BEYOND OVERPARAMETERIZATION: PROVABLE LEARNING OF INPUT-CONVEX MULTI-LAYER POLYNOMIAL NETWORKS WITH ACTIVE QUERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

The theoretical understanding of multi-layer neural networks is largely confined to overparameterized settings, which obscure parameter identifiability and incur high sample complexity. Neural tangent kernel (NTK) provides a general theory for wide networks, but does not offer efficient sample-complexity guarantees. Recent feature-learning results go beyond kernel methods for single-neuron, multi-index, and hierarchical targets. However, the analysis is often restricted to shallow or specific architectures and to the overparameterized regime. We break this paradigm to achieve parameter-level recovery of deep target networks, albeit by using active data queries. Specifically, we study L-layer polynomial networks with even degree-k monomial activations and nonnegative higher-layer weights. This structure makes the target network input-convex, while the optimization landscape remains highly nonconvex with respect to the parameters. Leveraging input convexity and active queries, we propose **ASPIRE** (**A**ctive **S**am**P**ling for **I**terative **R**ecovery via **E**igendirections), a layerwise sampling-based diagonalization algorithm that recovers all network parameters to δ-accuracy with sample complexity $\widetilde{O}_{k,L}\left(d^{L^2+O(L)}\delta^{-2e}\right)$ in polynomial time. To our knowledge, this is the *first* parameter-recovery guarantee for deep target networks whose exponent grows only polynomially with depth, as well as the *first* justification for the effectiveness of using high-quality data in neural network training, with a remarkably *exponential* separation.

1 INTRODUCTION

Deep neural networks have achieved remarkable empirical success, yet their theoretical understanding remains incomplete, in part because training is highly nonconvex. A major line of theory studies the overparameterized regime, where this difficulty becomes more tractable. In particular, the neural tangent kernel (NTK) framework shows that sufficiently wide networks can be approximated by kernel methods, yielding analyzable training dynamics and provable guarantees (Jacot et al., 2018; Lee et al., 2019; Jiang et al., 2025; Guillen et al., 2026).

However, kernel approximations largely suppress feature learning and can therefore fail to exploit the hierarchical structure of deep networks (Yang & Hu, 2021). Beyond the kernel regime, recent theory has established provable learning of low-dimensional and hierarchical features (Damian et al., 2022; Wang et al., 2024b; Fu et al., 2025; Ren et al., 2026a). However, these analyses are often restricted to shallow or specific network architectures and to the overparameterized regime. Under overparameterization, the model may lose the exact structure of the target, often leading to high sample complexity by ignoring inherent structures, as well as a lack of robustness under diverse environments. This makes exact-parameterization learning particularly appealing. However, there are technical barriers. Existing works typically can only explore the low-dimensional structures in the first or second layer, whereas the structures in the remaining layers cannot be explored. This is because that the input distributions for the rest layers are often much more complex (see more discussion in Section 3.2) and the core technique, Stein's Lemma (Stein, 1981; Yang et al., 2017) cannot be applyied. Thus, it remains widely open how to learn multi-layer networks under exact

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.

parameterization, even when the activation function is polynomial and the input data distribution is designed.

In recent years, the dominant success of large language models has witnessed the effectiveness of using useful data for training, as data quality, selection, and organization can substantially affect model training and downstream performance (Gu et al., 2025; Peng et al., 2025; Sachdeva et al., 2026; Subramanyam et al., 2026), with data access itself treated as part of the learning algorithm. Theoretically, it would be of great interest to provide separation results demonstrating the efficiency of data for neural network training. We study a strong oracle model where the learner can query the target function at arbitrary inputs. We further consider the target drawn from a depth-L polynomial network with non-increasing widths, monomial activations, and nonnegative higher-layer weights. These assumptions endow the target network with a convex structure (Amos et al., 2017), allowing us to explore the strength of data. Input convex networks have been extensively used in scientific problems, where neural networks provide a flexible alternative to traditional methods for approximating operators. Input-convex neural networks offer a principled way to approximate convex functions, with wide-ranging applications spanning optimal transport (Makkuva et al., 2020), single-cell perturbation modeling (Bunne et al., 2023), model predictive control (Wang et al., 2024a), and power-system reliability (Christianson et al., 2025).

The input convexity does not make the training problem convex (Amos et al., 2017), it preserves the core optimization difficulty of deep learning. Instead, we propose a layerwise sampling-based diagonalization algorithm that succesfully recovers all parameters to δ-accuracy with sample complexity $\widetilde{O}_{k,L}\left(d^{L^2+O(L)}\delta^{-2e}\right)$ in polynomial time. This substantially improves the dimension dependence over previous provable approaches for this multi-layer setting. Treating the network as an unstructured degree-k^{L-1} polynomial leads to an $O(d^{k^{L-1}})$ sample complexity through direct polynomial interpolation. Similarly, using kernel methods requires $d^{k^{L-1}-o(1)}$ samples in the non-degenerate case (Ghorbani et al., 2021). Many existing feature-learning guarantees rely on a small intrinsic dimension $r \ll d$. When $r \asymp d$, however, dimension reduction alone does not remove the exponential-in-L growth of the exponent of d in the sample complexity.

The main idea of our proposed algorithm is to construct two second-order moment matrices from active samples drawn uniformly from chosen convex sublevel sets. Under this sampling distribution, the even structure of the network induces coordinatewise sign symmetry, which diagonalizes the corresponding latent moments and allows the joint spectral structure of the two matrices to reveal the hidden directions. This enables layerwise recovery through spectral decomposition, with a two-Hessian variant for the final hidden layer and output weights. We further establish finite-sample error bounds and validate the theory numerically.

As a result, our theory breaks the *doubly exponential*-in-depth sample-complexity barrier of the overparameterized paradigm, reducing the exponent from k^{L-1} to $L^2 + O(L)$. It also demonstrates that exact parameter recovery of certain deep networks is provably achievable under active quesries and provides the *first* justification for the effectiveness of using high-quality data in neural network training with an exponential separation.

2 Related Work

NTK and Feature Learning. Foundational work established the NTK and lazy-training perspective for sufficiently wide neural networks (Jacot et al., 2018; Lee et al., 2019; Arora et al., 2019; Du et al., 2019; Allen-Zhu et al., 2019; Chizat et al., 2019). Subsequent work has further developed this framework by studying initialization, kernel evolution, and finite-width corrections (Chen et al., 2024; Jiang et al., 2025; Guillen et al., 2026). Beyond this regime, provable feature learning encompasses single-neuron and multi-index models (Yehudai & Shamir, 2020; Wu, 2022; Damian et al., 2022; Mousavi-Hosseini et al., 2023), as well as nonlinear and hierarchical features in deeper compositional models (Nichani et al., 2023; Wang et al., 2024b; Fu et al., 2025; Tabanelli et al., 2026; Ren et al., 2026a;b). These results primarily concern prediction, representations, or hidden subspaces. We take the next step toward estimating the target network itself: our algorithm is able to recover all parameters to arbitrarily small error.

Parameter Recovery in Neural Networks. Prior work establishes parameter identifiability up to inherent symmetries for deep neural networks under suitable conditions (Fefferman, 1994; Bona-Pellissier et al., 2023). Turning these structural guarantees into provably efficient reconstruction algorithms, however, is substantially more demanding. For ReLU networks, Rolnick & Körding (2020) exploit the geometry of activation boundaries to reconstruct architectures and parameters, although their analysis does not establish provably efficient end-to-end recovery. For polynomial networks, tensor-decomposition arguments establish generic finite identifiability of deep architectures under suitable conditions (Usevich et al., 2025). Related work provides efficient black-box reconstruction of sums of powers of low-degree polynomials under suitable degree and non-degeneracy conditions (Garg et al., 2020). For deep polynomial networks, existing results either establish identifiability without an efficient recovery procedure, or efficiently recover polynomial components without exploiting their shared hidden-layer structure. Our algorithm bridges both gaps by exploiting this shared structure iteratively: each recovered layer unlocks query access to the next, yielding end-to-end recovery of all network parameters with rigorous complexity guarantees.

Input-Convex Architectures and Sampling Methods. A standard way to guarantee input convexity is to impose architectural conditions such as convex nondecreasing activations and nonnegative higher-layer weights (Amos et al., 2017). Such architectures have found applications in optimal transport, computational biology, chemical engineering, and energy systems (Makkuva et al., 2020; Bunne et al., 2023; Wang et al., 2024a; Christianson et al., 2025). In our setting, convexity permits sampling from convex body using Hit-and-Run, whose polynomial-time mixing is classical (Lovász, 1999; Lovász & Vempala, 2006). Recent analyses sharpen its guarantees through localization and Poincaré-based bounds (Chen & Eldan, 2026; Zhang, 2026; Kook & Vempala, 2026). Our noisy Hit-and-Run analysis further makes this sampling step *robust* to the approximation errors introduced by prefix inversion, enabling its repeated use throughout parameter recovery.

3 Problem Setup and Preliminaries

Consider a depth-L fully connected multi-layer polynomial neural network with architecture

$$d = r_0 \geq r = r_1 \geq r_2 \geq \cdots \geq r_{L-1} > r_L = 1,$$

no biases, and monomial activation $\sigma(t) = t^k$. We focus on non-increasing-width networks with at least two neurons per hidden layer, as it's a common structural condition for ensuring parameter identifiability (Phuong & Lampert, 2020; Usevich et al., 2025). The bias-free formulation yields a homogeneous polynomial network, a simplified but canonical setting widely studied in the theory of polynomial neural networks (Kileel et al., 2019; Kubjas et al., 2024; Finkel et al., 2025; Usevich et al., 2025). Let weight matrices $\mathbf{W}_\ell \in \mathbb{R}^{r_{\ell-1} \times r_\ell}, 1 \leq \ell \leq L-1$, and $\mathbf{W}_L := \mathbf{a} \in \mathbb{R}^{r_{L-1} \times 1}$. Define

$$\mathbf{h}_1(\mathbf{x}) = \mathbf{W}_1^\top \mathbf{x}, \quad \mathbf{h}_\ell(\mathbf{x}) = \mathbf{W}_\ell^\top \sigma(\mathbf{h}_{\ell-1}(\mathbf{x})), \quad 2 \leq \ell \leq L-1,$$

where the activation is applied coordinatewise. The target function is $f^\star(\mathbf{x}) = \mathbf{a}^\top \sigma(\mathbf{h}_{L-1}(\mathbf{x}))$. We are given only real function-value access to $f^\star$: for any $\mathbf{x} \in \mathbb{R}^d$, one query returns $f^\star(\mathbf{x})$. Our goal is to *recover* $\mathbf{W}_1, \ldots, \mathbf{W}_{L-1}$ and $\mathbf{a}$ from such queries, i.e., using a network with the identical architecture to learn every parameter of the target function. To describe multi-layer networks more conveniently, we introduce prefix and suffix networks as follows.

Definition 3.1 (Prefix and suffix networks)**.** For $1 \leq \ell \leq L-1$, define the prefix representation

$$\boldsymbol{\phi}_\ell(\mathbf{x}) := \begin{cases} \mathbf{x}, & \ell = 1, \\ \sigma(\mathbf{h}_{\ell-1}(\mathbf{x})), & 2 \leq \ell \leq L-1. \end{cases}$$

We further define the suffix polynomials recursively by $h^{(L-1)}(\mathbf{u}) := \mathbf{a}^\top \mathbf{u}^{\odot k}$, and, for $1 \leq \ell \leq L-2$, $h^{(\ell)}(\mathbf{u}) := h^{(\ell+1)}\left(\mathbf{W}_{\ell+1}^\top \mathbf{u}^{\odot k}\right)$. Finally, define $F_\ell(\mathbf{y}) := h^{(\ell)}(\mathbf{W}_\ell^\top \mathbf{y}), \mathbf{y} \in \mathbb{R}^{r_{\ell-1}}$. Then, for every $1 \leq \ell \leq L-1$, $f^\star(\mathbf{x}) = F_\ell(\boldsymbol{\phi}_\ell(\mathbf{x}))$.

3.1 Assumptions

In this section, we state the assumptions needed for our analysis.

Assumption 3.2 (Evenness and nonnegativity)**.** *The activation degree $k \geq 4$ is an even integer. Moreover, all higher-layer weight matrices are entrywise nonnegative: $\mathbf{W}_\ell \geq \mathbf{0}, 2 \leq \ell \leq L$.*

The even monomial activation together with nonnegative higher-layer weights ensures that the network is input-convex. Convex activation and nonnegativity is a standard mechanism for enforcing input convexity in neural networks (Amos et al., 2017; Makkuva et al., 2020; Bunne et al., 2023; Wang et al., 2024a; Christianson et al., 2025).We further exclude $k = 2$, since the final suffix then reduces to a quadratic form whose decomposition into the final-hidden-layer directions and output weights is generally non-unique (Ottaviani & Turatti, 2025; Usevich et al., 2025).

Assumption 3.3 (Normalization)**.** *We normalize the first-layer columns by $\|\mathbf{w}_{1,j}\|_2 = 1, 1 \leq j \leq r$. For every $2 \leq \ell \leq L$, we impose $\mathbf{1}_{r_{\ell-1}}^\top \mathbf{w}_{\ell,j} = 1, 1 \leq j \leq r_\ell$.*

Assumption 3.3 serves two purposes. For the hidden layers, it selects a canonical representative from the layerwise rescaling equivalence class induced by homogeneity (Kileel et al., 2019; Usevich et al., 2025). The normalization of the output weights additionally fixes the overall scale of the realized function, in line with standard scale-normalization conventions in neural-network theory (Bach, 2017; Yehudai & Shamir, 2020; Damian et al., 2022). Together with Assumption 3.2, these normalizations keep the network at a dimension-independent scale under Gaussian inputs:

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[f^\star(\mathbf{x})^2\right] = \Theta_{k,L}(1).$$

Assumption 3.4 (Parameter conditioning and relative row balance)**.** *There exist constants $\kappa_0, \mu > 0$ such that each $\mathbf{W}_\ell$, $1 \leq \ell \leq L$, has full column rank, and*

$$\mathrm{cond}(\mathbf{W}_\ell) \leq \kappa_0 \quad \textit{for all } 1 \leq \ell \leq L-1, \qquad \min_{1\leq i\leq r_{\ell-1}} \left(\mathbf{W}_\ell \mathbf{1}_{r_\ell}\right)_i \geq \mu \frac{r_\ell}{r_{\ell-1}} \quad \textit{for all } 2 \leq \ell \leq L.$$

Since column normalization fixes the average row sum at $r_\ell/r_{\ell-1}$, the relative row-balance condition simply prevents any hidden neuron from having vanishingly small aggregate influence on the next layer. The condition-number bound further prevents hidden directions from becoming nearly linearly dependent. Comparable quantitative non-degeneracy assumptions are standard in provable feature-learning analyses (Damian et al., 2022; Mousavi-Hosseini et al., 2023; Nichani et al., 2023; Wang et al., 2024b; Fu et al., 2025).

Assumption 3.5 (Uniform relative spectral gap)**.** *For each $1 \leq \ell \leq L-2$, define the latent sublevel set $H_\ell := \left\{\mathbf{u} \in \mathbb{R}^{r_\ell} : h^{(\ell)}(\mathbf{u}) \leq 1\right\}$, and let $\mathbf{U}_\ell \sim \mathrm{Unif}(H_\ell)$. For $1 \leq i \leq r_\ell$, define $\theta_{\ell,i} := \mathbb{E}[U_{\ell,i}^2]\,\mathbb{E}\left[\left(\partial_i h^{(\ell)}(\mathbf{U}_\ell)\right)^2\right]$. We assume that there exists $\gamma \geq 1$ such that*

$$\frac{\min_{i\neq j} |\theta_{\ell,i} - \theta_{\ell,j}|}{\max_i \theta_{\ell,i}} \geq \gamma^{-1}, \qquad 1 \leq \ell \leq L-2.$$

Heuristically, it requires sufficiently strong mixing among the neurons in the network, thereby ruling out certain non-identifiable cases. If the normalized spectral signatures are in random position, their minimum spacing is typically of order r_ℓ^{-2}. This suggests that it is reasonable to assume that γ scales polynomially with r. Thus, in this sense, it does not affect the dominant term in the complexity bound of theorem 5.1. In Appendix H, we will construct a concrete example that satisfies all of these assumptions, demonstrating that they are non-vacuous.

3.2 Challenges

Despite the structural assumptions above, two fundamental challenges remain: the optimization difficulties of direct gradient methods and the limitations of overparameterized approaches.

Difficulties of Direct Gradient Methods. Although the target network is convex in its input, direct fitting in parameter space remains highly nonconvex. We show that, under our assumptions, there still exist local minima and high-order saddle points in the learning objective, which implies that local search methods such as gradient descent are inefficient.

Proposition 3.6 (Local minima and fourth-order saddle barriers in direct gradient-based fitting)**.** *Consider realizable matched two-layer teacher–student networks with quartic activation $\sigma(t) = t^4$ trained under empirical squared loss. There exist instances satisfying all applicable assumptions in Subsection 3.1 for which direct parameter fitting exhibits, respectively, a non-global local minimum and a non-optimal stationary point whose first descent appears only at fourth order.*

For the quartic teacher $f^\star(\mathbf{x}) = \frac{1}{2}(\mathbf{e}_1^\top \mathbf{x})^4 + \frac{1}{2}(\mathbf{e}_3^\top \mathbf{x})^4$, we construct one instance with a strict non-global local minimum and another with a fourth-order saddle satisfying $\mathcal{L}(t) = \frac{1}{16} - \frac{1}{8}t^4 + \frac{1}{16}t^8$. Thus direct fitting can face both bad local minima and higher-order saddles; more generally, finding a fourth-order local minimum is NP-hard (Anandkumar & Ge, 2016). See Appendix I.1 for details.

Limitations of Overparameterization. Overparameterized methods are not naturally suited to exact parameter recovery: while they may achieve good prediction, the enlarged parameterization obscures the underlying target parameters. Moreover, fixed-kernel methods face a severe sample-complexity barrier. Building on the lower bound of Ghorbani et al. (2021, Theorem 3) for rotationally invariant kernels, we show that the infinite-width NTK of our setting requires $d^{k^{L-1}-o(1)}$ samples.

Proposition 3.7 (Barrier for the NTK)**.** *For every fixed $\eta > 0$, learning our target class with the infinite-width NTK to constant Gaussian population error requires $\Omega_{k,L,\eta}(d^{k^{L-1}-\eta})$ samples.*

The proof, given in Appendix I.2, constructs an admissible target with a nonvanishing degree-k^{L-1} spherical-harmonic component, applies the rotationally invariant kernel lower bound on the sphere, and transfers it to Gaussian inputs using homogeneity.

Recent feature-learning methods improve this scaling by using low-dimensional or hierarchical structure (Damian et al., 2022; Wang et al., 2024b; Fu et al., 2025), but they do not remove the exponential dependence on depth in the exponent of d. In their analysis, by assuming a low-dimensional structure of the first layer, they extract the subspace of the first-layer weights, typically using Stein's Lemma . However, the lifted space loses information needed to identify the original hidden directions. As only the span of the weights can be discovered, one needs to overparameterize the weights in the first layer to guarantee approximation. This makes the input distributions for the subsequent layers complex, as the input often lies on a low-dimensional manifold where Stein's Lemma can no longer be applied to further extract the span of the weights for the remaining dimensions.

4 Algorithm

In this section, we present our recovery algorithm and state our main results. The algorithm recovers each non-final hidden layer via column-space recovery followed by ASPIRE, and then recovers the final hidden layer and output weights via two-Hessian decomposition. We begin with some useful oracles derived from function-value queries, which serve as building blocks for both stages.

4.1 Derived Oracles from Function-Value Queries

Since $f^\star$ is a homogeneous polynomial, function-value access yields several useful derived oracles.

Directional derivatives. We first describe how to recover directional derivatives. For a degree-q polynomial p, fix a point $\mathbf{x}$, a direction $\mathbf{v}$, and an order $0 \le m \le q$. The restriction $\psi(t) := p(\mathbf{x} + t\mathbf{v})$ is univariate of degree at most q, so its values at any $q+1$ distinct nodes $t_0, \dots, t_q$ determine ψ exactly, and hence its m-th derivative $\psi^{(m)}(0) = D^m_{\mathbf{v}} p(\mathbf{x})$. Thus each directional derivative is recovered exactly using $q+1$ function-value queries; in particular, gradients and Hessians follow from first- and second-order directional derivatives.

Complex queries. Function-value access at real inputs can also be used to evaluate the complexification of a homogeneous polynomial. Let p be homogeneous of degree q, and consider a complex point $\mathbf{z} = \mathbf{u} + i\mathbf{v}$. Choose an integer $N > 2q$ and let $\theta_j = 2\pi j/N$, $j = 0, \dots, N-1$. The degree-q

Fourier coefficient of these evaluations then recovers the complex value exactly:

$$p_{\mathbb{C}}(\mathbf{z}) = \frac{2^q}{N} \sum_{j=0}^{N-1} e^{iq\theta_j} p(\mathbf{u}\cos\theta_j + \mathbf{v}\sin\theta_j).$$

Thus, a complex query to $p_{\mathbb{C}}$ can be simulated using $O(q)$ real function-value queries.

Suffix oracles. Given the preceding layers, we can further simulate direct query access to the suffix polynomial F_ℓ. Fix $2 \le \ell \le L-1$, suppose that $\mathbf{W}_1, \dots, \mathbf{W}_{\ell-1}$ are known, and consider a desired suffix query $\mathbf{y} \in \mathbb{R}^{r_{\ell-1}}$. Starting from $\mathbf{z}_{\ell-1} = \mathbf{y}$, we invert the prefix recursively from layer $\ell-1$ down to layer 1. At layer j, choose coordinatewise principal k-th roots $\boldsymbol{\xi}_j \in \mathbb{C}^{r_j}$ satisfying $\boldsymbol{\xi}_j^{\odot k} = \mathbf{z}_j$, and then set $\mathbf{z}_{j-1} = \mathbf{W}_j(\mathbf{W}_j^\top \mathbf{W}_j)^{-1}\boldsymbol{\xi}_j$. Because $\mathbf{W}_j$ has full column rank, this choice satisfies $\mathbf{W}_j^\top \mathbf{z}_{j-1} = \boldsymbol{\xi}_j$, and hence $(\mathbf{W}_j^\top \mathbf{z}_{j-1})^{\odot k} = \mathbf{z}_j$. Repeating this recursion produces a complex input $\mathbf{z}_0 \in \mathbb{C}^d$ whose forward propagation through the first $\ell-1$ layers reaches exactly $\mathbf{y}$. Combining this prefix inversion with the preceding complex-evaluation procedure, we can evaluate $f^\star_{\mathbb{C}}(\mathbf{z}_0) = F_\ell(\mathbf{y})$ using only real function-value queries to $f^\star$.

4.2 Layerwise recovery

We now turn to the core layerwise recovery procedure. Assuming the preceding weights $\mathbf{W}_1, \dots, \mathbf{W}_{\ell-1}$ have been recovered sufficiently accurately, we recover $\mathbf{W}_\ell$ in two stages: first estimate its column space from gradient information, and then apply ASPIRE within the recovered subspace to identify the individual hidden directions.

Stage I: Column-Space Recovery. At suitable query points, the chain rule yields suffix gradients spanning $\mathrm{col}(\mathbf{W}_\ell)$. We choose query points $\mathbf{x}_1, \dots, \mathbf{x}_m$ such that the recovered prefix Jacobians $\widehat{\mathbf{J}}_{\ell,i} := D\widehat{\boldsymbol{\phi}}_\ell(\mathbf{x}_i)$ have full row rank. For each query point $\mathbf{x}_i$, we estimate the suffix gradient without explicitly evaluating the full gradient $\nabla f^\star(\mathbf{x}_i)$. For every $j = 1, \dots, r_{\ell-1}$, set $\mathbf{v}_{i,j} := \widehat{\mathbf{J}}_{\ell,i}^\top \mathbf{e}_j$ and recover the directional derivative $b_{i,j} := D_{\mathbf{v}_{i,j}} f^\star(\mathbf{x}_i)$ from function-value queries using polynomial interpolation stated in Subsection 4.1. Collecting these derivatives into $\mathbf{b}_i := (b_{i,1}, \dots, b_{i,r_{\ell-1}})^\top$, we compute $\widehat{\mathbf{g}}_i := (\widehat{\mathbf{J}}_{\ell,i}\widehat{\mathbf{J}}_{\ell,i}^\top)^{-1}\mathbf{b}_i$, which approximates the corresponding suffix gradient $\mathbf{g}_i$. We then form the empirical second moment $\widehat{\mathbf{M}}_{\mathrm{span},\ell} := \frac{1}{m}\sum_{i=1}^m \widehat{\mathbf{g}}_i\widehat{\mathbf{g}}_i^\top$. Because the true suffix gradients lie in $\mathrm{col}(\mathbf{W}_\ell)$, and sufficiently diverse query points make these gradients span this subspace, the leading r_ℓ-dimensional eigenspace of $\widehat{\mathbf{M}}_{\mathrm{span},\ell}$ estimates $\mathrm{col}(\mathbf{W}_\ell)$. We therefore let $\widehat{\mathbf{B}}_\ell \in \mathbb{R}^{r_{\ell-1}\times r_\ell}$ contain the orthonormal eigenvectors corresponding to the r_ℓ largest eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{span},\ell}$, yielding $\mathrm{col}(\widehat{\mathbf{B}}_\ell) \approx \mathrm{col}(\mathbf{W}_\ell)$.

Stage II: Individual-Direction Recovery. Recovering the column space does not identify the individual hidden directions within it. Writing $\mathbf{W}_\ell = \mathbf{B}_\ell\mathbf{C}_\ell$, ASPIRE (Algorithm 1) resolves this ambiguity by actively designing a sampling distribution that creates an identifiable spectral structure. We sample approximately uniformly from a convex sublevel set of the reduced suffix and form position and gradient second moments. After mapping the samples into the latent coordinates induced by $\mathbf{C}_\ell$, the sublevel set becomes symmetric under independent sign flips of each coordinate. This symmetry cancels all cross-coordinate interactions in both moments, so they share a common diagonal structure in the latent basis. As a result, combining the two moments isolates the latent coordinate directions, which correspond exactly to the columns of $\mathbf{C}_\ell$; Assumption 3.5 ensures that these directions are separated and hence identifiable.

4.3 Final hidden layer and output weights

The final suffix has coincident spectral signatures, so we use two Hessians. Take $\mathbf{y}^{(2)} = \mathbf{1}_{r_{L-2}}$ and $\mathbf{y}^{(1)} = \mathbf{1}_{r_{L-2}} + \tau\mathbf{u}$, where $\mathbf{u} \sim \mathrm{Unif}(S^{r_{L-2}-1})$ and $\tau > 0$ is sufficiently small and fixed. Construct symmetric estimates $\widehat{\mathbf{H}}_j \approx \nabla^2 F_{L-1}(\mathbf{y}^{(j)})$, $j = 1, 2$, by fixed-node interpolation of approximate suffix queries (Subsection 4.1).

Algorithm 1 **A**ctive **S**am**P**ling for **I**terative **R**ecovery via **E**igendirections (**ASPIRE**)

Input: Layer $1 \le \ell \le L-2$, recovered subspace basis $\widehat{\mathbf{B}}_\ell$, simulated suffix oracle $\widehat{F}_\ell$, number of samples m, and sampling accuracy parameters.

Let $G_\ell(\mathbf{z}) := F_\ell(\widehat{\mathbf{B}}_\ell \mathbf{z})$ and $K_\ell^\sharp := \{\mathbf{z} : G_\ell(\mathbf{z}) \le 1\}$ denote the reduced suffix and target sampling body.

Define the available reduced oracle $\widetilde{F}_\ell(\mathbf{z}) := \widehat{F}_\ell(\widehat{\mathbf{B}}_\ell \mathbf{z})$.

for $i = 1, \dots, m$ **do**

Independently run noisy Hit-and-Run from $\mathbf{0}$ using $\widetilde{F}_\ell$, as in Lemma B.4, and obtain endpoint $\mathbf{z}_i$.

Estimate $\nabla G_\ell(\mathbf{z}_i)$ by $\mathbf{q}_i$, using evaluations of $\widetilde{F}_\ell$ and the stable interpolation procedure of Lemma B.6.

end for

Form the empirical moments $\widehat{\mathbf{\Sigma}}_{x,\ell} := \frac{1}{m}\sum_{i=1}^m \mathbf{z}_i \mathbf{z}_i^\top$ and $\widehat{\mathbf{\Sigma}}_{g,\ell} := \frac{1}{m}\sum_{i=1}^m \mathbf{q}_i \mathbf{q}_i^\top$.

Compute normalized eigenvectors $\widehat{\mathbf{v}}_{\ell,1}, \dots, \widehat{\mathbf{v}}_{\ell,r_\ell}$ of $\widehat{\mathbf{A}}_\ell := \widehat{\mathbf{\Sigma}}_{g,\ell}\widehat{\mathbf{\Sigma}}_{x,\ell}$.

for $j = 1, \dots, r_\ell$ **do**

Set $\widehat{\mathbf{u}}_{\ell,j} := \widehat{\mathbf{B}}_\ell \widehat{\mathbf{v}}_{\ell,j}$.

if $\ell = 1$ **then**

Set $\widehat{\mathbf{w}}_{1,j} := \widehat{\mathbf{u}}_{1,j}/\|\widehat{\mathbf{u}}_{1,j}\|_2$, with an arbitrary choice of sign.

else

Set $\widehat{\mathbf{w}}_{\ell,j} := |\widehat{\mathbf{u}}_{\ell,j}|/(\mathbf{1}^\top|\widehat{\mathbf{u}}_{\ell,j}|)$, where the absolute value is coordinatewise.

end if

end for

return $\widehat{\mathbf{W}}_\ell := (\widehat{\mathbf{w}}_{\ell,1}, \dots, \widehat{\mathbf{w}}_{\ell,r_\ell})$.

Output: An estimate of $\mathbf{W}_\ell$, up to column permutation and, for $\ell = 1$, column signs.

Let $\widehat{\mathbf{B}} \in \mathbb{R}^{r_{L-2} \times r_{L-1}}$ contain orthonormal eigenvectors for the r_{L-1} largest eigenvalues of $\widehat{\mathbf{H}}_2$, and set $\widehat{\mathbf{M}}_j = \widehat{\mathbf{B}}^\top \widehat{\mathbf{H}}_j \widehat{\mathbf{B}}$, $j = 1, 2$. Under the prefix-accuracy condition of Appendix F, $\widehat{\mathbf{M}}_2$ is invertible. The eigenvectors of $\widehat{\mathbf{M}}_1 \widehat{\mathbf{M}}_2^{-1}$ estimate the projected hidden directions. Mapping them through $\widehat{\mathbf{B}}$, taking coordinatewise absolute values, and normalizing columns to unit ℓ_1-norm yields $\widehat{\mathbf{W}}_{L-1}$. Finally, set $\widehat{\mathbf{a}} := \dfrac{1}{k(k-1)} \operatorname{diag}\left(\widehat{\mathbf{W}}_{L-1}^\dagger \widehat{\mathbf{H}}_2 \widehat{\mathbf{W}}_{L-1}^{\dagger\top}\right)$.

5 Main results

Combining the preceding steps, we are now ready to state our main parameter recovery theorem, together with two corollaries on generalization error.

Theorem 5.1 (Parameter recovery)**.** *Let $k \ge 4$ be even and $L \ge 3$. Define the conditioning envelope $\kappa := \max\{2, \kappa_0, \mu^{-1}\}$. Suppose that Assumptions 3.2, 3.3, 3.4, and 3.5 hold. For every $0 < \delta < \dfrac{1}{2}$, our algorithm [1] makes at most $\widetilde{O}_{k,L}\left(dr + r^{L^2+O(L)}\gamma^{O(L)}\kappa^{O(L^2)}\delta^{-2\mathrm{e}}\right)$ real function-value queries to $f^\star$ and, with probability at least 0.99, returns $\widehat{\mathbf{W}}_1, \cdots, \widehat{\mathbf{W}}_{L-1}, \widehat{\mathbf{a}}$ satisfying*

$$\max_{1 \le \ell \le L-1} \|\widehat{\mathbf{W}}_\ell - \mathbf{W}_\ell\|_2 \le \delta, \qquad \|\widehat{\mathbf{a}} - \mathbf{a}\|_1 \le \delta,$$

after a consistent relabeling of the hidden neurons and a suitable choice of the first-layer column signs. For fixed k, L, the algorithm runs in polynomial time [2].

Using $r \le d$ and suppressing the dependence on γ and κ gives a simpler bound appearing in abstract and introduction. Gaussian stability then converts parameter accuracy into prediction accuracy, yielding the following corollary.

Corollary 5.2 (Gaussian generalization error)**.** *Under the assumptions of Theorem 5.1, for every $0 < \varepsilon < \dfrac{1}{2}$, our algorithm is able to make at most $\widetilde{O}_{k,L}\left(dr + r^{L^2+O(L)}\gamma^{O(L)}\kappa^{O(L^2)}\varepsilon^{-\mathrm{e}}\right)$ real function-value queries to $f^\star$ and, with probability at least 0.99, returns a polynomial network $\widehat{f}$ with the target architecture satisfying $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left(\widehat{f}(\mathbf{x}) - f^\star(\mathbf{x})\right)^2\right] \le \varepsilon$.*

[1] The algorithm is given the target architecture and the parameters κ_0, μ, γ from Subsection 3.1.

[2] We work in the unit-cost exact real-arithmetic model.

The recovered polynomial network can also be converted into a ReLU network by replacing its monomial activations with piecewise-linear approximations.

Corollary 5.3 (ReLU network). *Under the assumptions of Theorem 5.1, for every $0 < \varepsilon < \frac{1}{2}$, our algorithm, with ReLU postprocessing, makes at most $\widetilde{O}_{k,L}\left(dr + r^{L^2+O(L)}\gamma^{O(L)}\kappa^{O(L^2)}\varepsilon^{-\mathrm{e}}\right)$ real function-value queries to $f^\star$ and, with probability at least 0.99, returns a ReLU network f_{ReLU} with biases satisfying $\mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left(f_{\mathrm{ReLU}}(\mathbf{x}) - f^\star(\mathbf{x})\right)^2\right] \le \varepsilon$. The network has L layers with hidden widths $r_\ell^{\mathrm{ReLU}} = \widetilde{O}_{k,L}\left(r_\ell \varepsilon^{-1/4}\right)$.*

6 Proof Sketch

In this section, we provide a proof sketch of the main parameter recovery theorem. To establish the correctness of our algorithm, we first consider a simple setting that captures the essential idea, and then explain how the sample complexity of the algorithm is derived.

6.1 A simple setting

We consider the population-level first-layer recovery when $d = r$, ignoring column-space estimation, prefix inversion, and finite-sample errors, and directly analyze the correctness of ASPIRE.

Let $K_1 := \{\mathbf{x} : f^\star(\mathbf{x}) \le 1\}$ and draw $\mathbf{X} \sim \mathrm{Unif}(K_1)$. Since $f^\star(\mathbf{x}) = h^{(1)}(\mathbf{W}_1^\top \mathbf{x})$, the change of variables $\mathbf{U} := \mathbf{W}_1^\top \mathbf{X}$ gives $\mathbf{U} \sim \mathrm{Unif}(H_1)$, where $H_1 := \{\mathbf{u} : h^{(1)}(\mathbf{u}) \le 1\}$. Because k is even, $h^{(1)}$ is invariant under coordinatewise sign flips. Hence the distribution of $\mathbf{U}$ has the same symmetry. Let $\mathbf{\Lambda}_1 := \mathbb{E}[\mathbf{U}\mathbf{U}^\top]$. For $i \ne j$, flipping only the i-th coordinate leaves the distribution unchanged but changes U_iU_j to $-U_iU_j$. Therefore $\mathbb{E}[U_iU_j] = -\mathbb{E}[U_iU_j] = 0$, so $\mathbf{\Lambda}_1$ is diagonal, with diagonal entries $\lambda_{1,i} := \mathbb{E}[U_i^2]$. Similarly, let $\mathbf{D}_1 := \mathbb{E}[\nabla h^{(1)}(\mathbf{U})\nabla h^{(1)}(\mathbf{U})^\top]$. From $h^{(1)}(\mathbf{S}\mathbf{u}) = h^{(1)}(\mathbf{u})$ for every diagonal sign matrix $\mathbf{S}$, we have $\nabla h^{(1)}(\mathbf{S}\mathbf{u}) = \mathbf{S}\nabla h^{(1)}(\mathbf{u})$. Thus, for $i \ne j$, flipping the i-th coordinate changes $\partial_i h^{(1)}(\mathbf{U})\partial_j h^{(1)}(\mathbf{U})$ to its negative while preserving its distribution. Hence $(\mathbf{D}_1)_{ij} = 0$, so $\mathbf{D}_1$ is diagonal, with $d_{1,i} := \mathbb{E}[(\partial_i h^{(1)}(\mathbf{U}))^2]$. Since $\mathbf{X} = \mathbf{W}_1^{-\top}\mathbf{U}$ and $\nabla f^\star(\mathbf{X}) = \mathbf{W}_1 \nabla h^{(1)}(\mathbf{U})$, we obtain $\mathbf{\Sigma}_{x,1} = \mathbf{W}_1^{-\top}\mathbf{\Lambda}_1\mathbf{W}_1^{-1}$ and $\mathbf{\Sigma}_{g,1} = \mathbf{W}_1\mathbf{D}_1\mathbf{W}_1^\top$. Therefore

$$\mathbf{\Sigma}_{g,1}\mathbf{\Sigma}_{x,1} = \mathbf{W}_1 \,\mathrm{diag}(\theta_{1,1}, \dots, \theta_{1,r})\mathbf{W}_1^{-1}, \qquad \theta_{1,i} := \lambda_{1,i}d_{1,i}.$$

By Assumption 3.5, the $\theta_{1,i}$'s are distinct. Hence the eigendirections of $\mathbf{\Sigma}_{g,1}\mathbf{\Sigma}_{x,1}$ are exactly the columns of $\mathbf{W}_1$. After normalization, ASPIRE recovers $\mathbf{W}_1$ up to column permutation and sign.

6.2 Dimension and accuracy dependence

Throughout this subsection, we treat k and L as fixed constants and suppress dependence on κ, γ, logarithmic factors, and failure probabilities, focusing on dimension and accuracy. The full proof is technically involved, so we present only the main ideas here and defer the details to the appendix.

Proposition 6.1 (Stable column-space recovery). *Fix $2 \le \ell \le L-2$, and $\varepsilon_B < 1/2$. Suppose all preceding layers satisfy $\|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \lesssim \varepsilon_B r^{-7/2}$. Then algorithm can recover an orthonormal basis $\widehat{\mathbf{B}}_\ell$ such that, after orthogonal alignment, $\|\widehat{\mathbf{B}}_\ell - \mathbf{B}_\ell\|_2 \le \varepsilon_B$ using $\widetilde{O}(r^7\varepsilon_B^{-2})$ queries.*

We choose query points so that the estimated first-layer activations are uniform on $[1,2]^r$, keeping coordinates comparable and the prefix Jacobian well-conditioned. See Proposition E.1.

Decomposing the moment errors. Suppress the layer index. Let $\widehat{\pi}$ be the endpoint distribution of noisy Hit-and-Run. Write $\mathbf{g} = \nabla G$, let $\widehat{\mathbf{g}}$ be its fixed-node interpolation estimate from an oracle with multiplicative error ξ, and let $\mathbb{P}_m$ denote empirical averaging over m independent samples from $\widehat{\pi}$, supported on K. The population moments are $\mathbf{\Sigma}_x = \mathbb{E}_\pi[\mathbf{z}\mathbf{z}^\top]$ and $\mathbf{\Sigma}_g = \mathbb{E}_\pi[\mathbf{g}\mathbf{g}^\top]$. Their finite-sample errors decompose exactly as

$$\widehat{\mathbf{\Sigma}}_x - \mathbf{\Sigma}_x = \underbrace{(\mathbb{P}_m - \mathbb{E}_{\widehat{\pi}})[\mathbf{z}\mathbf{z}^\top]}_{\text{empirical error}} + \underbrace{(\mathbb{E}_{\widehat{\pi}} - \mathbb{E}_\pi)[\mathbf{z}\mathbf{z}^\top]}_{\text{distribution error}},$$

$$\widehat{\mathbf{\Sigma}}_g - \mathbf{\Sigma}_g = \underbrace{(\mathbb{P}_m - \mathbb{E}_{\widehat{\pi}})[\widehat{\mathbf{g}}\widehat{\mathbf{g}}^\top]}_{\text{empirical error}} + \underbrace{\mathbb{E}_{\widehat{\pi}}[\widehat{\mathbf{g}}\widehat{\mathbf{g}}^\top - \mathbf{g}\mathbf{g}^\top]}_{\text{gradient-estimation error}} + \underbrace{(\mathbb{E}_{\widehat{\pi}} - \mathbb{E}_{\pi})[\mathbf{g}\mathbf{g}^\top]}_{\text{distribution error}}.$$

Proposition 6.2 (Accurate estimation of the two moments)**.** *For $\eta < 1/2$, suppose $d_{\mathrm{TV}}(\widehat{\pi}, \pi) := \nu \lesssim \eta r^{-2}$ and the reduced suffix oracle has uniform multiplicative error $\xi \lesssim \eta r^{-5/2}$. Then $m = \widetilde{O}(r^2\eta^{-2})$ independent samples as above give $\|\widehat{\mathbf{\Sigma}}_a - \mathbf{\Sigma}_a\|_2 \le \eta\|\mathbf{\Sigma}_a\|_2$ for both $a = x, g$.*

The distribution and interpolation errors are $O(r^2\nu)$ and $O(r^{5/2}\xi)$, while matrix Bernstein controls the empirical error (Lemma B.7). This gives $O(r^3\eta)$ error per eigendirection (Lemmas B.5 and A.6).

Sampling to the required accuracy. Classical Hit-and-Run mixes in $\widetilde{O}(r^4)$ steps for bodies with radius ratio $O(\sqrt{r})$ (Lovász & Vempala, 2006). For our inexact suffix oracle, we use the following noisy variant.

Proposition 6.3 (Noisy Hit-and-Run)**.** *For $\nu < 1/2$, suppose the reduced oracle has multiplicative error $\xi \lesssim \nu^{1+1/L} r^{-O(1)}$. Starting from the origin, noisy Hit-and-Run produces an endpoint after $O(r^4 \log(r/\nu))$ steps whose distribution has total variation distance at most ν from $\mathrm{Unif}(K)$.*

The multivariate Markov inequality (Wilhelmsen, 1974) controls the polynomial Hessian; together with random-direction anti-concentration and chord comparisons, this yields negative-moment bounds on boundary depth along the exact chain. See Lemmas B.3 and B.4.

Proposition 6.4 (Stable recovery of the first layer)**.** *For any $\varepsilon < 1/2$, the first hidden layer can be recovered with $\|\widehat{\mathbf{W}}_1 - \mathbf{W}_1\|_2 \le \varepsilon$, up to inherent symmetries, using $\widetilde{O}(dr + r^{13}\varepsilon^{-2})$ queries.*

By Proposition C.1, r generic gradients recover the first-layer column space almost surely using $O(dr)$ queries, after which the reduced oracle and gradients are exact. Converting columnwise direction error to operator norm costs $\sqrt{r}$, so take $\eta \asymp \varepsilon r^{-7/2}$ and $\nu \asymp \eta r^{-2}$. Propositions 6.2 and 6.3 then give $m = \widetilde{O}(r^9\varepsilon^{-2})$ endpoints, each costing $\widetilde{O}(r^4)$ queries, proving the claim (Appendix C). Only this initial column-space recovery depends on d; deeper layers use the same sampling argument with approximate prefix inversion.

Proposition 6.5 (Stable recovery of one intermediate layer)**.** *Fix $2 \le \ell \le L - 2$, $\varepsilon < 1/2$, and $\alpha := 1 + 1/L$. Suppose preceding layers are recovered with operator-norm error at most ε_W. If $\varepsilon_W \lesssim \varepsilon^\alpha r^{-\ell/2-O(1)}$, then column-space recovery followed by ASPIRE returns $\widehat{\mathbf{W}}_\ell$ satisfying $\|\widehat{\mathbf{W}}_\ell - \mathbf{W}_\ell\|_2 \le \varepsilon$, up to column permutation, using $\widetilde{O}(r^{14}\varepsilon^{-2})$ additional queries.*

Choose $\varepsilon_B \asymp \varepsilon/r$ and $\eta \asymp \varepsilon r^{-4}$. The prefix condition gives column-space error ε_B, while mapping back and normalizing the recovered directions yields $O(r(\varepsilon_B + r^3\eta)) \le \varepsilon$ (Proposition E.2). Approximate prefix inversion gives $\xi_\ell \lesssim r^{\ell/2}\varepsilon_W$ (Lemma D.3). Taking $\nu \asymp \eta r^{-2} \asymp \varepsilon r^{-6}$ in Proposition 6.3 gives $\varepsilon_W \lesssim \varepsilon^\alpha r^{-\ell/2-O(1)}$, which also ensures gradient accuracy. Proposition 6.2 then needs $m = \widetilde{O}(r^{10}\varepsilon^{-2})$ endpoints, for total cost $\widetilde{O}(r^{14}\varepsilon^{-2})$. Thus $\alpha = 1 + 1/L$ comes from the sampling guarantee, not gradient estimation.

Closing the recovery and accumulating errors. By Proposition F.1, the final two-Hessian stage succeeds with $O(r^2)$ additional queries provided the preceding layers are accurate to $\varepsilon_W \lesssim r^{-L/2-O(1)}\delta$. Set $\tau_{L-2} = c\, r^{-L/2-O(1)}\delta$ and propagate tolerances backwards via Proposition 6.5: $\tau_{\ell-1} \asymp \tau_\ell^\alpha r^{-\ell/2-O(1)}$, where $\alpha = 1 + 1/L$. Unrolling gives $p_\star = \alpha^{L-3} < e$ and $\tau_1^{-1} = \widetilde{O}\Big(r^{L^2/2+O(L)}\delta^{-p_\star}\Big)$. Since τ_1 is smallest, it dominates the recovery costs, yielding $Q_{\mathrm{tot}} \le \widetilde{O}\big(dr + r^{14}\tau_1^{-2}\big) = \widetilde{O}\Big(dr + r^{L^2+O(L)}\delta^{-2p_\star}\Big)$. Finally, $p_\star < e$ gives Theorem 5.1; full bookkeeping is deferred to Appendix G.

7 Conclusion

Our results provide a class of convex functions with multi-layer compositional structure that are efficiently learnable from active queries beyond the NTK regime. This opens a new path to provable learning of deep networks, suggesting active queries can break barriers in overparameterized regime. Numerical experiment details are provided in Appendix J.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and the generation of LaTeX and Python code. All AI-assisted content was carefully checked and verified by the authors, who take full responsibility for the final content and the correctness of the work.

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.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 2017.

Animashree Anandkumar and Rong Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 81–102. PMLR, 2016.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Joachim Bona-Pellissier, François Bachoc, and François Malgouyres. Parameter identifiability of a deep feedforward ReLU neural network. *Machine Learning*, 112(11):4431–4493, 2023. doi: 10.1007/s10994-023-06355-4.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Guhan Chen, Yicheng Li, and Qian Lin. On the impacts of the random initialization in the neural tangent kernel theory. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35909–35944, 2024.

Yuansi Chen and Ronen Eldan. Hit-and-run mixing via localization schemes. *Discrete & Computational Geometry*, 75:747–794, 2026. doi: 10.1007/s00454-025-00808-4.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Nicolas Christianson, Wenqi Cui, Steven Low, Weiwei Yang, and Baosen Zhang. Fast and reliable $N-k$ contingency screening with input-convex neural networks. In *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pp. 527–539. PMLR, 2025.

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

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 2019.

Charles Fefferman. Reconstructing a neural net from its output. *Revista Matemática Iberoamericana*, 10(3):507–555, 1994. doi: 10.4171/RMI/160.

Bella Finkel, Jose Israel Rodriguez, Chenxi Wu, and Thomas Yahl. Activation degree thresholds and expressiveness of polynomial neural networks. *Algebraic Statistics*, 16(2):113–130, 2025. doi: 10.2140/astat.2025.16.113.

Hengyu Fu, Zihao Wang, Eshaan Nichani, and Jason D. Lee. Learning hierarchical polynomials of multiple nonlinear features. In *International Conference on Learning Representations*, 2025.

Ankit Garg, Neeraj Kayal, and Chandan Saha. Learning sums of powers of low-degree polynomials in the non-degenerate case. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 889–899. IEEE, 2020. doi: 10.1109/FOCS46700.2020.00087.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021. doi: 10.1214/20-AOS1990.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. Data selection via optimal control for language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Max Guillen, Philipp Misof, and Jan E. Gerken. Finite-width neural tangent kernels from feynman diagrams. In *Proceedings of the International Conference on Machine Learning*, 2026.

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.

Kaiqi Jiang, Jeremy Cohen, and Yuanzhi Li. Understanding the evolution of the neural tangent kernel at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 38, pp. 36087–36127, 2025.

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yunbum Kook and Santosh S. Vempala. Spectral gaps of hit-and-run and coordinate hit-and-run, 2026.

Kaie Kubjas, Jiayi Li, and Maximilian Wiesmann. Geometry of polynomial neural networks. *Algebraic Statistics*, 15(2):295–328, 2024. doi: 10.2140/astat.2024.15.295.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

László Lovász and Santosh Vempala. Hit-and-run from a corner. *SIAM Journal on Computing*, 35(4): 985–1005, 2006. doi: 10.1137/S009753970544727X.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681. PMLR, 2020.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *International Conference on Learning Representations*, 2023.

Eshaan Nichani, Alex Damian, and Jason D. Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Giorgio Ottaviani and Ettore Teixeira Turatti. Generalized identifiability of sums of squares. *Journal of Algebra*, 661:641–656, 2025. doi: 10.1016/j.jalgebra.2024.07.052.

Ru Peng, Kexin Yang, Yawen Zeng, Junyang Lin, Dayiheng Liu, and Junbo Zhao. DataMan: Data manager for pre-training large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mary Phuong and Christoph H. Lampert. Functional vs. parametric equivalence of ReLU networks. In *International Conference on Learning Representations*, 2020.

Yunwei Ren, Yatin Dandi, Florent Krzakala, and Jason D. Lee. Provable learning of random hierarchy models and hierarchical shallow-to-deep chaining. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5556–5597. PMLR, 2026a.

Yunwei Ren, Zihao Wang, and Jason D. Lee. Algorithmic separation between constant-depth and logarithmic-depth neural networks, 2026b.

David Rolnick and Konrad P. Körding. Reverse-engineering deep ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8178–8187. PMLR, 2020.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Cheng. How to train data-efficient LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Anirudh Subramanyam, Yuxin Chen, and Robert L. Grossman. Scaling laws revisited: Modeling the role of data quality in language model pretraining. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hugo Tabanelli, Yatin Dandi, Luca Pesce, and Florent Krzakala. Deep learning of compositional targets with hierarchical spectral methods. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Konstantin Usevich, Ricardo Borsoi, Clara Dérand, and Marianne Clausel. Identifiability of deep polynomial neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Wenlong Wang, Haohao Zhang, Yujia Wang, Yuhe Tian, and Zhe Wu. Fast explicit machine learning-based model predictive control of nonlinear processes using input convex neural networks. *Industrial & Engineering Chemistry Research*, 63(40):17279–17293, 2024a. doi: 10.1021/acs.iecr.4c02257.

Zihao Wang, Eshaan Nichani, and Jason D. Lee. Learning hierarchical polynomials with three-layer neural networks. In *International Conference on Learning Representations*, 2024b.

Don R. Wilhelmsen. A markov inequality in several dimensions. *Journal of Approximation Theory*, 11(3):216–220, 1974. doi: 10.1016/0021-9045(74)90012-4.

Lei Wu. Learning a single neuron for non-monotonic activation functions. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4178–4197. PMLR, 2022.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737, 2021.

Zhuoran Yang, Krishnakumar Balasubramanian, Zhaoran Wang, and Han Liu. Estimating high-dimensional non-gaussian multiple index models via stein's lemma. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3756–3786. PMLR, 2020.

Ruizhe Zhang. Hit-and-run mixes as fast as the ball walk, 2026.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Problem Setup and Preliminaries **3**

3.1 Assumptions 3

3.2 Challenges 4

4 Algorithm **5**

4.1 Derived Oracles from Function-Value Queries 5

4.2 Layerwise recovery 6

4.3 Final hidden layer and output weights 6

5 Main results **7**

6 Proof Sketch **8**

6.1 A simple setting 8

6.2 Dimension and accuracy dependence 8

7 Conclusion **9**

A Technical Tools for Stability Analysis **14**

A.1 Consequences of normalization and conditioning 14

A.2 Basic perturbation tools 15

A.3 Perturbation of two-matrix diagonalization 16

B Convex-body sampling and moment estimation **18**

B.1 Geometry of the reduced sampling body 18

B.2 Hit-and-Run with a multiplicative oracle 20

B.3 Population moments and spectral conditioning 25

B.4 Stable gradient and empirical moment estimation 27

C Recovery of the first hidden layer **30**

D Stability of the simulated suffix oracle **32**

E Recovery of intermediate hidden layers **36**

E.1 Stable recovery of the intermediate-layer column space 36

E.2 Stable recovery within the recovered subspace 42

F Recovery of the final hidden layer and output weights **43**

G Proof of the main theorem **47**

H A Weight-Based Sufficient Condition for the Assumptions **52**

H.1 Normalization and nonnegativity 52

H.2 Conditioning and relative row balance 52

H.3 Uniform relative spectral gap 53

H.4 An explicit nonempty family 56

I Proof of Challenges **56**

I.1 Local minima and fourth-order saddle barriers 56

I.2 Sample-complexity barrier for the NTK 57

J Experiment Details **61**

Roadmap of the appendix. The appendix is organized as follows. Appendix A collects several mostly standard perturbation bounds and numerical linear-algebra tools. Appendix B contains the main sampling machinery of the paper. In particular, we develop a quantitative analysis of Hit-and-Run when the defining homogeneous polynomial is accessed only through a multiplicatively noisy oracle, and use it to obtain the moment estimates required by the spectral recovery procedure. Appendix C applies these tools to the first hidden layer, where the relevant column space can be recovered exactly. Appendix D then studies the error introduced by approximately inverting the recovered prefix, which allows us to simulate access to the suffix networks at subsequent layers. Appendix E combines the subspace, sampling, and suffix-oracle analyses to establish stable recovery of the intermediate hidden layers. Appendix F treats the final hidden layer and the output weights separately using a two-Hessian spectral argument. Finally, Appendix G combines the layerwise guarantees to prove Theorem 5.1, Corollaries 5.2 and 5.3.

A Technical Tools for Stability Analysis

Throughout the section, we work under the assumptions stated in subsection 3.1. We first record several quantitative consequences that will be used in the perturbation analysis.

A.1 Consequences of normalization and conditioning

The following lemma translates the normalization and conditioning assumptions into explicit singular-value and right-inverse bounds for the hidden weight matrices.

Lemma A.1 (Singular-value bounds)**.** *We have* $\frac{1}{\kappa_0} \le \sigma_{\min}(\mathbf{W}_1) \le 1 \le \sigma_{\max}(\mathbf{W}_1) \le \kappa_0$, *and for every* $2 \le \ell \le L-1$, $\frac{1}{\kappa_0}\sqrt{\frac{r_\ell}{r_{\ell-1}}} \le \sigma_{\min}(\mathbf{W}_\ell) \le \sigma_{\max}(\mathbf{W}_\ell) \le \kappa_0$. *Consequently,*

$$\left\|\mathbf{W}_\ell(\mathbf{W}_\ell^\top \mathbf{W}_\ell)^{-1}\right\|_2 \le \kappa_0\sqrt{\frac{r_{\ell-1}}{r_\ell}}, \qquad 2 \le \ell \le L-1.$$

Proof. For $\mathbf{W}_1$, every column has been normalized, therefore $\|\mathbf{W}_1\|_F^2 = \sum_{j=1}^r \|\mathbf{w}_{1,j}\|_2^2 = r$. Since the squared Frobenius norm equals the sum of the squared singular values, $\|\mathbf{W}_1\|_F^2 = \sum_{j=1}^r \sigma_j(\mathbf{W}_1)^2$. Consequently, the average squared singular value is one, so $\sigma_{\min}(\mathbf{W}_1) \le 1 \le \sigma_{\max}(\mathbf{W}_1)$. Since $\mathrm{cond}(\mathbf{W}_1) = \frac{\sigma_{\max}(\mathbf{W}_1)}{\sigma_{\min}(\mathbf{W}_1)} \le \kappa_0$, we obtain

$$\frac{1}{\kappa_0} \le \sigma_{\min}(\mathbf{W}_1) \le 1 \le \sigma_{\max}(\mathbf{W}_1) \le \kappa_0.$$

Now fix $2 \le \ell \le L-1$. Since $\mathbf{W}_\ell \ge \mathbf{0}$ and every column has unit ℓ_1-norm, $\mathbf{W}_\ell^\top \mathbf{1}_{r_{\ell-1}} = \mathbf{1}_{r_\ell}$. Testing $\mathbf{W}_\ell^\top$ on the all-ones vector yields

$$\sigma_{\max}(\mathbf{W}_\ell) \ge \frac{\|\mathbf{W}_\ell^\top \mathbf{1}_{r_{\ell-1}}\|_2}{\|\mathbf{1}_{r_{\ell-1}}\|_2} = \sqrt{\frac{r_\ell}{r_{\ell-1}}}.$$

The condition-number bound therefore gives $\sigma_{\min}(\mathbf{W}_\ell) \ge \frac{1}{\kappa_0}\sqrt{\frac{r_\ell}{r_{\ell-1}}}$. For each column, $\|\mathbf{w}_{\ell,j}\|_2 \le \|\mathbf{w}_{\ell,j}\|_1 = 1$, and hence $\|\mathbf{W}_\ell\|_F^2 \le r_\ell$. At the same time,

$$\|\mathbf{W}_\ell\|_F^2 = \sum_{j=1}^{r_\ell} \sigma_j(\mathbf{W}_\ell)^2 \ge r_\ell \sigma_{\min}(\mathbf{W}_\ell)^2 \ge \frac{r_\ell}{\kappa_0^2}\sigma_{\max}(\mathbf{W}_\ell)^2.$$

Combining the two displays gives $\sigma_{\max}(\mathbf{W}_\ell) \le \kappa_0$. The right-inverse bound follows from

$$\left\|\mathbf{W}_\ell(\mathbf{W}_\ell^\top \mathbf{W}_\ell)^{-1}\right\|_2 = \sigma_{\min}(\mathbf{W}_\ell)^{-1}.$$

Thus the relative conditioning is dimension independent, while the absolute singular scale tracks the aspect ratio $r_\ell/r_{\ell-1}$. □

The next lemma makes the scale of the output weights explicit from normalization and the relative row-balance condition.

Lemma A.2 (Explicit output-weight scale)**.** *Under the unified normalization and relative row-balance condition, let*

$$s := r_{L-1}, \qquad a_{\min} := \min_{1\le i\le s} a_i, \qquad a_{\max} := \max_{1\le i\le s} a_i.$$

Then

$$\frac{\mu}{s} \le a_{\min} \le \frac{1}{s} \le a_{\max} \le 1 - \frac{(s-1)\mu}{s} \le 1.$$

Proof. For $\ell = L$, Assumption 3.4 reads $a_i \ge \mu/s$ for every i, which gives the first inequality. The middle two inequalities follow from $\sum_i a_i = 1$. Finally, after fixing an index attaining $a_{\max}$, the remaining $s-1$ entries each contribute at least μ/s, and therefore $a_{\max} \le 1 - \frac{(s-1)\mu}{s} \le 1$. □

A.2 Basic perturbation tools

The following lemma controls how errors in a full-column-rank matrix propagate to its pseudoinverse.

Lemma A.3 (Perturbation of the pseudoinverse)**.** *Let $\mathbf{A}, \mathbf{B}$ be full column rank. If $\|\mathbf{A}-\mathbf{B}\|_2 \le \frac{1}{2}\sigma_{\min}(\mathbf{A})$, then $\|\mathbf{A}^\dagger - \mathbf{B}^\dagger\|_2 \le \frac{6\|\mathbf{A}-\mathbf{B}\|_2}{\sigma_{\min}(\mathbf{A})^2}$.*

Proof. By Weyl's inequality, $\sigma_{\min}(\mathbf{B}) \ge \sigma_{\min}(\mathbf{A}) - \|\mathbf{A}-\mathbf{B}\|_2 \ge \frac{1}{2}\sigma_{\min}(\mathbf{A})$, and hence $\|\mathbf{B}^\dagger\|_2 \le 2\|\mathbf{A}^\dagger\|_2$. For full-column-rank matrices,

$$\mathbf{B}^\dagger - \mathbf{A}^\dagger = -\mathbf{B}^\dagger(\mathbf{B}-\mathbf{A})\mathbf{A}^\dagger + \mathbf{B}^\dagger\mathbf{B}^{\dagger\top}(\mathbf{B}-\mathbf{A})^\top(\mathbf{I}-\mathbf{A}\mathbf{A}^\dagger).$$

Since $\mathbf{I}-\mathbf{A}\mathbf{A}^\dagger$ is an orthogonal projector,

$$\|\mathbf{A}^\dagger - \mathbf{B}^\dagger\|_2 \le \left(\|\mathbf{A}^\dagger\|_2\|\mathbf{B}^\dagger\|_2 + \|\mathbf{B}^\dagger\|_2^2\right)\|\mathbf{A}-\mathbf{B}\|_2.$$

Substituting the preceding bound on $\|\mathbf{B}^\dagger\|_2$ and $\|\mathbf{A}^\dagger\|_2 = \sigma_{\min}(\mathbf{A})^{-1}$ proves the claim. □

The next lemma converts relative errors in a collection of spanning vectors into an error bound for the recovered orthonormal basis of their column space.

Lemma A.4 (Noisy recovery of an orthonormal basis)**.** *Let* $\mathbf{g}_1, \dots, \mathbf{g}_m \in \mathbb{R}^n$ *satisfy* $\dim(\mathrm{span}\{\mathbf{g}_1, \dots, \mathbf{g}_m\}) = p$*, and define* $\mathbf{G} = [\mathbf{g}_1, \dots, \mathbf{g}_m] \in \mathbb{R}^{n\times m}$*. Let* $\mathbf{B} \in \mathbb{R}^{n\times p}$ *be an orthonormal basis of* $\mathrm{col}(\mathbf{G})$*, so that* $\mathbf{B}^\top\mathbf{B} = \mathbf{I}_p$*. Suppose that we observe* $\widehat{\mathbf{g}}_1, \dots, \widehat{\mathbf{g}}_m$ *satisfying*

$$\|\widehat{\mathbf{g}}_i - \mathbf{g}_i\|_2 \le \eta\|\mathbf{g}_i\|_2, \qquad i = 1, \dots, m.$$

Let $\mathrm{cond}(\mathbf{G}) := \dfrac{\sigma_1(\mathbf{G})}{\sigma_p(\mathbf{G})}$, $\Delta_{\mathrm{sub}} := \eta\,\mathrm{cond}(\mathbf{G})\sqrt{p}$*. If* $\Delta_{\mathrm{sub}} < 1$*, then one can recover an orthonormal matrix* $\widehat{\mathbf{B}} \in \mathbb{R}^{n\times p}$*, with* $\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}} = \mathbf{I}_p$*, such that* $\|\widehat{\mathbf{B}} - \mathbf{B}\|_2 \le \dfrac{\sqrt{2}\,\eta\,\mathrm{cond}(\mathbf{G})\sqrt{p}}{1 - \eta\,\mathrm{cond}(\mathbf{G})\sqrt{p}}$*. Here the basis is understood up to an orthogonal transformation within the recovered* p*-dimensional subspace.*

Proof. Define

$$\widehat{\mathbf{G}} = [\widehat{\mathbf{g}}_1, \dots, \widehat{\mathbf{g}}_m] = \mathbf{G} + \mathbf{E}, \qquad \mathbf{E} = [\widehat{\mathbf{g}}_1 - \mathbf{g}_1, \dots, \widehat{\mathbf{g}}_m - \mathbf{g}_m].$$

By the assumed relative error bound, $\|\mathbf{E}\|_2 \le \|\mathbf{E}\|_F \le \eta\|\mathbf{G}\|_F$. Since $\mathrm{rank}(\mathbf{G}) = p$, $\|\mathbf{G}\|_F \le \sqrt{p}\,\sigma_1(\mathbf{G})$, and therefore

$$\|\mathbf{E}\|_2 \le \eta\sqrt{p}\,\sigma_1(\mathbf{G}) = \Delta_{\mathrm{sub}}\,\sigma_p(\mathbf{G}).$$

Compute the top p left singular vectors of $\widehat{\mathbf{G}}$ and collect them as

$$\widetilde{\mathbf{B}} \in \mathbb{R}^{n\times p}, \qquad \widetilde{\mathbf{B}}^\top\widetilde{\mathbf{B}} = \mathbf{I}_p.$$

Let $\widehat{\mathcal{S}} = \mathrm{col}(\widetilde{\mathbf{B}})$ and $\mathcal{S} = \mathrm{col}(\mathbf{B})$. By the singular-subspace perturbation theorem,

$$\|\mathbf{P}_{\widehat{\mathcal{S}}} - \mathbf{P}_{\mathcal{S}}\|_2 \le \frac{\|\mathbf{E}\|_2}{\sigma_p(\mathbf{G}) - \|\mathbf{E}\|_2} \le \frac{\Delta_{\mathrm{sub}}}{1 - \Delta_{\mathrm{sub}}}.$$

Finally, align the two orthonormal bases by an orthogonal Procrustes transformation. Namely, let $\mathbf{B}^\top\widetilde{\mathbf{B}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a singular value decomposition and set $\mathbf{Q} = \mathbf{V}\mathbf{U}^\top$, $\widehat{\mathbf{B}} = \widetilde{\mathbf{B}}\mathbf{Q}$. Then

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_2 = 2\sin\left(\frac{\theta_{\max}}{2}\right) \le \sqrt{2}\sin\theta_{\max} = \sqrt{2}\|\mathbf{P}_{\widehat{\mathcal{S}}} - \mathbf{P}_{\mathcal{S}}\|_2,$$

where $\theta_{\max}$ is the largest principal angle. Substitution of the projector bound proves the claim. □

The following corollary records the simpler linear error bound used when the perturbation is sufficiently small.

Corollary A.5. *Under the assumptions of Lemma A.4, if* $\eta\,\mathrm{cond}(\mathbf{G})\sqrt{p} \le \dfrac{1}{2}$*, then thereexists an orthonormal basis* $\widehat{\mathbf{B}}$ *of the recovered subspace such that*

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_2 \le 2\sqrt{2}\,\eta\,\mathrm{cond}(\mathbf{G})\sqrt{p} = O(\eta\,\mathrm{cond}(\mathbf{G})\sqrt{p})\,.$$

A.3 Perturbation of two-matrix diagonalization

The final lemma provides the spectral perturbation guarantee used to recover hidden directions from noisy pairs of moment matrices.

Lemma A.6 (Noisy two-matrix diagonalization)**.** *Let* $n \ge 2$ *be an integer and let*

$$\mathbf{M}_1 = \mathbf{C}^{-\top}\mathbf{\Lambda}\mathbf{C}^{-1}, \mathbf{M}_2 = \mathbf{C}\mathbf{D}\mathbf{C}^\top,$$

where $\mathbf{C} \in GL_n(\mathbb{R})$ *and* $\mathbf{D}, \mathbf{\Lambda}$ *are positive diagonal matrices. Write* $\mathbf{D}\mathbf{\Lambda} = \mathrm{diag}(\theta_1, \dots, \theta_n)$*, and suppose that* $\gamma \ge 1$ *satisfies the inverse-gap condition*

$$\frac{\min_{i\ne j}|\theta_i - \theta_j|}{\max_i|\theta_i|} \ge \gamma^{-1}.$$

Let $\mathbf{v}_i := \dfrac{\mathbf{c}_i}{\|\mathbf{c}_i\|_2}, \mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n]$*, and define* $\rho := \dfrac{\|\mathbf{M}_1\|_2\|\mathbf{M}_2\|_2}{\max_i|\theta_i|}$*. Suppose that* $\widehat{\mathbf{M}}_1, \widehat{\mathbf{M}}_2 \in \mathbb{R}^{n\times n}$ *satisfy* $\|\widehat{\mathbf{M}}_j - \mathbf{M}_j\|_2 \le \eta\|\mathbf{M}_j\|_2, j = 1, 2$*. If* $\eta \le \dfrac{c}{\gamma\,\mathrm{cond}(\mathbf{V})^2\rho}$ *for a sufficiently small*

absolute constant $c > 0$, then the eigenvectors of $\widehat{\mathbf{M}}_2\widehat{\mathbf{M}}_1$ can be labeled so that, after choosing their signs,

$$\max_{1\le i\le n} \|\widehat{\mathbf{v}}_i - \mathbf{v}_i\|_2 \le C\gamma\,\mathrm{cond}(\mathbf{V})^2\rho\eta,$$

where $\widehat{\mathbf{v}}_i$ denotes the normalized recovered eigenvector and $C > 0$ is an absolute constant.

Proof. Set $\mathbf{A} = \mathbf{M}_2\mathbf{M}_1 = \mathbf{C}(\mathbf{D}\mathbf{\Lambda})\mathbf{C}^{-1}$. Thus the column directions of $\mathbf{C}$ are precisely the eigendirections of $\mathbf{A}$. Write $\widehat{\mathbf{M}}_j = \mathbf{M}_j + \mathbf{E}_j$, $\|\mathbf{E}_j\|_2 \le \eta\|\mathbf{M}_j\|_2$. Then $\widehat{\mathbf{M}}_2\widehat{\mathbf{M}}_1 = \mathbf{A} + \mathbf{\Delta}$, where

$$\mathbf{\Delta} = \mathbf{M}_2\mathbf{E}_1 + \mathbf{E}_2\mathbf{M}_1 + \mathbf{E}_2\mathbf{E}_1.$$

Since $\max_i |\theta_i| \le \|\mathbf{A}\|_2 \le \|\mathbf{M}_2\|_2\|\mathbf{M}_1\|_2$, we have $\rho \ge 1$. Together with $\gamma \ge 1$ and $\mathrm{cond}(\mathbf{V}) \ge 1$, the assumed upper bound on η therefore implies $\eta \le 1$ after taking $c \le 1$. Hence

$$\|\mathbf{\Delta}\|_2 \le 3\eta\|\mathbf{M}_1\|_2\|\mathbf{M}_2\|_2 = 3\rho\eta\max_i |\theta_i|.$$

The matrix $\mathbf{A}$ is diagonalizable with normalized eigenvector matrix $\mathbf{V}$: $\mathbf{A} = \mathbf{V}\mathbf{\Theta}\mathbf{V}^{-1}, \mathbf{\Theta} = \mathrm{diag}(\theta_1, \dots, \theta_n)$. Let $g := \min_{i\ne j} |\theta_i - \theta_j|$. The inverse-gap hypothesis gives $g \ge \gamma^{-1}\max_i |\theta_i|$. Set $\mathbf{E} := \mathbf{V}^{-1}\mathbf{\Delta}\mathbf{V}$. Then $\|\mathbf{E}\|_2 \le \mathrm{cond}(\mathbf{V})\|\mathbf{\Delta}\|_2$. By the assumed upper bound on η, after decreasing c if necessary, $\|\mathbf{E}\|_2 \le \dfrac{g}{4}$. Since $\mathbf{A} + \mathbf{\Delta} = \mathbf{V}(\mathbf{\Theta} + \mathbf{E})\mathbf{V}^{-1}$, it suffices to study the eigenvectors of $\mathbf{\Theta} + \mathbf{E}$. By the Bauer–Fike theorem, every eigenvalue of $\mathbf{\Theta} + t\mathbf{E}$, $0 \le t \le 1$, lies in

$$\bigcup_{i=1}^{n} \{z \in \mathbb{C} : |z - \theta_i| \le \|\mathbf{E}\|_2\}.$$

Since $\|\mathbf{E}\|_2 \le g/4$ and the centers are separated by at least g, these disks are pairwise disjoint. For every i, let $\Gamma_i := \{z \in \mathbb{C} : |z - \theta_i| = g/3\}$, oriented counterclockwise. The circle Γ_i remains disjoint from the spectrum of $\mathbf{\Theta} + t\mathbf{E}$ for every $t \in [0, 1]$. Hence the Riesz projection

$$\mathbf{\Pi}_i(t) := \frac{1}{2\pi i}\int_{\Gamma_i} \big(z\mathbf{I} - (\mathbf{\Theta} + t\mathbf{E})\big)^{-1}\, dz$$

is well defined and continuous in t. Its rank equals its trace, which is continuous and integer valued, and therefore constant. At $t = 0$ it has rank one, so each disk contains exactly one eigenvalue $\widehat{\theta}_i$, counted with algebraic multiplicity, and $|\widehat{\theta}_i - \theta_i| \le \|\mathbf{E}\|_2$. Because $\mathbf{\Theta} + \mathbf{E}$ is real, nonreal eigenvalues occur in conjugate pairs. Each disk contains exactly one eigenvalue and is centered on the real axis; hence every $\widehat{\theta}_i$ is real and the associated eigenvector may be chosen real. Let $\mathbf{y}_i \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ be an eigenvector associated with $\widehat{\theta}_i$, and let $\mathbf{Q}_i := \mathbf{I} - \mathbf{e}_i\mathbf{e}_i^\top$. From the eigenvalue equation,

$$(\theta_j - \widehat{\theta}_i)(\mathbf{y}_i)_j = -(\mathbf{E}\mathbf{y}_i)_j, \qquad j \ne i,$$
$$\min_{j\ne i} |\theta_j - \widehat{\theta}_i| \ge g - \|\mathbf{E}\|_2 \ge \frac{3g}{4}.$$

Consequently, summing the squared coordinate bounds gives

$$\|\mathbf{Q}_i\mathbf{y}_i\|_2 \le \frac{4}{3g}\|\mathbf{Q}_i\mathbf{E}\mathbf{y}_i\|_2 \le \frac{2\|\mathbf{E}\|_2}{g}\|\mathbf{y}_i\|_2. \tag{1}$$

The corresponding eigenvector of $\mathbf{A} + \mathbf{\Delta}$ is proportional to $\mathbf{V}\mathbf{y}_i$. Since $\mathbf{v}_i = \mathbf{V}\mathbf{e}_i$ is a unit vector, we have

$$\sin\angle(\widehat{\mathbf{v}}_i, \mathbf{v}_i) \le \frac{\|\mathbf{V}\mathbf{Q}_i\mathbf{y}_i\|_2}{\|\mathbf{V}\mathbf{y}_i\|_2} \le 2\frac{\mathrm{cond}(\mathbf{V})^2\|\mathbf{\Delta}\|_2}{g} \le C\gamma\,\mathrm{cond}(\mathbf{V})^2\rho\eta.$$

Finally, after choosing the sign of each normalized recovered eigenvector,

$$\|\widehat{\mathbf{v}}_i - \mathbf{v}_i\|_2 \le \sqrt{2}\sin\angle(\widehat{\mathbf{v}}_i, \mathbf{v}_i),$$

which proves the claim. □

B CONVEX-BODY SAMPLING AND MOMENT ESTIMATION

This section develops the convex-body sampling and moment-estimation mechanism underlying ASPIRE, including a noisy Hit-and-Run algorithm that remains provably accurate under multiplicative errors in the membership oracle.

Notation. Throughout the remainder of the appendix, whenever an orthonormal matrix $\widehat{\mathbf{B}}_\ell \in \mathbb{R}^{r_{\ell-1}\times r_\ell}$ is under consideration, define $\mathbf{C}_\ell^\sharp := \widehat{\mathbf{B}}_\ell^\top \mathbf{W}_\ell, G_\ell(\mathbf{z}) := F_\ell(\widehat{\mathbf{B}}_\ell \mathbf{z}) = h^{(\ell)}\Big((\mathbf{C}_\ell^\sharp)^\top \mathbf{z}\Big)$, and $K_\ell^\sharp := \{\mathbf{z} \in \mathbb{R}^{r_\ell} : G_\ell(\mathbf{z}) \le 1\}$. In the first-layer stage, where the column space is recovered exactly, we use the same notation with $\widehat{\mathbf{B}}_1 := \mathbf{B}_1$.

B.1 GEOMETRY OF THE REDUCED SAMPLING BODY

We first control the conditioning and Euclidean geometry of the reduced coordinates in which ASPIRE performs convex-body sampling.

The following lemma shows that an approximate column-space basis preserves the conditioning of the reduced weight matrix.

Lemma B.1 (Conditioning of the reduced weight matrix)**.** *Let $\mathbf{B}_\ell$ and $\widehat{\mathbf{B}}_\ell$ have orthonormal columns, suppose $\mathrm{col}(\mathbf{B}_\ell) = \mathrm{col}(\mathbf{W}_\ell)$, and assume, after orthogonal alignment, that $\|\widehat{\mathbf{B}}_\ell - \mathbf{B}_\ell\|_2 \le \varepsilon_{B,\ell} < 1$. For $\mathbf{C}_\ell^\sharp := \widehat{\mathbf{B}}_\ell^\top \mathbf{W}_\ell$, one has*

$$(1-\varepsilon_{B,\ell})\sigma_{\min}(\mathbf{W}_\ell) \le \sigma_{\min}(\mathbf{C}_\ell^\sharp) \le \sigma_{\max}(\mathbf{C}_\ell^\sharp) \le (1+\varepsilon_{B,\ell})\sigma_{\max}(\mathbf{W}_\ell).$$

Consequently, $\mathrm{cond}(\mathbf{C}_\ell^\sharp) \le \dfrac{1+\varepsilon_{B,\ell}}{1-\varepsilon_{B,\ell}}\kappa_0$. *In particular, if $\varepsilon_{B,\ell} \le 1/2$, then* $\mathrm{cond}(\mathbf{C}_\ell^\sharp) \le 3\kappa_0$.

Proof. Write $\mathbf{W}_\ell = \mathbf{B}_\ell \mathbf{C}_\ell, \mathbf{C}_\ell := \mathbf{B}_\ell^\top \mathbf{W}_\ell$. Because $\mathbf{B}_\ell$ has orthonormal columns, $\mathbf{C}_\ell$ has exactly the nonzero singular values of $\mathbf{W}_\ell$. Define $\mathbf{Q}_\ell := \widehat{\mathbf{B}}_\ell^\top \mathbf{B}_\ell$. Then $\mathbf{C}_\ell^\sharp = \mathbf{Q}_\ell \mathbf{C}_\ell$. Moreover,

$$\|\mathbf{Q}_\ell - \mathbf{I}\|_2 = \|\widehat{\mathbf{B}}_\ell^\top \mathbf{B}_\ell - \mathbf{B}_\ell^\top \mathbf{B}_\ell\|_2 \le \|\widehat{\mathbf{B}}_\ell - \mathbf{B}_\ell\|_2 \le \varepsilon_{B,\ell}.$$

Weyl's inequality therefore gives

$$1-\varepsilon_{B,\ell} \le \sigma_{\min}(\mathbf{Q}_\ell) \le \sigma_{\max}(\mathbf{Q}_\ell) \le 1+\varepsilon_{B,\ell}.$$

The singular-value product inequalities now imply

$$\sigma_{\min}(\mathbf{C}_\ell^\sharp) \ge \sigma_{\min}(\mathbf{Q}_\ell)\sigma_{\min}(\mathbf{C}_\ell) \ge (1-\varepsilon_{B,\ell})\sigma_{\min}(\mathbf{W}_\ell),$$

and

$$\sigma_{\max}(\mathbf{C}_\ell^\sharp) \le \sigma_{\max}(\mathbf{Q}_\ell)\sigma_{\max}(\mathbf{C}_\ell) \le (1+\varepsilon_{B,\ell})\sigma_{\max}(\mathbf{W}_\ell).$$

Taking the ratio gives

$$\mathrm{cond}(\mathbf{C}_\ell^\sharp) \le \frac{1+\varepsilon_{B,\ell}}{1-\varepsilon_{B,\ell}}\,\mathrm{cond}(\mathbf{W}_\ell) \le \frac{1+\varepsilon_{B,\ell}}{1-\varepsilon_{B,\ell}}\kappa_0.$$

When $\varepsilon_{B,\ell} \le 1/2$, the ratio in front of κ_0 is at most three. □

The next lemma converts this conditioning control into explicit inner and outer Euclidean radii for the reduced sampling body.

Lemma B.2 (Euclidean radii of the sampling body)**.** *Fix $1 \le \ell \le L-2$, and let*

$$q_\ell := k^{L-\ell}, \qquad S_\ell := \sum_{j=1}^{L-\ell-1} k^j = \frac{q_\ell - k}{k-1}.$$

Let $\widehat{\mathbf{B}}_\ell$ satisfy the hypotheses of Lemma B.1 with $\varepsilon_{B,\ell} \le 1/2$. Then $\frac{1}{(1+\varepsilon_{B,\ell})\sigma_{\max}(\mathbf{W}_\ell)} B_2^{r_\ell} \subseteq K_\ell^\sharp$ *and* $K_\ell^\sharp \subseteq \frac{\mu^{-(S_\ell+1)/q_\ell}\sqrt{r_\ell}}{(1-\varepsilon_{B,\ell})\sigma_{\min}(\mathbf{W}_\ell)} B_2^{r_\ell}$. *Consequently, one may take radii satisfying*

$$r_{\mathrm{in},\ell} \ge \frac{1}{(1+\varepsilon_{B,\ell})\sigma_{\max}(\mathbf{W}_\ell)}, R_{\mathrm{out},\ell} \le \frac{\mu^{-(S_\ell+1)/q_\ell}\sqrt{r_\ell}}{(1-\varepsilon_{B,\ell})\sigma_{\min}(\mathbf{W}_\ell)}.$$

In particular,

$$\omega_\ell := \frac{R_{\mathrm{out},\ell}}{r_{\mathrm{in},\ell}} \le 3\kappa_0 \mu^{-(S_\ell+1)/q_\ell}\sqrt{r_\ell} \le C_{k,L}\kappa^4\sqrt{r_\ell} \le C_{k,L}\kappa^4\sqrt{r}.$$

Proof. For $\mathbf{u} \in \mathbb{R}^{r_\ell}$, define the nonnegative forward quantities

$$\mathbf{s}_\ell := \mathbf{u}^{\odot k}, \qquad \mathbf{s}_j := \left(\mathbf{W}_j^\top \mathbf{s}_{j-1}\right)^{\odot k}, \quad \ell+1 \le j \le L-1.$$

Then $h^{(\ell)}(\mathbf{u}) = \mathbf{a}^\top \mathbf{s}_{L-1}$.We first derive an upper bound that uses only the column-stochastic normalization. If $\mathbf{v} \ge \mathbf{0}$, every coordinate of $\mathbf{W}_j^\top \mathbf{v}$ is a convex combination of the coordinates of $\mathbf{v}$, and therefore $\|\mathbf{W}_j^\top \mathbf{v}\|_\infty \le \|\mathbf{v}\|_\infty$. It follows inductively that $\|\mathbf{s}_{L-1}\|_\infty \le \|\mathbf{u}\|_\infty^{q_\ell}$. Since $\mathbf{a} \ge \mathbf{0}$ and $\mathbf{1}^\top \mathbf{a} = 1$, $h^{(\ell)}(\mathbf{u}) = \mathbf{a}^\top \mathbf{s}_{L-1} \le \|\mathbf{s}_{L-1}\|_\infty \le \|\mathbf{u}\|_\infty^{q_\ell} \le \|\mathbf{u}\|_2^{q_\ell}$. For the reverse inequality, define the average mass at layer j by

$$\overline{s}_j := \frac{\|\mathbf{s}_j\|_1}{r_j}, \qquad \ell \le j \le L-1.$$

At the initial step, the comparison between ℓ_k- and ℓ_2-norms gives $\overline{s}_\ell = \frac{\|\mathbf{u}\|_k^k}{r_\ell} \ge \left(\frac{\|\mathbf{u}\|_2}{\sqrt{r_\ell}}\right)^k$. For every $j > \ell$, the relative row-balance condition and nonnegativity yield

$$\frac{1}{r_j}\|\mathbf{W}_j^\top \mathbf{s}_{j-1}\|_1 = \frac{1}{r_j}(\mathbf{W}_j \mathbf{1}_{r_j})^\top \mathbf{s}_{j-1} \ge \frac{\mu}{r_{j-1}}\|\mathbf{s}_{j-1}\|_1 = \mu \overline{s}_{j-1}.$$

Jensen's inequality then gives

$$\overline{s}_j = \frac{1}{r_j}\|\mathbf{W}_j^\top \mathbf{s}_{j-1}\|_k^k \ge \left(\frac{1}{r_j}\|\mathbf{W}_j^\top \mathbf{s}_{j-1}\|_1\right)^k \ge \mu^k \overline{s}_{j-1}^k.$$

Iterating through the remaining hidden layers gives $\overline{s}_{L-1} \ge \mu^{S_\ell}\left(\frac{\|\mathbf{u}\|_2}{\sqrt{r_\ell}}\right)^{q_\ell}$. The row-balance condition for $\mathbf{W}_L = \mathbf{a}$ gives $a_i \ge \mu/r_{L-1}$, and hence

$$h^{(\ell)}(\mathbf{u}) = \mathbf{a}^\top \mathbf{s}_{L-1} \ge \frac{\mu}{r_{L-1}}\|\mathbf{s}_{L-1}\|_1 = \mu \overline{s}_{L-1} \ge \mu^{S_\ell+1}\left(\frac{\|\mathbf{u}\|_2}{\sqrt{r_\ell}}\right)^{q_\ell}.$$

Let $H_\ell := \{\mathbf{u} \in \mathbb{R}^{r_\ell} : h^{(\ell)}(\mathbf{u}) \le 1\}$. The two polynomial bounds imply $B_2^{r_\ell} \subseteq H_\ell \subseteq \mu^{-(S_\ell+1)/q_\ell}\sqrt{r_\ell}\, B_2^{r_\ell}$. By definition, $K_\ell^\sharp = (\mathbf{C}_\ell^\sharp)^{-\top} H_\ell$. For an invertible matrix $\mathbf{C}$, the image $(\mathbf{C}^{-\top})(\rho B_2)$ contains $(\rho/\sigma_{\max}(\mathbf{C}))B_2$, whereas the image of RB_2 is contained in $(R/\sigma_{\min}(\mathbf{C}))B_2$. Applying Lemma B.1 gives the two radius bounds in the statement.

Taking their ratio yields

$$\omega_\ell \le \frac{1+\varepsilon_{B,\ell}}{1-\varepsilon_{B,\ell}}\,\mathrm{cond}(\mathbf{W}_\ell)\mu^{-(S_\ell+1)/q_\ell}\sqrt{r_\ell}.$$

Since $\varepsilon_{B,\ell} \le 1/2$, $\mathrm{cond}(\mathbf{W}_\ell) \le \kappa_0$, and $\frac{S_\ell+1}{q_\ell} = \frac{q_\ell - 1}{(k-1)q_\ell} < \frac{1}{k-1}$, the dimension-free factor is at most $C_{k,L}\kappa^4$. Finally, $r_\ell \le r$, proving the stated simplification. □

B.2 HIT-AND-RUN WITH A MULTIPLICATIVE ORACLE

We next develop the geometric and perturbative ingredients needed to sample from the true convex body when only a multiplicatively inaccurate oracle is available.

The following Markov-type bound controls the curvature of a polynomial on a convex body from its sup norm and inradius.

Lemma B.3 (Markov bound for second derivatives)**.** *Let $K \subset \mathbb{R}^n$ be a convex body satisfying*

$$\rho B_2^n \subseteq K$$

for some $\rho > 0$, and let $p : \mathbb{R}^n \to \mathbb{R}$ be a polynomial of degree at most $q \geq 2$. Then

$$\sup_{\mathbf{x}\in K} \|\nabla^2 p(\mathbf{x})\|_{\mathrm{op}} \leq \frac{4q^2(q-1)^2}{\rho^2}\|p\|_{L^\infty(K)} \leq \frac{4q^4}{\rho^2}\|p\|_{L^\infty(K)}.$$

Proof. Let $w(K)$ denote the minimum width of K. Since $\rho B_2^n \subseteq K$, we have $w(K) \geq 2\rho$. By the multivariate Markov inequality of Wilhelmsen (1974), every polynomial g of degree at most m satisfies

$$\|\nabla g\|_{L^\infty(K)} \leq \frac{4m^2}{w(K)}\|g\|_{L^\infty(K)} \leq \frac{2m^2}{\rho}\|g\|_{L^\infty(K)}.$$

Applying this first to p gives

$$\|\nabla p\|_{L^\infty(K)} \leq \frac{2q^2}{\rho}\|p\|_{L^\infty(K)}.$$

Now fix $\mathbf{u} \in S^{n-1}$. The directional derivative $D_{\mathbf{u}}p := \mathbf{u}^\top \nabla p$ has degree at most $q-1$, and therefore another application of the same inequality yields

$$\|\nabla(D_{\mathbf{u}}p)\|_{L^\infty(K)} \leq \frac{2(q-1)^2}{\rho}\|D_{\mathbf{u}}p\|_{L^\infty(K)} \leq \frac{2(q-1)^2}{\rho}\|\nabla p\|_{L^\infty(K)} \leq \frac{4q^2(q-1)^2}{\rho^2}\|p\|_{L^\infty(K)}.$$

Since $\nabla(D_{\mathbf{u}}p)(\mathbf{x}) = \nabla^2 p(\mathbf{x})\mathbf{u}$, taking the supremum over $\mathbf{u} \in S^{n-1}$ and $\mathbf{x} \in K$ proves the claim. □

Using this curvature control, the next lemma shows that noisy Hit-and-Run remains close to uniform sampling from the true body under sufficiently small multiplicative oracle error.

Lemma B.4 (Hit-and-Run with multiplicative oracle noise)**.** *Let $F : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a positive definite convex homogeneous polynomial of degree $q \geq 2$, and define $K := \{\mathbf{x} \in \mathbb{R}^n : F(\mathbf{x}) \leq 1\}$. Assume*

$$r_{\mathrm{in}} B_2^n \subseteq K \subseteq R_{\mathrm{out}} B_2^n, \qquad \omega := \frac{R_{\mathrm{out}}}{r_{\mathrm{in}}},$$

and suppose that the available oracle $\widehat{F} : \mathbb{R}^n \to \mathbb{R}$ is Borel measurable and satisfies

$$(1-\xi)F(\mathbf{x}) \leq \widehat{F}(\mathbf{x}) \leq (1+\xi)F(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^n,$$

for some $0 < \xi \leq 1/2$.

Fix $\delta \in (0, 1]$. Then there exist an absolute constant $C > 0$ and a constant $c_\delta > 0$, depending only on δ, such that, for every $\varepsilon \in (0, 1/2)$, if

$$\xi \leq \frac{c_\delta\, \varepsilon^{1+\delta}}{q^2 n^{4+3\delta}\omega^{4+2\delta}\left(\log\dfrac{2\omega}{\varepsilon}\right)^{1+\delta}},$$

then the noisy Hit-and-Run chain constructed in the proof, initialized at $\widehat{\mathbf{X}}_0 = \mathbf{0}$, satisfies

$$d_{\mathrm{TV}}\left(\mathcal{L}(\widehat{\mathbf{X}}_T), \mathrm{Unif}(K)\right) \leq \varepsilon$$

for $T = \left\lceil Cn^3\omega^2 \log\dfrac{2\omega}{\varepsilon}\right\rceil$. With

$$\tau := 1 - \left(\frac{1-\xi}{1+\xi}\right)^{1/q}, \qquad \eta := \frac{\varepsilon}{4T},$$

the construction uses $O\left(T \log\dfrac{\omega}{\eta\tau}\right)$ evaluations of $\widehat{F}$.

Proof. Set $\rho := \left(\frac{1-\xi}{1+\xi}\right)^{1/q}$, then$\tau = 1-\rho$. By homogeneity, $\rho K \subseteq \{\mathbf{x} : \widehat{F}(\mathbf{x}) \le 1-\xi\} \subseteq K$. Indeed, if $\mathbf{x} \in \rho K$, then $F(\mathbf{x}) \le \rho^q = \frac{1-\xi}{1+\xi}$, and hence $\widehat{F}(\mathbf{x}) \le (1+\xi)F(\mathbf{x}) \le 1-\xi$. Conversely,

$$\widehat{F}(\mathbf{x}) \le 1-\xi \Longrightarrow (1-\xi)F(\mathbf{x}) \le 1-\xi \Longrightarrow F(\mathbf{x}) \le 1.$$

Moreover,

$$q\tau \le \log\frac{1+\xi}{1-\xi} \le 3\xi. \tag{2}$$

Let P denote the exact Hit-and-Run transition kernel on K, and let $\mathbf{X}_0 = \mathbf{0}, \mathbf{X}_{t+1} \sim P(\mathbf{X}_t, \cdot)$. Define the boundary depth $D(\mathbf{x}) := 1 - F(\mathbf{x})$. Since ∂K has Lebesgue measure zero, $D(\mathbf{X}_t) > 0$ almost surely. We first establish a uniform negative-moment bound for $D(\mathbf{X}_t)$. Set $\alpha := \frac{1}{1+\delta}, s := 1-\alpha = \frac{\delta}{1+\delta}$, and define $\vartheta_\delta := \left(\frac{s}{2}\right)^{1/s}$. Let $C_0 = \sqrt{\frac{2}{\pi}}$. Then for every unit vector $\mathbf{v} \in \mathbb{R}^n$, every $a \in (0,1)$, and $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1})$,

$$\mathbb{P}\left(|\langle \mathbf{v}, \boldsymbol{\theta}\rangle| \le \frac{a}{\sqrt{n}}\right) \le C_0 a. \tag{3}$$

This is the standard spherical small-ball estimate.

Set $a_\delta := \frac{\vartheta_\delta}{2C_0}, \Delta_\delta := \frac{a_\delta^2}{32\, n\omega^2 q^2}, d_0 := \frac{\vartheta_\delta}{2}\Delta_\delta$. Thus

$$d_0 = \frac{c_0 \vartheta_\delta^3}{n\omega^2 q^2}, \qquad c_0 := \frac{1}{256 C_0^2}, \tag{4}$$

where $c_0 > 0$ is an absolute constant.

We claim that whenever $\mathbf{x} \in \mathrm{int}(K)$ and $D(\mathbf{x}) \le d_0$,

$$\mathbb{P}\left(D(\mathbf{X}_{t+1}) \ge d_0 \,|\, \mathbf{X}_t = \mathbf{x}\right) \ge 1 - \vartheta_\delta. \tag{5}$$

Since $d_0 \le 1/2$, $F(\mathbf{x}) \ge \frac{1}{2}$. Euler's identity and $\|\mathbf{x}\|_2 \le R_{\mathrm{out}}$ give

$$\mathbf{x}^\top \nabla F(\mathbf{x}) = qF(\mathbf{x}), \qquad \|\nabla F(\mathbf{x})\|_2 \ge \frac{q}{2R_{\mathrm{out}}}.$$

Since F is positive definite and $K = \{\mathbf{z} : F(\mathbf{z}) \le 1\}$, we have $\|F\|_{L^\infty(K)} = 1$. Hence Lemma B.3 yields

$$\mathcal{H} := \sup_{\mathbf{z}\in K} \|\nabla^2 F(\mathbf{z})\|_{\mathrm{op}} \le \frac{4q^4}{r_{\mathrm{in}}^2}.$$

Let $\mathbf{v} = \frac{\nabla F(\mathbf{x})}{\|\nabla F(\mathbf{x})\|_2}$. By (3), with probability at least $1 - \vartheta_\delta/2$, $|\langle \mathbf{v}, \boldsymbol{\theta}\rangle| \ge \frac{a_\delta}{\sqrt{n}}$. Choose the orientation of the sampled line for which $A := -\nabla F(\mathbf{x})^\top \boldsymbol{\theta} > 0$. On the above event, $A \ge \frac{q a_\delta}{2R_{\mathrm{out}}\sqrt{n}}$. Set $f(t) := F(\mathbf{x} + t\boldsymbol{\theta}), t_* := \frac{A}{\mathcal{H}}$, and let $b := \sup\{t \ge 0 : \mathbf{x} + s\boldsymbol{\theta} \in K \text{ for every } s \in [0,t]\}$ be the distance from $\mathbf{x}$ to the boundary of K along the chosen orientation of the line. Since K is compact and $\mathbf{x} \in \mathrm{int}(K)$, we have

$$0 < b < \infty, \qquad \mathbf{x} + b\boldsymbol{\theta} \in \partial K, \qquad f(b) = 1.$$

For every $0 \le t \le \min\{b, t_*\}$, the segment $\{\mathbf{x} + s\boldsymbol{\theta} : 0 \le s \le t\}$ is contained in K. Hence

$$f'(t) = f'(0) + \int_0^t \boldsymbol{\theta}^\top \nabla^2 F(\mathbf{x} + s\boldsymbol{\theta})\boldsymbol{\theta}\, ds \le -A + \mathcal{H}t.$$

Suppose, for contradiction, that $b < t_*$. Integrating the preceding inequality over $[0, b]$ gives

$$f(b) \le f(0) - Ab + \frac{\mathcal{H}}{2}b^2 < f(0) \le 1,$$

because $b < A/\mathcal{H}$. This contradicts $f(b) = 1$. Therefore $b \ge t_*$. In particular, $\mathbf{x} + t_*\boldsymbol{\theta} \in K$, and another integration yields $f(t_*) \le f(0) - At_* + \frac{\mathcal{H}}{2}t_*^2 = f(0) - \frac{A^2}{2\mathcal{H}}$. Consequently, $\frac{A^2}{2\mathcal{H}} \ge \frac{a_\delta^2}{32\, n\omega^2 q^2} = \Delta_\delta$, so the chord contains a point $\mathbf{z}$ satisfying $F(\mathbf{z}) \le 1 - \Delta_\delta$. Let $\mathbf{z}_-$ and $\mathbf{z}_+$ be the two endpoints of this chord. Since $F(\mathbf{z}_-) = F(\mathbf{z}_+) = 1$, convexity implies, for either sign and $t \in [0, 1]$,

$$F((1-t)\mathbf{z} + t\mathbf{z}_\pm) \le (1-t)F(\mathbf{z}) + tF(\mathbf{z}_\pm) \le 1 - (1-t)\Delta_\delta.$$

Hence $F((1-t)\mathbf{z} + t\mathbf{z}_\pm) \le 1 - d_0$ whenever $t \le 1 - \frac{d_0}{\Delta_\delta}$. Therefore the proportion of the entire chord on which $D \ge d_0$ is at least $1 - \frac{d_0}{\Delta_\delta} = 1 - \frac{\vartheta_\delta}{2}$. Combining this with the directional event yields

$$\mathbb{P}\left(D(\mathbf{X}_{t+1}) \ge d_0 \,|\, \mathbf{X}_t = \mathbf{x}\right) \ge \left(1 - \frac{\vartheta_\delta}{2}\right)^2 \ge 1 - \vartheta_\delta,$$

proving (5).

For the exact Hit-and-Run chain, we first record a deterministic homothetic chord comparison. Let $0 \le \sigma_0 < \lambda \le 1, \mathbf{x} \in \sigma_0 K$, and let I_K and $I_{\lambda K}$ be the intersections of a fixed line through $\mathbf{x}$ with K and λK, respectively. Then

$$\frac{|I_{\lambda K}|}{|I_K|} \ge \frac{\lambda - \sigma_0}{1 - \sigma_0}. \tag{6}$$

Indeed, write $\mathbf{x} = \sigma_0\mathbf{y}$ for some $\mathbf{y} \in K$, and set $\beta := \frac{\lambda - \sigma_0}{1 - \sigma_0}$. For every point $\mathbf{z} \in I_K$,

$$\mathbf{x} + \beta(\mathbf{z} - \mathbf{x}) = (1-\beta)\sigma_0\mathbf{y} + \beta\mathbf{z} = \lambda\left(\frac{(1-\beta)\sigma_0}{\lambda}\mathbf{y} + \frac{\beta}{\lambda}\mathbf{z}\right).$$

The two coefficients in parentheses are nonnegative and sum to one, so the point in parentheses belongs to K. Therefore $\mathbf{x} + \beta(I_K - \mathbf{x}) \subseteq I_{\lambda K}$, which proves (6).

Now fix $\mathbf{x} \in \mathrm{int}(K)$ and write $d := D(\mathbf{x}) \in (0, 1]$. For $0 < u < d$, set $s_d := (1-d)^{1/q}, \lambda_u := (1-u)^{1/q}$. Then $\mathbf{x} \in s_d K$ and $s_d < \lambda_u$. Conditional on any sampled direction, the event $D(\mathbf{X}_{t+1}) \le u$ is contained, up to chord endpoints of one-dimensional measure zero, in the part of I_K outside $I_{\lambda_u K}$. Thus (6), followed by averaging over the direction, gives

$$P\big(\mathbf{x}, \{\mathbf{y} \in K : D(\mathbf{y}) \le u\}\big) \le \frac{1 - \lambda_u}{1 - s_d}.$$

Since $v \mapsto 1 - (1-v)^{1/q}$ is convex and vanishes at 0,

$$P\big(\mathbf{x}, \{\mathbf{y} \in K : D(\mathbf{y}) \le u\}\big) \le \frac{u}{d}, \qquad 0 < u < d. \tag{7}$$

Define the one-step negative moment at $\mathbf{x}$ by $\mathcal{M}(\mathbf{x}) := \int_K D(\mathbf{y})^{-\alpha}\, P(\mathbf{x}, d\mathbf{y})$. If $d \le d_0$, then (5) and (7) imply

$$P\big(\mathbf{x}, \{\mathbf{y} : D(\mathbf{y}) \le u\}\big) \le \min\left\{\vartheta_\delta, \frac{u}{d}\right\}, \qquad 0 < u \le d_0.$$

Here, when $u \ge d$, the second bound follows from the trivial estimate $P(\mathbf{x}, \cdot) \le 1 \le u/d$. For $0 < Y \le 1$, $\mathbb{E}[Y^{-\alpha}] = 1 + \alpha\int_0^1 u^{-\alpha-1}\mathbb{P}(Y \le u)\, du$. Applying this identity to the one-step law $P(\mathbf{x}, \cdot)$, and splitting the integral at $u = \vartheta_\delta d$ and $u = d_0$, gives

$$\mathcal{M}(\mathbf{x}) \le 1 + \frac{\alpha}{d}\int_0^{\vartheta_\delta d} u^{-\alpha}\, du + \alpha\vartheta_\delta \int_{\vartheta_\delta d}^{d_0} u^{-\alpha-1}\, du + \alpha\int_{d_0}^1 u^{-\alpha-1}\, du \le \frac{\vartheta_\delta^{1-\alpha}}{1-\alpha}d^{-\alpha} + d_0^{-\alpha}.$$

Thus

$$\mathcal{M}(\mathbf{x}) \le \beta_\delta d^{-\alpha} + d_0^{-\alpha}, \tag{8}$$

where $\beta_\delta := \dfrac{\vartheta_\delta^{1-\alpha}}{1-\alpha}$. By the definition of ϑ_δ, $\beta_\delta = \dfrac{\vartheta_\delta^s}{s} = \dfrac{1}{2}$. If $d > d_0$, then (7) gives, for $u \le d_0$,

$$P\big(\mathbf{x}, \{\mathbf{y} : D(\mathbf{y}) \le u\}\big) \le \frac{u}{d} \le \frac{u}{d_0}.$$

Therefore

$$\mathcal{M}(\mathbf{x}) \le 1 + \frac{\alpha}{d_0}\int_0^{d_0} u^{-\alpha}\,du + \alpha\int_{d_0}^1 u^{-\alpha-1}\,du = \frac{1}{1-\alpha}d_0^{-\alpha} = \frac{1}{s}d_0^{-\alpha}.$$

Hence

$$\mathcal{M}(\mathbf{x}) \le \frac{1}{s}d_0^{-\alpha}. \tag{9}$$

The two cases imply the pointwise bound $\mathcal{M}(\mathbf{x}) \le \dfrac{1}{2}D(\mathbf{x})^{-\alpha} + \dfrac{1}{s}d_0^{-\alpha}$. Writing $M_t := \mathbb{E}[D(\mathbf{X}_t)^{-\alpha}]$, and using the Markov property, we obtain $M_{t+1} = \mathbb{E}[\mathcal{M}(\mathbf{X}_t)] \le \dfrac{1}{2}M_t + \dfrac{1}{s}d_0^{-\alpha}$. Since $D(\mathbf{X}_0) = 1$,

$$\sup_{t\ge 0}\mathbb{E}[D(\mathbf{X}_t)^{-\alpha}] \le \frac{2}{s}d_0^{-\alpha}. \tag{10}$$

Use the values of τ and η from the statement, and set $\delta_{\mathrm{bis}} := \eta\tau r_{\mathrm{in}}$. At a current point $\mathbf{x} \in K$, draw $\boldsymbol{\theta} \sim \mathrm{Unif}(S^{n-1})$. For each $\sigma \in \{+1, -1\}$, initialize $L_0^\sigma := 0, \qquad U_0^\sigma := 3R_{\mathrm{out}}$. Given L_j^σ, U_j^σ, set $m_j^\sigma := \dfrac{L_j^\sigma + U_j^\sigma}{2}$ and update

$$(L_{j+1}^\sigma, U_{j+1}^\sigma) = \begin{cases} (m_j^\sigma, U_j^\sigma), & \widehat{F}(\mathbf{x} + \sigma m_j^\sigma \boldsymbol{\theta}) \le 1-\xi, \\ (L_j^\sigma, m_j^\sigma), & \widehat{F}(\mathbf{x} + \sigma m_j^\sigma \boldsymbol{\theta}) > 1-\xi. \end{cases}$$

The procedure stops once $U_j^\sigma - L_j^\sigma \le \delta_{\mathrm{bis}}$, and returns $\widehat{b}_\sigma := L_j^\sigma$. The next state is sampled uniformly on the resulting approximate chord:

$$\widehat{t} \sim \mathrm{Unif}([-\widehat{b}_-, \widehat{b}_+]), \qquad \widehat{\mathbf{X}}_{t+1} = \mathbf{x} + \widehat{t}\boldsymbol{\theta}.$$

Whenever a midpoint is accepted, it lies in K, since $\{\widehat{F} \le 1-\xi\} \subseteq K$. Moreover $\mathbf{x} \in K$, so convexity of K implies that the entire approximate chord is contained in K.

To compare the noisy and exact chords, fix $\mathbf{x} \in K$ and a direction $\boldsymbol{\theta}$. Let $I_K = [-b_-, b_+]$ be the exact chord of K through $\mathbf{x}$ in direction $\boldsymbol{\theta}$. Whenever $\mathbf{x} \in \rho K$, let $I_{\rho K} = [-b_-^\rho, b_+^\rho]$ be the corresponding chord of ρK.

We first record the following deterministic property of the above bisection procedure. For each $\sigma \in \{+1, -1\}$, if $\mathbf{x} \in \rho K$, then throughout the bisection,

$$L_j^\sigma \le b_\sigma, \qquad U_j^\sigma \ge b_\sigma^\rho. \tag{11}$$

Indeed, the inequalities hold initially. Since $K \subseteq R_{\mathrm{out}}B_2^n$ and $\|\mathbf{x}\|_2 \le R_{\mathrm{out}}$, the point $\mathbf{x} + \sigma(3R_{\mathrm{out}})\boldsymbol{\theta}$ lies outside K, so

$$3R_{\mathrm{out}} \ge b_\sigma \ge b_\sigma^\rho.$$

Suppose (11) holds at step j. If m_j^σ is accepted, then $\mathbf{x} + \sigma m_j^\sigma \boldsymbol{\theta} \in K$, and therefore $m_j^\sigma \le b_\sigma$. Hence replacing L_j^σ by m_j^σ preserves the first invariant.

If m_j^σ is rejected, then $\mathbf{x} + \sigma m_j^\sigma \boldsymbol{\theta} \notin \rho K$, because every point of ρK is necessarily accepted by the sandwich $\rho K \subseteq \{\widehat{F} \le 1-\xi\}$. Since $\mathbf{x} \in \rho K$ and ρK is convex, the intersection of the ray $\{\mathbf{x} + \sigma t\boldsymbol{\theta} : t \ge 0\}$ with ρK is exactly the interval $[0, b_\sigma^\rho]$. Thus rejection implies $m_j^\sigma \ge b_\sigma^\rho$, so replacing U_j^σ by m_j^σ preserves the second invariant. This proves (11).

Consequently, upon termination,

$$b_\sigma^\rho - \delta_{\mathrm{bis}} \le \widehat{b}_\sigma \le b_\sigma, \qquad \sigma \in \{+1, -1\}.$$

In particular, if $\widehat{I} := [-\widehat{b}_-, \widehat{b}_+]$, then

$$\widehat{I} \subseteq I_K, \qquad |\widehat{I}| \ge |I_{\rho K}| - 2\delta_{\mathrm{bis}}. \tag{12}$$

Notice that this argument uses only $\rho K \subseteq \{\widehat{F} \le 1 - \xi\} \subseteq K$ and the convexity of K and ρK; no convexity or linewise monotonicity of the noisy sublevel set $\{\widehat{F} \le 1 - \xi\}$ is required.

Let $\widehat{P}$ denote the noisy transition kernel and define

$$e(\mathbf{x}) := d_{\mathrm{TV}}\big(P(\mathbf{x}, \cdot), \widehat{P}(\mathbf{x}, \cdot)\big).$$

For the one-step kernel bound, set $d := D(\mathbf{x}) = 1 - F(\mathbf{x})$. If $d \le q\tau$, then trivially

$$e(\mathbf{x}) \le 1 = \min\left\{1, \frac{q\tau}{d}\right\} \le \min\left\{1, \frac{q\tau}{d}\right\} + \eta.$$

Suppose now that $d > q\tau$. Since $\rho^q = (1-\tau)^q \ge 1 - q\tau$, we have $F(\mathbf{x}) = 1 - d < 1 - q\tau \le \rho^q$, and hence $\mathbf{x} \in \rho K$. Thus the preceding bisection invariants apply.

Write $s_d := (1-d)^{1/q}$. The same homothetic chord comparison used above gives $\dfrac{|I_{\rho K}|}{|I_K|} \ge \dfrac{\rho - s_d}{1 - s_d}$. Therefore

$$1 - \frac{|I_{\rho K}|}{|I_K|} \le \frac{1-\rho}{1-s_d} = \frac{\tau}{1 - (1-d)^{1/q}}.$$

Since $g(d) := 1 - (1-d)^{1/q}$ is convex, $g(0) = 0$, and $g'(0) = 1/q$, $1 - (1-d)^{1/q} \ge \dfrac{d}{q}$. Hence

$$1 - \frac{|I_{\rho K}|}{|I_K|} \le \frac{q\tau}{d}. \tag{13}$$

We also need a lower bound on the exact chord length. Since $\mathbf{x} \in \rho K$, write $\mathbf{x} = \rho\mathbf{y}$ for some $\mathbf{y} \in K$. For every $\mathbf{z} \in r_{\mathrm{in}} B_2^n \subseteq K$, convexity gives $\rho\mathbf{y} + \tau\mathbf{z} \in K$. Thus $\mathbf{x} + \tau r_{\mathrm{in}} B_2^n \subseteq K$, and consequently, in every direction, $|I_K| \ge 2\tau r_{\mathrm{in}}$. Using $\delta_{\mathrm{bis}} = \eta\tau r_{\mathrm{in}}$,

$$\frac{2\delta_{\mathrm{bis}}}{|I_K|} \le \eta. \tag{14}$$

Conditional on the sampled direction $\boldsymbol{\theta}$, the exact and noisy Hit-and-Run transitions are respectively uniform on I_K and $\widehat{I}$. Since $\widehat{I} \subseteq I_K$,

$$d_{\mathrm{TV}}\left(\mathrm{Unif}(I_K), \mathrm{Unif}(\widehat{I})\right) = 1 - \frac{|\widehat{I}|}{|I_K|}.$$

By (12), (13), and (14),

$$1 - \frac{|\widehat{I}|}{|I_K|} \le 1 - \frac{|I_{\rho K}|}{|I_K|} + \frac{2\delta_{\mathrm{bis}}}{|I_K|} \le \frac{q\tau}{d} + \eta.$$

Averaging over $\boldsymbol{\theta}$ therefore yields

$$e(\mathbf{x}) \le \min\left\{1, \frac{q\tau}{D(\mathbf{x})}\right\} + \eta. \tag{15}$$

Since $\min\{1, a\} \le a^\alpha$, $a \ge 0$, (10) and (15) imply

$$\sup_{t \ge 0} \mathbb{E}_{\mathbf{X}_t \sim \delta_{\mathbf{0}} P^t}[e(\mathbf{X}_t)] \le \frac{2}{s}\left(\frac{q\tau}{d_0}\right)^\alpha + \eta. \tag{16}$$

Using the telescoping identity $\widehat{P}^T - P^T = \sum_{j=0}^{T-1} P^j(\widehat{P} - P)\widehat{P}^{T-1-j}$ and contraction of total variation under Markov kernels, (16) gives

$$d_{\mathrm{TV}}\left(\delta_{\mathbf{0}}\widehat{P}^T, \delta_{\mathbf{0}}P^T\right) \le \frac{2T}{s}\left(\frac{q\tau}{d_0}\right)^{\alpha} + T\eta. \tag{17}$$

By Corollary 1.2 of Lovász & Vempala (2006), the exact Hit-and-Run chain on a convex body containing a Euclidean ball of radius r_{in} and contained in a Euclidean ball of radius R_{out}, when initialized at a point $\mathbf{x}_0 \in K$, satisfies

$$d_{\mathrm{TV}}\left(\delta_{\mathbf{x}_0}P^T, \mathrm{Unif}(K)\right) \le \frac{\varepsilon}{2}$$

provided

$$T \ge Cn^3 \frac{R_{\mathrm{out}}^2}{r_{\mathrm{in}}^2} \log \frac{2R_{\mathrm{out}}}{\varepsilon\ \mathrm{dist}(\mathbf{x}_0, \partial K)},$$

for an absolute constant $C > 0$. In our setting $\mathbf{x}_0 = \mathbf{0}$, and $r_{\mathrm{in}}B_2^n \subset K$ implies $\mathrm{dist}(\mathbf{0}, \partial K) \ge r_{\mathrm{in}}$. Therefore, with $\omega := \dfrac{R_{\mathrm{out}}}{r_{\mathrm{in}}}$, it suffices to take $T = \left\lceil Cn^3\omega^2 \log \dfrac{2\omega}{\varepsilon} \right\rceil$, for which

$$d_{\mathrm{TV}}\left(\delta_{\mathbf{0}}P^T, \mathrm{Unif}(K)\right) \le \frac{\varepsilon}{2}.$$

Since $T\eta = \dfrac{\varepsilon}{4}$, it remains to impose $\dfrac{2T}{s}\left(\dfrac{q\tau}{d_0}\right)^{\alpha} \le \dfrac{\varepsilon}{4}$. By (2), it is sufficient that $\xi \le c\, d_0 \left(\dfrac{s\varepsilon}{T}\right)^{1/\alpha}$. Using $\dfrac{1}{\alpha} = 1 + \delta$, $d_0 = \dfrac{c_0\vartheta_\delta^3}{n\omega^2 q^2}$, and $T = O\left(n^3\omega^2 \log \dfrac{2\omega}{\varepsilon}\right)$, we obtain

$$\xi \le \frac{c_\delta\, \varepsilon^{1+\delta}}{q^2 n^{4+3\delta}\omega^{4+2\delta}\left(\log \dfrac{2\omega}{\varepsilon}\right)^{1+\delta}},$$

where

$$c_\delta = cs^{1+\delta}\vartheta_\delta^3 = c\left(\frac{\delta}{1+\delta}\right)^{1+\delta}\left(\frac{\delta}{2(1+\delta)}\right)^{3(1+\delta)/\delta}.$$

The triangle inequality together with (17) therefore gives $d_{\mathrm{TV}}\left(\mathcal{L}(\widehat{\mathbf{X}}_T), \mathrm{Unif}(K)\right) \le \varepsilon$.

Finally, each endpoint bisection requires at most $\left\lceil \log_2 \dfrac{3R_{\mathrm{out}}}{\delta_{\mathrm{bis}}} \right\rceil$ oracle evaluations. Hence the total number of oracle calls is $O\left(T \log \dfrac{\omega}{\eta\tau}\right)$. □

B.3 Population moments and spectral conditioning

We now quantify the population position and gradient moments generated by uniform sampling and verify that the resulting spectral pencil is sufficiently well conditioned for diagonalization.

The following lemma provides the moment scales, envelope bounds, and spectral conditioning parameters used in the perturbation analysis of ASPIRE.

Lemma B.5 (Population moment bounds and spectral conditioning)**.** *Fix $1 \le \ell \le L-2$, and write $n := r_\ell, q := q_\ell, G := G_\ell, K := K_\ell^\sharp$. Let $\rho B_2^n \subseteq K \subseteq RB_2^n, \omega := \dfrac{R}{\rho}$, and let $\mathbf{Z} \sim \mathrm{Unif}(K)$. Define*

$$\mathbf{\Sigma}_x := \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top], \qquad \mathbf{\Sigma}_g := \mathbb{E}[\nabla G(\mathbf{Z})\nabla G(\mathbf{Z})^\top].$$

Then

$$\frac{\rho^2}{n+2} \le \|\mathbf{\Sigma}_x\|_2 \le R^2, \qquad \frac{q^2}{R^2(n+2q-2)} \le \|\mathbf{\Sigma}_g\|_2 \le \frac{q^2}{\rho^2}.$$

In particular,

$$\frac{\sup_{\mathbf{z}\in K}\|\mathbf{z}\mathbf{z}^\top\|_2}{\|\mathbf{\Sigma}_x\|_2}\le (n+2)\omega^2,\qquad \frac{\sup_{\mathbf{z}\in K}\|\nabla G(\mathbf{z})\nabla G(\mathbf{z})^\top\|_2}{\|\mathbf{\Sigma}_g\|_2}\le (n+2q-2)\omega^2. \tag{18}$$

Let $\mathbf{C}:=\mathbf{C}_\ell^\sharp$, $\mathbf{U}:=\mathbf{C}^\top\mathbf{Z}$, *and write* $\mathbf{\Sigma}_x=\mathbf{C}^{-\top}\mathbf{\Lambda}\mathbf{C}^{-1}$, $\mathbf{\Sigma}_g=\mathbf{C}\mathbf{D}\mathbf{C}^\top$, *where*

$$\mathbf{\Lambda}=\operatorname{diag}(\lambda_1,\dots,\lambda_n),\qquad \mathbf{D}=\operatorname{diag}(d_1,\dots,d_n)$$

. *If* $\theta_i:=\lambda_i d_i$, $\mathbf{v}_i:=\dfrac{\mathbf{c}_i}{\|\mathbf{c}_i\|_2}$, $\mathbf{V}:=[\mathbf{v}_1,\dots,\mathbf{v}_n]$, *then*

$$\operatorname{cond}(\mathbf{V})\le\operatorname{cond}(\mathbf{C})^2\le 9\kappa_0^2,\qquad \rho_\ell:=\frac{\|\mathbf{\Sigma}_x\|_2\|\mathbf{\Sigma}_g\|_2}{\max_i\theta_i}\le (n+q)^2\omega^2. \tag{19}$$

Under the standing assumptions, this simplifies to $\operatorname{cond}(\mathbf{V})\le 9\kappa^2$, $\rho_\ell\le C_{k,L}\kappa^8 r^3$.

Proof. Set $p(\mathbf{z}):=G(\mathbf{z})^{1/q}$. Because G is positive definite, convex, even, and q-homogeneous, p is a norm whose unit ball is K. Homogeneity of Lebesgue measure then gives $\mathbb{P}(p(\mathbf{Z})\le t)=t^n, 0\le t\le 1$. Consequently, for every $a>-n$,

$$\mathbb{E}[p(\mathbf{Z})^a]=\int_0^1 t^a\,d(t^n)=n\int_0^1 t^{n+a-1}\,dt=\frac{n}{n+a}.$$

The inclusions $\rho B_2^n\subseteq K\subseteq RB_2^n$ are equivalent to $\rho p(\mathbf{z})\le\|\mathbf{z}\|_2\le Rp(\mathbf{z}), \mathbf{z}\in\mathbb{R}^n$. Therefore

$$\operatorname{tr}(\mathbf{\Sigma}_x)=\mathbb{E}\|\mathbf{Z}\|_2^2\ge\rho^2\mathbb{E}[p(\mathbf{Z})^2]=\rho^2\frac{n}{n+2}.$$

Since $\mathbf{\Sigma}_x\succeq\mathbf{0}$,

$$\|\mathbf{\Sigma}_x\|_2\ge\frac{1}{n}\operatorname{tr}(\mathbf{\Sigma}_x)\ge\frac{\rho^2}{n+2}.$$

On the other hand, $\mathbf{Z}\in K\subseteq RB_2^n$ almost surely, so $\mathbf{0}\preceq\mathbf{Z}\mathbf{Z}^\top\preceq R^2\mathbf{I}$ and hence $\|\mathbf{\Sigma}_x\|_2\le R^2$. These estimates also give

$$\frac{\sup_{\mathbf{z}\in K}\|\mathbf{z}\mathbf{z}^\top\|_2}{\|\mathbf{\Sigma}_x\|_2}\le(n+2)\frac{R^2}{\rho^2}=(n+2)\omega^2.$$

For $\mathbf{z}\neq\mathbf{0}$, differentiation of $G=p^q$ gives $\nabla G(\mathbf{z})=qp(\mathbf{z})^{q-1}\nabla p(\mathbf{z})$. Let p_* denote the dual norm. At every differentiability point of the norm p, $p_*(\nabla p(\mathbf{z}))=1$. The inclusion $\rho B_2^n\subseteq K$ implies

$$p_*(\mathbf{v})=\sup_{p(\mathbf{x})\le 1}\mathbf{v}^\top\mathbf{x}\ge\rho\|\mathbf{v}\|_2.$$

Thus $\|\nabla p(\mathbf{z})\|_2\le\dfrac{1}{\rho}$, and consequently $\|\nabla G(\mathbf{z})\|_2\le\dfrac{q}{\rho}p(\mathbf{z})^{q-1}\le\dfrac{q}{\rho}, \mathbf{z}\in K$. This proves $\|\mathbf{\Sigma}_g\|_2\le\dfrac{q^2}{\rho^2}$. For the reverse estimate, Euler's identity gives $\mathbf{z}^\top\nabla G(\mathbf{z})=qG(\mathbf{z})=qp(\mathbf{z})^q$. Since $\|\mathbf{z}\|_2\le Rp(\mathbf{z})$, Cauchy–Schwarz yields $\|\nabla G(\mathbf{z})\|_2\ge\dfrac{q}{R}p(\mathbf{z})^{q-1}$. Therefore

$$\operatorname{tr}(\mathbf{\Sigma}_g)=\mathbb{E}\|\nabla G(\mathbf{Z})\|_2^2\ge\frac{q^2}{R^2}\mathbb{E}[p(\mathbf{Z})^{2q-2}]=\frac{q^2}{R^2}\frac{n}{n+2q-2}.$$

Hence

$$\|\mathbf{\Sigma}_g\|_2\ge\frac{1}{n}\operatorname{tr}(\mathbf{\Sigma}_g)\ge\frac{q^2}{R^2(n+2q-2)}.$$

Combining the upper envelope $\|\nabla G(\mathbf{z})\|_2^2\le q^2/\rho^2$ with this lower bound gives

$$\frac{\sup_{\mathbf{z}\in K}\|\nabla G(\mathbf{z})\nabla G(\mathbf{z})^\top\|_2}{\|\mathbf{\Sigma}_g\|_2}\le(n+2q-2)\omega^2.$$

This proves (18).

We next control the spectral pencil. The random vector $\mathbf{U} = \mathbf{C}^\top \mathbf{Z}$ is uniform on $H_\ell := \{\mathbf{u} \in \mathbb{R}^n : h^{(\ell)}(\mathbf{u}) \le 1\}$. Coordinatewise sign symmetry implies that $\mathbf{\Lambda} = \mathbb{E}[\mathbf{U}\mathbf{U}^\top]$ and

$$\mathbf{D} = \mathbb{E}[\nabla h^{(\ell)}(\mathbf{U}) \nabla h^{(\ell)}(\mathbf{U})^\top]$$

are diagonal. Let $h := h^{(\ell)}$. On ∂H_ℓ, $h(\mathbf{u}) = 1$. Euler's identity gives $\mathbf{u}^\top \nabla h(\mathbf{u}) = q, \qquad \mathbf{u} \in \partial H_\ell$, so ∇h doesn't vanish on the boundary. The divergence theorem is applicable. For every coordinate i,

$$\int_{H_\ell} \partial_i (u_i h(\mathbf{u}))\, d\mathbf{u} = \int_{\partial H_\ell} u_i h(\mathbf{u}) \nu_i(\mathbf{u})\, dS(\mathbf{u}) = \int_{\partial H_\ell} u_i \nu_i(\mathbf{u})\, dS(\mathbf{u}) = |H_\ell|.$$

Expanding the derivative on the left gives

$$\int_{H_\ell} h(\mathbf{u})\, d\mathbf{u} + \int_{H_\ell} u_i \partial_i h(\mathbf{u})\, d\mathbf{u} = |H_\ell|.$$

The same radial integration used above gives $\frac{1}{|H_\ell|}\int_{H_\ell} h(\mathbf{u})\, d\mathbf{u} = \frac{n}{n+q}$. It follows that $\mathbb{E}[U_i \partial_i h(\mathbf{U})] = 1 - \frac{n}{n+q} = \frac{q}{n+q}$. By Cauchy–Schwarz,

$$\theta_i = \mathbb{E}[U_i^2]\mathbb{E}[(\partial_i h(\mathbf{U}))^2] \ge \big(\mathbb{E}[U_i \partial_i h(\mathbf{U})]\big)^2 = \frac{q^2}{(n+q)^2}.$$

Consequently,

$$\rho_\ell = \frac{\|\mathbf{\Sigma}_x\|_2 \|\mathbf{\Sigma}_g\|_2}{\max_i \theta_i} \le \frac{R^2(q^2/\rho^2)}{q^2/(n+q)^2} = (n+q)^2 \omega^2.$$

Finally, define $\mathbf{S} := \mathrm{diag}(\|\mathbf{c}_1\|_2, \dots, \|\mathbf{c}_n\|_2)$. Then $\mathbf{V} = \mathbf{C}\mathbf{S}^{-1}$. For every i, $\sigma_{\min}(\mathbf{C}) \le \|\mathbf{c}_i\|_2 \le \sigma_{\max}(\mathbf{C})$, so

$$\|\mathbf{S}\|_2 \le \sigma_{\max}(\mathbf{C}), \qquad \|\mathbf{S}^{-1}\|_2 \le \sigma_{\min}(\mathbf{C})^{-1}.$$

Therefore

$$\mathrm{cond}(\mathbf{V}) \le \mathrm{cond}(\mathbf{C})\,\mathrm{cond}(\mathbf{S}) \le \mathrm{cond}(\mathbf{C})^2.$$

Lemma B.1 gives $\mathrm{cond}(\mathbf{C}) \le 3\kappa_0$, and hence $\mathrm{cond}(\mathbf{V}) \le 9\kappa_0^2$. Together with

$$\omega \le C_{k,L} \kappa^4 \sqrt{r}, \qquad n \le r, \qquad q = O_{k,L}(1),$$

the pencil bound becomes $\rho_\ell \le C_{k,L} \kappa^8 r^3$. This proves (19) and its stated simplification. □

B.4 Stable gradient and empirical moment estimation

We finally convert multiplicative suffix-oracle accuracy and approximate sampling into accurate empirical estimates of the two moment matrices used by ASPIRE.

The next lemma shows that polynomial interpolation converts a multiplicative function-value error into a controlled additive error in the recovered gradient.

Lemma B.6 (Stable gradient recovery from a multiplicative oracle)**.** *Suppose an available oracle satisfies*

$$|\widehat{G}_\ell(\mathbf{z}) - G_\ell(\mathbf{z})| \le \xi_\ell G_\ell(\mathbf{z}), \qquad \mathbf{z} \in \mathbb{R}^{r_\ell}.$$

Then, for every $\mathbf{z} \in K_\ell^\sharp$, one can construct $\widehat{\mathbf{g}}_\ell(\mathbf{z})$ using $O(q_\ell r_\ell)$ evaluations of $\widehat{G}_\ell$ such that

$$\|\widehat{\mathbf{g}}_\ell(\mathbf{z}) - \nabla G_\ell(\mathbf{z})\|_2 \le C_{q_\ell} \frac{\sqrt{r_\ell}}{r_{\mathrm{in},\ell}} \xi_\ell.$$

Proof. Write $n := r_\ell, q := q_\ell, \rho := r_{\mathrm{in},\ell}$. Fix distinct nodes $t_0, \dots, t_q \in [-1, 1]$, and let $L_0, \dots, L_q$ be the associated Lagrange basis polynomials. Set $\beta_a := L_a'(0)$. For $1 \le j \le n$, define $P_j(t) := G_\ell\Big(\mathbf{z} + \frac{\rho}{2} t \mathbf{e}_j\Big)$. Since G_ℓ has degree q, $P_j'(0) = \sum_{a=0}^q \beta_a P_j(t_a)$. Define

$$\widehat{g}_{\ell,j}(\mathbf{z}) := \frac{2}{\rho} \sum_{a=0}^q \beta_a \widehat{G}_\ell\Big(\mathbf{z} + \frac{\rho}{2} t_a \mathbf{e}_j\Big), \qquad p_\ell := G_\ell^{1/q}$$

Since $p_\ell(\mathbf{z}) \le 1, p_\ell(\rho \mathbf{e}_j) \le 1$, we have $p_\ell\Big(\mathbf{z} + \frac{\rho}{2} t_a \mathbf{e}_j\Big) \le \frac{3}{2}$. Hence $G_\ell\left(\mathbf{z} + \frac{\rho}{2} t_a \mathbf{e}_j\right) \le \left(\frac{3}{2}\right)^q$. Therefore

$$|\widehat{g}_{\ell,j}(\mathbf{z}) - \partial_j G_\ell(\mathbf{z})| \le C_q \frac{\xi_\ell}{\rho}.$$

Summing over the coordinates proves $\|\widehat{\mathbf{g}}_\ell(\mathbf{z}) - \nabla G_\ell(\mathbf{z})\|_2 \le C_q \frac{\sqrt{n}}{\rho} \xi_\ell$. □

Combining gradient stability, total-variation control of the sampling distribution, and matrix concentration, the final lemma gives simultaneous relative-error guarantees for the two empirical moments.

Lemma B.7 (Empirical recovery of the two moments)**.** *Fix $1 \le \ell \le L-2$, write $n := r_\ell, q := q_\ell, \omega := \omega_\ell$, and let $\pi_\ell := \mathrm{Unif}(K_\ell^\sharp)$. Suppose the endpoint distribution $\widehat{\pi}_\ell$ of each independent noisy Hit-and-Run chain satisfies*

$$d_{\mathrm{TV}}(\widehat{\pi}_\ell, \pi_\ell) \le \nu,$$

and suppose that

$$|\widehat{G}_\ell(\mathbf{z}) - G_\ell(\mathbf{z})| \le \xi_\ell G_\ell(\mathbf{z}), \qquad \mathbf{z} \in \mathbb{R}^n.$$

Let $\eta, \zeta \in (0, 1/2)$. There are constants $c_q, C_q > 0$, depending only on q, such that if

$$\nu \le \frac{c_q \eta}{(n+2q)\omega^2}, \qquad \xi_\ell \le \frac{c_q \eta}{\sqrt{n}\,(n+2q)\omega^2}, \tag{20}$$

and

$$m \ge C_q (n+2q)\omega^2 \eta^{-2} \log \frac{2n}{\zeta}, \tag{21}$$

then m independent endpoints are sufficient to construct

$$\widehat{\mathbf{\Sigma}}_{x,\ell} = \frac{1}{m} \sum_{i=1}^m \mathbf{Z}_i \mathbf{Z}_i^\top, \qquad \widehat{\mathbf{\Sigma}}_{g,\ell} = \frac{1}{m} \sum_{i=1}^m \widehat{\mathbf{g}}_\ell(\mathbf{Z}_i) \widehat{\mathbf{g}}_\ell(\mathbf{Z}_i)^\top$$

such that, with probability at least $1-\zeta$,

$$\|\widehat{\mathbf{\Sigma}}_{x,\ell} - \mathbf{\Sigma}_{x,\ell}\|_2 \le \eta \|\mathbf{\Sigma}_{x,\ell}\|_2, \qquad \|\widehat{\mathbf{\Sigma}}_{g,\ell} - \mathbf{\Sigma}_{g,\ell}\|_2 \le \eta \|\mathbf{\Sigma}_{g,\ell}\|_2.$$

In particular, under the standing assumptions it is sufficient that

$$\nu \le c_{k,L} \kappa^{-8} \eta r^{-2}, \qquad \xi_\ell \le c_{k,L} \kappa^{-8} \eta r^{-5/2}, \qquad m \ge C_{k,L} \kappa^8 r^2 \eta^{-2} \log \frac{2r}{\zeta}.$$

Proof. Write $\rho := r_{\mathrm{in},\ell}, R := R_{\mathrm{out},\ell}, \mathbf{g}(\mathbf{z}) := \nabla G_\ell(\mathbf{z}), \mathbf{e}(\mathbf{z}) := \widehat{\mathbf{g}}_\ell(\mathbf{z}) - \mathbf{g}(\mathbf{z})$. We first control the bias caused by approximate sampling and by the multiplicative suffix oracle. For the position moment, every point of $K_\ell^\sharp$ has norm at most R, so $\|\mathbf{Z}\mathbf{Z}^\top\|_2 \le R^2$. The total-variation bound therefore implies $\big\|\mathbb{E}_{\widehat{\pi}_\ell}[\mathbf{Z}\mathbf{Z}^\top] - \mathbb{E}_{\pi_\ell}[\mathbf{Z}\mathbf{Z}^\top]\big\|_2 \le 2R^2\nu$. Using $\|\mathbf{\Sigma}_{x,\ell}\|_2 \ge \frac{\rho^2}{n+2}$, we obtain

$$\frac{\big\|\mathbb{E}_{\widehat{\pi}_\ell}[\mathbf{Z}\mathbf{Z}^\top] - \mathbf{\Sigma}_{x,\ell}\big\|_2}{\|\mathbf{\Sigma}_{x,\ell}\|_2} \le 2(n+2)\omega^2\nu.$$

For the gradient moment, Lemmas B.5 and B.6 give, for every $\mathbf{z} \in K_\ell^\sharp$, $\|\mathbf{g}(\mathbf{z})\|_2 \le \frac{q}{\rho}$ and $\|\mathbf{e}(\mathbf{z})\|_2 \le C_q \frac{\sqrt{n}}{\rho} \xi_\ell$. The second inequality in (20) implies, after decreasing c_q, that $\|\mathbf{e}(\mathbf{z})\|_2 \le \frac{q}{\rho}$. Consequently,

$$\|\widehat{\mathbf{g}}_\ell(\mathbf{z})\|_2 \le \|\mathbf{g}(\mathbf{z})\|_2 + \|\mathbf{e}(\mathbf{z})\|_2 \le \frac{2q}{\rho}.$$

Moreover,

$$\big\|\widehat{\mathbf{g}}_\ell(\mathbf{z})\widehat{\mathbf{g}}_\ell(\mathbf{z})^\top - \mathbf{g}(\mathbf{z})\mathbf{g}(\mathbf{z})^\top\big\|_2 \le \|\mathbf{g}(\mathbf{z})\mathbf{e}(\mathbf{z})^\top\|_2 + \|\mathbf{e}(\mathbf{z})\mathbf{g}(\mathbf{z})^\top\|_2 + \|\mathbf{e}(\mathbf{z})\mathbf{e}(\mathbf{z})^\top\|_2 \le C_q \frac{\sqrt{n}}{\rho^2} \xi_\ell.$$

Define the population moment under the noisy endpoint distribution by $\overline{\mathbf{\Sigma}}_{g,\ell} := \mathbb{E}_{\widehat{\pi}_\ell}\left[\widehat{\mathbf{g}}_\ell(\mathbf{Z})\widehat{\mathbf{g}}_\ell(\mathbf{Z})^\top\right]$. Insert and subtract $\mathbb{E}_{\widehat{\pi}_\ell}[\mathbf{g}(\mathbf{Z})\mathbf{g}(\mathbf{Z})^\top]$. The gradient-oracle contribution is bounded by

$$\left\|\mathbb{E}_{\widehat{\pi}_\ell}[\widehat{\mathbf{g}}_\ell\widehat{\mathbf{g}}_\ell^\top - \mathbf{g}\mathbf{g}^\top]\right\|_2 \le C_q\frac{\sqrt{n}}{\rho^2}\xi_\ell.$$

The total-variation contribution is bounded by

$$\left\|\mathbb{E}_{\widehat{\pi}_\ell}[\mathbf{g}\mathbf{g}^\top] - \mathbb{E}_{\pi_\ell}[\mathbf{g}\mathbf{g}^\top]\right\|_2 \le 2\frac{q^2}{\rho^2}\nu.$$

Since $\|\mathbf{\Sigma}_{g,\ell}\|_2 \ge \dfrac{q^2}{R^2(n+2q-2)}$, we conclude that

$$\frac{\|\overline{\mathbf{\Sigma}}_{g,\ell} - \mathbf{\Sigma}_{g,\ell}\|_2}{\|\mathbf{\Sigma}_{g,\ell}\|_2} \le C_q(n+2q)\omega^2\left(\nu + \sqrt{n}\,\xi_\ell\right).$$

Together with the position-moment bound, the two conditions in (20) imply

$$\left\|\mathbb{E}_{\widehat{\pi}_\ell}[\mathbf{Z}\mathbf{Z}^\top] - \mathbf{\Sigma}_{x,\ell}\right\|_2 \le \frac{\eta}{2}\|\mathbf{\Sigma}_{x,\ell}\|_2, \qquad \|\overline{\mathbf{\Sigma}}_{g,\ell} - \mathbf{\Sigma}_{g,\ell}\|_2 \le \frac{\eta}{2}\|\mathbf{\Sigma}_{g,\ell}\|_2.$$

It remains to control finite-sample fluctuations. Set $\overline{\mathbf{\Sigma}}_{x,\ell} := \mathbb{E}_{\widehat{\pi}_\ell}[\mathbf{Z}\mathbf{Z}^\top]$. The population-bias estimates imply

$$\|\overline{\mathbf{\Sigma}}_{x,\ell}\|_2 \le \left(1+\frac{\eta}{2}\right)\|\mathbf{\Sigma}_{x,\ell}\|_2 \le \frac{3}{2}\|\mathbf{\Sigma}_{x,\ell}\|_2, \qquad \|\overline{\mathbf{\Sigma}}_{g,\ell}\|_2 \le \frac{3}{2}\|\mathbf{\Sigma}_{g,\ell}\|_2.$$

We first consider the position moment. Let $\mathbf{X}_i := \mathbf{Z}_i\mathbf{Z}_i^\top, \mathbf{S}_i := \mathbf{X}_i - \overline{\mathbf{\Sigma}}_{x,\ell}$. Because each noisy Hit-and-Run endpoint lies in $K_\ell^\sharp \subseteq RB_2^n$, $\mathbf{0} \preceq \mathbf{X}_i \preceq R^2\mathbf{I}$. Furthermore, $\mathbf{X}_i^2 = \|\mathbf{Z}_i\|_2^2\mathbf{X}_i \preceq R^2\mathbf{X}_i$. We have $\mathbb{E}\mathbf{S}_i = \mathbf{0}$ and $\|\mathbf{S}_i\|_2 \le R^2$. The matrix variance satisfies

$$\mathbb{E}[\mathbf{S}_i^2] = \mathbb{E}[\mathbf{X}_i^2] - \overline{\mathbf{\Sigma}}_{x,\ell}^2 \preceq \mathbb{E}[\mathbf{X}_i^2] \preceq R^2\overline{\mathbf{\Sigma}}_{x,\ell}.$$

Hence

$$\left\|\sum_{i=1}^m \mathbb{E}[\mathbf{S}_i^2]\right\|_2 \le mR^2\|\overline{\mathbf{\Sigma}}_{x,\ell}\|_2.$$

The self-adjoint matrix Bernstein inequality gives, for every $t > 0$,

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m \mathbf{X}_i - \overline{\mathbf{\Sigma}}_{x,\ell}\right\|_2 \ge t\right) \le 2n\exp\left(-\frac{mt^2}{2R^2\|\overline{\mathbf{\Sigma}}_{x,\ell}\|_2 + \frac{2}{3}R^2t}\right).$$

Take $t := \dfrac{\eta}{2}\|\mathbf{\Sigma}_{x,\ell}\|_2$. Using $\eta < 1/2$ and the bound on $\|\overline{\mathbf{\Sigma}}_{x,\ell}\|_2$, we obtain

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}_{x,\ell} - \overline{\mathbf{\Sigma}}_{x,\ell}\|_2 \ge \frac{\eta}{2}\|\mathbf{\Sigma}_{x,\ell}\|_2\right) \le 2n\exp\left(-cm\eta^2\frac{\|\mathbf{\Sigma}_{x,\ell}\|_2}{R^2}\right),$$

where $c > 0$ is an absolute constant. Since $\dfrac{R^2}{\|\mathbf{\Sigma}_{x,\ell}\|_2} \le (n+2)\omega^2$, it follows that $m \ge C_q(n+2q)\omega^2\eta^{-2}\log\dfrac{2n}{\zeta}$ makes the position fluctuation at most $\dfrac{\eta}{2}\|\mathbf{\Sigma}_{x,\ell}\|_2$ with failure probability at most $\zeta/2$, after changing C_q by an absolute factor.

We next consider the gradient moment. Let $\mathbf{Y}_i := \widehat{\mathbf{g}}_\ell(\mathbf{Z}_i)\widehat{\mathbf{g}}_\ell(\mathbf{Z}_i)^\top, \mathbf{T}_i := \mathbf{Y}_i - \overline{\mathbf{\Sigma}}_{g,\ell}$. The uniform gradient bound gives $\mathbf{0} \preceq \mathbf{Y}_i \preceq \dfrac{4q^2}{\rho^2}\mathbf{I}$ and $\mathbf{Y}_i^2 = \|\widehat{\mathbf{g}}_\ell(\mathbf{Z}_i)\|_2^2\mathbf{Y}_i \preceq \dfrac{4q^2}{\rho^2}\mathbf{Y}_i$. Thus $\mathbb{E}\mathbf{T}_i = \mathbf{0}$, $\|\mathbf{T}_i\|_2 \le \dfrac{4q^2}{\rho^2}$, and $\mathbb{E}[\mathbf{T}_i^2] = \mathbb{E}[\mathbf{Y}_i^2] - \overline{\mathbf{\Sigma}}_{g,\ell}^2 \preceq \mathbb{E}[\mathbf{Y}_i^2] \preceq \dfrac{4q^2}{\rho^2}\overline{\mathbf{\Sigma}}_{g,\ell}$. Consequently,

$$\left\|\sum_{i=1}^m \mathbb{E}[\mathbf{T}_i^2]\right\|_2 \le \frac{4q^2m}{\rho^2}\|\overline{\mathbf{\Sigma}}_{g,\ell}\|_2.$$

Matrix Bernstein now yields

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}_{g,\ell} - \overline{\mathbf{\Sigma}}_{g,\ell}\|_2 \geq t\right) \leq 2n \exp\left(-\frac{mt^2}{C_q \rho^{-2}(\|\overline{\mathbf{\Sigma}}_{g,\ell}\|_2 + t)}\right).$$

Taking $t := \frac{\eta}{2}\|\mathbf{\Sigma}_{g,\ell}\|_2$ gives

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}_{g,\ell} - \overline{\mathbf{\Sigma}}_{g,\ell}\|_2 \geq \frac{\eta}{2}\|\mathbf{\Sigma}_{g,\ell}\|_2\right) \leq 2n \exp\left(-c_q m \eta^2 \rho^2 \|\mathbf{\Sigma}_{g,\ell}\|_2\right).$$

By Lemma B.5, $\|\mathbf{\Sigma}_{g,\ell}\|_2 \geq \dfrac{q^2}{R^2(n+2q-2)}$. Therefore

$$\frac{1}{\rho^2 \|\mathbf{\Sigma}_{g,\ell}\|_2} \leq \frac{R^2(n+2q-2)}{\rho^2 q^2} \leq C_q(n+2q)\omega^2.$$

The same sample-size condition thus makes the gradient fluctuation at most $\dfrac{\eta}{2}\|\mathbf{\Sigma}_{g,\ell}\|_2$ with failure probability at most $\zeta/2$.

Combining the two concentration events with the two population-bias bounds and applying a union bound gives, with probability at least $1-\zeta$,

$$\|\widehat{\mathbf{\Sigma}}_{x,\ell} - \mathbf{\Sigma}_{x,\ell}\|_2 \leq \eta \|\mathbf{\Sigma}_{x,\ell}\|_2, \qquad \|\widehat{\mathbf{\Sigma}}_{g,\ell} - \mathbf{\Sigma}_{g,\ell}\|_2 \leq \eta \|\mathbf{\Sigma}_{g,\ell}\|_2.$$

Finally, Lemma B.2 gives $\omega^2 \leq C_{k,L}\kappa^8 r$. Since $n \leq r$ and $q = O_{k,L}(1)$, $(n+2q)\omega^2 \leq C_{k,L}\kappa^8 r^2$ and $\sqrt{n}\,(n+2q)\omega^2 \leq C_{k,L}\kappa^8 r^{5/2}$. Substituting these two inequalities into (20) and (21) gives the three sufficient conditions stated at the end of the lemma. □

C Recovery of the first hidden layer

This section establishes recovery of the first hidden layer, from identifying its column space to recovering the individual weight directions.

Proposition C.1 (Exact recovery of the first-layer column space)**.** *Suppose that $\mathbf{W}_1 \in \mathbb{R}^{d\times r}$ has full column rank and*

$$f^\star(\mathbf{x}) = h^{(1)}(\mathbf{W}_1^\top \mathbf{x}),$$

where $h^{(1)} : \mathbb{R}^r \to \mathbb{R}$ is a positive definite homogeneous polynomial. Assume exact function-value queries to $f^\star$, so that $\nabla f^\star(\mathbf{x})$ can be recovered exactly by polynomial interpolation.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$, $m \geq r$, be independently sampled from an absolutely continuous distribution whose support contains a nonempty open set. Define

$$\mathbf{g}_i := \nabla f^\star(\mathbf{x}_i), \qquad \boldsymbol{G}_1 := [\mathbf{g}_1, \ldots, \mathbf{g}_m].$$

Then, with probability one, $\mathrm{col}(\boldsymbol{G}_1) = \mathrm{col}(\mathbf{W}_1)$. *Hence* $\mathrm{col}(\mathbf{W}_1)$ *is recovered exactly from the top r left singular vectors of $\boldsymbol{G}_1$.*

Proof. By the chain rule, $\mathbf{g}_i = \mathbf{W}_1 \nabla h^{(1)}(\mathbf{W}_1^\top \mathbf{x}_i)$, and therefore $\mathrm{col}(\boldsymbol{G}_1) \subseteq \mathrm{col}(\mathbf{W}_1)$. Since $h^{(1)}$ is positive definite and homogeneous,

$$\mathrm{span}\left\{\nabla h^{(1)}(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^r\right\} = \mathbb{R}^r.$$

Indeed, otherwise there would exist $\mathbf{c} \neq \mathbf{0}$ such that $\mathbf{c}^\top \nabla h^{(1)}(\mathbf{u}) \equiv 0$. The function $t \mapsto h^{(1)}(\mathbf{u} + t\mathbf{c})$ would then have zero derivative for every $\mathbf{u}$ and t, and hence would be constant. Taking $\mathbf{u} = \mathbf{0}$ would give $h^{(1)}(t\mathbf{c}) = h^{(1)}(\mathbf{0}) = 0$ for every t, contradicting positive definiteness.

It follows that there exist $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^r$ for which

$$\det\left[\nabla h^{(1)}(\mathbf{u}_1), \ldots, \nabla h^{(1)}(\mathbf{u}_r)\right] \neq 0.$$

Define the polynomial on $(\mathbb{R}^d)^r$ by

$$\Psi(\mathbf{x}_1, \dots, \mathbf{x}_r) := \det\Big[\nabla h^{(1)}(\mathbf{W}_1^\top \mathbf{x}_1), \dots, \nabla h^{(1)}(\mathbf{W}_1^\top \mathbf{x}_r)\Big].$$

Because $\mathbf{W}_1^\top$ is surjective, for each $\mathbf{u}_i$ one can choose $\mathbf{x}_i$ satisfying $\mathbf{W}_1^\top \mathbf{x}_i = \mathbf{u}_i$. Hence Ψ is not the zero polynomial. The zero set of a nonzero real polynomial has Lebesgue measure zero, while the joint law of $(\mathbf{x}_1, \dots, \mathbf{x}_r)$ is absolutely continuous. Therefore $\mathbb{P}\big(\Psi(\mathbf{x}_1, \dots, \mathbf{x}_r) = 0\big) = 0$. Thus the first r latent gradients are linearly independent almost surely. Since $m \geq r$,

$$\operatorname{rank}\Big[\nabla h^{(1)}(\mathbf{W}_1^\top \mathbf{x}_1), \dots, \nabla h^{(1)}(\mathbf{W}_1^\top \mathbf{x}_m)\Big] = r$$

almost surely, which yields $\operatorname{col}(\boldsymbol{G}_1) = \operatorname{col}(\mathbf{W}_1)$. □

Thus, an orthonormal basis $\mathbf{B}_1$ of $\operatorname{col}(\mathbf{W}_1)$ can be recovered exactly.

Proposition C.2 (Stable recovery of the first hidden layer)**.** *Let $d \geq 2$ and $\varepsilon, \zeta \in (0, 1/2)$. Suppose that the standing assumptions hold. Then*

$$\widetilde{O}_{k,L}\left(dr + \frac{\gamma^2 r^{13} \kappa^{O(1)}}{\varepsilon^2}\right) \tag{22}$$

real function-value queries suffice to recover $\widehat{\mathbf{W}}_1$ such that, with probability at least $1 - \zeta$, after suitably relabeling the recovered columns and choosing their signs, $\|\widehat{\mathbf{W}}_1 - \mathbf{W}_1\|_2 \leq \varepsilon$.

Proof. By Proposition C.1, $\operatorname{col}(\mathbf{W}_1)$ is recovered exactly with probability one. Let $\mathbf{B}_1 \in \mathbb{R}^{d \times r}$ be an orthonormal basis of this subspace and set

$$\mathbf{C}_1 := \mathbf{B}_1^\top \mathbf{W}_1,\ \mathbf{C}_1^\sharp := \mathbf{C}_1,\ G_1(\mathbf{z}) := f^\star(\mathbf{B}_1 \mathbf{z}),\ K_1^\sharp := K_1 := \{\mathbf{z} : G_1(\mathbf{z}) \leq 1\}.$$

Then

$$\mathbf{W}_1 = \mathbf{B}_1 \mathbf{C}_1,\ \operatorname{cond}(\mathbf{C}_1) = \operatorname{cond}(\mathbf{W}_1) \leq \kappa_0 \leq \kappa,\ \|\mathbf{c}_{1,i}\|_2 = 1.$$

Moreover, Lemma B.2 gives

$$\omega_1 := \frac{R_{\text{out},1}}{r_{\text{in},1}} \leq C_{k,L} \kappa^4 \sqrt{r}.$$

Let $\mathbf{Z} \sim \operatorname{Unif}(K_1)$ and define

$$\boldsymbol{\Sigma}_{x,1} := \mathbb{E}[\mathbf{Z}\mathbf{Z}^\top], \qquad \boldsymbol{\Sigma}_{g,1} := \mathbb{E}[\nabla G_1(\mathbf{Z}) \nabla G_1(\mathbf{Z})^\top].$$

Lemma B.5 yields

$$\boldsymbol{\Sigma}_{x,1} = \mathbf{C}_1^{-\top} \boldsymbol{\Lambda}_1 \mathbf{C}_1^{-1}, \qquad \boldsymbol{\Sigma}_{g,1} = \mathbf{C}_1 \mathbf{D}_1 \mathbf{C}_1^\top,$$

where $\mathbf{D}_1, \boldsymbol{\Lambda}_1$ are positive diagonal matrices and $\mathbf{D}_1 \boldsymbol{\Lambda}_1 = \operatorname{diag}(\theta_{1,1}, \dots, \theta_{1,r})$. Furthermore, $\rho_1 := \dfrac{\|\boldsymbol{\Sigma}_{x,1}\|_2 \|\boldsymbol{\Sigma}_{g,1}\|_2}{\max_i |\theta_{1,i}|} \leq C_{k,L} \kappa^8 r^3$. Set $\eta := c_{k,L} \dfrac{\varepsilon}{\gamma \kappa^{12} r^{7/2}}, \nu := c_{k,L} \kappa^{-8} \eta r^{-2}$. Since G_1 is available exactly, it satisfies

$$(1 - \xi) G_1(\mathbf{z}) \leq G_1(\mathbf{z}) \leq (1 + \xi) G_1(\mathbf{z})$$

for every $\xi \in (0, 1/2]$. Let $q_1 := k^{L-1}$, and apply Lemma B.4 with $\delta_{\text{HR}} = 1$ and target total-variation error ν. Let $T_{\text{HR}} := \left\lceil C_{\text{HR}} r^3 \omega_1^2 \log \dfrac{2\omega_1}{\nu} \right\rceil$ be the number of Hit-and-Run steps for each independent chain, where $C_{\text{HR}} > 0$ is the absolute constant supplied by Lemma B.4, and set $\eta_{\text{HR}} := \dfrac{\nu}{4T_{\text{HR}}}$. Define

$$\xi_{\text{aux}} := \min\left\{\frac{1}{4},\ \frac{c_1 \nu^2}{2 q_1^2 r^7 \omega_1^6 \left(\log \frac{2\omega_1}{\nu}\right)^2}\right\},$$

where $c_1 > 0$ is the constant supplied by Lemma B.4 for $\delta_{\text{HR}} = 1$. Then the multiplicative-oracle hypothesis of that lemma holds with $\xi = \xi_{\text{aux}}$. Moreover, if

$$\tau_{\text{aux}} := 1 - \left(\frac{1 - \xi_{\text{aux}}}{1 + \xi_{\text{aux}}}\right)^{1/q_1},$$

then, since $0 < \xi_{\mathrm{aux}} \le 1/4$, $\tau_{\mathrm{aux}} \ge c_{q_1}\xi_{\mathrm{aux}}$ for a constant $c_{q_1} > 0$ depending only on q_1. Thus $\log \dfrac{\omega_1}{\eta_{\mathrm{HR}}\tau_{\mathrm{aux}}}$ is logarithmic in $r, \kappa, \gamma, \varepsilon^{-1}, \zeta^{-1}$. Using $\omega_1^2 \le C_{k,L}\kappa^8 r$, Lemma B.4 therefore generates each sample within total variation ν of $\mathrm{Unif}(K_1)$ using $\widetilde{O}_{k,L}(\kappa^8 r^4)$ function-value queries.

Since the oracle is exact, the gradients ∇G_1 are recovered exactly by polynomial interpolation. Applying Lemma B.7 with $\widehat{G}_1 = G_1$ and $\xi_1 = 0$ shows that

$$m = O_{k,L}\left(\kappa^8 r^2 \eta^{-2} \log \frac{2r}{\zeta}\right) = O_{k,L}\left(\frac{\gamma^2 \kappa^{32} r^9}{\varepsilon^2} \log \frac{2r}{\zeta}\right)$$

samples suffice to obtain relative η-accurate estimates of both population moment matrices, with probability at least $1 - \zeta$.

The normalized eigenvector matrix in Lemma A.6 is precisely $\mathbf{C}_1$, because its columns have unit Euclidean norm, and hence its condition number is at most κ. Assumption 3.5 and the choice of η therefore imply that, after relabeling and choosing signs,

$$\max_i \|\widehat{\mathbf{c}}_{1,i} - \mathbf{c}_{1,i}\|_2 \le C_{k,L}\gamma\kappa^{12} r^3 \eta \le c_{k,L}\frac{\varepsilon}{\sqrt{r}}.$$

Therefore

$$\|\widehat{\mathbf{C}}_1 - \mathbf{C}_1\|_2 \le \|\widehat{\mathbf{C}}_1 - \mathbf{C}_1\|_F \le \sqrt{r}\max_i \|\widehat{\mathbf{c}}_{1,i} - \mathbf{c}_{1,i}\|_2 \le \varepsilon.$$

Setting $\widehat{\mathbf{W}}_1 := \mathbf{B}_1\widehat{\mathbf{C}}_1$ and using $\mathbf{B}_1^\top \mathbf{B}_1 = \mathbf{I}_r$ gives the claimed operator-norm error.

Finally, exact recovery of the first-layer column space uses r gradients, each obtained with $O_{k,L}(d)$ real function-value queries, for a total of $O_{k,L}(dr)$. The sampling stage uses

$$m\,\widetilde{O}_{k,L}(\kappa^8 r^4) = \widetilde{O}_{k,L}\left(\frac{\gamma^2\kappa^{40} r^{13}}{\varepsilon^2}\log\frac{1}{\zeta}\right)$$

queries. This proves the claim. □

D STABILITY OF THE SIMULATED SUFFIX ORACLE

This section quantifies how errors in the recovered prefix propagate through prefix inversion and the resulting simulated suffix oracle.

Lemma D.1 (Stability of approximate prefix inversion). *Fix $2 \le \ell \le L-1$. Suppose that, for every $1 \le j \le \ell - 1$,*

$$\|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \le \varepsilon_{<\ell}.$$

Define

$$\widehat{\mathbf{R}}_j := \widehat{\mathbf{W}}_j(\widehat{\mathbf{W}}_j^\top \widehat{\mathbf{W}}_j)^{-1}.$$

For $\mathbf{y} \in \mathbb{R}^{r_{\ell-1}}$, set $\mathbf{z}_{\ell-1} = \mathbf{y}$, and for $j = \ell-1, \dots, 1$, choose coordinatewise principal k-th root $\boldsymbol{\xi}_j \in \mathbb{C}^{r_j}$ such that $\boldsymbol{\xi}_j^{\odot k} = \mathbf{z}_j$, and set $\mathbf{z}_{j-1} = \widehat{\mathbf{R}}_j\boldsymbol{\xi}_j$. Starting from $\widetilde{\mathbf{z}}_0 = \mathbf{z}_0$, define the true forward iterates

$$\widetilde{\mathbf{z}}_j = (\mathbf{W}_j^\top \widetilde{\mathbf{z}}_{j-1})^{\odot k}, \qquad 1 \le j \le \ell - 1.$$

Let

$$\vartheta_k := \frac{k-1}{2k}, \qquad P_j := \sqrt{\frac{r}{r_j}}\prod_{t=1}^{j} r_t^{\vartheta_k}, \qquad 1 \le j \le \ell - 1.$$

There are constants $c_{k,L}, C_{k,L} > 0$, independent of the conditioning parameters, such that if

$$\kappa^{O(\ell)} P_{\ell-1}\varepsilon_{<\ell} \le c_{k,L},$$

then

$$\|\widetilde{\mathbf{z}}_{\ell-1} - \mathbf{y}\|_2 \le C_{k,L}\kappa^{O(\ell)} P_{\ell-1}\varepsilon_{<\ell}\|\mathbf{y}\|_2.$$

Since $P_{\ell-1} \le r^{\ell/2}$, it is sufficient that

$$\kappa^{O(\ell)} r^{\ell/2}\varepsilon_{<\ell} \le c_{k,L}, \tag{23}$$

in which case

$$\|\widetilde{\mathbf{z}}_{\ell-1} - \mathbf{y}\|_2 \le C_{k,L}\kappa^{O(\ell)} r^{\ell/2}\varepsilon_{<\ell}\|\mathbf{y}\|_2. \tag{24}$$

Proof. If $\mathbf{y} = \mathbf{0}$, choose every coordinatewise k-th root in the backward recursion to be zero. Then all backward and forward iterates are zero, and the conclusion is immediate. Hence assume $\mathbf{y} \neq \mathbf{0}$.

Lemma A.1 gives

$$\sigma_{\min}(\mathbf{W}_1) \geq \kappa_0^{-1}, \qquad \sigma_{\min}(\mathbf{W}_j) \geq \kappa_0^{-1}\sqrt{\frac{r_j}{r_{j-1}}}, \quad 2 \leq j \leq \ell - 1,$$

and $\|\mathbf{W}_j\|_2 \leq \kappa_0$. Since the widths are nonincreasing,

$$\sqrt{\frac{r_{j-1}}{r_j}} \leq \sqrt{\frac{r}{r_{\ell-1}}} \leq P_{\ell-1}.$$

Thus the assumed smallness condition, after decreasing $c_{k,L}$, makes $\varepsilon_{<\ell} \leq \sigma_{\min}(\mathbf{W}_j)/2$ for every $j < \ell$. Weyl's inequality gives full column rank of each $\widehat{\mathbf{W}}_j$, together with

$$\|\widehat{\mathbf{R}}_1\|_2 \leq 2\kappa_0, \qquad \|\widehat{\mathbf{R}}_j\|_2 \leq 2\kappa_0\sqrt{\frac{r_{j-1}}{r_j}}, \quad 2 \leq j \leq \ell - 1.$$

In particular, the nonzero terminal value implies that every backward iterate is nonzero.

Set

$$\varepsilon := \varepsilon_{<\ell}, \qquad \mathbf{e}_j := \widetilde{\mathbf{z}}_j - \mathbf{z}_j, \qquad E_j := \frac{\|\mathbf{e}_j\|_2}{\|\mathbf{z}_j\|_2}.$$

Then $E_0 = 0$. Define the layer factors

$$b_1 := r^{\vartheta_k}, \qquad b_j := \sqrt{\frac{r_{j-1}}{r_j}}\, r_j^{\vartheta_k}, \quad j \geq 2.$$

Their products telescope:

$$\prod_{t=1}^{j} b_t = \sqrt{\frac{r}{r_j}} \prod_{t=1}^{j} r_t^{\vartheta_k} = P_j.$$

Because $\widehat{\mathbf{W}}_j^\top \mathbf{z}_{j-1} = \boldsymbol{\xi}_j$, we may write

$$\mathbf{W}_j^\top \widetilde{\mathbf{z}}_{j-1} = \boldsymbol{\xi}_j + \boldsymbol{\Delta}_j,$$

where

$$\boldsymbol{\Delta}_j := (\mathbf{W}_j - \widehat{\mathbf{W}}_j)^\top \mathbf{z}_{j-1} + \mathbf{W}_j^\top \mathbf{e}_{j-1}.$$

Consequently,

$$\|\boldsymbol{\Delta}_j\|_2 \leq (\varepsilon + \kappa E_{j-1})\|\mathbf{z}_{j-1}\|_2.$$

The right-inverse bounds and the comparison between ℓ_2 and ℓ_{2k} give

$$\|\mathbf{z}_{j-1}\|_2 \leq C\kappa b_j \|\boldsymbol{\xi}_j\|_{2k}.$$

Thus

$$\|\boldsymbol{\Delta}_j\|_{2k} \leq C\kappa b_j(\varepsilon + \kappa E_{j-1})\|\boldsymbol{\xi}_j\|_{2k}.$$

For $a, b \in \mathbb{C}$,

$$|(a+b)^k - a^k| \leq k|b|(|a| + |b|)^{k-1}.$$

Applying this coordinatewise and using Hölder's inequality gives

$$\left\|(\boldsymbol{\xi}_j + \boldsymbol{\Delta}_j)^{\odot k} - \boldsymbol{\xi}_j^{\odot k}\right\|_2 \leq k\|\boldsymbol{\Delta}_j\|_{2k}\left(\|\boldsymbol{\xi}_j\|_{2k} + \|\boldsymbol{\Delta}_j\|_{2k}\right)^{k-1}.$$

Set

$$a_j := C\kappa b_j(\varepsilon + \kappa E_{j-1}).$$

The preceding estimate gives

$$\|\boldsymbol{\Delta}_j\|_{2k} \leq a_j\|\boldsymbol{\xi}_j\|_{2k}.$$

Whenever $a_j \leq 1$,

$$\|\mathbf{e}_j\|_2 = \left\|(\boldsymbol{\xi}_j + \boldsymbol{\Delta}_j)^{\odot k} - \boldsymbol{\xi}_j^{\odot k}\right\|_2 \leq k a_j(1 + a_j)^{k-1}\|\boldsymbol{\xi}_j\|_{2k}^k \leq C_k\kappa b_j(\varepsilon + \kappa E_{j-1})\|\boldsymbol{\xi}_j\|_{2k}^k.$$

Since $\|\mathbf{z}_j\|_2 = \|\boldsymbol{\xi}_j\|_{2k}^k$, we obtain

$$E_j \le C_k \kappa b_j (\varepsilon + \kappa E_{j-1}). \tag{25}$$

We verify the temporary smallness condition and solve the recursion simultaneously. We claim that constants $A_0, \dots, A_{\ell-1}$, depending only on k, L, can be chosen so that

$$E_j \le A_j \kappa^{3j} P_j \varepsilon, \qquad 0 \le j \le \ell - 1. \tag{26}$$

where $P_0 := 1$. The assertion is immediate for $j = 0$. If it holds through $j-1$, then $b_j P_{j-1} = P_j$, $P_{j-1} \ge 1$, and

$$C\kappa b_j(\varepsilon + \kappa E_{j-1}) \le C_{k,L} \kappa^{3j} P_j \varepsilon.$$

The assumed smallness condition makes the last quantity at most one, validating (25). Applying that recursion once more yields the claimed bound at j.

Taking $j = \ell - 1$ proves the profile-dependent estimate. Finally, because $r_j \le r$, $r_{\ell-1} \ge 1$, and $\vartheta_k = (k-1)/(2k) < 1/2$,

$$P_{\ell-1} = \sqrt{\frac{r}{r_{\ell-1}}} \prod_{j=1}^{\ell-1} r_j^{\vartheta_k} \le r^{1/2 + \vartheta_k(\ell-1)} \le r^{\ell/2}.$$

This proves the simplified statements. □

Lemma D.2 (Relative perturbation of a homogeneous polynomial)**.** *Let $G : \mathbb{R}^n \to \mathbb{R}$ be a positive-definite, even, homogeneous polynomial of degree q, and suppose that $p(\mathbf{x}) := G(\mathbf{x})^{1/q}$ is a norm on $\mathbb{R}^n$. Let $G_{\mathbb{C}}$ denote the complexification of G, and define*

$$p_{\mathbb{C}}(\mathbf{u} + i\mathbf{v}) := p(\mathbf{u}) + p(\mathbf{v}), \qquad \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

Then there exists a constant $C_q > 0$, depending only on q, such that for every $\mathbf{z} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and every $\boldsymbol{\Delta} \in \mathbb{C}^n$ satisfying $p_{\mathbb{C}}(\boldsymbol{\Delta}) \le p(\mathbf{z})$, one has

$$\frac{|G_{\mathbb{C}}(\mathbf{z} + \boldsymbol{\Delta}) - G(\mathbf{z})|}{G(\mathbf{z})} \le C_q \frac{p_{\mathbb{C}}(\boldsymbol{\Delta})}{p(\mathbf{z})}.$$

Proof. Let $A : (\mathbb{R}^n)^q \to \mathbb{R}$ be the symmetric q-linear form associated with G, so that $G(\mathbf{x}) = A(\mathbf{x}, \dots, \mathbf{x})$. For $\mathbf{x}_1, \dots, \mathbf{x}_q \in \mathbb{R}^n$, the real polarization identity is

$$A(\mathbf{x}_1, \dots, \mathbf{x}_q) = \frac{1}{2^q q!} \sum_{\varepsilon \in \{-1,1\}^q} \left(\prod_{j=1}^q \varepsilon_j \right) G\left(\sum_{j=1}^q \varepsilon_j \mathbf{x}_j \right). \tag{27}$$

Suppose first that $p(\mathbf{x}_j) \le 1$ for every j. The triangle inequality for p gives $p\Big(\sum_{j=1}^q \varepsilon_j \mathbf{x}_j\Big) \le q$, and hence $\Big|G\Big(\sum_{j=1}^q \varepsilon_j \mathbf{x}_j\Big)\Big| \le q^q$. Applying this bound to (27) yields $|A(\mathbf{x}_1, \dots, \mathbf{x}_q)| \le \dfrac{q^q}{q!}$. If some $\mathbf{x}_j = \mathbf{0}$, the desired bound is immediate. Otherwise, apply the preceding estimate to $\mathbf{x}_j / p(\mathbf{x}_j)$ for each j and use multilinearity. Thus, for arbitrary $\mathbf{x}_1, \dots, \mathbf{x}_q$,

$$|A(\mathbf{x}_1, \dots, \mathbf{x}_q)| \le C_q^{(0)} \prod_{j=1}^q p(\mathbf{x}_j), \qquad C_q^{(0)} := \frac{q^q}{q!}. \tag{28}$$

Let $A_{\mathbb{C}}$ be the complex multilinear extension of A. Write $\mathbf{w}_j = \mathbf{u}_j + i\mathbf{v}_j$, and set $\mathbf{x}_{j,0} := \mathbf{u}_j, \quad \mathbf{x}_{j,1} := \mathbf{v}_j$. Multilinearity and (28) give

$$|A_{\mathbb{C}}(\mathbf{w}_1, \dots, \mathbf{w}_q)| \le C_q^{(0)} \sum_{\alpha \in \{0,1\}^q} \prod_{j=1}^q p(\mathbf{x}_{j,\alpha_j}) = C_q^{(0)} \prod_{j=1}^q \big(p(\mathbf{u}_j) + p(\mathbf{v}_j)\big) = C_q^{(0)} \prod_{j=1}^q p_{\mathbb{C}}(\mathbf{w}_j).$$

Therefore,

$$|G_{\mathbb{C}}(\mathbf{z} + \boldsymbol{\Delta}) - G(\mathbf{z})| = \left| \sum_{j=1}^q \binom{q}{j} A_{\mathbb{C}}\big(\mathbf{z}^{\otimes(q-j)}, \boldsymbol{\Delta}^{\otimes j}\big) \right| \le C_q^{(0)} \sum_{j=1}^q \binom{q}{j} p(\mathbf{z})^{q-j} p_{\mathbb{C}}(\boldsymbol{\Delta})^j.$$

Set $t := \dfrac{p_{\mathbb{C}}(\boldsymbol{\Delta})}{p(\mathbf{z})}$. Since $t \le 1$, $(1+t)^q - 1 \le (2^q - 1)t$. Hence

$$|G_{\mathbb{C}}(\mathbf{z}+\boldsymbol{\Delta}) - G(\mathbf{z})| \le C_q p(\mathbf{z})^q \frac{p_{\mathbb{C}}(\boldsymbol{\Delta})}{p(\mathbf{z})}.$$

Finally, $p(\mathbf{z})^q = G(\mathbf{z})$, which proves the claim. □

Lemma D.3 (Multiplicative error of the simulated suffix oracle)**.** *Fix $2 \le \ell \le L-2$, and suppose $\widehat{\mathbf{B}}_\ell$ satisfies the hypotheses of Lemma B.2. Let*

$$\varepsilon_{<\ell} := \max_{1 \le j \le \ell-1} \|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2.$$

For $\mathbf{z} \in \mathbb{R}^{r_\ell}$, set $\mathbf{y} = \widehat{\mathbf{B}}_\ell \mathbf{z}$, construct $\mathbf{z}_0 \in \mathbb{C}^d$ by approximate prefix inversion, and define

$$\widehat{G}_\ell(\mathbf{z}) := \operatorname{Re} f^\star_{\mathbb{C}}(\mathbf{z}_0).$$

There are constants $c_{k,L}, C_{k,L} > 0$ such that if

$$\kappa^{O(\ell)} r^{\ell/2} \varepsilon_{<\ell} \le c_{k,L}, \tag{29}$$

then

$$|\widehat{G}_\ell(\mathbf{z}) - G_\ell(\mathbf{z})| \le C_{k,L} \kappa^{O(\ell)} r^{\ell/2} \varepsilon_{<\ell} G_\ell(\mathbf{z}), \qquad \mathbf{z} \in \mathbb{R}^{r_\ell}. \tag{30}$$

Proof. Let

$$\vartheta_k := \frac{k-1}{2k}, \qquad P_{\ell-1} := \sqrt{\frac{r}{r_{\ell-1}}} \prod_{j=1}^{\ell-1} r_j^{\vartheta_k}, \qquad \mathbf{e}_{\ell-1} := \widetilde{\mathbf{z}}_{\ell-1} - \mathbf{y}.$$

The profile-dependent form of Lemma D.1 gives

$$\|\mathbf{e}_{\ell-1}\|_2 \le C_{k,L} \kappa^{O(\ell)} P_{\ell-1} \varepsilon_{<\ell} \|\mathbf{y}\|_2.$$

Define the induced perturbation in reduced coordinates by

$$\boldsymbol{\Delta}_\ell := (\mathbf{C}_\ell^\sharp)^{-\top} \mathbf{W}_\ell^\top \mathbf{e}_{\ell-1}.$$

Then $(\mathbf{C}_\ell^\sharp)^\top \boldsymbol{\Delta}_\ell = \mathbf{W}_\ell^\top \mathbf{e}_{\ell-1}$. Since $\mathbf{y} = \widehat{\mathbf{B}}_\ell \mathbf{z}$ and $\mathbf{C}_\ell^\sharp = \widehat{\mathbf{B}}_\ell^\top \mathbf{W}_\ell$, $(\mathbf{C}_\ell^\sharp)^\top \mathbf{z} = \mathbf{W}_\ell^\top \mathbf{y}$. Consequently,

$$(\mathbf{C}_\ell^\sharp)^\top (\mathbf{z} + \boldsymbol{\Delta}_\ell) = \mathbf{W}_\ell^\top (\mathbf{y} + \mathbf{e}_{\ell-1}).$$

The true forward pass from the approximately inverted input reaches $\mathbf{y} + \mathbf{e}_{\ell-1}$ immediately before layer ℓ, and hence

$$f^\star_{\mathbb{C}}(\mathbf{z}_0) = G_{\ell,\mathbb{C}}(\mathbf{z} + \boldsymbol{\Delta}_\ell).$$

To bound the reduced perturbation without introducing an artificial inverse aspect ratio, write

$$\mathbf{W}_\ell = \mathbf{B}_\ell \mathbf{C}_\ell, \qquad \mathbf{Q}_\ell := \widehat{\mathbf{B}}_\ell^\top \mathbf{B}_\ell.$$

Then $\mathbf{C}_\ell^\sharp = \mathbf{Q}_\ell \mathbf{C}_\ell$, and the exact identity

$$(\mathbf{C}_\ell^\sharp)^{-\top} \mathbf{W}_\ell^\top = \mathbf{Q}_\ell^{-\top} \mathbf{B}_\ell^\top$$

holds. Lemma B.1 gives $\sigma_{\min}(\mathbf{Q}_\ell) \ge 1 - \varepsilon_{B,\ell} \ge 1/2$, so

$$\|\boldsymbol{\Delta}_\ell\|_2 \le 2 \|\mathbf{e}_{\ell-1}\|_2 \le C_{k,L} \kappa^{O(\ell)} P_{\ell-1} \varepsilon_{<\ell} \|\mathbf{z}\|_2.$$

Here we used the orthonormality of $\widehat{\mathbf{B}}_\ell$, which gives $\|\mathbf{y}\|_2 = \|\mathbf{z}\|_2$.

If $\mathbf{z} = \mathbf{0}$, choose the zero k-th root at every step of the approximate prefix inversion. Then both sides vanish exactly. Henceforth assume $\mathbf{z} \ne \mathbf{0}$. Because G_ℓ is positive definite, convex, even, and q_ℓ-homogeneous,

$$p_\ell(\mathbf{z}) := G_\ell(\mathbf{z})^{1/q_\ell}$$

is a norm with unit ball $K_\ell^\sharp$. The radius inclusions imply

$$\frac{\|\mathbf{z}\|_2}{R_{\mathrm{out},\ell}} \le p_\ell(\mathbf{z}) \le \frac{\|\mathbf{z}\|_2}{r_{\mathrm{in},\ell}}.$$

For $\mathbf{w} = \mathbf{u} + i\mathbf{v}$, define $p_{\ell,\mathbb{C}}(\mathbf{w}) := p_\ell(\mathbf{u}) + p_\ell(\mathbf{v})$. Then $\dfrac{p_{\ell,\mathbb{C}}(\boldsymbol{\Delta}_\ell)}{p_\ell(\mathbf{z})} \le C_{k,L}\omega_\ell \kappa^{O(\ell)} P_{\ell-1}\varepsilon_{<\ell}$. Lemma B.2 and the monotonicity $r_\ell \le r_{\ell-1}$ give

$$\omega_\ell P_{\ell-1} \le C_{k,L}\kappa^4\sqrt{r_\ell}\sqrt{\frac{r}{r_{\ell-1}}}\prod_{j=1}^{\ell-1} r_j^{\vartheta_k} \le C_{k,L}\kappa^4\sqrt{r}\prod_{j=1}^{\ell-1} r_j^{\vartheta_k} \le C_{k,L}\kappa^4 r^{1/2+\vartheta_k(\ell-1)} \le C_{k,L}\kappa^4 r^{\ell/2}.$$

Thus

$$\frac{p_{\ell,\mathbb{C}}(\boldsymbol{\Delta}_\ell)}{p_\ell(\mathbf{z})} \le C_{k,L}\kappa^{O(\ell)} r^{\ell/2}\varepsilon_{<\ell}.$$

Condition (29) makes this quantity at most one. Lemma D.2 yields

$$\frac{|G_{\ell,\mathbb{C}}(\mathbf{z}+\boldsymbol{\Delta}_\ell) - G_\ell(\mathbf{z})|}{G_\ell(\mathbf{z})} \le C_{k,L}\kappa^{O(\ell)} r^{\ell/2}\varepsilon_{<\ell}.$$

Finally, $\widehat{G}_\ell(\mathbf{z}) = \mathrm{Re}\, f_{\mathbb{C}}^\star(\mathbf{z}_0)$ and $G_\ell(\mathbf{z})$ is real, so taking the real part cannot increase the absolute error. This proves (30). □

E Recovery of intermediate hidden layers

This section establishes the inductive recovery step for each intermediate hidden layer, combining column-space estimation with convex-body sampling and spectral diagonalization.

E.1 Stable recovery of the intermediate-layer column space

We first show that suitably chosen query points and directional derivatives recover the column space of the next hidden weight matrix despite errors in the previously recovered prefix.

Proposition E.1 (Stable recovery of the intermediate-layer column space)**.** *Fix $2 \le \ell \le L-2$ and $d \ge 2$. Suppose that the standing assumptions hold and that, after the consistent relabeling inherited from the previous stages,*

$$\max_{1\le j<\ell} \|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \le \varepsilon_W.$$

For any $\varepsilon_B, \zeta \in (0, 1/2)$, it is sufficient that

$$\varepsilon_W \le \frac{c_{k,L}\varepsilon_B}{\kappa^{O(L)} r^{7/2}} \tag{31}$$

and

$$m \ge C_{k,L}\kappa^{O(L)}\frac{r^6}{\varepsilon_B^2}\log\frac{2r}{\zeta}. \tag{32}$$

Under these conditions, one can construct an orthonormal matrix $\widehat{\mathbf{B}}_\ell$ such that, with probability at least $1-\zeta$, after orthogonal alignment,

$$\|\widehat{\mathbf{B}}_\ell - \mathbf{B}_\ell\|_2 \le \varepsilon_B.$$

The total number of additional real function-value queries is at most

$$\widetilde{O}_{k,L}\left(r^7\kappa^{O(L)}\varepsilon_B^{-2}\right). \tag{33}$$

Proof. Throughout this proof, C_L and c_L denote generic positive quantities that may change from line to line and satisfy

$$1 \le C_L \le C_{k,L}\kappa^{O(L)}, \qquad c_{k,L}\kappa^{-O(L)} \le c_L \le 1.$$

This convention records every use of κ_0 and μ^{-1} without suppressing their polynomial dependence. If $r_\ell = r_{\ell-1}$, then $\mathbf{W}_\ell$ is square and invertible by Assumption 3.4, so $\mathrm{col}(\mathbf{W}_\ell) = \mathbb{R}^{r_{\ell-1}}$. Hence no subspace estimation is needed: we may simply take $\widehat{\mathbf{B}}_\ell = \mathbf{I}_{r_{\ell-1}}$, and the asserted basis error is zero after orthogonal alignment. Thus, in the remainder of the proof, we may assume $r_\ell < r_{\ell-1}$. Condition on the previously recovered matrices $\widehat{\mathbf{W}}_1, \dots, \widehat{\mathbf{W}}_{\ell-1}$ satisfying the assumed error bound, with the first-layer signs chosen as in the statement. All probabilities below are with respect to the fresh samples used in this stage.

Let

$$\mathcal{I} = [1, 2], \qquad \mathbf{Z}_1, \dots, \mathbf{Z}_m \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathcal{I}^r),$$

and write $\mathbf{T}_i := \mathbf{Z}_i^{\odot 1/k}$. Define

$$\widehat{\mathbf{R}}_1 := \widehat{\mathbf{W}}_1(\widehat{\mathbf{W}}_1^\top \widehat{\mathbf{W}}_1)^{-1}, \qquad \mathbf{x}_i := \widehat{\mathbf{R}}_1 \mathbf{T}_i.$$

By Lemma A.1 and Weyl's inequality, $\widehat{\mathbf{W}}_1$ has full column rank and $\|\widehat{\mathbf{R}}_1\|_2 \le C_L$ whenever ε_W is sufficiently small. Hence $\|\mathbf{x}_i\|_2 \le C_L\sqrt{r}$. Moreover,

$$\widehat{\mathbf{W}}_1^\top \mathbf{x}_i = \mathbf{T}_i, \qquad \sigma(\widehat{\mathbf{W}}_1^\top \mathbf{x}_i) = \mathbf{Z}_i.$$

Thus the query points are obtained by inverting only the first layer.

We first establish uniform regularity of the true prefix on these query points. Set $\mathbf{A} := \mathbf{W}_1^\top \widehat{\mathbf{R}}_1$. Since $\widehat{\mathbf{W}}_1^\top \widehat{\mathbf{R}}_1 = \mathbf{I}_r$, we have

$$\mathbf{A} - \mathbf{I}_r = (\mathbf{W}_1 - \widehat{\mathbf{W}}_1)^\top \widehat{\mathbf{R}}_1,$$

and therefore $\|\mathbf{A} - \mathbf{I}_r\|_2 \le C_L \varepsilon_W$. For $\mathbf{T} \in [1, 2^{1/k}]^r$, $\|(\mathbf{A} - \mathbf{I}_r)\mathbf{T}\|_\infty \le C_L\sqrt{r}\,\varepsilon_W$. Since the assumed upper bound on ε_W implies $\sqrt{r}\,\varepsilon_W \le c_L$, after decreasing c_L if necessary, $\dfrac{1}{2} \le (\mathbf{AT})_s \le 3, 1 \le s \le r$. For $\mathbf{T} \in [1, 2^{1/k}]^r$, define the true forward maps

$$\mathbf{H}_1(\mathbf{T}) := \mathbf{AT}, \qquad \mathbf{\Phi}_2(\mathbf{T}) := \mathbf{H}_1(\mathbf{T})^{\odot k},$$

and recursively, for $2 \le j \le \ell - 1$, $\mathbf{H}_j(\mathbf{T}) := \mathbf{W}_j^\top \mathbf{\Phi}_j(\mathbf{T}), \mathbf{\Phi}_{j+1}(\mathbf{T}) := \mathbf{H}_j(\mathbf{T})^{\odot k}$. Thus $\mathbf{\Phi}_\ell(\mathbf{T}) = \boldsymbol{\phi}_\ell(\widehat{\mathbf{R}}_1\mathbf{T})$. Since every $\mathbf{W}_j$, $j \ge 2$, is nonnegative and satisfies $\mathbf{W}_j^\top \mathbf{1} = \mathbf{1}$, each coordinate of $\mathbf{W}_j^\top \mathbf{v}$ is a convex combination of the coordinates of $\mathbf{v}$. Therefore, by induction, there exist constants $0 < c_L < C_L < \infty$ such that $c_L \le (\mathbf{H}_j(\mathbf{T}))_s \le C_L$ for every $1 \le j \le \ell - 1$, every admissible coordinate s, and every $\mathbf{T} \in [1, 2^{1/k}]^r$. Consequently, if

$$\mathbf{D}_j(\mathbf{T}) := \mathrm{diag}\left(\mathbf{H}_j(\mathbf{T})^{\odot(k-1)}\right), \qquad 1 \le j \le \ell - 1,$$

then $c_L\mathbf{I} \preceq \mathbf{D}_j(\mathbf{T}) \preceq C_L\mathbf{I}$. Furthermore,

$$D\mathbf{\Phi}_\ell(\mathbf{T}) = k^{\ell-1}\mathbf{D}_{\ell-1}(\mathbf{T})\mathbf{W}_{\ell-1}^\top \cdots \mathbf{D}_2(\mathbf{T})\mathbf{W}_2^\top \mathbf{D}_1(\mathbf{T})\mathbf{A}.$$

By Lemma A.1, $\sigma_{\min}(\mathbf{W}_j) \ge \dfrac{1}{\kappa_0}\sqrt{\dfrac{r_j}{r_{j-1}}}$, $\|\mathbf{W}_j\|_2 \le \kappa_0, 2 \le j \le \ell - 1$, and $\mathbf{A}$ has all its singular values in a fixed neighborhood of one. For compatible full-row-rank matrices $\mathbf{P} \in \mathbb{R}^{m\times n}$ and $\mathbf{Q} \in \mathbb{R}^{n\times p}$, $\sigma_m(\mathbf{PQ}) \ge \sigma_m(\mathbf{P})\sigma_n(\mathbf{Q})$, since

$$\mathbf{PQQ}^\top\mathbf{P}^\top \succeq \sigma_n(\mathbf{Q})^2\mathbf{PP}^\top.$$

Applying this inequality successively to the preceding product, and using the telescoping product of aspect ratios, gives

$$\sigma_{r_{\ell-1}}\left(D\mathbf{\Phi}_\ell(\mathbf{T})\right) \ge c_L \prod_{j=2}^{\ell-1} \sqrt{\frac{r_j}{r_{j-1}}} = c_L\sqrt{\frac{r_{\ell-1}}{r}}.$$

Define $\mathbf{U}_\ell(\mathbf{T}) := \mathbf{W}_\ell^\top \mathbf{\Phi}_\ell(\mathbf{T})$. Then $D\mathbf{U}_\ell(\mathbf{T}) = \mathbf{W}_\ell^\top D\mathbf{\Phi}_\ell(\mathbf{T})$, and another application of the same product inequality yields

$$\sigma_{r_\ell}\left(D\mathbf{U}_\ell(\mathbf{T})\right) \ge c_L\sqrt{\frac{r_\ell}{r_{\ell-1}}}\sqrt{\frac{r_{\ell-1}}{r}} = c_L\sqrt{\frac{r_\ell}{r}}.$$

We also need uniform positivity of the hidden states appearing in the suffix network. For $\ell \le j \le L-2$, define recursively

$$\mathbf{U}_{j+1}(\mathbf{T}) := \mathbf{W}_{j+1}^\top \mathbf{U}_j(\mathbf{T})^{\odot k}.$$

The bound on $\mathbf{\Phi}_\ell(\mathbf{T})$, together with the fact that $\mathbf{W}_\ell^\top \mathbf{1} = \mathbf{1}$, first gives $c_L \le (\mathbf{U}_\ell(\mathbf{T}))_s \le C_L$. If the same bounds hold for $\mathbf{U}_j(\mathbf{T})$, then each coordinate of $\mathbf{U}_{j+1}(\mathbf{T})$ is a convex combination of the coordinates of $\mathbf{U}_j(\mathbf{T})^{\odot k}$. Hence, after adjusting the constants depending only on k, L, $c_L \le (\mathbf{U}_j(\mathbf{T}))_s \le C_L, \qquad \ell \le j \le L-1$, uniformly in $\mathbf{T} \in [1, 2^{1/k}]^r$.

We require a uniform lower bound for the Hessian of the suffix network. Recall that $F_\ell(\mathbf{y}) = h^{(\ell)}(\mathbf{W}_\ell^\top \mathbf{y})$. We claim that, for every $\ell \le j \le L-1$, $\nabla h^{(j)}(\mathbf{U}_j(\mathbf{T})) \ge \dfrac{c_L}{r_j}\mathbf{1}$ coordinatewise and $\nabla^2 h^{(j)}(\mathbf{U}_j(\mathbf{T})) \succeq \dfrac{c_L}{r_j}\mathbf{I}$ uniformly in $\mathbf{T}$. For $j = L-1$, $h^{(L-1)}(\mathbf{u}) = \mathbf{a}^\top \mathbf{u}^{\odot k}$, so $\nabla h^{(L-1)}(\mathbf{u}) = k\,\mathbf{a} \odot \mathbf{u}^{\odot(k-1)}$, and $\nabla^2 h^{(L-1)}(\mathbf{u}) = k(k-1)\operatorname{diag}\left(\mathbf{a} \odot \mathbf{u}^{\odot(k-2)}\right)$. The coordinates of $\mathbf{U}_{L-1}(\mathbf{T})$ are uniformly bounded below, and Assumption 3.4 at $\ell = L$ gives $a_i \ge \mu / r_{L-1}$, proving the claim at $j = L-1$.

Suppose it holds at $j+1$. Write

$$\mathbf{u} := \mathbf{U}_j(\mathbf{T}), \qquad \mathbf{v} := \mathbf{U}_{j+1}(\mathbf{T}) = \mathbf{W}_{j+1}^\top \mathbf{u}^{\odot k}, \qquad \mathbf{D}_\mathbf{u} := \operatorname{diag}(\mathbf{u}^{\odot(k-1)}).$$

By the chain rule,

$$\nabla h^{(j)}(\mathbf{u}) = k\mathbf{D}_\mathbf{u}\mathbf{W}_{j+1}\nabla h^{(j+1)}(\mathbf{v}).$$

Since $\mathbf{W}_{j+1}\mathbf{1}_{r_{j+1}} \ge \mu\dfrac{r_{j+1}}{r_j}\mathbf{1}_{r_j}$, the induction hypothesis gives $\mathbf{W}_{j+1}\nabla h^{(j+1)}(\mathbf{v}) \ge \dfrac{c_L}{r_j}\mathbf{1}$. Together with the uniform positive lower bound on $\mathbf{u}$, this proves the gradient claim. A second differentiation gives

$$\nabla^2 h^{(j)}(\mathbf{u}) = k^2\mathbf{D}_\mathbf{u}\mathbf{W}_{j+1}\nabla^2 h^{(j+1)}(\mathbf{v})\mathbf{W}_{j+1}^\top\mathbf{D}_\mathbf{u} + k(k-1)\operatorname{diag}\Big(\mathbf{u}^{\odot(k-2)} \odot \mathbf{W}_{j+1}\nabla h^{(j+1)}(\mathbf{v})\Big).$$

The first term is positive semidefinite, while the second is bounded below by $c_L r_j^{-1}\mathbf{I}$, proving the claim.

Define $\mathbf{q}_\ell(\mathbf{T}) := \nabla h^{(\ell)}(\mathbf{U}_\ell(\mathbf{T})) \in \mathbb{R}^{r_\ell}$. By the chain rule, $D\mathbf{q}_\ell(\mathbf{T}) = \nabla^2 h^{(\ell)}(\mathbf{U}_\ell(\mathbf{T}))D\mathbf{U}_\ell(\mathbf{T})$. The preceding Hessian and Jacobian lower bounds imply

$$\sigma_{r_\ell}(D\mathbf{q}_\ell(\mathbf{T})) \ge \frac{c_L}{r_\ell}\sqrt{\frac{r_\ell}{r}} = \frac{c_L}{\sqrt{r r_\ell}},$$

uniformly over $\mathbf{T} \in [1, 2^{1/k}]^r$. This pointwise differential lower bound yields a population second-moment lower bound. Let $\mathbf{T} = \mathbf{Z}^{\odot 1/k}$, where $\mathbf{Z} \sim \operatorname{Unif}([1,2]^r)$. The coordinates of $\mathbf{T}$ are independent and have density $\rho(t) = kt^{k-1}, 1 \le t \le 2^{1/k}$. In particular, ρ is bounded above and below by positive constants depending only on k. Let $D_{k,L}$ be an upper bound, depending only on k, L, for the degree of each coordinate of $\mathbf{q}_\ell(\mathbf{T})$ as a polynomial in $\mathbf{T}$. For example, one may take $D_{k,L} = k^{L-1}$. Set $I_k := [1, 2^{1/k}], h_k := 2^{1/k} - 1$, and define the explicit constant

$$C_{\mathrm{inv}}(k, L) := \frac{2^{(k-1)/k} D_{k,L}(D_{k,L}+1)^2(D_{k,L}+2)}{(2^{1/k}-1)^2}.$$

With the concrete choice $D_{k,L} = k^{L-1}$, this becomes

$$C_{\mathrm{inv}}(k, L) = \frac{2^{(k-1)/k} k^{L-1}(k^{L-1}+1)^2(k^{L-1}+2)}{(2^{1/k}-1)^2}.$$

We claim that every univariate polynomial p of degree at most $D_{k,L}$ satisfies

$$\int_{I_k} |p'(t)|^2 \rho(t)\,dt \le C_{\mathrm{inv}}(k, L)\int_{I_k}\left|p(t) - \int_{I_k} p(s)\rho(s)\,ds\right|^2 \rho(t)\,dt.$$

To prove this, let

$$q(t) := p(t) - \int_{I_k} p(s)\rho(s)\,ds.$$

Then $q' = p'$ and $\deg q \le D_{k,L}$. Make the affine change of variables

$$t = 1 + \frac{h_k}{2}(x+1), \qquad Q(x) := q(t), \qquad -1 \le x \le 1.$$

Write $D := D_{k,L}$, let P_n denote the n-th Legendre polynomial, and set

$$\varphi_n(x) := \sqrt{\frac{2n+1}{2}}\, P_n(x), \qquad 0 \le n \le D.$$

Then $\{\varphi_0, \dots, \varphi_D\}$ is an orthonormal basis for the polynomials of degree at most D in $L^2([-1,1])$. Write $Q = \sum_{n=0}^{D} a_n \varphi_n$. The Legendre identity

$$P_n'(x) - P_{n-2}'(x) = (2n-1)P_{n-1}(x), \qquad n \ge 2,$$

together with $P_1' = P_0$, gives by iteration

$$P_n' = \sum_{\substack{0 \le j \le n-1 \\ n-j \text{ odd}}} (2j+1) P_j.$$

Using the orthogonality relation $\int_{-1}^{1} P_j(x)^2\, dx = \dfrac{2}{2j+1}$, we obtain

$$\|P_n'\|^2_{L^2(-1,1)} = n(n+1), \qquad \|\varphi_n'\|^2_{L^2(-1,1)} = \frac{(2n+1)n(n+1)}{2}.$$

Hence, by the triangle inequality followed by Cauchy–Schwarz,

$$\begin{aligned} \|Q'\|^2_{L^2(-1,1)} &\le \left(\sum_{n=0}^{D} |a_n| \|\varphi_n'\|_{L^2(-1,1)} \right)^2 \\ &\le \left(\sum_{n=0}^{D} a_n^2 \right) \left(\sum_{n=0}^{D} \|\varphi_n'\|^2_{L^2(-1,1)} \right) \\ &= \frac{D(D+1)^2(D+2)}{4} \|Q\|^2_{L^2(-1,1)}. \end{aligned}$$

Rescaling from $[-1,1]$ back to I_k therefore gives

$$\int_{I_k} |q'(t)|^2\, dt \le \frac{D(D+1)^2(D+2)}{h_k^2} \int_{I_k} |q(t)|^2\, dt.$$

Finally, on I_k, $k \le \rho(t) = kt^{k-1} \le k\, 2^{(k-1)/k}$. Consequently,

$$\begin{aligned} \int_{I_k} |q'(t)|^2 \rho(t)\, dt &\le k\, 2^{(k-1)/k} \int_{I_k} |q'(t)|^2\, dt \\ &\le C_{\mathrm{inv}}(k,L) k \int_{I_k} |q(t)|^2\, dt \\ &\le C_{\mathrm{inv}}(k,L) \int_{I_k} |q(t)|^2 \rho(t)\, dt, \end{aligned}$$

which proves the claimed one-dimensional inverse inequality.

Fix $\mathbf{v} \in \mathbb{R}^{r\ell}$ with $\|\mathbf{v}\|_2 = 1$, and set $p_{\mathbf{v}}(\mathbf{T}) := \mathbf{v}^\top \mathbf{q}_\ell(\mathbf{T})$. Then

$$\nabla p_{\mathbf{v}}(\mathbf{T}) = D\mathbf{q}_\ell(\mathbf{T})^\top \mathbf{v}, \qquad \|\nabla p_{\mathbf{v}}(\mathbf{T})\|_2 \ge \frac{c_L}{\sqrt{r r_\ell}}.$$

For each $1 \le s \le r$, condition on all coordinates of $\mathbf{T}$ except T_s and apply the preceding one-dimensional inequality to the polynomial $T_s \mapsto p_{\mathbf{v}}(\mathbf{T})$. Since the conditional law of T_s is still $\rho(t)\, dt$,

$$\mathbb{E}[(\partial_s p_{\mathbf{v}}(\mathbf{T}))^2] \le C_{\mathrm{inv}}(k,L) \operatorname{Var}(p_{\mathbf{v}}(\mathbf{T})).$$

Summing over s gives

$$\mathbb{E}\|\nabla p_{\mathbf{v}}(\mathbf{T})\|_2^2 \le C_{\mathrm{inv}}(k,L) r \,\mathrm{Var}(p_{\mathbf{v}}(\mathbf{T})).$$

On the other hand,

$$\mathbb{E}\|\nabla p_{\mathbf{v}}(\mathbf{T})\|_2^2 \ge \frac{c_L}{r r_\ell}.$$

Therefore, after absorbing the constant $C_{\mathrm{inv}}(k,L)$ into c_L, $\mathrm{Var}(p_{\mathbf{v}}(\mathbf{T})) \ge \frac{c_L}{r^2 r_\ell}$. Since $\mathbf{v}$ was arbitrary, $\mathrm{Cov}(\mathbf{q}_\ell(\mathbf{T})) \succeq \frac{c_L}{r^2 r_\ell}\mathbf{I}_{r_\ell}$. Consequently, $\mathbb{E}[\mathbf{q}_\ell(\mathbf{T})\mathbf{q}_\ell(\mathbf{T})^\top] \succeq \frac{c_L}{r^2 r_\ell}\mathbf{I}_{r_\ell}$. For

$$\mathbf{g}_\ell(\mathbf{T}) := \nabla F_\ell(\mathbf{\Phi}_\ell(\mathbf{T})) = \mathbf{W}_\ell \mathbf{q}_\ell(\mathbf{T}),$$

define the population moment $\mathbf{M}_\ell := \mathbb{E}[\mathbf{g}_\ell(\mathbf{T})\mathbf{g}_\ell(\mathbf{T})^\top]$. Then $\mathbf{M}_\ell = \mathbf{W}_\ell \mathbb{E}[\mathbf{q}_\ell(\mathbf{T})\mathbf{q}_\ell(\mathbf{T})^\top]\mathbf{W}_\ell^\top$, and therefore $\mathrm{col}(\mathbf{M}_\ell) = \mathrm{col}(\mathbf{W}_\ell)$, while Lemma A.1 gives

$$\lambda_{r_\ell}(\mathbf{M}_\ell) \ge \frac{c_L}{r^2 r_\ell}\frac{r_\ell}{r_{\ell-1}} = \frac{c_L}{r^2 r_{\ell-1}}, \qquad \lambda_{r_\ell+1}(\mathbf{M}_\ell) = 0.$$

For the recovered-gradient error, fix a query point $\mathbf{x} = \widehat{\mathbf{R}}_1\mathbf{T}$, let $\mathbf{J}_\ell(\mathbf{x}) := D\boldsymbol{\phi}_\ell(\mathbf{x})$, and let $\widehat{\mathbf{J}}_\ell(\mathbf{x})$ denote the Jacobian obtained by replacing the prefix matrices by their recovered versions. We prove simultaneously by induction that

$$\max_{1\le j<\ell} \|\widehat{\mathbf{h}}_j(\mathbf{x}) - \mathbf{h}_j(\mathbf{x})\|_2 \le C_L\sqrt{r}\,\varepsilon_W \tag{34}$$

and that every coordinate of each true and estimated hidden state up to layer $\ell-1$ lies in a fixed interval $[-C_L, C_L]$. At the first layer, $\widehat{\mathbf{h}}_1(\mathbf{x}) = \mathbf{T}, \mathbf{h}_1(\mathbf{x}) = \mathbf{A}\mathbf{T}$, so

$$\|\widehat{\mathbf{h}}_1(\mathbf{x}) - \mathbf{h}_1(\mathbf{x})\|_2 \le \varepsilon_W\|\mathbf{x}\|_2 \le C_L\sqrt{r}\,\varepsilon_W.$$

Suppose the claims hold through layer $j-1$. The map $t \mapsto t^k$ is Lipschitz on the fixed interval containing the true and estimated coordinates, and $\|\sigma(\widehat{\mathbf{h}}_{j-1})\|_2 \le C_L\sqrt{r}$. Using $\|\mathbf{W}_j\|_2 \le \kappa_0$ gives

$$\|\widehat{\mathbf{h}}_j(\mathbf{x}) - \mathbf{h}_j(\mathbf{x})\|_2 \le C_L\sqrt{r}\,\varepsilon_W.$$

The prefix condition makes $\sqrt{r}\,\varepsilon_W$ sufficiently small, which also preserves the fixed coordinate interval and closes the induction.

For $1 \le j < \ell$, write

$$\mathbf{D}_j(\mathbf{x}) := \mathrm{diag}\left(\mathbf{h}_j(\mathbf{x})^{\odot(k-1)}\right), \qquad \widehat{\mathbf{D}}_j(\mathbf{x}) := \mathrm{diag}\left(\widehat{\mathbf{h}}_j(\mathbf{x})^{\odot(k-1)}\right).$$

The true and recovered prefix Jacobians are

$$\mathbf{J}_\ell(\mathbf{x}) = k^{\ell-1}\mathbf{D}_{\ell-1}\mathbf{W}_{\ell-1}^\top \cdots \mathbf{D}_1\mathbf{W}_1^\top$$

and

$$\widehat{\mathbf{J}}_\ell(\mathbf{x}) = k^{\ell-1}\widehat{\mathbf{D}}_{\ell-1}\widehat{\mathbf{W}}_{\ell-1}^\top \cdots \widehat{\mathbf{D}}_1\widehat{\mathbf{W}}_1^\top.$$

On the fixed coordinate interval established above, the derivative of $t \mapsto t^{k-1}$ is uniformly bounded. Hence

$$\|\widehat{\mathbf{D}}_j - \mathbf{D}_j\|_2 \le C_L\|\widehat{\mathbf{h}}_j - \mathbf{h}_j\|_\infty \le C_L\sqrt{r}\,\varepsilon_W.$$

For compatible matrix products,

$$\mathbf{A}_m\cdots\mathbf{A}_1 - \mathbf{B}_m\cdots\mathbf{B}_1 = \sum_{s=1}^{m} \mathbf{A}_m\cdots\mathbf{A}_{s+1}(\mathbf{A}_s - \mathbf{B}_s)\mathbf{B}_{s-1}\cdots\mathbf{B}_1.$$

All remaining factors in the two Jacobian products have operator norm at most C_L. Applying the preceding identity gives

$$\|\widehat{\mathbf{J}}_\ell(\mathbf{x}) - \mathbf{J}_\ell(\mathbf{x})\|_2 \le C_L\sqrt{r}\,\varepsilon_W.$$

Since $\mathbf{\Phi}_\ell(\mathbf{T}) = \boldsymbol{\phi}_\ell(\widehat{\mathbf{R}}_1\mathbf{T}), \mathbf{x} = \widehat{\mathbf{R}}_1\mathbf{T}$, the chain rule gives $D\mathbf{\Phi}_\ell(\mathbf{T}) = \mathbf{J}_\ell(\mathbf{x})\widehat{\mathbf{R}}_1$. Using $\|\widehat{\mathbf{R}}_1\|_2 \le C_L$ and the lower bound already obtained for $D\mathbf{\Phi}_\ell(\mathbf{T})$,

$$\sigma_{r_{\ell-1}}(\mathbf{J}_\ell(\mathbf{x})) \ge \frac{\sigma_{r_{\ell-1}}(D\mathbf{\Phi}_\ell(\mathbf{T}))}{\|\widehat{\mathbf{R}}_1\|_2} \ge c_L\sqrt{\frac{r_{\ell-1}}{r}}.$$

The prefix condition makes $\|\widehat{\mathbf{J}}_\ell - \mathbf{J}_\ell\|_2 \le \frac{1}{2}\sigma_{r_{\ell-1}}(\mathbf{J}_\ell)$. Weyl's inequality therefore yields $\sigma_{r_{\ell-1}}(\widehat{\mathbf{J}}_\ell(\mathbf{x})) \ge c_L\sqrt{\frac{r_{\ell-1}}{r}}$, after adjusting the constants c_L, C_L. Recovering the full gradient $\nabla f^\star(\mathbf{x}) \in \mathbb{R}^d$ is unnecessary. For $1 \le j \le r_{\ell-1}$, define $\mathbf{v}_j(\mathbf{x}) := \widehat{\mathbf{J}}_\ell(\mathbf{x})^\top \mathbf{e}_j \in \mathbb{R}^d$. By Subsection 4.1, the directional derivative $b_j(\mathbf{x}) := D_{\mathbf{v}_j(\mathbf{x})} f^\star(\mathbf{x})$ can be recovered exactly using $O_{k,L}(1)$ real function-value queries. Moreover,

$$b_j(\mathbf{x}) = \mathbf{e}_j^\top \widehat{\mathbf{J}}_\ell(\mathbf{x}) \nabla f^\star(\mathbf{x}).$$

Thus, with

$$\mathbf{b}(\mathbf{x}) := (b_1(\mathbf{x}), \dots, b_{r_{\ell-1}}(\mathbf{x}))^\top,$$

we recover exactly

$$\mathbf{b}(\mathbf{x}) = \widehat{\mathbf{J}}_\ell(\mathbf{x}) \nabla f^\star(\mathbf{x})$$

using $O_{k,L}(r_{\ell-1})$ real function-value queries. Define

$$\widehat{\mathbf{g}}_\ell(\mathbf{x}) := (\widehat{\mathbf{J}}_\ell \widehat{\mathbf{J}}_\ell^\top)^{-1} \mathbf{b}(\mathbf{x}).$$

Equivalently,

$$\widehat{\mathbf{g}}_\ell(\mathbf{x}) = (\widehat{\mathbf{J}}_\ell \widehat{\mathbf{J}}_\ell^\top)^{-1} \widehat{\mathbf{J}}_\ell \nabla f^\star(\mathbf{x}).$$

As in the original chain-rule calculation,

$$\widehat{\mathbf{g}}_\ell - \mathbf{g}_\ell = (\widehat{\mathbf{J}}_\ell^\top)^\dagger (\mathbf{J}_\ell - \widehat{\mathbf{J}}_\ell)^\top \mathbf{g}_\ell.$$

Hence

$$\|\widehat{\mathbf{g}}_\ell(\mathbf{x}) - \mathbf{g}_\ell(\mathbf{x})\|_2 \le C_L \frac{r}{\sqrt{r_{\ell-1}}} \varepsilon_W \|\mathbf{g}_\ell(\mathbf{x})\|_2.$$

It remains to bound the gradient envelope without an output-weight ratio. At the final suffix layer, the fixed coordinate bounds and $\|\mathbf{a}\|_1 = 1$ give

$$\|\nabla h^{(L-1)}(\mathbf{U}_{L-1})\|_1 \le C_L.$$

For every nonnegative vector $\mathbf{v}$, column normalization gives

$$\|\mathbf{W}_j \mathbf{v}\|_1 = \mathbf{1}^\top \mathbf{W}_j \mathbf{v} = (\mathbf{W}_j^\top \mathbf{1})^\top \mathbf{v} = \|\mathbf{v}\|_1.$$

The diagonal derivative factors are uniformly bounded, so backward induction yields

$$\|\mathbf{q}_\ell(\mathbf{T})\|_2 \le \|\mathbf{q}_\ell(\mathbf{T})\|_1 \le C_L, \qquad \|\mathbf{g}_\ell(\mathbf{T})\|_2 \le C_L.$$

It follows that, uniformly in i,

$$\left\|\widehat{\mathbf{g}}_\ell(\mathbf{T}_i)\widehat{\mathbf{g}}_\ell(\mathbf{T}_i)^\top - \mathbf{g}_\ell(\mathbf{T}_i)\mathbf{g}_\ell(\mathbf{T}_i)^\top\right\|_2 \le C_L \frac{r}{\sqrt{r_{\ell-1}}} \varepsilon_W.$$

Define $\mathbf{M}_{\ell,m} := \frac{1}{m}\sum_{i=1}^m \mathbf{g}_\ell(\mathbf{T}_i)\mathbf{g}_\ell(\mathbf{T}_i)^\top$ and $\widehat{\mathbf{M}}_{\ell,m} := \frac{1}{m}\sum_{i=1}^m \widehat{\mathbf{g}}_\ell(\mathbf{T}_i)\widehat{\mathbf{g}}_\ell(\mathbf{T}_i)^\top$. Then deterministically,

$$\|\widehat{\mathbf{M}}_{\ell,m} - \mathbf{M}_{\ell,m}\|_2 \le C_L \frac{r}{\sqrt{r_{\ell-1}}} \varepsilon_W.$$

Furthermore,

$$\|\mathbf{g}_\ell(\mathbf{T}_i)\mathbf{g}_\ell(\mathbf{T}_i)^\top\|_2 \le C_L.$$

Matrix Bernstein's inequality therefore implies that, with probability at least $1 - \zeta$,

$$\|\mathbf{M}_{\ell,m} - \mathbf{M}_\ell\|_2 \le C_L \left(\sqrt{\frac{\log(r/\zeta)}{m}} + \frac{\log(r/\zeta)}{m}\right).$$

The assumed sample size makes this at most $\frac{c_L \varepsilon_B}{r^2 r_{\ell-1}}$. Similarly, because $r_{\ell-1} \le r$, the prefix condition gives $C_L \frac{r}{\sqrt{r_{\ell-1}}} \varepsilon_W \le \frac{c_L \varepsilon_B}{r^2 r_{\ell-1}}$. Consequently, $\|\widehat{\mathbf{M}}_{\ell,m} - \mathbf{M}_\ell\|_2 \le \frac{c_L \varepsilon_B}{r^2 r_{\ell-1}}$. Since $\lambda_{r_\ell}(\mathbf{M}_\ell) \ge$

$\frac{c_L}{r^2 r_{\ell-1}}, \lambda_{r_\ell+1}(\mathbf{M}_\ell) = 0$, Davis–Kahan's theorem yields $\left\|\widehat{\mathbf{B}}_\ell \widehat{\mathbf{B}}_\ell^\top - \mathbf{B}_\ell \mathbf{B}_\ell^\top\right\|_2 \le C_L \varepsilon_B$, where $\widehat{\mathbf{B}}_\ell$ consists of the eigenvectors corresponding to the r_ℓ largest eigenvalues of $\widehat{\mathbf{M}}_{\ell,m}$. For two orthonormal bases of subspaces with projector distance at most $\eta < 1$, an orthogonal alignment $\mathbf{R}$ exists with

$$\|\widehat{\mathbf{B}}_\ell \mathbf{R} - \mathbf{B}_\ell\|_2 \le \sqrt{2}\, \|\widehat{\mathbf{B}}_\ell \widehat{\mathbf{B}}_\ell^\top - \mathbf{B}_\ell \mathbf{B}_\ell^\top\|_2.$$

After decreasing the constant in the assumptions and absorbing $\mathbf{R}$ into the choice of recovered basis, this gives $\|\widehat{\mathbf{B}}_\ell - \mathbf{B}_\ell\|_2 \le \varepsilon_B$. Finally, for each sample $\mathbf{x}_i$, the vector $\widehat{\mathbf{J}}_\ell(\mathbf{x}_i)\nabla f^\star(\mathbf{x}_i) \in \mathbb{R}^{r_{\ell-1}}$ is recovered coordinatewise from directional derivatives. By Subsection 4.1, this uses $O_{k,L}(r_{\ell-1}) \le O_{k,L}(r)$ real function-value queries per sample. Together with (32), the total number of additional queries is therefore

$$\widetilde{O}_{k,L}\left(r^7 \kappa^{O(L)} \varepsilon_B^{-2}\right).$$

□

E.2 Stable recovery within the recovered subspace

Given an accurate basis for the intermediate-layer column space, we next use the simulated suffix oracle, convex-body sampling, and two-matrix diagonalization to recover the individual hidden directions and hence the full weight matrix.

Proposition E.2 (Stable recovery by convex-body sampling)**.** *Fix $2 \le \ell \le L-2$, $d \ge 2$, and $\varepsilon, \zeta \in (0, 1/2)$. Suppose that, after consistent relabeling, $\max_{1\le j<\ell} \|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \le \varepsilon_W$, $\|\widehat{\mathbf{B}}_\ell - \mathbf{B}_\ell\|_2 \le \varepsilon_B$, where the second inequality is after orthogonal alignment. Set $\alpha := 1 + \frac{1}{L}$, $\mathcal{L}(\varepsilon, \gamma, \zeta) := 1 + \log\left(e + \frac{\gamma r^{13/2}}{\varepsilon\zeta}\right)$. It is sufficient that*

$$\varepsilon_B \le c_{k,L} \frac{\varepsilon}{\kappa^2 r} \tag{35}$$

and

$$\varepsilon_W \le \frac{c_{k,L}(\varepsilon/\gamma)^\alpha}{\kappa^{O(L)} r^{\,\ell/2+12+10/L} \mathcal{L}(\varepsilon,\gamma,\zeta)^\alpha}. \tag{36}$$

Then

$$\widetilde{O}_{k,L}\left(\frac{\gamma^2 r^{14} \kappa^{O(L)}}{\varepsilon^2}\right) \tag{37}$$

additional real function-value queries suffice to recover $\widehat{\mathbf{W}}_\ell$ satisfying, with probability at least $1-\zeta$, $\|\widehat{\mathbf{W}}_\ell - \mathbf{W}_\ell\|_2 \le \varepsilon$ after a suitable relabeling of the recovered columns.

Proof. By the preceding approximate-prefix and simulated-suffix lemmas, the available suffix oracle satisfies

$$|\widehat{G}_\ell(\mathbf{z}) - G_\ell(\mathbf{z})| \le \xi_\ell G_\ell(\mathbf{z}), \qquad \xi_\ell \le C_{k,L} \kappa^{O(\ell)} r^{\ell/2} \varepsilon_W.$$

Fix an absolute integer $A_0 \ge 20$, large enough to dominate the fixed conditioning exponents in Lemmas B.5 and A.6, and set $\eta := c_{k,L} \frac{\varepsilon}{\gamma \kappa^{A_0 L} r^4}$, $\nu := c_{k,L} \frac{\varepsilon}{\gamma \kappa^{A_0 L+8} r^6}$. Combining the assumed bound on ε_W with the preceding bound on ξ_ℓ gives

$$\xi_\ell \le \frac{c_{k,L} \nu^{1+1/L}}{\kappa^{O(L)} r^{6+4/L} \left(\log \frac{C_{k,L}\sqrt{r}}{\nu}\right)^{1+1/L}}.$$

Since $r_\ell \le r, q_\ell = O_{k,L}(1), \omega_\ell \le C_{k,L}\kappa^4\sqrt{r}$, this is sufficient for Lemma B.4 with $\delta = 1/L$ and target total-variation error ν. It also implies

$$\xi_\ell \le c_{k,L} \kappa^{-8} \eta r^{-5/2}.$$

Thus Lemma B.7 yields, from

$$m = O_{k,L}\left(\kappa^8 r^2 \eta^{-2} \log \frac{2r}{\zeta}\right) = O_{k,L}\left(\frac{\gamma^2 r^{10} \kappa^{2A_0 L+8}}{\varepsilon^2} \log \frac{2r}{\zeta}\right)$$

independent endpoints, moment matrices satisfying

$$\|\widehat{\mathbf{\Sigma}}_{x,\ell} - \mathbf{\Sigma}_{x,\ell}\|_2 \le \eta\|\mathbf{\Sigma}_{x,\ell}\|_2, \qquad \|\widehat{\mathbf{\Sigma}}_{g,\ell} - \mathbf{\Sigma}_{g,\ell}\|_2 \le \eta\|\mathbf{\Sigma}_{g,\ell}\|_2$$

with probability at least $1-\zeta$. By Lemma B.5, $\mathrm{cond}(\mathbf{V}_\ell) \le 9\kappa^2$, $\rho_\ell \le C_{k,L}\kappa^8 r^3$. Assumption 3.5 supplies the inverse relative gap γ. Lemma A.6 therefore implies that the normalized eigenvectors of $\widehat{\mathbf{\Sigma}}_{g,\ell}\widehat{\mathbf{\Sigma}}_{x,\ell}$ can be labeled so that

$$\min_{s_i\in\{\pm1\}} \|\widehat{\mathbf{v}}_i - s_i\mathbf{v}_i\|_2 \le C_{k,L}\gamma\kappa^{12}r^3\eta \le c_{k,L}\frac{\varepsilon}{\kappa^4 r}.$$

For each recovered direction define $\widehat{\mathbf{u}}_i := \widehat{\mathbf{B}}_\ell\widehat{\mathbf{v}}_i, \widehat{\mathbf{w}}_{\ell,i} := \dfrac{|\widehat{\mathbf{u}}_i|}{\mathbf{1}^\top|\widehat{\mathbf{u}}_i|}$, where the absolute value is coordinatewise. Let $\mathbf{w}_i := \mathbf{w}_{\ell,i}, \mathbf{u}_i^\star := \dfrac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2}$. Since $\mathbf{W}_\ell \ge \mathbf{0}$, the vector $\mathbf{u}_i^\star$ is coordinatewise nonnegative. Write $\mathbf{P}_\ell$ and $\widehat{\mathbf{P}}_\ell$ for the projectors onto the true and recovered column spaces. Because $\widehat{\mathbf{B}}_\ell\mathbf{v}_i = \dfrac{\widehat{\mathbf{P}}_\ell\mathbf{w}_i}{\|\widehat{\mathbf{P}}_\ell\mathbf{w}_i\|_2}$, and $\|\widehat{\mathbf{P}}_\ell - \mathbf{P}_\ell\|_2 \le C\varepsilon_B$, normalization of the projected vector gives

$$\min_{s_i\in\{\pm1\}} \|s_i\widehat{\mathbf{u}}_i - \mathbf{u}_i^\star\|_2 \le C_{k,L}\varepsilon_B + c_{k,L}\frac{\varepsilon}{\kappa^4 r}.$$

The coordinatewise inequality $\big||a|-b\big| \le |a-b|$ for $b \ge 0$ yields the same bound after taking absolute values.

The normalization $\mathbf{1}^\top\mathbf{w}_i = 1$ gives $\mathbf{1}^\top\mathbf{u}_i^\star = \dfrac{1}{\|\mathbf{w}_i\|_2} \ge 1$. Hence, for the constants in the hypotheses sufficiently small,

$$\|\widehat{\mathbf{w}}_{\ell,i} - \mathbf{w}_i\|_2 \le C_{k,L}\sqrt{r_{\ell-1}}\left(\varepsilon_B + \frac{\varepsilon}{\kappa^4 r}\right).$$

Consequently,

$$\|\widehat{\mathbf{W}}_\ell - \mathbf{W}_\ell\|_2 \le \|\widehat{\mathbf{W}}_\ell - \mathbf{W}_\ell\|_F \le C_{k,L}\sqrt{r_\ell r_{\ell-1}}\left(\varepsilon_B + \frac{\varepsilon}{\kappa^4 r}\right) \le C_{k,L}r\left(\varepsilon_B + \frac{\varepsilon}{\kappa^4 r}\right) \le \varepsilon,$$

after decreasing $c_{k,L}$ if necessary.

For the query count, each independent endpoint requires, by Lemma B.4, $\widetilde{O}_{k,L}(\kappa^8 r^4)$ evaluations of the simulated suffix oracle. The gradient estimator in Lemma B.6 requires only $O_{k,L}(r)$ further evaluations per endpoint. Each simulated suffix evaluation is implemented using $O_{k,L}(1)$ real evaluations of $f^\star$. Multiplying by the above value of m gives

$$\widetilde{O}_{k,L}\left(\frac{\gamma^2 r^{14}\kappa^{2A_0L+16}}{\varepsilon^2}\log\frac{1}{\zeta}\right)$$

additional real function-value queries. □

F Recovery of the final hidden layer and output weights

This section analyzes the final recovery stage, where a two-Hessian decomposition identifies the last hidden layer and the output weights.

Proposition F.1 (Stable recovery of the final hidden layer and output weights)**.** *Suppose that after the consistent relabeling inherited from the previous stages,*

$$\max_{1\le j\le L-2} \|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \le \varepsilon_W.$$

There exists $c_{k,L} > 0$, independent of the conditioning parameters, such that, for every $\varepsilon, \zeta \in (0, 1/2)$, it is sufficient that

$$\varepsilon_W \le c_{k,L}\frac{\zeta\varepsilon}{r^{L/2+6}\kappa^{O(L)}}. \tag{38}$$

Then, using $O_{k,L}(r^2)$ additional real function-value queries, one can construct $\widehat{\mathbf{W}}_{L-1}$ and $\widehat{\mathbf{a}}$ such that, with probability at least $1-\zeta$, after a suitable relabeling,

$$\|\widehat{\mathbf{W}}_{L-1} - \mathbf{W}_{L-1}\|_2 \le \varepsilon, \qquad \|\widehat{\mathbf{a}} - \mathbf{a}\|_1 \le \varepsilon.$$

Proof. Throughout this proof, C_L and c_L are generic quantities satisfying

$$1 \le C_L \le C_{k,L}\kappa^{O(L)}, \qquad c_{k,L}\kappa^{-O(L)} \le c_L \le 1.$$

Write

$$n := r_{L-2}, \qquad s := r_{L-1}, \qquad \mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_s),$$

and consider the final suffix $F(\mathbf{y}) := F_{L-1}(\mathbf{y}) = \sum_{i=1}^{s} a_i(\mathbf{w}_i^\top \mathbf{y})^k, \mathbf{y} \in \mathbb{R}^n$. Lemma A.2 gives

$$\frac{\mu}{s} \le a_{\min} \le \frac{1}{s} \le a_{\max} \le 1.$$

We first construct accurate Hessians of F at two points using the approximately recovered prefix. In the notation of Lemma D.1, let $\vartheta_k := \dfrac{k-1}{2k}, P_{L-2} = \sqrt{\dfrac{r}{n}} \prod_{j=1}^{L-2} r_j^{\vartheta_k}$. For $\mathbf{y}$ in a fixed bounded positive box, the simulated suffix oracle has the form $\widehat{F}(\mathbf{y}) = \operatorname{Re} F_{\mathbb{C}}(\mathbf{y} + \mathbf{e}(\mathbf{y}))$, where

$$\|\mathbf{e}(\mathbf{y})\|_2 \le C_L P_{L-2}\varepsilon_W \|\mathbf{y}\|_2.$$

On the box considered below, $\|\mathbf{y}\|_2 \le C\sqrt{n}$, and therefore

$$P_{L-2}\|\mathbf{y}\|_2 \le C\sqrt{r} \prod_{j=1}^{L-2} r_j^{\vartheta_k} \le C r^{1/2+\vartheta_k(L-2)} \le C r^{(L-1)/2}.$$

Thus

$$\|\mathbf{e}(\mathbf{y})\|_2 \le C_L r^{(L-1)/2}\varepsilon_W.$$

Since $\mathbf{W} \ge \mathbf{0}$, each column has unit ℓ_1-norm, and all points lie in a fixed positive box, every $\mathbf{w}_i^\top \mathbf{y}$ remains in a fixed positive interval. The identity

$$(z+\Delta)^k - z^k = \Delta \sum_{j=0}^{k-1} (z+\Delta)^{k-1-j} z^j$$

and $\sum_i a_i = 1$ give $|\widehat{F}(\mathbf{y}) - F(\mathbf{y})| \le C_L\|\mathbf{e}(\mathbf{y})\|_2 \le C_L r^{(L-1)/2}\varepsilon_W$. All complex evaluations involved in the simulated suffix oracle are implemented exactly by real function-value queries.

Set $\mathbf{y}^{(2)} := \mathbf{1}_n$. Let $\mathbf{u} \sim \mathrm{Unif}(S^{n-1})$, choose a sufficiently small constant $\tau = \tau_{k,L} > 0$, and define

$$\mathbf{y}^{(1)} := \mathbf{1}_n + \tau\mathbf{u}.$$

Choose fixed interpolation nodes sufficiently close to zero so that all points used to evaluate directional second derivatives around $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ remain in the same positive box. For $\mathbf{y} \in \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}\}$ and $\mathbf{v} \in \mathbb{R}^n$, the function $t \mapsto F(\mathbf{y} + t\mathbf{v})$ has degree at most k, so its second derivative at zero is a fixed linear combination of $k+1$ function values. Replacing these values by simulated suffix values gives an estimator of $\mathbf{v}^\top \nabla^2 F(\mathbf{y})\mathbf{v}$ with error at most $C_L r^{(L-1)/2}\varepsilon_W$.

Applying this to

$$\mathbf{v} = \mathbf{e}_p, \qquad \mathbf{v} = \mathbf{e}_p + \mathbf{e}_q, \qquad 1 \le p < q \le n,$$

and using polarization, we construct symmetric matrices $\widehat{\mathbf{H}}_j, \mathbf{H}_j := \nabla^2 F(\mathbf{y}^{(j)}), j = 1, 2$, satisfying

$$\|\widehat{\mathbf{H}}_j - \mathbf{H}_j\|_2 \le \delta_H, \qquad \delta_H := C_L r^{(L+1)/2}\varepsilon_W.$$

The construction uses $O_k(n^2) \le O_k(r^2)$ simulated suffix evaluations, and hence $O_{k,L}(r^2)$ real function-value queries.

At $\mathbf{y}^{(2)} = \mathbf{1}_n$, column normalization gives $\mathbf{w}_i^\top \mathbf{y}^{(2)} = 1$, and therefore

$$\mathbf{H}_2 = k(k-1)\mathbf{W}\,\mathrm{diag}(\mathbf{a})\mathbf{W}^\top.$$

If $s = n$, then $\mathbf{W}$ is square and invertible, so no subspace estimation is needed and we set $\widehat{\mathbf{B}} := \mathbf{I}_n$. If $s < n$, Lemmas A.1 and A.2 give

$$\lambda_s(\mathbf{H}_2) \ge k(k-1)a_{\min}\sigma_{\min}(\mathbf{W})^2 \ge c_L \frac{1}{s}\frac{s}{n} = \frac{c_L}{n}, \qquad \lambda_{s+1}(\mathbf{H}_2) = 0.$$

Let $\widehat{\mathbf{B}} \in \mathbb{R}^{n\times s}$ contain the eigenvectors corresponding to the s largest eigenvalues of $\widehat{\mathbf{H}}_2$. Davis–Kahan's theorem yields

$$\|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{W}}\|_2 \le C_L n\delta_H \le C_L r\delta_H.$$

Thus the same projector bound holds in either case. Choose an orthonormal basis

$$\mathbf{B} \in \underset{\substack{\widetilde{\mathbf{B}}^\top\widetilde{\mathbf{B}}=\mathbf{I}_s \\ \mathrm{col}(\widetilde{\mathbf{B}})=\mathrm{col}(\mathbf{W})}}{\arg\min} \|\widehat{\mathbf{B}} - \widetilde{\mathbf{B}}\|_2.$$

By the standard relation between principal angles and optimally chosen orthonormal bases,

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_2 \le \sqrt{2}\,\|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{W}}\|_2 \le C_L n\delta_H.$$

In the square case $s = n$, this convention simply corresponds to taking $\mathbf{B} = \widehat{\mathbf{B}} = \mathbf{I}_n$. Set $\varepsilon_{B,\mathrm{fin}} := \|\widehat{\mathbf{B}} - \mathbf{B}\|_2$. Since $\delta_H = C_L r^{(L+1)/2}\varepsilon_W$, $n \le r$, we have $\varepsilon_{B,\mathrm{fin}} \le C_L r^{(L+3)/2}\varepsilon_W$. Under (38), after decreasing $c_{k,L}$, and enlarging if necessary the exponent hidden in $\kappa^{O(L)}$, it follows that $\varepsilon_{B,\mathrm{fin}} \le \frac{1}{2}$. Define $\mathbf{C} := \widehat{\mathbf{B}}^\top\mathbf{W} \in \mathbb{R}^{s\times s}$. Lemma B.1, applied to the above choice of $\mathbf{B}$ and $\widehat{\mathbf{B}}$, gives

$$\sigma_{\min}(\mathbf{C}) \ge (1 - \varepsilon_{B,\mathrm{fin}})\sigma_{\min}(\mathbf{W}) \ge \frac{1}{2\kappa_0}\sqrt{\frac{s}{n}},$$

and

$$\sigma_{\max}(\mathbf{C}) \le (1 + \varepsilon_{B,\mathrm{fin}})\sigma_{\max}(\mathbf{W}) \le \frac{3}{2}\kappa_0.$$

Moreover,

$$\mathrm{cond}(\mathbf{C}) \le \frac{1 + \varepsilon_{B,\mathrm{fin}}}{1 - \varepsilon_{B,\mathrm{fin}}}\,\mathrm{cond}(\mathbf{W}) \le 3\kappa_0.$$

Thus, in the generic c_L, C_L notation used in this proof,

$$c_L\sqrt{\frac{s}{n}} \le \sigma_{\min}(\mathbf{C}) \le \sigma_{\max}(\mathbf{C}) \le C_L, \qquad \mathrm{cond}(\mathbf{C}) \le C_L.$$

Define $\mathbf{M}_j := \widehat{\mathbf{B}}^\top\mathbf{H}_j\widehat{\mathbf{B}}, \widehat{\mathbf{M}}_j := \widehat{\mathbf{B}}^\top\widehat{\mathbf{H}}_j\widehat{\mathbf{B}}$. Then $\|\widehat{\mathbf{M}}_j - \mathbf{M}_j\|_2 \le \delta_H$, and

$$\mathbf{M}_2 = k(k-1)\mathbf{C}\,\mathrm{diag}(\mathbf{a})\mathbf{C}^\top,$$

while

$$\mathbf{M}_1 = k(k-1)\mathbf{C}\,\mathrm{diag}\big(a_i(1 + \tau\mathbf{w}_i^\top\mathbf{u})^{k-2}\big)\,\mathbf{C}^\top.$$

Consequently,

$$\mathbf{A} := \mathbf{M}_1\mathbf{M}_2^{-1} = \mathbf{C}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_s)\mathbf{C}^{-1},$$

where $\lambda_i := (1 + \tau\mathbf{w}_i^\top\mathbf{u})^{k-2}$. For the spectral separation, Lemma A.1 gives $\|\mathbf{w}_i - \mathbf{w}_j\|_2 \ge c_L\sqrt{\frac{s}{n}}, i \ne j$. The spherical small-ball bound implies, for every $i \ne j$,

$$\mathbb{P}\left(|(\mathbf{w}_i - \mathbf{w}_j)^\top\mathbf{u}| \le t\frac{\|\mathbf{w}_i - \mathbf{w}_j\|_2}{\sqrt{n}}\right) \le Ct.$$

Choosing $t = c\zeta/s^2$ and applying a union bound over all pairs gives, with probability at least $1 - \zeta$,

$$\min_{i\ne j}|(\mathbf{w}_i - \mathbf{w}_j)^\top\mathbf{u}| \ge c_L\frac{\zeta}{s^{3/2}n} \ge c_L\frac{\zeta}{r^{5/2}}.$$

By the choice of τ, $1/2 \le 1 + \tau\mathbf{w}_i^\top\mathbf{u} \le 3/2$. The mean-value theorem therefore gives, on the same event,

$$\Delta := \min_{i\ne j}|\lambda_i - \lambda_j| \ge c_L\frac{\zeta}{r^{5/2}}.$$

Moreover,

$$\lambda_{\min}(\mathbf{M}_2) \ge \frac{c_L}{n}, \qquad \|\mathbf{A}\|_2 \le C_L.$$

The prefix condition makes $\delta_H \le c_L/n$, so $\widehat{\mathbf{M}}_2$ is invertible and $\|\widehat{\mathbf{M}}_2^{-1}\|_2 \le C_L n$. Let $\mathbf{E}_j := \widehat{\mathbf{M}}_j - \mathbf{M}_j$. Using $\mathbf{A}\mathbf{M}_2 = \mathbf{M}_1$, the perturbation admits the relative factorization

$$\mathbf{E}_{\mathbf{A}} := \widehat{\mathbf{M}}_1\widehat{\mathbf{M}}_2^{-1} - \mathbf{M}_1\mathbf{M}_2^{-1} = (\mathbf{E}_1 - \mathbf{A}\mathbf{E}_2)\widehat{\mathbf{M}}_2^{-1}.$$

Hence

$$\|\mathbf{E}_{\mathbf{A}}\|_2 \le C_L n\delta_H \le C_L r\delta_H.$$

Let $\mathbf{V}_{\mathbf{C}}$ denote the matrix whose columns are the normalized columns of $\mathbf{C}$. The conditioning bounds give $\mathrm{cond}(\mathbf{V}_{\mathbf{C}}) \le C_L$. Applying the simple-spectrum eigenvector argument used in Lemma A.6 shows that the eigenvectors of $\widehat{\mathbf{M}}_1\widehat{\mathbf{M}}_2^{-1}$ may be labeled so that, after choosing signs,

$$\left\|\widehat{\mathbf{v}}_i - \frac{\mathbf{c}_i}{\|\mathbf{c}_i\|_2}\right\|_2 \le C_L\frac{r\delta_H}{\Delta} \le C_L\frac{r^{L/2+4}}{\zeta}\varepsilon_W.$$

Set $\widehat{\mathbf{u}}_i := \widehat{\mathbf{B}}\widehat{\mathbf{v}}_i$, $\mathbf{u}_i^\star := \dfrac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2}$. Combining the eigenvector and subspace estimates yields

$$\min_{\epsilon_i\in\{\pm1\}} \|\epsilon_i\widehat{\mathbf{u}}_i - \mathbf{u}_i^\star\|_2 \le C_L\frac{r^{L/2+4}}{\zeta}\varepsilon_W.$$

Since $\mathbf{w}_i > \mathbf{0}$, taking coordinatewise absolute values preserves this bound. Define $\widehat{\mathbf{w}}_i := \dfrac{|\widehat{\mathbf{u}}_i|}{\mathbf{1}^\top|\widehat{\mathbf{u}}_i|}$. Because $\mathbf{1}^\top\mathbf{u}_i^\star = 1/\|\mathbf{w}_i\|_2 \ge 1$, $\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \le C_L\dfrac{r^{L/2+9/2}}{\zeta}\varepsilon_W$. Consequently, $E_W := \|\widehat{\mathbf{W}} - \mathbf{W}\|_2 \le \|\widehat{\mathbf{W}} - \mathbf{W}\|_F \le C_L\dfrac{r^{L/2+5}}{\zeta}\varepsilon_W \le c_{k,L}\dfrac{\varepsilon}{r} \le \varepsilon$. The stronger ε/r bound is used below. Enlarging the exponent hidden in $\kappa^{O(L)}$, it also ensures $E_W \le \dfrac{1}{2}\sigma_{\min}(\mathbf{W})$, because $\sigma_{\min}(\mathbf{W}) \ge \kappa_0^{-1}\sqrt{s/n} \ge (\kappa_0\sqrt{r})^{-1}$. Hence $\widehat{\mathbf{W}}$ has full column rank. To recover $\mathbf{a}$, use

$$\mathbf{H}_2 = k(k-1)\mathbf{W}\,\mathrm{diag}(\mathbf{a})\mathbf{W}^\top.$$

Define $\widehat{\mathbf{a}} := \dfrac{1}{k(k-1)}\,\mathrm{diag}\Big(\widehat{\mathbf{W}}^\dagger\widehat{\mathbf{H}}_2\widehat{\mathbf{W}}^{\dagger\top}\Big)$. Let $\mathbf{T}_W := \widehat{\mathbf{W}}^\dagger\mathbf{W}$. Since $\widehat{\mathbf{W}}^\dagger\widehat{\mathbf{W}} = \mathbf{I}_s$,

$$\mathbf{T}_W - \mathbf{I}_s = -\widehat{\mathbf{W}}^\dagger(\widehat{\mathbf{W}} - \mathbf{W}),$$

and therefore

$$\|\mathbf{T}_W - \mathbf{I}_s\|_2 \le C_L\sqrt{\frac{n}{s}}\,E_W.$$

Using $a_{\max} \le 1$,

$$\|\mathbf{T}_W\,\mathrm{diag}(\mathbf{a})\mathbf{T}_W^\top - \mathrm{diag}(\mathbf{a})\|_2 \le C_L\sqrt{\frac{n}{s}}\,E_W.$$

The Hessian perturbation contributes

$$\left\|\widehat{\mathbf{W}}^\dagger(\widehat{\mathbf{H}}_2 - \mathbf{H}_2)\widehat{\mathbf{W}}^{\dagger\top}\right\|_2 \le C_L\frac{n}{s}\delta_H.$$

Taking the ℓ_1-norm of the diagonal and using $\sqrt{ns} \le r$ and $n \le r$, we obtain

$$\|\widehat{\mathbf{a}} - \mathbf{a}\|_1 \le C_L r(E_W + \delta_H) \le \varepsilon,$$

where the final inequality follows from the prefix condition and the bounds for E_W and δ_H. This completes the proof. □

G PROOF OF THE MAIN THEOREM

The preceding propositions isolate four sources of complexity. The input dimension d in the query complexity appears only in the initial recovery of the first-layer feature subspace, which costs $O_{k,L}(dr)$ queries. Once this subspace has been identified, all subsequent gradient information is accessed through hidden-dimensional directional derivatives, and all sampling, moment estimation, and prefix-inversion costs are governed by the maximal hidden width r. The uniform inverse relative spectral gap contributes powers of γ, while composing $O(L)$ conditioned stages produces the factor $\kappa^{O(L^2)}$.

Proof of Theorem 5.1. Set $\alpha := 1+\frac{1}{L}, p_\star := \alpha^{L-3}$. Since $p_\star = (1+1/L)^{L-3} < (1+1/L)^L < e$, it is enough to prove the sharper query bound with accuracy exponent $2p_\star$. Let $N := 2L-4, \zeta_0 := \frac{0.01}{N}$. The number N upper-bounds the randomized recovery stages: there is one first-layer stage, two stages for every $2 \le \ell \le L-2$, and one final stage. We will show

$$Q_{\mathrm{tot}} \le \widetilde{O}_{k,L}\Big(dr + r^{L^2+O(L)}\gamma^{O(L)}\kappa^{O(L^2)}\delta^{-2p_\star}\Big). \tag{39}$$

Accuracy schedule. For $2 \le \ell \le L-2$, define $\beta_\ell := \frac{\ell}{2} + 12 + \frac{10}{L}$. Choose an absolute constant $A \ge 1$ sufficiently large that κ^{AL} dominates every factor denoted by $\kappa^{O(L)}$ in Propositions E.1, E.2, and F.1. Define the terminal intermediate-layer accuracy by

$$\tau_{L-2} := c_{k,L}\frac{\zeta_0\delta}{r^{L/2+6}\kappa^{AL}}.$$

For $\ell = L-2, L-3, \dots, 2$, define recursively

$$\tau_{\ell-1} := \frac{c_{k,L}\tau_\ell^\alpha}{\gamma^\alpha\kappa^{AL}r^{\beta_\ell}\left[1+\log\left(e+\frac{\gamma r^{13/2}}{\zeta_0\tau_\ell}\right)\right]^\alpha}. \tag{40}$$

After decreasing $c_{k,L}$, if necessary, $0 < \tau_1 \le \tau_2 \le \cdots \le \tau_{L-2} < \frac{1}{2}$. Indeed, $\tau_\ell^\alpha \le \tau_\ell$, $\gamma \ge 1$, $\beta_\ell > 9/2$, and the logarithmic factor is at least one. The terminal value is also at most $1/2$ after the same choice of constant.

First-layer recovery. Apply Proposition C.2 with target accuracy τ_1 and failure probability ζ_0. Its query count is

$$\widetilde{O}_{k,L}\Big(dr + \gamma^2 r^{13}\kappa^{O(1)}\tau_1^{-2}\Big).$$

With probability at least $1-\zeta_0$, the result satisfies, after a permutation of the columns and a choice of their inherent signs, $\|\widehat{\mathbf{W}}_1 - \mathbf{W}_1\|_2 \le \tau_1$.

Intermediate-layer recovery. Fix $2 \le \ell \le L-2$ and suppose inductively that, after propagating the relabelings selected in the preceding stages, $\|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \le \tau_j, 1 \le j < \ell$. Since the sequence (τ_j) is nondecreasing,

$$\max_{1\le j<\ell}\|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \le \tau_{\ell-1}.$$

Set $\varepsilon_{B,\ell} := c_{k,L}\frac{\tau_\ell}{\kappa^2 r}$. The recurrence, together with $\tau_\ell^\alpha \le \tau_\ell$ and $\beta_\ell > 9/2$, gives

$$\tau_{\ell-1} \le c_{k,L}\frac{\tau_\ell}{\kappa^{O(L)}r^{9/2}} = c_{k,L}\frac{\varepsilon_{B,\ell}}{\kappa^{O(L)}r^{7/2}}.$$

Thus Proposition E.1 applies with target basis accuracy $\varepsilon_{B,\ell}$. With probability at least $1-\zeta_0$, it returns an orthonormal matrix $\widehat{\mathbf{B}}_\ell$ such that, after orthogonal alignment, $\|\widehat{\mathbf{B}}_\ell - \mathbf{B}_\ell\|_2 \le \varepsilon_{B,\ell}$. The corresponding query count is

$$\widetilde{O}_{k,L}\Big(r^7\kappa^{O(L)}\varepsilon_{B,\ell}^{-2}\Big) = \widetilde{O}_{k,L}\Big(r^9\kappa^{O(L)}\tau_\ell^{-2}\Big).$$

The basis condition in Proposition E.2 is satisfied by construction. Moreover, recurrence (40) is exactly the prefix condition in that proposition with target accuracy τ_ℓ, failure probability ζ_0, and the uniform inverse-gap parameter γ. Hence the convex-body stage applies, with query count

$$\widetilde{O}_{k,L}\left(\gamma^2 r^{14}\kappa^{O(L)}\tau_\ell^{-2}\right).$$

With probability at least $1-\zeta_0$, it returns, after a suitable column relabeling, $\|\widehat{\mathbf{W}}_\ell - \mathbf{W}_\ell\|_2 \le \tau_\ell$. This closes the induction for every $2 \le \ell \le L-2$.

Last hidden layer and output layer recovery. After recovery through layer $L-2$,

$$\max_{1\le j\le L-2} \|\widehat{\mathbf{W}}_j - \mathbf{W}_j\|_2 \le \tau_{L-2} = c_{k,L}\frac{\zeta_0\delta}{r^{L/2+6}\kappa^{AL}}.$$

By the choice of A, this is the sufficient condition in Proposition F.1. That proposition uses $O_{k,L}(r^2)$ additional real function-value queries and returns, with probability at least $1-\zeta_0$,

$$\|\widehat{\mathbf{W}}_{L-1} - \mathbf{W}_{L-1}\|_2 \le \delta, \qquad \|\widehat{\mathbf{a}} - \mathbf{a}\|_1 \le \delta.$$

Because $\tau_\ell \le \tau_{L-2} \le \delta$ for $1 \le \ell \le L-2$, the intersection of the successful events satisfies all parameter-recovery guarantees in the theorem. The permutations selected at successive stages are propagated through adjacent weight matrices, producing one consistent hidden-unit relabeling. The first-layer column signs may be selected arbitrarily because k is even.

Solving the accuracy recurrence. Up to factors suppressed by $\widetilde{O}_{k,L}$, write

$$\tau_\ell^{-1} \le \widetilde{O}_{k,L}\left(r^{s_\ell}\gamma^{g_\ell}\kappa^{t_\ell}\delta^{-p_\ell}\right). \tag{41}$$

The terminal values are $p_{L-2}=1, s_{L-2}=\dfrac{L}{2}+6, g_{L-2}=0, t_{L-2}=AL$. The recurrence gives $p_{\ell-1}=\alpha p_\ell, s_{\ell-1}=\alpha s_\ell+\beta_\ell, g_{\ell-1}=\alpha g_\ell+\alpha, t_{\ell-1}=\alpha t_\ell+AL$. Therefore $p_1=\alpha^{L-3}=p_\star$. Unrolling the width recurrence yields

$$s_1 = p_\star\left(\frac{L}{2}+6\right) + \sum_{\ell=2}^{L-2}\alpha^{\ell-2}\beta_\ell. \tag{42}$$

Set $c_L := 12 + 10/L$. Then $\beta_\ell = \ell/2 + c_L$. The geometric identities $\sum_{\ell=2}^{L-2}\alpha^{\ell-2} = L(p_\star-1)$, and $\sum_{\ell=2}^{L-2}\ell\alpha^{\ell-2} = L^2 - L - 2Lp_\star$ give

$$s_1 = p_\star\left(\frac{L}{2}+6\right) + \frac{1}{2}\left(L^2 - L - 2Lp_\star\right) + c_L L(p_\star-1) = \frac{L^2}{2} + O(L).$$

Similarly,

$$g_1 = \sum_{j=1}^{L-3}\alpha^j = (L+1)(p_\star-1) = O(L),$$

and $t_1 = ALp_\star + AL^2(p_\star-1) = O(L^2)$. The logarithmic factors generated by the recurrence are polynomial in $\log r$, $\log\gamma$, $\log(1/\delta)$, and $\log\kappa$. The first three are suppressed by $\widetilde{O}_{k,L}$, while $\log\kappa$ is absorbed into the displayed factor $\kappa^{O(L^2)}$, since $\kappa \ge 2$. Substituting these bounds into (41) gives

$$\tau_1^{-1} \le \widetilde{O}_{k,L}\left(r^{L^2/2+O(L)}\gamma^{g_1}\kappa^{O(L^2)}\delta^{-p_\star}\right), \tag{43}$$

and consequently

$$\tau_1^{-2} \le \widetilde{O}_{k,L}\left(r^{L^2+O(L)}\gamma^{2g_1}\kappa^{O(L^2)}\delta^{-2p_\star}\right).$$

Total complexity. For every $1 \le \ell \le L-2$, $\tau_1 \le \tau_\ell$. The preceding stage bounds are

$$\widetilde{O}_{k,L}\Big(dr + \gamma^2 r^{13} \kappa^{O(1)} \tau_1^{-2}\Big)$$

for the first layer,

$$\widetilde{O}_{k,L}\Big(r^9 \kappa^{O(L)} \tau_1^{-2} + \gamma^2 r^{14} \kappa^{O(L)} \tau_1^{-2}\Big)$$

for each intermediate layer, and $O_{k,L}(r^2)$ for the final stage. Since $r \ge 1$, $\gamma \ge 1$, $\tau_1 < 1$, and there are only $O(L)$ stages, all hidden-dimensional contributions are dominated, up to $\widetilde{O}_{k,L}$-factors, by the convex-body term. Hence

$$Q_{\mathrm{tot}} \le \widetilde{O}_{k,L}\Big(dr + \gamma^2 r^{14} \kappa^{O(L)} \tau_1^{-2}\Big).$$

Using (43), absorbing the fixed width overhead into $r^{O(L)}$, and observing that $2g_1 + 2 = O(L)$, we obtain

$$Q_{\mathrm{tot}} = \widetilde{O}_{k,L}\Big(dr + r^{L^2+O(L)} \gamma^{O(L)} \kappa^{O(L^2)} \delta^{-2p_\star}\Big).$$

This is (39). Since $0 < \delta < 1$ and $p_\star < \mathrm{e}$, $\delta^{-2p_\star} \le \delta^{-2\mathrm{e}}$, proving the query bound in Theorem 5.1.

Success probability. Each stage uses fresh randomness. Conditioning successively on all preceding successful stages, the relevant recovery proposition fails with probability at most ζ_0. Hence the union bound gives

$$\mathbb{P}(\text{some recovery stage fails}) \le N\zeta_0 = 0.01.$$

Thus all parameter-recovery guarantees hold simultaneously with probability at least 0.99. □

Lemma G.1 (Gaussian stability under parameter perturbations)**.** *Let $f^\star$ satisfy the normalization and positivity assumptions of Theorem 5.1, and let $\widehat{f}$ have the same architecture. Suppose that, after a consistent relabeling and a suitable choice of the first-layer column signs,*

$$\max_{1 \le \ell \le L-1} \|\widehat{\mathbf{W}}_\ell - \mathbf{W}_\ell\|_2 \le \delta, \qquad \|\widehat{\mathbf{a}} - \mathbf{a}\|_1 \le \delta,$$

where $0 < \delta < 1/2$. Assume that the recovered hidden weights are normalized as in the recovery algorithm: the columns of $\widehat{\mathbf{W}}_1$ have unit Euclidean norm, while, for $2 \le \ell \le L-1$, the columns of $\widehat{\mathbf{W}}_\ell$ are nonnegative and have unit ℓ_1-norm. Then

$$\mathbb{E}_{\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \Big[\big(\widehat{f}(\mathbf{X}) - f^\star(\mathbf{X})\big)^2\Big] \le C_{k,L} r \delta^2.$$

Proof. Write $\mathbf{H}_1 = \mathbf{W}_1^\top \mathbf{X}, \widehat{\mathbf{H}}_1 = \widehat{\mathbf{W}}_1^\top \mathbf{X}$, and, for $2 \le \ell \le L-1$, $\mathbf{H}_\ell = \mathbf{W}_\ell^\top \mathbf{H}_{\ell-1}^{\odot k}, \widehat{\mathbf{H}}_\ell = \widehat{\mathbf{W}}_\ell^\top \widehat{\mathbf{H}}_{\ell-1}^{\odot k}$. We first record uniform moment bounds. For $p \ge 1$, define

$$M_{\ell,p} := \max_j \max\Big\{\|(\mathbf{H}_\ell)_j\|_{L_p}, \|(\widehat{\mathbf{H}}_\ell)_j\|_{L_p}\Big\}.$$

Because every column of $\mathbf{W}_1$ and $\widehat{\mathbf{W}}_1$ has unit Euclidean norm, $(\mathbf{H}_1)_j$, $(\widehat{\mathbf{H}}_1)_j \sim \mathcal{N}(0,1)$, and hence $M_{1,p} \le C_p$. For $\ell \ge 2$, every column of both $\mathbf{W}_\ell$ and $\widehat{\mathbf{W}}_\ell$ is a probability vector. Therefore Minkowski's inequality gives $M_{\ell,p} \le M_{\ell-1,kp}^k$. Since k and L are fixed, $M_{\ell,p} \le C_{k,L,p}, 1 \le \ell \le L-1$, for every fixed $p < \infty$.

Next define $D_{\ell,p} := \max_j \|(\widehat{\mathbf{H}}_\ell)_j - (\mathbf{H}_\ell)_j\|_{L_p}$. For the first layer, $(\widehat{\mathbf{H}}_1 - \mathbf{H}_1)_j = (\widehat{\mathbf{w}}_{1,j} - \mathbf{w}_{1,j})^\top \mathbf{X}$. Hence

$$D_{1,p} \le C_p \max_j \|\widehat{\mathbf{w}}_{1,j} - \mathbf{w}_{1,j}\|_2 \le C_p \delta.$$

Fix $2 \le \ell \le L-1$. For every j,

$$(\widehat{\mathbf{H}}_\ell - \mathbf{H}_\ell)_j = (\widehat{\mathbf{w}}_{\ell,j} - \mathbf{w}_{\ell,j})^\top \widehat{\mathbf{H}}_{\ell-1}^{\odot k} + \mathbf{w}_{\ell,j}^\top \Big(\widehat{\mathbf{H}}_{\ell-1}^{\odot k} - \mathbf{H}_{\ell-1}^{\odot k}\Big).$$

The first term satisfies

$$\Big\|(\widehat{\mathbf{w}}_{\ell,j} - \mathbf{w}_{\ell,j})^\top \widehat{\mathbf{H}}_{\ell-1}^{\odot k}\Big\|_{L_p} \le \|\widehat{\mathbf{w}}_{\ell,j} - \mathbf{w}_{\ell,j}\|_2 \Big\|\|\widehat{\mathbf{H}}_{\ell-1}^{\odot k}\|_2\Big\|_{L_p} \le C_{k,L,p}\sqrt{r}\,\delta.$$

For the second term, using $\mathbf{w}_{\ell,j} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{w}_{\ell,j} = 1$, we obtain

$$\left\| \mathbf{w}_{\ell,j}^\top \left(\widehat{\mathbf{H}}_{\ell-1}^{\odot k} - \mathbf{H}_{\ell-1}^{\odot k} \right) \right\|_{L_p} \leq \max_i \left\| (\widehat{H}_{\ell-1,i})^k - (H_{\ell-1,i})^k \right\|_{L_p}.$$

Using $|u^k - v^k| \leq k|u-v| \left(|u|^{k-1} + |v|^{k-1} \right)$ and Hölder's inequality, together with the preceding uniform moment bounds, gives

$$\left\| (\widehat{H}_{\ell-1,i})^k - (H_{\ell-1,i})^k \right\|_{L_p} \leq C_{k,L,p} D_{\ell-1,2p}.$$

Consequently, $D_{\ell,p} \leq C_{k,L,p} \left(\sqrt{r}\, \delta + D_{\ell-1,2p} \right)$. Iterating this inequality through the finitely many layers yields

$$D_{\ell,p} \leq C_{k,L,p} \sqrt{r}\, \delta, \qquad 1 \leq \ell \leq L-1.$$

Finally,

$$f^\star(\mathbf{X}) = \mathbf{a}^\top \mathbf{H}_{L-1}^{\odot k}, \qquad \widehat{f}(\mathbf{X}) = \widehat{\mathbf{a}}^\top \widehat{\mathbf{H}}_{L-1}^{\odot k},$$

and therefore

$$\widehat{f}(\mathbf{X}) - f^\star(\mathbf{X}) = (\widehat{\mathbf{a}} - \mathbf{a})^\top \widehat{\mathbf{H}}_{L-1}^{\odot k} + \mathbf{a}^\top \left(\widehat{\mathbf{H}}_{L-1}^{\odot k} - \mathbf{H}_{L-1}^{\odot k} \right).$$

The first term satisfies

$$\left\| (\widehat{\mathbf{a}} - \mathbf{a})^\top \widehat{\mathbf{H}}_{L-1}^{\odot k} \right\|_{L_2} \leq \|\widehat{\mathbf{a}} - \mathbf{a}\|_1 \max_i \|(\widehat{H}_{L-1,i})^k\|_{L_2} \leq C_{k,L} \delta.$$

Since $\mathbf{a} \geq \mathbf{0}$ and $\mathbf{1}^\top \mathbf{a} = 1$,

$$\left\| \mathbf{a}^\top \left(\widehat{\mathbf{H}}_{L-1}^{\odot k} - \mathbf{H}_{L-1}^{\odot k} \right) \right\|_{L_2} \leq \max_i \left\| (\widehat{H}_{L-1,i})^k - (H_{L-1,i})^k \right\|_{L_2} \leq C_{k,L} D_{L-1,4} \leq C_{k,L} \sqrt{r}\, \delta.$$

Hence

$$\|\widehat{f} - f^\star\|_{L_2(\mathcal{N}(\mathbf{0}, \mathbf{I}_d))} \leq C_{k,L} \sqrt{r}\, \delta,$$

which proves

$$\mathbb{E}_{\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \left[\left(\widehat{f}(\mathbf{X}) - f^\star(\mathbf{X}) \right)^2 \right] \leq C_{k,L} r \delta^2.$$

□

Proof of Corollary 5.2. Fix $0 < \varepsilon < 1/2$, and choose

$$\delta := c_{k,L} \sqrt{\frac{\varepsilon}{r}},$$

where $c_{k,L} > 0$ is sufficiently small that $C_{k,L} r \delta^2 \leq \varepsilon$. After decreasing the constant once more if necessary, $0 < \delta < 1/2$. Apply Theorem 5.1 with this value of δ. With probability at least 0.99, the returned parameters satisfy the hypotheses of Lemma G.1, and therefore

$$\mathbb{E}_{\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)} \left[(\widehat{f}(\mathbf{X}) - f^\star(\mathbf{X}))^2 \right] \leq C_{k,L} r \delta^2 \leq \varepsilon.$$

By Theorem 5.1, the number of real function-value queries is

$$\begin{aligned} \widetilde{O}_{k,L} \Big(dr + r^{L^2 + O(L)} \gamma^{O(L)} \kappa^{O(L^2)} \delta^{-2\mathrm{e}} \Big) &= \widetilde{O}_{k,L} \Big(dr + r^{L^2 + O(L) + \mathrm{e}} \gamma^{O(L)} \kappa^{O(L^2)} \varepsilon^{-\mathrm{e}} \Big) \\ &= \widetilde{O}_{k,L} \Big(dr + r^{L^2 + O(L)} \gamma^{O(L)} \kappa^{O(L^2)} \varepsilon^{-\mathrm{e}} \Big), \end{aligned}$$

because the fixed factor r^{e} is absorbed into $r^{O(L)}$ for $L \geq 3$. This proves the corollary. □

Proof of Corollary 5.3. Fix $0 < \varepsilon < \dfrac{1}{2}$ and run our recovery algorithm with $\delta = c_{k,L} \sqrt{\dfrac{\varepsilon}{r}}$, where $c_{k,L} > 0$ is sufficiently small. On the successful recovery event, which has probability at least 0.99, Lemma G.1 gives $\mathbb{E}[(\widehat{f}(\mathbf{X}) - f^\star(\mathbf{X}))^2] \leq \dfrac{\varepsilon}{4}$. Fix any returned parameters on this event. All expectations below are over an independent $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The recovered first-layer columns have unit Euclidean norm, and the higher hidden-layer columns are nonnegative and sum to one. Moreover, $\|\widehat{\mathbf{a}}\|_1 \leq \|\mathbf{a}\|_1 + \delta \leq 2$; no positivity assumption on $\widehat{\mathbf{a}}$ is needed.

A bounded piecewise-linear activation. Write $q := k^{L-1}$, and let $R \geq 1$ and an even integer $M \geq 2$ be chosen below. Let P_M be the linear interpolant of t^k at the knots $t_j = -1 + \dfrac{2j}{M}$, $0 \leq j \leq M$, extended by the constant 1 outside $[-1, 1]$. Since k is even, this extension is continuous and $0 \leq P_M(t) \leq 1$ on $\mathbb{R}$. The linear interpolation remainder, using $\sup_{|t|\leq 1} |(t^k)''| \leq k(k-1)$, gives $\sup_{|t|\leq 1} |P_M(t) - t^k| \leq C_k M^{-2}$. If b_j is the jump in the slope of P_M at t_j, then

$$P_M(t) = 1 + \sum_{j=0}^{M} b_j \,\mathrm{ReLU}(t - t_j). \tag{44}$$

Indeed, both sides equal 1 to the left of all knots and have the same slope on every interval.

Define f_{ReLU} by replacing each power activation in $\widehat{f}\Big(\dfrac{\mathbf{x}}{R}\Big)$ with P_M and multiplying the output by R^q. Without these replacements, homogeneity would give $R^q \widehat{f}\Big(\dfrac{\mathbf{x}}{R}\Big) = \widehat{f}(\mathbf{x})$. All parameters of this construction are determined by the recovered network and the interpolation knots.

Approximation on bounded features. Set $Z_j := \widehat{\mathbf{w}}_{1,j}^{\top}\mathbf{X}$ and $V := \max_{1\leq j\leq r} |Z_j|$. On $\{V \leq R\}$, the first preactivations of the rescaled polynomial network lie in $[-1, 1]$. Every higher hidden linear map is a convex combination and is therefore nonexpansive in the maximum norm. Thus both the original and modified activation outputs remain in $[0, 1]$ at every layer on this event. Since $t \mapsto t^k$ is k-Lipschitz on $[-1, 1]$, the maximum coordinatewise activation errors satisfy $e_1 \leq C_k M^{-2}$ and $e_\ell \leq C_k M^{-2} + k e_{\ell-1}$ for $2 \leq \ell \leq L-1$. Iteration and $\|\widehat{\mathbf{a}}\|_1 \leq 2$ yield

$$|f_{\mathrm{ReLU}}(\mathbf{X}) - \widehat{f}(\mathbf{X})| \leq C_{k,L} R^q M^{-2} \qquad \text{on } \{V \leq R\}. \tag{45}$$

Gaussian tails and the error budget. The same column normalization gives $|\widehat{f}(\mathbf{X})| \leq 2V^q$ everywhere, while the bounded range of P_M gives $|f_{\mathrm{ReLU}}(\mathbf{X})| \leq 2R^q$. Hence the squared approximation error is at most $16V^{2q}$ on $\{V > R\}$. Each Z_j is standard Gaussian, although the coordinates need not be independent. For a standard Gaussian variable Z, integrating its density and using $t^{2q} e^{-\frac{t^2}{4}} \leq C_q$ gives $\mathbb{E}\left[|Z|^{2q} \mathbf{1}_{\{|Z|>R\}}\right] \leq C_q e^{-R^2/4}$ for $R \geq 1$. Since $V^{2q}\mathbf{1}_{\{V>R\}} \leq \sum_{j=1}^{r} |Z_j|^{2q} \mathbf{1}_{\{|Z_j|>R\}}$, combining this tail estimate with (45) gives

$$\mathbb{E}\left[\left(f_{\mathrm{ReLU}}(\mathbf{X}) - \widehat{f}(\mathbf{X})\right)^2\right] \leq C_{k,L}\left(\frac{R^{2q}}{M^4} + r e^{-R^2/4}\right). \tag{46}$$

Choose $R = \sqrt{C_{k,L} \log\left(\dfrac{2r}{\varepsilon}\right)}$ and $M = 2\left\lceil C_{k,L} R^{q/2} \varepsilon^{-1/4}\right\rceil$, with sufficiently large constants. The two contributions in (46) are then at most $\dfrac{\varepsilon}{8}$ each. Consequently,

$$\mathbb{E}\left[\left(f_{\mathrm{ReLU}}(\mathbf{X}) - f^\star(\mathbf{X})\right)^2\right] \leq 2\mathbb{E}\left[\left(f_{\mathrm{ReLU}}(\mathbf{X}) - \widehat{f}(\mathbf{X})\right)^2\right] + 2\mathbb{E}\left[\left(\widehat{f}(\mathbf{X}) - f^\star(\mathbf{X})\right)^2\right] \leq \varepsilon.$$

Depth, width, and query complexity. By (44), each scalar replacement uses $M + 1$ parallel ReLU units followed by an affine readout. This readout is absorbed into the next layer's affine map, or into the output layer at the final stage. The constant extensions are already included in the same representation, so no additional clipping layer is needed. The input and output rescalings are absorbed into the first and final affine maps. Thus the resulting network has exactly $L - 1$ hidden ReLU layers and an affine output layer, with biases and widths bounded by

$$r_\ell^{\mathrm{ReLU}} \leq (M+1) r_\ell = O_{k,L}\left(r_\ell \varepsilon^{-1/4}\left[\log\left(\frac{2r}{\varepsilon}\right)\right]^{q/4}\right) = \widetilde{O}_{k,L}\left(r_\ell \varepsilon^{-1/4}\right). \tag{47}$$

For fixed k, L, these parameters can be constructed in polynomial time. No additional target queries or randomness are used, so the success probability remains at least 0.99. Finally, substituting $\delta = c_{k,L}\sqrt{\dfrac{\varepsilon}{r}}$ into Theorem 5.1 gives

$$\widetilde{O}_{k,L}\Big(dr + r^{L^2+O(L)} \gamma^{O(L)} \kappa^{O(L^2)} \varepsilon^{-\mathrm{e}}\Big)$$

queries, with the additional factor r^e absorbed into $r^{O(L)}$. This proves the corollary. □

H A WEIGHT-BASED SUFFICIENT CONDITION FOR THE ASSUMPTIONS

We construct a class of networks to demonstrate that the assumptions in Section 3 are not vacuous. Throughout this section, the hidden layers have a common width r, so that $d = r_0 \geq r_1 = \cdots = r_{L-1} = r \geq 3$ and $r_L = 1$.

Proposition H.1 (Ordered diagonal separation)**.** *Fix an even integer $k \geq 4$ and a depth $L \geq 3$. There exist constants $c_0, c_1 \in (0, 1]$, depending only on k, L, with the following property. Let $\mathbf{W}_1 \in \mathbb{R}^{d\times r}$ satisfy $\mathbf{W}_1^\top \mathbf{W}_1 = \mathbf{I}_r$, and set $\mathbf{W}_L = \mathbf{a} = \frac{1}{r}\mathbf{1}_r$. For each $2 \leq m \leq L-1$, choose the off-diagonal entries of $\mathbf{W}_m \in \mathbb{R}^{r\times r}$ to satisfy $(\mathbf{W}_m)_{ij} = (\mathbf{W}_m)_{ji} > 0$ for $i \neq j$, and set $(\mathbf{W}_m)_{ii} = 1 - \sum_{j\neq i}(\mathbf{W}_m)_{ij}$. Suppose that, in these same coordinate labels at all hidden layers, there are $\varepsilon, \Delta > 0$ such that, for every $2 \leq m \leq L-1$,*

$$\max_{1\leq i\leq r}\big(1 - (\mathbf{W}_m)_{ii}\big) \leq \varepsilon, \qquad (\mathbf{W}_m)_{ii} - (\mathbf{W}_m)_{i+1,i+1} \geq \Delta \quad (1 \leq i < r), \tag{48}$$

and

$$r^2\varepsilon^2 \leq c_0\Delta. \tag{49}$$

Then Assumptions 3.2, 3.3, 3.4 and 3.5 all hold, with the choices

$$\kappa_0 = 2, \qquad \mu = 1, \qquad \gamma = \frac{1}{c_1\Delta}. \tag{50}$$

In particular, every hidden weight matrix has full column rank, and

$$\frac{\min_{i\neq j}|\theta_{\ell,i} - \theta_{\ell,j}|}{\max_i \theta_{\ell,i}} \geq c_1\Delta, \qquad 1 \leq \ell \leq L-2. \tag{51}$$

The next three subsections prove the proposition. All constants denoted by C below depend only on k, L, may increase from line to line, and are independent of $d, r, \varepsilon, \Delta$, and the particular weights. The constant c_0 is chosen sufficiently small at the end.

H.1 NORMALIZATION AND NONNEGATIVITY

The identity $\mathbf{W}_1^\top \mathbf{W}_1 = \mathbf{I}_r$ gives $\|\mathbf{w}_{1,j}\|_2 = 1$ for every j. The definition of the higher-layer diagonal entries gives $\mathbf{W}_m\mathbf{1}_r = \mathbf{1}_r$. Their symmetry therefore also gives $\mathbf{1}_r^\top\mathbf{W}_m = \mathbf{1}_r^\top$, so every column is normalized as required.

To check positivity of the diagonal entries, telescope the ordered gaps in any one higher hidden layer. Since every off-diagonal entry is positive, $0 < 1 - (\mathbf{W}_m)_{ii} \leq \varepsilon$, and hence $(r-1)\Delta \leq \varepsilon$. Together with (49), this implies

$$\varepsilon \leq \frac{c_0}{r^2(r-1)} \leq \frac{1}{18} < \frac{1}{4}. \tag{52}$$

Consequently $(\mathbf{W}_m)_{ii} \geq 1 - \varepsilon > 0$, and all the higher hidden matrices are strictly positive and doubly stochastic. Finally, $a_i = \frac{1}{r} > 0$ and $\mathbf{1}_r^\top\mathbf{a} = 1$, which verifies the output-layer requirements as well. Notice also that $0 < \Delta \leq \varepsilon < 1$; thus the value of γ in (50) is at least one.

H.2 CONDITIONING AND RELATIVE ROW BALANCE

All singular values of $\mathbf{W}_1$ equal one. For $2 \leq m \leq L-1$, symmetry and the definition of the diagonal give $\|\mathbf{W}_m - \mathbf{I}_r\|_1 = \|\mathbf{W}_m - \mathbf{I}_r\|_\infty \leq 2\varepsilon$, and therefore $\|\mathbf{W}_m - \mathbf{I}_r\|_2 \leq 2\varepsilon$. Moreover, double stochasticity gives $\|\mathbf{W}_m\|_2 \leq 1$, while $\mathbf{W}_m\mathbf{1}_r = \mathbf{1}_r$ gives the reverse inequality. It follows that $\sigma_{\max}(\mathbf{W}_m) = 1$ and $\sigma_{\min}(\mathbf{W}_m) \geq 1 - 2\varepsilon > 0$. Thus $\mathrm{cond}(\mathbf{W}_m) \leq \frac{1}{1-2\varepsilon} \leq 2$. This proves full column rank and the conditioning bound with $\kappa_0 = 2$.

For the row-balance condition, the higher hidden layers satisfy $(\mathbf{W}_m \mathbf{1}_r)_i = 1 = \dfrac{r_m}{r_{m-1}}$ for every i. At the output layer, $(\mathbf{W}_L \mathbf{1}_1)_i = a_i = \dfrac{1}{r} = \dfrac{r_L}{r_{L-1}}$. Hence the relative row-balance lower bound holds with equality at every relevant layer when $\mu = 1$.

H.3 Uniform relative spectral gap

Proof of the spectral-gap assertion. Fix $1 \le \ell \le L-2$, and write $q = k^{L-\ell}$. For $m = \ell+1, \dots, L-1$, consider the interpolation $\mathbf{W}_m(t) = \mathbf{I}_r + \dfrac{t}{\varepsilon}(\mathbf{W}_m - \mathbf{I}_r)$, $0 \le t \le 1$. Every such matrix is nonnegative and doubly stochastic: its diagonal entries are at least $1-t$, and its off-diagonal entries are nonnegative. Let $h_t^{(\ell)}$ be the corresponding suffix, with the output weights kept equal to $\dfrac{1}{r}\mathbf{1}_r$, and set $G_t = r h_t^{(\ell)}$. Then $G_0(\mathbf{u}) = Q(\mathbf{u}) := \sum_{j=1}^r u_j^q$, and $G_\varepsilon = r h^{(\ell)}$. Write $\theta_{\ell,i}(t)$ for the spectral signature defined from $h_t^{(\ell)}$ exactly as in Assumption 3.5. In particular, $\theta_{\ell,i}(\varepsilon) = \theta_{\ell,i}$. At $t = 0$, permutation symmetry makes these signatures equal; denote their common positive value by $\theta_\ell^{(0)}$.

Exponential-integral representation. A positive coordinate rescaling does not change a spectral signature. Indeed, replacing $g(\mathbf{u})$ by $g(s_1 u_1, \dots, s_r u_r)$, with all $s_j > 0$, multiplies its i-th position moment by s_i^{-2} and its i-th gradient moment by s_i^2, by the corresponding change of variables in the uniform distribution. Multiplying a degree-q homogeneous polynomial by a positive scalar is a special case of this rescaling. Consequently, G_t and $h_t^{(\ell)}$ have the same signatures.

For a positive-definite homogeneous polynomial g of degree q and a homogeneous polynomial p, integration over the sublevel sets of g gives

$$\int_{\mathbb{R}^r} p(\mathbf{u}) e^{-g(\mathbf{u})}\, d\mathbf{u} = \Gamma\!\left(1 + \frac{r + \deg p}{q}\right) \int_{\{g \le 1\}} p(\mathbf{u})\, d\mathbf{u}. \tag{53}$$

Indeed, write $e^{-g(\mathbf{u})} = \int_{g(\mathbf{u})}^\infty e^{-v}\, dv$ and interchange the integrals. In the inner integral over $\{g \le v\}$, rescale $\mathbf{u} = v^{\frac{1}{q}} \mathbf{w}$. The remaining scalar integral is the displayed gamma factor. Absolute integrability justifies the same argument when p is not nonnegative.

Let $Z_t = \int_{\mathbb{R}^r} e^{-G_t(\mathbf{u})}\, d\mathbf{u}$, and write $\mathbb{E}_t$ and Cov_t for expectation and covariance under the density $Z_t^{-1} e^{-G_t}$. Define $A_i(t) = \mathbb{E}_t[X_i^2]$ and $B_i(t) = \mathbb{E}_t[\partial_{ii} G_t(\mathbf{X})]$. Integration by parts gives $\mathbb{E}_t[(\partial_i G_t(\mathbf{X}))^2] = B_i(t)$. All these integrals are well defined for t near zero; a uniform coercive bound is established below. At $t = 0$, the density is a product density, with one-dimensional moments $m_q(v) := \mathbb{E}_0[|X_i|^v] = \dfrac{\Gamma\!\left(\dfrac{v+1}{q}\right)}{\Gamma\!\left(\dfrac{1}{q}\right)}$ for $v \ge 0$. Thus $A_i(0) = m_q(2)$ and $B_i(0) = q(q-1)m_q(q-2)$.

Applying (53) to $p = 1$, $p = u_i^2$, and $p = (\partial_i g)^2$, the gamma factors cancel between t and zero, leaving

$$\frac{\theta_{\ell,i}(t)}{\theta_\ell^{(0)}} = \frac{A_i(t) B_i(t)}{m_q(2)\, q(q-1) m_q(q-2)}. \tag{54}$$

First-order response to the weights. A dot denotes differentiation with respect to t. Differentiating the normalized exponential integrals at zero gives

$$\begin{aligned} A_i'(0) &= -\operatorname{Cov}_0(X_i^2, \dot G_0(\mathbf{X})), \\ B_i'(0) &= \mathbb{E}_0[\partial_{ii} \dot G_0(\mathbf{X})] - q(q-1)\operatorname{Cov}_0(X_i^{q-2}, \dot G_0(\mathbf{X})). \end{aligned} \tag{55}$$

By (54), the normalized derivative of the signature is $\dfrac{A_i'(0)}{m_q(2)} + \dfrac{B_i'(0)}{q(q-1)m_q(q-2)}$, which is linear in $\dot G_0$. If a term in $\dot G_0$ is $u_i^s u_j^{q-s}$, with $i \ne j$ and even $s, q-s \ge 4$, its contribution to this derivative

is

$$\chi_q(s) := m_q(q-s)\left[2m_q(s) - \frac{m_q(s+2)}{m_q(2)} - \frac{(s-1)(q-s-1)m_q(s-2)}{q(q-1)m_q(q-2)}\right]. \tag{56}$$

Indeed, insert the monomial into (55), use independence, and simplify with $m_q(v+q) = \dfrac{v+1}{q} m_q(v)$. The contribution to coordinate j is $\chi_q(q-s)$, and the contributions to all other coordinates are zero. A pure-power term u_j^q contributes zero to every coordinate, since changing its coefficient in Q is a coordinate rescaling.

We next check the sign in (56). The gamma function is log-convex, by Hölder's inequality applied to its defining integral. Therefore $x \mapsto \dfrac{\Gamma(x+a)}{\Gamma(x)}$ is nondecreasing for fixed $a > 0$. Using this fact with $a = \dfrac{2}{q}$, together with the gamma recurrence, gives

$$\frac{m_q(s+2)}{m_q(2)m_q(s)} = \frac{3(s+1)}{s+3} \frac{\Gamma\left(1+\dfrac{s+3}{q}\right)\Gamma\left(1+\dfrac{1}{q}\right)}{\Gamma\left(1+\dfrac{s+1}{q}\right)\Gamma\left(1+\dfrac{3}{q}\right)} \geq \frac{3(s+1)}{s+3} > 2 \qquad (s \geq 4). \tag{57}$$

The first two terms in the bracket in (56) therefore have negative sum, and its last term is also negative. Hence $\chi_q(s) < 0$ whenever $4 \leq s \leq q-4$ and s is even.

For $\ell + 1 \leq m \leq L-1$, write $s_m = k^{m-\ell}$. Differentiating the suffix recursion at the identity matrices yields

$$\dot{G}_0(\mathbf{u}) = \frac{1}{\varepsilon} \sum_{m=\ell+1}^{L-1} \frac{q}{s_m} \left[\sum_{i \neq j} (\mathbf{W}_m)_{ij} u_i^{s_m} u_j^{q-s_m} - \sum_i \big(1 - (\mathbf{W}_m)_{ii}\big) u_i^q\right]. \tag{58}$$

For completeness, at the identity matrices the input to $\mathbf{W}_m^\top$ has coordinates $u_i^{s_m}$; the remaining network is the sum of the $\dfrac{q}{s_m}$-th powers of its coordinates. Differentiating one matrix entry gives precisely the corresponding term in (58).

Symmetry makes both the off-diagonal row sum and the off-diagonal column sum at coordinate i equal to $1 - (\mathbf{W}_m)_{ii}$. Thus the single-edge calculation gives

$$\frac{d}{dt} \frac{\theta_{\ell,i}(t)}{\theta_\ell^{(0)}}\bigg|_{t=0} = -\frac{1}{\varepsilon} \sum_{m=\ell+1}^{L-1} b_{\ell m}\big(1 - (\mathbf{W}_m)_{ii}\big), \qquad b_{\ell m} := -\frac{q}{s_m}\big(\chi_q(s_m) + \chi_q(q-s_m)\big) > 0. \tag{59}$$

Here $s_m \geq k \geq 4$ and $q - s_m \geq 4$, so the strict sign follows from (57). The finitely many coefficients $b_{\ell m}$ depend only on k, L, ℓ, m, not on the width or the weights.

A uniform second-order remainder. We prove that, for some $\delta > 0$ depending only on k, L,

$$\left|\frac{d^2}{dt^2} \frac{\theta_{\ell,i}(t)}{\theta_\ell^{(0)}}\right| \leq Cr^2, \qquad 0 \leq t \leq \frac{\delta}{r}. \tag{60}$$

This estimate does not use any spectral separation.

We first bound the polynomial and its derivatives. Every nonnegative doubly stochastic matrix is a contraction in each ℓ_p norm, $1 \leq p < \infty$: Jensen's inequality gives $\sum_j |(\mathbf{W}^\top \mathbf{v})_j|^p \leq \sum_{i,j} W_{ij}|v_i|^p = \sum_i |v_i|^p$. Applying this observation through the suffix, or inducting directly through its recursion, gives $0 \leq G_\tau(\mathbf{u}) \leq Q(\mathbf{u})$ for $0 \leq \tau \leq 1$. At the all-ones vector every hidden coordinate remains one, so $G_\tau(\mathbf{1}_r) = r$ and $\partial_i G_\tau(\mathbf{1}_r) = q$.

The degree of G_t as a polynomial in t is at most $\sum_{m=\ell+1}^{L-1} \dfrac{q}{s_m} = \dfrac{q-k}{k-1}$, which depends only on k, L. Fixed-node univariate interpolation on $[0,1]$ therefore expresses each of $\dot{G}_t$ and $\ddot{G}_t$ as a linear

combination of finitely many polynomials G_τ, with the sum of the absolute coefficients at most C, uniformly for $t \in [0,1]$. Consequently,

$$|\dot{G}_t(\mathbf{u})| + |\ddot{G}_t(\mathbf{u})| \le CQ(\mathbf{u}), \qquad (1-Ct)Q(\mathbf{u}) \le G_t(\mathbf{u}) \le Q(\mathbf{u}). \tag{61}$$

The lower bound follows by integrating the bound on $\dot{G}_t$ from zero.

We also need local derivative bounds without a factor of r. Each G_τ has nonnegative coefficients and is homogeneous of degree q. Its second derivative $\partial_{ii}G_\tau$ thus has nonnegative coefficients whose sum is $\partial_{ii}G_\tau(\mathbf{1}_r) \le (q-1)\partial_i G_\tau(\mathbf{1}_r) = q(q-1)$. Applying the same interpolation formula after differentiating in u_i twice shows that, for $j = 0,1,2$, the sum of the absolute coefficients of $\partial_{ii}\partial_t^j G_t$ is at most C. All these polynomials have degree $q-2$.

Choose δ small enough that $Ct \le \frac{1}{2}$ when $0 \le t \le \frac{\delta}{r}$. By (61) and homogeneity, $1 \le \frac{Z_t}{Z_0} \le (1-Ct)^{-\frac{r}{q}} \le C$. For a polynomial p of bounded degree whose sum of absolute coefficients is bounded independently of r, we then have

$$\mathbb{E}_t\big[|p(\mathbf{X})|Q(\mathbf{X})^j\big] \le Cr^j, \qquad j = 0,1,2. \tag{62}$$

Here the bound is uniform for the polynomials just considered. To verify it explicitly, take a monomial with multi-index α, write $|\mathbf{u}|^\alpha = \prod_{i=1}^r |u_i|^{\alpha_i}$, and use

$$\frac{1}{Z_0}\int_{\mathbb{R}^r} |\mathbf{u}|^\alpha Q(\mathbf{u})^j e^{-aQ(\mathbf{u})}\,d\mathbf{u} = a^{-\frac{r+|\alpha|}{q}-j}\frac{\Gamma\!\left(\frac{r+|\alpha|}{q}+j\right)}{\Gamma\!\left(\frac{r+|\alpha|}{q}\right)}\prod_{i=1}^r m_q(\alpha_i). \tag{63}$$

For $j=0$ this is a product of one-dimensional integrals; differentiating in a gives the identities for $j=1,2$. With $a = 1-Ct$, the power of a is uniformly bounded because $rt \le \delta$, the gamma ratio is $O_{k,L}(r^j)$, and the last product is bounded because $|\alpha|$ is bounded. Summing over monomials proves (62).

To differentiate the moments, use the following identities, with all integrals taken over $\mathbb{R}^r$:

$$\begin{aligned}\frac{d}{dt}\int p_t e^{-G_t}\,d\mathbf{u} &= \int (\dot{p}_t - p_t\dot{G}_t)e^{-G_t}\,d\mathbf{u},\\ \frac{d^2}{dt^2}\int p_t e^{-G_t}\,d\mathbf{u} &= \int \Big[\ddot{p}_t - 2\dot{p}_t\dot{G}_t + p_t\big((\dot{G}_t)^2 - \ddot{G}_t\big)\Big]e^{-G_t}\,d\mathbf{u}.\end{aligned} \tag{64}$$

The coercive lower bound in (61) and the polynomial envelopes above justify differentiation under the integral sign, including the derivatives used earlier at zero. Apply (64) to $p_t = 1$, $p_t = u_i^2$, and $p_t = \partial_{ii}G_t$. For each of these unnormalized integrals, the absolute value of its j-th derivative, divided by Z_t, is at most Cr^j for $j = 0,1,2$, by (62). The quotient rule, using $\frac{|Z_t'|}{Z_t} \le Cr$ and $\frac{|Z_t''|}{Z_t} \le Cr^2$, therefore gives $|A_i^{(j)}(t)| + |B_i^{(j)}(t)| \le Cr^j$ for $j = 0,1,2$. The denominator in (54) is a positive constant depending only on q. The product rule now proves (60).

By (52), decreasing c_0 ensures $\varepsilon \le \frac{\delta}{r}$. Taylor's theorem and (59) consequently yield the uniform expansion

$$\frac{\theta_{\ell,i}}{\theta_\ell^{(0)}} = 1 - \sum_{m=\ell+1}^{L-1} b_{\ell m}\big(1 - (\mathbf{W}_m)_{ii}\big) + O_{k,L}(r^2\varepsilon^2). \tag{65}$$

From ordered diagonals to a relative gap. Since there are only finitely many positive coefficients $b_{\ell m}$, there is a constant $b > 0$, depending only on k, L, such that $\sum_{m=\ell+1}^{L-1} b_{\ell m} \ge b$ for every $1 \le \ell \le L-2$. For $i < j$, the common ordering gives $(\mathbf{W}_m)_{ii} - (\mathbf{W}_m)_{jj} \ge (j-i)\Delta \ge \Delta$ in every higher layer. Subtracting the two instances of (65) therefore gives

$$\frac{\theta_{\ell,i} - \theta_{\ell,j}}{\theta_\ell^{(0)}} \ge b\Delta - Cr^2\varepsilon^2 \ge (b - Cc_0)\Delta \ge \frac{b\Delta}{2}, \tag{66}$$

where the last inequality follows by choosing c_0 sufficiently small. The sign of the first-order term in (65) also gives $\max_i \theta_{\ell,i} \le (1 + Cr^2\varepsilon^2)\theta_\ell^{(0)} \le 2\theta_\ell^{(0)}$, after the same choice of c_0. Dividing the last two bounds proves (51) with $c_1 = \min\{1, \frac{b}{4}\}$, uniformly over all $1 \le \ell \le L-2$. This completes the proof of Proposition H.1. □

H.4 An explicit nonempty family

The conditions hold for every $r \ge 3$, not just for a particular small width. Choose any $\mathbf{W}_1$ with orthonormal columns, set $\mathbf{a} = \frac{1}{r}\mathbf{1}_r$, and, for $2 \le m \le L-1$, take

$$(\mathbf{W}_m)_{ij} = \begin{cases} \dfrac{t(i+j)}{r^2}, & i \ne j, \\ 1 - \dfrac{t}{r^2}\left((r-2)i + \dfrac{r(r+1)}{2}\right), & i = j, \end{cases} \qquad 0 < t \le \frac{c_0}{3r^3}. \tag{67}$$

These matrices are symmetric, and the diagonal entries are precisely one minus the corresponding off-diagonal row sums. Their largest off-diagonal row sum is $\varepsilon = \dfrac{3(r-1)t}{2r}$, and their consecutive diagonal gaps equal $\Delta = \dfrac{(r-2)t}{r^2}$. Moreover,

$$\frac{r^2\varepsilon^2}{\Delta} = \frac{9tr^2(r-1)^2}{4(r-2)} \le 3tr^3 \le c_0, \tag{68}$$

where the first inequality uses $3(r-1)^2 \le 4r(r-2)$ for $r \ge 3$. Thus Proposition H.1 applies. In particular, setting $t = \dfrac{c_0}{3r^3}$ gives $\Delta = \dfrac{c_0(r-2)}{3r^5}$ and $\gamma = \dfrac{3r^5}{c_0 c_1 (r-2)} = O_{k,L}(r^4)$, while $\kappa_0 = 2$ and $\mu = 1$.

The first-layer orthonormality is used only for its normalization and conditioning, not for the spectral-gap argument. It can be replaced by unit-norm columns and any prescribed bound on $\mathrm{cond}(\mathbf{W}_1)$; the conclusion then uses $\kappa_0 = \max\{2, \mathrm{cond}(\mathbf{W}_1)\}$. The restriction to $r \ge 3$ is natural for this construction: a symmetric doubly stochastic 2×2 matrix necessarily has equal diagonal entries. Finally, the condition is local: (52) requires coupling of order at most r^{-3} for fixed k, L. No spectral-gap claim is made here for symmetric matrices outside this weak-coupling regime.

I Proof of Challenges

I.1 Local minima and fourth-order saddle barriers

Proof of Proposition 3.6. We give two explicit realizable instances. In both cases, let $f^\star(\mathbf{x}) = \frac{1}{2}(\mathbf{e}_1^\top \mathbf{x})^4 + \frac{1}{2}(\mathbf{e}_3^\top \mathbf{x})^4$. Thus the teacher has $\mathbf{W}_1^\star = [\mathbf{e}_1, \mathbf{e}_3]$ and $\mathbf{a}^\star = (1/2, 1/2)$. Its first-layer columns are orthonormal, and its output weights are positive, normalized, and row balanced, so all assumptions applicable to the two-layer case are satisfied.

A non-global local minimum. Consider the student point $\bar{\mathbf{w}}_1 = \mathbf{e}_2$, $\bar{\mathbf{w}}_2 = \mathbf{e}_3$, and $\bar{\mathbf{a}} = (1/2, 1/2)$. We work in the normalized parameterization and use the local coordinates $\boldsymbol{\xi} = (p, q, s, t, u)$ defined by $\mathbf{w}_1 = p\mathbf{e}_1 + \sqrt{1-p^2-q^2}\mathbf{e}_2 + q\mathbf{e}_3$, $\mathbf{w}_2 = s\mathbf{e}_1 + t\mathbf{e}_2 + \sqrt{1-s^2-t^2}\mathbf{e}_3$, $a_1 = 1/2 + u$, and $a_2 = 1/2 - u$. The point $\bar{\boldsymbol{\theta}}$ corresponds to $\boldsymbol{\xi} = \mathbf{0}$.

Let $c = 2^{1/4}$ and introduce the four samples $\mathcal{G} = \{\mathbf{e}_2, -\mathbf{e}_2, c\mathbf{e}_1 + \mathbf{e}_2, -c\mathbf{e}_1 + \mathbf{e}_2\}$. At $\boldsymbol{\xi} = \mathbf{0}$, their residuals are respectively $1/2, 1/2, -1/2, -1/2$. For a point $\mathbf{x} = (x_1, x_2, 0) \in \mathcal{G}$, the first variation of its residual is $Dr_{\mathbf{x}}(\mathbf{0})[\boldsymbol{\xi}] = 2x_2^3 x_1 p + x_2^4 u$. Hence $\sum_{\mathbf{x}\in\mathcal{G}} r_{\mathbf{x}} Dr_{\mathbf{x}}(\mathbf{0}) = 0$, so these four samples contribute zero gradient at $\bar{\boldsymbol{\theta}}$.

We next add exact-fit samples to make this stationary point locally stable. Let $\mathcal{A} = \{(1,1,0), (-1,1,0), (1,1,1), (-1,1,1), (1,1,2)\}$. For every $\mathbf{x} \in \mathcal{A}$, one has $(\mathbf{e}_1^\top \mathbf{x})^4 = (\mathbf{e}_2^\top \mathbf{x})^4$, and therefore the teacher and the student agree exactly at $\bar{\boldsymbol{\theta}}$. Writing an anchor point as $(\varepsilon, 1, z)$ with

$\varepsilon \in \{\pm 1\}$, its residual Jacobian at $\boldsymbol{\xi} = \mathbf{0}$ is $(2\varepsilon,\ 2z,\ 2\varepsilon z^3,\ 2z^3,\ 1 - z^4)$. For the five points above, the resulting Jacobian matrix is

$$\mathbf{J} = \begin{pmatrix} 2 & 0 & 0 & 0 & 1 \\ -2 & 0 & 0 & 0 & 1 \\ 2 & 2 & 2 & 2 & 0 \\ -2 & 2 & -2 & 2 & 0 \\ 2 & 4 & 16 & 16 & -15 \end{pmatrix},$$

whose determinant is 384. Thus $\mathbf{J}$ is invertible and $\mathbf{J}^\top \mathbf{J} \succ 0$.

Form an empirical data multiset by taking the four points in $\mathcal{G}$ once and each point in $\mathcal{A}$ exactly M times. Let $\mathbf{H}_{\mathcal{G}}$ denote the Hessian contribution of the four points in $\mathcal{G}$ at $\boldsymbol{\xi} = \mathbf{0}$. Since all anchor residuals vanish, their Hessian contribution is precisely $\mathbf{J}^\top \mathbf{J}$ per copy, up to the common positive normalization of the empirical loss. Therefore the total Hessian is proportional to $\mathbf{H}_{\mathcal{G}} + M\mathbf{J}^\top \mathbf{J}$. Choosing $M > \|\mathbf{H}_{\mathcal{G}}\|_{\mathrm{op}} / \lambda_{\min}(\mathbf{J}^\top \mathbf{J})$ makes this matrix positive definite. Hence $\bar{\boldsymbol{\theta}}$ is a strict local minimum in the normalized parameterization.

It is not global: its four residuals on $\mathcal{G}$ are nonzero, so its empirical loss is strictly positive, whereas the teacher parameters achieve zero loss on every sample. This proves the existence of a non-global local minimum.

A fourth-order saddle. For the second instance, use only the two samples $\mathbf{x}_1 = \mathbf{e}_1$ and $\mathbf{x}_2 = \mathbf{e}_3$ and consider the student point $\widetilde{\mathbf{w}}_1 = \mathbf{e}_1$, $\widetilde{\mathbf{w}}_2 = \mathbf{e}_2$, and $\widetilde{\mathbf{a}} = (1/2, 1/2)$. The two residuals are $r_1 = 0$ and $r_2 = -1/2$. Since both student preactivations vanish at $\mathbf{x}_2$, the quartic activation gives $Dr_2 = D^2 r_2 = D^3 r_2 = 0$ at this point. For the empirical squared loss $\mathcal{L} = \frac{1}{4}(r_1^2 + r_2^2)$, it follows that $\nabla \mathcal{L}(\widetilde{\boldsymbol{\theta}}) = 0$ and $\nabla^2 \mathcal{L}(\widetilde{\boldsymbol{\theta}}) = \frac{1}{2} \nabla r_1 \nabla r_1^\top \succeq 0$.

Moreover, if a direction lies in the nullspace of this Hessian, then the first-order variation of r_1 vanishes, so r_1^2 changes only at fourth order, while $r_2 = -1/2 + O(\|\boldsymbol{\delta}\|^4)$. Hence no descent term of order one, two, or three can occur along a Hessian-flat direction.

Finally, perturb only the second hidden neuron along the feasible normalized path $\mathbf{w}_2(t) = \sqrt{1 - t^2}\mathbf{e}_2 + t\mathbf{e}_3$. The first sample remains fitted exactly, while on the second sample the student output is $\frac{1}{2}t^4$. Consequently, $\mathcal{L}(\widetilde{\boldsymbol{\theta}}(t)) = \frac{1}{4}(-\frac{1}{2} + \frac{1}{2}t^4)^2 = \frac{1}{16} - \frac{1}{8}t^4 + \frac{1}{16}t^8$. Thus the loss decreases for all sufficiently small nonzero t, and its first descent term is quartic. Since the Hessian also has positive directions, this point is a genuine fourth-order saddle. □

I.2 Sample-complexity barrier for the NTK

We prove Proposition 3.7 by applying the spherical kernel lower bound of Ghorbani et al. (2021, Theorem 3). The Gaussian-to-sphere reduction and the verification of a nonvanishing highest-degree component for our admissible targets are supplied below.

The spherical result. Let $\nu_d = \mathrm{Unif}(S^{d-1}(\sqrt{d}))$. Write $P_{\le m}$ for the orthogonal projection in $L^2(\nu_d)$ onto restrictions of polynomials of total degree at most m, and set $P_{>m} = I - P_{\le m}$. For a fixed rotationally invariant kernel

$$K_d(\mathbf{u}, \mathbf{v}) = \psi_d\left(\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{d}\right), \qquad \mathbf{u}, \mathbf{v} \in S^{d-1}(\sqrt{d}),$$

and independent $\mathbf{U}_i \sim \nu_d$, define

$$R_{K,d}(f; \mathbf{U}_{1:n}) := \inf_{\boldsymbol{\beta} \in \mathbb{R}^n} \left\| f - \sum_{i=1}^{n} \beta_i K_d(\cdot, \mathbf{U}_i) \right\|_{L^2(\nu_d)}^2. \tag{69}$$

Theorem 3 of Ghorbani et al. (2021) states that, for a fixed integer $m \ge 0$, if

$$n \le d^{m+1-\delta_d}, \qquad \delta_d > 0, \qquad \delta_d^2 \log d \longrightarrow \infty,$$

then the following holds for any deterministic target sequence $f_d \in L^2(\nu_d)$ and any fixed kernel sequence whose spherical sections are square integrable. For every fixed $\epsilon > 0$, with probability

tending to one,

$$\begin{aligned}&\left|R_{K,d}(f_d;\mathbf{U}_{1:n}) - R_{K,d}(P_{\le m}f_d;\mathbf{U}_{1:n}) - \|P_{>m}f_d\|^2_{L^2(\nu_d)}\right| \\ &\le \epsilon\|f_d\|_{L^2(\nu_d)}\|P_{>m}f_d\|_{L^2(\nu_d)}.\end{aligned} \tag{70}$$

The kernel and target are fixed before the sample is drawn. The risk in (69) minimizes over all coefficients, rather than over coefficients obtained by a particular training algorithm. No assumption that all low-degree kernel eigenvalues are positive is required for this lower bound.

Proof of Proposition 3.7. Throughout the proof, k, L, and hence $q = k^{L-1}$, are fixed. We abbreviate K_d^{NTK} by K_d.

Step 1: NTK symmetry, homogeneity, and integrability. For a finite-width bias-free network with the specified activation, let $F_{\boldsymbol{\vartheta}}$ denote its output as a function of all trainable parameters $\boldsymbol{\vartheta}$. Induction through its $L-1$ activations gives

$$F_{\boldsymbol{\vartheta}}(s\mathbf{x}) = s^q F_{\boldsymbol{\vartheta}}(\mathbf{x}).$$

Differentiating in the parameters shows that every tangent feature has the same degree of input homogeneity:

$$\nabla_{\boldsymbol{\vartheta}} F_{\boldsymbol{\vartheta}}(s\mathbf{x}) = s^q \nabla_{\boldsymbol{\vartheta}} F_{\boldsymbol{\vartheta}}(\mathbf{x}).$$

The finite-width tangent kernel, and therefore its deterministic infinite-width limit, consequently satisfy

$$K_d(s\mathbf{x}, t\mathbf{z}) = s^q t^q K_d(\mathbf{x}, \mathbf{z}), \qquad s, t > 0. \tag{71}$$

An orthogonal change of input coordinates can be absorbed into an orthogonal change of the first-layer Gaussian weights. This preserves their initialization law and the Euclidean inner product of parameter gradients. The finite-width kernel is thus rotation invariant in distribution, and its deterministic limit satisfies

$$K_d(\mathbf{O}\mathbf{x}, \mathbf{O}\mathbf{z}) = K_d(\mathbf{x}, \mathbf{z}), \qquad \mathbf{O} \in O(d).$$

This is also the multilayer NTK invariance described in Ghorbani et al. (2021, Section 3). Its restriction to $S^{d-1}(\sqrt{d})$ therefore has the form used in (69). The diagonal values of this fixed-depth monomial NTK are finite, since its covariance and derivative-covariance recursions involve only finite Gaussian polynomial moments. Positive semidefiniteness and rotational invariance give

$$|K_d(\mathbf{u}, \mathbf{v})|^2 \le K_d(\mathbf{u}, \mathbf{u}) K_d(\mathbf{v}, \mathbf{v}) = \psi_d(1)^2 < \infty$$

on the sphere. Thus the square-integrability hypothesis of (70) holds.

Step 2: an exact Gaussian-to-sphere reduction. For the Gaussian training inputs, write

$$R_i := \frac{\|\mathbf{X}_i\|_2}{\sqrt{d}}, \qquad \mathbf{U}_i := \frac{\sqrt{d}\,\mathbf{X}_i}{\|\mathbf{X}_i\|_2}, \qquad \mathbf{X}_i = R_i \mathbf{U}_i.$$

The vectors $\mathbf{U}_i$ are independent with law ν_d, and are independent of the radii R_i, which are strictly positive almost surely. By (71),

$$K_d(\cdot, \mathbf{X}_i) = R_i^q K_d(\cdot, \mathbf{U}_i).$$

Consequently, the Gaussian samples and their normalized directions generate exactly the same kernel span:

$$\operatorname{span}\{K_d(\cdot, \mathbf{X}_i) : 1 \le i \le n\} = \operatorname{span}\{K_d(\cdot, \mathbf{U}_i) : 1 \le i \le n\}. \tag{72}$$

Every function in this span is q-homogeneous. An independent test input has the decomposition $\mathbf{X} = R\mathbf{U}$, with $\mathbf{U} \sim \nu_d$ independent of R. For each integer $j \ge 0$, set

$$A_{d,j} := \mathbb{E}[R^{2j}] = \frac{2^j \Gamma(d/2 + j)}{d^j \Gamma(d/2)} = \prod_{a=0}^{j-1} \left(1 + \frac{2a}{d}\right), \tag{73}$$

where the empty product is one. For any two q-homogeneous functions v, f whose restrictions to the sphere are square integrable, independence and homogeneity give the exact identity

$$\mathbb{E}_{\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)}[(v(\mathbf{X}) - f(\mathbf{X}))^2] = A_{d,q} \|v - f\|^2_{L^2(\nu_d)}. \tag{74}$$

In particular, $A_{d,q} \geq 1$ and $A_{d,q} = 1 + O_q(d^{-1})$. Combining (72) and (74) yields, pathwise,

$$\inf_{\boldsymbol{\alpha}\in\mathbb{R}^n} \mathbb{E}_{\mathbf{X}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)} \left[\left(f_d^\star(\mathbf{X}) - \sum_{i=1}^n \alpha_i K_d(\mathbf{X}, \mathbf{X}_i) \right)^2 \right] = A_{d,q} R_{K,d}(f_d^\star; \mathbf{U}_{1:n}). \tag{75}$$

Because the infimum permits every coefficient vector, this identity supplies a lower bound even when the learned coefficients use all sample radii, labels, and additional randomness. It does not assert equality of the Gaussian and spherical empirical ridge-regression objectives.

Step 3: an admissible constant-width target family. Let c_0, c_1 be the constants in Proposition H.1, and fix $t = c_0/162$, independently of d. For every $d \geq 3$, choose

$$\mathbf{W}_1 = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3), \qquad \mathbf{a} = \frac{1}{3}\mathbf{1}_3, \qquad \mathbf{W}_j = \mathbf{M}_t \quad (2 \leq j \leq L-1),$$

where

$$\mathbf{M}_t = \begin{pmatrix} 1 - 7t/9 & t/3 & 4t/9 \\ t/3 & 1 - 8t/9 & 5t/9 \\ 4t/9 & 5t/9 & 1 - t \end{pmatrix}. \tag{76}$$

The maximum off-diagonal row sum is t, and the consecutive diagonal gaps equal $t/9$. Since

$$3^2 t^2 \leq c_0(t/9),$$

Proposition H.1 applies and verifies all three assumptions in the statement, with

$$\kappa_0 = 2, \qquad \mu = 1, \qquad \gamma = \frac{9}{c_1 t}.$$

These parameters depend only on k, L, not on d. The resulting target has the form

$$f_d^\star(\mathbf{x}) = g(x_1, x_2, x_3), \tag{77}$$

where g is a fixed homogeneous polynomial of degree q. In particular, g itself does not change with d. Write $g(\mathbf{u}) = \sum_{|\boldsymbol{\beta}|=q} b_{\boldsymbol{\beta}} \mathbf{u}^{\boldsymbol{\beta}}$, and set $b_\star := b_{(q,0,0)} = g(1,0,0)$. Keeping only the path that uses the first neuron at every layer gives

$$b_\star \geq \frac{1}{3}\left(1 - \frac{7t}{9}\right)^{s_{k,L}} > 0, \qquad s_{k,L} := \sum_{j=1}^{L-2} k^j. \tag{78}$$

Column normalization, positivity, and the even activation also give

$$0 \leq g(\mathbf{u}) \leq \max_{1\leq j\leq 3} |u_j|^q.$$

Hence, for independent standard Gaussian G_1, G_2, G_3,

$$\mathbb{E}[g(G_1, G_2, G_3)^2] \leq 3\mathbb{E}[|G_1|^{2q}] = 3(2q-1)!!.$$

Using (74) with $v = 0$ yields

$$\sup_{d\geq 3} \|f_d^\star\|_{L^2(\nu_d)} \leq M, \qquad M := \sqrt{3(2q-1)!!}. \tag{79}$$

Step 4: nonvanishing degree-q spherical-harmonic energy. Let $P_q = P_{\leq q} - P_{\leq q-1}$. Its range is the space $V_{d,q}$ of restrictions of homogeneous harmonic polynomials of degree q, with dimension

$$B(d,q) = \binom{d+q-1}{q} - \binom{d+q-3}{q-2}.$$

Use the Gegenbauer normalization of Ghorbani et al. (2021, Section 5.2), and define

$$Z_{q,d}(\mathbf{u}) := \sqrt{B(d,q)}\, Q_q^{(d)}(\sqrt{d}\, u_1).$$

Then $Z_{q,d} \in V_{d,q}$ and $\|Z_{q,d}\|_{L^2(\nu_d)} = 1$, by Eq. (32) of that paper. Consequently,

$$\|P_q f_d^\star\|^2_{L^2(\nu_d)} \geq |\langle f_d^\star, Z_{q,d}\rangle_{L^2(\nu_d)}|^2. \tag{80}$$

Eq. (43) of Ghorbani et al. (2021) gives, for fixed q, the following coefficientwise convergence of polynomials:

$$\sqrt{B(d,q)}Q_q^{(d)}(\sqrt{d}\, s) \longrightarrow \frac{\mathrm{He}_q(s)}{\sqrt{q!}}. \tag{81}$$

Here He_q is the probabilists' Hermite polynomial, normalized by $\mathbb{E}[\mathrm{He}_q(G_1)^2] = q!$.

For completeness, convergence of all joint moments needed below follows from an exact radial identity. For $\boldsymbol{\beta} \in \mathbb{Z}^3_{\geq 0}$, writing $j = |\boldsymbol{\beta}|$, the decomposition $\mathbf{X} = R\mathbf{U}$ and (73) imply

$$\mathbb{E}_{\nu_d}\left[U_1^{2\beta_1}U_2^{2\beta_2}U_3^{2\beta_3}\right] = \frac{\prod_{a=1}^3 (2\beta_a - 1)!!}{A_{d,j}} \longrightarrow \prod_{a=1}^3 (2\beta_a - 1)!!. \tag{82}$$

We use $(-1)!! = 1$; moments with an odd exponent in any coordinate vanish by reflection symmetry. Since g and the polynomials in (81) have fixed degree, coefficientwise convergence and (82) yield

$$\langle f_d^\star, Z_{q,d}\rangle_{L^2(\nu_d)} \longrightarrow \frac{\mathbb{E}[g(G_1,G_2,G_3)\,\mathrm{He}_q(G_1)]}{\sqrt{q!}}. \tag{83}$$

Hermite orthogonality gives

$$\mathbb{E}[G_1^j\,\mathrm{He}_q(G_1)] = 0 \quad (0 \leq j < q), \qquad \mathbb{E}[G_1^q\,\mathrm{He}_q(G_1)] = q!.$$

Every monomial of g has total degree q. Thus only $b_\star u_1^q$ contributes to the expectation in (83); independence of the Gaussian coordinates gives

$$\mathbb{E}[g(G_1,G_2,G_3)\,\mathrm{He}_q(G_1)] = b_\star q!.$$

Combining this with (80) proves

$$\liminf_{d\to\infty} \|P_q f_d^\star\|^2_{L^2(\nu_d)} \geq q! b_\star^2 > 0. \tag{84}$$

In particular, with

$$E_0 := \frac{q! b_\star^2}{2} > 0,$$

we have $\|P_q f_d^\star\|^2_{L^2(\nu_d)} \geq E_0$ for all sufficiently large d.

Step 5: apply the spherical theorem and conclude. Fix $\eta \in (0, q)$, and suppose $n \leq d^{q-\eta}$. Apply (70) with

$$m = q - 1, \qquad \delta_d = \eta.$$

The integer m is fixed, and $\delta_d^2 \log d = \eta^2 \log d \to \infty$. Since $f_d^\star$ has degree q, $P_{>q-1} f_d^\star = P_q f_d^\star$. The nonnegativity of $R_{K,d}(P_{\leq q-1} f_d^\star; \mathbf{U}_{1:n})$ therefore gives, with probability tending to one,

$$R_{K,d}(f_d^\star; \mathbf{U}_{1:n}) \geq E_d - \epsilon \|f_d^\star\|_{L^2(\nu_d)}\sqrt{E_d}, \qquad E_d := \|P_q f_d^\star\|^2_{L^2(\nu_d)}.$$

Choose the fixed theorem tolerance $\epsilon = \sqrt{E_0}/(2M)$. Equations (79) and (84) imply

$$\epsilon \|f_d^\star\|_{L^2(\nu_d)}\sqrt{E_d} \leq \frac{\sqrt{E_0}\sqrt{E_d}}{2} \leq \frac{E_d}{2}.$$

Consequently,

$$\mathbb{P}\left(R_{K,d}(f_d^\star; \mathbf{U}_{1:n}) \geq \frac{E_0}{2}\right) \longrightarrow 1. \tag{85}$$

The exact identity (75), together with $A_{d,q} \geq 1$, transfers this bound to the best possible predictor in the Gaussian sample-centered kernel span. Set

$$\varepsilon_\star := \min\left\{\frac{1}{4}, \frac{E_0}{4}\right\}.$$

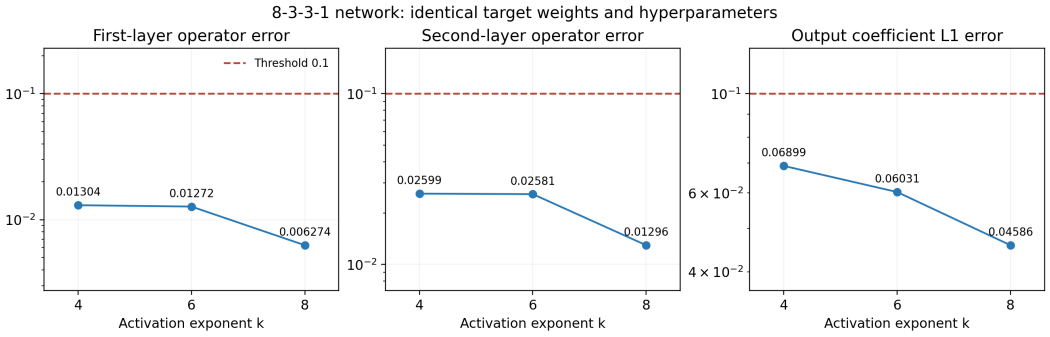


Figure 1: Recovery errors comparison for three activation degrees.

Then every predictor stated above has Gaussian risk at least $2\varepsilon_\star$ on the same event. The event depends only on the sample directions and holds for all coefficient vectors simultaneously, so additional algorithmic randomness does not affect the lower bound.

In particular, the probability of achieving risk at most $\varepsilon_\star$ tends to zero whenever $n \le d^{q-\eta}$. For a learner with success probability at least 0.99 for all sufficiently large d, even a subsequence satisfying this sample-size inequality is impossible. It follows that its worst-case sample complexity is $\Omega_{k,L,\eta}(d^{q-\eta})$ for every fixed $\eta \in (0, q)$. Equivalently, its necessary sample-size exponent satisfies

$$\liminf_{d\to\infty} \frac{\log n(d)}{\log d} \ge q.$$

□

Scope of the conclusion. The hard targets are chosen before the data are sampled and satisfy all of the manuscript's assumptions; no additional highest-degree energy assumption is imposed on the entire target class. The lower bound concerns fixed, sample-centered infinite-width NTK regression, including zero-initialized kernel gradient descent. It does not rule out passive feature-learning algorithms, data-dependent kernels, or active queries. The use of Theorem 3 also requires k, L to remain fixed as $d \to \infty$; the exponent statement alone is not a claim of an $\Omega(d^q)$ lower bound with no exponent slack.

J Experiment Details

We present a numerical illustration of parameter recovery on an $8 \to 3 \to 3 \to 1$ polynomial network with activation $\sigma(t) = t^k$ for $k \in \{4, 6, 8\}$. All configurations use the same target weights and numerical settings, with only the activation degree changed.

The first hidden layer is recovered by column-space estimation followed by ASPIRE, using empirical position and gradient moments constructed from independent Hit-and-Run chains. The treatment of the second hidden layer and output layer differs slightly from the procedure analyzed in the main text: we use multiple Hessian evaluations for the second hidden layer and ordinary least squares for the output weights. Thus, the experiment is primarily intended to validate the core ASPIRE mechanism for recovering hidden directions.

Before evaluation, we align hidden-unit permutations across adjacent layers and resolve first-layer column signs. Figure 2 visualizes the recovered parameters for $k = 4$, while Figure 1 reports the recovery errors for all three activation degrees. The results show accurate recovery across all tested settings, providing a numerical validation of the core ASPIRE procedure.

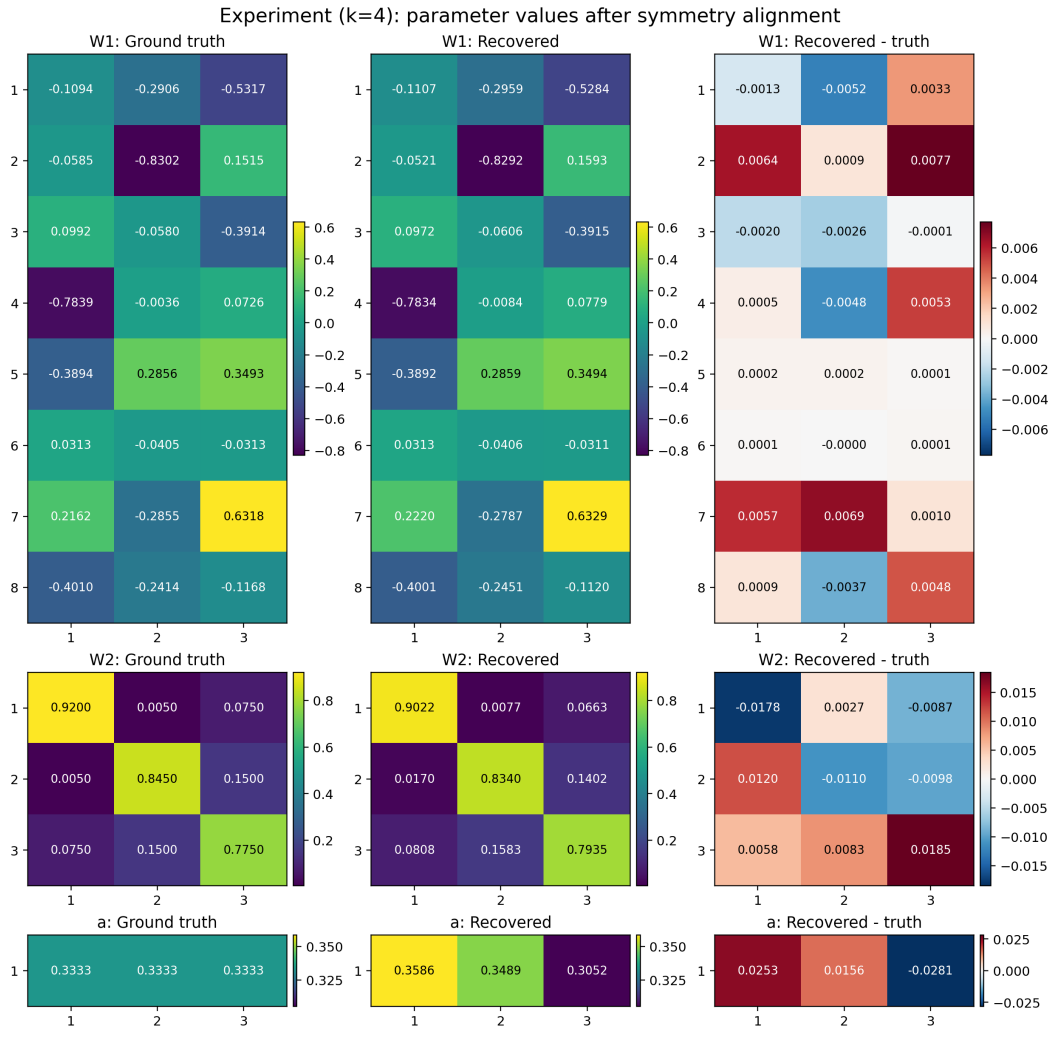


Figure 2: Parameter recovery for the $8 \to 3 \to 3 \to 1$ network with fourth-power activation ($k = 4$).