On the Asymptotics of Self-Supervised Pre-training: Two-Stage M-Estimation and Representation Symmetry

Mohammad Tinati TINATI@USC.EDU
Department of Electrical and Computer Engineering, University of Southern California

Stephen Tu STEPHEN.TU@USC.EDU
Department of Electrical and Computer Engineering, University of Southern California

Editors: Steve Hanneke and Tor Lattimore

Abstract

Self-supervised pre-training, where large corpora of unlabeled data are used to learn representations for downstream fine-tuning, has become a cornerstone of modern machine learning. While a growing body of work has begun to analyze this paradigm, existing bounds leave open the question of how sharp current rates are, and whether they accurately capture the complex interaction between pre-training and fine-tuning. In this paper, we address this gap by developing an asymptotic theory of pre-training via two-stage M-estimation. A key challenge is that the pre-training estimator is often identifiable only up to a group symmetry, a feature common in representation learning that requires careful treatment. We address this issue using tools from Riemannian geometry to study the *intrinsic* parameters of the pre-training representation, which we link with the downstream predictor through a notion of *orbit-invariance*, precisely characterizing the limiting distribution of the downstream test risk. We apply our results to spectral pre-training, factor models, and Gaussian mixture models, obtaining substantial improvements in problem-specific factors over prior art when applicable.

Keywords: Self-supervised pre-training, two-stage M-estimation, Riemannian CLT.

1. Introduction

Self-supervised pre-training has emerged as a powerful paradigm for learning representations from large corpora of unlabeled data, which are subsequently adapted to downstream tasks via fine-tuning. This approach has achieved striking empirical success across a wide range of domains in modern machine learning. For instance in computer vision, a growing body of contrastive, masked reconstruction, and self-distillation methods (Chen et al., 2020b; Zbontar et al., 2021; Grill et al., 2020; Oord et al., 2018; He et al., 2020; Wang and Isola, 2020; Chen and He, 2021; Bardes et al., 2022; He et al., 2022) have demonstrated that high-quality features can be learned without manual annotation and transferred effectively across tasks. More broadly, large language models and vision-language models trained on massive unlabeled or weakly labeled corpora have shown that pre-training can endow models with general-purpose capabilities that substantially reduce the amount of labeled data required for downstream adaptation (Devlin et al., 2019; Brown et al., 2020; Radford et al., 2021; Oquab et al., 2024).

Motivated by these empirical advances, a growing body of theoretical work has begun to investigate self-supervised pre-training, including contrastive learning and other variants, from a statistical perspective (Saunshi et al., 2019; Tosh et al., 2021; Lee et al., 2021; HaoChen et al., 2021; Cabannes et al., 2023; Saunshi et al., 2022; Ge et al., 2024; Lin and Mei, 2025). Despite the varying problem setups, loss functions, and structural assumptions in these works, a central question across much of this literature is: when does the two-stage pipeline of pre-training on unlabelled data

followed by fine-tuning on downstream task data *provably outperform training from scratch on the downstream data alone?* A closely related question involves the marginal value of pre-training data: when is the downstream task error fundamentally bottlenecked by the amount of labeled fine-tuning data, so that additional pre-training samples yield *diminishing improvement for downstream task performance?*

While recent works have made some progress towards answering these questions, we still lack an instance-optimal theory that precisely characterizes the role of pre-training loss, data distribution, and representation properties in downstream task performance. Indeed, much of the existing theory focuses on sufficient conditions and upper bounds, leaving open the question of how sharply current rates capture true behavior. Moreover, available results are typically not instance-adaptive: they do not explicitly reflect the interaction between the specific pre-training and fine-tuning distributions, losses, models, and representation structure. Contrast this to standard supervised learning, where classical M-estimation theory provides instance-specific asymptotic characterization of the excess risk; these bounds then serve as a benchmark for deriving sharp non-asymptotic results (Spokoiny, 2012; Frostig et al., 2015; Ostrovskii and Bach, 2021).

Our contribution. In this paper, we take a step towards bridging this gap by developing an asymptotic theory of self-supervised pre-training followed by fine-tuning with linear regression, in the joint limit of pre-training and fine-tuning data. Let $\alpha := m/n$ denote the ratio of pre-training samples to downstream labeled samples. Our main result shows that the scaled downstream excess risk of the two-stage estimator decomposes into two leading contributions: an intrinsic fine-tuning term, corresponding to well-specified least-squares regression on the limiting representation, and a pre-training interaction term that decays as $1/\alpha$ (see Figure 1). Thus, as the amount of pre-training data grows relative to downstream data, the pre-training interaction term vanishes, and the risk approaches the well-specified linear regression floor. Importantly, our result identifies a precise threshold characterization (i.e., $\alpha > \alpha_0$) where pre-training plus fine-tuning provides strict improvement over a downstream-only baseline. Applying our main result to several case studies—including pre-training with a spectral loss, a latent factor model, and a Gaussian mixture model—illuminates substantial gaps in problem-specific factors in prior art.

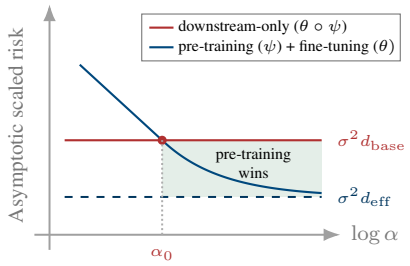

Figure 1: Schematic risk crossover. Here, d_{base} and d_{eff} denote the effective numbers of parameters in the downstream-only and known-representation limits, respectively.

A key technical challenge which arises in our analysis is that pre-training estimators are often identifiable only up to a group symmetry, which complicates the direct application of two-stage M-estimation theory (see e.g., Pagan, 1986; Newey and McFadden, 1994). We address this challenge for a general pre-training loss that learns a representation used in downstream linear regression. We first establish asymptotic normality of the *intrinsic* pre-training representation by building on recent results in Riemannian M-estimation (Brunel, 2023). We then link pre-training and downstream regression together via a key conceptual step that identifies structural conditions on the learned features that ensure *orbit-invariance* of the downstream predictor. Our general proof strategy should be broadly applicable beyond analyzing self-supervised pre-training pipelines, and we discuss future extensions in the conclusion.

2. Related Work

Our work draws on ideas from self-supervised learning, the asymptotic theory of two-stage estimators, and Riemannian M-estimation. We defer a more comprehensive discussion of related work to Appendix A, and focus here on the prior work most directly relevant to our contributions.

Most related to our work are Cabannes et al. (2023); Ge et al. (2024); Zhai et al. (2024). First, Cabannes et al. (2023) studies a VICReg-style (Balestriero and LeCun, 2022) pre-training loss combined with downstream RKHS regression. They control the downstream test risk by the (scaled) pretrain loss, which they bound using Rademacher complexity arguments. While their downstream RKHS setup is more flexible, our analysis holds for more general pre-training losses. Next, Ge et al. (2024) combine MLE pre-training with ERM fine-tuning. Their κ-*informative condition* shares high-level similarity with our goal of quantifying invariance in pre-training; however, the details differ substantially from our geometric approach. Finally, Zhai et al. (2024) study downstream error through the lens of RKHS approximation, showing that downstream error is influenced by two key terms: (a) the complexity of the RKHS induced by the augmentation distribution, and (b) how well the pre-trained encoder approximates the induced augmentation RKHS. In the aforementioned works, whether or not the upper bounds achieve optimal dependence on problem-specific constants is left open. In Section 6, we show non-trivial gaps between the upper bounds provided by Cabannes et al. (2023); Ge et al. (2024) and those that arise from our asymptotic analysis in several settings.[1]

3. Problem Formulation

Let μ_{pre} and μ_{down} be probability measures on input spaces $\mathcal{Z}$ and $\mathcal{X}$, respectively. We consider two training datasets: *(i)* a *pre-training* dataset $D_{\mathrm{pre}}^{(m)} := \{z_j\}_{j=1}^m$, where $z_j \overset{\text{i.i.d.}}{\sim} \mu_{\mathrm{pre}}$, and *(ii)* a *downstream* dataset $D_{\mathrm{down}}^{(n)} := \{(x_i, y_i)\}_{i=1}^n$, where $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} (X, Y)$ and $X \sim \mu_{\mathrm{down}}$. The datasets $D_{\mathrm{pre}}^{(m)}$ and $D_{\mathrm{down}}^{(n)}$ are drawn independently. The pair (X, Y) is further assumed to satisfy:

$$Y = f_\star(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X] = 0, \qquad \sigma^2 := \mathbb{E}[\varepsilon^2 \mid X] < \infty, \tag{3.1}$$

for some unknown regression function $f_\star : \mathcal{X} \mapsto \mathbb{R}$. Towards parameterizing $f_\star$, we fix a feature dimension $p \in \mathbb{N}_+$, and consider a differentiable representation $\psi(x, w) \in \mathbb{R}^p$, where $w \in \mathbb{R}^{q_0}$ is the representation parameter. For each w, define the linear hypothesis class $\mathcal{H}_w := \{f_{\theta,w} \mid \theta \in \mathbb{R}^p\}$ with $f_{\theta,w}(x) := \langle \theta, \psi(x, w)\rangle$. We assume that $f_\star$ in (3.1) is *well-specified* with respect to $\mathcal{F} := \bigcup_{w \in \mathbb{R}^{q_0}} \mathcal{H}_w$, i.e., $f_\star \in \mathcal{F}$. Let $(w_\star, \theta_\star)$ denote a pair such that $f_\star = f_{\theta_\star, w_\star}$.

Notation. Throughout, $L^2 := L^2(\mu_{\mathrm{down}})$ denotes the Hilbert space of real-valued square-integrable functions $g : \mathcal{X} \mapsto \mathbb{R}$ with inner product $\langle g, h\rangle := \mathbb{E}_{X \sim \mu_{\mathrm{down}}}[g(X)h(X)]$. The notation $\overset{d}{\rightsquigarrow}$ denotes convergence in distribution, and $\overset{\mathbb{P}}{\to}$ denotes convergence in probability. The set $B_d(w, r) := \{w \in \mathbb{R}^d \mid \|w\| \leqslant r\}$ denotes the closed ℓ_2-ball of radius r in $\mathbb{R}^d$; we drop the subscript d when the dimension is implicit. The set $O(p)$ denotes the orthogonal group $O(p) := \{Q \in \mathbb{R}^{p \times p} \mid Q^\top Q = QQ^\top = I\}$. Finally, $(\cdot)^+$ denotes the Moore-Penrose pseudo-inverse for a matrix.

1. The bounds from Zhai et al. (2024) are not directly comparable, as discussed further in Appendix A.

3.1. Pre-training Loss, Downstream Least-Squares Estimation, and Final Test Risk

Pre-training objective. Let $\ell_{\mathrm{pre}} : \mathbb{R}^{q_0} \times \mathcal{Z} \mapsto \mathbb{R}$ denote a pre-training loss which is twice continuously differentiable with respect to w for almost every Z, and let $L_{\mathrm{pre}}(w) := \mathbb{E}_{Z\sim\mu_{\mathrm{pre}}}\big[\ell_{\mathrm{pre}}(w; Z)\big]$ denote the corresponding population-level pre-training loss. The pre-training stage solves:

$$\hat{w}_m \in \operatorname*{arg\,min}_{w\in\mathbb{R}^{q_0}} \hat{L}_{\mathrm{pre}}(w; D_{\mathrm{pre}}^{(m)}) := \frac{1}{m}\sum_{j=1}^{m} \ell_{\mathrm{pre}}(w; z_j). \tag{3.2}$$

Our notation deliberately abstracts away the specific form of the pre-training loss; the analysis applies broadly to standard contrastive and representation-learning losses used in practice.

Downstream estimation. The downstream estimator uses both the pre-trained parameter $\hat{w}_m \in \mathbb{R}^{q_0}$ and the downstream training data $D_{\mathrm{down}}^{(n)}$ to compute:[2]

$$\hat{\theta}_{m,n} \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^{p}} \hat{L}_{\mathrm{down}}(\theta; \hat{w}_m, D_{\mathrm{down}}^{(n)}) := \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \langle\theta, \psi(x_i, w)\rangle\big)^2. \tag{3.3}$$

The resulting predictor for the downstream task is then $\hat{f}_{m,n}(\cdot) := \langle\hat{\theta}_{m,n}, \psi(\cdot, \hat{w}_m)\rangle \in \mathcal{H}_{\hat{w}_m}$.

Final test risk. Let $(X_{\mathrm{new}}, Y_{\mathrm{new}})$ be an independent test pair with the same distribution as (X, Y) (cf. (3.1)). We focus on a *conditional* notion of test-time risk that conditions on the realized pre-training dataset and downstream design, while still averaging over downstream label noise. Specifically, write $X_{1:n} := (X_1, \ldots, X_n)$ and define the (conditional) test-time risk:

$$R(D_{\mathrm{pre}}^{(m)}, X_{1:n}) := \mathbb{E}[\big(Y_{\mathrm{new}} - \hat{f}_{m,n}(X_{\mathrm{new}})\big)^2 \mid D_{\mathrm{pre}}^{(m)}, X_{1:n}]. \tag{3.4}$$

With this notation in place, the main goal of this work is the following asymptotic characterization.

Goal: Characterize as $(m, n) \to (\infty, \infty)$, with $m/n \to \alpha \in (0, \infty)$, the joint-sample limit:

$$\mathcal{E}_{m,n} := n\Big(R(D_{\mathrm{pre}}^{(m)}, X_{1:n}) - \sigma^2\Big) \overset{d}{\rightsquigarrow} \mathcal{E}_\alpha. \tag{3.5}$$

Interpreting the distributional limit (3.5). From (3.5), several important implications follow. By Fatou's lemma, we have the lower bound $\mathbb{E}[\mathcal{E}_\alpha] \leqslant \lim_{m,n}\mathbb{E}[\mathcal{E}_{m,n}]$ (here, $\lim_{m,n}$ is understood as $(m, n) \to \infty$ with $m/n \to \alpha$), which gives statistical *lower bounds* on the downstream performance. If the sequence $\{\mathcal{E}_{m,n}\}_{m,n}$ can be shown to be uniformly integrable, then this lower bound can be upgraded to equality, which yields an *exact characterization* of the asymptotic excess risk in expectation. Absent uniform integrability, we can still compute exact asymptotic high-probability upper bounds: since $\mathbb{P}(\mathcal{E}_\alpha \geqslant t) = \lim_{m,n}\mathbb{P}(\mathcal{E}_{m,n} \geqslant t)$ for any $t > 0$ (assuming $\mathcal{E}_\alpha$ is a continuous distribution), letting $t(\delta)$ be such that $\mathbb{P}(\mathcal{E}_\alpha \geqslant t(\delta)) = \delta$, we have that $\mathbb{P}(\mathcal{E}_{m,n} \geqslant t(\delta)) = \delta + o_{m,n}(1)$. In this work, we do not show uniform integrability, as this generally requires additional small-ball assumptions (Mourtada, 2022) on the features, which are not actually needed for (3.5) to hold (cf. Remark B.2).

The risk definition (3.4) takes the conditional expectation over the randomness in $(X_{\mathrm{new}}, Y_{\mathrm{new}})$ and over the downstream label noise in $D_{\mathrm{down}}^{(n)}$ (i.e., over $(Y_1, \ldots, Y_n)$ conditional on $X_{1:n}$), holding

2. When $\hat{\theta}_{m,n}$ is not unique, we define it as the minimum Euclidean norm solution.

fixed $D_{\text{pre}}^{(m)}$ and $X_{1:n}$. This is intentional, as it allows the limit analysis to focus on the interaction between the pre-train and fine-tune covariates $(D_{\text{pre}}^{(m)}, X_{1:n})$; see Appendix F.3 for more discussion.

4. Symmetries of the Two-Stage Pipeline

Section 3.1 defines a two-stage M-estimation procedure via (3.2) and (3.3), which in principle can be analyzed using the standard toolkit for classical M-estimation: consistency, asymptotic normality, and delta-method expansions (Van der Vaart, 2000). However, a crucial technical hurdle is that many pre-training objectives are invariant under symmetries, i.e., there exists a compact Lie group G acting *smoothly* on feature parameters $\mathbb{R}^{q_0}$ such that

$$\ell_{\text{pre}}(g \cdot w; z) = \ell_{\text{pre}}(w; z), \text{ for all } g \in G,\ w \in \mathbb{R}^{q_0},\ z \in \mathcal{Z}. \tag{4.1}$$

Concretely, consider a simple setting where $\mathcal{Z} = \mathbb{R}^d$ and we aim to learn a linear representation $\psi(x, A) = Ax$ with $A \in \mathbb{R}^{p \times d}$ in pre-training. Now, consider any family of pre-training losses (e.g., contrastive losses such as SimCLR (Chen et al., 2020b; Oord et al., 2018)) that act on this representation through a similarity measure $\text{sim}(x, x'; A) := \langle \psi(x, A), \psi(x', A) \rangle$. This similarity measure, and hence the pre-training loss, is invariant under any orthogonal transform $Q \in O(p)$, i.e., $\text{sim}(x, x'; A) = \text{sim}(x, x'; QA)$.

Consequently, population minimizers are typically not identifiable: if w minimizes $L_{\text{pre}}(w)$, then the entire orbit $[w] := \{g \cdot w \mid g \in G\}$ does as well. This lack of identifiability rules out both consistency and asymptotic normality for the direct parameter w. One of our key contributions is to make explicit the types of symmetry encountered in self-supervised pre-training—in a way that remains compatible with asymptotic analysis—utilizing concepts from Riemannian geometry and smooth manifolds (Lee, 2018).

4.1. Manifold Identifiability and Asymptotic Normality

The invariance (4.1) implies that the intrinsic parameter is an orbit $[w]$, i.e., an element of the quotient $\mathbb{R}^{q_0}/G$. Rather than work directly with the quotient, we represent it through a *descriptor map* $D : \mathbb{R}^{q_0} \mapsto \mathbb{R}^q$ that is *(a)* constant along orbits: $D(g \cdot w) = D(w)$ for all $g \in G,\ w \in \mathbb{R}^{q_0}$, and *(b)* separates orbits locally around $w_\star$: $\exists r > 0$ such that for $w, w' \in B(w_\star, r)$, we have $D(w) = D(w')$ iff $w' \in [w]$. We will also require that $\mathcal{M} := D\big(B(w_\star, r')\big) \subset \mathbb{R}^q$, for some $0 < r' \leqslant r$, is a well-defined C^2 embedded sub-manifold with $\Omega_\star := D(w_\star)$ in its interior; Assumption E.1 details the minimal assumptions needed for D to satisfy these requirements. We endow $\mathcal{M}$ with the Riemannian metric inherited from the ambient Euclidean space $\mathbb{R}^q$. Accordingly, for $\Omega \in \mathcal{M}$ we write $T_\Omega \mathcal{M}$ for the tangent space, and denote by $\operatorname{grad} L_{\text{pre}}(\Omega) \in T_\Omega \mathcal{M}$ and $\operatorname{Hess} L_{\text{pre}}(\Omega) : T_\Omega \mathcal{M} \mapsto T_\Omega \mathcal{M}$ the Riemannian gradient and Hessian of L_{pre} restricted to $\mathcal{M}$.

Identifiability in descriptor coordinates. We will study pre-training through the induced estimator $\hat{\Omega}_m := D(\hat{w}_m)$, rather than through the non-identifiable representative $\hat{w}_m$ itself. We start by assuming the population minimizer $\Omega_\star := \arg\min_{\Omega \in D(\mathbb{R}^{q_0}) := \{D(w) \mid w \in \mathbb{R}^{q_0}\}} L_{\text{pre}}(\Omega)$ is unique in the descriptor space. We overload the notation $L_{\text{pre}}(\Omega) := L_{\text{pre}}(w)$ for any $w \in D^{-1}(\Omega)$, which is well-defined as L_{pre} is G-invariant (cf. (4.1)). Since $\Omega_\star$ is unique and lies in the interior of the manifold $\mathcal{M}$, we have $\operatorname{grad} L_{\text{pre}}(\Omega_\star) = 0$. To obtain local quadratic control and a well-posed second-order expansion on $\mathcal{M}$, we further assume that $\operatorname{Hess} L_{\text{pre}}(\Omega_\star)$ is invertible on the tangent space: there exists $\mu > 0$ such that for all $v \in T_{\Omega_\star} \mathcal{M}$, $\big\langle v,\ \operatorname{Hess} L_{\text{pre}}(\Omega_\star)\, v \big\rangle \geqslant \mu \|v\|^2$.

Asymptotic normality on the descriptor manifold. Let $H_\star := \mathrm{Hess}\, L_{\mathrm{pre}}(\Omega_\star)$ denote the Hessian of the population pre-training loss, and let $\Sigma_\star := \mathbb{E}\big[\mathrm{grad}\, \ell_{\mathrm{pre}}(\Omega_\star; Z)\mathrm{grad}\, \ell_{\mathrm{pre}}(\Omega_\star; Z)^\top\big]$ denote the Fisher Information of the pre-training score. Writing $v_m := \log_{\Omega_\star}(\hat{\Omega}_m) \in T_{\Omega_\star}\mathcal{M}$, then under the regularity conditions stated formally in Assumption D.5, we show (Theorem D.7) that:

$$\sqrt{m}\, v_m \overset{d}{\rightsquigarrow} \mathcal{N}\big(0,\ H_\star^{-1}\Sigma_\star H_\star^{-1}\big) \qquad \text{in } T_{\Omega_\star}\mathcal{M}. \tag{4.2}$$

Equation (4.2) is the manifold analogue of classical asymptotic normality for the pre-training estimator. It follows by extracting out the asymptotic normality argument from Brunel (2023), replacing geodesic convexity assumptions with local smooth regularity conditions that apply to our pre-training setting; see Appendix D.4 for a detailed discussion.

4.2. Relating Pre-training and Downstream Estimation via Orthogonal Equivariance

As detailed in Section 3, the downstream stage depends on the pre-training stage through the representation map $\psi(\cdot, \hat{w}_m)$. However, as the symmetry in (4.1) precludes $\hat{w}_m$ from asymptotically converging to a fixed optimal parameter value, this implies that the convergence of both $\psi(\cdot, \hat{w}_m)$ and its induced linear hypothesis class $\mathcal{H}_{\hat{w}_m}$ are not well-defined without extra structure. To handle this issue, we assume that the symmetry of pre-training is compatible with downstream prediction in the sense that the induced hypothesis class is *orbit-invariant*: $\mathcal{H}_{g\cdot w} = \mathcal{H}_w$ for all $g \in G,\ w \in \mathbb{R}^{q_0}$. This condition expresses that different representatives within the same orbit $[w]$ yield the same family of predictors. However, this assumption alone is insufficient: when the OLS minimizer (3.3) is not unique, different pre-training parameters $\hat{w}_m$ within the same orbit can still lead to different minimum-norm choices. Therefore, we introduce the following condition, which we call *orthogonal equivariance* to address this issue. Specifically, we assume there exists a homomorphism $\rho : G \mapsto O(p)$ (i.e., $\rho(g_1 g_2) = \rho(g_1)\rho(g_2)$ for all $g_1, g_2 \in G$) such that

$$\psi(x, g \cdot w) = \rho(g)\, \psi(x, w), \ \text{for all } g \in G,\ w \in \mathbb{R}^{q_0},\ x \in \mathcal{X}. \tag{4.3}$$

Under condition (4.3), since different representatives w in the same orbit correspond to orthogonal coordinate changes in feature space, the corresponding OLS minimizer will be constant on the orbit.

Lemma 4.1 (Orbit-invariance of the minimum-norm downstream predictor) *Assume the orthogonal equivariance condition (4.3). Fix a downstream dataset $D_{\mathrm{down}}^{(n)}$ and a parameter $w \in \mathbb{R}^{q_0}$. Let $\hat{\theta}_w$ denote the minimum norm solution of the downstream OLS (3.3) with features $\psi(\cdot, w)$, and $\hat{f}_w(x) := \langle \hat{\theta}_w, \psi(x, w)\rangle$. Then for every $g \in G$, $\hat{f}_{g\cdot w}(\cdot) = \hat{f}_w(\cdot)$.*

Lemma 4.1 states that under (4.3), an *intrinsic* downstream feature map can be naturally defined on the orbit $[w]$ of each parameter. To see this, since D is constant on orbits, it induces a map on the quotient, and we use the descriptor $\Omega := D(w)$ as a representative coordinate for the orbit $[w]$. Then, Lemma 4.1 implies the feature map $\phi(x, \Omega) := \psi\big(x, s(\Omega)\big)$ is intrinsic for $\Omega \in \mathcal{M}$, where $s : \mathcal{M} \cap B(\Omega_\star, r'') \mapsto B(w_\star, r')$ is a C^2 local *lift* for some $r'' > 0$, satisfying $D(s(\Omega)) = \Omega$.[3]

5. Main Result: Asymptotic Behavior of the Test Risk

We now state our main result, which characterizes the joint-sample asymptotic behavior of the conditional test risk (3.4). We first introduce the operator notation used in the theorem statement.

3. The choice of s is not unique; Appendix E.2 develops a vector-bundle viewpoint showing that *(i)* the induced representation is well-defined on the quotient and *(ii)* the feature map $\phi(x, \Omega)$ is differentiable whenever s and ψ are.

5.1. Operator Notation and First-Order Residual Expansions

Let $\mathcal{H}_\Omega \subset L^2$ denote the downstream function class induced by $\phi(\cdot, \Omega)$ and Π_Ω be the population L^2-orthogonal projector onto $\mathcal{H}_\Omega$. Define the residual $e_\Omega := (I - \Pi_\Omega) f_\star$ and $\mathrm{Rep}(\Omega) := \|e_\Omega\|_{L^2}^2$. Note that under our well-specified setting $f_\star \in \mathcal{F}$, $e_{\Omega_\star} = 0$. Define the effective dimension $d_{\mathrm{eff}}(\Omega) := \dim(\mathcal{H}_\Omega) = \mathrm{rank}(\Sigma(\Omega))$ for a descriptor Ω.

The pretraining contribution is controlled by the first-order behavior of the residual e_Ω around $\Omega_\star$. We assume that this residual admits a first-order expansion in normal coordinates, i.e., there exists a bounded linear map $\mathcal{L} : T_{\Omega_\star}\mathcal{M} \mapsto L^2$ such that $e_{\exp_{\Omega_\star}(v)} = \mathcal{L}(v) + o(\|v\|)$ in L^2. Appendix G.2 gives sufficient structural conditions for such an expansion, which we now highlight. Let $N_\star := \ker(T_{\Omega_\star})$, $E_\star := N_\star^\perp$, $\mathcal{A}_v := \mathrm{Im}\left(T_{\exp_{\Omega_\star}(v)}|_{E_\star}\right)$ and define the *activated null-direction span*:

$$\mathcal{B}_v := \mathrm{Im}\left((I - \Pi_{\mathcal{A}_v}) T_{\exp_{\Omega_\star}(v)}|_{N_\star}\right).$$

We assume that there exists a stable limiting span of null directions (Assumption G.4), meaning that the projectors $\Pi_{\mathcal{B}_v}$ converge in operator norm to a limiting projector $\Pi_{\mathcal{B}_0}$, with $\mathcal{B}_0 \subseteq \mathcal{H}_{\Omega_\star}^\perp$. Define the active dimension $d_{\mathrm{act}}(\Omega_\star) := \dim\left(\mathcal{H}_{\Omega_\star} \oplus \mathcal{B}_0\right)$, which denotes the total limiting downstream degrees of freedom: the original effective dimension $d_{\mathrm{eff}}(\Omega_\star)$ plus the activated null directions. Let $\Pi_\star := \Pi_{\Omega_\star}$, and let $\theta_\star \in E_\star$ denote the unique coefficient vector with $T_{\Omega_\star}\theta_\star = f_\star$. Under these conditions, the residual linearization has the form:

$$\mathcal{L}(v) = -\left(I - \Pi_\star - \Pi_{\mathcal{B}_0}\right) DT_{\Omega_\star}[v]\theta_\star, \qquad v \in T_{\Omega_\star}\mathcal{M}. \tag{5.1}$$

The projector $I - \Pi_\star - \Pi_{\mathcal{B}_0}$ removes both the original well-specified span and the limiting active null-direction span; only the remaining first-order variation contributes to representation error. For the regular case $\mathcal{B}_0 = \{0\}$, the active and effective dimension coincide, i.e., $d_{\mathrm{act}}(\Omega) = d_{\mathrm{eff}}(\Omega)$.

5.2. Asymptotic Behavior of the Conditional Excess Test Risk

With this first-order expansion in place, we now turn to our main object of study: the scaled excess test-risk $\mathcal{E}_{m,n}$ in (3.5), obtained by conditioning on the realized pre-training sample and downstream design, and averaging only over the downstream label noise and the test pair. Accordingly, $\mathcal{E}_{m,n}$ is a random variable measurable with respect to the joint law $(D_{\mathrm{pre}}^{(m)}, X_{1:n})$, which we take all the convergences below with respect to. The following is our main result, which characterizes the asymptotic behavior of the conditional excess test risk.

Theorem 5.1 (Main result: asymptotic behavior of the conditional excess test risk) *Assume both the pre-training regularity conditions in Assumption D.5, in addition to the downstream regularity conditions Assumption G.1–G.4. Then, along any joint sequence $(m, n) \to (\infty, \infty)$ with $m/n \to \alpha \in (0, \infty)$, the (scaled) conditional excess risk $\mathcal{E}_{m,n}$ admits the distributional limit*

$$\mathcal{E}_{m,n} = n\left(R(D_{\mathrm{pre}}^{(m)}, X_{1:n}) - \sigma^2\right) \overset{d}{\rightsquigarrow} \underbrace{\sigma^2\, d_{\mathrm{act}}(\Omega_\star)}_{\text{OLS term on active limiting class}} + \underbrace{\alpha^{-1}\, \|\mathcal{L}(Z)\|_{L^2}^2}_{\text{Pre-training interaction term}}, \tag{5.2}$$

where $Z \sim \mathcal{N}(0, V)$ with $V := H_\star^{-1}\Sigma_\star H_\star^{-1}$, where $H_\star$ (resp. $\Sigma_\star$) denotes the Hessian (resp. Fisher Information matrix) of the pre-training loss (cf. Section 4.1).

Theorem 5.1 shows that the (scaled) conditional excess risk $\mathcal{E}_{m,n}$ converges in distribution to a random variable with two distinct terms. The first term, $\sigma^2 d_{\mathrm{act}}(\Omega_\star)$, is the downstream OLS degrees-of-freedom contribution on the limiting active class $\mathcal{H}_{\Omega_\star} \oplus \mathcal{B}_0$. The second term $\alpha^{-1}\|\mathcal{L}(Z)\|_{L^2}^2$ on the other hand captures the *interaction* between the pre-training loss and the downstream regression problem. As the pre-training data starts to dominate the downstream data (i.e., $\alpha \to \infty$), the contribution of the second term correctly vanishes. The remaining term is the downstream OLS degrees-of-freedom contribution on the limiting active class; in the regular case $\mathcal{B}_0 = \{0\}$, this reduces to the risk of well-specified OLS on $\mathcal{H}_{\Omega_\star}$. On the other hand, when pre-training data is relatively scarce compared to downstream (i.e., $\alpha \to 0$), the second term is dominant in (5.2), as the bias of the learned pre-training features becomes the limiting factor in the two-stage estimator. In Section 6, we instantiate Theorem 5.1 on several examples to illustrate how the assumptions translate over to concrete problem instances, and how the pre-training interaction term scales in practice.

Downstream-only baseline. To further interpret Theorem 5.1 and the risk-crossover picture in Figure 1, we compare to the empirical risk minimizer over $\mathcal{F}$ using only $D_{\mathrm{down}}^{(n)}$, in the regular case $\mathcal{B}_0 = \{0\}$. Specifically, define $f_n^{\mathrm{base}} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2$. For a regular realization $(\theta_\star, w_\star)$ of $f_\star$, define d_{base} as the rank of the linear map

$$(\delta\theta, \delta w) \mapsto \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} f_{\theta_\star + t\delta\theta,\, w_\star + t\delta w}.$$

Thus d_{base} is the dimension of the tangent space to the full downstream-only class $\mathcal{F}$ at $f_\star$. This should be contrasted with $d_{\mathrm{eff}}(\Omega_\star)$, which is the local dimension of the fixed representation class $\mathcal{H}_{\Omega_\star}$ appearing after pre-training. Assuming the standard regularity conditions for nonlinear least squares at the function level—i.e., differentiability of the parameterization, finite moments, and constant rank of the tangent map in a neighborhood of a regular realization $(\theta_\star, w_\star)$—the downstream-only baseline has leading scaled excess-risk limit $\sigma^2 d_{\mathrm{base}}$, whereas Theorem 5.1 gives the two-stage limit $\sigma^2 d_{\mathrm{eff}}(\Omega_\star) + \alpha^{-1}\|\mathcal{L}(Z)\|_{L^2}^2$. Therefore, if $d_{\mathrm{base}} > d_{\mathrm{eff}}(\Omega_\star)$,[4] then, under uniform integrability of the downstream-only and two-stage scaled excess risks, the two-stage procedure has smaller asymptotic risk in expectation whenever $\alpha > \alpha_0 := \frac{\mathbb{E}\|\mathcal{L}(Z)\|_{L^2}^2}{\sigma^2(d_{\mathrm{base}} - d_{\mathrm{eff}}(\Omega_\star))}$.

6. Examples

In this section, we apply our main result (Theorem 5.1) on a few concrete self-supervised learning examples. For each setting there are two key steps: *(i)* defining the minimal problem-specific structure to satisfy the assumptions of Theorem 5.1, and *(ii)* calculating the instance-specific bound from (5.2). Here, we describe the main setup and assumptions, deferring specific computations to the appendix. For our examples, we restrict attention to the regular case $\mathcal{B}_0 = \{0\}$. Geometrically, this rules out the activation of coefficient directions that are null at $\Omega_\star$ under infinitesimal perturbations of the descriptor. Under Assumptions G.2 and G.3, this is precisely the setting in which the population projector $\Omega \mapsto \Pi_\Omega$ is Fréchet differentiable at $\Omega_\star$. In this case, the residual linearization is given by

$$\mathcal{L}(v) = -D\Pi_{\Omega_\star}[v] f_\star = \big(I - \Pi_{\Omega_\star}\big) DT_{\Omega_\star}[v] f_\star, \qquad v \in T_{\Omega_\star}\mathcal{M}. \tag{6.1}$$

4. By Proposition G.23, if $d_{\mathrm{base}} = d_{\mathrm{eff}}(\Omega_\star)$, then $\mathcal{L} \equiv 0$. In this case, pre-training does not reduce the local downstream degrees of freedom; it only identifies an equivalent parametrization (e.g., invertible coordinate change).

6.1. Linear Spectral Pre-training

Inspired by the problem setting considered in Cabannes et al. (2023), we first consider a linear spectral contrastive objective; proofs for this case study are given in Appendix H.

Data model. Let $x \in \mathbb{R}^d$ denote a generic unlabeled pre-training sample with mean zero and covariance $\Sigma_{\mathrm{pre}} := \mathbb{E}[xx^\top] \in \mathbb{R}^{d\times d}$. A *positive pair* consists of two augmented views (x, x^+) of the same underlying instance. We define the cross-covariance matrix $\Sigma^+_{\mathrm{pre}} := \mathbb{E}[x^+x^\top]$ and assume that (x, x^+) is exchangeable which implies that Σ^+_{pre} is symmetric. A *negative sample* x^- is an independent copy of x, independent of (x, x^+). We group the samples as $z := (x, x^+, x^-)$.

Linear features and spectral loss. Fix a representation dimension $k \in [d]$ and consider linear representative-level feature maps $\psi(x, A) := Ax \in \mathbb{R}^k$ for $A \in \mathbb{R}^{k\times d}$. For any such A, define the Gram matrix (descriptor) $M_A := A^\top A \in \mathbb{R}^{d\times d}$. Here, we retain the notation A for the representative parameter (corresponding to w) and write M for the Gram descriptor (corresponding to Ω). Motivated by prior work on spectral contrastive objectives (HaoChen et al., 2021), we consider the following (single-negative) spectral loss and define the per-sample objective

$$\ell_{\mathrm{spec}}(A; z) := -2\,\langle \psi(x, A), \psi(x^+, A)\rangle + \langle \psi(x, A), \psi(x^-, A)\rangle^2. \tag{6.2}$$

Symmetry, quotient, and descriptor. The loss (6.2) depends on w only through $M_A = A^\mathsf{T}A$. Let $G = O(k)$ denote the group of $k \times k$ orthogonal matrices acting on $\mathbb{R}^{k\times d}$ by left multiplication. Then $\ell_{\mathrm{spec}}(QA; z) = \ell_{\mathrm{spec}}(A; z)$ for all $Q \in O(k)$ and all z, and hence $\hat{L}_{\mathrm{pre}}$ and L_{pre} are invariant under this action (cf. (4.1)). Moreover, the representative-level feature map is orthogonally equivariant (cf. (4.3)), that is, $\psi(x, Q \cdot A) = Q\psi(x, A)$ for all $x \in \mathbb{R}^d$ and $Q \in O(k)$. A natural orbit-invariant descriptor map is $D(A) := M_A = A^\mathsf{T}A$ which is constant along $O(k)$-orbits. On the regular regime $\mathrm{rank}(A) = k$, the action is free, and D locally identifies nearby $O(k)$-orbits with nearby points in the rank-k PSD cone $\mathcal{M}_{d,k} := \{M \in \mathbb{R}^{d\times d} \mid M \succcurlyeq 0,\ \mathrm{rank}(M) = k\}$. In particular, $\mathcal{M}_{d,k}$ is a smooth embedded submanifold of the space of symmetric matrices, and we may endow it with the induced Riemannian metric from the ambient Frobenius inner product. On a neighborhood of $M_\star$ there exists a C^2 local section s with $s(M)^\top s(M) = M$, and we define the quotient feature map by $\phi(x, M) := s(M)x$; see Appendix H.1 for the construction and smoothness.

Assumption 6.1 *Σ_{pre} is invertible and $\lambda_k(C) - \max\{\lambda_{k+1}(C), 0\} > 0$ for $C := \Sigma_{\mathrm{pre}}^{-1/2}\Sigma^+_{\mathrm{pre}}\Sigma_{\mathrm{pre}}^{-1/2}$.*

Under Assumption 6.1, the population descriptor minimizer $M_\star \in \mathcal{M}_{d,k}$ exists and is unique, which leads to the following verification of the assumptions.

Proposition 6.2 (Linear spectral model) *Assume Assumption 6.1 and $\mathbb{E}_{\mu_{\mathrm{pre}}}\|Z\|^4, \mathbb{E}_{\mu_{\mathrm{down}}}\|x\|^4 < \infty$. Then the linear spectral model satisfies the assumptions of Theorem 5.1.*

Concrete example. While the limiting expression in Theorem 5.1 admits an explicit characterization in this linear spectral model, its general form is involved. To obtain explicit closed-form expressions, we consider a simplified linear model inspired by Saunshi et al. (2022, Ex. 1). We focus on a diagonal setting that captures the essential spectral structure while allowing for precise calculations. Assume that $\Sigma_{\mathrm{pre}} = I_d$ and $\Sigma^+_{\mathrm{pre}} = \mathrm{diag}(1, 1/2, \ldots, 1/d)$, so that the whitened cross-covariance matrix is $C = \mathrm{diag}(1, 1/2, \ldots, 1/d)$. The top-$k$ population representation is therefore given by the first k coordinates. We consider a linear downstream target $f_\star(x) = \beta_\star^\top A_\star x$ where $A_\star = [\mathrm{diag}(1, 1/\sqrt{2}, \ldots, 1/\sqrt{k}), 0_{k\times(d-k)}]$ and $\beta_\star = (1, \ldots, 1)^\top$. This ensures that the task is realizable. We further assume the downstream data distribution satisfies $\mu_{\mathrm{down}} = \mathcal{N}(0, I_d)$.

Corollary 6.3 *Under the above setup, as* $(m, n) \to (\infty, \infty)$ *with* $m/n \to \alpha \in (0, \infty)$,

$$n\Big(R(D_{\text{pre}}^{(m)}, X_{1:n}) - \sigma^2\Big) \overset{d}{\rightsquigarrow} \sigma^2 k + \frac{2}{\alpha}\left(\sum_{i=1}^{k} i(1 + i^{-2} + \tau)\right)\chi^2_{d-k}, \quad \tau = \sum_{j=1}^{k} j^{-2}.$$

In particular, the pre-training interaction term $\frac{1}{m}\|\mathcal{L}(Z)\|_{L^2}^2$ *scales as* $\Theta\big(\frac{k^2(d-k)}{m}\big)$ *w.h.p.*

We compare Corollary 6.3 with Cabannes et al. (2023, Thm. 4). Applied to this setup, Cabannes et al. (2023, Thm. 4) yields (ignoring $\log n$ factors) that $\mathbb{E}[\mathcal{E}_{m,n}] \lesssim \sigma^2 k + \alpha^{-1}k(d-k)\|C^{-1}A_\star^\top \beta_\star\|_2^2 \asymp \sigma^2 k + \alpha^{-1}k^3(d-k)$ (see Appendix H.7). On the other hand, Corollary 6.3 implies that for large m, n, $\mathcal{E}_{m,n} \asymp \sigma^2 k + \alpha^{-2}k^2(d-k)$ w.h.p. Thus, our result improves the dependence in the second term by factor of k.

6.2. Factor Model Pre-training

We next specialize our main result to the latent factor example from Ge et al. (2024, Sec. 4). This setting is structurally similar to Section 6.1: the latent low-rank structure makes unsupervised pre-training natural, while a rotational invariance renders representation-level parameters non-identifiable. Proofs for this case study are presented in Appendix I.

Factor model and pre-training. Let $x \in \mathbb{R}^d$ be generated according to the factor model $x = A_\star h + \mu$ where $h \sim \mathcal{N}(0, I_k)$ is a latent factor, $\mu \sim \mathcal{N}(0, I_d)$ is independent noise, and $A_\star \in \mathbb{R}^{d\times k}$ is an unknown full-rank factor loading matrix. We assume that $k \ll d$. Pre-training observes m i.i.d. draws $\{z_i\}_{i=1}^m$ from the same distribution as x, and uses MLE to estimate the factor loading matrix, i.e., $\ell_{\text{pre}}(A; z_i) = \frac{1}{2}\left(\log\det(I_d + AA^\top) + z_i^\top (I_d + AA^\top)^{-1} z_i\right)$.

Downstream regression. We observe labeled samples (x, y) satisfying $y = \beta_\star^\top h + \nu$ where $\nu \sim \mathcal{N}(0, \sigma_\nu^2)$ and independent of (h, μ). The downstream learner does not observe h and fits a linear predictor based on features extracted from x using the pretrained representation. Specifically, under the Gaussian factor model, the regression function is linear and admits a closed form $f_\star(x) = \beta_\star^\top A_\star^\top (I_d + A_\star A_\star^\top)^{-1} x$. Accordingly, we can write the downstream labels as $Y = f_\star(X) + \varepsilon$ where $\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 := \sigma_\nu^2 + \beta_\star^\top (I_d + A_\star A_\star^\top)^{-1}\beta_\star$.

Reduction to the linear case. For any candidate loading matrix $A \in \mathbb{R}^{d\times k}$ we define the representative-level feature map $\psi(x, A) := W(A)x \in \mathbb{R}^k$ with $W(A) := A^\top(I_d + AA^\top)^{-1}$. Writing the orbit-invariant descriptor as $M = AA^\top \in \mathcal{M}_{d,k}$ and choosing any local section $s(M)$ with $s(M)s(M)^\top = M$, any representative A on the same orbit can be expressed as $A = s(M)Q$ for some $Q \in O(k)$, which yields $\psi(x, A) = Q^\top \phi(x, M)$ with $\phi(x, M) := s(M)^\top (I_d + M)^{-1}x$. Thus, passing from A to M removes the rotational non-identifiability. Consequently, this factor-model instance is a direct specialization of the linear example at the descriptor level, with the downstream class determined solely by the k-dimensional subspace $\text{range}(M)$. Let $M_\star = U_\star \,\text{diag}(\Sigma_\star, 0_{d-k})U_\star^\top$ with $\Sigma_\star = \text{diag}(\sigma_1, \ldots, \sigma_k)$ and $U_\star = [U_1\ U_2]$, where $U_1 \in \mathbb{R}^{d\times k}$ spans $\text{range}(M_\star)$.

Corollary 6.4 *For the factor model example, as* $(m, n) \to (\infty, \infty)$ *with* $m/n \to \alpha \in (0, \infty)$,

$$n\Big(R(D_{\text{pre}}^{(m)}, X_{1:n}) - \sigma^2\Big) \overset{d}{\rightsquigarrow} \sigma^2 k + \frac{1}{\alpha}\,\|(I_k + \Sigma_\star)^{-1/2} U_1^\top A_\star\, \beta_\star\|_2^2\, \chi^2_{d-k}.$$

Compared to Ge et al. (2024, Thm. 4.4), which states that w.h.p.,[5] $\mathcal{E}_{m,n} \lesssim D^4 k + \alpha^{-1} D^{12}(D^4 + \sigma_{\min}^{-4}(A_\star))d$ whenever $m \gtrsim D^4 d$, $n \geqslant D^4 k$ where $D := \max\{\|A_\star\|_{\mathrm{op}}, \|\beta_\star\|, 1\}$ (here, we ignore all $\log(1/\delta)$ for simplicity), we see that Corollary 6.4 provides a substantial improvement. In particular, it implies a sharper bound of the form $\mathcal{E}_{m,n} \asymp (\sigma_\nu^2 + D^2)k + \alpha^{-1} D^2(d-k)$ w.h.p. for large m, n, yielding a significant improvement on the dependence of D.

6.3. Gaussian Mixture Pre-training with Subspace-Aware Gating

Our final example considers unlabeled pretraining data drawn from a Gaussian mixture (MoG) with unknown centers, while the downstream predictor uses *subspace-aware* posterior responsibilities that depend only on a low-dimensional centered-mean subspace. This example is motivated by the latent classification models studied in e.g. Wei et al. (2021); Ge et al. (2024); Lin and Mei (2025), which we naturally extend to the regression setting. From a technical perspective, it also instantiates the quotient-descriptor viewpoint in a setting with discrete non-identifiability. As before, we discuss only the minimal ingredients needed to invoke Theorem 5.1, and we defer many details to Appendix J.

Pre-training data and loss. Fix $d \geqslant 1$ and $K \geqslant 2$. Let $U^\star = (u_1^\star, \dots, u_K^\star) \in (\mathbb{R}^d)^K$ be unknown centers and let $\tau \sim \mathrm{Unif}([K])$. The unlabeled distribution is the MoG $Z \mid (\tau = i) \sim \mathcal{N}(u_i^\star, I_d)$ for $i \in [K]$. Given m i.i.d. pre-training samples $Z_i \overset{\text{i.i.d.}}{\sim} \frac{1}{K}\sum_{i=1}^K \mathcal{N}(u_i^\star, I_d)$, we estimate $U^\star$ by MLE using $\ell_{\mathrm{pre}}(U; Z) := -\log(\frac{1}{K}\sum_{i=1}^K \varphi(Z - u_i))$ where $\varphi(\cdot)$ is the density of $\mathcal{N}(0, I)$.

Subspace-aware features. Define the empirical mean $\bar{u}(U) := \frac{1}{K}\sum_{i=1}^K u_i$ and centered second-moment matrix $S(U) := \sum_{i=1}^K \big(u_i - \bar{u}(U)\big)\big(u_i - \bar{u}(U)\big)^\top$. Let $r_\star := \mathrm{rank}(S(U_\star))$ and define $P_U \in O(d)$ to be the orthogonal projector onto the leading $r_\star$-dimensional eigenspace of $S(U)$ (on the regular neighborhood where this eigenspace is well-defined); centering via $S(U)$ removes an irrelevant global shift and isolates the *effective* subspace in which the mixture geometry varies. For U in the regular neighborhood, define responsibilities based on the *projected* mixture:

$$\pi_i(x; U) := \frac{\exp\big(\langle P_U u_i,\ P_U x\rangle - \frac{1}{2}\|P_U u_i\|_2^2\big)}{\sum_{j=1}^K \exp\big(\langle P_U u_j,\ P_U x\rangle - \frac{1}{2}\|P_U u_j\|_2^2\big)},\ i \in [K]. \tag{6.3}$$

These are exactly the Bayes posteriors $\pi_i(x; U) = \mathbb{P}_U(Z = i \mid P_U X = P_U x)$ for the projected model $P_U X \mid (Z = i) \sim \mathcal{N}(P_U u_i, I_{r_\star})$. Define the feature map $\psi_U : \mathbb{R}^d \mapsto \mathbb{R}^{K(d+1)}$ by

$$\psi_U(x) := \Big(\pi_1(x; U)P_U(x - u_1),\ \pi_1(x; U),\ \dots, \pi_K(x; U)P_U(x - u_K),\ \pi_K(x; U)\Big). \tag{6.4}$$

For parameters $\theta = (\theta_1, b_1, \dots, \theta_K, b_K)$ with $\theta_i \in \mathbb{R}^{r_\star}$ and $b_i \in \mathbb{R}$, the induced predictor is the linear model in features: $f_{\theta,U}(x) = \langle \theta, \psi_U(x)\rangle$.

Descriptor. The pretraining objective for the unlabeled mixture is invariant under permutations of the K components. We take $G = S_K$, the permutation group, and define its action on the parameter $U = (u_1, \cdots, u_k)$ by relabeling. The downstream hypothesis class depends on U only through its orbit $\underline{U} = [U] \in (\mathbb{R}^d)^K$. Assumption 6.5 below guarantees that $U_\star$ lies in the regular regime of the group action. Thus, we can define the quotient feature via any local lift, i.e., choice of ordering. See Appendix J.1 for the exact constructions.

5. Technically, Ge et al. (2024, Thm. 4.4) controls the excess risk *without* averaging over the conditional labels, as we do in (3.4). Since we can bound their excess risk definition by a constant factor of $R(D_{\mathrm{pre}}^{(m)}, X_{1:n})$ via Markov's inequality (on a constant probability event), we ignore this difference in our comparison.

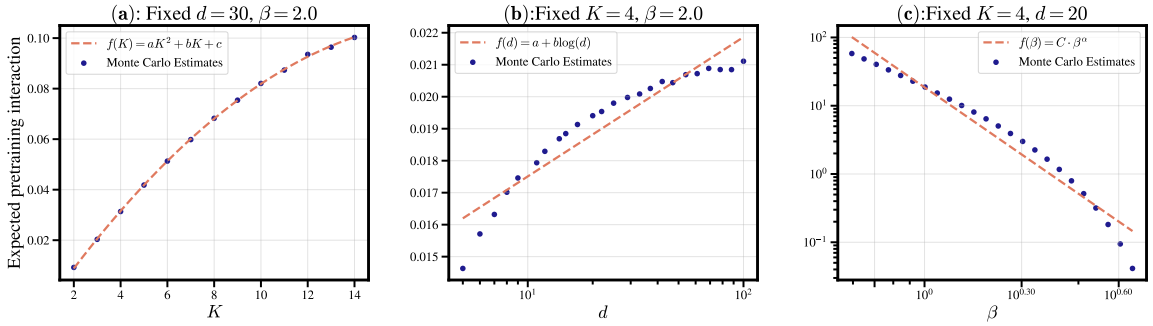

Figure 2: Monte Carlo evaluation of $\mathbb{E}[\|\mathcal{L}(Z)\|_{L^2}^2]$ in the MoG example with a block-structured downstream signal. (a) Varying K with $d = 30$ and $\beta = 2.0$. (b) Varying d with $K = 4$ and $\beta = 2.0$. (c) Varying β with $K = 4$ and $d = 20$. Dots denote Monte Carlo estimates and dashed curves denote simple fitted trends. See Appendix K for the full experimental setup and discussion.

Assumption 6.5 (Regular set for the mixture example) *The centers are distinct and the centered-mean subspace is locally stable, i.e., (i) $u_i^\star \neq u_j^\star$ for all $i \neq j$, (ii) the matrix has an eigengap between its $r_\star$-th and $(r_\star + 1)$-th eigenvalues, and (iii)* $\operatorname{rank}(\mathbb{E}[\psi_{U_\star}(X)\psi_{U_\star}(X)^\top]) = K(r_\star + 1)$.

Proposition 6.6 *Under Assumption 6.5, the MoG example satisfies the assumptions of Theorem 5.1.*

Explicit computations. While Proposition 6.6 allows us to invoke Theorem 5.1 in this MoG example, the limiting expression in (5.2) does not admit a simple closed form. The main difficulty is the term $\mathcal{L}(Z)$, which depends on derivatives of the projection operator Π_Ω and is not analytically tractable even in simple problem instances. Nevertheless, the limit in (5.2) can be evaluated numerically via Monte-Carlo simulation. Figure 2 reports such an evaluation for a simple instance with $u_i^\star = \beta e_i$, where $e_i \in \mathbb{R}^d$ denotes i-th coordinate. It shows a monotone, concave dependence on the number of blocks K, slow growth with the ambient dimension d, and rapid decay as β increases.

To complement the scaling study in Figure 2, we also provide a finite-sample distributional comparison of the prediction in Theorem 5.1. Figure 3 reports empirical cumulative distribution functions (CDFs) of the total scaled excess risk and its two leading components for a fixed GMM instance. As n grows with $m/n = \alpha$ held fixed, the empirical distributions move toward the asymptotic laws predicted by Theorem 5.1. Further experimental details are given in Appendix K. Code for reproducing the simulations is available at `https://github.com/mtinati/mog-subspace-jax`.

7. Conclusion and Discussion

We developed an asymptotic theory of self-supervised pre-training through a two-stage M-estimation framework, leveraging tools from Riemannian geometry to handle group symmetries arising in the pre-training stage. Our work opens up several promising future directions. On the technical side, a natural next step is to extend the downstream model to more general parametric regression settings; the main challenge lies in generalizing our notion of orthogonal equivariance (cf. Section 4.2) to accommodate orbit-invariance for estimators beyond OLS solutions. Another direction is developing non-asymptotic bounds for downstream risk whose leading-order terms match the asymptotic limits derived in this work. Extending our framework to pre-training objectives with auxiliary pre-training

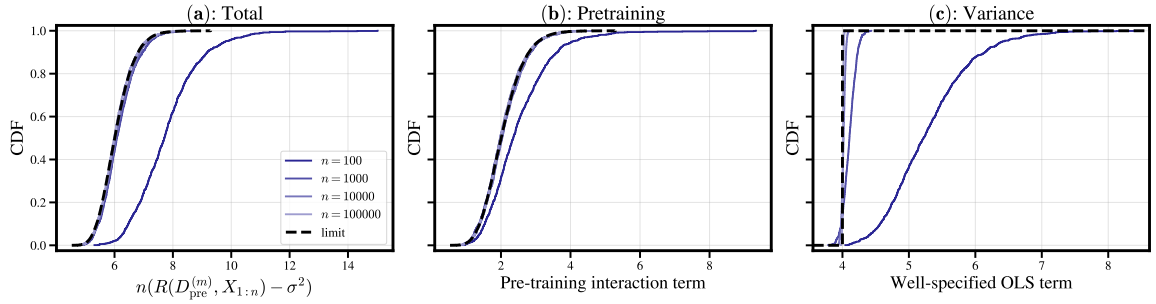

Figure 3: Finite-sample distributional convergence in the GMM example. We fix $K = 4$, $d = 20$, $\beta = 2$, and $\alpha = m/n = 2$, and compare empirical CDFs from repeated two-stage simulations with the asymptotic predictions of Theorem 5.1. Plot (a) shows the total scaled excess risk $n(R(D_{\text{pre}}^{(m)}, X_{1:n}) - \sigma^2)$ and its limiting law $\sigma^2 d_{\text{eff}} + \alpha^{-1}\|\mathcal{L}(Z)\|_{L^2}^2$. Plot (b) isolates the pretraining contribution and compares it with the limiting fluctuation $\alpha^{-1}\|\mathcal{L}(Z)\|_{L^2}^2$. Plot (c) shows the scaled variance contribution concentrating around the deterministic limit $\sigma^2 d_{\text{eff}}$. The CDF plots display representative values of n from a logarithmic grid; dashed black curves denote the corresponding asymptotic limits.

heads that are discarded before transfer is also of interest. Such objectives arise naturally in multi-task learning, where a shared representation w is trained with task-specific heads $\{\theta_k\}$,

$$\widehat{L}_{\text{pre}}(w, \theta_1, \ldots, \theta_K) = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} \ell_k(w, \theta_k; z_i),$$

and in teacher-student settings, where a student representation $\psi(\cdot, w)$ is matched to a teacher signal through an auxiliary readout θ. Since only w is transferred downstream, the natural object is the profiled loss

$$\widehat{Q}_m(w) = \inf_{\theta} \widehat{L}_{\text{pre}}(w, \theta),$$

viewed on the quotient space of representation parameters. Developing a quotient-level profile M-estimation theory for such objectives would allow auxiliary pre-training parameters to enter the asymptotic covariance through profile scores and Hessians, while keeping the downstream analysis centered on the learned representation descriptor.

More broadly, our asymptotic characterizations suggest a principled way to guide the design of pre-training losses and data-augmentation strategies, by directly optimizing the bound on the downstream risk over a diverse family of problem instances. This connection between asymptotic theory and practical pre-training design is an especially exciting avenue for future exploration.

Acknowledgments

We thank the anonymous reviewers for their thoughtful comments and suggestions, which helped improve the manuscript. This work was partially supported by a grant from Coefficient Giving and an Okawa Foundation Research Grant.

References

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26671–26685. Curran Associates, Inc., 2022.

Parikshit Bansal, Ali Kavis, and Sujay Sanghavi. Understanding contrastive learning via gaussian mixture models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Osbert Bastani. Asymptotic normality of generalized low-rank matrix sensing via riemannian geometry. *arXiv preprint arXiv:2407.10238*, 2024.

Abhishek Bhattacharya and Rabi Bhattacharya. Statistics on riemannian manifolds: Asymptotic distribution and curvature. *Proceedings of the American Mathematical Society*, 136(8):2959–2967, 2008.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds–i. *The Annals of Statistics*, 31(1):1–29, 2003.

Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds–ii. *The Annals of Statistics*, 33(3):1225–1259, 2005.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Victor-Emmanuel Brunel. Geodesically convex m-estimation in metric spaces. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2188–2210. PMLR, 2023.

Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann Lecun, and Alberto Bietti. The SSL interplay: Augmentations, inductive bias, and generalization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3252–3298. PMLR, 23–29 Jul 2023.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Yuyang Deng, Junyuan Hong, Jiayu Zhou, and Mehrdad Mahdavi. On the generalization ability of unsupervised pretraining. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4519–4527. PMLR, 02–04 May 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9588–9597, 2021.

Pascal Esser, Maximilian Fleissner, and Debarghya Ghoshdastidar. Non-parametric representation learning with kernels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11): 11910–11918, Mar. 2024.

Roy Frostig, Rong Ge, Sham M. Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 728–763, Paris, France, 03–06 Jul 2015. PMLR.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024.

Gene H Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, 20(1):1–58, 2010.

Stephan F. Huckemann. Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *The Annals of Statistics*, 39(2):1098–1124, 2011.

Atsushi Inoue and Gary Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561, 2010.

Daniel D. Johnson, Ayoub El Hanchi, and Chris J. Maddison. Contrastive learning can find an optimal basis for approximately view-invariant functions. In *The Eleventh International Conference on Learning Representations*, 2023.

Taj Jones-Mccormick, Aukosh Jagannath, and Subhabrata Sen. Provable benefits of unsupervised pre-training and transfer learning via single-index models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 28350–28376. PMLR, 13–19 Jul 2025.

Anders Klevmarken. Missing variables and two-stage least-squares estimation from more than one data set. Technical report, IUI Working Paper, 1982.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34: 309–323, 2021.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Licong Lin and Song Mei. A statistical theory of contrastive learning via approximate sufficient statistics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Kevin M. Murphy and Robert H. Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 3(4):370–379, 1985.

Whitney K. Newey. A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2):201–206, 1984. ISSN 0165-1765.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Kazusato Oko, Licong Lin, Yuhang Cai, and Song Mei. A statistical theory of contrastive pre-training and multimodal generative ai. *arXiv preprint arXiv:2501.04641*, 2025.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3–4):1175–1194, 2016. doi: 10.1007/s00440-016-0738-9.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Dmitrii M. Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance. *Electronic Journal of Statistics*, 15(1):326–391, 2021. doi: 10.1214/20-EJS1780.

David Pacini and Frank Windmeijer. Robust inference for the two-sample 2sls estimator. *Economics letters*, 146:50–54, 2016.

Adrian Pagan. Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, 25(1):221–247, 1984.

Adrian Pagan. Two stage and related estimators and their applications. *The Review of Economic Studies*, 53(4):517–538, 1986. ISSN 0034-6527. doi: 10.2307/2297604.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 09–15 Jun 2019.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19250–19286. PMLR, 17–23 Jul 2022.

Vladimir Spokoiny. Parametric estimation. Finite sample theory. *The Annals of Statistics*, 40(6): 2877 – 2909, 2012. doi: 10.1214/12-AOS1054.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research*, 22(281):1–31, 2021.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 13–18 Jul 2020.

Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, and Franziska Boenisch. Memorization in self-supervised learning improves downstream generalization. In *The Twelfth International Conference on Learning Representations*, 2024.

Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16158–16170. Curran Associates, Inc., 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR, 18–24 Jul 2021.

Runtian Zhai, Bingbin Liu, Andrej Risteski, J Zico Kolter, and Pradeep Kumar Ravikumar. Understanding augmentation-based self-supervised representation learning via RKHS approximation and regression. In *The Twelfth International Conference on Learning Representations*, 2024.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

Qi Zhang, Yifei Wang, and Yisen Wang. Identifiable contrastive learning with automatic feature importance discovery. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 58461–58477. Curran Associates, Inc., 2023.

Contents

Appendix A. More Detailed Related Work Discussion

Algorithmic approaches to pre-training and representation learning. Self-supervised pre-training has been developed through a diverse set of algorithmic paradigms, with the common goal of learning representations from unlabeled data that transfer effectively to downstream tasks. In natural language processing, masked-token prediction and autoregressive language modeling have been especially influential, as exemplified by BERT and GPT-style language models (Devlin et al., 2019; Brown et al., 2020). Analogous masked-prediction and reconstruction-based approaches have also been highly influential in vision (He et al., 2022). Contrastive learning methods—e.g., contrastive predictive coding (Oord et al., 2018), SimCLR (Chen et al., 2020b), MoCo (He et al., 2020), and CLIP (Radford et al., 2021)—learn representations by bringing together semantically related positive pairs while separating unrelated negatives. Algorithmic variants include modifying the comparison structure (Caron et al., 2020; Dwibedi et al., 2021; Zhang et al., 2023) and negative-sampling mechanism (Chuang et al., 2020; Robinson et al., 2021). A second line of joint-embedding methods avoids explicit negative samples, instead using architectural, statistical, or optimization mechanisms to prevent collapse; representative examples include BYOL (Grill et al., 2020), SimSiam (Chen and He, 2021), Barlow Twins (Zbontar et al., 2021), VICReg (Bardes et al., 2022), and DINOv2 (Oquab et al., 2024). Our work is complementary to algorithmic development; our framework is agnostic to the specific pre-training loss used in the first stage.

Theoretical studies of pre-training and fine-tuning. A growing theoretical literature seeks to explain when unsupervised pre-training produces representations that are useful for downstream prediction. Early theoretical work on contrastive learning introduced latent-class models and showed that representations with small contrastive loss can support sample-efficient downstream classification on related tasks (Saunshi et al., 2019). Related analyses study contrastive estimation in topic models, showing that contrastive representations contain topic-posterior information (Tosh et al., 2021), and pretext-task learning, where predicting one part of the input can provably reduce the labeled sample complexity of downstream learning under suitable conditional-independence assumptions (Lee et al., 2021). Spectral and graph-based perspectives have also played an important role: HaoChen et al. (2021) analyze a spectral contrastive loss through an augmentation graph, while Balestriero and LeCun (2022) connect various SSL objectives to spectral embedding methods. Bansal et al. (2025) analyze contrastive learning in Gaussian mixture models where the augmentation is biased towards the mixture component of the sample.

Many follow up works refine this picture by emphasizing that downstream performance is not determined by the SSL objective alone, but also by augmentations, model class, inductive bias, and memoization. Saunshi et al. (2022) show that analyses depending only on augmentation structure and contrastive loss can be vacuous without accounting for the function class and training algorithm. HaoChen and Ma (2023) further study how model-class inductive biases affect the structures recovered by contrastive learning. Chen et al. (2020a) give a group-theoretic treatment of data augmentation and show that it can provably lead to variance reduction. Wei et al. (2021) analyze why pretrained language models can help downstream head and prompt tuning using a latent-variable generative model. Wang et al. (2024) provide empirical evidence that some level of memorization improves generalization in SSL. Several works take an operator-theoretic approach to SSL: Cabannes et al. (2023) analyze the interplay between augmentations, inductive bias, and generalization using RKHS, Johnson et al. (2023) show that several existing contrastive learning methods are actually approximating a positive-pair kernel, Zhai et al. (2024) study SSL through the

lens of RKHS approximation, and Esser et al. (2024) develop kernelized SSL objectives. Ge et al. (2024) study a broad latent-variable framework in which maximum-likelihood pre-training is followed by empirical-risk minimization for downstream prediction. More general loss-transfer settings where the pre-training and downstream losses are allowed to have different orders is considered in Deng et al. (2024). Jones-Mccormick et al. (2025) shows that using PCA to initialize SGD improves the sample complexity of learning a single-index model. Finally, Lin and Mei (2025) bound downstream task error in terms of an approximate sufficiency loss, based on a theory of approximate sufficient statistics (Oko et al., 2025), and which is shown to be well-controlled with SimCLR pre-training.

As mentioned previously, the focus of our work is to develop an asymptotic analysis of the two-stage pre-training and fine-tuning pipeline, enabling precise computation of the limiting downstream risk. As discussed in Section 2, the works mentioned in the previous paragraph most directly related to our work are Ge et al. (2024); Cabannes et al. (2023); Zhai et al. (2024). These works provide sufficient conditions and upper bounds on the downstream risk. For Ge et al. (2024); Cabannes et al. (2023), we showed in Section 6 that the upper bounds are loose by problem-specific factors even in simple parametric examples; the bounds in our work however are not directly comparable to Zhai et al. (2024), as discussed in the next paragraph. Specific details aside, we believe that extending our analysis to cover the full generality of the nonparametric RKHS framework is exciting future work.

Detailed comparison with Zhai et al. (2024). Zhai et al. (2024) provides an RKHS approximation theory for augmentation-based SSL. Their analysis starts from the kernel induced by the augmentation distribution and assumes that the downstream target is soft invariant, or regular with respect to the corresponding RKHS. Their main downstream result (Zhai et al., 2024, Thm. 1) applies to an arbitrary d-dimensional encoder: the error bound decomposes into an approximation term, measuring how well the encoder aligns with the leading eigenspace of the augmentation-induced kernel, and an estimation term from fitting the downstream predictor with finitely many labeled samples. At the population level, the optimal d-dimensional representation—in their RKHS approximation sense—is the span of the top d eigenfunctions of the augmentation-induced kernel. They estimate this population augmentation kernel from finite pre-training data by taking the top d eigenfunctions of the corresponding empirical augmentation kernel.

Our work shares the high-level perspective that a population pre-training object determines the useful downstream representation. However, we measure the quality of pre-training in a different way. Although the bound of Zhai et al. (2024, Thm. 1) can in principle be evaluated for any fixed encoder, their analysis measures the encoder quality through augmentation-induced RKHS approximation quantities, such as alignment with the leading eigenspace or the associated trace gap. In contrast, our analysis measures pre-training quality through the statistical error of the empirical pre-training optimizer relative to its population counterpart. Thus, we study how finite sample fluctuations in the pre-training objective propagate to the downstream estimator. This also leads to different downstream guarantees. Zhai et al. (2024)'s analysis works under a weaker RKHS regularity condition and retains both an approximation term and a nonparametric downstream estimation term. Our downstream analysis, by contrast, imposes a stronger realizability structure assumption; under this stronger assumption, we obtain a faster downstream rate. Thus, these rates are not directly comparable, since they correspond to different assumptions and different notions of pre-training error.

Classical two-stage estimation in econometrics and statistics. Classical econometrics work on two-stage estimators provides the natural foundation for viewing pre-training followed by fine-tuning as a two-stage estimation problem. Pagan's work on generated regressors and two-stage estima-

tion (Pagan, 1984, 1986) highlights a key theme that remains prevalent in modern representation learning: the first-stage estimation errors generally propagate into the downstream *limiting* distribution of the second-stage estimator. Murphy and Topel (1985) derive asymptotically correct covariance expressions that account for the propagation of first-stage uncertainty, and Newey (1984) showed that sequential (e.g., two-stage) estimators can be interpreted within a method-of-moments framework. These broad ideas are encapsulated in Newey and McFadden's general large-sample theory (Newey and McFadden, 1994), which contains analysis of consistency, asymptotic normality, and variance estimation for two-stage estimation.

There are several key differences between the classic work and ours. On the one hand, classic two-stage theory does *not* assume that the data used in stage one is independent of the data used in stage two. On the other hand, the classic theory does typically assume that the first-stage estimator is identifiable in the direct parameter space, and does not address the structural issues of orbit invariance that we do in our work. Regarding the first point, there is related literature on two-sample, two-stage least squares (Klevmarken, 1982; Pacini and Windmeijer, 2016) and two-sample instrumental variable estimators (Inoue and Solon, 2010) where, similar to our pre-train and fine-tune pipeline, the data used in each stage is independent. Again, since these works do not deal with orbit identifiability, the asymptotic analysis is able to leverage e.g., arguments from Newey and McFadden (1994). Finally, we mention more recent work of Zhang et al. (2019) which studies asymptotic analysis of semi-supervised mean estimation.

Riemannian limit theorems and M-estimation. As mentioned in Section 2 and Section 4.1, the work of Brunel (2023) on geodesically-convex M-estimation provides key technical tools used to establish our limit theorems for the pre-training stage. Other prior works studying estimation on Riemannian manifolds include limit theorems for Fréchet means (Bhattacharya and Patrangenaru, 2003, 2005; Bhattacharya and Bhattacharya, 2008), geodesic principal components (Huckemann et al., 2010; Huckemann, 2011), Fréchet regression (Petersen and Müller, 2019), and low-rank matrix sensing (Bastani, 2024). Our work is complementary to this line of research, and can be viewed as leveraging these results (specifically the work of Brunel (2023)) to study a particular two-sample, two-stage M-estimation problem.

Appendix B. Structure of the Proof of Theorem 5.1

This appendix gives a high-level roadmap for the proof of Theorem 5.1. The purpose is to isolate the main statistical mechanisms behind the limit in (5.2); the full technical arguments are given in Appendix G.

B.1. Exact conditional risk decomposition

Recalling the pre-trained descriptor $\hat{\Omega}_m = \hat{\Omega}_m(D_{\mathrm{pre}}^{(m)}) \in \mathcal{M}$ from Section 4, we condition on $D_{\mathrm{pre}}^{(m)}$ and treat $\Omega := \hat{\Omega}_m$ as fixed. Assume $\mathbb{E}\|\phi(X,\Omega)\|_2^2 < \infty$. Define the population forward operator $T_\Omega : \mathbb{R}^p \to L^2_{\mathrm{down}}$ and its adjoint $T_\Omega^{\mathrm{adj}} : L^2_{\mathrm{down}} \to \mathbb{R}^p$ by

$$(T_\Omega\theta)(x) := \langle \theta, \phi(x,\Omega)\rangle, \qquad T_\Omega^{\mathrm{adj}} g := \mathbb{E}[g(X)\phi(X,\Omega)]. \tag{B.1}$$

These operators induce the population feature covariance

$$\Sigma(\Omega) := T_\Omega^{\mathrm{adj}} T_\Omega = \mathbb{E}[\phi(X,\Omega)\phi(X,\Omega)^\top].$$

Let $\mathcal{H}_\Omega = \mathrm{Im}(T_\Omega) \subseteq L^2_{\mathrm{down}}$ be the downstream function class induced by Ω. The population L^2_{down}-orthogonal projector onto $\mathcal{H}_\Omega$ is

$$\Pi_\Omega \coloneqq T_\Omega \, \Sigma(\Omega)^+ \, T_\Omega^{\mathrm{adj}}.$$

We use this projector to define the residual and representation error

$$e_\Omega \coloneqq (I - \Pi_\Omega) f_\star, \qquad \mathrm{Rep}(\Omega) \coloneqq \|e_\Omega\|^2_{L^2_{\mathrm{down}}}. \tag{B.2}$$

Since the model is well-specified at the population descriptor, we have $e_{\Omega_\star} = 0$.

Similarly, we can define the (downstream) empirical adjoint and empirical covariance as

$$T^{\mathrm{adj}}_{\Omega,n} g \coloneqq \frac{1}{n} \sum_{i=1}^n g(x_i) \phi(x_i, \Omega),$$

and $\Sigma_n(\Omega) \coloneqq \frac{1}{n} \sum_{i=1}^n \phi(x_i, \Omega) \phi(x_i, \Omega)^\top$, and the empirical projector

$$\Pi_{\Omega,n} g \coloneqq T_\Omega \, \Sigma_n(\Omega)^+ \, T^{\mathrm{adj}}_{\Omega,n} g.$$

By construction, the minimum-norm OLS predictor $\hat{f}_{\Omega,n} \coloneqq T_\Omega \hat{\theta}_{\Omega,n}$ satisfies $\hat{f}_{\Omega,n} = \Pi_{\Omega,n} y$ for any $y(x_i) = y_i$ for $i \in [n]$ (cf. Appendix C). Similarly, let $\varepsilon(\cdot)$ denote the noise function on the design, defined by $\varepsilon(x_i) = \varepsilon_i$ for $i \in [n]$. Then $y = f_\star + \varepsilon$ on $\{x_i\}_{i=1}^n$.

With this notation, the following exact finite-sample decomposition is the starting point of the proof.

Proposition B.1 (Exact conditional risk decomposition) *For the minimum-norm OLS predictor $\hat{f}_{\Omega,n}$, the conditional test risk admits the decomposition*

$$\begin{aligned} \mathbb{E}\Big[\big(Y_{\mathrm{new}} - \hat{f}_{\Omega,n}(X_{\mathrm{new}})\big)^2 \,\Big|\, D^{(m)}_{\mathrm{pre}}, X_{1:n} \Big] &= \sigma^2 + \mathrm{Rep}(\Omega) \\ &+ \underbrace{\mathbb{E}\Big[\big(\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star\big)(X_{\mathrm{new}})^2 \,\Big|\, D^{(m)}_{\mathrm{pre}}, X_{1:n} \Big]}_{=: \mathrm{Leakage}_n(\Omega)} + \underbrace{\frac{\sigma^2}{n} \mathrm{tr}\big(\Sigma(\Omega) \Sigma_n(\Omega)^+\big)}_{=: \mathrm{Var}_n(\Omega)}. \end{aligned} \tag{B.3}$$

The three terms in (B.3) have distinct roles. The representation term $\mathrm{Rep}(\Omega)$ measures the population approximation error of the learned feature class $\mathcal{H}_\Omega$. This term is zero at $\Omega_\star$, but is typically nonzero for the finite-sample pre-training estimator $\hat{\Omega}_m$. The leakage term $\mathrm{Leakage}_n(\Omega)$ measures the discrepancy between the empirical and population projection operators on $f_\star$. Under a standard well-posedness condition that the empirical inner product is non-degenerate on $\mathcal{H}_\Omega$, this term can be written as

$$\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star = \Pi_{\Omega,n} e_\Omega,$$

so leakage is caused by projecting the population residual e_Ω using the empirical downstream geometry. Finally, $\mathrm{Var}_n(\Omega)$ is the usual least-squares variance contribution from fitting the downstream labels. See Appendix F for the proof.

B.2. Main proof idea

We apply Proposition B.1 with $\Omega = \hat{\Omega}_m$. Subtracting σ^2 and multiplying by n gives

$$\mathcal{E}_{m,n} = n\,\mathrm{Var}_n(\hat{\Omega}_m) + n\,\mathrm{Leakage}_n(\hat{\Omega}_m) + n\,\mathrm{Rep}(\hat{\Omega}_m).$$

The proof of Theorem 5.1 analyzes these three terms separately under the joint limit $m, n \to \infty$ with $m/n \to \alpha$.

First, the variance term behaves as a well-specified least-squares variance term on the *activated* limiting feature space. In the stable-null-span regime, perturbations of directions that are null at $\Omega_\star$ may activate an additional limiting subspace $\mathcal{B}_0 \subseteq \mathcal{H}_{\Omega_\star}^{\perp}$. Thus the relevant limiting degrees of freedom are

$$d_{\mathrm{act}}(\Omega_\star) := \dim\big(\mathcal{H}_{\Omega_\star} \oplus \mathcal{B}_0\big),$$

rather than merely $d_{\mathrm{eff}}(\Omega_\star) = \dim(\mathcal{H}_{\Omega_\star})$. By Assumption G.4 and concentration of the downstream empirical covariance,

$$n\,\mathrm{Var}_n(\hat{\Omega}_m) = \sigma^2\,\mathrm{tr}\Big(\Sigma(\hat{\Omega}_m)\Sigma_n(\hat{\Omega}_m)^{+}\Big) \xrightarrow{\mathbb{P}} \sigma^2 d_{\mathrm{act}}(\Omega_\star).$$

Thus the downstream label noise contributes the same leading constant one would obtain from ordinary least squares on the limiting activated representation $\mathcal{H}_{\Omega_\star} \oplus \mathcal{B}_0$. In the regular special case $\mathcal{B}_0 = \{0\}$, this reduces to the usual $\sigma^2 d_{\mathrm{eff}}(\Omega_\star)$ term; see Section G.3.3.

Second, the leakage term is asymptotically negligible at the n^{-1} scale. Indeed, in the compatible regime $e_{\Omega_\star} = 0$, and the residual continuity implied by the representation-side assumptions gives

$$e_{\hat{\Omega}_m} \to 0 \qquad \text{in } L^2(\mu_{\mathrm{down}}).$$

Since the empirical projection is uniformly well behaved on $\mathcal{H}_{\hat{\Omega}_m}$, the conditional second moment of the projected residual is $o_{\mathbb{P}}(n^{-1})$, and hence

$$n\,\mathrm{Leakage}_n(\hat{\Omega}_m) \xrightarrow{\mathbb{P}} 0.$$

This shows that empirical projection error does not contribute to the limiting distribution; see Section G.3.4.

The remaining term is the representation error. This is where pre-training randomness enters. By the Riemannian M-estimation CLT for the descriptor,

$$\sqrt{m}\,\log_{\Omega_\star}(\hat{\Omega}_m) \overset{d}{\rightsquigarrow} Z, \qquad Z \sim \mathcal{N}(0, V), \qquad V = H_\star^{-1}\Sigma_\star H_\star^{-1}.$$

Since $f_\star \in \mathcal{H}_{\Omega_\star}$, the residual $e_{\Omega_\star} = (I - \Pi_{\Omega_\star})f_\star$ vanishes. The relevant first-order object is therefore the linearization of the residual map

$$e_\Omega = (I - \Pi_\Omega)f_\star,$$

rather than necessarily the full projector map $\Omega \mapsto \Pi_\Omega$. In normal coordinates, the structural condition in Appendix G.2 gives

$$e_{\exp_{\Omega_\star}(v)} = \mathcal{L}(v) + o(\|v\|) \qquad \text{in } L^2(\mu_{\mathrm{down}}),$$

where $\mathcal{L} : T_{\Omega_\star}\mathcal{M} \to L^2(\mu_{\text{down}})$ is the first-order residual map. Under the stable limiting span of null directions (Assumption G.4), this map is

$$\mathcal{L}(v) = -(I - \Pi_\star - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\theta_\star.$$

In the stronger special case where the projector map $\Omega \mapsto \Pi_\Omega$ is Fréchet differentiable at $\Omega_\star$ in operator norm, this reduces to

$$\mathcal{L}(v) = -D\Pi_{\Omega_\star}[v]f_\star.$$

Consequently,

$$m\,\text{Rep}(\hat{\Omega}_m) = m\|e_{\hat{\Omega}_m}\|^2_{L^2(\mu_{\text{down}})} \overset{d}{\rightsquigarrow} \|\mathcal{L}(Z)\|^2_{L^2(\mu_{\text{down}})}.$$

See Appendix G.2.

Combining the three limits, and using $n/m \to 1/\alpha$, gives

$$n\,\text{Rep}(\hat{\Omega}_m) = \frac{n}{m}\big(m\,\text{Rep}(\hat{\Omega}_m)\big) \overset{d}{\rightsquigarrow} \alpha^{-1}\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\text{down}})}.$$

Slutsky's theorem then yields

$$\mathcal{E}_{m,n} \overset{d}{\rightsquigarrow} \sigma^2 d_{\text{eff}}(\Omega_\star) + \alpha^{-1}\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\text{down}})}, \tag{B.4}$$

which is precisely (5.2).

Remark B.2 (From conditional to fully averaged risk)

The distributional limit in (B.4) is stated for the conditional excess risk (3.4), where we condition on the realized pre-training sample and downstream design. Passing from this conditional statement to a fully averaged risk statement requires interchanging limits and expectations. Indeed,

$$\mathbb{E}[\mathcal{E}_{m,n}] = n\left(\mathbb{E}\Big[(Y_{\text{new}} - \hat{f}_{m,n}(X_{\text{new}}))^2\Big] - \sigma^2\right),$$

since

$$\mathbb{E}\Big[(Y_{\text{new}} - \hat{f}_{m,n}(X_{\text{new}}))^2\Big] = \mathbb{E}\Big[R(D^{(m)}_{\text{pre}}, X_{1:n})\Big].$$

This upgrade is typically delicate because the conditional expansion contains inverse empirical covariance terms, with

$$\widehat{\Sigma}_{n,m} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i, \hat{\Omega}_m)\phi(x_i, \hat{\Omega}_m)^\top.$$

When $\widehat{\Sigma}_{n,m}$ is ill-conditioned, its smallest nonzero eigenvalue can be very small with non-negligible probability, producing heavy-tailed inverse-covariance contributions.

A sufficient route to convergence of the fully averaged risk is to establish uniform integrability of $\{\mathcal{E}_{m,n}\}_{m,n}$. For instance, sufficiently strong lower-tail or anti-concentration bounds for $\widehat{\Sigma}_{n,m}$, as in the small-ball analysis of random design least squares (cf. Mourtada, 2022), can control the inverse-covariance terms and allow the conditional expansion to be integrated. In that case,

$$\mathbb{E}[\mathcal{E}_{m,n}] \longrightarrow \mathbb{E}[\mathcal{E}_\alpha],$$

so the same limiting expression also characterizes the scaled fully averaged excess risk.

Appendix C. Empirical Projection Operators and Linear-Algebra Preliminaries

This appendix collects basic facts used repeatedly in the proofs of Proposition B.1 and Theorem 5.1. The statements are standard; we include them to (i) make explicit which inner product each projection is taken with respect to, and (ii) clarify what is unique (fitted values on the sample) versus what is a chosen convention (a minimum-norm representative in coefficient space) when the design is rank-deficient. All statements in this appendix are deterministic once the design points $X_{1:n}$ are fixed; in particular, they are tailored to our conditional-on-design viewpoint in the main text. We also list the norm conventions used for matrices, feature operators, and descriptor derivatives at the end of this appendix.

C.1. Empirical inner products and evaluation maps

Let $\mathcal{X}$ be the input space (e.g. $\mathcal{X} = \mathbb{R}^d$ in the linear–Gaussian example), and let $\mathcal{G}$ be a class of measurable functions $g : \mathcal{X} \to \mathbb{R}$ (for instance $\mathcal{G} = L^2(\mu_{\mathrm{down}})$, or any function space containing the hypothesis classes used in the main text).

Fix design points $x_{1:n} \coloneqq (x_1, \dots, x_n) \in \mathcal{X}^n$. For $g, h \in \mathcal{G}$ such that $g(x_i), h(x_i) \in \mathbb{R}$ for all $i \in [n]$, define the empirical bilinear form

$$\langle g, h \rangle_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} g(x_i)\, h(x_i), \qquad \|g\|_n^2 \coloneqq \langle g, g \rangle_n.$$

In general, $\|\cdot\|_n$ is only a *seminorm*: it depends only on the values of a function on the finite set $\{x_1, \dots, x_n\}$. Consequently, the fitted values of any empirical least-squares projection are uniquely determined only through these n values.

To isolate this dependence, define the evaluation map at $x_{1:n}$ by

$$\mathrm{Ev}_n : \mathcal{G} \to \mathbb{R}^n, \qquad \mathrm{Ev}_n(g) \coloneqq \big(g(x_1), \dots, g(x_n)\big)^\top.$$

Whenever $g(x_i) \in \mathbb{R}$ for all i, we have $\mathrm{Ev}_n(g) \in \mathbb{R}^n$ and

$$\langle g, h \rangle_n = \frac{1}{n} \mathrm{Ev}_n(g)^\top \mathrm{Ev}_n(h), \qquad \|g\|_n^2 = \frac{1}{n} \|\mathrm{Ev}_n(g)\|_2^2.$$

Thus, empirical least-squares projection statements can be proved equivalently in the finite-dimensional space $\mathbb{R}^n$ by working with the vectors of function values $\mathrm{Ev}_n(g)$.

C.2. Pseudoinverse identities

We use $(\cdot)^+$ to denote the Moore–Penrose pseudoinverse.

Lemma C.1 (Basic pseudoinverse identities) *For any matrix A (not necessarily square),*

$$AA^+A = A, \qquad A^+AA^+ = A^+, \qquad (AA^+)^\top = AA^+, \qquad (A^+A)^\top = A^+A.$$

Moreover, AA^+ is the Euclidean orthogonal projector onto $\mathrm{Im}(A)$ and A^+A is the Euclidean orthogonal projector onto $\mathrm{Im}(A^\top)$.

Lemma C.2 (Trace equals rank for symmetric PSD matrices) *If $S \in \mathbb{R}^{p\times p}$ is symmetric positive semidefinite, then*

$$\operatorname{tr}(S^{+}S) = \operatorname{rank}(S).$$

Remark C.3 (On conditioning and inverse-eigenvalue effects) *The identities in this section are purely algebraic and hold deterministically for any realization of the design. When taking expectations over random designs, quantities involving S^{+} can be sensitive to small eigenvalues of S, and additional tail control is typically needed to justify interchanging limits and expectations. This sensitivity motivates the conditional-on-design risk formulation adopted in the main text.*

C.3. Design matrices and hat matrices

Fix a feature parameter $\Omega \in \mathbb{R}^q$ and design points $x_{1:n}$. Define the feature matrix

$$\Phi_\Omega \in \mathbb{R}^{n\times p}, \qquad (\Phi_\Omega)_{i,:} \coloneqq \phi(x_i,\Omega)^\top,$$

and the empirical covariance

$$\Sigma_n(\Omega) = \frac{1}{n}\Phi_\Omega^\top\Phi_\Omega.$$

Define the hat matrix

$$H_{\Omega,n} \;\coloneqq\; \Phi_\Omega(\Phi_\Omega^\top\Phi_\Omega)^{+}\Phi_\Omega^\top \;=\; \frac{1}{n}\,\Phi_\Omega\Sigma_n(\Omega)^{+}\Phi_\Omega^\top.$$

Lemma C.4 (Hat matrix is an orthogonal projector) *$H_{\Omega,n}$ is the Euclidean orthogonal projector in $\mathbb{R}^n$ onto $\operatorname{Im}(\Phi_\Omega)$. In particular, $H_{\Omega,n}^2 = H_{\Omega,n}$, $H_{\Omega,n}^\top = H_{\Omega,n}$, and*

$$\operatorname{tr}(H_{\Omega,n}) = \operatorname{rank}(\Phi_\Omega) = \operatorname{rank}(\Sigma_n(\Omega)).$$

Proof By Lemma C.1, $\Phi_\Omega(\Phi_\Omega^\top\Phi_\Omega)^{+}\Phi_\Omega^\top$ is the orthogonal projector onto $\operatorname{Im}(\Phi_\Omega)$. The trace of an orthogonal projector equals its rank. Finally, $\operatorname{rank}(\Phi_\Omega) = \operatorname{rank}(\Phi_\Omega^\top\Phi_\Omega) = \operatorname{rank}(\Sigma_n(\Omega))$. ■

C.4. Empirical least-squares projection onto $\mathcal{H}_\Omega$

Recall the linear class

$$\mathcal{H}_\Omega \coloneqq \{T_\Omega\theta : \theta \in \mathbb{R}^p\}, \qquad (T_\Omega\theta)(x) \coloneqq \langle\theta, \phi(x,\Omega)\rangle.$$

Parameterization versus functions. The parametrization $\theta \mapsto T_\Omega\theta$ need not be injective: if $v \in \mathbb{R}^p$ satisfies $\langle v, \phi(x,\Omega)\rangle = 0$ for all $x \in \mathcal{X}$ (equivalently, $T_\Omega v \equiv 0$), then $T_\Omega(\theta+v) = T_\Omega\theta$ as functions on $\mathcal{X}$. Thus, even when the induced function is unique, the coefficient vector representing it may not be.

Given design points $x_{1:n}$, define the empirical adjoint

$$T_{\Omega,n}^{\mathrm{adj}} g \coloneqq \frac{1}{n}\sum_{i=1}^{n} g(x_i)\phi(x_i,\Omega) \in \mathbb{R}^p.$$

Lemma C.5 (Empirical adjoint identity) *For any $\theta \in \mathbb{R}^p$ and any function g, we have*

$$\langle T_\Omega\theta, g\rangle_n = \langle\theta, T_{\Omega,n}^{\mathrm{adj}} g\rangle_{\mathbb{R}^p}.$$

Proof This is simply a consequence of the following equalities:

$$\langle T_\Omega \theta, g\rangle_n = \frac{1}{n}\sum_{i=1}^n \langle \theta, \phi(x_i,\Omega)\rangle g(x_i) = \left\langle \theta, \frac{1}{n}\sum_{i=1}^n g(x_i)\phi(x_i,\Omega)\right\rangle = \langle \theta, T^{\mathrm{adj}}_{\Omega,n} g\rangle_{\mathbb{R}^p}.$$

■

Empirical projection: what is unique and what is a convention. Since $\|\cdot\|_n$ is only a seminorm, the empirical least-squares problem

$$\min_{h\in\mathcal{H}_\Omega} \|g-h\|_n^2 \qquad \left(\, \|g-h\|_n^2 = \frac{1}{n}\sum_{i=1}^n (g(x_i)-h(x_i))^2 \,\right)$$

can determine only the values of the minimizer on the sample points. In particular, the objective depends on h only through the vector $\mathrm{Ev}_n(h) = (h(x_1),\dots,h(x_n))^\top$, and any two functions that agree on $\{x_i\}_{i=1}^n$ are indistinguishable under $\|\cdot\|_n$. Writing $h = T_\Omega\theta$ so that $\mathrm{Ev}_n(h) = \Phi_\Omega\theta$, the problem reduces to Euclidean least squares in $\mathbb{R}^n$.

Lemma C.6 (Least-squares equivalence) *The empirical projection problem is equivalent to the Euclidean least-squares problem*

$$\min_{\theta\in\mathbb{R}^p} \frac{1}{n}\|\mathrm{Ev}_n(g) - \Phi_\Omega\theta\|_2^2.$$

Proof By definition,

$$\|g-h\|_n^2 = \frac{1}{n}\sum_{i=1}^n (g(x_i)-h(x_i))^2 = \frac{1}{n}\|\mathrm{Ev}_n(g) - \mathrm{Ev}_n(h)\|_2^2.$$

Every $h\in\mathcal{H}_\Omega$ can be written as $h = T_\Omega\theta$ for some $\theta\in\mathbb{R}^p$, and then

$$\mathrm{Ev}_n(h) = \big((T_\Omega\theta)(x_1),\dots,(T_\Omega\theta)(x_n)\big)^\top = \big(\langle\theta,\phi(x_1,\Omega)\rangle,\dots,\langle\theta,\phi(x_n,\Omega)\rangle\big)^\top = \Phi_\Omega\theta.$$

Substituting this identity into the empirical objective gives

$$\|g - T_\Omega\theta\|_n^2 = \frac{1}{n}\|\mathrm{Ev}_n(g) - \Phi_\Omega\theta\|_2^2,$$

which proves the claim. ■

Even when Φ_Ω is rank-deficient, the *fitted values* are unique: the Euclidean orthogonal projection of $\mathrm{Ev}_n(g)$ onto $\mathrm{Im}(\Phi_\Omega)$

$$\hat{g}_{1:n} \;\coloneqq\; H_{\Omega,n}\,\mathrm{Ev}_n(g) \in \mathbb{R}^n \tag{C.1}$$

is uniquely determined by Lemma C.4.

In contrast, the coefficient vector θ achieving these fitted values need not be unique when Φ_Ω is rank-deficient: if $\theta^\star$ is one minimizer then all minimizers are $\theta^\star + v$ with $v\in\ker(\Phi_\Omega)$, and they all

satisfy $\Phi_\Omega \theta = \hat{g}_{1:n}$. Whether these distinct minimizers define the same function on $\mathcal{X}$ depends on the feature representation: if $\ker(\Phi_\Omega) \subseteq \ker(T_\Omega)$ then all minimizers induce the same function in $\mathcal{H}_\Omega$, whereas if there exists $v \neq 0$ with $\Phi_\Omega v = 0$ but $T_\Omega v \not\equiv 0$, then different minimizers agree on the sample points but can differ off-sample.

Throughout the paper, we follow the convention fixed in the main text: the empirical projector $\Pi_{\Omega,n}$ is defined via the Moore–Penrose pseudoinverse (see Section 5), so that the associated coefficient vector is the minimum-Euclidean-norm least-squares solution. Importantly, all fitted-value identities below hold regardless of this convention.

Definition C.7 (Canonical empirical projector) *For g with finite evaluations on $\{x_i\}_{i=1}^n$, define*

$$\hat{\theta}_{\Omega,n}(g) \;:=\; \Sigma_n(\Omega)^+ \, T_{\Omega,n}^{\mathrm{adj}} g \in \mathbb{R}^p,$$

and set

$$\Pi_{\Omega,n} g \;:=\; T_\Omega \, \hat{\theta}_{\Omega,n}(g) \in \mathcal{H}_\Omega.$$

Lemma C.8 (Least-squares characterization and empirical orthogonality) *For any g we have*

$$\Pi_{\Omega,n} g \in \operatorname*{arg\,min}_{h \in \mathcal{H}_\Omega} \; \|g - h\|_n^2.$$

Moreover, the fitted values satisfy

$$\mathrm{Ev}_n(\Pi_{\Omega,n} g) = H_{\Omega,n} \, \mathrm{Ev}_n(g),$$

and the residual is empirically orthogonal to $\mathcal{H}_\Omega$ in the sense that

$$\langle g - \Pi_{\Omega,n} g, \, h \rangle_n = 0 \qquad \forall h \in \mathcal{H}_\Omega.$$

Proof Write $h = T_\Omega \theta$. Then

$$\|g - h\|_n^2 = \frac{1}{n} \|\mathrm{Ev}_n(g) - \Phi_\Omega \theta\|_2^2.$$

Thus minimizing over $h \in \mathcal{H}_\Omega$ is equivalent to least squares in $\mathbb{R}^n$. By Definition C.7, $\hat{\theta}_{\Omega,n}(g)$ is the Moore–Penrose minimum-norm least-squares solution, hence $\Pi_{\Omega,n} g = T_\Omega \hat{\theta}_{\Omega,n}(g)$ is a minimizer. Furthermore,

$$\mathrm{Ev}_n(\Pi_{\Omega,n} g) = \Phi_\Omega \hat{\theta}_{\Omega,n}(g) = \Phi_\Omega (\Phi_\Omega^\top \Phi_\Omega)^+ \Phi_\Omega^\top \, \mathrm{Ev}_n(g) = H_{\Omega,n} \mathrm{Ev}_n(g).$$

For empirical orthogonality, let $r := \mathrm{Ev}_n(g) - \mathrm{Ev}_n(\Pi_{\Omega,n} g)$. Since $\mathrm{Ev}_n(\Pi_{\Omega,n} g) = H_{\Omega,n} \mathrm{Ev}_n(g)$ and $H_{\Omega,n}$ is the orthogonal projector onto $\mathrm{Im}(\Phi_\Omega)$, we have $r \perp \mathrm{Im}(\Phi_\Omega)$. For any $h = T_\Omega \theta \in \mathcal{H}_\Omega$, $\mathrm{Ev}_n(h) = \Phi_\Omega \theta \in \mathrm{Im}(\Phi_\Omega)$, hence

$$n \langle g - \Pi_{\Omega,n} g, \, h \rangle_n = r^\top \mathrm{Ev}_n(h) = 0.$$

■

Lemma C.9 (Reproducing property on the sample points) *For any $h \in \mathcal{H}_\Omega$,*

$$\mathrm{Ev}_n(\Pi_{\Omega,n} h) = \mathrm{Ev}_n(h), \qquad \textit{and hence} \qquad \|h - \Pi_{\Omega,n} h\|_n = 0.$$

Proof If $h \in \mathcal{H}_\Omega$ then $\mathrm{Ev}_n(h) \in \mathrm{Im}(\Phi_\Omega)$, so $H_{\Omega,n}\mathrm{Ev}_n(h) = \mathrm{Ev}_n(h)$ by Lemma C.4. Now apply Lemma C.8. ■

Definition C.10 *For any vector $v \in \mathbb{R}^n$, we define $\mathrm{lift}_n(v)$ as any arbitrary measurable function $g : \mathcal{X} \to \mathbb{R}$ that satisfies $\mathrm{Ev}_n(g) = v$ (the values of g off $\{x_i\}_{i=1}^n$ are irrelevant). Since $\Pi_{\Omega,n}$ depends on an input only through its evaluations on the design points, $\Pi_{\Omega,n}\mathrm{lift}_n(v)$ is well-defined.*

Let $y_{1:n} = (y_1, \cdots, y_n)^\top$, and recall the minimum-norm least-squares solution $\hat{\theta}_{\Omega,n}$ and $\hat{f}_{\Omega,n} = T_\Omega \hat{\theta}_{\Omega,n}$.

Lemma C.11 (OLS equals empirical projection) *The OLS predictor satisfies $\hat{f}_{\Omega,n} = \Pi_{\Omega,n}\mathrm{lift}_n(y_{1:n})$, and hence*

$$\mathrm{Ev}_n(\hat{f}_{\Omega,n}) = H_{\Omega,n}\,(y_1, \ldots, y_n)^\top.$$

Proof By definition, $\hat{\theta}_{\Omega,n}$ minimizes $\|\mathrm{lift}_n(y_{1:n}) - T_\Omega\theta\|_n^2$, which is exactly the least-squares problem encoded in Definition C.7 with $g = \mathrm{lift}_n(y_{1:n})$. Therefore $T_\Omega\hat{\theta}_{\Omega,n}$ coincides with $\Pi_{\Omega,n}\mathrm{lift}_n(y_{1:n})$, and evaluating yields the hat-matrix identity. ■

C.5. Effective dimension and degrees of freedom

Recall the empirical effective dimension

$$d_{\mathrm{eff},n}(\Omega) := \mathrm{tr}\big(\Sigma_n(\Omega)^+\Sigma_n(\Omega)\big) = \mathrm{rank}(\Sigma_n(\Omega)).$$

Lemma C.12 (Equivalent characterizations of $d_{\mathrm{eff},n}$) *For any Ω and design points $x_{1:n}$,*

$$d_{\mathrm{eff},n}(\Omega) = \mathrm{rank}(\Sigma_n(\Omega)) = \mathrm{rank}(\Phi_\Omega) = \mathrm{tr}(H_{\Omega,n}).$$

Proof Combine Lemma C.2 with Lemma C.4. ■

Lemma C.13 (Projected-noise identity conditional on the design) *Let $\varepsilon_{1:n} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ with $\mathbb{E}[\varepsilon_{1:n} \mid x_{1:n}] = 0$ and $\mathbb{E}[\varepsilon_{1:n}\varepsilon_{1:n}^\top \mid x_{1:n}] = \sigma^2 I_n$, and assume $\varepsilon_{1:n}$ is conditionally independent of $x_{1:n}$. Then, for any Ω,*

$$\mathbb{E}\left[\frac{1}{n}\|H_{\Omega,n}\varepsilon_{1:n}\|_2^2 \,\middle|\, x_{1:n}\right] = \frac{\sigma^2}{n}\,\mathrm{tr}(H_{\Omega,n}) = \frac{\sigma^2}{n}\,d_{\mathrm{eff},n}(\Omega).$$

Proof Condition on $x_{1:n}$ and use $H_{\Omega,n}^\top = H_{\Omega,n}$ and $H_{\Omega,n}^2 = H_{\Omega,n}$:

$$\mathbb{E}\big[\|H_{\Omega,n}\varepsilon_{1:n}\|_2^2 \,\big|\, x_{1:n}\big] = \mathbb{E}\Big[\varepsilon_{1:n}^\top H_{\Omega,n}\varepsilon_{1:n} \,\Big|\, x_{1:n}\Big] = \mathrm{tr}\Big(H_{\Omega,n}\,\mathbb{E}[\varepsilon_{1:n}\varepsilon_{1:n}^\top \mid x_{1:n}]\Big) = \sigma^2\,\mathrm{tr}(H_{\Omega,n}).$$

Divide by n and invoke Lemma C.12. ■

Remark C.14 *Lemma C.12 is the linear-algebraic reason the leading noise-estimation constant is a degrees-of-freedom term: under homoskedastic noise, the conditional expected squared norm of the projected noise depends on the design only through $\mathrm{tr}(H_{\Omega,n})$, which equals $d_{\mathrm{eff},n}(\Omega)$ by Lemma C.12.*

C.6. Benefits of an operator-theoretic formulation

The downstream stage does not use Ω as an object in isolation; it uses only the linear function class it induces together with the least-squares fit of the labels onto this class. Introducing the linear map T_Ω packages all downstream quantities in a coordinate-free form.

(i) *Invariance becomes explicit.* If two feature parametrizations induce the same subspace $\mathcal{H}_\Omega \subseteq L^2_{\text{down}}$, then downstream predictions after refitting are identical. In the operator language, all population objects depend on Ω only through $\mathcal{H}_\Omega$, summarized by the projector $\Pi_\Omega = T_\Omega \Sigma(\Omega)^+ T_\Omega^{\text{adj}}$.

(ii) *Population and empirical stages have the same algebraic structure.* The empirical projector $\Pi_{\Omega,n}$ is obtained from Π_Ω by replacing expectations with sample averages, i.e., T_Ω^{adj} by $T_{\Omega,n}^{\text{adj}}$ and $\Sigma(\Omega)$ by $\Sigma_n(\Omega)$.

(iii) *Bridge to the manifold pre-training limit theory.* Our m-asymptotics enter through how population quantities change with Ω near $\Omega_\star$. When $\mathcal{M}$ is a Riemannian manifold, the pre-training estimator satisfies a log-map CLT

$$\sqrt{m}\, \log_{\Omega_\star}(\hat{\Omega}_m) \overset{d}{\rightsquigarrow} \mathcal{N}(0, V) \quad \text{in } T_{\Omega_\star}\mathcal{M},$$

under the assumptions in the main text. The operator viewpoint makes it natural to apply a delta-method argument to maps of the form $\Omega \mapsto \Pi_\Omega$ and $\Omega \mapsto \text{Rep}(\Omega)$, after passing to the appropriate descriptor/quotient parametrization described in Section 4.

C.7. Norm conventions

Throughout the paper, $\|\cdot\|_2$ denotes the Euclidean norm on finite-dimensional spaces. For a matrix A, $\|A\|_{\text{op}}$ denotes the induces Euclidean operator norm

$$\|A\|_{\text{op}} = \sup_{\|\theta\|_2=1} \|A\theta\|_2.$$

More generally, if $S : F \to E$ is a bounded linear map between normed spaces, $\|S\|_{\text{op}}$ denotes the induced operator norm. In particular, for a feature map $T_\Omega : \mathbb{R}^p \to L^2$,

$$\|T_\Omega\|_{\text{op}}^2 = \sup_{\|\theta\|_2=1} \|T_\Omega\theta\|_{L^2}^2 = \sup_{\|\theta\|_2=1} \mathbb{E}_X[\langle \theta, \phi(X, \Omega)\rangle].$$

Appendix D. Riemannian Geometry and M-estimation Background

This appendix collects the minimal differential-geometric and M-estimation background used in our quotient/descriptor-manifold asymptotic analysis. Our Riemannian conventions follow Lee (2018). For classical references on Euclidean M-estimation, see Van der Vaart (2000).

D.1. Basic Riemannian notions

Smooth manifolds and tangent spaces. A smooth q-dimensional manifold $\mathcal{M}$ is a Hausdorff, second-countable topological space equipped with a smooth atlas $\{(U_\alpha, \varphi_\alpha)\}$, where $\varphi_\alpha : U_\alpha \to$

$\varphi_\alpha(U_\alpha) \subseteq \mathbb{R}^q$ is a homeomorphism and all transition maps $\varphi_\beta \circ \varphi_\alpha^{-1}$ are smooth on overlaps. For $\Omega \in \mathcal{M}$, the tangent space $T_\Omega\mathcal{M}$ is a q-dimensional real vector space. For a smooth map $F : \mathcal{M} \to \mathcal{N}$, we write $dF_\Omega : T_\Omega\mathcal{M} \to T_{F(\Omega)}\mathcal{N}$ for its differential. A smooth vector field X assigns to each $\Omega \in \mathcal{M}$ a vector $X(\Omega) \in T_\Omega\mathcal{M}$ smoothly in charts.

Riemannian metrics and induced metrics. A Riemannian metric on $\mathcal{M}$ is a choice of inner product $\langle \cdot, \cdot \rangle_\Omega$ on each tangent space $T_\Omega\mathcal{M}$ that depends smoothly on Ω. The pair $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ is called a Riemannian manifold. If $\mathcal{M} \subseteq \mathbb{R}^{q_0}$ is an embedded C^k submanifold of $\mathbb{R}^{q_0}$, the induced (ambient) metric is $\langle u, v \rangle_\Omega \coloneqq u^\top v$ for $u, v \in T_\Omega\mathcal{M} \subseteq \mathbb{R}^{q_0}$.

For a piecewise C^1 curve $\gamma : [0, 1] \to \mathcal{M}$, its length is

$$L(\gamma) = \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\, dt, \qquad \|v\|_\Omega \coloneqq \sqrt{\langle v, v \rangle_\Omega}.$$

The associated Riemannian distance is defined by

$$d_{\mathcal{M}}(\Omega_1, \Omega_2) = \inf_{\gamma(0)=\Omega_1,\ \gamma(1)=\Omega_2} L(\gamma),$$

where the infimum ranges over piecewise C^1 curves $\gamma : [0, 1] \to \mathcal{M}$.

Levi–Civita connection, geodesics, gradients, and Hessians. There is a unique connection ∇ on $\mathcal{M}$ (the Levi–Civita connection) that is torsion-free and metric-compatible. A C^2 curve γ is a geodesic if it satisfies $\nabla_{\dot\gamma(t)}\dot\gamma(t) = 0$ for all t.

For a smooth function $f : \mathcal{M} \to \mathbb{R}$, the Riemannian gradient $\operatorname{grad} f(\Omega) \in T_\Omega\mathcal{M}$ is defined by the identity

$$df_\Omega(v) = \langle \operatorname{grad} f(\Omega), v \rangle_\Omega, \qquad v \in T_\Omega\mathcal{M}.$$

The Riemannian Hessian at point Ω is the linear map $\operatorname{Hess} f(\Omega) : T_\Omega\mathcal{M} \to T_\Omega\mathcal{M}$ given by

$$\operatorname{Hess} f(\Omega)[v] \coloneqq \nabla_V(\operatorname{grad} f)(\Omega),$$

where V is any smooth local extension of $v \in T_\Omega\mathcal{M}$. We also use the associated bilinear form as

$$\operatorname{Hess} f(\Omega)(u, v) = \langle u, \operatorname{Hess} f(\Omega)[v] \rangle_\Omega.$$

Exponential map, normal neighborhoods, and logarithm map. Fix $\Omega \in \mathcal{M}$. For each $v \in T_\Omega\mathcal{M}$, let $\gamma_v : [0, 1] \to \mathcal{M}$ denote the unique geodesic with $\gamma_v(0) = \Omega$ and $\dot\gamma_v(0) = v$. The exponential map at Ω is

$$\exp_\Omega : T_\Omega\mathcal{M} \to \mathcal{M}, \qquad \exp_\Omega(v) \coloneqq \gamma_v(1).$$

Moreover, there exists $\delta_\Omega > 0$ such that $\exp_\Omega$ restricts to a diffeomorphism from $B(\Omega, \delta_\Omega) \subset T_\Omega\mathcal{M}$ onto its image. On any set where this restriction is invertible, we write $\log_\Omega$ for the local inverse of $\exp_\Omega$.

Definition D.1 (Normal neighborhood and normal coordinates) *An open subset $U \subseteq \mathcal{M}$ is a* normal neighborhood *of Ω if there exists a $\delta > 0$ such that $U = \exp_\Omega(B(\Omega, \delta))$ and the restriction $\exp_\Omega : B(\Omega, \delta) \to U$ is a diffeomorphism. On such a U, the logarithm map at Ω is the inverse chart $\log_\Omega : U \to T_\Omega\mathcal{M}$, and the resulting chart $\log_\Omega$ defines the* normal coordinates *at Ω.*

D.2. Taylor expansions in normal coordinates

This subsection records the Taylor expansions we use to linearize first-order optimality conditions on the descriptor manifold.

Normal coordinates and pullbacks. Let $U \subseteq \mathcal{M}$ be a normal neighborhood of Ω (Definition D.1). Given a function $f : \mathcal{M} \to \mathbb{R}$, we pull f back to the tangent space via

$$\tilde{f}(v) \coloneqq f(\exp_\Omega(v)), \qquad v \in B(\Omega, \delta) \subseteq T_\Omega\mathcal{M}.$$

Equivalently, for $\Omega' \in U$, we write $v = \log_\Omega(\Omega')$ and view $\tilde{f}$ as f expressed in normal coordinates around Ω.

First- and second-order expansions of a smooth function. Let U be a normal neighborhood of Ω_1 and write $\Omega_2 = \exp_{\Omega_1}(v)$ with $v \in T_{\Omega_1}\mathcal{M}$. If f is $C^1(\mathcal{M}, \mathbb{R})$ on U, then as $v \to 0$ in $T_{\Omega_1}\mathcal{M}$,

$$f(\exp_{\Omega_1}(v)) = f(\Omega_1) + \langle \operatorname{grad}, f(\Omega_1), v\rangle_{\Omega_1} + o(\|v\|_{\Omega_1}). \tag{D.1}$$

If f is $C^2(\mathcal{M}, \mathbb{R})$ on U, then as $v \to 0$ in $T_{\Omega_1}\mathcal{M}$,

$$f(\exp_{\Omega_1}(v)) = f(\Omega_1) + \langle \operatorname{grad}, f(\Omega_1), v\rangle_{\Omega_1} + \frac{1}{2}\big\langle v,\ \operatorname{Hess}, f(\Omega_1)[v]\big\rangle_{\Omega_1} + o(\|v\|^2_{\Omega_1}). \tag{D.2}$$

The restriction to a normal neighborhood ensures that nearby points admit a unique representation via the logarithm map, so that the above expansions are well-defined and intrinsic.

Taylor expansion of the gradient via parallel transport. Fix $\Omega_1 \in \mathcal{M}$ and let U be a normal neighborhood of Ω_1. For $v \in T_{\Omega_1}\mathcal{M}$ sufficiently small, define the unique geodesic

$$\gamma(t) \coloneqq \exp_{\Omega_1}(tv), \qquad t \in [0, 1],$$

so that $\gamma(0) = \Omega_1$ and $\gamma(1) = \Omega_2 \coloneqq \exp_{\Omega_1}(v)$.

Parallel transport. Let $\mathcal{P}_{\Omega_1 \to \Omega_2} : T_{\Omega_1}\mathcal{M} \to T_{\Omega_2}\mathcal{M}$ denote parallel transport along γ. Given $w \in T_{\Omega_1}\mathcal{M}$, let $W(t) \in T_{\gamma(t)}\mathcal{M}$ be the unique vector field along γ that has the following property

$$\nabla_{\dot\gamma(t)} W(t) = 0, \qquad W(0) = w,$$

and set $\mathcal{P}_{\Omega_1 \to \Omega_2} w \coloneqq W(1)$. We write $\mathcal{P}_{\Omega_2 \to \Omega_1} \coloneqq \mathcal{P}^{-1}_{\Omega_1 \to \Omega_2}$.

Gradient expansion. If $f : \mathcal{M} \to \mathbb{R}$ is C^2 on U, then as $v \to 0$,

$$\mathcal{P}_{\Omega_2 \to \Omega_1}\big(\operatorname{grad} f(\Omega_2)\big) = \operatorname{grad} f(\Omega_1) + \operatorname{Hess} f(\Omega_1)[v] + r_f(v), \tag{D.3}$$

where $r_f(v) = o(\|v\|_{\Omega_1})$. If, in addition, $\operatorname{Hess} f$ is locally Lipschitz on U (with respect to $d_{\mathcal{M}}$), then there exist constants $C > 0$ and $\varepsilon > 0$ such that

$$\|r_f(v)\|_{\Omega_1} \leqslant C \|v\|^2_{\Omega_1} \qquad \text{for all } \|v\|_{\Omega_1} \leqslant \varepsilon. \tag{D.4}$$

Empirical objectives. The expansion (D.3) applies to empirical objectives of the form

$$\hat{f} \coloneqq \frac{1}{m}\sum_{i=1}^{m} f_i,$$

provided each f_i is C^2 on a common normal neighborhood U of $\Omega_\star$, and the corresponding derivatives admit local bounds on U (so that the remainder is uniform for $\|v\|_{\Omega_\star}$ small). In particular, for $v \in T_{\Omega_\star}\mathcal{M}$ small and $\hat{\Omega} \coloneqq \exp_{\Omega_\star}(v)$,

$$\mathcal{P}_{\hat{\Omega}\to\Omega_\star}\big(\operatorname{grad} \hat{f}(\hat{\Omega})\big) = \operatorname{grad} \hat{f}(\Omega_\star) + \operatorname{Hess} \hat{f}(\Omega_\star)[v] + r_{\hat{f}}(v), \tag{D.5}$$

where $r_{\hat{f}}(v) = o(\|v\|_{\Omega_\star})$ as $v \to 0$. If, in addition, $\operatorname{Hess} \hat{f}$ is locally Lipschitz on U, then $\|r_{\hat{f}}(v)\|_{\Omega_\star} = O(\|v\|_{\Omega_\star}^2)$.

Remark (normal coordinates vs. parallel transport). In normal coordinates at Ω, one may view (D.3) as an ordinary Euclidean Taylor expansion of the pullback $\tilde{f}(v) = f(\exp_\Omega(v))$ at $v = 0$. The parallel-transport form is convenient because it compares vectors in a single space $T_\Omega\mathcal{M}$.

D.3. Euclidean M-estimation: consistency and asymptotic normality

We first review standard Euclidean M-estimation. Note that all assumptions and results presented in this section are classical (Van der Vaart, 2000); our purpose for recording these arguments here is that we will follow their structure closely when generalizing to the Riemannian setting in Appendix D.4.

Let $(\mathcal{Z}, \mathcal{G})$ be a measurable space and let $Z_1, \dots, Z_m$ be i.i.d. with law $\mathbb{P}$ on $\mathcal{Z}$. Let $\Theta \subseteq \mathbb{R}^p$ be the parameter set. Given a measurable loss $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$, define

$$\hat{L}_m(\theta) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \ell(\theta; Z_i), \qquad L(\theta) \coloneqq \mathbb{E}[\ell(\theta; Z)],$$

where $Z \sim \mathbb{P}$ and we assume $L(\theta)$ is well-defined (possibly $+\infty$) for all $\theta \in \Theta$. An *M-estimator* is any measurable selection $\hat{\theta}_m$ from the set of empirical minimizers,

$$\hat{\theta}_m \in \operatorname*{arg\,min}_{\theta\in\Theta} \hat{L}_m(\theta),$$

whenever this set is nonempty.

Assumption D.2 (Euclidean M-estimation conditions) *There exist $\theta_\star \in \Theta$, an open set $U \subseteq \mathbb{R}^p$ with $\theta_\star \in U$, and $r > 0$ with $\overline{B}(\theta_\star, r) \subseteq U \cap \Theta$ such that:*

(i) ***(Identification and separation).*** *$\theta_\star$ is the unique minimizer of L on Θ and for every $\epsilon > 0$,*

$$\inf_{\theta\in\Theta:\ \|\theta-\theta_\star\|\geqslant\epsilon} \big(L(\theta) - L(\theta_\star)\big) > 0.$$

(ii) ***(Uniform LLN on a compact set).*** *On the compact set $\overline{B}(\theta_\star, r)$, we have*

$$\sup_{\theta\in\overline{B}(\theta_\star,r)} \big|\hat{L}_m(\theta) - L(\theta)\big| \xrightarrow{\ \mathbb{P}\ } 0,$$

and $\hat{\theta}_m \in \overline{B}(\theta_\star, r)$ with probability tending to one.

(iii) ***(Local C^2 smoothness and score moments).*** *For $\mathbb{P}$-a.e. z, the map $\theta \mapsto \ell(\theta; z)$ is C^2 on U, and $\mathbb{E}[\|\nabla \ell(\theta_\star; Z)\|^2] < \infty$.*

(iv) ***(Nondegenerate minimizer).*** *The matrix $H_\star := \nabla^2 L(\theta_\star)$ is invertible.*

(v) ***(Uniform Hessian convergence on $\overline{B}(\theta_\star, r)$).***

$$\sup_{\theta \in \overline{B}(\theta_\star, r)} \left\| \nabla^2 \hat{L}_m(\theta) - \nabla^2 L(\theta) \right\| \xrightarrow{\mathbb{P}} 0.$$

Define

$$\Sigma_\star := \mathrm{Var}\big(\nabla \ell(\theta_\star; Z)\big) = \mathbb{E}\Big[\nabla \ell(\theta_\star; Z) \nabla \ell(\theta_\star; Z)^\top\Big],$$

where $\mathbb{E}[\nabla \ell(\theta_\star; Z)] = \nabla L(\theta_\star) = 0$.

Proposition D.3 (Euclidean M-estimator consistency) *Assume Assumption D.2 (i)–(ii). Then*

$$\hat{\theta}_m \xrightarrow{\mathbb{P}} \theta_\star.$$

Proof The argument follows standard M-estimation proofs (see, e.g., Van der Vaart (2000)); we include it as a reference, since we will soon generalize this argument to Riemannian manifolds.

Fix $\epsilon > 0$ and define the separation gap

$$\Delta_\epsilon := \inf_{\theta \in \Theta:\ \|\theta - \theta_\star\| \geqslant \epsilon} \big(L(\theta) - L(\theta_\star)\big),$$

so that $\Delta_\epsilon > 0$ by Assumption D.2 (i). Consider the event

$$E_m := \left\{ \sup_{\theta \in \overline{B}(\theta_\star, r)} \big|\hat{L}_m(\theta) - L(\theta)\big| \leqslant \frac{\Delta_\epsilon}{3} \right\} \cap \{\hat{\theta}_m \in \overline{B}(\theta_\star, r)\}.$$

By the uniform law of large numbers on $\overline{B}(\theta_\star, r)$ and the localization $\mathbb{P}(\hat{\theta}_m \in \overline{B}(\theta_\star, r)) \to 1$, we have $\mathbb{P}(E_m) \to 1$.

On the event E_m, for any $\theta \in \overline{B}(\theta_\star, r)$ with $\|\theta - \theta_\star\| \geqslant \epsilon$ we have

$$\hat{L}_m(\theta) \ \geqslant\ L(\theta) - \frac{\Delta_\epsilon}{3} \ \geqslant\ L(\theta_\star) + \Delta_\epsilon - \frac{\Delta_\epsilon}{3} \ =\ L(\theta_\star) + \frac{2\Delta_\epsilon}{3},$$

while

$$\hat{L}_m(\theta_\star) \ \leqslant\ L(\theta_\star) + \frac{\Delta_\epsilon}{3}.$$

Hence, on E_m,

$$\inf_{\theta \in \overline{B}(\theta_\star, r):\ \|\theta - \theta_\star\| \geqslant \epsilon} \hat{L}_m(\theta) \ >\ \hat{L}_m(\theta_\star).$$

In particular, any minimizer of $\hat{L}_m$ over $\overline{B}(\theta_\star, r)$ must lie in $B(\theta_\star, \epsilon)$. Since $\hat{\theta}_m \in \overline{B}(\theta_\star, r)$ on E_m and $\hat{\theta}_m$ is (by assumption) an empirical minimizer, we conclude that $\|\hat{\theta}_m - \theta_\star\| < \epsilon$ on E_m. Therefore,

$$\mathbb{P}\big(\|\hat{\theta}_m - \theta_\star\| \geqslant \epsilon\big) \ \leqslant\ \mathbb{P}(E_m^c) \ \to\ 0,$$

which proves $\hat{\theta}_m \to \theta_\star$ in probability. ∎

Theorem D.4 (Euclidean M-estimator CLT) *Under Assumption D.2,*

$$\sqrt{n}\,(\hat{\theta}_n - \theta_\star) \overset{d}{\rightsquigarrow} \mathcal{N}\big(0,\ H_\star^{-1}\Sigma_\star H_\star^{-1}\big).$$

Proof The argument follows standard M-estimation proofs (see, e.g., Van der Vaart (2000)); we include it as a reference, since we will soon generalize this argument to Riemannian manifolds.

By Proposition D.3 and Assumption D.2 (i)–(ii), we have

$$\hat{\theta}_m \xrightarrow{\mathbb{P}} \theta_\star.$$

In particular, since $\hat{\theta}_m \in \overline{B}(\theta_\star, r)$ with probability tending to one by Assumption D.2 (ii), all arguments below may be restricted to the event $\{\hat{\theta}_m \in \overline{B}(\theta_\star, r)\}$.

On the event $\{\hat{\theta}_m \in \overline{B}(\theta_\star, r)\}$, the first-order condition for the empirical minimizer gives

$$\nabla \hat{L}_m(\hat{\theta}_m) = 0.$$

Since $\ell(\,\cdot\,; z)$ is C^2 on U for $\mathbb{P}$-a.e. z by Assumption D.2 (iii), the map $\theta \mapsto \nabla \hat{L}_m(\theta)$ is differentiable on U and we may apply the mean-value form of Taylor's theorem: there exists a point $\tilde{\theta}_m$ on the line segment between $\theta_\star$ and $\hat{\theta}_m$ such that

$$0 = \nabla \hat{L}_m(\hat{\theta}_m) = \nabla \hat{L}_m(\theta_\star) + \nabla^2 \hat{L}_m(\tilde{\theta}_m)\,(\hat{\theta}_m - \theta_\star). \tag{D.6}$$

Rearranging yields

$$\sqrt{m}\,(\hat{\theta}_m - \theta_\star) = -\Big(\nabla^2 \hat{L}_m(\tilde{\theta}_m)\Big)^{-1} \sqrt{m}\,\nabla \hat{L}_m(\theta_\star), \tag{D.7}$$

on the event that $\nabla^2 \hat{L}_m(\tilde{\theta}_m)$ is invertible.

Since $\tilde{\theta}_m$ lies on the segment between $\theta_\star$ and $\hat{\theta}_m$, we have $\tilde{\theta}_m \in \overline{B}(\theta_\star, r)$ whenever $\hat{\theta}_m \in \overline{B}(\theta_\star, r)$. Moreover, by consistency $\hat{\theta}_m \to \theta_\star$ in probability, hence $\tilde{\theta}_m \to \theta_\star$ in probability as well. By the uniform Hessian convergence in Assumption D.2 (v),

$$\sup_{\theta \in \overline{B}(\theta_\star, r)} \big\|\nabla^2 \hat{L}_m(\theta) - \nabla^2 L(\theta)\big\| \xrightarrow{\mathbb{P}} 0,$$

and therefore

$$\nabla^2 \hat{L}_m(\tilde{\theta}_m) - \nabla^2 L(\tilde{\theta}_m) \xrightarrow{\mathbb{P}} 0.$$

Since L is twice differentiable at $\theta_\star$ and $\tilde{\theta}_m \to \theta_\star$ in probability, we also have $\nabla^2 L(\tilde{\theta}_m) \to \nabla^2 L(\theta_\star) = H_\star$ in probability. Combining these gives

$$\nabla^2 \hat{L}_m(\tilde{\theta}_m) \xrightarrow{\mathbb{P}} H_\star. \tag{D.8}$$

By Assumption D.2 (iv), $H_\star$ is invertible, hence by continuity of matrix inversion,

$$\Big(\nabla^2 \hat{L}_m(\tilde{\theta}_m)\Big)^{-1} \xrightarrow{\mathbb{P}} H_\star^{-1}. \tag{D.9}$$

In particular, $\nabla^2 \hat{L}_m(\tilde{\theta}_m)$ is invertible with probability tending to one.

By definition,

$$\nabla \hat{L}_m(\theta_\star) = \frac{1}{m}\sum_{i=1}^{m} \nabla \ell(\theta_\star; Z_i).$$

Since $\theta_\star$ is a minimizer of L and L is differentiable at $\theta_\star$, we have $\mathbb{E}[\nabla \ell(\theta_\star; Z)] = \nabla L(\theta_\star) = 0$, hence the summands are mean-zero. By Assumption D.2 (iii), $\mathbb{E}[\|\nabla \ell(\theta_\star; Z)\|^2] < \infty$, so the multivariate CLT yields

$$\sqrt{m}\, \nabla \hat{L}_m(\theta_\star) \overset{d}{\rightsquigarrow} \mathcal{N}(0, \Sigma_\star). \tag{D.10}$$

Combining (D.7), (D.9), and (D.10), and applying Slutsky's theorem, we obtain

$$\sqrt{m}\,(\hat{\theta}_m - \theta_\star) \overset{d}{\rightsquigarrow} -H_\star^{-1} G, \qquad G \sim \mathcal{N}(0, \Sigma_\star).$$

Since $-H_\star^{-1} G \sim \mathcal{N}(0, H_\star^{-1}\Sigma_\star H_\star^{-1})$, this proves the claim. ■

D.4. M-estimation on a Riemannian manifold

Let $(\mathcal{M}, \langle\cdot,\cdot\rangle)$ be a finite-dimensional C^2 Riemannian manifold. Let $(\mathcal{Z}, \mathcal{G})$ be a measurable space and let $Z_1, \ldots, Z_m$ be i.i.d. with common law $\mathbb{P}$ on $\mathcal{Z}$. Let $f : \mathcal{M} \times \mathcal{Z} \to \mathbb{R}$ be a measurable loss such that the expectations below are well-defined. Define

$$\hat{F}_m(\Omega) \coloneqq \frac{1}{m}\sum_{i=1}^{m} f(\Omega; Z_i), \qquad F(\Omega) \coloneqq \mathbb{E}[f(\Omega; Z)], \qquad \Omega \in \mathcal{M}.$$

An *M-estimator* is a measurable map $\hat{\Omega}_m = \hat{\Omega}_m(Z_1, \ldots, Z_m)$ such that $\hat{\Omega}_m \in \arg\min_{\Omega\in\mathcal{M}} \hat{F}_m(\Omega)$ whenever the argmin set is nonempty.

Fix $\Omega_\star \in \mathcal{M}$. For $\epsilon > 0$, define the tangent-space ball

$$B(\Omega_\star, \epsilon) \coloneqq \{v \in T_{\Omega_\star}\mathcal{M} : \ \|v\|_{\Omega_\star} < \epsilon\}, \qquad \overline{B}(\Omega_\star, \epsilon) \coloneqq \{v \in T_{\Omega_\star}\mathcal{M} : \ \|v\|_{\Omega_\star} \leqslant \epsilon\}.$$

Assumption D.5 (Riemannian M-estimation conditions) *There exist $\Omega_\star \in \mathcal{M}$, a normal neighborhood $U = \exp_{\Omega_\star}\big(B(\Omega_\star, \epsilon_0)\big)$, and $\epsilon' \in (0, \epsilon_0)$ such that $\exp_{\Omega_\star}\big(\overline{B}(\Omega_\star, \epsilon')\big) \subseteq U$ and:*

(i) ***(Identification and separation).*** *$\Omega_\star$ is the unique minimizer of F on $\mathcal{M}$ and for every $\epsilon > 0$,*

$$\inf_{\Omega\in\mathcal{M}:\ d_{\mathcal{M}}(\Omega,\Omega_\star)\geqslant\epsilon} \big(F(\Omega) - F(\Omega_\star)\big) > 0.$$

(ii) ***(Uniform LLN on a compact set).*** *On the compact set $\exp_{\Omega_\star}\big(\overline{B}(\Omega_\star, \epsilon')\big)$, we have*

$$\sup_{\Omega\in\exp_{\Omega_\star}(\overline{B}(\Omega_\star,\epsilon'))} \big|\hat{F}_m(\Omega) - F(\Omega)\big| \xrightarrow{\mathbb{P}} 0,$$

and $\hat{\Omega}_m \in \exp_{\Omega_\star}\big(\overline{B}(\Omega_\star, \epsilon')\big)$ with probability tending to one.

(iii) ***(Local C^2 smoothness and score moments).*** *For $\mathbb{P}$-a.e. z, the map $\Omega \mapsto f(\Omega; z)$ is C^2 on U, and*

$$\mathbb{E}\big[\|\operatorname{grad} f(\Omega_\star; Z)\|_{\Omega_\star}^2\big] < \infty.$$

(iv) ***(Nondegenerate minimizer).*** *The linear map* $H_\star := \operatorname{Hess} F(\Omega_\star) : T_{\Omega_\star}\mathcal{M} \to T_{\Omega_\star}\mathcal{M}$ *is invertible.*

(v) ***(Uniform transported Hessian convergence on*** $\exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon'))$***).*** *For* $\Omega \in \exp_{\Omega_\star}\big(\overline{B}(\Omega_\star, \epsilon')\big)$*, define*

$$\widetilde{H}_m(\Omega) := \mathcal{P}_{\Omega \to \Omega_\star} \circ \operatorname{Hess} \hat{F}_m(\Omega) \circ \mathcal{P}_{\Omega_\star \to \Omega}, \qquad \widetilde{H}(\Omega) := \mathcal{P}_{\Omega \to \Omega_\star} \circ \operatorname{Hess} F(\Omega) \circ \mathcal{P}_{\Omega_\star \to \Omega}.$$

Then

$$\sup_{\Omega \in \exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon'))} \big\|\widetilde{H}_m(\Omega) - \widetilde{H}(\Omega)\big\| \xrightarrow{\mathbb{P}} 0.$$

Define the covariance operator $\Sigma_\star$ on $T_{\Omega_\star}\mathcal{M}$ by

$$\langle u,\ \Sigma_\star v\rangle_{\Omega_\star} = \mathbb{E}\Big[\langle u,\ \operatorname{grad} f(\Omega_\star; Z)\rangle_{\Omega_\star}\ \langle v,\ \operatorname{grad} f(\Omega_\star; Z)\rangle_{\Omega_\star}\Big], \qquad u, v \in T_{\Omega_\star}\mathcal{M}.$$

Proposition D.6 (Riemannian M-estimator consistency) *Assume Assumption D.5 (i)–(ii). Then*

$$\hat{\Omega}_m \xrightarrow{\mathbb{P}} \Omega_\star.$$

Proof Fix $\epsilon > 0$ and define the separation gap

$$\Delta_\epsilon := \inf_{\Omega \in \mathcal{M}:\ d_{\mathcal{M}}(\Omega, \Omega_\star) \geqslant \epsilon} \big(F(\Omega) - F(\Omega_\star)\big),$$

so that $\Delta_\epsilon > 0$ by Assumption D.5 (i). Consider the event

$$E_m := \left\{ \sup_{\Omega \in \exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon'))} \big|\hat{F}_m(\Omega) - F(\Omega)\big| \leqslant \frac{\Delta_\epsilon}{3} \right\} \cap \left\{ \hat{\Omega}_m \in \exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon')) \right\}.$$

By the uniform law of large numbers on $\exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon'))$ and the localization in Assumption D.5 (ii), we have $\mathbb{P}(E_m) \to 1$.

On the event E_m, for any $\Omega \in \exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon'))$ with $d_{\mathcal{M}}(\Omega, \Omega_\star) \geqslant \epsilon$ we have

$$\hat{F}_m(\Omega) \ \geqslant\ F(\Omega) - \frac{\Delta_\epsilon}{3} \ \geqslant\ F(\Omega_\star) + \Delta_\epsilon - \frac{\Delta_\epsilon}{3} \ =\ F(\Omega_\star) + \frac{2\Delta_\epsilon}{3},$$

while

$$\hat{F}_m(\Omega_\star) \ \leqslant\ F(\Omega_\star) + \frac{\Delta_\epsilon}{3}.$$

Hence, on E_m,

$$\inf_{\Omega \in \exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon')):\ d_{\mathcal{M}}(\Omega, \Omega_\star) \geqslant \epsilon} \hat{F}_m(\Omega) \ >\ \hat{F}_m(\Omega_\star).$$

In particular, any minimizer of $\hat{F}_m$ over $\exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon'))$ must lie in the metric ball

$$B_{d_{\mathcal{M}}}(\Omega_\star, \epsilon) := \{\Omega \in \mathcal{M} :\ d_{\mathcal{M}}(\Omega, \Omega_\star) < \epsilon\}.$$

Since $\hat{\Omega}_m \in \exp_{\Omega_\star}(\overline{B}(\Omega_\star, \epsilon'))$ on E_m and $\hat{\Omega}_m$ is (by definition) an empirical minimizer, we conclude that $d_{\mathcal{M}}(\hat{\Omega}_m, \Omega_\star) < \epsilon$ on E_m. Therefore,

$$\mathbb{P}\big(d_{\mathcal{M}}(\hat{\Omega}_m, \Omega_\star) \geqslant \epsilon\big) \ \leqslant\ \mathbb{P}(E_m^c) \ \to\ 0,$$

which proves $\hat{\Omega}_m \to \Omega_\star$ in probability. ∎

Theorem D.7 (Riemannian M-estimator CLT) *Under Assumption D.5, we have*

$$\sqrt{m}\,\log_{\Omega_\star}(\hat{\Omega}_m) \overset{d}{\rightsquigarrow} \mathcal{N}\big(0,\ H_\star^{-1}\Sigma_\star H_\star^{-1}\big),$$

where

$$H_\star \coloneqq \operatorname{Hess} F(\Omega_\star) : T_{\Omega_\star}\mathcal{M} \to T_{\Omega_\star}\mathcal{M}$$

and

$$\Sigma_\star \coloneqq \operatorname{Var}\big(\operatorname{grad} f(\Omega_\star; Z)\big) = \mathbb{E}\Big[\operatorname{grad} f(\Omega_\star; Z)\operatorname{grad} f(\Omega_\star; Z)^\top\Big],$$

with $\mathbb{E}[\operatorname{grad} f(\Omega_\star; Z)] = \operatorname{grad} F(\Omega_\star) = 0$.

Remark D.8 *In Theorem D.7, the logarithm map $\log_{\Omega_\star}$ is defined on a normal neighborhood of $\Omega_\star$. While $\hat{\Omega}_m$ need not belong to this neighborhood for each finite m, consistency ensures that $\hat{\Omega}_m \to \Omega_\star$ in probability. Consequently, $\log_{\Omega_\star}(\hat{\Omega}_m)$ is well-defined with probability tending to one.*

Proof By Proposition D.6 and Assumption D.5 (i)–(ii), we have

$$\hat{\Omega}_m \xrightarrow{\mathbb{P}} \Omega_\star.$$

Fix a normal neighborhood $U = \exp_{\Omega_\star}(B_{\Omega_\star}(\epsilon_0))$ and define $A_m \coloneqq \{\hat{\Omega}_m \in U\}$. Since $\hat{\Omega}_m \xrightarrow{\mathbb{P}} \Omega_\star$ and U is an open neighborhood of $\Omega_\star$, we have $\mathbb{P}(A_m) \to 1$. In Definition D.1, we defined the logarithm map $\log_{\Omega_\star} : U \to T_{\Omega_\star}\mathcal{M}$ as the unique inverse of the exponential map $\exp_{\Omega_\star}$. Define

$$v_m \coloneqq \begin{cases} \log_{\Omega_\star}(\hat{\Omega}_m) \in T_{\Omega_\star}\mathcal{M}, & \text{on } A_m, \\ 0 \in T_{\Omega_\star}\mathcal{M}, & \text{on } A_m^c. \end{cases}$$

Then, v_m is well-defined globally and satisfies $v_m \to 0$ in probability.

On the event A_m, the first-order condition for an empirical minimizer gives

$$\operatorname{grad} \hat{F}_m(\hat{\Omega}_m) = 0.$$

Applying the transported gradient expansion (D.5) with $\Omega_1 = \Omega_\star$, $\Omega_2 = \hat{\Omega}_m = \exp_{\Omega_\star}(v_m)$, and $f = \hat{F}_m$, we obtain

$$0 = \mathcal{P}_{\hat{\Omega}_m \to \Omega_\star}\big(\operatorname{grad} \hat{F}_m(\hat{\Omega}_m)\big) = \operatorname{grad} \hat{F}_m(\Omega_\star) + \operatorname{Hess} \hat{F}_m(\Omega_\star)[v_m] + r_m,$$

where r_m denotes the Taylor remainder on A_m, and we define $r_m = 0$ on A_m^c. Moreover,

$$\|r_m\|_{\Omega_\star}\mathbf{1}_{A_m} = o_{\mathbb{P}}\big(\|v_m\|_{\Omega_\star}\big).$$

Rearranging yields

$$\sqrt{m}\,v_m = -\Big(\operatorname{Hess} \hat{F}_m(\Omega_\star)\Big)^{-1}\sqrt{m}\operatorname{grad} \hat{F}_m(\Omega_\star)\mathbf{1}_{A_m} - \Big(\operatorname{Hess} \hat{F}_m(\Omega_\star)\Big)^{-1}\sqrt{m}\,r_m\mathbf{1}_{A_m}, \quad \text{(D.11)}$$

on the event $\{\operatorname{Hess} \hat{F}_m(\Omega_\star) \text{ is invertible}\}$.

By Assumption D.5 (v) applied at $\Omega = \Omega_\star$ (so that $\mathcal{P}_{\Omega_\star \to \Omega_\star} = \mathrm{Id}$),

$$\big\|\operatorname{Hess} \hat{F}_m(\Omega_\star) - \operatorname{Hess} F(\Omega_\star)\big\| = \big\|\widetilde{H}_m(\Omega_\star) - \widetilde{H}(\Omega_\star)\big\| \xrightarrow{\mathbb{P}} 0.$$

Hence,

$$\operatorname{Hess} \hat{F}_m(\Omega_\star) \xrightarrow{\mathbb{P}} H_\star. \tag{D.12}$$

By Assumption D.5 (iv), $H_\star$ is invertible, and by continuity of inversion we have

$$\Big(\operatorname{Hess} \hat{F}_m(\Omega_\star)\Big)^{-1} \xrightarrow{\mathbb{P}} H_\star^{-1}. \tag{D.13}$$

In particular, $\operatorname{Hess} \hat{F}_m(\Omega_\star)$ is invertible with probability tending to one.

By definition,

$$\operatorname{grad} \hat{F}_m(\Omega_\star) = \frac{1}{m} \sum_{i=1}^{m} \operatorname{grad} f(\Omega_\star; Z_i).$$

Since $\Omega_\star$ is a minimizer of F and F is differentiable at $\Omega_\star$, we have $\mathbb{E}[\operatorname{grad} f(\Omega_\star; Z)] = \operatorname{grad} F(\Omega_\star) = 0$. By Assumption D.5 (iii), $\mathbb{E}[\|\operatorname{grad} f(\Omega_\star; Z)\|_{\Omega_\star}^2] < \infty$, so the multivariate CLT yields

$$\sqrt{m} \operatorname{grad} \hat{F}_m(\Omega_\star) \overset{d}{\rightsquigarrow} \mathcal{N}(0, \Sigma_\star). \tag{D.14}$$

Since $\|r_m\|_{\Omega_\star} \mathbf{1}_{A_m} = o\big(\|v_m\|_{\Omega_\star}\big)$ in probability on the event A_m, we have

$$\frac{\|r_m\|_{\Omega_\star}}{\|v_m\|_{\Omega_\star}} \mathbf{1}_{A_m} \xrightarrow{\mathbb{P}} 0,$$

with the convention that the ratio is set to 0 on the event $\{v_m = 0\}$. Moreover, from (D.11) and (D.13) we have $\big(\operatorname{Hess} \hat{F}_m(\Omega_\star)\big)^{-1} = O_{\mathbb{P}}(1)$ and $\sqrt{m} \operatorname{grad} \hat{F}_m(\Omega_\star) = O_{\mathbb{P}}(1)$, hence

$$\sqrt{m}\, \|v_m\|_{\Omega_\star} = O_{\mathbb{P}}(1).$$

Consequently,

$$\sqrt{m}\, \|r_m\|_{\Omega_\star} \mathbf{1}_{A_m} = \big(\sqrt{m}\, \|v_m\|_{\Omega_\star}\big) \cdot \frac{\|r_m\|_{\Omega_\star}}{\|v_m\|_{\Omega_\star}} \mathbf{1}_{A_m} \xrightarrow{\mathbb{P}} 0,$$

and therefore

$$\sqrt{m}\, r_m \mathbf{1}_{A_m} \xrightarrow{\mathbb{P}} 0. \tag{D.15}$$

Combining (D.11), (D.13), (D.14), and (D.15), and applying Slutsky's theorem, using that $\mathbf{1}_{A_m} \xrightarrow{\mathbb{P}} \mathbf{1}$, we obtain

$$\sqrt{m}\, v_m \overset{d}{\rightsquigarrow} -H_\star^{-1} G, \qquad G \sim \mathcal{N}(0, \Sigma_\star).$$

Finally, on A_m, we have $v_m = \log_{\Omega_\star}(\hat{\Omega}_m)$, while $\mathbb{P}(A_m) \to 1$. Hence, v_m and $\log_{\Omega_\star}(\hat{\Omega}_m)$ agree with probability tending to one, so they have the same asymptotic distribution. Since $-H_\star^{-1} G \sim \mathcal{N}(0, H_\star^{-1} \Sigma_\star H_\star^{-1})$, this proves the claim. ∎

Remark (Euclidean case as a special case). If $\mathcal{M} = \mathbb{R}^p$ with the Euclidean metric, then $\exp_{\theta_\star}(v) = \theta_\star + v$, $\log_{\theta_\star}(\theta) = \theta - \theta_\star$, geodesics are line segments, and parallel transport is the identity. In this case $\widetilde{H}_n(\theta) = \nabla^2 \hat{L}_n(\theta)$.

Relation to Brunel (2023). Theorem D.7 is closely related to and inspired by the asymptotic-normality result of Brunel (2023) for geodesically convex M-estimators. Brunel proves a tangent-space Gaussian limit on a complete Riemannian manifold under the following conditions: the sample losses are geodesically convex, the population objective has a unique minimizer, the population objective is twice differentiable at that minimizer with positive-definite Hessian, and the loss admits measurable subgradients satisfying a local second-moment condition.

The key difference is the mechanism used to obtain the local quadratic expansion underlying asymptotic normality. In Brunel's theorem, geodesic convexity is the structural assumption that allows for nonsmooth empirical objectives. The proof uses subgradient inequalities and convexity to compare the localized empirical objective in normal coordinates with a random quadratic approximation. Thus, geodesic convexity replaces the need for differentiability of the empirical criterion, a Taylor expansion of the empirical first-order condition, and uniform convergence of empirical Hessians. It does not however replace population-level second-order differentiability: Brunel still assumes that the population objective is twice differentiable at the minimizer with positive-definite Hessian.

Theorem D.7 takes a complementary route. We do not assume geodesic convexity of L_{pre}. Instead, after passing to the descriptor manifold that quotients out the symmetry, we directly impose the local regularity conditions needed for a smooth Riemannian M-estimation CLT. These conditions (cf. Assumption D.5) include local identification and separation of the population minimizer, localization and a uniform law of large numbers on a compact normal neighborhood, local C^2 smoothness of the sample loss, a nonsingular population Riemannian Hessian, score moment control, and uniform convergence of transported empirical Hessians. Under these assumptions, Theorem D.7 follows by Taylor expanding the transported first-order condition in $T_{\Omega_\star}\mathcal{M}$.

Thus, Theorem D.7 can be viewed as abstracting the local smooth assumptions needed for asymptotic normality away from the particular sufficient condition of geodesic convexity. When ℓ_{pre} is C^2 smooth, geodesic convexity together with additional empirical Hessian regularity verifies Assumption D.5, and the two approaches yield the same sandwich-form tangent-space CLT. However, Brunel's assumptions do not necessarily imply Assumption D.5, since Brunel's theorem allows nonsmooth empirical losses through measurable subgradients. Conversely, Theorem D.7 does not require geodesic convexity: any smooth descriptor-manifold pre-training problem satisfying Assumption D.5, whether geodesically convex or not, falls under Theorem D.7. Our formulation is necessary for the quotient-manifold pre-training problems considered here, where the natural objective is locally smooth after passing to descriptor coordinates, but need not be globally geodesically convex.

Appendix E. Symmetry, Identifiability, and Quotient Geometry

This appendix formalizes how symmetry-induced non-identifiability in pretraining is handled in our analysis.

E.1. Quotients by group actions and local descriptor charts

Smooth group actions. Let G be a Lie group acting smoothly on a smooth manifold $\mathcal{A}$. We write the action as $(g, a) \mapsto g \cdot a$. For $a \in \mathcal{A}$, the orbit is $[a] = \{g \cdot a : g \in G\}$ and the stabilizer (isotropy subgroup) is $G_a = \{g \in G : g \cdot a = a\}$. The orbit space (quotient set) is $\mathcal{A}/G = \{[a] : a \in \mathcal{A}\}$ with the quotient topology, and the canonical projection is denoted $\pi : \mathcal{A} \to \mathcal{A}/G$, $\pi(a) = [a]$.

Regular neighborhoods and smooth quotients. A smooth action is called *free* if $G_a = \{e\}$ for all $a \in \mathcal{A}$. It is called *proper* if the map $G \times \mathcal{A} \to \mathcal{A} \times \mathcal{A}$, $(g, a) \mapsto (a, g \cdot a)$ is proper. A sufficient condition for properness is that G is compact (e.g., an orthogonal group) or finite (e.g., a permutation group).

If the action of G on $\mathcal{A}$ is smooth, free, and proper on an open set $\mathcal{U} \subseteq \mathcal{A}$, then the orbit space

$$\mathcal{B} \coloneqq \mathcal{U}/G$$

admits a unique smooth manifold structure such that the projection $\pi : \mathcal{U} \to \mathcal{B}$ is a smooth submersion. In this case, $\pi : \mathcal{U} \to \mathcal{B}$ is a principal G-bundle. If the action is proper but not free, one may restrict attention to a regular stratum on which the orbit type is constant; on such a neighborhood the quotient is again a smooth manifold. This is the regime implicitly used in the main text.

Local quotient charts via invariant descriptors. Rather than working directly with the abstract quotient manifold $\mathcal{B}$, we represent a local neighborhood of $\mathcal{B}$ using an orbit-invariant *descriptor map*. The following assumption makes this precise.

Assumption E.1 (Local quotient chart via an invariant descriptor) *Let $\mathcal{U} \subseteq \mathcal{A}$ be an open set on which the action of G is smooth, free, and proper, and let $\mathcal{B} = \mathcal{U}/G$ with projection $\pi : \mathcal{U} \to \mathcal{B}$. There exists a map $D : \mathcal{U} \to \mathbb{R}^q$ such that:*

(i) ***(Orbit-constancy).*** *For all $a \in \mathcal{U}$ and $g \in G$,*

$$D(g \cdot a) = D(a).$$

(ii) ***(Local chart for the quotient).*** *There exist open neighborhoods $V \subseteq \mathcal{B}$ and $W \subseteq \mathbb{R}^q$, and a C^k diffeomorphism $\bar{D} : V \to W$, such that on $\pi^{-1}(V)$ we have*

$$D = \bar{D} \circ \pi.$$

Consequences. Under Assumption E.1, the following hold.

(i) **Local orbit separation.** For $a, a' \in \pi^{-1}(V)$,

$$D(a) = D(a') \quad \Longleftrightarrow \quad \pi(a) = \pi(a') \quad \Longleftrightarrow \quad a' \in [a].$$

(ii) **Descriptor manifold.** We identify the local quotient neighborhood $V \subseteq \mathcal{B}$ with its descriptor coordinates $W \subseteq \mathbb{R}^q$ via $\bar{D}$, and write

$$\mathcal{M} \coloneqq W.$$

Thus $\mathcal{M}$ is a smooth manifold (indeed, an open subset of $\mathbb{R}^q$) equipped with the induced Euclidean metric.

(iii) **Existence of smooth lifts.** Since $\pi : \mathcal{U} \to \mathcal{B}$ is a principal bundle, there exists a smooth local section $\sigma : V \to \mathcal{U}$. Defining

$$s \coloneqq \sigma \circ \bar{D}^{-1} : \mathcal{M} \to \mathcal{U},$$

we obtain a C^k *lift* satisfying $D(s(M)) = M$ for all $M \in \mathcal{M}$.

(iv) **Well-defined induced objectives.** Any G-invariant function on $\mathcal{U}$ induces a well-defined function on $\mathcal{M}$ by evaluation at any representative in $D^{-1}(M) \cap \pi^{-1}(V)$.

E.2. Vector-bundle viewpoint: quotient-level features and the coordinate feature map $\phi(x, M)$

This subsection formalizes how equivariant representative-level features induce intrinsic quotient-level features, and how coordinate feature maps arise from choosing local lifts.

E.2.1. SETUP: PRINCIPAL BUNDLE AND EQUIVARIANT FEATURES

Let $\mathcal{U} \subseteq \mathbb{R}^{q_0}$ be an open set on which a Lie group G acts smoothly, freely, and properly, and let $\mathcal{B} = \mathcal{U}/G$ with projection $\pi : \mathcal{U} \to \mathcal{B}$. Fix a feature dimension $p \in \mathbb{N}_+$ and let $\rho : G \to O(p)$ be a smooth group homomorphism. Let $\psi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^p$ be measurable in x and C^k in its second argument.

We assume the orthogonal equivariance condition

$$\psi(x, g \cdot A) = \rho(g)\, \psi(x, A), \qquad x \in \mathcal{X},\ A \in \mathcal{U},\ g \in G.$$

E.2.2. ASSOCIATED VECTOR BUNDLE AND INTRINSIC FEATURE SECTION

Define an equivalence relation on $\mathcal{U} \times \mathbb{R}^p$ by

$$(A, v) \sim (g \cdot A, \rho(g)v), \qquad g \in G.$$

The associated rank-p vector bundle over $\mathcal{B}$ is

$$\mathcal{E} \coloneqq (\mathcal{U} \times \mathbb{R}^p)/\sim,$$

with projection $\pi_{\mathcal{E}}([A, v]) = [A]$. The Euclidean inner product on $\mathbb{R}^p$ descends to a well-defined fiberwise inner product on $\mathcal{E}$.

For each $x \in \mathcal{X}$, define the intrinsic feature section

$$\Phi_x : \mathcal{B} \to \mathcal{E}, \qquad \Phi_x([A]) \coloneqq [A, \psi(x, A)].$$

Proposition E.2 (Well-definedness and smoothness) *Under the equivariance assumption above, Φ_x is well-defined. If $A \mapsto \psi(x, A)$ is C^k on $\mathcal{U}$, then Φ_x is a C^k section of $\mathcal{E}$.*

Proof If $A' = g \cdot A$, then by equivariance,

$$[A', \psi(x, A')] = [g \cdot A, \rho(g)\psi(x, A)] = [A, \psi(x, A)],$$

so Φ_x depends only on the orbit. Smoothness follows by expressing Φ_x in local trivializations induced by smooth local sections of the principal bundle $\pi : \mathcal{U} \to \mathcal{B}$. ■

E.2.3. DESCRIPTOR COORDINATES AND THE COORDINATE FEATURE MAP

Let $\mathcal{M} \subseteq \mathbb{R}^q$ be the descriptor manifold provided by Assumption E.1, and let $s : \mathcal{M} \to \mathcal{U}$ be the associated C^k lift, i.e., $D(s(\Omega)) = \Omega$ for all Ω in a local neighborhood.

Coordinate feature map. Define

$$\phi : \mathcal{X} \times \mathcal{M} \to \mathbb{R}^p, \qquad \phi(x, M) \coloneqq \psi(x, s(M)).$$

Lemma E.3 (Differentiability of ϕ) *If $A \mapsto \psi(x, A)$ is C^k and s is C^k, then for each fixed $x \in \mathcal{X}$ the map $M \mapsto \phi(x, M)$ is C^k on $\mathcal{M}$.*

Proof Fix $x \in \mathcal{X}$ and define $\psi_x : \mathcal{U} \to \mathbb{R}^p$ by $\psi_x(A) \coloneqq \psi(x, A)$. By assumption, ψ_x is a C^k map on $\mathcal{U}$, and s is C^k on $\mathcal{M}$. Hence $\phi_x : \mathcal{M} \to \mathbb{R}^p$ given by

$$\phi_x(M) \coloneqq \phi(x, M) = \psi_x\big(s(M)\big)$$

is the composition $\phi_x = \psi_x \circ s$ of two C^k maps between smooth manifolds, and is therefore C^k. ■

Gauge transformations. If s' is another C^k lift on $\mathcal{M}$, then for each $M \in \mathcal{M}$ there exists a unique $g(M) \in G$ such that $s'(M) = g(M) \cdot s(M)$. The map $M \mapsto g(M)$ is C^k.

Lemma E.4 (Gauge transformation rule) *If ϕ and ϕ' are induced by lifts s and s' respectively, then*

$$\phi'(x, M) = \rho(g(M))\, \phi(x, M), \qquad x \in \mathcal{X},\ M \in \mathcal{M}.$$

Proof By equivariance,

$$\phi'(x, M) = \psi(x, s'(M)) = \psi(x, g(M) \cdot s(M)) = \rho(g(M))\, \phi(x, M).$$

■

Intrinsic meaning. The intrinsic object is the bundle section Φ_x. Choosing a lift s identifies each fiber with $\mathbb{R}^p$ and yields the coordinate representation $\phi(x, M)$. Different lifts correspond to orthogonal changes of coordinates.

E.2.4. ORBIT-INVARIANCE OF MINIMUM-NORM OLS

The proof of Lemma 4.1 in the main text follows directly from orthogonal equivariance and is given below for completeness.

Proof [Proof of Lemma 4.1] Fix a downstream dataset $D^{(n)}_{\text{down}} = \{(x_i, y_i)\}_{i=1}^n$ and write $Y \coloneqq (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. For any $w \in \mathbb{R}^{q_0}$, define the design matrix $\Psi_w \in \mathbb{R}^{n \times p}$ by

$$(\Psi_w)_{i,:} \coloneqq \psi(x_i, w)^\top.$$

The minimum Euclidean norm solution of the OLS problem is given by

$$\hat{\theta}_w = \Psi_w^+ Y,$$

where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse.

By the orthogonal equivariance condition (4.3), for any $g \in G$ and each i,

$$\psi(x_i, g \cdot w)^\top = (\rho(g)\psi(x_i, w))^\top = \psi(x_i, w)^\top \rho(g)^\top.$$

Therefore,

$$\Psi_{g\cdot w} = \Psi_w \, \rho(g)^\top.$$

For any matrix $M \in \mathbb{R}^{n\times p}$ and any orthogonal matrix $Q \in O(p)$,

$$(MQ)^+ = Q^\top M^+.$$

Applying this identity with $M = \Psi_w$ and $Q = \rho(g)^\top$ yields

$$\hat{\theta}_{g\cdot w} = \Psi_{g\cdot w}^+ Y = (\Psi_w \rho(g)^\top)^+ Y = \rho(g)\Psi_w^+ Y = \rho(g)\hat{\theta}_w.$$

For any $x \in \mathcal{X}$,

$$\begin{aligned}
\hat{f}_{g\cdot w}(x) &= \langle \hat{\theta}_{g\cdot w}, \psi(x, g\cdot w)\rangle \\
&= \langle \rho(g)\hat{\theta}_w, \rho(g)\psi(x, w)\rangle \\
&= \langle \hat{\theta}_w, \psi(x, w)\rangle \\
&= \hat{f}_w(x),
\end{aligned}$$

where we used the orthogonality of $\rho(g)$ in the third equality. This shows that the minimum-norm downstream predictor depends on w only through its orbit $[w]$. ■

Appendix F. Proof of Proposition B.1

Recall the downstream regression model in Equation (3.1)

$$Y = f_\star(X) + \varepsilon, \qquad X \sim \mu_{\text{down}}, \qquad \mathbb{E}[\varepsilon \mid X] = 0, \qquad \sigma^2 := \mathbb{E}[\varepsilon^2 \mid X] < \infty.$$

Let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. copies of (X, Y) and write $D_{\text{down}}^{(n)} = \{(x_i, y_i)\}_{i=1}^n$ for the labeled sample and $X_{1:n} = (x_1, \ldots, x_n)$ for the downstream design. Let $(X_{\text{new}}, Y_{\text{new}})$ be an independent copy of (X, Y). Define $\varepsilon_i := y_i - f_\star(x_i)$ for $i \in [n]$ and $\varepsilon_{\text{new}} := Y_{\text{new}} - f_\star(X_{\text{new}})$.

Manifold-valued feature parameters and quenched conditioning. In the main text, the feature parameter is learned in pre-training: $\Omega = \hat{\Omega}_m(D_{\text{pre}}^{(m)})$, and Ω may take values on a Riemannian manifold $\mathcal{M}$. All identities below are deterministic once $(D_{\text{pre}}^{(m)}, X_{1:n})$ is fixed, because conditioning on $D_{\text{pre}}^{(m)}$ freezes Ω, and conditioning on $X_{1:n}$ freezes the empirical projection operator $\Pi_{\Omega,n}$. This viewpoint isolates the downstream label noise and the fresh test pair randomness, without averaging over pre-training and downstream design.

F.1. Empirical projection notation

Recall the empirical inner product

$$\langle g, h\rangle_n = \frac{1}{n}\sum_{i=1}^n g(x_i)h(x_i).$$

Let $\mathcal{H}_\Omega = \{T_\Omega \theta : \theta \in \mathbb{R}^p\}$ denote the induced linear class and $\Pi_{\Omega,n}$ be the Moore–Penrose empirical least-squares map defined in Appendix C (Definition C.7). We will invoke the following properties (Lemma C.8 and Lemma C.9): for any g with finite evaluations on $\{x_i\}_{i=1}^n$,

1. $\Pi_{\Omega,n} g \in \mathcal{H}_\Omega$ and $\langle g - \Pi_{\Omega,n} g,\, h\rangle_n = 0$ for all $h \in \mathcal{H}_\Omega$;
2. $\mathrm{Ev}_n(\Pi_{\Omega,n} h) = \mathrm{Ev}_n(h)$ for all $h \in \mathcal{H}_\Omega$ (equivalently, $\|h - \Pi_{\Omega,n} h\|_n = 0$).

When $\langle \cdot, \cdot \rangle_n$ is non-degenerate on $\mathcal{H}_\Omega$ (equivalently, Ev_n is injective on $\mathcal{H}_\Omega$), these properties imply $\Pi_{\Omega,n} h = h$ for all $h \in \mathcal{H}_\Omega$, i.e. $\Pi_{\Omega,n}$ is the unique empirical orthogonal projector onto $\mathcal{H}_\Omega$.

OLS as empirical projection. Let $\hat{\theta}_{\Omega,n}$ be the minimum-norm OLS solution and set $\hat{f}_{\Omega,n} = T_\Omega \hat{\theta}_{\Omega,n}$. Write the sample-value vectors

$$y_{1:n} \coloneqq (y_1, \dots, y_n)^\top, \qquad \varepsilon_{1:n} \coloneqq (\varepsilon_1, \dots, \varepsilon_n)^\top, \qquad f_{\star,1:n} \coloneqq \big(f_\star(x_1), \dots, f_\star(x_n)\big)^\top,$$

so that $y_{1:n} = f_{\star,1:n} + \varepsilon_{1:n}$. By Lemma C.11,

$$\hat{f}_{\Omega,n} = \Pi_{\Omega,n} \,\mathrm{lift}_n(y_{1:n}).$$

By linearity of $\Pi_{\Omega,n}$ and $y_{1:n} = f_{\star,1:n} + \varepsilon_{1:n}$,

$$\hat{f}_{\Omega,n} = \Pi_{\Omega,n} f_\star + \Pi_{\Omega,n} \,\mathrm{lift}_n(\varepsilon_{1:n}). \tag{F.1}$$

For brevity, we will write $\Pi_{\Omega,n}\varepsilon$ instead of $\Pi_{\Omega,n} \,\mathrm{lift}_n(\varepsilon_{1:n})$, with the understanding that this means applying $\Pi_{\Omega,n}$ to any measurable representative whose evaluations on $\{x_i\}_{i=1}^n$ equal $\varepsilon_{1:n}$.

F.2. Population decomposition and orthogonality

Recall the population projector $\Pi_\Omega = T_\Omega \Sigma(\Omega)^+ T_\Omega^{\mathrm{adj}}$ onto $\mathcal{H}_\Omega$ in $L^2(\mu_{\mathrm{down}})$ and define

$$e_\Omega \coloneqq (I - \Pi_\Omega) f_\star, \qquad \mathrm{Rep}(\Omega) \coloneqq \|e_\Omega\|^2_{L^2(\mu_{\mathrm{down}})}.$$

Since Π_Ω is the $L^2(\mu_{\mathrm{down}})$-orthogonal projector onto $\mathcal{H}_\Omega$, we have

$$e_\Omega \perp \mathcal{H}_\Omega \quad \text{in } L^2(\mu_{\mathrm{down}}), \qquad \text{i.e.} \qquad \langle e_\Omega, h\rangle_{L^2(\mu_{\mathrm{down}})} = 0 \;\; \forall h \in \mathcal{H}_\Omega.$$

We will use the decomposition

$$f_\star = \Pi_\Omega f_\star + e_\Omega. \tag{F.2}$$

F.3. Exact risk decomposition conditional on $(D_{\mathrm{pre}}^{(m)}, X_{1:n})$

We now restate and prove Proposition B.1.

Proposition B.1 (Exact conditional risk decomposition) *For the minimum-norm OLS predictor $\hat{f}_{\Omega,n}$, the conditional test risk admits the decomposition*

$$\begin{aligned}\mathbb{E}\Big[\big(Y_{\mathrm{new}} - \hat{f}_{\Omega,n}(X_{\mathrm{new}})\big)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}\Big] &= \sigma^2 + \mathrm{Rep}(\Omega) \\ &+ \underbrace{\mathbb{E}\Big[\big(\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star\big)(X_{\mathrm{new}})^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}\Big]}_{=:\,\mathrm{Leakage}_n(\Omega)} + \underbrace{\frac{\sigma^2}{n}\,\mathrm{tr}\big(\Sigma(\Omega)\Sigma_n(\Omega)^+\big)}_{=:\,\mathrm{Var}_n(\Omega)}.\end{aligned} \tag{B.3}$$

Proof Start from the identity $Y_{\text{new}} = f_\star(X_{\text{new}}) + \varepsilon_{\text{new}}$ and write

$$Y_{\text{new}} - \hat{f}_{\Omega,n}(X_{\text{new}}) = \varepsilon_{\text{new}} + \big(f_\star - \hat{f}_{\Omega,n}\big)(X_{\text{new}}).$$

Using (F.1), we have

$$\hat{f}_{\Omega,n} = \Pi_{\Omega,n} f_\star + \Pi_{\Omega,n}\varepsilon.$$

Substituting and then adding and subtracting $\Pi_\Omega f_\star$ yields

$$\begin{aligned} Y_{\text{new}} - \hat{f}_{\Omega,n}(X_{\text{new}}) &= \varepsilon_{\text{new}} + \big(f_\star - \Pi_{\Omega,n} f_\star - \Pi_{\Omega,n}\varepsilon\big)(X_{\text{new}}) \\ &= \varepsilon_{\text{new}} + \big((f_\star - \Pi_\Omega f_\star) - (\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star) - \Pi_{\Omega,n}\varepsilon\big)(X_{\text{new}}) \\ &= \varepsilon_{\text{new}} + \big(e_\Omega - (\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star) - \Pi_{\Omega,n}\varepsilon\big)(X_{\text{new}}), \end{aligned} \tag{F.3}$$

where in the last step we used (F.2).

Square (F.3) and take conditional expectation given $(D_{\text{pre}}^{(m)}, X_{1:n})$:

$$\begin{aligned} &\mathbb{E}\Big[\big(Y_{\text{new}} - \hat{f}_{\Omega,n}(X_{\text{new}})\big)^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] \\ &= \mathbb{E}\Big[\varepsilon_{\text{new}}^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] \\ &\quad + \mathbb{E}\Big[\Big(e_\Omega - (\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star) - \Pi_{\Omega,n}\varepsilon\Big)(X_{\text{new}})^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] \\ &\quad + 2\,\mathbb{E}\Big[\varepsilon_{\text{new}}\Big(e_\Omega - (\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star) - \Pi_{\Omega,n}\varepsilon\Big)(X_{\text{new}}) \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big]. \end{aligned} \tag{F.4}$$

Cross term with ε_{new}. Condition on $(D_{\text{pre}}^{(m)}, X_{1:n}, X_{\text{new}}, \varepsilon_{1:n})$. The bracketed term in (F.4) evaluated at X_{new} is measurable with respect to $(D_{\text{pre}}^{(m)}, X_{1:n}, X_{\text{new}}, \varepsilon_{1:n})$, while $\mathbb{E}[\varepsilon_{\text{new}} \mid X_{\text{new}}] = 0$. Therefore, we have

$$\begin{aligned} &\mathbb{E}\Big[\varepsilon_{\text{new}}\Big(e_\Omega - (\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star) - \Pi_{\Omega,n}\varepsilon\Big)(X_{\text{new}}) \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] \\ &= \mathbb{E}\left[\mathbb{E}\Big[\varepsilon_{\text{new}}\Big(e_\Omega - (\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star) - \Pi_{\Omega,n}\varepsilon\Big)(X_{\text{new}}) \,\Big|\, X_{\text{new}}, \varepsilon_{1:n}\Big] \,\middle|\, D_{\text{pre}}^{(m)}, X_{1:n}\right] \\ &= \mathbb{E}\left[\mathbb{E}[\varepsilon_{\text{new}} \mid X_{\text{new}}, \varepsilon_{1:n}]\,\Big(e_\Omega - (\Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star) - \Pi_{\Omega,n}\varepsilon\Big)(X_{\text{new}}) \,\middle|\, D_{\text{pre}}^{(m)}, X_{1:n}\right] = 0. \end{aligned}$$

Since $(X_{\text{new}}, \varepsilon_{\text{new}})$ is an independent copy of (X, ε) and is independent of $(D_{\text{pre}}^{(m)}, X_{1:n}, \varepsilon_{1:n})$,

$$\mathbb{E}\Big[\varepsilon_{\text{new}}^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] = \mathbb{E}[\varepsilon^2] = \sigma^2.$$

Set

$$g \coloneqq \Pi_{\Omega,n} f_\star - \Pi_\Omega f_\star \in \mathcal{H}_\Omega, \qquad u \coloneqq \Pi_{\Omega,n}\varepsilon \in \mathcal{H}_\Omega.$$

Then pointwise,

$$(e_\Omega - g - u)^2 = e_\Omega^2 + g^2 + u^2 - 2e_\Omega g - 2e_\Omega u + 2gu.$$

Evaluating at X_{new} and taking $\mathbb{E}[\cdot \mid D_{\text{pre}}^{(m)}, X_{1:n}]$ gives

$$\begin{aligned}\mathbb{E}\Big[(e_\Omega - g - u)(X_{\text{new}})^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] &= \mathbb{E}\Big[e_\Omega(X_{\text{new}})^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] + \mathbb{E}\Big[g(X_{\text{new}})^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] \\ &\quad - 2\,\langle e_\Omega, g\rangle_{L^2(\mu_{\text{down}})} - 2\,\langle e_\Omega, u\rangle_{L^2(\mu_{\text{down}})} \\ &\quad + 2\,\mathbb{E}\Big[g(X_{\text{new}})u(X_{\text{new}}) \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] \\ &\quad + \mathbb{E}\Big[u(X_{\text{new}})^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big]. \end{aligned} \tag{F.5}$$

Orthogonality kills the e_Ω cross terms. Since $g, u \in \mathcal{H}_\Omega$ and $e_\Omega \perp \mathcal{H}_\Omega$ in $L^2(\mu_{\text{down}})$, we have

$$\langle e_\Omega, g\rangle_{L^2(\mu_{\text{down}})} = \langle e_\Omega, u\rangle_{L^2(\mu_{\text{down}})} = 0.$$

The remaining cross term averages to zero. Condition on $(D_{\text{pre}}^{(m)}, X_{1:n}, X_{\text{new}})$. Given $(D_{\text{pre}}^{(m)}, X_{1:n})$, the function g is deterministic (because Ω is $D_{\text{pre}}^{(m)}$-measurable and $\Pi_{\Omega,n}$ depends only on $(\Omega, X_{1:n})$), while $u = \Pi_{\Omega,n}\varepsilon$ is linear in the noise vector $\varepsilon_{1:n}$. Using $\mathbb{E}[\varepsilon_{1:n} \mid X_{1:n}] = 0$ and independence of $D_{\text{pre}}^{(m)}$ from the downstream noise, we obtain

$$\mathbb{E}\Big[u(X_{\text{new}}) \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}, X_{\text{new}}\Big] = 0,$$

and hence

$$\mathbb{E}\Big[g(X_{\text{new}})u(X_{\text{new}}) \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}, X_{\text{new}}\Big] = 0.$$

Therefore,

$$\mathbb{E}\Big[g(X_{\text{new}})u(X_{\text{new}}) \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] = 0.$$

Conclusion. Using $\mathbb{E}[e_\Omega(X_{\text{new}})^2] = \|e_\Omega\|^2_{L^2(\mu_{\text{down}})} = \text{Rep}(\Omega)$ in (F.5) and substituting into (F.4) yields (B.3). ∎

F.4. Well-posedness specialization

Corollary F.1 (Well-posedness implies $\Pi_{\Omega,n}f_\star - \Pi_\Omega f_\star = \Pi_{\Omega,n}f_\Omega$) *Assume $\langle\cdot,\cdot\rangle_n$ is non-degenerate on $\mathcal{H}_\Omega$ (equivalently, Ev_n is injective on $\mathcal{H}_\Omega$). Then $\Pi_{\Omega,n}h = h$ for all $h \in \mathcal{H}_\Omega$, and therefore*

$$\Pi_{\Omega,n}f_\star - \Pi_\Omega f_\star = \Pi_{\Omega,n}e_\Omega.$$

Consequently, (B.3) *reduces to*

$$\begin{aligned}\mathbb{E}\Big[\big(Y_{\text{new}} - \hat{f}_{\Omega,n}(X_{\text{new}})\big)^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] &= \mathbb{E}[\varepsilon^2] + \text{Rep}(\Omega) + \mathbb{E}\Big[(\Pi_{\Omega,n}e_\Omega)(X_{\text{new}})^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big] \\ &\quad + \mathbb{E}\Big[(\Pi_{\Omega,n}\varepsilon)(X_{\text{new}})^2 \,\Big|\, D_{\text{pre}}^{(m)}, X_{1:n}\Big].\end{aligned}$$

Proof Under the stated condition, $\Pi_{\Omega,n}h = h$ holds for all $h \in \mathcal{H}_\Omega$. Since $\Pi_\Omega f_\star \in \mathcal{H}_\Omega$ and $f_\star = \Pi_\Omega f_\star + e_\Omega$, we have

$$\Pi_{\Omega,n}f_\star = \Pi_{\Omega,n}(\Pi_\Omega f_\star + e_\Omega) = \Pi_{\Omega,n}(\Pi_\Omega f_\star) + \Pi_{\Omega,n}e_\Omega = \Pi_\Omega f_\star + \Pi_{\Omega,n}e_\Omega,$$

which implies $\Pi_{\Omega,n}f_\star - \Pi_\Omega f_\star = \Pi_{\Omega,n}e_\Omega$. Substituting this identity into (B.3) gives the claimed formula. ∎

Bridge to the Riemannian log-map CLT. In the main text we set $\Omega = \hat{\Omega}_m(D_{\mathrm{pre}}^{(m)}) \in \mathcal{M}$. The conditional risk is

$$R(D_{\mathrm{pre}}^{(m)}, X_{1:n}) = \mathbb{E}\Big[\big(Y_{\mathrm{new}} - \hat{f}_{\hat{\Omega}_m,n}(X_{\mathrm{new}})\big)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}\Big],$$

which is obtained by substituting $\Omega = \hat{\Omega}_m$ into (B.3). The Riemannian structure enters when analyzing the fluctuations of $\hat{\Omega}_m$ around $\Omega_\star$ through the log map: under the assumptions stated in the main text,

$$\sqrt{m}\,\log_{\Omega_\star}(\hat{\Omega}_m) \overset{d}{\rightsquigarrow} Z, \qquad Z \sim \mathcal{N}(0, V) \text{ in } T_{\Omega_\star}\mathcal{M}.$$

Since the decomposition (B.3) holds conditionally on $(D_{\mathrm{pre}}^{(m)}, X_{1:n})$ for each m, n, one can combine this log-map CLT with a delta-method argument for the map $\Omega \mapsto R(D_{\mathrm{pre}}^{(m)}, X_{1:n})$ in a normal neighborhood of $\Omega_\star$, without taking expectations over $D_{\mathrm{pre}}^{(m)}$.

Appendix G. Proof of Theorem 5.1

This appendix proves a more general version of Theorem 5.1 by analyzing the three terms in the exact conditional risk decomposition in Proposition B.1. The proof separates the contribution of the pre-training randomness from the two downstream estimation effects. The representation term captures how the random pre-trained feature parameter $\hat{\Omega}_m(D_{\mathrm{pre}}^{(m)})$ changes the population approximation error. The variance and leakage terms capture, respectively, the usual downstream noise-fitting effect and the additional finite-sample interaction between the downstream empirical projection and the pre-trained residual.

The statement in the main text corresponds to the regular case $\mathcal{B}_0 = \{0\}$. In the more general stable-null-span regime treated below, directions that are null at $\Omega_\star$ may have a limiting activated span $\mathcal{B}_0$, and the downstream OLS degrees-of-freedom term is then

$$\sigma^2 d_{\mathrm{act}}, \qquad d_{\mathrm{act}} := \dim(\mathcal{H}_{\Omega_\star} \oplus \mathcal{B}_0).$$

Specializing to $\mathcal{B}_0 = \{0\}$ gives $d_{\mathrm{act}} = d_{\mathrm{eff}}(\Omega_\star)$, recovering Theorem 5.1. We analyze all terms along coupled sequences $m, n \to \infty$ with $m/n \to \alpha$. Throughout, we work on a coupled probability space supporting three mutually independent objects:

(i) An i.i.d. pre-training sequence $(Z_j^{\mathrm{pre}})_{j \geqslant 1}$,

(ii) An i.i.d. downstream sequence $(X_i, \varepsilon_i)_{i \geqslant 1}$ with $X_i \sim \mu_{\mathrm{down}}$ and $\mathbb{E}[\varepsilon_i \mid X_i] = 0$,

(iii) An independent test covariate $X_{\mathrm{new}} \sim \mu_{\mathrm{down}}$ (and an independent noise $\varepsilon_{\mathrm{new}}$)

For each m, define the pre-training dataset $D_{\mathrm{pre}}^{(m)} := (Z_1^{\mathrm{pre}}, \ldots, Z_m^{\mathrm{pre}})$ and the feature parameter $\Omega_m := \hat{\Omega}_m(D_{\mathrm{pre}}^{(m)})$.

Conditioning convention. All conditional expectations in this appendix are taken *given* $D_{\mathrm{pre}}^{(m)}$ (and, when appropriate, given $X_{1:n}$ as well). Once we condition on $D_{\mathrm{pre}}^{(m)}$, Ω_m is fixed, and the downstream sample $(X_i, \varepsilon_i)_{i=1}^n$ remains i.i.d. and independent of $D_{\mathrm{pre}}^{(m)}$.

We first state the regularity assumptions in Appendix G.1. We then analyze the representation term in Appendix G.2, followed by the downstream variance and leakage terms in Appendix G.3.

G.1. Setup and standing regularity

For each n, define the population and empirical covariances

$$\Sigma_m \coloneqq \Sigma(\Omega_m) \coloneqq \mathbb{E}\Big[\phi(X,\Omega_m)\phi(X,\Omega_m)^\top \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big],$$
$$\Sigma_{m,n} \coloneqq \Sigma_n(\Omega_m) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\phi(X_i,\Omega_m)\phi(X_i,\Omega_m)^\top,$$

where the expectation in Σ_m is over $X \sim \mu_{\mathrm{down}}$.

Assumption G.1 (Well-posedness for identifying the leakage term) *With probability tending to 1 as $m \to \infty$ (under the joint law of $(D_{\mathrm{pre}}^{(m)}, X_{1:n})$), the empirical inner product $\langle\cdot,\cdot\rangle_n$ is non-degenerate on $\mathcal{H}_{\Omega_m}$ (equivalently, the evaluation map on $\{X_i\}_{i=1}^n$ is injective on $\mathcal{H}_{\Omega_m}$). On this event, $\Pi_{\Omega_m,n}h = h$ for all $h \in \mathcal{H}_{\Omega_m}$, and hence (G.7) holds (Appendix F.4).*

Assumption G.2 (Local-uniform moment and leverage bounds near $\Omega_\star$) *There exist $\delta > 0$, $\eta > 0$, a neighborhood $\mathcal{U}$ of $\Omega_\star$ in $\mathcal{M}$, and constants $C_\phi, C_q, C_e < \infty$ such that*

$$\sup_{\Omega\in\mathcal{U}} \mathbb{E}\Big[\|\phi(X,\Omega)\|^{4+\delta}\Big] \leqslant C_\phi.$$

For each $\Omega \in \mathcal{U}$, define the population leverage score

$$q_\Omega(X) \coloneqq \phi(X,\Omega)^\top \Sigma(\Omega)^+ \phi(X,\Omega). \tag{G.1}$$

Assume the local leverage moment bound

$$\sup_{\Omega\in\mathcal{U}} \mathbb{E}\big[q_\Omega(X)^{2+\eta}\big] \leqslant C_q.$$

Moreover, for the signal term, assume the leverage-weighted signal bound

$$\sup_{\Omega\in\mathcal{U}} \mathbb{E}\big[e_\Omega(X)^2 q_\Omega(X)^{1+\eta}\big] \leqslant C_e.$$

Assumption G.3 (Local C^1 regularity of the feature map in Ω with moment control) *Work in normal coordinates on a neighborhood $\mathcal{U}$ of $\Omega_\star$. Assume that for μ_{down}-a.e. x, the map $\Omega \mapsto \phi(x,\Omega) \in \mathbb{R}^p$ is continuously differentiable on $\mathcal{U}$. Moreover, there exist $\delta > 0$ and a constant $C_{\partial\phi} < \infty$ such that*

$$\sup_{\Omega\in\mathcal{U}} \mathbb{E}\Big[\|D_\Omega\phi(X,\Omega)\|_{\mathrm{op}}^{4+\delta}\Big] \leqslant C_{\partial\phi}, \tag{G.2}$$

where $D_\Omega\phi(x,\Omega) : T_\Omega\mathcal{M} \to \mathbb{R}^p$ denotes the derivative in normal coordinates.

The preceding assumptions control the empirical downstream quantities and the local smoothness of the feature map. We now introduce the geometric notation needed to analyze the representation term. Since $f_\star \in \mathcal{H}_{\Omega_\star}$, choose $\theta_\star \in \mathbb{R}^p$ such that $f_\star = T_{\Omega_\star}\theta_\star$. The map $T_{\Omega_\star}$ need not be injective. Let

$$N_\star \coloneqq \ker(T_{\Omega_\star}), \qquad E_\star \coloneqq N_\star^\perp.$$

Thus $\mathbb{R}^p = N_\star \oplus E_\star$: directions in $E_\star$ are identifiable at $\Omega_\star$, whereas directions in $N_\star$ are invisible at $\Omega_\star$.

For a nearby descriptor Ω, one could similarly define $N_\Omega = \ker(T_\Omega)$ and $E_\Omega = N_\Omega^\perp$. If T_Ω has locally constant rank, these subspaces vary smoothly with Ω, and the image spaces $\mathcal{H}_\Omega = \mathrm{Im}(T_\Omega)$ vary smoothly as well. This is the standard regime in which the full projector map $\Omega \mapsto \Pi_\Omega$ is Fréchet differentiable. In this appendix, however, we do not impose local constant rank. Instead of tracking the moving decomposition $N_\Omega \oplus E_\Omega$, we fix the splitting at $\Omega_\star$ and use it to describe how the perturbed image space $\mathcal{H}_\Omega$ is generated.

For $v \in T_{\Omega_\star}\mathcal{M}$ near 0, set

$$\Omega(v) \coloneqq \exp_{\Omega_\star}(v), \qquad T_v \coloneqq T_{\Omega(v)}, \qquad \mathcal{H}_v \coloneqq \mathcal{H}_{\Omega(v)}, \qquad \Pi_v \coloneqq \Pi_{\Omega(v)}.$$

The identifiable directions $E_\star$ generate the subspace $\mathcal{A}_v \coloneqq \mathrm{Im}\left(T_v\big|_{E_\star}\right)$. Since $T_{\Omega_\star}\big|_{E_\star}$ is injective, this is the stable part of the perturbed image: under the local smoothness assumptions, $\mathcal{A}_v$ converges to $\mathcal{H}_{\Omega_\star}$ as $v \to 0$.

The remaining issue is the contribution of the directions that were null at $\Omega_\star$. Because $\mathbb{R}^p = E_\star \oplus N_\star$, the full perturbed image $\mathcal{H}_v$ is generated by $T_v E_\star$ together with $T_v N_\star$. The subspace generated by $T_v E_\star$ is $\mathcal{A}_v$. We therefore isolate the orthogonal residual contribution of the perturbed null directions by defining

$$\mathcal{B}_v \coloneqq \mathrm{Im}\left((I - \Pi_{\mathcal{A}_v})T_v\big|_{N_\star}\right), \qquad v \neq 0.$$

By construction, $\mathcal{B}_v \subseteq \mathcal{A}_v^\perp$, and the image space decomposes as $\mathcal{H}_v = \mathcal{A}_v \oplus \mathcal{B}_v$. Thus $\mathcal{A}_v$ records the part of $\mathcal{H}_v$ generated by directions already identifiable at $\Omega_\star$, while $\mathcal{B}_v$ records the additional directions produced by perturbing directions that were null at $\Omega_\star$. The stable null-span assumption below requires only this second component to have a limiting span. It is therefore weaker than requiring smooth variation of the full image space, or equivalently differentiability of the full projector map $\Omega \mapsto \Pi_\Omega$.

Assumption G.4 (Stable limiting span of null directions) *There exists a finite-dimensional subspace $\mathcal{B}_0 \subseteq \mathcal{H}_{\Omega_\star}^\perp$ such that*

$$\|\Pi_{\mathcal{B}_v} - \Pi_{\mathcal{B}_0}\|_{\mathrm{op}} \to 0 \qquad \textit{as } v \to 0,$$

where $\Pi_{\mathcal{B}_v}$ and $\Pi_{\mathcal{B}_0}$ denote the $L^2(\mu_{\mathrm{down}})$-orthogonal projectors onto $\mathcal{B}_v$ and $\mathcal{B}_0$, respectively.

Lemma G.5 (Differentiability of the feature operator) *Under Assumption G.3, the map $\Omega \mapsto T_\Omega$ is Fréchet differentiable at $\Omega_\star$ as a map from $\mathcal{M}$ to $\mathcal{B}(\mathbb{R}^p, L^2)$, the space of bounded linear operators form $\mathbb{R}^p$ to L^2. In normal coordinates around $\Omega_\star$, its derivative is given by*

$$(DT_{\Omega_\star}[v]\theta)(x) = \theta^\top D_\Omega \phi(x, \Omega_\star)[v], \qquad v \in T_{\Omega_\star}\mathcal{M}.$$

Proof Fix $v \in T_{\Omega_\star}\mathcal{M}$ sufficiently small and define

$$\rho_v(x) \coloneqq \frac{\phi(x, \Omega(v)) - \phi(x, \Omega_\star) - D_\Omega \phi(x, \Omega_\star)[v]}{\|v\|}.$$

By Assumption G.3, for μ_{down}-a.e. x,

$$\rho_v(x) \to 0 \qquad \text{as } v \to 0.$$

We first show that $\rho_v \to 0$ in $L^2(\mu_{\text{down}}; \mathbb{R}^p)$. By the fundamental theorem of calculus in normal coordinates,

$$\rho_v(x) = \int_0^1 \left(D_\Omega \phi(x, \Omega(tv)) - D_\Omega \phi(x, \Omega_\star)\right) \left[\frac{v}{\|v\|}\right] dt.$$

Therefore,

$$\begin{aligned}
\|\rho_v(x)\| &= \left\| \int_0^1 \left(D_\Omega \phi(x, \Omega(tv)) - D_\Omega \phi(x, \Omega_\star)\right) \left[\frac{v}{\|v\|}\right] dt \right\| \\
&\leqslant \int_0^1 \left\| \left(D_\Omega \phi(x, \Omega(tv)) - D_\Omega \phi(x, \Omega_\star)\right) \left[\frac{v}{\|v\|}\right] \right\| dt \\
&\leqslant \int_0^1 \|D_\Omega \phi(x, \Omega(tv)) - D_\Omega \phi(x, \Omega_\star)\|_{\text{op}} \, dt.
\end{aligned}$$

Let $q = 4 + \delta$, where $\delta > 0$ is from Assumption G.3. Jensen's inequality and the elementary bound $\|A - B\|^q \leqslant 2^{q-1}(\|A\|^q + \|B\|^q)$ give

$$\mathbb{E}\|\rho_v(X)\|^q \leqslant 2^{q-1} \int_0^1 \mathbb{E}\|D_\Omega \phi(X, \Omega(tv))\|_{\text{op}}^q \, dt + 2^{q-1} \mathbb{E}\|D_\Omega \phi(X, \Omega_\star)\|_{\text{op}}^q.$$

The right-hand side is uniformly bounded for v sufficiently small by Assumption G.3. Hence the family $\{\|\rho_v(X)\|^2\}$ is uniformly integrable. Since $\rho_v(X) \to 0$ almost surely, Vitali's theorem yields

$$\mathbb{E}\|\rho_v(X)\|^2 \to 0.$$

Thus,

$$\|\phi(\cdot, \Omega(v)) - \phi(\cdot, \Omega_\star) - D_\Omega \phi(\cdot, \Omega_\star)[v]\|_{L^2(\mu_{\text{down}}; \mathbb{R}^p)} = o(\|v\|).$$

Now let $\theta \in \mathbb{R}^p$ with $\|\theta\|_2 = 1$. Then

$$\begin{aligned}
&\left\| \left(T_{\Omega(v)} - T_{\Omega_\star} - DT_{\Omega_\star}[v]\right) \theta \right\|_{L^2} \\
&\leqslant \|\phi(\cdot, \Omega(v)) - \phi(\cdot, \Omega_\star) - D_\Omega \phi(\cdot, \Omega_\star)[v]\|_{L^2(\mu_{\text{down}}; \mathbb{R}^p)} = o(\|v\|).
\end{aligned}$$

Taking the supremum over $\|\theta\|_2 = 1$ gives

$$\|T_{\Omega(v)} - T_{\Omega_\star} - DT_{\Omega_\star}[v]\|_{\text{op}} = o(\|v\|).$$

Thus $\Omega \mapsto T_\Omega$ is Fréchet differentiable at $\Omega_\star$, with derivative

$$(DT_{\Omega_\star}[v]\theta)(x) = \theta^\top D_\Omega \phi(x, \Omega_\star)[v].$$

■

G.2. Pretraining fluctuations

This subsection analyzes the representation term (B.2) in the compatible regime, where $\mathrm{Rep}(\Omega_\star) = 0$. The key step is to linearize the residual $e_\Omega = (I - \Pi_\Omega) f_\star$ around $\Omega_\star$. Under the stable-null-span condition (Assumption G.4), this residual admits a first-order expansion along perturbations of the pre-trained descriptor, and the $1/m$-scale limit of $\mathrm{Rep}(\Omega_m)$ follows by applying the pre-training CLT.

This approach is weaker than requiring Fréchet differentiability of the full projector map $\Omega \mapsto \Pi_\Omega$ as an operator on $L^2(\mu_{\mathrm{down}})$. We only need first-order control of the single target $f_\star$, or equivalently of the residual e_Ω. After proving the representation limit under the stable null-span condition (Assumption G.4), we record constant-rank differentiability of Π_Ω as a simpler sufficient condition used in examples.

G.2.1. MAIN STATEMENTS

Proposition G.6 (Representation term: distributional limit in the compatible case) *Assume the pretraining distribution and loss function satisfy Assumption D.5*

$$Z_m := \sqrt{m}\, \log_{\Omega_\star}(\Omega_m) \overset{d}{\rightsquigarrow} Z, \qquad Z \sim \mathcal{N}(0, V),$$

in $T_{\Omega_\star}\mathcal{M}$. Moreover, assume Assumptions G.2–G.4. Let $\mathcal{L}$ be the first-order residual map defined as

$$\mathcal{L}(v) := -(I - \Pi_{\Omega_\star} - \Pi_{\mathcal{B}_0}) DT_{\Omega_\star}[v]\theta_\star. \tag{G.3}$$

Then

$$m\, \mathrm{Rep}(\Omega_m) \overset{d}{\rightsquigarrow} \|\mathcal{L}(Z)\|^2_{L^2(\mu_{\mathrm{down}})}.$$

G.2.2. LINEARIZATION IN NORMAL COORDINATES

Lemma G.7 (Residual expansion under stable limiting span of null directions) *Assume Assumptions G.3 and G.4, and recall the linear map $\mathcal{L} : T_{\Omega_\star}\mathcal{M} \to L^2$ (G.3). Then, in normal coordinates around $\Omega_\star$,*

$$e_{\Omega(v)} = \mathcal{L}(v) + o(\|v\|)$$

in $L^2(\mu_{\mathrm{down}})$.

Proof Since $f_\star = T_{\Omega_\star}\theta_\star$, we have

$$f_\star = T_v\theta_\star + (T_{\Omega_\star} - T_v)\theta_\star.$$

We apply Π_v to both sides and use $T_v\theta_\star \in \mathcal{H}_{\Omega(v)}$ to get

$$\Pi_v f_\star = \Pi_v T_v\theta_\star + \Pi_v(T_{\Omega_\star} - T_v)\theta_\star = T_v\theta_\star + \Pi_v(T_{\Omega_\star} - T_v)\theta_\star.$$

Therefore

$$e_{\Omega(v)} = f_\star - \Pi_v f_\star = -(I - \Pi_v)(T_v - T_{\Omega_\star})\theta_\star.$$

By Lemma G.5, we have

$$T_v = T_{\Omega_\star} + DT_{\Omega_\star}[v] + r_T(v)$$

where $\|r_T(v)\|_{\mathrm{op}} = o(\|v\|)$. Since $I - \Pi_{\Omega(v)}$ is an orthogonal projector, and hence it is non-expansive, we have

$$e_{\Omega(v)} = -(I - \Pi_v)(DT_{\Omega_\star}[v] + r_T(v))\theta_\star = -(I - \Pi_v)(DT_{\Omega_\star}[v])\theta_\star + o(\|v\|).$$

We now identify the limit of $\Pi_{\Omega(v)}$. Since $\mathbb{R}^p = E_\star \oplus N_\star$, every element of $\mathcal{H}_{\Omega(v)} = \mathrm{Im}(T_v)$ can be written as

$$T_v\theta_E + T_v\theta_N$$

where $\theta_E \in E_\star$ and $\theta_N \in N_\star$. The first term belongs to $\mathcal{A}_v = \mathrm{Im}\left(T_v\big|_{E_\star}\right)$. For the second term,

$$T_v\theta_N = \Pi_{\mathcal{A}_v} T_v\theta_N + (I - \Pi_{\mathcal{A}_v})T_v\theta_N$$

where the first summand lies in $\mathcal{A}_v$ and the second lies in $\mathcal{B}_v = \mathrm{Im}((I - \Pi_{\mathcal{A}_v})T_v\big|_{N_\star})$. Hence, $\mathcal{H}_v = \mathcal{A}_v + \mathcal{B}_v$. Moreover, by construction, $\mathcal{B}_v \subseteq \mathcal{A}_v^\perp$, so we have

$$\mathcal{H}_v = \mathcal{A}_v \oplus \mathcal{B}_v.$$

Consequently, we have

$$\Pi_v = \Pi_{\mathcal{A}_v} + \Pi_{\mathcal{B}_v}.$$

It remains to identify the limiting behavior of the two projections. We first consider $\mathcal{A}_v$. The restriction $T_{\Omega_\star}\big|_{E_\star}$ is injective, because $E_\star = N_\star^\perp$ and $N_\star = \ker(T_{\Omega_\star})$. Since $E_\star$ is finite-dimensional, this injectivity is stable under small operator-norm perturbations. In particular, by Lemma G.5,

$$T_v\big|_{E_\star} \to T_{\Omega_\star}\big|_{E_\star}$$

in operator norm, and hence $T_v\big|_{E_\star}$ remains injective for all sufficiently small v. Thus $\mathcal{A}_v$ has the same dimension as $\mathcal{H}_\star$ for small v. Consequently, $\mathcal{A}_v$ converge to $\mathcal{H}_\star$ in the Grassmannian topology, or equivalently,

$$\|\Pi_{\mathcal{A}_v} - \Pi_{\Omega_\star}\|_{\mathrm{op}} \to 0.$$

The second projection is controlled directly by Assumption G.4, which gives

$$\|\Pi_{\mathcal{B}_v} - \Pi_{\mathcal{B}_0}\|_{\mathrm{op}} \to 0.$$

Since $\mathcal{H}_v = \mathcal{A}_v \oplus \mathcal{B}_v$, by orthogonality, we have

$$\Pi_{\Omega(v)} = \Pi_{\mathcal{A}_v} + \Pi_{\mathcal{B}_v}.$$

Combining the two projection limits yields

$$\Pi_{\Omega(v)} \to \Pi_{\Omega_\star} + \Pi_{\mathcal{B}_0}$$

in operator norm.

For small enough v, Assumption G.3 yields

$$\|DT_{\Omega_\star}[v]\theta_\star\|_{L^2}^2 = \mathbb{E}_X\left(\theta_\star^\top D_\Omega\phi(X, \Omega_\star)[v]\right)^2 \leqslant \|D_{\Omega_\star}\phi\|_{\mathrm{op}}^2 \|\theta_\star\|_2^2 \|v\|_2^2 = O(\|v\|_2^2).$$

Therefore,

$$(I - \Pi_{\Omega(v)})DT_{\Omega_\star}[v]\theta_\star = (I - \Pi_{\Omega_\star} - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\theta_\star + o(\|v\|).$$

Combining this with the earlier expansion for $e_{\Omega(v)}$ yields

$$e_{\Omega(v)} = -(I - \Pi_{\Omega_\star} - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\theta_\star + o(\|v\|) = \mathcal{L}(v) + o(\|v\|).$$

■

Lemma G.8 (Well-definedness of $\mathcal{L}$) *Assume Assumptions G.2, G.3, and G.4. The map*

$$\mathcal{L}(v) \coloneqq -(I - \Pi_{\Omega_\star} - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\theta_\star$$

does not depend on the choice of $\theta_\star \in \mathbb{R}^p$ satisfying $f_\star = T_{\Omega_\star}\theta_\star$.

Proof Let $\theta'_\star \in \mathbb{R}^p$ be another coefficient vector such that $f_\star = T_{\Omega_\star}\theta'_\star$. Then

$$\eta \coloneqq \theta'_\star - \theta_\star \in \ker(T_{\Omega_\star}) = N_\star.$$

It suffices to show that, for every $\eta \in N_\star$,

$$(I - \Pi_{\Omega_\star} - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\eta = 0.$$

Fix $\eta \in N_\star$. Since $T_{\Omega_\star}\eta = 0$, Lemma G.5 gives

$$T_v\eta = T_{\Omega_\star}\eta + DT_{\Omega_\star}[v]\eta + r_T(v)\eta = DT_{\Omega_\star}[v]\eta + r_T(v)\eta, \qquad \|r_T(v)\eta\|_{L^2} = o(\|v\|_2).$$

On the other hand, by the orthogonal decomposition $\mathcal{H}_v = \mathcal{A}_v \oplus \mathcal{B}_v$, the projector onto $\mathcal{H}_v$ is

$$\Pi_v = \Pi_{\mathcal{A}_v} + \Pi_{\mathcal{B}_v}.$$

Since $T_v\eta \in \mathcal{H}_v$, we have

$$(I - \Pi_{\mathcal{A}_v} - \Pi_{\mathcal{B}_v})T_v\eta = 0.$$

Substituting the expansion of $T_v\eta$ yields

$$(I - \Pi_{\mathcal{A}_v} - \Pi_{\mathcal{B}_v})DT_{\Omega_\star}[v]\eta = -(I - \Pi_{\mathcal{A}_v} - \Pi_{\mathcal{B}_v})r_T(v)\eta = o(\|v\|_2),$$

because $I - \Pi_{\mathcal{A}_v} - \Pi_{\mathcal{B}_v}$ is an orthogonal projector and therefore has operator norm at most one.

By the convergence

$$\Pi_{\mathcal{A}_v} \to \Pi_{\Omega_\star}, \qquad \Pi_{\mathcal{B}_v} \to \Pi_{\mathcal{B}_0}$$

in operator norm, and since

$$\|DT_{\Omega_\star}[v]\eta\|_{L^2} = O(\|v\|_2),$$

we also have

$$\begin{aligned}&(I - \Pi_{\Omega_\star} - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\eta \\ &\quad = (I - \Pi_{\mathcal{A}_v} - \Pi_{\mathcal{B}_v})DT_{\Omega_\star}[v]\eta + (\Pi_{\mathcal{A}_v} - \Pi_{\Omega_\star} + \Pi_{\mathcal{B}_v} - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\eta = o(\|v\|_2).\end{aligned}$$

The left-hand side is linear in v. A linear map that is $o(\|v\|_2)$ at the origin must be identically zero. Therefore

$$(I - \Pi_{\Omega_\star} - \Pi_{\mathcal{B}_0})DT_{\Omega_\star}[v]\eta = 0$$

for every $v \in T_{\Omega_\star}\mathcal{M}$ and every $\eta \in N_\star$. Thus replacing $\theta_\star$ by $\theta_\star + \eta$ does not change $\mathcal{L}(v)$, and $\mathcal{L}$ is well-defined. ■

Corollary G.9 (Deterministic first-order remainder bound) *Assume Assumptions G.3 and G.4. Then there exists a function $\omega : [0,\infty) \to [0,\infty)$ with $\omega(t) \to 0$ as $t \to 0$ such that for all $v \in T_{\Omega_\star}\mathcal{M}$ in a neighborhood of 0,*

$$\left\| e_{\exp_{\Omega_\star}(v)} - \mathcal{L}(v) \right\|_{L^2(\mu_{\mathrm{down}})} \leqslant \omega(\|v\|)\,\|v\|.$$

Proof By Lemma G.7, $e_{\exp_{\Omega_\star}(v)} = \mathcal{L}(v) + r(v)$ where $r(v) = o(\|v\|)$. Define

$$\omega(t) := \sup_{0<\|v\|\leqslant t} \frac{\|r(v)\|_{L^2(\mu_{\mathrm{down}})}}{\|v\|}.$$

Then $\omega(t) \to 0$ as $t \to 0$, and the desired bound follows. ∎

G.2.3. DISTRIBUTIONAL LIMIT OF THE REPRESENTATION TERM

By Corollary G.9,

$$e_{\Omega_m} = e_{\exp_{\Omega_\star}(v_m)} = \mathcal{L}(v_m) + r(v_m), \qquad \|r(v_m)\|_{L^2(\mu_{\mathrm{down}})} \leqslant \omega(\|v_m\|)\,\|v_m\|.$$

Multiplying by $\sqrt{m}$ yields

$$\sqrt{m}\, e_{\Omega_m} = \mathcal{L}(Z_m) + \sqrt{m}\, r(v_m), \qquad \|\sqrt{m}\, r(v_m)\|_{L^2(\mu_{\mathrm{down}})} \leqslant \omega(\|v_m\|)\,\|Z_m\|.$$

Lemma G.10 (The linearization remainder is negligible in probability) *Assume Assumptions D.5, G.3, and G.4. Then*

$$\|\sqrt{m}\, r(v_m)\|_{L^2(\mu_{\mathrm{down}})} \to 0 \qquad \textit{in probability.}$$

Proof Since $\Omega_m \to \Omega_\star$ in probability and $v_m = \log_{\Omega_\star}(\Omega_m)$ on a normal neighborhood, we have $\|v_m\| \to 0$ in probability. Moreover, $\|Z_m\|$ is tight by the CLT. Fix $\varepsilon > 0$ and $\delta > 0$. By tightness of $\|Z_m\|$, choose $0 < M < \infty$ such that

$$\limsup_{m\to\infty} \mathbb{P}(\|Z_m\| > M) \leqslant \delta.$$

Since $\omega(t) \to 0$, choose $t_0 > 0$ such that

$$\omega(t) \leqslant \frac{\varepsilon}{M} \qquad \text{for all } 0 \leqslant t \leqslant t_0.$$

On the event

$$\{\|v_m\| \leqslant t_0\} \cap \{\|Z_m\| \leqslant M\},$$

we have

$$\|\sqrt{m}\, r(v_m)\|_{L^2(\mu_{\mathrm{down}})} \leqslant \omega(\|v_m\|)\|Z_m\| \leqslant \varepsilon.$$

Therefore

$$\mathbb{P}\big(\|\sqrt{m}\, r(v_m)\|_{L^2(\mu_{\mathrm{down}})} > \varepsilon\big) \leqslant \mathbb{P}(\|v_m\| > t_0) + \mathbb{P}(\|Z_m\| > M).$$

Taking $\limsup_{m\to\infty}$ gives

$$\limsup_{m\to\infty} \mathbb{P}\big(\|\sqrt{m}\, r(v_m)\|_{L^2(\mu_{\mathrm{down}})} > \varepsilon\big) \leqslant \delta.$$

Since $\delta > 0$ is arbitrary, the claim follows. ∎

Lemma G.11 (Quadratic approximation) *Assume Assumptions D.5, G.3, and G.4. Then*

$$m\,\mathrm{Rep}(\Omega_m) - \|\mathcal{L}(Z_m)\|^2_{L^2(\mu_{\mathrm{down}})} \to 0 \qquad \textit{in probability.}$$

Proof Expanding the square,

$$m\,\mathrm{Rep}(\Omega_m) = \|\sqrt{m}\,e_{\Omega_m}\|^2_{L^2} = \|\mathcal{L}(Z_m)\|^2_{L^2} + 2\langle \mathcal{L}(Z_m), \sqrt{m}\,r(v_m)\rangle_{L^2} + \|\sqrt{m}\,r(v_m)\|^2_{L^2}.$$

Since $\mathcal{L}$ is bounded,

$$\|\mathcal{L}(Z_m)\|_{L^2} \leqslant \|\mathcal{L}\|_{\mathrm{op}}\|Z_m\|.$$

Since $\|Z_m\|$ is tight, $\|\mathcal{L}(Z_m)\|_{L^2}$ is tight. By Lemma G.10, $\|\sqrt{m}\,r(v_m)\|_{L^2} \to 0$ in probability, which implies both the cross term and the squared term vanish in probability. ■

Proof [Proof of Proposition G.6] By Lemma G.11, it suffices to identify the limit law of $\|\mathcal{L}(Z_m)\|^2_{L^2}$. Since $Z_m \overset{d}{\rightsquigarrow} Z$ in $T_{\Omega_\star}\mathcal{M}$ and $z \mapsto \|\mathcal{L}(z)\|^2_{L^2}$ is continuous, the continuous mapping theorem yields

$$\|\mathcal{L}(Z_m)\|^2_{L^2} \overset{d}{\rightsquigarrow} \|\mathcal{L}(Z)\|^2_{L^2}.$$

Combining with Lemma G.11 gives the claim. ■

G.2.4. A CONSTANT-RANK SUFFICIENT CONDITION

We next connect the residual expansion above to the more familiar regular constant-rank setting. Under the stable-null-span condition, the special case $\mathcal{B}_0 = \{0\}$ rules out the activation of directions that are null at $\Omega_\star$. In this case, the downstream feature covariance has locally constant rank and a positive eigengap; consequently, the Moore–Penrose inverse is smooth on the corresponding rank stratum, and the full population projector map $\Omega \mapsto \Pi_\Omega$ is Fréchet differentiable in operator norm. Thus, the abstract residual map in the expansion above agrees with the usual projector derivative:

$$\mathcal{L}(v) = -D\Pi_{\Omega_\star}[v]f_\star = -\big(I - \Pi_{\Omega_\star}\big)DT_{\Omega_\star}[v]\theta_\star.$$

Lemma G.12 *Assume Assumptions G.2, G.3, and G.4. Suppose that $\mathcal{B}_0 = \{0\}$. Then, there exists a neighborhood $\mathcal{U}_0 \subset \mathcal{U}$ of $\Omega_\star$ and a constant $\kappa > 0$ such that, for all $\Omega \in \mathcal{U}_0$,*

$$\mathrm{rank}\big(\Sigma(\Omega)\big) = \mathrm{rank}\big(\Sigma(\Omega_\star)\big) =: r, \qquad \lambda_r(\Sigma(\Omega)) \geqslant \kappa.$$

Proof By Lemma G.5,

$$\|T_v - T_\star\|_{\mathrm{op}} \to 0 \qquad \text{as } v \to 0.$$

First, we show the local constancy of the rank. Since $\mathcal{B}_0 = \{0\}$, we have $\Pi_{\mathcal{B}_0} = 0$. By Assumption G.4, we have

$$\|\Pi_{\mathcal{B}_v}\|_{\mathrm{op}} = \|\Pi_{\mathcal{B}_v} - \Pi_{\mathcal{B}_0}\|_{\mathrm{op}} \to 0.$$

But, an orthogonal projector has operator norm either 0 or 1. Hence, for all sufficiently small non-zero v, we have $\Pi_{\mathcal{B}_v} = 0$, and therefore, we have $\mathcal{B}_v = 0$. For such v,

$$\mathcal{H}_v = \mathcal{A}_v \oplus \mathcal{B}_v = \mathcal{A}_v.$$

Moreover, $T_\star|_{E_\star}$ is injective, and injectivity is stable under small operator-norm perturbations on a finite-dimensional domain. Hence, for all sufficiently small v,

$$T_v|_{E_\star}$$

remains injective. Therefore

$$\operatorname{rank}(T_v) = \dim \mathcal{A}_v = \dim E_\star = \operatorname{rank}(T_{\Omega_\star})$$

for all sufficiently small v. Since

$$\operatorname{rank}(\Sigma(\Omega(v))) = \operatorname{rank}(T_v^* T_v) = \operatorname{rank}(T_v),$$

we obtain

$$\operatorname{rank}(\Sigma(\Omega(v))) = r$$

for all sufficiently small v.

It remains to prove a uniform positive lower bound on the r-th positive eigenvalue. Let $s_r(T_\star) > 0$ denote the smallest nonzero singular value of $T_\star$. Since

$$\|T_v - T_\star\|_{\mathrm{op}} \to 0,$$

the singular-value perturbation bound gives, for sufficiently small v, we have $s_r(T_v) \geqslant \frac{1}{2} s_r(T_\star)$. Because

$$\lambda_r(\Sigma(\Omega(v))) = s_r(T_v)^2,$$

we get

$$\lambda_r(\Sigma(\Omega(v))) \geqslant \frac{1}{4} s_r(T_\star)^2$$

for all sufficiently small v. Taking

$$\kappa \coloneqq \frac{1}{4} s_r(T_\star)^2$$

and shrinking the normal neighborhood if necessary proves the claim. ■

Lemma G.13 (Differentiability of the pseudoinverse on the stable-rank region) *Let $A \succcurlyeq 0$ be symmetric with $\operatorname{rank}(A) = r$ and $\lambda_r(A) \geqslant \kappa > 0$. Then the restriction of the Moore–Penrose map to the stable-rank region*

$$\mathcal{R}_{r,\kappa} \coloneqq \{B \succcurlyeq 0 : \ B = B^\top, \ \operatorname{rank}(B) = r, \ \lambda_r(B) \geqslant \kappa\}$$

is Fréchet differentiable at A (in operator norm). Its derivative in direction H is

$$D(A^+)[H] = -A^+ H A^+ + A^{+2} H (I - A A^+) + (I - A^+ A) H A^{+2}. \tag{G.4}$$

In particular, there exists a constant $C_\kappa < \infty$ depending only on κ such that

$$\|D(A^+)[H]\|_{\mathrm{op}} \leqslant C_\kappa \, \|H\|_{\mathrm{op}}.$$

Proof This is a standard result in perturbation theory. For the proof see Golub and Pereyra (1973). ■

Proposition G.14 (Differentiability of $\Omega \mapsto \Pi_\Omega$ on a constant-rank region) *Fix $\Omega_\star \in \mathcal{M}$. Assume Assumptions G.2–G.4. Assume moreover that $\mathcal{B}_0 = 0$. Then, the map $\Omega \mapsto \Pi_\Omega$ is Fréchet differentiable at $\Omega_\star$ as a map into $\mathcal{B}(L^2(\mu_{\mathrm{down}}))$. In particular, for $v \in T_{\Omega_\star}\mathcal{M}$,*

$$\left\|\Pi_{\exp_{\Omega_\star}(v)} - \Pi_{\Omega_\star} - D\Pi_{\Omega_\star}[v]\right\|_{\mathrm{op}} = o(\|v\|) \qquad \text{as } v \to 0.$$

Proof Write $\Omega(v) := \exp_{\Omega_\star}(v)$. By Lemma G.5, the map $\Omega \mapsto T_\Omega$ is Fréchet differentiable at $\Omega_\star$ as a map into $\mathcal{B}(\mathbb{R}^p, L^2(\mu_{\mathrm{down}}))$. Since taking adjoints is a bounded linear map between operator spaces, it follows that $\Omega \mapsto T_\Omega^{\mathrm{adj}}$ is Fréchet differentiable at $\Omega_\star$ as a map into $\mathcal{B}(L^2(\mu_{\mathrm{down}}), \mathbb{R}^p)$. Next, using the identity $\Sigma(\Omega) = T_\Omega^{\mathrm{adj}} T_\Omega$ and the product rule for bounded operators, the map $\Omega \mapsto \Sigma(\Omega)$ is Fréchet differentiable at $\Omega_\star$ as a map into $\mathbb{R}^{p\times p}$.

By Lemma G.12, the nonzero spectrum of $\Sigma(\Omega)$ is bounded away from zero on $\mathcal{U}$. Therefore, by Lemma G.13, the map $\Omega \mapsto \Sigma(\Omega)^+$ is Fréchet differentiable at $\Omega_\star$. Finally, for every $\Omega \in \mathcal{U}$, the population projection onto $\mathcal{H}_\Omega = \mathrm{Im}(T_\Omega)$ is $\Pi_\Omega = T_\Omega \Sigma(\Omega)^+ T_\Omega^{\mathrm{adj}}$. The right-hand side is a composition and product of Fréchet differentiable maps between Banach spaces. Hence

$$\Omega \mapsto T_\Omega \Sigma(\Omega)^+ T_\Omega^{\mathrm{adj}}$$

is Fréchet differentiable at $\Omega_\star$ as a map into $\mathcal{B}(L^2(\mu_{\mathrm{down}}))$. Therefore $\Omega \mapsto \Pi_\Omega$ is Fréchet differentiable at $\Omega_\star$, which gives

$$\left\|\Pi_{\Omega(v)} - \Pi_{\Omega_\star} - D\Pi_{\Omega_\star}[v]\right\|_{\mathrm{op}} = o(\|v\|).$$

■

G.3. Downstream estimation terms

This subsection studies the two downstream *estimation* terms that appear in the exact conditional risk decomposition (Proposition B.1) where *both* the pre-training and downstream sample sizes diverge with an asymptotic constant rate $\frac{m}{n} \to \alpha \in (0, \infty)$.

G.3.1. ESTIMATION TERMS AND THE COMPATIBLE REGIME

The conditional risk is

$$R(D_{\mathrm{pre}}^{(m)}, X_{1:n}) := \mathbb{E}\left[\left(Y_{\mathrm{new}} - \hat{f}_{\Omega_m,n}(X_{\mathrm{new}})\right)^2 \,\middle|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}\right].$$

Recall from Proposition B.1 that the two downstream estimation terms are

$$\mathrm{Var}_n := \mathbb{E}\left[\left(\Pi_{\Omega_m,n}\varepsilon\right)(X_{\mathrm{new}})^2 \,\middle|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}\right], \tag{G.5}$$

$$\mathrm{Leakage}_n := \mathbb{E}\left[\left(\Pi_{\Omega_m,n} f_\star - \Pi_{\Omega_m} f_\star\right)(X_{\mathrm{new}})^2 \,\middle|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}\right]. \tag{G.6}$$

Here Π_{Ω_m} is the population $L^2(\mu_{\mathrm{down}})$ projector onto $\mathcal{H}_{\Omega_m}$ and $\Pi_{\Omega_m,n}$ is the canonical empirical projector from Appendix C (Definition C.7).

Define the population residual $e_\Omega := (I - \Pi_\Omega) f_\star$, so that $e_\Omega \perp \mathcal{H}_\Omega$ in $L^2(\mu_{\mathrm{down}})$. Under the well-posedness condition in Assumption G.1, one has

$$\Pi_{\Omega_m,n} f_\star - \Pi_{\Omega_m} f_\star = \Pi_{\Omega_m,n} e_{\Omega_m}, \tag{G.7}$$

and hence $\mathrm{Leakage}_n$ reduces to a residual-leakage term.

Compatible limit. In the main theorem we work in the *compatible* regime where $f_\star \in \mathcal{H}_{\Omega_\star}$ for the limit feature parameter $\Omega_\star$, equivalently

$$e_{\Omega_\star} = (I - \Pi_{\Omega_\star}) f_\star = 0. \tag{G.8}$$

As a consequence of the residual expansion proved in Appendix G.2, we have

$$\|e_{\Omega_m}\|_{L^2(\mu_{\mathrm{down}})} \xrightarrow{\mathbb{P}} 0.$$

This residual vanishing is the ingredient that makes the signal estimation term $\mathrm{Leakage}_n$ vanish at the scale n.

G.3.2. PERTURBATION LEMMAS FOR PSEUDOINVERSES

Lemma G.15 (Empirical span is contained in the population span) *Fix m and condition on $D_{\mathrm{pre}}^{(m)}$. Assume $\mathbb{E}[\|\phi(X,\Omega_m)\|^2 \mid D_{\mathrm{pre}}^{(m)}] < \infty$, so that Σ_m is well-defined. Let $S := \mathrm{Im}(\Sigma_m)$. Then $\phi(X,\Omega_m) \in S$ almost surely. In particular, $\mathrm{Im}(\Sigma_{m,n}) \subseteq S$ and $\mathrm{rank}(\Sigma_{m,n}) \leqslant \mathrm{rank}(\Sigma_m)$ almost surely.*

Proof Let $K := \ker(\Sigma_m)$ and let v be any unit vector in K. Then

$$0 = v^\top \Sigma_m v = \mathbb{E}\left[(v^\top \phi(X,\Omega_m))^2\right],$$

so $v^\top \phi(X,\Omega_m) = 0$ almost surely. Hence $\phi(X,\Omega_m) \in K^\perp = \mathrm{Im}(\Sigma_m)$ almost surely. Applying the same argument to each X_i yields the claims for $\Sigma_{m,n}$. ∎

Lemma G.16 (Relative empirical covariance from leverage moments) *Assume Assumption G.2. Let $S_m := \mathrm{Im}(\Sigma_m)$ and $P_m := \Pi_{S_m}$ where P_m is the Euclidean orthogonal projector onto S_m. Define*

$$C_{m,n} := \Sigma_m^{+/2} \Sigma_{m,n} \Sigma_m^{+/2}.$$

If $\Omega_m \to \Omega_\star$ in probability, then

$$\left\|C_{m,n} - P_m\right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0.$$

Proof Fix m, n and condition on $D_{\mathrm{pre}}^{(m)}$. Write

$$\phi_m(X) := \phi(X,\Omega_m), \qquad \widetilde{\phi}_m(X) := \Sigma_m^{+/2} \phi_m(X).$$

Then

$$C_{m,n} = \frac{1}{n} \sum_{i=1}^n \widetilde{\phi}_m(X_i) \widetilde{\phi}_m(X_i)^\top.$$

Moreover,

$$\mathbb{E}\left[\widetilde{\phi}_m(X) \widetilde{\phi}_m(X)^\top \,\middle|\, D_{\mathrm{pre}}^{(m)}\right] = \Sigma_m^{+/2} \Sigma_m \Sigma_m^{+/2} = P_m.$$

Therefore

$$C_{m,n} - P_m = \frac{1}{n} \sum_{i=1}^n U_{m,i},$$

where

$$U_{m,i} \coloneqq \widetilde{\phi}_m(X_i)\widetilde{\phi}_m(X_i)^\top - P_m$$

are conditionally i.i.d. mean-zero matrices.

For each entry (j,k), write $U_{m,i}^{(jk)}$ for the (j,k)-entry. On the event $\{\Omega_m \in \mathcal{U}\}$, Assumption G.2 gives

$$\mathbb{E}\Big[\|\widetilde{\phi}_m(X)\|^4 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] = \mathbb{E}\Big[q_{\Omega_m}(X)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] \leqslant C_q^{2/(2+\eta)}.$$

Hence there exists a finite constant C, depending only on C_q, such that on $\{\Omega_m \in \mathcal{U}\}$,

$$\mathbb{E}\Big[\big(U_{m,1}^{(jk)}\big)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] \leqslant C$$

for all j,k. Therefore, by Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n U_{m,i}^{(jk)}\right| > \frac{\varepsilon}{p} \,\middle|\, D_{\mathrm{pre}}^{(m)}\right) \leqslant \frac{Cp^2}{n\varepsilon^2}$$

on $\{\Omega_m \in \mathcal{U}\}$. Taking a union bound over all p^2 entries gives

$$\mathbb{P}\Big(\|C_{m,n} - P_m\|_{\mathrm{F}} > \varepsilon \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big) \leqslant \frac{Cp^4}{n\varepsilon^2}$$

on $\{\Omega_m \in \mathcal{U}\}$. Since

$$\|C_{m,n} - P_m\|_{\mathrm{op}} \leqslant \|C_{m,n} - P_m\|_{\mathrm{F}},$$

we obtain

$$\mathbb{P}(\|C_{m,n} - P_m\|_{\mathrm{op}} > \varepsilon) \leqslant \mathbb{P}(\Omega_m \notin \mathcal{U}) + \frac{Cp^4}{n\varepsilon^2}.$$

Because $\Omega_m \to \Omega_\star$ in probability and $\mathcal{U}$ is a neighborhood of $\Omega_\star$, the first term tends to zero. The second term also tends to zero as $n \to \infty$. Hence

$$\|C_{m,n} - P_m\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0.$$

■

Lemma G.17 (Relative trace approximation without an eigengap) *Assume Assumption G.2. If $\Omega_m \to \Omega_\star$ in probability, then*

$$\mathrm{tr}(\Sigma_m \Sigma_{m,n}^+) - \mathrm{rank}(\Sigma_m) \xrightarrow{\mathbb{P}} 0.$$

Proof Let

$$S_m \coloneqq \mathrm{Im}(\Sigma_m), \qquad P_m \coloneqq \Pi_{S_m}, \qquad r_m \coloneqq \mathrm{rank}(\Sigma_m),$$

and write

$$C_{m,n} = \Sigma_m^{+/2}\Sigma_{m,n}\Sigma_m^{+/2}.$$

By Lemma G.15, $\mathrm{Im}(\Sigma_{m,n}) \subseteq S_m$ almost surely.

Fix $0 < \rho < 1$, and work on the event

$$\mathcal{E}^{\mathrm{rel}}_{m,n}(\rho) := \{\|C_{m,n} - P_m\|_{\mathrm{op}} \leqslant \rho\}.$$

On S_m, the projector P_m is the identity. Hence every eigenvalue of $C_{m,n}|_{S_m}$ lies in

$$[1-\rho, 1+\rho].$$

In particular, $C_{m,n}|_{S_m}$ is invertible.

Since $\mathrm{Im}(\Sigma_{m,n}) \subseteq S_m$, we have the factorization

$$\Sigma_{m,n} = \Sigma_m^{1/2} C_{m,n} \Sigma_m^{1/2}$$

on S_m, and both sides vanish on $S_m^{\perp}$. Therefore, on S_m,

$$\Sigma^+_{m,n} = \Sigma_m^{+/2} \big(C_{m,n}|_{S_m}\big)^{-1} \Sigma_m^{+/2},$$

and both sides vanish on $S_m^{\perp}$. Consequently,

$$\begin{aligned}\mathrm{tr}(\Sigma_m \Sigma^+_{m,n}) &= \mathrm{tr}\Big(\Sigma_m \Sigma_m^{+/2} \big(C_{m,n}|_{S_m}\big)^{-1} \Sigma_m^{+/2}\Big) = \mathrm{tr}\Big(\Sigma_m^{+/2} \Sigma_m \Sigma_m^{+/2} \big(C_{m,n}|_{S_m}\big)^{-1}\Big) \\ &= \mathrm{tr}\Big(P_m \big(C_{m,n}|_{S_m}\big)^{-1}\Big) = \mathrm{tr}\Big(\big(C_{m,n}|_{S_m}\big)^{-1}\Big).\end{aligned}$$

Let $\lambda_1, \dots, \lambda_{r_m}$ be the eigenvalues of $C_{m,n}|_{S_m}$. Then

$$\mathrm{tr}(\Sigma_m \Sigma^+_{m,n}) = \sum_{j=1}^{r_m} \frac{1}{\lambda_j}.$$

Hence, on $\mathcal{E}^{\mathrm{rel}}_{m,n}(\rho)$,

$$\big|\mathrm{tr}(\Sigma_m \Sigma^+_{m,n}) - r_m\big| = \left|\sum_{j=1}^{r_m} \left(\frac{1}{\lambda_j} - 1\right)\right| \leqslant \sum_{j=1}^{r_m} \frac{|\lambda_j - 1|}{\lambda_j} \leqslant r_m \frac{\rho}{1-\rho} \leqslant p \frac{\rho}{1-\rho}.$$

By Lemma G.16,

$$\mathbb{P}\Big(\mathcal{E}^{\mathrm{rel}}_{m,n}(\rho)\Big) \to 1$$

for every fixed $0 < \rho < 1$. Since $\rho > 0$ can be taken arbitrarily small,

$$\mathrm{tr}(\Sigma_m \Sigma^+_{m,n}) - r_m \xrightarrow{\mathbb{P}} 0.$$

This proves the claim. ∎

Lemma G.18 (Active effective dimension under stable null span) *Assume Assumptions G.2–G.4. Let*

$$d_\star := \dim \mathcal{H}_{\Omega_\star} = d_{\mathrm{eff}}(\Omega_\star), \qquad b_0 := \dim \mathcal{B}_0,$$

and define

$$d_{\mathrm{act}} := \dim(\mathcal{H}_{\Omega_\star} \oplus \mathcal{B}_0) = d_\star + b_0.$$

Then

$$\mathrm{rank}(\Sigma_m) \xrightarrow{\mathbb{P}} d_{\mathrm{act}}.$$

Proof Since $T_{\Omega_\star}|_{E_\star}$ is injective and $E_\star$ is finite-dimensional, injectivity is stable under sufficiently small operator-norm perturbations. By Lemma G.5,

$$T_v|_{E_\star} \to T_{\Omega_\star}|_{E_\star}$$

in operator norm. Therefore, for all sufficiently small v,

$$\dim \mathcal{A}_v = \dim \mathcal{H}_{\Omega_\star} = d_\star.$$

By Assumption G.4,

$$\|\Pi_{\mathcal{B}_v} - \Pi_{\mathcal{B}_0}\|_{\mathrm{op}} \to 0.$$

For orthogonal projectors, if $\|P - Q\|_{\mathrm{op}} < 1$, then P and Q have the same rank. Hence, for all sufficiently small nonzero v,

$$\dim \mathcal{B}_v = \dim \mathcal{B}_0 = b_0.$$

Therefore, for all sufficiently small nonzero v,

$$\dim \mathcal{H}_v = \dim \mathcal{A}_v + \dim \mathcal{B}_v = d_\star + b_0 = d_{\mathrm{act}}.$$

Finally,

$$\mathrm{rank}(\Sigma(\Omega(v))) = \mathrm{rank}(T_v^{\mathrm{adj}} T_v) = \mathrm{rank}(T_v) = \dim \mathcal{H}_v.$$

Thus, for all sufficiently small nonzero v,

$$\mathrm{rank}(\Sigma(\Omega(v))) = d_{\mathrm{act}}.$$

Applying this with $v = \log_{\Omega_\star}(\Omega_m)$, and using $\log_{\Omega_\star}(\Omega_m) \to 0$ in probability gives

$$\mathrm{rank}(\Sigma_m) \xrightarrow{\mathbb{P}} d_{\mathrm{act}}.$$

■

G.3.3. NOISE ESTIMATION TERM

Lemma G.19 (Closed form for the noise term) *For each* (n, m),

$$\mathrm{Var}_{n,m} = \frac{\sigma^2}{n}\, \mathrm{tr}\big(\Sigma_m\, \Sigma_{m,n}^{+}\big).$$

Proof Fix n and condition on $(D_{\mathrm{pre}}^{(m)}, X_{1:n}, X_{\mathrm{new}})$. By Definition C.7 (Appendix C),

$$\Pi_{\Omega_m,n}\varepsilon = T_{\Omega_m}\hat{\theta}_\varepsilon, \qquad \hat{\theta}_\varepsilon = \Sigma_{m,n}^{+}\,\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\,\phi(X_i, \Omega_m).$$

Let $g_m := \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(X_i, \Omega_m)$, so $\hat{\theta}_\varepsilon = \Sigma_{m,n}^{+} g_m$ and

$$(\Pi_{\Omega_m,n}\varepsilon)(X_{\mathrm{new}}) = \phi(X_{\mathrm{new}}, \Omega_m)^\top \Sigma_{m,n}^{+} g_m.$$

Since $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ and $\mathbb{E}[\varepsilon_i^2 \mid X_i] = \sigma^2$,

$$\mathbb{E}[g_m g_m^\top \mid D_{\mathrm{pre}}^{(m)}, X_{1:n}] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[\varepsilon_i^2 \mid X_i] \phi(X_i, \Omega_m) \phi(X_i, \Omega_m)^\top = \frac{\sigma^2}{n} \Sigma_{m,n}.$$

Therefore,

$$\begin{aligned} \mathrm{Var}_{n,m} &= \mathbb{E}\left[\phi(X_{\mathrm{new}}, \Omega_m)^\top \Sigma_{m,n}^+ g_m g_m^\top \Sigma_{m,n}^+ \phi(X_{\mathrm{new}}, \Omega_m) \,\middle|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}, X_{\mathrm{new}} \right] \\ &= \frac{\sigma^2}{n} \, \phi(X_{\mathrm{new}}, \Omega_m)^\top \Sigma_{m,n}^+ \phi(X_{\mathrm{new}}, \Omega_m), \end{aligned}$$

using $\Sigma_{m,n}^+ \Sigma_{m,n} \Sigma_{m,n}^+ = \Sigma_{m,n}^+$. Taking conditional expectation over X_{new} yields

$$\mathrm{Var}_{n,m} = \frac{\sigma^2}{n} \, \mathrm{tr}\big(\Sigma_m \, \Sigma_{m,n}^+\big).$$

■

Corollary G.20 (Noise estimation term under stable null span) *Assume Assumptions G.2, G.3, and G.4. Let*

$$d_{\mathrm{act}} := \dim(\mathcal{H}_{\Omega_\star} \oplus \mathcal{B}_0) = d_{\mathrm{eff}}(\Omega_\star) + \dim \mathcal{B}_0.$$

Then, along $n, m \to \infty$ with $m/n \to \alpha$,

$$n \, \mathrm{Var}_{n,m} \xrightarrow{\mathbb{P}} \sigma^2 d_{\mathrm{act}}.$$

Proof By Lemma G.19,

$$n \, \mathrm{Var}_{n,m} = \sigma^2 \, \mathrm{tr}(\Sigma_m \Sigma_{m,n}^+).$$

By Lemma G.17,

$$\mathrm{tr}(\Sigma_m \Sigma_{m,n}^+) - \mathrm{rank}(\Sigma_m) \xrightarrow{\mathbb{P}} 0.$$

By Lemma G.18,

$$\mathrm{rank}(\Sigma_m) \xrightarrow{\mathbb{P}} d_{\mathrm{act}}.$$

Therefore,

$$\mathrm{tr}(\Sigma_m \Sigma_{m,n}^+) \xrightarrow{\mathbb{P}} d_{\mathrm{act}}.$$

Multiplying by σ^2 yields

$$n \, \mathrm{Var}_{n,m} \xrightarrow{\mathbb{P}} \sigma^2 d_{\mathrm{act}}.$$

■

G.3.4. LEAKAGE ESTIMATION TERM

Under Assumption G.1, (G.7) yields

$$\mathrm{Leakage}_n = \mathbb{E}\Big[\big(\Pi_{\Omega_m,n} e_{\Omega_m}\big)(X_{\mathrm{new}})^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}, X_{1:n}\Big].$$

Define the (random) population matrix

$$\Sigma_{e,m} \coloneqq \Sigma_e(\Omega_m) \coloneqq \mathbb{E}\Big[e_{\Omega_m}(X)^2\, \phi(X,\Omega_m)\phi(X,\Omega_m)^\top \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big], \qquad X \sim \mu_{\mathrm{down}}.$$

Lemma G.21 (Vanishing whitened signal covariance in the compatible limit) *Assume Assumptions G.2, G.3, and G.4. Then*

$$\mathrm{tr}(\Sigma_m^+ \Sigma_{e,m}) \xrightarrow{\mathbb{P}} 0.$$

Proof Write

$$e_m(X) \coloneqq e_{\Omega_m}(X), \qquad \phi_m(X) \coloneqq \phi(X,\Omega_m), \qquad q_m(X) \coloneqq q_{\Omega_m}(X).$$

By definition,

$$\mathrm{tr}(\Sigma_m^+ \Sigma_{e,m}) = \mathbb{E}\Big[e_m(X)^2 \phi_m(X)^\top \Sigma_m^+ \phi_m(X) \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] = \mathbb{E}\Big[e_m(X)^2 q_m(X) \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big].$$

Apply Hölder's inequality with conjugate exponents $\frac{1+\eta}{\eta}$ and $1+\eta$. Then

$$\begin{aligned}\mathbb{E}\Big[e_m(X)^2 q_m(X) \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] &= \mathbb{E}\Big[\big(e_m(X)^2\big)^{\frac{\eta}{1+\eta}} \big(e_m(X)^2 q_m(X)^{1+\eta}\big)^{\frac{1}{1+\eta}} \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] \\ &\leqslant \Big(\mathbb{E}\Big[e_m(X)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big]\Big)^{\frac{\eta}{1+\eta}} \Big(\mathbb{E}\Big[e_m(X)^2 q_m(X)^{1+\eta} \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big]\Big)^{\frac{1}{1+\eta}}.\end{aligned}$$

Let $v_m = \log_{\Omega_\star}(\Omega_m)$. By the residual expansion in Corollary G.9,

$$e_{\Omega_m} = \mathcal{L}(v_m) + r(v_m), \qquad \|r(v_m)\|_{L^2(\mu_{\mathrm{down}})} \leqslant \omega(\|v_m\|)\|v_m\|.$$

Since $v_m \to 0$ in probability and $\mathcal{L}$ is bounded, this implies

$$\mathbb{E}\Big[e_m(X)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] = \|e_{\Omega_m}\|^2_{L^2(\mu_{\mathrm{down}})} \xrightarrow{\mathbb{P}} 0.$$

On the event $\{\Omega_m \in \mathcal{U}\}$, Assumption G.2 gives

$$\mathbb{E}\Big[e_m(X)^2 q_m(X)^{1+\eta} \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] \leqslant C_{\mathrm{lev}}.$$

Therefore, on $\{\Omega_m \in \mathcal{U}\}$,

$$\mathrm{tr}(\Sigma_m^+ \Sigma_{e,m}) \leqslant C_{\mathrm{lev}}^{\frac{1}{1+\eta}} \Big(\mathbb{E}\Big[e_m(X)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big]\Big)^{\frac{\eta}{1+\eta}}.$$

Because $\Omega_m \to \Omega_\star$ in probability, we have $\mathbb{P}(\Omega_m \notin \mathcal{U}) \to 0$. Combining the preceding display with

$$\mathbb{E}\Big[e_m(X)^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] \xrightarrow{\mathbb{P}} 0$$

yields

$$\mathrm{tr}(\Sigma_m^+ \Sigma_{e,m}) \xrightarrow{\mathbb{P}} 0.$$

■

Proposition G.22 (Leakage estimation term under stable null span) *Assume Assumptions G.2–G.4. Then, along $m, n \to \infty$ with $m/n \to \alpha$,*

$$n \operatorname{Leakage}_n \xrightarrow{\mathbb{P}} 0.$$

Proof Fix n and condition on $(D_{\mathrm{pre}}^{(m)}, X_{1:n})$. Under Assumption G.1, (G.7) gives

$$\operatorname{Leakage}_n = \mathbb{E}\Big[\big(\Pi_{\Omega_m,n} e_{\Omega_m}\big)(X_{\mathrm{new}})^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}, X_{1:n} \Big].$$

Write

$$e_m(X) \coloneqq e_{\Omega_m}(X), \qquad \phi_m(X) \coloneqq \phi(X, \Omega_m),$$

and define

$$g_m \coloneqq \frac{1}{n} \sum_{i=1}^{n} e_m(X_i)\phi_m(X_i).$$

By the definition of the canonical empirical projector,

$$\big(\Pi_{\Omega_m,n} e_m\big)(X_{\mathrm{new}}) = \phi_m(X_{\mathrm{new}})^\top \Sigma_{m,n}^{+} g_m.$$

Therefore,

$$\begin{aligned} \operatorname{Leakage}_n &= \mathbb{E}\Big[g_m^\top \Sigma_{m,n}^{+} \phi_m(X_{\mathrm{new}}) \phi_m(X_{\mathrm{new}})^\top \Sigma_{m,n}^{+} g_m \,\Big|\, D_{\mathrm{pre}}^{(m)}, X_{1:n} \Big] \\ &= g_m^\top \Sigma_{m,n}^{+} \Sigma_m \Sigma_{m,n}^{+} g_m. \end{aligned}$$

Thus

$$n \operatorname{Leakage}_n = (\sqrt{n}\, g_m)^\top \Sigma_{m,n}^{+} \Sigma_m \Sigma_{m,n}^{+} (\sqrt{n}\, g_m).$$

Define the whitened residual-correlation vector

$$\widetilde{g}_m \coloneqq \Sigma_m^{+/2} g_m.$$

We first show that

$$\sqrt{n}\, \widetilde{g}_m \xrightarrow{\mathbb{P}} 0.$$

Condition only on $D_{\mathrm{pre}}^{(m)}$. The random vectors

$$\widetilde{W}_{m,i} \coloneqq e_m(X_i) \Sigma_m^{+/2} \phi_m(X_i), \qquad i = 1, \dots, n,$$

are conditionally i.i.d. Since $e_m \perp \mathcal{H}_{\Omega_m}$ in $L^2(\mu_{\mathrm{down}})$, and each coordinate of ϕ_m belongs to $\mathcal{H}_{\Omega_m}$, we have

$$\mathbb{E}\Big[e_m(X_i)\phi_m(X_i) \,\Big|\, D_{\mathrm{pre}}^{(m)} \Big] = 0.$$

Hence

$$\mathbb{E}\Big[\widetilde{W}_{m,i} \,\Big|\, D_{\mathrm{pre}}^{(m)} \Big] = 0.$$

Moreover,

$$\sqrt{n}\, \widetilde{g}_m = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{W}_{m,i}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\Big[\|\sqrt{n}\,\widetilde{g}_m\|^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] &= \mathbb{E}\Big[\|\widetilde{W}_{m,1}\|^2 \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] \\ &= \mathbb{E}\Big[e_m(X)^2\phi_m(X)^\top\Sigma_m^+\phi_m(X) \,\Big|\, D_{\mathrm{pre}}^{(m)}\Big] \\ &= \mathrm{tr}(\Sigma_m^+\Sigma_{e,m}).\end{aligned}$$

By Lemma G.21,

$$\mathrm{tr}(\Sigma_m^+\Sigma_{e,m}) \xrightarrow{\mathbb{P}} 0.$$

For every $\varepsilon > 0$,

$$\mathbb{P}(\|\sqrt{n}\,\widetilde{g}_m\| > \varepsilon) \leqslant \mathbb{E}\left[\min\left\{\frac{\mathbb{E}[\|\sqrt{n}\,\widetilde{g}_m\|^2 \mid D_{\mathrm{pre}}^{(m)}]}{\varepsilon^2}, 1\right\}\right].$$

The conditional second moment is $\mathrm{tr}(\Sigma_m^+\Sigma_{e,m}) \to_{\mathbb{P}} 0$, and the expression inside the expectation is bounded by 1. Hence dominated convergence along convergence in probability gives

$$\sqrt{n}\,\widetilde{g}_m \to_{\mathbb{P}} 0.$$

It remains to control the quadratic form. Let

$$S_m \coloneqq \mathrm{Im}(\Sigma_m), \qquad P_m \coloneqq \Pi_{S_m}, \qquad C_{m,n} \coloneqq \Sigma_m^{+/2}\Sigma_{m,n}\Sigma_m^{+/2}.$$

By Lemma G.15,

$$\mathrm{Im}(\Sigma_{m,n}) \subseteq S_m$$

almost surely. Work on the event

$$\mathcal{E}_{m,n}^{\mathrm{rel}} \coloneqq \left\{\|C_{m,n} - P_m\|_{\mathrm{op}} \leqslant \frac{1}{2}\right\}.$$

By Lemma G.16,

$$\mathbb{P}(\mathcal{E}_{m,n}^{\mathrm{rel}}) \to 1.$$

On this event, $C_{m,n}|_{S_m}$ is invertible and

$$\left\|\big(C_{m,n}|_{S_m}\big)^{-1}\right\|_{\mathrm{op}} \leqslant 2.$$

As in Lemma G.17,

$$\Sigma_{m,n}^+ = \Sigma_m^{+/2}\big(C_{m,n}|_{S_m}\big)^{-1}\Sigma_m^{+/2}$$

on S_m, and both sides vanish on $S_m^\perp$. Therefore,

$$\Sigma_{m,n}^+\Sigma_m\Sigma_{m,n}^+ = \Sigma_m^{+/2}\big(C_{m,n}|_{S_m}\big)^{-2}\Sigma_m^{+/2}.$$

Substituting into the leakage quadratic form yields

$$n\,\mathrm{Leakage}_n = (\sqrt{n}\,\widetilde{g}_m)^\top\big(C_{m,n}|_{S_m}\big)^{-2}(\sqrt{n}\,\widetilde{g}_m).$$

Hence, on $\mathcal{E}_{m,n}^{\mathrm{rel}}$,

$$0 \leqslant n\,\mathrm{Leakage}_n \leqslant 4\|\sqrt{n}\,\widetilde{g}_m\|^2.$$

Since

$$\sqrt{n}\,\widetilde{g}_m \xrightarrow{\mathbb{P}} 0$$

and $\mathbb{P}(\mathcal{E}_{m,n}^{\mathrm{rel}}) \to 1$, we conclude that

$$n\,\mathrm{Leakage}_n \xrightarrow{\mathbb{P}} 0.$$

■

G.4. Dimension gap

We record a simple observation explaining the role of the dimension gap in the downstream-only comparison. Work locally around a regular realization $(\theta_\star, \Omega_\star)$ of $f_\star$, so that

$$f_{\theta,\Omega}(x) = \langle \theta, \phi(x, \Omega)\rangle, \qquad f_\star = f_{\theta_\star,\Omega_\star}.$$

Let

$$\mathcal{H}_{\Omega_\star} \coloneqq \{\langle \theta, \phi(\cdot, \Omega_\star)\rangle : \theta \in \mathbb{R}^p\} \subset L^2(\mu_{\mathrm{down}})$$

be the fixed-representation downstream class of the optimal representation $\Omega_\star$. We have define

$$d_{\mathrm{eff}}(\Omega_\star) \coloneqq \dim \mathcal{H}_{\Omega_\star}.$$

The downstream-only tangent space at $(\theta_\star, \Omega_\star)$ is

$$\mathcal{T}_{\mathrm{base}} \coloneqq \left\{ \langle \delta\theta, \phi(\cdot, \Omega_\star)\rangle + \langle \theta_\star, D\phi(\cdot, \Omega_\star)[v]\rangle \,\middle|\, \delta\theta \in \mathbb{R}^p,\ v \in T_{\Omega_\star}\mathcal{M} \right\} \subset L^2(\mu_{\mathrm{down}}),$$

and we define

$$d_{\mathrm{base}} \coloneqq \dim \mathcal{T}_{\mathrm{base}}.$$

Finally, recall the linear map for the case $\mathcal{B}_0 = \{0\}$

$$\mathcal{L}(v) \coloneqq -D\Pi_{\Omega_\star}[v] f_\star, \qquad v \in T_{\Omega_\star}\mathcal{M}.$$

Proposition G.23 *We have $d_{\mathrm{base}} \geqslant d_{\mathrm{eff}}(\Omega_\star)$. Moreover,*

$$d_{\mathrm{base}} = d_{\mathrm{eff}}(\Omega_\star) \qquad \Longleftrightarrow \qquad \mathcal{L} \equiv 0.$$

Proof The directions obtained by varying only the readout parameter are

$$\{\langle \delta\theta, \phi(\cdot, \Omega_\star)\rangle : \delta\theta \in \mathbb{R}^p\} = \mathcal{H}_{\Omega_\star}.$$

These directions are contained in $\mathcal{T}_{\mathrm{base}}$, hence $\mathcal{H}_{\Omega_\star} \subseteq \mathcal{T}_{\mathrm{base}}$, which implies

$$d_{\mathrm{base}} \geqslant d_{\mathrm{eff}}(\Omega_\star).$$

For the equivalence, fix $v \in T_{\Omega_\star}\mathcal{M}$ and define

$$h_v := \langle \theta_\star, D\phi(\cdot, \Omega_\star)[v] \rangle.$$

By definition,

$$\mathcal{T}_{\text{base}} = \mathcal{H}_{\Omega_\star} + \{h_v : v \in T_{\Omega_\star}\mathcal{M}\}.$$

Since $\mathcal{H}_{\Omega_\star} \subseteq \mathcal{T}_{\text{base}}$, we have $d_{\text{base}} = d_{\text{eff}}(\Omega_\star)$ if and only if

$$h_v \in \mathcal{H}_{\Omega_\star} \qquad \text{for every } v \in T_{\Omega_\star}\mathcal{M}.$$

Now let $\Omega_t = \exp_{\Omega_\star}(tv)$. Since $f_{\theta_\star,\Omega_t} \in \mathcal{H}_{\Omega_t}$, we have

$$\Pi_{\Omega_t} f_{\theta_\star,\Omega_t} = f_{\theta_\star,\Omega_t}.$$

Differentiating at $t = 0$ gives

$$D\Pi_{\Omega_\star}[v] f_\star + \Pi_{\Omega_\star} h_v = h_v.$$

Therefore,

$$D\Pi_{\Omega_\star}[v] f_\star = (I - \Pi_{\Omega_\star}) h_v,$$

and hence

$$\mathcal{L}(v) = -D\Pi_{\Omega_\star}[v] f_\star = -(I - \Pi_{\Omega_\star}) h_v.$$

Thus $\mathcal{L}(v) = 0$ if and only if $h_v \in \mathcal{H}_{\Omega_\star}$. Since this holds for every $v \in T_{\Omega_\star}\mathcal{M}$, we conclude that

$$\mathcal{L} \equiv 0 \qquad \Longleftrightarrow \qquad d_{\text{base}} = d_{\text{eff}}(\Omega_\star).$$

The final equivalence follows by taking the contrapositive together with $d_{\text{base}} \geqslant d_{\text{eff}}(\Omega_\star)$. ■

Appendix H. Proofs of Section 6.1

This appendix verifies that the linear spectral contrastive model of Section 6.1 satisfies the standing assumptions of Theorem 5.1. We separate the verification into two parts. First, we verify the quotient geometry and downstream assumptions (Assumptions G.1–G.4). Second, we verify the pre-training CLT assumptions, namely Assumption D.5. Once these checks are in place, Proposition 6.2 follows by a direct invocation of Theorem 5.1.

H.1. Geometry, descriptor, and quotient feature map

Quotient feature map $\phi(x, M)$ via a local section. To express downstream prediction purely in quotient coordinates, we define a feature map that depends on the descriptor $M \in \mathcal{M}_{d,k}$. Fix a regular point $M_\star \in \mathcal{M}_{d,k}$ and a neighborhood $\mathcal{W} \subset \mathcal{M}_{d,k}$ of $M_\star$ on which there exists a C^2 *local section*

$$s : \mathcal{W} \to \mathbb{R}^{k \times d}, \qquad s(M)^\mathsf{T} s(M) = M \ \text{ for all } M \in \mathcal{W},$$

as constructed in Appendix E.2. We then define the *quotient feature map*

$$\phi(x, M) := s(M)\, x \in \mathbb{R}^k, \qquad (x, M) \in \mathbb{R}^d \times \mathcal{W}. \tag{H.1}$$

Any two choices of local section differ by a left-multiplication $s'(M) = Q(M)s(M)$ with $Q(M) \in O(k)$, hence they generate features related by an orthogonal transform.

Concrete local choice of the section $s(M)$. Fix a reference point $M_\star \in \mathcal{M}_{d,k}$ and choose an orthonormal basis $U_\star \in \mathbb{R}^{d\times k}$ of $\mathrm{range}(M_\star)$ (so $U_\star^\top U_\star = I_k$). For M in a sufficiently small neighborhood $\mathcal{W}$ of $M_\star$, let $P(M)$ denote the orthogonal projector onto $\mathrm{range}(M)$. Define

$$B(M) \coloneqq U_\star^\top P(M) U_\star \in \mathbb{R}^{k\times k},\; U(M) \coloneqq P(M) U_\star\, B(M)^{-1/2} \in \mathbb{R}^{d\times k},$$

so that $U(M)^\top U(M) = I_k$ and $\mathrm{range}(U(M)) = \mathrm{range}(M)$. Next set

$$\Lambda(M) \coloneqq U(M)^\top M U(M) \in \mathbb{R}^{k\times k},$$

which is positive definite on $\mathcal{W}$ and satisfies $M = U(M)\Lambda(M)U(M)^\top$. A concrete local section is then

$$s(M) \coloneqq \Lambda(M)^{1/2} U(M)^\top \in \mathbb{R}^{k\times d},$$

so that $s(M)^\top s(M) = M$. Consequently, the quotient-level feature map can be written explicitly as

$$\phi(x, M) = s(M)x = \Lambda(M)^{1/2} U(M)^\top x \in \mathbb{R}^k.$$

All smoothness claims (existence of $\mathcal{W}$ and differentiability of $M \mapsto s(M)$) are proved in Appendix E.2.

H.2. Population descriptor problem and regularity of $M_\star$

Population loss in descriptor space. The following lemma expresses the population objective in terms of $M = A^\top A$.

Lemma H.1 *Assume $\mathbb{E}\|x\|^4 < \infty$. Then*

$$L_{\mathrm{spec}}(M) = -2\,\mathrm{tr}\big(M\Sigma_{\mathrm{pre}}^+\big) + \mathrm{tr}\big((M\Sigma_{\mathrm{pre}})^2\big). \tag{H.2}$$

If furthermore Σ_{pre} is full-rank, then

$$L_{\mathrm{spec}}(M) = \big\|\Sigma_{\mathrm{pre}}^{1/2} M \Sigma_{\mathrm{pre}}^{1/2} - C\big\|_F^2 - \|C\|_F^2, \tag{H.3}$$

where $C \coloneqq \Sigma_{\mathrm{pre}}^{-1/2}\,\Sigma_{\mathrm{pre}}^{+}\,\Sigma_{\mathrm{pre}}^{-1/2}$.

Proof Let $\Sigma \coloneqq \Sigma_{\mathrm{pre}} = \mathbb{E}[xx^\top]$. Since x^- is an independent copy of x, we have

$$\begin{aligned} L_{\mathrm{spec}}(M) &= -2\mathbb{E}[x^\top M x^+] + \mathbb{E}[(x^\top M x^-)^2] = -2\,\mathrm{tr}(\mathbb{E}[Mx^+x^\top) + \mathrm{tr}(\mathbb{E}[Mx(x^-)^\top Mxx^\top]) \\ &= -2\,\mathrm{tr}(M\Sigma_{\mathrm{pre}}^+) + \mathrm{tr}(M\Sigma M\Sigma) \end{aligned}$$

which proves (H.2). If Σ is full rank, write

$$\mathrm{tr}\big((M\Sigma)^2\big) = \|\Sigma^{1/2} M \Sigma^{1/2}\|_F^2, \qquad \mathrm{tr}(M\Sigma^+) = \big\langle \Sigma^{1/2} M \Sigma^{1/2},\, C\big\rangle_F,$$

with $C = \Sigma^{-1/2}\Sigma^+\Sigma^{-1/2}$. Completing the square gives (H.3). ∎

Remark H.2 *Even though Σ_{pre} and M are PSD, the matrix C can be indefinite: augmentations may induce negative correlations along some directions.*

Uniqueness of the descriptor minimizer. By Lemma H.1, the population minimization in descriptor space reduces to

$$M_\star \in \operatorname*{arg\,min}_{M \in \mathcal{M}_{d,k}} \left\| \Sigma_{\mathrm{pre}}^{1/2} M \Sigma_{\mathrm{pre}}^{1/2} - C \right\|_F^2. \tag{H.4}$$

We now state a simple eigengap condition that guarantees uniqueness.

Assume throughout $\Sigma_{\mathrm{pre}} \succ 0$ and Assumption 6.1. Recall the whitened matrix

$$C \coloneqq \Sigma_{\mathrm{pre}}^{-1/2} \Sigma_{\mathrm{pre}}^{+} \Sigma_{\mathrm{pre}}^{-1/2},$$

and the descriptor manifold $\mathcal{M}_{d,k} \coloneqq \{M \succcurlyeq 0 : \operatorname{rank}(M) = k\}$. Thus, the population minimizer $M_\star$ exists, is unique, and satisfies

$$\Sigma_{\mathrm{pre}}^{1/2} M_\star \Sigma_{\mathrm{pre}}^{1/2} = U_k \Lambda_k U_k^\top,$$

where $U_k \Lambda_k U_k^\top$ is the rank-k truncation onto the top-k *positive* eigenvalues of C. In particular, $M_\star \in \mathcal{M}_{d,k}$ is a regular point of the fixed-rank PSD manifold.

H.3. Verification of Assumptions G.1–G.4

This subsection verifies all assumptions of Theorem 5.1 except the pre-training CLT assumption D.5.

H.3.1. LOCAL FEATURE REGULARITY AND MOMENTS

Fix the regular point $M_\star \in \mathcal{M}_{d,k}$ and let $s(\cdot)$ be the C^2 local section from Appendix E.2 (equivalently, the concrete construction in Section 6.1), defined on a neighborhood $\mathcal{W} \subset \mathcal{M}_{d,k}$ of $M_\star$ and satisfying $s(M)^\top s(M) = M$. Define the quotient feature map

$$\phi(x, M) \coloneqq s(M)x \in \mathbb{R}^k, \qquad (x, M) \in \mathbb{R}^d \times \mathcal{W}.$$

By Appendix E.2, the map $M \mapsto s(M)$ is C^2 on $\mathcal{W}$, hence $M \mapsto \phi(x, M)$ is C^1 for each fixed x.

Lemma H.3 (Local-uniform feature and derivative moments) *Assume $\mathbb{E}\|X\|^{4+\delta} < \infty$ for some $\delta > 0$, where $X \sim \mu_{\mathrm{down}}$. Then there exists a neighborhood $\mathcal{U} \subseteq \mathcal{W}$ of $M_\star$ and a constant $C < \infty$ such that*

$$\sup_{M \in \mathcal{U}} \mathbb{E}\|\phi(X, M)\|^{4+\delta} \leqslant C, \qquad \sup_{M \in \mathcal{U}} \mathbb{E}\|D_M \phi(X, M)\|_{\mathrm{op}}^{4+\delta} \leqslant C.$$

In particular, the local C^1 feature regularity condition in Assumption G.3 is satisfied.

Proof Since $s(\cdot)$ is C^2 on $\mathcal{W}$, its operator norm and the operator norm of its derivative are locally bounded. Choose a relatively compact neighborhood $\mathcal{U} \Subset \mathcal{W}$ of $M_\star$ so that

$$\sup_{M \in \mathcal{U}} \|s(M)\|_{\mathrm{op}} < \infty, \qquad \sup_{M \in \mathcal{U}} \|D_M s(M)\|_{\mathrm{op}} < \infty.$$

Then $\|\phi(X, M)\|_2 \leqslant \|s(M)\|_{\mathrm{op}} \|X\|_2$ and $\|D_M \phi(X, M)\|_{\mathrm{op}} \leqslant \|D_M s(M)\|_{\mathrm{op}} \|X\|_2$. Raising to the power $4 + \delta$ and taking expectations yields the stated bounds. ∎

H.3.2. DOWNSTREAM COVARIANCE, STABLE RANK, AND EFFECTIVE DIMENSION

Write $\Sigma_{\mathrm{down}} := \mathbb{E}[XX^\top]$ and assume $\Sigma_{\mathrm{down}} \succ 0$. For $M \in \mathcal{U}$, define the downstream feature covariance

$$\Sigma(M) := \mathbb{E}[\phi(X,M)\phi(X,M)^\top] = \mathbb{E}[s(M)XX^\top s(M)^\top] = s(M)\Sigma_{\mathrm{down}}s(M)^\top \in \mathbb{R}^{k\times k}.$$

Lemma H.4 (Stable rank/eigengap for $\Sigma(M)$ near $M_\star$) *Assume $\Sigma_{\mathrm{down}} \succ 0$ and let $\mathcal{U}$ be as in Lemma H.3. Then there exist constants $\kappa_\Sigma, K_\Sigma \in (0,\infty)$ such that, for all $M \in \mathcal{U}$,*

$$\mathrm{rank}(\Sigma(M)) = k, \qquad \lambda_k(\Sigma(M)) \geqslant \kappa_\Sigma, \qquad \|\Sigma(M)\|_{\mathrm{op}} \leqslant K_\Sigma.$$

In particular, $d_{\mathrm{eff}}(M) = \mathrm{tr}(\Sigma(M)\Sigma(M)^+) = k$ for all $M \in \mathcal{U}$.

Proof For each $M \in \mathcal{U}$, the matrix $s(M) \in \mathbb{R}^{k\times d}$ has rank k because $s(M)^\top s(M) = M$ and $\mathrm{rank}(M) = k$. Since $\Sigma_{\mathrm{down}} \succ 0$, for every nonzero $u \in \mathbb{R}^k$,

$$u^\top\Sigma(M)u = u^\top s(M)\Sigma_{\mathrm{down}}s(M)^\top u = \|\Sigma_{\mathrm{down}}^{1/2}s(M)^\top u\|_2^2 > 0.$$

Hence $\Sigma(M) \succ 0$ and $\mathrm{rank}(\Sigma(M)) = k$.

Shrink $\mathcal{U}$, if necessary, so that its closure $\overline{\mathcal{U}}$ is compact. Since $M \mapsto s(M)$ is continuous, the maps

$$M \mapsto \lambda_{\min}(\Sigma(M)), \qquad M \mapsto \|\Sigma(M)\|_{\mathrm{op}}$$

are continuous on $\overline{\mathcal{U}}$. The first is strictly positive on $\overline{\mathcal{U}}$, and the second is finite there. Therefore there exist $\kappa_\Sigma, K_\Sigma \in (0,\infty)$ such that, for all $M \in \mathcal{U}$,

$$\lambda_k(\Sigma(M)) \geqslant \kappa_\Sigma, \qquad \|\Sigma(M)\|_{\mathrm{op}} \leqslant K_\Sigma.$$

Finally, since $\Sigma(M)$ is invertible,

$$d_{\mathrm{eff}}(M) = \mathrm{tr}(\Sigma(M)\Sigma(M)^+) = \mathrm{tr}(I_k) = k.$$

■

H.3.3. LOCAL LEVERAGE AND LEVERAGE-WEIGHTED SIGNAL MOMENTS

Recall that for $M \in \mathcal{U}$, we defined the population leverage score as

$$q_M(X) := \phi(X,M)^\top\Sigma(M)^+\phi(X,M).$$

Lemma H.5 (Local-uniform leverage and signal moments) *Assume $\mathbb{E}\|X\|^{4+\delta} < \infty$ for some $\delta > 0$. Then there exist $\eta > 0$ and constants $C_q, C_e < \infty$ such that, after possibly shrinking $\mathcal{U}$,*

$$\sup_{M\in\mathcal{U}} \mathbb{E}\big[q_M(X)^{2+\eta}\big] \leqslant C_q, \qquad \sup_{M\in\mathcal{U}} \mathbb{E}\big[e_M(X)^2 q_M(X)^{1+\eta}\big] \leqslant C_e.$$

Thus, the local-uniform moment and leverage conditions in Assumption G.2 are satisfied.

Proof By Lemma H.4, $\|\Sigma(M)^+\|_{\mathrm{op}} \leqslant \kappa_\Sigma^{-1}$ uniformly over $M \in \mathcal{U}$. Hence

$$q_M(X) \leqslant \kappa_\Sigma^{-1}\|\phi(X,M)\|^2 \leqslant C\|X\|^2$$

uniformly over $M \in \mathcal{U}$. Choose any $\eta \in (0, \delta/2]$. Then

$$\sup_{M\in\mathcal{U}} \mathbb{E}q_M(X)^{2+\eta} \leqslant C\,\mathbb{E}\|X\|^{4+2\eta} \leqslant C\,\mathbb{E}(1+\|X\|^{4+\delta}) < \infty.$$

It remains to control the signal term. Since $f_\star(x) = \theta_\star^\top \phi(x, M_\star)$ and $\phi(x, M) = s(M)x$ with $s(M)$ uniformly bounded on $\mathcal{U}$, we have $|f_\star(X)| \leqslant C\|X\|$. Moreover,

$$\Pi_M f_\star(x) = \phi(x,M)^\top \Sigma(M)^+ \,\mathbb{E}[\phi(X,M)f_\star(X)].$$

The vector $\Sigma(M)^+\mathbb{E}[\phi(X,M)f_\star(X)]$ is uniformly bounded over $M \in \mathcal{U}$ by Lemma H.4, Lemma H.3, and Cauchy–Schwarz. Therefore $|\Pi_M f_\star(X)| \leqslant C\|X\|$ uniformly over $M \in \mathcal{U}$, and hence $|e_M(X)| \leqslant C\|X\|$ uniformly over $M \in \mathcal{U}$. Combining this bound with $q_M(X) \leqslant C\|X\|^2$ gives

$$e_M(X)^2 q_M(X)^{1+\eta} \leqslant C\|X\|^{4+2\eta}.$$

Taking expectations and using $2\eta \leqslant \delta$ proves the claim. ■

H.3.4. WELL-POSEDNESS OF THE EMPIRICAL PROJECTOR

Recall that $\mathcal{H}_M = \{x \mapsto \theta^\top \phi(x, M) : \theta \in \mathbb{R}^k\}$ is a k-dimensional linear class. The well-posedness assumption of Appendix G.3 is equivalent to requiring that the empirical inner product is non-degenerate on $\mathcal{H}_M$, or equivalently that the $k \times k$ empirical covariance $\Sigma_n(M)$ has full rank.

Lemma H.6 (Well-posedness holds almost surely for nondegenerate designs (Assumption G.1)) *Assume μ_{down} is nondegenerate in the sense that X has a density on $\mathbb{R}^d$ and $\Sigma_{\mathrm{down}} \succ 0$. Fix any $M \in \mathcal{U}$ and any $n \geqslant k$. Then, with probability one over $X_{1:n}$,*

$$\mathrm{rank}\Big(\Sigma_n(M)\Big) = k, \qquad \Sigma_n(M) := \frac{1}{n}\sum_{i=1}^n \phi(X_i,M)\phi(X_i,M)^\top.$$

Consequently, the empirical projector $\Pi_{M,n}$ acts as the identity on $\mathcal{H}_M$.

Proof Write $\Phi \in \mathbb{R}^{n\times k}$ for the design matrix with rows $\phi(X_i, M)^\top$. Then $\Sigma_n(M) = \frac{1}{n}\Phi^\top\Phi$ has rank k if and only if Φ has rank k. Since $\phi(X_i, M) = s(M)X_i$ and $s(M)$ has rank k, the random vector $\phi(X_i, M)$ has a density on $\mathbb{R}^k$ (because X_i has a density on $\mathbb{R}^d$ and $s(M)$ is a surjective linear map $\mathbb{R}^d \to \mathbb{R}^k$). For i.i.d. vectors in $\mathbb{R}^k$ with a density, the event that k of them fall into a common proper hyperplane has probability 0, so Φ has rank k almost surely when $n \geqslant k$. The final claim is exactly the well-posedness implication used in Appendix G.3. ■

Since M_m is independent of the downstream sample, Lemma H.6 applies conditionally on M_m and yields the well-posedness requirement along the triangular array (M_m, n) used in the master theorem.

Next, we show that the moment condition $\mathbb{E}[\|X\|^4] < \infty$ is in fact sufficient for Assumption G.1.

Proposition H.7 (Oliveira (2016)) *Fix a $\delta \in (0,1)$ and suppose $\|x\|_{L^4} < \infty$. Define the constants:*

$$C_X \coloneqq \sup_{v \in \mathbb{S}^{d-1}} \sqrt{\mathbb{E}[\langle (\Sigma^+)^{1/2} x, v\rangle^4]}, \quad k \coloneqq \operatorname{rank}(\Sigma). \tag{H.5}$$

Suppose that $n \geqslant c_0 C_X^2 (k + \log(1/\delta))$ for a universal c_0. Then with probability at least $1-\delta$:

$$\Sigma_n \succcurlyeq \frac{1}{4}\Sigma.$$

On this event, we also have that $\operatorname{Col}(\Sigma_n) = \operatorname{Col}(\Sigma)$.

Proof The first part of the claim, that $\Sigma_n \succcurlyeq \frac{1}{4}\Sigma$, is immediate from Oliveira (2016, Theorem 3.1). To finish, suppose that $\Sigma_n \succcurlyeq \frac{1}{4}\Sigma$ holds. Now, let $q \in \operatorname{Kern}(\Sigma_n)$. By the above, this implies that

$$0 = q^\mathsf{T} \Sigma_n q \geqslant \frac{1}{4} q^\mathsf{T} \Sigma q \geqslant 0,$$

and hence $q^\mathsf{T}\Sigma q = 0$, which implies $q \in \operatorname{Kern}(\Sigma^{1/2}) = \operatorname{Kern}(\Sigma)$. Therefore, $\operatorname{Kern}(\Sigma_n) \subseteq \operatorname{Kern}(\Sigma)$, which by Lemma G.15 implies $\operatorname{Col}(\Sigma_n) = \operatorname{Col}(\Sigma)$. ■

H.3.5. STABLE LIMITING SPAN OF NULL DIRECTIONS

Lemma H.8 (Stable limiting span of null directions) *Assume $\Sigma_{\text{down}} \succ 0$. Then the stable limiting span condition in Assumption G.4 holds with limiting null-residual span equal to $\{0\}$.*

Proof For every $M \in \mathcal{U}$, Lemma H.4 gives $\Sigma(M) \succ 0$. Recall that $(T_M\theta)(x) = \theta^\top \phi(x, M)$. Thus, for any $\theta \neq 0$, we have

$$\|T_M\theta\|_{L^2}^2 = \mathbb{E}\big[(\theta^\top \phi(X, M))^2\big] = \theta^\top \Sigma(M)\theta > 0.$$

Equivalently, the map $T_M : \mathbb{R}^k \to L^2(\mu_{\text{down}})$ is injective. Thus the population null space $N_M = \ker(T_M)$ is $\{0\}$ for every $M \in \mathcal{U}$, and in particular $N_\star = \{0\}$. The residual subspace generated by perturbing null directions is therefore identically zero:

$$\mathcal{B}_v = \operatorname{Im}\Big((I - \Pi_{\mathcal{A}_v}) T_v\big|_{N_\star}\Big) = \{0\}.$$

Hence the required limiting span exists and equals $\{0\}$. ■

H.4. Pre-training consistency and manifold CLT for the linear spectral loss

This subsection verifies Assumption D.5 for the linear spectral descriptor estimator $\hat{M}_m$.

Model and loss. A pre-training observation is $z = (x, x^+, x^-) \in (\mathbb{R}^d)^3$, where (x, x^+) is a positive pair and x^- is an independent negative: x^- is an independent copy of x, independent of (x, x^+). Assume $\mathbb{E}[x] = 0$ and define

$$\Sigma_{\text{pre}} := \mathbb{E}[xx^\top], \qquad \Sigma^+_{\text{pre}} := \mathbb{E}[x^+(x)^\top].$$

For $M \in \mathcal{M}_{d,k}$, recall the per-sample loss (well-defined for all symmetric M)

$$\ell_{\text{spec}}(M; z) = -2\, x^\top M x^+ + (x^\top M x^-)^2.$$

We minimize the empirical loss over the rank-k PSD manifold $\mathcal{M}_{d,k} = \{M \succcurlyeq 0 : \text{rank}(M) = k\}$:

$$\hat{L}_m(M) := \frac{1}{m}\sum_{j=1}^m \ell_{\text{spec}}(M; z_j), \qquad \hat{M}_m \in \operatorname*{arg\,min}_{M \in \mathcal{M}_{d,k}} \hat{L}_m(M).$$

Let $L(M) := \mathbb{E}[\ell_{\text{spec}}(M; z)]$ be the population loss.

Moment assumption (for LLN and CLT of derivatives). Assume there exists $\delta > 0$ such that

$$\mathbb{E}\|x\|^{8+\delta} < \infty, \qquad \mathbb{E}\big[\|x\|^{4+\delta}\|x^+\|^{4+\delta}\big] < \infty. \tag{H.6}$$

This ensures integrability of the score and Hessian random fields used below.

Let $\text{sym}(A) := (A + A^\top)/2$. Viewing $\ell_{\text{spec}}(\cdot; z)$ as a function on $\mathcal{M}_{d,k}$ with the Frobenius inner product, its Euclidean gradient is

$$\nabla_M \ell_{\text{spec}}(M; z) = -2\ \text{sym}(x(x^+)^\top) + 2\,(x^\top M x^-)\ \text{sym}(x(x^-)^\top). \tag{H.7}$$

Its Euclidean Hessian is the linear map $H \mapsto D(\nabla_M \ell_{\text{spec}})(M; z)[H]$ given by

$$D(\nabla_M \ell_{\text{spec}})(M; z)[H] = 2\,(x^\top H x^-)\ \text{sym}(x(x^-)^\top). \tag{H.8}$$

Taking expectations and using x^- independent of (x, x^+) with $\mathbb{E}[x^-(x^-)^\top] = \Sigma_{\text{pre}}$, we obtain

$$\nabla L(M) = -2\,\Sigma^+_{\text{pre}} + 2\,\Sigma_{\text{pre}}\, M\, \Sigma_{\text{pre}}, \tag{H.9}$$

$$D(\nabla L)(M)[H] = 2\,\Sigma_{\text{pre}}\, H\, \Sigma_{\text{pre}}. \tag{H.10}$$

Let

$$\varphi(z) := \text{Proj}_{T_{M_\star}\mathcal{M}_{d,k}}\big(\nabla_M \ell_{\text{spec}}(M_\star; z)\big) \in T_{M_\star}\mathcal{M}_{d,k}, \qquad \Sigma_\star := \text{Cov}(\varphi(z)). \tag{H.11}$$

Also define

$$H_\star := \text{Hess}\, L(M_\star) : T_{M_\star}\mathcal{M}_{d,k} \to T_{M_\star}\mathcal{M}_{d,k}. \tag{H.12}$$

Lemma H.9 (Verification of Assumption D.5 for the linear spectral loss) *Assume $\Sigma_{\text{pre}} \succ 0$, Assumption 6.1, and (H.6). Then the empirical objective*

$$\hat{L}_m(M) = \frac{1}{m}\sum_{j=1}^m \ell_{\text{spec}}(M; z_j)$$

satisfies Assumption D.5 on $\mathcal{M}_{d,k}$ at $M_\star$.

Proof We verify the five parts of Assumption D.5.

(i) Identification and separation. By Lemma H.1, the population objective satisfies

$$L(M) = \left\| \Sigma_{\mathrm{pre}}^{1/2} M \Sigma_{\mathrm{pre}}^{1/2} - C \right\|_F^2 - \|C\|_F^2.$$

Under Assumption 6.1, the rank-k positive truncation of C is unique. Hence the population minimizer $M_\star \in \mathcal{M}_{d,k}$ is unique. Moreover, since $\Sigma_{\mathrm{pre}} \succ 0$, the objective is coercive on $\mathcal{M}_{d,k}$, that is, $L(M) \to \infty$ whenever $\|M\|_F \to \infty$ within $\mathcal{M}_{d,k}$. Therefore, all sufficiently low sublevel sets are compact after closure. If the separation condition failed for some $\epsilon > 0$, then there would exist $M_j \in \mathcal{M}_{d,k}$ such that

$$d_{\mathcal{M}}(M_j, M_\star) \geqslant \epsilon, \qquad L(M_j) \downarrow L(M_\star).$$

By coercivity, a subsequence is bounded and has a limit point in the closure of the rank-k PSD stratum. The continuity of the squared-distance objective would make this limit point another minimizer, contradicting the uniqueness of the rank-k positive truncation under Assumption 6.1. Hence, for every $\epsilon > 0$,

$$\inf_{M \in \mathcal{M}_{d,k} : d_{\mathcal{M}}(M, M_\star) \geqslant \epsilon} \big(L(M) - L(M_\star)\big) > 0.$$

(ii) Uniform LLN on a compact set and localization. Let $U = \exp_{M_\star}(B(M_\star, \epsilon_0))$ be a normal neighborhood of $M_\star$, and fix $\epsilon' \in (0, \epsilon_0)$. Set

$$K_{\epsilon'} := \exp_{M_\star}\left(\overline{B}(M_\star, \epsilon')\right).$$

By Newey and McFadden (1994, Lemma 2.4), it is enough to verify that the class

$$\mathcal{F} := \{\ell_{\mathrm{spec}}(M; \cdot) : M \in K_{\epsilon'}\}$$

is pointwise continuous in M and dominated by an integrable envelope. The continuity is immediate because $M \mapsto \ell_{\mathrm{spec}}(M; z)$ is a polynomial for every fixed z. Since $K_{\epsilon'}$ is compact, there is $R_K < \infty$ such that $\sup_{M \in K_{\epsilon'}} \|M\|_{\mathrm{op}} \leqslant R_K$. Therefore, for all $M \in K_{\epsilon'}$,

$$|\ell_{\mathrm{spec}}(M; z)| \leqslant 2R_K \|x\| \, \|x^+\| + R_K^2 \|x\|^2 \|x^-\|^2.$$

The right-hand side is integrable under (H.6), using that x^- is an independent copy of x. Hence

$$\sup_{M \in K_{\epsilon'}} |\hat{L}_m(M) - L(M)| \xrightarrow{\mathbb{P}} 0. \tag{H.13}$$

Equation H.13 with the separation in part (i), gives the argmin consistency:

$$\hat{M}_m \xrightarrow{\mathbb{P}} M_\star.$$

Consequently, after fixing $\epsilon' \in (0, \epsilon_0)$,

$$\mathbb{P}\Big(\hat{M}_m \in \exp_{M_\star}\left(\overline{B}(M_\star, \epsilon')\right)\Big) \to 1.$$

(iii) Local C^2 smoothness and score moments. For every $z = (x, x^+, x^-)$, the map $M \mapsto \ell_{\mathrm{spec}}(M; z)$ is a polynomial in M. Hence, in any normal coordinate chart around $M_\star$, it is C^2. The first two ambient derivatives are given by (H.7) and (H.8).

Since the Riemannian gradient is the tangent projection of the Euclidean gradient,

$$\|\operatorname{grad} \ell_{\mathrm{spec}}(M_\star; z)\|_{M_\star} \leqslant \|\nabla_M \ell_{\mathrm{spec}}(M_\star; z)\|_F.$$

Using (H.7),

$$\|\nabla_M \ell_{\mathrm{spec}}(M_\star; z)\|_F \leqslant C\Big(\|x\|\|x^+\| + \|x\|^2\|x^-\|^2\Big),$$

where C depends only on $M_\star$. The moment condition (H.6), together with independence of x^-, implies

$$\mathbb{E}\big[\|\operatorname{grad} \ell_{\mathrm{spec}}(M_\star; Z)\|_{M_\star}^2\big] < \infty.$$

Thus, the score moment condition in Assumption D.5(iii) holds.

(iv) Nondegenerate minimizer. By (H.3), the population objective is

$$L_{\mathrm{spec}}(w) = \big\|\Sigma_{\mathrm{pre}}^{1/2} M \Sigma_{\mathrm{pre}}^{1/2} - C\big\|_F^2 - \|C\|_F^2,$$

restricted to $\mathcal{M}_{d,k}$. The map $M \mapsto \Sigma_{\mathrm{pre}}^{1/2} M \Sigma_{\mathrm{pre}}^{1/2}$ is a local diffeomorphism on $\mathcal{M}_{d,k}$. Under Assumption 6.1, the rank-k positive truncation is separated by an eigengap. Therefore the restricted squared-distance objective has a nondegenerate strict local minimum at $M_\star$. Equivalently, the Riemannian Hessian

$$H_\star = \operatorname{Hess} L(M_\star) : T_{M_\star}\mathcal{M}_{d,k} \to T_{M_\star}\mathcal{M}_{d,k}$$

is positive definite, and in particular invertible.

(v) Uniform transported Hessian convergence. Let

$$K_{\epsilon'} = \exp_{M_\star}\big(\overline{B}(M_\star, \epsilon')\big)$$

as above. In normal coordinates, the pulled-back loss $v \mapsto \ell_{\mathrm{spec}}(\exp_{M_\star}(v); z)$ has second derivatives that are continuous in v. Since $K_{\epsilon'}$ is compact and the exponential map and its derivatives are bounded on $\overline{B}(M_\star, \epsilon')$, these second derivatives admit an integrable envelope of the form

$$C_K\Big(\|x\|\|x^+\| + \|x\|^2\|x^-\|^2\Big),$$

again by (H.7)–(H.8). Therefore, by Newey and McFadden (1994, Lemma 2.4),

$$\sup_{M \in K_{\epsilon'}} \big\|\widetilde{H}_m(M) - \widetilde{H}(M)\big\| \xrightarrow{\mathbb{P}} 0,$$

where

$$\widetilde{H}_m(M) = \mathcal{P}_{M \to M_\star} \circ \operatorname{Hess} \hat{L}_m(M) \circ \mathcal{P}_{M_\star \to M}, \qquad \widetilde{H}(M) = \mathcal{P}_{M \to M_\star} \circ \operatorname{Hess} L(M) \circ \mathcal{P}_{M_\star \to M}.$$

This is precisely Assumption D.5(v).

Combining parts (i)–(v), Assumption D.5 holds. ∎

Proposition H.10 (Descriptor CLT for the linear spectral estimator) *Assume $\Sigma_{\mathrm{pre}} \succ 0$, Assumption 6.1, and (H.6). Let $\hat{M}_m$ be a measurable empirical minimizer of $\hat{L}_m$ over $\mathcal{M}_{d,k}$. Then*

$$\sqrt{m}\,\log_{M_\star}(\hat{M}_m) \overset{d}{\rightsquigarrow} Z, \qquad Z \sim \mathcal{N}(0, V), \quad V := H_\star^{-1}\Sigma_\star H_\star^{-1},$$

as a random element in $T_{M_\star}\mathcal{M}_{d,k}$.

Proof By Lemma H.9, Assumption D.5 holds for the linear spectral empirical objective on $\mathcal{M}_{d,k}$. The abstract Riemannian M-estimation CLT therefore gives

$$\sqrt{m}\,\log_{M_\star}(\hat{M}_m) \overset{d}{\rightsquigarrow} \mathcal{N}(0, H_\star^{-1}\Sigma_\star H_\star^{-1}),$$

where $\Sigma_\star = \mathrm{Cov}(\varphi(Z))$ and $H_\star$ is the Riemannian Hessian of L at $M_\star$. ■

H.5. Proof of Proposition 6.2

Proof [Proof of Proposition 6.2] In Lemmas H.3–H.8, we proved that all non-CLT assumptions of Theorem 5.1 hold for the quotient feature map $\phi(x, M) = s(M)x$. By Proposition H.10, the pre-training CLT assumption D.5 holds for the descriptor estimator $\hat{M}_m$. Therefore, all assumptions of Theorem 5.1 are satisfied. ■

H.6. Explicit calculations for a concrete example

In this section, we first derive an asymptotic distribution for the spectral pre-training estimator in a general Gaussian model by computing the closed-form characterization of the limiting covariance operator $V_\star = H_\star^{-1}\Sigma_\star H_\star^{-1}$ in Proposition H.10. We then specialize the result to the concrete diagonal example presented in Section 6.1.

Fully Gaussian assumption. Assume (X, X^+) is jointly Gaussian, and X^- is an independent copy of X, independent of (X, X^+). In particular,

$$\mathbb{E}[XX^\top] = \Sigma_{\mathrm{pre}}, \qquad \mathbb{E}[X(X^+)^\top] = \Sigma_{\mathrm{pre}}^+.$$

Bilinear form for the score covariance. Let $\mathcal{S}^k$ be the space of d by d symmetric matrices. For a symmetric direction $H \in \mathcal{S}^d$, define the scalar score functional at $M_\star$ by

$$S_H(Z) := \langle \nabla_M \ell_{\mathrm{spec}}(M_\star; Z), H\rangle = -2\,X^\top H X^+ + 2\,(X^\top M_\star X^-)\,(X^\top H X^-).$$

where $Z = (X, X^+, X^-)$. Recall that $\Sigma_\star = \mathrm{Cov}(\varphi(Z))$ with $\varphi(Z) = \mathrm{Proj}_T(\nabla_M \ell_{\mathrm{spec}}(M_\star; Z))$ where Proj_T is the orthogonal projection onto $T_{M_\star}\mathcal{M}_{d,k}$. We know that Proj_T is self-adjoint and $\mathbb{E}[\varphi(Z)] = 0$. Thus, for all $v_1, v_2 \in T_{M_\star}\mathcal{M}_{d,k}$ we have

$$\langle v_1, \Sigma_\star v_2\rangle = \mathrm{Cov}\big(S_{v_1}(Z), S_{v_2}(Z)\big). \tag{H.14}$$

The following proposition gives $\mathrm{Cov}(S_{H_1}, S_{H_2})$ in closed form under joint Gaussianity.

Proposition H.11 (Exact score covariance under joint Gaussianity) *Under the fully Gaussian assumption, for any $H_1, H_2 \in \mathcal{S}^d$,*

$$\mathrm{Cov}\big(S_{H_1}(Z), S_{H_2}(Z)\big) = 4\Big(C_{aa}(H_1, H_2) - C_{ab}(H_1, H_2) - C_{ab}(H_2, H_1) + C_{bb}(H_1, H_2)\Big), \tag{H.15}$$

where the terms are given as follows:

(i) Positive-pair term:

$$C_{aa}(H_1, H_2) \coloneqq \mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}}) + \mathrm{tr}(H_1\Sigma_{\mathrm{pre}}^{+}H_2\Sigma_{\mathrm{pre}}^{+}). \tag{H.16}$$

(ii) Cross term: Let $P_i \coloneqq M_\star\Sigma_{\mathrm{pre}}H_i$. Then

$$C_{ab}(H_1, H_2) \coloneqq \mathrm{tr}(P_2\Sigma_{\mathrm{pre}}H_1\Sigma_{\mathrm{pre}}^{+}) + \mathrm{tr}(P_2\Sigma_{\mathrm{pre}}^{+}H_1\Sigma_{\mathrm{pre}}). \tag{H.17}$$

(iii) Negative-sample term: Let $Q \coloneqq M_\star\Sigma_{\mathrm{pre}}M_\star$. Then

$$\begin{aligned} C_{bb}(H_1, H_2) = {} & \mathrm{tr}(P_1\Sigma_{\mathrm{pre}})\,\mathrm{tr}(P_2\Sigma_{\mathrm{pre}}) + 2\,\mathrm{tr}(P_1\Sigma_{\mathrm{pre}}P_2\Sigma_{\mathrm{pre}}) + \mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}})\,\mathrm{tr}(Q\Sigma_{\mathrm{pre}}) \\ & + 4\,\mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}}Q\Sigma_{\mathrm{pre}}). \end{aligned} \tag{H.18}$$

We will repeatedly use the following standard lemmas in the proof.

Lemma H.12 (Bilinear–bilinear moment) *Let (U, V) be jointly Gaussian with $\mathbb{E}[U] = \mathbb{E}[V] = 0$, $\mathbb{E}[UU^\top] = \Sigma_U$, $\mathbb{E}[VV^\top] = \Sigma_V$, and $\mathbb{E}[UV^\top] = \Sigma_{UV}$. Then for any $A, B \in \mathbb{R}^{d\times d}$,*

$$\mathbb{E}\big[(U^\top AV)(U^\top BV)\big] = \mathrm{tr}(A\Sigma_{UV})\,\mathrm{tr}(B\Sigma_{UV}) + \mathrm{tr}(A\Sigma_U B\Sigma_V) + \mathrm{tr}(A\Sigma_{UV}B^\top\Sigma_{UV}^\top). \tag{H.19}$$

Consequently,

$$\mathrm{Cov}(U^\top AV, U^\top BV) = \mathrm{tr}(A\Sigma_U B\Sigma_V) + \mathrm{tr}(A\Sigma_{UV}B^\top\Sigma_{UV}^\top). \tag{H.20}$$

Lemma H.13 (Quadratic–quadratic moment) *Let $U \sim \mathcal{N}(0, \Sigma)$ and let $A, B \in \mathbb{R}^{d\times d}$. Then*

$$\mathbb{E}\big[(U^\top AU)(U^\top BU)\big] = \mathrm{tr}(A\Sigma)\,\mathrm{tr}(B\Sigma) + \mathrm{tr}(A\Sigma B\Sigma) + \mathrm{tr}(A\Sigma B^\top\Sigma). \tag{H.21}$$

Consequently,

$$\mathrm{Cov}(U^\top AU, U^\top BU) = \mathrm{tr}(A\Sigma B\Sigma) + \mathrm{tr}(A\Sigma B^\top\Sigma). \tag{H.22}$$

Proof [Proof of Lemmas H.12–H.13] Both identities follow by expanding the products componentwise and applying Isserlis' formula to fourth moments. For instance, for Lemma H.12,

$$\mathbb{E}[(U^\top AV)(U^\top BV)] = \sum_{i,j,p,q} A_{ij}B_{pq}\,\mathbb{E}[U_iV_jU_pV_q],$$

and Isserlis gives $\mathbb{E}[U_iV_jU_pV_q] = \mathbb{E}[U_iV_j]\mathbb{E}[U_pV_q] + \mathbb{E}[U_iU_p]\mathbb{E}[V_jV_q] + \mathbb{E}[U_iV_q]\mathbb{E}[V_jU_p]$, which sums to (H.19). Lemma H.13 is analogous. ■

Proof [Proof of Proposition H.11] Write $S_H(Z) = -2a_H(Z) + 2b_H(Z)$ with

$$a_H(Z) \coloneqq X^\top H X^+, \qquad b_H(Z) \coloneqq (X^\top M_\star X^-)(X^\top H X^-).$$

Then, for any $H_1, H_2 \in \mathcal{S}^d$,

$$\mathrm{Cov}(S_{H_1}(Z), S_{H_2}(Z)) = 4\Big(\mathrm{Cov}(a_{H_1}, a_{H_2}) - \mathrm{Cov}(a_{H_1}, b_{H_2}) - \mathrm{Cov}(a_{H_2}, b_{H_1}) + \mathrm{Cov}(b_{H_1}, b_{H_2})\Big). \tag{H.23}$$

We compute the four covariance terms under joint Gaussianity using Isserlis' formula.

First, we will compute $\mathrm{Cov}(a_{H_1}, a_{H_2})$. Apply Lemma H.12 with $(U, V) = (X, X^+)$, $\Sigma_U = \Sigma_V = \Sigma_{\mathrm{pre}}$ and $\Sigma_{UV} = \Sigma_{\mathrm{pre}}^+$. Since H_1, H_2 are symmetric and Σ_{pre}^+ is symmetric, (H.20) yields

$$C_{aa}(H_1, H_2) = \mathrm{Cov}(a_{H_1}, a_{H_2}) = \mathrm{tr}(H_1 \Sigma_{\mathrm{pre}} H_2 \Sigma_{\mathrm{pre}}) + \mathrm{tr}(H_1 \Sigma_{\mathrm{pre}}^+ H_2 \Sigma_{\mathrm{pre}}^+).$$

Next, we will compute the cross term $\mathrm{Cov}(a_{H_1}, b_{H_2})$. Fix H_1, H_2 and write

$$b_{H_2} = (X^\top M_\star X^-)(X^\top H_2 X^-) = (X^-)^\top (M_\star X X^\top H_2) X^-.$$

Condition on (X, X^+) and use that X^- is independent of (X, X^+) with covariance Σ_{pre}:

$$\mathbb{E}[b_{H_2} \mid X, X^+] = \mathbb{E}\big[(X^\top M_\star X^-)(X^\top H_2 X^-) \mid X, X^+\big] = X^\top M_\star \Sigma_{\mathrm{pre}} H_2 X. \tag{H.24}$$

Hence,

$$\mathbb{E}[a_{H_1} b_{H_2}] = \mathbb{E}\Big[(X^\top H_1 X^+)\, \mathbb{E}[b_{H_2} \mid X, X^+]\Big] = \mathbb{E}\Big[(X^\top H_1 X^+)\, (X^\top P_2 X)\Big], \tag{H.25}$$

where $B_2 \coloneqq M_\star \Sigma_{\mathrm{pre}} H_2$.

We next compute the mixed moment in (H.25). Expand

$$(X^\top H_1 X^+)(X^\top P_2 X) = \sum_{i,j,p,q} (H_1)_{ij} (P_2)_{pq}\, X_i X_j^+ X_p X_q.$$

Apply Isserlis to $\mathbb{E}[X_i X_j^+ X_p X_q]$ for the jointly Gaussian pair (X, X^+):

$$\mathbb{E}[X_i X_j^+ X_p X_q] = \mathbb{E}[X_i X_j^+]\mathbb{E}[X_p X_q] + \mathbb{E}[X_i X_p]\mathbb{E}[X_j^+ X_q] + \mathbb{E}[X_i X_q]\mathbb{E}[X_j^+ X_p].$$

Using $\mathbb{E}[X_i X_j^+] = (\Sigma_{\mathrm{pre}}^+)_{ij}$ and $\mathbb{E}[X_p X_q] = (\Sigma_{\mathrm{pre}})_{pq}$, this yields

$$\mathbb{E}\big[(X^\top H_1 X^+)(X^\top P_2 X)\big] = \mathrm{tr}(H_1 \Sigma_{\mathrm{pre}}^+)\, \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}}) + \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}} H_1 \Sigma_{\mathrm{pre}}^+) + \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}}^+ H_1 \Sigma_{\mathrm{pre}}). \tag{H.26}$$

Moreover,

$$\mathbb{E}[a_{H_1}] = \mathrm{tr}(H_1 \Sigma_{\mathrm{pre}}^+), \qquad \mathbb{E}[b_{H_2}] = \mathbb{E}[X^\top P_2 X] = \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}}).$$

Therefore, subtracting $\mathbb{E}[a_{H_1}]\mathbb{E}[b_{H_2}]$ from (H.26) gives

$$C_{ab}(H_1, H_2) = \mathrm{Cov}(a_{H_1}, b_{H_2}) = \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}} H_1 \Sigma_{\mathrm{pre}}^+) + \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}}^+ H_1 \Sigma_{\mathrm{pre}}).$$

Finally, we will compute the negative-sample term $\mathrm{Cov}(b_{H_1}, b_{H_2})$. Write

$$b_{H_i} = (X^\top M_\star X^-)(X^\top H_i X^-) = (X^-)^\top A_i(X) X^-, \qquad A_i(X) \coloneqq M_\star X X^\top H_i.$$

From the total law of covariance, we have

$$C_{bb}(H_1, H_2) = \mathrm{Cov}(b_{H_1}, b_{H_2}) = \mathbb{E}[\mathrm{Cov}(b_{H_1}, b_{H_2}|X)] + \mathrm{Cov}(\mathbb{E}[b_{H_1}|X], \mathbb{E}[b_{H_2}|X]).$$

We apply Lemma H.13 to $X^- \sim \mathcal{N}(0, \Sigma_{\mathrm{pre}})$:

$$\begin{aligned}
\mathrm{Cov}(b_{H_1}, b_{H_2} \mid X) &= \mathrm{tr}(A_1(X)\Sigma_{\mathrm{pre}} A_2(X)\Sigma_{\mathrm{pre}}) + \mathrm{tr}(A_1(X)\Sigma_{\mathrm{pre}} A_2(X)^T \Sigma_{\mathrm{pre}}) \\
&= (X^\mathsf{T} M_\star \Sigma_{\mathrm{pre}} H_1 X X^\mathsf{T} M_\star \Sigma_{\mathrm{pre}} H_2 X) + (X^\mathsf{T} H_1 \Sigma_{\mathrm{pre}} H_2 X X^\mathsf{T} M_\star \Sigma_{\mathrm{pre}} M_\star X) \\
&= (X^\mathsf{T} P_1 X X^\mathsf{T} P_2 X) + (X^\mathsf{T} H_1 \Sigma_{\mathrm{pre}} H_2 X X^\mathsf{T} Q X).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\mathrm{Cov}(b_{H_1}, b_{H_2}|X)] = \mathrm{tr}(P_1\Sigma_{\mathrm{pre}})\,\mathrm{tr}(P_2\Sigma_{\mathrm{pre}}) + \mathrm{tr}(P_1\Sigma_{\mathrm{pre}}P_2\Sigma_{\mathrm{pre}}) + \mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}})\,\mathrm{tr}(Q\Sigma_{\mathrm{pre}}) \\
+ 3\,\mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}}Q\Sigma_{\mathrm{pre}}).
\end{aligned}$$

Similarly,

$$\mathrm{Cov}(\mathbb{E}[b_{H_1}|X], \mathbb{E}[b_{H_2}|X]) = \mathrm{Cov}(X^\mathsf{T} P_1 X, X^\mathsf{T} P_2 X) = \mathrm{tr}(P_1\Sigma_{\mathrm{pre}}P_2\Sigma_{\mathrm{pre}}) + \mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}}Q\Sigma_{\mathrm{pre}}).$$

Therefore, we get

$$\begin{aligned}
C_{bb}(H_1, H_2) = \mathrm{tr}(P_1\Sigma_{\mathrm{pre}})\,\mathrm{tr}(P_2\Sigma_{\mathrm{pre}}) + 2\,\mathrm{tr}(P_1\Sigma_{\mathrm{pre}}P_2\Sigma_{\mathrm{pre}}) + \mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}})\,\mathrm{tr}(Q\Sigma_{\mathrm{pre}}) \\
+ 4\,\mathrm{tr}(H_1\Sigma_{\mathrm{pre}}H_2\Sigma_{\mathrm{pre}}Q\Sigma_{\mathrm{pre}}).
\end{aligned}$$

■

Corollary H.14 *Under the fully Gaussian assumption, for any $H_1, H_2 \in \mathcal{S}^d$, we have*

$$\langle H_1, V_\star H_2 \rangle = C_{aa}(H_1', H_2') - C_{ab}(H_1', H_2') - C_{ab}(H_2', H_1') + C_{bb}(H_1', H_2') \tag{H.27}$$

where $H_i' = \Sigma_{\mathrm{pre}}^{-1} H_i \Sigma_{\mathrm{pre}}^{-1}$ for $i = \{1, 2\}$.

Proof For any $v \in \mathcal{S}^d$, we have

$$H_\star v = 2\Sigma_{\mathrm{pre}} v \Sigma_{\mathrm{pre}}, \qquad H_\star^{-1} v = \frac{1}{2}\Sigma_{\mathrm{pre}}^{-1} v \Sigma_{\mathrm{pre}}^{-1}.$$

Therefore, we have

$$\begin{aligned}
\langle H_1, V_\star H_2 \rangle &= \langle H_1, H_\star^{-1}\Sigma_\star H_\star^{-1} H_2 \rangle = \langle H_\star^{-1} H_1, \Sigma_\star H_\star^{-1} H_2 \rangle \\
&= \frac{1}{4}\langle \Sigma_{\mathrm{pre}}^{-1} H_1 \Sigma_{\mathrm{pre}}^{-1}, \Sigma_\star H_\star^{-1}\Sigma_{\mathrm{pre}}^{-1} H_2 \Sigma_{\mathrm{pre}}^{-1} \rangle = \frac{1}{4}\langle H_1', \Sigma_\star H_2' \rangle = \frac{1}{4}\,\mathrm{Cov}\big(S_{H_1'}(Z), S_{H_2'}(Z)\big).
\end{aligned}$$

Equation (H.27) follows after plugging this equation into Equation (H.15). ■

Linearization of the representation error. Fix M in a neighborhood of $M_\star$ and let $A := s(M) \in \mathbb{R}^{k\times d}$ be the local section so that $A^\top A = M$. Consider the operator $T_M : \mathbb{R}^k \to L^2(\mu_{\text{down}})$ defined by

$$(T_M b)(x) := \langle b, Ax\rangle = b^\top Ax, \qquad b \in \mathbb{R}^k.$$

Its adjoint $T_M^{\text{adj}} : L^2(\mu_{\text{down}}) \to \mathbb{R}^k$ satisfies, for any $g \in L^2(\mu_{\text{down}})$,

$$T_M^{\text{adj}} g = \mathbb{E}\big[g(X)\, AX\big] = A\,\mathbb{E}\big[g(X)\, X\big],$$

where $X \sim \mu_{\text{down}}$ and we used linearity of A.

Let

$$\Sigma_{\text{down}} := \mathbb{E}[XX^\top], \qquad \Sigma(M) := T_M^{\text{adj}} T_M = \mathbb{E}[(AX)(AX)^\top] = A\Sigma_{\text{down}}A^\top.$$

The orthogonal projector onto $\text{Range}(T_M)$ is given by

$$\Pi_M = T_M\, \Sigma(M)^+\, T_M^{\text{adj}}.$$

Therefore, for any $g \in L^2(\mu_{\text{down}})$,

$$\begin{aligned}(\Pi_M g)(x) &= \big(T_M \Sigma(M)^+ T_M^{\text{adj}} g\big)(x) = \Big\langle \Sigma(M)^+\, T_M^{\text{adj}} g,\ Ax\Big\rangle \\ &= \big\langle \Sigma(M)^+\, A\,\mathbb{E}[g(X)X],\ Ax\big\rangle = x^\top A^\top \Sigma(M)^+ A\,\mathbb{E}[g(X)X]. \end{aligned} \tag{H.28}$$

Define the induced matrix

$$P(M) := A^\top \Sigma(M)^+ A \in \mathbb{R}^{d\times d}.$$

Then (H.28) becomes the compact form

$$(\Pi_M g)(x) = x^\top P(M)\, \mathbb{E}[g(X)X].$$

We know that the target function is $f_\star(x) = \beta_\star^\top A_\star x$ for some $\beta_\star \in \mathbb{R}^k$ and $A_\star := s(M_\star)$. Then

$$\mathbb{E}[f_\star(X)X] = \mathbb{E}[XX^\top]A_\star^\top \beta_\star = \Sigma_{\text{down}} A_\star^\top \beta_\star,$$

and therefore

$$(\Pi_M f_\star)(x) = x^\top P(M)\, \Sigma_{\text{down}}\, A_\star^\top \beta_\star. \tag{H.29}$$

We linearize (H.29) around $M_\star$ using Proposition G.14. In the constant-rank region, the map $\Omega \mapsto \Pi_\Omega$ is Fréchet differentiable at $\Omega_\star$, and for $v \in T_{\Omega_\star}\mathcal{M}$,

$$\Pi_{\exp_{\Omega_\star}(v)} = \Pi_{\Omega_\star} + D\Pi_{\Omega_\star}[v] + o(\|v\|).$$

Define, for any $M \in \mathcal{S}_+^d$ with $\text{rank}(M) = k$, the *whitened descriptor*

$$B(M) := \Sigma_{\text{down}}^{1/2}\, M\, \Sigma_{\text{down}}^{1/2} \in \mathcal{S}_+^d.$$

Let $\Pi(B)$ denote the Euclidean orthogonal projector onto $\text{range}(B)$. Then the induced matrix in (H.28) admits the M–only representation

$$P(M) = \Sigma_{\text{down}}^{-1/2}\, \Pi\big(B(M)\big)\, \Sigma_{\text{down}}^{-1/2} = \Sigma_{\text{down}}^{-1/2}\, B(M)\, B(M)^+\, \Sigma_{\text{down}}^{-1/2}. \tag{H.30}$$

For the fixed target $f_\star$, define $m_\star := \mathbb{E}[f_\star(X)X] \in \mathbb{R}^d$. Since $m_\star$ is independent of M, we have for any tangent direction $v \in T_{M_\star}\mathcal{M}_{d,k}$,

$$\begin{aligned} D\Pi_{M_\star}[v]\, f_\star(x) &= x^\top DP(M_\star)[v]\, m_\star, \\ (\mathcal{L}(v) f_\star)(x) &:= -D\Pi_{M_\star}[v]\, f_\star(x) = -\,x^\top DP(M_\star)[v]\, m_\star. \end{aligned} \tag{H.31}$$

Derivative of $P(M)$. Let $B_\star := B(M_\star) = \Sigma_{\text{down}}^{1/2} M_\star \Sigma_{\text{down}}^{1/2}$ and $\dot{B} := DB(M_\star)[v] = \Sigma_{\text{down}}^{1/2} v \Sigma_{\text{down}}^{1/2}$. Differentiating (H.30) yields

$$DP(M_\star)[v] = \Sigma_{\text{down}}^{-1/2}\, D\Pi(B_\star)[\dot{B}]\, \Sigma_{\text{down}}^{-1/2}. \tag{H.32}$$

Assume an eigengap at k for $B_\star$, so that $\Pi(\cdot)$ is Fréchet differentiable at $B_\star$. Let $B_\star = U \operatorname{diag}(\Lambda_1, \Lambda_2) U^\top$ with $U = [U_1, U_2]$, $\Lambda_1 = \operatorname{diag}(\lambda_1, \dots, \lambda_k)$, $\Lambda_2 = \operatorname{diag}(\lambda_{k+1}, \dots, \lambda_d)$, and $\min_{i \leqslant k < j} |\lambda_i - \lambda_j| > 0$.

Lemma H.15 *For any symmetric direction $\dot{B} \in \mathcal{S}^d$,*

$$D\Pi(B_\star)[\dot{B}] = U_2 \left(G \oslash \Delta\right) U_1^\top + U_1 \left(G^\top \oslash \Delta^\top\right) U_2^\top, \tag{H.33}$$

where

$$G := U_2^\top \dot{B}\, U_1 \in \mathbb{R}^{(d-k)\times k}, \qquad \Delta_{ai} := \lambda_i - \lambda_{k+a},$$

and $\oslash$ is the elementwise division.

Proof We work in the eigenbasis of $B_\star$, i.e. replace $\dot{B}$ by $\widetilde{\dot{B}} := U^\top \dot{B}\, U$ and write the projector in this basis. Since $\Pi(B)$ is orthogonally equivariant, it suffices to prove the claim for $B_\star = \operatorname{diag}(\Lambda_1, \Lambda_2)$, in which case $\Pi(B_\star) = P_\star := \operatorname{diag}(I_k, 0)$.

Let Γ be a positively oriented contour in the complex plane enclosing $\{\lambda_1, \dots, \lambda_k\}$ and excluding $\{\lambda_{k+1}, \dots, \lambda_d\}$. By the Riesz projector formula,

$$\Pi(B) = \frac{1}{2\pi i} \oint_\Gamma (zI - B)^{-1}\, dz \tag{H.34}$$

for all B in a neighborhood of $B_\star$. Differentiating (H.34) at $B_\star$ in direction $\dot{B}$ and using the standard resolvent identity

$$D\big((zI - B)^{-1}\big)\Big|_{B=B_\star}[\dot{B}] = (zI - B_\star)^{-1}\, \dot{B}\, (zI - B_\star)^{-1},$$

we obtain

$$D\Pi(B_\star)[\dot{B}] = \frac{1}{2\pi i} \oint_\Gamma (zI - B_\star)^{-1}\, \dot{B}\, (zI - B_\star)^{-1}\, dz. \tag{H.35}$$

Since $B_\star = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$ is diagonal in this basis, $(zI - B_\star)^{-1}$ is also diagonal with entries $(z - \lambda_j)^{-1}$. Therefore, the (p, q) entry of the integrand in (H.35) equals

$$\left[(zI - B_\star)^{-1}\, \dot{B}\, (zI - B_\star)^{-1}\right]_{pq} = \frac{\dot{B}_{pq}}{(z - \lambda_p)(z - \lambda_q)}.$$

Hence

$$\left[D\Pi(B_\star)[\dot{B}]\right]_{pq} = \dot{B}_{pq} \cdot \frac{1}{2\pi i} \oint_\Gamma \frac{dz}{(z - \lambda_p)(z - \lambda_q)}. \tag{H.36}$$

Now consider cases.

(i) $p, q \leqslant k$: If both λ_p and λ_q lie inside of the contour Γ, then we have

$$I_{pq} \coloneqq \frac{1}{2\pi i}\oint_\Gamma \frac{dz}{(z-\lambda_p)(z-\lambda_q)} = \frac{1}{2\pi i}\oint_\Gamma \frac{1}{\lambda_p-\lambda_q}(\frac{1}{z-\lambda_p}-\frac{1}{z-\lambda_q})dz$$
$$= \frac{1}{2\pi i(\lambda_p-\lambda_q)}(1-1) = 0.$$

(ii) $p, q > k$: If both λ_p and λ_q lie outside of the contour Γ, then the integrand in (H.36) is holomorphic inside Γ and the contour integral vanishes.

(iii) $p \leqslant k < q$: Then λ_p is inside Γ and λ_q is outside, so the integrand in (H.36) has a single pole inside Γ at $z = \lambda_p$ with residue $1/(\lambda_p - \lambda_q)$. Therefore

$$\frac{1}{2\pi i}\oint_\Gamma \frac{dz}{(z-\lambda_p)(z-\lambda_q)} = \frac{1}{\lambda_p-\lambda_q},$$

and thus

$$\left[D\Pi(B_\star)[\dot{B}]\right]_{pq} = \frac{\dot{B}_{pq}}{\lambda_p-\lambda_q}.$$

By symmetry, the case $q \leqslant k < p$ gives the transpose.

Writing this in block form with respect to the split $\mathbb{R}^d = \mathbb{R}^k \oplus \mathbb{R}^{d-k}$ yields

$$D\Pi(B_\star)[\dot{B}] = \begin{pmatrix} 0 & H^\top \\ H & 0 \end{pmatrix}, \qquad H_{ai} = \frac{\dot{B}_{k+a,i}}{\lambda_i - \lambda_{k+a}}.$$

Returning to the original basis (undoing the U-conjugation) gives (H.33) with $G := U_2^\top \dot{B}\, U_1$ and $\Delta_{ai} := \lambda_i - \lambda_{k+a}$. ■

pre-training fluctuations. We will derive the asymptotic distribution of $m\|\mathcal{L}(\log_{M_\star}(\hat{M}_m)\|$.

$$\begin{aligned} m\|\mathcal{L}(\log_{M_\star}(\hat{M}_m))\|^2_{L^2(\mu_{\text{down}})} &= \|x^\top DP(M_\star)[\sqrt{m}\log_{M_\star}(\hat{M}_m)]\ m_\star\|^2_{L^2(\mu_{\text{down}})} \\ &= \beta_\star^\top A_\star \Sigma_{\text{down}} DP(M_\star)[\sqrt{m}\log_{M_\star}(\hat{M}_m)]^T DP(M_\star)[\sqrt{m}\log_{M_\star}(\hat{M}_m)]\Sigma_{\text{down}} A_\star^\top \beta_\star \\ &\overset{d}{\rightsquigarrow} \beta_\star^\top A_\star \Sigma_{\text{down}} DP(M_\star)[Z]^T DP(M_\star)[Z]\Sigma_{\text{down}} A_\star^\top \beta_\star \qquad \text{(H.37)} \end{aligned}$$

where the exact form of $DP(M_\star)$ is calculated in Equation (H.32) and (H.33), and Z is a mean-zero Gaussian with the covariance $V_\star$ (Equation (H.27)).

H.7. Comparison to Cabannes et al. (2023)

We now study the following concrete example to compare our asymptotic result with the general upper bound of Cabannes et al. (2023, Thm. 4). Assume that $\Sigma_{\text{down}} = I_d$ and

$$\Sigma_{\text{pre}} = \begin{pmatrix} \Sigma_{\text{pre},1} & 0 \\ 0 & \Sigma_{\text{pre},2} \end{pmatrix}, \qquad \Sigma_{\text{pre}}^+ = \begin{pmatrix} \Sigma_{\text{pre},1}^+ & 0 \\ 0 & \Sigma_{\text{pre},2}^+ \end{pmatrix}$$

with

$$\Sigma_{\text{pre},1} = \text{diag}(a_1, \ldots, a_k), \qquad \Sigma_{\text{pre},2} = \text{diag}(b_1, \ldots, b_{d-k}),$$

and

$$\Sigma^+_{\mathrm{pre},1} = \mathrm{diag}(c_1, \dots, c_k), \qquad \Sigma^+_{\mathrm{pre},2} = \mathrm{diag}(e_1, \dots, e_{d-k}).$$

Furthermore, we assume that all diagonal entries are strictly positive and

$$\frac{c_i}{a_i} > \frac{e_j}{b_j}, \qquad \forall\, i \in \{1, \dots, k\},\ \forall\, j \in \{1, \dots, d-k\}.$$

In this case, the whitened cross-covariance matrix $C = \Sigma_{\mathrm{pre}}^{-1/2} \Sigma^+_{\mathrm{pre}} \Sigma_{\mathrm{pre}}^{-1/2}$ is diagonal and given by

$$C = \mathrm{diag}\left(\frac{c_1}{a_1}, \dots, \frac{c_k}{a_k}, \frac{e_1}{b_1}, \dots, \frac{e_{d-k}}{b_{d-k}} \right).$$

By the ordering assumption, the top-k eigenvalues of C are $\frac{c_1}{a_1}, \dots, \frac{c_k}{a_k}$, so the constrained population minimizer over $\mathcal{M}_{d,k}$ is

$$M_\star = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$$

where $R := \mathrm{diag}(r_1, \dots, r_k)$ and $r_i := \frac{c_i}{a_i^2}$. Any tangent matrix $H \in T_{M_\star}\mathcal{M}_{d,k}$ has the block form

$$H = \begin{pmatrix} A & B \\ B^\top & 0 \end{pmatrix}, \qquad A \in \mathbb{S}^k, \quad B \in \mathbb{R}^{k \times (d-k)}.$$

Let $\mu_i := \frac{e_i}{b_i}$, $\lambda_i := \frac{c_i}{a_i}$, and $\tau := \sum_{i=1}^k \lambda_i^2$. By Corollary (H.27), the covariance operator $V_\star$ is given by

$$\langle H_1, V_\star H_2 \rangle = C_{aa}(H_1', H_2') - C_{ab}(H_1', H_2') - C_{ab}(H_2', H_1') + C_{bb}(H_1', H_2'),$$

where

$$H_\ell' = \Sigma_{\mathrm{pre}}^{-1} H_\ell \Sigma_{\mathrm{pre}}^{-1} = \begin{pmatrix} \Sigma_{\mathrm{pre},1}^{-1} A_\ell \Sigma_{\mathrm{pre},1}^{-1} & \Sigma_{\mathrm{pre},1}^{-1} B_\ell \Sigma_{\mathrm{pre},2}^{-1} \\ \Sigma_{\mathrm{pre},2}^{-1} B_\ell^\top \Sigma_{\mathrm{pre},1}^{-1} & 0 \end{pmatrix}.$$

We now compute the three terms in (H.16)–(H.18). Using the block-diagonal form of Σ_{pre} and Σ^+_{pre}, a direct multiplication gives

$$H_1' \Sigma_{\mathrm{pre}} = \begin{pmatrix} \Sigma_{\mathrm{pre},1}^{-1} A_1 & \Sigma_{\mathrm{pre},1}^{-1} B_1 \\ \Sigma_{\mathrm{pre},2}^{-1} B_1^\top & 0 \end{pmatrix}, \qquad H_2' \Sigma_{\mathrm{pre}} = \begin{pmatrix} \Sigma_{\mathrm{pre},1}^{-1} A_2 & \Sigma_{\mathrm{pre},1}^{-1} B_2 \\ \Sigma_{\mathrm{pre},2}^{-1} B_2^\top & 0 \end{pmatrix}.$$

Hence

$$\mathrm{tr}(H_1' \Sigma_{\mathrm{pre}} H_2' \Sigma_{\mathrm{pre}}) = \mathrm{tr}(\Sigma_{\mathrm{pre},1}^{-1} A_1 \Sigma_{\mathrm{pre},1}^{-1} A_2) + 2\,\mathrm{tr}(\Sigma_{\mathrm{pre},1}^{-1} B_1 \Sigma_{\mathrm{pre},2}^{-1} B_2^\top).$$

Similarly,

$$\mathrm{tr}(H_1' \Sigma^+_{\mathrm{pre}} H_2' \Sigma^+_{\mathrm{pre}}) = \mathrm{tr}(\Sigma_{\mathrm{pre},1}^{-1} A_1 \Sigma^+_{\mathrm{pre},1} \Sigma_{\mathrm{pre},1}^{-1} A_2 \Sigma^+_{\mathrm{pre},1}) + 2\,\mathrm{tr}(\Sigma_{\mathrm{pre},1}^{-1} B_1 \Sigma^+_{\mathrm{pre},2} \Sigma_{\mathrm{pre},2}^{-1} B_2^\top \Sigma^+_{\mathrm{pre},1} \Sigma_{\mathrm{pre},1}^{-1}).$$

Since all these matrices are diagonal, this becomes entrywise

$$C_{aa}(H_1', H_2') = \sum_{i,j=1}^k \frac{1 + \lambda_i \lambda_j}{a_i a_j} (A_1)_{ij} (A_2)_{ij} + 2 \sum_{i=1}^k \sum_{j=1}^{d-k} \frac{1 + \lambda_i \mu_j}{a_i b_j} (B_1)_{ij} (B_2)_{ij}.$$

Since $M_\star = \begin{pmatrix} R\Sigma_{\mathrm{pre}} & 0 \\ 0 & 0 \end{pmatrix}$, we obtain

$$P_\ell = M_\star \Sigma_{\mathrm{pre}} H_\ell' = \begin{pmatrix} RA_\ell \Sigma_{\mathrm{pre},1}^{-1} & RB_\ell \Sigma_{\mathrm{pre},2}^{-1} \\ 0 & 0 \end{pmatrix}.$$

Then

$$\begin{aligned} C_{ab}(H_1', H_2') &= \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}} H_1' \Sigma_{\mathrm{pre}}^+) + \mathrm{tr}(P_2 \Sigma_{\mathrm{pre}}^+ H_1' \Sigma_{\mathrm{pre}}) \\ &= \sum_{i,j=1}^{k} \frac{r_i(\lambda_i + \lambda_j)}{a_j} (A_1)_{ij}(A_2)_{ij} + 2 \sum_{i=1}^{k} \sum_{j=1}^{d-k} \frac{\lambda_i^2}{a_i b_j} (B_1)_{ij}(B_2)_{ij}. \end{aligned}$$

By symmetry,

$$C_{ab}(H_2', H_1') = \sum_{i,j=1}^{k} \frac{r_j(\lambda_i + \lambda_j)}{a_i} (A_1)_{ij}(A_2)_{ij} + 2 \sum_{i=1}^{k} \sum_{j=1}^{d-k} \frac{\lambda_i^2}{a_i b_j} (B_1)_{ij}(B_2)_{ij}.$$

We have

$$Q := M_\star \Sigma_{\mathrm{pre}} M_\star = \begin{pmatrix} R\Sigma_{\mathrm{pre},1} R & 0 \\ 0 & 0 \end{pmatrix},$$

so that $\mathrm{tr}(Q\Sigma_{\mathrm{pre}}) = \sum_{i=1}^{k} \lambda_i^2 = \tau$. Also, $\mathrm{tr}(P_\ell \Sigma_{\mathrm{pre}}) = \sum_{i=1}^{k} r_i (A_\ell)_{ii}$. Further,

$$\mathrm{tr}(P_1 \Sigma_{\mathrm{pre}} P_2 \Sigma_{\mathrm{pre}}) = \sum_{i,j=1}^{k} r_i r_j \, (A_1)_{ij}(A_2)_{ij}.$$

Finally, using again that all matrices are diagonal in block form,

$$\mathrm{tr}(H_1' \Sigma_{\mathrm{pre}} H_2' \Sigma_{\mathrm{pre}} Q \Sigma_{\mathrm{pre}}) = \sum_{i,j=1}^{k} \frac{\lambda_i^2}{a_i a_j} (A_1)_{ij}(A_2)_{ij} + \sum_{i=1}^{k} \sum_{j=1}^{d-k} \frac{\lambda_i^2}{a_i b_j} (B_1)_{ij}(B_2)_{ij}.$$

Substituting into (H.18), we obtain

$$\begin{aligned} C_{bb}(H_1', H_2') = &\Big(\sum_{i=1}^{k} r_i (A_1)_{ii} \Big) \Big(\sum_{j=1}^{k} r_j (A_2)_{jj} \Big) \\ &+ \tau \sum_{i,j=1}^{k} \frac{1}{a_i a_j} (A_1)_{ij}(A_2)_{ij} + 2\tau \sum_{i=1}^{k} \sum_{j=1}^{d-k} \frac{1}{a_i b_j} (B_1)_{ij}(B_2)_{ij} \\ &+ 4 \sum_{i,j=1}^{k} \frac{\lambda_i^2}{a_i a_j} (A_1)_{ij}(A_2)_{ij} + 4 \sum_{i=1}^{k} \sum_{j=1}^{d-k} \frac{\lambda_i^2}{a_i b_j} (B_1)_{ij}(B_2)_{ij}. \end{aligned}$$

Substituting the previous expressions into Corollary (H.27) yields

$$\begin{aligned} \langle H_1, V_\star H_2 \rangle = &\Big(\sum_{i=1}^{k} r_i (A_1)_{ii} \Big) \Big(\sum_{j=1}^{k} r_j (A_2)_{jj} \Big) + \sum_{i=1}^{k} \frac{1 + \tau + 3\lambda_i^2}{a_i^2} (A_1)_{ii}(A_2)_{ii} \\ &+ 2 \sum_{1 \leqslant i < j \leqslant k} \frac{1 + \tau + \lambda_i^2 + \lambda_i \lambda_j + \lambda_j^2}{a_i a_j} (A_1)_{ij}(A_2)_{ij} + 2 \sum_{i=1}^{k} \sum_{j=1}^{d-k} \frac{1 + \tau + \lambda_i^2}{a_i b_j} (B_1)_{ij}(B_2)_{ij}. \end{aligned}$$

Therefore, if

$$Z = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{12}^\top & 0 \end{pmatrix} \sim \mathcal{N}(0, V_\star),$$

then the covariance structure is as follows:

(i) The diagonal entries of Z_{11} are jointly Gaussian with covariance

$$\mathrm{Diag}\left(\frac{1+\tau+3\lambda_i^2}{a_i^2}\right)_{i=1}^k + rr^\top, \qquad r = (r_1, \dots, r_k)^\top.$$

(ii) The off-diagonal entries $(Z_{11})_{ij}$, $1 \leqslant i < j \leqslant k$, are independent with variances

$$\frac{2(1+\tau+\lambda_i^2+\lambda_i\lambda_j+\lambda_j^2)}{a_i a_j}.$$

(iii) The entries of Z_{12} are independent with variances

$$\frac{2(1+\tau+\lambda_i^2)}{a_i b_j}.$$

We now specialize (H.37) to this setting. In this case, $B(M) = \Sigma_{\mathrm{down}}^{1/2} M \Sigma_{\mathrm{down}}^{1/2} = M$ so that the derivative of the spectral projector simplifies to

$$DP(M_\star)[H] = D\Pi(M_\star)[H].$$

By Equation (H.33), the Fréchet derivative of the projector at $M_\star$ in direction Z is

$$DP(M_\star)[Z] = \begin{pmatrix} 0 & R^{-1}Z_{12} \\ Z_{12}^\top R^{-1} & 0 \end{pmatrix}.$$

Consequently,

$$DP(M_\star)[Z]^\top DP(M_\star)[Z] = \begin{pmatrix} R^{-1}Z_{12}Z_{12}^\top R^{-1} & 0 \\ 0 & Z_{12}^\top R^{-2} Z_{12} \end{pmatrix}.$$

Substituting this into (H.37) yields

$$L := \beta_\star^\top A_\star \Sigma_{\mathrm{down}} DP(M_\star)[Z]^\top DP(M_\star)[Z] \Sigma_{\mathrm{down}} A_\star^\top \beta_\star = \beta_\star^\top A_\star DP(M_\star)[Z]^\top DP(M_\star)[Z] A_\star^\top \beta_\star.$$

Writing

$$m_\star := A_\star^\top \beta_\star = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \qquad m_1 \in \mathbb{R}^k, \quad m_2 \in \mathbb{R}^{d-k},$$

we obtain

$$L = m_1^\top R^{-1} Z_{12} Z_{12}^\top R^{-1} m_1 + m_2^\top Z_{12}^\top R^{-2} Z_{12} m_2.$$

In particular, in the aligned basis where $m_\star$ lies entirely in the leading k-dimensional eigenspace, $m_\star = \begin{pmatrix} R^{1/2}\beta_\star \\ 0 \end{pmatrix}$, the second term vanishes and therefore

$$L = \beta_\star^\top R^{-1/2} Z_{12} Z_{12}^\top R^{-1/2} \beta_\star = \|Z_{12}^\top R^{-1/2}\beta_\star\|_2^2 = \sum_{j=1}^{d-k} \left(\sum_{i=1}^{k} \frac{\beta_{\star,i}}{\sqrt{r_i}} (Z_{12})_{ij} \right)^2. \tag{H.38}$$

By the explicit covariance structure of $V_\star$, the entries of Z_{12} are independent centered Gaussians with variances

$$\mathrm{Var}\big((Z_{12})_{ij}\big) = \frac{2(1+\tau+\lambda_i^2)}{a_i b_j}, \qquad 1 \leqslant i \leqslant k, \quad 1 \leqslant j \leqslant d-k.$$

Hence, for each j,

$$Y_j := \sum_{i=1}^{k} \frac{\beta_{\star,i}}{\sqrt{r_i}} (Z_{12})_{ij}$$

is centered Gaussian with variance

$$\mathrm{Var}(Y_j) = \sum_{i=1}^{k} \frac{\beta_{\star,i}^2}{r_i} \frac{2(1+\tau+\lambda_i^2)}{a_i b_j}.$$

Using $r_i = \lambda_i / a_i$, this simplifies to

$$\mathrm{Var}(Y_j) = \frac{2}{b_j} \sum_{i=1}^{k} \beta_{\star,i}^2 \frac{(1+\tau+\lambda_i^2)}{\lambda_i}.$$

Therefore, $L = \sum_{j=1}^{d-k} Y_j^2$ is a weighted chi-square random variable, namely

$$L \overset{d}{=} \sum_{j=1}^{d-k} \sigma_j^2 \chi_j^2(1), \qquad \sigma_j^2 = \frac{2}{b_j} \sum_{i=1}^{k} \beta_{\star,i}^2 \frac{(1+\tau+\lambda_i^2)}{\lambda_i}, \tag{H.39}$$

where the $\chi_j^2(1)$ are independent chi-square random variables with one degree of freedom.

Proof [Proof of Corollary 6.3] The result follows by a direct substitution into the general expression for the pre-training interaction term given in Equation (H.39).

Under the diagonal model specified in Section 6.1, we have

$$b_j = 1, \qquad \beta_{\star,i} = 1, \qquad \lambda_i = \frac{1}{i}, \qquad \text{and} \qquad \tau = \sum_{j=1}^{k} j^{-2}.$$

Substituting these quantities into Equation (H.39), we obtain

$$\sigma_j^2 = 2 \sum_{i=1}^{k} i\big(1 + i^{-2} + \tau\big),$$

which does not depend on j. Therefore, the weighted chi-square random variable reduces to

$$L = \left(2\sum_{i=1}^{k} i\big(1 + i^{-2} + \tau\big)\right) \sum_{j=1}^{d-k} \chi_j^2(1) = \left(2\sum_{i=1}^{k} i\big(1 + i^{-2} + \tau\big)\right) \chi^2_{(d-k)},$$

establishing the stated convergence in distribution.

The scaling of the expectation follows immediately from the above expression, yielding the claimed order.

■

Pre-training asymptotic sub-optimality. Let

$$\mathcal{Q} := \frac{1}{2}\big\langle Z,\, 2\Sigma_{\mathrm{pre}} Z \Sigma_{\mathrm{pre}} \big\rangle_F = \big\langle Z, \Sigma_{\mathrm{pre}} Z \Sigma_{\mathrm{pre}} \big\rangle_F.$$

We note that the population Hessian operator at $M_\star$ is given by the map $H \mapsto \langle H, 2\Sigma_{\mathrm{pre}} H \Sigma_{\mathrm{pre}}\rangle$. A second-order Taylor expansion of L_{spec} around $M_\star$, together with the first-order optimality condition, yields

$$m\big(L_{\mathrm{spec}}(\hat{M}_m) - L_{\mathrm{spec}}(M_\star)\big) \overset{d}{\rightsquigarrow} \mathcal{Q}.$$

We have

$$\begin{aligned}
\mathcal{Q} &= \sum_{i,j=1}^{k} a_i a_j (Z_{11})_{ij}^2 + 2\sum_{i=1}^{k}\sum_{u=1}^{d-k} a_i b_u (Z_{12})_{iu}^2 \\
&= \sum_{i=1}^{k} a_i^2 (Z_{11})_{ii}^2 + 2\sum_{1 \leqslant i < j \leqslant k} a_i a_j (Z_{11})_{ij}^2 + 2\sum_{i=1}^{k}\sum_{u=1}^{d-k} a_i b_u (Z_{12})_{iu}^2.
\end{aligned}$$

We now compute the law of each term.

Let

$$z_{\mathrm{diag}} := \big((Z_{11})_{11}, \ldots, (Z_{11})_{kk}\big)^\top.$$

By assumption,

$$z_{\mathrm{diag}} \sim \mathcal{N}(0, K), \qquad K = \mathrm{Diag}\left(\frac{1 + \tau + 3\lambda_i^2}{a_i^2}\right)_{i=1}^{k} + rr^\top, \qquad r_i = \frac{\lambda_i}{a_i}.$$

Hence

$$\sum_{i=1}^{k} a_i^2 (Z_{11})_{ii}^2 = z_{\mathrm{diag}}^\top \,\mathrm{Diag}(a_1^2, \ldots, a_k^2)\, z_{\mathrm{diag}}.$$

If $\xi \sim \mathcal{N}(0, I_k)$, this quadratic form may be written as

$$z_{\mathrm{diag}}^\top \,\mathrm{Diag}(a_1^2, \ldots, a_k^2)\, z_{\mathrm{diag}} \overset{d}{=} \xi^\top M \xi,$$

where

$$M = \mathrm{Diag}(a_1, \ldots, a_k)\, K\, \mathrm{Diag}(a_1, \ldots, a_k) = \mathrm{Diag}(1 + \tau + 3\lambda_i^2)_{i=1}^k + \lambda\lambda^\top,$$

with $\lambda = (\lambda_1, \dots, \lambda_k)^\top$. Therefore, if $\rho_1, \dots, \rho_k$ denote the eigenvalues of M, then

$$\sum_{i=1}^{k} a_i^2 (Z_{11})_{ii}^2 \overset{d}{=} \sum_{\ell=1}^{k} \rho_\ell \chi_\ell^2(1).$$

For $1 \leqslant i < j \leqslant k$, the variables $(Z_{11})_{ij}$ are independent centered Gaussians with variance

$$\mathrm{Var}\big((Z_{11})_{ij}\big) = \frac{2(1+\tau+\lambda_i^2+\lambda_i\lambda_j+\lambda_j^2)}{a_i a_j}.$$

Hence

$$2a_i a_j (Z_{11})_{ij}^2 \overset{d}{=} 4\big(1+\tau+\lambda_i^2+\lambda_i\lambda_j+\lambda_j^2\big)\chi_{ij}^2(1),$$

where the $\chi_{ij}^2(1)$ are independent.

For $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant d-k$, the variables $(Z_{12})_{ij}$ are independent centered Gaussians with variance

$$\mathrm{Var}\big((Z_{12})_{ij}\big) = \frac{2(1+\tau+\lambda_i^2)}{a_i b_j}.$$

Therefore

$$2a_i b_j (Z_{12})_{ij}^2 \overset{d}{=} 4(1+\tau+\lambda_i^2)\widetilde{\chi}_{ij}^2(1),$$

where the $\widetilde{\chi}_{ij}^2(1)$ are independent. Using Gaussian independence of the orthogonal coordinates, we obtain

$$\mathcal{Q} \overset{d}{=} \sum_{\ell=1}^{k} \rho_\ell \chi_\ell^2(1) + \sum_{1\leqslant i<j\leqslant k} 4\big(1+\tau+\lambda_i^2+\lambda_i\lambda_j+\lambda_j^2\big)\chi_{ij}^2(1) + \sum_{i=1}^{k}\sum_{j=1}^{d-k} 4(1+\tau+\lambda_i^2)\widetilde{\chi}_{ij}^2(1), \quad \text{(H.40)}$$

where $\rho_1, \dots, \rho_k$ are the eigenvalues of

$$M = \mathrm{Diag}(1+\tau+3\lambda_i^2)_{i=1}^k + \lambda\lambda^\top.$$

Taking expectations yields

$$\begin{aligned}\mathbb{E}[\mathcal{Q}] &= \mathrm{tr}(M) + 4\sum_{1\leqslant i<j\leqslant k}\big(1+\tau+\lambda_i^2+\lambda_i\lambda_j+\lambda_j^2\big) + 4\sum_{i=1}^{k}\sum_{u=1}^{d-k}(1+\tau+\lambda_i^2) \\ &= k(4d-2k-1)(1+\tau) + 2(2d-1)\tau + 2(\sum_{i=1}^{k}\lambda_i)^2. \quad \text{(H.41)}\end{aligned}$$

Connection to Cabannes et al. (2023). We compare our result to the generalization bound of Cabannes et al. (2023, Thm. 4). In their result, the contribution of pre-training appears through a product of a conditioning factor and a sub-optimality term, of the form

$$\|T_\lambda^{-1}\Pi_{\mathcal{F}_\lambda} f_\star\|^2 \big(L_{\mathrm{spec}}(\hat{M}) - L_{\mathrm{spec}}(M_\star)\big).$$

We now show how the conditioning factor $\|T_\lambda^{-1}\Pi_{\mathcal{F}_\lambda} f_\star\|^2$ reduces in our setting.

In the present model with $\lambda = 0$, the operator T_λ coincides with the covariance matrix $C = \mathrm{diag}(R, 0)$. Moreover, the projection onto the feature space is given by

$$\Pi_{\mathcal{F}_\lambda} = \mathrm{diag}(I_k, 0).$$

Since the target function satisfies $f_\star(x) = \beta_\star^\top A_\star x$, its projection onto the feature space depends only on the first k coordinates. Therefore, the action of $T_\lambda^{-1}\Pi_{\mathcal{F}_\lambda}$ amounts to inverting R on the top-k subspace, yielding

$$\|T_\lambda^{-1}\Pi_{\mathcal{F}_\lambda} f_\star\|^2 = \|C^{-1}A_\star^\top \beta_\star\|_2^2 = \|R^{-\frac{1}{2}}\beta_\star\|_2^2.$$

Consequently,

$$m\|T_\lambda^{-1}\Pi_{\mathcal{F}_\lambda} f_\star\|^2 \big(L_{\mathrm{spec}}(\hat{M}) - L_{\mathrm{spec}}(M_\star)\big) \overset{d}{\rightsquigarrow} \|R^{-\frac{1}{2}}\beta_\star\|_2^2\, \mathcal{Q}.$$

In the setting where $\Sigma_{\mathrm{pre}} = I_d$, $\Sigma_{\mathrm{pre}}^+ = \mathrm{diag}(1, 1/2, \ldots, 1/d)$, and $\beta_\star = (1, \cdots, 1)^\top$, we have

$$\|R^{-\frac{1}{2}}\beta_\star\|_2^2 = \sum_{i=1}^{k} i = \Theta(k^2),$$

while

$$\mathbb{E}[\mathcal{Q}] = k(4d - 2k - 1)(1 + \tau) + 2(2d - 1)\tau + 2(\sum_{i=1}^{k}\lambda_i)^2 = \Theta\big(k(d-k)\big).$$

Consequently, the expected pre-training contribution satisfies

$$\mathbb{E}\Big[\|R^{-\frac{1}{2}}\beta_\star\|_2^2\, \mathcal{Q}\Big] = \|R^{-\frac{1}{2}}\beta_\star\|_2^2\, \mathbb{E}[\mathcal{Q}] = O\big(k^3(d-k)\big).$$

Thus, the pre-training contribution scales as $O(k^3(d-k))$ in this model.

Appendix I. Proofs of Section 6.2

This appendix verifies that the factor-model example in Section 6.2 satisfies the assumptions of Theorem 5.1. The downstream verification is reduced to the linear-feature argument used in Appendix H: after passing to the quotient descriptor $M = AA^\top$, the feature map can be written as

$$\phi(x, M) = s(M)^\top (I_d + M)^{-1} x = B_M^\top x, \qquad B_M := (I_d + M)^{-1} s(M),$$

where $M \mapsto B_M$ is smooth and B_M has rank k locally around $M_\star$. Consequently, the quotient geometry and Assumptions G.1–G.4 follow by the same linear-feature verification, which we summarize below. The only model-specific pretraining step is to verify the Riemannian M-estimation condition, Assumption D.5, for the quotient estimator $\hat{M}_m$. After these checks, the corollary follows by directly invoking Theorem 5.1.

Throughout this appendix,

$$M_\star := A_\star A_\star^\top, \qquad \Sigma_x := I_d + M_\star,$$

and $A_\star \in \mathbb{R}^{d\times k}$ has full column rank. Thus $M_\star \in \mathcal{M}_{d,k} := \{M \succcurlyeq 0 : \mathrm{rank}(M) = k\}$ and $\Sigma_x \succ 0$.

We now derive the exact maximum likelihood estimator of the quotient-level parameter $M = AA^\top$. Recall the unlabeled pretraining model

$$X = A_\star Z + \mu, \qquad Z \sim \mathcal{N}(0, I_k), \quad \mu \sim \mathcal{N}(0, I_d), \tag{I.1}$$

with Z and μ independent and $k \ll d$. By marginalizing over the latent factor Z, the pretraining covariate X is Gaussian with covariance

$$X \sim \mathcal{N}\big(0,\ \Sigma_\star\big), \qquad \Sigma_\star = A_\star A_\star^\top + I_d = M_\star + I_d, \tag{I.2}$$

where $M_\star = A_\star A_\star^\top$ is a rank-k positive semidefinite matrix.

The negative log-likelihood of the pretraining sample $\{X_i\}_{i=1}^m$ as a function of $M \succcurlyeq 0$ is

$$\hat{L}_m(M) = \frac{1}{2}\Big(\log\det(I_d + M) + \mathrm{tr}\big(S_m(I_d + M)^{-1}\big)\Big), \tag{I.3}$$

up to an additive constant independent of M. Here, we defined $S_m \coloneqq \frac{1}{m}\sum_{i=1}^m X_i X_i^\top$. This maximum likelihood problem is exactly *probabilistic principal component analysis* (PPCA); see Tipping and Bishop (1999).

Let $S_m = U\Lambda U^\top$ be the eigenvalue decomposition of the sample covariance, and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ where $\lambda_1 \geqslant \cdots \geqslant \lambda_d$. Let $U_k \in \mathbb{R}^{d\times k}$ be the k-dimensional leading eigenspace.

Lemma I.1 (PPCA maximum likelihood estimator for M) *For the model $X \sim \mathcal{N}(0, I_d + M_\star)$ with $\mathrm{rank}(M_\star) = k$, a maximum likelihood estimator is*

$$\hat{M}_m \ = \ U_k\big(\mathrm{diag}((\lambda_1 - 1)_+ \cdots, (\lambda_k - 1)_+)\big)U_k^\top. \tag{I.4}$$

Proof See Tipping and Bishop (1999). ■

Lemma I.2 (Downstream reduction for the factor model) *The factor-model quotient feature map*

$$\phi(x, M) = s(M)^\top (I_d + M)^{-1} x$$

satisfies Assumptions G.1–G.4 by the same argument as the linear spectral feature map, with $s(M)$ replaced by $B_M^\top$, where

$$B_M \coloneqq (I_d + M)^{-1} s(M).$$

Proof The map $M \mapsto s(M)$ is C^2 by the local-section construction, and $M \mapsto (I_d + M)^{-1}$ is smooth because $I_d + M \succ 0$ for all $M \succcurlyeq 0$. Hence

$$M \mapsto B_M = (I_d + M)^{-1} s(M)$$

is C^2 locally. Since $s(M)$ has rank k and $I_d + M$ is invertible, B_M has rank k locally. The quotient feature map is therefore a rank-k linear feature map in the covariate:

$$\phi(x, M) = B_M^\top x.$$

All downstream verifications are then identical to the linear spectral case. Indeed, local boundedness of B_M and $D_M B_M$ gives the local-uniform feature and derivative moment bounds from Gaussian moments of X (Lemma H.3). The downstream covariance is

$$\Sigma(M) = \mathbb{E}[\phi(X, M)\phi(X, M)^\top] = B_M^\top \Sigma_x B_M,$$

which is positive definite because B_M has rank k and $\Sigma_x \succ 0$. Thus, the stable rank/eigengap condition holds locally and $d_{\text{eff}}(M) = k$ (Lemma H.4). The leverage and leverage-weighted signal moment bounds follow as in the linear case from

$$q_M(X) \leqslant C\|\phi(X, M)\|^2 \leqslant C\|X\|^2$$

and from the fact that both $f_\star$ and $\Pi_M f_\star$ are linear functions of X with locally bounded coefficients (Lemma H.5). Empirical well-posedness follows because $B_M^\top X$ has a density on $\mathbb{R}^k$, so the feature design has rank k almost surely for $n \geqslant k$. Finally, since $\Sigma(M) \succ 0$, the coefficient null space $\ker(T_M)$ is $\{0\}$, so the stable null-span condition holds trivially (Lemma H.8). ■

Verification of Assumption D.5. Next, we will prove the maximum likelihood estimator in Equation I.4 satisfies Assumption D.5.

Lemma I.3 *Assume $A_\star$ has full column rank, so that $M_\star = A_\star A_\star^\top \in \mathcal{M}_{d,k}$ and $\lambda_k(M_\star) > 0$. Let $L(M) = \mathbb{E}[\hat{L}_m(M)]$. Then Assumption D.5 holds at $M_\star$.*

Proof We verify the five parts of Assumption D.5.

(i) Identification and separation. The population objective is

$$L(M) = \frac{1}{2}\left[\log\det(I_d + M) + \operatorname{tr}\big(\Sigma_x(I_d + M)^{-1}\big)\right], \qquad \Sigma_x = I_d + M_\star.$$

This is the Gaussian negative log-likelihood for covariance $I_d + M$, up to an additive constant. Equivalently,

$$L(M) - L(M_\star) = \frac{1}{2}\left[\operatorname{tr}\big(\Sigma_x(I_d + M)^{-1}\big) - \log\det\big(\Sigma_x(I_d + M)^{-1}\big) - d\right].$$

The bracketed term is nonnegative and vanishes if and only if $I_d + M = \Sigma_x$, i.e. $M = M_\star$. Thus $M_\star$ is the unique minimizer of L on $\mathcal{M}_{d,k}$.

The separation condition follows from uniqueness and coercivity. Indeed, $L(M) \to \infty$ whenever $\|M\|_F \to \infty$ within $\mathcal{M}_{d,k}$, because $\log\det(I_d + M) \to \infty$. If separation failed for some $\epsilon > 0$, there would exist $M_j \in \mathcal{M}_{d,k}$ such that

$$d_{\mathcal{M}}(M_j, M_\star) \geqslant \epsilon, \qquad L(M_j) \downarrow L(M_\star).$$

By coercivity, a subsequence is bounded. Since the PSD cone is closed, a further subsequence converges to some $M_\infty \succcurlyeq 0$. By continuity of the Gaussian likelihood on the PSD cone, $L(M_\infty) = L(M_\star)$, so the uniqueness statement above gives $M_\infty = M_\star$, contradicting $d_{\mathcal{M}}(M_j, M_\star) \geqslant \epsilon$. Hence, for every $\epsilon > 0$,

$$\inf_{M \in \mathcal{M}_{d,k} : d_{\mathcal{M}}(M, M_\star) \geqslant \epsilon} \big(L(M) - L(M_\star)\big) > 0.$$

(ii) Uniform LLN on a compact set. Let $K_{\epsilon'} := \exp_{M_\star}\big(\overline{B}(M_\star, \epsilon')\big)$ be the compact normal-coordinate set appearing in Assumption D.5. For each fixed x, the map

$$M \mapsto \ell_{\mathrm{pre}}(M; x) = \frac{1}{2}\Big[\log\det(I_d + M) + x^\top (I_d + M)^{-1} x\Big]$$

is continuous on $K_{\epsilon'}$. Moreover, since $K_{\epsilon'}$ is compact, $\log\det(I_d + M)$ is uniformly bounded on $K_{\epsilon'}$, and $\|(I_d + M)^{-1}\|_{\mathrm{op}} \leqslant 1$ for all $M \succcurlyeq 0$. Therefore

$$\sup_{M \in K_{\epsilon'}} |\ell_{\mathrm{pre}}(M; x)| \leqslant C_K(1 + \|x\|^2).$$

The right-hand side is integrable because $X \sim \mathcal{N}(0, \Sigma_x)$. Thus, by Newey and McFadden (1994, Lemma 2.4),

$$\sup_{M \in K_{\epsilon'}} |\hat{L}_m(M) - L(M)| \xrightarrow{\mathbb{P}} 0.$$

It remains to verify localization of $\hat{M}_m$. We prove consistency of the PPCA estimator. By the strong law of large numbers, $S_m \to \Sigma_x$ almost surely in operator norm. Fix an outcome in this almost-sure event and write $\Delta_m := S_m - \Sigma_x$. Thus, $\|\Delta_m\|_{\mathrm{op}} \to 0$. Let $P_\star$ denote the orthogonal projector onto the top-k eigenspace of Σ_x, and let $\hat{P}_m := \sum_{j=1}^k \hat{u}_j \hat{u}_j^\top$ be the projector onto the top-k eigenspace of S_m. By the Davis–Kahan $\sin\Theta$ theorem (Davis and Kahan, 1970), for all m large enough,

$$\|\hat{P}_m - P_\star\|_{\mathrm{op}} \leqslant \frac{2\|\Delta_m\|_{\mathrm{op}}}{\lambda_k(\Sigma_x) - \lambda_{k+1}(\Sigma_x)} = \frac{2\|\Delta_m\|_{\mathrm{op}}}{\lambda_k(\Sigma_x) - 1} \longrightarrow 0. \tag{I.5}$$

Thus $\hat{P}_m \to P_\star$ in operator norm, hence also in Frobenius norm.

Let $\gamma := \lambda_k(\Sigma_x) - 1 > 0$. By Weyl's inequality, $|\hat{\lambda}_i - \lambda_i(\Sigma_x)| \leqslant \|\Delta_m\|_{\mathrm{op}}$ for all $i \in [d]$. Hence, for all m large enough so that $\|\Delta_m\|_{\mathrm{op}} \leqslant \gamma/2$,

$$\hat{\lambda}_k \geqslant \lambda_k(\Sigma_x) - \|\Delta_m\|_{\mathrm{op}} \geqslant 1 + \gamma/2.$$

Therefore $\hat{\lambda}_i > 1$ for all $i \leqslant k$, and the $(\cdot)_+$ truncation in the PPCA estimator is inactive.

Define the rank-k truncation map

$$\Pi_k(B) := \sum_{j=1}^k \lambda_j(B) u_j(B) u_j(B)^\top.$$

Then, $\hat{M}_m = \Pi_k(S_m) - \hat{P}_m$. Similarly, since $\lambda_{k+1}(\Sigma_x) = \cdots = \lambda_d(\Sigma_x) = 1$,

$$\Pi_k(\Sigma_x) = M_\star + P_\star, \qquad M_\star = \Pi_k(\Sigma_x) - P_\star.$$

Consequently,

$$\hat{M}_m - M_\star = \big(\Pi_k(S_m) - \Pi_k(\Sigma_x)\big) - \big(\hat{P}_m - P_\star\big). \tag{I.6}$$

The projector term converges to zero by (I.5). For the truncation term, the top-k spectral truncation is continuous at Σ_x because of the eigengap at k. Hence

$$\|\Pi_k(S_m) - \Pi_k(\Sigma_x)\|_F \to 0.$$

Combining this with (I.6) gives $\hat{M}_m \xrightarrow{\mathbb{P}} M_\star$. Therefore, for the compact normal-coordinate set $K_{\epsilon'}$,

$$\mathbb{P}(\hat{M}_m \in K_{\epsilon'}) \to 1$$

after fixing $\epsilon' > 0$ sufficiently small.

(iii) Local C^2 smoothness and score moments. For every x, the map

$$M \mapsto \ell_{\mathrm{pre}}(M; x) = \frac{1}{2}\left[\log\det(I_d + M) + x^\top (I_d + M)^{-1} x\right]$$

is C^2 on a neighborhood of $M_\star$. Its Euclidean gradient is

$$\nabla_M \ell_{\mathrm{pre}}(M; x) = \frac{1}{2}\left[(I_d + M)^{-1} - (I_d + M)^{-1} x x^\top (I_d + M)^{-1}\right].$$

Thus, at $M_\star$,

$$\|\operatorname{grad} \ell_{\mathrm{pre}}(M_\star; X)\|_{M_\star} \leqslant C(1 + \|X\|^2).$$

Since X is Gaussian,

$$\mathbb{E}\left[\|\operatorname{grad} \ell_{\mathrm{pre}}(M_\star; X)\|_{M_\star}^2\right] < \infty.$$

(iv) Nondegenerate minimizer. The population gradient is

$$\nabla L(M) = \frac{1}{2}\left[(I_d + M)^{-1} - (I_d + M)^{-1}\Sigma_x (I_d + M)^{-1}\right],$$

which vanishes at $M = M_\star$. The population Hessian at $M_\star$ has quadratic form

$$D^2 L(M_\star)[H, H] = \frac{1}{2}\operatorname{tr}\left(\Sigma_x^{-1} H \Sigma_x^{-1} H\right), \qquad H \in T_{M_\star}\mathcal{M}_{d,k}.$$

Because $\Sigma_x \succ 0$, this quadratic form is strictly positive for every nonzero $H \in T_{M_\star}\mathcal{M}_{d,k}$. Hence the Riemannian Hessian

$$H_\star = \operatorname{Hess} L(M_\star) : T_{M_\star}\mathcal{M}_{d,k} \to T_{M_\star}\mathcal{M}_{d,k}$$

is positive definite, and in particular invertible.

(v) Uniform transported Hessian convergence. On the compact normal-coordinate set $K_{\epsilon'}$, the transported Hessian entries are continuous functions of M. Moreover, the derivatives of the exponential map, parallel transport, and the local coordinate maps are bounded on $K_{\epsilon'}$, while the first two derivatives of $\ell_{\mathrm{pre}}(M; X)$ are dominated by an integrable envelope of the form $C_K(1 + \|X\|^2)$. Applying Newey and McFadden (1994, Lemma 2.4) entrywise to the finite-dimensional matrix representation of the transported Hessian gives

$$\sup_{M \in K_{\epsilon'}} \left\|\widetilde{H}_m(M) - \widetilde{H}(M)\right\| \xrightarrow{\mathbb{P}} 0,$$

where

$$\widetilde{H}_m(M) = \mathcal{P}_{M \to M_\star} \circ \operatorname{Hess} \hat{L}_m(M) \circ \mathcal{P}_{M_\star \to M},$$

and

$$\widetilde{H}(M) = \mathcal{P}_{M \to M_\star} \circ \operatorname{Hess} L(M) \circ \mathcal{P}_{M_\star \to M}.$$

This verifies Assumption D.5(v). ■

Explicit form of the Gaussian tangent limit. Lemma I.3 gives the abstract descriptor CLT. We now record the explicit form of the limiting Gaussian tangent vector Z, which is used in the evaluation of the pretraining interaction term. Let $X \sim \mathcal{N}(0, \Sigma_x)$ with

$$\Sigma_x = I_d + M_\star, \qquad \operatorname{rank}(M_\star) = k.$$

First, we write the eigendecomposition

$$\Sigma_x = U_\star \begin{pmatrix} I_k + D & 0 \\ 0 & I_{d-k} \end{pmatrix} U_\star^\top, \qquad D = \operatorname{diag}(\sigma_1, \dots, \sigma_k), \;\; \sigma_k > 0,$$

so $M_\star = U_\star \left(\begin{smallmatrix} D & 0 \\ 0 & 0 \end{smallmatrix}\right) U_\star^\top$. Let the sample covariance be $S_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^\top$. Recall the (fixed-$k$) PPCA MLE can be written as

$$\hat{M}_m = \hat{U}_k \, \operatorname{diag}(\hat{\lambda}_1 - 1, \dots, \hat{\lambda}_k - 1) \, \hat{U}_k^\top,$$

where $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_d$ and $\hat{U}_k$ are the top-k eigenpairs of S_m. Since $\lambda_k(\Sigma_x) = 1 + \sigma_k > 1$, we have $\hat{\lambda}_i > 1$ for $i \leqslant k$ with probability $\to 1$, so the "$(\cdot)_+$" truncation is asymptotically inactive and the map is smooth at Σ_x.

The Gaussian fourth-moment identity gives

$$\sqrt{m} \operatorname{vec}(S_m - \Sigma_x) \overset{d}{\rightsquigarrow} \mathcal{N}\big(0, \; (I + K)(\Sigma_x \otimes \Sigma_x)\big),$$

where K is the commutation matrix.

Define the smooth map (near Σ_x)

$$g(\Sigma) \coloneqq \text{"best rank-}k \text{ PSD approximation of } \Sigma - I_d\text{"}, \qquad \hat{M}_m = g(S_m), \quad M_\star = g(\Sigma_x).$$

Then

$$Z_m \coloneqq \sqrt{m} \, (\hat{M}_m - M_\star) \overset{d}{\rightsquigarrow} Z, \qquad Z = D g_{\Sigma_x}[G],$$

where G is the Gaussian matrix limit of $\sqrt{m}(S_m - \Sigma_x)$.

Let $U_\star = [U_1 \; U_2]$ with $U_1 \in \mathbb{R}^{d \times k}$ spanning the signal subspace and $U_2 \in \mathbb{R}^{d \times (d-k)}$ spanning its orthogonal complement, and rotate

$$\tilde{G} \coloneqq U_\star^\top G U_\star = \begin{pmatrix} \tilde{G}_{11} & \tilde{G}_{12} \\ \tilde{G}_{21} & \tilde{G}_{22} \end{pmatrix}.$$

A first-order perturbation calculation (using the eigengap $\sigma_k > 0$) yields the simple derivative

$$D g_{\Sigma_x}[G] = U_\star \begin{pmatrix} \tilde{G}_{11} & \tilde{G}_{12} \\ \tilde{G}_{21} & 0 \end{pmatrix} U_\star^\top.$$

Therefore the asymptotic fluctuation has the explicit form

$$Z = U_\star \begin{pmatrix} \tilde{G}_{11} & \tilde{G}_{12} \\ \tilde{G}_{21} & 0 \end{pmatrix} U_\star^\top, \qquad \tilde{G} = U_\star^\top G U_\star,$$

and in particular Z is mean-zero Gaussian (as a vector in $\mathbb{R}^{d(d+1)/2}$).

If we define the linear operator $\mathcal{P} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ by

$$\mathcal{P}(Y) \coloneqq U_\star \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & 0 \end{pmatrix} U_\star^\top \quad \text{(blocks taken in the } U_\star\text{-basis)},$$

then $\mathrm{vec}(Z) = (\mathcal{P} \otimes I)\mathrm{vec}(G)$ and hence

$$\mathrm{vec}(Z) \sim \mathcal{N}\big(0,\ (\mathcal{P} \otimes I)\,(I + K)(\Sigma_x \otimes \Sigma_x)\,(\mathcal{P} \otimes I)^\top\big).$$

Equivalently, Let $P_\star$ be the orthogonal projector onto $\mathrm{range}(M_\star)$. Then, we have

$$Z \ = \ P_\star G + G P_\star - P_\star G P_\star. \tag{I.7}$$

Let $W \in \mathbb{R}^{d\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries. Thus,

$$G \ \coloneqq \ \frac{1}{\sqrt{2}}\,\Sigma_x^{1/2}(W + W^\top)\Sigma_x^{1/2}. \tag{I.8}$$

Substituting the expression for G gives the explicit representation of Z as a linear transform of a standard Gaussian matrix:

$$Z = \frac{1}{\sqrt{2}}\Big[P_\star\Sigma_x^{1/2}(W + W^\top)\Sigma_x^{1/2} + \Sigma_x^{1/2}(W + W^\top)\Sigma_x^{1/2}P_\star - P_\star\Sigma_x^{1/2}(W + W^\top)\Sigma_x^{1/2}P_\star\Big]. \tag{I.9}$$

Pretraining fluctuations. Fix $\Sigma_x \coloneqq I_d + M_\star$ and write $f_\star(x) = w_\star^\top x$ with

$$w_\star = \Sigma_x^{-1}A_\star\beta_\star.$$

At $M_\star$, choose a local section with $s(M_\star) = A_\star$ and define

$$U_\star \coloneqq (I_d + M_\star)^{-1}A_\star = \Sigma_x^{-1}A_\star.$$

Let $P_\star$ be the Σ_x-orthogonal projector onto $\mathrm{range}(U_\star)$:

$$P_\star = U_\star\big(U_\star^\top\Sigma_x U_\star\big)^{-1}U_\star^\top\Sigma_x.$$

Then $(\Pi_{M_\star}f_\star)(x) = (P_\star^{(\Sigma_x)}w_\star)^\top x$ and $P_\star^{(\Sigma_x)}w_\star = w_\star$.

Recall $\mathcal{L}(v) \coloneqq -D\Pi_{M_\star}[v]\,f_\star$. Since $D\Pi_{M_\star}[v]\,f_\star(x) = (DP_\star^{(\Sigma_x)}[v]\,w_\star)^\top x$, we have

$$\mathcal{L}(v)(x) = -(DP_\star^{(\Sigma_x)}[v]\,w_\star)^\top x.$$

Define for each M

$$U_M \coloneqq (I_d + M)^{-1}s(M) \in \mathbb{R}^{d\times k}, \qquad \mathcal{S}_M \coloneqq \mathrm{range}(U_M),$$

and let $P_M^{(\Sigma_x)}$ denote the Σ_x-orthogonal projector onto $\mathcal{S}_M$. By definition of an orthogonal projector onto $\mathcal{S}_M$, we have the matrix identity

$$P_M^{(\Sigma_x)}U_M = U_M. \tag{I.10}$$

Fix $v \in T_{M_\star}\mathcal{M}_{d,k}$ and consider a smooth curve $M(t)$ in $\mathcal{M}_{d,k}$ with $M(0) = M_\star$ and $\dot{M}(0) = v$. Define $U(t) \coloneqq U_{M(t)}$ and $P(t) \coloneqq P_{M(t)}^{(\Sigma_x)}$. Then (I.10) becomes

$$P(t)\, U(t) = U(t) \qquad \text{for all } t \text{ near } 0.$$

Differentiating at $t = 0$ and applying the product rule gives

$$\dot{P}(0)\, U_\star + P_\star\, \dot{U}(0) = \dot{U}(0),$$

where $U_\star \coloneqq U_{M_\star}$ and $P_\star \coloneqq P_{M_\star}^{(\Sigma_x)}$. Rearranging yields the key identity

$$\dot{P}(0)\, U_\star = (I_d - P_\star)\, \dot{U}(0). \tag{I.11}$$

Interpreting $\dot{P}(0) = DP_\star^{(\Sigma_x)}[v]$ and $\dot{U}(0) = DU_\star[v]$, we can rewrite (I.11) as

$$DP_\star^{(\Sigma_x)}[v]\; U_\star = (I_d - P_\star)\; DU_\star[v]. \tag{I.12}$$

Finally, since in our model $w_\star \in \mathcal{S}_{M_\star}$ we can write

$$w_\star = U_\star \beta_\star$$

for the same $\beta_\star$ appearing in $f_\star(x) = \beta_\star^\top U_\star^\top x$. Multiplying (I.12) on the right by $\beta_\star$ gives

$$DP_\star^{(\Sigma_x)}[v]\; w_\star = DP_\star^{(\Sigma_x)}[v]\; U_\star \beta_\star = (I_d - P_\star)\; DU_\star[v]\; \beta_\star. \tag{I.13}$$

Moreover $U_M = (I_d + M)^{-1} s(M)$ implies

$$DU_\star[v]\, \beta_\star = -\Sigma_x^{-1} v\, \Sigma_x^{-1} A_\star \beta_\star + \Sigma_x^{-1} Ds(M_\star)[v]\, \beta_\star.$$

The section-dependent term does not change $\mathrm{range}(U_M)$ and is killed by $(I_d - P_\star^{(\Sigma_x)})$. Consequently,

$$DP_\star^{(\Sigma_x)}[v]\, \Sigma_x^{-1} A_\star \beta_\star = -(I_d - P_\star^{(\Sigma_x)}) \Sigma_x^{-1} v\, \Sigma_x^{-1} A_\star \beta_\star,$$

and therefore, for $v = Z$,

$$\mathcal{L}(Z)(x) = x^\top (I_d - P_\star^{(\Sigma_x)}) \Sigma_x^{-1} Z\, \Sigma_x^{-1} A_\star \beta_\star.$$

Define the coefficient vector

$$g(Z) \;\coloneqq\; (I_d - P_\star^{(\Sigma_x)}) \Sigma_x^{-1} Z\, \Sigma_x^{-1} A_\star \beta_\star \in \mathbb{R}^d,$$

so that $\mathcal{L}(Z)(x) = x^\top g(Z)$. Then

$$\|\mathcal{L}(Z)\|_{L^2(\mu_{\mathrm{down}})}^2 = \mathbb{E}\big[(x^\top g(Z))^2\big] = g(Z)^\top \Sigma_x\, g(Z).$$

Equivalently, expanding $g(Z)$,

$$\|\mathcal{L}(Z)\|_{L^2(\mu_{\mathrm{down}})}^2 = \Big(\Sigma_x^{-1} A_\star \beta_\star\Big)^\top Z^\top \Sigma_x^{-1}\, (I_d - P_\star^{(\Sigma_x)})^\top \Sigma_x (I_d - P_\star^{(\Sigma_x)})\, \Sigma_x^{-1} Z \Big(\Sigma_x^{-1} A_\star \beta_\star\Big).$$

Since $P_\star^{(\Sigma_x)}$ is Σ_x-self-adjoint and idempotent,

$$(P_\star^{(\Sigma_x)})^\top \Sigma_x = \Sigma_x P_\star^{(\Sigma_x)}, \qquad (P_\star^{(\Sigma_x)})^2 = P_\star^{(\Sigma_x)},$$

we have the simplification

$$(I_d - P_\star^{(\Sigma_x)})^\top \Sigma_x (I_d - P_\star^{(\Sigma_x)}) = \Sigma_x (I_d - P_\star^{(\Sigma_x)}).$$

Moreover, at $M_\star$, we have $P_\star^{(\Sigma_x)} = P_\star$ is the orthogonal projection onto the span of $\mathrm{image}(M_\star)$. Therefore,

$$\begin{aligned}\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\mathrm{down}})} &= \left(\Sigma_x^{-1} A_\star \beta_\star\right)^\top Z^\top \Sigma_x^{-1} \left(I_d - P_\star\right) \Sigma_x^{-1} Z \left(\Sigma_x^{-1} A_\star \beta_\star\right) \\ &= \|\left(I_d - P_\star\right) \Sigma_x^{-1} Z \left(\Sigma_x^{-1} A_\star \beta_\star\right)\|^2 = \|\left(I_d - P_\star\right) Z \left(\Sigma_x^{-1} A_\star \beta_\star\right)\|^2. \end{aligned} \tag{I.14}$$

Since $\Sigma_\star^{-1}$ acts as identity on $I - P_\star$. Now, we plug Eq. (I.7) into Equation (I.14) to get

$$\begin{aligned}\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\mathrm{down}})} &= \|\left(I_d - P_\star\right) \left(P_\star G + G P_\star - P_\star G P_\star\right) \left(\Sigma_x^{-1} A_\star \beta_\star\right)\|^2 \\ &= \|\left(I_d - P_\star\right) G P_\star \Sigma_x^{-1} A_\star \beta_\star\|^2. \end{aligned} \tag{I.15}$$

Finally, we plug Equation (I.8) into Equation (I.15)

$$\begin{aligned}\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\mathrm{down}})} &= \|(I_d - P_\star)\,(\frac{1}{\sqrt{2}}\,\Sigma_x^{1/2}(W + W^\top)\Sigma_x^{1/2}) P_\star \Sigma_x^{-1} A_\star \beta_\star\|^2 \\ &= \frac{1}{2}\|(I_d - P_\star)\,(W + W^\top) P_\star \Sigma_x^{-1/2} A_\star \beta_\star\|^2. \end{aligned} \tag{I.16}$$

Let $M_\star = U_\star \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} U_\star^\top$ with $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ and $U_\star = [U_1\ U_2]$, where $U_1 \in \mathbb{R}^{d \times k}$ spans $\mathrm{range}(M_\star)$. Then

$$P_\star = U_1 U_1^\top = U_\star \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} U_\star^\top, \qquad \Sigma_x = I_d + M_\star = U_\star \begin{pmatrix} I_k + D & 0 \\ 0 & I_{d-k} \end{pmatrix} U_\star^\top,$$

so

$$\Sigma_x^{-1/2} = U_\star \begin{pmatrix} (I_k + D)^{-1/2} & 0 \\ 0 & I_{d-k} \end{pmatrix} U_\star^\top.$$

Rotate the standard Gaussian matrix by $\tilde{W} \coloneqq U_\star^\top W U_\star$ (still i.i.d. $\mathcal{N}(0,1)$). Then

$$(I_d - P_\star)(W + W^\top) P_\star \Sigma_x^{-1/2} A_\star \beta_\star = U_\star \begin{pmatrix} 0 & 0 \\ (\tilde{W}_{21} + \tilde{W}_{12}^\top) & 0 \end{pmatrix} U_\star^\top\, U_\star \begin{pmatrix} (I_k + D)^{-1/2} & 0 \\ 0 & 0 \end{pmatrix} U_\star^\top A_\star \beta_\star.$$

Since $A_\star \beta_\star \in \mathrm{range}(U_1)$, letting $a \coloneqq U_1^\top A_\star \beta_\star \in \mathbb{R}^k$ gives

$$(I_d - P_\star)(W + W^\top) P_\star \Sigma_x^{-1/2} A_\star \beta_\star = U_2\,(\tilde{W}_{21} + \tilde{W}_{12}^\top)\,(I_k + D)^{-1/2}\,a.$$

Because U_2 is orthonormal, $\|U_2 v\|_2 = \|v\|_2$, hence the original quantity simplifies to

$$\frac{1}{2}\left\|(I_d - P_\star)(W + W^\top) P_\star \Sigma_x^{-1/2} A_\star \beta_\star\right\|_2^2 = \frac{1}{2}\left\|(\tilde{W}_{21} + \tilde{W}_{12}^\top)\,(I_k + D)^{-1/2}\,a\right\|_2^2, \qquad a = U_1^\top A_\star \beta_\star.$$

In particular, since $\tilde{W}_{21}$ has i.i.d. $\mathcal{N}(0,1)$ entries and $\tilde{W}_{12}$ is independent with the same law, the matrix $\frac{1}{\sqrt{2}}(\tilde{W}_{21}+\tilde{W}_{12}^\top)$ has i.i.d. $\mathcal{N}(0,1)$ entries. Therefore

$$\frac{1}{2}\big\|(\tilde{W}_{21}+\tilde{W}_{12}^\top)\,(I_k+D)^{-1/2}\,a\big\|_2^2 \;\overset{d}{=}\; \big\|\Xi\,(I_k+D)^{-1/2}\,U_1^\top A_\star\,\beta_\star\big\|_2^2, \qquad \Xi_{ij}\overset{\text{iid}}{\sim}\mathcal{N}(0,1).$$

Equivalently, we have

$$\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\text{down}})} \overset{d}{=} \|(I_k+D)^{-1/2}\,U_1^\top A_\star\,\beta_\star\|_2^2\,\chi^2_{d-k}. \tag{I.17}$$

Appendix J. Proofs of Section 6.3

This appendix verifies that the Gaussian-mixture example from Section 6.3 satisfies the hypotheses of Theorem 5.1, and hence yields Corollary 6.6. The verification is modular: (i) local-uniform moment bounds for the feature map, (ii) rank stability/eigengap for the population feature second-moment $\Sigma(M)$, (iii) a manifold CLT for the (pretraining) descriptor estimator $\underline{U}_m = D(\hat{U}_m)$ via an M-estimation CLT and a delta method, and (iv) verification of the hypotheses of the general projector differentiability result proved in Appendix G.2.

Throughout, we work on the regular set from Assumption 6.5 and write $\underline{U}_\star := D(U^\star)$.

J.1. Model, features, and the quotient-level descriptor

Unlabeled distribution. Let $K \geqslant 2$ and $d \geqslant 1$. The unlabeled distribution is the spherical Gaussian mixture

$$X \mid (Z=i) \sim \mathcal{N}(u_i^\star, I_d), \qquad Z \sim \text{Unif}([K]),$$

with unknown centers $U^\star = (u_1^\star,\dots,u_K^\star) \in (\mathbb{R}^d)^K$.

Centered-mean subspace. Define the empirical mean and centered second-moment matrix

$$\bar{u}(U) := \frac{1}{K}\sum_{i=1}^K u_i, \qquad S(U) := \sum_{i=1}^K \big(u_i-\bar{u}(U)\big)\big(u_i-\bar{u}(U)\big)^\top.$$

Let $r_\star := r(U^\star)$ and define P_U to be the orthogonal projector onto the leading $r_\star$-dimensional eigenspace of $S(U)$ (on the regular neighborhood where this eigenspace is well-defined). Write $P_\star := P_{U^\star}$ and $V_\star := \text{Im}(P_\star)$.

Subspace-aware responsibilities. For $i \in [K]$, define

$$\pi_i(x;U) = \frac{\exp\big(\langle P_U u_i, P_U x\rangle - \frac{1}{2}\|P_U u_i\|_2^2\big)}{\sum_{j=1}^K \exp\big(\langle P_U u_j, P_U x\rangle - \frac{1}{2}\|P_U u_j\|_2^2\big)}.$$

These satisfy $0 \leqslant \pi_i \leqslant 1$ and $\sum_{i=1}^K \pi_i = 1$.

Feature map and hypothesis class. The induced feature map has dimension $p(U) = K(r(U)+1)$:

$$\psi_U(x) = \Big(\pi_1(x;U)\big(P_U x - P_U u_1\big),\, \pi_1(x;U),\dots,\pi_K(x;U)\big(P_U x - P_U u_K\big),\, \pi_K(x;U)\Big).$$

Let $\mathcal{H}_U = \{\langle\theta,\psi_U(\cdot)\rangle : \theta \in \mathbb{R}^{p(U)}\}$ and $\Pi_U : L^2(\mu_{\text{down}}) \to L^2(\mu_{\text{down}})$ be the orthogonal projector onto $\mathcal{H}_U$.

Group action and quotient parameter. Fix $U = (u_1, \dots, u_K) \in (\mathbb{R}^d)^K$. The unlabeled mixture likelihood is invariant under relabeling of mixture components. Accordingly, we consider the action of the permutation group S_K on $(\mathbb{R}^d)^K$ defined by

$$(g \cdot U)_i := u_{g^{-1}(i)}, \qquad g \in S_K.$$

We write

$$\underline{U} := [U] \ \in \ (\mathbb{R}^d)^K / S_K$$

for the equivalence class (orbit) of U under this action, and refer to $\underline{U}$ as the *quotient-level parameter*.

Regular regime and local lifts. Throughout this appendix we restrict attention to the regular regime in which the centers are distinct:

$$u_i \neq u_j \qquad \text{for all } i \neq j.$$

In this regime the action of S_K is free, and the quotient $(\mathbb{R}^d)^K / S_K$ is a smooth manifold of dimension Kd. In particular, for any fixed $\underline{U}^\star = [U^\star]$ there exists a neighborhood $\mathcal{U}$ and a smooth local section

$$s : \mathcal{U} \to (\mathbb{R}^d)^K, \qquad s(\underline{U}) \in \underline{U},$$

which amounts to choosing a deterministic ordering of the centers in a neighborhood of $U^\star$. All constructions below are independent of the specific choice of section.

Permutation-invariant geometric quantities. Both $\bar{u}(U)$ and P_U are invariant under the action of S_K, and therefore depend only on the quotient parameter $\underline{U}$. We may thus unambiguously write $\bar{u}(\underline{U})$ and $P_{\underline{U}}$.

Responsibilities and equivariant feature map. For a representative $U \in \underline{U}$, define the projected responsibilities

$$\pi_i(x; U) := \frac{\exp\big(\langle P_U u_i,\ P_U x\rangle - \frac{1}{2}\|P_U u_i\|_2^2\big)}{\sum_{j=1}^K \exp\big(\langle P_U u_j,\ P_U x\rangle - \frac{1}{2}\|P_U u_j\|_2^2\big)}, \qquad i \in [K].$$

Let $\psi_U : \mathbb{R}^d \to \mathbb{R}^{K(r_\star+1)}$ denote the block-structured feature map

$$\psi_U(x) = \Big(\pi_1(x;U)(P_U x - P_U u_1),\ \pi_1(x;U);\ \dots;\ \pi_K(x;U)(P_U x - P_U u_K),\ \pi_K(x;U)\Big).$$

For any $g \in S_K$, let $\rho(g)$ denote the block-permutation matrix acting on $\mathbb{R}^{K(r_\star+1)}$. A direct computation shows that the feature map is S_K-equivariant:

$$\psi_{g\cdot U}(x) = \rho(g)\, \psi_U(x).$$

To be more precise, $\rho(G)$ consists of those permutations in S_{2K} that relabel $\pi_i(x, U)$ and $\pi_i(x; U)(P_U x - P_U u_i)$ two the same label. Therefore, we can define a feature map $\phi(\cdot, \underline{U})$ on the quotient manifold that satisfies all the assumptions of Appendix E.2.

J.2. Local-uniform moment bounds for $\psi(\cdot, \underline{U})$

We verify the local-uniform moment bounds from Assumption G.2 for this model (with $X \sim \mu_{\mathrm{down}}$ equal to the Gaussian mixture described in the example).

Lemma J.1 (Local-uniform polynomial moments of the Gaussian-mixture features(Assumption G.2) *Fix a compact neighborhood $\mathcal{U}$ of $U^\star$ in $(\mathbb{R}^d)^K$. Then for every $q \geqslant 1$,*

$$\sup_{U \in \mathcal{U}} \mathbb{E}\big[\|\psi_U(X)\|^q\big] < \infty.$$

Consequently, for every neighborhood $\underline{\mathcal{U}}$ of $\underline{U}_\star = [U^\star]$ contained in the quotient image $[\mathcal{U}] \subset (\mathbb{R}^d)^K/S_K$ and every $q \geqslant 1$

$$\sup_{\underline{U} \in \underline{\mathcal{U}}} \mathbb{E}\big[\|\psi_{s(\underline{U})}(X)\|^q\big] < \infty,$$

where $s : \underline{\mathcal{U}} \to (\mathbb{R}^d)^K$ is any local section (lift) with $s(\underline{U}) \in \underline{U}$.

Proof Fix $U \in \mathcal{U}$. Since $0 \leqslant \pi_i(x; U) \leqslant 1$ and $\|P_U x\|_2 \leqslant \|x\|_2$, we have for each i

$$\big\|\pi_i(X; U)\big(P_U X - P_U u_i\big)\big\|_2 \leqslant \|X\|_2 + \|u_i\|_2, \qquad |\pi_i(X; U)| \leqslant 1.$$

Thus there exists a constant $C < \infty$ depending only on K such that

$$\|\psi_U(X)\| \leqslant C\Big(1 + \|X\|_2 + \max_{i \in [K]} \|u_i\|_2\Big).$$

Taking q-th moments and using $\sup_{U \in \mathcal{U}} \max_i \|u_i\|_2 < \infty$, it remains to note that all polynomial moments of X are finite under a spherical Gaussian mixture, giving the stated uniform bound.

For the quotient-level statement, let $\underline{\mathcal{U}} \subset [\mathcal{U}]$ and let s be a local section on $\underline{\mathcal{U}}$. Then $s(\underline{U}) \in \mathcal{U}$ for all $\underline{U} \in \underline{\mathcal{U}}$, so the same bound applies uniformly to $\psi_{s(\underline{U})}(X)$. ∎

J.3. Rank stability and eigengap for the downstream second moment

Define the population second moment (feature covariance)

$$\Sigma(\underline{U}) \;\coloneqq\; \mathbb{E}\big[\phi_{\underline{U}}(X)\phi_{\underline{U}}(X)^\top\big] \in \mathbb{R}^{p \times p},$$

where p is the feature dimension. To invoke Theorem 5.1, we need rank stability and an eigengap at 0 for $\Sigma(\underline{U})$ on a neighborhood of $\underline{U}_\star$. We reduce this to (a) continuity of $\underline{U} \mapsto \Sigma(\underline{U})$ and (b) an eigengap at $\underline{U}_\star$.

Lemma J.2 (Continuity of $\underline{U} \mapsto \Sigma(\underline{U})$ in operator norm) *Let $\underline{\mathcal{U}}$ be a comapct neighborhood of $\underline{U}_\star$ such that the conclusion of Lemma J.1 holds on $s(\underline{\mathcal{U}})$ for some exponent $4 + \delta$. Then $\underline{U} \mapsto \Sigma(\underline{U})$ is continuous at $\underline{U}_\star$ in operator norm.*

Proof Fix $\underline{U} \in \underline{\mathcal{U}}$ and write $Y_{\underline{U}} \coloneqq \phi_{\underline{U}}(X)\phi_{\underline{U}}(X)^\top$. By Lemma J.1 (with exponent $1 + \delta/4$) and Cauchy–Schwarz,

$$\sup_{\underline{U} \in \underline{\mathcal{U}}} \mathbb{E}\big[\|Y_{\underline{U}}\|_{\mathrm{op}}^{1+\delta/4}\big] \;\leqslant\; \sup_{\underline{U} \in \underline{\mathcal{U}}} \mathbb{E}\big[\|\phi_{\underline{U}}(X)\|^{2+\delta/2}\big] \;<\; \infty,$$

so $\{\|Y_{\underline{U}}\|_{\mathrm{op}} : \underline{U} \in \underline{\mathcal{U}}\}$ is uniformly integrable. Since $\phi_{\underline{U}}$ is a continuous function of $\underline{U}$ in $\underline{\mathcal{U}}$, $Y_{\underline{U}} \to Y_{\underline{U}_\star}$ almost surely as $\underline{U} \to \underline{U}_\star$. Uniform integrability then yields

$$\big\|\Sigma(\underline{U}) - \Sigma(\underline{U}_\star)\big\|_{\mathrm{op}} = \Big\|\mathbb{E}\big[Y_{\underline{U}} - Y_{\underline{U}_\star}\big]\Big\|_{\mathrm{op}} \leqslant \mathbb{E}\big[\|Y_{\underline{U}} - Y_{\underline{U}_\star}\|_{\mathrm{op}}\big] \to 0,$$

proving operator-norm continuity at $\underline{U}_\star$. ■

Lemma J.3 (Local uniform invertibility from full rank at $\underline{U}_\star$) *Assume $\Sigma(\underline{U}_\star) \succ 0$. If $\underline{U} \mapsto \Sigma(\underline{U})$ is continuous at $\underline{U}_\star$ in operator norm (e.g. by Lemma J.2), then there exist a neighborhood $\underline{\mathcal{U}}$ of $\underline{U}_\star$ and a constant $\kappa > 0$ such that for all $\underline{U} \in \underline{\mathcal{U}}$,*

$$\lambda_{\min}\big(\Sigma(\underline{U})\big) \geqslant \kappa.$$

In particular, $\Sigma(\underline{U})$ is invertible on $\underline{\mathcal{U}}$ and $\Sigma(\underline{U})^+ = \Sigma(\underline{U})^{-1}$.

Proof Since $\Sigma(\underline{U}_\star) \succ 0$, we have $\lambda_{\min}(\Sigma(\underline{U}_\star)) > 0$. Let $\kappa := \frac{1}{2}\lambda_{\min}(\Sigma(\underline{U}_\star))$. By continuity in operator norm, for $\underline{U}$ close enough to $\underline{U}_\star$ we have $\|\Sigma(\underline{U}) - \Sigma(\underline{U}_\star)\|_{\mathrm{op}} \leqslant \kappa$. Weyl's inequality then yields

$$\lambda_{\min}\big(\Sigma(\underline{U})\big) \geqslant \lambda_{\min}\big(\Sigma(\underline{U}_\star)\big) - \|\Sigma(\underline{U}) - \Sigma(\underline{U}_\star)\|_{\mathrm{op}} \geqslant 2\kappa - \kappa = \kappa,$$

proving the claim. ■

J.4. Verification of Assumption D.5

We now verify the Riemannian M-estimation condition, Assumption D.5, for the quotient estimator $\underline{\hat{U}}_m = [\hat{U}_m]$. We first record the score/Hessian formulas and the local consistency input, and then verify the five parts of Assumption D.5 one by one.

Pretraining estimator. For concreteness, let $\hat{U}_m$ be a (measurable) local minimizer of the empirical negative log-likelihood for the mixture with equal weights and known covariance:

$$\ell(U; x) = -\log\Big(\frac{1}{K}\sum_{i=1}^{K}\exp\big(-\tfrac{1}{2}\|x - u_i\|_2^2\big)\Big), \qquad \hat{U}_m \in \arg\min_U \frac{1}{m}\sum_{j=1}^{m}\ell(U; X_j).$$

Define the quotient estimator $\underline{\hat{U}}_m := [\hat{U}_m] \in (\mathbb{R}^d)^K/S_K$.

Lemma J.4 (Score and Hessian for the spherical equal-weight MoG likelihood) *Let*

$$p_U(x) := \frac{1}{K}\sum_{i=1}^{K}\varphi_d(x - u_i), \qquad \varphi_d(t) := (2\pi)^{-d/2}\exp\Big(-\tfrac{1}{2}\|t\|_2^2\Big), \qquad \ell(U; x) := -\log p_U(x).$$

Define the responsibilities

$$\pi_i(x; U) := \frac{\varphi_d(x - u_i)}{\sum_{j=1}^{K}\varphi_d(x - u_j)} = \frac{\exp(-\frac{1}{2}\|x - u_i\|_2^2)}{\sum_{j=1}^{K}\exp(-\frac{1}{2}\|x - u_j\|_2^2)}, \qquad i \in [K].$$

Then $\ell(U;x)$ is C^∞ in U, and the gradient blocks satisfy

$$\nabla_{u_i}\ell(U;x) = -\pi_i(x;U)\,(x-u_i), \qquad i \in [K].$$

Moreover, the block Hessian $\nabla^2_{u_iu_j}\ell(U;x) \in \mathbb{R}^{d\times d}$ is given by

$$\nabla^2_{u_iu_j}\ell(U;x) = \begin{cases} \pi_i(x;U)\,I_d - \pi_i(x;U)\big(1-\pi_i(x;U)\big)\,(x-u_i)(x-u_i)^\top, & i=j,\\ \pi_i(x;U)\pi_j(x;U)\,(x-u_i)(x-u_j)^\top, & i\neq j. \end{cases}$$

In particular, for any bounded set $\mathcal{U} \subset (\mathbb{R}^d)^K$ there exists $C<\infty$ (depending on $\mathcal{U}$ and K) such that for all $U\in\mathcal{U}$ and all $x\in\mathbb{R}^d$,

$$\|\nabla\ell(U;x)\|_2 \leqslant C\big(1+\|x\|_2\big), \qquad \|\nabla^2\ell(U;x)\|_{\mathrm{op}} \leqslant C\big(1+\|x\|_2^2\big).$$

Proof The smoothness of $U\mapsto\ell(U;x)$ follows from $p_U(x)>0$ and smoothness of $u\mapsto\varphi_d(x-u)$. Write $Z(U;x) \coloneqq \sum_{j=1}^K \varphi_d(x-u_j)$ so that $\ell(U;x) = -\log Z(U;x) + \log K$. Since $\nabla_{u_i}\varphi_d(x-u_i) = \varphi_d(x-u_i)(x-u_i)$, we have

$$\nabla_{u_i}\ell(U;x) = -\frac{\nabla_{u_i}Z(U;x)}{Z(U;x)} = -\frac{\varphi_d(x-u_i)}{\sum_{j=1}^K\varphi_d(x-u_j)}(x-u_i) = -\pi_i(x;U)(x-u_i).$$

For the Hessian, first note that

$$\nabla_{u_j}\pi_i(x;U) = \pi_i(x;U)\big(\delta_{ij}-\pi_j(x;U)\big)(x-u_j),$$

which follows by differentiating the softmax form of π_i. Using $\nabla_{u_j}(x-u_i) = -\delta_{ij}I_d$, we obtain

$$\nabla^2_{u_iu_j}\ell(U;x) = -\nabla_{u_j}\big(\pi_i(x;U)(x-u_i)\big) = -\big(\nabla_{u_j}\pi_i(x;U)\big)(x-u_i)^\top + \pi_i(x;U)\delta_{ij}I_d.$$

Substituting the expression for $\nabla_{u_j}\pi_i$ yields the displayed block formulas for $i=j$ and $i\neq j$.

Finally, the bounds use $0\leqslant\pi_i\leqslant 1$ and, on a bounded set $\mathcal{U}$, the uniform bound $\max_i\|u_i\|_2 \leqslant C_{\mathcal{U}}$:

$$\|\nabla_{u_i}\ell(U;x)\|_2 \leqslant \|x-u_i\|_2 \leqslant \|x\|_2 + C_{\mathcal{U}},$$

and each Hessian block is a sum of terms bounded by 1 and $\|x-u_i\|_2\|x-u_j\|_2$, hence by $C(1+\|x\|_2^2)$ uniformly over $U\in\mathcal{U}$. Summing over the $K\times K$ blocks gives the stated operator-norm bound. ■

Lemma J.5 (Local uniqueness of the MLE after fixing the permutation symmetry) *Let $L(U) = \mathbb{E}[\ell(U;X)]$ and $\hat{L}_m(U) = \frac{1}{m}\sum_{j=1}^m \ell(U;X_j)$, where $X\sim p_{U^\star}$. Fix a gauge-fixing set $\Theta\subset(\mathbb{R}^d)^K$ that removes the permutation symmetry locally (e.g. a local ordering rule), so that $U^\star\in\Theta$ and no nontrivial permutation of $U^\star$ belongs to Θ. Assume:*

(i) if $p_U = p_{U^\star}$ then U equals a permutation of $U^\star$ (identifiability up to permutation);

(ii) L is twice differentiable on Θ and there exist $\mu>0$ and $\rho>0$ such that $\nabla^2 L(U) \succcurlyeq \mu I$ for all $U\in\Theta\cap B(U^\star,\rho)$;

(iii) $\sup_{U\in\Theta\cap B(U^\star,\rho)}\|\nabla^2\hat{L}_m(U) - \nabla^2 L(U)\|_{\mathrm{op}} \to 0$ in probability.

Then, with probability tending to one, $\hat{L}_m$ is $\mu/2$-strongly convex on $\Theta \cap B(U^\star, \rho)$ and has a unique minimizer there, denoted $\hat{U}_m$, and $\hat{U}_m \to U^\star$ in probability.

Proof This is identical to the standard strong-convexity plus uniform-Hessian-consistency argument; we include it here for completeness. By (iii), with probability tending to one,

$$\sup_{U \in \Theta \cap B(U^\star, \rho)} \left\| \nabla^2 \hat{L}_m(U) - \nabla^2 L(U) \right\|_{\mathrm{op}} \leqslant \mu/2.$$

On this event, (ii) implies $\nabla^2 \hat{L}_m(U) \succcurlyeq (\mu/2) I$ for all $U \in \Theta \cap B(U^\star, \rho)$, hence $\hat{L}_m$ is $\mu/2$-strongly convex there and has a unique minimizer. Consistency follows from (i) together with uniform convergence of $\hat{L}_m$ to L on $\Theta \cap \overline{B}(U^\star, \rho)$, which is implied by the same regularity underlying (iii) and the polynomial tail control of X under the mixture. ■

Lemma J.6 *Assume Assumption 6.5 and the conditions of Lemma J.5 for a gauge-fixing set Θ. Let*

$$F(\underline{U}) := \mathbb{E}[\ell(s(\underline{U}); X)], \qquad \hat{F}_m(\underline{U}) := \frac{1}{m} \sum_{j=1}^{m} \ell(s(\underline{U}); X_j),$$

where s is any local section of the quotient map near $\underline{U}_\star = [U^\star]$. Then Assumption D.5 holds for the quotient objective F at $\underline{U}_\star$.

Proof We verify the five parts of Assumption D.5.

(i) Identification and separation. The population objective is the cross-entropy

$$F(\underline{U}) = \mathbb{E}_{X \sim p_{U^\star}} \big[-\log p_{s(\underline{U})}(X) \big].$$

Equivalently,

$$F(\underline{U}) - F(\underline{U}_\star) = \mathrm{KL}\big(p_{U^\star} \,\|\, p_{s(\underline{U})}\big) \geqslant 0.$$

Equality holds if and only if $p_{s(\underline{U})} = p_{U^\star}$. By the identifiability condition in Lemma J.5, this happens if and only if $s(\underline{U})$ equals a permutation of $U^\star$, i.e. $\underline{U} = \underline{U}_\star$. Thus $\underline{U}_\star$ is the unique minimizer of F on the quotient.

The separation condition follows from the same identifiability and the local gauge fixing. Indeed, if for some $\epsilon > 0$ the separation failed, there would exist $\underline{U}_j$ such that

$$d(\underline{U}_j, \underline{U}_\star) \geqslant \epsilon, \qquad F(\underline{U}_j) \downarrow F(\underline{U}_\star).$$

Choosing the local gauge representative whenever $\underline{U}_j$ lies near $\underline{U}_\star$, the compactness of local sublevel sets and continuity of the likelihood give a limit point $\underline{U}_\infty$ satisfying $F(\underline{U}_\infty) = F(\underline{U}_\star)$. By the uniqueness just proved, $\underline{U}_\infty = \underline{U}_\star$, contradicting the distance lower bound. Hence, for every $\epsilon > 0$,

$$\inf_{\underline{U}:\, d(\underline{U}, \underline{U}_\star) \geqslant \epsilon} \big(F(\underline{U}) - F(\underline{U}_\star) \big) > 0.$$

(ii) Uniform LLN on a compact set. Let $K_{\epsilon'} := \exp_{\underline{U}_\star}\left(\overline{B}(\underline{U}_\star, \epsilon')\right)$ be the compact normal-coordinate set appearing in Assumption D.5. On this set, choose a smooth local section s. For each fixed x, the map $\underline{U} \mapsto \ell(s(\underline{U}); x)$ is continuous. Moreover, the centers $s(\underline{U})$ remain in a bounded subset of $(\mathbb{R}^d)^K$ as $\underline{U}$ ranges over $K_{\epsilon'}$. Hence there exists $C_K < \infty$ such that

$$\sup_{\underline{U} \in K_{\epsilon'}} |\ell(s(\underline{U}); x)| \leqslant C_K(1 + \|x\|^2).$$

The right-hand side is integrable under the spherical Gaussian mixture. Thus, by Newey and McFadden (1994, Lemma 2.4),

$$\sup_{\underline{U} \in K_{\epsilon'}} |\hat{F}_m(\underline{U}) - F(\underline{U})| \xrightarrow{\mathbb{P}} 0.$$

It remains to check localization. By Lemma J.5, after fixing the local gauge, the empirical likelihood has a unique local minimizer $\hat{U}_m$ with probability tending to one and $\hat{U}_m \xrightarrow{\mathbb{P}} U^\star$. Therefore

$$\underline{\hat{U}}_m = [\hat{U}_m] \xrightarrow{\mathbb{P}} [U^\star] = \underline{U}_\star.$$

Consequently, $\mathbb{P}\left(\underline{\hat{U}}_m \in K_{\epsilon'}\right) \to 1$ for every sufficiently small fixed $\epsilon' > 0$.

(iii) Local C^2 smoothness and score moments. By Lemma J.4, $U \mapsto \ell(U; x)$ is C^∞ for every x. Since the local section s is smooth, the quotient-coordinate map $\underline{U} \mapsto \ell(s(\underline{U}); x)$ is C^2 on a normal neighborhood of $\underline{U}_\star$.

The Riemannian gradient in quotient coordinates is obtained from the Euclidean score in the local gauge by a smooth change of coordinates. On bounded gauge neighborhoods, Lemma J.4 gives

$$\|\nabla \ell(U; x)\|_2 \leqslant C(1 + \|x\|).$$

Therefore,

$$\|\operatorname{grad} \ell(\underline{U}_\star; X)\|_{\underline{U}_\star} \leqslant C(1 + \|X\|).$$

Since X has finite moments of all orders under the Gaussian mixture,

$$\mathbb{E}\left[\|\operatorname{grad} \ell(\underline{U}_\star; X)\|^2_{\underline{U}_\star}\right] < \infty.$$

Thus Assumption D.5(iii) holds.

(iv) Nondegenerate minimizer. By Lemma J.5, in the gauge-fixed coordinates there exist $\mu > 0$ and $\rho > 0$ such that

$$\nabla^2 L(U) \succcurlyeq \mu I \qquad \text{for all } U \in \Theta \cap B(U^\star, \rho).$$

In particular, the Hessian at $U^\star$ is positive definite in the gauge-fixed coordinates. Since the gauge chart is a smooth local coordinate chart for the quotient manifold, this is equivalent to positive definiteness of the Riemannian Hessian

$$H_\star = \operatorname{Hess} F(\underline{U}_\star) : T_{\underline{U}_\star}\left((\mathbb{R}^d)^K / S_K\right) \to T_{\underline{U}_\star}\left((\mathbb{R}^d)^K / S_K\right).$$

Thus $H_\star$ is invertible.

(v) Uniform transported Hessian convergence. On the compact normal-coordinate set $K_{\epsilon'}$, the local section, the exponential map, and parallel transport are smooth with uniformly bounded derivatives. Therefore the transported Hessian entries are finite-dimensional linear combinations of the Euclidean Hessian entries in the gauge chart with smooth bounded coefficients.

By Lemma J.4, on bounded gauge neighborhoods,

$$\|\nabla^2 \ell(U; X)\|_{\mathrm{op}} \leqslant C(1 + \|X\|^2),$$

and the right-hand side is integrable under the Gaussian mixture. Hence, again by Newey and McFadden (1994, Lemma 2.4), applied entrywise to the transported Hessian matrix,

$$\sup_{\underline{U} \in K_{\epsilon'}} \left\| \widetilde{H}_m(\underline{U}) - \widetilde{H}(\underline{U}) \right\| \xrightarrow{\mathbb{P}} 0,$$

where

$$\widetilde{H}_m(\underline{U}) = \mathcal{P}_{\underline{U} \to \underline{U}_\star} \circ \operatorname{Hess} \hat{F}_m(\underline{U}) \circ \mathcal{P}_{\underline{U}_\star \to \underline{U}},$$

and

$$\widetilde{H}(\underline{U}) = \mathcal{P}_{\underline{U} \to \underline{U}_\star} \circ \operatorname{Hess} F(\underline{U}) \circ \mathcal{P}_{\underline{U}_\star \to \underline{U}}.$$

This is exactly Assumption D.5(v). ■

J.5. Fréchet differentiability of $\underline{U} \mapsto \Pi_{\underline{U}}$

The master theorem requires Fréchet differentiability of $\underline{U} \mapsto \Pi_{\underline{U}}$ at $\underline{U}_\star$ in operator norm on $L^2(\mu_{\mathrm{down}})$. We do not re-prove this here; instead we verify the hypotheses of the general differentiability result proved in Appendix G.2.

Lemma J.7 (Verification of the projector differentiability hypotheses) *Let $\mathcal{V}$ be a neighborhood of $\underline{U}_\star$ on which Lemma J.1 holds. Then:*

(i) In the local chart fixing the permutation ambiguity, $\underline{U} \mapsto \phi(\cdot, \underline{U})$ is C^1 on $\mathcal{V}$.

(ii) There exist $\delta > 0$ and $C_{\partial\phi} < \infty$ such that

$$\sup_{\underline{U} \in \mathcal{V}} \mathbb{E}\big[\|D_{\underline{U}} \phi(X, \underline{U})\|_{\mathrm{op}}^{4+\delta}\big] \leqslant C_{\partial\phi}.$$

(iii) If, additionally, $\Sigma(\underline{U})$ has stable rank/eigengap on $\mathcal{V}$ (as in Lemma J.3), then $\underline{U} \mapsto \Pi_{\underline{U}}$ is Fréchet differentiable at $\underline{U}_\star$ in operator norm.

Proof (i) On the regular set, $U \mapsto P_U$ and $U \mapsto (P_U u_i)$ are smooth, and $(x, U) \mapsto \pi_i(x; U)$ is a softmax of smooth functions. In a local chart around $U^\star$ fixing the permutation ambiguity, these objects become smooth functions of $\underline{U}$, hence $\underline{U} \mapsto \phi(x, \underline{U})$ is C^1 for each x.

(ii) Differentiating ϕ produces finite sums of terms involving π_i, $D\pi_i$, P_U, and DP_U, multiplied by x and the bounded centers. On any bounded neighborhood of $U^\star$, these derivatives are bounded by polynomials in $\|x\|_2$, while $0 \leqslant \pi_i \leqslant 1$. Since X under the mixture has finite moments of all orders, we obtain the stated bound for some $\delta > 0$.

(iii) This is exactly the implication of the general projector differentiability result in Appendix G.2 once the moment bounds and the stable-rank/eigengap condition are in place. ■

J.6. Proof of Corollary 6.6

Proof [Proof of Corollary 6.6] Under Assumption 6.5 and the model-specific verifications above:

- Lemma J.1 gives the local-uniform moment bounds in Assumption G.2.
- Lemma J.2 and Lemma J.3 give rank stability/eigengap for $\Sigma(M)$ on a neighborhood of $M_\star$.
- Lemma J.6 yields the manifold CLT $Z_m = \sqrt{m}\,\log_{M_\star}(\hat{M}_m) \overset{d}{\rightsquigarrow} Z \sim \mathcal{N}(0, V)$.
- Lemma J.7 verifies the hypotheses needed to invoke the general differentiability theorem for $M \mapsto \Pi_M$ from Appendix G.2.

Assuming compatibility $\mathrm{Rep}(M_\star) = 0$, the hypotheses of the master theorem in the compatible regime (Theorem 5.1 in the main text) hold for this model. Applying that theorem along any joint limit $(m, n) \to (\infty, \infty)$ with $m/n \to \alpha \in (0, \infty)$ yields the claimed limit in Corollary 6.6, with $\mathcal{L}(v) = -D\Pi_{M_\star}[v] f_\star$. ■

Appendix K. Experiment Details

This section provides the full setup underlying Figure 2. Our goal is to numerically evaluate, in the synthetic Gaussian-mixture model of Section 6.3, the pretraining interaction quantity

$$\mathbb{E}\Big[\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\mathrm{down}})}\Big],$$

and to study how it depends on the number of mixture components K, the ambient dimension d, and the separation parameter β.

Gaussian-mixture model. We consider an equally weighted K-component Gaussian mixture in $\mathbb{R}^d$ with identity covariance and component means

$$U^\star = \beta I_{K,d},$$

where $I_{K,d}$ denotes the $K \times d$ rectangular identity matrix. Thus the kth component is centered at βe_k, for $k = 1, \dots, K$, and in particular we require $d \geqslant K$.

Downstream signal. The downstream target is taken to be linear in the projected feature map, with a block structure aligned with the mixture components. For each $i \in [K]$, we associate a parameter block $(\theta_i^\star, b_i^\star)$, with block magnitude proportional to $1/i$. Equivalently, the concatenated parameter vector

$$(\theta_1^\star, b_1^\star, \dots, \theta_K^\star, b_K^\star)$$

is proportional to

$$\left(\mathbf{1}_{d+1}, \frac{1}{2}\mathbf{1}_{d+1}, \dots, \frac{1}{K}\mathbf{1}_{d+1}\right),$$

and is then normalized to have unit Euclidean norm.

Monte Carlo procedure. For each parameter triple (K, d, β), we estimate $\mathbb{E}[\|\mathcal{L}(Z)\|^2_{L^2(\mu_{\text{down}})}]$ using three Monte Carlo stages:

1. estimation of the Fisher information matrix;
2. estimation of the projection coefficients defining the downstream predictor;
3. estimation of the final quadratic quantity using fresh evaluation samples.

In our main experiments, we use $100{,}000$ samples for the Fisher-information estimate, $1{,}000{,}000$ samples for the projection step, and $1{,}000{,}000$ fresh samples for the evaluation step.

Descriptive fits and observed trends. To summarize the empirical scaling behavior, we overlay simple descriptive fits in each panel. In panel (a), the dependence on K is monotone increasing and concave over the plotted range, and is well captured by a quadratic fit $aK^2 + bK + c$ with $R^2 = 0.99$. In panel (b), the dependence on d shows slow growth and is reasonably described by a logarithmic fit $a + b \log d$ with $R^2 = 0.89$. In panel (c), the dependence on β decays rapidly and is well summarized by a power-law fit $C\beta^\alpha$ with $R^2 = 0.90$. These fits should be interpreted as compact summaries of the numerical trends rather than as formal asymptotic claims.

Finite-sample distributional validation. We also simulate the full two-stage procedure in order to check the distributional prediction of Theorem 5.1. We fix a representative GMM instance with

$$K = 4, \qquad d = 20, \qquad \beta = 2, \qquad \alpha := m/n = 2.$$

For each downstream sample size n, we set $m = \lfloor \alpha n \rfloor$, draw m unlabeled pretraining samples from the mixture, estimate the centers by EM, draw an independent downstream design $X_{1:n}$, and compute the conditional scaled excess risk

$$n\big(R(D^{(m)}_{\text{pre}}, X_{1:n}) - \sigma^2\big).$$

The conditional expectation over downstream label noise is computed analytically using the exact conditional risk decomposition according to Proposition B.1, rather than by repeatedly resampling labels. Thus, each Monte Carlo repetition produces the total scaled risk together with its finite-sample variance and pretraining components.

The asymptotic comparison is made against the limiting random variable

$$E_\alpha = \sigma^2 d_{\text{eff}}(\Omega_\star) + \alpha^{-1}\|\mathcal{L}(Z)\|^2_{L^2}, \qquad Z \sim \mathcal{N}(0, H_\star^{-1}\Sigma_\star H_\star^{-1}).$$

For the pretraining component alone, the corresponding limiting law is $\alpha^{-1}\|\mathcal{L}(Z)\|^2_{L^2}$, whereas the variance contribution converges in probability to the deterministic limit $\sigma^2 d_{\text{eff}}(\Omega_\star)$. Figure 3 displays empirical CDFs for a representative subset of sample sizes from a logarithmic grid. The dashed black curves denote the asymptotic laws. For visual clarity, the CDF panels show only a subset of the simulated n-values, while the quantitative metrics below are computed on the full logarithmic grid. For the finite-sample validation, we use 1000 repetitions per n, 2×10^4 Gaussian draws for the asymptotic CDFs, 10^5 samples for the Fisher estimate, and 2×10^5 samples for each projection, evaluation, and test-risk computation.

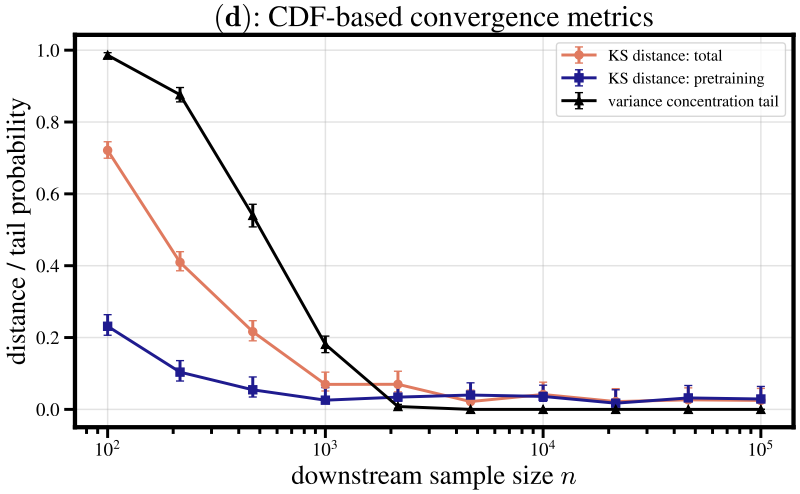

Figure 4: CDF-based convergence metrics for the GMM distributional validation. We plot the Kolmogorov distances $D_{\mathrm{KS}}^{\mathrm{tot}}(n)$ and $D_{\mathrm{KS}}^{\mathrm{pre}}(n)$ between the empirical CDFs and the corresponding asymptotic CDFs for the total scaled excess risk and pretraining contribution. Since the scaled variance contribution has the degenerate limit $\sigma^2 d_{\mathrm{eff}}(\Omega_\star)$, we instead plot the concentration probability $p_{\mathrm{var}}(n;\varepsilon)$. Metrics are computed over the logarithmic grid $n \in \mathrm{round}\big(\mathrm{geomspace}(100, 10^5, 10)\big)$.

Distributional metrics. To quantify the convergence observed in the CDF plots, we report Kolmogorov distances for the non-degenerate limiting distributions. For the total scaled excess risk, let $\widehat{F}_{n,\mathrm{tot}}$ denote the empirical CDF of $n\big(R(D_{\mathrm{pre}}^{(m)}, X_{1:n}) - \sigma^2\big)$ over repeated two-stage simulations, and let $\widehat{F}_{\infty,\mathrm{tot}}$ denote the Monte Carlo CDF of the limiting law E_α. We define

$$D_{\mathrm{KS}}^{\mathrm{tot}}(n) := \sup_t \left|\widehat{F}_{n,\mathrm{tot}}(t) - \widehat{F}_{\infty,\mathrm{tot}}(t)\right|.$$

Similarly, for the pretraining component, we define

$$D_{\mathrm{KS}}^{\mathrm{pre}}(n) := \sup_t \left|\widehat{F}_{n,\mathrm{pre}}(t) - \widehat{F}_{\infty,\mathrm{pre}}(t)\right|,$$

where $\widehat{F}_{n,\mathrm{pre}}$ denotes the empirical CDF of $n\mathrm{Rep}(\Omega_m)$ and $\widehat{F}_{\infty,\mathrm{pre}}$ denotes the Monte Carlo CDF of $\alpha^{-1}\|\mathcal{L}(Z)\|_{L^2}^2$.

For the variance contribution, the limiting distribution is the point mass at $\sigma^2 d_{\mathrm{eff}}(\Omega_\star)$. Since the limiting CDF is discontinuous, an ordinary Kolmogorov distance to this point mass is not a stable convergence metric. Instead, we report the concentration probability

$$p_{\mathrm{var}}(n;\varepsilon) := \widehat{\mathbb{P}}\left(\left|V_{m,n} - \sigma^2 d_{\mathrm{eff}}(\Omega_\star)\right| > \varepsilon\right),$$

where $V_{m,n}$ denotes the scaled variance contribution and we take

$$\varepsilon = 0.05\,\sigma^2 d_{\mathrm{eff}}(\Omega_\star).$$

The quantities $D_{\mathrm{KS}}^{\mathrm{tot}}(n)$, $D_{\mathrm{KS}}^{\mathrm{pre}}(n)$, and $p_{\mathrm{var}}(n;\varepsilon)$ are reported in Figure 4. Decreasing values across the logarithmic n-grid provide a quantitative summary of the distributional convergence shown in Figure 3.

Appendix L. AI Tool Usage

OpenAI's ChatGPT was used to assist in identifying relevant papers during the literature review, brainstorm high-level proof strategies, proofread mathematical arguments for clarity, edit portions of the manuscript, and assist in writing code for experiments. Anthropic's Claude was used to assist with generating schematic illustrations.