GENERALIZATION BELOW THE EDGE OF STABILITY: THE ROLE OF DATA GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding generalization in overparameterized neural networks hinges on the interplay between the data geometry, neural architecture, and training dynamics. In this paper, we theoretically explore how data geometry controls this implicit bias. This paper presents theoretical results for overparametrized two-layer ReLU networks trained *below the edge of stability*. First, for data distributions supported on a mixture of low-dimensional balls, we derive generalization bounds that provably adapt to the intrinsic dimension. Second, for a family of isotropic distributions that vary in how strongly probability mass concentrates toward the unit sphere, we derive a spectrum of bounds showing that rates deteriorate as the mass concentrates toward the sphere. These results instantiate a unifying principle: When the data is harder to "shatter" with respect to the activation thresholds of the ReLU neurons, gradient descent tends to learn representations that capture shared patterns and thus finds solutions that generalize well. On the other hand, for data that is easily shattered (e.g., data supported on the sphere) gradient descent favors memorization. Our theoretical results consolidate disparate empirical findings that have appeared in the literature.

1 INTRODUCTION

How does gradient descent (GD) discover well-generalized representations in overparameterized neural networks, when these models possess more than enough capacity to simply memorize the training data? Conventional wisdom in statistical learning attributes this to explicit capacity control via regularization such as weight decay. However, this view has been profoundly challenged by empirical findings that neural networks generalize remarkably even without explicit regularizers, yet can also fit randomly labeled data with ease, even with strong regularization (Zhang et al., 2017).

This paradox forces a critical re-evaluation of how we should characterize the effective capacity of neural networks, which appears to be implicitly constrained by the optimizer's preferences (Zhang et al., 2017; Arpit et al., 2017). A powerful lens for examining this *implicit regularization* is to inspect the properties of solutions to which GD can stably converge, since these stable points are the only solutions that the training dynamics can practically reach and maintain. This direction is strongly motivated by the empirical discovery of the "*Edge of Stability*" (EoS) regime, where GD with large learning rates operates in a critical regime where the step size is balanced by the local loss curvature (Cohen et al., 2020). This observation is further supported by theoretical analyses of GD's dynamical stability (Wu et al., 2018; Nar & Sastry, 2018; Mulayoff et al., 2021; Nacson et al., 2023; Damian et al., 2024), confirming that the curvature constraint imposed by stability provides a tractable proxy for this implicit regularization.

While the EoS regime offers a valuable proxy, a fundamental question remains: how precisely does this stability-induced regularization lead to generalization? Recent breakthroughs have established that for two-layer ReLU networks, this implicit regularizationthis implicit regularization acts like a data-dependent penalty on the network's complexity. Technically, this is captured by a weighted path norm, where the weight function is determined by the training dataset itself (Liang et al., 2025; Qiao et al., 2024; Nacson et al., 2023; Mulayoff et al., 2021). This resulting *data-dependent regularity* provides an ideal theoretical microcosm to probe how data geometry governs effective capactity (Arpit et al., 2017). For example, for uniform distribution on a ball, it implies generalization but also a curse of dimensionality (Liang et al., 2025). However, this prediction of a curse is at odds

with the empirical success of deep learning. This contradiction forces the question: how can we predict which data geometries will generalize well under implicit regularization, and which will not?

Contributions. In this work, we argue that the effectiveness of this data-dependent regularity is governed by a single, unifying principle, which we term *data shatterability*

The less shatterable the data geometry, the stronger the implicit regularization of EoS becomes.

We formalize this principle with the following theoretical results:

- **Provable Adaptation to Low-dimensionalty.** Assuming that the input feature only support on a mixture of m-dimensional subspaces in $\mathbb{R}^d$ with $m < d$, we prove the generalization bound of $\tilde{O}(n^{-\frac{1}{2m+4}})$ that adapts to the *intrinsic dimension* as Theorem 3.2, where $\tilde{O}$ hides the constants that mildly depends on the number of subspaces (at most linear), and the logarithmic factor of the probability term. Our empirical results on sythetic data also support the theoretical analysis.
- **A Spectrum of Generalization on Isotropic Data.** We show that generalization performance smoothly degrades as data concentrates near the boundary of its support. We provide precise upper bounds that depend on the dimension d and a concentration parameter α (Theorem 3.5), as well as lower bounds (Theorem 3.6). This analysis culminates in the limit of extreme boundary concentration (data on a sphere), where we provide a concrete construction of a network that perfectly interpolates any dataset at the BEoS regime, see Theorem 3.7. In particular, the "neural shattering" phenomenon, identified by Liang et al. (2025) for the uniform ball distribution, represents one special point of the broader generalization spectrum we uncover.

Our theoretical results for both subspace mixtures and isotropic distributions demonstrate how the principle of data shatterability operates in two distinct ways.

On the one hand, for data residing on a mixture of subspaces, its low-dimensional nature inherently limits its shatterability. We demonstrate that the network's complex decision boundaries, formed by combinations of half-spaces, are fundamentally constrained by the data's intrinsic low-dimensional structure. For example, when data lies on a line within $\mathbb{R}^d$, a ReLUs' complex hyperplane boundaries reduce to a series of knots and entire complexity is defined by the locations and magnitudes of these knots. The cornerstone of our proof is showing that the stability induced, data-dependent implicit regularization is adaptive to this nature.

On the other hand, in the isotropic case, shatterability is governed by the data's radial concentration. As more mass concentrates towards the spherical shell, the data geometry becomes more shatterable. This is because the more the data is concentrated near the sphere, the more non-overlapping caps the network can create across the sphere, while keeping the mass of data inside each cap constant. This allows the network to partition the data into a large number of disjoint, sparsely populated regions, a key feature of high shatterability, and dedicate different neurons to memorizing the labels within each region. Our theoretical analysis and empirical observation confirms that such solutions can be dynamically stable, meaning they can be favored by gradient descent when the data geometry permits.

From a representation learning perspective, our constructions for the lower bound and flat interpolation (Theorems 3.6 and 3.7) demonstrate that data shatterability directly governs the model's learning strategy. High shatterability enables the formation of "memorizing neurons" that activate on only a few examples. Conversely, low shatterability makes this memorization strategy difficult, implicitly forcing the optimizer to learn robust, shared representations (see Figure 2). We thus propose data shatterability as a foundational principle explaining the representations favored by gradient descent, bridging a critical gap between learning theory and practice.

Related work and novelty. We build upon a recent line of work (Qiao et al., 2024; Liang et al., 2025) that theoretically study the generalizaton of neural networks in Edge-of-Stability regime (Cohen et al., 2020) from a function space perspective (Mulayoff et al., 2021; Nacson et al., 2023). We add to this literature a two brand new dimensions: how a concentration coefficient of data

distribution affects generalization by EoS and how gradient descent with large step-size adapts to low-dimensional structures in data.

More broaderly, our work is inspired by the seminal work of Zhang et al. (2017) on "rethinking generalization". Our results provide new theoretical justification that rigorously explains several curious phenomena (such as why real data are harder to overfit than random Gaussian data) reported therein. Compared to other existing work inspired by Zhang et al. (2017), e.g., those that study the implicit bias of gradient descent from various alternative angles (dynamics (Arora et al., 2019; Mei et al., 2019; Jin & Montúfar, 2023), algorithmic stability (Hardt et al., 2016), large-margin (Soudry et al., 2018), benign overfitting (Joshi et al., 2024; Kornowski et al., 2024)), our work has more end-to-end generalization bounds and requires (morally, since the settings are not all compatible) weaker assumptions. On the practical front, we provide new theoretical insight into how "mix-up" data augmentation (Zhang et al., 2018; 2021) and "activation-based pruning" (Hu et al., 2016; Ganguli & Chong, 2024) work. A more detailed discussion of the related work and the implications of our results can be found in Appendix B.

Disclaimers and limitations. It is important to note that our reseacher focus on gradient descent training of overparameterized neural networks in the feature-learning regime (a.k.a "rich" learning regime). Formally analyzing the gradient dynamics is notoriously difficult once training enters the feature learning regime. This challenge is a core motivation for our work: we sidestep the dynamics and analyze the properties of the set of stable solutions instead. At a cost, our theoretical bounds do not apply to the early-phase of training. However, the benefit is that the function space characterization derived from this stability condition allows for a width-agnostic analysis, which we leverage for our generalization upper bounds that apply to networks of *arbitrary* finite width. Only one requirement for sufficiently large width is invoked for our negative results. Specifically, for the lower-bound construction of "hard-to-learn" functions (Theorem 3.6) and the existence of stable interpolating solutions (Theorem 3.7), the network width K must be at least on the order of the sample size n.

Our theoretical results are derived for two-layer fully-connected ReLU networks. While this architecture is a cornerstone for theoretical analysis, modern deep learning employs a much wider array of designs. Extending our analysis to deeper networks, or architectures with specific inductive biases like local connectivity (e.g., CNNs), is a significant undertaking left for future work.

2 Preliminaries and Notations

Neural network, data, and loss. We consider two-layer ReLU networks

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \phi(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k) + \beta, \quad \phi(z) = \max\{z, 0\}, \tag{1}$$

with parameters $\boldsymbol{\theta} = \{(v_k, \boldsymbol{w}_k, b_k)\}_{k=1}^{K} \cup \{\beta\} \in \mathbb{R}^{(d+2)K+1}$. Let $\boldsymbol{\Theta}$ be the parameter set of such $\boldsymbol{\theta}$ for arbitrary $K \in \mathbb{R}$. We also assume $\boldsymbol{w}_k \neq \mathbf{0}$ for all k in this form, otherwise we may absorb it into the output bias β. Given data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ with $\boldsymbol{x}_i$ in a bounded domain $\Omega \subset \mathbb{R}^d$ with $d > 1$, the training loss is $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i\big)^2$. We assume $\|\boldsymbol{x}_i\| \leq R$ and $|y_i| \leq D$ for all i.

"Edge of Stability" regime. Empirical and theoretical research (Cohen et al., 2020; Damian et al., 2024) has established the critical role of the linear stability threshold in the dynamics of gradient descent. In GD's trajectory, there is an initial phase of "progressive sharpening" where $\lambda_{\max}(\nabla^2 \mathcal{L}(\boldsymbol{\theta}_t))$ increases. This continues until the GD process approaches the "Edge of Stability", a state where $\lambda_{\max}(\nabla^2 \mathcal{L}(\theta_t)) \approx 2/\eta$, where η is the learning rate. In this paper, all the GD refers to vanilla GD with learning rate η.

Definition 2.1 (Below Edge of Stability (Qiao et al., 2024, Definition 2.3))**.** *We define the trajectory of parameters $\{\boldsymbol{\theta}_t\}_{t=1,2,\cdots}$ generated by gradient descent with a learning rate η as* Below-Edge-of-Stability (BEoS) *if there exists a time $t^* > 0$ such that for all $t \geq t^*$, $\lambda_{\max}(\nabla^2 \mathcal{L}(\boldsymbol{\theta}_t)) \leq \frac{2}{\eta}$. Any parameter state $\boldsymbol{\theta}_t$ with $t \geq t^*$ is thereby referred to as a BEoS solution.*

This condition applies to any twice-differentiable solution found by GD, even when the optimization process does not converge to a local or global minimum. Moreover, BEoS is empirically verified

to hold during both the "progressive sharpening" phase and the subsequent oscillatory phase at the EoS.

Our work aims to analyze the generalization properties of any solutions that satisfy the BEoS condition (Definition 2.1). The set of solutions defined as:

$$\mathbf{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D}) := \left\{ \boldsymbol{\theta} \;\middle|\; \lambda_{\max}(\nabla^2 \mathcal{L}(\boldsymbol{\theta})) \leq \frac{2}{\eta} \right\}. \tag{2}$$

Data-dependent weighted path norm. Given a weight function $g : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$, where $\mathbb{S}^{d-1} := \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\| = 1\}$, the *g-weighted path norm* of a neural network $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \phi(\boldsymbol{w}_k^\top \boldsymbol{x} - b_k) + \beta$ is defined to be

$$\|f_{\boldsymbol{\theta}}\|_{\mathrm{path},g} = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2 \cdot g\left(\frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_2}, \frac{b_k}{\|\boldsymbol{w}_k\|_2} \right). \tag{3}$$

The link between the EoS regime and weighted path norm constrain is presented in the following data-dependent weight function (Liang et al., 2025; Nacson et al., 2023; Mulayoff et al., 2021). Fix a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, we consider a weight function $g_{\mathcal{D}} : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ defined by $g_{\mathcal{D}}(\boldsymbol{u}, t) := \min\{\tilde{g}_{\mathcal{D}}(\boldsymbol{u}, t), \tilde{g}_{\mathcal{D}}(-\boldsymbol{u}, -t)\}$, where

$$\tilde{g}_{\mathcal{D}}(\boldsymbol{u}, t) := \mathbb{P}(\boldsymbol{X}^\top \boldsymbol{u} > t)^2 \cdot \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{u} - t \mid \boldsymbol{X}^\top \boldsymbol{u} > t] \cdot \sqrt{1 + \|\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{X}^\top \boldsymbol{u} > t]\|^2}. \tag{4}$$

Here, $\boldsymbol{X}$ is a random vector drawn uniformly at random from the training examples $\{\boldsymbol{x}_i\}_{i=1}^n$. Specifically, we may also consider its population level $g_{\mathcal{P}}$ by viewing $\boldsymbol{X}$ as a random variable

Proposition 2.2. *For any* $\boldsymbol{\theta} \in \mathbf{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})$, $\|f_{\boldsymbol{\theta}}\|_{\mathrm{path},g_{\mathcal{D}}} \leq \frac{1}{\eta} - \frac{1}{2} + (R+1)\sqrt{2\mathcal{L}(\boldsymbol{\theta})}$.

The proof of this proposition refers to (Liang et al., 2025, Corollary 3.3). The non-parametric characterization of stable minima via bounded weighted variation norm refers to (Liang et al., 2025; Nacson et al., 2023).

Supervised statistical learning and generalization gap. We consider a supervised learning problem where i.i.d. samples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are drawn from an unknown distribution $\mathcal{P}$. In this paper, we assume the feature space is a compact subset of Euclidean space, $\Omega \subset \mathbb{R}^d$, the label space is $\mathbb{R}$, and the data is supported on $\Omega \times [-D, D]$. We use the squared loss, defined as $\ell(f, \boldsymbol{x}, y) = \frac{1}{2}(f(\boldsymbol{x}) - y)^2$. The performance of a predictor f is measured by its population risk $R_{\mathcal{P}}(f) = \mathbb{E}_{(\boldsymbol{X},Y)\sim\mathcal{P}}\, \ell(f, \boldsymbol{X}, Y)$, while we optimize the empirical risk $\widehat{R}_{\mathcal{D}}(f) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}} \ell(f, \boldsymbol{x}_i, y_i)$. The difference between these two quantities is the generalization gap $\mathrm{Gap}_{\mathcal{P}}(f; \mathcal{D}) = |R_{\mathcal{P}}(f) - \widehat{R}_{\mathcal{D}}(f)|$. Our work focuses on the hypothesis classes the BEoS class $\mathbf{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})$ and the bounded weighted-path norm class $\mathbf{\Theta}_g(\Omega; M, C)$,

$$\mathbf{\Theta}_g(\Omega; M, C) = \left\{ \boldsymbol{\theta} \in \mathbf{\Theta} \;\middle|\; \|f_{\boldsymbol{\theta}}\|_{\Omega}\|_{L^\infty} \leq M,\ \|f_{\boldsymbol{\theta}}\|_{\mathrm{path},g} \leq C \right\}. \tag{5}$$

where g can be the weight function $g_{\mathcal{D}}$ associated to the empirical distribution $\mathcal{D}$ or the weight function $g_{\mathcal{P}}$ associated to the population distribution $\mathcal{P}$, see Section E for more details.

3 Main Results

In this section, we present our main theoretical results concerning the properties of stable solutions found by gradient descent. Section 3.1 establishes a generalization bound for data exhibiting intrinsic low-dimensional structure. Section 3.2 then derives a spectrum of generalization bounds for a tunable family of isotropic distributions, which connects the data's radial mass concentration to generalization performance. Finally, Section 3.3 investigates the behavior of stable solutions in the limiting case where data is supported entirely on the unit sphere. All the detail proofs are deffered to the appendix.

3.1 Provable Adaptation to Intrinsic Low-Dimensionality

We begin with our main positive result, considering the case where data possesses an underlying low-dimensional structure, a common feature of real-world datasets. We show that in this context, the generalization performance of stable networks adapts to this intrinsic dimension rather than the ambient one.

Assumption 3.1 (Mixture of Low-Dimensional Balls). *Let $\{V_j\}_{j=1}^J$ be a finite collection of J distinct m-dimensional (affine) linear subspaces within $\mathbb{R}^d$. Let $\mathcal{P}$ be a joint distribution over $\mathbb{R}^d \times \mathbb{R}$. The marginal distribution of the features $\boldsymbol{x}$ under $\mathcal{P}$, denoted $\mathcal{P}_{\boldsymbol{X}}$, is a mixture distribution given by*

$$\mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{j=1}^{J} p_j \mathcal{P}_{\boldsymbol{X},j}(\boldsymbol{x}), \quad \mathcal{P}_{\boldsymbol{X},j}(\boldsymbol{x}) = \mathcal{P}_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{x} \in V_j), \tag{6}$$

where $p_j > 0$ are the mixture probabilities $\mathbb{P}(\boldsymbol{x} \in V_j)$ satisfying $\sum_{j=1}^J p_j = 1$. Each component distribution $\mathcal{P}_j$ is the uniform distribution on the unit disc $\mathbb{B}_1^{V_j} := \{\boldsymbol{x} \in V_j : \|\boldsymbol{x}\|_2 \le 1\}$. The corresponding labels y are generated from a conditional distribution $\mathcal{P}(y|\boldsymbol{x})$ and are assumed to be bounded, i.e., $|y| \le D$ for some constant $D > 0$. Similarly, we define $\mathcal{P}_j(\boldsymbol{x}, y) = \mathcal{P}(\boldsymbol{x}, y \mid \boldsymbol{x} \in V_j)$.

Under the structural conditions of Assumption 3.1, we establish a generalization bound whose sample complexity depends on the intrinsic dimension m.

Theorem 3.2 (Generalization Bound for Mixture Models). *Let the data distribution $\mathcal{P}$ be as defined in Assumption 3.1. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples drawn from $\mathcal{P}$. Then, with probability at least $1 - \delta$,*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})} \mathrm{Gap}_{\mathcal{P}}(f_{\boldsymbol{\theta}}; \mathcal{D}) \lessapprox_d \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{m}{m^2+4m+3}} M^2 J^{\frac{4}{m}} n^{-\frac{1}{2m+4}} + M^2 J \sqrt{\frac{1}{2n}}. \tag{7}$$

where $M := \max\{D, \|f_{\boldsymbol{\theta}}|_{\mathbb{B}_1^V}\|_{L^\infty}, 1\}$ and $\lessapprox_d$ hides constants (which could depend on d) and logarithmic factors in J/δ and n.

The proof appears in Appendix F. The core strategy is to decompose the problem by analyzing the network's behavior on each subspace V_j individually and then aggregating the results. If we restrict the network to a single m-dimensional subspace V_j, a neuron's activation is governed not by its full weight vector $\boldsymbol{w}_k$, but solely by $\mathrm{proj}_{V_j} \boldsymbol{w}_k$, since the component of $\boldsymbol{w}_k$ that is orthogonal to V_j is "invisible to the data on V_j". However, a critical question then emerges: we know EoS provides a constraint on the global data-dependent regularity, which involves the full weights w_k, but how does this global constraint translate to local constraint on each individual subspace? To resolve this question, we prove that the global weight function g dominates the local weight function g_j that is only determined by the data points on V_j (Lemma F.3). This formally establishes that the implicit regularization is adaptive to the data's geometry, allowing us to derive a final bound that scales with the intrinsic dimension m.

3.2 A Spectrum of Generalization on Isotropic Distributions

To explore the transition from a generalizing to a memorizing regime, we now analyze a family of isotropic distributions parameterized by a term that controls the concentration of data mass near the boundary of the unit ball. This allows for a precise characterization of how generalization degrades as data points become more radially exposed.

Definition 3.3 (Isotropic Beta-radial distributions). *Let $\boldsymbol{X}$ be a d-dimensional random vector in $\mathbb{R}^d$. For any $\alpha \in (0, \infty)$, the isotropic α-powered-radial distribution is defined by the generation process*

$$\boldsymbol{X} = h(R)\boldsymbol{U} \sim \mathcal{P}_{\boldsymbol{X}}(\alpha), \tag{8}$$

where $R \sim \mathrm{Uniform}[0, 1]$ is a random variable drawn from a continuous uniform distribution on the interval $[0, 1]$, $\boldsymbol{U} \sim \mathrm{Uniform}(\mathbb{S}^{d-1})$ is a random vector drawn uniformly from the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$ and $h(r) = 1 - (1 - r)^{1/\alpha}$ is a radial profile.

Note that as $\alpha \to 0$, the distribution $\mathcal{P}(\alpha)$ will be closer to the uniform distribution on the sphere.

Assumption 3.4. *Fix $\alpha \in (0, \infty)$. Let $\mathcal{P}(\alpha)$ be a joint distribution over $\mathbb{R}^d \times \mathbb{R}$ such that The marginal distribution of the features x under $\mathcal{P}_{\boldsymbol{X}}(\alpha)$. The corresponding labels y are generated from a conditional distribution $\mathcal{P}(y|\boldsymbol{x})$ and are assumed to be bounded, i.e., $|y| \leq D$ for some constant $D > 0$. Similarly, we define $\mathcal{P}_j(\boldsymbol{x}, y) = \mathcal{P}(\boldsymbol{x}, y \mid \boldsymbol{x} \in V_j)$.*

This framework enables the derivation of a generalization upper bound that depends explicitly on the parameter α. The proof of the following theorem can be found in Appendix G.

Theorem 3.5. *Fix a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where each $(\boldsymbol{x}_i, y_i)$ is drawn i.i.d. from $\mathcal{P}(\alpha)$ defined in Assumption 3.4. Then, with probability at least $1 - \delta$, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})$,*

$$\mathrm{Gap}_{\mathcal{P}}(f_{\boldsymbol{\theta}}; \mathcal{D}) \lessapprox_d \begin{cases} \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{\alpha d}{d^2+4d+3}} M^{\frac{2d^2+7\alpha d+6\alpha}{d^2+4\alpha d+3\alpha}} n^{-\frac{\alpha(d+3)}{2(d^2+4\alpha d+3\alpha)}}, & \alpha \geq \frac{3d}{2d-3}; \\ \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{\alpha d}{d^2+4d+3}} M^{\frac{2d^2+7\alpha d+6\alpha}{d^2+4\alpha d+3\alpha}} n^{-\frac{\alpha}{2d+4\alpha}}, & \alpha < \frac{3d}{2d-3}, \end{cases} \tag{9}$$

and for where $M := \max\{D, \|f_{\boldsymbol{\theta}}|_{\mathbb{B}_1^d}\|_{L^\infty}, 1\}$ and $\lessapprox_d$ hides constants (which could depend on d) and logarithmic factors in n and $(1/\delta)$.

To demonstrate the tightness of this result, we establish a corresponding lower bound. For this purpose, we consider the class of neural networks with bounded g-weighted path norm $\boldsymbol{\Theta}_g(\mathbb{B}_1^d; 1, 1)$, where g is the population version of the weighted.

Theorem 3.6 (Generalization Gap Lower Bound)**.** *Let $\mathcal{P}$ denote any joint distribution of $(\boldsymbol{x}, y)$ where the marginal distribution of $\boldsymbol{x}$ is $\mathcal{P}_{\boldsymbol{X}}(\alpha))$ and y is supported on $[-1, 1]$. Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^n$ be a dataset of n i.i.d. samples from $\mathcal{P}$. Let $\widehat{R}_{\mathcal{D}_n}(f)$ be any empirical risk estimator for the true risk $R_{\mathcal{P}}(f) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}}[(f(\boldsymbol{x}) - y)^2]$. Then,*

$$\inf_{\widehat{R}} \sup_{\mathcal{P}} \mathbb{E}_{\mathcal{D}_n} \left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_g(\mathbb{B}_1^d; 1, 1)} \left| R_{\mathcal{P}}(f_{\boldsymbol{\theta}}) - \widehat{R}_{\mathcal{D}_n}(f_{\boldsymbol{\theta}}) \right| \right] \gtrsim_{d,\alpha} n^{-\frac{2\alpha}{d-1+2\alpha}}.$$

Crucially, the proof is constructive and, as detailed in Appendix H, leverages the "neural shattering" observation found by (Liang et al., 2025). The method involves constructing a large family of distinct two-layer ReLU networks, and then showing that with high probability, this family contains at least one pair of networks that are indistinguishable to any learning algorithm that only has access to the training data. This construction exhibit how data geometry (more precisely, shatterability) connect to statistical uncertainty: (1) For bounded isotropic distributions with fixed boundary concentration (fixed α) in high dimensions, the input space offers an exponential number of distinct directions; (2) any training set of finite size n can only cover a fraction of these directions, leaving vast regions of the input space unsampled. Our proof exploits this by designing networks whose neurons activate only in localized, disjoint regions near the boundary of the unit ball. The construction ensures that, with high probability, a significant number of these regions contain no training data. By having the two networks differ only in these empty regions, they become identical on every point of the training set, yet remain substantially different in their population risk.

3.3 Flat Interpolation of Spherically Supported Data

The previous analysis indicates that generalization degrades as data concentrates toward a boundary. We now investigate the limiting case of this phenomenon, where the data support is confined to the unit sphere. In this setting, we show that the stability condition at the EoS is insufficient to prevent the network from perfectly interpolating the training data.

Theorem 3.7 (Flat interpolation with width $\leq n$)**.** *Assume that $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is a dataset with $\boldsymbol{x}_i \in \mathbb{S}^{d-1}$ and pairwise distinct inputs, there exists a width $K \leq n$ network of the form (1) that interpolates the dataset and whose Hessian operator norm satisfies*

$$\lambda_{\max}\left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}\right) \leq 1 + \frac{D^2 + 2}{n}. \tag{10}$$

If we remove the output bias parameter β in (1), then $\lambda_{\max}\left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}\right) \leq \frac{D^2+2}{n}$.

Remark 3.8. *Both Theorem 3.7 our work and (Wen et al., 2023) construct explicit memorization networks to illustrate a* negative *result: even under natural "flatness/stability" proxies, a ReLU network can interpolate while failing to generalize. In (Wen et al., 2023), they exhibit trace-minimizing ("flattest") interpolators that perform poorly on the distribution $\mathcal{P}_{\mathrm{xor}}$, which is different from our setting.*

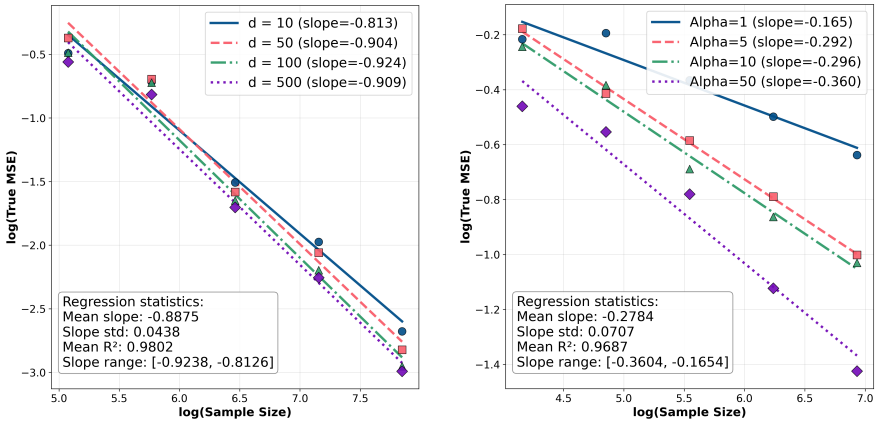
The existence of an interpolating network that remains stable demonstrates that the EoS condition alone does not preclude memorization. This result establishes a clear boundary for the effectiveness of the stability-induced implicit bias, highlighting that its success is critically dependent on the properties of the data's support.

4 Experiments

In this section, we present empirical verification of both our theoretical claims and *proof strategies*.

4.1 Empirical Verification of the Generalization Upper Bounds

We test two predictions of our theory using synthetic data and two-layer ReLU networks of width 1000 trained with MSE loss and vanilla GD with learning rate 0.4 for 20000 epochs. The synthetic training data is produced by fixing a ground-truth function f (ReLU networks or quadratic functions) to noisy labels $y_i = f(\boldsymbol{x}_i) + \xi_i$, where ξ_i is an i.i.d Gaussian noise. Generalization gap is measured by the *true MSE* $\mathbb{E}_{\mathcal{D}}[(\hat{f}(\boldsymbol{X}) - f(\boldsymbol{X}))^2]$ on the training set. In other words, this measures the resistence to memorize noise. Theory predicts Error $\lesssim n^{-c}$ with a geometry-dependent exponent c, so we plot $\log$(clean MSE) against $\log n$ and estimate the slope by OLS. For each sample size n, we train on n i.i.d. examples and report their true MSE. Each set-up sweep 6 random seeds and take averages. The results are summarized in Figure 1.

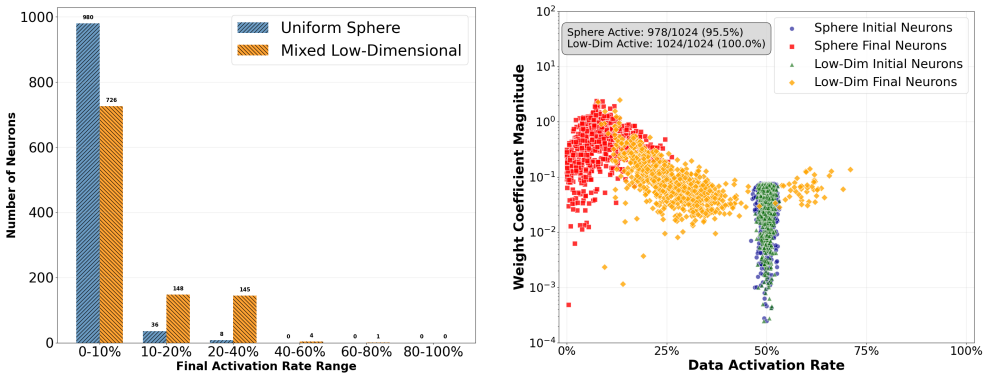


(a) Adaptation to intrinsic dimension (b) Radial concentration parameter α

Figure 1: **How data geometry controls generalization.** *(a)* Union of $J = 20$ lines ($m = 1$) embedded in $\mathbb{R}^d$ with $d \in \{10, 50, 100, 500\}$. The regression slopes remain nearly constant across different d, showing that generalization adapts to intrinsic rather than ambient dimension. *(b)* Fixed ambient dimension $d = 5$ with isotropic Beta-radial distributions (Definition G.1) for $\alpha \in \{1, 5, 10, 50\}$. Larger α yields steeper slopes in the log–log error curve, consistent with improved rates as probability mass concentrates away from the boundary.

4.2 How Data Geometry Affects Representation Learning

We study how data geometry shapes the *representation* selected by GD at the BEoS regime through *data activation rate* of neurons. Given a neuron $v_k\phi(\boldsymbol{w}_k^\top \boldsymbol{x} - b_k)$ in the neuron network, its data activation rate is define to $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\boldsymbol{w}_k^\top \boldsymbol{x}_i > b_k\}$, which is exactly the probability term in the definition of the weight function g in (25). Low data activation rate means the neuron fires on a small portion of the data.



(a) Activation-rate histogram (b) Weight magnitude vs. activation rate

Figure 2: **Neuron activation statistics under different geometries.** (a) On the uniform sphere, most neurons fire on less than 10% of the data, indicating highly specialized ReLUs as we predict in Theorem 3.6 and Theorem 3.7. On the low-dimensional mixture, many neurons fire on 10-40% of the data, reflecting broader feature reuse. (b) Scatter of weight coefficient magnitude versus activation rate. On the sphere, GD produces many low-activation neurons with large coefficients. On the low-dimensional mixture, neurons spread to medium activation rates with moderate coefficients.

To verify it empirically, we compare two input distribution in $\mathbb{R}^{50}$: (i) uniform distribution on a sphere and (ii) a union of lines of 20 lines by training ReLU networks the same recipe and initialization . As a result, the ReLU network trained on the sphere interpolates the noisy label quickly with final true MSE 1.0249 $\approx$ noise level ($\sigma^2 = 1$), while the the ReLU network trained on a union of lines resist to overfitting with final true MSE $0.07 \approx 0$ (more details appears in Appendix C). Notably, the trained representations are presented in Figure 2. In particular, GD empirically finds our lower bound construction below the edge of stability.

4.3 Empirical Evidence for the Data Shatterability Principle

Our theory assumes data supported exactly on a mixture of low-dimensional subspaces. In practice, real datasets are only approximately low-dimensional, as highlighted in the literature on subspace clustering (Vidal et al., 2016; Elhamifar & Vidal, 2013). For instance, MNIST images do not perfectly lie on a union of lines or planes, but still exhibit strong correlations that concentrate them near such structures. Our experiments (Figure 3, more details in Appendix C.2) show that even this approximate structure has a pronounced effect: compared to Gaussian data of the same size, GD on MNIST requires orders of magnitude more iterations before mildly overfitting solutions emerge. This demonstrates that our theoretical prediction is not fragile: generalization benefits from low-dimensional structure across a spectrum.

This experimental result validates not only our main theorems but also the core techniques of our proof. To illustrate this, we introduce **Tukey depth** $\mathrm{depth}_{\mathcal{P}_{\boldsymbol{X}}}(\boldsymbol{x}) := \inf_{\|\boldsymbol{u}\|_2=1} \mathbb{P}(\boldsymbol{u}^\top \boldsymbol{X} \geq \boldsymbol{u}^\top \boldsymbol{x})$, which measures the centrality of a point x by finding the minimum data mass (either population or empirical) on one side of any hyperplane passing through it (Tukey, 1975). Our key claim is that deeper regions of a distribution are hard-to-shatter. For a ReLU ridge to introduce non-linearity, or "wiggleness", within a deep region $\Omega_T := \{\boldsymbol{x} : \mathrm{depth}_{\mathcal{P}_{\boldsymbol{X}}}(\boldsymbol{x}) \geq T\}$, its decision boundary must pass through that region. By the very definition of Tukey depth, the corresponding neuron is then guaranteed to activate on at least a T-fraction of the data. This provides a lower bound on the EoS weight function g for these specific neurons that contribute to the function's nonlinearity in the Ω_T.

Consequently, within this deep region, the stability-induced weighted path norm constraint effectively becomes a more traditional unweighted path norm bound for the part of the network creating local complexity. Such function classes are known to ensure generalization (Parhi & Nowak, 2023; Neyshabur et al., 2015). Outside this core, the BEoS constraint provides no meaningful control for neurons activating there, as g can be vanishingly small, allowing for large unweighted norms that facilitate memorization. Heuristically, deeper regions should generalize better. The right panel of Figure 3 provides a striking visual confirmation of this principle on a real-world dataset.

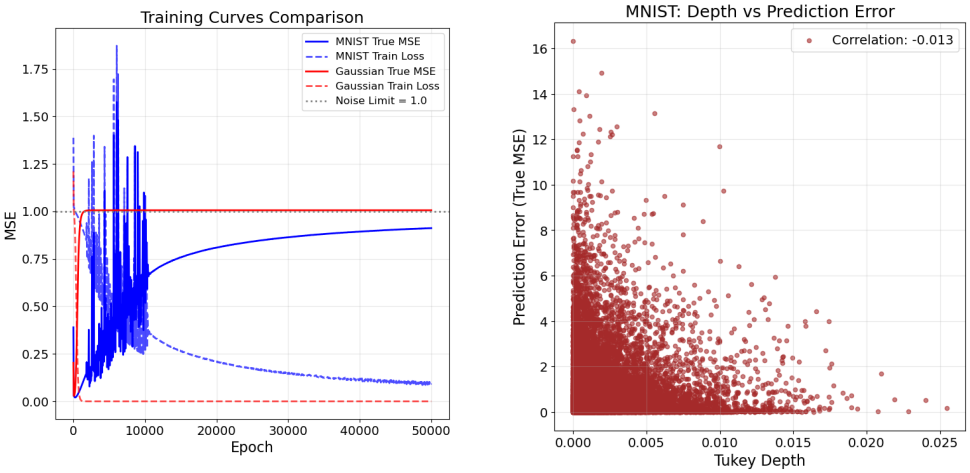


Figure 3: **Data geometry and memorization on MNIST. Left pannel:** Comparison of training curves under the same ground-truth predictor with Gaussian inputs versus MNIST inputs ($n = 30000$). GD on the Gaussian data set quickly interpolates, while MNIST resists overfitting for tens of thousands of steps. **Right pannel:** Prediction error against Tukey depth for MNIST samples. Shallow points (low depth) exhibit larger errors. This regions refers to "highly shatterable region".

Our upper-bound proof technique, which refines the domain decomposition strategy from (Liang et al., 2025), operationalizes this split. We partition the data space into a deep core Ω_T and its shallow complement. We control the generalization error in the core, while conservatively bounding the error in the shallow region by its total probability mass. The final bound arises from optimizing the trade-off in selecting T. Crucially, this strategy is consistent with our lower-bound construction, as well as the "neural shattering" phenomenon analyzed in (Liang et al., 2025). The "hard-to-learn" functions in Construction H.4 are built with neurons that activate exclusively in the low-depth regions. This shows that surrendering the shallow region in the upper bound is not a mere mathematical trick, but reflects a fundamental characteristic of the BEoS regime. In the extreme case of data on a sphere, where all points have zero depth, this technique becomes inapplicable—which aligns perfectly with our finding that stable interpolation is possible in that setting.

5 Discussion and Further Questions

In this work, we present a mechanism explaining *how* data geometry governs the implicit bias of neural networks trained below the Edge of Stability. We introduce the principle of "data shatterability," demonstrating that geometries resistant to shattering guide gradient descent towards discovering shared, generalizable representations. Conversely, we show that easily shattered geometries, such as data concentrated on a sphere, permit stable solutions that memorize the training data.

Our framework opens several promising avenues for future research. A central question is the connection between shatterability and optimization. The observation that a flip side of being prone to overfitting is often faster optimization leads to a natural hypothesis: are high-shatterability distributions easier to optimize? This, in turn, raises further questions about the role of normalization techniques. For instance, do normalization techniques like Batch Norm accelerate training precisely by enforcing more isotropic, and thus more shatterable, representations at each layer? This line of inquiry extends naturally to deep networks, where hidden layers not only sense the initial data geometry but actively create a new "representation geometry". Can our principles be translated to the understanding of representation geometry? Finally, this framework may offer a new lens to understand architectural inductive biases. For example, do CNNs generalize well precisely because their local receptive fields impose an architectural constraint that inherently reduces the model's ability to shatter the data, forcing it to learn local, reusable features? Answering such questions, alongside developing a quantifiable metric for shatterability, remains a key direction.

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International conference on machine learning*, pp. 247–257. PMLR, 2022.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024. PMLR, 2022.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 233–242. PMLR, 2017.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Alexander Cloninger and Timo Klock. A deep network construction that adapts to intrinsic dimensionality beyond the domain. *Neural Networks*, 141:404–419, 2021.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sIE2rI3ZPs.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2024.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

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

Tushar Ganguli and Edwin K. P. Chong. Activation-based pruning of neural networks. *Algorithms*, 17(1):48, 2024.

Boris Hanin and David Rolnick. Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, 2019a. URL https://api.semanticscholar.org/CorpusID:59291990.

Boris Hanin and David Rolnick. Deep ReLU networks have surprisingly few activation patterns. In *Advances in Neural Information Processing Systems*, volume 32, 2019b.

Boris Hanin, Ryan Jeong, and David Rolnick. Deep relu networks preserve expected length. *ArXiv*, abs/2102.10492, 2021. URL https://api.semanticscholar.org/CorpusID:231986303.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

David Haussler. Decision-theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint*, 2016. URL https://arxiv.org/abs/1607.03250.

Hui Jin and Guido Montúfar. Implicit bias of gradient descent for mean squared error regression with two-layer wide neural networks. *Journal of Machine Learning Research*, 24(137):1–97, 2023.

Nirmit Joshi, Gal Vardi, and Nathan Srebro. Noisy interpolation learning with shallow univariate ReLU networks. In *International Conference on Learning Representations*, 2024.

Michael Kohler, Adam Krzyżak, and Sophie Langer. Estimation of a function of low local dimensionality by deep neural networks. *IEEE transactions on information theory*, 68(6):4032–4042, 2022.

Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. *Advances in neural information processing systems*, 33:2625–2638, 2020.

Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From tempered to benign overfitting in relu neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 888–896, Naha, Okinawa, Japan, 2019. PMLR. URL https://proceedings.mlr.press/v89/liang19a.html.

Tongtong Liang, Dan Qiao, Yu-Xiang Wang, and Rahul Parhi. Stable minima of relu neural networks suffer from the curse of dimensionality: The neural shattering phenomenon, 2025. URL https://arxiv.org/abs/2506.20779.

Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pp. 2388–2464. PMLR, 2019.

Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Neural Information Processing Systems*, 2014. URL https://api.semanticscholar.org/CorpusID:5941770.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34:17749–17761, 2021.

Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *International Conference on Learning Representations*, 2023.

Kamil Nar and Shankar Sastry. Step size matters in deep learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1376–1401, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Neyshabur15.html.

Rahul Parhi and Robert D. Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

Rahul Parhi and Robert D. Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 69(2):1125–1139, 2023.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *Advances in Neural Information Processing Systems 34*, volume 34. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/995f5e03890b029865f402e83a81c29d-Abstract.html.

Tomaso Poggio and Qianli Liao. Theory ii: Landscape of the empirical risk in deep learning. *arXiv preprint arXiv:1703.09833*, 2017.

Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94163–94208, 2024.

Thiago Serra, Christian Tjandraatmadja, and Srikumar Ramalingam. Bounding and counting linear regions of deep neural networks. *ArXiv*, abs/1711.02114, 2017. URL https://api.semanticscholar.org/CorpusID:34019680.

Jonathan W. Siegel and Jinchao Xu. Characterization of the variation spaces corresponding to shallow neural networks. *Constructive Approximation*, pp. 1–24, 2023.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018.

Saket Tiwari and George Konidaris. Effects of data geometry in early deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25010–25023, 2022.

John W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pp. 523–531, Vancouver, 1975. Canadian Mathematical Congress.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

René Vidal, Yi Ma, and Shankar Sastry. *Generalized Principal Component Analysis*. Springer, 2016.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, pp. 37656–37684. PMLR, 2023.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Linjun Zhang, Kenji Kawaguchi, Amirata Ghorbani, Zhun Deng, and James Zou. How does mixup help with robustness and generalization? In *International Conference on Learning Representations (ICLR)*, 2021.

Xiao Zhang and Dongrui Wu. Empirical studies on the properties of linear regions in deep neural networks. *ArXiv*, abs/2001.01072, 2020. URL https://api.semanticscholar.org/CorpusID:209862298.

CONTENTS

A The Use of Large Language Models (LLMs)

In accordance with the ICLR 2026 policy on the responsible use of LLMs, we disclose the following. We employed commercial LLM services during manuscript preparation. Specifically, we used Gemini 2.5 Pro, ChatGPT 5, and DeepSeek to assist with language polishing, literature search, and consistency checks of theoretical derivations. We further used Claude 4 and Cursor to help generate experimental code templates. Importantly, all research ideas, theoretical results, and proof strategies originated entirely from the authors. The LLMs were used solely as productivity aids and did not contribute novel scientific content.

B More Related Works

How we "rethink" generalization. Our shatterability principle provides a theoretical account of the discrepancy noted by Zhang et al. (2017): networks fit Gaussian noise much faster than real images with random labels. Gaussian inputs concentrate on a thin spherical shell and are highly shatterable, while CIFAR-10 exhibits unknown low-dimensional structure that resists shattering. Strong generalization arises in practice because gradient descent implicitly exploits this non-shatterable geometry of the real world data. We conduct a similar experiment from the perspective of generalization in Section 4.3 (see Figure 3).

Revisit data augmentation. Mixup forms convex combinations of inputs and labels and encourages approximately linear predictions along these segments (Zhang et al., 2018). The added in-between samples penalize solutions that memorize isolated points with sharply varying piecewise-linear behavior. For example, on spherical-like data that ReLU units can easily shatter, such memorization incurs high loss on the mixed samples, which suppresses shattering-type separators. Prior work mostly views Mixup as a data-dependent regularizer that improves generalization and robustness (Zhang et al., 2021). Our analysis complements this view by tracing the effect to the implicit bias of gradient descent near the edge of stability and by linking the gains to a reduction in data shatterability induced by interpolation in low-density regions.

Activation-based network pruning. Empirical works have shown that pruning strategies based on neuron activation frequency, such as removing neurons with low activation counts, can even improve the test performance after retraining (Hu et al., 2016; Ganguli & Chong, 2024). This coincide with our theory: such rare-firing neurons may be harmful to generalization and pruning these neurons help models to learn more generalizable features.

Subspace and manifold hypothesis. A common modeling assumption in high-dimensional learning is that data lies on or near one or several low-dimensional subspaces embedded in the ambient space, especially in image datasets where pixel values are constrained by geometric structure and are well-approximated by local subspaces or unions of subspaces (Vidal et al., 2016). In particular, results in sparse representation and subspace clustering demonstrate that such structures enable efficient recovery and segmentation of high-dimensional data into their intrinsic subspaces (Elhamifar & Vidal, 2013). This also extends to a more general framework of the manifold hypothesis (Fefferman et al., 2016).

Capacity of neural networks. The subspace and manifold hypotheses have important implications for the capacity and generalization of neural networks. When data lies near low-dimensional sub-

spaces and manifolds, networks can achieve expressive power with significantly fewer parameters, as the complexity of the function to be learned is effectively constrained by the subspace dimension rather than the ambient dimension (Poggio & Liao, 2017; Cloninger & Klock, 2021; Kohler et al., 2022). However, these results focus only on expressivity and the existence of neural networks to learn efficiently on this data.

Interpolation, Benign overfitting and data geometry. Benign-overfitting (Bartlett et al., 2020) studies the curious phenomenon that one can interpolate noisy labels (i.e., 0 training loss) while consistently learn (excess risk $\to 0$ as n gets larger). Joshi et al. (2024) establishes that overfitting in ReLU Networks is not benign in general, but it could become more benign as the input dimension grows (Kornowski et al., 2024) in the isotropic Gaussian data case. Our results suggest that such conclusion may be fragile under *low-dimensional or structured* input distributions. On a positive note, our results suggest that in these cases, generalization may follow from edge-of-stability, which applies without requiring interpolation.

Implicit bias of gradient descent. A rich line of work analyzes the implicit bias of (stochastic) gradient descent (GD), typically through optimization dynamics or limiting kernels (Arora et al., 2019; Mei et al., 2019; Jin & Montúfar, 2023). In contrast, we do not analyze the time evolution per se; we characterize the *function spaces* that GD tends to realize at solutions. Our results highlight a strong dependence on the *input distribution*: even for the same architecture and loss, the induced hypothesis class (and thus generalization) changes as the data geometry changes, complementing prior dynamics-centric views.

Edge of Stability (EoS) and minima stability. The EoS literature primarily seeks to explain when and why training operates near instability and how optimization proceeds there (Cohen et al., 2020; Kong & Tao, 2020; Arora et al., 2022; Ahn et al., 2022; Damian et al., 2024). Central flows offer an alternative viewpoint on optimization trajectories that also emphasizes near-instability behavior (Cohen et al., 2025). Closest to our work is the line on *minima stability* (Ma & Ying, 2021; Mulayoff et al., 2021; Nacson et al., 2023; Wu & Su, 2023; Qiao et al., 2024), which links Hessian spectra and training noise to the geometry of solutions but largely leaves generalization out of scope. We leverage the EoS/minima-stability phenomena to *define* and analyze a data-distribution-aware notion of stability, showing adaptivity to low-dimensional structure and making explicit how distributional geometry shapes which stable minima GD selects.

Flatness vs. generalization. Whether (and which notion of) flatness predicts generalization remains debated. Several works argue sharp minima can still generalize (Dinh et al., 2017), propose information-geometric or Fisher–Rao–based notions (Liang et al., 2019), or develop relative/scale-invariant flatness measures (Petzka et al., 2021). We focus on the *largest* curvature direction (i.e., $\lambda_{\max}$) motivated by EoS/minima-stability. Our results rigorously prove that flatness in this notion does imply generalization (note that there is no contradiction with Dinh et al. (2017)), but it depends on data distribitton.

Linear regions of neural networks. Our research connects to a significant body of work that investigates the shattering capability of neural networks by quantifying their linear activation regions (Hanin & Rolnick, 2019a;b; Hanin et al., 2021; Montúfar et al., 2014; Serra et al., 2017). Other empirical work has meticulously characterized the geometric properties of linear regions shaped by different optimizers (Zhang & Wu, 2020). Particularly, (Tiwari & Konidaris, 2022) consider the how these linear regions intersect with data manifolds. These analyses primarily leverage the number of regions to characterize the expressive power of deep networks, while our work shifts the focus on the generalization performance of shallow networks at the EoS regime.

C Details of Experiments

C.1 Experimental Detials for Section 4.2

Here we provide the full experimental details of the discussion in Section 4.2.

We worked in ambient dimension $d = 50$ with $n = 2000$ training examples. For the *Sphere* condition, samples were drawn uniformly from the unit sphere. For the *Low-dimensional mixture*, we generated data from a mixture of 20 randomly oriented 1-dimensional subspaces uniformly.

Labels were produced by a fixed quadratic teacher function with added Gaussian noise of variance 1.

We trained a two-layer ReLU network with hidden width 1024. All models were trained with GD for 10000 epochs using learning rate 0.4 and gradient clipping at 50. The loss function was the squared error against noisy labels, while generalization performance was evaluated by the *true MSE* against the noiseless teacher. For comparability, both datasets shared the same initialization of parameters.

We monitored (i) training loss and true MSE, (ii) Hessian spectral norm estimated by power iteration on random minibatches, and (iii) neuron-level statistics such as activation rate and coefficient magnitude. The training curves are shown in Figure 4 and $\lambda_{\max}(\nabla_{\boldsymbol{\theta}}\mathcal{L})$-curves are shown in Figure 5.

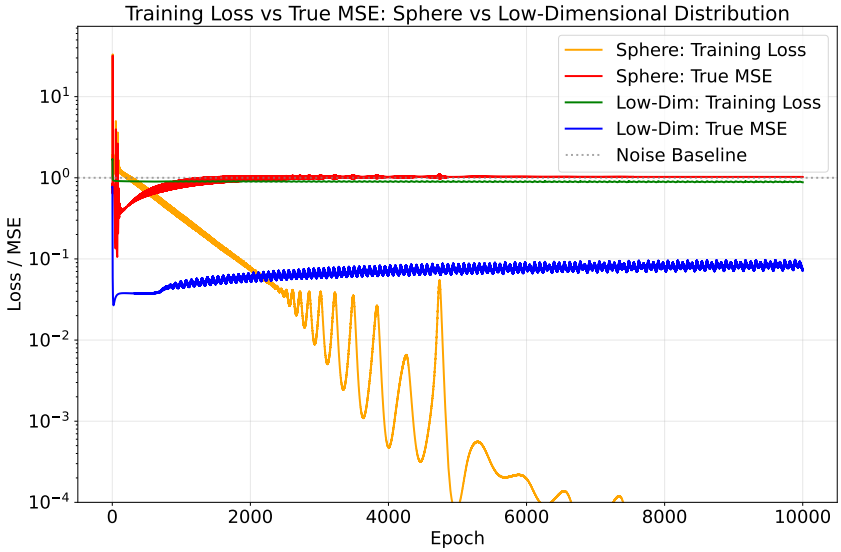


Figure 4: **Training curves on different geometries.** Training loss and clean MSE on Sphere vs. Low-dimensional mixture. We can see GD on sphere interpolate very quickly (before the 2000-th epoch) while the mixed low-dimensional data resist to overfitting.

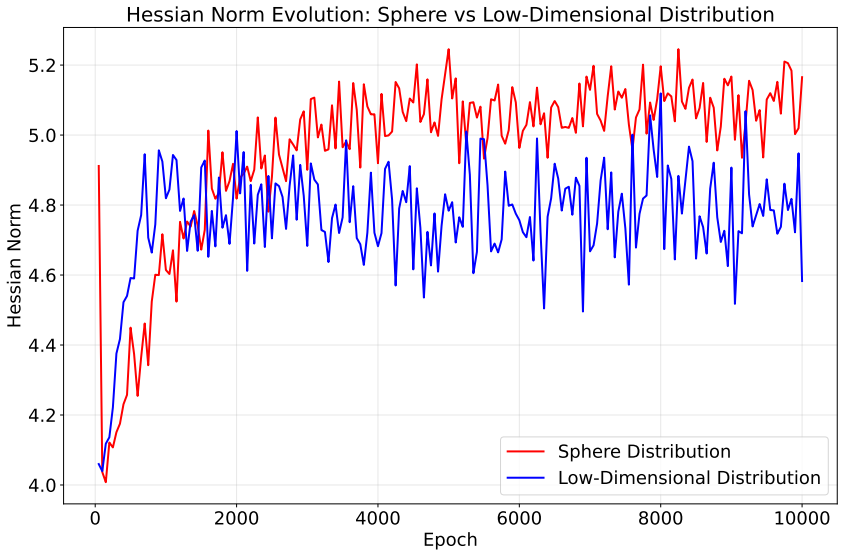


Figure 5: $\lambda_{\max}(\nabla_{\boldsymbol{\theta}}\mathcal{L})$**-curves.** Both of the curves oscillates around $2/\eta = 5$, signaling the edge of stability regime.

C.2 Experimental Details of Section 4.3

We complement the main experiments with a controlled comparison between real data (MNIST) and synthetic Gaussian noise under the same ground-truth function. The goal is to illustrate how the geometry of real-world data affects the speed and nature of memorization by GD.

We fix a ground-truth predictor f (a two-layer ReLU network) and generate noisy labels

$$y_i = f(\boldsymbol{x}_i) + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, 1).$$

We then compare two input distributions of size $n = 30000$:

(i) Gaussian inputs $\boldsymbol{x}_i \sim \mathcal{N}(0, I_d)$ with $d = 784$, and

(ii) MNIST images $\boldsymbol{x}_i \in [0, 1]^{784}$ after normalization by $1/255$.

Both datasets are trained with identical architecture (two-layer ReLU neuron network of 512 neurons), initialization, learning rate $\eta = 0.2$, gradient clip threshold 50.

We track both the empirical training loss and the *true MSE* $\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))^2$, which measures generalization. The horizontal dotted line at $y = 1$ corresponds to the noise variance and represents the interpolation limit.

Figure 6 shows training curves over the first 5000 epochs. On Gaussian inputs, GD rapidly interpolates: the training loss vanishes and the clean MSE rises to the noise limit within a few hundred steps. On MNIST inputs, GD initially decreases both training loss and clean MSE, entering a prolonged BEoS regime where interpolation is resisted. Only after thousands of epochs does the clean MSE start to increase, suggesting that memorization occurs at a much slower rate.

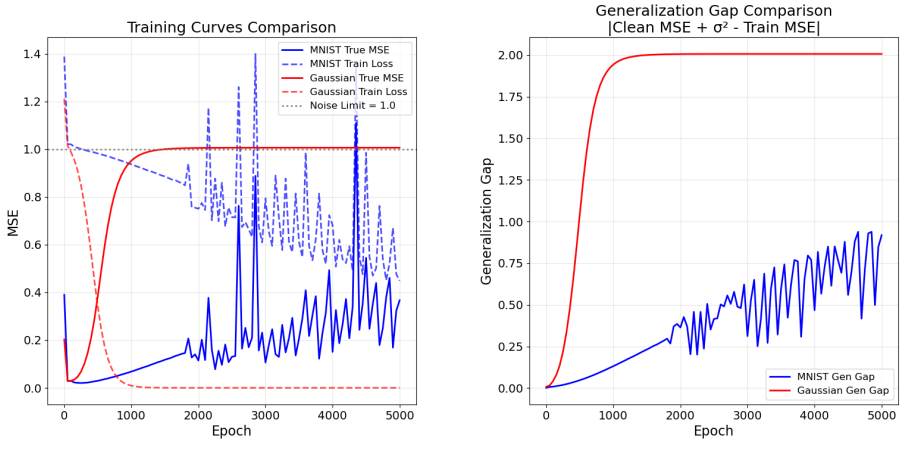
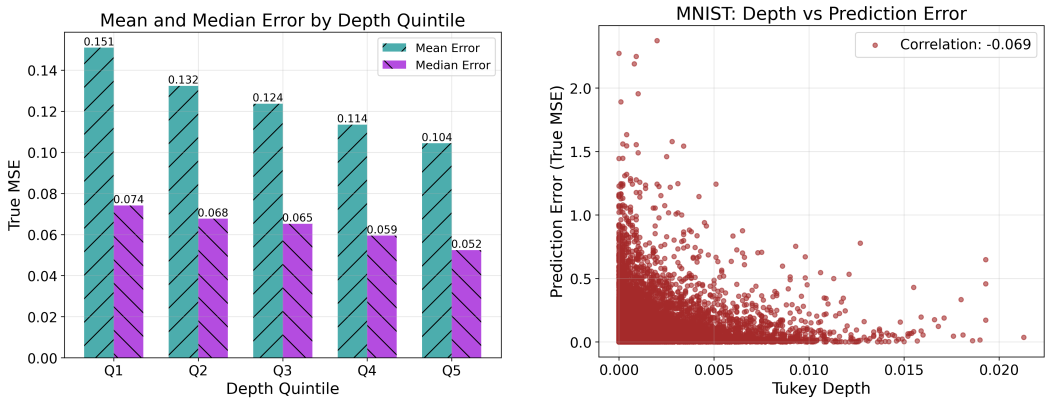


Figure 6: Training curves for Gaussian noise vs. MNIST over the first 5000 epochs. Gaussian quickly interpolates, while MNIST remains in a BEoS regime where clean MSE stays well below the noise level.



(a) Mean/median error by depth quintile (lower is better)

(b) Sample-wise error versus Tukey depth at epoch 5,000

Figure 7: MNIST at 5,000 epochs: deeper points have smaller error. The shallow region produces a long upper tail of errors, consistent with the annulus-interior decomposition used in our upper bounds.

D FUNCTIONAL ANALYSIS OF SHALLOW RELU NETWORKS

D.1 PATH-NORM AND VARIATION SEMI-NORM OF RELU NETWORKS

In this section, we summarize some result in (Parhi & Nowak, 2023) and (Siegel & Xu, 2023).

Definition D.1. *Let $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k\, \phi(\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} - b_k) + \beta$ be a two-layer neural network. The (unweighted) path-norm of $f_{\boldsymbol{\theta}}$ is defined to be*

$$\|f_{\boldsymbol{\theta}}\|_{\mathrm{path}} := \sum_{k=1}^{K} |v_k|\, \|\boldsymbol{w}_k\|_2 \,. \tag{11}$$

Dictionary representation of ReLU networks. By the positive 1-homogeneity of ReLU, each neuron can be rescaled without changing the realized function:

$$v_k\, \phi(\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} - b_k) = a_k\, \phi(\boldsymbol{u}_k^{\mathsf{T}}\boldsymbol{x} - t_k), \quad \boldsymbol{u}_k := \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_2} \in \mathbb{S}^{d-1},\ t_k := \frac{b_k}{\|\boldsymbol{w}_k\|_2},\ a_k := v_k\, \|\boldsymbol{w}_k\|_2 \,.$$

Hence f_θ admits the normalized finite-sum form

$$f(\boldsymbol{x}) = \sum_{k=1}^{K'} a_k\, \phi(\boldsymbol{u}_k^{\mathsf{T}}\boldsymbol{x} - t_k) + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} + c_0. \tag{12}$$

Let the (ReLU) ridge dictionary be $\mathscr{D}_\phi := \left\{\phi(\boldsymbol{u}^{\mathsf{T}} \cdot - t) :\ \boldsymbol{u} \in \mathbb{S}^{d-1},\ t \in \mathbb{R}\right\}$. We study the *over-parametrized, width-agnostic* class given by the *union over all finite widths*

$$\mathcal{F}_{\mathrm{fin}} := \bigcup_{K \geq 1} \left\{ \sum_{k=1}^{K} a_k\, \phi(\boldsymbol{u}_k^{\mathsf{T}} \cdot - t_k) + \boldsymbol{c}^{\mathsf{T}}(\cdot) + c_0 \right\}, \tag{13}$$

and measure complexity by the minimal path-norm needed to realize f:

$$\|f\|_{\mathrm{path,min}} := \inf \left\{ \|f_{\boldsymbol{\theta}}\|_{\mathrm{path}} :\ f_{\boldsymbol{\theta}} \equiv f \text{ of the form (12)} \right\}.$$

From finite sums to a *width-agnostic* integral representation. To analyze $\mathcal{F}_{\mathrm{fin}}$ without committing to a fixed width K, we pass to a convex, measure-based description that *represents the closure/convex hull of* (13). Specifically, let ν be a finite signed Radon measure on $\mathbb{S}^{d-1} \times [-R, R]$ and consider

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1} \times [-R,R]} \phi(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{x} - t)\ \mathrm{d}\nu(\boldsymbol{u}, t) + c^{\mathsf{T}}\boldsymbol{x} + c_0. \tag{14}$$

Any finite network (12) corresponds to the *sparse* measure $\nu = \sum_{k=1}^{K} a_k\, \delta_{(\boldsymbol{u}_k, t_k)}$, and conversely sparse measures yield finite networks. Thus, (14) is a *width-agnostic relaxation* of (11), not an assumption of an infinite-width limit.

Definition D.2. *The (unweighted) variation (semi)norm*

$$|f|_{\mathrm{V}} := \inf \left\{ \|\nu\|_{\mathcal{M}} :\ f \text{ admits (12) for some } (\nu, c, c_0) \right\}, \tag{15}$$

where $\|\nu\|_{\mathcal{M}}$ is the total variation of ν.

For the compact region $\Omega = \mathbb{B}_R^d$, we define the bounded variation function class as

$$\mathrm{V}_C(\Omega) := \left\{ f : \Omega \to \mathbb{R} \mid f = \int_{\mathbb{S}^{d-1} \times [-R,R]} \phi(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{x} - t)\, \mathrm{d}\nu(\boldsymbol{u}, t) + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} + b,\ |f|_{\mathrm{V}} \leq C \right\}. \tag{16}$$

Specifically, by identifying (12) with the atomic measure $\nu = \sum_k a_k \delta_{(\boldsymbol{u}_k, t_k)}$, we have

$$|f|_{\mathrm{V}} \leq \sum_k |a_k| = \|f_{\boldsymbol{\theta}}\|_{\mathrm{path}}, \quad \text{hence} \quad |f|_{\mathrm{V}} \leq \|f\|_{\mathrm{path,min}}.$$

Conversely, the smallest variation needed to represent f equals the smallest path-norm across all finite decompositions,

$$\|f\|_{\mathrm{path,min}} = |f|_{\mathrm{V}}. \tag{17}$$

Thus, the variation seminorm (15) is the *nonparametric* counterpart of the path-norm, which captures the same notion of complexity but *without fixing the width K*.

Remark D.3 ("Arbitrary width" $\neq$ "infinite width"). *Our analysis concerns $\mathcal{F}_{\text{fin}}$ in (13), i.e., the union over all* finite *widths. The integral model (14) is a convexification/closure of this union that facilitates analysis and regularization; it* does not *assume an infinite-width limit. In variational training with a total-variation penalty on ν, first-order optimality ensures sparse solutions (finite support of ν), which correspond to* finite-width *networks. Thus, all results in this paper apply to* arbitrary (but finite) width*, and the continuum measure is only a device to characterize and control $\|f\|_{\text{path,min}}$.*

D.2 Total Variation Semi-norm on Radon Domain

We now connect the (unweighted) variation semi-norm of shallow ReLU networks to an analytic description on the *Radon domain*. Our presentation follows (Parhi & Nowak, 2021; 2023).

Definition D.4. *For a function $f : \mathbb{R}^d \to \mathbb{R}$ and $(\boldsymbol{u}, t) \in \mathbb{S}^{d-1} \times \mathbb{R} := \mathbb{S}^{d-1} \times \mathbb{R}$, the Radon transform and its dual are defined by*

$$\mathscr{R} f(\boldsymbol{u}, t) = \int_{\{\boldsymbol{x}:\, \boldsymbol{u}^\mathsf{T}\boldsymbol{x}=t\}} f(\boldsymbol{x})\, \mathrm{d}s(\boldsymbol{x})$$

$$\mathscr{R}^* \{\Phi\}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} \Phi\big(\boldsymbol{u},\, \boldsymbol{u}^\mathsf{T}\boldsymbol{x}\big)\, \mathrm{d}\sigma(\boldsymbol{u}).$$

The Radon framework encodes a function f by its integrals over affine hyperplanes faithfully in the senses that the Radon transform is invertible up to a known dimension-dependent constant via a one-dimensional "ramp" filter in t.

Proposition D.5 (Filtered backprojection (Radon inversion)). *There exists $c_d > 0$ such that*

$$c_d\, f \;=\; \mathscr{R}^* \left\{ \Lambda_{d-1} \mathscr{R} f \right\},$$

where Λ_{d-1} acts in the t-variable with Fourier symbol $\widehat{\Lambda_{d-1}\Phi}(\boldsymbol{u}, \omega) = i^{\,d-1}\, |\omega|^{\,d-1}\, \hat{\Phi}(\boldsymbol{u}, \omega)$.

The inversion formula motivates measuring the "ridge-curvature" of f by differentiating in the Radon offset t after filtering, and aggregating its magnitude over all orientations and offsets.

The next definition is the sole norm we need on the Radon domain; it specializes all higher-order variants to the ReLU case.

Definition D.6 (Second-order Radon total variation (ReLU case)). *The (second-order) Radon total-variation seminorm is*

$$\mathscr{R}\mathrm{TV}^2(f) := \left\| \mathscr{R}\left\{ (-\Delta)^{\frac{d+1}{2}} f \right\} \right\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})},$$

where the fractional power is understood in the tempered-distribution sense. The null space of $\mathscr{R}\mathrm{TV}^2(\cdot)$ is the set of affine functions on $\mathbb{R}^d$.

Proposition D.7 (Equivalence of seminorms on bounded domains (Parhi & Nowak, 2021)). *Let $\mathcal{B} = \mathbb{B}_R^d$. For any $f : \mathcal{B} \to \mathbb{R}$ with finite variation seminorm, its canonical extension f_{ext} to $\mathbb{R}^d$ satisfies*

$$|f|_{\mathrm{V}} \;=\; \mathscr{R}\mathrm{TV}^2(f_{\text{ext}}),$$

and, in particular, for any finite two-layer ReLU network in reduced form $f_\theta(\boldsymbol{x}) = \sum_{k=1}^K v_k\, \phi(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0$,

$$\mathscr{R}\mathrm{TV}^2(f_{\boldsymbol{\theta}}) \;=\; \sum_{k=1}^K |v_k|\; \|\boldsymbol{w}_k\|_2\,,$$

which equals the minimal (unweighted) path-norm needed to realize f_θ on $\mathbb{B}_R^d$.

The key structural reason is simple: $\partial_t^2 \Lambda_{d-1} \mathscr{R}$ turns each ReLU ridge $\phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)$ into a Dirac mass at $(\boldsymbol{u}, t)$ on $\mathbb{S}^{d-1} \times \mathbb{R}$, so superpositions of ridges correspond exactly to finite signed measures on $\mathbb{S}^{d-1} \times \mathbb{R}$, and the total-variation of that measure coincides with both the variation seminorm and $\mathscr{R}\mathrm{TV}^2(\cdot)$ after fixing the affine null space.

Remark D.8 (Takeaway). *For ReLU networks on bounded domains, the three viewpoints*

$$\textit{path-norm } \|f\|_{\text{path}} \;\longleftrightarrow\; \textit{unweighted variation } |f|_{\mathrm{V}} \;\longleftrightarrow\; \textit{Radon-TV } \mathscr{R}\mathrm{TV}(f)$$

are equivalent up to the affine null space. We will freely switch between them in the sequel.

D.3 The Metric Entropy of Variation Spaces

Metric entropy quantifies the compactness of a set A in a metric space (X, ρ_X). Below we introduce the definition of covering numbers and metric entropy.

Definition D.9 (Covering Number and Entropy)**.** *Let A be a compact subset of a metric space (X, ρ_X). For $t > 0$, the* covering number $N(A, t, \rho_X)$ *is the minimum number of closed balls of radius t needed to cover A:*

$$N(t, A, \rho_X) := \min \left\{ N \in \mathbb{N} : \exists\, x_1, \ldots, x_N \in X \text{ s.t. } A \subset \bigcup_{i=1}^{N} \mathbb{B}(x_i, t) \right\}, \tag{18}$$

where $\mathbb{B}(x_i, t) = \{y \in X : \rho_X(y, x_i) \le t\}$. The metric entropy *of A at scale t is defined as:*

$$H_t(A)_X := \log N(t, A, \rho_X). \tag{19}$$

The metric entropy of the bounded variation function class has been studied in previous works. More specifically, we will directly use the one below in future analysis.

Proposition D.10 (Parhi & Nowak 2023, Appendix D)**.** *The metric entropy of $\mathrm{V}_C(\mathbb{B}_R^d)$ (see Definition D.2) with respect to the $L^\infty(\mathbb{B}_R^d)$-distance $\|\cdot\|_\infty$ satisfies*

$$\log N(t, \mathrm{V}_C(\mathbb{B}_R^d), \|\cdot\|_\infty) \lessapprox_d \left(\frac{C}{t}\right)^{\frac{2d}{d+3}}. \tag{20}$$

where $\lessapprox_d$ hides constants (which could depend on d) and logarithmic factors.

D.4 Generalization Gap of Unweighted Variation Function Class

As a middle step towards bounding the generalization gap of the weighted variation function class, we first bound the generalization gap of the unweighted variation function class according to a metric entropy analysis.

Lemma D.11. *Let $\mathcal{F}_{M,C} = \{f \in \mathrm{V}_C(\mathbb{B}_R^d) \mid \|f\|_\infty \le M\}$ with $M \ge D$. Then let $\mathcal{D} \sim \mathcal{P}^{\otimes n}$ be a sampled data set of size n, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}_{M,C}} \left| R(f) - \widehat{R}_{\mathcal{D}}(f) \right| \lesssim_d C^{\frac{d}{2d+3}} M^{\frac{3(d+2)}{2d+3}} n^{-\frac{d+3}{4d+6}} + M^2 \left(\frac{\log(4/\delta)}{n}\right)^{-\frac{1}{2}}. \tag{21}$$

Proof. According to Proposition D.10, one just needs $N(t)$ balls to cover $\mathcal{F}$ in $\|\cdot\|_\infty$ with radius $t > 0$ such that where

$$\log N(t) \lessapprox_d \left(\frac{C}{t}\right)^{\frac{2d}{d+3}}.$$

Then for any $f, g \in \mathcal{F}_{M,C}$ and any $(\boldsymbol{x}, y)$,

$$\left|(f(\boldsymbol{x}) - y)^2 - (g(\boldsymbol{x}) - y)^2\right| = |f(\boldsymbol{x}) - g(\boldsymbol{x})|\,|f(\boldsymbol{x}) + g(\boldsymbol{x}) - 2y| \le 4M \|f - g\|_\infty.$$

Hence replacing f by a centre f_i within t changes both the empirical and true risks by at most $4Mt$.

For any fixed centre $\bar{f}$ in the covering, Hoeffding's inequality implies that with probability at least $\ge 1 - \delta$, we have

$$|R(\bar{f}) - \widehat{R}_{\mathcal{D}}(\bar{f})| \le 4M^2 \sqrt{\frac{\log(2/\delta)}{n}} \tag{22}$$

because each squared error lies in $[0, 4M^2]$. Then we take all the centers with union bound to deduce that with probability at least $1 - \delta/2$, for any center $\bar{f}$ in the set of covering index, we have

$$\begin{aligned} |R(\bar{f}) - \widehat{R}_{\mathcal{D}}(\bar{f})| &\le 4M^2 \sqrt{\frac{\log(4N(t)/\delta)}{n}} \\ &\lesssim M^2 \cdot \left(\frac{C}{t}\right)^{\frac{d}{d+3}} \left(\frac{1}{n}\right)^{-\frac{1}{2}} + M^2 \left(\frac{\log(4/\delta)}{n}\right)^{-\frac{1}{2}} \\ &\lessapprox_d M^2 \cdot \left(\frac{C}{t}\right)^{\frac{d}{d+3}} \left(\frac{1}{n}\right)^{-\frac{1}{2}}, \end{aligned} \tag{23}$$

where $\lessapprox_d$ hides the logarithmic factors about $1/\delta$ and constants.

According to the definition of covering sets, for any $f \in \mathcal{F}_{M,C}$, we have that $\|f - \bar{f}\|_\infty \le t$ for some center $\bar{f}$. Then we have

$$\begin{aligned} &|R(f) - \widehat{R}_{\mathcal{D}}(f)| \\ \lessapprox_d\ &|R(\bar{f}) - \widehat{R}_{\mathcal{D}}(\bar{f})| + O(Mt) \\ \lesssim_d\ &M^2 \cdot \left(\frac{C}{t}\right)^{\frac{d}{d+3}} n^{-\frac{1}{2}} + O(Mt). \end{aligned} \tag{24}$$

After tuning t to be the optimal choice, we deduce that (21). □

E Data-Dependent Regularity from Edge-of-Stability

This section summarizes the *data-dependent regularity* induced by minima stability for two-layer ReLU networks.

E.1 Function Space Viewpoint of Neural Networks Below the Edge of Stability

Recall the notations: given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, we define the data-dependent weight function $g_{\mathcal{D}} : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ by

$$g_{\mathcal{D}}(\boldsymbol{u}, t) \coloneqq \min\{\tilde{g}_{\mathcal{D}}(\boldsymbol{u}, t), \tilde{g}_{\mathcal{D}}(-\boldsymbol{u}, -t)\},$$

where

$$\tilde{g}_{\mathcal{D}}(\boldsymbol{u}, t) \coloneqq \mathbb{P}_{\mathcal{D}}(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} > t)^2 \cdot \mathbb{E}_{\mathcal{D}}[\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t \mid \boldsymbol{X}^\mathsf{T}\boldsymbol{u} > t] \cdot \sqrt{1 + \|\mathbb{E}_{\mathcal{D}}[\boldsymbol{X} \mid \boldsymbol{X}^\mathsf{T}\boldsymbol{u} > t]\|^2}. \tag{25}$$

Here, $\boldsymbol{X}$ denotes a random draw uniformly sampled from $\{\boldsymbol{x}_i\}_{i=1}^n$, so that $\mathbb{P}_{\mathcal{D}}, \mathbb{E}_{\mathcal{D}}$ refer to probability and expectation under the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}$. When the dataset $\mathcal{D}$ is fixed and clear from context, we will simply write g in place of $g_{\mathcal{D}}$.

Then the curvature constrain on the loss landscape of $\mathcal{L}$ is converted into a weighted path norm constrain in the following sense.

Proposition E.1 (Finite-sum version of Theorem 3.2 in (Liang et al., 2025))**.** *Suppose that $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^K v_k\, \phi(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + \beta$ is two-layer neural network such that the loss $\mathcal{L}$ is twice differentiable at $\boldsymbol{\theta}$. Then*

$$\sum_{k=1}^K |v_k|\, \|\boldsymbol{w}_k\| \cdot g\Big(\frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|}, \frac{b_k}{\|\boldsymbol{w}_k\|}\Big) \le \frac{\lambda_{\max}(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}))}{2} - \frac{1}{2} + (R+1)\sqrt{2\mathcal{L}(\boldsymbol{\theta})}. \tag{26}$$

If we write $f_{\boldsymbol{\theta}}$ into a reduced form in (12), then we have

$$\sum_{k=1}^{K'} a_k \cdot g\Big(\boldsymbol{u}_k, t_k\Big) \le \frac{\lambda_{\max}(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}))}{2} - \frac{1}{2} + (R+1)\sqrt{2\mathcal{L}(\boldsymbol{\theta})}. \tag{27}$$

Therefore, we bring up the definition the g-weighted path norm and variation norm are introduced as prior work introduced (Liang et al., 2025; Nacson et al., 2023).

Definition E.2. *Let $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^K v_k\, \phi(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + \beta$ be a two-layer neural network. The (g-)weighted path-norm of $f_{\boldsymbol{\theta}}$ is defined to be*

$$\|f_{\boldsymbol{\theta}}\|_{\mathrm{path},g} \coloneqq \sum_{k=1}^K |v_k|\, \|\boldsymbol{w}_k\|_2 \cdot g\Big(\frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|}, \frac{b_k}{\|\boldsymbol{w}_k\|}\Big). \tag{28}$$

Similarly, for functions of the form

$$f_{\nu,\boldsymbol{c},c_0}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-R,R]} \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)\, \mathrm{d}\nu(\boldsymbol{u}, t) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0, \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{29}$$

where $R > 0$, $\boldsymbol{c} \in \mathbb{R}^d$, and $c_0 \in \mathbb{R}$, we define the g-weighted variation (semi)norm *as*

$$|f|_{\mathrm{V}_g} := \inf_{\substack{\nu \in \mathcal{M}(\mathbb{S}^{d-1} \times [-R,R]) \\ \boldsymbol{c} \in \mathbb{R}^d, c_0 \in \mathbb{R}}} \|g \cdot \nu\|_{\mathcal{M}} \quad \text{s.t.} \quad f = f_{\nu, \boldsymbol{c}, c_0}, \tag{30}$$

where, if there does not exist a representation of f in the form of (29), then the seminorm is understood to take the value $+\infty$. Here, $\mathcal{M}(\mathbb{S}^{d-1} \times [-R, R])$ denotes the Banach space of (Radon) measures and, for $\mu \in \mathcal{M}(\mathbb{S}^{d-1} \times [-R, R])$, $\|\mu\|_{\mathcal{M}} := \int_{\mathbb{S}^{d-1} \times [-R,R]} \mathrm{d}|\mu|(\boldsymbol{u}, t)$ is the measure-theoretic total-variation norm.

With this seminorm, we define the Banach space of functions $\mathrm{V}_g(\mathbb{B}_R^d)$ on the ball $\mathbb{B}_R^d := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \le R\}$ as the set of all functions f such that $|f|_{\mathrm{V}_g}$ is finite. When $g \equiv 1$, $|\cdot|_{\mathrm{V}_g}$ and $\mathrm{V}_g(\mathbb{B}_R^d)$ coincide with the variation (semi)norm and variation norm space of Bach (2017).

For convenience, we introduce the notation of bounded weighted variation class

$$\mathcal{F}_g(\Omega; M, C) := \left\{ f : \Omega \to \mathbb{R} \middle| |f|_{\mathrm{V}_g} \le C,\ \|f|_{\Omega}\|_{L^\infty} \le M \right\}. \tag{31}$$

In particular, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}_g(\Omega; M, C)$, we have $f_{\boldsymbol{\theta}} \in \mathcal{F}_g(\Omega; M, C)$.

Within this framework together with the connection between $|\cdot|_{\mathrm{V}}$ and $\mathscr{R}\mathrm{TV}^2(\cdot)$ as summarized in Section D.2, we show the functional characterization of stable minima.

Theorem E.3. *For any $f_{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})$, $|f_{\boldsymbol{\theta}}|_{\mathrm{V}_g} = \|g \cdot \mathscr{R}(-\Delta)^{\frac{d+1}{2}} f_{\boldsymbol{\theta}}\|_{\mathcal{M}} \le \frac{1}{\eta} - \frac{1}{2} + (R+1)\sqrt{2\mathcal{L}(\boldsymbol{\theta})}$.*

The detailed explanation and proof can be found in (Liang et al., 2025, Theorem 3.2, Corollary 3.3, Theorem 3.4, Appendix C, D).

E.2 Empirical Process for the Weight Function g

The implicit regularization of Edge-of-Stability induces a *data-dependent* regularity weight on the cylinder $\mathbb{S}^{d-1} \times \mathbb{R} := \mathbb{S}^{d-1} \times [-1, 1]$. Denote this empirical weight by $g_{\mathcal{D}}$ for a dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$. Directly analyzing generalization through the random, data-dependent class weighted by $g_{\mathcal{D}}$ is conceptually delicate, since the hypothesis class itself depends on the sample. To separate *statistical* from *algorithmic* randomness, we adopt the following paradigm.

(I) Fix an underlying distribution $\mathcal{P}$ for X with only the support assumption $\mathrm{supp}(\mathcal{P}) \subseteq \mathbb{B}_R^d := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \le R\}$. Define a *population* reference weight $g_{\mathcal{P}}$ on $\mathbb{S}^{d-1} \times \mathbb{R}$ (see below). This anchors a distribution-level notion of regularity independent of the particular sample.

(ii) For a realized dataset $\mathcal{D} \sim \mathcal{P}^{\otimes n}$, form the empirical plug-ins that define $g_{\mathcal{D}}$ on the same index set $\mathbb{S}^{d-1} \times \mathbb{R}$.

(iii) Use empirical-process theory to control the uniform deviation $\|g_{\mathcal{D}} - g_{\mathcal{P}}\|_\infty$ with high probability over the draw of $\mathcal{D}$. After this step, we can *condition on the high-probability event* and regard $\mathcal{D}$ as fixed in any subsequent analysis.

Let $\boldsymbol{X} \sim \mathcal{P}$ with $\mathrm{supp}(\mathcal{P}) \subseteq \mathbb{B}_R^d$. For $(\boldsymbol{u}, t) \in \mathbb{S}^{d-1} \times \mathbb{R}$ define

$$p_{\mathcal{P}}(\boldsymbol{u}, t) := \mathcal{P}\big(\boldsymbol{X}^\top \boldsymbol{u} > t\big), \qquad s_{\mathcal{P}}(\boldsymbol{u}, t) := \mathbb{E}_{\boldsymbol{X} \sim \mathcal{P}}\big[(\boldsymbol{X}^\top \boldsymbol{u} - t)_+\big].$$

On the unit ball we have $0 \le (\boldsymbol{X}^\top \boldsymbol{u} - t)_+ \le 2$ and $\|\mathbb{E}_{\mathcal{P}}[X \mid \boldsymbol{X}^\top \boldsymbol{u} > t]\| \le 1$, which yields the *pointwise equivalence*

$$g_{\mathcal{P}}(\boldsymbol{u}, t) \asymp p_{\mathcal{P}}(\boldsymbol{u}, t)\, s_{\mathcal{P}}(\boldsymbol{u}, t) \quad \text{(with absolute constants)}. \tag{32}$$

Given a dataset $\mathcal{D} = \{x_i\}_{i=1}^n$, let $\mathbb{P}_{\mathcal{D}}, \mathbb{E}_{\mathcal{D}}$ denote probability and expectation under the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Define

$$p_{\mathcal{D}}(\boldsymbol{u}, t) := \mathbb{P}_{\mathcal{D}}(\boldsymbol{X}^\top \boldsymbol{u} > t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\boldsymbol{x}_i^\top \boldsymbol{u} > t\}, \quad s_{\mathcal{D}}(\boldsymbol{u}, t) := \mathbb{E}_{\mathcal{D}}\big[(\boldsymbol{X}^\top \boldsymbol{u} - t)_+\big] = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{u} - t)_+,$$

and the empirical weight

$$g_{\mathcal{D}}(\boldsymbol{u}, t) \asymp p_{\mathcal{D}}(\boldsymbol{u}, t)\, s_{\mathcal{D}}(\boldsymbol{u}, t). \tag{33}$$

Lemma E.4 (Uniform deviation for halfspaces). *There exists a universal constant $C > 0$ such that, for every $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\sup_{u \in \mathbb{S}^{d-1},\, t \in [-1,1]} \left| p_{\mathcal{D}}(\boldsymbol{u}, t) - p_{\mathcal{P}}(\boldsymbol{u}, t) \right| > C\sqrt{\frac{d + \log(1/\delta)}{n}} \right) \leq \delta.$$

Proof. The class $\{(\boldsymbol{x} \mapsto \mathbb{1}\{\boldsymbol{x}^\mathsf{T}\boldsymbol{u} > t\}) : \boldsymbol{u} \in \mathbb{S}^{d-1},\, t \in \mathbb{R}\}$ has VC-dimension $d + 1$. Apply the VC-uniform convergence inequality for $\{0, 1\}$-valued classes (e.g., Vapnik (1998)) to the index set $\mathbb{S}^{d-1} \times [-1, 1]$ to obtain the stated bound. □

Lemma E.5 (Uniform deviation for ReLU). *There exists a universal constant $C > 0$ such that, for every $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\sup_{u \in \mathbb{S}^{d-1},\, t \in [-1,1]} \left| s_{\mathcal{D}}(\boldsymbol{u}, t) - s_{\mathcal{P}}(\boldsymbol{u}, t) \right| > C\sqrt{\frac{d + \log(1/\delta)}{n}} \right) \leq \delta.$$

Proof. Let $\mathcal{F} := \{f_{\boldsymbol{u},t}(\boldsymbol{x}) = (\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)_+ : \boldsymbol{u} \in \mathbb{S}^{d-1},\, t \in [-1, 1]\}$. Since $\|\boldsymbol{x}\| \leq 1$ and $t \in [-1, 1]$, every $f \in \mathcal{F}$ takes values in $[0, 2]$. Consider the subgraph class

$$\mathsf{subG}(\mathcal{F}) = \left\{ (\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R} :\ y \leq (\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)_+ \right\}.$$

For any $(\boldsymbol{x}, y)$ with $y \leq 0$, membership in $\mathsf{subG}(\mathcal{F})$ holds for all parameters, hence such points do not contribute to shattering. For points with $y > 0$, the condition $y \leq (\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)_+$ is equivalent to $\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t - y \geq 0$, i.e., an affine halfspace in $\mathbb{R}^{d+1}$ with variables $(\boldsymbol{x}, y)$. Therefore the family $\mathsf{subG}(\mathcal{F})$ is (up to the immaterial fixed set $\{y \leq 0\}$) parametrized by affine halfspaces in $\mathbb{R}^{d+1}$, whose VC-dimension is at most $d+2$. By the standard equivalence $\mathrm{Pdim}(\mathcal{F}) = \mathrm{VCdim}(\mathsf{subG}(\mathcal{F}))$, we obtain

$$\mathrm{Pdim}(\mathcal{F}) \ \leq\ d + 2.$$

Then by (Haussler, 1992, Theorem 3, Theorem 6, Theorem 7), we

$$\sup_{(\boldsymbol{u},t)} \left| s_{\mathcal{D}}(\boldsymbol{u}, t) - s_{\mathcal{P}}(\boldsymbol{u}, t) \right| \ \leq\ C\sqrt{\frac{d + \log(1/\delta)}{n}}$$

with probability at least $1 - \delta$ for some universal constant C, which is the claimed bound. □

Theorem E.6 (Distribution-free uniform deviation for $\hat{g}_n$). *There exists a universal constant $C > 0$ such that, for every $\delta \in (0, 1)$,*

$$\mathbb{P}\left(\sup_{\boldsymbol{u} \in \mathbb{S}^{d-1},\, t \in [-1,1]} \left| g_{\mathcal{D}}(\boldsymbol{u}, t) - g_{\mathcal{P}}(\boldsymbol{u}, t) \right| > C\sqrt{\frac{d + \log(1/\delta)}{n}} \right) \leq 2\delta.$$

Proof. By (32) and (33), it suffices (up to a universal factor) to control $\left| p_{\mathcal{D}} s_{\mathcal{D}} - p_{\mathcal{P}} s_{\mathcal{P}} \right|$. Using $0 \leq s_{\mathcal{D}}, s_{\mathcal{P}} \leq 2$ and $0 \leq p_{\mathcal{D}}, p_{\mathcal{P}} \leq 1$,

$$\left| p_{\mathcal{D}} s_{\mathcal{D}} - p_{\mathcal{P}} s_{\mathcal{P}} \right| \leq\ \left| p_{\mathcal{D}} - p_{\mathcal{P}} \right| s_{\mathcal{P}} + \left| s_{\mathcal{D}} - s_{\mathcal{P}} \right| p_{\mathcal{P}} + \left| p_{\mathcal{D}} - p_{\mathcal{P}} \right| \left| s_{\mathcal{D}} - s_{\mathcal{P}} \right|$$

Taking the supremum over $(\boldsymbol{u}, t) \in \mathbb{S}^{d-1} \times [-1, 1]$ and applying Lemmas E.4 and E.5 with a union bound yields

$$\mathbb{P}\left(\sup_{\boldsymbol{u},t} \left| p_{\mathcal{D}} s_{\mathcal{D}} - p_{\mathcal{P}} s_{\mathcal{P}} \right| \gtrsim \sqrt{\frac{d + \log(1/\delta)}{n}} \right) \leq 2\delta.$$

Finally, the equivalence $g \asymp p\, s$ transfers this bound to $\left| g_{\mathcal{D}} - g_{\mathcal{P}} \right|$ at the cost of an absolute multiplicative factor and one more failure event. □

F GENERALIZATION UPPER BOUND: MIXTURE OF LOW-DIMENSIONAL BALLS

In this section, we present the proof of Theorem 3.2. First, we prove the simple case of singe-subspace assumption ($J = 1$) via Theorem F.2.

F.1 CASE: UNIFORM DISTRIBUTION ON UNIT DISC OF A LINEAR SUBSPACE

Fix an m-dimensional subspace $V \subset \mathbb{R}^d$ and write $\mathbb{B}_1^V := \{\boldsymbol{x} \in V : \|\boldsymbol{x}\|_2 \le 1\}$, the canonical linear projection $\mathrm{proj}_V : \mathbb{R}^d \to V$. Recall the notations in (1): the parameters $\boldsymbol{\theta} := \{(v_k, \boldsymbol{w}_k, b_k)_{k=1}^K,\ \beta\}$ with $\boldsymbol{w}_k \neq \mathbf{0}$, define a two-layer neural network

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k\, \phi(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + \beta, \qquad \bar{\boldsymbol{w}}_k := \frac{\boldsymbol{w}_k}{\|\boldsymbol{w}_k\|_2}, \quad \bar{b}_k := \frac{b_k}{\|\boldsymbol{w}_k\|_2}.$$

Then we define neuronwise projection operator from neural networks to neural networks

$$\mathrm{proj}_V^* :\ f_{\boldsymbol{\theta}}(\boldsymbol{x}) \mapsto \sum_{k=1}^{K} v_k\, \phi\big((\mathrm{proj}_V\, \boldsymbol{w}_k)^\mathsf{T}\boldsymbol{x} - b_k\big) + \beta. \tag{34}$$

Lemma F.1 (Projection reduction)**.** *Fix $\mathcal{F}$ a hyothesis class of two-layer neural networks. Let $\mathcal{P}$ be a joint distribution on $(\boldsymbol{x}, y)$ supported on $\mathbb{R}^d \times [-D, D]$ such that the marginal distribution $\mathcal{P}_{\boldsymbol{X}}$ of $\boldsymbol{x}$ supports on V. For any dataset $\mathcal{D} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ drawn i.i.d. from $\mathcal{P}$,*

$$\sup_{f \in \mathcal{F}} \mathrm{Gap}_{\mathcal{P}}(f; \mathcal{D}) = \sup_{f \in \mathcal{F}} \mathrm{Gap}_{\mathcal{P}}(\mathrm{proj}_V^*\, f; \mathcal{D}). \tag{35}$$

Proof. Because $\boldsymbol{x} \in V$ almost surely and in the sample, we have $f(\boldsymbol{x}) = (f \circ \mathrm{proj}_V)(\boldsymbol{x})$ for every f and every $\boldsymbol{x} \in \mathbb{B}_1^V$. Using the identity $\boldsymbol{w}_k^\mathsf{T}(\mathrm{proj}_V\, \boldsymbol{x}) = (\mathrm{proj}_V\, \boldsymbol{w}_k)^\mathsf{T}\boldsymbol{x}$, we obtain $f \circ \mathrm{proj}_V = \mathrm{proj}_V^*\, f$ pointwise on $\mathbb{B}_1^V$. Hence for any $f \in \mathcal{F}$, $\mathrm{Gap}_{\mathcal{P}}(f; \mathcal{D}) = \mathrm{Gap}_{\mathcal{P}}(\mathrm{proj}_V^*\, f; \mathcal{D})$. □

Theorem F.2. *Let $\mathcal{P}$ denote the joint distribution of $(\boldsymbol{x}, y)$. Assume that $\mathcal{P}$ is supported on $\mathbb{B}_1^d \times [-D, D]$ for some $D > 0$ and that the marginal distribution of $\boldsymbol{x}$ is $\mathrm{Uniform}(\mathbb{B}_1^V)$. Fix a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where each $(\boldsymbol{x}_i, y_i)$ is drawn i.i.d. from $\mathcal{P}$. Then, with probability $\ge 1 - \delta$,*

$$\sup_{f_{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{g_{\mathcal{D}}}(\mathbb{B}_1^{V_j}; M, C)} \mathrm{Gap}_{\mathcal{P}}(f_{\boldsymbol{\theta}}; \mathcal{D}) \ \lesssim_d\ C^{\frac{m}{m^2+4m+3}}\, M^2\, n^{-\frac{1}{2m+4}} + M^2 \left(\frac{\log(4/\delta)}{n}\right)^{-\frac{1}{2}},$$

where $M := \max\{D, \|f_{\boldsymbol{\theta}}\|_{L^\infty(\mathbb{B}_1^V)}, 1\}$ and $\lesssim_d$ hides constants (which could depend on d).

Proof. By Lemma F.1, it remains to consider the case of $\mathrm{proj}_V^*\, f_{\boldsymbol{\theta}}$. Similarly, for any $\boldsymbol{u} \in \mathbb{S}^{d-1}$ and any data set $\mathcal{D} \subset V$, we have $g(\boldsymbol{u}, t) = g(\mathrm{proj}_V(\boldsymbol{u}), t)$. Therefore, we just need to consider the generalization gap with respect to the $\boldsymbol{\Theta}_{g_{\mathcal{D}}}^V(\mathbb{B}_1^{V_j}; M, C) = \left\{\mathrm{proj}_V^*\, f_{\boldsymbol{\theta}} :\ f_{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{g_{\mathcal{D}}}(\mathbb{B}_1^{V_j}; M, C)\right\}$. Therefore, we just need consider the case where the whole algorithm with any dataset sample from V operates in V and we get the result from (Liang et al., 2025, Theorem F.8) by replacing $\mathbb{R}^d$ with $V \cong \mathbb{R}^m$. □

F.2 PROOF OF THEOREM 3.2

In this section, we extend the generalization analysis from a single low-dimensional subspace to a more complex and practical scenario where the data is supported on a finite union of such subspaces. This setting is crucial for modeling multi-modal data, where distinct clusters can each be approximated by a low-dimensional linear structure. Our main result demonstrates that the sample complexity of stable minima adapts to the low intrinsic dimension of the individual subspaces, rather than the high ambient dimension of the data space.

F.2.1 Analysis of the Global Weight Function

A critical step in our proof is to understand the relationship between the global weight function $g(\boldsymbol{u}, t)$, which is induced by the mixture distribution $\mathcal{P}$, and the local weight functions $g_j(\boldsymbol{u}, t)$, each induced by a single component distribution $\mathcal{P}_j$ defined on V_j, which should be understood as the distribution conditioned to $\boldsymbol{x} \in V_j$. Fix a dataset $\mathcal{D}$, the function class $\boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta; \mathcal{D})$ is defined by the properties of the global function g. To analyze the performance on a specific subspace V_j, we must ensure that the global regularity constraint is sufficiently strong when viewed locally. The following lemma provides this crucial guarantee.

Lemma F.3 (Global-to-Local Weight Domination)**.** *For any mixed distribution $\mathcal{P}_{\boldsymbol{X}} = \sum_{j=1}^{J} p_j \mathcal{P}_{\boldsymbol{X},j}$ with $\mathrm{supp}(\mathcal{P}_{\boldsymbol{X},j}) = V_j$. Let g be the global weight induced by the mixture $\mathcal{P}_{\boldsymbol{X}}$, and g_j the weight induced by $\mathcal{P}_{\boldsymbol{X},j}$. For every $j \in \{1, \dots, J\}$,*

$$g(\boldsymbol{u}, t) \geq \frac{p_j^2}{\sqrt{2}} g_j(\boldsymbol{u}, t), \quad \text{for all } (\boldsymbol{u}, t) \in \mathbb{S}^{d-1} \times \mathbb{R}. \tag{36}$$

Consequently, for any $M, C > 0$,

$$\mathcal{F}_g(\mathbb{B}_1^{V_j}; M, C) \;\subseteq\; \mathcal{F}_{g_j}\big(\mathbb{B}_1^{V_j}; M, \sqrt{2}\, C/p_j^2\big)\,. \tag{37}$$

Proof. Fix j and the activation event $A := \{\boldsymbol{x} : \boldsymbol{u}^\mathsf{T}\boldsymbol{x} > t\}$. By definition of g (global) and g_j (local) we can write

$$g(\boldsymbol{u}, t) = \mathcal{P}_{\boldsymbol{X}}(A)^2 \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X}}}[\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t \mid A] \cdot \sqrt{1 + \| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X}}}[\boldsymbol{X} \mid A]\|_2^2}$$

$$g_j(\boldsymbol{u}, t) = \mathcal{P}_{\boldsymbol{X}}(A \mid \boldsymbol{x} \in V_j)^2 \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X}}}[\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t \mid A, \boldsymbol{x} \in V_j] \cdot \sqrt{1 + \| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X}}}[\boldsymbol{X} \mid A, \boldsymbol{x} \in V_j]\|_2^2}$$

$$= \mathcal{P}_{\boldsymbol{X},j}(A)^2 \cdot \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t \mid A] \cdot \sqrt{1 + \| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[\boldsymbol{X} \mid A]\|_2^2}$$

Using the law of total probability and total expectation for the mixture distribution $\mathcal{P}_{\boldsymbol{X}} = \sum_{i=1}^{J} p_i \mathcal{P}_{\boldsymbol{X},i}$, and the non-negativity of $(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A$, we get

$$\mathcal{P}_{\boldsymbol{X}}(A) \;\geq\; p_j\, \mathcal{P}_{\boldsymbol{X},j}(A), \qquad \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X}}}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A] \;\geq\; p_j \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A].$$

Hence, by combining the first two terms of $g(\boldsymbol{u}, t)$ as $\mathcal{P}_{\boldsymbol{X}}(A)\, \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X}}}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A]$, we have:

$$g(\boldsymbol{u}, t) \;\geq\; \big(p_j \mathcal{P}_{\boldsymbol{X},j}(A)\big) \cdot \big(p_j \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A]\big) \cdot 1 \;=\; p_j^2\, \mathcal{P}_{\boldsymbol{X},j}(A) \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A].$$

For the local weight function g_j, the same algebra gives

$$g_j(\boldsymbol{u}, t) = \mathcal{P}_{\boldsymbol{X},j}(A) \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A] \cdot \sqrt{1 + \| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[\boldsymbol{X} \mid A]\|_2^2}\,.$$

Since the support of $\mathcal{P}_{\boldsymbol{X},j}$ is $\mathbb{B}_1^{V_j}$, we have $\|\boldsymbol{X}\|_2 \leq 1$ almost surely under $\mathcal{P}_{\boldsymbol{X},j}$. This implies $\| \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[\boldsymbol{X} \mid A]\|_2 \leq 1$, and therefore $\sqrt{1 + \| \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[\boldsymbol{X} \mid A]\|_2^2} \leq \sqrt{2}$.

Combining these results, we establish the lower bound:

$$g(\boldsymbol{u}, t) \;\geq\; \frac{p_j^2}{\sqrt{2}} \left(\mathcal{P}_{\boldsymbol{X},j}(A) \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{u} - t)\mathbb{1}_A] \cdot \sqrt{1 + \| \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_{\boldsymbol{X},j}}[\boldsymbol{X} \mid A]\|_2^2} \right) \;=\; \frac{p_j^2}{\sqrt{2}}\, g_j(\boldsymbol{u}, t),$$

which proves (36). The class embedding (37) follows directly from the definition of the weighted variation seminorm. □

Proposition F.4. *Let $\mathcal{P}$ be a distribution defined in Assumption 3.1 and recall that $\mathcal{P}_j$ is $\mathcal{P}$ conditional to $\boldsymbol{x} \in V_j$. Fix $j \in \{1, \dots, J\}$ and a data set $\mathcal{D} \sim \mathcal{P}^{\otimes n}$. Let $\mathcal{D}_j := \mathcal{D} \cap V_j$ and $n_j := |\mathcal{D}_j|$. Then with probability $1 - \delta$,*

$$\sup_{f_{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})} \mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j) \;\lesssim_d\; \left(\frac{\frac{1}{\eta} - \frac{1}{2} + 4M}{p_j^2}\right)^{\frac{m}{m^2+4m+3}} M^2\, n_j^{-\frac{1}{2m+4}} + M^2 \left(\frac{\log(4/\delta)}{n}\right)^{-\frac{1}{2}}. \tag{38}$$

where $M := \max\{D, \|f_{\boldsymbol{\theta}}\|_{L^\infty(\mathbb{B}_1^{V_j})}, 1\}$ and $\lesssim_d$ hides constants (which could depend on d).

Proof. Note that the notation $\mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j)$ can be expanded into

$$\begin{aligned}\mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j) &= \left| R_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}) - \widehat{R}_{\mathcal{D}_j}(f_{\boldsymbol{\theta}}) \right| \\ &= \left| \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{P}_j} \left[(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2 \right] - \widehat{R}_{\mathcal{D}_j}(f_{\boldsymbol{\theta}}) \right| \\ &= \left| \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{P}} \left[(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2 \mid \boldsymbol{x} \in V_j \right] - \widehat{R}_{\mathcal{D}_j}(f_{\boldsymbol{\theta}}) \right|\end{aligned}$$

Let $C = \frac{1}{\eta} - \frac{1}{2} + 4M$. According to (Liang et al., 2025, Corollary 3.3), we have that

$$f_{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{g_{\mathcal{D}}}(\mathbb{B}_1^{V_j}; M, C), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta; \mathcal{D}).$$

Then by Lemma F.3, we conclude that

$$\boldsymbol{\Theta}_g(\mathbb{B}_1^{V_j}; M, C) \subseteq \boldsymbol{\Theta}_{g_j}(\mathbb{B}_1^{V_j}; M, \sqrt{2}C/p_j^2),$$

where the weight functions g and g_j can be either empirical or population.

Therefore,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta;\mathcal{D})} \mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j) \leq \sup_{f\in\boldsymbol{\Theta}_{g_j}(\mathbb{B}_1^{V_j}; M, \sqrt{2}C/p_j^2)} \mathrm{Gap}_{\mathcal{P}_j}(f; \mathcal{D}_j)$$

Then by Theorem F.2, we may conclude that

$$\sup_{f_{\boldsymbol{\theta}}\in\mathcal{F}_{g_j}(\mathbb{B}_1^{V_j}; M, \sqrt{2}C/p_j^2)} \mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j) \lesssim_d \left(\frac{\frac{1}{\eta} - \frac{1}{2} + 4M}{p_j^2} \right)^{\frac{m}{m^2+4m+3}} M^2 \, n_j^{-\frac{1}{2m+4}}$$

□

Theorem F.5 (Generalization Bound for Mixture Models)**.** *Let the data distribution $\mathcal{P}$ be as defined in Assumption 1. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples drawn from $\mathcal{P}$. Then, with probability at least $1 - 2\delta$,*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta,\mathcal{D})} \mathrm{Gap}_{\mathcal{P}}(f_{\boldsymbol{\theta}}; \mathcal{D}) \lesssim_d \left(\frac{1}{\eta} - \frac{1}{2} + 4M \right)^{\frac{m}{m^2+4m+3}} M^2 \, J^{\frac{4}{m}} \, n^{-\frac{1}{2m+4}} + M^2 \, J \sqrt{\frac{\log(4J/\delta)}{2n}}. \tag{39}$$

where $M := \max\{D, \|f_{\boldsymbol{\theta}}\|_{\mathbb{B}_1^V}\|_{L^\infty}, 1\}$ and $\lesssim_d$ hides constants (which could depend on d).

The proof proceeds in several steps. First, we establish a high-probability event where the number of samples drawn from each subspace is close to its expected value. Second, we decompose the total generalization gap into several terms. Finally, we bound each of these terms, showing that the dominant term is determined by the generalization performance on the individual subspaces, which scales with the intrinsic dimension m.

Proof. Let $n_j = \sum_{i=1}^n \mathbb{1}_{\{x_i \in V_j\}}$ be the number of samples from the dataset $\mathcal{D}$ that fall into the subspace V_j. Each n_j is a random variable following a Binomial distribution, $n_j \sim \mathrm{Bin}(n, p_j)$. We need to ensure that for all subspaces simultaneously, the empirical proportion n_j/n is close to the true probability p_j.

We use Hoeffding's inequality for each $j \in \{1, \ldots, J\}$. For any $\epsilon > 0$, $\mathbb{P}\left(\left|\frac{n_j}{n} - p_j\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}$. To ensure this holds for all J subspaces at once, we apply a union bound. Let δ_j be the failure probability allocated to the j-th subspace. The total failure probability is at most $\sum_{j=1}^J \delta_j = \delta$, so we set $\delta_j = \delta/J$ and yields $\epsilon = \sqrt{\frac{\log(2J/\delta)}{2n}}$.

Let $\mathcal{E}$ be the event that $|\frac{n_j}{n} - p_j| \leq \epsilon$ holds for all $j = 1, \ldots, J$. We have shown that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. The remainder of our proof is conditioned on this event $\mathcal{E}$. A direct consequence of this event is a lower bound on each n_j

$$n_j \geq np_j - n\epsilon = np_j - \sqrt{\frac{n}{2}\log\frac{2J}{\delta}}. \tag{40}$$

Now we decompose the generalization gap using the law of total expectation for the true risk and by partitioning the empirical sum for the empirical risk.

Let $\mathcal{P}_j$ denote the distribution $\mathcal{P}$ conditioned on $x \in V_j$, and let $\mathcal{D}_j = \mathcal{D} \cap V_j\}$.

$$\begin{aligned}
\mathrm{Gap}_{\mathcal{P}}(f_{\boldsymbol{\theta}}; \mathcal{D}) &= \left| R(f_{\boldsymbol{\theta}}) - \widehat{R}_{\mathcal{D}}(f_{\boldsymbol{\theta}}) \right| \\
&= \left| \sum_{j=1}^{J} p_j \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}}[(f(\boldsymbol{x}) - y)^2 \mid \boldsymbol{x} \in V_j] - \sum_{j=1}^{J} \frac{n_j}{n}\frac{1}{n_j} \sum_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}_j} (f(\boldsymbol{x}_i) - y_i)^2 \right| \\
&\leq \left| \sum_{j=1}^{J} p_j \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{P}_j}[(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2] - \sum_{j=1}^{J} p_j \frac{1}{n_j} \sum_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}_j} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2 \right| \\
&\quad + \left| \sum_{j=1}^{J} p_j \frac{1}{n_j} \sum_{(x_i, y_i)\in\mathcal{D}_j} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2 - \sum_{j=1}^{J} \frac{n_j}{n}\frac{1}{n_j} \sum_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}_j} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2 \right| \\
&\leq \sum_{j=1}^{J} p_j \left| R_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}) - \widehat{R}_{\mathcal{D}_j}(f_{\boldsymbol{\theta}}) \right| + \sum_{j=1}^{J} \left| p_j - \frac{n_j}{n} \right| \widehat{R}_{\mathcal{D}_j}(f_{\boldsymbol{\theta}}) \\
&= \underbrace{\sum_{j=1}^{J} p_j \mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j)}_{\text{Term A}} + \underbrace{\sum_{j=1}^{J} \left| p_j - \frac{n_j}{n} \right| \widehat{R}_{\mathcal{D}_j}(f_{\boldsymbol{\theta}})}_{\text{Term B}}
\end{aligned}$$

where $\widehat{R}_{\mathcal{D}_j}(f) = \frac{1}{n_j}\sum_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}_j}(f(\boldsymbol{x}_i) - y_i)^2$.

- **Bounding the Weighted Sum of Conditional Gaps (Term A):** According to Proposition F.4, with probability at least $1 - \delta$, for each j,

$$\mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j) \lesssim_d \left(\frac{\frac{1}{\eta} - \frac{1}{2} + 4M}{p_j^2}\right)^{\frac{m}{m^2+4m+3}} M^2 \, n_j^{-\frac{1}{2m+4}} + M^2 \left(\frac{\log(4J/\delta)}{n}\right)^{-\frac{1}{2}}.$$

  Conditioned on $\mathcal{E}$, we use the lower bound on n_j from (40) , $n_j \leq np_j(1 - \epsilon/p_j)$.

$$\begin{aligned}
\text{Term A} &= \sum_{j=1}^{J} p_j \mathrm{Gap}_{\mathcal{P}_j}(f_{\boldsymbol{\theta}}; \mathcal{D}_j) \\
&\lessapprox_d \sum_{j=1}^{J} p_j \left(\frac{\frac{1}{\eta} - \frac{1}{2} + 4M}{p_j^2}\right)^{\frac{m}{m^2+4m+3}} M^2 (np_j(1 - \epsilon/p_j))^{-\frac{1}{2m+4}} \\
&= \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{m}{m^2+4m+3}} M^2 n^{-\frac{1}{2m+4}} \sum_{j=1}^{J} p_j \cdot (p_j^{-2})^{\frac{m}{m^2+4m+3}} \cdot \left(p_j - \sqrt{\frac{\log(2J/\delta)}{2n}}\right)^{-\frac{1}{2m+4}} \\
&\lessapprox_d \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{m}{m^2+4m+3}} M^2 n^{-\frac{1}{2m+4}} \sum_{j=1}^{J} p_j^{1 - \frac{2m}{m^2+4m+3} - \frac{1}{2m+4}}.
\end{aligned}$$

  The exponent of p_j simplifies to

$$1 - \frac{2m}{(m+1)(m+3)} - \frac{1}{2m+4} = \frac{2m^3 + 7m^2 + 10m + 9}{2(m+1)(m+2)(m+3)}. \tag{41}$$

For positive integers m, (41) is strictly increasing and bounded above by 1. In particular, when $m = 1$, (41) $= \frac{7}{12}$. Therefore, a brute-force upper bound is

$$\sum_{j=1}^{J} p_j^{\frac{2m^3+7m^2+10m+9}{2(m+1)(m+2)(m+3)}} \leq J$$

and thus

$$\begin{aligned}\text{Term A} &\lesssim_d \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{m}{m^2+4m+3}} M^2 n^{-\frac{1}{2m+4}} \sum_{j=1}^{J} p_j^{1-\frac{2m}{m^2+4m+3}-\frac{1}{2m+4}} \\ &\lesssim_d \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{m}{m^2+4m+3}} M^2\, J\, n^{-\frac{1}{2m+4}}.\end{aligned}$$

Note that the dependence of Term A on J is very mild. Indeed, if we denote

$$\alpha(m) = 1 - \frac{2m}{m^2+4m+3} - \frac{1}{2m+4},$$

then

$$\sum_{j=1}^{J} p_j^{\alpha(m)} \leq J^{\,1-\alpha(m)} \leq J^{\frac{2m}{m^2+4m+3}+\frac{1}{2m+4}} \leq J^{\frac{4}{m}},$$

since $\sum_j p_j = 1$. For large m, the exponent $\alpha(m)$ is close to 1, hence $\sum_j p_j^{\alpha(m)}$ remains essentially of order one. Consequently, the bound on Term A grows at most linearly with J, and in practice the J-dependence is negligible in high m. Here we use the power $4/m$ upper for clean format.

- **Bounding the Sampling Deviation Error (Term B):** Conditioned on the event $\mathcal{E}$, we have $|p_j - n_j/n| \leq \epsilon$ for all j. The empirical risk term is bounded because $\max\{|f(\boldsymbol{x})|, |y|\} \leq M$, which implies $|\frac{1}{n_j}\sum_{(\boldsymbol{x}_i,y_i)\in\mathcal{D}_j}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2| \leq 4M^2$. Thus, Term B is bounded by:

$$\text{Term B} \leq \sum_{j=1}^{J} \epsilon 4M^2 = 4M^2\epsilon = 4JM^2\sqrt{\frac{\log(4J/\delta)}{2n}}. \tag{42}$$

The total generalization gap is bounded by the sum of the bounds for Term A and Term B.

$$\text{Gap}_{\mathcal{P}}(f_{\boldsymbol{\theta}}; \mathcal{D}) \lesssim_d \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{m}{m^2+4m+3}} M^2\, J^{\frac{4}{m}}\, n^{-\frac{1}{2m+4}} + M^2\, J\sqrt{\frac{\log(4J/\delta)}{2n}}.$$

This completes the proof. □

G Generalization Upper Bounds: Isotropic Beta Family

In this section, the data generalization process is considered to be a family of isotropic Beta-radial distributions.

Definition G.1 (Isotropic Beta-radial distributions)**.** *Let $\boldsymbol{X}$ be a d-dimensional random vector in $\mathbb{R}^d$. For any $\alpha \in (0, \infty)$, the isotropic α-powered-radial distribution is defined by the generation process*

$$\boldsymbol{X} = h(R)\boldsymbol{U} \sim \mathcal{P}_{\boldsymbol{X}}(\alpha), \tag{43}$$

where $R \sim \text{Uniform}[0, 1]$ is a random variable drawn from a continuous uniform distribution on the interval $[0, 1]$, $\boldsymbol{U} \sim \text{Uniform}(\mathbb{S}^{d-1})$ is a random vector drawn uniformly from the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$ and $h(r) = 1 - (1 - r)^{1/\alpha}$ is a radial profile.

Lemma G.2. *Let $\mathcal{P}_{\boldsymbol{X}}(\alpha)$ be the isotropic α-powered-radial distribution in Definition G.1. For $\boldsymbol{X} \sim \mathcal{P}_{\boldsymbol{X}}(\alpha)$ any $t \in [0, 1]$, $\mathbb{P}(\|\boldsymbol{X}\| > 1 - t) = t^{\alpha}$. In particular, $\|\boldsymbol{X}\|_2$ is a $\text{Beta}(1, \alpha)$ distribution.*

Proof. The proof follows from a direct calculation based on the properties of the data-generating process.

First, the norm simplifies to: $\|\boldsymbol{X}\| = \|h(R)\boldsymbol{U}\| = h(R)$. Next, it is equivalent to calculating the probability that the scalar random variable $h(R)$ is greater than $1-t$:

$$\mathbb{P}\left(\|\boldsymbol{X}\| > 1-t\right) = \mathbb{P}\left(h(R) > 1-t\right).$$

To proceed, we need to apply the inverse of the function h to both sides of the inequality. The function $h(r)$ is monotonically increasing for $r \in [0,1]$, so applying its inverse preserves the direction of the inequality. Note that the inverse function is $h^{-1}(y) = 1-(1-y)^{\alpha}$.

Applying the inverse function h^{-1} to the inequality $h(R) > 1-t$, we get

$$R > h^{-1}(1-t).$$

Substituting the expression for h^{-1}:

$$R > 1-(1-(1-t))^{\alpha} = 1-t^{\alpha}.$$

Finally, we compute the probability of this event for the random variable R. By our initial assumption, R is uniformly distributed on the interval $[0,1]$, i.e., $R \sim \mathrm{Uniform}[0,1]$. The cumulative distribution function (CDF) of R is $F_R(x) = x$ for $x \in [0,1]$. The tail probability is therefore,

$$\mathbb{P}(R > x) = 1 - F_R(x) = 1 - x.$$

Applying this to our inequality $R > 1-t^{\alpha}$:

$$\mathbb{P}\left(R > 1-t^{\alpha}\right) = 1-(1-t^{\alpha}) = t^{\alpha}.$$

Combining all steps, we have rigorously shown that

$$\mathbb{P}\left(\|\boldsymbol{X}\| > 1-t\right) = t^{\alpha}.$$

To show that this implies $\|\boldsymbol{X}\|$ is a $\mathrm{Beta}(1,\alpha)$ distribution, we can examine its cumulative distribution function (CDF). Let $Y = \|\boldsymbol{X}\|$. The CDF is $F_Y(y) = \mathbb{P}(Y \le y)$. Substituting $y = 1-t$, we have $t = 1-y$. Then the tail probability becomes:

$$\mathbb{P}(\|\boldsymbol{X}\| > y) = (1-y)^{\alpha}.$$

From this, the CDF can be derived as

$$F_Y(y) = \mathbb{P}(\|\boldsymbol{X}\| \le y) = 1 - \mathbb{P}(\|\boldsymbol{X}\| > y) = 1-(1-y)^{\alpha}.$$

This is the characteristic CDF of a $\mathrm{Beta}(1,\alpha)$ distribution, thus completing the proof. □

Assumption G.3. *Fix $\alpha \in (0,\infty)$. Let $\mathcal{P}(\alpha)$ be a joint distribution over $\mathbb{R}^d \times \mathbb{R}$ such that The marginal distribution of the features x under $\mathcal{P}_{\boldsymbol{X}}(\alpha)$. The corresponding labels y are generated from a conditional distribution $\mathcal{P}(y|\boldsymbol{x})$ and are assumed to be bounded, i.e., $|y| \le D$ for some constant $D > 0$. Similarly, we define $\mathcal{P}_j(\boldsymbol{x}, y) = \mathcal{P}(\boldsymbol{x}, y \mid \boldsymbol{x} \in V_j)$.*

G.1 Characterization of the Weight Function for a Custom Radial Distribution

In this section, we analyze the properties of the weight function $g_\alpha(\boldsymbol{u},t) = g_{\mathcal{P}_{\boldsymbol{X},\alpha}}(\boldsymbol{u},t)$ with respect to the population distribution $\mathcal{P}_{\boldsymbol{X},\alpha}$ we defined in Definition G.1 and Assumption G.3. Recall that $g_\alpha(\boldsymbol{u},t) = \min\left(\tilde{g}_\alpha(\boldsymbol{u},t), \tilde{g}_\alpha(-\boldsymbol{u},-t)\right)$, where

$$\tilde{g}_\alpha(\boldsymbol{u},t) \coloneqq \mathbb{P}_{\mathcal{P}_{\boldsymbol{X},\alpha}}(\boldsymbol{X}^\top\boldsymbol{u} > t)^2 \cdot \mathbb{E}_{\mathcal{P}_{\boldsymbol{X},\alpha}}[\boldsymbol{X}^\top\boldsymbol{u} - t \mid \boldsymbol{X}^\top\boldsymbol{u} > t] \cdot \sqrt{1 + \left\|\mathbb{E}_{\mathcal{P}_{\boldsymbol{X},\alpha}}[\boldsymbol{X} \mid \boldsymbol{X}^\top\boldsymbol{u} > t]\right\|^2}. \tag{44}$$

Due to rotational symmetry, we analyze the projection $X_d = \boldsymbol{X}^\top\boldsymbol{e}_d$ without loss of generality. Our primary goal is to establish rigorous bounds on the tail probability $Q(t) \coloneqq \mathbb{P}(X_d > t)$ and the conditional expectation for t in a specific range close to 1.

Proposition G.4 (Tail Probability). *Let $\boldsymbol{X}$ be a random vector from the distribution defined above. Let X_d be its projection onto a fixed coordinate, and let its tail probability be $Q(t) = \mathbb{P}(X_d > t)$ for $t \in (-1, 1)$. Then there exists a fixed $t_0 \in [0, 1)$ such that for all $t \in [t_0, 1)$:*

$$c_2(\alpha, d)(1-t)^{\alpha+\frac{d-1}{2}} \leq Q(t) \leq c_3(\alpha, d)(1-t)^{\alpha+\frac{d-1}{2}},$$

where $c_2(\alpha, d)$ and $c_3(\alpha, d)$ are positive constants depending on α and d.

Proof. The tail probability is $Q(t) = \mathbb{P}(h(R)U_d > t)$. We compute this by integrating over the distribution of $R \sim \text{Uniform}[0, 1]$:

$$Q(t) = \int_{h^{-1}(t)}^{1} \mathbb{P}(U_d > t/h(r))\,\mathrm{d}r,$$

where the lower limit $h^{-1}(t) = 1 - (1-t)^\alpha$ ensures $h(r) > t$. The term $\mathbb{P}(U_d > x)$ is the normalized surface area of a spherical cap on $\mathbb{S}^{d-1}$. For $x \in [0, 1)$, this area can be bounded. Let $\theta_0 = \arccos(x)$. The area is proportional to $\int_0^{\theta_0} (\sin\phi)^{d-2}\,\mathrm{d}\phi$. For $\phi \in [0, \pi/2]$, we have $2\phi/\pi \leq \sin\phi \leq \phi$. This provides lower and upper bounds on the cap area

$$C_{d,L}(1-x)^{(d-1)/2} \leq \mathbb{P}(U_d > x) \leq C_{d,U}(1-x)^{(d-1)/2},$$

where $C_{d,L}$ and $C_{d,U}$ are constants depending on d. Let's apply this to our integral, substituting $x = t/h(r)$:

$$Q(t) \geq \int_{h^{-1}(t)}^{1} C_{d,L}\left(1 - \frac{t}{h(r)}\right)^{(d-1)/2} \mathrm{d}r.$$

We analyze this for $t \to 1^-$. Let $t = 1 - \epsilon$. The lower limit is $1 - \epsilon^\alpha$. For $r \in [1 - \epsilon^\alpha, 1]$, $h(r)$ is close to 1. Let's choose t_0 such that for $t \in [t_0, 1)$, $h(r) \geq h(t_0) > 1/2$. Then $h(r)$ is bounded away from 0. The term $1 - t/h(r) = (h(r) - t)/h(r)$. Let's bound the denominator: $h(t_0) \leq h(r) \leq 1$.

$$Q(t) \geq C_{d,L} \int_{1-\epsilon^\alpha}^{1} (h(r) - (1-\epsilon))^{(d-1)/2}\,\mathrm{d}r.$$

The integrand is $h(r) - (1-\epsilon) = \epsilon - (1-r)^{1/\alpha}$. The integral becomes:

$$\int_{1-\epsilon^\alpha}^{1} \left(\epsilon - (1-r)^{1/\alpha}\right)^{(d-1)/2} \mathrm{d}r.$$

Let $y = (1-r)^{1/\alpha}$, so $r = 1 - y^\alpha$ and $\mathrm{d}r = -\alpha y^{\alpha-1}\,\mathrm{d}y$. Limits for y are $[\epsilon, 0]$.

$$\int_{\epsilon}^{0} (\epsilon - y)^{(d-1)/2}(-\alpha y^{\alpha-1}\,\mathrm{d}y) = \alpha \int_0^{\epsilon} (\epsilon - y)^{(d-1)/2} y^{\alpha-1}\,\mathrm{d}y.$$

Let $y = \epsilon z,\ \mathrm{d}y = \epsilon\,\mathrm{d}z$. Limits for z are $[0, 1]$.

$$\alpha \int_0^1 (\epsilon - \epsilon z)^{(d-1)/2}(\epsilon z)^{\alpha-1}\epsilon\,\mathrm{d}z = \alpha\epsilon^{\alpha+\frac{d-1}{2}} B\left(\alpha, \frac{d+1}{2}\right).$$

Combining all constants, we establish the lower bound $Q(t) \geq c_2(\alpha, d)(1-t)^{\alpha+\frac{d-1}{2}}$. The upper bound follows an identical procedure, absorbing the $1/h(r)$ term into the constant $c_3(\alpha, d)$. □

Proposition G.5 (Conditional Expectation). *For $t \in [t_0, 1)$, the conditional expectation $\mathbb{E}[X_d \mid X_d > t]$ is bounded by*

$$1 - c_5(\alpha, d)(1-t) \leq \mathbb{E}[X_d \mid X_d > t] \leq 1 - c_4(\alpha, d)(1-t),$$

where $c_4(\alpha, d)$ and $c_5(\alpha, d)$ are positive constants.

Proof. We analyze $\mathbb{E}[1 - X_d \mid X_d > t] = \frac{1}{Q(t)}\int_t^1 (1-s) f_{X_d}(s)\,\mathrm{d}s$, where $f_{X_d}(s) = -Q'(s)$. From Proposition G.4, we know $f_{X_d}(s) \propto (1-s)^{\alpha+\frac{d-3}{2}}$. The numerator is:

$$N(t) = \int_t^1 (1-s) f_{X_d}(s)\,\mathrm{d}s.$$

Bounding the constant of proportionality for $f_{X_d}(s)$ by $c_{1,L}$ and $c_{1,U}$:

$$c_{1,L}\int_t^1 (1-s)^{\alpha+\frac{d-1}{2}}\,\mathrm{d}s \le N(t) \le c_{1,U}\int_t^1 (1-s)^{\alpha+\frac{d-1}{2}}\,\mathrm{d}s.$$

The integral evaluates to $\frac{(1-t)^{\alpha+\frac{d+1}{2}}}{\alpha+\frac{d+1}{2}}$. So, $N(t) \propto (1-t)^{\alpha+\frac{d+1}{2}}$. Dividing $N(t)$ by $Q(t) \propto (1-t)^{\alpha+\frac{d-1}{2}}$, we get:

$$\mathbb{E}[1-X_d \mid X_d > t] \propto \frac{(1-t)^{\alpha+\frac{d+1}{2}}}{(1-t)^{\alpha+\frac{d-1}{2}}} = 1-t.$$

By carefully tracking the constants c_2, c_3 from Proposition G.4 and the constants from the integration of $f_{X_d}(s)$, we can construct explicit (though complex) expressions for c_4 and c_5 that provide rigorous two-sided bounds for t in the specified range $[t_0, 1)$. □

Proposition G.6 (Asymptotic Behavior of $g_\alpha^+(t)$)**.** *Let the function $g_\alpha^+(t)$ be defined as in (44). Then for $t \in [t_0, 1)$, we have:*

$$c_L^{(g)}(\alpha,d)(1-t)^{2\alpha+d} \le g_\alpha^+(t) \le c_U^{(g)}(\alpha,d)(1-t)^{2\alpha+d},$$

where $c_L^{(g)}(\alpha,d)$ and $c_U^{(g)}(\alpha,d)$ are positive constants.

Proof. Let $Q(t) = \mathbb{P}(X_d > t)$ and $E(t) = \mathbb{E}[X_d \mid X_d > t]$. The function is $g_\alpha^+(t) = Q(t)^2 \cdot (E(t)-t) \cdot \sqrt{1+E(t)^2}$. We establish bounds for $t \in [t_0, 1)$ for a sufficiently large t_0.

1. **Bounds for** $Q(t)^2$: From Proposition G.4, we have:

   $$(c_2(\alpha,d))^2(1-t)^{2\alpha+d-1} \le Q(t)^2 \le (c_3(\alpha,d))^2(1-t)^{2\alpha+d-1}.$$

   Let $A_L(\alpha,d) = (c_2(\alpha,d))^2$ and $A_U(\alpha,d) = (c_3(\alpha,d))^2$.

2. **Bounds for** $E(t)-t$: This is $\mathbb{E}[X_d - t \mid X_d > t]$. From Proposition G.5, we have $(1-t) - c_5(1-t) \le E(t)-t \le (1-t) - c_4(1-t)$. This gives:

   $$B_L(\alpha,d)(1-t) \le E(t)-t \le B_U(\alpha,d)(1-t),$$

   where $B_L(\alpha,d) = 1-c_5(\alpha,d)$ and $B_U(\alpha,d) = 1-c_4(\alpha,d)$. We can choose t_0 close enough to 1 to ensure these constants are positive.

3. **Bounds for** $\sqrt{1+E(t)^2}$: For $t \in [t_0, 1)$, we have $t_0 \le t < E(t) \le 1$. By choosing, for instance, $t_0 = 3/4$, we have $3/4 \le E(t) \le 1$. Thus,

   $$\sqrt{1+(3/4)^2} \le \sqrt{1+E(t)^2} \le \sqrt{1+1^2}.$$

   This gives constant bounds $C_L = 5/4$ and $C_U = \sqrt{2}$.

Combining these three bounds, for $t \in [t_0, 1)$:

$$A_L B_L C_L (1-t)^{2\alpha+d-1}(1-t) \le g_\alpha^+(t) \le A_U B_U C_U (1-t)^{2\alpha+d-1}(1-t).$$

This simplifies to the final result:

$$c_L^{(g)}(\alpha,d)(1-t)^{2\alpha+d} \le g_\alpha^+(t) \le c_U^{(g)}(\alpha,d)(1-t)^{2\alpha+d},$$

where the bounding constants are given by $c_L^{(g)}(\alpha,d) = A_L B_L C_L$ and $c_U^{(g)}(\alpha,d) = A_U B_U C_U$. □

G.2 Proof of Theorem 3.5

Theorem G.7 (Restate Theorem 3.5)**.** *Fix a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where each $(\boldsymbol{x}_i, y_i)$ is drawn i.i.d. from $\mathcal{P}(\alpha)$ defined in Assumption 3.4. Then, with probability at least $1-\delta$, for any $f_{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})$,*

$$\mathrm{Gap}_{\mathcal{P}}(f_{\boldsymbol{\theta}}; \mathcal{D}) \lessapprox_d \begin{cases} \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{\alpha d}{d^2+4d+3}} M^{\frac{2d^2+7\alpha d+6\alpha}{d^2+4\alpha d+3\alpha}} n^{-\frac{\alpha(d+3)}{2(d^2+4\alpha d+3\alpha)}}, & \alpha \geq \frac{3d}{2d-3}; \\ \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{\alpha d}{d^2+4d+3}} M^{\frac{2d^2+7\alpha d+6\alpha}{d^2+4\alpha d+3\alpha}} n^{-\frac{\alpha}{2d+4\alpha}}, & \alpha < \frac{3d}{2d-3}, \end{cases} \tag{45}$$

and for where $M \coloneqq \max\{D, \|f_{\boldsymbol{\theta}}\|_{L^\infty(\mathbb{B}_1^d)}, 1\}$ and $\lessapprox_d$ hides constants (which could depend on d) and logarithmic factors in n and $(1/\delta)$.

Proof. For convenience, we let $A = \frac{1}{\eta} - \frac{1}{2} + 4M$ and we have that

$$f_{\boldsymbol{\theta}} \in \mathcal{F}_{g_{\mathcal{D}}}(\mathbb{B}_1^{V_j}; M, C), \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta; \mathcal{D}).$$

For any fixed $\varepsilon < 1$, we may decompose $\mathbb{B}_1^d$ into *ε-annulus* $\mathbb{A}_\varepsilon^d \coloneqq \{\boldsymbol{x} \in \mathbb{B}_1^d \mid \|\boldsymbol{x}\|_2 \geq 1-\varepsilon\}$ and the closure of its complement is called *ε-strict interior*denoted by $\mathbb{I}_\varepsilon^d$.

$$\mathbb{B}_1^d = \mathbb{A}_\varepsilon^d \cup \mathbb{I}_\varepsilon^d.$$

According to the law of total expectation, the population risk is decomposed into

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{P}}\left[(f(\boldsymbol{x}) - y)^2\right] = \mathbb{P}(\boldsymbol{x} \in \mathbb{A}_\varepsilon^d) \cdot \mathbb{E}_{\mathbb{A}}\left[(f(\boldsymbol{x}) - y)^2\right] + \mathbb{P}(\boldsymbol{x} \in \mathbb{I}_\varepsilon^d) \cdot \mathbb{E}_{\mathbb{I}}\left[(f(\boldsymbol{x}) - y)^2\right], \tag{46}$$

where $\mathbb{E}_{\mathbb{A}}$ means that $\{\boldsymbol{x}, y\}$ is a new sample from the data distribution conditioned on $\boldsymbol{x} \in \mathbb{A}_\varepsilon^d$ and $\mathbb{E}_{\mathbb{I}}$ means that $(\boldsymbol{x}, y)$ is a new sample from the data distribution conditioned on $\boldsymbol{x} \in \mathbb{I}_\varepsilon^d$.

Similarly, we also have this decomposition for empirical risk

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{x}_i) - y_i)^2 &= \frac{1}{n}\left(\sum_{i\in I}(f(\boldsymbol{x}_i) - y_i)^2 + \sum_{j\in A}(f(\boldsymbol{x}_i) - y_i)^2\right) \\ &= \frac{n_I}{n}\frac{1}{n_I}\sum_{i\in I}(f(\boldsymbol{x}_i) - y_i)^2 + \frac{n_A}{n}\frac{1}{n_A}\sum_{j\in A}(f(\boldsymbol{x}_i) - y_i)^2, \end{aligned} \tag{47}$$

where I is the set of data points with $\boldsymbol{x}_i \in \mathbb{I}_\varepsilon^d$ and A is the set of data points with $\boldsymbol{x}_i \in \mathbb{A}_\varepsilon^d$. Then the generalization gap can be decomposed into

$$|R(f) - \widehat{R}_{\mathcal{D}}(f)| \leq \mathbb{P}(\boldsymbol{x} \in \mathbb{A}_\varepsilon^d) \cdot \mathbb{E}_{\mathbb{A}}\left[(f_{\boldsymbol{\theta}}(\boldsymbol{x}) - y)^2\right] + \frac{n_A}{n}\frac{1}{n_A}\sum_{j\in A}(f(\boldsymbol{x}_i) - y_i)^2 \tag{48}$$

$$+ \left|\mathbb{P}(\boldsymbol{x} \in \mathbb{I}_\varepsilon^d) - \frac{n_I}{n}\right| \frac{1}{n_I}\sum_{i\in I}(f(\boldsymbol{x}_i) - y_i)^2 \tag{49}$$

$$+ \mathbb{P}(\boldsymbol{x} \in \mathbb{I}_\varepsilon^d) \cdot \left|\mathbb{E}_{\mathbb{I}}\left[(f(\boldsymbol{x}) - y)^2\right] - \frac{1}{n_I}\sum_{i\in I}(f(\boldsymbol{x}_i) - y_i)^2\right|. \tag{50}$$

Using the property that the marginal distribution of $\boldsymbol{x}$ is $\mathrm{Uniform}(\mathbb{B}_1^d)$ and its concentration property, with probability at least $1-\delta$,

$$(48) \lessapprox_d O(M^2\varepsilon^\alpha), \tag{51}$$

where $\lessapprox_d$ hides the constants that could depend on d and logarithmic factors of $1/\delta$.

For the term (49), with probability $1-\delta$

$$\begin{cases} \left|\mathbb{P}(\boldsymbol{x} \in \mathbb{I}_\varepsilon^d) - \frac{n_I}{n}\right| & \lesssim \sqrt{\frac{\varepsilon^\alpha \log(1/\delta)}{n}}, \\ \frac{1}{n_I}\sum_{i\in I}(f(\boldsymbol{x}_i) - y_i)^2 & \leq 4M^2 \end{cases} \tag{52}$$

so we may also conclude that

$$(49) \lesssim M^2 \sqrt{\frac{\varepsilon \log(1/\delta)}{n}} \tag{53}$$

For the part of the interior (50), the scalar $\mathbb{P}(\boldsymbol{x} \in \mathbb{I}_\varepsilon^d)$ is less than 1 with high-probability. Therefore, we just need to deal with the term

$$\mathbb{E}_{\mathbb{I}}\left[(f(\boldsymbol{x}) - y)^2\right] - \frac{1}{n_I}\sum_{i \in I}(f(\boldsymbol{x}_i) - y_i)^2. \tag{54}$$

Since both the distribution and sample points only support in $\mathbb{I}_\varepsilon^d$, we may consider f by its restrictions in $\mathbb{I}_\varepsilon^d$, which are denoted by f^ε. Furthermore, according to the definition, we have

$$\begin{aligned} f(\boldsymbol{x}) &= \int_{\mathbb{S}^{d-1}\times[-1,1]} \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\nu(\boldsymbol{u},t) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + b \\ &= \int_{\mathbb{S}^{d-1}\times[-1+\varepsilon,1-\varepsilon]} \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\nu(\boldsymbol{u},t) + \underbrace{\int_{\mathbb{S}^{d-1}\times[-1,-1+\varepsilon)\cup(1-\varepsilon,1]} \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\nu(\boldsymbol{u},t)}_{\text{Annulus ReLU}} \\ &\quad + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + b \end{aligned} \tag{55}$$

where the Annulus ReLU term is totally linear in the strictly interior i.e. there exists $\boldsymbol{c}', b'$ such that

$$\boldsymbol{c}'^\mathsf{T}\boldsymbol{x} + b' = \int_{\mathbb{S}^{d-1}\times[-1,-1+\varepsilon)\cup(1-\varepsilon,1]} \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\nu(\boldsymbol{u},t), \quad \forall \boldsymbol{x} \in \mathbb{I}_\varepsilon^d. \tag{56}$$

Therefore, we may write

$$f(\boldsymbol{x}) = f^\varepsilon(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-1+\varepsilon,1-\varepsilon]} \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\nu(\boldsymbol{u},t) + (\boldsymbol{c}+\boldsymbol{c}')^\mathsf{T}\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{b'}, \quad \boldsymbol{x} \in \mathbb{I}_\varepsilon^d. \tag{57}$$

According to the definition, we have that

$$|f^\varepsilon|_{\mathrm{V}(\mathbb{I}_\varepsilon^d)} \le \int_{\mathbb{S}^{d-1}\times[-1+\varepsilon,1-\varepsilon]} |\,\mathrm{d}\nu|. \tag{58}$$

From empirical process we discussed in Section E.2, especically Theorem E.6, we know that with probability at least $1-\delta$,

$$\sup_{\boldsymbol{u},t} |g_{\mathcal{D}}(\boldsymbol{u},t) - g_\alpha(\boldsymbol{u},t)| \lesssim_d \sqrt{\frac{d + \log(2/\delta)}{n}} =: \epsilon_n. \tag{59}$$

This implies a lower bound on the empirical minimum weight in the core with probability at least $1 - \delta/3$,

$$g_{\mathcal{D},\min} = \inf_{|t|\le 1-\varepsilon} g_{\mathcal{D}}(\boldsymbol{u},t) \ge \inf_{|t|\le 1-\varepsilon} g_\alpha(\boldsymbol{u},t) - \epsilon_n = g_{\alpha,\min} - \epsilon_n. \tag{60}$$

Here, $g_{\alpha,\min} \asymp \varepsilon^{d+2\alpha}$ is the minimum of the population weight function in the core.

For the bound $|f^\varepsilon|_{\mathrm{V}} \le A/g_{\mathcal{D},\min} \le A/(g_{\alpha,\min} - \epsilon_n)$ to be meaningful with high probability, we must operate in a regime where $g_{\alpha,\min} \ge \epsilon_n$. We enforce a stricter **validity condition** for our proof

$$g_{\alpha,\min} \ge 2\epsilon_n \quad \Longrightarrow \quad \varepsilon^{d+2\alpha} \gtrsim_d \sqrt{\frac{d + \log(6/\delta)}{n}}. \tag{61}$$

Under this condition, we have $g_{\mathcal{D},\min} \ge g_{\alpha,\min} - \epsilon_n \ge g_{\alpha,\min}/2 \asymp \varepsilon^{d+2\alpha}$. Thus, for any $f \in \boldsymbol{\Theta}_{\mathrm{BEoS}}(\eta, \mathcal{D})$, its restriction f^ε has a controlled unweighted variation norm with high probability:

$$|f^\varepsilon|_{\mathrm{V}(\mathbb{B}_{1-\varepsilon}^d)} \le \frac{A}{g_{\mathcal{D},\min}} \le \frac{A}{g_{\alpha,\min}/2} \asymp \frac{A}{\varepsilon^{d+2\alpha}} =: C_\varepsilon.$$

According to the assumption, we have that $|f|_{\mathrm{V}_g(\mathbb{B}_1^d)} \le A$, and thus we have

$$\int_{\mathbb{S}^{d-1}\times[-1+\varepsilon,1-\varepsilon]} g_{\mathcal{D}}|\,\mathrm{d}\nu| \le \int_{\mathbb{S}^{d-1}\times[-1,1]} g_{\mathcal{D}}|\,\mathrm{d}\nu| \le A. \tag{62}$$

Suppose the validity condition (61) holds (we will verify it later), we have $g(\boldsymbol{u},t) \gtrapprox_d \varepsilon^{d+2\alpha}$ when $t \le 1-\varepsilon$ with probability $1-\delta/3$, we may use (62) to deduce that

$$\varepsilon^{d+2\alpha} \cdot \int_{\mathbb{S}^{d-1}\times[-1+\varepsilon,1-\varepsilon]} |\,\mathrm{d}\nu| \le \int_{\mathbb{S}^{d-1}\times[-1+\varepsilon,1-\varepsilon]} g_{\mathcal{D}} |\,\mathrm{d}\nu| \le A. \tag{63}$$

Combining (58) and (63), we deduce that

$$|f^{\varepsilon}|_{\mathrm{V}(\mathbb{B}^d_{1-\varepsilon})} \lessapprox_d \frac{A}{\varepsilon^{d+2\alpha}} =: C.$$

Therefore, we may leverage Lemma D.11 to $f^{\varepsilon} \in \mathrm{V}_C(\mathbb{B}^d_{1-\varepsilon})$, we may conclude that with probability at least $1-\delta$,

$$(50) \lessapprox_d C^{\frac{d}{2d+3}} M^{\frac{3(d+2)}{2d+3}} n^{-\frac{d+3}{4d+6}}, \tag{64}$$

where $\lessapprox_d$ hides the constants that could depend on d and logarithmic factors of $1/\delta$.

Now we combine the upper bounds (51), (53) and (64) to deduce an upper bound of the generalization gap. We have for any fixed $\epsilon > 0$, with probability $1-\delta$,

$$|R(f) - \widehat{R}_{\mathcal{D}}(f)| \lessapprox_d M^2\varepsilon^{\alpha} + \left(\frac{A}{\varepsilon^{d+2\alpha}}\right)^{\frac{d}{2d+3}} M^{\frac{3(d+2)}{2d+3}} n^{-\frac{d+3}{4d+6}}. \tag{65}$$

Then we may choose the optimal ε^* such that

$$M^2(\varepsilon^*)^{\alpha} = \left(\frac{A}{(\varepsilon^*)^{d+2\alpha}}\right)^{\frac{d}{2d+3}} M^{\frac{3(d+2)}{2d+3}} n^{-\frac{d+3}{4d+6}}$$

and by direct computation, we get

$$\varepsilon^* = \left(A^{\frac{d}{d^2+4\alpha d+3\alpha}} M^{-\frac{d}{d^2+4\alpha d+3\alpha}} n^{-\frac{d+3}{2(d^2+4\alpha d+3\alpha)}}\right).$$

To satisfy the validity condition (61), we require

$$(\varepsilon^*)^{d+2\alpha} = O\left(n^{-\frac{d+3}{2(d^2+4\alpha d+3\alpha)}}\right)^{d+2\alpha} \ge \tilde{O}(n^{-\frac{1}{2}}). \tag{66}$$

By adjusting some universal constants, it suffices to show whether

$$\frac{(d+3)(d+2\alpha)}{2(d^2+4\alpha d+3\alpha)} < \frac{1}{2}. \tag{67}$$

After direct computation, (67) is equivalent to $\alpha \in (\frac{3d}{2d-3}, \infty)$. With this assumption, we may evaluate the optimal ε^* in the inequality (65) to deduce the optimal results that

$$|R(f) - \widehat{R}_n(f)| \lessapprox_d \left(\frac{1}{\eta} - \frac{1}{2} + 4M\right)^{\frac{\alpha d}{d^2+4d+3}} M^{\frac{2d^2+7\alpha d+6\alpha}{d^2+4\alpha d+3\alpha}} n^{-\frac{\alpha(d+3)}{2(d^2+4\alpha d+3\alpha)}}. \tag{68}$$

In the case where $\alpha \le \frac{3d}{d+2\alpha}$, we set

$$\varepsilon^* = \tilde{O}\left(n^{-\frac{1}{2d+4\alpha}}\right)$$

and adjust some universal constant to satisfy the validaty condition. Then (65) has the form

$$|R(f) - \widehat{R}_{\mathcal{D}}(f)| \le \tilde{O}\left(n^{-\frac{2\alpha}{2d+4\alpha}}\right) + \tilde{O}\left(n^{-\frac{3}{4d+6}}\right).$$

Then assumption $\alpha < \frac{3d}{d+2\alpha}$ implies that $n^{-\frac{2\alpha}{2d+4\alpha}} > n^{-\frac{3}{4d+6}}$ and thus

$$|R(f) - \widehat{R}_{\mathcal{D}}(f)| \le \tilde{O}\left(n^{-\frac{2\alpha}{2d+4\alpha}}\right).$$

Note that the other constants in the front of $1/n$ does not change, so we finish the proof. □

H GENERALIZATION GAP LOWER BOUND VIA POISSONIZATION

This section provides a self-contained proof for a lower bound on the generalization gap in a noiseless setting. We employ the indistinguishability method, where the core technical challenge is to construct two functions that are identical on a given training sample yet significantly different in population. The Poissonization technique is the key tool that simplifies the probabilistic analysis required to guarantee the existence of such a pair. The paradigm is almost the same as the one in (Liang et al., 2025, Appendix H & I), but the assumption on distributions are different.

H.1 CONSTRUCTION OF "HARD-TO-LEARN" NETWORKS

Our strategy relies on functions localized on small, disjoint regions near the boundary of the unit ball. We first establish key geometric properties of these regions, called spherical caps. Let $\boldsymbol{u} \in \mathbb{S}^{d-1}$ be a unit vector. Let $\varepsilon \in \mathbb{R}_+$ be a constant with $\varepsilon \leq 1/2$. Consider the ReLU atom:

$$\varphi_{\boldsymbol{u},\varepsilon^2}(\boldsymbol{x}) = \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - (1-\varepsilon^2)). \tag{69}$$

Lemma H.1. *The $L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))$-norm of $\varphi_{\boldsymbol{u},\varepsilon^2}$, where the measure $\mathcal{P}_{\boldsymbol{X}}(\alpha)$ is defined in Definition G.1, is given by*

$$c_L(d,\alpha)\varepsilon^{\frac{d+3+2\alpha}{2}} \leq \|\varphi_{\boldsymbol{u},\varepsilon^2}\|_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} \leq c_U(d,\alpha)\varepsilon^{\frac{d+3+2\alpha}{2}}, \tag{70}$$

where $c_L(d,\alpha)$ and $c_U(d,\alpha)$ are constants that depend on the dimension d and the parameter α.

Before the formal proof, we offer a geometric justification for the result. The squared norm is an integral of $(\phi(\dots))^2$, and we can estimate its value as the product of the integrand's average magnitude and the measure of the small domain where it is non-zero. We estimate the measure of this "active" domain, where $r\boldsymbol{u}^\mathsf{T}\boldsymbol{U} > 1-\varepsilon^2$, using a polar coordinate perspective.

- **Integrand's Magnitude:** Within the active domain, the term $r\boldsymbol{u}^\mathsf{T}\boldsymbol{U} - (1-\varepsilon^2)$ represents the positive "height" above the activation threshold. This height varies from 0 to a maximum on the order of $O(\varepsilon^2)$. A reasonable estimate for the squared term's average value is thus $O((\varepsilon^2)^2) = O(\varepsilon^4)$.
- **Measure of the Domain:** We decompose the domain's volume into radial and angular parts.
  - **Radial Measure:** The condition requires the radius r to be near 1. For the $\mathcal{P}_{\boldsymbol{X}}(\alpha)$ distribution, this confines r to a region of length $\Delta r \sim O(\varepsilon^{2\alpha})$.
  - **Angular Measure:** The vector $\boldsymbol{U}$ is confined to a small spherical cap around $\boldsymbol{u}$. A cap defined by a "height" of $h \sim \mathcal{O}(\varepsilon^2)$ has a surface area on $\mathbb{S}^{d-1}$ of order $\mathcal{O}(h^{(d-1)/2})$. This gives an angular measure of $\Delta\Omega \sim O((\varepsilon^2)^{(d-1)/2}) = O(\varepsilon^{d-1})$.

Combining these estimates, the squared norm I scales as the product of the integrand's magnitude and the two components of the domain's measure:

$$I \approx \underbrace{O(\varepsilon^4)}_{\text{Integrand}} \times \underbrace{O(\varepsilon^{2\alpha})}_{\text{Radial}} \times \underbrace{O(\varepsilon^{d-1})}_{\text{Angular}} = \mathcal{O}(\varepsilon^{d+3+2\alpha}).$$

Taking the square root provides the claimed scaling for the L^2-norm. The formal proof makes this geometric heuristic rigorous.

Proof. The squared L^2 norm of $\varphi_{\boldsymbol{u},\varepsilon^2}$ over the distribution $\mathcal{P}_{\boldsymbol{X}}(\alpha)$ is defined by the expectation

$$I = \|\varphi_{\boldsymbol{u},\varepsilon^2}\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} = \mathbb{E}_{\boldsymbol{X}\sim\mathcal{P}_{\boldsymbol{X}}(\alpha)}\left[|\varphi_{\boldsymbol{u},\varepsilon^2}(\boldsymbol{X})|^2\right]$$

Substituting the definition of $\varphi_{\boldsymbol{u},\varepsilon^2}(\boldsymbol{x})$ and using the property of the ReLU function, we get

$$\begin{aligned} I &= \mathbb{E}_{R,\boldsymbol{U}}\left[\left(\phi(h(R)\boldsymbol{u}^\mathsf{T}\boldsymbol{U} - (1-\varepsilon^2))\right)^2\right] \\ &= \mathbb{E}_{R,\boldsymbol{U}}\left[\mathbb{1}_{\{h(R)\boldsymbol{u}^\mathsf{T}\boldsymbol{U} > 1-\varepsilon^2\}}(h(R)\boldsymbol{u}^\mathsf{T}\boldsymbol{U} - (1-\varepsilon^2))^2\right] \end{aligned} \tag{71}$$

where $R \sim \text{Uniform}[0,1]$ and $\boldsymbol{U} \sim \text{Uniform}(\mathbb{S}^{d-1})$.

Due to the rotational symmetry of the distribution of $\boldsymbol{U}$, we can perform a rotation of the coordinate system such that $\boldsymbol{u}$ aligns with the d-th standard basis vector $\boldsymbol{e}_d = (0,\dots,0,1)$ without changing the value of the integral. In these new coordinates, $\boldsymbol{u}^\mathsf{T}\boldsymbol{U} = U_d$. The expectation becomes an iterated integral:

$$I = \int_0^1 \mathbb{E}_{\boldsymbol{U}}\left[\mathbb{1}_{\{h(r)U_d > 1-\varepsilon^2\}}(h(r)U_d - (1-\varepsilon^2))^2\right] \mathrm{d}r$$

Let $z = U_d$. The probability density function of z is $p(z) = C_d(1-z^2)^{(d-3)/2}$ for $z \in [-1,1]$, where $C_d = \frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}$. The integral is non-zero only if $h(r) > 1-\varepsilon^2$, which implies $r > 1-\varepsilon^{2\alpha}$.

$$I = C_d \int_{1-\varepsilon^{2\alpha}}^1 \int_{\frac{1-\varepsilon^2}{h(r)}}^1 (h(r)z - (1-\varepsilon^2))^2 (1-z^2)^{\frac{d-3}{2}} \,\mathrm{d}z\,\mathrm{d}r \tag{72}$$

We perform a change of variable $z = 1-t$, so $\mathrm{d}z = -\mathrm{d}t$ and the integration limits change from $[\frac{1-\varepsilon^2}{h(r)}, 1]$ to $[1-\frac{1-\varepsilon^2}{h(r)}, 0]$.

$$\begin{aligned}\text{Inner integration of (72)} &= C_d \int_{1-\frac{1-\varepsilon^2}{h(r)}}^0 (h(r)(1-t) - (1-\varepsilon^2))^2 (1-(1-t)^2)^{\frac{d-3}{2}}(-\mathrm{d}t) \\ &= C_d \int_0^{t_0(r)} (h(r)t_0(r) - h(r)t)^2 (2t-t^2)^{\frac{d-3}{2}} \,\mathrm{d}t\end{aligned} \tag{73}$$

where $t_0(r) = 1-\frac{1-\varepsilon^2}{h(r)} = \frac{h(r)-1+\varepsilon^2}{h(r)}$. Since $r \in [1-\varepsilon^{2\alpha}, 1]$ and for small ε, $h(r)$ is close to 1, we know $t_0(r)$ is small. For a sufficiently small ε, we can ensure $t \le t_0(r) < 1/4$. Thus, we can bound the term $2-t$ as $7/4 \le 2-t \le 2$. This gives bounds on $(2t-t^2)^{(d-3)/2} = ((2-t)t)^{(d-3)/2}$:

$$\left(\frac{7}{4}\right)^{\frac{d-3}{2}} t^{\frac{d-3}{2}} \le (2t-t^2)^{\frac{d-3}{2}} \le 2^{\frac{d-3}{2}} t^{\frac{d-3}{2}}$$

The integral I is therefore bounded by:

$$\underline{C_d} \int_{1-\varepsilon^{2\alpha}}^1 J(r)\,\mathrm{d}r \le I \le \overline{C_d} \int_{1-\varepsilon^{2\alpha}}^1 J(r)\,\mathrm{d}r \tag{74}$$

where $\underline{C_d}, \overline{C_d}$ are new constants and $J(r) = \int_0^{t_0(r)} (h(r)t_0(r) - h(r)t)^2 t^{\frac{d-3}{2}}\,\mathrm{d}t$.

Consider the integral $J(r)$ and change variable by setting $t = t_0(r)s$, then $\mathrm{d}t = t_0(r)\,\mathrm{d}s$.

$$\begin{aligned}J(r) &= \int_0^1 (h(r)t_0(r) - h(r)t_0(r)s)^2 (t_0(r)s)^{\frac{d-3}{2}} (t_0(r)\,\mathrm{d}s) \\ &= (h(r)t_0(r))^2 (t_0(r))^{\frac{d-3}{2}} t_0(r) \int_0^1 (1-s)^2 s^{\frac{d-3}{2}}\,\mathrm{d}s \\ &= h(r)^2 (t_0(r))^{\frac{d+3}{2}} \underbrace{\left(\int_0^1 (1-s)^2 s^{\frac{d-3}{2}}\,\mathrm{d}s\right)}_{\text{constant}}\end{aligned} \tag{75}$$

To analyze $t_0(r)^{\frac{d+3}{2}}$, we let $r = 1-\delta$, so $\mathrm{d}r = -\mathrm{d}\delta$ and the integration limits for δ are $[\varepsilon^{2\alpha}, 0]$. $h(r) = 1-(1-(1-\delta))^{1/\alpha} = 1-\delta^{1/\alpha}$. As $\delta \to 0$, $h(r) \to 1$. $t_0(r) = \frac{(1-\delta^{1/\alpha})-1+\varepsilon^2}{1-\delta^{1/\alpha}} = \frac{\varepsilon^2-\delta^{1/\alpha}}{1-\delta^{1/\alpha}}$. For small δ, $1-\delta^{1/\alpha}$ is close to 1, providing upper and lower bounds. Thus I is bounded by integrals of the form

$$C\int_{\varepsilon^{2\alpha}}^0 \left(\varepsilon^2 - \delta^{1/\alpha}\right)^{\frac{d+3}{2}} (-\mathrm{d}\delta) = C\int_0^{\varepsilon^{2\alpha}} \left(\varepsilon^2 - \delta^{1/\alpha}\right)^{\frac{d+3}{2}} \mathrm{d}\delta$$

for some mild constant C.

Now we perform a new change-of-variable by setting $\delta^{1/\alpha} = \varepsilon^2 v$. This gives $\delta = (\varepsilon^2 v)^\alpha = \varepsilon^{2\alpha} v^\alpha$ and $\mathrm{d}\delta = \alpha\varepsilon^{2\alpha} v^{\alpha-1}\,\mathrm{d}v$. The limits for v become

$$\begin{aligned}\int_0^{\varepsilon^{2\alpha}} \left(\varepsilon^2 - \delta^{1/\alpha}\right)^{\frac{d+3}{2}}\,\mathrm{d}\delta &= \int_0^1 (\varepsilon^2 - \varepsilon^2 v)^{\frac{d+3}{2}} (\alpha\varepsilon^{2\alpha} v^{\alpha-1}\,\mathrm{d}v) \\ &= (\varepsilon^2)^{\frac{d+3}{2}} \varepsilon^{2\alpha} \int_0^1 (1-v)^{\frac{d+3}{2}} \alpha v^{\alpha-1}\,\mathrm{d}v \\ &= \varepsilon^{d+3+2\alpha} \underbrace{\left(\alpha \int_0^1 (1-v)^{\frac{d+3}{2}} v^{\alpha-1}\,\mathrm{d}v\right)}_{\text{constant}}\end{aligned} \tag{76}$$

The squared norm I is bounded by constants times $\varepsilon^{d+3+2\alpha}$. The L^2-norm is the square root of I:

$$c_L(d,\alpha)\,\varepsilon^{\frac{d+3+2\alpha}{2}} \le \left\|\varphi_{\boldsymbol{u},\varepsilon^2}\right\|_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} = \sqrt{I} \le c_U(d,\alpha)\,\varepsilon^{\frac{d+3+2\alpha}{2}} \tag{77}$$

where $c_9(d,\alpha)$ and $c_{10}(d,\alpha)$ are constants that absorb all factors depending on d and α from the bounds established in the derivation. This completes the proof. □

Lemma H.2 (Cap mass at angular scale ε)**.** *For $\varepsilon \in (0, \frac{1}{2}]$ and $\boldsymbol{u} \in \mathbb{S}^{d-1}$, define the thin cap*

$$C(\boldsymbol{u},\varepsilon) = \{x \in \mathbb{B}_1^d :\ u^\mathsf{T} x > 1 - \varepsilon^2\}.$$

There exist constants depending only on (d,α), such that $\mathcal{P}_{\boldsymbol{X}}\big(C(\boldsymbol{u},\varepsilon)\big) \asymp \varepsilon^{\,d-1+2\alpha}$.

Sketch proof. The result and the proof almost the same as the ones about Lemma H.1. We omit the calculation detials. □

Lemma H.3 (Disjoint Cap Packing)**.** *For any $\varepsilon \in (0, 1/2]$, there exists a set of N unit vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\} \subset \mathbb{S}^{d-1}$, with $N \asymp \varepsilon^{-(d-1)}$, such that the caps $\{C(\boldsymbol{u}_i,\varepsilon)\}_{i=1}^N$ are pairwise disjoint.*

Sketch proof. The angular radius of the cap $C(\boldsymbol{u},\varepsilon)$ is $\vartheta = \arccos(1-\varepsilon^2) \asymp \varepsilon$. For two caps to be disjoint, the angular separation between their centers must be at least 2ϑ. The maximum number of such points is the packing number $M(\mathbb{S}^{d-1}, 2\vartheta)$. A standard volumetric argument provides the upper bound $M(\mathbb{S}^{d-1}, 2\vartheta) = O(\varepsilon^{-(d-1)})$. The lower bound is established by relating the packing number to the covering number $N(\mathbb{S}^{d-1}, \alpha)$, which is known to scale as $N(\mathbb{S}^{d-1}, \alpha) \asymp \alpha^{-(d-1)}$, thus yielding the asserted scaling for N. □

We now formally establish the family of functions used to construct the adversarial pair. This family resides within a function class $\mathcal{F}_g(\mathbb{B}_1^d; 1, 1)$ and is built upon normalized ReLU atoms localized on the disjoint spherical caps.

Construction H.4 (Adversarial Function Family)**.** *Recall that $\varphi_{\boldsymbol{u},\varepsilon^2}(\boldsymbol{x}) = \phi(\boldsymbol{u}^\mathsf{T}\boldsymbol{x} - (1-\varepsilon^2))$. We define its normalized version as $\Phi_{\boldsymbol{u},\varepsilon^2} := \varepsilon^{-2}\varphi_{\boldsymbol{u},\varepsilon^2}$. By construction, $\|\Phi_{\boldsymbol{u},\varepsilon^2}\|_{L^\infty(\mathbb{B}_1^d)} \le 1$ and $\left\|\Phi_{\boldsymbol{u},\varepsilon^2}\right\|_{\mathrm{path},g} \asymp \varepsilon^{-2}\varepsilon^{2d+4\alpha} = \varepsilon^{2(d-1+2\alpha)}$. We assume that these normalized atoms, for a sufficiently small ε, belong to our function class $\mathcal{F}_g(\mathbb{B}_1^d; 1, C)$.*

Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$ be the set of vectors from Lemma H.3 that define a disjoint cap packing. We define a family of candidate functions indexed by sign vectors $\xi \in \{\pm 1\}^N$. For each ξ, the function $f_\xi \in \mathcal{F}$ is given by:

$$f_\xi(\boldsymbol{x}) = \sum_{i=1}^N \xi_i \Phi_i(\boldsymbol{x}), \quad \text{where } \Phi_i := \Phi_{\boldsymbol{u}_i,\varepsilon^2}.$$

As the atoms Φ_i have disjoint supports, the squared $L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))$ distance between any two distinct functions f_ξ and $f_{\xi'}$ can be computed as:

$$\|f_\xi - f_{\xi'}\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} = \sum_{i=1}^N (\xi_i - \xi_i')^2 \|\Phi_i\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} = 4 \sum_{i:\xi_i \ne \xi_i'} \|\Phi_i\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))}.$$

Referring to the cap mass properties in Lemma H.1 (which implies $\|\Phi_i\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} \asymp \varepsilon^{d-1+2\alpha}$), this simplifies to the final distance scaling

$$\|f_\xi - f_{\xi'}\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} \asymp_{d,\alpha} \varepsilon^{d-1+2\alpha} d_H(\xi, \xi'),$$

where $d_H(\xi, \xi')$ is the Hamming distance.

H.2 Proof of Theorem 3.6

A key step in our proof is to find a large number of caps that contain no data points from the dataset $\mathcal{D}$. In the standard fixed-sample-size setting, the number of points in each disjoint cap, say $Z_i := \#\{\boldsymbol{x}_j \in C(\boldsymbol{u}_i, \varepsilon)\}$, follows a multinomial distribution. The counts $(Z_1, \ldots, Z_N)$ are negatively correlated because their sum is fixed to n. This dependence complicates the analysis of finding many empty caps simultaneously.

To circumvent this difficulty, we employ **Poissonization**. We replace the fixed sample size n with a random sample size N_{poi} drawn from a Poisson distribution with mean n. This means the occupancy counts Z_i become independent Poisson random variables. This independence allows for the direct use of standard concentration inequalities like the Chernoff bound.

Proposition H.5 (Abundance of Empty Caps under Poissonization)**.** *Let $\{C(\boldsymbol{u}_i, \varepsilon)\}_{i=1}^N$ be the set of disjoint caps from Lemma H.3. Let the sample size be $N_{poi} \sim \text{Poi}(n)$. Let Z_i be the number of samples falling into cap $C(\boldsymbol{u}_i, \varepsilon)$. Define the expected number of points per cap as $\lambda := n \cdot \mathcal{P}_{\boldsymbol{X}}(C(\boldsymbol{u}_1, \varepsilon))$. If we choose ε such that $\lambda \asymp 1$, then there exists a constant $c > 0$ such that with probability at least $1 - \exp(-cN)$:*

$$\#\{i \in \{1, \ldots, N\} : Z_i = 0\} \geq \frac{1}{2} e^{-\lambda} N.$$

Proof. Under Poissonization, the random variables $Z_i = \#\{\boldsymbol{x}_j \in C(\boldsymbol{u}_i, \varepsilon)\}$ are independent Poisson variables with mean $\lambda_i = n \cdot \mathcal{P}_{\boldsymbol{X}}(C(\boldsymbol{u}_i, \varepsilon))$. By Lemma H.2 and our choice of scale, $\lambda_i = \lambda \asymp 1$ for all i.

Let $Y_i = \mathbb{1}\{Z_i = 0\}$ be the indicator that the i-th cap is empty. The variables $Y_1, \ldots, Y_N$ are i.i.d. Bernoulli random variables. The probability of success (a cap being empty) is:

$$p := \mathbb{P}(Y_i = 1) = \mathbb{P}(Z_i = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}.$$

Since $\lambda \asymp 1$, p is a positive constant. The expected number of empty caps is $\mathbb{E}[\sum Y_i] = Np = Ne^{-\lambda}$. By a standard Chernoff bound on the sum of i.i.d. Bernoulli variables, we have that for any $\delta \in (0, 1)$:

$$\mathbb{P}\left(\sum_{i=1}^N Y_i < (1-\delta)Np\right) \leq \exp\left(-\frac{\delta^2 Np}{2}\right).$$

Choosing $\delta = 1/2$, we find that the number of empty caps is at least $\frac{1}{2}Np = \frac{1}{2}e^{-\lambda}N$ with probability at least $1 - \exp(-cN)$ for some constant $c > 0$. □

The condition $\lambda \asymp 1$ is central. It balances the sample size n with the geometric scale ε. Using Lemma H.2, this balance is achieved when:

$$n \cdot \varepsilon^{d-1+2\alpha} \asymp 1 \quad \Longleftrightarrow \quad \varepsilon \asymp n^{-1/(d-1+2\alpha)}. \tag{78}$$

With this choice, Proposition H.5 guarantees that a constant fraction of the $N \asymp \varepsilon^{-(d-1)}$ caps are empty with overwhelmingly high probability. Informlly speaking, this hints appearance of the neural network with dedicated neurons, each of which has at most one activation point. This paradigm aligns with our construction stable/flat interpolation neural network discussed in Appendix I.

Armed with the guarantee of many empty caps, we can now construct our adversarial pair of functions, f and f'. These functions will be designed to agree on all non-empty caps but disagree on a large number of empty caps. Since by definition no data lies in the empty caps, the functions will be identical on the training data. However, their disagreement on a substantial portion of the space will create a large gap in their population risks.

Proposition H.6 (Indistinguishable yet Separated Pair). *Work under the scale choice* $\varepsilon \asymp n^{-1/(d-1+2\alpha)}$ *and on the high-probability event from Proposition H.5 where at least* $\frac{1}{2}e^{-\lambda}N$ *caps are empty. There exist two functions* $f, f' \in \mathcal{F}$ *from Construction H.4 such that*

1. ***Indistinguishability on Data:*** $f(\boldsymbol{x}_j) = f'(\boldsymbol{x}_j)$ *for all points* $\boldsymbol{x}_j$ *in the Poisson-drawn sample.*
2. ***Separation in Population:*** $\|f - f'\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} \asymp n^{-\frac{2\alpha}{d-1+2\alpha}}$.

Proof. Let $\mathcal{J} \subset \{1, \dots, N\}$ be the set of indices corresponding to empty caps, with $|\mathcal{J}| \geq \frac{1}{2}e^{-\lambda}N \asymp N$. Construct two sign vectors $\xi, \xi' \in \{\pm 1\}^N$ as follows:

- For $i \in \mathcal{J}$, set $\xi_i = 1$ and $\xi'_i = -1$.
- For $i \notin \mathcal{J}$, set $\xi_i = \xi'_i = 1$.

Let $f = f_\xi$ and $f' = f_{\xi'}$.

1. **Indistinguishability:** The function difference is $f - f' = \sum_{i \in \mathcal{J}} 2\Phi_i$. The support of this difference is $\bigcup_{i \in \mathcal{J}} C(\boldsymbol{u}_i, \varepsilon)$. Since all caps indexed by $\mathcal{J}$ are empty, no data point $\boldsymbol{x}_j$ falls into this support. Thus, $(f - f')(\boldsymbol{x}_j) = 0$ for all j, which implies $f(\boldsymbol{x}_j) = f'(\boldsymbol{x}_j)$.
2. **Separation:** The Hamming distance is $d_H(\xi, \xi') = |\mathcal{J}| \asymp N$. Using the result from Construction H.4:

   $$\|f - f'\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} \asymp \varepsilon^{d-1+2\alpha} \cdot d_H(\xi, \xi') \asymp \varepsilon^{d-1+2\alpha} \cdot N \asymp \varepsilon^{d-1+2\alpha} \cdot \varepsilon^{-(d-1)} = \varepsilon^{2\alpha}.$$

   Substituting our choice of scale $\varepsilon \asymp n^{-1/(d-1+2\alpha)}$ yields the desired separation:

   $$\|f - f'\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} \asymp \left(n^{-1/(d-1+2\alpha)}\right)^{2\alpha} = n^{-\frac{2\alpha}{d-1+2\alpha}}.$$

□

The final step is to transfer the result from the Poissonized model back to the original fixed-sample-size model. This is justified by the strong concentration of the Poisson distribution around its mean.

Lemma H.7 (De-Poissonization). *Let* $N_{poi} \sim \mathrm{Poi}(n)$. *For any* $\eta \in (0, 1)$, $\mathbb{P}(N_{poi} \notin [(1-\eta)n, (1+\eta)n]) \leq 2\exp(-c_\eta n)$ *for some constant* $c_\eta > 0$. *The conclusions of Proposition H.6 hold for a fixed sample size* n.

Proof. The existence of a large fraction of empty caps is an event that is monotone with respect to the sample size (fewer samples lead to more empty caps). The high-probability conclusion from Proposition H.5 holds for any sample size k within the concentration interval $[(1-\eta)n, (1+\eta)n]$, as changing n to k only alters the key parameter λ by a constant factor, which does not affect the asymptotic analysis. Since N_{poi} falls in this interval with probability $1 - o(1)$, the event of finding an indistinguishable pair also occurs with probability $1 - o(1)$ for a Poisson sample. This high-probability statement can be transferred back to the fixed-n setting, yielding the same rate for the lower bound. □

The existence of an indistinguishable pair allows us to establish a lower bound on the minimax risk for estimation in the noiseless setting. This intermediate result is the foundation for the final generalization gap bound.

Let $\mathcal{F}_{\text{pack}}$ be the adversarial class defined in Construction H.4 with ε defined in Proposition H.6.

Corollary H.8 (Minimax Lower Bound). *In the noiseless setting where* $y_i = f(\boldsymbol{x}_i)$, *the minimax risk for any estimator* $\hat{f}$ *over the adversarial class* $\mathcal{F}_{pack}$ *is bounded below*

$$\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_{pack}} \mathbb{E}\left[\|\hat{f} - f_0\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))}\right] \gtrsim n^{-\frac{2\alpha}{d-1+2\alpha}}.$$

Proof. Let E be the event that an indistinguishable pair $(f, f') \in \mathcal{F}_{\text{pack}}$ exists for a fixed sample size n. From Proposition H.6 and Lemma H.7, we know that $\mathbb{P}(E) = 1 - o(1)$. On this event E, let the true function f_0 be chosen uniformly at random from $\{f, f'\}$.

Any estimator $\hat{f}$ receives the dataset $\mathcal{D}_n$ of size n. Since $f(\boldsymbol{x}_i) = f'(\boldsymbol{x}_i)$ for all $\boldsymbol{x}_i \in \mathcal{D}_n$, the generated data is identical whether $f_0 = f$ or $f_0 = f'$. The estimator thus has no information to distinguish between f and f'. The expected risk of any estimator, conditioned on the event E, can be lower-bounded

$$\mathbb{E}\left[\|\hat{f} - f_0\|^2 \middle| E\right] = \frac{1}{2}\|\hat{f} - f\|^2 + \frac{1}{2}\|\hat{f} - f'\|^2 \geq \frac{1}{4}\|f - f'\|^2,$$

where the inequality is a standard result for a choice between two points. The worst-case risk for an estimator over $f_0 \in \{f, f'\}$ is thus at least $\frac{1}{4}\|f - f'\|^2$.

Taking the expectation over the sampling of D_n:

$$\begin{aligned}\inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_{\text{pack}}} \mathbb{E}\left[\|\hat{f} - f_0\|^2\right] &\geq \inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_{\text{pack}}} \mathbb{E}\left[\|\hat{f} - f_0\|^2 \middle| E\right] \mathbb{P}(E) \\ &\geq \frac{1}{4}\mathbb{E}\left[\|f - f'\|^2 \middle| E\right] \mathbb{P}(E).\end{aligned}$$

Since on the event E, the separation $\|f - f'\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))} \asymp n^{-\frac{2\alpha}{d-1+2\alpha}}$ and $\mathbb{P}(E) \to 1$ as $n \to \infty$, the result follows. □

Finally, we connect the minimax risk lower bound to the generalization gap. The argument reduces the problem of bounding the generalization gap to the minimax estimation problem we just solved.

Theorem H.9 (Generalization Gap Lower Bound)**.** *Let $\mathcal{P}$ denote any joint distribution of $(\boldsymbol{x}, y)$ where the marginal distribution of $\boldsymbol{x}$ is $\mathcal{P}_{\boldsymbol{X}}(\alpha))$ and y is supported on $[-1, 1]$. Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^n$ be a dataset of n i.i.d. samples from $\mathcal{P}$. Let $\widehat{R}_{\mathcal{D}_n}(f)$ be any empirical risk estimator for the true risk $R_{\mathcal{P}}(f) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}}[(f(\boldsymbol{x}) - y)^2]$. Then,*

$$\inf_{\widehat{R}} \sup_{\mathcal{P}} \mathbb{E}_{\mathcal{D}_n}\left[\sup_{f \in \mathcal{F}_g(\mathbb{B}_1^d;1,1)} \left|R_{\mathcal{P}}(f) - \widehat{R}_{\mathcal{D}_n}(f)\right|\right] \gtrsim_{d,\alpha} n^{-\frac{2\alpha}{d-1+2\alpha}}.$$

Proof. We lower-bound the supremum over all distributions $\mathcal{P}$ by restricting it to a worst-case family of deterministic distributions $\mathcal{P}_{f_0}$, where labels are given by $y = f_0(\boldsymbol{x})$ for some f_0 from our adversarial packing set, $\mathcal{F}_{\text{pack}}$. The proof proceeds via a chain of inequalities.

$$\inf_{\widehat{R}} \sup_{\mathcal{P}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|R_{\mathcal{P}}(f) - \widehat{R}_{\mathcal{D}_n}(f)\right|\right] \tag{79}$$

$$\geq \inf_{\widehat{R}} \sup_{f_0 \in \mathcal{F}_{\text{pack}}} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|R_{\mathcal{P}_{f_0}}(f) - \widehat{R}_{\mathcal{D}_n}(f)\right|\right] \tag{80}$$

$$\geq \sup_{f_0 \in \mathcal{F}_{\text{pack}}} \frac{1}{2} \inf_{\widehat{R}} \mathbb{E}\left[R_{\mathcal{P}_{f_0}}(\hat{f}_{\text{ERM}}) - R_{\mathcal{P}_{f_0}}(f_0)\right] \tag{81}$$

$$\geq \frac{1}{2} \inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_{\text{pack}}} \mathbb{E}\left[R_{\mathcal{P}_{f_0}}(\hat{f}) - R_{\mathcal{P}_{f_0}}(f_0)\right] \tag{82}$$

$$= \frac{1}{2} \inf_{\hat{f}} \sup_{f_0 \in \mathcal{F}_{\text{pack}}} \mathbb{E}\left[\|\hat{f} - f_0\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))}\right] \tag{83}$$

$$\text{Corollary H.8} \implies \gtrsim n^{-\frac{2\alpha}{d-1+2\alpha}} \tag{84}$$

The steps are justified as follows

- **Inequality (81):** This step uses a standard result relating the generalization gap to the excess risk of an Empirical Risk Minimizer (ERM), $\hat{f}_{\text{ERM}} := \arg\min_{f \in \mathcal{F}} \widehat{R}_{\mathcal{D}_n}(f)$. By

definition, $\widehat{R}(\hat{f}_{\text{ERM}}) \le \widehat{R}(f_0)$. This leads to the decomposition

$$R(\hat{f}_{\text{ERM}}) - R(f_0) = \Big(R(\hat{f}_{\text{ERM}}) - \widehat{R}(\hat{f}_{\text{ERM}})\Big) + \Big(\widehat{R}(\hat{f}_{\text{ERM}}) - \widehat{R}(f_0)\Big) + \Big(\widehat{R}(f_0) - R(f_0)\Big)$$
$$\le 2 \sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)|.$$

- **Inequation (82)** The infimum over all risk estimators $\widehat{R}$ (which induces a corresponding ERM) is lower-bounded by the infimum over all possible estimators $\hat{f}$ of the function f_0. This transitions the problem to the standard minimax framework.
- **Equation (83):** In this noiseless setting with a deterministic labeling function f_0, the population risk of f_0 is $R_{\mathcal{P}_{f_0}}(f_0) = 0$. The excess risk $R_{\mathcal{P}_{f_0}}(\hat{f})$ is precisely the squared L_2 distance $\|\hat{f} - f_0\|^2_{L^2(\mathcal{P}_{\boldsymbol{X}}(\alpha))}$. The expression becomes the definition of the minimax risk over the class $\mathcal{F}_{\text{pack}}$.

This completes the proof. □

I Flat Interpolating Two-Layer ReLU Networks on the Unit Sphere

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be a dataset with $\boldsymbol{x}_i \in \mathbb{S}^{d-1}$, $d > 1$, and pairwise distinct inputs. Assume labels are uniformly bounded, i.e., $|y_i| \le D$ for all i. Consider width-K two-layer ReLU models

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \phi(\boldsymbol{w}_k^\top \boldsymbol{x} - b_k) + \beta. \tag{85}$$

Theorem I.1 (Flat interpolation with width $\le n$)**.** *Under the set-up above, there exists a width $K \le n$ network of the form (85) that interpolates the dataset and whose Hessian operator norm satisfies*

$$\lambda_{\max}\left(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}\right) \;\le\; 1 + \frac{D^2 + 2}{n}. \tag{86}$$

Construction I.2 (Flat interpolation ReLU network)**.** *Let $I_{\neq 0} := \{i : y_i \neq 0\}$ and set the width $K := |I_{\neq 0}| \le n$. For each $k \in I_{\neq 0}$ define*

$$\rho_k \;:=\; \max_{k \neq i} \boldsymbol{x}_i^\top \boldsymbol{x}_k \;<\; 1, \qquad b_k \in (\rho_k,\, 1) \;\text{(e.g., } b_k = \frac{1+\rho_k}{2}\text{)}, \qquad \boldsymbol{w}_k := \boldsymbol{x}_k. \tag{87}$$

Then for any sample index i,

$$\boldsymbol{w}_k^\top \boldsymbol{x}_i - b_k = \begin{cases} 1 - b_k \;>\; 0, & i = k, \\ \le \rho_k - b_k \;<\; 0, & i \neq k, \end{cases} \tag{88}$$

so the k-th unit activates on $\boldsymbol{x}_k$ and is inactive on all $\boldsymbol{x}_i$ with $i \neq k$. Set the output weight

$$v_k \;:=\; \frac{y_k}{1 - b_k}. \tag{89}$$

By (88) and (89), the model interpolates on nonzero labels because $f(\boldsymbol{x}_k) = a_k(1 - b_k) = y_k$ for $k \in I_{\neq 0}$, and it also interpolates zero labels since all constructed units are inactive on $\boldsymbol{x}_i$ when $i \notin J_{\neq 0}$, hence $f(\boldsymbol{x}_i) = 0 = y_i$.

For each constructed unit, define

$$\tilde{v}_k := \operatorname{sign}(v_k) \in \{\pm 1\}, \qquad \tilde{\boldsymbol{w}}_k := |v_k|\, \boldsymbol{w}_k, \qquad \tilde{b}_k := |a_k|\, b_k. \tag{90}$$

Then for any input $\boldsymbol{x}$,

$$\tilde{v}_k \, \phi(\tilde{\boldsymbol{w}}_k^\top \boldsymbol{x} - \tilde{b}_k) = \operatorname{sign}(v_k)\, \phi\big(|v_k|(\boldsymbol{w}_k^\top \boldsymbol{x} - b_k)\big) = v_k \, \phi(\boldsymbol{w}_k^\top \boldsymbol{x} - b_k), \tag{91}$$

so interpolation is preserved. Moreover, the activation pattern on the dataset is unchanged because (88) has strict inequalities and $|a_i| > 0$. At $\boldsymbol{x}_i$ we have the (post-rescaling) pre-activation

$$\tilde{z}_k \;:=\; \tilde{\boldsymbol{w}}_k^\top \boldsymbol{x}_k - \tilde{b}_k = |a_k|\,(1 - b_k) = |y_k| \;>\; 0, \qquad |\tilde{v}_i| = 1. \tag{92}$$

In what follows we work with the reparameterized network and drop tildes for readability, implicitly assuming $|v_k| = 1$ for all $k \in I_{\neq 0}$ and $z_k := \boldsymbol{w}_k^\top \boldsymbol{x}_k - b_k = |y_k|$.

Proposition I.3. *Let $\boldsymbol{\theta}$ be the model in Construction I.2. Then*

$$\lambda_{\max}(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}) \le 1 + \frac{D^2+2}{n}.$$

Proof. By direct computation, the Hessian $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}$ is given by

$$\nabla^2_{\boldsymbol{\theta}}\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_i)\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_i)^{\mathsf{T}} + \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i)-y_i)\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{x}_i). \tag{93}$$

Since the model interpolates $f(\boldsymbol{x}_i) = y_i$ for all i, we have

$$\nabla^2_{\boldsymbol{\theta}}\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_i)\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_i)^{\mathsf{T}}. \tag{94}$$

Denote the tangent features matrix by

$$\boldsymbol{\Phi} = \left[\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_1), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_2), \cdots, \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_n)\right]. \tag{95}$$

Then $\nabla^2_{\boldsymbol{\theta}}\mathcal{L}$ in (94) can be expressed by $\nabla^2_{\boldsymbol{\theta}}\mathcal{L} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}}/n$, and the operator norm is computed by

$$\lambda_{\max}(\nabla^2_{\boldsymbol{\theta}}\mathcal{L}) = \max_{\boldsymbol{\gamma}\in\mathbb{S}^{(d+2)K}} \frac{1}{n}\|\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\gamma}\|^2 = \max_{\boldsymbol{u}\in\mathbb{S}^{n-1}} \frac{1}{n}\|\boldsymbol{\Phi}\boldsymbol{u}\|^2 \tag{96}$$

From direct computation we obtain

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}) = \begin{pmatrix} \nabla_{\boldsymbol{W}}(f) \\ \nabla_{\boldsymbol{b}}(f) \\ \nabla_{\boldsymbol{\omega}}(f) \\ \nabla_{\beta}(f) \end{pmatrix} \tag{97}$$

For the parameters $[\boldsymbol{w}_k, b_k, v_k]$ associated to the neuron of index j,

$$\frac{\partial f(\boldsymbol{x})}{\partial v_k} = \mathbb{1}\{\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} > b_k\}\left(\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} - b_k\right), \qquad \frac{\partial f(\boldsymbol{x}_i)}{\partial \boldsymbol{w}_k} = \mathbb{1}\{\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} > b_k\}\, v_k\, \boldsymbol{x},$$
$$\frac{\partial f(\boldsymbol{x}_i)}{\partial b_k} = \mathbb{1}\{\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x} > b_k\}\, v_k, \qquad \frac{\partial f(\boldsymbol{x}_i)}{\partial \beta} = 1.$$

By the one-to-one activation property (88), each sample $\boldsymbol{x}_i$ activates exactly one unit (the unit with the same index k when $k \in I_{\neq 0}$), and activates none when $i \notin I_{\neq 0}$. Hence the sample-wise gradient $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_k)$ has support only on the parameter triplet $(\boldsymbol{w}_k, b_k, v_k, \beta)$ for $k \in I_{\neq 0}$, and is zero for other parameters. Writing the nonzero gradient block explicitly (recall $|v_k| = 1$),

$$\nabla_{(\boldsymbol{w}_k, b_k, v_k, \beta)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_k) = \begin{pmatrix} \nabla_{(\boldsymbol{w}_k, b_k, v_k)} f_{\boldsymbol{\theta}} \\ 1 \end{pmatrix},$$
$$\nabla_{(\boldsymbol{w}_k, b_k, v_k)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_k) = \begin{cases} \begin{pmatrix} v_k\, \boldsymbol{x}_k \\ v_k \\ y_k \end{pmatrix}, & (k \in I_{\neq 0}), \\ \mathbf{0}, & (k \notin I_{\neq 0}), \end{cases} \tag{98}$$

After row permutation and subsistion by (98), (96) is of the form

$$\boldsymbol{\Phi} = \begin{pmatrix} \nabla_{(\boldsymbol{w}_1,b_1,v_1)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \nabla_{(\boldsymbol{w}_2,b_2,v_2)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_2) & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \vdots \\ \vdots & \vdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \nabla_{(\boldsymbol{w}_n,b_n,v_n)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_n) \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{99}$$

$$= \begin{pmatrix} \begin{pmatrix} v_1 \boldsymbol{x}_1 \\ v_1 \\ y_1 \end{pmatrix} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} v_2 \boldsymbol{x}_2 \\ v_2 \\ y_2 \end{pmatrix} & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \vdots \\ \vdots & \vdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \begin{pmatrix} v_n \boldsymbol{x}_n \\ v_n \\ y_n \end{pmatrix} \\ 1 & 1 & \cdots & 1 \end{pmatrix}. \tag{100}$$

Let $\boldsymbol{u} = (u_1, \cdots, u_n) \in \mathbb{S}^{n-1}$ and plug (100) in (96) to have

$$\lambda_{\max}(\nabla^2_{\boldsymbol{\theta}} \mathcal{L}) = \max_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \frac{1}{n} \|\boldsymbol{\Phi} \boldsymbol{u}\|^2 \tag{101}$$

$$= \frac{1}{n} \max_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \left\| \begin{pmatrix} u_1 \nabla_{(\boldsymbol{w}_1,b_1,v_1)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_1) \\ u_2 \nabla_{(\boldsymbol{w}_2,b_2,v_2)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_2) \\ \vdots \\ u_n \nabla_{(\boldsymbol{w}_n,b_n,v_n)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_n) \\ \sum_{i=1}^n u_i \end{pmatrix} \right\|_2^2$$

$$= \frac{1}{n} \max_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \sum_{i=1}^n u_i^2 \left\| \nabla_{(\boldsymbol{w}_i,b_i,v_i)} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \right\|_2^2 + \left(\sum_{i=1}^n u_i \right)^2 \tag{102}$$

$$= \frac{1}{n} \max_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \sum_{i=1}^n u_i^2 \left(\|\boldsymbol{x}_i\|_2^2 + 1 + y_i^2 \right) + \left(\sum_{i=1}^n u_i \right)^2 \tag{103}$$

$$\leq \frac{1}{n} \left(\max_{i \in [n]} \left(\|\boldsymbol{x}_i\|_2^2 + 1 + y_i^2 \right) + \max_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \left(\sum_{i=1}^n u_i \right)^2 \right) \tag{104}$$

$$\leq \frac{1}{n} \left(D^2 + 2 + n \right) = 1 + \frac{D^2 + 2}{n}$$

If we remove the output bias term β from the parameters, then the bottom row of 100 will be remove and thus term $\sum_i u_i$ in (102) will be removed. □

J Technical Lemmas

Lemma J.1 (Concentration of a Poisson Random Variable). *Let $N_{poi} \sim \mathrm{Poi}(n)$ be a Poisson random variable with mean n. Then for any $\eta \in (0, 1)$,*

$$\mathbb{P}\left(|N_{poi} - n| \geq \eta n \right) \leq 2 \exp\left(-\frac{\eta^2 n}{3} \right).$$

Proof. The proof employs the Chernoff bounding method. The Moment Generating Function (MGF) of $N_{\text{poi}} \sim \text{Poi}(n)$ is given by:

$$\mathbb{E}\left[e^{tN_{\text{poi}}}\right] = e^{n(e^t-1)}.$$

We will bound the upper and lower tails separately.

We want to bound $\mathbb{P}(N_{\text{poi}} \geq (1+\eta)n)$. For any $t > 0$, Markov's inequality implies:

$$\begin{aligned}\mathbb{P}(N_{\text{poi}} \geq (1+\eta)n) &= \mathbb{P}\left(e^{tN_{\text{poi}}} \geq e^{t(1+\eta)n}\right) \\ &\leq \frac{\mathbb{E}\left[e^{tN_{\text{poi}}}\right]}{e^{t(1+\eta)n}} \\ &= \frac{e^{n(e^t-1)}}{e^{t(1+\eta)n}} = \exp\left(n(e^t-1) - tn(1+\eta)\right).\end{aligned}$$

To obtain the tightest bound, we minimize the exponent with respect to t. The optimal t is found by setting the derivative to zero, which yields $e^t = 1+\eta$, or $t = \ln(1+\eta)$. Substituting this value back into the bound gives:

$$\mathbb{P}(N_{\text{poi}} \geq (1+\eta)n) \leq \exp\left(n((1+\eta)-1) - n(1+\eta)\ln(1+\eta)\right) = \exp\left(n[\eta - (1+\eta)\ln(1+\eta)]\right).$$

We now use the standard inequality: $\ln(1+x) \geq x - \frac{x^2}{2}$ for $x \geq 0$. A more specific inequality for this context is $\eta - (1+\eta)\ln(1+\eta) \leq -\frac{\eta^2}{2(1+\eta/3)}$. For $\eta \in (0,1]$, this further simplifies. A widely used bound derived from this expression is:

$$\exp\left(n[\eta - (1+\eta)\ln(1+\eta)]\right) \leq \exp\left(-\frac{\eta^2 n}{3}\right).$$

Next, we bound $\mathbb{P}(N_{\text{poi}} \leq (1-\eta)n)$. For any $t > 0$, we have:

$$\begin{aligned}\mathbb{P}(N_{\text{poi}} \leq (1-\eta)n) &= \mathbb{P}\left(e^{-tN_{\text{poi}}} \geq e^{-t(1-\eta)n}\right) \\ &\leq \frac{\mathbb{E}\left[e^{-tN_{\text{poi}}}\right]}{e^{-t(1-\eta)n}} \\ &= \frac{e^{n(e^{-t}-1)}}{e^{-t(1-\eta)n}} = \exp\left(n(e^{-t}-1) + tn(1-\eta)\right).\end{aligned}$$

The optimal t is found by setting $e^{-t} = 1-\eta$, or $t = -\ln(1-\eta)$. Substituting this value gives:

$$\mathbb{P}(N_{\text{poi}} \leq (1-\eta)n) \leq \exp\left(n((1-\eta)-1) - n(1-\eta)\ln(1-\eta)\right) = \exp\left(n[-\eta - (1-\eta)\ln(1-\eta)]\right).$$

Using the inequality $-\eta - (1-\eta)\ln(1-\eta) \leq -\frac{\eta^2}{2}$ for $\eta \in (0,1)$, we get a simple bound:

$$\exp\left(n[-\eta - (1-\eta)\ln(1-\eta)]\right) \leq \exp\left(-\frac{\eta^2 n}{2}\right).$$

Since for $\eta \in (0,1)$, we have $\exp(-\eta^2 n/2) \leq \exp(-\eta^2 n/3)$, the lower tail is also bounded by $\exp(-\eta^2 n/3)$.

Using the union bound, we combine the probabilities for the two tails:

$$\begin{aligned}\mathbb{P}\left(|N_{\text{poi}} - n| \geq \eta n\right) &= \mathbb{P}(N_{\text{poi}} \geq (1+\eta)n) + \mathbb{P}(N_{\text{poi}} \leq (1-\eta)n) \\ &\leq \exp\left(-\frac{\eta^2 n}{3}\right) + \exp\left(-\frac{\eta^2 n}{2}\right) \\ &\leq 2\exp\left(-\frac{\eta^2 n}{3}\right).\end{aligned}$$

This completes the proof. □