'(\alpha) \neq 0$. Every hybrid network is a limit, uniformly on compact sets, of ordinary σ-networks with the same hidden widths.*

Proof. For $\lambda \neq 0$, define

$$\iota_\lambda(s) = \frac{\sigma(\alpha + \lambda s) - \sigma(\alpha)}{\lambda\sigma'(\alpha)}.$$

Differentiability at α implies $\iota_\lambda \to \mathrm{id}$ uniformly on every bounded interval as $\lambda \to 0$. Indeed, write $\sigma(\alpha + u) = \sigma(\alpha) + \sigma'(\alpha)u + u\,\gamma(u)$ with $\gamma(u) \to 0$ as $u \to 0$. Then $\iota_\lambda(s) - s = s\,\gamma(\lambda s)/\sigma'(\alpha)$, and $|\lambda s| \leq |\lambda| R \to 0$ uniformly for $|s| \leq R$.

The coefficients of the rescaling grow as $\lambda \to 0$, so we write the affine maps explicitly. Write a hybrid layer as $x_\ell = \phi_\ell(A_\ell x_{\ell-1} + b_\ell)$, where each coordinate of ϕ_ℓ is the identity except for at most one coordinate, on which it is σ. We construct ordinary layers with states y_ℓ, together with affine images $\widehat{y}_\ell$ of y_ℓ that approximate x_ℓ; set $\widehat{y}_0 = y_0 = x$. If coordinate j is an identity coordinate of the hybrid layer, the ordinary layer uses the preactivation

$$\alpha + \lambda(A_\ell \widehat{y}_{\ell-1} + b_\ell)_j,$$

and we set

$$(\widehat{y}_\ell)_j = \frac{(y_\ell)_j - \sigma(\alpha)}{\lambda \sigma'(\alpha)} = \iota_\lambda((A_\ell \widehat{y}_{\ell-1} + b_\ell)_j).$$

On the σ coordinate, if there is one, the ordinary layer uses the preactivation $(A_\ell \widehat{y}_{\ell-1} + b_\ell)_j$, and we set $(\widehat{y}_\ell)_j = (y_\ell)_j$. Because $\widehat{y}_{\ell-1}$ is an affine function of $y_{\ell-1}$, all these preactivations are affine functions of $y_{\ell-1}$, whatever the size of their coefficients. Thus, exactly,

$$\widehat{y}_\ell = \phi_{\ell,\lambda}(A_\ell \widehat{y}_{\ell-1} + b_\ell),$$

where the identity coordinates of $\phi_{\ell,\lambda}$ are ι_λ and its σ coordinate is σ.

Fix a compact set of inputs. The images of this set under the first ℓ hybrid layers, and under the maps $A_\ell(\cdot) + b_\ell$ applied to them, are compact; enlarge them slightly. Since $\phi_{\ell,\lambda} \to \phi_\ell$ uniformly on these compact sets, an induction over the layers, using the fixed affine maps A_ℓ and the uniform continuity of σ on compact sets, proves $\widehat{y}_\ell \to x_\ell$ uniformly at every layer. The affine output layer of the ordinary network first maps the last state y_L to $\widehat{y}_L$ and then applies the output map of the hybrid network. The resulting ordinary networks have the same hidden widths and converge uniformly on the given compact set. □

Theorem 18 (Diffeomorphisms with one extra neuron)**.** *Let $r \geq 1$. Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous and nonaffine and is differentiable at some α with $\sigma'(\alpha) \neq 0$. Then every C^2 diffeomorphism $F : \mathbb{R}^r \to \mathbb{R}^r$ is a limit, uniformly on compact sets, of ordinary σ-networks of hidden width at most $r + 1$.*

Proof. First assume $r \geq 2$. Teshima et al. (Teshima et al., 2020, Theorem 4) (stated there for dimension at least two) show that, for a compact K, the restriction $F|_K$ is a finite composition of invertible affine maps and compactly supported C^2 single-coordinate transformations $T(x, y) = (x, \tau(x, y))$, with coordinate permutations absorbed into the affine factors. Each such factor satisfies the hypotheses of Lemma 15: $\tau(x, \cdot)$ is a bijection of $\mathbb{R}$ because T is a bijection fixing x; $DT(z)$ is invertible and fixes the first $r - 1$ coordinates, so $\partial_y \tau \neq 0$; and $\partial_y \tau$ is continuous and equals one off a compact set, so it is positive on the connected space $\mathbb{R}^r$. Each factor is approximable by a hybrid width-$(r + 1)$ network by Lemma 16; affine maps and coordinate permutations are absorbed into adjacent affine layers. An induction over the finite composition, using uniform continuity of each factor on a compact neighborhood of the intermediate images, gives a hybrid width-$(r + 1)$ approximation to $F|_K$. Lemma 17 converts it to an ordinary σ-network of the same width.

For $r = 1$, the derivative F' is continuous and never vanishes, since F^{-1} is differentiable. If $F' > 0$, then $\tau = F$ satisfies the hypotheses of Lemma 15 with $T = F$. If $F' < 0$, apply this to $z \mapsto F(-z)$ on $-K$ and absorb the reflection into the first affine layer.

Applying the construction on the balls of radius n with error $1/n$ gives a single sequence converging to F uniformly on compact sets. □

For smooth nonpolynomial activations, a width-$(r + 1)$ theorem for diffeomorphisms of $\mathbb{R}^r$ was previously obtained with approximation of the derivatives as well, under a $C^{1,1}$ assumption (Hwang, 2025a, Theorem 4.1). Theorem 18 instead needs only continuity and differentiability at one point with nonzero derivative for uniform approximation, and it also includes all nonaffine polynomial activations. Hwang and Kang's polynomial $d + k + 1$ bound (Hwang & Kang, 2026, Corollary 18) is recovered below, and the bound $w(d, k) + 1$ can be smaller.

B.4 From diffeomorphisms to the width bound

Recall from Section 2 the maps $q_{d,r}(x) = (x, 0)$ and $p_{r,k}$, the smooth embeddings of the cube, and the geometric width $w(d, k)$ defined in (3).

Corollary 19 (General width bound)**.** *Under the assumptions of Theorem 18,*

$$W_\sigma(d,k) \le w(d,k)+1 \le \min\{d+k+1,\ \max\{2d+1,k\}+1\}.$$

In particular, σ-networks of width $w(d,k)+1$ are dense in $C([0,1]^d,\mathbb{R}^k)$.

Proof. Put $r = w(d,k)$, fix $f \in C([0,1]^d,\mathbb{R}^k)$ and $\varepsilon > 0$, and choose a smooth embedding $e : [0,1]^d \to \mathbb{R}^r$ of the cube such that $\|p_{r,k}\circ e - f\|_{C([0,1]^d)} < \varepsilon/2$. By compactness, e is an injective immersion on a box $[-\delta, 1+\delta]^d$. Let $s(t) = t(1+(t/b)^{2p})^{-1/(2p)}$ with $b = \frac{1}{2}+\delta$; it is smooth, has $s' > 0$ and $|s| < b$, and is uniformly close to the identity on $[-\frac{1}{2},\frac{1}{2}]$ for large p. With $c = (\frac{1}{2},\dots,\frac{1}{2})$, the map $\Theta(x) = c + s(x-c)$ (coordinatewise) is a smooth injective immersion of $\mathbb{R}^d$ into that box, so $\psi = e\circ\Theta$ is a smooth injective immersion of $\mathbb{R}^d$, within $\varepsilon/4$ of e on the cube for large p. Palais's theorem (Palais, 1960, Theorem B and Corollary 2), applied to the restriction of ψ to a closed ball containing the cube, supplies a smooth diffeomorphism $G : \mathbb{R}^r \to \mathbb{R}^r$ such that

$$\psi = G \circ q_{d,r} \qquad \text{on } [0,1]^d.$$

Theorem B covers $d < r$. If $d = r$, the two embeddings may induce opposite orientations; Corollary 2 covers this case, because $\mathbb{R}^r$ admits an orientation-reversing diffeomorphism.

By Theorem 18, approximate G on $q_{d,r}([0,1]^d)$ within $\varepsilon/4$ by a width-$(r+1)$ ordinary σ-network N. The map $p_{r,k}\circ N\circ q_{d,r}$ is a d-input, k-output network of the same hidden width and approximates f within ε.

Finally, $w(d,k) \le d+k$: if g is a smooth approximation of f, then $x \mapsto (g(x),x)$ is a smooth embedding of the cube into $\mathbb{R}^{k+d}$ whose first k coordinates are g. Also, $w(d,k) \le \max\{2d+1,k\}$. Indeed, let $r' = \max\{2d+1,k\}$. The torus argument in the proof of Theorem 9 approximates $(f,0) : [0,1]^d \to \mathbb{R}^{r'}$ uniformly by smooth embeddings e of the cube, and then $p_{r',k}\circ e$ approximates f. The torus is used because, for noncompact manifolds, Hirsch's density theorem holds only among proper maps. The better of these two bounds gives the displayed inequality. □

C Proofs of the rigidity results

This appendix contains the proofs moved from Section 4; the statements are repeated by reference.

Proof of Lemma 2. Let P_N be the set on which every hidden preactivation is strictly positive. On this activation pattern, all preceding ReLUs are identities. Consequently, each hidden preactivation is an affine function of the original input. Thus P_N is the intersection of finitely many strict affine half-spaces, so it is open and convex, and N is affine on P_N.

We show that $C_N(Q)$ has Hausdorff dimension less than r. Let g be a composition of hidden layers of width at most r, that is, a network without its final affine map, and let P_g be the set on which all its preactivations are positive. We prove by induction over the number of layers that $\dim_H g(Q\setminus P_g) < r$ for every compact Q in the domain of g. For one layer $g(x) = \rho(Ax)$, with A affine and $n \le r$ outputs, every $x \notin P_g$ has some $(Ax)_i \le 0$, so $g(x)_i = 0$; thus $g(Q\setminus P_g)$ lies in the union of the n coordinate hyperplanes of $\mathbb{R}^n$, which have dimension $n-1 \le r-1$. For a composition $g = g_2\circ g_1$, we have $P_g = P_{g_1}\cap T_1^{-1}(P_{g_2})$, where T_1 is the affine formula of g_1 on P_{g_1}, and

$$g(Q\setminus P_g) \subset g_2\big(g_1(Q\setminus P_{g_1})\big) \cup g_2\big(g_1(Q)\setminus P_{g_2}\big).$$

The map g_2 is Lipschitz, so it does not increase Hausdorff dimension; the first set has dimension less than r by induction, and so does the second, by induction applied to g_2 on the compact set $g_1(Q)$. The final affine map is Lipschitz as well, so $\dim_H C_N(Q) < r$. A subset of $\mathbb{R}^r$ with nonempty interior has Hausdorff dimension r; hence $C_N(Q)$ has empty interior. Finally, $Q\setminus P_N$ is compact and N is continuous, so $C_N(Q)$ is compact, and therefore nowhere dense. □

Proof of Theorem 3. For each j, apply Lemma 2 with the same compact set Q, and abbreviate $P_j = P_{N_j}$ and $C_j = C_{N_j}(Q)$. Since F is a homeomorphism, every $F^{-1}(C_j)$ is closed and nowhere dense. The open set U is a Baire space, and hence

$$\mathcal{R} := U \setminus F^{-1}\left(\bigcup_{j=1}^{\infty} C_j\right)$$

is dense in U.

Fix two points $x_0, x_1 \in \mathcal{R}$ and write $y_i = F(x_i)$ for $i \in \{0, 1\}$. Choose $\eta > 0$ such that the cubes $B_i = \{z : \|z - x_i\|_\infty \le \eta\}$ lie in U. The closed 1-neighborhood K of $F(\bigcup_i B_i)$ is compact, so F^{-1} is uniformly continuous on K; choose $\delta \in (0, 1]$ such that $w, w' \in K$ and $\|w - w'\| < \delta$ imply $\|F^{-1}(w) - F^{-1}(w')\|_\infty \le \eta$. For all sufficiently large j, $\|N_j - F\|_{C(Q)} < \delta$, and then $h(z) = x_i + z - F^{-1}(N_j(z))$ maps B_i into itself, because $\|F^{-1}(N_j(z)) - F^{-1}(F(z))\|_\infty \le \eta$. Brouwer's fixed point theorem gives $x_{ij} \in B_i \subset U$ with $F^{-1}(N_j(x_{ij})) = x_i$, that is, $N_j(x_{ij}) = y_i$. By the definition of $\mathcal{R}$, y_i lies outside C_j. Since $x_{ij} \in U \subset Q$, it follows that

$$x_{ij} \in P_j.$$

These preimages converge to the selected points. Indeed,

$$\|F(x_{ij}) - F(x_i)\| = \|F(x_{ij}) - N_j(x_{ij})\| \le \|F - N_j\|_{C(Q)} \longrightarrow 0.$$

Every subsequential limit of x_{ij} lies in the compact set Q and is therefore equal to x_i by continuity and injectivity of F. Thus $x_{ij} \to x_i$.

Let $\lambda \in [0, 1]$. Both P_j and U are convex, so

$$z_j := \lambda x_{0j} + (1 - \lambda) x_{1j} \in P_j \cap U.$$

Because N_j is a single affine map on P_j,

$$N_j(z_j) = \lambda N_j(x_{0j}) + (1 - \lambda) N_j(x_{1j}) = \lambda F(x_0) + (1 - \lambda) F(x_1).$$

Moreover, $z_j \to z := \lambda x_0 + (1 - \lambda) x_1$. Since $\|F(z_j) - N_j(z_j)\| \le \|F - N_j\|_{C(Q)} \to 0$ and $F(z_j) \to F(z)$ by continuity of F,

$$F\big(\lambda x_0 + (1 - \lambda) x_1\big) = \lambda F(x_0) + (1 - \lambda) F(x_1). \tag{9}$$

Both sides of (9) are continuous in (x_0, x_1), and $\mathcal{R}$ is dense in U, so (9) holds for all $x_0, x_1 \in U$ and $\lambda \in [0, 1]$. Fix a base point $p \in U$. For $v \in \mathbb{R}^r$ and $s > 0$ with $p + sv \in U$, the identity shows that $L(v) = (F(p + sv) - F(p))/s$ does not depend on s; midpoints give $L(-v) = -L(v)$ and $L(v + w) = L(v) + L(w)$, and $L(tv) = tL(v)$ for $t > 0$. Hence L is linear and $F(x) = F(p) + L(x - p)$ for all $x \in U$. □

D Proof of Proposition 11: no good direction for $r = 2d + 1$

At $r = 2d + 1$, the spans V_{ab} in the proof of Lemma 7 can be all of $\mathbb{R}^r$, so the lemma no longer applies. Proposition 11, stated in Section 5, shows that some embeddings have no good direction, and neither has any continuous map uniformly close to them. It limits the method of Section 5; it is *not* a lower bound for general ReLU networks.

Proof of Proposition 11. Write $e_1, \dots, e_r$ for the standard basis. For each i, list the remaining $2d$ basis vectors as $e_{i,1}, \dots, e_{i,2d}$, and consider the affine patches

$$A_i(s) = \tfrac{1}{2} e_i + \sum_{m=1}^{d} s_m e_{i,m}, \qquad B_i(t) = -\tfrac{1}{2} e_i + \sum_{m=1}^{d} t_m e_{i,d+m}, \qquad s, t \in [-2, 2]^d. \tag{10}$$

The patches need not be disjoint, because they are only approximated. Choose $2r$ pairwise disjoint small cubes in the open unit ball of $\mathbb{R}^d$, called parameter cubes, with affine charts $a_i, b_i : [-2, 2]^d \to \mathbb{D}^d$ onto them, and a continuous map $f : \mathbb{D}^d \to \mathbb{R}^r$ with $f \circ a_i = A_i$ and $f \circ b_i = B_i$ (use continuous cutoffs with disjoint supports). Since $r \ge 2d + 1$, smooth embeddings of $\mathbb{D}^d$ are dense in $C(\mathbb{D}^d, \mathbb{R}^r)$, by the torus argument in the proof of Theorem 9 (Hirsch, 1976, Chapter 2, Theorems 2.6 and 2.13). Below, $\|\cdot\|_\infty$ denotes the maximum norm on a coordinate space and the corresponding supremum norm of a map; this is the norm of Section 2. Choose a smooth embedding $E : \mathbb{D}^d \to \mathbb{R}^r$ with $\|E - f\|_\infty < \theta = \frac{1}{10}$, and take $\eta = \frac{1}{10}$.

Fix a nonzero $v \in \mathbb{R}^r$ and choose i with $|v_i| = \max_j |v_j| > 0$. For $z = (s, -t) \in \mathcal{Q} = [-2, 2]^{2d}$, the difference of the two patches is

$$H_0(z) := A_i(s) - B_i(t) = e_i + \sum_{m=1}^{2d} z_m e_{i,m}.$$

Thus every $z \in \mathcal{Q}$ determines two source points $a_i(s)$ and $b_i(t)$, which are distinct because the two parameter cubes are disjoint.

Let $\Lambda_v : \mathbb{R}^r \to \mathbb{R}^{2d}$ send y to $(y_j - (y_i/v_i)v_j)_{j\neq i}$, with the coordinates $j \neq i$ ordered as the vectors $e_{i,m}$. It is linear with kernel $\operatorname{span}(v)$, and $\|\Lambda_v y\|_\infty \leq 2\|y\|_\infty$ because $|v_j| \leq |v_i|$. On $\mathcal{Q}$, $\Lambda_v H_0(z) = z - c_v$ with $c_v = (v_j/v_i)_{j\neq i}$ and $\|c_v\|_\infty \leq 1$. Let $E' : \mathbb{D}^d \to \mathbb{R}^r$ be continuous with $\|E' - E\|_\infty < \eta$. Its difference map $H'(z) = E'(a_i(s)) - E'(b_i(t))$ is within $2(\theta + \eta)$ of H_0, so $G = \Lambda_v \circ H'$ satisfies $\|G(z) - (z - c_v)\|_\infty \leq 4(\theta + \eta) \leq 1$ on $\mathcal{Q}$. Hence $z \mapsto z - G(z)$ maps the cube $\{z : \|z - c_v\|_\infty \leq 1\} \subset \mathcal{Q}$ into itself, and Brouwer's fixed point theorem gives a zero of G: distinct source points $a \in a_i([-2, 2]^d)$ and $b \in b_i([-2, 2]^d)$ with $E'(a) - E'(b) \in \operatorname{span}(v)$. The constants θ and η do not depend on v. Finally, Palais's theorem (Palais, 1960, Theorem B), which applies because $d < r$, supplies a diffeomorphism D of $\mathbb{R}^r$ with $E = D \circ q_{d,r}$ on $\mathbb{D}^d$. □

The proof uses E' only on the $2r$ parameter cubes. Choosing these cubes inside a cube C contained in the open unit ball, and restricting E to C, gives the same statement for the cube $[0, 1]^d$ in place of $\mathbb{D}^d$, after an affine change of variables.

Proposition 11 shows that the construction of Theorem 8 cannot approximate E when $r = 2d + 1$. Indeed, the block B constructed in the proof of Lemma 6 keeps $y = \pi_v(z)$ unchanged, that is, $\pi_v \circ B|_M = \pi_v|_M$. Hence $B|_M$ is injective, and v is a good direction for $B(M)$. An invertible affine map A applied afterwards is injective and turns v into the good direction Av, where A acts on v through its linear part. The flat set $q_{d,r}(\mathbb{D}^d)$ itself has good directions, because $d < r$. By induction, a composition N of such blocks and invertible affine maps, as in the proof of Theorem 8, is injective on $q_{d,r}(\mathbb{D}^d)$, and the image $N(q_{d,r}(\mathbb{D}^d))$ has a good direction v. If $E' = N \circ q_{d,r}$ satisfied $\|E' - E\|_{C(\mathbb{D}^d)} < \eta$, Proposition 11 would give distinct a, b with $E'(a) - E'(b) \in \operatorname{span}(v)$ and, by injectivity, $E'(a) \neq E'(b)$, so v would not be a good direction. Hence $\|N \circ q_{d,r} - E\|_{C(\mathbb{D}^d)} \geq \eta$.

This conclusion applies only to the construction used in Theorem 8. A general width-r ReLU network need not decompose into such blocks, and its restriction to the flat cube need not have a good direction. Consequently the minimum width for $k = 2d + 1$ remains open:

$$2d + 1 \;\leq\; W_{\mathrm{ReLU}}(d, 2d + 1) \;\leq\; 2d + 2.$$

E ReLU-emulating activations

Recall from Section 5 that a continuous $\sigma : \mathbb{R} \to \mathbb{R}$ *emulates ReLU on compact sets* if there are affine maps $\alpha_n, \beta_n : \mathbb{R} \to \mathbb{R}$ such that $\beta_n \circ \sigma \circ \alpha_n \to \mathrm{ReLU}$ uniformly on compact intervals.

Lemma 20 (Sufficient conditions)**.** *1. If $\sigma(y)/y \to c \neq 0$ as $y \to +\infty$ and $\sigma(y)/y \to 0$ as $y \to -\infty$, then $\sigma(\lambda x)/(c\lambda) \to \mathrm{ReLU}(x)$ uniformly on compact intervals as $\lambda \to \infty$. This covers GELU, SiLU, Mish, Softplus, ELU, CELU, and SELU.*

2. If σ is affine with slope $s \neq 0$ on an interval $[a, b]$, $a < b$, and constant on $(-\infty, a]$, then for every $\lambda > 0$ the function $x \mapsto \lambda\big(\sigma(a + x/\lambda) - \sigma(a)\big)/s$ agrees with ReLU on $(-\infty, \lambda(b - a)]$; hence σ emulates ReLU on compact sets. This covers ReLU6, hard-tanh, and hard-sigmoid.

Proof. (1) Fix $R > 0$ and $\eta > 0$. Choose $Y > 0$ such that $|\sigma(y)/y - c| < \eta$ for $y \geq Y$ and $|\sigma(y)/y| < \eta$ for $y \leq -Y$, and let $S = \sup_{|y|\leq Y} |\sigma(y)|$, which is finite by continuity. Let $|x| \leq R$ and $\lambda > 0$. If $\lambda x \geq Y$, then $|\sigma(\lambda x)/(c\lambda) - x| = |x|\,|\sigma(\lambda x)/(c\lambda x) - 1| \leq R\eta/|c|$. If $\lambda x \leq -Y$, then $\mathrm{ReLU}(x) = 0$ and $|\sigma(\lambda x)/(c\lambda)| = |x|\,|\sigma(\lambda x)/(\lambda x)|/|c| \leq R\eta/|c|$. If $|\lambda x| < Y$, then $\mathrm{ReLU}(x) \leq |x| < Y/\lambda$ and $|\sigma(\lambda x)/(c\lambda)| \leq S/(|c|\lambda)$. Hence, for every $\lambda > 0$,

$$\sup_{|x|\leq R} \Big| \frac{\sigma(\lambda x)}{c\lambda} - \mathrm{ReLU}(x) \Big| \leq \frac{R\eta}{|c|} + \frac{S/|c| + Y}{\lambda}.$$

Let $\lambda \to \infty$ and then $\eta \to 0$. The listed activations satisfy the two limits with $c = 1$, except SELU, whose limit c is its output scale. (2) If $\sigma(t) = \sigma(a) + s(t - a)$ on $[a, b]$ with $s \neq 0$ and $\sigma \equiv \sigma(a)$ on $(-\infty, a]$, then $\lambda\big(\sigma(a + x/\lambda) - \sigma(a)\big)/s = \mathrm{ReLU}(x)$ for all $x \leq \lambda(b - a)$. □

Proof of Corollary 10. Let $N = A_{L+1} \circ \rho \circ A_L \circ \cdots \circ \rho \circ A_1$ be a ReLU network of hidden width at most m, where $\rho = \mathrm{ReLU}$ acts coordinatewise, and let $K \subset \mathbb{R}^d$ be compact. Put $\rho_n := \beta_n \circ \sigma \circ \alpha_n$,

acting coordinatewise, and let N_n be the network obtained from N by replacing every ρ by ρ_n. Since α_n and β_n are affine, N_n is an ordinary σ-network with the same hidden widths: absorb α_n into A_ℓ and β_n into $A_{\ell+1}$.

Let x_ℓ and $y_\ell^{(n)}$ denote the hidden states of N and N_n after ℓ layers on a common input in K, with $x_0 = y_0^{(n)}$. We show by induction on ℓ that $\sup_K \|y_\ell^{(n)} - x_\ell\| \to 0$ as $n \to \infty$. Suppose this holds for $\ell - 1$. The set $A_\ell(x_{\ell-1}(K))$ is compact; let C_ℓ be its closed unit neighborhood. For all large n, $A_\ell y_{\ell-1}^{(n)} \in C_\ell$ on K, and

$$\big\|\rho_n(A_\ell y_{\ell-1}^{(n)}) - \rho(A_\ell x_{\ell-1})\big\| \le \sup_{C_\ell} \|\rho_n - \rho\| + \big\|\rho(A_\ell y_{\ell-1}^{(n)}) - \rho(A_\ell x_{\ell-1})\big\|.$$

Both terms tend to zero uniformly on K, the first by uniform convergence of ρ_n on the compact set of coordinate values of C_ℓ, the second because ReLU is uniformly continuous on that compact set and $\|A_\ell y_{\ell-1}^{(n)} - A_\ell x_{\ell-1}\| \le \|A_\ell\|\, \|y_{\ell-1}^{(n)} - x_{\ell-1}\|$. Continuity of A_{L+1} completes the induction, so N is a uniform limit on K of width-m σ-networks. Therefore density of width-m ReLU networks in $C([0,1]^d, \mathbb{R}^k)$ implies density of width-m σ-networks, that is, $W_\sigma(d,k) \le W_{\mathrm{ReLU}}(d,k)$. The remaining assertions follow from Theorem 9, the lower bound $W_\sigma(d,k) \ge k$, which holds for every activation by the last paragraph of the proof of Theorem 9, and the bound $W_{\mathrm{ReLU}}(d,k) \le \min\{d+k, w(d,k)+1\}$ of Section 5. □

Remark 21 (Fixed ELU and SELU)**.** *The proof above uses only that* ReLU *is continuous: if $\beta_n \circ \sigma \circ \alpha_n \to \rho$ uniformly on compact intervals for some continuous $\rho : \mathbb{R} \to \mathbb{R}$, the same induction gives $W_\sigma(d,k) \le W_\rho(d,k)$. For $\alpha > 0$ let $\mathrm{ELU}_\alpha(x) = x$ for $x \ge 0$ and $\mathrm{ELU}_\alpha(x) = \alpha(\mathrm{e}^x - 1)$ for $x < 0$, and let $\ell_\alpha(x) = x$ for $x \ge 0$ and $\ell_\alpha(x) = \alpha x$ for $x < 0$. For $-R \le x \le 0$ and $h > 0$, $|\mathrm{ELU}_\alpha(hx)/h - \alpha x| = \alpha|\mathrm{e}^{hx} - 1 - hx|/h \le \alpha h R^2/2$, so $\mathrm{ELU}_\alpha(hx)/h \to \ell_\alpha(x)$ uniformly on compact intervals as $h \downarrow 0$, and $W_{\mathrm{ELU}_\alpha}(d,k) \le W_{\ell_\alpha}(d,k)$. For $\alpha \ne 1$, ℓ_α is Leaky-ReLU with the fixed slope α, and $W_{\ell_\alpha}(d,k) = w(d,k)$: Hwang (2025b, Corollary 4.6) proves this for Leaky-ReLU with layerwise slopes, and Yang et al. (2025, Theorem 14) transfer it to every fixed slope in $(0,1) \cup (1,\infty)$. Since ELU_α is increasing and continuous, Hwang (2025b, Theorem 4.5) gives $W_{\mathrm{ELU}_\alpha}(d,k) \ge w(d,k)$, as does Proposition 12, and output scaling covers $\mathrm{SELU}_{\lambda,\alpha} = \lambda\,\mathrm{ELU}_\alpha$. Therefore*

$$W_{\mathrm{ELU}_\alpha}(d,k) = W_{\mathrm{SELU}_{\lambda,\alpha}}(d,k) = w(d,k) \qquad (\lambda > 0,\ \alpha > 0,\ \alpha \ne 1),$$

which equals k for $k > 2d$ and covers the standard SELU parameters. This value is not a contribution of this paper, but it is also not the statement of Yang et al. (2025, Theorems 16 and 17): there, ELU and SELU denote families whose parameter may change from layer to layer, and the proof of their Theorem 16 uses different parameters in successive layers. The standard ELU_1, whose rescaling limit is the identity, is covered only by Corollary 10; its case $k = 2d+1$ is open, as for ReLU.

F WIDTH THREE FOR CURVES IN $\mathbb{R}^3$ ($d = 1$, $k = 3$)

Proposition 11 limits the construction of Section 5. For $d = 1$ and $k = 3$, this appendix shows what ReLU networks of width three do achieve. We call a curve g in $\mathbb{R}^3$ *projectable* if $\pi \circ g$ is injective for some linear map $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ of rank two, that is, if the curve has a good direction.

Proposition 22 (Projectable curves)**.** *Let $K \subset \mathbb{R}$ be compact and $g \in C(K, \mathbb{R}^3)$. If some linear map $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ of rank two is such that $\pi \circ g$ is injective on K, then g is a uniform limit on K of ordinary ReLU networks of hidden width three. The same holds for every uniform limit of such maps.*

Proof. Put $y = \pi \circ g$ and $Y = y(K)$; choose a nonzero $v \in \ker \pi$ and a linear right inverse $J : \mathbb{R}^2 \to \mathbb{R}^3$ of π. Since K is compact and y is injective, $y : K \to Y$ is a homeomorphism, and $g(t) - Jy(t) \in \ker \pi$ for every $t \in K$. Hence there is a unique continuous $h : Y \to \mathbb{R}$ with $g = Jy + v\,(h \circ y)$ on K; by Tietze's theorem, extend h to $H \in C(\mathbb{R}^2)$. Fix $\varepsilon > 0$ and a compact neighborhood Q of Y. Because $W_{\mathrm{ReLU}}(1,2) \le 3$ (Hanin & Sellke, 2017) and H is uniformly continuous on Q, there is a width-three network $A : \mathbb{R} \to \mathbb{R}^2$ with $A(K) \subset Q$ and $\sup_K \|JA + v\,(H \circ A) - g\| < \varepsilon/2$. The graph form of the max–min construction recalled before Lemma 6, with input dimension two, gives a width-three network $B : \mathbb{R}^2 \to \mathbb{R}^3$ whose hidden state on Q is $(z + c, \widehat{H}(z) + C)$ with $\sup_Q |\widehat{H} - H| < \varepsilon/(2\|v\|)$; decode it by $(z', s') \mapsto J(z' - c) + v(s' - C)$,

so that $\sup_Q \|B - (J + vH)\| < \varepsilon/2$. Absorb the affine output of A into the first affine map of B. The composition is an ordinary network of hidden width three within ε of g on K. A diagonal argument proves the closure statement. □

Proposition 23 (An individual target of exact width three)**.** *On $K = [0, 3]$, the polygonal curve*

$$p(t) = \big(t,\ \mathrm{ReLU}(t-1) - \mathrm{ReLU}(t-2),\ \mathrm{ReLU}(t-2)\big)$$

is represented exactly by a ReLU network with one hidden layer of width three, and no sequence of ReLU networks of hidden width at most two converges to p uniformly on K.

Proof. The hidden layer $(\mathrm{ReLU}(t), \mathrm{ReLU}(t-1), \mathrm{ReLU}(t-2))$ followed by the affine map $(h_1, h_2, h_3) \mapsto (h_1, h_2 - h_3, h_3)$ represents p on K, where $t \geq 0$. The values $p(0), p(1), p(2), p(3)$ are $(0, 0, 0)$, $(1, 0, 0)$, $(2, 1, 0)$, $(3, 1, 1)$; the determinant of their differences from $p(0)$ equals one, so they are affinely independent. A network of hidden width at most two, or with no hidden layer, has its image in an affine plane, so its values at $0, 1, 2, 3$ are coplanar. Coplanar quadruples form the zero set of a determinant, a closed subset of $(\mathbb{R}^3)^4$ at positive distance from the target quadruple. Uniform convergence is therefore impossible. □

Width-three networks approximate more than the uniform limits of projectable curves. The next proposition exhibits an embedded arc that, like the embedding of Proposition 11, is not projectable even after small perturbations, and that is nevertheless a uniform limit of width-three networks. The construction applies one additional ReLU layer to a projectable curve.

Proposition 24 (A nonprojectable arc of width three)**.** *There is a polygonal embedding $E : [0, 1] \to \mathbb{R}^3$ with the following two properties.*

1. *E is a uniform limit of ordinary ReLU networks of hidden width three, and no sequence of ReLU networks of hidden width at most two converges uniformly to E.*
2. *If $E' : [0, 1] \to \mathbb{R}^3$ is continuous and $\|E' - E\|_\infty < 1/8$, where $\|\cdot\|_\infty$ is the supremum over $[0, 1]$ of the maximum norm on $\mathbb{R}^3$, then, for every linear map $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ of rank two, $\pi \circ E'$ is not injective.*

Consequently, the uniform closure of the curves in Proposition 22 is a proper subset of the uniform closure of width-three networks in $C([0, 1], \mathbb{R}^3)$.

Proof. A projectable precursor. Let $q : [0, 13] \to \mathbb{R}^3$ interpolate affinely between the vertices $q(j)$, $j = 0, \ldots, 13$, in the following table, and let $\pi_0(x, y, z) = (y, 10z - x)$.

j	$q(j)$	$\pi_0(q(j))$	j	$q(j)$	$\pi_0(q(j))$
0	$(-40, 0, 2)$	$(0, 60)$	7	$(10, 19/2, 2)$	$(19/2, 10)$
1	$(-40, 0, -2)$	$(0, 20)$	8	$(8, 21/2, 0)$	$(21/2, -8)$
2	$(-40, -3, -2)$	$(-3, 20)$	9	$(12, 21/2, 0)$	$(21/2, -12)$
3	$(1, -3, -3)$	$(-3, -31)$	10	$(10, 18, -1/2)$	$(18, -15)$
4	$(1, -2, 0)$	$(-2, -1)$	11	$(10, 22, -1/2)$	$(22, -15)$
5	$(1, 2, 0)$	$(2, -1)$	12	$(8, 20, 1/2)$	$(20, -3)$
6	$(10, 19/2, -2)$	$(19/2, -30)$	13	$(12, 20, 1/2)$	$(20, -7)$

The polygon $\pi_0 \circ q$, with coordinates (u, v), is simple. Its first five segments are vertical, horizontal, vertical, a segment in $-3 \leq u \leq -2$ with $v \leq -1$, and horizontal at $v = -1$; from the listed coordinates, nonconsecutive ones are disjoint. From $j = 5$ to $j = 10$, u is nondecreasing, and the only vertical segments lie on the distinct lines $u = 19/2$ and $u = 21/2$, so again nonconsecutive segments are disjoint. The last three segments lie in $18 \leq u \leq 22$ and meet only at consecutive vertices. These three parts satisfy $u \leq 2$, $2 \leq u \leq 18$, and $u \geq 18$, and attain $u = 2$ and $u = 18$ only at the shared vertices $j = 5$ and $j = 10$. Hence $\pi_0 \circ q$ is injective, and Proposition 22 gives width-three networks $N_n \to q$ uniformly on $[0, 13]$.

One clipping layer. Put $C(x, y, z) = (\mathrm{ReLU}(x), y, z)$ and $E(t) = C(q(13t))$ for $t \in [0, 1]$. Since $y \geq -3$ and $z \geq -3$ on $q([0, 13])$, the map C agrees there with

$$L(x, y, z) = \big(\mathrm{ReLU}(x),\ \mathrm{ReLU}(y + 5) - 5,\ \mathrm{ReLU}(z + 5) - 5\big),$$

one hidden layer of width three between two affine maps. Appending L to N_n and absorbing adjacent affine maps, including the reparametrization $t \mapsto 13t$, gives ordinary networks of hidden width three, and $\sup_t \|L(N_n(13t)) - E(t)\| \le \|N_n - q\|_{C([0,13])} \to 0$ because L is 1-Lipschitz. Moreover, E is an embedding. The first coordinate of q is negative exactly on $[0, 2 + \frac{40}{41})$, where E traces the simple planar chain $(0,0,2) \to (0,0,-2) \to (0,-3,-2) \to (0,-3,-\frac{122}{41})$ in $\{x = 0\}$; on the remaining parameters the first coordinate of q is nonnegative, E coincides with the injective map $q(13\,\cdot)$, and the image lies in $\{x > 0\}$ except at the junction $(0,-3,-\frac{122}{41})$, which is the only point shared by the two parts.

The lower bound. The points $E(0)$, $E(\frac{1}{13})$, $E(\frac{2}{13})$, $E(\frac{3}{13})$ are $(0,0,2)$, $(0,0,-2)$, $(0,-3,-2)$, $(1,-3,-3)$; the determinant of their differences from the first is -12. The argument of Proposition 23 shows that no networks of hidden width at most two converge uniformly to E. This proves (1).

Pairs of segments. The image of E contains the six segments

$$\begin{aligned} A_1(s) &= (1, s, 0), & B_1(t) &= (0, 0, t), \\ A_2(s) &= (10 + s, \tfrac{21}{2}, 0), & B_2(t) &= (10, \tfrac{19}{2}, t), & s, t \in [-2, 2], \\ A_3(s) &= (10 + s, 20, \tfrac{1}{2}), & B_3(t) &= (10, 20 + t, -\tfrac{1}{2}), \end{aligned}$$

the images of the pairwise disjoint parameter intervals $[4,5]$, $[0,1]$, $[8,9]$, $[6,7]$, $[12,13]$, $[10,11]$ of q. The differences $A_i(s) - B_i(t)$ equal $(1, s, -t)$, $(s, 1, -t)$, and $(s, -t, 1)$, as for the patches (10) in the proof of Proposition 11 with $d = 1$.

Small perturbations. Fix a nonzero $w \in \mathbb{R}^3$ and choose i with $|w_i| = \max_j |w_j| > 0$; let $j < l$ be the other two indices. With $\zeta = (\zeta_1, \zeta_2) = (s, -t)$, the i-th difference is $H_0(\zeta) = e_i + \zeta_1 e_j + \zeta_2 e_l$ on $\mathcal{Q} = [-2, 2]^2$. Let $\Lambda_w(\xi) = (\xi_j - (\xi_i/w_i)w_j,\ \xi_l - (\xi_i/w_i)w_l)$ for $\xi \in \mathbb{R}^3$; it is linear with kernel $\mathbb{R}w$, satisfies $\|\Lambda_w \xi\|_\infty \le 2\|\xi\|_\infty$, and $\Lambda_w H_0(\zeta) = \zeta - c_w$ with $c_w = (w_j/w_i, w_l/w_i)$ and $\|c_w\|_\infty \le 1$. If $\|E' - E\|_\infty < 1/8$, the corresponding difference map of E' on the two parameter intervals is uniformly within $1/4$ of H_0, so its image $G(\zeta)$ under Λ_w is within $1/2$ of $\zeta - c_w$ on $\mathcal{Q}$. The map $\zeta \mapsto \zeta - G(\zeta)$ therefore sends the cube $\{\zeta : \|\zeta - c_w\|_\infty \le 1\} \subset \mathcal{Q}$ into itself, and Brouwer's fixed point theorem gives a parameter ζ with $G(\zeta) = 0$, that is, two source points whose images under E' differ by a multiple of w. The two source points lie in disjoint parameter intervals, so they are distinct, and they have the same image under every rank-two π with $\ker \pi = \mathbb{R}w$. As w was arbitrary, (2) follows.

Finally, (2) places a uniform neighborhood of E outside the class of Proposition 22, so E is not in its closure, whereas (1) places E in the closure of width-three networks. □

Proposition 11 with $d = 1$ shows that projectable curves are not dense in $C(\mathbb{D}^1, \mathbb{R}^3)$, so Proposition 22 alone does not give $W_{\mathrm{ReLU}}(1,3) = 3$. Proposition 23 gives a target that needs width three; this only confirms the general lower bound $W_{\mathrm{ReLU}}(1,3) \ge 3$ and does not decide whether width four is needed. Proposition 24 shows in addition that the absence of a good direction, even after small perturbations, does not force width four, even for embedded arcs: one additional ReLU layer applied to a projectable curve reaches a curve at uniform distance at least $1/8$ from every projectable curve. A proof of $W_{\mathrm{ReLU}}(1,3) = 3$ must approximate every such arc, by steps other than the graph updates of Lemma 6; a proof of $W_{\mathrm{ReLU}}(1,3) = 4$ must separate some target from all width-three networks by a property other than projectability. Since $w(1,3) = 3$, the value $W_{\mathrm{ReLU}}(1,3) = 4$ would refute the conjecture of Section 6, whereas $W_{\mathrm{ReLU}}(1,3) = 3$ would confirm only that case. The analogue of Proposition 22 for $d \ge 2$ would need $W_{\mathrm{ReLU}}(d, 2d) = 2d + 1$, which is itself open (Table 1).

G Density of max–min strings

Section 5 uses the density of max–min strings (Hanin & Sellke, 2017, Proposition 3). The printed proof treats input dimension two in detail and describes higher dimensions by rotating the planar pictures; its Lemma 5 also yields the error 2ε instead of ε on the new triangle. We give a complete argument in every dimension. A *max–min string* on $\mathbb{R}^n$ is an affine function or $\max(\ell, g)$ or $\min(\ell, g)$ with ℓ affine and g a max–min string. We write $|\cdot|$ and $\langle \cdot, \cdot \rangle$ for the Euclidean norm and inner product and $\mathbb{B}(r)$ for the closed ball of radius r.

Proposition 25 (Density of max–min strings)**.** *For every compact $K \subset \mathbb{R}^n$, every $f \in C(K)$, and every $\varepsilon > 0$ there is a max–min string g with $|g - f| < \varepsilon$ on K.*

Proof. Reduction. Restrictions to K of globally Lipschitz functions form a lattice that contains the affine functions and therefore separates points strongly; by the lattice Stone–Weierstrass theorem there is an L-Lipschitz F on $\mathbb{R}^n$ with $|F - f| < \varepsilon/2$ on K. Choose R with $K \subset \mathbb{B}(R)$ and put $\eta = \varepsilon/2$. It suffices to find a string within η of F on $\mathbb{B}(R)$. If $\eta \geq LR$, the constant $F(0)$ works; assume $\eta < LR$ and put

$$r_0 = \frac{\eta}{L}, \qquad t = \frac{\eta^2}{49L^2R}, \qquad \mu = \frac{14L^2R}{\eta},$$

so that $0 < t < R/49$, $\mu^2 t = 4L^2R$, and $7L\sqrt{Rt} = \eta$.

Cut lemma. Let S, N be sets, $a \in \mathbb{R}$, and ℓ a function. If $|g - F| \leq \eta$ and $|F - a| \leq \ell + \eta$ on S, and $|F - a| + |\ell| \leq \eta$ on N, then $\hat{g} = \max(a - \ell, \min(a + \ell, g))$ satisfies $|\hat{g} - F| \leq \eta$ on $S \cup N$. Indeed, on S both $a + \ell$ and g are at least $F - \eta$, while $a - \ell$ and g are at most $F + \eta$; on N one checks $|\hat{g} - a| \leq |\ell|$ separately for $\ell \geq 0$ and $\ell < 0$. If g is a string and ℓ affine, $\hat{g}$ is a string.

One cone. Fix a unit vector u and $0 < r' \leq R$; put $c = r' + t$, $Z = cu$, $\ell(x) = \mu(c - \langle u, x\rangle)$, and $C(u, r') = \{x \in \mathbb{B}(r') : \langle u, x\rangle \geq r' - t\}$. For $x \in \mathbb{B}(r')$ let $s = c - \langle u, x\rangle \geq t$. Then $|x - Z|^2 \leq r'^2 - 2c\langle u, x\rangle + c^2 \leq 2cs$, so $L^2|x - Z|^2 \leq 4L^2Rs = \mu^2 ts \leq \mu^2 s^2$, that is, $L|x - Z| \leq \ell(x)$. For $x \in C(u, r')$ also $|x - Z|^2 \leq t^2 + 2ct \leq 5Rt$ and $s \leq 2t$, so $L|x - Z| + |\ell(x)| \leq 3L\sqrt{Rt} + 2\mu t = 7L\sqrt{Rt} = \eta$. With $a = F(Z)$, the cut lemma therefore extends any string that is η-close on a set $S \subset \mathbb{B}(r')$ to one that is η-close on $S \cup C(u, r')$.

Caps cover a shell. By compactness of the unit sphere, choose finitely many unit vectors $\mathcal{U}$ such that every unit vector v has $u \in \mathcal{U}$ with $|u - v|^2 \leq t/R$. Let $r_0 \leq r < R$ and $r' = \min(r + t/2, R)$. If $r < |x| \leq r'$ and $v = x/|x|$, then $\langle u, v\rangle \geq 1 - t/(2R) > 0$ and $\langle u, x\rangle \geq r(1 - t/(2R)) \geq r - t/2 \geq r' - t$, so $x \in C(u, r')$.

Rounds. If a string is η-close on $\mathbb{B}(r)$, applying the cone step once for each $u \in \mathcal{U}$ keeps every intermediate set inside $\mathbb{B}(r')$, so no earlier cap is lost, and yields a string η-close on $\mathbb{B}(r')$. The constant $F(0)$ is η-close on $\mathbb{B}(r_0)$, and after at most $2(R - r_0)/t$ rounds the radius reaches R. □