WHEN IS LEARNING A PLANTED RELU NETWORK TRACTABLE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning neural networks from data is a fundamental problem, yet the tractability of even one-hidden-layer networks remains poorly understood. Worst-case hardness is known when the output-layer weights may have arbitrary signs, but a basic and surprisingly unresolved case remains: can one efficiently learn a sum of ReLU units with positive output weights, even under standard Gaussian covariates? We study recovery of the hidden weights in planted networks of the form $f_W(x) = \sum_{\ell=1}^{k} \mathrm{ReLU}(\langle w_\ell, x\rangle)$, $x \sim \mathcal{N}(0, I_d)$. We first identify when the parameters are statistically determined, up to the natural permutation symmetry. In particular, Gaussian inputs allow the identification of the hidden weights under mild nondegeneracy conditions: no two weights are collinear, and no nonempty subset of the weights sums to zero. However, we show that identifiability alone is far from enough for efficient learning. Even under bounded norms, angular separation, and a strict positive subset-sum margin, any randomized polynomial-time algorithm that exactly recovers the parameters from membership queries at rational points would imply a randomized polynomial-time algorithm for an NP-hard vector subset-sum problem; hence no such algorithm exists unless $NP \subseteq RP$. This hardness result reveals the central obstruction: Gaussian even-order moments can recover the hidden hyperplanes, but orienting those hyperplanes can encode a hard discrete problem. On the positive side, we identify natural geometric conditions that make the orientation barrier tractable. Specifically, we show that a continuous analogue of the subset-sum margin, in the form of a positive-independence condition, enables, for any fixed angular-separation constant, exact recovery from polynomially many samples in polynomial time, given an exact solver for a polynomial-size linear program. Our method first combines implicit Gaussian Hermite moment estimation with random polynomial features and a low-dimensional matrix pencil to recover the network weights approximately up to sign. It then uses convex sign orientation based on the first moment to identify the correct signs, followed by anchored regression to refine the estimates to exact recovery.

1 INTRODUCTION

Neural networks are the basic computational engine behind modern Artificial Intelligence (AI). Yet our theoretical understanding of even the simplest nonlinear neural networks remains surprisingly incomplete. A particularly fundamental model is the one-hidden-layer ReLU network

$$f(x) = \sum_{\ell=1}^{k} v_\ell \mathrm{ReLU}(w_\ell^\top x),$$

where the input distribution is known and the labels are generated by a planted set of hidden weights. Despite the apparent simplicity of this model, learning its parameters can be computationally intractable. Classical results show that training small neural networks is NP-hard in the worst case (Blum & Rivest, 1988). More recent work shows that these computational obstructions do not disappear merely by using random Gaussian covariates, planted models, or overparameterized training. In fact, arbitrary output signs can lead to strong lower bounds for learning one-hidden-layer ReLU networks under Gaussian marginals (Diakonikolas et al., 2020; Goel et al., 2020).

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.

These hardness results leave open a basic and natural question. If all output weights are positive, then the model becomes a sum of ReLU functions rather than a signed combination of them. Since ReLU is positively homogeneous, the positive coefficients can be absorbed into the hidden weights, reducing the problem to learning

$$f_W(x_i) = \sum_{\ell=1}^{k} \mathrm{ReLU}(w_\ell^\top x_i) \tag{1}$$

from n independent samples (x_i, y_i), where $x_i \sim \mathcal{N}(0, I_d)$. This removes the most obvious source of cancellation created by arbitrary signs. Nevertheless, the computational status of this positive model is not well understood. Positive results are known in special or limiting regimes, for instance under full-rank and width-at-most-dimension assumptions, specially designed losses, vector-valued outputs, or other structural conditions (Ge et al., 2018; Bakshi et al., 2019). More recent algorithms also obtain learnability guarantees for restricted parameter regimes such as small width (Chen et al., 2023). However, a general characterization of when planted positive ReLU networks can be recovered in polynomial time and from polynomially many Gaussian samples remains open.

This leads to the central question of this paper:

> *When is exact ecovery of the hidden weights in a planted one-hidden-layer ReLU network with positive output weights computationally and statistically tractable?*

We study this question from both the statistical and computational perspectives. Our goal is not only to give another sufficient condition for recovery, but to isolate the obstruction that separates identifiability from efficient learning. Our contributions are as follows.

- **Statistical identifiability under mild nondegeneracy.** We first characterize when the planted weights are statistically determined by the label distribution. We show that, under Gaussian covariates, the hidden weights are identifiable up to permutation under two mild nondegeneracy conditions: no two weights are collinear, and no nonempty subset of the weights sums to zero. Thus, statistically, Gaussian data contain enough information to determine the hidden weights.

- **Hardness despite identifiability.** We next show that identifiability is far from sufficient for efficient learning. Even if the above nondegeneracy conditions are robustified by bounded norms, angular separation, and a strict positive subset-sum margin, parameter recovery remains computationally hard. In particular, we prove that a randomized polynomial-time algorithm for parameter recovery under these assumptions would imply a randomized polynomial-time algorithm for an NP-hard vector subset-sum problem. This establishes that the difficulty is not merely statistical or numerical.

- **A tractable regime via positive independence.** Finally, we identify a natural geometric regime in which learning becomes tractable. The condition is a continuous analogue of the discrete subset-sum margin: all positive linear combinations of the planted weights must remain quantitatively bounded away from zero, equivalently the weights lie in a common open half-space with margin. Under this positive-independence condition, we give a polynomial-time and polynomial-sample recovery algorithm. The algorithm first combines implicit Gaussian Hermite moment estimation with random polynomial features and a low-dimensional matrix pencil to recover the hidden directions approximately up to sign. Then it uses the first moment and a convex box relaxation to orient the signs; finally, an anchored regression step refines the approximate estimates to recover the planted weights.

Our results give a sharper conceptual picture of when learning planted networks are hard or tractable with random data. Gaussian inputs provide rich moment information about the planted network, and even-order moments can reveal the hidden directions of the weights. However, recovering directions is not the same as recovering oriented weights: each recovered direction comes with an unknown sign. We show that this seemingly small orientation ambiguity is the core computational barrier. Identifying this orientation ambiguity as the central barrier also points to the right minimal algorithmic assumptions, allowing us to obtain tractable polynomial-time recovery algorithms.

2 IDENTIFIABILITY

Before addressing the core question of computational tractability, we must first establish a rigorous baseline for statistical learnability: under what conditions are the parameters uniquely determined by the population data distribution? For networks with mixed output signs, cancellation effects can completely destroy identifiability. In our setting, where output weights are strictly positive, we show that remarkably simple, mild nondegeneracy conditions suffice to guarantee global identifiability up to the natural permutation symmetry of the hidden units. We formalize these geometric configurations through the following two assumptions.

Assumption 2.1 (Non-collinearity)**.** *Within the hidden layer representation, no two weight vectors point along the same or opposite lines. That is, $w_\ell \neq c \cdot w_j$ for all $\ell \neq j$ and any scalar $c \in \mathbb{R}$.*

Assumption 2.2 (No subset cancellation)**.** *No nonempty subset of the planted weights sums to the zero vector: $\sum_{\ell \in S} w_\ell \neq 0$ for every nonempty subset $S \subseteq [k]$.*

Assumption 2.1 prevents trivial redundancies where parallel hidden neurons could be merged or reshaped arbitrarily, while Assumption 2.2 rules out structural discrete dependency loops that could mask the network's true width. Together, these assumptions establish that the hidden units form a distinct, well-posed geometric constellation. Our first main result demonstrates that under these minimal constraints, the standard Gaussian distribution provides sufficiently rich information to fully resolve the parameters.

Theorem 2.3 (Global Identifiability)**.** *Let $f_W(x) = \sum_{\ell=1}^{k} \mathrm{ReLU}(w_\ell^\top x)$ and $f_U(x) = \sum_{\ell=1}^{k} \mathrm{ReLU}(u_\ell^\top x)$, where all hidden weights $w_\ell, u_\ell \in \mathbb{R}^d$ are nonzero. Assume both networks satisfy the Non-collinearity (Assumption 2.1) and No subset cancellation (Assumption 2.2) properties. If their network outputs match almost everywhere under Gaussian covariates, namely*

$$\mathbb{E}_{x \sim \mathcal{N}(0, I_d)} \left[(f_W(x) - f_U(x))^2 \right] = 0, \tag{2}$$

then the underlying parameters are identical up to ordering: there exists a permutation π of $[k]$ such that $u_{\pi(\ell)} = w_\ell$ for all $\ell \in [k]$.

This identifiability conclusion is a specialization of the more general result of Stock & Gribonval (2023). We include a Hermite-moment proof in our setting to expose the orientation ambiguity underlying our computational results.

Theorem 2.3 provides an encouraging statistical baseline: it guarantees that the population distribution contains a uniquely defined target for our learners, removing any ambiguity under infinite data. However, this leaves open the critical question of algorithmic tractability. As we will see in the next section, passing from information-theoretic uniqueness to a polynomial-time recovery algorithm introduces an entirely separate layer of discrete complexity.

3 HARDNESS OF LEARNING UNDER SUBSET-SUM MARGIN ASSUMPTIONS

While Gaussian inputs provide a firm foundation for global identifiability, mathematical uniqueness does not automatically translate to algorithmic tractability. To test the true limits of this learning problem, we analyze a quantitative, robustly scaled variant of the no-subset-cancellation condition. Specifically, we enforce a strict geometric separation where all hidden weight norms are bounded, their angular paths are explicitly well-separated, and their discrete subset combinations are strictly bounded away from zero. Intuitively, these constraints rule out all clear sources of numerical instability or near-cancellation, creating an idealized environment for parameter recovery.

We formalize this robust parameter regime using the following three structural criteria.

Assumption 3.1 (Norm bounds)**.** *There exists a constant $r > 0$ such that the planted hidden weights satisfy*

$$r \leq \|w_\ell\|_2 \leq 1 \qquad \text{for all } \ell \in [k]. \tag{3}$$

Assumption 3.2 (Robust non-collinearity)**.** *There exists an angular separation constant $\rho \in (0, 1)$ preventing hidden directions from mirroring or collapsing into one another:*

$$\left| \frac{\langle w_\ell, w_j \rangle}{\|w_\ell\|_2 \, \|w_j\|_2} \right| \leq \rho \qquad \text{for all } \ell \neq j. \tag{4}$$

Assumption 3.3 (Positive subset-sum margin)**.** *The planted hidden weights maintain a strict discrete subset-sum margin $\eta > 0$, bounded away from the origin:*

$$\min_{\emptyset \neq S \subseteq [k]} \left\| \sum_{\ell \in S} w_\ell \right\|_2 \geq \eta. \tag{5}$$

By enforcing these strict geometric boundaries, we guarantee robust statistical identifiability under finite sample sizes. Surprisingly, however, this quantitative margin is completely insufficient for computational efficiency.

To formally state our computational lower bounds, we consider two distinct algorithmic access models that capture varying degrees of learner capability:

Adaptive membership queries. The learner can evaluate the target network at any chosen query point $x \in \mathbb{R}^d$—possibly choosing queries sequentially based on previous responses—and receives the value $f_W(x)$. This represents an idealized, highly powerful active learning setting.

Gaussian samples or moments. The learner passively observes random pairs (X, Y), where $X \sim \mathcal{N}(0, I_d)$ and $Y = f_W(X)$, or equivalently enjoys direct access to the population moments. This matches our baseline model for statistical identifiability.

Proving a hardness result for the adaptive query framework yields a strong lower bound: it automatically rules out any passive Gaussian learner, as passive sampling can be perfectly simulated within an adaptive membership pipeline. Our main computational contribution demonstrates that even within this highly regularized, well-separated parameter space, parameter recovery remains fundamentally intractable.

Theorem 3.4 (Adaptive membership-query hardness)**.** *Assume there exists a randomized polynomial-time adaptive membership-query algorithm that, for every homogeneous positive ReLU network*

$$f_W(x) = \sum_{\ell=1}^{k} \mathrm{ReLU}(\langle w_\ell, x \rangle) \tag{6}$$

whose hidden weights satisfy Assumption 3.1 with $r = 1/12$, Assumption 3.2 with $\rho = 1/2$, and Assumption 3.3 with $\eta = 1/12$, recovers the unordered parameter set $\{w_1, \ldots, w_k\}$ with probability at least $2/3$. If the algorithm can choose rational query points adaptively and receive noise-free labels, then NP $\subseteq$ RP.

As a direct consequence of this active-learning barrier, the standard passive learning paradigm inherits the same computational barrier.

Corollary 3.5 (Passive hardness in the Gaussian real-arithmetic model)**.** *Suppose a randomized algorithm in the Gaussian real-arithmetic model (see Definition B.2 in Appendix B.5) uses polynomially many examples and operations and exactly recovers, with probability at least $2/3$, every planted network satisfying Assumption 3.1 with $r = 1/12$, Assumption 3.2 with $\rho = 1/2$, and Assumption 3.3 with $\eta = 1/12$. Then X3C has a polynomial-operation randomized algorithm with one-sided error in that same model. This statement does not assert* NP $\subseteq$ RP *in the ordinary finite-bit Turing model.*

This hardness landscape exposes a deeper obstruction: the core difficulty is neither statistical nor numerical, but strictly structural. While even-order Gaussian moments provide clear information to recover the lines of the hidden hyperplanes, orienting those hyperplanes correctly requires navigating an adversarial discrete space. The subset-sum margin successfully keeps the problem numerically stable, but it fails to eliminate this underlying combinatorial complexity, leaving the orientation step as a fundamental barrier to efficient learning.

4 Tractable Learning Under Positive Independence

To overcome the discrete orientation barrier revealed in the previous section, we must transition from a discrete subset-sum constraint to a continuous, space-spanning geometric framework. While

a discrete margin proves insufficient for tractability, we show that extending this margin requirement to encompass all continuous, positive linear combinations completely eliminates the hard sign-combinatorics. This shift enables us to bridge the gap between structural identifiability and efficient parameter recovery.

In this tractable regime, we assume that the planted weights satisfy the norm bounds of Assumption 3.1 and the robust non-collinearity condition of Assumption 3.2. Crucially, we also replace the discrete subset-sum margin with a continuous analogue.

Assumption 4.1 (Positive independence)**.** *There exists a quantitative margin constant $\eta > 0$ such that for all non-negative coefficients $\alpha_\ell \geq 0$, the hidden weights satisfy*

$$\left\| \sum_{\ell=1}^{k} \alpha_\ell w_\ell \right\|_2 \geq \eta \sum_{\ell=1}^{k} \alpha_\ell \left\| w_\ell \right\|_2 , \tag{7}$$

which is geometrically equivalent to stating that all weights w_ℓ reside within a common open half-space and preserve a margin $\langle w_\ell, q \rangle \geq \eta \left\| w_\ell \right\|_2$ along a shared reference direction $q \in \mathbb{R}^d$ with $\left\| q \right\|_2 = 1$.

Under these natural geometric conditions, exact learning becomes provably tractable. To achieve this, we introduce Algorithm Ω introduced in Table 1.

Throughout, $\otimes$ denotes the tensor (outer) product. For $g \sim \mathcal{N}(0,1)$, let h_t denote the probabilists' Hermite polynomial of degree t, and define $c_t := \mathbb{E}[\mathrm{ReLU}(g) h_t(g)]$. Let $\mathcal{H}_t(x)$ denote the corresponding multivariate Hermite tensor, characterized by $\langle \mathcal{H}_t(x), v^{\otimes t} \rangle = h_t(v^\top x)$ for every unit vector v. For each $\ell \in [k]$, we use $r_\ell = \left\| w_\ell \right\|_2$ and $u_\ell = w_\ell / r_\ell$ to denote the norm and directions of the weights. For every even t with $c_t \neq 0$, the normalized Hermite moment satisfies

$$M_t := \frac{1}{c_t} \mathbb{E}[y \mathcal{H}_t(x)] = \sum_{\ell=1}^{k} r_\ell u_\ell^{\otimes t}.$$

Overview of the Interlocking Steps:

- **Unoriented Atom Recovery (Steps 1 & 2):** The basic idea in these two steps is that Gaussian Hermite polynomials let us estimate the moment tensors of the hidden directions directly from samples. Indeed, for any test tensor $\mathcal{T}$, the quantity $\langle \mathcal{T}, M_t \rangle$ is the expectation of the observable random variable $c_t^{-1} y \langle \mathcal{T}, \mathcal{H}_t(x) \rangle$, and can therefore be estimated by empirical averaging. The advantage of going to higher even orders is that the hidden directions become increasingly separated: after lifting u_ℓ to $u_\ell^{\otimes m}$, the inner product between two distinct components is $(u_i^\top u_j)^m$. Thus, once m is sufficiently large, the lifted directions are nearly orthogonal even when the original directions are not.

  The difficulty is that explicitly forming M_m would require working in a space whose dimension grows exponentially with m. We avoid this using the implicit moment method of Diakonikolas & Kane (2025), which lets us estimate only the particular contractions of the high-order moments that we need. Random polynomial features provide a compact representation of the lifted directions. In this representation, a random symmetric weighting of the last two indices of M_{2m+2} gives a small matrix pencil whose eigenvectors separate the hidden components. Repeating the same construction with the coordinate matrices $H^{ij} = \frac{1}{2}(e_i e_j^\top + e_j e_i^\top)$ for $(i < j)$ reveals the entries of $u_\ell u_\ell^\top$, and hence u_ℓ up to sign, while the lower-order moment M_m determines the corresponding magnitude r_ℓ. The exact construction and finite-sample analysis are given in Appendix C.
- **Sign Orientation via Convex Relaxation (Steps 3 & 4):** The true population first-order moment matches the cumulative direction vector $m_1 = 2\mathbb{E}[yx] = \sum_{\ell=1}^{k} w_\ell$. To orient our unoriented base elements, we must identify an assignment of signs $s_\ell \in \{-1, +1\}$ that balances to match this overall direction. While finding a binary vector is typically NP-hard, relaxing the problem to the continuous infinity-norm box $z \in [-1, 1]^k$ yields an efficient box-constrained least-squares problem. Crucially, the positive independence condition (Assumption 4.1) guarantees that this box relaxation has the same sign as the discrete target signs.
- **Anchored Regression Clean-Up (Step 5):** Spectral tensor techniques and sample averages naturally carry finite-sample residual noise. To push our approximations to a precise parameter match,

Algorithm Ω: Orientation alignment from MomEnts and Geometric Anchoring (OMEGA)

Input: Samples $(x_i, y_i)_{i=1}^n$, where $x_i \sim \mathcal{N}(0, I_d)$ and $y_i = \sum_{\ell=1}^k \mathrm{ReLU}(\langle w_\ell, x_i \rangle)$. n is partitioned into n_T, n_o, and n_A samples for each of the steps below.

Step 1 **Implicit Higher-Order Moment Estimation:** Choose an even lifting degree m. For $t \in \{m, 2m, 2m+2\}$, implicitly approximate the population Hermite moment

$$M_t = \frac{1}{c_t}\mathbb{E}[y\mathcal{H}_t(x)] = \sum_{\ell=1}^k r_\ell u_\ell^{\otimes t}$$

from the samples by evaluating, for each required test tensor $\mathcal{T}$,

$$\left\langle \mathcal{T}, \widehat{M}_t \right\rangle = \frac{1}{n_\mathcal{T} c_t} \sum_{i=1}^{n_\mathcal{T}} y_i \left\langle \mathcal{T}, \mathcal{H}_t(x_i) \right\rangle .$$

Using the implicit moment procedure of Diakonikolas & Kane (2025), evaluate the random polynomial-feature contractions of $\widehat{M}_m$, $\widehat{M}_{2m}$, and $\widehat{M}_{2m+2}$ needed in Step 2, without explicitly forming these high-order tensors. For $\widehat{M}_{2m+2}$, also contract the final two indices with an independent isotropic Gaussian symmetric matrix $H_\star$ and with the symmetric coordinate matrices

$$H^{ii} = e_i e_i^\top, \qquad H^{ij} = \frac{1}{2}(e_i e_j^\top + e_j e_i^\top) \quad (i < j).$$

Step 2 **Unoriented Atom Recovery:** Use the random-feature contractions of $\widehat{M}_{2m}$ and the $H_\star$-weighted contractions of $\widehat{M}_{2m+2}$ to form a low-dimensional matrix pencil that separates the hidden components. Use the contractions weighted by H^{ij} to recover $\widehat{U}_\ell \approx u_\ell u_\ell^\top$, and let $\widehat{u}_\ell$ be its leading eigenvector. Use the lower-order moment $\widehat{M}_m$ to estimate the corresponding magnitude $\widehat{r}_\ell$, and output $\widehat{v}_\ell = \widehat{r}_\ell \widehat{u}_\ell, \widehat{V} = [\widehat{v}_1, \ldots, \widehat{v}_k]$.

Step 3 **First Moment Estimation:** Compute the empirical first-order moment over a batch of n_o samples:

$$\widehat{m}_1 = \frac{2}{n_o} \sum_{i=1}^{n_o} y_i x_i.$$

Step 4 **Sign Orientation:** Resolve the hidden combinatorial sign patterns by solving the globally convex ℓ_∞-relaxed least-squares program:

$$\widehat{z} \in \arg\min_{\|z\|_\infty \le 1} \left\| \widehat{V} z - \widehat{m}_1 \right\|_2 .$$

Project and round the signs to form the initial oriented parameters $\widehat{w}_\ell = \mathrm{sign}(\widehat{z}_\ell)\widehat{r}_\ell \widehat{u}_\ell$.

Step 5 **Anchored Regression Clean-Up:** Gather a final independent batch of n_A samples and establish the anchor matrix $A = [\widehat{w}_1, \ldots, \widehat{w}_k]$ and normalized version $B = [\frac{\widehat{w}_1}{\|\widehat{w}_1\|_2}, \ldots, \frac{\widehat{w}_k}{\|\widehat{w}_k\|_2}]$. Solve the convex linear programming refinement model to yield the parameters:

$$\max_{\overline{W} \in \mathbb{R}^{d \times k}} \left\langle B, \overline{W} \right\rangle \quad \text{subject to} \quad \sum_{\ell=1}^k \mathrm{ReLU}(\overline{w}_\ell^\top x_i) \le y_i \quad \forall i \in [n_A].$$

Table 1: Overview of our Ω Algorithm.

we leverage the initial estimates as a geometric anchor matrix A to guide a globally convex optimization formulation (Bahmani & Romberg, 2019). Due to the sub-linear, positively homogeneous architecture of positive ReLUs, this anchored optimization problem can be solved via a linear program. It cleans up any remaining finite-sample noise by maximizing alignment with the anchor subject to strict pointwise sample boundaries, converging directly to the planted weights w_ℓ.

With these three algorithmic stages—implicit moment recovery, convex sign synchronization, and anchored regression—in place, we can now state our main positive result. Robust angular separation makes the unoriented recovery stage tractable, while positive independence ensures that the sign-synchronization step avoids the combinatorial obstruction highlighted by our lower bounds. The final anchored-regression stage then upgrades the resulting approximate oriented estimates to exact recovery.

Theorem 4.2 (Informal: Exact Recovery via Algorithm Ω)**.** *Suppose the planted weights of a positive ReLU network satisfy the norm bounds (Assumption 3.1), robust non-collinearity (Assumption 3.2 with angular-separation constant ρ), and positive independence (Assumption 4.1 with mar-*

gin η). Then, with high probability, Algorithm Ω achieves exact parameter recovery of the hidden parameters $\{w_1, \ldots, w_k\}$. For every fixed angular-separation constant $\rho < 1$, the overall procedure runs in polynomial time and requires a number of samples polynomial in d, k, $1/r$, $1/\eta$. The precise finite-sample statement is given in Appendix C.

This theorem provides an operational counterweight to the cryptographic and combinatorial hardness barriers that usually dominate the shallow network literature. By showing that a stable continuous margin (η) allows an ℓ_∞-relaxation to find true discrete signs, Theorem 4.2 pinpoints the precise geometric property needed to bypass the orientation obstruction. We defer the precise statement of the theorem above to Theorem C.1 in the appendix.

5 Related Work

Hardness and Computational Complexity. Worst-case computational hardness for training even very small neural networks goes back to Blum & Rivest (1988), who showed that training a 3-node network of linear threshold units is NP-complete. Under Gaussian inputs, Diakonikolas et al. (2020) establish correlational statistical query (CSQ) lower bounds for PAC learning one-hidden-layer ReLU networks with arbitrary real output coefficients, while Goel et al. (2020) prove related superpolynomial lower bounds for gradient descent in regression and SQ algorithms in classification. More recently, Li et al. (2025) use the Continuous Learning with Errors (CLWE) assumption to establish cryptographic hardness for learning polynomial-size one-hidden-layer ReLU networks under standard Gaussian inputs and polynomially small Gaussian noise. These function-learning lower bounds concern different output-sign, noise, and access models from our noiseless positive-output parameter-recovery problem. In the overparameterized regime, convergence to an interpolating solution does not itself guarantee recovery of the planted parameters (Oymak & Soltanolkotabi, 2019).

Multi-Index Models and Tensor Decomposition. Method-of-moments and tensor techniques have long been used for latent-variable and shallow-network learning (Anandkumar et al., 2014; Janzamin et al., 2015). Classical index-model results include the Gaussian single-index analysis of Brillinger (1982), while Yang et al. (2017) develop second-order Stein methods for high-dimensional multi-index models. In overcomplete tensor decomposition, fourth-order blind-identification methods such as FOOBI can recover underdetermined mixtures with a number of components roughly quadratic in the observation dimension, under suitable identifiability conditions (De Lathauwer et al., 2007). Recent work by Diakonikolas and collaborators has substantially broadened the algorithmic theory of multi-index models. Diakonikolas et al. (2024) give query-based algorithms for agnostically learning broad classes of MIMs and, as a special case, resolve Feldman's question on agnostically learning a halfspace with queries. Diakonikolas et al. (2025) develop iterative subspace approximation methods for robustly learning finite-range MIMs.

PAC Learning under Gaussian Inputs. For one-hidden-layer ReLU networks with general output signs, Chen & Narayanan (2024) and Diakonikolas & Kane (2024) obtain PAC-learning algorithms with runtimes $(dkR/\varepsilon)^{O(k^2)}$ and $(dk/\varepsilon)^{O(k)}$, respectively, where R bounds the total absolute output weight for unit hidden directions in the former result and the latter uses a normalized target class. These results concern learning the network function rather than recovering the planted parameters: Chen and Narayanan output a one-hidden-layer ReLU hypothesis with small L_2 error, while Diakonikolas and Kane allow an improper hypothesis. Their runtimes are polynomial in d and $1/\varepsilon$ for fixed k, but not polynomial jointly in k, d, and $1/\varepsilon$.

For positive output coefficients, Diakonikolas et al. (2020) give a proper PAC learner with runtime $\mathrm{poly}(d/\varepsilon) + (k/\varepsilon)^{O(k^2)}$, without rank or conditioning assumptions on the hidden weights, and Diakonikolas & Kane (2020) improve this to $\mathrm{poly}(dk/\varepsilon) + (k/\varepsilon)^{O(\log^2 k)}$ using an improper hypothesis. Both results measure error relative to the target and noise scales. More recently, Diakonikolas & Kane (2025) obtain a $\mathrm{poly}(d,k)2^{\mathrm{poly}(1/\varepsilon)}$-time learner for unit hidden directions and nonnegative output coefficients summing to one, giving polynomial dependence on d, k at fixed normalized accuracy. For our parameterization, this normalization applies to f_W/R, where $R = \sum_\ell \|w_\ell\|_2$; absolute L_2 error ε for f_W therefore requires normalized error ε/R. These results do not provide a runtime jointly polynomial in all parameters, and their function-approximation guarantees do not identify the particular planted representation.

Gradient-Based Optimization. The optimization dynamics of neural networks have been extensively studied. In the infinite-width regime, Jacot et al. (2018) characterize neural-network training through the Neural Tangent Kernel (NTK). Related work establishes global convergence of gradient methods for sufficiently overparameterized networks under various assumptions (Du et al., 2019; Allen-Zhu et al., 2019; Soltanolkotabi et al., 2019). For parameter recovery in teacher/student models, Zhong et al. (2017) show that tensor-based initialization followed by gradient descent recovers the planted parameters for smooth homogeneous activations under undercompleteness and conditioning assumptions, while Zhang et al. (2019) prove linear convergence up to statistical error for multi-neuron ReLU networks via tensor initialization followed by gradient descent under suitable conditioning assumptions. Other work studies global convergence for a single ReLU (Soltanolkotabi, 2017), moderate overparameterization (Oymak & Soltanolkotabi, 2020), representation learning beyond the kernel regime (Damian et al., 2022), and robustness to label noise through early stopping (Li et al., 2020).

Comparison to Other Poly-time Approaches. The most closely related tractable parameter-recovery frameworks are those of Bakshi et al. (2019) and Ge et al. (2018). These works impose rank conditions on different weight matrices: Ge et al. (2018) consider an undercomplete regime with $k \le d$ and a full-rank hidden weight matrix, while Bakshi et al. (2019) consider vector-valued outputs and achieve perfect recovery in the noiseless setting under a full-column-rank condition on the output weight matrix. In contrast, our recovery guarantee allows the overcomplete regime $k > d$. Rather than requiring the original hidden directions to be linearly independent, we assume only pairwise angular separation. Our positive-independence condition is satisfied in the standard well-conditioned undercomplete setting. In particular, if $W = [w_1, \dots, w_k]$ has $\sigma_{\min}(W) > 0$ and $\|w_\ell\|_2 \le 1$, then for every $\alpha \succeq 0$,

$$\|W\alpha\|_2 \ge \sigma_{\min}(W)\,\|\alpha\|_2 \ge \frac{\sigma_{\min}(W)}{\sqrt{k}} \sum_{\ell=1}^{k} \alpha_\ell \,\|w_\ell\|_2 ,$$

so Assumption 4.1 holds with $\eta \ge \sigma_{\min}(W)/\sqrt{k}$. Assumption 4.1 can nevertheless hold when W is rank deficient, including when $k > d$.

Finally, unlike the function-learning guarantees of Chen & Narayanan (2024) and Diakonikolas & Kane (2024), our objective is recovery of the planted parameters themselves: after approximate unoriented recovery and convex sign orientation, we use the anchored-regression framework of Bahmani & Romberg (2019) to refine informative anchors to exact recovery under the conditions of Theorem C.1. More broadly, we believe that the structural insights developed here may also lead to a better understanding of function recovery for shallow ReLU networks, beyond the parameter-recovery setting considered in this work.

6 Proof Overview and Key Ideas

In our proof the Hermite polynomial expansion of the population loss plays an important role.

The Landscape and Hermite Expansion. As elegantly established by (Ge et al., 2018), evaluating a planted positive ReLU network f_W and a learner network $f_{\widetilde{W}}$ on standard Gaussian covariates allows us to decompose the expected L_2 population risk using probabilists' Hermite polynomials. If we parameterize the hidden weights by their magnitudes $r_\ell = \|w_\ell\|_2$ and unit-norm directions $u_\ell = w_\ell/\|w_\ell\|_2$ (and similarly $\tilde{r}_\ell = \|\tilde{w}_\ell\|_2$ and $\tilde{u}_\ell = \tilde{w}_\ell/\|\tilde{w}_\ell\|_2$ for $f_{\widetilde{W}}$), the population risk cleanly decouples into an infinite sum of tensor matching objectives:

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}\left[(f_W(x) - f_{\widetilde{W}}(x))^2\right] = \sum_{q=0}^{\infty} \hat{\sigma}_q^2 \left\| \sum_{\ell=1}^{k} r_\ell u_\ell^{\otimes q} - \sum_{\ell=1}^{k} \tilde{r}_\ell \tilde{u}_\ell^{\otimes q} \right\|_F^2 , \tag{8}$$

where $\hat{\sigma}_q$ represents the q-th Hermite coefficient of the ReLU activation function.

Identifiability via Even Tensors. This formula immediately reveals why the model is globally identifiable. For the ReLU function, every even-order Hermite coefficient is strictly non-zero ($\hat{\sigma}_{2p}^2 > 0$ for all $p \ge 0$). Thus, driving the population risk to zero requires perfectly matching an infinite sequence of even-order symmetric tensors: $\sum r_\ell u_\ell^{\otimes 2p} = \sum \tilde{r}_\ell \tilde{u}_\ell^{\otimes 2p}$. Because even-power tensors discard sign information (i.e., $v^{\otimes 2p} = (-v)^{\otimes 2p}$), this infinite sequence acts as a unique spatial signature that completely identifies the magnitudes r_ℓ and the *unoriented* lines $\pm u_\ell$.

The Core Obstruction: The Orientation Barrier. Equation 8 also hints at the main source of computational intractability. For the ReLU activation, all higher-order odd Hermite coefficients identically vanish: $\hat{\sigma}_{2p+1} = 0$ for all $p \geq 1$. Consequently, the loss landscape contains *zero* information about the orientation of the neurons from odd degrees > 1. The only constraint that fixes the signs $s_\ell \in \{-1, 1\}$ of the recovered unoriented lines comes strictly from the surviving $q = 1$ term:

$$\left\| \sum_{\ell=1}^{k} s_\ell r_\ell u_\ell - m_1 \right\|_2^2 = 0, \quad \text{where} \quad m_1 = \sum_{\ell=1}^{k} w_\ell. \tag{9}$$

This forces any learning algorithm to assign a binary sign to each recovered ray to match the first-order moment. If the parameters only satisfy a discrete subset-sum margin, this step seamlessly encodes an NP-hard subset-sum problem. This exposes the core algorithmic obstruction: exact recovery is intractable not because the hyperplanes are hidden, but because resolving their signs constitutes an adversarial combinatorial search.

Overcoming the Barrier: Implicit High-Order Moments and ℓ_∞ Relaxation. Isolating this orientation barrier points directly to our reconstruction algorithm. Algorithm Ω first recovers the hidden hyperplanes from higher even Hermite moments. The key observation is that tensor lifting amplifies angular separation: at degree m, the correlation between two lifted directions is $(u_i^\top u_j)^m$. We therefore choose a logarithmic even degree at which the lifted family is well conditioned. Rather than construct the corresponding tensors, we use the implicit moment estimator of Diakonikolas & Kane (2025) to evaluate only structured contractions of them. Random polynomial features turn these contractions into low-dimensional matrices, and a matrix-pencil construction separates the hidden components. Coordinate-weighted moments then recover each projector $u_\ell u_\ell^\top$, while a lower-order sketch recovers the associated magnitude. As a result we can extract the unoriented base vectors $\widehat{V} = [\widehat{v}_1, \ldots, \widehat{v}_k]$, where $\widehat{v}_\ell \approx r_\ell(\pm u_\ell)$.

To circumvent the NP-hard sign orientation problem without searching the boolean hypercube $\{-1, 1\}^k$, we relax the discrete constraints to a continuous convex box $z \in [-1, 1]^k$ and solve the ℓ_∞-relaxed least-squares program corresponding to Equation 9:

$$\widehat{z} \in \arg \min_{\|z\|_\infty \leq 1} \left\| \widehat{V} z - \widehat{m}_1 \right\|_2. \tag{10}$$

In general, this relaxation may return fractional, spurious solutions. Our *positive independence* condition (Assumption 4.1) prevents this by ensuring that no positive combination of the weights collapses to zero, forcing the optimizer to recover the correct discrete signs $\operatorname{sign}(\widehat{z}_\ell) \in \{-1, 1\}$.

Clean-Up via Anchored Regression. At this stage, we possess the correct signs and high-quality approximate estimates $\widehat{W}$ of the true weights. However, because these estimates are derived from finite-sample moment sketches, they inherently carry statistical noise. To bridge the gap from approximation to *exact* parameter recovery, we introduce an anchored regression clean-up phase. We treat the estimates $\widehat{W}$ as a geometric anchor matrix A and formulate a linear program that maximizes alignment with the normalized anchors B (obtained by normalizing columns of A):

$$\max_{\overline{W} \in \mathbb{R}^{d \times k}} \langle B, \overline{W} \rangle \quad \text{subject to} \quad \sum_{\ell=1}^{k} \operatorname{ReLU}(\overline{w}_\ell^\top x_i) \leq y_i \quad \forall i \in [n_A]. \tag{11}$$

Because the positive ReLU network is sub-linear and positively homogeneous, this constrained optimization operates as a globally convex funnel, scrubbing away residual empirical noise by snapping precisely to the true planted parameters.

Precise Proofs. The narrative above presents a conceptually clean, idealized roadmap. However, translating this into a mathematically rigorous finite-sample guarantee is somewhat involved. The rigorous argument must control three sources of error in the first two steps: approximation of the required high-order Hermite contractions from passive samples, conditioning of the random polynomial-feature representation, and perturbation of the resulting low-dimensional matrix pencil. Furthermore, for Step 3, we must prove that the ℓ_∞-relaxation remains strictly bound to the boolean hypercube even under adversarial, non-isotropic finite-sample noise. Managing these cascading errors and stability requirements forms the technical bulk of our proofs.

AI USE STATEMENT

In this work, we used generative AI tools for proving part of the mathematical claims. Specifically, an internal proprietary large language model provided substantial assistance in developing the computational hardness result, including the reduction from the hyperplane-orientation barrier to an NP-hard variant of vector subset sum and an initial proof of the reduction. The authors subsequently rewrote the argument, checked every step independently, and verified all mathematical claims, derivations, and proofs appearing in the final manuscript.

The positive results, including the design of Algorithm Ω, its overall structure, the implicit-moment recovery strategy, the sign-orientation procedure, and the associated proofs, were developed by the authors. Publicly accessible proprietary generative AI systems, including ChatGPT and Gemini, were used in this part only to assist with writing and polishing the exposition and to check or carry out routine algebraic calculations. All such calculations, derivations, and edits were independently verified by the authors.

We also used generative AI tools for retrieval and discovery, including identifying and locating potentially relevant prior work. These uses likewise involved publicly accessible proprietary systems, including ChatGPT and Gemini. All cited papers and all claims attributed to prior work were checked by the authors against the original sources.

The remaining required-disclosure categories were not used in this work. We reviewed all AI-assisted material and take responsibility for the final content of the paper, including its mathematical claims, proofs, citations, and presentation.

ETHICS STATEMENT

This is a theoretical work and does not involve human subjects, personal data, data collection, or deployment of machine-learning systems. We are not aware of specific ethical risks arising directly from the results. Our use of generative AI is disclosed separately in the AI use statement.

REPRODUCIBILITY STATEMENT

The paper states all assumptions and algorithms used in our results, and the appendix contains complete proofs of the hardness and tractability theorems, including the implicit-moment, sign-orientation, and anchored-regression arguments. No experimental data or external datasets are required.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 2019. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(80): 2773–2832, 2014. URL https://www.jmlr.org/papers/v15/anandkumar14b.html.

Sohail Bahmani and Justin Romberg. Solving equations of random convex functions via anchored regression. *Foundations of Computational Mathematics*, 19(4):813–841, 2019. doi: 10.1007/s10208-018-9401-4. URL https://link.springer.com/article/10.1007/s10208-018-9401-4.

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

Avrim Blum and Ronald L. Rivest. Training a 3-node neural network is NP-complete. In *Advances in Neural Information Processing Systems*, volume 1, 1988.

David R. Brillinger. A generalized linear model with "Gaussian" regressor variables. In Peter J. Bickel, Kjell A. Doksum, and Joseph L. Hodges, Jr. (eds.), *A Festschrift for Erich L. Lehmann*, pp. 97–114. Wadsworth, Belmont, California, 1982.

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

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

Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5413–5452. PMLR, 2022.

Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Transactions on Signal Processing*, 55(6, Part 2):2965–2973, 2007. doi: 10.1109/TSP.2007.893943.

Ilias Diakonikolas and Daniel M. Kane. Small covers for near-zero sets of polynomials and learning latent variable models. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 184–195. IEEE, 2020. URL https://arxiv.org/abs/2012.07774.

Ilias Diakonikolas and Daniel M. Kane. Efficiently learning one-hidden-layer ReLU networks via Schur polynomials. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1364–1378. PMLR, 2024. URL https://proceedings.mlr.press/v247/diakonikolas24c.html.

Ilias Diakonikolas and Daniel M. Kane. Implicit high-order moment tensor estimation and learning latent variable models. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1228–1247. IEEE, 2025. URL https://arxiv.org/abs/2411.15669.

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. URL https://proceedings.mlr.press/v125/diakonikolas20d.html.

Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostically learning multi-index models with queries. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1931–1952. IEEE, 2024.

Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Nikos Zarifis. Robust learning of multi-index models via iterative subspace approximation. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 2231–2239. IEEE, 2025. URL https://arxiv.org/abs/2502.09525.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3587–3596. PMLR, 2020. URL https://proceedings.mlr.press/v119/goel20a.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger (eds.), *Complexity of Computer Computations*, pp. 85–103. Springer, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4313–4324. PMLR, 2020. URL https://proceedings.mlr.press/v108/li20j.html.

Shuchen Li, Ilias Zadik, and Manolis Zampetakis. On the hardness of learning one hidden layer neural networks. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 700–701. PMLR, 2025. URL https://proceedings.mlr.press/v272/li25a.html.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4951–4960. PMLR, 2019. URL https://proceedings.mlr.press/v97/oymak19a.html.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Mahdi Soltanolkotabi. Learning ReLUs via gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019. doi: 10.1109/TIT.2018.2854560.

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, 1990.

Pierre Stock and Rémi Gribonval. An embedding of ReLU networks and an analysis of their identifiability. *Constructive Approximation*, 57(2):853–899, 2023. doi: 10.1007/s00365-022-09578-1. URL https://link.springer.com/article/10.1007/s00365-022-09578-1.

Leslie G. Valiant and Vijay V. Vazirani. NP is as easy as detecting unique solutions. *Theoretical Computer Science*, 47(1):85–93, 1986.

Zhuoran Yang, Krishna Balasubramanian, Zhaoran Wang, and Han Liu. Learning non-Gaussian multi-index model via second-order Stein's method. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer ReLU networks via gradient descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1524–1534. PMLR, 2019. URL https://proceedings.mlr.press/v89/zhang19g.html.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 4140–4149. PMLR, 2017. URL https://proceedings.mlr.press/v70/zhong17a.html.

A PROOFS FOR IDENTIFIABLITY (PROOF OF THEOREM 2.3)

In this section, we provide the full proof of Theorem 2.3, establishing that the parameters of a positive ReLU network are globally identifiable up to permutation under Gaussian covariates and our nondegeneracy assumptions.

Let $g \sim \mathcal{N}(0,1)$ be a standard normal random variable, and let $\{h_q\}_{q\geq 0}$ denote the probabilists' Hermite polynomials, normalized such that $\mathbb{E}[h_a(g)h_b(g)] = a!\delta_{ab}$. The ReLU activation function, $\phi(t) = \max\{t, 0\}$, admits the Hermite expansion

$$\phi(g) = \sum_{q=0}^{\infty} a_q h_q(g), \quad \text{where} \quad a_q = \frac{1}{q!}\mathbb{E}[\phi(g)h_q(g)]. \tag{12}$$

It is a standard fact that for the ReLU function, the first Hermite coefficient is $a_1 = 1/2 \neq 0$, all even coefficients a_{2p} are strictly non-zero for every $p \geq 0$, and all higher-order odd coefficients $a_{2p+1} = 0$ for every $p \geq 1$.

For each hidden weight w_ℓ and u_ℓ, we can decouple its magnitude from its direction. Let $r_\ell = \|w_\ell\|_2 > 0$ and $\widehat{w}_\ell = w_\ell/r_\ell \in \mathbb{S}^{d-1}$. Similarly, let $\rho_\ell = \|u_\ell\|_2 > 0$ and $\widehat{u}_\ell = u_\ell/\rho_\ell \in \mathbb{S}^{d-1}$. Since the input covariates are standard Gaussian, $x \sim \mathcal{N}(0, I_d)$, the one-dimensional projection $v^\top x \sim \mathcal{N}(0,1)$ for any fixed unit vector $v \in \mathbb{S}^{d-1}$. By the positive homogeneity of the ReLU function, we can express the network outputs as

$$f_W(x) = \sum_{\ell=1}^{k} r_\ell \phi(\widehat{w}_\ell^\top x) = \sum_{q=0}^{\infty} a_q \sum_{\ell=1}^{k} r_\ell h_q(\widehat{w}_\ell^\top x), \tag{13}$$

$$f_U(x) = \sum_{\ell=1}^{k} \rho_\ell \phi(\widehat{u}_\ell^\top x) = \sum_{q=0}^{\infty} a_q \sum_{\ell=1}^{k} \rho_\ell h_q(\widehat{u}_\ell^\top x). \tag{14}$$

The theorem assumes that $\mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[(f_W(x) - f_U(x))^2] = 0$, which implies $f_W(x) = f_U(x)$ almost surely. Consequently, the Hermite expansions of f_W and f_U must match at every degree. The degree-q Hermite component of $x \mapsto h_q(v^\top x)$ corresponds to the symmetric tensor $v^{\otimes q}$. Thus, for every degree q where $a_q \neq 0$, matching the degree-q components yields the tensor equality

$$\sum_{\ell=1}^{k} r_\ell \widehat{w}_\ell^{\otimes q} = \sum_{\ell=1}^{k} \rho_\ell \widehat{u}_\ell^{\otimes q}. \tag{15}$$

In particular, since $a_{2p} \neq 0$ for all $p \geq 0$, Equation (15) holds for all even degrees $q = 2p$. We can interpret this infinite sequence of polynomial identities measure-theoretically. Consider two finite discrete measures defined on the projective space $\mathbb{S}^{d-1}/\{\pm 1\}$:

$$\mu_W = \sum_{\ell=1}^{k} r_\ell \delta_{\{\pm\widehat{w}_\ell\}} \quad \text{and} \quad \mu_U = \sum_{\ell=1}^{k} \rho_\ell \delta_{\{\pm\widehat{u}_\ell\}}. \tag{16}$$

Equation (15) for $q = 2p$ indicates that all even tensor moments of these two measures perfectly agree:

$$\int z^{\otimes 2p}\, d\mu_W(z) = \int z^{\otimes 2p}\, d\mu_U(z) \quad \text{for all } p \geq 0. \tag{17}$$

Equivalently, $\int P(z)\, d\mu_W(z) = \int P(z)\, d\mu_U(z)$ for every even polynomial P on the sphere. Because even polynomials separate points on the projective space, and the projective space is compact, this moment-matching condition guarantees that the two measures are strictly identical, yielding $\mu_W = \mu_U$.

This equality dictates that the atoms (the unoriented directions) and their respective masses (the norms) must match. Under the Non-collinearity condition (Assumption 2.1), no two planted weights within W are collinear, and similarly for U. Thus, all atoms within each measure are distinct. Hence, there exists a unique permutation π of $[k]$ such that $\{\pm\widehat{u}_{\pi(\ell)}\} = \{\pm\widehat{w}_\ell\}$ and $\rho_{\pi(\ell)} = r_\ell$ for all $\ell \in [k]$. This allows us to write

$$u_{\pi(\ell)} = \sigma_\ell w_\ell \quad \text{for some signs} \quad \sigma_\ell \in \{-1, 1\}. \tag{18}$$

It remains to orient these signs. We appeal to the degree-one Hermite component. Since $a_1 = 1/2 \neq 0$, Equation (15) also holds for $q = 1$, giving

$$\sum_{\ell=1}^{k} r_\ell \widehat{w}_\ell = \sum_{\ell=1}^{k} \rho_\ell \widehat{u}_\ell. \tag{19}$$

Multiplying the directions by their respective lengths, this simplifies to the matching of the first-order moments:

$$\sum_{\ell=1}^{k} w_\ell = \sum_{\ell=1}^{k} u_\ell. \tag{20}$$

Substituting $u_{\pi(\ell)} = \sigma_\ell w_\ell$ into the right-hand side, we find

$$\sum_{\ell=1}^{k} w_\ell = \sum_{\ell=1}^{k} \sigma_\ell w_\ell. \tag{21}$$

Subtracting the right-hand side from the left-hand side and isolating the terms where $\sigma_\ell = -1$ gives

$$\sum_{\ell:\sigma_\ell=-1} 2w_\ell = 0. \tag{22}$$

Let $S = \{\ell \in [k] \mid \sigma_\ell = -1\}$ be the subset of indices where the signs are inverted. The condition then becomes $\sum_{\ell \in S} w_\ell = 0$. By the No subset cancellation condition (Assumption 2.2), this sum cannot evaluate to zero for any nonempty subset S. This strictly forces $S = \emptyset$, confirming that $\sigma_\ell = 1$ for all $\ell \in [k]$.

We conclude that $u_{\pi(\ell)} = w_\ell$ for all $\ell \in [k]$, proving that the network weights are globally identifiable up to permutation.

B Proof of Computational Hardness (Theorem 3.4 and Corollary 3.5)

In this section, we provide the complete formal proof of the adaptive membership-query lower bound (Theorem 3.4) and its corollary for passive Gaussian recovery (Corollary 3.5). We establish a randomized polynomial-time Turing reduction from the NP-complete Exact Cover by 3-Sets (X3C) problem (Karp, 1972).

The reduction is structured into four main steps:

1. **Parsimonious Reduction to Signed Vector Subset Sum:** We map X3C to a signed Vector Subset Sum (VSS) problem.
2. **Randomized Isolation:** We apply the Valiant-Vazirani Isolation Lemma to ensure the VSS instance has a strictly unique sign solution if it is satisfiable. We encode the resulting parity constraints as integer equations using auxiliary slack variables without introducing collinearity.
3. **Geometric Robustification via Centered-Simplex Replication:** We introduce a replication gadget that embeds the integer instance into a higher-dimensional space. This structurally guarantees that the planted weights rigorously satisfy the continuous geometric margins (Assumptions 3.1, 3.2, and 3.3).
4. **Pointwise Oracle Simulation:** We exploit the algebraic properties of the positive ReLU activation to simulate an exact membership oracle using exclusively public information, effectively masking the discrete combinatorial signs.

B.1 Step 1: Parsimonious Reduction to Signed Vector Subset Sum

The Exact Cover by 3-Sets (X3C) problem asks whether a universe $\mathcal{X} = \{1, \ldots, d_X\}$, with $d_X = 3q$, can be partitioned by a sub-collection of the given 3-element subsets $C_1, \ldots, C_{k_X}$. After removing duplicate sets, which does not change satisfiability, we may assume that the sets C_j are

distinct. Let $u_j \in \{0,1\}^{d_X}$ denote the indicator vector of C_j. Then an exact cover is equivalently a vector $b \in \{0,1\}^{k_X}$ satisfying

$$\sum_{j=1}^{k_X} b_j u_j = \mathbf{1}_{d_X}, \tag{23}$$

where $\mathbf{1}_{d_X}$ is the all-ones vector.

We now convert the Boolean variables to sign variables by setting $s_j = 2b_j - 1 \in \{-1, 1\}$, or equivalently $b_j = (s_j + 1)/2$. Substituting this expression into the exact-cover constraint and multiplying by 2 gives

$$\sum_{j=1}^{k_X} s_j u_j = 2\mathbf{1}_{d_X} - \sum_{j=1}^{k_X} u_j := T_{\text{base}}. \tag{24}$$

Thus the X3C instance is mapped to the signed Vector Subset Sum (VSS) instance $(u_1, \ldots, u_{k_X}; T_{\text{base}})$, which asks whether there exists $s \in \{-1, 1\}^{k_X}$ such that

$$\sum_{j=1}^{k_X} s_j u_j = T_{\text{base}}. \tag{25}$$

This map is solution-preserving and one-to-one: an exact cover b gives a sign solution through $s_j = 2b_j - 1$, and a sign solution s gives an exact cover through $b_j = (s_j + 1)/2$. These two maps are inverse to each other.

Finally, each vector u_j is nonzero because C_j has size 3. The vectors u_j are also pairwise non-collinear. Indeed, if $u_i = cu_j$ for some scalar c, then comparing ℓ_1 norms gives $3 = 3|c|$, so $|c| = 1$. Since both vectors have nonnegative coordinates and are nonzero, this forces $c = 1$, hence $u_i = u_j$, contradicting the assumption that $C_i \neq C_j$. At this stage the signed VSS instance may have many sign solutions; the next step uses randomized isolation to reduce to the unique-solution case.

B.2 Step 2: Randomized Isolation and Integer Encoding

The signed VSS instance obtained in Step 1 may have many sign solutions. For the geometric margin argument in the next step, we need a unique sign solution: if a nonempty signed sub-sum vanished, then flipping the signs on that sub-sum would create a second solution. We therefore first apply randomized isolation to the set of exact covers before embedding the resulting constraints into an ordinary integer signed VSS instance.

We use the Valiant-Vazirani isolation lemma (Valiant & Vazirani, 1986). Sample an integer h uniformly from $\{1, \ldots, k_X\}$. Then sample independent random bits $v_{i,j}, c_i \in \{0,1\}$ for each $i \in [h]$ and $j \in [k_X]$, and impose the random parity constraints

$$\sum_{j=1}^{k_X} v_{i,j} b_j \equiv c_i \pmod 2, \qquad i = 1, \ldots, h. \tag{26}$$

There is a universal constant $\gamma_{\text{iso}} > 0$ such that the following holds. If the original X3C instance is unsatisfiable, then the constrained instance is always unsatisfiable. If the original X3C instance is satisfiable, then with probability at least γ_{iso}/k_X over the random choice of h, $v_{i,j}$, and c_i, there is one exact cover $b \in \{0,1\}^{k_X}$ satisfying both the original X3C constraint and all the parity constraints.

We next encode these parity constraints as ordinary integer equations. Fix one constraint $i \in [h]$. Whenever

$$\sum_{j=1}^{k_X} v_{i,j} b_j \equiv c_i \pmod 2$$

is satisfied, the quantity $\sum_{j=1}^{k_X} v_{i,j} b_j - c_i$ is an even integer. It is also nonnegative: if $c_i = 0$ this is immediate, while if $c_i = 1$, the parity condition forces $\sum_{j=1}^{k_X} v_{i,j} b_j$ to be odd and hence at least 1.

Since this quantity is at most k_X, it can be written using binary slack bits $y_{i,p} \in \{0,1\}$ as

$$\sum_{j=1}^{k_X} v_{i,j} b_j - \sum_{p=0}^{L} 2^{p+1} y_{i,p} = c_i, \qquad L = \lfloor \log_2 k_X \rfloor. \tag{27}$$

The slack expression $\sum_{p=0}^{L} 2^{p+1} y_{i,p}$ represents every nonnegative even integer at most k_X, since $L = \lfloor \log_2 k_X \rfloor$ implies $2(2^{L+1} - 1) \geq k_X$. Its representation is unique because it is the ordinary binary representation of half of the even integer.

We now convert both the original variables and the slack variables into sign variables. Write

$$b_j = \frac{s_j + 1}{2}, \qquad y_{i,p} = \frac{r_{i,p}^{+} + 1}{2},$$

where $s_j, r_{i,p}^{+} \in \{-1, 1\}$. Substituting these expressions into (27) and multiplying by 2 gives

$$\sum_{j=1}^{k_X} v_{i,j} s_j - \sum_{p=0}^{L} 2^{p+1} r_{i,p}^{+} = 2c_i - \sum_{j=1}^{k_X} v_{i,j} + \sum_{p=0}^{L} 2^{p+1}. \tag{28}$$

It remains to encode the negative slack contribution in the standard signed VSS form $\sum_\ell \sigma_\ell W_\ell = T$, where all coefficients σ_ℓ are signs. We do this by introducing an auxiliary pair of vectors for each slack bit. Let the ambient dimension be

$$D = d_X + h + h(L+1).$$

The first d_X coordinates are the original X3C coordinates. The next h coordinates encode the parity equations. The remaining $h(L+1)$ coordinates are private coordinates, one for each pair (i,p). Define

$$H(i,p) = d_X + h + (i-1)(L+1) + p + 1$$

for $i \in [h]$ and $p \in \{0, \ldots, L\}$.

Equivalently, the augmented signed VSS instance is defined as follows.

- **Base vectors.** For each original set variable $j \in [k_X]$, define $U_j \in \mathbb{Z}^D$ by setting $U_j[a] = u_j[a]$ for $1 \leq a \leq d_X$, setting $U_j[d_X + i] = v_{i,j}$ for $i \in [h]$, and setting all remaining coordinates to 0.
- **Auxiliary vectors.** For each parity constraint $i \in [h]$ and each bit index $p \in \{0, \ldots, L\}$, introduce an auxiliary pair $V_{i,p}^{+}, V_{i,p}^{-} \in \mathbb{Z}^D$. The vector $V_{i,p}^{+}$ has entry -2^{p+1} in coordinate $d_X + i$, entry 1 in its private coordinate $H(i,p)$, and 0 elsewhere. The vector $V_{i,p}^{-}$ has entry 1 in coordinate $H(i,p)$ and 0 elsewhere.
- **Target vector.** The target vector $T \in \mathbb{Z}^D$ is defined by

$$T[a] = 2\mathbf{1}_{d_X}[a] - \sum_{j=1}^{k_X} u_j[a] \quad \text{for } 1 \leq a \leq d_X,$$

$$T[d_X + i] = 2c_i - \sum_{j=1}^{k_X} v_{i,j} + \sum_{p=0}^{L} 2^{p+1} \quad \text{for } i \in [h],$$

and

$$T[H(i,p)] = 0 \quad \text{for all } i \in [h],\ p \in \{0, \ldots, L\}.$$

Let $W_1, \ldots, W_K$ be the collection of all vectors

$$U_1, \ldots, U_{k_X}, \qquad V_{i,p}^{+}, V_{i,p}^{-} \quad (i \in [h],\ p \in \{0, \ldots, L\}),$$

where

$$K = k_X + 2h(L+1).$$

We interpret the signs assigned to U_j as s_j, the signs assigned to $V^+_{i,p}$ as $r^+_{i,p}$, and the signs assigned to $V^-_{i,p}$ as $r^-_{i,p}$. The private coordinate $H(i,p)$ enforces

$$r^+_{i,p} + r^-_{i,p} = 0,$$

and therefore $r^-_{i,p} = -r^+_{i,p}$. Consequently, in parity coordinate $d_X + i$, the auxiliary pair contributes

$$-2^{p+1} r^+_{i,p},$$

which reproduces the slack term in (28). Therefore the signed VSS equation

$$\sum_{\ell=1}^{K} \sigma_\ell W_\ell = T, \qquad \sigma_\ell \in \{-1, 1\}, \tag{29}$$

is equivalent to the original exact-cover constraint together with the random parity constraints and their slack-bit encodings.

The construction has polynomial size and polynomially bounded integer entries, because the largest coefficient is at most $2^{L+1} \le 2k_X$. Moreover, all vectors are nonzero and pairwise non-collinear. The original vectors U_j are nonzero and pairwise non-collinear because their first d_X coordinates are the distinct 3-set indicators u_j. No original vector U_j is collinear with an auxiliary vector, since U_j has a nonzero coordinate among the first d_X coordinates while every auxiliary vector is zero there. Finally, two distinct auxiliary vectors are not collinear because each auxiliary pair has its own private coordinate $H(i,p)$, and within the same pair $V^+_{i,p}$ has the additional parity-coordinate entry -2^{p+1} whereas $V^-_{i,p}$ does not.

Thus, after randomized isolation and integer encoding, we obtain a signed integer VSS instance $(W_1, \dots, W_K; T)$ with the following property. If the original X3C instance is unsatisfiable, then this VSS instance has no sign solution. If the original X3C instance is satisfiable, then with probability at least $\gamma_{\mathrm{iso}}/k_X$ the resulting VSS instance has one sign solution.

B.3 Step 3: Geometric Robustification via Centered-Simplex Replication

We take as input the signed integer VSS instance $(W_1, \dots, W_K; T)$ obtained in Step 2, where $W_\ell \in \mathbb{Z}^D$. In the good isolated case, this instance has a unique sign solution $s^* \in \{-1, 1\}^K$. The purpose of this step is to convert this integer VSS instance into a family of planted ReLU weights satisfying the norm, angular-separation, and subset-sum margin assumptions.

Let $\mathcal{W} = \{W_1, \dots, W_K\}$. Let B be a polynomial upper bound on the Euclidean norm of the vectors in $\mathcal{W}$, so that $\|W_\ell\|_2 \le B$ for all $\ell \in [K]$. Define the replication factor

$$m_{\mathrm{rep}} = \max\{\lceil B \rceil, 100\}.$$

For each $\ell \in [K]$, we introduce a private coordinate block $\mathbb{R}^{m_{\mathrm{rep}}}$. Inside each such block, define the centered simplex vectors

$$g_a = e_a - \frac{1}{m_{\mathrm{rep}}} \mathbf{1}_{m_{\mathrm{rep}}}, \qquad a \in [m_{\mathrm{rep}}]$$

where e_a denotes the a-th standard basis vector in $\mathbb{R}^{m_{\mathrm{rep}}}$, and $\mathbf{1}_{m_{\mathrm{rep}}}$ denotes the all-ones vector. Then we have

$$\sum_{a=1}^{m_{\mathrm{rep}}} g_a = 0, \qquad \|g_a\|_2^2 = 1 - \frac{1}{m_{\mathrm{rep}}}, \qquad \langle g_a, g_b \rangle = -\frac{1}{m_{\mathrm{rep}}} \quad (a \ne b).$$

We enlarge the ambient dimension to

$$d = D + K m_{\mathrm{rep}}.$$

For each $\ell \in [K]$ and $a \in [m_{\mathrm{rep}}]$, define $Z_{\ell,a} \in \mathbb{R}^d$ by

$$Z_{\ell,a} = \left(\frac{1}{m_{\mathrm{rep}}} W_\ell,\ 0, \dots, 0,\ 10 g_a \text{ in private block } \ell,\ 0, \dots, 0 \right). \tag{30}$$

Let

$$T' = (T, 0, \dots, 0) \in \mathbb{R}^d.$$

We first record the solution-preserving property of the replication gadget. Suppose signs $\tau_{\ell,a} \in \{-1, 1\}$ satisfy

$$\sum_{\ell=1}^{K} \sum_{a=1}^{m_{\text{rep}}} \tau_{\ell,a} Z_{\ell,a} = T'.$$

Looking at the private block corresponding to a fixed ℓ, we obtain

$$\sum_{a=1}^{m_{\text{rep}}} \tau_{\ell,a} g_a = 0.$$

Since

$$\sum_{a=1}^{m_{\text{rep}}} \tau_{\ell,a} g_a = \tau_\ell - \frac{1}{m_{\text{rep}}} \left(\sum_{a=1}^{m_{\text{rep}}} \tau_{\ell,a} \right) \mathbf{1}_{m_{\text{rep}}},$$

where $\tau_\ell = (\tau_{\ell,1}, \dots, \tau_{\ell,m_{\text{rep}}})$, this equality holds if and only if

$$\tau_{\ell,1} = \cdots = \tau_{\ell,m_{\text{rep}}}.$$

Thus every feasible replicated sign assignment is constant inside each block. Writing this common value as s_ℓ, the public coordinates reduce to

$$\sum_{\ell=1}^{K} s_\ell W_\ell = T.$$

Conversely, every sign solution $s \in \{-1, 1\}^K$ of the original VSS instance gives a replicated sign solution by setting $\tau_{\ell,a} = s_\ell$ for all a. Hence the replicated instance has sign solutions in bijection with the original VSS instance.

We now scale the replicated vectors by a fixed constant:

$$P_{\ell,a} = \frac{1}{11} Z_{\ell,a}, \qquad Q = \frac{1}{11} T'.$$

The total number of planted hidden weights is

$$k = K m_{\text{rep}}.$$

In the good isolated case, let s^* be the unique sign solution of the original VSS instance and define the planted weights

$$w^*_{\ell,a} = s^*_\ell P_{\ell,a}.$$

Then

$$\sum_{\ell=1}^{K} \sum_{a=1}^{m_{\text{rep}}} w^*_{\ell,a} = Q. \tag{31}$$

Indeed, the private components cancel inside every block because $\sum_a g_a = 0$, and the public coordinates sum to $\frac{1}{11} \sum_{\ell=1}^{K} s^*_\ell W_\ell = \frac{1}{11} T$.

step.

Lemma B.1 (Robust Geometric Margins)**.** *Assume the signed VSS instance $(W_1, \dots, W_K; T)$ has a unique sign solution s^*. Then the planted weights $w^*_{\ell,a} = s^*_\ell P_{\ell,a}$ satisfy:*

1. ***Norm bounds:***

$$\frac{1}{12} \le \left\| w^*_{\ell,a} \right\|_2 = \left\| P_{\ell,a} \right\|_2 \le 1 \qquad \textit{for all } \ell, a.$$

2. ***Robust non-collinearity:*** *for any distinct pairs $(\ell, a) \neq (\ell', b)$,*

$$\frac{\left| \left\langle w^*_{\ell,a}, w^*_{\ell',b} \right\rangle \right|}{\left\| w^*_{\ell,a} \right\|_2 \left\| w^*_{\ell',b} \right\|_2} = \frac{\left| \langle P_{\ell,a}, P_{\ell',b} \rangle \right|}{\left\| P_{\ell,a} \right\|_2 \left\| P_{\ell',b} \right\|_2} \le \frac{1}{2}.$$

3. ***Positive subset-sum margin, in Euclidean norm:***

$$\min_{\emptyset \neq S \subseteq [K]\times[m_{\mathrm{rep}}]} \left\| \sum_{(\ell,a)\in S} w^*_{\ell,a} \right\|_2 \geq \frac{1}{12}.$$

Thus Assumptions 3.1, 3.2, and 3.3 hold with $r = 1/12$, $\rho = 1/2$, *and* $\eta = 1/12$.

Proof. We first prove the norm bounds. For any ℓ, a, using (30),

$$\|Z_{\ell,a}\|_2^2 = \frac{1}{m_{\mathrm{rep}}^2} \|W_\ell\|_2^2 + 100\left(1 - \frac{1}{m_{\mathrm{rep}}}\right).$$

Since $m_{\mathrm{rep}} \geq \|W_\ell\|_2$ and $m_{\mathrm{rep}} \geq 100$,

$$0 \leq \frac{1}{m_{\mathrm{rep}}^2} \|W_\ell\|_2^2 \leq 1$$

and

$$99 \leq 100\left(1 - \frac{1}{m_{\mathrm{rep}}}\right) < 100.$$

Therefore

$$99 \leq \|Z_{\ell,a}\|_2^2 \leq 101.$$

After scaling,

$$\|P_{\ell,a}\|_2 = \frac{1}{11}\|Z_{\ell,a}\|_2 \in \left[\frac{\sqrt{99}}{11}, \frac{\sqrt{101}}{11}\right] \subseteq \left[\frac{1}{12}, 1\right].$$

Since $w^*_{\ell,a}$ differs from $P_{\ell,a}$ only by a sign, the same norm bounds hold for $w^*_{\ell,a}$.

We next prove robust non-collinearity. Consider two distinct pairs $(\ell, a) \neq (\ell', b)$. If $\ell = \ell'$ and $a \neq b$, then

$$\langle Z_{\ell,a}, Z_{\ell,b}\rangle = \frac{1}{m_{\mathrm{rep}}^2}\|W_\ell\|_2^2 - \frac{100}{m_{\mathrm{rep}}}.$$

The first term has absolute value at most 1, and the second term has absolute value at most 1, so

$$|\langle Z_{\ell,a}, Z_{\ell,b}\rangle| \leq 2.$$

If $\ell \neq \ell'$, the private blocks are orthogonal, and hence

$$|\langle Z_{\ell,a}, Z_{\ell',b}\rangle| = \frac{1}{m_{\mathrm{rep}}^2}|\langle W_\ell, W_{\ell'}\rangle| \leq \frac{1}{m_{\mathrm{rep}}^2}\|W_\ell\|_2\|W_{\ell'}\|_2 \leq 1.$$

In all cases,

$$|\langle Z_{\ell,a}, Z_{\ell',b}\rangle| \leq 2.$$

Since $\|Z_{\ell,a}\|_2^2 \geq 99$ and $\|Z_{\ell',b}\|_2^2 \geq 99$, we have

$$\|Z_{\ell,a}\|_2\|Z_{\ell',b}\|_2 \geq 99.$$

Thus

$$\frac{|\langle Z_{\ell,a}, Z_{\ell',b}\rangle|}{\|Z_{\ell,a}\|_2\|Z_{\ell',b}\|_2} \leq \frac{2}{99} \leq \frac{1}{2}.$$

Scaling by $1/11$ and multiplying columns by signs do not change this angular correlation, so the same bound holds for the planted weights.

It remains to prove the subset-sum margin. Let $S \subseteq [K] \times [m_{\mathrm{rep}}]$ be nonempty. For each $\ell \in [K]$, define

$$S_\ell = \{a \in [m_{\mathrm{rep}}] : (\ell, a) \in S\}.$$

First suppose that for some ℓ, the set S_ℓ is nonempty and is not all of $[m_{\mathrm{rep}}]$. Let $t = |S_\ell|$, so $1 \leq t \leq m_{\mathrm{rep}} - 1$. The private block ℓ of the subset sum $\sum_{(\ell',a)\in S} w^*_{\ell',a}$ is

$$\frac{10}{11}s^*_\ell \sum_{a\in S_\ell} g_a.$$

Its squared norm is

$$\left(\frac{10}{11}\right)^2 \left\| \sum_{a \in S_\ell} g_a \right\|_2^2 = \frac{100}{121} \cdot \frac{t(m_{\text{rep}} - t)}{m_{\text{rep}}}.$$

Since $1 \le t \le m_{\text{rep}} - 1$,

$$\frac{t(m_{\text{rep}} - t)}{m_{\text{rep}}} \ge \frac{m_{\text{rep}} - 1}{m_{\text{rep}}} \ge \frac{99}{100}.$$

Therefore the Euclidean norm of this private block is at least

$$\frac{\sqrt{99}}{11}.$$

The norm of the full subset sum is at least the norm of any one coordinate block, so

$$\left\| \sum_{(\ell', a) \in S} w^*_{\ell', a} \right\|_2 \ge \frac{\sqrt{99}}{11} > \frac{1}{12}.$$

It remains to handle the case where every S_ℓ is either empty or all of $[m_{\text{rep}}]$. Let

$$J = \{\ell \in [K] : S_\ell = [m_{\text{rep}}]\}.$$

Since S is nonempty, J is nonempty. The private simplex components cancel inside every full block, and hence

$$\sum_{(\ell,a) \in S} w^*_{\ell,a} = \sum_{\ell \in J} \sum_{a=1}^{m_{\text{rep}}} s^*_\ell P_{\ell,a} = \frac{1}{11} \sum_{\ell \in J} s^*_\ell W_\ell.$$

We claim that the integer vector $\sum_{\ell \in J} s^*_\ell W_\ell$ is nonzero. If it were zero, then flipping the signs of s^* on the nonempty set J would produce another sign vector $\widetilde{s} \neq s^*$ with

$$\sum_{\ell=1}^{K} \widetilde{s}_\ell W_\ell = \sum_{\ell=1}^{K} s^*_\ell W_\ell - 2 \sum_{\ell \in J} s^*_\ell W_\ell = T,$$

contradicting the uniqueness of the sign solution s^*. Hence $\sum_{\ell \in J} s^*_\ell W_\ell$ is a nonzero integer vector, so its Euclidean norm is at least 1. Therefore

$$\left\| \sum_{(\ell,a) \in S} w^*_{\ell,a} \right\|_2 \ge \frac{1}{11} > \frac{1}{12}.$$

This proves the subset-sum margin with $\eta = 1/12$ in Euclidean norm, not squared Euclidean norm. The lemma follows. □

B.4 Step 4: Pointwise Oracle Simulation and Exact Recovery (Proof of Theorem 3.4)

Suppose there exists a randomized polynomial-time algorithm $\mathcal{A}$ capable of recovering the hidden weights of any positive ReLU network satisfying Assumptions 3.1–3.3 via adaptive, noise-free membership queries.

Given an arbitrary X3C instance, we synthesize the public matrices $P_{\ell,a}$ alongside the unified target Q. We implicitly define the target network using the isolated true weights $w^*_{\ell,a} = s^*_\ell P_{\ell,a}$, where s^*_ℓ is the unique sign solution. To seamlessly simulate the active membership oracle empowering the algorithm without accidentally exposing the latent orientation signs s^*_ℓ, we exploit the homogeneous algebraic identity $\text{ReLU}(t) = \frac{1}{2}(|t| + t)$. For any requested query $x \in \mathbb{R}^d$, the continuous label

equals:

$$
\begin{aligned}
f_{W^*}(x) &= \sum_{\ell=1}^{K} \sum_{a=1}^{m_{\text{rep}}} \mathrm{ReLU}(\langle w_{\ell,a}^*, x \rangle) \\
&= \frac{1}{2} \sum_{\ell=1}^{K} \sum_{a=1}^{m_{\text{rep}}} |\langle s_\ell^* P_{\ell,a}, x \rangle| + \frac{1}{2} \sum_{\ell=1}^{K} \sum_{a=1}^{m_{\text{rep}}} \langle w_{\ell,a}^*, x \rangle \\
&= \frac{1}{2} \sum_{\ell=1}^{K} \sum_{a=1}^{m_{\text{rep}}} |\langle P_{\ell,a}, x \rangle| + \frac{1}{2} \langle Q, x \rangle . \qquad (32)
\end{aligned}
$$

Because $|s_\ell^*| = 1$, the absolute value terms dynamically filter out the hidden signs, relying explicitly only on the public vectors. Simultaneously, the linear term collapses into the public target Q. Therefore, we can transparently evaluate every adaptive query strictly using public information.

If the X3C instance is satisfiable, with probability $\Omega(1/k_X)$, our randomized isolation guarantees a unique target $w_{\ell,a}^*$ satisfying all geometric properties (Lemma B.1). The exact-recovery algorithm $\mathcal{A}$ smoothly evaluates this simulated oracle and outputs the true unordered set of internal weights $\{w_{\ell,a}^*\}$ with probability at least $2/3$. By structurally cross-matching the returned unorganized components against the public boundaries dictated by $P_{\ell,a}$, we uniquely and definitively decode the respective hidden signs s_ℓ^*. On the other hand, if this cross-matching step fails, or if the recovered signs are not constant within each replicated block, we reject the current trial. Also, if the X3C instance is unsatisfiable, or if the randomized isolation step does not leave a unique sign solution, then the simulated oracle may not come from a valid planted ReLU network. In that case, the algorithm $\mathcal{A}$ may output anything. This is harmless, because our reduction accepts only after checking that the decoded signs really form a valid exact cover.

Finally, we verify the outcome by checking whether the extracted signs correspond to a valid exact cover. If they do, we output "Satisfiable"; otherwise, we reject the current trial. If the original X3C instance is unsatisfiable, then no decoded vector $b \in \{0,1\}^{k_X}$ can satisfy the exact-cover equation. Hence the reduction never accepts an unsatisfiable instance. This remains true even if the simulated oracle does not correspond to a valid planted ReLU network, since every candidate output is filtered by the final exact-cover verification. If the original X3C instance is satisfiable, then with probability $\Omega(1/k_X)$ the randomized isolation step produces a unique sign solution. Conditioned on this event, the simulated oracle is a valid membership oracle for a network satisfying the required geometric assumptions, and $\mathcal{A}$ recovers the planted weights with probability at least $2/3$. Thus one trial accepts with probability $\Omega(1/k_X)$. Repeating the randomized reduction independently $O(k_X)$ times amplifies the acceptance probability to at least $2/3$, while preserving one-sided error.

Therefore, the existence of $\mathcal{A}$ would give a randomized polynomial-time algorithm for X3C with one-sided error. Since X3C is NP-complete, this implies $\mathrm{NP} \subseteq \mathrm{RP}$, proving Theorem 3.4. □

B.5 Passive Gaussian Recovery (Proof of Corollary 3.5)

All computations in this proof use the Gaussian real-arithmetic model below.

Definition B.2 (Computational models)**.** *In the* rational membership-query model*, the dimensions and a binary encoding-length bound b for the rational planted parameter matrix are supplied to the learner. Queries, labels, and output coordinates are rationals in binary representation. Running time includes the bit cost of producing queries, reading labels, and writing the output, and is polynomial in d, k, b. Every invocation is stopped after its stated polynomial time bound, including on oracles outside the promise.*

In the Gaussian real-arithmetic model*, examples have exact real coordinates and labels. Registers contain exact real numbers; addition, subtraction, multiplication, division by a nonzero number, comparison, and the nonnegative square root each cost one operation. The fixed constant π, independent standard Gaussian draws, and independent fair bits are available. Integer loop bounds can be constructed by doubling and comparison. No eigendecomposition, polynomial-root, or convex-optimization oracle is included.*

Perform the isolation, integer encoding, and replication construction from Sections B.1-B.3. Generate the Gaussian covariates required by the passive learner and assign each label using the public expression on the right-hand side of equation (32). This uses $O(kd)$ elementary operations per label.

On an isolated satisfiable trial, these examples have exactly the distribution of Gaussian examples from the promised planted network. The learner therefore returns the exact planted parameter set with probability at least $2/3$. Match its output against the public vectors $\pm P_{\ell,a}$, check blockwise sign consistency, and accept only after verifying the decoded exact cover. All these operations are polynomial in the construction size.

Each learner invocation is stopped after its stated polynomial operation bound. Thus invalid oracles, failed isolation, and arbitrary learner behavior outside the promise do not affect running time. They also cannot cause false acceptance, because acceptance requires an exact-cover certificate.

A satisfiable instance is accepted in one trial with probability $\Omega(1/k_X)$, whereas an unsatisfiable instance is never accepted. Independent $O(k_X)$ repetitions give constant success probability with one-sided error. The resulting X3C algorithm is polynomial in the same Gaussian real-arithmetic model. This argument supplies no finite-bit simulation of exact Gaussian draws or arbitrary real arithmetic, and therefore makes no claim that X3C belongs to ordinary RP.

C Proofs for tractability (precise version of Theorem 4.2 and its proof)

In this section, we provide the fully rigorous statement and proof of the tractability results for Algorithm Ω, corresponding to the informal Theorem 4.2. We do this by analyzing the three stages of Algorithm Ω. Even Hermite moments first determine the hidden weights approximately up to sign. Positive independence then makes the first-moment sign relaxation stable, and anchored regression turns the oriented estimates into exact recovery. The same angular-separation assumption controls both the high-order moment construction and the final regression step.

Throughout this appendix, $W = [w_1, \dots, w_k]$, $r_\ell = \|w_\ell\|_2$, and $u_\ell = w_\ell / r_\ell$. We use the Hermite normalization from the main body. Inner products of matrices and tensors are entrywise inner products; $\|\cdot\|_2$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_{\mathrm{F}}$ denote the Euclidean, operator, and Frobenius norms, respectively. All logarithms are natural. Constants denoted by C are universal and may be enlarged between displays.

C.1 Hermite moments and implicit estimation

For every positive even integer t, the population identity is

$$M_t = \frac{1}{c_t}\mathbb{E}[y\mathcal{H}_t(x)] = \sum_{\ell=1}^{k} r_\ell u_\ell^{\otimes t}, \qquad c_t = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[\mathrm{ReLU}(g)h_t(g)]. \tag{33}$$

Thus an empirical average of $c_t^{-1} y\mathcal{H}_t(x)$ is unbiased for M_t. The benefit of increasing t is geometric: tensor powers replace the inner product $u_i^\top u_j$ by $(u_i^\top u_j)^t$. The diagonal entries of the lifted Gram matrix remain one, while its off-diagonal entries become small. This separates directions that need not be linearly independent in $\mathbb{R}^d$.

Unbiasedness alone, however, does not give a useful high-dimensional error bound. Merely evaluating the raw empirical tensor without storing it does not remove its statistical noise. We instead use the data-dependent compressed estimator of Diakonikolas & Kane (2025). Below, an *implicit estimate* means the representation returned by that procedure, not the uncompressed empirical average.

The implicit-moment primitive. A *sequential tensor computation* is a circuit using additions, scalar multiplications, and successive tensor products with vectors (Diakonikolas & Kane, 2025, Definition 3.1). Proposition 3.2 of (Diakonikolas & Kane, 2025) estimates contractions of positive moment sequences from such circuits, without expanding their tensors. For the unit directions and coefficients $r_\ell \le 1$ in (33), we use the following quantitative specialization, obtained by tracing its proof. Suppose the unbiased source circuits at even orders through $2p$ have directional second moments bounded by $V \ge 1$. If an independent random query tensor $\mathcal{T}$ has the same bound, then

preprocessing

$$n \geq C\frac{(p+1)^7 d^2 k^6 V^4}{\tau\zeta^4} \tag{34}$$

independent labeled samples, split among the required source orders, gives an evaluator $\mathcal{E}_p$ for which, with probability at least $1-\tau$ over preprocessing,

$$\mathbb{E}_{\mathcal{T}}\big[\,|\mathcal{E}_p(\mathcal{T}) - \langle \mathcal{T}, M_p\rangle|^2 \mid \text{preprocessing}\big] \leq \zeta^2. \tag{35}$$

Here $p \geq 2$ is even, $0 < \zeta, \tau < 1$, and a directional second-moment bound means $\mathbb{E}[\langle \mathcal{T}, A\rangle^2] \leq V\|A\|_{\mathrm{F}}^2$ for every deterministic compatible tensor A. The explicit exponents in (34) are a conservative specialization, not the displayed statement of Proposition 3.2. For the query circuits below no auxiliary query averaging is needed. One preprocessed representation at each target order can therefore serve all its queries. The successful source-error event can be chosen jointly for the query distributions used below. Intermediate spaces of dimension less than k are retained in full, so the construction also covers $k > d$.

Application to ReLU observations. For an independent $z \sim \mathcal{N}(0, I_d)$, an appropriate unbiased source is

$$\mathcal{S}_t(x, y, z) = \frac{y}{c_t}\,\mathrm{Re}\left((x + \mathrm{i}z)^{\otimes t}\right). \tag{36}$$

This is the extended Hermite construction of Diakonikolas & Kane (2025, Definition 4.1), scaled to our convention; its real and imaginary parts have sequential circuits of size at most $8(t+1)$. The normalized Hermite tensor and coefficient in that paper both include $1/\sqrt{t!}$, which cancels in their ratio. Consequently $\mathbb{E}\mathcal{S}_t = M_t$. Their Hermite identities and moment bounds (Diakonikolas & Kane, 2025, Lemma 4.3, Corollary 4.9, and Lemma 4.12) give the sufficient bound

$$\mathbb{E}[\langle \mathcal{S}_t, A\rangle^2] \leq 64k^2(t+1)^{5/2}6^t\,\|A\|_{\mathrm{F}}^2\,. \tag{37}$$

We verify this numerical specialization when bounding the sketch errors. The preprocessing in (34) uses empirical averages of compressed source tensors; it does not return their raw empirical contractions.

C.2 Complete specification of Steps 1 and 2

The procedure takes a target relative error $\varepsilon_V \in (0, 1/4]$, a failure probability $\delta_{\mathcal{T}} \in (0, 1/4)$, the bounds r, ρ, the width k, and a batch of $n_{\mathcal{T}}$ independent labeled observations. These parameters are fixed before drawing the random features. The sample requirement is stated in Theorem C.2.

Degree and feature count. Choose

$$m = \min\{s \in 2\mathbb{N} : s \geq 4, \quad 8s^2(k-1)\rho^{s-2} \leq 1\}. \tag{38}$$

Set

$$q = \left\lceil \frac{2^{24}3^m k^4}{\delta_{\mathcal{T}}\varepsilon_V^2} \right\rceil. \tag{39}$$

The condition on m implies $(k-1)\rho^m \leq 1/8$ and will also control the active-gradient covariance (a quantity that will apear later in our proof). It is nonvacuous for every $\rho < 1$, including $k = 1$, when $m = 4$. For later complexity interpretation, the elementary bound

$$m \leq 4 + \frac{16}{1-\rho}\log\frac{16k}{1-\rho} \tag{40}$$

suffices: using $\log\rho \leq -(1-\rho)$, the even integer $2 + 2\lceil 8\log(16k/(1-\rho))/(1-\rho)\rceil$ satisfies the defining inequality in (38).

Random features and matrix weightings. Draw two independent Gaussian feature banks, indexed by $a \in [q]$ and $s \in [m]$, with $g^L_{a,s}, g^R_{a,s} \sim \mathcal{N}(0, I_d)$. Set

$$G^L_a = \bigotimes_{s=1}^{m} g^L_{a,s}, \qquad G^R_a = \bigotimes_{s=1}^{m} g^R_{a,s}. \tag{41}$$

These tensor products are stored as lists of vectors. Independently draw $H_\star = (Z + Z^\top)/2$, where the entries of Z are independent standard Gaussians, and define

$$H^{ii} = e_i e_i^\top, \qquad H^{ij} = \tfrac{1}{2}(e_i e_j^\top + e_j e_i^\top) \quad (1 \le i < j \le d). \tag{42}$$

Thus $H_\star$ is isotropic on the symmetric matrices: $\langle H_\star, A\rangle$ has variance $\|A\|_{\mathrm{F}}^2$ for symmetric A. The random matrix will assign different scalar labels to the components; the coordinate matrices will subsequently read out their coordinates.

Implicit moment sketches. Independently of the feature banks and $H_\star$, preprocess the observations for target orders $m, 2m, 2m+2$. Use independent source batches and the primitive above, with failure probability $\delta_{\mathcal{T}}/64$ at each order and root-mean-square query tolerance

$$\zeta = \frac{r^3 \varepsilon_V \delta_{\mathcal{T}}^{3/2} \sqrt{1-\rho^2}}{2^{32} k^3 d^2 \big(3\sqrt{2\log(64k/\delta_{\mathcal{T}})} + 1\big)}. \tag{43}$$

Construct $\widehat{b} \in \mathbb{R}^q$ and $\widehat{S}_0, \widehat{S}_{H_\star}, \widehat{S}_{H^{ij}} \in \mathbb{R}^{q\times q}$ by evaluating

$$\begin{aligned} \widehat{b}_a &= \mathcal{E}_m(G_a^L), \\ (\widehat{S}_0)_{ab} &= \mathcal{E}_{2m}(G_a^L \otimes G_b^R), \\ (\widehat{S}_H)_{ab} &= \mathcal{E}_{2m+2}(G_a^L \otimes G_b^R \otimes H), \quad H \in \{H_\star\} \cup \{H^{ij} : i \le j\}. \end{aligned} \tag{44}$$

Their population targets, used only in the analysis, are

$$b_a = \big\langle G_a^L, M_m \big\rangle, \qquad (S_0)_{ab} = \big\langle G_a^L \otimes G_b^R, M_{2m} \big\rangle, \tag{45}$$

$$(S_H)_{ab} = \big\langle G_a^L \otimes G_b^R \otimes H, M_{2m+2} \big\rangle. \tag{46}$$

A matrix weighting H thus means contracting the final two tensor indices against H. No unknown population tensor is used as an input.

Separation by a matrix pencil. Let $\widehat{S}_{0,k}$ be the rank-k singular-value truncation of $\widehat{S}_0$. Write $\dagger$ for the Moore–Penrose pseudoinverse and set

$$\beta = 2\sqrt{2\log\frac{64k}{\delta_{\mathcal{T}}}} + 1, \qquad \widehat{A}_\star = (\widehat{S}_{H_\star} + \beta \widehat{S}_0)\widehat{S}_{0,k}^\dagger. \tag{47}$$

Select the k eigenvalues of $\widehat{A}_\star$ with modulus greater than $1/2$, and let $\widehat{z}_1, \dots, \widehat{z}_k$ be corresponding real unit right eigenvectors. The proof in Diakonikolas & Kane (2025) shows that these eigenvalues are real and simple on the success event. Declare failure if the required rank or spectral conditions do not hold. Here, the role of the shift β is to keep the selected eigenvalues separate from the nullspace.

Directions and magnitudes. For $1 \le i \le j \le d$, form

$$\widehat{A}_{ij} = \widehat{S}_{H^{ij}} \widehat{S}_{0,k}^\dagger, \qquad (\widehat{U}_\ell)_{ij} = (\widehat{U}_\ell)_{ji} = \widehat{z}_\ell^\top \widehat{A}_{ij} \widehat{z}_\ell. \tag{48}$$

Take a unit eigenvector $\widehat{u}_\ell$ associated with the largest eigenvalue of the symmetric matrix $\widehat{U}_\ell$. Estimate the magnitudes by

$$\widehat{F} = \sqrt{q}\,[\widehat{z}_1, \dots, \widehat{z}_k], \qquad \widehat{c} \in \arg\min_{c\in\mathbb{R}^k} \left\| \widehat{F}c - \widehat{b} \right\|_2, \qquad \widehat{r}_\ell = |\widehat{c}_\ell|, \tag{49}$$

and return

$$\widehat{v}_\ell = \widehat{r}_\ell \widehat{u}_\ell, \qquad \widehat{V} = [\widehat{v}_1, \dots, \widehat{v}_k].$$

The eigenvector signs are arbitrary at this stage. Since the feature-column norms are nearly one the factor $\sqrt{q}$ in (49) is needed. We note that rank failure in this least-squares system is also treated as an algorithmic failure.

C.3 Formal Statement and Proof Architecture

We state the overall theorem first, followed by the guarantees for its three stages. Throughout, the lifted degree m refers to (38). Concretely, it is a function of k, ρ, not an additional geometric assumption.

Theorem C.1 (Rigorous version of Theorem 4.2)**.** *Assume the planted weights satisfy the Norm bounds (Assumption 3.1 with constant r), Robust non-collinearity (Assumption 3.2 with constant ρ), and Positive independence (Assumption 4.1 with margin η). Let $\delta \in (0, 1/4)$ and choose*

$$\varepsilon_V = \min\left\{\frac{\eta r}{8k}, \quad \frac{\sqrt{\pi/12}}{20^{3/2}k^{3/2}m^{15/4}}\right\}. \tag{50}$$

Use independent batches of sizes $n_{\mathcal{T}}, n_o, n_A$ for the three stages. If $n_{\mathcal{T}}$ satisfies (52) with $\delta_{\mathcal{T}}$ replaced by $\delta/3$, and

$$n_o \geq C\frac{k^2}{\eta^2 r^2}\left(d + \log\frac{3}{\delta}\right), \qquad n_A \geq Ck^8 m^{45/2}\left(kd + \log\frac{3}{\delta}\right), \tag{51}$$

then Algorithm Ω recovers W up to column permutation with probability at least $1-\delta$. The parameters and runtime for Steps 1–2 are those below with the same substitution $\delta_{\mathcal{T}} = \delta/3$. For each fixed $\rho < 1$, the sample requirements and the size of the subsequent convex programs are polynomial in $d, k, 1/r, 1/\eta, 1/\delta$.

Theorem C.2 (Steps 1 and 2: Unoriented atom recovery)**.** *Assume the Norm bounds (Assumption 3.1) and Robust non-collinearity (Assumption 3.2). Fix $\varepsilon_V \in (0, 1/4]$ and $\delta_{\mathcal{T}} \in (0, 1/4)$ and execute the procedure in Section C.2. There is a universal constant $C \geq 1$ such that the sufficient sample condition is*

$$\begin{aligned} n_{\mathcal{T}} \geq C\,&\frac{(2m+3)^7 k^{26} d^{10}(4m+5)^{10} 6^{16m+16}}{r^{12}\varepsilon_V^4 \delta_{\mathcal{T}}^7 (1-\rho^2)^2} \\ &\times \left(3\sqrt{2\log\frac{64k}{\delta_{\mathcal{T}}}} + 1\right)^4. \end{aligned} \tag{52}$$

Under this condition, with probability at least $1-\delta_{\mathcal{T}}$ the procedure succeeds and there are a permutation π and signs $s_1, \ldots, s_k \in \{-1, 1\}$ such that

$$\left\|\widehat{v}_\ell - s_\ell w_{\pi(\ell)}\right\|_2 \leq \varepsilon_V \left\|w_{\pi(\ell)}\right\|_2 \qquad (\ell \in [k]). \tag{53}$$

A conservative dense arithmetic-operation bound is

$$C\left[(2m+3)^2(kd+1)^4(n_{\mathcal{T}} + d^2q^2) + d^2q^3 + kd^3\right], \tag{54}$$

with m, q as in (38)–(39). The dependence on ρ is explicit through m and (52).

We note that in this result ρ is viewed as a fixed constant, we are not claiming a polynomial dependence on $1/(1-\rho)$ when ρ varies.

Theorem C.3 (Steps 3 and 4: Convex sign orientation)**.** *Assume the Norm bounds and Positive independence. After a column permutation, suppose $\widehat{V}$ satisfies $\|\widehat{v}_\ell - s_\ell w_\ell\|_2 \leq \varepsilon_V \|w_\ell\|_2$ for signs $s_\ell \in \{-1, 1\}$. Let $m_1 = \sum_\ell w_\ell$ and suppose $\|\widehat{m}_1 - m_1\|_2 \leq \varepsilon_m$. If*

$$2k\varepsilon_V + 2\varepsilon_m < \eta r, \tag{55}$$

then every minimizer

$$\widehat{z} \in \arg\min_{\|z\|_\infty \leq 1} \left\|\widehat{V}z - \widehat{m}_1\right\|_2$$

satisfies $\operatorname{sign}(\widehat{z}_\ell) = s_\ell$ for all ℓ. Hence $\widehat{w}_\ell = \operatorname{sign}(\widehat{z}_\ell)\widehat{v}_\ell$ obeys $\|\widehat{w}_\ell - w_\ell\|_2 \leq \varepsilon_V \|w_\ell\|_2$.

Theorem C.4 (Step 5: Anchored regression)**.** *Assume Robust non-collinearity. Let $A = [a_1, \ldots, a_k]$ be independent of the regression batch and satisfy $\|a_\ell - w_\ell\|_2 \leq \varepsilon_\ell \|w_\ell\|_2$ with $0 \leq \varepsilon_\ell < 1/2$. Write $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_k)^\top$. If*

$$\|\boldsymbol{\varepsilon}\|_2 \leq \frac{\sqrt{\pi/12}}{20^{3/2}km^{15/4}}, \tag{56}$$

then, for every $t > 0$, an independent Gaussian batch satisfying

$$n_A \geq Ck^8 m^{45/2}(kd + t^2) \tag{57}$$

has the following property with probability at least $1 - \exp(-2t^2)$: W is the unique solution of

$$\begin{aligned} \max_{\overline{W} \in \mathbb{R}^{d \times k}} \quad & \sum_{\ell=1}^{k} \langle a_\ell / \left\| a_\ell \right\|_2, \overline{w}_\ell \rangle \\ \textit{subject to} \quad & \sum_{\ell=1}^{k} \mathrm{ReLU}(\overline{w}_\ell^\top x_i) \leq y_i, \qquad i \in [n_A]. \end{aligned} \tag{58}$$

Below we will show how to combine these results to prove our main theorem (Theorem C.1) before proving each of these theorems individually.

C.3.1 Proof of Theorem C.1

We begin by noting that the accuracy in (50) serves two purposes. Its first term leaves a margin for sign rounding. Its second ensures $\sqrt{k}\,\varepsilon_V \leq \sqrt{\pi/12}/(20^{3/2} k m^{15/4})$, as needed for anchored regression. Since $\eta, r \leq 1$, it also satisfies $\varepsilon_V \leq 1/8$ and is an admissible input to Theorem C.2.

With $\delta_{\mathcal{T}}$ replaced by $\delta/3$ in the parameter choices and (52), Theorem C.2 supplies the required unoriented estimates with failure probability at most $\delta/3$. It remains to check the first-moment accuracy. We have $m_1 = 2\mathbb{E}[yx] = \sum_\ell w_\ell$. For every unit vector v, the centered variable $2y(v^\top x) - v^\top m_1$ is sub-exponential with parameter at most Ck: $y \leq \sum_\ell \left| w_\ell^\top x \right|$ has sub-Gaussian parameter at most Ck, and $v^\top x$ is standard Gaussian.

Bernstein's inequality and a $1/2$-net of the unit sphere, of cardinality at most 5^d, therefore give, for every $a > 0$,

$$\mathbb{P}(\left\| \widehat{m}_1 - m_1 \right\|_2 > a) \leq 2\,5^d \exp\left[-cn_o \min\left\{ a^2/k^2, a/k \right\} \right]$$

for a universal $c > 0$. Set $a = \eta r/8$. The first condition in (51) makes this probability at most $\delta/3$. On the two successful events,

$$2k\varepsilon_V + 2\left\| \widehat{m}_1 - m_1 \right\|_2 \leq 2k\frac{\eta r}{8k} + 2\frac{\eta r}{8} = \frac{\eta r}{2} < \eta r.$$

Theorem C.3 recovers every sign and preserves the relative accuracy of each weight. The resulting anchor matrix is independent of the final batch and satisfies (56). Apply Theorem C.4 with $t = \sqrt{\frac{1}{2}\log(3/\delta)}$. The second condition in (51) makes the final failure probability at most $\delta/3$. A union bound proves exact recovery with probability at least $1 - \delta$.

Finally, (40) implies, for every fixed $b \geq 1$,

$$b^m \leq b^4 \left(\frac{16k}{1-\rho} \right)^{16 \log b/(1-\rho)}.$$

This makes the fixed-ρ polynomial dependence of (52) and (54) explicit after substituting (50). Sign orientation is a box-constrained quadratic program. The anchored program is a linear program after introducing $n_A k$ variables $t_{i\ell}$ with $t_{i\ell} \geq 0$, $t_{i\ell} \geq \overline{w}_\ell^\top x_i$, and $\sum_\ell t_{i\ell} \leq y_i$. Both have polynomial size.

C.4 Proof of Theorem C.2 (Unoriented Atom Recovery)

The proof separates the population algebra from estimation error. The two random feature banks first give well-conditioned coordinates for the hidden lifted vectors. In these coordinates, matrix division cancels one feature bank, leaving an eigenvector problem. We then control the error in this calculation and verify that the implicit estimator provides the required accuracy.

Conditioning of the features. For the feature banks in (41), define the latent matrices $F^L, F^R \in \mathbb{R}^{q\times k}$ by

$$(F^L)_{a\ell} = \prod_{s=1}^{m} (g^L_{a,s})^\top u_\ell, \qquad (F^R)_{a\ell} = \prod_{s=1}^{m} (g^R_{a,s})^\top u_\ell.$$

We note that these are solely for analysis purposes and not quantities computed from the unknown weights. For either feature bank, the expected normalized Gram matrix is $K_m = [(u_i^\top u_j)^m]_{i,j}$. The degree condition gives

$$\tfrac{7}{8} I_k \preceq K_m \preceq \tfrac{9}{8} I_k.$$

Indeed, the sum of the absolute off-diagonal entries in each row is at most $(k-1)\rho^m \le 1/8$.

For one row and two columns, Gaussian fourth moments give

$$\mathbb{E}[((F^L)_{ai}(F^L)_{aj})^2] = \left(1 + 2(u_i^\top u_j)^2\right)^m \le 3^m.$$

Apply Chebyshev's inequality to every Gram entry with tolerance $1/(16k)$, and to every diagonal entry with tolerance $\varepsilon_V/(512k)$. A union bound over both banks, using (39), has failure probability at most

$$\frac{512\, 3^m k^4}{q} + \frac{2^{19} 3^m k^3}{q\varepsilon_V^2} \le \frac{\delta_{\mathcal{T}}}{16}.$$

Consequently, simultaneously,

$$\left\| q^{-1}(F^L)^\top F^L - K_m \right\|_{\mathrm{op}} \le \tfrac{1}{16}, \qquad \left\| q^{-1}(F^R)^\top F^R - K_m \right\|_{\mathrm{op}} \le \tfrac{1}{16}, \tag{59}$$

$$\left| q^{-1} \left\| F^L_{:\ell} \right\|_2^2 - 1 \right| \le \frac{\varepsilon_V}{512k}, \qquad \left| q^{-1} \left\| F^R_{:\ell} \right\|_2^2 - 1 \right| \le \frac{\varepsilon_V}{512k}. \tag{60}$$

In particular, both feature matrices have full column rank and

$$\sqrt{13q/16} \le \sigma_{\min}(F^L), \sigma_{\min}(F^R) \le \sigma_{\max}(F^L), \sigma_{\max}(F^R) \le \sqrt{19q/16}. \tag{61}$$

Population matrix pencil. Let $R = \mathrm{diag}(r_1, \dots, r_k)$. For a symmetric H, define $D_H = \mathrm{diag}(u_1^\top H u_1, \dots, u_k^\top H u_k)$. Expanding (45)–(46) gives

$$b = F^L (r_1, \dots, r_k)^\top, \qquad S_0 = F^L R (F^R)^\top, \qquad S_H = F^L R D_H (F^R)^\top.$$

On (61),

$$\sigma_k(S_0) \ge \tfrac{13}{16} rq, \qquad \|S_0\|_{\mathrm{op}} \le \tfrac{19}{16} q. \tag{62}$$

The full-column-rank factorizations imply $S_0^\dagger = ((F^R)^\dagger)^\top R^{-1} (F^L)^\dagger$, so

$$A_H = S_H S_0^\dagger = F^L D_H (F^L)^\dagger. \tag{63}$$

Thus $F^L_{:\ell}$ is a right eigenvector of A_H with eigenvalue $u_\ell^\top H u_\ell$. This cancellation is the purpose of the two-bank construction. Their independence will also keep the query second moments bounded.

A random weighting separates the eigenvalues. In this proof write

$$\theta_\ell = u_\ell^\top H_\star u_\ell, \qquad B = \sqrt{2 \log \frac{64k}{\delta_{\mathcal{T}}}}, \qquad \gamma = \frac{\delta_{\mathcal{T}} \sqrt{\pi(1-\rho^2)}}{64k^2}.$$

Each θ_ℓ is standard Gaussian. Also $\theta_i - \theta_j$ has variance $\left\| u_i u_i^\top - u_j u_j^\top \right\|_{\mathrm{F}}^2 = 2(1 - (u_i^\top u_j)^2) \ge 2(1-\rho^2)$. Gaussian tail and density bounds, followed by union bounds, give

$$\max_\ell |\theta_\ell| \le B,$$

$$|\theta_i - \theta_j| \ge \gamma \qquad (i \ne j),$$

with total failure probability at most $\delta_{\mathcal{T}}/8$. The second condition is vacuous when $k = 1$. The shifted population matrix is

$$A_\star = (S_{H_\star} + \beta S_0) S_0^\dagger = F^L \,\mathrm{diag}(\theta_\ell + \beta)_{\ell=1}^k (F^L)^\dagger.$$

Its k signal eigenvalues are at least $B + 1$, are pairwise separated by γ, and are separated from its zero eigenvalues.

Perturbation of the pencil. Condition on the preceding feature and Gaussian events. Set (in this paragraph only)

$$e = \frac{r^3 \gamma \varepsilon_V}{2^{20} kd(3B+1)},$$

and suppose

$$\left\|\widehat{S}_0 - S_0\right\|_{\mathrm{op}} \le eq, \qquad \left\|\widehat{S}_H - S_H\right\|_{\mathrm{op}} \le eq \quad (H = H_\star \text{ or } H^{ij}). \tag{64}$$

Singular-value truncation gives $\left\|\widehat{S}_{0,k} - S_0\right\|_{\mathrm{op}} \le 2eq$. Since $e \le r/32$, (62), Weyl's inequality, and the equal-rank pseudoinverse perturbation identity give

$$\left\|\widehat{S}_{0,k}^{\dagger}\right\|_{\mathrm{op}} \le \frac{2}{rq}, \qquad \left\|\widehat{S}_{0,k}^{\dagger} - S_0^{\dagger}\right\|_{\mathrm{op}} \le \frac{32e}{r^2 q}. \tag{65}$$

In particular, the latter follows by decomposing the pseudoinverse difference into its range, row-space, and nullspace terms and bounding each with $\left\|S_0^{\dagger}\right\|_{\mathrm{op}}, \left\|\widehat{S}_{0,k}^{\dagger}\right\|_{\mathrm{op}} \le 2/(rq)$; see Stewart & Sun (1990).

For coordinate matrices, $\|D_{H^{ij}}\|_{\mathrm{op}} \le 1$, whereas $\|S_{H_\star} + \beta S_0\|_{\mathrm{op}} \le 2q(3B+1)$. Expanding the products in (47) and (48) and using (65) therefore yields

$$\left\|\widehat{A}_{ij} - A_{H^{ij}}\right\|_{\mathrm{op}} \le \frac{128e}{r^2}, \qquad \left\|\widehat{A}_\star - A_\star\right\|_{\mathrm{op}} \le \frac{128(3B+1)e}{r^2} = \frac{r\gamma\varepsilon_V}{2^{13}kd}. \tag{66}$$

For clarity, the eigenvector perturbation here is for a diagonalizable, generally nonsymmetric matrix, not a symmetric-matrix perturbation theorem. Complete $F^L/\sqrt{q}$ with an orthonormal basis of its orthogonal complement to obtain a square diagonalizing matrix X. By (61), $\|X\|_{\mathrm{op}} \left\|X^{-1}\right\|_{\mathrm{op}} \le 2$. If $E = \widehat{A}_\star - A_\star$, then the perturbation in these coordinates has norm at most $2\|E\|_{\mathrm{op}}$. The disks of radius $2\|E\|_{\mathrm{op}}$ about the population signal eigenvalues are disjoint and separated from zero. Eigenvalue continuity along $A_\star + sE$, $0 \le s \le 1$, gives exactly one eigenvalue in each signal disk. Each is real because the matrix is real and a nonreal eigenvalue would bring its conjugate into the same disk. Also $\mathrm{rank}(\widehat{A}_\star) \le k$, so all remaining eigenvalues are zero.

In diagonal coordinates, the eigenvector equation bounds the coordinates outside the selected component by $4\|E\|_{\mathrm{op}}/\gamma$ times the full coordinate norm. Transforming back and normalizing gives the conservative bound $32\|E\|_{\mathrm{op}}/\gamma$ for the unit eigenvector error. Consequently, after relabeling and choosing signs $\xi_\ell \in \{-1, 1\}$,

$$\left\|\widehat{z}_\ell - \xi_\ell F_{:\ell}^L / \left\|F_{:\ell}^L\right\|_2\right\|_2 \le \frac{r\varepsilon_V}{256kd}. \tag{67}$$

This also verifies the spectral selection rule in the algorithm.

Recovering the directions. For the matrices in (42), $u^\top H^{ij} u = u_i u_j$. By (63), if $z_\ell = \xi_\ell F_{:\ell}^L / \left\|F_{:\ell}^L\right\|_2$, then $z_\ell^\top A_{H^{ij}} z_\ell = u_{\ell,i} u_{\ell,j}$. Since $\|A_{H^{ij}}\|_{\mathrm{op}} \le 2$, (66) and (67) imply

$$\left|(\widehat{U}_\ell)_{ij} - u_{\ell,i} u_{\ell,j}\right| \le \frac{128e}{r^2} + 4\|\widehat{z}_\ell - z_\ell\|_2 \le \frac{\varepsilon_V}{16d}.$$

Thus $\left\|\widehat{U}_\ell - u_\ell u_\ell^\top\right\|_{\mathrm{F}} \le \varepsilon_V/16$. The rank-one population projector has spectral gap one, so the usual symmetric eigenvector bound gives

$$\min_{s \in \{-1,1\}} \|\widehat{u}_\ell - s u_\ell\|_2 \le \frac{\varepsilon_V}{8}. \tag{68}$$

The signs in this display need not agree with the feature-eigenvector signs ξ_ℓ; the latter are only coordinates for the least-squares system.

Recovering the magnitudes. Define $F_{\mathrm{n}} = \sqrt{q}\,[z_1, \dots, z_k]$. The population vector satisfies $b = F_{\mathrm{n}} c^\star$, with $c_\ell^\star = \xi_\ell r_\ell \left\|F_{:\ell}^L\right\|_2 / \sqrt{q}$. Equations (59)–(60) give

$$\sigma_{\min}(F_{\mathrm{n}}) \geq \tfrac{3}{4}\sqrt{q}, \qquad \|c^\star\|_2 \leq 2\sqrt{k}, \qquad |\,|c_\ell^\star| - r_\ell| \leq \frac{\varepsilon_V}{512k} r_\ell.$$

Also, (67) implies $\left\|\widehat{F} - F_{\mathrm{n}}\right\|_{\mathrm{op}} \leq \sqrt{q}\, r\varepsilon_V/(256\sqrt{k}\, d)$ and $\sigma_{\min}(\widehat{F}) \geq \sqrt{q}/2$. Provided

$$\left\|\widehat{b} - b\right\|_2 \leq \frac{r\varepsilon_V \sqrt{q}}{256}, \tag{69}$$

the least-squares solution satisfies

$$\|\widehat{c} - c^\star\|_2 \leq \frac{2}{\sqrt{q}} \left(\left\|\widehat{b} - b\right\|_2 + \left\|\widehat{F} - F_{\mathrm{n}}\right\|_{\mathrm{op}} \|c^\star\|_2 \right) \leq \frac{r\varepsilon_V}{32}.$$

Taking absolute values and using $r \leq r_\ell$ yields

$$|\widehat{r}_\ell - r_\ell| \leq \frac{\varepsilon_V}{16} r_\ell. \tag{70}$$

Source moments and simultaneous sketch accuracy. We now verify (64) and (69) from the stated sample size. The ReLU coefficient has the value

$$c_t = \frac{(-1)^{t/2-1}(t-3)!!}{\sqrt{2\pi}} \quad (t \geq 2 \text{ even}), \qquad \frac{c_t^2}{t!} \geq \frac{1}{10 t^{5/2}}. \tag{71}$$

For $t = 2$ use $(-1)!! = 1$. For $t = 2p + 2 \geq 4$, the second inequality follows from

$$\frac{c_t^2}{t!} = \frac{1}{2\pi t(t-1)} \frac{\binom{2p}{p}}{4^p}, \qquad \frac{\binom{2p}{p}}{4^p} \geq \frac{1}{2\sqrt{p}}.$$

The binomial inequality follows by induction using the ratio $(2p+1)/(2p+2) \geq \sqrt{p/(p+1)}$; the case $t = 2$ is immediate. The coefficient formula itself follows by integrating the Hermite derivative identity twice on the positive half-line.

To check (37), let $P = \langle A, \mathrm{Re}((x + \mathrm{i}z)^{\otimes t})\rangle / \sqrt{t!}$. For symmetric A, the Gaussian pairing identity gives $\mathbb{E}P^2 = 2^{t-1}\|A\|_{\mathrm{F}}^2$; for an arbitrary A, replace it by its symmetrization, which cannot increase its norm. Gaussian hypercontractivity then gives $(\mathbb{E}P^4)^{1/2} \leq 3^t \mathbb{E}P^2$. Also $(\mathbb{E}y^4)^{1/2} \leq \sqrt{3}\, k^2$. Cauchy–Schwarz and (71) imply

$$\mathbb{E}[\langle \mathcal{S}_t, A\rangle^2] \leq 5\sqrt{3}\, k^2 t^{5/2} 6^t \|A\|_{\mathrm{F}}^2,$$

which is stronger than (37).

Every feature query in (44) has directional second moment at most one. For $G_a^L \otimes G_b^R$, this follows from independence and isotropy of its $2m$ Gaussian vectors. Appending a deterministic H^{ij} preserves the bound because $\left\|H^{ij}\right\|_{\mathrm{F}} \leq 1$; appending $H_\star$ preserves it because its covariance is the orthogonal projection onto symmetric matrices. This includes diagonal sketch entries $a = b$: the left and right banks are still independent.

All required source orders are at most $4m + 4$, so take $V = 64k^2(4m+5)^{5/2}6^{4m+4}$ in (34). Substitute the tolerance in (43), use preprocessing failure probability $\delta_{\mathcal{T}}/64$ at each target order, and sum the three sample requirements. The resulting sufficient condition is (52), after increasing its universal constant.

Condition on successful preprocessing. The common representations are independent of the feature banks and $H_\star$. By (35), each sketch entry has conditional squared error at most ζ^2. There are at most $3d^2$ matrix sketches, so

$$\mathbb{E}\left[\sum_H \left\|\widehat{S}_H - S_H\right\|_{\mathrm{F}}^2 \mid \text{preprocessing}\right] \leq 3d^2 q^2 \zeta^2, \qquad \mathbb{E}[\left\|\widehat{b} - b\right\|_2^2 \mid \text{preprocessing}] \leq q\zeta^2,$$

where the sum also includes S_0. With the e used in the perturbation argument,

$$\frac{\zeta}{e} = \frac{\sqrt{\delta_{\mathcal{T}}}}{64\sqrt{\pi}\, d}.$$

Markov's inequality and $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_{\mathrm{F}}$ therefore imply (64) with failure probability at most $3\delta_{\mathcal{T}}/(4096\pi) < \delta_{\mathcal{T}}/64$. The same inequality implies (69) with failure probability at most $\delta_{\mathcal{T}}/64$. This argument does not require the different sketch entries to be independent.

The feature, Gaussian, preprocessing, and sketch-error failures sum to less than $\delta_{\mathcal{T}}$. On their complement, choose the signs in (68). Equations (68) and (70) give

$$\|\widehat{v}_\ell - s_\ell w_\ell\|_2 \le |\widehat{r}_\ell - r_\ell| + \widehat{r}_\ell \,\|\widehat{u}_\ell - s_\ell u_\ell\|_2 \le \varepsilon_V r_\ell.$$

Restoring the relabeling proves (53).

Computational cost. The recursive source compression has dimension at most kd on each matrix side. For (36), each compressed source can be evaluated by propagating the real and imaginary parts of a rank-one complex tensor; a dense implementation is bounded by $C(2m+3)^2(kd+1)^4 n_{\mathcal{T}}$ operations, including the source-subspace decompositions. A feature query is a sequence of vector products. For a matrix weighting, use $H = \sum_{j=1}^{d} H_{:j} e_j^\top$ and append two vectors for each term. Its evaluation cost is bounded by $C(2m+d+3)(kd+1)^2$, which is dominated by $C(2m+3)^2(kd+1)^4$. There are at most $3d^2q^2 + q$ queries. Finally, dense sketch decompositions and coordinate products cost at most Cd^2q^3, and the k projector decompositions cost at most Ckd^3. This proves (54).

C.5 Proof of Theorem C.3 (Convex Sign Orientation)

After relabeling, write $\widehat{V} = WS + E_V$, where $S = \mathrm{diag}(s_1, \dots, s_k)$ and the columns of E_V have norm at most ε_V. At the true sign vector, the residual obeys

$$\left\|\widehat{V}s - \widehat{m}_1\right\|_2 \le \|E_V s\|_2 + \|\widehat{m}_1 - m_1\|_2 \le k\varepsilon_V + \varepsilon_m.$$

Every minimizer $\widehat{z}$ has at most this residual. Since $\|\widehat{z}\|_\infty \le 1$, the triangle inequality gives

$$\|WS(s-\widehat{z})\|_2 \le \left\|\widehat{V}\widehat{z} - \widehat{m}_1\right\|_2 + \|E_V\widehat{z}\|_2 + \|\widehat{m}_1 - m_1\|_2 \le 2k\varepsilon_V + 2\varepsilon_m. \tag{72}$$

The useful observation is that $\alpha = S(s-\widehat{z})$ has nonnegative entries: $\alpha_\ell = 1 - s_\ell\widehat{z}_\ell \ge 0$. Hence Positive independence applies to this error vector, giving

$$\|W\alpha\|_2 \ge \eta \sum_\ell \alpha_\ell \,\|w_\ell\|_2 \ge \eta r \,\|\alpha\|_\infty . \tag{73}$$

Combining (72), (73), and (55) yields $\|\alpha\|_\infty < 1$. Therefore $s_\ell\widehat{z}_\ell > 0$ for every ℓ. Rounding recovers all signs, and multiplying each estimated column by its recovered sign preserves its relative error. Notice that this argument identifies the signs of the minimizer; it does not require the noisy minimizer itself to be a vertex of the box.

C.6 Proof of Theorem C.4 (Anchored Regression Clean-Up)

The final step of Algorithm Ω converts the approximate oriented estimates into exact recovery through anchored regression (Bahmani & Romberg, 2019). We use the estimates as column-wise anchors $A = [\widehat{w}_1, \dots, \widehat{w}_k]$ and maximize alignment with their normalized directions subject to the observed sample constraints. To prove recovery, we show that every alternative matrix with objective value at least that of the planted matrix must violate one of these constraints.

Roadmap of the proof. The argument combines two ingredients. First, accurate anchors restrict the possible error directions. Lemmas C.6 and C.7 translate the relative errors into bounds on the normalized anchors, and Lemma C.8 then places every candidate error $H = \widetilde{W} - W$ in the cone $\mathcal{A}_{\mathrm{col}}$. Second, the sample constraints must detect every nonzero direction in this cone. Lemma C.5 uses angular separation to lower-bound the second moment of the directional derivative $\langle G_x, H\rangle$. Combining this bound with the cone restriction gives the positive directional margin in Lemma C.9.

Lemma C.10 converts that margin into a lower bound on the probability of a positive directional derivative and controls the complexity needed to make this guarantee uniform over the cone. The final concentration argument therefore ensures that every nonzero candidate error has $\langle G_{x_i}, H\rangle > 0$ for some sample i. Convexity then implies $f_{W+H}(x_i) > f_W(x_i) = y_i$, contradicting feasibility.

We begin with the active-gradient second-moment bound, which supplies the nondegeneracy needed by this argument. Define the uncentered second-moment matrix

$$G_x = [\mathbf{1}\{u_1^\top x \geq 0\}x, \ldots, \mathbf{1}\{u_k^\top x \geq 0\}x],$$
$$g_x = \mathrm{vec}(G_x), \qquad \Sigma = \mathbb{E}[g_x g_x^\top].$$

Although often called the active-gradient covariance, Σ is not centered. Its quadratic form is $\mathrm{vec}(H)^\top \Sigma\, \mathrm{vec}(H) = \mathbb{E}[\langle G_x, H\rangle^2]$.

Lemma C.5 (Active-gradient lower bound)**.** *Under Robust non-collinearity, the degree m in (38) satisfies*

$$\Sigma \succeq \frac{1}{20m^{5/2}} I_{dk}. \tag{74}$$

Proof. Fix $H = [h_1, \ldots, h_k]$ and write $Z_H(x) = \langle G_x, H\rangle = \sum_\ell \mathbf{1}\{u_\ell^\top x \geq 0\} h_\ell^\top x$. We lower-bound its squared L_2 norm by keeping just its degree-m Hermite component. Let Sym denote averaging over tensor-index permutations, and define

$$D_\ell = m\,\mathrm{Sym}(h_\ell \otimes u_\ell^{\otimes(m-1)}) - (m-1)(u_\ell^\top h_\ell) u_\ell^{\otimes m}.$$

Differentiating the single-neuron identity

$$\mathbb{E}[\mathrm{ReLU}(w^\top x)\mathcal{H}_m(x)] = c_m \|w\|_2 \, (w/\|w\|_2)^{\otimes m}$$

at $w = u_\ell$ in direction h_ℓ gives $\mathbb{E}[Z_H \mathcal{H}_m(x)] = c_m \sum_\ell D_\ell$. Differentiation under the expectation is valid: ReLU is Lipschitz, $u_\ell^\top x \neq 0$ almost surely, and the difference quotient is dominated by $\left|h_\ell^\top x\right|$ times a Gaussian polynomial.

Orthogonality of the unnormalized Hermite tensors implies

$$\mathbb{E}[Z_H^2] \geq \frac{c_m^2}{m!} \left\| \sum_\ell D_\ell \right\|_{\mathrm{F}}^2. \tag{75}$$

To control the tensor norm, write $a_\ell = u_\ell^\top h_\ell$, $h_\ell^\perp = h_\ell - a_\ell u_\ell$, and $\widetilde{h}_\ell = h_\ell^\perp + (a_\ell/m)u_\ell$. Then $D_\ell = m\,\mathrm{Sym}(\widetilde{h}_\ell \otimes u_\ell^{\otimes(m-1)})$, $\left\|\widetilde{h}_\ell\right\|_2 \leq \|h_\ell\|_2$, and direct symmetrization gives

$$\|D_\ell\|_{\mathrm{F}}^2 = m\left\|h_\ell^\perp\right\|_2^2 + a_\ell^2 \geq \|h_\ell\|_2^2.$$

For $i \neq j$, the same calculation gives

$$\langle D_i, D_j\rangle = m(\widetilde{h}_i^\top \widetilde{h}_j)(u_i^\top u_j)^{m-1} + m(m-1)(\widetilde{h}_i^\top u_j)(u_i^\top \widetilde{h}_j)(u_i^\top u_j)^{m-2}.$$

Its absolute value is at most $m^2\rho^{m-2}\|h_i\|_2\|h_j\|_2$. Therefore

$$\left\| \sum_\ell D_\ell \right\|_{\mathrm{F}}^2 \geq \left(1 - m^2(k-1)\rho^{m-2}\right)\|H\|_{\mathrm{F}}^2 \geq \tfrac{1}{2}\|H\|_{\mathrm{F}}^2,$$

where we used $\sum_{i\neq j}\|h_i\|_2\|h_j\|_2 \leq (k-1)\|H\|_{\mathrm{F}}^2$ and (38). Finally, (71) and (75) give $\mathbb{E}[Z_H^2] \geq \|H\|_{\mathrm{F}}^2/(20m^{5/2})$, proving the claim. □

The preceding lemma ensures that no nonzero perturbation has an identically vanishing directional derivative. To obtain the one-sided positivity needed to violate a sample constraint, we now use the

accuracy of the anchors. The next two lemmas control their normalized directions; these bounds will restrict candidate errors to $\mathcal{A}_{\text{col}}$ and prevent the mean directional derivative from being too negative.

Condition on the anchor matrix, which is independent of the regression data. Set $b_\ell = a_\ell / \|a_\ell\|_2$ and define

$$\delta_\ell = \sqrt{1-\varepsilon_\ell^2}, \qquad \alpha_\ell = \sqrt{2(1-\delta_\ell)}, \qquad \alpha_{\text{col}} = \left(\sum_\ell \alpha_\ell^2\right)^{1/2}.$$

The notation δ_ℓ here is an anchor-correlation quantity, distinct from the failure probabilities δ and $\delta_{\mathcal{T}}$.

Lemma C.6 (Anchor error)**.** *If $0 \le \varepsilon_\ell < 1/2$, then $\alpha_{\text{col}} \le \sqrt{2}\,\|\varepsilon\|_2$. In particular, (56) implies*

$$\alpha_{\text{col}} \le \frac{\sqrt{\pi/6}}{20^{3/2} k m^{15/4}}. \tag{76}$$

Proof. Rationalizing $1-\sqrt{1-\varepsilon_\ell^2}$ gives $\alpha_\ell^2 = 2\varepsilon_\ell^2/(1+\sqrt{1-\varepsilon_\ell^2}) \le 2\varepsilon_\ell^2$. Sum over the columns. □

Lemma C.7 (Column-wise anchor correlation)**.** *For every ℓ, $\langle b_\ell, u_\ell\rangle \ge \delta_\ell$ and $\|b_\ell - u_\ell\|_2 \le \alpha_\ell$.*

Proof. The relative error inequality gives

$$\langle b_\ell, u_\ell\rangle \ge \frac{1}{2}\left(\frac{\|a_\ell\|_2}{\|w_\ell\|_2} + \frac{1-\varepsilon_\ell^2}{\|a_\ell\|_2 / \|w_\ell\|_2}\right) \ge \sqrt{1-\varepsilon_\ell^2}.$$

The second step is the arithmetic–geometric mean inequality. Since b_ℓ, u_ℓ are unit vectors, $\|b_\ell - u_\ell\|_2^2 = 2 - 2\langle b_\ell, u_\ell\rangle \le \alpha_\ell^2$. □

Define the homogeneous cone

$$\mathcal{A}_{\text{col}} = \left\{ H = [h_1, \dots, h_k] : \sum_\ell \langle u_\ell, h_\ell\rangle \ge -\sum_\ell \alpha_\ell \|h_\ell\|_2 \right\}. \tag{77}$$

Lemma C.8 (Cone containment)**.** *If $\widetilde{W}$ has anchored objective at least that of W, then $H = \widetilde{W} - W$ belongs to $\mathcal{A}_{\text{col}}$.*

Proof. The objective inequality is $\sum_\ell \langle b_\ell, h_\ell\rangle \ge 0$. By Lemma C.7,

$$0 \le \sum_\ell \langle b_\ell, h_\ell\rangle \le \sum_\ell \big(\langle u_\ell, h_\ell\rangle + \alpha_\ell \|h_\ell\|_2\big),$$

which is (77). □

Lemma C.9 (Positive directional margin)**.** *Under Robust non-collinearity and (76),*

$$\mathfrak{m}_{\text{col}}(W, A) := \inf_{H \in \mathcal{A}_{\text{col}},\, \|H\|_{\text{F}}=1} \mathbb{E}[(\langle G_x, H\rangle)_+] \ge \frac{1}{4\sqrt{3}\, 20^{3/2} k m^{15/4}}. \tag{78}$$

Proof. Fix a unit-Frobenius $H \in \mathcal{A}_{\text{col}}$ and write $Z_H = \langle G_x, H\rangle$. The bound (74) gives $\mathbb{E}Z_H^2 \ge (20m^{5/2})^{-1}$. Also,

$$|Z_H| \le \sum_\ell \left|x^\top h_\ell\right| \le \sqrt{k}\, \left\|H^\top x\right\|_2.$$

For $H^\top x \sim \mathcal{N}(0, H^\top H)$,

$$\mathbb{E}\left\|H^\top x\right\|_2^4 = \operatorname{tr}(H^\top H)^2 + 2\operatorname{tr}((H^\top H)^2) \le 3.$$

Hence $\mathbb{E}Z_H^4 \le 3k^2$. Hölder interpolation yields

$$\mathbb{E}\,|Z_H| \ge \frac{(\mathbb{E}Z_H^2)^{3/2}}{(\mathbb{E}Z_H^4)^{1/2}} \ge \frac{1}{\sqrt{3}\, 20^{3/2} k m^{15/4}}.$$

On the other hand, $\mathbb{E}[\mathbf{1}\{u_\ell^\top x \geq 0\}x] = u_\ell/\sqrt{2\pi}$, so the cone condition gives

$$\mathbb{E}Z_H = \frac{1}{\sqrt{2\pi}} \sum_\ell \langle u_\ell, h_\ell \rangle \geq -\frac{\alpha_{\mathrm{col}}}{\sqrt{2\pi}} \geq -\frac{1}{2\sqrt{3}\, 20^{3/2} k m^{15/4}}.$$

Use $(Z_H)_+ = (|Z_H| + Z_H)/2$ to obtain (78). □

Lemma C.10 (Small-ball and complexity bounds)**.** *Under the conditions of Lemma C.9, every $H \in \mathcal{A}_{\mathrm{col}}$ with $\|H\|_{\mathrm{F}} = 1$ satisfies*

$$\mathbb{P}\left(\langle G_x, H\rangle \geq \frac{1}{8\sqrt{3}\, 20^{3/2} k m^{15/4}}\right) \geq \frac{1}{4k}\left(\frac{1}{4\sqrt{3}\, 20^{3/2} k m^{15/4}}\right)^2.$$

For independent Rademacher signs σ_i, independent of the data,

$$C_{n_A} := \mathbb{E} \sup_{\|H\|_{\mathrm{F}}=1} \left\langle H, \frac{1}{\sqrt{n_A}} \sum_{i=1}^{n_A} \sigma_i G_{x_i} \right\rangle \leq \sqrt{kd/2}. \tag{79}$$

Proof. Put $\mu = (4\sqrt{3}\, 20^{3/2} k m^{15/4})^{-1}$. The preceding lemma gives $\mathbb{E}(Z_H)_+ \geq \mu$, and $\mathbb{E}(Z_H)_+^2 \leq \mathbb{E}Z_H^2 \leq k$. Thus the Paley–Zygmund bound gives $\mathbb{P}(Z_H \geq \mu/2) \geq \mu^2/(4k)$. For the complexity bound, Frobenius duality and the independent signs give

$$C_{n_A} \leq \left(\mathbb{E}\,\|G_x\|_{\mathrm{F}}^2\right)^{1/2} = \sqrt{kd/2}.$$

The equality follows from Gaussian reflection symmetry: $\mathbb{E}[\mathbf{1}\{u_\ell^\top x \geq 0\}\,\|x\|_2^2] = d/2$ for every ℓ. □

Completion of the proof of Theorem C.4. The anchor condition and Lemma C.6 ensure that all the preceding lemmas apply. Set

$$\mu = \frac{1}{4\sqrt{3}\, 20^{3/2} k m^{15/4}}, \qquad \tau = \mu/2, \qquad p_\tau = \mu^2/(4k).$$

We spell out the final uniform argument to justify the use of the column-wise cone. Define $\psi_\tau(z) = \min\{1, \max\{0, z/\tau\}\}$. It is $1/\tau$-Lipschitz, takes values in $[0,1]$, and satisfies $\mathbb{E}\psi_\tau(Z_H) \geq p_\tau$ for every unit H in the cone. Symmetrization and the Rademacher contraction inequality bound the expected one-sided empirical deviation by $4C_{n_A}/(\tau\sqrt{n_A})$. Bounded differences then imply, with probability at least $1 - \exp(-2t^2)$,

$$\inf_{H \in \mathcal{A}_{\mathrm{col}},\, \|H\|_{\mathrm{F}}=1} \frac{1}{n_A} \sum_{i=1}^{n_A} \psi_\tau(\langle G_{x_i}, H\rangle) \geq p_\tau - \frac{4C_{n_A}/\tau + t}{\sqrt{n_A}}.$$

The right-hand side is at least $p_\tau/2 > 0$ whenever

$$n_A \geq 4\left(\frac{4C_{n_A} + t\tau}{\tau p_\tau}\right)^2. \tag{80}$$

Using (79), $\tau p_\tau = \mu^3/(8k)$, and $\mu \leq 1$, the right-hand side of (80) is at most $Ck^8 m^{45/2}(kd + t^2)$. Hence (57) suffices.

On this event, every nonzero $H \in \mathcal{A}_{\mathrm{col}}$ has some sample with $\langle G_{x_i}, H\rangle > 0$. But convexity gives

$$f_{W+H}(x_i) \geq f_W(x_i) + \langle G_{x_i}, H\rangle.$$

Thus no such $W + H$ is feasible for (58). Any candidate with objective at least that of W lies in the cone by Lemma C.8; consequently W is the unique optimizer. □