Pointwise Generalization in Deep Neural Networks

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

Abstract

We address the fundamental question of why deep neural networks generalize by establishing a pointwise generalization theory for fully connected networks. For each trained model, we characterize the hypothesis via a pointwise Riemannian Dimension, derived from the eigenvalues of the *learned feature representations* across layers. This approach establishes a principled framework for deriving hypothesis-dependent generalization bounds that accurately characterize the rich, nonlinear regime, overcome long-standing barriers, and reshape the statistical foundations of representation learning. These pointwise spectrum-aware bounds offer a systematic upgrade over approaches based on model size, products of norms, and infinite-width linearizations, yielding guarantees that are orders of magnitude tighter in both theory and experiment. Analytically, we identify the structural properties and mathematical principles that explain the tractability of deep networks. Empirically, the pointwise Riemannian Dimension exhibits substantial feature compression, decreases with increased over-parameterization, and captures the implicit bias of optimizers across standard architectures and datasets. Taken together, our results indicate that deep networks are mathematically tractable in practical regimes and that their generalization is sharply explained by pointwise, spectrum-aware complexity.

Contents

[*] In keeping with standard practice in mathematics and theory, authors are listed alphabetically. 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 scientific tasks. Yet, a fundamental paradox remains: while classical theory predicts severe overfitting for massive models, practice exhibits strong generalization. This gap has fueled a prevailing view that neural networks are opaque "black boxes" resistant to principled explanation. We narrow this gap by addressing the generalization problem for the canonical fully connected Deep Neural Networks (DNN). We demonstrate that, under minimal and verifiable spectral conditions on the *learned feature representations*, deep neural networks fall into a tractable regime with tight generalization guarantees. Crucially, these conditions impose no constraints on parameter count or weight sparsity, standing in sharp contrast to prior statistical conventions based on pure weight-space compression. Methodologically, our characterization leverages a pointwise generalization paradigm that fundamentally transcends classical uniform-convergence and covering-number approaches. To our knowledge, this offers one of the first fully rigorous treatments that establishes the tractability of deep neural networks by contemporary machine-learning standards. We hope this framework helps demystify generalization and provides a systematic methodology for analyzing nonlinear, overparameterized models in representation learning.

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

$$F_l(W, X) := \sigma_l\big(W_l\, F_{l-1}(W, X)\big) \in \mathbb{R}^{d_l \times n}, \qquad l = 1, \ldots, L, \tag{1.1}$$

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, mild logarithmic factors, and reasonable simplification (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 $W \in \mathbb{R}^{\sum_l d_l \cdot d_{l-1}}$,

$$\mathcal{L}_{\text{test}}(W) - \mathcal{L}_{\text{train}}(W) \ \lesssim\ \sqrt{\frac{1}{n}\sum_{l=1}^{L}\big(d_l + d_{l-1}\big)\, d_{\text{eff}}\Big(F_{l-1}(W, X)\, F_{l-1}(W, X)^\top\Big)}. \tag{1.2}$$

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 (1.2) 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 (1.2) 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 (1.2) can be used directly as a *regularizer*, leading to algorithms that adapt to the effective ranks around a benchmark $W^\star$ (Section 4.3). Under minimal spectral conditions on the learned feature representations, our theory renders fully connected networks tractable. 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], Neyshabur et al. [2018]), 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 grounded in unifying structural principles, clear links to prior frameworks, and qualified tightness discussions. By confronting the long-standing challenge of obtaining sharp, *nonlinear* guarantees for representation learning transcending uniform convergence—emphasized in Bartlett et al. [2021], Zhang et al. [2021], Nagarajan and Kolter [2019], Wilson [2025]—we aim to establish that generalization in deep neural networks is mathematically *tractable.*

1.1 Organization and Contributions

The paper is organized into three parts: (i) a pointwise generalization framework (Section 2); (ii) structural principles of deep networks (Sections 3 to 4); and (iii) empirical validation (Section 5). Related work appears in Appendix A.1, and all proofs are in Appendices B, C, D, E. Below we summarize the main novelties in each part.

Pointwise Generalization and Finite-Scale Geometry. We develop a pointwise framework that analyzes the *trained* hypothesis and yields generalization bounds with (qualified) matching upper and lower rates via a finite-scale notion of *pointwise dimension.* This fundamentally upgrades generic chaining and all covering–number approaches by assigning each hypothesis its *own complexity* that directly controls its error. The bounds can also be read as an optimally tuned PAC–Bayes objective specialized to deterministic predictors. Taken together, this framework reframes generalization as a study of pointwise geometry governed by dimension reduction at finite precision, clarifying why nonlinear models can generalize even in the absence of uniform convergence.

Structural Principles and Tight Bounds for Neural Networks. We develop a *non-perturbative* analysis based on exact telescoping identities (rather than Taylor linearizations) that preserves the finite-scale geometry of deep networks. This yields our first structural principle: *cross-layer correlations factor through the feature matrices, approximately preserving a pointwise-linear structure* (Section 3.2). Next, we show that bounding the pointwise dimension reduces, on each local chart, to the gold-standard *effective dimension*, and we lift this to a global statement by constructing an *ellipsoidal* covering of the Grassmannian of subspaces. This extension—beyond classical differential-geometric/Lie-algebraic treatments—establishes our second structural principle: *the complexity of the global atlas (covering reference eigenspaces) is commensurate with that of the local charts.* Building on these principles, we introduce the *Riemannian Dimension*—a spectrum-aware, pointwise effective complexity—which governs generalization at the trained model and yields tight, analyzable bounds (Section 3.3). We provide explicit generalization guarantees for deep neural networks and argue that the bounds are tight in a qualified sense; moreover, they *exponentially* sharpen spectral-norm bounds and suggest principles for algorithm design (Section 4).

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 new understandings on the nature of generalization.

2.1 Pointwise Dimension Strengthens PAC-Bayes and Generic Chaining

Let $\mathcal{F}$ be a hypothesis class, let $z \in \mathcal{Z}$ be a random data (e.g., input-label pair $z = (x, y)$), and let $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ be a real-valued loss. Denote by $\mathbb{P}$ the (unknown) population distribution, and by $\mathbb{P}_n$ the empirical distribution associated with an i.i.d. sample $S = \{Z_i\}_{i=1}^n \sim \mathbb{P}^{\otimes n}$. For any integrable function $g : \mathcal{Z} \to \mathbb{R}$, define the population and empirical averaging operators by

$$\mathbb{P}g := \mathbb{E}_{z\sim\mathbb{P}}[g(z)], \qquad \mathbb{P}_n g := \frac{1}{n}\sum_{i=1}^n g(z_i).$$

For convenience, we will often write $g(z)$ inside these operators (e.g., $g(z) = \ell(f; z)$). Our goal is to control the *generalization gap* $(\mathbb{P} - \mathbb{P}_n)\ell(f; z)$ in the following *pointwise generalization* manner: for $\delta \in (0, 1)$, with probability at least $1 - \delta$, uniformly over every $f \in \mathcal{F}$,

$$(\mathbb{P} - \mathbb{P}_n)\ell(f; z) \ := \ \mathbb{E}_{z\sim\mathbb{P}}\big[\ell(f; z)\big] \ - \ \frac{1}{n}\sum_{i=1}^n \ell(f; z_i) \le \ C\sqrt{\frac{d(f) + \log\frac{1}{\delta}}{n}}, \tag{2.1}$$

where $d(f)$ is a hypothesis-dependent complexity measure that aims to characterize the intrinsic complexity of every *trained* hypothesis f, different from class-wide, uniformly defined complexity measures. In Appendix B.1, we state necessary and sufficient conditions for pointwise generalization through the "uniform pointwise convergence" principle proposed in Xu and Zeevi [2020, 2025]. This stands in sharp contrast to the uniform-convergence paradigm [Vapnik, 2013, Zhang et al., 2021, Nagarajan and Kolter, 2019], which seeks a single worst-case bound on $\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\ell(f;z)$ rather than the precise, hypothesis-dependent (pointwise) control we pursue here.

In the spirit of (2.1), we introduce the central notion of this section, the *pointwise dimension*: a finite-scale analogue of ideas from fractal geometry [Falconer, 1997] and a pointwise counterpart distilled from generic chaining [Fernique, 1975]. Throughout the paper, "metric" ϱ means a *semi-metric*: 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 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:*

$$\log \frac{1}{\pi(B_\varrho(f,\varepsilon))}. \tag{2.2}$$

We define the loss-induced empirical $L_2(\mathbb{P}_n)$ metric $\varrho_{n,\ell}$ as

$$\varrho_{n,\ell}(f_1, f_2) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ell(f_1; z_i) - \ell(f_2; z_i))^2}.$$

Equipped with this data–dependent metric, we now state the following unified pointwise dimension generalization upper bound.

Theorem 1 (Pointwise Dimension Generalization Bound) *Let $\ell(f; z) \in [0, 1]$. There exists an absolute constant $C > 0$ such that for any data-independent prior π on $\mathcal{F}$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, uniformly over every $f \in \mathcal{F}$*

$$(\mathbb{P} - \mathbb{P}_n)\ell(f; z) \le C\left(\inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^1 \sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon\right\} + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right).$$

The concept of pointwise dimension and the unified generalization bound in Theorem 1 strengthen several established generalization methodologies such as PAC-Bayesian analysis, Kolmogorov complexity, generic chaining, and covering numbers. We elaborate on this unified strengthening in the next two paragraphs.

Theorem 1 sharpens best PAC–Bayes optimization. By the monotonicity of the pointwise dimension in ε, a direct relaxation of Theorem 1 yields the one–shot bound

$$(\mathbb{P} - \mathbb{P}_n)\,\ell(f; z) \;\le\; C\left(\inf_{\alpha\ge 0}\left\{\underbrace{\alpha}_{\text{bias (approximate } f)} + \underbrace{\sqrt{\frac{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\alpha))}}{n}}}_{\text{variance (PAC–Bayes term)}}\right\} + \sqrt{\frac{\log\bigl(\log(2n)/\delta\bigr)}{n}}\right). \tag{2.3}$$

Intuitively, the pointwise dimension uses prior mass over a *ball* around f, so it applies to general (uncountable) classes, overcoming limitations of hypothesis–by–hypothesis bounds such as Occam/description–length and Kolmogorov complexity [Lotfi et al., 2022]. Additionally, our perspective brings the best possible PAC-Bayesian mechanism: generalization is recast as a bias–variance tradeoff optimized over a user–chosen posterior, applies to *deterministic* hypotheses, and shows that the pointwise dimension optimally governs the complexity (see Section B.2 for this perspective). This clarifies and strengthens earlier PAC–Bayes approaches, which—rather than optimizing over general posteriors—adopt linear-in-weight Gaussian approximations to obtain computable, non-vacuous generalization bounds (e.g., [Hinton and Van Camp, 1993, Dziugaite and Roy, 2017]; see the second part of Section 2.2 for details).

Theorem 1 upgrades generic chaining and covering numbers to a pointwise form. The theorem extends generic chaining (notably the majorizing measure integral [Fernique, 1975, Talagrand, 1987]) to *pointwise* bounds, and is therefore strictly stronger than entropy–integral bounds based on *uniform* 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 Section 3 of Block et al. [2021] and Section 4.1 of Chen et al. [2024]. In particular, (B.18) in Appendix B.3 shows that

$$\inf_{\pi} \sup_{f\in\mathcal{F}} \frac{1}{\pi(B_{\varrho}(f,\varepsilon))} \tag{2.4}$$

is (up to absolute constants) equivalent to the *canonical covering number* of $\mathcal{F}$ with metric ϱ at scale ε. Consequently, Theorem 1 goes beyond classical covering analyses by (i) recasting covering-number complexity as the inverse-prior-density objective (2.4), and (ii) localizing this complexity *pointwise* in f. Through the results and tools developed in this paper, we advocate the "prior-density + localization" (i.e., pointwise dimension) viewpoint as a new paradigm for statistical complexity analysis—one that is markedly more flexible than the classical covering-number approaches prevalent in statistics and machine learning.

We also note that the multiscale integral is stronger than the one-shot bound (2.3): it applies to rich classes where the pointwise dimension can grow as $O\big(d(f)\,\varepsilon^{-2}\big)$ yet still yields a $\sqrt{d(f)/n}$ rate; by contrast, the one–shot relaxation (2.3) typically requires growth no worse than $O\big(d(f)\log(1/\varepsilon)\big)$ to achieve the same rate.

Finally, the integral upper bound in Theorem 1 is tight in the following qualified worst-case sense: no uniform improvement valid simultaneously for all hypotheses and all priors is possible. This is witnessed by a matching lower bound.

Theorem 2 (Worst-Case Lower Bound) *Let $\ell(f;z)\in[0,1]$. There exists absolute constants $c, c'>0$ so that*

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

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

The lower bound certifies the *worst-case tightness* of our pointwise-dimension upper bound in Theorem 1 (noting that fixing $\alpha=0$ relative to Theorem 1 only increase the lower bound). This is analogous to *minimax optimality* in frequentist decision theory—viewing the selection of a data-dependent pointwise complexity, or the search over posteriors in PAC–Bayes, as a statistical decision problem [Wald, 1945]. This worst-case tightness does not preclude sharper guarantees for a fixed hypothesis f. However, a strictly *pointwise* lower bound—one that conditions on the realized hypothesis f without the outer $\sup_{f\in\mathcal{F}}$—is generally unattainable, because any admissible prior π must be chosen independently of f (a "no free lunch" constraint).

We defer technical innovations and connections to existing methodologies to the Appendix—most notably the unified pointwise–generalization framework of Xu and Zeevi [2020, 2025] which we build upon (Appendix B.1), and the alternative PAC–Bayesian perspective (Appendix B.2). The key takeaway is that the proposed *pointwise dimension* is a powerful and precise descriptor that tightly characterizes pointwise generalization.

2.2 Necessity of Finite-Scale Pointwise Geometry and Structural Analysis

The transition from uniform convergence to the "prior-density + localization" (pointwise dimension) perspective offers a fundamental tightening over standard covering number approaches. However, translating this theoretical advantage into a practical framework for deep learning requires addressing two distinct challenges. First, we will distinguish the *geometric nature* of generalization from classical infinitesimal geometry: relevance lies not in the limit $\varepsilon \to 0$, but motivates a new program of finite-scale geometric analysis. Second, we must overcome the *computational intractability* of evaluating the pointwise dimension directly, which necessitates a dedicated structural analysis for deep neural networks.

Asymptotic vs. Finite-Scale Dimension. Although powerful in mathematics, standard differential-geometric tools (e.g., pointwise metrics and subspace angles) have not been systematically used in generalization theory, largely because they define dimension in infinitesimal notions. For instance, the *asymptotic pointwise dimension*—central to fractal and Riemannian geometry [Falconer, 1997, Jost, 2008] and used to characterize Hausdorff and packing dimensions (e.g., Theorem 3 of Lutz [2016])—is defined via a limit:

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

We argue that generalization is distinct from, and in some ways more challenging than, infinitesimal geometry: the nature of generalization in deep models lies in reducing geometric dimension at a *finite scale* of precision for each hypothesis. Crucially, the pointwise dimension $\log\frac{1}{\pi(B_\varrho(f,\varepsilon))}$ is monotonic: it naturally decreases as the resolution ε increases. Therefore, a finite-scale analysis reveals significant dimension reduction that infinitesimal analysis misses. In our one-shot bound (2.3), the objective is to identify the optimal finite scale $\varepsilon^\star$ where the trade-off between precision and pointwise complexity is minimized. At this scale, the effective dimension can be orders of magnitude smaller than the asymptotic dimension, explaining the tractability of overparameterized models. To the best of our knowledge, this distinction is novel; prior uses of geometric dimension in generalization (e.g., Birdal et al. [2021]) have largely emphasized globally uniform and infinitesimal notions. And the Neural Tangent Kernel (NTK) [Jacot et al., 2018] and Gaussian-process [Lee et al., 2018] viewpoints are valid only in an infinitesimal neighborhood of initialization (equivalently, in the infinite-width regime). A precise account of deep-model generalization thus calls for a shift from infinitesimal calculus to finite-scale, pointwise geometry.

Computation and the Necessity of Structural Analysis. Although tight, Theorem 1—like many abstract bounds (PAC-Bayes, mutual-information, generic chaining)—is generally not computationally tractable on its own; practical use requires adapting it to the function class at hand and introducing suitable relaxations. If we denote an effective dimension by $d(f) = \log \frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon^\star))}$ ($\varepsilon^\star$ tuned in the one-shot bound (2.3)), a brute-force Monte Carlo estimator using i.i.d. draws $f' \sim \pi$ would require on the order of $e^{d(f)}$ samples to obtain a single hit $f' \in B_{\varrho_{n,\ell}}(f, \varepsilon)$ with constant probability. For high-dimensional deep networks, where $d(f)$ should be moderate to large, this is computationally prohibitive.

This intractability helps explain why much of the PAC–Bayes literature pivoted to *inherent linearization* via closed-form calculations under Gaussian priors/posteriors on the parameter space: by relaxing the search over general posteriors on the nonlinear hypothesis class $\mathcal{F}$ to Gaussians over weights $W \in \mathbb{R}^p$, one obtains tractable objectives [Hinton and Van Camp, 1993, Dziugaite and Roy, 2017]; see Sections 3 and 6 in Dziugaite and Roy [2017] for these objectives. However, this strategy implicitly imposes a uniform linearization that discards the distinctive *pointwise* geometry of deep networks, effectively flattening a curved manifold. To retain the sharpness of pointwise dimension without incurring the simulation barrier, we therefore avoid black-box sampling and instead develop explicit *structural principles* of deep networks that allow analytic control of the pointwise dimension—yielding generalization guarantees that are both theoretically rigorous and practically computable.

3 Deep Neural Networks and Pointwise Riemannian Dimension

In this section we develop a systematical pointwise dimension analysis for deep neural networks. Section 3.1 formalizes the standard fully connected architecture and notation. Section 3.2 introduces a non-perturbative calculus (avoiding infinitesimal Taylor expansions) to analyze finite-scale behavior. Section 3.3 introduces a hierarchical covering scheme—our key technical innovation—that overcomes the well-known linear/kernel bottleneck in classical statistical learning and enables a principled treatment of genuinely nonlinear models.

3.1 Neural Network Setup

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 $d_0, \dots, d_L$ and weight matrices $W = \{W_1, \dots, W_L\}$ with $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ for $l = 1, \dots, L$. Let $\sigma_1, \dots, \sigma_L$ 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 $X = [x_1, \dots, x_n] \in \mathbb{R}^{d_0 \times n}$ collect the n training inputs as columns. For each layer $l \in \{1, \dots, L\}$, 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 (1.1)),

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

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.

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)

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

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.

3.2 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 $W' \in B_{\varrho_n}(W, \varepsilon)$

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

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.

$$\begin{aligned} &F_L(W', X) - F_L(W, X) \\ &= \sum_{l=1}^{L} [\underbrace{\sigma_L(W_L' \cdots W_{l+1}'}_{\text{controled by} M_{l\to L}} \underbrace{\sigma_l}_{\text{by1}} (W_l' \underbrace{F_{l-1}(W, X)}_{\text{learned feature}})) - \sigma_L(W_L' \cdots W_{l+1}' \sigma_l(W_l \underbrace{F_{l-1}(W, X)}_{\text{learned feature}}))]. \end{aligned} \tag{3.1}$$

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 $W' \in B_{\varrho_n}(W, \varepsilon)$,*

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

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 (2.2) we analyze in Section 2 (formalized as Lemma 17; metric domination lemma), 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).

3.3 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"):

$$\begin{aligned}
&G_{\mathrm{NP}}(W) = \mathrm{blockdiag}\left(\cdots, LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}^\top(W,X)\otimes I_{d_l},\cdots\right)\\
&\varrho_{G_{\mathrm{NP}}(W)}(W,W') = \mathrm{vec}(W'-W)^\top G_{\mathrm{NP}}(W)\mathrm{vec}(W'-W).
\end{aligned}\tag{3.2}$$

We are therefore interested in bounding the enlarged pointwise dimension under the pointwise ellipsoidal metric $\varrho_{G_{\mathrm{NP}}(W)}$:

$$\log\frac{1}{\pi(B_{\varrho_n}(f(W,\cdot),\varepsilon))}\le\log\frac{1}{\pi(B_{\varrho_{G_{\mathrm{NP}}(W)}}(W,\sqrt{n}\varepsilon))}.$$

This section offers a deep dive past classical effective dimension, shifting to hierarchical covering and a global geometric analysis.

3.3.1 Gold 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 $G(W)$, and the vectorized weights $W\in\mathbb{R}^p$ are restricted to the Euclidean ball $B_2(R):=\{w\in\mathbb{R}^p:\|w\|_2\le R\}$, then one can achieve a tight effective dimension as follows: define the *effective rank*

$$r_{\mathrm{eff}}(G(W),R,\varepsilon):=\max\{k:\lambda_k(G(W))R^2\ge n\varepsilon^2/2\},\tag{3.3}$$

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

$$d_{\mathrm{eff}}(G(W),R,\varepsilon):=\frac{1}{2}\sum_{k=1}^{r_{\mathrm{eff}}(G(W),R,\varepsilon)}\log\left(\frac{8R^2\lambda_k(G(W))}{n\varepsilon^2}\right).\tag{3.4}$$

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 $\mathcal{V}\subseteq\mathbb{R}^p$ the r-dimensional subspace corresponding to the top-r_{eff} eigenspace of $G(W)$.

3.3.2 Key Challenge: Prior Independence from W.

However, the main challenge is that the prior π must be chosen independently of the training data. This means that the construction of π cannot rely on knowledge of the learned weights W, including their 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*

$$\varrho_{\text{proj},G(W)}(\mathcal{V}, \bar{\mathcal{V}}) := \big\| \, G(W)^{1/2} \big(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}\big) \big\|_{\text{op}} \;\le\; \tfrac{\sqrt{n}\varepsilon}{4R}, \tag{3.5}$$

we have

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

In (3.5), $\mathcal{P}_{\mathcal{V}}$ denotes the orthogonal projector onto the subspace $\mathcal{V}$, and $\varrho_{\text{proj},G(W)}$ thus defines an ellipsoidal projection metric between subspaces. Further details are provided in the appendix.

3.3.3 Hierarchical covering (mixture prior over subspaces).

We introduce a hierarchical covering framework that pushes statistical learning beyond classical linear and kernel paradigms, providing a principled toolkit for genuinely nonlinear models—one of the central innovations of this work. It operates on two levels: a bottom-level local-chart covering that captures spectrum-aware behavior within a fixed subspace, and a top-level global geometric analysis over the Grassmannian.

(i) For each reference subspace $\bar{\mathcal{V}}$, placing a uniform prior on $\bar{\mathcal{V}}$ yields a tight pointwise-dimension bound for all "local" weights W whose top$-r$ eigenspace of $G(W)$ is well approximated by $\bar{\mathcal{V}}$ (see Lemma 2).

(ii) At the top level, we place a prior over reference subspaces $\bar{\mathcal{V}}$ and average the local priors, producing a data-independent prior and the final bound.

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:

$$\underbrace{\log \frac{1}{\mu(B_{\varrho_{\text{proj},G(W)}}(\mathcal{V},\sqrt{n\varepsilon/4R}))}}_{\text{covering Grassmannian (global atlas)}} + \underbrace{\sup_{\bar{\mathcal{V}} \in B_{\varrho_{\text{proj},G(W)}}(\mathcal{V},\sqrt{n\varepsilon/4R})} \log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W,\sqrt{n}\varepsilon))}}_{\text{covering local charts}}, \tag{3.6}$$

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 local charts.

Lemma 2 controls the local part, while the following new result (Lemma 3) on the *ellipsoidal* covering of the Grassmannian controls the global part.[1]

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*

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

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 (3.2), the l–th block is

$$G_l(W) = A_l(W) \otimes I_{d_l}, \text{ where } A_l(W) = LM^2_{l\to L}(W,\varepsilon) \cdot F_{l-1}(W,X)F_{l-1}(W,X)^\top \in \mathbb{R}^{d_{l-1}\times d_{l-1}}.$$

[1] Since the effective rank r of $\bar{\mathcal{V}}$ can take any value in $\{1,\ldots,p\}$, the top-level Grassmannian covering must range over all $\mathrm{Gr}(p,r)$. This adds only a negligible $O(\log p)$ overhead to the global-level cost. Accordingly, we construct a data-independent prior in three prior hierarchy: ("global-r") choose the rank r (paying the $\log p$ overhead), ("global-$\bar{\mathcal{V}}$") choose a reference subspace $\bar{\mathcal{V}} \in \mathrm{Gr}(p,r)$, and ("local") sample within $\bar{\mathcal{V}}$ using the local chart prior; see Figure 1 for an illustration. For conceptual clarity, Lemma 3 focuses on the Grassmannian covering cost at a fixed rank r; and we defer the layer–specific specialization (to each $d_{l-1} \times d_{l-1}$ feature–matrix block) to the calculation in (3.8).

Since the Kronecker factor is I_{d_l}, the spectrum of G_l consists of the eigenvalues of the feature Gram matrix $F_{l-1}F_{l-1}^\top$ (scaled by $LM^2_{l\to L}$), each repeated d_l times. Consequently, the *local–chart* (within–subspace) covering cost at layer l scales as

$$d_l \cdot d_{\mathrm{eff}}\Big(LM^2_{l\to L}(W,\varepsilon)\cdot\ F_{l-1}(W,X)F_{l-1}(W,X)^\top\Big), \tag{3.7}$$

while the *atlas* (subspace–selection) cost is the Grassmannian term over $\mathrm{Gr}\big(d_{l-1},\, r_{\mathrm{eff}}[W,l]\big)$, where $r_{\mathrm{eff}}[W,l]$ is the effective rank of $A_l(W)\in\mathbb{R}^{d_{l-1}\times d_{l-1}}$. By Lemma 3 (and the footnote preceding it), the *global-atlas* (choosing-subspace) covering cost at layer ℓ scales as

$$d_{l-1}\cdot d_{\mathrm{eff}}\Big(LM^2_{l\to L}(W,\varepsilon)\cdot\ F_{l-1}(W,X)F_{l-1}(W,X)^\top\Big) + \log(d_{l-1}). \tag{3.8}$$

Together, (3.7) and (3.8) yield a clean layerwise decomposition: the width d_l multiplies the spectral complexity of incoming features (local charts), whereas the input dimension d_{l-1} governs the Grassmannian covering (global atlas). This complementary, seemingly magical "duality" underlies the calculation below.

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 (3.2). Define the pointwise Riemannian Dimension*

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

where $A_l(W)$ is the the feature matrix $LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}^\top(W,X)$; 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*. On closer inspection, the effect hinges on the block–decomposable structure in (3.2). This structure is intrinsic to layered neural networks and typically absent in generic nonlinear models, which helps explain why DNN are particularly amenable to sharp generalization analysis.

For intuition, we illustrate the construction of the prior π in the single-layer case—via the schematic in Figure 1. From a top-down view, the prior π can be generated by first sampling the effective rank r, then a subspace $\bar{\mathcal{V}}$ on the Grassmannian, and finally a weight W inside that subspace. The general L-layer setting is then obtained by applying the same construction independently to each layer and taking a product measure, which is enabled by the layer-wise decomposable structure of neural networks (a consequence of our non-perturbative analysis).

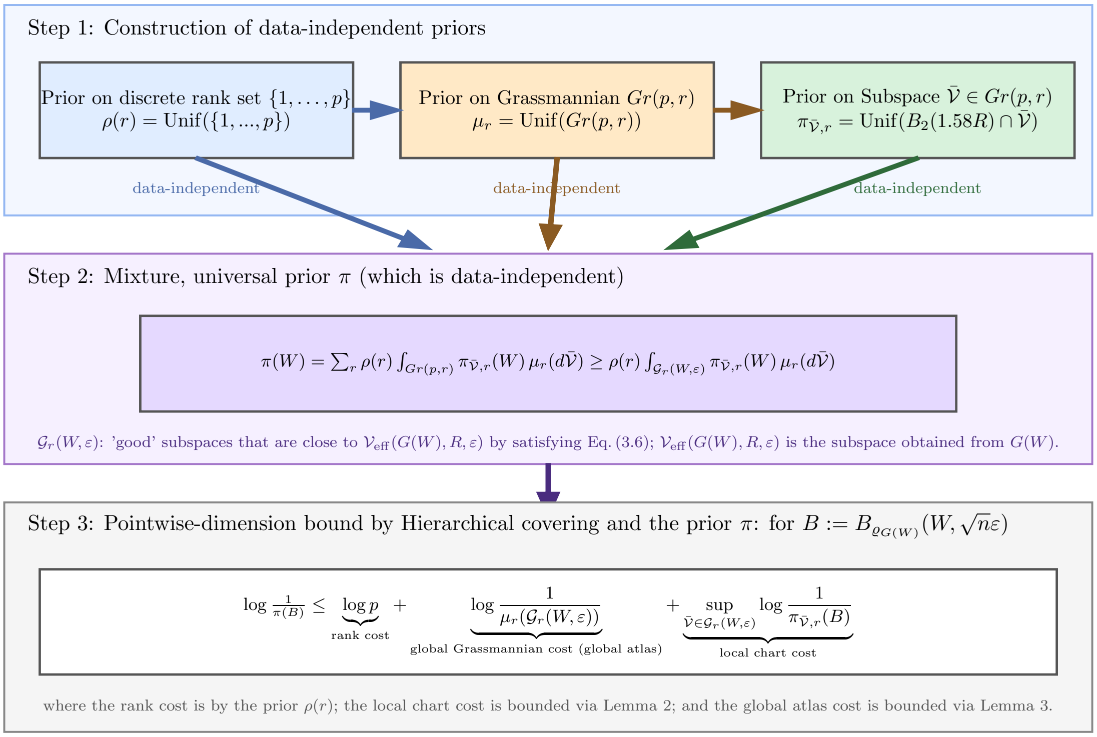
Step 1: Construction of data-independent priors
Prior on discrete rank set $\{1,\dots,p\}$ $\rho(r) = \mathrm{Unif}(\{1,...,p\})$
Prior on Grassmannian $Gr(p,r)$ $\mu_r = \mathrm{Unif}(Gr(p,r))$
Prior on Subspace $\bar{\mathcal{V}} \in Gr(p,r)$ $\pi_{\bar{\mathcal{V}},r} = \mathrm{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}})$
data-independent
data-independent
data-independent
Step 2: Mixture, universal prior π (which is data-independent)
$\pi(W) = \sum_r \rho(r) \int_{Gr(p,r)} \pi_{\bar{\mathcal{V}},r}(W)\, \mu_r(d\bar{\mathcal{V}}) \geq \rho(r) \int_{\mathcal{G}_r(W,\varepsilon)} \pi_{\bar{\mathcal{V}},r}(W)\, \mu_r(d\bar{\mathcal{V}})$
$\mathcal{G}_r(W,\varepsilon)$: 'good' subspaces that are close to $\mathcal{V}_{\mathrm{eff}}(G(W),R,\varepsilon)$ by satisfying Eq. (3.6); $\mathcal{V}_{\mathrm{eff}}(G(W),R,\varepsilon)$ is the subspace obtained from $G(W)$.
Step 3: Pointwise-dimension bound by Hierarchical covering and the prior π: for $B := B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon)$
$\log \frac{1}{\pi(B)} \leq \underbrace{\log p}_{\text{rank cost}} + \underbrace{\log \frac{1}{\mu_r(\mathcal{G}_r(W,\varepsilon))}}_{\text{global Grassmannian cost (global atlas)}} + \underbrace{\sup_{\bar{\mathcal{V}} \in \mathcal{G}_r(W,\varepsilon)} \log \frac{1}{\pi_{\bar{\mathcal{V}},r}(B)}}_{\text{local chart cost}}$
where the rank cost is by the prior $\rho(r)$; the local chart cost is bounded via Lemma 2; and the global atlas cost is bounded via Lemma 3.

Figure 1: Hierarchical construction of the data-independent prior π and its role in the pointwise-dimension bound (one single-layer case).

4 Generalization Bounds and Implications

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 1, 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 $\beta-$Lipschitz with respect to $f(W,x)$, for every $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over all $W \in B_{\mathsf{F}}(R)$,*

$$(\mathbb{P} - \mathbb{P}_n)\ell(f(W,x),y) \leq C_1 \left(\inf_{\alpha \geq 0} \left\{ \alpha + \frac{\beta}{\sqrt{n}} \int_\alpha^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)} d\varepsilon \right\} + \sqrt{\frac{\log \frac{\log(2n)}{\delta}}{n}} \right),$$

where the Riemannian Dimension is defined by

$$d_{\mathrm{R}}(W,\varepsilon) = \sum_{l=1}^{L}\Bigg((d_l + d_{l-1}) \sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \underbrace{\log \frac{8C_2^2 \lambda_k(F_{l-1}F_{l-1}^\top)}{n\varepsilon^2}}_{\text{spectrum of inner layers } 1:l\text{–}1}$$
$$+ (d_l + d_{l-1}) r_{\mathrm{eff}}[W,l] \cdot \underbrace{\log\Big(M_{l\to L}^2(W,\varepsilon) L \max\{||W||_{\mathsf{F}}^2, R^2/4^n\}\Big)}_{\text{spectrum of outer layers } l+1:L} + \log(d_{l-1}n)\Bigg), \tag{4.1}$$

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 (4.1) to the Informal Rate (1.2). 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 (1.2). In (4.1), 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 (4.1) 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 (1.2).

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.

1. **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 1), and the framework has a profound connection to finite-scale geometry—evidence that this is the right organizing principle.

2. **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 (3.1) is an exact *equality*, so the expansion is generally sharp (and fully avoid linearization).

3. **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 Comparison with Norm Bounds, VC, and NTK

We compare our generalization bound for fully connected DNN (Theorem 4) with three established lines of work: (i) bounds based on products of spectral norms, (ii) VC–dimension–type capacity bounds, and (iii) Neural Tangent Kernel (NTK) linearizations that are valid only in an infinitesimal neighborhood of initialization. Our framework yields *exponentially* tighter rates than norm–product bounds, refines VC–type statements into hypothesis– and data–dependent guarantees, and replaces infinitesimal linearization with a finite-scale, non-perturbative analysis that holds simultaneously for every trained hypothesis. For space, we defer the recovery of representative norm bounds to Appendix E.3.1 and a broader literature review to Appendix A.1.

Norm Bounds: Starting from the Riemannian–dimension term in Theorem 4, apply the elementary inequality

$$\log x \le \log(1+x) \le x, \quad \forall x > 0$$

together with $\sum_{k\ge 1} \lambda_k(F_{l-1}F_{l-1}^\top) = \|F_{l-1}(W,X)\|_{\mathsf{F}}^2$, we obtain: for each layer l,

$$\sum_{k\ge 1} \log\left(\frac{\lambda_k\big(F_{l-1}F_{l-1}^\top\big)\ \|W_l\|_{\mathsf{F}}^2\, L\, M_{l\to L}^2(W,\varepsilon)}{n\,\varepsilon^2}\right) \le \frac{\|F_{l-1}(W,X)\|_{\mathsf{F}}^2\,\|W_l\|_{\mathsf{F}}^2\, L\, M_{l\to L}^2(W,\varepsilon)}{n\,\varepsilon^2}.$$

Aggregating over layers and controlling $M_{l\to L}(W,\varepsilon)$ through $\prod_{i>l}\|W_i\|_{\mathrm{op}}$, Theorem 4 yields the following rank–free, spectrally normalized consequence: uniformly over all $W \in B_{\mathsf{F}}(R)$

$$(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \le O\left(\frac{\beta\|W\|_{\mathsf{F}}}{n}\sqrt{\sum_{l=1}^{L} L(d_l+d_{l-1})\prod_{i\ne l}\|W_i\|_{\mathrm{op}}^2}\right), \tag{4.2}$$

where the $O(\cdot)$ notation hides only nonessential terms (see Corollary 1 in Appendix E.3.1); moreover, since $\|X\|_{\mathsf{F}} \le \sqrt{n}\max_{1\le i\le n}\|x_i\|_2$, the bound in (4.2) typically scales as $n^{-1/2}$. Therefore, we illustrate that the Riemannian–dimension bound in Theorem 4 is *exponentially tighter* than (4.2), a representative spectral-norm bound in the style of Bartlett et al. [2017], Neyshabur et al. [2018], Golowich et al. [2020], Pinto et al. [2025], Ledent et al. [2025]. Appendix E.3.1 provides the full derivation and a detailed, side-by-side comparison.

VC Dimension: Let L be the number of layers and $P = \sum_{l=1}^{L} d_l d_{l-1}$ be the total number of weights, Bartlett et al. [2019] prove a nearly tight VC–dimension bound VCdim $\le O\left(PL\log P\right)$, supported by a lower bound VCdim $\ge \Omega\left(PL\log(P/L)\right)$. 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.

[2] The extra factor L beyond parameter count in VCdim is essentially unavoidable: for nonlinear compositional models, VC/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.

Neural Tangent Kernel (NTK): Our approach uses an exact, non-perturbative expansion that preserves the finite-scale geometry of deep networks, going beyond NTK's Taylor linearizations, which remain valid only in an infinitesimal neighborhood around initialization (or equivalently, in the infinite-width "lazy" regime) [Jacot et al., 2018, Arora et al., 2019]. Outside this regime the NTK approximation typically breaks down, limiting its explanatory power for practical networks. From a generalization standpoint, the initialization-centric, infinitesimal view suppresses the feature learning that actually drives generalization, and thus cannot account for why modern deep networks generalize well. In contrast, our results provide a finite-scale, pointwise theory that operates directly in practical regimes and explicitly captures feature learning through the spectra of the *learned* feature matrices.

4.3 Algorithmic Implications and Excess Risk Bound

Pointwise Dimension as Regularization and Excess Risk Bound. Our bounds imply a natural regularization strategy for algorithm design. Given the pointwise generalization inequality (2.1) (e.g., the Riemannian Dimension bound in Theorem 4), we consider a regularized ERM objective that explicitly minimizes this complexity measure:

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \left\{ \mathbb{P}_n \ell(f; z) + C\sqrt{\frac{d(f) + \log(2/\delta)}{n}} \right\}. \tag{4.3}$$

With probability at least $1-\delta$, its excess risk is bounded by (compared to any benchmark $f^\star \in \mathcal{F}$):

$$\begin{aligned} &\mathbb{P}\ell(\hat{f}; z) - \mathbb{P}\ell(f^\star; z) \\ &\le \inf_{f\in\mathcal{F}} \left\{ \mathbb{P}_n \ell(f; z) + C\sqrt{\frac{d(f) + \log(2/\delta)}{n}} \right\} - \mathbb{P}\ell(f^\star; z) \\ &\le (C + \sqrt{1/2})\sqrt{\frac{d(f^\star) + \log(2/\delta)}{n}}; \end{aligned} \tag{4.4}$$

see Appendix E.2 for full proof. Thus we obtain a problem–dependent oracle bound of order $\sqrt{d(f^\star)/n}$ that adapts to the optimal hypothesis $f^\star$.

From Explicit Regularization to Implicit Bias of Practical Algorithms. Since modern optimizers like SGD routinely drive empirical risk to near-zero, convergence analysis alone offers limited insight into generalization. The central theoretical challenge is therefore not determining *whether* a minimum is reached, but identifying *which* of the infinite interpolating solutions the optimizer selects. Plain ERM is insufficient for this task: without constraints on pointwise dimension, an empirical risk minimizer yields no guarantee of controlled excess risk. In contrast, our RD–regularized objective (4.3) explicitly enforces the low-complexity structure required for the generalization bound in (4.4). Although this intuition is rooted in the earliest practices of deep learning, our pointwise theory rigorously articulates the underlying mathematical reasoning.

This motivates a concrete agenda for optimization in deep learning: characterize algorithms whose implicit bias drives iterates toward solutions with *low pointwise complexity*, in particular low *Riemannian Dimension (RD)*. Analogous phenomena are well documented in linear and kernel settings: gradient descent converges to max–margin (logistic loss) or minimum–norm (least squares) solutions [Soudry et al., 2018, Gunasekar et al., 2018], iterate–averaged SGD behaves like ridge regression [Neu and Rosasco, 2018], and "ridgeless" kernel regression can generalize with an optimally zero ridge parameter [Liang and Rakhlin, 2020]; see Vardi [2023] for a survey. Our regularizer in (4.3), based on pointwise dimension and, in particular, the RD from Theorem 4, is strictly more informative than any single norm, making it a natural target for such analyses.

Empirically, we say an algorithm exhibits *Riemannian–Dimension implicit bias* if it preferentially returns solutions with small RD despite RD's large dynamic range; in Section 5.3 we observe that SGD indeed finds low-RD solutions.

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 $h \in \{2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}\}$. 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.2. The code to reproduce all experiments is available at our GitHub repository here.

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 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

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 4.2); 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

indicator of generalization in deep learning. In summary, increased parameterization is associated with reduced effective model complexity, and Riemannian Dimension faithfully characterizes this phenomenon.

5.2 Feature Learning Compresses Effective Rank

We investigate the dynamics of feature learning by monitoring the effective rank of the feature Gram matrices $F_{l-1}F_{l-1}^\top$ scaled by $L||W||_{\mathsf{F}}^2 \prod_{i>l} \|W_i\|_{\text{op}}^2$ (i.e., $F_{l-1}F_{l-1}^\top \cdot L||W||_{\mathsf{F}}^2 \prod_{i>l} \|W_i\|_{\text{op}}^2$), as dictated by our theory. Here, replacing the local Lipschitz constant $M_{l\to L}(W, \varepsilon)$ by the spectral-norm product $\prod_{i>l} \|W_i\|_{\text{op}}$ 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 2.

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 2. (2) Increased parameter

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

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 2. (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

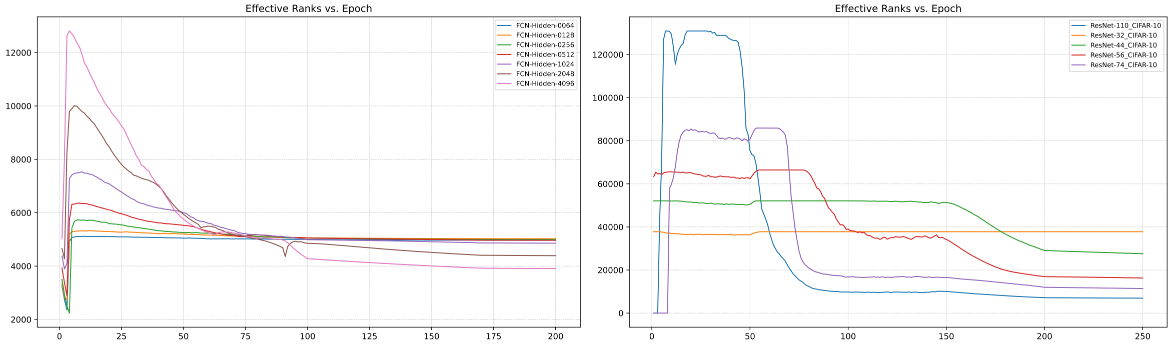

Figure 2: Effective Rank evolutions of FCNs on MNIST (left) and ResNets on CIFAR-10 (right) across the training

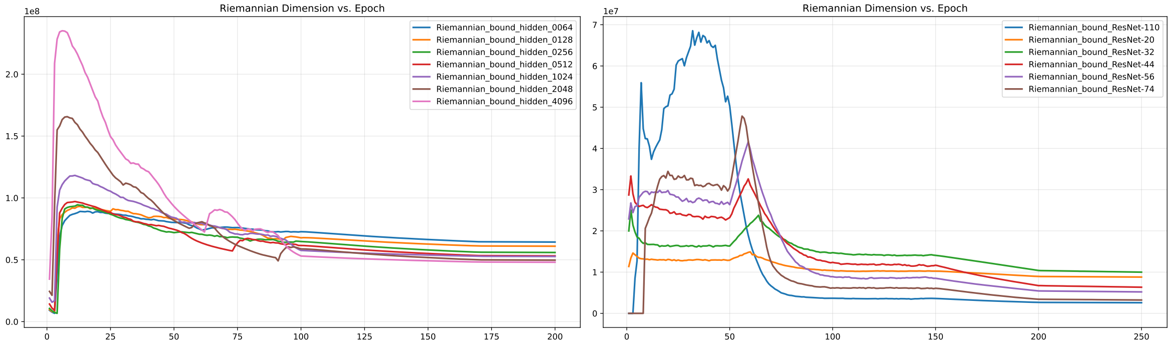

Figure 3: Riemannian Dimension evolutions of FCNs on MNIST (left) and ResNets on CIFAR-10 (right) across the training

whether this optimizer preferentially converge to solutions with lower Riemannian Dimension point, and the experimental results are presented in Figure 3.

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

Rayna Andreeva, Benjamin Dupuis, Rik Sarkar, Tolga Birdal, and Umut Simsekli. Topological generalization bounds for discrete-time stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 37:4765–4818, 2024.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pages 254–263. PMLR, 2018.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems 31*, pages 7245–7254, 2018.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2021.

Jean-Yves Audibert and Olivier Bousquert. Combining pac-bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8(4), 2007.

Jean-Yves Audibert and Olivier Bousquet. Pac-bayesian generic chaining. *Advances in neural information processing systems*, 16, 2003.

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

Laura Balzano, Tianjiao Ding, Benjamin D Haeffele, Soo Min Kwon, Qing Qu, Peng Wang, Zhangyang Wang, and Can Yaras. An overview of low-rank structures in the training and adaptation of large models. *arXiv preprint arXiv:2503.19859*, 2025.

Aristide Baratin, Thomas George, César Laurent, R Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. Implicit regularization via neural feature alignment. In *International Conference on Artificial Intelligence and Statistics*, pages 2269–2277. PMLR, 2021.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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

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

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

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.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776–6789, 2021.

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

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International conference on learning representations*, 2021.

Alexander Camuto, George Deligiannidis, Murat A Erdogdu, Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. *Advances in neural information processing systems*, 34:18774–18788, 2021.

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

Fan Chen, Dylan J Foster, Yanjun Han, Jian Qian, Alexander Rakhlin, and Yunbei Xu. Assouad, fano, and le cam with interaction: A unifying lower bound framework and characterization for bandit learnability. *Advances in Neural Information Processing Systems*, 37:75585–75641, 2024.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

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.

Benjamin Dupuis, George Deligiannidis, and Umut Simsekli. Generalization bounds using data-dependent fractal dimensions. In *International Conference on Machine Learning*, pages 8922–8968. PMLR, 2023.

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.

Alexander R Farhang, Jeremy D Bernstein, Kushal Tirumala, Yang Liu, and Yisong Yue. Investigating generalization by controlling normalized margin. In *International Conference on Machine Learning*, pages 6324–6336. PMLR, 2022.

Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35: 33054–33065, 2022.

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

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

Eugene Golikov, Eduard Pokonechnyy, and Vladimir Korviakov. Neural tangent kernel: A survey. *arXiv preprint arXiv:2208.13614*, 2022.

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.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1832–1841. PMLR, 2018.

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.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pages 4772–4784. PMLR, 2021.

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

Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1032–1041. PMLR, 2019.

Dmitry Kopitkov and Vadim Indelman. Neural spectrum alignment: Empirical study. In *International Conference on Artificial Neural Networks*, pages 168–179. Springer, 2020.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research*, 2022.

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.

Antoine Ledent, Rodrigo Alves, and Yunwen Lei. Generalization bounds for rank-sparse neural networks. *arXiv preprint arXiv:2510.21945*, 2025.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and generalization analysis of gradient methods for shallow neural networks. *Advances in neural information processing systems*, 35:38557–38570, 2022.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Jingbo Liu. Simple and sharp generalization bounds via lifting. *arXiv preprint arXiv:2508.18682*, 2025.

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.

Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim GJ Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *International Conference on Machine Learning*, pages 32801–32818. PMLR, 2024.

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

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

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

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

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.

Theodor Misiakiewicz and Andrea Montanari. Six lectures on linearized neural networks. *arXiv preprint arXiv:2308.13431*, 2023.

Vaishnavh Nagarajan and J. Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *arXiv preprint arXiv:1902.04742*, 2019.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Proceedings of the Conference on Learning Theory*, volume 75, pages 3222–3242. PMLR, 2018.

Behnam Neyshabur, Russ R Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems*, 28, 2015a.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015b.

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.

nLab contributors. Non-perturbative quantum field theory. `https://ncatlab.org/nlab/show/non-perturbative+quantum+field+theory`, 2025a. nLab (The n–Category Café).

nLab contributors. Strongly correlated system. `https://ncatlab.org/nlab/show/strongly+correlated+system`, 2025b. nLab (The n–Category Café).

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. 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.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

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.

Niket Patel and Ravid Shwartz-Ziv. Learning to compress: Local rank and information compression in deep neural networks. *arXiv preprint arXiv:2410.07687*, 2024.

Andrea Pinto, Akshay Rangamani, and Tomaso A. Poggio. On generalization bounds for neural networks with low rank layers. In *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272, pages 921–936. PMLR, 2025.

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

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics*, pages 868–874, 2024.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in neural information processing systems*, 34:8609–8621, 2021.

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

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 1232–1240. PMLR, 2016.

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.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

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

Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

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

Nathan Srebro, Noga Alon, and Tommi Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. *Advances In Neural Information Processing Systems*, 17, 2004.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

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.

Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer, Berlin, Heidelberg, 2014.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ieee information theory workshop*, pages 1–5. Ieee, 2015.

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

Vladimir Vapnik. *The nature of statistical learning theory.* Springer science & business media, 2013.

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.

Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46(2):265–280, 1945.

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. Minkowski inequality. `https://en.wikipedia.org/wiki/Minkowski_inequality`, 2025c.

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

Andrew Gordon Wilson. Deep learning is not so mysterious or different. *arXiv preprint arXiv:2503.02113*, 2025.

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.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 2524–2533, 2017.

Yunbei Xu and Assaf Zeevi. Towards problem-dependent optimal learning rates. *Advances in Neural Information Processing Systems*, 33:2196–2206, 2020.

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.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115, 2021.

A Related Works and Experimental Setup

A.1 Related Works

Given the breadth of work on generalization and its empirical proxies, the mathematical grounding of our approach, and its conceptual relevance to vision and language practice, we streamline the exposition by concentrating on the most relevant prior results.

Theoretical Generalization Bounds for DNN. A significant lineage of research anchors generalization bounds to various norms of network weights (e.g., path [Neyshabur et al., 2015a], spectral [Neyshabur et al., 2018, Bartlett et al., 2017, Arora et al., 2018], Frobenius [Neyshabur et al., 2015b, Golowich et al., 2020]). While offering conceptual insights, these bounds, often derived from globally uniform complexity measures like covering numbers or Rademacher complexity, frequently suffer from exponential dependencies on depth or layer norms, rendering them vacuous for practical, deep architectures. Compelling empirical evidence [Farhang et al., 2022, Razin and Cohen, 2020] further suggests that norm-based bounds alone are insufficient to fully elucidate the generalization phenomenon in deep learning. The kernel perspective [Belkin et al., 2018], epitomized by NTK theory [Jacot et al., 2018, Arora et al., 2019, Golikov et al., 2022], yields sharp guarantees by linearizing a network around its initialization—effectively casting training as kernel ridge regression with a fixed kernel. Within this linear/lazy regime, precise calculations explain both double descent [Belkin et al., 2019] and benign overfitting [Bartlett et al., 2020], and an eigenspace-projection viewpoint provides dimension-reduction and feature-compression insights [Bartlett et al., 2017]. Investigations beyond the lazy regime exist, but most analyses either study the two-layer infinite-width (mean-field) limit (e.g., [Mei et al., 2018, Chizat and Bach, 2018]) or remain in a neighborhood of initialization [Woodworth et al., 2020]. While insightful, these settings are idealized and struggle to capture the behavior of finite, deep networks (see Chapter 6 of [Misiakiewicz and Montanari, 2023]). More broadly, linear, lazy, or infinite-width approximations fail to reflect the feature learning that arises when parameters move far from initialization and representations evolve. This omission is widely viewed as a central bottleneck in current theory; indeed, the rich, representation-learning regime is often argued to be the key phenomenon distinguishing modern deep learning from long-standing frameworks (see, e.g., Wilson [2025]). Building on these directions, we establish—to our knowledge—the first pointwise generalization bounds for nonlinear DNN that are comparable in sharpness to prior linearization results and, crucially, remain valid in the practical feature-learning regime.

Other Theoretical Perspectives of Generalization. A growing line of work connects generalization to geometric notions of fractal dimension [Birdal et al., 2021, Dupuis et al., 2023, Simsekli et al., 2020, Andreeva et al., 2024, Camuto et al., 2021], typically through Hausdorff– or Minkowski–type dimensions of optimization trajectories or invariant measures. However, these fractal dimensions are globally uniform, infinitesimal-scale ($\varepsilon \to 0$) notions of complexity. In contrast, our theory is built on a pointwise, finite-scale notion of geometric dimension. Section 2.2 is precisely devoted to this distinction: we move from globally uniform to pointwise dimension and show that generalization is governed by the finite-scale pointwise dimension rather than its asymptotic limit. Several PAC–Bayesian approaches operate directly in parameter space W, endowing the weights with an explicit stochastic model and directly computing the KL divergence between a hand–designed prior and a posterior over W [Hinton and Van Camp, 1993, Dziugaite and Roy, 2017, Lotfi et al., 2022, 2024]; e.g., Gaussian distribution in Dziugaite and Roy [2017]. These parameter-space bounds are valuable for certifying that certain trained weight configurations admit non-vacuous PAC-Bayes guarantees, but they largely treat the network as a black box and do not directly capture how architecture and feature geometry control generalization. Alternative theoretical frameworks include algorithmic stability analyses, which are used primarily for one-hidden-layer networks and connected to the NTK/lazy-training viewpoint [Richards and Kuzborskij, 2021, Lei et al., 2022]; and VC-dimension methods [Bartlett et al., 2019], which has been discussed in Section 4.2.

Pointwise and Non-Perturbative Foundations. Our use of "pointwise" draws inspiration from several threads that emphasize hypothesis-specific complexity: the asymptotic pointwise dimension in fractal geometry [Falconer, 1997], PAC-Bayes analyses that tailor complexity to the chosen random posterior [McAllester, 1998, Alquier, 2024], and the Fernique–Talagrand integral in the majorizing-measure formulation of generic chaining [Fernique, 1975, Talagrand, 1987, Block et al., 2021]. The synthesis of PAC-Bayes bounds with generic chaining dates back to Audibert and Bousquet [2003], Audibert and Bousquert [2007], and mutual information based bounds have also been combined with chaining [Russo and Zou, 2016, Xu and Raginsky, 2017, Asadi et al., 2018, Liu, 2025]. In contrast, to our knowledge, a closed form pointwise bound for deterministic hypotheses together with an explicit connection to pointwise dimension is new. A generic conversion from classical (subset-homogeneous) uniform convergence to pointwise generalization bounds, established in Xu and Zeevi [2020, 2025], serves as a guiding principle and plays a central role in our proof of Theorem 1. The adjective "non-perturbative," borrowed from physics [nLab contributors, 2025a] and central to the study of strongly correlated systems [nLab contributors, 2025b], underscores that our theory remains valid far beyond infinitesimal neighborhoods of initialization—an essential property for deeply nonlinear, feature-learning DNN.

Connections to Differential Geometry and Lie Algebra. From a geometric perspective, Hausdorff dimension provides an asymptotic, covering-based notion of capacity (fundamental in geometric measure theory [Simon, 2018]), while differential and Riemannian geometry [Jost, 2008] develop the use of local charts and global atlases to analyze non-Euclidean manifolds. Our results motivate viewing generalization as a finite-scale problem in geometric analysis. The Grassmannian and families of orthogonal subspaces are traditionally studied via Lie groups; using differential-geometric tools, Szarek [1997], Pajor [1998] established finite-scale isotropic metric-entropy characterizations, which motivate our hierarchical covering viewpoint from local charts to a global atlas and our ellipsoidal entropy framework.

Empirical Indicators of Generalization. Complementing theory, much research has focused on empirical indicators that explain the generalization of deep learning. Phenomena like *Neural Collapse* [Papyan et al., 2020, Parker et al., 2023, Kothapalli, 2022] reveals the emergence of low-rank geometric structures in last-layer features. Studies on *Intrinsic Dimension* [Li et al., 2018, Huh et al., 2021] similarly suggest that deeper models exhibit an inductive bias toward low-rank last-layer feature representations. A line of work focuses on *Dynamic NTK variants* [Atanasov et al., 2021, Baratin et al., 2021, Fort et al., 2020, Kopitkov and Indelman, 2020] or related feature-gradient kernels [Radhakrishnan et al., 2024], where the kernels evolve along optimization trajectories, has empirically shown that the dynamic kernel evolution is linked to generalization behaviour. Other probes, examining Fisher information [Karakida et al., 2019, Jastrzebski et al., 2021], Hessian spectral properties [Ghorbani et al., 2019, Rahaman et al., 2019], and output-input Jacobians [Novak et al., 2018], offer another lens. Collectively, existing empirical probes offer valuable, though often partial, insights—typically from a specific layer perspective, or through a constructed similarity analysis—without a unifying formalism and a theory foundation. Our proposed empirical indicator, rooted in a mathematically sharp theory, resonates with their goals (our theory is in fact supported by many of their experiments) while advancing them. It provides a principled, formal measure for characterizing the generalization of neural networks.

Feature Compression in Deep Models for Vision and Language. Across vision and language, deep networks exhibit a robust layer–wise compression of representations. In computer vision, Ansuini et al. [2019] measure intrinsic dimensionality across convolutional layers and find early expansion followed by sharp reduction, with lower late–stage dimensionality correlating with stronger generalization; Feng et al. [2022] likewise show that feature matrices in CNNs and vision transformers become progressively low–rank with depth, at fixed width, indicating active compression of task–relevant information. Parallel trends appear in NLP: Cai et al. [2021] demonstrate that contextual embeddings (e.g., BERT) occupy narrow, anisotropic cones despite high nominal dimension, and Razzhigaev et al. [2024] document a two–phase training trajectory—initial expansion, then sustained compression. A complementary line grounded in the Information Bottleneck [Tishby and Zaslavsky, 2015] interprets these findings as the selective removal of task–irrelevant variability: Shwartz-Ziv and Tishby [2017] observe that networks spend most of training compressing internal features toward a prediction–compression trade–off, while Patel and Shwartz-Ziv [2024] show gradient descent reduces the local rank of intermediate activations. Balzano et al. [2025] provide a complementary tutorial on low-rank structures arising during the training and adaptation of large models, emphasizing how gradient-descent dynamics and implicit regularization generate low-rank representations. Taken together, these phenomena motivate our investigation: compression is not merely qualitative, but admits precise, hypothesis–specific complexity that governs generalization.

A.2 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 $\left[\sqrt{1/n},\ \max_{l=1,\ldots,L} \sqrt{\frac{2L\lambda_{\max}(F_{l-1}F_{l-1}^{\top})\cdot\|W\|_F^2 \prod_{i>l}\|W_i\|_{\mathrm{op}}^2}{n}}\right]$. The search selects the value that minimizes our one-shot Riemannian Dimension-based generalization bound (2.3). 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)

Much of this section is devoted to a full proof of Theorem 1 (the integral upper bound). Conceptually, the pointwise–dimension principle already follows from elementary PAC–Bayes arguments—see Lemma 6 and the subsequent remark in Appendix B.2. We present the full derivation to make explicit structural properties (e.g., unified blueprint, subset homogeneity, population–empirical isomorphism) that a rigorous proof requires.

B.1 The "Uniform Pointwise Convergence" Principle

In this section, we present a unified blueprint for establishing pointwise generalization bounds. We state necessary and sufficient conditions for pointwise generalization and show that, when applied carefully, the resulting pointwise bounds are no harder to obtain than classical uniform-convergence guarantees.

We begin by citing a general principle for converting *subset-homogeneous* uniform convergence guarantees—i.e., bounds in which the same pointwise complexity applies for every fixed subset $\mathcal{H} \subseteq \mathcal{F}$—into pointwise generalization bounds. This conversion, introduced by the name "uniform localized convergence" principle in [Xu and Zeevi, 2020] (short conference version) and Xu and Zeevi [2025] (full journal version), 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 4 ("Uniform Pointwise Convergence" Principle) (Proposition 1 in Xu and Zeevi [2020, 2025]). *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 $1 - \delta$,*

$$\sup_{f \in \mathcal{F}: d(f) \le r} (\mathbb{P} - \mathbb{P}_n)\ell(f; z) \le \psi(r; \delta). \tag{B.1}$$

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

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

This lemma provides a succinct proof that serves as a unifying principle to sharpen classical localization, building on Section 2 of Xu and Zeevi [2025]. A key advantage of this framework is its level of abstraction: it establishes subset homogeneity as the necessary and sufficient condition for pointwise generalization when the complexity functional $d(\cdot)$ is data–independent, and likewise when $d(\cdot)$ is swap–invariant and depends on both the observed sample $S = \{z_i\}_{i=1}^n$ and an i.i.d. ghost sample $S' = \{z_i'\}_{i=1}^n$. It also provides a clean treatment of data–dependent functionals and their induced (random) sublevel sets $\{f \in \mathcal{F} : d(f) \le r\}$, as outlined before Section4 and detailed in AppendixA.4 of Xu and Zeevi [2025], thereby avoiding the circularities that often arise when combining symmetrization with localization/offset arguments.

B.1.1 Necessary and Sufficient Conditions for Pointwise Generalization

We leverage this "uniform pointwise convergence" principle to streamline the derivation of our bounds. Let $d(\cdot)$ denote a pointwise complexity functional, which we categorize into data-independent forms and data-dependent forms. Let $\psi(\cdot; \delta)$ be a non-decreasing function (typically $\psi(r; \delta) \asymp \sqrt{(r + \log(1/\delta))/n}$). We provide a clean characterization of pointwise generalization.

Necessary Condition: Subset Homogeneity. A valid pointwise generalization guarantee (i.e., (2.1)) necessitates *subset homogeneity*. That is, if the pointwise inequality

$$(\mathbb{P} - \mathbb{P}_n)\ell(f; z) \leq \psi(d(f); \delta) \tag{B.3}$$

holds with probability at least $1 - \delta$, then (B.3) must imply that for every subset $\mathcal{H} \subseteq \mathcal{F}$,

$$\sup_{f \in \mathcal{H}} \big(\mathbb{P} - \mathbb{P}_n\big)\, \ell(f; z) \ \leq \ \sup_{f \in \mathcal{H}} \psi\big(d(f); \delta\big).$$

Crucially, the complexity evaluation d(f) must *not* depend on the chosen subset $\mathcal{H}$. For instance, for the pointwise dimension $\log \frac{1}{\pi(B_\varrho(f,\varepsilon))}$, the prior π (in particular, its support) should be independent of $\mathcal{H}$. This contrasts with classical empirical-process techniques—e.g., naive uses of Rademacher complexity and generic chaining—where the *pre-specified* index sets dictate the proxy $d(\cdot)$ via the chosen Rademacher expectation, admissible tree construction, or prior.

Subset homogeneity is thus the primary eligibility check for any candidate pointwise complexity functional. In Appendix B.3, we complete this check by establishing that the pointwise dimension is *ambiently equivalent*: using a prior $\pi \in \Delta(\mathcal{F})$ or its restriction $\pi \in \Delta(\mathcal{H})$ produces complexities that agree in order (up to absolute constants).

Sufficient Condition: Subset Homogeneity + Data-Independent (or Symmetrized) $d(\cdot)$. Assuming the following subset-homogeneity uniform convergence condition: for every fixed (i.e., data-independent) subset $\mathcal{H} \subseteq \mathcal{F}$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{H}} (\mathbb{P} - \mathbb{P}_n)\ell(f; z) \leq \sup_{f \in \mathcal{H}} \psi\big(d(f); \delta\big). \tag{B.4}$$

By taking the sublevel set

$$\mathcal{H} \ = \ \{f \in \mathcal{F} : \ d(f) \leq r\},$$

the condition (B.4) (applied to this fixed $\mathcal{H}$) directly implies the surrogate conditions (B.1) in Lemma 4, and hence the pointwise bound (B.2). Thus, subset homogeneity is a necessary and sufficient condition for a data-independent $d(\cdot)$ to imply a pointwise generalization bound.

Likewise, in Appendix B.4, we show that when the complexity $d(\cdot)$ may depend on both the observed sample $S = \{z_i\}_{i=1}^n$ and an i.i.d. ghost sample $S' = \{z'_i\}_{i=1}^n$, *provided it is swap–invariant in* (S, S') *(i.e., invariant under any exchange* $z_i \leftrightarrow z'_i$*)*, subset homogeneity suffices to yield a pointwise generalization bound via a final swap–symmetrization argument. This establishes Theorem 5: a pointwise generalization result in which the complexity is evaluated using both the observed sample S and the ghost sample S'.

Toward pointwise bounds using only observed sample. If one seeks bounds that are fully computable from the observed data $\{z_i\}_{i=1}^n$ alone (e.g., Theorem 1), without ghost sample or sample splitting, the analysis is more involved. A practical route is two–step: (i) first derive a symmetrized pointwise bound using a complexity functional based on (S, S') (which is already valid and sharp); (ii) then prove an isomorphism between the $L_2(\mathbb{P}_S)$– and $L_2(\mathbb{P}_{S'})$–induced pointwise complexities (following Appendix A.4 of Xu and Zeevi [2025]) so as to replace population or ghost–dependent terms by empirical ones, yielding a fully data–dependent bound. Theorem 1 will be fully proved in Appendix B.5.

B.2 The PAC-Bayes Optimization Problem

We illustrate why pointwise dimension is a natural consequence of *BEST* PAC-Bayes optimization.

Lemma 5 (PAC–Bayes Bound [Catoni, 2003]; see also Theorem 2.1 in Alquier [2024]) *Let π be a prior on a hypothesis class $\mathcal{F}$ independent to the data, and let $\ell\colon \mathcal{F}\times\mathcal{Z}\to[0,1]$ 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,*

$$(\mathbb{P}-\mathbb{P}_n)\langle\mu,\ell(f;z)\rangle \;\le\; \inf_{\eta>0}\left\{\frac{\mathrm{KL}\big(\mu,\pi\big)\;+\log\frac{1}{\delta}}{\eta n}+\frac{\eta}{8}\right\}=\sqrt{\frac{\mathrm{KL}\big(\mu,\pi\big)\;+\log\frac{1}{\delta}}{8n}}\,.$$

We now use the PAC-Bayes bound (which holds uniformly for every random posterior μ) to approximate a deterministic hypothesis f. On the event that the above PAC-Bayes bound holds, with probability at least $1-\delta$, we have that uniformly over every random $\mu\in\Delta(\mathcal{F})$ every deterministic $f\in\mathcal{F}$, for every $\eta>0$, the following uniform "deterministic hypothesis" bound holds:

$$\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)\ell(f;z)\\
=&\langle\mu,(\mathbb{P}-\mathbb{P}_n)\ell(\cdot;z)\rangle+\langle\mu,(\mathbb{P}_n-\mathbb{P})[\ell(\cdot;z)-\ell(f;z)]\rangle\\
\le&\frac{\eta}{8}+\frac{\mathrm{KL}\big(\mu,\pi\big)\;+\log\frac{1}{\delta}}{\eta n}+\langle\mu,\frac{1}{n}\sum_{i=1}^{n}|\ell(\cdot;z)-\ell(f;z)|\rangle+\langle\mu,\mathbb{E}|\ell(\cdot;z)-\ell(f;z)|\rangle\\
=&\frac{\eta}{8}+\frac{\mathrm{KL}\big(\mu,\pi\big)\;+\log\frac{1}{\delta}}{\eta n}+\langle\mu,\tilde{\varrho}(\cdot,f)\rangle,
\end{aligned}\tag{B.5}$$

where the metric $\tilde{\varrho}$ is defined as the sum of loss-induced $L_1(\mathbb{P}_n)$ metric and $L_1(\mathbb{P})$ metric:

$$\tilde{\varrho}(f',f)=\frac{1}{n}\sum_{i=1}^{n}|\ell(f';z)-\ell(f;z)|+\mathbb{E}|\ell(f';z)-\ell(f;z)|.\tag{B.6}$$

In (B.5), the inequality uses the PAC-Bayes bound (Lemma 5) to bound the first term, which we term the "variance" term, and use absolute values to bound the second term, which we term the "bias" term.

Motivated by the above bias-variance optimization (B.5) via PAC-Bayes, for a given prior π, metric ϱ, and confidence $\delta\in(0,1)$ we define the *PAC-Bayes optimization objective*

$$V(\mu,\eta,f,\varrho):=\underbrace{\frac{\eta}{8}+\frac{\mathrm{KL}(\mu,\pi)+\log\frac{1}{\delta}}{\eta n}}_{\text{Variance}}+\underbrace{\langle\mu,\varrho(\cdot,f)\rangle}_{\text{Bias}},\tag{B.7}$$

where $\eta>0$, n is the sample size, μ is a posterior over hypotheses. Here, the "Variance" term arises from a PAC-Bayes bound (Lemma 5) applied to μ, and the "Bias" term $\langle\mu,\varrho(\cdot,f)\rangle:=\mathbb{E}_{h\sim\mu}\big[\varrho(h,f)\big]$ measures how well the randomized μ approximates the target f.

Optimizing the Posterior μ for the Objective (B.7) The intuitive analysis (B.5) explains how the PAC-Bayesian optimization objective naturally bounds the generalization gap. We now minimize the posterior μ in (B.7). It is straightforward that (B.7) is minimized by the Gibbs posterior. To obtain a closed-form characterization of the optimized value, we proceed in two steps: (i) derive an explicit pointwise-dimension upper bound by taking μ to be the π-normalized distribution on the metric ball $B_\varrho(f,\varepsilon)$ (Lemma 6), and (ii) show that this choice is near-optimal (Lemma 7).

B.2.1 Pointwise Dimension Bound via Metric Ball

Given any prior π on $\mathcal{F}$ and any $f \in \mathcal{F}$, take μ to be the $\pi-$normalized distribution on the metric ball $B_\varrho(f,\varepsilon)$, i.e.,

$$\mu(A) = \frac{\pi(A \cap B_\varrho(f,\varepsilon))}{\pi(B_\varrho(f,\varepsilon))} \quad \text{for all measurable } A \subseteq \mathcal{F}. \tag{B.8}$$

This simple choice is essentially optimal in that it yields the same analytical upper bound as the Gibbs posterior that minimizes the bound (later presented in Lemma 7).

Lemma 6 (Pointwise Dimension and Pointwise Generalization Upper Bound) *For the PAC–Bayes objective* (B.7)*, let μ be $\pi-$normalized on $B_\varrho(f,\varepsilon)$, i.e.*

$$\frac{d\mu}{d\pi}(h) \;=\; \begin{cases} \dfrac{1}{\pi\big(B_\varrho(f,\varepsilon)\big)}, & h \in B_\varrho(f,\varepsilon), \\ 0, & h \notin B_\varrho(f,\varepsilon). \end{cases}$$

Then, with $\eta^\star = \sqrt{8\big(\mathrm{KL}(\mu,\pi) + \log(1/\delta)\big)/n}\,$,

$$V\big(\mu,\eta^\star,f,\varrho\big) \;\le\; \sqrt{\frac{\mathrm{KL}(\mu,\pi)+\log(1/\delta)}{2n}} \;+\; \varepsilon \;=\; \sqrt{\frac{\log\frac{1}{\pi(B_\varrho(f,\varepsilon))} + \log(1/\delta)}{2n}} \;+\; \varepsilon. \tag{B.9}$$

Combining the upper bound (B.9) *with* (B.5) *yields the pointwise generalization bound: for every $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over every $f \in \mathcal{F}$,*

$$(\mathbb{P}-\mathbb{P}_n)\ell(f;z) \le \inf_{\varepsilon>0}\left\{\sqrt{\frac{\log\frac{1}{\pi(B_{\tilde\varrho}(f,\varepsilon))} + \log(1/\delta)}{2n}} \;+\; \varepsilon\right\},$$

where $\tilde\varrho$ is the mixed $L_1(\mathbb{P}_n) + L_1(\mathbb{P})$ metric defined by $\tilde\varrho(f',f) = \frac{1}{n}\sum_{i=1}^n |\ell(f';z) - \ell(f;z)| + \mathbb{E}|\ell(f';z) - \ell(f;z)|$.

Remark (why this intuition matters). Since the mixed L_2–metric dominates the sum of empirical and population L_1–metrics, consider

$$\bar\varrho(f',f) \;:=\; \left(\frac{1}{n}\sum_{i=1}^n\big(\ell(f';z_i) - \ell(f;z_i)\big)^2 \;+\; \mathbb{E}\big[\big(\ell(f';Z) - \ell(f;Z)\big)^2\big]\right)^{1/2}. \tag{B.10}$$

By Lemma 17, pointwise dimension is monotone in the underlying metric; hence replacing $\tilde\varrho$ by the larger metric $\sqrt{2}\bar\varrho$ yields a valid pointwise generalization bound. For a trained predictor $\hat f$, this means we may estimate the bound using the observed sample $S = \{z_i\}_{i=1}^n$ together with an i.i.d. ghost sample $S' = \{z_i'\}_{i=1}^n$ to evaluate balls in the mixed metric (B.10).

Theorem 1 then *sharpens* this picture: it turns the one–shot PAC–Bayes bound into a chaining–integral and removes the need for the ghost sample by working solely with the empirical $L_2(\mathbb{P}_n)$–metric. The core spirit of Theorem 1 remains the same with the PAC–Bayes bias–variance optimization; the practical differences are (i) integral vs. one–shot control and (ii) using S alone instead of (S, S').

Proof of Lemma 6: For the choice (B.8),

$$\mathrm{KL}(\mu,\pi) = \int_{\mathcal{F}} \log\Big(\frac{d\mu}{d\pi}(h)\Big)\,\mu(dh) = \int_{B_\varrho(f,\varepsilon)} \log\Big(\tfrac{1}{\pi(B_\varrho(f,\varepsilon))}\Big)\,\mu(dh) = \log\frac{1}{\pi(B_\varrho(f,\varepsilon))}. \tag{B.11}$$

Moreover, by construction,

$$\langle \mu, \varrho(\cdot, f)\rangle = \int_{B_\varrho(f,\varepsilon)} \varrho(h,f)\,\mu(dh) \le \varepsilon.$$

Plugging (B.11) into (B.7) and minimizing $\frac{\eta}{8}+\frac{\mathrm{KL}(\mu,\pi)+\log(1/\delta)}{\eta n}$ over $\eta > 0$ gives $\sqrt{(\mathrm{KL}(\mu,\pi) + \log(1/\delta))/(2n)}$, which together with the bias bound $\langle \mu, \varrho(\cdot, f)\rangle \le \varepsilon$ yields the claimed bound (B.9).

□

B.2.2 Lower Bound and Optimality of PAC-Bayes Optimization

The following lemma indicates that the uniform-ball 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 7 (Optimality of Pointwise Dimension in PAC-Bayes Optimization) *For the PAC–Bayes optimization objective $V(\mu, \eta, f, \varrho)$ defined in (B.7), we have that for every $f \in \mathcal{F}$, $\eta > 0$, and $\varepsilon > 0$,*

$$\inf_{\mu} V(\mu,\eta,f,\varrho) \;\ge\; \frac{\eta}{8} + \frac{\log\frac{1}{\delta}}{\eta n} + \min\Big\{\frac{1}{\eta n}\log\frac{1}{\pi(B_\varrho(f,\varepsilon))},\ \varepsilon\Big\} \;-\; \frac{\log 2}{\eta n}. \tag{B.12}$$

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

$$\frac{\eta}{8} + \frac{\log\frac{1}{\delta}}{\eta n} + \min\Big\{\frac{\log\frac{1}{\pi(B_\varrho(f,\varepsilon))}}{\eta n},\ \varepsilon\Big\} \;-\; \frac{\log 2}{\eta n} \le \inf_{\mu} V(\mu,\eta,f,\varrho) \;\le\; \frac{\eta}{8} + \frac{\log\frac{1}{\pi(B_\varrho(f,\varepsilon))} + \log\frac{1}{\delta}}{\eta n} + \varepsilon. \tag{B.13}$$

Proof of Lemma 7 The upper bound in (B.13) is already proved in Lemma 6, so we only need to prove the lower bound (B.12). The Donsker–Varadhan variational identity states that for any measurable h,

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

Apply it with $h = -\eta n \varrho(\cdot, f)$ to obtain

$$-\log\int e^{-\eta n\varrho(\cdot,f)}\,d\pi \;= \inf_{\mu}\Big\{\mathrm{KL}(\mu,\pi) \;+\; \int \eta n\varrho(\cdot,f)\,d\mu\Big\},$$

which implies that

$$\frac{\eta}{8}+\frac{\log\frac{1}{\delta}}{\eta n}-\frac{1}{\eta n}\log\int e^{-\eta n\varrho(\cdot,f)}\,d\pi \;=\inf_{\mu}\Big\{\frac{\eta}{8}+\frac{\mathrm{KL}\big(\mu,\pi\big)\;+\log\frac{1}{\delta}}{\eta n}+\langle\mu,\varrho(\cdot,f)\rangle\Big\}. \tag{B.14}$$

By splitting the dual integral,

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

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

$$\inf_{\mu} V(\mu,\eta,f,\varrho)\;\ge\;\frac{\eta}{8}+\frac{\log\frac{1}{\delta}}{\eta n}-\frac{1}{\eta n}\log\Big(\pi(B_\varrho(f,\varepsilon))+e^{-\eta n\varepsilon}\Big). \tag{B.15}$$

The simplified form (B.12) follows from $a+b\le 2\max\{a,b\}$ or equivalently $-\log(a+b)\ge-\log 2+\min\{-\log a,-\log b\}$ on (B.15). Combining (B.7), (B.11) and (B.12) yields the sandwich (B.13).

□

B.3 Subset Homogeneity of Pointwise Dimension

We show that, for any $f\in\mathcal{H}\subseteq\mathcal{F}$, the pointwise–dimension functional defined with a prior π is unchanged in order (up to absolute constants) whether π is supported on $\mathcal{H}$ or on the ambient class $\mathcal{F}$. Hence one may take $\pi\in\Delta(\mathcal{F})$ without restricting it to any particular subset, which suffices to meet the subset–homogeneity condition in Appendix 2.

Lemma 8 (Ambient Equivalence of Pointwise Dimension) *Let $(\mathcal{F},\varrho)$ be a metric space and let $\mathcal{H}\subseteq\mathcal{F}$ be a subset. Consider a nearest-point selector $p:\mathcal{F}\to f$ satisfying $\varrho(f,p(f))=\min_{h\in\mathcal{H}}\varrho(f,h)$ for all $f\in\mathcal{F}$, and the pushforward measure induced by the nearest-point selector:*

$$\pi_{\mathcal{H}}(h):=\int\pi(f)\mathbb{1}\{p(f)=h\}df.$$

Then for every $\varepsilon>0$ we have

$$\pi_{\mathcal{H}}(B_\varrho(f,2\varepsilon))\ge\pi(B_\varrho(f,\varepsilon)),\quad \log\frac{1}{\pi_{\mathcal{H}}(B_\varrho(f,2\varepsilon))}\le\log\frac{1}{\pi(B_\varrho(f,\varepsilon))}.$$

Consequently, for $a>0$, $b>0$, $\mu\in\Delta(\mathcal{F})$, $f\in\mathcal{H}$, define the majorizing measure integral

$$I(\pi,f,\varrho):=\inf_{\alpha\ge0}\left\{\alpha+\frac{1}{\sqrt{n}}\int_{\alpha}^{+\infty}\sqrt{\log\frac{1}{\pi\big(B_\varrho(f,\varepsilon)\big)}}\,d\varepsilon\right\}$$

Then we have

$$\frac{1}{2}\inf_{\mu\in\Delta(\mathcal{H})}\sup_{f\in\mathcal{H}}I(\mu,f,\varrho)\le\inf_{\pi\in\Delta(\mathcal{F})}\sup_{f\in\mathcal{H}}I(\pi,f,\varrho)\le\inf_{\mu\in\Delta(\mathcal{H})}\sup_{f\in\mathcal{H}}I(\mu,f,\varrho). \tag{B.16}$$

Proof of Lemma 8: The upper bound in (B.16) is immediate since $\Delta(\mathcal{H}) \subset \Delta(\mathcal{F})$: taking μ supported on $\mathcal{H}$ gives $\inf_{\pi\in\Delta(\mathcal{F})} \sup_{f\in\mathcal{H}} I(\pi, f, \varrho) \le \inf_{\mu\in\Delta(\mathcal{H})} \sup_{f\in\mathcal{H}} I(\mu, f, \varrho)$.

For the lower bound in (B.16), take $\pi_{\mathcal{H}}$ to be the pushforward induced by the nearest-point selector. For any $f \in \mathcal{H}$ and $\varepsilon > 0$, if $f' \in B_\varrho(f, \varepsilon)$ then

$$\varrho\big(p(f'), f\big) \;\le\; \varrho\big(p(f'), f'\big) + \varrho(f', f) \;=\; \min_{f\in\mathcal{H}} \varrho(f', f) + \varrho(f', f) \;\le\; 2\varepsilon,$$

hence $p(f') \in B_\varrho(f, 2\varepsilon)$ and

$$\pi_{\mathcal{H}}\big(B_\varrho(f, 2\varepsilon)\big) \;\ge\; \pi\big(B_\varrho(f,\varepsilon)\big), \quad \log\frac{1}{\pi_{\mathcal{H}}(B_\varrho(f, 2\varepsilon))} \le \log\frac{1}{\pi(B_\varrho(f,\varepsilon))}. \tag{B.17}$$

Therefore,

$$\begin{aligned}
I(\pi, f, \varrho) \;&=\; \inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{+\infty}\sqrt{\log\frac{1}{\pi(B_\varrho(f,\varepsilon))}}\,d\varepsilon\right\}\\
&\ge\; \inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{+\infty}\sqrt{\log\frac{1}{\pi_{\mathcal{H}}(B_\varrho(f,2\varepsilon))}}\,d\varepsilon\right\}\\
&= \tfrac{1}{2}\inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{+\infty}\sqrt{\log\frac{1}{\pi_{\mathcal{H}}(B_\varrho(f,\varepsilon))}}\,d\varepsilon\right\}\\
&= \frac{1}{2} I(\pi_{\mathcal{H}}, f, \varrho),
\end{aligned}$$

where the first inequality is by (B.17); the second equality is by the change of variables. Taking $\sup_{f\in\mathcal{H}}$ and then $\inf_{\pi\in\Delta(\mathcal{F})}$, $\inf_{\mu\in\Delta(\mathcal{H})}$ yields the desired lower bound.

□

Relationship to Fractional Covering Number Additionally, note that the minimax quantity

$$\mathrm{N}'(\mathcal{H}, \varrho, \varepsilon) := \inf_{\pi\in\Delta(\mathcal{F})} \sup_{f\in\mathcal{H}} \frac{1}{\pi\big(B_\varrho(f,\varepsilon)\big)}$$

is the *fractional covering number*; see Section 3 of Block et al. [2021] for its role in chaining; see also Chen et al. [2024] for connections to information-theoretic lower bounds (e.g., Fano's method, the Yang–Barron method, and local packing). In particular, with $\mathrm{N}(\mathcal{H}, \varrho, \varepsilon)$ denoting the (internal) covering number from Definition 5, we have the order equivalence (Lemma 8 in Block et al. [2021]; Lemma 14 in Chen et al. [2024])

$$\log \mathrm{N}(\mathcal{H}, \varrho, 2\varepsilon) \;\le\; \log \mathrm{N}'(\mathcal{H}, \varrho, \varepsilon) = \inf_{\pi\in\Delta(\mathcal{F})} \sup_{f\in\mathcal{H}} \log\frac{1}{\pi\big(B_\varrho(f,\varepsilon)\big)} \;\le\; \log \mathrm{N}(\mathcal{H}, \varrho, \varepsilon). \tag{B.18}$$

The covering number in Definition 5 does not depend on the ambient set $\mathcal{F}$, which in turn suggests that the pointwise dimension enjoys favorable ambient–equivalence properties.

Collapsing the Distinction between Chaining and Generic Chaining. A simple illustration of the strength of our pointwise blueprint is the multi–dimensional setting. Let $(d^{(1)},\dots,d^{(k)}) : \mathcal{F} \to (0,R]^k$ be coordinatewise nondecreasing complexities and let $\psi(\cdot;\delta)$ be monotone. Our blueprint makes no essential distinction between the two uniform forms

$$\begin{aligned} &(\text{sup–inside}) \quad \sup_{f\in\mathcal{H}} (\mathbb{P}-\mathbb{P}_n)\,\ell(f;z) \;\le\; \psi\Big(\sup_{f\in\mathcal{H}} d_n^{(1)}(f),\dots,\sup_{f\in\mathcal{H}} d_n^{(k)}(f);\ \delta\Big),\\ &(\text{sup–outside}) \quad \sup_{f\in\mathcal{H}} (\mathbb{P}-\mathbb{P}_n)\,\ell(f;z) \;\le\; \sup_{f\in\mathcal{H}}\ \psi\Big(d_n^{(1)}(f),\dots,d_n^{(k)}(f);\ \delta\Big), \end{aligned}$$

in the sense that *either* one leads to the same pointwise conclusion after peeling.

More precisely, fix a base scale $r_0 \in (0,R]$. Then with probability at least $1-\delta$, for every $f \in \mathcal{F}$,

$$(\mathbb{P}-\mathbb{P}_n)\,\ell(f;z) \;\le\; \psi\bigg(\Big(\cdots,\max\Big\{2d^{(j)}(f),\, r_0\Big\},\cdots\Big);\ \delta\Big(\log_2\frac{2R}{r_0}\Big)^{-k}\bigg). \tag{B.19}$$

The most straightforward proof uses essentially the same peeling argument as in Lemma 4, with the only change that we use a grid of size $(\log_2(2R/r_0))^k$ (partition each coordinate into $\log_2(2R/r_0)$ dyadic scales); see the short proof of Proposition 1 in Xu and Zeevi [2025]. Alternatively, this can proved by applying Lemma 4 for k times, where at each step we remove one dimension functional and divided confidence by $\log_2(2R/r_0)$. Moreover, the multi–dimensional pointwise bound (B.19) shows that its right–hand side, viewed as a *scalar* complexity, yields an equally tight pointwise bound. Hence the multi–dimensional formulation does not improve the best-achievable rates beyond a suitably defined one–dimensional complexity (as in generic chaining).

Conceptually, this shows that the apparent gap between classical chaining (entropy integral; sup–inside), generic chaining (majorizing measures; sup–outside), and our pointwise generic–chaining bound (Theorem 1) disappears within the blueprint: each is just a subset–homogeneous uniform statement that implies the same pointwise bound up to absolute constants and minor logarithms.

B.4 Pointwise Generalization Bound via Ghost Sample

In this section, we prove an easier variant of Theorem 1 that permits *swap-invariant* randomized priors depending on both the observed sample and its ghost counterpart. This setting subsumes—and strengthens—the conditional mutual information (CMI) framework of Steinke and Zakynthinou [2020].

Let $S = (z_1,\dots,z_n)$ and $S' = (z'_1,\dots,z'_n)$ be two i.i.d. samples drawn from $\mathbb{P}^n$, independent of each other. For each index $i \in \{1,\dots,n\}$, define the coordinate–swap map

$$\tau_i(S,S') \;:=\; \big((z_1,\dots,z_{i-1},z'_i,z_{i+1},\dots,z_n),\ (z'_1,\dots,z'_{i-1},z_i,z'_{i+1},\dots,z'_n)\big).$$

A (randomized) prior π on $\mathcal{F}$ is a mapping $\pi_{(\cdot,\cdot)} : \mathcal{Z}^{2n} \to \Delta(\mathcal{F})$; we write $\pi_{(S,S')} \in \Delta(\mathcal{F})$ for the realized prior conditioned on $\mathcal{F}$. We say that π is *swap–invariant* on (S,S'), if

$$\pi_{(S,S')} \;=\; \pi_{\tau_i(S,S')} \qquad \text{for all } i=1,\dots,n \text{ and for } \mathbb{P}^{2n}\text{-a.e. } (S,S').$$

Equivalently, π depends only on the unordered multiset $\{\{(z_i,z'_i)\}_{i=1}^n\}$ and not on which element of each pair is designated as "observed" versus "ghost."

Connection to CMI. This notion covers the conditional–mutual–information (CMI) framework of Steinke and Zakynthinou [2020]. In the CMI setup, one Draw paired data $Z = ((Z_i^{(0)}, Z_i^{(1)}))_{i=1}^n \overset{\text{i.i.d.}}{\sim} (\mathbb{P} \times \mathbb{P})^n$ and an independent selector $U \in \{0,1\}^n$. The training and ghost sets are $S_U = (Z_1^{(U_1)}, \ldots, Z_n^{(U_n)})$ and $S_{\bar{U}} = (Z_1^{(1-U_1)}, \ldots, Z_n^{(1-U_n)})$. Any prior π that is a function of Z only (independent of U) is swap–invariant, since flipping U_i implements τ_i. Conversely, swap–invariance for all i is equivalent to invariance under all coordinatewise flips of U, hence independence from U.

Throughout, let $S = \{z_i\}_{i=1}^n$ be the observed sample and $S' = \{z_i'\}_{i=1}^n$ an i.i.d. ghost sample, independent of S. We write $\mathbb{P}_S$ for the empirical measure $\mathbb{P}_n$ based on S, and $\varrho_{S,\ell}$ for the metric $\varrho_{n,\ell}$ from the main paper. Let $\mathbb{P}_{S'}$ denote the empirical measure based on S'. For any integrable function $g : \mathcal{Z} \to \mathbb{R}$ (we write $g(z)$ when convenient; e.g., $g(z) = \ell(f;z)$), define the empirical averaging operators

$$\mathbb{P}_S g := \frac{1}{n}\sum_{i=1}^n g(z_i), \qquad \mathbb{P}_{S'} g := \frac{1}{n}\sum_{i=1}^n g(z_i').$$

We use the shorthand $(\mathbb{P}_S \pm \mathbb{P}_{S'})g := \mathbb{P}_S g \pm \mathbb{P}_{S'} g$ for the sum/difference of the two sample–average operators, and the same notation when $\mathbb{P}_S \pm \mathbb{P}_{S'}$ appear inside norms or distances.

Theorem 5 (Pointwise Generalization via Ghost Sample) *Let $f(f;z) \in [0,1]$. There exists an absolute constant $C > 0$ such that for any swap-invariant prior $\pi_{(\cdot,\cdot)}$ on (S, S'), and any $\delta \in (0,1)$, with probability at least $1-\delta$ over S, uniformly in $f \in \mathcal{F}$,*

$$(\mathbb{P}_{S'} - \mathbb{P}_S)\,\ell(f;z) \;\le\; C\left(\inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{\sqrt{2}}\sqrt{\log\frac{1}{\pi_{(S,S')}\big(B_{\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}}\,d\varepsilon\right\} + \sqrt{\frac{\log\big(\log(2n)/\delta\big)}{n}}\right),$$

where $\varrho_{(S,S'),\ell}(f_1,f_2) = \sqrt{(\mathbb{P}_S + \mathbb{P}_{S'})\big(\ell(f_1;z) - \ell(f_2;z)\big)^2}$.

Proof of Theorem 5: The proof of the upper bound in Theorem 5 consists of three steps: 1. Subset-Homogeneous Uniform Convergence; 2. Generic Conversion to Pointwise Generalization Bound; 3. High-Probability Symmetrization.

Step 1: Subset-Homogeneous Uniform Convergence. Let $S = \{z_i\}_{i=1}^n$ be the observed sample, and $S' = \{z_i'\}_{i=1}^n$ be an i.i.d. ghost sample. We consider the symmetrized loss

$$\tilde{\ell}(f;(z,z')) = \ell(f;z') - \ell(f;z). \tag{B.20}$$

Since $\ell(f;z)$ is uniformly bounded in $[0,1]$, $\tilde{\ell}(f;(z,z'))$ is uniformly bounded in $[-1,1]$. We adopt the notation

$$\varrho_{S,\ell}(f_1,f_2) = \varrho_{n,\ell}(f_1,f_2) = \sqrt{\mathbb{P}_S\big(\ell(f_1;z) - \ell(f_2;z)\big)^2}.$$

from the main paper. Furthermore, we define the loss-induced $L_2(\mathbb{P}_n)$ metrics $\varrho_{S',\ell}$ and $\varrho_{(S,S'),\ell}$ by

$$\varrho_{S',\ell}(f_1,f_2) = \sqrt{\mathbb{P}_S(\ell(f_1;z) - \ell(f_2;z))^2},$$
$$\varrho_{(S,S'),\ell}(f_1,f_2) = \sqrt{(\mathbb{P}_S + \mathbb{P}_{S'})\,(\ell(f_1;z) - \ell(f_2;z))^2}.$$

By Minkowski's inequality (see, e.g., Wikipedia contributors [2025c]) and $\sqrt{a}+\sqrt{b} \le \sqrt{2(a+b)}$, we have

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{\ell}(f_1;(z_i,z_i')) - \tilde{\ell}(f_2;(z_i,z_i')))^2} \le \varrho_{n,\ell}(f_1,f_2) + \varrho_{S',\ell}(f_1,f_2) \le \sqrt{2}\varrho_{(S,S'),\ell}(f_1,f_2). \tag{B.21}$$

Now applying the truncated integral bound (Lemma 12) to the empirical Rademacher complexity: let $\{\xi_i\}_{i=1}^n$ be i.i.d. Rademacher variables, then conditioned on (S,S'), given any subset $\mathcal{H} \subseteq \mathcal{F}$, we have that for all $\delta \in (0,1)$, with probability at least $1-\delta$ (the randomness all comes from $\{\xi_i\}_{i=1}^n$),

$$\begin{aligned}\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\xi_i\tilde{\ell}(f;(z_i,z_i)) &\le \mathbb{E}_\xi\left[\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\xi_i\tilde{\ell}(f;(z_i,z_i))\right] + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}\\ &\le C_0\inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\inf_{\mu\in\Delta(\mathcal{H})}\sup_{f\in\mathcal{H}}\int_\alpha^2\sqrt{\log\frac{1}{\mu(B_{\tilde{\varrho}}(f,\varepsilon))}}d\varepsilon\right\} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}},\end{aligned}$$

where $C_0 > 0$ is an absolute constant, and $\tilde{\varrho}(f_1,f_2) := \sqrt{\frac{1}{n}\sum_{i=1}^n(\tilde{\ell}(f_1;(z_i,z_i')) - \tilde{\ell}(f_2;(z_i,z_i')))^2}$. Here, the first inequality is by Mcdiarmid's inequality (Lemma 14); and the second inequality is by Lemma 12; the integral is capped at 2 because $\tilde{\ell}$ is uniformly bounded by $\tilde{\varrho}$ is uniformly bounded by 2. By the ambient–equivalence of the pointwise–dimension functional (Lemma 8), we have

$$\inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\inf_{\mu\in\Delta(\mathcal{H})}\sup_{f\in\mathcal{H}}\int_\alpha^2\sqrt{\log\frac{1}{\mu(B_{\tilde{\varrho}}(f,\varepsilon))}}\,d\varepsilon\right\} \le \inf_{\alpha\ge 0} 2\left\{\alpha + \frac{1}{\sqrt{n}}\inf_{\pi\in\Delta(\mathcal{F})}\sup_{f\in\mathcal{H}}\int_\alpha^2\sqrt{\log\frac{1}{\pi\big(B_{\tilde{\varrho}}(f,\varepsilon)\big)}}\,d\varepsilon\right\}.$$

By (B.21) and the fact that pointwise dimension is monotone in the underlying metric (Lemma 17), we have that for any $\pi\in\Delta(\mathcal{F})$,

$$\int_\alpha^2\sqrt{\log\frac{1}{\pi\big(B_{\tilde{\varrho}}(f,\varepsilon)\big)}}\,d\varepsilon \le \int_\alpha^2\sqrt{\log\frac{1}{\pi\big(B_{\sqrt{2}\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}}\,d\varepsilon = \sqrt{2}\int_{\alpha/\sqrt{2}}^{\sqrt{2}}\sqrt{\log\frac{1}{\pi\big(B_{\sqrt{2}\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}}\,d\varepsilon.$$

Combining the above three inequalities, we prove the following *subset-homogeneous* uniform convergence argument when choosing an arbitrary $\pi\in\Delta(\mathcal{F})$: conditioned on (S,S'), for any $\mathcal{H}\subseteq\mathcal{F}$ and $\delta\in(0,1)$, with probability at least $1-\delta$ (the randomness all comes from $\{\xi_i\}_{i=1}^n$,

$$\sup_{f\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\xi_i(\ell(f;z_i)-\ell(f;z_i')) \le \sup_{f\in\mathcal{H}} C_1\inf_{\alpha\ge 0}\left\{\alpha + \frac{1}{\sqrt{n}}\int_\alpha^{\sqrt{2}}\sqrt{\log\frac{1}{\pi\big(B_{\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}}\,d\varepsilon\right\} + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}, \tag{B.22}$$

where $C_1 = 2\sqrt{2}C_0 > 0$ is an absolute constant.

Conditioned on (S,S'), for fixed $\pi_{(S,S')}\in\Delta(\mathcal{F})$ that is independent with $\{\xi_i\}_{i=1}^n$, define the pointwise complexity

$$d_{S,S'}(f) := \left(\inf_{\alpha\ge 0}\left\{\sqrt{n}\,\alpha + \int_\alpha^{\sqrt{2}}\sqrt{\log\frac{1}{\pi_{(S,S')}\big(B_{\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}}\,d\varepsilon\right\}\right)^2. \tag{B.23}$$

Then, by (B.22), conditioned on (S, S'), for any $\mathcal{H} \subseteq \mathcal{F}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ (the randomness all comes from $\{\xi_i\}_{i=1}^n$),

$$\sup_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \xi_i(\ell(f; z_i') - \ell(f; z_i)) \ \le \ \sup_{f \in \mathcal{H}} \left(C_1 \sqrt{\frac{d_{S,S'}(f)}{n}} \ + \ \sqrt{\frac{2\log\frac{2}{\delta}}{n}} \right). \tag{B.24}$$

As discussed in Appendix B.1.1, this condition is both necessary and sufficient to establish pointwise convergence when the complexity functional is the $\{\xi_i\}_{i=1}^n$-independent $d_{S,S'}(\cdot)$ when conditioned on (S, S').

Step 2: Generic Conversion to Pointwise Generalization Bound. All the analysis in this step is condition on (S, S'), thus all the randomness discussed here comes from $\{\xi_i\}_{i=1}^n$. For every $r \in (0, n]$, we take the subset

$$\mathcal{H} = \{f \in \mathcal{F} : d_{S,S'}(f) \le r\},$$

which, by (B.24), implies that $\forall \delta \in (0,1)$ and $\forall r \in (0, 2n]$, with probability at least $1-\delta$

$$\sup_{f : d_{S,S'}(f) \le r} \frac{1}{n} \sum_{i=1}^n \xi_i(\ell(f; z_i) - \ell(f; z_i')) \le C_1 \sqrt{\frac{r}{n}} + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}, \tag{B.25}$$

where and C_1 is an absolute constant. The inequality (B.25) is precisely the condition (B.1) in the generic conversion provided in Lemma 4 (here, the expectation (equal to 0) and the empirical average are taken for $\{\xi_i\}_{i=1}^n$). Thus applying Lemma 4 we have the pointwise generalization bound: conditioned on (S, S'), for any $\delta \in (0,1)$, by taking $r_0 = 1/n$, with probability at least $1-\delta$, uniformly over all $f \in \mathcal{F}$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \xi_i(\ell(f; z_i') - \ell(f; z_i)) \le & \frac{C_1}{\sqrt{n}} \sqrt{\max\left\{2d_{S,S'}(f), \frac{1}{n}\right\}} + \sqrt{\frac{2\log\frac{2\log_2(4n^2)}{\delta}}{n}} \\ \le & C_1 \sqrt{\frac{2d_{S,S'}(f)}{n}} + \frac{C_1}{n} + \sqrt{\frac{4\log\frac{2\log_2(2n)}{\delta}}{n}}. \\ \le & C_2 \left(\sqrt{\frac{d_{S,S'}(f)}{n}} + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}} \right), \end{aligned} \tag{B.26}$$

where $C_2 > 0$ is an absolute constant, where the second inequality is because there exists $C_2 \ge \sqrt{2}C_1$ such that for all positive integer n,

$$\frac{C_1}{n} + \sqrt{\frac{4\left(\log\frac{1}{\delta} + \log(\log 2 + \log n) + \log\frac{2}{\log 2}\right)}{n}} \le C_2 \sqrt{\frac{\log\frac{1}{\delta} + \log(\log 2 + \log n)}{n}}.$$

Thus we prove the pointwise generalization bound (B.26) for the complexity functional $d_{S,S'}(\cdot)$ defined in (B.23), under the randomness of $\{\xi_i\}_{i=1}^n$: for all $\pi \in \Delta(\mathcal{F})$, conditioned on (S, S'), for any $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over all $f \in \mathcal{F}$,

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i(\ell(f;z_i') - \ell(f;z_i)) \leq C_2 \left(\inf_{\alpha \geq 0} \left\{ \alpha + \frac{1}{\sqrt{n}} \int_{\alpha}^{\sqrt{2}} \sqrt{\frac{1}{\pi_{(S,S')}\big(B_{\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}} d\varepsilon \right\} + \sqrt{\frac{\log \frac{\log(2n)}{\delta}}{n}} \right). \tag{B.27}$$

Step 3: High-Probability Symmetrization. Recall that $S = \{z_i\}_{i=1}^n$ and $S' = \{z_i'\}_{i=1}^n$ are i.i.d. samples, independent of each other, and $\{\xi_i\}_{i=1}^n$ are i.i.d. Rademacher signs, independent of (S,S'). The mixed (ghost) metric

$$\varrho_{(S,S'),\ell}(f_1,f_2) = \sqrt{(\mathbb{P}_S + \mathbb{P}_{S'})\big(\ell(f_1;z) - \ell(f_2;z)\big)^2},$$

is swap-invariant to the pair (z_i, z_i') for each $i = 1, \cdots, n$. By the definition of swap-invariant prior before Theorem 5, the prior $\pi_{(S,S')} \in \Delta(\mathcal{F})$ is also swap-invariant to the pair (z_i, z_i'). Denote the functionals

$$\begin{aligned} &X(f;S,S',\{\xi_i\}_{i=1}^n;\delta) \\ :=&\frac{1}{n}\sum_{i=1}^{n}(\ell(f;z_i') - \ell(f;z_i)) - C_2 \left(\inf_{\alpha \geq 0} \left\{ \alpha + \frac{1}{\sqrt{n}} \int_{\alpha}^{\sqrt{2}} \sqrt{\frac{1}{\pi_{(S,S')}\big(B_{\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}} d\varepsilon \right\} + \sqrt{\frac{\log \frac{\log(2n)}{\delta}}{n}} \right), \end{aligned}$$

and

$$\begin{aligned} &Y(f;S,S',\{\xi_i\}_{i=1}^n;\delta) \\ :=&\frac{1}{n}\sum_{i=1}^{n}\xi_i(\ell(f;z_i') - \ell(f;z_i)) - C_2 \left(\inf_{\alpha \geq 0} \left\{ \alpha + \frac{1}{\sqrt{n}} \int_{\alpha}^{\sqrt{2}} \sqrt{\frac{1}{\pi_{(S,S')}\big(B_{\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}} d\varepsilon \right\} + \sqrt{\frac{\log \frac{\log(2n)}{\delta}}{n}} \right). \end{aligned}$$

Symmetry argument. We write $\overset{d}{=}$ to denote equality in distribution (i.e., the random variables have the same law, equivalently the same cumulative distribution function). For each $i \in \{1, \ldots, n\}$, let $\tau_i(S,S')$ be the pair obtained by swapping z_i and z_i'. Since $\varrho_{(S,S'),\ell}$ and $\pi_{(S,S')}$ are invariant under $(S,S') \mapsto \tau_i(S,S')$ and $\tau_i(S,S') \overset{d}{=} (S,S')$, we have, for all $t \in \mathbb{R}$,

$$\begin{aligned} &\Pr\big(Y(f;S,S',\xi;\delta) \leq t\big) \\ =&\tfrac{1}{2}\Pr\big(X(f;S,S';\delta) \leq t\big) + \tfrac{1}{2}\Pr\big(X(f;\tau_i(S,S');\delta) \leq t\big) \\ =&\Pr\big(X(f;S,S';\delta) \leq t\big), \end{aligned}$$

i.e. $Y(f;S,S',\xi;\delta) \overset{d}{=} X(f;S,S';\delta)$.

By the conclusion (B.27) in Step 2, for any $\delta \in (0,1)$,

$$\Pr\big(Y(f;S,S',\xi;\delta) \leq 0 \text{ for all } f \in \mathcal{F}\big) \ \geq \ 1-\delta.$$

By equality in distribution, the same high–probability event holds for X. Hence, with probability at least $1-\delta$ (over (S,S')), uniformly over $f \in \mathcal{F}$,

$$\frac{1}{n}\sum_{i=1}^{n}\big(\ell(f;z_i') - \ell(f;z_i)\big) \ \leq \ C_2 \left(\inf_{\alpha \geq 0} \left\{ \alpha + \frac{1}{\sqrt{n}} \int_{\alpha}^{\sqrt{2}} \sqrt{\log \frac{1}{\pi_{(S,S')}\big(B_{\varrho_{(S,S'),\ell}}(f,\varepsilon)\big)}} \, d\varepsilon \right\} + \sqrt{\frac{\log\big(\log(2n)/\delta\big)}{n}} \right),$$

where $\varrho_{(S,S'),\ell}(f_1,f_2)=\sqrt{(\mathbb{P}_S+\mathbb{P}_{S'})\big(\ell(f_1;z)-\ell(f_2;z)\big)^2}$, and $C_2>0$ is an absolute constant.

□

B.5 Proof of Theorem 1

B.6 Proof of Theorem 2

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 13 in the auxiliary lemma part for this classical result. To be specific, by Lemma 13 we have that

$$\begin{aligned}\mathbb{E}_z\left[\sup_{f\in\mathcal{F}}(\mathbb{P}-\mathbb{P}_n)\ell(f;z)\right] &\geq \frac{c_1}{\sqrt{\log n}}\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i(\ell(f;z_i)-\mathbb{E}_z[\ell(f;z)])\right]\\ &\geq \frac{c_1}{\sqrt{\log n}}\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)-\left|\frac{1}{n}\sum_{i=1}^n g_i\right|\cdot\sup_{\mathcal{F}}\mathbb{E}[\ell(f;z)]\right]\\ &= \frac{c_1}{\sqrt{\log n}}\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)\right]-\frac{c_1}{\sqrt{\log n}}\sqrt{\frac{2}{\pi n}}\sup_{\mathcal{F}}\mathbb{E}[\ell(f;z)],\end{aligned}\tag{B.28}$$

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 10 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 $\{z_i\}_{i=1}^n$,

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

taking expectation on both side yields

$$\mathbb{E}_{g,z}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n g_i\ell(f;z_i)\right]\geq c_2\mathbb{E}\inf_{\pi}\sup_{f\in\mathcal{F}}\int_0^1\sqrt{\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f,\varepsilon))}}d\varepsilon.\tag{B.29}$$

Combining (B.28) and (B.29), we have that there exist absolute constants $c,c'>0$ such that

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

This inequality implies the following result

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

where we have used the facts that $-\inf_x h(x)=\sup_x(-h(x))$ and $\sup_x h_1(x)-\sup_x h_2(x)\leq\sup_x(h_1(x)-h_2(x))$.

□

B.7 Background on Gaussian and Empirical Processes

It is now well understood that the supremum of Gaussian process can be tightly characterized by the majorizing measure integral via matching upper and lower bounds up to absolute constants [Fernique, 1975, Talagrand, 1987]; the goal of this section is to extend this characterization to (1) bounded empirical processes and (2) a truncated form of integral.

Background on Gaussian Processes. We begin by recalling 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 [Fernique, 1975, Talagrand, 1987]. Note that generic chaining have several equivament formulations [Talagrand, 2005], and the one closest to our purpose is through majorizing measure.

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. On the index set T, consider the semi-metric ϱ given by

$$\varrho(u,v) = \sqrt{\mathbb{E}[(X_u - X_v)^2]}. \tag{B.30}$$

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 integral upper bound.

Lemma 9 (Upper Bound of Gaussian Processes via Majorizing Measure, Fernique [1975]) *Given a Gaussian process $(X_t)_{t\in T}$ with its metric ϱ defined by (B.30), we have*

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

where $C > 0$ is an absolute constant.

A prior π that makes the right hand side in Lemma 9 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 9 is tight up to absolute constants; the upper bound in Lemma 9 is thus called the Fernique-Talagrand (majorizing measure) integral.

Lemma 10 (Lower Bound of Gaussian Processes via Majorzing Measure, Talagrand [1987]) *Given a Gaussian process $(X_t)_{t\in T}$ with its metric ϱ defined by (B.30), we have*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \ge c \inf_{\pi\in\Delta(T)} \sup_{t\in T} \int_0^{\infty} \sqrt{\log \frac{1}{\pi(B_{\varrho_2}(t,\varepsilon))}}\, d\varepsilon,$$

where $c > 0$ is an absolute constant.

Thus the Fernique-Talagrand integral gives a complete characterization to the supremum of Gaussian process.

Background on Empirical Processes. We now give several results on upper and lower bounding empirical process by Rademacher and Gaussian complexities Giné and Zinn [1984], Bartlett and Mendelson [2002].

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 $\{\xi_i\}_{i=1}^n$ are i.i.d. Rademacher variables; and define the Gaussian complexity of $\mathcal{F}$ as

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

where $\{g_i\}_{i=1}^n$ are i.i.d. standard Gaussian variables.

It is well-known that Rademacher and Gaussian complexities are upper bounds of empirical processes (see, e.g., Lemma 7.4 in Van Handel [2014]):

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*

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

where $R_n(\mathcal{F})$ and $G_n(\mathcal{F})$ are (expected) Rademacher and Gaussian complexities defined in Definition 2.

We state a truncated form of the Fernique-Talagrand integral, adapted from Theorem 3 of Block et al. [2021], and use it in the proof of Theorem 1. Up to absolute constants, this truncated form is equivalent to the classical (nontruncated) Fernique-Talagrand integral; throughout, we interpret both forms as placing the $\inf_\pi$ and $\sup_{f\in\mathcal{F}}$ outside the integral.[3] The truncated variant is often more convenient for deriving tighter relaxations—for example, when fixing a particular prior π rather than taking $\inf_\pi$, as used in Theorem 1.

Lemma 12 (Truncated integral bound) *Given a function class $\mathcal{F}$ that consists of mappings from $\mathcal{Z}$ to $[0,1]$. Define the empirical $L_2(\mathbb{P}_n)$ pseudometric*

$$\rho_n(f_1,f_2) \;:=\; \sqrt{\frac{1}{n}\sum_{i=1}^n (f_1(z_i)-f_2(z_i))^2}.$$

[3] Sketch: for the γ_2 functional, one may cap the chaining diameter at 1 at any scale $\alpha\in(0,1]$, absorbing finer scales into an additive α term. By the standard equivalences among the γ_2 functional, admissible trees, and the Fernique–Talagrand integral (see §6.2 of Talagrand [2014]), the truncated and nontruncated forms are equivalent up to absolute constants.

There exists an absolute constant $C > 0$ such that

$$\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \xi_i f(z_i)\right] \;\le\; C\,\inf_{\alpha\ge 0}\left\{\alpha\;+\;\frac{1}{\sqrt{n}}\inf_{\pi\in\Delta(\mathcal{F})}\sup_{f\in\mathcal{F}}\int_\alpha^1\sqrt{\log\frac{1}{\pi\big(B_{\rho_n}(f,\varepsilon)\big)}}\,d\varepsilon\right\},$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. Rademacher variables, and the left hand side of the above inequality is called the empirical Rademacher complexity.

Remarks. (i) Because $f \in [0,1]$, the diameter of $\mathcal{F}$ with ρ_n is bounded by 1, which justifies truncating the integral at 1 and adding the small–scale term α. (ii) An analogous bound holds for Gaussian processes; we state the Rademacher version since it directly controls empirical processes via symmetrization and is what we need for Theorem 1. (iii) The proof of Lemma 12 is a straightforward adaptation of Theorem 3 in Block et al. [2021], specializing their sequential argument to the classical i.i.d. setting (with only minor notational changes).

The following result illustrate that Gaussian and Rademacher complexities can also be used to lower bounding empirical processes.

Lemma 13 (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 13: 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].

□

Basic Concentration Inequalities. We state Mcdiarmid's inequality, Hoeffding's inquality, and Bernstein's inequality.

Lemma 14 (McDiarmid's inequality (bounded differences), McDiarmid [1998]) *Let $Z_1, \ldots, Z_n$ be independent random variables with $Z_i \in \mathcal{Z}_i$. Let $h : \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n \to \mathbb{R}$ be a measurable function satisfying the bounded difference property: there are constants $c_1, \ldots, c_n \ge 0$ such that for all $i \in \{1, \cdots, n\}$ and all $Z_1 \in \mathcal{Z}_1, \cdots, Z_n \in \mathcal{Z}_n$,*

$$\sup_{Z_i'\in\mathcal{Z}_i}\left|h(Z_1,\cdots,Z_{i-1},Z_i,Z_{i+1},\cdots,Z_n) - h(Z_1,\cdots,Z_{i-1},Z_i',Z_{i+1},\cdots,Z_n)\right| \;\le\; c_i.$$

Then for every $t \ge 0$,

$$\Pr\big(h(Z_1,\cdots,Z_n) - \mathbb{E}[h(Z_1,\cdots,Z_n)] \ge t\big) \;\le\; \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma 15 (Hoeffding's inequality, Chapter 2 in Vershynin [2018]) *Let $Z_1, \cdots, Z_n$ be independent random variables with $a_i \leq Z_i \leq b_i$ almost surely. Then for every $t \geq 0$,*

$$\Pr\left(\sum_{i=1}^{n} Z_i - \mathbb{E}[Z] \geq t\right) \;\leq\; \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Lemma 16 (Bernstein's inequality, Chapter 2 in Vershynin [2018]) *Let $Z_1, \cdots, Z_n$ be independent mean–zero random variables with $|Z_i| \leq M$ almost surely. Then for every $t \geq 0$,*

$$\Pr\left(\sum_{i=1}^{n} Z_i \geq t\right) \;\leq\; \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^{n}\mathbb{E}[Z_i^2] + \frac{1}{3}Mt}\right).$$

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

C.1 Proof of Lemma 1 (Non-Perturbative Feature Expansion)

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.

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

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

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

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

$$\begin{aligned}
&\|\sigma_L(W_L' \cdots W_{l+1}'\sigma_l(W_l'F_{l-1}(W,X))) - \sigma_L(W_L' \cdots W_{l+1}'\sigma_l(W_lF_{l-1}(W,X)))\|_{\mathsf{F}}\\
\leq& M_{l\to L}[W,\varepsilon]\|\sigma_l(W_l'F_{l-1}(W,X)) - \sigma_l(W_lF_{l-1}(W,X))\|_{\mathsf{F}}. && \text{(C.3)}
\end{aligned}$$

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

$$\|\sigma_l(W_l'F_{l-1}(W,X)) - \sigma_l(W_lF_{l-1}(W,X))\|_{\mathsf{F}} \leq \|(W_l' - W_l)F_{t-1}(W,X)\|_{\mathsf{F}}. \qquad \text{(C.4)}$$

Combining (C.1) (C.3) and (C.4), 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 17.

Lemma 17 (Metric Domination Lemma) *For two metrics ϱ_1, ϱ_2 defined on $\mathcal{W}$, if $\varrho_1(W', W) \leq \varrho_2(W', W)$ for all $W' \in B_{\varrho_2}(W, \varepsilon)$, then for any prior $\pi \in \Delta(\mathcal{W})$ 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 17: Because $\varrho_1(W', W) \leq \varrho_2(W', W)$ 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

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

this implies

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

□

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 18 (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 $W' \in \mathcal{V}$,*

$$(\varrho_1(W', W))^2 \leq (\varrho_2(W', \mathcal{P}_{\mathcal{V}}(W)))^2 + \varepsilon_1^2, \tag{C.5}$$

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

$$\log \frac{1}{\pi(B_{\varrho_1}(W, \varepsilon))} \leq \log \frac{1}{\pi(B_{\varrho_2}(\mathcal{P}_{\mathcal{V}}(W), \sqrt{\varepsilon^2 - \varepsilon_1^2}))}. \tag{C.6}$$

Proof of Lemma 18: By the condition (C.5), 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 (C.6) in Lemma 18.

□

C.3 Pointwise Dimension Bound with Reference Subspace

Set Up of Reference Effective Subspace Consider the weight space $B_2(R) \subset \mathbb{R}^p$ for vectorized weights W, where $B_2(R) := \{w \in \mathbb{R}^p : \|w\|_2 \leq R\}$. 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 $\mathcal{V}_{\text{eff}}(G(W)), R, \varepsilon)$ 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 $r_{\text{eff}}(G(W), R, \varepsilon)$, and $\mathcal{V}$ as an abbreviation of $\mathcal{V}_{\text{eff}}(G(W)), R, \varepsilon)$ 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 (D.6)), which is consistent with the earlier operator notation characterized by ℓ_2–distance in Lemma 18. Now define the projected metric $\varrho^{\bar{\mathcal{V}}}_{G(W)}$ as

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

By the subspace metric dominance lemma (Lemma 18), 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:

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

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 $\mathcal{V}_{\text{eff}}(G(W), R, \varepsilon)$). 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 $\pi_{\bar{\mathcal{V}}} = \text{Unif}\big(B_2(1.58R) \cap \bar{\mathcal{V}}\big)$. 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

$$\varrho_{\text{proj},G(W)}(\mathcal{V}, \bar{\mathcal{V}}) := \big\| G(W)^{1/2}\big(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}\big) \big\|_{\text{op}} \leq \tfrac{\sqrt{n}\varepsilon}{4R}, \tag{C.7}$$

we have

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

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

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

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

$$\begin{aligned} G(W) =& \mathcal{P}_{\mathcal{V}}^\top G(W)\mathcal{P}_{\mathcal{V}} + \mathcal{P}_{\mathcal{V}_\perp}^\top G(W)\mathcal{P}_{\mathcal{V}_\perp} \\ \preceq& \mathcal{P}_{\mathcal{V}}^\top G(W)\mathcal{P}_{\mathcal{V}} + \lambda_{r_{\text{eff}}+1} \cdot \mathcal{P}_{\mathcal{V}_\perp}^\top \mathcal{P}_{\mathcal{V}_\perp}, \end{aligned} \tag{C.8}$$

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

$$\mathcal{P}_{\mathcal{V}}^\top G(W)\mathcal{P}_{\mathcal{V}} \preceq 2\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}} + 2(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})^\top G(W)(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}). \tag{C.9}$$

Combining (C.8) and (C.9), we have the fundamental loewner order inequality

$$G(W) \preceq 2\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}} + 2(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})^\top G(W)(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}) + \lambda_{r_{\text{eff}}+1} \cdot \mathcal{P}_{\mathcal{V}_\perp}^\top \mathcal{P}_{\mathcal{V}_\perp}. \tag{C.10}$$

In order to apply the subspace metric domination lemma (Lemma 18), we hope to bound $\|W' - W\|_2^2$ and apply that bound to the two last reminder terms in the right hand side of (C.10).

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

$$\begin{aligned} \|W' - \mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2^2 \leq& \frac{(W' - \mathcal{P}_{\bar{\mathcal{V}}}(W))^\top \mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}}(W' - \mathcal{P}_{\bar{\mathcal{V}}}(W))}{\lambda_{r_{\text{eff}}}(\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})} \leq \frac{n\varepsilon^2}{16\lambda_{r_{\text{eff}}}(\mathcal{P}_{\bar{\mathcal{V}}}^\top G(W)\mathcal{P}_{\bar{\mathcal{V}}})} \\ \leq& \frac{n\varepsilon^2}{8\lambda_{r_{\text{eff}}} - 16\|G(W)^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}}^2} \leq \frac{1}{3}R^2, \end{aligned} \tag{C.11}$$

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 second inequality used the condition of $W' \in B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$; the third inequality uses Lemma 19; and the last inequality uses $\lambda_{r_{\text{eff}}} \geq \frac{n\varepsilon^2}{2R^2}$

(by definition (3.3) of effective rank) and the approximation error condition (C.7). On the other hand, we have that $\|W\|_2^2 \le R^2$, so that for every $W' \in B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$

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

combined with (C.11).

From the fundamental loewner order inequality (C.10), we establish the desired metric domination condition: for all $W' \in B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$ and $W \in B_2(R)$,

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

where the first inequality holds because of the loewner order inequality (C.10) and the property of operator norm: $x^\top Ax \le \|A\|_{\mathrm{op}} \cdot \|x\|_2^2$ (one could also apply Lemma 19 to validate $\|\mathcal{P}_{\mathcal{V}_\perp}^\top \mathcal{P}_{\mathcal{V}_\perp}\|_{\mathrm{op}} \le 1$); and the last inequality uses the fact $\lambda_{r_{\mathrm{eff}}+1} < \frac{n\varepsilon^2}{2R^2}$ (by definition 3.3 of effective rank) and the approximation error condition (C.7). Now we can apply the subspace metric domination lemma (Lemma 18) and obtain: for any $\pi \in \Delta(\bar{\mathcal{V}})$,

$$\log \frac{1}{\pi\big(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon)\big)} \le \log \frac{1}{\pi\big(B_{\sqrt{2}\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/\sqrt{6})\big)} \le \log \frac{1}{\pi\big(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)\big)}. \tag{C.12}$$

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 $W' \in B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$, (C.11) suggests $\|W' - \mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2^2 \le \frac{1}{3}R^2$, and 2) for very $W \in B_2(R)$, we have $\|\mathcal{P}_{\bar{\mathcal{V}}}(W)\|_2 \le \|W\|_2 \le R$. Combining this and the above inequality we have

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

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

$$\log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)} = \frac{\mathrm{Vol}(\bar{\mathcal{V}} \cap B_2(1.58R))}{\mathrm{Vol}(B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4))}. \tag{C.13}$$

By the change–of–variables theorem in multivariate calculus [Wikipedia contributors, 2025a], the linear map $T = G(W)^{\frac{1}{2}}$ implies the volume formula for ellipsoid $E = B_{\varrho^{\bar{\mathcal{V}}}_{G(W)}}(\mathcal{P}_{\bar{\mathcal{V}}}(W), \sqrt{n}\varepsilon/4)$ 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 $\sqrt{n}\varepsilon/4$

$$\mathrm{Vol}\big(E\big) \;=\; |\det T|^{-1}\,\mathrm{Vol}\big(T(E)\big) \;=\; (\det G(W))^{-1/2}\,\mathrm{Vol}\big(B_2(\sqrt{n}\varepsilon/4)\big) = \left(\prod_{k=1}^{r_{\mathrm{eff}}} \lambda_k\right)^{-1/2} \mathrm{Vol}\big(B_2(\sqrt{n}\varepsilon/4)\big),$$

Also by the change-of-variable theorem, we have that the volume of $r_{\text{eff}}-$dimensional isotropic ball $\mathcal{V} \cap B_2(2R)$ is

$$\text{Vol}(\bar{\mathcal{V}} \cap B_2(1.58R)) \;=\; \left(\frac{1.58R}{\sqrt{n}\varepsilon/4}\right)^{r_{\text{eff}}} \text{Vol}(B_2(\sqrt{n}\varepsilon/4)).$$

Hence, applying (C.12) (C.13) and combining it with the two above volume equalities, we have

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

Finally, since the prior construction $\pi_{\bar{\mathcal{V}}} = \text{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}})$ 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 (C.7),

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

which is the claimed bound. □

Proof of Lemma 19: 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 $\lambda_k(\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}) \le \lambda_k$ 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,

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

where the first inequality is because for every fixed S and W we have $W^\top(\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}})W \le W^\top(2\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}})W + \|\mathcal{P}_{\mathcal{V}}^\top \Sigma \mathcal{P}_{\mathcal{V}} - 2\mathcal{P}_{\bar{\mathcal{V}}}^\top \Sigma \mathcal{P}_{\bar{\mathcal{V}}}\|_{\mathrm{op}} \|W\|_2^2$; and the last inequality is due to (C.9). Therefore we have

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

□

C.4 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 (3.2) 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 (3.2); $G_{\mathrm{NP}}(W)$ 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.4.1 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 20 (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 20. It is straightforward to see that, first, the effective rank of the fixed matrix $G(W)$ is

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

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.4.2 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 $\|W\|_{\mathsf{F}}$. 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 4.

Lemma 21 (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 $W \in B_{\mathsf{F}}(R)$,*

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

where $C > 0$ is an absolute constant.

Proof of Lemma 21: 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 (3.6), 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) $\mathcal{V}_{\mathrm{eff}}(G(W), R, \varepsilon)$ to the precision required by (C.7):

$$\|G(W)^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\mathrm{op}} \leq \frac{\sqrt{n}\varepsilon}{4R}, \tag{C.14}$$

where $G(W)^{\frac{1}{2}}$ is the unique square root of PSD matrix $G(W)$ (see, e.g, [Wikipedia contributors, 2025d]). Then by Lemma 2 (Pointwise Dimension Bound for Nonlinear Manifold with Approximate Effective Subspace), for every $(W, \varepsilon) \in B_2(R) \times [0, \infty)$ such that $\bar{\mathcal{V}}$ approximates $\mathcal{V}_{\mathrm{eff}}(G(W), R, \varepsilon)$ to the precision (C.14), the prior $\pi_{\bar{\mathcal{V}}} = \mathrm{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}})$ satisfies

$$\log \frac{1}{\pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \leq d_{\mathrm{eff}}(G(W), \sqrt{5}R, \varepsilon) = \sum_{l=1}^{L} d_l \cdot d_{\mathrm{eff}}(A_l(W), \sqrt{5}R, \varepsilon), \tag{C.15}$$

where the first inequality is by Lemma 2 (see definition (3.4) of effective dimension); and the last equality is by the decomposition property of NN-surrogate metric tensor (Lemma 20).

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

$$\pi(W) = \sum_{\mathcal{V}} \mu(\mathcal{V}) \times \pi_{\mathcal{V}}(W), \tag{C.16}$$

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

$$\begin{aligned}
&\log \frac{1}{\pi(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
=&\log \frac{1}{\sum_{\mathcal{V}} \mu(\mathcal{V})\pi_{\mathcal{V}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
\leq&\log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies (C.14)}) \inf_{\bar{\mathcal{V}} \text{ satisfies (C.14)}} \pi_{\bar{\mathcal{V}}}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
\leq&\underbrace{\log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies (C.14)})}}_{\text{covering cost of the Grassmannian}} + \sum_{l=1}^{L} d_l \cdot d_{\text{eff}}(A_l(W), \sqrt{5}R, \varepsilon),
\end{aligned} \tag{C.17}$$

where the first equality is by definition (C.16) of the “universal” prior π; the first inequality is straightforward; and the last inequality is by (C.15) (the result of the “must pay” part in the hierarchical covering) and the equivalence between $B_2(R)$ and $B_{\mathsf{F}}(R)$.

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 *Grassmannian manifold.*

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

$$\varrho_{\text{proj},\Sigma}(\mathcal{V}, \bar{\mathcal{V}}) = \|\Sigma^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\text{op}}, \tag{C.18}$$

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

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

$$\log \frac{1}{\mu(B_{\varrho_{\text{proj},\Sigma}}(\mathcal{V}, \varepsilon_1))} \leq \frac{d-r}{2} \sum_{k=1}^{r} \log \frac{C_0 \max\{\lambda_k, \varepsilon_1^2\}}{\varepsilon_1^2} + \frac{r}{2} \sum_{k=1}^{d-r} \log \frac{C_0 \max\{\lambda_k, \varepsilon_1^2\}}{\varepsilon_1^2}, \tag{C.19}$$

where $C_0 > 0$ is an absolute constant. We will use the result (C.19) and (C.17) 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 (C.19) 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 $r_{\text{eff}}(A_l(W), R, \varepsilon_1) = r_l$, and

$\lambda_{r_l+1}(A_l(W)) \le c\varepsilon_1^2 \le \lambda_{r_l}(A_l(W))$ where $c \ge 1$ can be any absolute constants no smaller than 1 (later we will specialize to $c = 8$),

$$\log \frac{1}{\mu_l(\bar{\mathcal{V}} : \varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon), \bar{\mathcal{V}}) \le \varepsilon_1)} \le \frac{d_{l-1}}{2} \sum_{k=1}^{r_l} \log \frac{C_1 \lambda_k(A_l(W))}{\varepsilon_1^2}, \tag{C.20}$$

where $C_1 = c \max\{C_0, 1\} \ge 1$ 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). (C.20) 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 (C.19). Their cumulative effect is at most

$$\mathbb{1}\{d_{l-1} - r_l > r_l\} \cdot \frac{r_l}{2} \sum_{k=r_l+1}^{d_{l-1}-r_l} \log \frac{C_0 c \varepsilon_1^2}{\varepsilon_1^2} = \frac{r_l \max\{d_{l-1} - 2r_l, 0\}}{2} \log C_0 c \le \frac{r_l(d_{l-1} - r_l)}{2} \log C_0 c$$

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

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

Summing up the contributions two parts of the spectrum together, we get the right hand side of (C.20).

By the subspace decomposition property in Lemma 20, we have that for $\bar{\mathcal{V}} = (\cdots, \underbrace{\bar{\mathcal{V}}_l, \cdots, \bar{\mathcal{V}}_l}_{\text{repeat } d_l \text{ times}}, \cdots)$,

$$\begin{aligned} &\varrho_{\mathrm{proj},G(W)}(\mathcal{V}_{\mathrm{eff}}(G(W), R, \varepsilon), \bar{\mathcal{V}}) \\ =&\varrho_{\mathrm{proj},G(W)}(\prod_{l=1}^{L} \mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon)^{d_l}, \prod_{l=1}^{L} \bar{\mathcal{V}}_l^{d_l}) \\ =&\max_l \varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W), R, \varepsilon), \bar{\mathcal{V}}_l), \end{aligned} \tag{C.21}$$

where the first equality is by Lemma 20, 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 (3.3) 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 (C.20) with $c = 8$. Then for all layers $l = 1, \cdots, L$, given a fixed $\{r_1, \cdots, r_L\}$, by (C.20), we have that there exists a prior

$$\mu_{\{r_l\}_{l=1}^L} = \mu_1^{d_1} \otimes \cdots \otimes \mu_L^{d_L} = \prod_{l=1}^{L} (\underbrace{\mu_l \otimes \cdots \otimes \mu_l}_{d_l \text{ times}}) \tag{C.22}$$

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 (C.17) is bounded by

$$
\begin{aligned}
&\log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies (C.14)})} \\
=&\log \frac{1}{\mu_{\{r_l\}_{l=1}^L}(\bar{\mathcal{V}} : \varrho_{\mathrm{proj},G(W)}(\mathcal{V}_{\mathrm{eff}}(G(W),R,\varepsilon),\bar{\mathcal{V}}) \leq \frac{\sqrt{n}\varepsilon}{4R} = \varepsilon_1)} \\
\leq&\log \frac{1}{\mu_{\{r_l\}_{l=1}^L}((\cdots,\underbrace{\bar{\mathcal{V}}_l,\cdots,\bar{\mathcal{V}}_l}_{d_l \text{ times}},\cdots) : \varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W),R,\varepsilon),\bar{\mathcal{V}}_l) \leq \varepsilon_1, \quad \forall l \in [L])} \\
=&\sum_{l=1}^L \log \frac{1}{\mu_{\{r_l\}_{l=1}^L}(\cdots,\bar{\mathcal{V}}_l,\cdots) : \varrho_{\mathrm{proj},A_l(W)}(\mathcal{V}_{\mathrm{eff}}(A_l(W),R,\varepsilon),\bar{\mathcal{V}}_l) \leq \varepsilon_1)} \\
\leq&\sum_{l=1}^L \frac{d_{l-1}}{2} \sum_{k=1}^{r_l} \log \frac{C_1 \lambda_k(A_l(W))}{\varepsilon_1^2} \\
\leq&\sum_{l=1}^L d_{l-1} d_{\mathrm{eff}}(A_l(W),\sqrt{2C_1}R,\varepsilon),
\end{aligned}
\tag{C.23}
$$

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 (C.21); the second equality is by the choice of the product prior (C.22); the second inequality is by the layer-wise covering bound (C.20); and the last inequality is by the choice $\varepsilon_1 = \sqrt{n}\varepsilon/(4R)$, and definition (3.4) of effective dimension.

Note that (C.23) is uniformly over all $W \in B_{\mathsf{F}}(R)$ such that $r_{\mathrm{eff}}(A_l(W),R,\varepsilon) = r_l$, $\forall l \in [L]$, not uniformly over all $W \in B_{\mathsf{F}}(R)$. We would like to extend (C.23) 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 (C.16)) defined by

$$
\begin{aligned}
\mu(\mathcal{V}) &= \prod_{l=1}^L \underbrace{\mathrm{Unif}([d_{l-1}])}_{\text{prior of } r_l} \otimes \underbrace{\mu_{\{r_k\}_{k=1}^L}}_{\text{prior over product Grassmannian in (C.22)}}, \\
\pi(W) &= \sum_{\mathcal{V}} \underbrace{\mu(\mathcal{V})}_{\text{prior over subspaces defined above}} \otimes \underbrace{\mathrm{Unif}(B_2(1.58R) \cap \bar{\mathcal{V}})}_{\text{uniform prior constrained in subspace}}.
\end{aligned}
\tag{C.24}
$$

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

$$\begin{aligned}
&\log \frac{1}{\pi(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
\leq &\log \frac{1}{\mu(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies (C.14)})} + \sum_{l=1}^{L} d_l \cdot d_{\mathrm{eff}}(A_l(W), \sqrt{5}R, \varepsilon)) \\
\leq &\sum_{l=1}^{L} \log d_{l-1} + \log \frac{1}{\mu_{\{r_k\}_{k=1}^{L}}(\bar{\mathcal{V}} : \bar{\mathcal{V}} \text{ satisfies (C.14)})} + \sum_{l=1}^{L} d_l \cdot d_{\mathrm{eff}}(A_l(W), \sqrt{5}R, \varepsilon)) \\
\leq &\sum_{l=1}^{L} \log d_{l-1} + \sum_{l=1}^{L} d_{l-1} \cdot d_{\mathrm{eff}}(A_l(W), \sqrt{2C_1}R, \varepsilon) + \sum_{l=1}^{L} d_l \cdot d_{\mathrm{eff}}(A_l(W), \sqrt{5}R, \varepsilon)),
\end{aligned}$$

where $C_1 > 0$ is an absolute constant. Here the first inequality is by the hierarchical covering argument (C.17); the second inequality is by the prior construction (C.24); and the third inequality is by the Grassmannian covering bound (C.23) 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:

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

where C is a positive absolute constant. This finishes the proof of Lemma 21 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 4), 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

$$\tilde{\pi} = \underbrace{\mathrm{Unif}(\{R_0, \cdots, 2^{\log_2\lceil R/R_0 \rceil} R_0\})}_{\text{prior over upper bound } \tilde{R} \text{ of } \|W\|_{\mathsf{F}}} \otimes \underbrace{\pi_{\tilde{R}}}_{\text{prior defined via (C.24)}},$$

where $\pi_{\tilde{R}}$ is the prior defined via (C.24) in the proof of Lemma 21. Then for every $W \in B_{\mathsf{F}}(R)$ where $\|W\|_{\mathsf{F}} > R_0$, denote $k(W)$ to be the integer such that $2^{k(W)} R_0 < \|W\|_{\mathsf{F}} \leq 2^{k(W)+1} R_0$, then

$$\begin{aligned}
&\log \frac{1}{\tilde{\pi}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
\leq &\underbrace{\log \log_2 \lceil R/R_0 \rceil}_{\text{density of } 2^{k(W)+1} R_0} + \underbrace{\log \frac{1}{\pi_{2^{k(W)+1} R_0}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))}}_{\pi \text{ is constructed via (C.24), with global radius taken to be } 2^{k(W)+1} R_0} \\
\leq &\log \log_2 \lceil R/R_0 \rceil + \sum_{l=1}^{L} ((d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(A_l(W), C_1 2^{k(W)+1} R_0, \varepsilon) + \log d_{l-1}) \\
\leq &\log \log_2 \lceil R/R_0 \rceil + \sum_{l=1}^{L} ((d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(A_l(W), C_1 \cdot 2\|W\|_{\mathsf{F}}, \varepsilon) + \log d_{l-1}),
\end{aligned}$$

where the first inequality is due to the product construction of $\tilde{\pi}$; the second inequality is due to Lemma 21, with $C_1 > 0$ being an absolute constant; and the last inequality uses the fact $\|W\|_{\mathsf{F}} \le 2^{k(W)+1} R_0 \le 2\|W\|_{\mathsf{F}}$, with $C_1 > 0$.

The above bound assumes $\|W\|_{\mathsf{F}} > R_0$. When $\|W\|_{\mathsf{F}} \le R_0$, we directly apply Lemma 21 and obtain

$$\begin{aligned}
&\log \frac{1}{\tilde{\pi}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))} \\
&\le \underbrace{\log \log_2 \lceil R/R_0 \rceil}_{\text{density of } R_0} + \underbrace{\log \frac{1}{\pi_{R_0}(B_{\varrho_{G(W)}}(W, \sqrt{n}\varepsilon))}}_{\pi \text{ is constructed via (C.24), with global radius taken to be } R_0} \\
&\le \log \log_2 \lceil R/R_0 \rceil + \sum_{l=1}^{L} (d_l + d_{l-1}) \cdot d_{\text{eff}}(A_l(W), C_1 \cdot R_0, \varepsilon) + \log d_{l-1}).
\end{aligned}$$

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

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

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

Finally, by the sentence below (3.2) (which is a straightforward result from non-perturbative feature expansion for DNN (Lemma 1) and the metric domination lemma (Lemma 17)), we know that there exists a prior $\tilde{\pi}$ such that uniformly over all $W \in B_{\mathsf{F}}(R)$,

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

where $A_l(W) = L M^2_{l \to L}(W, \varepsilon) \cdot F_{l-1}(W, X) F^{\top}_{l-1}(W, X)$ when taking $G(W)$ to be $G_{\mathrm{NP}}(W)$ defined in (3.2). 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 (D.6)), consistent with the earlier operator notation characterized by ℓ_2–distance in Lemma 18. In the isotropic case $\Sigma = I_d$, the ellipsoidal projection metric reduces to the standard isotropic projection metric

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

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

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

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

where $C > 0$ is an absolute constant.

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

$$\left(\frac{c}{\varepsilon}\right)^{r(d-r)} \leq \text{N}(\text{Gr}(d,r), \varrho_{\text{proj}}, \varepsilon) \leq \left(\frac{C}{\varepsilon}\right)^{r(d-r)}. \tag{D.2}$$

Here $\text{N}(\mathcal{F}, \varrho, \varepsilon)$ 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 (D.2) 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.

Effective Rank vs. Full-Spectrum Complexity. Consider a covariance matrix Σ with eigenvalues $\lambda_1 \geq \cdots \lambda_d \geq 0$. By Definition 4, the ellipsoidal metric satisfies

$$\varrho_{\mathrm{proj},\Sigma}(\mathcal{V}, \bar{\mathcal{V}}) \leq \lambda_1^{\frac{1}{2}} \varrho_{\mathrm{proj}}(\mathcal{V}, \bar{\mathcal{V}}).$$

If one is willing to accept a coarser complexity scaling, then one could invoke existing Grassmannian covering results under the canonical isotropic metric (D.2) (taking $\mu = \mathrm{Unif}(\mathrm{Gr}(d,r))$) and obtain

$$\log \frac{1}{\mu\big(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon)\big)} \;\leq\; \log \frac{1}{\mu\big(B_{\varrho_{\mathrm{proj}}}(\mathcal{V},\varepsilon/\sqrt{\lambda_1})\big)} \;\leq\; \big(d - r_{\mathrm{eff}}(\Sigma, R, \varepsilon)\big)\, r_{\mathrm{eff}}(\Sigma, R, \varepsilon)\, \log \frac{C\lambda_1}{\varepsilon^2}. \tag{D.3}$$

However, this makes the *global atlas* cost dominate the *local chart* cost, yielding a suboptimal bound than the full–spectrum effective dimension in (D.1). The refined analysis in this section—also simplifying and strengthening the isotropic route—establishes the correct structural principle: the global–atlas cost must be balanced by the local–chart cost. Thus, while the effective–rank bound (D.3) serves as a useful sanity check, the full–spectrum treatment is what delivers the sharpened complexities required for our main results.

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

Fix integers $r \leq d$. 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)

$$\mathrm{Gr}(d,r) \;\cong\; O(d)\big/\big(O(r)\times O(d-r)\big) \;\cong\; \mathrm{Gr}\big(d,\, d-r\big), \tag{D.4}$$

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

$$\underbrace{\frac{d(d-1)}{2}}_{\dim O(d)} - \underbrace{\frac{r(r-1)}{2}}_{\dim O(r)} - \underbrace{\frac{(d-r)(d-r-1)}{2}}_{\dim O(d-r)} = r(d-r), \tag{D.5}$$

   which also appears as the dimension factor in the precise covering-number bounds (D.2).

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[4]

$$\mathcal{P}_{\mathcal{V}} := VV^\top \quad \in \mathbb{R}^{d\times d}. \tag{D.6}$$

Then $\mathcal{P}_{\mathcal{V}}$ 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 $\mathcal{V}, \bar{\mathcal{V}} \in \mathrm{Gr}(d,r)$,

$$\varrho_{\mathrm{proj},\Sigma}(\mathcal{V}, \bar{\mathcal{V}}) \ := \ \|\Sigma^{\frac{1}{2}}(\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}})\|_{\mathrm{op}}, \tag{D.7}$$

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

$$\varrho_{\mathrm{proj},\Sigma}(V, \bar{V}) \ := \ \varrho_{\mathrm{proj},\Sigma}\big(\mathrm{span}(V), \mathrm{span}(\bar{V})\big) \ = \ \|\Sigma^{\frac{1}{2}}(VV^\top - \bar{V}\bar{V}^\top)\|_{\mathrm{op}}. \tag{D.8}$$

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

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

[4] By elementary linear algebra, the matrix definition of the orthogonal projector $\mathcal{P}$ here coincides with the ℓ_2-projection characterized in Lemma 18; 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

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

define the four blocks

$$\underbrace{C}_{r\times r} \;=\; V^\top \bar{V}, \qquad \underbrace{C_\perp}_{r\times(d-r)} \;=\; V^\top \bar{V}_\perp, \tag{D.9}$$

$$\underbrace{S}_{(d-r)\times r} \;=\; V_\perp^\top \bar{V}, \qquad \underbrace{S_\perp}_{(d-r)\times(d-r)} \;=\; V_\perp^\top \bar{V}_\perp. \tag{D.10}$$

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

$$C = V^\top \bar{V} \;=\; Q_1 \, \mathrm{diag}(\cos\theta_1, \cdots, \cos\theta_r) \, W_1^\top, \quad Q_1, W_1 \in O(r), \tag{D.11}$$

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 $\bar{\mathcal{V}}$; and where $\{\cos\theta_1, \cdots, \cos\theta_r\}$ are the singular values of C. Simultaneously, we have that the eigenvalues of S, $C_\perp$, $S_\perp$ are (notation sepc means spectrum, the set of singular values)

$$\begin{aligned} \mathrm{spec}(S) &= \{\sin\theta_1, \cdots, \sin\theta_{\min\{r,d-r\}}, \underbrace{0, \cdots, 0}_{\max\{d-2r,0\}}\}, \\ \mathrm{spec}(C_\perp) &= \{\sin\theta_1, \cdots, \sin\theta_{\min\{r,d-r\}}, \underbrace{0, \cdots, 0}_{\max\{d-2r,0\}}\} \\ \mathrm{spec}(S_\perp) &= \{\cos\theta_1, \cdots, \cos\theta_{\min\{r,d-r\}}, \underbrace{1, \cdots, 1}_{\max\{d-2r,0\}}\}. \end{aligned} \tag{D.12}$$

The above representation in (D.11) and (D.12) 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 (D.11) and (D.12) motivates our notation C and S when defining block matrices in (D.9) and (D.10). 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

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

It is known that the projection metric defined in (D.7) and (D.8) 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])

$$\varrho_{\text{proj}} = \|\mathcal{P}_{\mathcal{V}} - \mathcal{P}_{\bar{\mathcal{V}}}\|_{\text{op}} \;=\; \max_{1\le k\le r} \sin\theta_k \;=\; \sin\theta_r. \tag{D.13}$$

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 $\bar{\mathcal{V}} \in \text{Gr}(d,r)$ and its parameterization matrix $\bar{V} \in \text{St}(d,r)$. 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*

$$\mathcal{V}(X) \;:=\; \text{span}\Big\{[\bar{V}\bar{V}_\perp]\begin{pmatrix} I_r \\ X\end{pmatrix}\Big\}, \qquad X \;\in\; \mathbb{R}^{(d-r)\times r}, \tag{D.14}$$

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 $\text{Gr}(d,r)$ by

$$\phi_{\bar{\mathcal{V}}}:\; X \;\longmapsto \mathcal{V}(X) \in\; \text{Gr}(d,r). \tag{D.15}$$

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 $\mathcal{P}_{\bar{\mathcal{V}}}$ restricted to $\mathcal{V}$ is a bijection $\mathcal{P}_{\bar{\mathcal{V}}}|\mathcal{V} : \mathcal{V} \to \bar{\mathcal{V}}$. In the orthonormal basis $[\bar{V}\ \bar{V}_\perp]$, this means every $v \in \mathcal{V}$ can be written uniquely as

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

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 22 (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*

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

Proof of Lemma 22: 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 (D.11), $A = \bar{V}^\top V$ 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 that $\phi_{\bar{\mathcal{V}}}(X_1) = \phi_{\bar{\mathcal{V}}}(X_2)$. 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 23 (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 (D.11). For the graph chart (D.15), we have*

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

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 23: Given the fact $\varrho_{\mathrm{proj}}(\mathcal{V}(X), \bar{\mathcal{V}}) = \sin\theta_r$ (which is already shown in (D.13)), 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:

$$\bar{V} = \begin{pmatrix} I_r \\ 0 \end{pmatrix} \in \mathrm{St}(d,r), \qquad \bar{V}_\perp = \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix} \in \mathrm{St}(d, d-r). \tag{D.16}$$

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):

$$V(X) = [\bar{V}\ \bar{V}_\perp] \begin{pmatrix} I_r \\ X \end{pmatrix} (I_r + X^\top X)^{-1/2} = I_d \begin{pmatrix} I_r \\ X \end{pmatrix} (I_r + X^\top X)^{-1/2} = \begin{pmatrix} I_r \\ X \end{pmatrix} (I_r + X^\top X)^{-1/2}, \tag{D.17}$$

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 (D.16).

Step 2: Projection Metric and Principal Angles. A fundamental result in matrix analysis, our equation (D.11), 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 24 (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 24: 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 than 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 $\bar{\mathcal{V}} \in \text{Gr}(d, r)$,

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

Note that ϱ_{proj} is not the target metric; our goal is the ellipsoidal metric $\varrho_{\text{proj},\Sigma}$. Taking $\varepsilon = 1/\sqrt{2}$, we obtain:

$$\log \frac{1}{\nu(B_{\varrho_{\text{proj}}}(\bar{\mathcal{V}}, 1/\sqrt{2}))} \le C_1 r(d-r), \tag{D.18}$$

where $C_1 > 0$ is an absolute constant. By Lemma 23, we have that inside the ball $B_{\varrho_{\text{proj}}}(\bar{\mathcal{V}}, 1/\sqrt{2})$, by choosing $\bar{\mathcal{V}}$ as the reference subspace, the graph parameterization X of $\mathcal{V}$ satisfies

$$\|X\|_{\text{op}} \le 1,$$

which follows from that if $\varrho_{\text{proj}}(\mathcal{V}(X), \bar{\mathcal{V}}) \le 1/\sqrt{2}$ (i.e., $\sin\theta_r \le 1/\sqrt{2}$), we have $\|X\|_{\text{op}} \le 1$. See (D.14) for the definition of this graph chart parameterization; the existence and uniqueness of the parameterization X is by Lemma 22 (local bijection of graph chart). Furthermore, again by Lemma 22 and Lemma 23, $\mathcal{X} = \{X \in \mathbb{R}^{(d-r)r} : \|X\|_{\text{op}} \le 1\}$ satisfies ($\cong$ means isomorphism/bijection)

$$B_{\varrho_{\text{proj}}}(\bar{\mathcal{V}}, 1/\sqrt{2}) \cong \mathcal{X} \subset \bar{\mathcal{X}} \cong B_{\varrho_{\text{proj}}}(\bar{\mathcal{V}}, 2/\sqrt{5}). \tag{D.19}$$

Let

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

Then we have

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

where the first inequality is by restricting $\bar{\mathcal{V}}$ to $B_{\varrho_{\text{proj}}}(\mathcal{V}, 1/\sqrt{2})$; and the second inequality is by (D.18) as well as the bijection stated in (D.19) and Lemma 22. 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 Lipschitz Properties inside Graph Chart

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

Lemma 25 (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 (D.14). 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 25: 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 (D.17) (which, without loss of generality, uses a convenient orthogonal basis specified by (D.16)), 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:

$$\begin{aligned} V(X) =& A(X)M(X)^{1/2}, \\ \mathcal{P}_{\mathcal{V}(X)} =& A(X)M(X)A(X)^\top = A(X)M(X)\begin{pmatrix} I_r & X^\top \end{pmatrix} \quad \text{(D.20)} \\ \mathcal{P}_{\mathcal{V}(X)} - \mathcal{P}_{\mathcal{V}(X')} =& A(X)M(X)A(X)^\top - A(X')M(X')A(X')^\top \\ A(X)M(X) =& \mathcal{P}_{\mathcal{V}(X)} \begin{pmatrix} I_r \\ 0 \end{pmatrix} \quad \text{(D.21)} \\ A(X)M(X)X^\top =& \mathcal{P}_{\mathcal{V}(X)} \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix}, \quad \text{(D.22)} \end{aligned}$$

where (D.21) and (D.22) are straightforward consequences of (D.20).

We begin with a non-perturbative decomposition:

$$\begin{aligned} &\mathcal{P}_{\mathcal{V}(X)} - \mathcal{P}_{\mathcal{V}(X')} \\ =& A(X)M(X)A(X)^\top - A(X')M(X')A(X')^\top \\ =& (A(X) - A(X'))M(X')A(X')^\top + A(X)(M(X) - M(X'))A(X')^\top + A(X)M(X)(A(X) - A(X'))^\top. \end{aligned} \quad \text{(D.23)}$$

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

$$\begin{aligned} &(A(X) - A(X'))M(X')A(X')^\top \\ =& \begin{pmatrix} 0 \\ X - X' \end{pmatrix} M(X')A(X')^\top \\ =& \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix} (X - X')M(X')A(X')^\top \\ =& \begin{pmatrix} 0 \\ I_{d-r} \end{pmatrix} (X - X') \begin{pmatrix} I_r & 0 \end{pmatrix} \mathcal{P}_{\mathcal{V}(X')}, \end{aligned} \quad \text{(D.24)}$$

where the last equality uses the fact (D.21) 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

$$\begin{aligned}
&A(X)(M(X)-M(X'))A(X')^{\top}\\
=&A(X)M(X)X^{\top}(X'-X)M(X')A(X')^{\top}+A(X)M(X)({X'}^{\top}-X^{\top})X'M(X')A(X')^{\top}\\
=&-\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}0\\ I_{d-r}\end{pmatrix}(X-X')\begin{pmatrix}I_r & 0\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')}-\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X^{\top}-{X'}^{\top})\begin{pmatrix}0 & I_{d-r}\end{pmatrix}\mathcal{P}_{\mathcal{V}(X')},
\end{aligned}\tag{D.25}$$

where the last equality uses the fact (D.21) and the fact (D.22).

Third, we have

$$\begin{aligned}
&A(X)M(X)(A(X)-A(X'))^{\top}\\
=&A(X)M(X)\begin{pmatrix}0 & X^{\top}-{X'}^{\top}\end{pmatrix}\\
=&\mathcal{P}_{\mathcal{V}(X)}\begin{pmatrix}I_r\\ 0\end{pmatrix}(X^{\top}-{X'}^{\top})\begin{pmatrix}0 & I_{d-r}\end{pmatrix},
\end{aligned}\tag{D.26}$$

where the last equality uses the fact (D.21).

Substituting (D.24), (D.25), (D.26) back into (D.23), we have

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

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 25, we have the following Lipschitz property of graph chart.

Lemma 26 (Lipschitz 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 (D.17). Then the ellipsoidal projection metric is Lipschitz to ellipsoidal spectral metrics as follows: for every rank-r PSD $\Sigma \in \mathbb{R}^{d\times d}$,*

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

Proof of Lemma 26: By Lemma 25, we have

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

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 19); 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 26 only reduces the effective dimensions of the ellipsoidal map, and does not increase the eigenvalues (up to absolute constants).

Lemma 27 (Spectral domination under contractions) *Let $\Sigma \succeq 0$ be a $d\times d$ PSD matrix with ordered eigenvalues $\lambda_1(\Sigma) \geq \cdots \geq \lambda_d(\Sigma)$. Let $A \in \mathbb{R}^{d\times m}$ for some $m \leq d$ and write $s := \|A\|_{\mathrm{op}}$. Denote by $\mu_1 \geq \cdots \geq \mu_m$ the eigenvalues of $A^\top\Sigma A$. Then, for every $k = 1, \ldots, m$,*

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

Proof of Lemma 27: By the Courant–Fischer–Weyl max-min characterization (see, e.g., [Wikipedia contributors, 2025b]), we have

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

□

D.6 Proof of the Main Result

From Lemma 24, 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, the main result for ellipsoidal Grassmannian covering.

Proof of Lemma 3: We present the proof in multiple parts.

Part 1: Applying Lemma 24. Define $\mathcal{X} = \left\{X \in \mathbb{R}^{(d-r)\times r} : \|X\|_{\mathrm{op}} \le 1\right\}$ and $\bar{\mathcal{X}} = \left\{X \in \mathbb{R}^{(d-r)\times r} : \|X\|_{\mathrm{op}} \le 2\right\}$ By Lemma 24 (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 $\varepsilon > 0$,

$$\log \frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))} \le C_1 r(d-r) + \sup_{X\in\mathcal{X}} \log \frac{1}{\mathrm{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_{\mathrm{proj},\Sigma}(\mathcal{V}(X),\mathcal{V}(X')) \le \varepsilon\}}, \tag{D.27}$$

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

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

By Lemma 26 (Lipschitz of Graph Chart), we have that

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

Part 2: Volumetric Arguments. We analyze the log density complexity in (D.27) via volumetric arguments.

A technical step: ball inclusion via thresholding In order to compute the log density complexity with the uniform prior, one needs the operator norm ball to be included in the support of the prior. 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:

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

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

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

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 $\|X'\|_{\text{op}} \leq \|X' - X\|_{\text{op}} + \|X\|_{\text{op}} \leq 2$ 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:

$$B_{\varrho_1+\varrho_2}(X, \varepsilon) \subseteq B_{\varrho_1}(X, \varepsilon) \cap B_{\varrho_2}(X, \varepsilon) \subseteq B_{\varrho_1}(X, \varepsilon) \cup B_{\varrho_2}(X, \varepsilon) \subseteq \bar{\mathcal{X}}. \tag{D.28}$$

Now we are ready to proceed with the main part of the proof. By Lemma 26 (Lipschitz 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 $X \in \mathcal{X}$,

$$\begin{aligned}
&\log \frac{1}{\text{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_{\text{proj},\Sigma}(\mathcal{V}(X), \mathcal{V}(X')) \leq \varepsilon\}} \\
\leq &\log \frac{1}{\text{Unif}(\bar{\mathcal{X}})\{X' \in \bar{\mathcal{X}} : \varrho_1(X, X') + \varrho_2(X, X') \leq \varepsilon\}} \\
= &\log \frac{1}{\text{Unif}(\bar{\mathcal{X}})\{B_{\varrho_1+\varrho_2}(X, \varepsilon)\}} &\text{(D.29)} \\
= &\frac{\text{Vol}(\bar{\mathcal{X}})}{\text{Vol}(B_{\varrho_1+\varrho_2}(X, \varepsilon))}, &\text{(D.30)}
\end{aligned}$$

where the first equality uses the ball-inclusion result (D.28).

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 $B(y, R)$ the ball in $\mathcal{Y}$ centered at $y \in \mathcal{Y}$ with radius R, and by $\text{N}(\mathcal{Z}, \|\cdot\|, \varepsilon)$ the covering number of a subset $\mathcal{Z} \subseteq \mathcal{Y}$. 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 $\|z - y\| \leq \varepsilon$. The (internal) covering number is*

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

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

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

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

$$\mathsf{N}_{\text{ext}}(\mathcal{Z}, \|\cdot\|, \varepsilon) \ \leq \ \mathsf{N}(\mathcal{Z}, \|\cdot\|, \varepsilon) \ \leq \ \mathsf{N}_{\text{ext}}(\mathcal{Z}, \|\cdot\|, \varepsilon/2). \tag{D.31}$$

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

Proof of Lemma 28: The left inequality in (D.31) is immediate since any internal cover is also an external cover. For the right inequality in (D.31), 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 $V_i := \{z \in \mathcal{Z} : \ \|z - y_i\| \le \varepsilon/2\}$. By the very definition of external $(\varepsilon/2)$-cover, every $z \in \mathcal{Z}$ is within distance $\varepsilon/2$ of some y_i; hence

$$\bigcup_{i=1}^{m} V_i \ = \ \mathcal{Z}.$$

If $V_i \neq \emptyset$, 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 $y \in \mathcal{Y}$,

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

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

$$\frac{\text{Vol}(\mathcal{Z})}{\text{Vol}(B(y,\varepsilon))} \le \text{N}(\mathcal{Z}, \|\cdot\|, \varepsilon) \le \frac{\text{Vol}(\mathcal{Z} + B(y, \frac{\varepsilon}{2}))}{\text{Vol}(B(y, \frac{\varepsilon}{2}))} \le \frac{\text{Vol}(\frac{3}{2}\mathcal{Z})}{\text{Vol}(B(y, \frac{\varepsilon}{2}))} = 3^p \frac{\text{Vol}(\mathcal{Z})}{\text{Vol}(B(y,\varepsilon))}, \tag{D.32}$$

where $\lambda\mathcal{A} := \{\lambda a : a \in \mathcal{A}\}$ for $\lambda > 0$. Lastly, when the normed space $\mathcal{Y}$ is p–dimensional, for every $\varepsilon \in (0, R]$, setting $\mathcal{Z} = B(0, R)$ turns the above inequality (D.32) into the optimal covering number bound

$$\left(\frac{R}{\varepsilon}\right)^p \ \le \ \text{N}\big(B(0,R), \ \|\cdot\|, \ \varepsilon\big) \ \le \ \left(\frac{3R}{\varepsilon}\right)^p. \tag{D.33}$$

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]).

A 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

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

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 $B^{\mathcal{S}}_{\mathrm{op},H_1,H_2}((X,X),R) = \{(X',X') \in \mathcal{S} : \|(X',X')-(X,X)\|_{\mathrm{op},H_1,H_2} \leq R\}$ (the ball constrained in $\mathcal{S}$). Because there is a bijective, distance-preserving (isometric) 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

$$\frac{\mathrm{Vol}(\bar{\mathcal{X}})}{\mathrm{Vol}(B_{\varrho_1+\varrho_2}(X,\varepsilon))} = \frac{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))}, \tag{D.34}$$

where the volume on $\mathcal{S}$ is defined via the surface area measure. (D.34) is exactly the objective we need to bound in (D.29).

Given $\varepsilon > 0$, by the property (D.32) of covering number, we have that for every $X \in \mathcal{X}$ and $\varepsilon > 0$,

$$\frac{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B^{\mathcal{S}}_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))} \leq \mathrm{N}\big(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon\big). \tag{D.35}$$

Remark on why 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 (D.35) and (D.34). 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 $\mathbb{R}^{(d-r)\times r} \times \mathbb{R}^{(d-r)\times r}$:

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

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 28, we have

$$\begin{aligned}
&\mathrm{N}\big(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon\big)\\
\leq&\mathrm{N}_{\mathrm{ext}}\big(B^{\mathcal{S}}_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big)\\
\leq&\mathrm{N}_{\mathrm{ext}}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big)\\
\leq&\mathrm{N}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big).
\end{aligned} \tag{D.36}$$

For every $X \in \mathcal{X}$, the ball-inclusion argument (D.28) 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 (D.32), and we have

$$\begin{aligned}\mathrm{N}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big) \leq & 3^{2(d-r)r}\frac{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon/2))}\\ =& 6^{2(d-r)r}\frac{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))}. \qquad \text{(D.37)}\end{aligned}$$

Part 3: 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

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

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

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

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

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

which implies

$$\frac{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)}((X,X),\varepsilon))} = \prod_{k=1}^{d-r}\lambda_k(\bar{H}_1(X))^{r/2}\prod_{k=1}^{r}\lambda_k(\bar{H}_2(X))^{(d-r)/2}\frac{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((X,X),\varepsilon))}. \qquad \text{(D.38)}$$

Part 4: 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 (D.32) and (D.33), we have

$$\frac{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((0,0),4))}{\mathrm{Vol}(B_{\mathrm{op},I_{d-r},I_r}((X,X),\varepsilon))} \leq \left(\frac{12}{\varepsilon}\right)^{2(d-r)r}. \tag{D.39}$$

Combining the above inequality (D.39) with (D.37) and (D.38), we have

$$\begin{aligned}
&\log \mathrm{N}\big(B_{\mathrm{op},I_{d-r},I_r}((0,0),4),\ \|\cdot\|_{\mathrm{op},\bar{H}_1(X),\bar{H}_2(X)},\ \varepsilon/2\big)\\
\leq& 2(d-r)r\log\frac{72}{\varepsilon} + \frac{r}{2}\sum_{k=1}^{d-r}\log\lambda_k(\bar{H}_1(X)) + \frac{d-r}{2}\sum_{k=1}^{r}\log\lambda_k(\bar{H}_2(X))\\
=&\frac{r}{2}\sum_{k=1}^{d-r}\log\frac{72^2\lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{72^2\lambda_k(\bar{H}_2(X))}{\varepsilon^2}.
\end{aligned} \tag{D.40}$$

Combing the above inequality (D.40) with (D.34), (D.35) and (D.36), we have that for all $X \in \mathcal{X}$,

$$\log\frac{\mathrm{Vol}(\bar{\mathcal{X}})}{\mathrm{Vol}(B_{\varrho_1+\varrho_2}(X,\varepsilon))} \leq \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{72^2\lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{72^2\lambda_k(\bar{H}_2(X))}{\varepsilon^2}. \tag{D.41}$$

Finally, combine the above inequality (D.41) with (D.27) and (D.29), 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 $\varepsilon > 0$,

$$\begin{aligned}
\log\frac{1}{\mu(B_{\varrho_{\mathrm{proj},\Sigma}}(\mathcal{V},\varepsilon))} \leq& C_1 r(d-r) + \frac{r}{2}\sum_{k=1}^{d-r}\log\frac{72^2\lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{72^2\lambda_k(\bar{H}_2(X))}{\varepsilon^2}\\
=&\frac{r}{2}\sum_{k=1}^{d-r}\log\frac{C\lambda_k(\bar{H}_1(X))}{\varepsilon^2} + \frac{d-r}{2}\sum_{k=1}^{r}\log\frac{C\lambda_k(\bar{H}_2(X))}{\varepsilon^2},
\end{aligned} \tag{D.42}$$

where $C > 0$ is an absolute constant.

We end the proof by applying Lemma 27 and Lemma 19: since

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

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 (D.42), 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 (3.2), we construct the metric tensor

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

By Lipschitz property of the loss function we have

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

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

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

Combining the above two inequalities we have

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

By the metric domination lemma (Lemma 17), we have the pointwise dimension bound: for every $W\in B_{\mathsf{F}}(R)$,

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

By Theorem 3 (Riemannian Dimension Bound for DNN), we have that there exists a prior π such that uniformly over every $W\in B_{\mathsf{F}}(R)$,

$$\log\frac{1}{\pi(B_{\varrho_{n,\ell}}(f(W,\cdot),\varepsilon))} \le \log\frac{1}{\pi(B_{\varrho_{G_{\mathrm{NP}}(W)}}(W,\sqrt{n}\varepsilon/\beta))} \le d_{\mathrm{R}}(W,\varepsilon/\beta), \tag{E.1}$$

where the definition of Riemannian Dimension d_{R} can be found in Theorem 3. By Theorem 1, we have that there exists an absolute constant C_1 such that with probability at least $1-\delta$, uniformly over all $W\in B_{\mathsf{F}}(R)$,

$$\begin{aligned}(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \le& C_1\left(\inf_{\alpha\ge0}\left\{\alpha+\frac{1}{\sqrt{n}}\int_\alpha^1\sqrt{\log\left(\frac{1}{\pi(B_{\varrho_{n,\ell}}(f(W,\cdot),\varepsilon))}\right)}d\varepsilon\right\}+\sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right)\\ \le& C_1\left(\inf_{\alpha\ge0}\left\{\alpha+\frac{1}{\sqrt{n}}\int_\alpha^1\sqrt{d_{\mathrm{R}}(W,\varepsilon/\beta)}d\varepsilon\right\}+\sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right)\\ =& C_1\left(\inf_{\alpha\ge0}\left\{\alpha+\frac{\beta}{\sqrt{n}}\int_\alpha^1\sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon\right\}+\sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right).\end{aligned} \tag{E.2}$$

where C_1 is an absolute constant; the first inequality uses Theorem 1; and the second inequality uses (E.1). 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

$$\begin{aligned} d_{\mathrm{R}}(W,\varepsilon) = \sum_{l=1}^{L} \Big(&(d_l + d_{l-1}) \cdot d_{\mathrm{eff}}(LM^2_{l\to L}(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}(W,X)^\top, C_2 \max\{\|W\|_{\mathsf{F}}, R/2^n\}, \varepsilon) \\ &+ \log(d_{l-1}n)\Big), \end{aligned} \tag{E.3}$$

where $R = \sup_{\mathcal{W}} \|W\|_{\mathsf{F}}$, C_2 is an absolute constant, and the effective dimension (defined via (3.4)) is

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

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

Combining the identities (E.3) and (E.4), we have the pointwise dimension bound

$$\begin{aligned} &d_{\mathrm{R}}(W,\varepsilon) \\ &= \sum_{l=1}^{L}\left((d_l + d_{l-1}) \sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \log \frac{8C_2^2 \lambda_k(F_{l-1}F_{l-1}^\top)\cdot \max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\} LM^2_{l\to L}(W,\varepsilon)}{n\varepsilon^2} + \log(d_{l-1}n)\right) \\ &= \sum_{l=1}^{L}\Bigg((d_l + d_{l-1}) \sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \log \frac{8C_2^2 \lambda_k(F_{l-1}F_{l-1}^\top)}{n\varepsilon^2} \\ &\qquad + (d_l + d_{l-1}) r_{\mathrm{eff}}[W,l] \cdot \log\left(M^2_{l\to L}(W,\varepsilon) L \max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}\right) + \log(d_{l-1}n)\Bigg) \end{aligned} \tag{E.5}$$

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

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

Combining the integral upper bound (E.2) and the Riemannain dimension expression (E.5) concludes the proof of Theorem 4.

□

E.2 Proof for Regularized ERM in Section 4.3

Lemma 29 (Excess Risk Bound for Regularized ERM) *Assume we have high-probability pointwise generalization bound in the form of* (2.1)*, and the loss* $\ell(f;z)$ *is uniformly bounded by* $[0,1]$*. Then for the regularized ERM*

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

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* $1-\delta$*,*

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

Proof of Lemma 29: By (2.1), 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

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

where the first inequality uses the bound of the form (2.1); the second inequality uses definition of $\hat{f}$; and the third inequality is an application of the Hoeffding's inequality (Lemma 15) at $f^\star$; the equality is by $\delta_1 = \delta_2 = \delta/2$; and the last inequality follows from the monotonicity of the square root function. Thus we have that the excess risk is bounded by

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

□

E.3 Improvement over Norm Bounds in Section 4.2

E.3.1 Exponential Improvement 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 $\log x \le \log(1+x) \le x$, the effective dimension factor in Theorem 4 can be relaxed to the dimension-independent bound

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

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 $(\mathbb{P} - \mathbb{P}_n)\,\ell\big(f(W,x),y\big)$ is bounded by*

$$O\left(\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l + d_{l-1})L\|F_{l-1}(W,X)\|_{\mathsf{F}}^2\|W\|_{\mathsf{F}}^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n} + \sqrt{\frac{\beta^2\sum_{l=1}^{L}\log(d_{l-1}n) + \log\frac{\log(2n)}{\delta}}{n}}\right). \tag{E.6}$$

Furthermore, (E.6) 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

$$O\left(\frac{\beta\|X\|_{\mathsf{F}}\|W\|_{\mathsf{F}}\cdot\sqrt{\sum_{l=1}^{L} L(d_l + d_{l-1})\ \prod_{i\neq l}\|W_i\|_{\mathrm{op}}^2}}{n} + \sqrt{\frac{\beta^2\sum_{l=1}^{L}\log(d_{l-1}n) + L\log\frac{n\log\max\{R,2\}}{\delta}}{n}}\right). \tag{E.7}$$

Here in both (E.6) and (E.7), 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 $\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2R^2\sup_{\varepsilon>0}M_{l\to L^2(W,\varepsilon)}}}{n2^n}$; and in (E.7), 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 (E.6) and (E.7), the data matrix 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 (E.7). Second, we offer a metric–tensor interpretation that clarifies the source of this improvement. Finally, we position (E.7) 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 (E.7). 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 (E.7) can be equivalently viewed as replacing the metric tensor G_{NP} (3.2) used in Theorem 4 by the diagonal metric tensor

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

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 (E.7) 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 (E.7) 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

$$(\mathbb{P}-\mathbb{P}_n)\,\ell\big(f(W,x),y\big) \;\leq\; \tilde{O}\Big(\frac{\beta\,\|X\|_{\mathsf{F}}\,\sqrt{\sum_{l=1}^{L} L^2(d_l+d_{l-1})\,\|W_l\|_{\mathsf{F}}^2\,\prod_{i\neq l}\|W_i\|_{\mathrm{op}}^2}}{n}+\sqrt{\frac{\log\frac{1}{\delta}}{n}}\Big). \tag{E.8}$$

Neither (E.7) nor (E.8) strictly dominates the other, since factors of the form $(\sum_l a_l)(\sum_l b_l)$ in (E.7) *vs.* factors of the form $L\sum_l a_l b_l$ in (E.8) can swap their relative order.

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

$$(\mathbb{P}-\mathbb{P}_n)\,\ell\big(f(W,x),y\big) \;\leq\; \tilde{O}\Big(\frac{\beta\,\|X\|_{\mathsf{F}}\,\big(\sum_l\|W_l\|_{2,1}^{2/3}\,\sum_l\big(\prod_{i\neq l}\|W_i\|_{\mathrm{op}}\big)^{2/3}\big)^{3/2}}{n}+\sqrt{\frac{\log\frac{1}{\delta}}{n}}\Big), \tag{E.9}$$

which improves on (E.7) and (E.8) 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.

(e) Pinto et al. [2025] impose explicit per-layer rank constraints on the weight matrices, thereby replacing the width factors in (E.8) with the corresponding ranks while leaving the product of spectral norms unchanged. Their bound includes an additional C^L factor, which is subsequently removed by Ledent et al. [2025]. Moreover, Ledent et al. [2025] seek to bridge the spectral–norm and parameter–count regimes by leveraging the Schatten–p framework of Golowich et al. [2020], which interpolates between the product-of-spectral-norm regime ($p \to \infty$) and layerwise low-rank scalings ($p \to 0$). In the extreme $p \to 0$ limit, a representative consequence (Theorem E.8 of Ledent et al., 2025) yields

$$(\mathbb{P}-\mathbb{P}n)\ell\big(f(W,x),y\big) \leq \widetilde{O}\left(\frac{\sup_i\|x_i\|_2^2}{\sqrt{n}}\sqrt{\sum_{l=1}^{L} L(d_l+d_{l-1})\,\mathrm{rank}(W_l)}\right).$$

Notably, the explicit dependence on the *ranks of the weight matrices*—rather than on spectrum-aware or *feature*-rank quantities—renders this result structurally similar to VC-dimension bounds (indeed, the proof proceeds via uniform covering numbers, and packing/VC dimensions for matrices are known to adapt to explicit rank constraints [Srebro et al., 2004]). As the authors acknowledge, this is a principal limitation: empirical evidence suggests that deep networks exhibit low rank in their *features* rather than their weights, a phenomenon this bound does not capture.

In any case, (E.7) 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 (E.7).

E.3.2 Proof of Corollary 1

The bound in Theorem 4 (or (E.2) in its proof) can be further upper bounded by the following form

$$(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \leq C_1\left(\frac{1}{\sqrt{n}}\int_0^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon + \sqrt{\frac{\log\frac{\log(2n)}{\delta}}{n}}\right), \tag{E.10}$$

where the integral

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

Building on this inequality, we structure the proof in 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 $\log x \leq \log(1+x) \leq x$ for $x > 0$, we have

$$\begin{aligned}
&\sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \log\Big(\frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\Big)\\
\leq &\sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\\
\leq &\sum_{k=1}^{d_{l-1}} \frac{8C_2^2\lambda_k(F_{l-1}F_{l-1}^\top)\cdot\max\{||W||_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\\
= &\frac{8C_2^2\|F_{l-1}\|_{\mathsf{F}}^2\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2},
\end{aligned} \tag{E.11}$$

where F_{l-1} is the abbreviation of $F_{l-1}(W,X)$; $r_{\mathrm{eff}}[W,l]$ is the abbreviation of $r_{\mathrm{eff}}(LM_{l\to L}^2(W,\varepsilon)\cdot F_{l-1}(W,X)F_{l-1}(W,X)^\top, C_2\max\{\|W\|_{\mathsf{F}}, R/2^n\}, \varepsilon)$; and C_2 is a positive absolute constant. Here the second inequality uses the definition that $r_{\mathrm{eff}}[W,l]$ 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

$$\sum_{k=1}^{d_{l-1}} \lambda_k(F_{l-1}F_{l-1}^\top) = \mathrm{Tr}(F_{l-1}F_{l-1}^\top) = \|F_{l-1}\|_{\mathsf{F}}^2, \tag{E.12}$$

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 (E.11) and Theorem 4 we have the Riemannian Dimension upper bound

$$d_{\mathrm{R}}(W,\varepsilon) \leq 8C_2^2\sum_{l=1}^{L}(d_l+d_{l-1})\frac{\|F_{l-1}\|_{\mathsf{F}}^2\max\{\|W\|_{\mathsf{F}}^2, R^2/4^n\}LM_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2} + \sum_{l=1}^{L}\log(d_{l-1}n), \tag{E.13}$$

where C_2 is a positive absolute constant.

Taking (E.13) to the integral $\int_\alpha^1 \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon$, we have

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

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 $\log x \leq \log(1+x) \leq x$ for $x > 0$, we have

$$
\begin{aligned}
&\sum_{k=1}^{r_{\mathrm{eff}}[W,l]} \log\left(\frac{8C_2^2 \lambda_k(F_{l-1}F_{l-1}^\top) \cdot \max\{||W||_F^2, R^2/4^n\} L M_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\right) \\
\leq & \sum_{k=1}^{d_{l-1}} \log\left(\frac{8C_2^2 \lambda_k(F_{l-1}F_{l-1}^\top) \cdot \max\{||W||_F^2, R^2/4^n\} L M_{l\to L}^2(W,\varepsilon)}{n\varepsilon^2}\right) \\
= & \sum_{k=1}^{d_{l-1}} \log\left(\frac{8C_2^2 \lambda_k(F_{l-1}F_{l-1}^\top) \cdot \max\{||W||_F^2, R^2/4^n\} L M_{l\to L}^2(W,\varepsilon)}{n\alpha^2}\right) + d_{l-1}\log\frac{\alpha^2}{\varepsilon^2} \\
\leq & \frac{8C_2^2 \sum_{k=1}^{d_{l-1}} \lambda_k(F_{l-1}F_{l-1}^\top) \cdot \max\{||W||_F^2, R^2/4^n\} L M_{l\to L}^2(W,\varepsilon)}{n\alpha^2} + d_{l-1}\log\frac{\alpha^2}{\varepsilon^2} \\
= & \frac{8C_2^2 \|F_{l-1}(W,X)\|_{\mathsf{F}}^2 \cdot \max\{||W||_F^2, R^2/4^n\} L M_{l\to L}^2(W,\varepsilon)}{n\alpha^2} + d_{l-1}\log\frac{\alpha^2}{\varepsilon^2}.
\end{aligned}
\tag{E.14}
$$

Taking (E.14) to the integral $\int_0^\alpha \sqrt{d_{\mathrm{R}}(W,\varepsilon)}d\varepsilon$, we have

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

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

$$\frac{1}{\sqrt{n}} \int_0^\infty \sqrt{d_{\mathrm{R}}(W,\varepsilon)} d\varepsilon \leq O\left(\frac{\sqrt{\sum_{l=1}^L (d_l + d_{l-1}) L \|F_{l-1}\|_{\mathsf{F}}^2 \|W\|_{\mathsf{F}}^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n} + \sqrt{\frac{\sum_{l=1}^L \log(d_{l-1}n)}{n}}\right),$$

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 (E.10)), we have with probability at least $1-\delta$, uniformly over all $W \in B_{\mathsf{F}}(R)$,

$$\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \\
\leq& O\left(\frac{\beta\sqrt{\sum_{l=1}^L (d_l + d_{l-1}) L \|F_{l-1}\|_{\mathsf{F}}^2 \|W\|_{\mathsf{F}}^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n} + \sqrt{\frac{\beta^2 \sum_{l=1}^L \log(d_{l-1}n) + \log\frac{\log(2n)}{\delta}}{n}}\right),
\end{aligned} \tag{E.15}$$

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" Lipschitz constant at W in the sense that

$$\begin{aligned}
&||F_L(F_l(W',X),\{W_i'\}_{i=l+1}^L) - F_L(F_l(W,X),\{W_i'\}_{i=l+1}^L)||_{\mathsf{F}} \\
\leq& \left(\sup_{\varepsilon} M_{l\to L}(W,\varepsilon)\right) ||F_l(W',X) - F_l(W,X)||_{\mathsf{F}}, \quad \forall W' \in B_{\mathsf{F}}(R).
\end{aligned}$$

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

$$\begin{aligned}
&\|F_{l-1}(W,X)\|_{\mathsf{F}} \\
=&\|\sigma_{l-1}(W_{l-1}\cdots W_2\sigma_1(W_1X))\|_{\mathsf{F}} \\
\leq&\prod_{i<l}\|W_i\|_{\mathrm{op}}\cdot\|X\|_{\mathsf{F}},
\end{aligned} \tag{E.16}$$

by the property of spectral norm ($\|AB\|_{\mathsf{F}} \leq \|A\|_{\mathrm{op}}\|B\|_{\mathsf{F}}$), and the fact that all activation functions are $1-$Lipschitz in column.

In the meanwhile, we know that

$$\left(\sup_{\varepsilon} M_{l\to L}(W,\varepsilon)\right) \leq \sup_{\varepsilon}\prod_{i>l}\|W_i'\|_{\mathrm{op}},$$

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 $1-$Lipschitz in column. This results in

$$\sup_{\varepsilon} \prod_{i>l} \|W_i'\|_{\text{op}} \leq \sup_{W \in B_{\mathsf{F}}(R)} \prod_{i>l} \|W_i\|_{\text{op}}. \tag{E.17}$$

Combining (E.16) and (E.17) together with (E.15), 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

$$\begin{aligned} &(\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \\ \leq & O\left(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2 \cdot \sum_{l=1}^{L}(d_l+d_{l-1}) \prod_{i<l}\|W_i\|_{\text{op}}^2 \sup_{W\in B_{\mathsf{F}}(R)} \prod_{i>l}\|W_i\|_{\text{op}}^2}}{n}\right. \\ & \left. + \sqrt{\frac{\beta^2 \sum_{l=1}^{L} \log(d_{l-1}n) + \log\frac{\log(2n)}{\delta}}{n}}\right), \end{aligned} \tag{E.18}$$

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 $\frac{\beta\sqrt{\sum_{l=1}^{L}(d_l+d_{l-1})L\|F_{l-1}\|_{\mathsf{F}}^2 R^2 \sup_{\varepsilon>0} M_{l\to L}^2(W,\varepsilon)}}{n2^n}$.

The next step is to use a multi-dimensional extension of the "uniform pointwise convergence" principle (resulting in pointwise generalization bound (B.19) in this paper) 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 (E.18) implies that for any $l = 1, \cdots, L$, $\forall t_l \in (0, (R/\sqrt{L-1})^{2(L-1)})]$, with probability at least $1-\delta$,

$$\begin{aligned} &\sup_{W:T_l(W)\leq t_l, \forall l\in[L]} (\mathbb{P}-\mathbb{P}_n)\ell(f(W,x),y) \\ \leq & O\left(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2 \cdot \sum_{l=1}^{L}(d_l+d_{l-1})t_l}}{n} + \sqrt{\frac{\beta^2 \sum_{l=1}^{L} \log(d_{l-1}n) + \log\frac{\log(2n)}{\delta}}{n}}\right). \end{aligned} \tag{E.19}$$

With the smallest radius r_0 chosen to be $r_0 = (R/\sqrt{L-1})^{2(L-1)})/\max\{R,2\}^n$, and a grid of size $(\log_2(2\max_{W,l}\{T_l(W)\}/r_0))^k$ (partition each coordinate into $\log_2(2\max_{W,\ell}\{T_l(W)\}/r_0)$ dyadic scales, we can prove that for any $\delta \in (0,1)$, with probability at least $1-\delta$, uniformly over every

$W \in B_{\mathsf{F}}(R)$,

$$\begin{aligned}
&(\mathbb{P} - \mathbb{P}_n)\ell(f(W,x),y)\\
\leq& O\left(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\sum_{l=1}^{L}(d_l + d_{l-1})\max\{4T_l^2(W), \frac{(R/\sqrt{L-1})^{2L-2}}{\max\{R,2\}^{2n}}\}}}{n}\right.\\
&\left.+\sqrt{\frac{\beta^2\sum_{l=1}^{L}\log(d_{l-1}n) + L\log\frac{n\log\max\{R,2\}}{\delta}}{n}}\right)\\
=& O\left(\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\cdot\sum_{l=1}^{L}(d_l + d_{l-1})\prod_{i\neq l}\|W_i\|_{\mathrm{op}}^2}}{n} + \sqrt{\frac{\beta^2\sum_{l=1}^{L}\log(d_{l-1}n) + L\log\frac{n\log\max\{R,2\}}{\delta}}{n}}\right),
\end{aligned} \tag{E.20}$$

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 $\frac{\beta\sqrt{L\|W\|_{\mathsf{F}}^2\|X\|_{\mathsf{F}}^2\sum_{l=1}^{L}(d_l+d_{l-1})}(R/\sqrt{L-1})^{L-1}}{n\max\{R,2\}^n}$. The proof of this multi-dimensional "uniform pointwise convergence" is essentially the same peeling argument as in Lemma 4, with the only change that we use multi-dimensional grid; alternatively, this can be proved by applying Lemma 4 for k times, where at each step we remove one dimension functional and divided confidence by $\log_2(2R/r_0)$. We omit the repetitive proof details.

Now we see from (E.16) and (E.17) that the derived norm-constraint bound (E.15) implies the spectrally normalized bound (E.20). This completes the proof.

□