# CODED MANIFOLDS PRESERVE NEURAL HARDNESS AT CRITICAL REACH

**Pahan Dewasurendra**
Johns Hopkins University

## ABSTRACT

Does sufficiently low curvature make neural networks easy to learn under the manifold hypothesis? Recent work nearly identified a sharp geometric boundary. It proved exponential hardness on manifolds of reach $O(n^\alpha)$ for every $\alpha < 1/2$, while reach $\omega(\sqrt{n})$ forces polynomial intrinsic volume and permits interpolation. The critical regime $\Theta(\sqrt{n})$ was left open.

We close this boundary on the hard side. For every large ambient dimension $n$, we construct a connected closed $C^\infty$ curve in $[0,1]^n$ with reach $\Theta(\sqrt{n})$ and a full-support, polynomial-time samplable distribution whose density is within constant factors of arclength. Learning linear-width one-hidden-layer ReLU networks on this distribution from noise-free labels requires $\tau^2 \exp(\Omega(n))$ full statistical queries of tolerance $\tau$. Under polynomial-modulus Learning with Rounding, no polynomial-time learner exists even with arbitrary output hypotheses.

The construction replaces coordinate repetition by coding. We traverse a reflected Gray code of $\mathbb{F}_2^m$, embed its messages with an explicit systematic small-bias code, and smoothly round the resulting polygon. Character orthogonality gives a restricted isometry for every sparse signed measure needed by a chord and tangent. Federer's tangent formula then amplifies constant abstract reach to $\Omega(\sqrt{n})$. The systematic coordinates preserve a linear Boolean prefix on all but an exponentially small fraction of local pieces, which yields an almost pairwise-independent family of continuous parity networks. This shows that the $\sqrt{n}$ threshold is a genuine critical point rather than a limitation of the previous construction.

## 1 INTRODUCTION

The manifold hypothesis is often invoked to explain why learning in a high-dimensional ambient space can avoid the curse of dimensionality. If data lie on a low-dimensional regular manifold, then sample complexity may depend primarily on intrinsic dimension and geometric condition numbers. This principle underlies manifold reconstruction, geometric inference, and many analyses of learning on structured data (Narayanan & Mitter, 2010; Fefferman et al., 2016; Niyogi et al., 2008).

Low intrinsic dimension and low curvature alone do not guarantee efficient learning. Kiani et al. (2024) constructed one-dimensional manifolds in the Boolean cube on which learning shallow ReLU networks is exponentially hard. Their curves follow a reflected Gray code and repeat every Boolean coordinate enough times to obtain a prescribed reach. If the reach is $R$, this repetition uses $\Theta(R^2)$ ambient coordinates per independent Boolean bit. It leaves $\Theta(n/R^2)$ hard bits and therefore proves exponential hardness whenever

$$R = O(n^\alpha), \qquad \alpha < \frac{1}{2}.$$

The same work proves a complementary geometric fact. For fixed intrinsic dimension, a manifold in $[0,1]^n$ with reach $\omega(\sqrt{n})$ has polynomial volume and is efficiently covered by samples. A simple interpolation argument then learns bounded-Lipschitz networks. The authors explicitly leave reach $\Theta(\sqrt{n})$ open.

At this boundary, coordinate repetition fails completely. It leaves only $O(1)$ independent bits. A different geometry is necessary. Our key observation is that repetition enforces separation coordinate by coordinate, while reach only asks that a small family of geometric vectors be separated in

Euclidean norm. A small-bias code can enforce the latter simultaneously for exponentially many local configurations using only linearly many coordinates.

We first form a smooth abstract curve through the orthogonal vertices $\{e_z : z \in \mathbb{F}_2^m\} \subset \mathbb{R}^{2^m}$ in reflected Gray order. Every point, tangent, and local derivative is supported on at most three adjacent messages. We then map $e_z$ to a systematic codeword

$$C(z) = \big(z, \mathbf{1}\{\langle R_1, z\rangle = 1\}, \dots, \mathbf{1}\{\langle R_N, z\rangle = 1\}\big),$$

where $R_1, \dots, R_N$ form a constant-bias multiset in $\mathbb{F}_2^m$. Such multisets of linear size have deterministic polynomial-time constructions (Naor & Naor, 1993; Ta-Shma, 2017). For a uniform character $R$ and any real signed measure $w = (w_z)$ of total mass zero, character orthogonality gives

$$\mathbb{E}_R \left( \sum_z w_z \mathbf{1}\{\langle R, z\rangle = 1\} \right)^2 = \frac{1}{4} \sum_z w_z^2. \tag{1}$$

Small bias turns this identity into a deterministic restricted isometry for every constant-sparse signed measure. Since the map scales every chord, tangent, and chord–tangent residual by $\Theta(\sqrt{N})$, Federer's characterization of reach scales the abstract curve's constant reach to $\Omega(\sqrt{n})$ (Federer, 1959; Aamari et al., 2019). The reverse bound is topological: a closed curve in the unit cube has reach at most its diameter, hence at most $\sqrt{n}$.

Coding does not destroy the hard labels. The first $m$ coordinates are the message itself. Along the reflected Gray order, the first $d = \lfloor m/2 \rfloor$ message bits are constant on blocks of $2^{m-d}$ consecutive vertices. Except for an exponentially small fraction of pieces near block boundaries, these coordinates remain exactly Boolean throughout the smooth curve. A triangular-wave ReLU network computes every parity of this Boolean prefix. Two independent decoded prefixes are linearly independent over $\mathbb{F}_2$ except with exponentially small probability, so uniform random parity labels are pairwise independent on almost every pair of inputs. The full statistical-query theorem of Chen et al. (2022) then gives exponential noise-free hardness.

**Contributions.** We prove the following results.

1. We answer the critical-reach question of Kiani et al. (2024): exponential learning hardness persists at reach $\Theta(\sqrt{n})$.
2. We introduce a small-bias character isometry for constant-sparse signed measures on the Boolean cube and a reach-transfer lemma for sparse smooth curves.
3. We obtain $\tau^2 \exp(\Omega(n))$ noise-free lower bounds for the full SQ model, not only correlational SQ, for linear-width one-hidden-layer ReLU networks.
4. The hard input law has full support and density comparable to arclength, and it can be sampled to arbitrary precision in polynomial time.
5. Under polynomial-modulus Learning with Rounding, the same geometry is hard for every polynomial-time learner of polynomial-size one-hidden-layer ReLU networks.

## 2 Related work

**Learning under the manifold hypothesis.** Reach is the largest radius of a tubular neighborhood on which nearest-point projection is unique. It controls both local curvature and global bottlenecks (Federer, 1959; Aamari et al., 2019). It is a standard regularity parameter in manifold reconstruction and testing (Niyogi et al., 2008; Narayanan & Mitter, 2010; Fefferman et al., 2016). Linear near-isometries can preserve reach, although existing general bounds depend on global singular values (Eftekhari & Wakin, 2017). Our proof instead needs an isometry only on sparse chord–tangent residuals. The direct predecessor to this paper is Kiani et al. (2024), which establishes the $o(\sqrt{n})$ hard regime, the $\omega(\sqrt{n})$ volume regime, and the open critical case. Our result changes only the critical equality case. It does not contradict their interpolation theorem, which needs reach asymptotically larger than $\sqrt{n}$.

Table 1: Closest geometric and neural-network hardness results. "Full SQ" permits arbitrary bounded queries of both inputs and labels.

| Result | Input law | Target | Guarantee |
|---|---|---|---|
| Diakonikolas et al. (2020) | Gaussian, noisy | one hidden layer | restricted/noisy SQ |
| Chen et al. (2022) | Gaussian, noise-free | two hidden layers | full SQ |
| Kiani et al. (2024) | reach $O(n^\alpha)$, $\alpha < 1/2$ | one hidden layer | full SQ |
| This paper | reach $\Theta(\sqrt{n})$, noise-free | one hidden layer | full SQ and LWR |

**Hardness of shallow networks.** Statistical-query lower bounds for one-hidden-layer ReLU networks were previously known under Gaussian inputs with additive noise or in restricted SQ models (Diakonikolas et al., 2020; Vempala & Wilmes, 2019). Chen et al. (2022) obtained full-SQ lower bounds in the noise-free Gaussian setting for two-hidden-layer networks by constructing almost pairwise-independent function families. Their general pairwise-independence theorem is the hardness engine used here. Our manifold makes a simpler one-hidden-layer parity family almost pairwise independent without adding label noise. Cryptographic hardness for improper neural-network learning follows from local pseudorandom generators (Daniely & Vardi, 2021) and from Learning with Rounding (Chen et al., 2022). We transport the latter Boolean-cube result through the same systematic coordinates.

## 3 Setting and Main Results

For a closed set $M \subset \mathbb{R}^n$, let $\operatorname{reach}(M)$ be the supremum of radii $r$ such that every point at distance less than $r$ from $M$ has a unique nearest point on $M$. For a smooth embedded manifold, Federer gives the tangent formula

$$\operatorname{reach}(M) = \inf_{p \neq q \in M} \frac{\|q-p\|_2^2}{2\,\operatorname{dist}(q-p, T_pM)}. \tag{2}$$

This form exposes exactly which vectors a linear embedding must preserve.

An SQ oracle for a labeled distribution $P$ accepts any query $\phi(x,y) \in [-1,1]$ and returns a value within tolerance $\tau$ of $\mathbb{E}_P\phi(X,Y)$. A learner succeeds to squared error $\epsilon$ if it outputs any measurable $h$ satisfying $\mathbb{E}(h(X)-Y)^2 \le \epsilon$. The output need not be a neural network.

**Theorem 1** (Hardness at critical reach)**.** *There are universal constants $c, C, \rho, \tau_0 > 0$ and, for every sufficiently large $n$, a connected closed $C^\infty$ curve $M_n \subset [0,1]^n$, a distribution $\mathcal{D}_n$ on $M_n$, and a class $\mathcal{H}_n$ of one-hidden-layer ReLU networks such that:*

1. *$c\sqrt{n} \le \operatorname{reach}(M_n) \le C\sqrt{n}$.*
2. *$\mathcal{D}_n$ has full support. Its density relative to normalized arclength lies in $[c, C]$, and it is samplable to $b$ bits of precision in time polynomial in $n + b$.*
3. *Every member of $\mathcal{H}_n$ has width at most $Cn$, output in $[-1,1]$ on $M_n$, and weights and biases bounded by $Cn$.*
4. *For every $0 < \tau \le \tau_0$, an SQ learner that, with success probability at least $2/3$, learns every $h \in \mathcal{H}_n$ from the noise-free distribution $(X, h(X))$, $X \sim \mathcal{D}_n$, to squared error at most $1/16$ needs at least*

$$c\tau^2 \exp(\rho n)$$

*queries of tolerance $\tau$.*

The input distribution is fixed for each $n$ and does not depend on the target network. The result also extends to any class containing $\mathcal{H}_n$, including wider and deeper networks.

**Corollary 2** (Conditional computational hardness)**.** *Assume polynomial-modulus Learning with Rounding is not solvable in polynomial time. There is no polynomial-time algorithm that learns polynomial-size one-hidden-layer ReLU networks under the distributions $\mathcal{D}_n$ of Theorem 1 to a sufficiently small constant squared error, even with arbitrary possibly randomized, efficiently evaluable output hypotheses.*

The SQ theorem is unconditional. Corollary 2 addresses non-SQ algorithms, including algorithms that inspect individual examples, under the same cryptographic assumption used for one-hidden-layer networks on the uniform Boolean cube by Chen et al. (2022).

## 4 A CODED SMOOTH GRAY CURVE

Fix a sufficiently small constant $r > 0$, let $m = \lfloor rn \rfloor$, $N = n - m$, and write $K = 2^m$. Take a cyclic binary reflected Gray ordering $z_0, \dots, z_{K-1}$ of $\mathbb{F}_2^m$, so consecutive messages differ in one bit (Gray, 1953). We separate the construction into an abstract curve and a code map.

### 4.1 THE ABSTRACT SPARSE CURVE

Let $e_z \in \mathbb{R}^K$ be the basis vector indexed by $z \in \mathbb{F}_2^m$. Start from the closed polygon that visits $e_{z_0}, \dots, e_{z_{K-1}}$ in Gray order. Truncate a fixed fraction of every edge and replace every corner by the same smooth planar rounding inside

$$\operatorname{conv}\{e_{z_{i-1}}, e_{z_i}, e_{z_{i+1}}\}.$$

The rounding agrees with both adjacent straight segments to every derivative order. Appendix A gives an explicit tangent-angle construction.

**Lemma 3** (Sparse smooth Gray curve)**.** *There is a universal $r_0 > 0$ such that the resulting $C^\infty$ curve $\Gamma_m \subset \mathbb{R}^K$ is embedded and satisfies $\operatorname{reach}(\Gamma_m) \geq r_0$. Every point is a probability vector supported on at most three adjacent messages. Every tangent and derivative has total coordinate sum zero and support at most three. On each fixed natural parameter interval, speed is bounded above and below by universal positive constants.*

The dimension-free reach is a consequence of sparsity. Points on nonlocal pieces have disjoint supports and are separated by a constant. Every local configuration is a copy of one fixed finite-dimensional template. A curvature bound controls the tangent quotient in (2) near the diagonal, and compactness controls the remainder.

### 4.2 SYSTEMATIC SMALL-BIAS CODE

Fix a sufficiently small constant $\varepsilon_0 > 0$. An $\varepsilon_0$-biased multiset $\mathcal{R} \subset \mathbb{F}_2^m$ satisfies

$$\left| \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} (-1)^{\langle R, a \rangle} \right| \leq \varepsilon_0 \qquad \text{for every } a \neq 0.$$

Explicit constructions have $|\mathcal{R}| = O_{\varepsilon_0}(m)$ and are computable in polynomial time (Naor & Naor, 1993; Ta-Shma, 2017). Choose $r$ small enough that $N = n - m$ contains at least $1/\varepsilon_0$ full copies of such a multiset. Form $R_1, \dots, R_N$ by repeating full copies and truncating one final copy. The resulting multiset has bias at most $\varepsilon = 2\varepsilon_0$. Define the binary codeword

$$C(z) = (z_1, \dots, z_m, \mathbf{1}\{\langle R_1, z \rangle = 1\}, \dots, \mathbf{1}\{\langle R_N, z \rangle = 1\}) \in \{0,1\}^n. \tag{3}$$

Let $F : \mathbb{R}^K \to \mathbb{R}^n$ be the linear map $Fe_z = C(z)$. Choose $a = 1/4$ and map every probability vector $u \in \Gamma_m$ to

$$T(u) = a\mathbf{1} + (1 - 2a)Fu. \tag{4}$$

Because $u$ is a probability vector, $T(u) \in [a, 1-a]^n$. We set $M_n = T(\Gamma_m)$.

The next lemma is the coding core. It differs from a standard restricted isometry because the columns are indexed by every Boolean message and the relevant vectors are signed measures of total mass zero.

**Lemma 4** (Small-bias character isometry)**.** *Fix a constant $s$. For a sufficiently small constant bias and $N \geq C_s m$, the explicit code above simultaneously satisfies, for every $w \in \mathbb{R}^K$ with $\sum_z w_z = 0$ and $|\operatorname{supp}(w)| \leq s$,*

$$c_s N \|w\|_2^2 \leq \|Fw\|_2^2 \leq C_s N \|w\|_2^2. \tag{5}$$

For the parity block, expand the squared norm and use small bias on every nonzero difference $z + z'$. The diagonal contribution is exactly $N\|w\|_2^2/4$, while the absolute off-diagonal contribution is at most

$$\frac{\varepsilon N}{4} \sum_{z \neq z'} |w_z w_{z'}| \leq \frac{\varepsilon N(s-1)}{4} \|w\|_2^2.$$

Choosing $\varepsilon \leq 1/(2(s-1))$ gives constant lower and upper bounds without a net or union bound. The systematic coordinates add a nonnegative quadratic form and at most $ms\|w\|_2^2$. Choosing $r$ small enough makes $N \geq C_s m$. Appendix B gives the details, including the harmless truncation of the last copy.

**Proposition 5** (Critical reach)**.** *For the code in Lemma 4 with $s = 12$, the curve $M_n$ is a closed $C^\infty$ embedding and*

$$c\sqrt{n} \leq \operatorname{reach}(M_n) \leq \sqrt{n}.$$

*Proof sketch.* Let $p, q \in \Gamma_m$, let $w = q - p$, and let $v$ span the tangent at $p$. The vectors $w$ and $w - tv$ have total mass zero and constant support for every $t \in \mathbb{R}$. Lemma 4 gives

$$\|Fw\|_2^2 \geq cN\|w\|_2^2, \qquad \operatorname{dist}(Fw, \operatorname{span} Fv) \leq C\sqrt{N} \operatorname{dist}(w, \operatorname{span} v).$$

Substitution into (2) and Lemma 3 give $\operatorname{reach}(M_n) \geq c\sqrt{N} = \Omega(\sqrt{n})$. The same lower inequality makes $T$ injective on the curve and regular on every tangent.

For the upper bound, suppose the reach exceeded the diameter. Nearest-point projection would continuously retract the convex hull onto the curve. The induced identity on the curve's nonzero first homology would then factor through the zero first homology of its convex hull, a contradiction. Thus $\operatorname{reach}(M_n) \leq \operatorname{diam}(M_n) \leq \sqrt{n}$. □

## 5 Boolean information survives the embedding

The first $m$ coordinates of $C(z)$ are systematic. This is essential. Small-bias code coordinates create reach, while systematic coordinates expose a Boolean prefix to the target network.

The curve has one corner piece and one straight piece for every cyclic Gray index. Define $\mathcal{D}_n$ by choosing the piece type with probability $1/2$, choosing its index uniformly from $[K]$, and choosing its fixed natural parameter uniformly. Every image speed lies between $c\sqrt{n}$ and $C\sqrt{n}$ by Lemma 4. Consequently, $\mathcal{D}_n$ has a density between universal positive constants relative to normalized arclength. It has full support.

This law is also efficiently samplable. A Gray word and its neighbors can be computed from an $m$-bit index. Multiplication by the $m \times N$ generator matrix, followed by evaluation of one fixed local curve template, takes polynomial time. The explicit flat function in Appendix A and its one-dimensional integral can be evaluated to $b$ bits in time polynomial in $b$. Uniform real parameters can be generated to any requested finite precision.

Let $d = \lfloor m/2 \rfloor$ and $t = m - d$. In a reflected Gray ordering, the first $d$ bits are constant on blocks of $2^t$ consecutive messages. Call a local piece *good* if all two or three messages in its support have the same first $d$ bits. Only a constant number of pieces adjacent to each of the $2^d$ block boundaries are bad.

**Lemma 6** (Uniform decoded prefixes)**.** *For $X \sim \mathcal{D}_n$, the probability that $X$ lies on a bad piece is $O(2^{-t})$. Conditioned on the piece type and on being good, the normalized prefix*

$$Z(X) = \left( \frac{X_1 - a}{1 - 2a}, \ldots, \frac{X_d - a}{1 - 2a} \right)$$

*is exactly uniform on $\{0,1\}^d$.*

The exact uniformity follows because every high-prefix block has the same length and loses the same number of local indices at its boundaries. This is stronger than merely bounding the mass of each code neighborhood.

## 6 NOISE-FREE FULL-SQ HARDNESS

For $u \in [0, d]$, define the period-two triangular wave

$$g(u) = 1 - 2u + 4\sum_{k=1}^{d}(-1)^{k+1}(u-k)_+. \tag{6}$$

It satisfies $g(k) = (-1)^k$ at every integer $k \in \{0, \dots, d\}$ and takes values in $[-1, 1]$ on this interval. For a mask $S \in \mathbb{F}_2^d$, set

$$h_S(x) = g\left(\sum_{j=1}^{d} S_j \frac{x_j - a}{1-2a}\right). \tag{7}$$

On $M_n$, the argument of $g$ lies in $[0, d]$. Equation (6) is a one-hidden-layer ReLU representation of width at most $d+1$. The linear term is $-2\,\mathrm{ReLU}(u)$ because $u \geq 0$ on $M_n$. Its hidden weights have norm $O(\sqrt{d})$, its biases are $O(d)$, and its output weights are bounded by four.

On a good piece, $Z(X) \in \{0,1\}^d$ is constant and

$$h_S(X) = (-1)^{\langle S, Z(X)\rangle}. \tag{8}$$

Thus the target is continuous everywhere, but it acts as an exact parity on almost all local pieces.

**Definition 7** (Almost pairwise sign independence)**.** A finite family $\mathcal{H}$ of real-valued functions is $(1-\eta)$-pairwise sign independent under $D$ if, with probability at least $1-\eta$ over independent $X, X' \sim D$, a uniform $h \in \mathcal{H}$ makes $(h(X), h(X'))$ uniform on $\{-1,1\}^2$. The functions may take other values on the exceptional input pairs.

**Lemma 8** (Parity independence)**.** *The family $\mathcal{H}_n = \{h_S : S \in \mathbb{F}_2^d\}$ is $(1-\eta_n)$-pairwise sign independent under $\mathcal{D}_n$, for $\eta_n \leq \exp(-\rho n)$.*

*Proof.* The probability that either draw is bad is $O(2^{-t})$. Conditioned on two good draws, their decoded prefixes are independent and uniform by Lemma 6. Two vectors in $\mathbb{F}_2^d$ fail to be linearly independent only if one is zero or they are equal, an event of probability at most $3 \cdot 2^{-d}$. Otherwise the map

$$S \longmapsto (\langle S, Z(X)\rangle, \langle S, Z(X')\rangle)$$

has rank two and is uniform on $\mathbb{F}_2^2$. Equation (8) finishes the proof. Since $d, t = \Omega(n)$, the failure probability is exponentially small. □

The full-SQ consequence is immediate but worth stating separately. The result is stronger than an orthogonality or correlational-SQ argument because the query may depend arbitrarily on both $x$ and $y$.

**Lemma 9** (Pairwise-independence SQ bound, adapted from Chen et al., 2022)**.** *Let $\mathcal{H}$ be $(1-\eta)$-pairwise sign independent under $D$. Distinguishing an unknown deterministic labeled law $(X, h(X))$, $h \in \mathcal{H}$, from the random-function mixture requires at least $\tau^2/(2\eta)$ bounded SQ queries of tolerance $\tau$ for a deterministic distinguisher. A randomized distinguisher with constant success probability requires $\Omega(\tau^2/\eta)$ queries.*

The random-function mixture first draws $X \sim D$ and then draws a fresh uniform $h \in \mathcal{H}$ to set the label $h(X)$. For completeness, if $\phi[h] = \mathbb{E}_X \phi(X, h(X))$, the same calculation as in Lemma F.3 of Chen et al. (2022) gives

$$\mathrm{Var}_{h \sim \mathrm{Unif}(\mathcal{H})}\, \phi[h] \leq 2\eta.$$

Indeed, expand the variance using two inputs and compare a common uniform $h$ with independent uniform $h, h'$. The two joint label laws agree on every nonexceptional input pair, even though labels on exceptional pairs need not be signs. The integrand is bounded by two elsewhere. An oracle can therefore answer every query by the mixture expectation. Chebyshev's inequality shows that one query excludes at most a $2\eta/\tau^2$ fraction of targets. The deterministic claim follows by a union bound over the resulting query path. Averaging over the target and the internal coins gives the randomized claim with a universal constant loss.

*Proof of Theorem 1.* Proposition 5 gives the geometric claim. The speed bounds above give all distributional claims. Equation (7) gives the stated architecture.

It remains to reduce learning to distinguishing. Clip the learner's output to $[-1, 1]$. Under a deterministic target, its expected squared loss is at most $1/16$. Under the random-function mixture, on every good input with nonzero decoded prefix the fresh label is a uniform sign independent of the input. Every fixed predictor then has conditional squared loss at least one. Bad and zero-prefix inputs have total mass $\exp(-\Omega(n))$. One final bounded SQ query for one quarter of the squared loss separates the cases with constant gap whenever $\tau \leq \tau_0$. Lemmas 8 and 9 imply the claimed $c\tau^2 \exp(\rho n)$ lower bound after absorbing this final query. □

## 7 Conditional hardness beyond SQ

Full-SQ hardness does not rule out algorithms that exploit individual examples. The systematic construction also transports a standard cryptographic lower bound.

Chen et al. (2022) show that if polynomial-size one-hidden-layer ReLU networks can be learned efficiently under the uniform Boolean distribution to constant squared error, then polynomial-modulus Learning with Rounding can be solved efficiently. Their result allows arbitrary efficiently evaluable output hypotheses.

Conditioned on a good piece, our decoded prefix is exactly uniform by Lemma 6. Given a Boolean prefix $z$, one can invert the reflected Gray order of $z$, select a uniformly random good local index inside its block, and evaluate the coded curve in polynomial time. This lifts a uniform Boolean example to the conditional good-piece law. Conversely, the full law and the good-piece law differ by $\exp(-\Omega(n))$ in total variation. A polynomial number of examples can be coupled with probability $1 - \exp(-\Omega(n))$.

Finally, replacing a Boolean input $z_j$ by $(x_j - a)/(1 - 2a)$ preserves one-hidden-layer architecture and polynomial weights. On bad pieces the resulting network is still polynomially bounded, so their exponentially small mass changes squared risk by $\exp(-\Omega(n))$. An efficient learner under $\mathcal{D}_n$ would therefore give the Boolean-cube learner excluded by the LWR assumption. This proves Corollary 2. Appendix F records the coupling formally.

## 8 Discussion and limitations

The $\sqrt{n}$ boundary is sharp in the following geometric sense. Reach $\omega(\sqrt{n})$ forces polynomial volume for fixed intrinsic dimension in the unit cube (Kiani et al., 2024). Theorem 1 gives exponential SQ hardness at $c\sqrt{n}$ for a fixed positive constant $c$. Thus no asymptotic improvement from $o(\sqrt{n})$ to $O(\sqrt{n})$ is possible without an additional assumption.

The coding viewpoint may be useful beyond this boundary question. Reach depends on chord residuals against tangent spaces, not on coordinatewise separation. A restricted isometry for sparse coefficient measures is therefore enough to preserve reach of any curve with a sparse local representation. The small-bias character map is especially compatible with Boolean labels because it leaves systematic message coordinates untouched.

Several limitations remain. The construction is worst case and has exponential length. It should not be interpreted as a model of a typical real-data manifold. The unconditional result excludes SQ algorithms, not every polynomial-time learner. The latter conclusion uses LWR. The hard manifold is one-dimensional. Finally, the theorem resolves reach up to constant factors at the critical order. Determining the largest attainable constant $c$ compatible with exponential volume and hardness is a separate coding-geometric problem.

### Reproducibility statement

All assumptions and complete proofs appear in the main text and Appendices A–F. The paper makes no empirical claims. The construction uses a standard explicit small-bias set and otherwise consists of the formulas given in Sections 4 and 5.

## REFERENCES


Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359–1399, 2019. doi: 10.1214/19-EJS1551.

Sitan Chen, Aravind Gollakota, Adam R. Klivans, and Raghu Meka. Hardness of noise-free learning for two-hidden-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10709–10724, 2022. doi: 10.52202/068431-0778.

Amit Daniely and Gal Vardi. From local pseudorandom generators to hardness of learning. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 1358–1394. PMLR, 2021.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer ReLU networks. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1514–1539. PMLR, 2020.

Armin Eftekhari and Michael B. Wakin. What happens to a manifold under a bi-lipschitz map? *Discrete & Computational Geometry*, 57(3):641–673, 2017. doi: 10.1007/s00454-016-9847-6.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959. doi: 10.1090/S0002-9947-1959-0110078-1.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852.

Frank Gray. Pulse code communication. *United States Patent 2,632,058*, 1953.

Bobak T. Kiani, Jason Wang, and Melanie Weber. Hardness of learning neural networks under the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 5661–5696, 2024. doi: 10.52202/079017-0184.

Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM Journal on Computing*, 22(4):838–856, 1993. doi: 10.1137/0222053.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39:419–441, 2008. doi: 10.1007/s00454-008-9053-2.

Amnon Ta-Shma. Explicit, almost optimal, epsilon-balanced codes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 238–251. ACM, 2017. doi: 10.1145/3055399.3055408.

Santosh Vempala and John Wilmes. Gradient descent for one-hidden-layer neural networks: Polynomial convergence and SQ lower bounds. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 3115–3117. PMLR, 2019.


## A THE ABSTRACT SPARSE CURVE

We prove Lemma 3. The only purpose of this appendix is to make the smoothness and dimension-free constants explicit. The precise numerical constants are immaterial.

Fix $0 < \alpha < 1/10$. Truncate every polygon edge at Euclidean distance $\alpha$ from each endpoint. At a vertex $e_i$, let $u = (e_i - e_{i-1})/\sqrt{2}$ and $v = (e_{i+1} - e_i)/\sqrt{2}$ be the incoming and outgoing unit tangents. They satisfy $\langle u, v\rangle = -1/2$, so their turning angle is $\theta = 2\pi/3$.

For a sufficiently small fixed rational $\delta > 0$, let

$$b_\delta(s) = Z_\delta^{-1} \exp\left(-\frac{\delta}{s(1-s)}\right), \qquad 0 < s < 1,$$

and set it to zero at the endpoints, where $Z_\delta$ normalizes its integral to one. This is a positive symmetric $C^\infty$ function that is flat at zero and one. As $\delta \downarrow 0$, it converges to the constant function one on compact subintervals of $(0, 1)$. Write $b = b_\delta$. For a length $L > 0$, define

$$\beta(s) = \theta \int_0^{s/L} b(r)\, dr, \qquad 0 \le s \le L.$$

Let $U(\varphi)$ be the unit vector obtained by rotating $u$ through angle $\varphi$ in the plane spanned by the three local basis vectors, and put

$$\gamma(s) = p + \int_0^s U(\beta(q))\, dq.$$

Symmetry gives $U(\beta(s)) + U(\beta(L-s))$ parallel to $u + v$. Hence $\gamma(L) - \gamma(0) = c_b L(u + v)$ for a constant $c_b > 0$. Choosing $L = \alpha/c_b$ makes the endpoints equal to the two truncated polygon points. For $b \equiv 1$, the formula is the circular arc tangent to both edges and lying in the local triangle. Positive smooth flat approximations preserve the two inward-pointing endpoint jets and converge to this circular arc away from the endpoints. Choosing one sufficiently close approximation therefore keeps the open arc in the triangle. This template and all its constants are fixed once and for all.

The arc has unit speed. Its curvature is $\beta'(s) = \theta b(s/L)/L$, which is uniformly bounded because $L$ is fixed. Every derivative of curvature vanishes at the endpoints. It therefore joins the straight segments to all orders. The resulting cyclic curve is $C^\infty$, and every point is a convex combination of at most three adjacent basis vectors.

To prove a uniform reach bound, classify pairs of local pieces by cyclic index distance. If their three-vertex supports are disjoint, each point is a probability vector with norm at least $1/\sqrt{3}$, so their distance is at least $\sqrt{2/3}$. If supports overlap, there are only finitely many relative configurations, each congruent to a fixed compact template. Embeddedness of the rounded orthogonal polygon gives positive separation away from the diagonal.

Near the diagonal, let $\gamma$ be any local unit-speed parametrization and let $\kappa_0$ bound curvature. Taylor's theorem gives, for small $|h|$,

$$\operatorname{dist}(\gamma(s+h) - \gamma(s), \operatorname{span} \gamma'(s)) \le \frac{\kappa_0}{2} h^2, \qquad \|\gamma(s+h) - \gamma(s)\| \ge \frac{|h|}{2}.$$

The quotient in (2) is therefore bounded below by a positive constant near the diagonal. Compactness handles every remaining local pair. The minimum of these finitely many constants and the disjoint-support bound is $r_0 > 0$.

## B SMALL-BIAS CHARACTER ISOMETRY

We prove Lemma 4. Let $\mathcal{R}_0 \subset \mathbb{F}_2^m$ be an explicit $\varepsilon_0$-biased multiset of size $N_0 \le C_{\varepsilon_0} m$ (Naor & Naor, 1993; Ta-Shma, 2017). Let $q = \lfloor N/N_0 \rfloor$. By taking $r$ sufficiently small, $q \ge 1/\varepsilon_0$. Repeat $\mathcal{R}_0$ exactly $q$ times and append any $N - qN_0 < N_0$ elements of one further copy. For every $a \ne 0$, the resulting empirical character average has absolute value at most

$$\frac{qN_0\varepsilon_0 + (N - qN_0)}{N} \le \varepsilon_0 + \frac{1}{q} \le 2\varepsilon_0.$$

Write $\varepsilon = 2\varepsilon_0$ for this final bias. Let $A \in \{0,1\}^{N \times K}$ be the parity part of $F$, so

$$A_{\ell,z} = \mathbf{1}\{\langle R_\ell, z\rangle = 1\}.$$

For $\chi_R(z) = (-1)^{\langle R, z\rangle}$,

$$\mathbf{1}\{\langle R, z\rangle = 1\} = \frac{1 - \chi_R(z)}{2}.$$

If $\sum_z w_z = 0$, then

$$\sum_z w_z \mathbf{1}\{\langle R, z\rangle = 1\} = -\frac{1}{2}\sum_z w_z \chi_R(z).$$

For a zero-mass vector $w$, expansion over the empirical law of the $R_\ell$ gives

$$\begin{aligned}\frac{1}{N}\|Aw\|_2^2 &= \frac{1}{4N}\sum_{\ell=1}^{N}\left(\sum_z w_z \chi_{R_\ell}(z)\right)^2 \\ &= \frac{1}{4}\|w\|_2^2 + \frac{1}{4}\sum_{z\neq z'} w_z w_{z'}\left(\frac{1}{N}\sum_{\ell=1}^{N}\chi_{R_\ell}(z+z')\right). \end{aligned} \tag{9}$$

Every $z + z'$ in the second sum is nonzero. If $w$ has support at most $s$, small bias and Cauchy–Schwarz give

$$\left|\frac{1}{N}\|Aw\|_2^2 - \frac{1}{4}\|w\|_2^2\right| \le \frac{\varepsilon}{4}\sum_{z\neq z'}|w_z w_{z'}| \le \frac{\varepsilon(s-1)}{4}\|w\|_2^2.$$

Choose $\varepsilon \le 1/(2(s-1))$. Then, simultaneously for every such $w$,

$$\frac{1}{8}N\|w\|_2^2 \le \|Aw\|_2^2 \le \frac{3}{8}N\|w\|_2^2. \tag{10}$$

It remains to add the systematic part. Let $B \in \{0,1\}^{m\times K}$ have column $z$ at index $z$. For every $s$-sparse $w$,

$$|(Bw)_j| \le \|w\|_1 \le \sqrt{s}\|w\|_2, \qquad \|Bw\|_2^2 \le ms\|w\|_2^2.$$

Since $F = (B, A)$, the lower bound in (10) remains valid and the systematic part changes only the upper constant when $m = O(N)$. This proves (5).

## C Reach transfer in detail

Let $q = 1 - 2a = 1/2$. The derivative of $T$ is $qF$. Fix distinct $p, u \in \Gamma_m$, let $w = u - p$, and let $v \neq 0$ span $T_p\Gamma_m$. By Lemma 3, $w$ has support at most six and $v$ has support at most three. Both have coordinate sum zero, and every $w - \lambda v$ has support at most six. Lemma 4 yields constants $c_0, C_0 > 0$ such that

$$\|qFw\|_2^2 \ge c_0 N\|w\|_2^2$$

and

$$\begin{aligned}\operatorname{dist}(qFw, \operatorname{span}(qFv)) &= \inf_{\lambda\in\mathbb{R}}\|qF(w-\lambda v)\|_2 \\ &\le C_0\sqrt{N}\inf_{\lambda\in\mathbb{R}}\|w-\lambda v\|_2 \\ &= C_0\sqrt{N}\operatorname{dist}(w, \operatorname{span} v).\end{aligned}$$

Consequently,

$$\frac{\|T(u)-T(p)\|_2^2}{2\operatorname{dist}(T(u)-T(p), T_{T(p)}M_n)} \ge \frac{c_0}{C_0}\sqrt{N}\frac{\|u-p\|_2^2}{2\operatorname{dist}(u-p, T_p\Gamma_m)}.$$

Taking the infimum and using (2) gives

$$\operatorname{reach}(M_n) \ge \frac{c_0}{C_0}\sqrt{N}\operatorname{reach}(\Gamma_m) \ge c\sqrt{n}.$$

The same restricted lower bound shows that distinct curve points have distinct images and nonzero tangents remain nonzero. Hence the image is a smooth embedding.

For completeness, suppose a compact set $M$ has reach greater than its diameter. Every point of $\operatorname{conv}(M)$ lies within $\operatorname{diam}(M)$ of $M$, so nearest-point projection is a continuous retraction from $\operatorname{conv}(M)$ to $M$. If $M$ is a closed embedded curve, the induced identity on $H_1(M)$ would then factor through $H_1(\operatorname{conv}(M)) = 0$, a contradiction. Therefore $\operatorname{reach}(M) \le \operatorname{diam}(M)$. Applying this to $M_n \subset [0,1]^n$ gives the upper bound $\sqrt{n}$.

## D GRAY COUNTING AND THE INPUT DENSITY

Write the reflected Gray word at index $i \in \{0, \dots, 2^m - 1\}$ as

$$G(i) = i \operatorname{xor} (i \gg 1).$$

For $d + t = m$, the first $d$ bits of $G(i)$ are constant for $2^t$ consecutive indices. The sequence of these block values is itself a reflected Gray order on $d$ bits. This follows directly from the recursive reflection construction.

There are $2^d$ cyclic blocks. A straight piece is bad only when it crosses a block boundary, so there are exactly $2^d$ bad straight indices. A corner piece can be bad only when its three-message window meets a boundary, so there are at most $2^{d+1}$ bad corner indices. With a uniform index and fixed type probabilities,

$$\Pr(\text{bad}) \le C\frac{2^d}{2^m} = C2^{-t}.$$

Every block loses the same number of indices for each fixed piece type. Conditional on that type and goodness, its prefix is therefore uniform on $\mathbb{F}_2^d$. Mixing the two types preserves uniformity and proves Lemma 6.

Let $s_j(u)$ denote the abstract parametrization of type $j \in \{\text{straight}, \text{corner}\}$ and index $i$. Its speed lies in a universal interval $[c_0, C_0]$. Since its tangent has total mass zero and support at most three, Lemma 4 gives

$$c\sqrt{n} \le \|(T \circ s_j)'(u)\|_2 \le C\sqrt{n}.$$

There are $2K$ parameter intervals, each selected with probability $1/(2K)$. The total image length is between $cK\sqrt{n}$ and $CK\sqrt{n}$. Dividing parameter density by speed and comparing with inverse total length shows that $d\mathcal{D}_n/d\,\mathrm{UnifArc}$ lies between two universal positive constants.

## E PAIRWISE INDEPENDENCE AND LEARNING REDUCTION

Let $E$ be the event that both independent draws lie on good pieces and their decoded prefixes are nonzero and distinct. Lemma 6 gives

$$\Pr(E^c) \le C2^{-t} + 2 \cdot 2^{-d} + 2^{-d} \le \exp(-\rho n).$$

Over $\mathbb{F}_2$, two nonzero vectors are linearly dependent exactly when they are equal. Thus, on $E$, the linear map from $S \in \mathbb{F}_2^d$ to its two parity bits has rank two. Uniform $S$ produces two independent uniform signs. This proves Lemma 8.

We also spell out the final validation query. Let $\widehat{h}$ be the learner's output and clip it pointwise to $[-1, 1]$. Query

$$\phi(x, y) = \frac{1}{4}(\widehat{h}(x) - y)^2 \in [0, 1].$$

For a learned deterministic target, $\mathbb{E}\phi \le 1/64$. Under the random-function mixture, condition on a good input with a nonzero decoded prefix. Its label is an independent uniform sign, and

$$\mathbb{E}_y(\widehat{h}(x) - y)^2 = 1 + \widehat{h}(x)^2 \ge 1.$$

Therefore the mixture expectation is at least $1/4 - \exp(-\rho n)$. Any fixed $\tau_0 < 1/16$ separates the two cases for large $n$.

## F CONDITIONAL LWR REDUCTION

We use Corollary 4.3 of Chen et al. (2022): an efficient learner for polynomial-size one-hidden-layer ReLU networks under the uniform law on $\{0, 1\}^d$, to constant squared error, yields an efficient algorithm for polynomial-modulus LWR.

Let $\mathcal{D}_n^{\text{good}}$ be $\mathcal{D}_n$ conditioned on a good piece. Its total variation distance from $\mathcal{D}_n$ is $\delta_n = \exp(-\Omega(n))$. A Boolean input $z \in \mathbb{F}_2^d$ can be lifted to $\mathcal{D}_n^{\text{good}}$ as follows. Invert the reflected Gray map on the high-prefix block $z$, choose the piece type according to its conditional probability,

choose uniformly among its good indices in that block, and choose the natural parameter uniformly. By the equal block counts proved in Appendix D, this generates exactly the conditional law. Gray inversion, code evaluation, and local curve evaluation are polynomial time.

Take a hard Boolean-cube network $f$ from the LWR reduction and replace each input $z_j$ by

$$y_j(x) = \frac{x_j - a}{1 - 2a}.$$

This affine substitution does not change depth or width and increases weights and biases only polynomially. Call the resulting network $\widetilde{f}$. On every good piece, $\widetilde{f}(x) = f(Z(x))$.

Suppose a polynomial-time learner $L$ learned these networks under $\mathcal{D}_n$. A transcript of $Q(n)$ examples from $\mathcal{D}_n$ can be coupled to one from $\mathcal{D}_n^{\mathrm{good}}$ with failure probability at most $Q(n)\delta_n = \exp(-\Omega(n))$. Run $L$ on lifted uniform Boolean examples. Given a Boolean test input $z$, draw a fresh good piece with decoded prefix $z$ and evaluate the hypothesis returned by $L$ there. The expected Boolean squared risk of this randomized predictor is exactly the hypothesis risk under $\mathcal{D}_n^{\mathrm{good}}$, which is at most its risk under $\mathcal{D}_n$ divided by $1 - \delta_n$. Thus this procedure learns the Boolean LWR class to the constant accuracy excluded by Corollary 4.3 of Chen et al. (2022). This proves Corollary 2.