Pointwise Generalization in Deep Neural Networks

Shaojie Li Yunbei Xu*
National University of Singapore
{li_sj,yunbei}@nus.edu.sg

Abstract

We address the long-standing question of why deep neural networks generalize by establishing a complete pointwise generalization theory for fully connected networks. For each trained model, we equip the hypothesis with a pointwise Riemannian Dimension through the effective ranks of the *learned* feature matrices across layers, and derive hypothesis- and data-dependent generalization bounds. These spectrum-aware bounds break long-standing barriers and are orders of magnitude tighter in theory and experiment, rigorously surpassing bounds based on model size, products of norms, and infinite-width linearizations. Analytically, we identify structural properties and mathematical principles that explain the tractability of deep nets. Empirically, the pointwise Riemannian Dimension exhibits substantial dimensionality reduction, decreases with increased over-parameterization, and captures feature learning and the implicit bias of optimizers across standard architectures and datasets. Taken together, these results provide evidence that deep networks are mathematically tractable in the practical regime and that their generalization is sharply explained by pointwise, spectrum-aware complexity.

Contents

*Authors are listed in alphabetical order. Yunbei Xu (yunbei@nus.edu.sg) is the corresponding author.

1 Introduction

Deep learning has ushered in a new era of AI, delivering striking generalization across a wide range of scientific tasks. Yet these successes are predominantly empirical; theory has not kept pace.

Paradoxically, despite massive overparameterization, especially for large language models, classical theory predicts severe overfitting, whereas practice shows strong generalization. The resulting gap has fueled a prevailing pessimism that neural networks are opaque "black boxes" resistant to principled explanation. This paper narrows that gap by addressing the generalization problem for the canonical neural network—fully connected deep neural network (DNN). Under minimal, verifiable spectral conditions, we prove that fully connected deep networks fall into the tractable family—on a rigorous footing comparable to sparse linear models and low-rank matrix factorization—rather than the unconstrained "general" overparameterized class. To our knowledge, a fully rigorous account that treats generalization in fully connected networks as tractable—by the learning-theory community's accepted standards—has remained limited. This work aims to help tackle this central challenge.

We study standard fully connected (feed-forward) networks on a dataset $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d_0 \times n}$, where each column is one input example. The network has widths $d_1, \ldots, d_L$, and weight matrices $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ for $l = 1, \ldots, L$. We define the *feature matrix* at layer l by the recursion

$$F_l(W, X) := \sigma_l\big(W_l\, F_{l-1}(W, X)\big) \in \mathbb{R}^{d_l \times n}, \qquad l = 1, \ldots, L, \tag{1}$$

where $F_0 := X$ and the nonlinear activation σ_l acts columnwise. Each *column* of F_l is the feature vector of one data point at layer l; each *row* of F_l is the activation of one neuron across the dataset.

Our focus is the *generalization gap*—the difference between test and training loss at the learned weights W. Informally—up to universal constants and mild logarithmic factors in the local Lipschitz constants (made precise in Theorem 4 with discussion on the feasibility of this simplification)—we prove that this gap is controlled by the *effective dimension* of the learned features: uniformly over every $W \in \mathbb{R}^{\sum_l d_l \cdot d_{l-1}}$,

$$\mathcal{L}_{\text{test}}(W) - \mathcal{L}_{\text{train}}(W) \;\lesssim\; \sqrt{\frac{1}{n} \sum_{l=1}^{L} \big(d_l + d_{l-1}\big)\, d_{\text{eff}}\Big(F_{l-1}(W, X)\, F_{l-1}(W, X)^\top\Big)}. \tag{2}$$

Here $d_{\text{eff}}(\cdot)$ denotes the (layerwise) *effective dimension*—a smoothed, spectrum-aware notion of rank—of the feature Gram matrix $F_{l-1}(W, X)F_{l-1}(W, X)^\top$, i.e., the number of meaningful directions the feature data actually occupies at that layer. Intuitively, each layer contributes a term proportional to its size $(d_l + d_{l-1})$ multiplied by how many directions its features $F_{l-1}(W, X)$ truly use, d_{eff}. When features are correlated, low rank, or exhibit a rapidly decaying spectrum (a few large eigenvalues dominating many small ones), d_{eff} is small, so the bound remains tight even for very wide/deep networks. Such "feature compression" phenomena is widely observed in modern deep learning [Huh et al., 2021, Wang et al., 2025, Parker et al., 2023]. Strikingly, in our experiments, increasing overparameterization often induces pronounced *feature-rank compression*: the bound (2) decreases as model size grows (Section 5); for example, in ResNet trained on CIFAR-10, a majority of layers compress to (near-)zero effective rank.

Inequality (2) yields a strong *uniform, hypothesis- and data-dependent* guarantee, which we term *pointwise generalization*. It tracks how features evolve across layers of the *trained* model and explains overparameterization in practice. Moreover, the right-hand side of (2) can be used directly as a *regularizer*, leading to algorithms that adapt to the effective ranks around a benchmark $W^\star$ (Section 4.2). Under minimal spectral conditions, our theory places fully connected networks in the same complexity class as sparse linear models and low-rank matrix factorization: generalization is governed by low *effective dimension* rather than full parameter count. The spectrum-aware effective-dimension notion we adopt is standard and minimax-sharp in linear and kernel settings [Even and Massoulié, 2021]. In contrast, existing bounds either (i) rely on infinite–width linearizations (the NTK line of work, e.g., Jacot et al. [2018]), (ii) blow up exponentially with products of norms (e.g., Bartlett et al. [2017]), or (iii) scale with model size (e.g., VC dimension [Bartlett et al., 2019]). Our bounds avoid these pathologies, providing a pointwise, spectrum–aware account with matching upper and lower rates. In the sense of accepted learning-theory standards (see, e.g., Section 7 of Bartlett et al. [2021]), our results help narrow the gap and provide evidence that generalization in fully connected deep networks is *tractable*.

1.1 Contributions

The paper is organized into three parts—(i) a pointwise generalization framework (Section 2, (ii) structural principles of deep networks (Section 3-4), and (iii) empirical validation (Section 5). We elaborate noveltes in each.

Pointwise Generalization and Finite-Scale Geometry. Classical tools (e.g., Rademacher complexity, uniform covers, products of norms) measure global complexity and are often too coarse for modern deep nets: they miss how a specific trained model uses its learned features across layers. We propose a pointwise framework that targets the model actually trained and yields bounds with matching upper and lower rates via a finite-scale notion of pointwise dimension—achieving the precision of generic chaining—while assigning each hypothesis a pointwise dimension that governs its error. This yields a geometric view of generalization: a *finite-scale*, spectrum-aware geometry driven by dimension reduction (as opposed to infinitesimal limits), which clarifies the nature of generalization and the sources of its difficulty. Our bounds can also be read as optimally analyzed PAC–Bayes: the analysis admits data-dependent priors and deterministic predictors while retaining posterior adaptivity, thereby overcoming core limitations of standard PAC–Bayes approaches.

Structual Principles and Tight Bounds for Neural Networks. We develop a *non-perturbative* approach that uses exact telescoping decompositions (rather than Taylor linearizations) to preserve the finite-scale geometry of deep networks. This yields our first structural principle: *cross-layer correlations factor through the feature matrices and approximately preserve a pointwise linear structure.* We then show that bounding the pointwise dimension reduces to the gold standard of *effective dimension* on local charts, and we extend this to a global statement by constructing an *ellipsoidal* covering over the set of subspaces (Grassmannian). This extension—novel beyond the classical differential–geometric/Lie–algebraic treatments—establishes our second structural principle: *the complexity of the global atlas (covering reference eigenspaces) remains commensurate with that of the local charts.* Building on these principles, we introduce *Riemannian Dimension*—a spectrum-aware, pointwise effective complexity—that governs generalization at the trained model and yields tight, analyzable bounds. We review each step and argue that the resulting bounds are tight in a qualified sense; moreover, they *unconditionally reduce* to spectral–norm bounds (see Appendix E.3.1).

Empirical Findings and Evidences. The experiments are designed to systematically examine three central questions in modern deep learning: (i) why does overparameterization often improve generalization? (ii) how does feature learning evolve during training? and (iii) what implicit regularization is encoded by the baseline optimizer? Across the experimental results, we observe that (i) the overparameterization impressively leads to decreasing Riemannian Dimension; (ii) feature learning compresses the effective ranks of learned features during the training; and (iii) SGD with momentum implicitly regularizes the Riemannian Dimension.

2 The Nature of Pointwise Generalization

In this section, we develop our pointwise framework for generalization analysis, which introduces a tight tool–pointwise dimension–to characterize generalization. We illustrate its advancement to existing methodologies and bring some new understandings on the nature of generalization.

2.1 Pointwise Generalization as *BEST* PAC-Bayes Optimization

Let $\mathcal{F}$ be a hypothesis class, z be random data drawn from an unknown distribution $\mathbb{P}$ (e.g., input-label pair $z = (x, y)$), and $\ell(f; z)$ be real-valued loss function. Denote by $\mathbb{P}_n$ the empirical distribution supported on an i.i.d. sample $\{z_i\}_{i=1}^n \sim \mathbb{P}^n$. Our goal is to control the *generalization gap* $(\mathbb{P} - \mathbb{P}_n)\ell(f; z)$ in the following manner: for $\delta \in (0, 1)$, with probability at least $1-\delta$, uniformly over every $f \in \mathcal{F}$,

$$(\mathbb{P} - \mathbb{P}_n)\ell(f; z) \ := \ \mathbb{E}_{z\sim\mathbb{P}}\big[\ell(f; z)\big] \ - \ \frac{1}{n}\sum_{i=1}^{n}\ell(f; z_i) \le \ C\sqrt{\frac{d(f) + \log\frac{1}{\delta}}{n}}, \tag{3}$$

where $d(f)$ is a hypothesis-dependent complexity measure that aims to characterize the intrinsic complexity of every *trained* hypothesis f, different from canonical uniform convergence analysis.

In the spirit of (3) we introduce the core concept in this section—*pointwise dimension*, a concept strengthen several established generalization methodologies such as PAC-Bayesian analysis, Kolmogorov complexity, and generic chaining (in particular, the formula of Fernique [1975]). We then illustrate its tightness in characterizing the generalization by two theorems. Throughout, "metric" ϱ means a *pseudometric*: all metric axioms hold except that $\varrho(f_1, f_2) = 0$ need not imply $f_1 = f_2$.

Definition 1 (Pointwise Dimension) *Given a function class $\mathcal{F}$, a metric ϱ on $\mathcal{F}$, and a data-dependent prior π over $\mathcal{F}$, the local dimension at f with scale ε is defined as the log inverse density of the $\varepsilon-$ ball $B_\varrho(f, \varepsilon) = \{f' \in \mathcal{F} : \varrho(f, f') \le \varepsilon\}$ centered at f:*

$$\log\frac{1}{\pi(B_\varrho(f, \varepsilon))}. \tag{4}$$

To prepare for our first theorem, we define the loss-induced empirical $L_2(\mathbb{P}_n)$ metric $\varrho_{n,\ell}$ as

$$\varrho_{n,\ell}(f_1, f_2) = \sqrt{\mathbb{E}_{z\sim\mathbb{P}_n}\big[(\ell(f_1; z) - \ell(f_2; z))^2\big]}.$$

Theorem 1 (One-Shot Bound) *Let π be any data-dependent prior on a function class $\mathcal{F}$, and loss $\ell(f; z)$ bounded by $[0, 1]$. Then for every $\delta \in (0, 1)$, with probability at least $1-\delta$ over n i.i.d. draws $z_1, \ldots, z_n \sim \mathbb{P}$, uniformly over all $f \in \mathcal{F}$,*

$$(\mathbb{P} - \mathbb{P}_n)\ell(f; z) \le \inf_{\varepsilon > \sqrt{1/n}}\left\{2\varepsilon + \sqrt{\frac{2\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}{n}}\right\} + \sqrt{\frac{4}{n}} + 3\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Intuitively, pointwise dimension not only concerns the prior mass on a single hypothesis but works for general, incountable $\mathcal{F}$ by taking the piror mass over a ball centered at f. This overcomes key limitations of previous hypothesis-dependent bounds such as Occam's razor bound and Kolmogorov complexity [Lotfi et al., 2022]. Additionally, our perspective brings the best possible PAC-Bayesian mechanism, where the PAC-Bayes is seen as a promising tool for sharpening generalization. The proof develops the generalization gap as an bias-variance optimization problem on the user-chosen posterior. Through novel techniques such as symmetrization and conditioning, we permits data-dependent priors (not allowed in classical PAC-Bayes) and makes the framework applicable to every single hypothesis beyond artificial randomization. This theorem alone, with proved optimality of chosen posterior and closed-form final expression, greatly improves relevant works in the area such as [Hinton and Van Camp, 1993] and [Dziugaite and Roy, 2017].

We present a second bound that strengthens the above one-shot bound via a multi-scale integral (known as generic chaining [Talagrand, 2005]). It covers rich classes whose pointwise dimension can grow as $O\big(d(f)\,\varepsilon^{-2}\big)$ while still achieving a generalization rate of $\sqrt{d(f)/n}$.

Theorem 2 (Generic Chaining Bound and Global Lower Bound) *For loss $\ell(f;z)$ bounded in $[0,1]$, (i) there exists an absolute constant $C>0$ such that for any data-dependent prior π on $\mathcal{H}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over every $f \in \mathcal{F}$*

$$(\mathbb{P}-\mathbb{P}_n)\ell(f;z) \le C\left(\frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log\left(\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}\right)}d\varepsilon + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right);$$

(ii) under mild measurability conditions there are absolute constants $c, c' > 0$ so that

$$\mathbb{E}\left[\inf_{\pi}\sup_{f\in\mathcal{F}}\left((\mathbb{P}-\mathbb{P}_n)\ell(f;z) - \frac{c}{\sqrt{n\log n}}\int_0^\infty \sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon\right) + \frac{c'\sup_{\mathcal{F}}\mathbb{E}[\ell(f;z)]}{\sqrt{n\log n}}\right] \ge 0,$$

where notation $\mathbb{E}$ means taking expectation over samples.

The integral upper bound in Theorem 2 is *tight* in the following sense: no uniform improvement (valid simultaneously for all hypotheses and priors) is possible. This is witnessed by a matching lower bound. A strictly *pointwise* lower bound (depending on the realized hypothesis) is generally unattainable, because the prior π must be hypothesis–blind (a "no free lunch" constraint). Theorem 2 extends Talagrand's celebrated generic chaining to *pointwise* generalization bounds. Consequently, it is fundamentally stronger than classical entropy–integral bounds based on *global* covering numbers—e.g., Dudley's integral—whose integrand takes a supremum over the entire class $\mathcal{F}$ rather than localizing at the realized hypothesis (see, e.g., Section 3 of Block et al. [2021]).

We defer technical innovations and connections to existing methodologies—especially PAC–Bayesian analysis and the unified pointwise generalization framework of Xu and Zeevi [2025]—to the appendix. The key takeaway is that the proposed *pointwise dimension* is a powerful and precise descriptor: it tightly characterizes pointwise generalization.

2.2 Generalization *IS* Finite-Scale Dimension

We advocate the viewpoint that the nature of generalization is a *finite-scale* notion of dimension. Concretely, our proposed pointwise dimension (4) is evaluated at a finite resolution—capturing the model's intrinsic, spectrum-aware complexity and drives dimension reduction at this scale. This stands in sharp contrast to infinitesimal-scale geometric notions—often reducible to model-size[1] measures such as the Hausdorff dimension [Lutz, 2016] (explained below)—which therefore fail to capture the structure that governs predictive performance.

Asymptotic vs. finite-scale dimension. A central notion to geometry is *asymptotic pointwise dimension*, and a classical definition is

$$\lim_{\varepsilon \to 0} \frac{\log \pi(B_\varrho(f, \varepsilon))}{\log \varepsilon},$$

which is essential to fractal geometry (e.g., Chapter 10.1 in [Falconer, 1997]) and Riemannian geometry [Jost, 2008], e.g., in the classical characterizations of Hausdorff and packing dimensions (see Theorem 3 of Lutz [2016]). According to this definition, geometric dimension is inherently infinitesimal: it studies limit behavior $\varepsilon \to 0$ at the point f. A key point that distinguishes generalization to geometry is that generalization studies the finite-scale dimension; and our pointwise dimension $\log \frac{1}{\pi(B_\varrho(f,\varepsilon))}$ clearly reduces as ε increases; thus finite-scale study of geometry leads to significant dimension reduction. In Theorem 1, the goal of generalization is to identify the best finite scale ($\varepsilon^\star \approx$ resulted bound), and at this scale our pointwise dimension (4) could be much smaller than the asymptotic dimensions, which allows tractable generalization in overparameterized models.

3 Deep Neural Networks and Riemannian Dimension

We consider fully connected (feed-forward) networks that map an input $x \in \mathbb{R}^{d_0}$ to an output $f_L(W, x) \in \mathbb{R}^{d_L}$. The architecture is specified by widths $d_0, \ldots, d_L$ and weight matrices $W = \{W_1, \ldots, W_L\}$ with $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ for $l = 1, \ldots, L$. Let $\sigma_1, \ldots, \sigma_L$ be nonlinear activations (e.g., ReLU), acting componentwise on column vectors, and each $\sigma_l : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ is assumed 1-Lipschitz. The network's forward map is the composition

$$f_L(W, x) \;:=\; \sigma_L\Big(W_L\, \sigma_{L-1}\big(W_{L-1} \cdots \sigma_1(W_1 x)\big)\Big).$$

Let $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d_0 \times n}$ collect the n training inputs as columns. For each layer $l \in \{1, \ldots, L\}$, define the depth-l map and the corresponding *feature matrix*

$$f_l(W, x) \;:=\; \sigma_l\Big(W_l\, \sigma_{l-1}\big(W_{l-1} \cdots \sigma_1(W_1 x)\big)\Big), F_l(W, X) \;:=\; \big[f_l(W, x_1) \;\cdots\; f_l(W, x_n)\big] \in \mathbb{R}^{d_l \times n}.$$

Equivalently (full, non-recursive form consist with (1)),

$$F_l(W, X) \;=\; \sigma_l\Big(W_l\, \sigma_{l-1}\big(W_{l-1} \cdots \sigma_1(W_1 X)\big)\Big),$$

where for a matrix $A = [a_1, \ldots, a_n]$ we write $\sigma_l(A) := [\,\sigma_l(a_1), \ldots, \sigma_l(a_n)\,]$. Thus $F_L(W, X)$ collects the network outputs on the dataset X.

[1] Throughout the paper we use "model size" to mean capacity measures such as VC dimension, metric entropy (via covering/packing numbers), and Hausdorff/packing dimension—not the raw parameter count. These notions are defined via $\varepsilon-$coverings or shattering, and when the parameter–to–hypothesis map lacks sufficient Lipschitz regularity, they can exceed the number of parameters. The study of Hausdorff dimension, which is distinct from the ambient Euclidean (parameter-space) dimension, is central in geometric measure theory (see, e.g., Chapter 1 of Simon [2018]).

We denote $\|\cdot\|_{\mathsf{F}}$ for the Frobenius norm, $\|\cdot\|_{\mathrm{op}}$ for the spectral norm, and $\|\cdot\|_2$ for the Euclidean norm on vectors. We abbreviate norm balls by $B_{\mathsf{F}}(R)$, $B_{\mathrm{op}}(R)$, and $B_2(R)$ (all centered at 0; radius being R). The empirical $L_2(\mathbb{P}_n)$ distance between two hypotheses W, W' is (a $1/\sqrt{n}$ scaling is used to keep consistency with Section 2)

$$\varrho_n(W, W') \ := \ \sqrt{\left\|F_L(W, X) - F_L(W', X)\right\|_{\mathsf{F}}^2 / n}\,.$$

The function-level empirical metric and generalization statements in Section 2 for the loss $x \mapsto \ell(f_L(W, x), y)$ at data–label pairs $z = (x, y)$ specialize, on the dataset X, to the metric ϱ_n defined above. We assume the loss $\ell(\cdot, y)$ is β-Lipschitz in its first argument with respect to $f_L(W, x)$, which bridges the metric $\varrho_{n,\ell}$ studied in Section 2 to ϱ_n defined on the weight space.

3.1 Non-Perturbative Expansion and Layer-wise Correlation

Throughout, our finite-scale analysis relies on *non-perturbative* expansions. Borrowing terminology from theoretical physics, "non-perturbative" here means we avoid Taylor/derivative expansions and instead use exact, telescoping algebraic identities that hold at finite scale. For example,

$$W_2'W_1' - W_2W_1 = W_2'(W_1' - W_1) + (W_2' - W_2)W_1, \qquad \Sigma'^{-1} - \Sigma^{-1} = \Sigma'^{-1}(\Sigma - \Sigma')\Sigma^{-1},$$

with analogous decompositions used throughout. This viewpoint preserves the full finite-scale geometry of deep networks, rather than linearizing around an infinitesimal neighborhood.

To present our non-perturbative expansion for DNN, we define *local Lipchitz constant* $M_{l\to L}(W, \varepsilon)$, which characterizes the sensitivity of the layer L output, F_L, to variations in layer l's output, within a neighborhood around F_l. Formally, we assume that for every $W' \in B_{\varrho_n}(W, \varepsilon)$

$$||F_L(F_l(W', X), \{W_i'\}_{i=l+1}^L) - F_L(F_l(W, X), \{W_i'\}_{i=l+1}^L)||_{\mathsf{F}} \leq M_{l\to L}(W, \varepsilon)||F_l(W', X) - F_l(W, X)||_{\mathsf{F}}.$$

Local Lipschitz constants are typically much smaller than products of spectral norms and can be computed by formal–verification toolchains [Shi et al., 2022]. In our bounds these constants appear only inside *logarithmic factors*, so they do not affect the leading rates. For completeness, we discuss them carefully in Appendix E.3.1. We propose a telescoping decomposition to replace conventional Taylor expansion, where in each summand the only difference lie in W_l' and W_l.

$$\begin{aligned} &F_L(W', X) - F_L(W, X) \\ =&\sum_{l=1}^{L} [\underbrace{\sigma_L(W_L' \cdots W_{l+1}'}_{\text{controled by} M_{l\to L}} \underbrace{\sigma_l}_{\text{by1}} (W_l' \underbrace{F_{l-1}(W, X)}_{\text{learned feature}})) - \sigma_L(W_L' \cdots W_{l+1}' \sigma_l(W_l \underbrace{F_{l-1}(W, X)}_{\text{learned feature}}))], \end{aligned} \tag{5}$$

Note that this is a *non-perturbative* expansion that holds unconditionally and does not rely on infinitesimal approximation, and crucially keeps the *learned* feature $F_{l-1}(W, X)$ at the *trained* weight W. From this decomposition and applying basic inequalities, we have the following key lemma.

Lemma 1 (Non-Perturbative Feature Expansion) *For all $W' \in B_{\varrho_n}(W, \varepsilon)$,*

$$||F(W', X) - F(W, X)||_{\mathsf{F}}^2 \leq \sum_{l=1}^{L} L \cdot M_{l\to L}[W, \varepsilon]^2 \cdot ||(W_l' - W_l)F_{l-1}(W, X)||_{\mathsf{F}}^2. \tag{6}$$

The lemma captures the first structural principle of fully connected DNN: cross-layer correlations mostly pass through the feature matrices, preserving an approximate pointwise linear structure.

Since enlarging the metric only shrinks metric balls and hence *increases* the pointwise dimension (4) we analyze in Section 2 (formalized as Lemma 13), it suffices to analyze pointwise dimension under the *pointwise ellipsoidal metric* that appears on the right-hand side of Lemma 1. Concretely, $F_{l-1}(W,X)\,F_{l-1}(W,X)^\top$, the feature Gram matrix from layer $l-1$, faithfully encodes the spectral information induced by the network–data geometry at layer l. Working with the corresponding pointwise ellipsoidal metric yields sharp, *pointwise, spectrum-aware* bounds with the desired properties for deep networks, and underpins our tractability results (with the structural principles and technical innovations to developed in the next subsection).

3.2 Hierarchical Covering from Local Chart to Global Atlas

Lemma 1 suggests that the following *pointwise ellipsoidal metric* dominates $n \cdot \varrho_n$ at every W (here, NP stands for "non-perturbative"):

$$\begin{aligned} G_{\mathrm{NP}}(W) &= \mathrm{blockdiag}\left(\cdots, LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}^\top(W,X)\otimes I_{d_l},\cdots\right) \\ \varrho_{G_{\mathrm{NP}}(W)}(W,W') &= \mathrm{vec}(W'-W)^\top G_{\mathrm{NP}}(W)\mathrm{vec}(W'-W). \end{aligned} \tag{7}$$

We are therefore interested in bounding the enlarged pointwise dimension under the pointwise ellipsoidal metric $\varrho_{G_{\mathrm{NP}}(W)}$: $\log \frac{1}{\pi(B_{\varrho_n}(f(W,\cdot),\varepsilon))} \le \log \frac{1}{\pi(B_{\varrho_{G_{\mathrm{NP}}(W)}}(W,\sqrt{n}\varepsilon))}$.

Golden standard: effective dimension. Classical studies of static ellipsoidal metrics suggest that if π is chosen to be uniformly constrained on the top-r eigenspace of a PSD matrix $G(W)$, then one can achieve a tight effective dimension as follows: define the *effective rank*

$$r_{\mathrm{eff}}(G(W),R,\varepsilon) := \max\{k : \lambda_k(G(W)), R^2 \ge n\varepsilon^2/2\}, \tag{8}$$

where the eigenvalues $\{\lambda_k(G(W))\}$ are ordered nonincreasingly; and define the spectrum-aware *effective dimension*

$$d_{\mathrm{eff}}(G(W),R,\varepsilon) := \frac{1}{2}\sum_{k=1}^{r_{\mathrm{eff}}(G(W),R,\varepsilon)} \log\left(\frac{8R^2\lambda_k(G(W))}{n\varepsilon^2}\right). \tag{9}$$

This definition serves as a gold standard for static ellipsoidal metrics and is asymptotically tight, as established by the covering number of the unit ball with ellipsoids in Dumer et al. [2004]. For brevity, we write r for $r_{\mathrm{eff}}(G(W),R,\varepsilon)$, and denote by $\mathcal{V}\subset\mathbb{R}^p$ the r-dimensional subspace corresponding to the top-r_{eff} eigenspace of $G(W)$.

Key challenge: prior independence from W. However, the main challenge is that the construction of π cannot rely on knowledge of W, including its top-r_{eff} eigenspace, yet still capture the underlying geometric structure. The next lemma extends classical results on static ellipsoidal metrics by showing that a uniform prior over a reference subspace $\bar{\mathcal{V}}$ suffices to bound the pointwise dimension for all W whose top-r eigenspace of $G(W)$ can be approximated by $\bar{\mathcal{V}}$.

Lemma 2 (Pointwise Dimension via Reference Subspace) *Consider the weight space $B_2(R)\subset\mathbb{R}^p$ for vectorized weights, and a pointwise ellipsoidal metric defined via PSD $G(W)$. Let $\bar{\mathcal{V}}\subseteq\mathbb{R}^p$*

be a fixed r-dimensional subspace. Define the prior $\pi_{\bar{\mathcal{V}}} = \mathrm{Unif}\big(B_2(1.58R) \cap \bar{\mathcal{V}}\big)$. Then, uniformly over all (W, ε) such that the top-r eigenspace $\mathcal{V}$ of $G(W)$ can be approximated by $\bar{\mathcal{V}}$ to precision

$$\varrho_{\mathrm{proj},G(W)}(\mathcal{V}, \bar{\mathcal{V}}) := \big\| \, G(W)^{1/2}\big(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}\big) \big\|_{\mathrm{op}} \;\leq\; \tfrac{\sqrt{n}\varepsilon}{4R}, \tag{10}$$

we have

$$\log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \;\leq\; \frac{1}{2} \sum_{k=1}^{r_{\mathrm{eff}}(G(W),R,\varepsilon)} \log\!\left(\frac{40R^2 \lambda_k(G(W))}{n\varepsilon^2} \right) \;=\; d_{\mathrm{eff}}(G(W), \sqrt{5}R, \varepsilon).$$

In (10), $\mathcal{P}_{\mathcal{V}}$ denotes the orthogonal projector onto the subspace $\mathcal{V}$, and $\varrho_{\mathrm{proj},G(W)}$ thus defines an ellipsoidal projection metric between subspaces. Further details are provided in the appendix.

Hierarchical covering (mixture prior over subspaces). We employ a hierarchical covering argument. For each reference subspace $\bar{\mathcal{V}}$, the bottom-level prior (uniform on $\bar{\mathcal{V}}$) can achieve a tight pointwise dimension bound for all "local" weights W whose top-r eigenspace of $G(W)$ can be well-approximated by $\bar{\mathcal{V}}$. At the top level, we then construct a prior over $\bar{\mathcal{V}}$. By combining these two levels of priors, we obtain a pointwise dimension bound using a prior π that is completely blind to the choice of W. To formalize this, we introduce a top-level distribution μ over the Grassmannian

$$\mathrm{Gr}(p, r) \;:=\; \big\{ r\text{–dimensional linear subspaces of } \mathbb{R}^p \big\}$$

the collection of all r-dimensional subspaces, and define

$$\pi(W) \;=\; \sum_{\mathcal{V}} \mu(\mathcal{V}) \, \pi_{\mathcal{V}}(W).$$

We refer to this two-stage construction as the hierarchical covering argument. Under the resulting prior π, the following bound holds uniformly over all (vectorized) $W \in B_2(R)$, the pointwise dimension $\log \frac{1}{\pi(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))}$ is bounded by two parts:

$$\underbrace{\log \frac{1}{\mu(B_{\varrho_{\mathrm{proj},G(W)}}(\mathcal{V}, \sqrt{n\varepsilon/4R}))}}_{\text{covering Grassmannian (global atlas)}} + \underbrace{\sup_{\bar{\mathcal{V}} \in B_{\varrho_{\mathrm{proj},G(W)}}(\mathcal{V}, \sqrt{n\varepsilon/4R})} \log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))}}_{\text{covering local charts}}, \tag{11}$$

In differential–geometric terms, our argument has two components. *Local (chart) analysis:* fixing a reference subspace $\bar{\mathcal{V}}$, we use effective dimension as the gold standard to determine the metric entropy of the corresponding local chart. *Global (atlas) covering:* we cover the Grassmannian by such reference subspaces, i.e., we bound the metric entropy of the global atlas and account for the cost of transitioning across charts. Lemma 2 controls the local part, while the following new result on the *ellipsoidal* Grassmannian controls the global part:

Lemma 3 (Ellipsoidal Covering of the Grassmannian manifold) *Consider the Grassmannian $\mathrm{Gr}(d, r)$. For uniform prior $\mu = \mathrm{Unif}(\mathrm{Gr}(d, r))$, we have that for every $\mathcal{V} \in \mathrm{Gr}(d, r)$, every $\varepsilon > 0$ and every PSD matrix Σ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$, we have the pointwise dimension bound*

$$\log \frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V}, \varepsilon))} \leq \frac{d-r}{2} \sum_{k=1}^{r} \log \frac{C \max\{\lambda_k, \varepsilon^2\}}{\varepsilon^2} + \frac{r}{2} \sum_{k=1}^{d-r} \log \frac{C \max\{\lambda_k, \varepsilon^2\}}{\varepsilon^2},$$

where $C > 0$ is an absolute constant.

The result above is mathematically significant in its own right. It extends the classical metric-entropy (covering number) theory for the Grassmannian—where log covering number $\asymp r(d-r)\log(C/\varepsilon)$ under the *isotropic* projection metric— to an *ellipsoidal* (anisotropic) metric that captures feature– and model–induced geometry. This generalization translates the traditional differential-geometric and Lie-algebraic treatments (see Appendix D) and, we believe, illustrates a two–way exchange: deep mathematical structure is essential to understanding generalization in modern neural networks, and, conversely, generalization theory can motivate new questions and results in pure mathematics.

Leveraging the block–decomposable structure in (7), whose l-th block tensor product is a $d_{l-1}\times d_{l-1}$ feature matrix replicated across d_l neurons, we obtain the following explicit calculation.

Theorem 3 (Riemannian Dimension for DNN) *Consider the weight space $B_{\mathsf{F}}(R)$, and a pointwise ellipsoidal metric defined via the ellipsoidal metric $G_{\mathrm{NP}}(W)$ defined in (7). Define the pointwise Riemannian Dimension*

$$d_{\mathrm{R}}(W,\varepsilon)=\sum_{l=1}^{L}\Big(\underbrace{d_l\cdot d_{\mathrm{eff}}(A_l(W))}_{\text{covering local charts}}+\underbrace{d_{l-1}\cdot d_{\mathrm{eff}}(A_l(W))}_{\text{covering global atlas}}+\underbrace{\log(d_{l-1}n)}_{\text{covering discrete } r_{\mathrm{eff}}}\Big),$$

where $A_l(W)$ is the the feature matrix $LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F^{\top}_{l-1}(W,X)$; and $d_{\mathrm{eff}}(A_l(W))$ is abbreviation of $d_{\mathrm{eff}}(A_l(W), C\max\{\|W\|_{\mathsf{F}}, R/2^n\},\varepsilon)$ with $C>0$ an absolute constant. Then we have the pointwise dimension bound: there exists a prior π such that uniformly over all $W\in B_{\mathsf{F}}(R)$,

$$\log\frac{1}{\pi(B_{\varrho_n}(f(W,\cdot),\varepsilon))}\le d_{\mathrm{R}}(W,\varepsilon).$$

This concludes our program for fully connected networks: we establish *Riemannian Dimension* as a principled complexity measure that explains—and sharply bounds—generalization. We summarize the *second structural principle of fully connected DNN*: The complexity of the *global atlas* (covering the space of reference top eigenspaces) remains commensurate with the layerwise, spectrum–aware complexity of covering the *local charts.*

4 Generalization Bounds and Comparison

4.1 Generalization Bound for DNN

We are now ready to state our generalization bound for fully connected DNN here. Combining Theorem 3 and Theorem 2, we establish the following theorem.

Theorem 4 (Generalization Bound for DNN) *Let the loss $\ell(f(W,x),y)$ be bounded in $[0,1]$ and $\beta-$Lipchitz with respect to $f(W,x)$, for every $\delta\in(0,1)$, with probability at least $1-\delta$, uniformly over all $W\in B_{\mathsf{F}}(R)$,*

$$(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y)\le C_1\Big(\frac{\beta}{\sqrt{n}}\int_0^{\infty}\sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon+\sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\Big),$$

where the Riemannian Dimension is defined by

$$d_{\mathrm{R}}(W,\varepsilon) = \sum_{l=1}^{L} \Bigg((d_l + d_{l-1}) \sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \underbrace{\log \frac{8C_2^2 \lambda_k(F_{l-1}F_{l-1}^\top)}{n\varepsilon^2}}_{\text{spectrum of inner layers } 1:l-1}$$
$$+ (d_l + d_{l-1}) r_{\mathrm{eff}}[W,l] \cdot \underbrace{\log \Big(M_{l\to L}^2(W,\varepsilon) L \max\{||W||_{\mathsf{F}}^2, R^2/4^n\} \Big)}_{\text{spectrum of outer layers } l+1:L} + \log(d_{l-1}n) \Bigg), \tag{12}$$

where F_{l-1} is learned feature $F_{l-1}(W,X)$; and the effective rank $r_{\mathrm{eff}}[W,l]$ is the abbreviation of $r_{\mathrm{eff}}(LM_{l\to L}^2(W,\varepsilon)F_{l-1}F_{l-1}^\top, C_2\max\{||W||_{\mathsf{F}}, R/2^n\}, \varepsilon)$, where $C_1, C_2 > 0$ are absolute constants.

Interpreting (12) to the informal rate (2). Although $r_{\mathrm{eff}}[W,l]$ incorporates local Lipschitz factors—specifically, the effective rank is computed for $LM_{l\to L}^2(W,\varepsilon)\, F_{l-1}F_{l-1}^\top$ rather than $F_{l-1}F_{l-1}^\top$ alone—when $F_{l-1}F_{l-1}^\top$ exhibits rapidly decaying eigenvalues this dependence is strongly suppressed; it disappears entirely under strict low rank (as also observed in our experiments). Consequently, under mild low-rank or spectral-decay conditions, the bound aligns with the informal rate (2). In (12), the first and second parts correspond to the inner and outer layers, respectively. For each layer. For each layer l, the first ("log–eigenvalues") term in (12) quantifies the contribution of the inner layers $1{:}l{-}1$ via the feature Gram $F_{l-1}F_{l-1}^\top$, while the second ("log–Lipschitz") term captures the influence of the outer layers $l{+}1{:}L$ through $M_{l\to L}$—making explicit how the outer layers enter the bound and restoring inner/outer symmetry. Together, these terms provide a complete layerwise account of the effective dimension in the informal rate (2).

Tightness of each step and resulting bounds. We conclude by reviewing our comprehensive theory for generalization in fully connected networks and justifying the tightness of the resulting bounds. **First**, in Section 2 we develop a framework based on *pointwise dimension*. The upper and lower bounds match in a qualified (non-uniform) sense (see remarks after Theorem 2), and the framework has a profound connection to finite-scale geometry—evidence that this is the right organizing principle. **Second**, Section 3 introduces a *non-perturbative* expansion. Lemma 1 applies Cauchy–Schwarz layerwise (treating each layer as a block). While there may be room to improve depth dependence, the telescoping decomposition (5) is an exact *equality*, so the expansion is generally sharp (and fully avoid linearization). **Third**, the hierarchical covering argument shows that the resulting *Riemannian Dimension* bound matches the gold standard of *effective dimension*. Thus our pointwise, spectrum-aware bounds achieve the optimal form dictated by static ellipsoid theory, now in strongly correlated deep networks.

4.2 Implicit Bias and Algorithmic Implication

Pointwise Dimension as Regularization and Implicit Bias. A central tenet in deep learning generalization is implicit bias (regularization favored by the algorithms) [Vardi, 2023]. From our theory, we see any pointwise generalization bound results in a regularization (and thus an implicit bias) for algorithm design. For a bound in the form (3), considering the regularized ERM: $\hat{f} = \mathrm{argmin}_f\{\mathbb{P}_n\ell(f;z) + C\sqrt{\frac{d(f)+\log(2/\delta)}{n}}\}$, with probability at least $1-\delta$, its excess risk is bounded by (compared to any benchmark $f^\star \in \mathcal{F}$):

$$\begin{aligned}
&\mathbb{P}\ell(\hat{f}; z) - \mathbb{P}\ell(f^\star; z) \\
&\leq \inf_{f \in \mathcal{F}} \left\{ \mathbb{P}_n \ell(f; z) + C\sqrt{\frac{d(f) + \log(2/\delta)}{n}} \right\} - \mathbb{P}\ell(f^\star; z) \leq (C + \sqrt{1/2})\sqrt{\frac{d(f^\star) + \log(2/\delta)}{n}};
\end{aligned}$$

see Appendix E.2. This gives a problem-dependent upper bound $\sqrt{d(f^\star)/n}$ (adapts to the optimal hypothesis $f^\star$). Sparse linear models and matrix factorization explicitly impose sparsity/low-rank assumptions on $f^\star$; without strong convexity, the resulting statistical rates are unimprovable. This provides a sanity check that, under minimal and verifiable spectral conditions, our theory places deep neural networks in the same statistical complexity class as these well-understood models.

4.3 Comparison with Norm Bound, VC, and NTK

Norm bound: Invoking the elementary inequality

$$\log x \leq \log(1 + x) \leq x,$$

we rigorously show that the Riemannian-dimension–based bound in Theorem 4 is *exponentially* tighter than a representative spectral-norm bound in the style of [Bartlett et al., 2017, Neyshabur et al., 2018]; see Appendix E.3.1 for the detailed derivation and comparison.

VC dimension: Let P be the number of weights and L be the number of layers, Bartlett et al. [2019] prove a nearly tight VC–dimension bound $\text{VCdim} \leq O(PL \log P)$, supported by a lower bound $\text{VCdim} \geq \Omega(PL \log(P/L))$. This VC dimension bound is roughly equivalent to be $L \sum_{l=1}^{L} d_l d_{l-1}$.[2] Our Riemannian Dimension bound, by contrast, substantially sharpens this rate: it removes the explicit dependence on depth L and replaces the crude width factor with a (layerwise) effective-rank term.

Neural Tangent Kernel (NTK): Our approach uses exact non-perturbative expansion, which preserves the finite-scale geometry of deep networks, beyond NTK's Taylor linearizations that remains only valid in the infinitesimal-scale neighborhood around initialization. In other words, our results yields a finite-scale, pointwise theory in practical regimes, whereas NTK breaks down beyond the linear (or "lazy") regime—a fundamental limitation that largely undermines its practical relevance.

5 Experiments

We evaluate our Riemannian Dimension on two standard architectures—Fully Connected Networks (FCNs) and ResNets, using two benchmark datasets—MNIST [LeCun et al., 1998] and CIFAR-10 [Krizhevsky, 2009], respectively. We consider a 9-hidden-layer FCN architecture, where, except for the fixed layers, hidden layers share a common width h, with $h \in \{2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$.

[2] The extra factor L beyond parameter count in VCdim is essentially unavoidable: for nonlinear compositional models, VC/Hausdorff/packing dimensions depend on the logarithm of a global worst-case Lipschitz constant, and in depth$-L$ networks that constant grows multiplicatively across layers, yielding an additional linear dependence on L.

Table 1: Final-epoch Metrics of FCNs on MNIST. Supplementary explanations of columns: 1) Width$-2^{\star}$ means $h = 2^{\star}$; 2) Train Error; 3) Generalization gap is defined as test error minus train error; 4) The spectral norm is the spectrally normalized margin bound of [Bartlett et al., 2017]. It is the tightest norm–based bound in the literature to our knowledge; 5) Parameter Counts of the network; 6) VC dimension. We adopt a nearly tight VC–dimension bound from [Bartlett et al., 2019] and report $PL \log P$ for brevity (see Section 4.3); 7) R-D means our Riemannian Dimension.

Model	Train	Gen	Spectral Norm	# Parameters	VC dimension	R-D
Width-2^6	0.0002	0.0205	3.146×10^{15}	5.961×10^{6}	9.299×10^{8}	6.433×10^{7}
Width-2^7	0.0002	0.0187	2.695×10^{15}	6.167×10^{6}	9.641×10^{8}	6.097×10^{7}
Width-2^8	0.0000	0.0191	2.093×10^{15}	6.726×10^{6}	1.057×10^{9}	5.589×10^{7}
Width-2^9	0.0000	0.0186	2.401×10^{15}	8.434×10^{6}	1.345×10^{9}	5.316×10^{7}
Width-2^{10}	0.0000	0.0215	4.816×10^{15}	1.421×10^{7}	2.340×10^{9}	5.266×10^{7}
Width-2^{11}	0.0000	0.0160	1.001×10^{16}	3.520×10^{7}	6.116×10^{9}	4.972×10^{7}
Width-2^{12}	0.0000	0.0210	1.466×10^{16}	1.149×10^{8}	2.133×10^{10}	4.803×10^{7}

Table 2: Final-Epoch Metrics of ResNets on CIFAR-10

Model	Train Error	Gen Gap	# Parameters	VC dimension	R-D
ResNet-20	0.0016	0.0752	2.690×10^{5}	6.727×10^{7}	8.801×10^{6}
ResNet-32	0.0003	0.0695	4.630×10^{5}	1.933×10^{8}	9.992×10^{6}
ResNet-44	0.0001	0.0627	6.570×10^{5}	3.872×10^{8}	6.339×10^{6}
ResNet-56	0.0000	0.0637	8.510×10^{5}	6.507×10^{8}	5.200×10^{6}
ResNet-74	0.0000	0.0615	1.142×10^{6}	1.179×10^{9}	3.237×10^{6}
ResNet-110	0.0000	0.0576	1.724×10^{6}	2.723×10^{9}	2.583×10^{6}

Increasing h monotonically enlarges both layer widths and model sizes. We adopt canonical ResNet architectures—ResNet-20, ResNet-32, ResNet-44, ResNet-56, ResNet-74 and ResNet-110—which differ only in the number of residual blocks per stage while maintaining the same overall architecture (three-stage, basic-block design) as introduced by [He et al., 2016]. These ResNet architectures provides a clean capacity sweep via depth. In what follows, we organize experiments around the two complementary regimes—width scaling on FCNs and depth scaling on ResNets.

This design lets us systematically study three central questions in modern deep learning: (i) why does overparameterization often improve generalization? (ii) how does feature learning evolve during training? and (iii) what implicit regularization is encoded by the baseline optimizer? Detailed experimental setups are deferred to Appendix A.

5.1 Riemannian Dimension Explains Overparameterization

This section studies why does overparameterization—despite exploding model capacity—often improve generalization. We investigate this paradox by tracking our Riemannian Dimension across models with varying parameter counts, asking whether more parameters truly enlarge capacity or instead deduce complexity.

Final-epoch metrics of FCNs on MNIST and ResNets on CIFAR-10 are reported in Table 1 and

Table 3: Final-epoch Effective Ranks for FCNs on MNIST, where Width$-2^{\star}$ means $h = 2^{\star}$, and where for the form A/B, A represents the effective rank and B represents the original dimension, and where Layer-1 means the input layer.

Metric	Width-2^6	Width-2^7	Width-2^8	Width-2^9	Width-2^{10}	Width-2^{11}	Width-2^{12}
Layer-1	713/763	712/763	710/763	710/763	707/763	707/763	704/763
Layer-2	2048/2048	2044/2048	2042/2048	2048/2048	2047/2048	2048/2048	2048/2048
Layer-3	2048/2048	2045/2048	2037/2048	2019/2048	1925/2048	1460/2048	1009/2048
Layer-4	61/64	97/128	92/256	85/512	79/1024	79/2048	59/4096
Layer-5	23/64	43/128	34/256	33/512	28/1024	26/2048	22/4096
Layer-6	20/64	24/128	20/256	21/512	19/1024	18/2048	15/4096
Layer-7	15/64	18/128	17/256	15/512	15/1024	14/2048	13/4096
Layer-8	15/64	14/128	15/256	11/512	13/1024	13/2048	12/4096
Layer-9	14/64	14/128	15/256	13/512	13/1024	12/2048	12/4096
Layer-10	13/64	13/128	12/256	14/512	12/1024	13/2048	14/4096
Total	4970	5024	4994	4969	4858	4390	3908

Table 4: Final-epoch Effective Ranks for ResNets on CIFAR-10, where for the form A/B, A represents the effective rank and B represents the original dimension, and where Layer-0% means the input layer.

Metric	ResNet-20	ResNet-32	ResNet-44	ResNet-56	ResNet-74	ResNet-110
Layer-0%	384/3072	384/3072	17/3072	0/3072	0/3072	0/3072
Layer-25%	2048/16384	2048/16384	7/16384	1/16384	0/16384	0/16384
Layer-50%	1024/8192	1024/8192	1024/8192	227/8192	0/8192	0/8192
Layer-75%	512/4096	512/4096	512/4096	512/4096	58/4096	0/4096
Layer-100%	8/64	8/64	8/64	8/64	8/64	8/64
Total	23432	37768	27564	16294	11401	6925

Table 2, respectively. In these Tables, the train error quickly collapses to zero for sufficiently large models, confirming their expressive capacity. Consistently, the generalization can continue to be improved as parameters increase, especially on ResNets (Table 2). This phenomenon means the overfitting does not appear and reflects a paradoxical truth of deep learning: over-parameterization is not a curse, but can benefit the generalization. However, classical complexity measures—e.g., the spectral norm and the VC dimension, often scale exponentially as the parameter count grows. Notably, the spectral norm is about 10^6 times larger than the VC dimension and seems to be a worse complexity measure (see Table 1). The two measures therefore struggle to explain the generalization of modern overparameterized networks. In contrast, our Riemannian Dimension exhibits a consistent downward trend as model size grows—both under width scaling (last column of Table 1) and depth scaling (last column of Table 2), and it is about 10^3 times smaller than the VC dimension, suggesting that the effective dimension—not raw parameter count—is the true indicator of generalization in deep learning. In summary, increased parameterization is associated with reduced effective model complexity, and Riemannian Dimension faithfully characterizes this phenomenon.

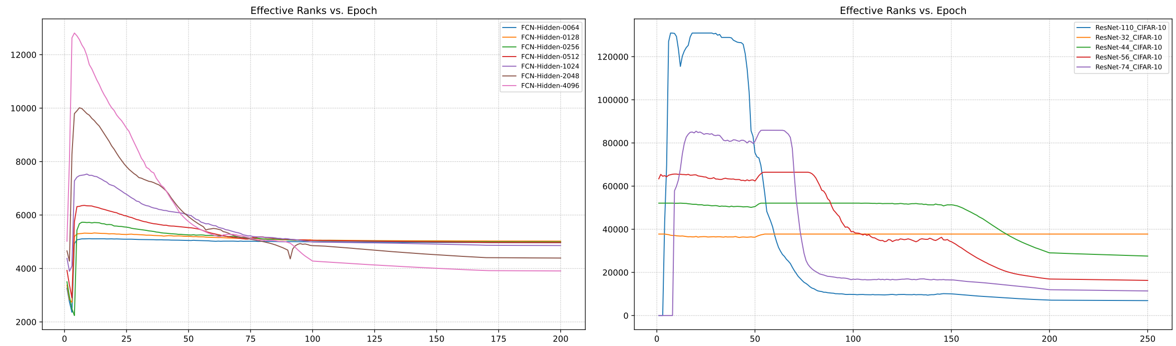

Figure 1: Effective Rank evolutions of FCNs on MNIST (left) and ResNets on CIFAR-10 (right) across the training

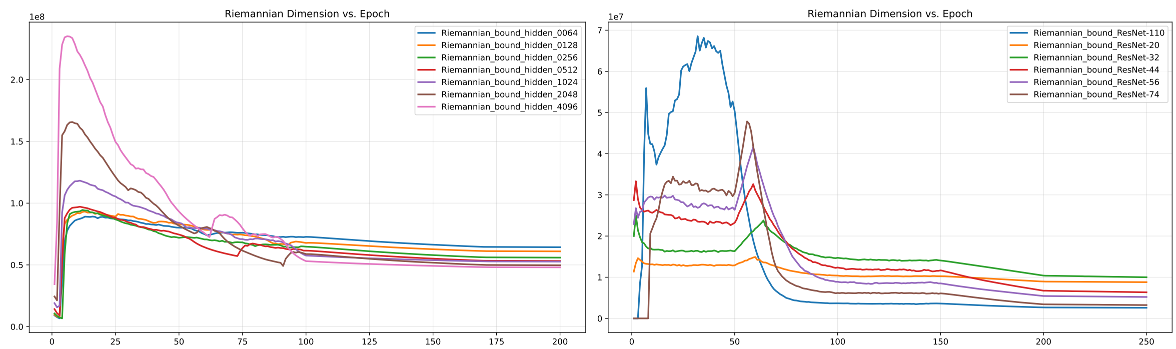

Figure 2: Riemannian Dimension evolutions of FCNs on MNIST (left) and ResNets on CIFAR-10 (right) across the training

5.2 Feature Learning Compresses Effective Rank

We investigate the dynamics of feature learning by monitoring the effective rank of the feature Gram matrices $F_{l-1}F_{l-1}^{\top}$, with the normalization $\cdot L\|W\|_F^2 \prod_{i>l} \|W_i\|_{\mathrm{op}}^2$ dictated by our theory. Here, replacing the local Lipschitz constant $M_{l\to L}(W,\varepsilon)$ by the spectral-norm product $\prod_{i>l}\|W_i\|_{\mathrm{op}}$ is conservative: state-of-the-art formal-verification toolchains [Shi et al., 2022] can compute local Lipschitz constants much more sharply—with well-developed packages and rigorous numerical guarantees—than this crude product bound, and could therefore further strengthen all our empirical results (an active research area). On the other hand, this relaxation—dropping the $\varepsilon-$dependence when making the conservative substitution—can be justified rigorously (see the Step 4 in the proof of Corollary 1 in Appendix E.3.2), and we adopt this simplification in our experiments. We report our empirical results in Tables 3, 4 and Figure 1.

Experimental results reveal some clear patterns: (1) As training proceeds, the effective ranks of feature grams decreases sharply after a short transient; refer to Figure 1. (2) Increased parameter counts, both under width scaling (FCNs) and depth scaling (ResNets), foster compressing effective ranks of feature grams in both the rate and the degree; refer to Figure 1. (3) On the largest FCN, the degree of effective rank compression can reach as much as 1/300, which explains why the Riemannian Dimension can achieve such a significant improvement over the VC dimension; refer to Table 3. While on the largest ResNet, the effective ranks of the vast majority of layers compress to zero, which explains why deeper networks can, paradoxically, exhibit a smaller Riemannian Dimension; refer to Table 4. These experimental results indicates that feature learning steadily reduces the intrinsic dimensionality of features over training and aim to learn a lower-dimensional feature manifold, and the overparameterization intensifies this reduction.

5.3 SGD Finds Low Riemannian Dimension Point

Related literature has shown that various norms are implicit bias of optimizers, but typically limited to linear models [Vardi, 2023]. This section studies whether SGD with momentum, in modern deep learning, implicit regularized Riemannian Dimension across training dynamics. We examine whether this optimizer preferentially converge to solutions with lower Riemannian Dimension point, and the experimental results are presented in Figure 2.

Empirical results show a repeatable pattern across the architectures: SGD with momentum drives the networks toward solutions with lower intrinsic Riemannian Dimension complexity, after an early transient; refer to Figure 2. Notably, Riemannian Dimension drops by orders of magnitude, whereas VC dimension remains essentially unchanged. The alignment between optimization dynamics and complexity control supports the view that SGD with momentum implicitly regularizes the Riemannian Dimension. Therefore, optimization is not merely as a mechanism for convergence; it is a primary driver of generalization through its systematic preference for low-complexity solutions. Riemannian Dimension provides a practical and theoretically grounded lens through which the implicit bias of optimizes in machine learning can be quantitatively assessed.

6 Conclusion

We have developed a coherent, geometry-aware foundation for tractable and predictive generalization in fully connected neural networks. Meeting this challenge required several technical innovations: a pointwise generalization framework, a non-perturbative calculus for network mappings, a hierarchical covering scheme, and an ellipsoidal entropy theory for the Grassmannian—yielding structural insights into cross-weight correlations and global geometric organization. The results strengthen the case that deep-learning generalization admits rigorous explanation and motivate a broader program in finite-scale geometric analysis of strongly correlated learning systems. Our experiments support the theory: the proposed Riemannian Dimension consistently tracks benign overparameterization, feature learning, and the optimizer's implicit bias. Important directions ahead include integrating the analysis with optimization and algorithmic perspectives, extending it to modern architectures, and translating the theory into concrete design principles for new deep-learning systems.

References

Pierre Alquier et al. User-friendly introduction to pac-bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024.

Steve Awodey. *Category theory*, volume 52. OUP Oxford, 2010.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Thomas Bendokat, Ralf Zimmermann, and P-A Absil. A grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50(1):6, 2024.

Adam Block, Yuval Dagan, and Alexander Rakhlin. Majorizing measures, sequential complexities, and online learning. In *Conference on Learning Theory*, pages 587–590. PMLR, 2021.

Olivier Catoni. A pac-bayesian approach to adaptive classification. *preprint*, 840(2):6, 2003.

Karel Devriendt, Hannah Friedman, Bernhard Reinke, and Bernd Sturmfels. The two lives of the grassmannian. *arXiv preprint arXiv:2401.03684*, 2024.

Ilya Dumer, Mark S Pinsker, and Vyacheslav V Prelov. On coverings of ellipsoids in euclidean spaces. *IEEE transactions on information theory*, 50(10):2348–2356, 2004.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Mathieu Even and Laurent Massoulié. Concentration of non-isotropic random tensors with applications to learning and empirical risk minimization. In *Conference on Learning Theory*, pages 1847–1886. PMLR, 2021.

Kenneth J Falconer. *Techniques in fractal geometry*, volume 3. Wiley Chichester, 1997.

X Fernique. Regularite des trajectoires des fonctions aleatoires gaussiennes. *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*, pages 1–96, 1975.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9(2):473–504, 2020.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jürgen Jost. *Riemannian geometry and geometric analysis*, volume 42005. Springer, 2008.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

Neil Lutz. A note on pointwise dimensions. *arXiv preprint arXiv:1612.05849*, 2016.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Vitali D Milman and Gideon Schechtman. *Asymptotic theory of finite dimensional normed spaces: Isoperimetric inequalities in riemannian manifolds*, volume 1200. Springer Science & Business Media, 1986.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Christopher C Paige and Musheng Wei. History and generality of the cs decomposition. *Linear Algebra and its Applications*, 208:303–326, 1994.

Alain Pajor. Metric entropy of the grassmann manifold. *Convex Geometric Analysis*, 34(181-188): 0942–46013, 1998.

Liam Parker, Emre Onal, Anton Stengel, and Jake Intrater. Neural collapse in the intermediate hidden layers of classification neural networks. *arXiv preprint arXiv:2308.02760*, 2023.

Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

Alessandro Rinaldo and Enxu Yan. Lecture 14: Uniform Bound via Rademacher Complexity. Scribed notes for CMU 36-755: Advanced Statistical Theory I, October 2016. URL `https://www.stat.cmu.edu/~arinaldo/Teaching/36755/F16/Scribed_Lectures/AST_Oct19_Scribe.pdf`.

Zhouxing Shi, Yihan Wang, Huan Zhang, J Zico Kolter, and Cho-Jui Hsieh. Efficiently computing local lipschitz constants of neural networks via bound propagation. *Advances in Neural Information Processing Systems*, 35:2350–2364, 2022.

Leon Simon. Introduction to geometric measure theory. `https://math.stanford.edu/~lms/ntu-gmt-text.pdf`, 2018. NTU Lecture Notes, updated 7 Mar 2018.

Stanislaw J Szarek. Metric entropy of homogeneous spaces. *arXiv preprint math/9701213*, 1997.

Michel Talagrand. Regularity of gaussian processes. *Acta Mathematica*, 159:99–149, 1987.

Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2005.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3): 2–3, 2014.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Peng Wang, Xiao Li, Can Yaras, Zhihui Zhu, Laura Balzano, Wei Hu, and Qing Qu. Understanding deep representation learning via layerwise feature compression and discrimination. *arXiv preprint arXiv:2311.02960*, 2025.

Wikipedia contributors. Change of variables. `https://en.wikipedia.org/wiki/Change_of_variables`, 2025a.

Wikipedia contributors. Min–max theorem. `https://en.wikipedia.org/wiki/Min-max_theorem`, 2025b.

Wikipedia contributors. Square root of a matrix. `https://https://en.wikipedia.org/wiki/Square_root_of_a_matrix`, 2025c.

Wikipedia contributors. Volume of an n-ball. `https://en.wikipedia.org/wiki/Volume_of_an_n-ball`, 2025d.

David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *Mathematics of Operations Research*, 50(1):40–67, 2025.

A Experimental Setup

We introduce detailed experimental setups. We evaluate our Riemannian Dimension bound on two standard architectures—Fully Connected Networks (FCNs) and ResNets, using two benchmark datasets—MNIST [LeCun et al., 1998] and CIFAR-10 [Krizhevsky, 2009], respectively. The architecture of FCNs: we consider a 9-hidden-layer FCN in which the first two hidden layers have width 2^{11} and the remaining seven hidden layers share a common width h, with $h \in \{2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$. The output layer is a linear classifier mapping to 10 logits, and we use ReLU as the activation and use PyTorch's default initialization (Kaiming uniform for ReLU). Increasing h monotonically enlarges both layer widths and the total parameter count, yielding a clean capacity sweep at fixed depth. The architecture of ResNets: we adopt the canonical ResNet architectures, ResNet-20, ResNet-32, ResNet-44, ResNet-56, ResNet-74, and ResNet-110, which differ only in the number of residual blocks per stage while maintaining the same overall architecture (three-stage, basic-block design) as introduced by [He et al., 2016]. Following the practice of [He et al., 2016], we apply BatchNorm and ReLU after each convolution, with shortcut connections added as needed, and a global average pooling layer precedes the final linear classifier. These ResNet architectures provides a clean capacity sweep via depth.

We adopt standard training pipelines widely used in the benchmarks. (1) The training Protocol of FCNs is: SGD with momentum optimizer where momentum $= 0.9$, learning rate $= 0.01$, and weight decay $= 5 \times 10^{-4}$; 200 epochs and 128 batch size; a step decay at epochs $\{100, 170\}$, where the learning rate is scaled by $\times 0.1$. (2) The training Protocol of ResNets is: SGD with momentum optimizer where momentum $= 0.9$, learning rate $= 0.1$, and weight decay $= 5 \times 10^{-4}$; 250 epochs and 128 batch size; a step decay at epochs $\{50, 150, 200\}$, where the learning rate is scaled by $\times 0.1$; Following practical training conditions, we apply standard data augmentation on CIFAR-10: random horizontal flips and 4-pixel random crops with zero-padding.

In the experiments of FCNs and ResNets, to enable layerwise analysis of the evolving feature representations and support our computation of Riemannian Dimension, we register forward hooks on all nonlinearity layers. For layers followed by pooling, we replace the last recorded ReLU activation with the corresponding pooled output. We also pre-register the input hook to capture the feature matrix of the data. These hooks ensures precise extraction of nonlinearity activations at each depth throughout training. We set the hyper-parameter ε via a one–dimensional ternary-search procedure: at the end of each training stage we perform a 500-step ternary search for FCNs and a 50-step ternary search for ResNets over the admissible interval $[\sqrt{1/n},\ \max_{l=1,\ldots,L} \sqrt{\frac{2L\lambda_{\max}(F_{l-1}F_{l-1}^\top)\cdot\|W\|_F^2 \prod_{i>l}\|W_i\|_{\mathrm{op}}^2}{n}}]$. The search selects the value that minimizes our one-shot Riemannian Dimension-based generalization bound (Theorem 1). We note that tighter bounds could be achieved with more refined optimization procedures on ε. For FCNs, we compute full feature gram matrices. While for ResNets, the feature matrix F is formed by flattening each activation map into a vector of dimension $d = C \cdot H \cdot W$, where C, H, W are the channel, height, and width of the feature map respectively. To align with our theory, we simplify ResNets to fully connected (feed-forward) networks when computing our bound; we apply the same simplification to the associated VC-dimension and parameter-count calculations to maintain consistency. To avoid out-of-memory in computing full feature gram matrices in high-dimensional convolutional layers, we use the standard Gaussian sketching approximation, where each feature gram matrix uses a Gaussian sketch with parameter $r = \min(8192, \lfloor d/8 \rfloor)$ [Woodruff et al., 2014]. By standard subspace-embedding guarantees, such Gaussian sketches preserve Gram quadratic forms—and hence the spectra—of the feature matrices with high probability, introducing only negligible distortion and leaving our conclusions unchanged [Woodruff et al., 2014].

B Proofs for Pointwise Generalization Framework (Section 2)

B.1 Proof of Theorem 1 (One-Shot Generalization Bound)

To prove our pointwise-dimension generalization bound, we start with a novel combination of symmetrization and conditional PAC-Bayes, which enables data-dependent prior and empirical L_1/L_2 metric control. In the next step, we show how a standard analysis of symmetrized loss extends this result to original losses, incurring at most a constant factor of 2 in the bound.

B.1.1 Symmetrized Losses and Conditional PAC-Bayes

As discussed above, we start with a one-shot approximation bound for symmetrized loss. Now define $g(f; z, \xi) = \xi\ell(f; z)$, where ξ is an independent Rademacher variable, a symmetric binary random variable taking values with $\Pr(\xi = +1) = \Pr(\xi = -1) = \frac{1}{2}$. Denote the population and empirical distributions of ξ as $\mathbb{Q}$ and $\mathbb{Q}_n$. By independence assumption we know that $\mathbb{Q}$ is also the conditional distribution of (z, ξ) conditioned on z, thus we can use the product distribution (such as $\mathbb{Q}_n \otimes \mathbb{P}_n$) to denote the joint distributions of (z, ξ). A key property of symmetrized loss is that it preserves the empirical L_2 and L_1 metric of the original loss:

$$\varrho_{n,\ell}(f, f') := \sqrt{\mathbb{P}_n[(\ell(f; z) - \ell(f'; z))^2]} = \sqrt{\mathbb{Q}_n \otimes \mathbb{P}_n[(g(f; z, \xi) - g(f'; z, \xi))^2]}, \tag{13}$$

$$L_1(\mathbb{P}_n)(f, f') = \mathbb{P}_n|\ell(f; z) - \ell(f'; z)| = \mathbb{Q}_n \otimes \mathbb{P}_n|\ell(f; z) - \ell(f'; z)|, \tag{14}$$

because for Rdemacher variables ξ, $(\ell(f; z) - \ell(f'; z))^2 = [\xi(\ell(f; z) - \ell(f'; z))]^2 = (g(f; z, \xi) - g(f'; z, \xi))^2$, and $|\ell(f; z) - \ell(f'; z)| = |\xi(\ell(f; z) - \ell(f'; z))| = |g(f; z, \xi) - g(f'; z, \xi)|$.

We are interested in the conditional generalization gap (conditioned on every fixed realization of $\{z_i\}_{i=1}^n$ regardless of $\mathbb{P}$):

$$\mathbb{Q} \otimes \mathbb{P}_n g(f; z, \xi) - \mathbb{Q}_n \otimes \mathbb{P}_n g(f; z, \xi) \;=\; 0 \;-\; \frac{1}{n}\sum_{i=1}^n g\big(f; z_i, \xi_i\big).$$

For technical reasons (the later application of McDiarmid's inequality via bounded difference to prove Theorem 1), we work with $L_1(\mathbb{P}_n)$ metric now (which is a smaller metric than $L_2(\mathbb{P}_n)$ so the bound is more general).

Lemma 4 *Let π be any data-dependent prior on a hypothesis class $\mathcal{F}$ (π can depend on $z_1, \cdots, z_n$), and let $\ell\colon \mathcal{F} \times \mathcal{Z} \to [0, 1]$ be a bounded loss. Fix confidence $\delta \in (0, 1)$ and sample size n. Then for all $\varepsilon > 0$, with probability at least $1 - \delta$ over n i.i.d. draws $\xi_1, \ldots, \xi_n$ from $\mathbb{Q}$, uniformly over all $f \in \mathcal{F}$,*

$$0 - \frac{1}{n}\sum_{i=1}^n g(f; z_i, \xi_i) \le \varepsilon + \sqrt{\frac{\log \frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f, \varepsilon))} + \log \frac{1}{\delta}}{2n}}.$$

Proof of Lemma 4: The proof consists of three steps: 1. Conditional PAC-Bayes; 2. Approximating the Deterministic Hypothesis; and 3. Taking Uniform Posterior in $\varepsilon-$Ball.

1. Conditional PAC-Bayes. Applying Catoni's PAC-Bayes uniform bound [Catoni, 2003] (Lemma 9 in our Auxiliary Lemma Part) to the population and empirical distribution of ξ, we have that with probability at least $1-\delta$, uniformly over every $\mu \in \Delta(\mathcal{F})$, we have the "random hypothesis" bound

$$\langle \mu, 0 - \mathbb{Q}_n \otimes \mathbb{P}_n g(\cdot; z, \xi) \rangle \leq \sqrt{\frac{\mathrm{KL}\big(\mu, \pi\big) \ + \log \frac{1}{\delta}}{2n}}, \tag{15}$$

where the absolute constant comes from the loss $g \in [-1, 1]$. The KL divergence is defined by $\mathrm{KL}(\mu \| \pi) = \int_{\mathcal{F}} \log\Big(\frac{d\mu}{d\pi}(f)\Big)\, \mu(df)$.

2. Approximating the Deterministic Hypothesis. So on the event that the above conditional PAC-Bayes inequality (15) holds, with probability at least $1-\delta$, we have that uniformly over every random $\mu \in \Delta(\mathcal{F})$ and at the same time uniformly over every deterministic $f \in \mathcal{F}$, for every $\eta > 0$, we have the "deterministic hypothesis" bound:

$$\begin{aligned}
&0 - \mathbb{Q}_n \otimes \mathbb{P}_n g(f; z, \xi) \\
=&\langle \mu, 0 - \mathbb{Q}_n \otimes \mathbb{P}_n g(\cdot; z, \xi) \rangle + \langle \mu, \mathbb{Q}_n \otimes \mathbb{P}_n [g(\cdot; z, \xi) - g(f; z, \xi)] \rangle \\
\leq&\frac{\eta}{8} + \frac{\mathrm{KL}\big(\mu, \pi\big) \ + \log \frac{1}{\delta}}{\eta n} + \langle \mu, \mathbb{Q}_n \otimes \mathbb{P}_n [|g(\cdot; z, \xi) - g(f; z, \xi)|] \rangle \\
=&\frac{\eta}{8} + \frac{\mathrm{KL}\big(\mu, \pi\big) \ + \log \frac{1}{\delta}}{\eta n} + \langle \mu, L_1(\mathbb{P}_n)(\cdot, f) \rangle,
\end{aligned} \tag{16}$$

where the first identity holds because $\langle \mu, \mathbb{Q}_n \otimes \mathbb{P}_n g(f; z, \xi) \rangle = g(f; z, \xi)$; the inequality uses conditional PAC-Bayes (15) to bound the first term and use absolute values to bound the second term; and the last identity uses metric preservation of symmetrization (identity (13)). On the event that (15) holds, the new inequality (16) holds simultaneously not only for every random μ (due to PAC-Bayes) but also for every deterministic f, because on such event, the new inequality (16) does not use any new probabilistic inequality besides (15).

3. Taking Uniform Posterior in $\varepsilon-$ball. Given any prior π on $\mathcal{F}$, for any $f \in \mathcal{F}$, we take μ to be the uniform distribution over $B_{L_1(\mathbb{P}_n)}(f, \varepsilon)$:

$$\mu = \mathrm{Unif}(B_{L_1(\mathbb{P}_n)}(f, \varepsilon)). \tag{17}$$

We emphasize that using this uniform measure over the ε-ball is "essentially optimal" in that it reproduces the same analytical upper bound—namely the conclusion of Lemma 4—as would be obtained by employing the Gibbs distribution that minimizes the bound in (16).

This "essentially optimal" choice (17) of μ yields

$$\frac{d\mu}{d\pi}(f) = \begin{cases} \dfrac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f, \varepsilon))}, & f \in B_{L_1(\mathbb{P}_n)}(f, \varepsilon), \\ 0, & f \notin B_{L_1(\mathbb{P}_n)}(f, \varepsilon). \end{cases}$$

The KL divergence is

$$\begin{aligned}
\mathrm{KL}(\mu\|\pi) &= \int_{\mathcal{F}} \log\Big(\tfrac{d\mu}{d\pi}(f)\Big)\,\mu(df) \\
&= \int_{B_{L_1(\mathbb{P}_n)}(f,\varepsilon)} \log\Big(\tfrac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}\Big)\,\mu(df) \\
&= \log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}\,\underbrace{\int_{B_{L_1(\mathbb{P}_n)}(f,\varepsilon)} \mu(df)}_{=1} \\
&= \log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}. \qquad (18)
\end{aligned}$$

Because (16) holds simultaneously for every f and every $\mu = \mathrm{Unif}(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))$, we can apply (16) and substitute KL equality (18) back to (16). Minimizing over η, this gives:

$$0 - \mathbb{Q}_n g(f;z,\xi) \le \sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))} + \log\frac{1}{\delta}}{2n}} + \varepsilon.$$

□

In summary, the proof develops the generalization gap as an optimization problem on bias and variance. Through introducing symmetrized losses and a conditioning technique, we transformer the bias-variance optimization to a PAC-Bayes optimization problem on the user-chosen posterior. The following lemma indicates that the uniform posterior is optimal up to the min–max gap: the lower bound $\min\{a,\varepsilon\}$ and the upper bound $\max\{a,\varepsilon\}$ bracket the optimum, coincide when $a=\varepsilon$, and have the same order whenever a and ε are comparable.

Lemma 5 (Optimality of the ε-ball uniform posterior for the PAC–Bayes optimization) *Define the PAC–Bayes optimization functional*

$$V_f(\mu;\pi) \;:=\; \inf_{\mu\ll\pi}\Big\{\frac{\eta}{8} + \frac{\mathrm{KL}(\mu,\pi) + \log\frac{1}{\delta}}{\eta n} + \langle \mu, L_1(\mathbb{P}_n)(\cdot,f)\rangle\Big\}.$$

Then for every $f\in\mathcal{F}$, $\eta>0$, and $\varepsilon>0$,

$$V_f(\mu;\pi) \;\ge\; \frac{\eta}{8} + \frac{\log\frac{1}{\delta}}{\eta n} + \min\Big\{\frac{1}{\eta n}\log\frac{1}{\pi\big(B_{L_1(\mathbb{P}_n)}(f,\varepsilon)\big)},\ \varepsilon\Big\} \;-\; \frac{\log 2}{\eta n}. \qquad (19)$$

Consequently, for every $f\in\mathcal{F}$, $\eta>0$, and $\varepsilon>0$,

$$\frac{\eta}{8} + \frac{\log\frac{1}{\delta}}{\eta n} + \min\Big\{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{\eta n},\ \varepsilon\Big\} \;-\; \frac{\log 2}{\eta n} \;\le\; V_f(\mu;\pi) \;\le\; \frac{\eta}{8} + \frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))} + \log\frac{1}{\delta}}{\eta n} + \varepsilon. \qquad (20)$$

Proof of Lemma 5 The Donsker–Varadhan variational identity states that for any measurable h,

$$-\log\int e^h\,d\pi \;=\; \inf_{\mu\ll\pi}\Big\{D_{\mathrm{KL}}(\mu\|\pi) \;-\; \int h\,d\mu\Big\}.$$

Apply it with $h = -\eta n L_1(\mathbb{P}_n)(\cdot, f)$ to obtain

$$-\log \int e^{-\eta n L_1(\mathbb{P}_n)(\cdot,f)}\, d\pi \;=\; \inf_{\mu \ll \pi} \Big\{ D_{\mathrm{KL}}(\mu \| \pi) \;+\; \int \eta n L_1(\mathbb{P}_n)(\cdot, f)\, d\mu \Big\},$$

which implies that

$$\frac{\eta}{8} + \frac{\log \frac{1}{\delta}}{\eta n} - \frac{1}{\eta n} \log \int e^{-\eta n L_1(\mathbb{P}_n)(\cdot,f)}\, d\pi \;=\; \inf_{\mu \ll \pi} \Big\{ \frac{\eta}{8} + \frac{\mathrm{KL}\big(\mu, \pi\big) \; + \log \frac{1}{\delta}}{\eta n} + \langle \mu, L_1(\mathbb{P}_n)(\cdot, f) \rangle \Big\}. \quad (21)$$

By splitting the dual integral,

$$\begin{aligned}
\int e^{-\eta n L_1(\mathbb{P}_n)(\cdot,f)}\, d\pi &= \int_{B_{L_1(\mathbb{P}_n)}(f,\varepsilon)} e^{-\eta n L_1(\mathbb{P}_n)(\cdot,f)}\, d\pi + \int_{B_{L_1(\mathbb{P}_n)}(f,\varepsilon)^c} e^{-\eta n L_1(\mathbb{P}_n)(\cdot,f)}\, d\pi \\
&\le \pi\big(B_{L_1(\mathbb{P}_n)}(f,\varepsilon)\big) + e^{-\eta n \varepsilon}\big(1 - \pi\big(B_{L_1(\mathbb{P}_n)}(f,\varepsilon)\big)\big) \\
&\le \pi\big(B_{L_1(\mathbb{P}_n)}(f,\varepsilon)\big) + e^{-\eta n \varepsilon},
\end{aligned}$$

where $B_{L_1(\mathbb{P}_n)}(f,\varepsilon)^c$ is complement of $B_{L_1(\mathbb{P}_n)}(f,\varepsilon)$; where we have used $e^{-\eta n L_1(\mathbb{P}_n)(\cdot,f)} \le 1$ on $B_{L_1(\mathbb{P}_n)}(f,\varepsilon)$ and $e^{-\eta n L_1(\mathbb{P}_n)(\cdot,f)} \le e^{-\eta n \varepsilon}$ on $B_{L_1(\mathbb{P}_n)}(f,\varepsilon)^c$. Hence

$$V_f(\mu;\pi) \;\ge\; \frac{\eta}{8} + \frac{\log \frac{1}{\delta}}{\eta n} - \frac{1}{\eta n} \log\Big(\pi\big(B_{L_1(\mathbb{P}_n)}(f,\varepsilon)\big) + e^{-\eta n \varepsilon}\Big). \quad (22)$$

The simplified form (19) follows from $a + b \le 2\max\{a, b\}$ or equivalently $-\log(a+b) \ge -\log 2 + \min\{-\log a, -\log b\}$ on (22). Combining (16), (18) and (19) yields the sandwich (20). □

B.1.2 Original Losses

In the next step, via symmetrization, we extend the bound for symmetrized losses (Lemma 4) to original losses.

Proof of Theorem 1: The proof consists of three steps, following the standard symmetrization routine in controlling empirical processes: 1. Obtain In-Expectation Bound for Offset Symmetrized Losses; 2. Bounding Offset Symmetrized Losses by Offset Original Losses; 3. Applying McDiarmid's Inequality to Get High-Probability Bound.

1. Obtain In-Expectation Bound for Offset Symmetrized Losses. The bound in Lemma 4 implies

$$\Pr\left(\sup_{f \in \mathcal{F}} \left\{ 0 - \frac{1}{n} \sum_{i=1}^{n} g(f; z_i, \xi_i) - \sqrt{\frac{\log \frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))} + \log \frac{1}{\delta}}{2n}} - \varepsilon \right\} > 0 \right) \le \delta$$

which implies (by $\mathbb{Q} \otimes \mathbb{P} g(f; z, \xi) = 0$ and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$)

$$\Pr\left(\sup_{f \in \mathcal{F}} \left\{ (\mathbb{Q} \otimes \mathbb{P} - \mathbb{Q}_n \mathbb{P}_n) g(f; z, \xi) - \sqrt{\frac{\log \frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}} - \varepsilon \right\} > \sqrt{\frac{\log \frac{1}{\delta}}{n}} \right) \le \delta.$$

Denoted by $\sup_{f\in\mathcal{F}}\left\{(\mathbb{Q}\otimes\mathbb{P}-\mathbb{Q}_n\mathbb{P}_n)g(f;z,\xi)-\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-\varepsilon\right\}:=Y$, then

$$\mathbb{E}Y\leq\mathbb{E}Y_+=\int_0^\infty \Pr(Y>t)dt\leq\int_0^\infty e^{-nt^2}dt=\sqrt{\frac{\pi}{4n}}<\sqrt{\frac{1}{n}}.$$

Thus, we now get an expectation bound.

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{(\mathbb{Q}\otimes\mathbb{P}-\mathbb{Q}_n\mathbb{P}_n)g(f;z,\xi)-\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-\varepsilon-\sqrt{\frac{1}{n}}\right\}\leq 0. \tag{23}$$

2. Bounding Offset Symmetrized Losses by Offset Original Losses. Now we apply the following analysis, standard in using symmetrization to control empirical processes:

$$\begin{aligned}
&\mathbb{E}_z\sup_{f\in\mathcal{F}}\left\{(\mathbb{P}-\mathbb{P}_n)\ell(f;z)-2\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-2\varepsilon-\sqrt{\frac{4}{n}}\right\}\\
=&\mathbb{E}_z\sup_{f\in\mathcal{F}}\mathbb{E}_{z'}\left\{\mathbb{P}_n[(\ell(f;z')-\ell(f;z))]-2\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-2\varepsilon-\sqrt{\frac{4}{n}}\right\}\\
\leq&\mathbb{E}_{z,z'}\sup_{f\in\mathcal{F}}\left\{\mathbb{P}_n[(\ell(f;z')-\ell(f;z))]-2\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-2\varepsilon-\sqrt{\frac{4}{n}}\right\}\\
=&\mathbb{E}_{z,z',\xi}\sup_{f\in\mathcal{F}}\left\{\mathbb{Q}_n\otimes\mathbb{P}_n[\xi(\ell(f;z')-\ell(f;z))]-2\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-2\varepsilon-\sqrt{\frac{4}{n}}\right\}\\
\leq&\mathbb{E}_{z',\xi}\sup_{f\in\mathcal{F}}\left\{\mathbb{Q}_n\otimes\mathbb{P}_n\xi\ell(f;z')-\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-\varepsilon-\sqrt{\frac{1}{n}}\right\}\\
&+\mathbb{E}_{z,\xi}\sup_{f\in\mathcal{F}}\left\{\mathbb{Q}_n\otimes\mathbb{P}_n(-\xi\ell(f;z))-\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-\varepsilon-\sqrt{\frac{1}{n}}\right\}\\
=&2\mathbb{E}_{z,\xi}\sup_{f\in\mathcal{F}}\left\{(\mathbb{Q}\otimes\mathbb{P}-\mathbb{Q}_n\otimes\mathbb{P}_n)g(f;z,\xi)-\sqrt{\frac{\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{2n}}-\varepsilon-\sqrt{\frac{1}{n}}\right\}\\
\leq&0,
\end{aligned}$$

where the first inequality is because sup is a convex function; the second equality is because $\ell(f;z')-\ell(f;z)$ has a symmetric distribution; the second inequality is because $\sup_x(h_1(x)+h_2(x))\leq$

$\sup_x h_1(x) + \sup_x h_2(x)$; and the last equality is because $\xi\ell(f;z')$ has the same distribution with $\xi\ell(f,z)$; and the last inequality is by (23). We conclude that

$$\mathbb{E}_z \sup_{f\in\mathcal{F}} \left\{ (\mathbb{P}-\mathbb{P}_n)\ell(f;z) - \inf_{\varepsilon>0}\left\{ \sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} - \sqrt{\frac{4}{n}} \right\} \le 0.$$

3. Applying McDiarmid's Inequality to Get High-Probability Bound. Finally, we apply McDiarmid's inequality (Lemma 10) to get the high probability bound. By an arbitrary change of the sample $S = \{z_1, ..., z_n\}$ to $S' = \{z_1, ..., z_j', ..., z_n\}$, the bounded difference of $\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\ell(f;z) - \inf_{\varepsilon>1/n}\left\{\sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{n}} + 2\varepsilon\right\} - \sqrt{\frac{4}{n}}$ is at most $3/n$. Specifically, define

$$\psi(f;S) = (\mathbb{P}-\mathbb{P}_n)\ell(f;z) - \inf_{\varepsilon>1/n}\left\{ \sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} - \sqrt{\frac{4}{n}},$$

we want to show

$$|\sup_f \psi(f;S) - \sup_f \psi(f;S')| \le \frac{3}{n}. \tag{24}$$

Clearly, to prove (24), it suffices to show that for every fixed $f\in\mathcal{F}$,

$$|\psi(f,S) - \psi(f,S')| \le \frac{3}{n}, \quad \forall \text{ fixed } f.$$

Denote $L_1(S)$ and $L_1(S')$ to be the empirical L_1 norms with S and S', respectively (abbreviated from $L_1(\mathbb{P}_n(S))$ and $L_1(\mathbb{P}_n(S'))$). We have that $|\psi(f;S) - \psi(f;S')|$ is bounded by

$$\left| \frac{\ell(f;z') - \ell(f;z)}{n} + \inf_{\varepsilon>1/n}\left\{ \sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(S')}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} - \inf_{\varepsilon>1/n}\left\{ \sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(S)}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} \right|, \tag{25}$$

For all $\varepsilon > 1/n$, as ℓ is bounded in $[0,1]$,

$$B_{L_1(S)}(f,\varepsilon-1/n) \subseteq B_{L_1(S')}(f,\varepsilon) \subseteq B_{L_1(S)}(f,\varepsilon+1/n),$$

so we have

$$\log\frac{1}{\pi(B_{L_1(S)},\varepsilon+1/n)} \le \log\frac{1}{\pi(B_{L_1(S')},\varepsilon)} \le \log\frac{1}{\pi(B_{L_1(S)},\varepsilon-1/n)}$$

and thus (taking $\varepsilon_1 = \varepsilon - 1/n$ and $\varepsilon_2 = \varepsilon + 1/n$.

$$\begin{aligned}\inf_{\varepsilon_1>2/n}\left\{ \sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(S)}(f,\varepsilon_1))}}{n}} + 2(\varepsilon_1 - 1/n) \right\} &\le \inf_{\varepsilon>1/n}\left\{ \sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(S')}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} \\ &\le \inf_{\varepsilon_2>0}\left\{ \sqrt{\frac{2\log\frac{1}{\pi(B_{L_1(S')}(f,\varepsilon_2))}}{n}} + 2(\varepsilon+1/n) \right\}.\end{aligned}$$

This sandwich bound implies that

$$\left| \inf_{\varepsilon > 1/n} \left\{ \sqrt{\frac{2 \log \frac{1}{\pi(B_{L_1(S')}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} - \inf_{\varepsilon > 1/n} \left\{ \sqrt{\frac{2 \log \frac{1}{\pi(B_{L_1(S)}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} \right| \leq \frac{2}{n}.$$

Combining the above bound with (25) we prove the bounded difference property (24). Thus, by McDiarmid's inequality (Lemma 10), from the bounded difference property (24), with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left\{ (\mathbb{P} - \mathbb{P}_n)\ell(f;z) - \inf_{\varepsilon > 1/n} \left\{ \sqrt{\frac{2 \log \frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} - \sqrt{\frac{4}{n}} \right\} \\
\leq & \mathbb{E}_z \sup_{f\in\mathcal{F}} \left\{ (\mathbb{P} - \mathbb{P}_n)\ell(f;z)) - \inf_{\varepsilon > 1/n} \left\{ \sqrt{\frac{2 \log \frac{1}{\pi(B_{L_1(\mathbb{P}_n)}(f,\varepsilon))}}{n}} + 2\varepsilon \right\} - \sqrt{\frac{4}{n}} \right\} + \ 3\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \\
\leq & 3\sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\end{aligned}$$

Finally, we convert the stronger $L_1(\mathbb{P}_n)$ bound to the weaker $L_2(\mathbb{P}_n)$ bound we stated in the main paper: combining the above $L_1(\mathbb{P}_n)$ bound with the fact $L_1(\mathbb{P}_n) \leq L_2(\mathbb{P}_n)$, we have that with probability at least $1-\delta$, uniformly over every $f \in \mathcal{F}$,

$$(\mathbb{P} - \mathbb{P}_n)\ell(f;z) \leq \inf_{\varepsilon > \sqrt{1/n}} \left\{ 2\varepsilon + \sqrt{\frac{2 \log \frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}{n}} + \frac{4}{n} + 3\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right\}.$$

$\square$

B.2 Proof for Theorem 2 (Generic Chaining Upper and Lower Bounds)

Theorem 2 is a corollary of established matching upper and lower bounds for Gaussian Processes [Fernique, 1975, Talagrand, 1987]. Note that generic chaining have several equivament formulations, and the one closest to our purpose is through majorizing measure. On the upper bound side, we extend it to pointwise generalization bound; on the lower bound side, we re-derive it to empirical processes.

B.2.1 Background on Gaussian Processes

We now recall several key results from a series of seminal papers by Talagrand, Fernique, and others, which introduces the majorizing-measure formulation of the generic chaining framework.

The statements below follow Talagrand's original exposition [Talagrand, 1987, "Regularition of Gaussian Processes," Acta Mathematica, 1987] closely, with only minor adjustments in notation and wording.

A *centered Gaussian random variable* X is a real-valued measurable function on the outcome space such that the law of X has density

$$(2\pi\sigma^2)^{-1/2}\,\exp\Big(-\tfrac{x^2}{2\sigma^2}\Big).$$

The law of X is thus determined by $\sigma = (\mathbb{E}[X^2])^{1/2}$. If $\sigma = 1$, X is called *standard normal*.

A *Gaussian process* is a family $\{X_t\}_{t\in T}$ of random variables indexed by some set T, such that every finite linear combination $\sum_{j=1}^{k}\alpha_j X_{t_j}$ is Gaussian. Its covariance function

$$\Gamma(u,v) \;=\; \mathbb{E}[\,X_u\,X_v\,], \quad (u,v)\in T\times T,$$

determines $\mathbb{E}\big[\sum_j \alpha_j X_{t_j}\big]^2$, and hence the joint law of $\{X_t\}_{t\in T}$. On the index set T, consider the pseudo-distance ϱ given by

$$\varrho(u,v) = \sqrt{\mathbb{E}[(X_u - X_v)^2]}.$$

Gaussian processes are thus a very rigid class of stochastic processes, with exceptionally nice properties that have been fully developed in the literature.

Fernique [1975] proved the following result.

Lemma 6 (Upper Bound of Gaussian Processes via Majorizing Measure, Fernique [1975]) *For any Gaussian process $(X_t)_{t\in T}$, and any prior π on T, we have*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le C\sup_{x\in T}\int_0^\infty \sqrt{\log\frac{1}{\pi(B_\varrho(x,\varepsilon))}}d_\varepsilon,$$

where $C > 0$ is an absolute constant.

A prior π that makes the right hand side in Lemma 6 finite is called a *majorizing measure*. Fernique conjectured as early as 1974 that the existence of majorizing measures might characterize the boundedness of Gaussian processes. He proved a number of important partial results, and his determination eventually motivated the Talagrand to attack the problem in 1987. Talagrand [1987] proved that the integral in Lemma 6 is tight up to absolute constants. We define the following minor measurability condition:

Assumption 1 (Minor Measurability Condition) *The Gaussian process satisfies* $\sup_{t\in D}|X_t| < \infty$ *almost surely for each countable subset $D \subseteq T$.*

Lemma 7 (Lower Bound of Gaussian Processes via Majorzing Measure, Talagrand [1987]) *For every Gaussian process that satisfies a mild measurability condition (Assumption 1), there exists a prior π on T such that*

$$\mathbb{E}\left[\sup_{x\in T} X_t\right] \ge c\sup_{x\in T}\int_0^\infty \sqrt{\log\frac{1}{\pi(B_\varrho(x,\varepsilon))}}d\varepsilon,$$

where $c > 0$ is an absolute constant.

Thus the pointwise dimension integral gives a complete characterization to the supremum of Gaussian process. We now use the upper and lower bounds in Lemma 6 and Lemma 7 to prove Theorem 2.

B.2.2 Proof of Theorem 2

I: Proof of the Upper Bound. We begin by citing a general principle for converting (localized) uniform convergence guarantees into pointwise generalization bounds. This conversion, introduced as "uniform localized convergence" principle in [Xu and Zeevi, 2025], provides a direct mechanism for obtaining the type of pointwise generalization bounds central to our work. We state this result as "uniform pointwise convergence" principle.

Lemma 8 ("Uniform Pointwise Convergence" Principle, Xu and Zeevi [2025]) *(A Generic Conversion from Uniform Convergence to Pointwise Generalization.) For a function class $\mathcal{F}$ and functional $d : \mathcal{F} \to (0, R]$, assume there is a function $\psi(r; \delta)$, which is non-decreasing with respect to r, non-increasing with respect to δ, and satisfies that $\forall \delta \in (0, 1)$, $\forall r \in [0, R]$, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}: d(f) \leq r} (\mathbb{P} - \mathbb{P}_n)\ell(f; z) \leq \psi(r; \delta). \tag{26}$$

Then, given any $\delta \in (0, 1)$ and $r_0 \in (0, R]$, with probability at least $1 - \delta$, uniformly over all $f \in \mathcal{F}$,

$$(\mathbb{P} - \mathbb{P}_n)\ell(f; z) \leq \psi\left(\max\{2d(f), r_0\}; \delta\left(\log_2 \frac{2R}{r_0}\right)^{-1}\right). \tag{27}$$

The proof of the upper bound in Theorem 2 consists of three steps: 1. Bounding Empirical Process by Gaussian Process; 2. Applying Integral Upper Bound; 3. Generic Conversion to Pointwise Generalization Bound.

Step 1: Bounding Empirical Process by Gaussian Process. We start with the classical result that Gaussian complexity of a function class is an upper bound for expected uniform convergence, see Definition 2 and Lemma 11 in the auxiliary lemma part for this classical result. To be specific, by Lemma 11 (and $\sqrt{2\pi} < 3$) we have

$$\mathbb{E}_z\left[\sup_{f \in \mathcal{F}} (\mathbb{P} - \mathbb{P}_n)\ell(f; z)\right] \leq 3\mathbb{E}_{g,z}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n g_i \ell(f; z_i)\right],$$

where g_i are i.i.d. standard Gaussian variables. Applying Mcdiarmid's inequality (Lemma 10) twice, we have that with probability at least $1 - \delta$, the following two inequality simultaneously hold:

$$\sup_{f \in \mathcal{F}} (\mathbb{P} - \mathbb{P}_n)\ell(f; z) - \mathbb{E}_z\left[\sup_{f \in \mathcal{F}} (\mathbb{P} - \mathbb{P}_n)\ell(f; z)\right] \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}},$$

$$\mathbb{E}_{g,z}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n g_i \ell(f; z_i)\ell(f; z)\right] - \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n g_i \ell(f; z_i)\ell(f; z) \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Combining the above three inequalities we have

$$\sup_{f \in \mathcal{F}} (\mathbb{P} - \mathbb{P}_n)\ell(f; z) \leq 3 \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n g_i \ell(f; z_i) + \sqrt{\frac{8 \log \frac{2}{\delta}}{n}},$$

taking expectation for $\{g_i\}_{i=1}^n$ on both sides, we obtain

$$\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\ell(f;z) \leq 3\mathbb{E}_g\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)\right] + \sqrt{\frac{8\log\frac{2}{\delta}}{n}}, \tag{28}$$

Step 2: Applying Integral Upper Bound. Applying Lemma 6 to the Gaussian process $\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)$ (with fixed $\{z_i\}$), we have that for all data-dependent prior π,

$$\mathbb{E}_g\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)\right] \leq \frac{C_1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon, \tag{29}$$

where $C_1 > 0$ is an absolute constant.

Step 3: Generic Conversion to Pointwise Generalization Bound. Combining (28) and (29), we have: for all $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\ell(f;z) \leq \frac{3C_1}{\sqrt{n}}\sup_{f\in\mathcal{F}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon + \sqrt{\frac{8\log\frac{2}{\delta}}{n}},$$

which implies that $\forall\delta \in (0,1)$ and $\forall r \in (0,n]$, with probability at least $1-\delta$

$$\sup_{f:d(f)\leq r}(\mathbb{P}-\mathbb{P}_n)\ell(f;z) \leq C_2\sqrt{\frac{r}{n}} + \sqrt{\frac{8\log\frac{2}{\delta}}{n}}, \tag{30}$$

where the functional $d : \mathcal{F} \to [0,n]$ is defined by

$$d(f) = \min\left\{\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon, \sqrt{n}\right\}^2,$$

and $C_2 = \max\{3C_1, 1\}$ is an absolute constant (C_1 controls (30) for $r < n$ and 1 controls (30) for $r = n$). The inequality (30) is precisely the condition (26) in the generic conversion provided in Lemma 8. Thus applying Lemma 8 we have the pointwise generalization bound: for any $\delta \in (0,1)$, and $r_0 = 1/n$, with probability at least $1-\delta$, uniformly over all $f \in \mathcal{F}$,

$$\begin{aligned}(\mathbb{P}-\mathbb{P}_n)\ell(f;z) \leq& \frac{C_2}{\sqrt{n}}\max\left\{2\min\left\{\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon, \sqrt{n}\right\}, \frac{1}{n}\right\} + \sqrt{\frac{8\log\frac{\log_2(4n^2)}{\delta}}{n}} \\ \leq& \frac{2C_2}{\sqrt{n}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon + \frac{C_2}{n^{1.5}} + \sqrt{\frac{8\log\frac{\log_2(4n^2)}{\delta}}{n}}. \\ \leq& C\left(\frac{1}{\sqrt{n}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right),\end{aligned}$$

where $C > 0$ is an absolute constant, where the second inequality uses $\min\{\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon, \sqrt{n}\} \leq \sqrt{n}$.

II: Proof of the Lower Bound. We use the classical result that the expected uniform convergence is lower bounded by Gaussian complexity of the centered class, up to a $\sqrt{\log n}$ factor, see Definition 2 and Lemma 12 in the auxiliary lemma part for this classical result. To be specific, by Lemma 12 we have that

$$\begin{aligned}\mathbb{E}_z\left[\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\ell(f;z)\right] &\geq \frac{c_1}{\sqrt{\log n}}\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i(\ell(f;z_i)-\mathbb{E}_z[\ell(f;z)])\right]\\ &\geq \frac{c_1}{\sqrt{\log n}}\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)-\left|\frac{1}{n}\sum_{i=1}^n g_i\right|\cdot\sup_{\mathcal{F}}\mathbb{E}[\ell(f;z)]\right]\\ &= \frac{c_1}{\sqrt{\log n}}\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)\right]-\frac{c_1}{\sqrt{\log n}}\sqrt{\frac{2}{\pi n}}\sup_{\mathcal{F}}\mathbb{E}[\ell(f;z)],\end{aligned}\tag{31}$$

where $c_1>0$ is an absolute constant, and the equality use the fact that $\mathbb{E}[|Y|]=\sqrt{\frac{2}{\pi n}}$ for $Y\sim N(0,1/n)$.

Now applying Lemma 7 to lower bounding the Gaussian process $\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)$ by the integral, we have for any $\{z_i\}_{i=1}^n$,

$$\mathbb{E}_g\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)\right]\geq c_2\inf_\pi\sup_{f\in\mathcal{F}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon,$$

taking expectation on both side yields

$$\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)\right]\geq c_2\mathbb{E}\inf_\pi\sup_{f\in\mathcal{F}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon.\tag{32}$$

Combining (31) and (32), we have that there exist absolute constants $c,c'>0$ such that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\ell(f;z)\right]\geq\frac{c}{\sqrt{n\log n}}\mathbb{E}\inf_\pi\sup_{f\in\mathcal{F}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon-\frac{c'\sup_{\mathcal{F}}\mathbb{E}[\ell(f;z)]}{\sqrt{n\log n}}.$$

This inequality implies the following result

$$\mathbb{E}\left[\inf_\pi\sup_{f\in\mathcal{F}}\left((\mathbb{P}-\mathbb{P}_n)\ell(f;z)-\frac{c}{\sqrt{n\log n}}\int_0^\infty\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon\right)+\frac{c'\sup_{\mathcal{F}}\mathbb{E}[\ell(f;z)]}{\sqrt{n\log n}}\right]\geq 0,$$

where we have used the fact that $\sup_x h_1(x)-\sup_x h_2(x)\leq\sup_x(h_1(x)-h_2(x))$. $\square$

B.3 Auxiliary Lemmas for Section 2

Lemma 9 (PAC–Bayes Bound [Catoni, 2003]; see also Theorem 2.1 in Alquier et al. [2024]) *Let π be a prior on a hypothesis class $\mathcal{F}$ independent to the data, and let $\ell\colon\mathcal{F}\times\mathcal{Z}\to[0,B]$ be a bounded loss. Fix confidence $\delta\in(0,1)$ and sample size n. Then for every $\eta>0$, with probability at least $1-\delta$ over n i.i.d. draws $z_1,\dots,z_n\sim\mathbb{P}$, for* every *distribution μ on $\mathcal{F}$ simultaneously,*

$$(\mathbb{P}-\mathbb{P}_n)\ell(f;z)\ \leq\ \sqrt{\frac{B^2\big(\mathrm{KL}(\mu,\pi)\ +\log\frac{1}{\delta}\big)}{8n}}.$$

Lemma 10 (McDiarmid's inequality [McDiarmid, 1998]) *Suppose that $z_1, ..., z_n \in \mathcal{Z}$ are independent, and $h : \mathcal{Z}^n \to \mathbb{R}$. Let $c_1, ..., c_n$ satisfy*

$$\sup_{z_1,...,z_n,z_i'} |h(z_1, ..., z_{i-1}, z_i, z_{i+1}, ..., z_n) - h(z_1, ..., z_{i-1}, z_i', z_{i+1}, .., z_n)| \le c_i,$$

for $i = 1, ..., n$. Then

$$\Pr(h(z_1, \cdots, z_n) - \mathbb{E}[h(z_1, \cdots, z_n)] \ge t) \le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Definition 2 (Rademacher and Gaussian complexities) *For a function class $\mathcal{F}$ that consists of mappings from $\mathcal{Z}$ to $\mathbb{R}$, define the Rademacher complexity of $\mathcal{F}$ as*

$$R_n(\mathcal{F}) := \mathbb{E}_{z,\xi}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \xi_i f(z_i)\right],$$

where ξ_i are i.i.d. Rademacher variables; and define the Gaussian complexity of $\mathcal{F}$ as

$$G_n(\mathcal{F}) := \mathbb{E}_{z,\xi}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n g_i f(z_i)\right],$$

where g_i are i.i.d. standard Gaussian variables.

Lemma 11 (Upper Bounds with Rademacher and Gaussian Complexities) *For any function class $\mathcal{F}$ that consists of mappings from $\mathcal{Z}$ to $\mathbb{R}$, we have*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(\mathbb{P} - \mathbb{P}_n)f(z)\right] \le 2R_n(\mathcal{F}) \le \sqrt{2\pi}G_n(\mathcal{F}).$$

Lemma 11 can be found in, e.g., Lemma 7.4 in Van Handel [2014].

Lemma 12 (Lower Bounds with Rademacher and Gaussian Complexities) *For any function class $\mathcal{F}$ that consists of mappings from $\mathcal{Z}$ to $\mathbb{R}$, defined its centered class $\tilde{\mathcal{F}}$ as $\{f - \mathbb{E}[f(z)] : f \in \mathcal{F}\}$. We have*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}(\mathbb{P} - \mathbb{P}_n)f(z)\right] \ge \frac{1}{2}R_n(\tilde{\mathcal{F}}) \ge \frac{c}{\sqrt{\log n}}G_n(\tilde{\mathcal{F}}),$$

where $c > 0$ is an absolute constant.

Proof of Lemma 12: Both the fact that uniform convergence admit a lower bound in terms of the Rademacher complexity of the centered class, and the result that Rademacher complexity itself is bounded below by Gaussian complexity up to a factor of $\sqrt{\log n}$, are classical and admit simple proofs. For a full proof of the first inequality, see Theorem 14.3 in Rinaldo and Yan [2016]; for a reference and proof sketch of the second inequality, see Problem 7.1 in Van Handel [2014].

$\square$

C Proofs for Deep Neural Networks and Riemannian Dimension (Section 3)

C.1 Proof of Lemma 1 in Section 3.1

We start with the telescoping decomposition presented in the main paper, which serves as a non-perturvative replacement of conventional Taylor expansion, where in each summand the only difference lie in W_l' and W_l.

$$
\begin{aligned}
&F_L(W', X) - F_L(W, X)\\
=&\sum_{l=1}^{L}[\underbrace{\sigma_L(W_L' \cdots W_{l+1}'}_{\text{controled by} M_{l\to L}} \underbrace{\sigma_l}_{\text{by1}}(W_l' \underbrace{F_{l-1}(W,X)}_{\text{learned feature}})) - \sigma_L(W_L' \cdots W_{l+1}' \sigma_l(W_l \underbrace{F_{l-1}(W,X)}_{\text{learned feature}}))],
\end{aligned}
$$

Applying Cauchy-Schwartz inequality to the above identity, we have

$$
\begin{aligned}
&||F(W', X) - F(W, X)||_{\mathsf{F}}^2 &(33)\\
\leq &\sum_{l=1}^{L} L\|\sigma_L(W_L' \cdots W_{l+1}' \sigma_l(W_l' F_{l-1}(W,X))) - \sigma_L(W_L' \cdots W_{l+1}' \sigma_l(W_l F_{l-1}(W,X)))\|_{\mathsf{F}}^2 &(34)
\end{aligned}
$$

By the definition of local Lipchitz constant in Section 3, for all $W' \in B_{\varrho_n}(W, \varepsilon)$,

$$
\begin{aligned}
&\|\sigma_L(W_L' \cdots W_{l+1}' \sigma_l(W_l' F_{l-1}(W,X))) - \sigma_L(W_L' \cdots W_{l+1}' \sigma_l(W_l F_{l-1}(W,X)))\|_{\mathsf{F}}\\
\leq &M_{l\to L}[W,\varepsilon]\|\sigma_l(W_l' F_{l-1}(W,X)) - \sigma_l(W_l F_{l-1}(W,X))\|_{\mathsf{F}}. \qquad (35)
\end{aligned}
$$

Because the activation function σ_l is $1-$Lipchitz for each column, we have

$$
\|\sigma_l(W_l' F_{l-1}(W,X)) - \sigma_l(W_l F_{l-1}(W,X))\|_{\mathsf{F}} \leq \|(W_l' - W_l)F_{t-1}(W,X)\|_{\mathsf{F}}. \qquad (36)
$$

Combining (33) (35) and (36), we prove that

$$
\|F(W', X) - F(W, X)\|_{\mathsf{F}}^2 \leq \sum_{l=1}^{L} L \cdot M_{l\to L}[W,\varepsilon]^2 \cdot \|(W_l' - W_l)F_{l-1}(W,X)\|_{\mathsf{F}}^2.
$$

□

C.2 Metric Domination Lemma

Our non-perturbative expansion facilitates bounding the pointwise dimension of complex geometries via metric comparison. By constructing a simpler, dominating metric (i.e., one that is pointwise larger), we establish that the pointwise dimension of the original geometry is upper bounded by that of this new, more structured geometry. This "enlargement" for analytical tractability, a concept with roots in comparison geometry and majorization principles, is operationalized in Lemma 13.

Lemma 13 (Metric Domination Lemma) *For two metrics ϱ_1, ϱ_2 defined on $\mathbb{R}^p$, if $\varrho_1(W', W) \leq \varrho_2(W', W)$ for all $W' \in B_{\varrho_2}(W, \varepsilon)$, then for any prior $\pi \in \Delta(\mathbb{R}^p)$ and any $\varepsilon > 0$, we have*

$$
\log \frac{1}{\pi(B_{\varrho_1}(W,\varepsilon))} \leq \log \frac{1}{\pi(B_{\varrho_2}(W,\varepsilon))}.
$$

Proof of Lemma 13: Because $\varrho_1(W', W) \le \varrho_2(W', W)$ for all $W' \in B_{\varrho_2}(W, \varepsilon)$, we have that

$$B_{\varrho_1}(W, \varepsilon) \supseteq B_{\varrho_2}(W, \varepsilon).$$

So for any prior π on $\mathbb{R}^p$, monotonicity of measures gives

$$\pi(B_{\varrho_1}(W, \varepsilon)) \ge \pi(B_{\varrho_2}(W, \varepsilon)),$$

this implies

$$\log \frac{1}{\pi(B_{\varrho_1}(W, \varepsilon))} \le \log \frac{1}{\pi(B_{\varrho_2}(W, \varepsilon))}.$$

$\square$

We then state an extension of the metric domination lemma, which turns pointwise dimension in a high-dimensional space into a lower-dimensional subspace.

Lemma 14 (Subspace Metric Domination Lemma) *Given a metric ϱ_1 defined on $\mathbb{R}^p$ a subspace $\mathcal{V} \subseteq \mathbb{R}^p$, and a metric ϱ_2 defined on $\mathcal{V}$. Define the orthogonal projector to subspace $\mathcal{V}$ as $\mathcal{P}_{\mathcal{V}}(W) := \arg\min_{\tilde{W} \in \mathcal{V}} \|\tilde{W} - W\|_2$. If there exists $\varepsilon_1 \in (0, \varepsilon)$ such that for every $W' \in \mathcal{V}$,*

$$(\varrho_1(W', W))^2 \le (\varrho_2(W', \mathcal{P}_{\mathcal{V}}(W)))^2 + \varepsilon_1^2, \tag{37}$$

then for any prior $\pi \in \Delta(\mathcal{V})$, we have

$$\log \frac{1}{\pi(B_{\varrho_1}(W, \varepsilon))} \le \log \frac{1}{\pi(B_{\varrho_2}(\mathcal{P}_{\mathcal{V}}(W), \sqrt{\varepsilon^2 - \varepsilon_1^2}))}. \tag{38}$$

Proof of Lemma 14: By the condition (37), we know

$$B_{\varrho_1}(W, \varepsilon) \supseteq B_{\varrho_1}(W, \varepsilon) \cap \mathcal{V} \supseteq B_{\varrho_2}(\mathcal{P}_{\mathcal{V}}(W), \sqrt{\varepsilon^2 - \varepsilon_1^2}),$$

and this gives the desired conclusion (38) in Lemma 14.

$\square$

C.3 Auxiliary Lemmas for Volumetric Arguments

The following result on volumes of $d-$dimensional ℓ_2 ball with radius r is classical (e.g., see [Wikipedia contributors, 2025d]).

Lemma 15 (Volumes of ℓ_2 Ball, [Wikipedia contributors, 2025d]) *Let d be a positive integer and $r > 0$. Define*

$$B_2(r) = \{\, x \in \mathbb{R}^d : \|x\|_2 \le r \}.$$

Then its volume is

$$\mathrm{Vol}\big(B_2(r)\big) \;=\; \frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2} + 1\big)}\, r^d.$$

We now state the result on the volume of an ellipsoid, which is also classical. We give a proof here.

Lemma 16 (Volume of a Σ–Ellipsoid) *Let $\Sigma \in \mathbb{R}^{d\times d}$ be symmetric positive definite with eigenvalues $\lambda_1, \ldots, \lambda_d > 0$. For any $r > 0$, define the ellipsoid*

$$E := \{ w \in \mathbb{R}^d : w^\top \Sigma w \le r^2 \}.$$

Then its d-dimensional volume is

$$\mathrm{Vol}(E) \;=\; \frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2}+1\big)}\, r^d \left(\prod_{i=1}^{d} \lambda_i\right)^{-\frac{1}{2}}.$$

Proof of Lemma 16: Denote $\Sigma^{1/2}$ to be the square root of Σ: the unique positive semidefinite matrix satisfying $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$ (see, e.g., [Wikipedia contributors, 2025c]). Define the linear map

$$T : \mathbb{R}^d \to \mathbb{R}^d, \qquad z \;=\; T(w) := \Sigma^{1/2} w.$$

For $w \in E$ we have $w^\top \Sigma w \le r^2$, hence $\|z\|_2^2 = w^\top \Sigma w \le r^2$. Thus

$$T\big(E\big) = B_2(r) \;:=\; \{ z \in \mathbb{R}^d : \|z\|_2 \le r \},$$

the Euclidean ball of radius r. By the change–of–variables theorem in multivariate calculus [Wikipedia contributors, 2025a], the linear map T scales Lebesgue measure by $|\det T|$. Thus

$$\mathrm{Vol}\big(E\big) \;=\; |\det T|^{-1}\, \mathrm{Vol}\big(T(E)\big) \;=\; (\det \Sigma)^{-1/2}\, \mathrm{Vol}\big(B_2(r)\big),$$

where the second identity holds since $T = \Sigma^{1/2}$,

$$|\det T| \;=\; \det(\Sigma^{1/2}) \;=\; (\det \Sigma)^{1/2}.$$

The standard formula, Lemma 15, gives

$$\mathrm{Vol}\big(B_2(r)\big) = \frac{\pi^{d/2}}{\Gamma\big(\frac{d}{2}+1\big)}\, r^d.$$

This yields the claimed expression

$$\mathrm{Vol}\big(E\big) \;=\; (\det \Sigma)^{-1/2} \frac{\pi^{d/2}}{\Gamma\big(\frac{d}{2}+1\big)}\, r^d \;=\; \frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2}+1\big)}\, r^d \left(\prod_{i=1}^{d} \lambda_i\right)^{-\frac{1}{2}}.$$

□

C.4 Pointwise Dimension Bound with Approximate Effective Subspace

Set Up of Approximate Effective Subspace Given any fixed $p \times p$ PSD matrix $G(W)$, order the eigenvalues $\lambda_1(G(W)), \cdots, \lambda_p(G(W))$ nonincreasingly. For notational convenience, we suppress the dependence on $G(W)$ and write simply λ_k when no confusion can arise. We denote $\mathcal{V}_{\mathrm{eff}}(G(W)), R, \varepsilon)$ to be the *effective subspace*—the true top-r_{eff} eigenspace—of $G(W)$. For notiaional convenience, we use r_{eff} as the abbreviation of $r_{\mathrm{eff}}(G(W), R, \varepsilon)$, and $\mathcal{V}$ as an abbreviation of $\mathcal{V}_{\mathrm{eff}}(G(W)), R, \varepsilon)$ when no confusion can arise.

Assume there is another $r-$dimensional subspace $\bar{\mathcal{V}}$. We will show that if $\bar{\mathcal{V}}$ approximates $\mathcal{V}$, then using a prior supported on $\bar{\mathcal{V}}$ still yields a valid effective-dimension bound. This observation underpins the hierarchical covering argument in Theorem 3. For a self-contained introduction to subspaces (collectively known as the Grassmannian) and their frame parameterizations (the Stiefel manifold), see Section D.1, where we translate algebraic and differential-geometric insights into machine learning terminology.

Motivation of Approximate Effective Subspace. We can view the orthogonal projector to a subspace as a matrix (see the definition via the Stiefel parameterization in (60)), which is consistent with the earlier operator notation characterized by ℓ_2–distance in Lemma 14. Now define the projected metric $\varrho_{G(W)}^{\bar{\mathcal{V}}}$ as

$$\varrho_{G(W)}^{\bar{\mathcal{V}}}(W_1, W_2) = \sqrt{(\mathcal{P}_{\bar{\mathcal{V}}}(W_1) - \mathcal{P}_{\bar{\mathcal{V}}}(W_1))^\top G(W)(\mathcal{P}_{\bar{\mathcal{V}}}(W_2) - \mathcal{P}_{\bar{\mathcal{V}}}(W_2))} = \sqrt{(W_1 - W_2)^\top \mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}}(W_1 - W_2)}.$$

By the subspace metric dominance lemma (Lemma 14), if $\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}}$ approximates $G(W)$, we can use prior over $\bar{\mathcal{V}}$ to bound the pointwise dimension and achieve dimension reduction.

We will require the following approximation error condition:

$$\varrho_{\mathrm{proj},G(W)}(\mathcal{V}, \bar{\mathcal{V}}) = \|G(W)^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\mathrm{op}} \leq \frac{\sqrt{n}\varepsilon}{4R}.$$

In Section D, we systematically study the ellipsoidal covering of Grassmannian, and establish that we can *always* find $\bar{\mathcal{V}}$ that approximates $\mathcal{V}$ to the desired precision, with an additional covering cost of the Grassmannain bound in the Riemannain Dimension. This generalizes the canonical projection metric between subspaces into ellipsoidal set-up.

Effective Dimension Bound for Approximate Effective Subspace. We now present the lemma that establish effective dimension bound using prior supported on approximate effective subspace $\bar{\mathcal{V}}$ (not necessarily the true effective subspace $\mathcal{V}_{\mathrm{eff}}(G(W), R, \varepsilon)$). We state the main result of this subsection (Lemma 2) in the main paper.

Consider the weight space $B_2(R) \subset \mathbb{R}^p$ for vectorized weights, and a pointwise ellipsoidal metric defined via PSD $G(W)$. Let $\bar{\mathcal{V}} \subseteq \mathbb{R}^p$ be a fixed r-dimensional subspace. Define the prior $\pi_{\bar{\mathcal{V}}} = \mathrm{Unif}\big(B_2(1.58R) \cap \bar{\mathcal{V}}\big)$. Then, uniformly over all (W, ε) such that the top-r eigenspace $\mathcal{V}$ of $G(W)$ can be approximated by $\bar{\mathcal{V}}$ to precision

$$\varrho_{\mathrm{proj},G(W)}(\mathcal{V}, \bar{\mathcal{V}}) := \big\| G(W)^{1/2}\big(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}\big) \big\|_{\mathrm{op}} \leq \tfrac{\sqrt{n}\varepsilon}{4R}, \tag{39}$$

we have

$$\log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \leq \frac{1}{2} \sum_{k=1}^{r_{\mathrm{eff}}(G(W),R,\varepsilon)} \log\left(\frac{40R^2\lambda_k(G(W))}{n\varepsilon^2}\right) = d_{\mathrm{eff}}(G(W), \sqrt{5}R, \varepsilon).$$

Proof of Lemma 2: Given a fixed PSD matrix $G(W)$ with eigenvalues $\lambda_1 \geq \cdots \lambda_p$, denote $r_{\mathrm{eff}} = r_{\mathrm{eff}}(G(W), R, \varepsilon)$, and the projected metric $\varrho_{G(W)}^{\bar{\mathcal{V}}}$ on $\bar{\mathcal{V}}$:

$$\varrho_{G(W)}^{\bar{\mathcal{V}}}(W_1, W_2) = \sqrt{(W_1 - W_2)^\top \mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}}(W_1 - W_2)}.$$

Since $\mathcal{V}$ is the top-r_{eff} eigenspace of $G(W)$, by the elementary property of eigendecomposition we have that

$$\begin{aligned} G(W) =& \mathcal{P}_{\mathcal{V}}^\top G(W)\mathcal{P}_{\mathcal{V}} + \mathcal{P}_{\mathcal{V}_\perp}^\top G(W)\mathcal{P}_{\mathcal{V}_\perp} \\ \preceq& \mathcal{P}_{\mathcal{V}}^\top G(W)\mathcal{P}_{\mathcal{V}} + \lambda_{r_{\mathrm{eff}}+1} \cdot \mathcal{P}_{\mathcal{V}_\perp}^\top \mathcal{P}_{\mathcal{V}_\perp}, \end{aligned} \tag{40}$$

where $\mathcal{V}_\perp$ is orthogonal complement of $\mathcal{V}$. It is also straightforward to see

$$\mathcal{P}_{\mathcal{V}}^\top G(W)\mathcal{P}_{\mathcal{V}} \preceq 2\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}} + 2(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})^\top G(W)(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}). \tag{41}$$

Combining (40) and (41), we have the fundamental loewner order inequality

$$G(W) \preceq 2\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}} + 2(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})^\top G(W)(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}) + \lambda_{r_{\text{eff}}+1} \cdot \mathcal{P}_{\mathcal{V}_\perp}^\top \mathcal{P}_{\mathcal{V}_\perp}. \tag{42}$$

In order to apply the subspace metric domination lemma (Lemma 14), we hope to bound $\|W' - W\|_2^2$ and apply that bound to the two last reminder terms in the right hand side of (42).

To bound $\|W' - W\|_2^2$, we firstly state the following lemma on the eigenvalue of $\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}}$, whose proof is deferred until after the current proof.

Lemma 17 (Eigenvalue Bound for Projected Metric Tensor) *Assume $\mathcal{V}$ is the top-r eigenspace of a PSD matrix Σ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$, then for a $r-$dimensional subspace $\bar{\mathcal{V}}$ we have that for $k = 1, 2, \cdots, r$,*

$$\lambda_k \geq \lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) \geq \lambda_k/2 - \|\Sigma^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}}^2.$$

For every $W' \in B_{\varrho_{G(W)}^{\mathcal{V}}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$, we have $\forall k = 1, \cdots, r_{\text{eff}}$,

$$\begin{aligned}\|W' - \mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2^2 \leq& \frac{(W' - \mathcal{P}_{\bar{\mathcal{V}}}(W))^\top \mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}}(W' - \mathcal{P}_{\bar{\mathcal{V}}}(W))}{\lambda_{r_{\text{eff}}}(\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})} \leq \frac{n\varepsilon^2}{16\lambda_{r_{\text{eff}}}(\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})} \\ \leq& \frac{n\varepsilon^2}{8\lambda_{r_{\text{eff}}} - 16\|G(W)^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}}^2} \leq \frac{1}{3}R^2,\end{aligned} \tag{43}$$

where the first inequality holds because if A is a symmetric positive definite matrix, then for all vectors x, we have $x^\top Ax \geq \lambda_{\min}(A)\|x\|_2^2$; the third inequality uses Lemma 17; and the last inequality uses $\lambda_{r_{\text{eff}}} \geq \frac{n\varepsilon^2}{2R^2}$ (by definition (8) of effective rank) and the approximation error condition (39). On the other hand, we have that $\|W\|_2^2 \leq R^2$, so that for every $W' \in B_{\varrho_{G(W)}^{\bar{\mathcal{V}}}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$

$$\|W' - W\|_2^2 = \|W' - \mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2^2 + \|\mathcal{P}_{\bar{\mathcal{V}}_\perp}(W)\|_2^2 \leq \frac{4}{3}R^2.$$

From the fundamental loewner order inequality (42), we establish the desired metric domination condition: for all $W' \in B_{\varrho_{G(W)}^{\bar{\mathcal{V}}}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$ and $W \in B_2(R)$,

$$\begin{aligned}&(W' - W)^\top G(W)(W' - W) \\ \leq&(W' - W)^\top(2\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})(W' - W) + (2\|G(W)^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}}^2 + \lambda_{r_{\text{eff}}+1})\|W' - W\|_2^2 \\ \leq&2\varrho_{G(W)}^{\bar{\mathcal{V}}}(W', \mathcal{P}_{\bar{\mathcal{V}}}(W))^2 + \frac{5n\varepsilon^2}{6},\end{aligned}$$

where the first inequality holds because of the loewner order inequality (42) and the property of operator norm: $x^\top Ax \leq \|A\|_{\text{op}} \cdot \|x\|_2^2$ (one could also apply Lemma 17 to validate $\|\mathcal{P}_{\mathcal{V}_\perp}^\top \mathcal{P}_{\mathcal{V}_\perp}\|_{\text{op}} \leq 1$); and the last inequality uses the fact $\lambda_{r_{\text{eff}}+1} < \frac{n\varepsilon^2}{2R^2}$ (by definition 8 of effective rank) and the approximation error condition (39). Now we can apply the subspace metric domination lemma (Lemma 14) and obtain: for any $\pi \in \Delta(\bar{\mathcal{V}})$,

$$\log \frac{1}{\pi\big(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon)\big)} \leq \log \frac{1}{\pi\big(B_{\sqrt{2}\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/\sqrt{6})\big)} \leq \log \frac{1}{\pi\big(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)\big)}. \tag{44}$$

In particular, we choose π to be the uniform prior over $\bar{\mathcal{V}}$:

$$\pi_{\bar{\mathcal{V}}} = \mathrm{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}}).$$

Then we aim to prove that $B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4) \subseteq \bar{\mathcal{V}} \cap B_2(1.58R)$. This is true because: 1) for every $W' \in B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$, (43) suggests $\|W' - \mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2^2 \leq \frac{1}{3}R^2$, and 2) for very $W \in B_2(R)$, we have $\|\mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2 \leq \|W\|_2^2 \leq R^2$. Combining this and the above inequality we have

$$\|W'\|_2 \leq \|W' - \mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2 + \|\mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2 \leq (\sqrt{1/3} + 1)R < 1.58R.$$

This proves that $B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4) \subseteq \bar{\mathcal{V}} \cap B_2(1.58R)$, so we have

$$\log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)} = \frac{\mathrm{Vol}(\bar{\mathcal{V}} \cap B_2(1.58R))}{\mathrm{Vol}(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(W, \sqrt{n}\varepsilon/4))}. \tag{45}$$

Applying Lemma 16 about the volume of ellipsoid, with dimension r_{eff}, eigenvalues $\{\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})\}_{k=1}^{r_{\mathrm{eff}}}$ and radius $r = \sqrt{n}\varepsilon/4$, we have:

$$\begin{aligned}
\mathrm{Vol}(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)) =& \frac{\pi^{r_{\mathrm{eff}}/2}}{\Gamma(r_{\mathrm{eff}}/2+1)} \Big(\sqrt{n} \cdot \varepsilon/4\Big)^{r_{\mathrm{eff}}} \Big(\prod_{k=1}^{r_{\mathrm{eff}}} \lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})\Big)^{-1/2} \\
\geq& \frac{\pi^{r_{\mathrm{eff}}/2}}{\Gamma(r_{\mathrm{eff}}/2+1)} \Big(\sqrt{n} \cdot \varepsilon/4\Big)^{r_{\mathrm{eff}}} \Big(\prod_{k=1}^{r_{\mathrm{eff}}} \lambda_k\Big)^{-1/2},
\end{aligned}$$

where the inequality uses Lemma 17 to upper bound $\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})$ by λ_k. The standard formula for the volume of unit ball, Lemma 16, suggests that the volume of $\mathcal{V} \cap B_2(2R)$ is

$$\mathrm{Vol}(\bar{\mathcal{V}} \cap B_2(1.58R)) \ = \ \frac{\pi^{r_{\mathrm{eff}}/2}}{\Gamma(r_{\mathrm{eff}}/2+1)}(1.58R)^{r_{\mathrm{eff}}}.$$

Hence, applying (44) (45) and combining it with the two above volume bounds, we have

$$\begin{aligned}
&\log \frac{1}{\pi\big(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon)\big)} \leq \log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4))} = \log \frac{\mathrm{Vol}(\bar{\mathcal{V}} \cap B_2(1.58R))}{\mathrm{Vol}(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4))} \\
&\leq \frac{1}{2} \log \frac{(1.58R)^{2r_{\mathrm{eff}}} \prod_{k=1}^{r_{\mathrm{eff}}} \lambda_k}{(\sqrt{n}\varepsilon/4)^{2r_{\mathrm{eff}}}} \leq \frac{1}{2} \sum_{k=1}^{r_{\mathrm{eff}}} \log \frac{40R^2\lambda_k}{n\varepsilon^2} \\
&= d_{\mathrm{eff}}(G(W), \sqrt{5}R, \varepsilon).
\end{aligned}$$

Finally, since the prior construction $\pi_{\bar{\mathcal{V}}} = \mathrm{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}})$ only depends on $\bar{\mathcal{V}}$ rather than W and ε, we have that uniformly over all $(W, \varepsilon) \in B_2(R) \times [0, \infty)$ such that $\bar{\mathcal{V}}$ approximates $\mathcal{V}_{\text{eff}}(G(W), R, \varepsilon)$ to the precision (39),

$$\log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \leq d_{\text{eff}}(G(W), \sqrt{5}R, \varepsilon).$$

□

Proof of Lemma 17: The Courant–Fischer–Weyl max-min characterization [Wikipedia contributors, 2025b] states that for any Hermitian (i.e. symmetric for real matrices studying here) matrix,

$$\lambda_k(\Sigma) \;=\; \max_{\substack{\mathcal{S} \subseteq \mathbb{R}^p \\ \dim \mathcal{S} = k}} \; \min_{\substack{W \in \mathcal{S} \\ W \neq 0}} \frac{W^\top \Sigma W}{\|W\|_2^2},$$

and we have that for any $r-$dimensional subspace $\bar{\mathcal{V}}$,

$$\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) \;=\; \max_{\substack{\mathcal{S} \subseteq \bar{\mathcal{V}} \\ \dim \mathcal{S} = k}} \; \min_{\substack{W \in \mathcal{S} \\ W \neq 0}} \frac{W^\top \mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}} W}{\|W\|_2^2},$$

so we have $\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) \leq \lambda_k$ for $k = 1, 2, \cdots, r$.

Moreover, by the elementary property of eigendecomposition we have $\lambda_k = \lambda_k(\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}})$, and by the Courant–Fischer–Weyl max-min characterization we know that,

$$\begin{aligned}
\lambda_k(\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}}) &= \max_{\substack{\mathcal{S} \subseteq \mathbb{R}^p \\ \dim \mathcal{S} = k}} \; \min_{\substack{W \in \mathcal{S} \\ W \neq 0}} \frac{W^\top (\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}}) W}{\|W\|_2^2} \\
&\leq \max_{\substack{\mathcal{S} \subseteq \mathbb{R}^p \\ \dim \mathcal{S} = k}} \; \min_{\substack{W \in \mathcal{S} \\ W \neq 0}} \frac{W^\top (2\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) W + \|\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}} - 2\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}\|_{\text{op}} \|W\|_2^2}{\|W\|_2^2} \\
&= 2\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) + \|\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}} - 2\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}\|_{\text{op}} \\
&\leq 2\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) + 2\|(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})^\top \Sigma (\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}},
\end{aligned}$$

where the first inequality is because for every fixed S and W we have $W^\top (\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}}) W \leq W^\top (2\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) W + \|\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}} - 2\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}\|_{\text{op}} \|W\|_2^2$; and the last inequality is due to (41). Therefore we have

$$\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) \geq \lambda_k / 2 - \|\Sigma^{\frac{1}{2}} (\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}}^2.$$

□

C.5 Proof of Riemannian Dimension Bound for DNN (Theorem 3)

In the language of Riemannian geometry [Jost, 2008], we regard a pointwise PSD, matrix-valued function $G(W)$ as a (possibly degenerate) *metric tensor*; such a G(W) endows the parameter space $\mathbb{R}^{\sum_{l=1}^L d_{l-1} d_l}$ with a (semi-)Riemannian manifold structure. The pointwise ellipsoidal metric in (7) belongs to the following family of block-decomposable metric tensors.

Definition 3 (Metric Tensor of NN-surrogate Type) *A metric tensor $G(W)$ (pointwise PSD-valued function of size $\sum_{l=1}^{L} d_{l-1}d_l \times \sum_{l=1}^{L} d_{l-1}d_l$) is of "NN-surrogate" type if $G(W)$ is in the form*

$$G(W) = \mathrm{blockdiag}(A_1(W) \otimes I_{d_1}, \cdots, A_l(W) \otimes I_{d_l}, \cdots, A_L(W) \otimes I_{d_L})$$

where $A_l(W) \in \mathbb{R}^{d_{l-1} \times d_{l-1}}$.

By Lemma 1, the non-perturbative feature expansion gives rise to the metric tensor $G_{\mathrm{NP}}(W)$ defined in (7); $G_{\mathrm{NP}}(W)$ belongs to the "NN-surrogate" class. We first record some elementary decomposition properties for this family of NN-surrogate metric tensors, and then prove Theorem 3.

C.5.1 Decomposition Properties of NN-surrogate Metric Tensor

The NN-surrogate metric tensor $G(W)$ in Definition 3 has decomposition properties described by the next lemma.

Lemma 18 (Decomposition Properties of NN-surrogate Metric Tensor) *Given a NN-surrogate metric tensor $G(W)$ defined in Definition 3, for every W, we have the following decomposition properties: First, the effective rank and dimension decompose to*

$$r_{\mathrm{eff}}(G(W), R, \varepsilon) = \sum_{l=1}^{L} d_l \cdot r_{\mathrm{eff}}(A_l(W), R, \varepsilon);$$
$$d_{\mathrm{eff}}(G(W), R, \varepsilon) = \sum_{l=1}^{L} d_l \cdot d_{\mathrm{eff}}(A_l(W), R, \varepsilon).$$

Second, denote $\mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon)$ the effective subspace (i.e., the top-$r_{\mathrm{eff}}(A_l(W), R, \varepsilon)$ eigenspace) of $A_l(W)$. Then the effective subspace of $G(W)$ is

$$\mathcal{V}_{\mathrm{eff}}(G(W), R, \varepsilon) = \mathcal{V}_{\mathrm{eff}}(A_1(W), R, \varepsilon)^{d_1} \times \cdots \times \mathcal{V}_{\mathrm{eff}}(A_L(W), R, \varepsilon)^{d_L}.$$

Proof of Lemma 18. It is straightforward to see that, first, the effective rank of the fixed matrix $G(W)$ is

$$\begin{aligned}
&r_{\mathrm{eff}}(G(W), R, \varepsilon) \\
=& \max\{k : 2\lambda_k(G(W))R^2 \geq n\varepsilon^2\} \\
=& \sum_{l=1}^{L} \max\{k : 2\lambda_k(A_l(W) \otimes I_{d_l})R^2 \geq n\varepsilon^2\} \\
=& \sum_{l=1}^{L} d_l \max\{k : 2\lambda_k(A_l(W))R^2 \geq n\varepsilon^2\} \\
=& \sum_{l=1}^{L} d_l \cdot r_{\mathrm{eff}}(A_l(W), R, \varepsilon);
\end{aligned}$$

and the effective dimension of the fixed matrix $G(W)$ is

$$\begin{aligned}
&d_{\text{eff}}(G(W), R, \varepsilon)\\
=&\frac{1}{2}\sum_{k=1}^{r_{\text{eff}}(G(W),R,\varepsilon)} \log\left(\frac{8R^2\lambda_k(G(W))}{n\varepsilon^2}\right)\\
=&\sum_{l=1}^{L}\frac{1}{2}\sum_{k=1}^{r_{\text{eff}}(A_l(W)\otimes I_{d_l},R,\varepsilon)} \log\left(\frac{8R^2\lambda_k(A_l(W)\otimes I_{d_l})}{n\varepsilon^2}\right)\\
=&\sum_{l=1}^{L} d_l\cdot\frac{1}{2}\sum_{k=1}^{r_{\text{eff}}(A_l(W),R,\varepsilon)} \log\left(\frac{8R^2\lambda_k(A_l(W))}{n\varepsilon^2}\right)\\
=&\sum_{l=1}^{L} d_l\cdot d_{\text{eff}}(A_l(W), R, \varepsilon).
\end{aligned}$$

Second, as the effective subspace of the matrix tensor product $A_l(W)\otimes I_{d_l}$ is subspace tensor product $\mathcal{V}_{\text{eff}}(A_l(W), R, \varepsilon)^{d_l}$, the effective subspace for NN-surrogate metric tensor $G(W) = \text{blockdiag}(\cdots; A_l(W)\otimes I_{d_l}; \cdots)$ is

$$\mathcal{V}_{\text{eff}}(G(W), R, \varepsilon) := \mathcal{V}_{\text{eff}}(A_1(W), R, \varepsilon)^{d_1}\times\cdots\times\mathcal{V}_{\text{eff}}(A_L(W), R, \varepsilon)^{d_L}.$$

□

C.5.2 Proof of Theorem 3

We firstly prove the following result, which is almost Theorem 3, with the only difference being that the radius in the effective dimension depends on the global radius R rather than the pointwise Frobenious norm $\|W\|$. Extending this result to Theorem 3 can be achieved via a simple application of the "uniform pointwise convergence" principle [Xu and Zeevi, 2025] illustrated in Lemma 8.

Lemma 19 (Riemannian Dimension for NN-surrogate Metric Tensor—Global Radius Version) *Consider the NN-surrogate metric tensor in Definition 3, and the weight space $B_{\mathsf{F}}(R)$. Then we have that the pointwise dimension is bounded by the pointwise Riemannian Dimension as the following: there exists a prior π such that uniformly over all $W\in B_{\mathsf{F}}(R)$,*

$$\log\frac{1}{\pi(B_{\varrho_{G(W)}}(W,\sqrt{n}\varepsilon))}\le\sum_{l=1}^{L}\Bigg(\underbrace{d_l\cdot d_{\text{eff}}(A_l(W), CR, \varepsilon)}_{\text{“must pay” cost at each } W}+\underbrace{d_{l-1}\cdot d_{\text{eff}}(A_l(W), CR, \varepsilon)}_{\text{covering cost of Grassmannian}}+\underbrace{\log(d_{l-1})}_{\text{covering cost of } r_{\text{eff}}\in[d_{l-1}]}\Bigg),$$

where $C>0$ is an absolute constant.

Proof of Lemma 19: The proof has two key steps: 1. Hierarchical covering argument, and 2. Bound covering Cost of the Grassmannian. A crucial lemma about the ellipsoidal covering of the Grassmannian, which is new even in the pure mathematics context, is deferred to Section D.

Step 1: Hierarchical Covering. As explained the main paper, the major difficulty is that the prior measure $\pi_{\mathcal{V}}$ it constructed, is defined over the effective subspace $\mathcal{V}$, which itself encodes information of the point W and $\varepsilon > 0$. The goal of our proof is to construct a "universal" prior π that does not depend on $\mathcal{V}$. This is achieved via a hierarchical covering argument (11), which we make rigorous below.

The key idea of hierarchical covering is as follows: Firstly, for all W, we search for subspace $\bar{\mathcal{V}}$ that approximates the true effective subspace (top-r_{eff} eigenspace) $\mathcal{V}_{\text{eff}}(G(W), R, \varepsilon)$ to the precision required by (39):

$$\|G(W)^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}} \le \frac{\sqrt{n}\varepsilon}{4R}, \tag{46}$$

where $G(W)^{\frac{1}{2}}$ is the unique square root of PSD matrix $G(W)$ (see, e.g, [Wikipedia contributors, 2025c]). Then by Lemma 2 (Pointwise Dimension Bound for Non-Linear Manifold with Approximate Effective Subspace), for every $(W, \varepsilon) \in B_2(R) \times [0, \infty)$ such that $\bar{\mathcal{V}}$ approximates $\mathcal{V}_{\text{eff}}(G(W), R, \varepsilon)$ to the precision (46), the prior $\pi_{\bar{\mathcal{V}}} = \text{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}})$ satisfies

$$\log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \le d_{\text{eff}}(G(W), \sqrt{5}R, \varepsilon) = \sum_{l=1}^{L} d_l \cdot d_{\text{eff}}(A_l(W), \sqrt{5}R, \varepsilon), \tag{47}$$

where the first inequality is by Lemma 2 (see definition (9) of effective dimension); and the last equality is by the decomposition property of NN-surrogate metric tensor (Lemma 18).

Secondly, we put a prior μ over all possible subspaces $\mathcal{V}$ and construct the "universal" prior

$$\pi(W) = \sum_{\mathcal{V}} \mu(\mathcal{V}) \times \pi_{\mathcal{V}}(W), \tag{48}$$

which implies that uniformly over all $W \in B_{\mathsf{F}}(R)$,

$$\begin{aligned}
&\log \frac{1}{\pi(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
=&\log \frac{1}{\sum_{\mathcal{V}} \mu(\mathcal{V}) \pi_{\mathcal{V}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
\le&\log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies } (46)) \inf_{\bar{\mathcal{V}} \text{ satisfies } (46)} \pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
\le& \underbrace{\log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies } (46))}}_{\text{covering cost of the Grassmannian}} + \sum_{l=1}^{L} d_l \cdot d_{\text{eff}}(A_l(W), \sqrt{5}R, \varepsilon),
\end{aligned} \tag{49}$$

where the first equality is by definition (48) of the "universal" prior π; the first inequality is straightforward; and the last inequality is by (47), the result of the "must pay" part in the hierarchical covering.

The above hierarchical covering argument successfully gives a valid Riemannian Dimension, with the cost of the additional covering cost given by the subspace prior μ. This explains our basic proof idea. The remaining proof executes this basic proof idea.

Step 2: Bounding Covering Cost of the Grassmannian. Section D provides a systematical study to the ellipsoidal metric entropy of Grassmannian manifold, which we detail the conclusion below.

Define

$$\mathrm{Gr}(d,r) \ := \ \left\{ r\text{–dimensional linear subspaces of } \mathbb{R}^d \right\}$$

as the *Grassmann manifold*.

Given a $d\times d$ PSD Σ, define the anisometric projection metric between two subspaces by (labeled as Definition 4 in Section D)

$$\varrho_{\mathrm{proj},\Sigma}(\mathcal{V},\bar{\mathcal{V}}) = \|\Sigma^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\mathrm{op}}, \tag{50}$$

where $\Sigma^{\frac{1}{2}}$ is the square root of the PSD matrix Σ (see, e.g., [Wikipedia contributors, 2025c]).

Lemma 3 states that (note that we use ε_1 and C_0 here instead of ε and C in the original statement of Lemma 3), given a Grassmannian $\mathrm{Gr}(d,r)$, for uniform prior $\mu = \mathrm{Unif}(\mathrm{Gr}(d,r))$, we have that for every $\mathcal{V}\in\mathrm{Gr}(d,r)$, every $\varepsilon_1 > 0$ and PSD matrix $\Sigma\in\mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1 \geq \cdots \lambda_d \geq 0$, we have the pointwise dimension bound

$$\log\frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon_1))} \leq \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{C_0\max\{\lambda_k,\varepsilon_1^2\}}{\varepsilon_1^2} + \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{C_0\max\{\lambda_k,\varepsilon_1^2\}}{\varepsilon_1^2}, \tag{51}$$

where $C_0 > 0$ is an absolute constant. We will use the result (51) and (49) to prove Theorem 3.

For a particular layer l, $d_{l-1}\times d_{l-1}$ PSD matrix $A_l(W)$, and a fixed rank r_l denote $\mathrm{Gr}(d_{l-1},r_l)$ as a Grassmannian (the collection of all r_l-dimensional in $\mathbb{R}^{d_{l-1}}$). By (51) we have that there exists a prior μ_l over $\mathrm{Gr}(d_{l-1},r_l)$ such that for every (W,ε_1) such that $r_{\mathrm{eff}}(A_l(W),R,\varepsilon_1) = r_l$, and $\lambda_{r_l+1}(A_l(W)) \leq c\varepsilon_1^2 \leq \lambda_{r_l}(A_l(W))$ where $c\geq 1$ can be any absolute constants no smaller than 1 (later we will specialize to $c=8$),

$$\log\frac{1}{\mu_l(\bar{\mathcal{V}}:\varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W),R,\varepsilon),\bar{\mathcal{V}})\leq\varepsilon_1)} \leq \frac{d_{l-1}}{2}\sum_{k=1}^{r_l}\log\frac{C_1\lambda_k(A_l(W))}{\varepsilon_1^2}, \tag{52}$$

where $C_1 = c\max\{C_0,1\}\geq 1$ is an absolute constant depending only on the absolute constant c (later we take $c=8$ so $C_1 = 8\max\{C_0,1\}$ is indeed an absolute constant). This is because: 1) all eigenvalues with index at least r_l+1 (each no larger than $c\,\varepsilon_1^2$) contribute only through the second term in (51). Their cumulative effect is at most

$$\mathbb{1}\{d_{l-1}-r_l > r_l\}\cdot\frac{r_l}{2}\sum_{k=r_l+1}^{d_{l-1}-r_l}\log\frac{c\varepsilon_1^2}{\varepsilon_1^2} = \frac{r_l\max\{d_{l-1}-2r_l,0\}}{2}\log c \leq \frac{r_l(d_{l-1}-r_l)}{2}\log c$$

unaffected to the spectrum, and we absorb this into the absolute constant C_1. And 2) all eigenvalues with index at most r_l's contribution leads to at most

$$\frac{d_{l-1}-r_l}{2}\sum_{k=1}^{r_l}\log\frac{C_0\lambda_k(A_l(W))}{\varepsilon_1^2} + \frac{r_l}{2}\sum_{k=1}^{\max\{r_l,d_{l-1}-r_l\}}\log\frac{C_0\lambda_k(A_l(W))}{\varepsilon_1^2} \leq \frac{d_{l-1}}{2}\sum_{k=1}^{r_l}\log\frac{\max\{C_0,1\}\lambda_k(A_l(W))}{\varepsilon_1^2}.$$

Summing up the contributions two parts of the spectrum together, we get the right hand side of (52).

By the subspace decomposition property in Lemma 18, we have that for $\bar{\mathcal{V}} = (\cdots, \underbrace{\bar{\mathcal{V}}_l, \cdots, \bar{\mathcal{V}}_l}_{\text{repeat } d_l \text{ times}}, \cdots)$,

$$
\begin{aligned}
&\varrho_{\mathrm{proj},G(W)}(\mathcal{V}_{\mathrm{eff}}(G(W), R, \varepsilon), \bar{\mathcal{V}}) \\
=&\varrho_{\mathrm{proj},G(W)}(\prod_{l=1}^{L} \mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon)^{d_l}, \prod_{l=1}^{L} \bar{\mathcal{V}}_l^{d_l}) \\
=&\max_l \varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon), \bar{\mathcal{V}}_l),
\end{aligned} \tag{53}
$$

where the first equality is by Lemma 18, and the second equality is by the properties of the spectral norm: $\|\mathrm{blockdiag}(A, B)\|_{\mathrm{op}} = \max\{\|A\|_{\mathrm{op}}, \|B\|_{\mathrm{op}}\}$ and $\|A \otimes I_d\|_{\mathrm{op}} = \|A\|_{\mathrm{op}}$.

Taking $\varepsilon_1 = \frac{\sqrt{n}\varepsilon}{4R}$, by definition (8) on the threshold to determine effective rank, we obtain $\lambda_{r_l+1}(A_l(W)) \le 8\varepsilon_1^2 = n\varepsilon^2/(2R^2) \le \lambda_{r_l}(A_l(w))$, thus this particular choice satisfies the required eigenvalue condition to establish (52) with $c = 8$. Then for all layers $l = 1, \cdots, L$, given a fixed $\{r_1, \cdots, r_L\}$, by (52), we have that there exists a prior

$$
\mu_{\{r_l\}_{l=1}^L} = \mu_1^{d_1} \times \cdots \times \mu_L^{d_L} = \prod_{l=1}^{L} (\underbrace{\mu_l \times \cdots \times \mu_l}_{d_l \text{ times}}) \tag{54}
$$

over the product Grassmannian $\mathrm{Gr}(d_0, r_1)^{d_1} \times \cdots \times \mathrm{Gr}(d_{L-1}, r_L)^{d_L}$ such that uniformly over all $W \in B_{\mathsf{F}}(R)$ such that $r_{\mathrm{eff}}(A_l(W), R, \varepsilon) = r_l$, $\forall l \in [L]$ (here $[L]$ is the notation of $\{1, 2, \cdots, L\}$), the "Grassmannian covering cost" term in (49) is bounded by

$$
\begin{aligned}
&\log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies } (46))} \\
=&\log \frac{1}{\mu_{\{r_l\}_{l=1}^L}(\bar{\mathcal{V}} : \varrho_{\mathrm{proj},G(W)}(\mathcal{V}_{\mathrm{eff}}(G(W), R, \varepsilon), \bar{\mathcal{V}}) \le \frac{\sqrt{n}\varepsilon}{4R} = \varepsilon_1)} \\
\le&\log \frac{1}{\mu_{\{r_l\}_{l=1}^L}((\cdots, \underbrace{\bar{\mathcal{V}}_l, \cdots, \bar{\mathcal{V}}_l}_{d_l \text{ times}}, \cdots) : \varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon), \bar{\mathcal{V}}_l) \le \varepsilon_1, \quad \forall l \in [L])} \\
=&\sum_{l=1}^{L} \log \frac{1}{\mu_{\{r_l\}_{l=1}^L}((\cdots, \underbrace{\bar{\mathcal{V}}_l, \cdots, \bar{\mathcal{V}}_l}_{d_l \text{ times}}, \cdots) : \varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon), \bar{\mathcal{V}}_l) \le \varepsilon_1)} \\
\le&\sum_{l=1}^{L} \frac{d_{l-1}}{2} \sum_{k=1}^{r_l} \log \frac{C_1 \lambda_k(A_l(W))}{\varepsilon_1^2} \\
\le&\sum_{l=1}^{L} d_{l-1} d_{\mathrm{eff}}(A_l(W), \sqrt{2C_1} R, \varepsilon),
\end{aligned} \tag{55}
$$

where the first inequality is by restricting $\bar{\mathcal{V}}$ to the form $\prod_{l=1}^{L} \bar{\mathcal{V}}_l^{d_l}$ and using (53); the second equality is by the choice of the product prior (54); the second inequality is by the layer-wise covering bound (52); and the last inequality is by the choice $\varepsilon_1 = \sqrt{n}\varepsilon/(4R)$, and definition (9) of effective dimension.

Note that (55) is uniformly over all $W \in B_{\mathsf{F}}(R)$ such that $r_{\text{eff}}(A_l(W), R, \varepsilon) = r_l$, $\forall l \in [L]$, not uniformly over all $W \in B_{\mathsf{F}}(R)$. We would like to extend (55) to all $W \in B_{\mathsf{F}}(R)$ over uniform prior over possible integer values of r_l. Now assign uniform prior over $[d_{l-1}] = \{1, \cdots, d_{l-1}\}$ for r_l, we obtain the "universal" prior π (as we have pursued in in our hierarchical covering argument (48)) defined by

$$
\begin{aligned}
\mu(\mathcal{V}) &= \prod_{l=1}^{L} \underbrace{\text{Unif}([d_{l-1}])}_{\text{prior of } r_l} \times \underbrace{\mu_{\{r_k\}_{k=1}^L}}_{\text{prior over product Grassmannian in (54)}}, \\
\pi(W) &= \underbrace{\mu(\mathcal{V})}_{\text{prior over subspaces defined above}} \times \underbrace{\text{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}})}_{\text{uniform prior constrained in subspace}}.
\end{aligned} \tag{56}
$$

Then we have that uniformly over all $W \in B_{\mathsf{F}}(R)$,

$$
\begin{aligned}
&\log \frac{1}{\pi(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
&\leq \log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies (46)})} + \sum_{l=1}^{L} d_l \cdot d_{\text{eff}}(A_l(W), \sqrt{5}R, \varepsilon)) \\
&\leq \sum_{l=1}^{L} \log d_{l-1} + \log \frac{1}{\mu_{\{r_k\}_{k=1}^L}(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies (46)})} + \sum_{l=1}^{L} d_l \cdot d_{\text{eff}}(A_l(W), \sqrt{5}R, \varepsilon)) \\
&\leq \sum_{l=1}^{L} \log d_{l-1} + \sum_{l=1}^{L} d_{l-1} \cdot d_{\text{eff}}(A_l(W), \sqrt{2C_1}R, \varepsilon) + \sum_{l=1}^{L} d_l \cdot d_{\text{eff}}(A_l(W), \sqrt{5}R, \varepsilon)),
\end{aligned}
$$

where $C_1 > 0$ is an absolute constant. Here the first inequality is by the hierarchical covering argument (49); the second inequality is by the prior construction (56); and the third inequality is by the Grassmannian covering bound (55) for fixed $\{r_k\}_{k=1}^L$. This shows that for NN-surrogate metric tensor $G(W)$, the pointwise dimension is bounded by the Riemannian Dimension as the following:

$$
\log \frac{1}{\pi(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \leq \sum_{l=1}^{L} (d_l + d_{l-1}) \cdot d_{\text{eff}}(A_l(W), CR, \varepsilon) + \log(d_{l-1}),
$$

where C is a positive absolute constant. This finishes the proof of Lemma 19 with R in effective dimension being a global upper bound of $\|W\|_{\mathsf{F}}$.

□

Proof of Theorem 3: Motivated by the "uniform pointwise convergence" principle (proposed in Xu and Zeevi [2025] and illustrated in Lemma 8), we apply a peeling argument to adapt the Riemannian Dimension to $\|W\|_{\mathsf{F}}$. Given any $R_0 \in (0, R]$, we take $R_k = 2^k R_0$ for $k = 0, 1, \cdots \log_2\lceil R/R_0 \rceil$. Taking a uniform prior on these R_k, and set

$$
\tilde{\pi} = \underbrace{\text{Unif}(\{R_0, \cdots, 2^{\log_2\lceil R/R_0 \rceil} R_0\})}_{\text{prior over upper bound } \tilde{R} \text{ of } \|W\|_{\mathsf{F}}} \times \underbrace{\pi_{\tilde{R}}}_{\text{prior defined via (56)}},
$$

where $\pi_{\tilde{R}}$ is the prior defined via (56) in the proof of Lemma 19. Then for every $W \in B_{\mathsf{F}}(R)$ where $\|W\|_{\mathsf{F}} > R_0$, denote $k(W)$ to be the integer such that $2^{k(W)}R_0 < \|W\|_{\mathsf{F}} \le 2^{k(W)+1}R_0$, then

$$
\begin{aligned}
&\log \frac{1}{\tilde{\pi}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
&\le \underbrace{\log\log_2\lceil R/R_0\rceil}_{\text{density of } 2^{k(W)+1}R_0} + \underbrace{\log \frac{1}{\pi_{2^{k(W)+1}R_0}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))}}_{\pi \text{ is constructed via (56), with global radius taken to be } 2^{k(W)+1}R_0} \\
&\le \log\log_2\lceil Rn\rceil + \sum_{l=1}^{L}((d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(A_l(W), C_1 2^{k_0+1}R_0, \varepsilon) + \log d_{l-1}) \\
&\le \log\log_2\lceil Rn\rceil + \sum_{l=1}^{L}((d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(A_l(W), C_1 \cdot 2\|W\|_{\mathsf{F}}, \varepsilon) + \log d_{l-1})
\end{aligned}
$$

where the first inequality is due to the product construction of $\tilde{\pi}$; the second inequality is due to Lemma 19, with $C_1 > 0$ being an absolute constant; and the last inequality uses the fact $\|W\|_{\mathsf{F}} \le 2^{k_0+1}R_0 \le 2\|W\|_{\mathsf{F}}$, with $C_1 > 0$.

The above bound assumes $\|W\|_{\mathsf{F}} > R_0$. When $\|W\|_{\mathsf{F}} \le R_0$, we directly apply Lemma 19 and obtain

$$
\begin{aligned}
&\log \frac{1}{\tilde{\pi}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
&\le \underbrace{\log\log_2\lceil R/R_0\rceil}_{\text{density of } R_0} + \underbrace{\log \frac{1}{\pi_{R_0}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))}}_{\pi \text{ is constructed via (56), with global radius taken to be } R_0} \\
&\le \log\log_2\lceil Rn\rceil + \sum_{l=1}^{L}(d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(A_l(W), C_1 \cdot R_0, \varepsilon) + \log d_{l-1}).
\end{aligned}
$$

Combining the two cases discussed above, we conclude that the pointwise dimension for NN-surrogate metric tensor $G(W)$ in Definition 3 is bounded by the Riemmanin Dimension

$$
\begin{aligned}
&\log \frac{1}{\tilde{\pi}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \le d_{\mathrm{R}}(W, \varepsilon) \\
&= \sum_{l=1}^{L}(d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(A_l(W), C \max\{\|W\|_{\mathsf{F}}, R_0\}) + \log(d_{l-1}\log_2\lceil R/R_0\rceil),
\end{aligned}
$$

where $C = 2C_1$ is a positive absolute constant.

Finally, by the sentence below (7) (which is a straightforward result from non-perturbative feature expansion for DNN (Lemma 1) and the metric domination lemma (Lemma 13)), we know that there exists a prior $\tilde{\pi}$ such that uniformly over all $W \in B_{\mathsf{F}}(R)$,

$$
\begin{aligned}
&\log \frac{1}{\tilde{\pi}(B_{\varrho_n}(W, \sqrt{n}\varepsilon))} \le \log \frac{1}{\tilde{\pi}(B_{\varrho_{G_{\mathrm{NP}}(W)}}(W, \sqrt{n}\varepsilon))} \\
&\le d_{\mathrm{R}}(W, \varepsilon) = \sum_{l=1}^{L}(d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(A_l(W), C \max\{\|W\|_{\mathsf{F}}, R_0\}) + \log(d_{l-1}\log_2\lceil R/R_0\rceil),
\end{aligned}
$$

where $A_l(W) = LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F^\top_{l-1}(W,X)$ when taking $G(W)$ to be $G_{\mathrm{NP}}(W)$ defined in (7). Taking $R_0 = R/2^n$ proves Theorem 3.

□

D Ellipsoidal Covering of the Grassmannian (Lemma 3)

The central goal of this section is to prove the following result on the ellipsoidal metric entropy of the Grassmannian manifold. The definition for Gr (Grassmannian manifold), St (Stiefel parameterization manifold) are temporarily deferred to Section D.1.

Definition 4 (Ellipsoidal Projection Metric) *For two subspaces $\mathcal{V},\bar{\mathcal{V}} \in \mathrm{Gr}(d,r)$, and a positive semidefinite matrix Σ, define the ellipsoidal projection metric $\varrho_{\mathrm{proj},\Sigma}$ by*

$$\varrho_{\mathrm{proj},\Sigma}(\mathcal{V},\bar{\mathcal{V}}) = \|\Sigma^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\mathrm{op}},$$

where $\mathcal{P}_{\mathcal{V}}$ and $\mathcal{P}_{\bar{\mathcal{V}}}$ are orthogonal projectors to subspace $\mathcal{V}$ and $\bar{\mathcal{V}}$, respectively.

We view orthogonal projectors as matrices (see the definition via the Stiefel parameterization in (60)), consistent with the earlier operator notation characterized by ℓ_2–distance in Lemma 14. In the isotropic case $\Sigma = I_d$, the ellipsoidal projection metric reduces to the standard isotropic projection metric

$$\varrho_{\mathrm{proj}}(\mathcal{V},\bar{\mathcal{V}}) = \left\|\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}\right\|_{\mathrm{op}}.$$

We now state our main result in this section (Lemma 3 in the main paper).

Consider the Grassmannian $\mathrm{Gr}(d,r)$ and the uniform prior $\mu = \mathrm{Unif}(\mathrm{Gr}(d,r))$, then for every $\mathcal{V} \in \mathrm{Gr}(d,r)$, every $\varepsilon > 0$ and every PSD matrix Σ with eigenvalues $\lambda_1 \geq \cdot\lambda_d \geq 0$, we have

$$\log\frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))} \leq \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{C\max\left\{\lambda_k,\varepsilon^2\right\}}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{C\max\left\{\lambda_k,\varepsilon^2\right\}}{\varepsilon^2},$$

where $C > 0$ is an absolute constant.

Recall that the traditional covering number bound for the Grassmannian manifold states that

$$\left(\frac{c}{\varepsilon}\right)^{r(d-r)} \leq \mathrm{N}(\mathrm{Gr}(d,r),\varrho_{\mathrm{proj}},\varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^{r(d-r)}. \tag{57}$$

Here $\mathrm{N}(\mathcal{F},\varrho,\varepsilon)$ is the standard covering number— the smallest size of an ε-net that covers $\mathcal{F}$ under the metric ϱ; see Definition 5 for details. In comparison, Lemma 3 is much more challenging than proving classical isotropic covering number bounds (57) because

- 1) we consider ellipsoidal metric;
- 2) we require the prior μ to be independent with Σ and ε.

We need to firstly understand how such classical results are proved, and then proceed to generalized them. This suggests that deep mathematical insights are necessary for the purpose to study neural networks generalization, as we will introduce below.

From Pure Mathematics to Machine Learning Language. Understanding the classical proof for the Grassmannian and generalizing them to prove Lemma 3 necessitate the a deep dive in to the geometry and algebra of subspaces and Grassmannians. In fact, traditional treatments to study Grassmannian manifold often invoke advanced machinery—ranging from differential geometry [Bendokat et al., 2024] and Lie-group theory [Szarek, 1997] to algebraic geometry [Devriendt et al., 2024], and the seminal covering number proof [Szarek, 1997] is particularly stated in Lie-algebra and differential-geometry language.

Motivated by the subsequent covering number proof [Pajor, 1998] that uses relatively more elementary language, we give an exposition that is elementary and entirely self-contained, relying only on matrix-analysis and learning-theoretic techniques familiar from machine learning. In particular, every "advanced" fact—for example, the group theory of continuous symmetries traditionally handled via Lie groups—is derived by elementary means (explicit matrix parameterizations, principal-angle/cosine-sine representations, and basic spectral arguments) while preserving the high-level geometric intuition. We hope that this versatile framework—and our novel contributions (e.g., Definition 4 and Lemma 3), which are new even in a pure-mathematics setting—will establish subspaces, the Grassmannian, and their underlying algebraic structures as powerful tools for future machine learning applications.

D.1 Grassmannian Manifold, Stiefel Parameterization, and Orthogonal Groups

Fix integers $r \le d$. Define

$$\mathrm{Gr}(d,r) \;:=\; \big\{ r\text{–dimensional linear subspaces of } \mathbb{R}^d \big\}$$

as the *Grassmann manifold*. Write

$$\mathrm{St}(d,r) \;:=\; \big\{\, V \in \mathbb{R}^{d\times r} \;:\; V^\top V = I_r \,\big\}$$

for the *Stiefel manifold* of r orthonormal columns in $\mathbb{R}^d$. $\mathrm{St}(d,r)$ is a convenient *parameterization* of that class $\mathrm{Gr}(d,r)$.

If for subspace $\mathcal{V} \in \mathrm{Gr}(d,r)$ and matrix $V \in \mathrm{St}(d,r)$ we have $\mathcal{V} = \mathrm{span}(V)$, then we say V is a *parameterization matrix* of $\mathcal{V}$. Though such parameterization is not unique, the associated orthogonal projector and projection metric are both unique. Moreover, the anisometric projection we define in Definition 4 is also unique. We will prove these shortly.

Write

$$O(r) := \{Q \in \mathbb{R}^{r\times r} : Q^\top Q = QQ^\top = I_r\}$$

to be the *orthogonal group*. Optionally, we also state that (in the real setting)

$$\mathrm{Gr}(d,r) \;\cong\; O(d)\big/\big(O(r)\times O(d-r)\big) \;\cong\; \mathrm{Gr}\big(d,\, d-r\big), \tag{58}$$

where "/" denotes the *quotient* and "≅" denotes a canonical *isomorphism* (indeed, a *diffeomorphism* of smooth manifolds or a *homeomorphism* of topological manifolds; see, e.g., Chapter 1.5 in [Awodey, 2010]). Moreover, $\mathrm{Gr}(d,r)$ can be regarded as a standard *algebraic variety* [Devriendt et al., 2024]. We do not aim to explain these notions in detail, but merely note that:

1. The geometric properties of $\mathrm{Gr}(d,r)$ coincide with those of $\mathrm{Gr}(d,d-r)$ under this isomorphism (geometric equivalence).

2. The number of degrees of freedom of $\mathrm{Gr}(d, r)$ is

$$\underbrace{\frac{d(d-1)}{2}}_{\dim O(d)} - \underbrace{\frac{r(r-1)}{2}}_{\dim O(r)} - \underbrace{\frac{(d-r)(d-r-1)}{2}}_{\dim O(d-r)} = r(d-r), \tag{59}$$

which also appears as the dimension factor in the precise covering-number bounds (57).

We now define the orthogonal projector and the projection metric on the Grassmannian manifold.

Definition of Orthogonal Projector. For $V \in \mathrm{St}(d, r)$ and its column-space $\mathcal{V} = \mathrm{span}(V)$, define the rank-$r$ orthogonal projector[3]

$$P_{\mathcal{V}} := VV^\top \quad \in \mathbb{R}^{d\times d}. \tag{60}$$

Then $P_{\mathcal{V}}$ depends *only* on the subspace $\mathcal{V}$. Indeed, if $Q \in O(r)$ then $(VQ)(VQ)^\top = VQQ^\top V^\top = VV^\top$, so V and VQ represent the same subspace. Hence the map

$$\Psi : \mathrm{St}(d, r) \longrightarrow \mathrm{Gr}(d, r), \quad V \mapsto \mathrm{span}(V),$$

is an $O(r)-$*quotient*: two frames give the same subspace iff they differ by a right orthogonal factor.

Ellipsoidal Projection Metric. Following Definition 4, for $\mathcal{V}, \bar{\mathcal{V}} \in \mathrm{Gr}(d, r)$,

$$\varrho_{\mathrm{proj},\Sigma}(\mathcal{V}, \bar{\mathcal{V}}) := \|\Sigma^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\mathrm{op}}, \tag{61}$$

where $\mathcal{P}_{\mathcal{V}} := VV^\top$ for *any* V such that $\mathrm{span}(V) = \mathcal{V}$ (similarly $P_{\bar{\mathcal{V}}}$). Because $P_{\mathcal{V}}$ is unique for each subspace, $\varrho_{\mathrm{proj},\Sigma}$ is well defined (independent of the chosen V). The metric can be pulled back to $\mathrm{St}(d, r)$:

$$\varrho_{\mathrm{proj},\Sigma}(V, \bar{V}) := \varrho_{\mathrm{proj},\Sigma}\big(\mathrm{span}(V), \mathrm{span}(\bar{V})\big) = \|\Sigma^{\frac{1}{2}}(VV^\top - \bar{V}\bar{V}^\top)\|_{\mathrm{op}}. \tag{62}$$

D.2 Principal Angles between Subspaces

W study how metrics and angles between images $\mathcal{V}$ and $\bar{\mathcal{V}}$ affect their spectral properties. We introduce principal angles and the cosine–sine (CS) decomposition—standard tools for analyzing subspaces (see, e.g., Chapter 6.4.3 in [Golub and Van Loan, 2013]).

Principle Angles and Cosine-Sine representation. Let U and $\bar{U}$ be two $d \times d$ orthogonal matrix, and V and $\bar{V}$ be the first r columns of U and $\bar{U}$, respectively. We are interested in studying the metrics and angles between $r-$dimensional subspaces $\mathcal{V} = \mathrm{span}(V)$ and $\bar{\mathcal{V}} = \mathrm{span}(\bar{V})$. Formally, denote

$$U, \bar{U} \in O(d), \qquad U = \big[V \;\; V_\perp\big], \;\; \bar{U} = \big[\bar{V} \;\; \bar{V}_\perp\big],$$

[3] By elementary linear algebra, the matrix definition of the orthogonal projector $\mathcal{P}$ here coincides with the ℓ_2-projection characterized in Lemma 14; thus the notation is consistent.

where

$$V, \bar{V} \in \mathbb{R}^{d\times r}, \quad V^\top V = I_r, \quad \bar{V}^\top \bar{V} = I_r,$$

and

$$V_\perp, \bar{V}_\perp \in \mathbb{R}^{d\times(d-r)}, \quad V_\perp^\top V_\perp = I_{d-r}, \quad \bar{V}_\perp^\top \bar{V}_\perp = I_{d-r}.$$

Since $U, \bar{U} \in O(d)$, their product $U^\top \bar{U}$ is itself orthogonal. Writing

$$U^\top \bar{U} \;=\; \begin{pmatrix} V^\top \\ V_\perp^\top \end{pmatrix} \, [\bar{V} \;\; \bar{V}_\perp] \;=\; \begin{pmatrix} V^\top \bar{V} & V^\top \bar{V}_\perp \\ V_\perp^\top \bar{V} & V_\perp^\top \bar{V}_\perp \end{pmatrix},$$

define the four blocks

$$\underbrace{C}_{r\times r} \;=\; V^\top \bar{V}, \qquad \underbrace{C_\perp}_{r\times(d-r)} \;=\; V^\top \bar{V}_\perp, \tag{63}$$

$$\underbrace{S}_{(d-r)\times r} \;=\; V_\perp^\top \bar{V}, \qquad \underbrace{S_\perp}_{(d-r)\times(d-r)} \;=\; V_\perp^\top \bar{V}_\perp. \tag{64}$$

Thus

$$U^\top \bar{U} \;=\; \begin{pmatrix} C & C_\perp \\ S & S_\perp \end{pmatrix} \;\in\; O(d).$$

Now we introduce principal angles between $\mathcal{V} = \mathrm{span}(V)$ and $\bar{\mathcal{V}} = \mathrm{span}(\bar{V})$ by writing

$$C = V^\top \bar{V} \;=\; Q_1 \,\mathrm{diag}(\cos\theta_1, \cdots, \cos\theta_r)\, W_1^\top, \quad Q_1, W_1 \in O(r), \tag{65}$$

where

$$0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_r \le \pi/2$$

are called the principle angles between subspaces $\mathcal{V}$ and $\bar{\mathcal{V}}$. Simultaneously, we have that the eigenvalues of S, $C_\perp$, $S_\perp$ are (notation sepc means spectrum, the set of singular values)

$$\begin{aligned} \mathrm{spec}(S) &= \{-\sin\theta_1, \cdots, -\sin\theta_{\min\{r,d-r\}}, \underbrace{0, \cdots, 0}_{\max\{d-2r,0\}}\}, \\ \mathrm{spec}(C_\perp) &= \{\sin\theta_1, \cdots, \sin\theta_{\min\{r,d-r\}}, \underbrace{0, \cdots, 0}_{\max\{d-2r,0\}}\} \\ \mathrm{spec}(S_\perp) &= \{\cos\theta_1, \cdots, \cos\theta_{\min\{r,d-r\}}, \underbrace{1, \cdots, 1}_{\max\{d-2r,0\}}\}. \end{aligned} \tag{66}$$

The above representation in (65) and (66) are without loss of generality: if $r \le d-r$, then all the four spectrum contain all r principal angles; if $r > d-r$, then only first $d-r$ principal angles $\{\theta_k\}_{k=1}^{d-r}$ can be smaller than $\pi/2$ and $\theta_k = 0$ for all $d-r+1 \le k \le r$.

The cosine–sine representation of the eigenvalues in (65) and (66) motivates our notation C and S when defining block matrices in (63) and (64). This representation is an immediate consequence of the classical CS decomposition for orthogonal matrices [Paige and Wei, 1994, Golub and Van Loan, 2013], and we henceforth regard the resulting eigenvalue characterization as given.

Projection Metric via Principal Angles. For subspaces $\mathcal{V}$ and $\bar{\mathcal{V}}$, recall that for orthogonal projectors

$$P_{\mathcal{V}} = VV^\top, \quad P_{\bar{\mathcal{V}}} = \bar{V}\bar{V}^\top,$$

It is known that the projection metric defined in (61) and (62) are equal to $\sin\theta_r$, sine of the largest principal angle between the two subspaces. Formally, there is the fact (see, e.g., the last equation in Section 6.4.3 in [Golub and Van Loan, 2013])

$$\varrho_{\text{proj}} = \|P_{\mathcal{V}} - P_{\bar{\mathcal{V}}}\|_{\text{op}} \;=\; \max_{1\le k\le r} \sin\theta_k \;=\; \sin\theta_r. \tag{67}$$

Here θ_i is the i-th principal-angle between $\mathcal{V}$ and $\bar{\mathcal{V}}$, and the spectral norm of the difference of two projectors equals the largest of these sines.

D.3 Local Charts of the Grassmannian

In differential geometry, a *chart* is a single local coordinate map. An *atlas* is the whole collection of charts that covers the manifold. We introduce a useful atlas that consists of finite graph charts, which only rely on elementary linear algebra and avoid more advanced Lie algebra and exponential map techniques in Szarek [1997].

Choose a reference subspace $\bar{\mathcal{V}} \in \mathrm{Gr}(d,r)$ and its parameterization matrix $\bar{V} \in \mathrm{St}(d,r)$. Denote $X \in \mathbb{R}^{(d-r)\times r}$ to be mappings from $r-$dimensional subspace $\bar{\mathcal{V}}$ to $(d-r)-$dimensional subspace $\bar{\mathcal{V}}_\perp$. Every r–dimensional subspace close to $\bar{\mathcal{V}}$ can be written as the *graph*

$$\mathcal{V}(X) \;:=\; \operatorname{span}\Big\{[\bar{V}\bar{V}_\perp]\begin{pmatrix} I_r \\ X\end{pmatrix}\Big\}, \qquad X \;\in\; \mathbb{R}^{(d-r)\times r}, \tag{68}$$

where $\mathcal{V}(X)$ is the subspace spanned by the columns of $[\bar{V}\ \bar{V}_\perp]\begin{pmatrix} I_r \\ X\end{pmatrix}$ (the matrix multiplication). Given the reference subspace $\bar{\mathcal{V}}$, define the local *graph chart* from $\mathbb{R}^{(d-r)\times r}$ to $\mathrm{Gr}(d,r)$ by

$$\phi_{\bar{\mathcal{V}}}:\ X \;\longmapsto \mathcal{V}(X) \in\ \mathrm{Gr}(d,r). \tag{69}$$

Note that for the $(d-r)\times r$ zero matrix (denoted as 0), we have $\phi_{\bar{\mathcal{V}}}(0) = \bar{\mathcal{V}}$.

Intuition for the graph chart. If a subspace $\mathcal{V}$ is close to $\bar{\mathcal{V}}$—specifically, $\varrho_{\text{proj}}(\mathcal{V},\bar{\mathcal{V}}) = \sin\theta_r < 1$—then all principal angles between $\mathcal{V}$ and $\bar{\mathcal{V}}$ satisfy $\theta_i < \pi/2$. Equivalently, the orthogonal projection $P_{\bar{\mathcal{V}}}$ restricted to $\mathcal{V}$ is a bijection $P_{\bar{\mathcal{V}}}|\mathcal{V} : \mathcal{V} \to \bar{\mathcal{V}}$. In the orthonormal basis $[\bar{V}\ \bar{V}_\perp]$, this means every $v \in \mathcal{V}$ can be written uniquely as

$$v = [\bar{V}\bar{V}_\perp]\begin{pmatrix} \bar{v} \\ X\,\bar{v}\end{pmatrix}, \qquad \bar{v} \in \operatorname{span}\left\{\begin{pmatrix} I_r \\ 0\end{pmatrix}\right\},$$

for a linear map $X \in \mathbb{R}^{(d-r)\times r}$. Thus, locally around $\bar{\mathcal{V}}$ (all principal angles $< \pi/2$), every $r-$plane admits—and is uniquely determined by—its graph parameter X. We call X the *graph parameterization* of $\mathcal{V}(X)$ in this image. This is formalized as the following lemma.

Lemma 20 (Local Bijection of Graph Chart) *Fix an orthonormal decomposition $\mathbb{R}^d = \bar{\mathcal{V}} \oplus \bar{\mathcal{V}}_\perp$ with basis $[\bar{V}\ \bar{V}_\perp]$. Then every $r-$dimensional subspace $\mathcal{V}$ such that $\varrho_{\text{proj}}(\mathcal{V},\bar{\mathcal{V}}) < 1$ (i.e., all principal angles $< \pi/2$) can be written uniquely as a graph*

$$\mathcal{V} = \phi_{\bar{\mathcal{V}}}(X) = \operatorname{span}\Big\{[\bar{V}\ \bar{V}_\perp]\begin{pmatrix} I_r \\ X\end{pmatrix}\Big\}, \qquad X \in \mathbb{R}^{(d-r)\times r}.$$

Proof of Lemma 20: If $V \in \mathrm{St}(d,r)$ spans $\mathcal{V}$, block it in the $[\bar{V}\ \bar{V}_\perp]$ basis: denote

$$\begin{pmatrix} A \\ B \end{pmatrix} := \begin{pmatrix} \bar{V}^\top \\ \bar{V}_\perp^\top \end{pmatrix} V \quad (A \in \mathbb{R}^{r\times r},\ B \in \mathbb{R}^{(d-r)\times r}).$$

Then by the principal angle representation (65), $A = \bar{V}^\top V$ is invertible iff all principal angles $< \pi/2$, and choosing

$$X = B\,A^{-1}$$

leads to

$$\mathcal{V} = \mathrm{span}(V) = \mathrm{span}\left\{[\bar{V}\bar{V}_\perp]\begin{pmatrix} A \\ B \end{pmatrix}\right\} = \mathrm{span}\left\{[\bar{V}\bar{V}_\perp]\begin{pmatrix} I_r \\ X \end{pmatrix}\right\},$$

where the last equality is because for invertible A one always have $\mathrm{span}(ZA) = \mathrm{span}(Z)$ for any matrix Z.

We have already shown existence. For uniqueness, assuming there are two different X_1, X_2 such taht $\phi_{\bar{\mathcal{V}}}(X_1) = \phi_{\bar{\mathcal{V}}}(X_2)$. Because two bases of the same r–dimensional subspace differ by an invertible change of coordinates, so there exists an invertible $r \times r$ matrix Y such that

$$[\bar{V}\bar{V}_\perp]\begin{pmatrix} I_r \\ X_1 \end{pmatrix} Y = [\bar{V}\bar{V}_\perp]\begin{pmatrix} I_r \\ X_2 \end{pmatrix},$$

which results in $Y = I_r$ and $X_1 = X_2$. Thus the parameterization X of $\mathcal{V}$ is unique.

□

Sine-tangent Relationship in Graph Chart. We will show that there is a sine-tangent relationship between $\varrho_{\mathrm{proj}}(\mathcal{V}, \bar{\mathcal{V}})$ and $\|X\|_{\mathrm{op}}$. To be specific, we have the following lemma.

Lemma 21 (Sine-Tangent Relationship in Graph Chart) *Denote θ_r is the maximal principal angle between the subspaces $\mathcal{V}(X)$ and $\bar{\mathcal{V}}$, defined in (65). For the graph chart (69), we have*

$$\varrho_{\mathrm{proj}}(\mathcal{V}(X), \bar{\mathcal{V}}) = \sin\theta_r, \quad \|X\|_{\mathrm{op}} = \tan\theta_r.$$

The above relationship immediately implies that

$$\varrho_{\mathrm{proj}}(\mathcal{V}(X), \bar{\mathcal{V}}) = \|X\|_{\mathrm{op}}/\sqrt{1 + \|X\|_{\mathrm{op}}^2}.$$

Proof of Lemma 21: Given the fact $\varrho_{\mathrm{proj}}(\mathcal{V}(X), \bar{\mathcal{V}}) = \sin\theta_r$ (which is already shown in (67)), where θ_r is the largest principal angle between the subspaces $\mathcal{V}(X)$ and the reference subspace $\bar{\mathcal{V}}$, we want to show $\|X\|_{\mathrm{op}} = \tan\theta_r$.

Step 1: Setup and Simplification. The projection metric is invariant under orthogonal transformations of the ambient space $\mathbb{R}^d$. We can therefore choose a coordinate system that simplifies the calculations without loss of generality. We choose a basis such that the reference frame $\bar{V}$ and its orthogonal complement $\bar{V}_\perp$ are represented as:

$$\bar{V} = \begin{pmatrix} I_r \\ 0 \end{pmatrix} \in \mathrm{St}(d,r), \qquad \bar{V}_\perp = \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix} \in \mathrm{St}(d, d-r). \tag{70}$$

In this basis, the reference subspace is $\bar{\mathcal{V}} = \operatorname{span}(\bar{V})$. The parameterization matrix (orthonormal basis) $V(X)$ for the subspace $\mathcal{V}(X)$ simplifies to (here $(I_r + X^\top X)^{-1/2}$ normalize $V(X)$ to be an orthogonal matrix):

$$V(X) = [\bar{V}\ \bar{V}_\perp] \begin{pmatrix} I_r \\ X \end{pmatrix} (I_r + X^\top X)^{-1/2} = I_d \begin{pmatrix} I_r \\ X \end{pmatrix} (I_r + X^\top X)^{-1/2} = \begin{pmatrix} I_r \\ X \end{pmatrix} (I_r + X^\top X)^{-1/2}, \quad (71)$$

where the second equality follows from our choice of basis without loss of generality: the reference frame $\bar{V}$ and its complement $\bar{V}_\perp$ are represented as block identity matrices as in (70).

Step 2: Projection Metric and Principal Angles. A fundamental result in matrix analysis, our equation (65), states that the cosines of the principal angles, $\cos\theta_i$, between two subspaces spanned by orthonormal bases V and $\bar{V}$ are the singular values of $V^\top \bar{V}$. In our case, the principal angles between $\mathcal{V}(X)$ and $\bar{\mathcal{V}}$ are determined by the singular values of $V(X)^\top \bar{V}$—which are, equivalently, the singular values of $\bar{V}^\top V(X)$.

Step 3: Calculation of $\cos\theta_i$. Let's compute the matrix product $\bar{V}^\top V(X)$ using our simplified forms:

$$\begin{aligned}
\bar{V}^\top V(X) &= \begin{pmatrix} I_r & 0 \end{pmatrix} \left[\begin{pmatrix} I_r \\ X \end{pmatrix} (I_r + X^\top X)^{-1/2} \right] \\
&= \left(\begin{pmatrix} I_r & 0 \end{pmatrix} \begin{pmatrix} I_r \\ X \end{pmatrix} \right) (I_r + X^\top X)^{-1/2} \\
&= I_r \cdot (I_r + X^\top X)^{-1/2} \\
&= (I_r + X^\top X)^{-1/2}.
\end{aligned}$$

To find the singular values of this matrix, we use the Singular Value Decomposition (SVD) of X. Let $X = U\Sigma W^\top$, where $U \in \mathbb{R}^{(d-r)\times(d-r)}$ and $W \in \mathbb{R}^{r\times r}$ are orthogonal, and $\Sigma \in \mathbb{R}^{(d-r)\times r}$ is a rectangular diagonal matrix with the singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ on its diagonal. The spectral norm is $\|X\|_{\text{op}} = \lambda_1$.

Then, $X^\top X = (U\Sigma W^\top)^\top (U\Sigma W^\top) = W\Sigma^\top U^\top U\Sigma W^\top = W\Sigma_r^2 W^\top$, where Σ_r^2 is the $r \times r$ diagonal matrix with entries λ_i^2. So, the matrix $I_r + X^\top X = W(I_r + \Sigma_r^2)W^\top$. Its inverse square root is: $(I_r + X^\top X)^{-1/2} = W(I_r + \Sigma_r^2)^{-1/2} W^\top$.

The singular values of $\bar{V}^\top V(X)$ are the diagonal entries of $(I_r + \Sigma_r^2)^{-1/2}$, which are: $s_i = \frac{1}{\sqrt{1+\lambda_i^2}}$. These singular values are the values of $\cos\theta_i$. The largest principal angle, θ_r, corresponds to the smallest cosine value. This occurs when the singular value λ_i is largest, i.e., for $\lambda_1 = \|X\|_{\text{op}}$. Thus,

$$\cos\theta_r = \frac{1}{\sqrt{1 + \|X\|_{\text{op}}^2}}.$$

Step 4: Deriving $\tan\theta_r$. Using the fundamental trigonometric identity $\sin^2\theta + \cos^2\theta = 1$ and the fact that principal angles lie in $[0, \pi/2]$, we have:

$$\tan\theta_r = \|X\|_{\text{op}}.$$

We have shown that for graph charts, there is the relationship $\varrho_{\text{proj}}(\mathcal{V}(X), \bar{\mathcal{V}}) = \sin\theta_r$ and $\|X\|_{\text{op}} = \tan\theta_r$. This suggests

$$\varrho_{\text{proj}}(\mathcal{V}(X), \bar{\mathcal{V}}) = \frac{\|X\|_{\text{op}}}{\sqrt{1 + \|X\|_{\text{op}}^2}}.$$

□

D.4 Global Atlas of Graph Charts

For the Grassmannian $\text{Gr}(d, r)$ we have that for all $\varepsilon > 0$, we have the coarse covering number bound $\text{N}(\text{Gr}(d, r), \varrho_{\text{proj}}, \varepsilon) \leq C^{\frac{r(d-r)}{\varepsilon}}$, where $C > 0$ is an absolute constant. This is a coarse bound—its dependence is exponential in $1/\varepsilon$ (hence not rate–optimal; the optimal dependence is polynomial)—and we use it only as a preliminary supporting estimate. This coarse estimate suggests that, a finite $O(e^{r(d-r)})$ number of graph charts are sufficient to cover the entire $\text{Gr}(d, r)$ such that every subspace $\mathcal{V} \in \text{Gr}(d, r)$ is contained in the image of a graph chart with its graph parameterization X satisfies $\|X\|_{\text{op}} \leq 1$. From this intuition, we have the following lemma.

Lemma 22 (Pointwise Dimension Consequence of Finite Global Atlas) *The uniform prior $\mu = \text{Unif}(\text{Gr}(d, r))$ satisfies that for every $\mathcal{V} \in \text{Gr}(d, r)$, every PSD matrix Σ and every $\varepsilon > 0$,*

$$\log \frac{1}{\mu(B_{\varrho_{\text{proj},\Sigma}}(\mathcal{V}, \varepsilon))} \leq C_1 r(d - r) + \sup_{X \in \mathcal{X}} \log \frac{1}{\text{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_{\text{proj},\Sigma}(\mathcal{V}(X), \mathcal{V}(X')\} \leq \varepsilon)},$$

where $\mathcal{X} = \{X \in \mathbb{R}^{(d-r)r} : \|X\|_{\text{op}} \leq 1\}$ and $\bar{\mathcal{X}} = \{X \in \mathbb{R}^{(d-r)r} : \|X\|_{\text{op}} \leq 2\}$ (we make $\bar{\mathcal{X}}$ slightly larger than $\mathcal{X}$ for later technical derivation), $\text{Unif}(\bar{\mathcal{X}})\{\cdot\}$ is the uniform measure over $\bar{\mathcal{X}}$, and $C_1 > 0$ is an absolute constant.

Proof of Lemma 22: Proposition 6 in [Pajor, 1998] prove a coarse covering number bound

$$\text{N}(\text{Gr}(d, r), \varrho_{\text{proj}}, \varepsilon) \leq C^{\frac{r(d-r)}{\varepsilon}}$$

where $C > 0$ is an absolute constant; this coarse estimate is exponential rather then polynomial in ε, so it is used only for preliminary supporting purposes. For every $\mathcal{V} \in \text{Gr}(d, r)$, by the homogeneity of the Grassmannian (under the action of $O(d)$), the ϱ_{proj}-ball $B_{\text{proj}}(\mathcal{V}, \varepsilon)$ has volume independent of its center. We therefore refer to this common value as the volume of an ε–ϱ_{proj} ball, written as $\text{Vol}(\varepsilon - \varrho_{\text{proj}} \text{ ball})$. By the definition of covering number (see Definition 5 and the subsequent inequality for background), we have that

$$\text{N}(\text{Gr}(d, r), \varrho_{\text{proj}}, \varepsilon) \cdot \text{Vol}(\varepsilon - \varrho_{\text{proj}} \text{ ball}) \geq \text{Vol}(\text{Gr}(d, r)),$$

then for the uniform prior $\nu = \text{Unif}(\text{Gr}(d, r))$, we have that for every $\bar{\mathcal{V}} \in \text{Gr}(d, r)$,

$$\log \frac{1}{\nu(B_{\varrho_{\text{proj}}}(\bar{\mathcal{V}}, \varepsilon))} = \log \frac{\text{Vol}(\text{Gr}(d, r))}{\text{Vol}(\varepsilon - \varrho_{\text{proj}} \text{ ball})} \leq r(d - r)\frac{\log C}{\varepsilon}.$$

Taking $\varepsilon = 1/\sqrt{2}$, we obtain:

$$\log \frac{1}{\nu(B_{\varrho_{\text{proj}}}(\bar{\mathcal{V}}, 1/\sqrt{2}))} \leq C_1 r(d - r), \tag{72}$$

where $C_1 > 0$ is an absolute constant. By Lemma 21, we have that inside the ball $B_{\varrho_{\mathrm{proj}}}(\bar{\mathcal{V}}, 1/\sqrt{2})$, by choosing $\bar{\mathcal{V}}$ as the reference subspace, the graph parameterization X of $\mathcal{V}$ satisfies

$$\|X\|_{\mathrm{op}} \leq 1.$$

See (68) for the definition of this graph chart parameterization; the existence and uniqueness of the parameterization X is by Lemma 20 (local bijection of graph chart). Furthermore, again by Lemma 20 and Lemma 21, $\mathcal{X} = \{X \in \mathbb{R}^{(d-r)r} : \|X\|_{\mathrm{op}} \leq 1\}$ satisfies ($\cong$ means isomorphism/bijection)

$$B_{\varrho_{\mathrm{proj}}}(\bar{\mathcal{V}}, 1/\sqrt{2}) \cong \mathcal{X} \subset \bar{\mathcal{X}} \cong B_{\varrho_{\mathrm{proj}}}(\bar{\mathcal{V}}, 2/\sqrt{5}). \tag{73}$$

Let

$$\mu_{\bar{\mathcal{V}}} = \mathrm{Unif}(B_{\mathrm{proj}}(\bar{\mathcal{V}}, 2/\sqrt{5})), \quad \mu(\mathcal{V}) = \int \nu(\bar{\mathcal{V}})\mu_{\bar{\mathcal{V}}}(\mathcal{V})d\bar{\mathcal{V}} = \mathrm{Unif}(\mathrm{Gr}(d,r)).$$

Then we have

$$\begin{aligned}
\log \frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))} &= \log \frac{1}{\int \nu(\bar{\mathcal{V}})\mu_{\bar{\mathcal{V}}}(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))d\bar{\mathcal{V}}} \\
&= \log \frac{1}{\int \nu(\bar{\mathcal{V}})\mu_{\bar{\mathcal{V}}}(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon) \cap B_{\mathrm{proj}}(\bar{\mathcal{V}}, 2/\sqrt{5}))d\bar{\mathcal{V}}} \\
&\leq \log \frac{1}{\nu(B_{\varrho_{\mathrm{proj}}}(\mathcal{V}, 1/\sqrt{2})) \min\limits_{\bar{\mathcal{V}} \in B_{\varrho_{\mathrm{proj}}}(\mathcal{V}, 1/\sqrt{2})} \mu_{\bar{\mathcal{V}}}(X' \in \bar{\mathcal{X}} : \varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X), \mathcal{V}(X')) \leq \varepsilon)} \\
&\leq C_1 r(d-r) + \sup_{X \in \mathcal{X}} \log \frac{1}{\mathrm{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X), \mathcal{V}(X')) \leq \varepsilon\}},
\end{aligned}$$

where the first inequality is by restricting $\bar{\mathcal{V}}$ to $B_{\varrho_{\mathrm{proj}}}(\mathcal{V}, 1/\sqrt{2})$; and the second inequality is by (72) as well as the bijection stated in (73) and Lemma 20. Note that we use different radius here than in $\mu_{\bar{\mathcal{V}}}$ to enusre that the set $\bar{\mathcal{X}}$ for X', which is inside the uniform distribution in the final bound, to be larger than the domain $\mathcal{X}$ for X to take sup. This will help later technical derivation.

□

D.5 Decomposition and Lipchitz Properties inside Graph Chart

We apply a non-perturbative analysis to the ellipsoidal projection metric.

Lemma 23 (Non-Perturbative Decomposition of Projector Difference) *Let $X, X' \in \mathbb{R}^{(d-r)\times r}$ be two matrices. Given any reference subspace $\bar{\mathcal{V}}$, consider the graph chart $\phi_{\bar{\mathcal{V}}} : X \mapsto \mathcal{V}(X)$ defined in (68). Then the difference between two projectors $\mathcal{P}_{\mathcal{V}(X)}$, $\mathcal{P}_{\mathcal{V}(X')}$ be decomposed as follows:*

$$\begin{aligned}
&\mathcal{P}_{\mathcal{V}(X)} - \mathcal{P}_{\mathcal{V}(X')} \\
=&\mathcal{P}_{\mathcal{V}(X)_\perp} \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix} (X - X') \begin{pmatrix} I_r & 0 \end{pmatrix} \mathcal{P}_{\mathcal{V}(X')} + \mathcal{P}_{\mathcal{V}(X)} \begin{pmatrix} I_r \\ 0 \end{pmatrix} (X^\top - X'^\top) \begin{pmatrix} 0 & I_{d-r} \end{pmatrix} \mathcal{P}_{\mathcal{V}(X')_\perp}.
\end{aligned}$$

Proof of Lemma 23: The projector is invariant under orthogonal transformations of the ambient space R^d. We can therefore choose a coordinate system that simplifies the calculations without loss of generality. By the matrix representation (71) (which, without loss of generality, uses a convenient orthogonal basis specified by (70)), we denote

$$A(X) = \begin{pmatrix} I_r \\ X \end{pmatrix}, \quad M(X) = (I_r + X^\top X)^{-1},$$

and have the following facts:

$$\begin{aligned} V(X) =& A(X)M(X)^{1/2}, \\ \mathcal{P}_{\mathcal{V}}(X) =& A(X)M(X)A(X)^\top = A(X)M(X)\begin{pmatrix} I_r & X^\top \end{pmatrix} \quad (74) \\ \mathcal{P}_{\mathcal{V}(X)} - \mathcal{P}_{\mathcal{V}(X')} =& A(X)M(X)A(X)^\top - A(X')M(X')A(X')^\top \\ A(X)M(X) =& \mathcal{P}_{\mathcal{V}(X)} \begin{pmatrix} I_r \\ 0 \end{pmatrix} \quad (75) \\ A(X)M(X)X^\top =& \mathcal{P}_{\mathcal{V}(X)} \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix}, \quad (76) \end{aligned}$$

where (75) and (76) are straightforward consequences of (74).

We begin with a non-perturbative decomposition:

$$\begin{aligned} & \mathcal{P}_{\mathcal{V}(X)} - \mathcal{P}_{\mathcal{V}(X')} \\ =& A(X)M(X)A(X)^\top - A(X')M(X')A(X')^\top \\ =& (A(X) - A(X'))M(X')A(X')^\top + A(X)(M(X) - M(X'))A(X')^\top + A(X)M(X)(A(X) - A(X'))^\top. \end{aligned} \quad (77)$$

We continue to decompose each term non-perturbatively. First,

$$\begin{aligned} & (A(X) - A(X'))M(X')A(X')^\top \\ =& \begin{pmatrix} 0 \\ X - X' \end{pmatrix} M(X')A(X')^\top \\ =& \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix} (X - X')M(X')A(X')^\top \\ =& \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix} (X - X') \begin{pmatrix} I_r & 0 \end{pmatrix} \mathcal{P}_{\mathcal{V}(X')}, \end{aligned} \quad (78)$$

where the last equality uses the fact (75) and symmetry of $\mathcal{P}_{\mathcal{V}(X)}$.

Second, because we have the non-perturbative decomposition

$$\begin{aligned} & M(X) - M(X') \\ =& (I_r + X^\top X)^{-1} \left((I_r + X'^\top X') - (I_r + X^\top X) \right) (I_r + X'^\top X')^{-1} \\ =& (I_r + X^\top X)^{-1} \left(X'^\top X' - X^\top X \right) (I_r + X'^\top X')^{-1} \\ =& (I_r + X^\top X)^{-1} \left(X^\top (X' - X) + (X'^\top - X^\top) X' \right) (I_r + X'^\top X')^{-1} \\ =& M(X)X^\top (X' - X)M(X') + M(X)(X'^\top - X^\top)X'M(X'), \end{aligned}$$

we have

$$
\begin{aligned}
&A(X)(M(X)-M(X'))A(X')^\top \\
=&A(X)M(X)X^\top(X'-X)M(X')A(X')^\top + A(X)M(X)(X'-X)X'M(X')A(X')^\top \\
=&-\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}(X-X')\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')} - \mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X-X')\begin{pmatrix}0 & I_{d-r}\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')},
\end{aligned} \tag{79}
$$

where the last equality uses the fact (75) and the fact (76).

Third, we have

$$
\begin{aligned}
&A(X)M(X)(A(X)-A(X'))^\top \\
=&A(X)M(X)\begin{pmatrix}0 & X^\top - X'^\top\end{pmatrix} \\
=&\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X^\top - X'^\top)\begin{pmatrix}0 & I_{d-r}\end{pmatrix},
\end{aligned} \tag{80}
$$

where the last equality uses the fact (75).

Substituting (78), (79), (80) back into (77), we have

$$
\begin{aligned}
&\mathcal{P}_{\mathcal{V}(X)} - \mathcal{P}_{\mathcal{V}(X')} \\
=&\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}(X-X')\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')} \\
&-\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}(X-X')\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')} - \mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X-X')\begin{pmatrix}0 & I_{d-r}\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')} \\
&+\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X^\top - X'^\top)\begin{pmatrix}0 & I_{d-r}\end{pmatrix} \\
=&\mathcal{P}_{\mathcal{V}(X)_\perp}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}(X-X')\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')} + \mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X^\top - X'^\top)\begin{pmatrix}0 & I_{d-r}\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')_\perp},
\end{aligned}
$$

where the last equality uses $I_d - \mathcal{P}_{\mathcal{V}(X)} = \mathcal{P}_{\mathcal{V}(X)_\perp}$ and $I_d - \mathcal{P}_{\mathcal{V}(X')} = \mathcal{P}_{\mathcal{V}(X')_\perp}$.

□

Building upon the non-perturbative decomposition in Lemma 23, we have the following Lipchitz property of graph chart.

Lemma 24 (Lipchitz of Graph Chart) *Let $X, X' \in \mathbb{R}^{(d-r)\times r}$ be two matrices. Given any reference subspace $\bar{\mathcal{V}}$, consider the graph chart defined in (71). Then the ellipsoidal projection metric is Lipschitz to ellipsoidal spectral metrics as follows: for every rank-r PSD $\Sigma \in \mathbb{R}^{d\times d}$,*

$$
\begin{aligned}
&\varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X), \mathcal{V}(X')) \\
\le&\left\|\left(\begin{pmatrix}0 & I_{d-r}\end{pmatrix}\mathcal{P}_{\mathcal{V}(X)}^\top \Sigma \mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}\right)^{\frac{1}{2}}(X-X')\right\|_{\mathrm{op}} + \left\|\left(\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X)_\perp}^\top \Sigma \mathcal{P}_{\mathcal{V}(X)_\perp}\begin{pmatrix}I_r\\ 0\end{pmatrix}\right)^{\frac{1}{2}}(X^\top - X'^\top)\right\|_{\mathrm{op}}.
\end{aligned}
$$

Proof of Lemma 24: By Lemma 23, we have

$$\begin{aligned}
&\varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X),\mathcal{V}(X')) = \left\|\Sigma^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}(X)} - \mathcal{P}_{\mathcal{V}(X')})\right\|_{\mathrm{op}} \\
&= \left\|\Sigma^{\frac{1}{2}}\mathcal{P}_{\mathcal{V}(X)_\perp}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}(X-X')\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')} + \Sigma^{\frac{1}{2}}\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X^\top - X'^\top)\begin{pmatrix}0 & I_{d-r}\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')_\perp}\right\|_{\mathrm{op}} \\
&\le \left\|\Sigma^{\frac{1}{2}}\mathcal{P}_{\mathcal{V}(X)_\perp}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}(X-X')\right\|_{\mathrm{op}} + \left\|\Sigma^{\frac{1}{2}}\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X^\top - X'^\top)\right\|_{\mathrm{op}} \\
&= \left\|\left(\begin{pmatrix}0 & I_{d-r}\end{pmatrix}\mathcal{P}_{\mathcal{V}(X)}^\top\Sigma\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}\right)^{\frac{1}{2}}(X-X')\right\|_{\mathrm{op}} + \left\|\left(\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X)_\perp}^\top\Sigma\mathcal{P}_{\mathcal{V}(X)_\perp}\begin{pmatrix}I_r\\ 0\end{pmatrix}\right)^{\frac{1}{2}}(X^\top - X'^\top)\right\|_{\mathrm{op}}.
\end{aligned}$$

where the inequality follows from the triangle inequality and the facts that the spectral norms of $\mathcal{P}_{\mathcal{V}(X')}$, $\mathcal{P}_{\mathcal{V}(X')_\perp}$, and the two block–identity matrices are all at most 1 (the fact that spectral norms of projectors are at most 1 can be proved via the first inequality in Lemma 17); and the last equality is because for any matrices A, B we have

$$\|\Sigma^{\frac{1}{2}}AB\|_{\mathrm{op}} = \sqrt{\|B^\top A^\top \Sigma AB\|_{\mathrm{op}}} = \|(A^\top\Sigma A)^{\frac{1}{2}}B\|_{\mathrm{op}}.$$

□

We continue to present the following lemma, which implies that the projectors and the block-identity matrices in Lemma 24 only reduces the effective dimensions of the ellipsoidal map, and does not increase the eigenvalues (up to absolute constants).

Lemma 25 (Spectral domination under contractions) *Let $\Sigma \succeq 0$ be a $d\times d$ PSD matrix with ordered eigenvalues $\lambda_1(\Sigma) \ge \cdots \ge \lambda_d(\Sigma)$. Let $A \in \mathbb{R}^{d\times m}$ for some $m \le d$ and write $s := \|A\|_{\mathrm{op}}$. Denote by $\mu_1 \ge \cdots \ge \mu_m$ the eigenvalues of $A^\top\Sigma A$. Then, for every $k = 1, \ldots, m$,*

$$\mu_m \;\le\; s^2\,\lambda_m(\Sigma).$$

Proof of Lemma 25: By the Courant–Fischer–Weyl max-min characterization (see, e.g., [Wikipedia contributors, 2025b]), we have

$$\begin{aligned}
\lambda_k(A^\top\Sigma A) \;&=\; \min_{\substack{S\subset\mathbb{R}^d\\ \dim S = d-k+1}} \sup\{\|A^\top\Sigma^{\frac{1}{2}}x\|_2^2 :\; x\in S,\; \|x\|_2 = 1\} \\
&\le s^2\cdot \min_{\substack{S\subset\mathbb{R}^d\\ \dim S = d-k+1}} \sup\{\|\Sigma^{1/2}x\|_2 :\; x\in S,\; \|x\|_2 = 1\} \\
&= s^2\lambda_k(\Sigma).
\end{aligned}$$

□

D.6 Proof of the Main Result

From Lemma 22, to cover $\mathrm{Gr}(d,r)$ it suffices to cover the unit ball of $(d-r)\times r$ matrices under the ellipsoidal spectral metric. We are now ready to prove Lemma 3, our main result for ellipsoidal Grassmannian covering.

Proof of Lemma 3: Define $\mathcal{X} = \{X \in \mathbb{R}^{(d-r)\times r} : \|X\|_{\mathrm{op}} \le 1\}$ and $\bar{\mathcal{X}} = \{X \in \mathbb{R}^{(d-r)\times r} : \|X\|_{\mathrm{op}} \le 2\}$. By Lemma 22 (Pointwise Dimension Consequence of Finite Global Atlas), for $\mu = \mathrm{Unif}(\mathrm{Gr}(d,r))$, we have that for all $\mathcal{V} \in \mathrm{Gr}(d,r)$ and all $\varepsilon > 0$,

$$\log \frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))} \le C_1 r(d-r) + \sup_{X\in\mathcal{X}} \log \frac{1}{\mathrm{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X),\mathcal{V}(X')) \le \varepsilon\}}, \tag{81}$$

where $C_1 > 0$ is an absolute constant.

Define the $(d-r)\times(d-r)$ positive definite matrices $H_1(X)$ and the $r \times r$ positive definite matrix $H_2(X)$ as the following

$$\begin{aligned} H_1(X) &= \begin{pmatrix} 0 & I_{d-r} \end{pmatrix} \mathcal{P}_{\mathcal{V}(X)}^\top \Sigma \mathcal{P}_{\mathcal{V}(X)} \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix}, \\ H_2(X) &= \begin{pmatrix} I_r & 0 \end{pmatrix} \mathcal{P}_{\mathcal{V}(X)_\perp}^\top \Sigma \mathcal{P}_{\mathcal{V}(X)_\perp} \begin{pmatrix} I_r \\ 0 \end{pmatrix}. \end{aligned}$$

By Lemma 24 (Lipchitz of Graph Chart), we have that

$$\varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X),\mathcal{V}(X')) \le \|H_1(X)^{\frac{1}{2}}(X'-X)\|_{\mathrm{op}} + \|H_2(X)^{\frac{1}{2}}(X'-X)^\top\|_{\mathrm{op}}.$$

An technical step: ball inclusion via thresholding. Given a PSD matrix $H \in \mathbb{R}^{m\times m}$ and an eigenvalue threshold α, assume its eigendecomposition is $H = U\,\mathrm{diag}(\beta_1,\cdots,\beta_m)U^\top$, define the thresholding function T_α by

$$T_\alpha(H) = U\,\mathrm{diag}(\max\{\beta_1,\alpha\},\cdots,\max\{\beta_m,\alpha\})U^\top.$$

Clearly this function only increases the metric. We further define the following two ellipsoidal metrics:

$$\begin{aligned} \varrho_1^2(X,X') &= \|(X'-X)^\top \bar{H}_1(X)(X'-X)\|_{\mathrm{op}}, \quad \bar{H}_1(X) = T_{\varepsilon^2}(H_1(X)) \\ \varrho_2^2(X,X') &= \|(X'-X)\bar{H}_2(X)(X-X')^\top\|_{\mathrm{op}}, \quad \bar{H}_2(X) = T_{\varepsilon^2}(H_2(X)) \end{aligned}$$

We note that the two balls $B_{\varrho_1}(X,\varepsilon)$, $B_{\varrho_2}(X,\varepsilon)$ are contained in $\bar{\mathcal{X}}$, as we have applied the thresholding function to ensure this inclusion. For example, for the first ball, from

$$X' - X = \big(\bar{H}_1(X)\big)^{-1/2} \underbrace{\big(\bar{H}_1(X)\big)^{\frac{1}{2}}(X'-X)}_{\text{spectral norm} \le \varepsilon \text{ for } X' \in B_{\varrho_1}(X,\varepsilon)},$$

we have (by using the ε estimate from the second underbraced term above, and combining it with the thresholding guarantee $\lambda_{\min}(\bar{H}_1(X)) \ge \varepsilon^2$)

$$\|X'-X\|_{\mathrm{op}} \le \lambda_{\min}(\bar{H}_1(X))^{-1/2}\cdot\varepsilon \le 1,$$

which resulting in $\|X'\|_{\mathrm{op}} \le \|X'-X\|_{\mathrm{op}} + \|X\|_{\mathrm{op}} \le 2$ and thus $B_{\varrho_1}(X,\varepsilon) \subseteq \bar{\mathcal{X}}$. Similarly, we can show $B_{\varrho_2}(X,\varepsilon) \subseteq \bar{\mathcal{X}}$. this gives us the auxiliary ball-inclusion result:

$$B_{\varrho_1+\varrho_2}(X,\varepsilon) \subseteq B_{\varrho_1}(X,\varepsilon) \cap B_{\varrho_2}(X,\varepsilon) \subseteq B_{\varrho_1}(X,\varepsilon) \cup B_{\varrho_2}(X,\varepsilon) \subseteq \bar{\mathcal{X}}. \tag{82}$$

Now we are ready to proceed with the main part of the proof. By Lemma 24 (Lipchitz of Graph Chart) and the fact that threholding only increase the spectral norm, the ellipsoidal projection metric is bounded by $\varrho_1 + \varrho_2$, so for any $X \in \mathcal{X}$,

$$\begin{aligned}
&\log \frac{1}{\mathrm{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X), \mathcal{V}(X')) \le \varepsilon\}} \\
\le &\log \frac{1}{\mathrm{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_1(X, X') + \varrho_2(X, X') \le \varepsilon\}} \\
= &\log \frac{1}{\mathrm{Unif}(\bar{\mathcal{X}})\{B_{\varrho_1+\varrho_2}(X, \varepsilon)\}} \qquad (83)\\
= &\frac{\mathrm{Vol}(\bar{\mathcal{X}})}{\mathrm{Vol}(B_{\varrho_1+\varrho_2}(X, \varepsilon))}, \qquad (84)
\end{aligned}$$

where the first equality uses the ball-inclusion result (82).

Background on covering number. Classical volume-ratio arguments give the following results on the covering number of balls in general normed space $\mathcal{Y}$ for a p-dimensional normed space equipped with the metric associated to its norm $\|\cdot\|$, we denote by by $B(y, R)$ the ball in $\mathcal{Y}$ centered at $y \in \mathcal{Y}$ with radius R, and by $\mathsf{N}(\mathcal{Z}, \|\cdot\|, \varepsilon)$ the covering number of the $p-$dimensional set K. Formally, we give the definition of covering number as follows.

Definition 5 (Covering numbers) *Let $(\mathcal{Y}, \|\cdot\|)$ be a normed space and let $\mathcal{Z} \subseteq \mathcal{Y}$. For $\varepsilon > 0$, a set $\mathcal{N} \subseteq \mathcal{Z}$ is an* internal ε–cover *of $\mathcal{Z}$ if for every $z \in \mathcal{Z}$ there exists $y \in \mathcal{N} \subseteq \mathcal{Z}$ with $\|z - y\| \le \varepsilon$. The (internal) covering number is*

$$\mathsf{N}(\mathcal{Z}, \|\cdot\|, \varepsilon) := \min\{\, m : \ \exists \text{ internal } \varepsilon\text{–cover of } \mathcal{Z} \text{ with size } m \,\}.$$

A set $\mathcal{N}_{\mathrm{ext}} \subseteq \mathcal{Y}$ (not necessarily inside $\mathcal{Z}$) is an external ε–cover *of $\mathcal{Z}$ if for every $z \in \mathcal{Z}$ there exists $y \in \mathcal{N}_{\mathrm{ext}}$ with $\|z - y\| \le \varepsilon$. The external covering number is*

$$\mathsf{N}_{\mathrm{ext}}(\mathcal{Z}, \|\cdot\|, \varepsilon) := \min\{\, m : \ \exists \text{ external } \varepsilon\text{–cover of } \mathcal{Z} \text{ with size } m \,\}.$$

Internal covering numbers depend only on the metric induced on $\mathcal{Z}$, while external covering numbers also depend on the ambient space $\mathcal{Y}$. Throughout the paper, "covering number" means the internal one unless otherwise stated.

We now relate the internal and external covering numbers, showing they are equivalent up to a constant factor in the radius—and thus interchangeable for our purposes.

Lemma 26 (Properties of External Covering Number) *For every $\varepsilon > 0$ and $\mathcal{Z} \subseteq \mathcal{Y}$,*

$$\mathsf{N}_{\mathrm{ext}}(\mathcal{Z}, \|\cdot\|, \varepsilon) \ \le \ \mathsf{N}(\mathcal{Z}, \|\cdot\|, \varepsilon) \ \le \ \mathsf{N}_{\mathrm{ext}}(\mathcal{Z}, \|\cdot\|, \varepsilon/2). \qquad (85)$$

And the external covering number enjoys monotonicity under set inclusion: if $\mathcal{Z}_1 \subseteq \mathcal{Z}_2$ then $\mathsf{N}_{\mathrm{ext}}(\mathcal{Z}_1, \|\cdot\|, \varepsilon) \ \le \ \mathsf{N}_{\mathrm{ext}}(\mathcal{Z}_2, \|\cdot\|, \varepsilon)$.

Proof of Lemma 26: The left inequality in (85) is immediate since any internal cover is also an external cover. For the right inequality in (85), let $\{y_1, \ldots, y_m\} \subseteq \mathcal{Y}$ be an external $(\varepsilon/2)$–cover of $\mathcal{Z}$. For each i, define the (possibly empty) cell $V_i := \{z \in \mathcal{Z} : \ \|z - y_i\| \le \varepsilon/2\}$ and, if $V_i \ne \emptyset$, pick a representative $z_i \in V_i$. Then for any $z \in V_i$,

$$\|z - z_i\| \ \le \ \|z - y_i\| + \|y_i - z_i\| \ \le \ \varepsilon/2 + \varepsilon/2 \ = \ \varepsilon,$$

so the selected $\{z_i\} \subseteq \mathcal{Z}$ form an internal ε–cover. Hence $\mathsf{N}(\mathcal{Z}, \|\cdot\|, \varepsilon) \le m = \mathsf{N}_{\text{ext}}(\mathcal{Z}, \|\cdot\|, \varepsilon/2)$. Lastly, the monotonicity under set inclusion for the external covering number is a straightforward consequence of its definition.

□

Proposition 4.2.10 in Vershynin [2018] (the proof is elementary and clearly holds true for general metric in a normed space) states that for $\mathcal{Z} \subseteq \mathcal{Y}$ and general metric $\|\cdot\|$, we have that for any $y \in \mathcal{Y}$,

$$\frac{\mathrm{Vol}(\mathcal{Z})}{\mathrm{Vol}(B(y,\varepsilon))} \le \mathrm{N}(\mathcal{Z}, \|\cdot\|, \varepsilon) \le \frac{\mathrm{Vol}(\mathcal{Z} + B(y, \frac{\varepsilon}{2}))}{\mathrm{Vol}(B(y, \frac{\varepsilon}{2}))},$$

where the set $\mathcal{A} + \mathcal{B} := \{a + b : a \in \mathcal{A}, b \in \mathcal{B}\}$. When $\mathcal{Z}$ is convex and $B(y, \varepsilon) \subseteq \mathcal{Z}$, we further have

$$\frac{\mathrm{Vol}(\mathcal{Z})}{\mathrm{Vol}(B(y,\varepsilon))} \le \mathrm{N}(\mathcal{Z}, \|\cdot\|, \varepsilon) \le \frac{\mathrm{Vol}(\mathcal{Z} + B(y, \frac{\varepsilon}{2}))}{\mathrm{Vol}(B(y, \frac{\varepsilon}{2}))} \le \frac{\mathrm{Vol}(\frac{3}{2}\mathcal{Z})}{\mathrm{Vol}(B(y, \frac{\varepsilon}{2}))} = 3^p \frac{\mathrm{Vol}(\mathcal{Z})}{\mathrm{Vol}(B(y,\varepsilon))}, \tag{86}$$

where $\lambda\mathcal{A} := \{\lambda a : a \in \mathcal{A}\}$ for $\lambda > 0$. Lastly, when the normed space $\mathcal{Y}$ is p–dimenional, for every $\varepsilon \in (0, R]$, setting $\mathcal{Z} = B(0, R)$ turns the above inequality (86) into the optimal covering number bound

$$\left(\frac{R}{\varepsilon}\right)^p \ \le \ \mathrm{N}\big(B(0,R), \ \|\cdot\|, \ \varepsilon\big) \ \le \ \left(\frac{3R}{\varepsilon}\right)^p. \tag{87}$$

Note that this result is for general normed space, not only for the ℓ_2 norm in Euclidean space (see, e.g., display (1) in Pajor [1998]; see also Milman and Schechtman [1986], Pisier [1999]).

An technical step: lifting to product space. Consider the product space $\mathbb{R}^{(d-r)\times r} \times \mathbb{R}^{(d-r)\times r}$ (of dimension $2 \times (d-r) \times r$). Given any $(d-r) \times (d-r)$ positive definite matrix H_1 and $r \times r$ positive definite matrix H_2, define the modified spectral norm by

$$\|(X_1, X_2) - (X_1', X_2')\|_{\mathrm{op}, H_1, H_2} = \|H_1^{\frac{1}{2}}(X_1 - X_1')\|_{\mathrm{op}} + \|H_2^{\frac{1}{2}}(X_2^\top - {X_2'}^\top)\|_{\mathrm{op}}.$$

Consider the constrained set

$$\mathcal{S} := \{(X_1, X_2) \in \mathbb{R}^{(d-r)\times r} \times \mathbb{R}^{(d-r)\times r} : X_1 = X_2\} = \{(X, X) : X \in \mathbb{R}^{(d-r)\times r}\},$$

which is a normed space with dimension $(d-r) \times r$ (isomorphic to $\mathbb{R}^{(d-r)\times r}$), equipped with the modifed spectral norm

$$\|(X, X) - (X', X')\|_{\mathrm{op}, H_1, H_2} = \|H_1^{\frac{1}{2}}(X - X')\|_{\mathrm{op}} + \|H_2^{\frac{1}{2}}(X^\top - {X'}^\top)\|_{\mathrm{op}}.$$

Denote $B^{\mathcal{S}}_{\mathrm{op},H_1,H_2}((X,X),R) = \{(X',X') \in \mathcal{S} : \|(X',X') - (X,X)\|_{\mathrm{op},H_1,H_2} \le R\}$ (the ball constrained in $\mathcal{S}$). Because there is a bijective, distance-preserving map between $B_{\varrho_1+\varrho_2}(X,\varepsilon)$ and $B^{\mathcal{S}}_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon)$, and likewise $B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4)$ and $\bar{\mathcal{X}}$ (here 0 denotes the $(d-r)\times r$ 0 matrix), we obtain

$$\frac{\mathrm{Vol}(\bar{\mathcal{X}})}{\mathrm{Vol}(B_{\varrho_1+\varrho_2}(X,\varepsilon))} = \frac{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))}, \tag{88}$$

where the volume on $\mathcal{S}$ is defined via the surface area measure. (88) is exactly the objective we need to bound in (83).

Given $\varepsilon > 0$, by the property (86) of covering number, we have that for every $X \in \mathcal{X}$ and $\varepsilon > 0$,

$$\frac{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))} \le \mathrm{N}\big(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon\big). \tag{89}$$

How lifting to product space double the degree of freedom. We now lift the $\mathcal{S}-$constrained ball $B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4)$ to the product space $\bar{\mathcal{X}}\times\bar{\mathcal{X}}$, using the covering number of the lifted product space to bound the covering number of the original space, in order to obtain an upper bound on (89) and (88). This is the reason why our final bound will scale (in the isotropic case) in the order $O((d-r)r\log\frac{1}{\varepsilon^2}) = O(2(d-r)r\log\frac{1}{\varepsilon})$ rather than the classical optimal order $\Theta((d-r)r\log\frac{1}{\varepsilon})$—the lifting to product space increase the number of freedom by a multiplicative factor of 2. Nevertheless, such difference is negligible in our theory.

For every $(X_1,X_2) \in \mathbb{R}^{(d-r)\times r} \times \mathbb{R}^{(d-r)\times r}$, every $(d-r)\times(d-r)$ matrix $H_1 \succ 0$, and every $r\times r$ matrix $H_2 \succ 0$, and radius R, denote $B_{\mathrm{op},H_1,H_2}((X_1,X_2),R)$ to be the unconstrained ball in $\mathbb{R}^{(d-r)\times r}\times\mathbb{R}^{(d-r)\times r}$:

$$B_{\mathrm{op},H_1,H_2}((X_1,X_2),R) := \{(X_1',X_2') \in \mathbb{R}^{(d-r)\times r}\times\mathbb{R}^{(d-r)\times r} : \|(X_1,X_2) - (X_1' - X_2')\|_{\mathrm{op},H_1,H_2} \le R\}.$$

Lifting to the product space can only increase the external covering number (monotonicity under set inclusion), and the external covering number is equivalent to the internal covering number up to a constant factor in the radius. To be specific, by Lemma 26, we have

$$\begin{aligned}
&\mathrm{N}\big(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon\big)\\
\le&\mathrm{N}_{\mathrm{ext}}\big(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big)\\
\le&\mathrm{N}_{\mathrm{ext}}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big)\\
\le&\mathrm{N}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big).
\end{aligned} \tag{90}$$

For every $X \in \mathcal{X}$, the ball-inclusion argument (82) is strong enough to imply that the unconstrained ball $B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon) \subseteq \mathbb{R}^{(d-r)\times r}\times\mathbb{R}^{(d-r)\times r}$ is also included in the lifted ball $B_{\mathrm{op},I_{d-r},I_r}((0,0),4)$, which gives that

$$B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon/2) \subset B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon) \subseteq B_{\mathrm{op},I_{d-r},I_r}((0,0),4).$$

This satisfies the inclusion condition required to establish (86), and we have

$$\begin{aligned}
\mathrm{N}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big) \le& 3^{2(d-r)r}\frac{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon/2))}\\
=&6^{2(d-r)r}\frac{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))}
\end{aligned} \tag{91}$$

Applying change of variable and calculating the Jacobian determinant. Applying the standard change of variables

$$Y_1 = \bar{H}_1(X)^{1/2} X_1, \qquad Y_2 = X_2\, \bar{H}_2(X)^{1/2},$$

the map on vectorized variables is

$$\operatorname{vec}(Y_1) = (I_r \otimes \bar{H}_1(X)^{1/2}) \operatorname{vec}(X_1), \qquad \operatorname{vec}(Y_2) = (\bar{H}_2(X)^{\top 1/2} \otimes I_{d-r}) \operatorname{vec}(X_2),$$

and the total Jacobian is

$$J(X) = \begin{pmatrix} I_r \otimes \bar{H}_1(X)^{1/2} & 0 \\ 0 & \bar{H}_2(X)^{\top 1/2} \otimes I_{d-r}). \end{pmatrix}$$

The two block–diagonal Jacobian determinants are

$$\left|\det\big(I_r \otimes \bar{H}_1(X)^{1/2}\big)\right| = \big(\det \bar{H}_1(X)^{1/2}\big)^{r} = \det\!\big(\bar{H}_1(X)\big)^{r/2},$$
$$\left|\det\big(\bar{H}_2(X)^{\top 1/2} \otimes I_{d-r}\big)\right| = \big(\det \bar{H}_2(X)^{1/2}\big)^{d-r} = \det\!\big(\bar{H}_2(X)\big)^{(d-r)/2}.$$

Multiplying the two factors, the total Jacobian of the linear change of variables is

$$\det(J(X)) = \det\!\big(\bar{H}_1(X)\big)^{r/2} \det\!\big(\bar{H}_2(X)\big)^{(d-r)/2}.$$

(We used $\det(B^\top) = \det(B)$ and that $\bar{H}_1(X), \bar{H}_2(X) \succ 0$, so determinants are positive.) By the change of variable formula in integration (see, e.g., Wikipedia contributors [2025a]), we have

$$\begin{aligned} &\operatorname{Vol}\big(B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon)\big) \\ =\ &\operatorname{Vol}\big(B_{\mathrm{op},I_{d-1},I_r}((X,X),\varepsilon)\big)\, (\det(J(X)))^{-1} \\ =\ &\operatorname{Vol}\big(B_{\mathrm{op},I_{d-1},I_r}((X,X),\varepsilon)\big) \prod_{k=1}^{d-r} \lambda_k(\bar{H}_1(X))^{-r/2} \prod_{k=1}^{r} \lambda_k(\bar{H}_2(X))^{-(d-r)/2}, \end{aligned}$$

which implies

$$\frac{\operatorname{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\operatorname{Vol}(B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))} = \prod_{k=1}^{d-r} \lambda_k(\bar{H}_1(X))^{r/2} \prod_{k=1}^{r} \lambda_k(\bar{H}_1(X))^{(d-r)/2} \frac{\operatorname{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\operatorname{Vol}(B_{\mathrm{op},I_{d-r},I_r}((X,X),\varepsilon))}, \tag{92}$$

Proving the final bound. For all $X \in \mathcal{X}$ and $\varepsilon \leq 1$, we have that $B_{\mathrm{op},I_{d-r},I_r}((X,X),\varepsilon) \subseteq B_{\mathrm{op},I_{d-r},I_r}((0,0),4)$ and thus by (86) and (87), we have

$$\frac{\operatorname{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\operatorname{Vol}(B_{\mathrm{op},I_{d-r},I_r}((X,X),\varepsilon))} \leq \left(\frac{12}{\varepsilon}\right)^{2(d-r)r}. \tag{93}$$

Combining the above inequality (93) with (91) and (92), we have

$$\begin{aligned} &\log \mathrm{N}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},H_1,H_2},\ \varepsilon\big) \\ \leq& 2(d-r)r \log\frac{72}{\varepsilon} + \frac{r}{2}\sum_{k=1}^{d-r} \log \lambda_k(\bar{H}_1(X)) + \frac{d-r}{2}\sum_{k=1}^{r} \log \lambda_k(\bar{H}_2(X)) \\ =& \frac{r}{2}\sum_{k=1}^{d-r} \log\frac{72^2 \lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r} \log\frac{72^2 \lambda_k(\bar{H}_2(X))}{\varepsilon^2}. \end{aligned} \tag{94}$$

Combing the above inequality (94) with (88), (89) and (90), we have that for all $X \in \mathcal{X}$,

$$\log \frac{\mathrm{Vol}(\bar{\mathcal{X}})}{\mathrm{Vol}(B_{\varrho_1+\varrho_2}(X,\varepsilon))} \leq \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{72^2\lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{72^2\lambda_k(\bar{H}_2(X))}{\varepsilon^2}. \tag{95}$$

Finally, combine the above inequality (95) with (81) and (83), we prove that for $\mu = \mathrm{Unif}(\mathrm{Gr}(d,r))$, we have that for all $\mathcal{V} \in \mathrm{Gr}(d,r)$ and all $\varepsilon > 0$,

$$\begin{aligned}\log \frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))} \leq & C_1 r(d-r) + \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{72^2\lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{72^2\lambda_k(\bar{H}_2(X))}{\varepsilon^2} \\ = & \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{C\lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{C\lambda_k(\bar{H}_2(X))}{\varepsilon^2},\end{aligned} \tag{96}$$

where $C > 0$ is an absolute constant.

We end the proof by applying Lemma 25 and Lemma 17: since

$$\begin{aligned}&\lambda_k(H_1(X)) \leq \lambda_k(\mathcal{P}_{\mathcal{V}(X)}^\top \Sigma \mathcal{P}_{\mathcal{V}(X)}) \leq \lambda_k, \quad k = 1, \cdots, d-r;\\ &\lambda_k(H_2(X)) \leq \lambda_k(\mathcal{P}_{\mathcal{V}(X)_\perp}^\top \Sigma \mathcal{P}_{\mathcal{V}(X)_\perp}) \leq \lambda_k, \quad k = 1, \cdots, r,\end{aligned}$$

we have

$$\begin{aligned}&\lambda_k(\bar{H}_1(X)) \leq \max\{\lambda_k, \varepsilon^2\}, \quad k = 1, \cdots, d-r;\\ &\lambda_k(\bar{H}_2(X)) \leq \max\{\lambda_k, \varepsilon^2\}, \quad k = 1, \cdots, r,\end{aligned}$$

Substituting this bound to (96), we prove that for $\mu = \mathrm{Unif}(\mathrm{Gr}(d,r))$, we have that for all $\mathcal{V} \in \mathrm{Gr}(d,r)$ and all $\varepsilon > 0$,

$$\log \frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))} \leq \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{C\max\left\{\lambda_k,\varepsilon^2\right\}}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{C\max\left\{\lambda_k,\varepsilon^2\right\}}{\varepsilon^2},$$

where $C > 0$ is an absolute constant.

□

E Proofs for Generalization Bounds and Comparison (Section 4)

E.1 Proof of Theorem 4 in Section 4.1

The proof consists of two steps: 1. Obtaining the Integral Bound on Generalization Gap; and 2. Obtaining the Expression of Riemannian Dimension.

Step 1: Obtaining the Integral Bound on Generalization Gap. As presented in (7), we construct the metric tensor

$$G_{\mathrm{NP}}(W) := \mathrm{blockdiag}\left(\cdots, LM_{l\to L}^2(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}^\top(W,X)\otimes I_{d_l}, \cdots\right).$$

By Lipchitz property of the loss function we have

$$\begin{aligned}\varrho_{n,\ell}(W', W) &= \sqrt{\mathbb{P}_n(\ell(W'; (x, y)) - \ell(W; (x, y)))^2} \\ &\leq \beta\sqrt{\mathbb{P}_n\|f(W', x) - f(W, x)\|_2^2} = \beta\varrho_n(W', W)\end{aligned}$$

By Lemma 1 we have the metric dominating relationship: for every $W \in B_{\mathsf{F}}(R)$,

$$\sqrt{n}\varrho_n(W', W) \leq \varrho_{G_{\mathrm{NP}}(W)}(W', W), \quad \forall W' \in B_{\mathsf{F}}(R).$$

Combining the above two inequalities we have

$$\varrho_{n,\ell}(W', W) \leq \frac{\beta}{\sqrt{n}}\varrho_{G_{\mathrm{NP}}(W)}(W', W), \quad \forall W' \in B_{\mathsf{F}}(R).$$

By the metric domination lemma (Lemma 13), we have the pointwise dimension bound: for every $W \in B_{\mathsf{F}}(R)$,

$$\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(W, \varepsilon))} \leq \log\frac{1}{\pi(B_{G_{\mathrm{NP}}(W)}(W, \sqrt{n}\varepsilon/\beta))},$$

By Lemma 19 (Riemannian Dimension Bound for DNN), we have that there exists a prior π such that uniformly over every $W \in B_{\mathsf{F}}(R)$,

$$\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(W, \varepsilon))} \leq \log\frac{1}{\pi(B_{G_{\mathrm{NP}}(W)}(W, \sqrt{n}\varepsilon/\beta))} \leq d_{\mathrm{R}}(W, \varepsilon/\beta), \tag{97}$$

where the definition of Riemannian Dimension d_{R} can be found in Lemma 19. By Theorem 2, we have that there exists an absolute constant C_1 such that with probability at least $1-\delta$, uniformly over all $W \in B_{\mathsf{F}}(R)$,

$$\begin{aligned}(\mathbb{P} - \mathbb{P}_n)\ell(f(W, x), y) \leq& C_1\left(\frac{1}{\sqrt{n}}\int_0^\infty \sqrt{\log\left(\frac{1}{\pi(B_{\varrho_{n,\ell}}(W, \varepsilon))}\right)}d\varepsilon + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right) \\ \leq& C_1\left(\frac{1}{\sqrt{n}}\int_0^\infty \sqrt{d_{\mathrm{R}}(W, \varepsilon/\beta)}d\varepsilon + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right) \\ =& C_1\left(\frac{\beta}{\sqrt{n}}\int_0^\infty \sqrt{d_{\mathrm{R}}(W, \varepsilon)}d\varepsilon + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right).\end{aligned} \tag{98}$$

where C_1 is an absolute constant; the first inequality uses Theorem 2; and the second inequality uses (97). This finishes the first part of Theorem 4 (integral upper bound).

Step 2: Obtaining the Expression of Riemannian Dimension. It remains to express the Riemannian Dimension d_{R} by Theorem 3 and prove the second part of Theorem 4. By Theorem 3, we have that the expression of Riemannian Dimension is

$$\begin{aligned}d_{\mathrm{R}}(W, \varepsilon) =& \sum_{l=1}^{L}\Big((d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(LM_{l\to L}^2(W, \varepsilon) \cdot F_{l-1}(W, X)F_{l-1}(W, X)^\top, C_2\max\{\|W\|_{\mathsf{F}}, R/2^n\}, \varepsilon) \\ &+ \log(d_{l-1}n)\Big),\end{aligned} \tag{99}$$

where $R = \sup_{\mathcal{W}} \|W\|_{\mathsf{F}}$, C_2 is an absolute constant, and the effective dimension (defined via (9)) is

$$\begin{aligned} & d_{\text{eff}}(LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}(W,X)^\top, C_2\max\{\|W\|_{\mathsf{F}}, R/2^n\},\varepsilon) \\ & \quad = \frac{1}{2}\sum_{k=1}^{r_{\text{eff}}[W,l]} \log\frac{8C_2^2\max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)\lambda_k(F_{l-1}F_{l-1}^\top)}{n\varepsilon^2}, \end{aligned} \tag{100}$$

where F_{l-1} is the abbreviation of $F_{l-1}(W,X)$ and $r_{\text{eff}}[W,l]$ is the abbreviation of $r_{\text{eff}}(LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}(W,X)^\top, C_2\max\{\|W\|_{\mathsf{F}}, R/2^n\},\varepsilon)$.

Combining the identities (99) and (100), we have the pointwise dimension bound

$$\begin{aligned} & d_{\mathrm{R}}(W,\varepsilon) \\ = & \sum_{l=1}^{L}\left((d_l+d_{l-1})\sum_{k=1}^{r_{\text{eff}}[W,l]}\log\frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\varepsilon^2}+\log(d_{l-1}n)\right) \\ = & \sum_{l=1}^{L}\Bigg((d_l+d_{l-1})\sum_{k=1}^{r_{\text{eff}}[W,l]}\log\frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)}{n\varepsilon^2} \\ & \qquad +(d_l+d_{l-1})r_{\text{eff}}[W,l]\cdot\log\left(M^2_{l\to L}(W,\varepsilon)L\max\{||W||_{\mathsf{F}}^2, R^2/4^n\}\right)+\log(d_{l-1}n)\Bigg) \end{aligned} \tag{101}$$

where F_{l-1} is the abbreviation of $F_{l-1}(W,X)$; $r_{\text{eff}}[W,l]$ is the abbreviation of $r_{\text{eff}}(LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}(W,X)^\top, C_2\max\{\|W\|_{\mathsf{F}}, R/2^n\},\varepsilon)$; and C_2 is an absolute constant.

This finishes the second part of Theorem 4 (expression of Riemannian Dimension).

Combining the integral upper bound (98) and the Riemannain dimension expression (101) concludes the proof of Theorem 4.

□

E.2 Proof for Regularized ERM in Section 4.2

Lemma 27 (Excess Risk Bound for Regularized ERM) *Assume we have high-probability pointwise generalization bound in the form of (3), and the loss $\ell(f;z)$ is uniformly bounded by $[0,1]$. Then for the regularized ERM*

$$\hat{f} = \text{argmin}_f \left\{\mathbb{P}_n\ell(f;z) + C\sqrt{\frac{d(f)+\log(2/\delta)}{n}}\right\},$$

we have the excess risk bound against the population risk minimizer $f^\star := \arg\min_{\mathcal{F}}\mathbb{P}\ell(f;z)$: with probability at least $1-\delta$,

$$\begin{aligned} \mathbb{P}\ell(\hat{f};z) - \mathbb{P}\ell(f^\star;z) \le & \inf_{f\in\mathcal{F}}\left\{\mathbb{P}_n\ell(f;z) + C\sqrt{\frac{d(f)+\log(2/\delta)}{n}}\right\} - \mathbb{P}\ell(f^\star;z) \\ \le & (C+\sqrt{1/2})\sqrt{\frac{d(f^\star)+\log(2/\delta)}{n}}. \end{aligned}$$

Proof of Lemma 27: by (3), for every $\delta \in (0,1)$, take $\delta_1 = \delta_2 = \delta/2$, we have that with probability at least $1-\delta_1-\delta_2 = 1-\delta$, we have

$$
\begin{aligned}
\mathbb{P}\ell(\hat{f};z) \leq \inf_{f\in\mathcal{F}} & \left\{\mathbb{P}_n\ell(f;z) + C\sqrt{\frac{d(f)+\log(1/\delta_1)}{n}}\right\} \\
\leq & \mathbb{P}_n\ell(f^\star;z) + C\sqrt{\frac{d(f^\star)+\log(1/\delta_1)}{n}} \\
\leq & \mathbb{P}\ell(f^\star;z) + \sqrt{\frac{\log(1/\delta_2)}{2n}} + C\sqrt{\frac{d(f^\star)+\log(1/\delta_1)}{n}} \\
= & \mathbb{P}\ell(f^\star;z) + \sqrt{\frac{\log(2/\delta)}{2n}} + C\sqrt{\frac{d(f^\star)+\log(2/\delta)}{n}} \\
\leq & \mathbb{P}\ell(f^\star;z) + (C+\sqrt{1/2})\sqrt{\frac{d(f^\star)+\log(2/\delta)}{n}}.
\end{aligned}
$$

where the first inequality uses the bound of the form (3); the second inequality uses definition of $\hat{f}$; and the third inequality is an application of the Mcdiarmid inequality (Lemma 10) at $f^\star$; the equality is by $\delta_1 = \delta_2 = \delta/2$; and the last inequality is a straightforward implicaition of applying Jensen's inequality to the square root function. Thus we have that the excess risk is bounded by

$$
\begin{aligned}
\mathbb{P}\ell(\hat{f};z) - \mathbb{P}\ell(f^\star;z) \leq \inf_{f\in\mathcal{F}} & \left\{\mathbb{P}_n\ell(f;z) + C\sqrt{\frac{d(f)+\log(2/\delta)}{n}}\right\} - \mathbb{P}\ell(f^\star;z) \\
\leq & (C+\sqrt{1/2})\sqrt{\frac{d(f^\star)+\log(2/\delta)}{n}}
\end{aligned}
$$

□

E.3 Improvement over Norm Bounds in Section 4.3

E.3.1 Exponential Relaxation to a Norm Bound and Comparison

We now provide norm-constrained bound from Theorem 4 without any expression r_{eff} and d_{eff} in the bound. Invoking the elementary bound $\log x \leq \log(1+x) \leq x$, the effective dimension factor in Theorem 4 can be relaxed to the dimension-independent bound

$$
\begin{aligned}
\sum_{k=1}^{\infty} \log\Bigl(\tfrac{\lambda_k\left(F_{l-1}F_{l-1}^\top\right)\,\|W\|_{\mathsf{F}}^2\,L\,M_{l\to L}^2(W,\varepsilon)}{n\,\varepsilon^2}\Bigr) \leq & \frac{\sum_{k=1}^{\infty}\lambda_k(F_{l-1}F_{l-1}^\top)\,\|W\|_{\mathsf{F}}^2\,L\,M_{l\to L}^2(W,\varepsilon)}{n\,\varepsilon^2} \\
\leq & \frac{\|F_{l-1}(W,X)\|_{\mathsf{F}}^2\,\|W\|_{\mathsf{F}}^2\,L\,M_{l\to L}^2(W,\varepsilon)}{n\,\varepsilon^2},
\end{aligned}
$$

and one arrives at the following rank–free consequence.

Corollary 1 (Norm-constrained bound) *Theorem 4 is never worse than: uniformly over all $W \in B_{\mathsf{F}}(R)$, the generalization gap $(\mathbb{P}-\mathbb{P}_n)\,\ell\bigl(f(W,x),y\bigr)$ is bounded by*

$$
O\left(\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}(W,X)\|_{\mathsf{F}}^2\|W\|_{\mathsf{F}}^2\sup_{\varepsilon>0}M_{l\to L}^2(W,\varepsilon)}}{n} + \sqrt{\frac{\beta^2\sum_{l=1}^{L}\log(d_{l-1}n)+\log\frac{\log(2n)}{\delta}}{n}}\right). \tag{102}
$$

Furthermore, (102) *implies the spectrally normalized bound: uniformly over* $W \in B_{\mathsf{F}}(R)$*, the generalization gap* $(\mathbb{P} - \mathbb{P}_n)\,\ell\big(f(W,x),y\big)$ *is bounded by*

$$O\left(\frac{\beta\|X\|_{\mathsf{F}}\|W\|_{\mathsf{F}} \cdot \sqrt{\sum_{l=1}^{L} L(d_l + d_{l-1})\ \prod_{i\neq l}\|W_i\|_{\mathrm{op}}^2}}{n} + \sqrt{\frac{\beta^2 \sum_{l=1}^{L}\log(d_{l-1}n) + L\log\frac{n\log\max\{R,2\}}{\delta}}{n}}\right). \tag{103}$$

Here in both (102) *and* (103)*,* O *hides multiplicative absolute constants and two ignorable high-order terms:* $\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})d_{l-1}}}{n^{5.5}}$ *and* $\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2 R^2 \sup_{\varepsilon>0} M_{l\to L^2(W,\varepsilon)}}}{n2^n}$*; and in* (103)*,* O *additionally hides an ignorable high-order term* $\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\sum_{l=1}^{L}(d_l+d_{l-1})(R/\sqrt{L-1})^{L-1}}}{n\max\{R,2\}^n}$.

Note that (102) and (103), the feature matrices $F_{l-1}(W;X)$ and X contain n features vectors so their Frobenius norms scales with $\sqrt{n}$, making the order of both bounds to be $n^{-1/2}$.

Discussion of Corollary 1: We proceed in three paragraphs of discussion. First, we show that the Riemannian Dimension bound in Theorem 4 is *exponentially* tighter than the spectrally normalized bound in (103). Second, we offer a metric–tensor interpretation that clarifies the source of this improvement. Finally, we position (103) relative to the most representative spectrally normalized bounds (SNB) in the existing literature.

I: Why the improvement is exponential. Empirically one observes

$$\|F_{l-1}\|_{\mathsf{F}} \ll \prod_{i<l}\|W_i\|_{\mathrm{op}}\,\|X\|_{\mathsf{F}}, \qquad M_{l\to L}(W,\varepsilon) \le \sup_{W'\in B_{\varrho n}(W,\varepsilon)}\prod_{i>l}\|W_i'\|_{\mathrm{op}}.$$

Combining this dramatic improvement with the *already–exponential* gain that comes *solely* from the elementary inequality $\log x \le \log(1+x) \le x$ (for $x \ge 0$), we conclude that Theorem 4 is *exponentially tighter* than (103). Therefore, Theorem 4 improves on Corollary 1 by an exponential factor.

II: Metric tensor interpretation. For understand the improvement deeper, we highlight that the spectral norm bound (103) can be equivalently viewed as replacing the metric tensor G_{NP} (7) used in Theorem 4 by the diagonal metric tensor

$$G_{\mathrm{SNB}}(W) = \mathrm{blockdiag}\Big(\dots, L\sup_{W'\in B_{\mathsf{F}}(R)}\prod_{k\neq l}\|W_k'\|_{\mathrm{op}}\,\|X\|_{\mathsf{F}}^2 \otimes I_{d_l\times d_{l-1}}, \dots\Big),$$

which is a far coarser relaxation that completely discards the learned feature $F_l(W,X)$.

III: Relation to existing spectrally normalized bounds. The bound in (103) is structurally close to the classical SNB results of Bartlett et al. [2017] and Neyshabur et al. [2018]; the three bounds differ only in the *global ball* used to constraint the hypothesis class.

(a) Our bound (103) controls *all* layers simultaneously via the global Frobenius norm $\|W\|_{\mathsf{F}}$, hence the factor $\|W\|_{\mathsf{F}}$ in the numerator.

(b) Neyshabur et al. [2018] bounds each layer l separately by its Frobenius norm $\|W_l\|_{\mathsf{F}}$. Strengthening their argument with Dudley's entropy integral (one-shot optimization in the original paper) gives

$$(\mathbb{P}-\mathbb{P}_n)\,\ell\big(f(W,x),y\big) \;\le\; \tilde{O}\Big(\frac{\beta\,\|X\|_{\mathsf{F}}\,\sqrt{\sum_{l=1}^{L} L^2(d_l+d_{l-1})\,\|W_l\|_{\mathsf{F}}^2\,\prod_{i\neq l}\|W_i\|_{\mathrm{op}}^2}}{n}+\sqrt{\frac{\log\frac{1}{\delta}}{n}}\Big). \tag{104}$$

Neither (103) nor (104) strictly dominates the other, since factors of the form $(\sum_l a_l)(\sum_l b_l)$ in (103) *vs.* factors of the form $L\sum_l a_l b_l$ in (104) can swap their relative order.

(c) Bartlett et al. [2017] replaces each Frobenius norm by the $\|\cdot\|_{2,1}$ norm, obtaining the tighter

$$(\mathbb{P}-\mathbb{P}_n)\,\ell\big(f(W,x),y\big) \;\le\; \tilde{O}\Big(\frac{\beta\,\|X\|_{\mathsf{F}}\,\big(\sum_l\|W_l\|_{2,1}^{2/3}\,\sum_l\big(\prod_{i\neq l}\|W_i\|_{\mathrm{op}}\big)^{2/3}\big)^{3/2}}{n}+\sqrt{\frac{\log\frac{1}{\delta}}{n}}\Big), \tag{105}$$

which improves on (103) and (104) thanks to the sharper $2,1$ norm. Extending our Riemannian-dimension analysis to the $2,1$ norm setting is an interesting direction for future work.

(d) Size-independent SNB bounds (pioneered by Golowich et al. [2020]) remove all depth/width dependence at the price of a worse scaling in n; incorporating their technique is left for future research.

In any case, (103) is a representative SNB bound, and the key message in this subsection is that our Riemannian-Dimension result in Theorem 4 is *exponentially* sharper than (103).

E.3.2 Proof of Corollary 1

Riemannain Dimension can be expressed in the following equivalent form

$$\int_0^\infty \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon = \inf_{\alpha>0}\left(\int_0^\alpha \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon + \int_\alpha^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon\right).$$

We organize the proof with four steps.

Step 1: Bounding the Dominating Integral. As we will take α to be very small so that the $\int_0^\alpha \sqrt{d_{\mathrm{R}}(W,\varepsilon)}$ will be not exceed the order of $\int_\alpha^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)}$, we firstly prove $\int_\alpha^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon$. By the basic inequality $\log x \le \log(1+x) \le x$ for $x>0$, we have

$$\begin{aligned}
&\sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \log\Big(\frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\Big)\\
\le&\sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\\
\le&\sum_{k=1}^{d_{l-1}} \frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\\
=&\frac{8C_2^2\|F_{l-1}\|_{\mathsf{F}}^2\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2},
\end{aligned} \tag{106}$$

where F_{l-1} is the abbreviation of $F_{l-1}(W,X)$; $r_{\text{eff}}[W,l]$ is the abbreviation of $r_{\text{eff}}(LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}(W,X)^\top, C_2\max\{\|W\|_{\mathsf{F}}, R/2^n\},\varepsilon)$; and C_2 is a positive absolute constant. Here the second inequality uses the definition that $r_{\text{eff}}[W,l]$ as the effective rank of a $d_{l-1}\times d_{l-1}$ matrix, is no larger than the matrix width d_{l-1}; the first equality is because

$$\sum_{k=1}^{d_{l-1}} \lambda_k(F_{l-1}F_{l-1}^\top) = \mathrm{Tr}(F_{l-1}F_{l-1}^\top) = \|F_{l-1}\|_{\mathsf{F}}^2, \tag{107}$$

a well-known property of the Frobenius norm (the squared Frobenius norm $\|F_{l-1}\|_{\mathsf{F}}^2$ equals trace of $F_{l-1}F_{l-1}^\top$). By (106) and Theorem 4 we have the Riemannian Dimension upper bound

$$d_{\mathrm{R}}(W,\varepsilon) \le 8C_2^2\sum_{l=1}^{L}(d_l+d_{l-1})\frac{\|F_{l-1}\|_{\mathsf{F}}^2\max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\varepsilon^2} + \sum_{l=1}^{L}\log(d_{l-1}n), \tag{108}$$

where C_2 is a positive absolute constant.

Taking (108) to the integral $\int_\alpha^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon$, we have

$$\begin{aligned}
&\int_\alpha^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon \\
\le& 2\sqrt{2}C_2\int_\alpha^1\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})\frac{\|F_{l-1}\|_{\mathsf{F}}^2\max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\varepsilon^2}}d\varepsilon + (1-\alpha)\sqrt{\sum_{l=1}^{L}\log(d_{l-1}n)} \\
\le& C_3\sqrt{\frac{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2\max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}\sup_{\varepsilon>0}M^2_{l\to L}(W,\varepsilon)}{n}}\log\frac{1}{\alpha} + (1-\alpha)\sqrt{\sum_{l=1}^{L}\log(d_{l-1}n)},
\end{aligned}$$

where $C_3>0$ is an absolute constant.

Step 2: Bounding the Rest Integral. We then prove $\int_0^\alpha \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon$. Again, by the basic inequality $\log x \le \log(1+x)\le x$ for $x>0$, we have

$$\begin{aligned}
&\sum_{k=1}^{r_{\text{eff}}[W,l]}\log\Big(\frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_F^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\varepsilon^2}\Big) \\
\le&\sum_{k=1}^{d_{l-1}}\log\Big(\frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_F^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\varepsilon^2}\Big) \\
=&\sum_{k=1}^{d_{l-1}}\log\Big(\frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_F^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\alpha^2}\Big) + d_{l-1}\log\frac{\alpha^2}{\varepsilon^2} \\
\le&\frac{8C_2^2\sum_{k=1}^{d_{l-1}}\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_F^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\alpha^2} + d_{l-1}\log\frac{\alpha^2}{\varepsilon^2} \\
=&\frac{8C_2^2\|F_{l-1}(W,X)\|_{\mathsf{F}}^2\cdot\max\{||W||_F^2, R^2/4^n\}LM^2_{l\to L}(W,\varepsilon)}{n\alpha^2} + d_{l-1}\log\frac{\alpha^2}{\varepsilon^2}.
\end{aligned}\tag{109}$$

Taking (109) to the integral $\int_0^\alpha \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon$, we have

$$
\begin{aligned}
&\int_0^\alpha \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon \\
\le& 2\sqrt{2}C_2 \int_0^\alpha \sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})\frac{\|F_{l-1}\|_{\mathsf{F}}^2 \max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\} L M_{l\to L}^2(W,\varepsilon)}{n\alpha^2}}d\varepsilon + \int_0^\alpha \sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})d_{l-1}\log\frac{\alpha^2}{\varepsilon^2}}d\varepsilon \\
\le& C_4\left(\sqrt{\frac{\sum_{l=1}^{L}(d_l+d_{l-1})\|F_{l-1}\|_{\mathsf{F}}^2 \max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\} L \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}{n}} + \alpha\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})d_{l-1}}\right),
\end{aligned}
$$

where the second inequality holds by calculating the integral $\int_0^\alpha \sqrt{\log\left(\frac{\alpha^2}{\varepsilon^2}\right)}d\varepsilon = \alpha\sqrt{\frac{\pi}{2}}$, and $C_4 > 0$ is an absolute constant. Taking $\alpha = \frac{1}{n^5}$, the high-order term $\alpha\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})d_{l-1}}$ will be $\frac{\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})d_{l-1}}}{n^5}$ and is ignorable.

Step 3: Combing the Two Integrals. Combining Step 1 and Step 2, we get the full Riemannian Dimension integral upper bound

$$
\frac{1}{\sqrt{n}}\int_0^\infty \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon \le O\left(\frac{\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2\|W\|_{\mathsf{F}}^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n} + \sqrt{\frac{\sum_{l=1}^{L}\log(d_{l-1}n)}{n}}\right),
$$

where O hides multiplicative absolute constants and two ignorable high-order terms: $\frac{\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})d_{l-1}}}{n^{5.5}}$ and $\frac{\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2 R^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n2^n}$.

Put this bound into Theorem 4 (or (98) in its proof), we have with probability at least $1-\delta$, uniformly over all $W \in B_{\mathsf{F}}(R)$,

$$
\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \\
\le& O\left(\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2\|W\|_{\mathsf{F}}^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n} + \sqrt{\frac{\beta^2\sum_{l=1}^{L}\log(d_{l-1}n) + \log\frac{\log(2n)}{\delta}}{n}}\right),
\end{aligned}
\tag{110}
$$

where O hides multiplicative absolute constants and two ignorable high-order terms: $\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})d_{l-1}}}{n^{5.5}}$ and $\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2 R^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n2^n}$. Note that here $F_{l-1}(W;X) \in R^{d_{l-1}\times n}$ contains n features vectors in dimension d_{l-1} so its Frobenius norm $\|F_{l-1}\|_{\mathsf{F}}$ scales with $\sqrt{n}$ with respect to sample size; and $\sup_{\varepsilon>0} M_{l\to L}(W,\varepsilon)$ is the "one-point" Lipchitz constant at W in the sense that

$$
\begin{aligned}
&||F_L(F_l(W',X),\{W_i'\}_{i=l+1}^L) - F_L(F_l(W,X),\{W_i'\}_{i=l+1}^L)||_{\mathsf{F}} \\
\le& \left(\sup_{\varepsilon} M_{l\to L}(W,\varepsilon)\right)||F_l(W',X) - F_l(W,X)||_{\mathsf{F}}, \quad \forall W' \in B_{\mathsf{F}}(R).
\end{aligned}
$$

This concludes the first generalization bound in Corollary 1.

Step 4: Prove the Second Generalization Bound. Now we continue to show that the bound in Corollary 1 is strictly better than the spectrally normalized bound. To see this, as we presented under Corollary 1, we have

$$\begin{aligned}
&\|F_{l-1}(W,X)\|_{\mathsf{F}} \\
=&\|\sigma_{l-1}(W_{l-1}\cdots W_2\sigma_1(W_1X))\|_{\mathsf{F}} \\
\leq&\prod_{i<l}\|W_i\|_{\text{op}}\cdot\|X\|_{\mathsf{F}},
\end{aligned} \tag{111}$$

by the property of spectral norm ($\|AB\|_{\mathsf{F}}\leq\|A\|_{\text{op}}\|B\|_{\mathsf{F}}$), and the fact that all activation functions are 1−Lipchitz in column.

In the meanwhile, we know that

$$\left(\sup_{\varepsilon}M_{l\to L}(W,\varepsilon)\right)\leq\sup_{\varepsilon}\prod_{i>l}\|W_i'\|_{\text{op}},$$

again by the property of spectral norm ($\|AB\|_{\mathsf{F}}\leq\|A\|_{\text{op}}\|B\|_{\mathsf{F}}$) and the fact that all activation functions are 1−Lipchitz in column. This results in

$$\sup_{\varepsilon}\prod_{i>l}\|W_i'\|_{\text{op}}\leq\sup_{W\in B_{\mathsf{F}}(R)}\prod_{i>l}\|W_i\|_{\text{op}}. \tag{112}$$

Combining (111) and (112) together with (110), we have that for any $\delta\in(0,1)$, with probability at least $1-\delta$, uniformly over every $W\in B_{\mathsf{F}}(R)$, we have

$$\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \\
\leq&O\left(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\cdot\sum_{l=1}^L(d_l+d_{l-1})\prod_{i<l}\|W_i\|_{\text{op}}^2\sup_{W\in B_{\mathsf{F}}(R)}\prod_{i>l}\|W_i\|_{\text{op}}^2}}{n}\right. \\
&\left.+\sqrt{\frac{\beta^2\sum_{l=1}^L\log(d_{l-1}n)+\log\frac{\log(2n)}{\delta}}{n}}\right),
\end{aligned} \tag{113}$$

where O hides multiplicative absolute constants and two ignorable high-order terms: $\frac{\beta\sqrt{\sum_{l=1}^L(d_l+d_{l-1})d_{l-1}}}{n^{5.5}}$ and $\frac{\beta\sqrt{\sum_{l=1}^L(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2R^2\sup_{\varepsilon>0}M_{l\to L}^2(W,\varepsilon)}}{n2^n}$.

The next step is to use Lemma 8 (the "uniform pointwise convergence" principle [Xu and Zeevi, 2025]) to give a conversion from the uniform convergence to the pointwise convergence. Denote the functional $T_l:B_{\mathsf{F}}(R)\to(0,R_l]$ is defined by

$$T_l(W)=\prod_{i\neq l}\|W_i\|_{\text{op}}^2.$$

Since $\sum_{i\neq l}\|W_i\|_{\mathsf{F}}^2\leq\|W\|_{\mathsf{F}}^2\leq R^2$, we have $T_l(W)=\prod_{i\neq l}\|W_i\|_{\text{op}}^2\leq(R/\sqrt{L-1})^{2(L-1)}$ according to the AM-GM inequality. The bound in (113) implies that for any $l=1,\cdots,L$, $\forall t_l\in$

$(0, (R/\sqrt{L-1})^{2(L-1)})]$, with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{W:T_l(W)\leq t_l, \forall l\in[L]} (\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \\
\leq & O\left(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\cdot\sum_{l=1}^L(d_l+d_{l-1})t_l}}{n} + \sqrt{\frac{\beta^2\sum_{l=1}^L\log(d_{l-1}n)+\log\frac{\log(2n)}{\delta}}{n}}\right).
\end{aligned} \tag{114}$$

The inequality (114) precisely match the condition in Lemma 8 for each functional $d_l(w)$. Thus applying Lemma 8 by L times, with the smallest radius r_0 in the statement of Lemma 8 chosen to be $r_0 = (R/\sqrt{L-1})^{2(L-1)})/\max\{R,2\}^n$, we have that for any $\delta\in(0,1)$, with probability at least $1-\delta$, uniformly over every $W\in B_{\mathsf{F}}(R)$,

$$\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \\
\leq & O\Bigg(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\sum_{l=1}^L(d_l+d_{l-1})\max\{4T_l^2(W),\frac{(R/\sqrt{L-1})^{2L-2}}{\max\{R,2\}^{2n}}\}}}{n} \\
&+\sqrt{\frac{\beta^2\sum_{l=1}^L\log(d_{l-1}n)+L\log\frac{n\log\max\{R,2\}}{\delta}}{n}}\Bigg) \\
= & O\left(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\cdot\sum_{l=1}^L(d_l+d_{l-1})\prod_{i\neq l}\|W_i\|_{\mathrm{op}}^2}}{n} + \sqrt{\frac{\beta^2\sum_{l=1}^L\log(d_{l-1}n)+L\log\frac{n\log\max\{R,2\}}{\delta}}{n}}\right),
\end{aligned} \tag{115}$$

where O hides multiplicative absolute constants and three ignorable high-order terms: $\frac{\beta\sqrt{\sum_{l=1}^L(d_l+d_{l-1})d_{l-1}}}{n^{5.5}}$, $\frac{\beta\sqrt{\sum_{l=1}^L(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2R^2\sup_{\varepsilon>0}M_{l\to L}^2(W,\varepsilon)}}{n2^n}$ and $\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\sum_{l=1}^L(d_l+d_{l-1})}(R/\sqrt{L-1})^{L-1}}{n\max\{R,2\}^n}$.

Now we see from (111) and (112) that the derived norm-constraint bound (110) implies the spectrally normalized bound (115). This completes the proof.

□