PROPER LEARNING OF SIGNED NARROW RELU NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cancellation among signed neurons complicates uniform learning guarantees for narrow ReLU networks. We address this challenge with a proper learner for normalized, bias-free networks under Gaussian inputs, without separation assumptions. Specifically, for width k and squared L_2 error ε, it uses $g(k)\varepsilon^{-O(k^2)}\widetilde{O}(d)$ noiseless examples and $g(k)\varepsilon^{-O(k^2)}\widetilde{O}(d^2)$ real-arithmetic operations. Thus, at any fixed width, sample complexity is nearly linear in dimension and running time is nearly quadratic, while both are polynomial in inverse accuracy. To prove this, we combine first-moment control with a scaled nonlinear label-distance statistic to lower-bound a second-moment signal using residual variance. A norm-preserving rotation then ensures approximation by at most k units in the original normalized class, yielding proper output.

1 INTRODUCTION

Learning useful features from high-dimensional data is a central problem in machine learning. In nonlinear regression, the response may depend on only a few unknown linear combinations of the input coordinates. Finding these combinations yields a compact representation for prediction. This is the motivation behind supervised dimension reduction (Li, 1991; Dalalyan et al., 2008) and learning models with few relevant directions (Chen & Meka, 2020).

A one-hidden-layer ReLU network with k neurons is a concrete model of this structure: it depends on at most k directions. We consider learning such networks from independent Gaussian examples in d dimensions. The output coefficients may have either sign. Neurons can then nearly cancel, so accurate prediction need not reveal their individual parameters (Chen et al., 2023; Chen & Narayanan, 2024). This makes function approximation the relevant learning objective.

Previous work provides several routes to learning these networks. Tensor and optimization methods recover parameters under suitable conditioning assumptions (Zhong et al., 2017; Ge et al., 2018). For function learning, nonnegative output coefficients permit fixed dimension exponents and polynomial inverse-accuracy dependence at each fixed width (Diakonikolas & Kane, 2020). With arbitrary signs, two approaches offer complementary guarantees. Filtered PCA keeps the exponent of d fixed, but its cost is exponential in inverse accuracy (Chen et al., 2022). Higher-order moment methods achieve polynomial inverse-accuracy dependence for each fixed k (Chen et al., 2023; Chen & Narayanan, 2024; Diakonikolas & Kane, 2024). Their dimension exponent, however, grows with k.

We address the question raised by Chen et al. (2023, Section 5): can these two guarantees be achieved together? For normalized, bias-free networks, we combine a fixed dimension exponent with polynomial inverse-accuracy dependence at each fixed width. The sample count is nearly linear in d, and the operation count is nearly quadratic. The learned network has at most k units and satisfies the same norm constraints as the target.

1.1 MAIN RESULT

For $k, d \geq 1$, let $\mathcal{F}_k$ consist of the functions

$$F(x) = \sum_{i=1}^{k} a_i \sigma(\langle u_i, x \rangle), \qquad \sigma(t) = \max(t, 0), \quad \|u_i\|_2 \leq 1, \quad \sum_i |a_i| \leq 1. \tag{1}$$

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.

The coefficients a_i may have either sign, and zero coefficients allow fewer than k units. We observe independent examples (X, Y) with $X \sim \mathcal{N}(0, I_d)$ and $Y = F(X)$. Throughout, ε denotes *squared* Gaussian L_2 error.

Theorem 1 (Proper learning of signed narrow ReLU networks)**.** *There is a randomized real-arithmetic algorithm that, for every $k, d \geq 1$, $F \in \mathcal{F}_k$, and $\varepsilon, \delta \in (0, 1)$, returns $\widehat{F} \in \mathcal{F}_k$ with $\mathbb{E}(F(X) - \widehat{F}(X))^2 \leq \varepsilon$, with probability at least $1 - \delta$, using independent examples $(X, F(X))$ with $X \sim \mathcal{N}(0, I_d)$. For an absolute constant C and a computable function g, its labeled-sample and operation counts are respectively at most*

$$N \leq g(k)\varepsilon^{-Ck^2} d \ \mathrm{polylog}\left(\frac{2d}{\varepsilon\delta}\right), \tag{2}$$

$$T \leq g(k)\varepsilon^{-Ck^2} d^2 \ \mathrm{polylog}\left(\frac{2d}{\varepsilon\delta}\right). \tag{3}$$

The guarantee is uniform over $\mathcal{F}_k$. The output is *proper*, with at most k units and the same norm constraints. For fixed k, both bounds are polynomial in $1/\varepsilon$, while the dimension exponents are independent of k; the width-only factor $g(k)$ remains explicit. The setting is noiseless and bias-free, and the guarantee concerns function learning in real arithmetic (Section 2). Section 6 sketches the proof; Appendix E.1 gives the full argument.

1.2 Proof Overview

We first learn a subspace and then fit a network within it. Progress is measured by the conditional variance of the label given the current projection, namely the best squared prediction error attainable from that projection. The proof has three components.

A second moment that survives cancellation. For a current subspace W, the activation patterns of the projected neurons determine at most 2^k possible affine intercepts. The distance from the label to this finite set is a nonlinear statistic whose expectation is at least the remaining conditional variance divided by a width factor. After controlling the first moment, we scale the missing input component and find a range admitting an affine approximation. The expected distance must then change across a finite list of scales. Gaussian density differentiation converts this change into a second moment of size at least $v/[g(k)\log(2/v)]$ when the residual variance exceeds v (Proposition 4). Either sign of the signal is useful.

Estimating directions from ordinary examples. Homogeneity expresses the scaled moments as weighted expectations of ordinary labeled examples; the change of measure acts only on W. A finite coefficient net covers the unknown intercepts, and one fresh batch estimates all candidates within each round (Section 4.1). A subspace-error bound controls every selected direction, including those selected after the residual is already small. This permits stopping when no detectable moment remains (Sections 5–6).

Returning a network in the original class. The conditional mean need not be a narrow ReLU network. Instead, we show that rotating the target weights into an enlarged output subspace preserves their norms and incurs squared error at most twice the conditional variance (Lemma 5). This existence argument justifies a finite parameter net; a final independent validation batch selects a proper hypothesis.

1.3 Related Work

Function learning. For nonnegative output coefficients, Diakonikolas et al. (2020); Diakonikolas & Kane (2020) give fixed dimension exponents and polynomial inverse-accuracy dependence at each fixed width. For arbitrary signs, filtered PCA gives proper learning with a fixed polynomial dependence on d (Chen et al., 2022). Higher-order moment methods successively improve the dependence on width (Chen et al., 2023; Chen & Narayanan, 2024; Diakonikolas & Kane, 2024); Table 1 summarizes their running times. Chen & Narayanan (2024) return a ReLU network, whereas the Schur-polynomial learner of Diakonikolas & Kane (2024) is improper. Our result combines

Table 1: Gaussian function learning from independent examples. Bounds are expressed in squared L_2 error ε; fixed norm bounds and confidence factors are suppressed. The homogeneous Lipschitz row assumes unit L_2 norm and a fixed Lipschitz bound, unlike our upper coefficient-norm bound. Our output stays in $\mathcal{F}_k$; the Schur-polynomial output is improper.

Work	Target class	Running time
Diakonikolas & Kane (2020)	Nonnegative ReLU	$\mathrm{poly}(d/\varepsilon) + (k/\varepsilon)^{O(\log^2 k)}$
Chen et al. (2022)	Signed ReLU	$\widetilde{O}(d^2)\exp(\mathrm{poly}(k, 1/\varepsilon))$
Chen et al. (2023)	Signed ReLU	$(d/\varepsilon)^{k^{O(\log^2 k)}}$
Chen & Narayanan (2024)	Signed ReLU	$(dk/\varepsilon)^{O(k^2)}$
Diakonikolas & Kane (2024)	Signed ReLU	$(dk/\varepsilon)^{O(k)}$
Diakonikolas et al. (2025)	Homogeneous Lipschitz	$\mathrm{poly}(d)2^{\mathrm{poly}(k/\varepsilon)}$
This work (Theorem 1)	Signed ReLU	$g(k)\varepsilon^{-O(k^2)}\widetilde{O}(d^2)$

fixed dimension exponents, polynomial inverse-accuracy dependence at fixed width, and output in the same normalized class.

Subspace estimation and nonlinear label statistics. Inverse regression and second-moment methods are classical (Li, 1991; 1992; Li & Wang, 2007); filtered PCA uses nonlinear label weights for low-dimensional polynomials and ReLU networks (Chen & Meka, 2020; Chen et al., 2022). Damian et al. (2024; 2025) study dimension dependence through nonlinear label information; for continuous piecewise-linear targets, the latter give subspace recovery from $C(P, K)d(\tau^{-1} + \tau^{-2})$ samples at subspace error τ, with target- and kernel-dependent $C(P, K)$. Our distinction is a bound uniform over the normalized shallow class and a proper function output. Diakonikolas et al. (2025) give a general multi-index framework whose realizable bound is polynomial in inverse accuracy and inverse distinguishing-moment strength. Its unit-L_2-normalized homogeneous Lipschitz application, however, retains exponential inverse-accuracy dependence. The quantitative issue here is therefore a uniform signal bound for the normalized ReLU class, supplied by Proposition 4, together with proper output.

Other assumptions and access models. Parameter-recovery results require conditioning or rank assumptions (Zhong et al., 2017; Ge et al., 2018; Bakshi et al., 2019); Awasthi et al. (2021) allow biases under nondegeneracy assumptions. Second-order Stein estimation beyond Gaussian inputs assumes a nondegenerate expected Hessian (Yang et al., 2017). Chosen-input labels permit fully polynomial ReLU learning (Chen et al., 2021), whereas we use only independent Gaussian examples. Correlational statistical query (CSQ) lower bounds constrain statistics linear in the label (Goel et al., 2020; Diakonikolas et al., 2020), not the nonlinear distance statistic used here or filtered PCA (Chen et al., 2022). Unlike the moment learners' independent zero-mean noise extensions (Chen et al., 2023; Diakonikolas & Kane, 2024), our guarantee is noiseless.

2 Preliminaries

Notation and normalization. We use the model in (1) and write $[k] = \{1, \dots, k\}$, $V = \mathrm{span}\{u_i : i \in [k]\}$, and $r = \dim V \le \min(k, d)$. The representation may be degenerate, including $F \equiv 0$. For a subspace W, P_W is the orthogonal projector and I_W the identity on W. Unsubscripted vector norms are Euclidean; $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ denote matrix operator and Frobenius norms, and $\|f\|_p = (\mathbb{E}|f(X)|^p)^{1/p}$ for $X \sim \mathcal{N}(0, I_d)$. Our normalization bounds coefficients from above, rather than fixing $\|F\|_2$. A known bound $\sum_i |a_i| \le R$ is handled by dividing labels by R and using squared accuracy ε/R^2; if $R^2 \le \varepsilon$, the zero network suffices.

Residual prediction error. For any W, define

$$v_W(F) = \mathbb{E}\,\mathrm{Var}(F(X) \mid P_W X) = \inf_h \|F - h \circ P_W\|_2^2 . \tag{4}$$

The infimum is over square-integrable predictors and is attained by the conditional mean. Thus $v_W(F)$ measures the information missing from the projection, decreases when W is enlarged, and vanishes if $V \subseteq W$.

To see what is lost, write $X = w + z$, where $w \sim \mathcal{N}(0, I_W)$ and $z \sim \mathcal{N}(0, I_{W^\perp})$ are independent. For a copy z' of z independent of (w, z), the conditional-variance identity gives

$$v_W(F) = \tfrac{1}{2}\mathbb{E}\big(F(w+z) - F(w+z')\big)^2.$$

This compares labels at a fixed observed projection. The conditional mean need not lie in $\mathcal{F}_k$, so small residual variance does not by itself provide a proper hypothesis. Lemma 5 supplies that additional step.

Gaussian identities. The norm constraint makes F 1-Lipschitz. For every centered Gaussian Z with $\mathrm{Cov}(Z) \preceq I$, Minkowski's inequality applied neuron by neuron gives $\|F(Z)\|_2 \le 1/\sqrt{2}$ and $\|F(Z)\|_4 \le (3/2)^{1/4}$. Almost everywhere,

$$\nabla F(x) = \sum_i a_i u_i \mathbf{1}\{\langle u_i, x\rangle > 0\} \in V, \qquad \|\nabla F(x)\| \le 1. \tag{5}$$

Applying Gaussian Poincaré (Ledoux, 2001) conditionally on $P_W X$ gives

$$v_W(F) \le \mathbb{E}\,\|P_{W^\perp}\nabla F(X)\|^2 \le \left(\sum_i |a_i|\,\|P_{W^\perp}u_i\|\right)^2.$$

The last inequality follows from (5) and the triangle inequality. Thus short missing weight components imply small prediction error. This is only an upper bound: cancellation can make the error much smaller. A lower bound on $v_W(F)$ therefore forces some missing weight components to be large, as used in the scale-selection argument.

The decomposition $\sigma(t) = (t + |t|)/2$ (Chen et al., 2023, Lemma 2) separates F into a linear term and an even function. Gaussian symmetry therefore gives

$$\mu := \mathbb{E}[YX] = \tfrac{1}{2}\sum_i a_i u_i \in V. \tag{6}$$

The first moment alone can miss a nonconstant target. For example, $F(x) = (\sigma(x_1)+\sigma(-x_1))/2 = |x_1|/2$ belongs to $\mathcal{F}_2$ and has $\mu = 0$, but $v_{\{0\}}(F) = (1 - 2/\pi)/4 > 0$. Estimating μ controls the linear component before the second-moment search. Positive homogeneity, $F(tx) = tF(x)$ for $t \ge 0$, will allow that search to use scaled distributions without querying new inputs.

For $Z \sim \mathcal{N}(0, I_S)$ on a subspace S, the centered quadratic $ZZ^\top - I_S$ has mean zero. Any integrable scalar weight independent of Z therefore contributes zero to its weighted expectation. Even with a nonnegative weight, the resulting matrix need not be positive semidefinite. Its operator norm detects eigenvalues of either sign. Dependence can also cancel; our construction addresses this with a nonlinear label statistic and a search over scales.

Complexity conventions. The factor $g(k)$ denotes a positive computable function of width alone and may change between bounds.

3 Residual Variance and Scaling

We now construct the label statistic used to find a missing direction. The first step is to relate its expected value to conditional variance; the second is to find scales at which that value changes.

3.1 Distance to the Projected Coefficients

For a subspace W, split $u_i = p_i + q_i$, with $p_i = P_W u_i$ and $q_i = P_{W^\perp} u_i$, and put $c_i = a_i p_i$. The possible affine intercepts form the finite set

$$S_c(w) = \left\{\sum_{i\in J}\langle c_i, w\rangle : J \subseteq [k]\right\}, \qquad c_i = a_i p_i. \tag{7}$$

Put $U = P_{W^\perp} V$ and take independent $w \sim \mathcal{N}(0, I_W)$ and $z \sim \mathcal{N}(0, I_U)$. The discarded coordinates affect neither $P_W X$ nor $F(X)$, and $\dim U \le k$. We use *absolute* distance:

$$D_c(w, y) = \operatorname{dist}(y, S_c(w)), \qquad H_c(t) = \mathbb{E} D_c(w, F(w + tz)), \quad t \ge 1. \tag{8}$$

Every subset-sum set contains zero, so $D_c(w, y) \le |y|$. The distance is 1-Lipschitz in y and obeys $D_c(sw, sy) = s D_c(w, y)$ for $s > 0$.

Lemma 2 (Distance to a finite set)**.** *Let $g : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz, $L > 0$, and $Z \sim \mathcal{N}(0, I_n)$. If S is nonempty with $|S| \le M$, then*

$$\mathbb{E} \operatorname{dist}(g(Z), S) \ge \frac{\operatorname{Var}(g(Z))}{\sqrt{2\pi}\, LM}. \tag{9}$$

Consequently, $H_c(1) \ge v_W(F)/(\sqrt{2\pi}\, 2^k)$.

Proof sketch. An antiderivative of $\operatorname{dist}(\cdot, S)$ grows by at least $|a - b|^2/(4M)$ between a and b. Applying Gaussian rotation (Pisier, 1986) to this antiderivative composed with g bounds the resulting difference by a constant times $L\mathbb{E} \operatorname{dist}(g(Z), S)$. Conditioning on w gives the network statement. Appendix A.1 gives the full proof.

3.2 A Scale with an Affine Approximation

The next lemma uses the finite number of neurons to find a scale where the learned and missing components can be separated.

Lemma 3 (Affine approximation over a range of scales)**.** *Suppose $k \ge 2$, $v_W(F) \ge v > 0$, $2\|P_{W^\perp}\mu\| \le \sqrt{v}/4$, $B \ge 1$, and $\kappa \ge 2$. Set $A = B\kappa^{3/2}$. Some $t \in \{4\kappa v^{-1/2} A^j : 0 \le j < k - 1\}$ admits $s(w, z) \in S_c(w)$ and $G(w, z)$, both independent of α, such that*

$$\sup_{1 \le \alpha \le B} \mathbb{E}|F(w + \alpha t z) - s(w, z) - \alpha t G(w, z)| \le \kappa^{-1}, \tag{10}$$

where $c_i = a_i P_W u_i$.

Proof sketch. Gaussian Poincaré and the small first moment force at least two missing components to have length at least $3\sqrt{v}/8$. Of the $k - 1$ intervals of projection ratios, one is therefore empty. At its corresponding scale, linearize each ReLU around its larger Gaussian component. Summing the remainders with $\sum_i |a_i| \le 1$ gives (10); the complete proof is in Appendix A.2.

The learner scans these $k - 1$ scales.

4 Detecting and Estimating a Second-Moment Signal

The scale lemma reduces the residual to an affine function of scale. We use this to find a nonzero second moment even when signed terms cancel at individual scales, then estimate that moment from ordinary examples. Define

$$M_c(t) = \mathbb{E} D_c(w, F(w + tz))(zz^\top - I_U), \qquad \overline{M}_c(T) = \frac{1}{\log T} \int_1^T M_c(t)\, \frac{dt}{t}. \tag{11}$$

Differentiating the Gaussian density gives

$$\frac{dH_c(t)}{d \log t} = \operatorname{tr} M_c(t), \qquad \operatorname{tr} \overline{M}_c(T) = \frac{H_c(T) - H_c(1)}{\log T}. \tag{12}$$

This differentiates a smooth density. For $k \ge 2$, a variance threshold $v \in (0, 1)$, and $Q = 2^k$, choose

$$m = \frac{v}{8Q}, \qquad B = 32Q, \qquad \kappa = 8/m, \qquad A = B\kappa^{3/2}, \tag{13}$$

$$T_* = 4\kappa v^{-1/2} A^{k-2}, \qquad \overline{T} = BT_*, \qquad \eta = \frac{m}{16k}, \qquad \lambda = \frac{m}{32k \log \overline{T}}. \tag{14}$$

For each $t_j = 4\kappa v^{-1/2} A^j$, $0 \le j < k - 1$, scan the endpoints t_j and $16nt_j$, $1 \le n \le 2Q$. Their set $\mathcal{T}$ has at most $(k - 1)(2Q + 1)$ elements and maximum $\overline{T}$.

Proposition 4 (One family of second moments). *Suppose $k \geq 2$, $v_W(F) \geq v$, $2\|P_{W^\perp}\mu\| \leq \sqrt{v}/4$, and $\sum_i \|\widehat{c}_i - a_i P_W u_i\| \leq k\eta$. Then*

$$\max_{T\in\mathcal{T}} \left\|\overline{M}_{\widehat{c}}(T)\right\|_{\mathrm{op}} > 8\lambda = \frac{m}{4k\log\overline{T}}. \tag{15}$$

Proof. For any nonempty S with $|S| \leq Q$ and $u, a \in \mathbb{R}$,

$$\frac{1}{2Q}\sum_{n=1}^{2Q} \mathrm{dist}(u + 16na, S) \geq 4|a|. \tag{16}$$

If $a \neq 0$, an open interval of radius $8|a|$ contains at most one progression point. At least Q points therefore have distance at least $8|a|$; the case $a = 0$ is immediate.

Write $H = H_{\widehat{c}}$ and $h = H(1) \geq m$ by Lemma 2. At the scale t from Lemma 3, set $b = t\mathbb{E}|G(w,z)|$. Matching subsets contributes at most $k\eta = m/16$, and the affine remainder is at most $\kappa^{-1} = m/8$. Equation (16) therefore gives

$$H(t) \leq b + 3m/16, \qquad \frac{1}{2Q}\sum_{n=1}^{2Q} H(16nt) \geq 4b - m/8. \tag{17}$$

Let $\Delta = \max_{T\in\mathcal{T}} |H(T) - h|$. Then $h - \Delta \leq b + 3m/16$ and $4b - m/8 \leq h + \Delta$. Eliminating b gives

$$5\Delta \geq 3h - 7m/8 \geq 17m/8, \qquad \Delta > m/4.$$

Equation (12) and $|\operatorname{tr}\overline{M}| \leq k\left\|\overline{M}\right\|_{\mathrm{op}}$ prove the proposition. □

The same $G(w,z)$ works at every multiplier; the learner need not know it. Since $\log\overline{T} \leq 5k^2\log(2/v)$, the signal also satisfies

$$\max_{T\in\mathcal{T}} \left\|\overline{M}_{\widehat{c}}(T)\right\|_{\mathrm{op}} > \frac{v}{160\cdot 2^k k^3 \log(2/v)}. \tag{18}$$

For fixed width, the bound is linear in v up to a logarithm. The largest scale grows as $v^{-3(k-1)/2}$. Thus a strong population signal alone does not determine the sample count: estimating its change-of-measure weights is also costly.

The signal need not be positive. For $F(w,z) = \sigma(z)$, $W = \mathrm{span}(e_1)$, and $c = 0$, $H_c(t) = M_{zz}(t) = t/\sqrt{2\pi} > 0$. In contrast, consider the signed pair

$$F(w,z) = \frac{\sigma(w+z) - \sigma(-w+z)}{2\sqrt{2}}, \qquad W = \mathrm{span}(e_1). \tag{19}$$

Here $F(0,z) = 0$ even though F depends on z. Appendix F shows that $H_c(t) \sim 1/(8\sqrt{\pi}\,t)$ and $M_{zz}(t)/H_c(t) \to -1$. Figure 1 illustrates both signs, which the matrix search treats alike. The accompanying checks concern these identities.

4.1 Ordinary Samples and Coefficient Nets

For an example (X, Y), put $X_W = P_W X$, $Z = P_{W^\perp}X$, $P = P_{W^\perp}$, and $\ell = \dim W$. Define

$$R_t(X_W) = t^\ell \exp\left(-\frac{t^2-1}{2}\|X_W\|^2\right), \qquad K_W(X) = ZZ^\top - P. \tag{20}$$

The ratio R_t compresses Gaussian variance only on W. Homogeneity gives the implementable identity

$$M_c(t) = \mathbb{E}R_t(X_W)tD_c(X_W, Y)K_W(X). \tag{21}$$

Averaging with $t = T^\xi$, $\xi \sim \mathrm{Unif}[0,1]$ independent of the example, implements $\overline{M}_c(T)$. We identify the matrices on U with their extensions by zero to $\mathbb{R}^d$. They vanish on W and have range in $U = P_{W^\perp}V$ for every tuple, by Gaussian independence outside $W + V$.

The initial first-moment estimate $\widehat{\mu}$ will have error at most $\zeta_0 \le \eta/2$ (Section 6 and Appendix B). Since $\sum_i c_i = 2P_W\mu$, enumerate an $\eta/2$-net of the unit ball of W for only $c_1, \dots, c_{k-1}$, and set

$$\widehat{c}_k = 2P_W\widehat{\mu} - \sum_{i<k} \widehat{c}_i, \qquad \sum_i \|\widehat{c}_i - c_i\| \le 2\sum_{i<k} \|\widehat{c}_i - c_i\| + 2\zeta_0 \le k\eta. \tag{22}$$

Thus $g(k)\eta^{-(k-1)\ell}$ tuples suffice for Proposition 4.

4.2 One Batch for All Matrices

Condition on all previous rounds, so that W and the candidate list are fixed. Let $J \le (k-1)(2Q+1)g(k)\eta^{-(k-1)\ell}$ count the scanned matrices. Each is $M = \mathbb{E}\omega K_W$, where the nonnegative scalar weight satisfies

$$\omega \le C_0|Y|, \qquad C_0 = \overline{T}^{\ell+1}, \qquad V_0 = \overline{T}^{\ell+2}.$$

For desired operator error $\zeta < 1$, set

$$L = 4C_0\sqrt{\log(64C_0/\zeta)}, \qquad \widehat{M} = \frac{1}{N_M}\sum_{s=1}^{N_M} \min(\omega_s, L)(Z_sZ_s^\top - P). \tag{23}$$

A shared batch of examples and uniform scales suffices for all J matrices, with

$$N_M = \left\lceil 2^{17}(V_0\zeta^{-2} + L\zeta^{-1})\left(d + \log\frac{2J}{\delta'}\right)\right\rceil. \tag{24}$$

With probability at least $1-\delta'$, all estimates have operator error at most ζ. Appendix B proves this using Gaussian label tails, $\mathbb{E}R_t^2 \le t^\ell$, and Bernstein's inequality. Every estimate is symmetric and vanishes on W.

For $\alpha_s = \min(\omega_s, L)$, the product

$$\widehat{M}a = \frac{1}{N_M}\sum_s \alpha_s\big((Z_s^\top a)Z_s - Pa\big) \tag{25}$$

costs $O(N_Md + kd)$ operations after the projections are stored.

5 Stable Directions and Proper Output

An estimated moment need not have range inside the target span. We first bound this error for every coefficient tuple, not only the nearby tuple guaranteed by the net. We then convert small conditional variance into a network in the original class. Measure subspace error by $e_W = \|(I-P_V)P_W\|_{\mathrm{op}}$. For fixed w, the distance is a piecewise affine function of Z, with weak gradient $t\theta_{c,t}P\nabla F(w+tZ)$ for some $|\theta_{c,t}| \le 1$. Gaussian integration by parts (Stein, 1981) therefore gives

$$M_c(t) = t\,\mathbb{E}\big[\theta_{c,t}P\nabla F(w+tZ)Z^\top\big], \qquad Z \sim \mathcal{N}(0,P). \tag{26}$$

Since $(I-P_V)P\nabla F = -(I-P_V)P_W\nabla F$, (5) and $\mathbb{E}|\langle a, Z\rangle| \le 1$ for unit a imply

$$\|(I-P_V)M_c(t)\|_{\mathrm{op}} \le te_W, \qquad \big\|(I-P_V)\overline{M}_c(T)\big\|_{\mathrm{op}} \le \frac{T-1}{\log T}\,e_W. \tag{27}$$

These identities hold for every tuple and every W.

Put $s_* = (\overline{T}-1)/\log\overline{T}$. If $\left\|\widehat{M} - \overline{M}_c(T)\right\|_{\mathrm{op}} \le \zeta$ and $\left\|\widehat{M}a\right\| \ge 2\lambda$ for a unit a, then

$$b = \frac{\widehat{M}a}{\left\|\widehat{M}a\right\|} \in W^\perp, \qquad \|(I-P_V)b\| \le \frac{\zeta + s_*e_W}{2\lambda}. \tag{28}$$

For $W' = W \oplus \mathrm{span}(b)$, Cauchy–Schwarz now gives

$$e_{W'} \le \beta e_W + \frac{\zeta}{2\lambda}, \qquad \beta = 1 + \frac{s_*}{2\lambda} = 1 + \frac{16k(\overline{T}-1)}{m}. \tag{29}$$

Lemma 6 finds an image of size at least $\left\|\widehat{M}\right\|_{\mathrm{op}}/2$ in $O(\log(d/\alpha))$ matrix products, with failure probability at most α, for either sign. Normalize after each product; reject a zero vector.

5.1 A Network in the Same Normalized Class

The conditional mean need not be a narrow ReLU network, and projecting the weights can lose the variance guarantee (Appendix D). The following rotation avoids both problems while preserving the class.

Lemma 5 (Proper approximation by rotation)**.** *For any subspace S with $\dim S \geq r$, there is $G \in \mathcal{F}_k$ with all its inner weights in S such that*

$$\|F - G\|_2^2 \leq 2v_S(F). \tag{30}$$

The output coefficients and the norm of each inner weight are unchanged.

Proof sketch. Choose principal-vector bases of V and a subspace of S, and rotate one basis to the other. This preserves every weight norm. The two Gaussian coordinate vectors admit a coupling that is conditionally i.i.d. given a shared Gaussian component; total variance then gives the factor two. The construction and full argument appear in Appendix D.1.

Once the subspace search stops at W, extend it to any S_W of dimension $s = \min(k, d)$. All moment estimates use W; only the final network search uses S_W. Choose

$$v = \varepsilon/128, \qquad \rho = \sqrt{\varepsilon}/8. \tag{31}$$

By ReLU homogeneity, absorb $|a_i|$ into the inner weight:

$$G(x) = \sum_i s_i \sigma(\langle v_i, x\rangle), \qquad s_i \in \{-1, 1\}, \quad \sum_i \|v_i\| \leq 1. \tag{32}$$

This standard rescaling (Bach, 2017) preserves the normalized class. Enumerate signs and a ρ-net in the metric $\sum_i \|v_i - \widehat{v}_i\|$. Rounding each coordinate toward zero preserves the norm constraint and constructs the net directly (Appendix D). Then $N_{\text{fit}} \leq g(k)\varepsilon^{-ks/2}$. Since $v_{S_W}(F) \leq v_W(F)$, Lemma 5 and ReLU Lipschitzness give a candidate H satisfying

$$\|F - H\|_2 \leq \sqrt{2v_W(F)} + \rho \leq \sqrt{\varepsilon}/4 \quad \text{if } v_W(F) \leq v. \tag{33}$$

For any normalized candidate H, Gaussian Poincaré gives

$$\operatorname{Var}((F-H)^2) \leq 4\mathbb{E}[(F-H)^2 \|\nabla(F-H)\|^2] \leq 16\|F-H\|_2^2. \tag{34}$$

Thus median-of-means validation needs only $O(\varepsilon^{-1}\log(2N_{\text{fit}}/\delta'))$ fresh examples: accept median loss at most $\varepsilon/4$. With probability at least $1-\delta'$, all candidates of risk at most $\varepsilon/16$ are accepted, and none of risk at least $\varepsilon/2$ is accepted (Appendix D). This test is performed only once, after the moment search has established $v_W(F) \leq v$.

6 The Learner and Proof of Theorem 1

For $k = 1$, nonzero labels give exact linear equations after removing their sign. Randomized projection updates attain the required accuracy using $O(d\log(2/(\varepsilon\delta)))$ examples and $O(d^2\log(2/(\varepsilon\delta)))$ operations (Appendix E). Below assume $k \geq 2$.

At $W = \{0\}$, homogeneity removes the scale search. Use

$$M_{\text{init}} = \mathbb{E}|Y|(XX^\top - I), \qquad \operatorname{tr} M_{\text{init}} = \mathbb{E}|Y|, \qquad \operatorname{range} M_{\text{init}} \subseteq V. \tag{35}$$

Indeed, $H_0(t) = t\mathbb{E}|Y|$ in (12). If $v_{\{0\}}(F) > v$, Lemma 2 with $S = \{0\}$ gives $\|M_{\text{init}}\|_{\text{op}} > v/(\sqrt{2\pi}k)$. On one fresh batch, estimate both M_{init} and μ to error ζ_0 using $C_0 = V_0 = 1$, $J = 2$ in (23)–(24). For μ, average $\operatorname{sgn}(Y)\min(|Y|, L)X$. Appendix B verifies the same concentration bound.

The factor β in (29) amplifies earlier errors through later extensions. We therefore estimate earlier moments more accurately. At round $j = \dim W < k$, choose the image threshold and estimation accuracy as

$$\lambda_0 = \frac{v}{32k}, \qquad \lambda_j = \lambda \ (j \geq 1), \qquad \zeta_j = \frac{\lambda_j\sqrt{v}}{k\beta^{k-1-j}}. \tag{36}$$

Algorithm 1 Gaussian narrow-ReLU learner

Require: $k \geq 2, d, \varepsilon, \delta$ and independent examples $(X, F(X))$
Set the parameters above; $W \leftarrow \{0\}$; $\delta' \leftarrow \delta/(3(k+1))$
for $j = 0, \ldots, \min(k, d) - 1$ **do**
  Set λ_j, ζ_j as in (36)
  if $j = 0$ **then**
    Estimate M_{init} and μ on a fresh batch, with $C_0 = V_0 = 1$, $J = 2$
    If $\|\widehat{\mu}\| \geq \sqrt{v}/16$, set $W \leftarrow \text{span}(\widehat{\mu})$ and **continue**
  else
    Complete tuples by (22); estimate all matrices on one fresh batch
  end if
  For each matrix estimate, use (44) with $\alpha = \delta'/J$
  If no resulting unit a has $\left\|\widehat{M}a\right\| \geq 2\lambda_j$, **break**
  Choose one such image, set $b = \widehat{M}a / \left\|\widehat{M}a\right\|$, and $W \leftarrow W \oplus \text{span}(b)$
end for
Extend W to S_W of dimension $\min(k, d)$; validate its width-k net on fresh examples
return a candidate of median loss $\leq \varepsilon/4$, or zero if none is accepted

If $\|\widehat{\mu}\| \geq \gamma := \sqrt{v}/16$, choose $b = \widehat{\mu}/\|\widehat{\mu}\|$ as the first direction. Otherwise use M_{init}. In either case $\|P_{W^\perp}\widehat{\mu}\| \leq \gamma$, so every subsequent W satisfies

$$\|P_{W^\perp}\mu\| \leq \zeta_0 + \gamma \leq 3\sqrt{v}/32. \tag{37}$$

Thus the first-moment premise of Proposition 4 holds at every later round. During the search, if the variance exceeds v, a detectable moment must exist.

Proof sketch of Theorem 1. For $k = 1$, use the procedure above. For $k \geq 2$, condition on accurate moment estimates and successful power iterations. The error recurrence gives $e_{W_j} \leq \sqrt{v}/2$ and $\dim W_j \leq r$ at every round. If $v_W(F) > v$, the initial matrix or Proposition 4 forces an accepted image. Thus early stopping certifies $v_W(F) \leq v$. At the dimension cap, projection gaps give the same variance bound. Lemma 5 and the parameter net then supply a candidate of risk at most $\varepsilon/16$. Independent validation returns risk below $\varepsilon/2$. A union bound over at most $r \leq k$ rounds and validation controls the failure probability by δ. Appendix E.1 gives the full proof, including the subspace invariant, stopping argument, and complexity bounds. □

Smaller accuracy exponents. An alternative in Appendix G adds a cross moment whose range lies in V, then retains strong directions from all previous images. It selects a direction, then estimates a vector mean after rescaling its W component. No factor β^{k-j} is needed. For $k \geq 4$, the sample and operation exponents become $3k^2/2 + 3k - 5/2$ and $5k^2/2 - 2k + 3/2$, with the same d and d^2 factors. At $k = 3$ both exponents are 20; at $k = 2$ we keep Algorithm 1.

7 Conclusion

For normalized signed, bias-free ReLU networks under Gaussian inputs, we obtain proper learning with width-independent dimension exponents and polynomial inverse-accuracy dependence at fixed width. The analysis connects residual prediction error to a scaled second-moment signal. Subspace-error control and norm-preserving rotation then yield an output in the original class.

These guarantees rely on assumptions tied to specific proof steps. Biases break the homogeneity behind (21), so an extension needs another way to estimate scaled moments from ordinary examples. For noisy labels, the obstacle instead lies in $D_c(w, Y)$: even independent, mean-zero noise need not average out through this nonlinear distance. Extending this analysis to noise therefore requires controlling the induced moment distortion relative to the target's residual variance. Finally, large scales inflate estimation variance bounds, while fine coefficient nets increase computation. Reducing these costs without sacrificing the dimension dependence is a further direction for improving the $O(k^2)$ accuracy exponent.

AI USE STATEMENT

OpenAI Codex and ChatGPT assisted with formulating claims, developing and revising proofs, literature search, writing, and the design, implementation, and interpretation of synthetic numerical checks. They also helped revise the notation and references. The authors are responsible for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Theorem 1, Sections 2–6, and the appendices specify the model, algorithm, and complexity bounds.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.html.

Pranjal Awasthi, Alex Tang, and Aravindan Vijayaraghavan. Efficient algorithms for learning depth-2 neural networks with general ReLU activations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13485–13496, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/700fdb2ba62d4554dc268c65add4b16e-Abstract.html.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017. URL https://jmlr.org/papers/v18/14-546.html.

Ainesh Bakshi, Rajesh Jayaram, and David P. Woodruff. Learning two layer rectified neural networks in polynomial time. In *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pp. 195–268. PMLR, 2019. URL https://proceedings.mlr.press/v99/bakshi19a.html.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973. doi: 10.1090/S0025-5718-1973-0348991-3.

Sitan Chen and Raghu Meka. Learning polynomials in few relevant dimensions. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1161–1227. PMLR, 2020. URL https://proceedings.mlr.press/v125/chen20a.html.

Sitan Chen and Shyam Narayanan. A faster and simpler algorithm for learning shallow networks. In *COLT*, volume 247 of *Proceedings of Machine Learning Research*, pp. 981–994. PMLR, 2024. URL https://proceedings.mlr.press/v247/chen24b.html.

Sitan Chen, Adam R. Klivans, and Raghu Meka. Efficiently learning any one hidden layer ReLU network from queries. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2111.04727.

Sitan Chen, Adam R. Klivans, and Raghu Meka. Learning deep ReLU networks is fixed-parameter tractable. In *2021 IEEE FOCS*, pp. 696–707. IEEE, 2022. doi: 10.1109/FOCS52979.2021.00073.

Sitan Chen, Zehao Dou, Surbhi Goel, Adam Klivans, and Raghu Meka. Learning narrow one-hidden-layer ReLU networks. In *COLT*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5580–5614. PMLR, 2023. URL https://proceedings.mlr.press/v195/chen23a.html.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, pp. 342–350, 2009. URL https://cseweb.ucsd.edu/~yoc002/paper/nips09_arccos.pdf.

Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9(53):1647–1678, 2008. URL https://jmlr.org/papers/v9/dalalyan08a.html.

Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. Computational-statistical gaps in Gaussian single-index models (extended abstract). In *COLT*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1262. PMLR, 2024. URL https://proceedings.mlr.press/v247/damian24a.html.

Alex Damian, Jason D. Lee, and Joan Bruna. The generative leap: Tight sample complexity for efficiently learning Gaussian multi-index models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0953.

Luc Devroye, Matthieu Lerasle, Gábor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016. doi: 10.1214/16-AOS1440.

Ilias Diakonikolas and Daniel M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE FOCS*, pp. 184–195. IEEE, 2020. doi: 10.1109/FOCS46700.2020.00026.

Ilias Diakonikolas and Daniel M. Kane. Efficiently learning one-hidden-layer ReLU networks via Schur polynomials. In *COLT*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1364–1378. PMLR, 2024. URL https://proceedings.mlr.press/v247/diakonikolas24c.html.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, and Nikos Zarifis. Algorithms and SQ lower bounds for PAC learning one-hidden-layer ReLU networks. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pp. 1514–1539. PMLR, 2020. URL https://proceedings.mlr.press/v125/diakonikolas20d.html.

Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Lisheng Ren. Algorithms and SQ lower bounds for robustly learning real-valued multi-index models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2177. URL https://papers.nips.cc/paper_files/paper/2025/hash/5debeab774db7ca61ea15f1117f7fa33-Abstract-Conference.html.

Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.00501.

Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3587–3596. PMLR, 2020. URL https://proceedings.mlr.press/v119/goel20a.html.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016. URL https://www.jmlr.org/papers/v17/14-273.html.

Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001. URL https://www.ams.org/books/surv/089/.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007. doi: 10.1198/016214507000000536.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992. doi: 10.1080/01621459.1992.10476258.

Alessandra Lunardi and Diego Pallara. Ornstein–Uhlenbeck semigroups in infinite dimension. *Philosophical Transactions of the Royal Society A*, 378(2185):20190620, 2020. doi: 10.1098/rsta.2019.0620. URL https://arxiv.org/abs/2005.09499.

Gilles Pisier. Probabilistic methods in the geometry of Banach spaces. In *Probability and Analysis*, volume 1206 of *Lecture Notes in Mathematics*, pp. 167–241. Springer, 1986. doi: 10.1007/BFb0076302.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. doi: 10.1214/aos/1176345632.

Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009. URL https://arxiv.org/abs/math/0702226.

Robert van de Geijn and Margaret Myers. *Advanced Linear Algebra: Foundations to Frontiers*. The University of Texas at Austin, 2026. URL https://www.cs.utexas.edu/~flame/laff/alaff/chapter11-Jacobi-method-spectral-decomposition.html. Online edition, January 26, 2026.

Larry Wasserman and Kathryn Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009. doi: 10.1214/08-AOS646. URL https://arxiv.org/abs/0704.1139.

Zhuoran Yang, Krishnakumar Balasubramanian, Zhaoran Wang, and Han Liu. Estimating high-dimensional non-gaussian multiple index models via Stein's lemma. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/4db0f8b0fc895da263fd77fc8aecabe4-Abstract.html.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 4140–4149. PMLR, 2017. URL https://proceedings.mlr.press/v70/zhong17a.html.

A Finite-Set Distance, Scaling, and Gaussian Moments

A.1 Proof of the Finite-Set Distance Lemma

Proof of Lemma 2. Set $q(s) = \int_0^s \operatorname{dist}(u, S)\, du$. The r-neighborhood of S has length at most $2Mr$. Integrating over r on the interval between a and b gives

$$|q(a) - q(b)| \geq \int_0^{|a-b|/(2M)} (|a-b| - 2Mr)\, dr = \frac{|a-b|^2}{4M}.$$

For independent standard Gaussians Z, Z', put $Z_\theta = Z\cos\theta + Z'\sin\theta$. At each θ, Z_θ and $\dot{Z}_\theta$ are independent standard Gaussians. Gaussian rotation (Pisier, 1986) therefore gives

$$\begin{aligned} 2\operatorname{Var}(g(Z)) &\leq 4M\mathbb{E}|q(g(Z)) - q(g(Z'))| \\ &\leq 4M\int_0^{\pi/2} \mathbb{E}\big[\operatorname{dist}(g(Z_\theta), S)|\langle \nabla g(Z_\theta), \dot{Z}_\theta\rangle|\big]\, d\theta \\ &\leq 4M\sqrt{\pi/2}\, L\mathbb{E}\operatorname{dist}(g(Z), S). \end{aligned} \tag{38}$$

The chain rule holds almost everywhere; Lipschitzness gives Gaussian integrability. Conditioning on w proves the final assertion. □

A.2 Proof of the Scale-Selection Lemma

Proof of Lemma 3. Write $p_i = P_W u_i$ and $q_i = P_U u_i$. Let h index a longest q_i and let b be the second largest length. Gaussian Poincaré (Ledoux, 2001) and the triangle inequality give

$$\sqrt{v} \leq \sum_i |a_i|\, \|q_i\| \leq \left\|\sum_i a_i q_i\right\| + 2\sum_{i\neq h} |a_i|\, \|q_i\| \leq \sqrt{v}/4 + 2b. \tag{39}$$

Thus two lengths are at least $3\sqrt{v}/8$. Put $r_i = \|q_i\| / \|p_i\|$, with $r_i = \infty$ when $p_i = 0 \neq q_i$ and $r_i = 0$ when $q_i = 0$. Since $\|p_i\| \leq 1$, at most $k-2$ ratios lie below $\sqrt{v}/4$. One of the $k-1$ intervals $(\sqrt{v}A^{-j-1}/4, \sqrt{v}A^{-j}/4)$, $0 \leq j < k-1$, is empty. At $t = 4\kappa v^{-1/2}A^j$, each ratio satisfies $tr_i \leq 1/(B\sqrt{\kappa})$ or $tr_i \geq \kappa$.

For independent centered Gaussians a, b with standard deviations $s_a > 0, s_b$, the residual below is nonnegative, so

$$\mathbb{E}|\sigma(a+b) - (a+b)\mathbf{1}\{a+b>0\}| = \frac{\sqrt{s_a^2 + s_b^2} - s_a}{\sqrt{2\pi}} \leq \frac{s_b^2}{s_a}. \tag{40}$$

Write $x_i = \langle p_i, w\rangle$ and $z_i = \langle q_i, z\rangle$. Choose $e_i = \mathbf{1}\{x_i > 0\}$ for small ratios and $e_i = \mathbf{1}\{z_i > 0\}$ for large ratios, and set

$$s = \sum_i a_i x_i e_i \in S_c(w), \qquad G = \sum_i a_i z_i e_i.$$

Apply (40) around x_i in the first case and around $\alpha t z_i$ in the second. The respective remainders are at most $\|p_i\|(\alpha t r_i)^2 \leq \kappa^{-1}$ and $\|p_i\|/(\alpha t r_i) \leq \kappa^{-1}$. Neurons with $p_i = 0$ or $q_i = 0$ are exact. Sum using $\sum_i |a_i| \leq 1$. □

A.3 Moment and Change-of-Measure Identities

Put $n = \dim U$. If $n = 0$, then $v_W(F) = 0$ and the signal proposition has no premise. Otherwise, differentiating the density $\varphi_{t^2 I_U}$ gives

$$\frac{\partial}{\partial \log t}\varphi_{t^2 I_U}(u) = \left(\frac{\|u\|^2}{t^2} - n\right)\varphi_{t^2 I_U}(u).$$

Apply this to $u \mapsto D_c(w, F(w+u))$, which has at most linear growth, then integrate in w. On compact positive scale intervals a Gaussian times a polynomial dominates the differentiated integrand, proving (12).

The scale parameters give $\log \overline{T} = 3(k-1)\log(1/v)/2 + (5k^2/2 + 11k - 15)\log 2 \leq 5k^2 \log(2/v)$ for integer $k \geq 2$, proving (18).

Implementing the Moments. Under R_t, the input has distribution $w/t + z$. Homogeneity of both the network and the finite set gives

$$tD_c(w/t, F(w/t + z)) = D_c(w, F(w + tz)). \tag{41}$$

The centered quadratic in z is unchanged, proving (21).

For arbitrary W, decompose $W^\perp = U \oplus (W+V)^\perp$. The scalar weight depends only on (w, z_U). The cross blocks containing $z_{(W+V)^\perp}$ have zero mean, and the remaining diagonal block has a centered quadratic independent of the weight. Thus every tuple gives range contained in $U = P_{W^\perp} V$.

B Shared Samples, Clipping, and Coefficient Nets

Fix a visited W and condition on all earlier rounds. For each sample, draw an independent $\xi_s \sim \mathrm{Unif}[0, 1]$. Every endpoint T uses $t_s = T^{\xi_s}$ on that same example. For each fixed candidate, these pairs are independent across s.

The dimension-free label bound is

$$\mathbb{E} \exp(Y^2/4) \le (1 + \sqrt{2})/2 < 2. \tag{42}$$

Apply convexity of $s \mapsto e^{s^2/4}$ to $|Y| \le \sum_i |a_i| \sigma(\langle u_i, X\rangle)$, adding any missing weight at zero. Each Gaussian ReLU has exponential moment at most $(1 + \sqrt{2})/2$. For every matrix, $\omega \le C_0|Y|$. For a unit b, put $Q_b = (b^\top Z)^2 - \|Pb\|^2$; then $\mathbb{E}Q_b^2 = 2\|Pb\|^4 \le 2$. Writing $h = L/C_0$, the bound $y^2 \le 8e^{y^2/8}$ implies

$$\mathbb{E}Y^2 \mathbf{1}\{|Y| > h\} \le 8e^{-h^2/8} \mathbb{E}e^{Y^2/4} \le 16e^{-h^2/8}.$$

Cauchy–Schwarz and $(\omega - L)_+^2 \le C_0^2 Y^2 \mathbf{1}\{|Y| > h\}$ give clipping bias at most

$$\sqrt{32}\, C_0 e^{-h^2/16} = \sqrt{32}\, \zeta/64 < \zeta/8.$$

Concentration with a Linear Dimension Factor. For fixed t, weighting by R_t^2 gives the exact identity

$$c_t = \mathbb{E}R_t^2 = \frac{t^{2\ell}}{(2t^2 - 1)^{\ell/2}} \le t^\ell.$$

Under the probability measure R_t^2/c_t, the input is Gaussian with covariance $(2t^2-1)^{-1}P_W + P \preceq I$. Write $\mathbb{E}_t$ for this expectation. Then $\mathbb{E}_t Y^4 \le 3/2$ and $\mathbb{E}_t \exp(|Q_b|/4) \le e^{1/4}\sqrt{2} < 2$. Consequently, for every integer $p \ge 2$,

$$\mathbb{E}_t[Y^2 |Q_b|^p] \le \sqrt{3(2p)!\, 4^{2p}} \le \sqrt{3}\, 8^p p!.$$

Put $\xi_b = \min(\omega, L) Q_b$. Using $\omega^2 \le R_t^2 t^2 Y^2$ and averaging over t gives $\mathbb{E}|\xi_b|^p \le \sqrt{3}\, 8^p p! V_0 L^{p-2}$. The centered variable $\widetilde{\xi}_b = \xi_b - \mathbb{E}\xi_b$ thus satisfies

$$\mathbb{E}|\widetilde{\xi}_b|^p \le \frac{p!}{2}(1024 V_0)(16L)^{p-2}. \tag{43}$$

The scalar Bernstein bound, obtained by expanding the moment-generating function, is

$$\mathbb{P}\left(\left| N^{-1} \sum_s \widetilde{\xi}_{b,s} \right| > u \right) \le 2 \exp\left(-\frac{N u^2}{2(1024 V_0 + 16 L u)} \right).$$

A $1/4$-net $\mathcal{Q}$ of the unit sphere has size at most 9^d, and every symmetric E obeys $\|E\|_{\mathrm{op}} \le 2 \max_{b \in \mathcal{Q}} |b^\top E b|$. Taking $u = \zeta/4$ and substituting (24) makes the sample-mean operator error at most $\zeta/2$ for all J candidates with probability at least $1 - \delta'$. The bias is below $\zeta/8$, so the total error is less than ζ.

The clipped average is symmetric and annihilates W on every sample outcome.

The First Moment. For the initial vector estimate, replace Q_b by $b^\top X$ and $\min(\omega, L)$ by $\mathrm{sgn}(Y) \min(|Y|, L)$. Then $\mathbb{E}Q_b^2 = 1$ and $\mathbb{E}e^{|Q_b|/4} \le e^{1/8} \mathbb{E}e^{Q_b^2/8} = 2e^{1/8}/\sqrt{3} < 2$. The same bias and moment estimates apply with $C_0 = V_0 = 1$. For a vector e, the same sphere net gives $\|e\| \le (4/3) \max_{b \in \mathcal{Q}} |b^\top e|$. Thus (24) with $J = 2$ controls the initial matrix and vector together, even though they share examples.

Coefficient Nets. For $\ell > 0$, put $M = \lceil 2\sqrt{\ell}/\eta \rceil$ and retain the points of $M^{-1}\mathbb{Z}^\ell$ in the unit ball. Rounding each coordinate toward zero stays in the ball and has Euclidean error at most $\sqrt{\ell}/M \le \eta/2$. There are at most $(2M+1)^\ell \le (C\sqrt{\ell}/\eta)^\ell$ points. Use it for $k-1$ vectors and complete the last one by (22). The required initial accuracy follows from $\overline{T} > 2$, $\beta \ge 1/m$, and $k \ge 2$:

$$\zeta_0 = \frac{v^{3/2}}{32k^2\beta^{k-1}} \le \frac{mv^{3/2}}{32k^2} \le \frac{\eta}{4}.$$

There are $g(k)\eta^{-(k-1)\ell}$ tuples. The completed vector need not have norm at most one: all moment, range, and concentration bounds hold for arbitrary tuples, since $0 \in S_c(w)$ still gives $D_c(w, Y) \le |Y|$. At $\ell = 0$, use the initial moment directly. The endpoint estimate in (17) uses $\mathbb{E}\sum_i |\langle c_i - \widehat{c}_i, w\rangle| = \sqrt{2/\pi}\sum_i \|c_i - \widehat{c}_i\| \le k\eta$. The law of w is unchanged by scaling z. At each example, the 2^k subset sums are computed after the learned coordinates are known. This costs a function of k.

C Gaussian Integration by Parts and Subspace Error

Lemma 6 (Finding a vector with a large image)**.** *For any nonzero symmetric $d \times d$ matrix A and $\alpha \in (0,1)$, draw $g \sim \mathcal{N}(0, I_d)$ and set*

$$q = \left\lceil \log_2(64d/\alpha^3) \right\rceil, \qquad a = \frac{A^q g}{\|A^q g\|}. \tag{44}$$

Then $\|Aa\| \ge \|A\|_{\mathrm{op}}/2$ with probability at least $1-\alpha$.

Proof. Write $\Lambda = \|A\|_{\mathrm{op}}$. With probability at least $1-\alpha$, the coordinate g_1 along an eigenvector of magnitude Λ satisfies $|g_1| \ge \alpha/4$, and $\|g\|^2 \le 2d/\alpha$. The squared mass of $A^q g$ in eigenvalues of magnitude below $\Lambda/\sqrt{2}$, divided by its total squared norm, is at most $2^{-q}\|g\|^2/g_1^2 \le 1/2$. Thus $\|Aa\|^2 \ge (\Lambda^2/2)(1/2)$. □

Subspace Error. Fix w and use orthonormal coordinates on $W^\perp$. The function $f(z) = D_c(w, F(w+tz))$ is Lipschitz and piecewise affine. On each open affine region, its gradient is $t\theta_{c,t}P\nabla F(w+tz)$, where $\theta_{c,t} \in \{-1, 0, 1\}$. The value zero covers regions where the label equals a root. The boundaries have measure zero; hence this is also its weak gradient. Gaussian integration by parts applied to $f(z)z_j$ gives

$$\mathbb{E}[f(Z)(Z_iZ_j - \delta_{ij})] = \mathbb{E}[\partial_i f(Z)Z_j], \tag{45}$$

which proves (26). Linear growth of f and boundedness of its gradient justify truncation at infinity.

Equation (5) then gives (27). The function $(T-1)/\log T$ increases for $T > 1$, since $\log T - 1 + 1/T \ge 0$. Thus s_* bounds this factor at every endpoint.

For b in (28), put $b_0 = (\zeta + s_* e_W)/(2\lambda)$. Every unit vector of $W' = W \oplus \mathrm{span}(b)$ is $w + ab$ with $\|w\|^2 + a^2 = 1$. Cauchy–Schwarz gives

$$e_{W'} \le \sqrt{e_W^2 + b_0^2} \le e_W + b_0 = \beta e_W + \zeta/(2\lambda). \tag{46}$$

This holds for every selected tuple.

Finally, when $\dim W = \dim V = r$, let B_W, B_V have orthonormal columns spanning the two spaces and put $C = B_V^\top B_W$. The squared projection gaps are the largest eigenvalues of $I - C^\top C$ and $I - CC^\top$. These agree because C is square. Consequently $\|(I - P_W)P_V\|_{\mathrm{op}} = e_W$, and

$$\|F - F \circ P_W\|_2 \le \sum_i |a_i|\,\|(I - P_W)u_i\| \le e_W,$$

as used in (59).

D Proper Approximation and Validation

D.1 Proof of the Rotation Lemma

Proof of Lemma 5. The case $r = 0$ is immediate. Choose matrices B, L with orthonormal columns spanning V and an r-dimensional subspace of S, respectively, such that $P_S B = LC$ and $C = \operatorname{diag}(\gamma_i)$, $0 \le \gamma_i \le 1$ (Björck & Golub, 1973). Write $F(x) = f(B^\top x)$ and set $G(x) = f(L^\top x)$. This replaces u_i by $LB^\top u_i$, preserving its norm.

Put $\xi = B^\top X$ and $\xi' = L^\top X$. The remaining coordinates in S are independent of (ξ, ξ'), so $v_S(F) = \mathbb{E}\operatorname{Var}(f(\xi) \mid \xi')$. For independent standard Gaussians Z_0, Z_1, Z_2, the joint law is

$$\xi = C^{1/2} Z_0 + (I - C)^{1/2} Z_1, \qquad \xi' = C^{1/2} Z_0 + (I - C)^{1/2} Z_2.$$

Given Z_0, the two vectors are independent and identically distributed, so $\|F - G\|_2^2 = 2\mathbb{E}\operatorname{Var}(f(\xi) \mid Z_0)$. For $h_0 = \mathbb{E}[f(\xi) \mid Z_0]$, conditional independence and total variance give

$$v_S(F) = \mathbb{E}\operatorname{Var}(f(\xi) \mid Z_0) + \mathbb{E}\operatorname{Var}(h_0 \mid \xi') \ge \tfrac{1}{2}\|F - G\|_2^2 . \tag{47}$$

□

D.2 Constructing the Output Space

The learner extends a basis of W by coordinate vectors using Gram–Schmidt until $\dim S_W = \min(k, d)$, in $g(k)d^2$ operations. The rotation in Lemma 5 proves existence; the learner only enumerates a parameter net in S_W. To obtain its bases, take orthonormal bases B_0, Q of V, S and a thin SVD $Q^\top B_0 = UCR^\top$. Then $B = B_0 R$ and $L = QU$ satisfy $P_S B = LC$.

The coupling in Lemma 5 is coordinatewise Gaussian averaging, as in the Ornstein–Uhlenbeck semigroup (Lunardi & Pallara, 2020). Its conditional-variance identity (47) applies to any square-integrable f.

Why Direct Restriction Can Lose Accuracy. For independent standard Gaussians (w, z, y), set

$$F_\alpha(w, z) = c_\alpha[\sigma(z + \alpha w) + \sigma(z - \alpha w) - 2\sigma(z)] = c_\alpha(\alpha|w| - |z|)_+, \qquad c_\alpha = \frac{1}{2(\sqrt{1 + \alpha^2} + 1)}.$$

After normalizing the three inner weights, their output coefficients have total absolute value one. For $W = \operatorname{span}(e_w)$ and $\alpha \downarrow 0$, we have

$$v_W(F_\alpha) \sim \frac{4c_\alpha^2 \alpha^3}{3\pi}, \qquad \|F_\alpha - F_\alpha \circ P_W\|_2^2 \sim c_\alpha^2 \alpha^2. \tag{48}$$

To verify the first asymptotic, let $A = \alpha|w|$ and let ϕ be the standard Gaussian density. Then

$$\mathbb{E}F_\alpha^2 = 2c_\alpha^2 \mathbb{E}\int_0^A (A - s)^2 \phi(s)\, ds \sim \frac{4c_\alpha^2 \alpha^3}{3\pi}, \qquad \mathbb{E}[F_\alpha \mid w] \le c_\alpha \phi(0) \alpha^2 w^2.$$

The squared conditional mean is $O(c_\alpha^2 \alpha^4)$ in expectation. For the second asymptotic, dominated convergence gives

$$\frac{(F_\alpha - F_\alpha \circ P_W)^2}{c_\alpha^2 \alpha^2} = \min\{|w|, |z|/\alpha\}^2 \longrightarrow w^2 \quad \text{almost surely, with upper bound } w^2.$$

Thus restriction to W need not have error $O(v_W(F))$. Moving the z direction to y instead gives a normalized three-unit network $G_\alpha(w, y) = F_\alpha(w, y)$ with $\|F_\alpha - G_\alpha\|_2^2 = 2v_W(F_\alpha)$, because the two labels are conditionally independent given w.

D.3 Normalized Parameter Nets

Equation (32) uses $v_i = |a_i| u_i$ and $s_i = \operatorname{sgn}(a_i)$, choosing $s_i = 1$ if $a_i = 0$. Conversely, set $a_i = s_i \|v_i\|$ and $u_i = v_i / \|v_i\|$, with $u_i = 0$ when $v_i = 0$. Thus the parameterization gives exactly the same normalized class.

Let $s = \dim S_W = \min(k, d)$, $\rho = \sqrt{\varepsilon}/8$, and $M = \lceil k\sqrt{s}/\rho \rceil$. Enumerate the 2^k sign choices and all grid tuples $(\widehat{v}_i) \in (M^{-1}\mathbb{Z}^s)^k$ satisfying $\sum_i \|\widehat{v}_i\| \le 1$. For every admissible (v_i), rounding coordinates toward zero gives

$$|\widehat{v}_{i,j}| \le |v_{i,j}|, \qquad \sum_i \|v_i - \widehat{v}_i\| \le \frac{k\sqrt{s}}{M} \le \rho.$$

Thus the rounded tuple remains admissible. The candidate count obeys

$$N_{\mathrm{fit}} \le 2^k (2M+1)^{ks} \le g(k)\varepsilon^{-ks/2} \le g(k)\varepsilon^{-k^2/2}. \tag{49}$$

The ReLU Lipschitz bound gives the entire rounding estimate:

$$\left\| \sum_i s_i [\sigma(\langle v_i, X\rangle) - \sigma(\langle \widehat{v}_i, X\rangle)] \right\|_2 \le \sum_i \|v_i - \widehat{v}_i\| \le \rho. \tag{50}$$

Together with Lemma 5, this proves (33).

Testing Losses. Put $R = \|F - H\|_2^2$ and $L = (F(X) - H(X))^2$. Both networks are 1-Lipschitz, so Gaussian Poincaré (Ledoux, 2001) gives (34). The weak gradient and all moments are integrable, justifying this application to the square. Take blocks of $n = \lceil 1024/\varepsilon \rceil$ examples and let $\overline{L}$ be a block mean. Chebyshev's inequality gives

$$\mathbb{P}\{|\overline{L} - R| \ge |R - \varepsilon/4|\} \le \frac{16R}{n(R - \varepsilon/4)^2} \le \frac{1}{8}$$

whenever $R \le \varepsilon/16$ or $R \ge \varepsilon/2$. The case $R = 0$ is immediate. For an odd number $b \ge 4\log(2N_{\mathrm{fit}}/\delta')$ of independent blocks, a majority bound makes each wrong decision have probability at most $e^{-9b/32} \le \delta'/N_{\mathrm{fit}}$. A union bound over candidates thus proves the stated test using

$$N_{\mathrm{val}} = O\left(\varepsilon^{-1} \log \frac{2N_{\mathrm{fit}}}{\delta'}\right) \tag{51}$$

examples, shared by all candidates (Devroye et al., 2016). Accept when the median block mean is at most $\varepsilon/4$. Every accepted candidate has risk below $\varepsilon/2$ on this event. If $v_W(F) \le v$, at least one is accepted.

E Complexity and the Real-Arithmetic Model

Functions $g_i(k)$ below are positive computable functions independent of d, ε, δ. For $k \ge 2$, the basic parameters satisfy

$$\overline{T} = g_1(k) v^{-3(k-1)/2}, \qquad \lambda^{-1} = g_2(k) v^{-1} \log \overline{T}, \qquad \eta^{-1} = g_3(k) v^{-1}. \tag{52}$$

Set $a = 3(k-1)/2$. The logarithm cancels from the recurrence factor: $\beta = 1 + 16k(\overline{T} - 1)/m \le g(k)v^{-(a+1)}$. For $0 \le j < k$, put $p_j = 3/2 + (k - 1 - j)(a + 1)$. Equation (36) gives $\zeta_j^{-1} \le g(k)v^{-p_j} \log \overline{T}$; the logarithm is unnecessary at $j = 0$. At round $j \ge 1$, the concentration bound is

$$N_j = O\left(\left[\frac{V_0}{\zeta_j^2} + \frac{C_0}{\zeta_j} \sqrt{\log \frac{64 C_0}{\zeta_j}} \right] \left(d + \log \frac{2J}{\delta'} \right) \right). \tag{53}$$

Here $C_0 = \overline{T}^{j+1}$, $V_0 = \overline{T}^{j+2}$, and $J \le g(k)v^{-(k-1)j}$. The variance term has exponent

$$A_j = a(j+2) + 2p_j = 3k^2 - k + 1 - (3k+1)j/2. \tag{54}$$

The linear term has exponent $a(j+1) + p_j < A_j$. Thus $N_j \le g(k)v^{-A_j} \log^2(2/v)(d + \log(2J/\delta'))$. The same batch serves all candidates, so only $\log J$ enters the sample count. Multiplying by J for runtime adds $(k-1)j$. Both A_j and $A_j + (k-1)j$ decrease with $j \ge 1$: their maxima are $3k^2 - 5k/2 + 1/2$ and $3k^2 - 3k/2 - 1/2$ at $j = 1$.

At $j = 0$, homogeneity gives (35) without any importance weights. Its coefficient tuple is zero and its range is in V. The first moment adds one estimate, so use $C_0 = V_0 = 1$, $J = 2$, with

$$\zeta_0^{-1} = \frac{k\beta^{k-1}}{\lambda_0\sqrt{v}} \le g(k)v^{-(3k^2-4k+4)/2}.$$

The initial sample and operation exponents are therefore at most $3k^2 - 4k + 4$. This dominates the sample exponent only at $k = 2$, where it equals 8; the operation exponent is largest at $j = 1$.

Computing and storing the projections costs $O(N_j kd)$ per round. For each candidate, scalar weights cost $g(k)N_j$, and (25) costs $O(N_j d + kd)$ per product. With $\alpha = \delta'/J$, the fixed power count is $O(\log(dJ(k+1)/\delta))$. Summing over J candidates and at most k extensions proves the d^2 operation factor. No dense $d \times d$ matrix is constructed. Random test vectors need no labels. The single final validation uses $O(\varepsilon^{-1}\log(2N_{\text{fit}}/\delta'))$ examples; evaluating all candidates costs at most $g(k)\varepsilon^{-(k^2/2+1)}d$ up to logarithmic factors. Both are smaller than the matrix bounds for $k \ge 2$. With $v = \varepsilon/128$, the total counts for $k \ge 2$ are

$$\begin{aligned} N &\le g(k)\varepsilon^{-b_k} d \ \operatorname{polylog}\frac{2d}{\varepsilon\delta}, \\ T &\le g(k)\varepsilon^{-(3k^2-3k/2-1/2)}d^2 \ \operatorname{polylog}\frac{2d}{\varepsilon\delta}. \end{aligned} \tag{55}$$

Only the shared examples and one candidate's scalar weights are stored. Here $b_k = \max\{3k^2 - 4k+4, 3k^2-5k/2+1/2\}$. The $(k-1)(2^{k+1}+1)$ scale endpoints contribute only a width-dependent factor. Linearization reduces their magnitudes.

One Neuron. Write $F = s\sigma(\langle u, x\rangle)$, with $s \in \{-1, 1\}$ and $\|u\| \le 1$, and start at $h_0 = 0$. On a nonzero label, $|Y| = \langle u, X\rangle$ and the sign determines s. Use the projection update, a randomized Kaczmarz step (Strohmer & Vershynin, 2009),

$$h_{n+1} = h_n + \mathbf{1}\{Y_n \ne 0\}\frac{|Y_n| - \langle h_n, X_n\rangle}{\|X_n\|^2}X_n, \tag{56}$$

where a zero label means no update. For $u \ne 0$, symmetry and $\mathbb{E}[XX^\top/\|X\|^2] = I_d/d$ give, for every fixed e,

$$\mathbb{E}\left[\frac{\langle e, X\rangle^2}{\|X\|^2}\mathbf{1}\{\langle u, X\rangle > 0\}\right] = \frac{\|e\|^2}{2d}.$$

Consequently, conditional on earlier examples,

$$\mathbb{E}[\|h_{n+1} - u\|^2 \mid h_n] = \left(1 - \frac{1}{2d}\right)\|h_n - u\|^2. \tag{57}$$

After $N = \lceil 2d\log(2/(\varepsilon\delta))\rceil$ examples, $\mathbb{E}\|h_N - u\|^2 \le \varepsilon\delta$. Project h_N onto the unit ball and return the ReLU with sign s. Projection cannot increase parameter error, and the ReLU Lipschitz bound followed by Markov's inequality gives risk at most ε with probability at least $1 - \delta$. If no nonzero label was seen, $h_N = 0$ and return zero; the same risk bound applies. If $F = 0$, every iterate is zero. Each update costs $O(d)$, proving the stated sample and operation bounds.

E.1 Proof of the Main Theorem

Proof of Theorem 1. For $k = 1$, use the procedure above. Assume $k \ge 2$.

Subspace error. Condition on accurate moment estimates and successful power iterations. The first direction has $e_{W_1} \le \zeta_0/(2\lambda_0)$: use $\mu \in V$ and $\gamma \ge 2\lambda_0$ for the vector, or (35) for the matrix. Equations (29) and (36) then give

$$e_{W_j} \le \frac{j\sqrt{v}}{2k\beta^{k-j}} \le \frac{\sqrt{v}}{2} < 1, \qquad 0 \le j \le \min(k, d). \tag{58}$$

This bound holds for every selected image, regardless of the residual variance. In particular, $\dim W_j \le r$: a larger space would contain a unit vector orthogonal to V.

Stopping and accuracy. If $v_W(F) > v$, (35) or Proposition 4 gives a population norm above $8\lambda_j$. Since $\zeta_j < \lambda_j$, Lemma 6 then supplies an image above $7\lambda_j/2$. Thus stopping for lack of an image gives $v_W(F) \le v$. At the dimension cap, either $W = \mathbb{R}^d$, or $\dim W = r = k$. In the latter case, equal-dimensional projection gaps and ReLU Lipschitzness give

$$v_W(F) \le \|F - F \circ P_W\|_2^2 \le \|(I - P_W)P_V\|_{\mathrm{op}}^2 = e_W^2 \le v/4 \tag{59}$$

(see Appendix C). In either case, (33) supplies a candidate of risk at most $\varepsilon/16$; the final validation returns one of risk below $\varepsilon/2$. There are at most r extensions.

Probability and sample reuse. At each round, condition on the past. The candidate list is fixed, so simultaneous moment estimation and all power iterations each fail with probability at most δ'. After the search, condition on the final W; one fresh validation batch has failure probability at most δ'. The total is at most $(2k+1)\delta' < \delta$. Samples are shared within a round. The loop cap and finite lists bound work on every outcome.

Complexity. At round $j \ge 1$, the leading sample and operation costs are proportional to $\overline{T}^{j+2}\zeta_j^{-2}$ and $v^{-(k-1)j}\overline{T}^{j+2}\zeta_j^{-2}$, respectively. Appendix E bounds these and the initial and validation costs by (2)–(3). Together with the accuracy and failure bounds above, this proves the theorem. □

F Numerical Identities and Reproducibility

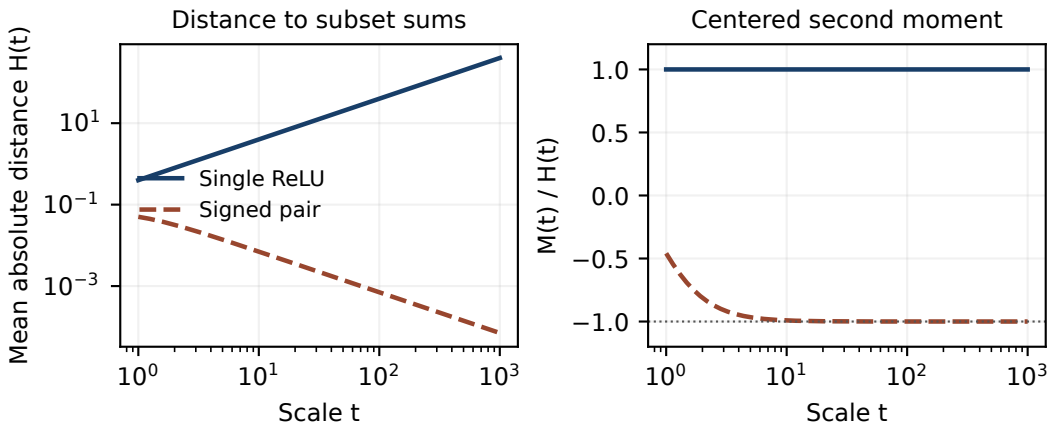


Figure 1: The same statistic detects increasing and decreasing distance. A single missing ReLU gives a positive second moment; the signed pair in (19) gives a negative one. These are population quantities.

F.1 Closed Forms for the Signed Example

For (19), write $w = R\cos\theta$, $z = R\sin\theta$. By symmetry it suffices to use the first quadrant, where the absolute residual on the unit circle is

$$h_t(\theta) = \begin{cases} \min\{t\sin\theta, \cos\theta - t\sin\theta\}/(2\sqrt{2}), & 0 \le \theta \le \arctan(1/t), \\ 0, & \arctan(1/t) < \theta \le \pi/2. \end{cases} \tag{60}$$

The switch is at $\alpha = \arctan(1/(2t))$. Since $\mathbb{E}R = \sqrt{\pi/2}$ and $\mathbb{E}R^3 = 3\mathbb{E}R$,

$$H_c(t) = \sqrt{2/\pi}\int_0^{\pi/2} h_t(\theta)\,d\theta, \qquad M_{zz}(t) = \sqrt{2/\pi}\int_0^{\pi/2} h_t(\theta)(3\sin^2\theta - 1)\,d\theta. \tag{61}$$

The subset-sum slopes are $\{0, 1/(2\sqrt{2}), 1/\sqrt{2}\}$. Integrating $\sin\theta$ and $\cos\theta$ over the two active intervals gives

$$H_c(t) = \frac{t + \sqrt{t^2+1} - \sqrt{4t^2+1}}{2\sqrt{\pi}}. \tag{62}$$

Differentiation gives

$$M_{zz}(t) = \frac{t}{2\sqrt{\pi}} \left(1 + \frac{t}{\sqrt{t^2+1}} - \frac{4t}{\sqrt{4t^2+1}} \right). \tag{63}$$

F.2 Checking the Identities

For the proper-output bound, an independent calculation uses the Gaussian ReLU kernel (Cho & Saul, 2009):

$$\mathbb{E}[\sigma(A)\sigma(B)] = \frac{\sigma_A \sigma_B}{2\pi} \left(\sqrt{1-\gamma^2} + (\pi - \arccos\gamma)\gamma \right), \tag{64}$$

where A, B are centered jointly Gaussian with correlation γ. The value is zero if either variance is zero. For the rotation $\widetilde{u}_i = LB^\top u_i$, this formula gives $\mathbb{E}F(X)G(X)$ using covariance $\langle u_i, \widetilde{u}_j \rangle$. It also gives $\mathbb{E}[\mathbb{E}(F(X) \mid P_S X)^2]$ using covariance $\langle P_S u_i, P_S u_j \rangle$ and the original marginal variances.

G Estimating Vectors in the Target Subspace

The matrices above have range in $P_{W^\perp} V$. Adding a cross moment places their range in V itself. Estimating a subspace from several nearly contained vectors is a standard approach (Dalalyan et al., 2008); here the cross moment supplies those vectors without gradient queries. We select a useful image, then estimate only that image on fresh examples. For $k \geq 2$, the resulting bounds are

$$\begin{aligned} N &\leq g(k)(\varepsilon^{-p_k} d + \varepsilon^{-q_k}) \operatorname{polylog} \frac{2d}{\varepsilon\delta}, \\ T &\leq g(k)(\varepsilon^{-s_k} d^2 + \varepsilon^{-q_k} d) \operatorname{polylog} \frac{2d}{\varepsilon\delta}, \end{aligned} \tag{65}$$

where

$$p_k = \frac{3}{2}k^2 + \frac{3}{2}, \qquad q_k = \frac{3}{2}k^2 + 3k - \frac{5}{2}, \qquad s_k = \frac{5}{2}k^2 - 2k + \frac{3}{2}. \tag{66}$$

All scale parameters are unchanged. As in Algorithm 1, fit and validate a network only after the subspace search stops.

G.1 A Cross Moment Restores the Missing Component

For any tuple c, choose a nearest subset J in $D_c(w, y)$, breaking ties in a fixed order, and put $c_J = \sum_{i \in J} c_i$ and $\theta = \operatorname{sgn}(y - \langle c_J, w \rangle)$. Write $f = D_c(w, F(w + tz))$, with $w \sim \mathcal{N}(0, P_W)$ and $z \sim \mathcal{N}(0, P)$ independent. Define

$$B_c(t) = M_c(t) + t\,\mathbb{E}[(fw + \theta c_J) z^\top] = t\,\mathbb{E}[\theta \nabla F(w + tz) z^\top]. \tag{67}$$

Indeed, almost everywhere $\nabla_w f = \theta(P_W \nabla F - c_J)$. Integration by parts in w, followed by (26), proves the equality. Ties between distinct affine pieces lie on null sets unless the pieces agree there; the weak-gradient formula also covers open regions where the distance is zero. Consequently, for every W and every tuple,

$$\operatorname{range} B_c(t) \subseteq V, \qquad P B_c(t) = M_c(t), \qquad \|B_c(t)\|_{\mathrm{op}} \leq t. \tag{68}$$

The last inequality uses $\|\nabla F\| \leq 1$ and $\mathbb{E}|\langle a, z \rangle| \leq 1$ for unit a. Use the same logarithmic average to define $\overline{B}_c(T)$.

These matrices require no extra label access. For an ordinary example, let $D = D_c(X_W, Y)$, and select J, θ at (X_W, Y). Homogeneity and the change of measure give

$$B_c(t) = \mathbb{E} R_t t \left[D(ZZ^\top - P) + (t^2 D X_W + \theta c_J) Z^\top \right]. \tag{69}$$

G.2 Select a Direction, Then Estimate Its Image

Project the initial estimate $\widehat{\mu}$ onto the ball of radius $1/2$; this cannot increase its error since $\|\mu\| \le 1/2$. With the completed coefficient net in (22), every subset satisfies $\|c_J\| \le K := 2k$.

On one batch, estimate all candidate $\overline{M}$ matrices by (23)–(24), with accuracy $\lambda/2$. Use Lemma 6, with failure probability δ'/J per candidate, to select (c, T, a) with $\|a\| = 1$ and $\left\|\widetilde{M}a\right\| \ge 3\lambda$. Then $\left\|\overline{M}_c(T)a\right\| \ge 5\lambda/2$ on the estimation event. Only this selected image is refined.

Condition on the selection and use a fresh batch. Thus (c, T, a) is fixed for concentration, a standard use of sample splitting (Wasserman & Roeder, 2009). We use an elementary vector version of median-of-means (Hsu & Sabato, 2016, Algorithm 2).

Lemma 7 (Selecting a block mean)**.** *Let independent vectors have mean ν and $\mathbb{E}\|Z_i - \nu\|^2 \le \sigma^2$. For $u > 0$ and $\delta' \in (0,1)$, take an odd number $b \ge 4\log(2/\delta')$ of disjoint blocks, each of size $n \ge \max\{1, \lceil 72\sigma^2/u^2\rceil\}$. Among their means, return one with the smallest median distance to the other means, including itself. Its error is at most u with probability at least $1 - \delta'$.*

Proof. A block mean lies within $u/3$ of ν with probability at least $7/8$, by Markov's inequality. Hoeffding's inequality makes a strict majority of the means this accurate except with probability $e^{-9b/32} \le \delta'$. Their pairwise distances are at most $2u/3$. The selected mean therefore has a ball of radius $2u/3$ containing a majority of all means. This ball contains an accurate mean, so its center lies within u of ν. □

To allow larger errors inside W, first multiply that component by

$$\alpha = \sqrt{v}/4, \qquad C_W = P + \alpha P_W. \tag{70}$$

For $t = T^\xi$, $\xi \sim \mathrm{Unif}[0,1]$, the ordinary-sample vector is

$$\Xi = R_t t\left[DK_W a + (t^2 DX_W + \theta c_J)(Z^\top a)\right], \qquad \mathbb{E}\Xi = \overline{B}_c(T)a. \tag{71}$$

Only a second-moment bound is needed:

$$\mathbb{E}\|C_W\Xi\|^2 \le \sigma_W^2 := 16\left(d\overline{T}^{\ell+2} + K^2 v\ell\overline{T}^{\ell+4}\right). \tag{72}$$

Indeed, under the measure R_t^2/c_t from Appendix B, $c_t \le t^\ell$ and $\mathbb{E}_t Y^4 \le 3/2$. For unit $z_0 \in W^\perp$ and $w_0 \in W$, Gaussian moments give

$$\mathbb{E}_t(z_0^\top K_W a)^4 \le 60, \qquad \mathbb{E}_t[(w_0^\top X_W)^4 (a^\top Z)^4] \le 9t^{-4}.$$

The first follows by polarization from $\mathbb{E}(G^2-1)^4 = 60$ for $G \sim \mathcal{N}(0,1)$; the second uses independence. Cauchy–Schwarz, $D \le |Y|$, and $\|c_J\| \le K$ now give

$$\mathbb{E}\|P\Xi\|^2 \le 10d\overline{T}^{\ell+2}, \qquad \mathbb{E}\|P_W\Xi\|^2 \le 3K^2\ell\overline{T}^{\ell+4}.$$

For the second inequality, $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ gives $(8\ell + 2K^2)\overline{T}^{\ell+4} \le 3K^2\ell\overline{T}^{\ell+4}$ when $\ell \ge 1$; the vector is zero when $\ell = 0$. Orthogonality and $\alpha^2 = v/16$ prove (72).

Apply Lemma 7 to $C_W\Xi$ with $u = \zeta/2$ and the smallest admissible odd b. It uses

$$N_{\mathrm{ref}} = b\left\lceil 288\sigma_W^2/\zeta^2\right\rceil \tag{73}$$

examples and returns $\widehat{b}$. Set

$$x = C_W^{-1}\widehat{b}, \qquad \left\|C_W(x - \overline{B}_c(T)a)\right\| \le \zeta/2 \tag{74}$$

on an event of probability at least $1 - \delta'$. Computing the block means and sorting their distance lists costs $O_k(N_{\mathrm{ref}}d + b^2(d + \log b))$ operations. All choices are finite.

G.3 Keeping Only Sufficiently Strong Directions

The following elementary construction controls the whole learned space without a recurrence.

Lemma 8. *Suppose $A = A_0 + E$ has n columns,* $\operatorname{range} A_0 \subseteq V$*, and* $\|E\|_F \le \sqrt{n}\,\zeta$*. For any $\tau > 0$, one can construct a subspace W such that*

$$\|(I - P_W)A\|_{\mathrm{op}} \le \sqrt{3/2}\,\tau, \qquad e_W \le \frac{\sqrt{2n}\,\zeta}{\tau}. \tag{75}$$

If the latter bound is below one, then $\dim W \le \dim V$.

Proof. Use plane rotations on $A^\top A$ to find an orthogonal O with

$$O^\top A^\top A O = D + R, \qquad D \text{ diagonal}, \quad \|R\|_{\mathrm{op}} \le \|R\|_F \le \tau^2/2.$$

Let $I = \{i : D_{ii} \ge \tau^2\}$ and $C = AO_I$, and take $W = \operatorname{range} C$. If I is empty, take $W = \{0\}$. Otherwise $C^\top C \succeq (\tau^2/2)I$, while $\|(I - P_V)C\|_{\mathrm{op}} \le \|E\|_{\mathrm{op}}$. This proves the second bound. The discarded columns have Gram matrix at most $(3\tau^2/2)I$, which proves the first. Finally, any space of dimension exceeding $\dim V$ contains a unit vector orthogonal to V. □

This uses finite Jacobi rotations (van de Geijn & Myers, 2026, Section 11.3). Zeroing a largest off-diagonal entry decreases the squared off-diagonal Frobenius norm by a factor at most $1-2/[n(n-1)]$ when $n > 1$. Thus $O(n^2 \log(2+n \max_i \|A_i\|^2/\tau^2))$ rotations suffice. Form the small Gram matrix, accumulate the rotations, and orthonormalize the retained columns. Even finding each largest entry by exhaustive comparison costs only a polynomial in n and the displayed logarithm, besides $O(dn^2)$ operations on the columns.

G.4 Parameters, Iteration, and Termination

Keep $v = \varepsilon/128$ and the scale parameters (13)–(14). Set

$$\tau = \lambda/8, \qquad n_* = \left\lceil 8k \log\left(2 + \frac{32k(\overline{T}+1)^2}{\tau^2}\right)\right\rceil,$$
$$\zeta = \frac{\tau\sqrt{v}}{16\sqrt{n_*+1}}, \qquad \zeta_\mu = \min\left\{\frac{\eta}{4}, \frac{\zeta\sqrt{v}}{64\tau}\right\}. \tag{76}$$

Estimate μ to error ζ_μ by Lemma 7, using $\mathbb{E}\,\|YX\|^2 \le 3d$, and project it onto its norm ball. Start the column matrix at $A = [x_0]$, where $x_0 = (64\tau/\sqrt{v})\widehat{\mu}$. Its distance to V is at most ζ, and $\|x_0\| < 1$. At each round perform the following steps.

1. Construct W from A by Lemma 8. Stop the search if $\dim W = \min(k, d)$.
2. Otherwise, use a fresh batch to select (c, T, a) with $\left\|\widetilde{M}a\right\| \ge 3\lambda$, as above. Stop the search if no such image is found.
3. On another fresh batch, estimate the weighted mean and append x from (74) to A.

At either normal stopping condition, extend W to S_W, and use one fresh batch to validate the proper-output net. Return an accepted candidate, or zero if none is accepted. Return zero on any of the following exceptional outcomes: n_* additions are reached, $\|x\| > \overline{T} + 1$, or a constructed space has dimension above $\min(k, d)$. These caps bound work on every outcome. Use $\delta' = \delta/[4(n_*+2)]$ for the first-moment estimate, for each round's coarse estimation, power iterations and refinement, and for the single final validation.

On their joint success event, induction shows that every column has distance at most ζ to V. The initial column has this property. If previous columns do, Lemma 8 gives

$$\dim W \le r, \qquad e_W \le \frac{\sqrt{v}}{8\sqrt{2}} < \frac{\sqrt{v}}{8}. \tag{77}$$

Since $\overline{B}_c(T)a \in V$ and $C_W^{-1} = P + \alpha^{-1}P_W$,

$$\|(I - P_V)x\| \le (1 + e_W/\alpha)\left\|C_W(x - \overline{B}_c(T)a)\right\| \le 3\zeta/4. \tag{78}$$

This closes the induction, with no amplification from round to round. Also $\left\|x - \overline{B}_c(T)a\right\| \le \zeta/(2\alpha) < 1$, so the norm cap is respected. The first column and the first bound in (75) give

$$\|P_{W^\perp}\mu\| \le \frac{\sqrt{v}}{64\tau}\|P_{W^\perp}x_0\| + \zeta_\mu \le \frac{\sqrt{v}}{32} + \zeta_\mu \le \frac{\sqrt{v}}{16}.$$

The first-moment condition therefore survives even if W changes. Coefficient completion remains valid since $\zeta_\mu \le \eta/4$. Whenever $v_W(F) > v$, Proposition 4 and Lemma 6 supply a coarse image above 3λ. Thus stopping for lack of an image gives $v_W(F) \le v$. Stopping at dimension $\min(k, d)$ gives the same bound by (59), or by $W = \mathbb{R}^d$. Every round that estimates matrices has $\ell < \min(k, d) \le k$.

To bound the number of rounds, put $S = AA^\top$ and $\Phi(A) = \log\det(I + S/\tau^2)$. The determinant identity used in standard log-determinant arguments (Abbasi-Yadkori et al., 2011, Lemma 11) gives

$$\Phi([A, x]) - \Phi(A) = \log(1 + x^\top(\tau^2 I + S)^{-1}x). \tag{79}$$

Here $\|P_{W^\perp}x\| \ge 5\lambda/2 - \zeta/2 \ge 2\lambda$. Cauchy–Schwarz and (75) give

$$x^\top(\tau^2 I + S)^{-1}x \ge \frac{\|Px\|^2}{\tau^2 + \|PA\|_{\mathrm{op}}^2} \ge \frac{4\lambda^2}{(5/2)\tau^2} > 1.$$

Each addition increases Φ by more than $\log 2$. Conversely, write $A_0 = P_V A$ and $E = (I - P_V)A$. Then $\operatorname{range} A_0 \subseteq V$, its column norms are at most $\overline{T} + 1$, and $\|E\|_F^2 \le n\zeta^2$. Use $A^\top A = A_0^\top A_0 + E^\top E$ and $\det(I + AA^\top/\tau^2) = \det(I + A^\top A/\tau^2)$. Since $\log\det(I + C + D) \le \log\det(I + C) + \operatorname{tr} D$ for $C, D \succeq 0$,

$$\Phi(A) \le k\log\left(1 + \frac{n(\overline{T} + 1)^2}{\tau^2}\right) + \frac{n\zeta^2}{\tau^2}. \tag{80}$$

Let $L_* = \log(2 + 32k(\overline{T} + 1)^2/\tau^2)$. For $n \le n_* + 1 \le 9kL_*$ the last bound is at most $2kL_* + 1$, whereas $n_* \log 2 \ge 8kL_* \log 2 > 2kL_* + 1$. Thus n_* additions are impossible on the success event. The search stops normally with $v_W(F) \le v$, and the final validation returns a proper network of risk below $\varepsilon/2$. There are at most $4(n_* + 2)$ conditional failure events, including the initial estimate and final validation; their union has probability at most δ.

G.5 Cost and Comparison

For fixed k, $n_* = O_k(\log(2/v))$ and $\zeta^{-2} = g(k)v^{-3}\operatorname{polylog}(2/v)$. Ignoring k factors and logarithms, at dimension $\ell \le k - 1$ the selection and refinement sample bounds are respectively

$$d\overline{T}^{\ell+2}v^{-2}, \qquad d\overline{T}^{\ell+2}v^{-3} + \overline{T}^{\ell+4}v^{-2}.$$

Their operation bounds multiply these by Jd and d, respectively, where $J \le g(k)v^{-(k-1)\ell}$. The distance comparisons add only $d\operatorname{polylog}(1/\delta')$. Substitute $\overline{T} = g(k)v^{-3(k-1)/2}$ and $\ell = k - 1$ to obtain (66); $s_k - p_k = k(k - 2) \ge 0$. The first moment, column selection, extra rounds, and output validation are smaller. This proves (65).

Since $q_k - p_k = 3k - 4 > 0$, we also have

$$N \le g(k)\varepsilon^{-q_k}\widetilde{O}(d), \qquad T \le g(k)\varepsilon^{-\max(q_k, s_k)}\widetilde{O}(d^2).$$

At $k = 3$, both exponents are 20, preserving the sample bound and improving the operation bound in Appendix E. For $k \ge 4$, both improve and $s_k - q_k = (k - 1)(k - 4) \ge 0$. Use Algorithm 1 at $k = 2$ and the earlier one-neuron procedure at $k = 1$.